ANALYSIS OF FATALITIES IN THE UNITED STATES

PROJECT REPORT SUBMITTED
IN FULFILMENT OF THE REQUIREMENTS FOR THE COURSE
STAT 250 – APPLIED STATISTICS

DEPARTMENT OF STATISTICS OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ECEHAN VERGİLİEL – 2429397

MEHMET EGEMEN GÜNDÜR – 2429090

BATUHAN SAYLAM – 2429264

BİRKAN ÖZYURT – 2290880

June 2022

ABSTRACT

The research report is mainly focused on car accident fatalities in all states of the U.S. between 1982 and 1988. Intending to study fatalities, we used "Fatalities" data. The data contains states, years, spirit consumption, unemployment rate, the employment-population ratio, income, beer tax, number of Baptist and Mormon, drink age, dry states, young driver rate, average miles per driver, breath test, jail and service sentence, fatality types and fatalities according to age, population, and population of age groups, total vehicle miles, U.S. unemployment rate, U.S. employment/population ratio, GSP rate of change.

We find solutions for our research questions which highlight the significance of the project's purpose: the effects on car accident fatalities and the causes of car accident fatalities, using statistical methods in R. The effects on car accidents are analyzed using income, beer tax, breath, jail, service, fatal, afatal columns, and the causes of car accidents are analyzed using drinkage, fatal1517, fatal1820, fatal2024 columns. The project is concluded with some significant causes and effects we analyzed.

1. Introduction

We made statistical inferences and interpretations of the "Fatalities" data. Since no N.A. values exist, the data is used without any arrangements. We apply the given information about drinking age, beer taxes, type of sentences, fatalities, fatalities of different age groups, population, and population of different age groups in every state of the U.S in our analysis.

We focused on per capita personal income, beer tax, alcohol-involved accidents, and jail or service sentences while examining the significance of the effects we chose with the multiple regression model and two-way ANOVA methods. In addition, fatalities in different age groups and populations in different age groups to obtain fatality rate, drinking age, and breath test variables are the main objectives in the course of causes of car accident fatalities analysis with the methods of hypothesis testing with one-sample and two-sample to make inferences about mean and proportions, simple linear regression model, one-way ANOVA and multiple comparison such as Tukey's test.

To conclude, we completed our research with statistical methods' results and plots of some outputs and drew conclusions accordingly.

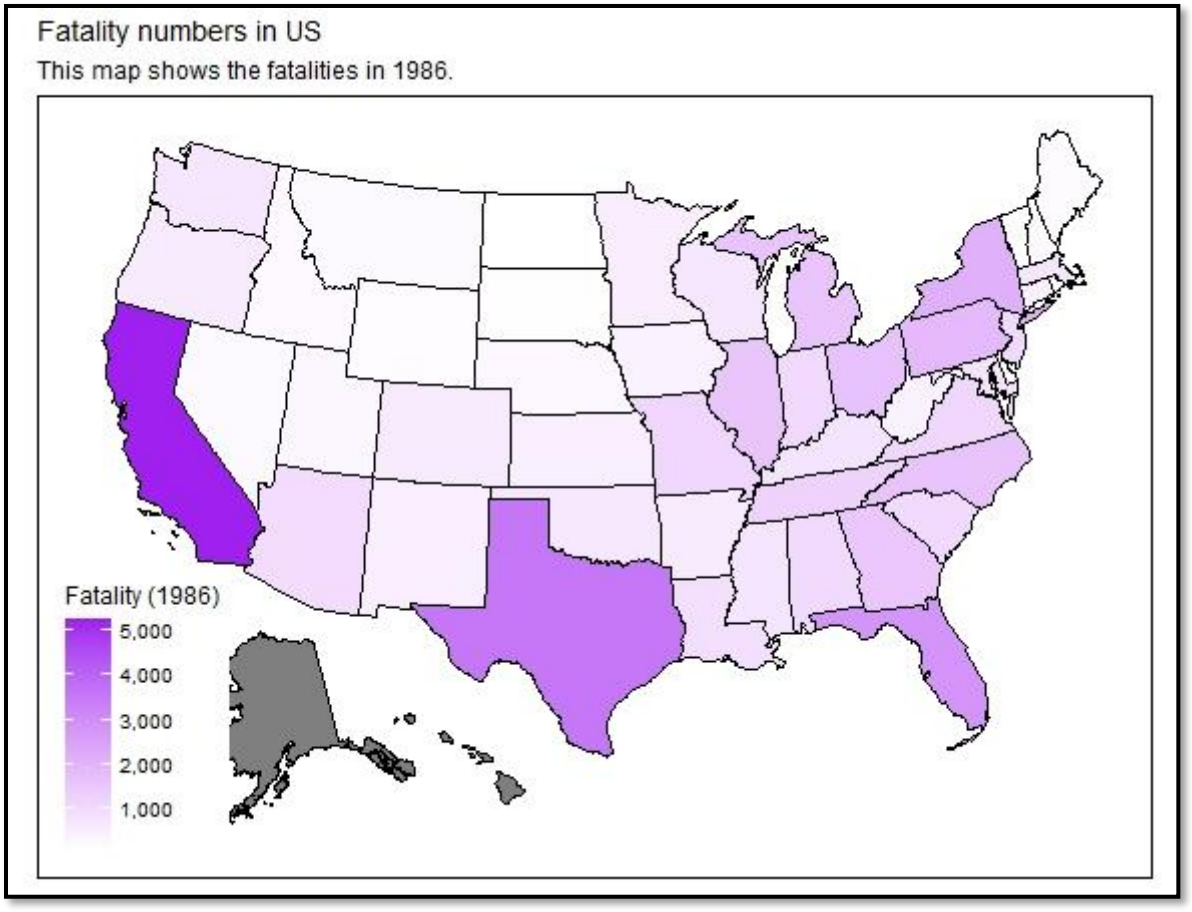
1.1.Data description

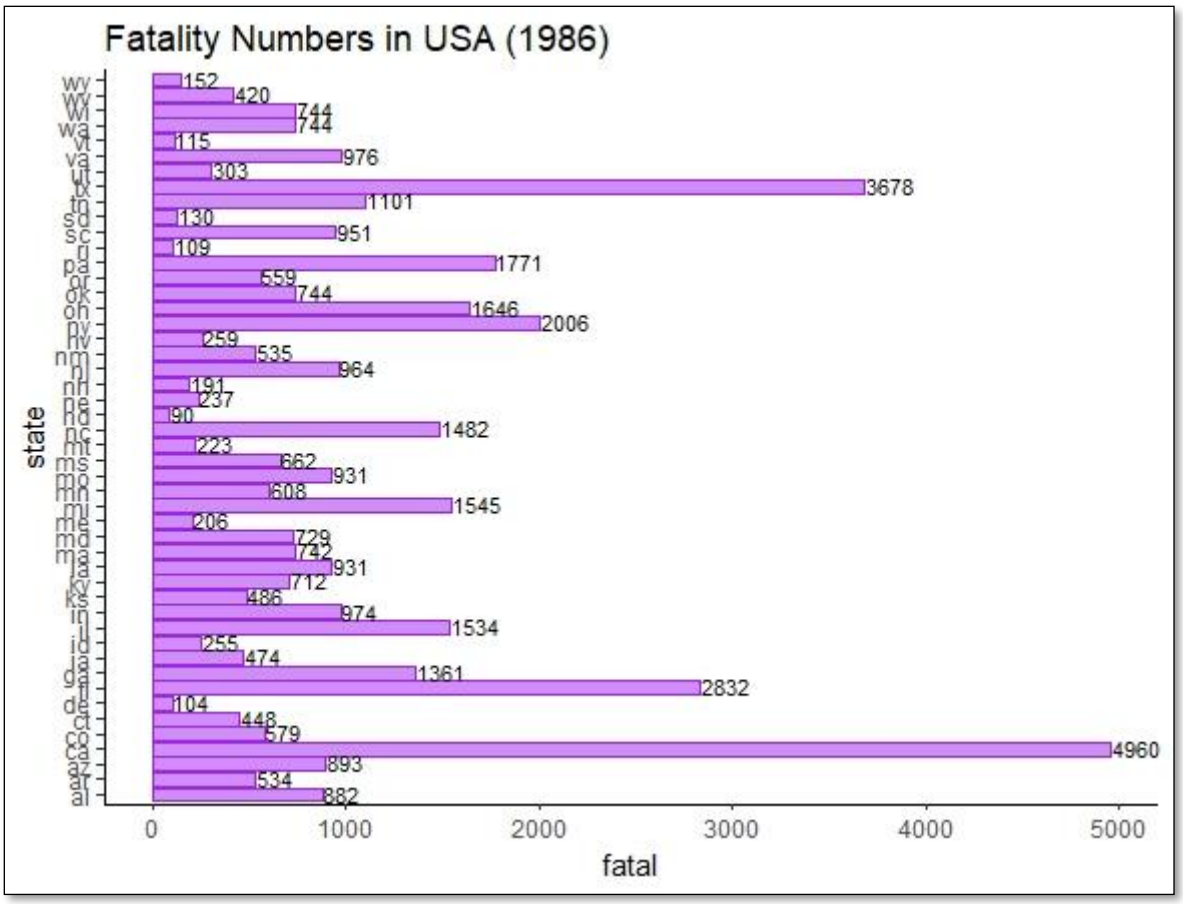
Dataset is obtained from the U.S. Department of Transportation Fatal Accident Reporting System. The dataset includes 335 observations and 35 variables; 4 of them are categorical, 21 of them are discrete, and 10 of them are continuous. The columns describe the factors or statistics that had an impact on accidents. Spirits consumption per person is measured in liters of pure alcohol per year. Beer tax is the tax determined according to the states' alcohol tax on one beer. The drinking age represents the minimum legal drink age for a given state. Jail and service are minimum sentencing for an initial drunk driving conviction for a given state. Total vehicle miles are traveled for a year in each state, which was obtained from the Department of Transportation. The U.S. Bureau of Economic Analysis provided data on personal income, and the U.S. Bureau of Labor Statistics provided data on the unemployment rate. The variables are explained in the table below.

Variable	**Description**	**Variable**	**Description**
state	U.S. state	nfatal	Number of night-time vehicle fatalities
year	year	sfatal	Number of single-vehicle fatalities
spirits	Spirits consumption	fatal1517	Number of vehicle fatalities, 15–17-year-olds
unemp	Unemployment rate	nfatal1517	Number of night-time vehicle fatalities, 15–17-year-olds
income	Per capita personal income in 1987 dollars	fatal1820	Number of vehicle fatalities, 18–20-year-olds
emppop	Employment/population ratio	nfatal1820	Number of night-time vehicle fatalities, 18–20-year-olds
beertax	Tax on the case of beer	fatal2124	Number of vehicle fatalities, 21–24-year-olds
baptist	Percent of the southern baptist	nfatal2124	Number of night-time vehicle fatalities, 21–24-year-olds
mormon	Percent of mormon	afatal	Number of alcohol-involved vehicle fatalities

drinkage	Minimum legal drinking age	pop	Population
dry	Percent residing in "dry" countries	pop1517	Population, 15–17-year-olds
youngdrivers	Percent of drivers aged 15–24	pop1820	Population, 18–20-year-olds
miles	Average miles per driver	pop2124	Population, 21–24-year-olds
breath	Preliminary breath test law?	milestot	Total vehicle miles (millions)
jail	Mandatory jail sentence?	unempus	U.S. unemployment rate
service	Mandatory community service?	emppopus	U.S. employment/population ratio
fatal	Number of vehicle fatalities	gsp	GSP rate of change

Visualization tools that help us to examine data more clearly are below.





1.2. Research questions

The questions of interest are:

- Does the 15-17-year-old population have an effect on the number of fatalities?
- Is the mean of number of alcohol-involved vehicle fatalities in states where preliminary breath test exists equal to the mean of number of alcohol-involved vehicle fatalities in states where preliminary breath test does not exist?
- Is the mean of drink age in the USA equal to 21?
- Do the beer tax and per capita personal income have an association with the number of alcohol-involved vehicle fatalities in states?
- Is there a significant difference between the means of the number of vehicle fatalities for 15–17-year-olds, the number of vehicle fatalities for 18–20-year-olds, and the number of vehicle fatalities for 21–24-year-olds?
- Is the fatality rate (the number of traffic deaths per person living in the U.S. between 1982-1988) equal to 0.0001?

- Do mandatory jail sentences and mandatory community service have any effect on the number of numbers of vehicle fatalities?
- Is the fatality rate, which is the number of 15–17-year-old traffic deaths per a 15–17-year-old person living in the U.S., equal to the fatality rate, that the number of 21–24-year-old traffic deaths per a 21-24-year-old person living in the U.S. between 1982-1988?

1.3.Aim of the study

This project aimed to analyze the effects and causes of fatalities in each state of the United States of America by using applications of statistics. We aim to examine the data and determine which subjects cause or affect car accidents and how we can categorize these subjects with their influences on fatalities. We created research questions to satisfy our purpose and analyze the data effectively. With this project, we can understand the real reasons for fatalities and try to decrease the effects of these subjects to reduce accidents.

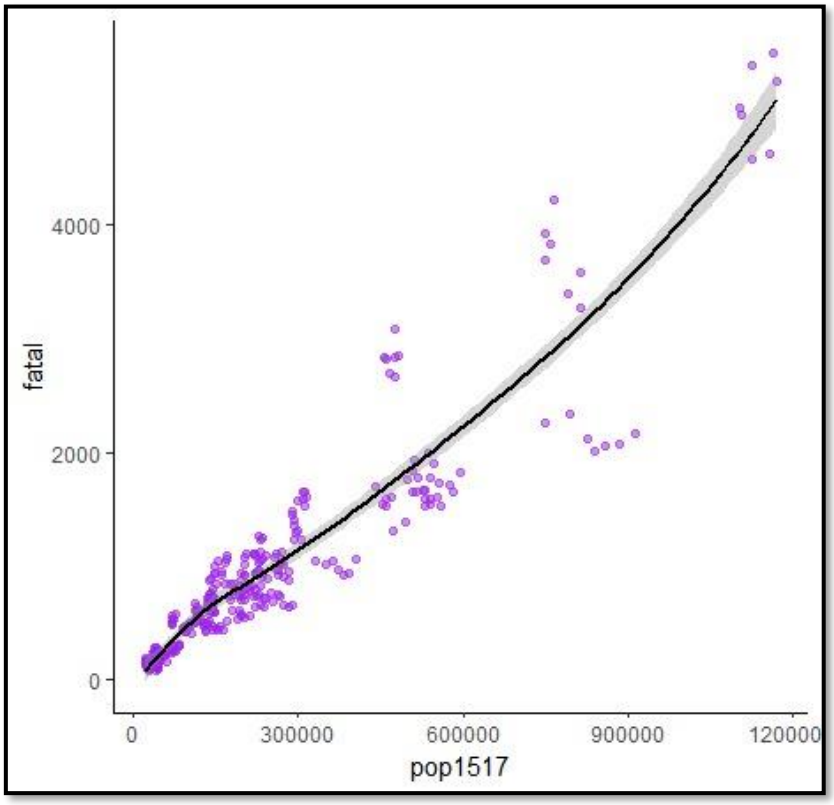
2. Methodology/Analysis

The methods we used in our research are simple linear regression, two-sample test hypotheses testing, one-sample hypothesis testing, multiple linear regression, one-way ANOVA, Tukey's test and two-way ANOVA. For the one-way ANOVA, we needed to create a new data frame to examine the means of three different types of drivers (see Appendix A). We used simple linear regression in the first question to observe the relationship between the dependent and independent variables. We used two-sample hypothesis testing in the second and eighth questions to compare means and proportions. In the third question and the sixth question, we used one-sample hypothesis testing to find out if the mean of the minimum legal drinking age equals to 21 and if the fatality rate equals 0.0001. For the fourth question, we used multiple linear regression to find out whether the beer tax and per capita personal income had an effect on the number of alcohol-involved vehicle fatalities. In the fifth and seventh questions, we took the logarithm of the numerical variables to normalize the dataset, and then, in the fifth question, we used the one-way ANOVA and Tukey's HSD (honestly significant difference) test to determine if there is a significant difference among the means. Finally, in the seventh question, we used the two-way ANOVA to find out whether mandatory jail sentences and mandatory community service have any effect on the number of numbers of vehicle fatalities. We used the

R software language and RStudio to generate the summary tables of the above tests and the required plots. Also, we used the car, dplyr, and ggplot2 packages in R.

3. Results and Findings

Question 1: Does the 15-17-year-old population have an effect on the number of vehicle fatalities?



As we have seen in the plot, there is a simple linear relationship between the 15-17-year-old population and the number of vehicle fatalities.

```
Call:
lm(formula = fatal ~ pop1517, data = fatal)

Residuals:
     Min       1Q   Median       3Q      Max
-1358.95   -95.62    -8.97   105.48  1246.86

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.158e+01  2.561e+01   2.014   0.0449 *
pop1517     3.800e-03  7.868e-05  48.296   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 331.1 on 334 degrees of freedom
Multiple R-squared:  0.8747,    Adjusted R-squared:  0.8744
F-statistic:  2332 on 1 and 334 DF,  p-value: < 2.2e-16

> cor(fatal$pop1517, fatal$fatal)
[1] 0.9352757
```

According to the summary table of the model, an estimated simple linear regression model is formed by the coefficients:

$$fatal = 51.58 + pop1517*0.0038$$

The way to interpret the coefficients is as follows:

When the 15-17-year-old population equals zero, the number of vehicle fatalities equals 51.58. Each additional one-unit increase in the 15-17-year-old population is associated with an average increase of 0.0038 points in the number of vehicle fatalities.

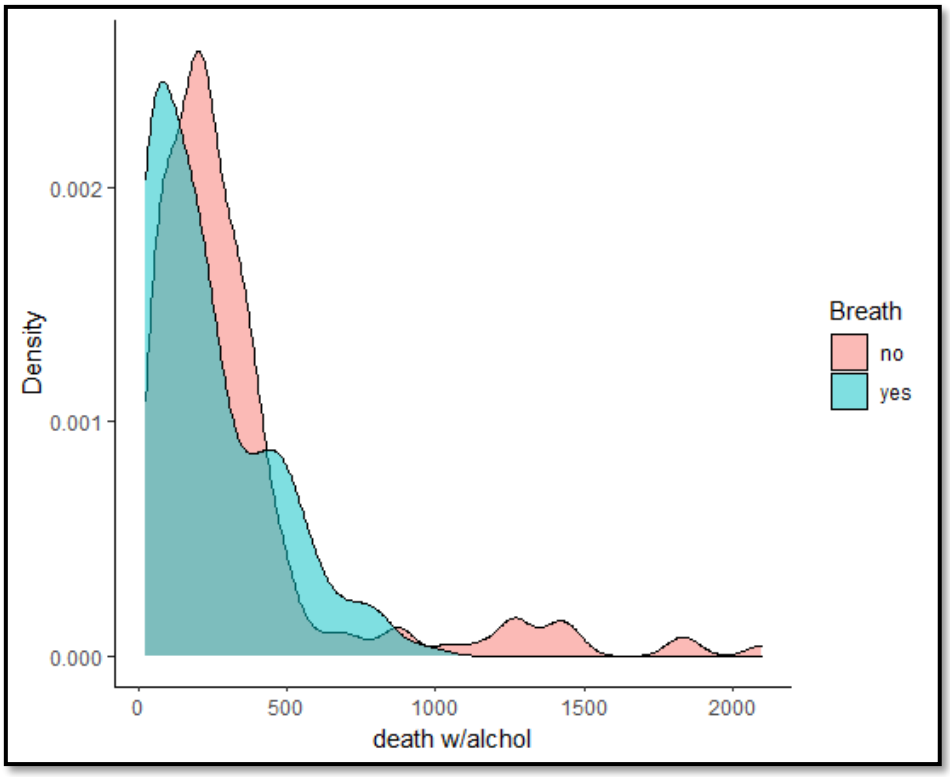
R-Square: 87.47 % of the variation in vehicle fatalities can be explained by the 15-17-year-old population.

Standard error: the observed values fall an average of 331.1 units from the regression line.

Significance F: the p-value (<2.2e-16) is less than 0.05, indicating that the 15-17-year-old population's explanatory variable has a statistically significant association with the number of vehicle fatalities.

Coefficient P-values: We can see that pop1517 (p<2e-16) is statistically significant at $\alpha = 0.05$.

Question 2: Is the mean of the number of alcohol-involved vehicle fatalities in states where preliminary breath test exists equal to the mean of the number of alcohol-involved vehicle fatalities in states where preliminary breath test does not exist?



According to the plot above, the density of the number of alcohol-involved vehicle fatalities in states where preliminary breath test does not exist is more than the density of the number of alcohol-involved vehicle fatalities in states where preliminary breath test exists for every value of the number of alcohol-involved vehicle fatalities.

Step 1: Define the hypotheses.

We will perform the two-sample t-test with the following hypotheses:

H0: $\mu 1 = \mu 2$ (the two population means are equal)

H1: $\mu 1 \neq \mu 2$ (the two population means are not equal)

Where $\mu 1$ is the mean of the number of alcohol-involved vehicle fatalities in states where preliminary breath test exists, and $\mu 2$ is the mean of the number of alcohol-involved vehicle fatalities in states where preliminary breath test does not exist.

Step 2: Calculate the p-value of the test statistic t

```
        Welch Two Sample t-test

data:  breathy[, "afatal"] and breathno[, "afatal"]
t = -3.0002, df = 296.29, p-value = 0.002928
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -155.52849  -32.31147
sample estimates:
mean of x mean of y
 240.0229  333.9429
```

According to the summary table of the t-test above, the p-value associated with t = -3.002 and degrees of freedom = 296.29 is 0.002928.

Step 3: Draw a conclusion.

As the p-value is less than our significance level $\alpha = 0.05$, we reject the null hypothesis (H0). We have enough evidence to say that the mean of number of alcohol-involved vehicle fatalities in states where preliminary breath test exists is different from the mean of number of alcohol-involved vehicle fatalities in states where preliminary breath test does not exist.

Question 3: Is the mean of drinkage in the U.S. equal to 21?

Step 1: Define the hypotheses.

We will perform the one-sample t-test with the following hypotheses:

H0: $\mu = 21$ (the population mean is equal to 21)

H1: $\mu \neq 21$ (the population mean is not equal to 21)

Where μ is mean of drinkage in the USA.

Step 2: Calculate the p-value of the test statistic t

```
        One Sample t-test

data:  Fatalities$drinkage
t = -11.105, df = 334, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 21
95 percent confidence interval:
 20.35729 20.55071
sample estimates:
mean of x
   20.454
```

According to the summary table of the t-test above, the p-value associated with t = -11.105 and degrees of freedom = 334 is less than 2.2e-16.

Step 3: Draw a conclusion.

As this p-value is less than our significance level $\alpha = 0.05$, we reject the null hypothesis(H0). We have enough evidence to say that the mean of drinkage in the USA is different from 21.

Question 4: Do the beer tax and per capita, personal income have an association with the number of alcohol-involved vehicle fatalities in states?

```
Call:
lm(formula = afatal ~ income + beertax, data = fatal)

Residuals:
    Min      1Q  Median      3Q     Max
-384.83 -175.94  -72.27   59.02 1808.08

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.079e+02  1.197e+02  -0.902  0.36783
income       2.517e-02  7.898e-03   3.187  0.00157 **
beertax      1.010e+02  3.724e+01   2.713  0.00702 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 298.9 on 333 degrees of freedom
Multiple R-squared:  0.03657,   Adjusted R-squared:  0.03078
F-statistic:  6.32 on 2 and 333 DF,  p-value: 0.002024
```

According to the summary table of the model, an estimated multiple linear regression model is formed by the coefficients:

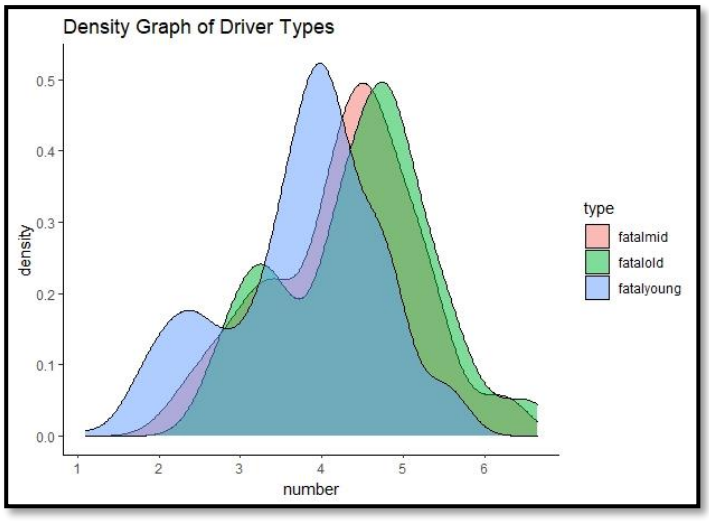
*afatal = -107.9 + income*0.02517 + beertax*101*

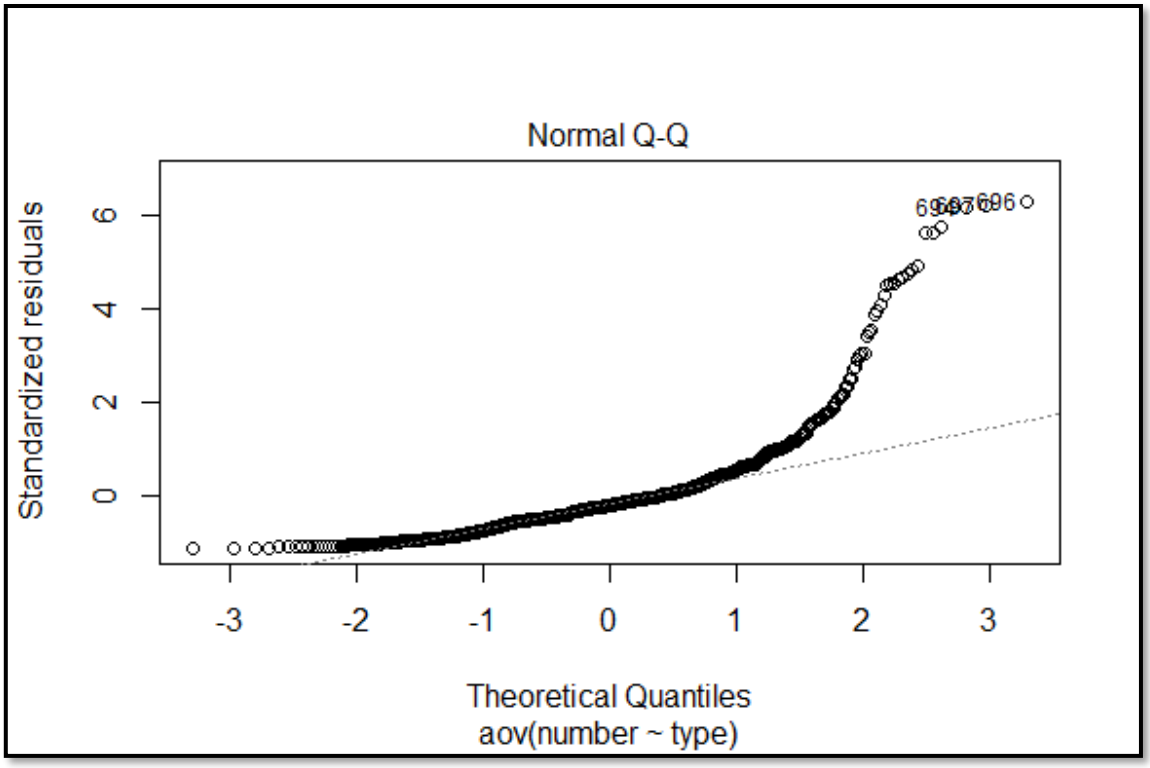
The way to interpret the coefficients is as follows:
When the beer tax and income are equal to zero, the alcohol-involved vehicle fatalities are equal to -107.9.
Each additional one-unit increase in income is associated with an average increase of 0.02517 points in alcohol-involved vehicle fatalities, assuming beer tax is held constant.
Each additional one-unit increase in beer tax is associated with an average decrease of 101 points in alcohol-involved vehicle fatalities, assuming income is held constant.

R-Square: 3.657 % of the variation in the alcohol-involved vehicle fatalities can be explained by income and beer tax.
Standard error: the observed values fall an average of 298.9 units from the regression line.
Significance F: the p-value is less than 0.05, which indicates that the explanatory variables income and beer tax combined have a statistically significant association with alcohol-involved vehicle fatalities.
Coefficient P-values: We can see that income (p=0.00157) and beertax (p=0.00702) are statistically significant at $\alpha = 0.05$.

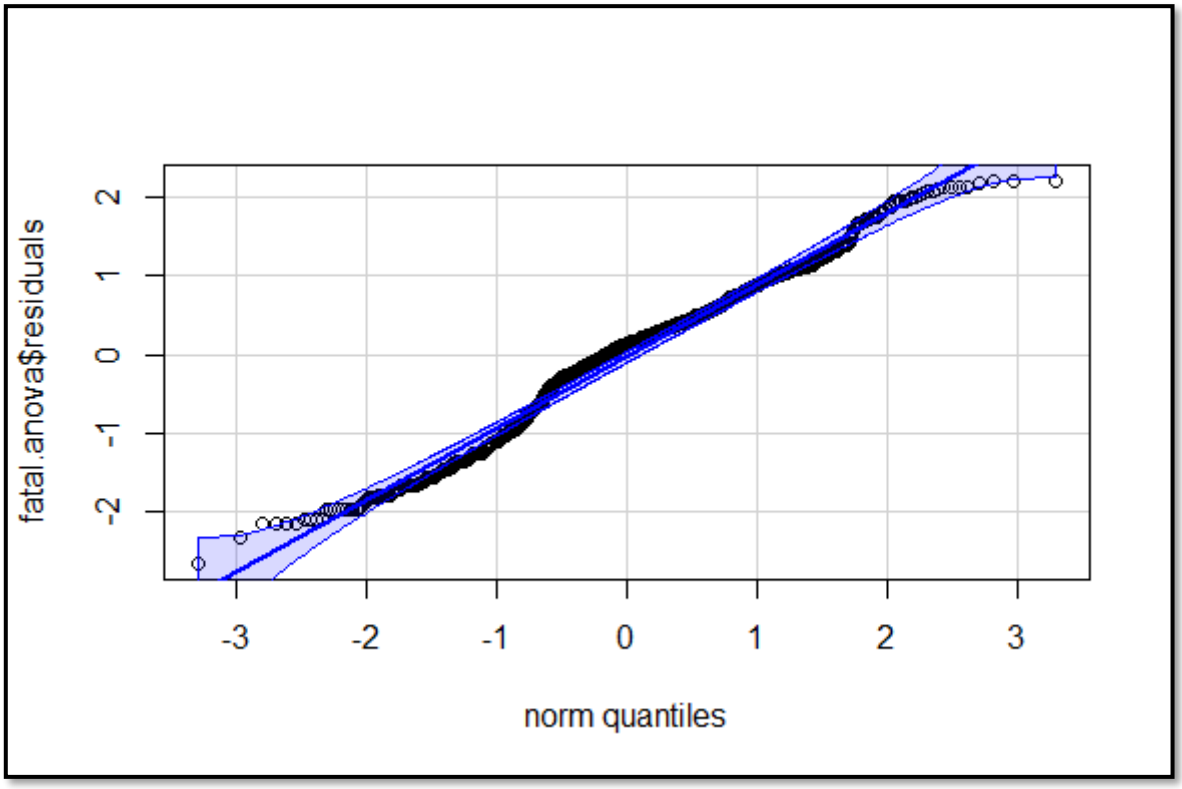
Question 5: Is there a significant difference between the means of the number of vehicle fatalities for 15–17-year-olds, the number of vehicle fatalities for 18–20-year-olds, and the number of vehicle fatalities for 21–24-year-olds?

For the one-way ANOVA, we created a new data frame (see Appendix A). In the type of column; fatalyoung = fatal1517, fatalmid = fatal1820, fatalold = fatal2124. Types should be factors; fatalities should be numeric.

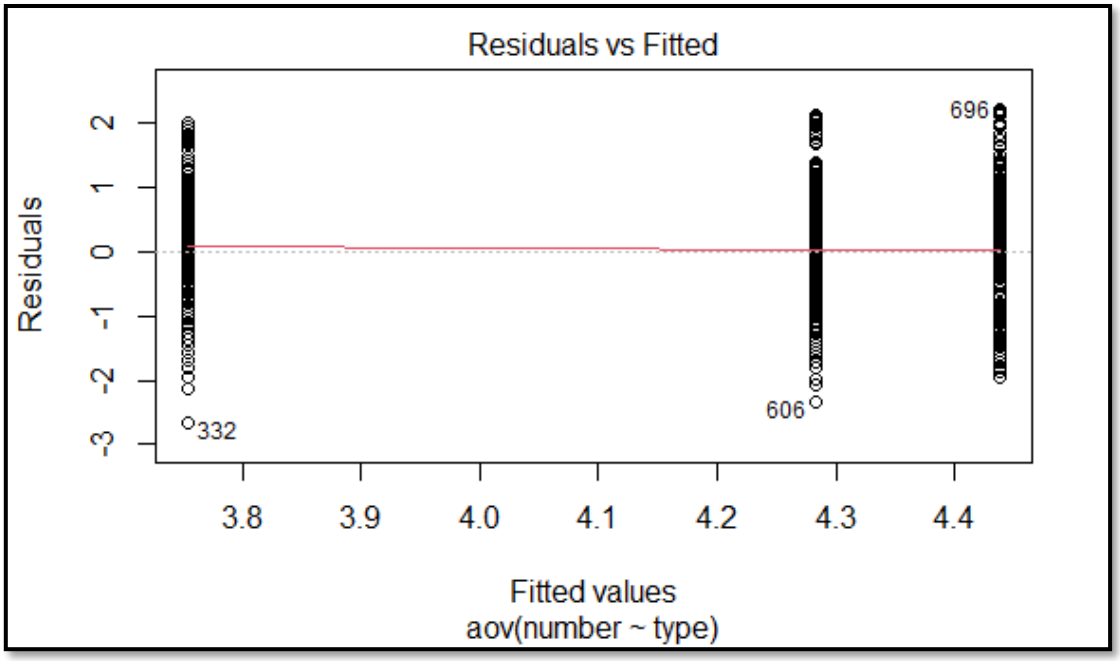




Since the points do not lie mainly along the straight diagonal line with deviations, we normalize the data by taking the logarithm of the numerical variables.



The points lie mostly along the straight diagonal line with minor deviations along each tail. Hence, the normality assumption of one-way ANOVA is met.



The Residuals vs. Fitted plot above allows us to check our equal variances assumption. As we have seen, the residuals are equally spread out for each level of the fitted values.

```
> leveneTest(number~type, data = b2)
Levene's Test for Homogeneity of Variance (center = median)
        Df F value Pr(>F)
group    2  0.0452 0.9558
      1005
```

Also, according to the output above, the p-value is greater than the significance level of 0.05. This means there is no evidence to suggest that the variance across groups is statistically significantly different. Therefore, we can assume the variances are equal in the different treatment groups. Hence, the equal variances assumption of one-way ANOVA is met.

Since the assumptions of one-way ANOVA are met, the one-way ANOVA can apply.

State the hypotheses.

H0 (null hypothesis): $\mu 1 = \mu 2 = \mu 3$ (all the population means are equal)
H1 (alternative hypothesis): at least one population mean is different from the rest
$\mu 1$, $\mu 2$, and $\mu 3$ are the means of fatalities according to the type column.

```
> summary(fatal.anova)
             Df Sum Sq Mean Sq F value Pr(>F)
type          2   86.1   43.07    50.4 <2e-16 ***
Residuals  1005  859.0    0.85
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

As we have seen in the summary table, the F test statistic is 50.4, and the corresponding p-value is less than 2e-16.

Make The Decision.

Since this p-value is less than 0.05, we reject the null hypothesis.

Summarize The Results.

This means we have sufficient evidence to say that there is a statistically significant difference between the mean of the fatalities of the three groups.

This question can be examined with another test: Tukey's test. This multiple comparison test is used to find means that are significantly different from each other.

```
> summary(glht(model1, mcp(type = "Tukey")))

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = number ~ type, data = df)

Linear Hypotheses:
                           Estimate Std. Error t value Pr(>|t|)
fatalold - fatalmid == 0    0.03807    0.01910   1.993    0.114
fatalyoung - fatalmid == 0 -0.14337    0.01910  -7.506   <1e-04 ***
fatalyoung - fatalold == 0 -0.18144    0.01910  -9.499   <1e-04 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- single-step method)
```

According to this test, there is a significant difference between means of fatalyoung and fatalold, fatalyoung and fatalmid. However, it seems like means of fatalold and fatalmid are not different, which can be caused by the test's problem. We checked these two types with a t-test.

```
        Welch Two Sample t-test

data:  df %>% filter(type == "fatalold") %>% pull() and df %>% filter
(type == "fatalmid") %>% pull()
t = 2.1913, df = 667.65, p-value = 0.02878
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.003957616 0.072191304
sample estimates:
mean of x mean of y
 1.467402  1.429328
```

As desired, the p-value is 0.02878, which is smaller than 0.05. We can conclude that means of these two types are different from each other.

Question 6: Is the fatality rate (the number of traffic deaths per person living in the U.S. between 1982-1988) equal to 0.0001?

State the hypotheses.

H0 (null hypothesis): the fatality rate (the number of traffic deaths per person living in the U.S. between 1982-1988) is equal to 0.0001

H1 (alternative hypothesis): the fatality rate (the number of traffic deaths per person living in the U.S. between 1982-1988) is not equal to 0.0001
Nevada state is used as a sample.

Find the critical values.

The critical value at significance level $\alpha = 0.05$ is 1.96 according to the z table.

```
> nevada_prop
    fatal       pop
176   280  877998.9
177   253  897000.9
178   249  916998.7
179   259  936001.3
180   233  967001.6
181   262 1006999.1
182   286 1054001.0
> p_hat
[1] 0.000274526
> p_zero
[1] 1e-04
> q_zero
[1] 0.9999
> z_score_nevada
[1] 17.01922
> mean(data$fatal / data$pop)
[1] 0.0002040444
```

According to the summary table of the z test above, the test value is 17.01922.

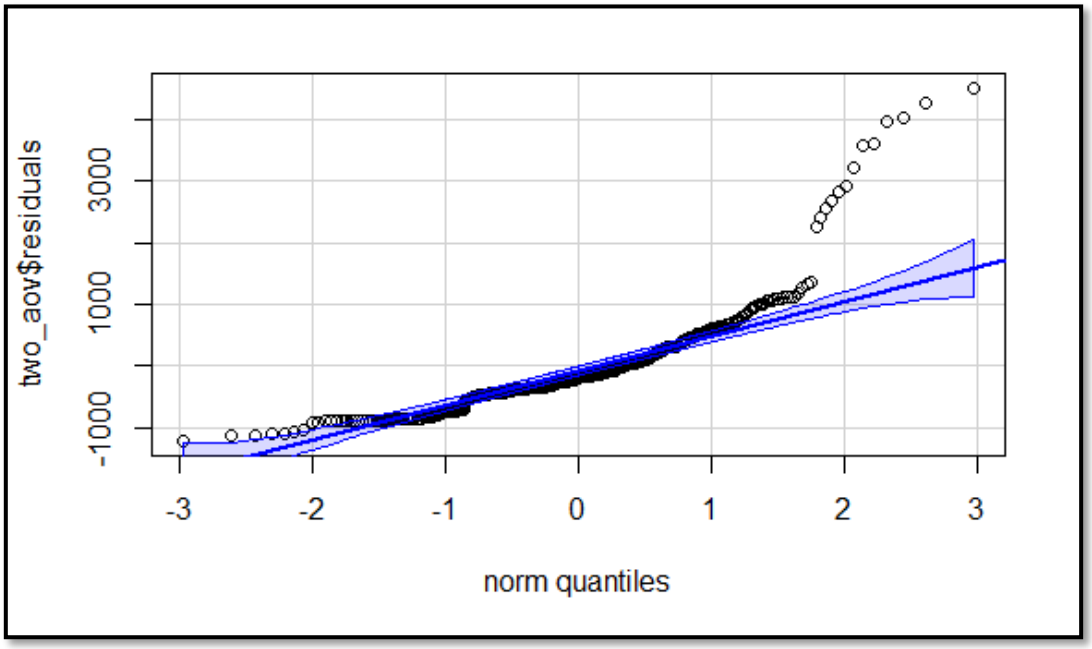
Make The Decision.

As the z score (17.01922) is greater than the critical value (1.96), we reject the null hypothesis (H0).

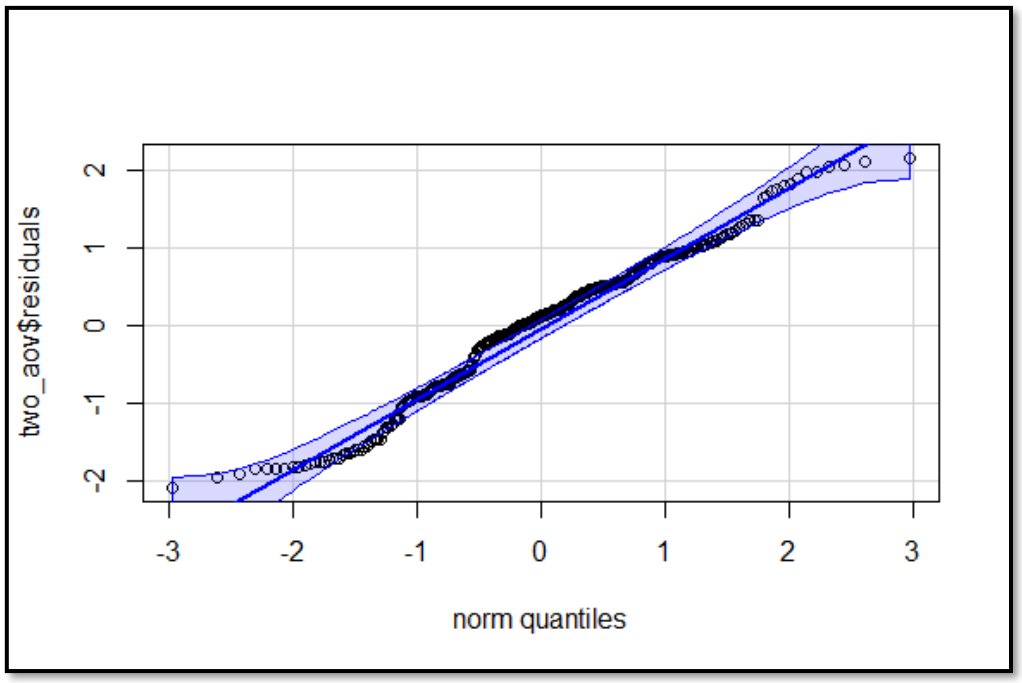
Summarize The Results.

We have enough evidence to say that the fatality rate (the number of traffic deaths per person living in the U.S. between 1982-1988) is not equal to 0.0001.

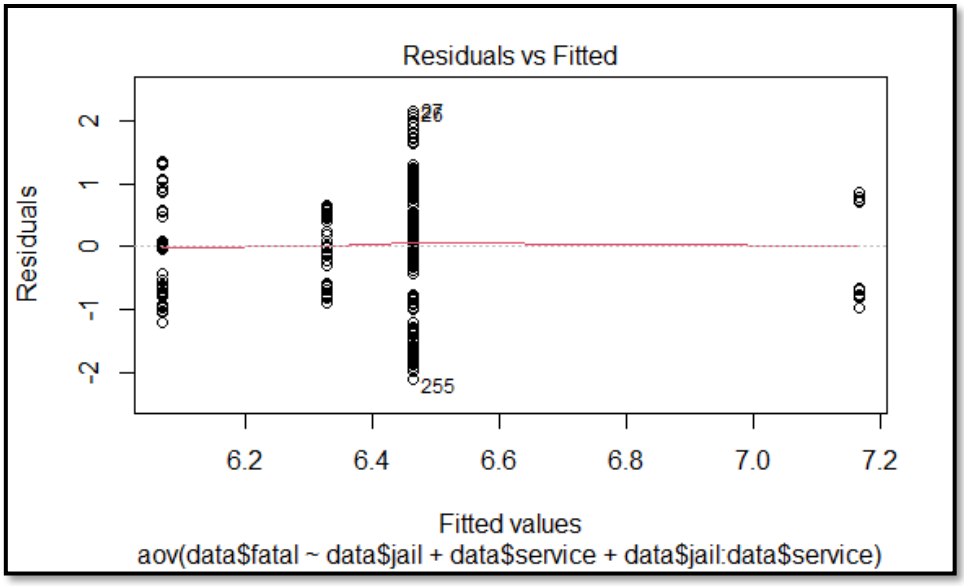
Question 7: Do mandatory jail sentences and mandatory community service have any effect on the number of vehicle fatalities?



Since the points do not lie mainly along the straight diagonal line, we normalize the data by taking the logarithm of the numerical variables.



The points lie mostly along the straight diagonal line with minor deviations along each tail. Hence, the normality assumption of one-way ANOVA is met.



The Residuals vs. Fitted plot above allows us to check our equal variances assumption. As we have seen, the residuals are equally spread out for each level of the fitted values.

```
> leveneTest(fatal~service*jail, data = fatal)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value   Pr(>F)
group   3  4.8269 0.002652 **
      331
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Also, according to the output above, the p-value is greater than the significance level of 0.001. This means there is no evidence to suggest that the variance across groups is statistically significantly different. Therefore, we can assume the variances are equal in the different treatment groups. Hence, the equal variances assumption of two-way ANOVA is met.

Since the assumptions of two-way ANOVA are met, the two-way ANOVA can apply.

State the hypotheses.

The hypotheses for the interaction:

H0: There is no interaction effect between a mandatory jail sentence and mandatory community service on the number of vehicle fatalities.

H1: There is an interaction effect between a mandatory jail sentence and mandatory community service on the number of vehicle fatalities.

The hypotheses for the mandatory jail sentence:

H0: There is no difference between the means of the number of vehicle fatalities for mandatory jail sentences.
H1: There is a difference between the means of the number of vehicle fatalities for mandatory jail sentences.

The hypotheses for the mandatory community service:

H0: There is no difference between the means of the number of vehicle fatalities for mandatory community service.
H1: There is a difference between the means of the number of vehicle fatalities for mandatory community service.

Find the critical values for each F test.

Factor A is designated as the mandatory jail sentence. It has two levels, yes or no; therefore, d.f.N = 2-1=1. Factor B is designated as the mandatory community service. It has two levels, yes or no; therefore, d.f.N=2-1=1. The degree of freedom for interaction effect is d.f.N. = (2-1)*(2-1)=1. The degree of freedom for within(error) is d.f.D. =331. The critical value for the F(A) test of Factor A is found by using $\alpha = 0.05$, d.f.N=1, and d.f.D. =331. In this case, F(A) =3.84. The critical value for Factor B's F(B) test is found using $\alpha = 0.05$, d.f.N=1, and d.f.D. =331. In this case, F(B) =3.84. Finally, the critical value for the F(AxB) test of interaction effect is found by using $\alpha = 0.05$,d.f.N=1 and d.f.D. =331. In this case, F(AxB) =3.84.

```
> summary(two_aov)
              Df Sum Sq Mean Sq F value  Pr(>F)
jail           1   6.21   6.215   7.582 0.00622 **
service        1   6.41   6.414   7.826 0.00545 **
jail:service   1   1.66   1.657   2.021 0.15605
Residuals    331 271.31   0.820
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
1 observation deleted due to missingness
```

Make The Decision.

Since F(A)=7.582 and F(B) =7.826 are greater than the critical value of 3.84, the null hypotheses concerning the mandatory jail sentences and the mandatory community service should be rejected.

Summarize The Results.

Since the null hypothesis for Factor A was rejected, it can be concluded that the mandatory jail sentence significantly affects the number of vehicle fatalities.

Since the null hypothesis for Factor B was rejected, it can be concluded that mandatory community service significantly affects the number of vehicle fatalities.

Since the null hypothesis for the interaction effect failed to reject, it can be concluded that combining mandatory jail sentences and mandatory community service does not affect the number of vehicle fatalities.

Question 8: Is the fatality rate, which is the number of 15–17-year-old traffic deaths per a 15–17-year-old person living in the U.S., equal to the fatality rate that the number of 21–24-year-old traffic deaths per a 21-24-year-old person living in the U.S. between 1982-1988?

State the hypotheses.

H0: $p1=p2$ and H1: $p1 \neq p2$

New York state is used as a sample.

Where p1 is the fatality rate which is the number of 15–17-year-old traffic deaths per a 15–17-year-old person living in the U.S. between 1982-1988, and p2 is the fatality rate that the number of 21–24-year-old traffic deaths per a 21-24-year-old person living in the U.S. between 1982-1988.

Find the critical values.

Since the significance level is $\alpha = 0.05$, the critical values are 1.96 and -1.96.

```
> new_york_fatal_pop
    fatal1517 fatal2124  pop1517 pop2124
204       129       298 913002.9 1218998
205       121       248 885999.1 1225000
206       108       246 858000.9 1224002
207       126       263 838000.0 1214001
208       146       257 827997.6 1108002
209       157       304 796000.7 1074000
210       126       281 748000.9 1138998
> p1
[1] 0.0001565867
> p2
[1] 0.0002323203
> p_bar
[1] 0.0001997156
> q_bar
[1] 0.9998003
> z_score_new_york
[1] -3.872306
```

Make The Decision.

Reject the null hypothesis since -3.872306 is less than -1.96.

Summarize the Results.

There is enough evidence to reject the claim that there is no difference between the fatality rate of 15-17-year-old and the fatality rate of 21-24-year-old.

4. Discussion/Conclusion

In this project, we investigated the car accident fatalities in every state of the U.S. between 1982 and 1988 in the "Fatalities" data. We created eight different research questions to understand the data correctly. As a result of our questions, we have found a linear relationship between the 15–17-year-old population and the fatality numbers, and the correlation between these variables is 0.93, which is quite close to 1. The result of our first question supports the idea that younger drivers may cause more car accident fatalities. Furthermore, we determined that the mean of the number of alcohol-involved vehicle fatalities in states where preliminary breath test exists is different from the mean of the number of alcohol-involved vehicle fatalities in states where preliminary breath test does not exist. Hence, the breath test may affect car accident fatalities. In addition, since the minimum drink age in the U.S. is known as 21, we encountered that it is changing from state to state, and the mean is not equal to 21. Moreover, we created a multiple linear model and noticed a significant association of income and beer tax and a number of alcohol-involved vehicle fatalities. Besides, the one-way ANOVA test showed us that there is a significant difference between the means of the fatalities of the three different age groups. In addition to this question, we conducted a multiple comparison test called

Tukey's. The test told us that the two groups' means are different from each other. We checked the third group with Welch's two-sample t-test, which supports our first finding from ANOVA. All in all, we found that means of these three types are different from each other. Likewise, we thought that fatalities caused by car accidents in a country could be 0.01%, and we stated our hypothesis with this idea. After the calculations, it can be shown that the fatality rate in the U.S. is not equal to 0.01%. Further, the two-way ANOVA test showed us that the mandatory jail sentence, and the mandatory community service sentence significantly affect the number of vehicle fatalities. On the other hand, the interaction of the mandatory jail sentence and community service sentence does not affect the number of vehicle fatalities. Lastly, we found no difference between the two different age groups' fatality rates. In conclusion, the fatality numbers of a country can be affected by many variables, as well as causes various circumstances.

References

Arel-Bundock, V. (2022). *Rdatasets: A collection of datasets originally distributed in various R packages*. R package version 1.0.0. https://vincentarelbundock.github.io/Rdatasets.

By Zach. (2019, April 29). *How to Conduct a One-Way ANOVA in R*. Statology. https://www.statology.org/one-way-anova-r/

Camli, O. (2022, May 13), *Linear Regression Models [Recitation Notes]*, METU, https://odtuclass2021s.metu.edu.tr/pluginfile.php/417829/mod_resource/content/0/Stat%20250-Rec%207.pdf

Camli, O. (2022, May 27), *Analysis of Variance [Recitation Notes]*, METU, https://odtuclass2021s.metu.edu.tr/pluginfile.php/433751/mod_resource/content/0/Stat%20250-Rec%209.pdf

Pipis, G. (2020). ANOVA vs Multiple Comparisons | R-bloggers. R-bloggers. Retrieved 2 July 2022, from https://www.r-bloggers.com/2020/10/anova-vs-multiple-comparisons/.

R Documentation (n.d.), *Fatalities*, https://www.rdocumentation.org/packages/AER/versions/1.2-10/topics/Fatalities

Two-Way ANOVA Test in R. (n.d.). STHDA. http://www.sthda.com/english/wiki/two-way-anova-test-in-r

Appendices

Appendix A: Part of the data frame that we created for one-way ANOVA test from our original data.

(Data frame is in csv file format)

type	number				
fatalyoung	53	fatalmid	62	fatalold	81
fatalyoung	71	fatalmid	65	fatalold	64
fatalyoung	49	fatalmid	72	fatalold	63
fatalyoung	66	fatalmid	56	fatalold	69
fatalyoung	82	fatalmid	66	fatalold	76
fatalyoung	94	fatalmid	72	fatalold	23
fatalyoung	66	fatalmid	61	fatalold	18
		fatalmid	567	fatalold	17

The summary and structure of the data frame:

```
> summary(b2)
         type         number
 fatalmid  :336   Min.   :  3.00
 fatalold  :336   1st Qu.: 33.00
 fatalyoung:336   Median : 70.00
                  Mean   : 98.71
                  3rd Qu.:123.00
                  Max.   :770.00
> str(b2)
'data.frame':   1008 obs. of  2 variables:
 $ type  : Factor w/ 3 levels "fatalmid","fatalold",..: 3 3 3 3 3 3 3 3 3 ...
 $ number: num  53 71 49 66 82 94 66 40 40 51 ...
```