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)

Comunmente conocido como regex, es una secuencia de caracteres que define un patrón de aceptación o búsqueda, en el ámbito de computer science se utiliza en la tokenización para un scanner que eventualmente conforma un compilador, en programación o data science es útil para buscar patrones, reemplazar caracteres dentro de una columna, etc.

1. Enumere y explique brevemente cuatro aplicaciones prácticas en las cuales las expresiones regulares son utilizadas. (5 pts)
   1. Búscar y reemplazar texto
   2. Procesamiento de lenguage natural (NLP)
   3. Validación de datos (formatos de correos eléctronicos, direcciones, etc)
   4. Tokenización para un scanner de un compilador
2. Explique brevemente las 3 condiciones que establecen que una tabla se encuentra en formato ***tidy.*** (5 pts)
   1. Cada variable tiene que ser una columna
   2. Cada observación de estas variables debe ser una fila
   3. Debe existir una tabla unidad de observación o clase de variable
3. Diagnostique y explique por qué la siguiente tabla no está en formato ***tidy.*** Luego, explique cómo convertirla a formato ***tidy*.** (7 pts)



Se está utilizando cada año como columnas, debería de existir una columna de año, una de las premisas de tiny define cada variable como una sola columna, esto no lo cumple.

Las observaciones de una misma variable están distribuidas en múltiples columnas, en tidy los valores asociados a una variable deben estar en una sola columna, debería de existir una columna llamada “Cantidad” o lo que sea que representan estos valores.

Entonces, para convertir esta tabla a Tiny deberia de tener el siguiente formato:

|  |  |  |
| --- | --- | --- |
| Country | Year | Units |
| Guatemala | 2008 | 5 |
| Guatemala | 2009 | 9 |
| … | … | … |

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



La columna “Jugador” contiene valores compuestos, es decir, está mezclando nombres y posición, incluso podría decirse que está mezclando primer nombre, segundo nombre y apellido, dado que no todos contienen los tres, al momento de extraer los valores no sabemos si es un apellido o nombre.

Para convertir esta tabla en formato tidy se debe de hacer lo siguiente:

* Separar la columna “Jugador” en 4 columnas
  + Primer nombre
  + Segundo nombre
  + Apellido
  + Posicion
* De esta forma cada fila representaría una observación unica a la variable de la columna.

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 está en formato de tidy porque:

* Se está utilizando variables como columnas
  + Los rangos de precio están distribuidos como columnas
  + Tanto la unidad de medida como el valor están combinados en la columna de “Producto”
  + La ubicación (Urbano o Rural) están distribuidos como columnas
* Se está utilizando una marca (“x”) para indicar la presencia de estas observaciones

Para convertir esta tabla en formato tidy se necesitaría:

* Separar el nombre del producto en una sola columna
* Separar la unidad de medida en una columna
* Separar el valor de la unidad de medida en otra columna
* Crear una columna de Ubicación
* Crear una columna de Rango de Precio
* Eliminar el uso de la “x” y representar la observación de las variables como su valor.

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

Ambas se utilizan para manejar diferencias de tiempo pero tienen distintos propositos y comportamientos:

**Period:** Se utiliza para el lapso de tiempo como nosotros lo vemos es decir, días, meses, años. Se ajusta al contexto en el que se calcula, por ejemplo un período de mes de febrero vs marzo podría variar respecto a cuantos días hay en el mes pero la función se ajusta al calculo según el calendario.

**Duration:** Lapso de tiempo exacto en segundos, independiente del contexto como el ejemplo anterior, esta función nos da el lapso entre dos tiempos exacto.

Entonces, duration es para medidas de lapso de tiempo exactas mientras que period es para medidas de tiempo que varian según el contexto.

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

Se utilizaría period en contextos donde las unidades de tiempo tienen que permanecer constantes, como programar que corra un script el primer dia de cada mes por ejemplo.

Duration cuando se necesita un lapso de tiempo con una precisión fija, para cálculos de matemáticas, física, etc. Por ejemplo la medición de tiempo de un experimento se utiliza duration para tener el lapso exacto en segundos.

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

Es una característica de los datos faltantes cuando la falta de datos de una variable no está relacionada con los valores de esta misma variable o de las otras variables, implica que la falta de datos es completamente aleatoria.

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

Depende de que tipo de variable estemos tratando pero si es una variable continua recomendaría una imputación por la media por ser fácil de implementar.

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.

Sería una elección entre la opción a y b, ya que la c y d son para outliers y no aplicaría en este caso. Dado que es una muestra pequeña de encuestas recomendaría la b (pairwise deletion) porque esta busca maximizar el uso de datos disponibles al eliminar los casos faltantes para cada variable, a diferencia del listwise que elimina el registro completo si hay algún valor faltante.

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.

Aquí se buscaría tratar los outliers y la mas adecuada es la d a. outliers cap via percentile approach ya que permite identificar y eliminar los valores más extremos, es decir, la demanda atípica, esto permitiría elaborar un modelo con base en la demanda más regular y permitiría una planificación más consistente.

1. ¿En qué contexto de Machine Learning se recomienda utilizar Min Max Scaling? (6 pts)
   1. Modelos sensibles a la escala
   2. Optimización de gradientes
   3. Visualización de datos
      1. Al tener todos los valores en un rango establecido permite una mejor legibilidad
   4. Variables con diferentes unidades de medida

No es recomendable cuando los datos tienen outliers muy significativos.

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)

La más utilizada es la transformación logarítmica ya que puede ayudar a estabilizar la varianza y hacer la distribución más simétrica.

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 necesitarían dos variables utilizando one hot encoding que pueda representar los 3 estados posibles, ya que se necesita 1- la cantidad de categorías posibles.

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

Se utiliza cuando tenemos una variable categórica que queremos convertir a un formato númerico para que pueda ser recibido por los modelos.

1. ¿Qué es un n-gram? (5 pts)

Es una secuencia de n elementos de un texto, comunmente utilizado en NLP donde podemos tener algo como: “Me gusta data wrangling” tendriamos un bigrama de “Me gusta”, un trigrama de “Me gusta data”, etc.

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)
2. Convierte los documentos en formatos que pueden ser procesados digitalmente
3. Descompone los textos en elementos como n-gramas
4. Hace referencia a una base de datos de trabajos entregados, artículos, libros, etc.
5. Utilizando algoritmos de coincidencia compara los textos con su base de datos.
6. 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)?