Process data for visualization

## 2. Data for visualization

Different feature treatment followed for visualization than for pre-modeling

2.1 Categorical features visualization

- o Changed all feature dtype to object.

- o Changed all but score categorical variables into dummies using a loop. Merged all new columns into one DF (label encoding maintains a column, labeling the categorical values as numbers).

- o Ploted all the categorical features without differenciating loan or not loan(pie plot/subplot)(graphic 1).

Graphic 1
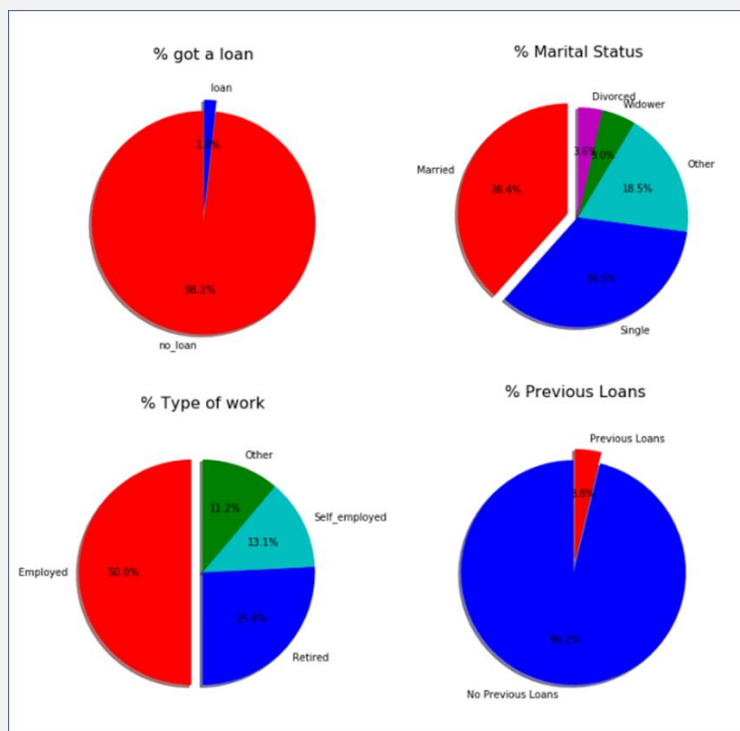
All the categorical feature plots can be found in the notebook

Insights:

- 1.8% of the population studied have contracted a personal loan.
- 54.3% of the population studied have levesl of studies 1 or 2 (no studies or up to basic studies).
- 50% employed and 25% retired.
- 38.4% married and 34.5% single.
- 71.5% belong to mass market segment.

- 14.7% have current mortgages.
- 3.8% have had previous personal loans.

2.1.1 Categorical features visualization (bar plot)

Same analysis than before but differentiating population with and without a personal loan (Graphic 2).

o Stacked bar graphs.

o Created a new column counting the clients for each group.

o Created a percentage for each group category.

o Dropped count to keep only the percentage column.

o Added the percentage column.

o Transposed the df to create the stacked bar plot.

o Transposed to show the detail table underneath and renamed the legend.

o Rounded the values in the table under bar plot.

Graphic 2
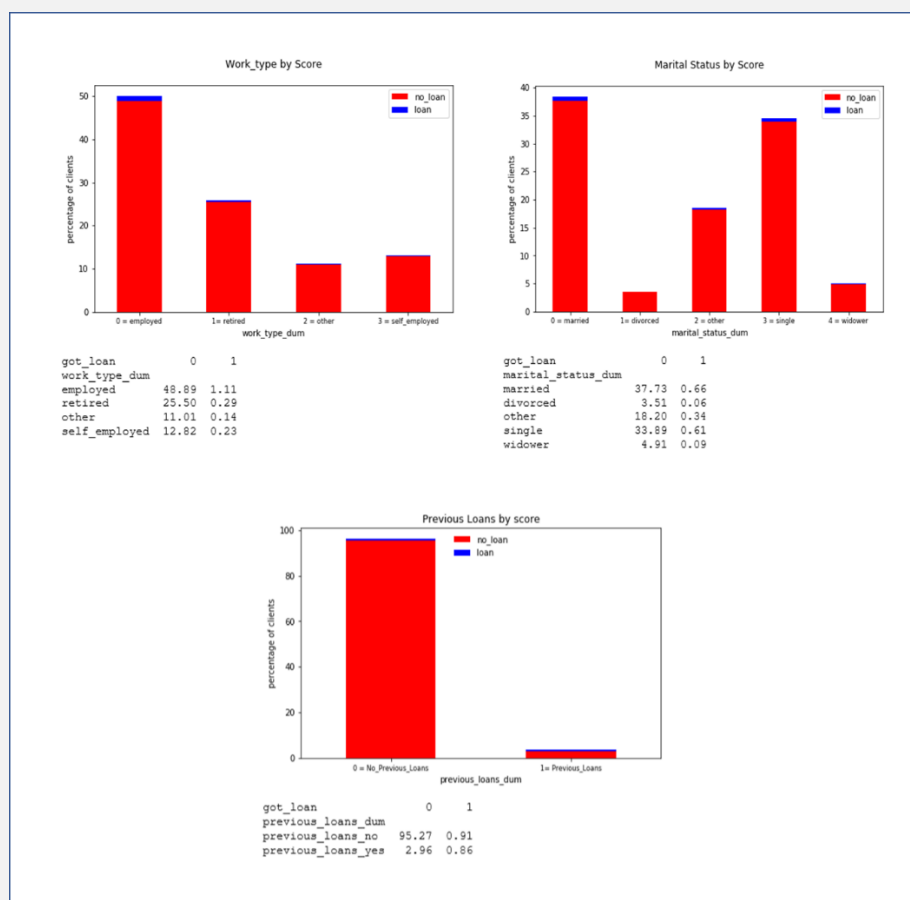
*All the categorical feature plots can be found in the notebook*

Insights:

- Out of the total population, 1.08% were men and contracted a loan vs 0.68% women.

- Out of the total population, 0.70% have study level 2 and contracted a loan followed by study level 3 with 0.4% (study level 1- 0.31%).

- Out of the total population, 1.11% are population employed and contracted a loan followed by retired 0.29%.

- Out of the total population, 0.66% are married and contracted a loan followed by 0.61% single.

- Out of the total population, 1.37% belong to the mass market segment and contracted a loan followed by self-employed with 0.24%.

- 1.45% with no mortage contracted a personal loan vs 0.32% with mortgage.

- 0.91% with previous personal loans contracted a loan vs 0.86% without previous personal loans.

2.1.2 Categorical features visualization (base 100)

Same as the previos analysis, but taking into account each population weight within the sample.

Each population with and without a loan was analysed to know the percentage in base 100 of each group (Graphic 3).

Graphic 3
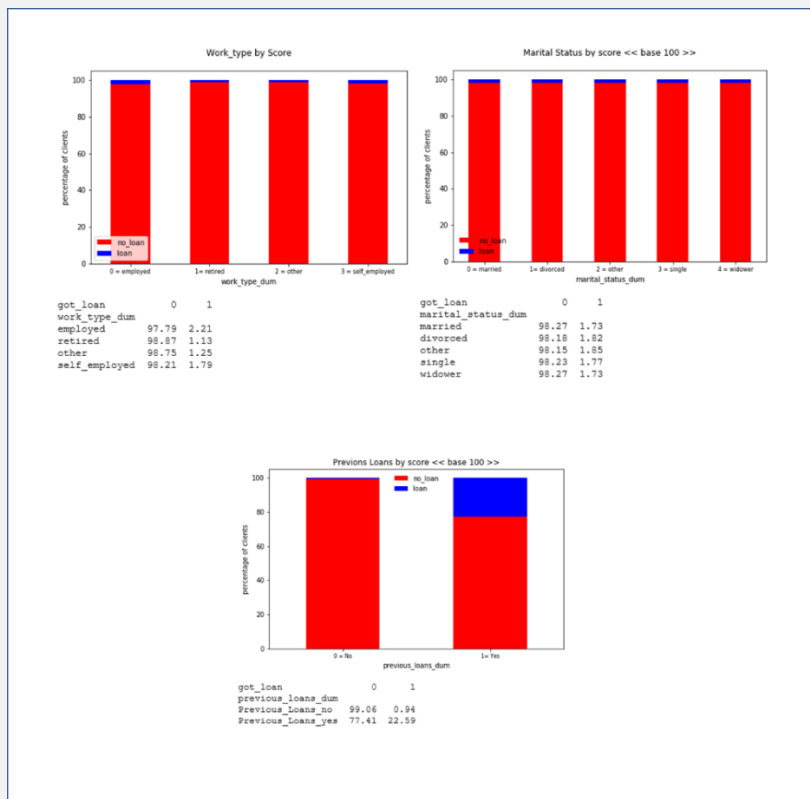
Insights:

- 1.86% of total men contracted a loan vs. 1.65% women.

- 2.16% of the population with study level 2, contracted a loan followed by study level 6 with 1.84% (study level 1 - 1.43%).

- 2.21% of the population employed contracted a loan followed by self-employed 1.79%.

- 1.82% of divorced population contracted a loan followed by 1.77% singles.

- 1.98% of the population belong to the self-employed segment contracted a loan followed by mass market with 1.92%.

- 2.19% with a mortage, contracted a personal loan vs. 1.70% without a mortgage .

- 22.59% with a previous personal loan contracted a loan vs. 0.94% without a previous personal loan.

2.2 Numercial visualization

Kdeplot from *seaborn* will be used. Kernel Density Estimate was used for visualizing of the Probability Density of a continuous variable.

- o Set the title and size.
- o Defined the feature to plot together with got loan values 0 and 1.Defined different color and width for got_loan=1 .
- o Inserted graph for 'aveg_amount_invest_funds_0m' and 'amount_trfr_0m' features to better distinguish values 0 and 1.
- o Set the frame and the legend positioning.
- o Names the plot.

Graphic 4
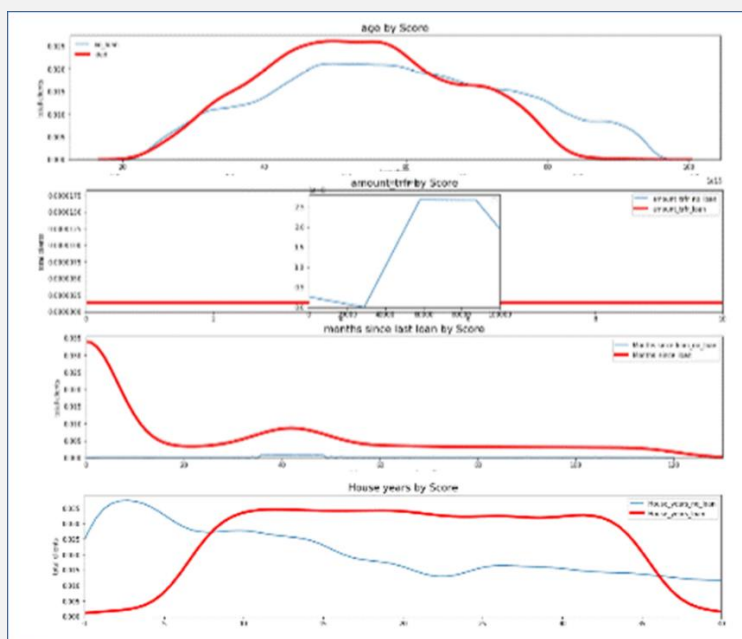
Insights:

- It seems that between 30 and 65, there is a higher probability to contract a personal loan.

- Up to 20 months since your last lown paymet and between 30 - 60 months after, there is a higher probability to contract a new personal loan.

- When the house is between 10-35 years old, the probability of contracting a personal loan increases.

- The probability of getting a personal loan is higher if the person spends between 100€ and 1.500€ on his/her ccard.

- Clients with no significant expenses on their personal credit cards, have more probability of contracting a personal loan. (Most likely due to risk aversion from lenders).

- Clients with more than 30k euros in funds have a higher probability of contracting a loan (tend to happen to avoid customer's descapitalization).

- Clients with 8-year-old vehicles have the highest probabiliy of contracting a personal loan to purchase a new car.

- Digital clients who check bank details through their Bank App have a higher probability of contrating a loan (most likely due to the online offer for these types of loans).

2.3 Customers Geolocalization

The objective is to see if there is a concentration either of total clients or clients who contracted a loan in a specific area (Graphic 5).

'Folium.Map' and 'Folium.Choropleth' were used to define 2 separates groups and folium 'LayerControl' used to add them into the same html graph.

- Downloaded the map 4 files to local folder and read de 'shp' file.

- Selected the column, with a significant value, from 'my file,change' column name to be used as key with the previous file 'polygon_pc' and grouped the clients by postal code.

- The map showed 2 separate groups. Total clients and clients who contracted loans.

- Weighed no loan clients by province.

- 2 separate groups were plotted. Choropleth is used for binding the data between Pandas DataFrames and JSON geometries. Sequential color schemes are built-into the library and can be passed to visualize different combinations. Creates a LayerControl object to be added on a folium map.

- Saved Info in html file.

Graphic 5

Clients with loan by Province
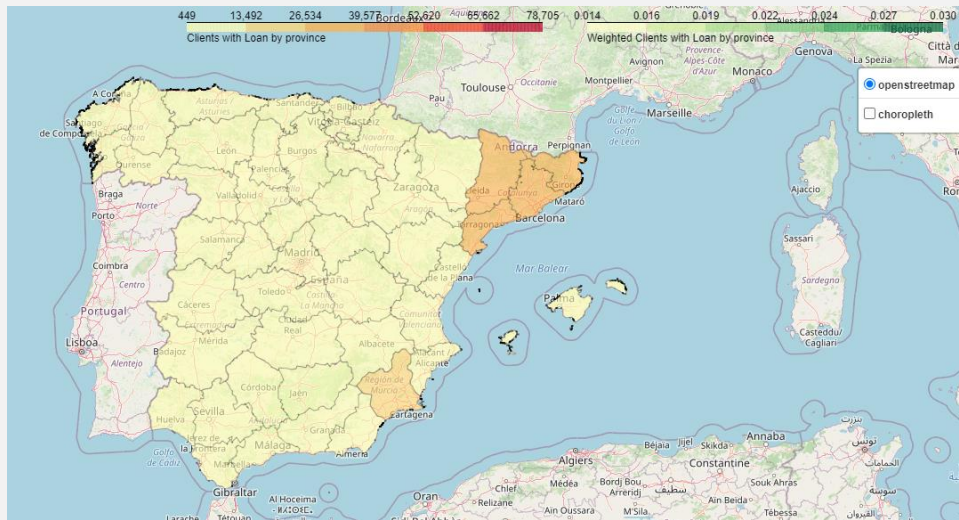
Weighted Clients with Personal Loan by Province

Insight:

- Comparing to the total number of clients in the population, the results obatained show that there are more clients from the Cataluña and Malaga areas.

- However, the success rate is the highest in the Northwest and Cataluña áreas (as shown in the previous graphic).

[1]Folium: https://python-visualization.github.io/folium/quickstart.html

Although the initial information included in the dataset is real, in order to present this work and to comply with current regulations, all data has been anonymized.