26 November, 2022

Brainnest Data Analysis Industry training

Lending Club Loan Data Analysis

- Diego Gules Butori
- Nairuhi Tovmasyan
- Jaswinder Singh
- Federico Andrés Gómez Quiroga
- Cian O'Sullivan

Table of Contents

LendingClub is the world's leading online marketplace for connecting borrowers and investors.

- Lending Club is a Peer-to-Peer lending company that utilizes a group of private investors to fund loan requests.
- Lending Club assigns each borrower a grade and subgrades based on their credit history.
- Investors are presented with a list of loan requests along with their grades and borrower details. Then they select loan requests they will fund/partially fund.
- Lending Club makes money by charging borrowers an origination fee and a service fee to investors.

The objective of this project is to clean, visualize and interpret the results of our analysis of the Lending club dataset under consideration.

Brief Overview of Dataset

The dataset contains 396030 rows and 27 columns. The columns in the dataset have different data types like int, float and object. A snippet of the dataset is shown in the picture below.

COLUMNS

- loan_amnt
- term
- interest_rate
- instalment
- grade/rank
- sub_grade
- emp_title
- emp_length
- home_ownership
- annual_income
- verification_status
- issue_date
- loan_status
- purpose
- title
- dti
- earliest_cr_line
- open_acc
- pub_rec
- revol_bal
- revol_util
- total_acc
- initial_list_status
- application_type
- mort_acc
- pub_rec_bankruptcies
- address'

loan_amnt	term	interest_rate	installment	grade/rank	sub_grade	emp_title	emp_length	home_ownership	annual_income	...	open_acc	pub_rec	revol_bal	revol_util
10000.0	36 months	11.44	329.48	B	B4	Marketing	10+ years	RENT	117000.0	...	16.0	0.0	36369.0	41.8
8000.0	36 months	11.99	265.68	B	B5	Credit analyst	4 years	MORTGAGE	65000.0	...	17.0	0.0	20131.0	53.3
15600.0	36 months	10.49	506.97	B	B3	Statistician	< 1 year	RENT	43057.0	...	13.0	0.0	11987.0	92.2
7200.0	36 months	6.49	220.65	A	A2	Client Advocate	6 years	RENT	54000.0	...	6.0	0.0	5472.0	21.5
24375.0	60 months	17.27	609.33	C	C5	Destiny Management Inc.	9 years	MORTGAGE	55000.0	...	13.0	0.0	24584.0	69.8

Objectives

The key objectives of this project are to utilise the analysis of the Lending Club Loan file to answer the following questions by interpreting the results:

- Do people tend to take short or long-term loans?
- What are the top 5 loan purposes?
- Does the verification status matter in granting a loan?
- Do people with low grades have higher interest rates?
- Is there a relationship between the employment length and the loan amount?

Data Cleaning

Step-1: Removing Irrelevant Data

- address is not necessary for our analysis as it does not provide any meaningful information for predicting any of the other variables. This could have been used if we were going to do some clustering based on the area but this is not our objective. So we drop this column.

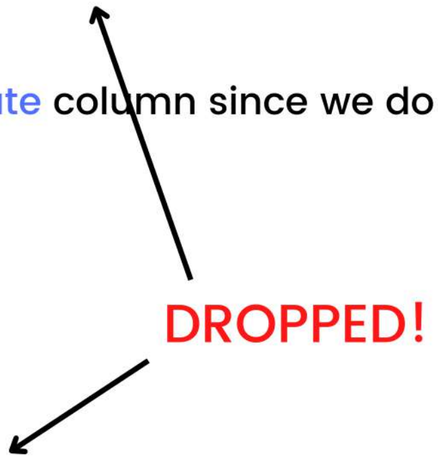
- When we show the counts of each category for application_type variable, we see below that it has three categories and the majority of the cases are of one category, namely INDIVIDUAL. Only a very small portion of the data is of the other categories.

```
INDIVIDUAL    395309
JOINT            425
DIRECT_PAY       285
Name: application_type, dtype: int64
```

Therefore, we can drop application_type column, as only one of its categories is common.

- We have to drop also issue_date column since we do not have the info on how it is represented as it is not a usual date but some number.

```
0    42005.0
1    42005.0
2    42005.0
3    41944.0
4    41365.0
Name: issue_date, dtype: float64
```



- term column has 'months' string after the number. This is irrelevant so we remove it and later, after dealing with missing values, make *term* to an integer datatype.

```
0     36 months
1     36 months
2     36 months
3     36 months
4     60 months
Name: term, dtype: object
```

BEFORE

```
0     36
1     36
2     36
3     36
4     60
Name: term, dtype: object
```

AFTER

- In emp_length we can drop years string in each row and group the entries into 3 categories making emp_length of an ordinal data type. We are doing this because we have 10+ and < 1 year which we cannot replace with an exact number if we want to make the emp_length of an integer data type.
  a. Hence, we can take years from 1 to 10 into one group, so that we will have 3 groups: less than 1, from 1 to 10, and more than 10 years which are ordered as in ordinal data type.

```
10+ years    126041
2 years       35827
< 1 year      31725
3 years       31665
5 years       26495
1 year        25882
4 years       23952
6 years       20841
7 years       20819
8 years       19168
9 years       15314
Name: emp_length, dtype: int64
```

```
10+    126041
2       35827
< 1     31725
3       31665
5       26495
1       25882
4       23952
6       20841
7       20819
8       19168
9       15314
Name: emp_length, dtype: int64
```

```
1 to 10    219963
10+        126041
< 1         31725
Name: emp_length, dtype: int64
```

- We make the column name term as term_months so that we will know later that the term is in months and emp_length as emp_length_years.

- loan_amnt
- term_months
- interest_rate
- instalment
- grade/rank
- sub_grade
- emp_title
- emp_length_years
- home_ownership
- annual_income
- verification_status
- loan_status
- purpose
- title
- dti
- earliest_cr_line
- open_acc
- pub_rec
- revol_bal
- revol_util
- total_acc
- initial_list_status
- mort_acc
- pub_rec_bankruptcies

Current columns

Current shape → (396030, 24)

Step-2: Data Deduplication

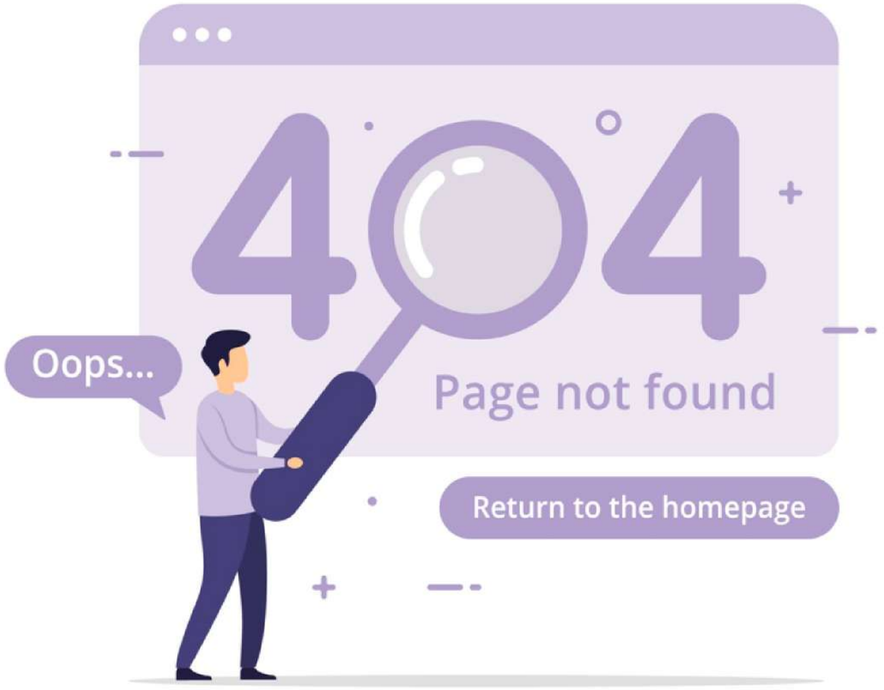
There were no duplicated data entries in our dataset.

```
df.duplicated().sum()
```

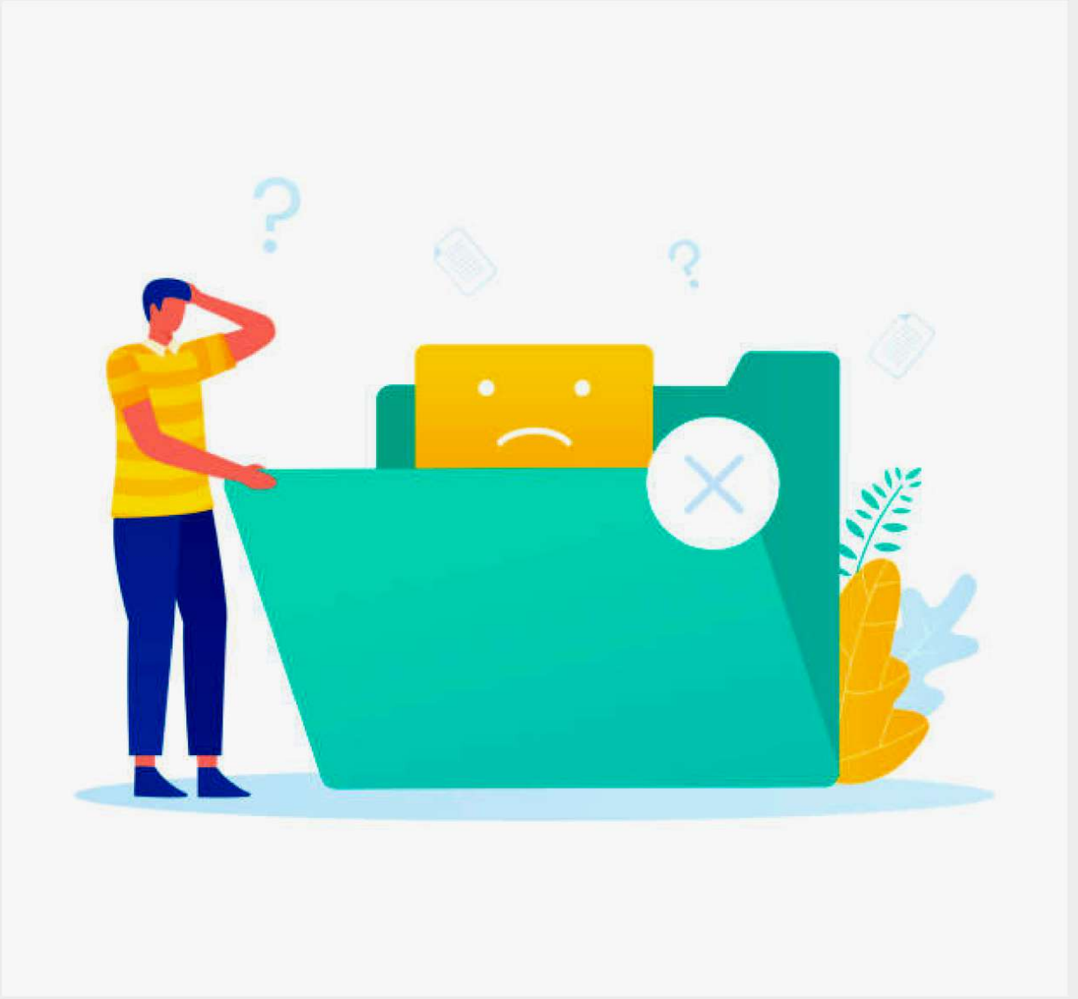
```
0
```

Step-3: Fixing Structural Errors

There were structural errors in our dataset!



Step-4: Dealing with Missing Data



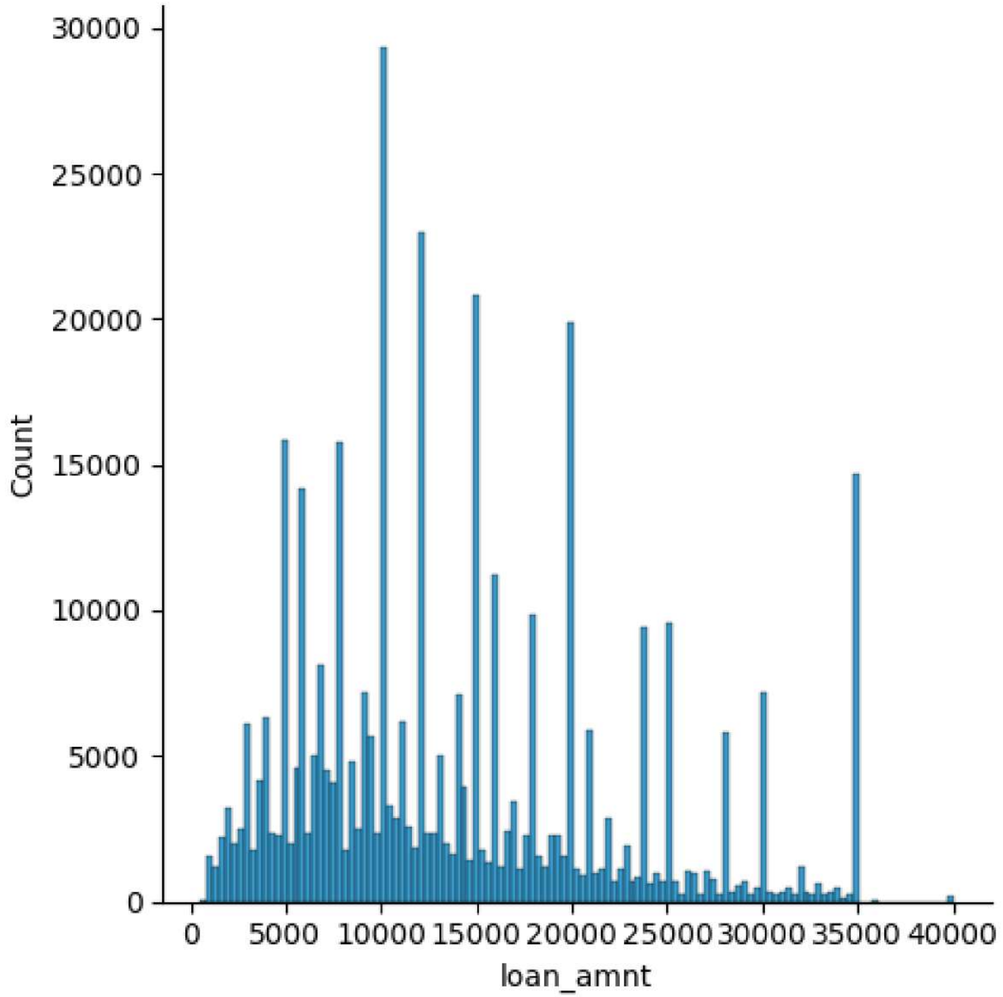
- Number and percentages of missing values in each column

	Percent	Count
loan_amnt	0.000505	2
term_months	0.000505	2
interest_rate	0.000505	2
installment	0.001010	4
sub_grade	0.000253	1
emp_title	5.789460	22928
emp_length_years	4.621115	18301
annual_income	0.000253	1
verification_status	0.000253	1
loan_status	0.001515	6
purpose	0.001768	7
title	0.444663	1761
dti	0.002778	11
earliest_cr_line	0.005303	21
open_acc	0.001768	7
pub_rec	0.000253	1
revol_bal	0.006060	24
revol_util	0.072974	289
total_acc	0.001515	6
initial_list_status	0.003030	12
mort_acc	9.543974	37797
pub_rec_bankruptcies	0.135343	536

We can see that the maximum number of missing values is in the range of 4% to 10%, therefore they are acceptable at first glance.

Columns with the percentage of missing values greater than 1

- We see that loan_amnt is not normally distributed, so we can replace its 2 NA values with median.




```
df['loan_amnt'] = df['loan_amnt'].fillna(df['loan_amnt'].median())
df['loan_amnt'].isna().sum()
```

✓ 0.9s

```
0
```

- term appears in two values: either 36 or 60. So, we can replace 2 NA values with the mode value and make the term an integer data type.

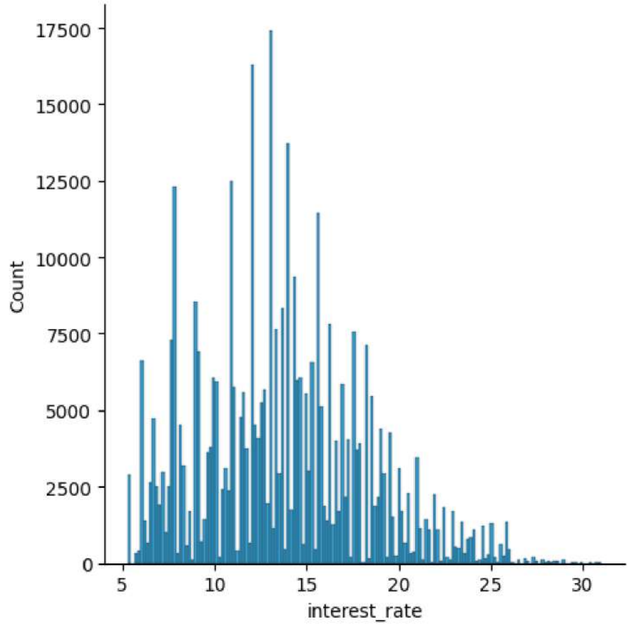
```
 36    302003
 60     94025
Name: term_months, dtype: int64
```

→

```
df['term_months'] = df['term_months'].fillna('36').apply(pd.to_numeric)
df['term_months'].isna().sum()
```

```
0
```

- Below we plot the distribution of the interest_rate and we can see that it is close to normal distribution so we can fill NAs with the mean.

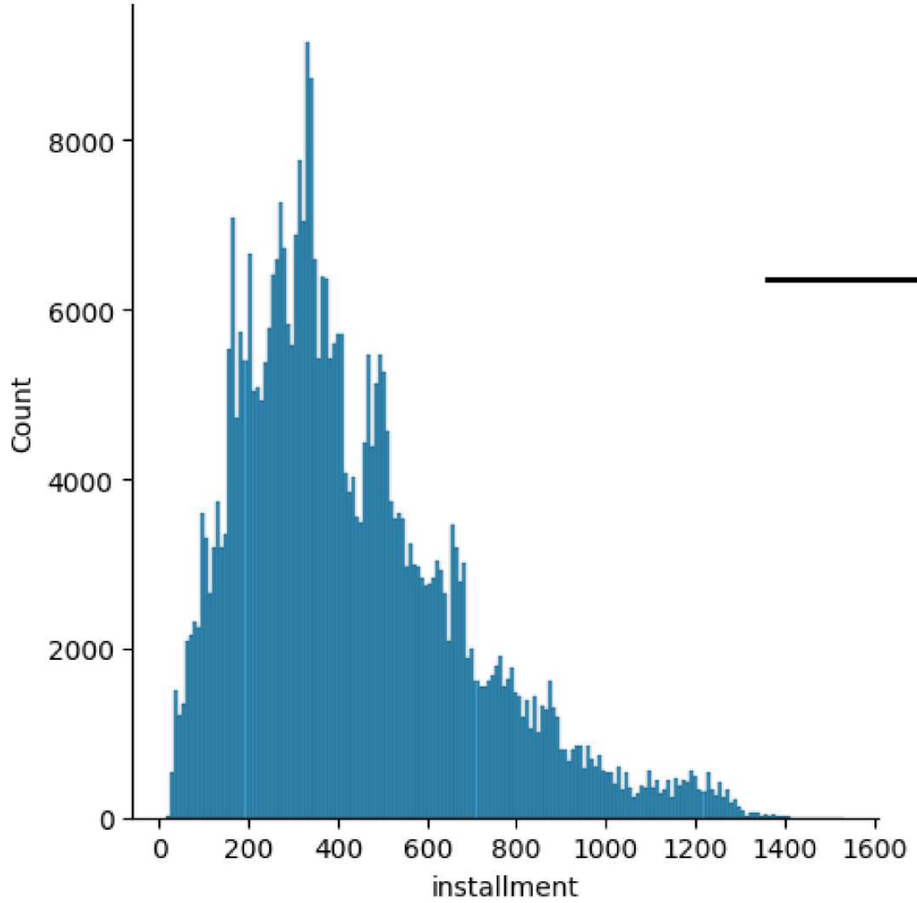


→

```
df['interest_rate'] = df['interest_rate'].fillna(df['interest_rate'].mean())
df['interest_rate'].isna().sum()
```

```
0
```

- Here we do the same thing for the installment but we replace NAs with median because the distribution is more skewed.




```
df['installment'] = df['installment'].fillna(df['installment'].median())
df['installment'].isna().sum()
```

```
0
```

- For annual_income, we replace the NA values by mean since it is a numeric type column.

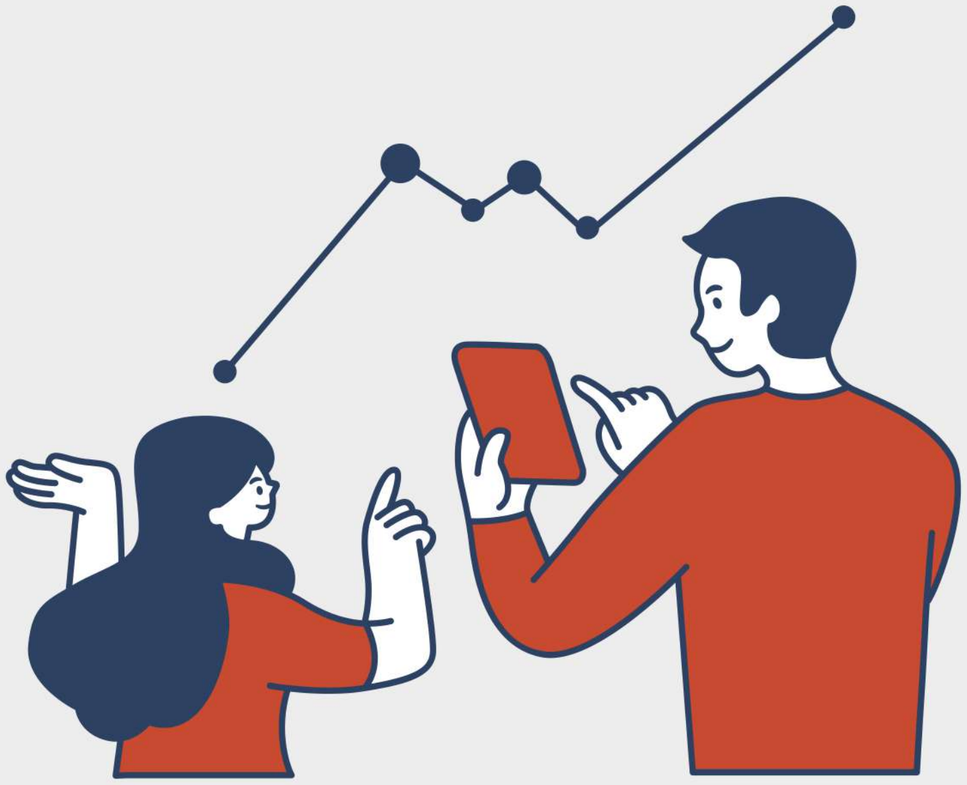
```
count    3.960290e+05
mean     7.420326e+04
std      6.163768e+04
min      0.000000e+00
25%      4.500000e+04
50%      6.400000e+04
75%      9.000000e+04
max      8.706582e+06
Name: annual_income, dtype: float64
```

```
df['annual_income'] = df['annual_income'].fillna(df['annual_income'].mean())
df['annual_income'].isna().sum()
```

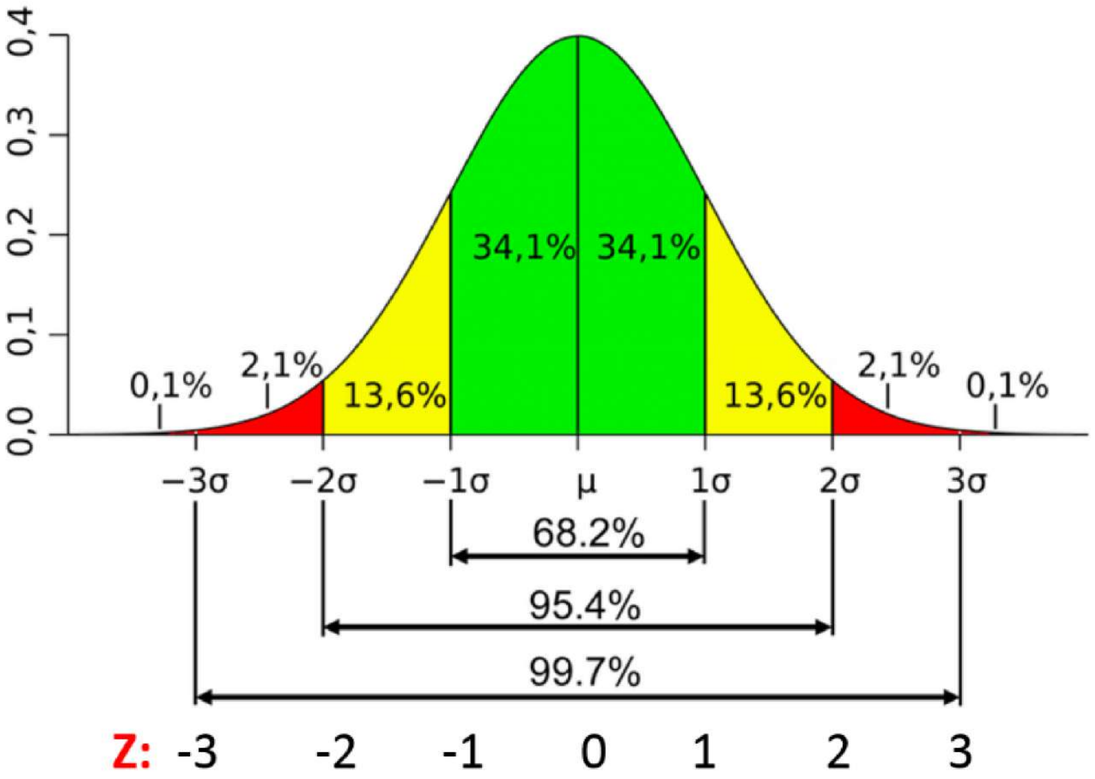
```
0
```

Step-5: Filtering Outliers

Identifying the Odd Balls in the Data

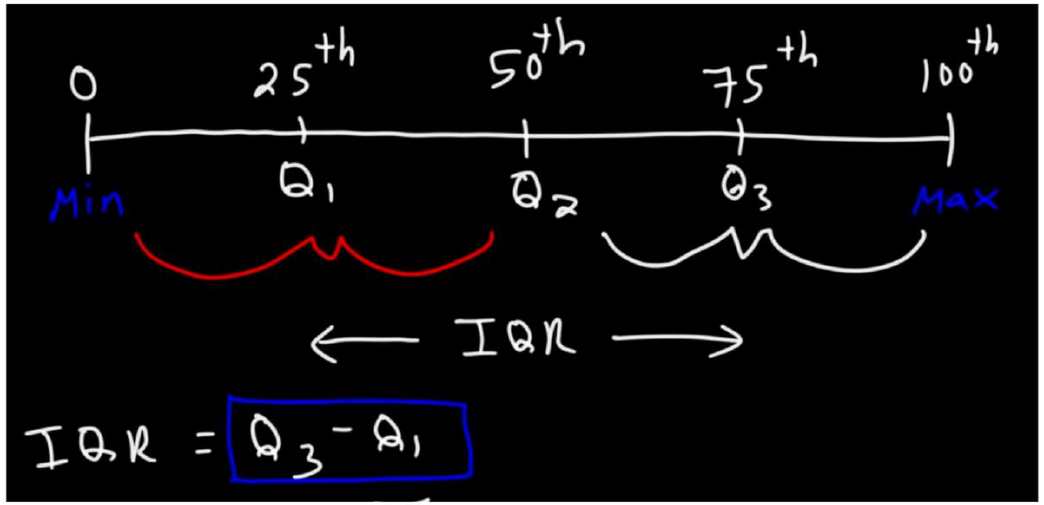


Standard Normal Distribution



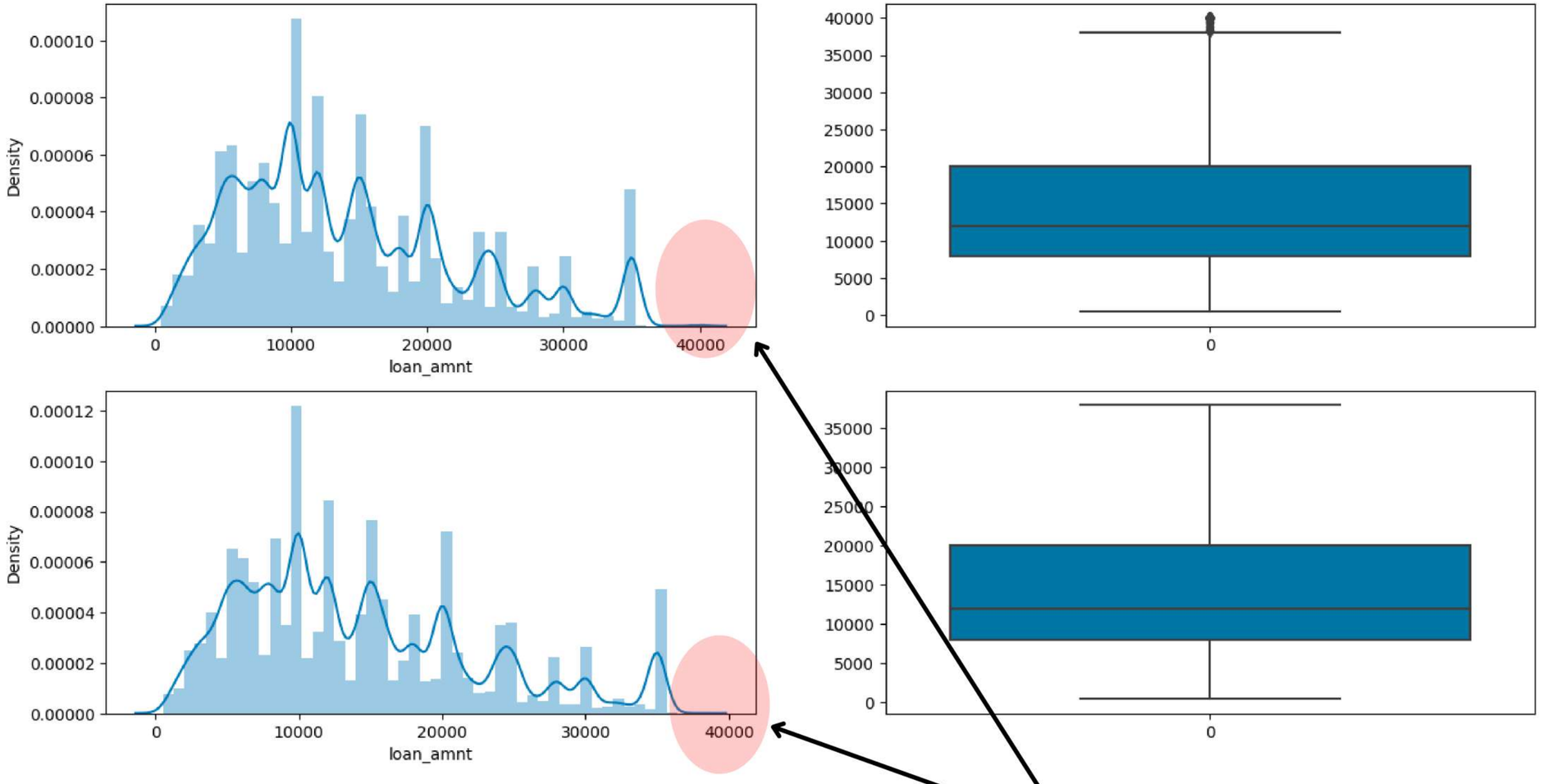
For Normal distributions: The data points which fall below mean-3*(standard_deviation) or above mean+3*(standard_deviation) are outliers.

Interquartile Range (IQR)



For skewed distributions: Use **Inter-Quartile Range (IQR) proximity rule**.
The data points which fall below Q1 – 1.5 IQR or above Q3 + 1.5 IQR are outliers, where Q1 and Q3 are the 25th and 75th percentile of the dataset respectively,

- We observe and trim the outliers in the column loan_amnt, since this column is the main focus of our analysis



Note: Trimmed Outliers!

Descriptive Analysis

Interpreting the Results

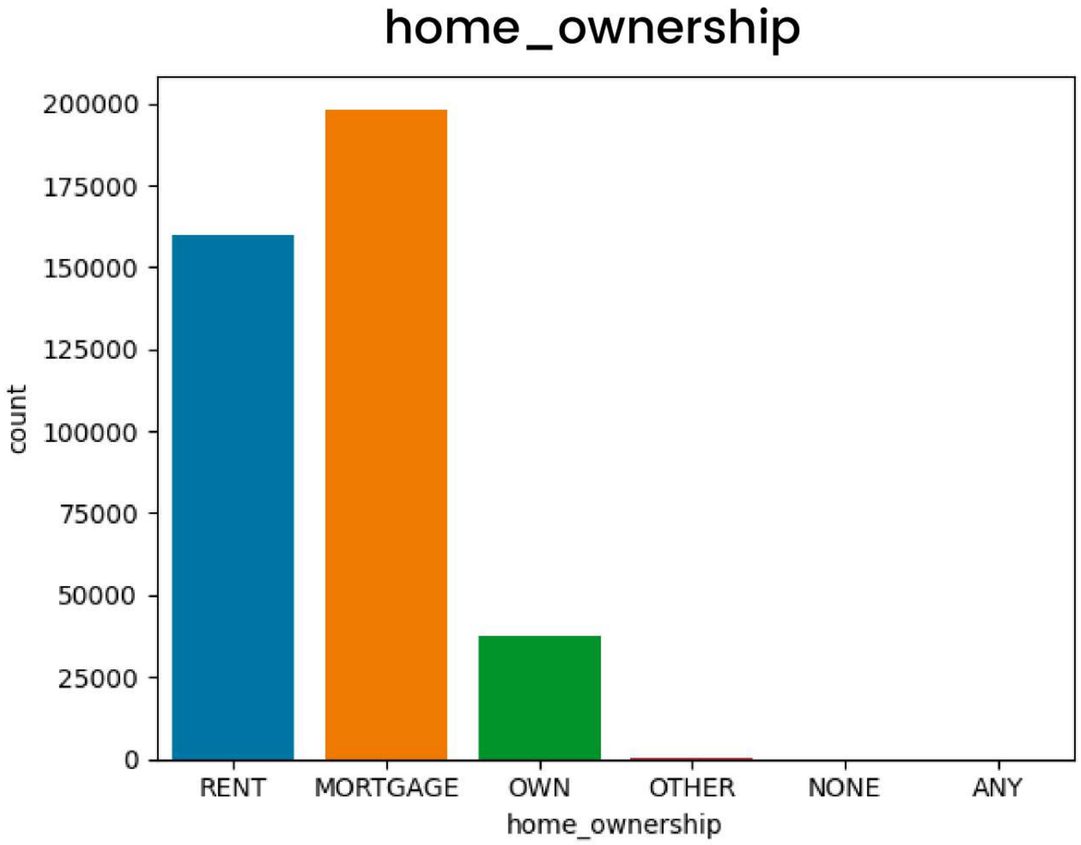
- Mean, median, Standard Deviation and count of few columns in the dataset

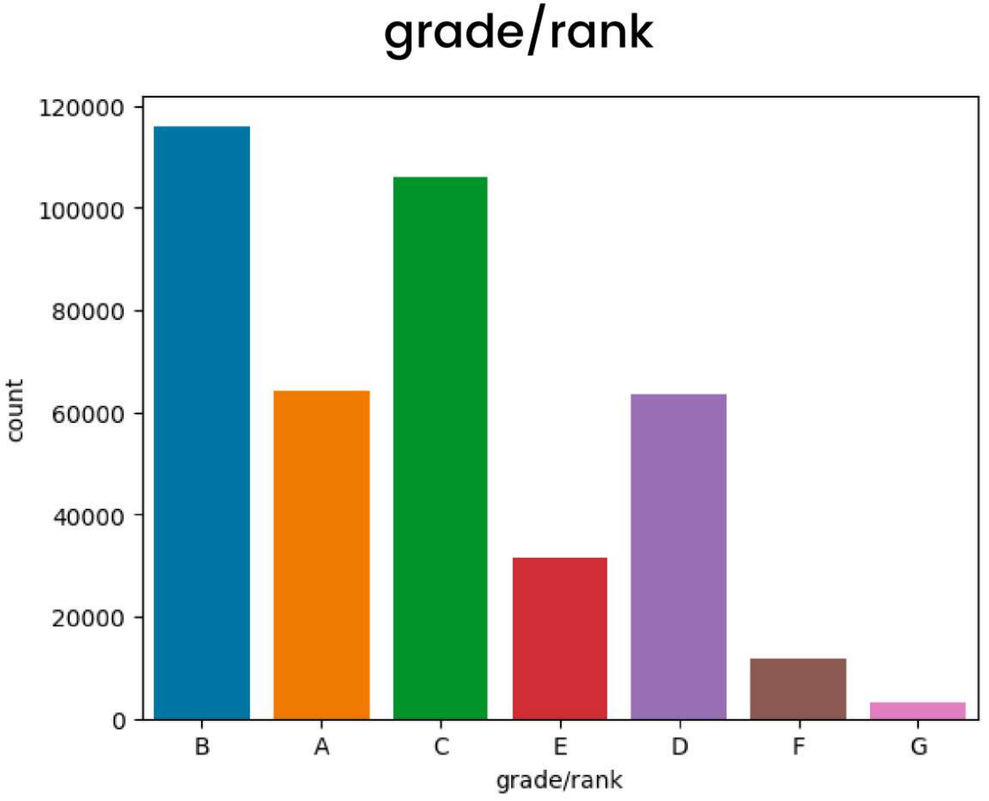
```
df.describe()
```

Python

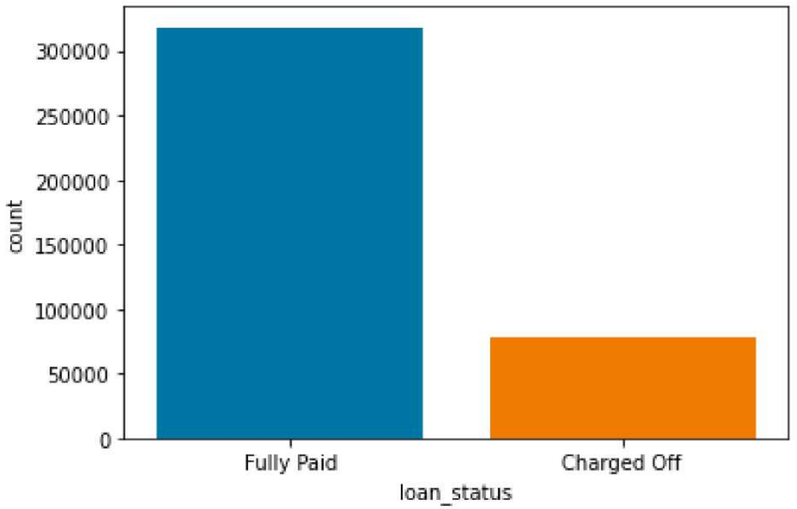
	loan_amnt	interest_rate	installment	annual_income	dti	earliest_cr_line	open_acc	pub_rec	revol_bal	revol_util	total_acc
count	396028.000000	396028.000000	396026.000000	3.960290e+05	396019.000000	396009.000000	396023.000000	396029.000000	3.960060e+05	395741.000000	396024.000000
mean	14113.898765	13.639413	431.847261	7.420326e+04	17.379524	35918.408488	11.311171	0.178191	1.584428e+04	53.791744	25.414662
std	8357.458375	4.472158	250.723864	6.163768e+04	18.019304	2629.915550	5.137659	0.530671	2.059207e+04	24.452117	11.886998
min	500.000000	5.320000	16.080000	0.000000e+00	0.000000	16072.000000	0.000000	0.000000	0.000000e+00	0.000000	2.000000
25%	8000.000000	10.490000	250.330000	4.500000e+04	11.280000	34608.000000	8.000000	0.000000	6.025000e+03	35.800000	17.000000
50%	12000.000000	13.330000	375.430000	6.400000e+04	16.910000	36404.000000	10.000000	0.000000	1.118100e+04	54.800000	24.000000
75%	20000.000000	16.490000	567.300000	9.000000e+04	22.980000	37712.000000	14.000000	0.000000	1.961975e+04	72.900000	32.000000
max	40000.000000	30.990000	1533.810000	8.706582e+06	9999.000000	41548.000000	90.000000	86.000000	1.743266e+06	892.300000	151.000000

Frequency Distributions

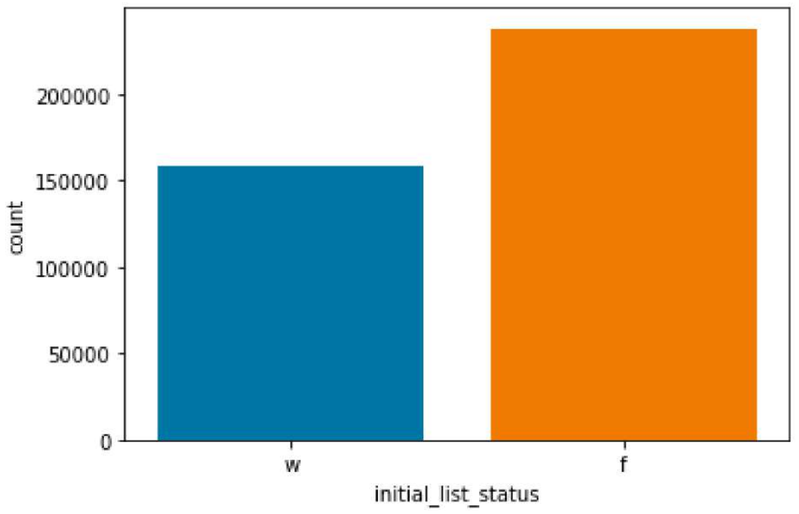




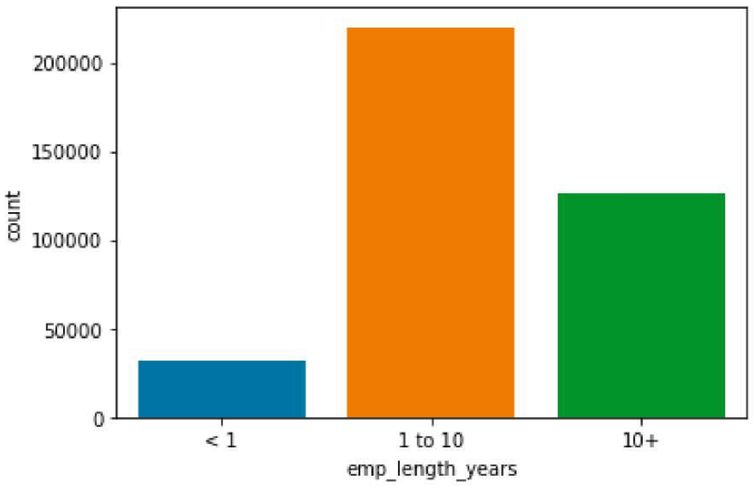
loan_status



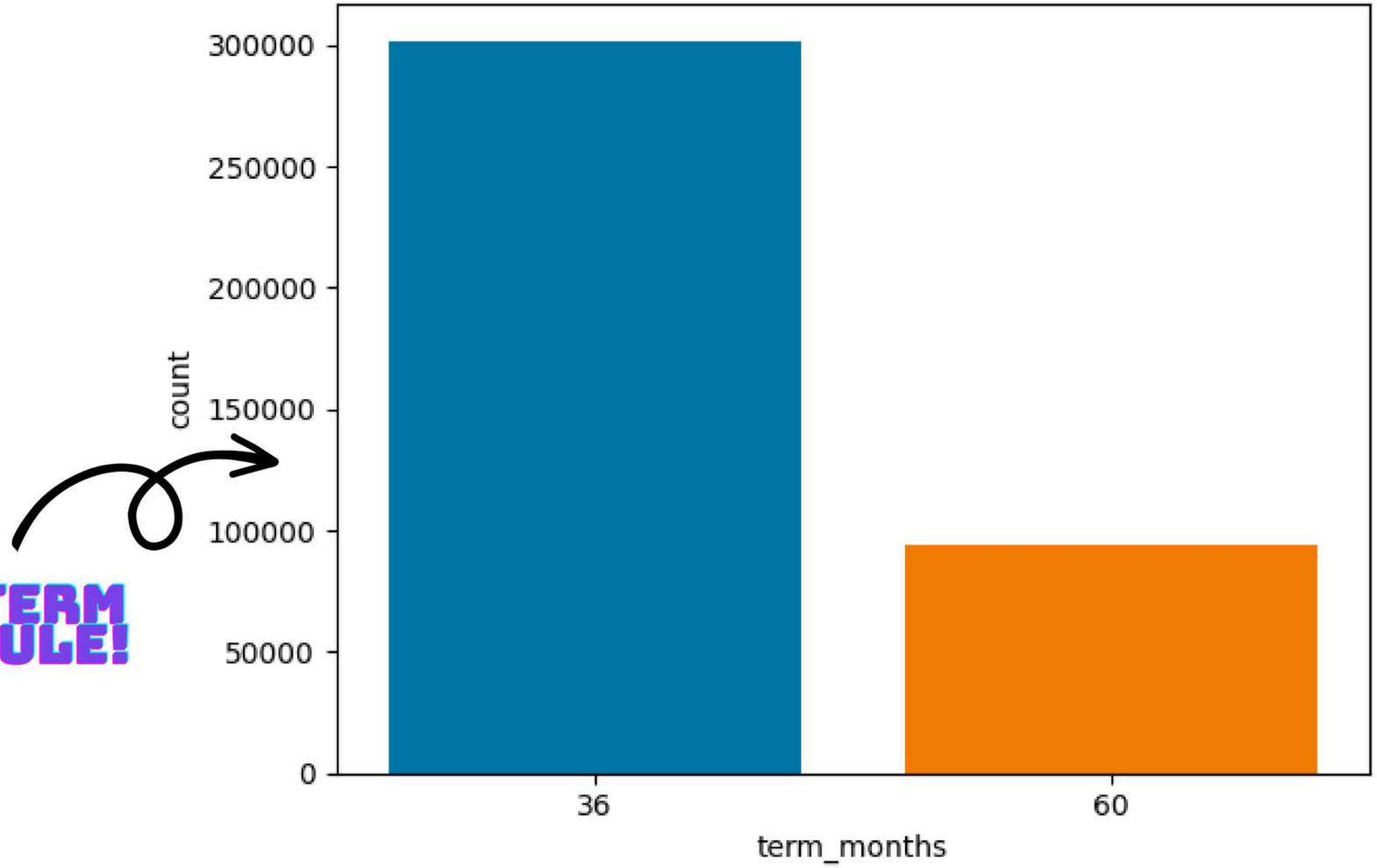
initial_list_status



emp_length_years

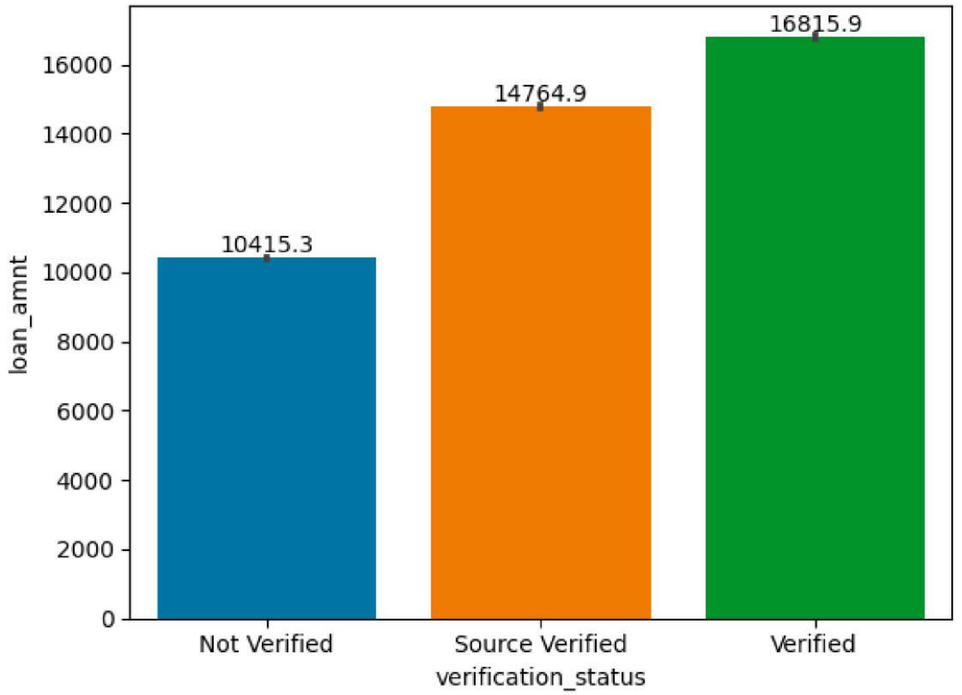


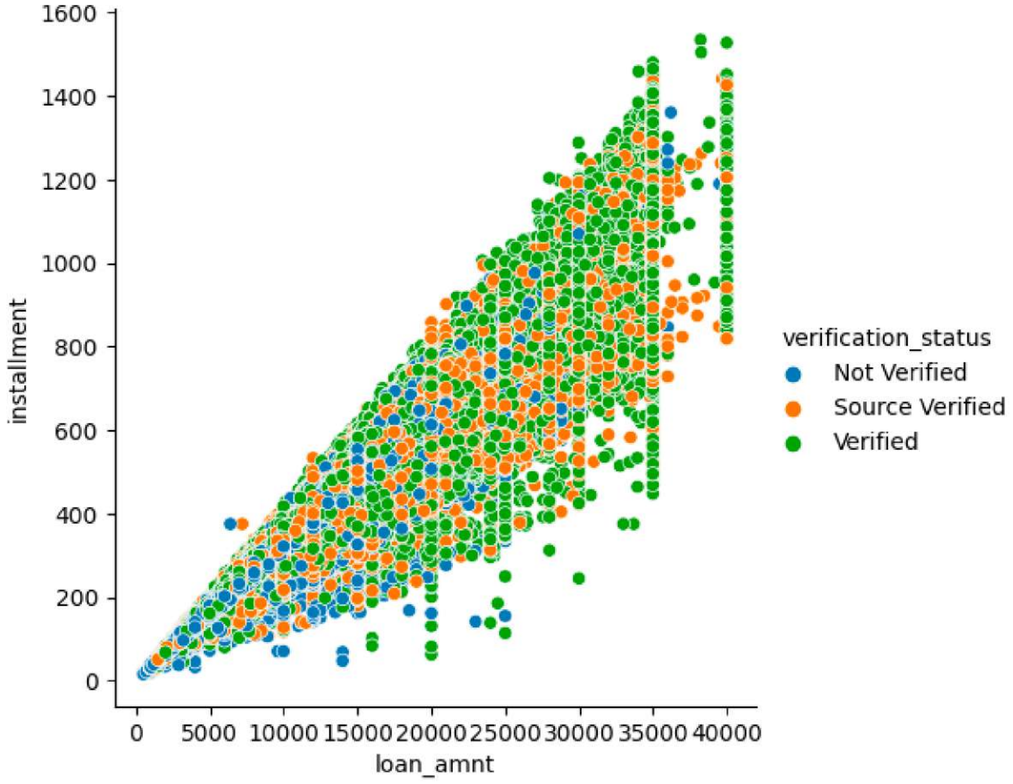
Do people tend to take out short-term or long-term loans?



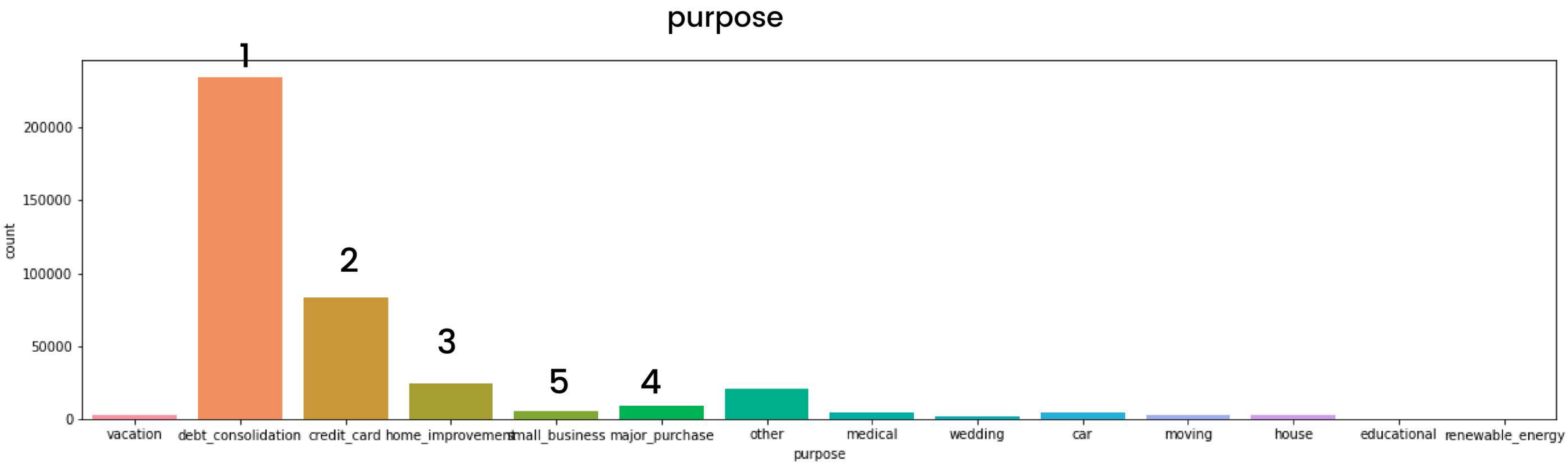
SHORT-TERM LOANS RULE!

verification_status



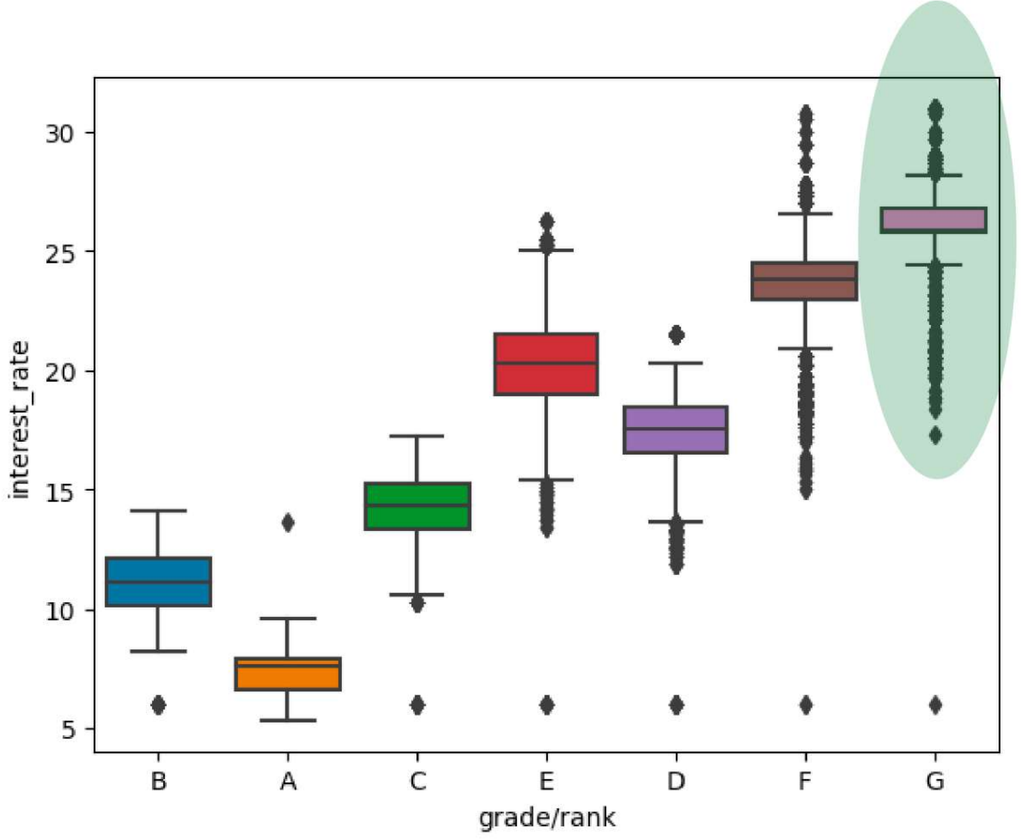


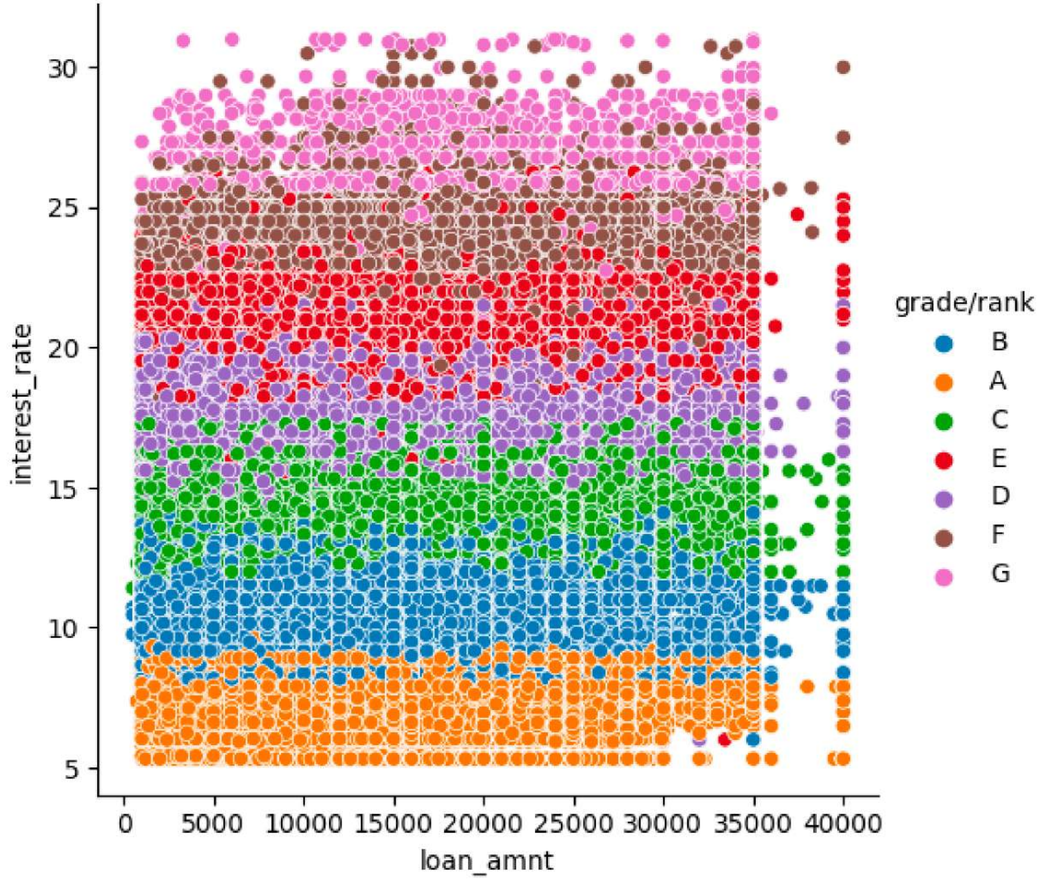
Does verification status matter in granting a loan?



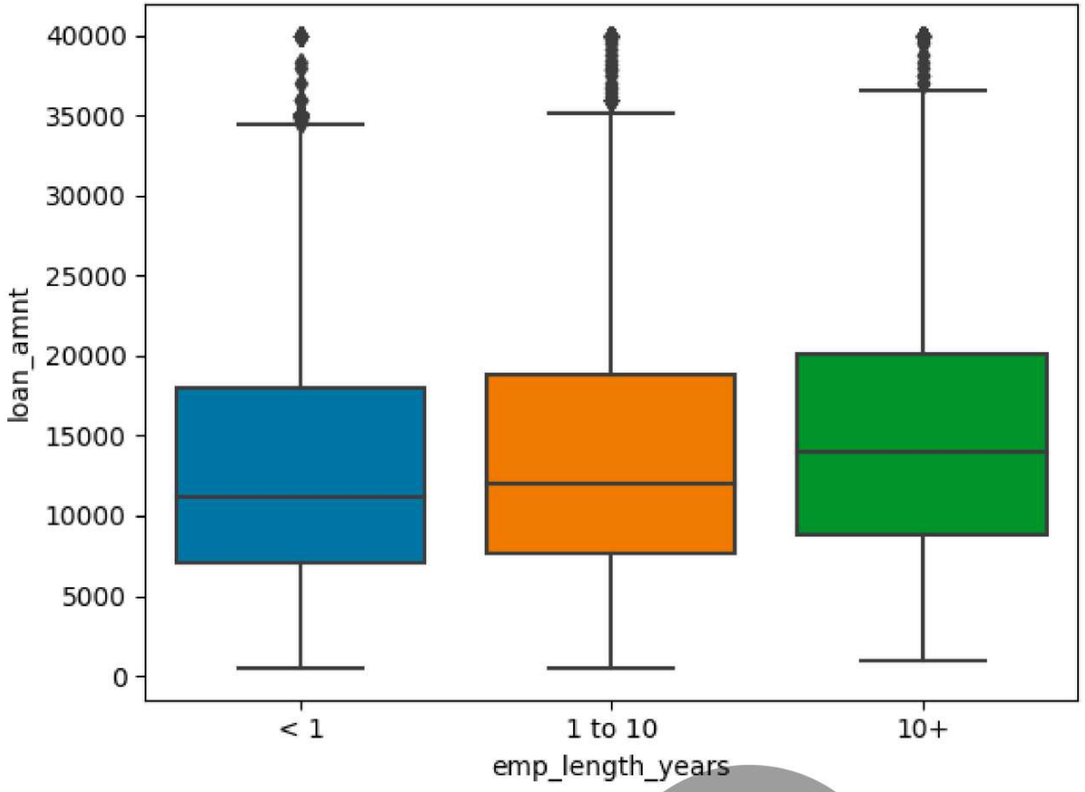
What are the top 5 loan purposes?

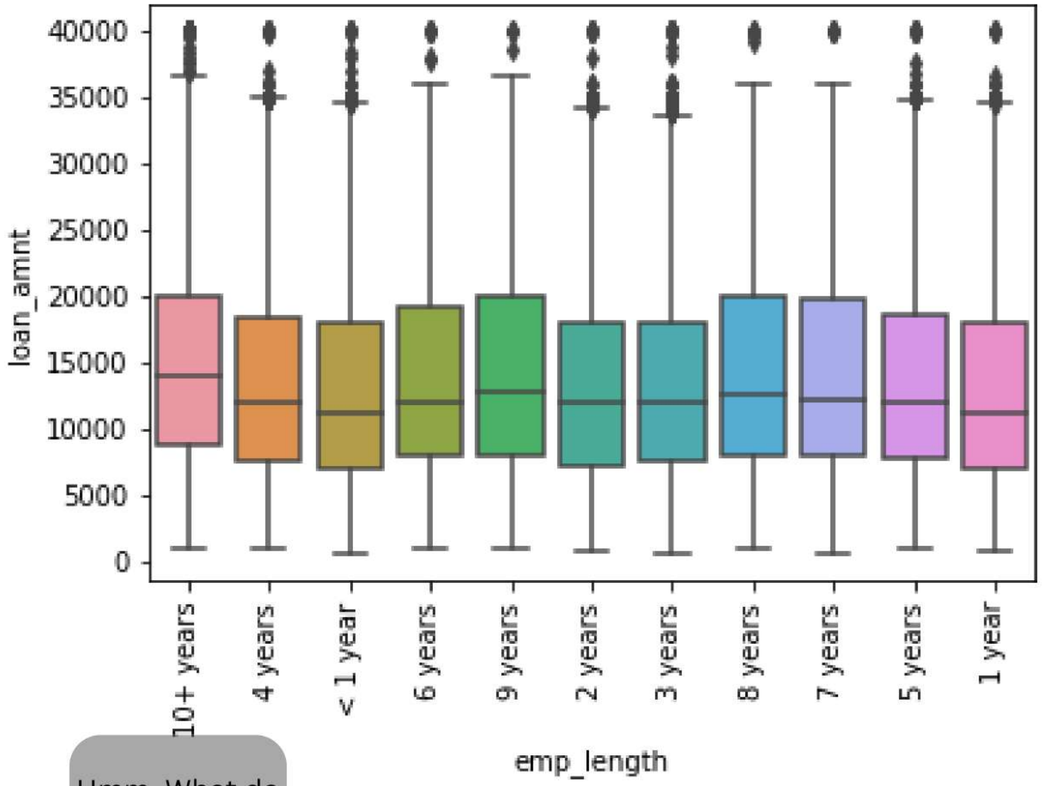
Do people with low grade have higher interest rates?

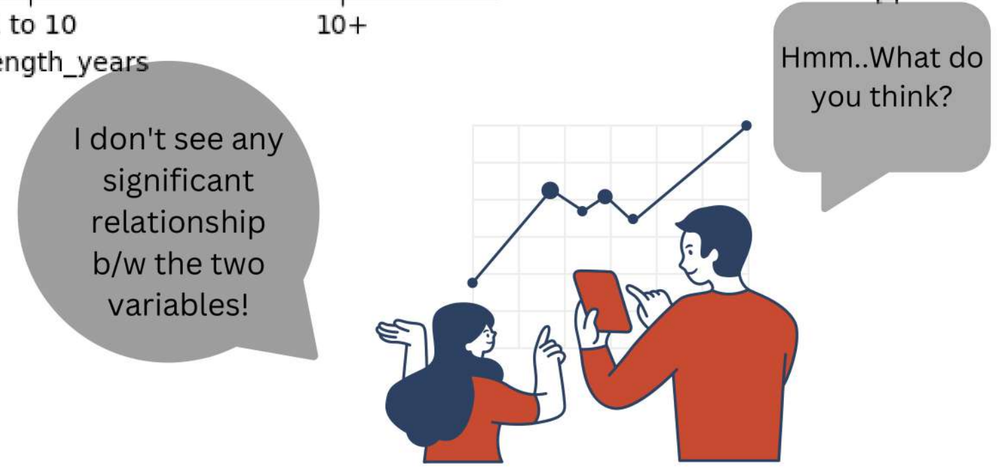




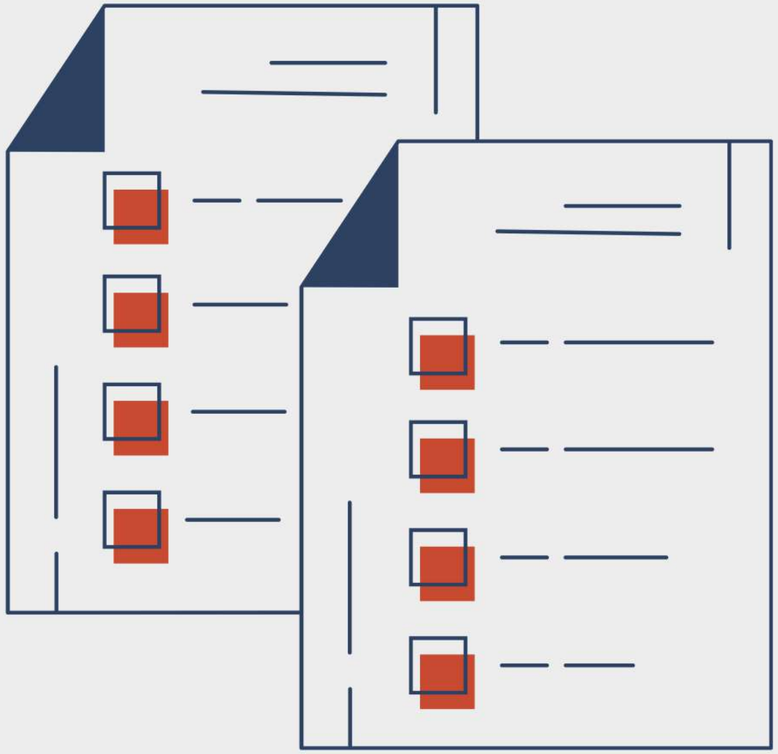
Is there a relationship between the employment length and the loan amount?







Cross Tabulation



loan_status	Charged Off	Fully Paid	Total
emp_length_years			
< 1	6563	25162	31725
1 to 10	42857	177101	219958
10+	23215	102825	126040
Total	72635	305088	377723

loan_status	Charged Off	Fully Paid	Total
verification_status			
Not Verified	0.046227	0.269611	0.315838
Source Verified	0.071243	0.260515	0.331759
Verified	0.078662	0.273742	0.352404
Total	0.196133	0.803867	1.000000

loan_status	Charged Off	Fully Paid	Total
purpose			
car	633	4063	4696
credit_card	13874	69140	83014
debt_consolidation	48639	185861	234500
educational	42	215	257
home_improvement	4087	19943	24030
house	434	1767	2201
major_purchase	1448	7342	8790
medical	911	3285	4196
moving	670	2184	2854
other	4495	16690	21185
renewable_energy	77	252	329
small_business	1679	4022	5701
vacation	464	1988	2452
wedding	219	1593	1812
Total	77672	318345	396017

Conclusion

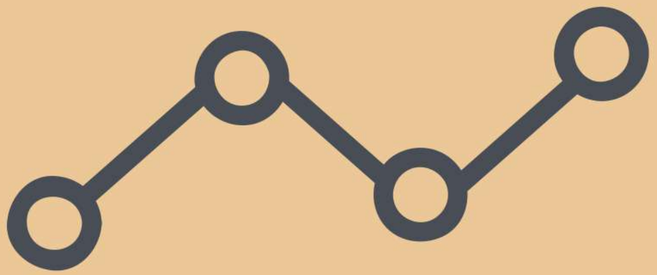
- We have cleaned, analyzed and transformed the given Lending Club loan dataset and found some interesting insights.
- The following set of questions was defined and answered during the analysis of the dataset:
  - Do people tend to take short or long-term loans?
    - Yes. People prefer short-term loan plans (36 months) rather than long-term (60 months).
  - Does the verification status matter in granting a loan?
    - The verification status does **NOT** appear to have a significant effect on the loan amount but the data shows that those customers who are verified tend to be granted a relatively higher loan amount than those who are not.
  - What are the top 5 loan purposes?
    - The top 5 loan purposes are debt consolidation, credit card, home improvements, major purchase and small business.
  - Do people with low grades have higher interest rates?
    - Yes. The people with the lowest grade 'G' have on average the highest interest rate
  - Is there a relationship between the employment length and the loan amount?
    - There does **NOT** seem to be any significant relationship between the two variables. We compared both the grouped and non-grouped employment length values boxplots to verify this.

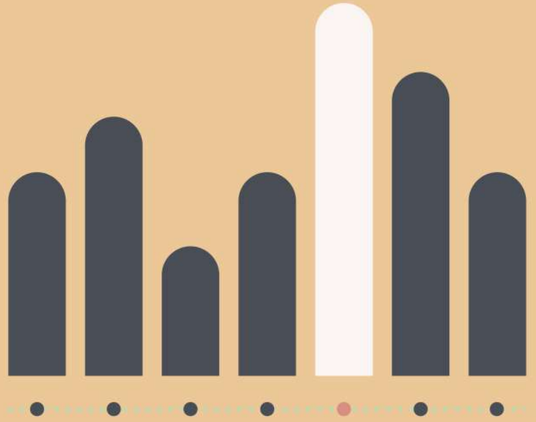
Thank you for listening

Questions?

NORMALITY TESTS FOR LENDING CLUB LOAN DATA

- Diego Gules Butori
- Nairuhi Tovmasyan
- Jaswinder Singh
- Federico Andrés Gómez Quiroga
- Cian O'Sullivan





OBJECTIVES

This presentation aims to test the normality of different variables in our dataset by using normality tests and descriptive plots. In particular, we will analyze the normality through the following methods:

- Histogram
- Q-Q and P-P plots
- Boxplots
- Skewness and Kurtosis
- z-scores
- Statistical Normality tests:
  - Shapiro-Wilk test
  - Kolmogorov Smirnov Test
  - Anderson-Darling test

Visual Methods

Descriptive Statistics based Methods

Statistical tests for Normality

Brief Overview of Cleaned Dataset

The dataset contains 335791 rows and 24 columns. The columns in the dataset have different data types like int, float and object. A snippet of the dataset is shown in the picture below.

COLUMNS

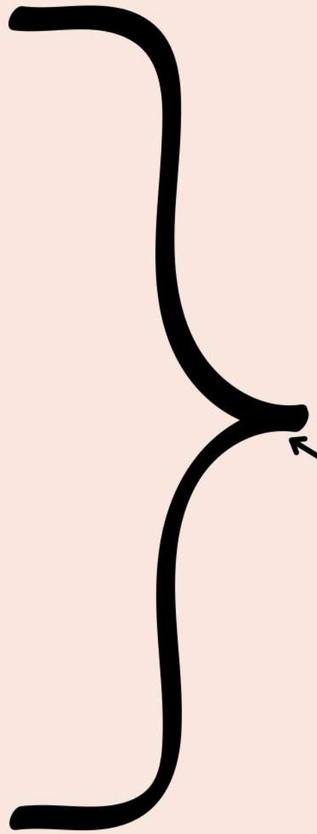
- loan_amnt
- term
- interest_rate
- instalment
- grade/rank
- sub_grade
- emp_title
- emp_length_years
- home_ownership
- annual_income
- verification_status
- issue_date
- loan_status
- purpose
- title
- dti
- earliest_cr_line
- open_acc
- pub_rec
- revol_bal
- revol_util
- total_acc
- initial_list_status
- mort_acc
- pub_rec_bankruptcies

	loan_amnt	term_months	interest_rate	installment	grade/rank	sub_grade	emp_title	emp_length_years	home_ownership	annual_income	...	open_acc	pub_rec	revol_bal
0	10000.0	36	11.44	329.48	B	B4	Marketing	10+	RENT	117000.0	...	16.0	0.0	36369.0
1	8000.0	36	11.99	265.68	B	B5	Credit analyst	1 to 10	MORTGAGE	65000.0	...	17.0	0.0	20131.0
2	15600.0	36	10.49	506.97	B	B3	Statistician	< 1	RENT	43057.0	...	13.0	0.0	11987.0
4	24375.0	60	17.27	609.33	C	C5	Destiny Management Inc.	1 to 10	MORTGAGE	55000.0	...	13.0	0.0	24584.0
5	20000.0	36	13.33	677.07	C	C3	HR Specialist	10+	MORTGAGE	86788.0	...	8.0	0.0	25757.0

NUMERIC AND CATEGORICAL COLUMNS

Numeric Columns:

- loan_amnt
- term_months
- interest_rate
- installment
- annual_income
- dti
- earliest_cr_line
- open_acc
- pub_rec
- revol_bal
- revol_util
- total_acc
- mort_acc
- pub_rec_bankruptcies

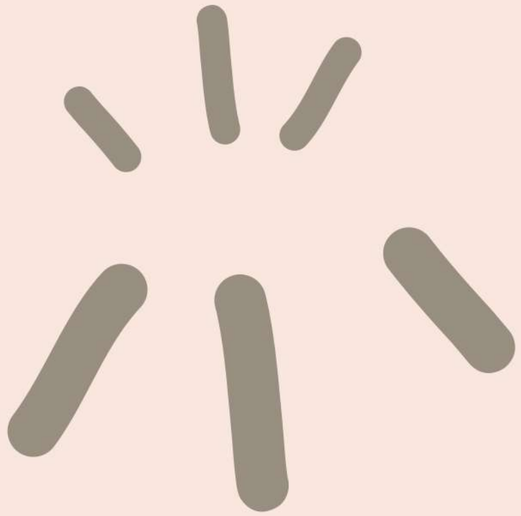


Categorical Columns:

- grade/rank
- sub_grade
- emp_title
- emp_length_years
- home_ownership
- verification_status
- loan_status
- purpose
- title
- initial_list_status

We'll be focussing on these!

WHY DO WE CARE ABOUT NORMALITY?



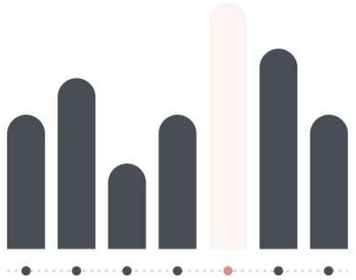
- Before we dive into methods of checking normality, let's see why "your data" being "normal" is good news:
    a. it has the infamous bell-shaped curve which is symmetric around the mean and makes it an appealing choice for linear regression models.
    b. and then because of the Central Limit Theorem, for a large number of samples, the normal distribution can approximate other known distributions.
    c. and also the mean, mode and median are equal (or approximately equal depending on how close your data is to a standard normal distribution)

The properties mentioned make normal distributions more analytically appealing and solvable!

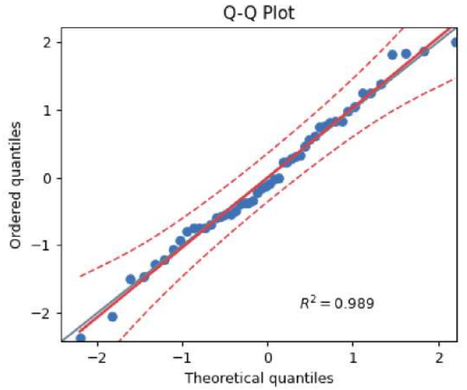
TESTS FOR NORMALITY

Looking at plots

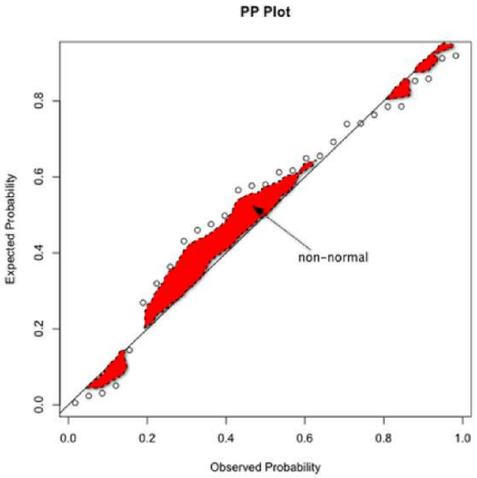
Histograms



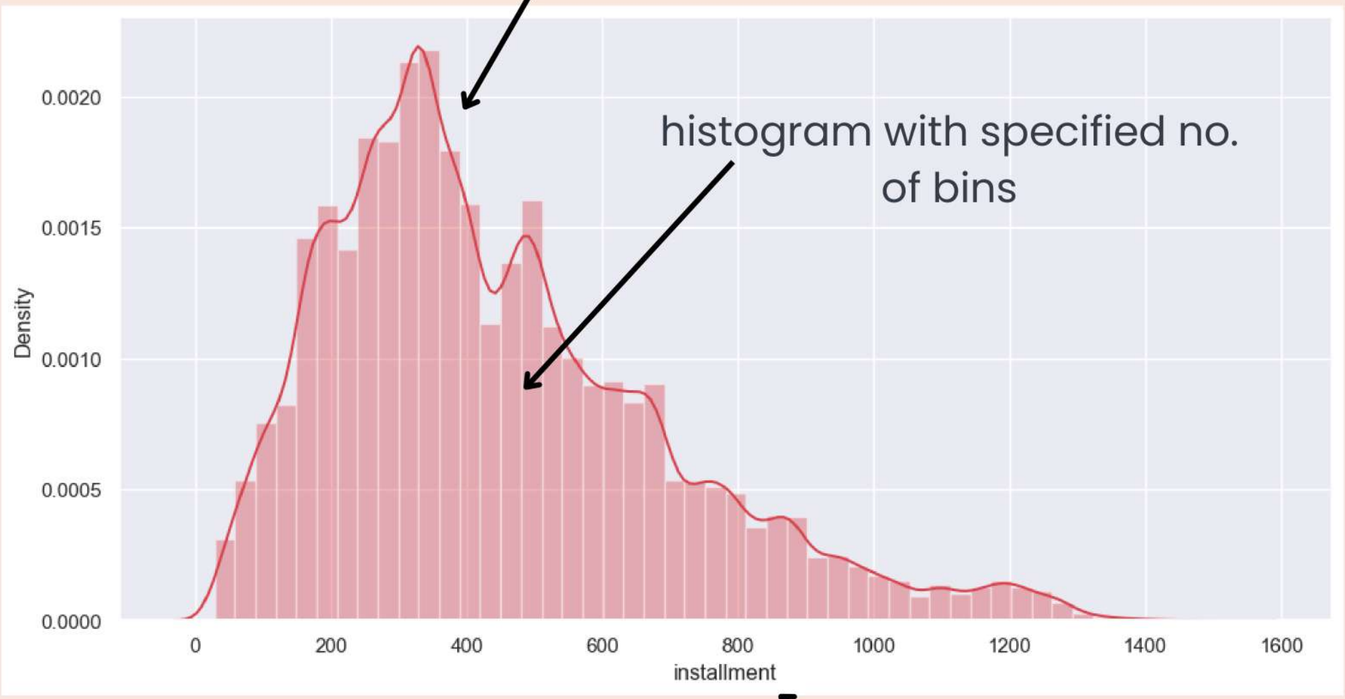
Q-Q plots



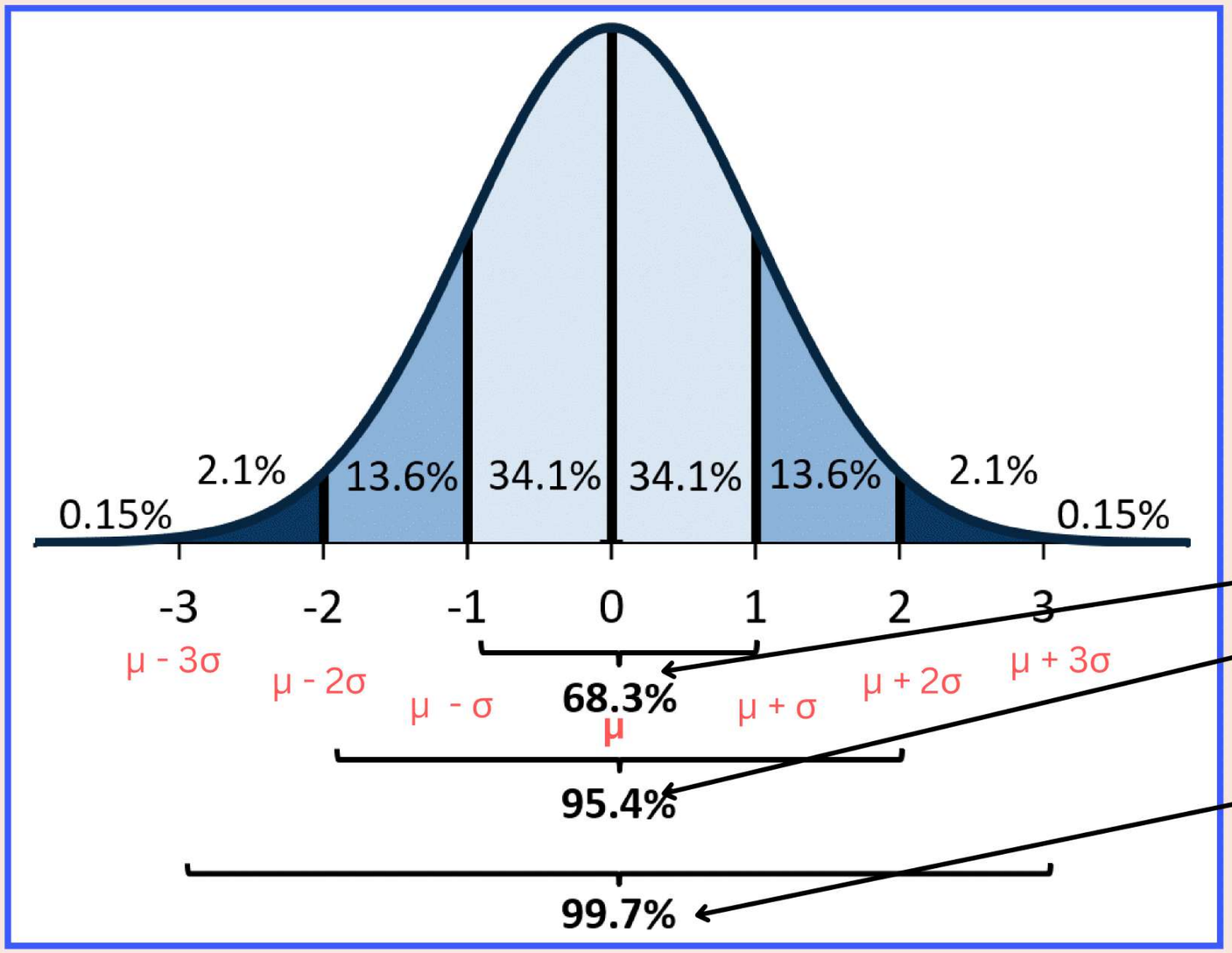
P-P plots



Create a histogram



- If the histogram is roughly "bell-shaped", then the data is assumed to be normally distributed
- We can get a better idea of the distribution of values by plotting a density curve over the histogram.
- For the values to be normally distributed, they should approximately follow the empirical rule for the family of normal distributions (see next slide)

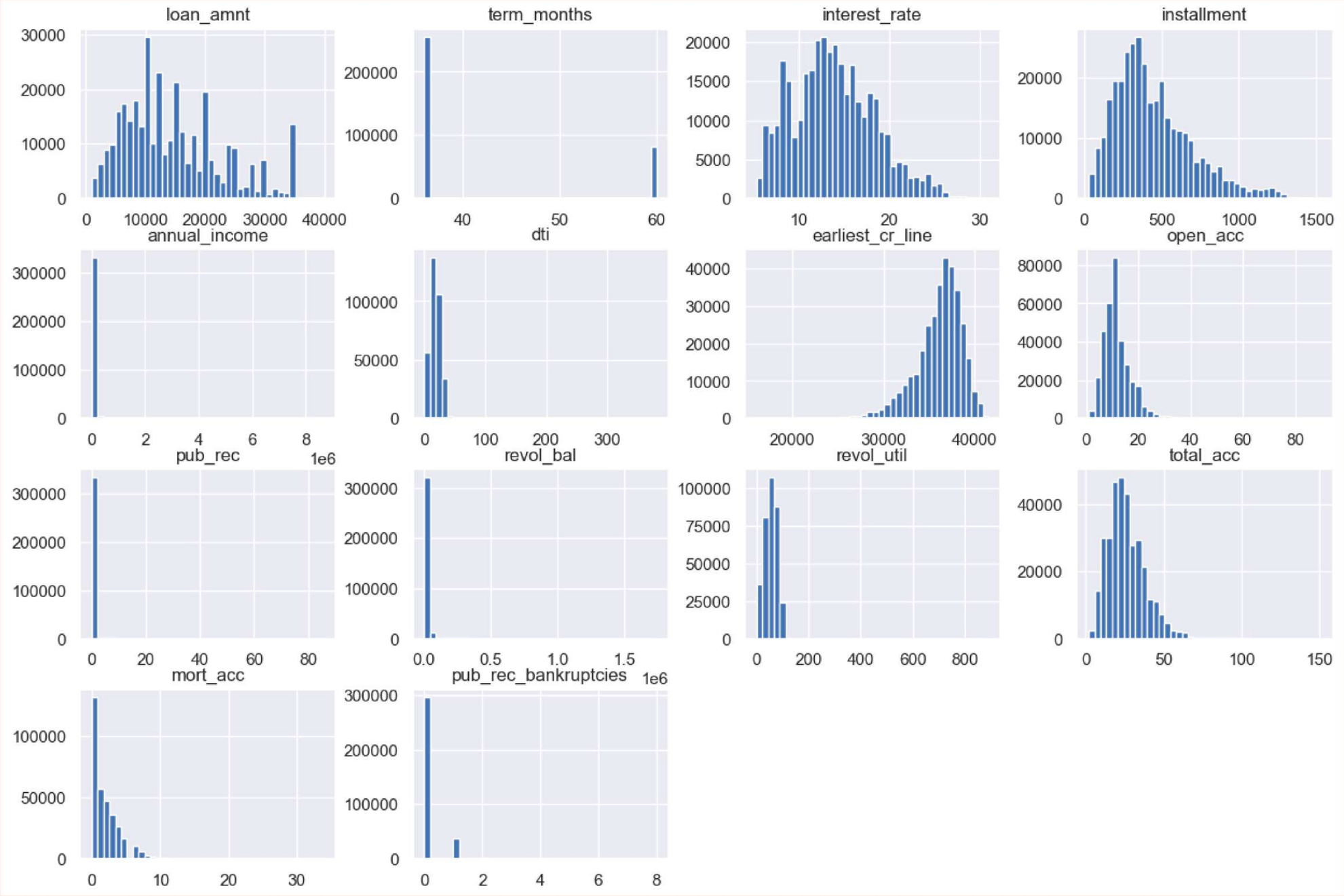


- The empirical rule (also called the "68-95-99.7 rule") is a guideline for how data is distributed in a normal distribution.

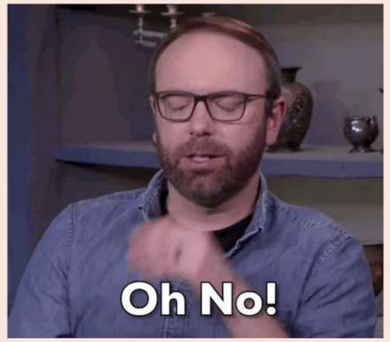
- The rule states that (approximately): - 68% of the data points will fall within one standard deviation of the mean. - 95% of the data points will fall within two standard deviations of the mean. - 99.7% of the data points will fall within three standard deviations of the mean

Well approximately...

Histograms of the (numerical) columns in our data

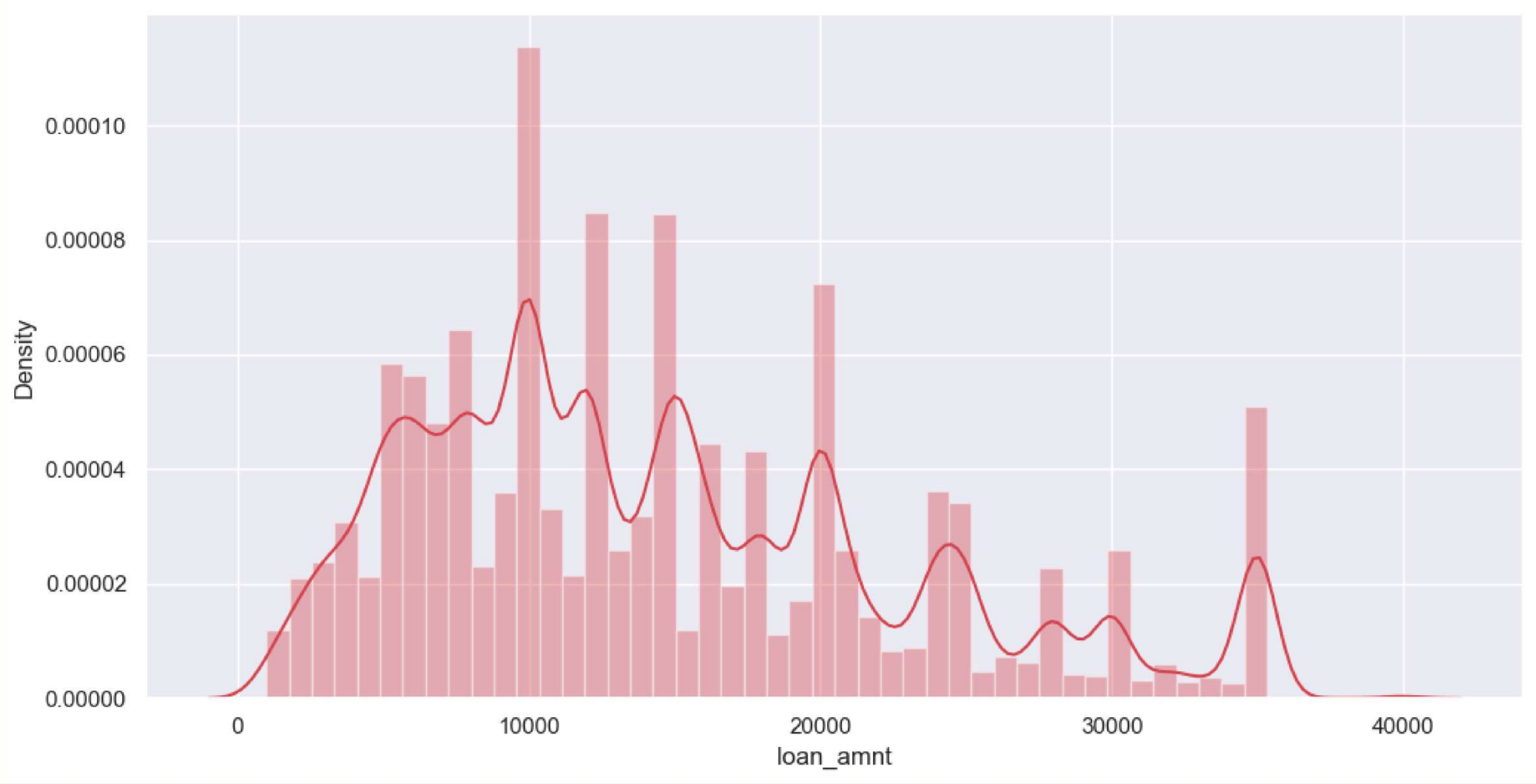


Most of them seem to be NOT normally distributed at first sight

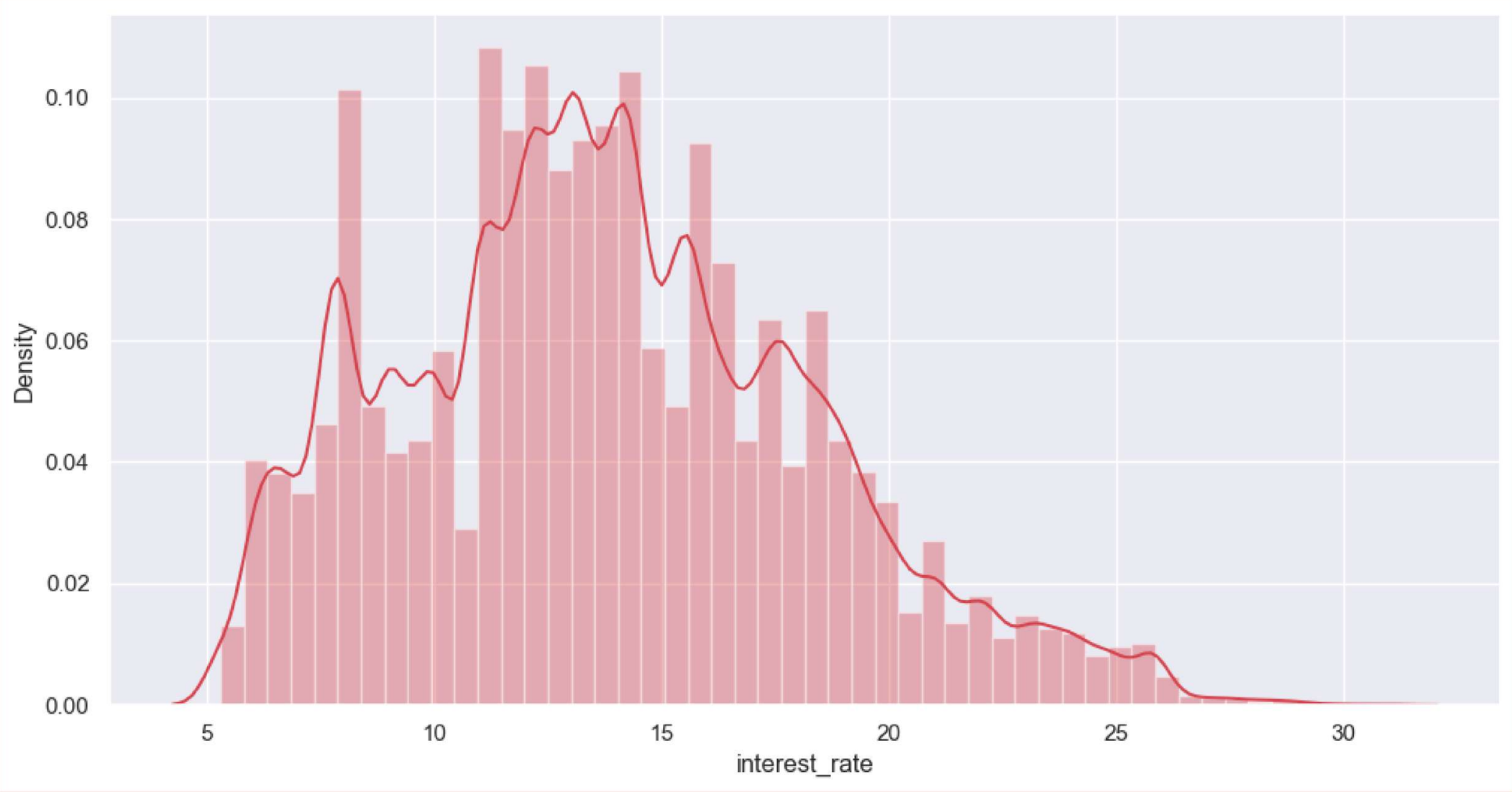


Let's zoom in a bit.
Shall we?

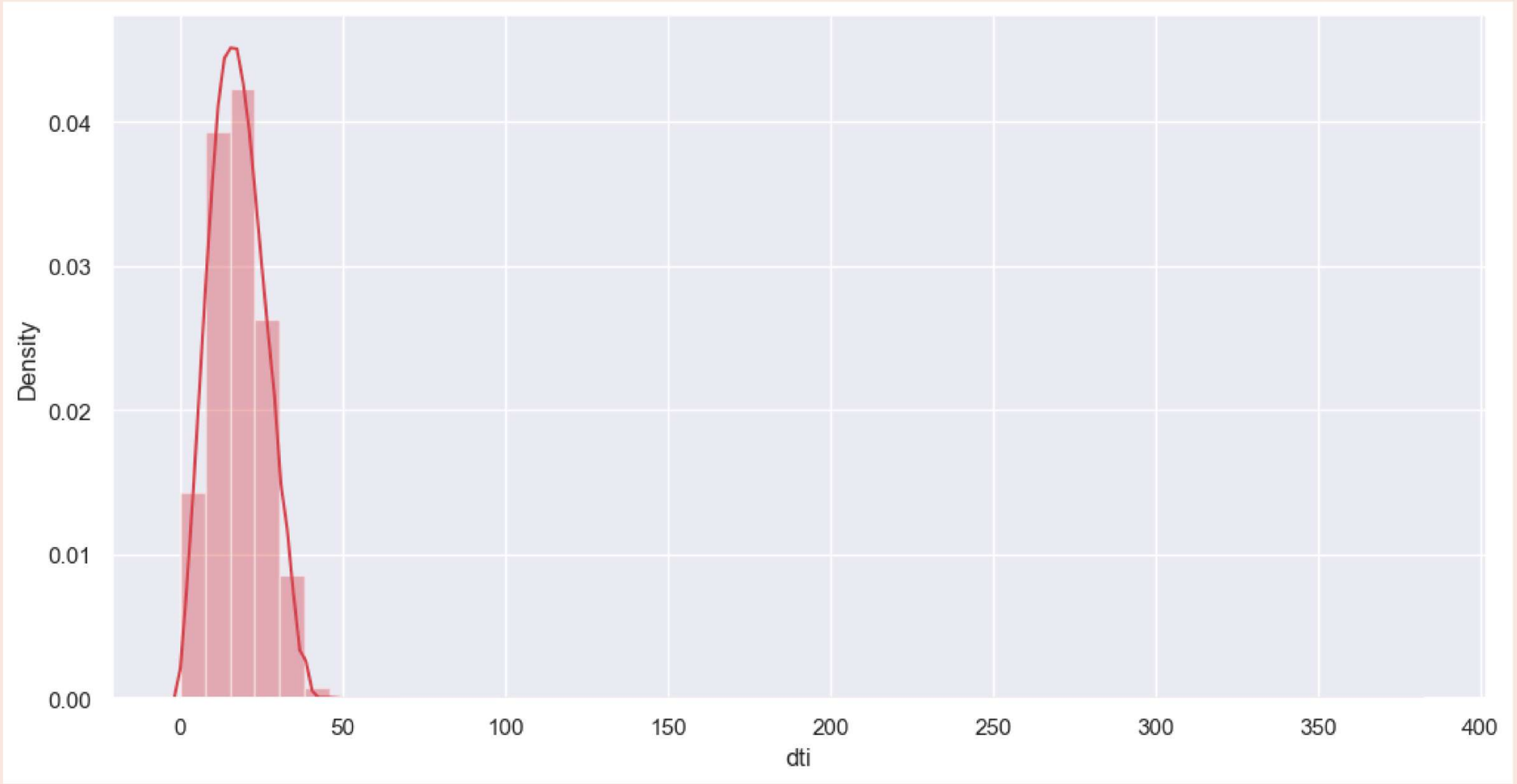
loan_amnt


0.00010
0.00008
0.00006
0.00004
0.00002
0.00000
Density
0
10000
20000
30000
40000
loan_amnt

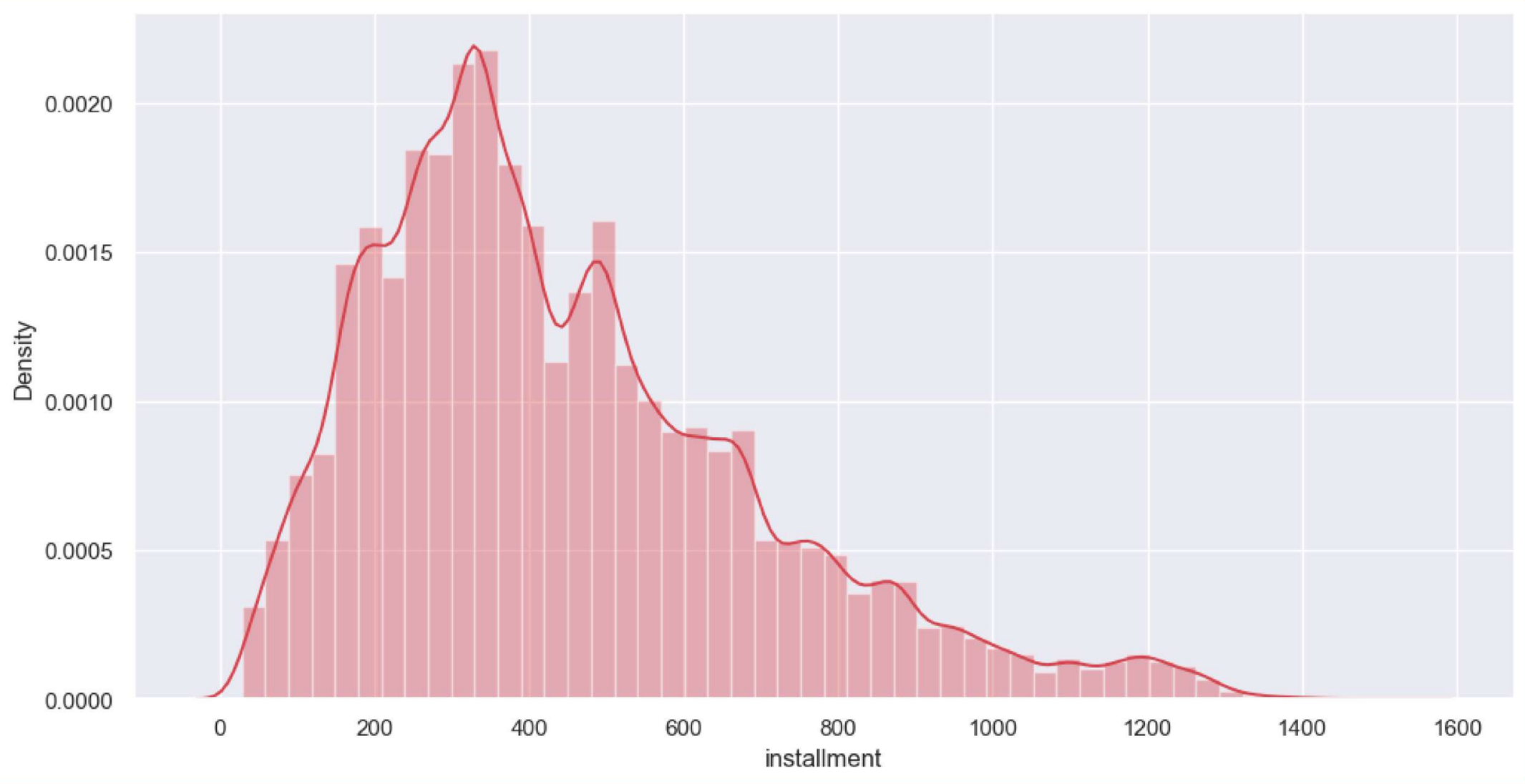
interest_rate



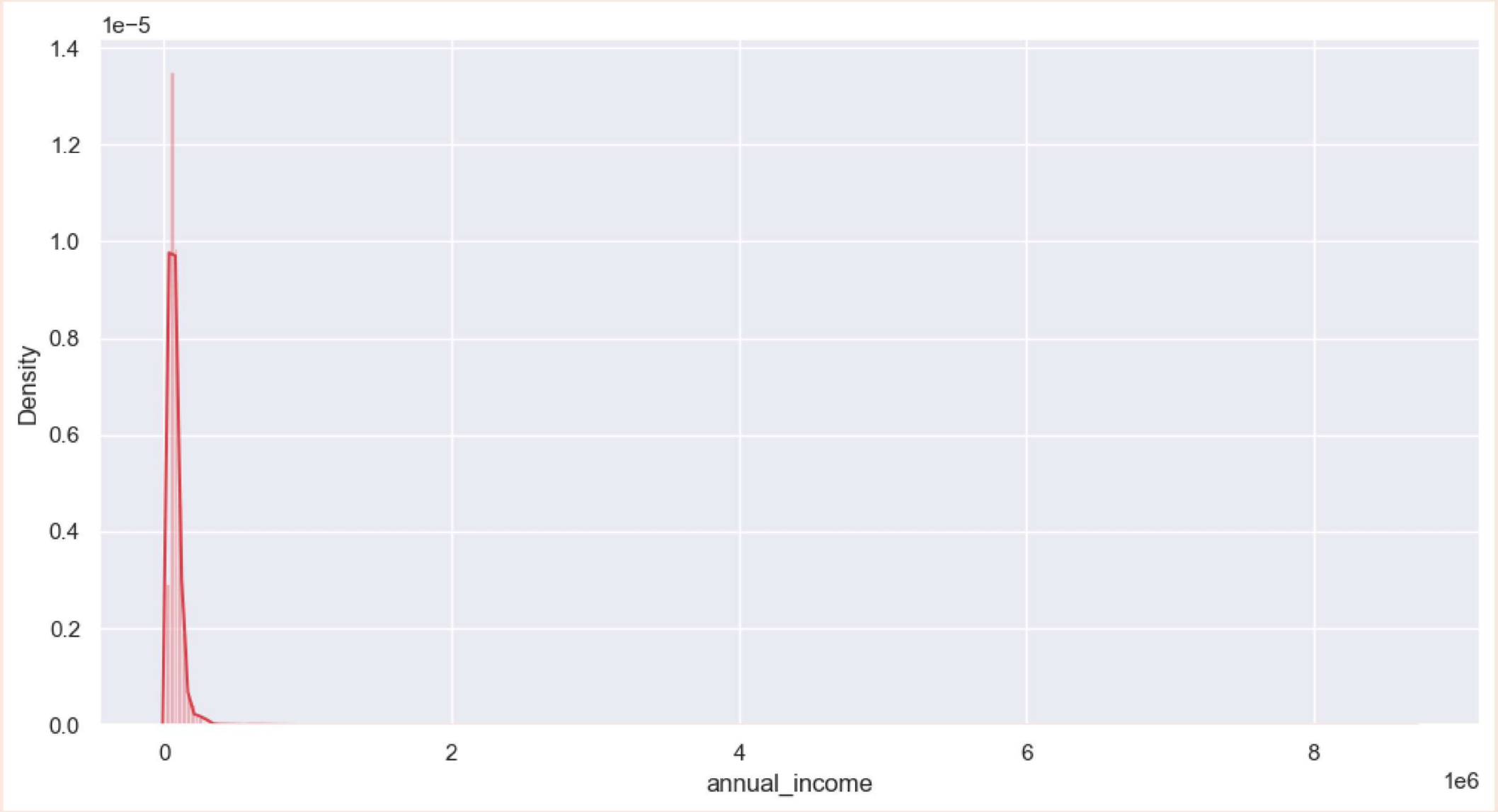
dti



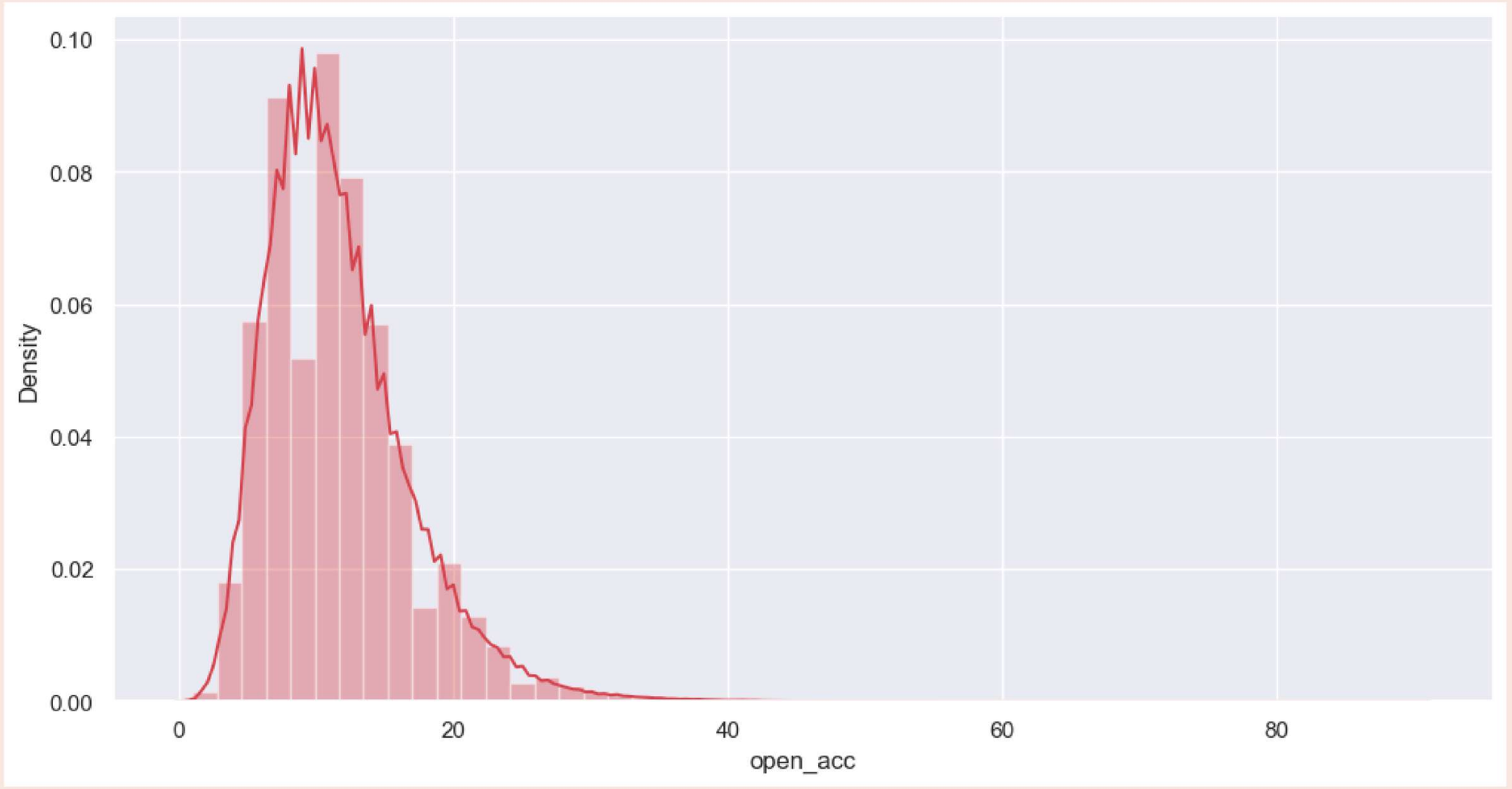
installment



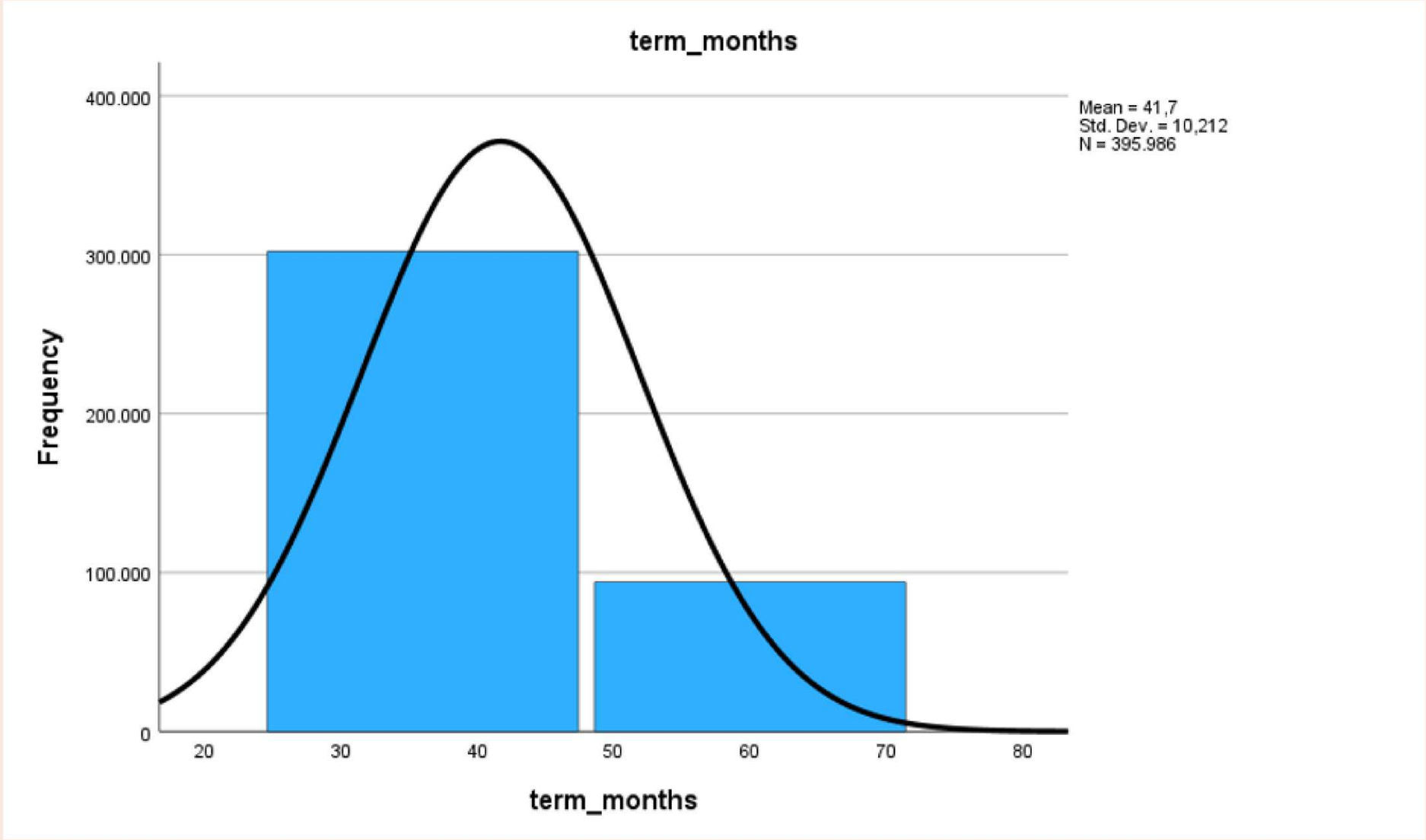
annual_income



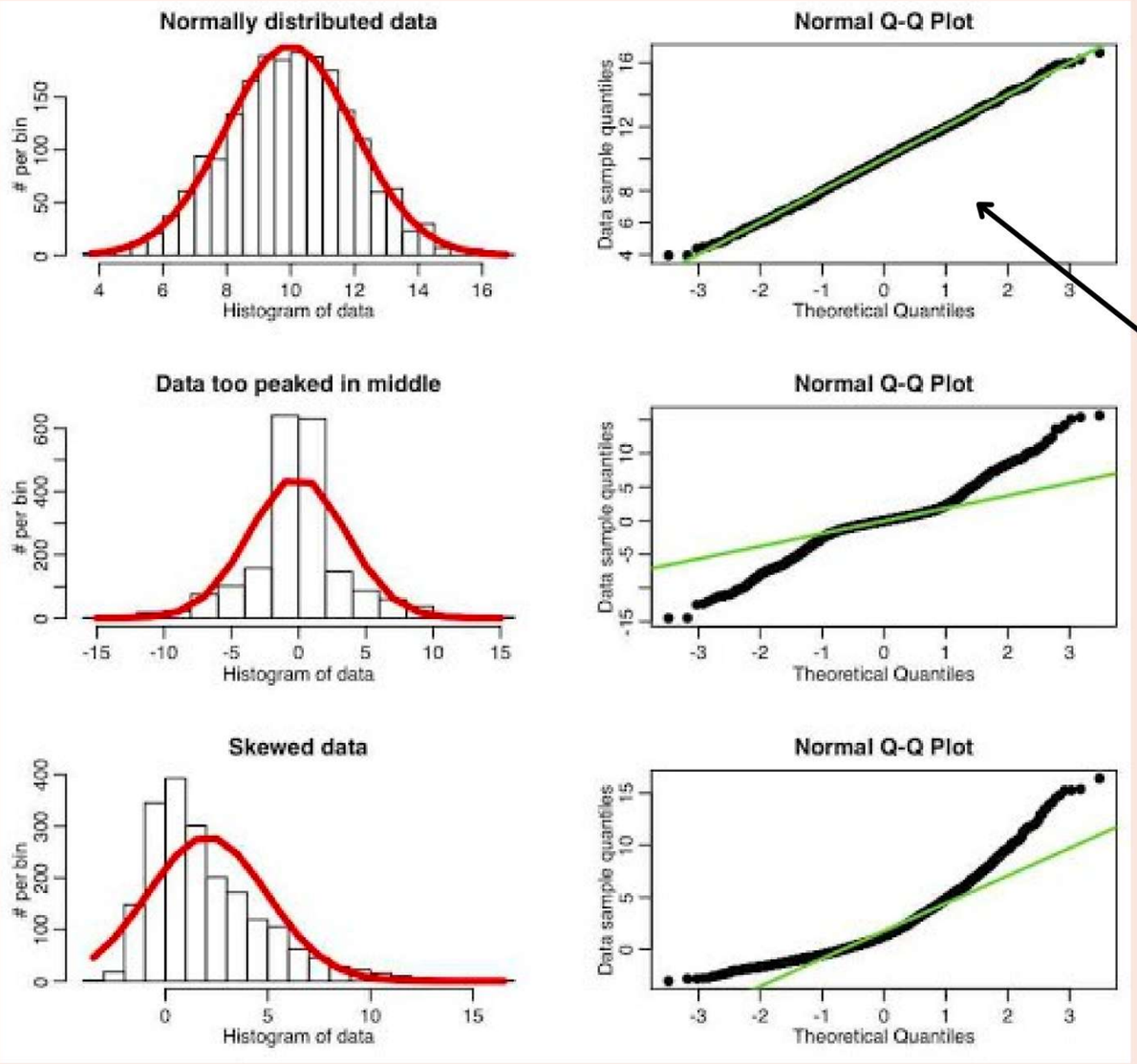
open_acc



term_months


term_months
Mean = 41,7
Std. Dev. = 10,212
N = 395.986
Frequency
400.000
300.000
200.000
100.000
0
20
30
40
50
60
70
80
term_months

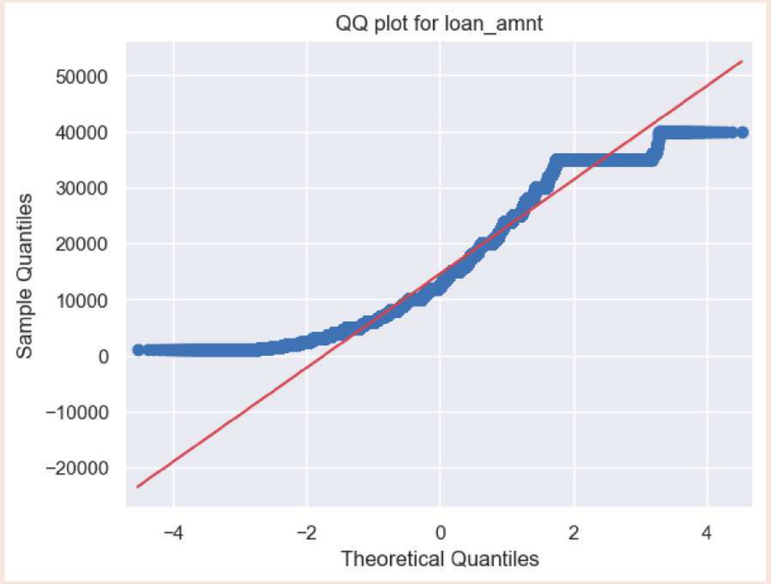
Q-Q plots

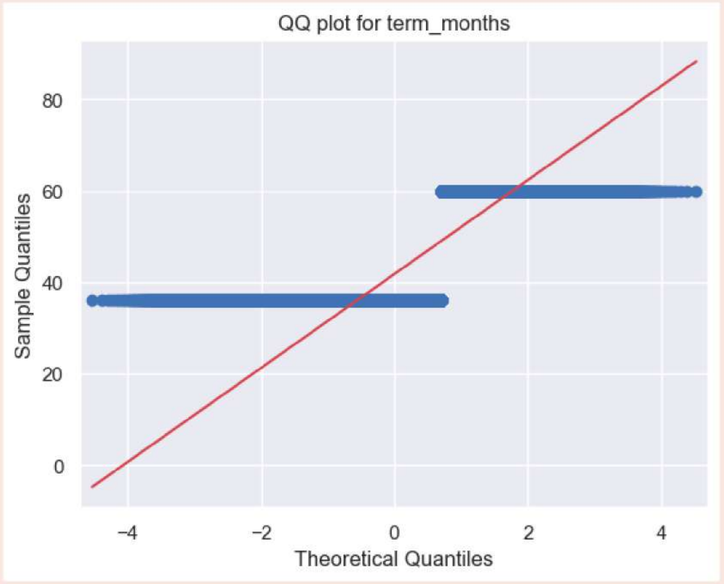


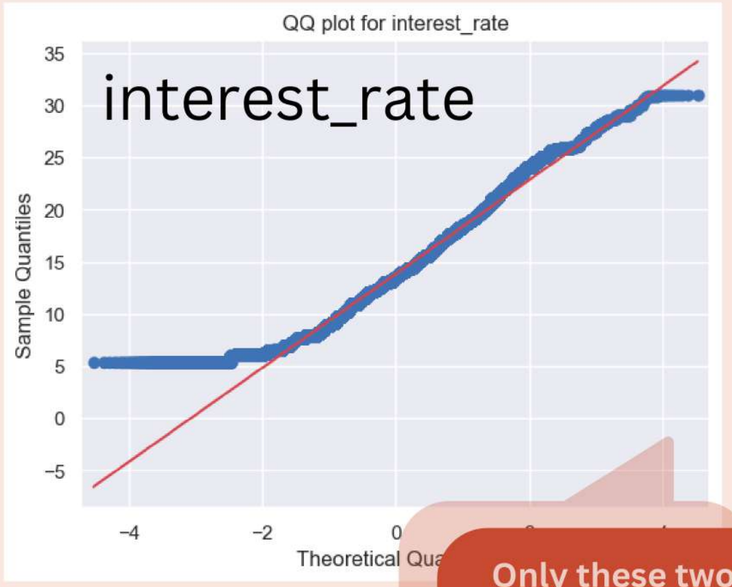
Good Q-Q plot

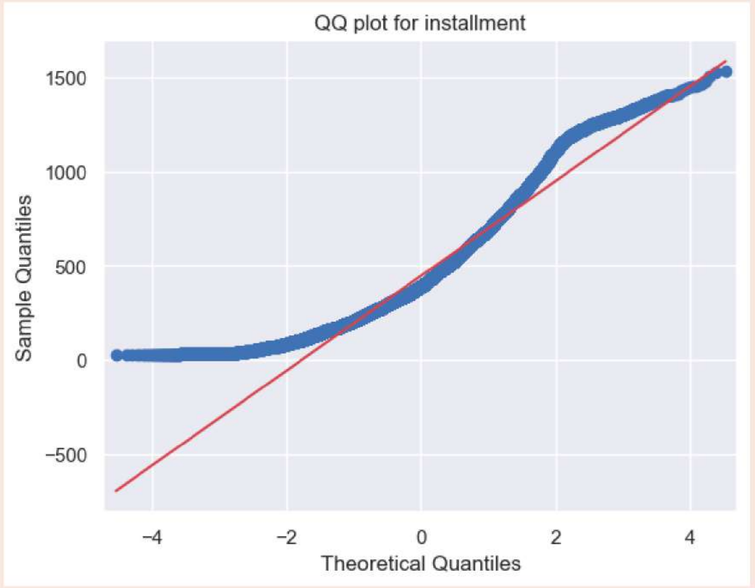
If the points in the plot roughly fall along a straight diagonal line, then the data is assumed to be normally distributed

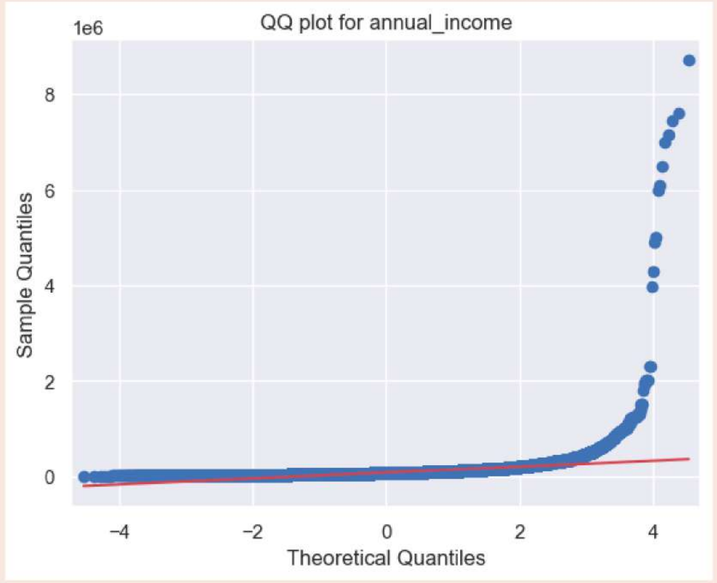
Q-Q plots of our dataset columns

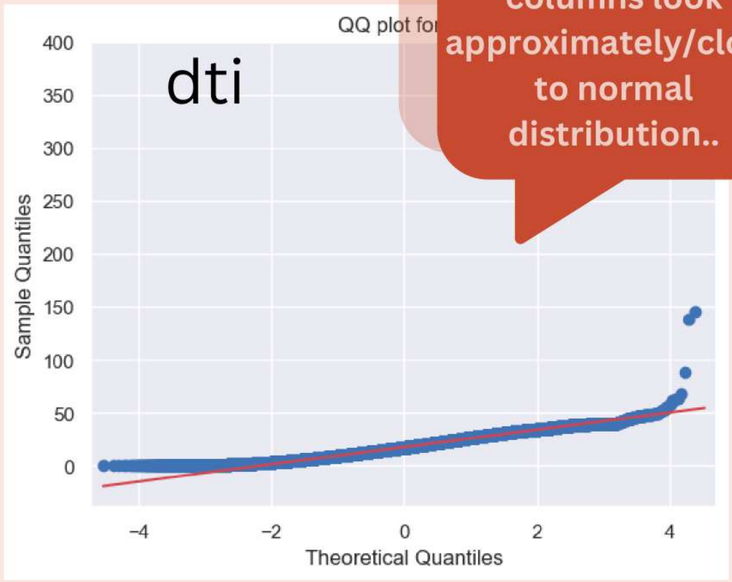




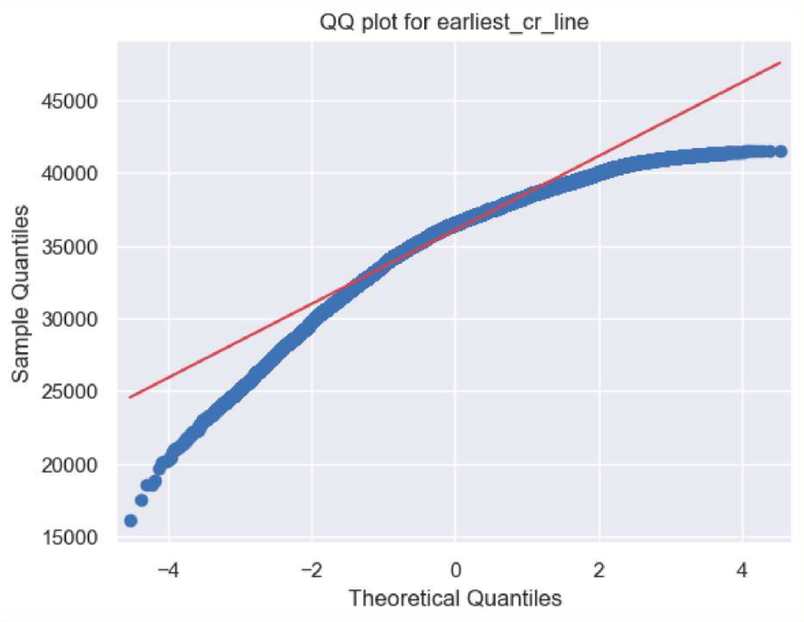


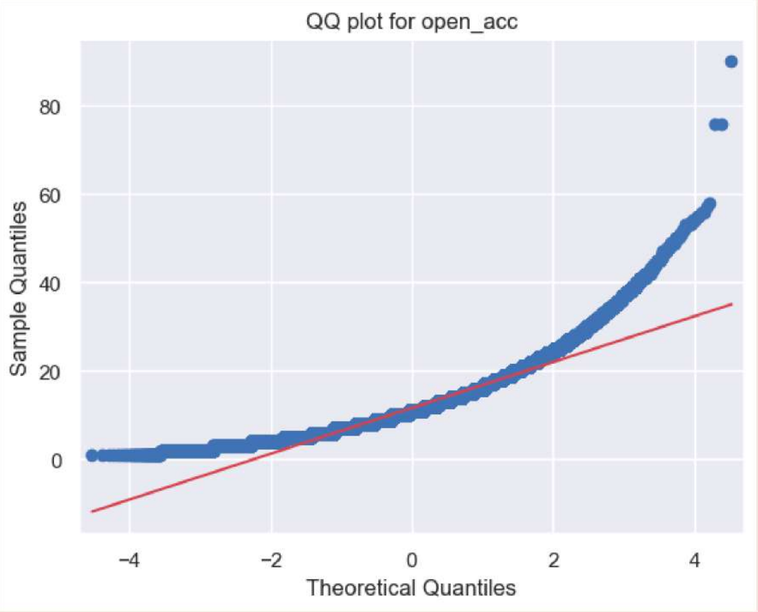


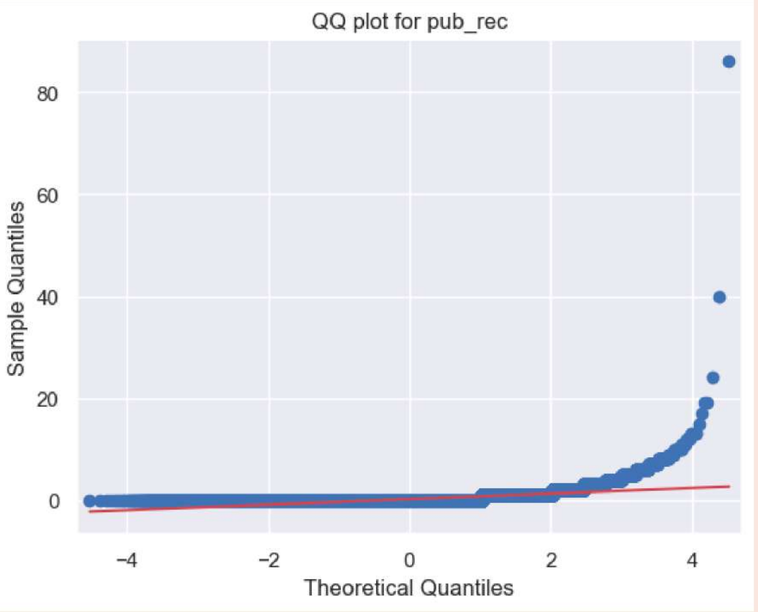


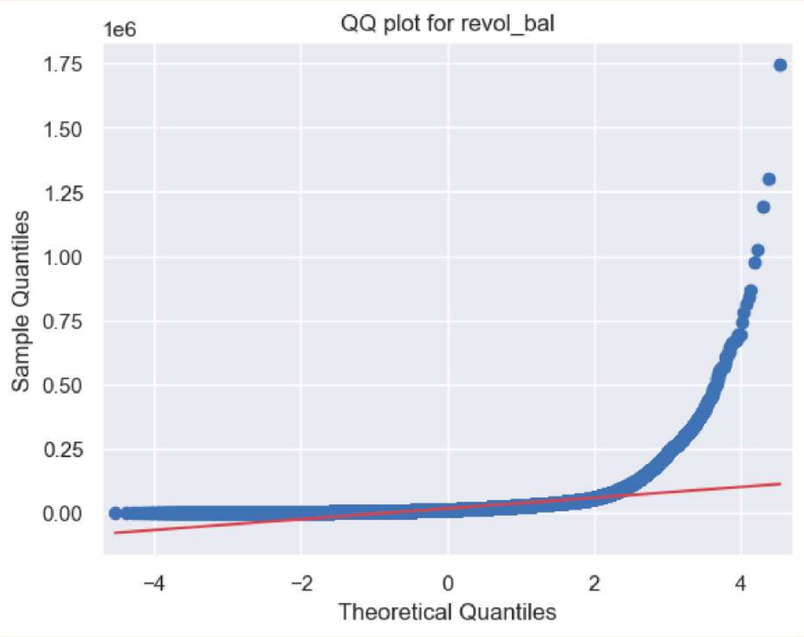


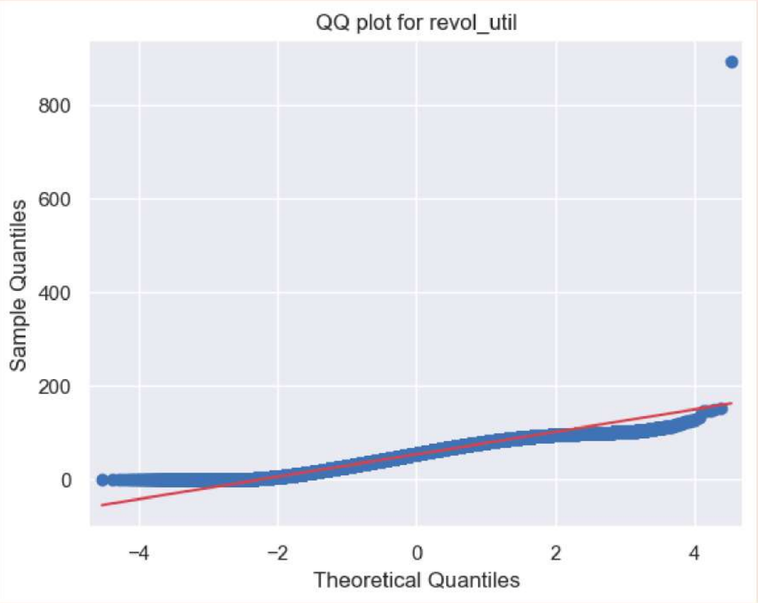
Only these two columns look approximately/close to normal distribution..

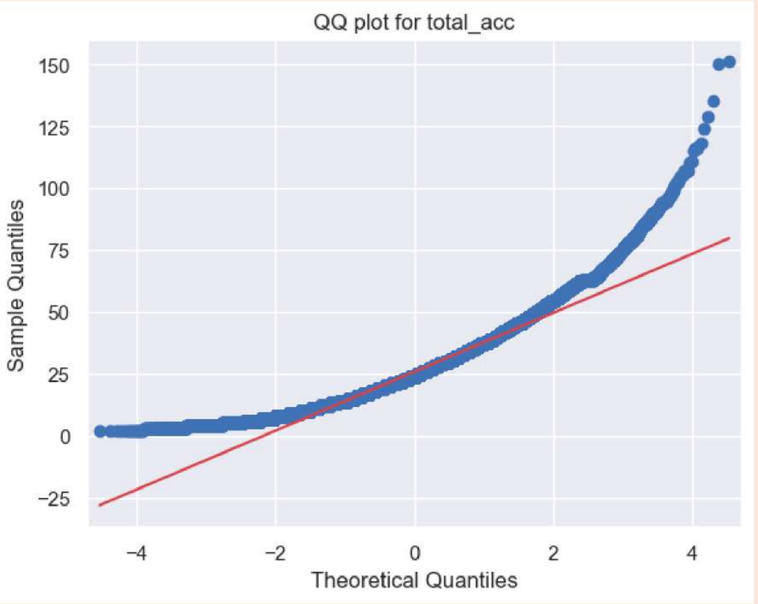

QQ plot for earliest_cr_line
Sample Quantiles
Theoretical Quantiles
45000
40000
35000
30000
25000
20000
15000
−4
−2
0
2
4

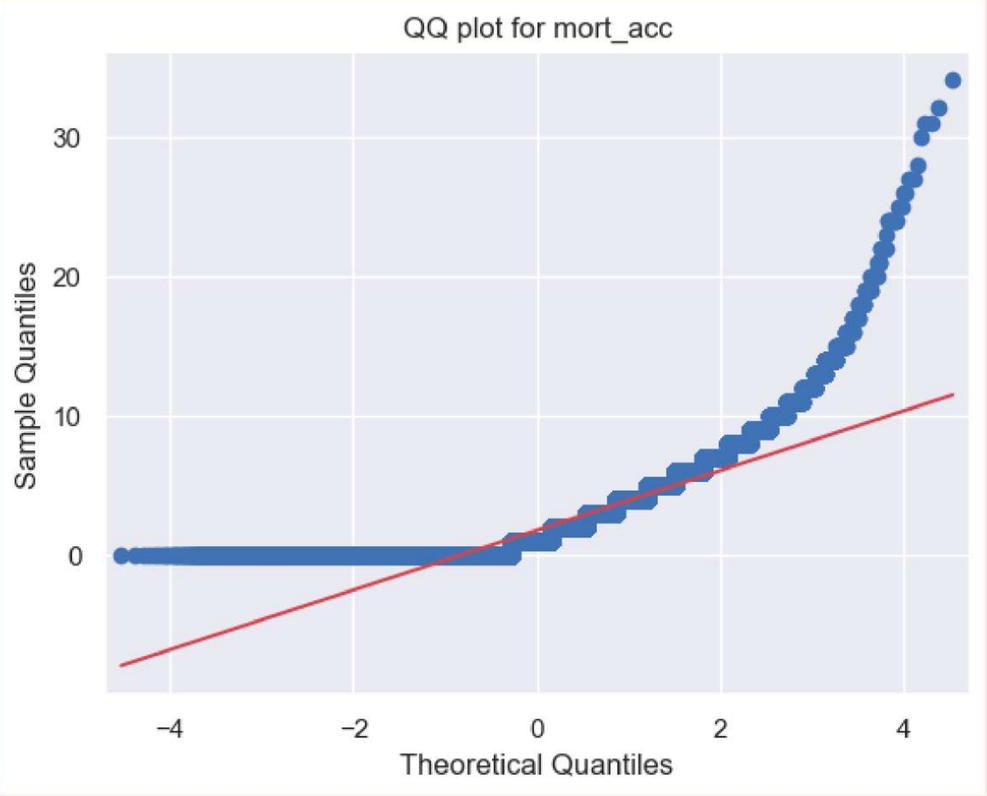

QQ plot for open_acc
Sample Quantiles
Theoretical Quantiles
80
60
40
20
0
−4
−2
0
2
4

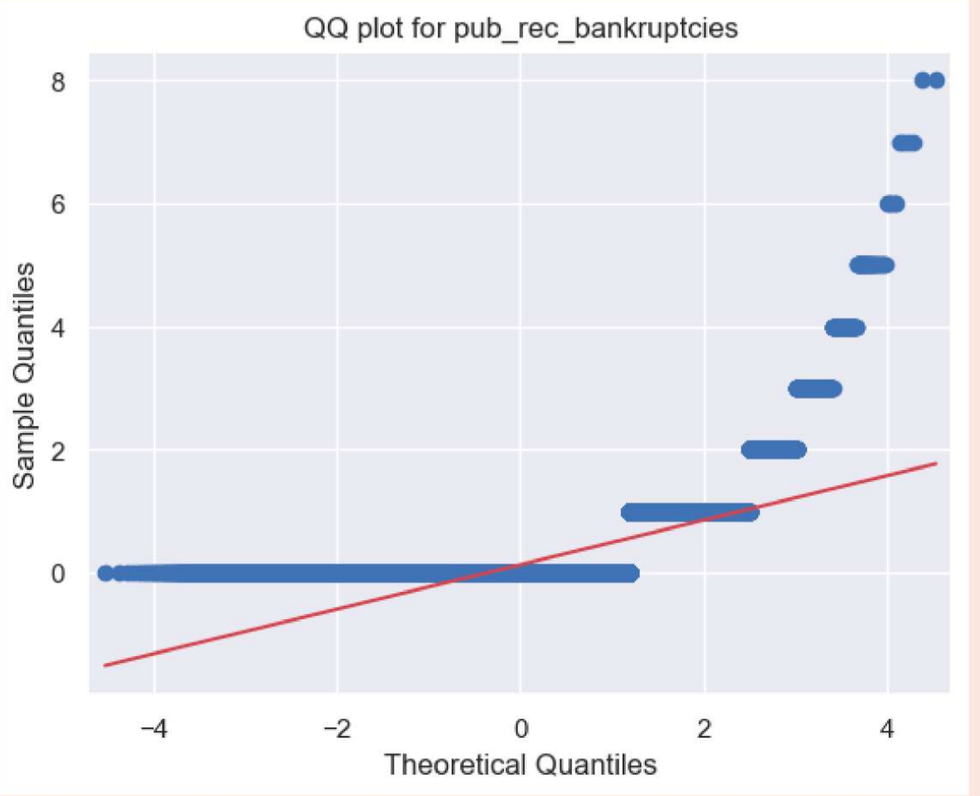

QQ plot for pub_rec
Sample Quantiles
Theoretical Quantiles
80
60
40
20
0
−4
−2
0
2
4


QQ plot for revol_bal
1e6
Sample Quantiles
Theoretical Quantiles
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
−4
−2
0
2
4

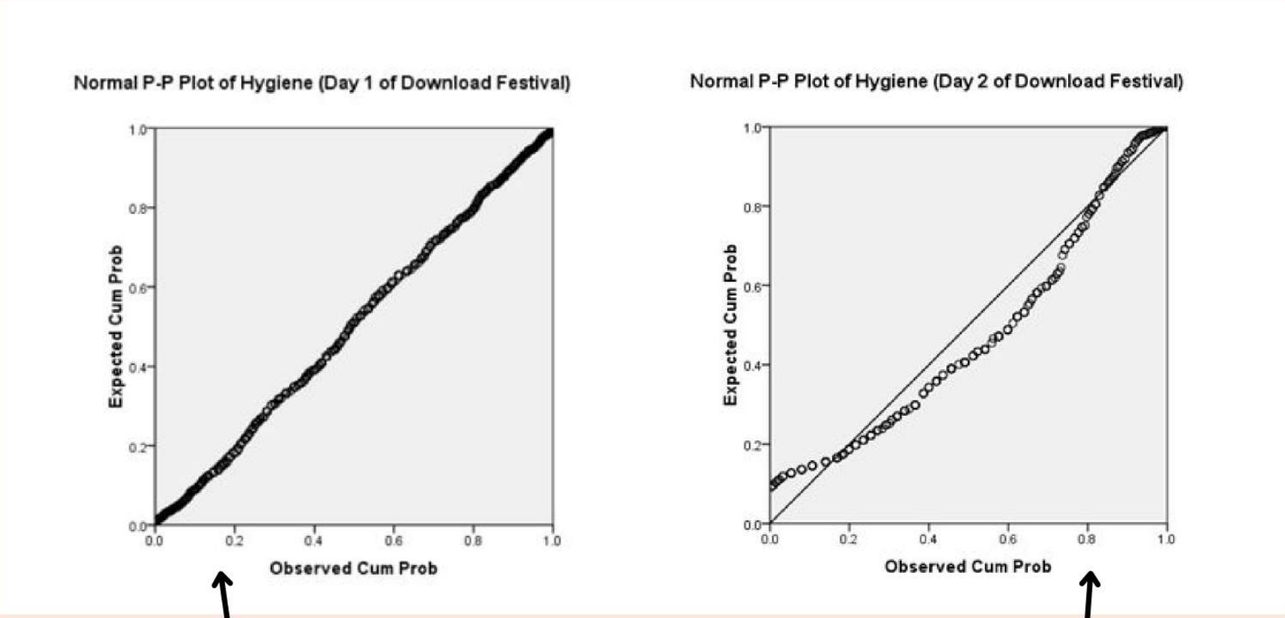

QQ plot for revol_util
Sample Quantiles
Theoretical Quantiles
800
600
400
200
0
−4
−2
0
2
4


QQ plot for total_acc
Sample Quantiles
Theoretical Quantiles
150
125
100
75
50
25
0
−25
−4
−2
0
2
4


QQ plot for mort_acc
Sample Quantiles
Theoretical Quantiles
30
20
10
0
−4
−2
0
2
4


QQ plot for pub_rec_bankruptcies
Sample Quantiles
Theoretical Quantiles
8
6
4
2
0
−4
−2
0
2
4

P-P plots

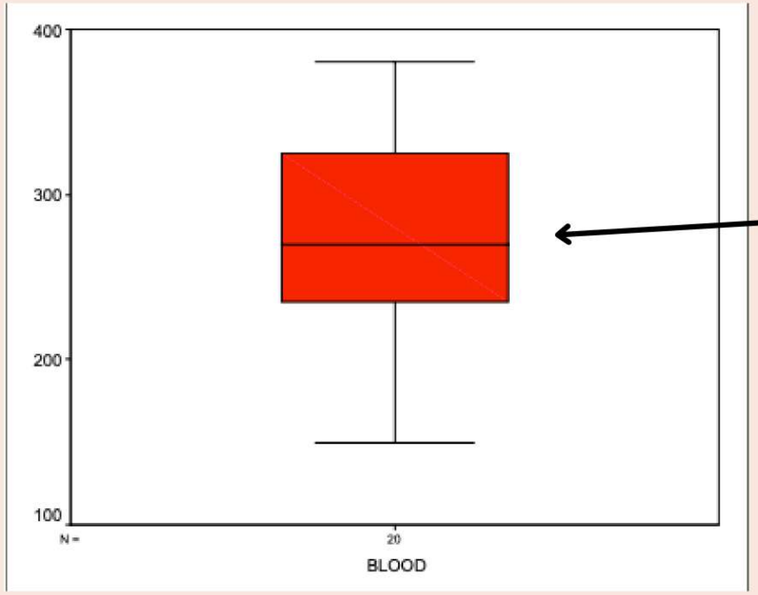


normal

not normal

- The Q-Q is plotting the quantiles–the actual values of X against the theoretical values of X under the normal distribution.
- A P-P plot, one the other hand, plots the corresponding areas under the curve (cumulative distribution function) for those values.
- If the points in the plot roughly fall along a straight diagonal line, then the data is assumed to be normally distributed

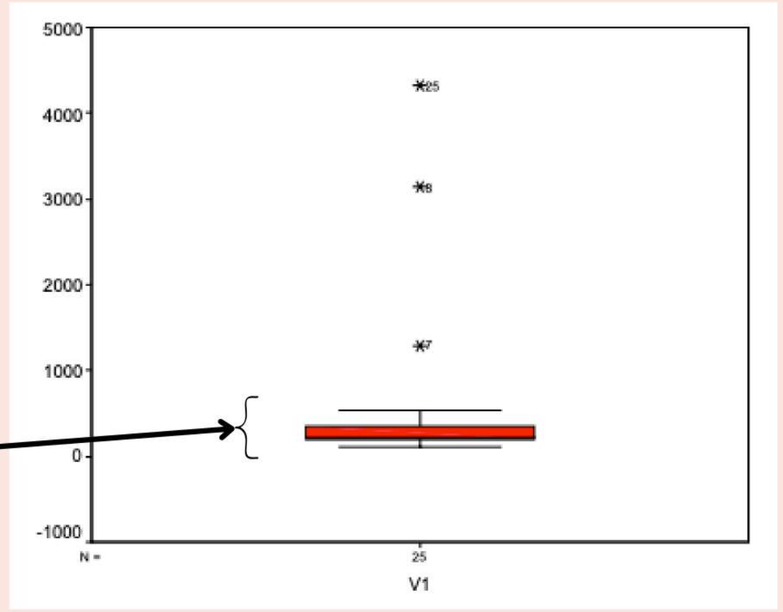
Boxplots



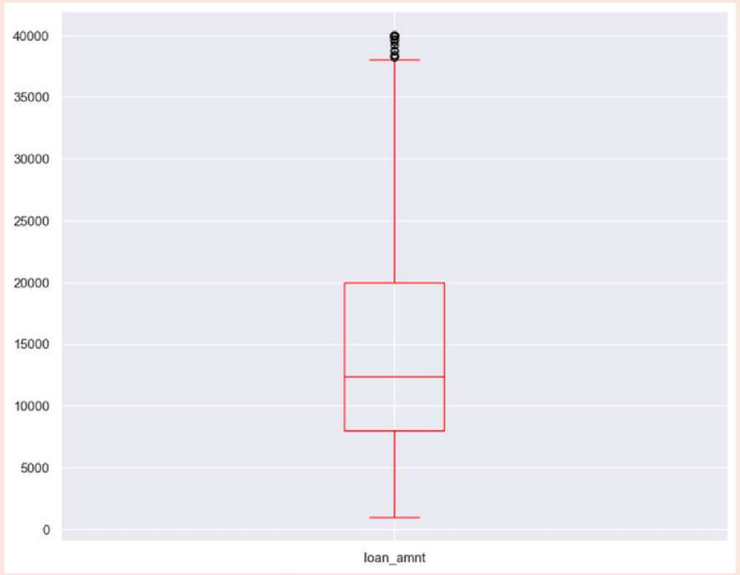
Although the box plot is not perfectly symmetric, there is no clear violation of normality.

- It is hard to detect normality using a box-plot. But, at the very least, look for symmetry. Severe skewness and/or outliers are indications of non-normality.

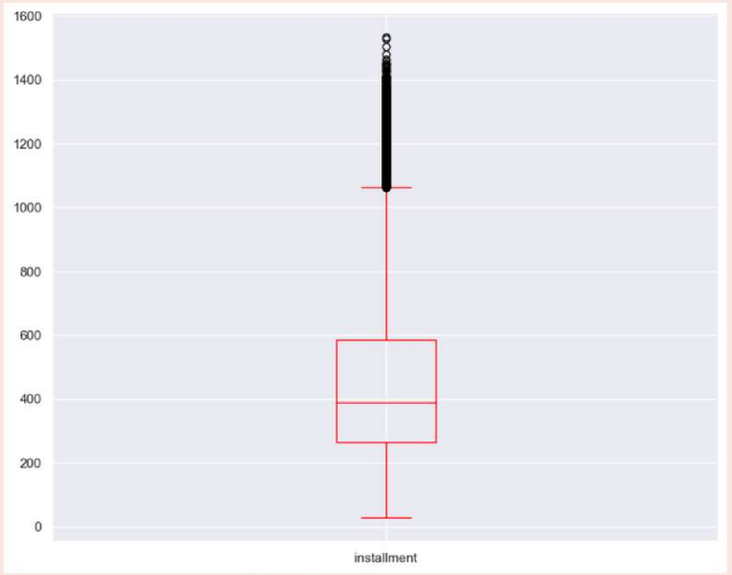
There is a clear indication that the data are right-skewed with some strong outliers. The assumption of normality is clearly violated.



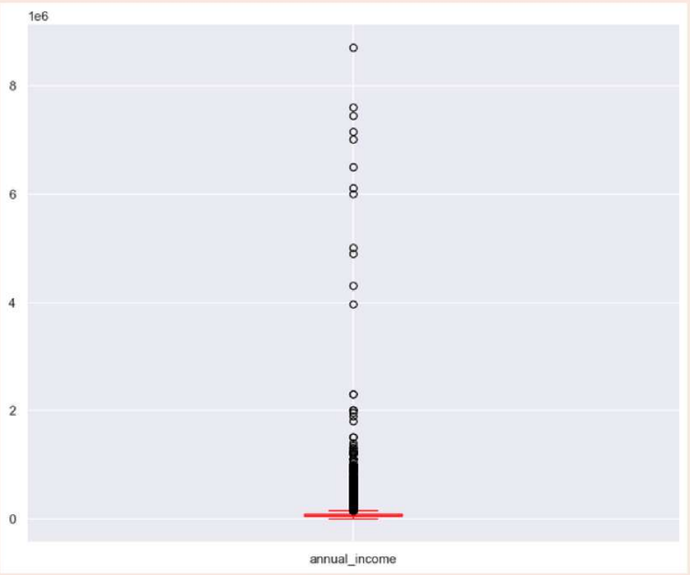
Boxplots of our data



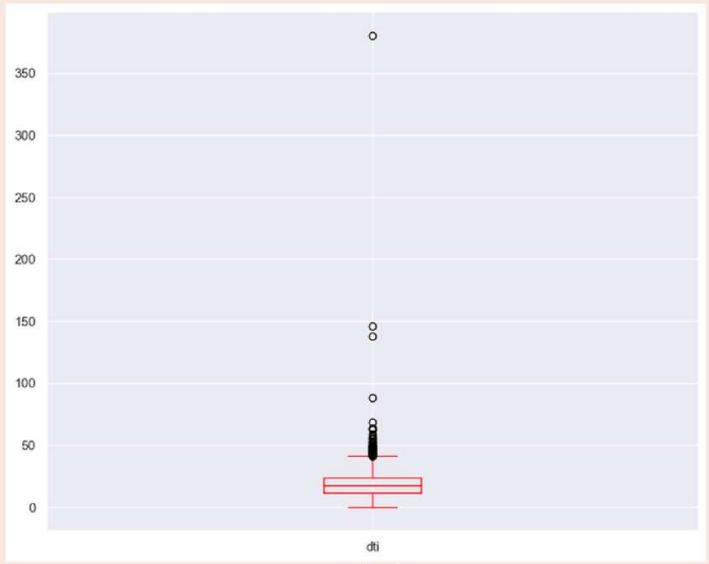
loan_amnt



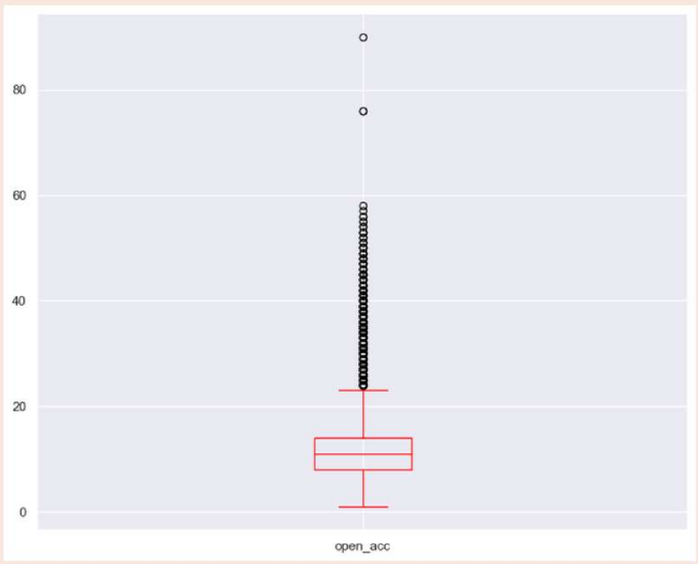
installment



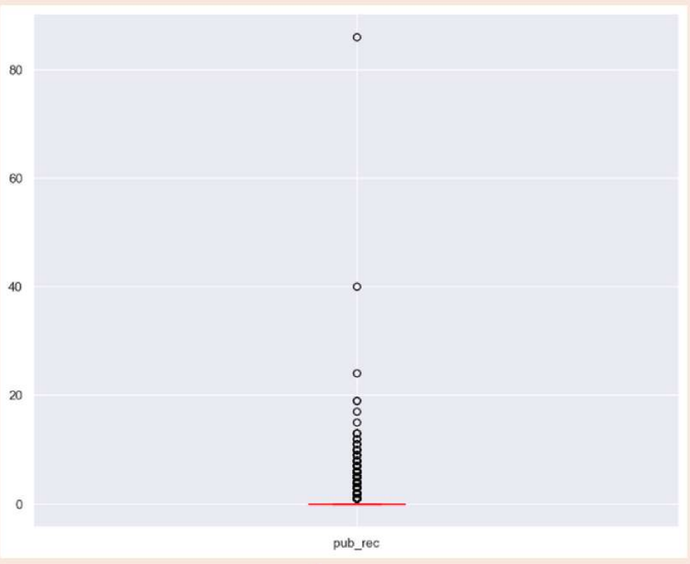
annual_income



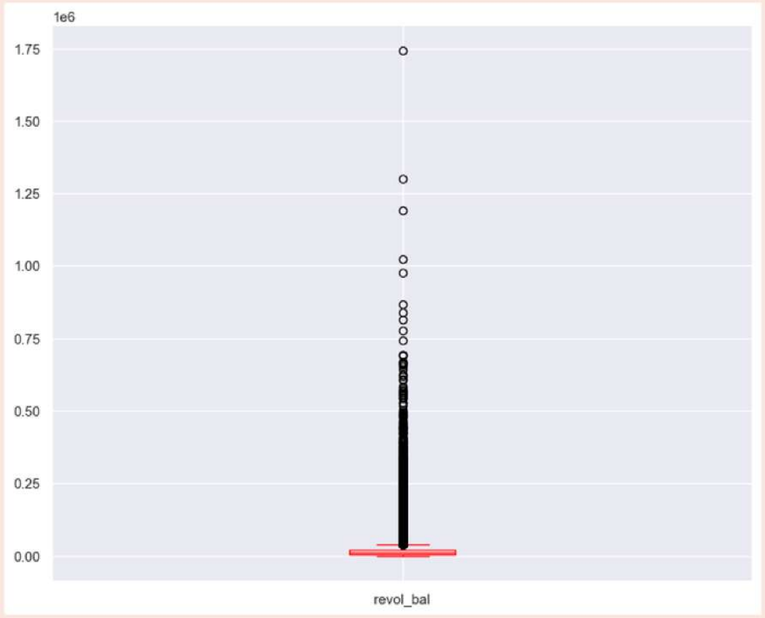
dti



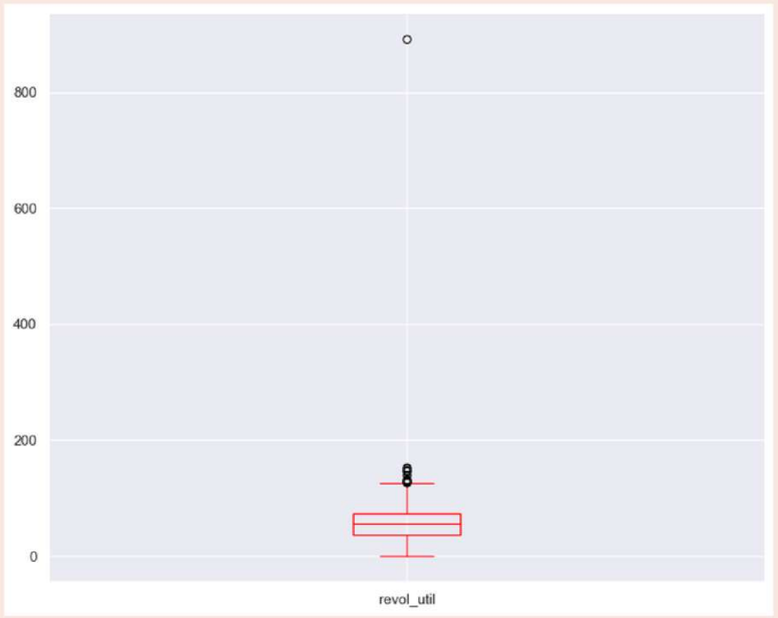
open_acc



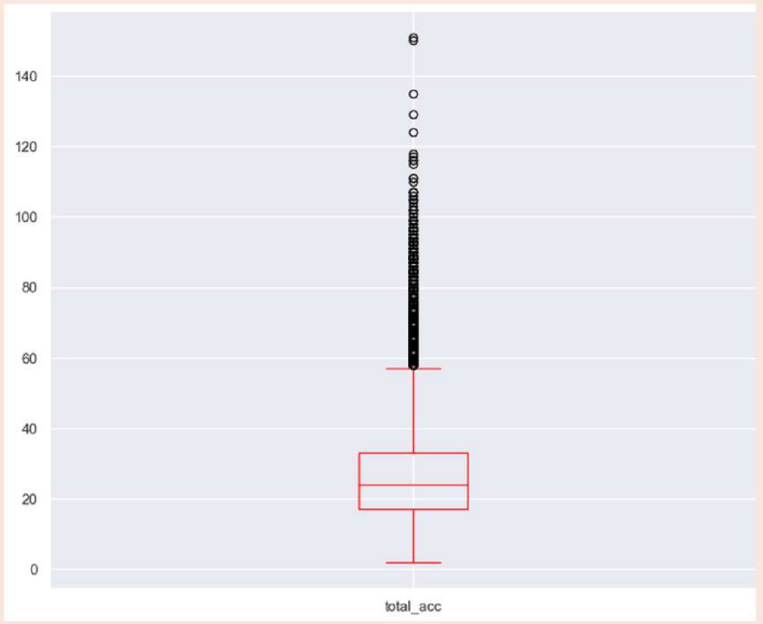
pub_rec



revol_bal

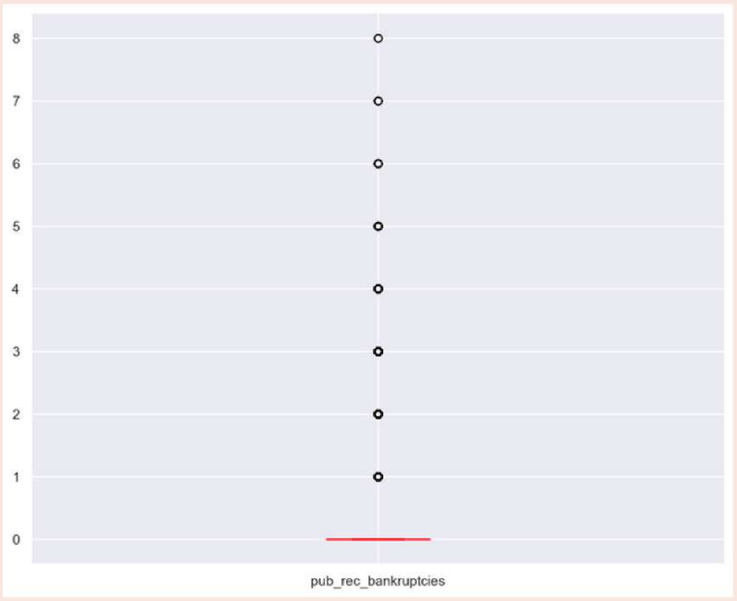


revol_util



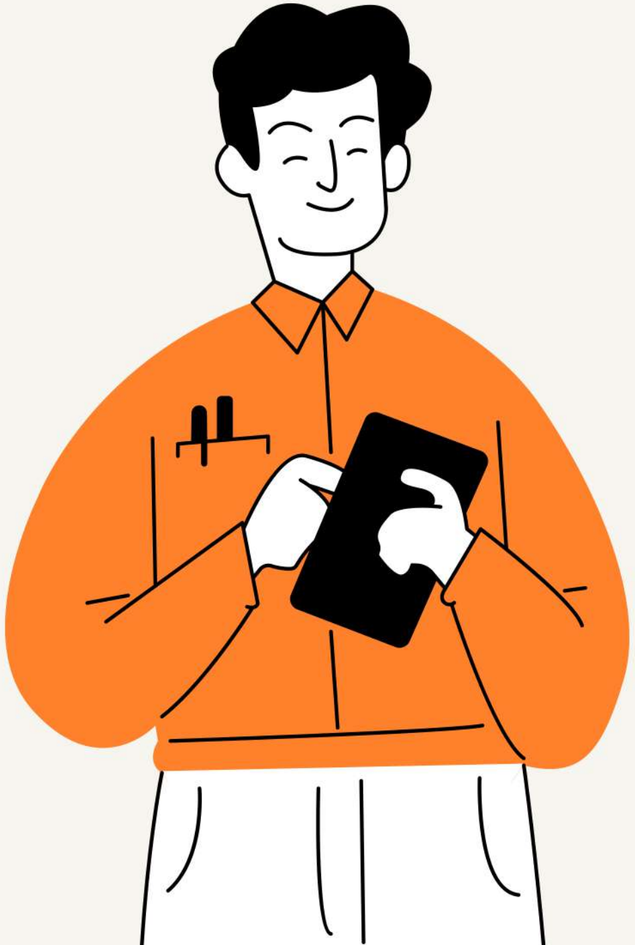
total_acc

mort_acc



pub_rec_bankruptcies

Descriptive Statistics

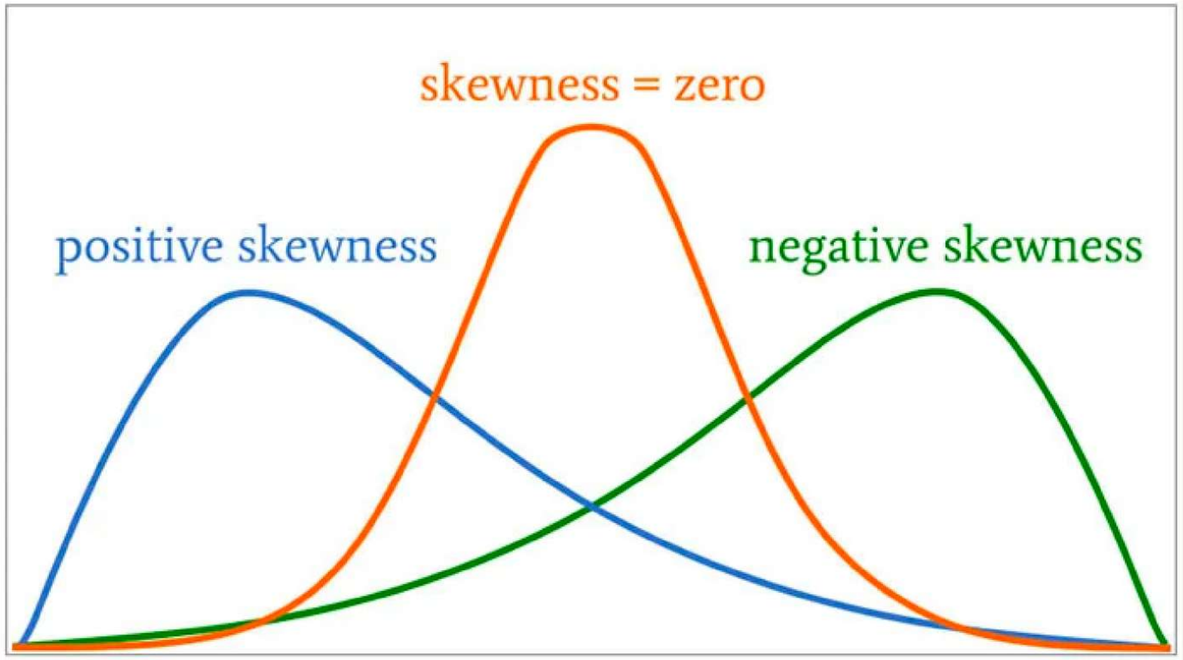


Skewness

Kurtosis

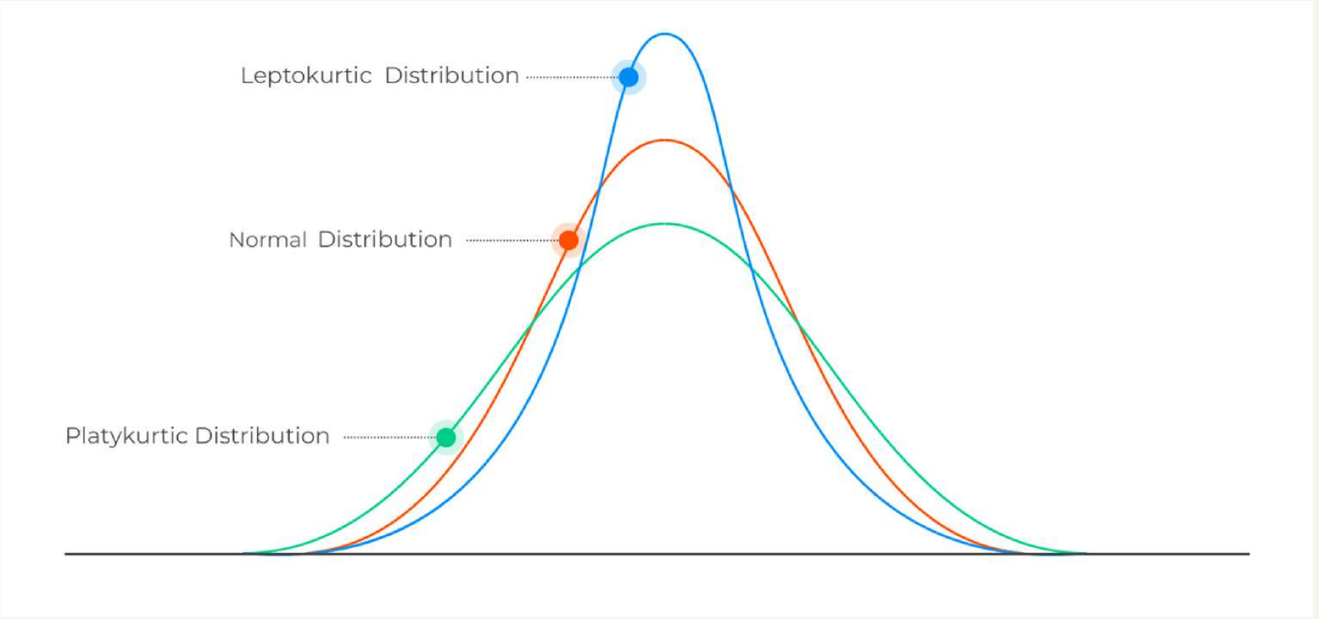
- Skewness is a measure of symmetry, or more precisely, the lack of symmetry.
- A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.

- Kurtosis is a measure of whether the data is heavy-tailed or light-tailed relative to a normal distribution.
- That is, data sets with high kurtosis tend to have heavy tails or outliers. Data sets with low kurtosis tend to have light tails or a lack of outliers.



skewness types

kurtosis types



	Skewness	Kurtosis
loan_amnt	0.740607	-0.138568
term_months	1.208303	-0.540007
interest_rate	0.408461	-0.167339
installment	0.957986	0.709928
annual_income	43.351710	4575.089373
dti	0.523821	11.534291
earliest_cr_line	-1.014706	1.568242
open_acc	1.225312	3.023047
pub_rec	17.582331	1993.086414
revol_bal	11.955063	392.228951
revol_util	-0.045809	3.696422
total_acc	0.868852	1.249050
mort_acc	1.601658	4.432527
pub_rec_bankruptcies	3.389503	17.833842

Moderately skewed

- As a general rule of thumb:
  - If skewness is less than -1 or greater than 1, the distribution is **highly skewed.**
  - If skewness is between -1 and -0.5 or between 0.5 and 1, the distribution is **moderately skewed.**
  - If skewness is between -0.5 and 0.5, the distribution is **approximately symmetric**

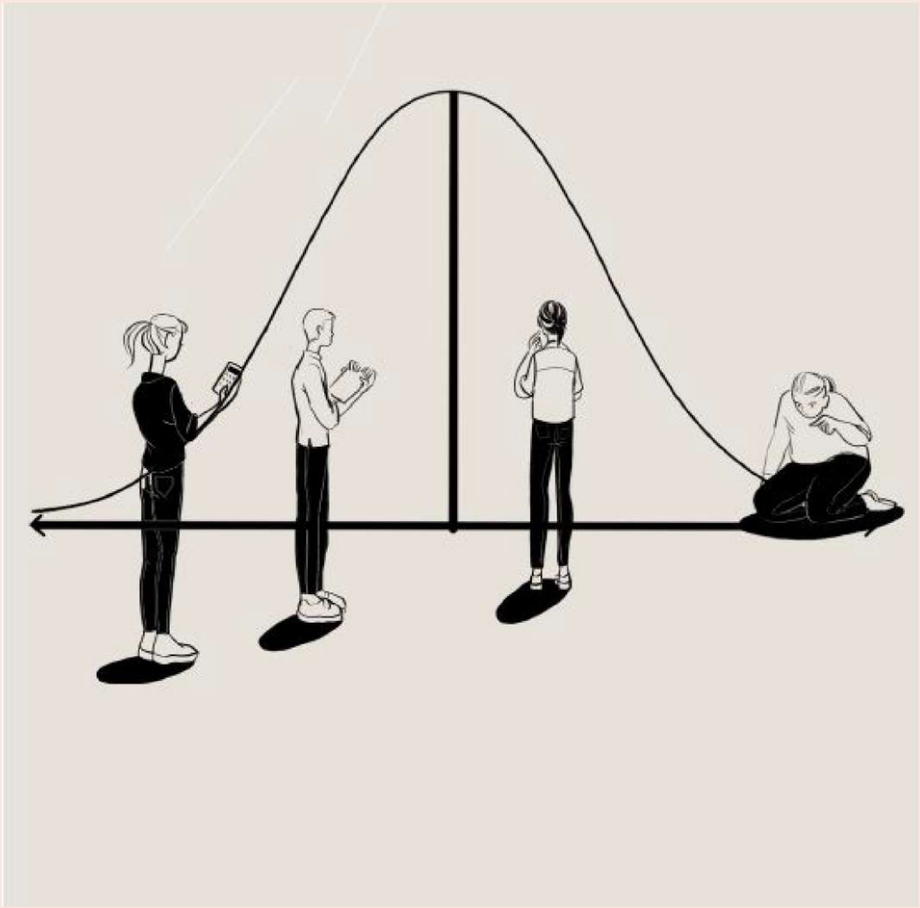
Extremely skewed

Highly skewed

Approximately symmetric

z-scores

probabilities for normal random variables



- A **z-score** or a "standardized score" is a measurement of how far away an individual value is from the mean within a normal distribution. To find the z-score of a value, the following formula is used:

$$Z = \frac{(observed\ value\ - mean)}{SD}$$

- A z-score is used for the normality test using skewness and kurtosis. A z-score could be obtained by dividing the skew values or excess kurtosis by their standard errors.

$$Z = \frac{\text{Skew value}}{SE_{\text{skewness}}}, Z = \frac{\text{Excess kurtosis}}{SE_{\text{excess kurtosis}}}$$

Therefore, critical values for rejecting the null hypothesis need to be different according to the sample size as follows:

- For **small samples** ($n < 50$), if absolute **z-scores for either skewness or kurtosis are larger than 1.96**, which corresponds with an alpha level of 0.05, then reject the null hypothesis and conclude the distribution of the sample is non-normal.

- For **medium-sized samples** ($50 < n < 300$), reject the null hypothesis at an **absolute z-value over 3.29**, which corresponds with an alpha level of 0.05, and conclude the distribution of the sample is non-normal.

- For **sample sizes greater than 300**, depending on the histograms and the absolute values of skewness and kurtosis without considering z-values. **Either an absolute skew value larger than 2 or an absolute kurtosis (proper) larger than 7** may be used as reference values for determining substantial non-normality.

Frequencies

Statistics

		loan_amnt	term_months	interest_rate	installment	annual_income
N	Valid	395986	395986	395986	395986	395986
	Missing	0	0	0	0	0
Mean		14113,77	41,70	13,639447478	431,8423	74202,8367
Median		12000,00	36,00	13,330000000	375,4300	64000,0000
Mode		10000	36	10,99000000	327,34	60000,00
Std. Deviation		8357,387	10,212	4,4721583121	250,72112	61638,83584
Variance		69845916,896	104,288	20,000	62861,078	3799346083,8
Skewness		,777	1,234	,421	,984	41,045
Std. Error of Skewness		,004	,004	,004	,004	,004
Kurtosis		-,063	-,477	-,144	,784	4238,688
Std. Error of Kurtosis		,008	,008	,008	,008	,008

Two variable analysis

grade_rank by interest_rate

graderank	Skewness	Std Error	Z-Value	Kurtosis	Std Error	Z-Value
A	-0.173	0.010	**-17.300**	-0.748	0.019	**-39.368**
B	-0.094	0.007	**-13.429**	-0.559	0.014	**-39.929**
C	0.230	0.008	**28.750**	-0.207	0.015	**-13.800**
D	-0.123	0.010	**-12.300**	0.635	0.019	**33.421**
E	-0.236	0.014	**-16.857**	2.265	0.028	**80.893**
F	-1.262	0.023	**-54.870**	5.840	0.045	**129.778**
G	-1.276	0.044	**-29.000**	6.683	0.890	**7.509**

All z-values are far away from +/- 3.96 (n>300).
Our example data is skewed and kurtotic for each grade_rank
and we can assume that our data is **NOT normal**.

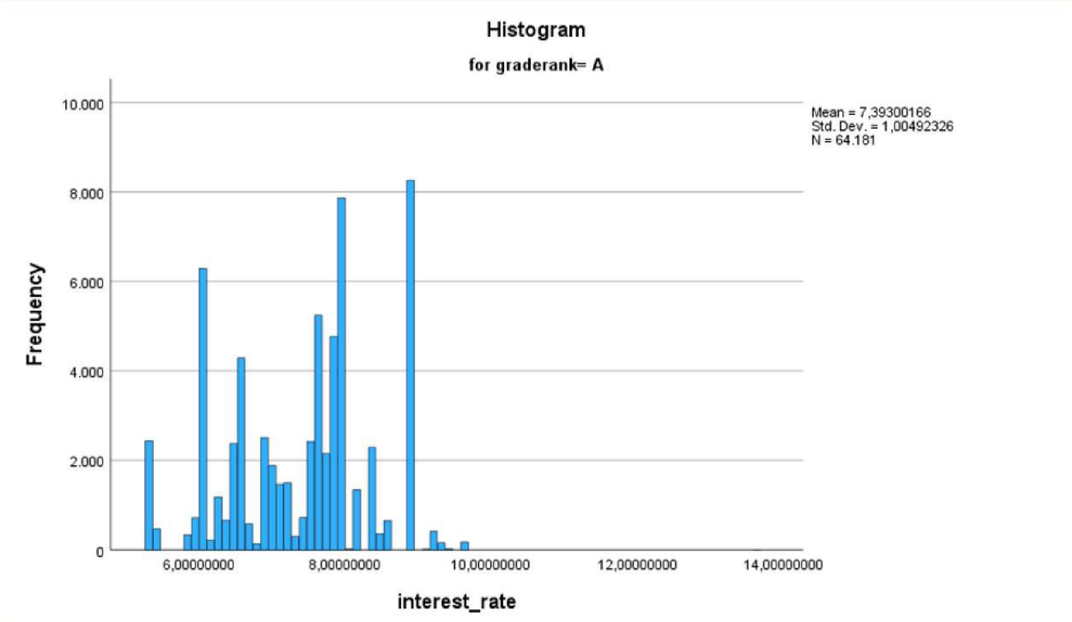
also confirmed by

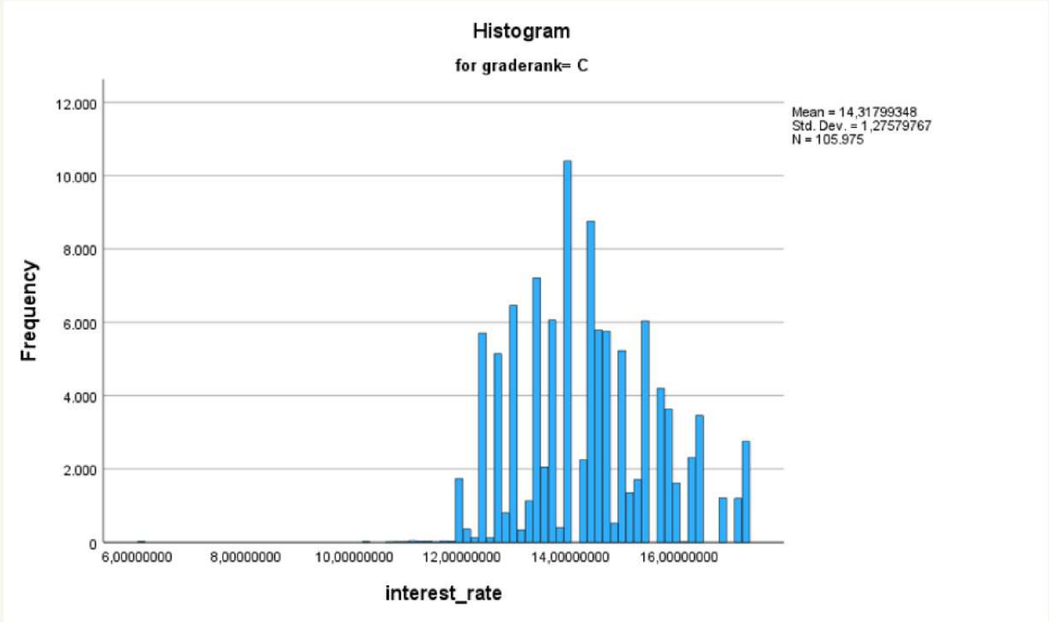
Tests of Normality

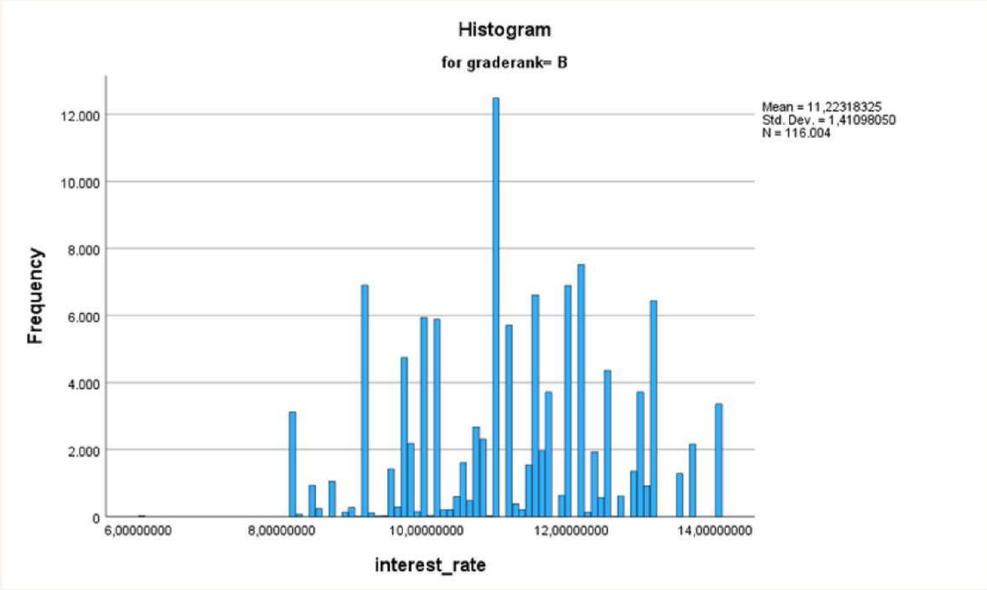
		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	graderank	Statistic	df	Sig.	Statistic	df	Sig.
interest_rate	A	,113	64181	<,001			
	B	,075	116004	<,001			
	C	,065	105975	<,001			
	D	,086	63517	<,001			
	E	,075	31484	<,001			
	F	,172	11771	<,001			
	G	,280	3054	<,001	,842	3054	<,001

a. Lilliefors Significance Correction

Histograms for different grades





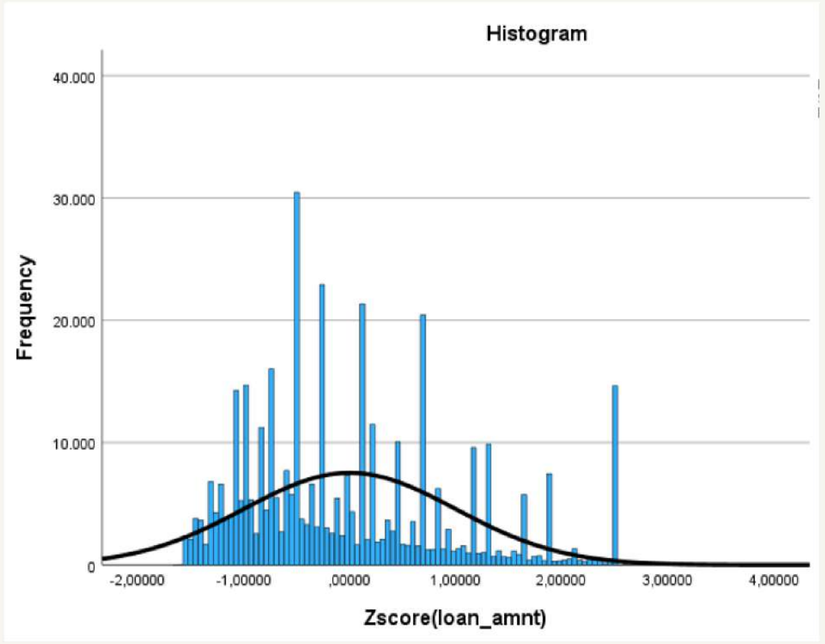


loan_amnt

Statistics

Zscore(loan_amnt)

N	Valid	395986
	Missing	0
Mean		,0000000
Median		-,2529228
Mode		-,49223
Std. Deviation		1,00000000
Variance		1,000
Skewness		,777
Std. Error of Skewness		,004
Kurtosis		-,063
Std. Error of Kurtosis		,008



loan_amnt	Zloan_amnt
10000	-,49223
8000	-,73154
15600	,17783
7200	-,82726
24375	1,22780
20000	,70431
18000	,46500
13000	-,13327
18900	,57269

z	0
+0	.50000
+0.1	.53983
+0.2	.57926
+0.3	.61791
+0.4	.65542
+0.5	.69146
+0.6	.72575
+0.7	.75804

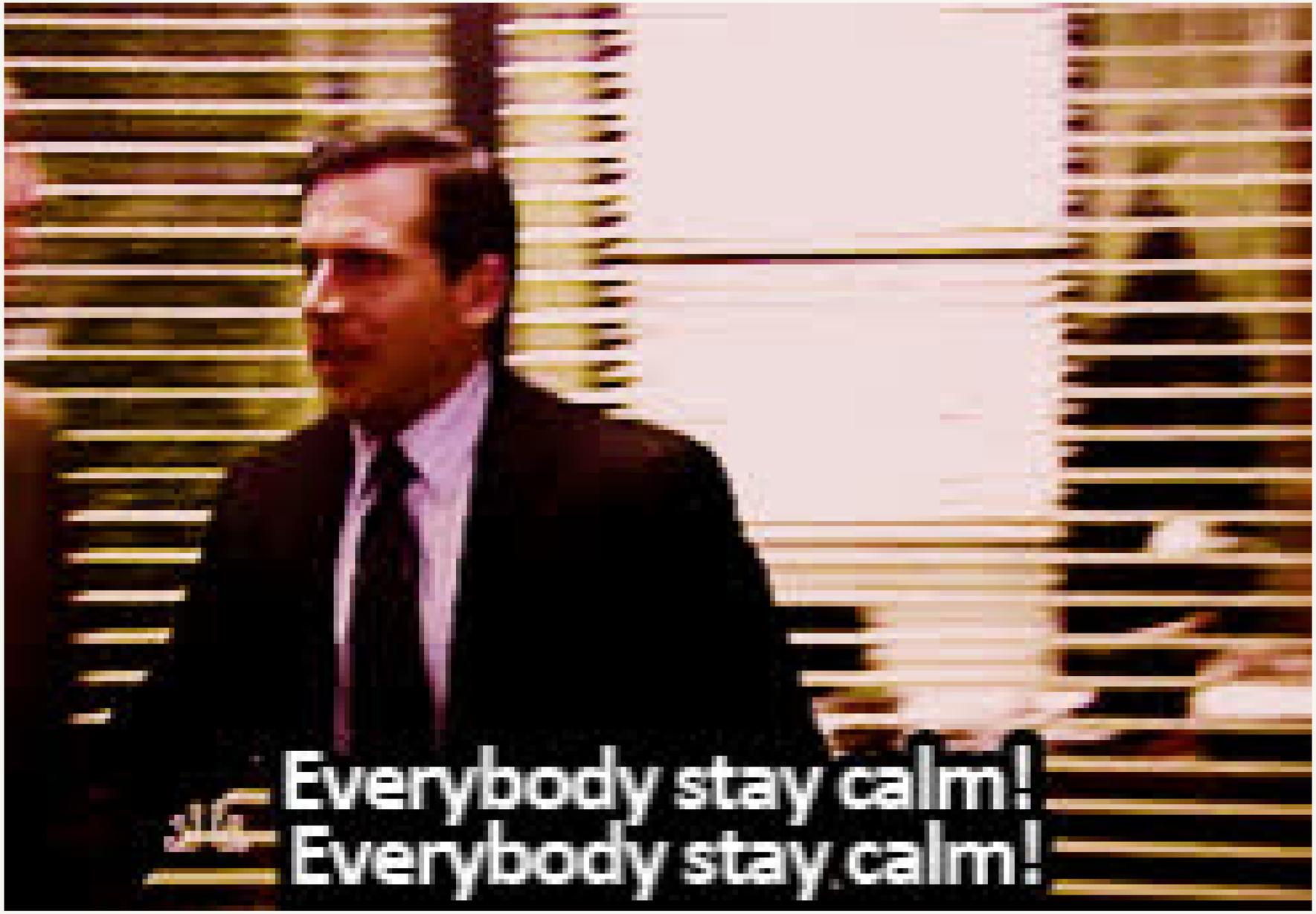
In the lending club loan dataset, what is the probability of granting a loan higher than 20,000?

P(X=20000) = P(z = 0.70431) = 75.80%

P(X>20000) = P(z > 0.70431) = 100% - 75.8%
= **24.20%**

Problem

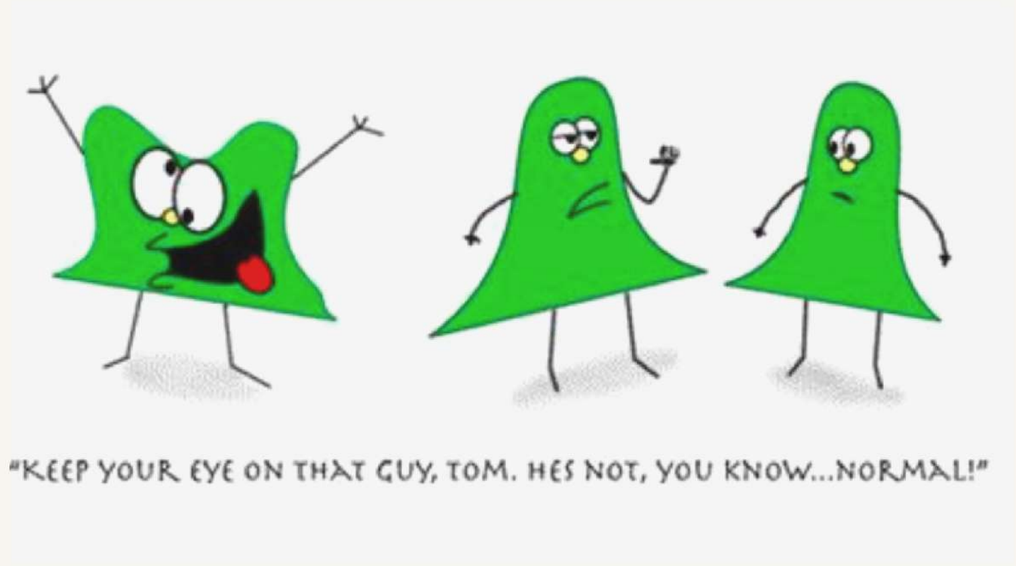
The problem with using the methods such as skewness, histograms, and Q-Q plots for assessing normality is that we cannot be absolutely certain whether the variable is truly normal or it "just seems like normal".


Everybody stay calm!
Everybody stay calm!

Solution

The good thing is we have a solution for reducing that uncertainty in our decision within statistical limits - "**Statistical Normality Tests**"

Statistical tests for Normality



Shapiro–Wilk test

Kolmogorov–Smirnov test

Anderson–Darling test

- Statistical tests for normality are more precise since actual probabilities are calculated.

- Tests for normality calculate the probability that the sample was drawn from a normal distribution.

- Each Test will return the following two metrics that will help us to interpret the results.
  - **Statistic**
  - **p-value**

- The Statistical Tests make an assumption that the sample is drawn from the Gaussian Distribution. This is called the null hypothesis or Ho. A threshold level known as alpha which is usually 5% is used to interpret the p-value.
  - **Ho: The sample data are not significantly different from a normal population.**
  - **Ha: The sample data are significantly different from a normal population.**

- When testing for normality:
  - If the p-value <= 0.05; the null hypothesis can be rejected (the variable is NOT normally distributed). •
  - If the p-value > 0.05; the null hypothesis cannot be rejected (the variable MAY BE normally distributed).

Broadly speaking, results with a larger p-value confirm that our sample was likely drawn from a Gaussian distribution.

Shapiro-Wilk test

- Fairly powerful omnibus test. Usually good with small samples or discrete data.
- Good power with symmetrical, short and long tails. Good with asymmetry.

Anderson-Darling test

- Similar in power to Shapiro-Wilk but has less power with asymmetry.
- Works well with discrete data.

Kolmogorov-Smirnov test

- All tend to have lower power. Data have to be very non-normal to reject Ho.
- These tests can outperform other tests when using discrete or grouped data.

Column_name	Shapiro Wilk		Komogorov Smirnov		Anderson-Darling	
	SW Statistic	p-value	KS Statistic	p-value	AD Statistic	critical_value (at 5% significance level)
loan_amnt	0.531320	<0.00	1.000000	<0.00	5424.416	0.787
term_months	0.983430	<0.00	1.000000	<0.00	82000.334	0.787
interest_rate	0.938586	<0.00	1.000000	<0.00	1041.709	0.787
installment	0.453948	<0.00	1.000000	<0.00	5422.288	0.787
annual_income	0.984930	<0.00	0.979990	<0.00	25425.719	0.787
dti	0.947193	<0.00	1.000000	<0.00	666.708	0.787
earliest_cr_line	0.929324	<0.00	0.996205	<0.00	4300.071	0.787
open_acc	0.355772	<0.00	0.500000	<0.00	4987.303	0.787

Column_name	Shapiro Wilk		Komogorov Smirnov		Anderson-Darling	
	SW Statistic	p-value	KS Statistic	p-value	AD Statistic	critical_value (at 5% significance level)
pub_rec	0.488958	<0.00	0.996752	<0.00	84095.460	0.787
revol_bal	0.981521	<0.00	0.985920	<0.00	31381.552	0.787
revol_util	0.958086	<0.00	0.999456	<0.00	996.788	0.787
total_acc	0.808839	<0.00	0.500000	<0.00	2801.732	0.787
mort_acc	0.372111	<0.00	0.500000	<0.00	19599.401	0.787
pub_rec_bankruptcies	0.942999	<0.00	1.000000	<0.00	101468.769	0.787

looks like none of the columns in our data passed the tests

So what's next?
P.S: it's not over yet..

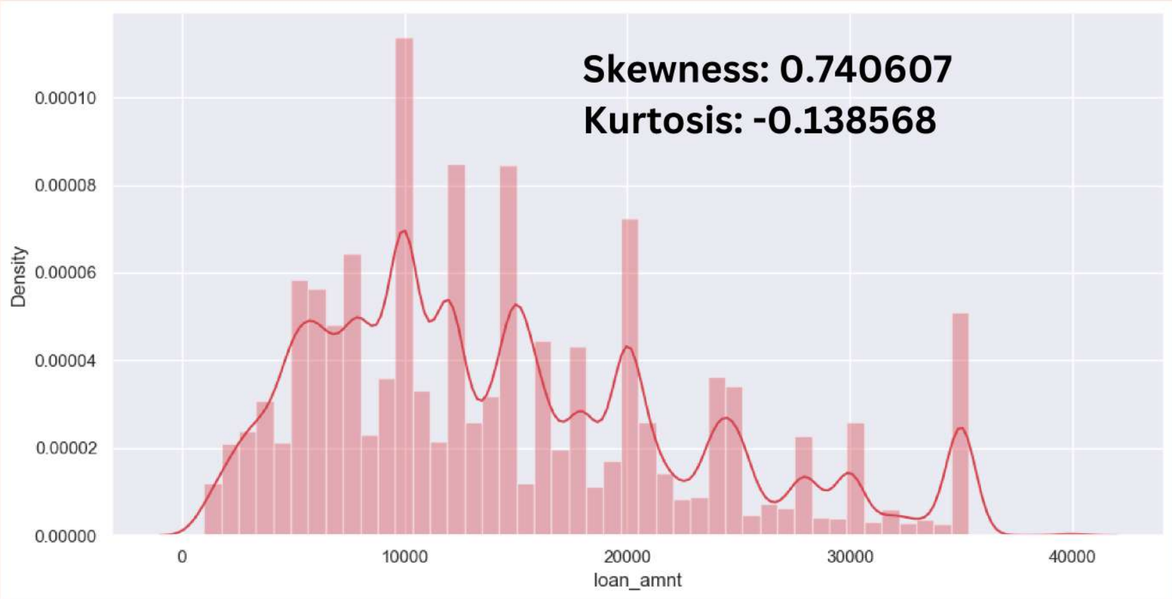
If a given dataset is **not normally distributed**, we can often perform one of the following transformations to make it more normally distributed:

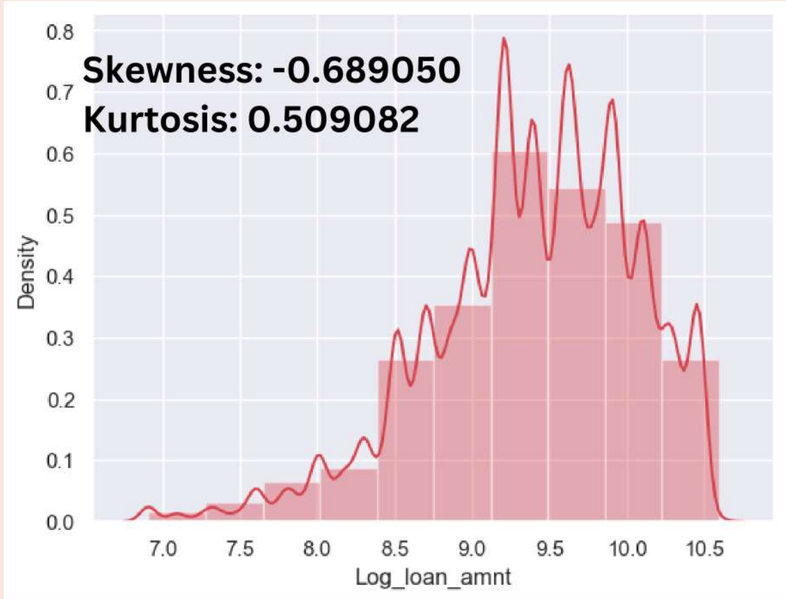
- **Log Transformation**: Transform the values from x to log(x).
- **Square Root Transformation**: Transform the values from x to √x.
- **Cube Root Transformation**: Transform the values from x to x1/3.

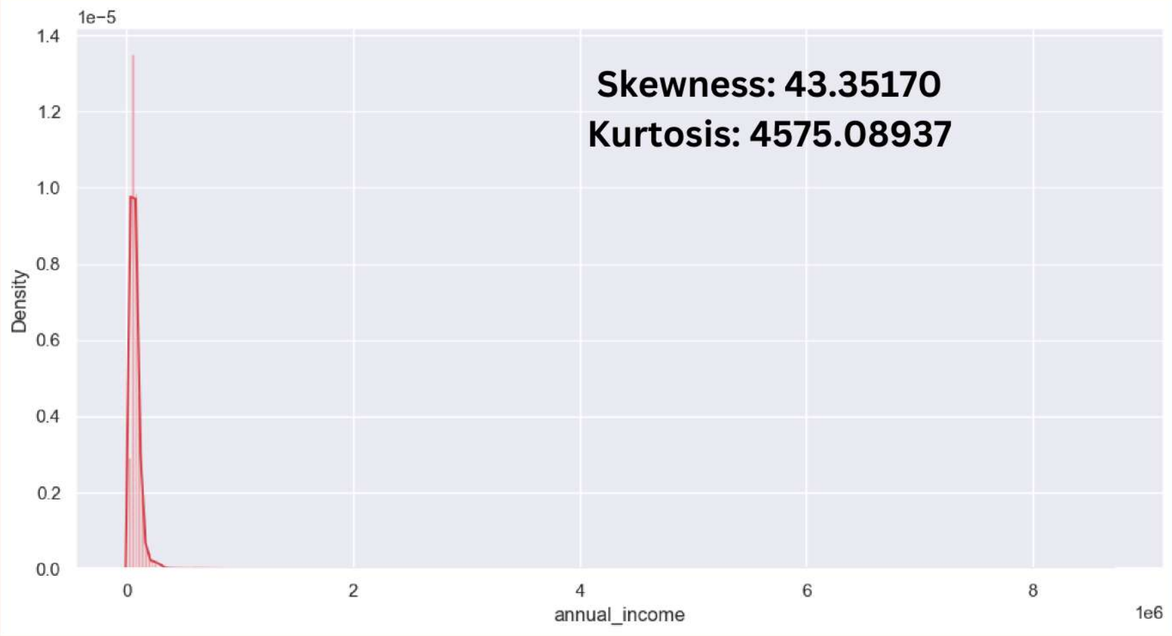
By performing these transformations, the dataset may become more normally distributed!

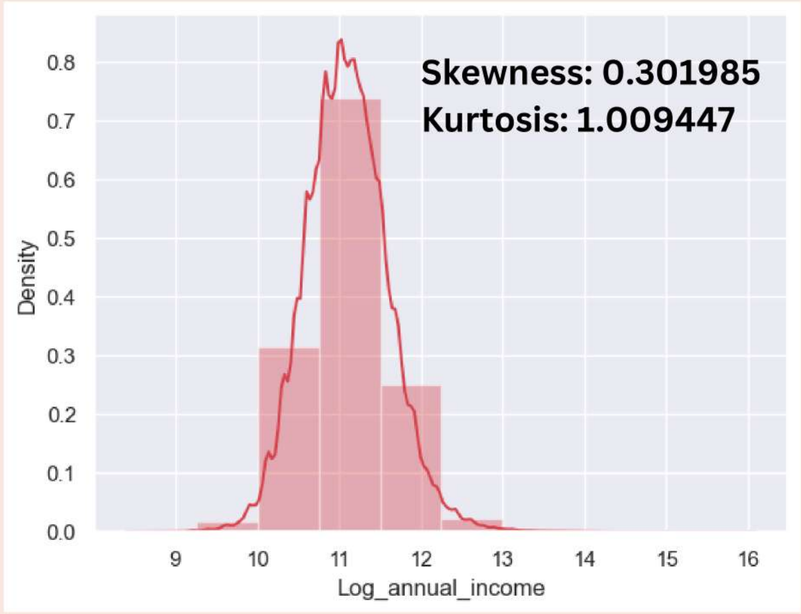
Log transformation of loan_amnt and annual_income

Before









After

Thank you for listening!

Questions?

Authors:

- **Diego Gules Butori**
- **Nairuhi Tovmasyan**
- **Jaswinder Singh**
- **Federico Andrés Gómez Quiroga**
- **Cian O'Sullivan**

17th December 2022

Hypothesis Testing for Lending Club Loan Data

Objectives

Goal

To perform the hypothesis testing on the lending club loan dataset and stating any necessary assumptions that were made during the tests

Stops!

- One-Sample t-test
- Paired-Samples t-test
- Unpaired t-test
- One-factorial ANOVA
- Two-factorial ANOVA

Introduction

01

What is hypothesis testing?

Hypothesis testing is a statistical method that is used in making statistical decisions using experimental data. Hypothesis Testing is basically an assumption that we make about the population parameter.

02

Why do we use it?

Hypothesis testing is the process used to evaluate the strength of evidence from the sample and provides a framework for making determinations related to the popultaion, i.e. it provides a method for understanding how reliably one can extrapolate observed findings ina sample under study to the larger population.

Steps involved in hypothesis testing

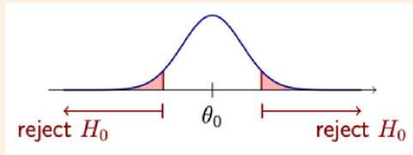
1. Defining Hypothesis:

First of all, we should understand which scientific question we are looking for an answer to, and it should be formulated in the form of the Null Hypothesis (H_0) and the Alternative Hypothesis (H_1 or H_a). Please remember that **H_0 and H_1 must be mutually exclusive**, and H_1 shouldn't contain equality

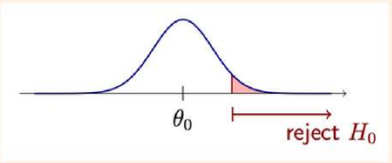
Forms of H_0 and H_1

- $H_0: \mu = x$, $H_1: \mu \neq x$
- $H_0: \mu \leq x$, $H_1: \mu > x$
- $H_0: \mu \geq x$, $H_1: \mu < x$

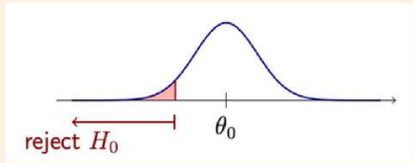
Two-tailed



Right-tailed



Left-tailed



2. Assumption Check

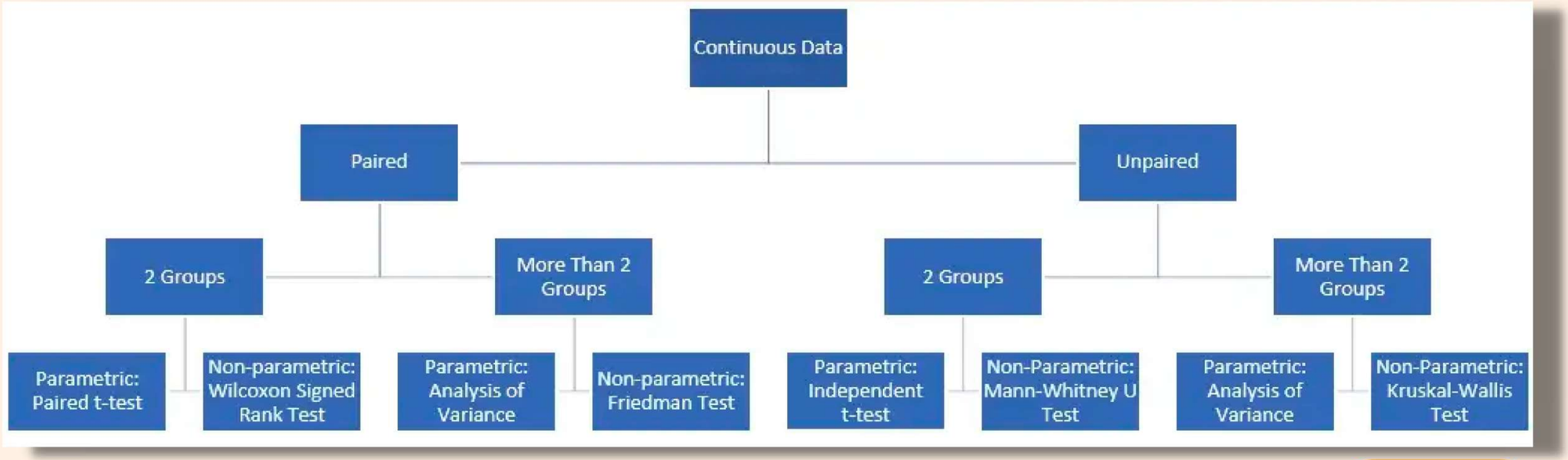
To decide whether to use the parametric or nonparametric version of the test, we should check the specific requirements listed below:

- Observations in each sample are **independent and identically distributed** (IID).
- Observations in each sample are **normally distributed.**
- Observations in each sample have the **same variance**.

For this presentation, we assume that all the assumptions are met for our data!

3. Selecting the proper test

Then we select the appropriate test to be used. When choosing the proper test, it is essential to analyze how many groups are being compared and whether the data are paired or not. To determine whether the data is matched, it is necessary to consider whether the data was collected from the same individuals.

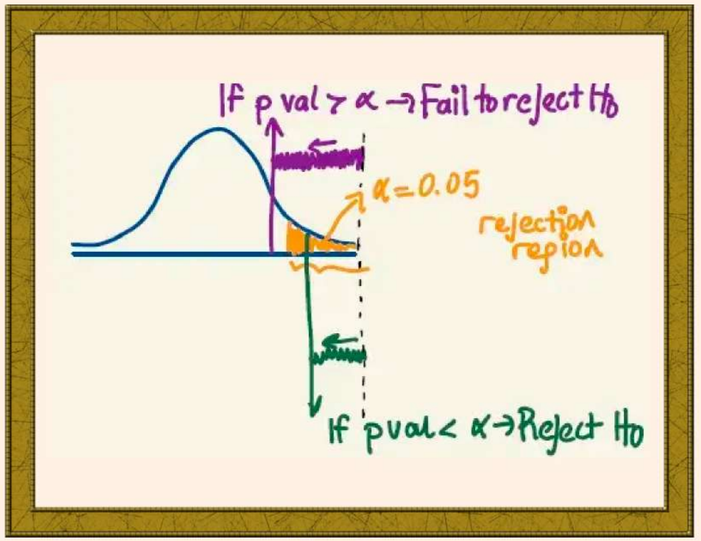


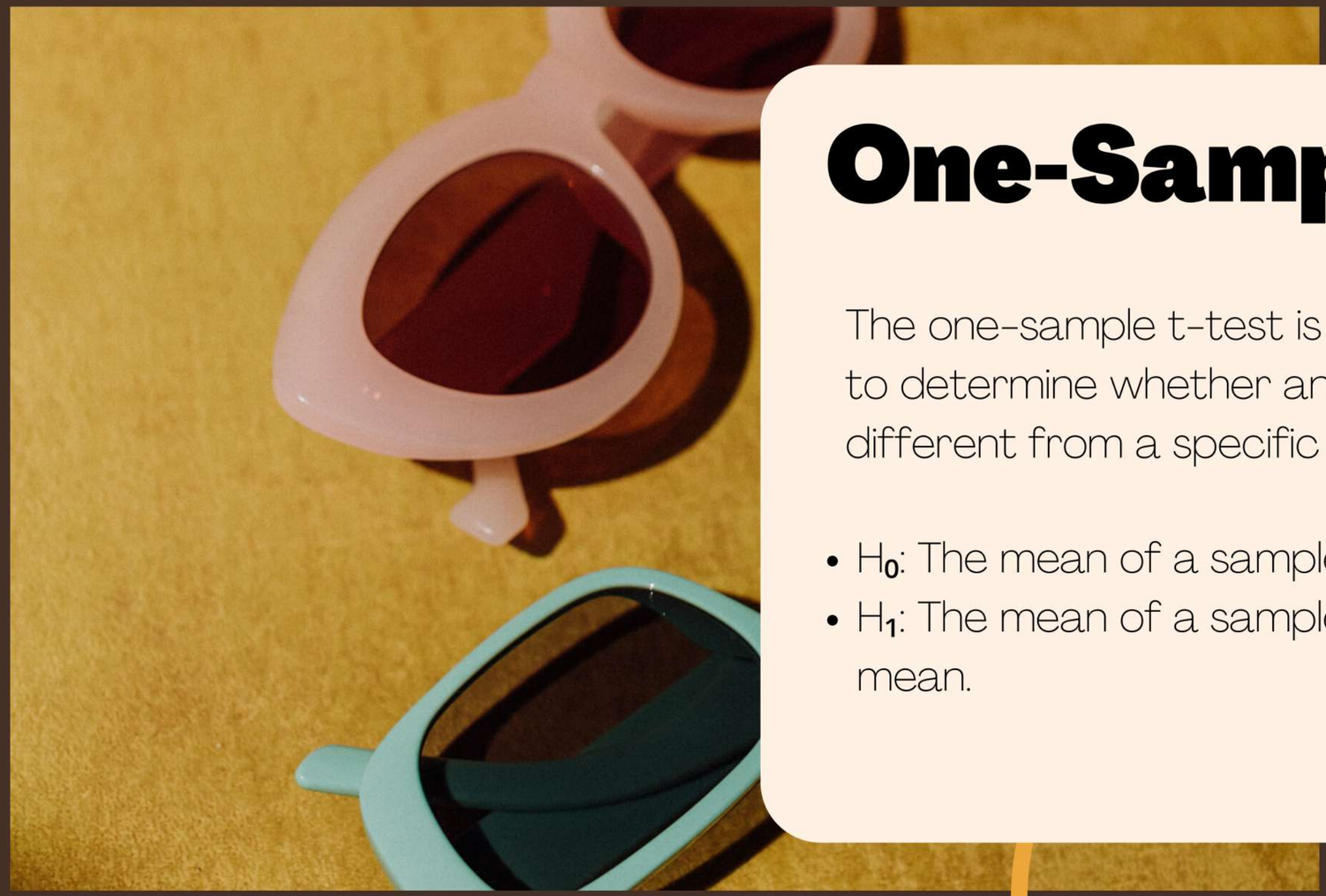
4. Decision and Conclusion

After performing the hypothesis testing, we obtain a related p-value that shows the significance of the test. If the p-value is smaller than the alpha (the significance level), in other words, there is enough evidence to prove H_0 is not valid; you can reject H_0. Otherwise, you fail to reject H_0.

- **Rejecting H_0 validates H_1**.
- However, **failing to reject H_0 does not mean H_0 is valid, nor does it mean H_1 is wrong**.

Remeber this when making decisions about your hypothesis!





One-Sample t-test

The one-sample t-test is a statistical hypothesis test used to determine whether an unknown population mean is different from a specific value.

- H_0: The mean of a sample is equal to the population mean.
- H_1: The mean of a sample is different from the population mean.

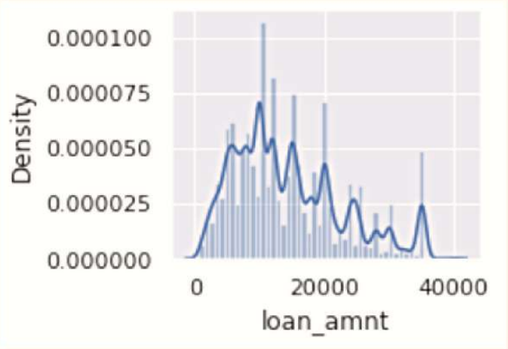
Column: 'loan_amnt'

The **average loan amount of the sector**, based on public statistics is **20,000**. Stakeholders want to know: if lending club loan is equal to that number to measure the effectiveness of the management against the competition.

- H_0: **the mean of the loan amount is equal to the population mean**
- H_1: **the mean of the loan amount is not equal to population mean**

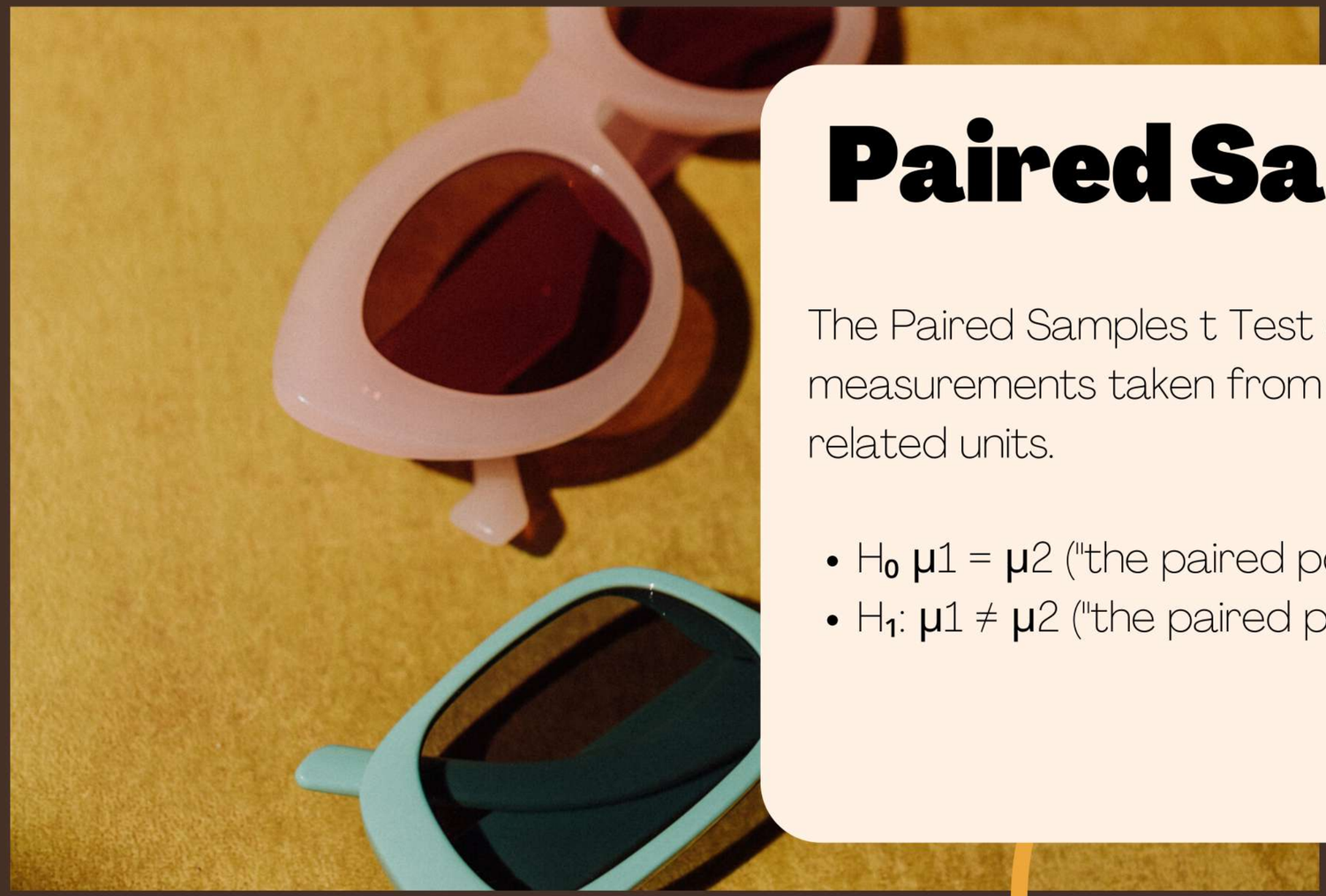
```
stats.ttest_1samp(df.loan_amnt,20000)

Ttest_1sampResult(statistic=-419.0498279544623, pvalue=0.0)
```



Result

As the p_value for the given problem is less than 0.05 which is the alpha value, we **reject the null hypothesis** and accept the alternative hypothesis (mean of lending club amount is not equal to the mean of the sector)



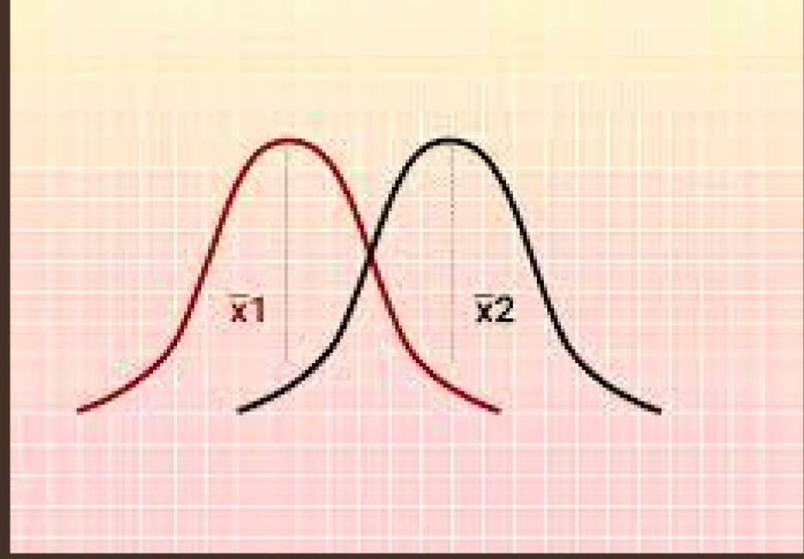
Paired Samples t-test

The Paired Samples t Test compares the means of two measurements taken from the same individual, object, or related units.

- H_0 $\mu 1 = \mu 2$ ("the paired population means are equal")
- H_1: $\mu 1 \neq \mu 2$ ("the paired population means are not equal")

We don't have any factor for our sample to measure if there is any difference between the members before and after the application of the new factor!

Unpaired t-test



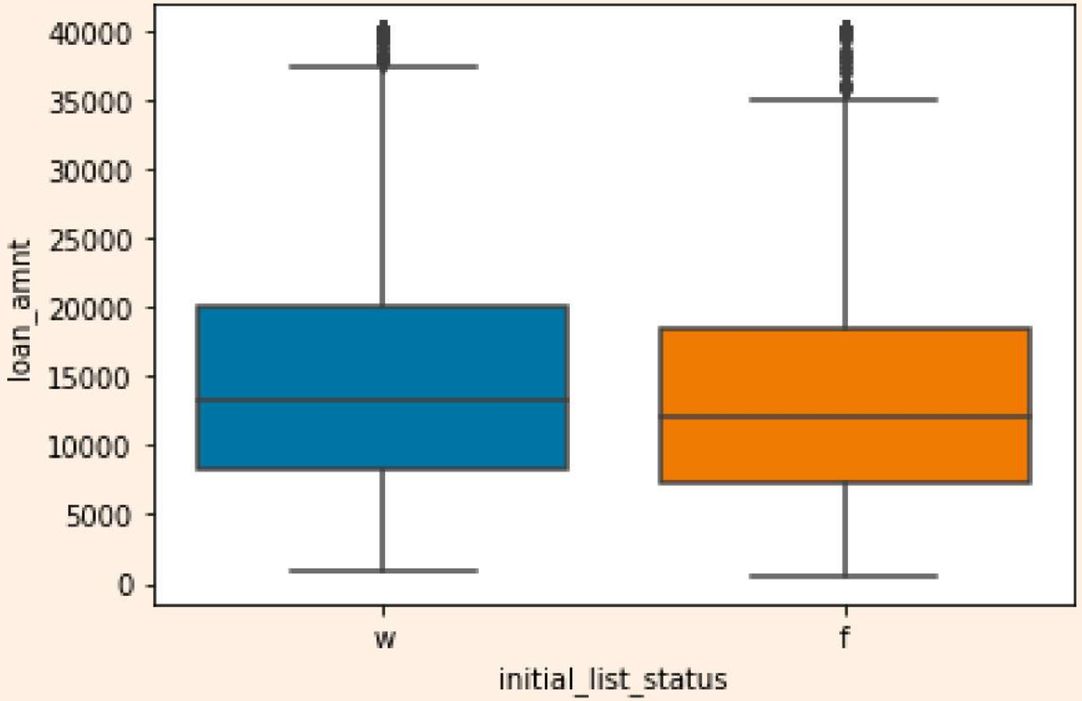
H_0

Two independent samples have similar mean values

H_1

Two independent samples **DO NOT** have similar mean values.

Hypothesis: The type of initial list status (w-whole, f-fraction) impacts the amount of loan?

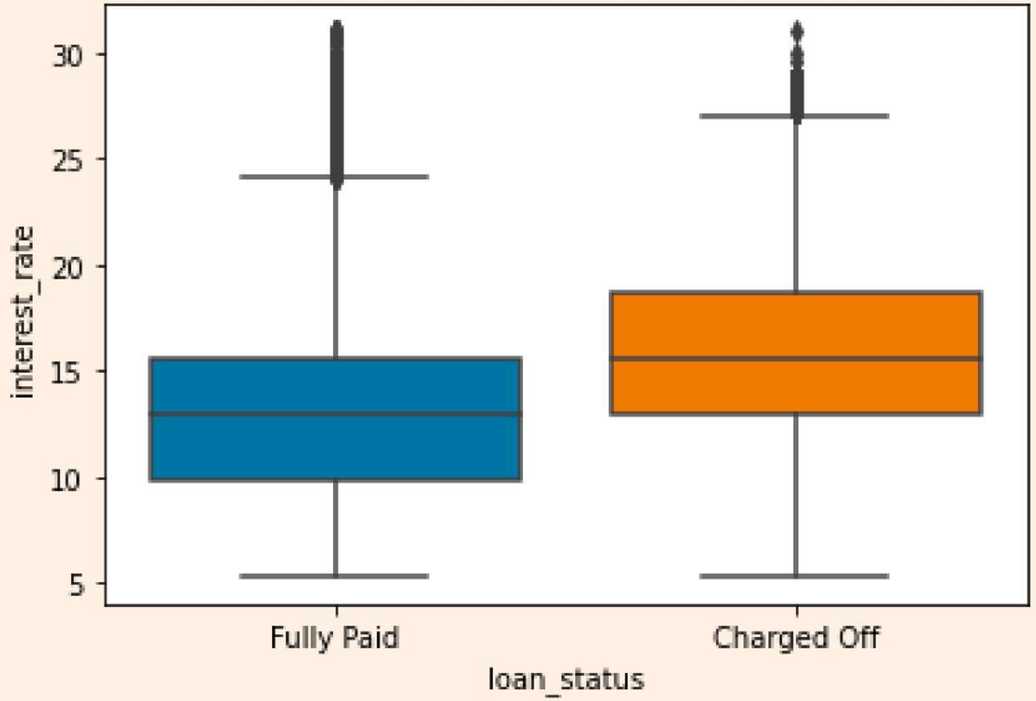



```
Ttest_indResult(statistic=-44.44857170504447, pvalue=0.0)
```

Result

Since the p_value is less than 0.05 which is the alpha value, we **reject the null hypothesis** and accept the alternative hypothesis. Therefore **initial_list_status** affects the **loan_amnt**

Hypothesis: Loan status has an influence on the interest rate.




```
Ttest_indResult(statistic=-127.09759633999799, pvalue=0.0)
```

Result

Since the p_value is less than 0.05 which is the alpha value, we **reject the null hypothesis** and accept the alternative hypothesis. Therefore **loan_status** affects the **interest_rate**

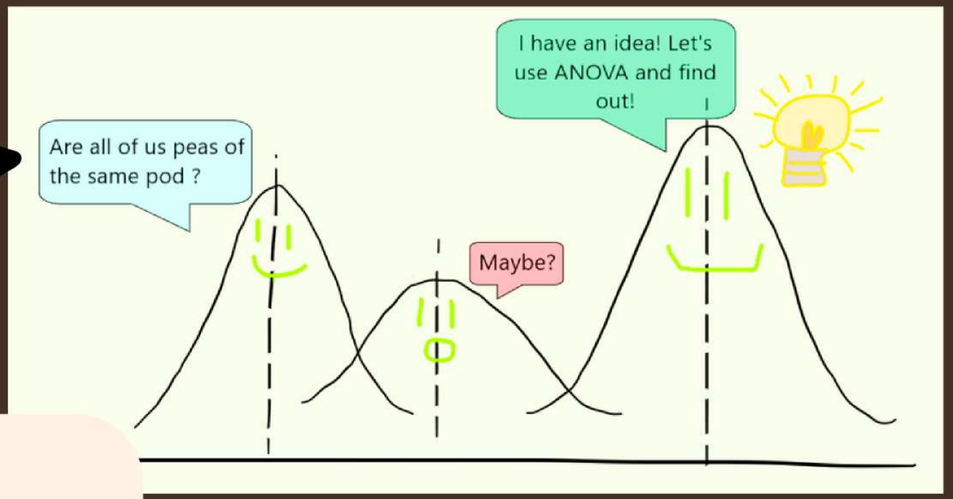
One-factorial ANOVA

One-Way ANOVA ("analysis of variance") compares the **means of two or more independent groups** in order to determine whether there is statistical evidence that the **associated population means are significantly different.**

Variables used

The variables used in this test are known as:

- **Dependent variable**
- **Independent variable** (also known as the **grouping variable**, or **factor**)
    - This variable divides cases into two or more mutually exclusive levels, or groups



Hypothesis

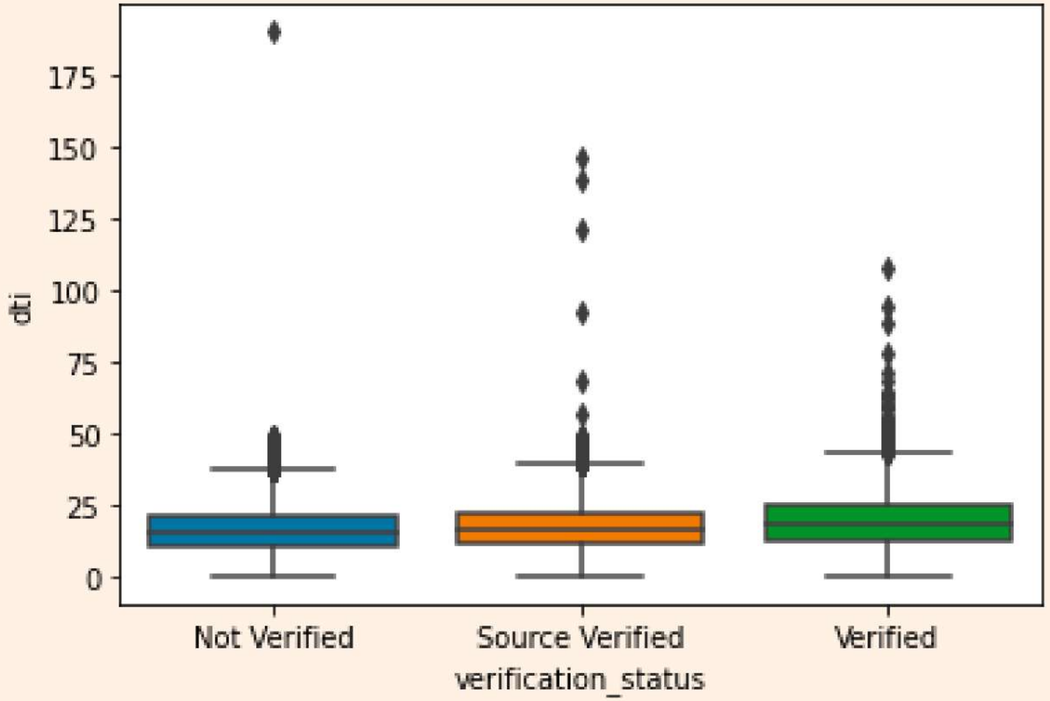
The null and alternative hypotheses of one-way ANOVA can be expressed as:

- H_0: $\mu 1 = \mu 2 = \mu 3 = \ldots = \mu k$ ("all k population means are equal")
- H_1: At least one μi different ("at least one of the k population means is not equal to the others")

where

μi is the population mean of the ith group (i = 1, 2, ..., k)

Hypothesis: Verification status has an influence on the debt to income ratio (dti).

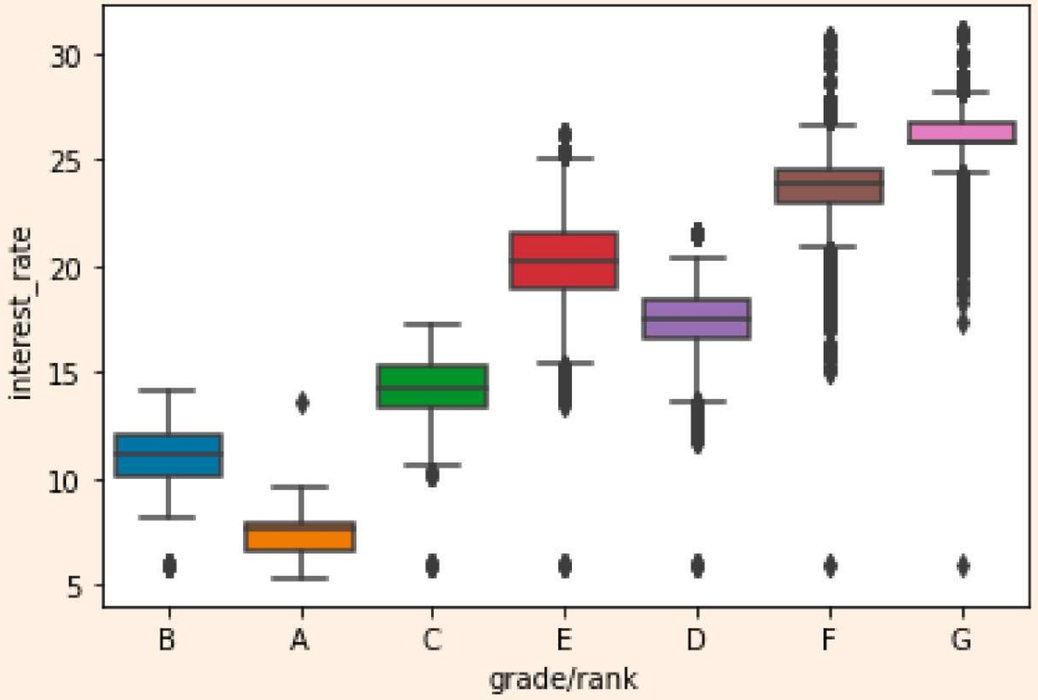



```
F_onewayResult(statistic=2734.7344723257797, pvalue=0.0)
```

Result

Since the p-value is less than 0.05 which is the alpha value, we **reject the null hypothesis** and accept the alternative hypothesis. Therefore **verification_status** affects the **dti**

Hypothesis: The grade has an influence on the interest rate

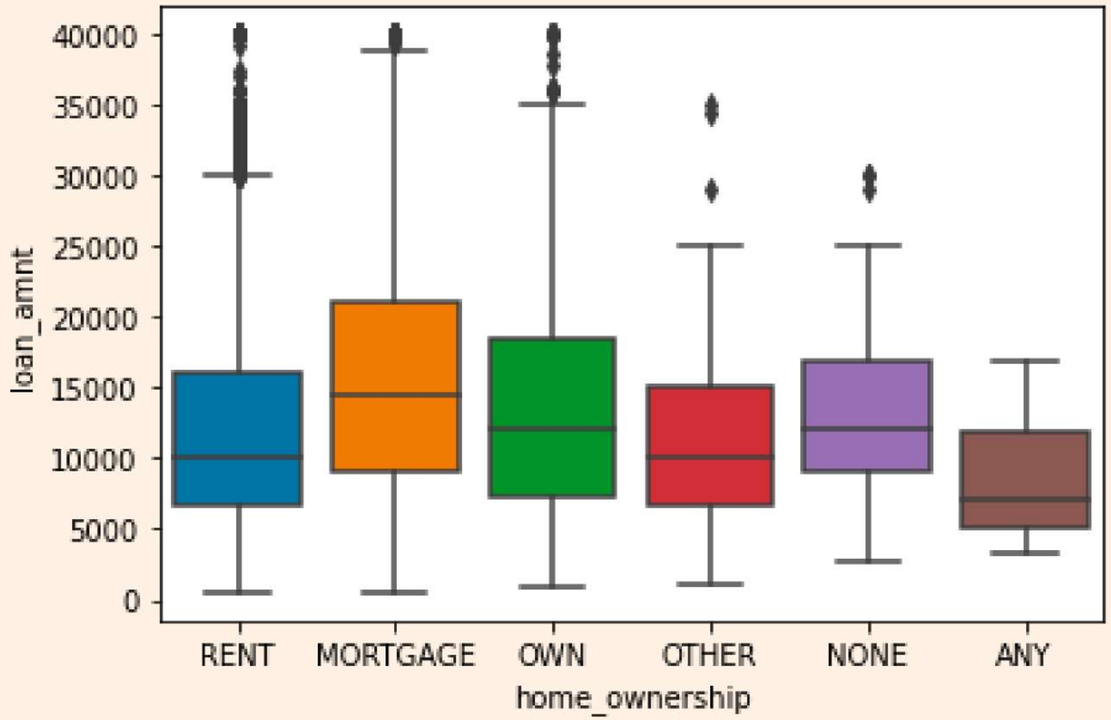



```
F_onewayResult(statistic=670525.7241141677, pvalue=0.0)
```

Result

Since the p-value is less than 0.05 which is the alpha value, we **reject the null hypothesis** and accept the alternative hypothesis. So, there is a significant difference in the mean between groups and so **grade** has an influence on the **interest rate**.

Hypothesis: Home ownership has an influence on the loan amount.




```
F_onewayResult(statistic=2913.291700148934, pvalue=0.0)
```

Result

Since the p-value is less than 0.05 which is the alpha value, we **reject the null hypothesis** and accept the alternative hypothesis. So, there is a significant difference in the mean between groups and so **home ownership** has an influence on the **loan amount**.

Two-factorial ANOVA

The **two-way ANOVA** compares the **mean differences between groups that have been split on two independent variables** (called **factors**). The primary purpose of a two-way ANOVA is to understand if there is an **interaction between the two independent variables** on the **dependent variable**

Variables used

The variables used in this test are known as:

- **Dependent variable**
- **Two Independent variables** (also known as the **grouping variables**, or **factors**)

Hypothesis

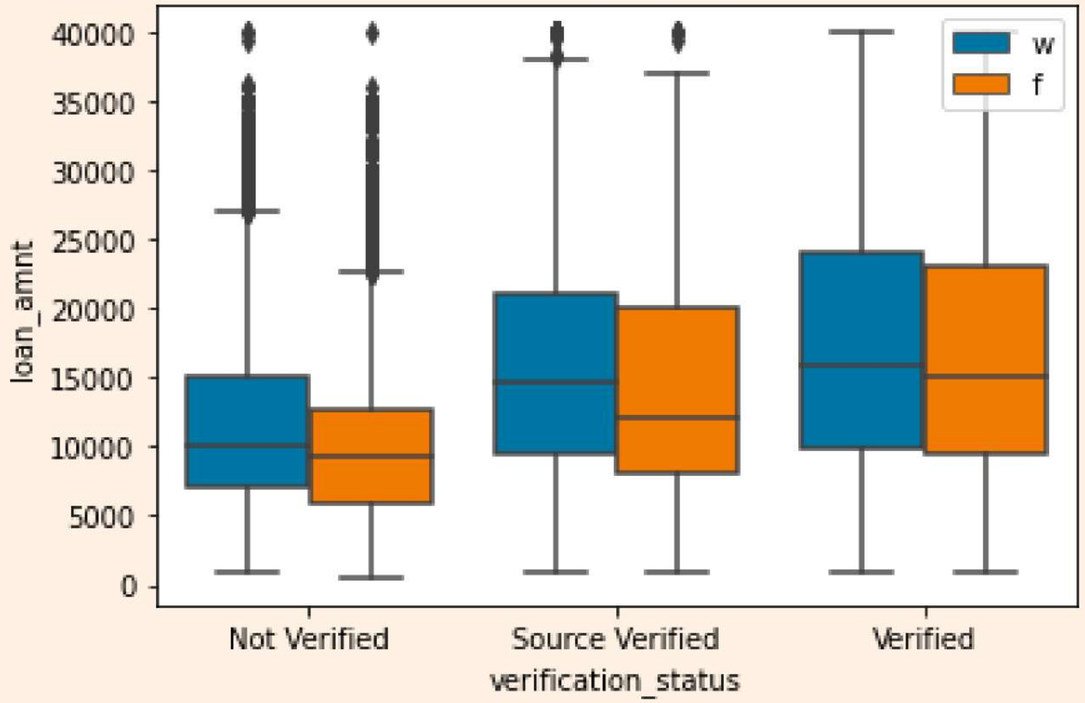
H_0

- There are **no significant differences** in the mean between the **groups** (**factor levels**) of the **first factor**.
- There are **no significant differences** in the mean between the **groups** (**factor levels**) of the **second factor.**
- One factor has **no effect** on the other factor.

H_1

- There is a **significant difference** in the mean between the **groups** (**factor levels**) of the **first factor**.
- There is a **significant difference** in the mean between the **groups** (**factor levels**) of the **second factor**.
- One factor **has an effect** on the other factor.

Dependent Variable: loan_amnt
Independent Variables: verification_status, initial_list_status

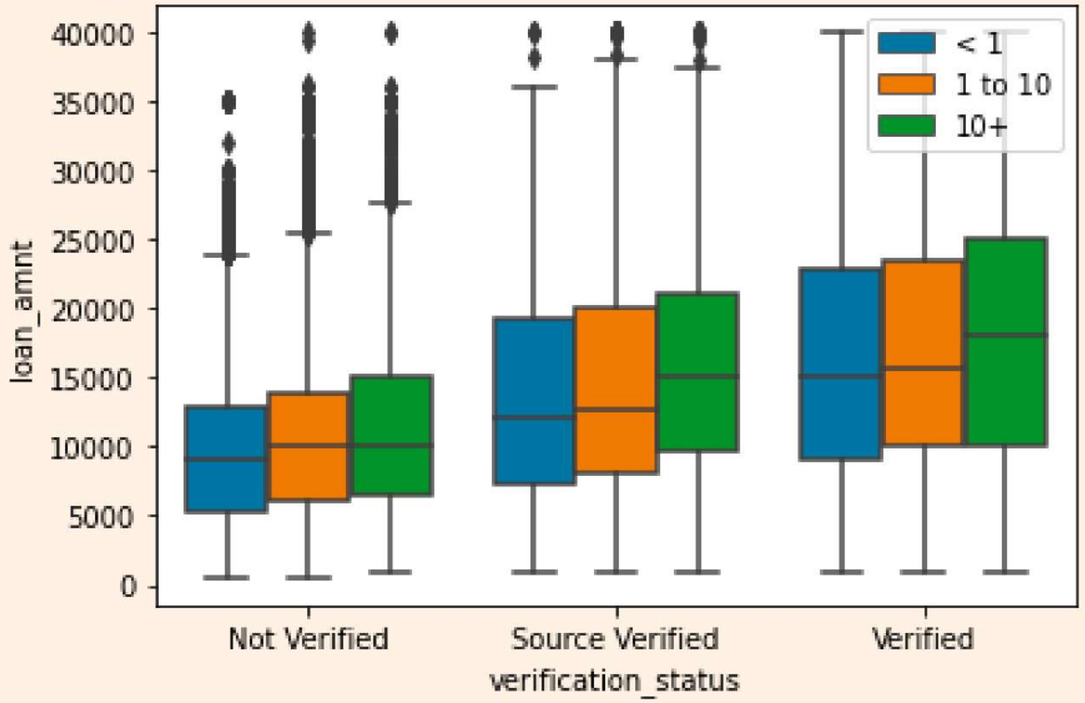


	sum_sq	df	F	PR(>F)
C(verification_status)	2.784884e+12	2.0	22339.368385	0.000000e+00
C(initial_list_status)	1.666610e+11	1.0	2673.792395	0.000000e+00
C(verification_status):C(initial_list_status)	2.456204e+10	2.0	197.028081	2.980827e-86
Residual	2.468370e+13	396008.0	NaN	NaN

Result

Since the p-values are less than 0.05 which is the alpha value, we **reject the null hypothesis** and accept the alternative hypothesis. Therefore there is a significant difference between the means of **verification_status** and **initial_list_status**

Dependent Variable: loan_amnt
Independent Variables: verification_status, emp_length_years

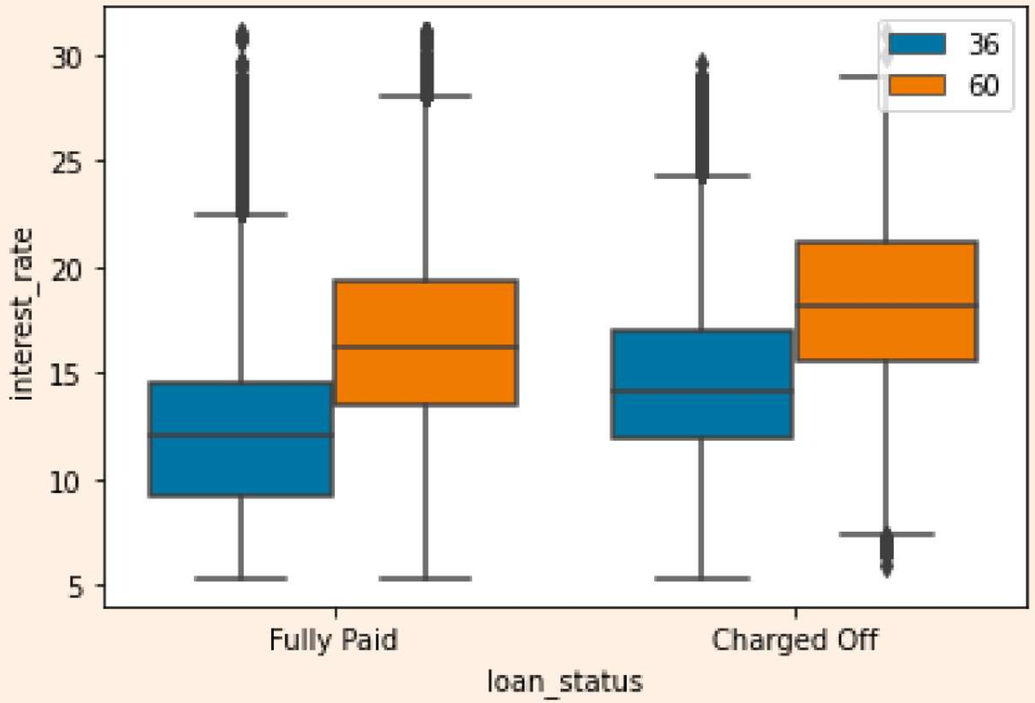


	sum_sq	df	F	PR(>F)
C(verification_status)	3.008396e+12	2.0	24449.239665	0.000000e+00
C(emp_length_years)	1.552537e+11	2.0	1261.747051	0.000000e+00
C(verification_status):C(emp_length_years)	6.164339e+09	4.0	25.048799	8.996275e-21
Residual	2.323846e+13	377718.0	NaN	NaN

Result

Since the p-values are less than 0.05 which is the alpha value, we **reject the null hypothesis** and accept the alternative hypothesis. Therefore there is a significant difference between the means of **verification_status** and **emp_length_years**

Dependent Variable: interest_rate
Independent Variables: loan_status, term_months



	sum_sq	df	F	PR(>F)
C(loan_status)	2.428870e+05	1.0	15564.067749	0.000000e+00
C(term_months)	1.252590e+06	1.0	80265.283029	0.000000e+00
C(loan_status):C(term_months)	1.526601e+03	1.0	97.823765	4.601532e-23
Residual	6.180092e+06	396017.0	NaN	NaN

Result

Since the p-values are less than 0.05 which is the alpha value, we **reject the null hypothesis** and accept the alternative hypothesis. Therefore there is a significant difference between the means of **loan_status** and **term_months**

Effect Size and Power Analysis

Power analysis is a statistical analysis based on significance level, effect size, statistical power, and sample size. We can use it to calculate any one of the four values given the other three.

Effect size is the **magnitude of the difference** between the **null hypothesis** and the **alternative hypothesis**.

Cohen's-d Measure

- Cohen’s d measures the distance between the mean of treatment and control group while considering the standard deviation.

  - **0.2** - **small effect size**,
  - **0.5** - **medium** effect size
  - **0.8** - **large effect size**.

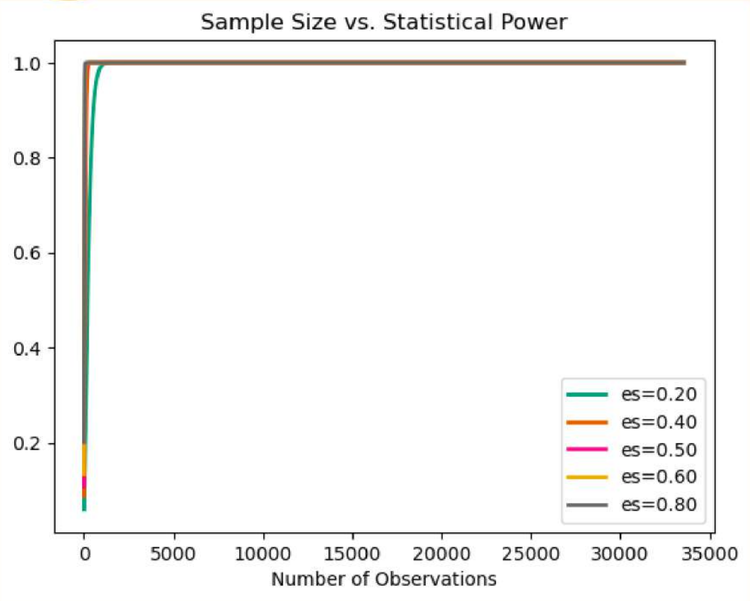
Statistical Power

- Statistical power is the probability of correctly rejecting the null hypothesis. Or we can say it’s the probability of detecting an effect if it exists. **Power equals one minus beta**, where beta is the **false negative rate or type II error**. 0.8 is a commonly used value for power.

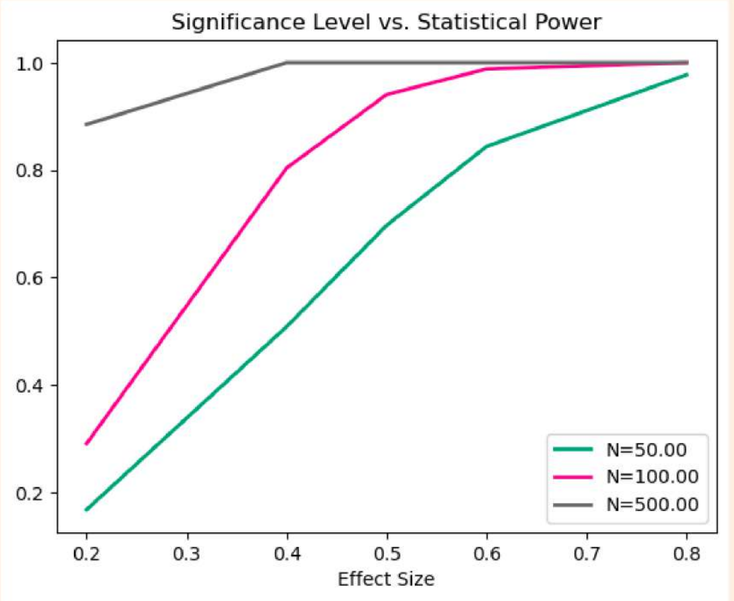
- The sample size is the **minimum sample size** needed per group.

Columns Name	Cohen's d	Point Estimate	90% Confidence Interval	
			Lower	Upper
mort_acc	72.7526	0.001	-.002	.004
total_acc	988.7020	0.011	008	.013
revol_util	4365.1153	-0.012	-0.014	-0.009
revol_bal	20453.4965	-.226	-.228	0.223
pub_rec	1163.8298	.003	0.00	0.006
open_acc	5448.2666	0.021	0.018	0.024
earliest_cr_line	6090.5875	-0.810	-0.813	-0.807
dti	103.4902	-0.70	-0.073	-0.068
loan_amnt	8357.4386	-0.704	-0.707	-0.701

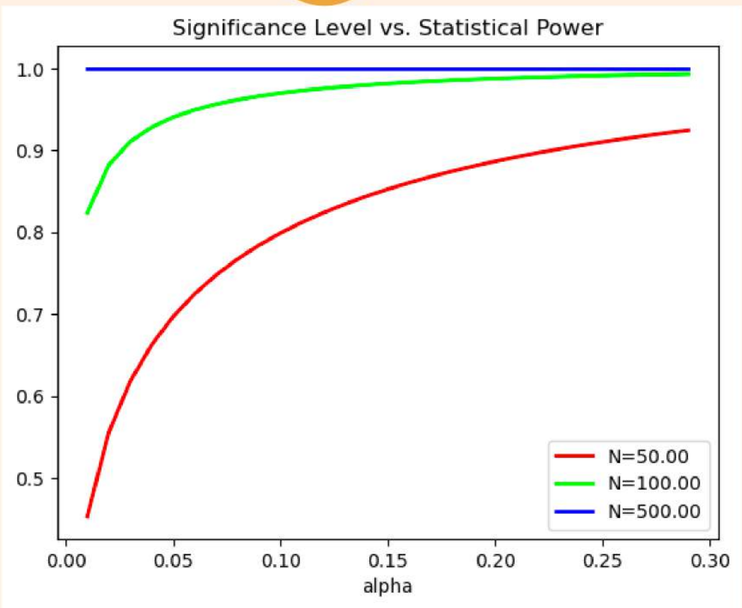
1


Sample Size vs. Statistical Power
1.0
0.8
0.6
0.4
0.2
0
5000
10000
15000
20000
25000
30000
35000
Number of Observations
es=0.20
es=0.40
es=0.50
es=0.60
es=0.80

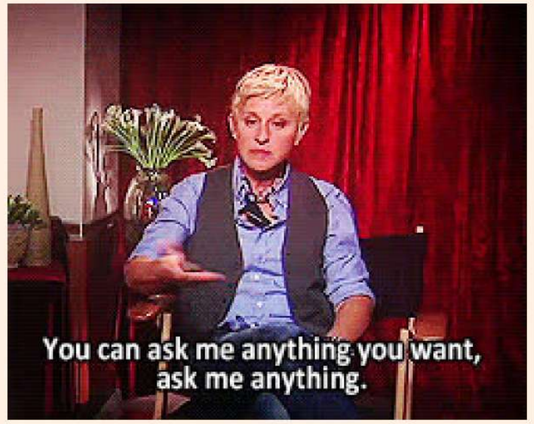
2


Significance Level vs. Statistical Power
1.0
0.8
0.6
0.4
0.2
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Effect Size
N=50.00
N=100.00
N=500.00

3


Significance Level vs. Statistical Power
1.0
0.9
0.8
0.7
0.6
0.5
0.00
0.05
0.10
0.15
0.20
0.25
0.30
alpha
N=50.00
N=100.00
N=500.00

Thank You!



17th December 2022

Non-paramteric Tests and Correlation Analysis

Authors:

1. Diego Gules Butori
2. Nairuhi Tovmasyan
3. Jaswinder Singh
4. Federico Andrés Gómez Quiroga
5. Cian O'Sullivan

Objectives

- The objective of this week's presentation is to introduce and conduct various **non-paramteric tests** for the lending club loan data.
- Besides that we will also perform a **correlation analysis** on the different columns in our data and build a **regression model** to predict the loan amount.

AGENDA

INTRODUCTION

- Choosing between parametric and non-paramteric tests
- Adavantages of parametric and non-paramteric tests

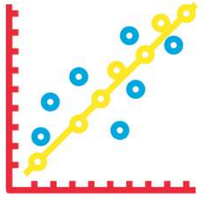
NON-PARAMETRIC TESTS

- Kruskal Wallis Test
- Mann Whitney Test

CORRELATION ANALYSIS

Studying relationships between the variables

REGRESSION



- Simple Linear Regression
- Multiple Regression

Chooosing between Parametric and Non-Parameteric test

Parameteric tests (means)

- 1-sample t test
- 2-sample t test
- One-Way ANOVA
- Two-way ANOVA

Pearson's Correlation

Non-Parameteric tests (medians)

- 1-sample Sign, 1-sample Wilcoxon
- Mann Whitney test
- Kruskal-Wallis, Mood's median test
- Friedman test

Spearman's Correlation

MORE ON THIS LATER!

Advantages of Paramteric and Non-parameteric Tests

Parameteric

- Parametric tests can provide trustworthy results with distributions that are skewed and nonnormal
- Parametric tests can provide trustworthy results when the groups have different amounts of variability
- Parametric tests have greater statistical power

MORE INFO ON NEXT SLIDE!

Non-Parametric

- Nonparametric tests assess the median which can be better for some study areas
- Nonparametric tests are valid when our sample size is small and your data is potentially nonnormal
- Nonparametric tests can analyze ordinal data, ranked data, and outliers

Sample Size Requirements for Paramteric Tests

Parametric analyses	Sample size guidelines for nonnormal data
1-sample t test	Greater than 20
2-sample t test	Each group should be greater than 15
One-Way ANOVA	• For 2-9 groups, each group should have more than 15 obs. • For 10-12 groups, each group should have more than 20 observations

p-value

NON-PARAMETRIC TESTS

Statistics

- Kruskal-Wallis
- Mann Whitney

results

box-plots

Kruskal Wallis Test

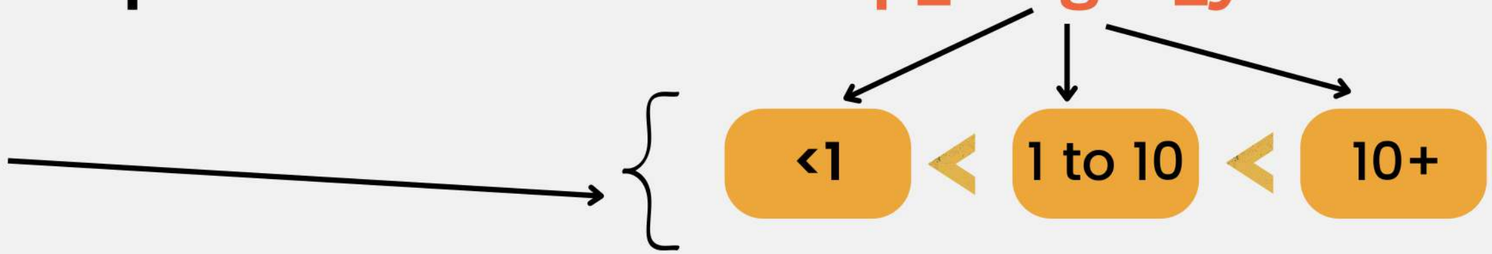
- Kruskal Wallis H test is a nonparametric test that is used to determine the statistical differences between the two or more groups of an independent variable.
- **Requirement:**
  - Data should be **ordinal**!
  - One independent variable with two or more levels (independent groups)

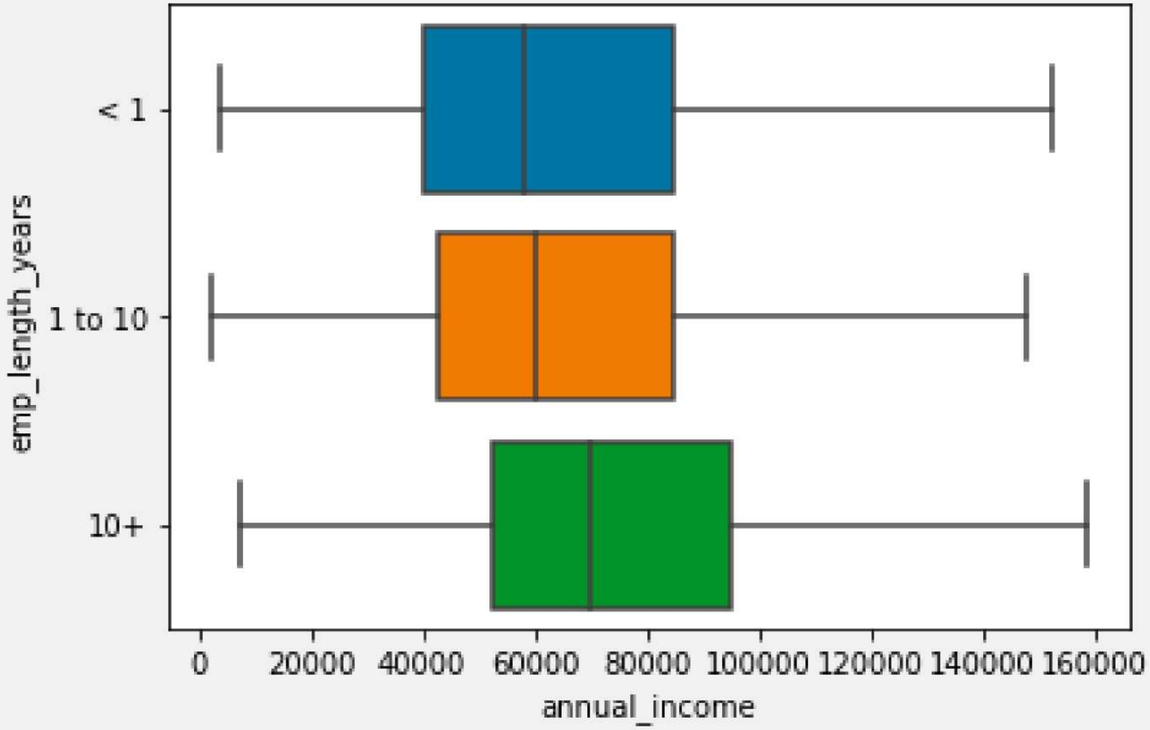
Null Hypothesis: The independent samples all have the **same central tendency** and therefore come from the same population. In other words, there is **no difference in the rank sums**.

Alternative Hypothesis: At least one of the independent samples **does not have the same central tendency** as the other samples and therefore originates from a different population, or in other words, **at least one group differs in rank sums**.

Dependent Variable: annual_income

Independent Variable: emp_length_years






```
stats.kruskal(df[df['emp_length_years']=='< 1']['annual_income'],
              df[df['emp_length_years']=='1 to 10']['annual_income'],
              df[df['emp_length_years']=='10+']['annual_income'])
```

```
KruskalResult(statistic=9494.726632514885, pvalue=0.0)
```

Conclusion: Since the p-value is less than 0.05, we can **reject the NULL hypothesis** that the groups have same central tendency (median). Therefore we have to accept the alternative hypothesis. Hence we can conclude that the **employment length has an influence on the annual_income.**

Mann Whitney Test

- When we want to check if there is any difference between two independent groups/samples, the Mann-Whitney test is applied.
- The Mann-Whitney test checks the existence of **difference in the rank sum**, whenever we have two independent samples.

Null Hypothesis: The **sum of the rankings in the two groups does not differ** in the population.

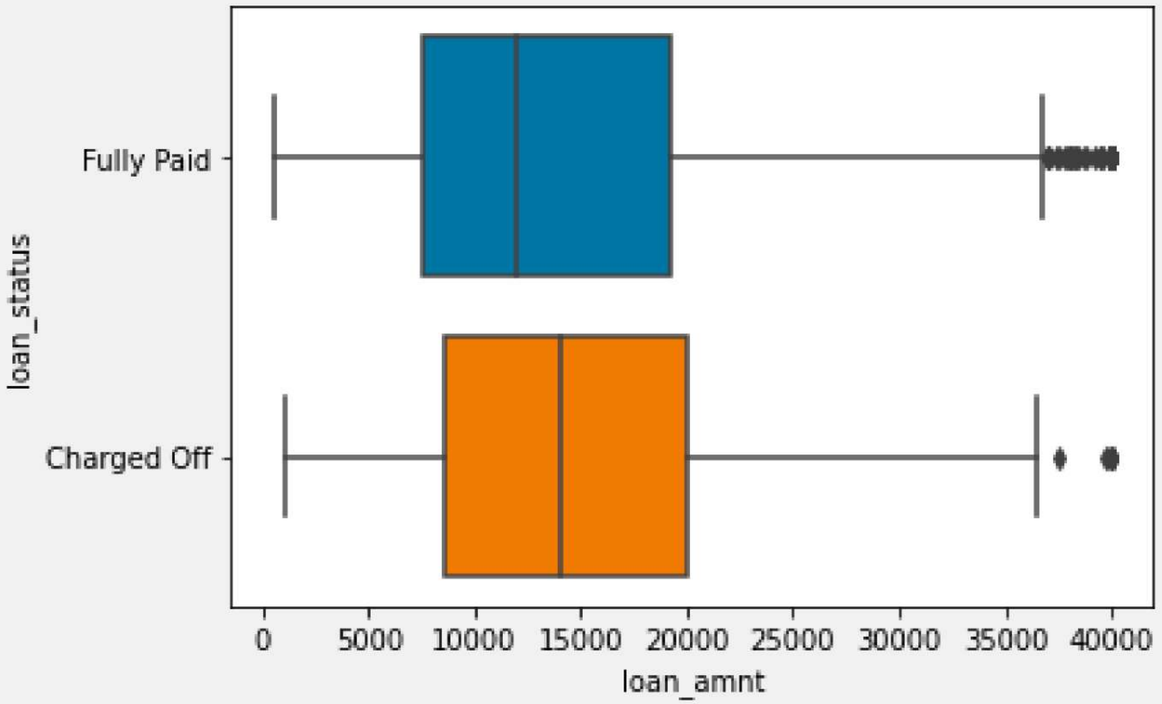
Alternative Hypothesis: The **sum of the rankings differs in the two groups** in the population.

Dependent Variable: loan_amnt

Independent Variable: loan_status

Fully paid

Charged Off




```
mannwhitneyu(df[df['loan_status']=='Fully Paid']['loan_amnt'],
             df[df['loan_status']=='Charged Off']['loan_amnt'])
```

```
MannwhitneyuResult(statistic=11235674768.0, pvalue=0.0)
```

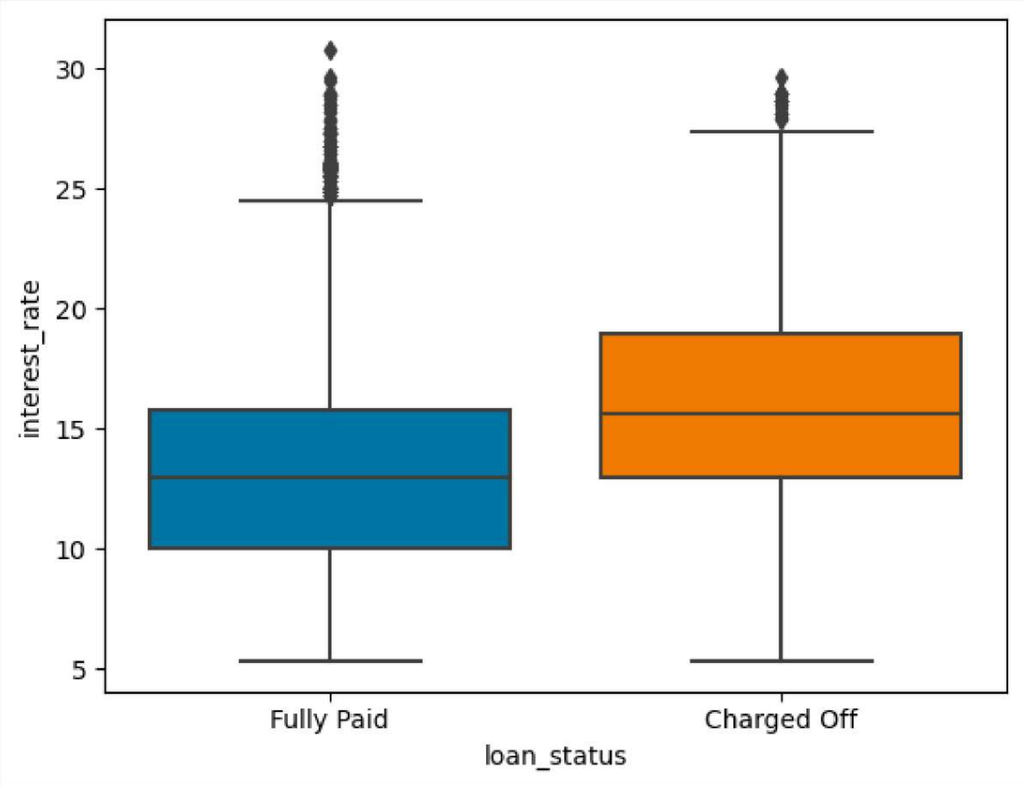
Conclusion: Since the p-value is less than O.O5, we can **reject the NULL hypothesis** that the sum of rankings in the two groups differ. Therefore we have to accept the alternative hypothesis. Hence we can conclude that the **loan_status has an influence on the loan amount**.

Dependent Variable: interest_rate

Independent Variable: loan_status

Fully paid

Charged Off




```python
# Mann Whitney
# Does the type of interest_rate impacts the amount of loan?

alpha = 0.05
stat, p_value = stats.mannwhitneyu(df[df['loan_status']=='Fully Paid']['interest_rate'].sample(n=len(df[df['loan_status']=='Charged Off'])),
                                   df[df['loan_status']=='Charged Off']['interest_rate'])

# Check the p-value against the significance level
if p_value < alpha:
    print("The samples are significantly different (p = {})".format(p_value))
else:
    print("The samples are not significantly different (p = {})".format(p_value))
```

✓ 0.6s

```
The samples are significantly different (p = 2.4300133568734786e-286)
```

Conclusion: Since the p-value is less than O.O5, we can **reject the NULL hypothesis** that the sum of rankings in the two groups differ. Therefore we have to accept the alternative hypothesis. Hence we can conclude that **loan_status has an influence on the interest rate**.

CORRELATION ANALYSIS

Spearman Correlation Coefficient

Correlation

In statistics, correlation is any statistical relationship, whether causal or not, between two random variables or bivariate data.

Correlation analysis is a statistical technique that shows how variables are related.

Correlation coefficients range between -1 and +1

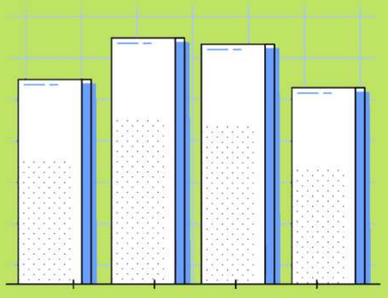
The correlation between two variables can be either:

- **Positive:** If variable X **increases**, theres a tendencie for Y to **increase**

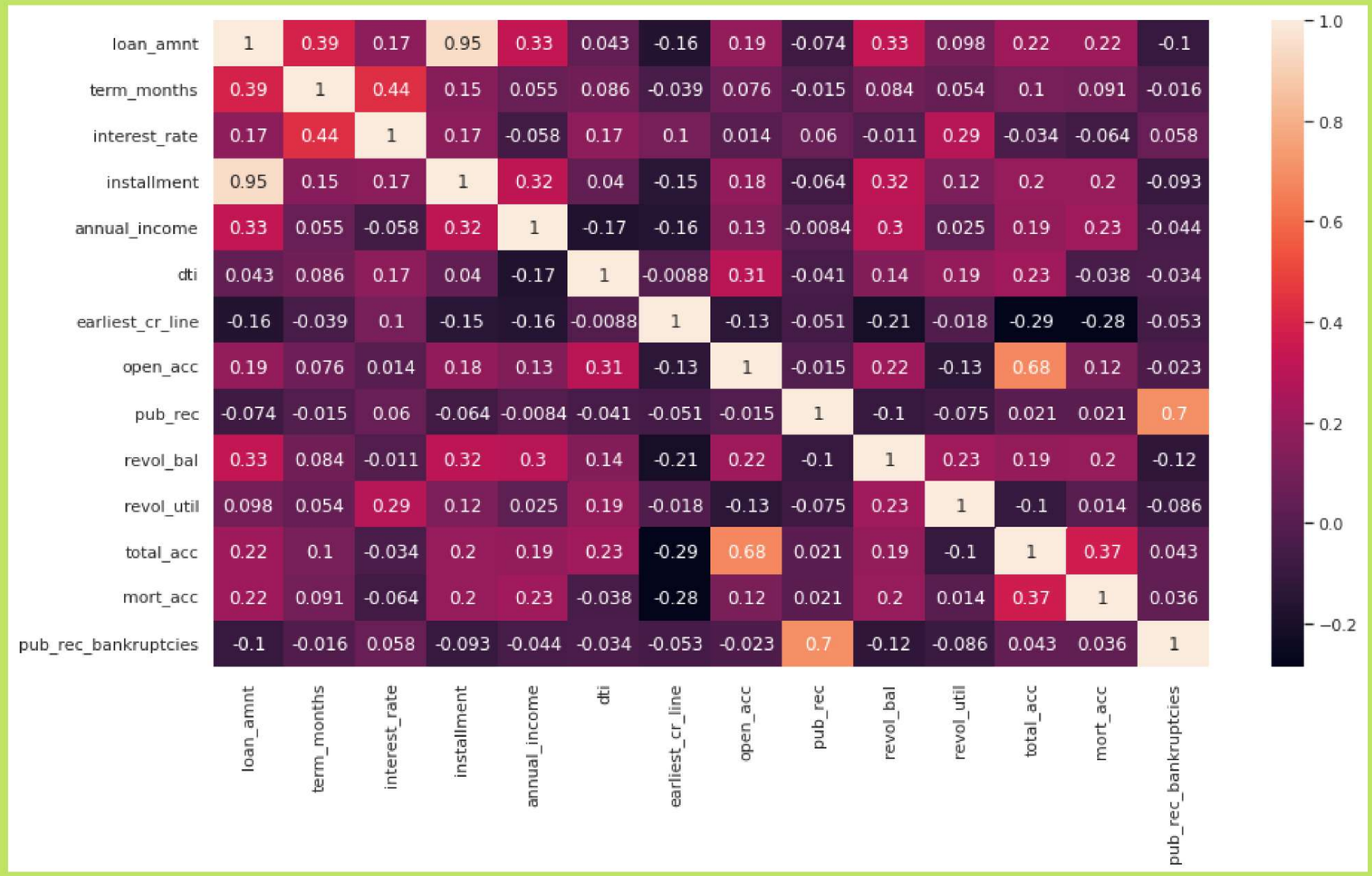
- **Negative:** If variable X **increases**, theres a tendencie for Y to **decrease**

Correlation between columns in the dataset

LendingClub



Correlation Heat Map



p-value

REGRESSION ANALYSIS

Independent Variables

results

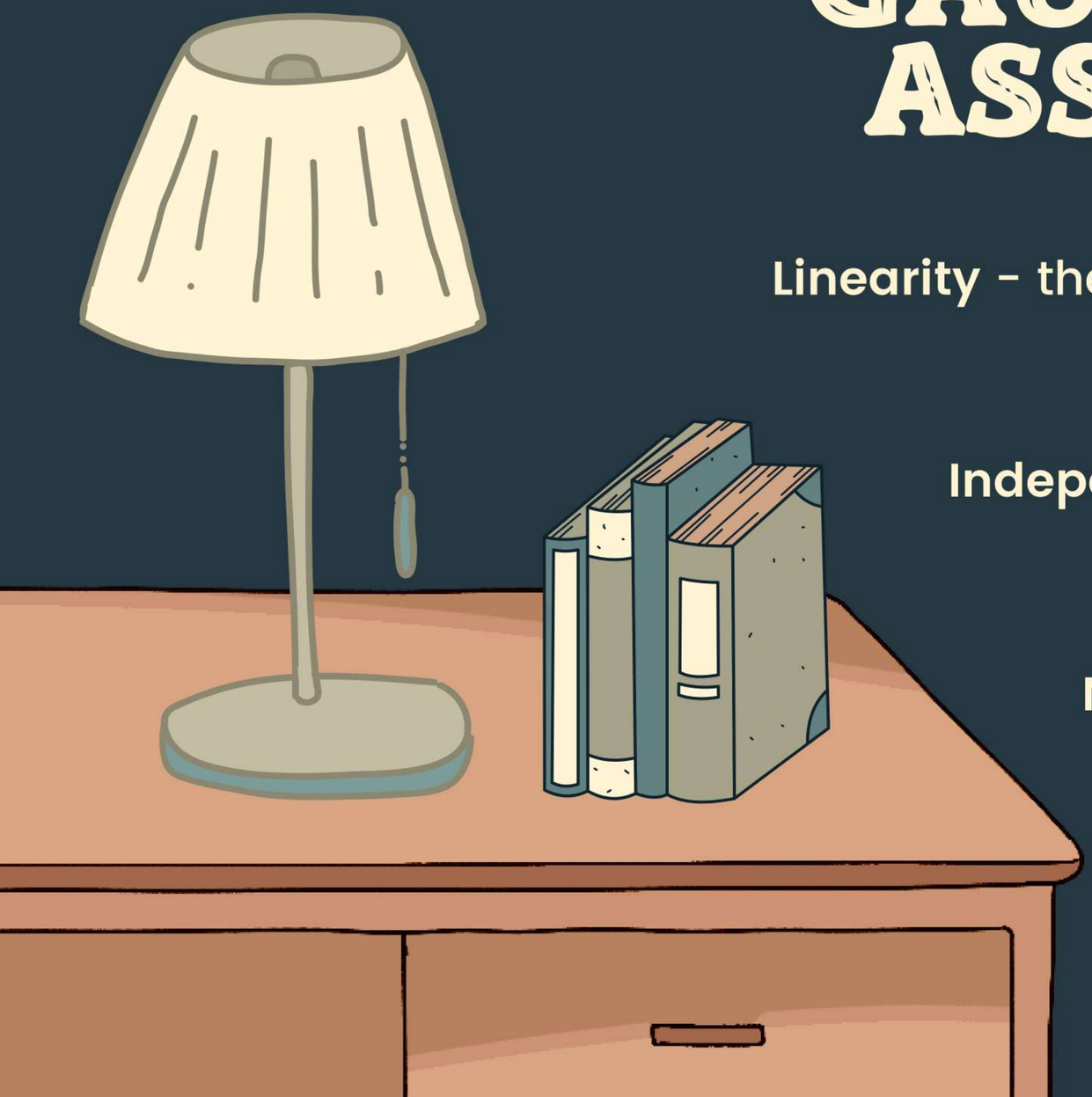
- Simple Linear Regression
- Multiple Regression

Gauss-Markov Assumptions

LINEAR REGRESSION

Regression is a tool for finding existence of an association, relationship between a dependent variable (Y) and one or more independent variables (X1, X2, ..., Xn) in a study. The relationship can be linear or non-linear. Regression is not designed to capture causality. Linear regression is a linear approach to modeling the relationship between a dependent variable and covariates.

GAUSS-MARKOV ASSUMPTIONS



Linearity - there is a linear relationship between dependent and independent variables

Independence - each row is an independent observation

Normality - errors are normally distributed

Equal variance - the phenomena of having constant variance over different values of the independent variable

PREDICT LOAN AMOUNT WITH INSTALLMENT

OLS Regression Results

Dep. Variable:	loan_amnt	**R-squared:**	0.910
Model:	OLS	**Adj. R-squared:**	0.910
Method:	Least Squares	**F-statistic:**	3.009e+06
Date:	Fri, 16 Dec 2022	**Prob (F-statistic):**	0.00
Time:	19:21:23	**Log-Likelihood:**	-2.7459e+06
No. Observations:	297020	**AIC:**	5.492e+06
Df Residuals:	297018	**BIC:**	5.492e+06
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	378.5107	9.152	41.357	0.000	360.573	396.449
installment	31.8048	0.018	1734.775	0.000	31.769	31.841

Omnibus:	76006.890	**Durbin-Watson:**	1.996
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	202212.446
Skew:	1.380	**Prob(JB):**	0.00
Kurtosis:	5.954	**Cond. No.**	994.

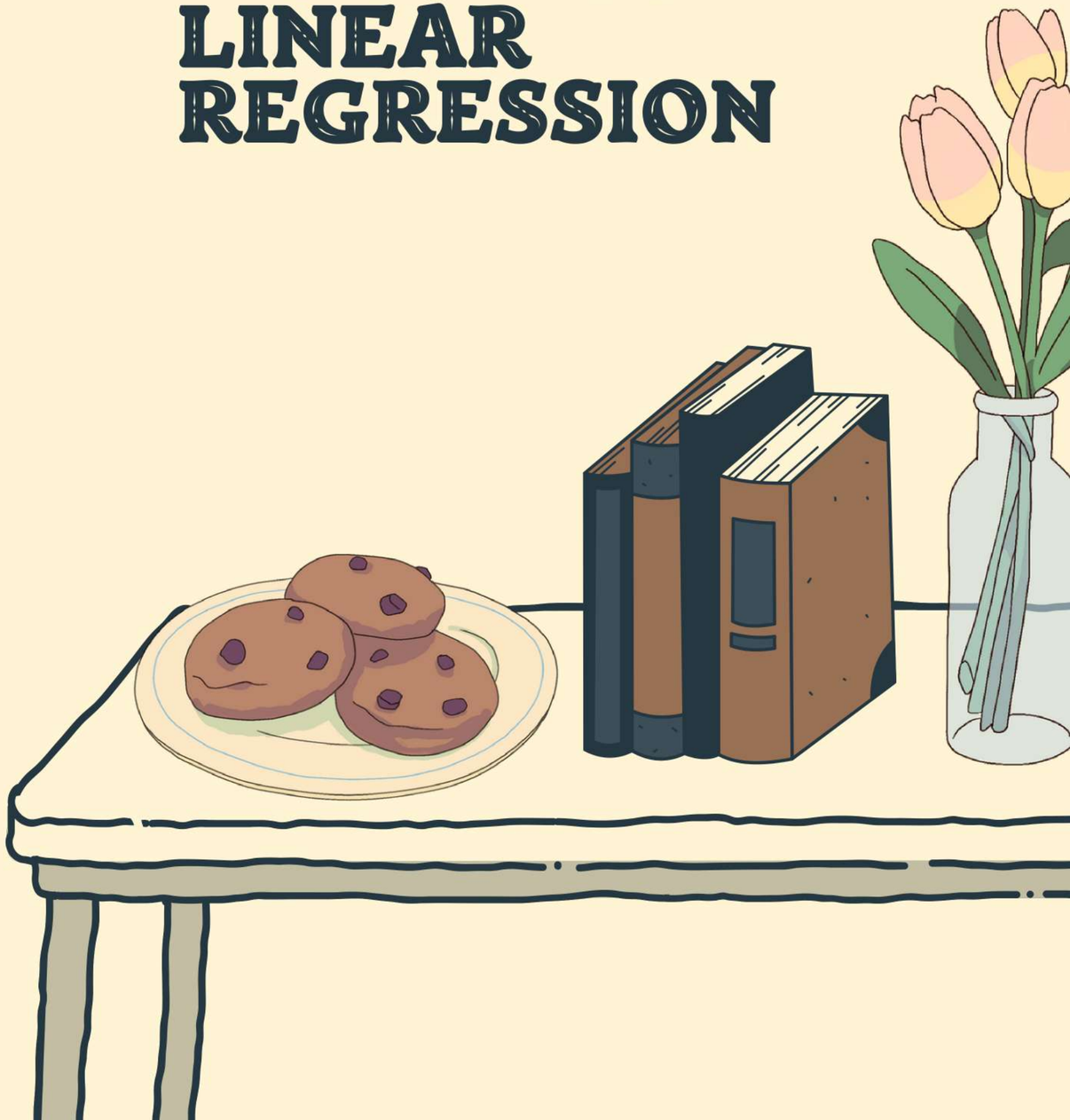
OLS Regression Results

Dep. Variable:	loan_amnt	**R-squared:**	0.983
Model:	OLS	**Adj. R-squared:**	0.983
Method:	Least Squares	**F-statistic:**	1.011e+06
Date:	Fri, 16 Dec 2022	**Prob (F-statistic):**	0.00
Time:	19:29:10	**Log-Likelihood:**	-3.3286e+06
No. Observations:	396027	**AIC:**	6.657e+06
Df Residuals:	396003	**BIC:**	6.658e+06
Df Model:	23		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3052.4611	136.071	22.433	0.000	2785.767	3319.156
interest_rate	-200.8979	1.318	-152.450	0.000	-203.481	-198.315
installment	30.7133	0.008	3831.933	0.000	30.698	30.729
annual_income	0.0007	3.18e-05	23.325	0.000	0.001	0.001
dti	-1.0980	0.241	-4.564	0.000	-1.569	-0.626
earliest_cr_line	-0.0076	0.001	-10.796	0.000	-0.009	-0.006
open_acc	-2.3358	0.379	-6.171	0.000	-3.078	-1.594
pub_rec	-49.9596	3.324	-15.032	0.000	-56.474	-43.445
revol_bal	0.0033	9.66e-05	34.042	0.000	0.003	0.003
revol_util	-0.9259	0.081	-11.474	0.000	-1.084	-0.768
mort_acc	11.6529	0.910	12.805	0.000	9.869	13.437
term_months_36	-1422.8073	68.074	-20.901	0.000	-1556.230	-1289.385
term_months_60	4475.2684	68.079	65.736	0.000	4341.835	4608.702
grade/rank_A	518.6426	22.995	22.555	0.000	473.574	563.711
grade/rank_B	658.5475	20.872	31.552	0.000	617.639	699.456
grade/rank_C	727.5503	20.019	36.343	0.000	688.313	766.787
grade/rank_D	713.8159	20.091	35.530	0.000	674.439	753.193
grade/rank_E	551.5314	21.075	26.169	0.000	510.224	592.839
grade/rank_F	210.6564	23.638	8.912	0.000	164.327	256.986
grade/rank_G	-328.2829	28.858	-11.376	0.000	-384.843	-271.723
emp_length_years_< 1	1037.7616	45.600	22.758	0.000	948.387	1127.136
emp_length_years_1 to 10	1007.9072	45.431	22.185	0.000	918.864	1096.951
emp_length_years_10+	1006.7924	45.424	22.164	0.000	917.763	1095.821
verification_status_Not Verified	1005.8015	45.430	22.140	0.000	916.760	1094.843
verification_status_Source Verified	975.2290	45.418	21.472	0.000	886.211	1064.247
verification_status_Verified	1071.4306	45.438	23.580	0.000	982.373	1160.488
initial_list_status_f	-1868.6262	312.234	-5.985	0.000	-2480.596	-1256.656
initial_list_status_w	-1823.4539	312.238	-5.840	0.000	-2435.431	-1211.476

Omnibus:	176625.105	**Durbin-Watson:**	2.007
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	4786766.581
Skew:	1.574	**Prob(JB):**	0.00
Kurtosis:	19.738	**Cond. No.**	3.52e+16

MULTIPLE LINEAR REGRESSION



THANK YOU FOR LISTENING!

Don't hesitate to ask any questions!