**Project Description:**

With the dataset provided for bank loan data, my job is to analyze it and provide insights that can help save valuable clients for the bank in question while also making sure that it does not lose any money because of defaulters. In the later pages of this report, I will apply statistical measures on the data which will help the bank understand what factors affect the payment probability of a loan applicant. The approach and the corresponding insights and results will be listed down in detail in the coming pages of this report.

**Tech-Stack Used:**

To implement this project, I have used Microsoft Excel 365 Version 2302. There are two reasons for this:

* MS Excel 365 allows a lot of features that are not available in the previous version such as there are a lot of automated charts which help me to quickly use them on my data without having to perform the calculations manually.
* This subscription is offered to me by the company I currently work in, so I am quite familiar with its ins and outs.

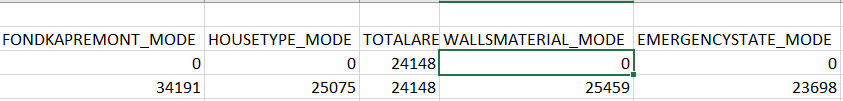
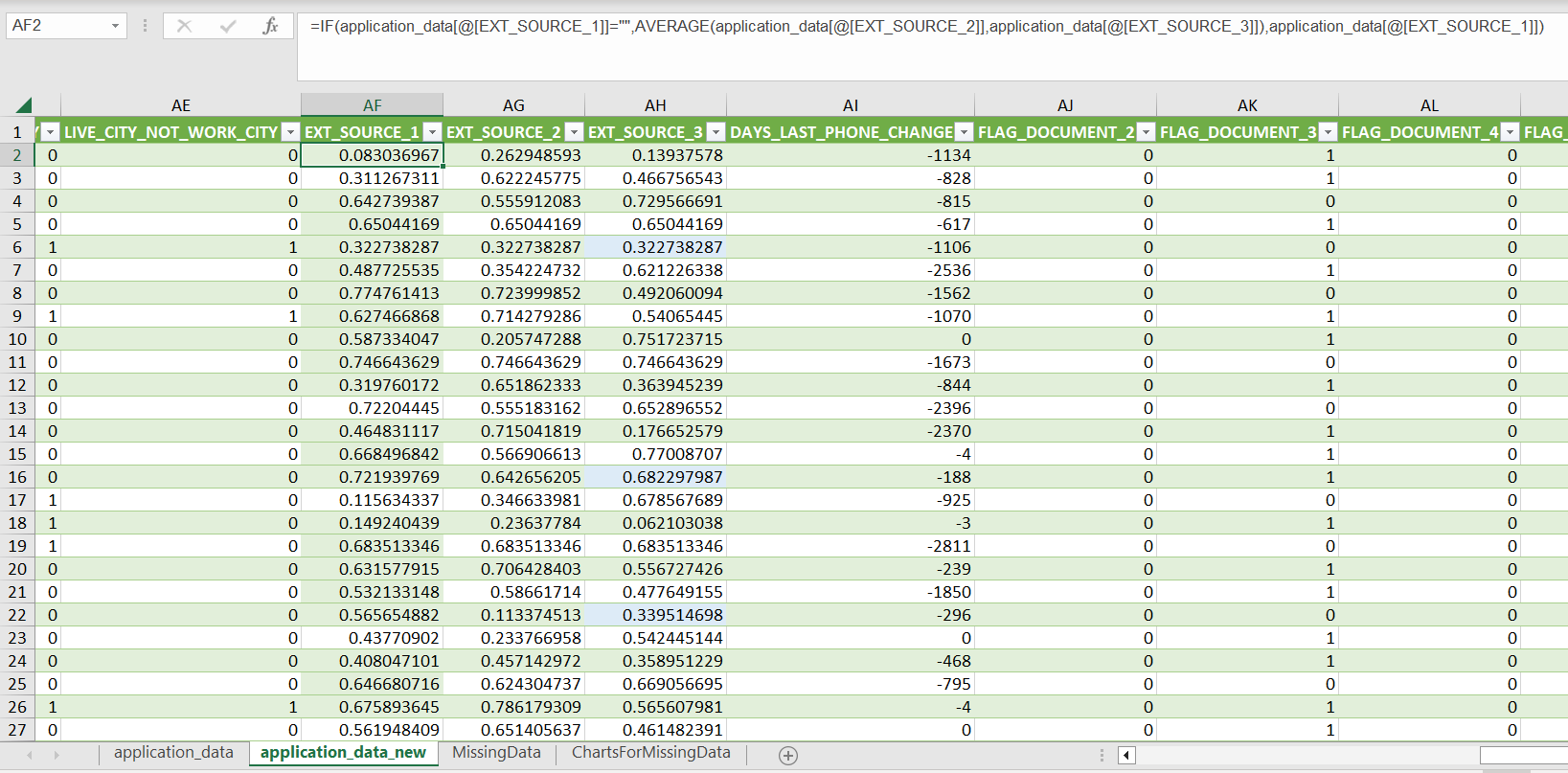
Kindly download the dataset from the link below and view in Excel to see all the charts and tables for the problems (Google sheets doesn’t show all the analytics done). In total there are 13 sheets in this workbook.

[application\_data\_analyzed](https://docs.google.com/spreadsheets/d/1GVhebUH5EhYqyJyiBetDTQvpkHOXd_3E/edit?usp=sharing&ouid=104611970421205783778&rtpof=true&sd=true)

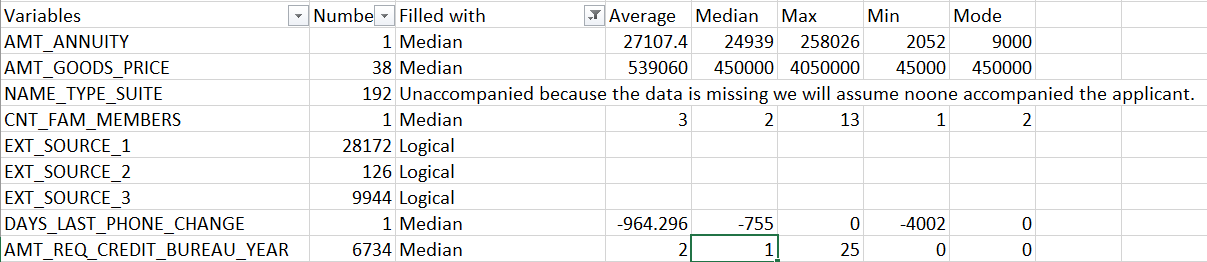
Problem Description No. 1:

Identify the missing data in the dataset and decide on an appropriate method to deal with it using Excel built-in functions and features.

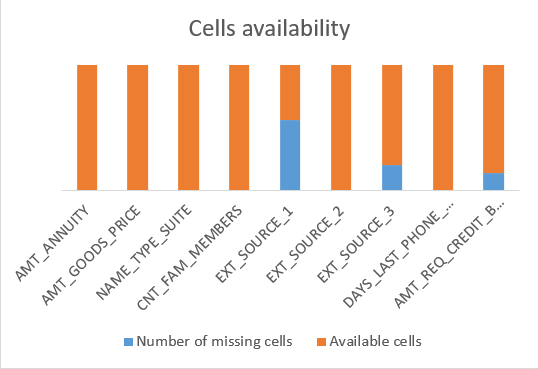
**Approach:**

1. For a first look at the data, I have highlighted the blank cells using conditional formatting option with the formula ISBLANK(range).
2. To count the number of blank cells in each column or attribute, I have used the COUNT(), IF() and ISBLANK() functions.
3. We observe that using the above functions doesn’t accurately count the blank cells because in excel a blank cell doesn’t mean an empty cell. Blank cells can have non-zero length strings whereas empty cells are actually empty.
4. This is why we will use COUNTBLANK() function which accurately counts the blank cells also in the “MissingData” sheet.  
   
5. Next, we must identify if these missing values are essential for our risk analysis and if so, how to fill in them.
6. For the numerical variables, we will first identify whether the distribution is normal or not. If yes, then we will fill the missing values with the average. If no, then with the median or the mode.
7. Another thing to note here is that **all the variables/attributes are not necessary** for our analysis. Hence, we will fill in only those missing variables which can help us derive meaningful insights.
8. After filtering out these variables, we will need to fill the missing data for only 15 variables.
9. In the excel sheet “application\_data\_new” I have removed all the columns that are not useful for my analysis.
10. Occupation\_type is removed since Name\_Income\_Type does the job. Similarly, I have removed the region wise rating and address and kept only the city wise information.
11. I have also taken only the credit enquiries about the client one year before the application since even for the clients for whom the data is available, most number of queries have been made one year before the application. We can also see that for some clients, there are no queries made at all. Hence, we will consider only one such query (one year before application) and fill it with the average/median.
12. After this, out of 122 we are left with only 56 columns for our analysis.
13. In the “ChartsForMissingData” sheet, the strategies to handle missing data have been defined.
14. For the “EXT\_SOURCE\_x” columns, I have taken the average of the values available and filled the missing cell with it. This is because for each customer the credit score varies greatly, and the most reliable way would be to take the average of the scores available from all the sources.  
    

Breakdown of the formula:  
If the original column from sheet “application\_data” is empty, fill the new column in sheet “application\_data\_new” with the average of the other two columns from “application\_data” or else fill it with the available value.

1. After filling these three columns, we will find some cells with #DIV0! Error. This is because for such records, no credit score is available. We will use the IFERROR() function for such cells and use “NA” where we get this error.
2. Finally, we have the following summary for filling the missing data for the useful variables:  
   

**Insights:**

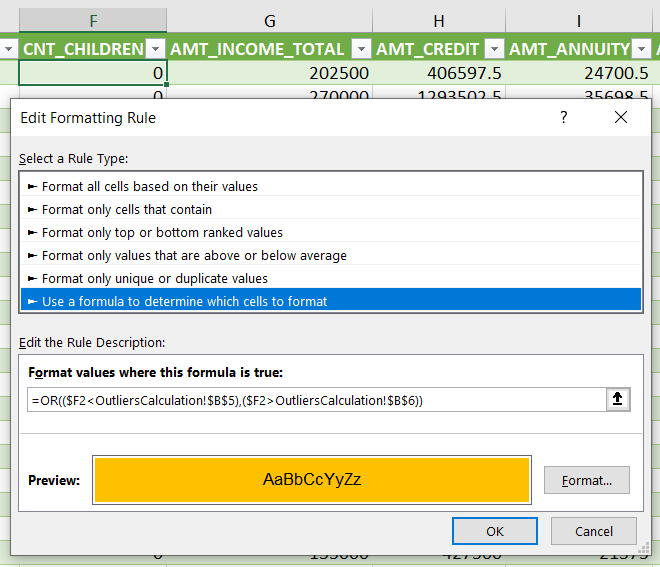
From the Missing Data information, I have created a stacked column chart to visualize the proportion of the missing data as compared to that which is available.  
  
Here, we can see that the missing data is almost negligible as compared to the available data for the variables that are significant for our analysis. Only for EXT\_SOURCE\_1, the missing data is higher as compared to the available data, but we don’t have to worry about that since later for credit score reliability we are going to consider the average of the scores available from all the three sources. And here we can see that for source 2 and 3 we mostly have data available.

Count of available data cells have been calculated simply by subtracting the count of missed data cells from the total number of records available (49999).

Problem Description No. 2:

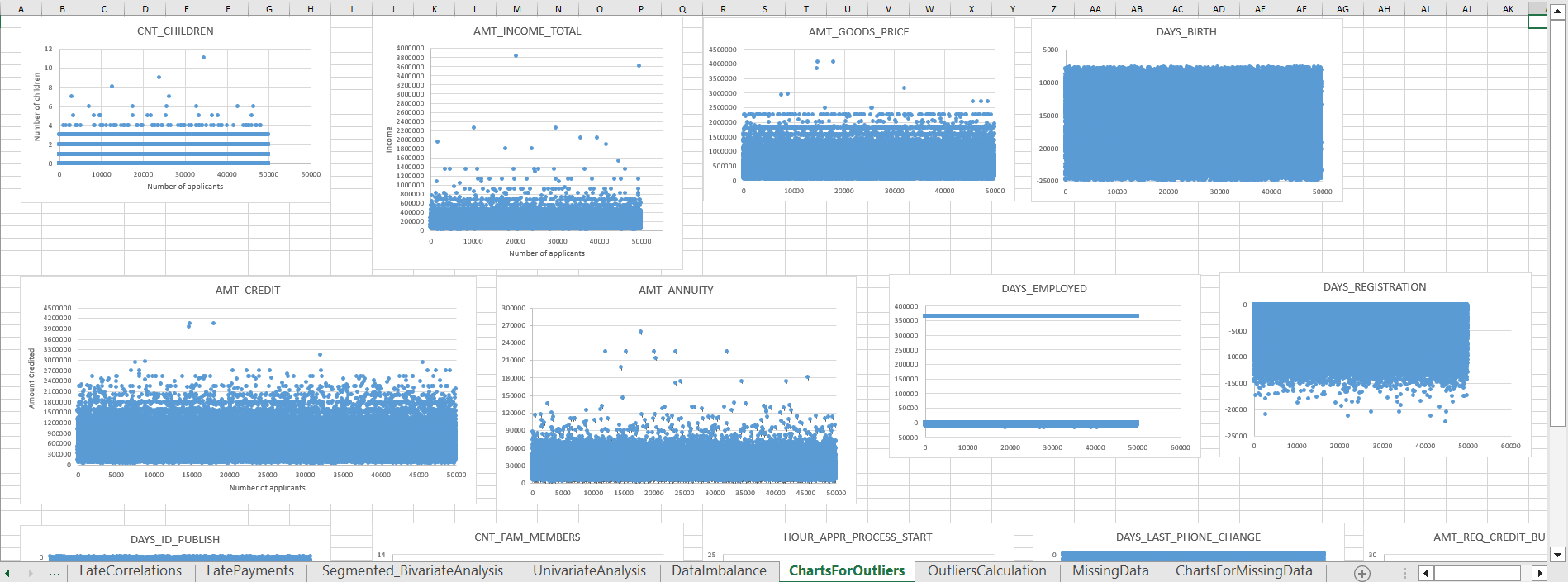
Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.

**Approach:**

1. We still have a lot of numerical variables when we consider the cleaned data obtained in the solution to problem 1. The first variable - SK\_ID\_CURR is not really a variable since it denotes the loan application ID, and it is understood that it won’t have any outliers impacting our analysis. Similarly, TARGET also has only two values – 1 and 0 to denote the payment probability.
2. We will calculate the IQR (Inter Quartile Range) for each of these numerical columns (16 in total) as given in sheet “OutliersCalculation”.
3. We can see that for the EXT\_SOURCE columns we get a DIV0! error but since it already is a normalized score, we don’t have to worry about the outliers here.
4. Next, with the lower and upper bounds identified, I have formatted the cells in the “application\_data\_new” sheet to visualize the outliers calculated above using the following formula:  
   
5. Similarly, I have copied this formula for the other columns and highlighted the outliers.

**Insights:**

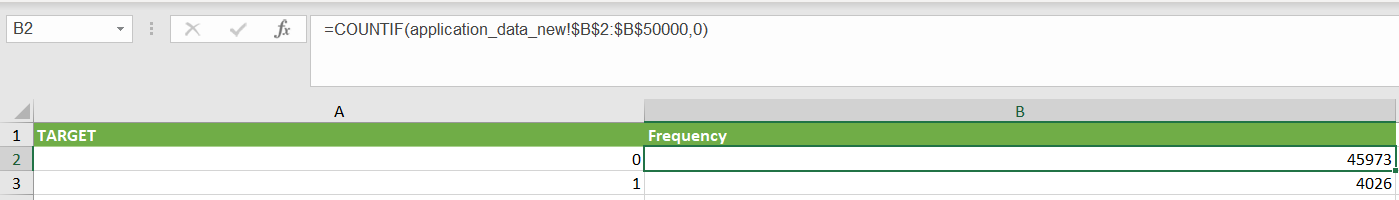
When we map the data on a scatter plot chart, we have the following findings for the outliers:

1. CNT\_CHILDREN: Even though IQR tells us that 4 is also an outlier, from the scatter plot we can see that there are a lot of applicants who have 4 children. Hence, we will only consider more than 4 children as outlier values.
2. AMT\_INCOME\_TOTAL: Here we have one income value 117000000 which comes as the only outlier when we visualize the data. But for further analysis, I have set the maximum value for vertical axis as 40,000,00.  
   Even though, the IQR tell us that any income beyond 3,37,500 is an outlier, from the scatter plot we can see that there are many applicants whose salary is beyond that value. Hence, we will only consider those incomes as outliers that are beyond 10,000,00.
3. AMT\_CREDIT: For the amount credited, the outlier calculation states that any amount above ~16,000,00 can be taken as an outlier value. But when we see the scatter plot, it is evident that there are more than just a few applicants who have received a credit of more than 16,000,00. This is why we will consider only the credit amount beyond 27,000,00 as outlier.
4. AMT\_ANNUITY: Here also, we will consider annuity above 1,20,000 as outlier.
5. Similarly, for the other variables the outliers can be clearly seen in the scatter plots I have created in the “ChartsForOutliers” sheet.  
   

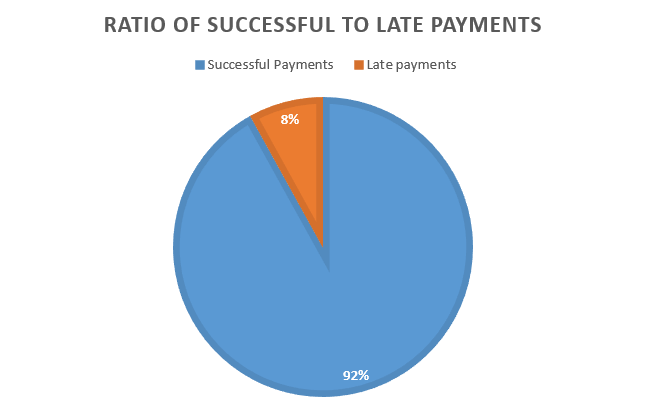
Problem Description No. 3:

Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.

**Approach:**

1. We will count the number of 0s and 1s in the TARGET column from the application\_data\_new sheet using the COUNTIF function.  
   
2. Next, to calculate the ratio of the number of successful payments to the number of payment difficulties we will simply divide B2 by B3 and round it to 2 decimal places.

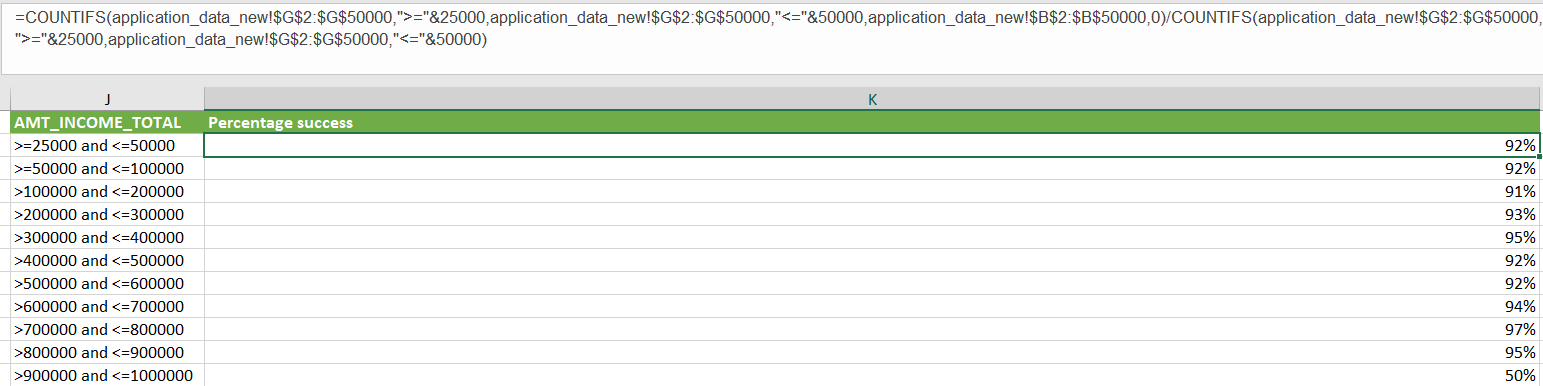
**Insights:**

From the ratio we can see that there is a clear data imbalance since the value is too high = 11.42. Even when we visualize the result, we can see that the number of successful payments is much more than the payments with difficulties.  


Problem Description No. 4:

Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features.

**Approach:**

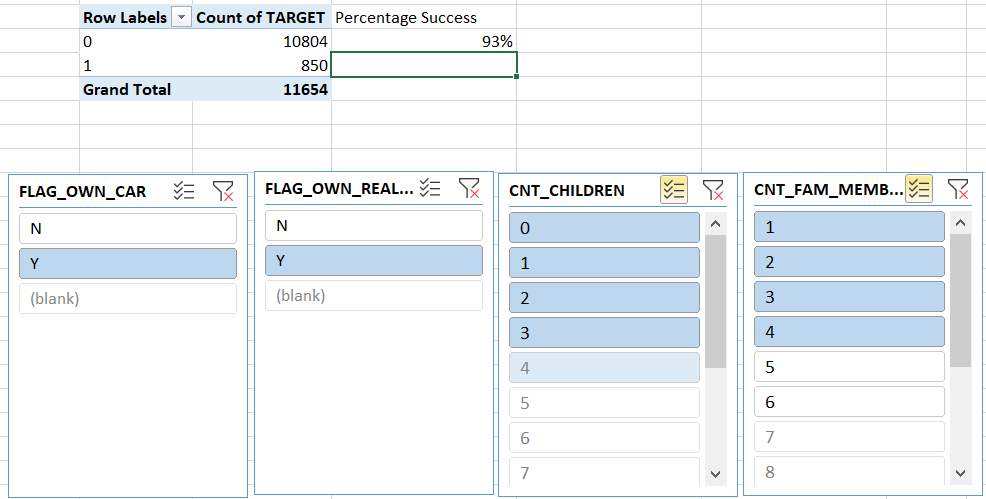
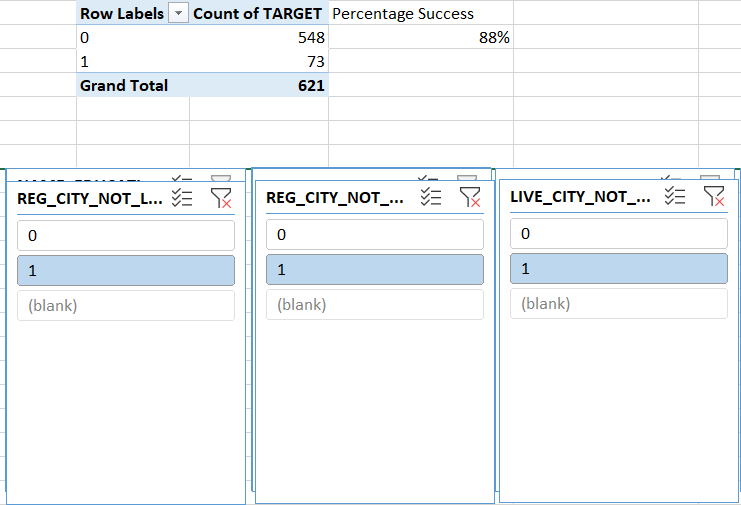
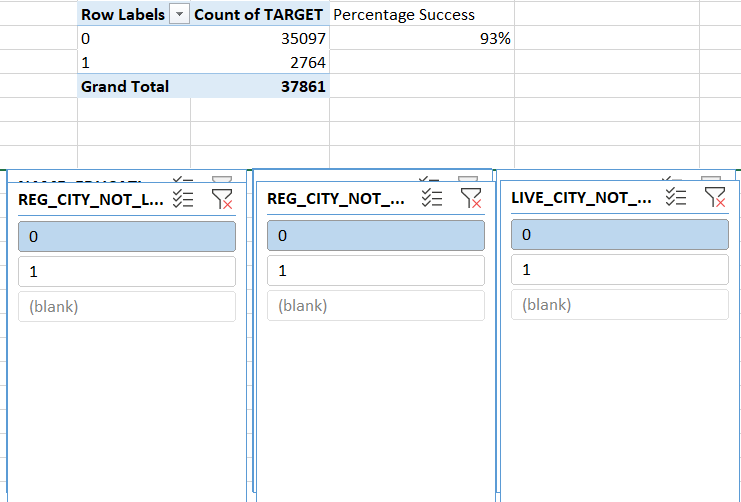
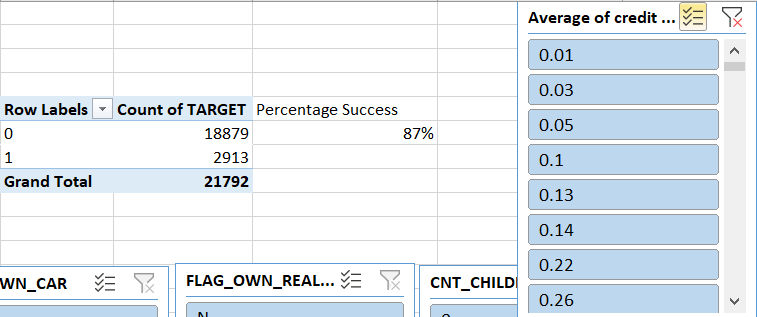
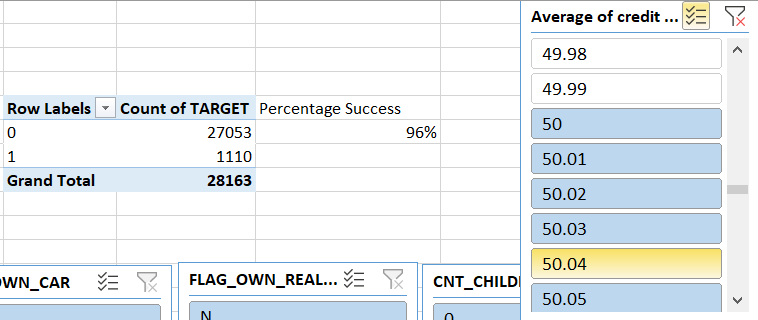
1. For each of the variables filtered in the application\_data\_new sheet, I have done the analysis based on which class had been successful paying off the loans and which class hasn’t been.
2. For each of these variables, I have used the COUNTIFS excel function in the sheet “UnivariateAnalysis” where the percentage success for each variable category is demonstrated.
3. Here, I have taken only the major variables for univariate analysis.  
   FLAG\_OWN\_CAR, FLAG\_OWN\_REALTY, CNT\_CHILDREN, AMT\_INCOME\_TOTAL, NAME\_INCOME\_TYPE, NAME\_EDUCATION\_TYPE, NAME\_FAMILY\_STATUS, NAME\_HOUSING\_TYPE, DAYS\_EMPLOYED, CNT\_FAM\_MEMBERS, REGION\_RATING\_CLIENT\_W\_CITY.
4. For example, for the AMT\_INCOME\_TOTAL I have used the following formula which gives me the success rate of paying off the loans for each income class.  
   
5. For segmented and bivariate analysis, I have simply created a pivot table with only the “COUNT of TARGET” as my analytical variable and inserted slicers for all the variables, which can be used to filter based on desired analysis.
6. Additionally, I have calculated the average of normalized credit scores from the 3 sources available for an overall analysis.

**Insights:**

1. From the percentage success rates we can see that applicants who own a car and their own home, have number of children less than or equal to 2, have total family members less than equal to 4 and are residents of tier 1 city are **more successful in paying off** their loans on time as compared to the other applicants.
2. Applicants having income in the bracket (700000, 800000] have the highest success payment rate whereas those in the bracket (900000, 1000000] have the lowest.
3. Students, Businessmen, and women on Maternity leave have 100% successful payment rate whereas it is clear from the data that the unemployed applicants have the lowest rate of 67%.
4. Applicants holding an academic degree have 100% successful payment rate whereas those having only the lower secondary education have the lowest rate of 88%.
5. From the NAME\_FAMILY\_STATUS, we can see that it doesn’t have much impact on the payment of loans even though widows and married applicants have a higher successful payment rate.
6. Applicants who live with their parents or in a rented apartment have lower successful payment rates than the other applicants.
7. From DAYS\_EMPLOYED, we can clearly see a **linear relationship between the successful payment rate and the number of days the applicant has been employed for**. If the applicant has been employed for a greater number of days before applying for a loan he/she has a higher chance of paying it off on time.

The results for each analysed variable have been visualized in the form of charts in the excel sheet attached.

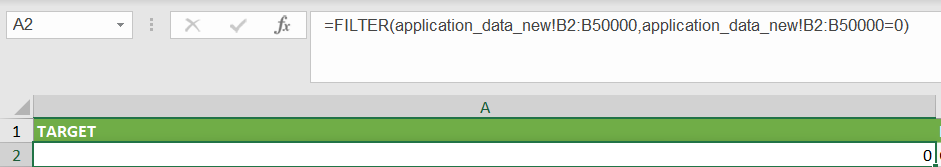
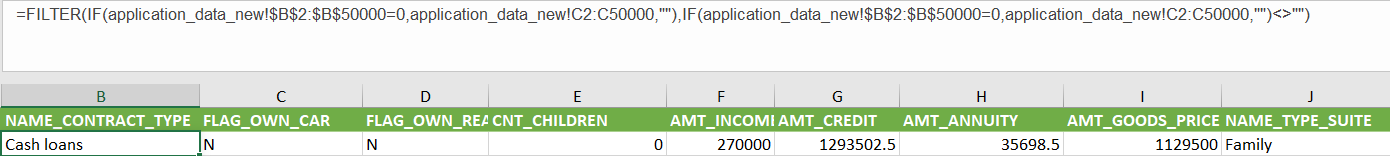
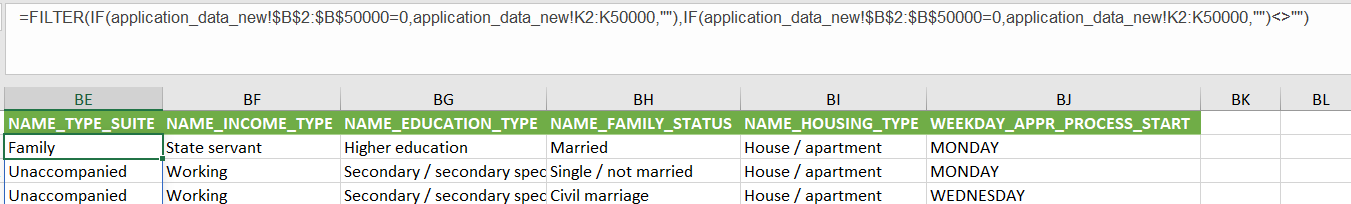
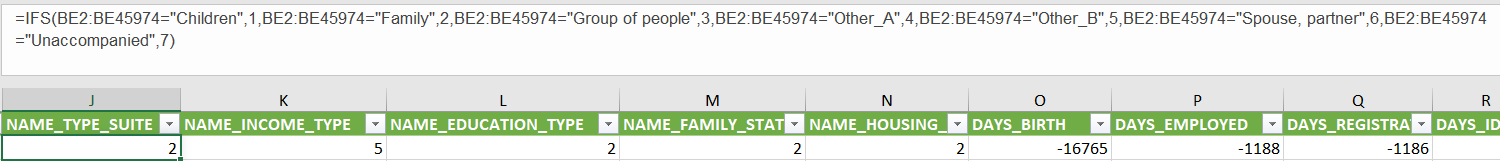
For the segmented and bivariate analysis, there can be n number of inferences that can be drawn because of too many variables. Here I will list down **three** such examples from the “Segmented\_BivariateAnalysis” sheet.

1. Applicants having their own car and home, with number of children less than 4 and total family members less than 5 have 93% successful payment rate.   
   
2. For applicants, where there is a mismatch between the addresses provided the successful payment rate is 88% whereas for those where this no mismatch between the addresses provided, it is 93%.  
    
3. From the average of the credit scores we can see that, the applicants having an average credit score greater than equal of 50 have 96% successful payment rate while those below 50 have only 87% successful payment rate.  
    

Problem Description No. 5:

Segment the dataset based on different scenarios (e.g., clients with payment difficulties and all other cases) and identify the top correlations for each segmented data using Excel functions.

**Approach:**

1. Create two new sheets – “SuccessfulPayments” and “LatePayments” by filtering out the target variable into 0 and 1 and the respective variables’ data.
2. To filter out the data I have used the FILTER and IF functions as demonstrated below:  
     
   
3. For successful payments the range is checked for the value 0 and for late payments it is checked for 1.
4. The second formula in snap 2 is copied across all variables to get the data for them.
5. For the external data sources for credit scores there are NA values which are texts, and since we need to calculate correlations, I have replaced these NA values with 0.
6. We have the following non-numerical variables NAME\_CONTRACT\_TYPE, NAME\_TYPE\_SUITE, NAME\_INCOME\_TYPE, NAME\_EDUCATION\_TYPE, NAME\_FAMILY\_STATUS, NAME\_HOUSING\_TYPE and WEEKDAY\_APPR\_PROCESS\_START.
7. For these variables we will replace each text value to a corresponding number and then replace it within the formula.
8. NAME\_CONTRACT\_TYPE: For Cash loans -> 1. For Revolving loans -> 0.
9. Similarly, for the other textual variables, I have just numbered the texts from 1 to n, where n = number of categories.
10. For the variables where we have more than 2 categories, I have Filtered the data outside my range that I will later use for calculating the correlation matrix.  
    
11. From these filtered rows I have replaced texts with numbers using the IFS() function.  
    
12. Finally to calculate the correlation, I have used the “Data Analysis” add-in available in the “Data” tab of the menu bar and created the correlation matrix in the sheet “SuccessfulCorrelations” for successful target variable and “LateCorrelations” for late target variable.
13. I have replaced the correlations of variables with themselves with a value of 1 with NA.
14. Finally, I have highlighted the top 10 correlations.

**Insights:**

From both the correlations sheet we can see that there is a high relation between the variables:

1. Highest correlation between the amount credited and the goods price = 1.0. This tells us that whatever the price of the goods is, the loan amount is credit based on that only.  
   
2. This is followed by the correlation 0.9 between the variables:  
     
   This is obvious, since more the number of children, more will be the count of family members.
3. We have a 0.8 correlation between the variables:  
     
   This tells us that the contact address and permanent address of applicant is most likely to be the same.

**Result:**

This was by far the most difficult project. The two greatest difficulties I faced were in terms of handling such a large dataset and secondly, getting familiar with the banking terms. Hence, this project helped me learn how to clean the data and use my critical thinking to understand what is really helpful and vital for my analysis. Moreover, it also helped me understand how Data Analytics can really predict the success of an organisation if applied correctly. Overall, this was a heavy project and this has given me even more confidence to deep dive into more real world problems for Data Analytics.