Análisis y Limpieza de Datos

Análisis del Presupuesto Familiar mensual

Profesor: Jorge Castillo

Alumno: José Pérez

# Introducción

El presente informe busca cumplir con 2 objetivos generales, a saber:

* Objetivo implícito: mostrar que se conoce y se tiene dominio teórico y práctico de los contenidos abordados en el curso “Análisis y Limpieza de Datos” del MDS2019, impartido en la UDD.
* Objetivo explícito: explorar, conocer y analizar el presupuesto familiar mensual, en Santiago de Chile. Y de esta forma poder responder preguntas de valor como, por ejemplo:
  1. ¿Podemos conocer – predecir -, el nivel socio-económico de una familia, dado su gasto mensual en el ítem alimentación?
  2. ¿Podemos conocer – predecir -, el gasto mensual familiar en el ítem telecomunicaciones?

Ambas son preguntas relevantes de responder para las distintas industrias de productos y servicios en el mercado. En el caso de las de las preguntas ejemplos planteadas, podrían estar asociadas al sector económico del retail y al de telecomunicaciones, respectivamente. Por cierto, hay muchas más preguntas que se pueden intentar responder mediante este análisis de datos.

# Contexto.

Para la elaboración del presente trabajo se usarán los DataSet asociados a la **"VIII Encuesta de presupuestos familiares"**, disponibilizados por el Instituto Nacional de Estadísticas (INE), para el período junio 2018. En su sitio web, INE señala respecto de esta encuesta:

*“La Encuesta de Presupuestos Familiares (EPF) es una encuesta socioeconómica aplicada a hogares, cuyo propósito es recopilar información sobre gastos en los que estos incurren y los ingresos que perciben en un período de tiempo determinado. Actualmente, su trabajo de campo tiene una duración de un año calendario, tiempo durante el cual se recolecta información en todas las capitales regionales del país y sus principales zonas conurbadas. Su información se produce cada cinco años.”*

Los DataSet pueden ser descargados desde el siguiente enlace:

<https://datosabiertos.ine.cl/datasets/187191/viii-epf/>

En ese mismo enlace se encuentra también la documentación de las características y tipos de variables que componen los DataSet, en el archivo PDF, “diccionario-de-variables-viii-encuesta-de-presupuestos-familiares-(epf)”

Entonces en lo que se refiere a la data, en específico, se tiene:

* **BASE\_PERSONAS\_VIII\_EPF.csv** : con 247 variables y 48.308 observaciones.
* **BASE\_GASTOS\_VIII\_EPF.csv** : con 14 variables y 1.064.239 observaciones.

Dada la cantidad de variables que componen los DataSet – en total 261 variables -, se efectúa un estudio previo de la documentación de la data, para entenderla mejor y preseleccionar aquellas características candidatas a ser usadas en este trabajo académico.

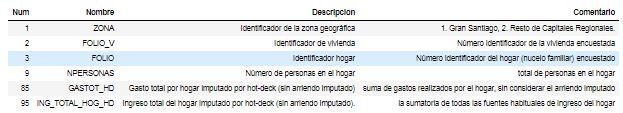
Luego de varios intentos y reflexiones se elige analizar la data a nivel de FAMILIAS y no a nivel de individuos. Esto se explica porque este Análisis de Datos, intentará responder preguntas asociadas con el GASTO FAMILIAR mensual, tal como se mencionó en los objetivos.

# Desarrollo.

A continuación, se efectúa una “presentación” de los DataSet usados, en total 3.

### Presentación del DataSet Personas y preselección de variables

Tal como se mencionó con anterioridad, desde el DataSet  **BASE\_PERSONAS\_VIII\_EPF.csv** se ha preseleccionado las columnas que han parecido relevantes de las personas, pero a nivel de "FAMILIA", en lugar de individuos:



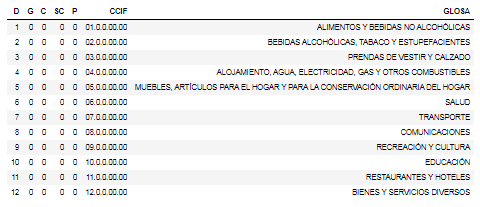
### Presentación del DataSet Gastos y preselección de variables

Del mismo modo, desde el DataSet **BASE\_GASTOS\_VIII\_EPF.csv** se han preseleccionado las siguientes variables:



### Presentación del DataSet CCIF

Para tener un mejor entendimiento de la encuesta, se presenta el código CCIF que corresponde a la "Clasificación Consumo Individual por Finalidad que contiene la descripción completa de todos los códigos de gasto, abiertos por División (D), Grupo (G), Clase (C), Sub\_Clase (SC) y Producto (P) En total son 12 divisiones que engloban a todos los posibles gatos mensuales de un grupo familiar :



A continuación, se realiza la exploración de la data.

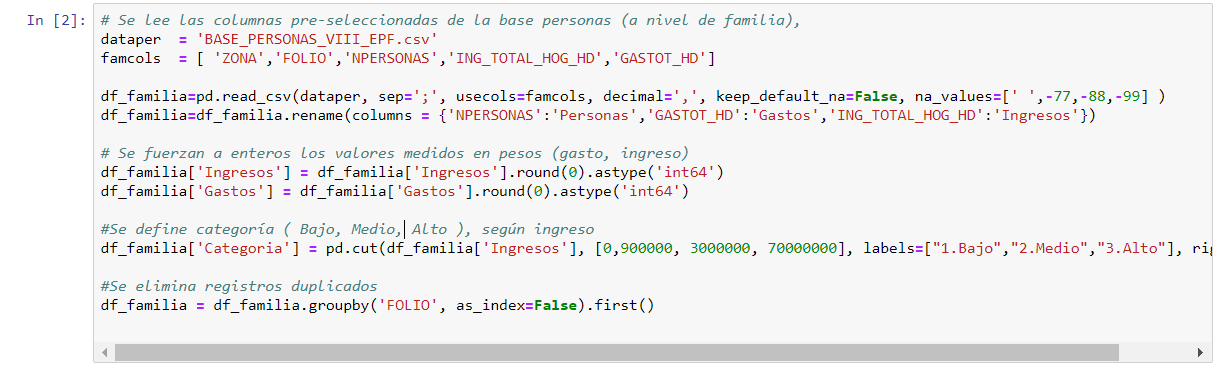
### Exploración de la Data

La exploración de datos se realiza en forma dirigida, según el objetivo presentado con anterioridad. Se realizarán búsquedas de datos relevantes intentando limpiar aquella data que no aporte un valor significativo, eliminando de esta forma el posible ruido que acompañe estos DataSet.



## DataSet FAMILIA (a partir de la base de personas)

* Se lee las columnas pre-seleccionadas de la base personas (a nivel de familia): ZONA', 'FOLIO', 'NPERSONAS', 'ING\_TOTAL\_HOG\_HD' y 'GASTOT\_HD'
* Al realizar la lectura del DataSet se aplican los siguientes parámetros
  + keep\_default\_na=False, na\_values=[‘ ‘,-77,-88,-99] esto para que todas las muestras que tengan esos valores, sean filtradas desde el DataSet leído y no generen ruido indeseado.
  + decimal=’,’ esto para que los valores numéricos que vienen en el DataSet con el separador decimal ‘,’ sean interpretados correctamente en la lectura y evitar un posterior tratamiento de formateo de los datos.
* Se fuerzan a tipo entero los valores medidos en pesos (gasto, ingreso)
* Se define de forma arbitraria una categoría (Bajo, Medio, Alto), para la familia, según su nivel de ingreso. Los valores dados, para esta definición de categoría, pueden ser ajustados y mejorados de acuerdo a nuevos criterios. Para ello se usa la función “cut()”
* Se eliminan registros duplicados (la base personas contiene información repetida por individuos). Se realiza una agrupación por la columna ‘FOLIO’



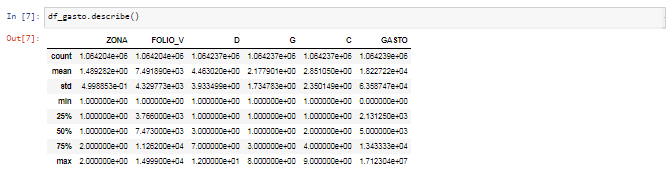
* Se asegura que el DataSet leído, tenga las columnas con el formato deseado, y además se mira valores descriptivos del DataSet, mediante las funciones “info()” y “describe()” respectivamente.

## 

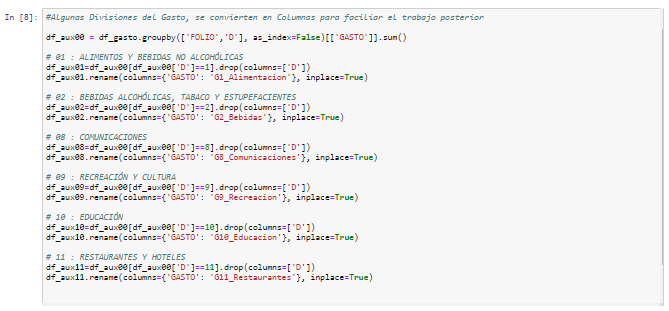
## DataSet GASTOS (solo en algunos ítems seleccionados)

* Se leen las columnas pre-seleccionadas de la base GASTOS (a nivel de familia) 'ZONA', 'FOLIO\_V', 'FOLIO', 'CCIF', 'D', 'G', 'C', 'GASTO' y 'GLOSA'
* Al realizar la lectura del DataSet se aplican los siguientes parámetros
  + keep\_default\_na=False, na\_values=[‘ ‘,-77] esto para que todas las muestras que tengan esos valores, sean filtradas desde el DataSet leído y no generen ruido indeseado.
  + decimal=’,’ esto para que los valores numéricos que vienen en el DataSet con el separador decimal ‘,’ sean interpretados correctamente en la lectura y evitar un posterior tratamiento de formateo de los datos.
* Se asegura que el DataSet leído, tenga las columnas con el formato deseado, y además se mira valores descriptivos del DataSet, mediante las funciones “info()” y “describe()” respectivamente.

## 



* Se seleccionan solo algunos ítems de gastos (Alimentación, Bebidas Alcohólicas, Comunicaciones, Recreación, Educación y Restaurantes) para continuar con su exploración y análisis posteriores.

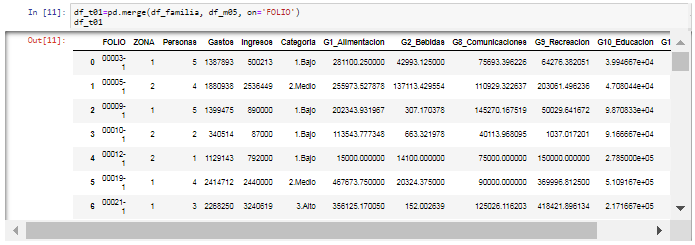


* Se aplica “Merge” para dar forma a un nuevo DataFrame del Gasto por Familia en los ítems preseleccionados (Alimentación, Bebidas Alcohólicas, Comunicaciones, Recreación, Educación y Restaurantes)

## 

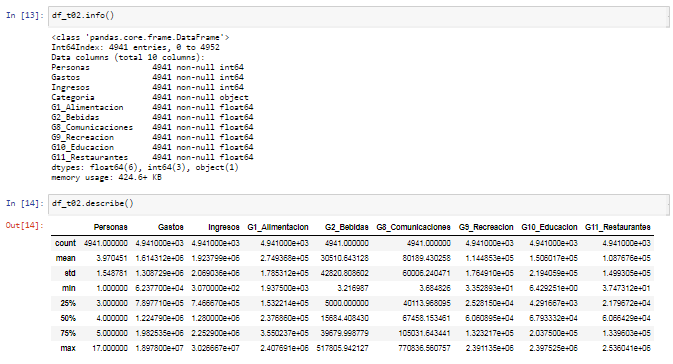
## DataSet GASTOS x FAMILIA

* Para construir esta relación, en primera instancia se combinan los 2 DataSet anteriores, con la información relevante, mediante la característica “Folio”. Se usa la función “merge()”



* En una segunda etapa son filtrados del DataFrame, solamente aquellas características que interesa relacionar.



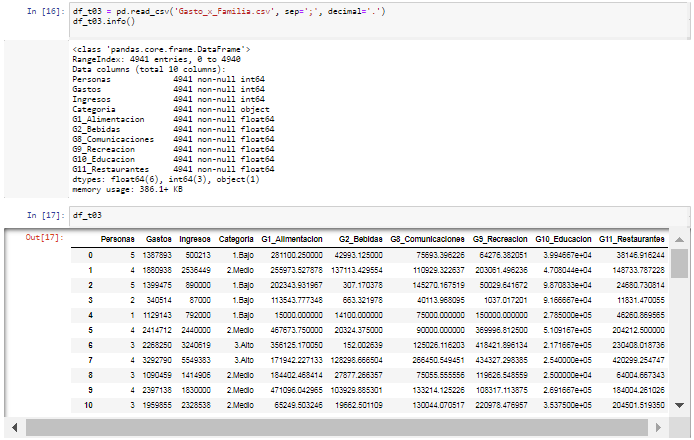


* Finalmente, se guarda el DataFrame como CSV, como medida de seguridad. Ahora este nuevo DataSet será usado como base para continuar con este análisis de datos.



## Validación

* Para validar el DataFrame guardado, se vuelve a leer y es comparado. Se observa que el DataFrame es el mismo que se generó.



# Conclusiones

Creo que se ha demostrado a lo largo del desarrollo de este informe que se cumple con el objetivo de haber adquirido conocimiento teórico y dominio práctico de los contenidos abordados el curso de Análisis de Datos.

Con relación al segundo objetivo planteado de responder las preguntas de valor realizadas, primero en lo que se refiere a la posibilidad de predecir si una familia clasifica dentro de una de las categorías definidas como Bajo, Medio o Alto, y segundo predecir cuál es el gasto familiar en el ítem comunicaciones, en ambos casos esto se puede abordar mediante modelos de aprendizaje supervisado. En el primer caso mediante un clasificador usando Support Vector Machine, y en el segundo caso usando una Regresión Lineal. Se entiende que los modelos de aprendizaje supervisado mencionados, forman parte del curso de Machine Learning, y en ese sentido, se sugieren como una alternativa a seguir para dar continuidad a este trabajo, pero que no se abordan dado que no pertenecen al ámbito del curso de Análisis de Datos.

Una de las mayores complejidades de este proyecto, ha sido la gran cantidad de variables de los DataSet usados. En el DataSet de personas se cuentan 247 variables y en el DataSet gastos hay 14 variables, es decir, en total se tienen 261 variables. El conocer los DataSet para posteriormente realizar una preselección de variables, ha representado un gran esfuerzo e inversión de tiempo. Junto con esta complejidad mencionada, viene aparejada también la enriquecedora experiencia de ser capaz de ir tomando decisiones para ir avanzando – en ocasiones retroceder y volver a avanzar – en la consecución básica de contar con un DataSet “adecuado”, según las variables que se eligen y la cruza de información realizada entre ambos DataSet.

Se ha comprendido la enorme importancia de conocer los DataSet que se están usando. Se ha abordado y solucionado técnicamente la adecuada homogeneidad de los formatos de los valores de las variables seleccionadas, por ejemplo, asegurar el formato de los valores numéricos tales como ingreso, gasto, etc.

Se ha limpiado la data de ambos DataSet que podría a la larga, representar ruido en términos del análisis que se quiere realizar.

Se ha agrupado información desde el DataSet personas, porque lo que se ha querido analizar es una familia y no datos de individuos.

Se ha relacionado información de 2 DataSet, personas y gastos, mediante la variable común ‘FOLIO’, construyendo así un nuevo DataSet con la información requerida, correctamente formateada, agrupada por familia, y sin valores que metan ruido, para un sugerido posterior modelamiento usando aprendizaje supervisado.

Por último, decir que, desde el punto de vista del desarrollo técnico, se ha ido depurando mediante sucesivas versiones de Jupyter-Notebooks las distintas operaciones realizadas sobre los DataSet. Por dar un ejemplo de ello, mencionar que, para la lectura de los DataSet, se realizó un estudio práctico de los parámetros que permite la función “pd.read\_csv()”, de tal forma que en la misma lectura se asegurase el formato de los datos decimales que en el archivo vienen con separador de decimal ‘,’ y no ‘.’, así evitando posteriores re-trabajos de formateo de la data.