Universidad Francisco Marroquín

Data Wrangling

Catedrático: Juan Carlos Girón

Auxiliar: José Josue

Examen Final Data Wrangling

**Cruz del Cid**

**20200394**

**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 expresión formal para representar una cadena de caracteres por medio de sus características. En una expresión regular podemos definir las características o el formato que debe tener un texto que vayamos a utilizar. Por ejemplo, como que inicie con tres números y que tenga una o muchas letras y termine con un numeral. Estas características definidas definidas en las expresiones regulares s

1. Enumere y explique brevemente cuatro aplicaciones prácticas en las cuales las expresiones regulares son utilizadas. (5 pts)
   1. Validación entradas en un programa, como contraseñas o textos en formularios. Por ejemplo, que la contraseña contenga una mayúscula o que el correo esté bien escrito (con @ antes del donminio).
   2. Tokenización de conversaciones para su análisis. Con esto se busca descifrar el desarrollo la conversación a lo largo del tiempo y cómo se han desarrollado la relación de quienes participan.
   3. Compilación de programas de un lenguaje a otro. La tokenización del código por medio de expresiones regulares es la primera etapa de la compilación. Que significa pasar un código de un lenguaje a otro manteniendo su funcionamiento.
   4. Análisis de reportes financieros para conocer si el resultado es bueno o malo. Esta aplicación está hecha para ahorrar el tiempo que se dedica a entender el estado financiero de una empresa a partir de unos análisis de texto. El problema es que en los reportes financieros se trata de esconder el estado de la empresa entre mucho texto.
2. Explique brevemente las 3 condiciones que establecen que una tabla se encuentra en formato ***tidy.*** (5 pts)
   1. Que cada variable debe estar representada en una columna. Las variables son todos los valores que pertenecen a la misma unidad. Por ejemplo, el color, en una tabla de automóviles.
   2. Cada observación debe ser una fila. Las observaciones son los valores de un mismo elemento u objetivo medido a través de varias unidades. Por ejemplo un automóvil, con modelo, año, color y placa.
   3. Y cada tabla debe estar compuesta por una sola unidad observacional. Un ejemplo de esto es que si tenemos una tabla con las canciones y el ranking semanal (dos unidades observacionales), se deben separar en dos tablas. Una solo para las canciones (su nombre y otra información) y otra con el ranking de las canciones de cada semana.
3. Diagnostique y explique por qué la siguiente tabla no está en formato ***tidy.*** Luego, explique cómo convertirla a formato ***tidy*.** (7 pts)



Problema: Los nombres de las columnas 2008, 2008 y 2010 son valores y no etiquetas de variables.

Pasos:

* En un nuevo dataframe crearía las columnas: country, year y value.
* Llenaría la columna de country con los nombres de los países.
* Pasaría los años a la columna de year, repetiría cada país en las observaciones la cantidad de años que hay, en este caso, hay 3 años. Por lo tanto, cada país aparecería 3 veces una vez por cada año.
* Por último, llenaría la columna de value con el valor de la celda interceptada entre país y año para cada fila generada anteriormente.

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



Problema: existen múltiples variables en una columna. El nombre de jugador y su posición están en la misma variable.

Pasos:

* Crear una nueva columna llamada **posición**.
* Hacer un split o una separación por “ - “ de la columna jugador.
* Asignar el primer elemento del split a la columna jugador (donde iría el nombre)
* Asignar el segundo elemento del split en la columna **posición**.

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



Problema: Los nombres de las columnas Urbano, Rural y el resto de columnas de rangos de dinero son valores y no etiquetas de variables.

Pasos:

* Crear las columnas área y rango.
* Colocar “urbano” en la columna **área** done la x esté marcada en la columna urbano y “rural” en la columna **área** done la x esté marcada en la columna rural.
* Haría lo mismo en la columna rango, pondría el nombre del nombre de la columna que esté marcada con “x” en la columna **rango**.

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

* Period: el periodo toma las fechas en términos relativos, en una compresión más parecida a la humana. Por ejemplo, hacer una cita para dentro de un mes, aunque la duración de los meses no siempre sea la misma, la fecha de la cita será el mismo día.
* Duration: Se utiliza para calcular la duración exacta de un evento en términos reales. Por ejemplo, que la llegada de un paquete sea en 30 días, aquí se toma independientemente de la duración de cada mes, por lo que la fecha de llegada no necesariamente es igual a la fecha (en día) en que el que se hace el pronóstico.

1. ¿En qué contexto utilizaría una función period y en cúal utilizaría una función duration? (5 pts)
   1. Period:
   2. Duration: utilizaría esta función en el caso en que necesite predecir en evento exacto, por ejemplo el paso de un cometa, por tengo que calcular la fecha sin importar la percepción.
2. Explique el concepto de data Missing Completely at Random (MCAR). (6 pts)

Este es un estimador que nos indica estadísticamente si la perdida de datos o los datos faltantes se debe simplemente por un efecto al azar o si existe algún problema o alguna perdida de información de otra forma. También esto significa que los datos faltantes pueden ser represtados correctamente por el resto de información que sí temenos.

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

La imputación que haría en este caso es por promedio, porque se sabe que el resto de datos de la columna o variable es representativa de estos datos que faltaran, por lo tanto no cambiaríamos muchos la distribución y tampoco sesgaríamos tanto la data.

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.

Utiliazaría esta opción en entre las que se propusieron porque temenos poca información 150 observaciones no son muchas para poder hace un listwise, por lo tanto conservaría los pares de información que sí están completes para utilizer la información disponible lo mejor porisble. Y con los otros 3 metodos se pierde mucha data.

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.

Utilizaría este método porque así conseguiría el 90% de las observaciones que es la demanda con la que quiero cumplir. El resto de información (los outliers) no serán tomados en cuenta.

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

Se recomienda cuando la diferencia entre las observaciones es muy grande, por ejemplo, que tengamos unidades y millones en la misma columna. Y busca reducir los saltos en **Poorly scaled function** para encontrar con menos iteraciones la minimización del error de los modelo. Cuando temenos la escala normal y esta es muy grande es más difícil minimizer el error del modelo.

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)

Una normalización logarítmica para pasar la distribución de los datos a algo más parecido a una distribución normal. El único requisito es que todos los datos sean positivos.

1. Si se tiene una variable categórica con tres niveles, cuantas variables dummy necesita para poder pasar la data a un modelo econométrico y de machine learning? (5 pts)

En un modelo econométrico solo se necesitan dos para cuidar si se colocan 3 se puede eliminar la heteroscedasticidad del modelo, que es lo mismo la independencia lineal de las variables, necesaria para regresiones lineales.

Para un modelo de machine learning utilizaría 3, una por cada categoría porque no es necesario tener independencia lineal en las variables, o el algo que no afecta en el modelo.

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

Cuando tenemos variables categóricas porque este tipo de variables no es entendible por sí sola la maquina, nosotros debemos darle un significado. Se puede hacer utilizando variables dummys o también asignando un valor de importancia a cada categoría si es posible.

1. ¿Qué es un n-gram? (5 pts)

Es una secuencia de n elementos. Dependiendo de la n, es el tamaño de elementos consecutivos que tiene el gran. Se utiliza mucho en los teclados de los teléfonos para saber qué palabra o elemento sigue después de otro en una oración y dar recomendaciones al usuario para que no tenga que escribir tanto.

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)

Una de los análisis que hace es ver que tan únicas son las combinaciones sucesivas de palabras o n-grams en el documento que se está analizando en comparación con el resto de documentos que están en la base de datos. Esto nos da como resultado un ratio, que nos indica la proporción de particularidad que hay en el documento escaneado, para decir que cuál es el porcentaje de plagio.