IDS Project

Analysis of Loans

Objective:

To clean and analyse the loan.csv dataset for understanding the various aspects of loan lending.

Overview:

Our dataset contains the details of all the loans disbursed in a 7 year period from 2011-2018 by one of the world's top lending companies.

Apart from the traditional fields such as amount lended, rate of lending, period of loan, our dataset also contains additional informative and insightful fields such as annual income of borrower, mode of ownership of current residence, details of employment,etc

Steps Done

- Reducing Dataset Size
- Checking for missing values
- Filling numerical missing values with mean
- Filling Categorical missing values using ffill
- Normalizing the Data
- Plotting Graphs
- Drawing Inferences
- Finding Correlations
- Hypothesis Testing

Filling missing values

Before and after cleaning

loan_amnt	0
funded_amnt	0
funded_amnt_inv	0
term	0
int_rate	0
installment	0
grade	484
sub_grade	484
emp_title	902
emp_length	459
home_ownership	538
annual_inc	0
verification_status	432
issue_d	0
loan_status	0
purpose	0
title	0
zip_code	0
addr_state	0

dtype: int64

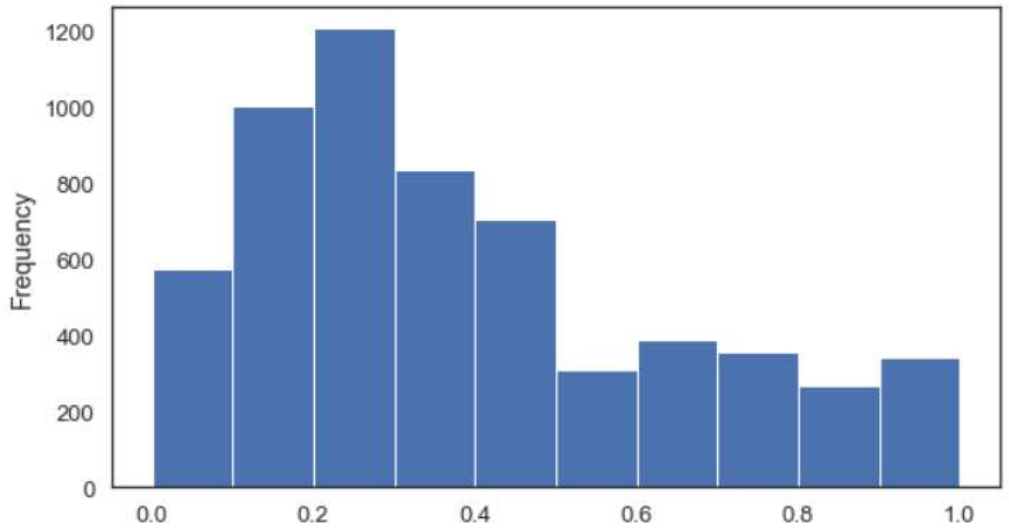
loan_amnt	0
funded_amnt	0
funded_amnt_inv	0
term	0
int_rate	0
installment	0
grade	0
sub_grade	0
emp_title	0
emp_length	0
home_ownership	0
annual_inc	0
verification_status	0
issue_d	0
loan_status	0
purpose	0
title	0
zip_code	0
addr_state	0

dtype: int64

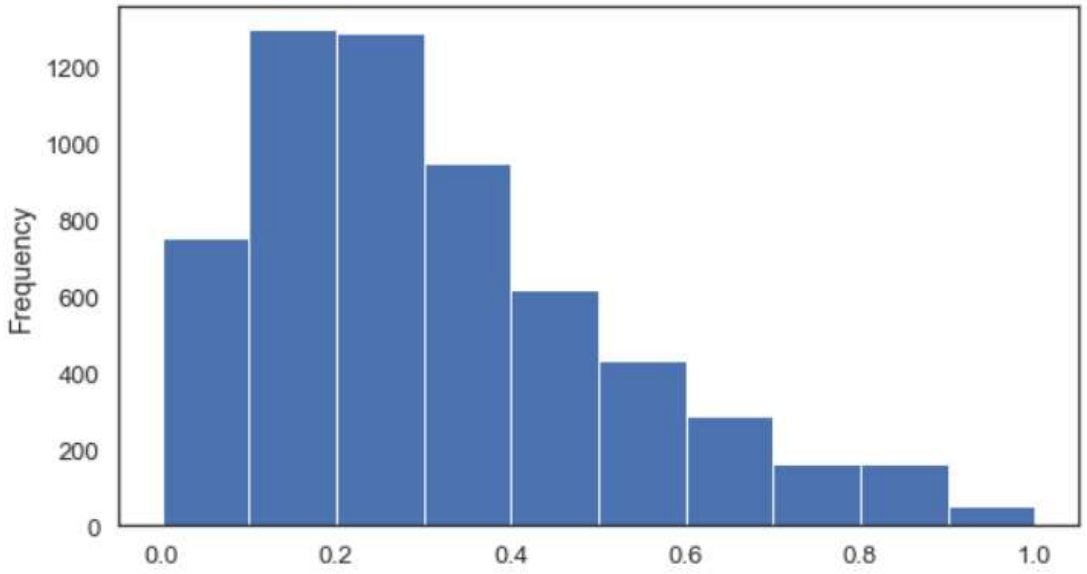
Importance of Normalization

- Dataset can be utilized much better when normalized
- Can detect duplicate values in dataset
- Data Analysis becomes easier (eg: Finding relations)
- Changing or updating data becomes easier.

Normalization Graph

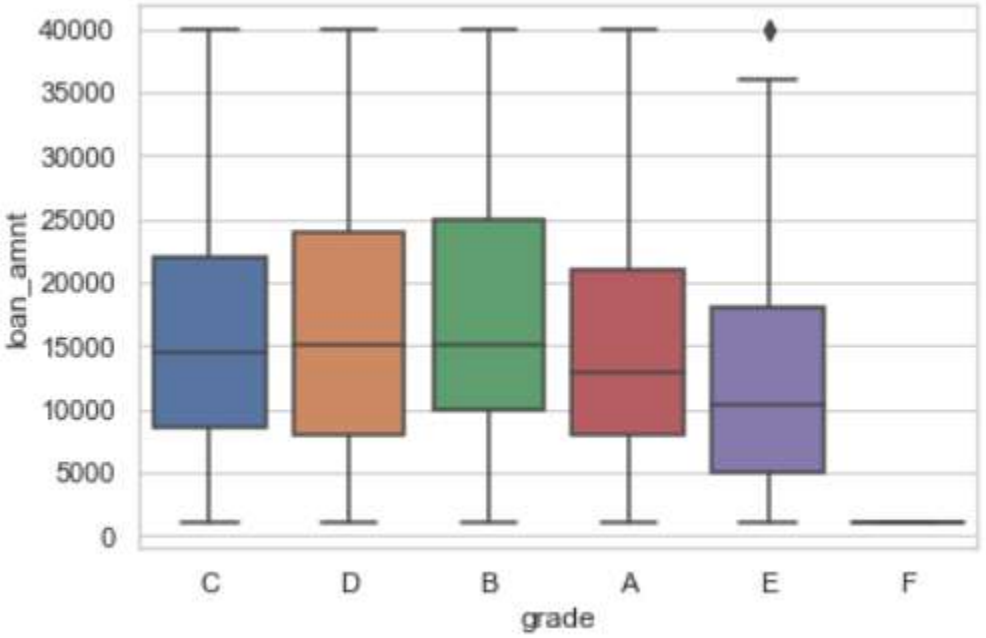


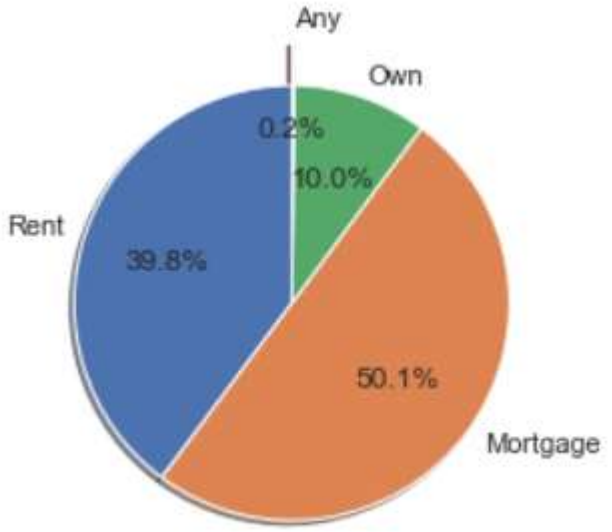
Loan Amount



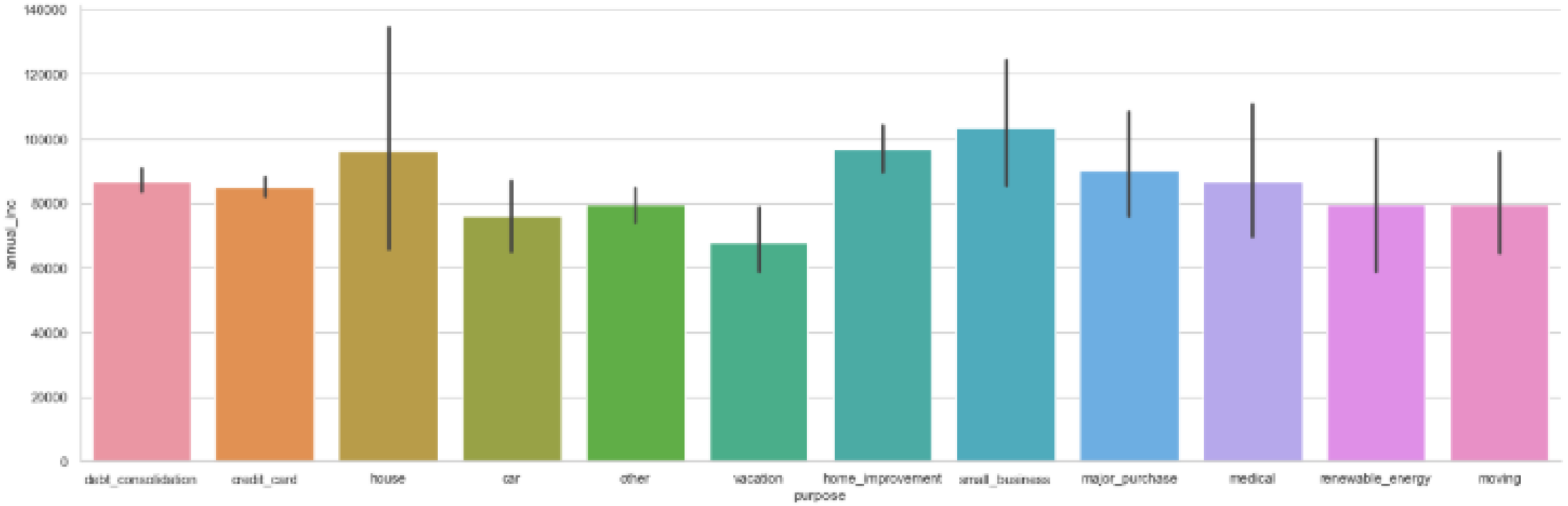
Installment

Loan Amount vs Grade & Loan Categories

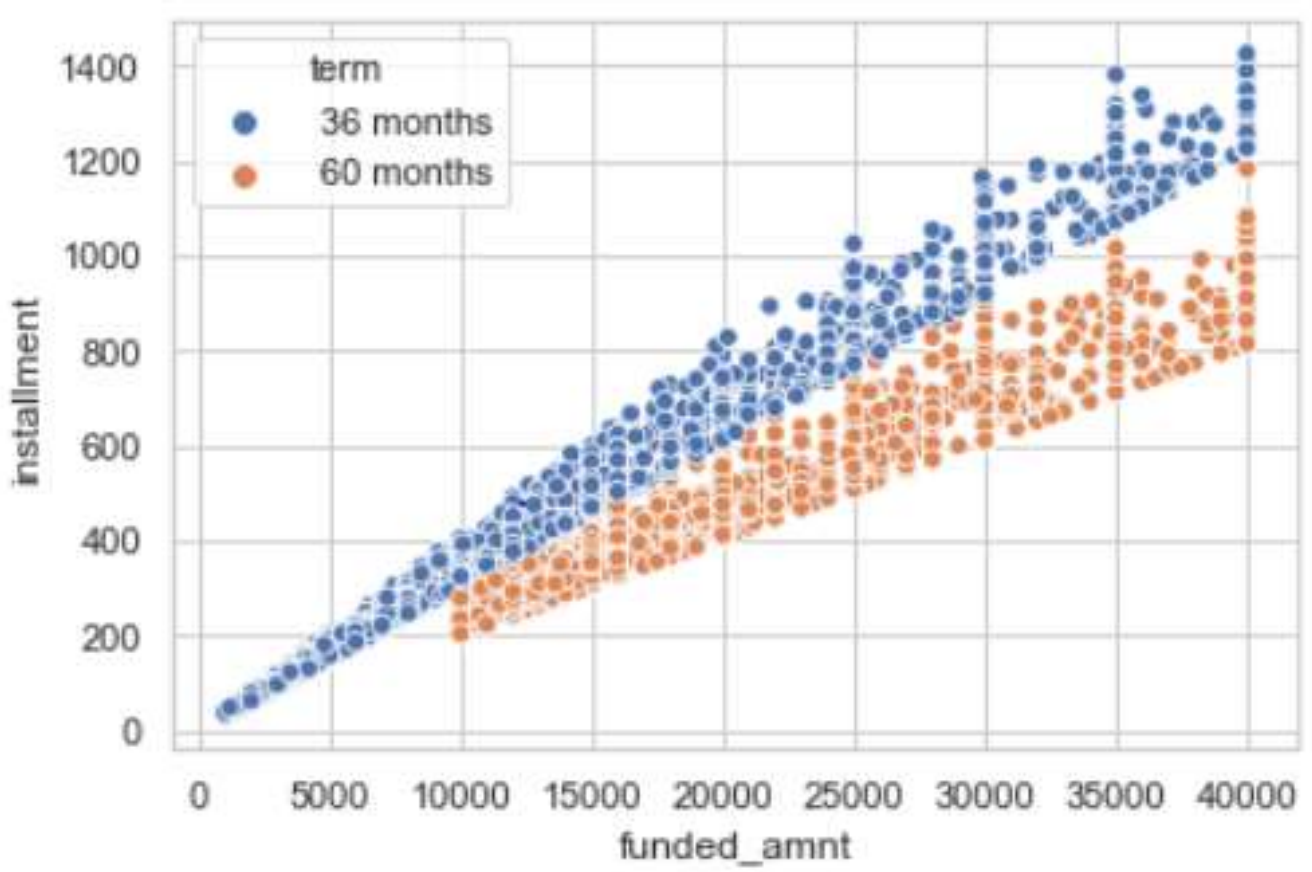




Purpose VS Annual Income



Funded Amount vs Installments



Inferences from Graphs

- Annual Income is highest for people applying for small business among all people applying for loans
- Bank does not lend F Grade Citizens
- Bank Lends highest amount for B Grade Citizens
- F4,F5 – Loans lent for people belonging to these categories is zero

Inferences (continued)

- In general, it is found that number of people who were not verified were more than those verified
- Mortage Loans constitute 50% of all loans

Correlations

Loan Amount VS Employment Length

	n	r	CI95%	r2	adj_r2	p-val	BF10	power
pearson	5999	0.043	[0.02, 0.07]	0.002	0.002	0.000876	4.1	0.914

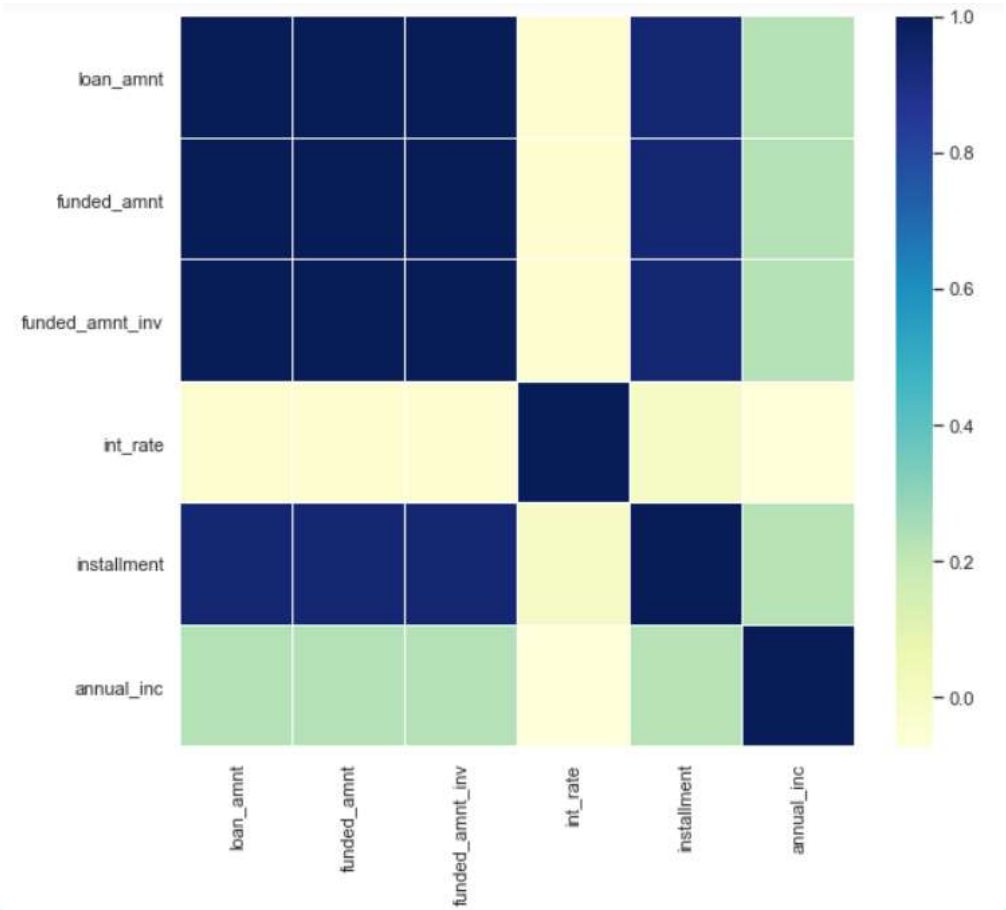
Interest Rates VS Grades

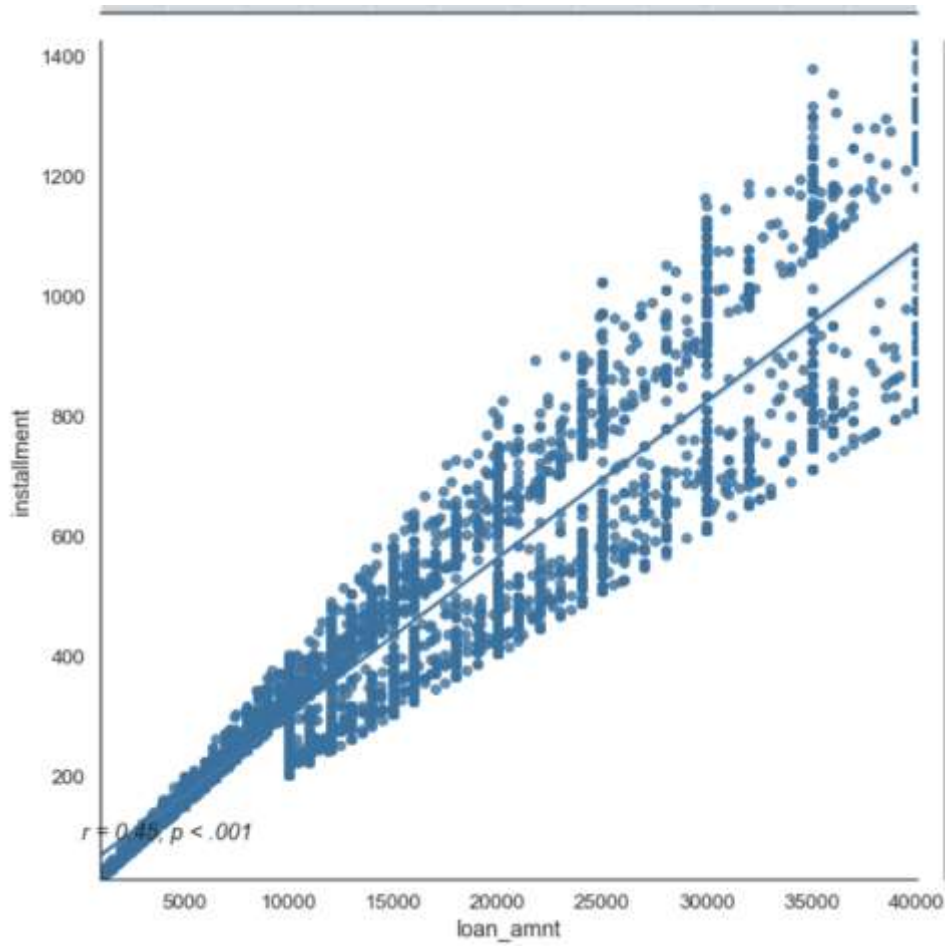
	n	r	CI95%	r2	adj_r2	p-val	BF10	power
pearson	5999	0.901	[0.9, 0.91]	0.812	0.812	0.0	inf	1.0

Loan Amount VS Installment

	n	r	CI95%	r2	adj_r2	p-val	BF10	power
pearson	5999	0.939	[0.94, 0.94]	0.882	0.882	0.0	inf	1.0

Plots related to correlations





Correlations

- No correlation between grades and employment lengths
- No correlation between employment lengths and interest rates
- No correlation between funded amount and annual income
- High correlation between loan amount and instalment
- High correlation between interest rates and grades

Hypothesis Testing

- 2 hypothesis tests were performed
- The first one is 2 tailed hypothesis testing for the population mean of the loan amount
- The second one is a 1 tailed hypothesis testing for the population mean of the annual incomes
- Alpha was taken to be 0.05

TWO-TAILED HYPOTHESIS TESTING

- Firstly, the population mean for the categorical column loan amount was described and from that a random sample of 1000 values was selected.
- The sample mean was tested with respect to the population mean and then the hypothesis test was conducted to check whether it is rejected or not.

```
H0 : μ = 16304.434
H1 : μ != 16304.434
alpha value is : 0.05

actual z value : 1.9599639845400545
hypothesis z value : 0.7061235995972134

Failed to reject NULL hypothesis
```

ONE-TAILED HYPOTHESIS TESTING

- The population mean for the categorical column annual income was described and from that a random sample of 1000 values was selected.
- The sample mean was tested with respect to the population mean and then the hypothesis test was conducted to check whether it is rejected or not.

```
H0 : μ <= 86407.41
H1 : μ > 86407.41
alpha value is : 0.05

actual z value : 1.6448536269514729
hypothesis z value : -0.11797554849777012

Failed to reject NULL hypothesis
```