Assignment 3

Miguel Perez-Luna

9/20/2020

Contents

```
library(tidyverse)
library(ggplot2)
```

Before starting, I am loading the person-level data I created, and I am going to recode some variables to make labeling axes easier later on.

```
transpo_data <- read.csv("transpo_data.csv") %>%
  mutate(meansTW = case_when(
    JWTR_label=="Bicycle"~"Bicycle",
    JWTR_label=="Bus or trolley bus"~"Bus or\ntrolley bus",
    JWTR_label=="Car, truck, or van"~"Car, truck,\nor van",
    JWTR_label=="Motorcycle"~"Motorcycle",
    JWTR_label=="Other method"~"Other",
    JWTR_label=="Railroad"~"Railroad",
    JWTR_label=="Streetcar or trolley car (carro publico in Puerto Rico)"~"Streetcar or\ntrolley car",
    JWTR_label=="Subway or elevated"~"Subway or\nelevated",
    JWTR_label=="Taxicab"~"Taxicab",
    JWTR_label=="Walked"~"Walk")) %>%
  mutate(edu = case_when(
    SCHL_label=="Some college, but less than 1 year"~"<1 year of college",
    SCHL_label=="Regular high school diploma"~"High school diploma",
    SCHL_label=="Bachelor's degree"~"Bachelor's degree",
    SCHL_label=="1 or more years of college credit, no degree"~"> 1 yr of college,\nno degree",
    SCHL_label=="Master's degree"~"Master's degree",
    SCHL_label=="GED or alternative credential"~"GED",
    SCHL_label=="Doctorate degree"~"Doctorate degree",
    SCHL_label=="Associate's degree"~"Associate's degree",
    SCHL_label=="Grade 8"~"Grade 8",
    SCHL_label=="Grade 7"~"Grade 7",
    SCHL_label=="Grade 10"~"Grade 10",
    SCHL_label=="12th grade - no diploma"~"12th grade -\nno diploma",
    SCHL_label=="Grade 6"~"Grade 6",
    SCHL_label=="Grade 11"~"Grade 11",
    SCHL_label=="Professional degree beyond a bachelor's degree"~"Professional degree",
    SCHL_label=="Grade 9"~"Grade 9",
    SCHL_label=="No schooling completed"~"No schooling",
    SCHL_label=="Grade 5"~"Grade 5",
    SCHL_label=="Grade 4"~"Grade 4",
    SCHL_label=="Nursery school, preschool"~"Preschool",
    SCHL_label=="Grade 3"~"Grade 3",
```

```
    SCHL_label=="Grade 1"~"Grade 1",
    SCHL_label=="Kindergarten"~"Kindergarten",
    SCHL_label=="Grade 2"~"Grade 2"))
```

In this assignment, I had 5 continuous variables, 1 binary categorical variable, and 2 categorical variables, one with 10 unique categories and the other with 24 unique categories. This means that I ran 10 Pearson's correlation tests, 5 2-sample t-tests, 10 ANOVA tests, and 3 chi-square tests.

My sample size remains the same as my last assignment: 6914 people in Colorado who pay rent, and spend time commuting to work.

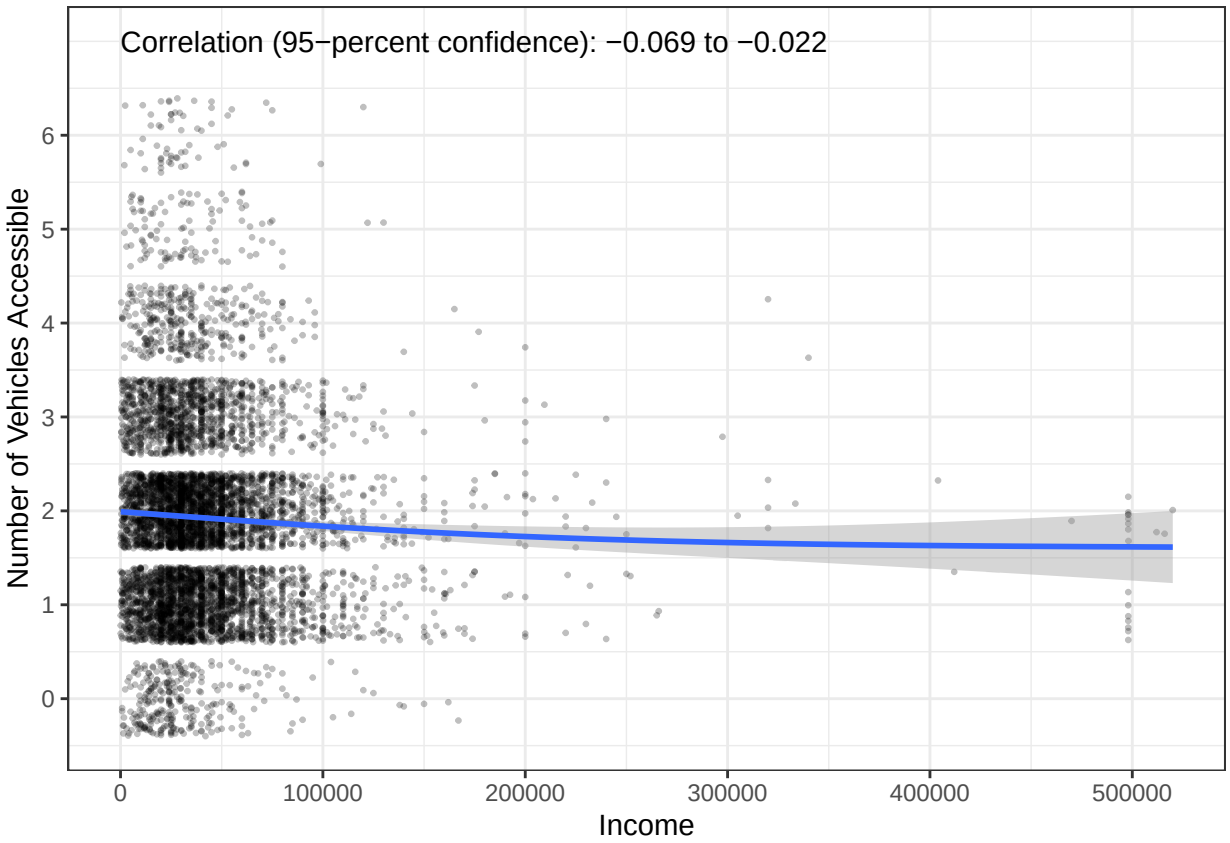
Relationships between two continuous variables

Number of Vehicles and Income

```
veh_inc_correlation <- cor.test(transpo_data$vehicle, transpo_data$PINCP)

veh_inc_correlation
```

```
##
##  Pearson's product-moment correlation
##
## data:  transpo_data$vehicle and transpo_data$PINCP
## t = -3.7988, df = 6912, p-value = 0.0001466
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.06914258 -0.02209666
## sample estimates:
##         cor
## -0.04564493
```

```
options(scipen = 999)
ggplot(transpo_data, aes(x = PINCP, y = vehicle)) +
  geom_point(size = 0.5, alpha = 0.25, position = "jitter") +
  geom_smooth() +
  theme_bw() +
  scale_x_continuous(name = "Income") +
  scale_y_continuous(name = "Number of Vehicles Accessible",
                     breaks = seq(0, 6, by = 1),
                     labels = paste(prettyNum(seq(0, 6, by = 1),
                                              big.mark = ","))) +
  annotate(geom = "text", x = 0, y = 7,
           label = paste("Correlation (95-percent confidence):",
                         prettyNum(veh_inc_correlation$conf.int[1], digits = 2),
                         "to",
                         prettyNum(veh_inc_correlation$conf.int[2], digits = 2)),
           hjust = 0)
```



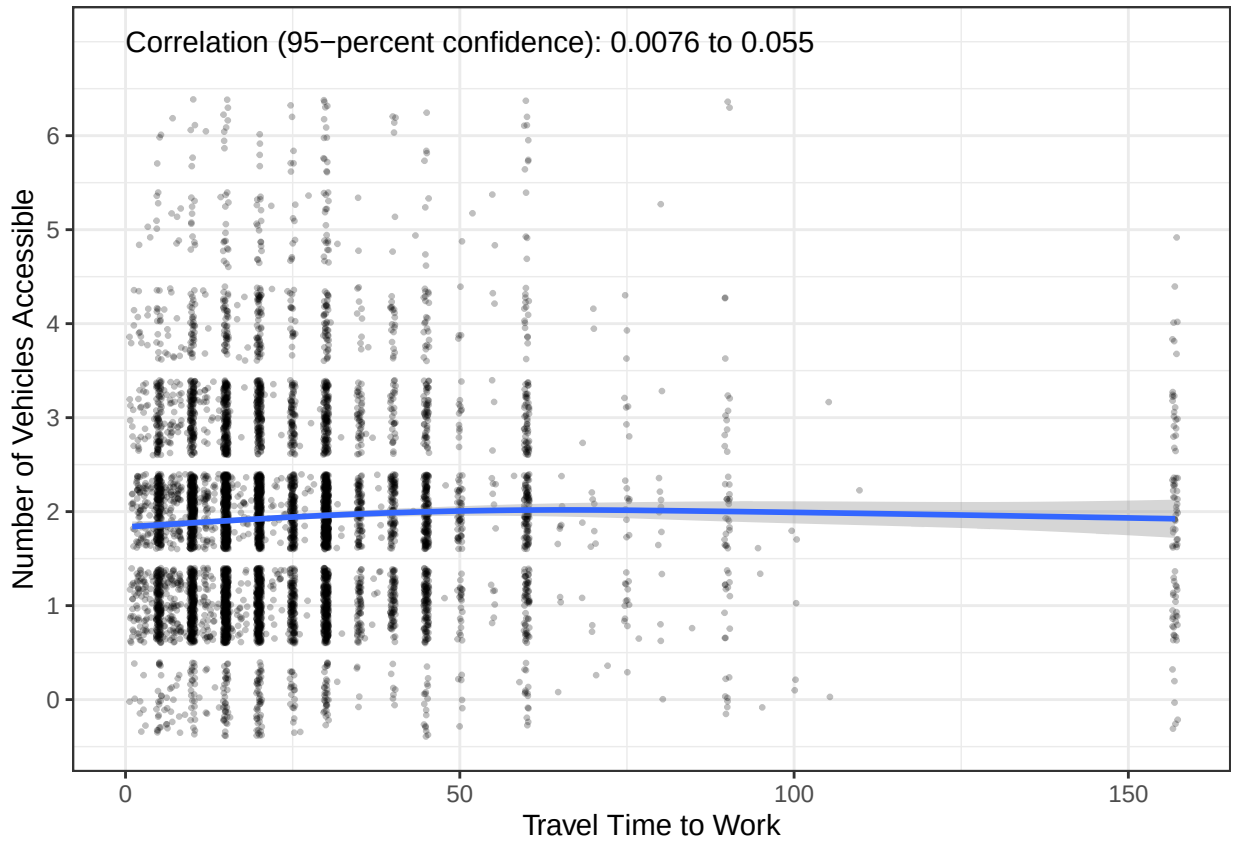
The correlation between income and number of vehicles accessible is surprisingly weak with an r-value of -0.04564493 , but still statistically significant (p = 0.0001466). This could be driven by the fact that some of the outliers in the income variable report having access to only 1 or 2 vehicles. It's also possible that regardless of how much income one earns, most homes in the Northern Colorado area only have one- to two-car garages.

Number of Vehicles and Travel Time to Work

```
veh_time_correlation <- cor.test(transpo_data$vehicle, transpo_data$JWMNP)

veh_time_correlation
```

```
##
##  Pearson's product-moment correlation
##
## data:  transpo_data$vehicle and transpo_data$JWMNP
## t = 2.5957, df = 6912, p-value = 0.009459
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.007639928 0.054738125
## sample estimates:
##        cor
## 0.03120635
```

```
options(scipen = 999)
ggplot(transpo_data, aes(x = JWMNP, y = vehicle)) +
  geom_point(size = 0.5, alpha = 0.25, position = "jitter") +
  geom_smooth() +
  theme_bw() +
  scale_x_continuous(name = "Travel Time to Work") +
  scale_y_continuous(name = "Number of Vehicles Accessible",
                     breaks = seq(0, 6, by = 1),
                     labels = paste(prettyNum(seq(0, 6, by = 1),
                                              big.mark = ","))) +
  annotate(geom = "text", x = 0, y = 7,
           label = paste("Correlation (95-percent confidence):",
                         prettyNum(veh_time_correlation$conf.int[1], digits = 2),
                         "to",
                         prettyNum(veh_time_correlation$conf.int[2], digits = 2)),
           hjust = 0)
```



The correlation between travel time to work and number of vehicles accessible is weak, with an r-value of 0.03120635. It is statistically significant, though (p = 0.009459).

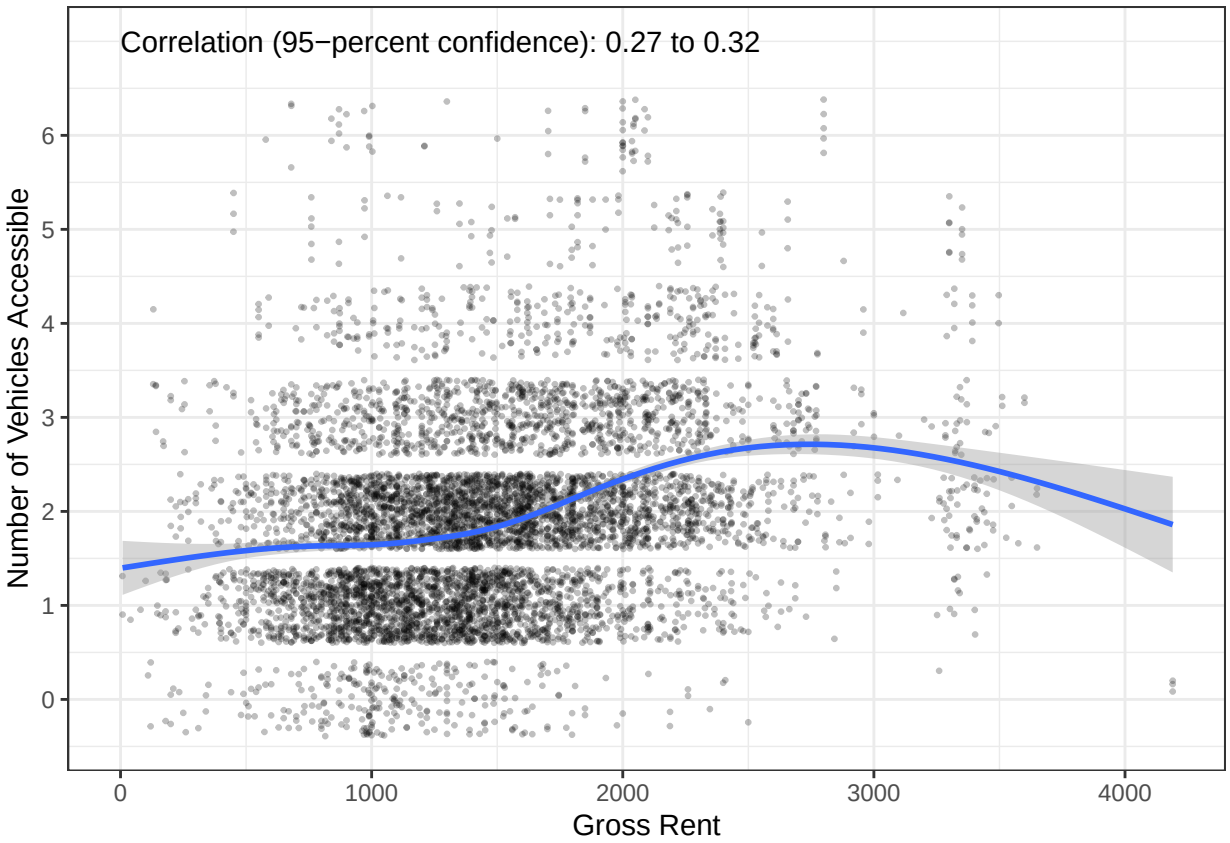
Number of Vehicles and Gross Rent

```
veh_rent_correlation <- cor.test(transpo_data$vehicle, transpo_data$GRNTP)
```

```
veh_rent_correlation
```

```
##
##  Pearson's product-moment correlation
##
## data:  transpo_data$vehicle and transpo_data$GRNTP
## t = 25.663, df = 6912, p-value < 0.00000000000000022
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2732750 0.3163199
## sample estimates:
##       cor
## 0.2949471
```

```
options(scipen = 999)
ggplot(transpo_data, aes(x = GRNTP, y = vehicle)) +
  geom_point(size = 0.5, alpha = 0.25, position = "jitter") +
  geom_smooth() +
  theme_bw() +
  scale_x_continuous(name = "Gross Rent") +
  scale_y_continuous(name = "Number of Vehicles Accessible",
                     breaks = seq(0, 6, by = 1),
                     labels = paste(prettyNum(seq(0, 6, by = 1),
                                              big.mark = ","))) +
  annotate(geom = "text", x = 0, y = 7,
           label = paste("Correlation (95-percent confidence):",
                         prettyNum(veh_rent_correlation$conf.int[1], digits = 2),
                         "to",
                         prettyNum(veh_rent_correlation$conf.int[2], digits = 2)),
           hjust = 0)
```



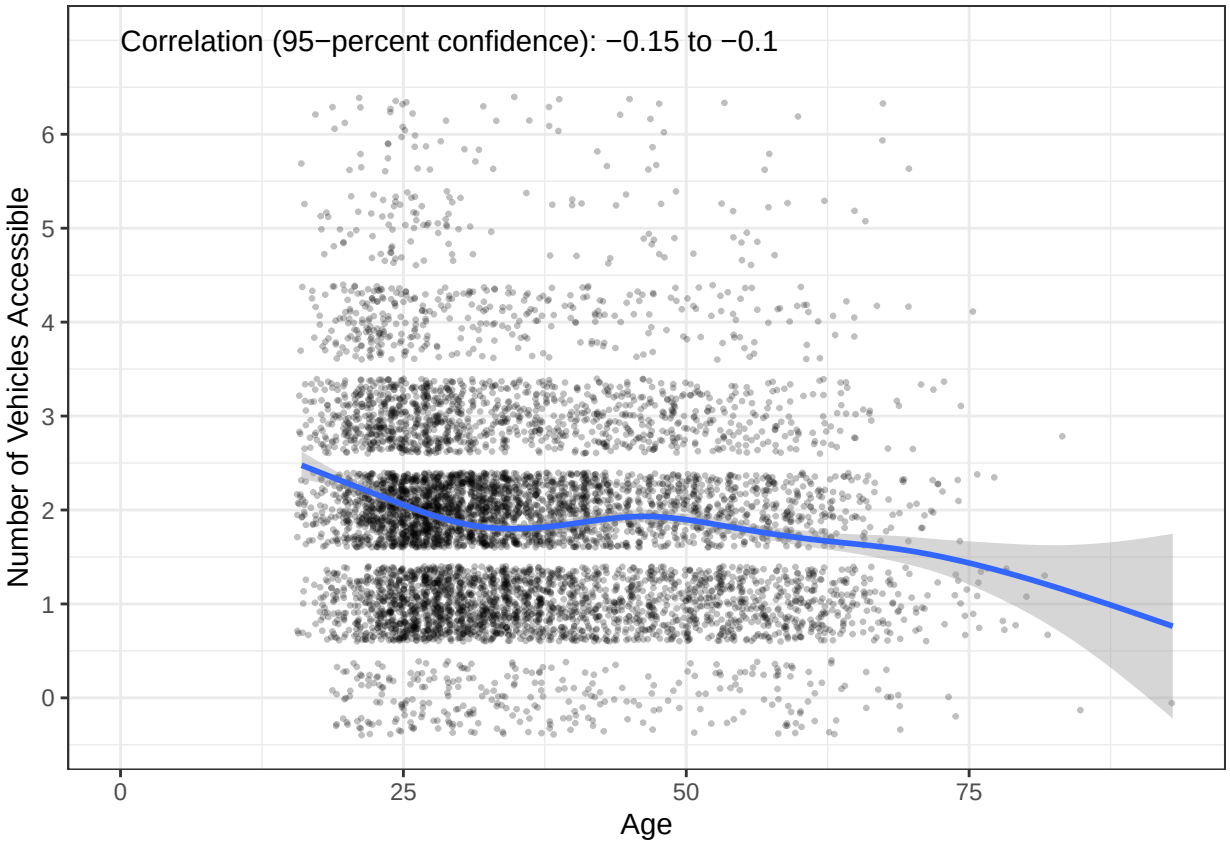
This correlation is the strongest one I've found yet, with an r-value of 0.2949471. It is also statistically significant. It's interesting to note that the trend line increases until about a rent of $3000, after which it decreases.

Number of Vehicles and Age

```
veh_age_correlation <- cor.test(transpo_data$vehicle, transpo_data$AGEP)

veh_age_correlation
```

```
##
##  Pearson's product-moment correlation
##
## data:  transpo_data$vehicle and transpo_data$AGEP
## t = -10.427, df = 6912, p-value < 0.00000000000000022
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1475839 -0.1011695
## sample estimates:
##        cor
## -0.1244448
```

```r
options(scipen = 999)
ggplot(transpo_data, aes(x = AGEP, y = vehicle)) +
  geom_point(size = 0.5, alpha = 0.25, position = "jitter") +
  geom_smooth() +
  theme_bw() +
  scale_x_continuous(name = "Age") +
  scale_y_continuous(name = "Number of Vehicles Accessible",
                     breaks = seq(0, 6, by = 1),
                     labels = paste(prettyNum(seq(0, 6, by = 1),
                                              big.mark = ","))) +
  annotate(geom = "text", x = 0, y = 7,
           label = paste("Correlation (95-percent confidence):",
                         prettyNum(veh_age_correlation$conf.int[1], digits = 2),
                         "to",
                         prettyNum(veh_age_correlation$conf.int[2], digits = 2)),
           hjust = 0)
```



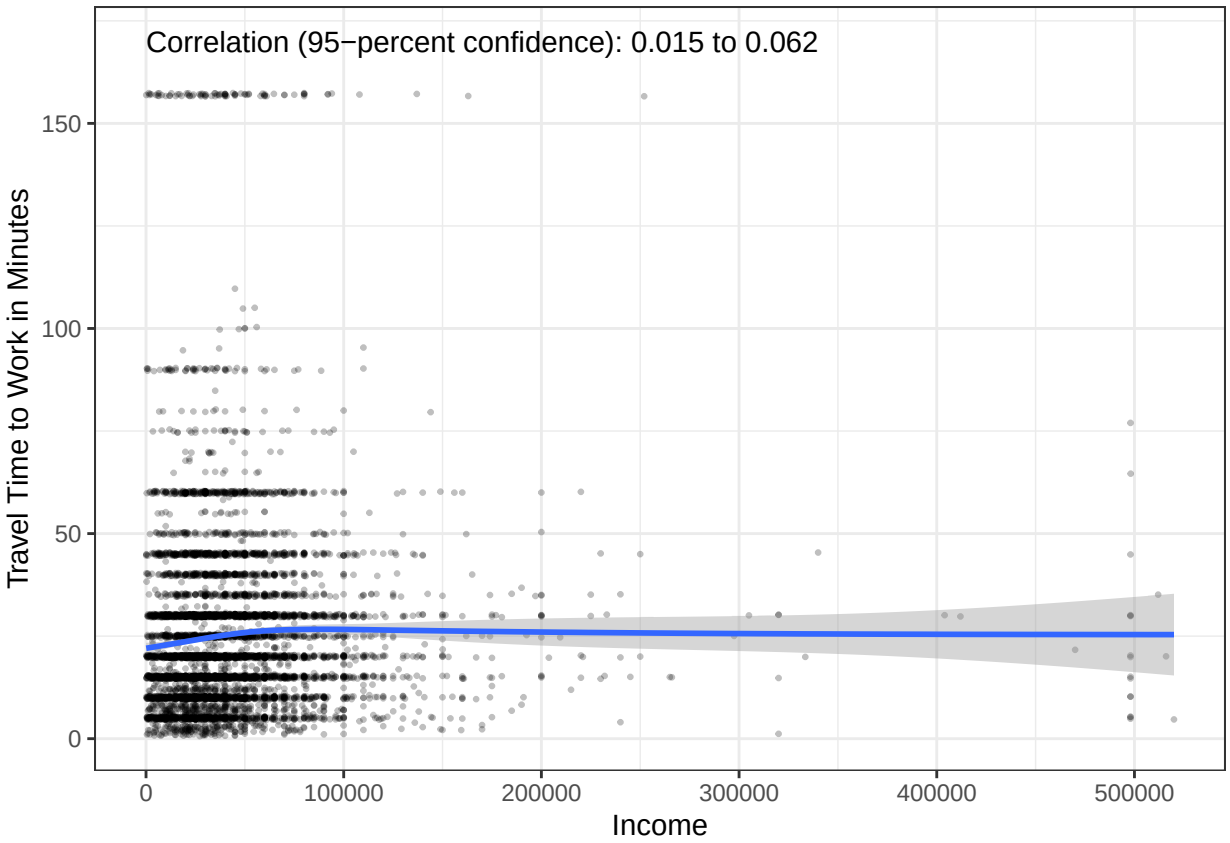
This correlation is weak with an r-value of -0.1244448. It is still statistically significant. The trend line suggests a general decline in the number of vehicles accessible as age increases, which makes some sense since senior citizens might choose to drive less as they get older.

Income and Travel Time to Work

```r
income_time_correlation <- cor.test(transpo_data$PINCP, transpo_data$JWMNP)

income_time_correlation
```

```
##
##  Pearson's product-moment correlation
##
## data:  transpo_data$PINCP and transpo_data$JWMNP
## t = 3.2073, df = 6912, p-value = 0.001346
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.01499023 0.06206430
## sample estimates:
##       cor
## 0.03854865
```

```r
options(scipen = 999)
ggplot(transpo_data, aes(x = PINCP, y = JWMNP)) +
  geom_point(size = 0.5, alpha = 0.25, position = "jitter") +
  geom_smooth() +
  theme_bw() +
  scale_x_continuous(name = "Income") +
  scale_y_continuous(name = "Travel Time to Work in Minutes") +
  annotate(geom = "text", x = 0, y = 170,
           label = paste("Correlation (95-percent confidence):",
                         prettyNum(income_time_correlation$conf.int[1], digits = 2),
                         "to",
                         prettyNum(income_time_correlation$conf.int[2], digits = 2)),
           hjust = 0)
```



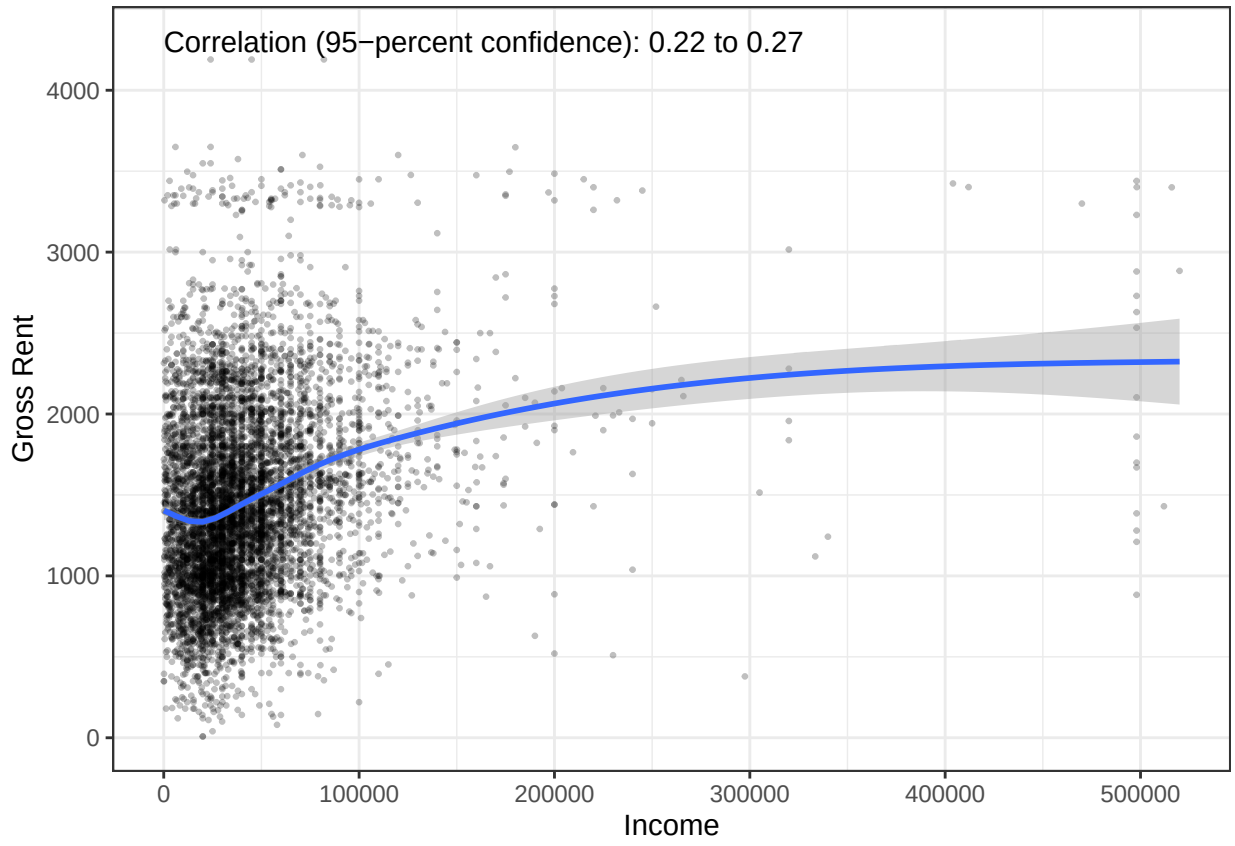
This correlation is weak (r = 0.03854865) though still statistically significant (p = 0.001346). The trendline is almost exactly flat, which means that a change in income does not necessarily predict a change in travel time to work.

Income and Gross Rent

```
income_rent_correlation <- cor.test(transpo_data$PINCP, transpo_data$GRNTP)

income_rent_correlation
```

```
##
##  Pearson's product-moment correlation
##
## data:  transpo_data$PINCP and transpo_data$GRNTP
## t = 20.869, df = 6912, p-value < 0.00000000000000022
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2211605 0.2655116
## sample estimates:
##       cor
## 0.2434633
```

```
options(scipen = 999)
ggplot(transpo_data, aes(x = PINCP, y = GRNTP)) +
  geom_point(size = 0.5, alpha = 0.25, position = "jitter") +
  geom_smooth() +
  theme_bw() +
  scale_x_continuous(name = "Income") +
  scale_y_continuous(name = "Gross Rent") +
  annotate(geom = "text", x = 0, y = 4300,
           label = paste("Correlation (95-percent confidence):",
                         prettyNum(income_rent_correlation$conf.int[1], digits = 2),
                         "to",
                         prettyNum(income_rent_correlation$conf.int[2], digits = 2)),
           hjust = 0)
```



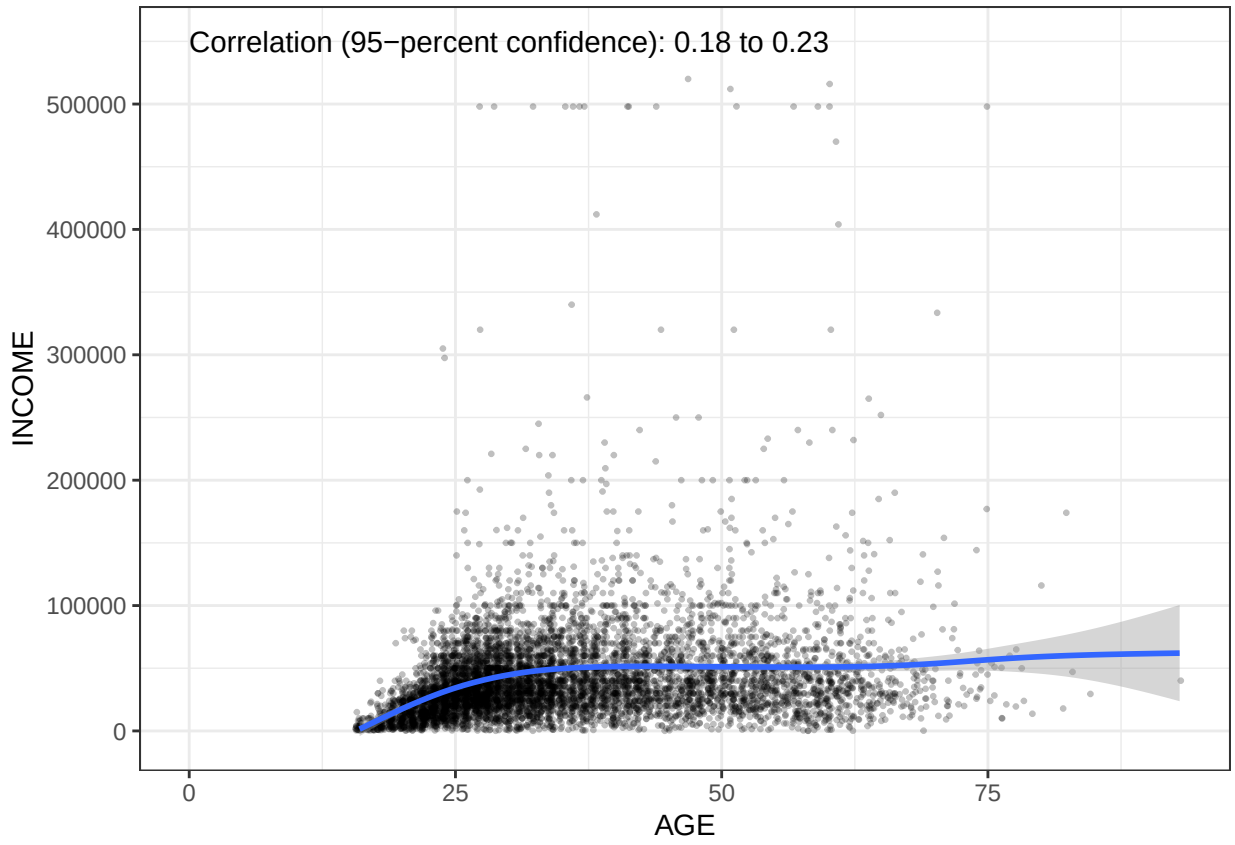
This correlation is fairly weak, though stronger than several of the ones I've found so far (r = 0.2434633). It is also statistically significant (p < 2.2e-16). This general relationship should be somewhat expected: the higher your income, the higher rent you can afford.

Income and Age

```
income_age_correlation <- cor.test(transpo_data$PINCP, transpo_data$AGEP)

income_age_correlation
```

```
##
##  Pearson's product-moment correlation
##
## data:  transpo_data$PINCP and transpo_data$AGEP
## t = 17.34, df = 6912, p-value < 0.00000000000000022
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1814715 0.2266514
## sample estimates:
##       cor
## 0.2041702
```

```
options(scipen = 999)
ggplot(transpo_data, aes(x = AGEP, y = PINCP)) +
  geom_point(size = 0.5, alpha = 0.25, position = "jitter") +
  geom_smooth() +
  theme_bw() +
  scale_x_continuous(name = "AGE") +
  scale_y_continuous(name = "INCOME",
                     breaks = seq(0, 600000, by = 100000)) +
  annotate(geom = "text", x = 0, y = 550000,
           label = paste("Correlation (95-percent confidence):",
                         prettyNum(income_age_correlation$conf.int[1], digits = 2),
                         "to",
                         prettyNum(income_age_correlation$conf.int[2], digits = 2)),
           hjust = 0)
```



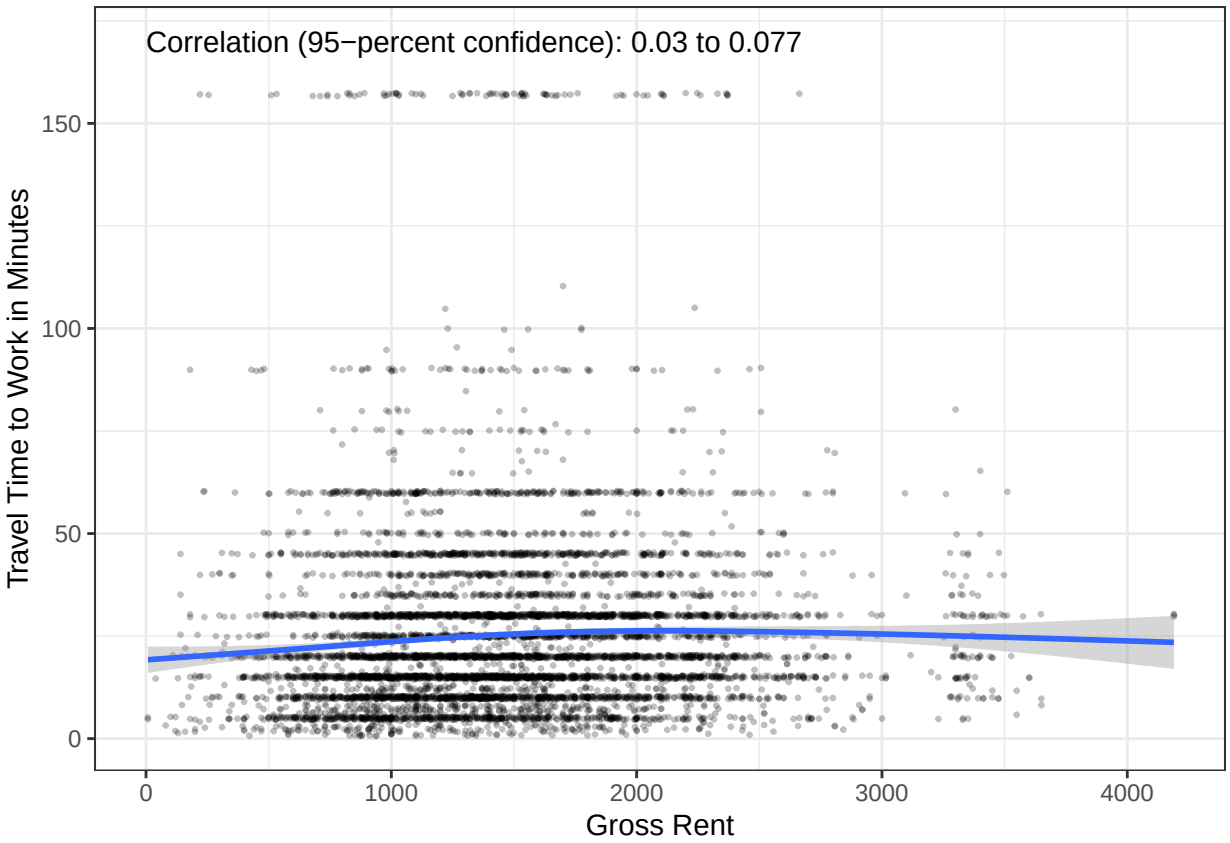
This correlation is fairly weak (r = 0.2041702), though it is statistically significant. The general trend makes some sense: it's expected that one's income will grow as they get older.

Travel time to work and Gross Rent

```
time_rent_correlation <- cor.test(transpo_data$GRNTP, transpo_data$JWMNP)

time_rent_correlation
```

```
##
##  Pearson's product-moment correlation
##
## data:  transpo_data$GRNTP and transpo_data$JWMNP
## t = 4.4475, df = 6912, p-value = 0.000008822
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.02988474 0.07689436
## sample estimates:
##        cor
## 0.05341915
```

```
options(scipen = 999)
ggplot(transpo_data, aes(x = GRNTP, y = JWMNP)) +
  geom_point(size = 0.5, alpha = 0.25, position = "jitter") +
  geom_smooth() +
  theme_bw() +
  scale_x_continuous(name = "Gross Rent") +
  scale_y_continuous(name = "Travel Time to Work in Minutes") +
  annotate(geom = "text", x = 0, y = 170,
           label = paste("Correlation (95-percent confidence):",
                         prettyNum(time_rent_correlation$conf.int[1], digits = 2),
                         "to",
                         prettyNum(time_rent_correlation$conf.int[2], digits = 2)),
           hjust = 0)
```



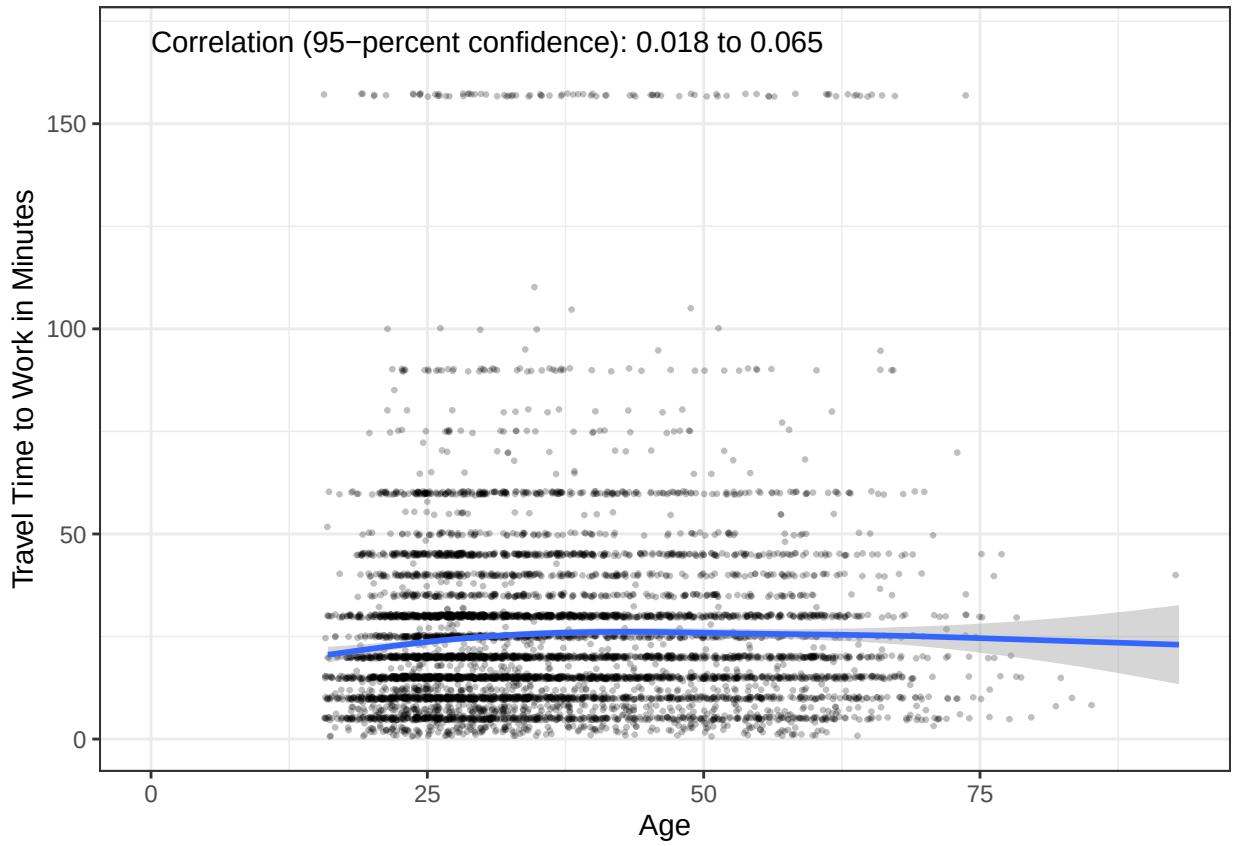
This correlation is weak (r = 0.05341915) though still statistically significant ($p < 0.5$). The trendline is almost exactly flat, which means that a change in rent does not necessarily predict a change in travel time to work.

Travel time to work and Age

```
time_age_correlation <- cor.test(transpo_data$AGEP, transpo_data$JWMNP)

time_age_correlation
```

```
##
##  Pearson's product-moment correlation
##
## data:  transpo_data$AGEP and transpo_data$JWMNP
## t = 3.4264, df = 6912, p-value = 0.0006152
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.01762312 0.06468731
## sample estimates:
##        cor
## 0.04117806
```

```
options(scipen = 999)
ggplot(transpo_data, aes(x = AGEP, y = JWMNP)) +
  geom_point(size = 0.5, alpha = 0.25, position = "jitter") +
  geom_smooth() +
  theme_bw() +
  scale_x_continuous(name = "Age") +
  scale_y_continuous(name = "Travel Time to Work in Minutes") +
  annotate(geom = "text", x = 0, y = 170,
           label = paste("Correlation (95-percent confidence):",
                         prettyNum(time_age_correlation$conf.int[1], digits = 2),
                         "to",
                         prettyNum(time_age_correlation$conf.int[2], digits = 2)),
           hjust = 0)
```



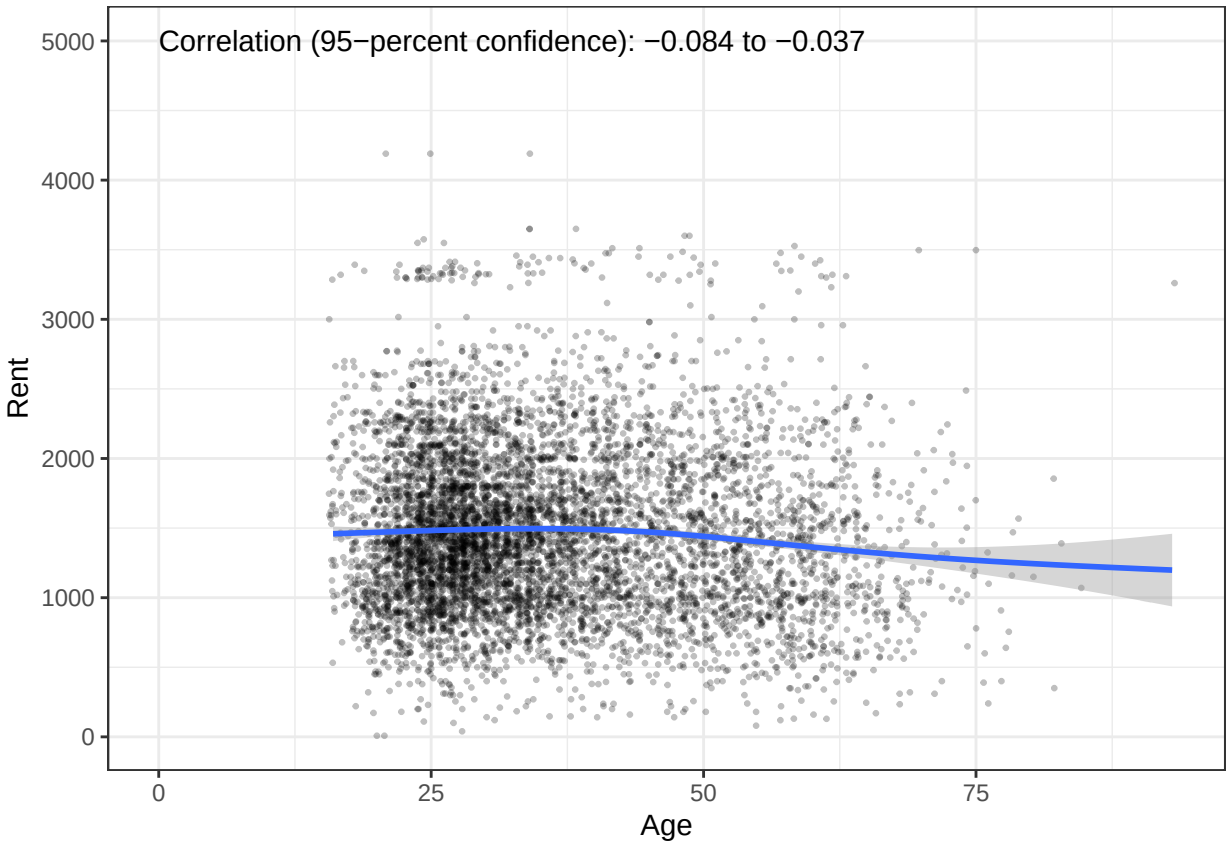
This correlation is weak (r = 0.04117806) though still statistically significant (p = 0.0006152). The trendline is almost exactly flat, which means that a change in age does not necessarily predict a change in travel time to work.

Age and Rent

```
age_rent_correlation <- cor.test(transpo_data$AGEP, transpo_data$GRNTP)

age_rent_correlation
```

```
##
##  Pearson's product-moment correlation
##
## data:  transpo_data$AGEP and transpo_data$GRNTP
## t = -5.0605, df = 6912, p-value = 0.0000004289
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.08420749 -0.03723733
## sample estimates:
##         cor
## -0.06075605
```

```
options(scipen = 999)
ggplot(transpo_data, aes(x = AGEP, y = GRNTP)) +
  geom_point(size = 0.5, alpha = 0.25, position = "jitter") +
  geom_smooth() +
  theme_bw() +
  scale_x_continuous(name = "Age") +
  scale_y_continuous(name = "Rent") +
  annotate(geom = "text", x = 0, y = 5000,
           label = paste("Correlation (95-percent confidence):",
                         prettyNum(age_rent_correlation$conf.int[1], digits = 2),
                         "to",
                         prettyNum(age_rent_correlation$conf.int[2], digits = 2)),
           hjust = 0)
```



This correlation is weak (r = -0.06075605) though still statistically significant ($p < 0.05$). The trendline is almost flat, with a slight decrease, which might suggest (not robustly) that as one gets older, they pay less for rent.

Relationship between a continuous variable and a binary variable

Sex and Number of Vehicles Accessible

By running this two sample t-test, I am attempting to answer: do males have access to more vehicles than females do?

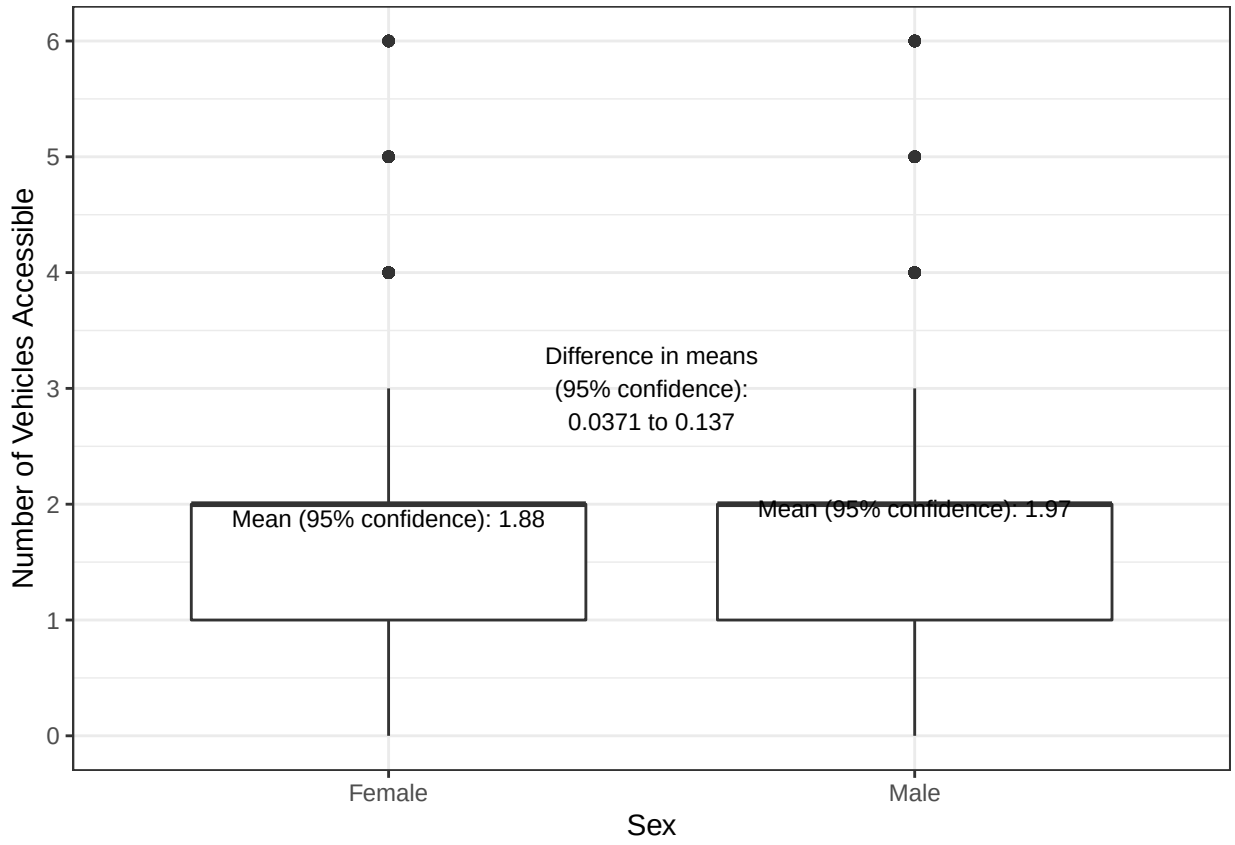
```
veh_difference = t.test(vehicle ~ SEX_label == "Male",
                        data = transpo_data)
veh_difference
```

```
##
##  Welch Two Sample t-test
##
## data:  vehicle by SEX_label == "Male"
## t = -3.4203, df = 6795.3, p-value = 0.0006293
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.13663313 -0.03707391
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            1.878664            1.965517
```

Males generally have access to more vehicles than females do. Males generally have access to 1.97 cars while females generally have access to 1.88 cars. This is statistically significant ($p = 0.0006293$). We can visualize this result with a box plot.

```
ggplot(transpo_data, aes(x = SEX_label, y = vehicle)) +
  geom_boxplot() +
  theme_bw() +
  scale_x_discrete(name = "Sex",
                   labels = c("Female", "Male")) +
   scale_y_continuous(name = "Number of Vehicles Accessible",
                      breaks = seq(0, 6, by = 1),
                      labels = paste(prettyNum(seq(0, 6, by = 1),
                                               big.mark = ","))) +
  annotate(geom = "text", x = 1.5, y = 3, size = 3,
           label = paste("Difference in means\n(95% confidence):\n",
                         prettyNum(abs(veh_difference$conf.int[2]),
                                   digits = 3), " to ",
                         prettyNum(abs(veh_difference$conf.int[1]),
                                   digits = 3),
                         sep = "")) +
  annotate(geom = "text", x = 1, y = veh_difference$estimate[1], size = 3,
           label = paste("Mean (95% confidence):",
                         prettyNum(veh_difference$estimate[1], digits = 3))) +
  annotate(geom = "text", x = 2, y = veh_difference$estimate[2], size = 3,
```

```
          label = paste("Mean (95% confidence):",
                        prettyNum(veh_difference$estimate[2], digits = 3)))
```



Although the difference in mean number of vehicles accessible by sex is statistically significant, it's practically not interesting since you can only have whole cars. Either value would likely be rounded up to 2 for practical purposes.

Sex and Income

By running this two sample t-test, I am attempting to answer: do males earn a higher income than females?

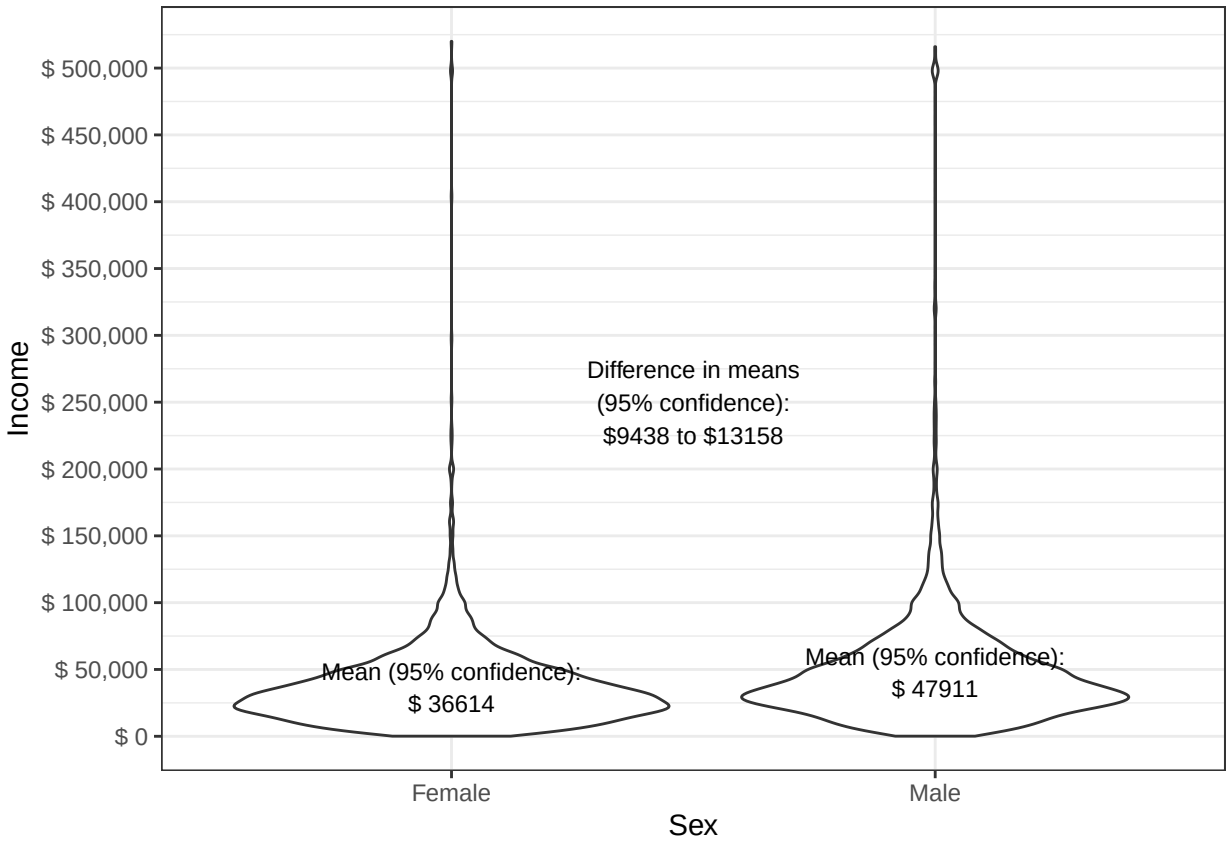
```
income_difference = t.test(PINCP ~ SEX_label == "Male",
                    data = transpo_data)
income_difference
```

```
##
##  Welch Two Sample t-test
##
## data:  PINCP by SEX_label == "Male"
## t = -11.907, df = 6705, p-value < 0.00000000000000022
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -13157.602  -9437.736
## sample estimates:
```

```
## mean in group FALSE  mean in group TRUE
##            36613.52            47911.19
```

Males generally earn a higher income than females do. Males, on average, earn 47,911.19 dollars while females earn 36,613.52 dollars. This is statistically significant ($p < 0.05$). We can visualize this result with a violin plot.

```
ggplot(transpo_data, aes(x = SEX_label, y = PINCP)) +
  geom_violin() +
  theme_bw() +
  scale_x_discrete(name = "Sex",
                   labels = c("Female", "Male")) +
   scale_y_continuous(name = "Income",
                      breaks = seq(0, 500000, by = 50000),
                      labels = paste("$",
                                     prettyNum(seq(0, 500000, by = 50000),
                                               big.mark = ","))) +
  annotate(geom = "text", x = 1.5, y = 250000, size = 3,
           label = paste("Difference in means\n(95% confidence):\n$",
                         prettyNum(abs(income_difference$conf.int[2]),
                                   digits = 0), " to $",
                         prettyNum(abs(income_difference$conf.int[1]),
                                   digits = 0),
                         sep = "")) +
  annotate(geom = "text", x = 1, y = income_difference$estimate[1], size = 3,
           label = paste("Mean (95% confidence):\n$",
                         prettyNum(income_difference$estimate[1], digits = 0))) +
  annotate(geom = "text", x = 2, y = income_difference$estimate[2], size = 3,
           label = paste("Mean (95% confidence):\n$",
                         prettyNum(income_difference$estimate[2], digits = 0)))
```



With the outliers in the data, the difference in means doesn't appear to be great, but according to the two-sample t-test, we can be about 95% confident that males earn between 9,438 and 13,158 more dollars than females, on average.

Sex and Travel Time to Work

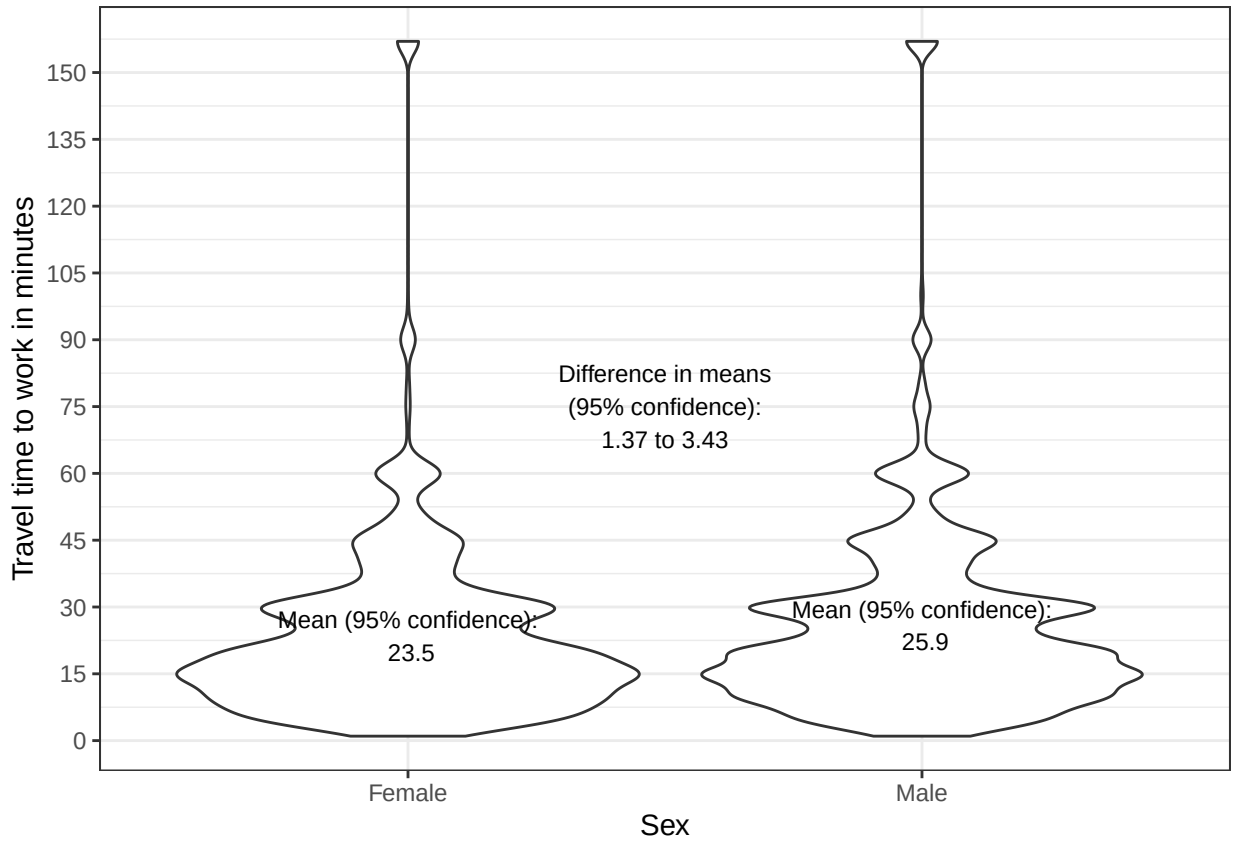
By running this two sample t-test, I am attempting to answer: do males have a longer travel time to work then females do?

```
time_difference = t.test(JWMNP ~ SEX_label == "Male",
                         data = transpo_data)
time_difference
```

```
##
##  Welch Two Sample t-test
##
## data:  JWMNP by SEX_label == "Male"
## t = -4.5831, df = 6871.9, p-value = 0.000004663
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.425799 -1.373147
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            23.48692            25.88639
```

Though just by little, males generally travel for a longer time than females when going to work. Males, on average, take 25.89 minutes to get to work, while females take 23.49 minutes to get to work. This is statistically significant ($p < 0.05$). We can visualize this result with a violin plot.

```
ggplot(transpo_data, aes(x = SEX_label, y = JWMNP)) +
  geom_violin() +
  theme_bw() +
  scale_x_discrete(name = "Sex",
                   labels = c("Female", "Male")) +
   scale_y_continuous(name = "Travel time to work in minutes",
                      breaks = seq(0, 150, by = 15),
                      labels = paste(prettyNum(seq(0, 150, by = 15),
                                               big.mark = ","))) +
  annotate(geom = "text", x = 1.5, y = 75, size = 3,
           label = paste("Difference in means\n(95% confidence):\n",
                         prettyNum(abs(time_difference$conf.int[2]),
                                   digits = 3), " to ",
                         prettyNum(abs(time_difference$conf.int[1]),
                                   digits = 3),
                         sep = "")) +
  annotate(geom = "text", x = 1, y = time_difference$estimate[1], size = 3,
           label = paste("Mean (95% confidence):\n",
                         prettyNum(time_difference$estimate[1], digits = 3))) +
  annotate(geom = "text", x = 2, y = time_difference$estimate[2], size = 3,
           label = paste("Mean (95% confidence):\n",
                         prettyNum(time_difference$estimate[2], digits = 3)))
```



With the outliers in the data, the difference in means doesn't appear to be great. Indeed, on average, males only travel for about 2.4 minutes longer than females to work.

Sex and Rent

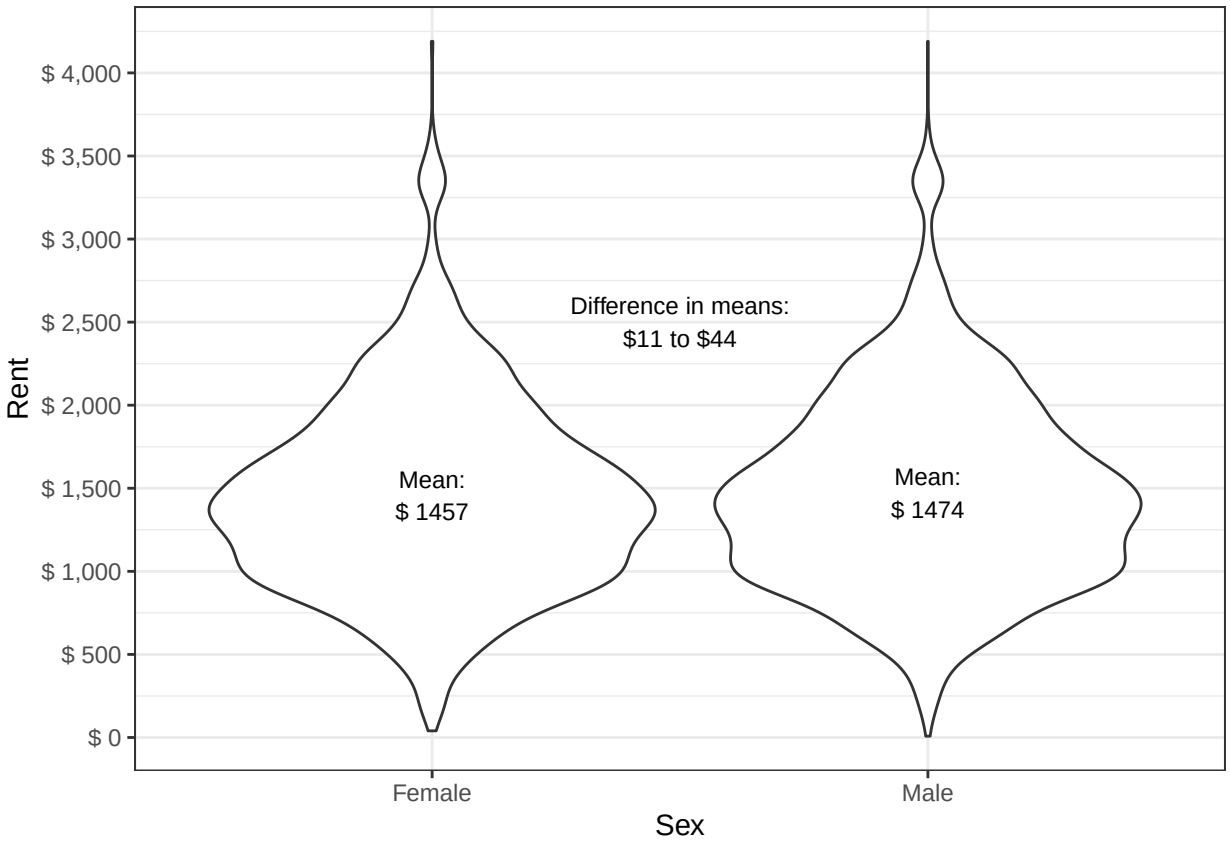
By running this two sample t-test, I am attempting to answer: do males payer higher rents then females do?

```
rent_difference = t.test(GRNTP ~ SEX_label == "Male",
                     data = transpo_data)
rent_difference
```

```
##
##  Welch Two Sample t-test
##
## data:  GRNTP by SEX_label == "Male"
## t = -1.1739, df = 6737.1, p-value = 0.2405
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -43.95413  11.02898
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            1457.158            1473.620
```

The difference in means is not statstically significant (p = 0.2405). This means that while the mean rent for males is 1473.62 and the mean rent for females is 1457.16, we can't be at least 95% confident that this difference didn't just occur by chance. In fact, if we look at a violin plot, we can see that the distributions and means are almost identical.

```
ggplot(transpo_data, aes(x = SEX_label, y = GRNTP)) +
  geom_violin() +
  theme_bw() +
  scale_x_discrete(name = "Sex",
                   labels = c("Female", "Male")) +
   scale_y_continuous(name = "Rent",
                      breaks = seq(0, 5000, by = 500),
                      labels = paste("$",
                                     prettyNum(seq(0, 5000, by = 500),
                                               big.mark = ","))) +
  annotate(geom = "text", x = 1.5, y = 2500, size = 3,
           label = paste("Difference in means:\n$",
                         prettyNum(abs(rent_difference$conf.int[2]),
                                   digits = 0), " to $",
                         prettyNum(abs(rent_difference$conf.int[1]),
                                   digits = 0),
                         sep = "")) +
  annotate(geom = "text", x = 1, y = rent_difference$estimate[1], size = 3,
           label = paste("Mean:\n$",
                         prettyNum(rent_difference$estimate[1], digits = 0))) +
  annotate(geom = "text", x = 2, y = rent_difference$estimate[2], size = 3,
           label = paste("Mean:\n$",
                         prettyNum(rent_difference$estimate[2], digits = 0)))
```



Sex and Age

By running this two sample t-test, I am attempting to answer: are males older than females? I have a strong suspicion that this difference in means will not be statistically significant.

```
age_difference = t.test(AGEP ~ SEX_label == "Male",
                        data = transpo_data)
age_difference
```

```
##
##  Welch Two Sample t-test
##
## data:  AGEP by SEX_label == "Male"
## t = 0.50473, df = 6593.2, p-value = 0.6138
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.4504762  0.7628869
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            35.97951            35.82331
```

As I predicted, the difference in means of ages between males and females is not statistically significant (p = 0.6138). We cannot be at least 95% sure that the difference didn't just occur by chance. This makes sense, as nowadays sex (at least in aggregate - this likely is different if we examine sex and race or sex and socioeconomic status) doesn't necessarily affect how old one grows to be.

Relationship between a continuous variable and a categorical variable (with three or more levels)

Means of transportation to work and Number of Accessible Vehicles

To see if there's a statistically significant association between the number of vehicles one has access to and one's means of transportation, we can do an analysis of variance.

```
means_veh_anova <- aov(vehicle ~ meansTW, data = transpo_data)

summary(means_veh_anova)
```

```
##               Df Sum Sq Mean Sq F value              Pr(>F)
## meansTW        9    294   32.72   30.47 <0.0000000000000002 ***
## Residuals   6904   7413    1.07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

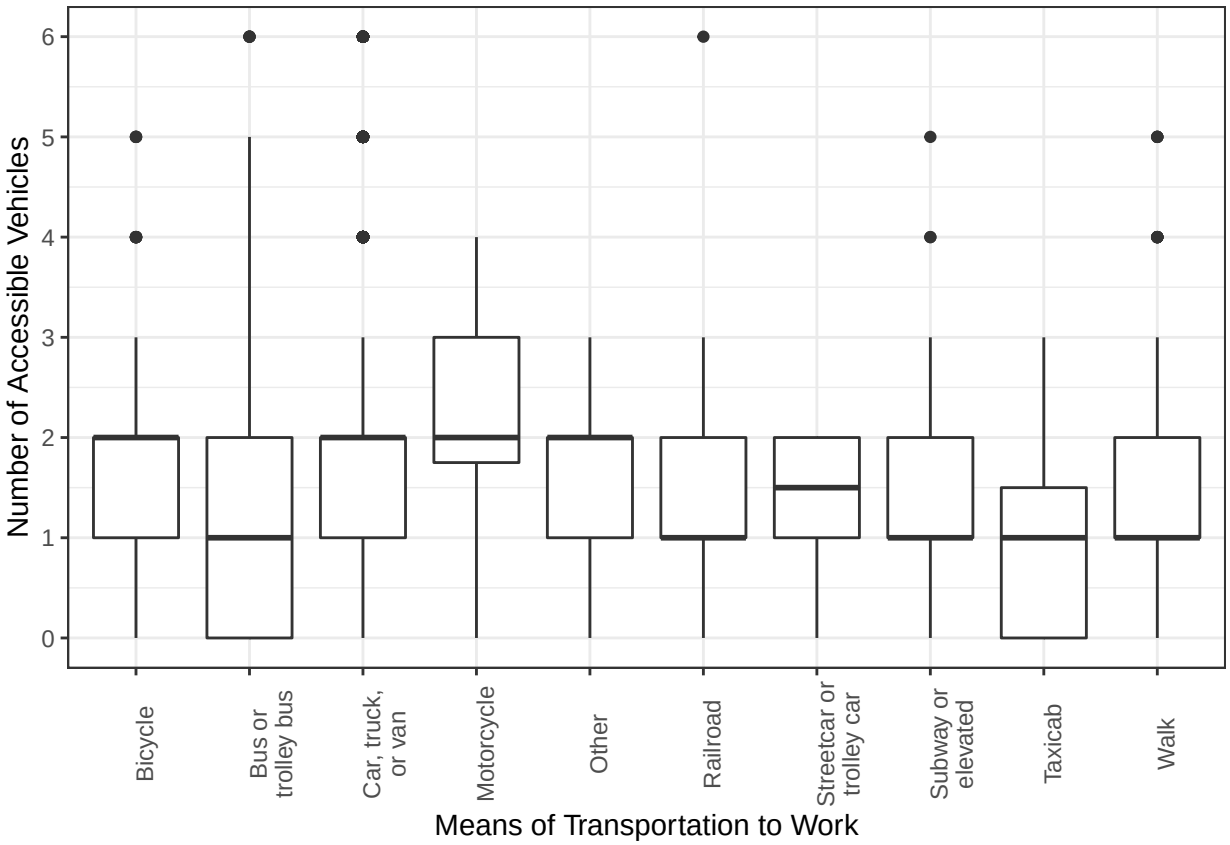
The p-value is much less than 0.05, so we can be 95% confident that there is a significant association between the means of transportation one takes to work and the number of vehicles one has access to. Tukey's Honestly Significant Difference test shows us the magnitude of the differences among the different categories.

```
means_veh_differences <- TukeyHSD(means_veh_anova)

as_tibble(cbind(pair = row.names(means_veh_differences$meansTW),
                means_veh_differences$meansTW))
```

```
## # A tibble: 45 x 5
##    pair                   diff         lwr          upr         `p adj`
##    <chr>                  <chr>        <chr>        <chr>       <chr>
##  1 "Bus or\ntrolley bus-~ -0.48204621~ -0.84823583~ -0.1158566~ 0.0012956125404~
##  2 "Car, truck,\nor van-~ 0.285170995~ -0.00662020~ 0.57696219~ 0.0618462911305~
##  3 "Motorcycle-Bicycle"   0.474321705~ -0.39488040~ 1.34352381~ 0.7802322642717~
##  4 "Other-Bicycle"        -0.04104714~ -0.55062276~ 0.46852847~ 0.99999992127209
##  5 "Railroad-Bicycle"     -0.30692829~ -0.95456985~ 0.34071326~ 0.8926766984664~
##  6 "Streetcar or\ntrolle~ -0.31317829~ -1.38965382~ 0.76329723~ 0.99584851056228
##  7 "Subway or\nelevated-~ -0.27200182~ -0.90419689~ 0.36019324~ 0.9386794293834~
##  8 "Taxicab-Bicycle"      -0.60791513~ -1.41375884~ 0.19792856~ 0.3334113735028~
##  9 "Walk-Bicycle"         -0.37207042~ -0.71077471~ -0.0333661~ 0.0183251405639~
## 10 "Car, truck,\nor van-~ 0.767217214~ 0.538081481~ 0.99635294~ 0.0000000000042~
## # ... with 35 more rows
```

We can visualize these differences with a box plot.

```
ggplot(transpo_data, aes(x = meansTW, y = vehicle)) +
  geom_boxplot() +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_x_discrete(name = "Means of Transportation to Work") +
  scale_y_continuous(name = "Number of Accessible Vehicles",
                     breaks = seq(0, 6, by = 1),
                     labels = paste(prettyNum(seq(0, 6, by = 1))))
```



I find it interesting that the Bus or Trolley Bus category has the largest distribution of accessible vehicles. I also find it interesting that the Motorcycle category also has a larger distribution than the Car, truck, or van category.

Means of transportation to work and Income

```
means_income_anova <- aov(PINCP ~ meansTW, data = transpo_data)

summary(means_income_anova)
```

```
##               Df         Sum Sq    Mean Sq F value Pr(>F)
## meansTW        9    34404094709 3822677190   2.302 0.0141 *
## Residuals   6904 11467191892516 1660948999
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

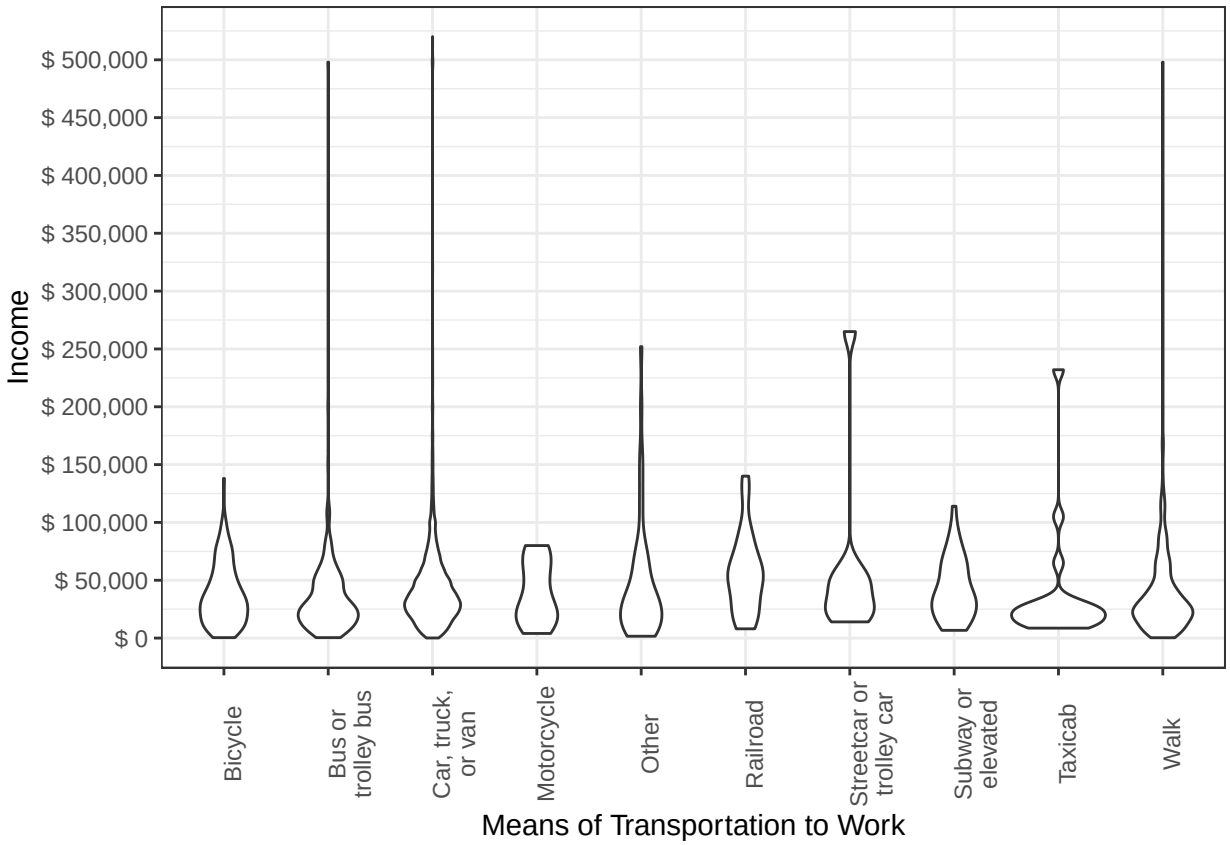
Since the p-value is less than 0.05, we can be 95% confident that there is a significant association between the means of transportation one takes to work and one's income. Tukey's Honestly Significant Difference test shows us the magnitude of the differences among the different categories.

```
means_income_differences <- TukeyHSD(means_income_anova)

as_tibble(cbind(pair = row.names(means_income_differences$meansTW),
                means_income_differences$meansTW))
```

```
## # A tibble: 45 x 5
##    pair                         diff         lwr          upr         `p adj`
##    <chr>                        <chr>        <chr>        <chr>       <chr>
##  1 "Bus or\ntrolley bus-Bicy~ -2109.26649~ -16511.4577~ 12292.9248~ 0.999985048~
##  2 "Car, truck,\nor van-Bicy~ 5853.218674~ -5622.89231~ 17329.3296~ 0.841631122~
##  3 "Motorcycle-Bicycle"       2008.929263~ -32176.6801~ 36194.5386~ 0.999999995~
##  4 "Other-Bicycle"            8669.021476~ -11372.5235~ 28710.5665~ 0.936671450~
##  5 "Railroad-Bicycle"         18481.42926~ -6990.23168~ 43953.0902~ 0.391441517~
##  6 "Streetcar or\ntrolley ca~ 21217.05426~ -21120.5926~ 63554.7011~ 0.855287534~
##  7 "Subway or\nelevated-Bicy~ 5267.642498~ -19596.5099~ 30131.7949~ 0.999662834~
##  8 "Taxicab-Bicycle"          487.0542635~ -31206.6771~ 32180.7857~ 1
##  9 "Walk-Bicycle"             503.8297737~ -12817.3673~ 13825.0269~ 0.999999999~
## 10 "Car, truck,\nor van-Bus ~ 7962.485165~ -1049.39435~ 16974.3646~ 0.137745555~
## # ... with 35 more rows
```

I can also represent these differences with a violin plot.

```
ggplot(transpo_data, aes(x = meansTW, y = PINCP)) +
  geom_violin() +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_x_discrete(name = "Means of Transportation to Work") +
  scale_y_continuous(name = "Income",
                     breaks = seq(0, 500000, by = 50000),
                     labels = paste("$",
                                    prettyNum(seq(0, 500000, by = 50000),
                                              big.mark = ",")))
```



I find it interesting that the few outliers that exist in the data set are not only found in the Car, truck, or van category (which makes sense, given that private vehicles are relatively expensive), but they are also found in the Bus or trolley bus and Walk categories (which are fairly cheap or free).

Means of transportation to work and Travel time to work

```
means_time_anova <- aov(JWMNP ~ meansTW, data = transpo_data)

summary(means_time_anova)
```

```
##               Df  Sum Sq Mean Sq F value              Pr(>F)
## meansTW        9  174081   19342   42.58 <0.0000000000000002 ***
## Residuals   6904 3136051     454
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

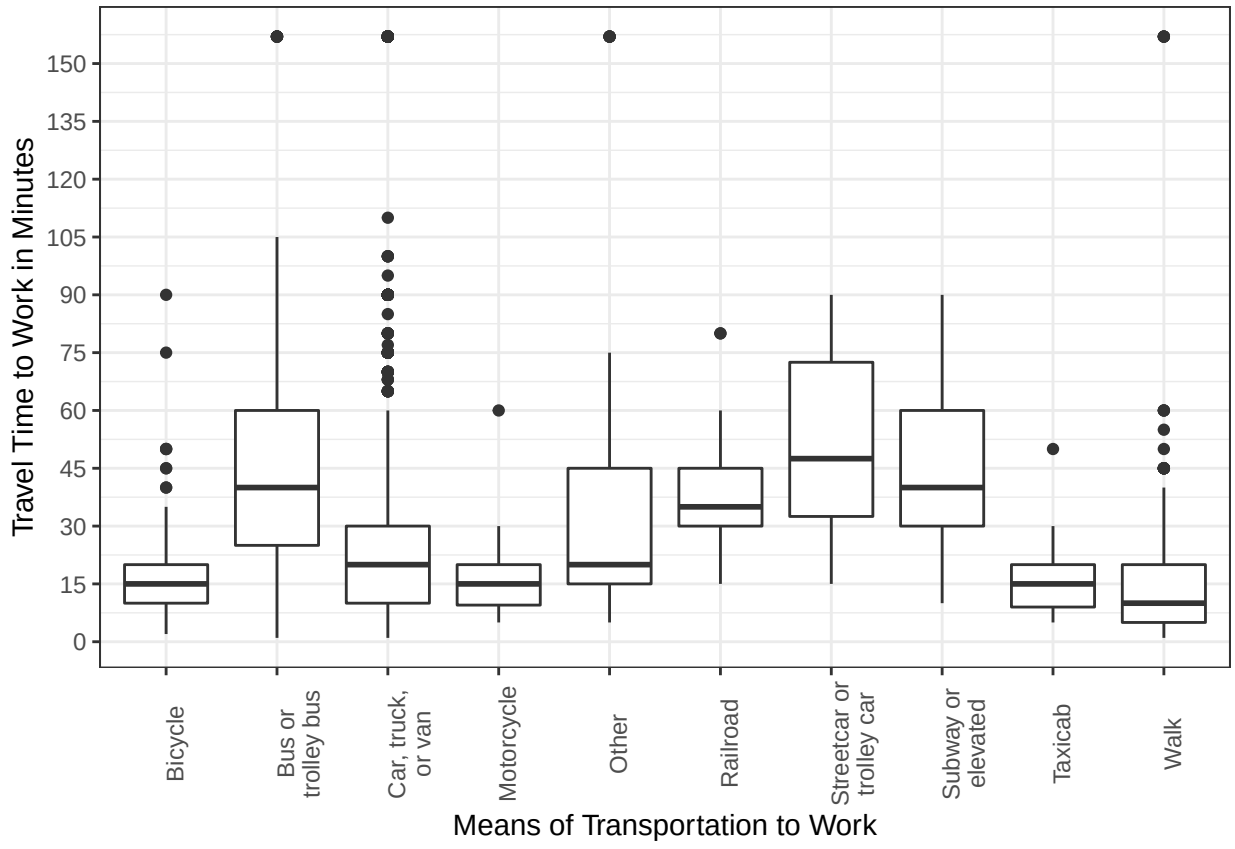
Since the p-value is less than 0.05, we can be 95% confident that there is a significant association between the means of transportation and the time it takes to get to work. Tukey's Honestly Significant Difference test shows us the magnitude of the differences among the different categories.

```
means_time_differences <- TukeyHSD(means_time_anova)

as_tibble(cbind(pair = row.names(means_time_differences$meansTW),
                means_time_differences$meansTW))
```

```
## # A tibble: 45 x 5
##    pair                    diff        lwr         upr         `p adj`
##    <chr>                   <chr>       <chr>       <chr>       <chr>
##  1 "Bus or\ntrolley bus-B~ 28.7219540~ 21.1902783~ 36.2536298~ 0.00000000000417~
##  2 "Car, truck,\nor van-B~ 7.19085532~ 1.18938340~ 13.1923272~ 0.00585046150564~
##  3 "Motorcycle-Bicycle"    -0.1153100~ -17.992792~ 17.7621725~ 1
##  4 "Other-Bicycle"         21.9728046~ 11.4920093~ 32.4535999~ 0.00000000157607~
##  5 "Railroad-Bicycle"      21.8221899~ 8.50169676~ 35.1426830~ 0.00000986190328~
##  6 "Streetcar or\ntrolley~ 33.2596899~ 11.1190710~ 55.4003087~ 0.00008850596362~
##  7 "Subway or\nelevated-B~ 25.6126310~ 12.6098366~ 38.6154255~ 0.00000002165434~
##  8 "Taxicab-Bicycle"       -1.1876784~ -17.762024~ 15.3866679~ 0.999999972195923
##  9 "Walk-Bicycle"          -2.0508057~ -9.0171718~ 4.91556040~ 0.995461793462294
## 10 "Car, truck,\nor van-B~ -21.531098~ -26.243892~ -16.818305~ 0.00000000000417~
## # ... with 35 more rows
```

I can also represent these differences with a box plot.

```
ggplot(transpo_data, aes(x = meansTW, y = JWMNP)) +
  geom_boxplot() +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_x_discrete(name = "Means of Transportation to Work") +
  scale_y_continuous(name = "Travel Time to Work in Minutes",
                     breaks = seq(0, 150, by = 15),
                     labels = paste(prettyNum(seq(0, 150, by = 15))))
```



Although the median travel times for cars, trucks, vans, and bicycles are lower than public transportation, they also have several outliers for longer trip times. I suspect the outlier in the walk category might be a mistake. The median travel times are highest for buses, streetcars, and subways. In the state of Colorado, which is where my sample exists, this makes sense bcause our cities are rarely built densely enough so that public transit is actually more efficient than driving.

Means of transportation to work and Rent

```
means_rent_anova <- aov(GRNTP ~ meansTW, data = transpo_data)

summary(means_rent_anova)
```

```
##               Df     Sum Sq Mean Sq F value  Pr(>F)
## meansTW        9    9268619 1029847   3.056 0.00117 **
## Residuals   6904 2326520636  336982
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value is less than 0.05, we can be 95% confident that there is a significant association between the means of transportation to work and one's rent. Tukey's Honestly Significant Difference test shows us the magnitude of the differences among the different categories.

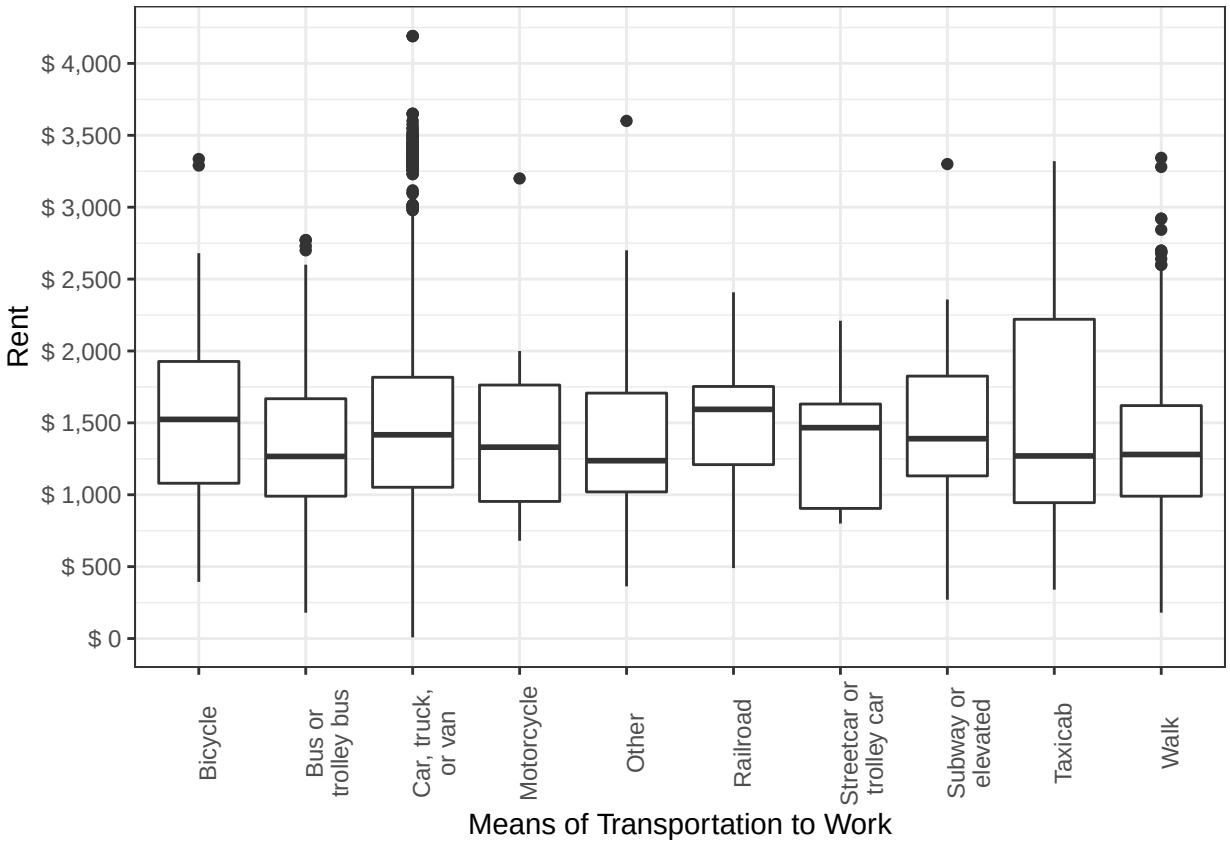
```
means_rent_differences <- TukeyHSD(means_rent_anova)

as_tibble(cbind(pair = row.names(means_rent_differences$meansTW),
                means_rent_differences$meansTW))
```

```
## # A tibble: 45 x 5
##    pair                      diff         lwr         upr          `p adj`
##    <chr>                     <chr>        <chr>       <chr>        <chr>
##  1 "Bus or\ntrolley bus-Bic~ -205.748793~ -410.89027~ -0.60731075~ 0.0486062845~
##  2 "Car, truck,\nor van-Bic~ -65.9336261~ -229.39669~ 97.52944510~ 0.9589329333~
##  3 "Motorcycle-Bicycle"      -120.168604~ -607.10051~ 366.7633105~ 0.9988420015~
##  4 "Other-Bicycle"           -141.057948~ -426.52508~ 144.4091843~ 0.8655538224~
##  5 "Railroad-Bicycle"        1.831395348~ -360.98105~ 364.6438437~ 1
##  6 "Streetcar or\ntrolley c~ -173.918604~ -776.96625~ 429.1290475~ 0.9961140102~
##  7 "Subway or\nelevated-Bic~ -84.8009575~ -438.96019~ 269.3582803~ 0.9990899041~
##  8 "Taxicab-Bicycle"         -49.2080783~ -500.64626~ 402.2301034~ 0.9999988587~
##  9 "Walk-Bicycle"            -180.535222~ -370.27927~ 9.208829015~ 0.0781066851~
## 10 "Car, truck,\nor van-Bus~ 139.8151672~ 11.4520392~ 268.1782951~ 0.0202553564~
## # ... with 35 more rows
```

I can also represent these differences with a box plot.

```
ggplot(transpo_data, aes(x = meansTW, y = GRNTP)) +
  geom_boxplot() +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_x_discrete(name = "Means of Transportation to Work") +
  scale_y_continuous(name = "Rent",
```

```
                    breaks = seq(0, 5000, by = 500),
                    labels = paste("$",
                                   prettyNum(seq(0, 5000, by = 500),
                                             big.mark = ",")))
```



There doesn't appear, visually at least, to be huge differences among the medians. However, the Car, truck, or van category has the most outliers for highest rents. Interestingly the highest median rent is in the railroad category.

Means of transportation to work and Age

```
means_age_anova <- aov(AGEP ~ meansTW, data = transpo_data)

summary(means_age_anova)
```

```
##               Df  Sum Sq Mean Sq F value      Pr(>F)
## meansTW        9    7412   823.6   5.082 0.000000712 ***
## Residuals   6904 1118894   162.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

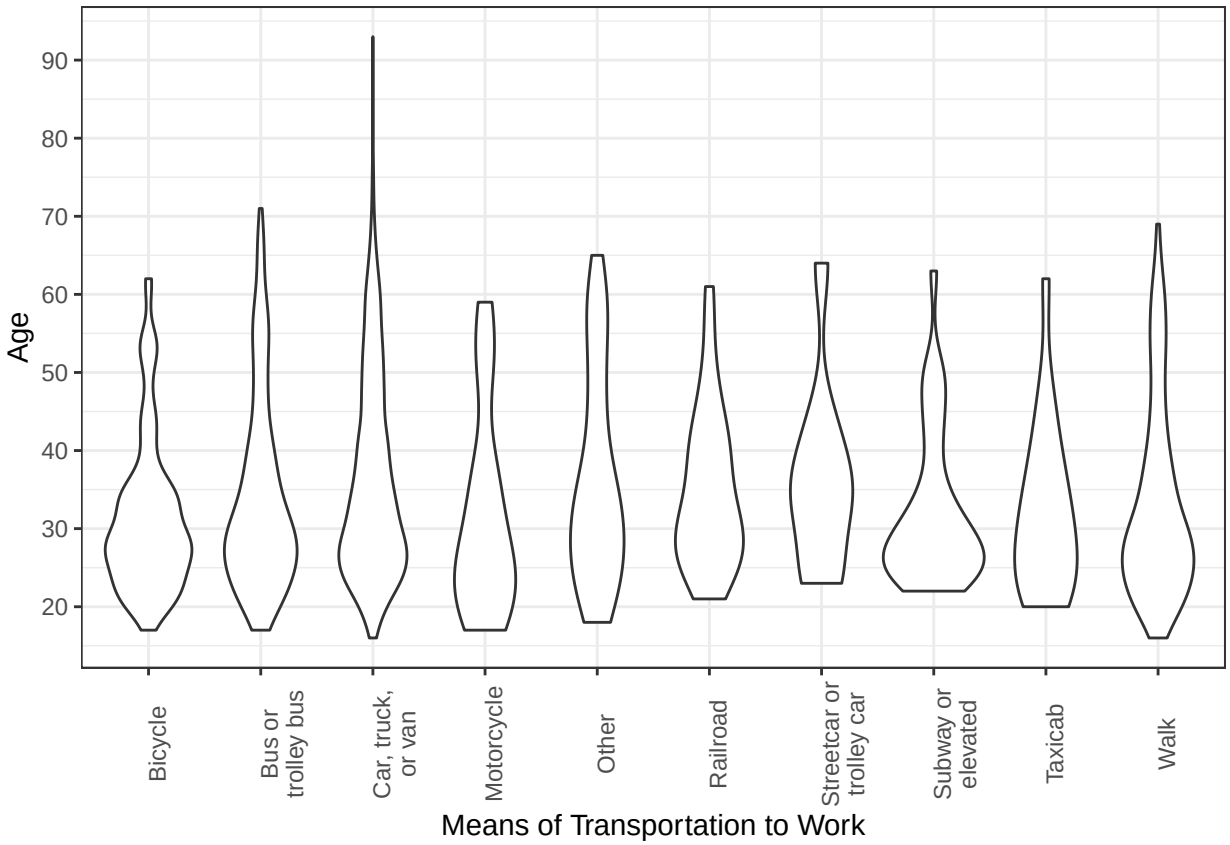
Since the p-value is less than 0.05, we can be 95% confident that there is a significant association between the means of transportation to work and one's age. Tukey's Honestly Significant Difference test shows us the magnitude of the differences among the different categories.

```r
means_age_differences <- TukeyHSD(means_age_anova)

as_tibble(cbind(pair = row.names(means_age_differences$meansTW),
                means_age_differences$meansTW))
```

```
## # A tibble: 45 x 5
##    pair                       diff         lwr          upr         `p adj`
##    <chr>                      <chr>        <chr>        <chr>       <chr>
##  1 "Bus or\ntrolley bus-Bic~ 2.618765540~ -1.88001128~ 7.11754236~ 0.7080878635~
##  2 "Car, truck,\nor van-Bic~ 4.745546282~ 1.160781918~ 8.33031064~ 0.0011722825~
##  3 "Motorcycle-Bicycle"      -0.19137596~ -10.8698501~ 10.4870981~ 1
##  4 "Other-Bicycle"           4.463337145~ -1.79699063~ 10.7236649~ 0.4181029082~
##  5 "Railroad-Bicycle"        3.339874031~ -4.61664560~ 11.2963936~ 0.9472629836~
##  6 "Streetcar or\ntrolley c~ 4.396124031~ -8.82878186~ 17.6210299~ 0.9889288743~
##  7 "Subway or\nelevated-Bic~ 0.731418148~ -7.03533556~ 8.49817186~ 0.9999996855~
##  8 "Taxicab-Bicycle"         0.496124031~ -9.40396835~ 10.3962164~ 0.9999999988~
##  9 "Walk-Bicycle"            1.755599249~ -2.40551010~ 5.91670860~ 0.9455836605~
## 10 "Car, truck,\nor van-Bus~ 2.126780742~ -0.68823774~ 4.94179922~ 0.3312297556~
## # ... with 35 more rows
```

I can also represent this with a violin plot.

```r
ggplot(transpo_data, aes(x = meansTW, y = AGEP)) +
  geom_violin() +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_x_discrete(name = "Means of Transportation to Work") +
  scale_y_continuous(name = "Age",
                     breaks = seq(0, 90, by = 10),
                     labels = paste(prettyNum(seq(0, 90, by = 10))))
```



The greatest spread of ages exists in the car, truck, or van category. It seems that generally, the mean ages are higher for the streetcar, railroad, and car, truck, or van categories.

Educational Attainment and Number of Accessible Vehicles

```
ed_veh_anova <- aov(vehicle ~ edu, data = transpo_data)

summary(ed_veh_anova)
```

```
##               Df Sum Sq Mean Sq F value       Pr(>F)
## edu           23     86   3.723   3.365 0.0000000933 ***
## Residuals   6890   7622   1.106
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

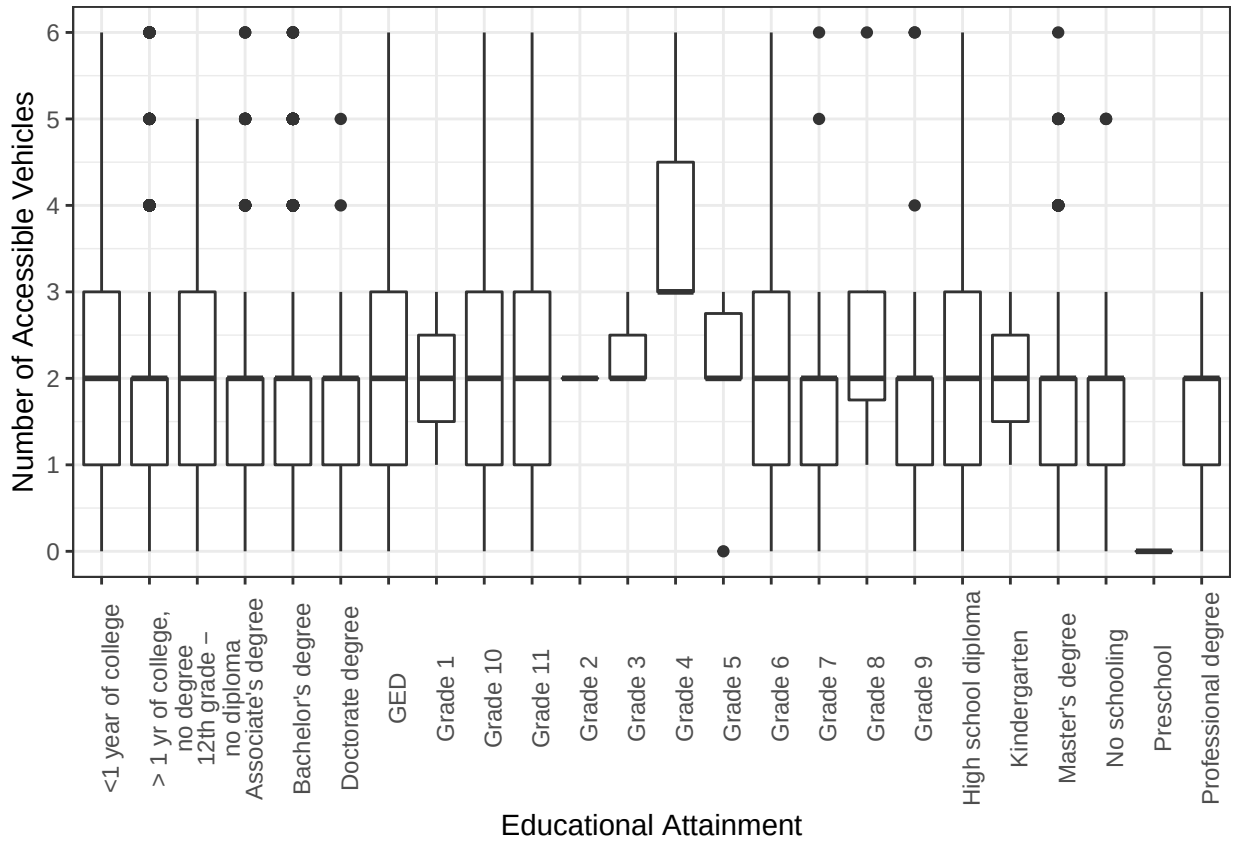
Since the p-value is less than 0.05, we can be 95% confident that there is a significant association between one's educational attainment and the number of accessible vehicles. Tukey's Honestly Significant Difference test shows us the magnitude of the differences among the different categories.

```
ed_veh_differences <- TukeyHSD(ed_veh_anova)

as_tibble(cbind(pair = row.names(ed_veh_differences$edu),
                ed_veh_differences$edu))
```

```
## # A tibble: 276 x 5
##    pair                          diff          lwr          upr          `p adj`
##    <chr>                         <chr>         <chr>        <chr>        <chr>
##  1 "> 1 yr of college,\nno de~ 0.0070356797~ -0.2020488~ 0.21612024~ 1
##  2 "12th grade -\nno diploma--~ -0.026329787~ -0.4287553~ 0.37609575~ 1
##  3 "Associate's degree-<1 yea~ -0.062676574~ -0.2978274~ 0.17247427~ 0.99999929~
##  4 "Bachelor's degree-<1 year~ -0.076152578~ -0.2740201~ 0.12171499~ 0.99952198~
##  5 "Doctorate degree-<1 year ~ -0.351584889~ -0.7765900~ 0.07342025~ 0.28778706~
##  6 "GED-<1 year of college"    0.0004559270~ -0.2801181~ 0.28102997~ 1
##  7 "Grade 1-<1 year of colleg~ 0.0361702127~ -2.6758258~ 2.74816627~ 1
##  8 "Grade 10-<1 year of colle~ 0.1240823006~ -0.3142402~ 0.56240482~ 0.99999776~
##  9 "Grade 11-<1 year of colle~ 0.0495035460~ -0.3094045~ 0.40841165~ 0.99999999~
## 10 "Grade 2-<1 year of colleg~ 0.0361702127~ -2.6758258~ 2.74816627~ 1
## # ... with 266 more rows
```

I can also represent this with a box plot.

```
ggplot(transpo_data, aes(x = edu, y = vehicle)) +
  geom_boxplot() +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_x_discrete(name = "Educational Attainment") +
  scale_y_continuous(name = "Number of Accessible Vehicles",
                     breaks = seq(0, 6, by = 1),
                     labels = paste(prettyNum(seq(0, 6, by = 1))))
```



For some odd reason, people who only ever went to school through grade 4 have the highest median number of cars. According to my dataset, only 3 people in my entire sample only completed school through grade 4, so those three people alone are making it appear that people who only completed school through 4th grade might have access to more cars than most people.

Educational Attainment and Income

```
ed_income_anova <- aov(PINCP ~ edu, data = transpo_data)

summary(ed_income_anova)
```

```
##               Df         Sum Sq     Mean Sq F value               Pr(>F)
## edu           23  1226400780121 53321773049   35.76 <0.0000000000000002 ***
## Residuals   6890 10275195207105  1491320059
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value is less than 0.05, we can be 95% confident that there is a significant association between one's educational attainment and their income. Tukey's Honestly Significant Difference test shows us the magnitude of the differences among the different categories.

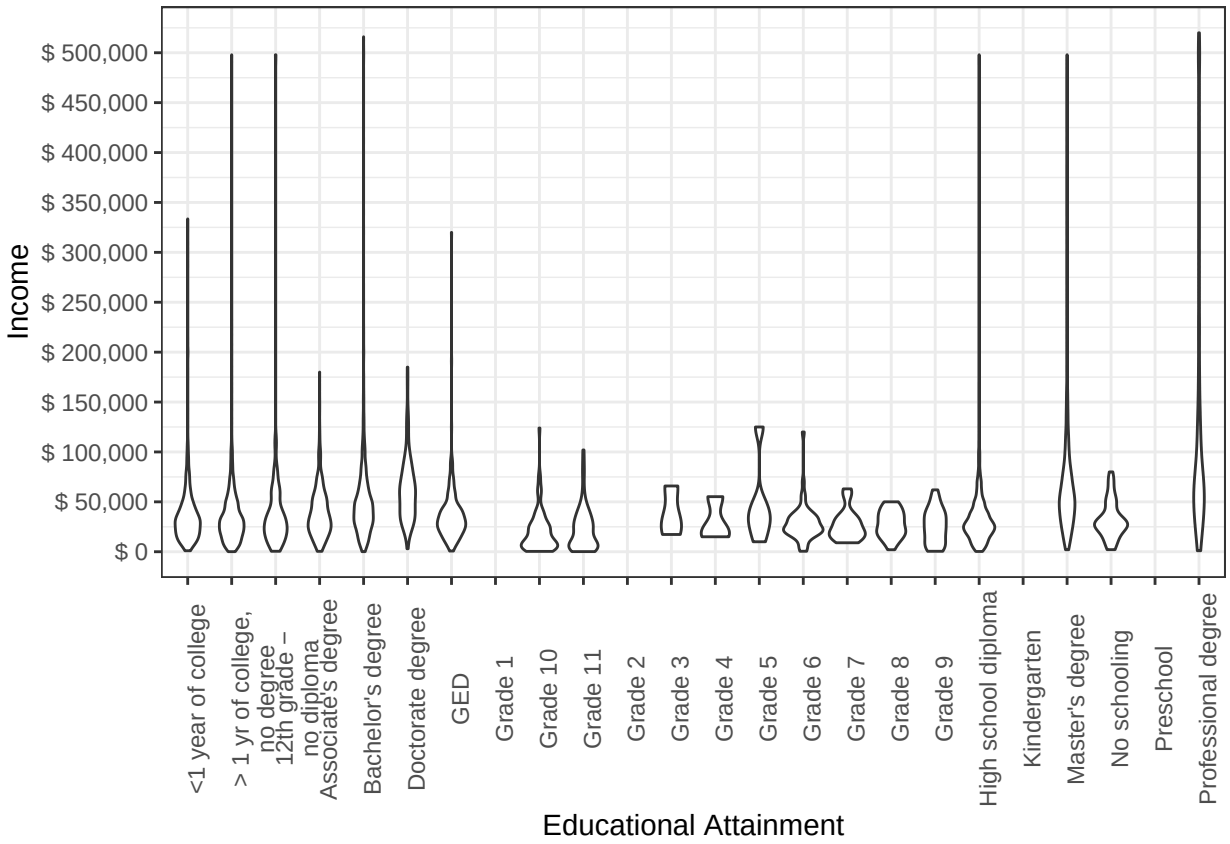
```
ed_income_differences <- TukeyHSD(ed_income_anova)

as_tibble(cbind(pair = row.names(ed_income_differences$edu),
                ed_income_differences$edu))
```

```
## # A tibble: 276 x 5
##    pair                        diff        lwr        upr        `p adj`
##    <chr>                       <chr>       <chr>      <chr>      <chr>
##  1 "> 1 yr of college,\nno d~ -2054.1360~ -9730.897~ 5622.6254~ 0.9999992378795~
##  2 "12th grade -\nno diploma~ 1795.89931~ -12979.58~ 16571.380~ 1
##  3 "Associate's degree-<1 ye~ 2074.67152~ -6559.141~ 10708.484~ 0.9999999077892~
##  4 "Bachelor's degree-<1 yea~ 14580.1551~ 7315.2372~ 21845.073~ 0.0000000000885~
##  5 "Doctorate degree-<1 year~ 25189.7386~ 9585.2238~ 40794.253~ 0.0000012162780~
##  6 "GED-<1 year of college"   -901.81983~ -11203.39~ 9399.7539~ 1
##  7 "Grade 1-<1 year of colle~ 542.595744~ -99031.21~ 100116.40~ 1
##  8 "Grade 10-<1 year of coll~ -16822.239~ -32915.71~ -728.7627~ 0.0283260103649~
##  9 "Grade 11-<1 year of coll~ -14224.537~ -27402.22~ -1046.845~ 0.0179645130007~
## 10 "Grade 2-<1 year of colle~ -17957.404~ -117531.2~ 81616.409~ 0.9999999997555~
## # ... with 266 more rows
```

I can represent this with a violin plot.

```
ggplot(transpo_data, aes(x = edu, y = PINCP)) +
  geom_violin() +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_x_discrete(name = "Educational Attainment") +
  scale_y_continuous(name = "Income",
                     breaks = seq(0, 500000, by = 50000),
```

```
                    labels = paste("$",
                                   prettyNum(seq(0, 500000, by = 50000),
                                             big.mark = ",")))
```



Interestingly the No Schooling category reaches quite high, and even higher than most of the grades. It seems that outliers are pulling the mean incomes of several variables up, including those who went to college for more than a year but earned no degree, those who finished 12th grade but without a high school diploma, those who have bachelor's degrees, high school diplomas, master's degrees, and professional degrees.

Educational Attainment and Travel Time to Work

```
ed_time_anova <- aov(JWMNP ~ edu, data = transpo_data)

summary(ed_time_anova)
```

```
##               Df  Sum Sq Mean Sq F value Pr(>F)
## edu           23   15792   686.6   1.436 0.0811 .
## Residuals   6890 3294340   478.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value is greater than 0.05, we can be at least 95% confident that there is a significant association between one's educational attainment and their travel time to work.

Educational Attainment and Rent

```
ed_rent_anova <- aov(GRNTP ~ edu, data = transpo_data)

summary(ed_rent_anova)
```

```
##               Df     Sum Sq Mean Sq F value              Pr(>F)
## edu           23  109518144 4761658   14.74 <0.0000000000000002 ***
## Residuals   6890 2226271110  323116
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

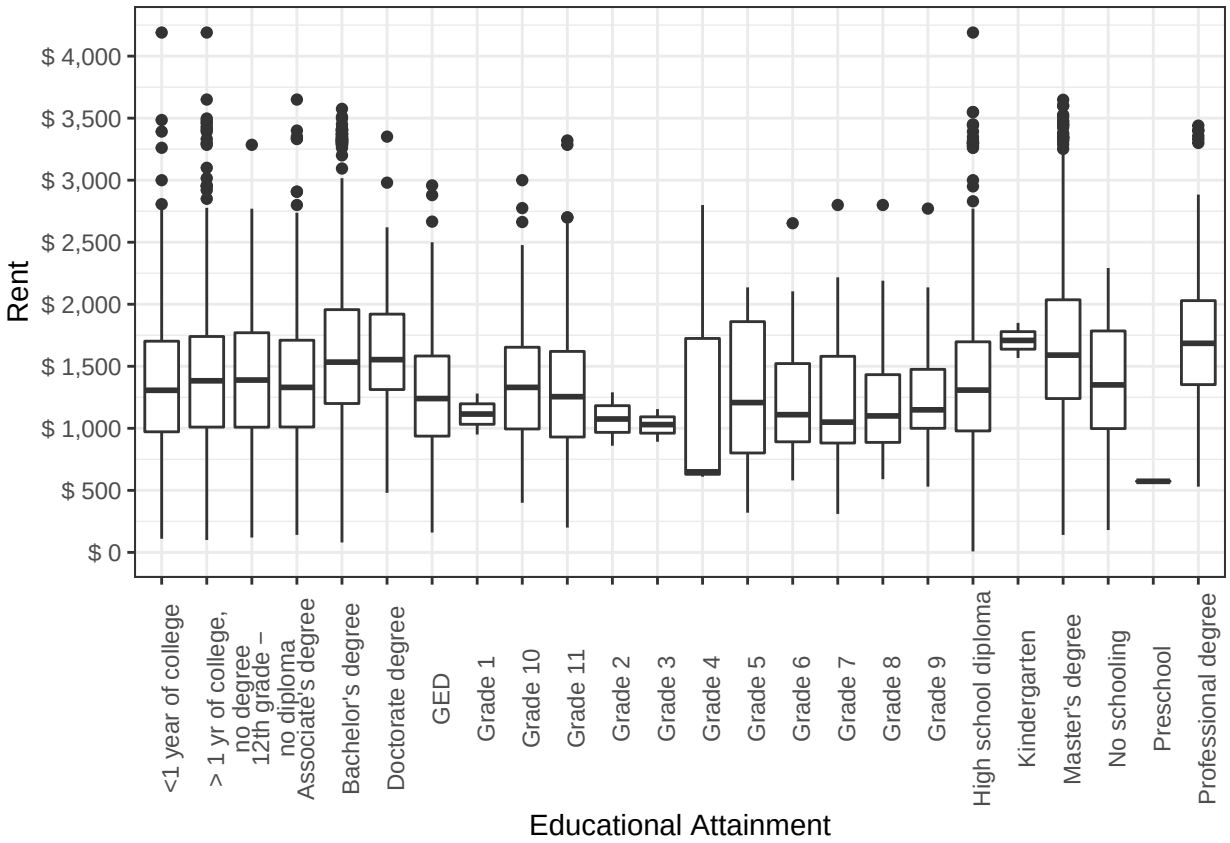
Since the p-value is less than 0.05, we can be 95% confident that there is a significant association between one's educational attainment and their monthly rent. Tukey's Honestly Significant Difference test shows us the magnitude of the differences among the different categories.

```
ed_rent_differences <- TukeyHSD(ed_rent_anova)

as_tibble(cbind(pair = row.names(ed_rent_differences$edu),
                ed_rent_differences$edu))
```

```
## # A tibble: 276 x 5
##    pair                        diff        lwr        upr        `p adj`
##    <chr>                       <chr>       <chr>      <chr>      <chr>
##  1 "> 1 yr of college,\nno d~ 61.4619701~ -51.53636~ 174.46030~ 0.9523798211348~
##  2 "12th grade -\nno diploma~ 73.8333206~ -143.6548~ 291.32147~ 0.99994012745025
##  3 "Associate's degree-<1 ye~ 39.3664902~ -87.71919~ 166.45217~ 0.9999881693356~
##  4 "Bachelor's degree-<1 yea~ 230.748465~ 123.81227~ 337.68465~ 0.0000000000034~
##  5 "Doctorate degree-<1 year~ 259.066739~ 29.375593~ 488.75788~ 0.0090706720066~
##  6 "GED-<1 year of college"   -68.005153~ -219.6394~ 83.629186~ 0.9954045416988~
##  7 "Grade 1-<1 year of colle~ -243.51489~ -1709.194~ 1222.1649~ 0.99999999995768
##  8 "Grade 10-<1 year of coll~ 17.2213701~ -219.6670~ 254.10980~ 1
##  9 "Grade 11-<1 year of coll~ -20.261560~ -214.2310~ 173.70789~ 1
## 10 "Grade 2-<1 year of colle~ -283.51489~ -1749.194~ 1182.1649~ 0.9999999989244~
## # ... with 266 more rows
```

I can represent these differences with a box plot.

```
ggplot(transpo_data, aes(x = edu, y = GRNTP)) +
  geom_boxplot() +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_x_discrete(name = "Educational Attainment") +
  scale_y_continuous(name = "Rent",
                     breaks = seq(0, 5000, by = 500),
                     labels = paste("$",
                                    prettyNum(seq(0, 5000, by = 500),
                                              big.mark = ",")))
```



The lowest median rent belongs to the preschool category, but it turns out there is only one person in my dataset who only ever finished preschool. The highest median rents tend to be found in the higher education degrees: Bachelor's, Master's, Doctorates, and other professional degrees. So - if anything, my MUP degree means I'll end up paying higher rent than others! I am just kidding, of course.

Educational Attainment and Age

```
ed_age_anova <- aov(AGEP ~ edu, data = transpo_data)

summary(ed_age_anova)
```

```
##               Df  Sum Sq Mean Sq F value             Pr(>F)
## edu           23   36378  1581.7   9.999 <0.0000000000000002 ***
## Residuals   6890 1089927   158.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value is less than 0.05, we can be 95% confident that there is a significant association between one's educational attainment and their age. Tukey's Honestly Significant Difference test shows us the magnitude of the differences among the different categories.

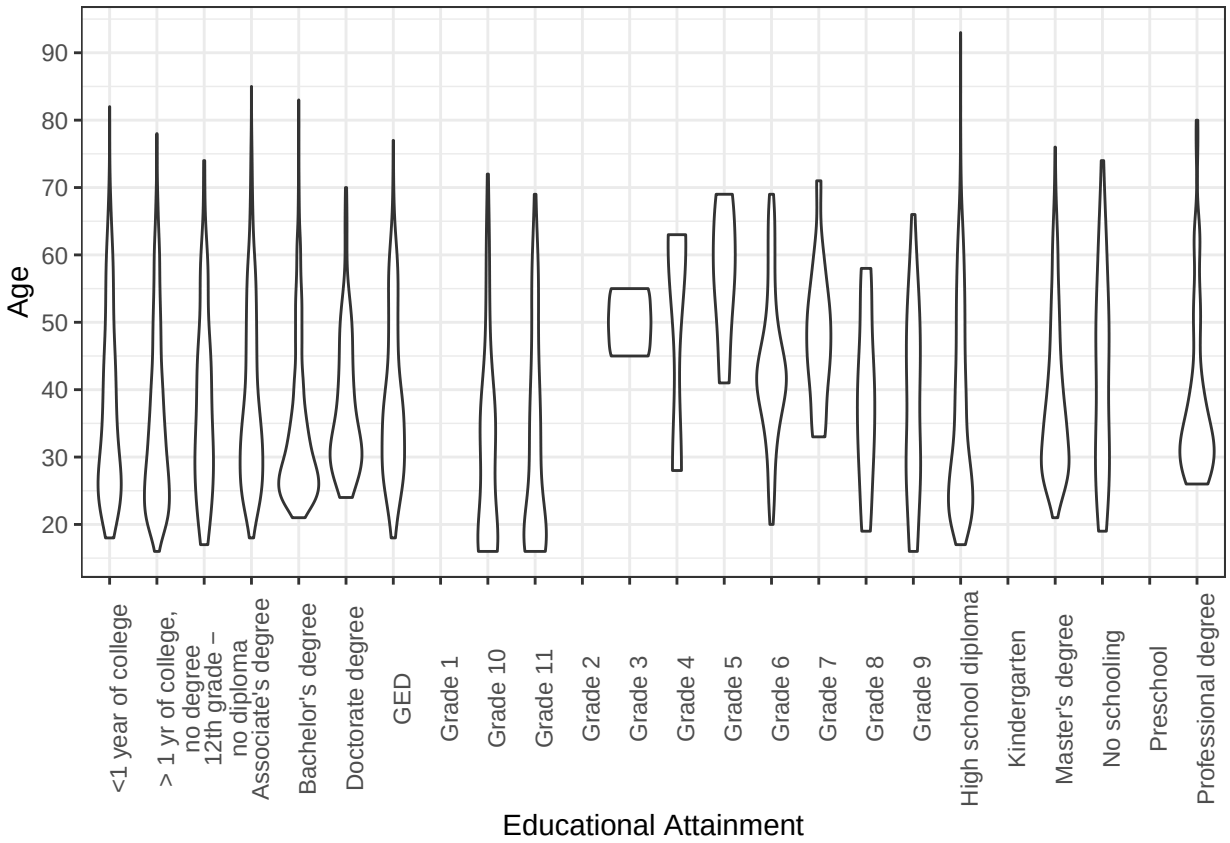
```
ed_age_differences <- TukeyHSD(ed_age_anova)
```

```
as_tibble(cbind(pair = row.names(ed_age_differences$edu),
                ed_age_differences$edu))
```

```
## # A tibble: 276 x 5
##    pair                          diff        lwr         upr         `p adj`
##    <chr>                         <chr>       <chr>       <chr>       <chr>
##  1 "> 1 yr of college,\nno deg~ -1.7614049~ -4.2616454~ 0.73883555~ 0.626299710~
##  2 "12th grade -\nno diploma-<~ 0.66816109~ -4.1440575~ 5.48037978~ 0.999999999~
##  3 "Associate's degree-<1 year~ 1.67754565~ -1.1343964~ 4.48948780~ 0.884676287~
##  4 "Bachelor's degree-<1 year ~ -2.7916262~ -5.1577336~ -0.4255188~ 0.004162250~
##  5 "Doctorate degree-<1 year o~ 0.58397742~ -4.4982489~ 5.66620378~ 1
##  6 "GED-<1 year of college"     1.66978447~ -1.6853297~ 5.02489873~ 0.982449399~
##  7 "Grade 1-<1 year of college" -6.2425531~ -38.672697~ 26.1875907~ 0.999999999~
##  8 "Grade 10-<1 year of colleg~ -6.2700257~ -11.511501~ -1.0285496~ 0.003224154~
##  9 "Grade 11-<1 year of colleg~ -5.2158865~ -9.5077222~ -0.9240508~ 0.002382054~
## 10 "Grade 2-<1 year of college" 15.7574468~ -16.672697~ 48.1875907~ 0.986907355~
## # ... with 266 more rows
```

I can represent these differences with a violin plot.

```
ggplot(transpo_data, aes(x = edu, y = AGEP)) +
  geom_violin() +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_x_discrete(name = "Educational Attainment") +
  scale_y_continuous(name = "Age",
                     breaks = seq(0, 90, by = 10),
                     labels = paste(prettyNum(seq(0, 90, by = 10))))
```



The shape of the Grade 3 distribution is interesting to me, though it appears there are only 3 people in my data set who only ever completed Grade 3. Despite the differences in age by educational attainment, I don't think this data necessarily highlights any interesting trends.

Relationship between two categorical variables

Sex and Means of Transportation to Work

By running this Chi-square test, I am attempting to answer: is there a relationship between one's sex and their means of transportation to work?

```
sex_means_chi_sq <- chisq.test(transpo_data$meansTW, transpo_data$SEX_label == "Male")

sex_means_chi_sq
```

```
##
##  Pearson's Chi-squared test
##
## data:  transpo_data$meansTW and transpo_data$SEX_label == "Male"
## X-squared = 31.544, df = 9, p-value = 0.0002387
```

The p-value is less than 0.05, so the relationship is significant at a 95% confidence interval. These are the values observed for each combination of variables.

```
sex_means_chi_sq$observed
```

```
##
## transpo_data$meansTW        FALSE TRUE
##   Bicycle                      36   93
##   Bus or\ntrolley bus          89  123
##   Car, truck,\nor van        2825 3233
##   Motorcycle                    3   13
##   Other                        26   35
##   Railroad                     11   21
##   Streetcar or\ntrolley car     2    8
##   Subway or\nelevated          11   23
##   Taxicab                       9   10
##   Walk                        161  182
```

And here are the values we would expect if the two variables were not related.

```
sex_means_chi_sq$expected
```

```
##
## transpo_data$meansTW              FALSE        TRUE
##   Bicycle                     59.201186   69.798814
##   Bus or\ntrolley bus         97.291872  114.708128
##   Car, truck,\nor van       2780.161122 3277.838878
##   Motorcycle                   7.342783    8.657217
##   Other                       27.994359   33.005641
##   Railroad                    14.685566   17.314434
##   Streetcar or\ntrolley car    4.589239    5.410761
##   Subway or\nelevated         15.603413   18.396587
##   Taxicab                      8.719555   10.280445
##   Walk                       157.410905  185.589095
```

Sex and Educational Attainment

By running this Chi-square test, I am attempting to answer: is there a relationship between one's sex and their educational attainment?

```
sex_edu_chi_sq <- chisq.test(transpo_data$edu, transpo_data$SEX_label == "Male")

sex_edu_chi_sq
```

```
##
##  Pearson's Chi-squared test
##
## data:  transpo_data$edu and transpo_data$SEX_label == "Male"
## X-squared = 82.147, df = 23, p-value = 0.00000001421
```

The p-value is less than 0.05, so the relationship is significant at a 95% confidence interval. These are the values observed for each combination of variables.

```
sex_edu_chi_sq$observed
```

```
##
## transpo_data$edu                FALSE TRUE
##   <1 year of college              223  247
##   > 1 yr of college,\nno degree   553  614
##   12th grade -\nno diploma         32   80
##   Associate's degree              307  300
##   Bachelor's degree               916  918
##   Doctorate degree                 43   55
##   GED                             110  198
##   Grade 1                           1    1
##   Grade 10                         42   49
##   Grade 11                         58   92
##   Grade 2                           0    2
##   Grade 3                           0    3
##   Grade 4                           1    2
##   Grade 5                           1    5
##   Grade 6                          21   31
##   Grade 7                           6    9
##   Grade 8                          10   14
##   Grade 9                          23   33
##   High school diploma             475  705
##   Kindergarten                      1    1
##   Master's degree                 289  284
##   No schooling                     16   34
##   Preschool                         0    1
##   Professional degree              45   63
```

And here are the values we would expect if the two variables were not related.

```
sex_edu_chi_sq$expected
```

```
##
## transpo_data$edu                      FALSE        TRUE
##   <1 year of college            215.6942436 254.3057564
##   > 1 yr of college,\nno degree 535.5642175 631.4357825
##   12th grade -\nno diploma       51.3994793  60.6005207
##   Associate's degree            278.5668209 328.4331791
##   Bachelor's degree             841.6664738 992.3335262
##   Doctorate degree               44.9745444  53.0254556
##   GED                           141.3485681 166.6514319
##   Grade 1                         0.9178478   1.0821522
##   Grade 10                       41.7620769  49.2379231
##   Grade 11                       68.8385884  81.1614116
##   Grade 2                         0.9178478   1.0821522
##   Grade 3                         1.3767718   1.6232282
##   Grade 4                         1.3767718   1.6232282
##   Grade 5                         2.7535435   3.2464565
##   Grade 6                        23.8640440  28.1359560
##   Grade 7                         6.8838588   8.1161412
##   Grade 8                        11.0141741  12.9858259
```

```
##   Grade 9                        25.6997397  30.3002603
##   High school diploma           541.5302285 638.4697715
##   Kindergarten                    0.9178478   1.0821522
##   Master's degree               262.9634076 310.0365924
##   No schooling                   22.9461961  27.0538039
##   Preschool                       0.4589239   0.5410761
##   Professional degree            49.5637836  58.4362164
```

Educational Attainment and Means of Transportation to Work

By running this Chi-square test, I am attempting to answer: is there a relationship between one's educational attainment and whether they take a car, truck, or van to work?

```
edu_means_chi_sq <- chisq.test(transpo_data$edu, transpo_data$meansTW)

edu_means_chi_sq
```

```
##
##  Pearson's Chi-squared test
##
## data:  transpo_data$edu and transpo_data$meansTW
## X-squared = 205.6, df = 207, p-value = 0.5145
```

The p-value is far greater than 0.05, so the relationship is not statistically significant at a 95% confidence interval. These are the values observed for each combination of variables.

```
edu_means_chi_sq$observed
```

```
##                                transpo_data$meansTW
## transpo_data$edu                Bicycle Bus or\ntrolley bus Car, truck,\nor van
##   <1 year of college                  1                  16                 426
##   > 1 yr of college,\nno degree      15                  31                1034
##   12th grade -\nno diploma            2                   3                 100
##   Associate's degree                  3                   8                 558
##   Bachelor's degree                  62                  66                1558
##   Doctorate degree                    7                   4                  78
##   GED                                 3                  15                 270
##   Grade 1                             0                   0                   2
##   Grade 10                            0                   2                  81
##   Grade 11                            2                   3                 136
##   Grade 2                             0                   0                   2
##   Grade 3                             0                   0                   3
##   Grade 4                             0                   0                   3
##   Grade 5                             0                   0                   5
##   Grade 6                             0                   2                  47
##   Grade 7                             1                   1                  13
##   Grade 8                             0                   0                  23
##   Grade 9                             0                   1                  53
##   High school diploma                12                  38                1039
##   Kindergarten                        0                   0                   2
##   Master's degree                    20                  17                 482
##   No schooling                        0                   4                  42
```

```
##   Preschool                            0                  0                   1
##   Professional degree                  1                  1                 100
##                                transpo_data$meansTW
## transpo_data$edu                Motorcycle Other Railroad
##   <1 year of college                     0     5        0
##   > 1 yr of college,\nno degree          7     6        8
##   12th grade -\nno diploma               0     0        0
##   Associate's degree                     2     6        4
##   Bachelor's degree                      3    14        9
##   Doctorate degree                       0     0        2
##   GED                                    0     3        0
##   Grade 1                                0     0        0
##   Grade 10                               0     0        0
##   Grade 11                               0     1        0
##   Grade 2                                0     0        0
##   Grade 3                                0     0        0
##   Grade 4                                0     0        0
##   Grade 5                                0     0        0
##   Grade 6                                0     1        0
##   Grade 7                                0     0        0
##   Grade 8                                0     1        0
##   Grade 9                                0     1        0
##   High school diploma                    3    16        1
##   Kindergarten                           0     0        0
##   Master's degree                        1     4        7
##   No schooling                           0     1        1
##   Preschool                              0     0        0
##   Professional degree                    0     2        0
##                                transpo_data$meansTW
## transpo_data$edu                Streetcar or\ntrolley car Subway or\nelevated
##   <1 year of college                                    1                   2
##   > 1 yr of college,\nno degree                         0                   3
##   12th grade -\nno diploma                              0                   1
##   Associate's degree                                    1                   2
##   Bachelor's degree                                     5                  14
##   Doctorate degree                                      0                   1
##   GED                                                   1                   1
##   Grade 1                                               0                   0
##   Grade 10                                              0                   0
##   Grade 11                                              0                   1
##   Grade 2                                               0                   0
##   Grade 3                                               0                   0
##   Grade 4                                               0                   0
##   Grade 5                                               0                   0
##   Grade 6                                               0                   0
##   Grade 7                                               0                   0
##   Grade 8                                               0                   0
##   Grade 9                                               0                   0
##   High school diploma                                   1                   4
##   Kindergarten                                          0                   0
##   Master's degree                                       0                   5
##   No schooling                                          0                   0
##   Preschool                                             0                   0
##   Professional degree                                   1                   0
```

```
##                                transpo_data$meansTW
## transpo_data$edu                Taxicab Walk
##   <1 year of college                  3   16
##   > 1 yr of college,\nno degree       3   60
##   12th grade -\nno diploma            0    6
##   Associate's degree                  2   21
##   Bachelor's degree                   1  102
##   Doctorate degree                    0    6
##   GED                                 0   15
##   Grade 1                             0    0
##   Grade 10                            0    8
##   Grade 11                            1    6
##   Grade 2                             0    0
##   Grade 3                             0    0
##   Grade 4                             0    0
##   Grade 5                             0    1
##   Grade 6                             0    2
##   Grade 7                             0    0
##   Grade 8                             0    0
##   Grade 9                             0    1
##   High school diploma                 6   60
##   Kindergarten                        0    0
##   Master's degree                     3   34
##   No schooling                        0    2
##   Preschool                           0    0
##   Professional degree                 0    3
```

And these are the expected values.

```
edu_means_chi_sq$expected
```

```
##                                transpo_data$meansTW
## transpo_data$edu                    Bicycle Bus or\ntrolley bus
##   <1 year of college             8.76916402         14.41133931
##   > 1 yr of college,\nno degree 21.77364767         35.78304889
##   12th grade -\nno diploma       2.08967313          3.43419150
##   Associate's degree            11.32528204         18.61209141
##   Bachelor's degree             34.21839745         56.23488574
##   Doctorate degree               1.82846399          3.00491756
##   GED                            5.74660110          9.44402661
##   Grade 1                        0.03731559          0.06132485
##   Grade 10                       1.69785942          2.79028059
##   Grade 11                       2.79866937          4.59936361
##   Grade 2                        0.03731559          0.06132485
##   Grade 3                        0.05597339          0.09198727
##   Grade 4                        0.05597339          0.09198727
##   Grade 5                        0.11194677          0.18397454
##   Grade 6                        0.97020538          1.59444605
##   Grade 7                        0.27986694          0.45993636
##   Grade 8                        0.44778710          0.73589818
##   Grade 9                        1.04483656          1.71709575
##   High school diploma           22.01619902         36.18166040
##   Kindergarten                   0.03731559          0.06132485
```

```
##   Master's degree               10.69091698         17.56956899
##   No schooling                   0.93288979          1.53312120
##   Preschool                      0.01865780          0.03066242
##   Professional degree            2.01504194          3.31154180
##                              transpo_data$meansTW
## transpo_data$edu               Car, truck,\nor van  Motorcycle        Other
##   <1 year of college                  411.8108186 1.087648250  4.146658953
##   > 1 yr of college,\nno degree      1022.5175007 2.700607463 10.296065953
##   12th grade -\nno diploma             98.1336419 0.259184264  0.988140006
##   Associate's degree                  531.8492913 1.404686144  5.355365924
##   Bachelor's degree                  1606.9383859 4.244142320 16.180792595
##   Doctorate degree                     85.8669367 0.226786231  0.864622505
##   GED                                 269.8675152 0.712756725  2.717385016
##   Grade 1                               1.7523865 0.004628290  0.017645357
##   Grade 10                             79.7335840 0.210587214  0.802863755
##   Grade 11                            131.4289847 0.347121782  1.323401793
##   Grade 2                               1.7523865 0.004628290  0.017645357
##   Grade 3                               2.6285797 0.006942436  0.026468036
##   Grade 4                               2.6285797 0.006942436  0.026468036
##   Grade 5                               5.2571594 0.013884871  0.052936072
##   Grade 6                              45.5620480 0.120335551  0.458779288
##   Grade 7                              13.1428985 0.034712178  0.132340179
##   Grade 8                              21.0286375 0.055539485  0.211744287
##   Grade 9                              49.0668209 0.129592132  0.494070003
##   High school diploma                1033.9080127 2.730691351 10.410760775
##   Kindergarten                          1.7523865 0.004628290  0.017645357
##   Master's degree                     502.0587214 1.326005207  5.055394851
##   No schooling                         43.8096616 0.115707261  0.441133931
##   Preschool                             0.8761932 0.002314145  0.008822679
##   Professional degree                  94.6288690 0.249927683  0.952849291
##                              transpo_data$meansTW
## transpo_data$edu                  Railroad Streetcar or\ntrolley car
##   <1 year of college             2.175296500               0.679780156
##   > 1 yr of college,\nno degree  5.401214926               1.687879664
##   12th grade -\nno diploma       0.518368528               0.161990165
##   Associate's degree             2.809372288               0.877928840
##   Bachelor's degree              8.488284640               2.652588950
##   Doctorate degree               0.453572462               0.141741394
##   GED                            1.425513451               0.445472953
##   Grade 1                        0.009256581               0.002892682
##   Grade 10                       0.421174429               0.131617009
##   Grade 11                       0.694243564               0.216951114
##   Grade 2                        0.009256581               0.002892682
##   Grade 3                        0.013884871               0.004339022
##   Grade 4                        0.013884871               0.004339022
##   Grade 5                        0.027769743               0.008678045
##   Grade 6                        0.240671102               0.075209719
##   Grade 7                        0.069424356               0.021695111
##   Grade 8                        0.111078970               0.034712178
##   Grade 9                        0.259184264               0.080995082
##   High school diploma            5.461382702               1.706682094
##   Kindergarten                   0.009256581               0.002892682
##   Master's degree                2.652010414               0.828753254
##   No schooling                   0.231414521               0.072317038
```

```
##   Preschool                      0.004628290              0.001446341
##   Professional degree            0.499855366              0.156204802
##                                transpo_data$meansTW
## transpo_data$edu                Subway or\nelevated     Taxicab        Walk
##   <1 year of college                    2.311252531 1.291582297 23.31645936
##   > 1 yr of college,\nno degree         5.738790859 3.206971362 57.89427249
##   12th grade -\nno diploma              0.550766561 0.307781313  5.55626266
##   Associate's degree                    2.984958056 1.668064796 30.11295921
##   Bachelor's degree                     9.018802430 5.039919005 90.98380098
##   Doctorate degree                      0.481920741 0.269308649  4.86172982
##   GED                                   1.514608042 0.846398612 15.27972230
##   Grade 1                               0.009835117 0.005496095  0.09921898
##   Grade 10                              0.447497830 0.250072317  4.51446341
##   Grade 11                              0.737633787 0.412207116  7.44142320
##   Grade 2                               0.009835117 0.005496095  0.09921898
##   Grade 3                               0.014752676 0.008244142  0.14882846
##   Grade 4                               0.014752676 0.008244142  0.14882846
##   Grade 5                               0.029505351 0.016488285  0.29765693
##   Grade 6                               0.255713046 0.142898467  2.57969338
##   Grade 7                               0.073763379 0.041220712  0.74414232
##   Grade 8                               0.118021406 0.065953139  1.19062771
##   Grade 9                               0.275383280 0.153890657  2.77813133
##   High school diploma                   5.802719121 3.242695979 58.53919583
##   Kindergarten                          0.009835117 0.005496095  0.09921898
##   Master's degree                       2.817761065 1.574631183 28.42623662
##   No schooling                          0.245877929 0.137402372  2.48047440
##   Preschool                             0.004917559 0.002748047  0.04960949
##   Professional degree                   0.531096326 0.296789124  5.35782470
```