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)

Las expresiones regulares son patrones, secuencias de caracteres que se utilizar para encontrar un patrón de búsqueda. Es la base para cualquier lenguaje ya que cada elemento de su alfabeto es una expresión, incluyendo una cadena vacía de caracteres.

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

Primero, se puede utilizar para validar contraseña. Dependiendo del parámetro utilizado, se pueden establecer condiciones que debe cumplir. Por ejemplo, que tenga por lo menos una letra mayúscula y minúscula, un número entero, un carácter especial, etc. Esto se puede realizar mediante las expresiones regulares.

Segundo, para validar un correo electrónico. Un correo electrónico debe cumplir con una estructura base y para ello se utilizan las expresiones regulares. Con ello se puede validar que tenga una arroba, un punto y termino con “.com”, “.edu”, etc.

Tercero, para la validación del número de una cuenta bancaria. Cada banco tiene una cantidad de dígitos específicos por cada tipo de cuenta. Dependiendo del banco, una cuenta de ahorro puede tener 10 y una monetaria puede tener 13 caracteres. Con las expresiones regulares se puede verificar esto y, además, que la cadena solo contenga números.

Cuarto, se puede utilizar para validar las placas de vehículos existentes y la asignación de una nueva. En Guatemala, las letras que se ven solamente son consonantes, por lo tanto, mediante las expresiones regulares se puede validar ese formato. Así como la cantidad de dígitos que tiene, y el largo de la cadena de una placa de vehículos.

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

Para que una tabla se encuentre en formato tidy, debe: (1) Tener cada variable como una columna. En ocasiones una variable puede contener dos variables, por lo que es importante separarlas, por ejemplo, si en una columna tenemos el nombre y la edad, es mejor separar en dos columnas diferentes. (2) Las observaciones deben formar una fila. (3) Las tablas se deben componer de una única unidad observacional. Si una tabla contiene la información de los clientes y las compras que ha realizado, es mejor separarlo y definir una columna única como primary key para identificar y poder relacionar las tablas.

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



La tabla contiene dos observaciones de análisis. Seria mejor separar la tabla en dos, en donde la primera contenga el nombre del país y el id de la región. En la segunda tabla sería mejor tener las observaciones por región y las columnas que representa la actividad por año.

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



La tabla no se encuentra en formato tidy ya que en la columna de jugador almacena dos tipos de variables diferentes. La tabla en sí almacena más de una unidad de análisis, ya que almacena la información del equipo y del jugador.

Para convertirla a formato tidy, primero hay que separar la columna de jugador en dos. La primera sería nombre\_jugador y la segunda posicion\_juego. Lo segundo a realizar es separar la tabla en dos, la primera sería “equipo” que contenga un id\_equipo, el nombre del equipo y la cantidad de jugadores en el mismo. La segunda tabla sería la información de jugador, que contenga el id\_jugador, el nombre del jugador, la posición del jugador en el campo, y el id\_equipo al que pertenece.

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



La tabla contiene datos que no son una única observación, es decir, analiza la distribución del producto por región y el rango de precios de venta del producto.

Para ello es mejor separa la tabla en dos, en donde la primera sea región y que contenga como columnas id\_producto y región, las observaciones serían los ids de los productos. La segunda tabla sería producto, en donde contenga columnas como id\_produto, nombre y precio exacto del producto, no es necesario poner un rango ya que es ambiguo e inexacto colocarlo así.

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

Las funciones de duration miden períodos de tiempo en números exactos de segundo. Mientras que un period no tiene una duración fija. En duration, 1 hora siempre representará un lapso de 1 hora, aún con posibles irregularidades en la línea de tiempo. Mientras que con period, 1 hora puede tener diferentes lapsos de tiempo cuando se mira una línea de tiempo después de un evento irregular.

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

Period se utilizaría en un contexto en que las fechas sean interpretar según como nos comunicamos. Por ejemplo, en un análisis de ventas por meses. En este caso es la misma fecha para todo como los fines de meses.

Durantion sería en un contexto de análisis como un cronómetro, cuantos días activos de trabajo se tuvieron, etc.

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

Trata de datos completa al azar. En el análisis de cualquier información, tiene la posibilidad de que pueda perderse. Lo que implica es que las causas de los datos que faltan no están relacionadas con los datos.

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

Recomendaría utilizar Pairwise ya que la eliminación por pares me permitiría no eliminar tantos datos en el caso que de la data no sea mucha.

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.

Para este caso sería mejor utilizar la opción d. Para que la data se acerque al rango intercuartílico definido. Ya que la data no es mucha es probable que con deletion se pierda mucha información y el análisis quede incompleto.

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 este caso sería la opción e. El min-max scaling le daría un mejor enfoque sobre el análisis a realizar ya que estaría escalado y no se vería tan dispersa la tabla de observaciones.

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

Lo utilizaría en sentido de que la data sea muy dispersa, es decir si se trata de cantidad existentes sobre un rango de 0 a 1000, sería mejor escalarlo un rango menor. Para ello ya no se tendría la data tan dispersa desde la variable en que se analizó, incluso el análisis puede ser más fácil al definir en una menor escala como a 1.

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)

Para este caso sería mejor utilizar una transformación logarítmica.

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)

Para un modelo econométrico solamente se necesitarían dos variables, en el caso de una variable categórica con tres niveles; la ausencia de un uno en las dos variables dummy indica que no pertenece a ninguna de esas opciones. Para un análisis de machine learning es mejor utilizar tres variables dummy.

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

Se utilizaría para transformar una variable categórica en variables dummy/binarias. En el contexto que lo utilizaría es sobre el sexo de una persona. Para un análisis más efectivo sería mejor definir la variable como binaria y que 1 sea mujer y 0 hombre.

1. ¿Qué es un n-gram? (5 pts)

Es una secuencia de caracteres que se utilizan para el análisis de oraciones y palabras. Estas se pueden utilizar sobre la base de fonemas, sílabas y letras. La “n” en sí representa la extensión del n-gram. Es similar en ese sentido a la clasificación de silabas ya que una monosílaba es de una sílaba, trisílaba es de tres y así sucesivamente. Dependiendo del uso de los n-gram, eisten los unigramas, bigramas, trigramas, etc. Por ejemplo, el unigrama para la siguiente oración: “Esta es una oración.” Sería: “Esta”, “sería”, “una”, “oración”.

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)

El funcionamiento de los analizadores de texto se basa principalmente en la utilización de n-gramas. Lo que hace es tomar una cierta cantidad de palabras y las comprar con los documentos que tiene como referencia para comparar con la cadena tomada originalmente.