

PROGRAMACIÓN DATA SCIENCE 2022

Integrantes

Asignatura: Programación Data Science

Profesor: Eduardo Celis

Nombre: Salomón Salinas

**Informe Programación para data Science**

**1.- Librerías a Utilizar en este Informe.-**

**Librerías para elaborar el proyecto, a través de la fuente recopilatorio de datos llamado 'BankData.csv', también se deja especificado que, si arroja algún error de las librerías, es porque, no existe la instalación del paquete (pip install ....).**

import pandas as pd

import numpy as np

import matplotlib as mpl

import matplotlib.pyplot as plt

import seaborn as sns

import warnings

warnings.filterwarnings("ignore")

%matplotlib inline

import itables

import itables.options as opt

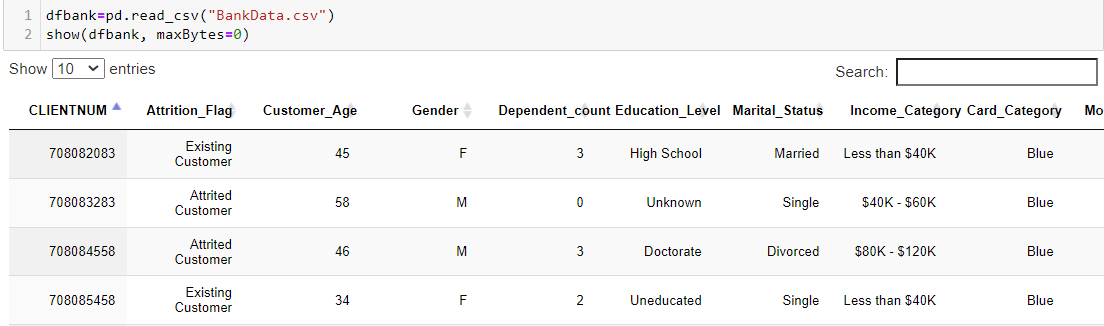
from itables import show

from itables import init\_notebook\_mode

init\_notebook\_mode(all\_interactive=True)

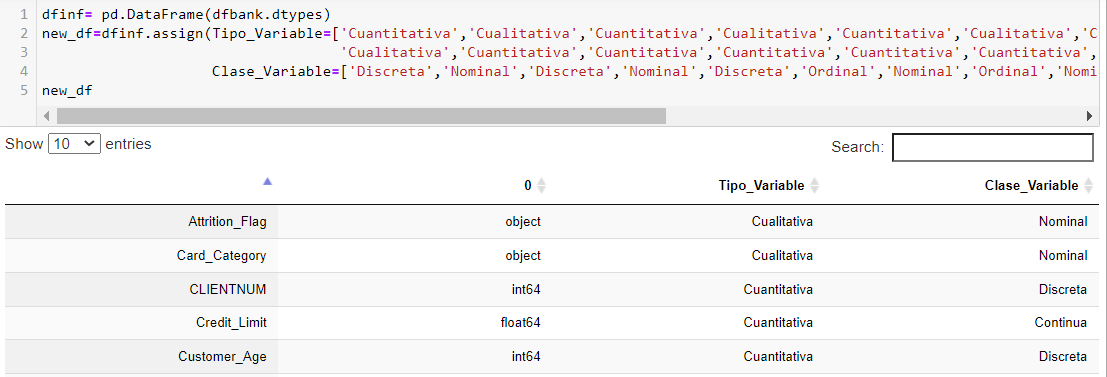
**2.- Lectura del DataSet.**

**Se presentará un dataframe dfbank que contiene los datos de la dataset llamado BankData.csv, para realizar el análisis correspondiente y descomponerla, para este caso se realizó una lectura de la misma y se muestra en pantalla.**



**2.1.- Clasificación de campos de data BankData.**

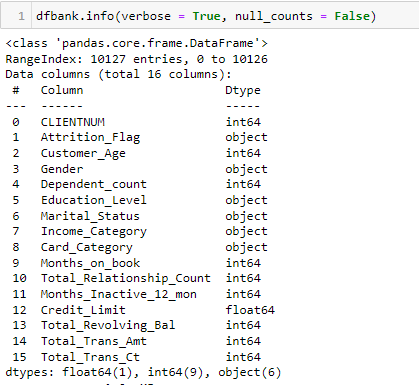
**Tabla de definición de campos, con clasificación de sus columnas, se construyó un dataframe para adicionarle unas columnas a la nueva dataframe principal llamado " new\_df "para identificar los diferentes tipos de campos, de modo que se pueda apreciar un diccionario de datos.**

****

**3.- Exploración de la dataset BankData.**

Se mostrará algunas funciones dentro de la exploración para definir y ver con que tipos de datos se ésta trabajando.

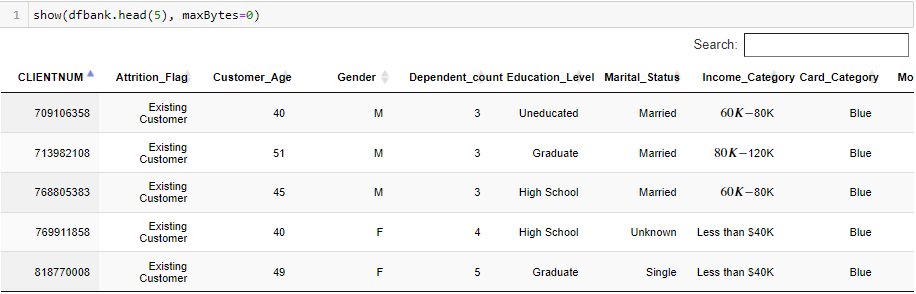
1.- INFO(), Lo primero que se realizará en la exploración, es verificar, que tipos de campos tenemos en el dataset para explorar, usando la función info().



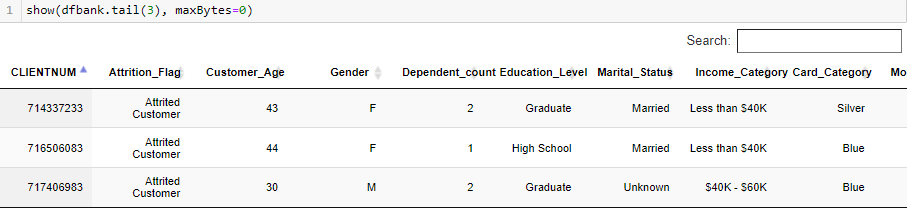
2.- SHAPE, Se Muestra a continuación el dimensionamiento de la dataframe dfbank, de modo que se pueda ver, en forma general, con cuantos datos y columnas se está, trabajando.

dfbank.shape obteniendo como resultado (10127, 16)

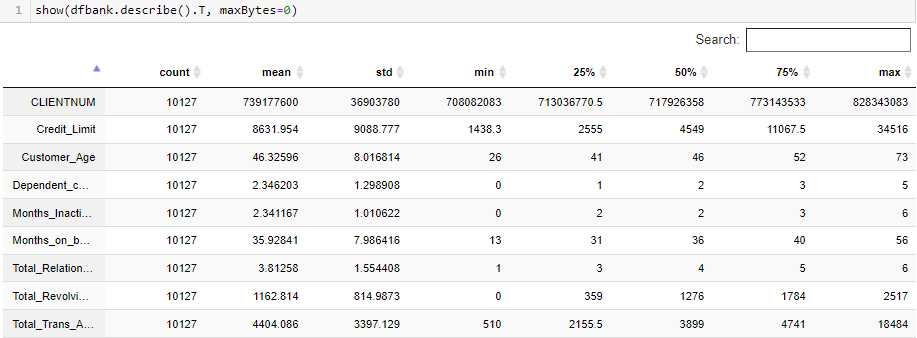
3.- HEAD(), A continuación, se mostrará las primeras 5 filas de la dataframe dfbank, que esta, explorando, usando la función de extracción head().



4.- TAIL(), Por consiguiente, se explorará las últimas filas de la dataset para corroborar toda la información obtenida, usando la función tail().

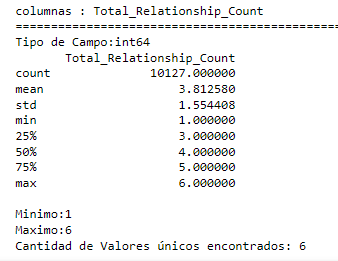


5.- DESCRIBE(), Por ende, para una mejor representación de los datos se hará un resumen estadístico de toda la dataframe bfbank, pertenecientes a la dataset BankData.csv, por lo cual, se tomará todos los datos numéricos y se realizará un resumen obteniendo datos de importancia, para esto de obtendrá, los resultados con la función describe(), por lo cual, nos entregará información relevante como la media,desviación standard, los mínimos, los máximos y además, percentiles etc.



**3.- Buscando Datos en la Exploración.**

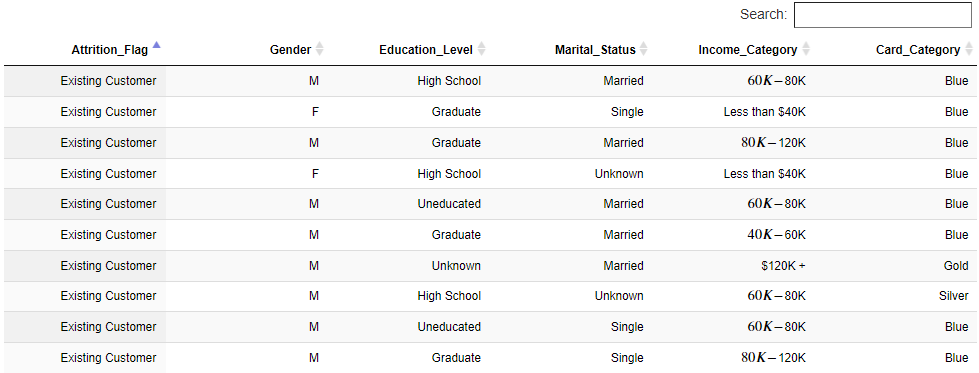
**Se realizó un control de toda la tabla para resumir sus características principales en la exploración, ya sea buscando datos de valores únicos, valores mínimos y máximos, además, verificando el tipo de datos, todo esto recorriendo la dataset.**



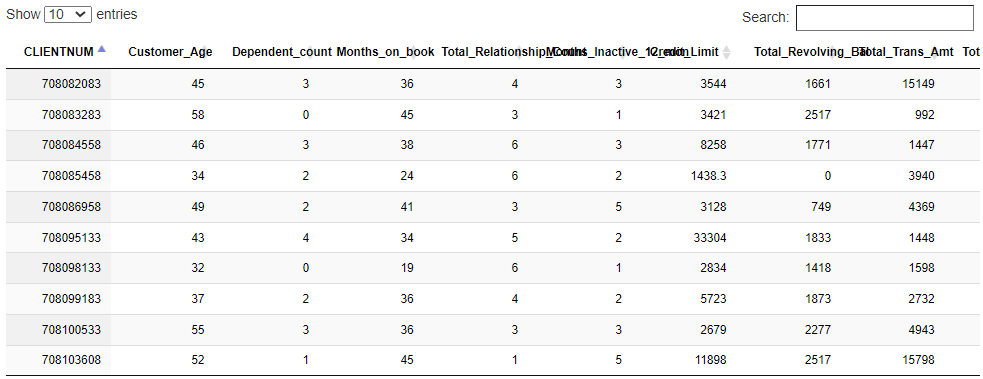
**3.2.- Apartado de columnas en String aparte de los datos Numéricos.**

**En este Apartado se hizo, una separación de campos dejando a un lado los campos de tipo string y al otro lado los de tipo númericos. A continuación, se verá las dos variables categóricas, ya sea, la variable Education\_Level y de la variable Gender.**

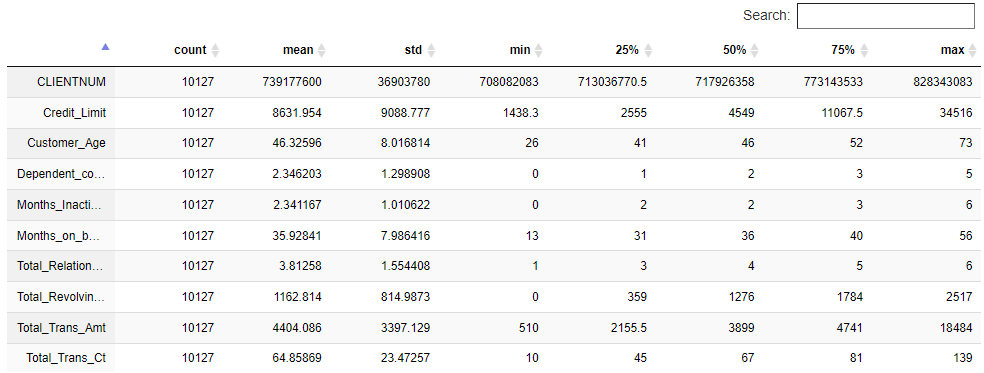
1.- VARIABLES DE TIPO OBJECT, Se separa todas las columnas pertenecientes a object o tipo String pertenecientes a los datos ya sea de tipo categóricos o no categóricos.



2.- VARIABLES DE TIPO INT Y FLOAT, Se vuelve a recorrer para separar las columnas, pero de tipo float y enteros, luego se realizará un resumen general.

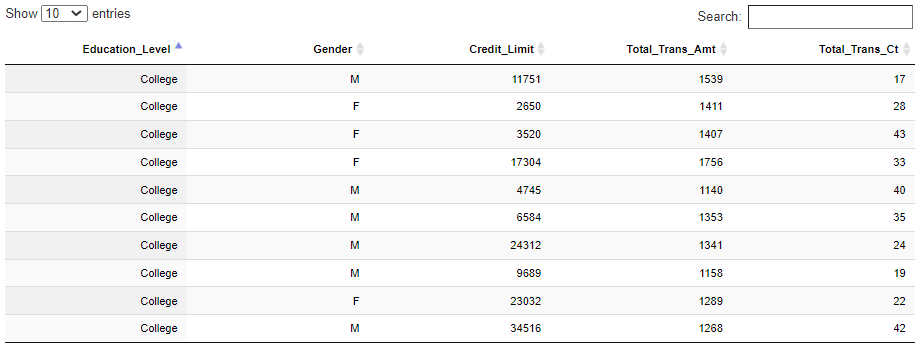


3.- RESUMEN ESTADISTICO, Se hace una estadística general de representación de los datos para conocer su distribución de forma general, comparativa a la anterior describe.



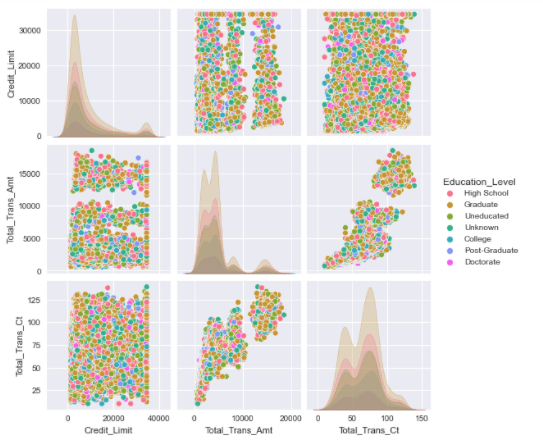
**4.- Selección de Cinco campos o Series a trabajar en la visualización.**

**En este Apartado se realizará, una exploración de los datos a presentar y se obtendrá una extracción de campos a trabajar en el informe y en la visualización de los datos.**



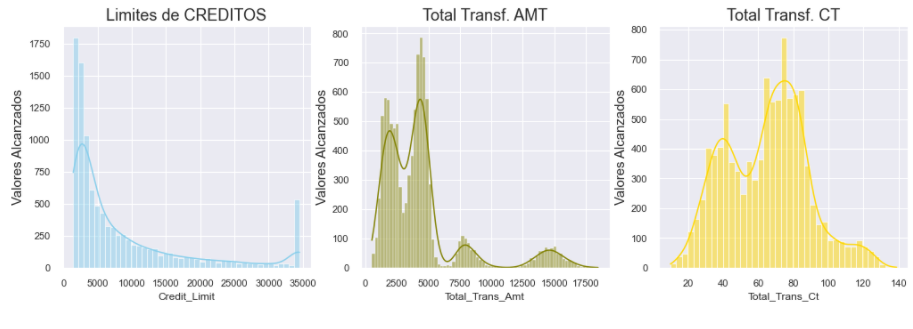
**4.1.- Gráfico General para las variables numéricas.**

**Se desplegarán los datos en diferentes gráficos, de esta manera tendremos una visualización General de nuestras variables numéricas con lo cual, de desplegarán diferentes visualizaciones, ya sea, con filtro variables categóricas como de Eduaction\_Level y también como la variable Gender.**

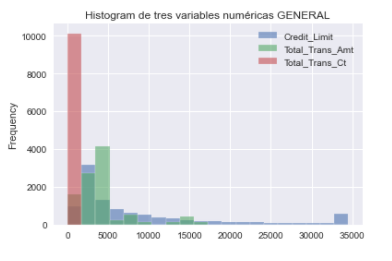


**4.2.- Desplegar el histograma de tres variables numéricas.**

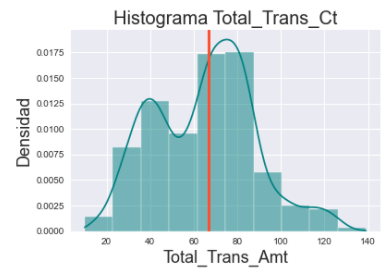
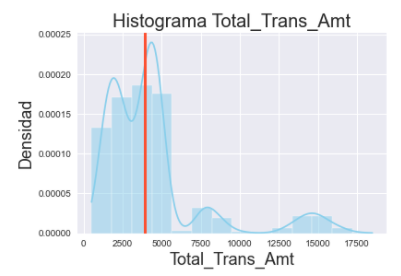
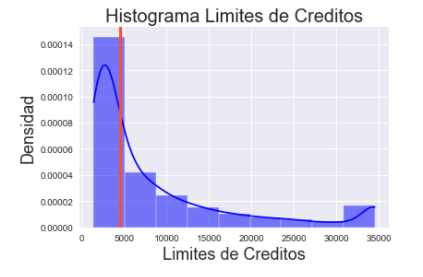
**En este apartado de desplegará diferentes formas de interpretar los datos en el histograma, para su comprensión y análisis de ello.**



1.- Visualización General Usando Seaborn para su construcción por cada una de las series numéricas usando los histogramas de seaborn, para este caso, se apartó las variables Credit\_Limit, Total\_Trans\_AMT y Total\_Trans\_Ct.

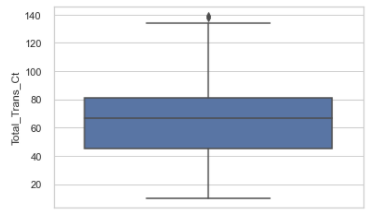
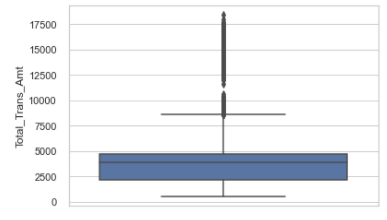
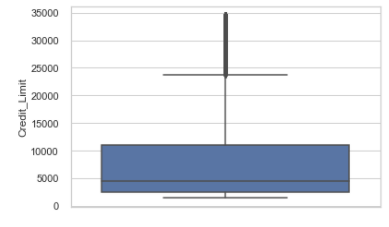
****

2.- A continuación, se desplegó, por separado las tres variables numéricas para verificar su distribución respectos a sus datos y su medida de  
 tendencia central.



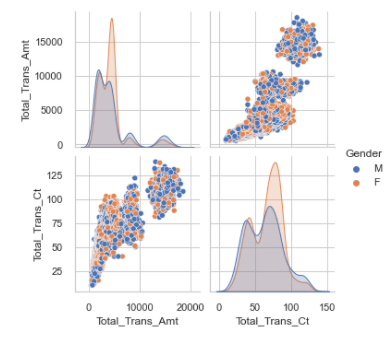
4.3.- Generación de visualización de tres variables numéricas con boxplot

Se realizó una visualización, con seaborn boxplot para representar las variables numéricas y visualizar su limites en sus datos, para este caso se apartó las variables Credit\_Limit, Total\_Trans\_AMT y Total\_Trans\_Ct.



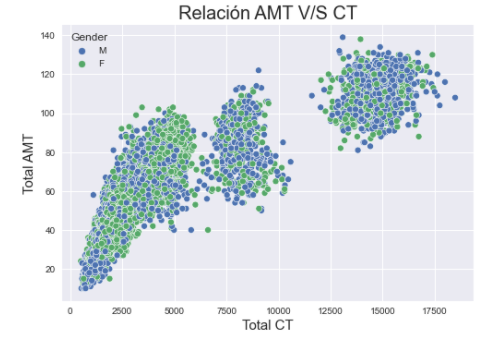
**4.4.- Gráfico General entre dos Variables numéricas (Scatter PLot).**

**Se realiza una comparativa en sus distribuciones de datos para representarlas en las categorías de Gender y Eduaction\_Level, para este caso se apartó las variables Total\_Trans\_AMT y Total\_Trans\_Ct, utilizando como herramienta de visualización el Scatter plot de seabron.**

****

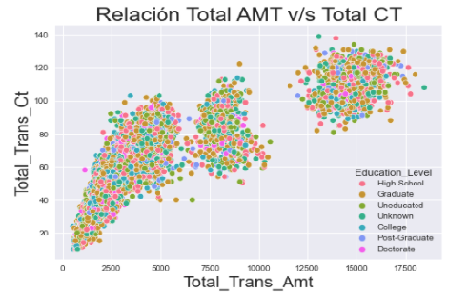
**4.5.- Representación la relación entres dos variables numéricas con scatter plot.**

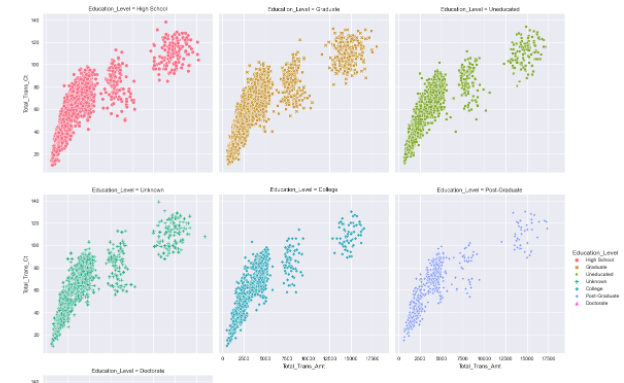
1.- Relacionando dos variables numéricas, con scatter plot, usando la librería matplotlib, para diferenciarlos con el tipo de variable y por el tipo sexo.

**

****

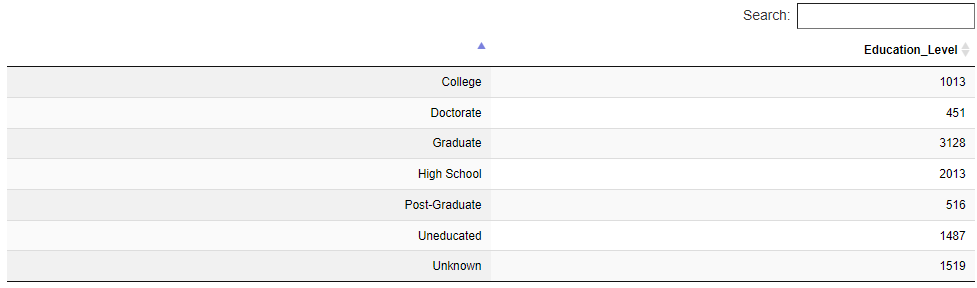
2.- Relacionando dos variables numéricas, con scatter plot, usando la librería matplotlib, para diferenciarlos con el tipo de Education\_Level.

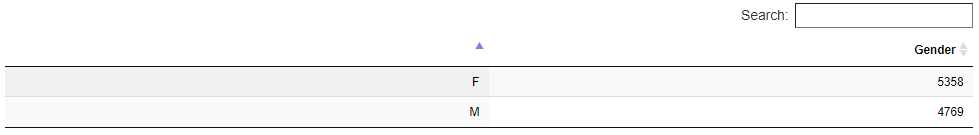




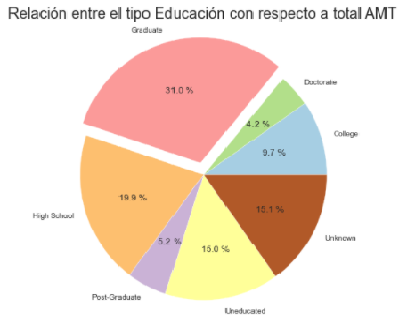
**5.- Desplegar resumen y gráficos para la visualización de PIE.**

Se desplegó unos dataframe idfbank y gdfbank para ver sus cantidades para luego optar por una gráfica de pastel para indagar en las variables categóricas y su porcentaje de datos*.*

****

****

Representación gráfica de pastel para indagar en las variables categóricas y su porcentaje de datos con las variables numéricas*.*

****



**6.- Conclusión General.**

Existen muchas formas de interpretar los datos que se está, analizando a través, de la exploración y además utilizando diferentes perspectivas de gráficos que representen una distribución de ellos, por lo tanto, se realizó varias comparativas entre grupos en diferentes flujos de datos, esto es, para comprender su límites y su expansión, incluido sus tamaños y dimensionalidad del problema que se está, estudiando en la muestra, ver como se procesó la información real de la data, el recorrido de esta, y por medios de diferentes elementos que nos permite visualizar la información que pueda transmitir el desarrollo de la solución, de la mejor y más sencilla forma de comprenderla y analizarla.