SML 201 Problem Set 2

Bill Haarlow & Weston Carpenter

2019-10-23

Problem set 2 is due by 11:59pm on Wednesday October 9. Please submit both a .Rmd and a .pdf file on Blackboard by the deadline **and** drop off a hard copy of the pdf file at 26 Prospect Avenue by 5pm of the **next day** of the due date. To look for the drop-off cabinet, after you enter the building turn to the left to enter the lounge area and the file cabinet is to your right with an open slot with the label "SML 201 Homework"; note that the building might be locked after 6pm and on the weekends. You are also welcome to bring your PDF copy to any lecture **before** the deadline and I will drop off the copy for you.

This problem set can be completed in groups of up to 3 students. It is okay to work by yourself, if this is preferable. You are welcome to get help (you can either ask questions on Piazza or talk to instructors in person during office hours) from instructors; however, please do not post code/solutions on Piazza on a public post.

When working in a group it is your responsibility to make sure that you are satisfied with all parts of the report and the submission is on time (e.g., we will not entertain arguments that deficiencies are the responsibility of other group members). We expect that you each work independently first and then compare your answers with each other once you all finish, or you all work together on your own laptops. Failing to make contributions and then putting your name on a report will be considered a violation of the honor code. **Please do not divide work among group mates.** Everyone in your group is responsible for being able to explain the solutions in the report.

For all parts of this problem set, you **MUST** use R commands to print the output as part of your R Markdown file. You are not permitted to find the answer in the R console and then copy and paste the answer into this document.

If you are completing this problem set in a group, please have only **one** person in your group turn in the .Rmd and .pdf files; the other person in your group should turn in the list of the people in your group in the *Text Submission* field on the submission page. This means that **everyone should make a submission**–either a file-upload or a text submission–regardless of whether you are working in a group or not.

Please type your name(s) after "Digitally signed:" below the honor pledge to serve as digital signature(s). Put the pledge and your signature(s) at the beginning of each document that you turn in.

> I pledge my honor that I have not violated the honor code when completing this assignment.

Digitally signed: Bill Haarlow & Weston Carpenter

In order to receive full credits, please have sensible titles and axis labels for all your graphs and adjust values of the relevant graphical parameters so that your plots are informative. Also, all answers must be written in complete sentences.

```
library(ggplot2)
```

Background info on Lending Club

In this problem set we will explore a subset of a dataset provided by Lending Club (https://www.lendingclub.com/). The dataset includes loan cases from 2008-2015. To produce graphs for this report you are free to use the basic `graphics` functions (i.e., the functions covered by Ch 2.2-2.3 in the lecture notes) in `R` as well as the functions in the `ggplot2` package that will be covered on Tuesday October 8.

The company

Lending Club is a US peer-to-peer lending company, headquartered in San Francisco, California. It was the first peer-to-peer lender to register its offerings as securities with the Securities and Exchange Commission (SEC), and to offer loan trading on a secondary market. Lending Club operates an online lending platform that enables borrowers to obtain a loan, and investors to purchase notes backed by payments made on loans. Lending Club is the world's largest peer-to-peer lending platform. The company claims that $ 15.98 billion in loans had been originated through its platform up to December 31, 2015.

(Ref: https://en.wikipedia.org/wiki/Lending_Club)

How it works

This is how Lending Club works (the steps below were taken from Lending Club's website):

- Customers interested in a loan complete a simple application at LendingClub.com
- [Lending Club] leverage[s] online data and technology to quickly assess risk, determine[s] a credit rating and assign[s] appropriate interest rates. Qualified applicants receive offers in just minutes and can evaluate loan options with no impact to their credit score
- Investors ranging from individuals to institutions select loans in which to invest and can earn monthly returns

You can read the details on https://www.lendingclub.com/public/how-peer-lending-works.action

Objectives of this problem set

In this problem set we would like to answer these questions:

- Why do people want to borrow?
- How big a loan do people usually apply for?
- Why does Lending Club welcome risky loans?

(9 pts) (3 pts each: code runs, code has annotations, answers are in complete sentences)

(1 pt) (Did not print out large datasets that serve no purpose for the report)

Read the data from the *loan_data.csv* file into Rtudio; name this object `loancase`. Make sure to check that you have 887,379 rows and 5 columns for `loancase`. Perform the usual steps that we did in lecture to get familiar with the dataset; e.g., you should find out the object type of `loancase`, the data types of the elements in each column and the summary statistics for each column etc. Also, you should take a look at the first and the last few rows of the dataset to get an idea on the structure of the dataset.

Below are the description for the variables in `loancase`:

- id - A unique Lending Club assigned ID for the loan listing
- funded_amnt - The total approved loan amount for the applicant at the time when the data was recored
- grade - Lending Club assigned loan grade
- int_rate - Interest Rate on the loan
- purpose - A category provided by the borrower for the loan request

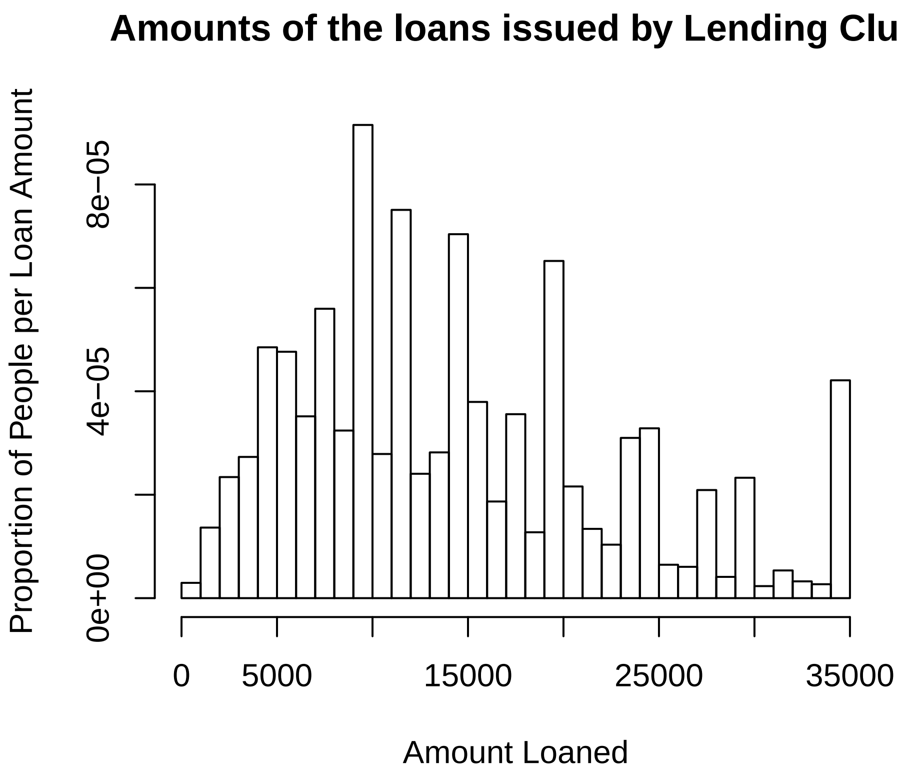
```r
# Reads in .csv file
loancase = read.csv(file = "/Users/billhaarlow/Desktop/SML201/Pset2/loan_data.csv")
class(loancase)  # Ouputs `loancase` object type
str(loancase)  # Outputs the structure of `loancase`
summary(loancase)  # Outputs numerical summaries for `loancase`
head(loancase, n = 10)  # Outputs the first 10 rows of `loancase`
tail(loancase, n = 10)  # Outputs the last 10 rows of `loancase`
```

Problem 1. The amounts of the loans issued by Lending Club. (20 points)

Part a (10 points)

Make a histogram that displays the distribution of the loan amounts. Make sure that the areas of the bins represent proportions.

```r
# Creates histogram without frequencies
hist(loancase$funded_amnt, freq = F, xlab = "Amount Loaned", ylab = "Proportion of People per Loa
    breaks = 30, main = "Amounts of the loans issued by Lending Club")
```



Part b (10 points)

Describe the shape of the loan amount distribution (i.e., is the distribution symmetric or skewed? If it is skewed, is it left- or right-skewed?). What numbers will you use to summarize the distributions of the sizes of the loans? Justify your choice and report the numbers.

Answer: The shape of the loan amount distribution is right-skewed. To summarize this distribution, we will use a 6 number summary because it is not as easily affected by outliers. These numbers can be found in the printed 6 number summary below:

```
summary(loancase$funded_amnt)   # Summary for extracted `funded_amnt` column
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    500    8000   13000   14742   20000   35000 
```

Problem 2. Why does Lending Club welcome risky loans? (30 points)

A *loan grade* takes into account a combination of factors; these factors include, but not limited to:

- Information provided on the loan application
- Information provided by credit bureaus
- Credit score, which predicts the likelihood that borrowers will make on time payments until loans are fully repaid
- Debt-to-income ratio

- Credit history length, the number of other accounts currently open, and usage and payment history with those accounts
- Recent credit activity, including how many other credit inquiries have been initiated over the past six months

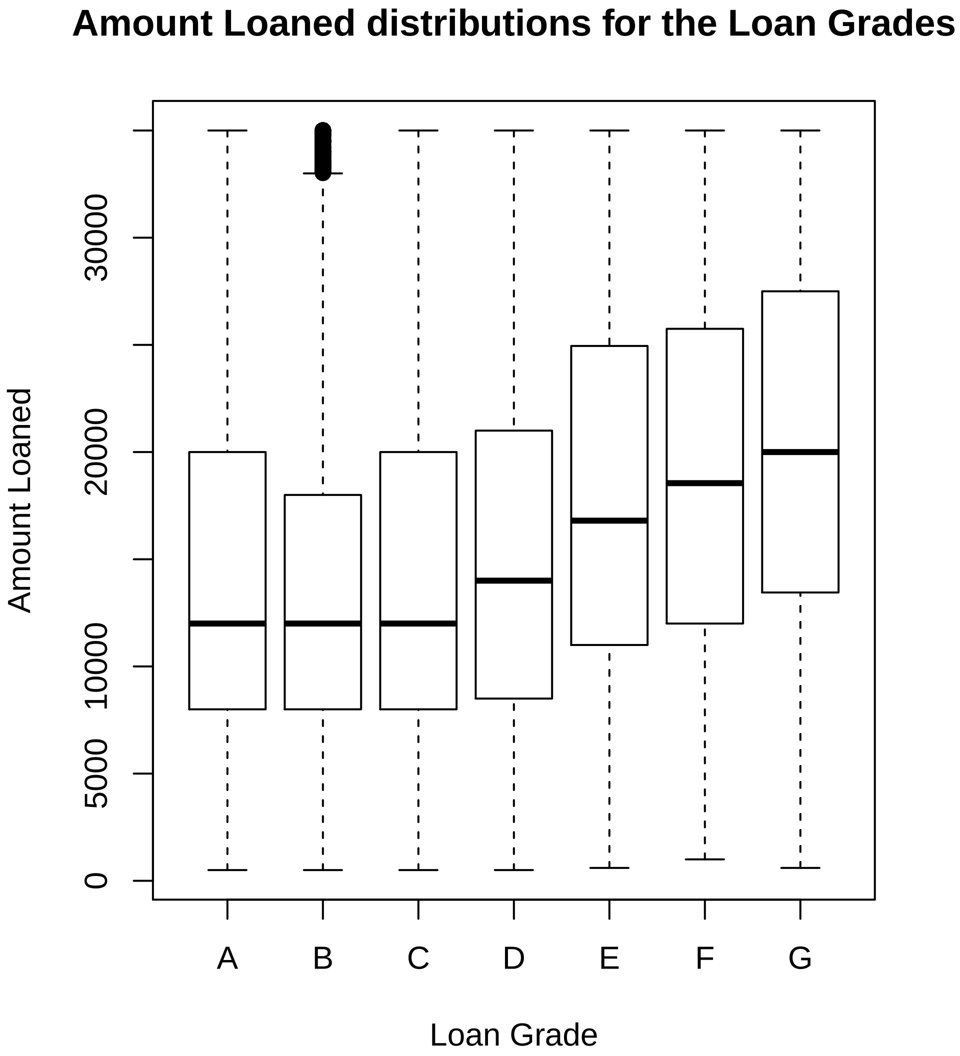
In general the higher the loan grade the more risky Lending Club thinks the loan is; e.g., grade A is for the least risky loans and grade G is for the most risky ones.

(See more details on https://www.lendingclub.com/foliofn/rateDetail.action)

Part a (7 points)

Compare the distributions of the loan amounts across different loan grades by using a side-by-side boxplot.

```r
# Creates side by side boxplots for visual comparison of loan
# amounts
boxplot(loancase$funded_amnt ~ loancase$grade, xlab = "Loan Grade",
    ylab = "Amount Loaned", main = "Amount Loaned distributions for the Loan Grades")
```



Part b (4 points)

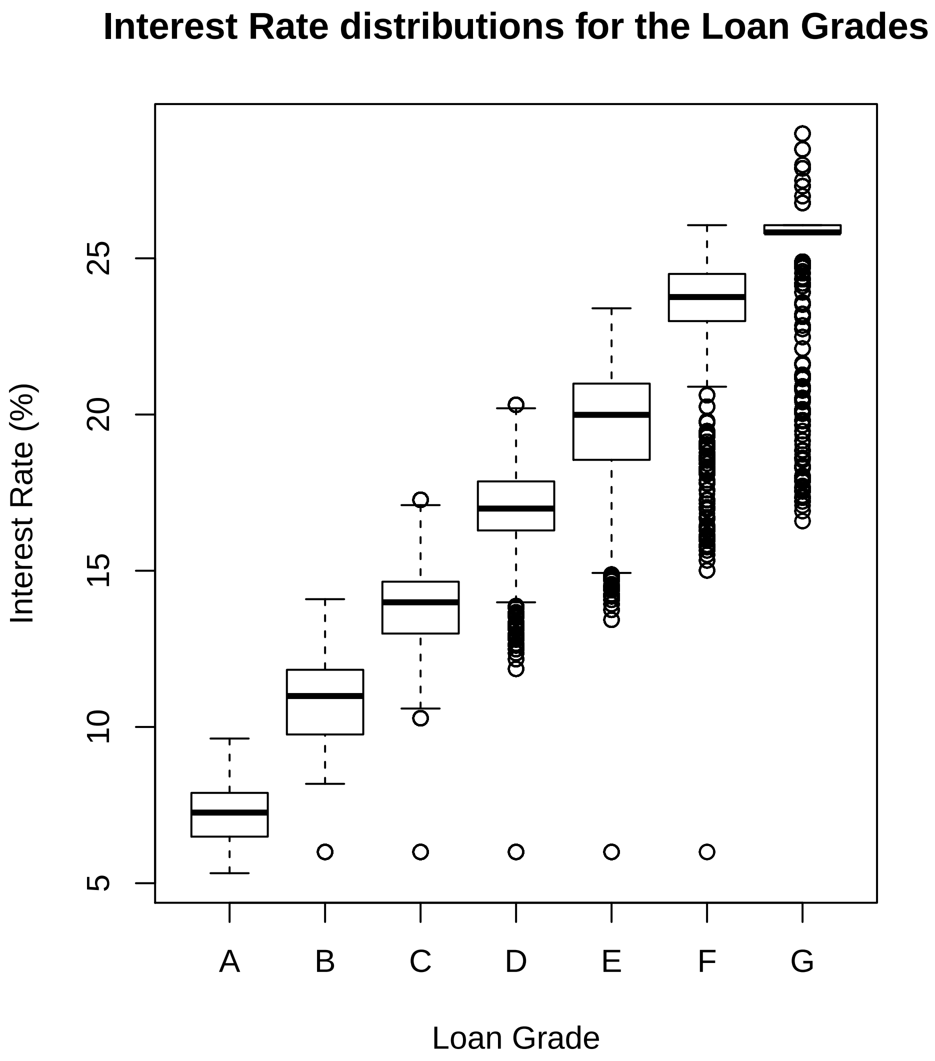
Use the plot in part a and comment on the trend of the data in terms of the relationship between the loan amount and the loan grade; in general do the risky loans tend to have higher or lower loan amounts?

Answer: In general, the riskier the loan, the higher the loan amount.

Part c (7 points)

Next we would like to find out why Lending Club issues high-risk-large-amount loans. Make a side-by-side boxplot to compare the interest rates across different loan grades; do the loans with higher risk tend to have higher or lower interest rates?

```r
# Creates side by side boxplots for visual comparison of
# interest rates
boxplot(loancase$int_rate ~ loancase$grade, xlab = "Loan Grade",
    ylab = "Interest Rate (%)", main = "Interest Rate distributions for the Loan Grades")
```



Answer: The loans with higher risk tend to have higher interest rates.

Part d (12 points: 4pts each)

Use the side-by-side boxplot that you produced in Part c to choose the best answer for the following questions.

(i.) The minimum interest rate for the loans in the most risky category is around

```
(A.) 16.5%
(B.) 20%
```

```
(C.) 25%
(D.) 26%
(E.) 29%
```

No explanation is required.

Answer: A

(ii.) The median interest rate for the loans with grade D is around

```
(A.) 12%
(B.) 14%
(C.) 15.5%
(D.) 17%
(E.) 20.5%
```

No explanation is required.

Answer: D

(iii.) For the interest rates of the loans with grade D the quantity, 1st quartile - 1.5 IQR, is around 11.5.

```
(A.) True
(B.) False
```

No explanation is required.

Answer: B

Problem 3. Why do people want to borrow?(40+3 points)

Part a (15 points)

What are the top three and bottom three reasons for people to apply for loans? Answer this question by finding out how many loan cases there are for each loan purpose category and list the categories in a decreasing order in terms of the number of loan cases (i.e., the category with the highest number of loan cases should be listed first). Please print out the numbers for all the categories. To demonstrate that you understand how the function `tapply()` works please use `tapply()` to solve this question and avoid using the function `table()`.

Hint: there is an example in the notes, Ch2.3_EDA_part_II, that shows that `table()` is just a special case of `tapply()`. See the `Some built-in functions in R` section in Ch2.3_EDA_part_II.

```r
# Turns `loancase$purpose` into a group
grp = loancase$purpose
# Uses the `tapply()` function to create a table of `grp`
loan.purpose = tapply(rep(1, times = length(grp)), INDEX = grp,
    FUN = sum)
sort(loan.purpose, decreasing = T)  # Sorts the table in decreasing order
debt_consolidation        credit_card   home_improvement
            524215             206182              51829
             other     major_purchase     small_business
             42894             17277             10377
               car           medical            moving
              8863              8540              5414
          vacation             house           wedding
              4736              3707              2347
  renewable_energy       educational
               575               423
```

Part b (25 points)

For each purpose category we want to investigate how the loan cases are divided into cases of different loan grades. For example, do credit card loans tend to be more risky or less risky compared to other loans?

Make a 14 by 7 matrix where the rows correspond to the categories in `purpose` and the columns correspond to the categories in `grade`. The row names should be the purpose categories in alphabetical order, and the column names should be the grade categories in alphabetical order. For each row display the percentages of the loan cases for the grades among all the loan cases that belong to the purpose category. E.g., the (1,1) entry of the matrix should be 26.64 and it represents the percentage of the grade A loan cases among all the loan cases within the purpose category `car`.

Express the numbers on the matrix in the unit of percentages (e.g. express .0003 as .03). For a neat display please round up all numbers to 2 digits after the decimal point; you can use the function `round()`; e.g., `round(43.8475, digit=2)` will give you 43.85. It is okay to just print out the table with your code; you do not need to repeat all the numbers in your report. Each row should add up to 100.

Hint.1: It might be easier if you first find out the number of cases across different grade for each purpose categories first.

Hint.2: R's ability to perform vectorized calculation and the recycling rule might be useful here. If you do not remember what vectorized calculation mean please see the example in Precept1 demo where we had a matrix `m` and we calculated:

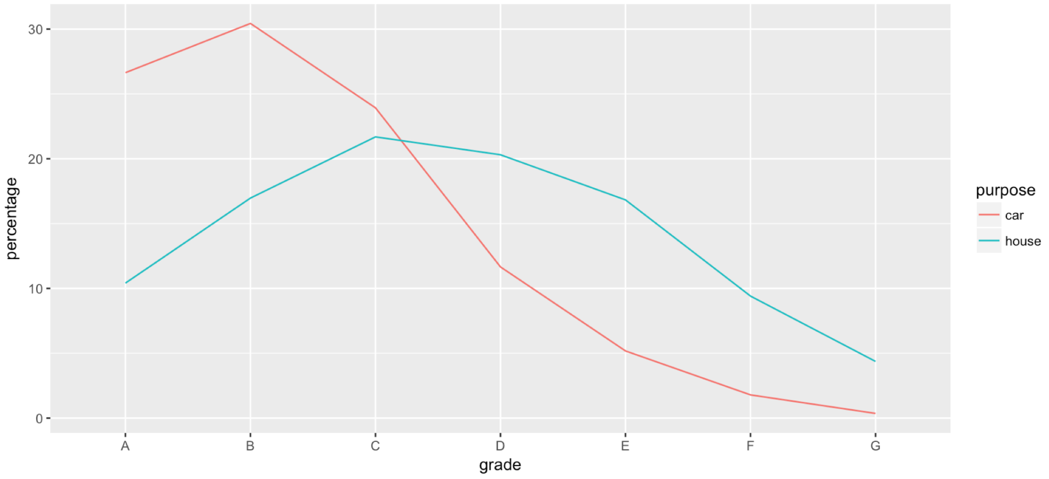
```
# Example matrix to show vectorized calculations
m = matrix(1:6, ncol = 2)
# Multiplies each element in each column with the corresponding
# element in the vector
m * c(1, 2, 3)
     [,1] [,2]
[1,]    1    4
[2,]    4   10
[3,]    9   18
```

```
# The following function creates a matrix of the # of loans per
# grade per purpose, then divides that matrix by the
# concatenated list of total loans per purpose, then multiplies
# the resulting matrix by 100 to output the percentages, and
# then rounds to 2 digits after the decimal point.
m = as.matrix(round((as.matrix(table(loancase$purpose, loancase$grade))/c(8863,
    206182, 524215, 423, 51829, 3707, 17277, 8540, 5414, 42894,
    575, 10377, 4736, 2347)) * 100, digit = 2))
m
```

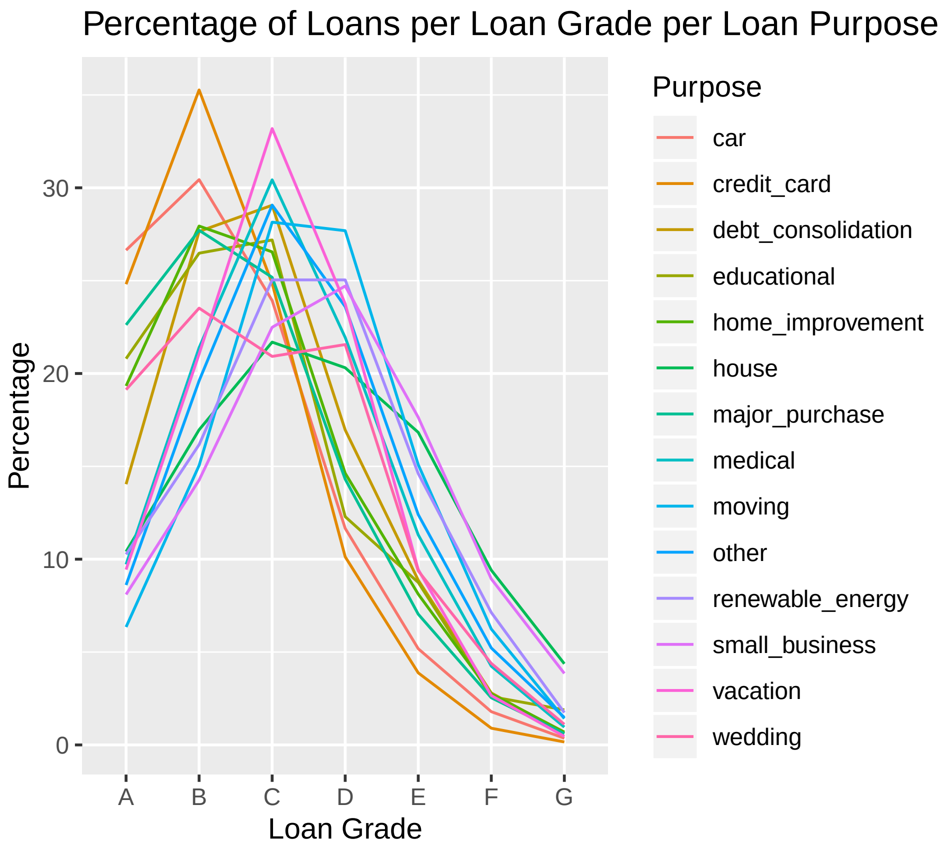
```
                       A     B     C     D     E    F    G
  car                26.64 30.44 23.93 11.66  5.18 1.79 0.36
  credit_card        24.81 35.27 24.85 10.13  3.88 0.90 0.16
  debt_consolidation 14.04 27.65 29.06 16.96  8.87 2.80 0.62
  educational        20.80 26.48 27.19 12.29  8.75 2.60 1.89
  home_improvement   19.33 27.94 26.55 14.63  8.12 2.75 0.69
  house              10.41 16.97 21.69 20.31 16.83 9.41 4.37
  major_purchase     22.62 27.71 25.18 14.31  7.04 2.54 0.61
  medical             9.72 21.39 30.43 21.94 11.31 4.24 0.96
  moving              6.35 15.05 28.15 27.69 15.09 6.24 1.42
  other               8.61 19.61 29.07 23.60 12.40 5.22 1.48
  renewable_energy   10.26 16.17 25.04 25.04 14.61 7.13 1.74
  small_business      8.10 14.26 22.49 24.71 17.64 8.94 3.85
  vacation            9.44 21.05 33.19 23.75  9.44 2.66 0.46
  wedding            19.13 23.52 20.92 21.56  9.37 4.39 1.11
```

Part c (bonus credit: 3 points credits applied to any points that you might have missed on this problem set; i.e., the maximum score of this problem set is 100 with the bonus credit.)

For each purpose category make a line plot to plot the percentages on each row in the matrix in part b; overlay the 14 lines so that they are all in one graph. E.g., if this were just for categories `car` and `house`, the graph should look like this if you use the `ggplot2` package; if you use one of the basic graphic functions in R instead the aesthetic elements of the graph will look different but the lines should have similar shapes.



```r
# Coerces the percentage matrix `m` into the data frame `d`
d = as.data.frame(m)
names(d)[1] = "Purpose"  # These three lines name the columns of `d`
names(d)[2] = "Grade"
names(d)[3] = "Percentage"
# Creates a plot of percentages of loans per loan grades per
# loan purposes using `ggplot2`
ggplot(d) + geom_line(aes(x = Grade, y = Percentage, group = Purpose,
    color = Purpose)) + xlab("Loan Grade") + ylab("Percentage") +
    ggtitle("Percentage of Loans per Loan Grade per Loan Purpose")
```


Percentage of Loans per Loan Grade per Loan Purpose
Percentage
A
B
C
D
E
F
G
Loan Grade
Purpose
car
credit_card
debt_consolidation
educational
home_improvement
house
major_purchase
medical
moving
other
renewable_energy
small_business
vacation
wedding