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 expresion regular es una secuencia de caracteres que tiene un patron en particular. Las expresiones regulares estan formadas de caracteres ordinarios y especiales esto con el fin de poder hacer patrones mas complejos.**

1. Enumere y explique brevemente cuatro aplicaciones prácticas en las cuales las expresiones regulares son utilizadas. (5 pts)
   1. **Validacion de datos: poder verificar que lo ingresado por el usuario coincide con lo solicitado.**
   2. **Busqueda de texto: de acuerdo a un patron puede encontrar texto que se repita dentro de un texto.**
   3. **Transformar el texto: realizar cambios en una palabra especifica o limpieza de datos.**
   4. **Copiladores: se usa en copiladores de codigo para verificar que se tenga el sintax correcto**
2. Explique brevemente las 3 condiciones que establecen que una tabla se encuentra en formato ***tidy.*** (5 pts)
3. **Que cada una de las variables del conjunto de datos sean columnas**
4. **Cada unidad observacional sea una tabla**
5. **Cada observacion sea una fila**
6. Diagnostique y explique por qué la siguiente tabla no está en formato ***tidy.*** Luego, explique cómo convertirla a formato ***tidy*.** (7 pts)



**En este caso no esta en formato tidy ya que la variable de año no es una columna y cada año no es una variable. Por lo que para volverla tidy se tendria que crear dos columnas nuevas para poder poner año y el valor que este tenga por ejemplo:**

|  |  |  |
| --- | --- | --- |
| **Country** | **Año** | **Value** |
| **Guatemala** | **2008** | **5** |

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 no se encuentra en formato tidy, ya que la variable (posicion) no esta como una columna sino que esta mezclada con la columna jugador, para realizarla tidy añadiria la columna posicion y separaria utilizando el – la posicion del nombre.**

**Por ejemplo:**

|  |  |  |
| --- | --- | --- |
| Equipo | Jugador | Posicion |
| Real Madrid | Federico Valverde | Mediocentro |

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



**Esta nuevamente no esta en formato tidy ya que el Urbano, rural y el rango de precios no son variables, por lo que no deberian de estar como columnas. Por lo que para volverlo tidy se tendria que 2 columnas nuevas para poder organizar la tabla, por ejemplo:**

|  |  |  |
| --- | --- | --- |
| Producto | Sector | Rango |
| Banano 12 und | Urbano | Q0-Q50 |

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

**La funcion period hace referencia una cantidad de tiempo definida en términos humanos, como días, meses o años. Mientras que la funcion duration hace referencia a unidades de tiempo más pequeñas y son ideales para operaciones matemáticas precisas, como calcular la diferencia exacta entre dos momentos en el tiempo.**

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

**Para tabular los datos de una encuesta seria util la funcion period ya que muchas veces en las encuentas las personas dan el tiempo en periodos como 6 meses o 1 año 3 meses. Mientras que la funcion duration se podria hacer para calcular el tiempo que una persona trabajo en la empresa, ya que se toma el dia de su contratacion y usando duration al dia actual podemos saber con precisión cuanto tiempo ya sea en termino de dias, semanas o años.**

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

**Este concepto describe una situación en la que la ausencia de datos en un conjunto de datos ocurre de manera completamente aleatoria, sin seguir ningún patrón discernible ni estar relacionada con ninguna variable medida o no medida en el conjunto de datos.**

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

**La imputacion recomendada en este caso seria la imputacion simple por la media, ya que es muy sencilla de implementar y al no tener un patron para predecir los datos faltantes esta resolveria el problema.**

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.

El tener la Perdida de tantos datos, es mejor utilizar los datos existentes para cada análisis. Ademas de que al no saber si hay un patron de perdida de datos ya que fueron en varias variables este funciona mejor.

* 1. outliers cap via standard deviation.
  2. outliers cap via percentile approach.

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.

Ya que utilizando este metodo se puede llegar a limitar los valores extornos para que aquellos valores que son atípicos no afecten en el objetivo del 90%

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

**El contexto ideal es cuando se puede conocer cual es el minimo y maximo que tenemos. Por ejemplo, para mejorar la calidad de una imagen ya que se puede conocer que pixeles tienen un max y un min.**

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)

**Utilizaria la transformacion logaritmica ya que esta tecnica permite reducir la variabilidad en los extremos de la distribución exponencial, logrando una forma más simétrica de 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)

**Al tener 3 variables categoricas necesitamos n-1 variables dummies por lo que serian 2. Cada variable dummy representa la presencia o ausencia de un nivel específico de la variable categórica.**

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

**Se utiliza one hot encoding para convertir variables categóricas en un formato numérico binario, representando cada categoría como una columna binaria.**

1. ¿Qué es un n-gram? (5 pts)

**Un n-gram es una secuencia de caracteres de una expresion regular que ayuda en que la secuencia de palabras es pertinente, tales como análisis de sentimiento, clasificación de texto y generación 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 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)

**Este tipo de herramienta es una expresión regular que compara los documentos o el texto que está dentro de estos para poder buscar si ese patrón ya existe, de ser así envía un resultado que valida que el patrón existe y con que ocurrencia o con que documento en específico, lo que significaría que si existe plagio.**

1. 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)?