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 es una secuencia de caracteres que forman un patrón de búsqueda. Son utilizadas principalmente en búsqueda de patrones en cadenas de caracteres y en operaciones de sustitución.**

1. Enumere y explique brevemente cuatro aplicaciones prácticas en las cuales las expresiones regulares son utilizadas. (5 pts)
2. **Verificación de contraseñas: Las expresiones regulares se pueden utilizar para asegurarse de que las contraseñas cumplen con ciertos criterios de seguridad.**
3. **Buscadores de texto: Los buscadores de texto como la función search en Python utilizan expresiones regulares para buscar patrones de texto en archivos o cadenas de texto.**
4. **Procesadores de texto: Los procesadores de texto avanzados, como Microsoft Word o Google Docs, permiten buscar y reemplazar texto utilizando expresiones regulares.**
5. **Búsqueda de cadenas de proteínas: En bioinformática, las expresiones regulares se utilizan para buscar patrones específicos en secuencias de proteínas o ácidos nucleicos.**
6. Explique brevemente las 3 condiciones que establecen que una tabla se encuentra en formato ***tidy.*** (5 pts)
7. **Cada variable es una columna: Esto significa que cada característica o variable que se está registrando debe estar representada en una columna separada.**
8. **Cada observación es una fila: Cada fila de la tabla debe representar una observación única o un caso individual. Esto implica que los datos relacionados con una misma entidad o evento deben agruparse en una sola fila.**
9. **Cada valor es una celda: Cada celda de la tabla debe contener un solo valor, lo que significa que no debe haber celdas que contengan múltiples valores separados por comas u otros delimitadores.**
10. 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* por lo siguiente: 1) En un formato tidy, cada variable debe formar una columna y cada observación una fila. Aquí, los años (2008, 2009, 2010) son valores de una variable, posiblemente "Año", pero se han utilizado como encabezados de columna. 2) Los encabezados de columnas deben ser nombres de variables, no valores de una variable. Los años deberían estar bajo una sola columna que represente el tiempo.**

**Para convertirla en un formato *tidy*, se deberían hacer los siguientes cambios: 1) Crear una sola columna denominada "Año" que contenga los valores 2008, 2009 y 2010 repetidos tantas veces como sea necesario para acomodar todas las observaciones. 2) La columna que contiene los valores numéricos, actualmente separados en distintas columnas por año, debería consolidarse en una única columna, digamos "Valor" o "Cantidad".**

**El resultado sería una tabla con tres columnas: "País", "Año" y "Valor", donde cada fila representa una observación única para un país en un año específico, con su correspondiente valor.**

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 está en formato *tidy* por las siguientes razones: 1) Cada celda debe contener una única variable. En este caso, la columna "Jugador" contiene tanto el nombre del jugador como la posición que juega, lo cual son dos variables distintas. 2) Los nombres de las columnas deberían representar variables, y las filas deberían representar observaciones. En este caso, para estar en formato *tidy*, se debería tener una columna separada para los nombres de los jugadores y otra para sus posiciones en el campo.**

**Para convertir esta tabla a formato *tidy*, se deberían hacer los siguientes cambios: 1) Crear una columna llamada "Nombre del Jugador" y otra llamada "Posición", donde "Nombre del Jugador" contiene solo los nombres de los jugadores y "Posición" contiene las posiciones (Mediocentro, Delantero, Defensa, Portero). 2) Asegurarse de que todas las posiciones estén nombradas consistentemente en toda la columna "Posición".**

**El resultado sería una tabla con tres columnas: "Equipo", "Nombre del Jugador" y "Posición", con cada fila representando la observación de un jugador específico, su equipo y su posición.**

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 está en formato *tidy* por las siguientes razones: 1) Las columnas "Urbano", "Rural", "Q0 - Q50", "Q50 - Q100", "Q100 - Q500", y "Q500 +" representan valores de dos variables posibles, que podrían ser "Ubicación" y "Rango de Precio". 2) En un formato *tidy*, cada observación debe ser una fila y cada variable una columna. Las marcas 'x' indican la presencia de un producto en una categoría, pero no hay una columna que registre directamente esa información.**

**Para convertir esta tabla en formato *tidy*, se necesitarían los siguientes cambios: 1) Se debería tener una columna para "Producto", una para "Ubicación" y otra para "Rango de Precio". 2) Cada 'x' se convertiría en una fila en la tabla, con la información correspondiente al "Producto", la "Ubicación" y el "Rango de Precio".**

**El resultado sería una tabla donde cada fila representa una observación única. Por ejemplo, si "Banano 12 und." está marcado bajo "Urbano" y "Q0 - Q50", habría una fila con "Banano 12 und.", "Urbano", "Q0 - Q50". Cada combinación de producto y sus categorías se representaría en una fila distinta.**

1. Sobre lubridate: Explique la diferencia entre las funciones period y las funciones duration. (5 pts)
2. ***Period* se utiliza para representar intervalos de tiempo en términos de unidades de tiempo comunes, como años, meses, días, horas, minutos y segundos. Esta funcion mide intervalos que tienen en cuenta la información de calendario.**
3. ***Duration* representa intervalos de tiempo en términos de unidades de tiempo estándar, como horas, minutos y segundos. En esta función las duraciones no consideran la información de calendario, por lo que son más adecuadas para medir intervalos de tiempo exactos y precisos.**
4. ¿En qué contexto utilizaría una función period y en cúal utilizaría una función duration? (5 pts)
5. ***Period:* Cuando necesite representar intervalos de tiempo que tienen en cuenta el calendario, como años, meses y días. Por ejemplo, si deseo calcular la diferencia entre dos fechas teniendo en cuenta la cantidad de meses o días completos entre ellas, o si deseo agregar o restar meses o años a una fecha.**
6. ***Duration*: Cuando necesite representar intervalos de tiempo precisos en términos de horas, minutos, segundos, etc., sin considerar el calendario. Esto sería útil para medir el tiempo de ejecución de una operación o calcular la diferencia exacta entre dos marcas de tiempo.**
7. Explique el concepto de data Missing Completely at Random (MCAR). (6 pts)

**Se refiere a una situación en la que la probabilidad de que falte un valor en un conjunto de datos no está relacionada con ninguna característica o variable en el conjunto de datos.**

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

**Utilizaría *listwise deletion*. Esta imputación no introduciría sesgo en el análisis y estadísticas, y no distorsionaría las propiedades de la muestra. Además, es fácil de implementar en cualquier lenguaje.**

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.

**El análisis se realizaría utilizando todas las observaciones que tienen datos disponibles para cada variable específica. Esto significa que se excluyen solo las observaciones con valores faltantes en las variables involucradas, sin tener una pérdida significativa de datos.**

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.

***Min-Max Scaling* me permitiría ajustar los valores los datos para que estén dentro de un rango específico. Esto me sería útil para asegurarte de que todas las variables estén en la misma escala antes de realizar un análisis de capacidad de transporte.**

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

**Se recomienda en el contexto de Machine Learning cuando se desean normalizar las variables de un conjunto de datos para que estén dentro de un rango específico, generalmente entre 0 y 1. Esto es especialmente útil cuando se utilizan algoritmos que son sensibles a la escala de las características, como SVM y los algoritmos de gradiente descendente. Al aplicar el escalado Min-Max, se asegura que todas las características tengan la misma influencia en el modelo, evitando que las características con valores más grandes dominen las contribuciones y garantizando que el rendimiento del modelo sea más estable y equitativo en todas las dimensiones de los datos.**

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)

**Utilizaría *Log Transformations.* Esta técnica consiste en aplicar el logaritmo natural a los datos. Debido a que los datos se presentan en un comportamiento exponencial y se distribuyen de manera sesgada hacia valores más altos, aplicar el logaritmo natural me permitiría que los valores más altos se compriman y que los datos se asemejen más a una distribución normal.**

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)

**Se necesitaría crear dos variables dummy para poder incorporar esta variable en un modelo econométrico o de machine learning. Se crearían dos variables dummy de la siguiente manera:**

* + - **Variable Dummy 1: Representa la primera categoría y toma el valor 1 si la observación pertenece a esa categoría, y 0 en caso contrario.**
    - **Variable Dummy 2: Representa la segunda categoría y toma el valor 1 si la observación pertenece a esa categoría, y 0 en caso contrario.**
    - **La tercera categoría se puede inferir de las dos variables dummy: si ambas variables son 0, entonces la observación pertenece a la tercera categoría.**

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

**El *one-hot encoding* se utiliza para transformar una variable categórica en vectores binarios donde cada vector corresponde a una categoría o nivel diferente de la variable categórica. Esta técnica se aplica en el contexto de procesamiento de datos y modelado de machine learning cuando se desea representar variables categóricas en una forma numérica que sea adecuada para su inclusión en algoritmos de aprendizaje automático que requieren valores numéricos en lugar de categorías.**

1. ¿Qué es un n-gram? (5 pts)

**Un *n-gram* es una secuencia continua de n elementos (como caracteres, palabras o símbolos) que se extrae de un texto o una cadena de datos.**

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 \_\_\_ JOIN B ON A.KEY = B.KEY \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

***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)

**Turnitin utiliza el procesamiento de texto, la manipulación de datos, y la extracción de información relevante de documentos para detectar plagio. Turnitin puede utilizar plataformas de bases de datos como SQL para gestionar grandes bases de datos de contenido textual y utilizar algoritmos de coincidencia de texto para comparar el texto del usuario con sus bases de datos. Es posible que Turnitin utilice expresiones regulares como parte de sus métodos de detección de plagio. Podrían utilizar expresiones regulares identificando patrones específicos o características de los documentos que son indicativas de plagio. Por ejemplo, se podría utilizar una expresión regular para buscar similitudes exactas entre el texto del estudiante y el texto en línea. La expresión regular podría buscar fragmentos de texto que coincidan palabra por palabra o incluso carácter por carácter con el texto de la fuente en línea.**

1. 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)?