Universidad Francisco Marroquín
Data Wrangling 2020
Catedrático: Juan Carlos Girón
Auxiliar: José Josue

# Examen Final Data Wrangling 2020

Manuel Alexander Palencia Gutierrez
20160391

## 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 el cual nos permite encontrar patrones dentro de una cadena de texto, este lenguaje tiene sus propias reglas, sintaxis y operadores (.,(),[],+,\). El objetivo final es la búsqueda de patrones dentro de un texto, una vez identificados estos patrones se pueden eliminar, comprobar, validar, sustituir entro muchas otras funcionalidades.

2. Enumere y explique brevemente cuatro aplicaciones prácticas en las cuales las expresiones regulares son utilizadas. (5 pts)
    1. Se utiliza mucho para saber si una contraseña cumple con las reglas de la plataforma es decir el tamaño mínimo y máximo, que contenga caracteres especiales, que contenga números etc.
    2. Buscar un conjunto de palabras especiales dentro de un texto para analizar el contexto, idea y sentimiento que este representa. Por ejemplo queremos saber en cuantos tweets se menciona “Trump” y que se tenga la palabra “disconformidad”.

3. Validar la extensión de un archivo para evitar que coloquen cualquier tipo de archivo, por ejemplo a la hora de subir un archivo a una plataforma no queremos que sean capaz de subir archivos .bat ya que pueden contener algún código malicioso.
4. Evitar SQL injection, se utilizan las expresiones regulares en las plataformas web para evitar que personas ingresen en los campos cierta sintaxis de SQL con el fin de hacker y obtener información sensible.
5. Validar reglas de un correo, por ejemplo, que este no pueda empezar con un numero, no pueda contener un carácter especial etc.

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

Para que una tabla cumpla con el formato tidy se tiene que cumplir que cada observación de esta tiene que ser medible, que cada observación sea una fila independiente, y por último que cada variable es una columna.

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 |

Esta tabla se puede ver que esta en un estado parecido a tidy sin embargo no es su totalidad se puede colocar los años en una misma columna y cada esta con sus respectivos cantidad. Con la función de melt de R se puede hacer, lo que haría seria crear una tabla con 3 columnas country, año y cantidad.

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

| **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 |

Esta tabla no cumple con tidy ya que en la columna de Jugador se encuentran múltiples características el nombre y la posición se tiene que separa esta para que la data sea fácilmente accesible y tidy. Para cambiar Esta tabla a Tidy con R podemos separar los valores de la tabla de jugador en dos tablas y aparte agrupar por rol del jugador.

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

| 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 | |

Esta tabla no esta en formato Tidy debido a que todas las columnas de precio se pueden expresar en una solo y de igual manera colocar un valor en lugar de la x. Para transformar esta a tidy se puede utilizar la función melt de r esta lo que haría seria colocar las columnas de los precios en una con donde se especifique el precio y otra colmna de ubicación para contener si es Urbano o Rural..

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

Ambas funciones nos facilitan diferentes funcionalidades con el tiempo entre dos momentos. Sin embargo, su diferenciación radica en la presión e estas funcionalidades. La función duration mide la cantidad exacta de tiempo entre dos momentos es decir que esta toma en cuenta los años bisiestos, la cantidad de segundo entre otros básicamente funciona como un cronometro que empieza en el momento 1 y termina en el momento 2. Por otro lado la función periods no toma en cuenta los años bisiestos, segundos exactos o day light saving simplemente ejecuta una suma, resta o cualquier función necesario de las fechas es decir si quiere restar dos fechas este simplemente resta los números exactos sin tomar en cuentas las demás variables de tiempo. Esta es su principal diferencia y es por esto que tienen distintos casos de uso.

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

Utilizaría la función duration cuando requiera conocer el tiempo exacto recorrido entre de dos fechas por ejemplo necesito saber cuanto se tardo exactamente el viaje de un barco entre dos fechas para conocer porque se tardo ese tiempo. Por el otro lado la función period la utilizaría para hacer operaciones relacionadas con tiempo donde no me importe el transcurso o la precisión del tiempo, por ejemplo

tengo un dataset donde todo los datos de fecha están atrasados por un día utilizaría la función period para sumar a mi fecha un día.

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

Este concepto nos explica como dentro de un dataset los valores cambiantes son completamente random es decir que estos valores faltantes son totalmente independientes, no tienen ni una relación con las otras variables o no tiene un patrón de estos valores faltantes. El concepto contrario a este es "Missing at random (MAR)". Con el concepto MCAR se asume que los datos faltantes no están relaciones a ninguna variable de estudio

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

Se recomienda utilizar el método de imputación constante ya que MCAR no esta relacionada a ni un patron únicamente se quiere utilizar eso para colocar un valor constante a los valores faltantes. O utilizar una imputación múltiple es decir eliminar las filas con estos datos faltantes pero pueden tener distintas desventajas.

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.

Se utiliza el pairwise deletion debido a que esta trata de colocar valores en los mising data en base a la correlación de múltiples variables.

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.

Se utiliza este método ya que con este se busca eliminar los datos atípicos del data set con fin que no arruine la precisión de estimación de nuestro predicción.

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

Se recomienda utilizarlo cuando la data esta pre-procesada es decir que las maginitudes de los valores se encuentre en escala de 0-1, de esta manera logramos tener rangos mas manipulables por la maquina y pudiendo tomar decisiones en base a estos.

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)

Utilizaría la técnica de log Transformación ya que esta busca transformar los datos a una forma normal y nos ayuda con skew features.

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)

Se necesitaría 3 distintas columnas para las variables dummy la tabla quedaría de la siguiente manera:

| Variable 1 | Variable 3 | Variable2 |
|---|---|---|
| 1 | 0 | 0 |
| 1 | 1 | 1 |
| 0 | 0 | 1 |

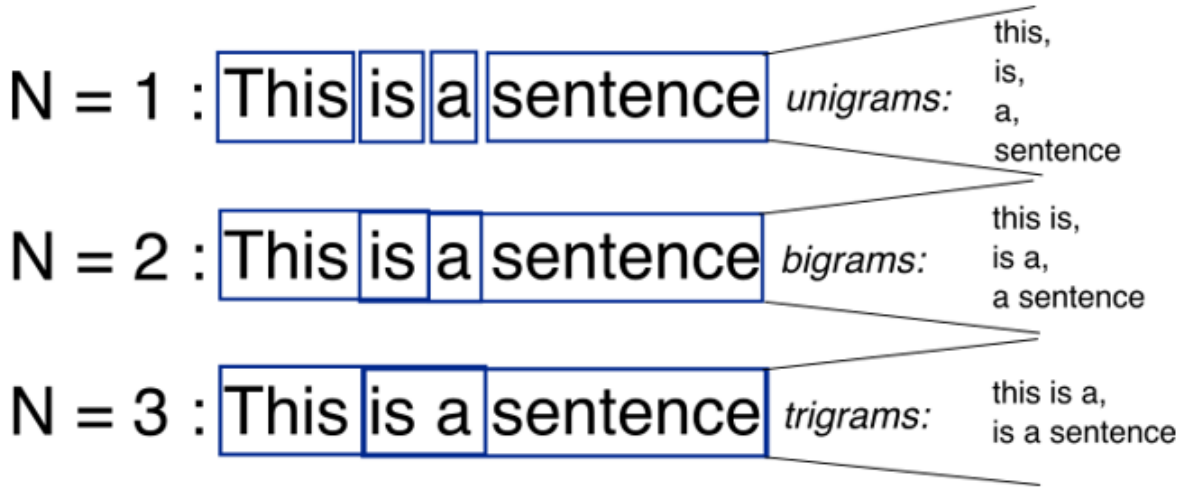
16.¿En cuál contexto utilizamos one hot encoding? (5 pts)

Lo utilizamos para transformar variables categóricas en un vector vinario es decir 0 y 1, a estos también se le llaman dummy variables. Estos se utilizan en campos de machine learning para que la computadora sea capaz de utilizar variables categóricas para entrenar y comprar distintos modelos.

| Variable 1 | Variable2 |
|---|---|
| 1 | 0 |
| 1 | 1 |
| 0 | 1 |

17.¿Qué es un n-gram? (5 pts)

Un n-gram es utilizado en el campo de la lingüística y probabilidad y es la secucia continua de tokens (palabras). Esto se utiliza en varios ámbitos como en el lenguaje natural, secuencia de genes etc.



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***