Homework1

Rachel Holman

2023-06-20

1. Download the dataset "students.txt" from Canvas. The dataset contains information on students taking an introductory statistics class at a large public university in the early 2000s. The columns of the data are:

- `Student` : ID number on survey
- `Gender` : gender of student (male / female)
- `Smoke` : whether the student smokes (yes / no)
- `Marijuan` : whether the student smokes marijuana (yes / no)
- `DrivDrnk` : whether the student has ever driven while drunk (yes / no)
- `GPA` : student's current GPA
- `PartyNum` : number of days per month the student parties
- `DaysBeer` : number of days per month the student has at least 2 alcoholic drinks
- `StudyHrs` : number of hours spent studying per week

For the questions below, you may use either base R operations or the dplyr operations (or even a combination of both).

```
students <- read.table("students.txt", header=TRUE)
head(students)
```

```
##   Student Gender Smoke Marijuan DrivDrnk  GPA PartyNum DaysBeer StudyHrs
## 1       1 female    No      Yes      Yes 3.40        4        6        7
## 2       2 female    No       No       No 3.45        4        0       20
## 3       3   male    No       No      Yes 3.89        9        4       30
## 4       4 female    No       No       No 3.75        6        3       12
## 5       5   male   Yes      Yes      Yes 2.30       10       15       14
## 6       6 female   Yes      Yes       No 2.80        2        5       10
```

a. **Looking at the variables above, is there a variable that will definitely not be part of any meaningful analysis? If yes, which one, and remove this variable from your data frame.**

The `Student` variable will definitly not be part of any meaningful analysis because it is simply an indexing variable.

```
students <- students%>%
  select(-Student)
head(students)
```

```
##   Gender Smoke Marijuan DrivDrnk  GPA PartyNum DaysBeer StudyHrs
## 1 female    No      Yes      Yes 3.40        4        6        7
## 2 female    No       No       No 3.45        4        0       20
## 3   male    No       No      Yes 3.89        9        4       30
## 4 female    No       No       No 3.75        6        3       12
## 5   male   Yes      Yes      Yes 2.30       10       15       14
## 6 female   Yes      Yes       No 2.80        2        5       10
```

b. **How many students are there in this data set?**

```
nrow(students)
```

```
## [1] 249
```

There are *249* students in this data set.

c. **How many students have a missing entry in at least one of the columns?**

```
missing <- students[!complete.cases(students),] #find rows with missing data
nrow(missing)
```

```
## [1] 12
```

There are *12* students that have a missing entry in at least one of the columns.

d. **Report the median values of the numeric variables.**

```
students%>%
  summarize(medGPA=median(GPA,na.rm = T),medPartyNum=median(PartyNum, na.rm=T),medDaysBe
er=median(DaysBeer, na.rm=T),medStudyHrs=median(StudyHrs,na.rm = T))
```

```
##   medGPA medPartyNum medDaysBeer medStudyHrs
## 1    3.2           8           8          14
```

The median GPA is 3.2, the median number of days per month spent at parties is 8, the median number of days per month students have at least 2 alcoholic drinks is 8, and the median number of hours spent studying per week is 14.

e. **Compare the mean, standard deviation, and median StudyHrs between female and male students. Based on these values, comment on what you can glean about time spent studying between female and male students.**

```
students%>%
  group_by(Gender)%>%
  summarize(meanStudyHrs=mean(StudyHrs,na.rm=T), sdStudyHrs=sd(StudyHrs, na.rm=T), medSt
udyHrs=median(StudyHrs, na.rm=T))
```

```
## # A tibble: 2 × 4
##   Gender meanStudyHrs sdStudyHrs medStudyHrs
##   <chr>         <dbl>      <dbl>       <dbl>
## 1 female         15.4       8.97          14
## 2 male           14.7      10.2           12
```

Based on the output produced above, it seems female students spend more hours studying both on average and as a median than male students do. Addisionally, female students have a smaller standard deviation which shows that there is less variation in the range of hours spent studying amongst females than males.

f. **Create a new variable called PartyAnimal, which takes on the value "yes" if PartyNum the student parties a lot (more than 8 days a month), and "no" other- wise.**

```
students<-students%>%
  mutate(PartyAnimal = ifelse(PartyNum > 8, "yes", "no"))
head(students)
```

```
##   Gender Smoke Marijuan DrivDrnk  GPA PartyNum DaysBeer StudyHrs PartyAnimal
## 1 female    No      Yes      Yes 3.40        4        6        7          no
## 2 female    No       No       No 3.45        4        0       20          no
## 3   male    No       No      Yes 3.89        9        4       30         yes
## 4 female    No       No       No 3.75        6        3       12          no
## 5   male   Yes      Yes      Yes 2.30       10       15       14         yes
## 6 female   Yes      Yes       No 2.80        2        5       10          no
```

g. **Create a new variable called GPA.cat, which takes on the following values**

- `low` if GPA is less than 3.0
- `moderate` if GPA is less than 3.5 and at least 3.0
- `high` if GPA is at least 3.5

```
students<-students%>%
  mutate(GPA.cat=cut(GPA, breaks = c(-Inf, 3.0, 3.5, Inf),
                     labels = c("low", "moderate", "high"),
                     right=FALSE)) #changing intervals to be closed on the left so 3.0
="moderate"

head(students)
```

```
##   Gender Smoke Marijuan DrivDrnk  GPA PartyNum DaysBeer StudyHrs PartyAnimal
## 1 female    No      Yes      Yes 3.40        4        6        7          no
## 2 female    No       No       No 3.45        4        0       20          no
## 3   male    No       No      Yes 3.89        9        4       30         yes
## 4 female    No       No       No 3.75        6        3       12          no
## 5   male   Yes      Yes      Yes 2.30       10       15       14         yes
## 6 female   Yes      Yes       No 2.80        2        5       10          no
##    GPA.cat
## 1 moderate
## 2 moderate
## 3     high
## 4     high
## 5      low
## 6      low
```

h. **Suppose we want to focus on students who have low GPAs (below 3.0), party a lot (more than 8 days a month), and study little (less than 15 hours a week). Create a data frame that contains these students. How many such students are there?**

```
badStudents <- students %>%
  filter(GPA.cat=="low" & PartyAnimal=="yes" & StudyHrs<15)
nrow(badStudents)
```

```
## [1] 29
```

There are *29* students who have low GPAs (below 3.0), party a lot (more than 8 days a month), and study little (less than 15 hours a week).

i. **Produce a frequency table of the number of students in each level of GPA.cat. If needed, be sure to arrange the order of the output appropriately. How many students are in each level of GPA.cat?**

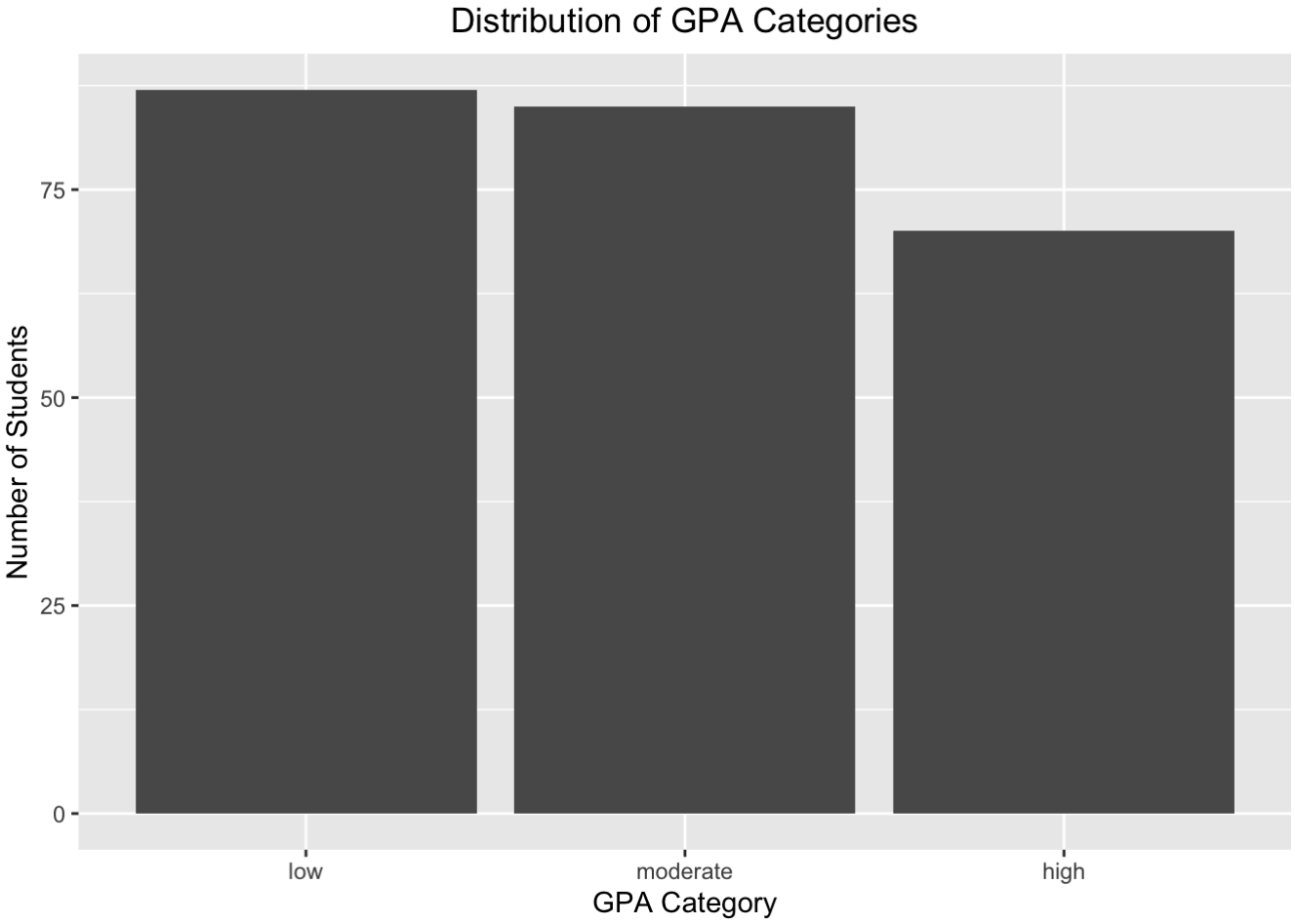
```
table(students$GPA.cat)
```

```
##
##      low moderate     high
##       87       85       70
```

There are 87 students with a GPA categorized as “low”, 85 with a GPA categorized as “moderate”, and 70 with “high” GPA values.

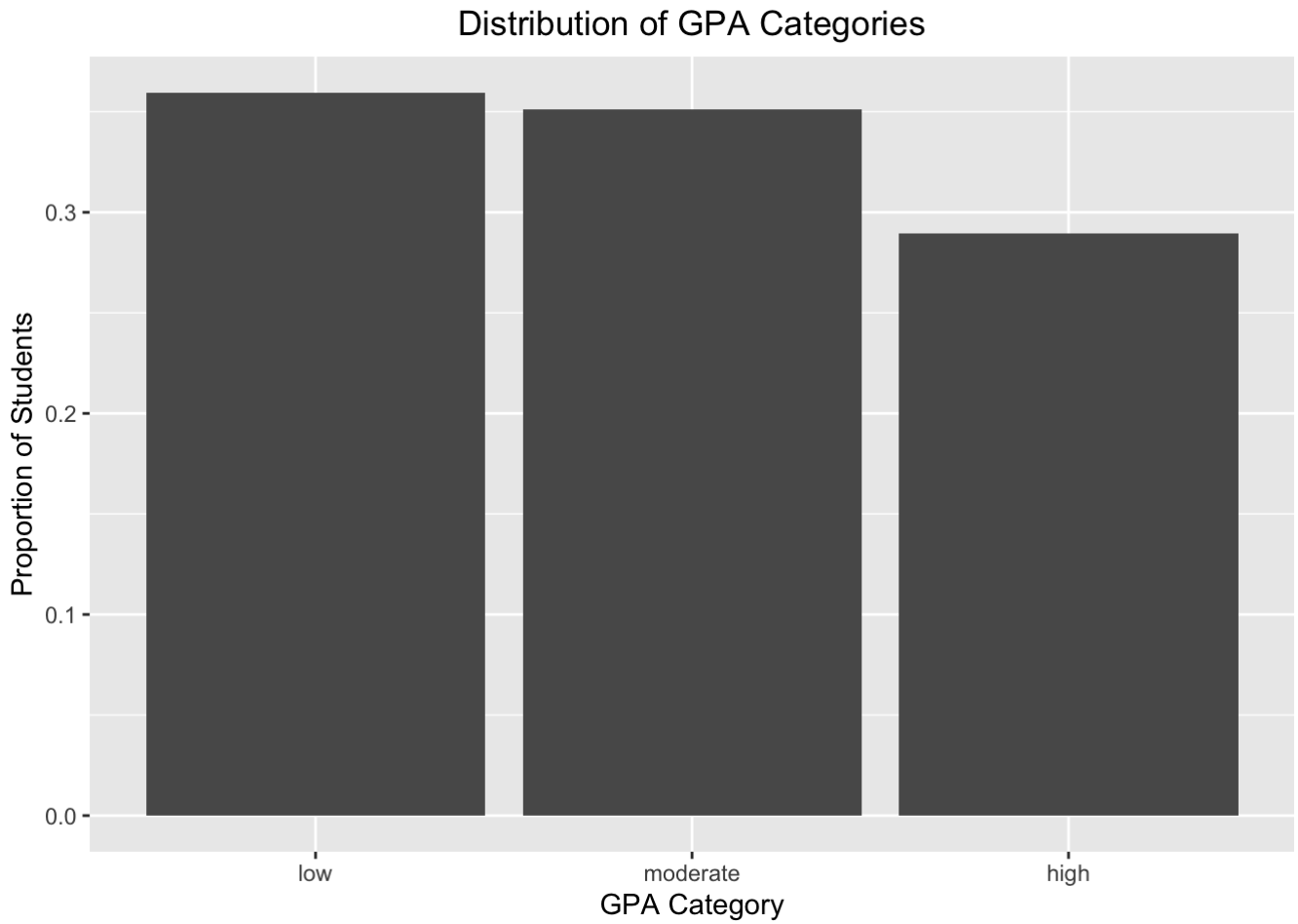
j. **Produce a bar chart that summarizes the number of students in each level of GPA.cat. Be sure to add appropriate labels and titles so that the bar chart conveys its message clearly to the reader. Be sure to remove the bar corresponding to the missing values.**

```
substudents <- subset(students, GPA.cat != "NA")
#substudents <- students %>% filter(!is.na(GPA.cat))

ggplot(substudents, aes(x=GPA.cat))+
  geom_bar()+
  theme(plot.title = element_text(hjust = 0.5))+
  labs(x="GPA Category", y="Number of Students", title="Distribution of GPA Categories")
```



k. **Create a similar bar chart as you did in part 1j, but with proportions instead of counts. Be sure to remove the bar corresponding to the missing values.**

```
pctGPA<-substudents%>%
  group_by(GPA.cat)%>%
  summarize(Counts=n())%>%
  mutate(Percent=Counts/nrow(substudents))

ggplot(pctGPA, aes(x=GPA.cat, y=Percent))+
  geom_bar(stat="identity")+
  theme(plot.title = element_text(hjust = 0.5))+
  labs(x="GPA Category", y="Proportion of Students", title="Distribution of GPA Categori
es")
```



l. **Produce a frequency table for the number of female and male students and the GPA category.**

```
genderGPAcnt <- table(students$Gender, students$GPA.cat)
genderGPAcnt
```

```
##
##          low moderate high
##   female  41       52   46
##   male    46       33   24
```

m. **Produce a table for the percentage of GPA category for each gender. For the percentages, round to 2 decimal places. Comment on the relationship between gender and GPA category.**

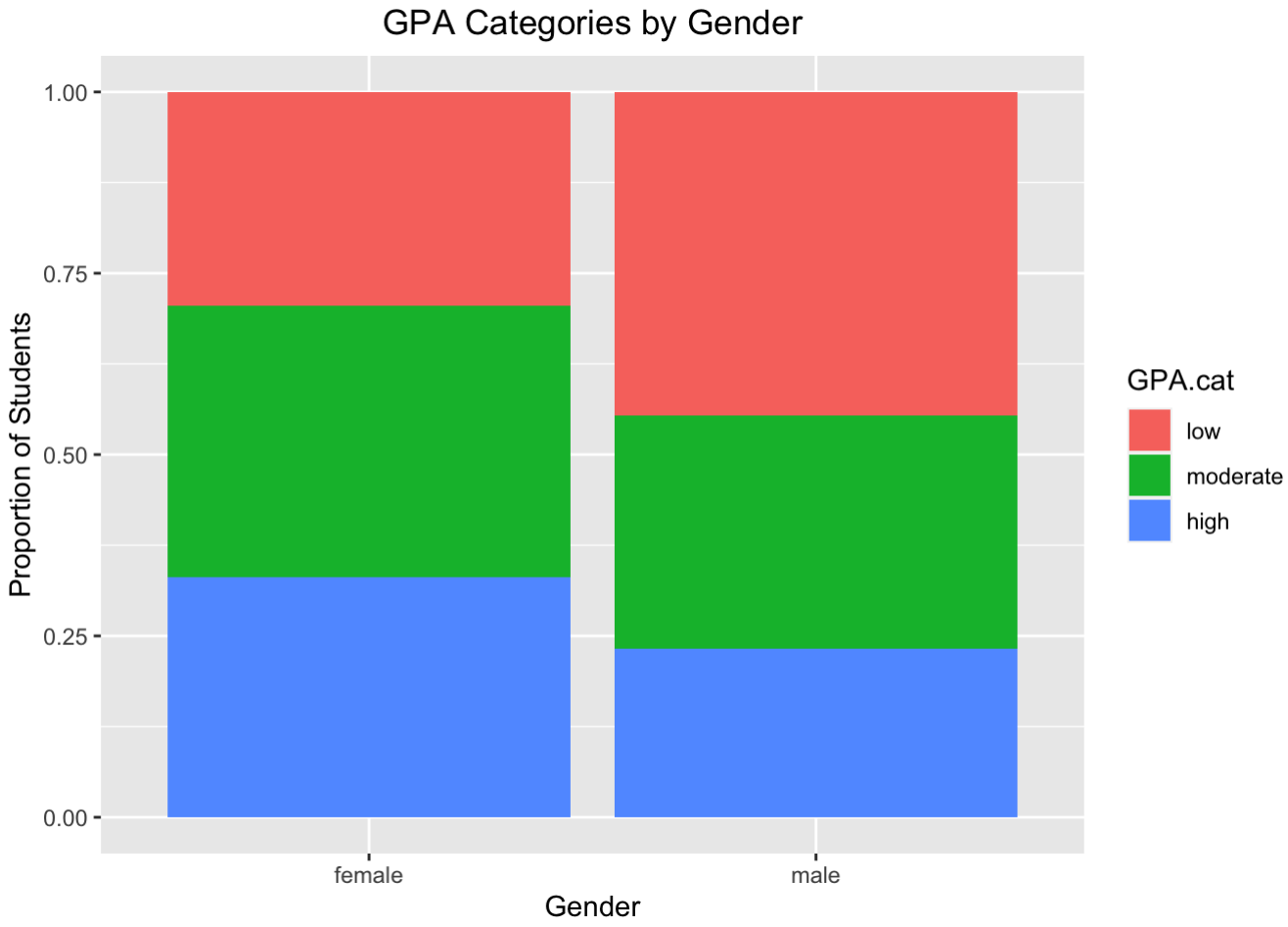
```
genderGPApct <- round(prop.table(genderGPAcnt, 1) * 100, 2)
genderGPApct
```

```
##
##            low moderate  high
##   female 29.50    37.41 33.09
##   male   44.66    32.04 23.30
```

Female students seem to be harder working than male students because they earn higher GPA's overall. More students with high and moderate GPA's are females while male students have more of the low GPA's.

n. **Create a bar chart to explore the proportion of GPA categories for female and male students. Be sure to remove the bar corresponding to the missing values.**

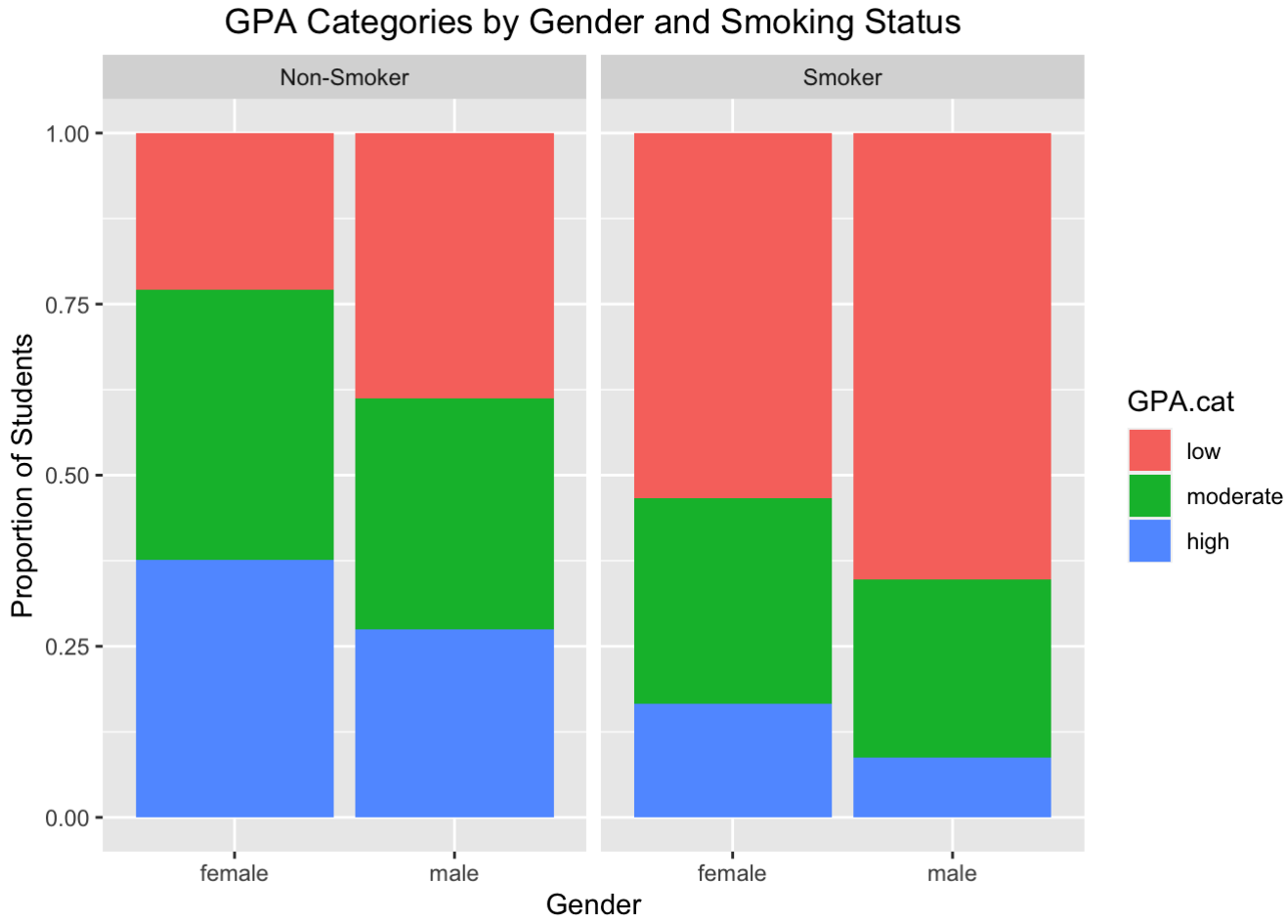
```
pctgenderGPA<-substudents%>%
  group_by(GPA.cat, Gender)%>%
  summarize(Counts=n())%>%
  mutate(Percent=Counts/nrow(substudents))

ggplot(pctgenderGPA, aes(x=Gender, y=Percent, fill=GPA.cat))+
  geom_bar(stat="identity", position="fill")+
  theme(plot.title = element_text(hjust = 0.5))+
  labs(x="Gender", y="Proportion of Students", title="GPA Categories by Gender")
```



The plot generated above shows that female students have a lower proportion of “low” GPA values than male students. Females have a larger proportion of “moderate” and “high” grades thsan male students.

o. **Create a similar bar chart similar to the bar chart in part 1n, but split by smoking status. Comment on this bar chart.**

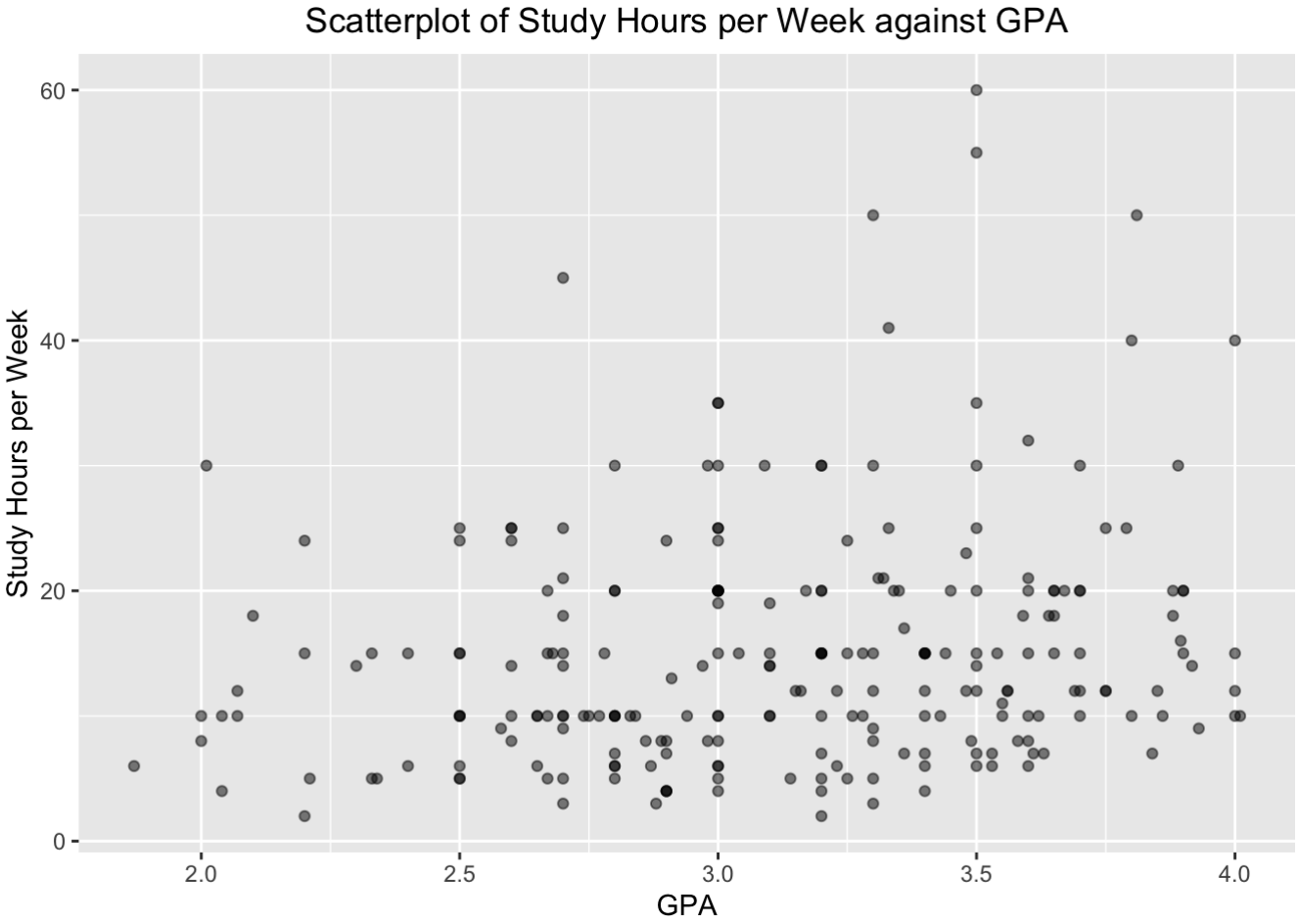
```
genderGPApctSmoke<-substudents%>%
  group_by(Gender, GPA.cat, Smoke)%>%
  summarize(Counts=n())%>%
  mutate(Percent=Counts/nrow(substudents))

genderGPApctSmoke$Smoke <- recode_factor(genderGPApctSmoke$Smoke, "No"="Non-Smoker", "Ye
s"="Smoker")

ggplot(genderGPApctSmoke, aes(x=Gender, y=Percent, fill=GPA.cat))+
  geom_bar(stat="identity", position="fill")+
  facet_grid(~Smoke) +
  theme(plot.title = element_text(hjust = 0.5))+
  labs(x="Gender", y="Proportion of Students", title="GPA Categories by Gender and Smoki
ng Status")
```



The plot above shows very clearly that students who smoke have a much larger proportion of “low” GPA values and smaller proportion of “high” GPA values. In both the smoking and non-smoking categories, female students have larger proportions od “moderate” and “high” GPAs.

p. **Create a scatterplot of GPA against the amount of hours spent studying a week. How would you describe the relationship between GPA and amount of time spent studying?**

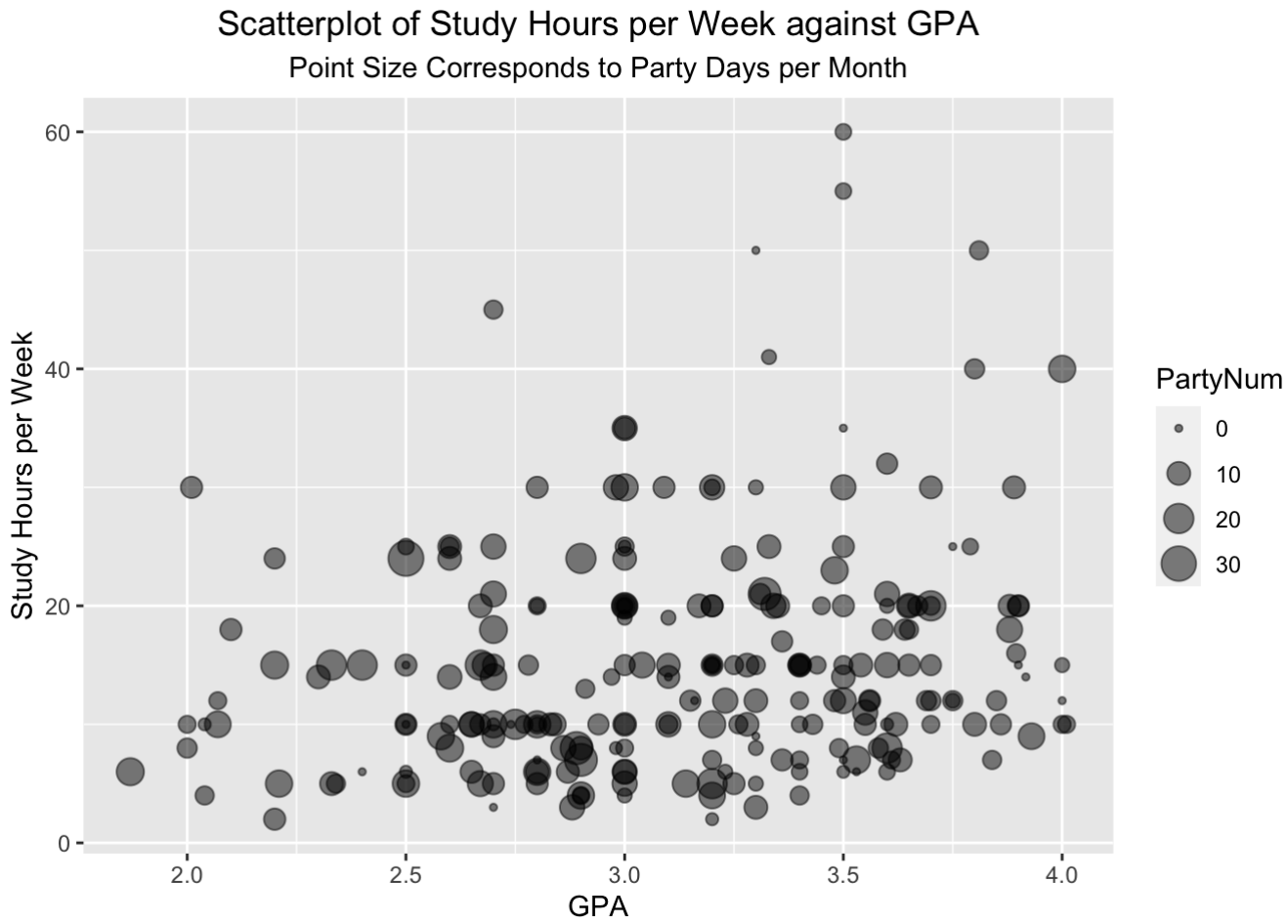
```
naomitstudents <- na.omit(students)

ggplot(naomitstudents, aes(x=GPA,y=StudyHrs))+
  geom_point(alpha=0.5)+
  theme(plot.title = element_text(hjust = 0.5))+
  labs(x="GPA", y="Study Hours per Week", title="Scatterplot of Study Hours per Week aga
inst GPA")
```



Looking at the scatter plot above, it seems that there is a very slight positive linear association between study hours and GPA. In other words, students who spent more hours studying per week seem to have higher GPA values.

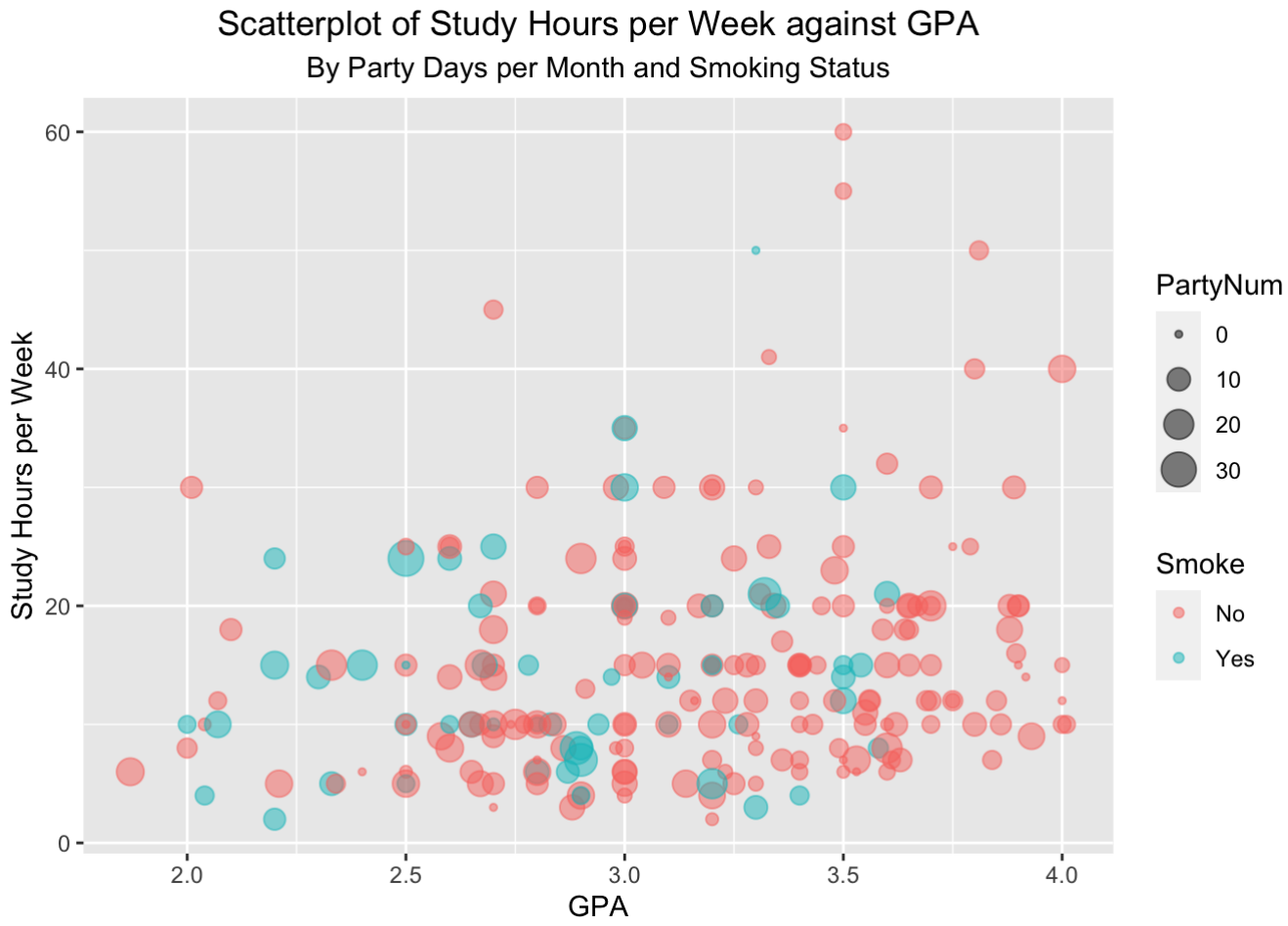
q. **Edit the scatterplot from part 1p to include information about the number of days the student parties in a month.**

```
ggplot(naomitstudents, aes(x=GPA,y=StudyHrs, size=PartyNum))+
  geom_point(alpha=0.5)+
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5))+
  labs(x="GPA", y="Study Hours per Week", title="Scatterplot of Study Hours per Week aga
inst GPA", subtitle = "Point Size Corresponds to Party Days per Month")
```



r. **Edit the scatterplot from part 1q to include information about whether the student smokes or not.**

```
ggplot(naomitstudents, aes(x=GPA,y=StudyHrs, size=PartyNum, color=Smoke))+
  geom_point(alpha=0.5)+
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5))+
  labs(x="GPA", y="Study Hours per Week", title="Scatterplot of Study Hours per Week aga
inst GPA", subtitle = "By Party Days per Month and Smoking Status")
```



2. Download the dataset UScovid.csv from Canvas. The dataset was released by The New York Times and contains data on cumulative (accruing) counts of coronavirus cases and deaths in the United States, at the state and county level, over each day from Jan 21, 2020 to June 3 2021. You may read more about the data and the variable descriptions here (please note the dataset is regularly updated, we will use the file on Canvas).

```
#clear R environment before next problem
#rm(list = ls())

covid <- read.csv("UScovid.csv", header=TRUE)
head(covid)
```

```
##         date    county      state  fips cases deaths
## 1 2020-01-21 Snohomish Washington 53061     1      0
## 2 2020-01-22 Snohomish Washington 53061     1      0
## 3 2020-01-23 Snohomish Washington 53061     1      0
## 4 2020-01-24      Cook   Illinois 17031     1      0
## 5 2020-01-24 Snohomish Washington 53061     1      0
## 6 2020-01-25    Orange California  6059     1      0
```

For this question, we focus on data at the county level.

a. **We are interested in the data on June 3 2021. Create a data frame called latest that:**

- has only rows pertaining to data from June 3 2021,
- removes rows pertaining to counties that are “Unknown”,
- removes the columns date and fips,
- is ordered by county and then state alphabetically

Use the head() function to display the first 6 rows of the data frame latest.

```
latest <- covid %>%
  filter(date == "2021-06-03" & county != "Unknown") %>%
  select(-c(date, fips)) %>%
  arrange(county, state)
head(latest)
```

```
##      county          state cases deaths
## 1 Abbeville South Carolina  2599     41
## 2    Acadia      Louisiana  6703    195
## 3  Accomack       Virginia  2862     43
## 4       Ada          Idaho 52964    475
## 5     Adair           Iowa   873     32
## 6     Adair       Kentucky  1944     54
```

b. **Calculate the case fatality rate (number of deaths divided by number of cases, and call it death.rate) for each county. Report the case fatality rate as a percent and round to two decimal places. Add death.rate as a new column to the data frame latest. Display the first 6 rows of the data frame latest.**

```
latest <- latest%>%
  mutate(death.rate = round((deaths/cases)*100, 2))
head(latest)
```

```
##      county          state cases deaths death.rate
## 1 Abbeville South Carolina  2599     41       1.58
## 2    Acadia      Louisiana  6703    195       2.91
## 3  Accomack       Virginia  2862     43       1.50
## 4       Ada          Idaho 52964    475       0.90
## 5     Adair           Iowa   873     32       3.67
## 6     Adair       Kentucky  1944     54       2.78
```

c. **Display the counties with the 10 largest number of cases. Be sure to also display the number of deaths and case fatality rates in these counties, as well as the state the counties belong to.**

```
mostcases <- latest %>%
  arrange(desc(cases))
head(mostcases, 10)
```

```
##            county      state   cases deaths death.rate
## 1     Los Angeles California 1245127  24375       1.96
## 2   New York City   New York  949986  33257       3.50
## 3            Cook   Illinois  554390  10893       1.96
## 4        Maricopa    Arizona  551509  10084       1.83
## 5      Miami-Dade    Florida  501925   6472       1.29
## 6          Harris      Texas  401345   6462       1.61
## 7          Dallas      Texas  303533   4082       1.34
## 8       Riverside California  300879   4614       1.53
## 9  San Bernardino California  298599   4760       1.59
## 10       San Diego California  280410   3760       1.34
```

d. **Display the counties with the 10 largest number of deaths. Be sure to also display the number of cases and case fatality rates in these counties, as well as the state the counties belong to.**

```
mostdeaths <- latest %>%
  arrange(desc(deaths))
head(mostdeaths, 10)
```

```
##            county      state   cases deaths death.rate
## 1   New York City   New York  949986  33257       3.50
## 2     Los Angeles California 1245127  24375       1.96
## 3            Cook   Illinois  554390  10893       1.96
## 4        Maricopa    Arizona  551509  10084       1.83
## 5      Miami-Dade    Florida  501925   6472       1.29
## 6          Harris      Texas  401345   6462       1.61
## 7          Orange California  272242   5070       1.86
## 8           Wayne   Michigan  164612   5048       3.07
## 9  San Bernardino California  298599   4760       1.59
## 10      Riverside California  300879   4614       1.53
```

e. **Display the counties with the 10 highest case fatality rates. Be sure to also display the number of cases and deaths in these counties, as well as the state the counties belong to. Is there sometime you notice about these counties?**

```
highestfatalty <- latest %>%
  arrange(desc(death.rate))
head(highestfatalty, 10)
```

```
##          county      state cases deaths death.rate
## 1         Grant   Nebraska    41      4       9.76
## 2        Sabine      Texas   524     45       8.59
## 3       Harding New Mexico    12      1       8.33
## 4     Petroleum    Montana    12      1       8.33
## 5         Foard      Texas   124     10       8.06
## 6       Hancock    Georgia   928     68       7.33
## 7      Glascock    Georgia   269     19       7.06
## 8        Motley      Texas   116      8       6.90
## 9       Candler    Georgia   978     67       6.85
## 10 Throckmorton      Texas    73      5       6.85
```

These counties all have fewer than 70 deaths total.

f. **Display the counties with the 10 highest case fatality rates among counties with at least 100,000 cases. Be sure to also display the number of cases and deaths in these counties, as well as the state the counties belong to.**

```
highestfatalty2 <- latest %>%
  filter(cases>100000) %>%
  arrange(desc(death.rate))
head(highestfatalty2, 10)
```

```
##           county         state  cases deaths death.rate
## 1  New York City      New York 949986  33257       3.50
## 2          Wayne      Michigan 164612   5048       3.07
## 3      Middlesex Massachusetts 134980   3761       2.79
## 4         Bergen    New Jersey 104301   2868       2.75
## 5         Macomb      Michigan 100190   2441       2.44
## 6   Philadelphia  Pennsylvania 153521   3692       2.40
## 7      St. Louis      Missouri 100195   2249       2.24
## 8      Fairfield   Connecticut 100093   2198       2.20
## 9           Pima       Arizona 116997   2406       2.06
## 10       Oakland      Michigan 118035   2368       2.01
```

g. **Display the number of cases, deaths, and case fatality rates for the following counties:**

i. Albemarle, Virginia

ii. Charlottesville city, Virginia

```
specific <- latest %>%
  filter(state=="Virginia" & (county == "Albemarle" | county == "Charlottesville city"))
head(specific)
```

```
##                 county    state cases deaths death.rate
## 1            Albemarle Virginia  5801     83       1.43
## 2 Charlottesville city Virginia  4014     57       1.42
```

3. This question is based on the same dataset from question 2. For this question, we focus on data at the state level. Note that the dataset has data on the 50 states, plus DC, Puerto Rico, Guam, Northern Mariana Islands, and the Virgin Islands. For the purpose of this question, we will consider DC, Puerto Rico, Guam, Northern Mariana Islands, and the Virgin Islands, as "states" as well.

a. **We are interested in the data on June 3 2021. Create a data frame called state.level that:**

- has 55 rows: 1 for each state, DC, and territory
- has 3 columns: name of the state, number of cases, number of deaths
- is ordered alphabetically by name of the state

Display the first 6 rows of the data frame state.level.

```
state.level <- covid %>%
  filter(date == "2021-06-03") %>%
  group_by(state) %>%
  summarise(cases=sum(cases), deaths=sum(deaths)) %>%
  arrange(state)

head(state.level)
```

```
## # A tibble: 6 × 3
##   state        cases deaths
##   <chr>        <int>  <int>
## 1 Alabama     545028  11188
## 2 Alaska       69826    352
## 3 Arizona     882691  17653
## 4 Arkansas    341889   5842
## 5 California 3793055  63345
## 6 Colorado    547961   6746
```

b. **Calculate the case fatality rate (call it state.rate) for each state. Report the case fatality rate as a percent and round to two decimal places. Add state.rate as a new column to the data frame state.level. Display the first 6 rows of the data frame state.level.**

```
state.level <- state.level%>%
  mutate(state.rate = round((deaths/cases)*100, 2))
head(state.level)
```

```
## # A tibble: 6 × 4
##   state        cases deaths state.rate
##   <chr>        <int>  <int>      <dbl>
## 1 Alabama     545028  11188       2.05
## 2 Alaska       69826    352       0.5
## 3 Arizona     882691  17653       2
## 4 Arkansas    341889   5842       1.71
## 5 California 3793055  63345       1.67
## 6 Colorado    547961   6746       1.23
```

c. **What is the case fatality rate in Virginia?**

```
virginia.rate<- state.level %>%
  filter(state=="Virginia") %>%
  select(state.rate)
virginia.rate
```

```
## # A tibble: 1 × 1
##   state.rate
##        <dbl>
## 1       1.66
```

The case fatality rate in Virginia is 1.66%. In other words, 1.66% of covid cases in Virginia resulted in death as of June 3rd 2021.

d. **What is the case fatality rate in Puerto Rico?**

```
pr.rate<- state.level %>%
  filter(state=="Puerto Rico") %>%
  select(state.rate)
pr.rate
```

```
## # A tibble: 1 × 1
##   state.rate
##        <dbl>
## 1         NA
```

There were no deaths recorded in the data set for Puerto Rico through June 3rd 2021, so the case fatality rate is 0%.

e. **Which states have the 10 highest case fatality rates?**

```
higheststaterate <- state.level %>%
  arrange(desc(state.rate))
head(higheststaterate, 10)
```

```
## # A tibble: 10 × 4
##    state                cases deaths state.rate
##    <chr>                <int>  <int>      <dbl>
##  1 New Jersey          1017044  26253       2.58
##  2 Massachusetts        707523  17893       2.53
##  3 New York            2102003  52811       2.51
##  4 Connecticut          347748   8245       2.37
##  5 District of Columbia  49041   1136       2.32
##  6 Mississippi          318048   7324       2.3
##  7 Pennsylvania        1208879  27349       2.26
##  8 Louisiana            472617  10605       2.24
##  9 New Mexico           203330   4275       2.1
## 10 Maryland             460406   9626       2.09
```

f. **Which states have the 10 lowest case fatality rates?**

```
loweststaterate <- state.level %>%
  arrange(state.rate)
head(loweststaterate, 10)
```

```
## # A tibble: 10 × 4
##    state                     cases deaths state.rate
##    <chr>                     <int>  <int>      <dbl>
##  1 Alaska                    69826    352       0.5
##  2 Utah                     406895   2308       0.57
##  3 Virgin Islands             3512     28       0.8
##  4 Vermont                   24240    255       1.05
##  5 Nebraska                 223517   2385       1.07
##  6 Idaho                    192704   2103       1.09
##  7 Northern Mariana Islands    183      2       1.09
##  8 Wisconsin                675152   7923       1.17
##  9 Wyoming                   60543    720       1.19
## 10 Colorado                 547961   6746       1.23
```

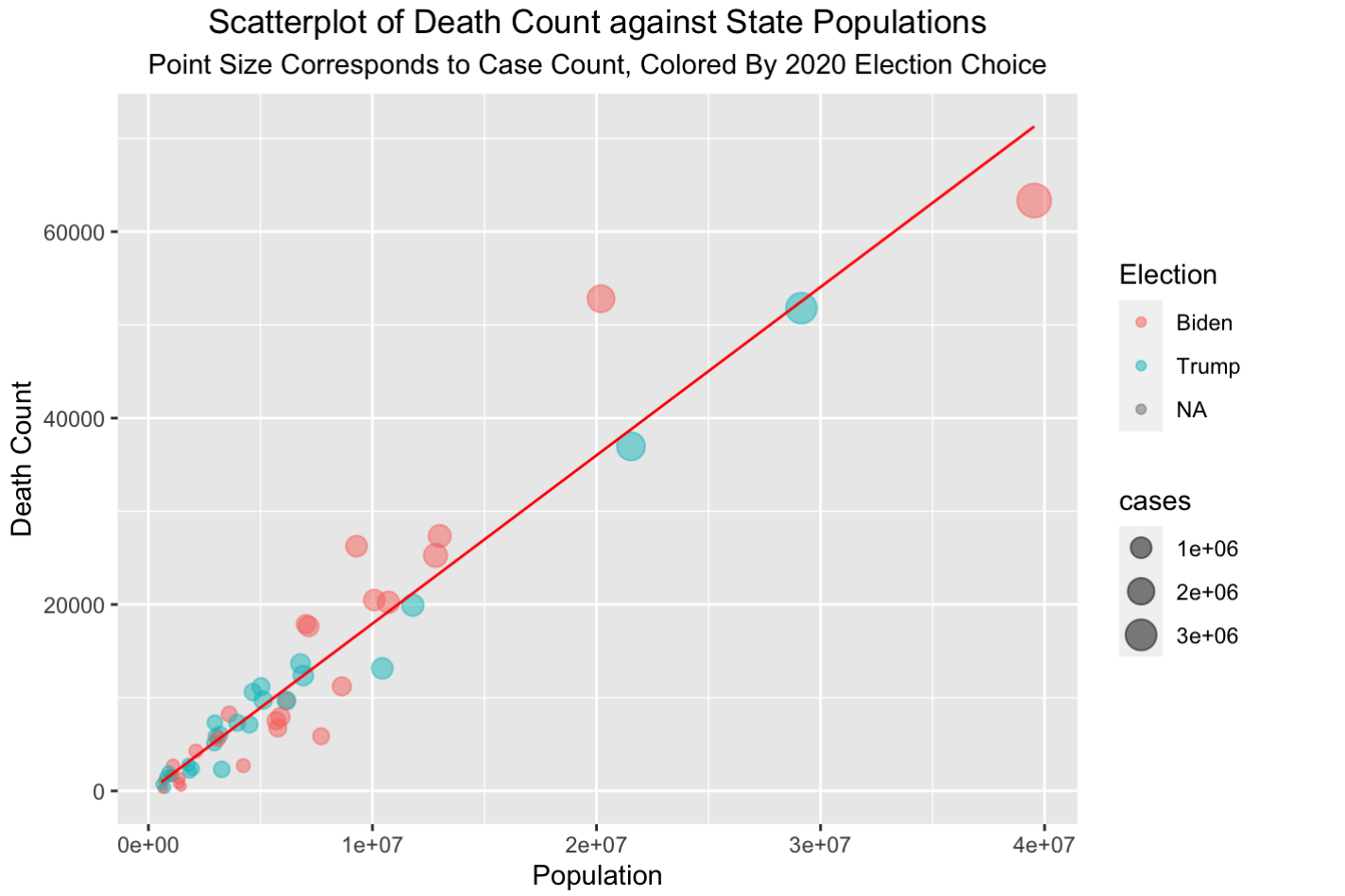
g. **There is a dataset on Canvas, called State_pop_election.csv. The dataset contains the population of the states from the 2020 census (50 states plus DC and Puerto Rico), as well as whether the state voted for Biden or Trump in the 2020 presidential elections. Merge State_pop_election.csv and the data frame state.level. Use the head() function to display the first 6 rows after merging these two datasets. Be sure to arrange the states alphabetically.**

```
election <- read.csv("State_pop_election.csv", header=TRUE)
election <- election %>%
  rename(state=State)

fullstatedata <- merge(state.level, election, by="state", all=T) %>%
  arrange(state)
head(fullstatedata)
```

```
##        state   cases deaths state.rate Population Election
## 1    Alabama  545028  11188       2.05    5024279    Trump
## 2     Alaska   69826    352       0.50     733391    Trump
## 3    Arizona  882691  17653       2.00    7151502    Biden
## 4   Arkansas  341889   5842       1.71    3011524    Trump
## 5 California 3793055  63345       1.67   39538223    Biden
## 6   Colorado  547961   6746       1.23    5773714    Biden
```

h. **Pick at least two variables from the dataset and create a suitable visualization of the variables. Comment on what the visualization reveals. You may create new variables based on existing variables, and describe how you created the new variables.**

```
ggplot(fullstatedata, aes(x=Population,y=deaths, size=cases, color=Election))+
  geom_point(alpha=0.5)+
  geom_smooth(method = "lm", se=FALSE, color="red", size=0.5)+
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5))+
  labs(x="Population", y="Death Count", title="Scatterplot of Death Count against State Populations", subtitle = "Point Size Corresponds to Case Count, Colored By 2020 Election Choice")
```



The visual created above shows that there is a positive linear association between population size and death count. Additionally, the states with larger death counts also have a larger case count on average.

---

4. We will look at a data set concerning adult penguins near Palmer station, Antarctica. The data set, penguins comes from the palmerpenguins package. Be sure to install and load the palmerpenguins package. I recommend reading the documentation of this data set by typing ?penguins

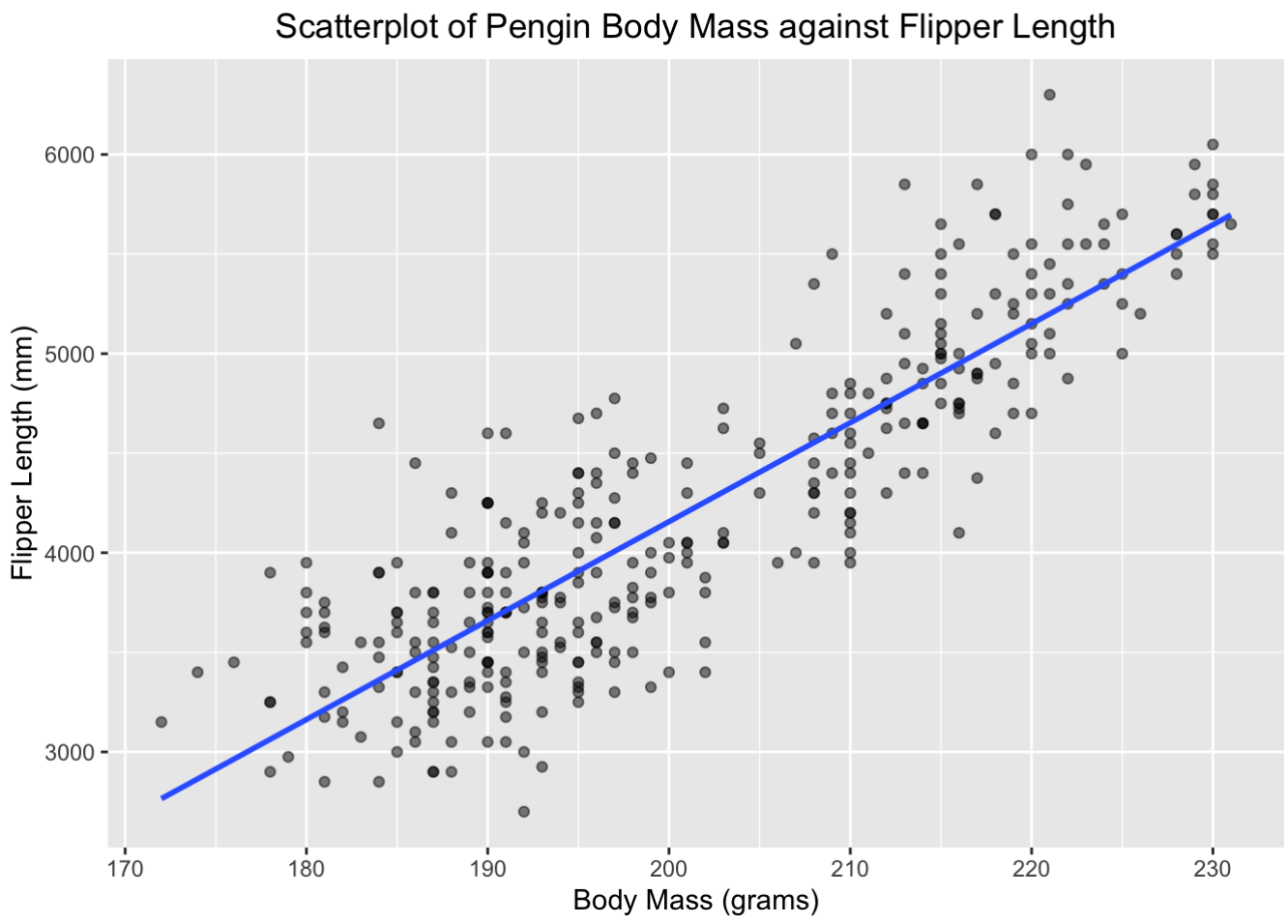
```
#clear R environment before next problem
#rm(list = ls())

library(palmerpenguins)
head(penguins)
```

```
## # A tibble: 6 × 8
##   species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
##   <fct>   <fct>              <dbl>         <dbl>             <int>       <int>
## 1 Adelie  Torgersen           39.1          18.7               181        3750
## 2 Adelie  Torgersen           39.5          17.4               186        3800
## 3 Adelie  Torgersen           40.3          18                 195        3250
## 4 Adelie  Torgersen           NA            NA                  NA          NA
## 5 Adelie  Torgersen           36.7          19.3               193        3450
## 6 Adelie  Torgersen           39.3          20.6               190        3650
## # ℹ 2 more variables: sex <fct>, year <int>
```

We will explore the relationship between the response variable body mass (in grams), body_mass_g, and the predictor length of the flippers (in mm), flipper_length_mm.

a. **Produce a scatterplot of the two variables. How would you describe the relationship between the two variables? Be sure to label the axes and give an appropriate title. Based on the appearance of the plot, does a simple linear regression appear reasonable for the data?**

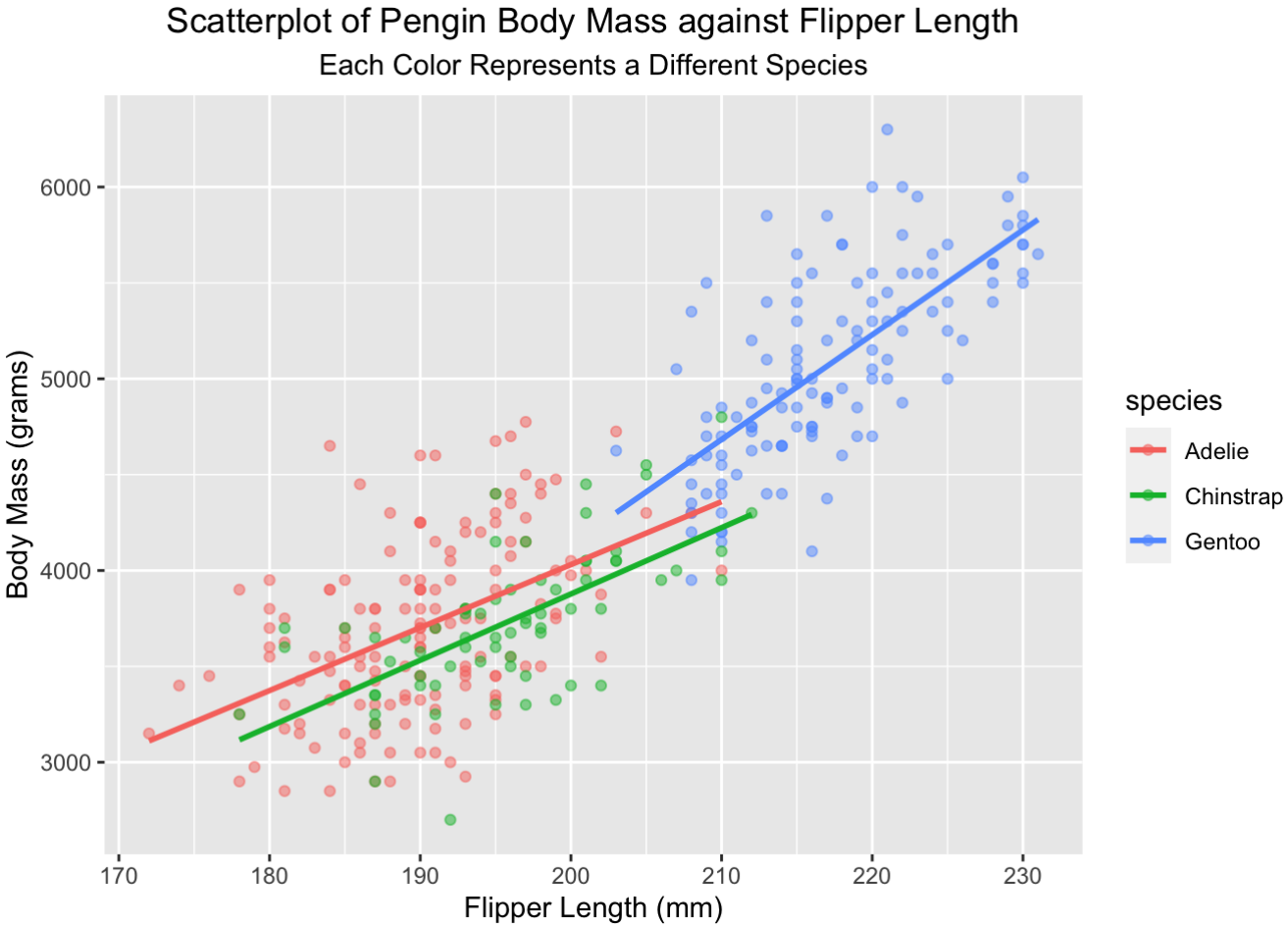
```
ggplot(penguins, aes(x=flipper_length_mm,y=body_mass_g,))+
  geom_point(alpha=0.5)+
  geom_smooth(method = "lm", se=FALSE)+
  theme(plot.title = element_text(hjust = 0.5))+
  labs(x="Body Mass (grams)", y="Flipper Length (mm)", title="Scatterplot of Pengin Body
Mass against Flipper Length")
```



Penguin body mass (in grams) and flipper length (in mm) seem to be very positively linearly associated, which suggests that a simple linear regression does appear reasonable for the data. Additionally, the observations are fairly evenly scattered on both sides of the regression line, so a linear association exists.

b. **Produce a similar scatterplot, but with different colored plots for each species. How does this scatterplot influence your answer to the previous part?**

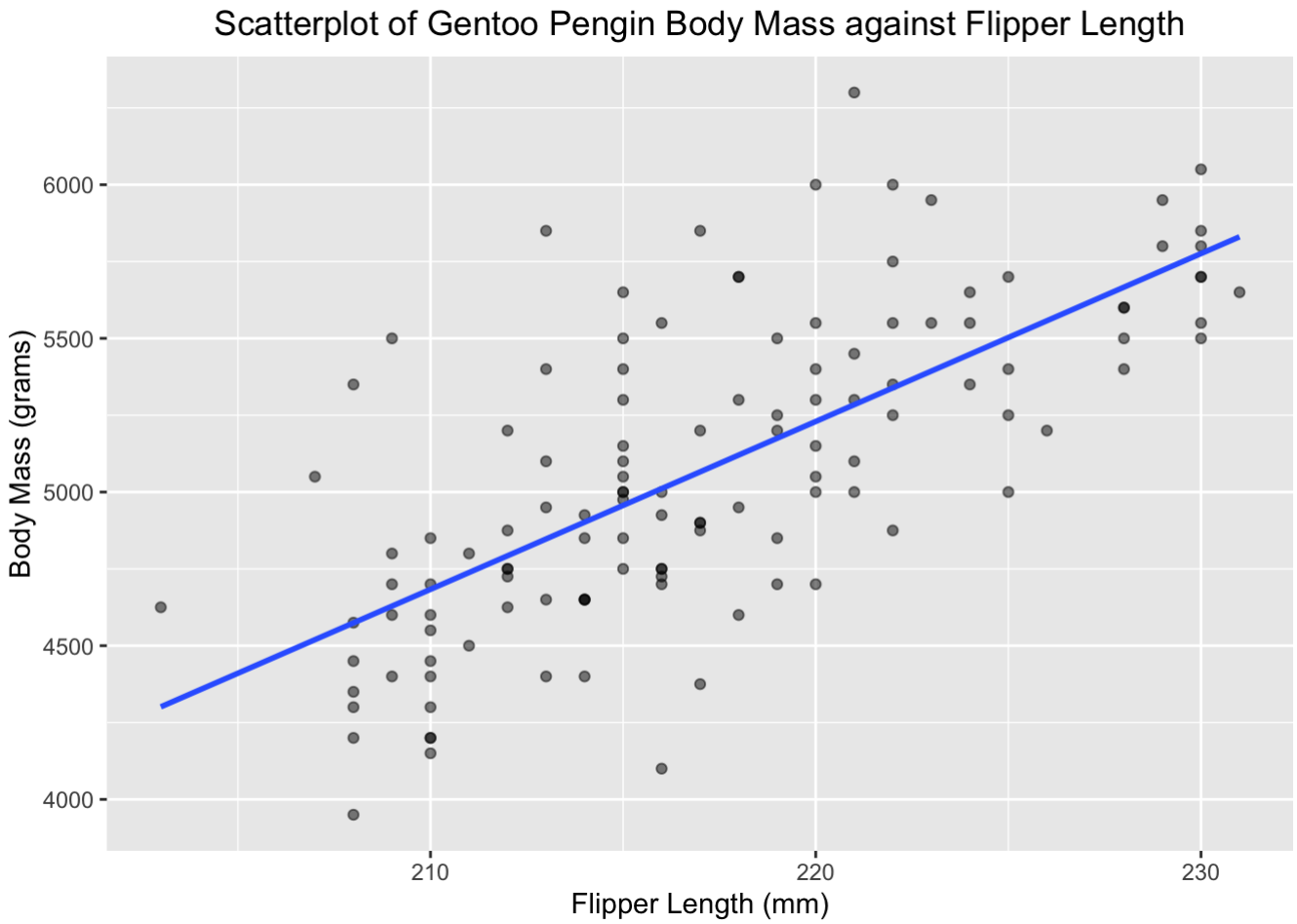
```
ggplot(penguins, aes(x=flipper_length_mm,y=body_mass_g, color=species))+
  geom_point(alpha=0.5)+
  geom_smooth(method = "lm", se=FALSE)+
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5))+
  labs(x="Flipper Length (mm)", y="Body Mass (grams)", title="Scatterplot of Pengin Body
Mass against Flipper Length", subtitle = "Each Color Represents a Different Species")
```



This scatterplot shows that although each species may be well suited for a simple linear regression, all the points seem to follow a nonlinear path overall.

c. **Regardless of your answer to the previous part, produce a scatterplot of body mass and flipper length for Gentoo penguins. Based on the appearance of the plot, does a simple linear regression appear reasonable for the data?**

```
gentoo <- penguins %>%
  filter(species=="Gentoo")

ggplot(gentoo, aes(x=flipper_length_mm,y=body_mass_g))+
  geom_point(alpha=0.5)+
  geom_smooth(method = "lm", se=FALSE)+
  theme(plot.title = element_text(hjust = 0.5))+
  labs(x="Flipper Length (mm)", y="Body Mass (grams)", title="Scatterplot of Gentoo Peng
in Body Mass against Flipper Length")
```



Gentoo penguin body mass (in grams) and flipper length (in mm) seem to be positively linearly associated, which suggests that a simple linear regression does appear reasonable for the data. Additionally, the observations are fairly evenly scattered on both sides of the regression line, so a linear association exists.

d. **What is the correlation between body mass and flipper length for Gentoo penguins. Interpret this correlation contextually. How reliable is this interpretation?**

```
cor(gentoo$body_mass_g, gentoo$flipper_length_mm, use = "complete.obs")
```

```
## [1] 0.7026665
```

The correlation between body mass and flipper length for Gentoo penguins is 0.7026665 which is a positive, moderately strong linear correlation. This is not entirely reliable because there were incomplete, or NA, observations which had to be omitted.

For the rest of the questions, assume the assumptions to perform linear regression on Gentoo penguins are met.

e. **Use the lm() function to fit a linear regression for body mass and flipper length for Gentoo penguins. Write out the estimated linear regression equation.**

```
gentoolm<-lm(body_mass_g~flipper_length_mm, data=gentoo)

summary(gentoolm)
```

```
##
## Call:
## lm(formula = body_mass_g ~ flipper_length_mm, data = gentoo)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -911.18 -235.76  -51.93  170.75 1015.71
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -6787.281   1092.552  -6.212 7.65e-09 ***
## flipper_length_mm    54.623      5.028  10.863  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 360.2 on 121 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.4937, Adjusted R-squared:  0.4896
## F-statistic:   118 on 1 and 121 DF,  p-value: < 2.2e-16
```

$$Y = \beta_0 + \beta_1 X + e$$
$$(\text{body\_mass\_g}) = -6787.281 + 54.623(\text{flipper\_length\_mm}) + 360.2$$

f. **Interpret the estimated slope contextually.**

```
##extract slope
summary(gentoolm)$coefficients[2,1]
```

```
## [1] 54.6225
```

$\hat{B_1}$ = 54.6225. The estimated slope informs us the the predicted body mass increases by 54.6225 grams per unit (1 mm) increase in flipper length.

g. **Does the estimated intercept make sense contextually?**

```
##extract intercept
summary(gentoolm)$coefficients[1,1]
```

```
## [1] -6787.281
```

$\hat{B_0}$ = -6787.281. For gentoo penguins with no flipper length, their predicted body mall is -6787.281 grams This value does **not** make sense contextually because a penguin cannot have 0 millimeter flippers. The minimum flipper length value in our data is 203 millimeters.

h. **Report the value of** $R^2 $ **from this linear regression, and interpret its value contextually.**

```
#extract r2
summary(gentoolm)$r.squared
```

```
## [1] 0.4937402
```

R^2 = 0.4937402. The coefficient of determination informs us that about 49.37% of the variation in body mass (grams) can be explained by flipper length (mm).

i. **What is the estimated value for the standard deviation of the error terms for this regression model,** $\hat{\sigma}$?

```
#extract sigma
summary(gentoolm)$sigma
```

```
## [1] 360.1676
```

s = 360.1676, is the estimate of the standard deviation of the error terms. This is reported as residual standard error in R. Squaring this gives the estimated variance.

j. **For a Gentoo penguin which has a flipper length of 220mm, what is its predicted body mass in grams?**

```
##create data point for prediction
newdata1<-data.frame(flipper_length_mm=220)
##predicted body mass when x=200
predict(gentoolm,newdata1)
```

```
##       1
## 5229.67
```

This gentoo penguin's predicted body mass is 5229.67 grams.

k. **Produce the ANOVA table for this linear regression. Using only this table, calculate the value of** R^2.

```
anova.tab<-anova(gentoolm)
anova.tab
```

```
## Analysis of Variance Table
##
## Response: body_mass_g
##                    Df   Sum Sq  Mean Sq F value    Pr(>F)
## flipper_length_mm   1 15308045 15308045  118.01 < 2.2e-16 ***
## Residuals         121 15696203   129721
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
SST<-sum(anova.tab$"Sum Sq")

R2 <- anova.tab$"Sum Sq"[1]/SST
R2
```

```
## [1] 0.4937402
```

The R^2 value of 0.4937402 was computed using only the ANOVA table by first computing the Sum of Squares Total ($SS_R + SS_{res}$), then dividing the SS_R by the SS_T.

l. **What are the null and alternative hypotheses for the ANOVA F test?**

$H_0 : \beta_1 = 0$ (The slope of the SLR is equal to 0, so the two variables are uncorrelated)
$H_A : \beta_1 \neq 0$ (The slope of the SLR is not equal to 0, so the two variables are correlated)

m. **Explain how the F statistic of 118.01 is found.**

The F statistic of 118.01 is found using by dividing the MS_R by the MS_{res}:

```
anova.tab$`Mean Sq`[1]/anova.tab$`Mean Sq`[2]
```

```
## [1] 118.0077
```

n. **Write an appropriate conclusion for the ANOVA F test for this simple linear regression model.**

```
critval <- qf(1-0.05, 1, 123-2)
critval
```

```
## [1] 3.919465
```

Since out test statistic is larger than the critical value, we reject the null hypothesis. Our data supports the claim that the slope is different from 0, or in other words, that there is a linear association between gentoo penguin flipper length (mm) and body mass (grams).

o. **Report the 95% confidence interval for the change in the predicted body mass (in grams) when flipper length increases by 1mm.**

```
confint(gentoolm,level = 0.95)
```

```
##                          2.5 %      97.5 %
## (Intercept)       -8950.27535 -4624.28587
## flipper_length_mm    44.66777    64.57724
```

The 95% CI for β_1 is (44.66777, 64.57724). We have 95% confidence that for each additional millimeter in gentoo flipper length, the predicted body mass increases between 44.66777 grams and 64.57724 grams.

p. **Are your results from parts 4n and 4o consistent? Briefly explain.**

Yes, the confidence interval excluded 0, so the results from parts 4n and 4o are consistant. Part 4n shows that a one unit increase in flipper length will have a non-zero impact on body mass. Part 4o shows that the slope, or change to body mass with a 1 unit increase in flipper length is between 44.66777 grams and 64.57724 grams. Since zero is not included in that confidence interval, we are 95% confident that the slope is not equal to 0.

q. **Estimate the mean body mass (in grams) for Gentoo penguins with flipper lengths of 200mm. Also report the 95% confidence interval for the mean body mass (in grams) for Gentoo penguins with flipper lengths of 200mm.**

```
newgentoo<-data.frame(flipper_length_mm=200)
predict(gentoolm,newgentoo,level=0.95, interval="confidence")
```

```
##       fit      lwr      upr
## 1 4137.22 3954.446 4319.993
```

The estimate for the mean body mass (in grams) for Gentoo penguins with flipper lengths of 200mm is 4137.22 grams.
The 95% CI for the mean body mass in grams for Gentoo penguins with a flipper length of 200 millimeters is (3954.446, 4319.993). We have 95% confidence the mean body mass in grams for Gentoo penguins with a flipper length of 200 millimeters is between 3954.446 grams and 4319.993 grams.

r. **Report the 95% prediction interval for the body mass (in grams) of a Gentoo penguin with flipper length of 200mm.**

```
predict(gentoolm,newgentoo,level=0.95, interval="prediction")
```

```
##       fit      lwr      upr
## 1 4137.22 3401.121 4873.319
```

We have 95% confidence that for a Gentoo penguin with a flipper length of 200 mm, this penguins's body mass is between 3401.121 grams and 4873.319 grams. We note that the 95% prediction interval is wider than the 95% confidence interval because prediction intervals account for more uncertainty for individual outcomes.

s. **A researcher hypothesizes that for Gentoo penguins, the predicted body mass increases by more than 50 g for each additional mm in flipper length. Conduct an appropriate hypothesis test. What is the null and alternative hypotheses, test statistic, and conclusion?**

```
n=123
b1 = 50
b1hat = summary(gentoolm)$coefficients[2,1]
seB1 = summary(gentoolm)$coefficients[2,2]
t = (b1hat - b1)/(seB1)
t
```

```
## [1] 0.9193074
```

```
pvalue= 1 - pt(t, n-2)
pvalue
```

```
## [1] 0.1798819
```

```
critval = qt(0.95, n-2)
critval
```

```
## [1] 1.657544
```

$H_0 : \beta_1 = 50$ (The slope of the SLR is equal to 50, or the predicted body mass increases by exactly 50 g for each additional mm in flipper length)
$H_A : \beta_1 > 50$ (The slope of the SLR is greater than 50, or the predicted body mass increases by more than 50 g for each additional mm in flipper length)

Test statistic (t) = 0.9193074
P-value = 0.1798819
critical value = 1.657544

Since the p-value is larger than $\alpha = 0.05$ and the critical value is larger than the test statistic, we fail to reject H_0. The data does not support H_A that the predicted body mass increases by more than 50 g for each additional mm in flipper length.

---

5. We will use the dataset “copier.txt” for this question. The Tri-City Office Equipment Corporation sells an imported copier on a franchise basis and performs preventive maintenance and repair service on this copier. The data have been collected from 45 recent calls on users to perform routine preventive maintenance service; for each call, Serviced is the number of copiers serviced and Minutes is the total number of minutes spent by the service person.

```
copier <- read.table("copier.txt", header = TRUE)
head(copier)
```

```
##   Minutes Serviced
## 1      20        2
## 2      60        4
## 3      46        3
## 4      41        2
## 5      12        1
## 6     137       10
```

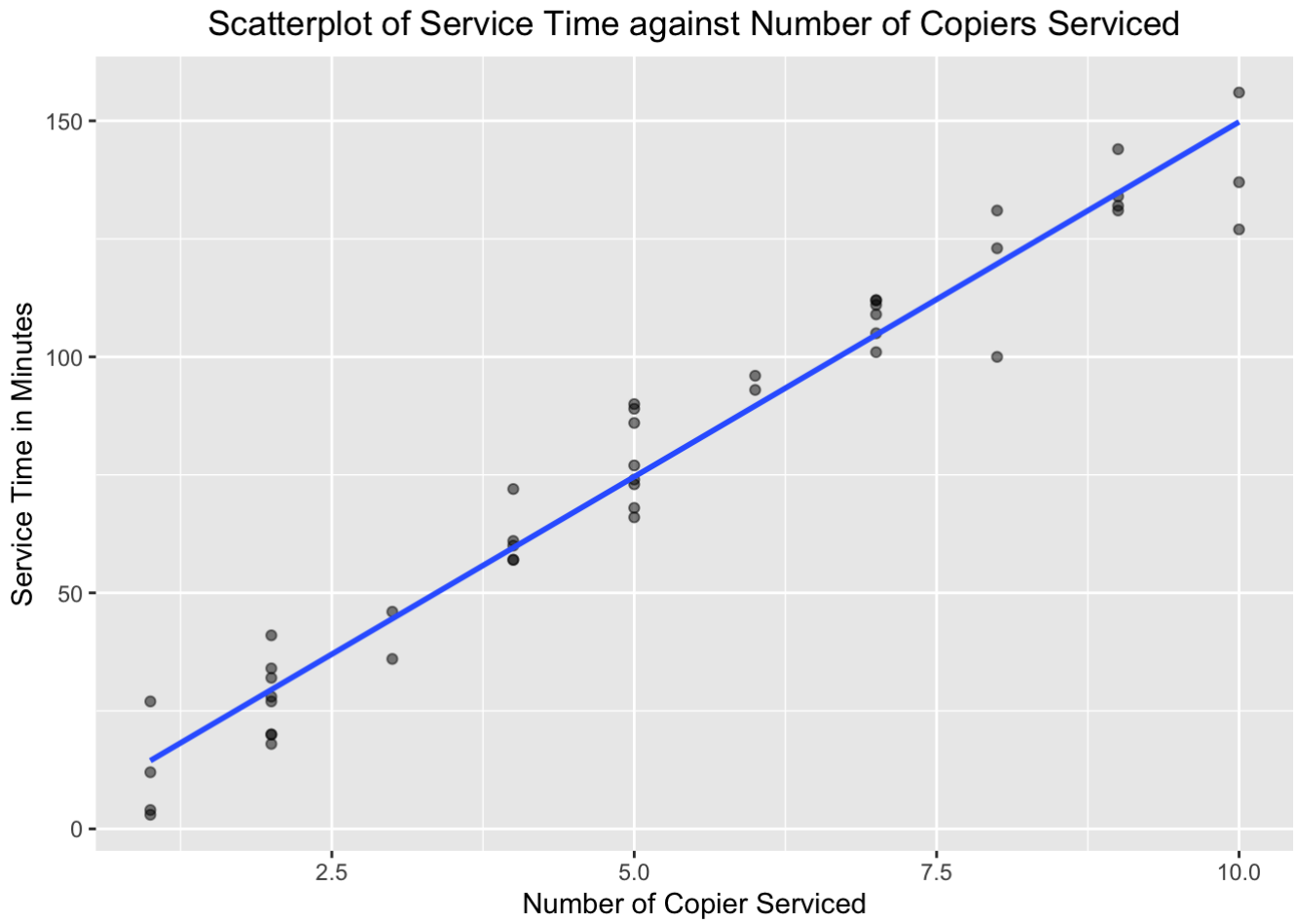
a. **What is the response variable in this analysis? What is predictor in this analysis?**

In this analysis, `Minutes` is the response variable and `Services` serves as the predictor.

b. **Produce a scatterplot of the two variables. How would you describe the relationship between the number of copiers serviced and the time spent by the service person?**

```
ggplot(copier, aes(x=Serviced,y=Minutes))+
  geom_point(alpha=0.5)+
  geom_smooth(method = "lm", se=FALSE)+
  theme(plot.title = element_text(hjust = 0.5))+
  labs(x="Number of Copier Serviced", y="Service Time in Minutes", title="Scatterplot of
Service Time against Number of Copiers Serviced")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



The number of copiers serviced and the time spent by the service person seem to be very strongly positively linearly associated, which suggests that a simple linear regression would be reasonable for the data. Additionally, the observations are fairly evenly scattered on both sides of the regression line, so a linear association exists.

c. **What is the correlation between the total time spent by the service person and the number of copiers serviced? Interpret this correlation contextually.**

```
cor(copier$Serviced, copier$Minutes)
```

```
## [1] 0.978517
```

The correlation between the total time spent by the service person and the number of copiers serviced is 0.978517 which is a very strong, positive linear association.

d. **Can the correlation found in part 5c be interpreted reliably? Briefly explain.**

The correlation of 0.978517 can be interpreted to show that there is a very strong, positive linear association between the total time spent by the service person and the number of copiers serviced. As the number of copiers to be serviced increases, so too does the amount of time a service person spends.

e. **Use the lm() function to fit a linear regression for the two variables. Where are the values of** $_1, _0, R^2, and ^2 $ **for this linear regression?**

```
copierlm <- lm(Minutes~ Serviced, data=copier)
summary(copierlm)$coefficients[1,1]
```

```
## [1] -0.5801567
```

```
summary(copierlm)$coefficients[2,1]
```

```
## [1] 15.03525
```

```
summary(copierlm)$r.squared
```

```
## [1] 0.9574955
```

```
(summary(copierlm)$sigma)^2
```

```
## [1] 79.45063
```

$\hat{\beta_0}$ = -0.5801567
$\hat{\beta_1}$ = 15.03525
R^2 = 0.9574955
$\hat{\sigma^2}$ = 79.45063

f. **Interpret the values of** $\hat{\beta_1}$ and $\hat{\beta_0}$ **contextually. Does the value of** $\hat{\beta_0}$ **make sense in this context?**

$\hat{\beta_1}$ = 15.03525. The estimated slope informs us the the predicted service time increases by 15.03525 minutes per one unit increase in the number of copiers that need to be serviced.

$\hat{\beta_0}$ = -0.5801567. When there are no copiers that need to be services, the total time spent by the service person is -0.5801567 minutes. This value does **not** make sense in this context because it is impossible to have a nugative amount of minutes, and the minimum number of copiers serviced in this dataset is 1. This is an instance of extrapolating.

g. **Use the anova() function to produce the ANOVA table for this linear regression. What is the value of the ANOVA F statistic? What null and alternative hypotheses are being tested here? What is a relevant conclusion based on this ANOVA F statistic?**

```
anova.tab<-anova(copierlm)
anova.tab
```

```
## Analysis of Variance Table
##
## Response: Minutes
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Serviced   1  76960   76960  968.66 < 2.2e-16 ***
## Residuals 43   3416      79
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
critval <- qf(1-0.05, 1, 123-2)
critval
```

```
## [1] 3.919465
```

$H_0 : \beta_1 = 0$ (The slope of the SLR is equal to 0, so the two variables are uncorrelated)
$H_A : \beta_1 \neq 0$ (The slope of the SLR is not equal to 0, so the two variables are correlated)

The ANOVA F statistic is 968.66 which is extremely high. Because it is larger than the critical value (3.919465), we reject the null hypothesis. Our data supports the claim that the slope is different from 0, or in other words, that there is a linear association between the total time spent by the service person and the number of copiers serviced.

h. **Suppose a service person is sent to service 5 copiers. Obtain an appropriate 95% interval that predicts the total service time spent by the service person.**

```
newdata2<-data.frame(Serviced=5)
predict(copierlm, newdata2, level=0.95, interval="confidence")
```

```
##        fit      lwr      upr
## 1 74.59608 71.91422 77.27794
```

We have 95% confidence that for a routine preventive maintenance service person with 5 copiers to service, the worker's total time spend servicing the copiers is between 71.91422 minutes and 77.27794 minutes.

6. (You may only use R as a simple calculator or to find p-values or critical values) Suppose that for n = 6 students, we want to predict their scores on the second quiz using scores from the first quiz. The estimated regression line is

$$\hat{y} = 20 + 0.8x.$$

a. **For each individual observation, calculate its predicted score on the second quiz $\hat{y_i}$ and the residual e_i. You may show your results in the table below.**

Predicted/Fitted values: $\hat{\hat{y}_i} = 20 + 0.8(x_i)$

Residuals: $e_i = y_i - \hat{\hat{y}_i}$

x_i	70	75	80	80	85	90
y_i	75	82	80	86	90	91
$\hat{y_i}$	76	80	84	84	88	92
e_i	-1	2	-4	2	2	-1

b. **Complete the ANOVA table for this dataset below.**

For reference:

	DF	SS	MS	F-stat	p-value
Regression	1	$\sum(\hat{\hat{y}_i} - \bar{y})^2$	$\frac{SS_R}{df_R}$	$\frac{MS_R}{MS_{res}}$	0.0099
Residual	n-2	$\sum(y_i - \hat{y})^2$	$\frac{SS_{res}}{df_{res}}$	***	***
Total	n-1	$\sum(y_i - \bar{y})^2$	***	***	***

Therefore:

	DF	SS	MS	F-stat	p-value
Regression	1	160	160	21.333	0.0099
Residual	4	30	7.5	***	***
Total	5	190	***	***	***

c. **Calculate the sample estimate of the variance σ^2 for the regression model.**

$s^2 = MS_{res} = 7.5$

d. **What is the value of R2 here? Interpret this value in context.**

$R^2 = \frac{SS_R}{SS_T} = \frac{160}{190} = 0.84210526$

R^2 = 0.84210526. The coefficient of determination informs us that about 84.21% of the variation in a students' second quiz score can be explained by the students' first quiz score.

e. **Carry out the ANOVA F test. What is an appropriate conclusion?**

$H_0 : \beta_1 = 0$ (The slope of the SLR is equal to 0, so the two variables are uncorrelated)
$H_A : \beta_1 \neq 0$ (The slope of the SLR is not equal to 0, so the two variables are correlated)

```
qf(1-0.05, 1, 6-2)
```

```
## [1] 7.708647
```

The ANOVA F statistic is 21.333. Because it is larger than the critical value (found to be 7.708647), we reject the null hypothesis. Our data supports the claim that the slope is different from 0, or in other words, that there is a linear association between a students first quiz grade and their second quiz grade.

---

7. (You may only use R as a simple calculator or to find p-values or critical values) A substance used in biological and medical research is shipped by airfreight to users in cartons of 1000 ampules. The data consist of 10 shipments. The variables are number of times the carton was transferred from one aircraft to another during the shipment route (transfer), and the number of ampules found to be broken upon arrival (broken). We want to fit a simple linear regression. A simple linear regression model is fitted using R. The corresponding output from R is shown next, with some values missing.

```
Call:
lm(formula = broken ~ transfer)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.2000     0.6633  ______ ________ ***
transfer      4.0000     0.4690  ______ ________ ***
Residual standard error: 1.483 on 8 degrees of freedom
...
Analysis of Variance Table
Response: broken
          Df Sum Sq Mean Sq F value    Pr(>F)
transfer   1  160.0   160.0  ______ ________ ***
Residuals  8   17.6     2.2
```

The following values are also provided for you, and may be used for the rest of this question:
$\bar{x} = 1, \sum_{i=1}^{10} (x_i - \bar{x})^2 = 10$

a. **Carry out a hypothesis test to assess if there is a linear relationship between the variables of interest.**

$H_0 : \beta_1 = 0$ (The slope of the SLR is equal to 0, so the two variables are uncorrelated)
$H_A : \beta_1 \neq 0$ (The slope of the SLR is not equal to 0, so the two variables are correlated)

```
# f-test: msr/msres
#f = 160/2.2
#f
#qf(1-(0.05/2), 1, 8)

# t-test
b1 = 0
b1hat = 4
seB1 = 0.4690

t = (b1hat-b1)/(seB1)
t
```

```
## [1] 8.528785
```

```
qt(1-(0.05/2), 8)
```

```
## [1] 2.306004
```

```
2*pt(-t,8)
```

```
## [1] 2.746895e-05
```

T-statistic (t) = 8.528785
critical value = 2.306004
p-value = 0.00002746895

Since the T-statistic of 8.528785 is larger than the critical value, we reject H_0 in favor of H_A. We have enough evidence to conclude that there is a linear relationship between the variables of interest.

b. **Calculate a 95% confidence interval that estimates the unknown value of the population slope.**

```
lwr <- 4 - qt(1-(0.05/2), 8)*0.4690
upr <- 4 + qt(1-(0.05/2), 8)*0.4690
cat("[", lwr, ",", upr,"]")
```

```
## [ 2.918484 , 5.081516 ]
```

c. **A consultant believes the mean number of broken ampules when no transfers are made is different from 9. Conduct an appropriate hypothesis test (state the hypotheses statements, calculate the test statistic, and write the corresponding conclusion in context, in response to his belief).**

$H_0 : \mu = 9$ (The mean number of broken ampules when no transfers are made is equal to 9)
$H_A : \mu \neq 9$ (The mean number of broken ampules when no transfers are made is different than 9)

```
b0 = 9
b0hat = 10.2
seB0 = 0.6633

t = (b0hat-b0)/(seB0)
t
```

```
## [1] 1.809136
```

```
qt(1-(0.05/2), 8)
```

```
## [1] 2.306004
```

```
2*pt(t,8)
```

```
## [1] 1.891967
```

Since the t-statistic of 1.809136 is smaller than the critical value and the p-value is larger than $\alpha = 0.05$, we fail to reject H_0. We do not have enough evidence to conclude that the mean number of broken ampules when no transfers are made is different than 9.

d. **Calculate a 95% confidence interval for the mean number of broken ampules and a 95% prediction interval for the number of broken ampules when the number of transfers is 2.**

```
#Confidence interval
muhat = 10.2 + 4*(2)
clwr <- muhat - (qt(.975, 8)*(1.483)*sqrt((1/10)+(((2 - 1)^2) / 10)))
cupr <- muhat + (qt(.975, 8)*(1.483)*sqrt((1/10)+(((2 - 1)^2) / 10)))
cat("[", clwr, ",", cupr,"]")
```

```
## [ 16.67062 , 19.72938 ]
```

The 95% CI for the mean number of broken ampules for a carton of 1000 ampules shipped with the number of transfers being 2 is (16.67062 , 19.72938). We have 95% confidence the mean number of broken ampules for a carton of 1000 ampules shipped with the number of transfers being 2 is between 16.67062 and 19.72938.

```
#Prediction interval
plwr <- muhat - (qt(.975, 8)*(1.483)*sqrt(1+ (1/10)+(((2 - 1)^2) / 10)))
pupr <- muhat + (qt(.975, 8)*(1.483)*sqrt(1+ (1/10)+(((2 - 1)^2) / 10)))
cat("[", plwr, ",", pupr,"]")
```

```
## [ 14.45379 , 21.94621 ]
```

We have 95% confidence that for a carton of 1000 ampules shipped when the number of transfers is 2, the number of broken ampules is between 14.45379 and 21.94621.

8. Derive the least squares estimators of the simple linear regression model

i.e. show that

$$\hat{\beta_0} = \bar{y} - \hat{\beta_1}\bar{x}$$

and

$$\hat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Recall that we want to minimize the sum of squared errors, i.e., minimize

$$SS_{res} = \sum_{i=1}^{n}(y_i - \hat{y_i})^2,$$

Hint 1: Note that $\hat{y_i} = \beta_0 + \beta_1 x_i$

Hint 2: Take partial derivatives of SS_{res} with reference to $\hat{\beta_1}$, and $\hat{\beta_0}$.

Hint 3: the following formulae may be useful, and may be used without proof:

$$\sum(x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2$$

,

$$\sum(x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n\bar{x}\bar{y}$$

.

Hint 4: Work through showing equations $\hat{\beta_0}$ and $\hat{\beta_1}$ in order.

$$SS_{res} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \longrightarrow \text{Remember: } \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$= \sum_{i=1}^{n}\left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right]^2$$

(1.) Take partial derivative with respect to $\hat{\beta}_0$:

$$\frac{\partial}{\partial \hat{\beta}_0}\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 = \sum_{i=1}^{n}\frac{\partial}{\partial \hat{\beta}_0}\left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right]^2$$

$$= \sum_{i=1}^{n} 2(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))(-1)$$

$$= -2\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))$$

(2.) Take partial derivative with respect to $\hat{\beta}_1$:

$$\frac{\partial}{\partial \hat{\beta}_1}\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 = \sum_{i=1}^{n}\frac{\partial}{\partial \hat{\beta}_1}\left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right]^2$$

$$= \sum_{i=1}^{n} 2(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))(-x_i)$$

$$= -2\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))(x_i)$$

Set partial derivatives equal to zero to solve for $\hat{\beta}_0$ and $\hat{\beta}_1$

(3.) $-2\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)) = 0$

$$\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)) = 0$$

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\hat{\beta}_0 - \sum_{i=1}^{n}\hat{\beta}_1 x_i = 0$$

$$\sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1\sum_{i=1}^{n} x_i = 0$$

$$n\hat{\beta}_0 = \sum_{i=1}^{n} y_i - \hat{\beta}_1\sum_{i=1}^{n} x_i$$

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i}{n} - \frac{\hat{\beta}_1\sum_{i=1}^{n} x_i}{n}$$

$$\boxed{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}}$$

(4.) $-2\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))(x_i) = 0$

$$\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))(x_i) = 0$$

Substitute $\hat{\beta}_0$ →

$$\sum_{i=1}^{n}(y_i - (\bar{y} - \hat{\beta}_1\bar{x} + \hat{\beta}_1 x_i))(x_i) = 0$$

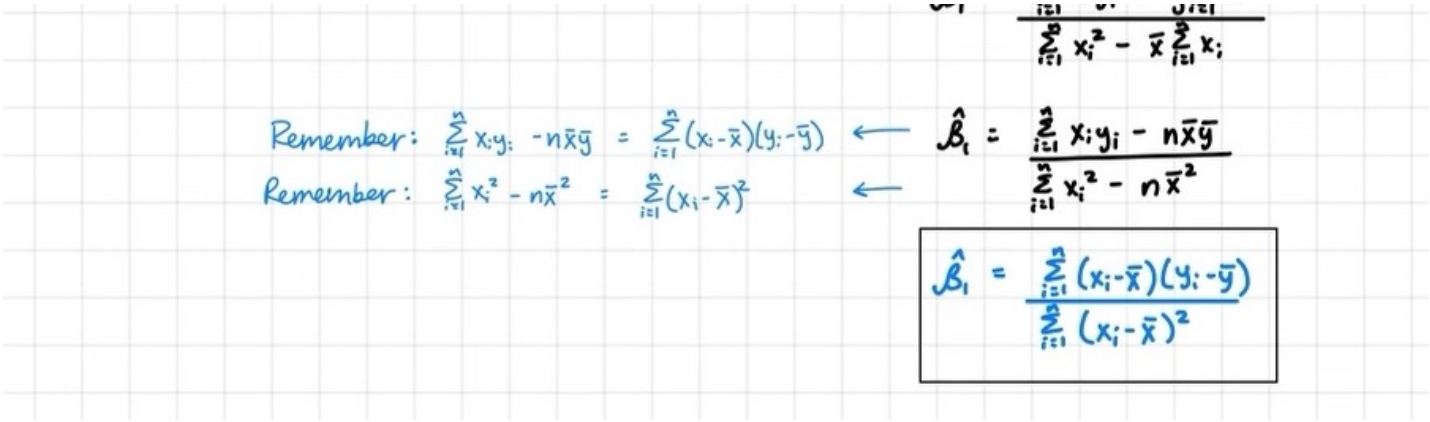
$$\sum_{i=1}^{n}(y_i - \bar{y} - \hat{\beta}_1(x_i - \bar{x}))x_i = 0$$

$$\sum_{i=1}^{n} x_i(y_i - \bar{y}) - \sum_{i=1}^{n} x_i\hat{\beta}_1(x_i - \bar{x}) = 0$$

$$\sum_{i=1}^{n} x_i(y_i - \bar{y}) = \hat{\beta}_1\sum_{i=1}^{n} x_i(x_i - \bar{x})$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i(y_i - \bar{y})}{\sum_{i=1}^{n} x_i(x_i - \bar{x})}$$

$$\hat{\beta}_1 = \frac{\sum x_i y_i - \bar{y}\sum x_i}{}$$



Problem 8 Written Solution