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 Expresión regular es lo que se conoce también como una REGEX. Estos son comandos que los buscas es encontrar patrones para poder minar texto. Resultan realmente útiles cuando lo que queremos ver es que dentro de un string tenemos palabras, frases o cadenas de texto que se repiten mucho.

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

* Estas pueden ser ùtiles en páginas web en donde lo que se quiere comprar es que en efecto el teléfono de una persona sea de un determinado país. Por ejemplo, en el caso de Guatemala tenemos que deben de empezar con el prefijo +502 porque de esta manera determinamos que en efecto es un teléfono de Guatemala.
* Lo mismo sucede con las placas, cada país tiene un estándar bajo el cual deben de registrar las placas. Entonces podemos usar las REGEX para determinar que en realidad estas placas sean del vehículo que queremos.
* Otra función que se tiene en algunas aplicaciones de videollamadas o juntas en línea es la que la REGEX busca que el correo electrónico termine con cierto dominio. Si no encuentra la terminación del correo electrónico que corresponde entonces manda una alerta o no otorga el permiso.
* También las podemos usar para minería de texto, por ejemplo si lo que hacemos es pasar encuestas de respuestas abiertas, podemos determinar mediante el uso de REGEX las palabras más usadas y las frases que más se repiten.

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

* Cada variable que se mida debe de estar en una columna: Quiere decir que debemos tener en claro de que las columnas se utilizan para variables de los datos. Es importante saber de que, si se tiene que ir agregando columnas al dataset, esta tabla no es tidy.
* Cada observaciòn que tengamos debe de estar en una fila. El propósito de las filas debe de ser registrar lo que le sucedió a la observación.
* Los datasets deben de crecer hacia abajo y no a los lados. Si el dataset crece hacia los lados podemos determinar que ese conjunto de datos no está en formato tidy.

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



* Esta table no està en formato tidy por el hecho de que si queremos guardar los datos del año 2011 tendrìamos que agregar una columna más. Esto haría de que el dataset creciera de lado.
* Para convertirla en formato tidy deberíamos de tener una columna de cuntry, en donde vayan los países repitiéndose. Una columna de año, y una columna de la variable numérica que se quiere observar.

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



* Lo que sucede con esta tabla es que en las columnas de jugador tenemos en realidad 2 características. Lo que debería de ser para que sea una tabla tidy es que deberíamos de tener las columnas de Equipo, Jugador y Posiciòn. De esta manera nos aseguramos de que cada celda tenga únicamente una característica de la observación.

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



* Tenemos de que en las columnas esos rangos no dejan ver claramente la observación. Además de esto, en el momento de querer analizar los datos obtendremos un problema por la manera en que están colocadas las xs, esto porque dentro de la misma fila tenemos de que se marcan más de dos opciones.

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

* Las diferencias más importantes que tenemos es que las funciones period miden periodos de tiempo como los humanos las entendemos. En combio, con las funciones de duration lo que hace es que podemos determinar la distancia de tiempo de un período a otro de manera exacta.

1. ¿En qué contexto utilizaría una función period y en cúal utilizaría una función duration? (5 pts)
2. Explique el concepto de data Missing Completely at Random (MCAR). (6 pts)

* Esto es un concepto que significa que los missing values dentro de un dataset son completamente aleatorios y no muestran un patròn. En realidad si nos enfrentamos a un problema de missing values esto hace las cosas más sencillas porque quiere decir que los métodos que usemos para rellenar los datos pueden ser determinados según la distribución que tenemos de estos.

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

* Como los datos hacen falta de manera aleatoria. Si tenemos que la variable tiene una distribución normal entonces tendríamos de que podría hacer una imputaciòn por la moda

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.

* Harìa la imputaciòn por medio de listwise deletion. Esto porque tengo conocimiento de las variables. A pesar de que que no tenga una observación, puede que tenga las demás observaciones en las otras columnas.

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.

* Para determinas con què población de datos trabajar usarìa el método de outliers cap via percentile approach. Porque lo que quiero es satisfacer el 90% de la demanda. Si bien en cierto que dejarè clientes sin atender, esto sería lo mejor en relación al costo beneficio, probablemente.

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

* Esto se recomienda cuando estamos entrenando los modelos. Si le damos a los modelos datos que van de 0 a 1 entonces estos podrán procesarlos de mejor manera ahorrando recursos de la máquina.

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)
2. 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)

* Usarìa 2 variables como máximo, de lo contraria si meto las tres podría ser que el modelo cayera en la trampa de la colinealidad.

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

* El one hot encoding es lo que debemos de hacer cuando tenemos variables tipo texto que podrían ser tomadas como categorías. Esto ayuda a que los modelos que estemos entrenando e incluso las gràficas que hagamos en algunos lenguajes puedan procesarse de manera más ràpida.

1. ¿Qué es un n-gram? (5 pts)

* Esta es una de las maneras que se utiliza para pasar de un texto que no está ordenado a un texto ordenada.

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)

* Si se busca hacer esto lo que se debe de hacer es que se tiene que hacer un left join para poder encontrar las filas que no encuentran en la tabla B. Quedarìa algo así:

*SELECT \* FROM A \_LEFT\_JOIN B ON A.KEY = B.KEY \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

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)

* Lo que sucede es que turniting es entrenado con los mismos textos que las universidades del mundo registra y al mismo tiempo con una gran cantidad de textos que se encuentran en internet. Entonces al momento que uno le pasa Turniting a un documento lo que hace este software es que busca frases, oraciones y similitudes con los documentos que este haya entrenado. Posterior a eso tira una alerta cuando encuentra similitudes y así es como los profesores se pueden dar cuenta de esto.

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)?