SML 201 Problem Set 2

Tyler Campbell

March 5, 2018

Problem set 2 is due by 11:59pm on Monday March 5 on Blackboard. Please submit both a .Rmd and a .pdf file on Blackboard. If the due date falls on a date that has a lecture on the next day please bring a hard copy of the pdf file to the first lecture after the due date; otherwise, please drop off the pdf copy at 26 Prospect Avenue (see the *Submitting Problem Sets and Projects* section under *Problem Sets and Projects* on the Syllabus for detailed instructions) by 5pm on the next day of the due date.

Make sure that you have all your digital signatures along with the honor pledge in each of these documents (there should be more than one signature if you work in groups).

This problem set can be completed in groups of up to 3 students. Unlike for projects, you can work with whoever you prefer for problem sets. 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 the instructors in person during office hours) from instructors for *problem sets*; however, please do not post code on a public post on Piazza.

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 the work on any given problem set or project contains approximately equal contributions from all members of the group; 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. Failing to make contributions and then putting your name on a project will be considered a violation of the honor code. Also, please do not divide work among your group mates.

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 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; other people in your group should turn in the list of the people in your group in the *Text Submission* field on the submission page.

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: Tyler Campbell

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

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 use the basic `graphics` functions in `R` and the functions in the `ggplot2` package; the choice between using a basic `graphics` function in R and its equivalent version in `ggplot2` package is up to you.

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 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:

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

- Why do people want to borrow?

Problem 1. Getting familiar with the dataset

Part a

Read in the data in the *loan_data.csv* file by using the `read.csv()` function; name this object `loancase`. Make sure to check that you have 887,379 rows and 5 columns for `loancase`. What is the object type of `loancase`?

Answer: data frame

```
# read in file into dataframe, check dimensions and
# object type
loancase = read.csv("loan_data.csv")
dim(loancase)
[1] 887379      5
class(loancase)
[1] "data.frame"
```

Part b

The file `LCDataDictionary.xlsx` includes the variable definitions for the variables in the original dataset. List the names and the definitions of the variables used in this problem set. You do not need to show any code for this part; it is fine to just copy and paste the definitions. The purpose of this part is to help you understand the meanings of the variables that you need for this problem set. Also, note that we use the variable `funded_amnt` rather than `loan_amnt` because `funded_amnt` is the approved loan amount for an applicant whereas `loan_amnt` is the loan amount requested by the applicant.

id: A unique LC assigned ID for the loan listing.

funded_amnt: The total amount committed to that loan at that point in time.

int_rate: Interest Rate on the loan

grade: LC assigned loan grade

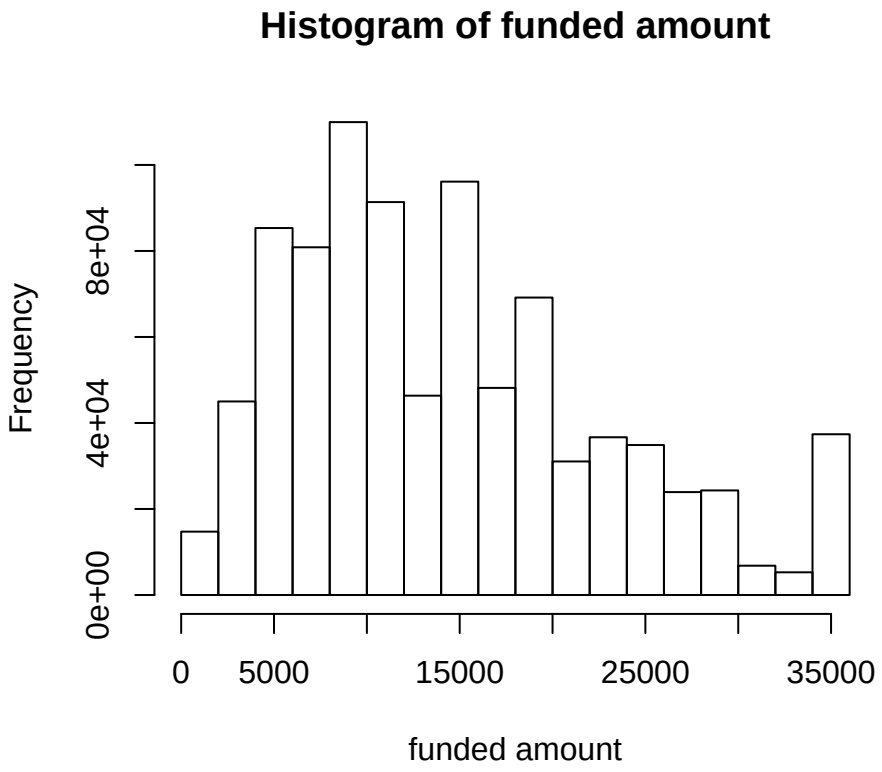
purpose: A category provided by the borrower for the loan request.

Problem 2. The sizes of the loans issued by Lending Club.

Part a

Make a histogram that displays the distribution of the loan amounts.

```r
# makes histogram
hist(loancase$funded_amnt, xlab = "funded amount",
    main = "Histogram of funded amount")
```



Part b

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. Report these numbers.

Answer: right skewed

Use 5 number summary and the mean to summarize dataset, cause it shows the shape of the distribution of a dataset and can show how the mean is affected by extreme values.

Min. 1st Qu. Median Mean 3rd Qu. Max. 500 8000 13000 14742 20000 35000

```r
# summray gives 5 number summaray and the mean of
# the vector
summary(loancase$funded_amnt)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    500    8000   13000   14742   20000   35000
```

Part c

Use the rule that we discussed in lecture for identifying outliers and check to see if there are any outliers in the loan amount values.

Answer: No outliers

```
# solve for IQR multiply by 1.5 and add/subtract
# from 3rd and 1st quintile to find cut off for
# outliers. check if the is any value below min
# cutoff or above max cutoff
IQRrange = IQR(loancase$funded_amnt)
IQRrange = IQRrange * 1.5
min = fivenum(loancase$funded_amnt)[2] - IQRrange
max = fivenum(loancase$funded_amnt)[4] + IQRrange
min(loancase$funded_amnt) < min
[1] FALSE
max(loancase$funded_amnt) > max
[1] FALSE
```

Problem 3. Why does Lending Club welcome risky loans?

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

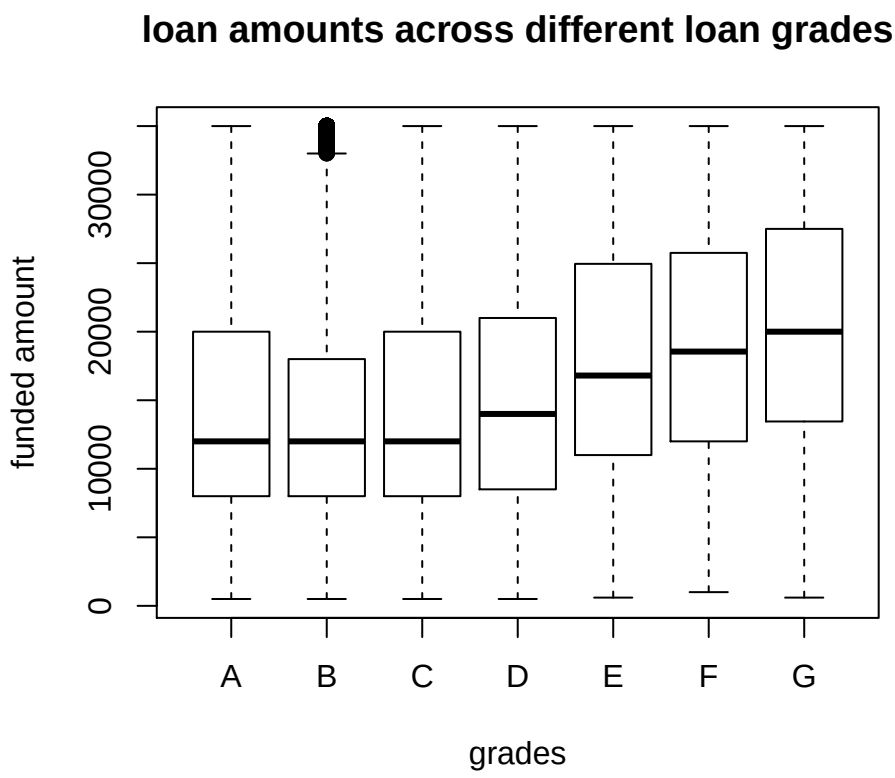
In general the higher the loan grade the more risky Lending Club thinks the loan is; e.g., grade A is for the least risky loans and grade G is for the most risky ones.

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

Part a

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

```r
# creates boxplot of funded amount by grade
boxplot(loancase$funded_amnt ~ loancase$grade, xlab = "grades",
    ylab = "funded amount", main = "loan amounts across different loan grades")
```

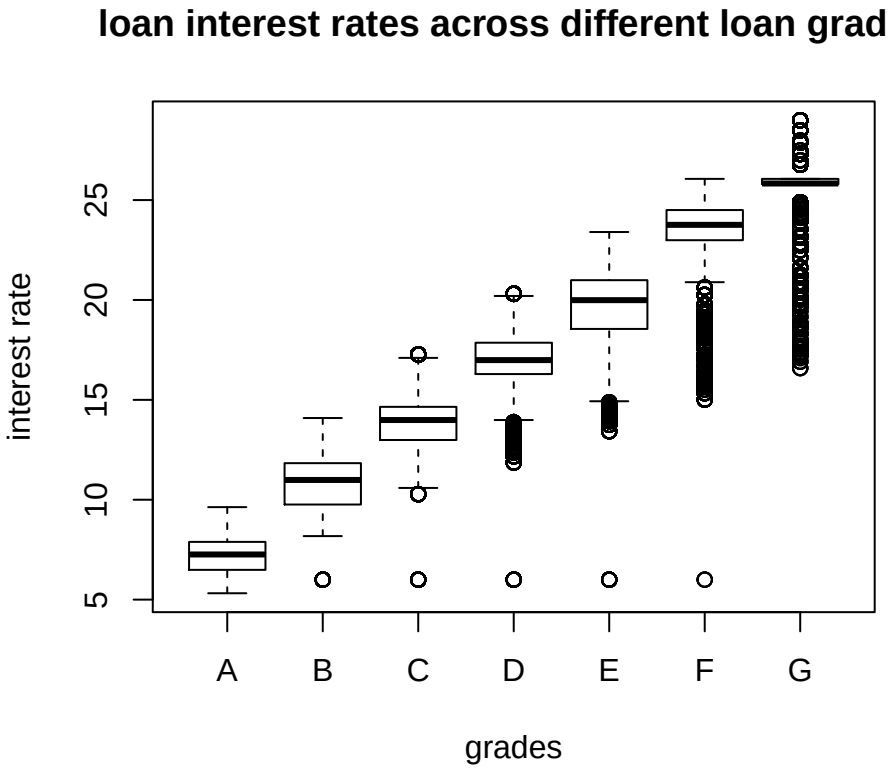


Part b

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

As the loan amount increases the grade decreases. Risky loans have higher amounts.

```r
# creates boxplot of interest rates by grade
boxplot(loancase$int_rate ~ loancase$grade, xlab = "grades",
    ylab = "interest rate", main = "loan interest rates across different loan grades")
```



Part c

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; are you still surprised by the result that you saw in part b? Why or why not?

Lower grades have higher interest rates. This does not surprise me because if Lending Club gives a high risk loan they would want a bug return since they are taking a chance.

Problem 4. Why do people want to borrow?

Part a

What are the top 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). Answer: 1 debt consolidation, 2 credit card, 3 home improvement

```
# creates table of each purpose and its freq then
# sorts by freq in decending order and extracts top
```

```
# 3
reason = table(loancase$purpose)
reason = sort(reason, decreasing = TRUE)
names(reason[1:3])
[1] "debt_consolidation" "credit_card"        "home_improvement"
```

Part b

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 be 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 might be useful here. If you do not remember what vectorized calculation mean please see example in Week2 Precept demo where we have a matrix `m` and we calculated:

```
# vector of purpose names
purpose = names(reason)
# vector of freq
freq = as.numeric(reason)
# creates data frame with columns of previous 2
# vectors sorts data by purpose in alphabetical
# order
df = data.frame(purpose = purpose, freq = freq)
df = df[order(df$purpose), ]
# grades vector A:G in alphabetical order
grades = sort(unique(as.vector(loancase$grade)))
# creates matrix and with column names of grades
# and row names with the purpose
m = matrix(nrow = 14, ncol = 7)
rownames(m) = sort(purpose)
colnames(m) = grades
```

```
# loop through each postion in matrix and solve for
# percentage using correct data from data frame
for (row in 1:nrow(m)) {
    for (col in 1:ncol(m)) {
        x = df$purpose[row]
        y = grades[col]
        n = nrow(loancase[loancase$grade == y & loancase$purpose ==
            x, ])
        m[row, col] = round(n/df$freq[row] * 100, digits = 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
```

You can use the table that you created in part b to answer the following questions.

Part c

Which two purpose categories are the most risk-taking categories (i.e., have the highest percentages of grade G loans)? Please provide code that outputs the answer. Answer: house and small business

```
# sort in decreasing order and extract first 2
# elements from 7th column where the grade is G
names(sort(m[, 7], decreasing = TRUE))[1:2]
```

```
[1] "house"          "small_business"
```

Part d

For credit card loans which loan grade has the highest percentage of loan cases? What about for small business loans? You do not have to provide any code to answer these questions; it is okay to just look at the numbers on your table if you prefer.

credit card = B

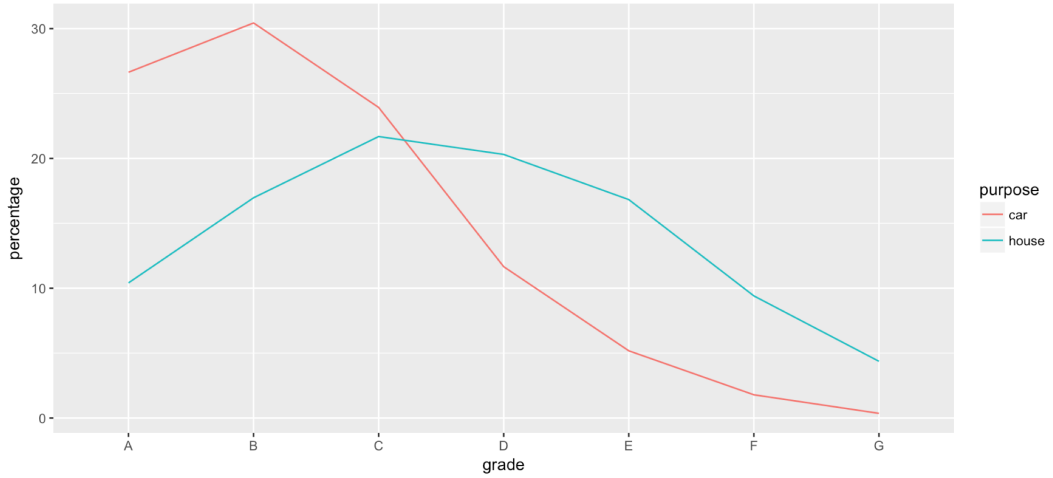
```
names(sort(m[2, ], decreasing = TRUE))[1]
[1] "B"
```

small business = D

```
names(sort(m[12, ], decreasing = TRUE))[1]
[1] "D"
```

(2 pts extra credit) Part e

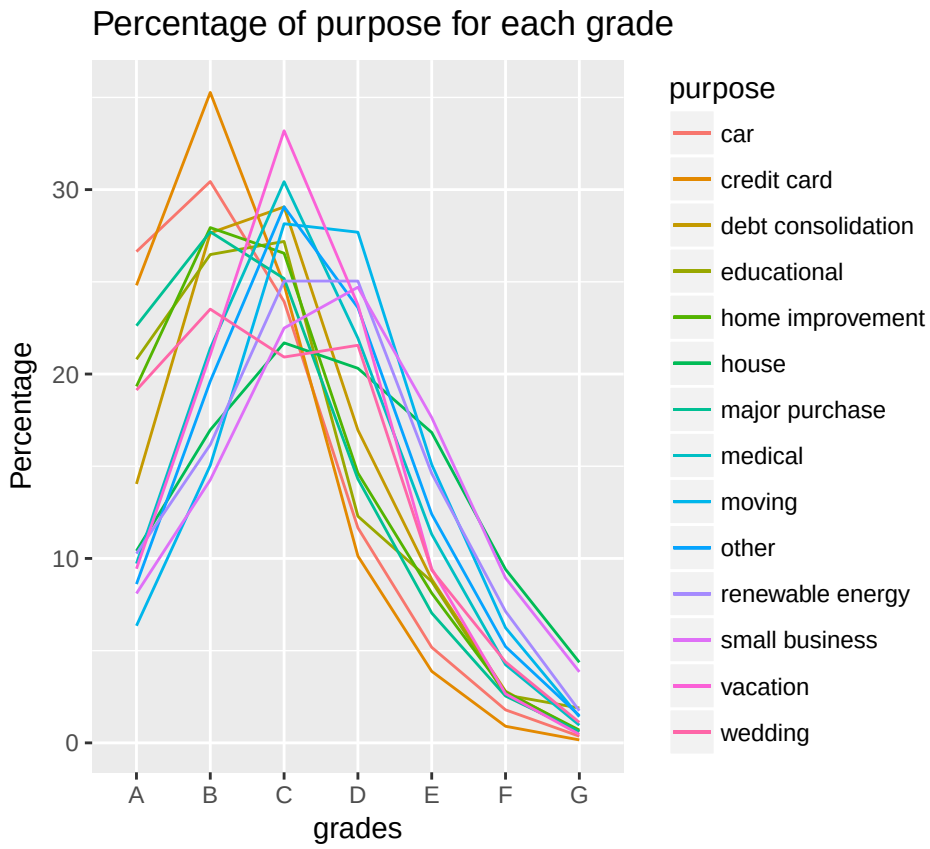
For each purpose category make a line plot to plot the percentages on each row in the matrix in part b; overlay the 14 lines so that they are all in one graph; e.g., if this were just for categories `car` and `house`, the graph should look like this.




```
# lineplot each purpose percentage per grade
library(ggplot2)
plot = as.data.frame(t(m))

ggplot(plot, aes(x = grades)) + geom_line(aes(y = car,
    group = 1, colour = "car")) + geom_line(aes(y = credit_card,
    group = 1, colour = "credit card")) + geom_line(aes(y = debt_consolidation,
    group = 1, colour = "debt consolidation")) + geom_line(aes(y = educational,
    group = 1, colour = "educational")) + geom_line(aes(y = home_improvement,
    group = 1, colour = "home improvement")) + geom_line(aes(y = house,
```

```
    group = 1, colour = "house")) + geom_line(aes(y = major_purchase,
    group = 1, colour = "major purchase")) + geom_line(aes(y = medical,
    group = 1, colour = "medical")) + geom_line(aes(y = moving,
    group = 1, colour = "moving")) + geom_line(aes(y = other,
    group = 1, colour = "other")) + geom_line(aes(y = renewable_energy,
    group = 1, colour = "renewable energy")) + geom_line(aes(y = small_business,
    group = 1, colour = "small business")) + geom_line(aes(y = vacation,
    group = 1, colour = "vacation")) + geom_line(aes(y = wedding,
    group = 1, colour = "wedding")) + ylab("Percentage") +
    labs(title = "Percentage of purpose for each grade") +
    scale_color_discrete(name = "purpose")
```



To receive full credits you must use the `geom_line()` function in the `ggplot2` package to make this graph. Confirm your answers for parts c and d with the results shown on the graph.