CI-0115 Probabilidad y Estadística
Prof. Sebastián Ruiz Blais
I-2024
Grupos 2 y 3

Proyecto – Entrega 2 (10%)
Fecha de entrega: *16 de junio del 2024*

Objetivo general
Profundizar el análisis de conjuntos de datos mediante el uso de gráficas descriptivas y análisis estadísticos para comparar grupos.

Objetivos específicos
1. Estudiar una temática específica utilizando varias bases de datos.
2. Visualizar y estudiar un conjunto de datos utilizando gráficas de dispersión y diagramas de cajas.
3. Estudiar las relaciones entre 3 o más variables y construir modelos lineales para realizar predicciones.
4. Comparar poblaciones utilizando análisis de varianza (ANOVA).

Especificación
1. Para esta entrega, continúe trabajando con la misma base de datos que la entrega anterior e incluya más bases de datos para enriquecer la cantidad de datos. Procure que las nuevas bases de datos puedan ser unificadas a las anteriores por medio de alguna columna en común (como en los labs 3-4). En caso de no ser relevante, muestre con código Python cómo haría para unir dos bases de datos (aunque no las utilice).
2. Realice una lista de las preguntas que guían su análisis. Pueden ser algunas de las mismas anteriores y otras nuevas. Algunas preguntas deben buscar predecir una variable en función de una combinación de 2 o más variables (ver punto 4). Por ejemplo, la pregunta "si el día está soleado y es viernes de quincena, hay mayor compra de helados?" buscaría predecir una variable en función a dos variables. Otras preguntas deben buscar comparar entre distintas categorías de datos, como por ejemplo en los labs 3-4, se investigó si existían diferencias entre países en varios grupos de *latitud* y varios grupos de *altitud*. Deben haber por lo menos tres preguntas que comparen categorías y dos preguntas que combinen variables.
3. Utilice gráficas de dispersión para describir las distribuciones y diagramas de cajas para las relaciones entre las variables de interés. Describa las distribuciones de todas las variables de interés (mínimo 5 variables) y comente sobre su distribución. Para las relaciones, solamente enfoque las que sean más significativas (3 o más relaciones, que se relacionen con alguna pregunta).
4. Para estudiar si 2 o más variables pueden predecir otra variable utilice la regresión lineal múltiple (regresión lineal). Esto implica usar la forma $Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n + K$.

Por ejemplo, en el lab 3-4, se puede predecir la esperanza de vida con base en la combinación de las variables latitud y altitud (en lugar de tratar ambos casos por aparte). Describa tres preguntas que desea resolver de esta manera y, para cada una, las variables que considera utilizar para generar las predicciones y las variables que se busca predecir.
5. Para implementar el punto anterior, siga el tutorial https://www.youtube.com/watch?v=ZmNDRnmYaPc. Implemente las operaciones paso a paso sin utilizar bibliotecas especializadas de Python. Solamente para manipulación de matrices.
6. Además de obtener los coeficientes de la regresión lineal múltiple, indique también el error de predicción para este modelo y el conjunto de datos estudiado, obtenga el coeficiente de determinación múltiple y el coeficiente ajustado de determinación múltiple (ver referencia).
7. Invente 3 datos nuevos para intentar, con valores para las variables X usadas en la predicción pero no para la variable Y. Por ejemplo, para el lab 3-4, se puede crear los datos de latitud 9.6516711629073 y altitud de 0 msnv, que corresponden a Puerto Viejo de Limón. La idea es obtener una predicción de la variable Y. En el ejemplo, la idea sería predecir la esperanza de vida a partir de la altitud y latitud. Obtenga el valor de la variable pronosticado por la combinación lineal para los 3 datos. Esto podría dar una predicción valiosa o no, lo cual depende de la capacidad predictiva del modelo. Explique y justifique si le parecen predicción valiosas o no. ¿Cómo podría mejorarse?
8. Pruebe otra combinación de variables para predecir la variable de interés. Si usó solamente 2 variables, pruebe añadir otra variable o si usó más de 2 variables, intente eliminar la variable que considera que contribuye menos en el modelo. Obtenga los coeficientes para este nuevo modelo, así como el error de predicción y el coeficiente ajustado de determinación múltiple. Compare entre este modelo y el anterior y explique cuál de los dos modelos es mejor y de qué forma. Busque en internet cómo saber si un modelo es mejor que otro para fundamentar su respuesta.
9. A partir del conjunto de datos, separe la población en varios grupos que desee comparar, similar a lo realizado en el laboratorio 4, para las diferentes latitudes o altitudes. Para cada pregunta de investigación (por lo menos 3 preguntas), plantee por lo menos 2 ó 3 grupos distintos, dependiendo de lo que tenga mayor sentido.
10. Utilice diagramas de cajas para estudiar la distribución de cada uno de los grupos de análisis. Al igual que en el lab 4, grafique las distribuciones de cada grupo, uno a la par del otro.
11. Al igual que en el lab 4, realice análisis de varianza para comparar las medias de los grupos y poder contestar a las preguntas investigadas. ¿Provienen los grupos de la misma distribución o de distribuciones distintas? Concluya respondiendo a estas preguntas.

Evaluación

- *Calidad del análisis*: logran implementar cada uno de los objetivos del proyecto con una base de datos lo suficientemente rica. Realizan gráficas correctamente etiquetadas. Realizan cada paso correctamente. Las preguntas son relevantes.
- *Justifican cada etapa*

- *Calidad de las explicaciones*: explican los pasos realizados y las preguntas realizadas. Contestan algunas de las preguntas basándose en el análisis de los datos. Presentan los pasos de manera ordenada y usando gráficas y visualizaciones para mejorar las explicaciones. El análisis es coherente y sigue una lógica clara. Se responde a las preguntas planteadas.

Referencias

Regresión lineal múltiple:
https://www.youtube.com/watch?v=ZmNDRnmYaPc