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)

En lenguajes normales, son una secuencia de caracteres que forman un patrón de búsqueda o una cadena con información. Esto permite que un programa especializado identifique dichos patrones y puede ser útil al buscar información dentro de textos.

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

* Búsqueda de patrones: permite buscar patrones y extraer de ellos información, como sentimientos, predicción de cómo va a terminar, entre otros.
* Sustitución: como en este Word, se puede ver una palabra con expresiones regulares y verificar que sí esté bien escrita. En caso contrario, se sustituye el carácter malo.
* Verificación de inputs: se puede usar una expresión regular de ciertos caracteres para validar que un correo electrónico tenga las tres partes importantes antes de ser enviado al servidor (forma de pre limpiar la data).
* Compiladores: se usan las expresiones regulares en compiladores de código para validar que una instrucción haya sido escrita en una oración entendible en el lenguaje que estamos programando, es decir, que termine adecuadamente, que lleve todos los paréntesis o llaves, entre otros.

1. Explique brevemente las 3 condiciones que establecen que una tabla se encuentra en formato ***tidy.*** (5 pts)
   1. Que cada variable del dataset sea una columna.
   2. Que cada observación del dataset sea una fila.
   3. Cada tabla se compone de una unidad observacional.
2. 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, los años no son variables, son valores de una variable llamada **año**, por lo que, para hacerla ***tidy***, crearía una nueva columna llamada **año** y allí agregaría los años descritos arriba. Respecto los valores, haría una columna llamada **valor** donde iría el valor por año y por país. Quedaría así:

|  |  |  |
| --- | --- | --- |
| Country | Year | Value |
| Guatemala | 2008 | 5 |
| Guatemala | 2009 | 9 |
| Guatemala | 2010 | 13 |
| United States | 2008 | 9 |
| United States | 2009 | 13 |
| United States | 2010 | 23 |

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



A pesar de que el equipo si está bien clasificado, en la columna **Jugador** se mezcla el nombre y la posición que este desempeña. Para hacerla ***tidy***, y aprovechando que todos los nombres y posiciones se separan por un -, separaría por ese – y dejaría el nombre en su columna y la **posición**, que es una variable, la colocaría en una columna aparte. Pasaría el dataset a tener tres columnas: **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)



Esta tabla tiene dos variables identificadas como columna: **zona** y **precio**. Para hacerlo ***tidy*** crearía estas dos columnas y detallaría en cada observación si es urbano o rural y el nivel de ingreso por cada producto. Quedaría así

|  |  |  |
| --- | --- | --- |
| Producto | Zona | Precio |
| Banano 12 und. | Urbano | Q0-50 |
| Café molido 1lb. | Urbano | Q0-50 |
| Televisión Samsung 32´´ | Rural | Q500 + |
| Carne Molida 5lb | Rural | Q50-100 |
| Licuadora 1lt | Urbano | Q100-500 |

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

Duration representa el timespan exacto al segundo, es decir, no permite irregularidades humanas, tal y como period si lo permite.

Start = ymd\_hms("2016-03-13 01:00:00")

Si le sumamos 1 día con duration: 2016-03-14 02:00:00

Si le sumamos 1 día con period: 2016-03-14 01:00:00

Entonces, period si abarca que 1 día más humanamente nos referimos a la misma hora pero el día siguiente, mientras que el otro le agrega 1 hora más porque es exacto.

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

Cuando esté revisando datos que personas ingresaron manualmente con contextos humanos (una encuesta) usaría el period, ya que abarca esas irregularidades humanas. Cuando esté trabajando con la hora de la máquina o fechas ingresadas por la máquina, incluso rangos, usaría duration, ya que en este caso si es exacto.

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

No existe una correlación directa entre n valores que se perdieron, es decir, el que los datos se hayan borrado o corrompido no tiene una explicación relacionada a los datos, simplemente el accidente borró los datos aleatoriamente y no hay forma de recuperarlos. No hay una probabilidad definida de que los datos en tal patrón no estén.

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

Yo usaría Listwise ya que, al no poder predecir dónde podemos tener datos faltantes, no podes ir borrando donde sepamos que si puede haber problemas, entonces como no hay relación “aparente” entre la data perdida, es mejor borrar toda la observación ya que es data difícil de recuperar.

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.

Si tengo muchos datos perdidos, prefiero usar los datos que tengo y no borrar toda la observación solo por un valor faltante, aún con eso puedo trabajar con las encuestas. Además, como son valores únicos a los que no les conozco una tendencia evidente, los outliers pueden borrar valores que si son reales.

* 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.

En este caso, puedo normalizar la data por medio de percentiles ya que conozco capacidades de los camiones, se que ciertos modelos se comportan similar y tienen valores parecidos, por lo que sacar los outliers por sector (que pueden ser modelos) me va a permitir sacar valores atípicos por región en la data y no excluir de más o dejar datos que no me sirven porque son anormales.

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

Por ejemplo, si estamos haciendo un modelo para mejorar la calidad de una imagen, podemos usar el Min Max Scaling ya que sabemos cuales son los valores mínimo y máximo, ya que conocemos cuales píxeles brillan con mayor intensidad y con menor. Sabiendo eso, podemos usar esta técnica.

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)

Usaría la normalización por ***Bins***, ya que con ellos puedo ir seleccionando rangos de datos para irlos normalizando sin perder el comportamiento original que ellos presentaban, ya que esa tendencia es importante conservarla. Al ir seleccionando por segmentos, puedo normalizar usando la tendencia de ese segmento y así no sesgo demasiado los valores como si lo usara con una media u otra estrategia.

1. Si se tiene una variable categórica con tres niveles, cúantas variables dummy necesita para poder pasar la data a un modelo econométrico y de machine learning? (5 pts)

Necesitaría tres variables dummy, una por cada nivel que la variable categórica posee, ya que necesito reflejar qué opción es esa observación y, al mismo tiempo, decir cuales no son.

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

Cuando necesito que el modelo tome un orden dependiendo de las elecciones que hizo y para darle variables categóricas a la computadora que ella si entiende (ella si reconoce un 1, 0, 0 en tres columnas a decir que es pollo y no pepino o manzana).

1. ¿Qué es un n-gram? (5 pts)

Es una secuencia de caracteres en una expresión regular que puede ayudar a la máquina a predecir qué debe seguir o qué le falta para ser una oración correcta. Puede ordenarse unigramas (de una palabra), bigramas (de dos palabras) y así sucesivamente.

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)

Esta herramienta utiliza expresiones regulares para buscar patrones en diferentes archivos. Esta herramienta busca patrones de escritura en diferentes documentos y resalta aquellas frases que detectó iguales o parecidas en otros documentos por medio de similitudes en las expresiones regulares que segmentó por cada documento. Esto permite sacarle todas las expresiones regulares posibles a un examen, indicar qué está igual con otros documentos y sacar un porcentaje de cuántas expresiones regulares son iguales o muy parecidas en su ordenamiento.