Fall 2020 Midterm I SOLUTIONS

1. [0.5 point] The overall rate of diabetes in county A is greater than in county B. Therefore, the rate of diabetes for each age group in county A must be greater than the rate of diabetes in the corresponding age group in county B. Is this statement true or false?

(a) True
(b) False

SOLUTION: False. By Simpsons's paradox, this relationship could reverse.

Question 2 [1 point total]

Use the data below on a study looking at the effects of anger (measured by the Speilbere Trait Anger Scale test) and coronary heart disease (CHD) to answer the following questions.

	Low Anger	Moderate Anger	High Anger	Total
CHD	53	110	27	190
No CHD	3057	4621	606	8284
Total	3110	4731	633	8474

2.1 [0.5 point] What percent of those individuals who were classified as "High Anger" developed CHD?

(a) Approximately 14%
(b) Approximately 4%
(c) Approximately 0.3%
(d) Approximately 7%

SOLUTION: (b) 27/633 = 4.27%

2.2 [0.5 point] This percent is part of the _____.

(a) Marginal distribution of CHD
(b) Conditional distribution of anger given CHD
(c) Marginal distribution of anger
(d) Conditional distribution of CHD given high anger

SOLUTION: (d) Conditional distribution of CHD given high anger

3. [0.5 point] You are given a dataset, `covid_data` which has 6 columns (`id`, `county`, `state`, `num_deaths`, `population` and `num_uninsured`). Which line of code could you run so that there are exactly 5 columns in the output data frame?

(a) `covid_data %>% rename(county_name = county)`
(b) `covid_data %>% select(- num_uninsured)`
(c) `covid_data %>% filter(state == "California")`
(d) `covid_data %>% select(county, population)`

SOLUTION: (b) `covid_data %>% select(-num_uninsured)`

4. [0.5 point] With `covid_data %>% arrange(state, -population)`, how will this line of code sort the data?

(a) Sort `state` in descending order first, then `population` in ascending order
(b) Sort `state` in ascending order first, then `population` in ascending order
(c) Sort `state` in ascending order first, then `population` in descending order
(d) Sort `population` in ascending order first, then `state` in descending order

SOLUTION: (c) Sort `state` in ascending order first, then `population` in descending order

5. [1 point] What functions are necessary to visualize the distribution of a categorical variable? Choose all that apply.

(a) geom_histogram()
(b) ggplot()
(c) geom_point()
(d) geom_bar()
(e) aes()
(f) geom_cat()

SOLUTION: (b) ggplot(), (d) geom_bar(), (e) aes()

Question 6 [2.5 points total]

In your job as an analyst, your supervisor asks you to analyze data from the National Survey on Drug Use and Health from the Substance Abuse and Mental Health Data Archive.

Each row in the dataset `drug_dat` corresponds to an age group, with variables summarizing drug use across ages. The variable `heroin_use` gives the percentage of heroin use for the corresponding age group. Here are the first six rows of `age` and `heroin_use`

```
drug_dat <- read_csv("./drug_use_by_age_shaziap1.csv")
drug_dat %>% select(age, heroin_use) %>% head()
```

```
## # A tibble: 6 x 2
##   age   heroin_use
##   <chr>      <dbl>
## 1 12         0.025
## 2 13         0.03
## 3 14         0.05
## 4 15         0.04
## 5 16         0.03
## 6 17         0.1
```

6.1 [0.5 point] What type of variable is Heroin Usage (%)? Select all that apply.

(a) Categorical
(b) Quantitative
(c) Nominal
(d) Ordinal
(e) Continuous
(f) Discrete

SOLUTION: (b) Quantitative and (e) Continuous

6.2 [2 points] I am interested in examining the relationship between heroin

use and age. _____ is the explanatory variable in this plot and will

go on the _____ axis. I will use geom_ _____ to make this

plot. From the data, one thing I can say about the plot without making it is

that the relationship is _____.

SOLUTION:

age

x

scatter (or line)

the relationship is positive (increasing)

Question 7 [3.5 points total]

The dataset `food_data` includes percentage of food intake for different categories of food, with a row of data for each of 170 countries. The dataset also includes the proportion of the country's population who are obese, the proportion undernourished, and and the % of COVID-19 cases.

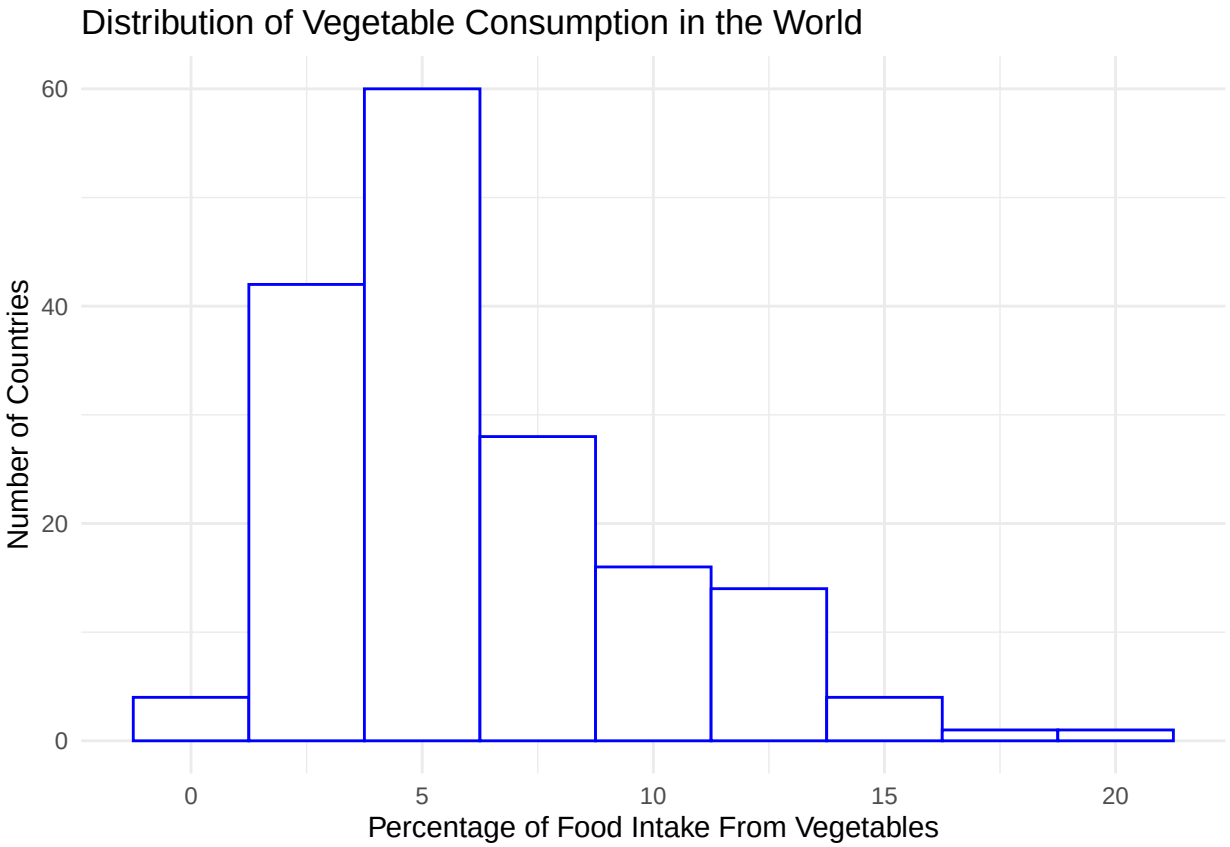
```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   Continent = col_character(),
##   Country = col_character(),
##   Undernourished = col_character(),
##   `Unit (all except Population)` = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

```
head(food_dat)
```

```
## # A tibble: 6 x 33
##   Continent Country             `Alcoholic Bever~ `Animal fats` `Animal Product~
##   <chr>     <chr>                           <dbl>         <dbl>            <dbl>
## 1 Asia      Afghanistan                    0.0014        0.197             9.43
## 2 Europe    Albania                        1.67          0.136            18.8
## 3 Africa    Algeria                        0.271         0.0282            9.63
## 4 Africa    Angola                         5.81          0.056             4.93
## 5 N America Antigua and Barbuda            3.58          0.0087           16.7
## 6 S America Argentina                      4.27          0.223            19.3
## # ... with 28 more variables: Aquatic Products, Other <dbl>,
## #   Cereals - Excluding Beer <dbl>, Eggs <dbl>, Fish, Seafood <dbl>,
## #   Fruits - Excluding Wine <dbl>, Meat <dbl>, Milk - Excluding Butter <dbl>,
## #   Miscellaneous <dbl>, Offals <dbl>, Oilcrops <dbl>, Pulses <dbl>,
## #   Spices <dbl>, Starchy Roots <dbl>, Stimulants <dbl>,
## #   Sugar & Sweeteners <dbl>, Sugar Crops <dbl>, Treenuts <dbl>,
## #   Vegetable Oils <dbl>, Vegetables <dbl>, Vegetal Products <dbl>, ...
```

Use this histogram to answer parts 1-4.



7.1 [1 point] Describe the distribution.

SOLUTION: Shape: unimodal (0.5 point), skewed to the right (0.5 point)

7.2 [0.5 point] Pick the sentence that is most correct.

(a) The mean is approximately equal than 5
(b) The mean is larger than 5
(c) The mean is smaller than 5
(d) Not enough information to choose

SOLUTION: (b) The mean is larger than 5

7.3 [1 point] Select all true statements based on the histogram and knowledge you've gained in this class.

(a) mean = median
(b) mean > median
(c) mean < median
(d) The mean is resistant to outliers.
(e) The median is resistant to outliers.

SOLUTION: (b) mean > median and (e) The median is resistant to outliers.

7.4 [1 point] What is the binwidth for this distribution?

SOLUTION: 2.5

Question 8 [2.5 points total]

The data set named `diabt` contains information about diabetic and non-diabetic patients. In particular, the variable `diabetes` equals 0 for individuals without diabetes, equals 1 for individuals with type 1 diabetes and equals 2 for individuals with type 2 diabetes.

Here is some information about these data:

```
dim(diabt)
```

```
## [1] 18  6
```

```
head(diabt)
```

```
## # A tibble: 6 x 6
##   nameid height_cm weight_kg sex   race  diabetes
##   <chr>      <dbl>     <dbl> <fct> <chr>    <dbl>
## 1 ADF          160        75 1     white        2
## 2 PUD          186        78 1     white        0
## 3 HYD          155        49 1     blakc        0
## 4 RFD          150        64 1     blakc        1
## 5 UDF          172        72 1     white        0
## 6 USR          174       123 1     blakc        1
```

8.1 [0.5 points] What type of variable is `diabetes`? Choose the best answer.

(a) continuous
(b) discrete
(c) categorical
(d) ordinal

SOLUTION: (c) categorical

8.2 [2 points] Write code to make a chart (histogram or bar) for the distribution of the types of diabetes, where there is a separately colored bar for men and women; these bars are next to each other, within each type of diabetes.

SOLUTION:

```
ggplot(diabt, aes(x=diabetes)) +
  geom_bar(aes(fill = sex), position = "dodge")
```

8.3 [2 points] The formula to calculate BMI is $\frac{weight(kg)}{height^2(m)}$. Add a variable named `bmi` to the `diabt` dataframe.

SOLUTION:

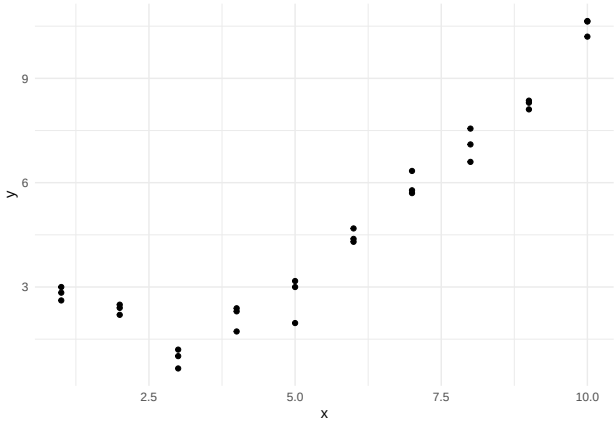
```
diabt <- diabt %>% mutate(bmi = weight_kg / (height_cm / 100)^2)
```

9. [0.5 points] True or False: Correlations for average measures are usually stronger than correlations based on individual data.

(a) True
(b) False

SOLUTION: (a) True. This follows the reasoning of how a least squares regression line works. If there are many points, they all influence the line of best fit and outliers can pull the line in a specific direction. With less influence from many more points pulling the line, the relationship will usually be stronger. Average values mask some of the individiaul to individual variation.

10. [0.5 points] The Pearson's correlation coefficient for this graph is likely close to:



a) 0.2
b) 0.4
c) 0.6
d) 0.8
e) You should not calculate Pearson's correlation for this relationship

SOLUTION: e) You should not calculate Pearson's correlation for this relationship.

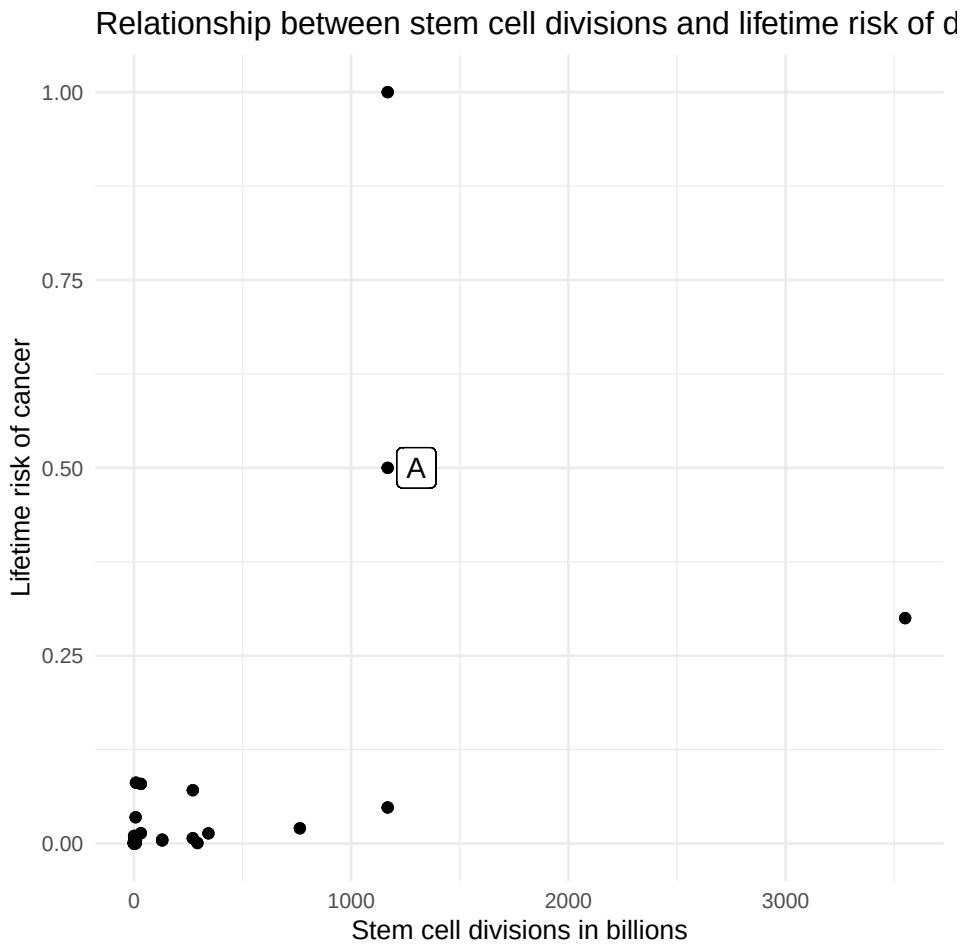
Question 11 [6 points total]

You are interested in visualizing the relationship between the number of stem cell divisions and one's lifetime risk of different types of cancer. To investigate, you have a dataset called `cancer_data`, with a row of data for each of various types of cancers:

```
head(cancer_data)
```

```
## # A tibble: 6 x 3
##   disease          lifetime_risk stem_cell_divisions
##   <chr>                    <dbl>               <dbl>
## 1 AM leukemia             0.0041               130.
## 2 Basal Cell              0.3                 3550
## 3 CL Leukemia             0.0052               130.
## 4 Colorectal              0.048               1168
## 5 FAP Colorectal          1                   1168
## 6 Lynch Colorectal        0.5                 1168
```

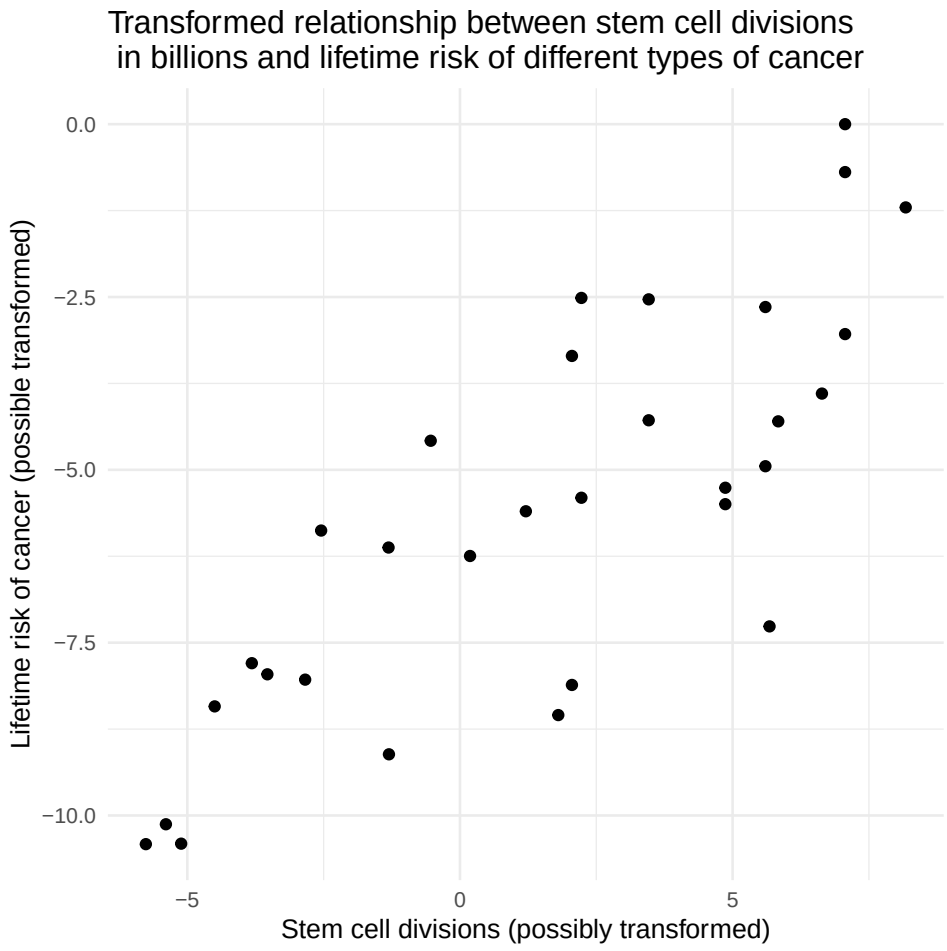
You create a scatterplot of `lifetime_risk` versus `stem_cell_divisions`:



11.1 [2 points] In 1-3 *brief* sentences, how would you describe these data? Would you want to use a linear model to summarize this relationship?

SOLUTION: Most of the points are clumped in the bottom left corner. No you wouldn't want to use a linear model.

You decide to transform your data and make a plot of the transformed relationship as shown below.



11.2 [3 points] What transformation did you likely perform on the the explanatory and/or the response variables to produce the second scatterplot? To make sure you picked the correct function, apply the transformation to your best guess of the x and y values for the point labeled A in the first plot and show that it roughly corresponds to a point on the second plot.

SOLUTION: log transformed both x and y. Take the datapoint at 1100 (approximately) and 0.5. log 1100 = 7, log 0.5 = -0.69. Instead of 1100, they can use a number approximately close since they cannot read it off the page.

11.3 [1 point] A classmate says that according to this plot, it is clear that the number of stem cell divisions directly affects the lifetime risk of cancer. What is one concept you learned about in class that provides an alternate explanation for the linear relationship between these variables?

SOLUTION: should say either confounding or lurking variable.

Question 12 [8 points total]

The dataset `bupa` contains information about liver disorders. It contains data on 345 individuals' blood test results and liver disorder status. The following table shows the first six rows of `bupa`. The variables `SGPT` and `GAMMAGT` are both measurements of the patients' liver condition in the unit IU/L.

```
## # A tibble: 6 x 7
##     MCV ALKPHOS  SGPT  SGOT GAMMAGT DRINKS disorder
##   <dbl>   <dbl> <dbl> <dbl>   <dbl>  <dbl>    <dbl>
## 1    85      92    45    27      31      0        1
## 2    85      64    59    32      23      0        2
## 3    86      54    33    16      54      0        2
## 4    91      78    34    24      36      0        2
## 5    87      70    12    28      10      0        2
## 6    98      55    13    17      17      0        2
```

12.1 [2 points] You made a scatter plot of `SGPT` vs. `GAMMAGT` and based on the plot, decide it might be better to build a linear regression model using the natural log transformed variables `log_SGPT` and `log_GAMMAGT`: $\log(\text{SGPT}) = a + b * \log(\text{GAMMAGT})$. Write code that adds two new variables to `bupa`, and fits the linear model (saved as `bupa_model`).

SOLUTION:

```
`bupa <- bupa %>% mutate(log_SGPT = log(SGPT),** <p>/<p>
                         **log_GAMMAGT = log(GAMMAGT))`

bupa_model <- lm(log_SGPT ~ log_GAMMAGT, data = bupa)
```

12.2 [1 point] The summary of the fitted linear model is:

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    2.06     0.0986      20.9 6.05e-63
## 2 log_GAMMAGT    0.369    0.0290      12.7 1.50e-30
```

Interpret the slope parameter.

SOLUTION: A one unit increase in the logarithm of GAMMAPT(IU/L) is associated with an increase of 0.369 in the logarithm of SGPT(IU/L).

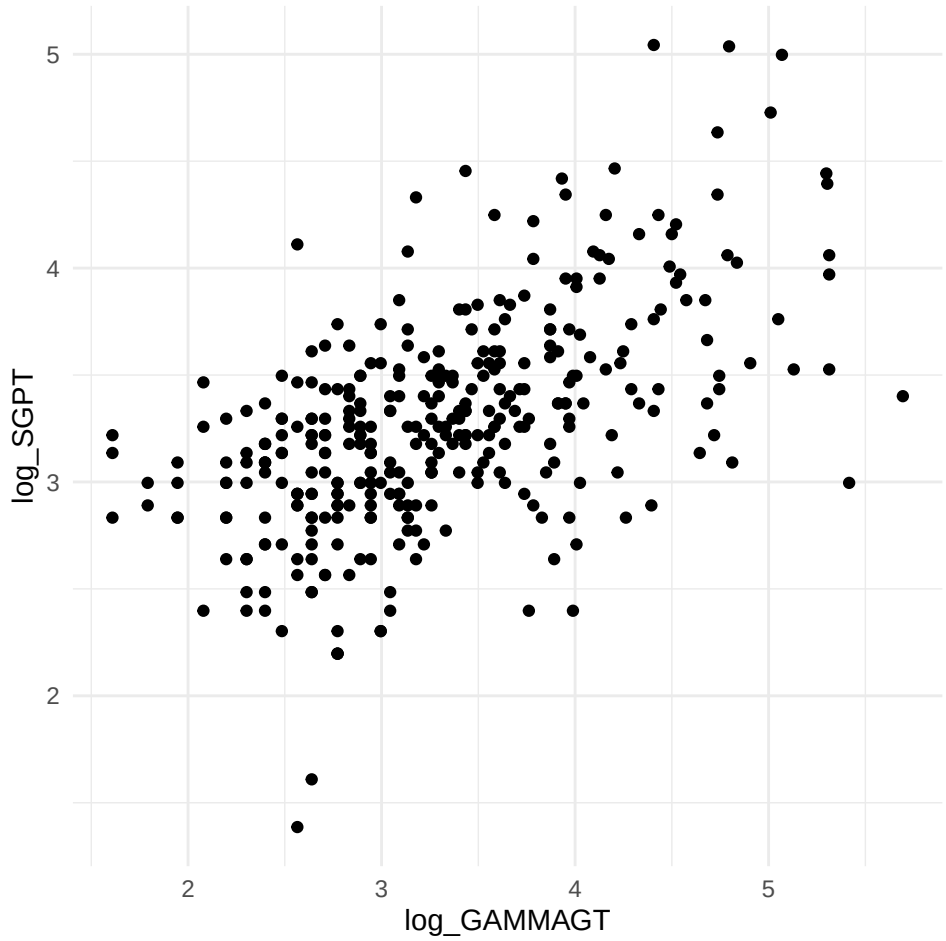
12.3 [1 point] Interpret the r-squared value based on the output below. Be specific.

```
glance(bupa_model)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     0.320         0.318 0.420      161. 1.50e-30     1  -189.  384.  396.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

SOLUTION: r_squared = 0.320 means that 32% of the variation in log_SGPT can be explained by the variations in log_GAMMAGT.

12.4 [2 points] Recall that you applied the log transformation on both variables. How would the r-squared value for the relationship between `SGPT` and `GAMMAGT` compare to the r-squared value for the relationship between `log_SGPT` and `log_GAMMAGT` and why? For reference, here is the scatter plot based on the transformed data:



SOLUTION: The original plot didn't look linear so if calculated the r-squared value for non-linear relationship, we suspect it would be lower than this clearly linear relationship.

12.5 [2 points] Explain why it is not a good idea to make a prediction of `SGPT` given `GAMMAGT`=5000 using the current data and model. Provide a calculation based on any information provided above to support your reasoning.

SOLUTION: extrapolation.

Calculations: log(5000) = 8.52, while the maximum value of log_GAMMAGT in the current dataset is 5.69. It is far from the bulk of the data and cannot be accurately predicted. Alternatively could compute exp(6) or exp(7) and see that is $<$ 1096 which is smaller than 5000.

13. [1 point] You have a dataset called `diet` that contains information on diet and the incidence of coronary heart disease (CHD) of individuals.

For reference, the variables in this dataset include:
`id`: subject identifier, numeric
`job`: occupation, that can take the values `Driver`, `Conductor`, and `Bank worker`
`energy`: total energy intake (kCal per day/100), numeric
`height`: in cm, numeric
`weight`: in kg, numeric
`fat`: fat intake (g), numeric
`chd`: CHD event, where the value 1 implies this individuals has had a CHD event, and 0 implies this individuals has had no CFD event

Write one line of code to create a new data frame called `diet_subset`, which only contains individuals who are drivers and have fat intakes larger than 100g.:

SOLUTION:

```
diet_subset <- diet %>% filter(job == "Driver", fat > 100)
```

Question 14 [6 points total]

The following data looks at the relationship between endometriosis and hypertension. A third variable included in this analysis is the genotype each woman has of a particular gene. The three levels are GG, GT, and TT.

```
## # A tibble: 6 x 6
##   endo_status genotype count count_with_ht genotype_prop percent_ht
##   <chr>       <chr>    <dbl>         <dbl>         <dbl>      <dbl>
## 1 Endo        GG          55            24         0.056       43.6
## 2 Endo        GT         768           344         0.784       44.8
## 3 Endo        TT         156            62         0.159       39.7
## 4 No Endo     GG        2401          1121         0.261       46.7
## 5 No Endo     GT        4393          2028         0.478       46.2
## 6 No Endo     TT        2395          1007         0.261       42.0
```

14.1 [1 point] Using the data, fill in the blanks of the following two-way table.

	Hypertension	No Hypertension	Total
Endo	430	A	979
No Endo	B	C	9194
Total	4586	D	10173

A:
B:

C:
D:

SOLUTION: A = 549, B = 4156, C = 5038, D = 5587

14.2 [1 point] What is the marginal distribution of endometriosis in this population? Round your answer to 2 decimal places.

SOLUTION: The percentage of women who have endometriosis is 979/10173 is 9.62%. The percentage of women who do not have endometriosis is 90.38%.

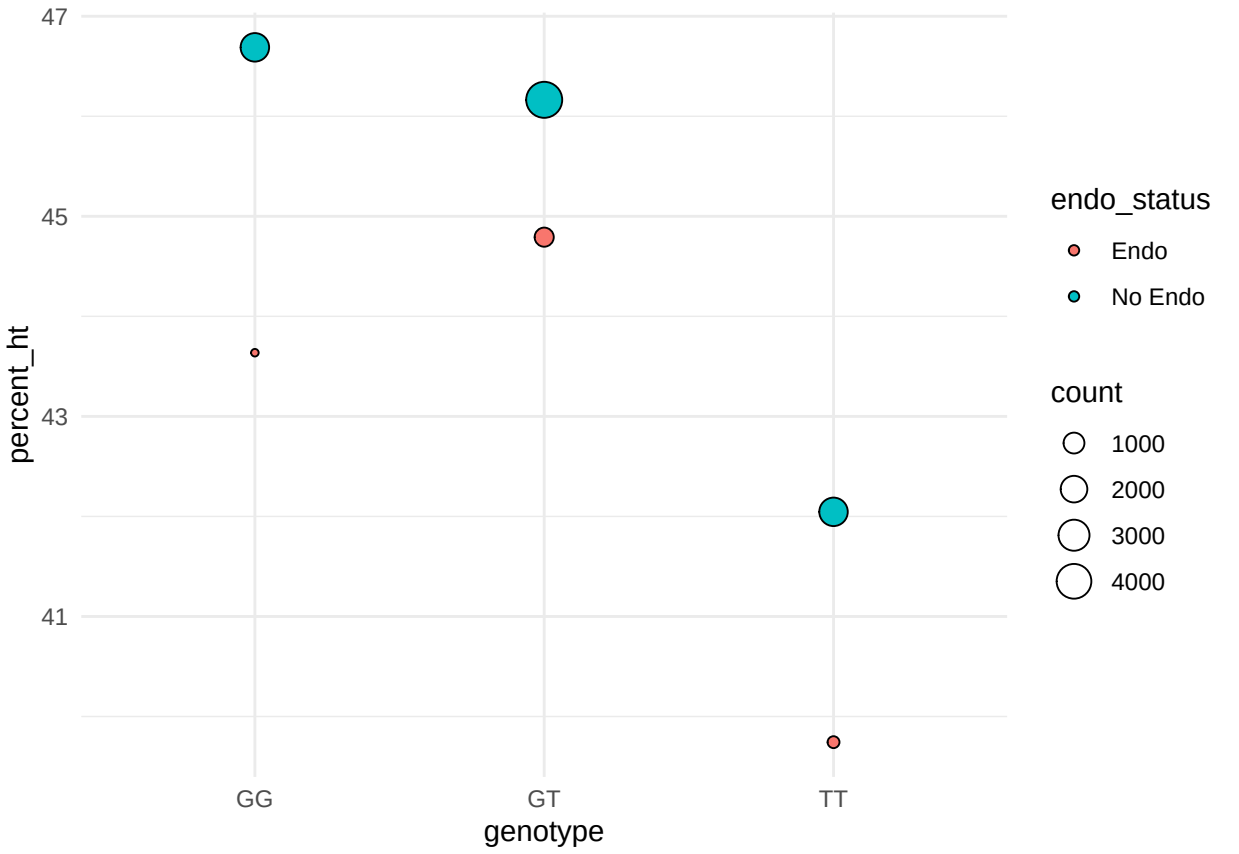
14.3 [1 point] What is the conditional distribution of hypertension among women with endometriosis?

SOLUTION: The percentage of women who have hypertension given they have endometriosis is 43.92%. The percentage of women who do not have hypertension given they have endometriosis is 56.08%.

14.4 [1 point] Which group has the highest overall rate of hypertension?

(a) Endo
(b) No endo

CORRECTED SOLUTION: (b) No endo has the highest overall rate of hypertension.
ORIGINAL SOLUTION: (a) Endo has the highest overall rate of hypertension.




```
## $x
## [1] "Genotype"
##
## $y
## [1] "Percent with Hypertension"
##
## attr(,"class")
## [1] "labels"
```

14.5 [2 points] From the visualization above, it is evident that within each genotype, there is a higher incidence of hypertension in the group without endo than the group with endo. In 1-3 *brief* sentences and using your answer in Part E, identify the cause of this phenomenon, and explain why that is the cause. Hint: Look at the variable `genotype_prop` in your dataframe.

CORRECTED SOLUTION: This question no longer applies with the corrected values input in the 2x2 table. You may want to change the numbers around so Simpson's paradox still applies to keep this question.

ORIGINAL SOLUTION: Overall, there is a higher incidence of hypertension in the endo group but within each strata of genotype, there is a higher incidence of hypertension in the non-endo group. This is because genotype is a confounding variable for the relationship between endo and hypertension. There is a different distribution of genotype among women with endo and women without leading us to believe that genotype has an effect on both endo and hypertension.