Project 1

Drunk Driving Deaths and SeatBelt Safety

1. Introduction

Set up

First load the packages you will use throughout the document. *Note: you can hide this step in the resulting output with the option* `echo=FALSE` .

Description of the datasets

The first dataset is called “US Seat Belts” that explores the relationship of traffic fatalities depending on the seat belt usage in different states over the years. This dataset interested me because sometimes, I honestly forget to wear my seat belt when driving until my car alerts me. I wondered what the effects of not wearing seat belts were in a greater scale. Each row represents each states from years 1983-1997. They also contain 12 different variables that include 6 numeric and 6 categorical variables. It is obtained from R Studio datasets: https://vincentarelbundock.github.io/Rdatasets/doc/AER/USSeatBelts.html (https://vincentarelbundock.github.io/Rdatasets/doc/AER/USSeatBelts.html).

The second dataset is called “Fatality” that explores how different drunk driving laws across the different states over the years has an effect of traffic fatality rate. It is obtained from R Studio datasets: https://vincentarelbundock.github.io/Rdatasets/doc/Ecdat/Fatality.html (https://vincentarelbundock.github.io/Rdatasets/doc/Ecdat/Fatality.html). This dataset particularly interested me since I wondered whether or not having a strict or loose alcohol laws had an effect on people dying in traffic due to drunk driving, especially since drinking is frequent in universities like UT. Each row also represents each states but from the years 1982 to 1988. It contains 10 variables that include 8 numeric and 2 categorical variables. This dataset was supported by another dataset to identify state ID code, which was obtained from a dataset called “stateID” : https://gist.github.com/dantonnoriega/bf1acd2290e15b91e6710b6fd3be0a53 (https://gist.github.com/dantonnoriega/bf1acd2290e15b91e6710b6fd3be0a53).

Based on these two datasets, I can join them with state and year, since they are the two variables that are common and present in both datasets. Moreover, by joining the two datasets with states and year, we can potentially explore the relationship between seatbelt usage from “US Seat Belts” and jail sentences or death due to drunk driving from “Fatality”. We can also explore possible relationship of drunk driving death in correlation to different income and ages.

```
# US Seat Belt Dataset
seatbelt <- read_csv('USSeatBelts.csv')
  # Examining types of variables
head(seatbelt)
```

```
## # A tibble: 6 × 13
##    ...1 state  year miles fatal…¹ seatb…² speed65 speed70 drink…³ alcohol income
##   <dbl> <chr> <dbl> <dbl>   <dbl>   <dbl> <chr>   <chr>   <chr>   <chr>    <dbl>
## 1     1 AK     1983  3358  0.0447      NA no      no      yes     no       17973
## 2     2 AK     1984  3589  0.0373      NA no      no      yes     no       18093
## 3     3 AK     1985  3840  0.0331      NA no      no      yes     no       18925
## 4     4 AK     1986  4008  0.0252      NA no      no      yes     no       18466
## 5     5 AK     1987  3900  0.0195      NA no      no      yes     no       18021
## 6     6 AK     1988  3841  0.0253      NA no      no      yes     no       18447
## # … with 2 more variables: age <dbl>, enforce <chr>, and abbreviated variable
## #   names ¹fatalities, ²seatbelt, ³drinkage
```

```
# Fatality Dataset
drunkdriving <- read_csv('Fatality.csv')
  # Examining types of variables
head(drunkdriving)
```

```
## # A tibble: 6 × 11
##    ...1 state  year mrall beertax  mlda jaild comserd vmiles unrate perinc
##   <dbl> <dbl> <dbl> <dbl>   <dbl> <dbl> <chr> <chr>    <dbl>  <dbl>  <dbl>
## 1     1     1  1982  2.13    1.54  19   no    no        7.23  14.4  10544.
## 2     2     1  1983  2.35    1.79  19   no    no        7.84  13.7  10733.
## 3     3     1  1984  2.34    1.71  19   no    no        8.26  11.1  11109.
## 4     4     1  1985  2.19    1.65  19.7 no    no        8.73   8.90 11333.
## 5     5     1  1986  2.67    1.61  21   no    no        8.95   9.80 11662.
## 6     6     1  1987  2.72    1.56  21   no    no        9.17   7.80 11944
```

```
# Supporting Dataset for "Fatality"
stateID <- read_csv('us-state-ansi-fips.csv')

stateID <- stateID %>%
  mutate(st = as.numeric(st), stusps = as.character(stusps)) %>%
  select(st,stusps) %>%
  rename(state=st)

head(stateID)
```

```
## # A tibble: 6 × 2
##   state stusps
##   <dbl> <chr>
## 1     1 AL
## 2     2 AK
## 3     4 AZ
## 4     5 AR
## 5     6 CA
## 6     8 CO
```

Research Question

Based on the two datasets that were chosen, one thing that will be explored is: “Is there a relationship between seatbelt usage and drunk driving deaths across the states in US?”

2. Tidying

The two datasets that were chosen are tidy as “every observation has its own row and every variable its own column”.

```
# Displaying that "seatbelt" is tidy
head(seatbelt)
```

```
## # A tibble: 6 × 13
##    ...1 state  year miles fatal…¹ seatb…² speed65 speed70 drink…³ alcohol income
##   <dbl> <chr> <dbl> <dbl>   <dbl>   <dbl> <chr>   <chr>   <chr>   <chr>    <dbl>
## 1     1 AK     1983  3358  0.0447      NA no      no      yes     no       17973
## 2     2 AK     1984  3589  0.0373      NA no      no      yes     no       18093
## 3     3 AK     1985  3840  0.0331      NA no      no      yes     no       18925
## 4     4 AK     1986  4008  0.0252      NA no      no      yes     no       18466
## 5     5 AK     1987  3900  0.0195      NA no      no      yes     no       18021
## 6     6 AK     1988  3841  0.0253      NA no      no      yes     no       18447
## # … with 2 more variables: age <dbl>, enforce <chr>, and abbreviated variable
## #   names ¹fatalities, ²seatbelt, ³drinkage
```

```
# Displaying that "drunkdriving" is tidy
head(drunkdriving)
```

```
## # A tibble: 6 × 11
##    ...1 state  year mrall beertax  mlda jaild comserd vmiles unrate perinc
##   <dbl> <dbl> <dbl> <dbl>   <dbl> <dbl> <chr> <chr>    <dbl>  <dbl>  <dbl>
## 1     1     1  1982  2.13    1.54  19   no    no        7.23  14.4  10544.
## 2     2     1  1983  2.35    1.79  19   no    no        7.84  13.7  10733.
## 3     3     1  1984  2.34    1.71  19   no    no        8.26  11.1  11109.
## 4     4     1  1985  2.19    1.65  19.7 no    no        8.73   8.90 11333.
## 5     5     1  1986  2.67    1.61  21   no    no        8.95   9.80 11662.
## 6     6     1  1987  2.72    1.56  21   no    no        9.17   7.80 11944
```

3. Joining/Merging

```
# ---BEFORE Joining---

# Total Observations in each dataset before joining
nrow(seatbelt)
```

```
## [1] 765
```

```
nrow(drunkdriving)
```

```
## [1] 336
```

```
# IDs that appear in one dataset but not the other
setdiff(names(seatbelt), names(drunkdriving))
```

```
##  [1] "miles"      "fatalities" "seatbelt"   "speed65"    "speed70"
##  [6] "drinkage"   "alcohol"    "income"     "age"        "enforce"
```

```
setdiff(names(drunkdriving), names(seatbelt))
```

```
## [1] "mrall"   "beertax" "mlda"    "jaild"   "comserd" "vmiles"  "unrate"
## [8] "perinc"
```

```
# IDs in common
intersect(names(drunkdriving), names(seatbelt))
```

```
## [1] "...1"  "state" "year"
```

```
# ---JOINING---


# Combining supporting dataset "stateID" to "drunkdriving" dataset
new_drunkdriving <- drunkdriving %>%
  left_join(stateID, by='state')

# Renaming state abbreviation to state in order to join with "seatbelt"
new_drunkdriving <- new_drunkdriving %>%
  select(!state) %>%
  rename(state = stusps)

head(new_drunkdriving)
```

```
## # A tibble: 6 × 11
##    ...1  year mrall beertax  mlda jaild comserd vmiles unrate perinc state
##   <dbl> <dbl> <dbl>   <dbl> <dbl> <chr> <chr>    <dbl>  <dbl>  <dbl> <chr>
## 1     1  1982  2.13    1.54  19   no    no        7.23  14.4  10544. AL
## 2     2  1983  2.35    1.79  19   no    no        7.84  13.7  10733. AL
## 3     3  1984  2.34    1.71  19   no    no        8.26  11.1  11109. AL
## 4     4  1985  2.19    1.65  19.7 no    no        8.73   8.90 11333. AL
## 5     5  1986  2.67    1.61  21   no    no        8.95   9.80 11662. AL
## 6     6  1987  2.72    1.56  21   no    no        9.17   7.80 11944  AL
```

```
# Joining the "new_drunkdriving" and "seatbelt" by state and year
joint_dataset <- seatbelt %>%
  left_join(new_drunkdriving, by= c('year', 'state')) %>%
  na.omit(seatbelt) # Omitting N/A values for variable seatbelt

# IDs that may have been left out after joining
setdiff(names(drunkdriving), names(joint_dataset))
```

```
## [1] "...1"
```

```
setdiff(names(seatbelt), names(joint_dataset))
```

```
## [1] "...1"
```

```
# Examining change in observations/rows
nrow(joint_dataset)
```

```
## [1] 114
```

Before joining, dataset “seatbelt” (US Seat Belt Dataset) has 765 observations and dataset “drunkdriving” (Fatality Dataset) has 336 observations.

The dataset “seatbelt” has following IDs that are not in “drunkdriving”: miles, fatalities, seatbelt, speed65, speed70, drinkage, alcohol, income, age, enforce.

The dataset “drunkdriving” has following IDs that are not in “seatbelt”: mrall, beertax, mlda, jaild, comserd, vmiles, unrate, and perinc.

The IDs that are in common are “state” and “year”.

No IDs were left out after joining

After joining, it can be seen that there are only 114 observations in the joint dataset. This can be because I omitted the data that had NA values for the variable “seatbelt” since I am trying to explore the relationship of seatbelt and drunk driving death, so not having a value for seatbelt wouldn’t make sense.

4. Wrangling

```
# Creating a categorical variable "safety" based on a numeric variable "seatbelt" by using
mutate
joint_dataset <- joint_dataset %>%
  mutate(safety = ifelse(seatbelt<0.5, 'Unsafe', 'Safe'))

# Summary table displaying state, comparing seatbelt safety and drunk driving death
joint_dataset %>%
  select(state, seatbelt, mrall) %>%
  group_by(state) %>%
  summarize(mean_seatbelt_safety = mean(seatbelt), mean_mrall = mean(mrall)) %>%
  arrange(desc(mean_seatbelt_safety))
```

```
## # A tibble: 31 × 3
##    state mean_seatbelt_safety mean_mrall
##    <chr>                <dbl>      <dbl>
##  1 NC                   0.522       2.47
##  2 NY                   0.477       1.21
##  3 OR                   0.470       2.45
##  4 WA                   0.468       1.66
##  5 MD                   0.465       1.73
##  6 IA                   0.430       1.72
##  7 CA                   0.425       1.93
##  8 WI                   0.411       1.66
##  9 FL                   0.398       2.44
## 10 VA                   0.394       1.79
## # … with 21 more rows
```

```
# Summary table, displaying drunk driving death across the years for those whose seatbelt safety is considered unsafe
joint_dataset %>%
  filter(safety=='Unsafe') %>%
  group_by(year) %>%
  summarize(mean_seatbelt_safety = mean(seatbelt), mean_mrall = mean(mrall)) %>%
  arrange(mean_mrall)
```

```
## # A tibble: 6 × 3
##    year mean_seatbelt_safety mean_mrall
##   <dbl>                <dbl>      <dbl>
## 1  1983                0.100       1.72
## 2  1984                0.150       1.83
## 3  1985                0.218       1.91
## 4  1987                0.332       1.94
## 5  1986                0.313       1.95
## 6  1988                0.379       2.06
```

```
# Another way of looking at data and having all states in columns.
joint_dataset %>%
  pivot_wider(names_from = state,
              values_from = seatbelt) %>%
  select(-c('miles','fatalities','speed65','speed70','drinkage','alcohol','income','age','enforce','beertax','mlda','jaild','comserd','vmiles','unrate','perinc')) #Exculding unwanted variables
```

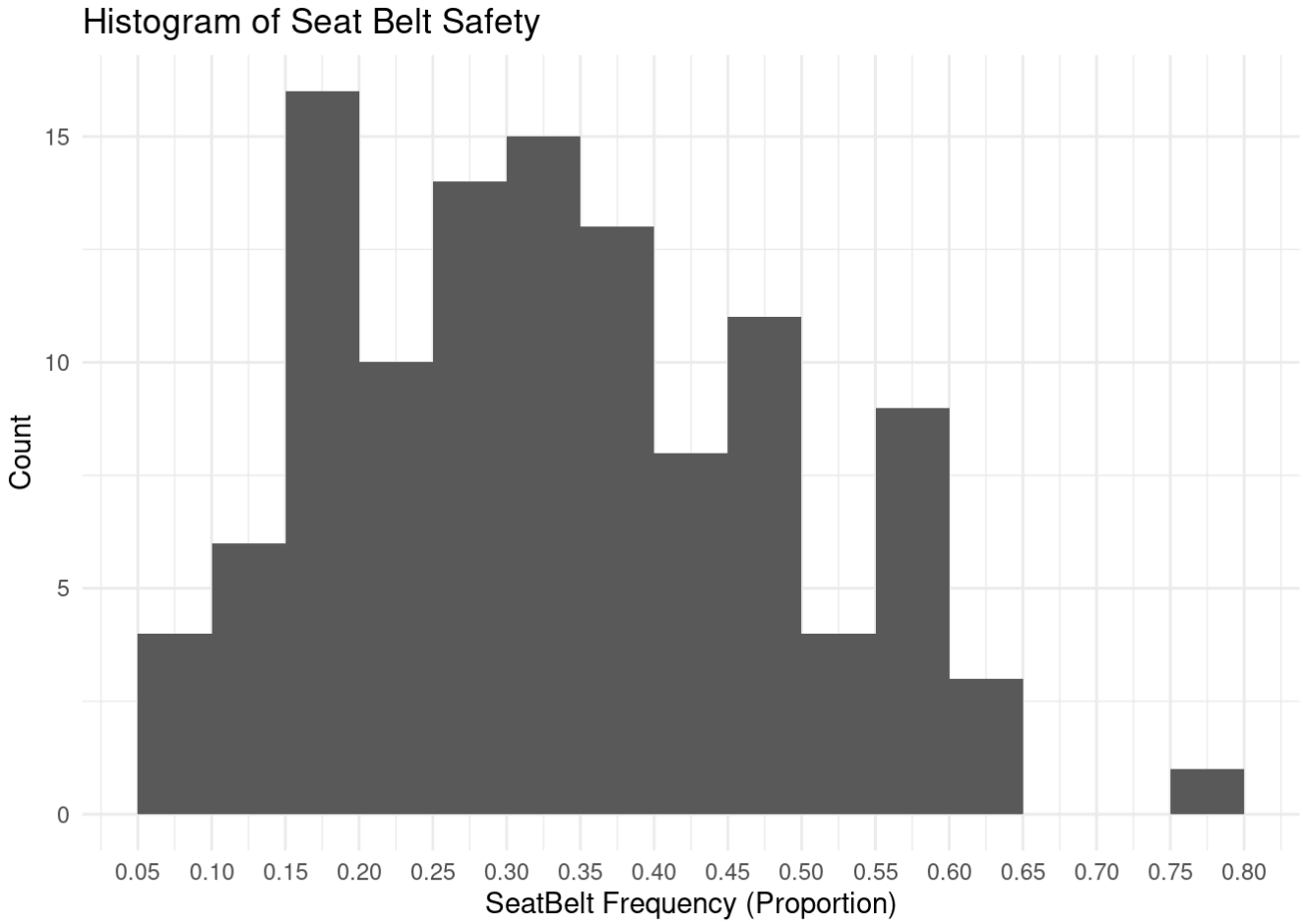
```
## # A tibble: 114 × 36
##    ...1.x  year ...1.y mrall safety     AL     AR     CA    FL    GA    IA    ID
##     <dbl> <dbl>  <dbl> <dbl> <chr>   <dbl>  <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>
##  1     17  1984      3  2.34 Unsafe  0.130 NA     NA        NA    NA    NA    NA
##  2     18  1985      4  2.19 Unsafe  0.170 NA     NA        NA    NA    NA    NA
##  3     19  1986      5  2.67 Unsafe  0.290 NA     NA        NA    NA    NA    NA
##  4     20  1987      6  2.72 Unsafe  0.210 NA     NA        NA    NA    NA    NA
##  5     21  1988      7  2.49 Unsafe  0.290 NA     NA        NA    NA    NA    NA
##  6     34  1986     19  2.54 Unsafe NA      0.198 NA        NA    NA    NA    NA
##  7     36  1988     21  2.55 Unsafe NA      0.301 NA        NA    NA    NA    NA
##  8     63  1985     25  1.88 Unsafe NA     NA      0.258    NA    NA    NA    NA
##  9     64  1986     26  1.95 Unsafe NA     NA      0.454    NA    NA    NA    NA
## 10     65  1987     27  1.99 Unsafe NA     NA      0.481    NA    NA    NA    NA
## # … with 104 more rows, and 24 more variables: IL <dbl>, IN <dbl>, KS <dbl>,
## #   KY <dbl>, LA <dbl>, MA <dbl>, MD <dbl>, MI <dbl>, MN <dbl>, MT <dbl>,
## #   NC <dbl>, NE <dbl>, NH <dbl>, NJ <dbl>, NV <dbl>, NY <dbl>, OH <dbl>,
## #   OK <dbl>, OR <dbl>, SC <dbl>, VA <dbl>, VT <dbl>, WA <dbl>, WI <dbl>
```

From the first summary table that examines seatbelt safety and drunk driving death across the states, it can be seen that South Carolina (SC), at the very bottom of the table, has a low seatbelt frequency of 0.128 or 12.8% and has a high drunk driving death cases of about 2.841 deaths every 10,000 people per year. Also, North Carolina (NC) at the very top of the table has seatbelt frequency of 52.3% and has less drunk driving death of aroud 2.466 deaths every 10,000 people per year. However, an interesting finding comes from the

second table where it displays seatbelt safety and drunk driving death across the years 1983-1988 only for unsafe seatbelt frequencies. It seems that as the years progress, seatbelt usage gets higher, but drunk driving death increases, but it is hard to tell because we are not looking at the data as a whole, and there are many other confounding variables that are not represented.

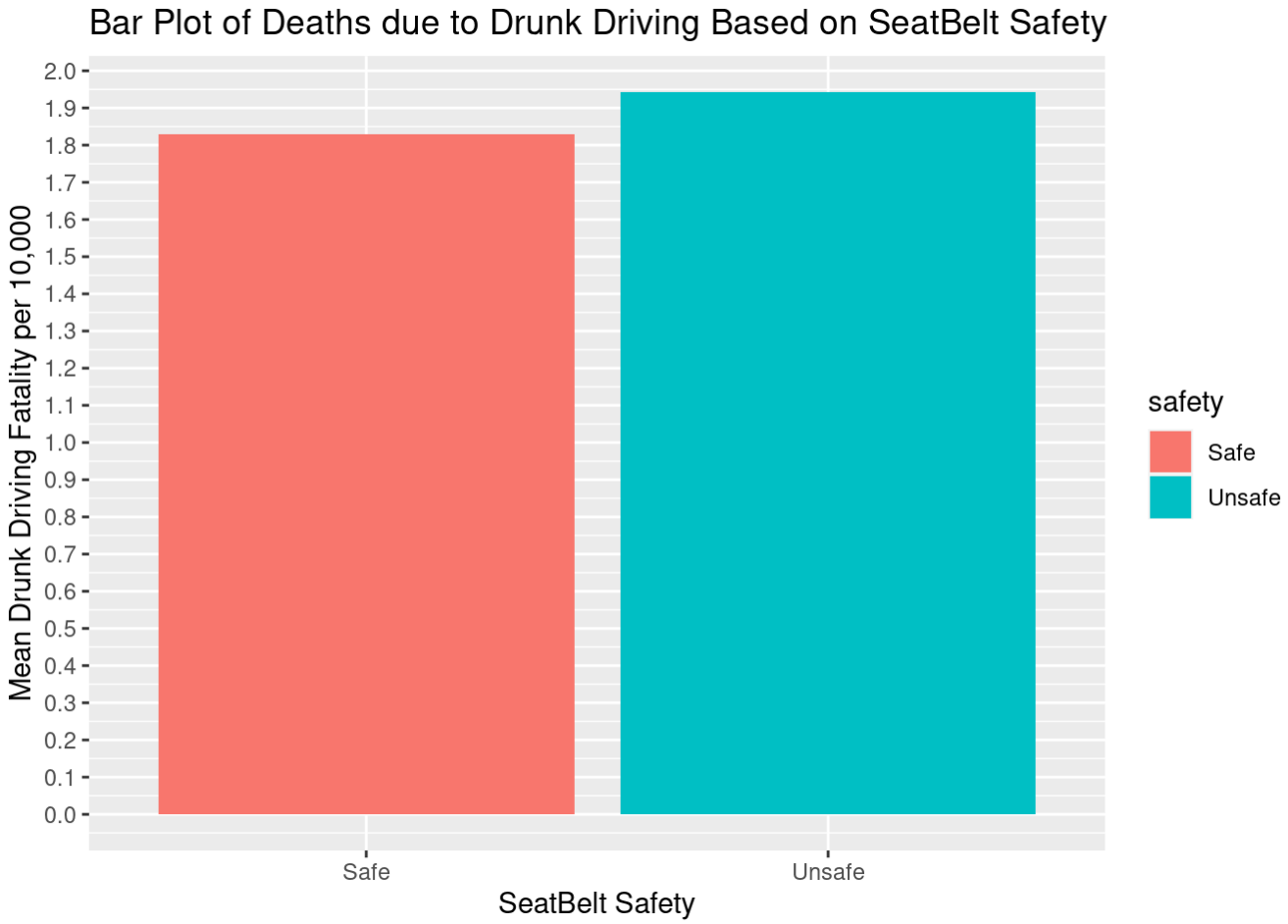
5. Visualize and analyze our data

```
# Histogram of seatbelt safety (1 variable)
joint_dataset %>%
  ggplot(aes(x=seatbelt)) +
  geom_histogram(binwidth = 0.05, center=0.025) +
  scale_x_continuous(breaks = seq(0,1,0.05)) + # Changing scale on x-axis
  scale_y_continuous(breaks = seq(0,30,5)) + # Changing scale on y-axis
  theme_minimal() + # Changing theme
  labs(title = 'Histogram of Seat Belt Safety',
       x='SeatBelt Frequency (Proportion)',
       y='Count')
```



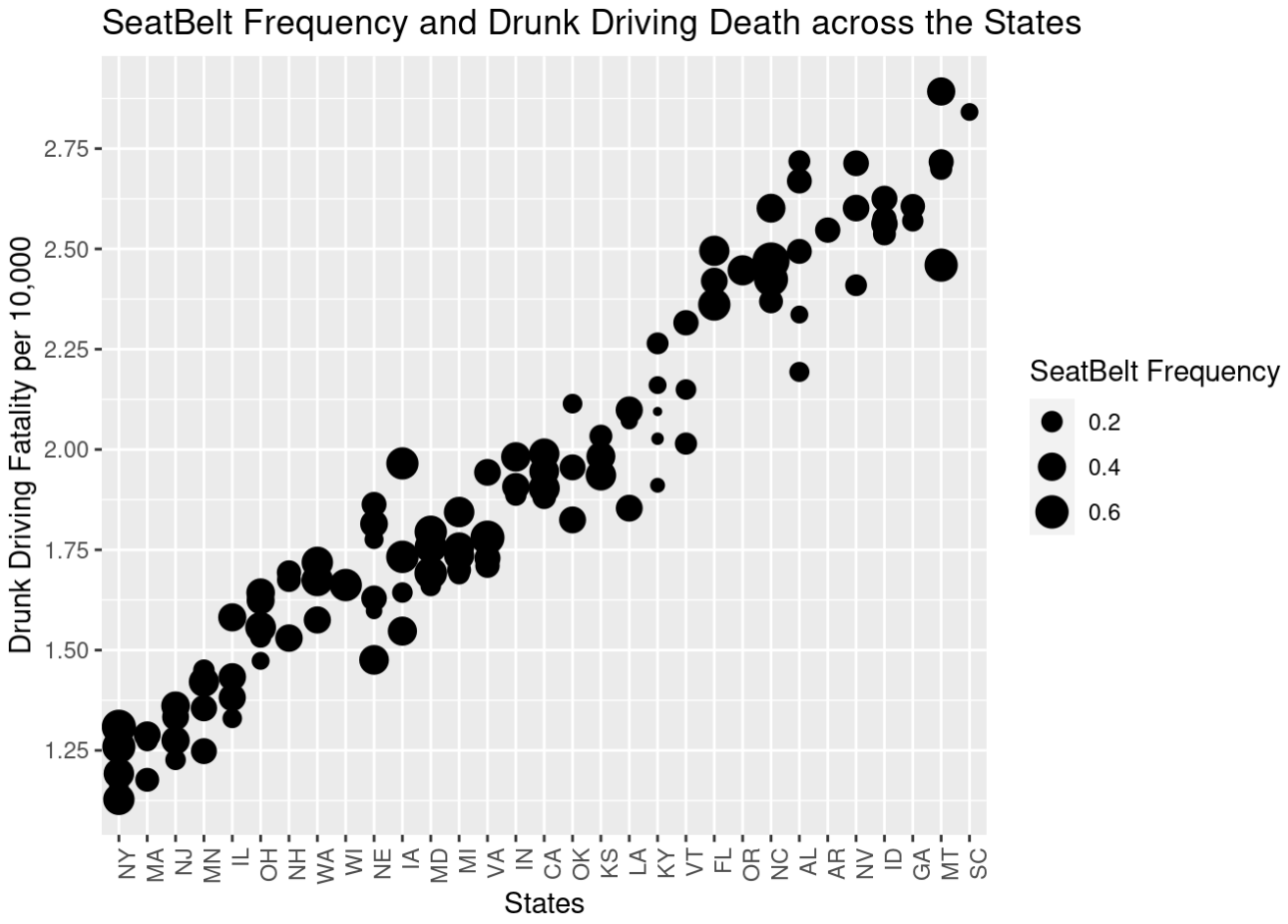
This histogram displays the count of seatbelt frequency in proportion in the joint dataset. It can be seen that seatbelt frequency of 0.15 (15%) was the more frequent in my dataset, and displays a right-skewed distribution. Moreover, from the histogram, we can see that there are more data in seatbelt frequency from 0.15 to 0.35.

```
# Relationship of seatbelt and drunk driving death (2 variables)
joint_dataset %>%
  ggplot(aes(x=safety, y=mrall, fill=safety)) +
  geom_bar(stat='summary') + # Using stat summary
  labs(title = 'Bar Plot of Deaths due to Drunk Driving Based on SeatBelt Safety',
       x='SeatBelt Safety',
       y='Mean Drunk Driving Fatality per 10,000') + # Adding labels
  theme_grey() + # Changing Theme
  scale_y_continuous(breaks = seq(0,2,0.1)) # Changing scale on y-axis
```



The bar plot above observes the mean drunk driving death based on seatbelt safety. It can be noted that the mean drunk driving deaths are very similar, but for seatbelt safety that is considered unsafe, the mean drunk driving death is slightly higher at about 1.95 deaths every 10,000 people than when seatbelt safety is considered safe which is about 1.8 deaths every 10,000 people.

```
# Relationship between seatbelt frequency, drunkdriving death, and states(3 variables)
joint_dataset %>%
  ggplot(aes(x=reorder(state, mrall), y=mrall)) + #Readjusting the order from least to gre
atest depending on death
  geom_point(aes(size=seatbelt)) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  scale_y_continuous(breaks = seq(0,3,0.25)) + # Chaning the scales in y-axis
  labs(title='SeatBelt Frequency and Drunk Driving Death across the States',
      x = 'States',
      y='Drunk Driving Fatality per 10,000',
      size = 'SeatBelt Frequency')
```



The three variable scatter plot notes the relationship of SeatBelt Frequency and Drunk Driving Death across the States with states as x-axis, drunk driving death as y-axis, and seatbelt frequency as size of the dots. Through the display, we can note that New York (NY) has the lowest drunk driving fatality while South Carolina (SC) has the highest. Moreover, in the extreme sides of both left and right, the size of the dot on the left seem a lot bigger, depicting that people do seatbelt more frequently when drunk driving fatalities are lower, but in the right side where the dots seem a little smaller, we can find that seatbelt frequency is a lot lower when drunk driving fatalities are higher.

6. Discussion

From these reports, it can be concluded that there is a very slight correlation between seatbelt safety and drunk driving deaths across the states. As can be seen in the “SeatBelt Frequency and Drunk Driving Death across the States” plot, as bigger dots, representing safer seatbelt frequency, are correlated with less

deaths due to drunk driving and smaller dots, representing more unsafe seatbelt frequency, are correlated with higher deaths due to drunk driving. Moreover, as the first summary table represents, it can be seen that South Carolina (SC) has the lowest seatbelt safety frequency as well as highest drunk driving death, but North Carolina (NC) is quite the opposite with highest seatbelt safety frequency with a little lower drunk driving death as well. But New York (NY) can be seen with the second highest seatbelt safety frequency with the lowest drunk driving death, showing how even if states change, the slight correlation between seatbelt safety and drunk driving death remains.

The most challenging part about conducting this process was playing around with data to display the statistics of my interest as well as trying to hit all the requirements needed for this project. I felt that some conclusions generated can't be fully explained due to limits we have on this project because there are just so many other variables that play a factor in a certain result. Couple numeric and a few categorical variables seems like it didn't cut it for the result of my desire.

7. Formatting

Remember to knit your file and produce a pdf with multiple pages to upload on Gradescope. It is ok to not follow the structure of the project in order (you might need to join before you tidy, or wrangle before you join, etc.). Just make sure to identify the page in which you apply the tidying functions, joining functions, wrangling functions, …