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)

Una expresión regular es un conjunto de caracteres que nos establecen un resultado esperado, es decir, nos permiten establecer que valores esperamos y cómo identificamos a ese conjunto de carateres.

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

* Verificación de constraseñas: en este caso, se usa la expresión para validar el “formato” permitido de la contraseña, por ejemplo: poner sólo 8 caracteres, con al menos un número y un símbolo (#$%&\_-) . Clave: 12Carlo$.
* Buscadores de texto: En este caso, buscamos palabras, frases, oraciones que contengan ciertas palabras, etc. La idea es establecer la expresion regular que buscamos encontrar o que contenga la cantidad de valores “especiales” que buscamos encontrar. A veces se usa para elminar palabras que son muy repetitivas que no deseamos analizar (el, por, que, ella, la, los, ect).
* Análisis léxico en un compiladores: en este caso, son las expresiones regulares, aquellos tokens que permite el lenguaje para su compilación, es decir, establecen palabras y hasta símbolos que se pueden manejar dentro del lenguaje.
* Códigos de estudiantes: en este caso, el carné para estudiantes se puede ver ya más detallado, porque puede ser una expresión regular que establezaque que primero debe de ir el año, luego iniciales del estudiante y por último el número de cuando fue insctrito. Esto permite un mejor manejo de los carnés.

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

* Las columas son variables: en este caso, buscamos que las columnas esten nombrados con las variables que se requieren en la tabla.
* Las filas deben de ser observaciones: al momento de ver la fila, estemos seguros que estamos teniendo una observación ordenada y clara de la tabla.
* Y la tabla se debe de componer de sólo una unidad observacional: en este caso, que todas las observaciones esten destinadas a un mismo “tipo de observación” es decir, que todas sean congruentes en qué tipo de información se ve por cada observación.

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



En este caso, esta poniendo en el nombre de la columna, valores, que en este caso son los años, pero no esta poniendo una etiqueta año, sino cada año. Por lo que hace falta una columna que contenga una cabal los valores que vemos en las intersecciones entre el país y el año.

El proceso para llegar a la tabla tidy, es: crear una columna donde este se llame: Año, donde sus valores en ella serán los que estaban anteriormente como encabezados de las columnas. Luego, se pondrían una columna que diga Valor: donde contiene los números que intersectan en la tabla anterior.

Country. |Año. |Valor

Guatemala|2008|5

…. Los demás 2008

Guatemala|2009|9

… así sucesivamente.

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



En este caso, la columna equipo esta bien, pero la columna Jugador, que de cierta forma si es buen encabezado, pero lo malo es que maneja dos valores en la misma columna. Por lo que se tendría que hacer es:

Crear una columna: Jugador y otra de Posición.

Para que la columna ya sólo contenga el nombre del jugador y luego la otra columan ya tenga la posición que tiene le jugador.

Ejemplos de tidy:

Equipo | Jugador | Posición

Real Madrid | Federico Valverde | Mediocentro

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



En este caso, contamos con muchos casos de mal manejo de tidy: primero la columna de producto cuenta con dos valores, por lo que se deben separar y crear una nueva columan. Luego en el caso de urbano y rural, son columnas nombradas con valores, se puede crear una columna que diga tipo o región o área por ejemplos, y sus valores ya sea si es rural o si es urbano. Por último, las siguientes columnas, son rangos de valores, y a pear de que en la intersección sólo contenga una x, es mejor reducirla a una columna que diga: Rango de precios, y que en ella se especifique en que rango cumple.

Ejemplo tabla tidy:

Producto | Unidad/Medida | Área | Rango de precio

Banano | 12 unidades. | Urbano | Q0-Q50

Televisión Samsung | 32 pulgadas. | Urbano | Q0-Q50

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

Las funciones period de cierta forma ignoran las irregularidades, por lo que, si saltamos 1 mes, no es lo mismo a 1 mes con las duraiton, porque la period saltan de fecha a fecha, mientras que el duration, se mueve por medio de los segundos, por lo que les permite ser constantes, entonces, al tomar en cuenta las irregularidades y moverse en segundos, si se mueven 1 mes con duration, no caes exactamenta a la misma fecha del siguiente mes, porque puede que hayan menos días en el mes que estás o más días por ejemplos, entonces, uno debe de considerar cúando usar period o durations.

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

Un ejemplo de las funciones period, es uno visto en clase, si queremos calcular el próximo eclipse, o algo que sabemos que cuenta con exactitud la unidad de tiempo en días, horas, minutos y segundos por ejemplo. Porque quiere calcularlo teniendo exactamente la fecha de inicios y como cuentas con valores exactos de dias, horas, etc, pueden hacer la operación sin problema.

Para las funciones durations, es una forma de hacer cronómetros, por lo que puede usarse para saber exáctamente cuanto falta de fecha a fecha, manejar algunas fechas y ver la diferencia entre ellas, para saber si esos procesos se tardaron el tiempo estimado por ejemplo.

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

Este concepto nos dice cuando la probabilidad de que los datos falten es la misma para todos los casos, decimos que es un MCAR, es decir, que, si los datos faltan, no tiene relación con los datos mismos, por lo que permite que se ignore el hecho de que falten los datos.

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

De cierta forma podemos hacer un fill, ya que nos implica que estos datos que hayan faltado, tengan que tener alguna relación importante para que tengan valor, por lo que agregar el valor que tiene el de arriba, no tiene mayor importancia hacer mayor enfuerzo, porque son datos que no tienen relación con los demás datos.

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.

Pairwise deletion: porque 150 individuos no lo considero como data suficiente para poder eliminarlos y trabajar solo con aquellos que están completamente llenos. De igual manera esto nos permite utilizar la mayoría de la data posible, ya que, si de los dos valores que voy a analizar, si están llenos, pero porque en otra columna esta vacío y lo eliminamos, estamos perdiendo data que si contábamos y es relevante.

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.

Outliers cap via percentile approach: debido a que, aquellos que sobrepacen nuestro percentil del 90%, los traemos hacia nuestro valor de 90%, porque al final, es lo que exáctamente se requiere, si tuvimos datos por encima de ello, fue ganancia y no es un problema si nos pasamos, entonces, nos mantenemos con la data que llega al 90% para analizarla.

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

Por ejemplos si estamos manejando precios, o algún tipo de data que contenga rangos muy grandes entre el valor mínimo y el valor máximo, y que por lo tanto, el modelo no pueda entender si la diferencia entre 1 y 1000 es significativa como 1 – 200000 por ejemplo, entonces, el hacer el min max scaling, trae todo a un rango de 0 - 1 y así entender si hay una buena diferencia entre esos datos.

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)

Se haría una normalización por medio de una transformación logarímica para poder hacercarla lo más posible a una distribusión normal.

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)

En el caso de un modelo de econometría, se debe de hacer una columna por cada una, porque, se requiere para que la regresión lineal pueda hacer su proceso de manera correcta. Ahora en el caso de machine learning, probablemente sólo requiere que se agregue una columna que identifique con número los valores de esa variable categórica.

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

El one hot encoding lo utlizamos cuando hacemos modelos de predicción que nos mermitan hacer un fit al modelo, como por ejemplo una regresión lineal, ya que esto le permite al modelo poder comprenderlo de la mejor manera.

1. ¿Qué es un n-gram? (5 pts)

*“La 'n' en n-gramas generalmente se considera un número. Entonces, por ejemplo, una búsqueda de 2 gramos encontrará todas las secuencias contiguas de 2 elementos en un corpus”* Es decir son diagramas que establecen lass combinaciones que desean buscar y te ayudan a obtenerlas de una manera más controlada. Un ejemplo puede ser en los teclado predictivos, ya que utilizan de cierta forma las posibles combinaciones que puede llevar una palabra y puede llegar a sugerirlas, pareciendo que la máquina entienda lo que uno está escribiendo.

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)