Module 1 Homework

Sarah Baker

12/19/2017

Module 1 Homework

(Background on student loans). We'll now look at some data on Perkins loans for ten nearby schools. Our primary data file is called student_loans_ts.csv

Day 1

Question 1

Please review the following functions shown in class today. You should use the help documentation for each function in answering these questions. To access the help use `?function` or help(function). Please describe the following for each function:

- What are the inputs to the function?
- What are the outputs from the function?
- When is this function useful?

Please discuss the following functions:

- select()
- mutate()
- read_csv()
- setwd()
- ggplot()
- aes()

Example Answer:

- sum()
- The sum() function takes one or more vectors of numbers as an argument. Additionally, it accepts the na.rm argument which allows sum() to be used even when NA values are present.
- The output of the sum() function is a single number, the sum of all values contained in the inputs
- The function is useful when looking to find the sum of multiple numbers. It is also useful with TRUE/FALSE values and can be used in creating complex filters.

Question 2

A) Read in the data.

```
# Answer
## Set your working directory so that Rstudio knows where your data files are
#loans_data <- read_csv( # fill in your code here
#)
v <- "william"
print(v)
```

```
## [1] "william"
```

B) What are the classes of each variable type? (Hint: reproduce the output below.)

C) Show the first 15 observations of the data. Your results should look like the output below.

D) What is one question you want to explore after looking at your data?

Question 3

A) Select out the variables for zip code and ID.

Since we have the school's state, we don't particularly care about it's zipcode. And since we have each school's name, we don't really need it's ID number. So let's select out those columns from `loans_data`, but still call our object `loans_data`.

```
# Answer
loans_data <- # fill in code here
```

B) Create a dataset containing only the data from Howard University from `loans_data`. Then, reproduce this plot of loan disbursements over time. Hint: you need to use `filter()` to achieve this.

```
# Answer
howard_data <- # fill in code here

ggplot(howard_data, # fill in code here
       )
```

C) What do you notice about this chart? (Hint: discuss any sharp drops or spikes and what might be causing them.)

Question 4

A) Create a new variable, disbursements per recipient, using `howard_data`. Hint: you will need to use `mutate()`.

```
# Answer
howard_data <- mutate(howard_data, dpr = # fill in code here
                      )
```

B) Plot disbursements per recipient (recreate the plot below.)

```
# Answer
ggplot(howard_data, # fill in code here
       )
```

C) How is this chart different from that in 2C? Does this chart tell the same story? (Again, discuss spikes and drops and their possible causes.)

Day 2

Question 1

Please review the following functions shown in class today. You should use the help documentation for each function in answering these questions. To access the help use `?function` or help(function). Please describe the following for each function:

- What are the inputs to the function?
- What are the outputs from the function?
- When is this function useful?

Please discuss the following functions:

- rename()
- %>%
- seq()
- group_by()
- summarize()
- scale_x_discrete()

Example Answer:

- sum()
- The sum() function takes one or more vectors of numbers as an argument. Additionally, it accepts the na.rm argument which allows sum() to be used even when NA values are present.
- The output of the sum() function is a single number, the sum of all values contained in the inputs
- The function is useful when looking to find the sum of multiple numbers. It is also useful with TRUE/FALSE values and can be used in creating complex filters.

Question 2

A) Let's add the loan disbursements per recipient for every school, using `loans_data`.

```
# Answer
loans_data <- loans_data %>%
  mutate( # fill in
  )
```

B) Let's examine the spread of disbursements per recipient using a histogram. Reproduce the chart below using `geom_histogram()` (hint: you will need to play with the value for `binwidth`.)

```
loans_data %>% ggplot(# fill in
  )
```

C) Uh oh. That warning means something is wrong with our variable `dpr`. Look at the values of `dpr` using the `unique()` function and describe the problem. Which school has the problem observations?

D) Replace the `NA` values with zero using the `if_else()` function and `loans_data`. Reproduce the same chart as in part B).

```
loans_data <- loans_data %>% mutate(dpr = ifelse(# fill in
  ))

loans_data %>% ggplot(# fill in
  )
```

E) Describe the spread of disbursements per recipient. (Hint: where is the highest count of observations? Where are the lowest?)

Question 3

A) Create a data set containing only DC schools from `loans_data`. Then, reproduce the chart below, showing loan disbursements per recipient for DC schools. (Note: a school is considered "in DC" if its main campus is located in a DC zipcode.)

```
# Answer
dc_schools <- # fill in code here

ggplot(dc_schools, # fill in code here
       )
```

B) How does Howard's disbursement pattern over time compare with that of other DC schools?

C) Let's see how DC schools compare to those in other states. Reproduce the faceted chart below using `loans_data` and the `facet_wrap` command.

D) How do the trends in funding over time compare in these three areas? What might drive some of the differences?

Day 3

Question 1

Please review the following functions shown in class today. You should use the help documentation for each function in answering these questions. To access the help use `?function` or help(function). Please describe the following for each function:

- What are the inputs to the function?
- What are the outputs from the function?
- When is this function useful?

Please discuss the following functions:

- facet_wrap()
- wtd.mean()
- scale_color_manual()
- unique()
- as.Date()

Example Answer:

- sum()
- The sum() function takes one or more vectors of numbers as an argument. Additionally, it accepts the na.rm argument which allows sum() to be used even when NA values are present.
- The output of the sum() function is a single number, the sum of all values contained in the inputs
- The function is useful when looking to find the sum of multiple numbers. It is also useful with TRUE/FALSE values and can be used in creating complex filters.

Question 2

It's possible that whether the school is public or private may be influencing disbursements in addition to, or perhaps more than, it's geographic location. Let's look into the average disbursement per recipient for public and private universities.

A) What are the values of `School.Type`? Do you foresee any problems with these values?

B) Use `case_when()` and `mutate()` to update the `School.Type` variable. Display the unique values of your updated `School.Type` variable.

```
loans_data <- loans_data %>% mutate(School.Type = # fill in
                                    )

unique(# fill in
  )
```

C) Use `group_by()` and `summarise()` to find the average loan disbursement per recipient for private versus public schools *for each year*. Show the first 10 observations (they should look like those below).

D) Reproduce the chart below. Note the title, labels, and the `breaks` of the x and y axes.

E) What does this chart say about loans to private versus public universities? (Hint: what is similar about the two lines? What is different?)

Question 3

Now we can also examine the spread of disbursements per recipient by school type.

A) Reproduce the box plot below using `loans_data`.

B) Interpret the chart from part (A). What is similar and what is different about the distributions of public versus private schools?

C) Name a type of chart that could help us further explore this relationship, and what that type of chart would show us that the boxplot does not.

D) One type of chart that could help us further explore this relationship is a density plot. Reproduce the chart below using `loans_data`.

E) Using the two charts above, describe the distribution of disbursements per recipient for private versus public schools. Offer a reason why their distributions are so different. (Hint: where is the highest concentration for each school type? What federal financial aid program might be causing the giant spike?)

Day 4

We will be using a cleaned cross-section of the loans data, student loans_xc.csv, for this section.

Question 1

Please review the following functions shown in class today. You should use the help documentation for each function in answering these questions. To access the help use `?function` or help(function). Please describe the following for each function:

- What are the inputs to the function?
- What are the outputs from the function?
- When is this function useful?

Please discuss the following functions:

- scale_x_date()
- summary()
- lm()
- tidy()
- augment()

Example Answer:

- sum()
- The sum() function takes one or more vectors of numbers as an argument. Additionally, it accepts the na.rm argument which allows sum() to be used even when NA values are present.
- The output of the sum() function is a single number, the sum of all values contained in the inputs
- The function is useful when looking to find the sum of multiple numbers. It is also useful with TRUE/FALSE values and can be used in creating complex filters.

Question 2

Let's use regression analysis to determine what factors affect the amount of disbursements per recipient.

A) Load the student_loans_xs data and create a regression model, `baseline_model`, that explores the effect of school type on disbursements per recipeint. Display these results using the summary function.

B) Interpret the coefficients from the model. Are these results statistically and economically significant? (Hint: What is the omitted group?)

C) Using the broom package, model the distribution of residuals by school type. Are we over- or under-predicting the disbursements per student in our current model?

Question 3

The previous model wasn't terrible, but we would like to improve it in order to better understand what affects the distribution of Perkins loans.

A) Create a new regression model, `revised_model`, that adds two dummy variables for each state the schools are loacated within, and display these results using the summary function.

B) Interpret the coefficients from the model. Are these results statistically and economically significant? (Hint: What are the omitted groups)

C) Create a new model, `revised_model2`, that includes an interaction between school type and state. Interpret at least one of the interaction terms from the new model, are these terms statistically and economically significant?

D) Use stargazer to create a regression table that includes all 3 of the models we have developed so far. Which of these models would you consider the "best"?

E) Use `group_by()` and `summarise()` to find the average loan disbursement per recipient for private versus public schools *for each state*. Compare the average loan disbursements to the coefficients from the interaction model, how do they compare? Do you believe we over fitted our third model, why or why not?

F) Dicuss one piece of data that you believe would be helpful in improving the accuracy of our model. How would this improve our model? How would you go about collecting this information?