Universidad Francisco Marroquín

Data Wrangling 2020

Catedrático: Juan Carlos Girón

Auxiliar: José Josue

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)

* Una expresión regular o Regex es una secuencia de caracteres que conforman un patrón, mas que nada se usa para limpiar, extraer o detectar información. Con el objetivo de extrapolar datos e insigths.

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

* Detectar un tipo de archivo, para no permitir formato incorrecto en una encuesta o solicitud de información.
* Validación de estructuración de contraseñas, a la hora de desarrollar algún sitio web con usuarios o bases de datos.
* Limpieza de una Data Frame lleno de caracteres extraños o corruptos.
* Detección de caracteres claves para el análisis de un libro, ensayo, texto o trabajo.

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

* Las variables se categorizan y juntan en una sola forma, y no se extienden en múltiples columnas.

1. Cada observación conforma una línea

* Cada observación de la data; persona, evento u objeto se clasifica en una línea que responde a las variables en las columnas.

1. Cada tipo de unidad de observación conforma una tabla.

* En una tabla no se mezclan observaciones de diferentes estudios o eventos.

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 anteriormente presentada no se encuentra en formato tidy, ya que (“2008”,”2009”,”2010”) conforman 3 columnas. Pero solo representan una variable. Para pasarlo a formato tidy , tendría que hacer un melt o pivot\_longer de esas 3 columnas y transformando la tabla a solo 3 columnas. (“Country”,” Year”,” Value”)

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 anteriormente presentada no se encuentra en formato tidy, ya que dentro de la columna Jugador se encuentra 2 diferentes variables. Para pasarlo a formato tidy, tendría que hacer un Split de la columna “Jugador” usando como separador el símbolo “-“ transformando la tabla a 3 columnas. (“Equipo”,” Jugador”,” Posicion”)

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 anteriormente presentada no se encuentra en formato tidy, ya que cuenta con 3 variables, pero 2 es estas están mal representadas en 7 columnas. Una variable está separada en las columnas 2 y 3. La otra se encuentra separada en las columnas 4, 5, 6 y 7. Yo haría una copia del Data Frame. En la primera copia solo usaría las columnas 1:3. Para pasarlo a formato long usando de ID la variable del producto. Quedaría una tabla de dos columnas (“Producto”, “Locación”). En la segunda copia usaría las siguientes columnas (1,c(4:7)) en otras palabras las que están relaciones don la variable de dinero. En esta haría el mismo proceso con un pivot\_longer para quedarme con una tabla de dos variables. (“Producto”,”Rango de Precio”). Para finalizar haría un merge de las dos copias usando la variable “Producto como ID”.

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

* La función Period mide o se refiere en el cambio de hora que ocurre entre dos instantes
* La función duration mide o se refiere la cantidad exacta de tiempo que ocurre entre dos instantes

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

* Si quiero medir algo como una carrera o un evento que no depende de los cambios en el calendario o tiempo uso duration.
* Si quiero medir algo que puede ser afectado por años bisiestos cambios de horarios u otras aberraciones al tiempo uso period.

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

* MCAR o Missing Completely at Random significa que no existe relación entre la falta de datos y cualquier valor, observado o perdido. Los puntos de datos que faltan son un subconjunto aleatorio de los datos.

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

* Si tengo un Data Set o alguna tabla donde logro verificar que mis missing values son MCAR´s , la imputación que haría o recomendaría dependería de otros factores. Sin más información podría ser viable una imputación con la media o moda incluso una regresión lineal.

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.

* Si estoy realizando un análisis de 150 observaciones en las cuales tengo datos faltantes en diferentes variables. Mi método a utilizar dependería de lo que yo considere como porcentaje relevante de información para poder trabajar, pero sin ir más a detalle usaría un pairwise deletion. Porque no quiero perder todas mis observaciones quiero usar toda la data que recolecte.

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.

* El método que utilizaría para determinar con que población de mis datos trabajar seria, outliers cap via percentile approach. Ya que este método me permitiría trabajar con el 90% de la demanda y posiblemente identificar si vale la pena en ciertas estacionalidades o épocas aumentar mis operaciones.

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

* El contexto para ulilizar Min Max Scaling en un programa o desarrollo de machine learning es cuando estoy trabajando con variables que iteran en diferentes rangos. Por lo cual no se pueden comparar directamente, se tiene que representar los valores de tal manera que se puede trabajar.

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)

* Si mis datos tienes a un comportamiento exponencial usaría transformaciones logarítmicas para normalizar los mas posible de esos datos.

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)

* Necesitaría 2 variables dummys.

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

* El contexto donde usamos one hot encondig es cuando tenemos una variable categórica que no contiene una relación ordinaria. Sirve para transformar la variable en vectores binarios.

1. ¿Qué es un n-gram? (5 pts)

* Un N-gram es un diagrama que permite visualizar como se relaciones diferentes palabras, en otras palabras, para analizar las secuencias de palabras que conforman oraciones.

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*