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 una sentencia de código que utiliza una notación mediante símbolos para representar patrones o secuencias de caracteres. Por ejemplo, la expresión regular *[A-Za-z0-9]+@ [A-Za-z0-9]+[\.com]* representa cómo identificar una dirección de correo electrónico.

2. Enumere y explique brevemente cuatro aplicaciones prácticas en las cuales las expresiones regulares son utilizadas. (5 pts)
   a. Identificar cuentas de correo electrónico
      i. En los campos de registro de un formulario, para identificar si ingresó un valor válido
   b. Identificar números de tarjeta de crédito
      i. En el momento de pago(checkout) de los sitios web para validar que el valor ingresado cumpla con determinadas condiciones, por ejemplo aquí en Guatemala, tener 16 dígitos números.
   c. Crear identificadores de productos
      i. Algunos almacenes crean un código para identificar un producto usando las letras iniciales del modelo, cifras del año y

otros atributos del mismo. Para revisar que el identificador fue creado con éxito se utilizan expresiones regulares.

d. Contraseñas seguras
    i. En muchos sitios, como banca en línea se revisa que las contraseñas de los cuentahabientes posean determinadas condiciones como uso de mayúsculas, símbolos y números. Para verificar que la contraseña es segura se usan expresiones regulares.

3. Explique brevemente las 3 condiciones que establecen que una tabla se encuentra en formato ***tidy.*** (5 pts)
    a. Cada variable debe de tener su propia columna
    b. Cada observación debe de ser atómica y contar con su propia fila
    c. Cada valor debe tener su propia celda.

Donde variables son los atributos de un objeto, observación son las instancias de determinado objeto con sus atributos y valor se refiere a la ocurrencia específica de cada observación con un atributo determinado.

Nombre	Especie	Color
Juan	Perro	Negro
Pablo	Gato	Blanco

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

Country	2008	2009	2010
Guatemala	5	9	13
United States	9	13	23
Belgium	7	13	18
Argentina	9	18	28
France	7	13	24
United Kingdom	3	3	5
Germany	10	15	27
Poland	1	2	2

La tabla no se encuentra en formato tidy porque cada año es un valor de un atributo. Lo correcto sería que se mostrara así:

Country	Año	Cantidad

GT	2008	5
USA	2009	13
GT	2009	9

Los pasos fueron:

- Convertir el año en un atributo
- El valor de la columna original de cada año se volvió otro atributo llamado cantidad.
- Cada país y su año es una observación independiente.

5. 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 está en formato tidy porque las observaciones están mal organizadas. El nombre del jugador y su posición están concatenados.
   Lo correcto sería:

Jugador	Equipo	Posición
Federico Valverde	RM	Mediocentro
Cristiano	Juventus	Delantero

Equipo	Jugador
Real Madrid	Federico Valverde - Mediocentro
Juventus	Cristiano Ronaldo - Delantero
Barcelona	Frenkie De Jong - Mediocentro
Manchester United	Marcus Rashford - Delantero
Manchester City	Eric García - Defensa
Liverpool	Alisson - Portero
Atlético de Madrid	Joao Félix - Delantero
AC Milan	Sandro Tonali - Mediocentro
Roma	Pedro - Delantero
Inter de Milan	Achraf Hakimi - Defensa
Sevilla	Lucas Ocampos - Delantero
Valencia	Jose Luis Gayá - Defensa
PSG	Neymar - Delantero
Monaco	Cesc Fábregas - Mediocentro
Bayern Munich	Alphonso Davies - Defensa

Para ello se:

- Separó el campo jugador en 2.

- Se hizo un atributo separado
- Se hizo una observación para cada uno.

6. 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 es tidy porque hay una columna individual para el rango de precio además de que se hizo una especie de encoding para el área. También el nombre presenta una agrupación de campos. Lo correcto es:

Producto	Unidad	Medida	Área	Precio
Banano	12	Unidades	Urbana	0-50

   Para ello:
   - Separar el campo del nombre del producto
   - Hacer una columna de atributo para el área
   - Hacer una columna de atributo para el rango de precios

Producto	Urabno	Rural	Q0 - Q50	Q50 - Q100	Q100 - Q500	Q500 +
Banano 12 und.	x		x			
Café molido 1 lb	x		x			
Televisión Samsung 32"		x				x
Carne Molida 5 lb		x		x		
Licuadora 1 lt	x				x	

7. Sobre lubridate: Explique la diferencia entre las funciones period y las funciones duration. (5 pts)
   a. Duración guarda el número exacto de segundos entre un intervalo de tiempo.
   b. Período almacena el tiempo del ciclo, por ejemplo, sí preserva años, meses y días en sus variaciones, por ejemplo, duración no contempla los meses con 28, 31 días.

8. ¿En qué contexto utilizaría una función period y en cúal utilizaría una función duration? (5 pts)
   a. Period la usaría con cuestiones de carácter preciso en sistema humano, por ejemplo, súmarle un mes al 31 de enero me dará el último día de febrero (28/29).
   b. Duration serviría para cosas de índole matemática: saber con qué frecuencia en días sucede un evento astronómico por ejemplo.

9. Explique el concepto de data Missing Completely at Random (MCAR). (6 pts)
   Es un enunciado que usan los analistas de la información para explicar que no existe una causa definida para los valores faltantes, estos están vacíos al azar, sin un patrón regular u observable.
10. Si logramos verificar que la data faltante es MCAR, ¿cuál imputación recomendaría utilizar? (5 pts)

Depende de la naturaleza de los datos ¿son datos que sí es posible imputar o reconstruir? Si la respuesta anterior es afirmativa recomendaría una imputación que respete la distribución actual de los datos, esto porque la ausencia de los datos no tiene causa pero esto no contradice que los datos presenten una distribución normal, estándar, de Poisson, etc. Imputar datos usando reconstrucciones históricas(regresiones o clasificaciones) tampoco es mala idea si los datos muestran tendencias obviamente marcadas.

11. 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)
    a. listwise deletion.
    b. pairwise deletion.
    c. outliers cap via standard deviation.
    d. outliers cap via percentile approach.

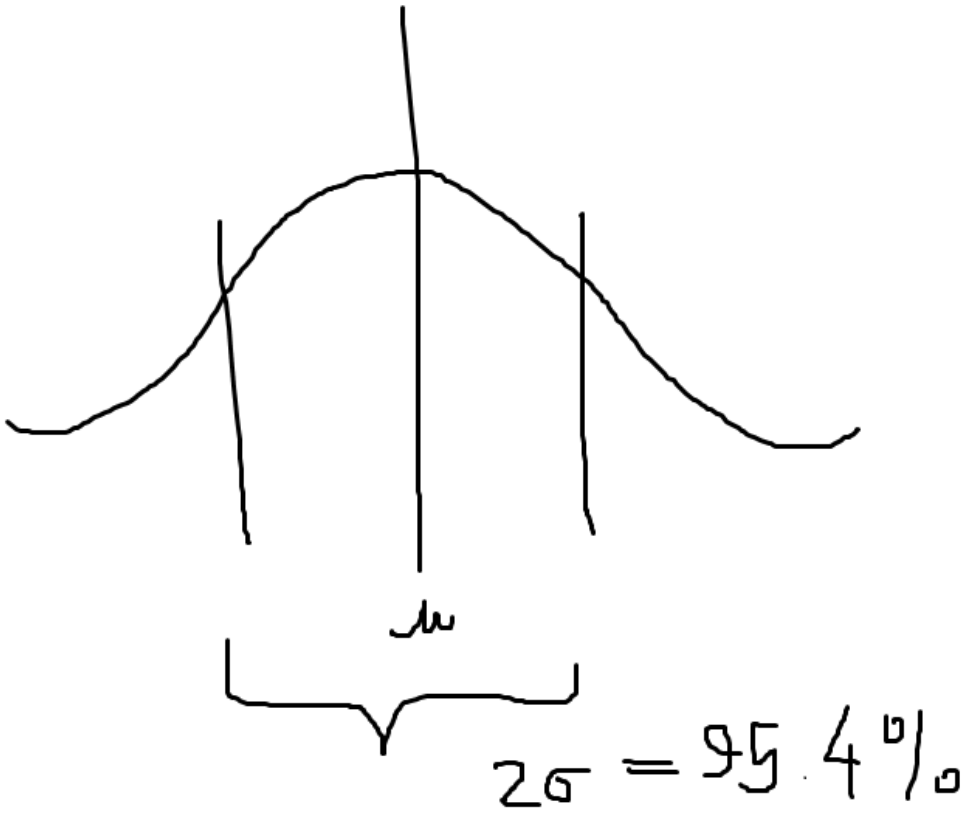
Primero, ¿es una encuesta abierta o cerrada? En una encuesta abierta no existe la posibilidad de imputación.
Segundo, cuál es la distribución de los datos, si es una distribución no normal, empobrecería el modelo eliminar los outliers. Tercero, qué tan importantes son el número de registros, ya que si no es importante puede eliminarse las observaciones, de lo contrario es mejor ignorar los atributos nulos.
Personalmente, yo utilizaría un pairwise deletion, porque las observaciones no son muchas. Eliminar los datos a rajatabla sería contraproducente para un análisis.

12. 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)

a. listwise deletion.
b. pairwise deletion.
c. outliers cap via standard deviation.
d. outliers cap via percentile approach.
e. min-max scaling.

Primero, eliminar outliers, ya que como necesitamos trabajar con el 90% de casos trabajaría con los percentiles representativos de una distribución estándar o ya sea con dos deviaciones estándar y tendría al menos el 95% de casos.



Segundo, como es demanda, un listwise sería sacrificar registros, pro lo que trabajaría con un pairwise, para obtener medias más significativas.

La teoría económica nos dice que en cuestiones de bienes el mínimo es 0 pero hay un máximo tanto como recursos del universo, por lo que no usaría un MinMaxScaling, probablemente usaría una escala estándar.

13. ¿En qué contexto de Machine Learning se recomienda utilizar Min Max Scaling? (6 pts)
    a. Primero, cuando se conocen los mínimos y máximos del atributo a estandarizar. (No es lo mismo estandarizar una demanda a estandarizar las medidas de un poste de luz que tiene un tope)
    b. Segundo, cuando nuestras variables contribuyen a traer sesgo a un modelo cuando se usan escalas muy diferenciadas, por ejemplo, modelos no jerárquicos como SVM o LDA.
    c. Tercero, deseamos reducir una dimensión, por ejemplo, no es lo mismo usar cifras grandes (898,990,786.655345) que números del 0 al 1.

14. 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 una librería con una estandarización Normal Z (StandardScaler en sklearn). Ya que al ser una distribución exponencial la media será de 0 y la desviación casi de una unidad.

15. 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)
    Usando un One Hot Encoding, si la variable tiene tres niveles se haría así:

Nivel	Variable-dummyOH1	Variable-dummyOH2
1	0	0
2	1	0
3	1/0 (Puede usarse 0 o 1)	1

    Depende del sistema se necesitarán al menos tantas variables dummy como categorías diferentes del atributo categórico.

16. ¿En cuál contexto utilizamos one hot encoding? (5 pts)
    En modelos donde necesitamos usar una variable de índole categórica pero esta posee varios niveles que no representan una jerarquía. Por ejemplo, colores: rojo, azul y verde, donde volverlas una escala 1, 2,3 nos sería contraproducente porque 2 > 1 y azul no es mayor que rojo. Entonces

ciframos la categoría en un grupo de columnas par, indicando su valor mediante ceros o unos.

Nivel	EncodeAzul	EncodeVerde
Rojo	0	0
Azul	1	0
Verde	0	1

17. ¿Qué es un n-gram? (5 pts)

Se refiere a la separación en grupos de n palabras cuando se analizará texto. Por ejemplo, separar la oración Anita lava la tina en un n-gram de 2 da como resultado:

Anita lava
Lava la
La tina

18. 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*