**MEMORIA**

**Hugo Meschini Rocha**

**Tema:**

Tras hacer muchas búsquedas de temas y datasets, he llegado al tema de Dataset de Consumo por perfil de hogares.

De forma inicial he obtenido 3 distintos datasets, pero cuando empecé mi análisis exploratorio he decidido trabajar solo con uno de ellos ya que los otros dos no tenían informaciones distintas que me llevaría a generar análisis distintos, por lo que he decidido trabajar solo con uno dataset que contiene mejores detalles.

**Exploratorio inicial:**

Considerando que ya he comentado en el punto anterior, solo destacar que he cogido 3 Datasets, pero que uno de ellos era más completo y detallado y que me iba permitir hacer distintos análisis, por lo que me quedé trabajando solo con uno.

**Cleaning:**

Todo el trabajo he generado desde un jupyter notebook. En este documento he podido indicar través del Markdown cada una de las etapas que he tenido en cuenta, desde el exploratorio de datos hasta la producción de los gráficos que he considerado para mi presentación.

Podéis verificar que he indicado em el jupyter las siguientes etapas:

* Generación de los primeros dataframes a partir del archivo “.csv”
* He identificado 3 errores (3 líneas) que se refieren a edades incorrectas, por lo qué las he borrado
* Generación de grupos por edad
  + El dataframe tenía una columna de “Year\_Birth”, podéis verificar que habían personas de casi todas las edades posibles, por lo que he decidido generar grupos por edad de cara a facilitar mi análisis
* Luego, ya con los grupos generados, he decidido eliminar y renombrar el nombre de algunas columnas que – importante destacar que en dicho momento no tenía claro si iba a utilizar todas las columnas, pero tenía claro las que no iban a impactar en mi análisis (estas columnas han sido borradas).
* Como parte del proceso de limpieza y preparación de datos, he identificado que habían 24 líneas con el contenido “NaN” en la columna “Income”, así que decidido rellenar estos espacios con la media de la columna.
* En la columna “Marital Status” he identificado 3 definiciones que sumaban 7 líneas cómo total, así que las he sustituido por otro valor ya definido en la columna, simplemente por estandarizar la columna.
  + Tras hacer las modificaciones, he aprovechado para sacar un primer grafico relativo al estado civil de los participantes.
* En este momento también he podido generar un grafico de barras relativo a los grupos por edades que he creado en el inicio del trabajo.

**Dataframes:**

He identificado 4 dataframes que he generado a partir del dataframe inicial, los mismo indican informaciones relativas a lo siguiente:

* Independiente (jovens sin hijos, entre 26 e 45 años)
* Familia (entre 26 y 55 años con hijos)
* Independiente Senior (a partir de 46 años sin hijos)
* Familia Senior (a partir de 46 años con hijos)

He generado un dataframe para cada grupo, el objetivo es hacer un análisis comparativo con los importes totales e identificar los gastos por perfiles.

**Análisis:**

* El primer análisis ha sido basado en los gastos de cada grupo por categorías *(Wine, Fruits, Meat, Fish, Sweet, Gold).*
* He identificado los gastos de cada grupo categoría y he hecho un análisis comparativo donde he identificado el % del Dataframe Total, y he hecho la comparación por grupo, por ejemplo, se indica que 50% de las personas gastan más con vinos, pero en el grupo X se nota que se gasta 40% (o 60%). Así que podemos identificar cuanto gasta cada perfil por categoría.
* Otro análisis ha sido relativa al numero de compras hechas por grupo, donde vemos cuál el porcentaje de compras hechas en tiendas físicas o por la web.
* Otro análisis ha sido del promedio de renta anual por grupo.
* El ultimo análisis ha sido del promedio de gastos por grupo en comparación con el promedio total.

**Mis consideraciones:**

Inicialmente me gustaría destacar que el proyecto ha sido muy interesante por ofrecer la posibilidad de hacer un análisis exploratorio de datos. Los puntos que me gustaría destacar:

* La búsqueda por dataframe: pensaba que seria lo más sencillo, y tras pasar días buscando informaciones relativa a otro tema, he decidido cambiarlo por no encontrar nada que me podría generar una cierta complejidad de análisis y también de códigos.
* He pensado en algunas hipótesis iniciales, pero mientras avanzaba con el análisis exploratorio he podido identificar que podría hacer otras cosas con las informaciones. En un primer momento no tenemos claro todo lo que podemos sacar del dataset, pero sí que se pude hacer muchas cosas. Finalmente he enfocado en algunas tantas análisis de cara a respetar el tiempo de presentación.
* Tuve muchísima dificultade en generar uno de los gráficos (stacked bar), incluso con la tutoría no he podido hacerlo. Así que he investigado mucho hasta llegar a una solución que me generó el grafico en una sola línea (pero antes tuve que generar un pequeño dataframe solo con las informaciones que eran necesarias poner en dicho grafico.
* Con respecto a la presentación, no tuve muchas dificultades ya que en los últimos 10 años vengo preparando powerpoints y presentando a los clientes de mi ultimo trabajo. En este caso me quedé enfocado en los gráficos y espero poder explicarlos bien, mi objetivo no es incluir mucho texto, pero sí que es importante construir un storytelling y tener una buena conclusión.

\* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \*