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

Nombre: Ricardo Sebastián Pineda de León
Carné: 20160164

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)

   Es una secuencia de caracteres que conforma un patrón de búsqueda.

2. Enumere y explique brevemente cuatro aplicaciones prácticas en las cuales las expresiones regulares son utilizadas. (5 pts)

   1. Desarrollar y ejecutar programas informáticos
   2. Construcción de compiladores
   3. Buscar secuencias determinadas de caracteres para identificarlas o incluso reemplazarlas por otra secuencia
   4. Limpieza de datos. Identificar caracteres que no sean validos dentro de una columna de texto por ejemplo, eliminar dichos caracteres para que la columna esté en un formato válido

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

    1. Todas las variables forman una columna
    2. Todas las observaciones forman una fila
    3. Cada tipo de unidad 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

El problema con esta tabla es que no se tiene ninguna indicación sobre que significa cada uno de los números. Se que Guatemala obtuvo, 5, 9 y 13 pero no sé a qué se refiere.
Además en esta tabla, las observaciones forman las columnas, no las filas, por lo que hay que efectuar una transposición a la tabla para que cumpla con las reglas tidy.

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

Esta tabla no está en formato Tidy debido a que la posición de los jugadores está en la misma columna que el nombre del jugador. Para convertirla a formato tidy, es necesario hacer una nueva columna 'posición' que contenga la posición de cada jugador, así como eliminar la posición de la columna 'jugador' y renombrar dicha columna a 'nombre'.

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	

En primer lugar hay un typo. Hay que renombrar la columna 'urabno' a 'urbano'. Luego sería mejor, en vez de tener x marcando las que si son, introducir una distinción binaria para identificarlos como True y False [0,1]

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

   La diferencia está en como maneja los tiempos. Por ejemplo, si tuviéramos la fecha 31 de enero, y a este le sumamos 1 mes, podríamos tener dos resultados distintos:

   31 de enero + 1 mes = 28 de febrero [Period]
   31 de enero + 1 mes = 3 de marzo [Duration]

   Depende de la aplicación que le demos, la función que hay que usar.

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

   Si se tiene una función que necesitamos que se corra el último día de cada mes, usaría Period pues se que solo hay que sumarle un mes para tener el último día del siguiente mes.

   Por otro lado, si necesito un cálculo que se realice cada 30 días, pues es la definición interpretada al hablar de 'un mes', entonces usaría duration, pues al sumarle un mes, le estaría sumando 30 días.

9. Explique el concepto de data Missing Completely at Random (MCAR). (6 pts)

   Indica que los valores faltantes están en puntos completamente aleatorios. Es decir, no hay nada que indique diferencias entre los valores faltantes y los que si están. La probabilidad de que las clases tengan valores faltantes es la misma.

10. Si logramos verificar que la data faltante es MCAR, ¿cuál imputación recomendaría utilizar? (5 pts)

    Imputación múltiple

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.

    La realidad es que tendría que saber exactamente el formato de la encuesta para dar una respuesta certera, pero dado que solo puedo asumir, argumentaría que la mejor decisión es usar pariwise por que solo se tienen 150 observaciones, y botar las filas nulas podría eliminar demasiada data y quitar toda relevancia estadística.
    Así que con pariwise no perderemos todos esos datos para todos los modelos, solo para aquellos en los que hayan datos faltantes.

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 este método para agarrar los datos relevantes para alcanzar como mínimo el 90% de la demanda. Si se quitan las colas de una distribución normal solo estaríamos perdiendo valores atípicos, logrando así un analisis mas certero para lo que nos interesa.

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

Min Max Scaler tiene un uso ideal cuando se conocen los limites superiores o inferiores. Por ejemplo en un rango de densidad de pixeles que van de 0 a 255 en la escala RGB.

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)

Transformación logaritmica

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)

Si se tuviera una variable categoría con tres variables, por ejemplo:

Colores
Rojo
Verde
Azul

Entonces, si aplicaramos variables dummies, la tabla quedaría de la siguiente manera:

Rojo	Verde	Azul
1	0	0
0	1	0
0	0	1

Sin embargo, se podría hacer una distinción binaria entre dos colores, e indicar que si no tiene asignado uno de los valores binarios, entonces es el tercer color. Es decir, Rojo = 0, Verde = 1, Azul = Na

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

Para datos categóricos que no tienen una relación 'ordinaria', se aplica un one hot encoding que tenga una representación numérica.
Dichas representaciones son usualmente llamadas dummies o variables dummy

17. ¿Qué es un n-gram? (5 pts)

Es una secuencia continua de *n* objetos de una muestra dada. Dicha muestra suele usualmente ser texto o una transcripción de speech [Speech to Text]
Son utilizados en NLP

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 ;*