Universidad Francisco Marroquín
Data Wrangling
Catedrático: Juan Carlos Girón
Auxiliar: José Josue
Gabriela Martínez - 20210219

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 secuencia de caracteres que forman un patrón de búsqueda para proporcionar una manera de buscar y reconocer cadenas de texto en un lenguaje formal.

2. Enumere y explique brevemente cuatro aplicaciones prácticas en las cuales las expresiones regulares son utilizadas. (5 pts)
   - Búsqueda de cadenas de proteínas
   - Procesadores y buscadores de texto
   - Analizadores léxicos de compiladores de lenguajes
   - Verificación de contraseñas

3. Explique brevemente las 3 condiciones que establecen que una tabla se encuentra en formato *tidy*. (5 pts)
   - Cada fila debe corresponder a una sola observación
   - Cada columna debe de tener data de una sola variable
   - Cada registro debe de corresponder a un evento

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

Se asume que esta tabla está hablando de ventas de un producto según el país. La tabla no se encuentra en formato tidy ya que se tienen diferentes columnas para describir una sola variable: ventas. Para convertirla a formato tidy primero hay que generar una columna titulada “año” donde se registre de que año se esta hablando (2008,2009,2010) y luego se generaría otra columna titulada “ventas” donde se meterían los datos de venta según el país y el año, permitiendo que la tabla crezca para abajo en vez de para los lados.

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 ya que en una sola columna se están registrando dos variables: "jugador" y "posición". Para poder pasarla a formato tidy primero hay que hacer un substr() para obtener el nombre de la posición. Se modifica la columna de "jugador" para que tome en cuenta todo antes del "-" (el nombre del jugador) y se genera otra columna llamada "posición" donde se coloca todo lo que se obtuvo con el substring después del "-", la posición en la que juega.

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	

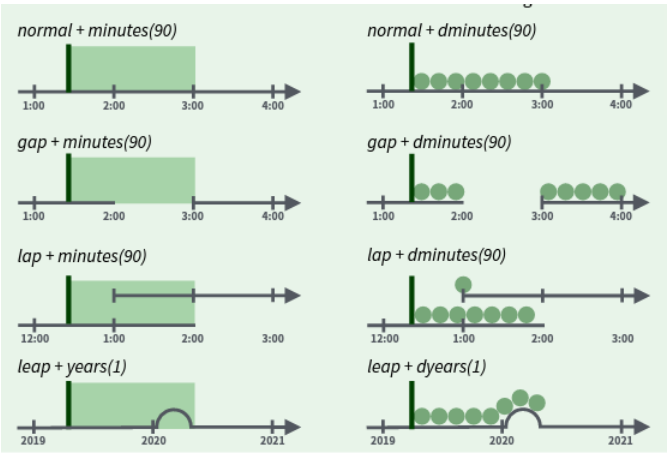
Esta tabla no se encuentra en formato tidy ya que los datos en las columnas no describen las características de las variables registradas y de la misma manera podría llevar a un caso de multicolinealidad. Para poder pasarlo a formato tidy hay que generar una columna llamada "zona" y colocar la

palabra “urbana” para todos los registros que tienen ”x” en la columna de “Urbano” y la palabra “rural” para todos los registros que tienen “x” en la columna de “Rural”. Luego se genera otra columna llamada “rango_precio” donde se pueden establecer categorías ya sea numéricas o con palabras para describir cada rango de precio por ejemplo: 1 de Q0-50, 2 de Q50-100, 3 de Q100-500 y 4 de Q500 o más. Luego se coloca el número de categoría en la columna “rango_precio” según en que columna este la “x” de cada registro.

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

   Period sigue los cambios según las horas del reloj, es decir, que ignora las irregularidades en una línea de tiempo. En cambio, duración mide el paso del tiempo físico, por lo que si puede desviarse del tiempo que marca el reloj a causa de alguna irregularidad.

Period duration



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

   Period se utiliza cuando se quiere medir algo según pasan las horas del reloj, es decir cuando se quiere medir el paso del tiempo sin que nada lo afecte o eventos que pasan según las horas del reloj, como por ejemplo cuando baja la bola de Año Nuevo en Times Square. En cambio duration se utilizaría cuando se quiere sumar o restar periodos de tiempo a procesos físicos, como la vida de una batería de un celular.

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

Este concepto explica un tipo de missing data que ocurre cuando la probabilidad de que la data se haya perdido es totalmente independiente de todas las variables en dataset que se esta trabajando, incluyendo la variable de missing data. Es decir que pasa de manera "random" y no depende de características observadas o no observadas de la data recolectada.

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

    Se recomendaría listwise o pairwise deletion.

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.

    Se recomienda pairwise porque no compromete la data ni sacrifica una gran cantidad de datos.

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.

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

    Se recomiendta utilizar min max scaling cuando se están utilizando modelos de aprendizaje automático basados en gradientes (como regresión lineal,

regresión logística) o cuando se utilizan algoritmos de aprendizaje que son sensibles al feature scaling, como k-nearest neighbors. En cuanto a los datos en general, se recomienda utilizar cuando hay una gran diferencia en la magnitud de las características del conjunto de datos o cuando los datos de una variable no están en una escala natural.

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)

    Se recomendaría log transformations 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)

    Se deberían de tener dos variables dummies para que describa los primeros dos niveles de la categoría. Como se vuelven categorías binarias si las dos tienen un valor de 0 se asume que ese registro pertenece a la tercera categoría. Esto se hace para evitar problemas de multicolinealidad.

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

    Se utiliza one hot encoding cuando se tienen categorías para diferentes variables pero estas están escritas de con palabras, por lo que no pueden ser ingresadas para entrenar un modelo. One hot encoding lo que permite es que se le agregue un valor numérico a cada una de estas categorías para poder ya entrenar al modelo con esa información.

17. ¿Qué es un n-gram? (5 pts)

    Una secuencia continua de n cantidad de objetos en un cuerpo específico de texto utilizada para capturar diferentes niveles de información de cierto lenguaje.

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 Excluding JOIN B ON A.KEY = B.KEY id

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)

    Analiza el texto subido y lo compara con una base de datos de textos académicos o ya presentados para averiguar si encuentra algún "match". Analiza los n-grams del texto subido para poder comparar secuencias de palabras y su similitud a otros papers o textos ya publicados. También analiza patrones de texto y vocabulario para poder identificar cualquier que sea sospechoso o de indicios de plagio y de la misma manera también analiza el estilo de escritura para observar si es similar a un patrón o estilo de algún otro autor en su base de datos.

20. Utilizando el dataset de "Student Performance", realice una presentación respondiendo alguna de las siguientes preguntas (10 pts)

    a. ¿Cuál es el efecto de la dieta del estudiante antes de la prueba?
    b. ¿Existe alguna diferencia entre grupo de estudiantes (gender/race) al estar previamente preparados?
    c. ¿Existe alguna relación entre los resultados de matemáticas, lectura y escritura para los diferentes grupos de estudiantes (gender/race)?