TSAR Project Assignment Part1 (10 points)

Exploratory Data Analysis with ggplot2

gwendolin.wilke@fhnw.ch

The Lending Club Data Set

For the first project part we use a version of the «Lending Club Loan Data». The Lending Club (LC) operates an online peer-to peer credit marketplace in the US.

> *"Peer-to-peer lending, also abbreviated as P2P lending, is the practice of lending money to individuals or businesses through online services that match lenders with borrowers. Peer-to-peer lending companies often offer their services online, and attempt to operate with lower overhead and provide their services more cheaply than traditional financial institutions. As a result, lenders can earn higher returns compared to savings and investment products offered by banks, while borrowers can borrow money at lower interest rates, even after the P2P lending company has taken a fee for providing the match-making platform and credit checking the borrower."* (Wikipedia, Peer-to-peer lending)

> *"Lending club is a financial services company headquartered in San Francisco, California. […] At its height, LendingClub was the world's largest peer-to-peer lending platform. The company reported that $15.98 billion in loans had been originated through its platform up to December 31, 2015."* (Wikipedia, Lending Club)

The data set published on Moodle contains real anonymized data describing personal loans issued through the Lending Club website. The data set contains historical data (i.e. loans from several years back). The original data set can be found on Kaggle. We use a modified version of this original data set.

The **data dictionary** provided on Moodle provides a description of the semantics of the included features.

Preliminaries

- Work through this quarto tutorial.
- Create a folder on your computer for TSAR project part 1.
- Download from Moodle the data set `LCdata.csv`, and store it in your project folder.
- Open RStudio, navigate to your project folder and set it as your working directory.
- Look up your group-ID in the groups list on Moodle.
- Create a new quarto document for pdf output using the Knitr Engine, and call it "`P1-<your group-ID>.qmd`". (E.g., "P1-12.qmd").
- Copy the following code in your quarto document to install and load the required packages:

```
# install the required packages if needed
if (!require("data.table")) install.packages("data.table")
if (!require("dplyr")) install.packages("dplyr")

# load them
library(data.table)
library(dplyr)
```

- Copy/paste the following code block, but insert your group-ID in the place of `<group-ID>` and execute it:

```
# Random sample your group's data set from the original LCdata data set
LCdata <- fread("LCdata.csv")
set.seed(<group-ID>)
myLCdata <- LCdata %>% sample(5) %>% slice_sample(prop = .5)
fwrite(myLCdata, file = "myLCdata.csv")
```

Based on your group-ID, the code random samples 5 out of 19 attributes of the data set `LCdata`, as well as 50% of the rows.

There are 11'628 possibilities to choose 5 out of 19 attributes, so the chance that two groups have exactly the same data set is quite small. You are encouraged to discuss with your group mates and classmates about the project as much as possible! Also go and find ideas on the web. The goal is that you learn as much as possible! You can expect similar questions in the exam as well.

In the code block above, we have written out your group's data set `myLCdata` using `fwrite()`. After you have run this code block once, you can out-comment it and load `myLCdata` again using `fread()`:

```
# # Random sample your group's data set from the original LCdata data set
# LCdata <- fread("LCdata.csv")
# set.seed(1)
# myLCdata <- LCdata %>% sample(5) %>% slice_sample(prop = .5)
# fwrite(myLCdata, file = "myLCdata.csv")
myLCdata <- fread("myLCdata.csv")
```

We do this to avoid having Quarto loading and random-sampling the original data set `LCData.csv` every time you render your pdf. Since `LCData.csv` is quite large, it takes much longer to load `LCData.csv`than loading the small sample `myLCdata`, which can get annoying.

Deliverables

As for all project parts, your deliverable consists of 2 parts:

- a **pdf** report that you generate using quarto,
- the corresponding working **.qmd** file that you used to render your report.

The **pdf report** should hold the solutions to the tasks listed below, including

- the code you used to solve the task,
- the code execution output, whenever it is relevant for your answer/solution (e.g., the plots you generated or the `skim()` output),
- the answers/interpretations asked for in natural language text.

The report must have a professional writing style, structure and layout. This includes, e.g.,

- title, author, date, and one subsection for each task;
- figures are numbered, they have captions describing them, and they are discussed in the text;
- warnings and messages of code executions are suppressed, except they are relevant for your answer.

Task 1. Business Understanding (1 Points)

Use `dplyr::glimpse()` and `str()` to get a first impression of your data set. Download the data dictionary from Moodle, find the attributes that are present in your data set `myLCdata` and read the corresponding discriptions. Try to understand their meaning in the context of peer-to-peer loans.

Task 2. Data Understanding Based on Summary Statistics (3 points)

Use `summary()` and `skimr::skim()` to review the main statistical metrics of your attributes. What do they tell you about the attributes? Interpret them not only statistically, but also in the domain context.

Examples:

- You may realize that a lot of NAs are present in one of the attributes. This is important for the data preparation step (which we will not cover in this course - you learned about it in the Machine Learning lecture.)
- For a numerical attribute, you might realize that the maximum seems too high considering the variable semantics (e.g. a monthly income of 1 million $ seems too high). This hints towards the presence of outliers due to erroneous inputs.
- For a numerical attribute, you might realize that the maximum is very high, while the median is very low. This hints towards a heavily right skewed distribution. It means that most customers have a very low value in this attribute. Then you can ponder how to interpret it considering variable semantics.
- For a categorical attribute, you might realize that the value frequencies are *unbalanced*. (E.g., 95% of the loan applications are granted, and only 5% are dismissed). This is a *data bias* that may lead to biased predictions of a classifier (which we will not cover here - you learned about it in the machine learning lecture).

Task 3. Univariate Exploratory Data Analysis (EDA) with ggplot2 (3 points)

Choose at least one suitable diagram type (e.g. a histogram, a scatterplot) for each of the 5 attributes in your data set and visualize it using ggplot2. Give your plots a professional look (e.g., add a title, make the axes labels readable, add the measurement units if available).

Interpret each of your plots:

- What information can you read off the diagram type you chose (e.g. in a histogram you can read off range, distribution, etc.)?
- What does it tell you about the attribute in the context of the attribute semantic and domain? Does it confirm your first assumptions about the attribute from task 2?

Task 4. Bivariate EDA with ggplot2 (3 points)

In task 3, you looked at each variable separately. Now the goal is to see if there are some obvious correlations or pattens when looking at pairs of variables.

- To get a first impression, (install if necessary and) load the `Ggally` package and create a pairs plot of your data set using `ggpairs()`.
- Select at least two different types of 2-dimensional subplots that `ggpairs()` is showing you. Select subplots that provide you with additional insight/knowledge about the data - something that was not obvious before. Explain what this is.
- Recreate the two plots, and turn them into professional looking stand-alone plots using `ggplot2`.(If you don't know which type of `geom` it is or how to use it, consult the ggplot2 cheatsheet on Moodle.)
- Choose at least one of these 2 plots and add a 3rd variable (e.g. using the color channel) to it. Argue why this third variable is of interest in this context. What additional insight/knowledge does it give you about the data that was not obvious before?