Midterm project - Esophageal Cancer Analysis

Abigail Castro

Exploratory data analysis is compared to detective work: it is the process of gathering evidence.

Please load the esoph dataset and other necessary packages in R and follow the following steps:

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.6     v dplyr   1.0.3
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
```

1. Read the description of the dataset and learn about its variables. List two main questions/hypotheses you want to test about your data.
2. Identify at least two variables that are relevant to your questions. Argue why other variables may not be relevant to the questions.
3. Search for evidence by visualising, transforming, and modeling your data.
4. Discuss the evidence to answer/refine your questions, test your hypotheses, and/or generate new questions. Summarize and conclude your findings.

Step 1. Describe what the dataset is about and list at least two main questions/hypotheses you want to test about the data. (10 points)

Question 1:

Which age group(s) have the highest percentage of cases?

Question 2:

Does tobacco consumption or alcohol consumption have a greater affect on the percentage of cases?

Question 3, 4, . . . if any

How does tobacco consumption affect the percentage of cases in each age group? How does alcohol consumption affect the percentage of cases in each age group?

Step 2. Identify at least two variables that are relevant to each question. (10 points)

Varaibles relevant to Q1

- What are the types of those variables? How do you determine that? Since we are using percentages, we must create a new column in our data set, which is done below.

```
d1<- esoph %>%  group_by(agegp) %>%
  summarise(count = n(), total_cases = sum(ncases), total_controls = sum(ncontrols),
            percentage=total_cases*100/total_controls)
d1
```

```
## # A tibble: 6 x 5
##   agegp count total_cases total_controls percentage
## * <ord> <int>       <dbl>          <dbl>      <dbl>
## 1 25-34    15           1            116      0.862
## 2 35-44    15           9            199      4.52
## 3 45-54    16          46            213     21.6
## 4 55-64    16          76            242     31.4
## 5 65-74    15          55            161     34.2
## 6 75+      11          13             44     29.5
```

By printing d1, we find that agegp is a categorical variable with type order factor with 6 levels, and percentage is a continuous variable from 0 to 100 with a double data type.

- Describe why these variables may be relevant to this question and why other variables are not relevant The age group variable is important to this question because we would like to see if there's a trend between age and the percentage of cancer cases. Since we are using a percentage of cancer cases, the variables, ncases and ncontrols, which are of type double will both be used. Percentages are used for comparison so that the data analysis is more meaningful since the number of participants in each age group varies. Alcohol and tobacco consumption is not relevant to this question, but will be used in the investigation of next question.

Variables relevant to Q2

- What are the types of those variables? How do you determine that?

```
esoph
```

```
##    agegp     alcgp    tobgp ncases ncontrols
## 1  25-34 0-39g/day 0-9g/day      0        40
## 2  25-34 0-39g/day    10-19      0        10
## 3  25-34 0-39g/day    20-29      0         6
```

```
## 4  25-34 0-39g/day      30+      0         5
## 5  25-34     40-79 0-9g/day      0        27
## 6  25-34     40-79    10-19      0         7
## 7  25-34     40-79    20-29      0         4
## 8  25-34     40-79      30+      0         7
## 9  25-34    80-119 0-9g/day      0         2
## 10 25-34    80-119    10-19      0         1
## 11 25-34    80-119      30+      0         2
## 12 25-34      120+ 0-9g/day      0         1
## 13 25-34      120+    10-19      1         1
## 14 25-34      120+    20-29      0         1
## 15 25-34      120+      30+      0         2
## 16 35-44 0-39g/day 0-9g/day      0        60
## 17 35-44 0-39g/day    10-19      1        14
## 18 35-44 0-39g/day    20-29      0         7
## 19 35-44 0-39g/day      30+      0         8
## 20 35-44     40-79 0-9g/day      0        35
## 21 35-44     40-79    10-19      3        23
## 22 35-44     40-79    20-29      1        14
## 23 35-44     40-79      30+      0         8
## 24 35-44    80-119 0-9g/day      0        11
## 25 35-44    80-119    10-19      0         6
## 26 35-44    80-119    20-29      0         2
## 27 35-44    80-119      30+      0         1
## 28 35-44      120+ 0-9g/day      2         3
## 29 35-44      120+    10-19      0         3
## 30 35-44      120+    20-29      2         4
## 31 45-54 0-39g/day 0-9g/day      1        46
## 32 45-54 0-39g/day    10-19      0        18
## 33 45-54 0-39g/day    20-29      0        10
## 34 45-54 0-39g/day      30+      0         4
## 35 45-54     40-79 0-9g/day      6        38
## 36 45-54     40-79    10-19      4        21
## 37 45-54     40-79    20-29      5        15
## 38 45-54     40-79      30+      5         7
## 39 45-54    80-119 0-9g/day      3        16
## 40 45-54    80-119    10-19      6        14
## 41 45-54    80-119    20-29      1         5
## 42 45-54    80-119      30+      2         4
## 43 45-54      120+ 0-9g/day      4         4
## 44 45-54      120+    10-19      3         4
## 45 45-54      120+    20-29      2         3
## 46 45-54      120+      30+      4         4
## 47 55-64 0-39g/day 0-9g/day      2        49
## 48 55-64 0-39g/day    10-19      3        22
## 49 55-64 0-39g/day    20-29      3        12
## 50 55-64 0-39g/day      30+      4         6
## 51 55-64     40-79 0-9g/day      9        40
## 52 55-64     40-79    10-19      6        21
## 53 55-64     40-79    20-29      4        17
## 54 55-64     40-79      30+      3         6
## 55 55-64    80-119 0-9g/day      9        18
## 56 55-64    80-119    10-19      8        15
## 57 55-64    80-119    20-29      3         6
```

```
## 58 55-64    80-119      30+      4         4
## 59 55-64      120+ 0-9g/day      5        10
## 60 55-64      120+    10-19      6         7
## 61 55-64      120+    20-29      2         3
## 62 55-64      120+      30+      5         6
## 63 65-74 0-39g/day 0-9g/day      5        48
## 64 65-74 0-39g/day    10-19      4        14
## 65 65-74 0-39g/day    20-29      2         7
## 66 65-74 0-39g/day      30+      0         2
## 67 65-74     40-79 0-9g/day     17        34
## 68 65-74     40-79    10-19      3        10
## 69 65-74     40-79    20-29      5         9
## 70 65-74    80-119 0-9g/day      6        13
## 71 65-74    80-119    10-19      4        12
## 72 65-74    80-119    20-29      2         3
## 73 65-74    80-119      30+      1         1
## 74 65-74      120+ 0-9g/day      3         4
## 75 65-74      120+    10-19      1         2
## 76 65-74      120+    20-29      1         1
## 77 65-74      120+      30+      1         1
## 78   75+ 0-39g/day 0-9g/day      1        18
## 79   75+ 0-39g/day    10-19      2         6
## 80   75+ 0-39g/day      30+      1         3
## 81   75+     40-79 0-9g/day      2         5
## 82   75+     40-79    10-19      1         3
## 83   75+     40-79    20-29      0         3
## 84   75+     40-79      30+      1         1
## 85   75+    80-119 0-9g/day      1         1
## 86   75+    80-119    10-19      1         1
## 87   75+      120+ 0-9g/day      2         2
## 88   75+      120+    10-19      1         1
```

By printing esoph, we find that tobgp and alcgp are both categorical variables with type order factor with 4 levels. Percentage will be used to create the chart in a later section and since percentage is computed as (ncases/ncontrols)*100, it will be a double.

- Describe why these variables may be relevant to this question and why other variables are not relevant
  We are comparing the the effects of alcohol consumption and tobacco consumption so both variables will be used in the analysis. Since we are using a percentage of cancer cases, the variables, ncases and ncontrols, which are of type double will both be used. Percentages are used for comparison so that the data analysis is more meaningful since the number of participants in each age group varies. Agegp will not be used in the investigation of this question, since it will distract from the original purpose of comparing alcohol and tobacco consumption by adding an extraneous layer. Additionally, agegp was investigated in the first question.

Step 3. Search for evidence by visualising, transforming, and modeling your data (60 points)

(Check RDS 3, 5, 7.3, 7.4, 7.5, 7.6 for ideas and inspiration)

3.1 What type of variation occurs within each variable? (30+ points)

```
summary(esoph)
```

```
##    agegp          alcgp         tobgp        ncases         ncontrols
##  25-34:15   0-39g/day:23   0-9g/day:24   Min.   : 0.000   Min.   : 1.00
##  35-44:15   40-79    :23   10-19   :24   1st Qu.: 0.000   1st Qu.: 3.00
##  45-54:16   80-119   :21   20-29   :20   Median : 1.000   Median : 6.00
##  55-64:16   120+     :21   30+     :20   Mean   : 2.273   Mean   :11.08
##  65-74:15                                3rd Qu.: 4.000   3rd Qu.:14.00
##  75+  :11                                Max.   :17.000   Max.   :60.00
```

```
summary(d1)
```

```
##    agegp       count         total_cases    total_controls    percentage
##  25-34:1   Min.   :11.00   Min.   : 1.00   Min.   : 44.0   Min.   : 0.8621
##  35-44:1   1st Qu.:15.00   1st Qu.:10.00   1st Qu.:127.2   1st Qu.: 8.7910
##  45-54:1   Median :15.00   Median :29.50   Median :180.0   Median :25.5709
##  55-64:1   Mean   :14.67   Mean   :33.33   Mean   :162.5   Mean   :20.3488
##  65-74:1   3rd Qu.:15.75   3rd Qu.:52.75   3rd Qu.:209.5   3rd Qu.:30.9401
##  75+  :1   Max.   :16.00   Max.   :76.00   Max.   :242.0   Max.   :34.1615
```

```
d1
```

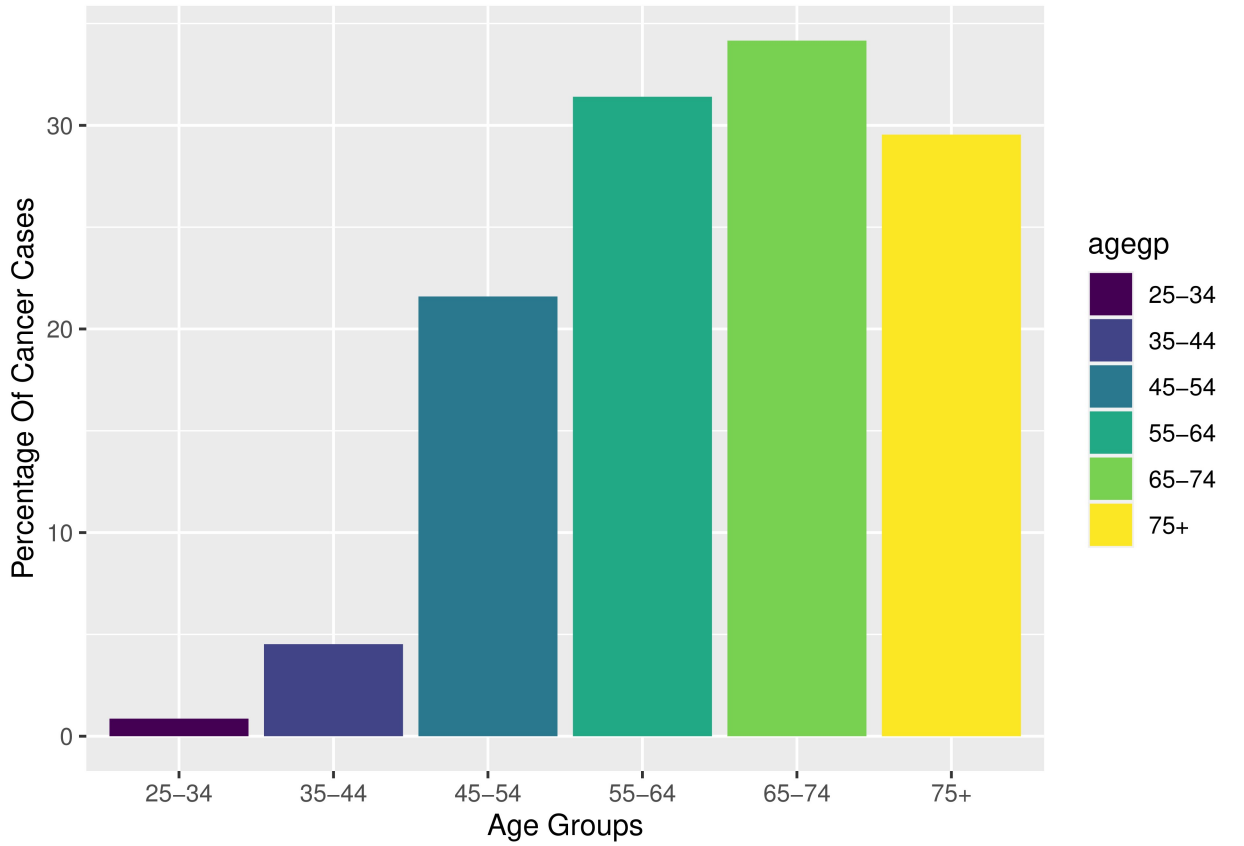
```
## # A tibble: 6 x 5
##   agegp count total_cases total_controls percentage
## * <ord> <int>       <dbl>          <dbl>      <dbl>
## 1 25-34    15           1            116      0.862
## 2 35-44    15           9            199      4.52
## 3 45-54    16          46            213     21.6
## 4 55-64    16          76            242     31.4
## 5 65-74    15          55            161     34.2
## 6 75+      11          13             44     29.5
```

As seen in the summary of esoph, the median for ncases is less than the mean which means that it is skewed right. Additionally, the range excluding outliers would be from -3.5 to 8.5. The median for ncontrols is less than the mean which means it is skewed right. Also, the range excluding outliers would be -13.5 to 27.5. Based on the summary of d1, the median is greater than the mean which means that it is skewed left. The range for percentage is excluding outliers would be from -24.43 to 64.16, and since the minimum and maximum of the percentage column are within this range, there are no outliers. Since agegp, tobgp, and alcgp are factors you cannot measure the variation in it, but based on the count, the number of observations of each factor are relatively close to each other.

3.1.1 Variable 1 (15 points)

3.1.1.1 Visualising distributions (Barcharts, Histograms) (5 points) Graph

```
ggplot(d1, aes(x=agegp, y=percentage,fill=agegp)) +
  geom_bar(stat="identity") +
  labs(x= 'Age Groups', y= 'Percentage Of Cancer Cases')
```



- Which values are the most common? Why? The graph shows that a larger percentage of participants between the ages of 45-75+ have cancer, most likely since the body has had more time for damage to build up in the cells eventually leading to cancer.

- Which values are rare? Why? Does that match your expectations? The graph shows that a smaller percentage of younger participants have cancer, which I believe makes sense since younger people tend to be healthier overall.
- Can you see any unusual patterns? What might explain them? One possibly unusual pattern is that the percentage of cancer cases decreased from the age group 65-74 to 75+. Since the average life expectancy in the US is about 73 years, those who live longer are probably a bit healthier, which could account for the slight drop in percentage.
- Are there clusters in the data? There do not seem to be any clusters in the data. If so,
  - How are the observations within each cluster similar to or different from each other?
  - How can you explain or describe the clusters?

The general trend displayed by the graph is that the percentage of cases increases as age increases.

3.1.1.2 Unusual values (2 points)

- Describe and demonstrate how you determine if there are unusual values in the data. E.g. too large, too small, negative, etc. Describe and demonstrate how you determine if they are outliers.

Using summary the summary of d1, you can find the IQR and the range for outliers. The range for percentage excluding outliers would be from -24.43 to 64.16, and since the minimum and maximum of the percentage column are within this range, there are no outliers and hence no unusual values.

- Show how do your distributions look like with and without the unusual values. N/A
- Discuss whether or not you need to remove unusual values and why. In general, unusual values can skew your data which makes your observations less meangingful, so outliers should be taken out if possible.

3.1.1.3 Missing values (RDS 5.2) (2 points)

- Does this variable include missing values? Demonstrate how you determine that.
- Demonstrate and discuss how you handle the missing values. E.g., removing, replacing with a constant value, or a value based on the distribution, etc.
- Show how your data looks in each case after handling missing values.Describe and discuss the distribution.

One way to determine if there is missing values is to use the summary(function) which will give you statistics such as mean, median, number of NAs, etc., you can also just look through the dataset if it's small enough. By using both of these mehtods, there aren't any explicitly missing values, but there are some implicitly missing values such as the observation for the age group 24-34, alcohol consumption group 80-119, and tobacco consumption group 20-29 is not included within the dataset, which means that they probably did not find any participants that matched that profile. Since they are implicitly missing and the we are using percentages, these values do not have much affect on the data analysis.

3.1.1.4 Does converting the type of this variable help exploring the distribution of its values or identifying outliers or missing values? (3)

- What type can the variable be converted to? How will the distribution look? Please demonstrate with appropriate plots.

Since the type of percentage is double, it can already be used to identify outliers or missing values, so it does not need to be converted.The distribution is shown by the graph found above which shows that the data is skewed left.

3.1.1.5 What new variables do you need to create from this? (RDS 5.5, 5.6, 5.7) (3)

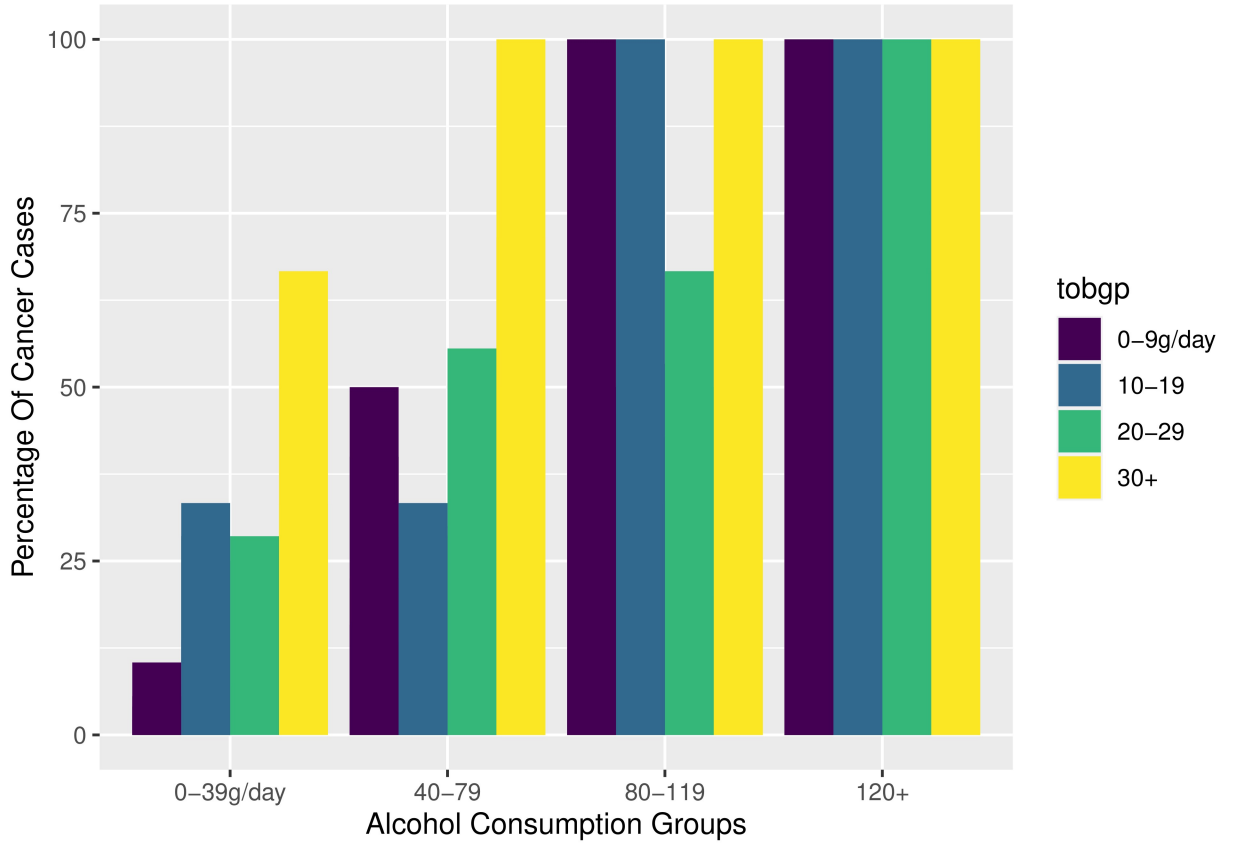
- List the variables
- Describe and discuss why they are needed and how you plan to use them.

No new variables need to be created to answer the question.

3.1.2 Variable 2 (15 points)

3.1.2.1 Visualising distributions (Barcharts, Histograms) (5 points) Graph

```
ggplot(esoph, aes(factor(alcgp), ncases*100/ncontrols, fill = tobgp)) +
  geom_bar(stat="identity", position = "dodge") +
  labs(x= 'Alcohol Consumption Groups', y= 'Percentage Of Cancer Cases')
```



- Which values are the most common? Why?

The graph shows that the percentage is largest for 90-120+ g/day of alcohol most likely because large alcohol consumption increase the risk of cancer significantly.

- Which values are rare? Why? Does that match your expectations?

The percentages for small amounts of alcohol and tobacco consumption are also small, which I believe also matches expectations since alcohol and tobacco are generally viewed as unhealthy, consuming a smaller amount should lead to better health overall.

- Can you see any unusual patterns? What might explain them?

An interesting pattern is that for 0-79 g/day of alcohol consumed, the tobacco consumption has a clear difference in the percentage of cases with larger consumption corresponding to higher percentages, but between 80-120+ g/day of alcohol consumed there is practically no difference between the tobacco consumption factors. The increase in alcohol consumption most likely significantly increases the percentage of cases because large amounts of alcohol can damage your DNA and prevent your body from repairing it. This damaged molecular material can eventual lead to cancer.

- Are there clusters in the data? If so,

There does not appear to be clusters in the data.

- How are the observations within each cluster similar to or different from each other?
- How can you explain or describe the clusters?

3.1.2.2 Unusual values (2 points)

- Describe and demonstrate how you determine if there are unusual values in the data. E.g. too large, too small, negative, etc.
- Describe and demonstrate how you determine if they are outliers.
- Show how do your distributions look like with and without the unusual values.
- Discuss whether or not you need to remove unusual values and why.

Since these are categorical variables, it is hard to determine unusual values.

3.1.2.3 Missing values (2 points)

- Does this variable include missing values? Demonstrate how you determine that. Demonstrate and discuss how you handle the missing values. E.g., removing, replacing with a constant value, or a value based on the distribution, etc.

One way to determine if there is missing values is to use the summary(function) which will give you statistics such as mean, median, number of NAs, etc., you can also just look through the dataset if it's small enough. By using both of these mehtods, there aren't any explicitly missing values, but there are some implicitly missing values such as the observation for the age group 24-34, alcohol consumption group 80-119, and tobacco consumption group 20-29 is not included within the dataset, which means that they probably did not find any participants that matched that profile. Since they are implicitly missing and the we are using percentages, these values do not have much affect on the data analysis.

- Show how your data looks in each case after handling missing values.Describe and discuss the distribution. N/A

3.1.2.4 Does converting the type of this variable help exploring the distribution of its values or identifying outliers or missing values? (3)

- What type can the variable be converted to?
- How will the distribution look? Please demonstrate with appropriate plots.

Since the type of percentage is double, it can already be used to identify outliers or missing values, so it does not need to be converted.The distribution is shown by the graph found above which demonstrates that the data is skewed left.

3.1.2.5 What new variables do you need to create? (3)

- List the variables
- Describe and discuss why they are needed and how you plan to use them.

No new variables need to be created to answer the question.

Variable 3, 4, ... if any (extra 15 points)

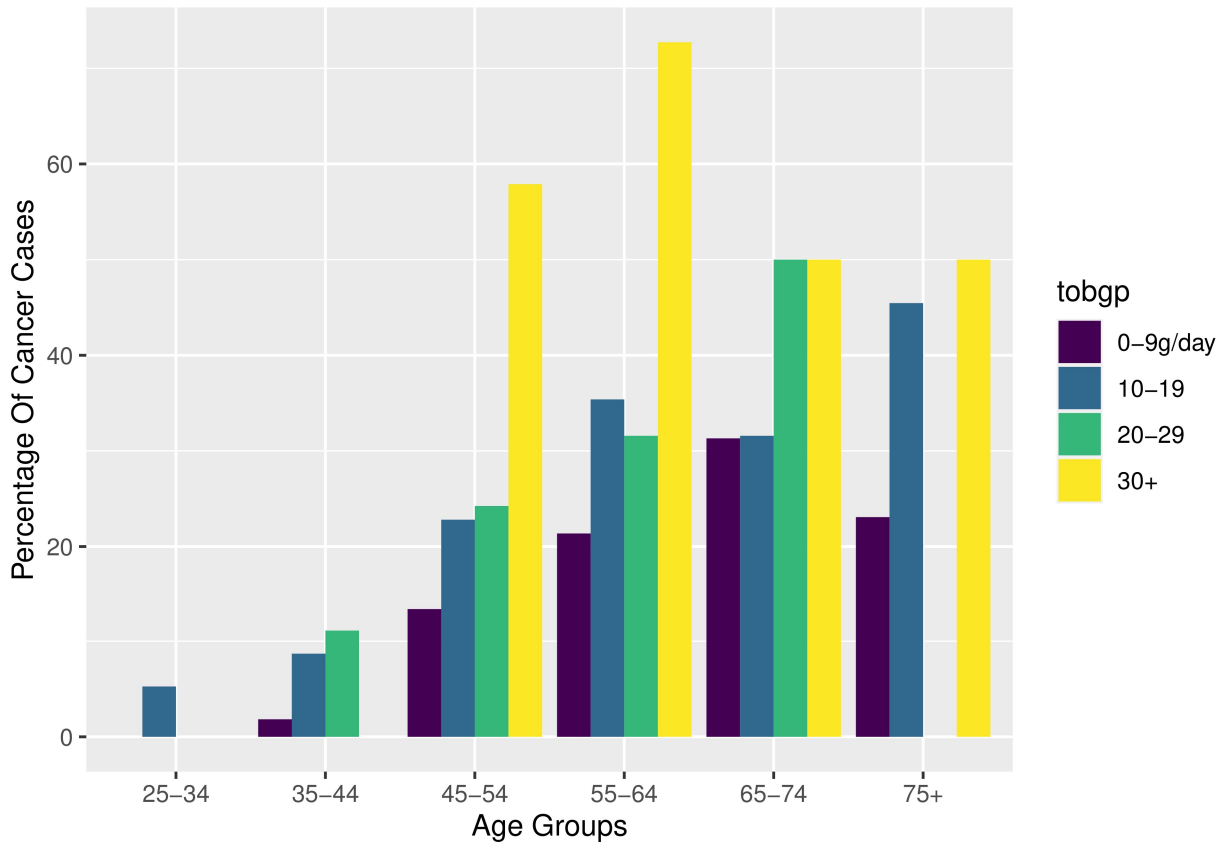
Repeat the same as above

Question 3

```
d2<- esoph %>%  group_by(agegp,tobgp) %>%
    summarise(count = n(), total_cases = sum(ncases), total_controls = sum(ncontrols),
              percentage_tob = total_cases*100/total_controls)
```

```
## `summarise()` has grouped output by 'agegp'. You can override using the `.groups` argument.
```

```
ggplot(d2, aes(agegp, percentage_tob, fill = tobgp)) +
   geom_bar(stat="identity", position = "dodge") +
   labs(x= 'Age Groups', y= 'Percentage Of Cancer Cases')
```



The general trend is that for each group as the tobacco consumption increases the percentage of cases also

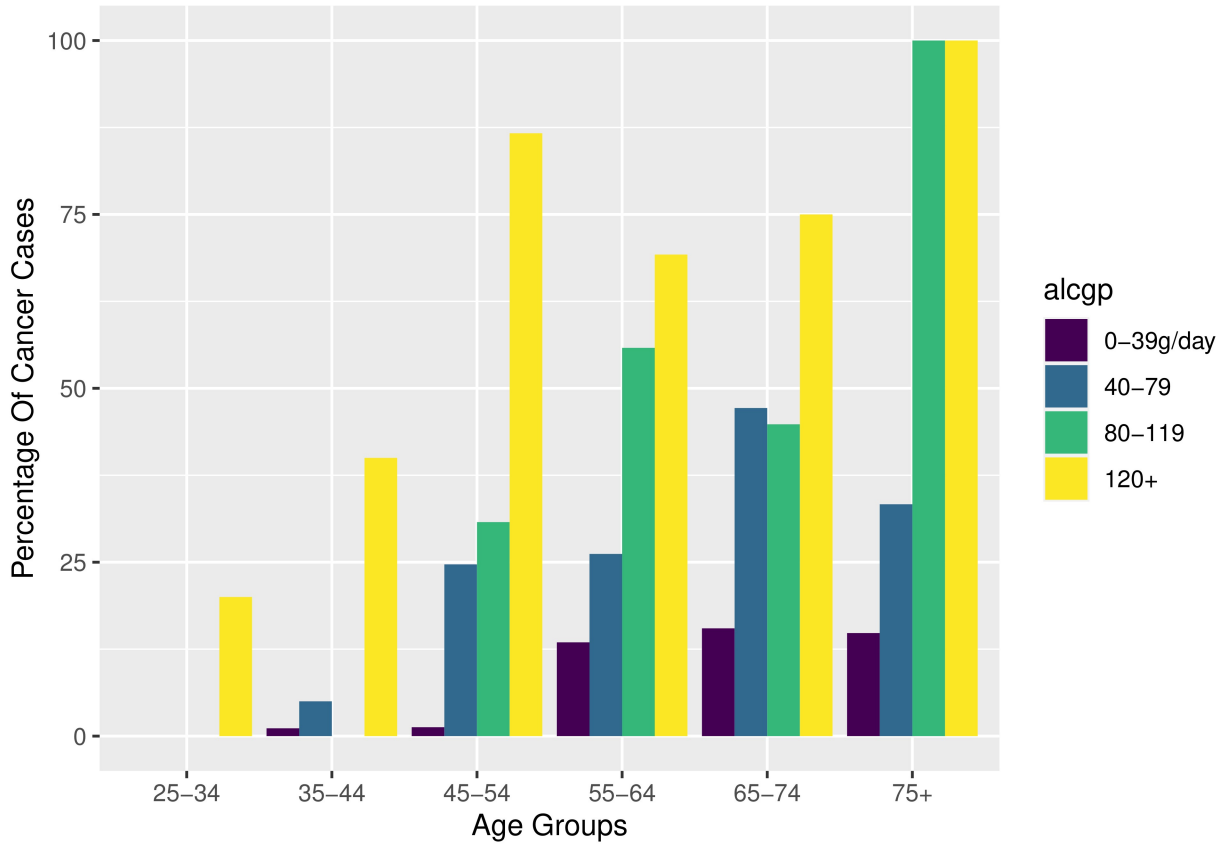
increase. As seen by the absence of some bars, there are some implicitly missing values, but the graph still demonstrates the generally increasing trend in percentage of cases.

Question 4

```
d3<- esoph %>%  group_by(agegp, alcgp) %>%
    summarise(count = n(), total_cases = sum(ncases), total_controls = sum(ncontrols),
              percentage_alc = total_cases*100/total_controls)
```

```
## `summarise()` has grouped output by 'agegp'. You can override using the `.groups` argument.
```

```
ggplot(d3, aes(agegp, percentage_alc, fill=alcgp)) +
   geom_bar(stat="identity", position = "dodge") +
   labs(x= 'Age Groups', y= 'Percentage Of Cancer Cases')
```



The general trend is that for each group as the alcohol consumption increases the percentage of cases also increase, but significantly more so from 80-119 to 120+ g/day. As seen by the absence of some bars, there are some implicitly missing values, but the graph still demonstrates the generally increasing trend in the percentage of cases.

3.2. What type of covariation occurs between the variables? (30+ points)

If you don't have variables of a certain type in the original dataset or among the created variables (features), you can further create them from the existing variables. See RDS chap. 5, 7.5 and 7.6.

3.2.1 Between a categorical and continuous variable (10 points) Question 1

- Describe what type of visualization you can use and why.

A bar chart or a box plot is the best visualization to show a relationship between a categorical and continuous variable. For the questions, I opted to use a bar chart since the variables given were counts rather than something like temperature or height, which would be more suitable for box plots. I also decided to use percentages to make more meaningful graphs since the number of observations for each group differed.

- Describe the patterns and relationships you observe. Could the identified patterns be due to coincidence (i.e. random chance)?

Each bar is clearly different than the others, so I don't think that this relationship could be due to coincidence.

- Describe the relationship implied by the pattern? (e.g., positive or negative correlation)

Since the age groups were ordered, there is a clear increase in the percentage of cases as age increased (positive correlation).

- Calculate the strength of the relationship implied by the pattern (e.g., correlation)

Finding correlation between a categorical variable and a continuous variable is generally tricky and has not really been covered in class, since the increase is so dramatic between the age groups, I believe that the positive correlation between age and percentage of cases is very strong.

- Discuss what other variables might affect the relationship.

One variable that could also affect this relationship is the person's predisposition to develop cancer based on their genetics.

- Does the relationship change if you look at individual subgroups of the data? Please discuss and demonstrate.

Questions 3 and 4 explore the subgroups of alcohol and tobacco consumption, and they show that for each age group the percentage of cases increases with the consumption of alcohol and tobacco.

- Demonstrate if converting the type of these variables help exploring the relationship.

Since the type of percentage is double, it can already be used to identify outliers or missing values, so it does not need to be converted.

- Discuss how the observed patterns support/reject your hypotheses or answer your questions.

Which age group(s) have the highest percentage of cases? The age groups of 65-74, 55-65, and 75+ had the first, second, and third highest percentage of cases respectively.

3.2.2 Between two categorical variables (10 points) Question 2

- Describe what type of visualization you can use and why.

Stacked bar charts or bar charts that use multiple bars for each category can be used to visualize and compare two categorical variables.

- Describe the patterns and relationships you observe. Could the identified patterns be due to coincidence (i.e. random chance)?

An interesting pattern is that for 0-79 g/day of alcohol consumed, the tobacco consumption has a clear difference in the percentage of cases with larger consumption corresponding to higher percentages, but between 80-120+ g/day of alcohol consumed there is practically no difference between the tobacco consumption factors. The increase in alcohol consumption most likely significantly increases the percentage of cases because large amounts of alcohol can damage your DNA and prevent your body from repairing it. This damaged molecular material can eventual lead to cancer.

- Describe the relationship implied by the pattern? (e.g., positive or negative correlation)

There seems to be a positive correlation between the two variables.

- Calculate the strength of the relationship implied by the pattern (e.g., correlation)

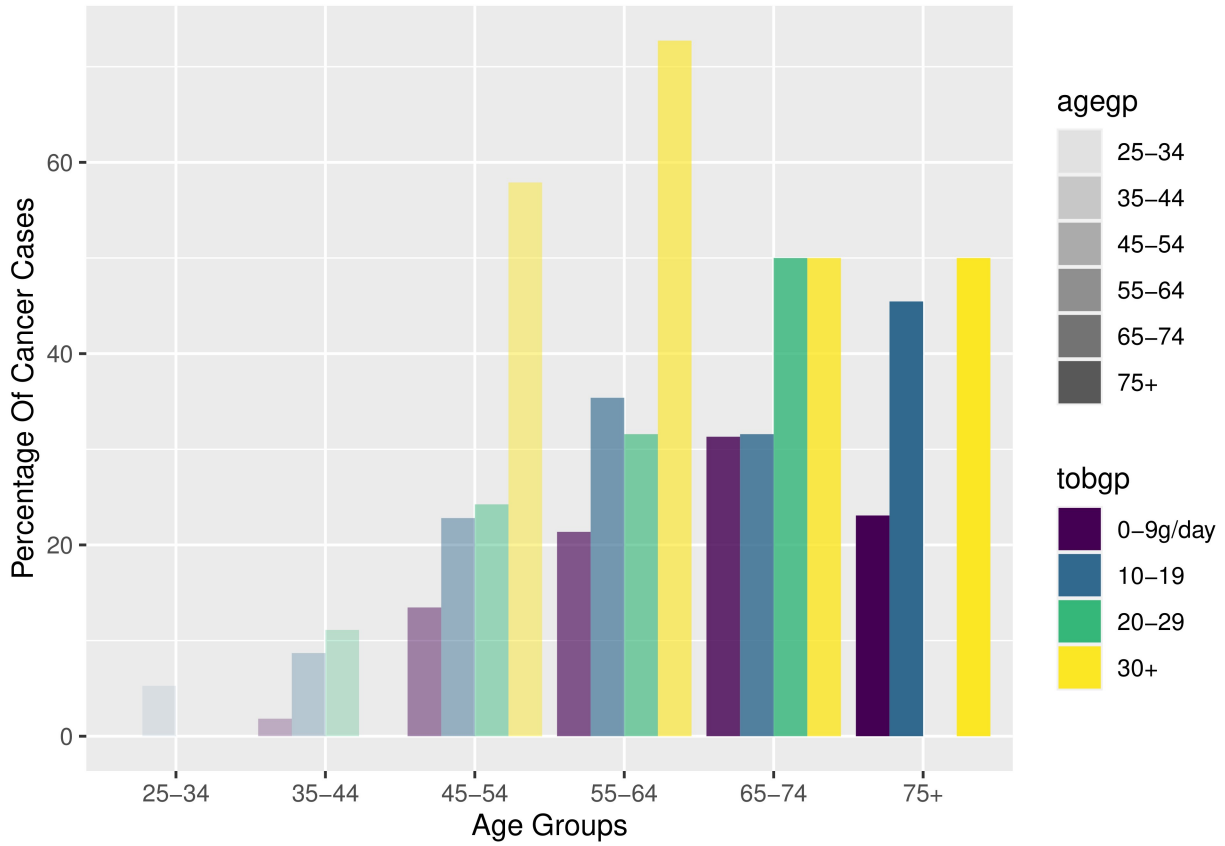
The height of the bars increase dramatically from left to right indicating a strong positive relationship.

- Discuss what other variables might affect the relationship.

The age group would be another interesting variable to take into account, but the graph is already heavily layered, so adding another layer may take away from the original findings. Another variable that is not available that could affect the relationship is the duration of consumption, since large consumption over a short period of time may be more harmful that smaller consumption over a longer period of time.

- Does the relationship change if you look at individual subgroups of the data? Please discuss and demonstrate.

```
ggplot(d2, aes(agegp, percentage_tob, fill = tobgp, alpha = agegp)) +
   geom_bar(stat="identity", position = "dodge") +
   labs(x= 'Age Groups', y= 'Percentage Of Cancer Cases')
```



The relationship stays the same with the addition of the subgroup agegp.

- Demonstrate if converting the type of these variables help exploring the relationship.

Since they are factors, you cannot convert these variables in a meaningful way without more data.

- Discuss how the observed patterns support/reject your hypotheses or answer your questions.

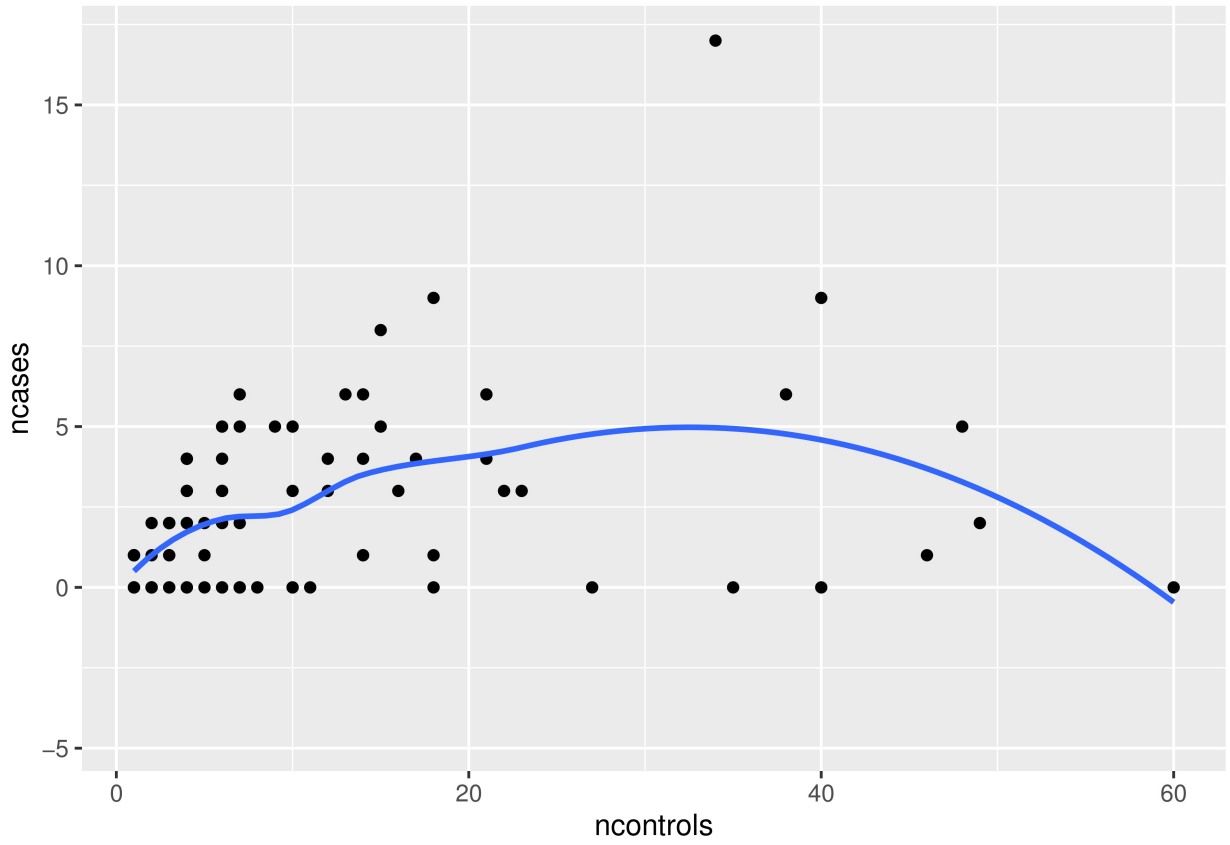
Does tobacco consumption or alcohol consumption have a greater affect on the percentage of cases? It shows that tobacco has a greater affect than alcohol for the first two alcohol consumption groups but makes little to no difference in the last two alcohol consumption groups, but I believe gathering more numerical observations rather than ranges would help to find the particular point where these two variables intersect and which has a greater affect.

3.2.3 Between two continuous variables (10 points)

Additional Graph for Two Continuous Variables Does increasing the number of controls also increase the number of cases reported?

```
ggplot(esoph, aes(x=ncontrols, y=ncases)) +
  geom_point() +
  geom_smooth(fill = NA)+
  labs(x= 'ncontrols', y= 'ncases')
```

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



- Describe what type of visualization you can use and why.

Scatter plots are one type of visualization you can use to graph two continuous variables because each variable column can be paired with another variable column to create a graph that could show correlation. The addition of fitted lines by a certain categorical variable can also allow you to see multiple sub trends within the original plot.

- Describe the patterns and relationships you observe. Could the identified patterns be due to coincidence (i.e. random chance)?

Between 0 and 30 ncontrols, the slope is upward, but it proceeds to dip back down, so the initial positive correlation could be due to coincidence.

- Describe the relationship implied by the pattern? (e.g., positive or negative correlation)

There does not seem a to be a relationship between ncases and ncontrols.

- Calculate the strength of the relationship implied by the pattern (e.g., correlation)

Since there does not seem to be a relationship, the strength of the correlation would be very small.

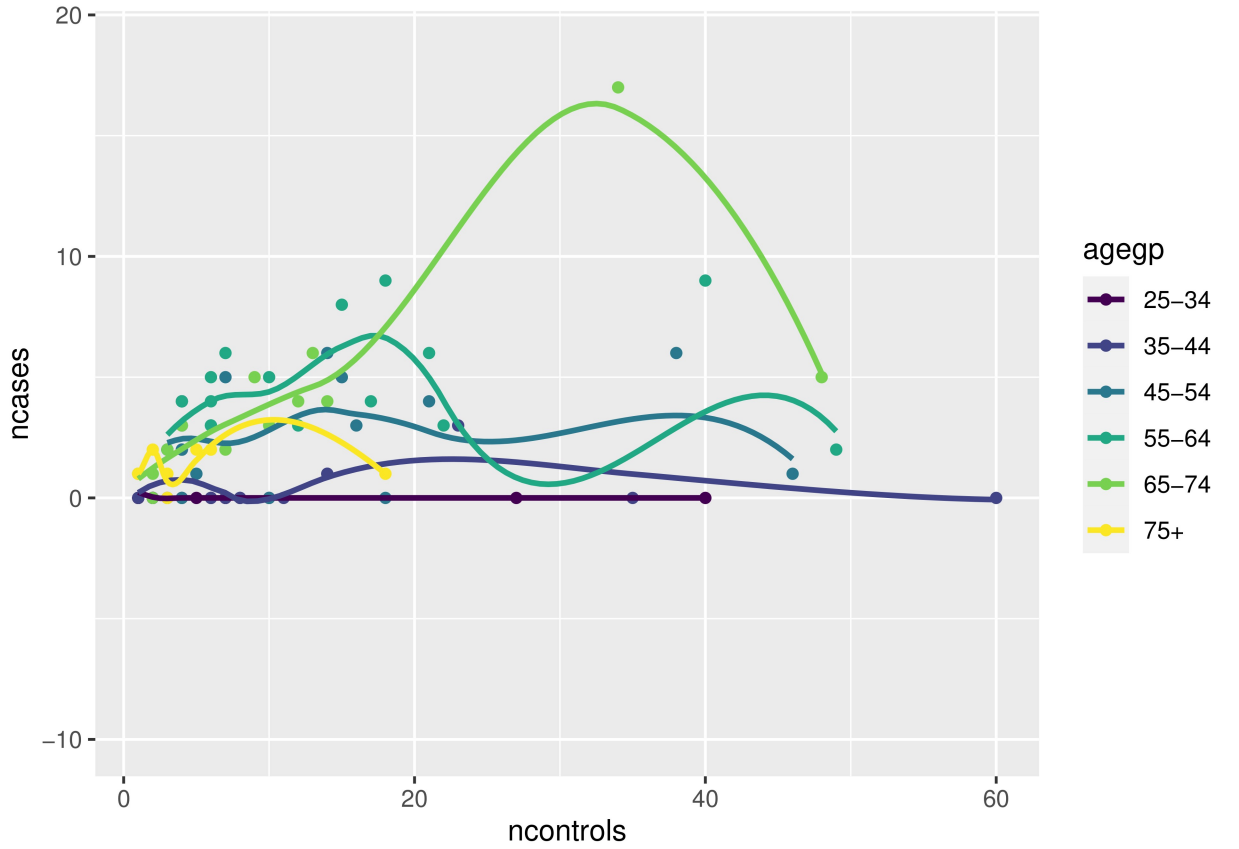
- Discuss what other variables might affect the relationship.

In general, increasing the number of observations could affect the relationship.

- Does the relationship change if you look at individual subgroups of the data? Please discuss and demonstrate.

```
options(warn=-1)
ggplot(esoph, aes(x=ncontrols, y=ncases,color=agegp)) +
  geom_point() +
  geom_smooth(fill = NA) +
  labs(x= 'ncontrols', y= 'ncases')
```

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



The 65-74 age group is generally above the other age groups which suggests that there are more cases for this age group. Since there aren't a lot data points and the other age groups are pretty close together this pattern could be due to coincidence. Most of the sub trends do not seem to have a positive or negative correlation except for the age group 65-74 which has a sharp upward slope for the first few data points but also has a sharp downward slope between the last two data points. The combination of these slopes in the age group 65-74, however, could also result in no net correlation.

- Demonstrate if converting the type of these variables help exploring the relationship.

Since the type of these variables is numeric, it can already be used to identify outliers or missing values, so it does not need to be converted.

- Discuss how the observed patterns support/reject your hypotheses or answer your questions.

The observed patterns do not support the intial thought that the number of controls increases the number of cases reported.

3.2.4 Between two other continuous/categorical variables (if any) (10 extra points)

- Please repeat the process above

Step 4. Summarize your findings (20 points)

- Summarize your findings about the questions you asked at the beginning. (5 points)

Question 1: Which age group(s) have the highest percentage of cases?

As age increases, the percentage of cases also increase most likely due to the introduction of damaging substances over time.

Question 2: Does tobacco consumption or alcohol consumption have a greater affect on the percentage of cases?

Alcohol consumption has a smaller affect on the percentage of cancer cases than tobacco for the first two alcohol consumption groups as seen by the difference of heights between the tobacco consumption bars. However, the effects of alcohol consumption catches up with tobacco consumption for the last two alcohol consumption groups since the percentage of cases did not significantly change depending on the amount of tobacco consumed.

- Describe and discuss how your observations support or reject your hypotheses or answer your questions. (5 points)

Since relationships were found in both graphs made for each question, I believe that my observations do answer my questions.

- Describe what new questions your analysis may generate. (5 points)

To find a greater correlation, more specific data should be collected about age, alcohol consumption, and tobacco consumption to prove and calculate the positive correlation between these factors and the percentage of cases. At one point does the affect of alcohol consumption overtake that of tobacco consumption?

- Discuss if you have enough evidence to make a conclusion about your analysis. (5 points)

I believe that the stark difference in the heights of the bars are enough to conclude there is a positive correlation between all of the groups. However, the correlation could be better calculated if we had actual observations for age, alcohol consumption, and tobacco consumption rather than groups of ranges.