Homework2part1

Ayush

9/12/2020

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts ------------------------------------------------------------------------------------ tidy
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(ggmap)
```

```
## Google's Terms of Service: https://cloud.google.com/maps-platform/terms/.

## Please cite ggmap if you use it! See citation("ggmap") for details.
```

```
library(maptools)
```

```
## Loading required package: sp

## Checking rgeos availability: FALSE
##      Note: when rgeos is not available, polygon geometry     computations in maptools depend on gpcli
##      which has a restricted licence. It is disabled by default;
##      to enable gpclib, type gpclibPermit()
```

```
library(maps)
```

```
##
## Attaching package: 'maps'

## The following object is masked from 'package:purrr':
##
##     map
```

```
library(dplyr)
library(corrr)
library(ggcorrplot)
```

(a) Use the read.csv() function to read the data into R, or the csv library to read in the data with python. In R you will load the data into a dataframe. In python you may store it as a list of lists or use the pandas dataframe to store your data. Call the loaded data college. Ensure that your column headers are not treated as a row of data.

```
college <- read.csv("College.csv")
view(college)
```

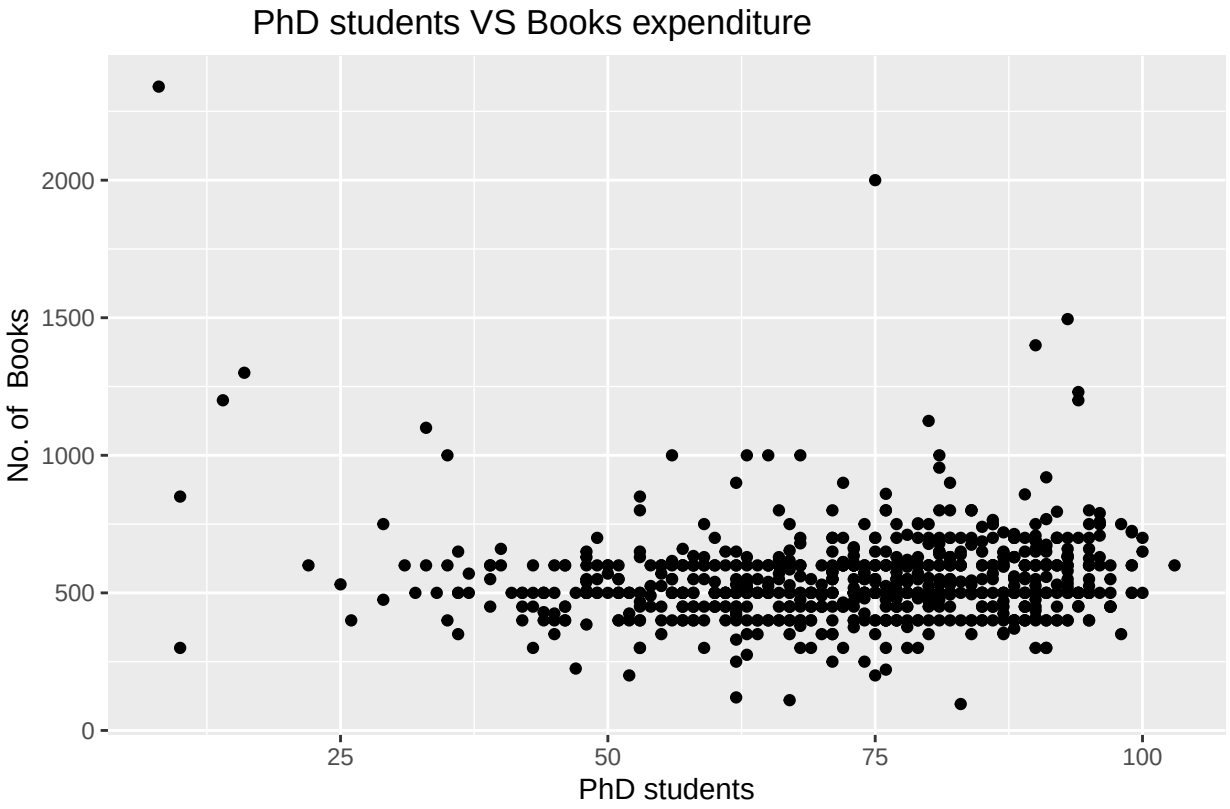
(b) Find the median cost of books for all schools in this dataset.

```
median(college$Books, na.rm = FALSE, )
```

```
## [1] 500
```

(c) Produce a scatterplot that shows a relationship between two numeric (not factor or boolean) features of your choice in the dataset. Ensure it has appropriate axis labels and a title.

```
CG <- ggplot(college, aes(x=PhD ,y = Books)) +geom_point()
CG <- CG +labs(title = "
               PhD students VS Books expenditure",x="PhD students", y="No. of  Books")
show(CG)
```

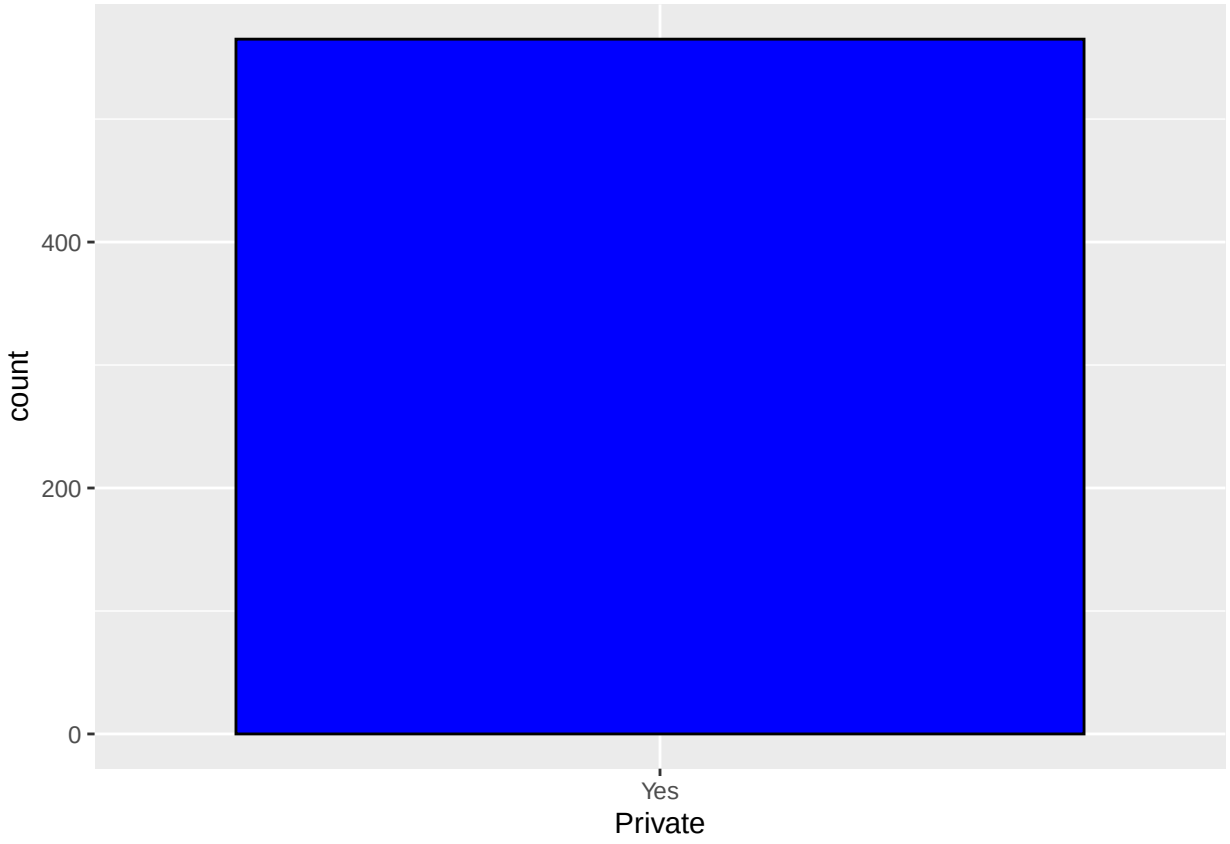


(d) Produce a histogram showing the overall enrollment numbers (P.Undergrad plus F.Undergrad) for both public and private (Private) schools. You may choose to show both on a single plot (using side by side bars) or produce one plot for public schools and one for private schools. Ensure whatever figures you produce have appropriate axis labels and a title.

```
hist <- college[c(2,8,9)]
private_students <- hist %>%
  select(Private,F.Undergrad,P.Undergrad) %>%
  filter(Private =="Yes")
UG <- ggplot(data=private_students,mapping=aes(Private)) + geom_histogram(position = "dodge"
                                                          ,color="black", fill="blue",stat = "count")
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
UG
```




```
  #Public Students
public_students <- hist %>%
  select(Private,F.Undergrad,P.Undergrad) %>%
  filter(Private =="No")
```

(e) Create a new qualitative variable, called Top, by binning the Top10perc variable into two categories (Yes and No). Specifically, divide the schools into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 75%.

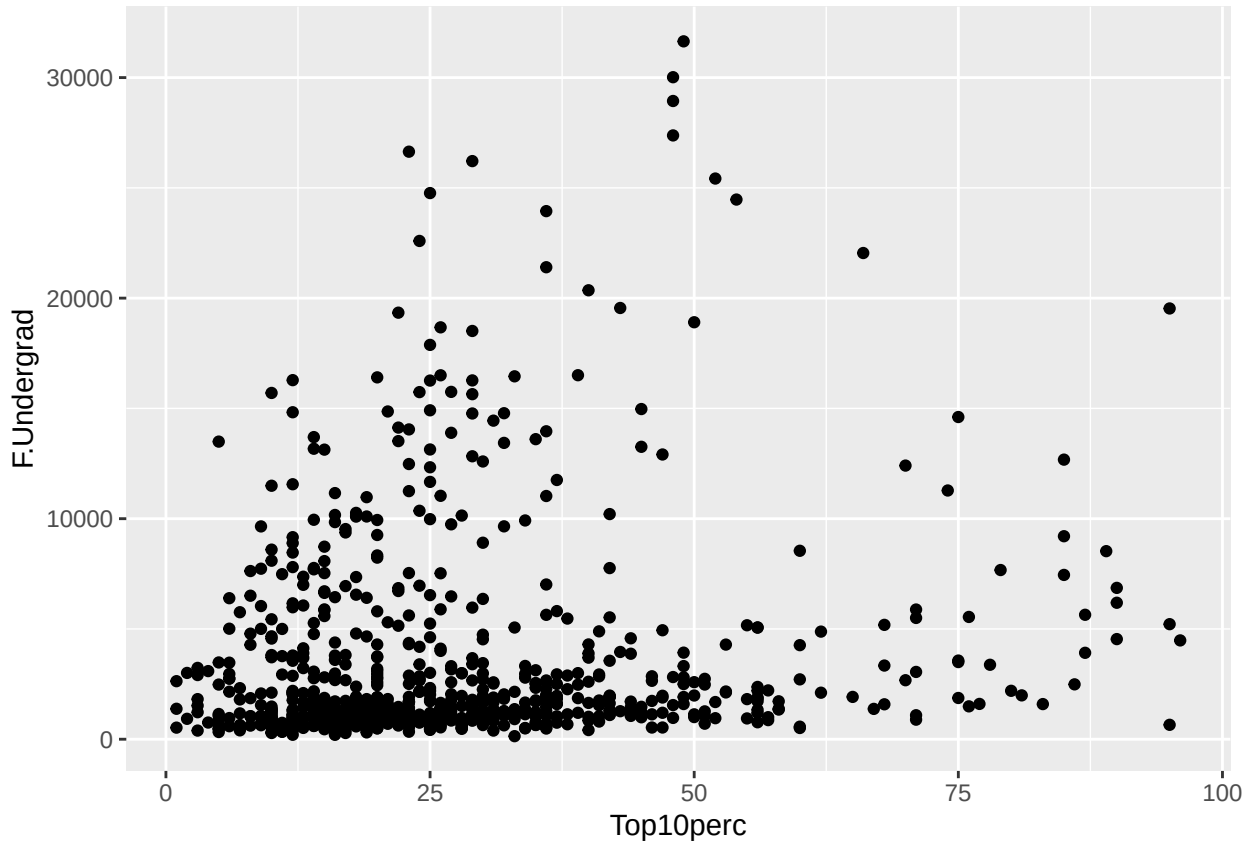
```
Top = rep("Yes",nrow(college))
Top[college$Top10perc < 75] = "No"
Top = as.factor(Top)
college$Top <- Top
view(college)
summary(Top)
```

```
##  No Yes
## 751  26
```

```
ggplot()
```


(f) Continue exploring the data, producing two new plots of any type, and provide a brief (one to two sentence) summary of your hypotheses and what you discover. Feel free to think outside the box on this one but if you want something to point you in the right direction, look at the summary statistics for various features, and think about what they tell you. Perhaps try plotting various features from the dataset against each other and see if any patterns emerge.

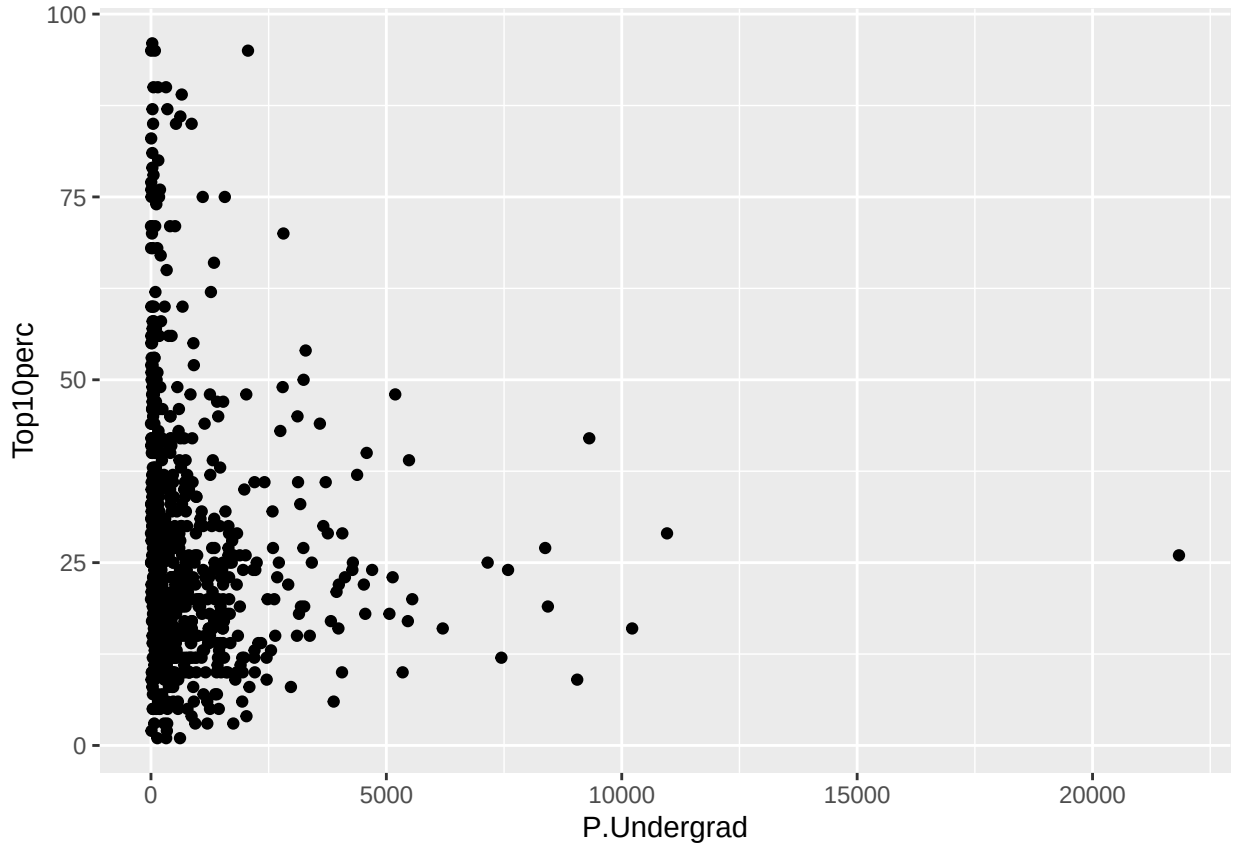
```
full_time <- college %>%
  select(Top10perc, F.Undergrad) %>%
  mutate(sch = Top10perc/F.Undergrad) #Sch = Scholarship opportunities for the top 10% for Full time st
full_time$sch = round(full_time$sch, digits = 3)
ggplot(full_time, aes(x=Top10perc, y= F.Undergrad)) + geom_point()
```




```
view(full_time)
mean(full_time$sch)
```

```
## [1] 0.01818147
```

```
part_time <- college %>%
  select(Top10perc,P.Undergrad) %>%
  mutate(sch = Top10perc/P.Undergrad) #Sch = Scholarship opportunities for the top 10% in Part time stu
part_time$sch = round(part_time$sch , digits = 3)
ggplot(college, aes(x=P.Undergrad, y=Top10perc)) + geom_point()
```




```
#view(part_time)
mean(part_time$sch)
```

```
## [1] 0.7002394
```