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)**

* Es una secuencia de caracteres que unidos forman un patrón de búsqueda con alguna estructura definida. Es una manera flexible de buscar y reconocer cadenas de texto.

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

* Validación de placas de un vehículo particular de Guatemala, la expresión regular tendría que ser capaz de identificar que la frase empiece con una letra P, seguida por 3 números entre el 0 y el 9, para finalizar con 3 letras entre la A-Z ignorando todas las vocales.
* Validación de contraseñas, tendría que ser capaz de asegurarse que la frase para contraseña cuente con por lo menos 8 caracteres, al menos una mayúscula y algún símbolo especial.
* Validación de archivos con terminación .pdf, tendría que ser capaz de identificar que dentro de la cadena de texto se encuentre un . y lo que sigue de ese punto sean las letras pdf. Pueden ser mayúsculas o minúsculas.
* Validación de carné UFM, tendría que ser capaz de identificar los primeros 4 números como dentro de los años 2001-2030, seguido por un 0, con terminación de una numeración entre 100-999. El carné debe tener una cantidad exacta de 8 números.

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

* Cada variable es una columna
* Cada observación forma una fila
* Cada tabla se compone de una unidad observacional

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



* Los años se están usando como variables mientras que son valores.
* Utilizaría la variable de año como columna y dentro de la misma se encontrarían 2008,2009,2010. Luego le asignaría una etiqueta a los valores númericos para que la tabla esté conformada por 3 columnas y N observaciones.

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



* Hay multiples variables dentro de una columna
* Agregaría una columna de ‘posición’ en la que tendría la segunda variable que actualmente se encuentra en la columan de jugador. Se eliminaría la segunda variable de la columa jugador para que quede unicamente una tabla de 3 columnas con N observaciones. Cada columna con una única variable por observación.

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



* Cuenta con valores como nombre de columna, tiene multiples variables dentro de una misma columna.
* Cambiaría los nombres de columnas a Producto, Cantidad,Área y Precio. Quitaría las multiples variables de la actual columna producto, pasando la cantidad o peso especificado a la nueva columna de cantidad. Pasaría las dos columnas de valores urbano y rural a ser una sola columna de Área y ya utilizando Urbano y Rural como sus observaciones, Precio sería la columna para los valores que se encuentran actualmente como nombres de columnas sobre el rango de precio.

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

* Duration calcula la cantidad exacta de tiempo entre dos momentos, también se le puede asociar con un cronómetros, mientras que period mide de forma precisa los tiempos del reloj. Period no considera los años biciestos, se le asocia más con la interpretación humana del tiempo.

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

* Usaría la función period para saber cuanto tiempo falta para una fecha y hora en específico y la función duration lo usaría para calcular la cantidad de tiempo que tarda una persona en terminar una tarea.

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

* Es cuando los datos faltantes son completamente aleatorios, por lo que la probabilidad de que un dato sea faltante es la misma para todos los demás datos. Los datos faltantes son completamente independientes y no están relacionados con ningún otro valor en el conjunto de datos.

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

* La imputación de la media ya que los datos faltantes son aleatorios, entonces el valor medio es el más probable.

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.**

* B, ya que hay poca información en general, utilizando pairwise deletion no se eliminan muchos datos a comparación de la información que se cuenta.

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.**

* D, porque se tomaría más en cuenta el porcentaje de cumplimiento que se quiere lograr.

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

* Cuando se necesita normalizar los datos para que estén dentro de un rango específico. Esto hace que mejore el rendimiento de los modelos al aumentar convergencia en el entrenamiento de los 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)**

* estandarización

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

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

* Se utiliza cuando se necesita representar datos categoricos de una manera que sea compatible con los algoritmos. Ayuda a transformar los datos categoricos en una matriz de datos en donde cada columna representa una categoría y cada fila un ejemplo.

1. **¿Qué es un n-gram? (5 pts)**

* En un conjunto de n caracteres consecutivos, n es usualmente considerado un número. Según el número n es la cantidad de caracteres consecutivos encadenados en un corpus.

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 \_\_\_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)**