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.

```r
# Answer
loans_data <- read.csv( # fill in your code here
)
```

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

```
## 'data.frame':    108 obs. of  9 variables:
##  $ OPE.ID       : int  143400 143700 144400 144500 144800 208300 210500 374900 374500 143400 ...
##  $ School       : chr  "American University (The)" "Catholic University of America (The)" "George Wa
##  $ State        : chr  "DC" "DC" "DC" "DC" ...
##  $ Zip.Code     : chr  "200168020" "200640002" "200522353" "200571252" ...
##  $ School.Type  : chr  "Private/Non-Profit" "Private/Non-Profit" "Private/Non-Profit" "Private/Non-P
##  $ Recipients   : int  556 796 2367 2051 119 289 215 379 1587 616 ...
##  $ Disbursements: int  1543612 1159214 6208395 9778697 426312 512583 381030 700457 4671716 1738142 .
##  $ Year         : int  2005 2005 2005 2005 2005 2005 2005 2005 2005 2006 ...
##  $ Date         : int  2005 2005 2005 2005 2005 2005 2005 2005 2005 2006 ...
```

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

```
##    OPE.ID                                    School State  Zip.Code
## 1  143400                 American University (The)    DC 200168020
## 2  143700      Catholic University of America (The)    DC 200640002
## 3  144400              George Washington University    DC 200522353
## 4  144500                     Georgetown University    DC 200571252
## 5  144800                         Howard University    DC 200590001
## 6  208300                   Morgan State University    MD 212510002
## 7  210500 University of Maryland - Baltimore County    MD 212500002
## 8  374900                   George Mason University    VA 220304444
## 9  374500                    University of Virginia    VA 229032600
## 10 143400                 American University (The)    DC 200168020
## 11 143700      Catholic University of America (The)    DC 200640002
## 12 144400              George Washington University    DC 200522353
## 13 144500                     Georgetown University    DC 200571252
## 14 144800                         Howard University    DC 200590001
## 15 208300                   Morgan State University    MD 212510002
##           School.Type Recipients Disbursements Year Date
## 1  Private/Non-Profit        556       1543612 2005 2005
## 2  Private/Non-Profit        796       1159214 2005 2005
## 3  Private/Non-Profit       2367       6208395 2005 2005
## 4  Private/Non-Profit       2051       9778697 2005 2005
## 5  Private/Non-Profit        119        426312 2005 2005
## 6              Public        289        512583 2005 2005
## 7              Public        215        381030 2005 2005
## 8              Public        379        700457 2005 2005
## 9              Public       1587       4671716 2005 2005
## 10 Private/Non-profit        616       1738142 2006 2006
## 11 Private/Non-profit        782       1253236 2006 2006
## 12 Private/Non-profit       2454       4758710 2006 2006
## 13 Private/Non-profit       1204       5465439 2006 2006
## 14 Private/Non-profit        102        441252 2006 2006
## 15             Public        289        509087 2006 2006
```

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.

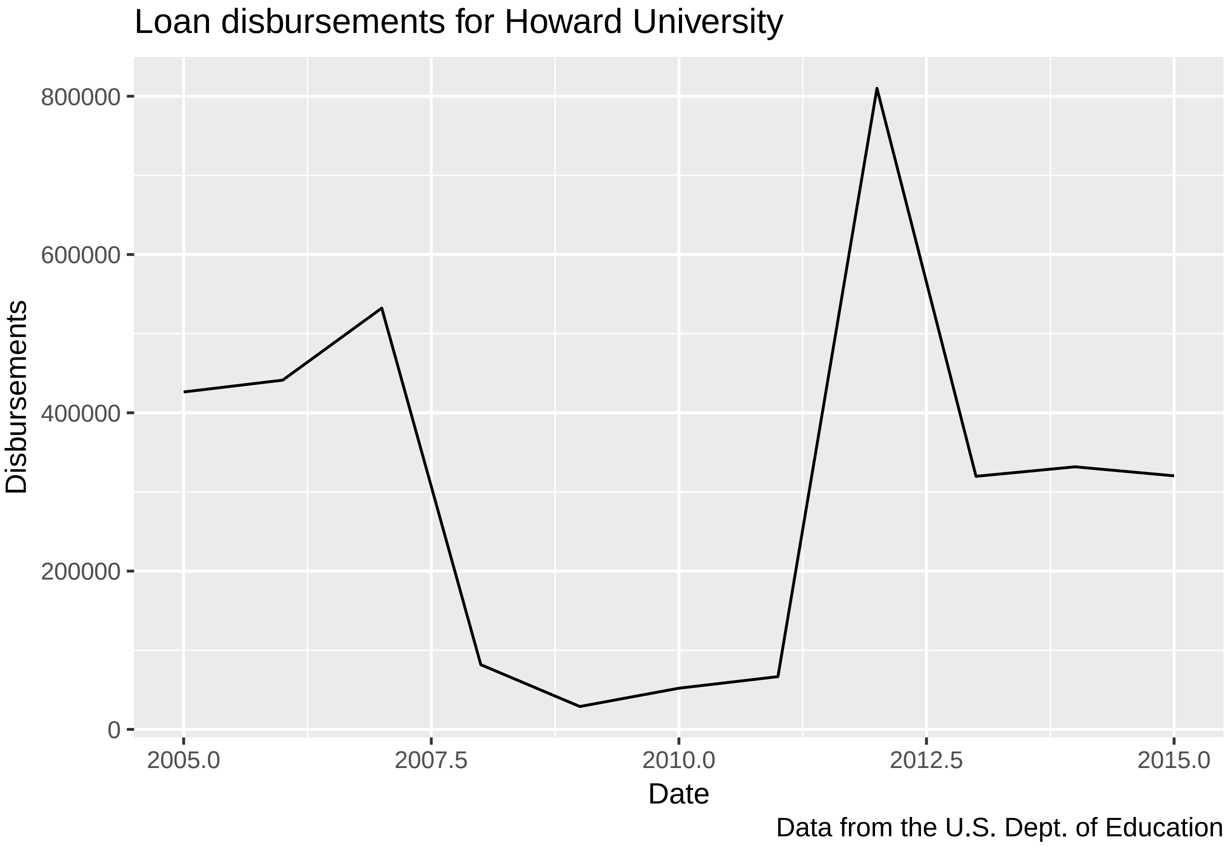
Since we have the school's state, we don't particularly care about it's zipcode. And since we have each school's name, we don't really need it's ID number. So let's select out those columns from `loans_data`, but still call our object `loans_data`.

```
# 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.)

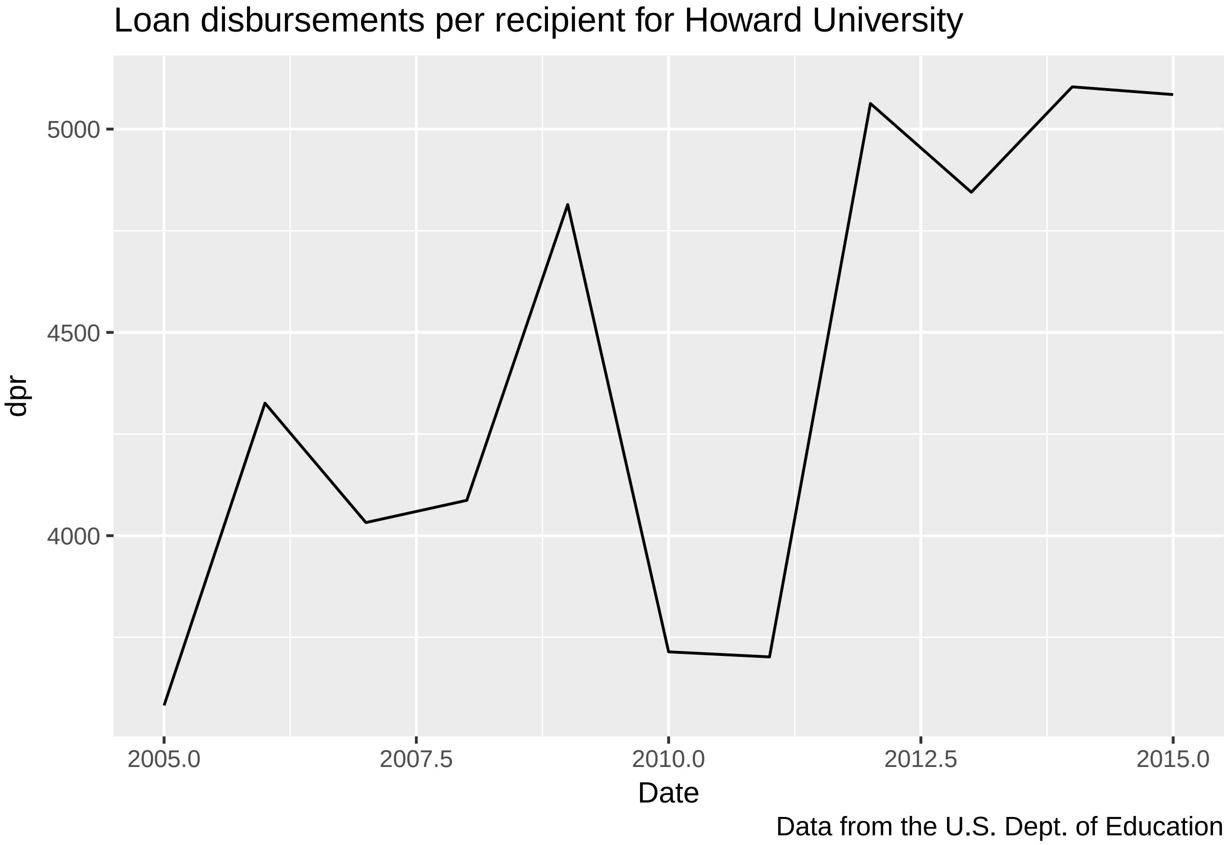
Question 4

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
       )
```


Loan disbursements per recipient for Howard University

Data from the U.S. Dept. of Education

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

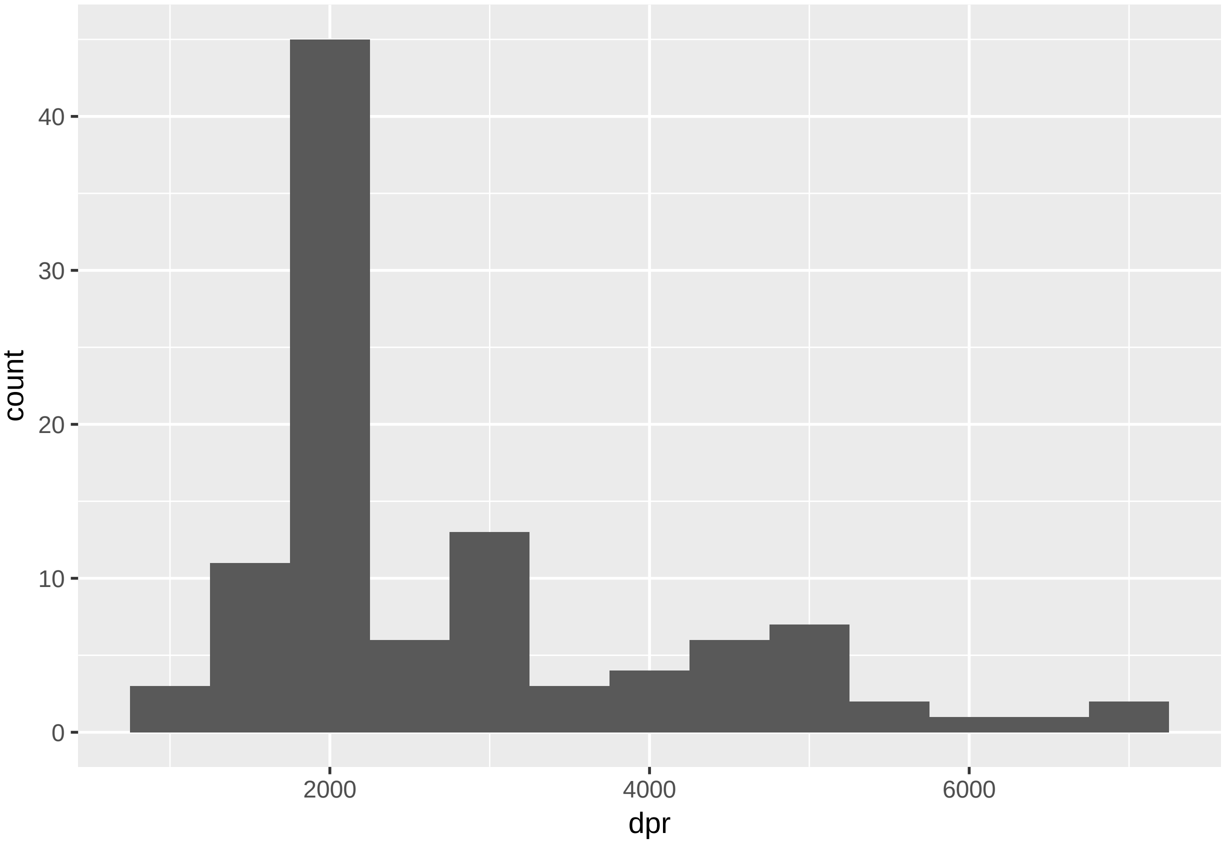
A) Let's add the loan disbursements per recipient for every school, using `loans_data`.

```
# 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
  )
```

```
## Warning: Removed 4 rows containing non-finite values (stat_bin).
```



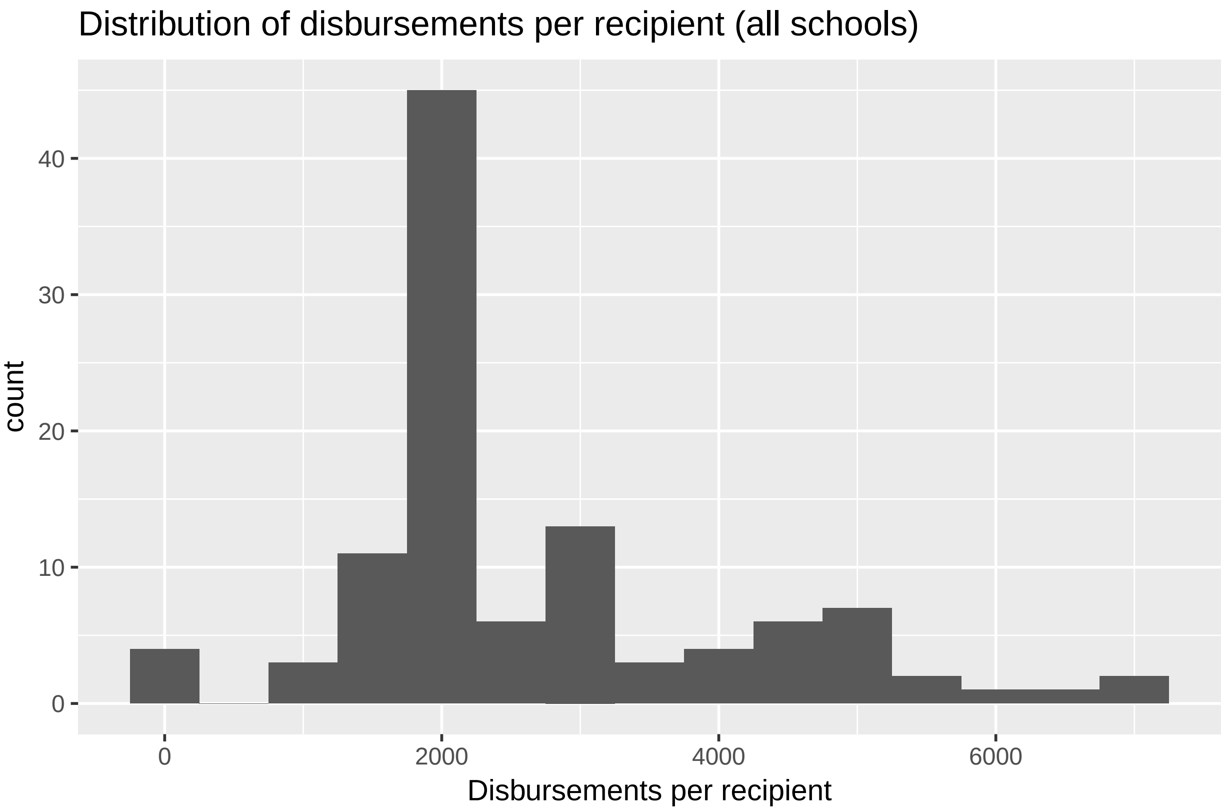
C) **Uh oh. That warning means something is wrong with our variable dpr. Look at the values of dpr using the unique() function and describe the problem. Which school has the problem observations?**

```
##   [1] 2776.281 1456.299 2622.896 4767.770 3582.454 1773.644 1772.233
##   [8] 1848.172 2943.740 2821.659 1602.604 1939.165 4539.401 4326.000
##  [15] 1761.547 1946.342 1844.292 2702.499 2830.896 1499.237 2206.861
##  [22] 4629.328 4032.280 1864.030 1991.426 1696.486 1844.301 2344.830
##  [29] 2828.568 1362.560 1917.110 4430.126 4086.950 1812.227 1893.636
##  [36] 1081.785 1897.382 1945.268 2857.612      NaN 1466.551 5318.735
##  [43] 4814.500 1870.833 1884.791 1492.249 1847.421 1941.151 2870.242
##  [50] 1652.017 5065.518 3714.286 1787.037 1584.574 1105.703 2110.389
##  [57] 2921.131 1865.760 1899.194 5665.471 3701.833 1894.463 2343.615
##  [64] 1134.628 2152.309 2878.313 1910.816 1271.149 5862.413 5062.869
##  [71] 1850.000 2158.273 1700.023 4315.792 2209.758 2965.439 1968.297
##  [78] 1856.184 6823.853 4844.697 1863.532 1929.367 2191.942 4458.768
##  [85] 2225.352 2927.432 1975.656 2251.963 6894.121 5103.908 1842.081
##  [92] 1923.843 2020.068 2198.655 3853.824 2881.309 2029.922 1997.681
##  [99] 6691.094 5084.889 2222.222 2959.195 1830.160 4017.978 2252.899
```

D) **Replace the NA values with zero using the ifelse() 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
  )
```


Distribution of disbursements per recipient (all schools)

Data from U.S. Dept. of Education

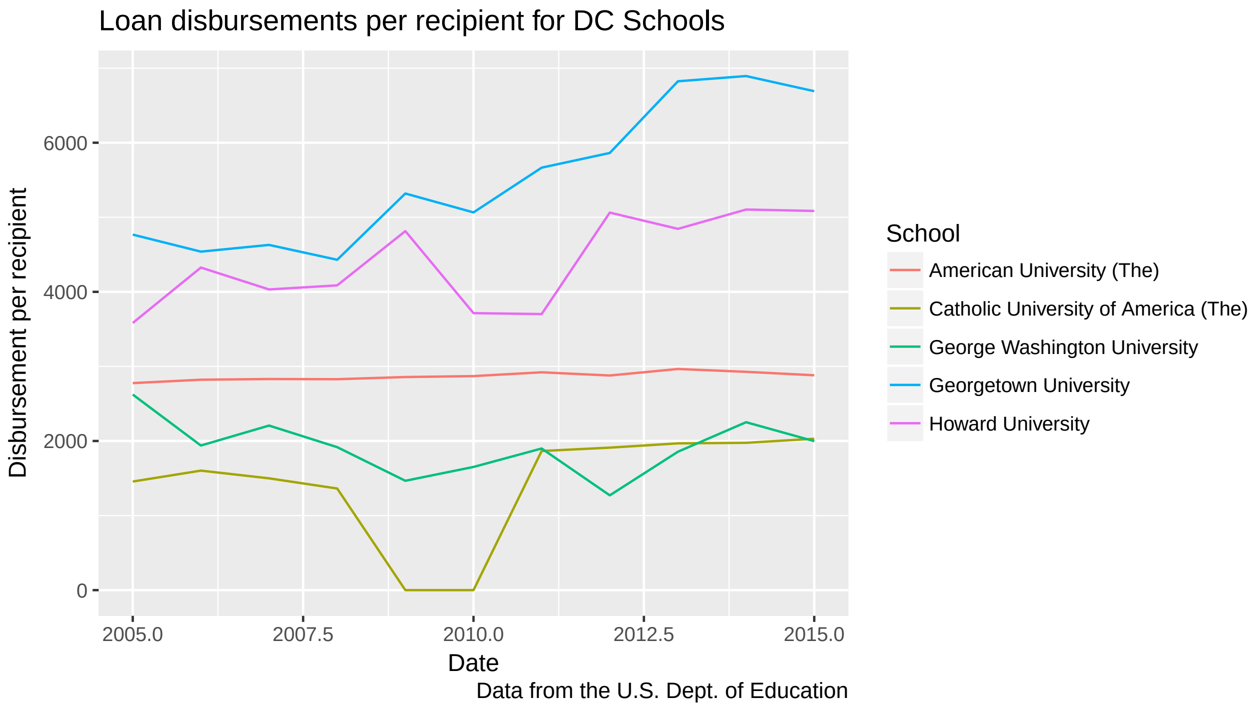
E) Describe the spread of disbursements per recipient. (Hint: where is the highest count of observations? Where are the lowest?)

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.)

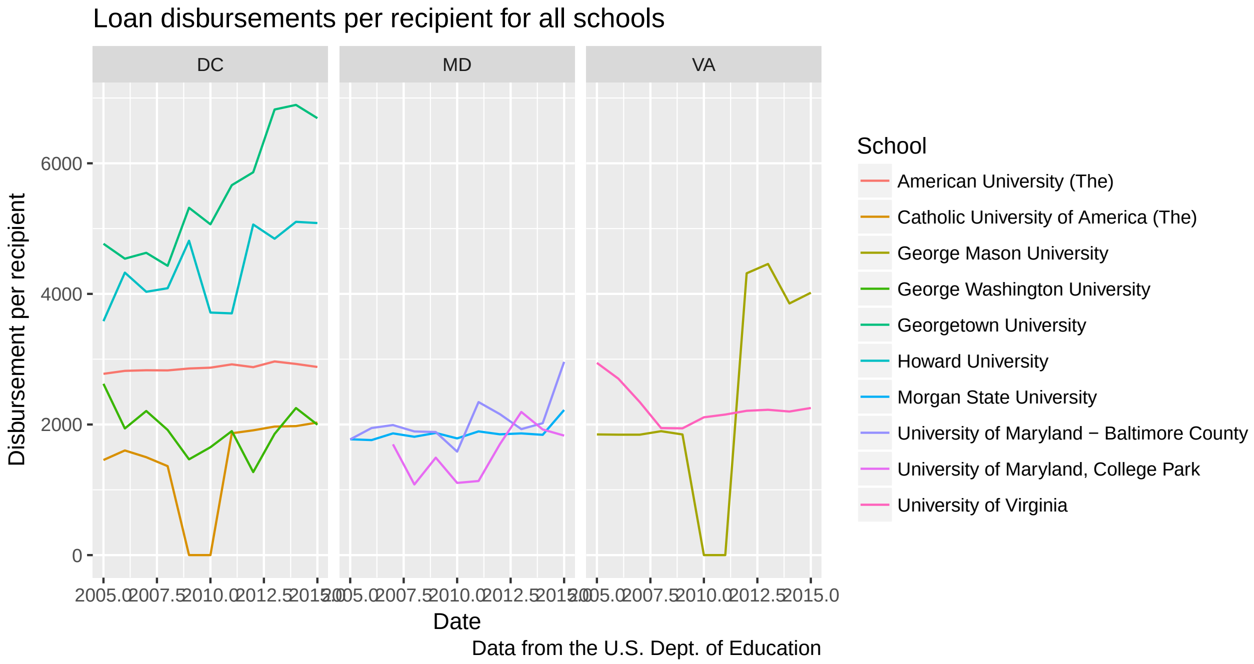
```
# Answer
dc_schools <- # fill in code here

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.

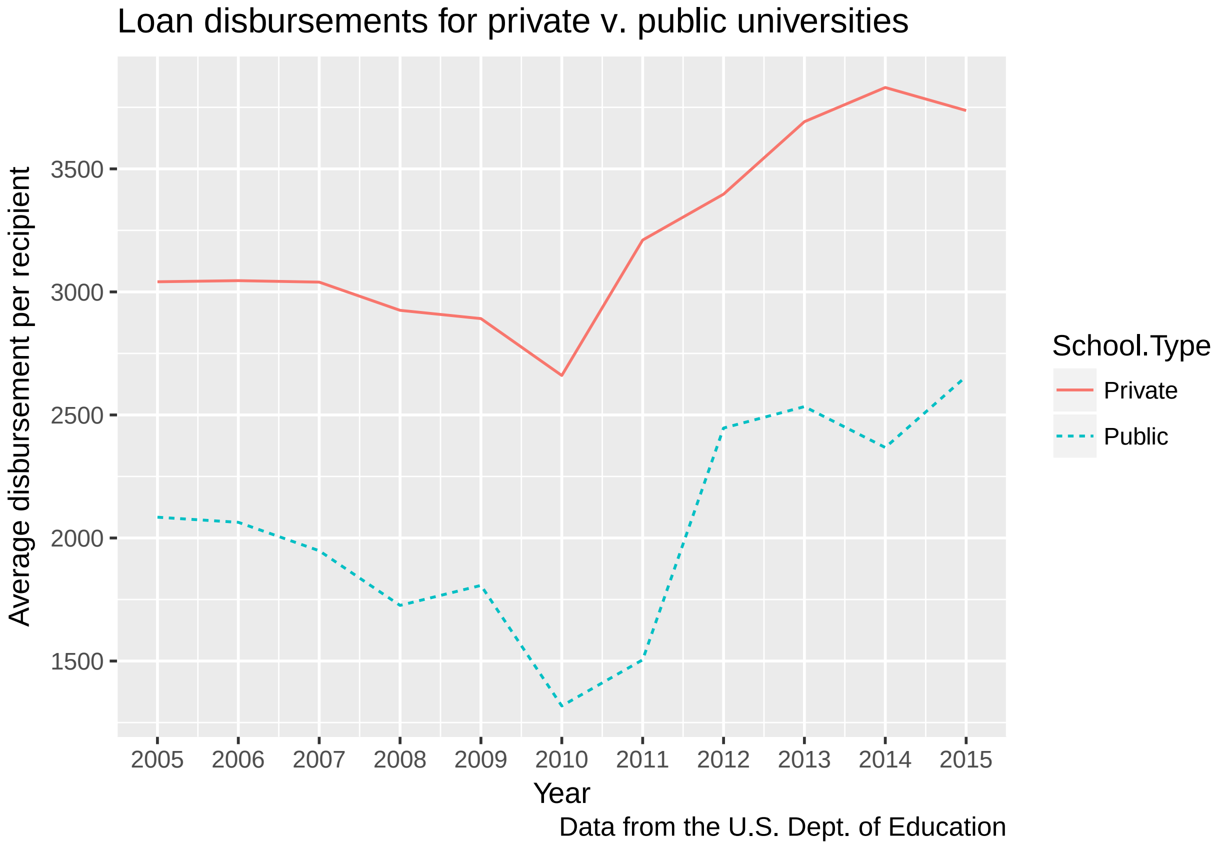
```
loans_data <- loans_data %>% mutate(School.Type = # fill in
                                      )

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).

```
## # A tibble: 10 x 3
## # Groups:   School.Type [1]
##    School.Type  Year avg_dpr
##    <chr>       <int>   <dbl>
##  1 Private      2005    3041
##  2 Private      2006    3046
##  3 Private      2007    3040
##  4 Private      2008    2925
##  5 Private      2009    2891
##  6 Private      2010    2660
##  7 Private      2011    3211
##  8 Private      2012    3397
##  9 Private      2013    3692
## 10 Private      2014    3831
```

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

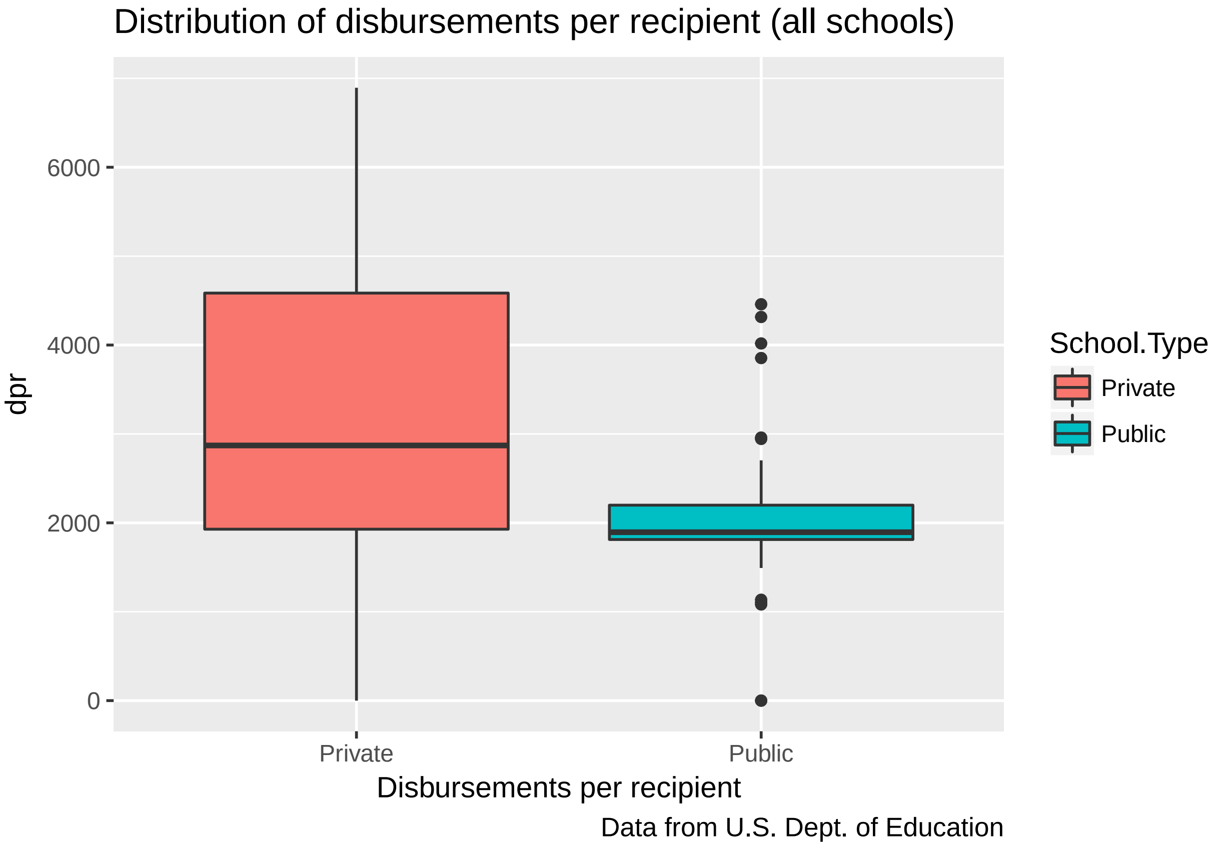


E) What does this chart say about loans to private versus public universities? (Hint: what is similar about the two lines? What is different?)

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`.

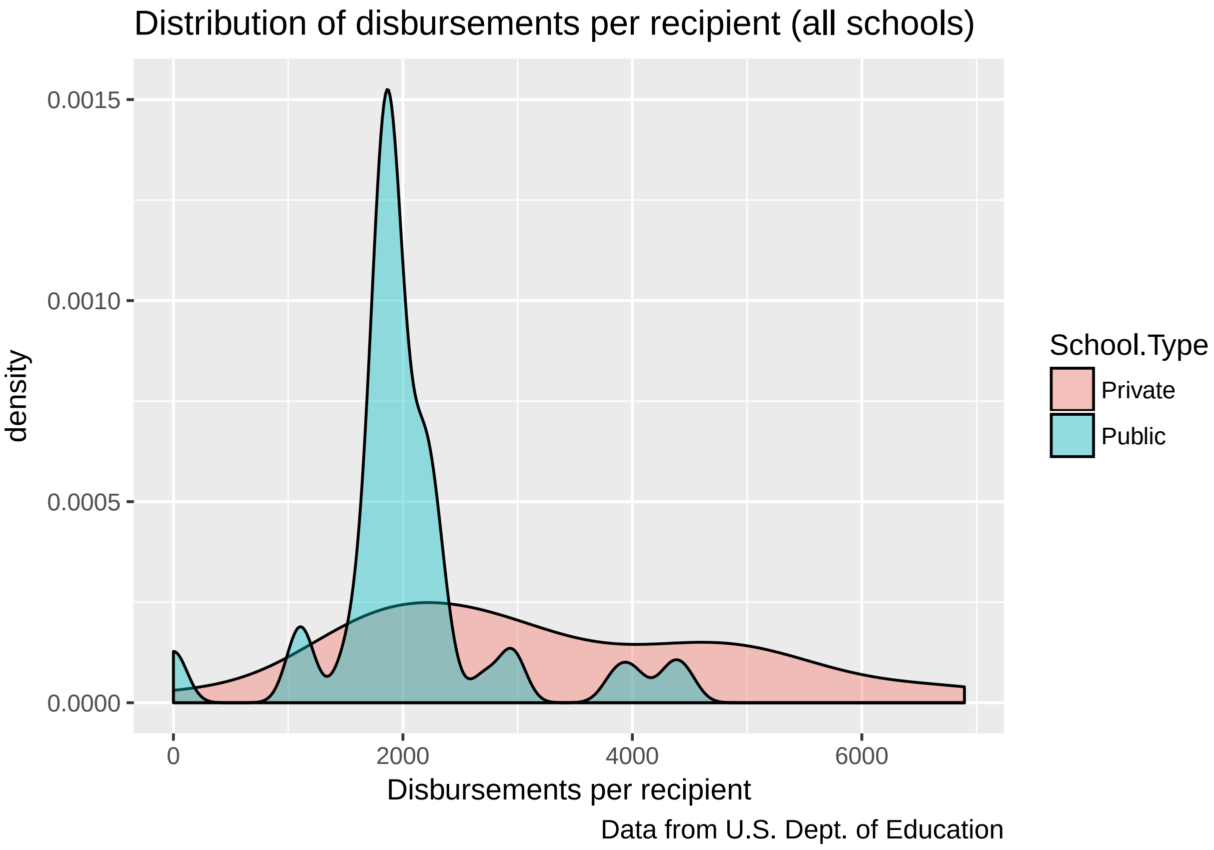


Data from U.S. Dept. of Education

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`.



Distribution of disbursements per recipient (all schools)

Data from U.S. Dept. of Education

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 required 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?