Hackathon - IE6600 - Sec 03 -Group5

Ankita Yadav, Chris Bolsinger & Zeeshan Ali

2/18/2022

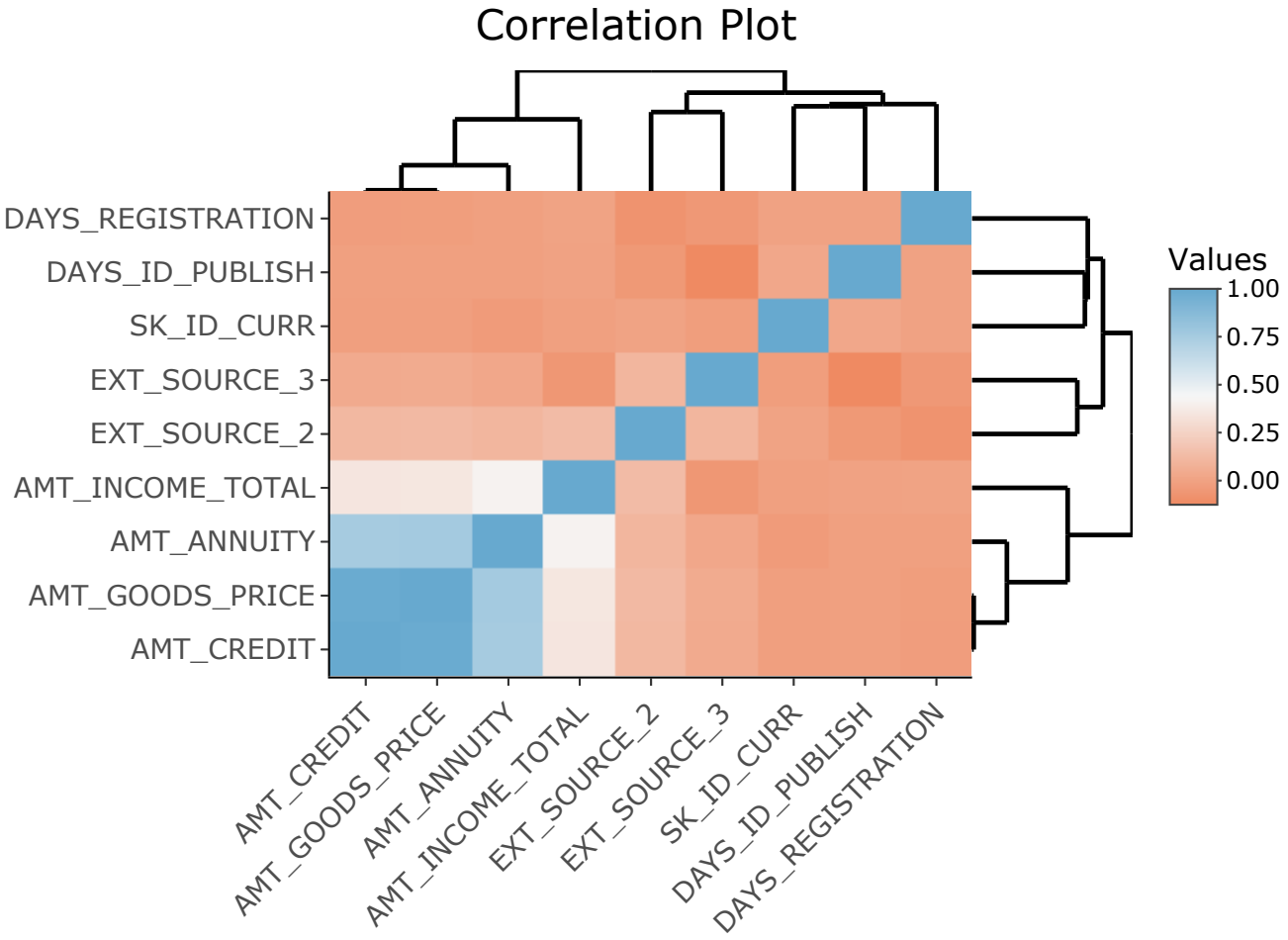
Introduction and Problem statement

Due to the weak or non-existent credit history of the clients, lending providers find it difficult to provide loans to their customers. As a result, some consumers take advantage of this by becoming defaulters. This project tries to find patterns that suggest if a client is having problems paying their installments, which can be utilized to take actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, and so on. This ensures that consumers who are capable of repaying the loan are not refused. The purpose of this case study is to identify such applications using EDA. In other words, this analysis and visualization will help the organisations to identify the reasons (or driver variables) that lead to loan default, i.e. the variables that are significant predictors of default. This knowledge can be used to the company's portfolio and risk assessment.

This data set was chosen from kaggle. After running the summary statistics, it was known that it has 8602 observations and 123 variables which is large enough for performing EDA. This data set was chosen as it had categorical values like Gender(F or M), Type of loan (Cash or Revolving), Educational level (Higher Level, Academic, etc.), Target (0 or 1 explaining the difficulty level). It also includes numerical values like Credit Amount, Income Amount, Amount Annuity, etc. All these values proves to be of great use for our analysis and visualizations to draw insights and figure patterns and reasons for defaulters as well as clients repaying on time. This data set had some outlines such as NA values, hence, data wrangling procedures were performed on the data set during the initial steps.

The project tries to determine whether a client's gender, education level, marital status, occupation type, Contract type (loan type), Application day of week affects the payment difficulty of these clients. In order to achieve this, the group will be running EDA on these variables and answer some business questions.

Q1. What is the correlation between different variables? Which variables are strongly correlated?

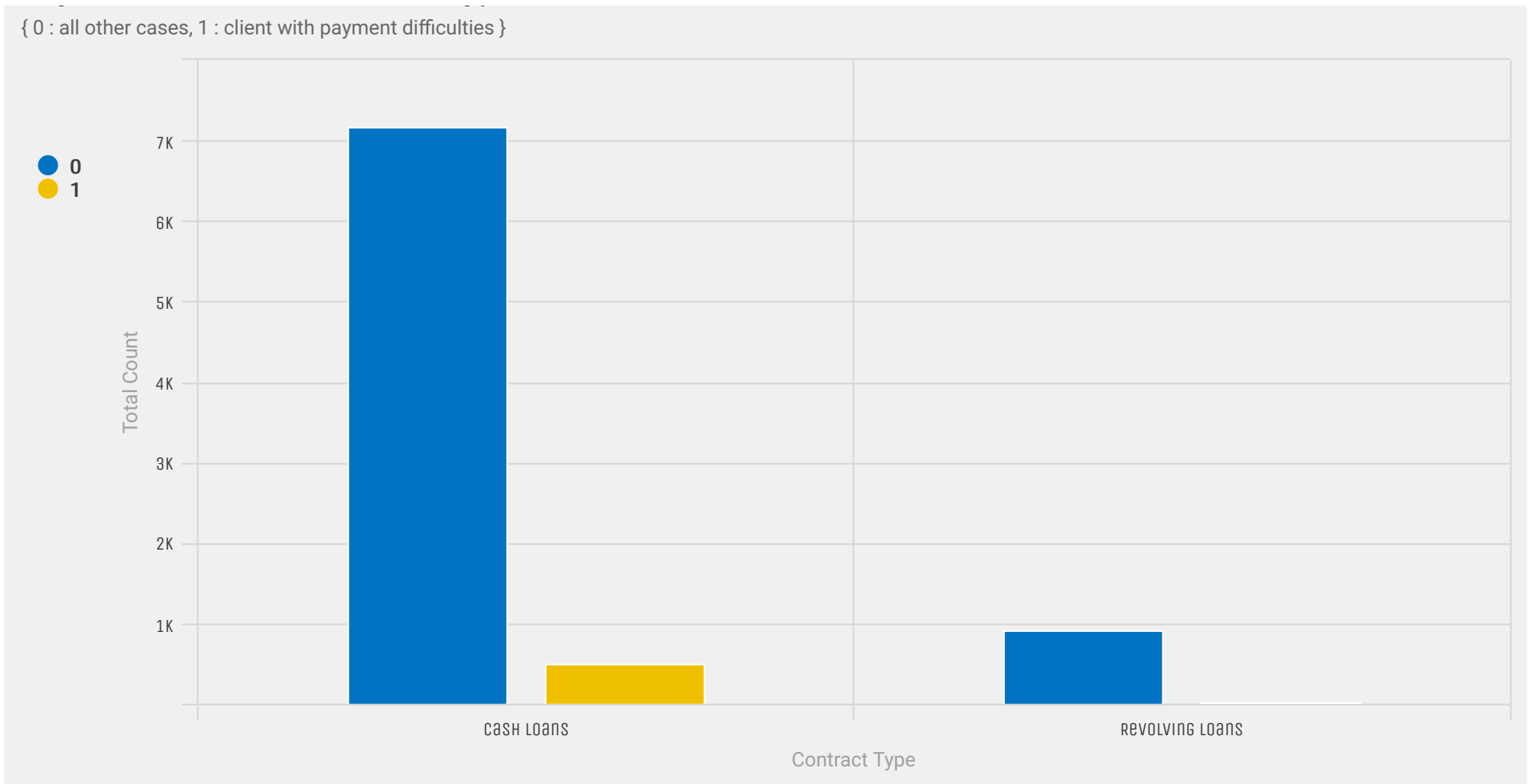


Conclusion:

The higher the correlation values the higher the variables are related to each other. In the plot above, each element on the principal diagonal of a correlation matrix is the correlation of a random variable with itself, which always equals 1. AMT_CREDIT has a correlation value of 0.98 with AMT_GOODS_PRICE and a correlation value of 0.75 with AMT_ANNUITY.

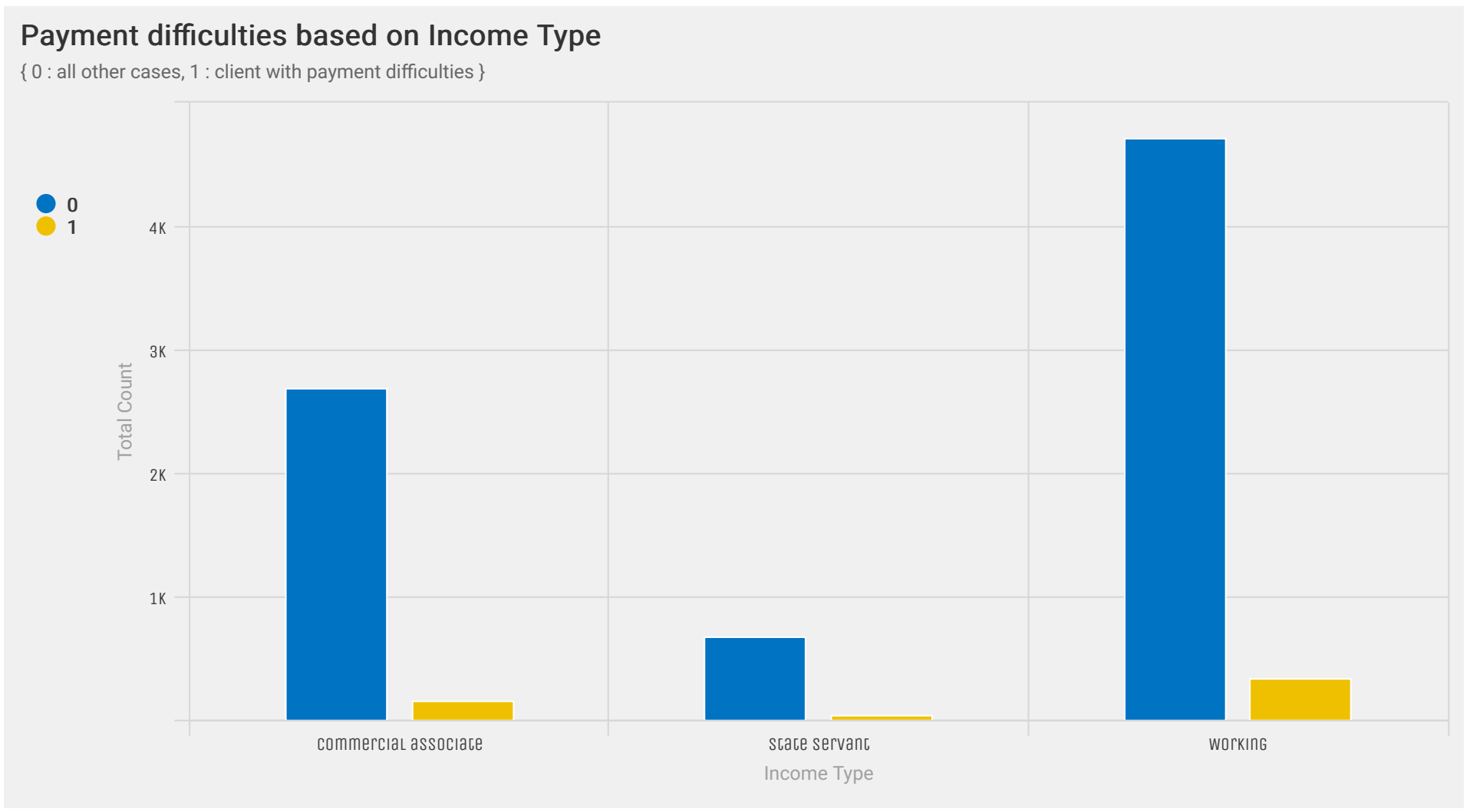
Q2. Does factors like Contract type,Income type, Application day of week, Organisation type influence the amount payment difficulties of the clients?

Payment difficulties based on Type of Contracts



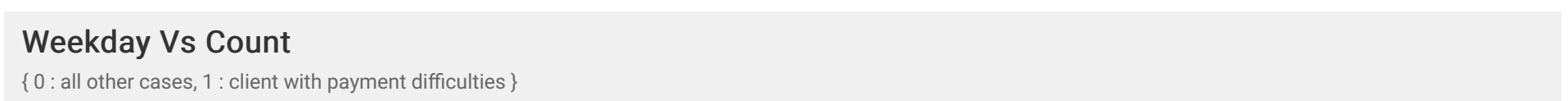
Conclusion:

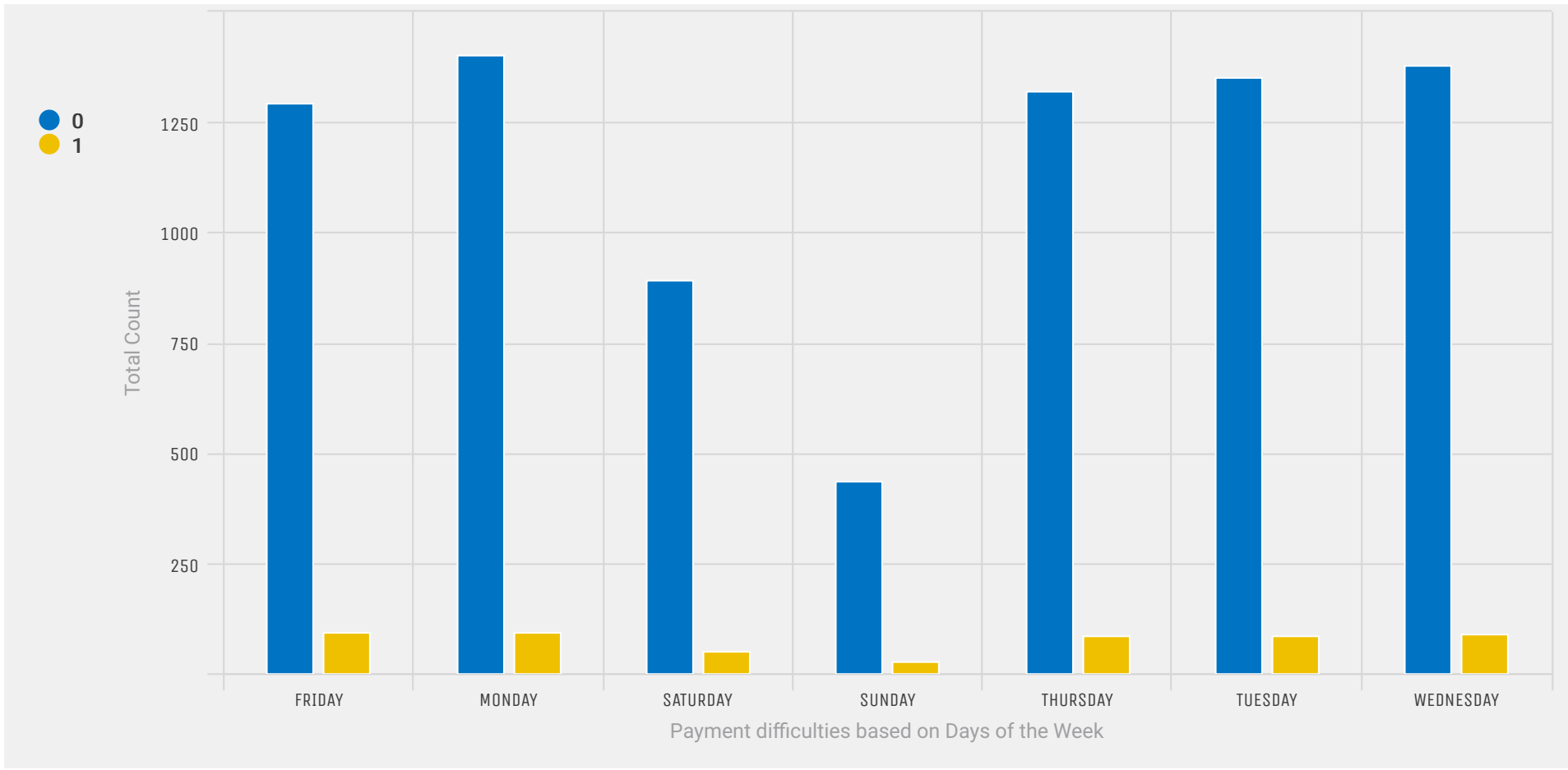
The bar chart here displays the type of contract (loan type) that people were more interested in. The proportion of people opting out for Cash loans and paying the amount back is more than the people opting for revolving loans.



Conclusion:

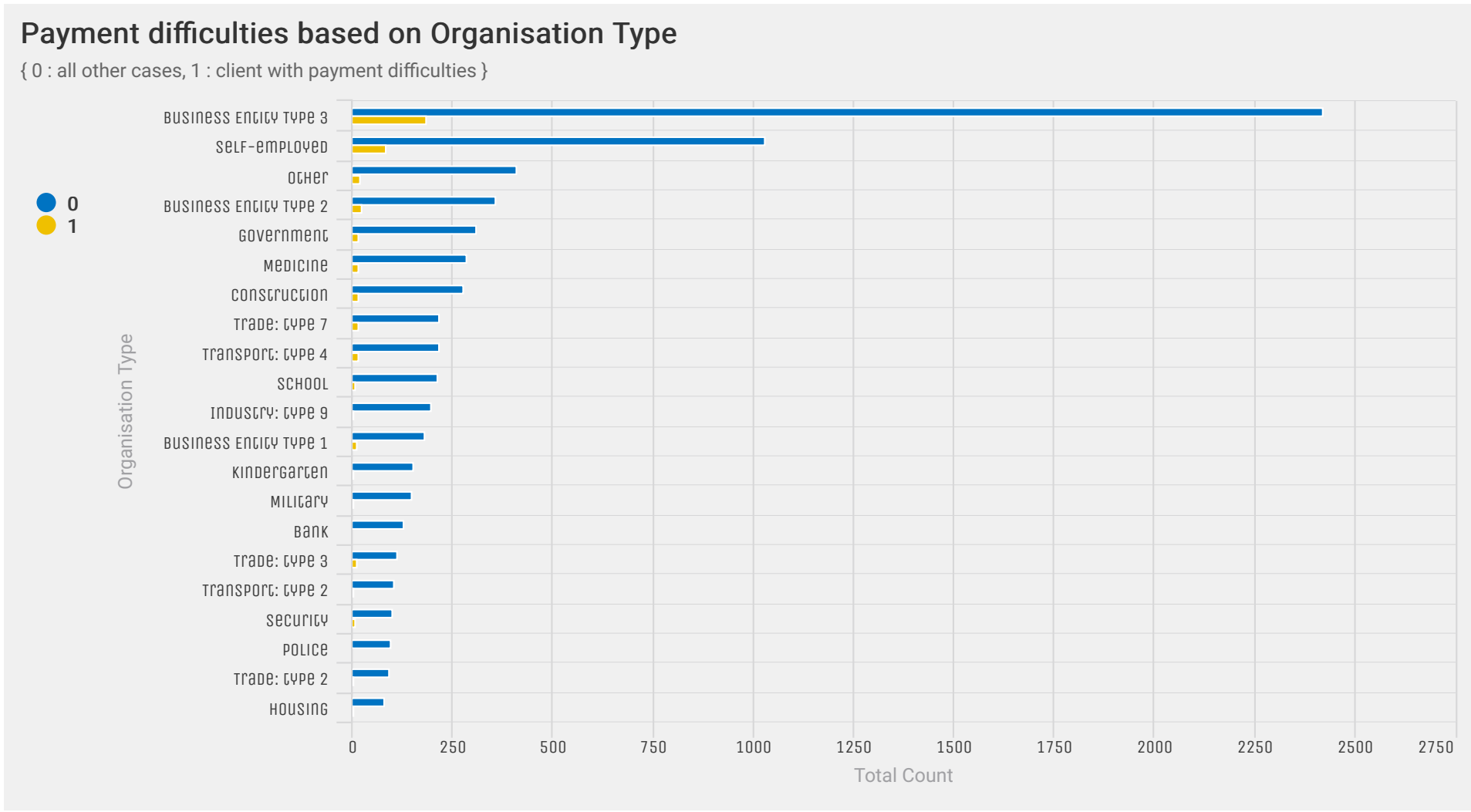
The number of people working for income are more than any other category. But the number of people having difficulty to pay the more are also from working category people. There are very negligible amount of applicants who are unemployed, student, business man or are on maternity leave who have applied for loans or who have difficulties to pay.





Conclusion:

The dataset had weekday values which was used to find out whether the application day affected their repayment difficulty or not. As seen in the bar chart above, Most people prefer to apply for a loan on Monday with 1404 applications followed by Wednesday with 1379 applications. The least prefered day was Sunday as most of the banks are closed. One of the possible reasons for the numbers showing up could be that these applications must be filled online.



Conclusion

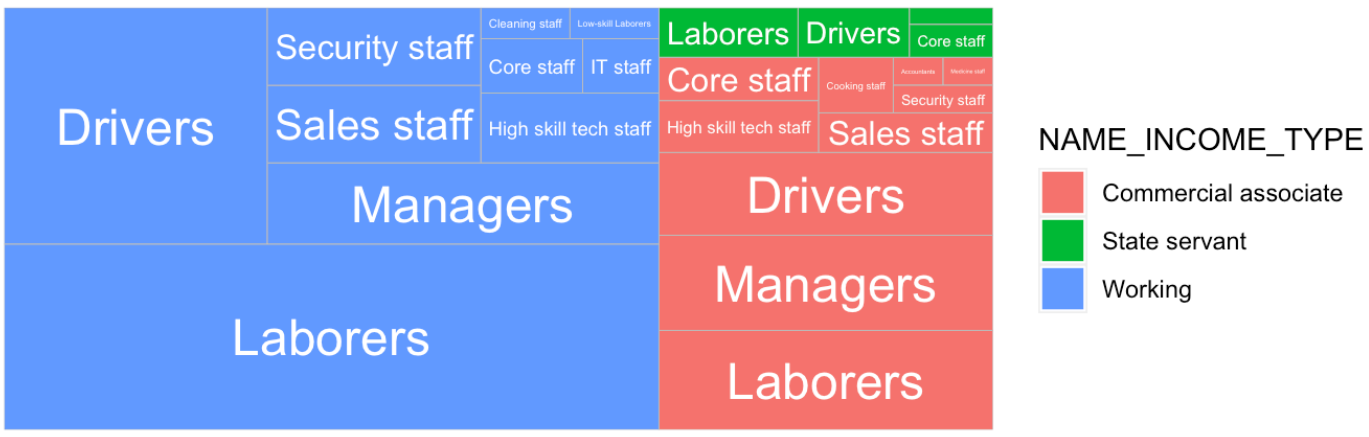
For this chart, as there were several Organisation types it was a bit difficult to adjust all of them in a single chart. Hence, only top 20 were considered and displayed in a descending order. According to the Grouped bar chart, people belonging to organizations like Business entity type 3 and self-employed have applied for the most number of loans, around 2421 and 1031 respectively out of which 186 for Business entity type 3 and 83 for self-employed are having difficulties in paying back.

Furthermore, according to the from the chart, the categories which were facing the least difficulties includes Bank, Cleaning, Electricity, Industry: type 10, Industry: type 2, Industry: type 3, Industry: type 6, Industry: type 8, Restaurant, Trade: type 4, Trade: type 6.

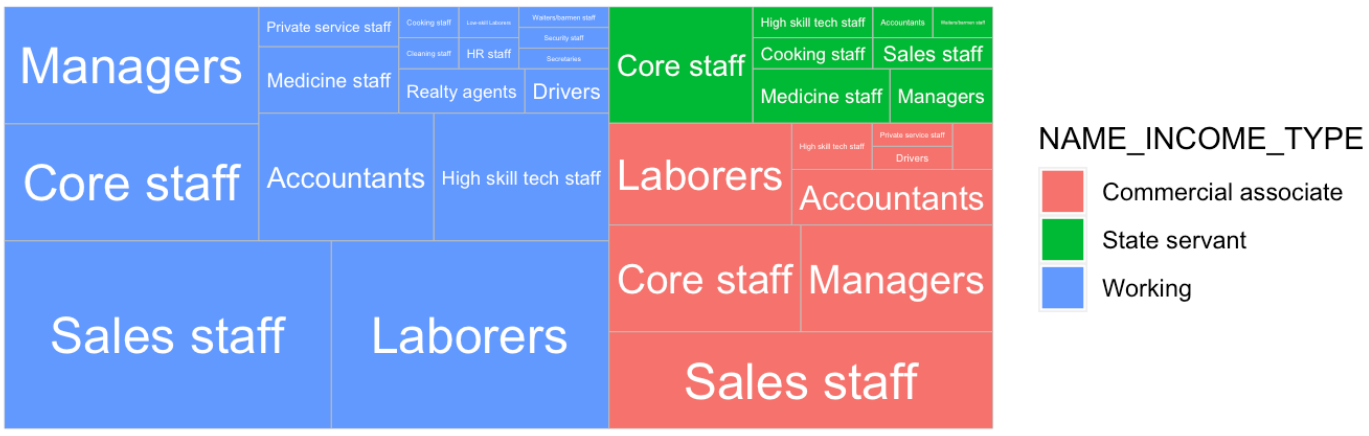
Q2. What income type of clients their occupation are most likely to have difficulties while making a payment? Does Gender play a role in it?

play a role in it?

Male of various occupations facing difficulty in repayment



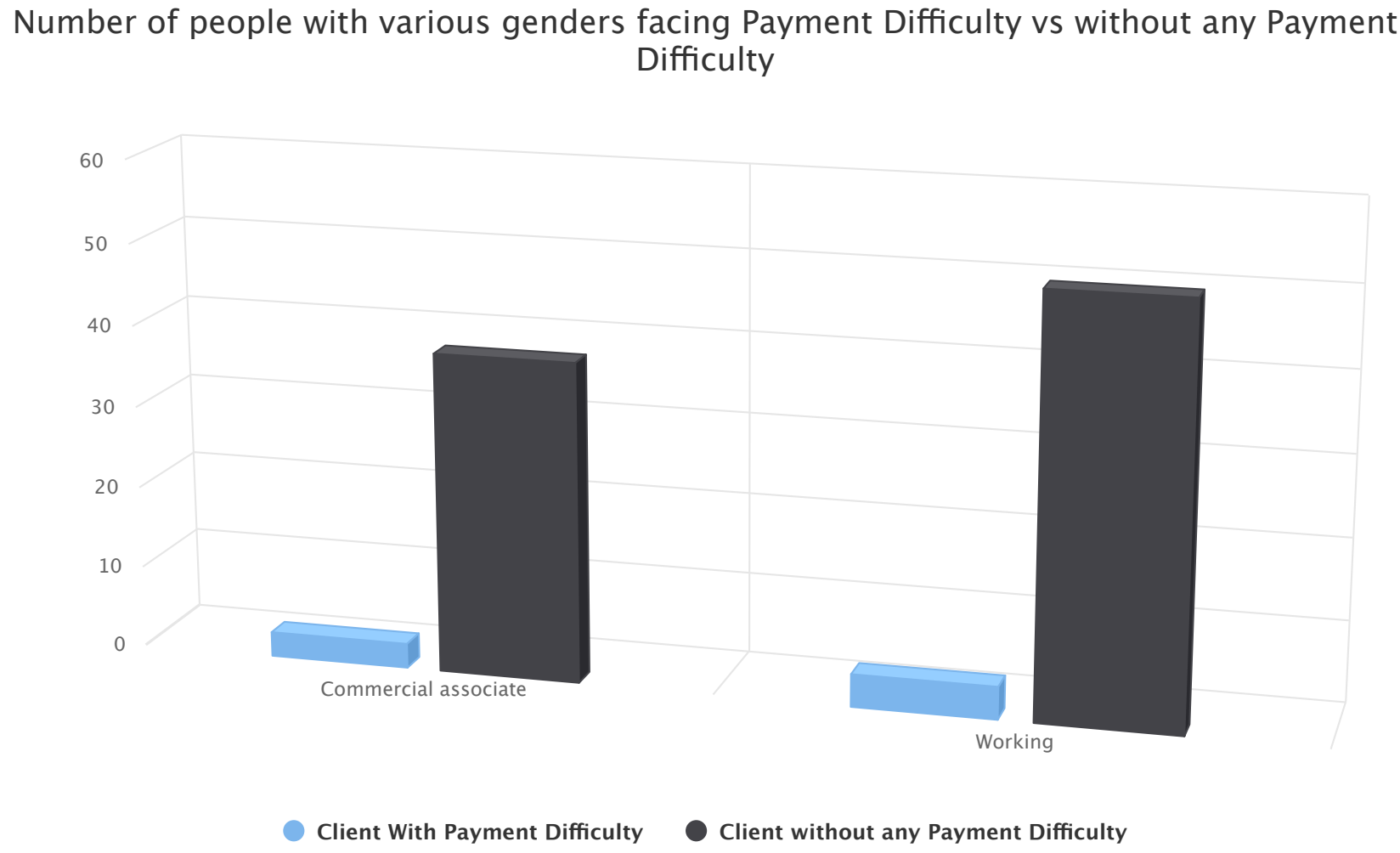
Females of various occupations facing diffuculty in repayment



Conclusion:

For this Tree-map, the group first filtered out the dataframe according to the target value and Gender. We then extracted the OCCUPATION_TYPE and NAME_INCOME_TYPE from the original dataset.As it can be seen in the Tree Map, overall, the working category was facing with the most difficulties for both Males and Females followed by Commercial associates and State Servants. In Working category, the Male Laborers were the ones who suffered the most while for females, it was Sales staff and Laborers.

Q3. What are the chances that people in Working and Commercial Categories giving their personal information such as Mobile phone, Work Phone, email, income type, are likely to face problems while repaying?



Conclusion:

The original dataset, the required columns had binary values, hence, those values were filtered out grouped according to NAME_INCOME_TYPE (Commercial associate or Working) and then summerised inorder to count them.

As it can be depicted from the Grouped Bar chart above, the personal details shared by both of them have no major affect as such but it can be

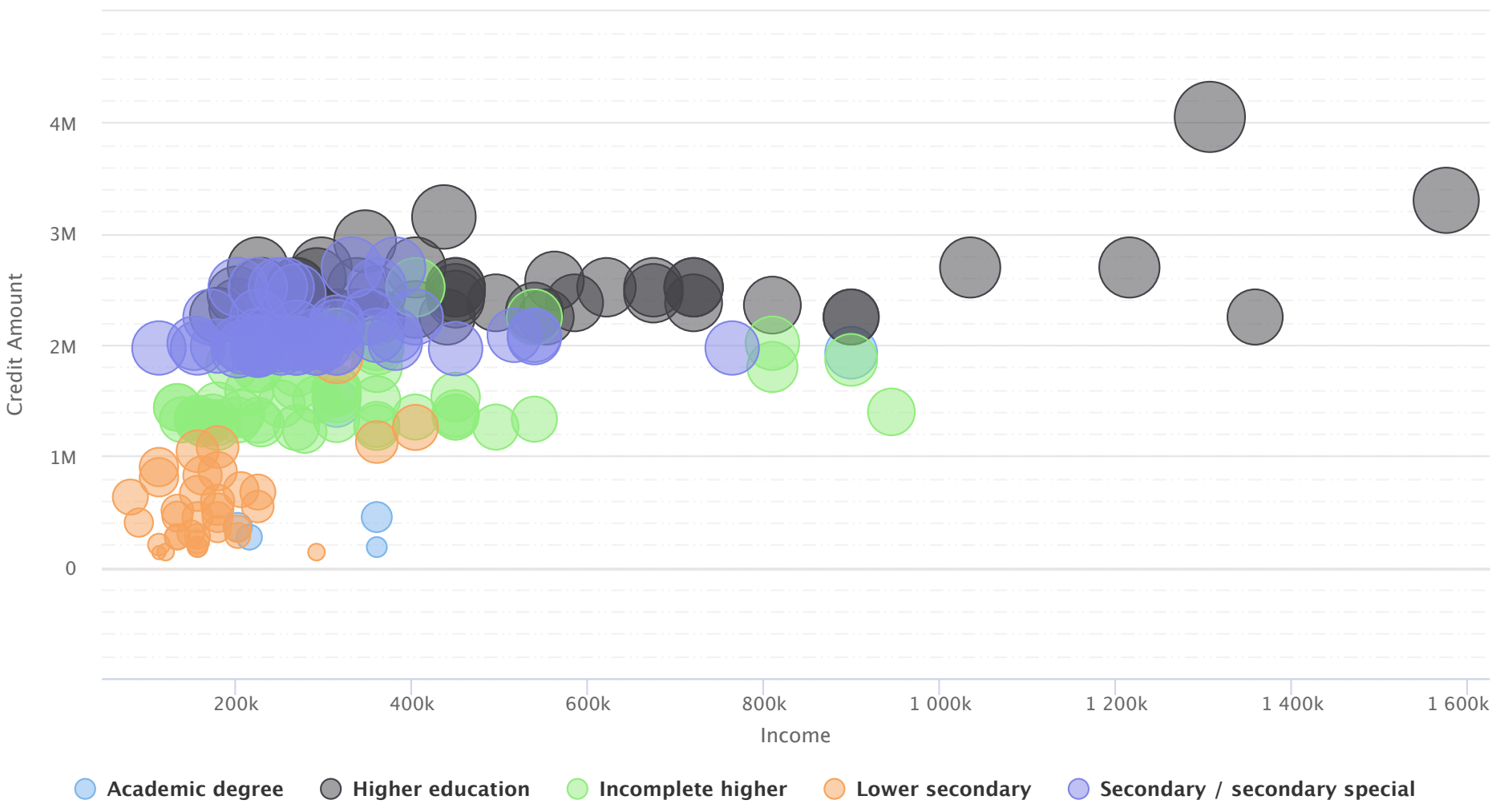
inferred that for Working class, Out of 55 people, 51 repay the loan amount without any problem whereas for Commercial associate, Out of 42 people, 39 repay the amount without difficulties.

Q5. Does the education level of the clients affect the credit amount ?

```
df_edu <- df_creditcard %>%
  select(NAME_EDUCATION_TYPE, AMT_INCOME_TOTAL, AMT_CREDIT) %>%
  group_by(NAME_EDUCATION_TYPE) %>%
  arrange(desc(AMT_CREDIT)) %>%
  slice(1:50) %>%
  drop_na()

h1 = df_edu %>%
  hchart("bubble",
         hcaes(x = AMT_INCOME_TOTAL, y = AMT_CREDIT, size=AMT_CREDIT , group = NAME_EDUCATION_TYPE),
         maxSize = "10%" ) %>%
  hc_xAxis(title = list(text = "Income"),
         opposite = FALSE) %>%
  hc_yAxis(title = list(text = "Credit Amount"),
           opposite = FALSE,
           minorTickInterval = "auto",
           minorGridLineDashStyle = "LongDashDotDot",
           showFirstLabel = FALSE,
           showLastLabel = FALSE,
           plotBands = list(
             list(from = 25, to = 80, color = "rgba(100, 0, 0, 0.1)")))

h1
```



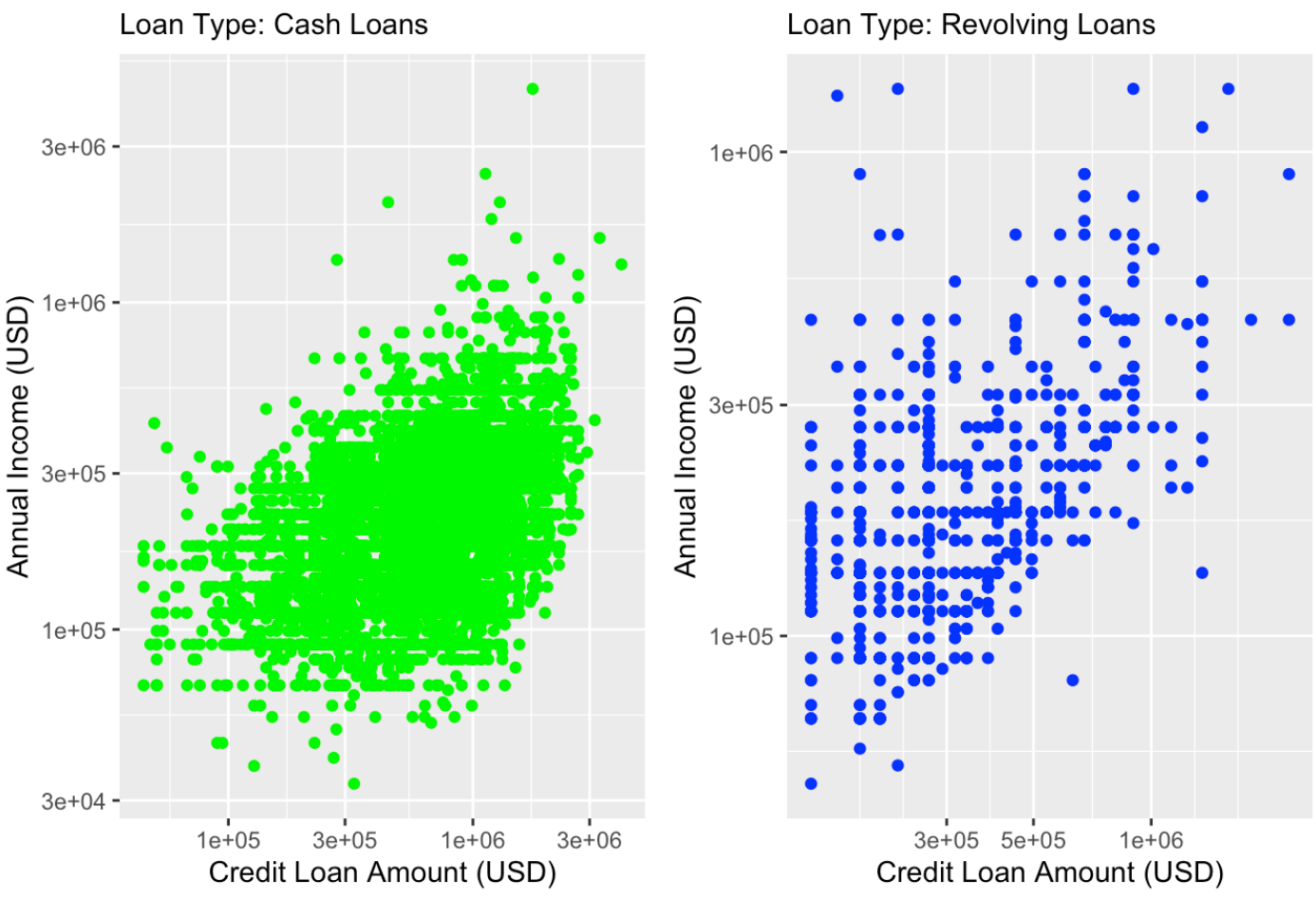
Conclusion:

From the original dataset, NAME_EDUCATION_TYPE, AMT_INCOME_TOTAL, and AMT_CREDIT was extracted and then grouped according to NAME_EDUCATION_TYPE and arranged according to their AMT_CREDIT (credit amount) to display the top 50.

The clients who have successfully completed the Higher Education receive the most benefit here followed by Secondary / secondary special. As seen in the scatter plot, the clients with higher education generally have high salaries hence they tend to have higher credit amount typically above 2 Million. On the contrary, the clients with lower secondary education level have a low credit amount which is around 50k.

Q6: Are higher loan amounts awarded to those with higher annual income? Is the correlation the same for both Cash and Revolving loan types?

Credit Loan vs. Annual Income



Conclusion:

In order to answer this question, we created side by side scatter plots. The plot on the left shows the correlation between Annual Income and Credit Loan Amount for Cash Loans whereas the plot on the right shows the correlation between Annual Income and Credit Loan Amount for Revolving Loans. From the data, one can infer that there is a stronger linear relationship between x & y for Cash Loans. However, we also see a highly dense section of data between $200,000 - $1,000,000 Credit Loan Amount and $100,000 - $350,000 Annual Income where it appears the credit loan amount is relatively arbitrary, which means for the middle quartile there is not a strong correlation between annual income and credit loan amount. This is in keeping with a common practice where banks give loans to people who cannot repay them in order to gain interest and potentially foreclose on their property

Q7: How does occupation impact the amount of credit an individual has? Does marital status play a role in this?

Conclusion:

To determine which occupation type secured the highest credit limits, we plotted different occupation types in box plots side by side grouped by marriage status. From the boxplots we can see that in general, there is not a large distinction between occupation types in terms of the amount of credit an individual has. However, when we look at marriage status, we can clearly see that people who are married tend to get high amounts of credit as compared to single, separated, or widowed people. This makes sense given married couples file under different tax laws, which also allows them to file for loans as a couple, combining their incomes for the risk evaluation. As such, the data shows that marital status actually plays a larger role in credit allowance than occupation type.

Project Conclusion

Our team evaluated a large data set, which focused on credit lending and descriptors of individuals that might impact their ability to secure a loan. We started by evaluating what the correlations are between variables to get a general understanding of our data. We then began analyzing many different aspects of our dataset to better classify what impacts individuals ability to secure a loan.

We evaluated what occupations face the largest difficulty in repaying loans, stratified by gender, where we saw that Laborers and Sales Staff tend to have the most difficult times repaying loans. Next, we evaluated clients who had difficult paying back loans based on their job classifier, Commercial Associate vs. Working. We found that in both categories the vast majority of clients were able to pay back their loans without difficulty. Further, we evaluated payment difficulty by type of loan and found that clients who took out cash loans tended to face more difficulty in repayment. We repeated a similar analysis based on weekday and organization type and found that among the most common organization types, self-employed individuals are some of the most likely individuals to apply for a loan. This tends to make sense given people who are starting businesses typically need seed money in order to get off the ground.

Next we look at loan amounts versus a number of different factors. First, we looked at loan amount versus education level. We found that in general, individuals with a higher education degree tended to have larger salaries and therefore a larger credit amount. Next, we evaluated the correlation strength between annual income and credit loan amount, stratified by loan type. We found strong correlations for both types of loans between the two aforementioned variables. Again, this tends to make sense given people who have a larger salary will be able to pay back more money in loans. Finally, we looked at amount of credit versus occupation type and found that in general, amount of credit was not highly dependent upon occupation type but it was highly dependent upon marital status. We would expect multi-income families to be able to payback more in loans than individuals.

In conclusion, this project not only explores and describes key factors in securing loans but it also takes advantage of many types of visualization techniques in order to demonstrate and display this data visually to all audiences. Performing this study is important because understanding how much money you should expect to qualify for is critical when making large purchasing decisions such as a house, car, or boat. Using this analysis, people can be better prepared to apply for a loan, know what to expect, and understand what they can afford.

(reference: https://wearecitizensadvice.org.uk/lending-people-money-they-cant-afford-is-too-profitable-8c97863b2834 (https://wearecitizensadvice.org.uk/lending-people-money-they-cant-afford-is-too-profitable-8c97863b2834))