clementiSummerReadExploration

Johnathan Clementi

3/10/2020

Data information: Juniata College’s freshman Summer Reading assignment data

Homework requirements: 1. Come up at least two questions you might have about your variables that you can try to answer with your exploration. 2. Visualize Variation in one quantitative variable (1 Figure) 3. Visualize variation in one categorical variable (1 Figure) 4. Visualize covariation for 3 groups of variables; graphs include: (3 Figures in total) a. 1 figure of 1 categorical variable x 1 categorical variable b. 1 figure of 1 quantitative variable x 1 quantitative variable c. 1 figure of 1 categorical variable x 1 quantitative varaible \* Label graphs appropriately and include 1 sentence description. \* Include relevant statistics for each figure 5. Describe how these figures answered the research questions \* Include new questions inspired by these analyses

Library load

library(tidyverse) #loads ggplot and the others  
library(readxl)#for reading in excel files  
library(readr)

Read in data

summerRead = read.csv("C:/Users/clemenj/Google Drive/Grad School/Juniata\_DataScience/DS500/Week7/SummerRead.csv")

str(summerRead)

## 'data.frame': 2296 obs. of 40 variables:  
## $ Gender : Factor w/ 4 levels "","Female","Male",..: 3 2 3 3 3 3 3 3 3 3 ...  
## $ USCitizen : Factor w/ 3 levels "","No","Yes": 3 3 3 3 3 3 3 3 3 3 ...  
## $ English : Factor w/ 3 levels "","No","Yes": 3 3 3 3 3 3 3 3 3 3 ...  
## $ POE : Factor w/ 6 levels "","Exploratory",..: 5 5 4 4 4 4 4 4 4 4 ...  
## $ BooksReadSummer : num 1 1 5 5 5 5 5 5 5 5 ...  
## $ ReadProgHS : Factor w/ 3 levels "","No","Yes": 3 2 2 2 2 2 2 2 2 2 ...  
## $ PartReadProgHS : Factor w/ 3 levels "","No","Yes": 3 2 2 2 2 2 2 2 2 2 ...  
## $ OwnBookSummer : Factor w/ 3 levels "","No","Yes": 3 3 3 3 3 3 3 3 3 3 ...  
## $ PercentReadSummer : num 15 75 0 0 0 0 0 0 0 0 ...  
## $ WhyNotRead : Factor w/ 42 levels "","Boring","Boring, Topic Dislike",..: 15 15 12 12 12 12 12 12 12 12 ...  
## $ OtherReasonsNot : Factor w/ 506 levels "","2 jobs conflicted with my time.",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ FamilyRead : Factor w/ 3 levels "","No","Yes": 2 2 2 2 2 2 2 2 2 2 ...  
## $ SummerSurvey : Factor w/ 2 levels "","Yes": 2 2 2 2 2 2 2 2 2 2 ...  
## $ QuizScore : num NA 0.5 0 0 0 0 0 0 0 0 ...  
## $ FirstGen : Factor w/ 3 levels "","No","Yes": 1 2 2 2 2 2 3 3 2 2 ...  
## $ PercentReadSummerFollowup: num NA 75 0 100 1 1 85 30 100 100 ...  
## $ PercentReadSemester : num NA 100 0 100 0 0 95 70 0 100 ...  
## $ Reread : Factor w/ 3 levels "","No","Yes": 1 2 2 3 2 2 3 2 2 3 ...  
## $ CWSRead : Factor w/ 6 levels "","All of the book",..: 1 5 5 2 5 5 4 2 5 2 ...  
## $ CWSWrite : Factor w/ 24 levels "","Blog post",..: 1 21 21 5 21 21 5 23 21 5 ...  
## $ Think : int NA 4 3 3 2 4 5 2 4 2 ...  
## $ Challenge : int NA 3 3 3 4 4 5 3 3 3 ...  
## $ EnjoyBook : int NA 2 3 3 3 4 5 5 5 3 ...  
## $ EventAdd : int NA NA NA NA NA 5 5 NA NA NA ...  
## $ CourseEffect : Factor w/ 3 levels "","No","Yes": 1 2 2 2 2 2 2 2 2 2 ...  
## $ FollowupSurvey : Factor w/ 2 levels "","Yes": 1 2 2 2 2 2 2 2 2 2 ...  
## $ Book : Factor w/ 5 levels "Handmaid's Tale",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ Year : int 2017 2017 2017 2017 2017 2017 2017 2017 2017 2017 ...  
## $ PercentReadSemesterM : num NA 100 0 100 1 1 95 70 100 100 ...  
## $ ReadSummer : Factor w/ 3 levels "","No","Yes": 2 2 2 2 2 2 2 2 2 2 ...  
## $ ReadFollowup : Factor w/ 3 levels "","No","Yes": 1 2 2 3 2 2 3 2 3 3 ...  
## $ ReadSemester : Factor w/ 3 levels "","No","Yes": 1 3 2 3 2 2 3 2 3 3 ...  
## $ Read : Factor w/ 3 levels "","No","Yes": 1 3 2 3 2 2 3 2 3 3 ...  
## $ AttendEvent : Factor w/ 3 levels "","No","Yes": 1 2 2 2 2 3 3 2 2 2 ...  
## $ ConversationsCount : int NA 0 2 3 2 5 NA 2 0 NA ...  
## $ ConversationsWith : Factor w/ 4 levels "","Group","Individual",..: 1 1 3 4 3 4 2 3 1 1 ...  
## $ BookLength : int 3 3 2 2 2 2 2 2 2 2 ...  
## $ BookDifficulty : int 3 4 2 2 2 2 2 2 2 2 ...  
## $ KnowSummerRead : Factor w/ 41 levels "","Classmate",..: 37 22 22 22 22 22 22 22 22 22 ...  
## $ WhichOrientation : Factor w/ 3 levels "","August","June": 3 3 2 2 2 2 2 2 2 2 ...

UN-TIDY DATA LIST: \* This is a running list of all of the issues I find with the dataset while working on it:

1. At the bottom of the dataset, there are approximately 41 records missing values for variables from Gender to SummerSurvey (Possibly someone accidentally deleted these values)
2. Variables do not line up with metadata variables - function of combining datasets
3. Year is stored as an integer, not a factor
4. POE variable has a category of [BLANK] - Fill with NA?
5. Answers to in WhyNotRead could be cleaned up and combined

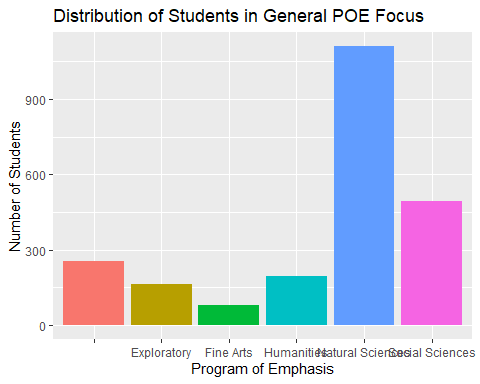
QUESTIONS: 1. Do first generation students tend to or not read the book? 1a. Is the status of first generation student serve as a predictor for how many books the student read over the summer?

1. Is gender a predictor of the students reading the book?
2. Are certain books less interesting to students as a whole? Compare Book and WhyNotRead
3. Does book length correlate with the percentage of the book that the student read?
4. Do the answers students gave to questions: ‘[book] challenged me to think’ and ‘[book] was an enjoyable book’ correlate?

VISUALIZATIONS

Variation in a single Categorical Variable:

ggplot(data = summerRead, aes(x = POE)) +  
 geom\_bar(mapping = aes(fill = POE)) + guides(fill = FALSE) +  
 labs(title = "Distribution of Students in General POE Focus", x = "Program of Emphasis", y = "Number of Students")



summary(summerRead$POE)

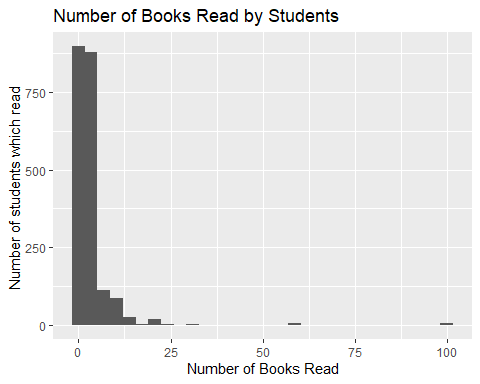
## Exploratory Fine Arts Humanities   
## 253 162 81 197   
## Natural Sciences Social Sciences   
## 1111 492

Figure 1. This is the distribution of students across general Program of Emphasis topics.

Variation in a single Quantitative Variable:

ggplot(data = summerRead) +  
 geom\_histogram(mapping = aes(x = BooksReadSummer), na.rm=TRUE) +  
 labs(title = "Number of Books Read by Students", x = "Number of Books Read", y = "Number of students which read ")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



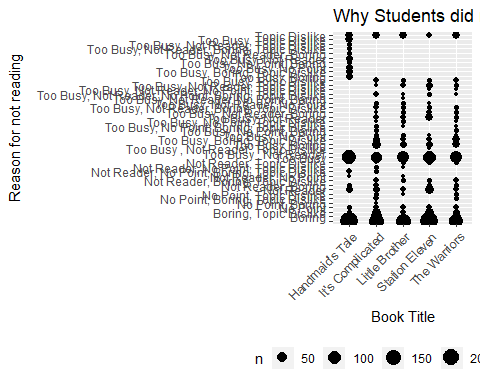
summary(summerRead$BooksReadSummer)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0.000 1.000 2.000 3.548 4.000 100.000 247

Figure 2. This figure illustrates the number of students which read a certain number of books.

Covariation between two categorical variables:

ggplot(data = summerRead) +  
 geom\_count(mapping = aes(x = Book, y = WhyNotRead)) +  
 labs(title = "Why Students did not read each book", x = "Book Title", y = "Reason for not reading") +  
 theme(legend.position = "bottom", axis.text.x = element\_text(angle = 45, hjust = 1))



summary(summerRead$WhyNotRead)

##   
## 942   
## Boring   
## 153   
## Boring, Topic Dislike   
## 83   
## No Point   
## 35   
## No Point, Boring   
## 1   
## No Point, Boring, Topic Dislike   
## 10   
## No Point, Topic Dislike   
## 6   
## Not Reader   
## 79   
## Not Reader, Boring   
## 13   
## Not Reader, Boring, Topic Dislike   
## 18   
## Not Reader, No Point   
## 2   
## Not Reader, No Point, Boring, Topic Dislike   
## 12   
## Not Reader, No Point, Topic Dislike   
## 3   
## Not Reader, Topic Dislike   
## 11   
## Too Busy   
## 448   
## Too Busy , Not Reader   
## 1   
## Too Busy , Not Reader, Topic Dislike   
## 1   
## Too Busy\t, Boring   
## 79   
## Too Busy\t, Boring, Topic Dislike   
## 55   
## Too Busy\t, No Point   
## 6   
## Too Busy\t, No Point, Boring   
## 4   
## Too Busy\t, No Point, Boring, Topic Dislike   
## 5   
## Too Busy\t, No Point, Topic Dislike   
## 4   
## Too Busy\t, Not Reader   
## 43   
## Too Busy\t, Not Reader, Boring   
## 13   
## Too Busy\t, Not Reader, Boring, Topic Dislike   
## 19   
## Too Busy\t, Not Reader, No Point   
## 5   
## Too Busy\t, Not Reader, No Point, Boring   
## 1   
## Too Busy\t, Not Reader, No Point, Boring, Topic Dislike   
## 10   
## Too Busy\t, Not Reader, No Point, Topic Dislike   
## 2   
## Too Busy\t, Not Reader, Topic Dislike   
## 23   
## Too Busy\t, Topic Dislike   
## 24   
## Too Busy, Boring   
## 14   
## Too Busy, Boring, Topic Dislike   
## 16   
## Too Busy, No Point   
## 17   
## Too Busy, No Point, Boring   
## 2   
## Too Busy, Not Reader   
## 17   
## Too Busy, Not Reader, Boring   
## 2   
## Too Busy, Not Reader, Boring, Topic Dislike   
## 3   
## Too Busy, Not Reader, Topic Dislike   
## 2   
## Too Busy, Topic Dislike   
## 21   
## Topic Dislike   
## 91

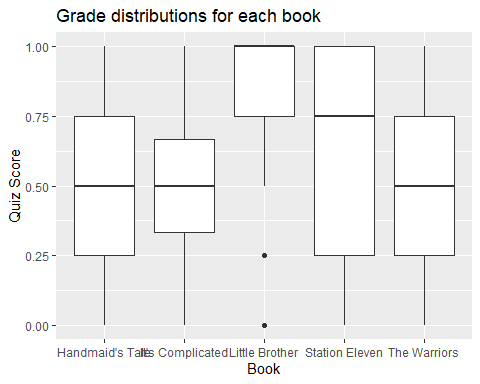
summary(summerRead$Book)

## Handmaid's Tale It's Complicated Little Brother Station Eleven   
## 540 472 361 524   
## The Warriors   
## 399

Figure 3. This figure illustrates the distribution of reasons provided for not reading the assigned summer-read by which book the student was supposed to read.

Covariation between a categorical variable and a quantitative variable:

ggplot(data = summerRead) +  
 geom\_boxplot(mapping = aes(x = Book, y = QuizScore), na.rm = TRUE) +  
 labs(title = "Grade distributions for each book", x = "Book", y = "Quiz Score")



summary(summerRead$Book)

## Handmaid's Tale It's Complicated Little Brother Station Eleven   
## 540 472 361 524   
## The Warriors   
## 399

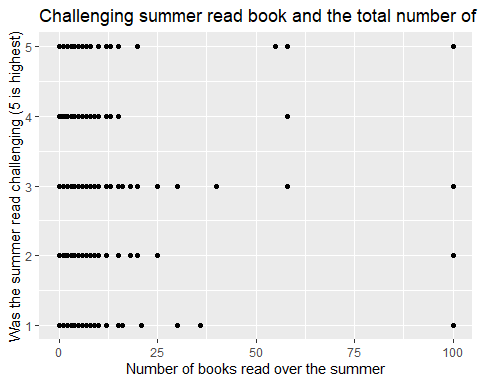
summary(summerRead$QuizScore)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0.0000 0.3333 0.5000 0.5833 0.7500 1.0000 639

Figure 4. Grade distributions for each book. Did students do well on with one book in particular?

Covariation between two quantitative variables:

ggplot(data = summerRead) +  
 geom\_point(mapping = aes(x = BooksReadSummer, y = Challenge), na.rm = TRUE) +  
 labs(title = "Challenging summer read book and the total number of books read", x = "Number of books read over the summer", y = "Was the summer read challenging (5 is highest)")



summary(summerRead$BooksReadSummer)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0.000 1.000 2.000 3.548 4.000 100.000 247

summary(summerRead$Challenge)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 1.000 2.000 3.000 2.875 4.000 5.000 723

Figure 5. Is there a relationship between number of books read, and finding the summer read challenging?