Case Study #1

Below is a data set that represents thousands of loans made through the Lending Club platform, which is a platform that allows individuals to lend to other individuals.

We would like you to perform the following using the language of your choice:

- Describe the dataset and any issues with it.
- Generate a minimum of 5 unique visualizations using the data and write a brief description of your observations. Additionally, all attempts should be made to make the visualizations visually appealing
- Create a feature set and create a model which predicts *interest rate* using at least 2 algorithms. Describe any data cleansing that must be performed and analysis when examining the data.
- Visualize the test results and propose enhancements to the model, what would you do if you had more time. Also describe assumptions you made and your approach.

Dataset

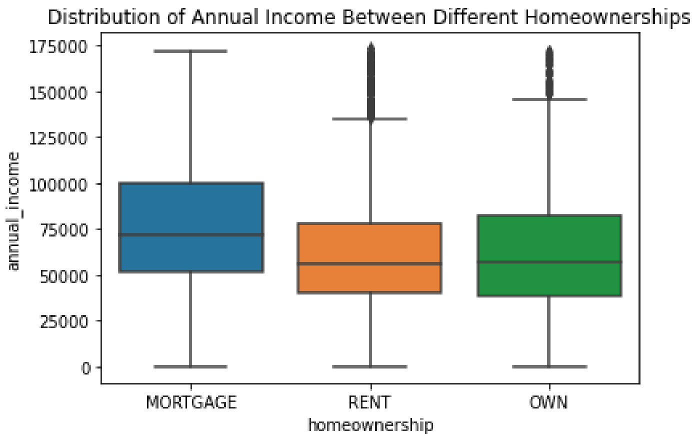
https://www.openintro.org/data/index.php?data=loans_full_schema

Output

An HTML website hosting all visualizations and documenting all visualizations and descriptions. All code hosted on GitHub for viewing. Please provide URL's to both the output and the GitHub repo.

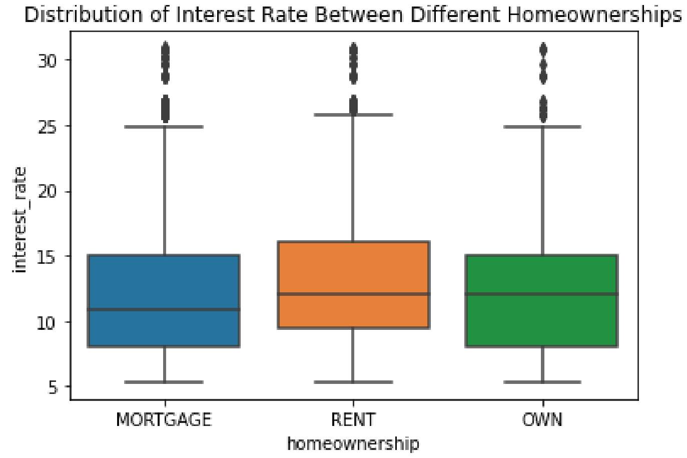
** If you submit a jupyter notebook, also submit the accompanying python file. You may use python(.py), R, and RMD (knit to HTML) files. Other languages are acceptable as well.*

Plot 1 The Distribution of Annual Income Between Different Homeownerships



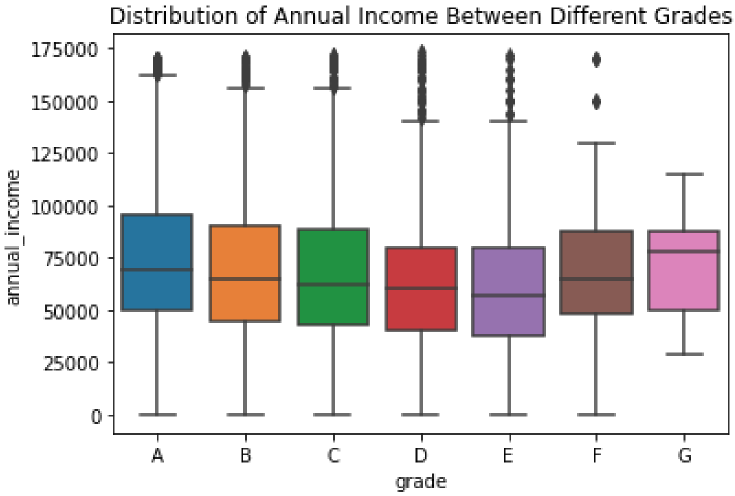
As we see from the boxplot, people whose homeownership type is mortgage have an overall higher annual income compared to the other 2 types of homeownership.

Plot 2 The Distribution of Interest Rate Between Different Homeownerships



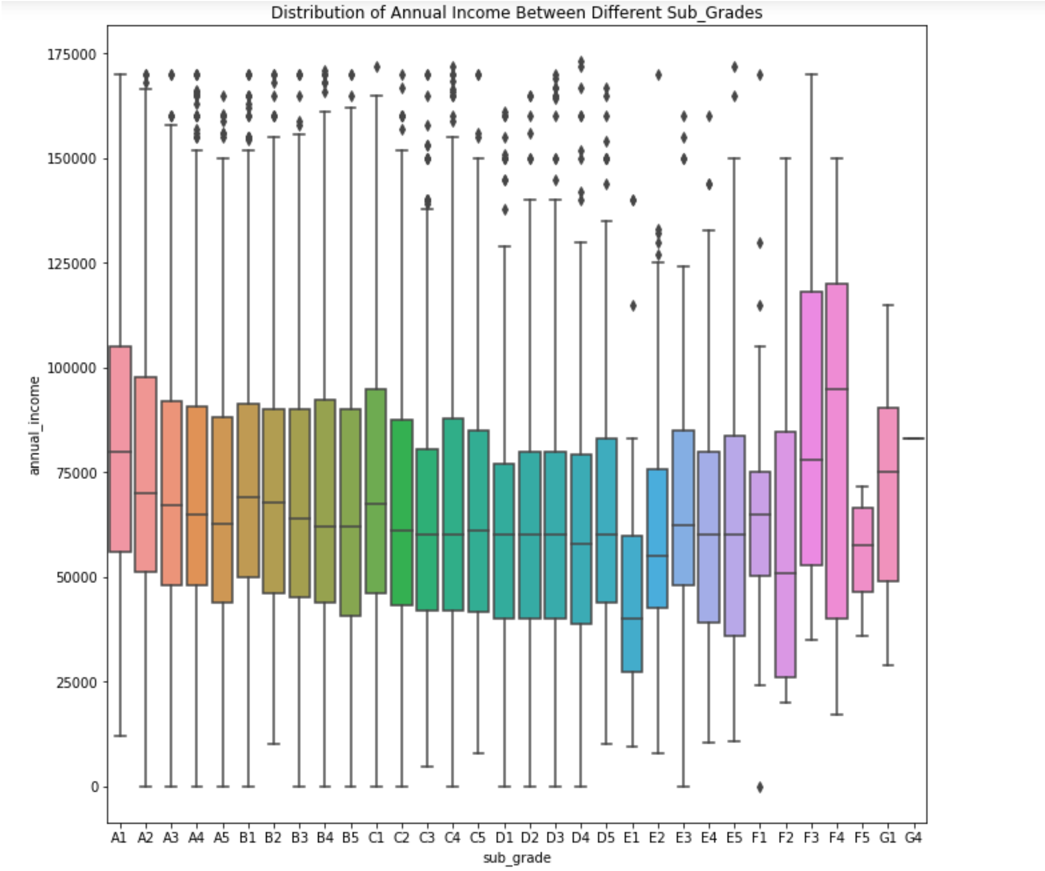
As the plot shows, mortgage has an overall relatively lower interest rate than other 2 types of homeownerships.

Plot 3 We also want to evaluate the annual income between different grades



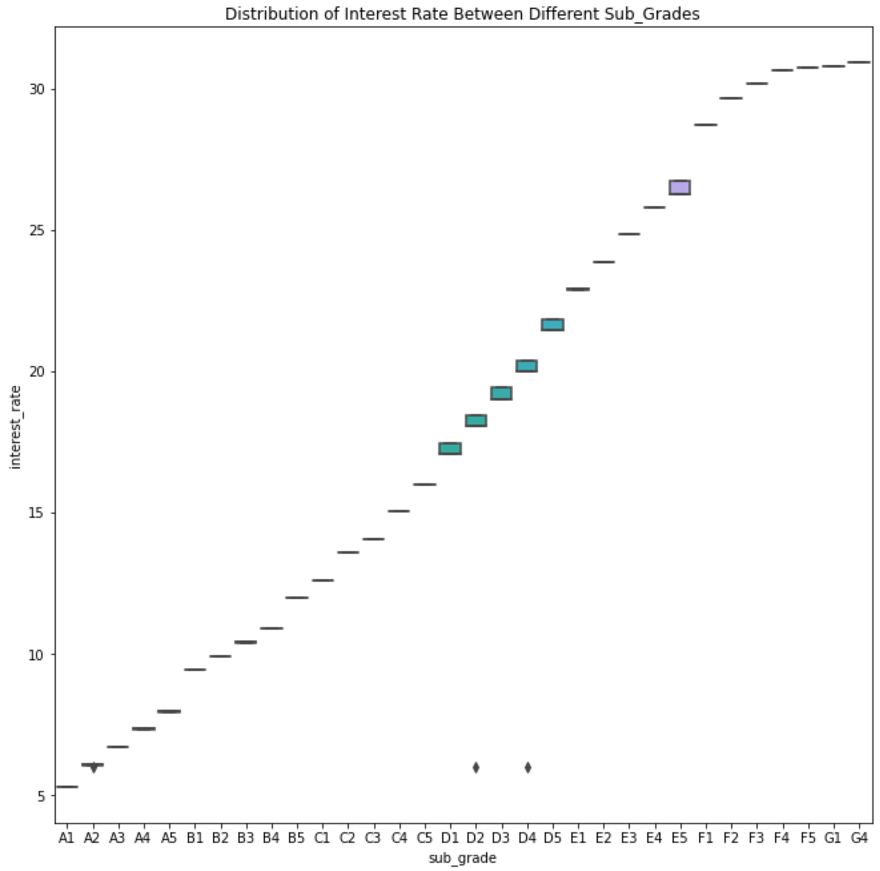
The plot shows that for people between grade A to E, the overall annual income decreases steadily. However, for grade F and G, people have higher annual income than other grades.

Plot 4 The Distribution of Annual Income Between Different Sub_grades



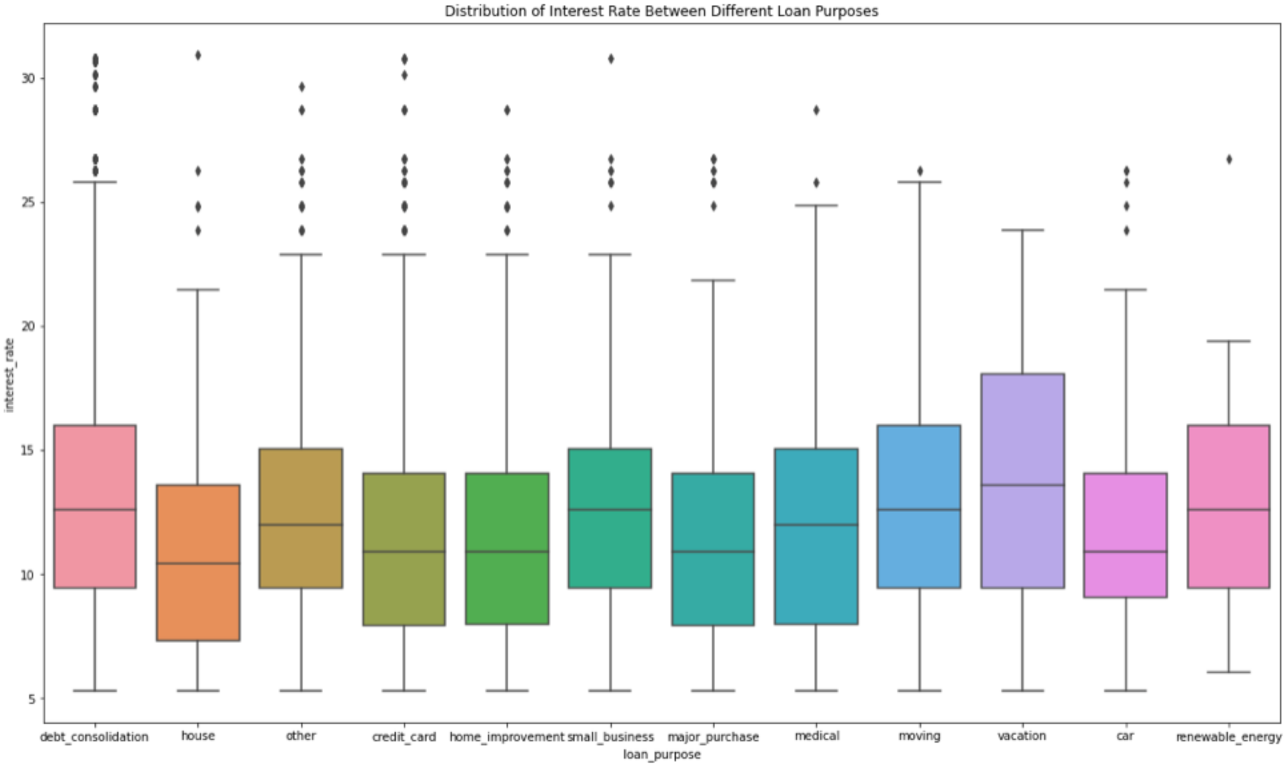
As we see, the annual income distribution pattern basically follows the one in plot 3.

Plot 5 Shows the interest rate for people in different sub grades



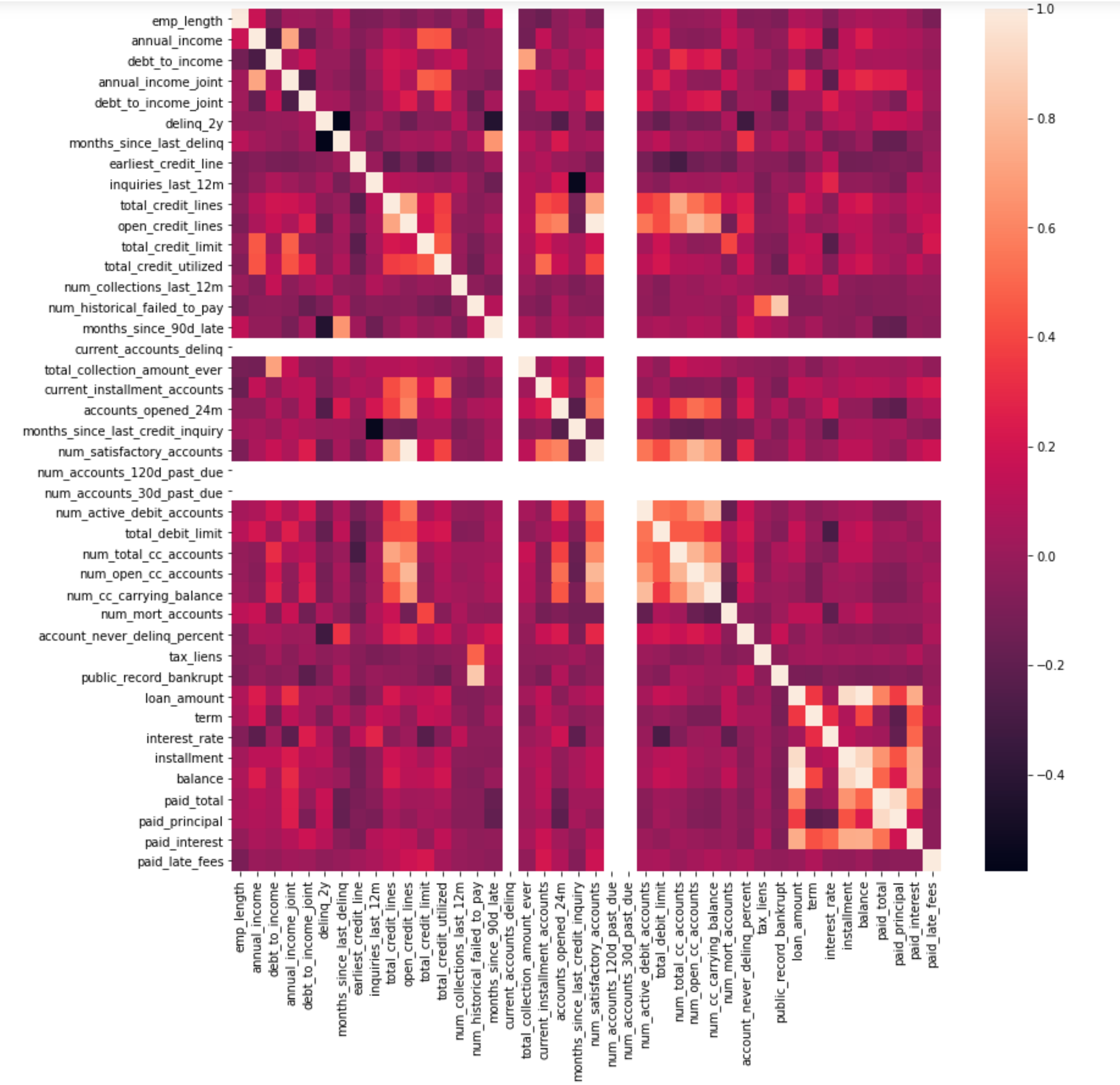
As we see, the lower grade a person holds, the higher interest rates he/she needs to pay.

Plot 6 We can also inspect the interest rates between different loan purposes



It shows the loan purpose for vacation, small business, renewable energy, moving and debt consolidation would charge for higher interest rates

7 Using heatmap to explore the correlationships between variables



From this plot, when we are focusing on interest rate, we can initially knowing interest rate is negatively related with annual income, annual income joint, total credit limit, months since last credit inquiry, total debit limit, number of mortgage accounts and account never delinquent percent. And it is positive related with paid interest and term.

Inspect the correlation between interest rate and other variables

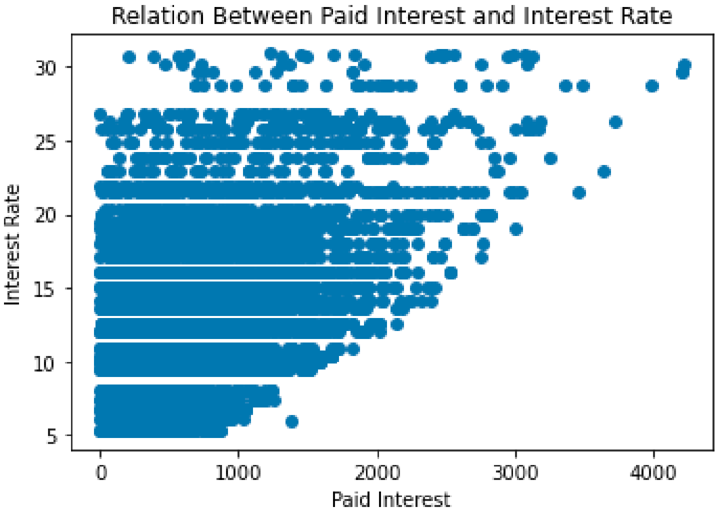
```
interest_rate                    1.000000
paid_interest                    0.522652
term                             0.355937
debt_to_income_joint             0.260277
debt_to_income                   0.136421
inquiries_last_12m               0.133352
installment                      0.124691
```

```
accounts_opened_24m                    0.123408
earliest_credit_line                   0.096935
delinq_2y                              0.089288
balance                                0.088443
loan_amount                            0.086502
num_cc_carrying_balance                0.084596

paid_total                             0.073469

total_credit_utilized                  0.056317
current_installment_accounts           0.050778
public_record_bankrupt                 0.044632
num_historical_failed_to_pay           0.035781
num_collections_last_12m               0.032151
```

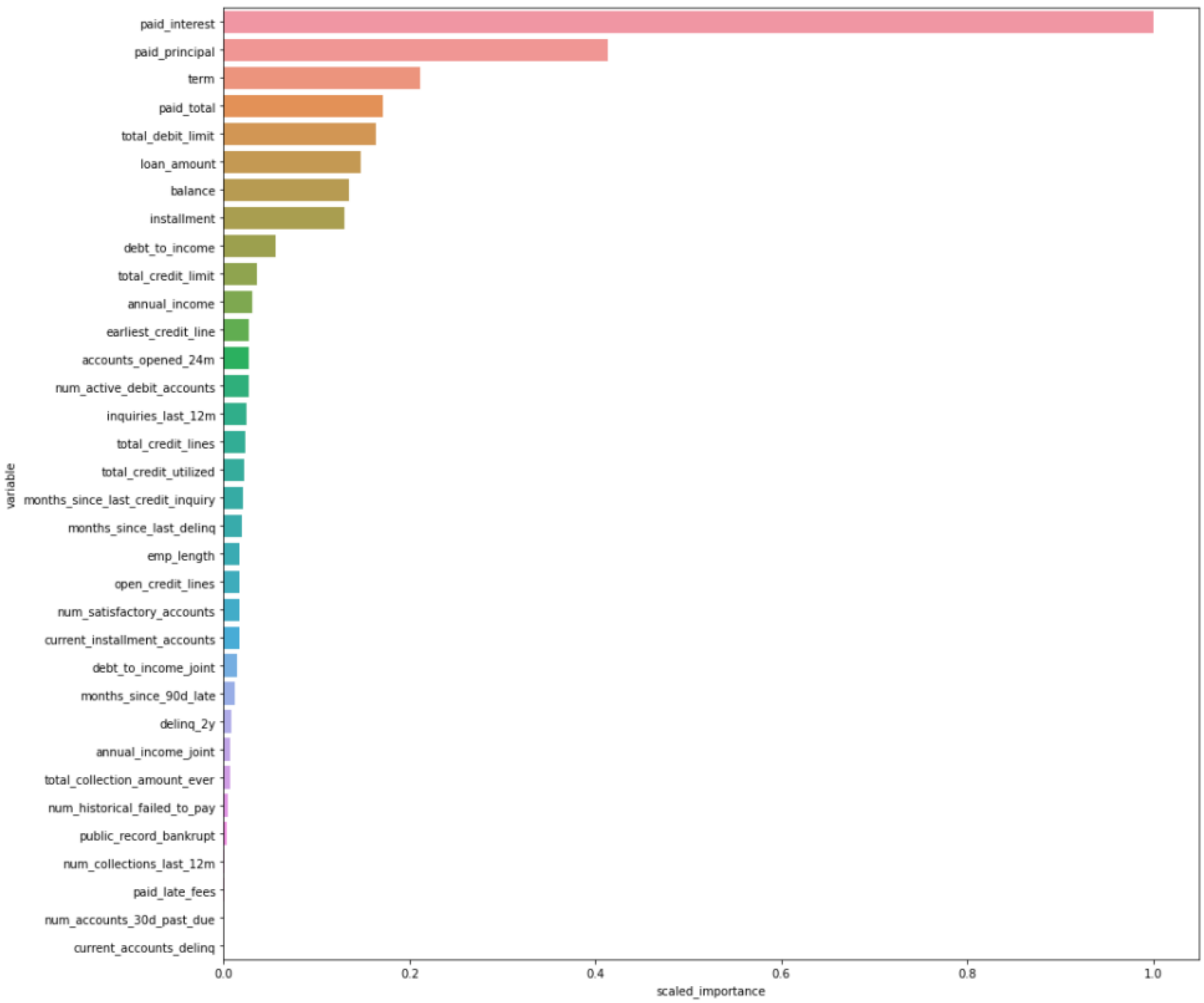
Plot 8 Inspect the high correlation between paid interest and interest rate with scatter plot



Predict Models Building

1. Using H2O package to bulid RandomForest Model to predict interest rate

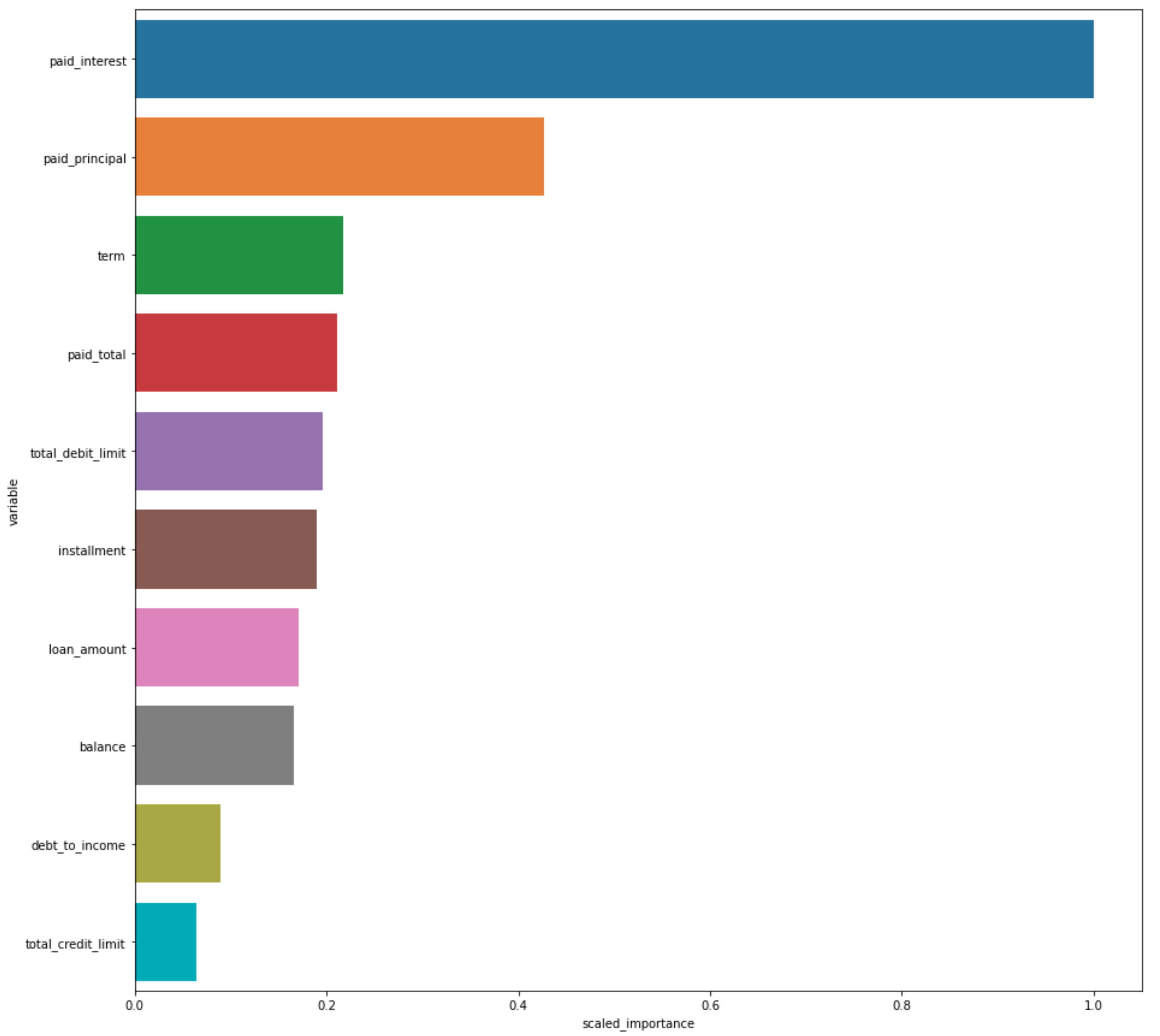
The importance of different predict features in the model.



We can see paid interest, paid principal, term, paid total, total debit limit are relative important variables for interest rate predicting.

Select top 10 important variables to rebuild the model.

The R squared of the new model for test dataset is 0.97.



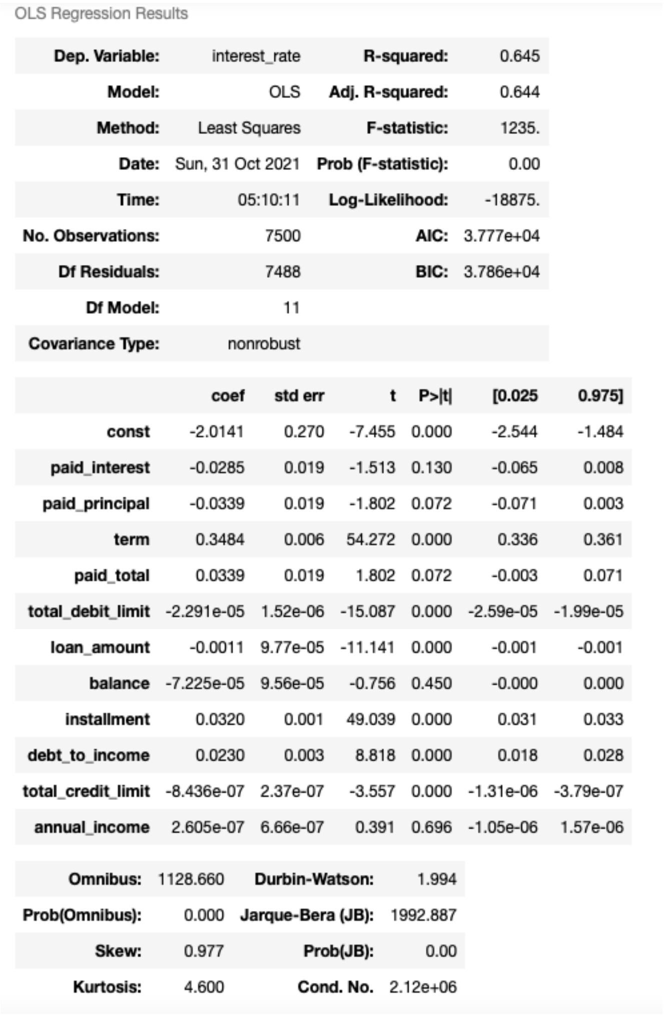
We can also inspect the importance of different variables for predicting interest rate in this model.

2. 2nd Model would apply multi-linear regression

Based on Random Forest Regression, develop a regression model with variables that are important.

I chose 'paid_interest','paid_principal','term','paid_total','total_debit_limit','loan_amount', 'balance','installment','debt_to_income','total_credit_limit', 'annual_income' as predict variables.

Below is the initial regression model result



OLS Regression Results

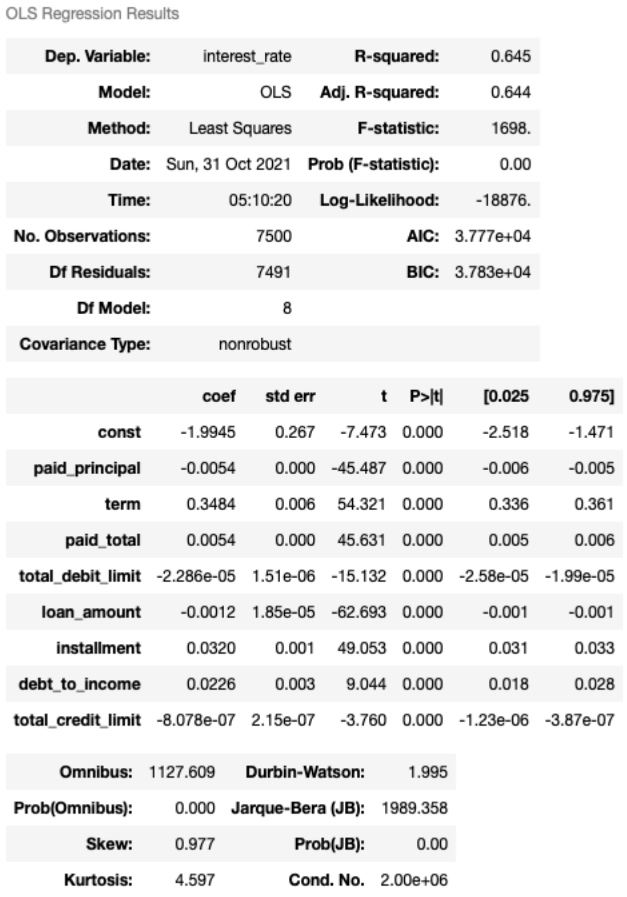
Dep. Variable:	interest_rate	**R-squared:**	0.645
Model:	OLS	**Adj. R-squared:**	0.644
Method:	Least Squares	**F-statistic:**	1235.
Date:	Sun, 31 Oct 2021	**Prob (F-statistic):**	0.00
Time:	05:10:11	**Log-Likelihood:**	-18875.
No. Observations:	7500	**AIC:**	3.777e+04
Df Residuals:	7488	**BIC:**	3.786e+04
Df Model:	11		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	-2.0141	0.270	-7.455	0.000	-2.544	-1.484
paid_interest	-0.0285	0.019	-1.513	0.130	-0.065	0.008
paid_principal	-0.0339	0.019	-1.802	0.072	-0.071	0.003
term	0.3484	0.006	54.272	0.000	0.336	0.361
paid_total	0.0339	0.019	1.802	0.072	-0.003	0.071
total_debit_limit	-2.291e-05	1.52e-06	-15.087	0.000	-2.59e-05	-1.99e-05
loan_amount	-0.0011	9.77e-05	-11.141	0.000	-0.001	-0.001
balance	-7.225e-05	9.56e-05	-0.756	0.450	-0.000	0.000
installment	0.0320	0.001	49.039	0.000	0.031	0.033
debt_to_income	0.0230	0.003	8.818	0.000	0.018	0.028
total_credit_limit	-8.436e-07	2.37e-07	-3.557	0.000	-1.31e-06	-3.79e-07
annual_income	2.605e-07	6.66e-07	0.391	0.696	-1.05e-06	1.57e-06

Omnibus:	1128.660	**Durbin-Watson:**	1.994
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1992.887
Skew:	0.977	**Prob(JB):**	0.00
Kurtosis:	4.600	**Cond. No.**	2.12e+06

Drop the variables which are not statistically significant one by one.
After that we only left:
'paid_principal','term','paid_total','total_debit_limit','loan_amount','installment','debt_to_income','total_credit_limit'
As predict variables.

Here is the regression model result after adjusted:



```
OLS Regression Results

    Dep. Variable:      interest_rate         R-squared:      0.645
           Model:                 OLS    Adj. R-squared:      0.644
          Method:       Least Squares       F-statistic:      1698.
            Date:    Sun, 31 Oct 2021  Prob (F-statistic):     0.00
            Time:            05:10:20    Log-Likelihood:     -18876.
No. Observations:                7500               AIC:  3.777e+04
    Df Residuals:                7491               BIC:  3.783e+04
        Df Model:                   8
 Covariance Type:           nonrobust

                       coef    std err        t   P>|t|     [0.025     0.975]
           const    -1.9945      0.267   -7.473   0.000     -2.518     -1.471
  paid_principal    -0.0054      0.000  -45.487   0.000     -0.006     -0.005
            term     0.3484      0.006   54.321   0.000      0.336      0.361
      paid_total     0.0054      0.000   45.631   0.000      0.005      0.006
total_debit_limit -2.286e-05  1.51e-06  -15.132   0.000   -2.58e-05  -1.99e-05
     loan_amount    -0.0012   1.85e-05  -62.693   0.000     -0.001     -0.001
     installment     0.0320      0.001   49.053   0.000      0.031      0.033
  debt_to_income     0.0226      0.003    9.044   0.000      0.018      0.028
total_credit_limit -8.078e-07  2.15e-07   -3.760   0.000   -1.23e-06  -3.87e-07

       Omnibus:  1127.609     Durbin-Watson:     1.995
Prob(Omnibus):     0.000  Jarque-Bera (JB):  1989.358
          Skew:     0.977          Prob(JB):      0.00
      Kurtosis:     4.597         Cond. No.  2.00e+06
```

Based on the model, we can see holding other variables unchanged, for 1 unit increase in paid principal, the interest rate will decrease -0.0054, while 1 unit increase of term, the interest rate will increase 0.3484, 1 unit increase of paid total, the interest rate will 0.0054, 1 unit increase of loan amount, the interest rate will decrease 0.0012, 1 unit increase of installment will increase interest rate for 0.032, 1 unit increase in debt to income ratio will increase interest rate for 0.0226.

The model has R squared score of 0.499 for the test dataset.

For model enhancements, it maybe helpful to try some combinations of different "ntrees" and "max_depth" to better optimize the parameters of the random forest regression model. It is also worth to transfer the categorical variables into dummy variables and apply into the linear regression models, to discover whether they are helpful to increase the model prediction accuracy.

For future investigation, I propose that we can record people's historical loan records over the time. Based on the change of a person's interest rate overtime, we may can create a time series model, to predict his/her interest rate changes in the future. Combining with the changes of other variables (such as paid interest, paid principal, paid total) overtime, we may can discover a clearer picture of the relationship between these variables and better foresee a person's loaning behavior

Case Study #2

There is 1 dataset(csv) with 3 years' worth of customer orders. There are 4 columns in the csv dataset: index, CUSTOMER_EMAIL (unique identifier as hash), Net Revenue, and Year.

For each year we need the following information:

- Total revenue for the current year
- New Customer Revenue **e.g., new customers not present in previous year only**
- Existing Customer Growth. To calculate this, use the Revenue of existing customers for current year –(minus) Revenue of existing customers from the previous year
- Revenue lost from attrition
- Existing Customer Revenue Current Year
- Existing Customer Revenue Prior Year
- Total Customers Current Year
- Total Customers Previous Year
- New Customers
- Lost Customers

Additionally, generate a few unique plots highlighting some information from the dataset. Are there any interesting observations?

Dataset

https://www.dropbox.com/sh/xhy2fzjdvg3ykhy/AADAVKH9tgD_dWh6TZtOd34ia?dl=0
customer_orders.csv

Output

An HTML website with the results of the data. Please highlight which year the calculations are for. All code should be hosted on GitHub for viewing. Please provide URL's to both the output and the GitHub repo.

** If you submit a jupyter notebook, also submit the accompanying python file. You may use python(.py), R, and RMD (knit to HTML) files. Other languages are acceptable as well.*

- Total revenue for the current year

When assuming current year is 2017, the Total Revenue for the Current Year is 31417495.03.
When assuming current year is 2016, the Total Revenue for the Current Year is 25730943.59.
When assuming current year is 2015, the Total Revenue for the Current Year is 29036749.19.

- New Customer Revenue **e.g., new customers not present in previous year only**

New Customer Revenue for 2015 is 29036749.19, for 2016 is 18245491.01, for 2017 is 28676607.64

- Existing Customer Growth. To calculate this, use the Revenue of existing customers for current year –(minus) Revenue of existing customers from the previous year

The Existing Customer Growth in 2016 is 20335.46.
The Existing Customer Growth is 2017 is 120238.74.

- Revenue lost from attrition

For 2016 Revenue lost from attrition is 21551296.61
For 2017 Revenue lost from attrition is 23089683.60

- Existing Customer Revenue Current Year
- Existing Customer Revenue Prior Year

When assuming 2015 is the current year, since it is the first year of the data, there is no Existing Customer for Revenue Current Year or Existing Customer Revenue for Prior Year.

When assuming 2016 is the current year, the Existing Customer Revenue for Current Year is 7485452.58, the Existing Customer Revenue for Prior Year is 7465117.12

When assuming 2017 is the current year, the Existing Customer Revenue for Current Year is 2740887.39, the Existing Customer Revenue for Prior Year is 2620648.65.

- Total Customers Current Year
- Total Customers Previous Year

When assuming current year is 2017, there are 249987 customers in current year. There are 204646 customers in 2016, and 376356 customers totally for pervious years.

When assuming current year is 2016, there are 204646 customers in current year. There are 231294 customers for pervious year.

When assuming current year is 2015, there are 231294 customers in current year. There are 0 customer for pervious year.

- New Customers

  New Customers for 2017 is 228262

  New Customers for 2016 is 145062

  New Customers for 2015 is 231294

- Lost Customers

  The lost customers for 2016 is 171710.

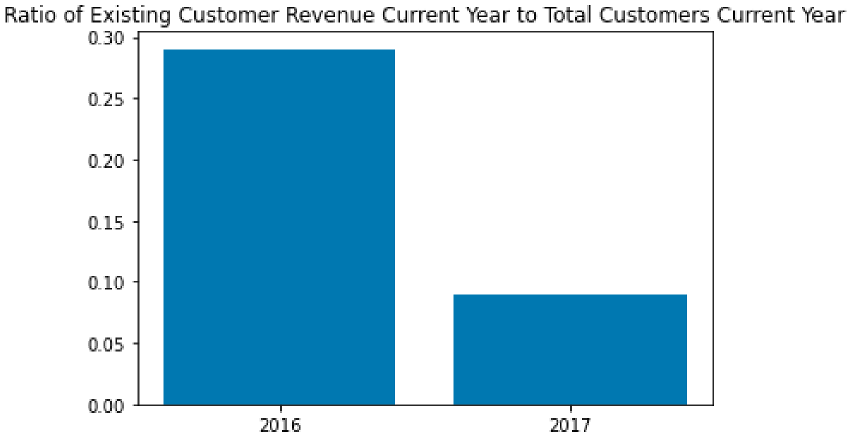
  The lost customers for 2017 is 183687.

  The total lost customer for 2017 since 2015 is 354631.

Additionally, generate a few unique plots highlighting some information from the dataset. Are there any interesting observations?

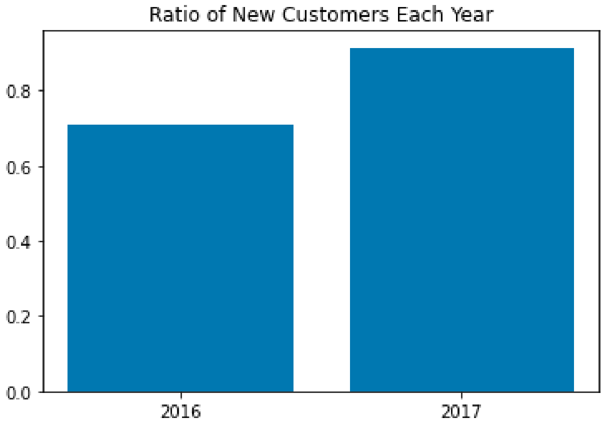
Plot 1

It is worth to investigate the existing customer ratio (which is defined as existing customers amount divided by total customer in a specific year) to see any fluctuation. As we see the existing customer ratio decreased in 2017.



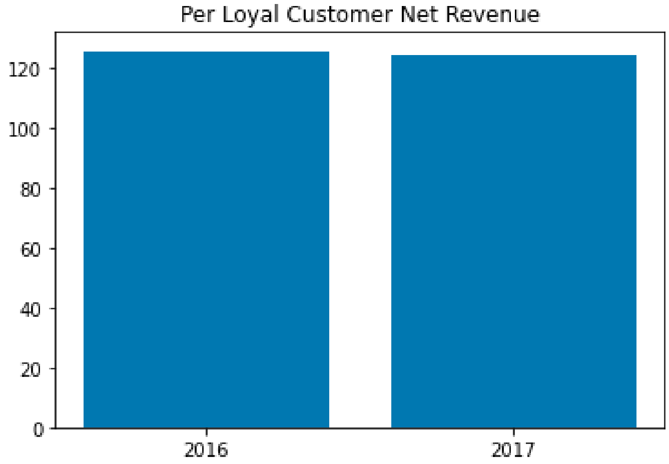
Plot 2 New customer ratio.

Further we can also see new customer ratio (which is defined by the new customer amount divided by total customer amount in that year) for each year.



As we see the ratio increased in 2017.

Plot 3 We can also investigate the net revenue for each customer who had purchased from 2015 all the way to 2017 (loyal customer), discover their net revenue changes.



For the chart we can see the net revenue for each "loyal customer" has not changed too much over the year (which is around 120).