ELINA SINGH

Bank Loan Case Study

Data Understanding:

Download the Dataset using the link given under dataset section on the right.

1. `application_data.csv` contains all the information of the client at the time of application.
   The data is about wheather a client has payment difficulties.
2. `previous_application.csv` contains information about the client's previous loan data. It contains the data whether the previous application had been Approved, Cancelled, Refused or Unused offer.
3. `columns_descrption.csv` is data dictionary which describes the meaning of the variables.

You are required to provide a detailed report for the below data record mentioning the answer to the questions that follows:

- Present the overall approach of the **analysis**. Mention the problem statement and the analysis approach briefly
- **Identify** the missing data and use appropriate method to deal with it. (Remove columns/or replace it with an appropriate value)
  Hint: *Note that in EDA, since it is not necessary to replace the missing value, but if you have to replace the missing value, what should be the approach. Clearly mention the approach.*
- Identify if there are **outliers** in the dataset. Also, mention why do you think it is an outlier. Again, remember that for this exercise, it is not necessary to remove any data points.
- Identify if there is data imbalance in the data. Find the ratio of data imbalance.
  Hint: *Since there are a lot of columns, you can run your analysis in loops for the appropriate columns and find the insights.*
- Explain the **results of univariate, segmented univariate, bivariate analysis, etc.** in business terms.
- Find the top 10 **correlation** for the Client with payment difficulties and all other cases (Target variable). Note that you have to find the top correlation by segmenting the data frame w.r.t to the target variable and then find the top correlation for each of the segmented data and find if any insight is there. Say, there are 5+1(target) variables in a dataset: Var1, Var2, Var3, Var4, Var5, Target. And if you have to find top 3 correlation, it can be: Var1 & Var2, Var2 & Var3, Var1 & Var3. Target variable will not feature in this correlation

as it is a categorical variable and not a continuous variable which is increasing or decreasing.

- **Include visualizations** and **summarize** the most important results in the presentation. You are free to choose the graphs which explain the numerical/categorical variables. Insights should explain why the variable is important for differentiating the clients with payment difficulties with all other cases.

Project Description:

The project is about carrying out Exploratory Data Analysis (EDA) for the bank to understand the defaulters and people who will be able to repay the loans.

Exploratory Data Analysis (EDA):

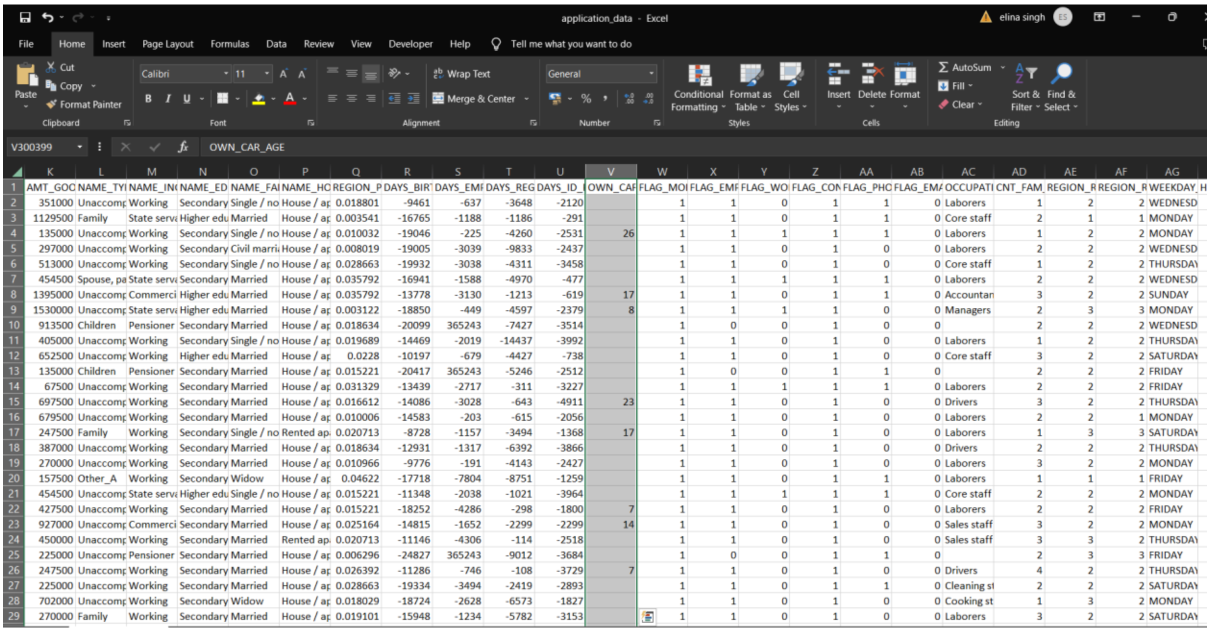
It is an approach to analyse the data using visual techniques. It is used to discover trends, patterns or to check assumptions with the help of statistical summary and graphical representations.

Approach:

- Analyse the data
- Clean the data
- Filling the missing values
- Finding the outliers
- Pivot tables
- Charts to visualise

REMOVING BLANK CELLS

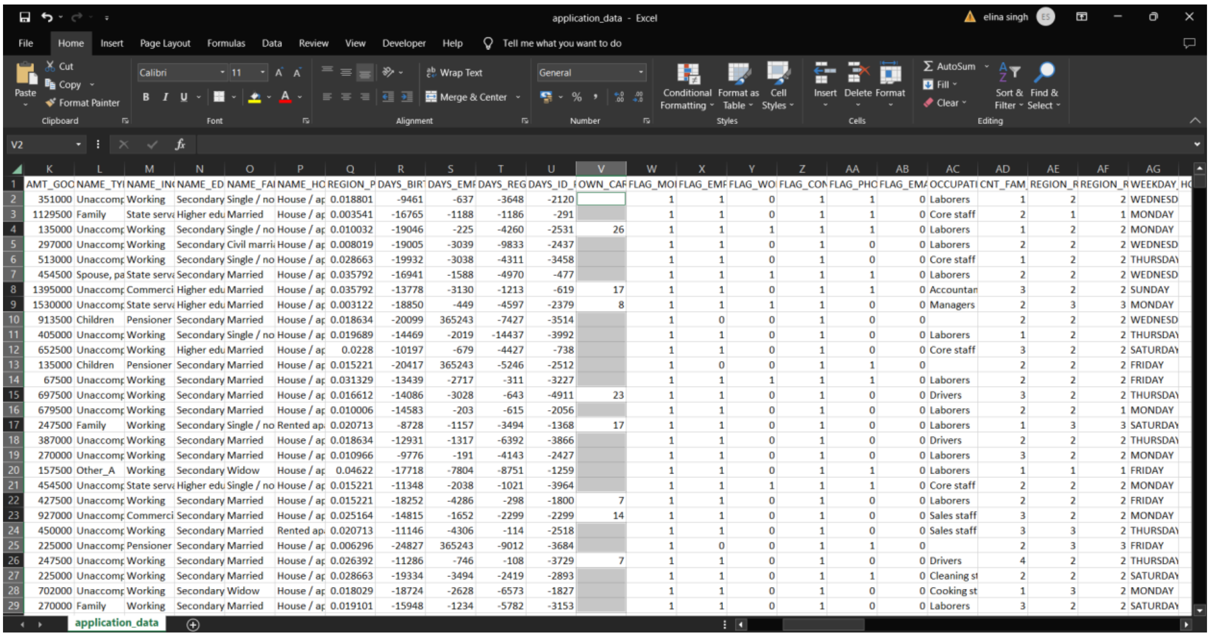
Select the cells the particular column you see blank cells as V1:V307512



V300399 | OWN_CAR_AGE

	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG
1	AMT_GOO	NAME_TYI	NAME_INC	NAME_ED	NAME_FAI	NAME_HO	REGION_P	DAYS_BIRT	DAYS_EMF	DAYS_REG	DAYS_ID_I	OWN_CAF	FLAG_MOI	FLAG_EMF	FLAG_WOI	FLAG_CON	FLAG_PHO	FLAG_EMA	OCCUPATI	CNT_FAM_	REGION_R	REGION_R	WEEKDAY_
2	351000	Unaccomp	Working	Secondary	Single / no	House / ap	0.018801	-9461	-637	-3648	-2120		1	1	0	1	1	0	Laborers	1	2	2	WEDNESD
3	1129500	Family	State serva	Higher edu	Married	House / ap	0.003541	-16765	-1188	-1186	-291		1	1	0	1	1	0	Core staff	2	1	1	MONDAY
4	135000	Unaccomp	Working	Secondary	Single / no	House / ap	0.010032	-19046	-225	-4260	-2531	26	1	1	1	1	1	0	Laborers	1	2	2	MONDAY
5	297000	Unaccomp	Working	Secondary	Civil marri	House / ap	0.008019	-19005	-3039	-9833	-2437		1	1	0	1	0	0	Laborers	2	2	2	WEDNESD
6	513000	Unaccomp	Working	Secondary	Single / no	House / ap	0.028663	-19932	-3038	-4311	-3458		1	1	0	1	0	0	Core staff	1	2	2	THURSDAY
7	454500	Spouse, pa	State serva	Secondary	Married	House / ap	0.035792	-16941	-1588	-4970	-477		1	1	1	1	1	0	Laborers	2	2	2	WEDNESD
8	1395000	Unaccomp	Commerci	Higher edu	Married	House / ap	0.035792	-13778	-3130	-1213	-619	17	1	1	0	1	1	0	Accountan	3	2	2	SUNDAY
9	1530000	Unaccomp	State serva	Higher edu	Married	House / ap	0.003122	-18850	-449	-4597	-2379	8	1	1	1	1	0	0	Managers	2	3	3	MONDAY
10	913500	Children	Pensioner	Secondary	Married	House / ap	0.018634	-20099	365243	-7427	-3514		1	0	0	1	0	0		2	2	2	WEDNESD
11	405000	Unaccomp	Working	Secondary	Single / no	House / ap	0.019689	-14469	-2019	-14437	-3992		1	1	0	1	0	0	Laborers	1	2	2	THURSDAY
12	652500	Unaccomp	Working	Higher edu	Married	House / ap	0.0228	-10197	-679	-4427	-738		1	1	0	1	0	0	Core staff	3	2	2	SATURDAY
13	135000	Children	Pensioner	Secondary	Married	House / ap	0.015221	-20417	365243	-5246	-2512		1	0	0	1	1	0		2	2	2	FRIDAY
14	67500	Unaccomp	Working	Secondary	Married	House / ap	0.031329	-13439	-2717	-311	-3227		1	1	1	1	1	0	Laborers	2	2	2	FRIDAY
15	697500	Unaccomp	Working	Secondary	Married	House / ap	0.016612	-14086	-3028	-643	-4911	23	1	1	0	1	0	0	Drivers	3	2	2	THURSDAY
16	679500	Unaccomp	Working	Secondary	Married	House / ap	0.010006	-14583	-203	-615	-2056		1	1	0	1	0	0	Laborers	2	2	1	MONDAY
17	247500	Family	Working	Secondary	Single / no	Rented ap	0.020713	-8728	-1157	-3494	-1368	17	1	1	0	1	0	0	Laborers	1	3	3	SATURDAY
18	387000	Unaccomp	Working	Secondary	Married	House / ap	0.018634	-12931	-1317	-6392	-3866		1	1	0	1	0	0	Drivers	2	2	2	THURSDAY
19	270000	Unaccomp	Working	Secondary	Married	House / ap	0.010966	-9776	-191	-4143	-2427		1	1	0	1	0	0	Laborers	3	2	2	MONDAY
20	157500	Other_A	Working	Secondary	Widow	House / ap	0.04622	-17718	-7804	-8751	-1259		1	1	0	1	1	0	Laborers	1	1	1	FRIDAY
21	454500	Unaccomp	State serva	Higher edu	Single / no	House / ap	0.015221	-11348	-2038	-1021	-3964		1	1	1	1	1	0	Core staff	2	2	2	MONDAY
22	427500	Unaccomp	Working	Secondary	Married	House / ap	0.015221	-18252	-4286	-298	-1800	7	1	1	0	1	0	0	Laborers	2	2	2	FRIDAY
23	927000	Unaccomp	Commerci	Secondary	Married	House / ap	0.025164	-14815	-1652	-2299	-2299	14	1	1	0	1	0	0	Sales staff	3	2	2	MONDAY
24	450000	Unaccomp	Working	Secondary	Married	Rented ap	0.020713	-11146	-4306	-114	-2518		1	1	0	1	0	0	Sales staff	3	3	2	THURSDAY
25	225000	Unaccomp	Pensioner	Secondary	Married	House / ap	0.006296	-24827	365243	-9012	-3684		1	0	0	1	1	0		2	3	3	FRIDAY
26	247500	Unaccomp	Working	Secondary	Married	House / ap	0.026392	-11286	-746	-108	-3729	7	1	1	0	1	0	0	Drivers	4	2	2	THURSDAY
27	225000	Unaccomp	Working	Secondary	Married	House / ap	0.028663	-19334	-3494	-2419	-2893		1	1	0	1	1	0	Cleaning st	2	2	2	SATURDAY
28	702000	Unaccomp	Working	Secondary	Widow	House / ap	0.018029	-18724	-2628	-6573	-1827		1	1	0	1	0	0	Cooking st	1	3	2	MONDAY
29	270000	Family	Working	Secondary	Married	House / ap	0.019101	-15948	-1234	-5782	-3153		1	1	0	1	0	0	Laborers	3	2	2	SATURDAY

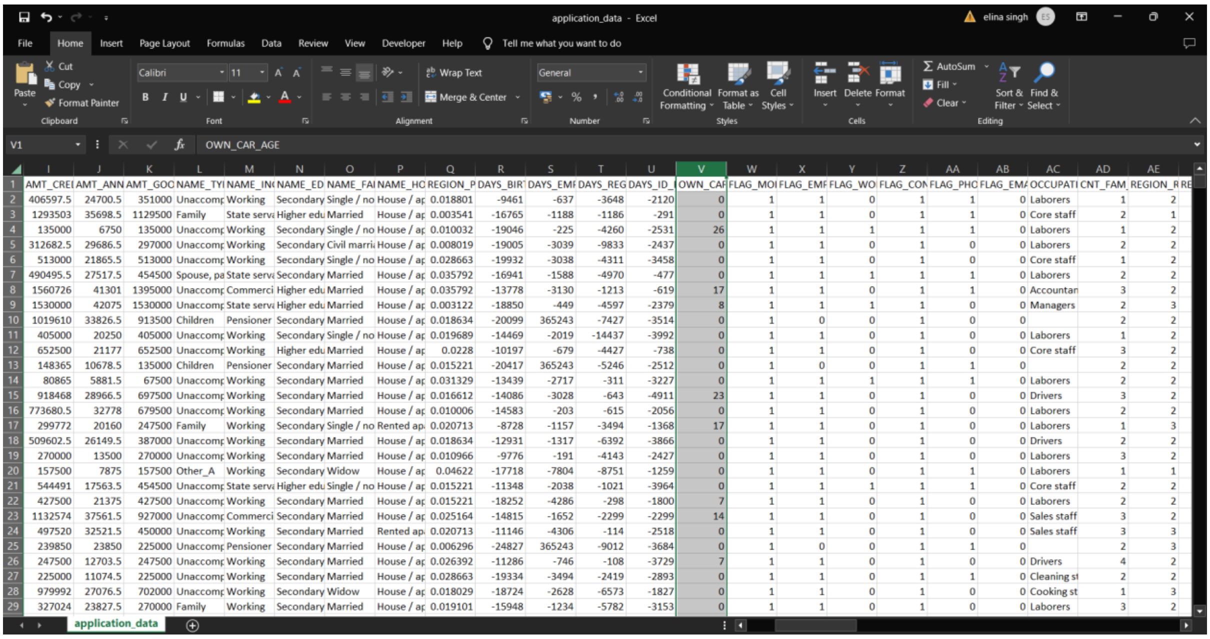
Click **F5** and go to **special** there select **blanks** and click **ok**



V2

	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG
1	AMT_GOO	NAME_TYI	NAME_IN(	NAME_ED	NAME_FAI	NAME_HO	REGION_P	DAYS_BIRT	DAYS_EMF	DAYS_REG	DAYS_ID_I	OWN_CAF	FLAG_MOI	FLAG_EMF	FLAG_WOI	FLAG_CON	FLAG_PHO	FLAG_EMA	OCCUPATI	CNT_FAM_	REGION_R	REGION_R	WEEKDAY_HO
2	351000	Unaccomp	Working	Secondary	Single / no	House / ap	0.018801	-9461	-637	-3648	-2120		1	1	0	1	1	0	Laborers	1	2	2	WEDNESD
3	1129500	Family	State serva	Higher edu	Married	House / ap	0.003541	-16765	-1188	-1186	-291		1	1	0	1	1	0	Core staff	2	1	1	MONDAY
4	135000	Unaccomp	Working	Secondary	Single / no	House / ap	0.010032	-19046	-225	-4260	-2531	26	1	1	1	1	1	0	Laborers	1	2	2	MONDAY
5	297000	Unaccomp	Working	Secondary	Civil marri	House / ap	0.008019	-19005	-3039	-9833	-2437		1	1	0	1	0	0	Laborers	2	2	2	WEDNESD
6	513000	Unaccomp	Working	Secondary	Single / no	House / ap	0.028663	-19932	-3038	-4311	-3458		1	1	0	1	0	0	Core staff	1	2	2	THURSDAY
7	454500	Spouse, pa	State serva	Secondary	Married	House / ap	0.035792	-16941	-1588	-4970	-477		1	1	1	1	1	0	Laborers	2	2	2	WEDNESD
8	1395000	Unaccomp	Commerci	Higher edu	Married	House / ap	0.035792	-13778	-3130	-1213	-619	17	1	1	0	1	1	0	Accountan	3	2	2	SUNDAY
9	1530000	Unaccomp	State serva	Higher edu	Married	House / ap	0.003122	-18850	-449	-4597	-2379	8	1	1	1	1	0	0	Managers	2	3	3	MONDAY
10	913500	Children	Pensioner	Secondary	Married	House / ap	0.018634	-20099	365243	-7427	-3514		1	0	0	1	0	0		2	2	2	WEDNESD
11	405000	Unaccomp	Working	Secondary	Single / no	House / ap	0.019689	-14469	-2019	-14437	-3992		1	1	0	1	0	0	Laborers	1	2	2	THURSDAY
12	652500	Unaccomp	Working	Higher edu	Married	House / ap	0.0228	-10197	-679	-4427	-738		1	1	0	1	0	0	Core staff	3	2	2	SATURDAY
13	135000	Children	Pensioner	Secondary	Married	House / ap	0.015221	-20417	365243	-5246	-2512		1	0	0	1	1	0		2	2	2	FRIDAY
14	67500	Unaccomp	Working	Secondary	Married	House / ap	0.031329	-13439	-2717	-311	-3227		1	1	1	1	1	0	Laborers	2	2	2	FRIDAY
15	697500	Unaccomp	Working	Secondary	Married	House / ap	0.016612	-14086	-3028	-643	-4911	23	1	1	0	1	0	0	Drivers	3	2	2	THURSDAY
16	679500	Unaccomp	Working	Secondary	Married	House / ap	0.010006	-14583	-203	-615	-2056		1	1	0	1	0	0	Laborers	2	2	1	MONDAY
17	247500	Family	Working	Secondary	Single / no	Rented ap	0.020713	-8728	-1157	-3494	-1368	17	1	1	0	1	0	0	Laborers	1	3	3	SATURDAY
18	387000	Unaccomp	Working	Secondary	Married	House / ap	0.018634	-12931	-1317	-6392	-3866		1	1	0	1	0	0	Drivers	2	2	2	THURSDAY
19	270000	Unaccomp	Working	Secondary	Married	House / ap	0.010966	-9776	-191	-4143	-2427		1	1	0	1	0	0	Laborers	3	2	2	MONDAY
20	157500	Other_A	Working	Secondary	Widow	House / ap	0.04622	-17718	-7804	-8751	-1259		1	1	0	1	1	0	Laborers	1	1	1	FRIDAY
21	454500	Unaccomp	State serva	Higher edu	Single / no	House / ap	0.015221	-11348	-2038	-1021	-3964		1	1	1	1	1	0	Core staff	2	2	2	MONDAY
22	427500	Unaccomp	Working	Secondary	Married	House / ap	0.015221	-18252	-4286	-298	-1800	7	1	1	0	1	0	0	Laborers	2	2	2	FRIDAY
23	927000	Unaccomp	Commerci	Secondary	Married	House / ap	0.025164	-14815	-1652	-2299	-2299	14	1	1	0	1	0	0	Sales staff	3	2	2	MONDAY
24	450000	Unaccomp	Working	Secondary	Married	Rented ap	0.020713	-11146	-4306	-114	-2518		1	1	0	1	0	0	Sales staff	3	3	2	THURSDAY
25	225000	Unaccomp	Pensioner	Secondary	Married	House / ap	0.006296	-24827	365243	-9012	-3684		1	0	0	1	1	0		2	3	3	FRIDAY
26	247500	Unaccomp	Working	Secondary	Married	House / ap	0.026392	-11286	-746	-108	-3729	7	1	1	0	1	0	0	Drivers	4	2	2	THURSDAY
27	225000	Unaccomp	Working	Secondary	Married	House / ap	0.028663	-19334	-3494	-2419	-2893		1	1	0	1	1	0	Cleaning st	2	2	2	SATURDAY
28	702000	Unaccomp	Working	Secondary	Widow	House / ap	0.018029	-18724	-2628	-6573	-1827		1	1	0	1	0	0	Cooking st	1	3	2	MONDAY
29	270000	Family	Working	Secondary	Married	House / ap	0.019101	-15948	-1234	-5782	-3153		1	1	0	1	0	0	Laborers	3	2	2	SATURDAY

All cells in that particular column are selected



V1 fx OWN_CAR_AGE

	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE
1	AMT_CREI	AMT_ANN	AMT_GOO	NAME_TYI	NAME_IN(	NAME_ED	NAME_FAI	NAME_HO	REGION_P	DAYS_BIRT	DAYS_EMF	DAYS_REG	DAYS_ID_I	OWN_CAF	FLAG_MOI	FLAG_EMF	FLAG_WOI	FLAG_CON	FLAG_PHO	FLAG_EMA	OCCUPATI	CNT_FAM_	REGION_R
2	406597.5	24700.5	351000	Unaccomp	Working	Secondary	Single / no	House / ap	0.018801	-9461	-637	-3648	-2120	0	1	1	0	1	1	0	Laborers	1	2
3	1293503	35698.5	1129500	Family	State serva	Higher edu	Married	House / ap	0.003541	-16765	-1188	-1186	-291	0	1	1	0	1	1	0	Core staff	2	1
4	135000	6750	135000	Unaccomp	Working	Secondary	Single / no	House / ap	0.010032	-19046	-225	-4260	-2531	26	1	1	1	1	1	0	Laborers	1	2
5	312682.5	29686.5	297000	Unaccomp	Working	Secondary	Civil marri	House / ap	0.008019	-19005	-3039	-9833	-2437	0	1	1	0	1	0	0	Laborers	2	2
6	513000	21865.5	513000	Unaccomp	Working	Secondary	Single / no	House / ap	0.028663	-19932	-3038	-4311	-3458	0	1	1	0	1	0	0	Core staff	1	2
7	490495.5	27517.5	454500	Spouse, pa	State serva	Secondary	Married	House / ap	0.035792	-16941	-1588	-4970	-477	0	1	1	1	1	1	0	Laborers	2	2
8	1560726	41301	1395000	Unaccomp	Commerci	Higher edu	Married	House / ap	0.035792	-13778	-3130	-1213	-619	17	1	1	0	1	1	0	Accountan	3	2
9	1530000	42075	1530000	Unaccomp	State serva	Higher edu	Married	House / ap	0.003122	-18850	-449	-4597	-2379	8	1	1	1	1	0	0	Managers	2	3
10	1019610	33826.5	913500	Children	Pensioner	Secondary	Married	House / ap	0.018634	-20099	365243	-7427	-3514	0	1	0	0	1	0	0		2	2
11	405000	20250	405000	Unaccomp	Working	Secondary	Single / no	House / ap	0.019689	-14469	-2019	-14437	-3992	0	1	1	0	1	0	0	Laborers	1	2
12	652500	21177	652500	Unaccomp	Working	Higher edu	Married	House / ap	0.0228	-10197	-679	-4427	-738	0	1	1	0	1	0	0	Core staff	3	2
13	148365	10678.5	135000	Children	Pensioner	Secondary	Married	House / ap	0.015221	-20417	365243	-5246	-2512	0	1	0	0	1	1	0		2	2
14	80865	5881.5	67500	Unaccomp	Working	Secondary	Married	House / ap	0.031329	-13439	-2717	-311	-3227	0	1	1	1	1	1	0	Laborers	2	2
15	918468	28966.5	697500	Unaccomp	Working	Secondary	Married	House / ap	0.016612	-14086	-3028	-643	-4911	23	1	1	0	1	0	0	Drivers	3	2
16	773680.5	32778	679500	Unaccomp	Working	Secondary	Married	House / ap	0.010006	-14583	-203	-615	-2056	0	1	1	0	1	0	0	Laborers	2	2
17	299772	20160	247500	Family	Working	Secondary	Single / no	Rented ap	0.020713	-8728	-1157	-3494	-1368	17	1	1	0	1	0	0	Laborers	1	3
18	509602.5	26149.5	387000	Unaccomp	Working	Secondary	Married	House / ap	0.018634	-12931	-1317	-6392	-3866	0	1	1	0	1	0	0	Drivers	2	2
19	270000	13500	270000	Unaccomp	Working	Secondary	Married	House / ap	0.010966	-9776	-191	-4143	-2427	0	1	1	0	1	0	0	Laborers	3	2
20	157500	7875	157500	Other_A	Working	Secondary	Widow	House / ap	0.04622	-17718	-7804	-8751	-1259	0	1	1	0	1	1	0	Laborers	1	1
21	544491	17563.5	454500	Unaccomp	State serva	Higher edu	Single / no	House / ap	0.015221	-11348	-2038	-1021	-3964	0	1	1	1	1	1	0	Core staff	2	2
22	427500	21375	427500	Unaccomp	Working	Secondary	Married	House / ap	0.015221	-18252	-4286	-298	-1800	7	1	1	0	1	0	0	Laborers	2	2
23	1132574	37561.5	927000	Unaccomp	Commerci	Secondary	Married	House / ap	0.025164	-14815	-1652	-2299	-2299	14	1	1	0	1	0	0	Sales staff	3	2
24	497520	32521.5	450000	Unaccomp	Working	Secondary	Married	Rented ap	0.020713	-11146	-4306	-114	-2518	0	1	1	0	1	0	0	Sales staff	3	3
25	239850	23850	225000	Unaccomp	Pensioner	Secondary	Married	House / ap	0.006296	-24827	365243	-9012	-3684	0	1	0	0	1	1	0		2	3
26	247500	12703.5	247500	Unaccomp	Working	Secondary	Married	House / ap	0.026392	-11286	-746	-108	-3729	7	1	1	0	1	0	0	Drivers	4	2
27	225000	11074.5	225000	Unaccomp	Working	Secondary	Married	House / ap	0.028663	-19334	-3494	-2419	-2893	0	1	1	0	1	1	0	Cleaning st	2	2
28	979992	27076.5	702000	Unaccomp	Working	Secondary	Widow	House / ap	0.018029	-18724	-2628	-6573	-1827	0	1	1	0	1	0	0	Cooking st	1	3
29	327024	23827.5	270000	Family	Working	Secondary	Married	House / ap	0.019101	-15948	-1234	-5782	-3153	0	1	1	0	1	0	0	Laborers	3	2

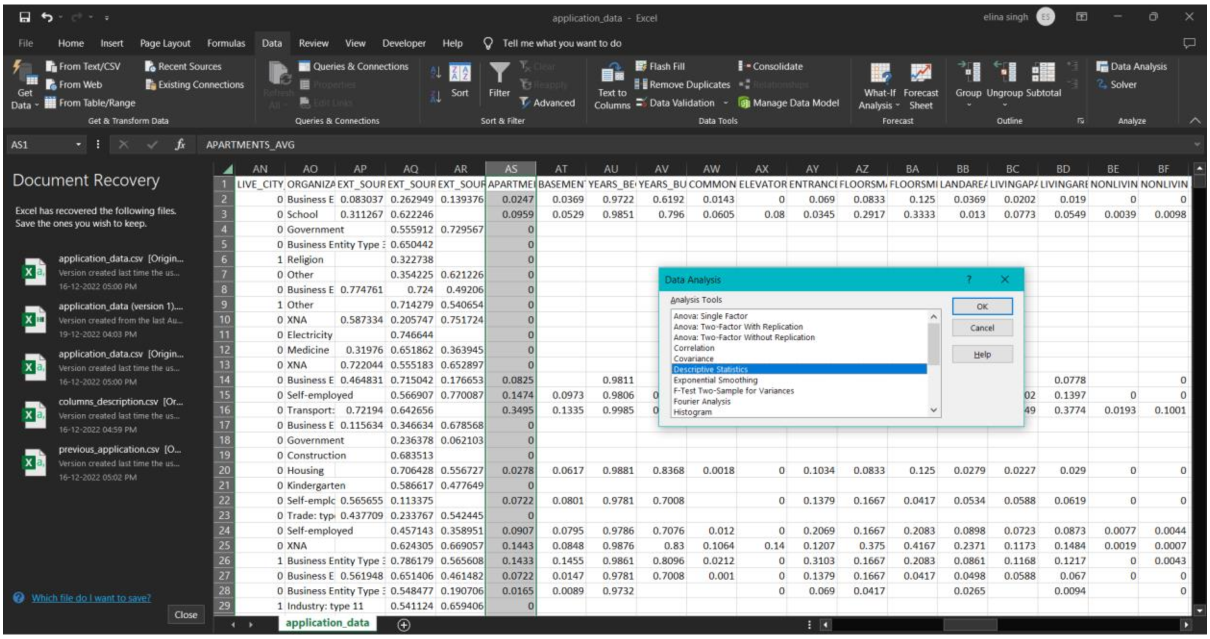
Now Press **F2** and write the value you wish to enter which we have taken **0** and then click **ctrl+enter** all blank cells get filled by 0 repeat the step in all data sets.

Approach to replacing the missing values:

First we select the column we would want to fill the missing values with the mean of the data.

Then we go to the Data toolbar and select Data Analysis from the Analyse tool. The following popup will show up.

Next step is to select Descriptive statistics to get the complete data of the column



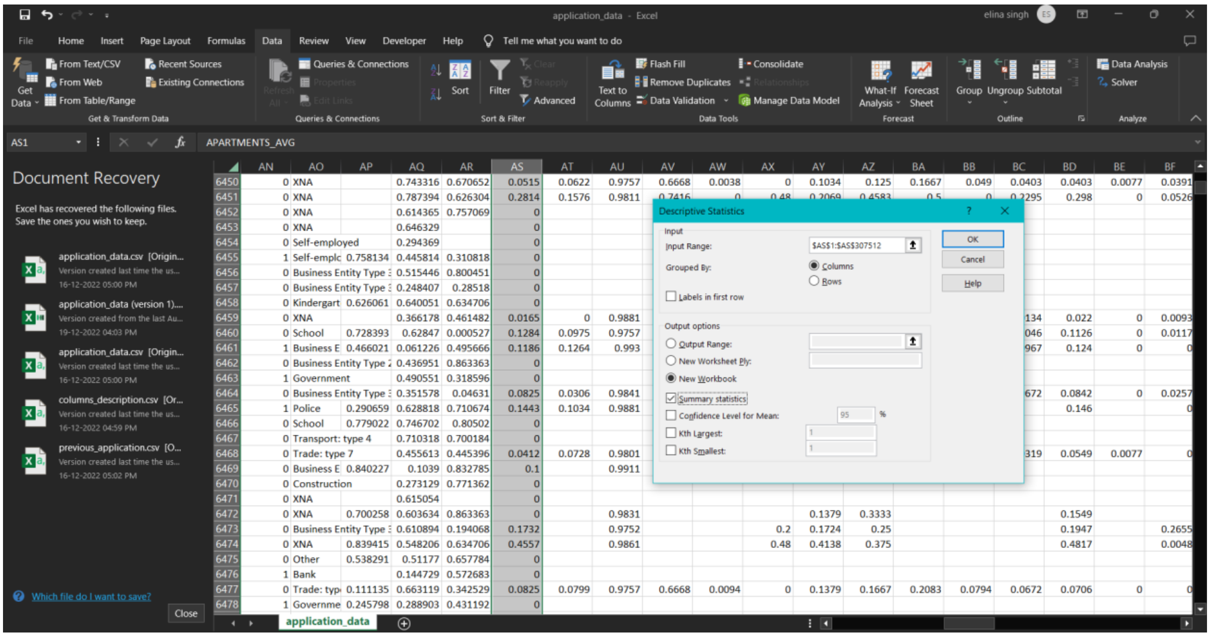
Data Analysis

Analysis Tools

- Anova: Single Factor
- Anova: Two-Factor With Replication
- Anova: Two-Factor Without Replication
- Correlation
- Covariance
- Descriptive Statistics
- Exponential Smoothing
- F-Test Two-Sample for Variances
- Fourier Analysis
- Histogram

Click Ok

Then screen opens which asks you to select the column and other attributes



Descriptive Statistics

Input

Input Range: $AS$1:$AS$307512

Grouped By: ● Columns ○ Rows

□ Labels in first row

Output options

○ Output Range:

○ New Worksheet Ply:

● New Workbook

☑ Summary statistics

□ Confidence Level for Mean: 95 %

□ Kth Largest: 1

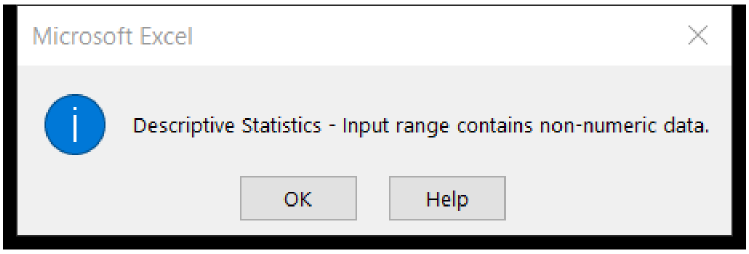
□ Kth Smallest: 1

Select summary Statistics and click ok

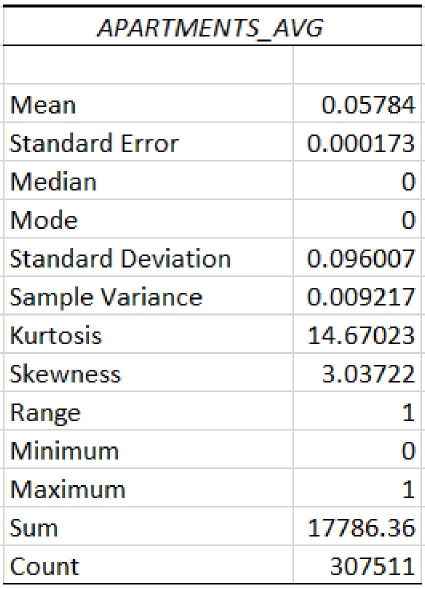
Descriptive Statistics is used to **summarize** a given **data** found from any study. It can provide **basic information** and the **internal relationship** between the **variables** in a dataset.

We have to convert the general formatted data to number format for the descriptive statistics too to work as it does not work on non numeric data.

For that we have to go to the home tab and change the column property from general to number format, then proceed with the analysis



Descriptive Statistics - Input range contains non-numeric data.



APARTMENTS_AVG	
Mean	0.05784
Standard Error	0.000173
Median	0
Mode	0
Standard Deviation	0.096007
Sample Variance	0.009217
Kurtosis	14.67023
Skewness	3.03722
Range	1
Minimum	0
Maximum	1
Sum	17786.36
Count	307511

Now select the colums with null values and fill it by mean, median or mode as per the requirement. Continue this process with all the datasets.

OUTLIERS in EXCEL

1. Review your data
2. Sort your data values
3. Analyse your values
4. Identify your data quartiles
5. Define the interquartile range
6. Calculate the upper and lower bound
7. Remove your outliers

What is an Outlier?

An outlier is a data point within a data set that lies outside of the range of most of the other data points. For example in the date set of ages 12, 13, 15, 16, 52, 14, and 11, you can see that 52 is the outlier age. This is because the other ages fall within the range of 11 to 16 and 52 is outside of that range.

It's important for statisticians to be able to identify outliers like this since they can dramatically alter their calculations. In this example, if you include the outlier the mean average age is 19. If you exclude the outlier, the mean is 13.5.

You need to investigate any outliers carefully before removing them. It's possible the outliers are simple mistakes that should be excluded. Alternatively, those outliers may contain important statistical information such as a new trend or a significant insight that needs to be considered.

How to find outliers?

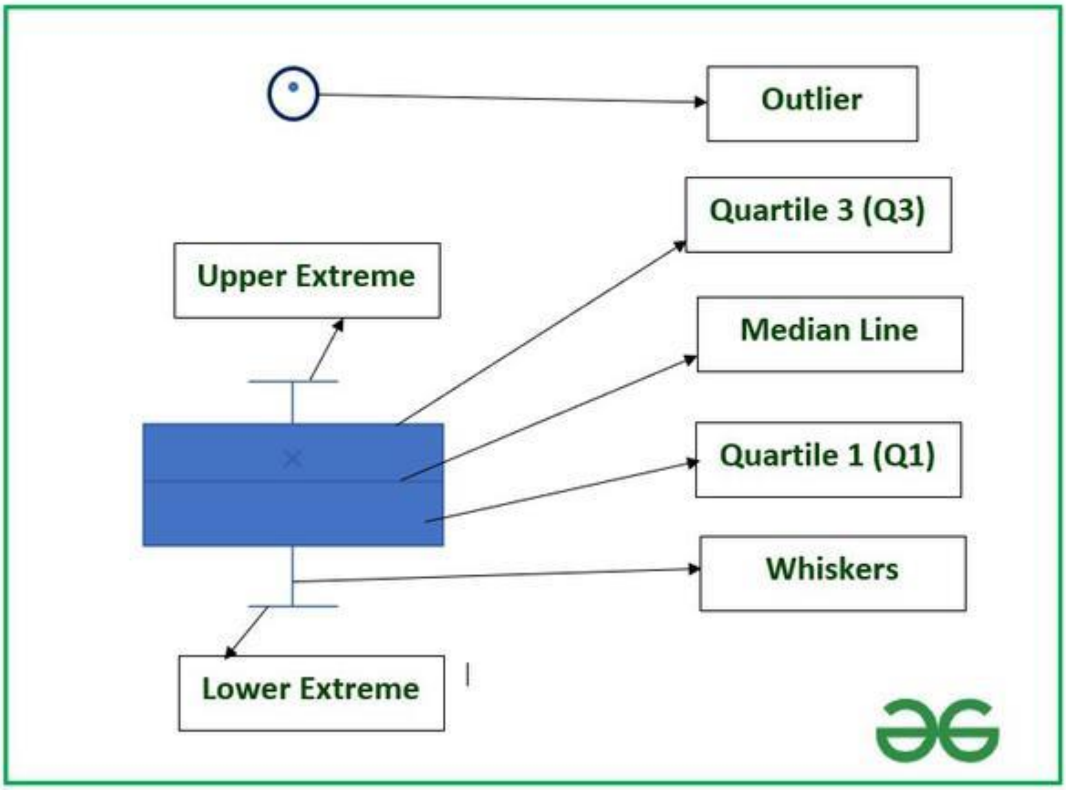
Statistical concepts to know outliers

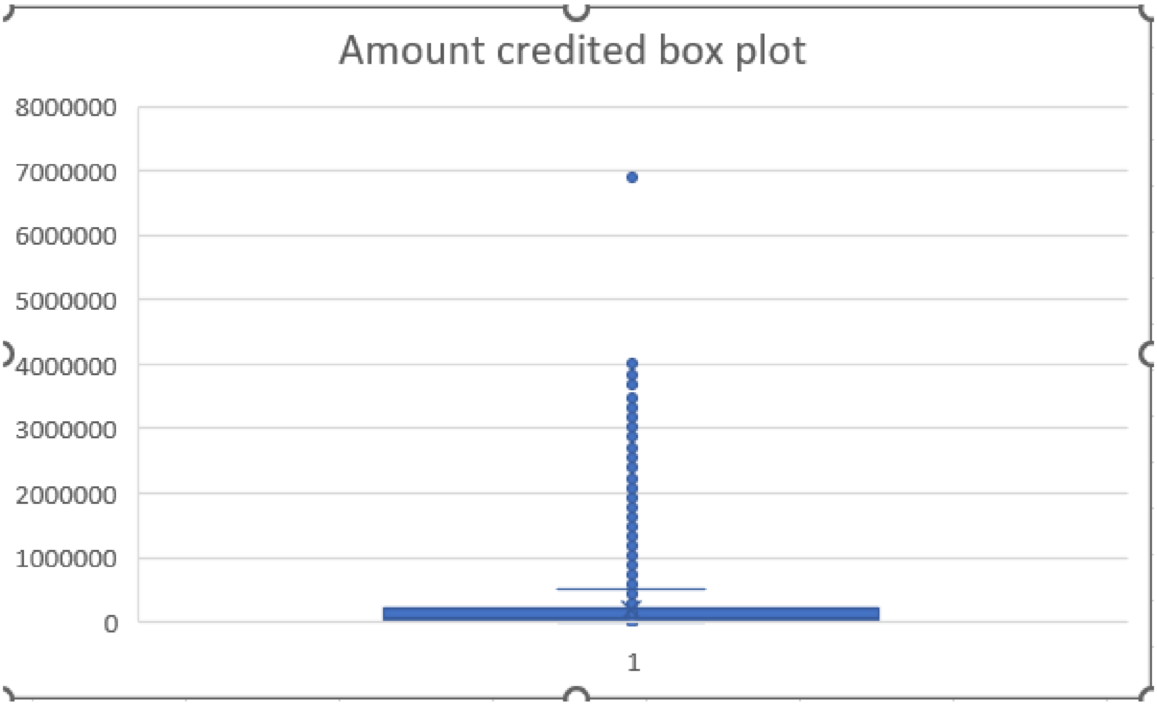
1. **Box Plots** – in the image below you can see that several points exist outside of the box. The box is the central tendency of the data. It is clustered around a middle value. The upper bound line is the limit of the centralization of that data.
2. **Quartiles**– represent how the data is broken up into quarters.

Plotting the box plot for the amount credited and identifying the outliers

A box plot shows 5 values:

- Minimum
- First quartile
- Median
- Third quartile
- Maximum





The above box plot is showing the outliers in the amount credit column similarly we perform the same for all the datasets

What is data imbalance

Imbalanced data refers to those types of datasets where the target class has an uneven distribution of observations, i.e one class label has a very high number of observations and the other has a very low number of observations. We can better understand imbalanced dataset handling with an example.

Let’s assume that XYZ is a bank that issues a credit card to its customers. Now the bank is concerned that some fraudulent transactions are going on and when the bank checks their data they found that for each 2000 transaction there are only 30 Nos of fraud recorded. So, the number of fraud per 100 transactions is less than 2%, or we can say more than 98% transaction is “No Fraud” in nature. Here, the class “No Fraud” is called the majority class, and the much smaller in size “Fraud” class is called the minority class.

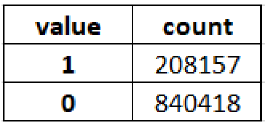
Finding data Imbalance:

Dataset: previous_application.csv

The column is distributed in binomial values which are 0 and 1

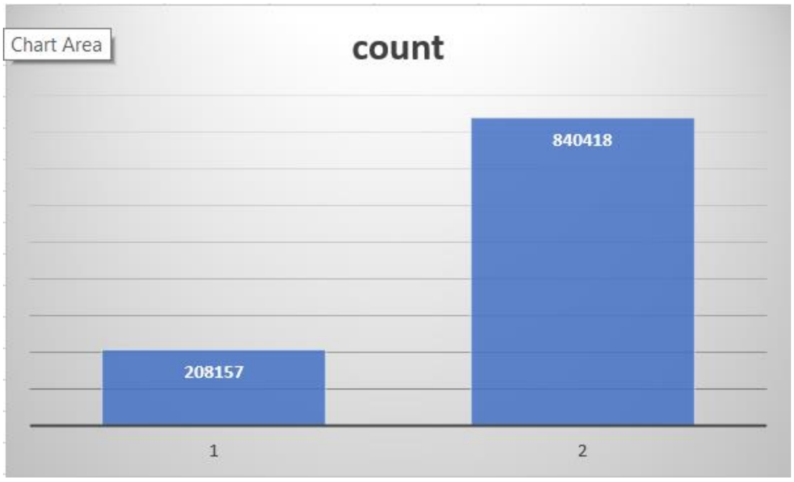
Considering the count of 1s and 0s we use countif() function

Here is the following result:



value	count
1	208157
0	840418

Now we plot a bar chart to see imbalance in the dataset:



Dataset: application.csv

Here we will consider AMT_REQ_CREDIT_BUREAU_YEAR column

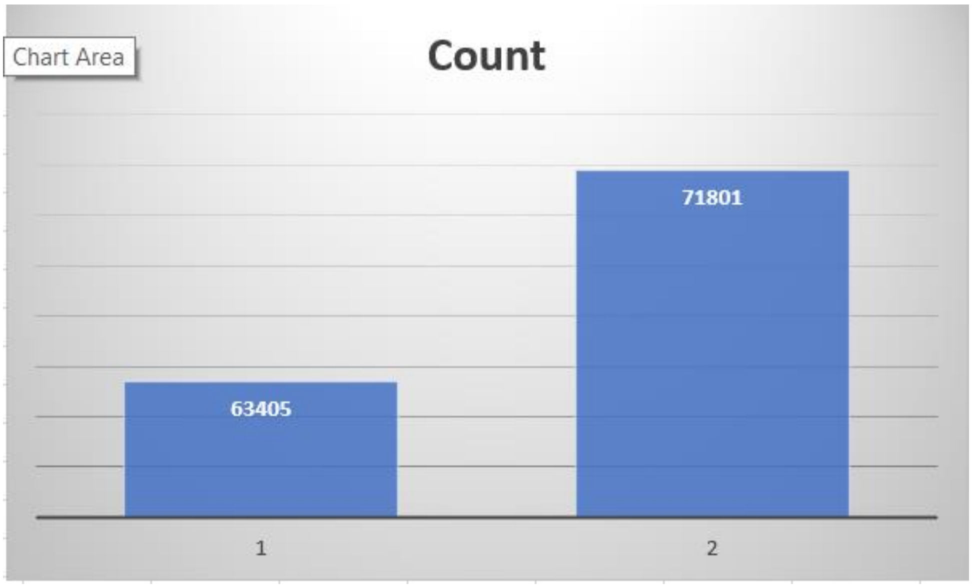
The column is distributed in binomial values which are 0 and 1

Considering the count of 1s and 0s we use countif() function

Here is the following result:

value	Count
1	63405
0	71801

Now we plot a bar chart to see imbalance in the dataset:



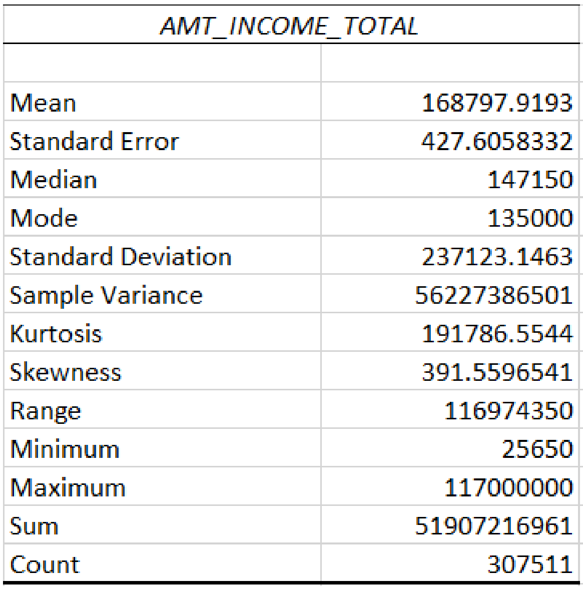
Results of univariate, segmented univariate, bivariate analysis

Univariate Analysis:

Uni means one and variate means variable, so in univariate analysis, there is only one dependable variable. The objective of univariate analysis is to derive the data, define and summarize it, and analyze the pattern present in it. In a dataset, it explores each variable separately. It is possible for two kinds of variables- Categorical and Numerical.

Some patterns that can be easily identified with univariate analysis are Central Tendency (mean, mode and median), Dispersion (range, variance), Quartiles (interquartile range), and Standard deviation.

For finding out univariate analysis of application.csv we will use **descriptive statistics** tool from **Data Analysis** in **Data tab**



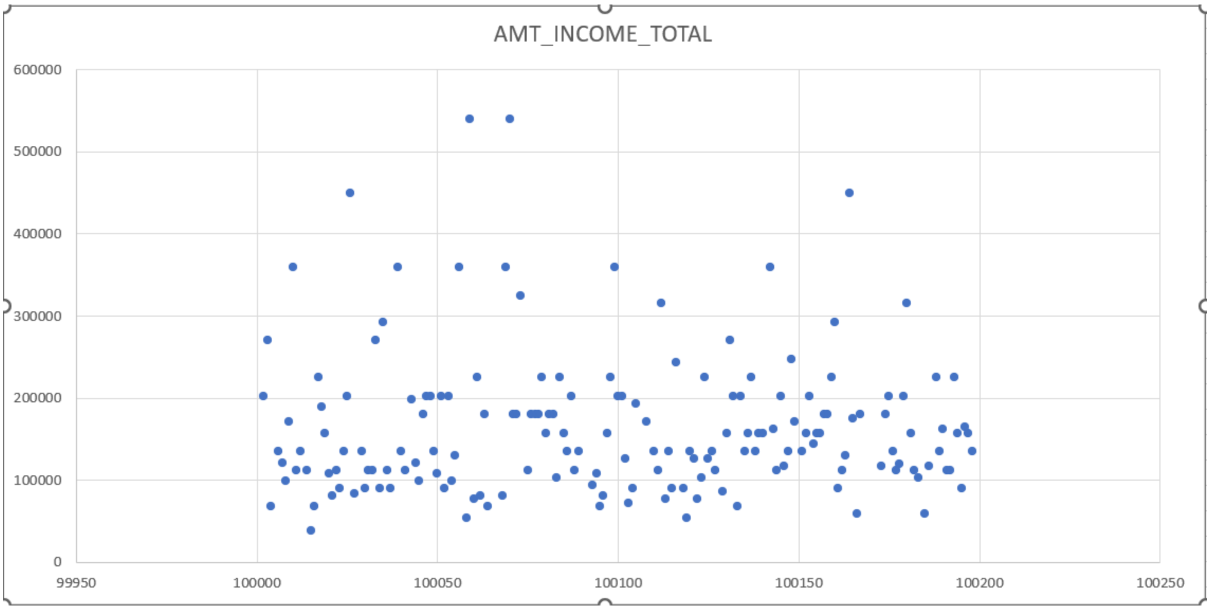
AMT_INCOME_TOTAL	
Mean	168797.9193
Standard Error	427.6058332
Median	147150
Mode	135000
Standard Deviation	237123.1463
Sample Variance	56227386501
Kurtosis	191786.5544
Skewness	391.5596541
Range	116974350
Minimum	25650
Maximum	117000000
Sum	51907216961
Count	307511

Bivariate Analysis:

Bivariate analysis is one of the statistical analysis where two variables are observed. One variable here is dependent while the other is independent. These variables are usually denoted by X and Y. So, here we analyse the changes occured between the two variables and to what extent. Apart from bivariate, there are other two statistical analyses, which are Univariate (for one variable) and Multivariate (for multiple variables).

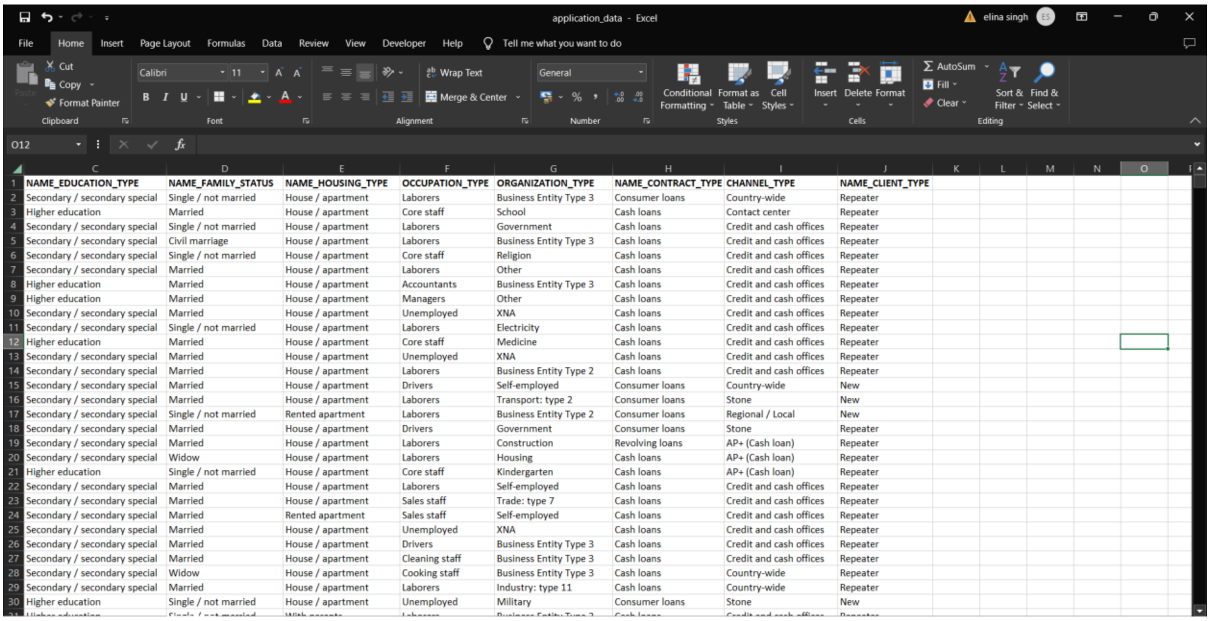
In statistics, we usually interpret the given set of data and make statements and predictions about it. During the research, an analysis attempts to determine the impact and cause in order to conclude the given variables.

Scatter Plot between AMT_INCOME_TOTAL and SK_ID_CURR



Find the top 10 correlation for the Client with payment difficulties and all other cases (Target variable).

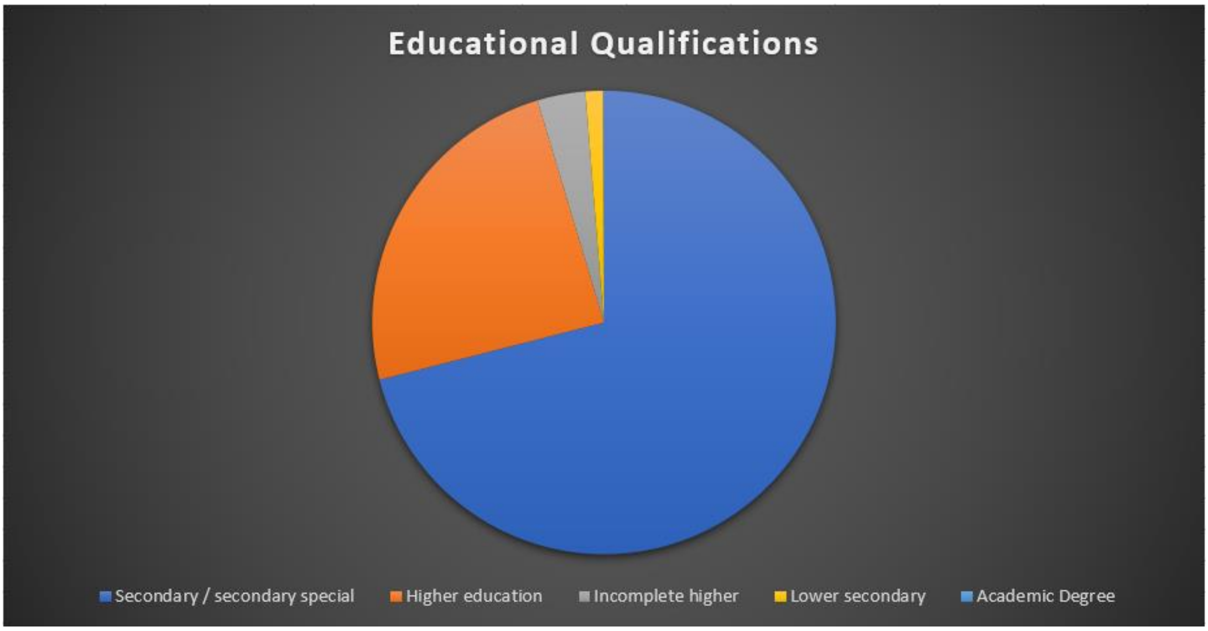
To find 10 top correlations we will include 10 columns from both previous_application.csv dataset and application.csv dataset



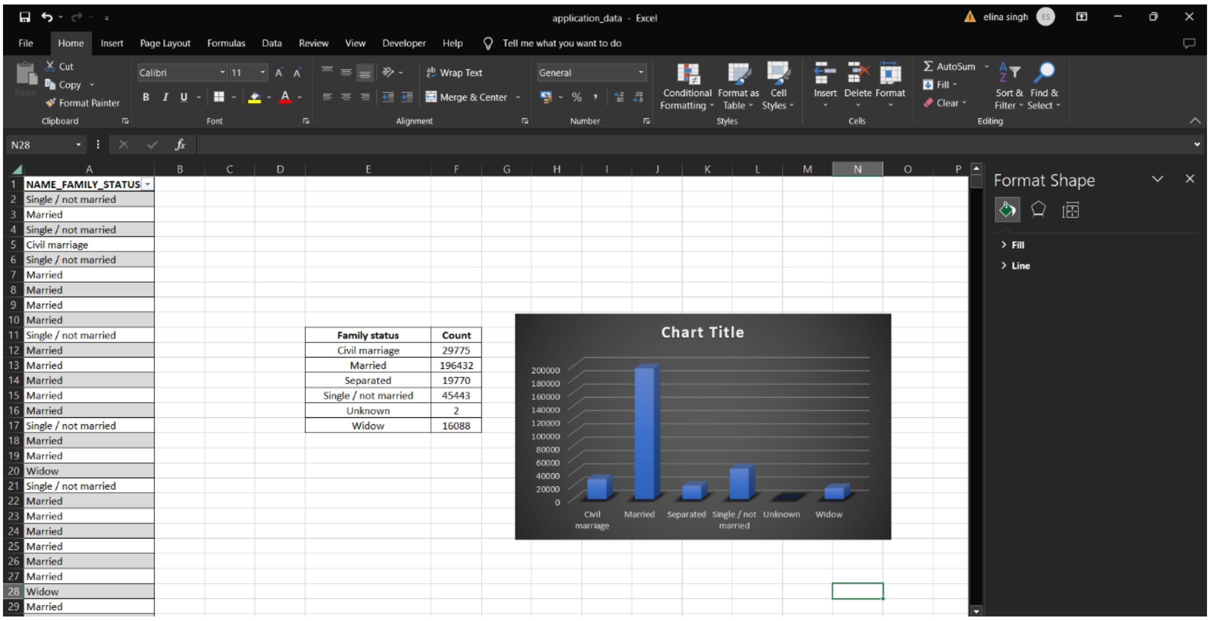
	C	D	E	F	G	H	I	J
1	NAME_EDUCATION_TYPE	NAME_FAMILY_STATUS	NAME_HOUSING_TYPE	OCCUPATION_TYPE	ORGANIZATION_TYPE	NAME_CONTRACT_TYPE	CHANNEL_TYPE	NAME_CLIENT_TYPE
2	Secondary / secondary special	Single / not married	House / apartment	Laborers	Business Entity Type 3	Consumer loans	Country-wide	Repeater
3	Higher education	Married	House / apartment	Core staff	School	Cash loans	Contact center	Repeater
4	Secondary / secondary special	Single / not married	House / apartment	Laborers	Government	Cash loans	Credit and cash offices	Repeater
5	Secondary / secondary special	Civil marriage	House / apartment	Laborers	Business Entity Type 3	Cash loans	Credit and cash offices	Repeater
6	Secondary / secondary special	Single / not married	House / apartment	Core staff	Religion	Cash loans	Credit and cash offices	Repeater
7	Secondary / secondary special	Married	House / apartment	Laborers	Other	Cash loans	Credit and cash offices	Repeater
8	Higher education	Married	House / apartment	Accountants	Business Entity Type 3	Cash loans	Credit and cash offices	Repeater
9	Higher education	Married	House / apartment	Managers	Other	Cash loans	Credit and cash offices	Repeater
10	Secondary / secondary special	Married	House / apartment	Unemployed	XNA	Cash loans	Credit and cash offices	Repeater
11	Secondary / secondary special	Single / not married	House / apartment	Laborers	Electricity	Cash loans	Credit and cash offices	Repeater
12	Higher education	Married	House / apartment	Core staff	Medicine	Cash loans	Credit and cash offices	Repeater
13	Secondary / secondary special	Married	House / apartment	Unemployed	XNA	Cash loans	Credit and cash offices	Repeater
14	Secondary / secondary special	Married	House / apartment	Laborers	Business Entity Type 2	Cash loans	Credit and cash offices	Repeater
15	Secondary / secondary special	Married	House / apartment	Drivers	Self-employed	Consumer loans	Country-wide	New
16	Secondary / secondary special	Married	House / apartment	Laborers	Transport: type 2	Consumer loans	Stone	New
17	Secondary / secondary special	Single / not married	Rented apartment	Laborers	Business Entity Type 2	Consumer loans	Regional / Local	New
18	Secondary / secondary special	Married	House / apartment	Drivers	Government	Consumer loans	Stone	Repeater
19	Secondary / secondary special	Married	House / apartment	Laborers	Construction	Revolving loans	AP+ (Cash loan)	Repeater
20	Secondary / secondary special	Widow	House / apartment	Laborers	Housing	Cash loans	AP+ (Cash loan)	Repeater
21	Higher education	Single / not married	House / apartment	Core staff	Kindergarten	Cash loans	AP+ (Cash loan)	Repeater
22	Secondary / secondary special	Married	House / apartment	Laborers	Self-employed	Cash loans	Credit and cash offices	Repeater
23	Secondary / secondary special	Married	House / apartment	Sales staff	Trade: type 7	Cash loans	Credit and cash offices	Repeater
24	Secondary / secondary special	Married	Rented apartment	Sales staff	Self-employed	Cash loans	Credit and cash offices	Repeater
25	Secondary / secondary special	Married	House / apartment	Unemployed	XNA	Cash loans	Credit and cash offices	Repeater
26	Secondary / secondary special	Married	House / apartment	Drivers	Business Entity Type 3	Cash loans	Credit and cash offices	Repeater
27	Secondary / secondary special	Married	House / apartment	Cleaning staff	Business Entity Type 3	Cash loans	Credit and cash offices	Repeater
28	Secondary / secondary special	Widow	House / apartment	Cooking staff	Business Entity Type 3	Cash loans	Country-wide	Repeater
29	Secondary / secondary special	Married	House / apartment	Laborers	Industry: type 11	Cash loans	Country-wide	Repeater
30	Higher education	Single / not married	House / apartment	Unemployed	Military	Consumer loans	Stone	New

NAME_EDUCATION_TYPE

Qualification	count
Secondary / secondary special	218391
Higher education	74863
Incomplete higher	10277
Lower secondary	3816
Academic Degree	164



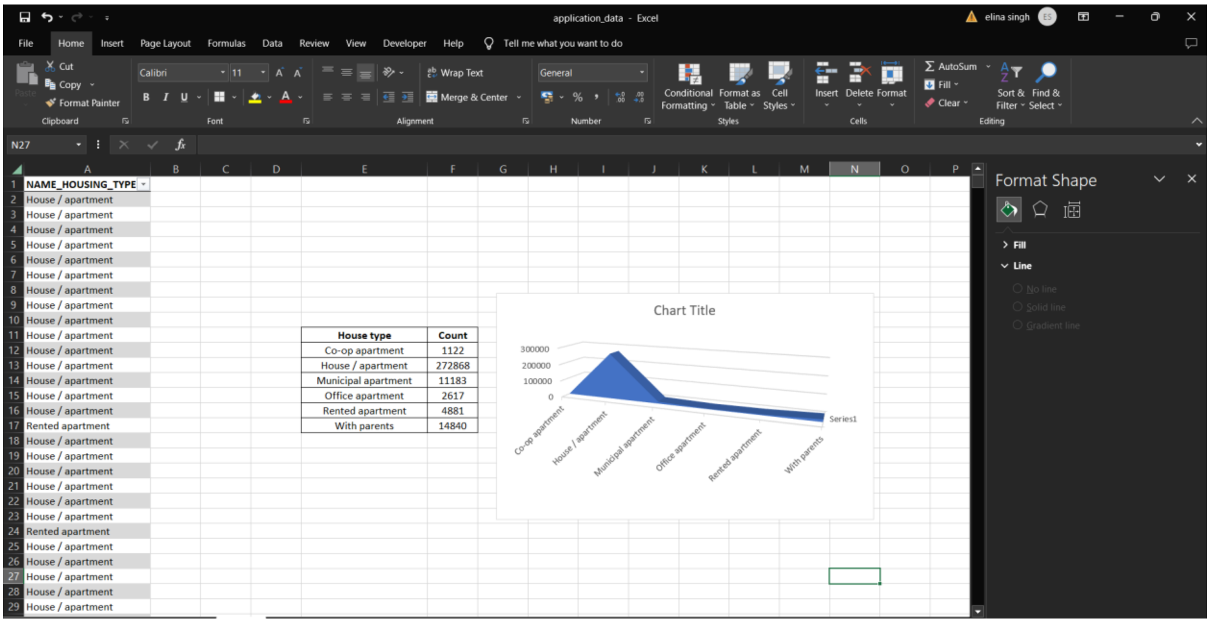
NAME_FAMILY_STATUS



	NAME_FAMILY_STATUS
2	Single / not married
3	Married
4	Single / not married
5	Civil marriage
6	Single / not married
7	Married
8	Married
9	Married
10	Married
11	Single / not married
12	Married
13	Married
14	Married
15	Married
16	Married
17	Single / not married
18	Married
19	Married
20	Widow
21	Single / not married
22	Married
23	Married
24	Married
25	Married
26	Married
27	Married
28	Widow
29	Married

Family status	Count
Civil marriage	29775
Married	196432
Separated	19770
Single / not married	45443
Unknown	2
Widow	16088

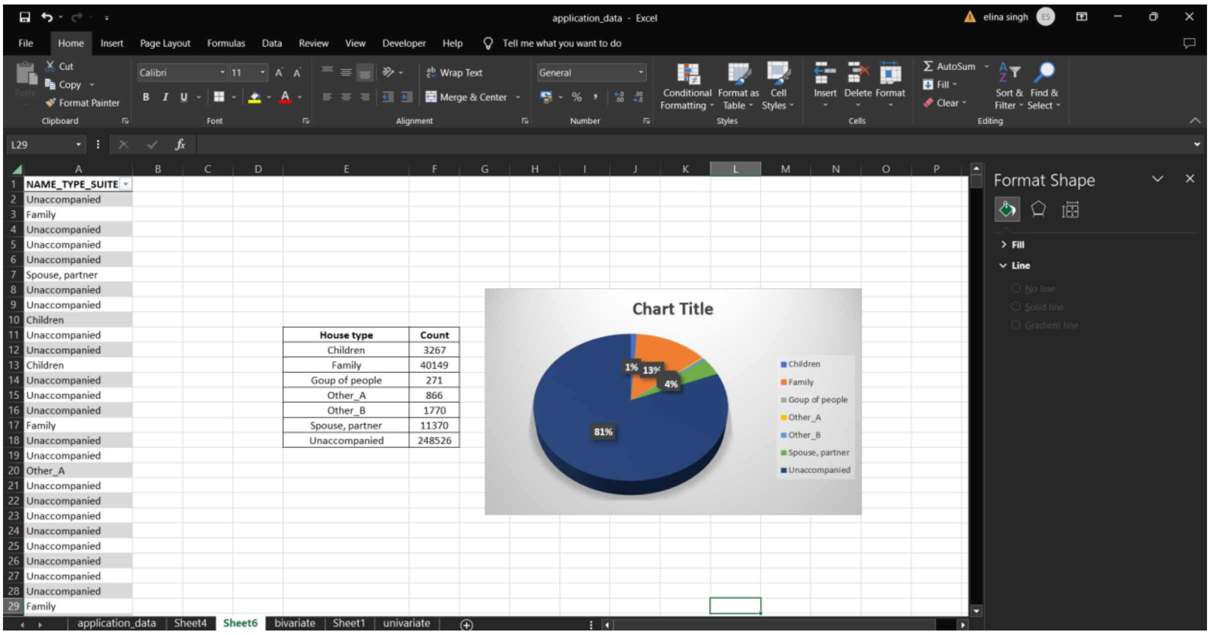
NAME_HOUSING_TYPE



	A
1	NAME_HOUSING_TYPE
2	House / apartment
3	House / apartment
4	House / apartment
5	House / apartment
6	House / apartment
7	House / apartment
8	House / apartment
9	House / apartment
10	House / apartment
11	House / apartment
12	House / apartment
13	House / apartment
14	House / apartment
15	House / apartment
16	House / apartment
17	Rented apartment
18	House / apartment
19	House / apartment
20	House / apartment
21	House / apartment
22	House / apartment
23	House / apartment
24	Rented apartment
25	House / apartment
26	House / apartment
27	House / apartment
28	House / apartment
29	House / apartment

House type	Count
Co-op apartment	1122
House / apartment	272868
Municipal apartment	11183
Office apartment	2617
Rented apartment	4881
With parents	14840

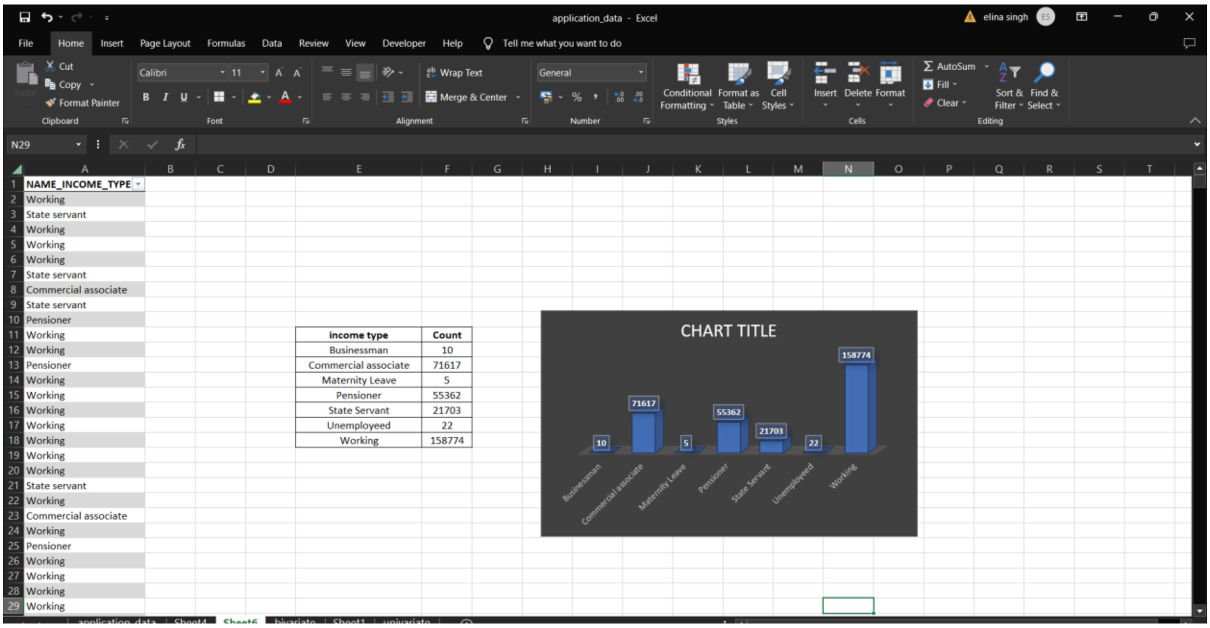
NAME_TYPE_SUITE



	A
1	NAME_TYPE_SUITE
2	Unaccompanied
3	Family
4	Unaccompanied
5	Unaccompanied
6	Unaccompanied
7	Spouse, partner
8	Unaccompanied
9	Unaccompanied
10	Children
11	Unaccompanied
12	Unaccompanied
13	Children
14	Unaccompanied
15	Unaccompanied
16	Unaccompanied
17	Family
18	Unaccompanied
19	Unaccompanied
20	Other_A
21	Unaccompanied
22	Unaccompanied
23	Unaccompanied
24	Unaccompanied
25	Unaccompanied
26	Unaccompanied
27	Unaccompanied
28	Unaccompanied
29	Family

House type	Count
Children	3267
Family	40149
Goup of people	271
Other_A	866
Other_B	1770
Spouse, partner	11370
Unaccompanied	248526

NAME_INCOME_TYPE



	NAME_INCOME_TYPE
2	Working
3	State servant
4	Working
5	Working
6	Working
7	State servant
8	Commercial associate
9	State servant
10	Pensioner
11	Working
12	Working
13	Pensioner
14	Working
15	Working
16	Working
17	Working
18	Working
19	Working
20	Working
21	State servant
22	Working
23	Commercial associate
24	Working
25	Pensioner
26	Working
27	Working
28	Working
29	Working

income type	Count
Businessman	10
Commercial associate	71617
Maternity Leave	5
Pensioner	55362
State Servant	21703
Unemployeed	22
Working	158774

Similarly we continue the same with all correlations and establish relations of variables.

Visualization using different graphs is also shown here.