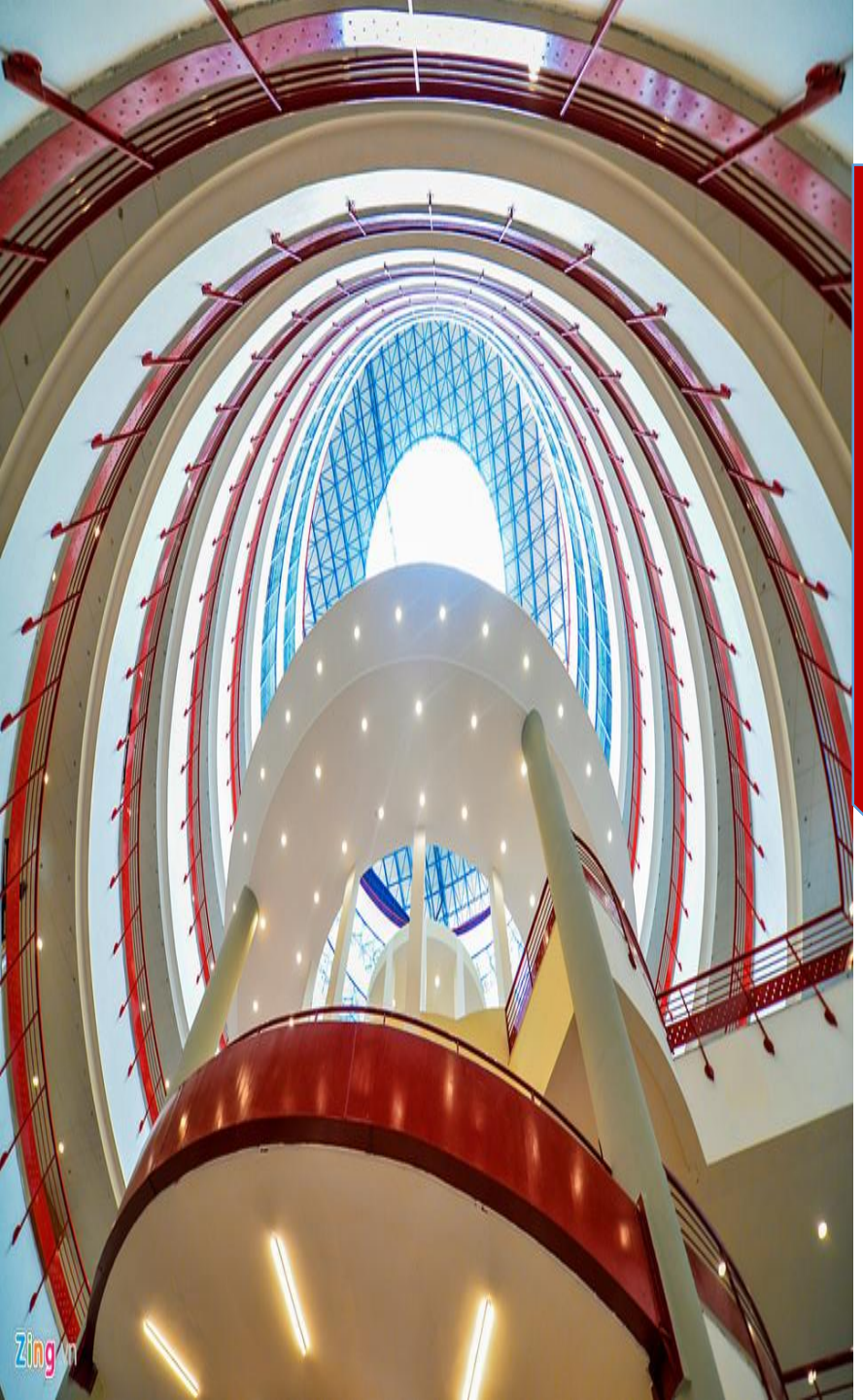


DATA PREPARATION AND VISUALIZATION

Department of Mathematical Economics

National Economics University
https://www.neu.edu.vn/

Data Exploration

A First Sight to Our Dataset

- Df.head()
- Df.info(): a quick overview of the number of records that are filled with values for each column
- Df.dtypes: take a look at the data types of the columns
- Df.unique(): see the distinct values of each attribute
- Df.shape: the total number of records and the total number of columns
- Df.isnull().sum(): the total of missing values in each attribute

SUMMARIZATION

- Summarization gives us summary for summary of each feature, giving us the basic statistics properties
- df.describe(): summary statistics
- Df.sum(): sum of values
- Df.cumsum(): cumulative sum of values
- Df.min()/df.max(): minimum/maximum values
- Df.idexmin / df.idexmax(): minimum/maximum index of value
- Df.mean(): mean of values
- Df.median(): median of values
- Df.value_counts()

Histogram

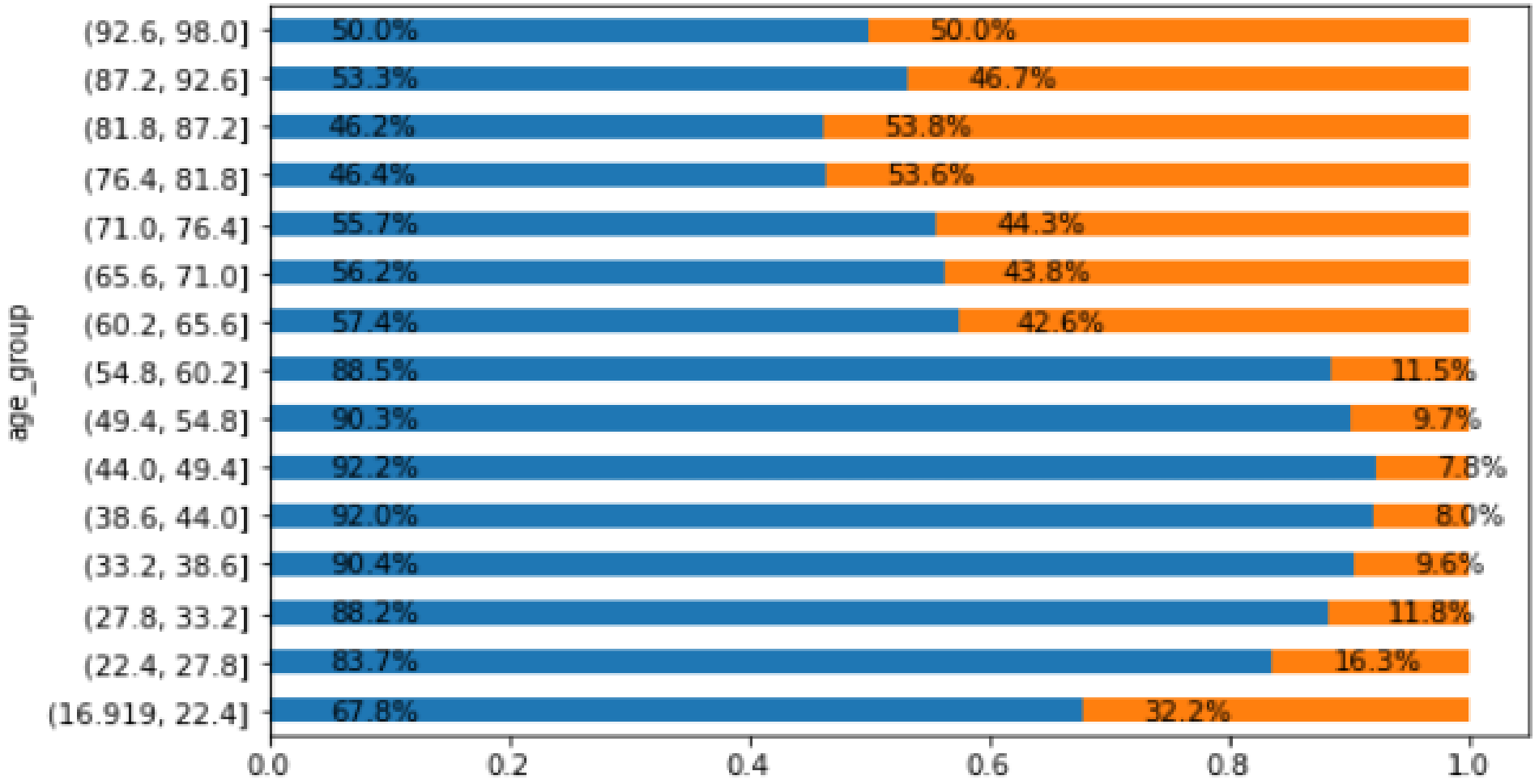
- Df.hist(): get information about the shape of the distribution and the skewness of the feature.
- df.boxplot(): check outliers

PairPlot

- Using seaborn package
- Sns.pairplot()

Example: Conversions versus non-conversions by Age

Age

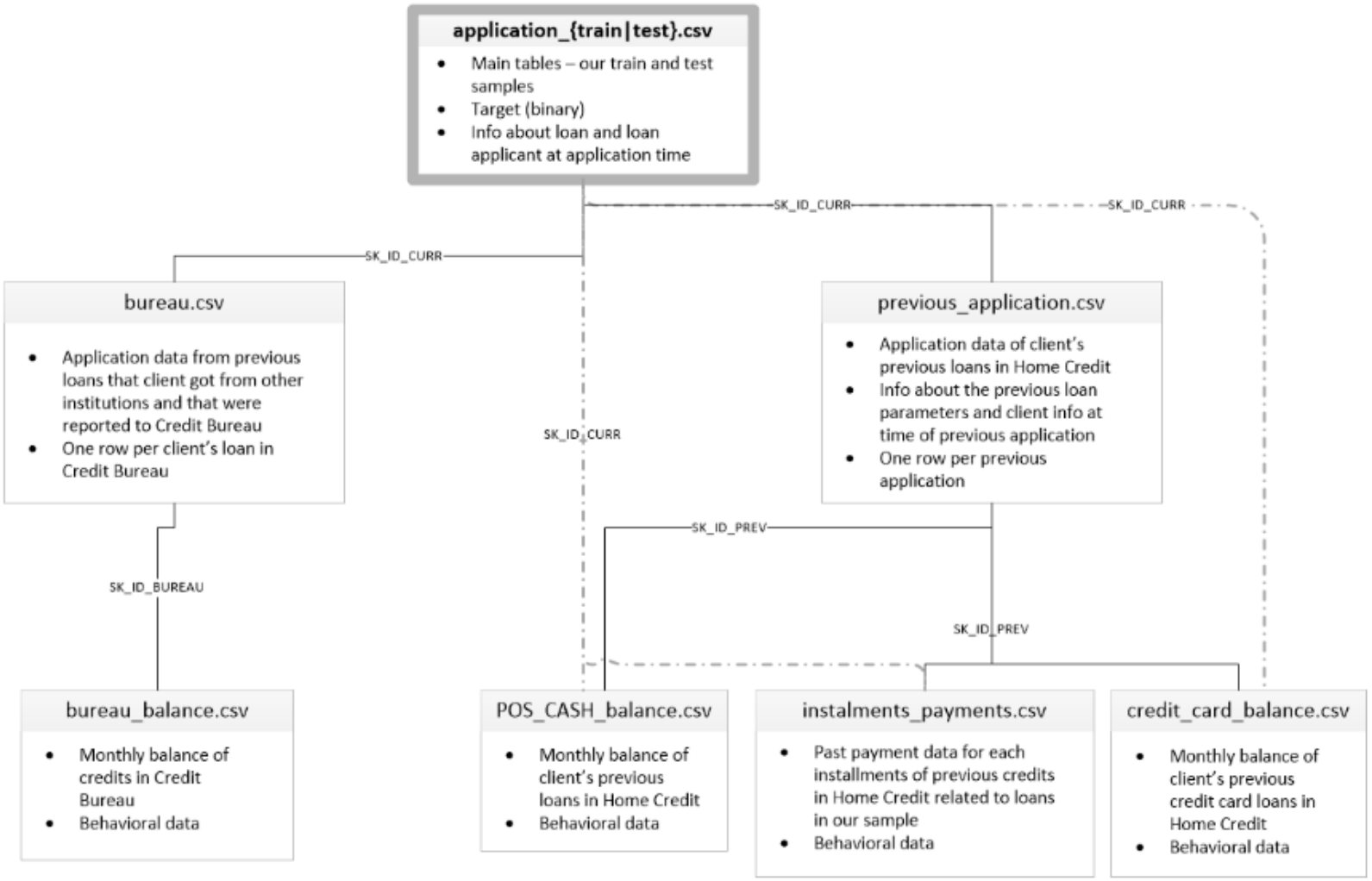


Practical suggestions

Some general, practical suggestions for performing EDA:

- Focus on data visualization whenever you can. Charts are more readable than tables of numbers.
- Use the pair plots for measuring the correlation between the features and between each feature and the target variable (for regression problems)
- Stacked histograms can be a very useful representation for classification problems
- Use Pandas profiling libraries if you want to save time
- A good EDA can give a very huge value to a project. Several insights can be extracted using the correct visualizations even before using any model
- Take your time to perform EDA. It's the first approach to a machine learning project and, if you apply it quickly and not focusing on the important things, this may affect the success of the entire project
- Don't forget to extract the information behind data, that is the purpose of data science
- Write proper storytelling that shows the analysis path you have followed. Your audience will appreciate it and understand the information much better.

Project – Credit Risk Analysis



Data Descriptions

- **application_{train|test}.csv**
  - This is the main table, broken into two files for Train (with TARGET) and Test (without TARGET).
  - Static data for all applications. One row represents one loan in our data sample.
- **bureau.csv**
  - All client's previous credits provided by other financial institutions that were reported to Credit Bureau (for clients who have a loan in our sample).
  - For every loan in our sample, there are as many rows as number of credits the client had in Credit Bureau before the application date.
- **bureau_balance.csv**
  - Monthly balances of previous credits in Credit Bureau.
  - This table has one row for each month of history of every previous credit reported to Credit Bureau – i.e the table has (#loans in sample * # of relative previous credits * # of months where we have some history observable for the previous credits) rows.

Data Descriptions

- **POS_CASH_balance.csv**
  - Monthly balance snapshots of previous POS (point of sales) and cash loans that the applicant had with Home Credit.
  - This table has one row for each month of history of every previous credit in Home Credit (consumer credit and cash loans) related to loans in our sample – i.e. the table has (#loans in sample * # of relative previous credits * # of months in which we have some history observable for the previous credits) rows.
- **credit_card_balance.csv**
  - Monthly balance snapshots of previous credit cards that the applicant has with Home Credit.
  - This table has one row for each month of history of every previous credit in Home Credit (consumer credit and cash loans) related to loans in our sample – i.e. the table has (#loans in sample * # of relative previous credit cards * # of months where we have some history observable for the previous credit card) rows.
- **previous_application.csv**
  - All previous applications for Home Credit loans of clients who have loans in our sample.
  - There is one row for each previous application related to loans in our data sample.
- **installments_payments.csv**
  - Repayment history for the previously disbursed credits in Home Credit related to the loans in our sample.
  - There is a) one row for every payment that was made plus b) one row each for missed payment.
  - One row is equivalent to one payment of one installment OR one installment corresponding to one payment of one previous Home Credit credit related to loans in our sample.
- **HomeCredit_columns_description.csv**
  - This file contains descriptions for the columns in the various data files.