Universidad Francisco Marroquín

Data Wrangling

Catedrático: Juan Carlos Girón

Auxiliar: José Josue

Examen Final Data Wrangling

**Alumna:** Luz Noemi Arévalo Milian – 20200392

**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)

Las expresiones regulares son secuencias de caracteres alfanuméricos o caracteres especiales (como corchetes, punto, barra inversa, etc.) que se utilizan para la búsqueda de patrones concretos en cadenas de texto.

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

-Reemplazar palabras de una cadena.

-En buscadores de texto.

-Validación de contraseñas.

-En análisis de textos definir la similitud entre dos documentos.

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

* **Cada variable es una columna**: significa que cada columna contiene todos los valores que miden un atributo de una unidad determinada.
* **Cada observación forma una fila**: contiene todos los valores medidos de una unidad a través de las distintas variables/atributos.
* **Cada tabla se compone de una unidad observacional**: significa que una tabla contiene observaciones con sus respectivos atributos sobre una misma unidad, aquella que se busca comprender o almacenar.

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



Las columnas 2008-2010 no son variables, sino valores observacionales. La forma de convertir esta tabla al formato *tidy* es transformarla en tres columnas: *Contry*, *Year*, *Value*, de forma que la tabla quedará en 3 variables con 24 observaciones (filas).

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



En la variable **jugador** hay más de un valor observacional. Para que quede en el formato estándar *tidy* es agregar una tercera columna para la variable **Posición\_en\_campo** en la que se agregarán los valores mediocentro, delantero, etc. De forma que la tabla quedará con las siguientes variables: Equipo, Jugador y Posición\_en\_campo.

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



Esta tabla contiene valores observacionales en las columnas. De acuerdo con el objetivo de uso la tabla podría transformarse en tidy de distintas maneras:

**Para facilitar la agrupación y graficas**: Organizar la tabla en 4 columnas, *Producto*, *Area (categórica)*, *precio\_inferior (numeric)*, *precio\_superior (numeric)*.

**Para la modelación de los datos**: como los precios son rangos, aquí podrían aplicarse *bins* para transformar todas las columnas precio en una sola, de forma que se agregue una columna que ayude a entender la data (por ejemplo de 0-50 es *low*, de 50-100 *medium*, ect.) Los valores de columnas urbano y rural pueden ser binarios, de forma que se obtengan variables *dummy* que permitan hacer regresiones. Como resultado la tabla queda con las siguientes variables: Producto, Urbano (*dummy*), Rural (*dummy*), Rango\_precio (categórica) y Bin\_precio (categórica)

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

En las clases temporales de lubridate encontramos las *period* y *duration*. La diferencia es la interpretación que hace la máquina. Una función ***duration*** es más precisa en cuanto al tiempo entre dos fechas, funciona como un cronometro, mientras que una ***period*** tiene una concepción más humana sobre el tiempo.

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

Si hacemos la suma “today() + years(**2**)” el resultado será "2024-11-16" 4, nos devuelve lo que se piensa en la mente humana, porque se está utilizando una función *period*, utilizarla cuando quieren tener fechas de la realidad. Mientras que si usamos ““today() + dyears(**2**)” obtendremos "2024-11-15 12:00:00 UTC”, porque toma en cuenta que el 2024 es un año bisiesto con un día adicional, utilizar *duration* cuando se quieren hacer operaciones exactas con fechas.

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

Cuando los valores faltantes de las observaciones se encuentran de esa manera por azar, es decir, que no dependen de la ausencia de otros valores o que no se ha hecho con premeditación.

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

Al determinar que la data faltante es MCAR, entonces recomendaría la imputación por la media (de tratarse de una variable numérica) o por la moda (de ser una categórica).

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.

Porque cada observación recogida en una encuesta cuesta y es valiosa, por lo que hacer suposiciones podría distorsionar la información y eliminarla sería un desperdicio de recursos y datos importantes, por lo tanto, lo recomendable es aprovechar los pares de variables que pueden ser relevantes.

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

Cuando deseamos optimizar (mejorar el rendimiento) el modelo y reducir la cantidad de iteraciones necesarias. Además, esto permite hacer comparaciones escaladas entre 0 y 1 lo que permite determinar si un cambio entre valores es realmente significativo.

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)

Utilizaría una transformación logarítmica (si se trata solo de valores positivos), para captar mejor las diferencias en variables que no tienen un comportamiento lineal.

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)

Una variable *dummy* por cada nivel de la variable categórica. En este caso 3 variables son requeridas.

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

Cuando tenemos variables categóricas y deseamos transformarlas a un lenguaje comprensible para las computadoras con las que ya se puedan realizar predicciones. Hacer esto también puede permitir que la máquina ordene naturalmente las distintas categorías si es que existe un orden entre ellas.

1. ¿Qué es un n-gram? (5 pts)

Herramienta que encuentra patrones de **n** palabras o secuencias continuas dentro de un corpus. Es útil para hacer predicciones de escritura de texto.

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 OUTER 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)

A través del factor TF-IDF (*Term frequency-inverse document text*) esta herramienta hace búsquedas en los documentos de la base de datos que coincidan con el documento objetivo. Con las palabras más frecuentes del documento objetivo, hace comparaciones con las palabras de los documentos de la base de datos. A medida que este factor se acerque a 0, entonces significa que el documento evaluado tiene alta similitud con alguno o varios documentos de la base de datos.