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

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)

Expresiones regulares son patrones en cadena para buscar y reconocer texto (caracteres). Tambien se usan para hacer operaciones de sustitucion. Hay diferentes operadores de expresiones regulares, como por ejemplo el punto, la barra inversa o corchetes.

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

I. En buscadores como Google uno puede buscar de forma que NO aparezca una palabra. Ej: "Ellen -degeneres" busca todo lo que sea ellen pero que no contenga degeneres. Si uno buscara únicamente escribiendo "ellen", los primeros resultados son de DeGeneres.
II. Se puede usar para encontrar palabaras y sustituirlas. Por ejemplo, para cambiar el nombre de archivos. Esto quiere decir que puede encontrar U, o Uni, o Universidad.

III. Se puede usar para validar el DPI, número de pasaporte, número de carnet de un estudiante, número de teleéfono, etc.
IV. Se puede usar al momento de registrar un correo para validar que la contraseña contenga el mínimo de caracteres, más una mayuscula, un número y un simbolo raro.

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

I. Cada variable es una columna. Es decir que no pueden haber por ejemplo varias columnas correlacionadas con variables dummy.
II. Cada observación forma una fila. Puede representar un problema si una observacioón esta dispersa en varias filas.
III. Cada tabla se conforma de una 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

No está en formato tidy porque deberiía ser una variable por columna, pero en este caso tenemos tres columnas con el resultado de tres años.
Para convertirlo a tidy en R usaría la libreria "tidyr" usaría gather() o mas bien pivot_longer(), ya que el anterior realmente ya no se usa.

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

No esta en formato tidy porque en una misma columana hay dos informaciones diferentes. En la columna "Jugador" dicen tanto el nombre del jugador como la posición en la que juega.
Convertiría esta columna en dos utilizando separate(). Luego convertirira la posición en una variable nominal para convertirla a un valor numérico.

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	

No esta en formato tidy por varias razones. Uno porque el lugar de produccion (urbano o rural) está separada en dos culumnas como variabales dummy, cuando debería ser solo una. Luego los rangos de precio también están distribuidos en varias columnas como dummys. Para converitlo usaria pivot_longer para crear una tambla con una columna con el nombre del producto, una columna con lugar de producción (rural o urbano) y una tercera columna con precios (0-50,50-100,etc).

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

"Period" se usa para convertir un valor a la unidad que uno necesite, ya sean segundos, horas, dias, semanas, etc.
"Duration" devuelve la cantidad de segundos que tiene un valor.
La diferencia es que period trabaja con unidades maás grandes que segundos, mientras que duration trabaja únicamente en segundos, aunque tambien presenta un resultado facil de entender para la persona,
Ej: duration("2d 2H 2M 2S") = "180122s (~2.08 days)".
Además, esto hace que period trabaje con intervalos de tiempo que son distintos, por ejemplo febrero no tiene la misma cantidad de dias que marzo o abril.

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

Se usa *duration* cuando se quieren tener resultados matemáticos con resultados precisos. Pero se usa *period* cuando se quiere tener en cuenta fluctuaciones en el tiempo, ya que como se explicó en la pregunta anterior el tiempo fluctua y un mes no es igual a otro. Por ejemplo si queremos tomar en cuenta que un año es bisiesto, se usa *period*.

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

Hay diferentes tipos de missing data. Este en particular quiere decir que los datos que faltan no faltan porque la cuestion estudiada no haya proporcionado los datos, sino que estos se perdieron en el camino de forma random, por ejemplo por un virus que se come los datos, o por un mal procesamiento de datos en el laboratorio.

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

Recomendaría una imputación general por la media.

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.

Eligiría b.) pairwise deltion. Esto me permite que aunque no pueda trabajar con una variable en especiífico, que pueda seguir analizando todas las otras variables de una tabla sin necesidad de borrar la observación completa.

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.

Eligiría c.) outliers cap via standard deviation
De esta forma puedo elegir cuantas desviaciones estandar mantener, por ejemplo puedo elegir dos desviaciones estandar y quedarme con el ~95% de los datos.

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

Min-Max Scaling producen rangos entre 0 y 1. Se usa cuando hay unidades que varían bastante en magnitud entre los valores mas pequeños y los grandes, de forma que por ejemplo al realizar un promedio, los numeros muy grandes tienen muchímo peso.

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)

Usaría coefficient of variation para normalizar la data.

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)

Como son tres categorías se necesitan tres variables binarias o dummy para poder pasar la data a un modelo econométrico o de machine learning.

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

Se usa one hot encoding para poder usar variables categoricas en modelos que requiern de un input numérico, como por ejemplo modelos de machine learning. Se usa cuando no existe ninguna relación entre las variables (ej: rojo, azul, verde).

17. ¿Qué es un n-gram? (5 pts)

Se usa en Text Mining. N-grams une palabras que están juntas para entender la secuencia de elementos. La N representa la cantidad de palabras, entonces un 2-gram sería "manzana roja", un 3-gram sería "sal y pimienta", etc. Esto sirve para entender la probabilidad de que una palabra suceda dada otra palabra. Por ejemplo en Whatsapp o Google que sugiere la siguiente palabra dependiendo de lo que se acaba de escribir.

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 ___ JOIN B ON A.KEY = B.KEY _____*

SELECT * FROM A LEFT JOIN B ON A.KEY = B.KEY IS NULL