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 se podria decir que es un conjunto de de reglas que sirve para que la computadora busque un tipo de texto especifico, esas reglas es una secuencia de caracteres predeterminada. Esto sirve para buscar y validar texto, siempre y cuando coincidan con el patron que se habia predeterminado.
2. Enumere y explique brevemente cuatro aplicaciones prácticas en las cuales las expresiones regulares son utilizadas. (5 pts)
   * La primera podria ser para la busqueda y reemplazo de texto en los editores.
   * La segunda es para filtrar y extraer información de bases de datos.
   * La tercera puede ser la validación de formulario web, como ingresar correo electronico o numero de celular.
   * La cuarta puede ser para encontrar patrones en redes sociales, es decir los trends o hashtags.
3. Explique brevemente las 3 condiciones que establecen que una tabla se encuentra en formato ***tidy.*** (5 pts)
   * Cada variable debe estar representada en un columna distinta.
   * Por lo tanto, cada observación representa una fila.
   * Cada celda tiene su propio valor.
4. 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 la tabla no cumple con las condiciones de que cada variable en su propia columna ni cada observacion en su propia fila, esto hara que esta tabla por cada año nuevo se agregue una columna nueva lo cual no es tidy.

En este caso se podria realizar de la siguiente manera como ejemplo y seguiria para abajo con cada observacion.:

|  |  |  |
| --- | --- | --- |
| Country | Year | Medallas de Bronce |
| Guatemala | 2008 | 5 |
| Guatemala | 2009 | 9 |
| Guatemala | 2010 | 13 |

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 no esta en formato tidy por que no cumple las condiciones que ya se han mencionado que es que cada variable este en su propia columna, ni cada observacion en su propia fila, y hay celdas con dos valores.

En este caso se debera de separar los valores para que la tabla crezca para abajo.

|  |  |  |
| --- | --- | --- |
| Equipo | Jugador | Posicion |
| Real Madrid | Federico Valverde | Mediocentro |
| Juventus | Cristiano Ronaldo | Delantero |
| Barcelona | Frenkie de Jong | Mediocentro |
| Manchester United | Marcus Rashford | Delantero |

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 no esta en formato tidy por que primero no tiene cada variable en su propia columna, cada observacion deberia de tener su propia fila pero hay algunos que tienen varias observaciones en una sola y las columnas no tienen un nombre en si.

La tabla podria ser tidy de la siguiente manera:

|  |  |  |  |
| --- | --- | --- | --- |
| Producto | Tipo de Poblacion | Unidad | Precio |
| Banano | Urbano | 12 unidades | 0-50 |
| Café Molido | Urbano | 1lb | 0-50 |
| Television Samsung | Rural | 12” | 500+ |

1. Sobre lubridate: Explique la diferencia entre las funciones period y las funciones duration. (5 pts)
   * La funcion period se usa para representar lapsos de tiempo en termino year, month, week, days, hour, minute and seconds. Esta comunmente se usa para representar periodos recurrentes o intervalos de tiempo. Su longitud tiende a ser irregular.
   * Por el otro lado, la funcion duration representan una duracion/cantidad de tiempo siempre es en longitud de segundos.
2. ¿En qué contexto utilizaría una función period y en cúal utilizaría una función duration? (5 pts)
   * Period se utilizaria para representar el cumpleaños de alguien, o calculos que impliquen intervalos de tiempo.
   * Duration: la podria utilizar para representar el tiempo que tardo un programa en ejecturase o si quiere hacer caluclos matematicos como calcular la velocidad de un objeto.
3. Explique el concepto de data Missing Completely at Random (MCAR). (6 pts)
   * El hecho de que falten datos es independiente de los datos observados o no observados. En otras palabras, la presencia o falta de datos es aleatoria. Se podria decir que no hay diferencias sistematicas entre los que tienen datos faltantes y aquellos que estan completos
4. Si logramos verificar que la data faltante es MCAR, ¿cuál imputación recomendaría utilizar? (5 pts)
   * En este caso podria ser cualquiera, pero se podria usar por medio ya que es un metodo simple y eficiente (depende de los outliers) o se podria ir por la mediana.
5. 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.
   * El pairwise solo eliminaria los casos que tengan un dato faltante en las variables que se estan analizando. Esto nos permite mantener la mayor parte de los datos y por lo tanto nos da una mejor estimación.
6. 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)
7. listwise deletion.
8. pairwise deletion.
9. outliers cap via standard deviation.
10. outliers cap via percentile approach.
11. min-max scaling.
    * Con este metodo se podrian eliminar los outliers de la demanda mensual y esto nos hace tener una mejor estimación. En este caso, si no se quitaran estos valores atipicos podria ser que el modelo estime una capacidad que no cumpla para atendere el 90% de la demanda mensual.
12. ¿En qué contexto de Machine Learning se recomienda utilizar Min Max Scaling? (6 pts)
    * Se podria usar cuando las variables tiene diferentes escalas por lo que puede ser dificil comparalas unas con las otras, por ejemplo una es centimetros y otra son libras. Tambien puede ayudar a mejorar el rendimeinto de algunos algoritmos que son más sensible a la escala de datos, como lo puede ser un arbol de desición. Por otro lado, hay algunos modelos lineales que requieren que las variables esten centradas en cero entonces esto puede ayudar.
13. 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)
    * En este caso se podria usar la transformación logaritmica, esto basicamente es tomar el log de cada valor de la variable.
14. 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 necesitarian dos variables dummy, por que si se hacen 3 entonces se crearia multiconlinealidad que puede llevar a tener estimaciones sesgadas.
15. ¿En cuál contexto utilizamos one hot encoding? (5 pts)
    * Esta tecnica de transformacion de datos se utiliza más que todo cuando hay variables categoricas que se pueden hacer de formato binario.
16. ¿Qué es un n-gram? (5 pts)
    * Es una secuencia de n elementos, estos se usan comunmente para el procesamiento del lenguaje natural para presentar las relaicones entre palabras. Esto se usa para implementaciones como el autocorrector.
17. 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)
   * Comparaciones de coincidencias, tiene acceso a bases de datos extensas con las que compara lo que sube el estudiante.
   * Análisis de similitud, esto puede ser frecuencia de palabras, estructura gramatical, etc. Esto también puede ser estilo, si es similar al de un texto conocido.
2. 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)?