Stat 6021: Homework 1

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).

   (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.

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

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

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

   (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.

   (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.

(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

(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?

(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`?

(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.

(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.

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

(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.

(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.

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

(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?

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

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

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).

   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`.

(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`.

(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.

(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.

(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?

(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.

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

i. Albemarle, Virginia
ii. Charlottesville city, Virginia

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`.

(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`.

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

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

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

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

(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.

(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.

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`

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?

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

(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?

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

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.

(f) Interpret the estimated slope contextually.

(g) Does the estimated intercept make sense contextually?

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

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

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

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

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

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

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

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

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

(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.

(r) Report the 95% prediction interval for the body mass (in grams) of a Gentoo penguin with 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?

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.

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

(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?

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

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

(e) Use the `lm()` function to fit a linear regression for the two variables. Where are the values of $\hat{\beta}_1$, $\hat{\beta}_0$, R^2, and $\hat{\sigma}^2$ for this linear regression?

(f) Interpret the values of $\hat{\beta}_1$, $\hat{\beta}_0$ contextually. Does the value of $\hat{\beta}_0$ make sense in this context?

(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?

(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.

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.

x_i	70	75	80	80	85	90
y_i	75	82	80	86	90	91
$\hat{y}_i$						
e_i						

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

	DF	SS	MS	F-stat	p-value
Regression					0.0099
Residual				***	***
Total			***	***	***

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

(d) What is the value of R^2 here? Interpret this value in context.

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

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.

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

(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).

(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.

8. Derive the least squares estimators of the simple linear regression model, i.e. show that

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \tag{1}$$

and

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2}$$

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

$$SS_{res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \tag{3}$$

Hint 1: Note that $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

Hint 2: Take partial derivatives of (3) w.r.t. $\hat{\beta}_1$ and $\hat{\beta}_0$.

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

$$\begin{aligned} \sum (x_i - \bar{x})^2 &= \sum x_i^2 - n\bar{x}^2, \\ \sum (x_i - \bar{x})(y_i - \bar{y}) &= \sum x_i y_i - n\bar{x}\bar{y}. \end{aligned}$$

Hint 4: Work through showing equations (1) and (2) in order.