GROUP C – LENDING CLUB LOAN

Part 1

DEFINITIONS

-Debt-to-Income Ratio
Your debt-to-income ratio (DTI) compares how much you owe each month to how much you earn. Specifically, it's the percentage of your gross monthly income (before taxes) that goes towards payments for rent, mortgage, credit cards, or other debt.

-Public Record
Public records and collections are derogatory items because they reflect financial obligations that were not paid as agreed.

-Revolving Balance
In credit card terms, a revolving balance is the portion of credit card spending that goes unpaid at the end of a billing cycle.
The amount can vary, going up or down depending on the amount borrowed and the amount repaid.

-Revolving Utilization
Revolving utilization measures the amount of revolving credit limits that you are currently using, and it accounts for a large portion of your credit score.

-Mortgage Account
Mortgage Account means an Eligible Account secured by a lien on real estate (e.g., a mortgage or deed of trust) on a 1-6-family residential property or a mixed- use property. A Mortgage Account includes a closed-end mortgage loan and a home equity loan and does not include a home equity line of credit.

-Public Bankruptcies Records
Unless you have your bankruptcy records sealed, any and all documents associated with the bankruptcy are public. The information contained in these cases is all public record.

ANALYSIS

(Ioannis Psomadakis)

Using the “Descriptive Statistics” function of SPSS for the numeric variables that we picked, mainly thevariables “Interest Rate”, “Instalment”, “Revolving Balance” and “Revolving Utilization” we got the following results.

Statistics

		Interest Rate	Installment	Revolving Balance	Revolving Utilization
N	Valid	396028	396026	396006	395741
	Missing	2	4	24	289
Mean		13,6394	431,8473	15844,2756	53,792
Median		13,3300	375,4300	11181,0000	54,800
Mode		10,99	327,34	,00	,0
Std. Deviation		4,47216	250,72386	20592,06900	24,4521
Variance		20,000	62862,456	424033305,53	597,906
Range		25,67	1517,73	1743266,00	892,3
Minimum		5,32	16,08	,00	,0
Maximum		30,99	1533,81	1743266,00	892,3
Percentiles	25	10,4900	250,3300	6025,0000	35,800
	50	13,3300	375,4300	11181,0000	54,800
	75	16,4900	567,3000	19620,0000	72,900

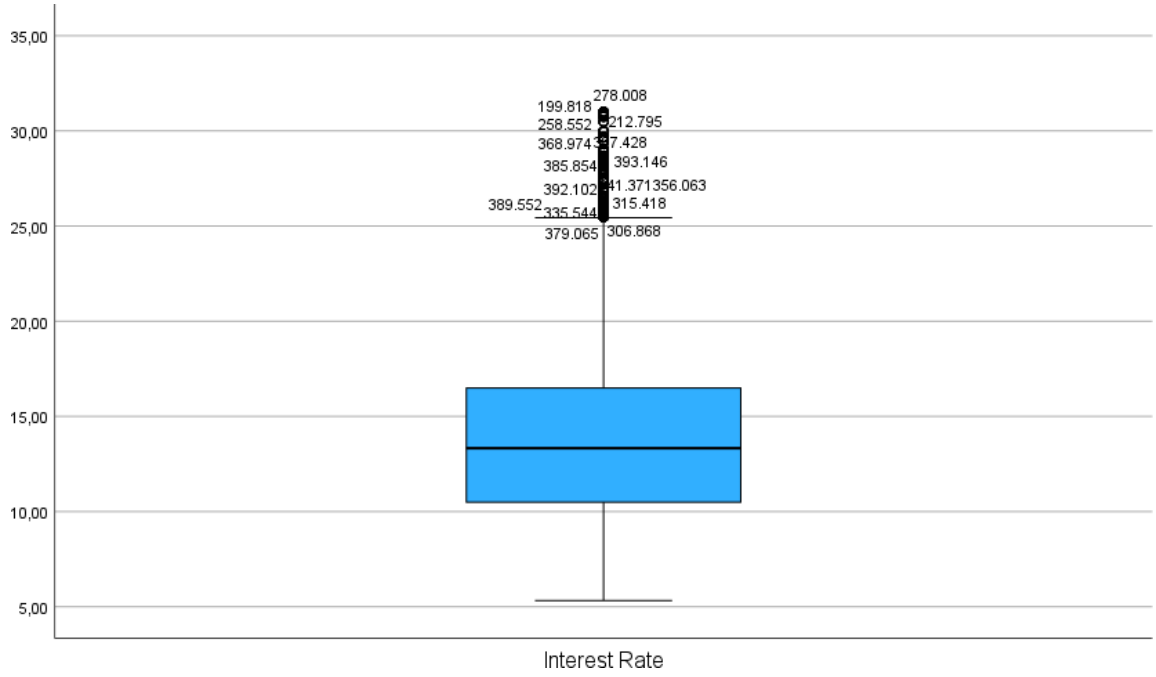
Missing Values

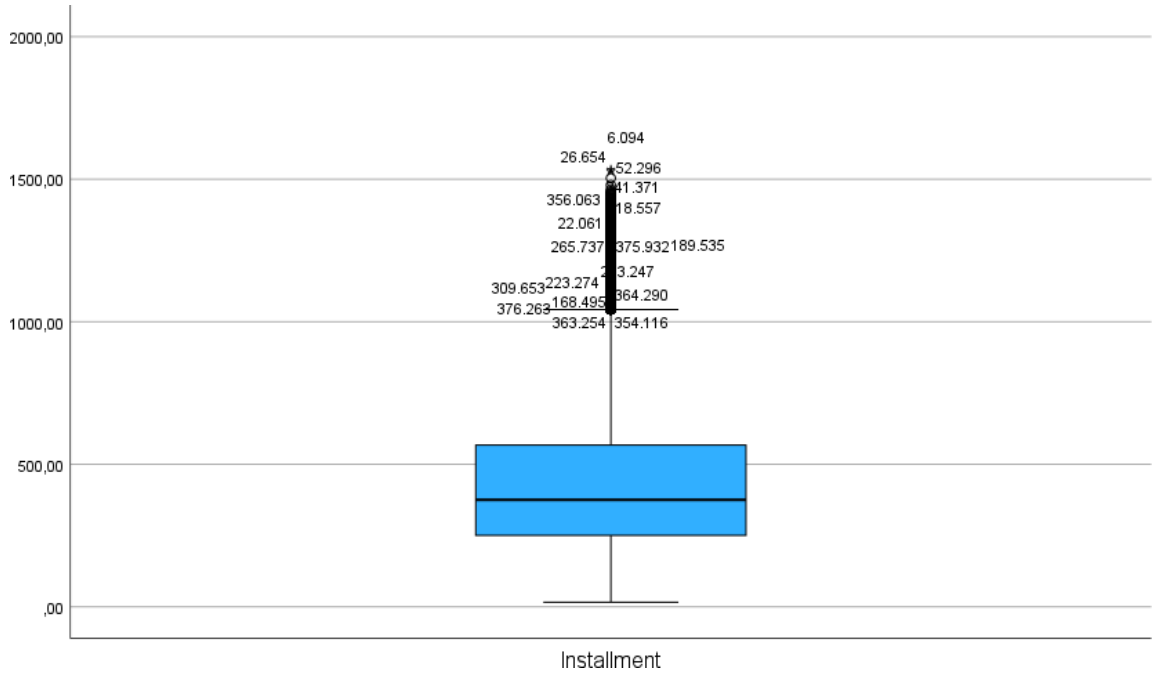
In all cases we can notice that the amount of missing values does not exceed the 10% mark. As a matterof fact, not one of the variables has missing values that exceed 1%. Also, the missing values can be explained by the definition of the variable and/or by the rest of the variables in our data set.

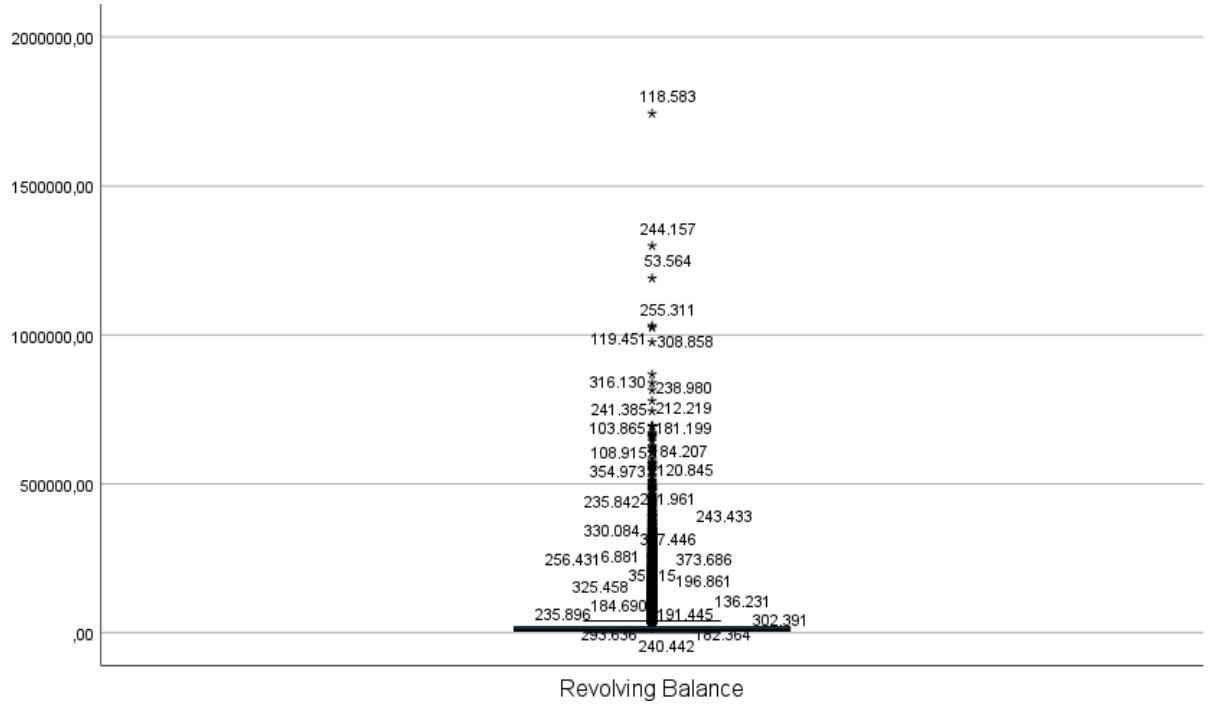
The missing values in the variables of "Interest Rate" and "Instalment" can be explained by the fact that the application hasn't been approved yet, or by the loan being paid of (Missing at random(MAR)). On the other hand, the same can be said for the variables of "Revolving Balance" and "Revolving Utilization". In credit card terms, a ***revolving balance*** is the portion of credit card spending that goes unpaid at the end of a billing cycle. The amount can vary, going up or down depending on the amount borrowed and the amount repaid. So, the missing values can very well mean that there is no portion of credit card spending that goes unpaid. ***Revolving utilization*** measures the amount of revolving credit limits that you are currently using, and it accounts for a large portion of your credit score. So, the missing values can mean that the client is not using any amount of revolving credit limits.

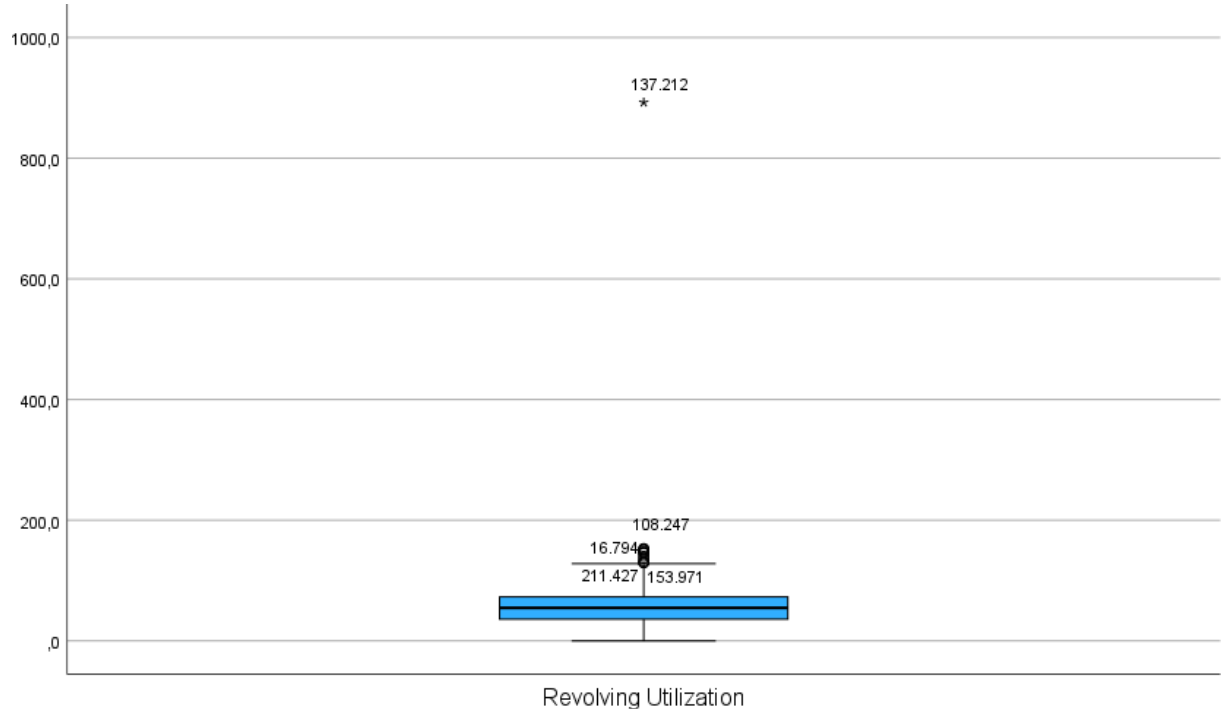
All of the above point to the fact that there is no need for us to further handle the missing values.

Outliers



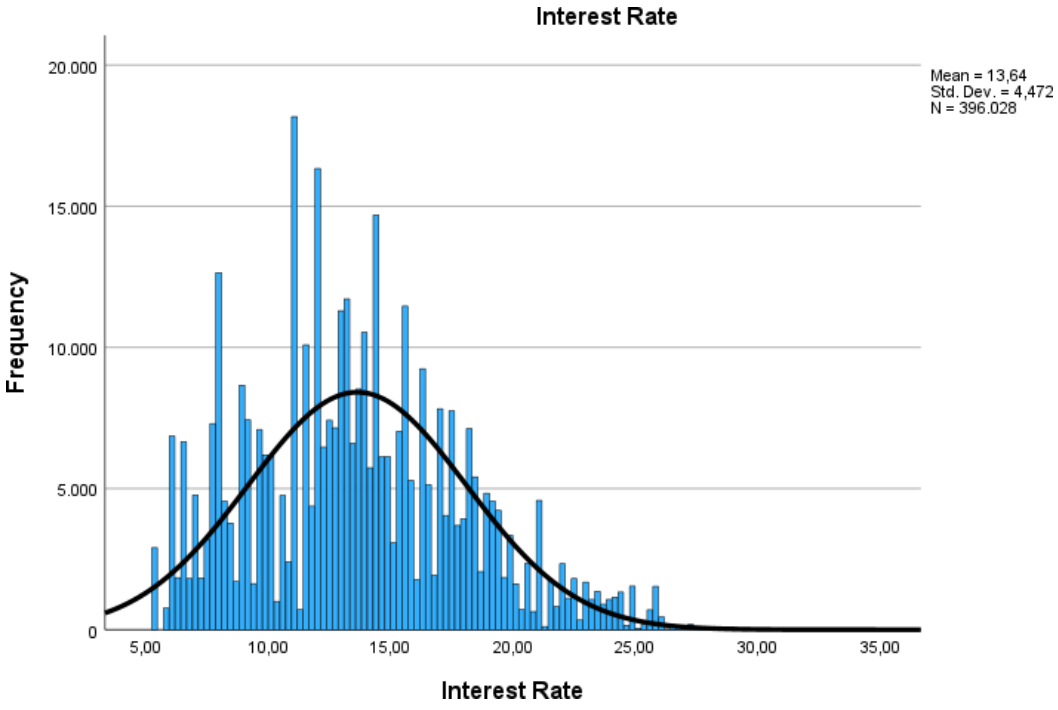

2000,00
1500,00
1000,00
500,00
,00
6.094
26.654
52.296
41.371
356.063
18.557
22.061
265.737
375.932
189.535
223.274
364.290
309.653
376.263
168.495
363.254
354.116
Installment


2000000,00
1500000,00
1000000,00
500000,00
,00
118.583
244.157
53.564
255.311
119.451
308.858
316.130
238.980
241.385
212.219
103.865
181.199
108.915
84.207
354.973
120.845
235.842
1.961
243.433
330.084
7.446
256.431
6.881
373.686
325.458
196.861
184.690
136.231
235.896
191.445
302.391
240.442
Revolving Balance

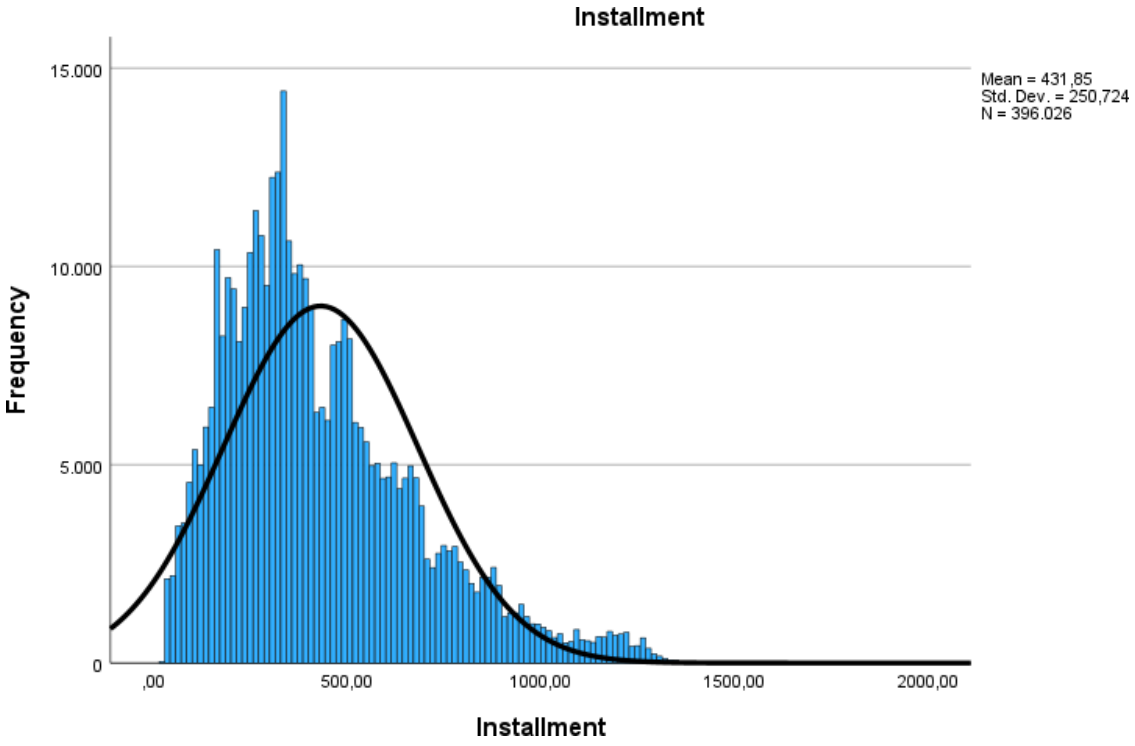


Here, even though we see a lot of instances of values that lie way above the rest of the observations, dueto the definition of each value we know that these values are natural variations in the population and they do not represent measurement errors, data entry errors or poor sampling errors, so these outliers should be left as are in our data set.

Early Assumptions by Looking at Descriptive Statistics



Our interest rate fluctuates mainly between the amounts of 11% and 17% with the mode being 10,99%. This shows us that our bank has exposed itself to clients with relatively bad credit and payment history, a not that good of an income, and a co-signer with a credit score of lower than 750. The percentiles of 25, 50 and 75 being 10,49, 13,33 and 16,49 and the standard deviation being 4,47216 further proves this assumption as more that 80% of our clients have an interest rate of 10% and higher. Looking at the mean(13,6394) and the median(13,33) being relatively the same and the normal distribution line of the histogram, we get hints of our variable following the normal distribution.



The instalment amount that our clients have to pay each month fluctuates mainly between 250 and 600 currency units. The mean is 431,85 and the median is 375,43, while the mode is 327,34. This means that the majority of our clients instalments remain relatively low when taking into consideration the high interest rates that the majority of their contracts have implemented. The percentiles of 25, 50 and 75 being 250,33, 375,43 and 567,30 and the standard deviation being 250,72386 further proves this assumption as more that 80% of our clients have an instalment of 570 currency units and lower. This drives us to make the assumption that maybe their income is relatively high or the loan amount is not that high or that the pay-out period spans over a long period of time.

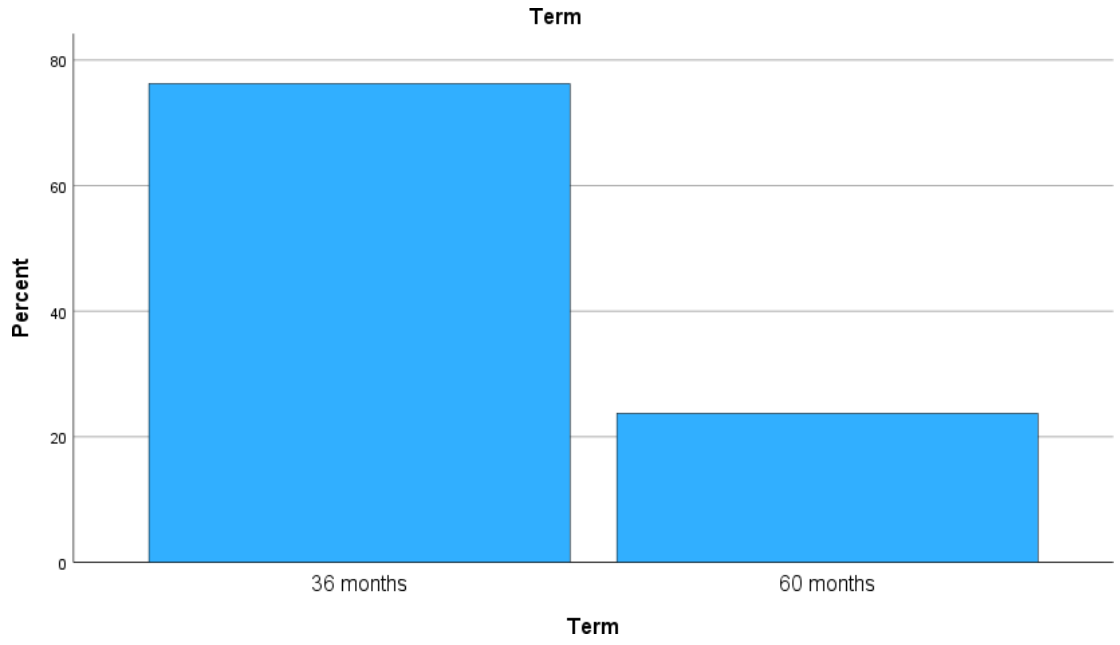
Statistics

Term

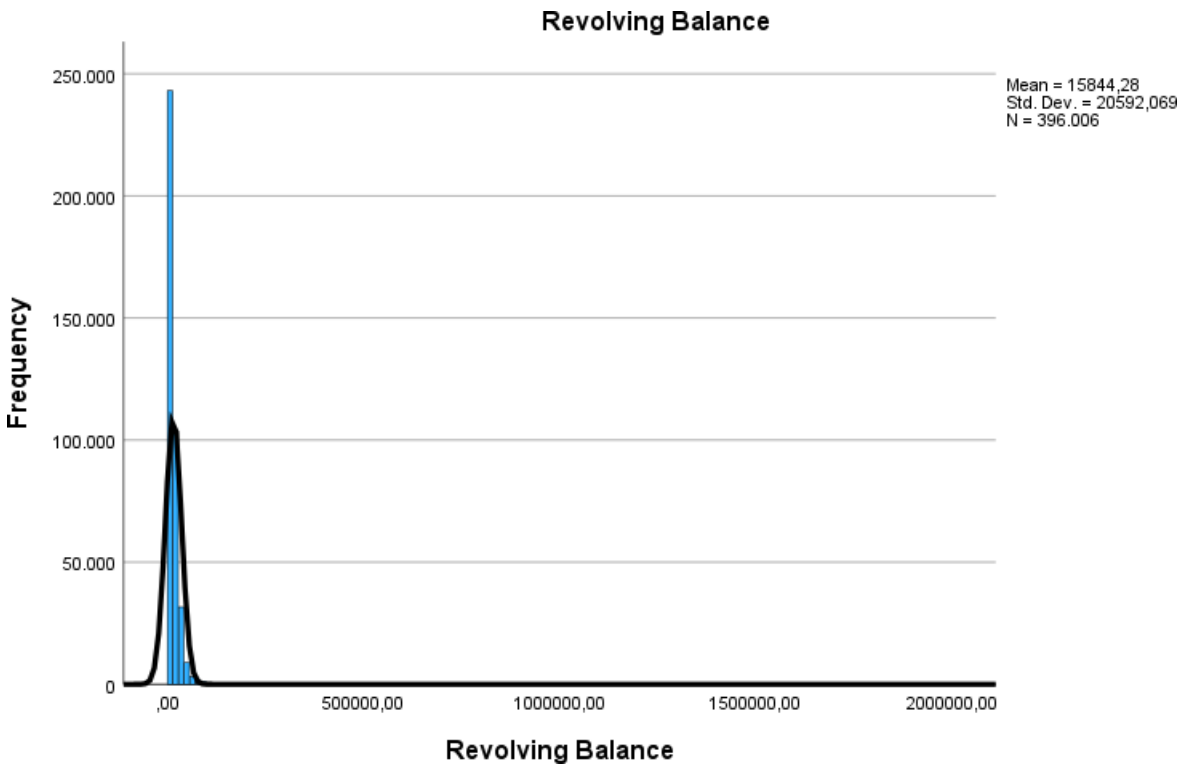
N	Valid	396028
	Missing	2

Term

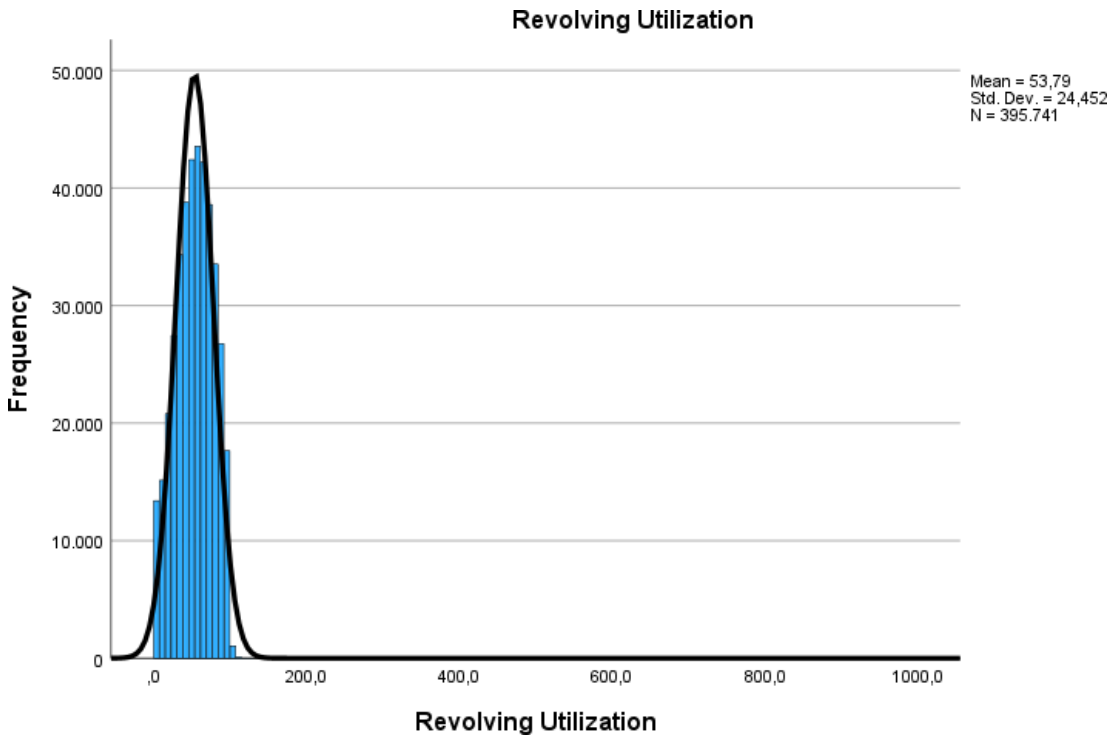
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	36 months	302003	76,3	76,3	76,3
	60 months	94025	23,7	23,7	100,0
	Total	396028	100,0	100,0	
Missing	99	2	,0		
Total		396030	100,0		



However this comes in contrast to the fact that 76% of our loans have a term time of 36 months(3 years) while only the 24% of our loans have a term time of 60 months(5 years). So, we can draw the conclusion that the relatively high interest rates and low instalments are not the result of high term times.



The mode statistic of the variable of revolving balance and the histogram above shows us that most of our clients have a revolving balance of 0. However, this is disrupted on some occasions where our clients have very big amounts in their revolving balance. This can be noticed by the value of the standard deviation being 20592,069 and the mean being 15844,28. Also, the value of the median being 1181 means that most values are relatively low currency-wise. From the above we can draw the assumption that most of our clients have 0 amount in their revolving balance and the amount that goes unpaid is mostly 0 which is very good news for the well-being of our bank.



Here, once again we see that the vast majority of our clients have a relatively low revolving utilization. The mean is 53,79 and the median is 54,8 while the mode is 0. The standard deviation is 24,452 which means that most observations are gathered around the mean. This fact when combined with the histogram and the normal distribution line that we see above makes us feel that there is a high chance that our data follow the normal distribution.

Crosstabulation

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Term * Interest Rate	396026	100,0%	4	0,0%	396030	100,0%

Term * Interest Rate Crosstabulation

Count

		Interest Rate																	
		5,32	5,42	5,79	5,93	5,99	6,00	6,03	6,17	6,24	6,39	6,49	6,54	6,62	6,68	6,76	6,89	6,91	6,92
Term	36 months	2439	465	327	431	278	49	6291	193	1184	656	2358	208	4040	574	136	1098	201	1157
	60 months	1	0	6	0	0	21	0	27	0	0	16	35	0	0	3	0	41	11
Total		2440	465	333	431	278	70	6291	220	1184	656	2374	243	4040	574	139	1098	242	1168

8,70	8,88	8,90	8,94	8,99	9,01	9,07	9,16	9,17	9,20	9,25	9,32	9,33	9,38	9,45	9,49	9,51	9,62	9,63	9,64
12	98	7722	242	240	7	11	717	5463	10	434	154	13	19	27	1246	5	90	286	13
0	29	297	0	22	0	0	70	645	0	79	0	0	0	0	167	0	62	23	0
12	127	8019	242	262	7	11	787	6108	10	513	154	13	19	27	1413	5	152	309	13

9,67	9,70	9,71	9,75	9,76	9,80	9,83	9,88	9,91	9,96	9,99	10,00	10,01	10,08	10,14	10,15	10,16	10,20	10,25	10,28
3095	5	1518	786	901	261	7	142	425	15	4398	164	5	35	7	2324	3222	16	175	16
50	0	68	116	76	44	0	0	47	0	850	40	0	0	0	203	127	0	0	0
3145	5	1586	902	977	305	7	142	472	15	5248	204	5	35	7	2527	3349	16	175	16

10,33	10,36	10,37	10,38	10,39	10,46	10,49	10,51	10,59	10,62	10,64	10,65	10,71	10,74	10,75	10,78	10,83	10,91	10,95	10,96
7	135	256	134	32	7	1287	17	240	177	1959	437	33	805	871	255	26	14	61	12
0	68	134	41	0	0	295	0	59	9	135	107	0	132	196	50	0	0	0	0
7	203	390	175	32	7	1582	17	299	186	2094	544	33	937	1067	305	26	14	61	12

24,49	24,50	24,52	24,59	24,70	24,74	24,76	24,83	24,89	24,99	25,09	25,11	25,28	25,29	25,44	25,49	25,57	25,65	25,69	25,78
23	350	1	0	1	10	2	4	7	294	6	7	3	23	7	3	104	7	15	20
53	885	2	1	122	16	3	75	192	968	28	7	99	55	14	24	498	14	37	155
76	1235	3	1	123	26	5	79	199	1262	34	14	102	78	21	27	602	21	52	175

25,80	25,83	25,88	25,89	25,99	26,06	26,14	26,24	26,49	26,57	26,77	26,99	27,31	27,34	27,49	27,79	27,88	27,99	28,14	28,18
68	47	25	23	16	7	7	2	12	24	26	13	16	17	15	3	9	1	11	4
482	359	63	295	214	199	8	14	24	46	132	38	106	52	39	26	70	14	27	23
550	406	88	318	230	206	15	16	36	70	158	51	122	69	54	29	79	15	38	27

28,34	28,49	28,67	28,69	28,88	28,99	29,49	29,67	29,96	29,99	30,49	30,74	30,79	30,84	30,89	30,94	30,99	Total
11	10	7	2	5	15	2	5	0	0	0	1	4	0	0	3	1	302001
26	51	27	12	15	62	10	10	8	7	5	3	5	1	3	0	12	94025
37	61	34	14	20	77	12	15	8	7	5	4	9	1	3	3	13	396026

Via this crosstabulation between the interest rate and the term of the loans we draw the assumption that loans with lower interest rate are being paid in the lowest available term, that being 36 months andas the interest rate increases so does the amount of loans that get the highest available term, that being60 months. This makes sense as a higher interest rate means that the client is less likely to be able to payback their loan, so the bank put itself in higher risk, so the term is higher than that of the client that gets a lower interest rate.

Normality Check

Descriptive Statistics

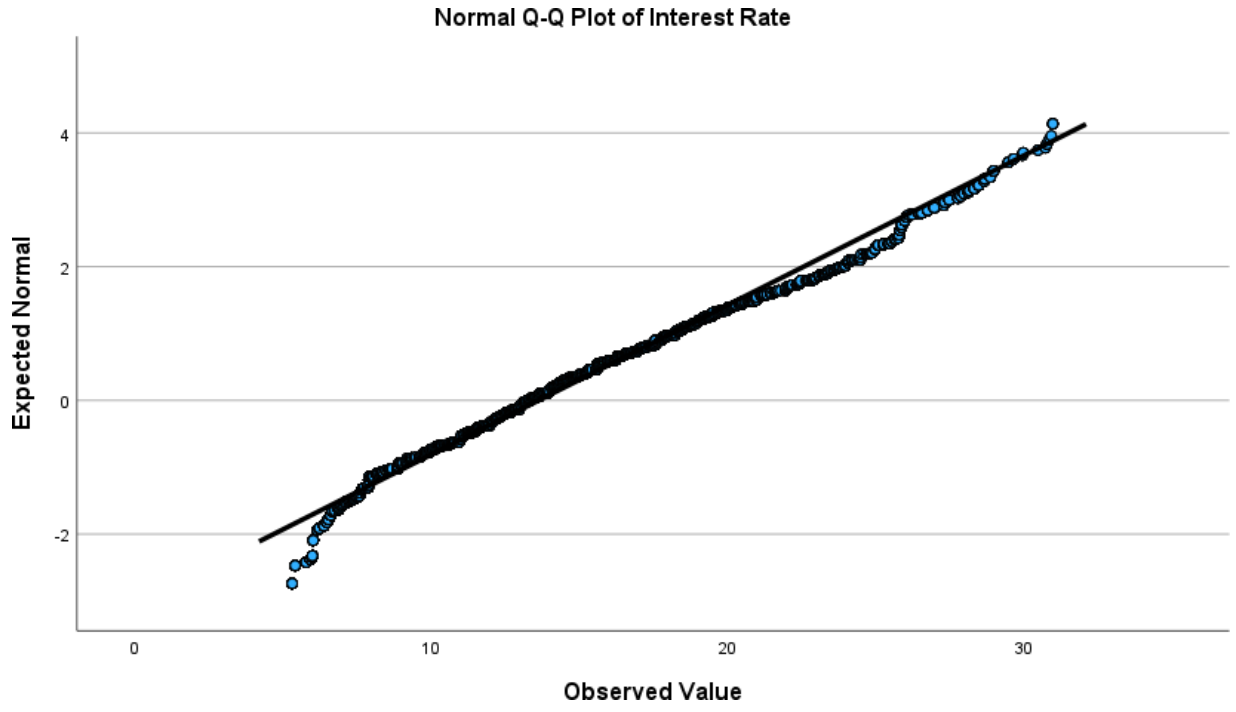
	N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Interest Rate	396028	25,67	5,32	30,99	13,6394	4,47216	20,000	,421	,004	-,144	,008
Installment	396026	1517,73	16,08	1533,81	431,8473	250,72386	62862,456	,984	,004	,784	,008
Revolving Balance	396006	1743266,00	,00	1743266,00	15844,2756	20592,06900	424033305,53	11,728	,004	384,228	,008
Revolving Utilization	395741	892,3	,0	892,3	53,792	24,4521	597,906	-,072	,004	2,712	,008
Valid N (listwise)	395711										

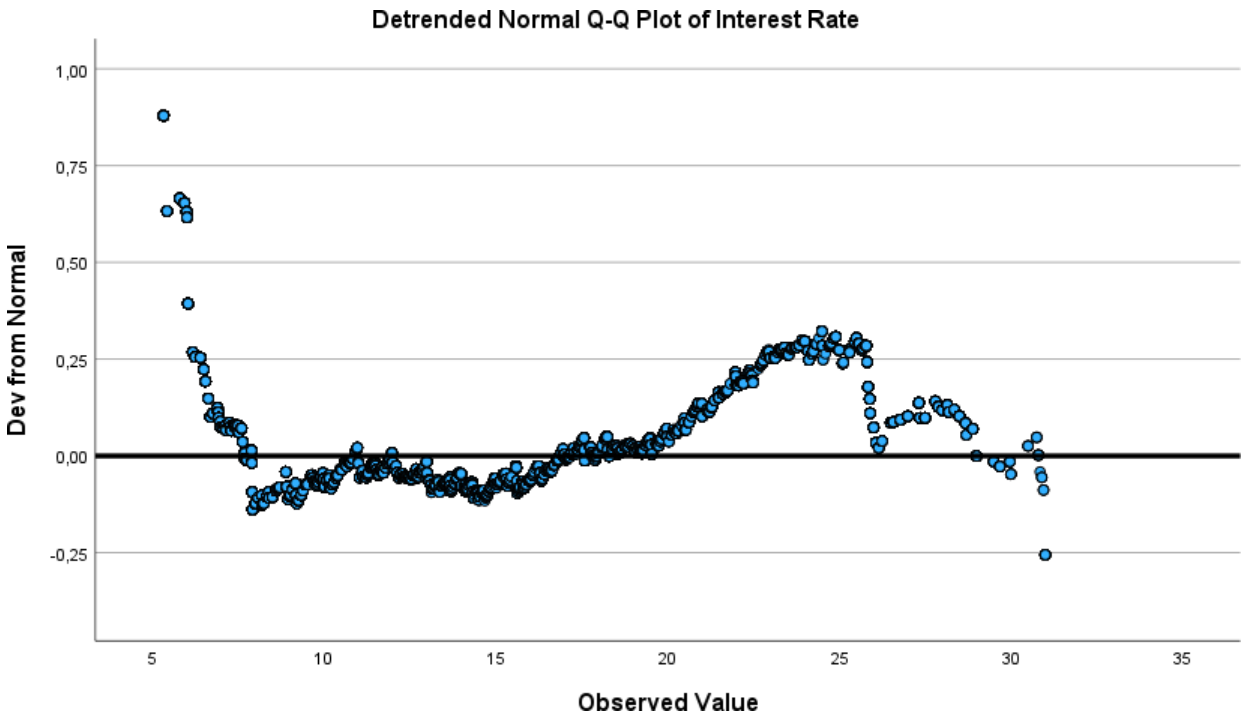
Tests of Normality

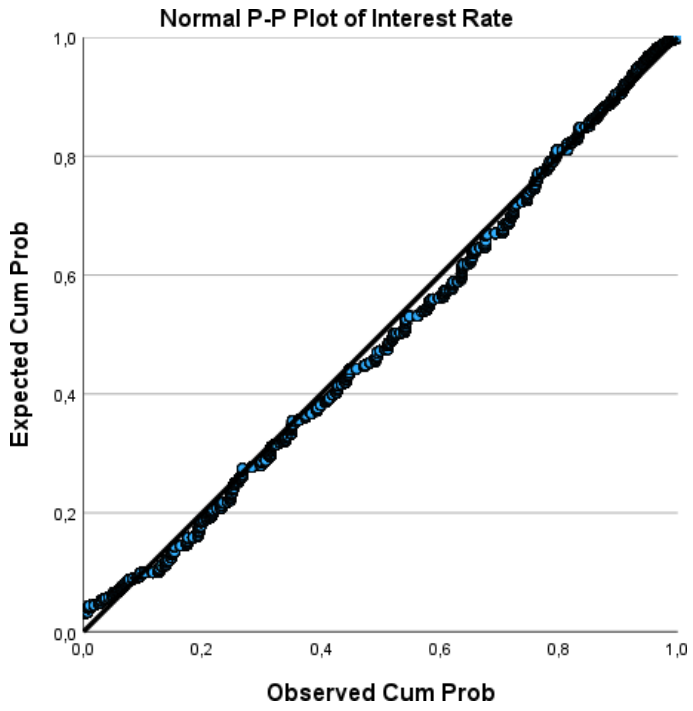
	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
Interest Rate	,046	395711	<,001
Installment	,094	395711	<,001
Revolving Balance	,221	395711	<,001
Revolving Utilization	,035	395711	<,001

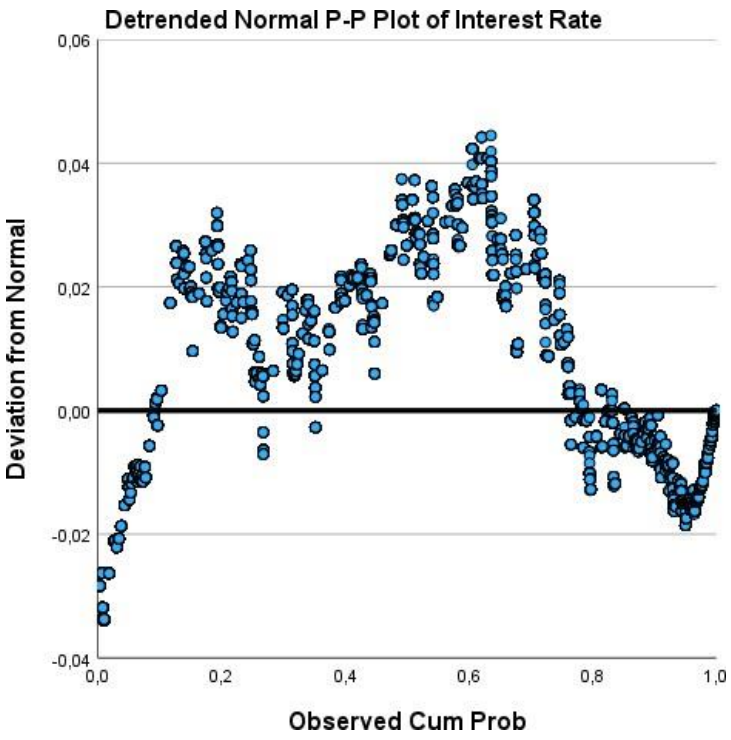
a. Lilliefors Significance Correction

Interest Rate




Detrended Normal Q-Q Plot of Interest Rate
Dev from Normal
1,00
0,75
0,50
0,25
0,00
-0,25
5
10
15
20
25
30
35
Observed Value


Normal P-P Plot of Interest Rate
Expected Cum Prob
1,0
0,8
0,6
0,4
0,2
0,0
0,0
0,2
0,4
0,6
0,8
1,0
Observed Cum Prob



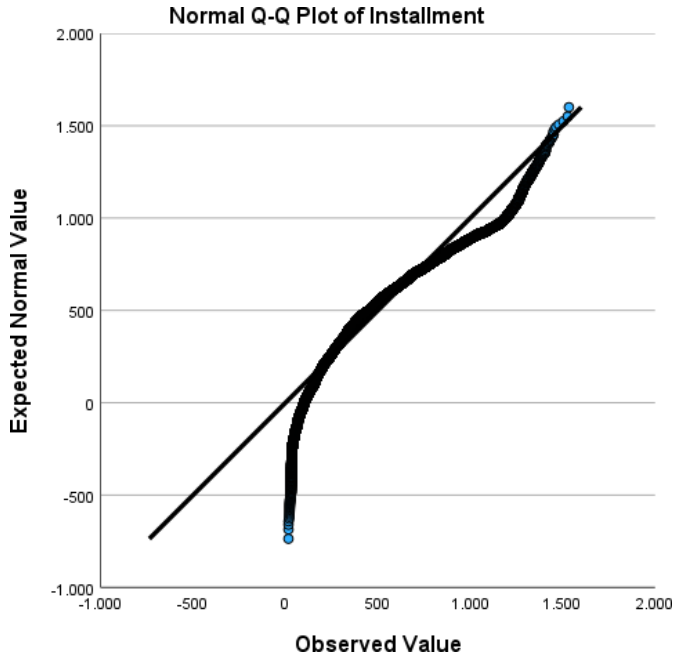
Z-Score:

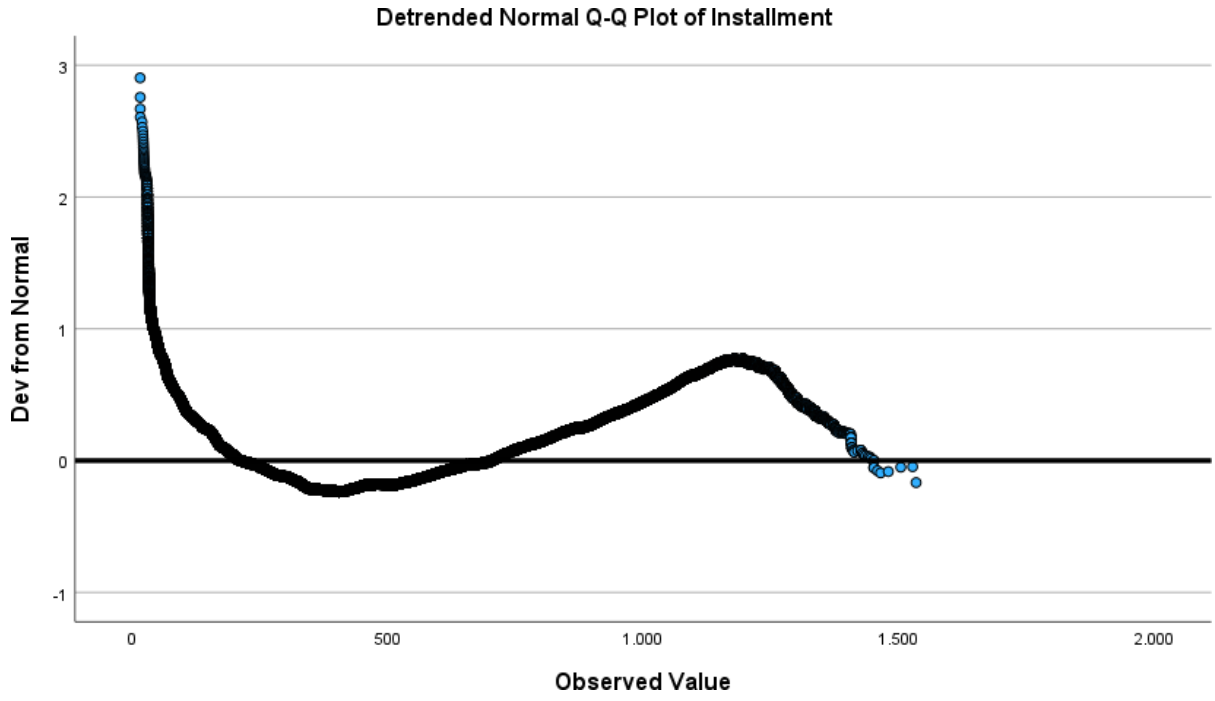
Skewness: (0,421)/(0,004)= 105,25(>1,96)

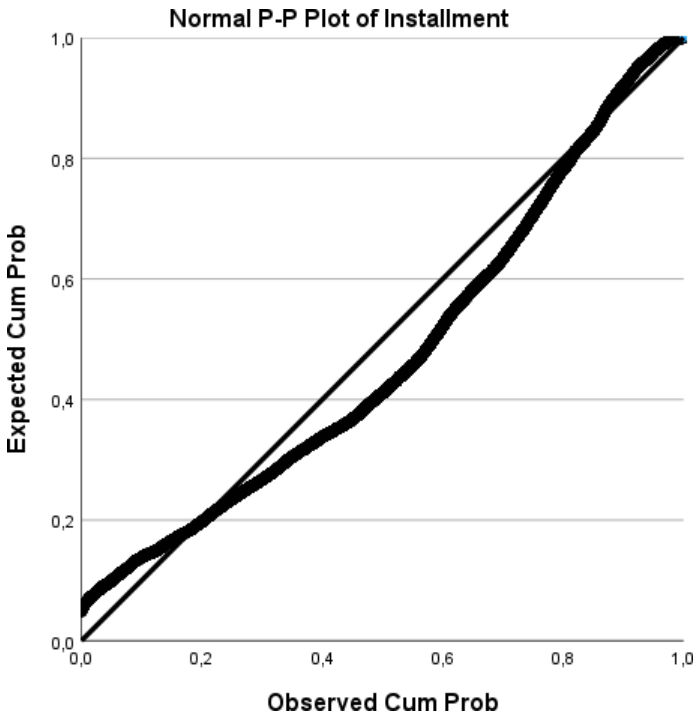
Kurtosis: (-0,144)/(0,008)= -18(<-1,96)

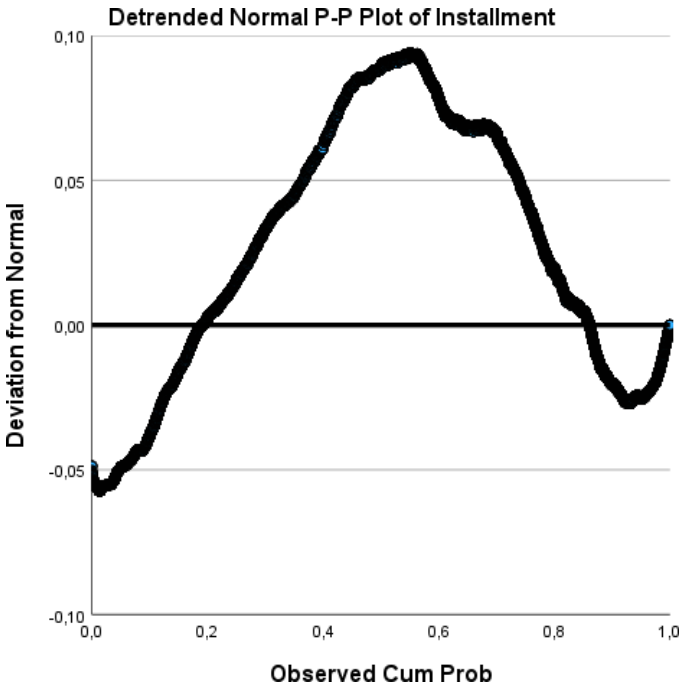
Even though the Q-Q plot, P-P plot, Box plot and Histogram alongside with the values of Skewness and Kurtosis were giving us the hint that our observations of the variable Interest Rate might be following the normal distribution, the Detrended Q-Q and P-P plots alongside with the z-values of Skewness and Kurtosis showed us that they do not follow the normal distribution.

Instalment


Normal Q-Q Plot of Installment
Expected Normal Value
Observed Value
2.000
1.500
1.000
500
0
-500
-1.000
-1.000
-500
0
500
1.000
1.500
2.000


Detrended Normal Q-Q Plot of Installment
Dev from Normal
Observed Value
3
2
1
0
-1
0
500
1.000
1.500
2.000





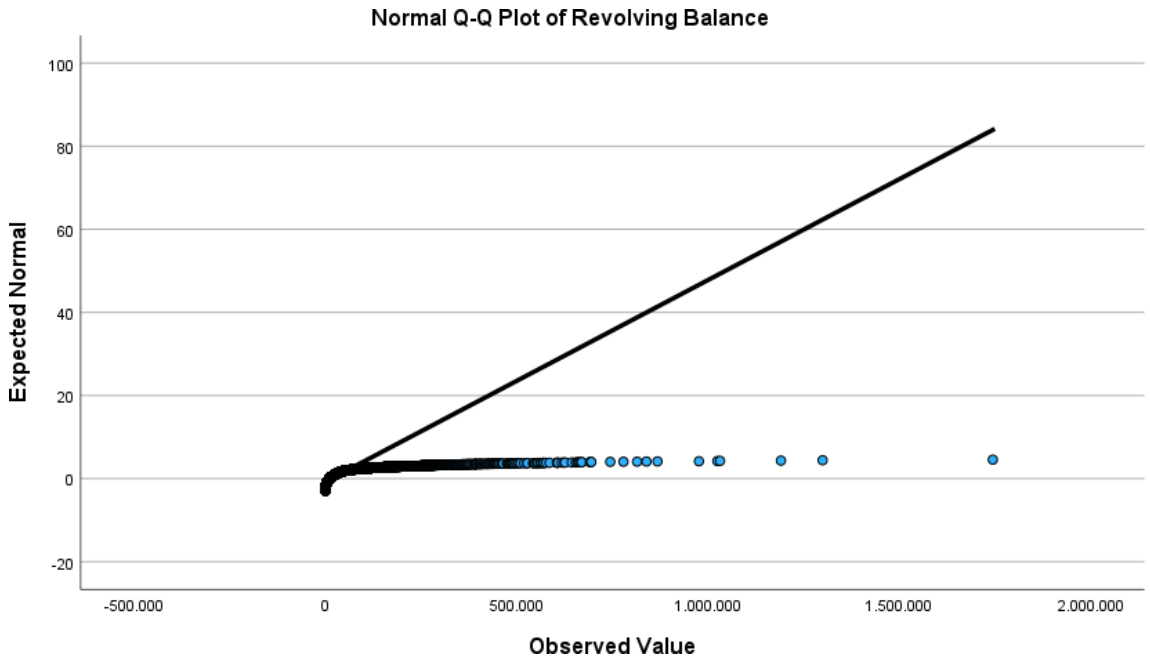
Z-Score:

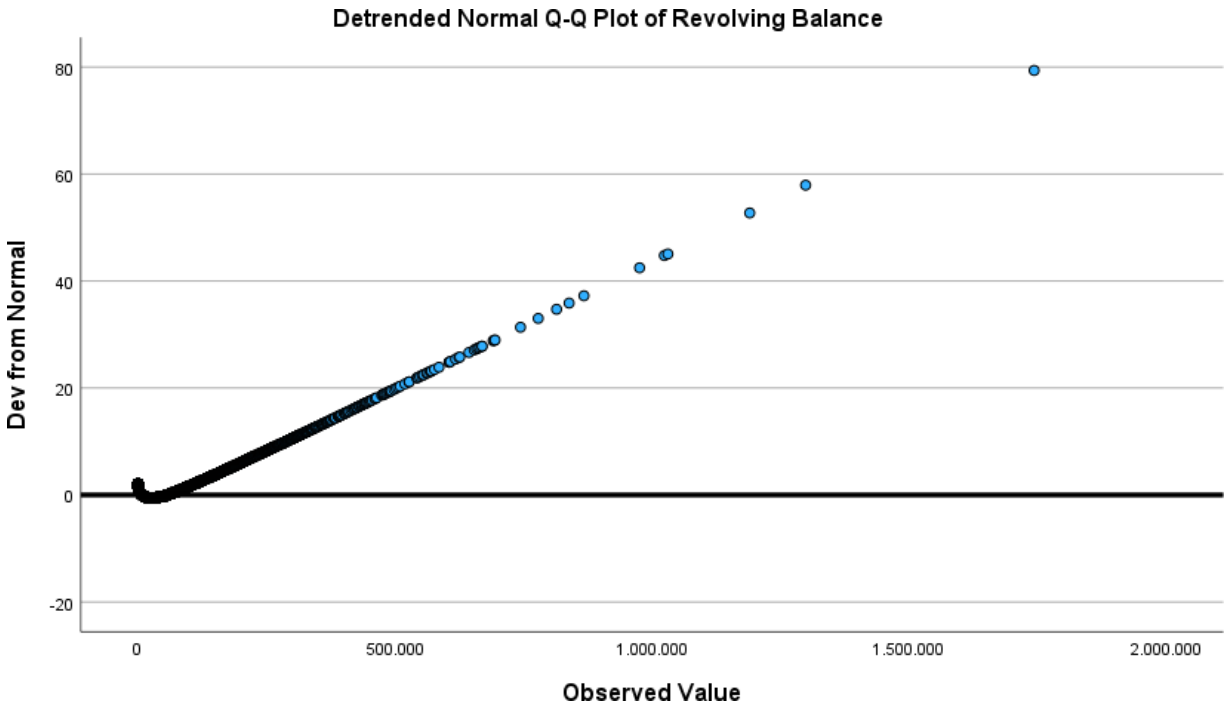
Skewness: (0,984)/(0,004)= 246(>1,96)

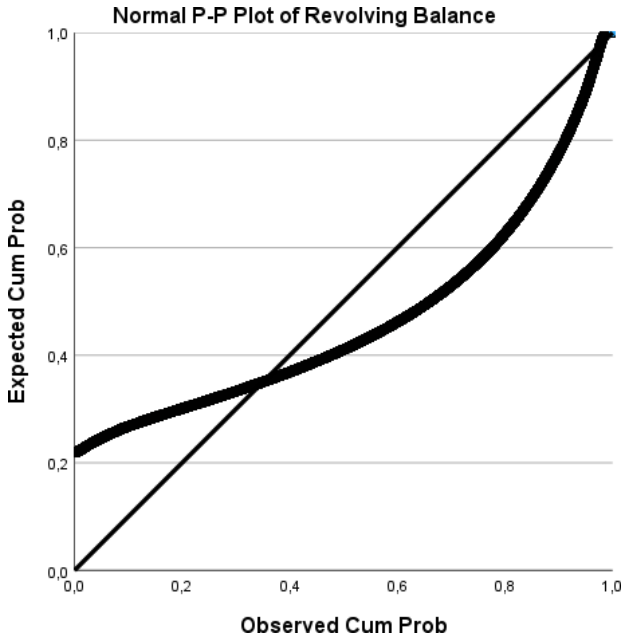
Kurtosis: (0,784)/(0,008)= 98(>1,96)

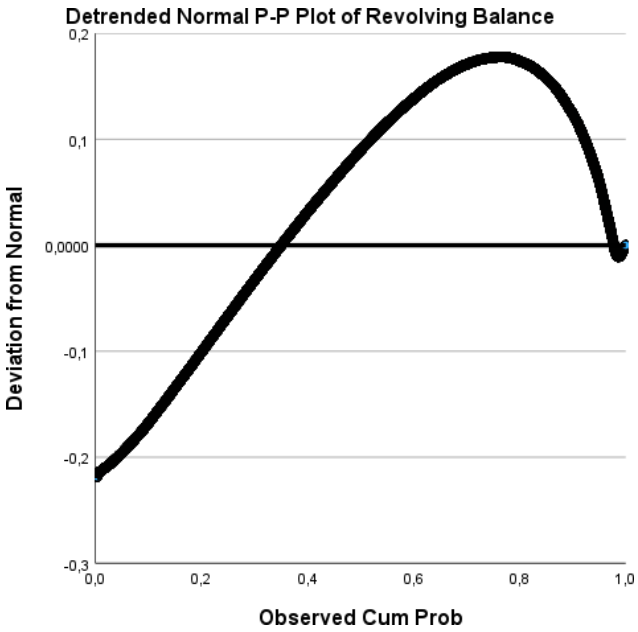
Visual and statistical tests, Kolmogorov-Smirnov, Histogram, and P-P plots and Q-Q plots alongside the z-values of Skewness and Kurtosis, showed that the observations for the variable Instalment weren't normally distributed.

Revolving Balance




Detrended Normal Q-Q Plot of Revolving Balance
Dev from Normal
80
60
40
20
0
-20
0
500.000
1.000.000
1.500.000
2.000.000
Observed Value


Normal P-P Plot of Revolving Balance
Expected Cum Prob
1,0
0,8
0,6
0,4
0,2
0,0
0,0
0,2
0,4
0,6
0,8
1,0
Observed Cum Prob



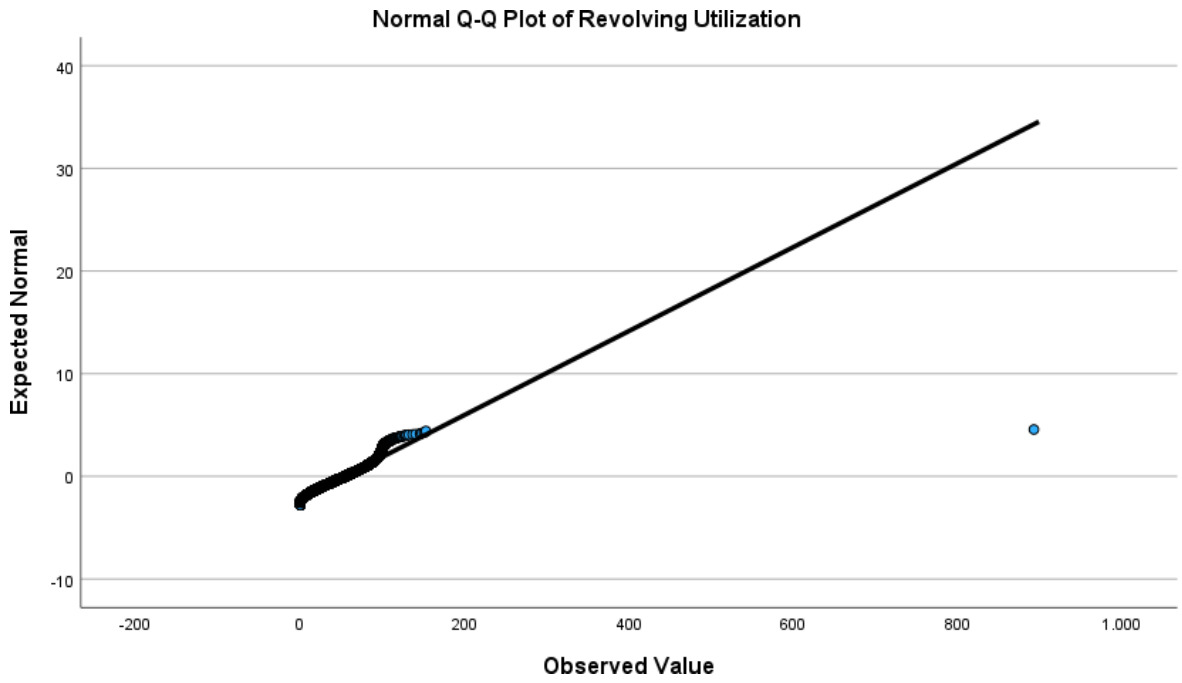
Z-Score:

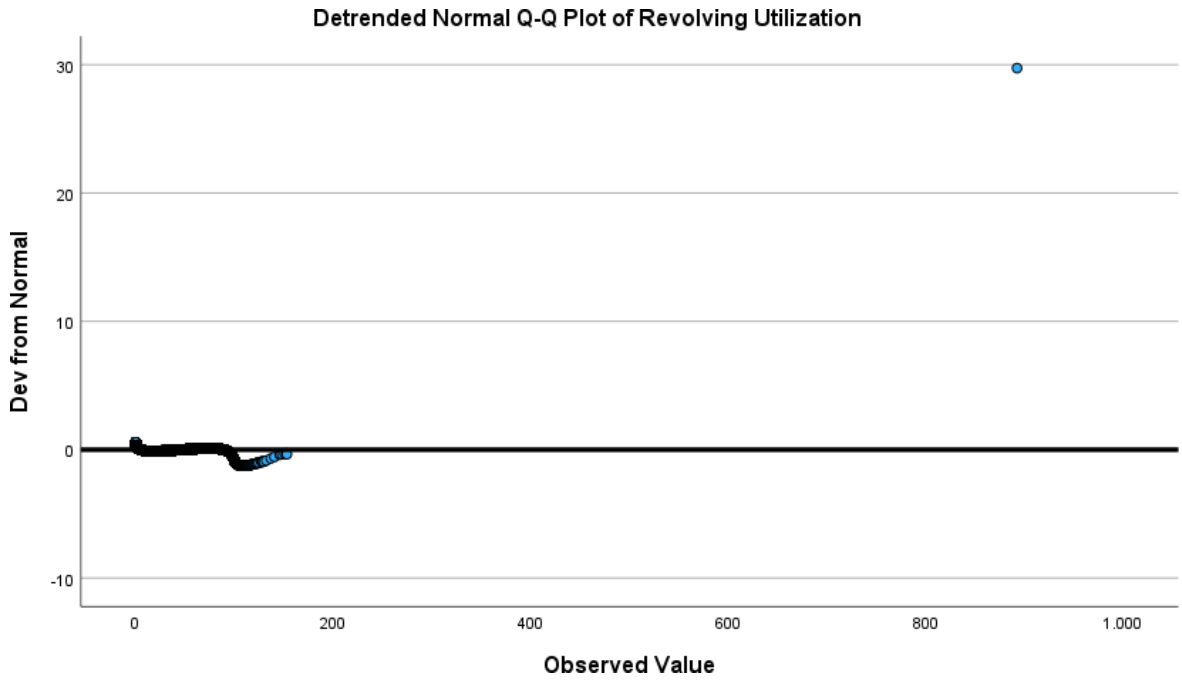
Skewness: (11,728)/(0,004)= 2932(>1,96)

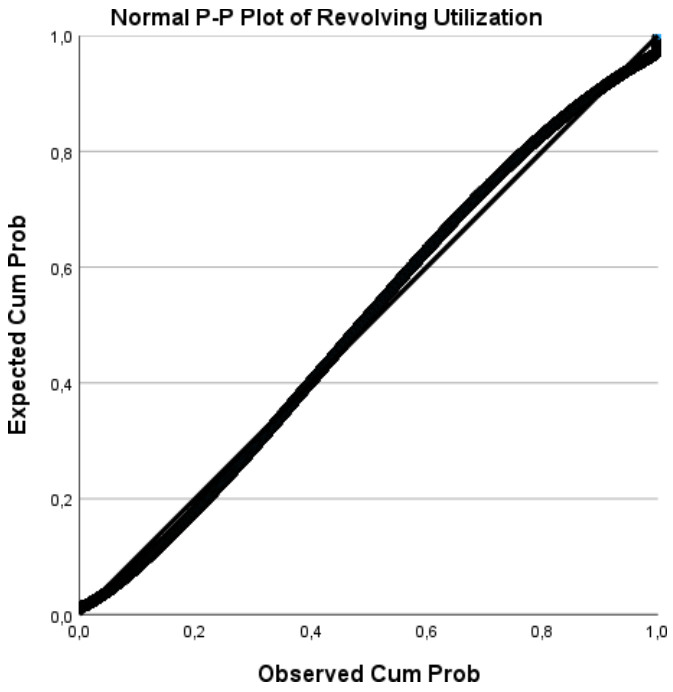
Kurtosis: (384,228)/(0,008)= 48028.5(>1,96)

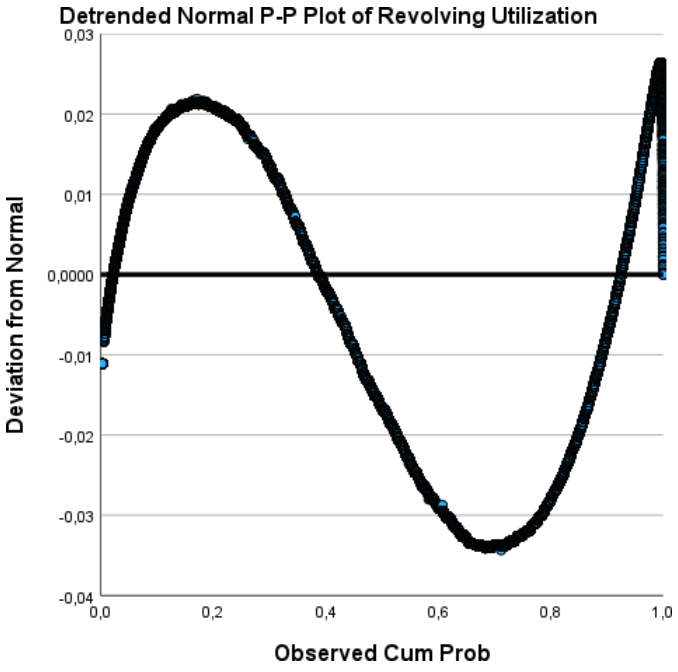
Visual and statistical tests, Kolmogorov-Smirnov, Histogram, and P-P plots and Q-Q plots alongside the z-values of Skewness and Kurtosis, showed that the observations for the variable Revolving Balance weren't normally distributed.

Revolving Utilization









Z-Score:

Skewness: (-0,072)/(0,004)= -18(<-1,96)

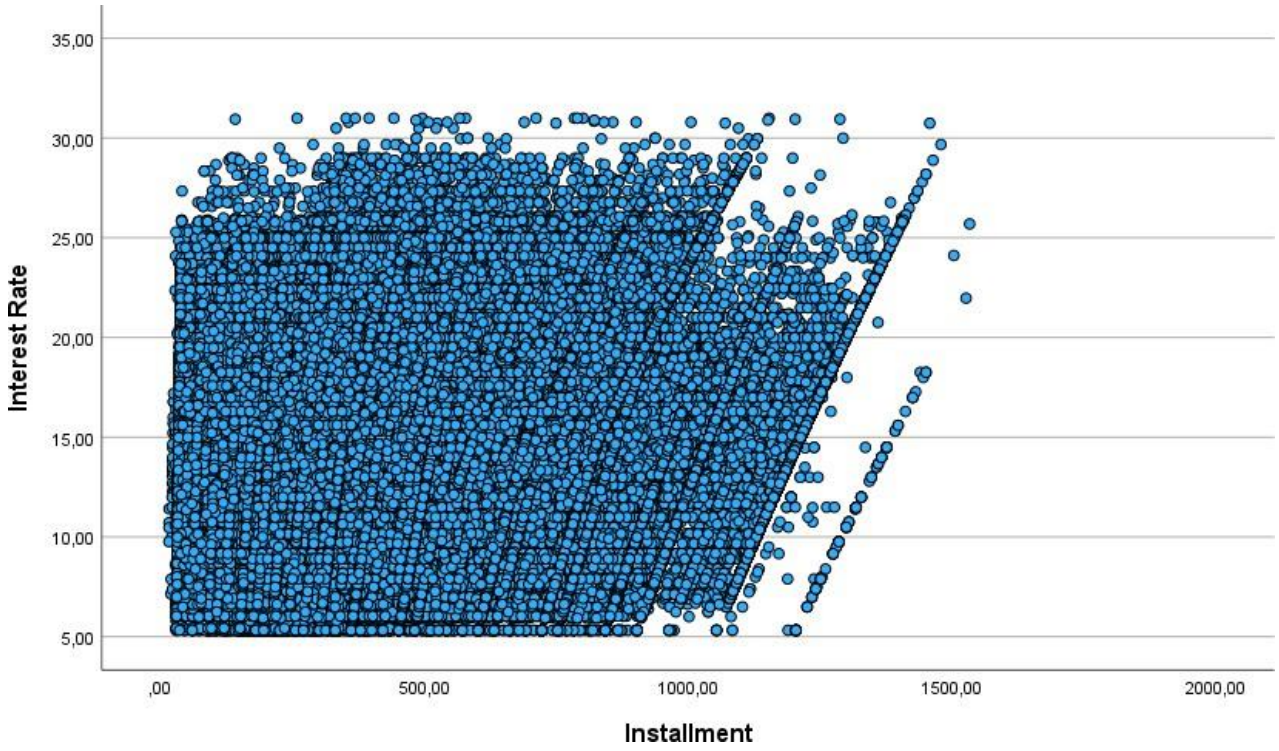
Kurtosis: (2,712)/(0,008)= 339(>1,96)

Even though the Q-Q plot, P-P plot, Box plot and Histogram alongside with the values of Skewness and Kurtosis were giving us the hint that our observations of the variable Revolving Utilization might be following the normal distribution, the Detrended Q-Q and P-P plots alongside the z-values of Skewness and Kurtosis showed us that they do not follow the normal distribution.

In addition, since our sample size is bigger than 300, the normality of the data depends on the histograms and the absolute values of skewness and kurtosis. None of the following requirements of either an absolute skewness value ≤2 or an absolute kurtosis (excess) ≤4 are true, so using the values asreference we can determine that there is no considerable normality in our data for these 4 variables.

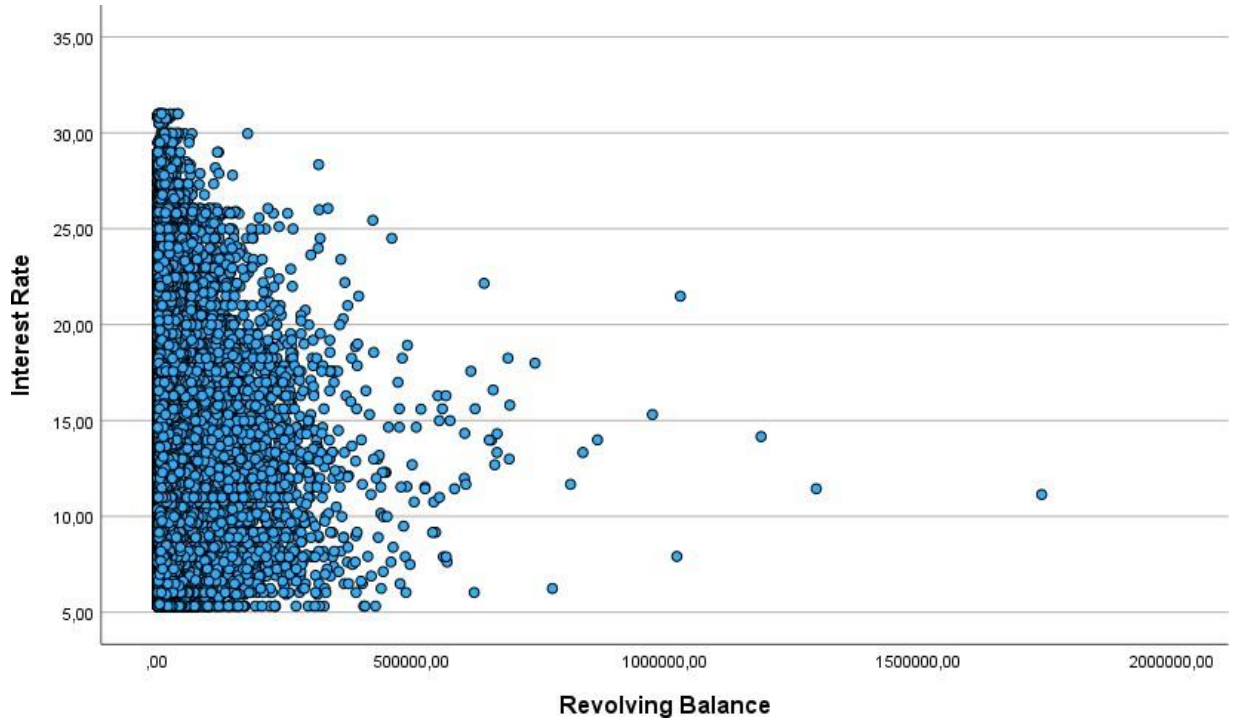
Correlation

Interest Rate – Instalment



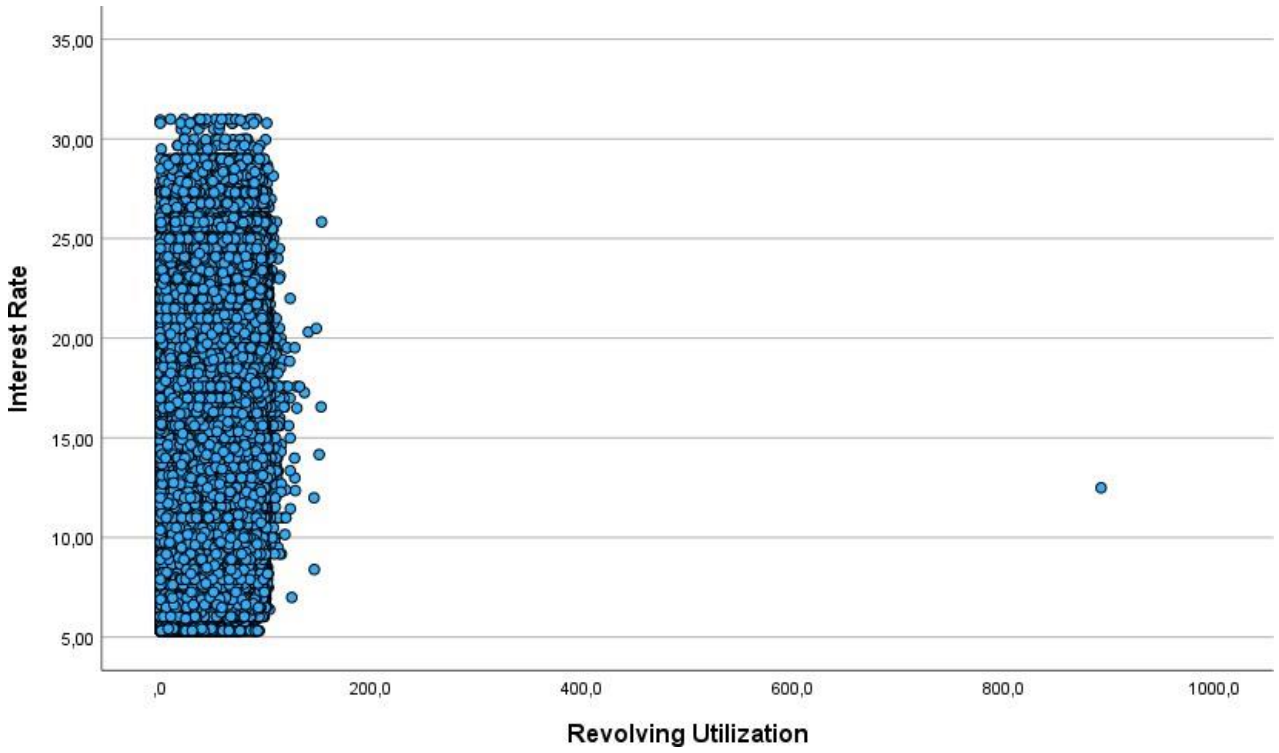
The scatterplot above although somewhat clustered, can show us that in some individual cases there seems to be some correlation between Interest Rate and the Instalment while in some other cases it seems that these two variables are not correlated at all. In the first case, we see that when the Interest Rate increases from 5% onwards the Instalments from 1000 currency units seem to increase as well(the same can be deduced in the instalments of 750 currency units), while in the case of very small Instalments(around 100 currency units) the increasing of the Interest Rate does not affect the amount of Instalment at all.

Interest Rate – Revolving Balance



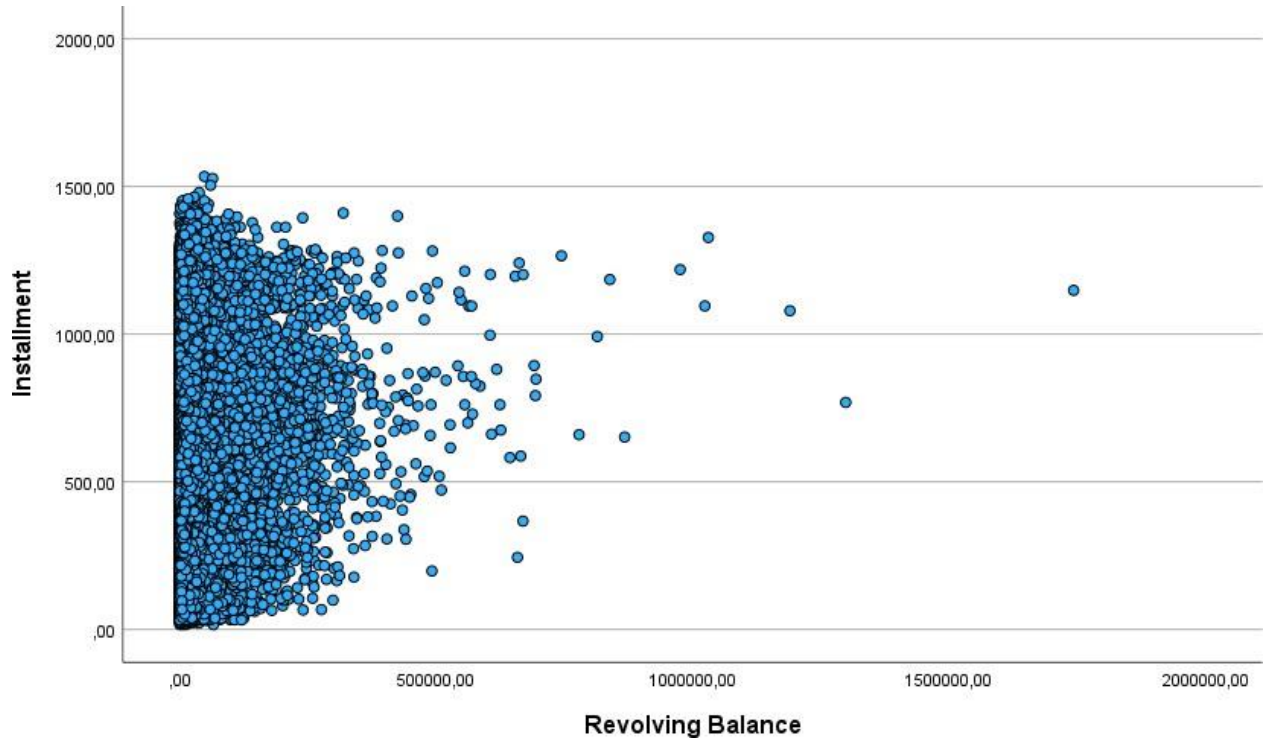
The scatterplot above although somewhat clustered, can show us that there seems not to be any kind of correlation between Interest Rate and Revolving Balance. We can see that in the case of very small Revolving Balance(around 0) the increasing of the Interest Rate does not affect the amount of Revolving Balance at all.

Interest Rate – Revolving Utilization



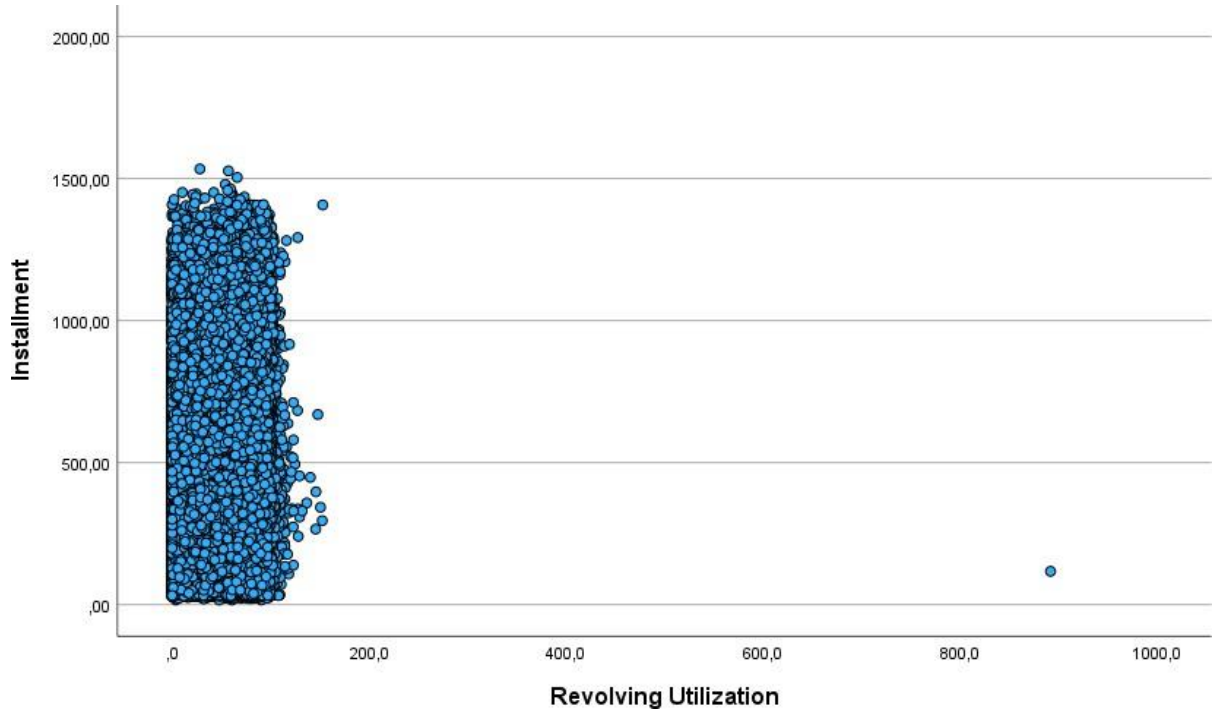
The scatterplot above although somewhat clustered, can show us that there seems not to be any kind of correlation between Interest Rate and Revolving Utilization. We can see that in all the cases present vertical lines can be spotted, this means that the Revolving Utilization remains the same irregardless of the increasing or decreasing of the Interest Rate.

Instalment – Revolving Balance



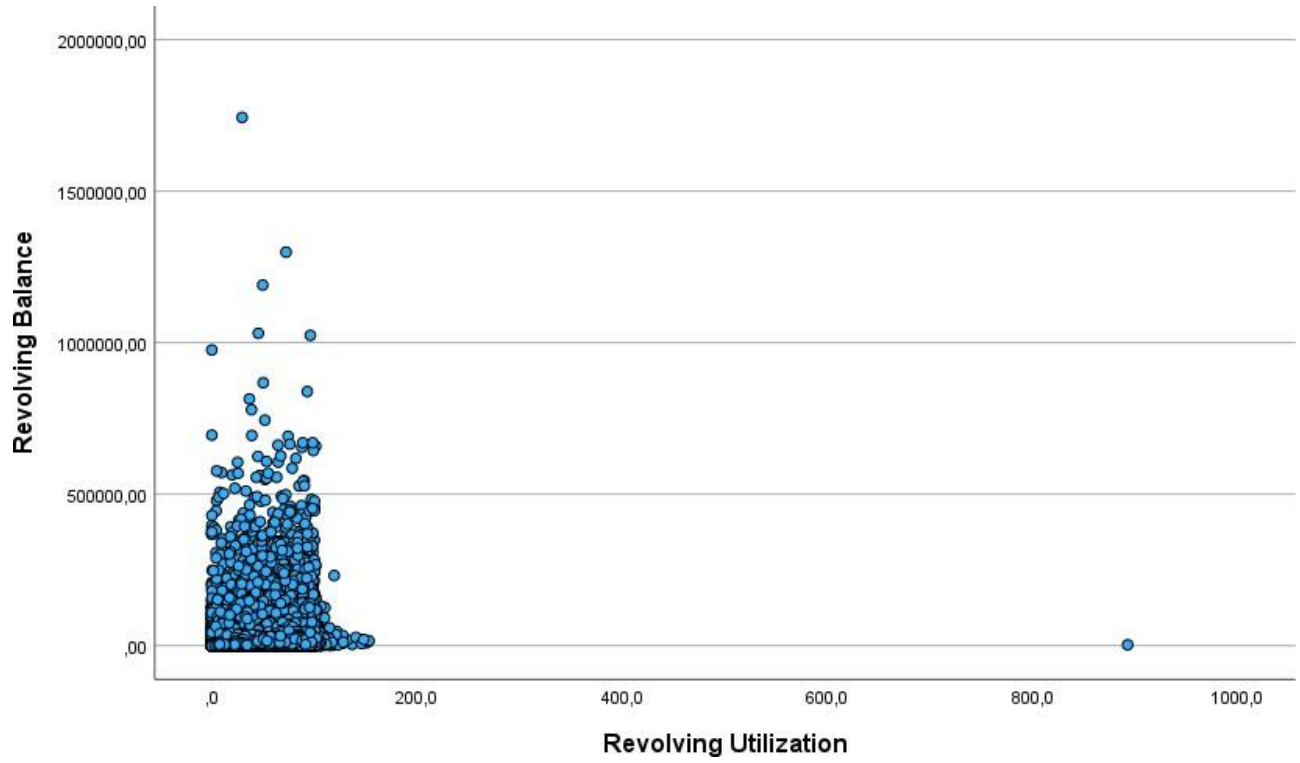
The scatterplot above although somewhat clustered, can show us that there seems not to be any kind of correlation between Instalment and Revolving Balance. We can see that in all the cases present vertical lines can be spotted, this means that the Revolving Balance remains the same irregardless of the increasing or decreasing of the amount of the Instalments.

Instalment – Revolving Utilization



The scatterplot above although somewhat clustered, can show us that there seems not to be any kind of correlation between Interest Rate and Revolving Utilization. We can see that in all the cases present vertical lines can be spotted, this means that the Revolving Utilization remains the same irregardless of the increasing or decreasing of the amount of Instalments.

Revolving Balance – Revolving Utilization



The scatterplot above although somewhat clustered, can show us that there seems not to be any kind of correlation between Revolving Balance and Revolving Utilization. We can see that in all the cases present both vertical and horizontal lines can be spotted, this means that the Revolving Utilization remains the same irregardless of the increasing or decreasing of the Revolving Balance and vice versa.

Probabilities

-What is the probability of a client whose interest rate is more than 19,84%?

P(x=19,84) = P(z=1,39) = 91,77%
P(x>19,84) = P(z>1,39) = 100%-91,77%=8,23%

-What is the probability of a client whose instalment is more than 706,01 currency units?

P(x=706,01) = P(z=1,09) = 86,21%
P(x>706,01) = P(z>1,09) = 100%-86,21%=13,79%

Next batch of variables analyzed were grade rank, verification status, debt to income ratio, bankrupt records, mortgage accounts.

(Santiago Hourcade)

GENERAL STATISTICS

From the “Descriptive Statistics” function of SPSS we analyze these variables in order to check MissingValues and Outliers existing.

Statistics

		Grade Rank	Verified or not	Ratio Debt to Income	Bankrupt Records	Mortage Account
N	Valid	396030	396030	396019	395494	358233
	Missing	0	0	11	536	37797
Mean				17,3795	,12	1,81
Median				16,9100	,00	1,00
Mode				,00	0	0
Std. Deviation				18,01930	,356	2,148
Variance				324,695	,127	4,614
Skewness				431,048	3,423	1,600
Std. Error of Skewness				,004	,004	,004
Range				9999,00	8	34
Minimum				,00	0	0
Maximum				9999,00	8	34

Missing Values

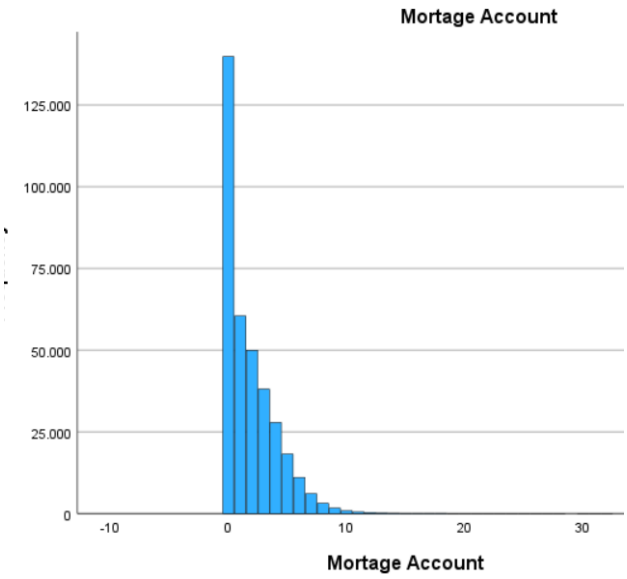
The results of the report expose that the amount of Missing Values is irrelevant except of the variable Mortage Account. Nevertheless it represent less than 10%, to be exactly 37.797 missing values that means 9.5% of total. So to handle with it, we should define if the missing values show a pattern. First, weneed to run descriptive statistics to check if the missing is related to a specific variable, case, or setting.

Mortage Account

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	139776	35,3	39,0	39,0
	1	60415	15,3	16,9	55,9
	2	49948	12,6	13,9	69,8
	3	38049	9,6	10,6	80,4
	4	27887	7,0	7,8	88,2
	5	18194	4,6	5,1	93,3
	31	2	,0	,0	100,0
	32	2	,0	,0	100,0
	34	1	,0	,0	100,0
	Total	358233	90,5	100,0	
Missing	System	37797	9,5		
Total		396030	100,0		

Mortage Account

N	Valid	358233
	Missing	37797
Mean		1,81
Median		1,00
Mode		0
Std. Deviation		2,148
Variance		4,614
Skewness		1,600
Std. Error of Skewness		,004
Range		34
Minimum		0
Maximum		34



As we see in the graph the variable doesn't shows a normal distribution. It seems to be a left distribution therefore we can define the variable as a MAR (Missing at Random). So in order to replace its we must apply Median.

Outliers

From analyzing Descriptive Analyze Report we can recognize the maximum value of ratio Deb to Incomeis 9999.00. As we are talking of a Ratio we realize there is an outlier so let's find out:

annual_inco	verification_status	issue_date	loan_status	purpose	title	dti	e
2500	Not Verified	1/11/2016	Fully Paid	debt_consolidation	Debt consolidation	189,90%	
8000	Source Verified	1/3/2016	Fully Paid	debt_consolidation	Debt consolidation	145,65%	
6672	Verified	1/2/2016	Fully Paid	other	Other	107,55%	
0	Not Verified	1/11/2015	Charged Off	credit_card	Credit card refinancing	9999,00%	
8700	Source Verified	1/12/2015	Fully Paid	debt_consolidation	Debt consolidation	120,66%	
16000	Source Verified	1/9/2016	Fully Paid	debt_consolidation	Debt consolidation	138,03%	
5000	Source Verified	1/11/2015	Fully Paid	other	Other	380,53%	
600	Source Verified	1/6/2016	Fully Paid	debt_consolidation	Debt consolidation	1622,00%	

Ratio Debt to Income

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	,00	313	,1	,1	,1
	,01	8	,0	,0	,1
	,02	12	,0	,0	,1
	,03	5	,0	,0	,1
	71,40	1	,0	,0	100,0
	77,95	1	,0	,0	100,0
	88,21	1	,0	,0	100,0
	92,13	1	,0	,0	100,0
	93,86	1	,0	,0	100,0
	107,55	1	,0	,0	100,0
	120,66	1	,0	,0	100,0
	138,03	1	,0	,0	100,0
	145,65	1	,0	,0	100,0
	189,90	1	,0	,0	100,0
	380,53	1	,0	,0	100,0
	1622,00	1	,0	,0	100,0
	9999,00	1	,0	,0	100,0
	Total	396019	100,0	100,0	
Missing	System	11	,0		
Total		396030	100,0		

Debt to Income is a ratio, so values must be between 0% and 100%. Exceptionally could be more than 100% in the case that Debt is more than the Incomes but they don't represent measurement errors, data entry errors or poor sampling errors.

For the last two values (1622% and 9999%) the issue is evident that the Income is too low or zero. So there, we need to focus in the reason why a person without income has receive a loan.

Early Assumptions

Bankrupt

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	350379	88,5	88,6	88,6
	1	42790	10,8	10,8	99,4
	2	1847	,5	,5	99,9
	3	351	,1	,1	100,0
	4	82	,0	,0	100,0
	5	32	,0	,0	100,0
	6	7	,0	,0	100,0
	7	4	,0	,0	100,0
	8	2	,0	,0	100,0
	Total	395494	99,9	100,0	
Missing	System	536	,1		
Total		396030	100,0		

The bankrupts represent only 11.5% (100%-88.5%) of total Loans, so we can assume that the risk taken by the bank is ok. Not several unpaid credits.

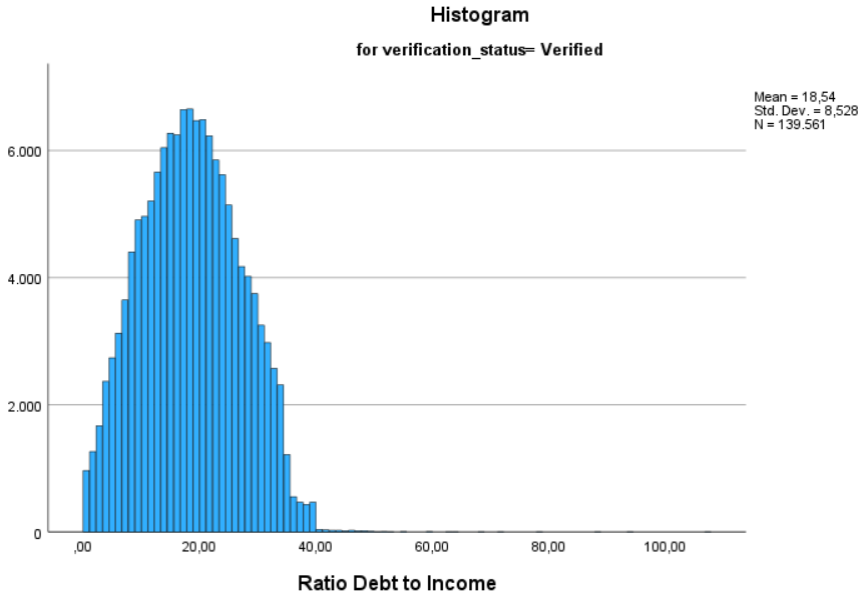
We can also confirm this assumption from the distribution of the rank categories:

- A, B and C categories represents 72.3% of total Debt Grade.
- B and C are the most frequently. This two represents 56% of of total Debt Grade.

Grade Rank

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	A	64187	16,2	16,2	16,2
	B	116018	29,3	29,3	45,5
	C	105987	26,8	26,8	72,3
	D	63524	16,0	16,0	88,3
	E	31488	8,0	8,0	96,3
	F	11772	3,0	3,0	99,2
	G	3054	,8	,8	100,0
	Total	396030	100,0	100,0	

From the crosstab function we can analyze the Relation between Verified Accounts and DTI: It shows a nearly Normal distribution around 18.54% (Mean). That could be a little bit High rate that we should analyze further.



Next batch of variables analysed were subrank, employment length, total number of credit lines, initial listing status, and number of public records. First, we ran frequency analysis on the data. The following are the results:

(Shamma Al Remeithi)

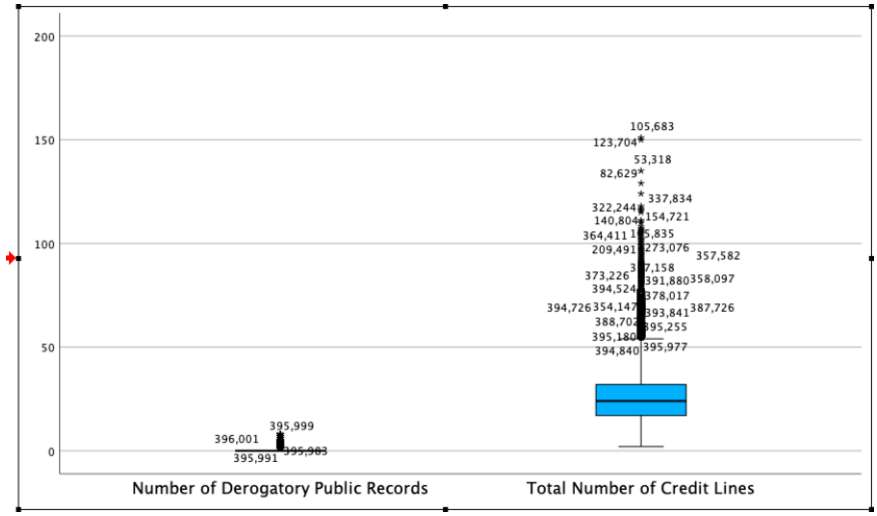
Statistics

		Sub Rank	Employment Length	Total Number of Credit Lines	Initial Listing Status of the Loan	Number of Derogatory Public Records
N	Valid	396029	377729	396024	396018	396017
	Missing	1	18301	6	12	13
Mean				25.41		.18
Median				24.00		.00
Mode				21		0
Std. Deviation				11.887		.494
Variance				141.301		.244
Skewness				.864		4.351
Std. Error of Skewness				.004		.004
Kurtosis				1.205		31.450
Std. Error of Kurtosis				.008		.008
Range				149		8
Minimum				2		0
Maximum				151		8
Percentiles	25			17.00		.00
	50			24.00		.00
	75			32.00		.00

Missing Values

As observed from the analysis, the only variable exceeding the 4% of missing values is employment length (4.6%), the rest have less than 1% missing value. The missing values of employment length are understandable as they correlate with missing values of employee title (MAR), which can mean the client is unemployed or retired. Missing values for initial listing status correlate to loan status being FULLY PAID(MAR) since Listing status is categorised as Whole or Fractional, meaning if the bank will buy the full loan, or a part of it, considering the loan is fully paid for by the client, it is understandable why there are some missing values. Missing values for sub rank are too few to find any pattern (MCAR), no pattern is found for missing values of public records either. For these reasons, there is no need to handle the missing values.

Outliers



Again, even though we see a lot of instances of values that lie above the rest of the observations, these values are natural variations in the population.

Normality Check

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Number of Derogatory Public Records	396017	100.0%	13	0.0%	396030	100.0%
Total Number of Credit Lines	396024	100.0%	6	0.0%	396030	100.0%

Extreme Values

			Case Number	Value
Number of Derogatory Public Records	Highest	1	3562	8
		2	8092	8
		3	14811	8
		4	19927	8
		5	21044	8[a]
	Lowest	1	396030	0
		2	396029	0
		3	396028	0
		4	396027	0
		5	396026	0[b]
Total Number of Credit Lines	Highest	1	105683	151
		2	123704	150
		3	53318	135
		4	82629	129
		5	96572	124
	Lowest	1	366731	2
		2	344369	2
		3	324748	2
		4	310336	2
		5	242958	2[c]

a. Only a partial list of cases with the value 8 are shown in the table of upper extremes.

b. Only a partial list of cases with the value 0 are shown in the table of lower extremes.

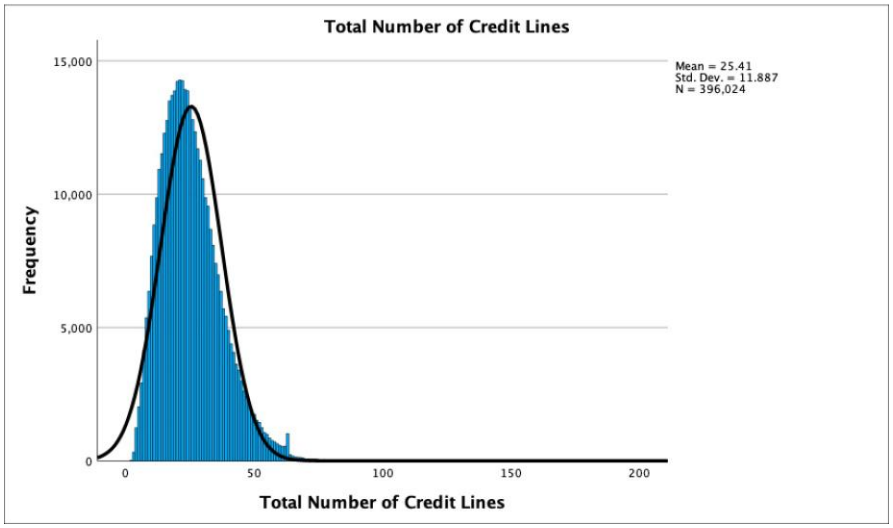
c. Only a partial list of cases with the value 2 are shown in the table of lower extremes.

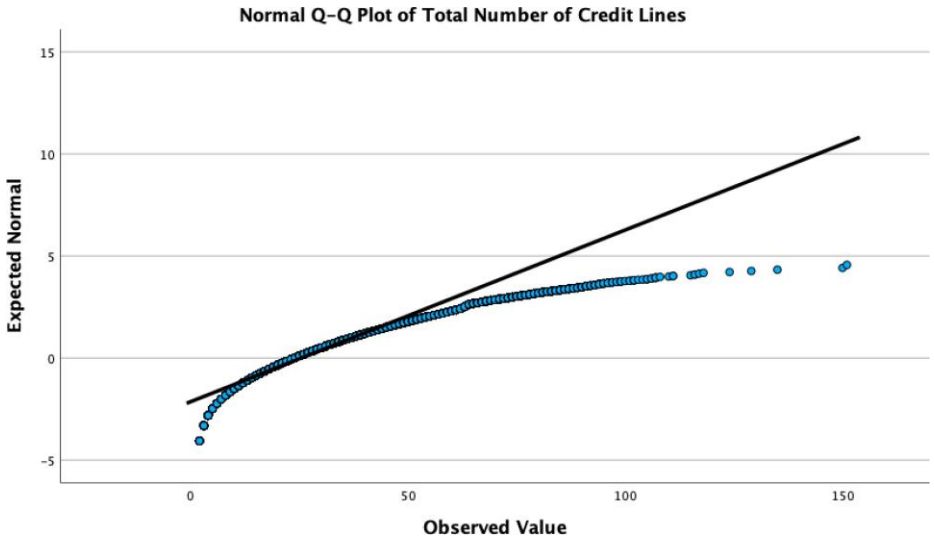
Tests of Normality

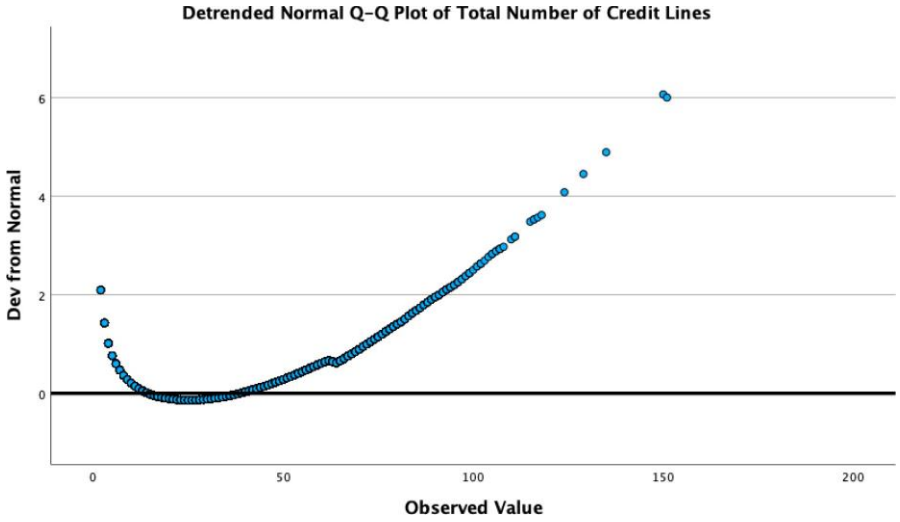
	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
Number of Derogatory Public Records	.494	396017	<.001
Total Number of Credit Lines	.072	396024	<.001

a. Lilliefors Significance Correction

Histogram


Total Number of Credit Lines
Mean = 25.41
Std. Dev. = 11.887
N = 396,024
Frequency
15,000
10,000
5,000
0
0
50
100
150
200
Total Number of Credit Lines


Normal Q-Q Plot of Total Number of Credit Lines
Expected Normal
15
10
5
0
-5
0
50
100
150
Observed Value



Even though the Q-Q plot, Box plot and Histogram alongside with the values of Skewness and Kurtosis and the z-plot were giving us the hint that our observations of the variable Total Number of Credit Line might be following the normal distribution, the Detrended Q-Q plots show us that they do not follow the normal distribution.

Cross Tabulations

Crosstabs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Number of Derogatory Public Records * Sub Rank	396016	100.0%	14	0.0%	396030	100.0%
Number of Derogatory Public Records * Initial Listing Status of the Loan	396005	100.0%	25	0.0%	396030	100.0%
Number of Derogatory Public Records * Employment Length	377717	95.4%	18313	4.6%	396030	100.0%
Total Number of Credit Lines * Sub Rank	396023	100.0%	7	0.0%	396030	100.0%
Total Number of Credit Lines * Initial Listing Status of the Loan	396012	100.0%	18	0.0%	396030	100.0%
Total Number of Credit Lines * Employment Length	377723	95.4%	18307	4.6%	396030	100.0%

Sub Rank * Employment Length Crosstabulation

Count

		Employment Length											
		< 1 year	1 year	10+ year	2 years	3 years	4 years	5 years	6 years	7 years	8 years	9 years	Total
Sub Rank	A1	770	600	3197	898	738	566	636	536	487	467	361	9256
	A2	822	588	3012	879	813	569	681	499	465	466	369	9163
	A3	894	710	3402	941	823	701	719	548	533	489	359	10119
	A4	1333	1043	4886	1479	1302	999	1136	874	797	717	563	15129
	A5	1541	1225	5939	1675	1532	1154	1228	955	963	871	731	17814
	B1	1433	1210	6232	1761	1541	1086	1284	1062	989	952	730	18280
	B2	1755	1451	7291	2011	1766	1373	1548	1194	1179	1068	857	21493
	B3	2138	1680	8408	2462	2139	1578	1828	1424	1468	1306	1042	25473
	B4	2023	1691	8054	2380	2029	1595	1771	1392	1336	1173	1037	24481
	B5	1755	1456	6881	2050	1748	1334	1525	1156	1171	1124	816	21016
	C1	1892	1635	7363	2131	1911	1483	1603	1255	1269	1113	906	22561
	C2	1897	1491	7126	2023	1787	1382	1473	1206	1201	1087	864	21537
	C3	1699	1405	6764	1915	1657	1290	1398	1084	1078	1054	837	20181
	C4	1663	1352	6564	1730	1633	1168	1241	1048	1085	1024	812	19320
	C5	1442	1217	5895	1591	1412	1087	1182	934	948	905	722	17335
	D1	1289	1056	5037	1521	1331	947	1065	853	805	721	566	15191
	D2	1091	926	4301	1271	1137	879	928	759	738	658	549	13237
	D3	973	849	3726	1100	1004	740	806	620	662	632	461	11573
	D4	952	785	3728	1028	954	633	795	602	631	528	467	11103
	D5	754	597	3163	877	730	626	615	482	517	460	368	9189
	E1	627	539	2534	721	621	497	494	388	396	393	325	7535
	E2	630	468	2349	685	589	389	490	378	421	383	325	7107
	E3	475	398	2012	536	502	366	423	328	310	320	248	5918
	E4	406	319	1750	467	423	339	380	269	299	256	208	5116
	E5	338	286	1495	397	358	276	288	228	239	254	212	4371
	F1	257	219	1153	300	270	220	256	190	210	155	144	3374
	F2	212	180	895	244	235	170	190	138	123	154	116	2657
	F3	177	136	745	229	179	133	154	119	127	124	78	2201
	F4	139	102	607	157	142	123	89	98	113	90	67	1727
	F5	109	82	491	118	107	75	99	68	65	75	53	1342
	G1	82	60	347	83	89	68	68	55	67	55	39	1013
	G2	61	52	259	62	54	46	39	34	48	36	32	723
	G3	46	36	184	51	42	20	32	32	42	22	22	529
	G4	29	21	139	32	32	20	17	20	21	17	14	362
	G5	21	17	112	22	35	19	14	13	16	19	14	302
Total		31725	25882	126041	35827	31665	23951	26495	20841	20819	19168	15314	377728

Number of Derogatory Public Records * Sub Rank Crosstabulation

Count

		Sub Rank													
		A1	A2	A3	A4	A5	B1	B2	B3	B4	B5	C1	C2	C3	C4
Number of Derogatory Public Records	0	9484	9199	9958	14622	16742	16625	19453	22991	21837	18273	19779	18790	17448	16779
	1	219	326	550	1032	1594	2295	2729	3210	3307	3281	3335	3275	3249	3001
	2	16	29	42	100	131	181	211	315	318	374	380	335	362	350
	3	6	9	15	22	40	45	65	87	85	96	102	106	103	80
	4	3	0	3	8	10	18	16	29	25	29	43	38	29	40
	5	1	1	3	3	4	11	8	12	15	19	10	18	16	15
	6	0	0	3	1	3	4	7	8	10	8	9	12	6	9
	7	0	1	0	1	2	3	2	3	4	2	2	5	3	4
	8	0	2	1	0	0	0	3	0	0	3	1	1	3	2
Total		9729	9567	10575	15789	18526	19182	22494	26655	25601	22085	23661	22580	21219	20280

C5	D1	D2	D3	D4	D5	E1	E2	E3	E4	E5	F1	F2	F3	F4	F5	G1
14947	13126	11625	10114	9713	8100	6566	6189	5202	4474	3861	2958	2291	1921	1503	1172	874
2807	2448	1984	1758	1638	1336	1116	1041	847	738	615	497	399	312	230	183	144
329	268	226	251	199	192	167	143	110	103	69	47	49	36	35	35	29
98	89	67	56	73	43	45	32	26	32	16	24	16	11	9	4	5
33	32	30	19	15	16	13	15	15	8	9	5	6	3	6	2	5
13	16	6	16	10	11	4	6	3	4	0	2	4	1	2	1	1
9	5	4	4	4	2	4	2	2	0	1	1	0	2	1	0	0
3	5	5	2	3	0	1	1	1	1	1	1	0	0	0	0	0
2	4	4	2	1	0	1	1	0	1	0	1	0	0	1	0	0
18241	15993	13951	12222	11656	9700	7917	7430	6206	5361	4572	3536	2765	2286	1787	1397	1058

G2	G3	G4	G5	Total
624	445	325	260	338270
111	79	39	45	49770
13	16	7	9	5477
4	6	3	1	1521
0	4	0	1	528
0	1	0	0	237
1	0	0	0	122
0	0	0	0	56
0	1	0	0	35
753	552	374	316	396016

According to the crosstabs, it is more likely for the client to be assigned a higher Sub grade the longer they are employed. Along this observation, we can see that the lower number of public records the client has, the likelihood increases in having a higher sub grade, as well as fewer number of credit lines.

Using the Frequencies statistics function on spss for the numeric variables that we picked, mainly the variable “home ownership, annual income, purpose, loan status, and open account, to get the following result.

Statistics

		Purpose	Loan_Status	Home_Ownership	Annual_Income	Open_Account
N	Valid	396030	396030	396030	396029	396023
	Missing	0	0	0	1	7
Mean					74203.2622	11.31
Median					64000.0000	10.00
Mode					60000.00	9
Std. Deviation					61637.67502	5.138
Variance					3799202981.3	26.396
Skewness					41.043	1.213
Std. Error of Skewness					.004	.004
Kurtosis					4238.546	2.967
Std. Error of Kurtosis					.008	.008

After running our data, we saw that purpose, loan status, and home ownership were not missing any missing values. As we looked at annual income where it had one missing value, and the open account was 7 missing values. Understanding that we have the majority of our data and only missing shows a few shows a piece of very accurate information. When we look at annual incomes, the highest income is around $74,203, $64,000 is the median, and the mode is

$60,000. The standard deviation is $61,638, which is the average annual income. As we look at the open account, we can see that for the highest income, they have 11.3 open accounts. The median is at 10, and the mode is at 9. Std. the deviation is at 5.1

Purpose

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid		7	.0	.0	.0
	car	4697	1.2	1.2	1.2
	credit_card	83016	21.0	21.0	22.1
	debt_consolidation	234503	59.2	59.2	81.4
	educational	257	.1	.1	81.4
	home_improvement	24030	6.1	6.1	87.5
	house	2201	.6	.6	88.1
	major_purchase	8790	2.2	2.2	90.3
	medical	4196	1.1	1.1	91.3
	moving	2854	.7	.7	92.1
	other	21185	5.3	5.3	97.4
	renewable_energy	329	.1	.1	97.5
	small_business	5701	1.4	1.4	98.9
	vacation	2452	.6	.6	99.5
	wedding	1812	.5	.5	100.0
	Total	396030	100.0	100.0	

\

As we look into the purpose of the loan to see the different reasons for our loan, we see that most of our loans are distributed from the car, credit card, debt consolidation, educational, homeimprovement, house improvements house, major purchase, medical, moving, other, renewable energy, small business, vacation, and wedding. We see that our most extensive loan distributionis in debt consolidation at $234,503, second being credit card at $83,016 and third being home improvement loan at $24,030

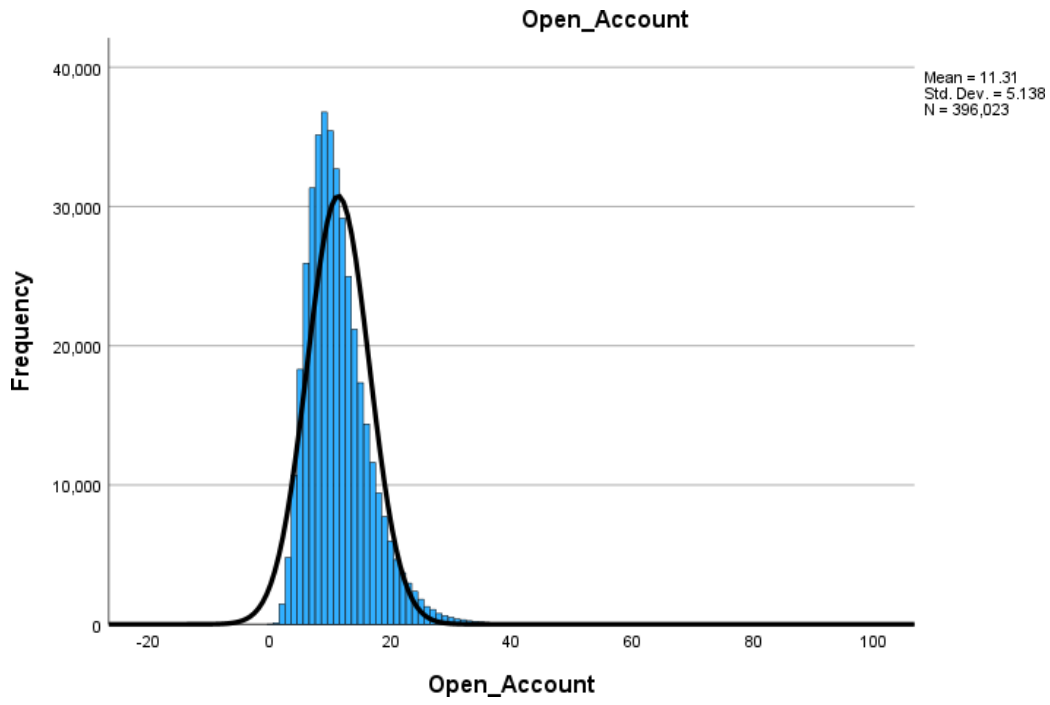
Loan_Status

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid		6	.0	.0	.0
	Charged Off	77673	19.6	19.6	19.6
	Fully Paid	318351	80.4	80.4	100.0
	Total	396030	100.0	100.0	

Home_Ownership

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	ANY	3	.0	.0	.0
	MORTGAGE	198348	50.1	50.1	50.1
	NONE	31	.0	.0	50.1
	OTHER	112	.0	.0	50.1
	OWN	37746	9.5	9.5	59.7
	RENT	159790	40.3	40.3	100.0
	Total	396030	100.0	100.0	

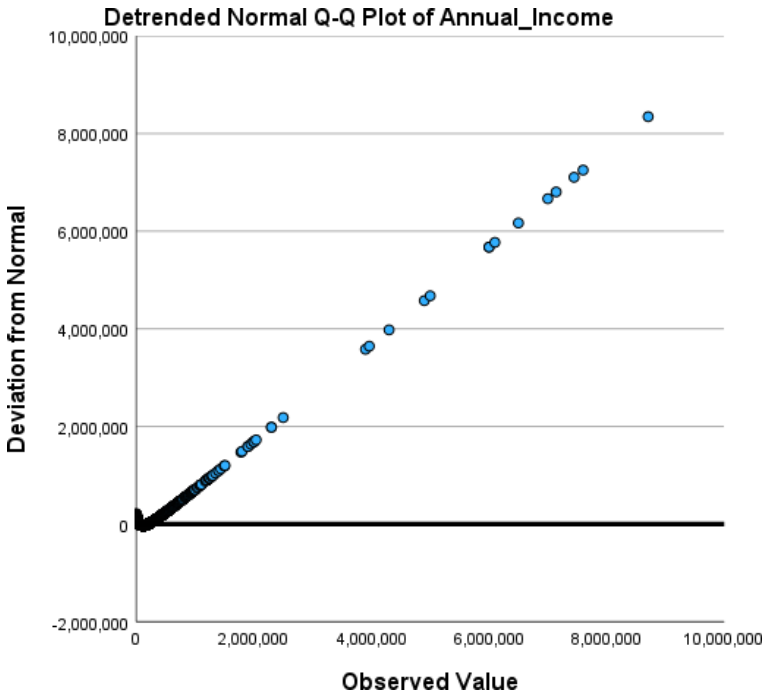
Histogram

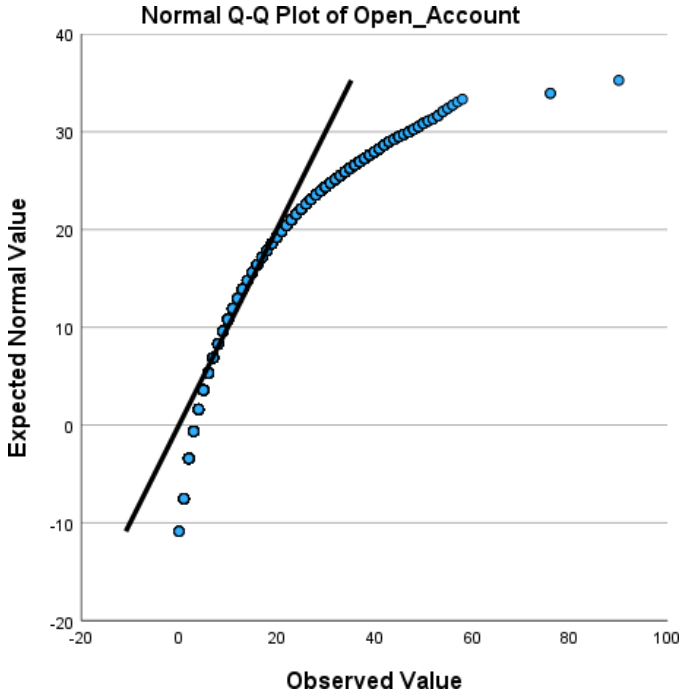


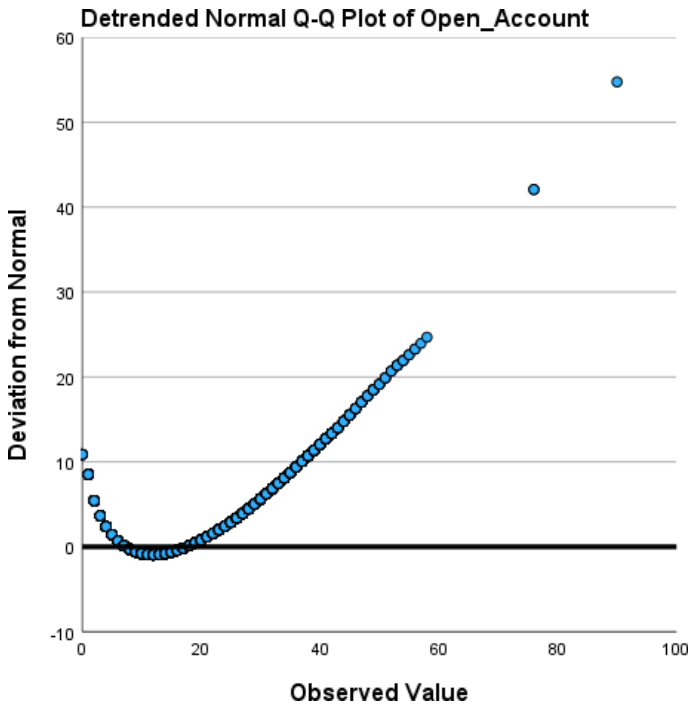
Purpose * Loan_Status Crosstabulation

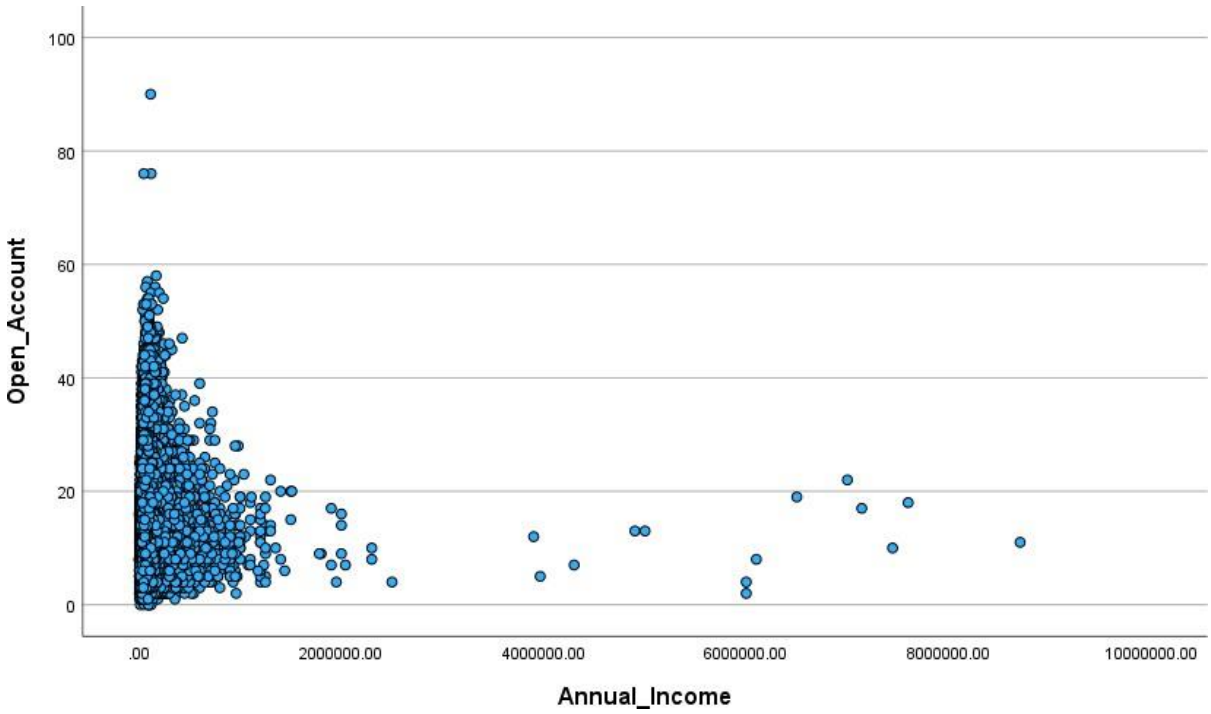
			Loan_Status		
			Charged Off	Fully Paid	Total
Purpose		Count	0[a]	1[a]	6[a]	7
		% within Purpose	0.0%	14.3%	85.7%	100.0%
		% within Loan_Status	0.0%	0.0%	0.0%	0.0%
		% of Total	0.0%	0.0%	0.0%	0.0%
	car	Count	1[a]	633[b]	4063[c]	4697
		% within Purpose	0.0%	13.5%	86.5%	100.0%
		% within Loan_Status	16.7%	0.8%	1.3%	1.2%
		% of Total	0.0%	0.2%	1.0%	1.2%
	credit_card	Count	2[a, b]	13874[b]	69140[a]	83016
		% within Purpose	0.0%	16.7%	83.3%	100.0%
		% within Loan_Status	33.3%	17.9%	21.7%	21.0%
		% of Total	0.0%	3.5%	17.5%	21.0%
	debt_consolidation	Count	3[a, b]	48639[b]	185861[a]	234503
		% within Purpose	0.0%	20.7%	79.3%	100.0%
		% within Loan_Status	50.0%	62.6%	58.4%	59.2%
		% of Total	0.0%	12.3%	46.9%	59.2%
	educational	Count	0[a]	42[a]	215[a]	257
		% within Purpose	0.0%	16.3%	83.7%	100.0%
		% within Loan_Status	0.0%	0.1%	0.1%	0.1%
		% of Total	0.0%	0.0%	0.1%	0.1%
	home_improvement	Count	0[a, b]	4087[b]	19943[a]	24030
		% within Purpose	0.0%	17.0%	83.0%	100.0%
		% within Loan_Status	0.0%	5.3%	6.3%	6.1%
		% of Total	0.0%	1.0%	5.0%	6.1%
	house	Count	0[a]	434[a]	1767[a]	2201
		% within Purpose	0.0%	19.7%	80.3%	100.0%
		% within Loan_Status	0.0%	0.6%	0.6%	0.6%
		% of Total	0.0%	0.1%	0.4%	0.6%
	major_purchase	Count	0[a, b]	1448[b]	7342[a]	8790
		% within Purpose	0.0%	16.5%	83.5%	100.0%
		% within Loan_Status	0.0%	1.9%	2.3%	2.2%
		% of Total	0.0%	0.4%	1.9%	2.2%
	medical	Count	0[a, b]	911[b]	3285[a]	4196
		% within Purpose	0.0%	21.7%	78.3%	100.0%
		% within Loan_Status	0.0%	1.2%	1.0%	1.1%
		% of Total	0.0%	0.2%	0.8%	1.1%
	moving	Count	0[a, b]	670[b]	2184[a]	2854
		% within Purpose	0.0%	23.5%	76.5%	100.0%
		% within Loan_Status	0.0%	0.9%	0.7%	0.7%
		% of Total	0.0%	0.2%	0.6%	0.7%
	other	Count	0[a, b]	4495[b]	16690[a]	21185
		% within Purpose	0.0%	21.2%	78.8%	100.0%
		% within Loan_Status	0.0%	5.8%	5.2%	5.3%
		% of Total	0.0%	1.1%	4.2%	5.3%
	renewable_energy	Count	0[a]	77[a]	252[a]	329
		% within Purpose	0.0%	23.4%	76.6%	100.0%
		% within Loan_Status	0.0%	0.1%	0.1%	0.1%
		% of Total	0.0%	0.0%	0.1%	0.1%
	small_business	Count	0[a, b]	1679[b]	4022[a]	5701
		% within Purpose	0.0%	29.5%	70.5%	100.0%
		% within Loan_Status	0.0%	2.2%	1.3%	1.4%
		% of Total	0.0%	0.4%	1.0%	1.4%
	vacation	Count	0[a]	464[a]	1988[a]	2452
		% within Purpose	0.0%	18.9%	81.1%	100.0%
		% within Loan_Status	0.0%	0.6%	0.6%	0.6%
		% of Total	0.0%	0.1%	0.5%	0.6%
	wedding	Count	0[a, b]	219[b]	1593[a]	1812
		% within Purpose	0.0%	12.1%	87.9%	100.0%
		% within Loan_Status	0.0%	0.3%	0.5%	0.5%
		% of Total	0.0%	0.1%	0.4%	0.5%
Total		Count	6	77673	318351	396030
		% within Purpose	0.0%	19.6%	80.4%	100.0%
		% within Loan_Status	100.0%	100.0%	100.0%	100.0%
		% of Total	0.0%	19.6%	80.4%	100.0%

Each subscript letter denotes a subset of Loan_Status categories whose column proportions do not differ significantly from each other at the .05 level.


Detrended Normal Q-Q Plot of Annual_Income
Deviation from Normal
10,000,000
8,000,000
6,000,000
4,000,000
2,000,000
0
-2,000,000
0
2,000,000
4,000,000
6,000,000
8,000,000
10,000,000
Observed Value


Normal Q-Q Plot of Open_Account
Expected Normal Value
40
30
20
10
0
-10
-20
-20
0
20
40
60
80
100
Observed Value


Detrended Normal Q-Q Plot of Open_Account
Deviation from Normal
60
50
40
30
20
10
0
-10
0
20
40
60
80
100
Observed Value


Open_Account
100
80
60
40
20
0
.00
2000000.00
4000000.00
6000000.00
8000000.00
10000000.00
Annual_Income

Group C - Lending Club Loan

Part 2

Introduction

The LendingClub is a fintech marketplace bank, which is designed to help customers pay less when borrowing and earn more when saving. This report is a continuation of our analysis on theloans given by the LendingClub where we are researching the various factors that go into deciding which customers are eligible to receive the loans as well as how the interest rates are determined among various others. The contents of this report will show our t-test and ANOVA analyses on the multiple variables that we have concluded so far have an impact on the decisions that are made in regards to the loans, and will raise hypotheses and check if they are accepted or rejected according to these tests. In the previous report, we have concluded that most of the variables in our dataset are not normally distributed. However, we will be continuing in this report with the assumptions that they are normal.

(collectively)

Hypotheses

We took the decision of further reducing our variables, so we decided to omit some of the variables that we concluded in our last report that do not impact our research.

T-tests

One-Sample t-test

- Interest Rate:

H_0:	There is **no considerable difference** between the mean value of the population and the specified mean value of the samplefor the variable of Interest Rate ($M_N = M_n$)
H_1:	There **is considerable difference** betweenthe mean value of the population and the specified mean value of the sample for the variable of Interest Rate ($M_N \neq M_n$)

After some research, we found that the average lending interest rate (%) in the United States was reported at **3.25%** in 2021, according to the World Bank collection of development indicators, compiled from officially recognized sources (Trading Economics, 2022). So the Test Value in our One Sample t-test that we are going to perform via SPSS, in this case, is going tobe 3.25 (M_N).

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Installment	396026	431,8473	250,72386	,39841

One-Sample Test

Test Value = 1.233

	t	df	Significance One-Sided p	Significance Two-Sided p	Mean Difference	95% Confidence Interval of the Difference Lower	95% Confidence Interval of the Difference Upper
Installment	1080,823	396025	<,001	<,001	430,61426	429,8334	431,3951

We can conclude that:

a) the value of our calculated t-value is bigger that the value of the criticalt-value(1461.964 > 1,645)

b) the value of p-value is less than 0,001 so less than 5%

c) between the range of the lower value(10.37) and the upper value(10.40) 0 is notincluded.

The assumption that we can draw from the above is that we **reject** the $\mathbf{H_0}$ hypothesis and we **accept** the $\mathbf{H_1}$ hypothesis and we conclude that "There is considerable difference between the mean value of the population and the specified mean value of the sample for the variable of Interest Rate ($M_N \neq M_n$)".

- Installment:

H_0:	There is **no considerable difference** between the mean value of the population and the specified mean value of the samplefor the variable of Installment ($M_N = M_n$)
H_1:	There **is considerable difference** betweenthe mean value of the population and the specified mean value of the sample for the variable of Installment ($M_N \neq M_n$)

After some research, we found that "when it comes to monthly expenses, consumers spend agood chunk of their income paying down debt: an average monthly total of 1.233, a LendingTree

study found (LendingTree, 2022). So the Test Value in our One Sample t-test that we are going to perform via SPSS, in this case, is going to be 1.233 (M_N).

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Installment	396026	431,8473	250,72386	,39841

One-Sample Test

Test Value = 1.233

	t	df	Significance One-Sided p	Significance Two-Sided p	Mean Difference	95% Confidence Interval of the Difference Lower	95% Confidence Interval of the Difference Upper
Installment	1080,823	396025	<,001	<,001	430,61426	429,8334	431,3951

We can conclude that:

a) the value of our calculated t-value is bigger that the value of the critical t-value(1080.823 > 1.645)

b) the value of p-value is less than 0.001 so less than 5%

c) between the range of the lower value (429.61) and the upper value (431.39) 0 is not included.

The assumption that we can draw from the above is that we **reject** the $\mathbf{H_0}$ hypothesis and we **accept** the $\mathbf{H_1}$ hypothesis and we conclude that "There is considerable difference between the mean value of the population and the specified mean value of the sample for the variable of Installment ($M_N \neq M_n$)".

- Total Number of Credit Lines:

H_0:	There is **no considerable difference** between the mean value of the population and the specified mean value of the samplefor the variable of Total Number of Credit Lines ($M_N = M_n$)

H_1:	There **is considerable difference** betweenthe mean value of the population and the specified mean value of the sample for the variable of Total Number of Credit Lines ($M_N \neq M_n$)

After some research, we found that the average ownership of credit accounts in the United States was reported at **3** in 2021. So the Test Value in our One Sample t-test that we are going to perform via SPSS, in this case, is going to be 3 (M_N).

These are the results we got:

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Total Number of Credit Lines	396024	25.41	11.887	.019

One-Sample Test

Test Value = 3

	t	df	Significance One-Sided p	Significance Two-Sided p	Mean Difference	95% Confidence Interval of the Difference Lower	95% Confidence Interval of the Difference Upper
Total Number of Credit Lines	1186.645	396023	<.001	<.001	22.415	22.38	22.45

The calculated **t-value** is 1186.645 the **degree of freedom(df)** is 396023 (**N-1**), the **p-value** is less than 0.001 and by extension, less than 5% and the **Mean Difference** between the Mean of population(M_N) and the Mean of our sample(M_n) is equal to 22.415.

We can conclude that:

a) the value of our calculated t-value is bigger that the value of the critical t-value(1186.645 > 1.645)

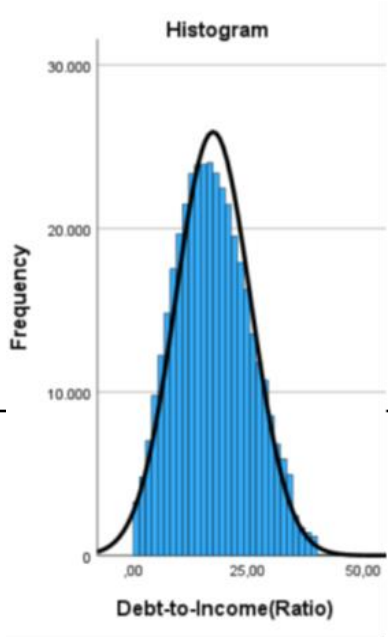
b) the value of p-value is less than 0.001 so less than 5%

c) between the range of the lower value (25.38) and the upper value (25.45) 0 is not included.

The assumption that we can draw from the above is that we **reject** the H_0 hypothesis and we **accept** the H_1 hypothesis and we conclude that "There is considerable difference between the mean value of the population and the specified mean value of the sample for the variable of Total Number of Credit Lines ($M_N \neq M_n$)".

- DTI:

Once we substitute outliers and out of the range values we observe that the variable DTI (Debt to Income) follows normal distribution.



H_0:	There is **no considerable difference** between the mean value of the population and the specified mean value of the sample for the variable of DTI ($M_N = M_n$)
H_1:	There **is considerable difference** between the mean value of the population and the specified mean value of the sample for the variable of DTI ($M_N \neq M_n$)

In order to find out the mean value of the population we found that the average DTI (%) in the United States was around **36%** according to Chase Bank (Source: https://www.chase.com/personal/credit-cards/education/basics/what-is-debt-to-income-ratio-and-why-it-is-important). So the Test Value in our One Sample t-test that we are going to perform via SPSS, in this case, is going to be 36(M_N).

These are the results that we got:

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Debt-to-Income(Ratio)	396019	17,3488	8,13301	,01292

One-Sample Test

Test Value = 36

	t	df	Significance One-Sided p	Significance Two-Sided p	Mean Difference	95% Confidence Interval of the Difference Lower	95% Confidence Interval of the Difference Upper
Debt-to-Income(Ratio)	-1443,160	396018	<,001	<,001	-18,65125	-18,6766	-18,6259

Even though the calculated **t-value** is negative -1443.16, the **p-value** is less than 0,001 and so less than 5%. Also p-value value is not include between de 95% Confidence Interval of the Difference. Therefore we can assume that we **reject** the **H_0** hypothesis and we **accept** the **H_1** hypothesis and we conclude that **There is considerable difference between the mean value**

of the population and the specified mean value of the sample for the variable of DTI ($M_N \neq M_n$)".

Paired-Samples t-test

For the variables that we have available, the theoretical necessary conditions do not apply so that we can proceed to the paired-samples t-test.

Unpaired-Samples t-test

- Interest Rate – Term

H_0:	There is **no considerable difference** between the mean value of the interest rateof the loans that have 36 months as a termtime and those that have 60 months as a term time
H_1:	There **is considerable difference** betweenthe mean value of the interest rate of the loans that have 36 months as a term time and those that have 60 months as a term time

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means							
						Significance				95% Confidence Interval of the Difference	
		F	Sig.	t	df	One-Sided p	Two-Sided p	Mean Difference	Std. Error Difference	Lower	Upper
Interest Rate	Equal variances assumed	1353,250	<,001	-303,668	396024	<,001	<,001	-4,56772	,01504	-4,59720	-4,53823
	Equal variances not assumed			-289,733	146247,464	<,001	<,001	-4,56772	,01577	-4,59862	-4,53682

We decided to use the other available parametric tests in our arsenal in order to draw conclusions about the variables of our sample.

- Total Number of Credit Accounts – Initial Listing Status

H_0:	There is **no considerable difference** between the mean value of the total numberof credit accounts for loan accounts that are listed as whole or fractional
H_1:	There **is considerable difference** between the mean value of the total number of credit accounts for loan accounts that are listed as whole or fractional

Group Statistics

	Initial Listing Status of the Loan	N	Mean	Std. Deviation	Std. Error Mean
Total Number of Credit Lines	Whole Program	157921	26.41	12.067	.030
	Fractional Program	157994	24.76	11.720	.029

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means							
						Significance				95% Confidence Interval of the Difference	
		F	Sig.	t	df	One-Sided p	Two-Sided p	Mean Difference	Std. Error Difference	Lower	Upper
Total Number of Credit Lines	Equal variances assumed	78.205	<.001	38.927	315913	<.001	<.001	1.648	.042	1.565	1.731
	Equal variances not assumed			38.926	315636.954	<.001	<.001	1.648	.042	1.565	1.731

The calculated **t-value** is 38.927 the **degree of freedom(df)** is 315913 (**N-1**), the **p-value** is less than 0.001 and by extension, less than 5% and the **Mean Difference** is 1.648

We can conclude that:

a) the value of our calculated t-value is bigger that the value of the critical t-value(38.927 > 1.645)

b) the value of p-value is less than 0.001 so less than 5%

c) between the range of the lower value (1.565) and the upper value (1.731) 0 is not included.

The assumption that we can draw from the above is that we **reject** the H_0 hypothesis and we **accept** the H_1 hypothesis and we conclude that "There **is considerable difference** between the mean value of the total number of credit accounts for loan accounts that are listed as whole or fractional".

ANOVA

One-factor-ANOVA

1. Does the interest rate implemented in the loan have a considerable effect on the amountof the monthly installment that is being paid by the client?

H_0:	The mean value of all groups of interest rate is the **same**.
H_1:	There are **differences** in the mean values of the groups.

ANOVA

Installment

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1355264195,6	565	2398697,691	40,297	<,001
Within Groups	23539761311	395458	59525,313		
Total	24895025507	396023			

We **reject** hypothesis H_0 and we **accept** hypothesis H_1. So, we draw the assumption that, the means of at least 2 groups of interest rate are not equal and that interest rate, in fact, has a considerable effect on the amount of monthly installment that our clients mustpay.

2. Does the total number of credit lines have a considerable effect on the initial listing status of the customer?

H_0:	The mean value of all groups of interest rate is the **same**.
H_1:	There are **differences** in the mean values of the groups.

ANOVA

Total Number of Credit Lines

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	257583.215	1	257583.215	1831.367	<.001
Within Groups	55699111.1	396010	140.651		
Total	55956694.3	396011			

We **reject** hypothesis H_0 and we **accept** hypothesis H_1. So, we draw the assumption thatthe means of at least 2 groups of credit lines are not equal and that the number of credit lines, in fact, has a considerable effect on the initial listing status. This supports the unpaired t-test result.

3. Does the DTI ratio of the persons have a considerable effect on the loan amount they receive?

H_0:	The mean value of all groups of DTI ratio is the **same**.
H_1:	There are **differences** in the mean values of the groups.

ANOVA

Debt-to-Income(Ratio)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	684644,852	1396	490,433	7,587	<,001
Within Groups	25510096,935	394620	64,645		
Total	26194741,787	396016			

As we can see here, p-value is less than 0,001 and by extend less than 0,05. That means thatwe **reject** hypothesis $\mathbf{H_0}$ and we **accept** hypothesis $\mathbf{H_1}$. So we can conclude that **DTI ratio has a considerable effect on the amount of loan** that clients receive. In fact, it has sense.

Two-factor-ANOVA

1. Do the interest rate implemented in the loan and the term in which the loan is to be paid back, both have considerable effect on the amount of the monthly installment that is being paid by the client?

H_0:	There are **no** significant differences between the factor levels of interest rate There are **no** significant differencesbetween the factor levels of term Interest rate has **no** effect on the effect ofterm
H_1:	There **are** significant differences between the factor levels of interest rate There **are** significant differences between the factor levels of term

	Interest rate **has** effect on the effect ofannual salary

Tests of Between-Subjects Effects

Dependent Variable: Installment

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[b]
Corrected Model	1783466504[a]	931	1915646,083	32,748	<,001	,072	30488,353	1,000
Intercept	1485605122,2	1	1485605122,2	25396,414	<,001	,060	25396,414	1,000
interestrate	728520277,64	565	1289416,421	22,043	<,001	,031	12454,051	1,000
term	128313,317	1	128313,317	2,194	,139	,000	2,194	,316
interestrate * term	255414865,38	365	699766,754	11,963	<,001	,011	4366,316	1,000
Error	23111499936	395091	58496,650					
Total	98750370466	396023						
Corrected Total	24894966439	396022						

a. R Squared = ,072 (Adjusted R Squared = ,069)

b. Computed using alpha = ,05

When taking into consideration the results that we got for **p-value**, we can say that term has no significant effect on the amount of monthly installments that our clients must pay(since 0,139>0,05), so for the variable of term we **accept H_0** and we **reject H_1**. However, in all our other cases we **reject H_0** and we **accept H_1**.

So we conclude that besides the factor levels of term, all the other factors have a significant effect on the amount of installment that our clients must pay each month. Except for the fact that the effect that the interest rate has on the effect of term, even though existent, is not of that high of significance.

2. Do the initial listing status of the loan and the total number of credit lines a customer hasboth have considerable effect on the grade of the loan?

H_0:	There are **no** significant differences between the factor levels of Total Numberof Credit Lines There are **no** significant differences between the factor levels of Initial Listing Status

	Total Number of Credit Lines has **no** effect on the Initial Listing Status
H_1:	There **are** significant differences between the factor levels of Total Number of CreditLines There **are** significant differences between the factor levels of Initial Listing Status Total Number of Credit Lines **has** effect on the Initial Listing Status

Tests of Between-Subjects Effects

Dependent Variable: Grade/Rank

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[b]
Corrected Model	3867.183[a]	219	17.658	9.975	<.001	.005	2184.509	1.000
Intercept	4388.132	1	4388.132	2478.785	<.001	.006	2478.785	1.000
initialliststatus	34.470	1	34.470	19.472	<.001	.000	19.472	.993
totalacc	2985.604	117	25.518	14.415	<.001	.004	1686.519	1.000
initialliststatus * totalacc	499.909	101	4.950	2.796	<.001	.001	282.390	1.000
Error	700660.899	395792	1.770					
Total	2019699.000	396012						
Corrected Total	704528.082	396011						

a. R Squared = .005 (Adjusted R Squared = .005)

b. Computed using alpha = .05

According to these results, for all cases, p-value is <0.001, so we accept H_1 and reject H_0.

3. Do the Debt rank and the public record bankruptcies, both have considerable effect on the DTI ratio of the clients?

H_0:	There are **no** significant differences between the factor levels of Debt rank There are **no** significant differences between the factor levels of public record bankruptcies

	DTI ratio has **no** effect on the effect of public record bankruptcies
H_1:	There **are** significant differences between the factor levels of Debtrank There **are** significant differences between the factor levels of publicrecord bankruptcies Debt rank **has** effect on the effect of public record bankruptcies

Tests of Between-Subjects Effects

Dependent Variable: Debt-to-Income(Ratio)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[b]
Corrected Model	957659,867[a]	56	17101,069	268,308	<,001	,037	15025,247	1,000
Intercept	25321,221	1	25321,221	397,278	<,001	,001	397,278	1,000
graderank	1880,101	6	313,350	4,916	<,001	,000	29,498	,993
pubrecbankruptcies	11473,807	9	1274,867	20,002	<,001	,000	180,019	1,000
graderank * pubrecbankruptcies	7450,912	41	181,730	2,851	<,001	,000	116,901	1,000
Error	25237316,986	395962	63,737					
Total	145388440,81	396019						
Corrected Total	26194976,853	396018						

As we can see from the results of the two-factor ANOVA the observed power of oursamples is 100%, except for debt rank which is 99.3% (Almost 100%).

When taking into consideration the results that we got for **p-value**,
we can see that all the variables (dependent and independents) are less than 5%even the relation between Debt Rank and Public Record Bankruptcies. So that's why we must **reject H_0 and we accept H_1.**

As a conclusion we can say that **all the factors have a significant effect on the DTIratio**

Non-Parametric Tests

Since most of our data are not normally distributed we proceed with the Non-parametric Tests inorder to draw assumptions for our data and variables.

In order to conduct some of the tests we transform the data from scale to ordinal.

Mann-Whitney U test

- Are there differences in the amount of the monthly installment our clients must pay between the 2 different term periods in which the repayment of the loan must occur?

H_0:	The sum of the rankings of installment inthe two groups **does not** differ in the population
H_1:	The sum of the rankings of installment **differs** in the two groups in the population

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The distribution of Installment is the same across categories of Term.	Independent-Samples Mann-Whitney U Test	<,001	Reject the null hypothesis.

a. The significance level is ,050.

b. Asymptotic significance is displayed.

As we can see, **p-value** is less that 5%, so we **reject H_0** and we **accept H_1**. So, the distribution of the monthly installment is not equal across the different terms that we have available.

- Are there differences in the amount of the interest rates of our clients contracts between the 2 different term periods in which the repayment of the loan must occur?

H_0:	The sum of the rankings of interest rate inthe two groups **does not** differ in the population
H_1:	The sum of the rankings of interest rate **differs** in the two groups in the population

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The distribution of Interest Rate is the same across categories of Term.	Independent-Samples Mann-Whitney U Test	<,001	Reject the null hypothesis.

a. The significance level is ,050.

b. Asymptotic significance is displayed.

As we can see, **p-value** is less that 5%, so we **reject H_0** and we **accept H_1**. So, the distribution of the interest rate across the 2 categories of term, 36 months and 60 months differs significantly.

- Are there differences in the number of credit lines our clients have between the 2 different initial listing status in which the bank provides the loan?

H_0:	The sum of the rankings of number of credit lines in the two groups **does not** differ in the population
H_1:	The sum of the rankings of number of credit lines **differs** in the two groups in the population

Mann-Whitney Test

Ranks

	Initial Listing Status of the Loan	N	Mean Rank	Sum of Ranks
Total Number of Credit Lines	Fractional Program	157958	151587.15	2.39E+10
	Whole Program	157957	164328.89	2.60E+10
	Total	315915		

Test Statistics[a]

	Total Number of Credit Lines
Mann-Whitney U	1.147E+10
Wilcoxon W	2.394E+10
Z	-39.279
Asymp. Sig. (2-tailed)	<.001

a. Grouping Variable: Initial Listing Status of the Loan

As we can see, **p-value** is less that 5%, so we **reject H_0** and we **accept H_1**. So, the number of credit lines is not equal across the different initial listing status that we have available.

KRUSKAL-WALLIS Test

Variables: Debt to Income ratio - Debt Grade

- **H_0:** There is **no difference** in the Debt to Income ratio of the clients according to the Debt Grade.
- **H_1:** There is a difference in the Debt to Income ratio of the clients according to the Debt Grade.

Nonparametric Tests

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The distribution of Debt-to-Income (Ratio) is the same across categories of Grade/Rank.	Independent-Samples Kruskal-Wallis Test	<,001	Reject the null hypothesis.

a. The significance level is ,050.

b. Asymptotic significance is displayed.

The result shows that **p-value** is less that 5%, so we **reject H_0** and we **accept H_1**. So it means that at least one group differs in ranks sums. There is a difference between groups (Debt Grade:A, B, C, D, E, F and G)

Conclusion: There is a **difference** in the Debt to Income ratio of the clients according the Debt Grade. It has sense, because the lower the risk is (Debt Rank = A) the lower the Debt to Income ratio should be. In other words: When the Debt Grade is "G" the Debt to Income ratio should be higher.

Correlation

- **Interest rate - Installment**

As these two variables do not follow normal distribution we should apply the Spearman test.

H_0:	There is **no** correlation between interest rate and installment
H_1:	There **is** correlation between interest rate and installment

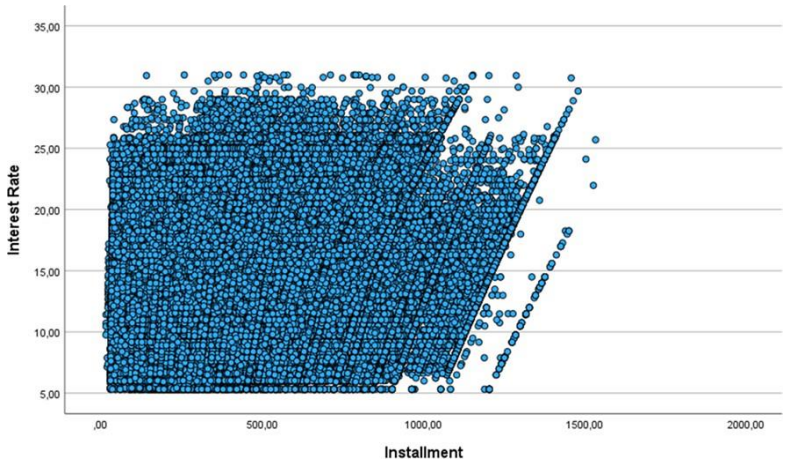
Correlations

			Interest Rate	Installment
Spearman's rho	Interest Rate	Correlation Coefficient	1,000	,137**
		Sig. (2-tailed)	.	<,001
		N	396028	396024
	Installment	Correlation Coefficient	,137**	1,000
		Sig. (2-tailed)	<,001	.
		N	396024	396026

**. Correlation is significant at the 0.01 level (2-tailed).

As we can see, **p-value** is less than 5%, so we **reject H_0** and we **accept H_1**. This means that there is correlation between interest rate and the installment that our clients must pay. Yet, we see that the correlation even though positive is weak, around 14%.

We expected this when we took a look at the scatter plot between Interest Rate and Installments at an earlier stage.



- **Initial Listing Status - Number of Credit Lines**

As these two variables do not follow normal distribution we should apply the Spearman test.

H_0:	There is **no** correlation between number of credit lines and initial listing status
H_1:	There **is** correlation between number of credit lines and initial listing status

Correlations

			Initial Listing Status of the Loan	Total Number of Credit Lines
Spearman's rho	Initial Listing Status of the Loan	Correlation Coefficient	1.000	.068**
		Sig. (2-tailed)	.	<.001
		N	396018	396012
	Total Number of Credit Lines	Correlation Coefficient	.068**	1.000
		Sig. (2-tailed)	<.001	.
		N	396012	396024

**. Correlation is significant at the 0.01 level (2-tailed).

As we can see, **p-value** is less than 5%, so we **reject H_0** and we **accept H_1**. This means thatthere is correlation between number of credit lines and initial listing status

Conclusion: Despite a correlation existing, the result shows that it is weak: positive 6.9%.

- **Debt to Income ratio – Annual Income**

As these two variables follow normal distribution we should apply Pearson test.

- **H_0:** There is **no correlation** between Debt to Income ratio and Annual Income.
- **H_1:** There is **correlation** between Debt to Income ratio and Annual Income.

Correlations

		Debt-to-Income(Ratio)	Annual Income
Debt-to-Income(Ratio)	Pearson Correlation	1	-,016**
	Sig. (2-tailed)		<,001
	N	396019	396018
Annual Income	Pearson Correlation	-,016**	1
	Sig. (2-tailed)	<,001	
	N	396018	396029

**. Correlation is significant at the 0.01 level (2-tailed).

As we can see, **p-value** is less than 5%, so we **reject H_0** and we **accept H_1**. This means that there is correlation between Debt to Income ratio and Annual Income.

Conclusion: Despite a correlation existing, the result shows that it is weak: negative 16%. We expected a stronger correlation.

Regression

Linear

- **Interest rate - Installment**

Above, we saw that there is a weak correlation between the interest rate and the monthly installment (around 14%), yet how does this appear in a mathematical formula? How does interest rate affect the monthly installment?

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,163[a]	,026	,026	247,38091

a. Predictors: (Constant), Interest Rate

b. Dependent Variable: Installment

The $\mathbf{R^2}$ leads us to conclude that our model has a small percentage of variability, so it will be able to explain only a small percentage of the variance of observations. The Adjusted $\mathbf{R^2}$ is very low, around 2.6% as well, which leads us to assume, again, that since the variability of our observations is low our model won't be sufficient in us deciphering the full effect that these 2 variables share. Most of the variances can't be covered by our model.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	307,385	1,262		243,627	<,001
	Interest Rate	9,125	,088	,163	103,814	<,001

a. Dependent Variable: Installment

When checking the **p-value** of the 2 coefficients, we see that they are less than 5% which means that both of them have a significant effect on our dependent variable of installment.

The mathematical formula for the effect of interest rate on installment is:

$$\hat{y} = 9{,}125x + 307{,}385$$

- **Initial Listing Status - Number of Credit Lines**

Above, we saw that there is a weak correlation between the Number of Credit Lines and Initial Listing Status. How is Initial Listing status affected?

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.069[a]	.005	.005	.499

a. Predictors: (Constant), Total Number of Credit Lines

b. Dependent Variable: Initial Listing Status of the Loan

The $\mathbf{R^2}$ leads us to conclude that our model has a small percentage of variability, so it will be able to explain only a small percentage of the variance of observations. The Adjusted $\mathbf{R^2}$ is very low, around 0.5% as well, which leads us to assume, again, that since the variability of our observations is low our model won't be sufficient in us deciphering the full effect that these 2 variables share. Most of the variances can't be covered by our model.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	.426	.002		202.747	<.001
	Total Number of Credit Lines	.003	.000	.069	38.901	<.001

a. Dependent Variable: Initial Listing Status of the Loan

When checking the **p-value** of the 2 coefficients, we see that they are less than 5% which means that both of them have a significant effect on our dependent variable of initial listing status.

The mathematical formula for the effect of interest rate on installment is:

$$\hat{y} = 0.003x + 0.426$$

- **Debt to Income ratio – Annual Income**

Above, we saw that there is a weak correlation between the Debt to Income ratio and Annual Income. But must Annual Income affect Debt to Income ratio?

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,016[a]	,000	,000	8,13195

a. Predictors: (Constant), Annual Income

b. Dependent Variable: Debt-to-Income(Ratio)

The result shows us that the Adjusted $\mathbf{R^2}$ is almost 1% so we can confirm that since the variability of our observations is low our model won't be sufficient in us deciphering the full effect that these 2 variables share. Most of the variances can't be covered by our model.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	17,375	,013		1318,817	<,001
	Annual Income	-2,289E-7	,000	-,016	-10,296	<,001

a. Dependent Variable: Debt-to-Income(Ratio)

When checking the **p-value** of the 2 coefficients, we see that they are less than 5% which means that the independent variable has effect on the dependent one.

Conclusion: The Annual Income has an effect/influence on Debt to Income ratio although a very weak negative one.

Multiple

Above, we saw that interest rate does in fact affect the amount of installment, so now, it's time for us to answer a more complicated question; Does interest rate and annual income affect the monthly installment that our clients are called to pay?

The addition of the variable term will most probably help us increase the percentage of Adjusted $\mathbf{R^2}$, which will lead to a better balanced model that will take into consideration all the factors that have a considerable effect on the amount of installments.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,171[a]	,029	,029	247,03889

a. Predictors: (Constant), Annual Income, Interest Rate

b. Dependent Variable: Installment

Still, we see that even though we added one more independent variable the percentage of both $\mathbf{R^2}$ and Adjusted $\mathbf{R^2}$ increased only marginally, from 2.6% to 2.9%. This means that still our model can't explain a big percentage of the variability of our observations and still more independent variables need to be added..

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	304,834	1,262		241,490	<,001
	Interest Rate	9,123	,088	,163	103,928	<,001
	Annual Income	2,239e-5	,000	,052	33,156	<,001

a. Dependent Variable: Installment

Yet, when checking the **p-value** of the 3 coefficients, we see that they are less than 5% which means that all of them have a significant effect on our dependent variable of installment. Though the annual income has a very weak effect on the amount of installments.

The mathematical formula for the effect of interest rate and annual income on installment is:

$$\hat{y} = 9{,}125x_1 + 0{,}00002239x_2 + 307{,}385$$

Conclusion

To conclude, we can see from the non-parametric, and regression tests that the variables do affect the specifics of the loan such as initial listing status and monthly installments. This is further supported when looking at the correlations between these variables with number of credit lines and interest rate. However, the effect is too small to conclude that they are deciding factors. More investigation needs to be done on a larger set of variables to see how the variables are being coordinated to affect these decisions. While that is being said, we can see that annual income and debt to income ratio specifically are a large part of this decision, and further analysis can show how they relate to the other variables in more detail.