Universidad Francisco Marroquín

Data Wrangling

Catedrático: Juan Carlos Girón

Auxiliar: José Josue

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)

Es una secuencia de caracteres que forma un patrón de búsqueda es principalmente utilizada para la búsqueda de patrones de cadenas de caracteres dentro de texto y operaciones de sustitución a esto me refiero que en la programación/procesamiento de testo para realizar búsquedas y remplazo de complejos de texto según el patrón que fue definido.

1. Enumere y explique brevemente cuatro aplicaciones prácticas en las cuales las expresiones regulares son utilizadas. (5 pts)
   1. Web Scraping: Las expresiones regulares son fundamentales para procesar datos textuales desordenados en la extracción de información de sitios web, como buscar títulos de artículos en blogs o extraer datos específicos de páginas web.

b. Validación de correos electrónicos: Permiten verificar que las direcciones de correo electrónico cumplan con un formato estándar, incluyendo la presencia de un dominio válido y la estructura correcta de la dirección.

c. Análisis léxico en un compilador: En la construcción de compiladores, las expresiones regulares se utilizan para definir patrones que el analizador léxico emplea para identificar tokens o lexemas válidos del lenguaje de programación.

d. Validación de formatos de contraseñas: Se utilizan para asegurar que las contraseñas cumplan con criterios específicos de seguridad, como la longitud mínima y máxima, la inclusión de caracteres especiales, mayúsculas, minúsculas y números.

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

a. Cada variable forma una columna: Esto significa que cada columna del conjunto de datos representa una variable diferente. Es decir que cada variable debe tener su propia columna

b. Cada observación forma una fila: Cada fila del conjunto de datos representa una observación distinta, que puede ser un punto de tiempo, una entrada, etc. Es decir que cada observación debe tener su propia fila

c. Cada tipo de unidad observacional forma una tabla: Si existen diferentes unidades observacionales, cada una debe estar en su propia tabla para mantener la claridad y la estructura.

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



La tabla no se encuentra en formato Tidy porque los nombres de las

columnas no son nombres de variables, sino que son valores de una

variable. En este caso los nombres de las columnas 2008, 2009 y 2010

representan valores de la variable año, los valores en dichas columnas

representan valores de la variable ‘casos’ y cada fila representa tres

observaciones no solo una. Para transformarla a formato tidy tendría que

crear nuevas columnas para asignarle variables de “año” y “casos” y

respectivamente hacer una sumatoria de los casos por año para que sea

reflejado en la columna de casos.

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



La tabla no es considerada tidy porque se puede observar que la

posición se encuentra a la par del nombre, por lo que debería tener una

variable aparte para establecer la posición del jugador ya que no se

pueden tener dos variables en la misma columna. Para arreglarla se

debería crear una columna con la variable posición. es decir, dividir la columna "jugador" en dos una para el nombre del jugador y otra pasa su posición, luego al dividir la columna renombrarlas como se debe podría ser: "Nombres del jugador" y "posición” garantizando que cada columna representa una sola variable y asegurarse que cada fila contenga una única observación. sin agrupar múltiples observaciones en una sola fila.

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



Encabezados de Columna como Valores: Las categorías de precios (Q0 - Q50, Q50 - Q100, etc.) son valores y no variables, y están representadas como encabezados de columna. Así mismo los indicadores 'Urbano' y 'Rural' parecen ser variables binarias que están codificadas como 'x' en las filas en lugar de estar representadas en columnas propias. y por ultimo Las filas no representan observaciones únicas. Por ejemplo, si 'x' significa disponibilidad, entonces cada 'x' es una observación distinta que no está claramente separada en la tabla. Cambiar las 'x' por valores binarios 0 o 1 en columnas separadas para indicar la disponibilidad del producto en áreas urbanas y rurales, convertir los rangos de precios en valores de una nueva columna 'Rangos de Precio’. Y como en el anterior asegurarse de que cada fila sea una observación única con el producto, su disponibilidad en urbano o rural, y el rango de precio asociado.

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

a. Duration mide la cantidad exacta de tiempo entre dos momentos,

funciona como un cronometro y es independiente a una fecha de inicio,

mientras que periods mide de forma precisa los tiempos del “reloj” o

“calendario” sin tomar en cuenta años bisiestros o ‘day light savings’

independiente a una fecha de inicio (interpretración más humana del

tiempo).

Duration se usa para medidas de tiempo precisas y consistentes, mientras que period se usa para medidas de tiempo que son más relevantes para el calendario y la forma en que las personas piensan acerca del tiempo.

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

Durations lo utilizaría cuando necesito saber tiempo exacto por ejemplo la

duración exacta con segundos de un tiempo de aterrizaje que mide esa precisión o el tiempo que un satélite tarda en orbitar la Tierra mientras que period

la utilizan en un contexto donde la fecha no sea de gran importancia la

exactitud, como por ejemplo calcular la fecha en la que se realizó cierta acción

Más Humana check in Airbnb)

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

MCAR consiste en la propensión de que falte algún punto de datos es

completamente aleatoria. Es decir, la probabilidad de que falten

valores en una variable es la misma para todas las muestras, No existe

una relación entre la ausencia de un dato y cualquier valor del dataset,

perdido u observado

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

Imputación por la media o la mediana: Para variables continuas, reemplazar los datos faltantes con la media o la mediana de los datos observados puede ser una opción simple y efectiva, especialmente cuando la proporción de datos faltantes no es muy grande.

1. 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)
   1. listwise deletion.
   2. pairwise deletion.
   3. outliers cap via standard deviation.
   4. outliers cap via percentile approach.

Elegiría listwise deletion (eliminación por listas) si los datos faltantes son mínimos y parecen ser MCAR, ya que este método es más directo y menos propenso a sesgos que pueden surgir con el pairwise deletion. El listwise deletion utiliza solo observaciones completas, lo cual es particularmente importante si se van a realizar análisis que requieren la integridad de todos los datos en todas las variables, como la regresión múltiple.

1. 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)
2. listwise deletion.
3. pairwise deletion.
4. outliers cap via standard deviation.
5. outliers cap via percentile approach.
6. min-max scaling.

Este enfoque permite limitar los valores extremos de demanda que podrían ser poco comunes. Al utilizar un límite basado en percentiles, específicamente el percentil 90 en este caso, se puede determinar la capacidad necesaria para cumplir con al menos el 90% de la demanda mensual histórica. Los valores atípicos que excedan este umbral podrían ser excepciones no representativas de la demanda regular y, por lo tanto, podrían no ser pertinentes para la planificación de la capacidad estándar.

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

En algoritmos que requieren o asumen que las características de entrada están en un rango específico definido.

1. 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)

La técnica más adecuada para utilizar sería la **transformación logarítmica.** Esta técnica es efectiva para manejar datos con distribuciones sesgadas, como las exponenciales.

La transformación logarítmica funciona al aplicar el logaritmo a cada valor de tus datos. Esto tiene el efecto de "aplastar" o reducir la variabilidad de los datos que están muy dispersos o que tienen un sesgo a la derecha (como es común en las distribuciones exponenciales). Como resultado, los datos transformados a menudo se alinean más estrechamente con una distribución normal, lo que facilita su análisis y la aplicación de técnicas estadísticas que asumen la normalidad.

1. 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)

Para una variable categórica con tres niveles, se necesita crear dos variables dummy para incorporarla en un modelo econométrico o de machine learning. Ya que la regla de "N-1", nos dice que N es el número de categorías. Se crea una variable dummy menos que el número total de categorías para evitar la multicolinealidad y asegurar que el modelo sea interpretable.

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

One hot encoding es utilizado para transformar variables categóricas a

vectores binarios en categorías. Un contexto donde se tendría que

utilizar one hot enconding sucede cuando la característica categórica

no es ordinal, también cuando la cantidad de características

categóricas es menor por lo que se puede aplicar one hot enconding de

manera efectiva.

1. ¿Qué es un n-gram? (5 pts)

Los n grams de texto se utilizan amplicamente en ‘text mining’ y en las

tareas de procesamiento de lenguaje natural. Son un conjunto de

palabras que coexisten dentro de una ventana determinada (secuencia

de palabras N), al calcular los n-grams normalmente se avanza una

palabra.

1. 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

1. 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)

Turnin extrae y limpia el texto de los documentos. Luego, aplica técnicas de procesamiento de texto, como tokenización y normalización, para descomponer el texto en elementos manejables.

Estas herramientas también emplean algoritmos de comparación y patrones de búsqueda, similares a las expresiones regulares que vimos en clase, para identificar coincidencias entre el texto suministrado y una extensa base de datos. La detección de similitudes y paráfrasis puede implicar algoritmos de comparación de cadenas y análisis de frecuencia de palabras, aspectos que se asemejan a las técnicas de análisis y manipulación de datos.

Finalmente, genera informes de similitud y el resumen de resultados es análoga a la visualización y presentación de datos en Data Wrangling, donde se transforman los datos crudos en información comprensible y útil.

1. Utilizando el dataset de “Student Performance”, realice una presentación respondiendo alguna de las siguientes preguntas (10 pts)
   1. ¿Cuál es el efecto de la dieta del estudiante antes de la prueba?
   2. ¿Existe alguna diferencia entre grupo de estudiantes (gender/race) al estar previamente preparados?
   3. ¿Existe alguna relación entre los resultados de matemáticas, lectura y escritura para los diferentes grupos de estudiantes (gender/race)?