Universidad Francisco Marroquín

Data Wrangling 2020

Catedrático: Juan Carlos Girón

Auxiliar: José Josue

Raúl Castellanos 20180052

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)

Es un conjunto de caracteres o de signos especiales que nos permiten el análisis de texto por medio de algún patrón de búsqueda de caracteres.

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

* Por medio de un patrón por ejemplo Sr. O Sra. Podemos definir el sexo de la persona en alguna encuesta.
* Podremos contar las palabras más repetidas y así tener un resumen breve del rating de algún producto o servicio.
* Convertir observaciones en variables dummy, por ejemplo si en la observación se detecta “Perro” será 1, si es “Gato” 0.
* Verificación de que se escribieron bien algunas palabras, por ejemplo, buscar cocer o coser, lo que nos puede ayudar para corregir el texto.

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

* Cada variable debe tener su propia columna.
* Las observaciones deben tener su propia fila.
* En cada celda debe aparecer el valor.

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



No esta en formato tidy, ya que las columnas que contienen los años pueden ser una sola columna demominada “año” y sus datos pueden ser otra columna con la variable “observación”.

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



No está en formato tidy porque la variable de jugador tiene dos valores, y el formato tidy permite que cada columna únicamente puede tomar un valor. Para convertirla a tidy se deberá separar esa columna entre “Jugador”con sus nombres y una nueva variable de “Posición” con la posición de cada jugador.

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 esta en formato tidy ya que sus columnas pueden ser los valores que toma la observación de los productos. Se puede convertir a tidy por medio de crear una nueva columna con la variable “Área” que es la que tomaría el valor de rural o urbano y otra variable de “Precio” la cual contendrá el rango de precio.

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

La diferencia en estas funciones es que period no considera como debe ser la duración completa de un año, mes, día o minuto, sino que period tiene un cálculo predeterminado por ejemplo un año con period comienza el 1 de enero del 2010 y termina el 1 de enero del 2011, en cambio duration considera los tiempos como deben ser entonces si el año comienza el 1 de enero del 2010 y termina el 31 de diciembre del 2010, lo que cambia son sus formatos predeterminados del cálculo entre el tiempo.

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

Period lo utilizaría para fijar una fecha calendario para cada mes de algún evento o pago salarial. Duration lo utilizaría cuando quiera calcular o analizar los tiempos de duración de algun evento o actividad.

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

Los datos faltantes en el dataset fueron determinados completamente aleatorios, sin alguna relación en que falten y otras variables.

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

Dependiendo de la variable podríamos rellenar estos datos faltantes por medio de imputación por media o moda, ó podríamos hacer un relleno por medio de una regresión lineal tomando en cuenta alguna relación entre dos variables.

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.
2. 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)
3. listwise deletion.
4. pairwise deletion.
5. outliers cap via standard deviation.
6. outliers cap via percentile approach.
7. min-max scaling.
8. ¿En qué contexto de Machine Learning se recomienda utilizar Min Max Scaling? (6 pts)

Cuando la variable que queremos analizar tiene valores con mucha dispersión entre ellos, ya que con Min Max Scaling se pueden normalizar para analizar los datos dentro de un rango entre ellos.

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)

Podría normalizarla por medio de la transformación logarítmica, porque la distribución exponencial tiene sólo datos positivos entonces por medio de la transformación logarítmica podría logar que los datos se acerquen a una distribución 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)

Se necesitarían 3, por ejemplo, si en una columna tenemos valores como Alto, Mediano, Bajo, necesitaríamos 3 nuevas columnas que tomarán el valor 0 o 1 cuando sea correcto su valor.

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

Cuando tenemos variables categóricas y sus valores están en formato texto las cuales se pueden transformar en binarias 0 y 1 para poder analizar su distribución o realizar una regresión lineal.

1. ¿Qué es un n-gram? (5 pts)

Es una interpretación de las secuencias en una oración. Por ejemplo, si tenemos la oración “Esta es la clase de data wrangling”. El output de un Ngram de 1 y 2 sería:

* N = 1
  + Esta
  + Es
  + La
  + Clase
  + De
  + Data
  + Wrangling
* N = 2
  + Esta es
  + Es la
  + La clase
  + Clase de
  + De data
  + Data wrangling

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\_\_\_*

Left Join toma solo lo que está en A, y donde la Key de B sea null, es donde B ya no tiene valores donde A sí los tiene.