Untitled

HW1_6120

2023-06-20

```
students = read.table("C:\\Users\\jacqu\\Downloads\\students.txt", header = T)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

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

Question 1

(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.

I think that every column is meaningful! Maybe the student ID column, but that could also help with identifying influential observations and such, so I would personally keep it.

(b) How many students are there in this data set?

```
nrow(students)
```

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

There are 249 students recorded in this dataset.

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

```
students%>%
  filter(!complete.cases(students))
```

```
##   Student Gender Smoke Marijuan DrivDrnk  GPA PartyNum DaysBeer StudyHrs
## 1      16 female    No       No       No   NA        4        0       16
## 2      17 female    No       No       No   NA        6        0       14
## 3      23   male    No       No       No 3.60       NA        8       20
## 4      25 female    No      Yes      Yes   NA       10       10       15
## 5      38 female   Yes      Yes      Yes 3.11       NA       10       10
## 6      78   male    No      Yes      Yes 3.54       NA       10       10
## 7     100 female    No       No       No   NA        8        8        7
## 8      105 female    No      Yes      Yes 3.98       NA        9       28
## 9      113 female    No       No       No   NA        3        0       30
## 10     192   male    No       No      Yes   NA        5        0       10
## 11     222 female    No       No       No 2.70       NA        6       20
## 12     241 female    No      Yes      Yes   NA       12       15       18
```

There are 12 students that have a missing entry in at last one column.

(d) Report the median values of the numeric variables.

```
students%>%
  summarise(mGPA=median(GPA, na.rm=T),mPN=median(PartyNum, na.rm=T),mDB=median(DaysBeer, na.rm=T),mSH=me
```

```
##   mGPA mPN mDB mSH
## 1  3.2   8   8  14
```

The median GPA is 3.2, median PartyNum is 8, median DaysBeer is 8, and median StudyHrs 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)%>%
  summarise(meanSH=mean(StudyHrs,na.rm=T),sdSH=sd(StudyHrs,na.rm=T),
            medSH=median(StudyHrs,na.rm=T))
```

```
## # A tibble: 2 x 4
##   Gender meanSH  sdSH medSH
##   <chr>   <dbl> <dbl> <dbl>
## 1 female   15.4  8.97    14
## 2 male     14.7 10.2     12
```

Generally, females spend more time studying compared to 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” otherwise.

```
students = students%>%
  mutate(PartyAnimal=ifelse(PartyNum>8,"yes","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,3,3.5,Inf),
                      right=F,
                      labels=c("low","moderate","high")))
```

(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?

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

```
##    Student Gender Smoke Marijuan DrivDrnk  GPA PartyNum DaysBeer StudyHrs
## 1        5   male   Yes      Yes      Yes 2.30       10       15       14
## 2        9 female    No      Yes      Yes 1.87       16       20        6
## 3       18 female    No      Yes      Yes 2.70        9        8        9
## 4       61 female   Yes      Yes      Yes 2.33       10       20        5
## 5       66 female   Yes      Yes      Yes 2.87        9       15        6
## 6       70 female    No      Yes      Yes 2.70       14       12       14
## 7       80 female    No       No       No 2.65        9        9        6
## 8       97 female   Yes      Yes      Yes 2.80       10       20        6
## 9       99   male    No      Yes       No 2.86       12       20        8
## 10     106   male    No      Yes      Yes 2.75       20       20       10
## 11     116   male    No      Yes      Yes 2.21       15       20        5
## 12     119   male    No       No       No 2.58       15       10        9
## 13     130 female   Yes       No      Yes 2.90       10        0        8
## 14     141   male   Yes      Yes      Yes 2.65       12       12       10
## 15     148   male    No       No      Yes 2.67       13       13        5
## 16     150   male   Yes      Yes      Yes 2.89       25       25        8
## 17     160 female   Yes       No       No 2.83        9        9       10
## 18     170   male    No      Yes      Yes 2.88       12       14        3
## 19     171   male    No      Yes      Yes 2.70       15       13       10
## 20     177   male    No      Yes      Yes 2.60       11       13       14
## 21     183   male    No       No      Yes 2.65       12       25       10
## 22     185   male    No      Yes      Yes 2.90       15       18        4
## 23     200   male    No       No       No 2.84        9       10       10
## 24     202 female    No      Yes       No 2.50       15       31        5
## 25     216 female   Yes      Yes      Yes 2.90       25       20        7
## 26     217   male    No      Yes      Yes 2.60       16       20        8
## 27     221   male    No      Yes      Yes 2.80       15       15        6
## 28     239   male   Yes      Yes      Yes 2.07       14       16       10
## 29     244   male    No      Yes      Yes 2.80       15       20       10
##    PartyAnimal GPA.cat
## 1          yes     low
## 2          yes     low
## 3          yes     low
## 4          yes     low
## 5          yes     low
## 6          yes     low
## 7          yes     low
## 8          yes     low
## 9          yes     low
## 10         yes     low
## 11         yes     low
## 12         yes     low
## 13         yes     low
## 14         yes     low
## 15         yes     low
## 16         yes     low
## 17         yes     low
## 18         yes     low
## 19         yes     low
## 20         yes     low
## 21         yes     low
## 22         yes     low
## 23         yes     low
## 24         yes     low
## 25          yes     low
## 26          yes     low
## 27          yes     low
## 28          yes     low
## 29          yes     low
```

There are 29 students that meet all three requirements.

(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)
```

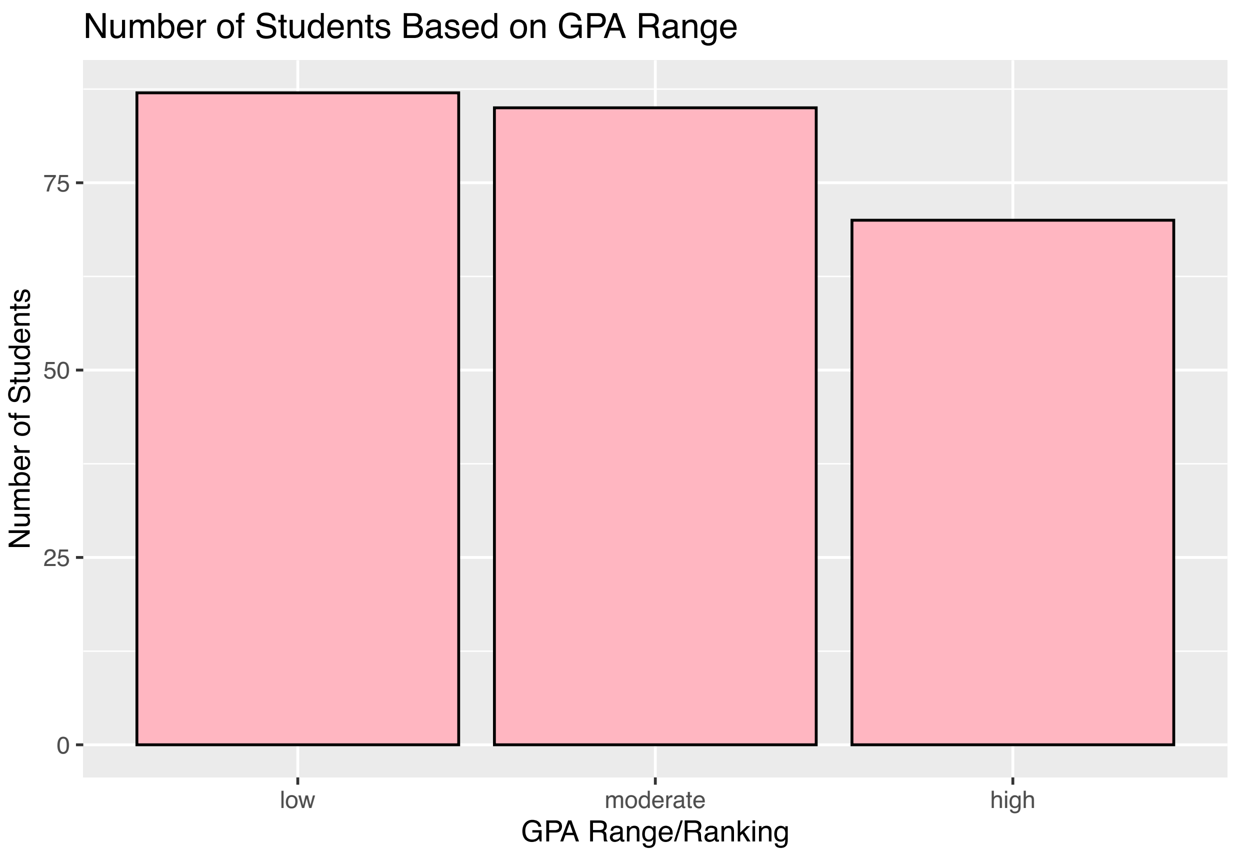
```
##
##      low moderate     high
##       87       85       70
```

87 students with low GPA, 85 with moderate GPA, and 70 with high GPA.

(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.

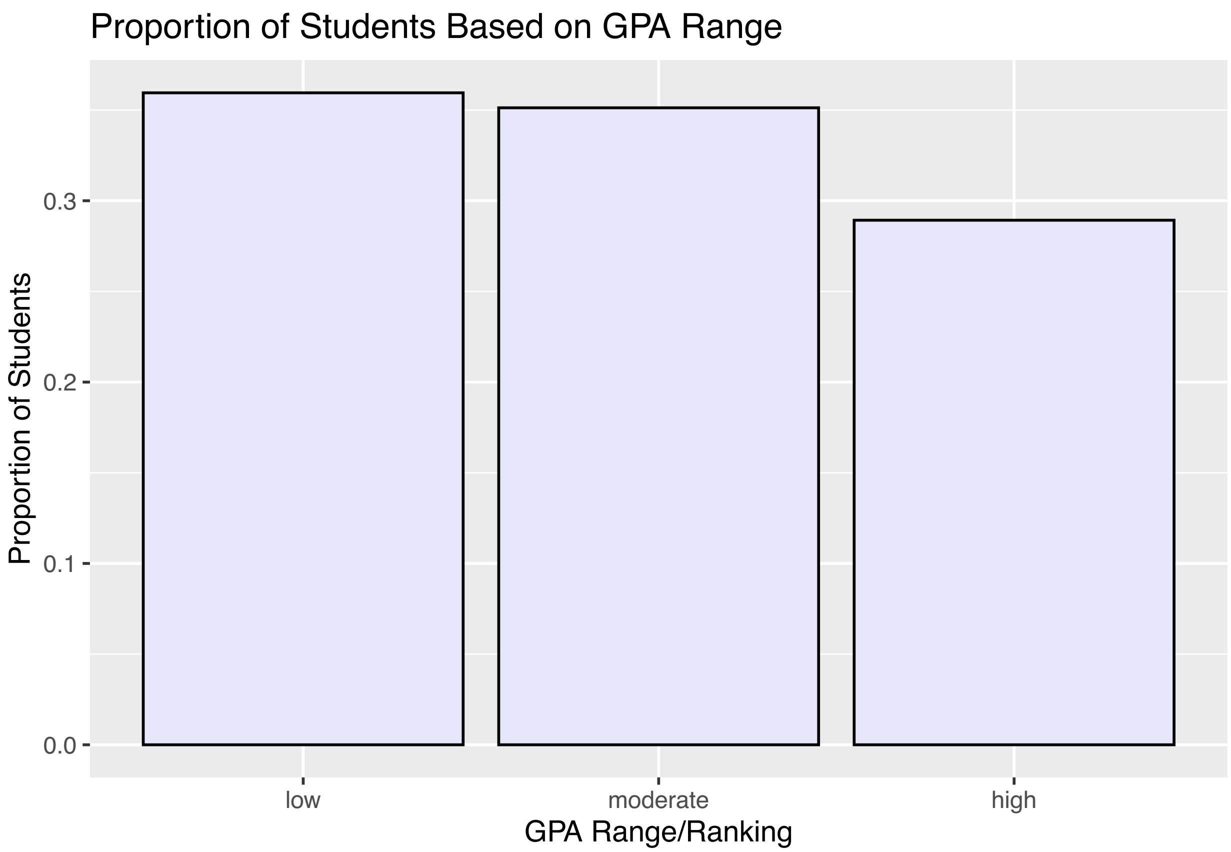
```
ggplot(students[complete.cases(students$GPA.cat),])+
  geom_bar(aes(x=GPA.cat,y=..count..),
           fill="lightpink", color="black")+
  ggtitle("Number of Students Based on GPA Range")+
  xlab("GPA Range/Ranking")+
  ylab("Number of Students")
```

```
## Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(count)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



(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.

```
ggplot(students[complete.cases(students$GPA.cat),])+
  geom_bar(aes(x=GPA.cat,y=..count../nrow(students[complete.cases(students$GPA.cat),])),
           fill="lavender", color="black")+
  ggtitle("Proportion of Students Based on GPA Range")+
  xlab("GPA Range/Ranking")+
  ylab("Proportion of Students")
```



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

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

```
##
##          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.

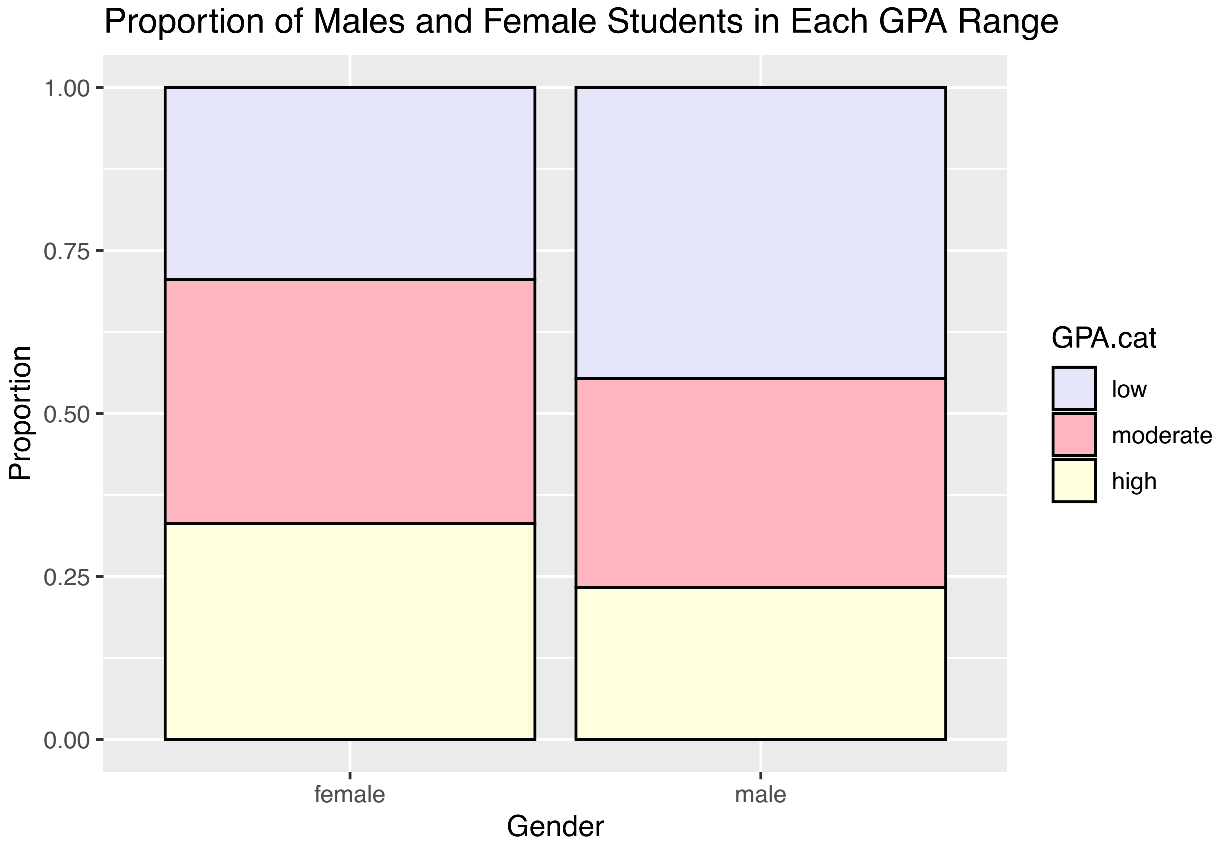
```
round(prop.table(table(students$Gender, students$GPA.cat),1)*100,2)
```

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

The percentage of females with moderate-high GPAs (~70%) is higher than the males (~55%) within their own population. So, female students tend to have higher GPAs than males.

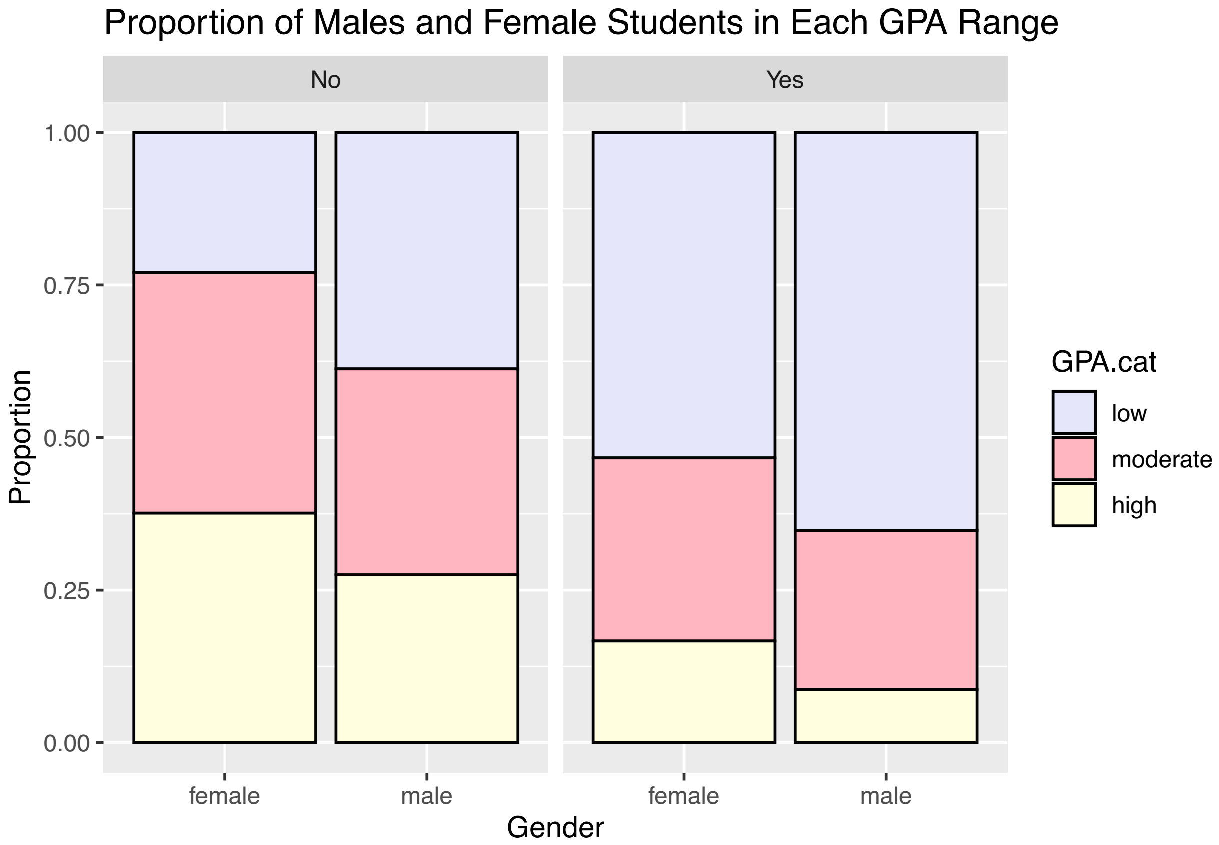
(n) Create a bar chart to explore the proportion of GPA categories for female and male students. Be sure to remove the bar corresponding to the missing values.

```
ggplot(students[complete.cases(students$GPA.cat),])+
  geom_bar(aes(x=Gender,y=..count../nrow(students[complete.cases(students$GPA.cat),]),
               fill=GPA.cat),colour="black", position="fill")+
  scale_fill_manual(values=c("lavender","lightpink","lightyellow"))+
  ggtitle("Proportion of Males and Female Students in Each GPA Range")+
  xlab("Gender")+
  ylab("Proportion")
```



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

```
ggplot(students[complete.cases(students$GPA.cat),])+
  geom_bar(aes(x=Gender,y=..count../nrow(students[complete.cases(students$GPA.cat),]),
               fill=GPA.cat),colour="black", position="fill")+
  scale_fill_manual(values=c("lavender","lightpink","lightyellow"))+
  facet_wrap(~Smoke)+
  ggtitle("Proportion of Males and Female Students in Each GPA Range")+
  xlab("Gender")+
  ylab("Proportion")
```

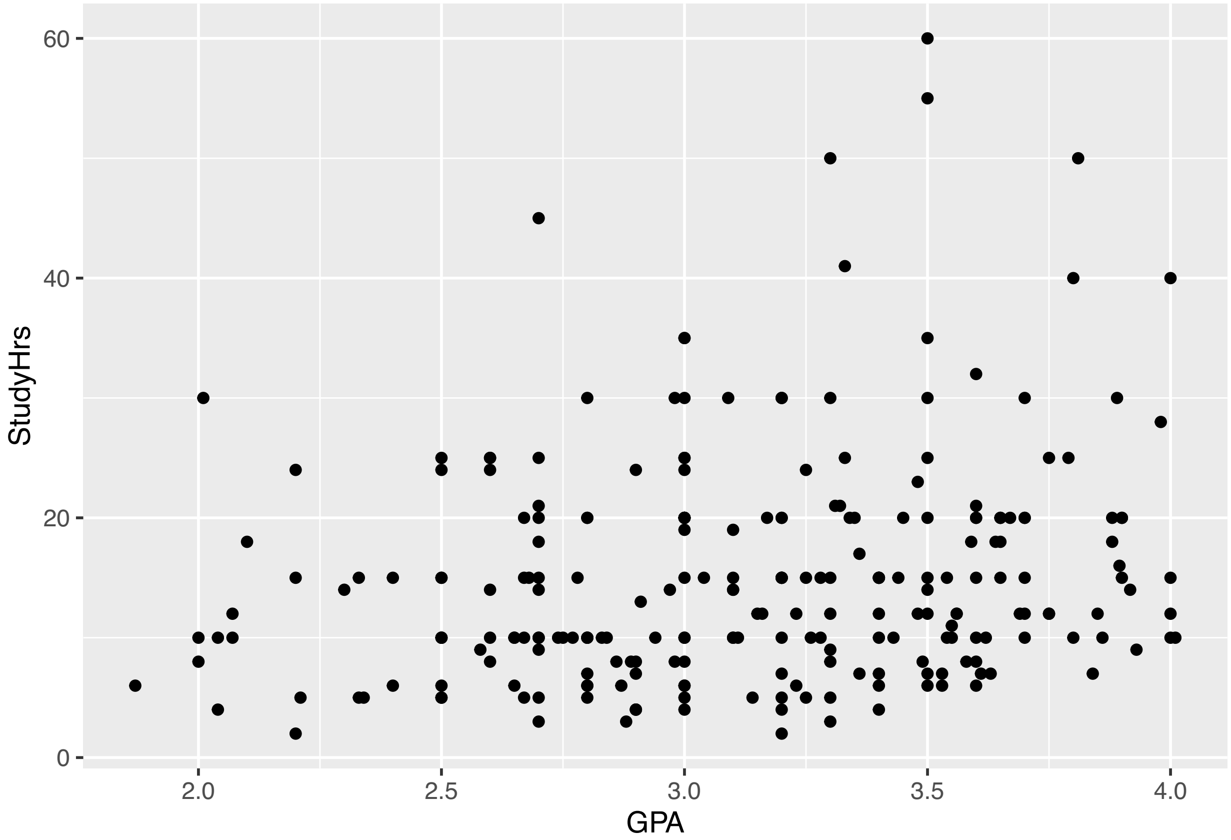


Regardless of smoking status, female students tend to have higher GPAs than males. In general though, students who smoke tend to have lower GPAs.

(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?

```
ggplot(students)+
  geom_point(aes(x=GPA,y=StudyHrs))
```

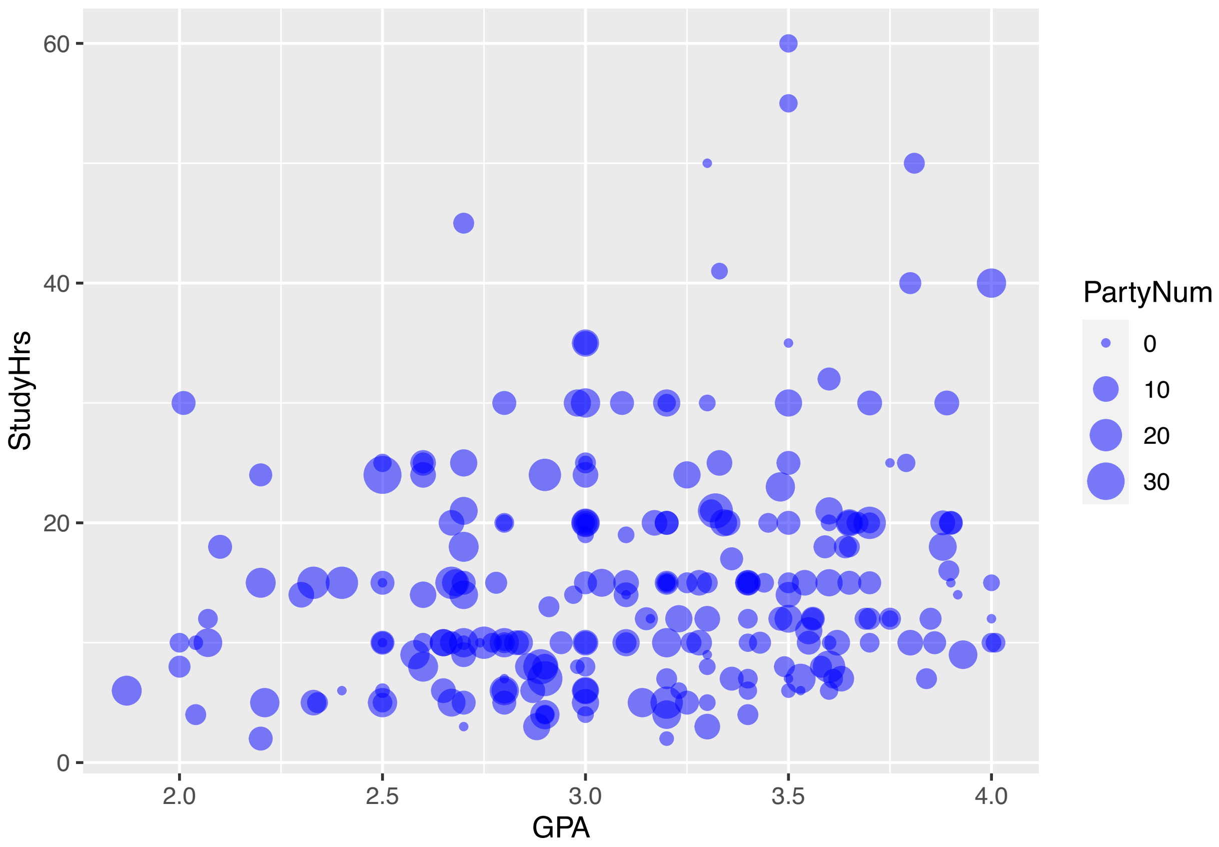
```
## Warning: Removed 7 rows containing missing values (`geom_point()`).
```



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

```
ggplot(students)+
  geom_point(aes(x=GPA,y=StudyHrs,size=PartyNum),
             alpha=0.5,color="blue")
```

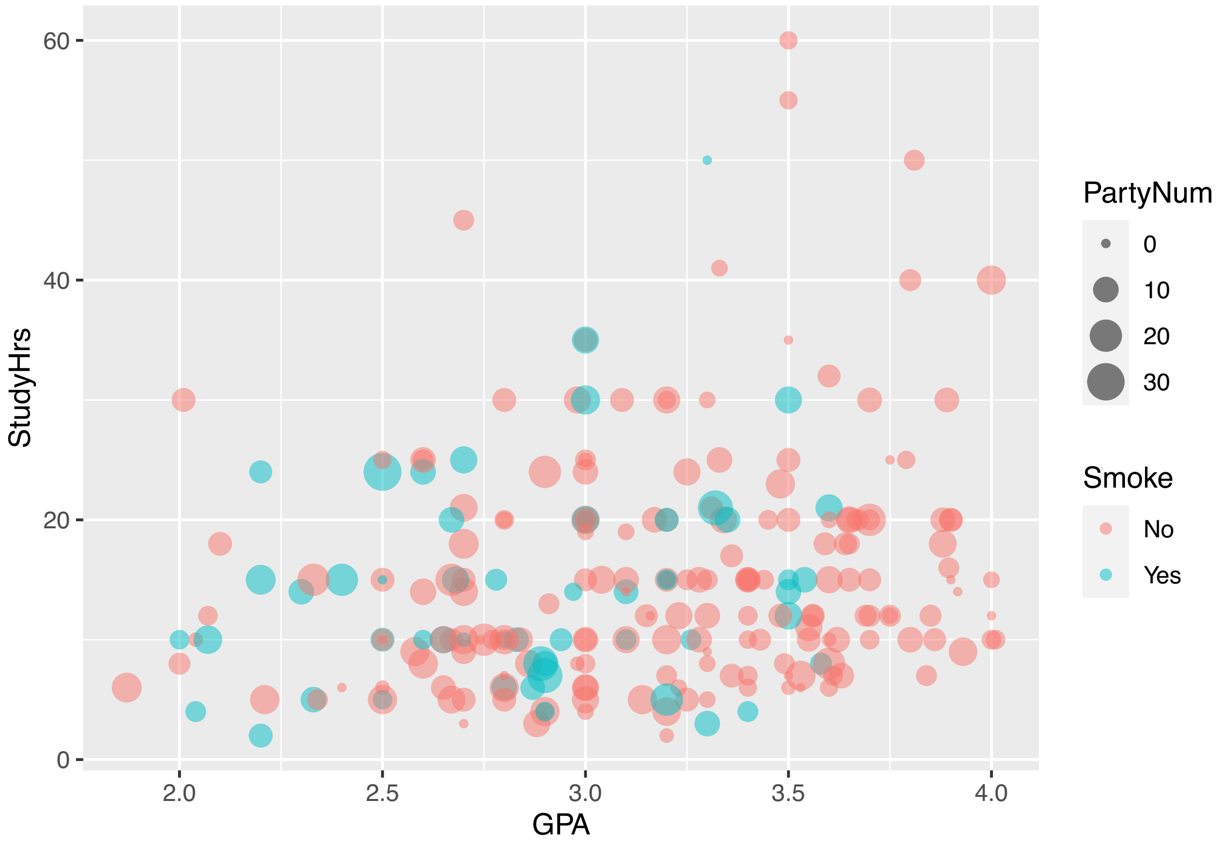
```
## Warning: Removed 12 rows containing missing values (`geom_point()`).
```



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

```
ggplot(students)+
  geom_point(aes(x=GPA,y=StudyHrs,size=PartyNum,color=Smoke),
             alpha=0.5)
```

```
## Warning: Removed 12 rows containing missing values (`geom_point()`).
```



Question 2

```
usc = read.csv("C:\\Users\\jacqu\\Downloads\\UScovid.csv", header=T)
```

(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

```
latest = usc%>%
  filter(date=="2021-06-03" & county!="Unknown")%>%
  select(c(2,3,5,6))%>%
  arrange(county,state)
```

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

```
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*100/cases,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.

```
latest%>%
  arrange(desc(cases))%>%
  head(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.

```
latest%>%
  arrange(desc(deaths))%>%
  head(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?

```
latest%>%
  arrange(desc(death.rate))%>%
  head(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
```

I noticed that a lot of these counties reside in Southern states with Texas and Georgia standing out the most. However, they also have relativily low number of cases.

(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.

```
latest%>%
  filter(cases>=100000)%>%
  arrange(desc(death.rate))%>%
  head(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

```
latest%>%
  filter(state=="Virginia")%>%
  filter(county=="Albemarle" | county=="Charlottesville city")
```

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

Question 3

(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=latest%>%
  select(-1)%>%
  group_by(state)%>%
  summarise(cases=sum(cases),deaths=sum(deaths))%>%
  arrange(state)
head(state.level)
```

```
## # A tibble: 6 x 3
##   state        cases deaths
##   <chr>        <int>  <int>
## 1 Alabama     545028  11188
## 2 Alaska       69534    352
## 3 Arizona     882691  17653
## 4 Arkansas    338986   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*100/cases,2))
head(state.level)
```

```
## # A tibble: 6 x 4
##   state        cases deaths state.rate
##   <chr>        <int>  <int>      <dbl>
## 1 Alabama     545028  11188       2.05
## 2 Alaska       69534    352       0.51
## 3 Arizona     882691  17653       2   
## 4 Arkansas    338986   5842       1.72
## 5 California 3793055  63345       1.67
## 6 Colorado    547961   6746       1.23
```

(c) What is the case fatality rate in Virginia?

```
state.level[state.level$state=="Virginia",]
```

```
## # A tibble: 1 x 4
##   state     cases deaths state.rate
##   <chr>     <int>  <int>      <dbl>
## 1 Virginia 676041  11216       1.66
```

1.66% fatality rate in Virginia.

(d) What is the case fatality rate in Puerto Rico?

```
state.level[state.level$state=="Puerto Rico",]
```

```
## # A tibble: 1 x 4
##   state        cases deaths state.rate
##   <chr>        <int>  <int>      <dbl>
## 1 Puerto Rico 166825     NA         NA
```

There is no data for the number of deaths in Puerto Rico, therefore, the fatality rate is currently unknown.

(e) Which states have the 10 highest case fatality rates?

```
state.level%>%
  arrange(desc(state.rate))%>%
  head(10)
```

```
## # A tibble: 10 x 4
##    state                  cases deaths state.rate
##    <chr>                  <int>  <int>      <dbl>
##  1 Massachusetts         660563  17881       2.71
##  2 New Jersey           1016219  26253       2.58
##  3 New York             2102003  52811       2.51
##  4 Connecticut           346564   8244       2.38
##  5 District of Columbia   49041   1136       2.32
##  6 Mississippi           318048   7324       2.3
##  7 Pennsylvania         1208879  27349       2.26
##  8 Louisiana             472222  10605       2.25
##  9 New Mexico            203330   4275       2.1
## 10 Maryland              460406   9587       2.08
```

(f) Which states have the 10 lowest case fatality rates?

```
state.level%>%
  arrange(state.rate)%>%
  head(10)
```

```
## # A tibble: 10 x 4
##    state                     cases deaths state.rate
##    <chr>                     <int>  <int>      <dbl>
##  1 Alaska                    69534    352       0.51
##  2 Utah                     405721   2286       0.56
##  3 Virgin Islands             3512     28       0.8
##  4 Vermont                   24218    255       1.05
##  5 Nebraska                 222317   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.

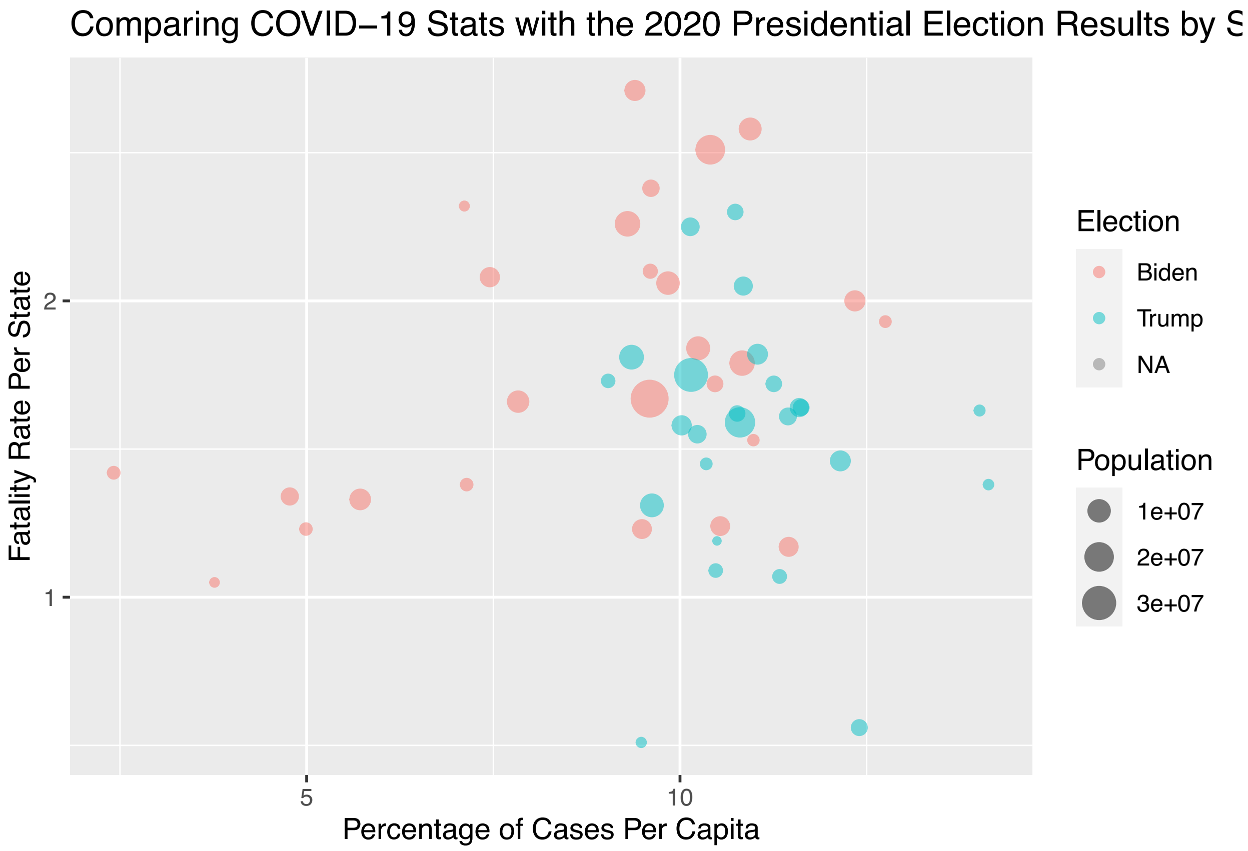
```
pope = read.csv("C:\\Users\\jacqu\\Downloads\\State_pop_election.csv", header=T)
pope_sl = state.level%>%
  mutate(State=state)%>%
  select(2,3,4,5)%>%
  left_join(pope, by="State")%>%
  arrange(State)
head(pope_sl)
```

```
## # A tibble: 6 x 6
##     cases deaths state.rate State      Population Election
##     <int>  <int>      <dbl> <chr>           <int> <chr>
## 1  545028  11188       2.05 Alabama       5024279 Trump
## 2   69534    352       0.51 Alaska         733391 Trump
## 3  882691  17653       2    Arizona       7151502 Biden
## 4  338986   5842       1.72 Arkansas      3011524 Trump
## 5 3793055  63345       1.67 California   39538223 Biden
## 6  547961   6746       1.23 Colorado      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 decribe how you created the new variables.

```
ggplot(pope_sl)+
  geom_point(aes(x=cases*100/Population,y=state.rate,color=Election,size=Population),alpha=0.5)+
  ggtitle("Comparing COVID-19 Stats with the 2020 Presidential Election Results by State")+
  xlab("Percentage of Cases Per Capita")+
  ylab("Fatality Rate Per State")
```

```
## Warning: Removed 3 rows containing missing values (`geom_point()`).
```



I didn't add any variables, but I did calculate the percentage of cases per capita by multiplying cases by 100 and the dividing by population. This is because in my graph, I wanted to somewhat normalise the number of cases for each state (600 cases in a population of 1000 is very different than 600 cases in a population of 1000000). Comparing that percentage to the fatality rate reveals the proportion of deaths to cases in terms of population. I also added the bubble sizes to show population size because even though it's calculated in (with the per capita), visually, it's hard to see it. Adding the size helps with showing the data in terms of population. Lastly, I added the colors to reflect the election results to see if there was any trends with COVID and election results.

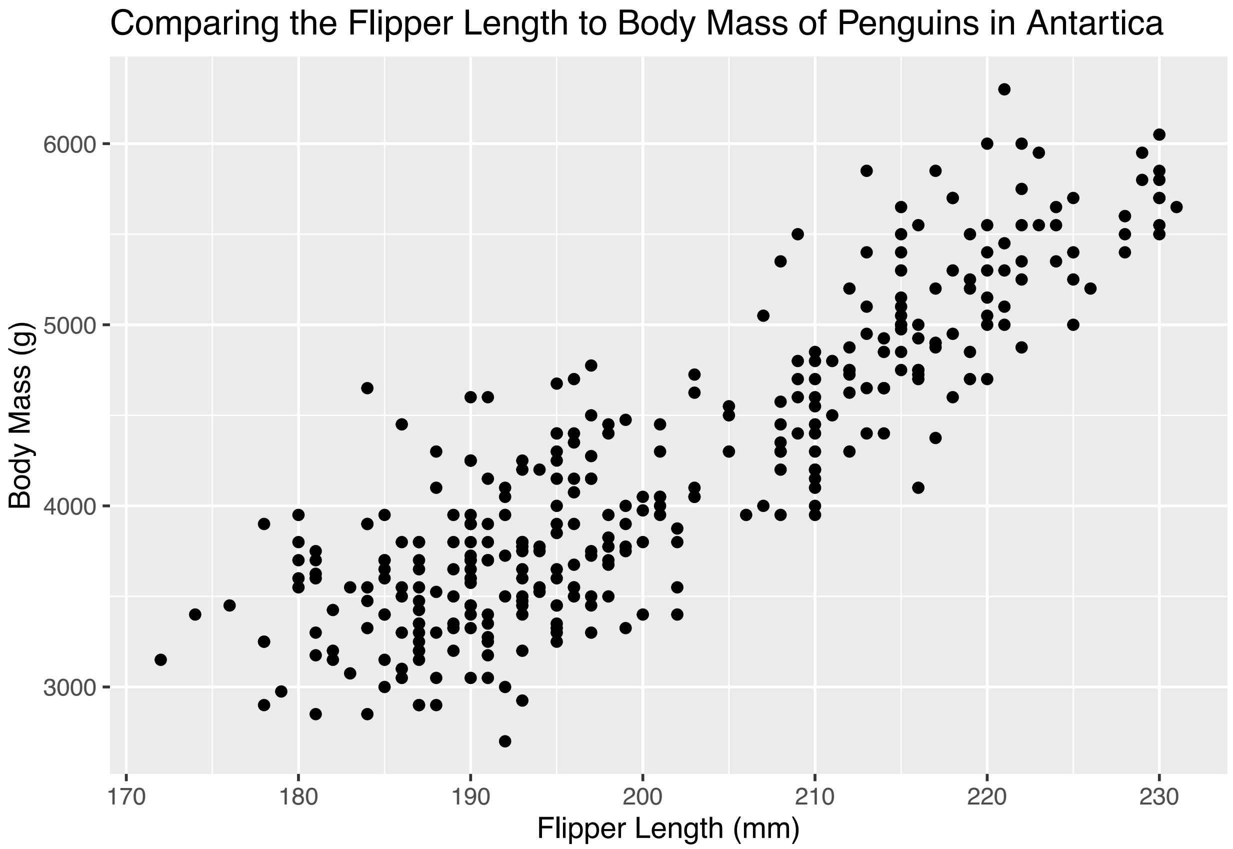
Question 4

```
library(palmerpenguins)
pen = palmerpenguins::penguins
```

(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(pen)+
  geom_point(aes(x=flipper_length_mm,y=body_mass_g))+
  ggtitle("Comparing the Flipper Length to Body Mass of Penguins in Antartica")+
  xlab("Flipper Length (mm)")+
  ylab("Body Mass (g)")
```

```
## Warning: Removed 2 rows containing missing values (`geom_point()`).
```

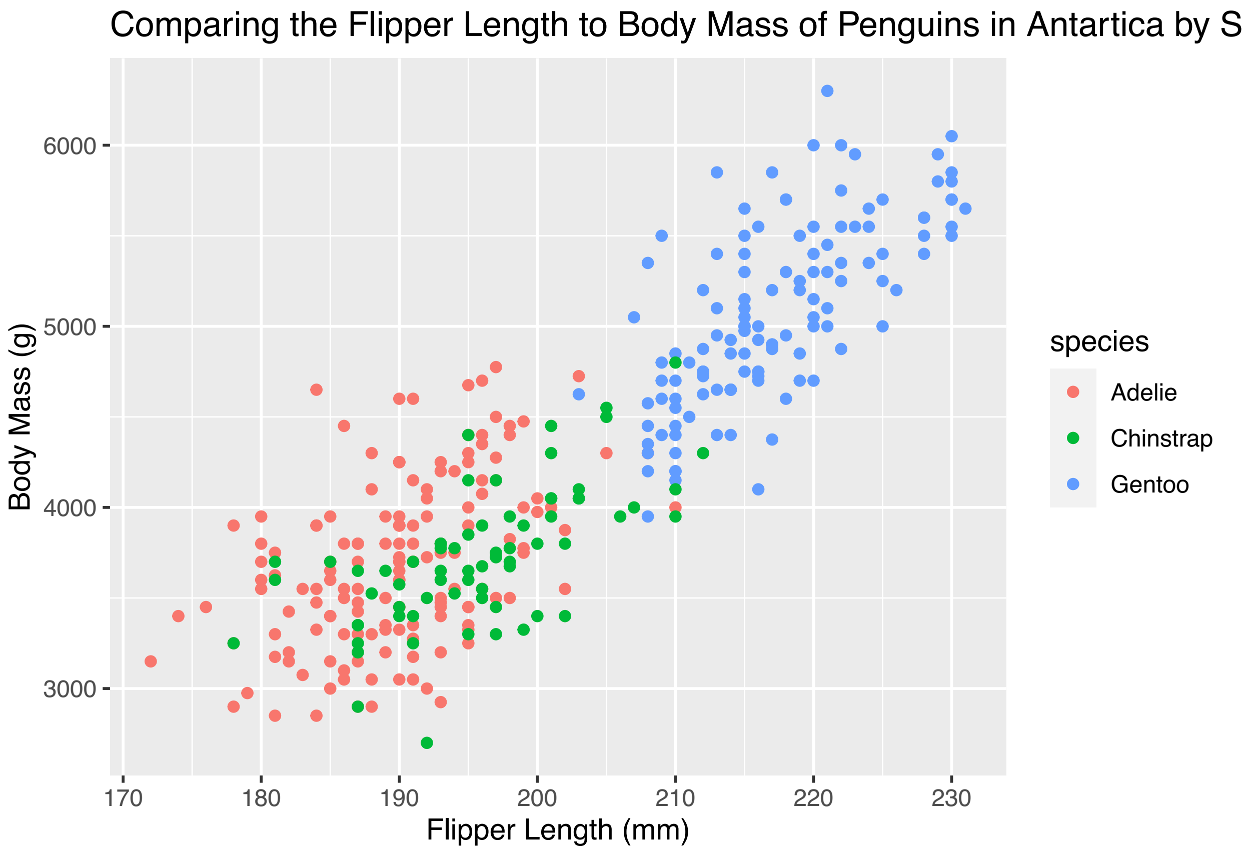


Yes, it seems reasonable! But there are different species, so we should check if a linear regression is appropriate based on species.

(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(pen)+
  geom_point(aes(x=flipper_length_mm,y=body_mass_g,color=species))+
  ggtitle("Comparing the Flipper Length to Body Mass of Penguins in Antartica by Species")+
  xlab("Flipper Length (mm)")+
  ylab("Body Mass (g)")
```

```
## Warning: Removed 2 rows containing missing values (`geom_point()`).
```

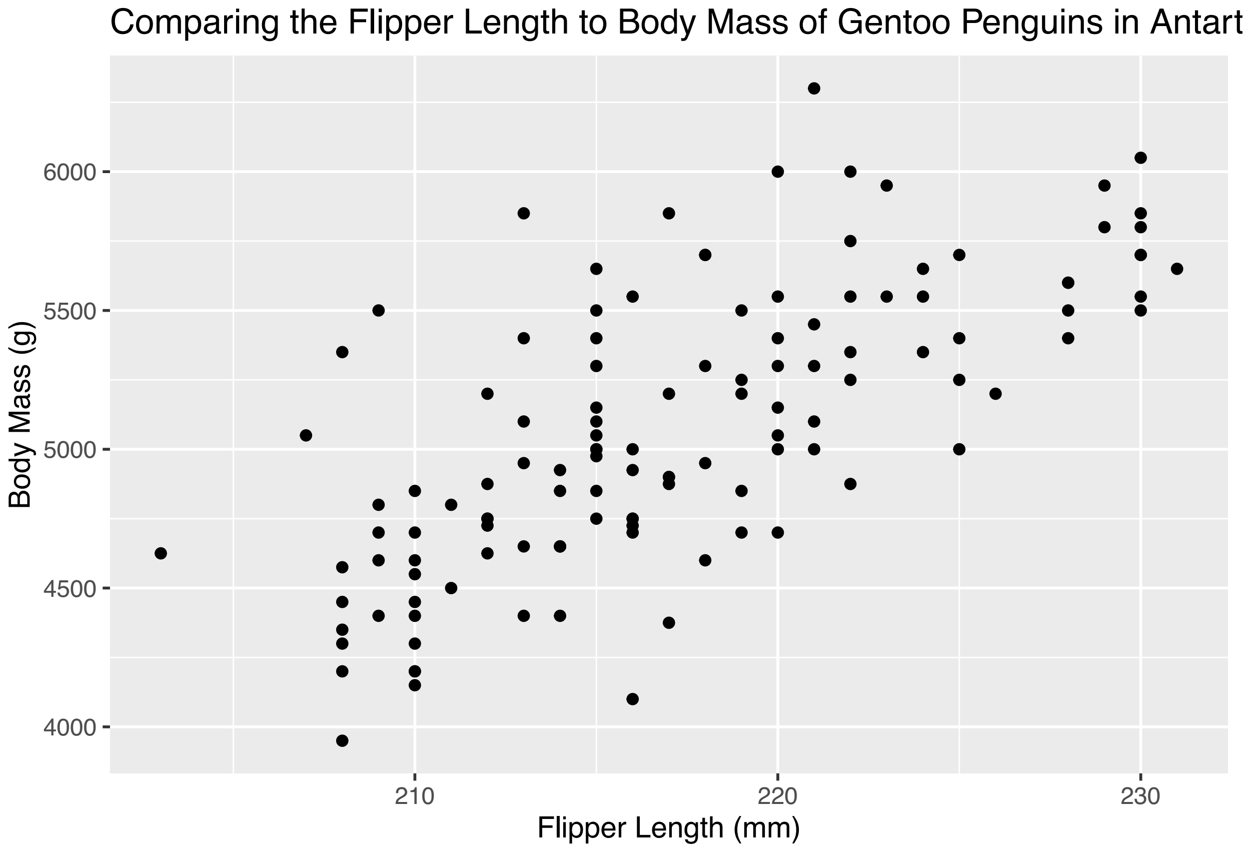


It doesn't, I still think a linear regression model is appropriate, provided that the model accounts for the categories in the predictor variable.

(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?

```
ggplot(pen[pen$species=="Gentoo",])+
  geom_point(aes(x=flipper_length_mm,y=body_mass_g))+
  ggtitle("Comparing the Flipper Length to Body Mass of Gentoo Penguins in Antartica")+
  xlab("Flipper Length (mm)")+
  ylab("Body Mass (g)")
```

```
## Warning: Removed 1 rows containing missing values (`geom_point()`).
```



Yes, it is appropriate to produce a SLM for this species.

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

```
genpen = pen[pen$species=="Gentoo",]
cor(genpen$flipper_length_mm, genpen$body_mass_g, use="complete.obs")
```

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

The correlation is 0.7026665. So there is a somewhat strong, positive correlation between the flipper length a Gentoo penguin has and the mass the penguin has. This makes sense because we're adding body mass to a penguin if we're adding more length to its' flipper.

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.

```
regpen = lm(body_mass_g~flipper_length_mm, data=genpen)
summary(regpen)
```

```
##
## Call:
## lm(formula = body_mass_g ~ flipper_length_mm, data = genpen)
##
## 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
```

The equation is (body mass) = -6787.28+54.62(flipper length) concerning Gentoo penguins.

(f) Interpret the estimated slope contextually.

For each mm added to the flipper length of a Gentoo penguin, the estimated body mass would increase by 54.62g.

(g) Does the estimated intercept make sense contextually?

No, it doesn't make sense because there is no penguin with a negative body mass. A penguin could be an amputee with 0 mm of flipper length, but at 0mm flipper length, the penguin would either have some body mass (assuming amputee) or wouldn't exist in general (0g body mass).

(h) Report the value of R2 from this linear regression, and interpret its value contextually.

The R2 value is 0.4937. Meaning that around 49% of the variability in body mass can be explained by the flipper length of a Gentoo penguin.

(i) What is the estimated value for the standard deviation of the error terms for this regression model, sigma-hat?

s = 360.2

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

```
-6787.28+(54.62*220)
```

```
## [1] 5229.12
```

The predicted body mass of a Gentoo penguin with 220mm flipper length is 5229.12g.

(k) Produce the ANOVA table for this linear regression. Using only this table, calculate the value of R2.

```
anova(regpen)
```

```
## 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
```

```
anova(regpen)$"Sum Sq"[1]/sum(anova(regpen)$"Sum Sq")
```

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

(l) What are the null and alternative hypotheses for the ANOVA F test?

H0: beta-hat1=0, (if the slope is 0) the model is inadequate at predicting body mass

HA: beta-hat1=/=0, (the slope is not 0) the model is adequate at predicting body mass

(m) Explain how the F statistic of 118.01 is found.

```
msr = anova(regpen)$"Mean Sq"[1]
msres = anova(regpen)$"Mean Sq"[2]
msr/msres
```

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

You can find the F-stat by taking MSR (the regression mean square) and dividing it by MSres (aka s^2, estimate of the varience of the error terms) ==> MSR/MSres

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

There is enough evidence to support the regression model using flipper length as a predictor for body mass aka the model is adequate.

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

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

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

We are 95% confident that as the flipper length increases by 1mm, the body mass increases between 44.67g and 64.58g.

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

Yes, they are consistent since the slope (beta-hat1) is within the confidence interval for the change in predicted body mass.

(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.

```
newdat = data.frame(species="Gentoo", flipper_length_mm=200)
predict(regpen, newdat, level=0.95, interval="confidence")
```

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

We are 95% confident that the mean body mass for Gentoo penguins with a flipper length of 200mm is between 3954.45g and 4319.99g.

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

```
predict(regpen, newdat, level=0.95, interval="prediction")
```

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

We are 95% confident that the body mass of a single Gentoo penguin is between 3401.12g and 4873.32g if they have a flipper length of 200mm.

(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?

H0:beta-hat1=50, the predicted body mass increases by 50g for every 1mm

HA:beta-hat1>50, the predicted body mass increases by more than 50g for every 1mm

t-stat: (beta-hat1-50)/se(beta-hat1) = (54.623-50)/5.028 = 0.9194511

t*: qt(0.95, 121) = 1.657544

p-val: 1-pt(0.9194511, 121) = 0.1798445

Conclusion: We do not have enough evidence to suggest that the predicted body mass does not increase by 50g for every 1mm.

Question 5

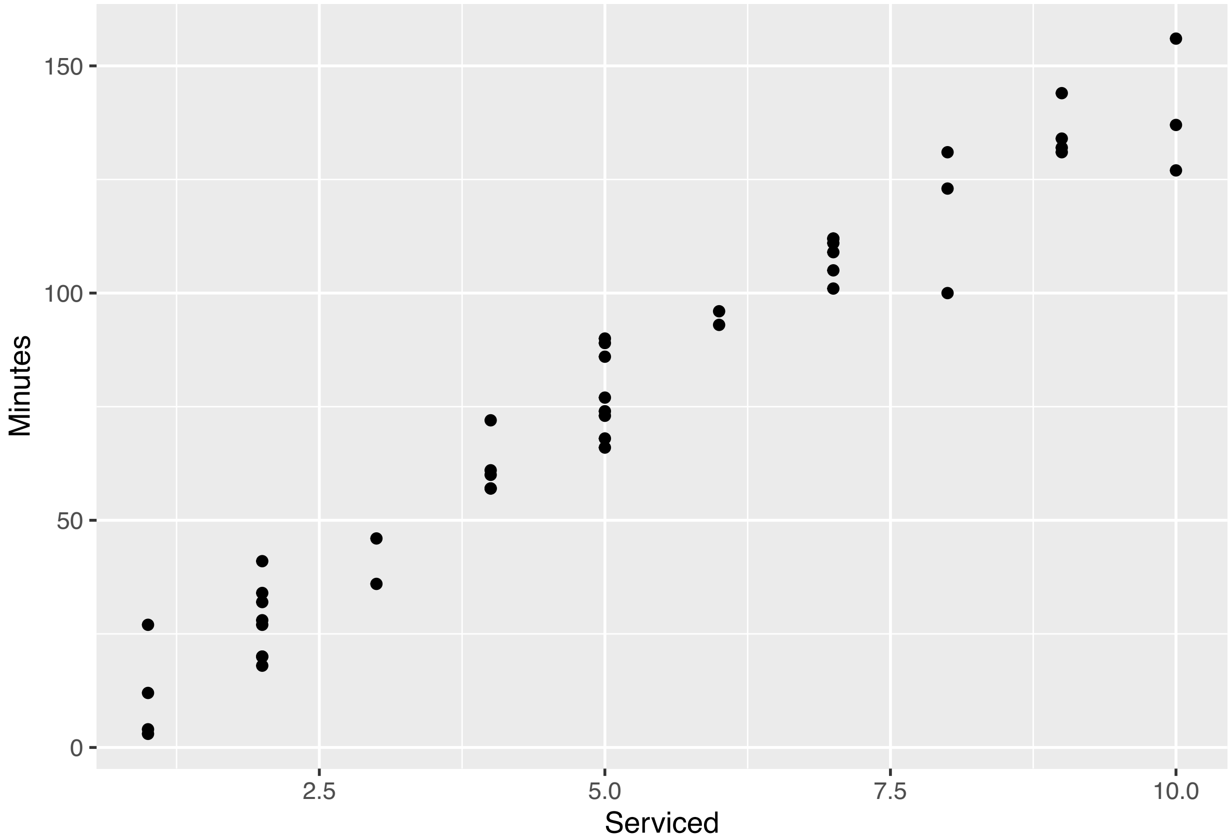
```
copier = read.table("C:\\Users\\jacqu\\Downloads\\copier.txt", header=T)
```

(a) What is the response variable in this analysis? What is predictor in this analysis?

The response variable is the total number of minutes spent by the service person. The predictor variable is the number of copiers serviced.

(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)+
  geom_point(aes(x=Serviced,y=Minutes))
```



I would describe this as a positive, linear relationship.

(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
```

There is a very strong, positive linear correlation between the number of copiers serviced and the total number of minutes spent by the service person.

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

I would say that the correlation can be interpreted reliably because the relationship between the two variables makes sense in this context.

(e) Use the lm() function to fit a linear regression for the two variables. Where are the values of beta-hat1, beta-hat0, R2, and sigma-hatˆ2 for this linear regression?

```
regcop = lm(Minutes~Serviced, data=copier)
summary(regcop)
```

```
##
## Call:
## lm(formula = Minutes ~ Serviced, data = copier)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.7723  -3.7371   0.3334   6.3334  15.4039
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.5802     2.8039  -0.207    0.837
## Serviced     15.0352     0.4831  31.123   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.914 on 43 degrees of freedom
## Multiple R-squared:  0.9575, Adjusted R-squared:  0.9565
## F-statistic: 968.7 on 1 and 43 DF,  p-value: < 2.2e-16
```

beta-hat1 = 15.0352

beta-hat0 = -0.5802

R2 = 0.9575

sigma-hatˆ2 = 8.914ˆ2 = 79.4594

(f) Interpret the values of beta-hat1, beta-hat0 contextually. Does the value of beta-hat0 make sense in this context?

beta-hat1: For every copier serviced, the estimated total number of minutes spent by the service person increases by about 15 minutes.

beta-hat0: If there are 0 copiers being serviced, the estimated total number of minutes spent by the service person is around -0.58 minutes.

The value of beta-hat0 does not make sense since a person physically can not spend negative amount of minutes on something. But at the same time, -0.5 is close to 0 minutes, so if we wanted to round to 0 minutes, it would then make sense.

(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(regcop)
```

```
## 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
```

H0: beta-hat1=0, the model is not adequate at predicting total number of minutes

HA: beta-hat1=/=0, the model is adequate at predicting total number of minutes

F-stat: 968.66

p-val: < 2.2e-16

Conclusion: There is enough evidence to support the model aka the model is adequate at predicting the total number of minutes using 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

```
newdat2 = data.frame(Serviced=5)
predict(regcop, newdat2, interval="prediction")
```

```
##        fit      lwr      upr
## 1 74.59608 56.42133 92.77084
```

We are 95% confident that the total service time spent by a person who serviced 5 copiers is between 56.42133 and 92.77084 minutes.

Question 6

```
q6df = data.frame(x=c(70,75,80,80,85,90), y=c(75,82,80,86,90,91))
```

yhat=20+0.8x

(a) For each individual observation, calculate its predicted score on the second quiz y-hati and the residual ei. You may show your results in the table below.

```
q6a = function(df){
  y.hat = list()
  ei = list()
  for (i in 1:nrow(df)){
    yh = 20+(0.8*df[i,1])
    e_i = df[i,2]-yh
    y.hat = append(y.hat,yh)
    ei = append(ei,e_i)
  }
  df$'y.hat' = unlist(y.hat)
  df$'ei' = unlist(ei)
  print(df)
}
```

```
df6 = q6a(q6df)
```

```
##    x  y y.hat ei
## 1 70 75    76 -1
## 2 75 82    80  2
## 3 80 80    84 -4
## 4 80 86    84  2
## 5 85 90    88  2
## 6 90 91    92 -1
```

(b) Complete the ANOVA table for this dataset below. Note: Cells with *** in them are typically left blank.

```
p = 2
n = 6
SSR = sum((df6$"y.hat"-mean(df6$"y"))^2)
SSres = sum((df6$"ei")^2)
MSR = SSR/(p-1)
MSres = sum((df6$"ei")^2)/(n-2)
SST = sum((df6$"y"-mean(df6$"y"))^2)
df_row = c(p-1,n-p,n-1)
SS = c(SSR, SSres, SST)
MS = c(MSR, MSres, "***")
Fstat = c(MSR/MSres, "***", "***")
atable = data.frame(df_row, SS, MS, Fstat)
atable
```

```
##   df_row  SS  MS            Fstat
## 1      1 160 160 21.3333333333333
## 2      4  30 7.5              ***
## 3      5 190 ***              ***
```

(c) Calculate the sample estimate of the variance sigma^2 for the regression model.

```
SSres/(n-2)
```

```
## [1] 7.5
```

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

```
SSR/SST
```

```
## [1] 0.8421053
```

84.21% of the variability in the quiz 2 score can be explained using the scores for quiz 1.

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

H0: beta-hat1=0, the model is not adequate at predicting the quiz 2 scores

HA: beta-hat1=/=0, the model is adequate at predicting the quiz 2 scores

F-stat: 21.3333

Conclusion: There is enough evidence to support the model aka the model is adequate at predicting the quiz 2 scores using the quiz 1 scores.

Question 7

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).

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

```r
(4-0)/0.4690
```

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

```r
2*(1-pt(8.528785,8))
```

```
## [1] 2.746894e-05
```

H0: beta-hat1=0, there is no linear relationship between transfer number and number of broken ampules

HA: beta-hat1=/=0, there is a linear relationship between transfer number and number of broken ampules

tstat = 8.528785

pval = 2.746894e-05

Conclusion: There is enough evidence to suggest that there is a linear relationship between the number of transfers and the number of amuples broken.

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

```r
4-(qt(0.975,8)*0.4690)
```

```
## [1] 2.918484
```

```r
4+(qt(0.975,8)*0.4690)
```

```
## [1] 5.081516
```

We are 95% confident that the change in number of ampules broken as the number of transfers increases by 1 is between 2.918484 and 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).

```r
(10.2-9)/0.6633
```

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

```r
2*(1-pt(1.809136,8))
```

```
## [1] 0.1080333
```

H0: beta-hat0=9, the number of broken ampules is 9 when no transfers are made

HA: beta-hat0=/=9, the number of broken ampules is not 9 when no transfers are made

tstat = 1.809136

pval = 0.1080333

Conclusion: There is enough evidence to suggest that the number of broken ampules when no transfers are made is not 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.

```r
# y=10.2+4x
10.2+(4*2)
```

```
## [1] 18.2
```

```
18.2-(qt(0.975,8)*1.483*sqrt((1/10)+(1/10)))
```

```
## [1] 16.67062
```

```
18.2+(qt(0.975,8)*1.483*sqrt((1/10)+(1/10)))
```

```
## [1] 19.72938
```

```
18.2-(qt(0.975,8)*1.483*sqrt(1+(1/10)+(1/10)))
```

```
## [1] 14.45379
```

```
18.2+(qt(0.975,8)*1.483*sqrt(1+(1/10)+(1/10)))
```

```
## [1] 21.94621
```

95% confidence interval for the mean number of broken ampules when the number of transfers is 2: (16.67062, 19.72938)

95% prediction interval for the number of broken ampules when the number of transfers is 2: (14.45379, 21.94621)

Q8 Prove $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ & $\hat{\beta}_1 = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\Sigma(x_i-\bar{x})^2}$

$SSE = \Sigma(y_i - \hat{y}_i)^2$ * $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

$= \Sigma\left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right]^2 \Rightarrow \Sigma(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$

$\frac{\partial}{\partial \hat{\beta}_0} \Sigma(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \Rightarrow \Sigma -2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$

$-\cancel{2}\,\Sigma(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = \cancel{0}$

$\Sigma(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$

$\Sigma y_i - \Sigma\hat{\beta}_0 - \Sigma\hat{\beta}_1 x_i = 0$

* $\Sigma y_i = n\bar{y}$

$\Sigma const. = n \cdot const.$

$\frac{n\bar{y} - n\hat{\beta}_0 - n\hat{\beta}_1\bar{x} = 0}{n}$

$y - \hat{\beta}_0 - \hat{\beta}_1 \bar{x} = 0$

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

$\frac{\partial}{\partial \hat{\beta}_1} \Sigma(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \Rightarrow \Sigma -2x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$

$-\cancel{2}\,\Sigma x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = \cancel{0}$

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

$-\Sigma x_i\left[y_i - (\bar{y} - \hat{\beta}_1\bar{x}) - \hat{\beta}_1 x_i\right] = 0$

$-\Sigma x_i y_i + \Sigma x_i \bar{y} - \Sigma x_i \hat{\beta}_1 \bar{x} + \Sigma \hat{\beta}_1 x_i^2 = 0$

const

$\hat{\beta}_1(\Sigma x_i^2 - n\bar{x}^2) = \Sigma x_i y_i - n\bar{x}\bar{y}$

$\boxed{\hat{\beta}_1 = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\Sigma(x_i-\bar{x})^2}}$