Universidad Francisco Marroquín
Data Wrangling 2020
Catedrático: Juan Carlos Girón
Auxiliar: José Josue

UFM UNIVERSIDAD FRANCISCO MARROQUÍN
VERITAS • LIBERTAS • JVSTITIA

Examen Final Data Wrangling 2020

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 exámen para los estudiantes involucrados.

Serie Única: Conteste a las siguientes preguntas

1. ¿Qué es una expresión regular? (5 pts)
   Análisis de una secuencia de caracteres para encontrar cierto comportamiento o validar que cumplan con características especificas. Estas se ajustan perfectamente a las necesidades del buscador pues se pueden formar completamente y así asegurarse de obtener resultados estrictamente como se están buscando.

2. Enumere y explique brevemente cuatro aplicaciones prácticas en las cuales las expresiones regulares son utilizadas. (5 pts)
   a. Buscadores de texto
   b. Validación de usuarios y contraseñas
   c. Análisis de redacción o uso del idioma
   d. Procesadores de texto

3. Explique brevemente las 3 condiciones que establecen que una tabla se encuentra en formato *tidy*. (5 pts)
   a. Cada variable forma una columna
   b. Cada observación forma una fila

c. Cada tipo de unidad observacional forma una tabla

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

La tabla no esta en formato tidy porque aun no ha sido limpiada y contiene columnas de mas. El formato tidy se asegura de que las tablas sean mas sencillas y comprensibles por lo que en este caso se convertiría en tan solo 3 columnas 1. País, 2. Año y 3. Numero. Con este formato cualquier usuario puede entender la tabla sin tener mayor conocimiento o perderse en las diversas columnas que al final son las mismas variables.

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 anterior no esta en formato tidy dado que las columnas no son directamente solo la variable indicada. En la columna de jugador esta el nombre y la posición, para que esta este en tidy format se deben establecer 3 variables: 1. Equipo, 2. Jugador, 3. Posición. Con estos ajustes también se vuelve mas sencillo analizar los datos e incluso agruparlos y graficarlos.

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 anterior no esta en formato tidy por todas las columnas con las que cuenta la misma variable. Este tipo de tablas accaban siendo confusas para quienes necesiten obtener información y realmente tener columnas separadas es completamente innecesario ya que ni siquiera indican un valor. En este caso se tendría que separar en una columna producto, una de ubicación (Urbano o Rural) y una de precio donde ponga el rango directamente. También se podría crear un

diccionario con los distintos rangos para reducir el contenido de esta columna de rango y entonces ya solo diga de 1-4.

7. Sobre lubridate: Explique la diferencia entre las funciones period y las funciones duration. (5 pts)
   Period: esta función establece un periodo de tiempo (meses, días, minutos, etc.) especifico que se repite constantemente. Un periodo va acorde con el tiempo real, es decir los días, meses, etc. Como van ocurriendo en la vida real.
   Duration: esta función mide la duración según lo que se necesite (meses, días, etc). La diferencia entre estas dos funciones es que duration toma el tiempo desde el momento que se inicia, sin importar si cuadra o no con los calendarios o relojes. Es mas como un temporizador.

8. ¿En qué contexto utilizaría una función period y en cúal utilizaría una función duration? (5 pts)
   Period se puede utilizar para medir por ejemplo los semestres, pues tiene que durar por ejemplo 5 meses desde el 8 de enero y entonces finaliza un día en mayo cuando ya pasaron los 5 meses.
   Duration se puede utilizar como el temporizador del celular, uno lo puede establecer como 15 minutos o 2 dias y te va a avisar cuando paso exactamente este periodo de tiempo, aunque sea a medio minuto.

9. Explique el concepto de data Missing Completely at Random (MCAR). (6 pts)
   MCAR se define como datos que no se tienen de distintas tablas sin un patrón especifico, la probabilidad de que falten datos es igual en todas las columnas. Pueden faltar datos de cualquier variable y se deben de manejar todas las faltantes por igual.

10. Si logramos verificar que la data faltante es MCAR, ¿cuál imputación recomendaría utilizar? (5 pts)
    La imputación recomendada seria Listwise deletion, que elimina la fila completa donde hay un faltante. Como no es representativo de que columna va a faltar se eliminan completas, con esto se asegura mantener el tamaño del dataset parejo en todas las variables.

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.

Dado que se cuenta con muy pocos datos es muy caro para quienes están analizando perder columnas enteras o lograr encontrar claramente comportamientos para normalizar el resto, por ende, yo utilizaría pairwise deletion, para asegurarse de perder solo los datos que no están, adicionalmente como son datos de individuos y no se va a profundizar dentro de cada uno sino dentro de cada variable, este método no afectaría el comportamiento de los resultados.

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 el método de outliers cap via standard deviation, con este modelo se conoce cual es el comportamiento de los datos según su desviación y se puede establecer en base a aquellos que van a cumplir la demanda, es decir dejando fuera a todos los otliers que no serian constantes. Es un método altamente confiable cuando se conoce a la población y se quieren establecer limites claros de para mayor confianza en los datos.

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

Min Max Scaling sirve para normalizar datos en una escala muy amplia y reducirlo para que sea más analizable, típicamente se pasa a una escala de 0-1. Esto sirve en algunos casos para medidas de distancia muy variables y distintas. También podría utilizarse para normalizar la escala de colores de una imagen que en muchos casos va de 0 a números mucho mayores.

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)

    Utilizaría el modelo de normalización de Log Transformation ya que con distribuciones logarítmicas las vuelve mas cercanas a una distribucion normal sesgando los datos a las características de las variables del dataset.

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)

    Se necesitarían tres variables dummy para marcar cada una de las variables categóricas del dataset, para que con los modelos de machine learning se puedan analizar y predecir los datos.

16. ¿En cuál contexto utilizamos one hot encoding? (5 pts)

    One Hot Encoding se utiliza para analizar cuando se quiere analizar variables categóricas separadas y que sea más fácil predecirlas para ciertos algoritmos. Un ejemplo de cuando se podría usar es en una venta de automóviles con distintos modelos y precios donde se quieren analizar estos mismos.

17. ¿Qué es un n-gram? (5 pts)

    Son las divisiones en que se separa el texto cuando se realizan análisis como de expresiones regulares. Por ejemplo, puede estar dividido individualmente (N = 1) o mas grams . Esto sirve para entender el comportamiento y poder predecir que palabras serán usadas más adelante según los datos de lo que se ha encontrado previamente.

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*