Open in Colab

Visualizing Data for Classification

In the previous lab, you explored the automotive price dataset to understand the relationships for a regression problem. In this lab you will explore the German bank credit dataset to understand the relationships for a **classification** problem. The difference being, that in classification problems the label is a categorical variable.

In other labs you will use what you learn through visualization to create a solution that predicts the customers with bad credit. For now, the focus of this lab is on visually exploring the data to determine which features may be useful in predicting customer's bad credit.

Visualization for classification problems shares much in common with visualization for regression problems. Colinear features should be identified so they can be eliminated or otherwise dealt with. However, for classification problems you are looking for features that help **separate the label categories**. Separation is achieved when there are distinctive feature values for each label category. Good separation results in low classification error rate.

Load and prepare the data set

Prepare data to a manageable format

Processing bson files

source:

- Kaggle
- Access and process nested objects, arrays or JSON

```
## Unziping (file on linux)
#unzip pantapa_api_development.zip
```