Universidad Francisco Marroquín

Data Wrangling 2020

Catedrático: Juan Carlos Girón

Auxiliar: José Josue

Examen Final Data Wrangling 2020

**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 exámen para los estudiantes involucrados.

**Serie Única: Conteste a las siguientes preguntas**

1. **¿Qué es una expresión regular? (5 pts)**

Una expresión regular es un lenguaje formal, una expresión regular esta conformada por una secuencia de caracteres que forman un patrón de búsqueda. En informática su uso radica en proporcionar una manera muy flexible de buscar y reconocer distintas cadenas de texto.

1. **Enumere y explique brevemente cuatro aplicaciones prácticas en las cuales las expresiones regulares son utilizadas. (5 pts)**
2. Validación de correo electrónico: Mediante la utilización de expresiones regulares podemos verificar si un correo electrónico es válido o no. Con este podemos verificar dominios por ejemplo como @yahoo.com que seria un dominio valido pero @asdfj.lkjh podría no ser un correo electrónico valido.
3. Contraseñas a sitios: Cuando iniciamos sesión particularmente contamos con una contraseña y estas contraseñas pueden cumplir ciertos requisitos que como desarrolladores queramos implementar por ejemplo el uso de una mayúscula como mínimo o algún carácter especial necesario para la validación. Si el usuario no cumple con estos requisitos en su contraseña no podrá registrarla ya que no es válida.
4. Validación de archivos: Mediante el uso de expresiones regulares podemos validar en un sistema el tipo de archivo que se esta registrando a una plataforma, por ejemplo. En aquellas plataformas en las que solo se requiera un .pdf no se aceptara un .xlsx por ejemplo con el objetivo de evitar errores.
5. Elaboración de un compilador: Un compilador es sumamente complejo, mediante expresiones regulares podemos llegar a reconocer ciertas palabras claves que nos permiten identificar palabras reservadas o más y también nos sirven en la creación de los autómatas.
6. **Explique brevemente las 3 condiciones que establecen que una tabla se encuentra en formato *tidy.* (5 pts)**
7. Cada variable es una columna: Cuando tenemos información en un Dataset las diferentes categorías, Atributos o también propiedades deben ir en columnas para distinguir los tipos de datos con los que contamos.
8. Cada observación es una fila: Cuando tenemos información en un Dataset las filas representan observaciones y estas observaciones son registros de cualidades o cantidades de una observación de un fenómeno, incluye cosas que podemos ver, medir, oir.
9. Cada tipo de observación medible forma una tabla:
10. **Diagnostique y explique por qué la siguiente tabla no está en formato *tidy.* Luego, explique cómo convertirla a formato *tidy*. (7 pts)**



Algunos de los errores que podemos observar en la tabla es que las columnas representan valores y no son nombres de variables como lo podría ser Año, por ejemplo. Podríamos separar estos años y agregarle su respectivo conteo como una observación en base al año.

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 podemos ver que en la columna de jugador cuenta con el nombre de la persona seguido de un carácter que define su posición. No podemos registrar múltiples observaciones en la misma tabla. Para volverla en formato tidy tendríamos que separar la columna de Jugador con una de posición. Las variables que tendríamos serían Equipo, Jugador y Posición.

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 tabla el precio del artículo podría estar únicamente en una sola columna. Para volverla en formato tidy las columnas tendrían que volverse observaciones de una variable llamada precio por ejemplo. Las observaciones cuentan con una X únicamente cuando debería de estar de forma binaria como mínimo.

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

La función periodo no tiene una duración exacta hasta que no se defina el momento específico, esto es porque el periodo preciso de un año, mes, día, etc. puede cambiar dependiendo si ocurre en un año bisiesto o no por ejemplo mientras que una duración es un método más generalizado tomando el tiempo como algo uniforme.

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

El periodo podríamos tomarlo como la duración del mes de febrero por ejemplo cuando es relevante si el año es bisiesto o no para determinar la cantidad de días que este mes tiene.

La función de duración lo podría tomar para ver cuánto se tomar un algoritmo en computar y arrojar un resultado final pues necesito minutos que son constantes.

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

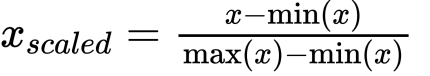
El concepto de MCAR nos dice que no existe una relación entre la ausencia de la data y ningunos de sus valores, observados o ausentes. Es decir que los datos que nos faltan son un subset aleatorio de la data. No existe la posibilidad de que de manera sistemática sea más probable que tengamos menos información sobre una variable que sobre otra.

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

Si estamos seguros de que la data es MCAR podemos utilizar la listwise deletion ya que esta es una imputación que se aplica cuando no existe una correlación entre valores múltiples.

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.
2. **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)**
3. listwise deletion.
4. pairwise deletion.
5. outliers cap via standard deviation.
6. outliers cap via percentile approach.
7. min-max scaling.
8. **¿En qué contexto de Machine Learning se recomienda utilizar Min Max Scaling? (6 pts)**

Podemos utilizar min Max scaling cuando tenemos bien definido nuestras variables dentro de un rango mínimo y máximo. Esto nos permite hacer un fitting más preciso al momento de realizar nuestro modelo sin crear un sesgo.



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)**

Log Transformations

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)**

Podríamos tener 3 nuevas variables utilizando una observación binaria para cada una de ellas.

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

Utilizamos One Hot Encoding cuando queremos transformar una variable categórica a distintas variables binarias ya que se nos hace más fácil para poder computar dentro de nuestro modelo.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Comida | Categoría | Calorios |  | Pollo | Pescado | Carne | Calorias |
| Pollo | 1 | 200 |  | 1 | 0 | 0 | 200 |
| Pescado | 2 | 150 |  | 0 | 1 | 0 | 150 |
| Carne | 3 | 300 |  | 0 | 0 | 1 | 300 |

ANTES 🡪 Despues

1. ¿Qué es un n-gram? (5 pts)

Un n-grama es una subsecuencia de n elementos de una secuencia dada. Es decir, un n-grama de 2 buscará los 2 items continuos en una secuencia. Se puede utilizar para estudio de secuencias de Aminoácidos en los genes, por ejemplo.

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*