Universidad Nacional Pedro Ruiz Gallo
Facultad de Ingeniería Civil, de Sistemas y Arquitectura
Escuela Profesional de Ingeniería de Sistemas

TESIS

R y Python experiencia en el aprendizaje de estadística descriptiva en la Universidad Nacional Pedro Ruiz Gallo

Para Obtener el Título Profesional de:

Ingeniera de Sistemas

Gayoso Rojas, Ynés Jossely
Autor

Dr. Ing. Diaz Plaza, Regis Jorge Alberto
Asesor

Lambayeque – Perú
2024

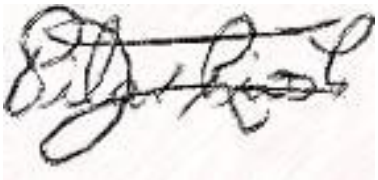
Universidad Nacional Pedro Ruiz Gallo
Facultad de Ingeniería Civil, de Sistemas y Arquitectura
Escuela Profesional de Ingeniería de Sistemas

TESIS

R y Python experiencia en el aprendizaje de estadística descriptiva en la UniversidadNacional Pedro Ruiz Gallo

Para Obtener el Título Profesional de
Ingeniera de Sistemas

Aprobado por los Miembros de Jurado:



Mg. Ing. Ríos Campos, Pilar del Rosario
Presidente de Jurado

Mg. Ing. Guzmán Valle, María de los Ángeles
Secretario

Ing. Guzmán Valle, César Augusto
Vocal

Lambayeque – Perú
2024

Universidad Nacional Pedro Ruiz Gallo
Facultad de Ingeniería Civil, de Sistemas y Arquitectura
Escuela Profesional de Ingeniería de Sistemas

TESIS

R y Python experiencia en el aprendizaje de estadística descriptiva en la Universidad Nacional Pedro Ruiz Gallo

Para Obtener el Título Profesional de
Ingeniera de Sistemas

Gayoso Rojas, Ynés Jossely
Autor

Dr. Ing. Diaz Plaza, Regis Jorge Alberto
Asesor

Lambayeque – Perú
2024

Dedicatoria

A mis padres, por su apoyo incondicional, gracias por desear y anhelar lo mejor para mi vida sin ustedes nada sería posible.

A mi esposo, tu amor y apoyo han sido la base de nuestro hogar. Esta tesis es un tributo a la colaboración, paciencia y comprensión que me has brindado a lo largo de este viaje académico.

A mi hijo, Ramón Stefano por ser mi motor y motivo para seguir adelante y ser mi fuerza para seguir cumpliendo mis sueños.

Gracias a la vida por este nuevo triunfo, gracias a todas las personas que me apoyaron y creyeron en mí.

Y, por último, me lo dedico a mí, porque los logros no son suerte, es esfuerzo y dedicación.

Agradecimiento

A la Universidad Nacional Pedro Ruiz Gallo por su invaluable apoyo en mi formación académica.

A mi asesor por su dedicación y paciencia, sin sus palabras y correcciones precisas no hubiese podido lograr llegar a esta instancia tan anhelada. Gracias por su guía y todos sus consejos.

Son muchos los docentes que han sido parte de mi camino universitario, y a todos ellos les quiero agradecer por transmitirme los conocimientos necesarios para hoy poder estar aquí.

RESUMEN

En la investigación se abordó el problema de mejorar el proceso de aprendizaje de técnicas de ciencia de datos a partir de la selección de un lenguaje de programación, con el planteamiento del objetivo de describir la experiencia de los estudiantes utilizando los lenguajes de programación R y Python, en el proceso de aprendizaje de técnicas de ciencia de datos.

El análisis se realizó mediante la evaluación del cálculo de las medidas de tendencia central, medidas de dispersión y la contrastación de hipótesis descriptivas, para lo cual se elaboró un caso que se desarrolló en un taller dirigido, se utilizó el cuestionario UEQ+ para el análisis descriptivo de las herramientas utilizadas Colab para Python y Posit CLOUD para R, a los que agregaron las preguntas para que el estudiante evalué si logro determinar los indicadores descriptivos, se utilizaron técnicas no paramétricas.

Palabras Claves : La ciencia de datos en la Universidad Nacional Pedro Ruiz Gallo

ABSTRACT

The research addressed the problem of improving the learning process of data science techniques from the selection of a programming language, with the objective of describing the experience of students using the programming languages R and Python. , in the process of learning data science techniques.

The analysis was carried out by evaluating the calculation of the measures of central tendency, measures of dispersion and the contrast of descriptive hypotheses, for which a case was developed that was developed in a directed workshop, the UEQ+ questionnaire was used for the descriptive analysis Of the tools used Colab for Python and Posit CLOUD for R, to which they added the questions for the student to evaluate whether they were able to determine the descriptive indicators, non-parametric techniques were used..

Keywords : Data science at the Pedro Ruiz Gallo National University

INDICE

INDICE DE TABLAS

INDICE DE FIGURAS

INTRODUCCION

Síntesis de la situación problemática

La ciencia de datos en la Universidad Nacional Pedro Ruiz Gallo

La Universidad Nacional Pedro Ruiz Gallo, tiene cuarenta y cinco programas (45) académicos a nivel de pregrado agrupadas en 14 facultades, de los cuales treinta y ocho (38) programas tienen el curso de Estadística o Estadística General, veintisiete (27) programas tienen al menos un curso adicional relacionado con el campo de Estadística, tres (3) programas no tienen curso de Estadística o relacionado, y dos (02) programas tienen al menos un curso relacionado con ciencias de datos.

Tabla 1
Programas con cursos de Estadística, cursos relacionados con estadística y Ciencia de datos

Facultad	Programa	Estadística General	Cursos relacionados con Estadística	Curso relacionado con ciencia de datos
Facultad de Agronomía	Agronomía	1	1	
Facultad de Biología	Biología		1	
	Botánica		1	
	Microbiología		1	
	Pesquería		1	
	Ciencias biológicas	1	1	
Facultad de Ciencias Económicas, Administrativas y Contables	Administración	1	2	
	Comercio y negocios internacionales	1	2	
	Contabilidad	1		

	Economía	1	2	
Facultad de ciencias físicas y matemáticas	Computación e Informática	1	1	
	Estadística	2	3	2
	Física	1	2	
	Matemáticas	1	1	
	Ingeniería electrónica	1		
Facultad de Ciencias Históricas Sociales y Educación	Arqueología	1		
	Artes plásticas	1		
	Teatro	1		
	Pedagogía Artística	1	1	
	Música	1		
	Danzas	1		
	Ciencias de la Comunicación	1	1	
	Psicología	1	2	
	Sociología	1	1	
	Educación inicial	1	1	
	Educación primaria	1	1	
	Ciencias naturales	1		
	Ciencias Históricos Sociales y Filosofía	1		
	Lengua y Literatura	1		
	Idiomas extranjeros	1		
	Matemática y computación	1	2	1
	Educación Física	1		
Facultad de Derecho y Ciencias Políticas	Derecho			
	Ciencias políticas	1		
Facultad de Enfermería	Enfermería			
Facultad de Ingeniería Agrícola	Ingeniería Agrícola	1		
Facultad de Ingeniería Civil, de Sistemas y Arquitectura	Ingeniería Civil	1	1	
	Arquitectura	1	1	
	Ingeniería de Sistemas	1	2	

Facultad de Ingeniería Mecánica y Eléctrica	Ingeniería Mecánica y Eléctrica	1	
Facultad de Ingeniería Química e Industrias Alimentarias	Ingeniería Química	1	1
	Industrias Alimentarias	1	1
Facultad de Medicina Humana	Medicina Humana		2
Facultad de Medicina Veterinaria	Medicina Veterinaria	1	1
Facultad de Zootecnia	Zootecnia		1

Tabla 2
Resumen de cantidad de programas con cursos Estadística, cursos relacionados con Estadística y Ciencia de datos

Categoría de Programas	**Cantidad**
Programas con curso de estadística	38
Programas con cursos relacionados con estadística	25
Programas con cursos relacionado con ciencia de datos	2
Programas sin curso de estadística	2

En la tabla 02 se observa que solamente 4.4% de los programas tiene un curso de ciencia de datos, es decir, en la Universidad Nacional Pedro Ruiz Gallo, también se observó que solamente se dictan tres (03) cursos relacionados con la ciencia de datos: Estadística computacional en R, y Estadística computacional en Python en el programa de Estadística y Estadística con R en el programa de Educación especialidad Matemática y Computación.

En la Universidad Nacional Pedro Ruiz Gallo, en la mayoría de los programas, las técnicas para las investigaciones consideran solamente el uso de datos tradicionales, es decir, datos utilizados recopilados por técnicas cuantitativas y

cualitativas, almacenados en formatos fijos. Las técnicas utilizadas para las investigaciones de pre grado y post grado en la Universidad Nacional Pedro Ruiz Gallo, son las tradicionales, y no se consideran las técnicas de ciencia de datos.

No se ha observado o no existe, un plan para la incorporación de estas técnicas en los diversos programas que ofrece la Universidad Nacional Pedro Ruiz Gallo.

Formulación del Problema Científico

A partir de las técnicas para el tratamiento de los datos las que se están utilizando en el mundo respecto a las técnicas utilizadas en la UNPRG, se formula el problema:

¿Cómo mejorar el proceso de aprendizaje de técnicas de ciencia de datos a partir de la selección de un lenguaje de programación?

Objetivos de la Investigación

Objetivo General

Con atención al problema formulado, la intervención sobre el objeto de estudio se realizará mediante un experimento, planteando el objetivo general:

Describir la experiencia vivida por los estudiantes utilizando los lenguajes de programación R y Python, en el proceso de aprendizaje de técnicas de ciencia de datos.

Objetivos Específicos

Analizar si el lenguaje R permite el aprendizaje de las técnicas descriptivas estadísticas.

Analizar si el lenguaje Python permite el aprendizaje de las técnicas descriptivas estadísticas.

CAPITULO I DISEÑO TEÓRICO

1.1 Antecedentes

La ciencia de datos a nivel mundial y el rol de la universidad

En el 2017, David Dohoho, presenta la historia de los análisis de datos, recordando que Jhon Tukey hace más de 50 años iniciaba el pedido de una reforma de las estadísticas académicas, una nueva ciencia "análisis de datos", reforzado por los pedidos de John Chambers, Jeff Wu, Bill Cleveland y Leo Breiman a que la estadística tradicional expandiera sus límites teóricos, cambiando inclusive el énfasis de los modelos estadísticos a la preparación y presentación de datos, y de la inferencia a la predicción, a lo que Cleveland y Wu proponen el nombre de "ciencia de datos". Es en el 2015, universidades como UC Berkeley, NYU, MIT y principalmente la Universidad de Michigan en el 2015, realizaron una inversión de cien (100) millones de dólares para la enseñanza de nuevos programas con uso de tecnología (Donoho, 2017).

La ciencia de datos se ha convertido en los principales términos en el campo empresarial, industrial o académico, esto está respaldo por el incremento del número de investigaciones y patentes cada año, incluyendo que en el 90% de los datos disponibles a utilizar se genera en los dos últimos años, cambiando drásticamente el mundo de la ciencia y de los negocios (Ley & Bordas, 2018).

El problema definido como "el desperdicio inherente tanto de datos de formato tradicional como de big data, que está siendo producido por la digitalización de la prestación de servicios institucionales, públicos y privados" en el África,

concluye con la importancia de la participación de las universidades para que asuma el nuevo rol como centro de datos abiertos, con el propósito de obtener datos abiertos que garanticen la calidad, confiabilidad, oportunidad y relevancia para las necesidades de la información,(Mutuku, 2019).

El interés por la aplicación de las técnicas de ciencia de datos, ha conducido a evaluar cuales son las tecnologías que servirían para un entorno de aprendizaje, los lenguajes de programación, como lo señala (Ozgur et al., 2021) MatLab se puede utilizar para enseñar matemáticas introductorias como cálculo, tanto Python como R se pueden usar para tomar decisiones involucrando grandes datos, Python es perfecto para enseñar conceptos básicos estadísticas en un entorno rico en datos mientras que R es un poco más complicado.

La selección del lenguaje de programación es una decisión muy importante hoy en día, y se realizan investigaciones para determinar las ventajas y desventajas de cada uno, (López, 2020) realizó un análisis entre los lenguajes Python y R enfocado en el desarrollo de aplicaciones para ciencia de datos, desarrollando casos de prueba y comparando los resultados en base al tiempo de ejecución, donde R tuvo mejor tiempo de ejecución.

La Universidad de la Rioja (Unir, 2020), presentó un análisis sobre ¿cuál es mejor para el análisis de datos, Python o R, encontrando diferencia entre Python y R, sin embargo, sobe las diferencias en el momento de aprender un lenguaje de programación destaca que, Python es mejor para principiantes, Python es

multipropósito , Python es escalable y Python es un lenguaje de programación orientado a objetos, R es para el ámbito académico, R cuenta con una sólida comunidad, R es intuitivo y Visualización de datos.

Los antecedentes presentados muestran la importancia la ciencia de datos en el tratamiento de los datos, y el constante crecimiento del número de investigaciones que se realizan sobre este campo de estudio.

1.2 Bases Teóricas

Estadística

El Dpto. de Estadística e Investigación Operativa de la Universidad de Alicante presenta el concepto de estadística como “método científico que mediante el análisis matemático nos permite obtener información sobre la realidad que nos rodea” (D.S. Gómez-Reverte, 2022), “El término estadística refiere a datos numéricos, tales como promedios, medianas, porcentajes e índices que caracterizar un objeto de estudio o de referencia”(Rafael Agacino, 2022).

Medidas de tendencia central

Las medidas de tendencia central describen en un número a un conjunto de valores de una población, intenta describir que tan cercanos al centro están distribuidos los valores, los parámetros de medida de tendencia central son:

Media; de forma correcta sería promedio aritmético, permite determinar el valor central de la distribución de los elementos

Mediana; permite determinar el valor que ocupa la mitad de las posiciones en los datos, previamente ordenados de mayor a menor.

Moda; el valor que más se repite en una muestra o población.

Medidas de dispersión

Las medidas de dispersión permiten describir la separación, conocido como variabilidad, de los datos respecto a la media, describe la distribución de los datos y permite analizar y una correcta selección de las técnicas de comparación de grupos.

Rango; es la diferencia entre el valor mayor y el valor menor de la muestra o población, describe la distancia entre los extremos.

Varianza; describe la distancia desde los puntos de datos hacia la media de la muestra o población, siempre considera las distancias como positivas.

Desviación Estándar; es la raíz cuadrada de la varianza, también describe la distancia desde los puntos de datos hacia la media de la muestra o población, pero considera las unidades de la muestra o población.

Medidas de simetría

Las medidas de simetría permiten describir la forma que toma la distribución de los datos alrededor de la media.

Coeficiente de asimetría; permite describir la simetría de la distribución de los datos alrededor de media, el resultado indica asimetrías a la derecha, izquierda de la

media, o una simetría a ambos lados que permite considerar a la distribución de los datos como una distribución normal.

Curtosis, describe que tan agrupados están los datos respecto a la media, pudiendo tender a un lejano agrupamiento platicúrtica, una alta concentración leptocúrtica o una concentración moderada mesocúrtica.

Hipótesis descriptiva

Son suposiciones que son comprobadas mediante comparaciones, pero que solamente permiten describir el cambio o diferencia de las características de la muestra o población, sin determinar si existe alguna relación entre la causa que origina el cambio o la diferencia y el resultado obtenido.

Ciencia de datos

Sobre la definición de la ciencia de datos no se concluido con precisión cual es el objeto de estudio, sin embargo, se describe el campo de acción, como la propuesta de (Skiena, 2017) al considerarla como la intersección entre informática, la estadística y los dominios de aplicación real.

(Donoho, 2017) sobre las actividades de tratamiento de datos, acoge la propuesta de Chambers, a una diferencia entre "menor ciencia de datos" (LDS) y el campo más grande "mayor ciencia de datos" (GDS), considerando a GDS como las actividades de la ciencia de datos: Recopilación, preparación y exploración de datos, Representación y transformación de datos, Informática con datos, Visualización y presentación de datos, Modelamiento de datos y Ciencia sobre Ciencia de Datos.

Sobre R

Existen diversos conceptos sobre lo que es el lenguaje R, como “El lenguaje R es un software libre que se usa en big data de forma habitual gracias, entre otros aspectos, a su manejabilidad y coherencia” (IMF, 2021, Pág. 1.), “R es un lenguaje de programación que se ha convertido en uno de los más usados para la Ciencia de Datos, así como una herramienta recurrente para las empresas dedicadas al análisis de datos y finanzas como Google, Facebook, Microsoft, Ford Motors, John Deere, Lloyds, entre otras” (Shinde et al., 2017).

Sobre Python

Python es un lenguaje de programación de propósito general, lo que significa que cualquiera puede usar el lenguaje y modificarlo para adaptarlo a sus necesidades específicas. Los dominios de aplicación de Python van desde desarrollo web, programación de teléfonos móviles y educación a GUI de escritorio, desarrollo de software y aplicaciones comerciales. (Python Software Foundation, 2021).

UEQ+

UEQ (User Experience Questionnaire) en español significa “Cuestionario de Experiencia de Usuario”, a su vez, UEQ +, es una extensión del cuestionario de experiencia del usuario (UEQ, ver Laugwitz, Schrepp & Held, 2008).

El UEQ determina la experiencia del usuario con seis módulos (atractivo, eficiencia, visibilidad, confiabilidad, estimulación y novedad). (Martin Schrepp J.T., 2019).

En el UEQ + tiene más escalas de experiencia de usuario, el investigador podría seleccionar las escalas sean importantes, UEQ+ no se definie como un cuestionario Experiencia Usuario, si no vendría hacer un tipo de herramienta constructor de cuestionarios concretos que se ajustan a los escenarios especiales de evaluación.

Tabla 3
UEQ+ Dimensiones UEQ+

Dimensión	Semántica	Atributos	
Atractivo	impresión general del producto. ¿Les gusta o no les gusta a los usuarios?	Desagradable Malo Incómodo Antipático	Agradable Bueno Cómodo Simpático
Eficiencia	El usuario tiene la sensación relativa que puede obtener las metas relacionadas con un menor esfuerzo del uso del producto	Lento Ineficiente No pragmático Ordenado	Rápido Eficiente Pragmático Sobrecargado
Claridad	El usuario tiene la sensación relativa de que es claro el aprendizaje y entendimiento del uso del producto	No entendible Difícil de aprender Complicado Confuso	Entendible Fácil de aprender Fácil Claro
Confianza	El usuario está seguro que datos están en buenas manos y no se utilizan indebidamente	Inseguro No confiable Dudoso Opaco	Seguro Confiable Fiable Transparente

Estímulo	El usuario tiene la sensación de que usar el producto es sugestivo y conmovedor. Es satisfactorio manejar y laborar con él	Aburrido No interesante Activante De poco valor	Emocionante Interesante Adormecedor Valioso
Novedad	La estructura de la apreciación del usuario observa un producto como nuevo y original	Sin imaginación Convencional Habitual Conservador	Creativo Inventivo Novedoso Innovador

CAPITULO II DISEÑO METODOLOGICO

2.1 Tipo de Investigación

La investigación realizada fue de tipo descriptiva, tecnológica, y pre experimental.

Tabla 4
Tabla de descripción de variable

Título	Variable de estudio	Objetivo General	Hipótesis general	Objetivos específicos	Hipótesis específicas	Hipótesis específicas
R y Python experiencia en el aprendizaje de estadística descriptiva en la Universidad Nacional Pedro Ruiz Gallo	Aprendizaje de estadística descriptiva	Describir la experiencia vivida por los estudiantes utilizando los lenguajes de programación R y Python, en el proceso de enseñanza aprendizaje de técnicas de ciencia de datos	¿R y Python permiten el aprendizaje de las técnicas descriptivas estadísticas?	Analizar si el lenguaje R permite el aprendizaje de las técnicas descriptivas estadísticas.	¿R permite el aprendizaje de las técnicas descriptivas estadísticas?	Medidas de tendencia central Medidas de dispersión Contrastación de hipótesis descriptivas
				Analizar si e lenguaje Python permite el aprendizaje de las técnicas descriptivas estadísticas	¿Python permite el aprendizaje de las técnicas descriptivas estadísticas?	Medidas de tendencia central Medidas de dispersión Contrastación de hipótesis descriptivas

2.2 Objeto de estudio

La observación se realiza sobre el proceso de aprendizaje de Estadística Descriptiva en la Universidad Nacional Pedro Ruiz Gallo.

2.3 Población y muestra

Criterio de selección:

Que los participantes no hayan participado de alguna actividad académica o de investigación relacionada con R y Python.

Población

La observación se realizó en los estudiantes que cursaban el primer ciclo de la Universidad Nacional Pedro Ruiz Gallo, durante el ciclo 2023 2, se matricularon 1159 estudiantes.

Muestra

La muestra fueron los estudiantes que culminaron el taller diseñado para el pre experimento, siento un total de 33 estudiantes los que culminaron el pre experimento.

2.4 Procedimiento de pre experimento

Se siguieron las siguientes etapas para el pre experimento:

Etapa preparación del Taller

Se diseñó una Hoja de Trabajo (Anexo 01) que sirvió de ruta durante el taller

Se diseñó un Caso (Anexo 02) que se desarrolló en Python y R, el caso consistió en determinar las medidas de tendencia central, medidas de dispersión y contrastas hipótesis descriptivas.

Se prepararon los datos (999) en la hoja de cálculo Excel (Anexo 03), y se entregaron por correo el primer día del taller a los estudiantes.

Se determinaron los comandos que se utilizarían en Python (Anexo 04) y en R (Anexo 05).

Etapa de convocatoria a Taller

Culminado la preparación se realizó la convocatoria (Anexo 06), al "Taller Introducción a Herramientas de tratamiento de datos" a través de los correos institucionales, se realizaron dos envíos para lograr la inscripción de los estudiantes, se inscribieron ciento cincuenta y cinco (155) estudiantes.

Etapa de desarrollo de taller

El taller se desarrolló en tres etapas:

Etapa de introducción, consistió en presentar los avances en Ciencia de Datos, Dataset, herramientas Colab y Posit Cloud, concepto de Estadística Descriptiva, se compartió el caso a desarrollar y el archivo con contenido los datos.

Etapa de Taller de Python, inició con la presentación de Colab, la explicación del formato de CSV, la posibilidad de acceder a Dataset de repositorios, el procesamiento en la nube, la carga de CSV, y el desarrollo de caso. El taller fue demostrativo y luego se supervisó a los participantes para que logren los objetivos específicos planteados.

Etapa de Taller de R, inició con la presentación de Posit Cloud, la carga de CSV, y el desarrollo de caso. El taller fue demostrativo y luego se supervisó a los participantes para que logren los objetivos específicos planteados.

Finalizado cada taller s aplicó una encuesta para obtener la experiencia de usuario y el logro de los objetivos planteados: obtener las medidas de tendencia central, medidas de dispersión y pruebas de hipótesis descriptivas.

CAPITULO III RESULTADOS

3.1 Índice de respuesta

Los correos fueron remitidos a todas las cuentas de correo de los estudiantes del primer ciclo de la UNPRG, se inscribieron 155 y finalizaron 33 estudiantes el taller de Python y Taller de R.

Tabla 5
Tasa de respuestas de estudiantes

Correos remitidos	Inscripciones	Culminaron el taller
1159	155	33

3.2 Análisis de consistencia interna

El análisis de consistencia interna mediante el coeficiente de Alfa de

Cronbach aplicado al cuestionario UEQ+

Tabla 6
Análisis de Consistencia Interna para el cuestionario UEQ+ aplicado a Python en Colab y R en Posit Cloud

Cuestionario	Indicador de Alpha de Cronbach	Interpretación
Colab	0.972	Excelente
Posit Cloud	0.914	Excelente

Tabla 7
Análisis de Consistencia Interna para evaluar las preguntas sobre logro del cálculo de las medidas de tendencia central, medidas de dispersión y contrastación de hipótesis en lenguaje Python y R

Cuestionario para	Indicador de Alpha de Cronbach	Interpretación
Python con Colab	0.874	Buena
R con Posit Cloud	0.791	Aceptable

3.3 Pruebas de normalidad

La prueba de normalidad Kolmogorov_Smirnov se aplicó a los dos cuestionarios Python con Colab y R con Posit Cloud, los resultados indicaron que todas las preguntas no siguen una distribución normal (Anexo 8).

3.4 Comparación descriptiva entre Python con Colab y R con Posit Cloud, por experiencia de usuario

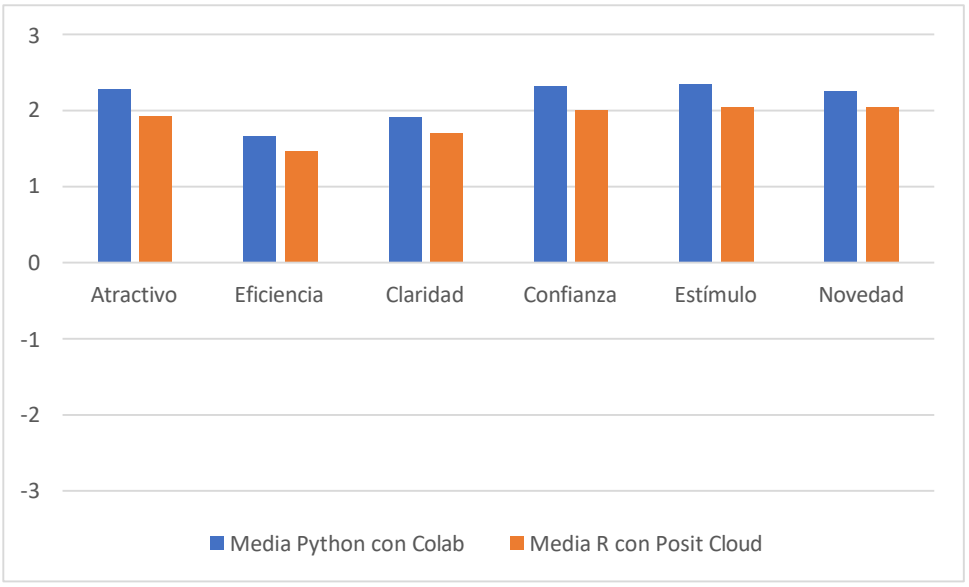
Resultado de la experiencia de usuario por dimensiones

La comparación de las experiencias de usuario entre Python con Colab y R con Posit Cloud, se realizó por las dimensiones del cuestionario UEQ+, se calcularon las medias de las escalas (media de todos los ítems). Los valores medios se transformaron de un rango de 1 a 7 a un rango de -3 a +3 para que sean compatibles con el formato de informe del UEQ.

Tabla 8
Experiencia de usuario de los estudiantes del primer ciclo de la UNPRG, por dimensión entre Python con Colab y R con Posit Cloud

Dimensión	Media de la experiencia Python con Colab	Media de la experiencia R con Posit Cloud
Atractivo	2.28	1.93
Eficiencia	1.66	1.47
Claridad	1.91	1.71
Confianza	2.33	2.00
Estímulo	2.35	2.05
Novedad	2.26	2.05

Figura 1
Medias de la experiencia de usuario de los estudiantes del primer ciclo de la UNPRG, por dimensión entre Python con Colab y R con Posit Cloud



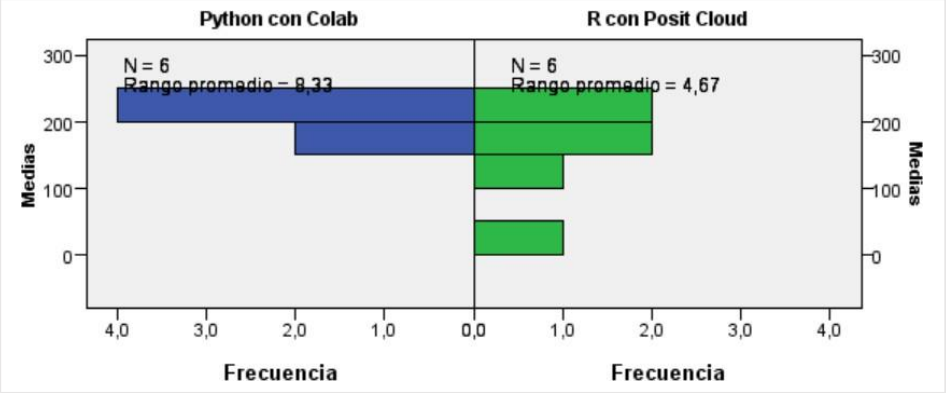
La tabla 8 y figura 1 presentan el resultado de las experiencias de usuarios de los estudiantes del primer ciclo de la UNPRG, los datos están en la escala UEQ+,

destaca que la dimensión Atractivo, Confianza y Dimensión en el uso de Python con Colab, y Novedad, Estímulo y Confianza en el uso de R con Posit Cloud.

Se observa que, en todas las dimensiones Python con Colab tiene mayor índice que R con Posit Cloud.

En ambos lenguajes la dimensión eficiencia fue la más baja.

Figura 2
Prueba U de Mann-Whitney entre dimensiones de experiencia de usuario al usar Lenguajes de Programación entre Python con Colab y R con Posit Cloud



NOTA: La figura 2 muestra la distribución de las medias de las experiencias de usuario, fue necesario realizar la prueba de hipótesis para determinar si existe diferencia entre ambas experiencias.

La diferencia entre las medias de las experiencias se analizó mediante el estadístico U de Mann-Whitney, y se obtuvo un p-valor de 0.093, y se determinó que no existen diferencia entre las experiencias de usuario al hacer uso para el aprendizaje con Python en Colab y R en Posit Cloud.

La interpretación de la prueba de U de Mann-Whitney determina que la experiencia de usuario que tuvieron los estudiantes del primer ciclo de la UNPRG, fue la misma en el momento usar Python con Colab y R con Posit Cloud.

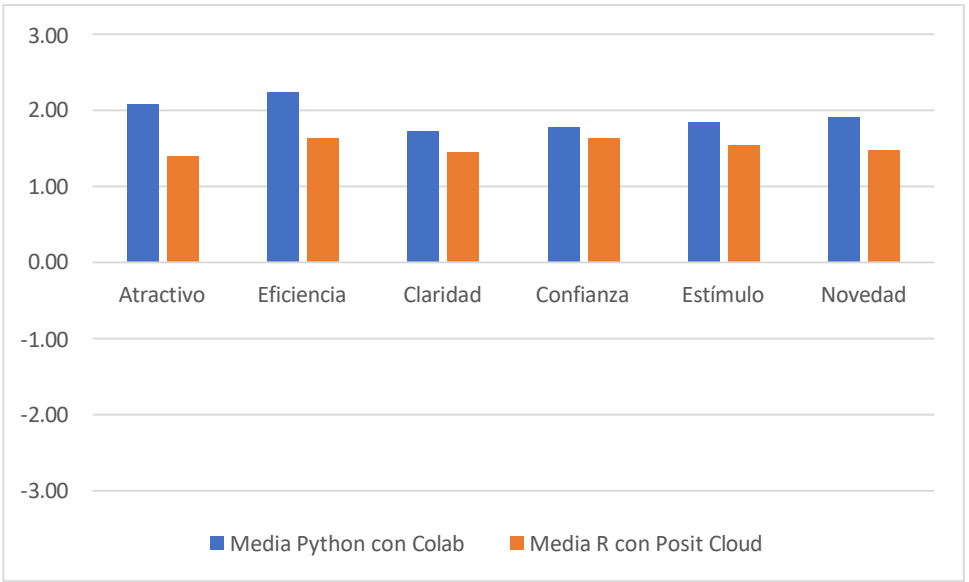
Asignación de importancia a las dimensiones

La percepción de la importancia asignada por los estudiantes a cada dimensión, se analizó por las dimensiones.

Tabla 9
Importancia asignada a la dimensión por estudiantes del primer ciclo de la UNPRG, al utilizar Python con Colab y R con Posit Cloud

Dimensión	Media de la importancia Python con Colab	Media de la importancia R con Posit Cloud
Atractivo	2.09	1.39
Eficiencia	2.24	1.64
Claridad	1.73	1.45
Confianza	1.79	1.64
Estímulo	1.85	1.55
Novedad	1.91	1.48

Figura 3
Importancia asignada a la dimensión por estudiantes del primer ciclo de la UNPRG, al utilizar Python con Colab y R con Posit Cloud

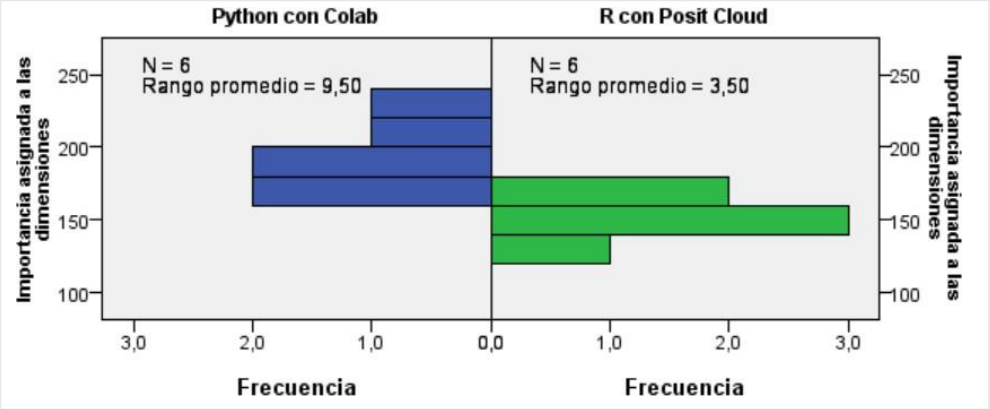


La tabla 9 y figura 3 presentan el resultado de las asignaciones de importancia a las dimensiones de experiencias de usuarios por los estudiantes del primer ciclo de la UNPRG, los datos están en la escala UEQ+, destaca que la dimensión Eficiencia es la que asignaron mayor importancia en Python con Colab, R con Posit Cloud.

Las asignaciones de importancia fueron mayores al lenguaje Python con Colab respecto a R con Posit Cloud.

La dimensión con menor asignación de importancia fue Calidad en Python con Colab respecto a R con Posit Cloud.

Figura 4
Prueba U de Mann-Whitney entre las dimensiones de las importancias asignadas al usar Lenguajes de Programación entre Python con Colab y R con Posit Cloud



NOTA: La figura 4 muestra la distribución de las medias de las importancias asignadas a las dimensiones, fue necesario realizar la prueba de hipótesis para determinar si existe diferencia entre las importancias asignadas.

La diferencia entre las medias de las importancias asignadas se analizó mediante el estadístico U de Mann-Whitney, y se obtuvo un p-valor de 0.02, y se determinó que si existe diferencia significativa al determinar la importancia de describir la experiencia de usuario Python en Colab y R en Posit Cloud.

La interpretación de la prueba de U de Mann-Whitney determina que la asignación de la importancia por parte de los estudiantes del primer ciclo de la UNPRG, no fue la misma en el momento usar Python con Colab y R con Posit Cloud.

La investigación no identifica las causas porque los estudiantes determinaron la diferencia de importancia.

Indicadores claves de rendimiento KPI

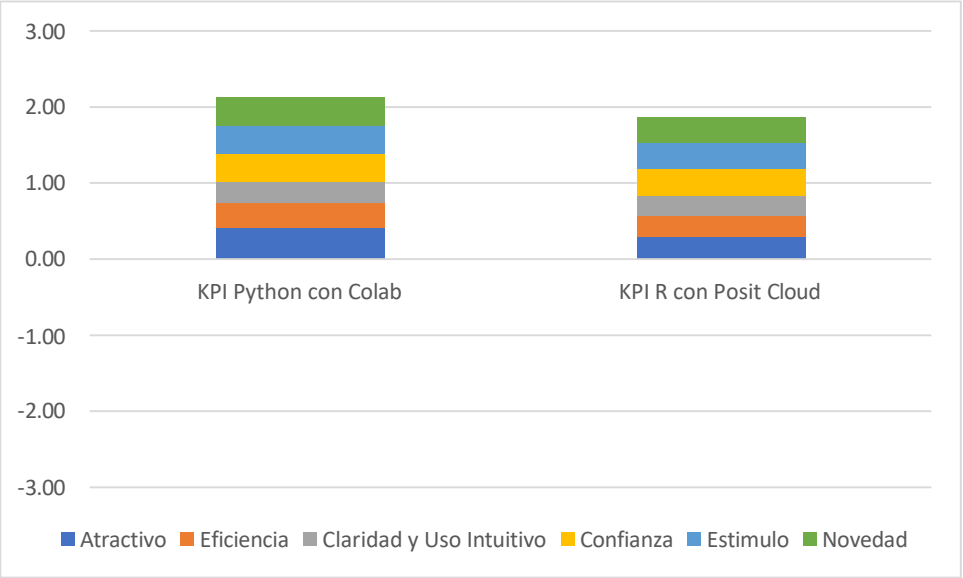
Los indicadores claves de rendimiento (KPI por sus siglas en inglés), expresión cuantificada de lo que un estudiante está haciendo y cómo interactúa con un sitio web, se observaron en la investigación.

Tabla 10
KPI por dimensión al utilizar Python con Colab y R con Posit Cloud, por los estudiantes del primer ciclo de la UNPRG

Dimensión	KPI Python con Colab	KPI R con Posit Cloud
Atractivo	0.41	0.30
Eficiencia	0.33	0.26
Claridad	0.28	0.27
Confianza	0.36	0.35

Estímulo	0.38	0.35
Novedad	0.37	0.33
KPI Global	**2.13**	**1.87**

Figura 5
Aporte al KPI por dimensiones: KPI Python con Colab y KPI R con Posit Cloud.

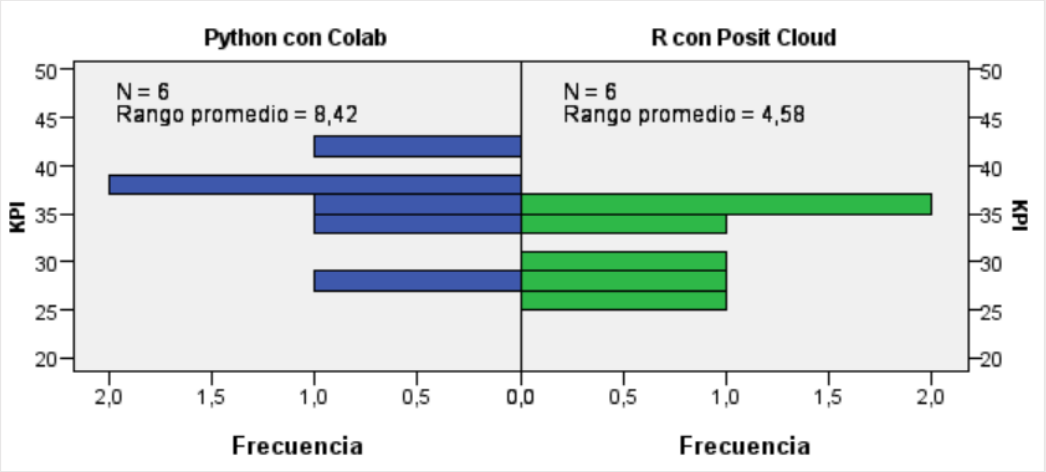


La tabla 10 y figura 5 presentan el KPI de Python con Colab 2.13 superior al KPI de R con Posit Cloud 1.87, y el aporte de las dimensiones para el KPI.

En el KPI de Python con Colab, la dimensión con mayor aporte fue Atractivo y la de menor aporte la dimensión Claridad.

En el KPI de R con Posit Cloud, la dimensión con mayor aporte fue Confianza y Estímulo, y la de menor aporte Eficiencia.

Figura 6
Prueba U de Mann-Whitney entre los KPI de las dimensiones entre Python con Colab y R con Posit Cloud



NOTA: La figura 6 muestra la distribución de las medias de los KPI, fue necesario realizar la prueba de hipótesis para determinar si existe diferencia entre los valores de los KPI.

La diferencia entre las medias de las importancias de los KPI se analizó mediante el estadístico U de Mann-Whitney, y se obtuvo un p-valor de 0.065, determinándose que no existe diferencia significativa entre los resultados obtenidos de los KPI de Python en Colab y R en Posit Cloud

3.5 Análisis de aprendizaje de las técnicas descriptivas estadísticas

Para la valoración de haber logrado determinar las medidas estadísticas, se diseñó la escala de Likert de 1 a 7, representado 1 como la imposibilidad y 7 la posibilidad total, de determinar las medidas estadísticas (Anexo 7 – Preguntas de resumen).

Se construyó la tabla de calificación a los valores de la escala de Likert para las contrastaciones de hipótesis Tabla 11.

Tabla 11
Tabla de calificación del valor de la escala de Likert para contrastación de hipótesis

Valor en la escala de Likert	Interpretación
1	No logró determinar resultado
2	
3	
4	Indiferencia
5	Determinó el resultado sin confianza
6	Logró determinar el resultado
7	

Se determinó tres grupos de técnicas descriptivas estadísticas: medidas de tendencia central, medidas de dispersión y Contrastación de hipótesis descriptivas para verificar el aprendizaje de las técnicas descriptivas.

Analizar si el lenguaje Python permite el aprendizaje de las técnicas descriptivas estadísticas

En al análisis sobre la posibilidad que el lenguaje Python permite el aprendizaje de las técnicas descriptivas, se formuló la hipótesis de investigación: ¿Python permite el aprendizaje de las técnicas descriptivas estadísticas?, y se construyeron las hipótesis estadísticas:

Tabla 12
Tabla de diseño de hipótesis para evaluar que los estudiantes lograron determinar los indicadores descriptivos utilizando Python con Colab.

Hipótesis de investigación	Hipótesis estadística
Los estudiantes del primer ciclo de la UNPRG determinan las Medidas de Tendencia Central utilizando Python con Colab	H_0: u<6 H_1: u>=6
Los estudiantes del primer ciclo de la UNPRG determinan las Medidas de Dispersión utilizando Python con Colab	H_0: u<6 H_1: u>=6
Los estudiantes del primer ciclo de la UNPRG contrastan hipótesis descriptivas utilizando Python con Colab	H_0: u<6 H_1: u>=6

La prueba estadística que se utilizó fue Wilcoxon, por que los datos no siguen una distribución normal y comparar los resultados con un parámetro.

Tabla 13
Resultado de la prueba de Wilcoxon para evaluar si los estudiantes lograron determinar los indicadores descriptivos utilizando Python con Colab.

Hipótesis de investigación	Media	Hipótesis estadística	Wilcoxon Significancia	Decisión
Los estudiantes del primer ciclo de la UNPRG determinan las Medidas de Tendencia Central utilizando Python con Colab	6	H_0: u=6 H_1: u≠6	0.580	Retener hipótesis nula
Los estudiantes del primer ciclo de la UNPRG determinan las	5.88	H_0: u=6 H_1: u≠6	0.917	Retener hipótesis nula

Medidas de Dispersión utilizando Python con Colab				
Los estudiantes del primer ciclo de la UNPRG contrastan hipótesis descriptivas utilizando Python con Colab	6	H_0: u=6 H_1: u≠6	0.695	Retener hipótesis nula

Interpretación: La tabla 13 indica mediante la significancia obtenida al aplica la prueba de Wilcoxon, que se debe retener la hipótesis nula, y con la tabla 11, se interpreta que los estudiantes si lograron determinar las medidas de tendencia central utilizando Python con Colab, para la determinación de las medidas de dispersión aunque la media (tabla 13) está por debajo del parámetro 6, esta diferencia no es significativa, y se interpreta que los estudiantes si lograron determinar las medidas de dispersión, y finalmente para los estudiantes si lograron contrastar hipótesis descriptivas.

Analizar si el lenguaje R permite el aprendizaje de las técnicas descriptivas estadísticas

En al análisis sobre la posibilidad que el lenguaje R permite el aprendizaje de las técnicas descriptivas, se formuló la hipótesis de investigación: ¿R permite el aprendizaje de las técnicas descriptivas estadísticas?, y se construyeron las hipótesis estadísticas:

Tabla 14
Tabla de diseño de hipótesis para evaluar que los estudiantes lograron determinar los indicadores descriptivos utilizando R con Posit Cloud.

Hipótesis de investigación	Hipótesis estadística
Los estudiantes del primer ciclo de la UNPRG determinan las Medidas de Tendencia Central utilizando R con Posit Cloud	H_0: u<6 H_1: u>=6
Los estudiantes del primer ciclo de la UNPRG determinan las Medidas de Dispersión utilizando R con Posit Cloud	H_0: u<6 H_1: u>=6
Los estudiantes del primer ciclo de la UNPRG contrastan hipótesis descriptivas utilizando R con Posit Cloud	H_0: u<6 H_1: u>=6

La prueba estadística que se utilizó fue Wilcoxon, porque los datos no siguen una distribución normal y comparar los resultados con un parámetro.

Tabla 15
Resultado de la prueba de Wilcoxon para evaluar si los estudiantes lograron determinar los indicadores descriptivos utilizando R con Posit Cloud.

Hipótesis de investigación	Media	Hipótesis estadística	Wilcoxon Significancia	Decisión
Los estudiantes del primer ciclo de la UNPRG determinan las Medidas de Tendencia Central utilizando R con Posit Cloud	6.06	H_0: u=6 H_1: u≠6	0.736	Retener hipótesis nula

Los estudiantes del primer ciclo de la UNPRG determinan las Medidas de Dispersión utilizando R con Posit Cloud	6.33	H_0: u=6 H_1: u≠6	0.037	Rechazar hipótesis nula
Los estudiantes del primer ciclo de la UNPRG contrastan hipótesis descriptivas utilizando R con Posit Cloud	6.21	H_0: u=6 H_1: u≠6	0.164	Retener hipótesis nula

Interpretación: La tabla 13 indica mediante la significancia obtenida al aplica la prueba de Wilcoxon, que se debe retener la hipótesis nula, y con la tabla 11, se interpreta que los estudiantes si lograron determinar las medidas de tendencia central utilizando R con Posit Colab, para la determinación de las medidas de dispersión aunque la decisión es rechazar la hipótesis nula se aprecia que la media 6.33 es superior a 6 interpretándose que los estudiantes si lograron determinar las medidas de dispersión, y finalmente los estudiantes si lograron contrastar hipótesis descriptivas utilizando R con Posit Cloud.

CAPITULO IV DISCUSION DE LOS RESULTADOS

El interés por el análisis de los datos desde el siglo pasado, se mantiene vigente en los investigadores y universidades del presente (Dohoho, 2017) (Ley & Bordas, 2018), análisis que puede ser desarrollado debido a la gran cantidad de datos generados y disponibles cada año por la prestación de servicios digitales, sin embargo, en países en vía de desarrollo como los ubicados América Latina o el continente africano, existe un desperdicio de datos, por lo que es importante desde la universidad formar investigadores en análisis de gran volumen de datos concordando con (Mutuku, 2019).

La selección del lenguaje de programación para iniciar el aprendizaje de las técnicas iniciales en ciencias de datos, Python y R, es muy importante, y desde sus apariciones, se inició el debate sobre que lenguaje es mejor. A diferencia de las pruebas realizadas por (López, 2020) sobre tiempo de ejecución, en esta investigación se comparó la posibilidad de determinar estadísticos descriptivos, y el resultado obtenido indicó no existió diferencia entre los lenguajes Python y R, estos resultados concuerdan con lo expuesto por (UNIR, 2020) como lenguajes potentes para el análisis de datos, dejando la preferencia o elección a otros factores.

El resultado del análisis descriptivo de la experiencia de usuario que vivieron los estudiantes del primer ciclo, no indicó preferencia por algún lenguaje, una de las características de la muestra, que al inicio de la investigación no fue considerada, fue que ningún participante había usado el lenguaje Python o R.

CONCLUSIONES

La investigación ha permitido analizar que el lenguaje R sirve para aprendizaje de las técnicas descriptivas estadísticas, como las medidas de tendencia central, medidas de dispersión, y contrastación de hipótesis descriptivas, en estudiantes del primer ciclo de universidad.

La investigación ha permitido analizar que el lenguaje Python sirve para aprendizaje de las técnicas descriptivas estadísticas, como las medidas de tendencia central, medidas de dispersión, y contrastación de hipótesis descriptivas, en estudiantes del primer ciclo de universidad.

La investigación presenta que la importancia asignada a las dimensiones de experiencia de usuario es más valorada en Python con Colab que en R con Posit CLOUD, la investigación no concluye las razones de esta diferencia, así mismo, permitió determinar que el indicador clave de rendimiento es similar entre Python con Colab y R con Posit CLOUD.

Finalmente, la investigación logró en un pre experimento, describir experiencia de usuario por los estudiantes utilizando los lenguajes de programación R y Python, en el proceso de enseñanza aprendizaje de técnicas de ciencia de datos.

RECOMENDACIONES

Con respecto al uso lenguaje Python con Colab y R con Posit CLOUD, se recomienda evaluar su aplicación en estudiantes que estén elaborando sus proyectos de investigación y necesiten el conocimiento de técnicas avanzadas de estadística.

Con respecto al uso lenguaje R con Posit CLOUD, se recomienda evaluar su aplicación en estudiantes que estén elaborando sus proyectos de investigación y necesiten el conocimiento de técnicas avanzadas de estadística.

La valoración de la importancia e indicadores claves de rendimiento, de los lenguajes de programación R y Python debe ser profundizada y analizada de acuerdo a los programas académicos.

Se propone evaluar la experiencia de los estudiantes utilizando los lenguajes de programación R y Python, mediante cuasi experimento.

REFERENCIAS

Donoho, D. (2017). 50 Years of Data Science. In Journal of Computational and Graphical Statistics (Vol. 26, Issue 4). https://doi.org/10.1080/10618600.2017.1384734

D.S. Gómez-Reverte, M. D. M. J. M. M. J. N. y A. P. (2022). Introducción a la Estadística. https://rua.ua.es/dspace/bitstream/10045/26617/1/Tema1.pdf

López R. (2020), Análisis comparativo de lenguajes de programación para el desarrollo de aplicaciones en Ciencia de Datos , Disponible en https://rinacional.tecnm.mx/bitstream/TecNM/4155/1/MC_Ricardo_Gudiel_Lopez_Perez_2020.pdf, México

IMF (2021). Ventajas y desventajas del lenguaje R . (Recuperado el 10 de Julio de 2021). Disponible en: https://blogs.imf-formacion.com/blog/tecnologia/ventajasy-desventajas-del-lenguaje-r-202007/ . Blog Tecnología. Bogotá. D.C.

Laugwitz, B., Held, T., & Schrepp, M. (2008). Construction and evaluation of a user experience questionnaire. Springer, Holzinger, A. (Ed.): USAB 2008, LNCS 5298, S. 63-76. Obtenido de http://dx.doi.org/10.1007/978-3-540-89350-9_6

Martin Schrepp,J. T. (2019). Handbook for the modular extension of the User Experience Questionnaire. Obtenido de www.ueq-online.org

Ley, C., & Bordas, S. P. A. (2018). What makes Data Science different? A discussion involving Statistics2.0 and Computational Sciences. International Journal of Data Science and Analytics, 6(3). https://doi.org/10.1007/s41060-017-0090-x

Mutuku, C. M. (2019). Engaging a Data Revolution: Open Science Data Hubs and the New Role for Universities in Africa. Open Information Science, 3(1). https://doi.org/10.1515/opis-2019-0008

Ozgur, C., Colliau, T., Rogers, G., & Hughes, Z. (2021). MatLab vs. Python vs. R. Journal of Data Science, 15(3). https://doi.org/10.6339/jds.201707_15(3).0001

Python Software Foundation, "Applications for Python," [Online]. Available: https://www.python.org/about/apps/. [Accessed 21 May 2020].

P. P. Shinde, K. S. Oza and R. K. Kamat, "Big data predictive analysis: Using R analytical tool," presented at 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, 2017

Rafael Agacino. (2022). Conceptos fundamentales de estadística. CEPAL. https://www.cepal.org/sites/default/files/presentations/2.2-conceptos-fundamentales-estadisticas-ambientales.pdf

Skiena, S. (2017) The Data Science Design Manual. Nueva York, Estados Unidos: Springer

UNIR, (2020), R vs Python: ¿cuál es mejor para el análisis de datos?, https://www.unir.net/ingenieria/revista/r-vs-python/

Anexos

Anexo 01 Hoja de trabajo

Taller Introducción a Herramientas de Tratamiento de Datos

HOJA DE TRABAJO

Trabajo

1. Averigüe como son los estudiantes de una Universidad
   Realizar la estadística descriptiva de los estudiantes del primer ciclo de una Universidad.
2. Tráigame información de una población de vegetales
   Realizar la estadística descriptiva de una población de vegetales.
3. Quiero, "para ayer", los datos de la población afectada
   Realizar la estadística descriptiva de una población expuesta a una enfermedad por su ambiente natural.
4. Etc.

Marco teórico:

Estadística Descriptiva

La estadística emplea métodos descriptivos y de inferencia estadística. Los primeros se ocupan de la recolección, organización, tabulación, presentación y reducción de la información.

https://www.esan.edu.pe/conexion-esan/que-es-la-estadistica-descriptiva

Descriptive statistics provide summarizing information of the characteristics and distribution of values in one or more datasets.

Jay Lee, in International Encyclopedia of Human Geography (Second Edition), 2020

https://www.sciencedirect.com/topics/social-sciences/descriptive-statistics

Pero ¿qué describimos? o ¿qué investigamos?

Para realizar una descripción, primero debes investigar → medir

NO es lo mismo medir talla que temperatura → cada uno requiere de un instrumento

¿se puede medir los sentimientos? → ¿podemos cuantificar?

Físico – abstracto

Centímetro – Balanza – Termómetro – Cuestionario – Ficha de observación etc.

¿Qué vamos a presentar?

Informes (información)

Contenido estadístico descriptivo

Medias de Tendencia Central

Las medidas de tendencia central son datos que informan cuál es el centro en torno al cual se ubica un conjunto de datos; estas se utilizan principalmente para resumir la información.

https://colombia.unir.net/actualidad-unir/medidas-tendencia-central/

- Media
- Mediana
- Moda
- Menor
- Mayor
- Suma

Medidas de distribución dispersión

- Rango de variación
- Varianza.
- Coeficiente de variación.
- Desviación estándar.

Medidas de asimetría y curtosis

Anexo 02 Caso

Taller Introducción a Herramientas de Tratamiento de Datos

CASO

Objetivo: realizar el análisis descriptivo en una población de 999 personas

Cuestionario

1. DEPARTAMENTO: _____
2. RANGO DE EDAD: a) 18 – 30 b) 31-40 c) 41-50 d) 51-60 e)60 a más
3. SEXO: Masculino _____Femenino _____
4. PESO: _____
5. TALLA: _____
6. LE GUSTA EL CINE marcar con una X: _____SI _____NO

Informe

Tabla 01
Total de personas por departamento

Departamento	Total

Tabla 02
Medias de Tendencia Central de Peso

Medida	Valor
Media	
Mediana	
Moda	
Menor	
Mayor	

Suma	

Tabla 03
Medias de Distribución de Peso

Medida	Valor
Rango de variación	
Varianza	
Coeficiente de variación	
Desviación estándar	

Tabla 04
Medidas de asimetría de Peso

Medida	Valor
Coeficiente de asimetría	
Curtosis	

Hipótesis descriptivas

a) Determinar si el peso de la población es mayor a 60Kg

Media vs un parámetro

b) Determinar si existe diferencia entre el peso de las mujeres y el peso de los hombres

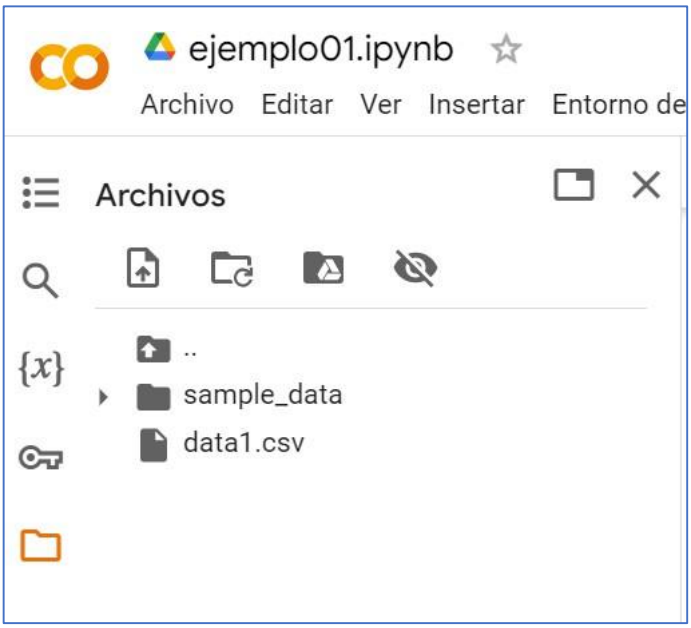
Media de un grupo vs Media de otro grupo

Anexo 03 Estructura del Archivo data.xlsx

A	B	C	D	E	F	G	H
item	DEPARTAMENTO	RANGO	SEXO	PESO	TALLA	CINE	CINE 1
1	CAJAMARCA	D	M	71	160	2	S
2	CAJAMARCA	B	F	62	170	1	S
3	LAMBAYEQUE	B	M	90	172	1	S
4	LA LIBERTAD	C	F	82	174	2	S
5	CAJAMARCA	B	F	93	178	2	S
6	LA LIBERTAD	E	F	91	173	2	S
7	LA LIBERTAD	C	M	97	177	1	S
8	LA LIBERTAD	A	M	97	174	1	S
9	LAMBAYEQUE	C	M	78	155	2	S
10	LA LIBERTAD	D	F	77	174	2	S
11	LA LIBERTAD	A	M	87	172	1	S
12	LA LIBERTAD	A	F	78	180	1	S
13	CAJAMARCA	B	F	81	157	2	S
14	LA LIBERTAD	C	F	62	175	2	S
15	LA LIBERTAD	A	F	84	162	1	S
16	CAJAMARCA	E	M	90	153	2	S
17	LA LIBERTAD	E	M	76	162	1	S
18	CAJAMARCA	D	M	82	165	1	S
19	LA LIBERTAD	C	F	88	169	1	S
20	LA LIBERTAD	D	F	71	171	1	S

Anexo 04 Comandos de Python

Cargar archivo



Paquetes

```
import pandas as pd
```

Dataset

```
datos = pd.read_csv("/content/data1.csv")
```

Descriptivos

```
datos.head()
datos.describe
datos.groupby('DEPARTAMENTO').count()
```

```
tabla=pd.pivot_table(datos,values='TALLA',index='DEPARTAMENTO',aggfunc=np.size)
tabla
```

```
datos['PESO'].describe()
```

```
datos['PESO'].median()
```

```
datos['PESO'].mode()
datos['PESO'].sum()
datos['PESO'].var()
datos['PESO'].std()
datos['PESO'].max()
datos['PESO'].min()
```

```
rango=datos['PESO'].max()-datos['PESO'].min()
rango
```

```
coeficiente_variacion=datos['PESO'].std()/datos['PESO'].mean()
coeficiente_variacion
```

```
datos['PESO'].skew()
datos['PESO'].kurt()
```

```
datos.groupby('DEPARTAMENTO')['TALLA','PESO'].mean()
datos.groupby('SEXO')['PESO'].mean()
```

Comandos para hipótesis descriptivas

```
from scipy import stats
```

```
shapiro_wilk=stats.shapiro(datos.PESO)
shapiro_wilk
```

```
kolmogorov_smirnov=stats.kstest(datos.PESO,'norm')
kolmogorov_smirnov
```

```
stats.ttest_1samp(datos.PESO,1.60)
```

```
sexos=datos['SEXO']
tallas=datos['TALLA']
```

```
tallas_m=tallas[sexos=='F']
tallas_h=tallas[sexos=='M']
stats.ttest_ind(tallas_m,tallas_h)
```

Anexo 05 Comandos de R

COMANDOS

x<-2

y<-3

z<-x+y

z

muestra el valor de Z

En que directorio me encuentro

getwd()

Limpiar pantalla (no borra el historial)

CTRL (+) L

Informarse sobre un comando

?read.csv()

Dataset

datos<-read.csv('data1.csv')

datos

DESCRIPTIVOS

mean(datos$PESO)

mean(datos$PESO)

promedio_peso=mean(datos$PESO)

promedio_peso

max(datos$PESO)

min(datos$PESO)

median(datos$PESO)

range(datos$PESO)

```
var(datos$PESO)
sd(datos$PESO)
summary(datos)
summary(datos$PESO)
```

Paquete - librerías

```
install.packages("Hmisc")
require(Hmisc)

library(pastecs)
detach("package:pastecs", unload = TRUE)
```

otros paquetes

```
install.packages("survival")
install.packages("lattice")
install.packages("ggplot2")
```

Proporción

```
describe(datos)
describe(datos$DEPARTAMENTO)
```

Descriptivos

stat.desc(datos)

res<-stat.desc(datos) comentario, los resultados quedan almacenado en “res”

round(res,2) comentario, los resultados se redondean a dos decimales

res<-stat.desc(datos[,-c(1)]) comentario, los resultados quedan almacenado en “res” menos la columna 1

round(res,2) comentario, los resultados se redondean a dos decimales, pero sin la columna 1

res<-stat.desc(datos)

res<-stat.desc(datos[,-c(3,4)])

Otro paquete

install.packages("psych")

require(psych)

psych::describe(datos)

psych::describeBy(datos,group = datos.DEPARTAMENTO)

También

Hmisc::describe(datos)

Prueba de hipótesis – un parámetro

```
t.test(datos$PESO,mu=60)
```

Prueba de hipótesis de dos muestras

```
peso_hombres=datos$PESO[datos$SEXO=='M']
peso_hombres
peso_mujeres=datos$PESO[datos$SEXO=='F']
peso_mujeres

shapiro.test(peso_hombres)
shapiro.test(peso_mujeres)

ks.test(peso_hombres, "pnorm")

var.test(peso_hombres,peso_mujeres)

t.test(peso_hombres,peso_mujeres)
t.test(peso_hombres,peso_mujeres,,paired = FALSE,var.equal = TRUE)

ks.test(peso_hombres,peso_mujeres)

wilcox.test(peso_hombres,peso_mujeres)
```

Anexo 06 Convocatoria

TALLER
INTRODUCCIÓN A HERRAMIENTAS DE TRATAMIENTO DE DATOS

Objetivo:
Iniciar a jóvenes universitarios en el uso de herramientas para investigación (primer ciclo)

Requisito
Ganas de aprender, mucha intuición, no importa si no sabes de computación o investigación, aquí empiezas, eso sí, *si no tienes ganas de aprender no te inscribas*.

Costo: gratuito

** Cupo Limitado por conexiones a Meet

Fechas:
29 de enero a las 9:00pm
31 de enero a las 9:00pm
02 de febrero a las 9:00pm
Y adicionales ...🤗🤗🤗

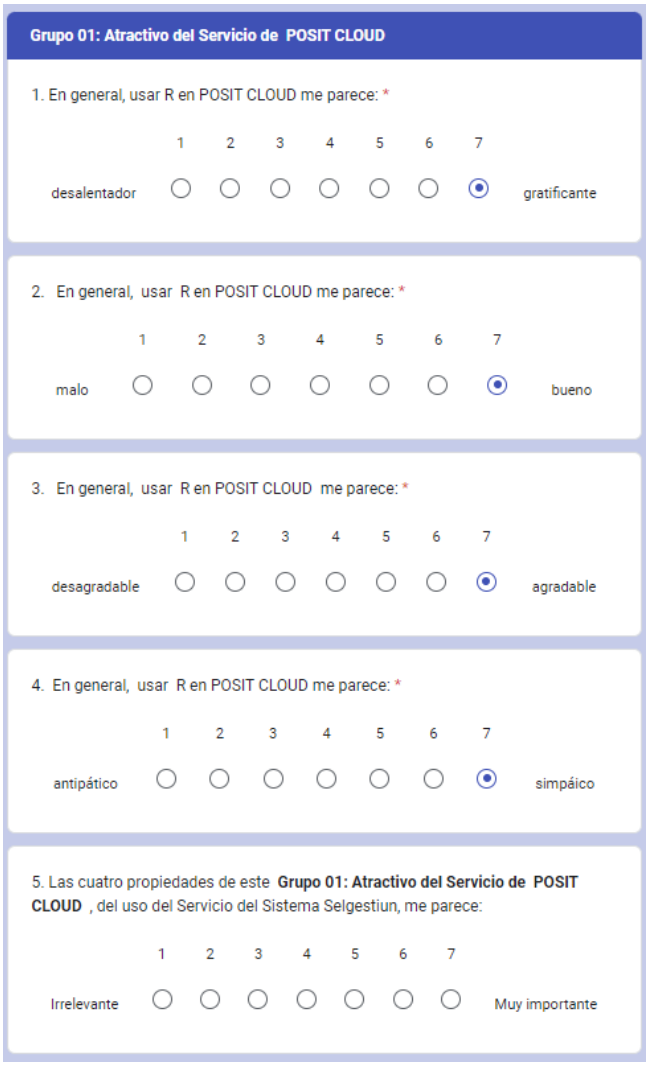
Recomendable: conectarse por laptop o PC

Logros: al culminar el taller tendrás nueva concepción de lo que es la universidad, y una ventaja competitiva para afrontar tu vida universitaria.

Una vez aceptada tu inscripción, se te enviará a tu correo el enlace para el taller

Anexos 07 - Cuestionario

"EXPERIENCIA DE USUARIO UTILIZANDO (PYTHON o R) EN COLAB PARA CALCULAR ESTADISTICOS DESCRIPTIVOS"



Grupo 01: Atractivo del Servicio de POSIT CLOUD

1. En general, usar R en POSIT CLOUD me parece: *

	1	2	3	4	5	6	7	
desalentador	○	○	○	○	○	○	◉	gratificante

2. En general, usar R en POSIT CLOUD me parece: *

	1	2	3	4	5	6	7	
malo	○	○	○	○	○	○	◉	bueno

3. En general, usar R en POSIT CLOUD me parece: *

	1	2	3	4	5	6	7	
desagradable	○	○	○	○	○	○	◉	agradable

4. En general, usar R en POSIT CLOUD me parece: *

	1	2	3	4	5	6	7	
antipático	○	○	○	○	○	○	◉	simpáico

5. Las cuatro propiedades de este **Grupo 01: Atractivo del Servicio de POSIT CLOUD** , del uso del Servicio del Sistema Selgestiun, me parece:

	1	2	3	4	5	6	7	
Irrelevante	○	○	○	○	○	○	○	Muy importante

Grupo 02: Eficiencia del Servicio de POSIT CLOUD

6. El manejo de R en POSIT CLOUD me parece: *

	1	2	3	4	5	6	7	
Lento	○	○	○	○	○	○	○	Rápido

7. El manejo de R en POSIT CLOUD , me parece: *

	1	2	3	4	5	6	7	
Ineficiente	○	○	○	○	○	○	○	Eficiente

8. El manejo de R en POSIT CLOUD me parece: *

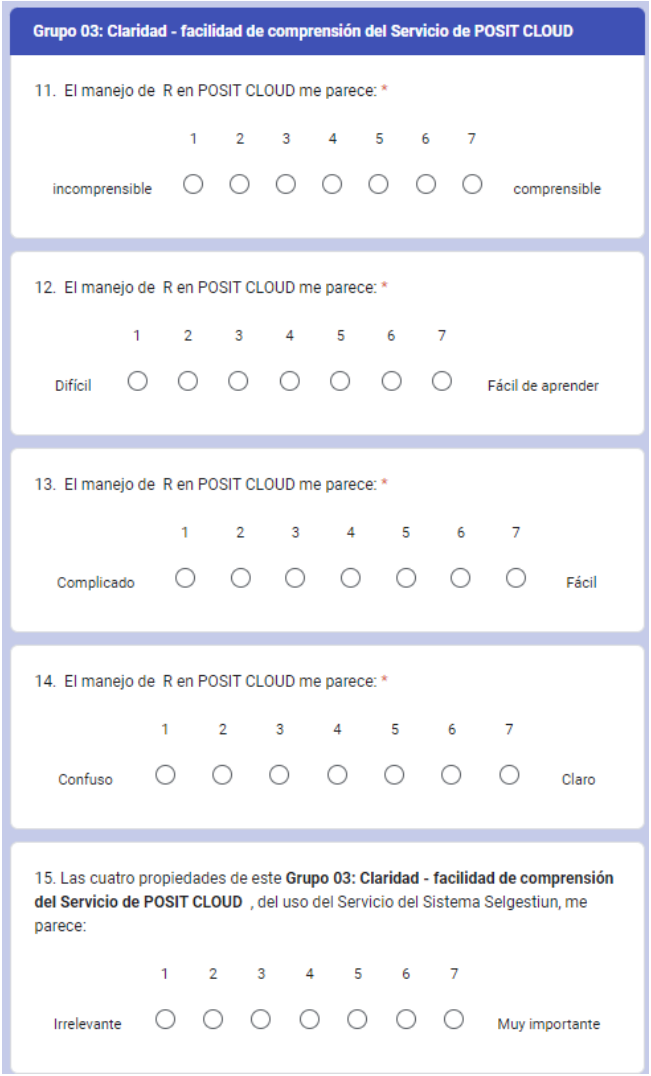
	1	2	3	4	5	6	7	
nada práctico	○	○	○	○	○	○	○	práctico

9. El manejo de R en POSIT CLOUD me parece: *

	1	2	3	4	5	6	7	
ordenado	○	○	○	○	○	○	○	sobrecargado

10. Las cuatro propiedades de este **Grupo 02: Eficiencia del Servicio de POSIT CLOUD**, del uso del Servicio del Sistema Selgestiun, me parece: *

	1	2	3	4	5	6	7	
Irrelevante	○	○	○	○	○	○	○	Muy importante



Grupo 03: Claridad - facilidad de comprensión del Servicio de POSIT CLOUD

11. El manejo de R en POSIT CLOUD me parece: *

	1	2	3	4	5	6	7	
incomprensible	○	○	○	○	○	○	○	comprensible

12. El manejo de R en POSIT CLOUD me parece: *

	1	2	3	4	5	6	7	
Difícil	○	○	○	○	○	○	○	Fácil de aprender

13. El manejo de R en POSIT CLOUD me parece: *

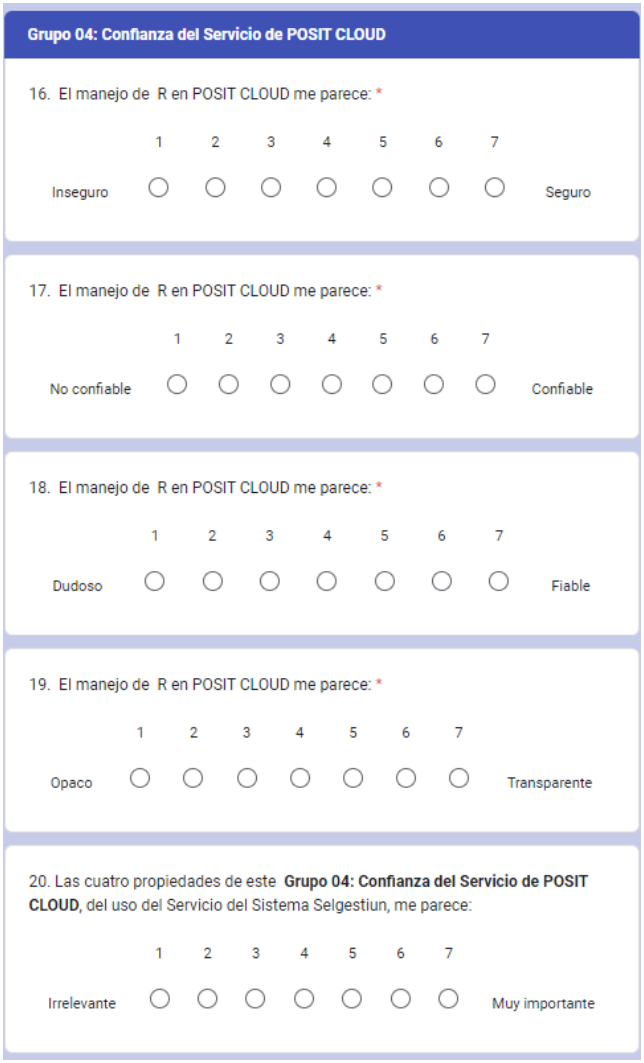
	1	2	3	4	5	6	7	
Complicado	○	○	○	○	○	○	○	Fácil

14. El manejo de R en POSIT CLOUD me parece: *

	1	2	3	4	5	6	7	
Confuso	○	○	○	○	○	○	○	Claro

15. Las cuatro propiedades de este **Grupo 03: Claridad - facilidad de comprensión del Servicio de POSIT CLOUD** , del uso del Servicio del Sistema Selgestiun, me parece:

	1	2	3	4	5	6	7	
Irrelevante	○	○	○	○	○	○	○	Muy importante



Grupo 04: Confianza del Servicio de POSIT CLOUD

16. El manejo de R en POSIT CLOUD me parece: *

	1	2	3	4	5	6	7	
Inseguro	○	○	○	○	○	○	○	Seguro

17. El manejo de R en POSIT CLOUD me parece: *

	1	2	3	4	5	6	7	
No confiable	○	○	○	○	○	○	○	Confiable

18. El manejo de R en POSIT CLOUD me parece: *

	1	2	3	4	5	6	7	
Dudoso	○	○	○	○	○	○	○	Fiable

19. El manejo de R en POSIT CLOUD me parece: *

	1	2	3	4	5	6	7	
Opaco	○	○	○	○	○	○	○	Transparente

20. Las cuatro propiedades de este **Grupo 04: Confianza del Servicio de POSIT CLOUD**, del uso del Servicio del Sistema Selgestiun, me parece:

	1	2	3	4	5	6	7	
Irrelevante	○	○	○	○	○	○	○	Muy importante

Grupo 05: Estímulo en el uso del Servicio de POSIT CLOUD

21. El manejo de R en POSIT CLOUD me parece: *

	1	2	3	4	5	6	7	
nada interesante	○	○	○	○	○	○	○	interesante

22. El manejo de R en POSIT CLOUD me parece: *

	1	2	3	4	5	6	7	
aburrido	○	○	○	○	○	○	○	entretenido

23. El manejo de R en POSIT CLOUD me parece: *

	1	2	3	4	5	6	7	
poco valioso	○	○	○	○	○	○	○	valioso

24. El manejo de R en POSIT CLOUD me parece: *

	1	2	3	4	5	6	7	
inclina al sueño	○	○	○	○	○	○	○	estimulante

25. Las cuatro propiedades de este **Grupo 05: Estímulo en el uso del Servicio de POSIT CLOUD**, del uso del Servicio del Sistema Selgestiun, me parece:

	1	2	3	4	5	6	7	
Irrelevante	○	○	○	○	○	○	○	Muy importante

Grupo 06: Originalidad - Novedad del Servicio POSIT CLOUD

26. El manejo de R en POSIT CLOUD me parece: *

	1	2	3	4	5	6	7	
falto de imaginación	○	○	○	○	○	○	○	creativo

27. El manejo de R en POSIT CLOUD me parece: *

	1	2	3	4	5	6	7	
convencional	○	○	○	○	○	○	○	original

28. El manejo de R en POSIT CLOUD me parece: *

	1	2	3	4	5	6	7	
tradicional	○	○	○	○	○	○	○	novedoso

29. El manejo de R en POSIT CLOUD me parece: *

	1	2	3	4	5	6	7	
conservador	○	○	○	○	○	○	○	innovador

30. Las cuatro propiedades de este **Grupo 06: Originalidad - Novedad del Servicio POSIT CLOUD**, del uso del Servicio del Sistema Selgestiun, me parece:

	1	2	3	4	5	6	7	
Irrelevante	○	○	○	○	○	○	○	Muy importante

Resumen

31. Finalmente pude determinar las medidas de tendencia central de un dataset utilizando R *

	1	2	3	4	5	6	7	
No pude dfinitivamente	○	○	○	○	○	○	○	Si pude definitivamente

32. Finalmente pude determinar las medidas de distribución y asimetría de un dataset utilizando R *

	1	2	3	4	5	6	7	
No pude dfinitivamente	○	○	○	○	○	○	○	Si pude definitivamente

33. Finalmente pude realizar el contraste hipótesis descriptivas de un dataset utilizando R *

	1	2	3	4	5	6	7	
No pude dfinitivamente	○	○	○	○	○	○	○	Si pude definitivamente

Anexo 08 Pruebas de normalidad

Pruebas de normalidad CUESTIONARIO PYTHON CON COLAB

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
p1	,373	33	,000	,614	33	,000
p2	,348	33	,000	,624	33	,000
p3	,346	33	,000	,626	33	,000
p4	,328	33	,000	,546	33	,000
p5	,323	33	,000	,629	33	,000
p6	,405	33	,000	,537	33	,000
p7	,364	33	,000	,588	33	,000
p8	,241	33	,000	,802	33	,000
p9	,261	33	,000	,669	33	,000
p10	,265	33	,000	,778	33	,000
p11	,252	33	,000	,839	33	,000
p12	,234	33	,000	,764	33	,000
p13	,391	33	,000	,584	33	,000
p14	,378	33	,000	,537	33	,000
p15	,341	33	,000	,607	33	,000
p16	,366	33	,000	,594	33	,000
p17	,380	33	,000	,547	33	,000
p18	,318	33	,000	,595	33	,000
p19	,364	33	,000	,464	33	,000
p20	,311	33	,000	,636	33	,000
p21	,344	33	,000	,482	33	,000
p22	,363	33	,000	,669	33	,000
p23	,307	33	,000	,676	33	,000
p24	,359	33	,000	,657	33	,000
pc01	,392	33	,000	,621	33	,000
pc02	,408	33	,000	,610	33	,000
pc03	,361	33	,000	,635	33	,000
pc04	,408	33	,000	,610	33	,000
pc05	,361	33	,000	,635	33	,000
pc06	,345	33	,000	,638	33	,000
tc	,258	33	,000	,752	33	,000
md	,264	33	,000	,793	33	,000
hi	,258	33	,000	,773	33	,000

a. Corrección de la significación de Lilliefors

Pruebas de normalidad con SPSS

Cuestionario R con POSIT CLOUD

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
p1	,212	33	,001	,841	33	,000
p2	,279	33	,000	,770	33	,000
p3	,262	33	,000	,811	33	,000
p4	,243	33	,000	,836	33	,000
p5	,266	33	,000	,760	33	,000
p6	,268	33	,000	,785	33	,000
p7	,246	33	,000	,827	33	,000
p8	,155	33	,043	,906	33	,008
p9	,248	33	,000	,799	33	,000
p10	,209	33	,001	,849	33	,000
p11	,248	33	,000	,843	33	,000
p12	,249	33	,000	,822	33	,000
p13	,281	33	,000	,806	33	,000
p14	,249	33	,000	,822	33	,000
p15	,230	33	,000	,837	33	,000
p16	,346	33	,000	,727	33	,000
p17	,285	33	,000	,633	33	,000
p18	,207	33	,001	,834	33	,000
p19	,323	33	,000	,606	33	,000
p20	,199	33	,002	,843	33	,000
p21	,243	33	,000	,829	33	,000
p22	,259	33	,000	,800	33	,000
p23	,232	33	,000	,836	33	,000
p24	,251	33	,000	,817	33	,000
pc01	,392	33	,000	,621	33	,000
pc02	,408	33	,000	,610	33	,000
pc03	,361	33	,000	,635	33	,000
pc04	,408	33	,000	,610	33	,000
pc05	,361	33	,000	,635	33	,000
pc06	,345	33	,000	,638	33	,000
tc	,281	33	,000	,806	33	,000
md	,298	33	,000	,739	33	,000
hi	,318	33	,000	,755	33	,000

a. Corrección de la significación de Lilliefors

Anexo 09 Hipótesis obtenidas con SPSS

Resultados con SPSS de prueba de Wilcoxon aplicadas a Tendencia Central TC, Medidas de Dispersión MD y Constrastación de Hipótesis HI de Python con Colab

Resumen de prueba de hipótesis

	Hipótesis nula	Test	Sig.	Decisión
1	La media de tc es igual a 6,000.	Prueba de Wilcoxon de los rangos con signo de una muestra	,580	Retener la hipótesis nula.
2	La media de md es igual a 6,000.	Prueba de Wilcoxon de los rangos con signo de una muestra	,917	Retener la hipótesis nula.
3	La media de hi es igual a 6,000.	Prueba de Wilcoxon de los rangos con signo de una muestra	,695	Retener la hipótesis nula.

Se muestran las significancias asintóticas. El nivel de significancia es ,05.

Resultados con SPSS de prueba de Wilcoxon aplicadas a Tendencia Central TC, Medidas de Dispersión MD y Constrastación de Hipótesis HI de R con Posit CLOUD

Resumen de prueba de hipótesis

	Hipótesis nula	Test	Sig.	Decisión
1	La media de tc es igual a 6,000.	Prueba de Wilcoxon de los rangos con signo de una muestra	,736	Retener la hipótesis nula.
2	La media de md es igual a 6,000.	Prueba de Wilcoxon de los rangos con signo de una muestra	,037	Rechazar la hipótesis nula.
3	La media de hi es igual a 6,000.	Prueba de Wilcoxon de los rangos con signo de una muestra	,164	Retener la hipótesis nula.

Se muestran las significancias asintóticas. El nivel de significancia es ,05.

ACTA DE SUSTENTACION

UNIVERSIDAD NACIONAL PEDRO RUIZ GALLO
FACULTAD DE INGENIERÍA CIVIL DE SISTEMAS Y DE ARQUITECTURA
DECANATO

ACTA DE SUSTENTACIÓN N°569-2024-FICSA-D

Siendo las 8:30am ~~horas~~ del día 22 de febrero del 2024, se reunieron de manera presencial los miembros de jurado de la tesis titulada: : **"R Y PYTHON EXPERIENCIA EN EL APRENDIZAJE DE ESTADÍSTICA DESCRIPTIVA EN LA UNIVERSIDAD NACIONAL PEDRO RUIZ GALLO" con código N° IS_V_2023_011,** designado por **Resolución Decanal Virtual N° 385-2023-UNPRG-FICSA** con la finalidad de Evaluar y Calificar la sustentación de la tesis antes mencionada, conformado por los siguientes docentes:

MSC. ING. PILAR DEL ROSARIO RIOS CAMPOS	**PRESIDENTE**
MSC. ING. MARIA DE LOS ÁNGELES GUZMÁN VALLE	**SECRETARIO**
ING. CÉSAR AUGUSTO GUZMÁN VALLE	**VOCAL**

Asesorado por DR. ING. REGIS JORGE ALBERTO DIAZ PLAZA

El acto de sustentación fue autorizado por **OFICIO VIRTUAL N° 037-2024-UIFICSA,** la tesis fue presentada y sustentada por la Bachiller: **GAYOSO ROJAS YNES JOSSELY,** tuvo una duración de 60 minutos. Después de la sustentación, y absueltas las preguntas y observaciones de los miembros del jurado; se procedió a la calificación respectiva:

	NUMERO	LETRAS	CALIFICATIVO
GAYOSO ROJAS YNES JOSSELY	16	DIECISEIS	BUENO

Por lo que quedan **APTOS** para obtener el Título Profesional de **INGENIERO (A) DE SISTEMAS** de acuerdo con la Ley Universitaria 30220 y la normatividad vigente de la Facultad de Ingeniería Civil De Sistemas y de Arquitectura de la Universidad Nacional Pedro Ruiz Gallo.

Siendo las 9:30 am Se dio por concluido el presente acto académico, dándose conformidad al presente acto, con la firma de los miembros del jurado.

MSC. ING. PILAR DEL ROSARIO RIOS CAMPOS
PRESIDENTE

MSC. ING. MARIA DE LOS ÁNGELES GUZMÁN VALLE
SECRETARIO

ING. CÉSAR AUGUSTO GUZMÁN VALLE
VOCAL

DR. ING. REGIS JORGE ALBERTO DIAZ PLAZA
ASESOR

FACULTAD DE INGENIERÍA CIVIL, SISTEMAS Y ARQUITECTURA
DECANO
UNIVERSIDAD NACIONAL PEDRO RUIZ GALLO

DR. ING. SERGIO BRAVO IDROGO
DECANO

"Año de la universalización de la salud".

CONSTANCIA DE APROBACION DE ORIGINALIDAD DE TESIS

Según Res. N° 659-2020-R

Yo, Dr. Ing. Regis Jorge Alberto Díaz Plaza**, asesor de tesis de la bachillera**:

YNES JOSSELY GAYOSO ROJAS

TITULADA:

R Y PYTHON EXPERIENCIA EN EL APRENDIZAJE DE ESTADÍSTICA DESCRIPTIVA EN LA UNIVERSIDAD NACIONAL PEDRO RUIZ GALLO

Luego de la revisión exhaustiva del documento constato que la misma tiene un índice de similitud de 15% verificable en el reporte de similitud del programa TURNITIN.

El suscrito analizó dicho reporte y concluyó que cada una de las coincidencias detectadas NO CONSTITUYEN PLAGIO. A mi leal saber y entender la tesis cumple con todas las normas para el uso de citas y referencias establecidas por la Universidad Nacional Pedro Ruiz Gallo.

Se expide la presente según lo dispuesto en la Resolución Nº 659-2020-R, de fecha 8 de setiembre de 2020 formativa para la obtención de Grados y Títulos de la UNPRG:

Lambayeque, 09 de febrero del 2023

ATENTAMENTE,

Dr. Ing. Regis Jorge Alberto Díaz Plaza
DNI. 16620941

Se adjunta:
Recibo digital de Turnitin
Revisión de informe en Turnitin

turnitin

Recibo digital

Este recibo confirma quesu trabajo ha sido recibido por Turnitin. A continuación podrá ver la información del recibo con respecto a su entrega.

La primera página de tus entregas se muestra abajo.

Autor de la entrega:	YNES JOSSELY GAYOSO ROJAS
Título del ejercicio:	pre_grado_sin_deposito
Título de la entrega:	TESIS_YNES_JOSSELY_GAYOSO_ROJAS
Nombre del archivo:	YNES_JOSSELY_GAYOSO_ROJAS.pdf
Tamaño del archivo:	1.11M
Total páginas:	68
Total de palabras:	8,452
Total de caracteres:	50,086
Fecha de entrega:	09-feb.-2024 10:02p. m. (UTC-0500)
Identificador de la entre...	2290957502

Dr. Ing. Regis Jorge Alberto Díaz Plaza
DNI. 16620941

TESIS_YNES_JOSSELY_GAYOSO_ROJAS

INFORME DE ORIGINALIDAD

15%	14%	4%	6%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	Submitted to Universidad Nacional Pedro Ruiz Gallo Trabajo del estudiante	3%
2	repositorio.unprg.edu.pe:8080 Fuente de Internet	2%
3	hdl.handle.net Fuente de Internet	1%
4	repositorio.une.edu.pe Fuente de Internet	1%

Dr. Ing. Regis Jorge Alberto Díaz Plaza
DNI. 16620941