Bases de Datos III
Tarea Nº1 (10%)
Prof: Ana Aguilera Faraco
Ayudante: Camila Araya
Agosto 2023

Instrucciones:

- Seleccione solamente un conjunto de datos entregados según estime conveniente, es libre de adecuar el dataset para el mejor entendimiento para el análisis.
- Si tiene alguna duda sobre el dataset seleccionado debe contactar al ayudante camila.arayamo@alumnos.uv.cl.
- La tarea Nº1 es grupal (máximo 3 integrantes, si su grupo de proyecto es de 4 integrantes pueden dividirse y formar grupos de 2 integrantes), y en caso de copia se aplicarán las sanciones correspondientes
- Utilice todos los recursos que ofrece Python para realizar un trabajo completo (librerías).
- El nombre del archivo debe ser "T1-NombreApellido.ipynb", cada NombreApellido debe ir separado con un guión "-" si trabaja con más de 1 integrante.
- Puntaje total: 100 puntos. Nota 4,0: 60 puntos.
- **Datasets**

Recordar las métricas vistas en clases:
- Medidas de tendencia central: Media, Mediana y Moda.
- Medidas de dispersión: Rango, varianza, desviación estándar, Máximo y Mínimo.
- Medidas de posición: Cuartiles, deciles y percentiles.
- Entre otras medidas

Presentación del dataset y cuaderno: Debe realizar una descripción del dataset que está trabajando, indicando cuáles son las variables que éste contiene, el tipo de dato de estas variables debe ser ordenado y explicativo mediante comentarios. En caso de usar librerías externas debe presentar documentación.

Apartado Nº1:
Debe realizar un análisis estadístico univariado al conjunto de datos el cual muestra el comportamiento de "al menos 2 variables estudiadas por separado" de medida y obtener métricas.

Luego debe averiguar cuántas filas de los datos contienen valores NaN. De ser así debe crear un nuevo subconjunto de datos que contenga filas solo con valores que sean significativos, para realizar esto debe establecer una estrategia de solución. Puede utilizar varios métodos de limpieza de datos NaN. Debe obtener nuevamente las métricas del nuevo conjunto de datos y trabajar con él en los siguientes apartados.

Además, debe presentar una pequeña conclusión sobre los valores obtenidos entre el primer conjunto de datos y el subconjunto de datos, e indicar cuál cree usted que es mejor para el análisis.
Debe agregar gráficos de las métricas para así obtener una mayor claridad visual del trabajo realizado, como por ejemplo gráfica de distribución normal, histogramas, gráficas de cajas, gráficas de torta etc.

Apartado Nº2:
Estudiar la relación entre dos variables del conjunto de datos mediante un análisis Bivariado y dar una hipótesis la cual debe ser planteada por usted (Recordar que para realizar este análisis se define por pasos) es pertinente que sea bien detallado en el código.

> ***Observación:*** Para medir la relación entre dos variables se define la covarianza, la cual indica si es positivo nos dice que estas se relacionarían de forma directa y si es negativa de forma inversa, la covarianza está presente en distintas fórmulas tal como se indicó en clases, una de ella es el coeficiente de correlación de Pearson. Valor que oscila entre -1 y 1, mientras más cerca a estos límites más fuerte será el grado de asociación inversa (-) o directa (+) de las dos variables.

Usando las funciones de varianza, media y covarianza hacer una recta de regresión (puede ser de utilidad un gráfico scatter). Saque conclusiones sobre su hipótesis con los resultados obtenidos.

Apartado Nº3:
En base a lo aprendido en clases realice un análisis estadístico de multivariable sobre el conjunto de datos, usted debe plantear y describir este análisis, es decir que hipótesis está realizando, cuáles son las variables involucradas etc. de tal forma que sea bien completo y explicativo, utilizar representación gráfica para mejor entendimiento de lo planteado y aplicar los contenidos vistos en clases.

Rúbrica de Evaluación

Presenta	Aspectos a evaluar	No aplica (0%)	Deficiente (30%)	Regular (60%)	Bueno (80%)	Destacado (100%)	Puntaje máximo del ítem
Calidad en la presentación del cuaderno	• Orden de código • Comentarios (buena redacción y entendibles) • Documentación (si utiliza librerías debe especificar)	No incluye estos aspectos.	Solo incluye solo comentarios y el orden de código no es adecuado.	1 aspecto no queda del todo claro y el código no tiene buen orden.	1 aspecto no queda del todo claro.	Todos los aspectos están claros.	5
Descripción DataSet	• Explica el Dataset (sobre qué tema trata) • Describe sus variables • Obtiene información del DataSet(variables,tamaño dataset etc).	No incluye estos aspectos.	No se detalla bien y no cumple con 2 aspectos.	2 aspectos están presentes, pero no detallados en profundidad.	Solo 1 aspecto no queda claro o no está presente.	Todos los aspectos están claros y detallados.	15
Análisis descriptivo univariado	• Realiza el análisis • Obtiene métricas • Limpia datos. • creación de subconjunto • Gráficos • Conclusiones	No incluye estos aspectos.	No queda claro el análisis y solo incluye 1 aspecto. (9 pto)	2 aspectos están presentes, pero no detallados en profundidad. (18 pto)	1 aspecto noqueda claro o no está presente. (24 pto)	Todos los aspectos están claros,son completos y consistentes.	30
Análisis descriptivo Bivariado	• Realiza el análisis • Obtiene métricas • Hipótesis • Descripción • Gráficos • Conclusiones	No incluye estos aspectos.	No queda claro el análisis y solo incluye 1 aspecto.	2 aspectos están presentes, pero no detallados en profundidad.	1 aspecto no queda claro o no está presente.	Todos los aspectos están claros,son completos y consistentes.	30
Análisis descriptivo Multivariado	• Realiza el análisis • Obtiene métricas • Hipótesis • Descripción • Gráficos • Conclusiones	No incluye conclusiones y no domina el tema.	Responde 1 de las 2 preguntas y concluye que son deficientes o no aplican.	Responde 2 preguntas y obtiene conclusiones, pero no domina bien el tema.	1 aspecto no queda claro o no está presente.	Todos los aspectos están claros y detallados.	20