Shyam-Project

Code

Shyam BV

December 7, 2016

Part 1 - Introduction

Some months back, I was trying to apply for a lending club loan. Interest rate was different me and my colleage. That made me to think more about Lending club(LC). LC is a US peer-to-peer lending company. Where investors provide funding and borrowers return back the payments. Lending club selects and approves the borrowers using many parameters. It is a sort of EBay for loans.

In this project, I'll be trying to predict the interest rate with various predictor variables. By performing this analysis we will know below information.

1. What parameters will impact my interest rate? ie., Is loan interest % predictive of FICO credit score alone?
2. Is loan funded amount are equal for different purpose of loan request? So the person can get loan in that particular loan type.
3. It is always mentioned that living state plays a important role in interest rate. This hypothesis will be validated.
4. There is a myth that home ownership will impact FICO scores. It will be validated via this dataset.
5. Did lending club receive equal number of loans in each month

Part 2 - Data

When we register as a lending club user, you will get access to the borrowers data from Lending loan website (https://www.lendingclub.com/info/download-data.action (https://www.lendingclub.com/info/download-data.action)).

This dataset has borrowers details(personal info will be removed) It has the funded amount, interest rate, fico credit score and about 115 variables. Also the row count is around 130K for Q1 2016.

For current analysis, I have taken a simple random sample of 1000 rows. These data are transformed and modified to perform analysis on data.

Data collection

Describe the method of data collection.

When we register as a lending club borrower/invester, you will get access to the borrowers data from Lending loan website (https://www.lendingclub.com/info/download-data.action (https://www.lendingclub.com/info/download-data.action)).

LC also provides loan rejection dataset. But for current analysis, we have used only the borrowers dataset.

Cases

This dataset is an Observational study. Consumers requested for loan in LC. Each row will contain borrowers information(personal info will be removed by LC) and their current loan status.

For this project, in some cases, I have used the comple dataset and for some analysis I have taken a simple random sample of 1000 rows.

Variables

It has the funded amount, interest rate, fico credit score and about 115 variables. Also there are around 140K observations for Q1 2016.

But for this current analysis, below are the variables used.

LoanStatNew	Description
loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
funded_amnt	The total amount committed to that loan at that point in time.
int_rate	Interest Rate on the loan
emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
home_ownership	The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are: RENT, OWN, MORTGAGE, OTHER
annual_inc	The self-reported annual income provided by the borrower during registration.
verification_status	Indicates if income was verified by LC, not verified, or if the income source was verified
addr_state	The state provided by the borrower in the loan application
issue_d	The month which the loan was funded
loan_status	Current status of the loan
purpose	A category provided by the borrower for the loan request.
total_pymnt_inv	Payments received to date for portion of total amount funded by investors
fico_range_high	The upper boundary range the borrower’s FICO at loan origination belongs to.
fico_range_low	The lower boundary range the borrower’s FICO at loan origination belongs to.
term	The number of payments on the loan. Values are in months and can be either 36 or 60.

Type of study

This is an observational study. We will arrive at conclusion by performing below tests on the mentioned variables.

1. Hypothesis Test - Reasoning whether the inference is just by chance.
2. F-Test - Compare multiple variables
3. Create Linear regression - Form the regression line with the available parameters. Check the values between predicted and observed outcome.
4. Create logarithmic regression - Create an model for non-linear data varaibles.

Scope of inference

1. Generalizability: Population of interest for this study is applicable only to US population and its territory. To borrow the loan from LC, it requires credit score info, personal info like home ownership, purpose of loan, employment length, annual income.

If these information is not available for the borrower, this study will not be applicable.

Also this analysis is perfomed only for lending club and other peer-to-peer borrowing companies. It may not be applicable for the banking interest rates.

2. Bias: Here the bias that prevents the generalizability is the borrower information. Only the person who has knowledge about LC and peer-to-peer investing, is requesting for a loan in LC. Bank might use another cofounding variable to get the interest

rate.

Causality

As this is an observational study we cannot derive any causal connections between the variables.

Code

Part 3 - Exploratory data analysis

Below are some exploratory data analysis charts to understand more about the data.

Explanatory

Below table summarizes the question, response and explanatory variable. It also shows whether it is Numerical or Categorical.

Question	Response Variable	Explanatory Variable
1. What parameters will impact my interest rate? ie., Is loan interest % predictive of FICO credit score alone?	Interest rate % (Numerical)	FICO Credit score (Numerical), home ownership (Categorical), Purpose (Categorical)

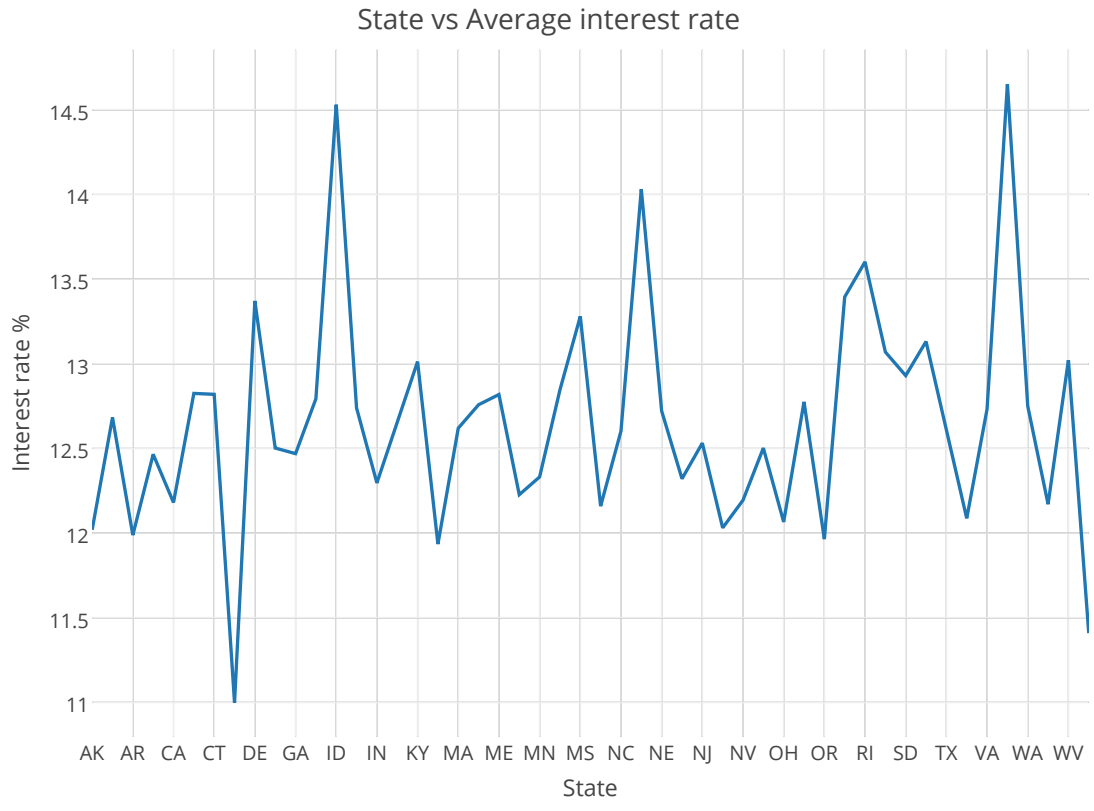
2. Is loan funded interest rate % are equal for different purpose of loan request? | Interest rate % (Numerical) | purpose (Categorical)
3. Does different states have equal interest rate? | Interest rate % (Numerical) | States (Categorical)
4. Does home ownership really impact FICO scores or just by chance? | FICO scores (Numerical) | Home Ownership (Categorical)
5. Did lending club receive equal number of loans in each month? | Loan count (Numerical) | - |

Charts

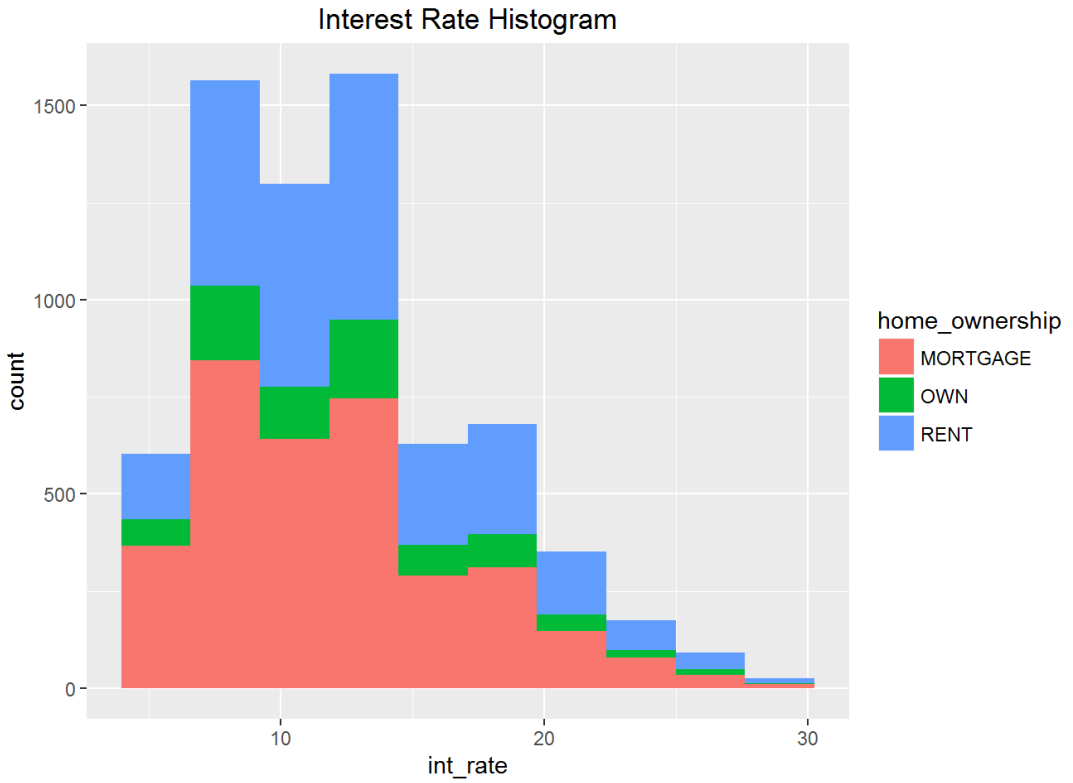
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

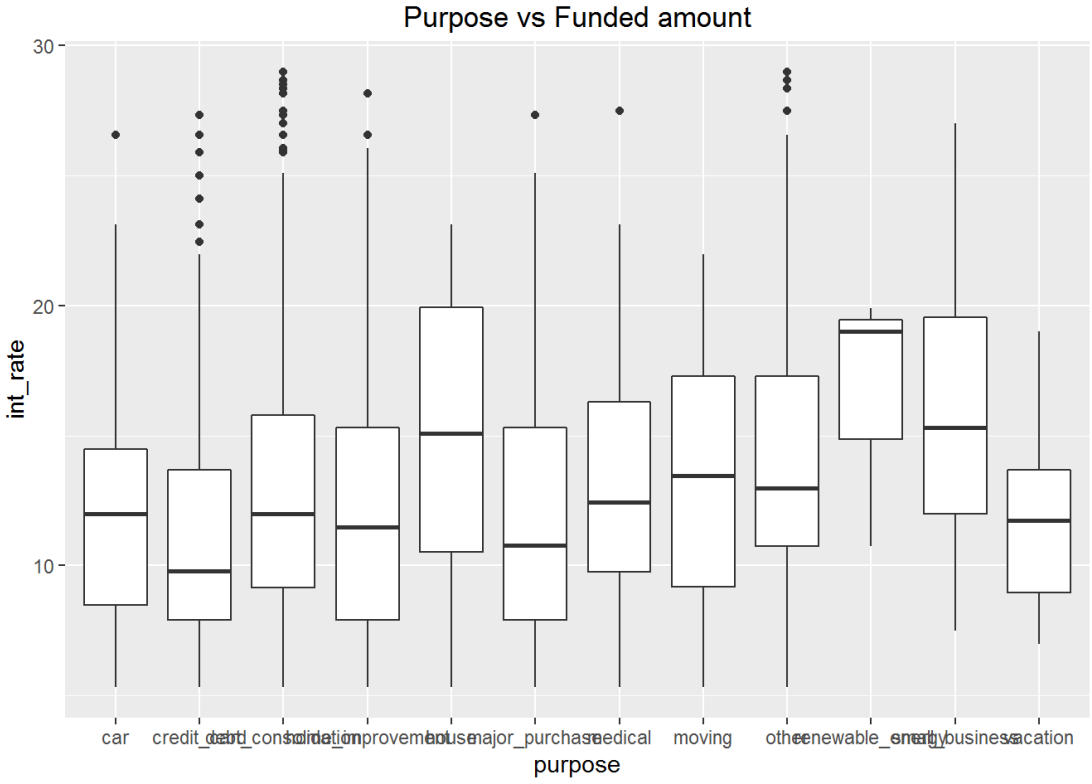
Code

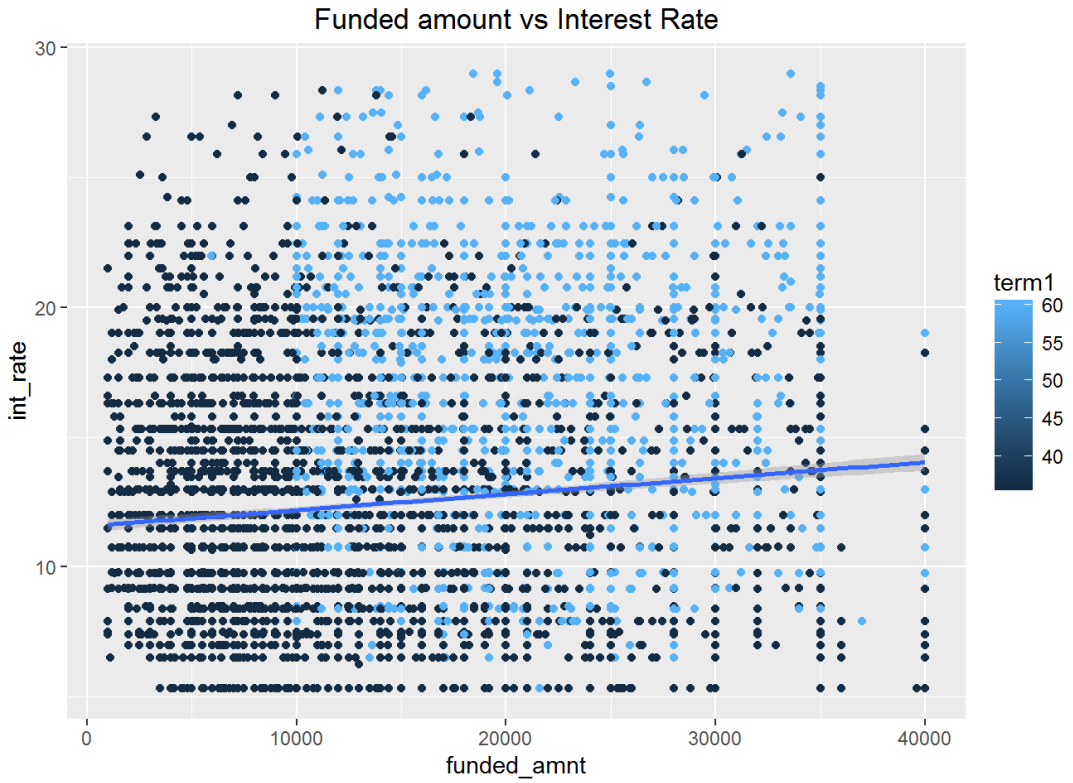
Code

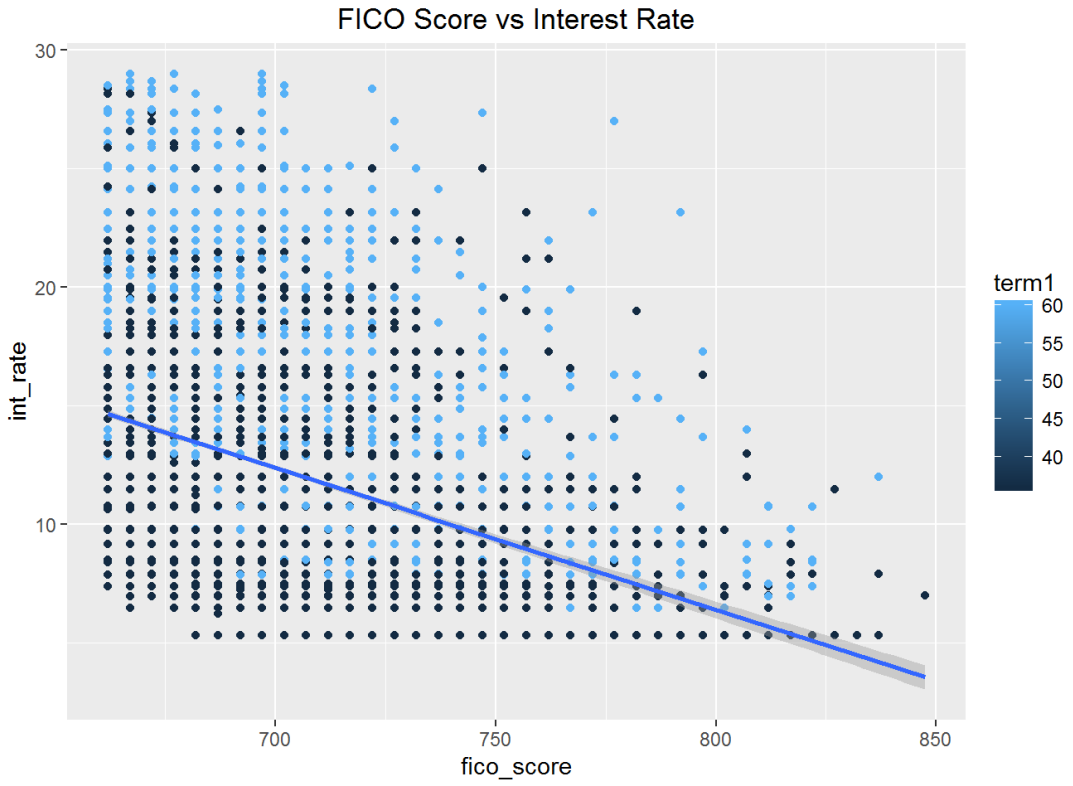


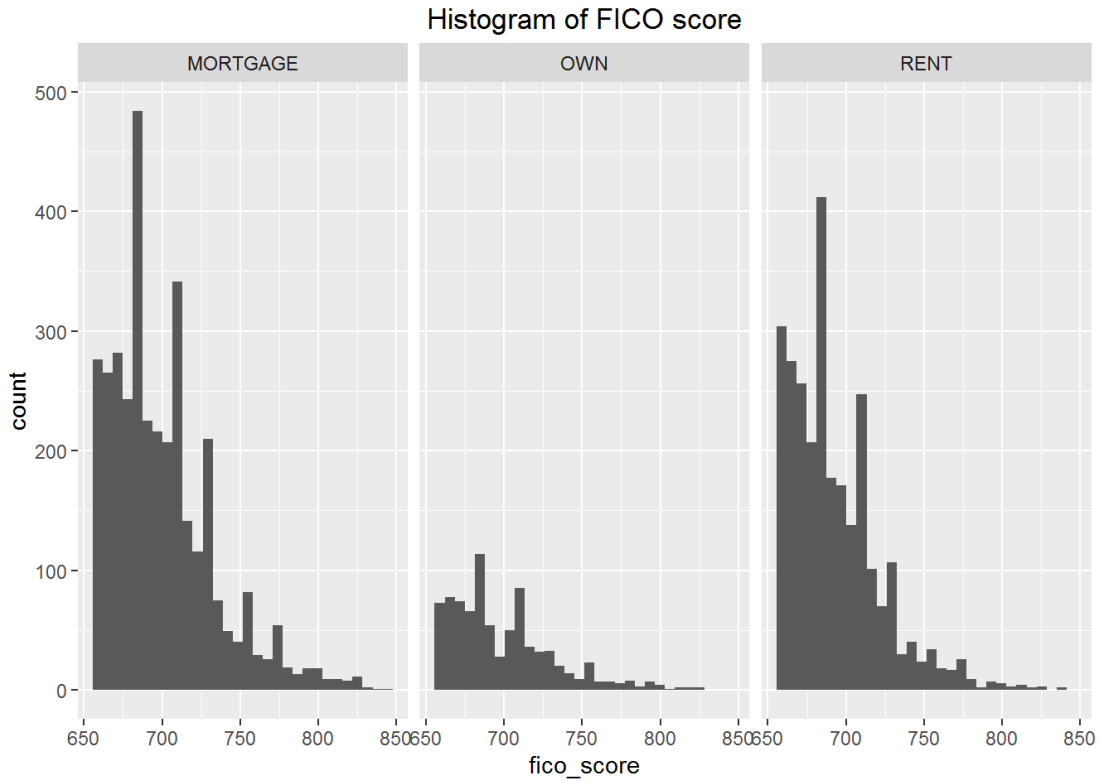
Code


Interest Rate Histogram
count
1500
1000
500
0
10
20
30
int_rate
home_ownership
MORTGAGE
OWN
RENT


Purpose vs Funded amount
int_rate
30
20
10
car
credit_card
debt_consolidation
home_improvement
house
major_purchase
medical
moving
other
renewable_energy
small_business
vacation
purpose


Funded amount vs Interest Rate
int_rate
30
20
10
0
10000
20000
30000
40000
funded_amnt
term1
60
55
50
45
40


FICO Score vs Interest Rate
int_rate
30
20
10
700
750
800
850
fico_score
term1
60
55
50
45
40



Statistics

Exploratory data analysis suggests below statistics.

Statistic	Variable	Value
Population	Mean Interest rate	12.4776311
Population	SD Interest rate	4.8290031
Sample Statistics	Mean Interest rate	12.5321657
Sample Statistics	SD Interest rate	4.8246797

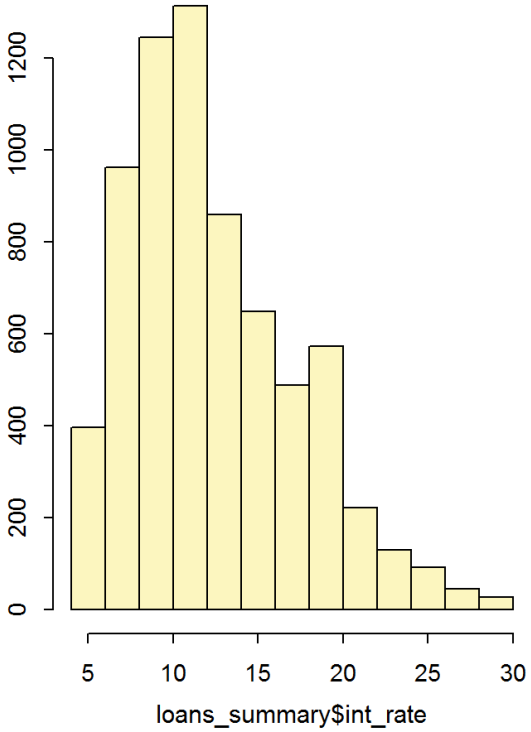
Confidence interval of the Interest Rate

Point estimate from the sample with the confidence interval is shown below

Code

```
## Single mean
## Summary statistics:
```

```
## mean = 12.5322 ;  sd = 4.8247 ;  n = 7000
## Standard error = 0.0577
## 95 % Confidence interval = ( 12.4191 , 12.6452 )
```



For this test lest validate the total sample size required.

If the margin of error to be 3%, we need to get the samples of around 6738.5430648. Code

- Some states have higher interest rates than another.
- Those who have home morgage has higher interest rate than others.
- Interest rate increases with funded amount.

Part 4- Inference

1. Is loan interest % predictive of credit score?

To perform this statement, we will use the linear model to validate it. Here we are performing forward elimination technique to get the maximum adjusted R-squared vale.

Code

```
##
## Call:
## lm(formula = int_rate ~ fico_score + home_ownership + purpose +
##     term1 + loan_amnt + annual_inc + emp_length + issue_d + pub_rec_bankruptcies +
##     dti, data = loans_summary)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.2082 -2.7615 -0.6272  2.1442 25.0656
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 4.724e+01  1.211e+00  39.006  < 2e-16 ***
## fico_score                 -5.995e-02  1.541e-03 -38.902  < 2e-16 ***
## home_ownershipOWN           5.643e-01  1.498e-01   3.768 0.000166 ***
## home_ownershipRENT          8.014e-01  1.048e-01   7.648 2.32e-14 ***
## purposecredit_card         -1.667e+00  4.784e-01  -3.485 0.000494 ***
## purposedebt_consolidation -2.176e-01  4.729e-01  -0.460 0.645473
## purposehome_improvement     3.371e-01  5.072e-01   0.665 0.506323
## purposehouse                2.565e+00  9.159e-01   2.801 0.005115 **
## purposemajor_purchase      -1.665e-01  5.645e-01  -0.295 0.768012
## purposemedical              1.043e+00  6.194e-01   1.684 0.092214 .
## purposemoving               1.548e+00  8.182e-01   1.892 0.058515 .
## purposeother                1.681e+00  5.050e-01   3.328 0.000878 ***
## purposerenewable_energy     5.802e+00  2.276e+00   2.549 0.010818 *
## purposesmall_business       3.500e+00  6.647e-01   5.266 1.44e-07 ***
## purposevacation             3.978e-01  7.951e-01   0.500 0.616889
## term1                       1.651e-01  4.622e-03  35.718  < 2e-16 ***
## loan_amnt                   5.996e-05  6.296e-06   9.524  < 2e-16 ***
## annual_inc                 -8.864e-06  7.649e-07 -11.589  < 2e-16 ***
## emp_length1 year           -1.198e-01  2.409e-01  -0.497 0.619093
## emp_length10+ years        -3.223e-01  1.821e-01  -1.769 0.076858 .
## emp_length2 years          -1.660e-01  2.261e-01  -0.734 0.462934
## emp_length3 years          -2.837e-01  2.274e-01  -1.247 0.212274
## emp_length4 years          -7.933e-02  2.520e-01  -0.315 0.752909
## emp_length5 years          -2.565e-01  2.494e-01  -1.028 0.303847
## emp_length6 years          -1.662e-01  2.778e-01  -0.598 0.549676
## emp_length7 years          -5.150e-01  3.055e-01  -1.686 0.091846 .
## emp_length8 years          -3.066e-01  2.652e-01  -1.156 0.247703
## emp_length9 years          -1.969e-01  2.953e-01  -0.667 0.504825
## emp_lengthNA                8.164e-01  2.454e-01   3.327 0.000883 ***
## issue_dJan-2016            -3.414e-01  1.260e-01  -2.710 0.006750 **
## issue_dMar-2016             9.206e-02  1.092e-01   0.843 0.399448
## pub_rec_bankruptcies       -2.038e-01  1.209e-01  -1.685 0.091983 .
## dti                         3.165e-04  3.862e-04   0.819 0.412613
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.851 on 6967 degrees of freedom
## Multiple R-squared:  0.3657, Adjusted R-squared:  0.3628
## F-statistic: 125.5 on 32 and 6967 DF,  p-value: < 2.2e-16
```

After mulitple iterations, we have come to the conclusion that below are the variables that affect interest rate %

Code

fico_score
home_ownership
purpose
term1
loan_amnt
annual_inc
emp_length
issue_d
pub_rec_bankruptcies
dti(debt to income ration)

Below are the conditions for least squared line

We are going to validate the conditions for least squared line.

1. Linearity

From the below chart, it shows that there is a downward relationship between `FICO credit score` and `interest rate` . But the linear model is not very strong due to large number of variability.

The correlation between the two variables is around `-0.383` .

Code

```
## [1] -0.3858804
```

2. Nearly normal residuals

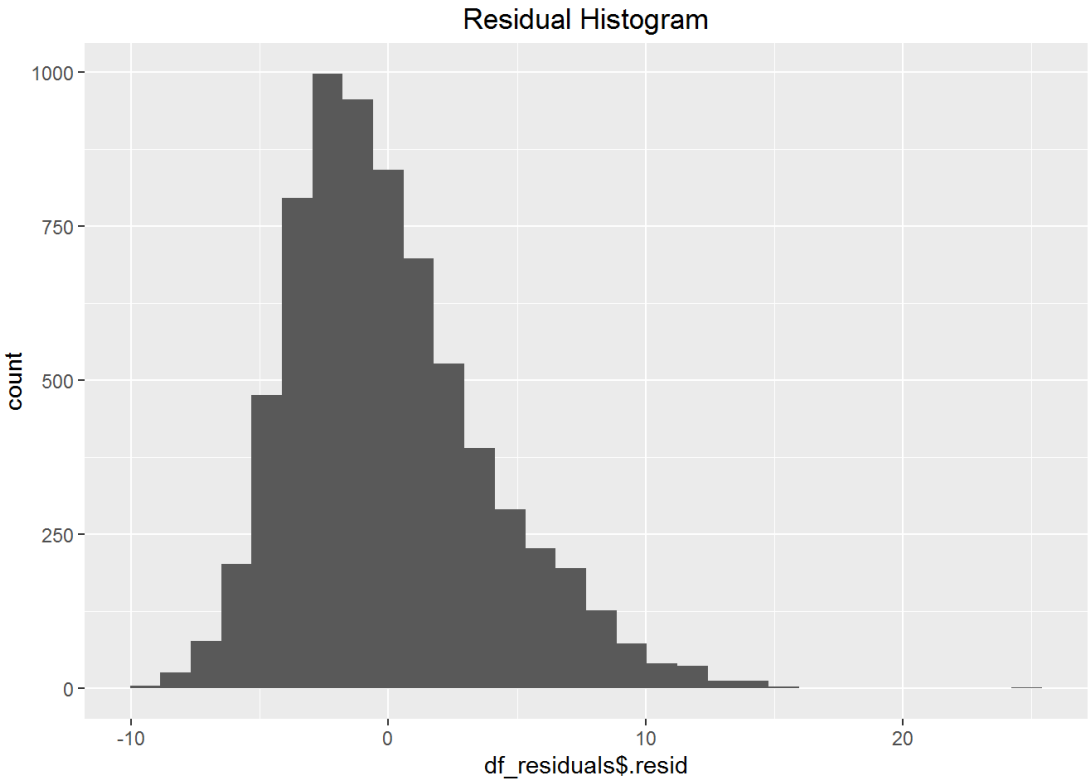
Let's check the residuals normality with histogram and qqplot.

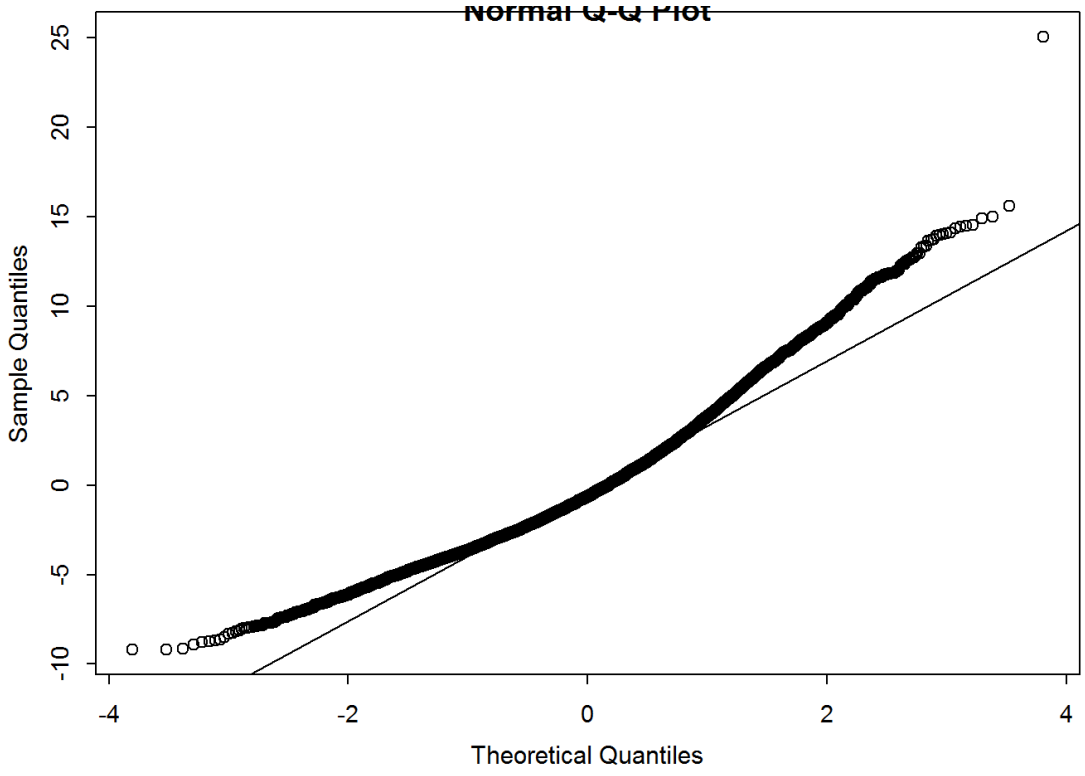
Code

Code

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Code

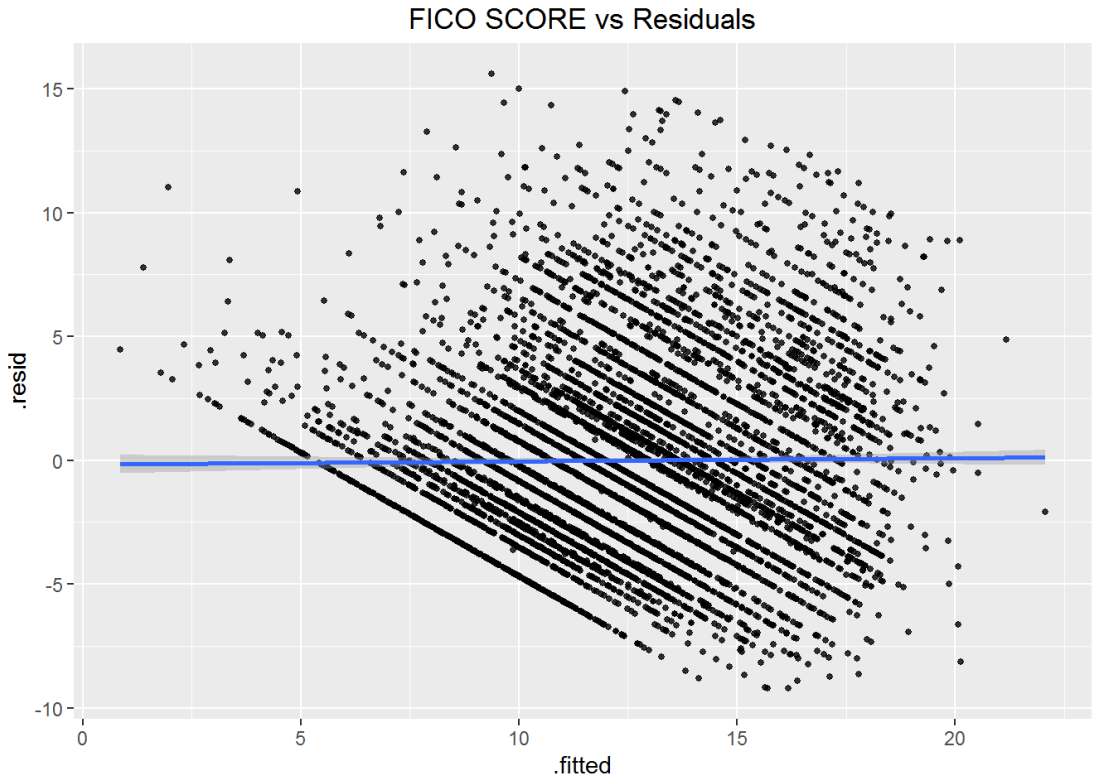




The plots show that residuals are slightly left skewed. But the residuals are normal

3. Constant Variability

Code



Above plot shows that there is a constant variability in the chart. 36.28% of interest rate % is explained by the model.

2. Different purpose of loan request

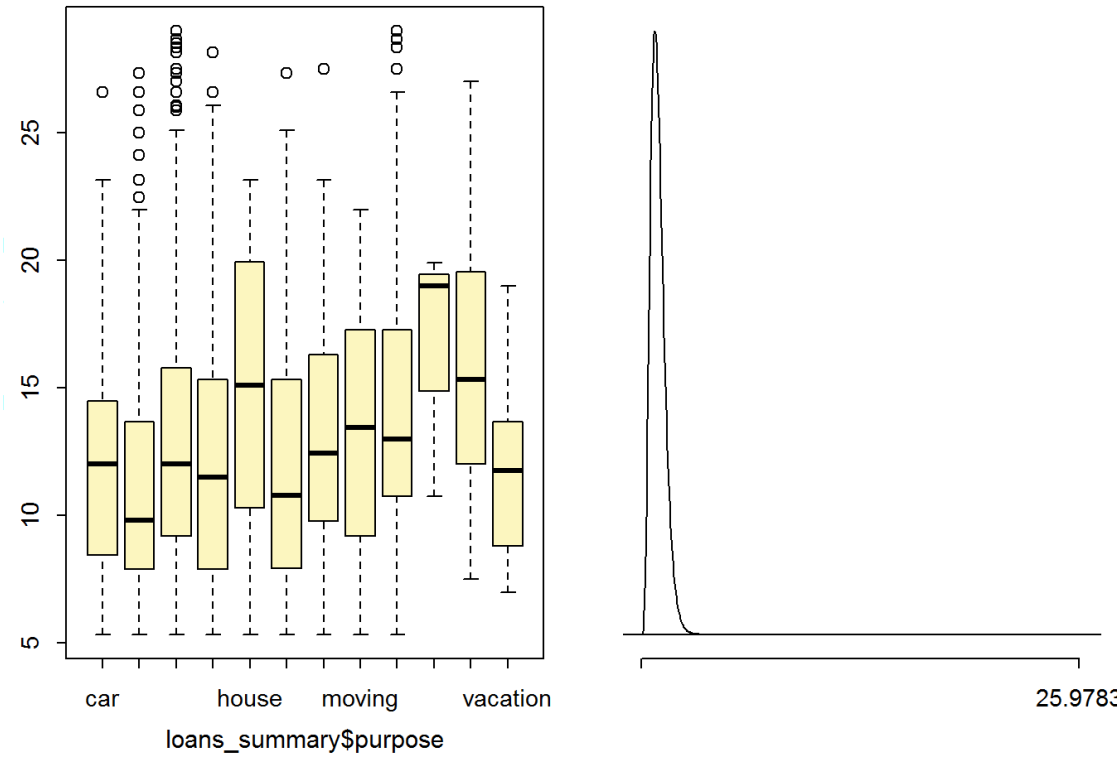
2. Is loan funded amount are equal for different purpose of loan request?

Let's validate if the purpose of loan interest rate varies or not.

Code

```
## Response variable: numerical, Explanatory variable: categorical
## ANOVA
##
## Summary statistics:
## n_car = 68, mean_car = 12.3943, sd_car = 4.7805
## n_credit_card = 1661, mean_credit_card = 11.0786, sd_credit_card = 4.1975
## n_debt_consolidation = 4051, mean_debt_consolidation = 12.9649, sd_debt_consolidation = 4.8732
## n_home_improvement = 410, mean_home_improvement = 12.0454, sd_home_improvement = 5.1758
## n_house = 24, mean_house = 15.2312, sd_house = 5.629
## n_major_purchase = 150, mean_major_purchase = 11.8111, sd_major_purchase = 4.8367
## n_medical = 90, mean_medical = 13.2814, sd_medical = 4.7381
## n_moving = 33, mean_moving = 13.7994, sd_moving = 4.44
## n_other = 407, mean_other = 14.0046, sd_other = 4.8329
## n_renewable_energy = 3, mean_renewable_energy = 16.5433, sd_renewable_energy = 5.0373
## n_small_business = 67, mean_small_business = 15.8237, sd_small_business = 5.2839
## n_vacation = 36, mean_vacation = 11.77, sd_vacation = 3.3051
```

```
## H_0: All means are equal.
## H_A: At least one mean is different.
## Analysis of Variance Table
##
## Response: y
##             Df Sum Sq Mean Sq F value    Pr(>F)
## x           11   6401  581.87  25.978 < 2.2e-16
## Residuals 6988 156519   22.40
##
## Pairwise tests: t tests with pooled SD
##                       car credit_card debt_consolidation home_improvement
## credit_card        0.0247          NA                 NA               NA
## debt_consolidation 0.3241      0.0000                 NA               NA
## home_improvement   0.5734      0.0002             0.0002               NA
## house              0.0116      0.0000             0.0194           0.0014
## major_purchase     0.3994      0.0695             0.0034           0.6040
## medical            0.2434      0.0000             0.5303           0.0249
## moving             0.1617      0.0011             0.3131           0.0406
## other              0.0094      0.0000             0.0000           0.0000
## renewable_energy   0.1373      0.0457             0.1905           0.1010
## small_business     0.0000      0.0000             0.0000           0.0000
## vacation           0.5222      0.3858             0.1315           0.7378
##                      house major_purchase medical moving  other
## credit_card             NA             NA      NA     NA     NA
## debt_consolidation      NA             NA      NA     NA     NA
## home_improvement        NA             NA      NA     NA     NA
## house                   NA             NA      NA     NA     NA
## major_purchase     0.0010             NA      NA     NA     NA
## medical            0.0730         0.0198      NA     NA     NA
## moving             0.2595         0.0289  0.5907     NA     NA
## other              0.2173         0.0000  0.1896 0.8107     NA
## renewable_energy   0.6508         0.0864  0.2403 0.3364 0.3546
## small_business     0.5987         0.0000  0.0009 0.0443 0.0036
## vacation           0.0055         0.9626  0.1054 0.0752 0.0066
##                    renewable_energy small_business
## credit_card                      NA             NA
## debt_consolidation               NA             NA
## home_improvement                 NA             NA
## house                            NA             NA
## major_purchase                   NA             NA
## medical                          NA             NA
## moving                           NA             NA
## other                            NA             NA
## renewable_energy                 NA             NA
## small_business               0.7967             NA
## vacation                     0.0933              0
```



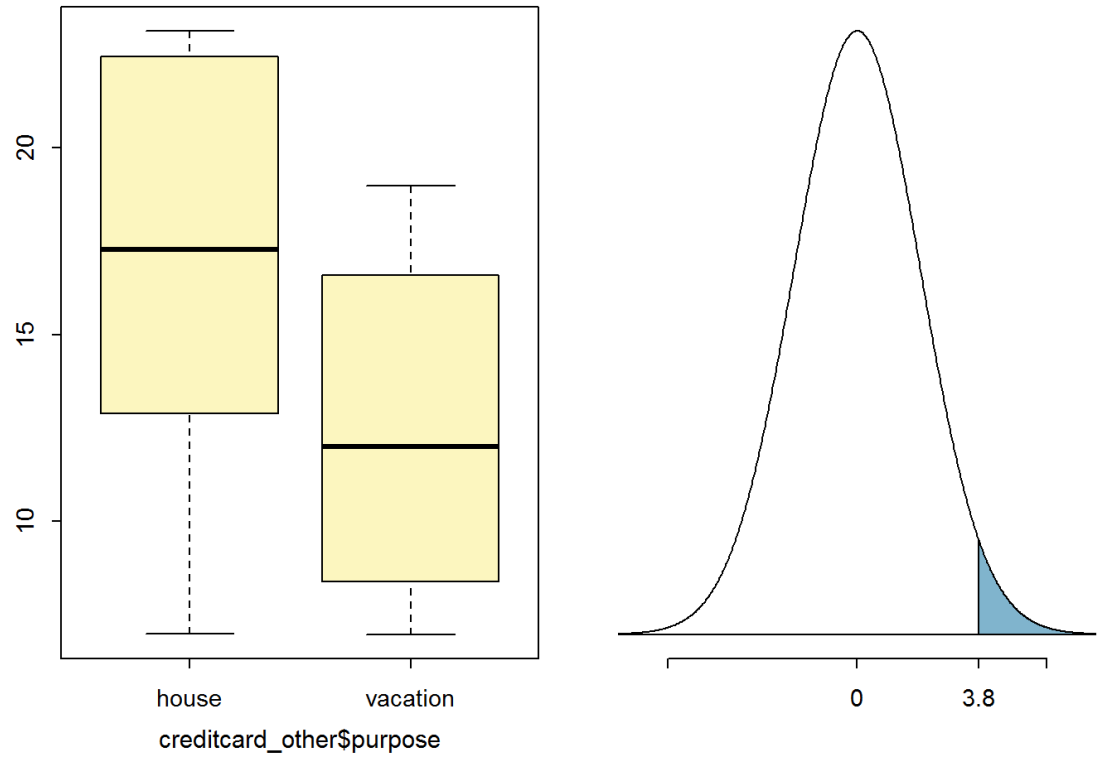
Above output shows that interest rate varies for each purpose of loan.

Our Model states that if the purpose of loan is `vacation` then the interest rate will be higher. Let's compare with other purpose `house`

Code

```
## Response variable: numerical, Explanatory variable: categorical
## Difference between two means
## Summary statistics:
## n_house = 13, mean_house = 16.3769, sd_house = 5.6391
## n_vacation = 13, mean_vacation = 12.5723, sd_vacation = 4.3593
```

```
## Observed difference between means (house-vacation) = 3.8046
##
## H0: mu_house - mu_vacation = 0
## HA: mu_house - mu_vacation > 0
## Standard error = 1.977
## Test statistic: T =  1.925
## Degrees of freedom:  12
## p-value =  0.0392
```



In output, the p-value is low. So we will reject null Hypothesis and conclude that the interest rate % is higher for `house` .

3. State vs interest rate

3. It is always mentioned that living state plays a important role in interest rate. This hypothesis will be validated.

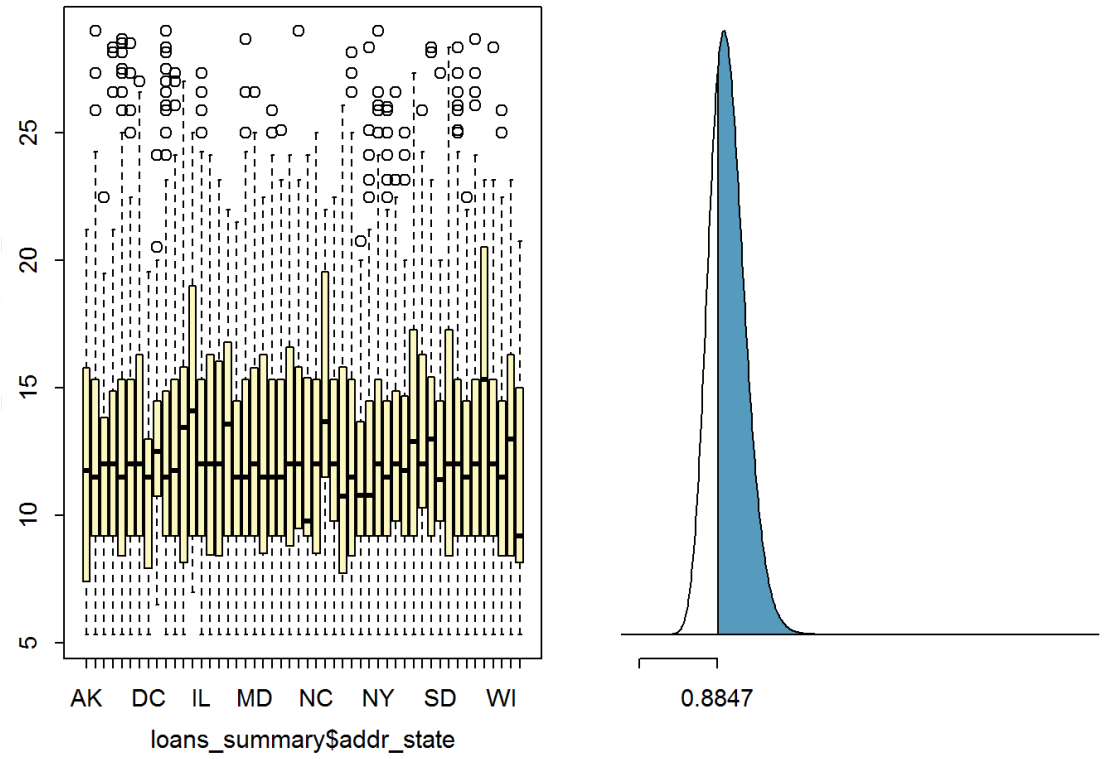
Null hypothesis $H0$: Interest rate is same for all states.

Alternative hypothesis HA:Interest rate is different at-least for one state.

Code

```
## Response variable: numerical, Explanatory variable: categorical
## ANOVA
##
## Summary statistics:
## n_AK = 34, mean_AK = 12.0197, sd_AK = 4.747
## n_AL = 95, mean_AL = 12.6826, sd_AL = 5.1936
## n_AR = 47, mean_AR = 11.9881, sd_AR = 3.9117
## n_AZ = 134, mean_AZ = 12.4653, sd_AZ = 4.6191
## n_CA = 965, mean_CA = 12.1806, sd_CA = 4.9903
## n_CO = 142, mean_CO = 12.8249, sd_CO = 5.0707
## n_CT = 104, mean_CT = 12.8195, sd_CT = 4.9946
## n_DC = 22, mean_DC = 10.9991, sd_DC = 3.7911
## n_DE = 26, mean_DE = 13.3712, sd_DE = 4.2513
## n_FL = 498, mean_FL = 12.5023, sd_FL = 4.7869
## n_GA = 216, mean_GA = 12.4691, sd_GA = 4.8793
## n_HI = 39, mean_HI = 12.7928, sd_HI = 5.4609
## n_ID = 26, mean_ID = 14.5304, sd_ID = 5.0331
## n_IL = 262, mean_IL = 12.7392, sd_IL = 4.8206
## n_IN = 124, mean_IN = 12.2962, sd_IN = 4.8372
## n_KS = 64, mean_KS = 12.6522, sd_KS = 4.6001
## n_KY = 64, mean_KY = 13.0127, sd_KY = 4.4486
## n_LA = 92, mean_LA = 11.9349, sd_LA = 3.8755
## n_MA = 149, mean_MA = 12.6195, sd_MA = 4.7527
## n_MD = 152, mean_MD = 12.7582, sd_MD = 4.8841
## n_ME = 29, mean_ME = 12.8183, sd_ME = 5.3032
## n_MI = 203, mean_MI = 12.2267, sd_MI = 4.6473
## n_MN = 115, mean_MN = 12.3316, sd_MN = 4.3078
## n_MO = 115, mean_MO = 12.8459, sd_MO = 4.9702
## n_MS = 43, mean_MS = 13.2798, sd_MS = 5.4124
## n_MT = 19, mean_MT = 12.1589, sd_MT = 5.1096
## n_NC = 209, mean_NC = 12.6026, sd_NC = 4.7808
## n_ND = 21, mean_ND = 14.031, sd_ND = 4.9149
## n_NE = 29, mean_NE = 12.7214, sd_NE = 4.4401
## n_NH = 35, mean_NH = 12.3209, sd_NH = 5.9575
## n_NJ = 259, mean_NJ = 12.5314, sd_NJ = 5.0718
## n_NM = 49, mean_NM = 12.031, sd_NM = 3.825
## n_NV = 97, mean_NV = 12.1941, sd_NV = 4.7798
## n_NY = 542, mean_NY = 12.5022, sd_NY = 4.5085
## n_OH = 249, mean_OH = 12.0657, sd_OH = 4.399
## n_OK = 82, mean_OK = 12.7751, sd_OK = 4.5988
## n_OR = 88, mean_OR = 11.9642, sd_OR = 4.529
## n_PA = 223, mean_PA = 13.3944, sd_PA = 5.1073
## n_RI = 24, mean_RI = 13.6021, sd_RI = 5.4027
## n_SC = 74, mean_SC = 13.0695, sd_SC = 5.173
## n_SD = 22, mean_SD = 12.93, sd_SD = 5.238
## n_TN = 121, mean_TN = 13.1315, sd_TN = 5.5413
## n_TX = 543, mean_TX = 12.6216, sd_TX = 4.7932
## n_UT = 47, mean_UT = 12.087, sd_UT = 4.662
## n_VA = 205, mean_VA = 12.732, sd_VA = 5.0258
## n_VT = 9, mean_VT = 14.6511, sd_VT = 6.1342
## n_WA = 148, mean_WA = 12.7499, sd_WA = 4.7428
## n_WI = 109, mean_WI = 12.1703, sd_WI = 4.8583
## n_WV = 25, mean_WV = 13.0212, sd_WV = 5.4999
## n_WY = 11, mean_WY = 11.4091, sd_WY = 5.0787
```

```
## H_0: All means are equal.
## H_A: At least one mean is different.
## Analysis of Variance Table
##
## Response: y
##             Df Sum Sq Mean Sq F value Pr(>F)
## x           49   1010  20.611  0.8847 0.7001
## Residuals 6950 161910  23.296
```



Higher p-value states that we failed to reject null hypothesis. It might also mention that interest rate is same for all states.

4. Homeownership vs FICO score

4. Does home ownership really impact FICO scores or just by chance?

We want to understand does the home ownership really impact FICO scores. Below we are performing ANOVA between the three variables (RENT VS MORTGAGE VS OWN)

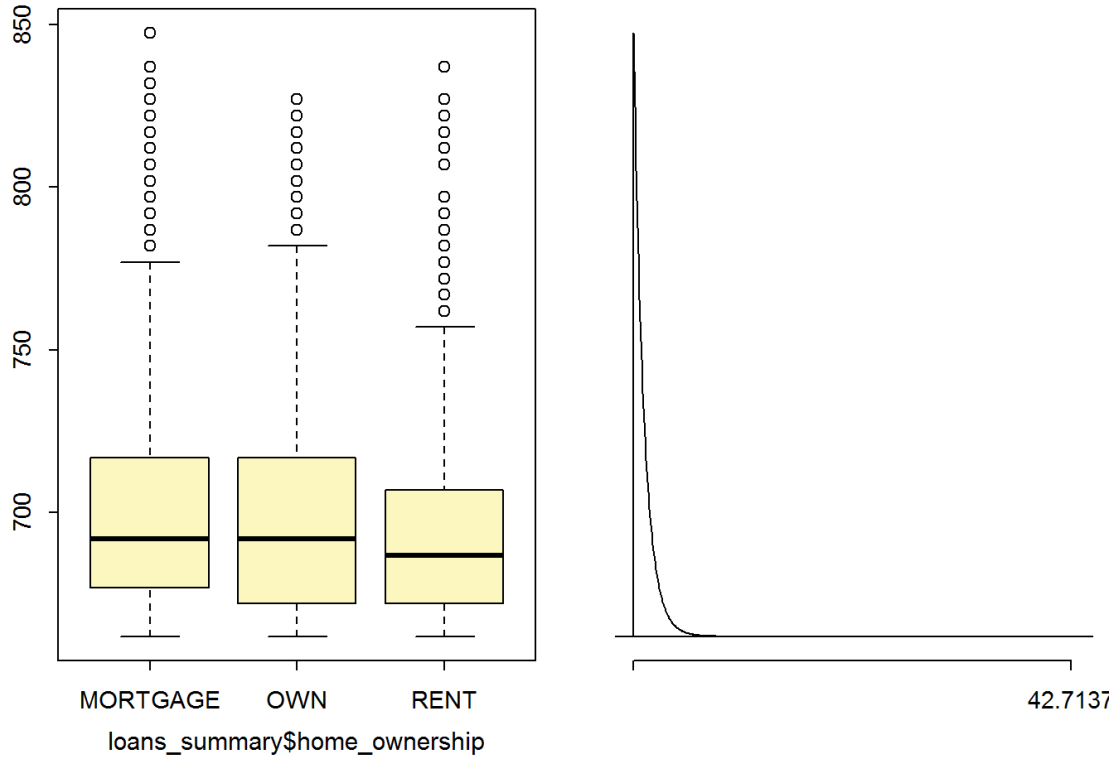
Null hypothesis $H0$: FICO scores are same between different home ownership. Scores do not dependent on home ownership.

Alternative hypothesis HA: FICO scores are different for each home ownership.

Code

```
## Response variable: numerical, Explanatory variable: categorical
## Summary statistics:
## n_MORTGAGE = 3470, mean_MORTGAGE = 700.4973, sd_MORTGAGE = 32.4716
## n_OWN = 838, mean_OWN = 698.9153, sd_OWN = 32.4277
## n_RENT = 2692, mean_RENT = 693.2351, sd_RENT = 28.3526
```

```
## H_0: All means are equal.
## H_A: At least one mean is different.
## Analysis of Variance Table
##
## Response: y
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## x            2   81815   40907  42.714 < 2.2e-16
## Residuals 6997 6701103     958
##
## Pairwise tests: t tests with pooled SD
##      MORTGAGE OWN
## OWN    0.1842  NA
## RENT   0.0000   0
```



In output, the p-value is low. So we will reject null Hypothesis and conclude that FICO scores are not impacted by home ownership

5. Loan issue date

5. Did lending club receive equal number of loans in each month?

Our model also states that if the loan interest rate depends on the issue month. So lets validate it.

$H0$: There is no incosistency between the observed and expected counts. Observed counts follow the same distribution as expected counts.

$H1$: There is an incosistency between the observed and expected counts. Observed counts do not follow the same distribution as expected counts.Some months are different than another.

Code

```
##
##  Chi-squared test for given probabilities
##
## data:  loans_count$observed_loans
## X-squared = 526.21, df = 2, p-value < 2.2e-16
```

Code

```
## # A tibble: 3 × 4
##   loan_month observed_loans  percent expected_loans
##        <int>          <int>    <dbl>          <dbl>
## 1          1           1738 33.33333       2333.333
## 2          2           2041 33.33333       2333.333
## 3          3           3221 33.33333       2333.333
```

The p-value is less than the significance level 5%. So we can reject null hypothesis. From the model and output of chi-test, we can say that if the loans applied on January will receive less interest rate than other months.

Part 5 - Conclusion

Dataset from Lending club is an interesting dataset. It is offen very difficult to get the insights of interest rate from Bank. This analysis provides interesting information about the interest rate which we get from Lending club for each person.

The interest rate which we receive depends on the various factors like FICO score, Homeownership, Purpose of loan, Team length of loan, loan amount requested, Annual income, Employee length, Issue month, Previous bankrupcies and Debt to income ratio.

If a person is wanting to get a good interest rate then he need to focus on above factors before applying for a lending club loan.