Data Analysis Project 1: Lending Club Loan Rate Criteria

Define the question

Our sample data set consists of 2,500 peer-to-peer loans issued through the Lending Club. The interest rate of these loans is determined by the Lending Club on the basis of characteristics of the person asking for the loan such as their employment history, credit history, and creditworthiness scores.

Objective

The purpose of this analysis is to identify and quantify associations between the interest rate of a loan and the other variables in the data set. In particular, we will consider whether any of these variables have an important association with interest rate after taking into account the applicant's FICO score. For example, if two people have the same FICO score, can the other variables explain a difference in interest rate between them?

Data

Define the ideal data set

Lending Club has been in business since 2007 providing investors and borrowers with an alternative to the traditional banking model. As of January 7, 2013, over 47,648 investment accounts have funded over $1.3 trillion in loans and received over $103,315,600 in interest payments. Access to the entire corpus of loans funded to date since June 2007 would be ideal for this analysis. Yet it will be interesting to see if our analysis on the sample set mimics the claims Lending Club has made pertaining to historical characteristics of qualified borrowers.

Determine what data you can access

The sample data for this exercise is accessible from the Amazon Cloud:

- CSV Format
- RDA Format

The sample data set is accompanied with a code book

Getting Started

Project Setup

We have established a root project folder, *dap1*, that contains the following sub-directories:

- assignment:

- code: organizing sandbox and finalize code for reproducibility
- data: storing sample sets and project workspaces (.rda)
- writing: dedicated for the text artifacts of this endeavor

Obtain the data

```
## Setup working environment and download sample data from Amazon EC2
setwd("~/Activity/Education/Coursera/dataanalaysis/dap1")
fileURL = "https://spark-
public.s3.amazonaws.com/dataanalysis/loansData.csv"
localFile = "./data/loansData.csv"
download.file(fileURL, localFile, method = "curl")
loans.df <- read.csv(localFile)
## Supress warning messages
options(warn = -1)
```

Handle required files and libraries

```
## Load Utility Functions
incFile <-
"~/Activity/Education/Coursera/dataanalaysis/dap1/code/finalcode/loanClubUtils.R"

fileHandle <- file(incFile, open = "r")
source(fileHandle)
close(fileHandle)
## Load Package Dependencies Utility Functions
library(lattice)
library(stringr)
```

Tidy Data Check

Our sample set conforms to most of the Tidy Data requirements:

- Variables in columns
- Observations in rows
- Tables holding elements of only one kind
- Column names are easy to use and informative
- Row names are easy to use and informative
- Variable values are internally consistent

However, we do need to address a few items:

- Mistakes in the data need to be removed
- Appropriate transformed variables have yet to be added

Remove noise

Out of 2500 loan records two (2) loans contained missing data, namely rows 367 and 1595. We removed these records from our sample set leaving a total of 2498 records for our analysis.

```
dim(loans.df)
```

```
## [1] 2500   14
```

```
cleanData <- complete.cases(loans.df)
sum(!cleanData)
```

```
## [1] 2
```

```
which(!cleanData)
```

```
## [1]  367 1595
```

```
loans.approved <- loans.df[cleanData, ]
dim(loans.approved)
```

```
## [1] 2498   14
```

Extract, Transform and Load

We need to perform some data munging in order to get some of our qualitative data into quantitative format for downstream exploration in the analysis pipeline.

Convert Interest.Rate(String) to ETL.Rate(Numeric)

```
etl.Rate <- data.frame(as.numeric(sub("%", "",
loans.approved$Interest.Rate)))
colnames(etl.Rate) <- c("ETL.Rate")
loans.approved <- cbind(loans.approved, etl.Rate)
```

Convert Debit to Income Ratio(String) to ETL.RatioDTI(Numeric)

```
etl.DTI <- data.frame(as.numeric(sub("%", "",
loans.approved$Debt.To.Income.Ratio)))
colnames(etl.DTI) <- c("ETL.RatioDTI")
loans.approved <- cbind(loans.approved, etl.DTI)
```

Since our FICO value per loan observation is given in a range, we will compute the mean FICO score per borrower so that we can group observations using quantitative values.

```
rangeMean <- function(x) {
    t <- strsplit(x, "-")
    mean(as.numeric(unlist(t)))
}
fico.list <- sapply(X = as.vector(loans.approved$FICO.Range), FUN =
rangeMean)
etl.MeanFICO <- data.frame(fico.list)
colnames(etl.MeanFICO) <- c("ETL.MeanFICO")
loans.approved <- cbind(loans.approved, etl.MeanFICO)
```

Exploratory Data Analysis

Question 1

Lending Club claims that their average FICO score for quality borrowers is 708. Is this still valid for our sample set?

```
round(mean(loans.approved$ETL.MeanFICO))
```

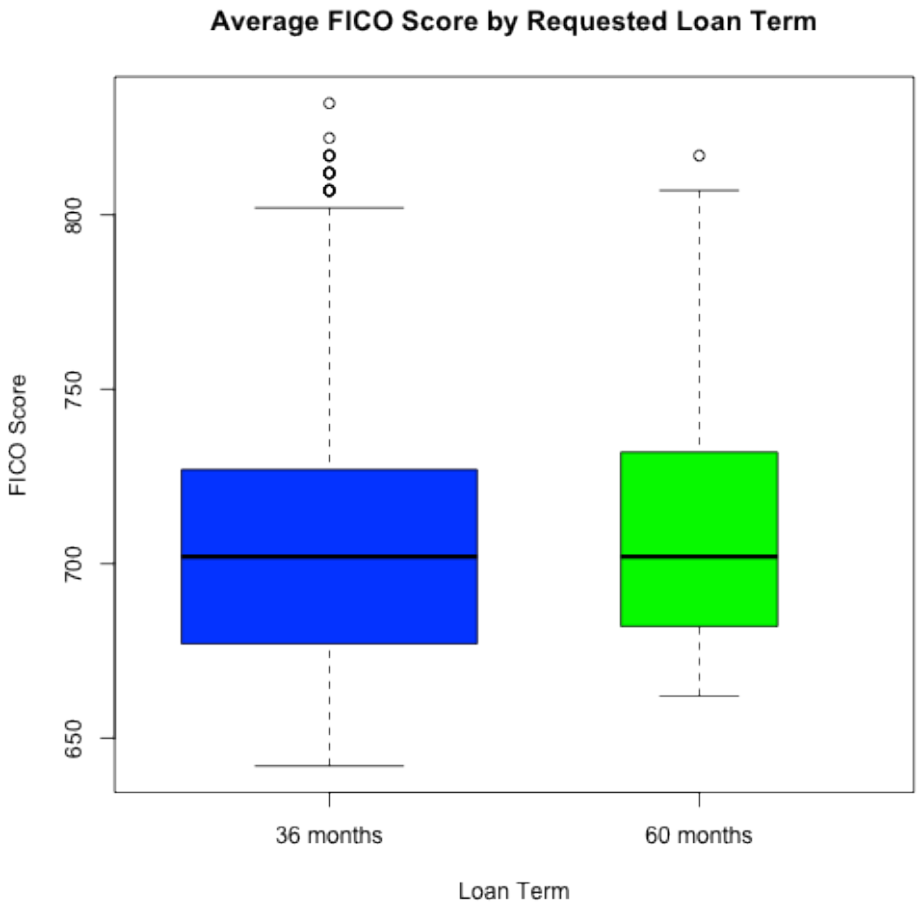
```
## [1] 708
```

Observation: It seems that our sample set is very representative of their historical records.

Question 2

Is there enough evidence in a relationship between interest rate and loan term to justify further exploration into the term variable?

```
boxplot(loans.approved$ETL.MeanFICO ~
as.factor(loans.approved$Loan.Length),
    col = c("blue", "green"), varwidth = TRUE, main = "Average FICO Score
by Requested Loan Term",
    xlab = "Loan Term", ylab = "FICO Score")
```

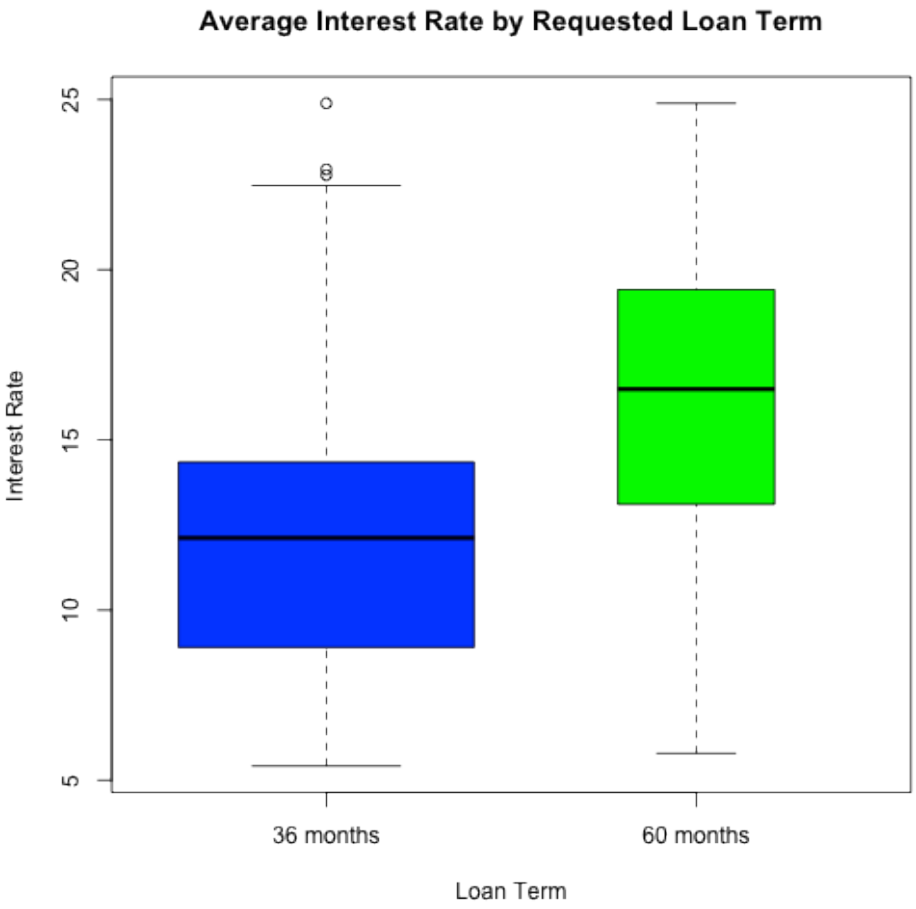


Observation: While there were more requests for 36 month loans, both loan types maintained a FICO range of 675-725 and a mean score of 708. As such, there seems to be no relationship between loan term and FICO score.

Question 3

What is the distribution of Interest Rate with respect to loan term?

```
boxplot(loans.approved$ETL.Rate ~ as.factor(loans.approved$Loan.Length),
col = c("blue",
    "green"), varwidth = TRUE, main = "Average Interest Rate by Requested
Loan Term",
    xlab = "Loan Term", ylab = "Interest Rate")
```

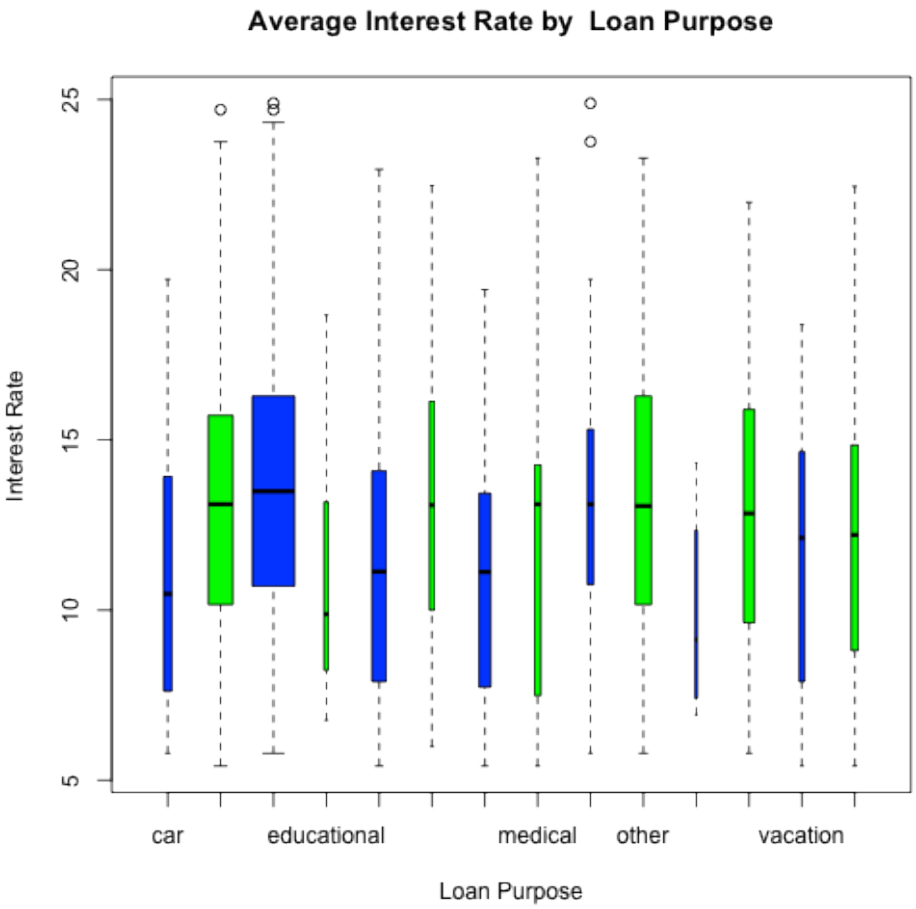


Observation: It seems that that loan term alone is a key factor on interest rate, regardless of FICO score. The mean interest rates for our two categories of loans terms, 36 and 60, are ~12.5% and ~17%, respectively.

Question 4

Is there enough evidence in a relationship between interest rate and loan purpose to justify further exploration into the purpose variable?

```
boxplot(loans.approved$ETL.Rate ~ as.factor(loans.approved$Loan.Purpose),
col = c("blue",
    "green"), varwidth = TRUE, main = "Average Interest Rate by  Loan
Purpose",
    xlab = "Loan Purpose", ylab = "Interest Rate", )
```



Observation: There does not seem to be any relationship between rate and purpose.

Question 5

Is there a correlation between the interest rate and the borrower's revolving credit balance?

```
eda.CreditBalance.Cor.df <- data.frame(loans.approved$ETL.Rate,
loans.approved$Revolving.CREDIT.Balance)
eda.CreditBalance.Cor.result <- cor.test(eda.CreditBalance.Cor.df[, 1],
eda.CreditBalance.Cor.df[,
    2])
print(eda.CreditBalance.Cor.result)
```

```
##
##  Pearson's product-moment correlation
##
## data:  eda.CreditBalance.Cor.df[, 1] and eda.CreditBalance.Cor.df[, 2]
## t = 3.059, df = 2496, p-value = 0.002246
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.02194 0.10009
## sample estimates:
##     cor
## 0.06111
```

```
str(eda.CreditBalance.Cor.result)
```

```
## List of 9
##  $ statistic  : Named num 3.06
##   ..- attr(*, "names")= chr "t"
##  $ parameter  : Named int 2496
##   ..- attr(*, "names")= chr "df"
##  $ p.value    : num 0.00225
##  $ estimate   : Named num 0.0611
##   ..- attr(*, "names")= chr "cor"
##  $ null.value : Named num 0
##   ..- attr(*, "names")= chr "correlation"
##  $ alternative: chr "two.sided"
##  $ method     : chr "Pearson's product-moment correlation"
##  $ data.name  : chr "eda.CreditBalance.Cor.df[, 1] and
eda.CreditBalance.Cor.df[, 2]"
##  $ conf.int   : atomic [1:2] 0.0219 0.1001
##   ..- attr(*, "conf.level")= num 0.95
##  - attr(*, "class")= chr "htest"
```

Observation: A correlation can only indicate the presence or absence of a relationship, not the nature of the relationship. This correlation analysis yields:

- There exists a slight correlation of only 6% between rate and balance of revolving credit.
- The correlation is positive implying that both variables move in the same direction. For example, if one increases so does the other.
- The p-value of 0.002 implies the statistical significance of the presences of a correlation.
- However, the effective size ($r^{2)}$ of <1% indicates 99% of the variations can not be explained by the relationship.

Question 6

Is there a correlation between the interest rate and the number of open credit lines?

```
eda.CreditLines.Cor.df <- data.frame(loans.approved$ETL.Rate,
loans.approved$Open.CREDIT.Lines)
eda.CreditLines.Cor.result <- cor.test(eda.CreditLines.Cor.df[, 1],
eda.CreditLines.Cor.df[,
    2])
print(eda.CreditLines.Cor.result)
```

```
##
##  Pearson's product-moment correlation
##
## data:  eda.CreditLines.Cor.df[, 1] and eda.CreditLines.Cor.df[, 2]
## t = 4.53, df = 2496, p-value = 6.169e-06
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.05127 0.12907
## sample estimates:
##     cor
## 0.09031
```

```
str(eda.CreditLines.Cor.result)
```

```
## List of 9
##  $ statistic  : Named num 4.53
##   ..- attr(*, "names")= chr "t"
##  $ parameter  : Named int 2496
##   ..- attr(*, "names")= chr "df"
##  $ p.value    : num 6.17e-06
##  $ estimate   : Named num 0.0903
##   ..- attr(*, "names")= chr "cor"
##  $ null.value : Named num 0
##   ..- attr(*, "names")= chr "correlation"
##  $ alternative: chr "two.sided"
##  $ method     : chr "Pearson's product-moment correlation"
##  $ data.name  : chr "eda.CreditLines.Cor.df[, 1] and
eda.CreditLines.Cor.df[, 2]"
##  $ conf.int   : atomic [1:2] 0.0513 0.1291
##   ..- attr(*, "conf.level")= num 0.95
##  - attr(*, "class")= chr "htest"
```

Observation: This correlation analysis yields:

- There exists a slight correlation of only 9% between rate and number of open credit lines.
- The correlation is positive implying that both variables move in the same direction. For example, if one increases so does the other.
- The p-value of 6.17e-06 implies the statistical significance of the presences of a correlation.
- However, the effective size ($r^{2)}$ of <1% indicates 99% of the variations can not be explained by the relationship.

Question 7

Is there a correlation between the interest rate and monthly income?

```
eda.MonthlyIncome.Cor.df <- data.frame(loans.approved$ETL.Rate,
loans.approved$Monthly.Income)
eda.MonthlyIncome.Cor.result <- cor.test(eda.MonthlyIncome.Cor.df[, 1],
eda.MonthlyIncome.Cor.df[,
    2])
print(eda.MonthlyIncome.Cor.result)
```

```
##
##  Pearson's product-moment correlation
##
## data:  eda.MonthlyIncome.Cor.df[, 1] and eda.MonthlyIncome.Cor.df[, 2]
## t = 0.6456, df = 2496, p-value = 0.5186
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.02631  0.05211
## sample estimates:
##     cor
## 0.01292
```

```
str(eda.MonthlyIncome.Cor.result)
```

```
## List of 9
##  $ statistic  : Named num 0.646
##   ..- attr(*, "names")= chr "t"
##  $ parameter  : Named int 2496
##   ..- attr(*, "names")= chr "df"
##  $ p.value    : num 0.519
##  $ estimate   : Named num 0.0129
##   ..- attr(*, "names")= chr "cor"
##  $ null.value : Named num 0
##   ..- attr(*, "names")= chr "correlation"
##  $ alternative: chr "two.sided"
##  $ method     : chr "Pearson's product-moment correlation"
##  $ data.name  : chr "eda.MonthlyIncome.Cor.df[, 1] and
eda.MonthlyIncome.Cor.df[, 2]"
##  $ conf.int   : atomic [1:2] -0.0263 0.0521
##   ..- attr(*, "conf.level")= num 0.95
##  - attr(*, "class")= chr "htest"
```

Observation: This correlation analysis yields:

- There exists a slight correlation of only 1% between rate and monthly income.
- The correlation is positive implying that both variables move in the same direction. For example, if one increases so does the other.
- The p-value of 0.519 implies no statistical significance of the presences of a correlation.
- The effective size ($r^{2)}$ of 3% indicates 97% of the variations can not be explained by the relationship.

Question 8

Is there a correlation between the interest rate and debit to income ratio?

```
eda.RatioDTI.Cor.df <- data.frame(loans.approved$ETL.Rate,
loans.approved$ETL.RatioDTI)
eda.RatioDTI.Cor.result <- cor.test(eda.RatioDTI.Cor.df[, 1],
eda.RatioDTI.Cor.df[,
    2])
print(eda.RatioDTI.Cor.result)
```

```
##
##  Pearson's product-moment correlation
##
## data:  eda.RatioDTI.Cor.df[, 1] and eda.RatioDTI.Cor.df[, 2]
## t = 8.734, df = 2496, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1339 0.2100
## sample estimates:
##    cor
## 0.1722
```

```
str(eda.RatioDTI.Cor.result)
```

```
## List of 9
##  $ statistic  : Named num 8.73
##   ..- attr(*, "names")= chr "t"
##  $ parameter  : Named int 2496
##   ..- attr(*, "names")= chr "df"
##  $ p.value    : num 0
##  $ estimate   : Named num 0.172
##   ..- attr(*, "names")= chr "cor"
##  $ null.value : Named num 0
##   ..- attr(*, "names")= chr "correlation"
##  $ alternative: chr "two.sided"
##  $ method     : chr "Pearson's product-moment correlation"
##  $ data.name  : chr "eda.RatioDTI.Cor.df[, 1] and eda.RatioDTI.Cor.df[,
2]"
##  $ conf.int   : atomic [1:2] 0.134 0.21
##   ..- attr(*, "conf.level")= num 0.95
##  - attr(*, "class")= chr "htest"
```

Observation: This correlation analysis yields:

- There exists a slight correlation of only 17% between rate and DTI.
- The correlation is positive implying that both variables move in the same direction. For example, if one increases so does the other.
- The p-value of 0 implies the statistical significance of the presences of a correlation.
- The effective size ($r^{2)}$ of <1% indicates 99% of the variations can not be explained by the relationship.

Question 9

While there does not seem to be any strong relationship between variables and the interest rate when looking across the entire corpus, could there be more significant correlations when we look at subsets of loans where the borrowers had the same mean FICO?

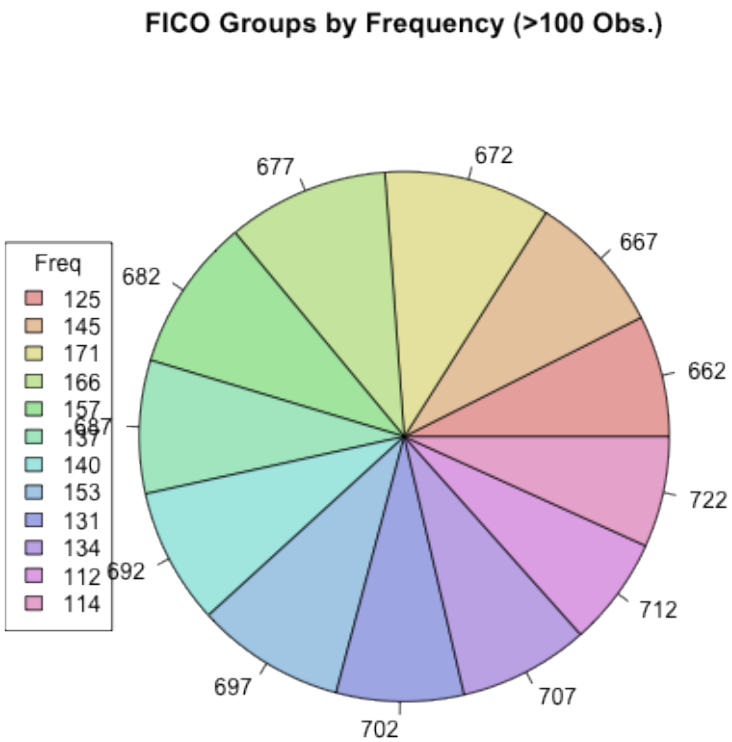
Lets look at distribution of loans per mean FICO.

```
eda.FICO.groups <- table(loans.approved$ETL.MeanFICO)
print(eda.FICO.groups)
```

```
##
## 642 647 652 657 662 667 672 677 682 687 692 697 702 707 712 717 722 727
##   5   3   1   4 125 145 171 166 157 137 140 153 131 134 112  93 114  94
## 732 737 742 747 752 757 762 767 772 777 782 787 792 797 802 807 812 817
##  94  65  53  54  61  46  46  36  17  22  28  19  20  13  12  11   8   6
## 822 832
##   1   1
```

Action: Lets test the our theory by looking at FICO groups that had at least 100 loans and retry some of the correlation tests to see if they increase from the baseline (looking across the entire corpus ignoring FICO score).

```
## Identify the subset of Mean FICO Groups with at least 100 loans
eda.FICO.groups.df <- as.data.frame(eda.FICO.groups)
eda.FICO.groups.OfInterest <- subset(eda.FICO.groups.df, Freq >= 100)
eda.Colors = rainbow(length(eda.FICO.groups.OfInterest$Freq), s = 0.3, v =
0.9,
    start = 0, end = 0.9)
pie(eda.FICO.groups.OfInterest$Freq, labels =
as.vector(eda.FICO.groups.OfInterest$Var1),
    main = "FICO Groups by Frequency (>100 Obs.)", col = eda.Colors)
legend("left", legend = eda.FICO.groups.OfInterest$Freq, title = "Freq",
ncol = 1,
    fill = eda.Colors, xjust = 1)
```



```
max(eda.FICO.groups.OfInterest$Freq)
```

```
## [1] 171
```

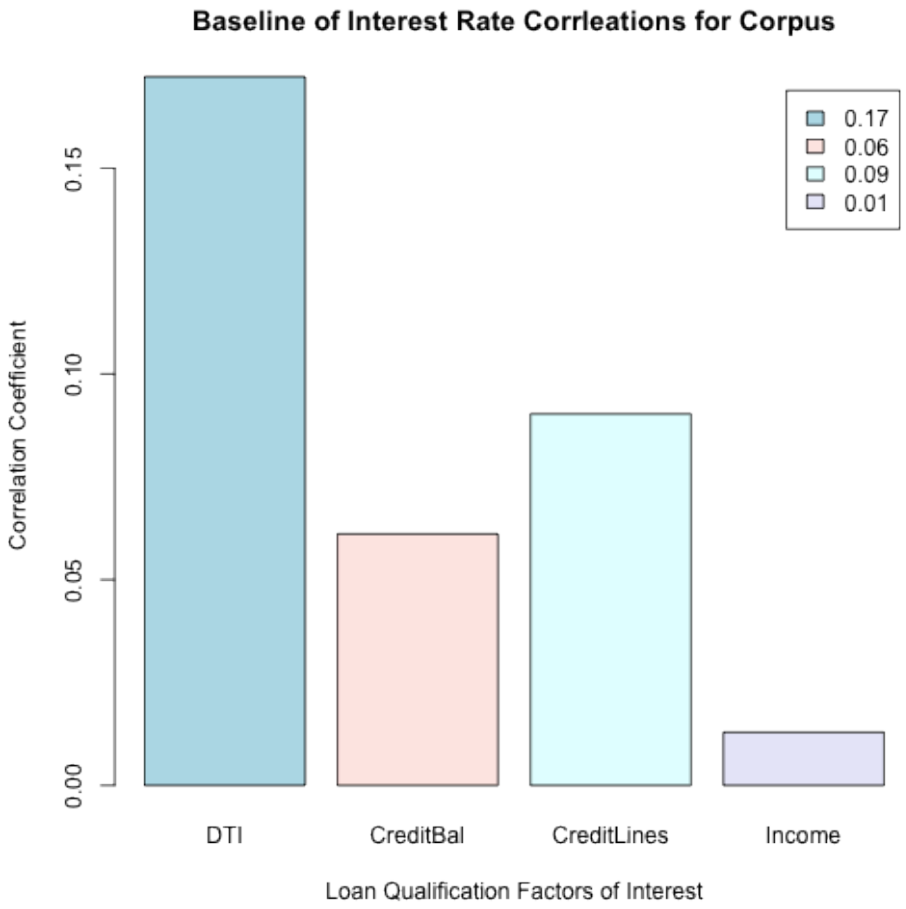
```
min(eda.FICO.groups.OfInterest$Freq)
```

```
## [1] 112
```

Observation: We have narrowed our focus down to 12 groups of FICO scores where there range of observations (loans) per group is between 112 and 171.

```
## Generate all correlation coefficients for all baseline observations
eda.baseline.correlations <- t(sapply(X = 1, FUN = baselineCorrelations))
colnames(eda.baseline.correlations) <- c("Rate.to.DTI",
"Rate.to.CreditBal",
    "Rate.to.CreditLines", "Rate.to.Income")

## Plot results for Baseline Correlations
eda.baseline.cordata = t(eda.baseline.correlations)
barplot(eda.baseline.correlations[1, ], names.arg = c("DTI", "CreditBal",
"CreditLines",
    "Income"), main = "Baseline of Interest Rate Corrleations for Corpus",
xlab = "Loan Qualification Factors of Interest",
    ylab = "Correlation Coefficient", legend = round(eda.baseline.cordata[,
        1], 2), col = c("lightblue", "mistyrose", "lightcyan", "lavender"))
```



Observation: When looking across the entire corpus of loans (ignoring FICO scores), Monthly Income seems to present the least evidence of a relationship to Interest Rate as compared to correlations for DTI, Credit Balance and Quantity of Credit Lines.

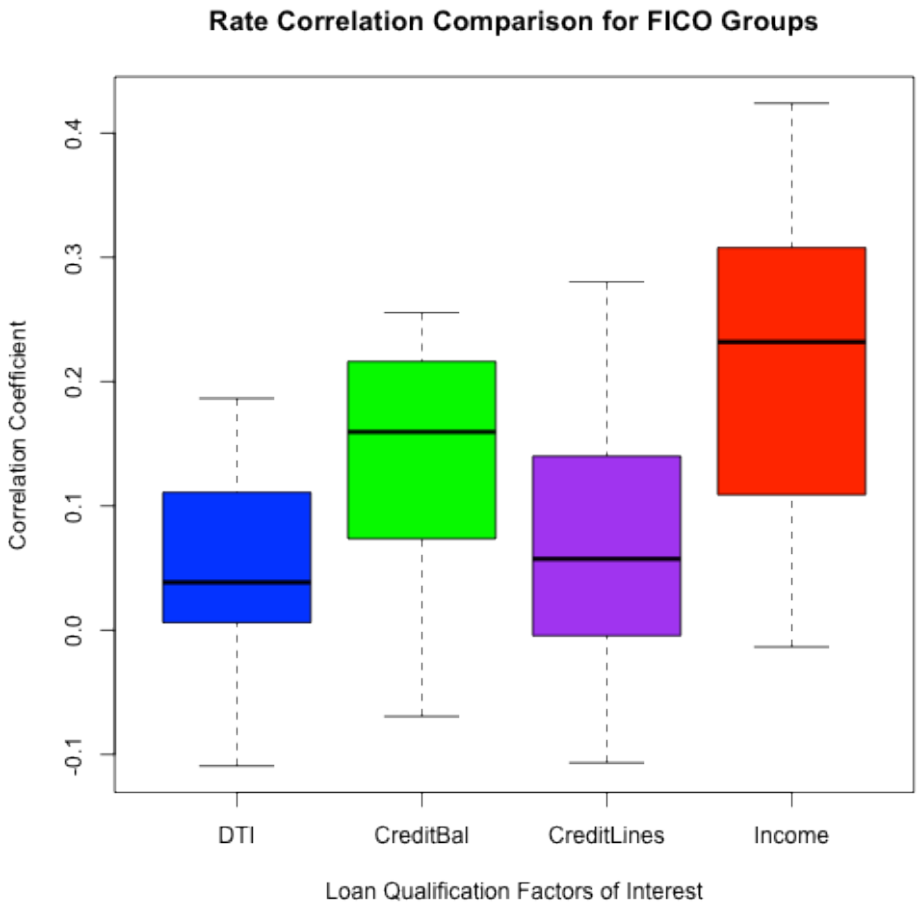
Actions: Lets generate correlation coefficients for all FICO Group observations for each of the loan qualification factors of interest.

```
## Generate all correlation coefficients for all FICO Group observations
eda.FICO.groups.allCorrelations <- t(sapply(X =
as.vector(eda.FICO.groups.OfInterest$Var1),
    FUN = groupCorrelations))
colnames(eda.FICO.groups.allCorrelations) <- c("Rate.to.DTI",
"Rate.to.CreditBal",
    "Rate.to.CreditLines", "Rate.to.Income", "MaxMean.State",
"MaxMean.Rate",
    "MaxMean.Income", "MinMean.State", "MinMean.Rate", "MinMean.Income")

## Plot results for FICO Group Correlations
eda.FICO.groups.corData <-
data.frame(as.numeric(eda.FICO.groups.allCorrelations[,
    1]), as.numeric(eda.FICO.groups.allCorrelations[, 2]),
as.numeric(eda.FICO.groups.allCorrelations[,
    3]), as.numeric(eda.FICO.groups.allCorrelations[, 4]))
colnames(eda.FICO.groups.corData) <- c("Rate.to.DTI", "Rate.to.CreditBal",
"Rate.to.CreditLines",
    "Rate.to.Income")
print(eda.FICO.groups.corData)
```

```
##    Rate.to.DTI Rate.to.CreditBal Rate.to.CreditLines Rate.to.Income
## 1       0.1243            0.0507              0.1216         0.1889
## 2       0.0977           -0.0694              0.0463         0.0341
## 3      -0.0248            0.2309              0.0636         0.3055
## 4       0.1865            0.1883              0.2801         0.2208
## 5       0.1383            0.1469              0.1586        -0.0135
## 6       0.0114            0.2017              0.0049         0.2431
## 7       0.0933            0.2555              0.0628         0.3215
## 8       0.0486            0.1253              0.0518         0.3103
## 9      -0.1093            0.2499              0.1689         0.4240
## 10      0.0005            0.1721             -0.0140         0.2470
## 11      0.0283           -0.0645             -0.1067         0.0646
## 12      0.0229            0.0963             -0.0828         0.1531
```

```
boxplot(eda.FICO.groups.corData, col = c("blue", "green", "purple", "red"),
    varwidth = TRUE, names = c("DTI", "CreditBal", "CreditLines",
"Income"),
    main = "Rate Correlation Comparison for FICO Groups", xlab = "Loan
Qualification Factors of Interest",
    ylab = "Correlation Coefficient")
```



Observation: When we rerun our correlation tests on the same loan variables for the 12 FICO Groups, Monthly Income seems to present the most evidence of a relationship to Interest Rate as compared to correlations for DTI, Credit Balance and Quantity of Credit Lines. This motivates us to explore the potential for a relationship between Interest Rate and Monthly Income after FICO score has been considered.

There seems to be evidence for an impact of Monthly Income on individuals with equal FICO scores.

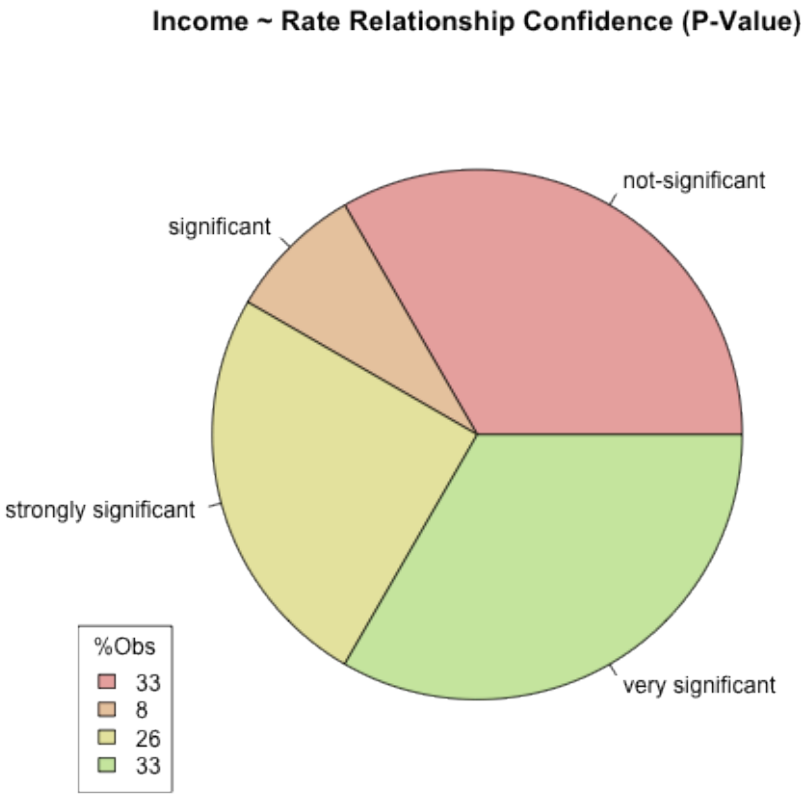
- When we compare the impact of 4 loan factors to the entire corpus, Monthly Income has the least potential for a relationship to Interest Rate while DTI has the greatest. Nevertheless, none of the factors by themselves demonstrate evidence of any significant relationship.
- Our baseline analysis across the entire corpus yielded a correlation coefficient of 1%. This implies an absence of a relationship to interest rates when the FICO scores vary.
- When we compare loans that have the same FICO score (minimum of 100 per observation), we see a change in the impact of Monthly Income to Interest Rate. There is a 22% mean correlation that Monthly Income has a relationship to the loan interest rate.
- We can infer that Monthly Income has a stronger impact on interest rate after FICO score than other loan factors.
- This observation is inline with claims from Lending Club whereby they state that the average borrower requires $69,924 personal annual income. In other words, the higher a borrowers income the more likely they are to qualify. However, this does not mean that income has a direct correlation to to interest rate.

Action: Let us isolate our exploration on Monthly Income. Let's measure the statistical significance of a relationship between Monthly Income and Interest Rate on each of the 12 FICO Groups.

```
## Generate Monthly Income correlation coefficients for all FICO Group
## observations
eda.FICO.groups.correlations <- t(sapply(X =
as.vector(eda.FICO.groups.OfInterest$Var1),
    FUN = groupIncomeCorrelations))
```

```
## [1] "0.188917889280368"  "0.0348595923144712" "significant"
## [1] "0.0341078912247989" "0.683804417649384"  "not-significant"
## [1] "0.305451655910619"    "4.85447539553263e-05" "very significant"
## [1] "0.220815930114509"    "0.00425122260055866"  "strongly significant"
## [1] "-0.0134739815837243" "0.866986891254902"   "not-significant"
## [1] "0.24311602993638"     "0.00420173277444835"  "strongly significant"
## [1] "0.321526840265477"    "0.000107229434847866" "very significant"
## [1] "0.310341557697412"    "9.45868294119911e-05" "very significant"
## [1] "0.424023319846619"    "4.49135040048176e-07" "very significant"
## [1] "0.246973163108538"    "0.00401678320702437"  "strongly significant"
## [1] "0.0645702186761655" "0.498794830927008"  "not-significant"
## [1] "0.153128828435626" "0.103830958402985" "not-significant"
```

```
colnames(eda.FICO.groups.correlations) <- c("Rate.to.Income", "p-Value",
"Confidence",
    "MaxMean.State", "MaxMean.Rate", "MaxMean.Income", "MinMean.State",
"MinMean.Rate",
    "MinMean.Income")
eda.FICO.groups.corMI.confidence <-
as.data.frame(table(eda.FICO.groups.correlations[,
    3]))
confidenceRates <- c(round(eda.FICO.groups.corMI.confidence$Freq[1]/12, 2)
*
    100, round(eda.FICO.groups.corMI.confidence$Freq[2]/12, 2) * 100,
round(eda.FICO.groups.corMI.confidence$Freq[3]/12,
    2) * 100 + 1, round(eda.FICO.groups.corMI.confidence$Freq[4]/12, 2) *
100)
eda.Colors = rainbow(length(eda.FICO.groups.OfInterest$Freq), s = 0.3, v =
0.9,
    start = 0, end = 0.9)
pie(eda.FICO.groups.corMI.confidence$Freq, labels =
as.vector(eda.FICO.groups.corMI.confidence$Var1),
    main = "Income ~ Rate Relationship Confidence (P-Value)", col =
eda.Colors)
legend("bottomleft", legend = confidenceRates, title = "%Obs", ncol = 1,
fill = eda.Colors)
```



Observation: Given our FICO Group segmentation of 12 groups, 67% (8) of the groups yielded a p-value inferring a significant confidence in a relationship between Monthly Income and Interest Rate. Additionally, 33% (4) of the groups yielded a p-value inferring a very significant confidence.

Action: Since there is confidence in an Income to Rate relationship within FICO Groups, can we assume that the relationship is the same across all states? Let's identify for each FICO Group observation the Min and Max State Incomes and Rates.

Question 10

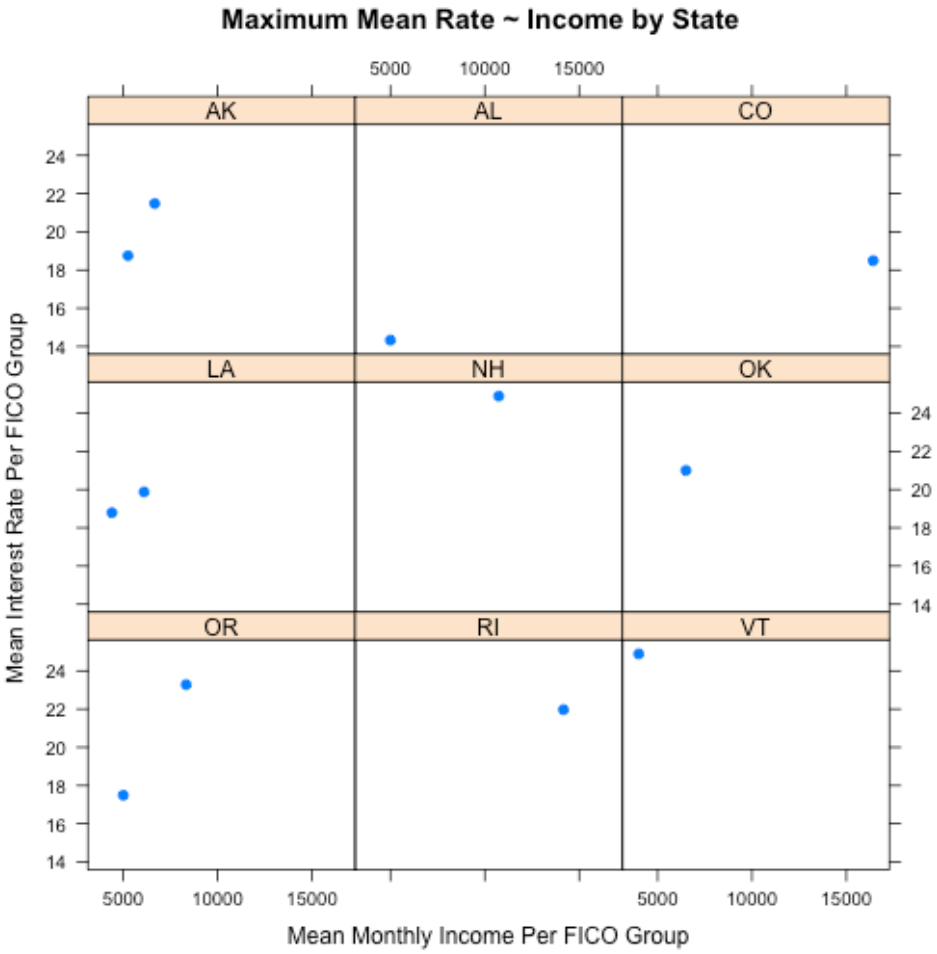
Does state of residence (a qualitative loan variable) also have an impact on Interest Rate?

```
## Create new Dataframe from Matrix
eda.FICO.groups.StateData <- data.frame(eda.FICO.groups.correlations[, 4],
as.numeric(eda.FICO.groups.correlations[,
    5]), as.numeric(eda.FICO.groups.correlations[, 6]),
eda.FICO.groups.correlations[,
    7], as.numeric(eda.FICO.groups.correlations[, 8]),
as.numeric(eda.FICO.groups.correlations[,
    9]))
colnames(eda.FICO.groups.StateData) <- c("MaxMean.State", "MaxMean.Rate",
"MaxMean.Income",
    "MinMean.State", "MinMean.Rate", "MinMean.Income")
print(eda.FICO.groups.StateData)
```

```
##     MaxMean.State MaxMean.Rate MaxMean.Income MinMean.State MinMean.Rate
## 662            NH        24.89          10738            MN        13.93
## 667            VT        24.89           4000            NC        13.86
## 672            LA        19.88           6101            MI        13.04
## 677            OR        23.28           8333            NM        12.12
## 682            LA        18.79           4400            WY        11.11
## 687            OK        21.00           6500            IN        10.59
## 692            RI        21.97          14167            AL        11.93
## 697            AK        21.48           6667            MO        10.16
## 702            CO        18.49          16417            MO         7.66
## 707            OR        17.49           5000            MN         9.91
## 712            AK        18.75           5250            LA         9.76
## 722            AL        14.33           5000            DC         6.62
##     MinMean.Income
## 662           6000
## 667           8215
## 672           3625
## 677           1312
## 682           2253
## 687           2083
## 692           3229
## 697           3500
## 702           6500
## 707           3667
## 712           5417
## 722           8161
```

```
# Generate Plotting Variabes
maxrate <- eda.FICO.groups.StateData$MaxMean.Rate
maxincome <- eda.FICO.groups.StateData$MaxMean.Income
state <- factor(eda.FICO.groups.StateData$MaxMean.State)

# Plot data
xyplot(maxrate ~ maxincome | state, data = eda.FICO.groups.StateData, main
= "Maximum Mean Rate ~ Income by State",
    xlab = "Mean Monthly Income Per FICO Group", ylab = "Mean Interest Rate
Per FICO Group",
    pch = 19, as.table = TRUE)
```



Observation: Given our 12 FICO Groups, 9 our of the 48 represented States accounted for the highest mean Interest Rates per Group. AK, LA and OR each yielded the highest mean rates for 6 FICO Groups. We can assert that the relationship observed in step 9 is different per state.

Statistical prediction/modeling

Correlation Theory P-Values

```
eda.Rate2Income.Relationship.Assessment <-
data.frame(strengthOfRelationship(as.vector(eda.FICO.groups.correlations[,
    1])), directionOfRelationship(as.vector(eda.FICO.groups.correlations[,
1])),
    as.vector(eda.FICO.groups.correlations[, 3]),
predictabilityOfVariation(as.vector(eda.FICO.groups.correlations[,
        1])))
colnames(eda.Rate2Income.Relationship.Assessment) <- c("Strength",
"Direction",
    "Significance", "Predictability")
print(eda.Rate2Income.Relationship.Assessment)
```

```
##    Strength Direction         Significance Predictability
## 1      weak      same          significant           weak
## 2      weak      same      not-significant           weak
## 3      weak      same     very significant           weak
## 4      weak      same strongly significant           weak
## 5      weak  opposite      not-significant           weak
## 6      weak      same strongly significant           weak
## 7      weak      same     very significant           weak
## 8      weak      same     very significant           weak
## 9      weak      same     very significant           weak
## 10     weak      same strongly significant           weak
## 11     weak      same      not-significant           weak
## 12     weak      same      not-significant           weak
```

Observation: Our corpus of observations, namely 12 FICO Groups each having a minimum of 100 approved loans yields the following evidence for a correlation between Monthly Income and Interest Rate for borrowers with similar FICO scores:

- 100% (12 out or 12) of our observations yielded the presence of a weak relationship since they all had correlation coefficients below 0.5.
- 92% (11 out of 12) of our observations yielded a positive correlation implying that both variables move in the same direction. For example, if one increases so does the other.
- 67% (8 out of 12) yielded a p-value < 0.05 which implies the statistical significance of the presences of a relationship.
- 100% (12 out of 12) of the observations yielded an effective size ($r^{2)}$ below 0.5 which indicates that this relationship can only predict <50% of the variations between Monthly Income and Interest Rate.

Interpret results

Our correlation analysis suggests that:

- there is a significant probability of relationship between Monthly Income and Interest Rates
- the relationship tends to cause both variables to move in the same direction
- the impact of this relationship by itself on the offering of a specific interest rate is not very strong. In other words, there are other mitigating factors, such as State of residence.

It would seem that Lending Club members use Monthly Income as a possible secondary factor for qualifying borrowers after they consider FICO Scores. The rates assigned per borrowers with similar FICO scores will differ based on Income thresholds that differ per State.

Challenge results

While statistical significance exists for a relationship between Monthly Income and Interest Rates within States, there is no evidence as to the exact qualification rules being applied by Lending Club members. All we can infer is that borrowers seeking a loans from a lender in one State will be judged by a orthogonal set of rules by a lender in another State.

Finalize Report

- Synthesize/write up results in a separate document using the following outline
  - Methods
    - Data Collection
    - Exploratory Analysis
    - Statistical Modeling
    - Reproducibility
  - Results
  - Conclusions
  - References