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 una secuencia de caracteres que define un patrón de búsqueda la cual se utiliza principalmente para la búsqueda y manipulación de texto basada en patrones.

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

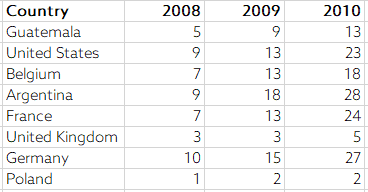
* Los corchetes. Se utilizan para definir un conjunto de caracteres posibles en un punto específico de una cadena. Por ejemplo [a-z] coincidirá con cualquier letra minúscula.
* La barra. Se utiliza para denotar una alternativa o elección. En otras palabras, funciona como un operador or. Por ejemplo gato|perro coincidirá con cualquier cadena que contenga la palabra "gato" o la palabra "perro".
* El circunflejo. Se utiliza para definir el inicio de una cadena. Por ejemplo ^abc coincidirá con cualquier cadena que comience con "abc".
* El signo de dólar. Se utiliza para definir el final de una cadena. Por ejemplo abc$ coincidirá con cualquier cadena que termine con "abc".

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

* Cada variable forma su propia columna.
* Cada observación forma su propia línea.
* Cada valor forma su propia celda.

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 porque las columnas 2008, 2009 y 2010 representan valores de la misma variable (años). Para solucionarlo se debería tener una columna para el año y otra columna para los valores.



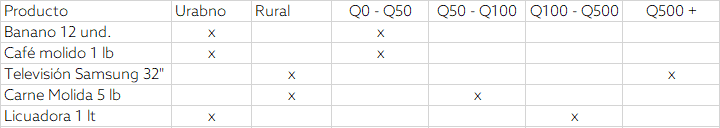
1. Diagnostique y explique por qué la siguiente tabla no está en formato ***tidy.*** Luego, explique cómo convertirla a formato ***tidy*.** (7 pts)

La tabla no se encuentra en formato tidy porque la columna de Jugador también contiene información sobre la posición del jugador. Para solucionarlo se debería agregar una tercera columna que especifique la posición del 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 se encuentra en formato tidy porque las columnas de urbano y rural representan valores de una misma variable (área), al igual que las columnas que indican el rango de precios deberían representan una sola variable (precio). Para solucionarlo se deberían cambiar las columnas de urbano y rural por una que especifique el área y cambiar las columnas de los rangos de precios por una sola que se llame precio y que contenga el valor específico del precio.



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

Duration se utiliza para representar medidas precisas e invariables de tiempo, mientras que period se utiliza para representar medidas relativas que pueden adaptarse a la estructura del calendario, como meses y años de diferentes longitudes.

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

Utilizaría la función duration para averiguar una cantidad exacta de tiempo en términos de días, horas, minutos, segundos, sin preocuparme por las variaciones en la longitud de los meses o años. Por otro lado, usaría duration si estuviese trabajando con fechas de calendario para representar intervalos de tiempo en términos de años, meses, días, etc.

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

Si la probabilidad de que falten datos es la misma para todos los casos, entonces se dice que los datos están completamente ausentes al azar.

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

Listwise deletion, ya que los datos faltantes son aleatorios y no están relacionados con ninguna característica específica, eliminar las observaciones con datos faltantes no introduce ningún sesgo en los resultados.

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.

Recomendaría utilizar listwise deletion, pues la data no se encuentra completamente disponible. Sin embargo, si los valores faltantes son una cantidad considerable para el análisis y quiero tomar toda la data posible en cuenta recomendaría usar pairwise deletion.

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.

Recomendaría usar Outliers cap via percentile approach, pues los outliers pueden distorsionar la distribución de los datos y, por lo tanto, pueden conducir a una estimación incorrecta de la capacidad necesaria.

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

Cuando las variables tienen rangos de valores muy diferentes

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)

Logs transformations.

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 dummy (k-1).

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

Para modelos de aprendizaje automático, bases de datos relacionales o la visualización de datos

1. ¿Qué es un n-gram? (5 pts)

Es una combinación de "n" elementos adyacentes de una secuencia.

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 \_\_\_ JOIN B ON A.KEY = B.KEY \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

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)  
   La herramienta podría basarse en la comparación de similitud con la base de datos que posee, para supervisar que nivel de similitud contiene. También es posible que realicen un búsqueda de patrones por medio de expresiones regulares en la forma en que escriben los estudiantes para posteriormente buscar patrones similares en la forma en que escriben los demás estudiantes
2. 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)?