Data Glacier

Team Members:

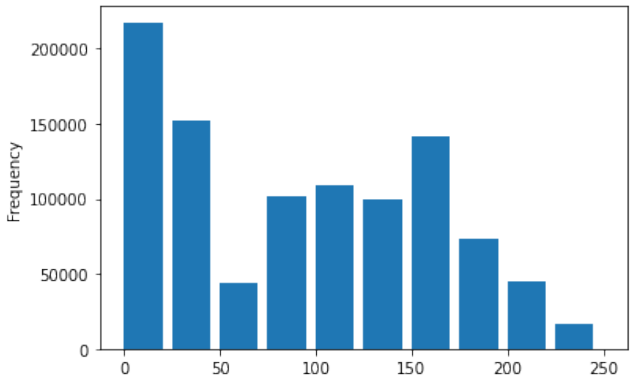
- Ray Ng
  - rayng2018@gmail.com
  - Canada
  - University of British Columbia
  - Specialization: Data Science
- Rita Uzoka
  - rita.uzoka@yahoo.com
  - United Kingdom
  - Sheffield Hallam University
  - Specialization: Data Science
- Fatemeh Bagheri
  - f.bagheri13@gmail.com
  - France
  - Université Jean Monnet St Etienne - Université de Lyon
  - Your Specialization: Data Science

Problem: Customer Segmentation - XYZ bank wants to roll out Christmas offers to their customers. But the bank does not want to roll out the same offer to all customers, instead they want to roll out personalized offers to particular sets of customers. If they manually start understanding the category of customer then this will not be efficient and also they will not be able to uncover the hidden pattern in the data (pattern which groups certain kinds of customer in one category). Bank approached ABC analytics company to solve their problem. Bank also shared information with ABC analytics that they don't want more than 5 groups as this will be inefficient for their campaign.

GitHub Repo Link:
https://github.com/faba13/VC.git

EDA Points:

- The majority of customers are located in Spain
- Province Madrid has both the highest share of customers in the data, and also the highest renta (household income)
- Most of the customers' belong to the sex group(V)
- Most of the customers' are those that have never worked for the bank and currently do not work for the bank.
- Most of the customers reside in the same country as the bank.
- The most commonly registered account types are Current Account, Particular Account and Direct Debit Account, while the least commonly registered types are Saving Account, Guarantees, and Derivada Account.
- Many customers became the first holder of a contract after 2011 compared to years before.
- The average customer seniority is 97 months, but the histogram below shows that there are also a large portion of customers with seniority below 50 months.



Exploratory Data Analysis or EDA is used to take insights from the data. Data Scientists and Analysts try to find different patterns, relations, and anomalies in the data using some statistical graphs and other visualization techniques. Following things are part of EDA :

1. Get maximum insights from a data set
2. Uncover underlying structure
3. Extract important variables from the dataset
4. Detect outliers and anomalies(if any)
5. Test underlying assumptions
6. Determine the optimal factor settings

For this project, some steps are done in previous weeks.

To explain more for Missing value imputation, We will now check for missing values in our dataset. In case there are any missing entries, we will impute them with appropriate values (mode in case of categorical feature, and median or mean in case of numerical feature). We will use the isnull() function for this purpose.