Universidad Francisco Marroquín

Data Wrangling

Catedrático: Juan Carlos Girón

Auxiliar: José Josue

Examen Final Data Wrangling

Javier Alejandro Mazariegos Godoy 20200223

**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 se utiliza para encontrar patrones en una cadena de caracteres. Por lo general las expresiones regulares se utilizan para detectar patrones de texto, es decir, buscar palabras o texto en un parrafo o sustituir fragmentos de texto.**

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

* **Validaciones de Email:**

Por lo general en los formularios de internet hay que ingresar un correo electrónico y se podría utilizar expresiones regulares para validar que el correo esté bien escrito.

* **Identificar o validar un número de teléfono:**

Se podría utilizar para validar que un número de teléfono este bien escrito dependiendo del país de donde sea.

* **Identificar palabras de un texto**

Se puede utilizar expresiones regulares cuando se quieren buscar ciertas palabras en un texto o tambien se podría utilizar para buscar ciertas palabras en mensajes de whatsapp.

* **Validación de campos en un formulario de Google Forms.**

Google Forms tiene un apartado para poder validar los campos con una expresión regular: se podría validar la fecha de nacimiento o respuestas directas que tienen que tener un formato en específico.

1. Explique brevemente las 3 condiciones que establecen que una tabla se encuentra en formato ***tidy.*** (5 pts)
2. **Cada variable tiene que ser una columna.**
3. **Cada observación debe formar una fila**
4. **Cada tabla debe de componerse de una unidad observacional.**
5. 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 se encuentra en formato tidy debido a que los nombres de las columnas son valores y no variables, y tambien que hay variables en columnas y filas.

Lo que habría que hacer para pasarlo a tidy es:

* + Pasar las columnas a filas
  + Crear una nueva columna con la frecuencia.
* Resultado

|  |  |  |
| --- | --- | --- |
| Country | Año | frecuencia |
| Guatemala | 2008 | 5 |
| Guatemala | 2009 | 9 |
| Guatemala | 2010 | 13 |
|  |  |  |
|  |  |  |

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



No está en formato tidy porque hay múltiples variables en una sola columna, en este caso el jugador y la posición en la que juega se encuentra en la misma columna de jugador.

Lo que habría que hacer aquí para convertirlo a tidy es:

* + Crear otra columna nueva para separar a jugador y posición.
  + Una columna llamda Jugador y otra columna llamada posición
* Resultado

|  |  |  |
| --- | --- | --- |
| Equipo | Jugador | Posicion |
| Real Madrid | Federico Valverde | Mediocentro |
| Juventus | Cristiano Ronaldo | Delantero |

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



No está en formato tidy debido a que los nombres de las columnas son valores y no variables. Para poder convertirlo a tidy lo que se debería de hacer es:

* + Convertir de columnas a filas.
  + crear tres columnas: Producto, población, precio y frecuencia.
* Resultado

|  |  |  |
| --- | --- | --- |
| Producto | Poblacion | Precio |
| Banano | Urbano | Q0-Q50 |
| Café molido 1lb | Urbano | Q0-Q50 |
|  |  |  |

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

La función de period da medidas de tiempo exactas, mientras que la función duration da medidas de tiempo que son relativas.

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

La función duration se utilizaría cuando se quiere calcular lapsos de tiempo, mientras que la función duration se utilizaría cuando se quieren calcular unidades humanas, es decir, cuando se quiere calcular timpos que son fácil de entender para nosotros:

Ejemplo: Si se quiere sumar 1 dia a”2016-03-13 01:00:00” con duration la respuesta seria ”2016-03-13 02:00:00” mientras que con period seria ”2016-03-13 02:00:00” debido a que nosotros entendemos que sumar un día nos estaría dando la misma hora que la fecha a la que le estamos sumando.

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

La probabilidad de que haya un missing value en una variable no depende de otras variables sino que la probabilidad es totalmente aleatoria.

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

Si la data el MCAR se recomendaría utilizar linear regression si la variable es numérica o logistic regresión si es categórica, debido a que no depende de otras variables entonces no existiría ningún sesgo.

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.

**Utilizaria pairwise deletion porque seria el metodo con el qué menos data perderia, debido a que unicamente tengo 150 observaciones que son relativamente pocas para realizar un analisis.**

**Debido a que pairwise deletion no eleimina todas las filas que tengan al menos un missing data sino que unicamente elimina las parejas de la regression que tienen un missing data.**

* 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.

Trabajaria con outliers cap via percentile approach porque puedo definer mi percentil, es decir, puedo definer cual es mi intervalo de confianza y si no quiero trabajar con los datos atipicos podria eliminar las colas para trabajar unicamente con el 90% de la demanda mensual.

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

Min Max se utiliza cuando hay diferencias muy grande dentro de mi data, es decir que tengo una variable que tiene un mínimo de 0 y máximos de 300,000. Entonces se utiliza Min Max para cambiar la escala y que mi intervalo quede entre 0 y 1. Entonces ya no hay que moverse tanto para encontrar el promedio, teniendo una escala de 0 y 1, el promedio debe estar dentro de esa escala.

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 log tranformation, cundo mi función no es lineal sino que se parece más a una exponencial es mejor utilizar log transformation para transformar los datos a una distribución normal. Si se utiliza una lineal el problema es que no se pasa casi por ningún punto, entonces se busca una exponencial para cubrir la mayoría de los puntos.

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)

Se necesitan tres variables dummy, cuando una variable es categórica es necesario utilizar One Hot Encoding para pasar variables dummy. Se utilizaría una variable dummy por cada nivel.

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

Se utiliza One Hot Encoding cuando no es importante el orden, entonces todas las variables se ponen al mismo nivel. Lo que nos dice una variable dummy es si la fila en la que aparece el valor es un veradero o un falso.

1. ¿Qué es un n-gram? (5 pts)

N-gram es una aplicación más de los encoders y consiste en calcular la probabilidad de que después de escribir una palabra venga otra que tenga reación. Los N-grams son muy utilizados en la actualidad por Google o Apple para dar sugerencias de que palabra se puede escribir después de haber escrito una.

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)

Primero se pasa por un proceso de análisis léxico, es decir, se tokenizan las palabras validas, en nuestro caso son las palabras en español. Luego se pasa por un proceso de Análisis sintactico, es decir que se examina la gramática, y luego se comparan los archivos y se ve si la gramática de otros archivos que tiene guardados turnitin es igual a la que se esta examinando, de esa forma se puede ver que tan parecido es un trabajo de otro.