Universidad Francisco Marroquín
Data Wrangling
Catedrático: Juan Carlos Girón
Auxiliar: José Josue

Mario Enrique Pisquiy Gómez Carné 20200399

Examen Final Data Wrangling

Instrucciones

- Usted tiene el período de la clase para resolver el examen final.
- La entrega del final, al igual que las tareas, es por medio de su cuenta de GitHub, adjuntando el link en el portal de MiU.
- Pueden hacer uso del material del curso e internet (stack overflow, etc.). Sin embargo, si encontramos algún indicio de copia, se anulará el examen para los estudiantes involucrados.

Serie Única: Conteste a las siguientes preguntas

1. ¿Qué es una expresión regular? (5 pts)

Una expresión regular es una cadena de caracteres que forman un patrón de búsqueda personalizado

2. Enumere y explique brevemente cuatro aplicaciones prácticas en las cuales las expresiones regulares son utilizadas. (5 pts)
    a. Compiladores: es usado para verificar que el código ingresado posea las palabras permitidas por el lenguaje
    b. Formularios: para verificación de que lo ingresado es valido para lo que se quiere obtener, por ejemplo, el campo de correo electrónico, verificar si en verdad es un correo electrónico
    c. Contraseñas: para verificar que contenga los caracteres permitidos y otros añadidos (como símbolos) para seguridad del usuario y la aplicación.
    d. Análisis de Texto: para encontrar patrones de palabras y combinaciones de alfanuméricos, así como combinaciones con símbolos

3. Explique brevemente las 3 condiciones que establecen que una tabla se encuentra en formato ***tidy.*** (5 pts)

    a. Cada variable es una columna
    b. Cada observación es una fila
    c. Cada tabla se compone una única unidad observacional

4. Diagnostique y explique por qué la siguiente tabla no está en formato ***tidy.*** Luego, explique cómo convertirla a formato ***tidy***. (7 pts)

Country	2008	2009	2010
Guatemala	5	9	13
United States	9	13	23
Belgium	7	13	18
Argentina	9	18	28
France	7	13	24
United Kingdom	3	3	5
Germany	10	15	27
Poland	1	2	2

Esta tabla no está en formato tidy porque las columnas 2008, 2009 y 2010 no son variables. Para pasarlo a formato tidy, lo primero que haría sería poner las columnas como filas, con un proceso de melting, es decir, que Guatemala tenga 3 filas, una con 2008, otra como 2009 y una última con 2010, con el nombre de columna "Year". Luego, en una tercera columna, podría los valores observados, quedando la tabla más o menos de esta forma:

Country	Year	Value
Guatemala	2008	5
Guatemala	2009	9
Guatemala	2010	13
United States	2008	9
United States	2009	13
United States	2010	23

5. Diagnostique y explique por qué la siguiente tabla no está en formato ***tidy.*** Luego, explique cómo convertirla a formato ***tidy***. (7 pts)

Equipo	Jugador
Real Madrid	Federico Valverde - Mediocentro
Juventus	Cristiano Ronaldo - Delantero
Barcelona	Frenkie De Jong - Mediocentro
Manchester United	Marcus Rashford - Delantero
Manchester City	Eric García - Defensa
Liverpool	Alisson - Portero
Atlético de Madrid	Joao Félix - Delantero
AC Milan	Sandro Tonali - Mediocentro
Roma	Pedro - Delantero
Inter de Milan	Achraf Hakimi - Defensa
Sevilla	Lucas Ocampos - Delantero
Valencia	Jose Luis Gayá - Defensa
PSG	Neymar - Delantero
Monaco	Cesc Fábregas - Mediocentro
Bayern Munich	Alphonso Davies - Defensa

La tabla no está en formato Tidy porque hay múltiples variables en una columna, en este caso, la de jugador, puesto que ahí también está presente la variable de "Posición". Para pasarlo a tidy, separaría los valores en la columna "Jugador", eliminando el guión (-) de paso. Quedando de esta forma:

Equipo	Jugador	Posición
Atlético de Madrid	Joao Félix	Delantero
Roma	Pedro	Delantero

6. Diagnostique y explique por qué la siguiente tabla no está en formato ***tidy.*** Luego, explique cómo convertirla a formato ***tidy.*** (7 pts)

Producto	Urabno	Rural	Q0 - Q50	Q50 - Q100	Q100 - Q500	Q500 +
Banano 12 und.	x		x			
Café molido 1 lb	x		x			
Televisión Samsung 32"		x				x
Carne Molida 5 lb		x		x		
Licuadora 1 lt	x				x	

La tabla no está en formato tidy porque los nombres de las columnas son valores, no variables, también porque en la columna producto, hay múltiples variables en la

columna, de número según forma de venta (unidades, libras, pulgadas, etc.) así como de clasificación según tipo de producto. Para pasarlo a tidy, con el Melting,lo primero que haría es pasar las columnas a filas, convirtiendo la otra columna en 2, la de valores y una columna que fue la que se "transformó", y también se separaría en dos tablas, una de producto, y otra para su clasificación, utilizando un id de producto para referencia del segundo , quedando de la siguiente manera:

Id_producto	Producto	Numero	Unidad
1	Banano	12	Und.
2	Café Molido	1	Lb

Id_producto		Check
1	Urbano	X
1	Rural	
1	Q0-Q50	X
1	Q50-Q100	
1	Q100-Q500	
1	Q500+	

7. Sobre lubridate: Explique la diferencia entre las funciones period y las funciones duration. (5 pts)

8. ¿En qué contexto utilizaría una función period y en cúal utilizaría una función duration? (5 pts)

9. Explique el concepto de data Missing Completely at Random (MCAR). (6 pts)
   El MCAR es cuando la Missing Data es que no hay relación entre la falta de data y cualquier observación, es decir, que no se puede encontrar algún patrón.

10. Si logramos verificar que la data faltante es MCAR, ¿cuál imputación recomendaría utilizar? (5 pts)
    Recomendaría utilizar la mediana, por ser el valor más céntrico de las observaciones para esas columnas, esto en el caso de ser una variable numérica, si es categórica (texto) sería la moda.

11. Si estamos realizando el análisis de una encuesta en la cual tenemos información sobre 150 individuos y tenemos valores faltantes en diferentes variables de nuestra tabla, ¿cúal de los siguientes métodos utilizaría y por qué? (6 pts)
    a. listwise deletion.
    b. pairwise deletion.
    c. outliers cap via standard deviation.
    d. outliers cap via percentile approach.

    Utilizaría el pairwise, porque hay poca cantidad de observaciones y desconozco cuantas observaciones podría perder con el listwise. Y utilizaría el outliers cap via percentile standard deviation, ya que, en este caso, puedo aproximar los valores fuera de los límites y poder utilizarlos en valores más cercanos al resto de observaciones y que no puedan alterar el modelo.

12. Usted se encuentra realizando un modelo sobre la capacidad necesaria que necesita para atender la demanda de transporte de un producto determinado. Se requiere que cumpla con el 90% de la demanda mensual. ¿Cúal de los siguientes métodos utilizaría para determinar con qué población de sus datos trabajar? (6 pts)
    a. listwise deletion.
    b. pairwise deletion.
    c. outliers cap via standard deviation.
    d. outliers cap via percentile approach.
    e. min-max scaling.

    Utilizaría pairwise deletion para la missing data, así no pierdo tantos valores de observaciones. Para los outiliers, utilizaría outliers cap via percentile approach, para poder discriminar los valores fuera de los límites superiores e inferiores.

13. ¿En qué contexto de Machine Learning se recomienda utilizar Min Max Scaling? (6 pts)

    Cuando los limites superior e inferior son conocidos, por ejemplo, la intensidad de pixeles que van en un rango de 0 a 255.

14. Si encuentra que la distribución de sus datos tiene un comportamiento exponencial, ¿cúal técnica de normalización utilizaría para transformar los datos a una distribución normal? (5 pts)

Una transformación logarítmica.

15. Si se tiene una variable categórica con tres niveles, cúantas variables dummy necesita para poder pasar la data a un modelo econométrico o de machine learning? (5 pts)
    3 variables, una para cada nivel de la variable categórica.

16. ¿En cuál contexto utilizamos one hot encoding? (5 pts)
    Cuando tenemos una variable categórica que no tienen una relación entre sí, principalmente de texto.

17. ¿Qué es un n-gram? (5 pts)
    Es una subsecuencia de n elementos de una secuencia data, por ejemplo, de palabras en una oración

18. Si quiero obtener como resultado las filas de la tabla A que no se encuentran en la tabla B, ¿cómo debería de completar la siguiente sentencia de SQL? (5 pts)

*SELECT * FROM A _LEFT_ JOIN B ON A.KEY = B.KEY _WHERE B.KEY IS NULL_*

19. Actualmente la UFM implementó la herramienta Turnitin, utilizada para detectar plagio en los entregables de los alumnos. Explique, basado en los conceptos visto en clase, el funcionamiento de este tipo de herramientas que analizan texto. (10 pts)
    En el caso de Turnitin, primero se vale de expresiones regulares para encontrar patrones en las palabras utilizadas, y luego, esos tokens reconocidos los usa para armar n-grams y analizar la estructura de los textos, esto porque no compara que el texto sea una copia 100% fiel de otro, sino que contengan palabras y estructuras similares, es decir, la misma idea puesta de otra forma, un plagio.