Quant_3rd_final

Jiwon Park

2020 9 23

Contents

Load Packages

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

Load Data

```
tractMA2018 <- read.csv("tractMA2018_3rd.csv")
```

The sample included census tracts of all counties in Massachusetts. Therefore, my level of analysis here is "tract".

I edited one categorical variable from the second assignment, which is the "Transportation Mode With the Highest Modal Share (categorical)". I changed it from the variable "Whether more than 10% of population commute by Non-Car" to make this variable have multiple sub-categories. I also decided to not use the variable "Total population (continuous)" as this is not directly related to the question I have.

Therefore, the variables included in this assignment are :

1. Median Income (continuous) : med_income
2. Monthly Housing Cost (continuous) : mon_hous
3. Percentage of Tenant Population (continuous) : pct_ten
4. Percentage of Foreign-born Population (continuous) : pct_foreign
5. Whether the Majority of Population is Foreign-born or Not : maj_foreign
6. Transportation Mode With the Highest Modal Share (categorical) : maj_tra
7. Race/Ethnicity With the Highest Ratio (not Hispanic or Latino) (categorical) : maj_race

Relationships between Two Continuous Variables (Pearson's Correlation)

1. Relationships between Median Income and Percentage of Tenant Population

```r
correlation1 <- cor.test(tractMA2018$med_incomeE, tractMA2018$pct_ten)
correlation1
```

```
##
##  Pearson's product-moment correlation
##
## data:  tractMA2018$med_incomeE and tractMA2018$pct_ten
## t = -18.551, df = 1460, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4773266 -0.3943089
## sample estimates:
##        cor
## -0.4367472
```

The relationship between median income and percentage of tenant population in the tracts of Massachusetts : weak ($< |.5|$), but statistically significant (p-value<0.05), negative correlation at a 95% confidence level.

2. Relationships between Median Income and Monthly Housing Cost

```r
correlation2 <- cor.test(tractMA2018$med_incomeE, tractMA2018$mon_housE)
correlation2
```

```
##
##  Pearson's product-moment correlation
##
## data:  tractMA2018$med_incomeE and tractMA2018$mon_housE
## t = 8.9656, df = 1460, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1792690 0.2764661
## sample estimates:
##       cor
## 0.2284367
```

The relationship between median income and monthly housing cost in the tracts of Massachusetts : weak ($< |.3|$), but statistically significant(p-value<0.05), positive correlation at a 95% confidence level.

3. Relationships between Median Income and Nativity

```r
correlation3 <- cor.test(tractMA2018$med_income, tractMA2018$pct_foreign)
correlation3
```

```
##
##  Pearson's product-moment correlation
##
## data:  tractMA2018$med_income and tractMA2018$pct_foreign
## t = -8.1359, df = 1460, p-value = 8.66e-16
```

```
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2567830 -0.1586845
## sample estimates:
##        cor
## -0.2082574
```

The relationship between median income and percentage of foreign-born population in the tracts of Massachusetts : very weak (< |.3|), but statistically significant(p-value<0.05), negative correlation at a 95% confidence level.

4. Relationships between Percentage of Tenant Population and Monthly Housing Cost

```
correlation4 <- cor.test(tractMA2018$pct_ten, tractMA2018$mon_housE)
correlation4
```

```
##
##  Pearson's product-moment correlation
##
## data:  tractMA2018$pct_ten and tractMA2018$mon_housE
## t = -4.8067, df = 1462, p-value = 1.692e-06
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.17484488 -0.07397056
## sample estimates:
##      cor
## -0.12473
```

The relationship between percentage of tenant population and monthly housing cost in the tracts of Massachusetts : extremely weak (< |.2|), but statistically significant(p-value<0.05), negative correlation at a 95% confidence level.

5. Relationships between Nativity and Monthly Housing Cost

```
correlation5 <- cor.test(tractMA2018$pct_foreign, tractMA2018$mon_housE)
correlation5
```

```
##
##  Pearson's product-moment correlation
##
## data:  tractMA2018$pct_foreign and tractMA2018$mon_housE
## t = -3.0005, df = 1462, p-value = 0.002741
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.12894661 -0.02710777
## sample estimates:
##         cor
## -0.07823127
```

The relationship between percentage of foreign-born population and monthly housing cost in the tracts of Massachusetts : extremely weak (< |.1|), but statistically significant(p-value<0.05), negative correlation at a 95% confidence level.

6. Relationships between Percentage of Tenant Population and Nativity

```
correlation6 <- cor.test(tractMA2018$pct_ten, tractMA2018$pct_foreign)
correlation6
```

```
##
##  Pearson's product-moment correlation
##
## data:  tractMA2018$pct_ten and tractMA2018$pct_foreign
## t = 29.569, df = 1462, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5786537 0.6428350
## sample estimates:
##       cor
## 0.6117501
```

The relationship between percentage of tenant population and percentage of foreign-born population in the tracts of Massachusetts : moderate ($> |.6|$), statistically significant(p-value<0.05), positive correlation at a 95% confidence level.

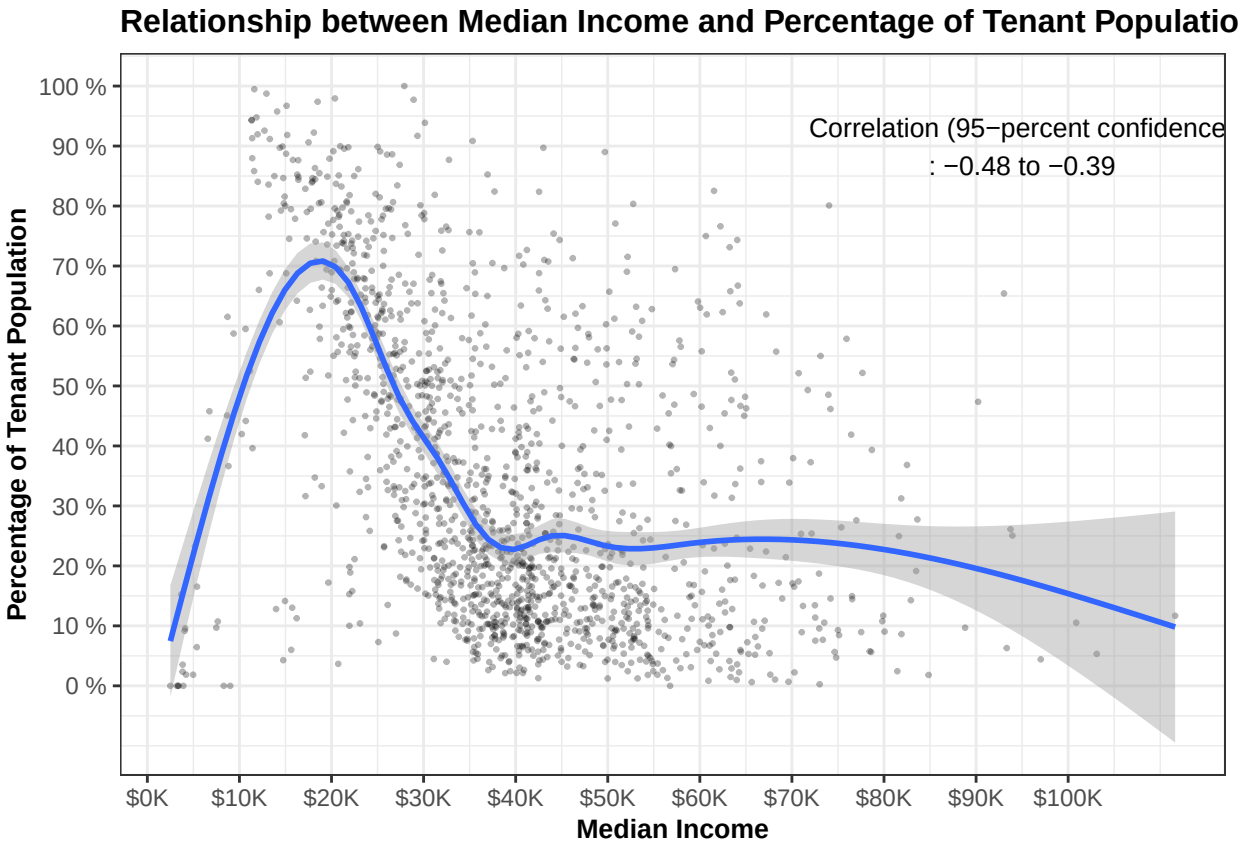
1-1. A scatter plot showing the relationships between median income and percentage of tenant population

```
ggplot(tractMA2018, aes(x=med_incomeE, y=pct_ten)) +
  geom_point(size = 0.5, alpha = 0.3, position = "jitter") +
  geom_smooth() +
  theme_bw() +
  scale_x_continuous(name = "Median Income",
                     breaks = seq(0, 100000, by = 10000),
                     labels = paste("$", prettyNum(seq(0, 100, by = 10), big.mark = ","), "K", sep = "")
  scale_y_continuous(name = "Percentage of Tenant Population",
                     breaks = seq(0, 1, by = 0.1),
                     labels = paste(prettyNum(seq(0, 100, by = 10)), "%")) +
  ggtitle("Relationship between Median Income and Percentage of Tenant Population") +
  theme(title = element_text(size=10, face = "bold")) +
  annotate("text", x = 95000, y = 0.9, size=3.5,
           label = paste("Correlation (95-percent confidence)\n:", prettyNum(correlation1$conf.int[1], d
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 16 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 16 rows containing missing values (geom_point).
```



Looking at this graph, it is understandable why the general correlation between two variables was weak. It is because the percentage of tenant population increases with high positive coefficient to the level of certain median income (around USD15,000), and then significantly drops thereafter until the median income reaches USD45,000.

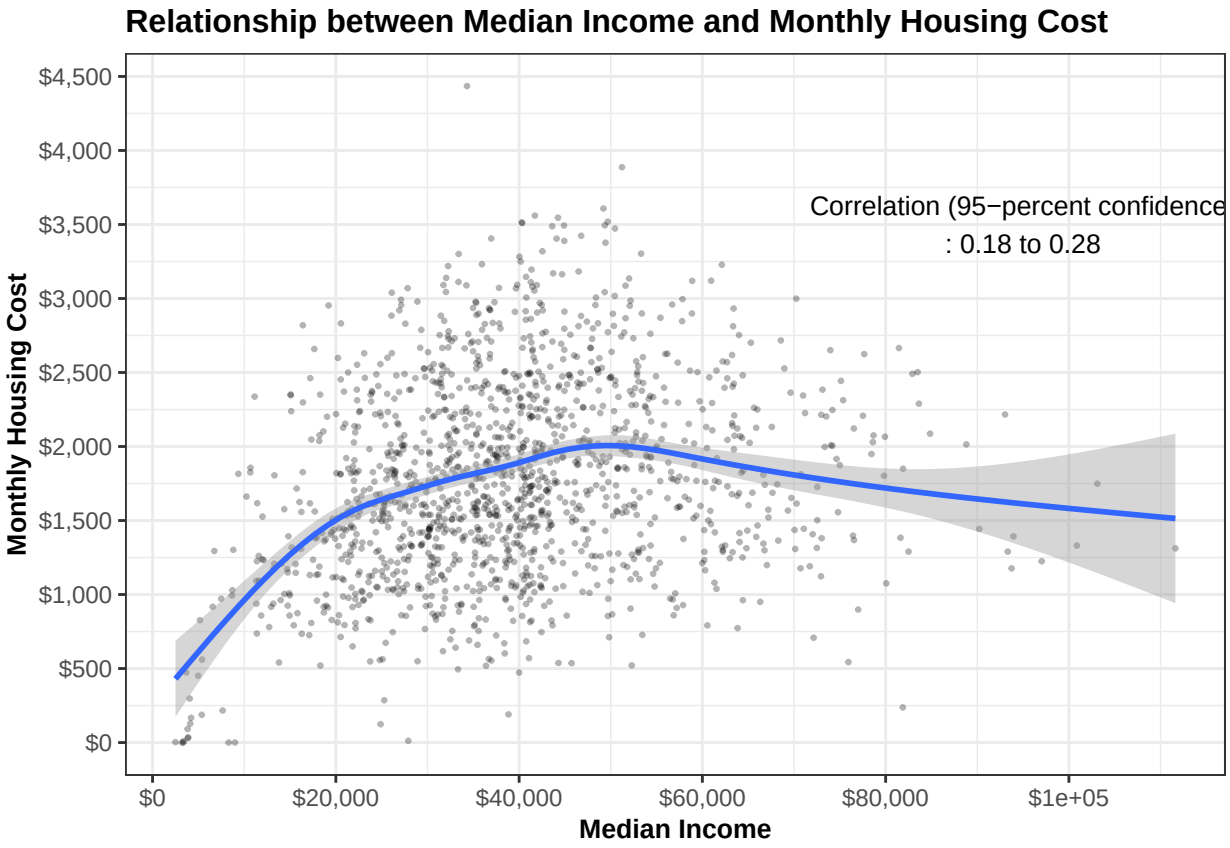
2-1. A scatter plot showing the relationships between median income and monthly housing cost

```
ggplot(tractMA2018, aes(x=med_incomeE, y=mon_housE)) +
  geom_point(size = 0.5, alpha = 0.3, position = "jitter") +
  geom_smooth() +
  theme_bw() +
  scale_x_continuous(name = "Median Income",
                     breaks = seq(0, 100000, by = 20000),
                     labels = paste("$", prettyNum(seq(0, 100000, by = 20000), big.mark = ","), sep = "
  scale_y_continuous(name = "Monthly Housing Cost",
                     breaks = seq(0, 5000, by = 500),
                     labels = paste( "$", prettyNum(seq(0, 5000, by = 500), big.mark = ","), sep = ""))
  ggtitle("Relationship between Median Income and Monthly Housing Cost") +
  theme(title = element_text(size=10, face = "bold")) +
  annotate("text", x = 95000, y = 3500, size=3.5,
           label = paste("Correlation (95-percent confidence)\n:", prettyNum(correlation2$conf.int[1], 
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 16 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 16 rows containing missing values (geom_point).
```



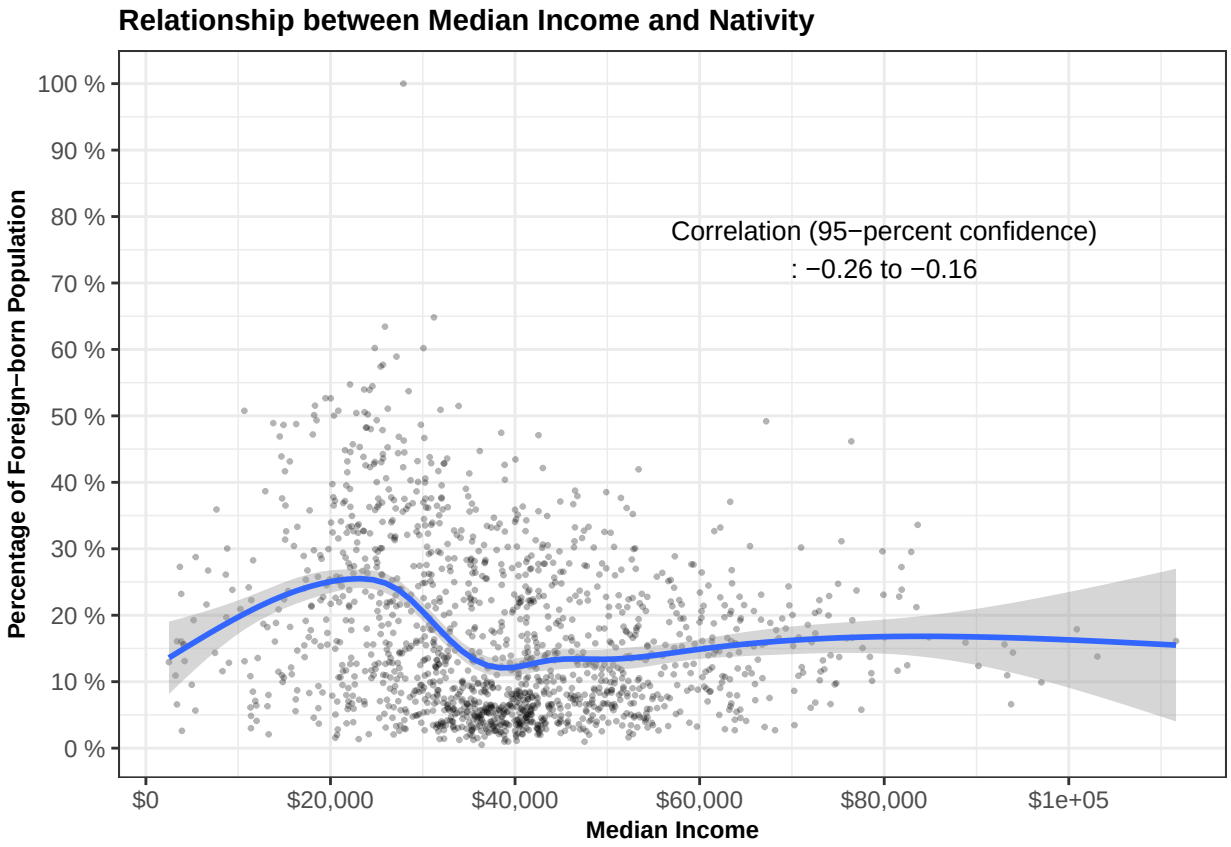
3-1. A scatter plot showing the relationships between median income and nativity

```
ggplot(tractMA2018, aes(x=med_incomeE, y=pct_foreign)) +
  geom_point(size = 0.5, alpha = 0.3, position = "jitter") +
  geom_smooth() +
  theme_bw() +
  scale_x_continuous(name = "Median Income",
                     breaks = seq(0, 100000, by = 20000),
                     labels = paste("$", prettyNum(seq(0, 100000, by = 20000), big.mark = ","), sep = "
  scale_y_continuous(name = "Percentage of Foreign-born Population",
                     breaks = seq(0, 1, by = 0.1),
                     labels = paste(prettyNum(seq(0, 100, by = 10), big.mark = ","), "%")) +
  ggtitle("Relationship between Median Income and Nativity") +
  theme(title = element_text(size=9, face = "bold")) +
  annotate("text", x = 80000, y = 0.75, size=3.5,
           label = paste("Correlation (95-percent confidence)\n:", prettyNum(correlation3$conf.int[1], 
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 16 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 16 rows containing missing values (geom_point).
```



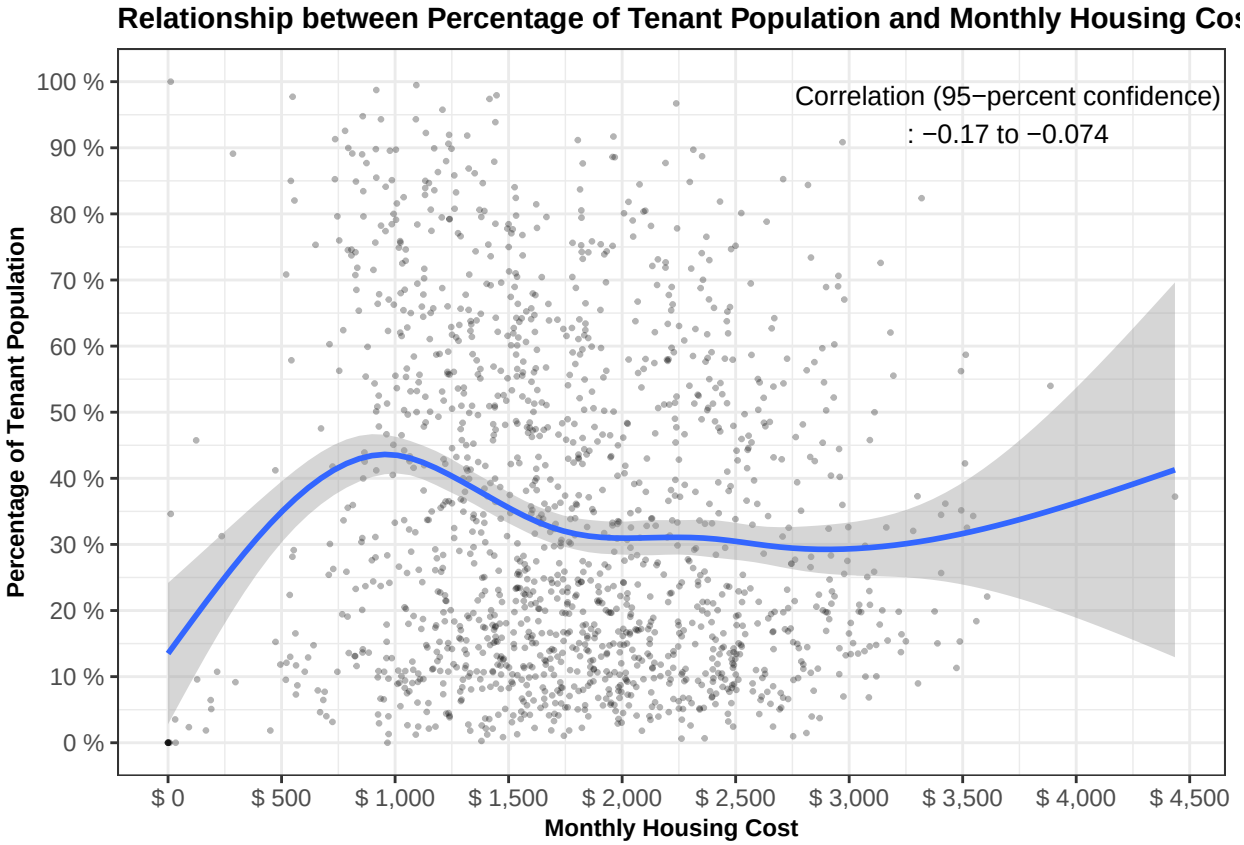
4-1. A scatter plot showing the relationships between percentage of tenant population and monthly housing cost

```
ggplot(tractMA2018, aes(x=mon_housE, y=pct_ten)) +
  geom_point(size = 0.5, alpha = 0.3, position = "jitter") +
  geom_smooth() +
  theme_bw() +
  scale_x_continuous(name = "Monthly Housing Cost",
                     breaks = seq(0, 5000, by = 500),
                     labels = paste("$", prettyNum(seq(0, 5000, by = 500), big.mark = ","))) +
  scale_y_continuous(name = "Percentage of Tenant Population",
                     breaks = seq(0, 1, by = 0.1),
                     labels = paste(prettyNum(seq(0, 100, by = 10)), "%")) +
  ggtitle("Relationship between Percentage of Tenant Population and Monthly Housing Cost") +
  theme(title = element_text(size=9, face = "bold")) +
  annotate("text", x = 3700, y = 0.95, size=3.5,
           label = paste("Correlation (95-percent confidence)\n:", prettyNum(correlation4$conf.int[1],
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 14 rows containing non-finite values (stat_smooth).

## Warning: Removed 14 rows containing missing values (geom_point).
```



Looking at this graph, it is understandable that why the general correlation between two variables was extremely weak. It is interesting (especially when I understood this statistics based on my urban experience in Korea), as often homeownership is deemed to lower the housing cost for households - though there are also a lot more benefits attached to it such as the freedom from eviction. However, this plot at least implies that there is is a low correlation between the housing cost and housing tenure in the tracts of Massachusetts. I checked that monthly housing cost used here contains both cash rent and mortgage payment with this source (https://www2.census.gov/programs-surveys/acs/tech_docs/subject_definitions/2018_ACSSubjectDefinitions.pdf).

5-1. A scatter plot showing the relationships between monthly housing cost and nativity

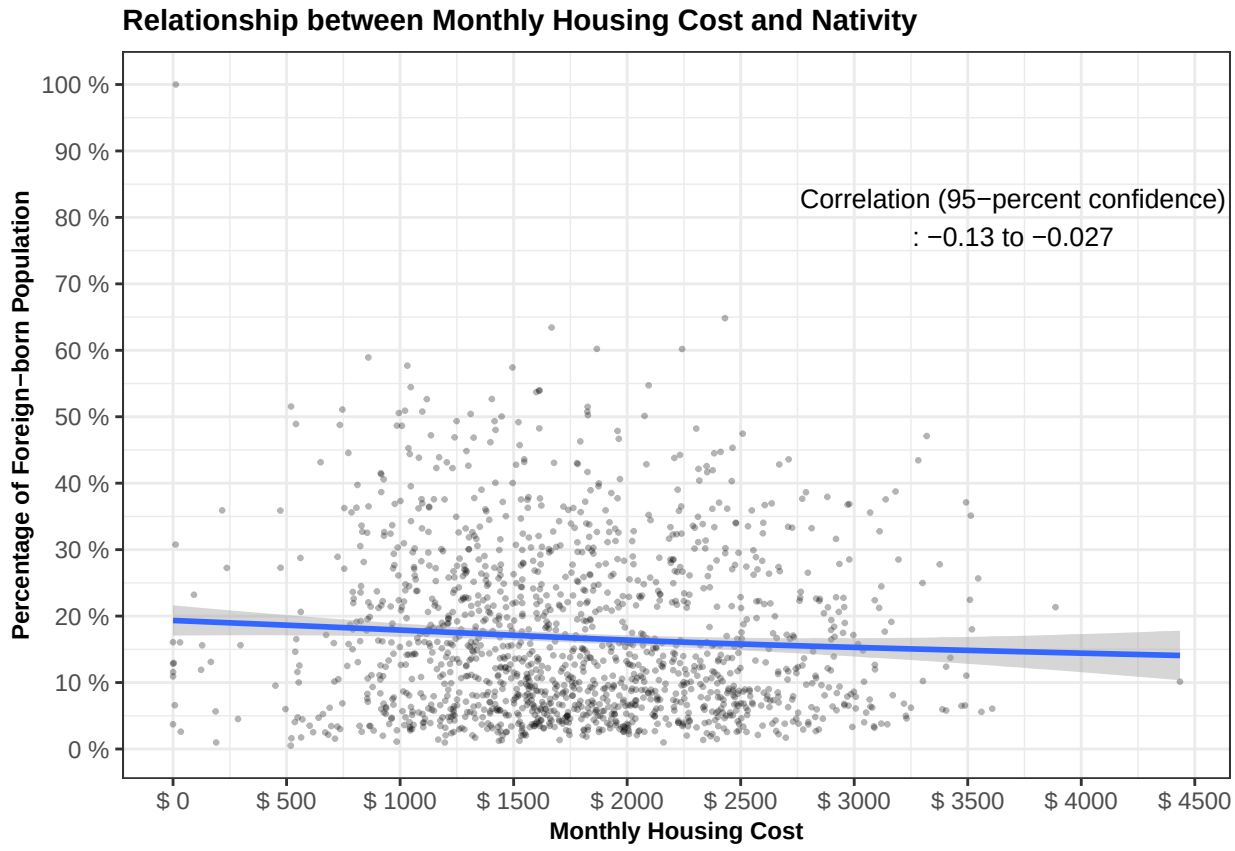
```
ggplot(tractMA2018, aes(x=mon_housE, y=pct_foreign)) +
  geom_point(size = 0.5, alpha = 0.3, position = "jitter") +
  geom_smooth() +
  theme_bw() +
  scale_x_continuous(name = "Monthly Housing Cost",
                     breaks = seq(0, 5000, by = 500),
                     labels = paste("$", prettyNum(seq(0, 5000, by = 500)))) +
  scale_y_continuous(name = "Percentage of Foreign-born Population",
                     breaks = seq(0, 1, by = 0.1),
```

```
                    labels = paste(prettyNum(seq(0, 100, by = 10)), "%")) +
  ggtitle("Relationship between Monthly Housing Cost and Nativity") +
  theme(title = element_text(size=9, face = "bold")) +
  annotate("text", x = 3700, y = 0.8, size=3.5,
           label = paste("Correlation (95-percent confidence)\n:", prettyNum(correlation5$conf.int[1], d
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 14 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 14 rows containing missing values (geom_point).
```



6-1. A scatter plot showing the relationships between percentage of tenant population and nativity

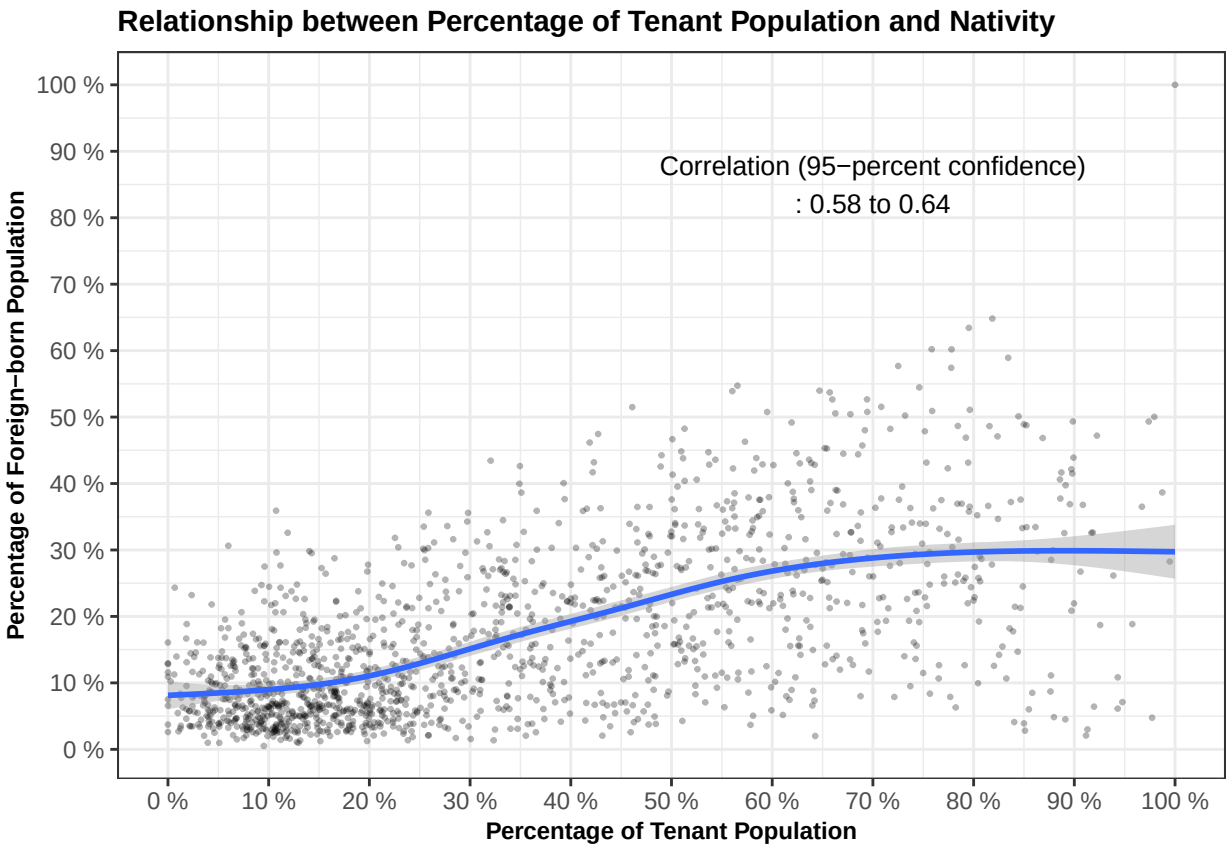
```
ggplot(tractMA2018, aes(x=pct_ten, y=pct_foreign)) +
  geom_point(size = 0.5, alpha = 0.3, position = "jitter") +
  geom_smooth() +
  theme_bw() +
  scale_x_continuous(name = "Percentage of Tenant Population",
                     breaks = seq(0, 1, by = 0.1),
                     labels = paste(prettyNum(seq(0, 100, by = 10)), "%")) +
```

```
  scale_y_continuous(name = "Percentage of Foreign-born Population",
                     breaks = seq(0, 1, by = 0.1),
                     labels = paste(prettyNum(seq(0, 100, by = 10)), "%")) +
  ggtitle("Relationship between Percentage of Tenant Population and Nativity") +
  theme(title = element_text(size=9, face = "bold")) +
  annotate("text", x = 0.7, y = 0.85, size=3.5,
           label = paste("Correlation (95-percent confidence)\n:", prettyNum(correlation6$conf.int[1], d
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 14 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 14 rows containing missing values (geom_point).
```



Relationships between a Continuous Variable and a Binary Variable (A two-sample t-test)

1. Relationships between percentage of tenant population and nativity

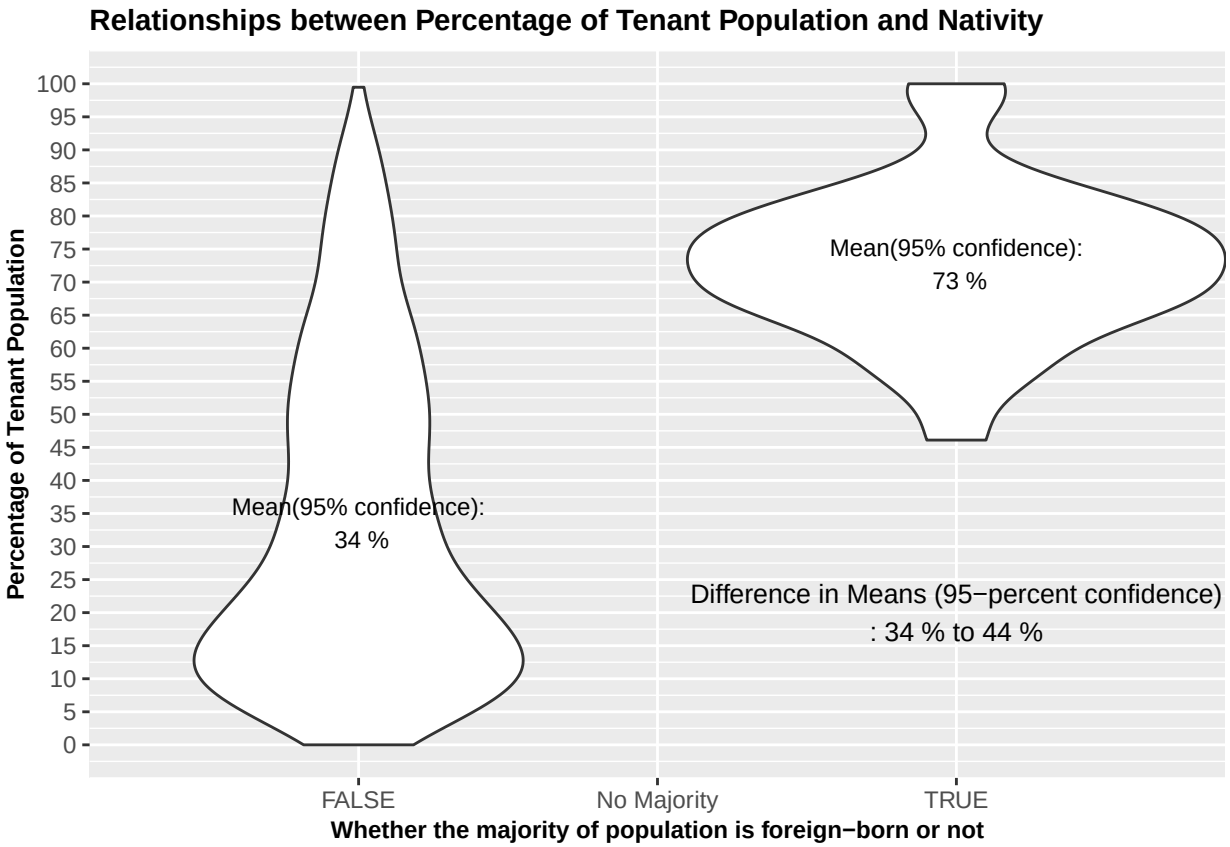
```
difference1 = t.test(pct_ten ~ maj_foreign == "TRUE",
                     data = tractMA2018)
difference1
```

```
##
##  Welch Two Sample t-test
##
## data:  pct_ten by maj_foreign == "TRUE"
## t = -16.166, df = 28.9, p-value = 5.154e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.4418356 -0.3425794
## sample estimates:
## mean in group FALSE  mean in group TRUE
##           0.3357957           0.7280033
```

There is significant relationship found between the percentage of tenant population and nativity; tracts with a majority foreign-born population have a larger percentage of tenant population than majority native-born tracts do.

```
ggplot(tractMA2018, aes(x = maj_foreign, y = pct_ten)) +
  geom_violin() +
  theme_gray() +
  scale_x_discrete(name = "Whether the majority of population is foreign-born or not") +
  scale_y_continuous(name = "Percentage of Tenant Population",
                     breaks = seq(0, 1, by = 0.05),
                     labels = paste(prettyNum(seq(0, 100, by = 5), "%"))) +
  ggtitle("Relationships between Percentage of Tenant Population and Nativity") +
  theme(title = element_text(size=9, face = "bold")) +
  annotate(geom = "text", x=1, y=difference1$estimate[1], size=3,
           label = paste("Mean(95% confidence):\n", prettyNum(difference1$estimate[1]*100, digits=2), "
  annotate(geom = "text", x=3, y=difference1$estimate[2], size=3,
           label = paste("Mean(95% confidence):\n", prettyNum(difference1$estimate[2]*100, digits=2), "
  annotate("text", x = 3, y = 0.2, size=3.5,
           label = paste("Difference in Means (95-percent confidence)\n:", prettyNum(abs(difference1$co
```

```
## Warning: Removed 14 rows containing non-finite values (stat_ydensity).
```



2. Relationships between monthly housing cost and nativity

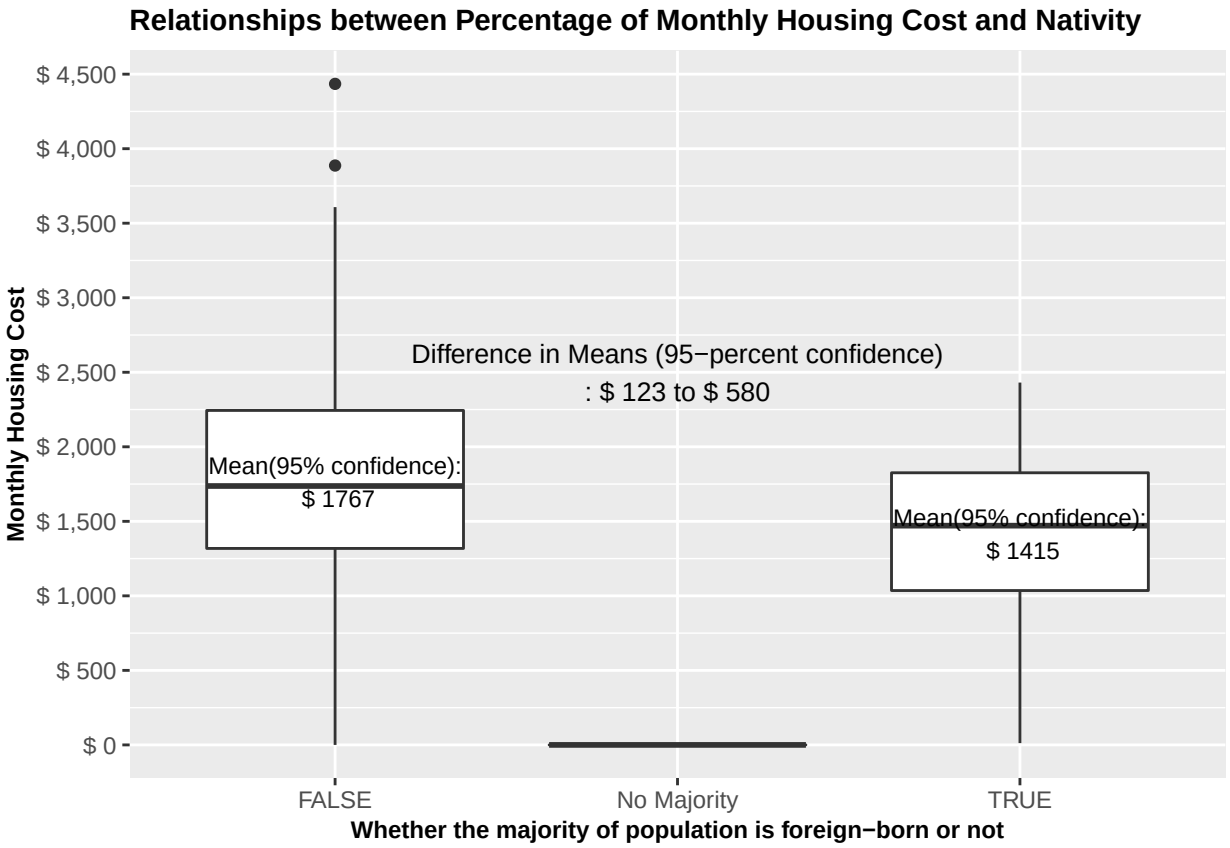
```
difference2 = t.test(mon_housE ~ maj_foreign == "TRUE",
                     data = tractMA2018)
difference2
```

```
##
##  Welch Two Sample t-test
##
## data:  mon_housE by maj_foreign == "TRUE"
## t = 3.1585, df = 26.332, p-value = 0.003955
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  122.9242 580.3045
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            1766.614            1415.000
```

There is significant relationship found between monthly housing cost and nativity; tracts with a majority foreign-born population have less monthly housing cost than majority native-born tracts do.

```
ggplot(tractMA2018, aes(x = maj_foreign, y = mon_housE)) +
  geom_boxplot() +
```

```
  theme_gray() +
  scale_x_discrete(name = "Whether the majority of population is foreign-born or not") +
  scale_y_continuous(name = "Monthly Housing Cost",
                     breaks = seq(0, 5000, by = 500),
                     labels = paste("$", prettyNum(seq(0, 5000, by = 500), big.mark = ","))) +
  ggtitle("Relationships between Percentage of Monthly Housing Cost and Nativity") +
  theme(title = element_text(size=9, face = "bold")) +
  annotate(geom = "text", x=1, y=difference2$estimate[1], size=3,
           label = paste("Mean(95% confidence):\n", "$", prettyNum(difference2$estimate[1], digits=0))) +
  annotate(geom = "text", x=3, y=difference2$estimate[2], size=3,
           label = paste("Mean(95% confidence):\n", "$", prettyNum(difference2$estimate[2], digits=0))) +
  annotate("text", x = 2, y = 2500, size=3.5,
           label = paste("Difference in Means (95-percent confidence)\n:", "$", prettyNum(difference2$c
```



3. Relationships between median income and nativity

```
difference3 = t.test(med_incomeE ~ maj_foreign == "TRUE",
                     data = tractMA2018)
difference3
```

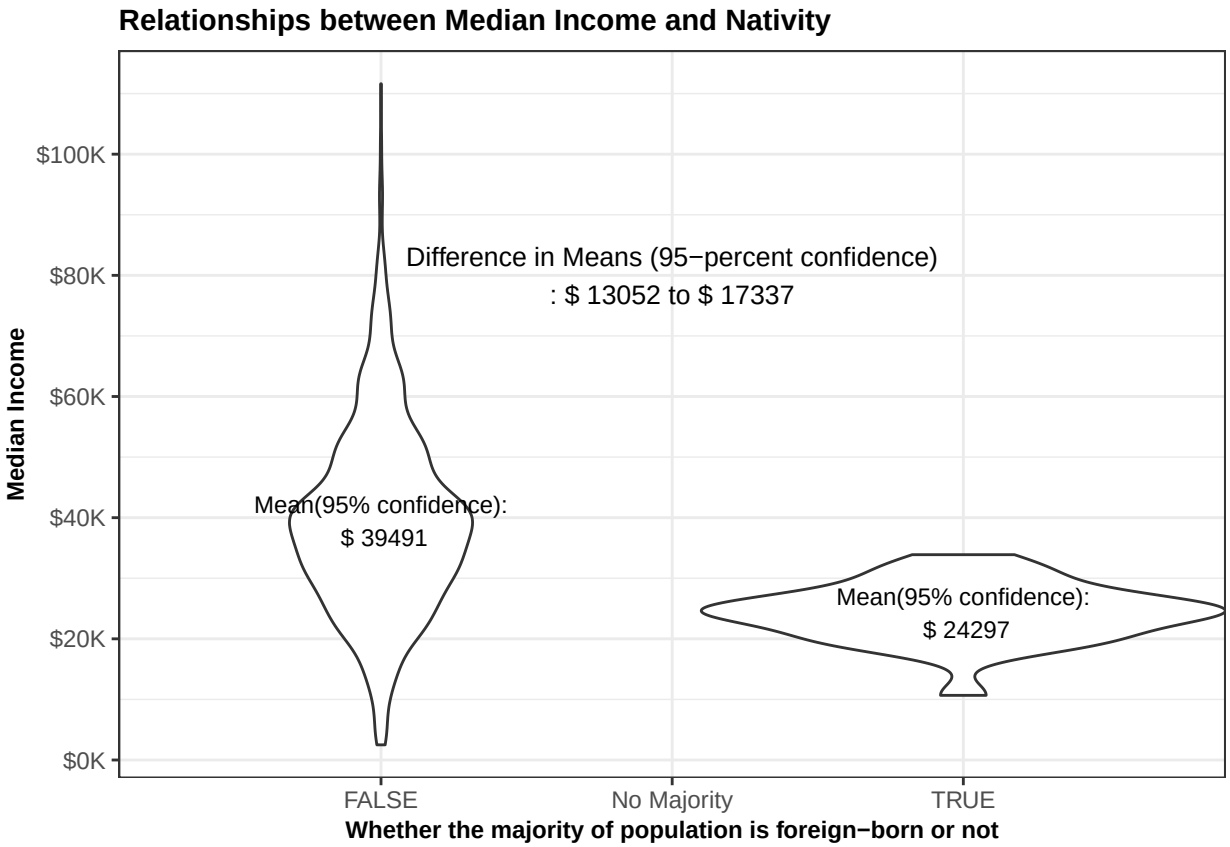
```
##
##  Welch Two Sample t-test
##
```

```
## data:  med_incomeE by maj_foreign == "TRUE"
## t = 14.401, df = 34.625, p-value = 3.476e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  13051.56 17337.10
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            39491.40            24297.08
```

There is significant relationship found between median income and nativity; tracts with a majority foreign-born population have less median income than majority native-born tracts do.

```
ggplot(tractMA2018, aes(x = maj_foreign, y = med_incomeE)) +
  geom_violin() +
  theme_bw() +
  scale_x_discrete(name = "Whether the majority of population is foreign-born or not") +
  scale_y_continuous(name = "Median Income",
                     breaks = seq(0, 100000, by = 20000),
                     labels = paste("$", prettyNum(seq(0, 100, by = 20)), "K", sep = "")) +
  ggtitle("Relationships between Median Income and Nativity") +
  theme(title = element_text(size=9, face = "bold")) +
  annotate("text", x=1, y=difference3$estimate[1], size=3,
           label = paste("Mean(95% confidence):\n", "$", prettyNum(difference3$estimate[1], digits = 1)
  annotate("text", x=3, y=difference3$estimate[2], size=3,
           label = paste("Mean(95% confidence):\n", "$", prettyNum(difference3$estimate[2], digits = 1)
  annotate("text", x = 2, y = 80000, size=3.5,
           label = paste("Difference in Means (95-percent confidence)\n:", "$", prettyNum(difference3$c
```

```
## Warning: Removed 16 rows containing non-finite values (stat_ydensity).
```



This graph is specifically interesting in that the distribution of median income is highly different by nativity.

I don't test the relationship between the nativity and percentage of foreign-population, as the relationship between the two is already obvious.

Relationships between a Continuous Variable and a Categorical Variable with three or more levels (ANOVA)

1. Relationships between percentage of tenant population and transportation mode choice

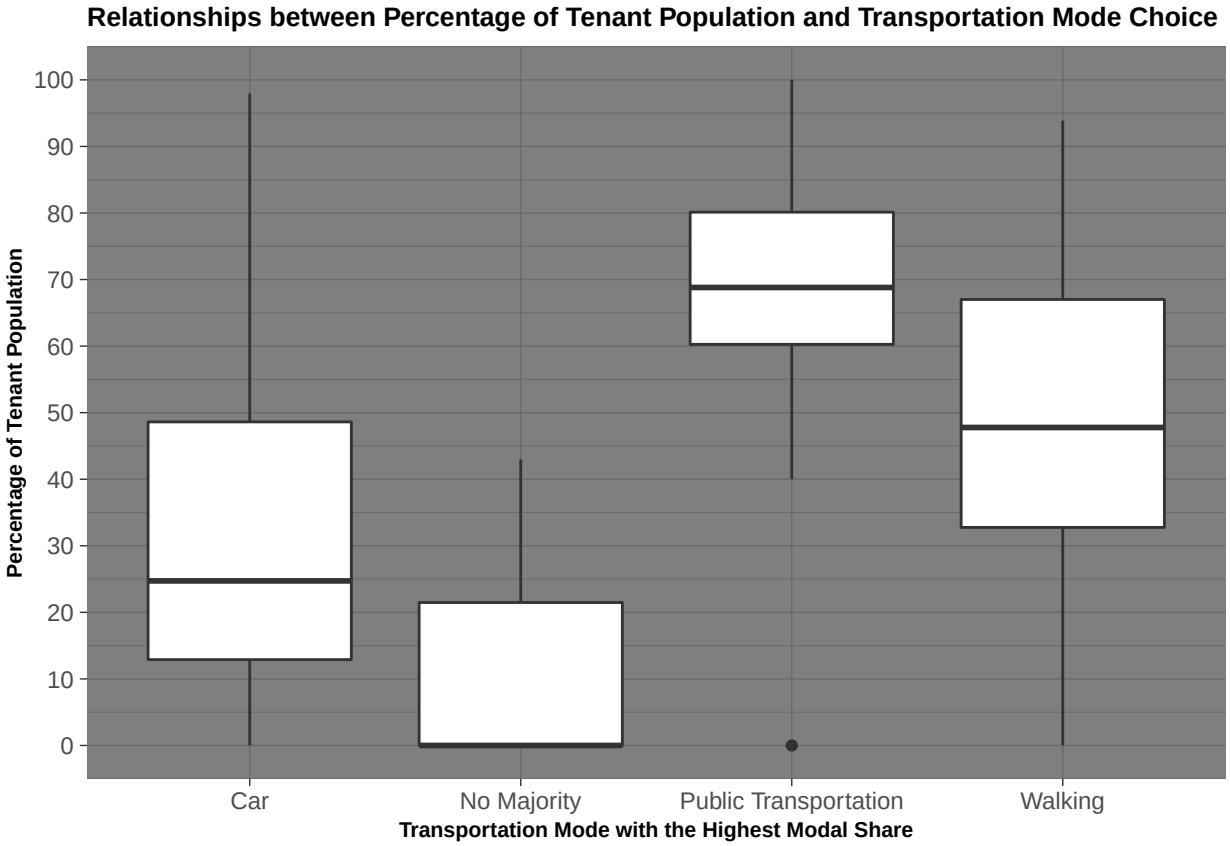
```
anova1 <- aov (pct_ten ~ maj_tra, data = tractMA2018)
summary(anova1)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## maj_tra        3   9.59   3.196   58.48 <2e-16 ***
## Residuals   1460  79.80   0.055
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 14 observations deleted due to missingness
```

There is a significant relationship between percentage of tenant population and transportation mode with the highest modal share.

```r
ggplot(tractMA2018, aes(x = maj_tra, y = pct_ten)) +
  geom_boxplot() +
  theme_dark() +
  ggtitle("Relationships between Percentage of Tenant Population and Transportation Mode Choice") +
  theme(title = element_text(size=8, face = "bold")) +
  scale_x_discrete(name = "Transportation Mode with the Highest Modal Share") +
  scale_y_continuous(name = "Percentage of Tenant Population",
                     breaks = seq(0, 1, by = 0.1),
                     labels = paste(prettyNum(seq(0, 100, by = 10), "%")))
```

```
## Warning: Removed 14 rows containing non-finite values (stat_boxplot).
```



In the tracts in MA where public transportation and walking are the transport modes with the highest share, there is a tendency that the percentage of tenant population is higher than other tracts where not.

2. Relationships between nativity and transportation mode choice

```r
anova2 <- aov (pct_foreign ~ maj_tra, data = tractMA2018)
summary(anova2)
```

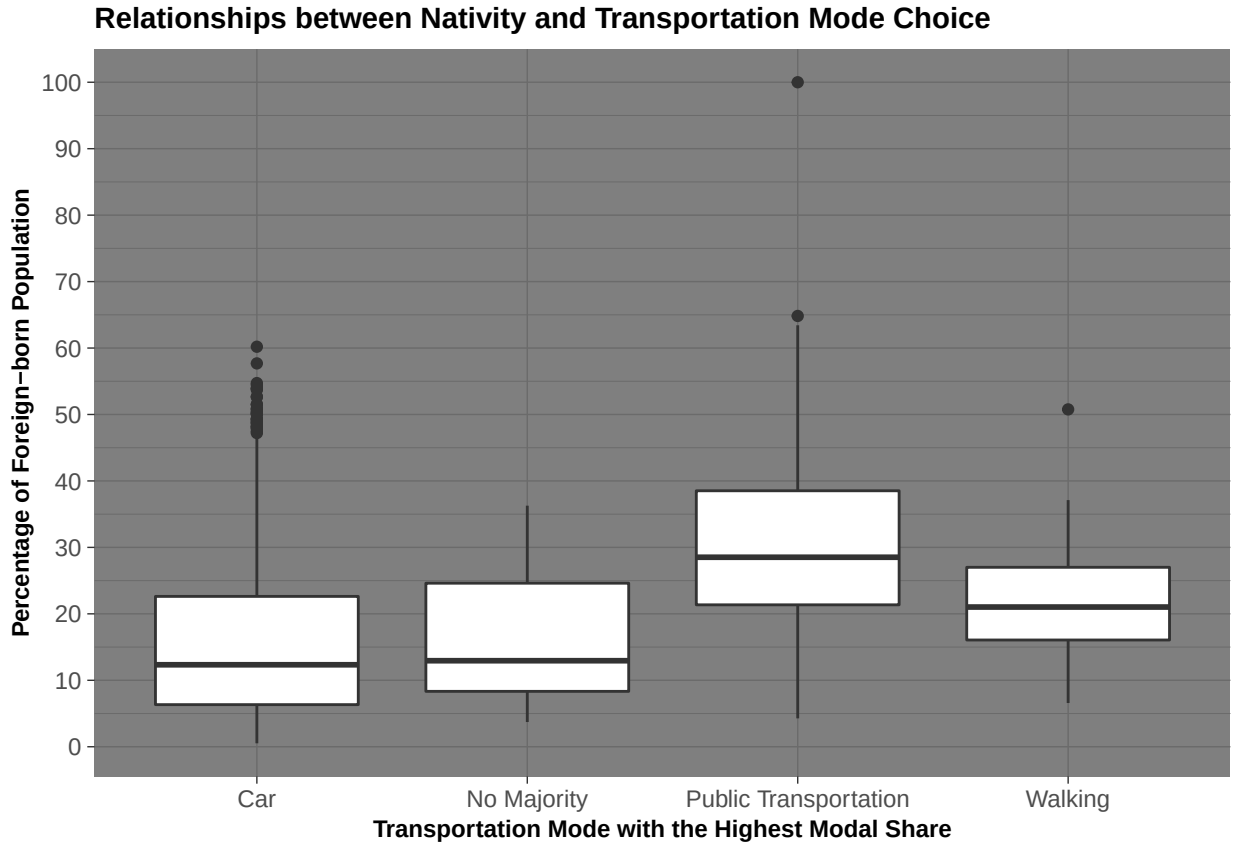
```
##               Df Sum Sq Mean Sq F value Pr(>F)
## maj_tra        3  1.868  0.6227   42.96 <2e-16 ***
## Residuals   1460 21.163  0.0145
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 14 observations deleted due to missingness
```

There is a significant relationship between percentage of foreign-born population and transportation mode with the highest modal share.

```
ggplot(tractMA2018, aes(x = maj_tra, y = pct_foreign)) +
  geom_boxplot() +
  theme_dark() +
  ggtitle("Relationships between Nativity and Transportation Mode Choice") +
  theme(title = element_text(size=9, face = "bold")) +
  scale_x_discrete(name = "Transportation Mode with the Highest Modal Share") +
  scale_y_continuous(name = "Percentage of Foreign-born Population",
                     breaks = seq(0, 1, by = 0.1),
                     labels = paste(prettyNum(seq(0, 100, by = 10), "%")))
```

```
## Warning: Removed 14 rows containing non-finite values (stat_boxplot).
```



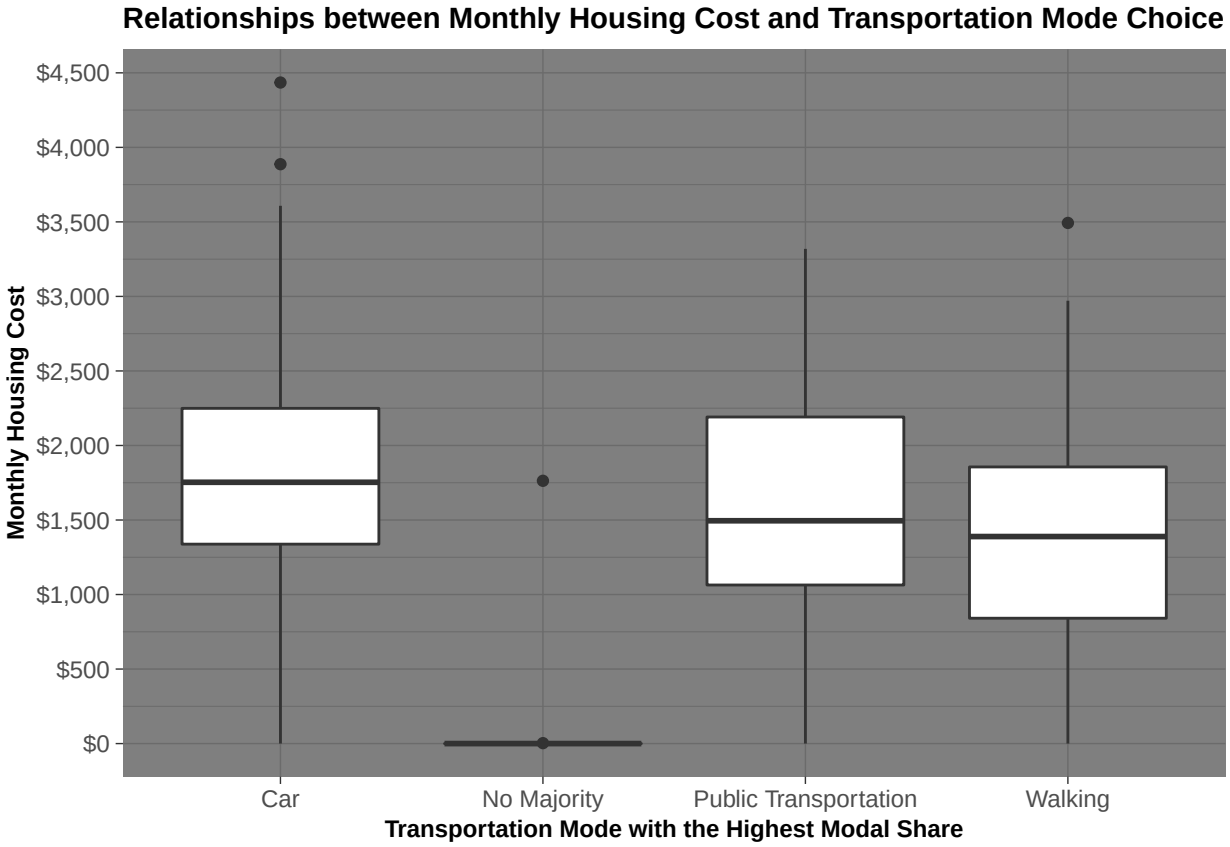
3. Relationships between monthly housing cost and transportation mode choice

```
anova3 <- aov (mon_housE ~ maj_tra, data = tractMA2018)
summary(anova3)
```

```
##                Df    Sum Sq  Mean Sq F value Pr(>F)
## maj_tra         3  59550200 19850067   47.15 <2e-16 ***
## Residuals    1474 620618440   421044
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is a significant relationship between monthly housing cost and transportation mode with the highest modal share.

```
ggplot(tractMA2018, aes(x = maj_tra, y = mon_housE)) +
  geom_boxplot() +
  theme_dark() +
  ggtitle("Relationships between Monthly Housing Cost and Transportation Mode Choice") +
  theme(title = element_text(size=9, face = "bold")) +
  scale_x_discrete(name = "Transportation Mode with the Highest Modal Share") +
  scale_y_continuous(name = "Monthly Housing Cost",
                     breaks = seq(0, 5000, by = 500),
                     labels = paste( "$", prettyNum(seq(0, 5000, by = 500), big.mark = ","), sep = ""))
```



4. Relationships between median income and transportation mode choice

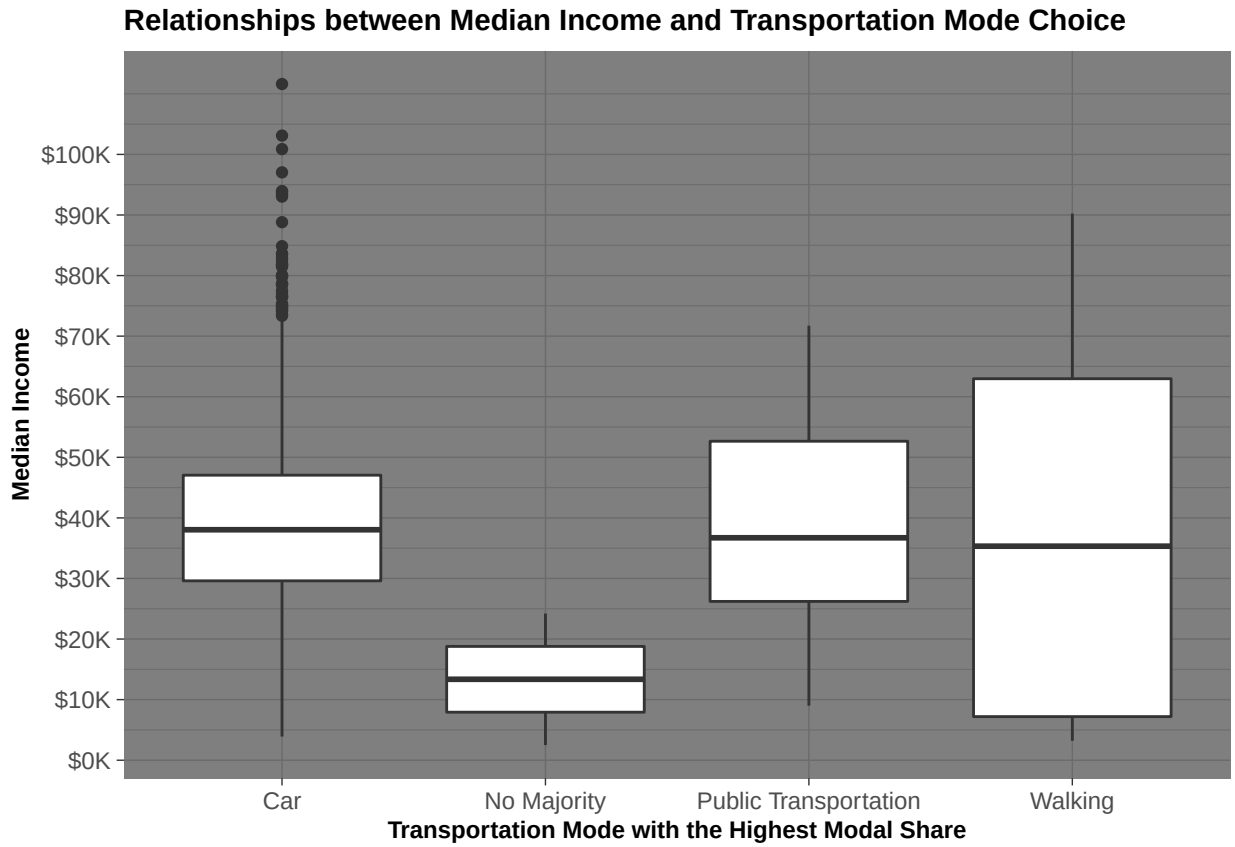
```
anova4 <- aov (med_incomeE ~ maj_tra, data = tractMA2018)
summary(anova4)
```

```
##                Df    Sum Sq   Mean Sq F value Pr(>F)
## maj_tra         3 1.832e+09 610627173   2.544 0.0547 .
## Residuals    1458 3.500e+11 240021564
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 16 observations deleted due to missingness
```

There is no statistically meaningful relationship between median income and transportation mode with the highest modal share.

```
ggplot(tractMA2018, aes(x = maj_tra, y = med_incomeE)) +
  geom_boxplot() +
  theme_dark() +
  ggtitle("Relationships between Median Income and Transportation Mode Choice") +
  theme(title = element_text(size=9, face = "bold")) +
  scale_x_discrete(name = "Transportation Mode with the Highest Modal Share") +
  scale_y_continuous(name = "Median Income",
                     breaks = seq(0, 100000, by = 10000),
                     labels = paste("$", prettyNum(seq(0, 100, by = 10)), "K", sep = ""))
```

```
## Warning: Removed 16 rows containing non-finite values (stat_boxplot).
```



This graph is interesting in that the tracts where walking is the highest modal share have relatively wide distribution of median income level than other modes.

5. Relationships between median income and race/ethnicity

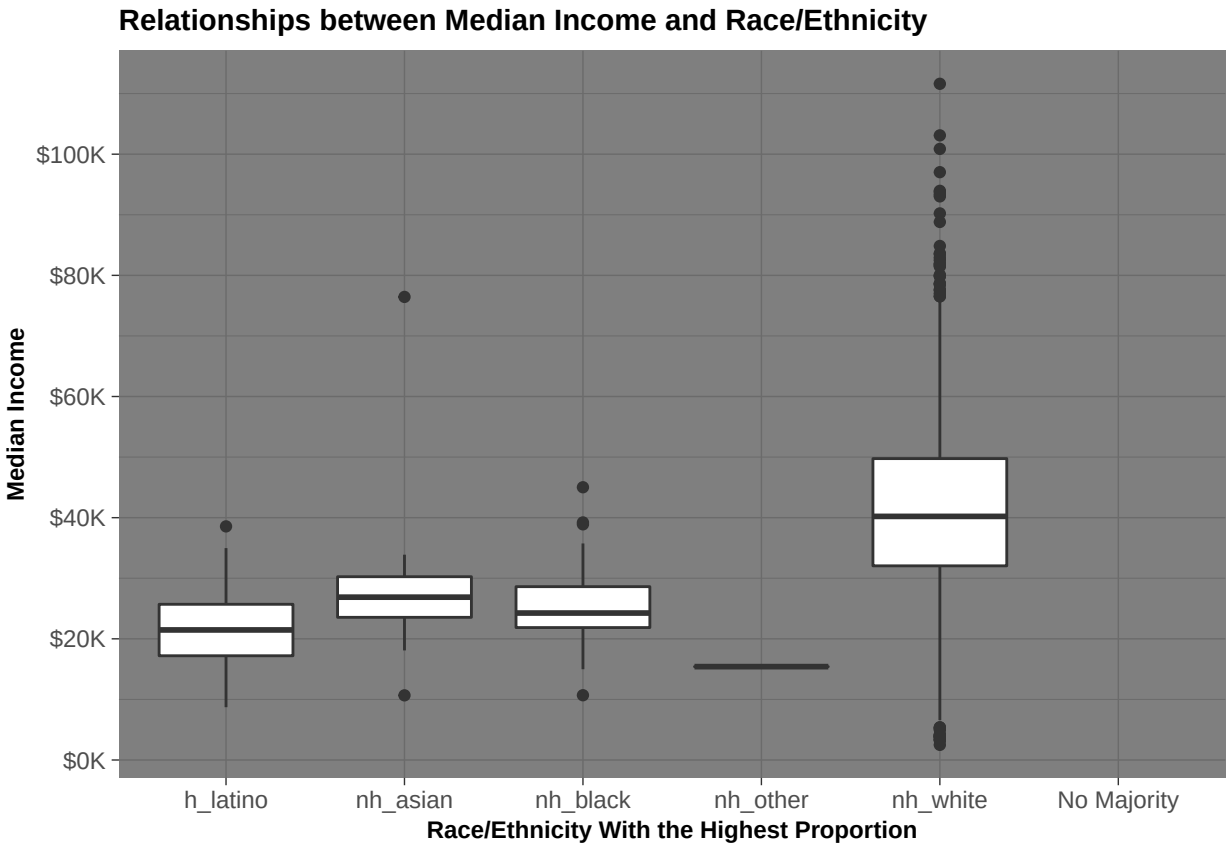
```
anova5 <- aov (med_incomeE ~ maj_race, data = tractMA2018)
summary(anova5)
```

```
##               Df    Sum Sq   Mean Sq F value Pr(>F)
## maj_race       4 5.569e+10 1.392e+10   68.51 <2e-16 ***
## Residuals   1457 2.961e+11 2.032e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 16 observations deleted due to missingness
```

There is a significant relationship between median income and race/ethnicity with the highest proportion.

```
ggplot(tractMA2018, aes(x = maj_race, y = med_incomeE)) +
  geom_boxplot() +
  theme_dark() +
  ggtitle("Relationships between Median Income and Race/Ethnicity") +
  theme(title = element_text(size=9, face = "bold")) +
  scale_x_discrete(name = "Race/Ethnicity With the Highest Proportion") +
  scale_y_continuous(name = "Median Income",
                     breaks = seq(0, 100000, by = 20000),
                     labels = paste("$", prettyNum(seq(0, 100, by = 20)), "K", sep = ""))
```

```
## Warning: Removed 16 rows containing non-finite values (stat_boxplot).
```



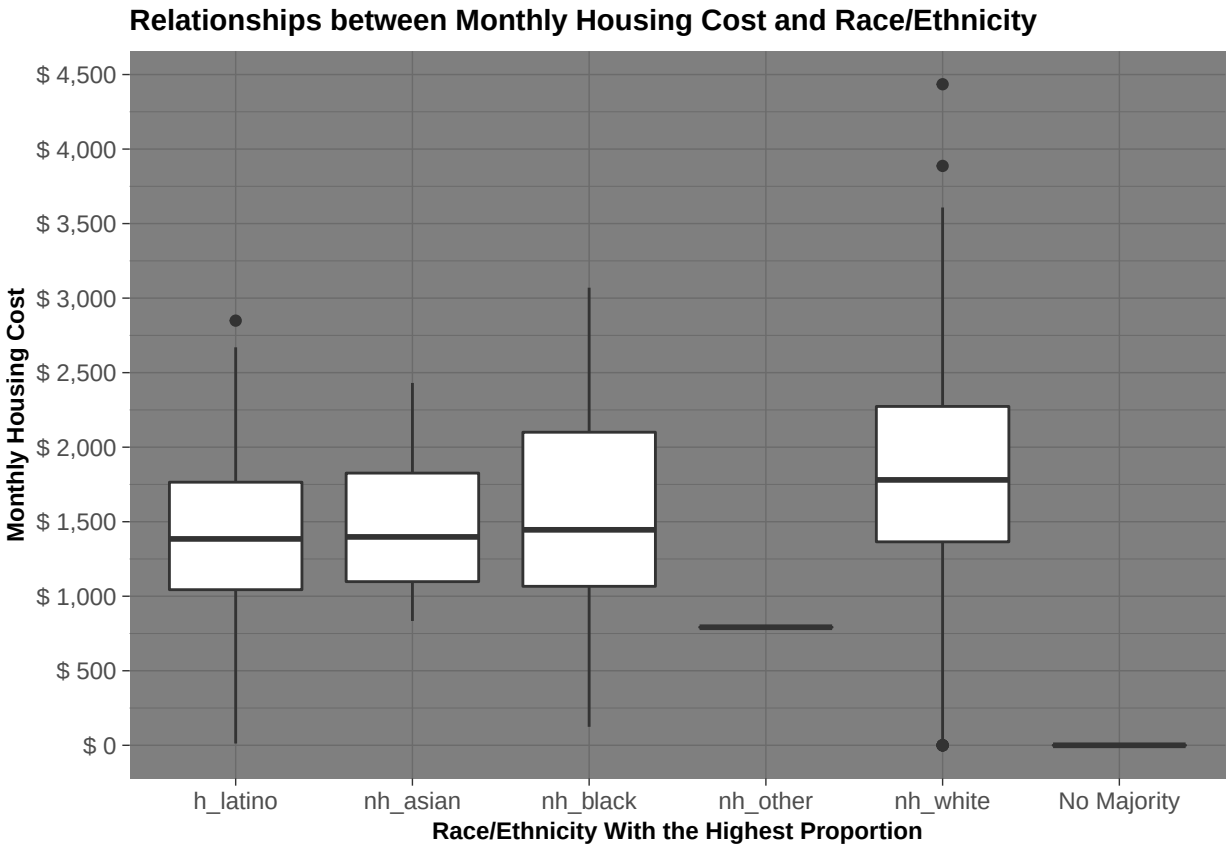
6. Relationships between monthly housing cost and race/ethnicity

```
anova6 <- aov (mon_housE ~ maj_race, data = tractMA2018)
summary(anova6)
```

```
##               Df    Sum Sq  Mean Sq F value Pr(>F)
## maj_race       5  64866388 12973278   31.04 <2e-16 ***
## Residuals   1472 615302252   418004
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is a significant relationship between monthly housing cost and race/ethnicity with the highest proportion.

```
ggplot(tractMA2018, aes(x = maj_race, y = mon_housE)) +
  geom_boxplot() +
  theme_dark() +
  ggtitle("Relationships between Monthly Housing Cost and Race/Ethnicity") +
  theme(title = element_text(size=9, face = "bold")) +
  scale_x_discrete(name = "Race/Ethnicity With the Highest Proportion") +
  scale_y_continuous(name = "Monthly Housing Cost",
                     breaks = seq(0, 5000, by = 500),
                     labels = paste("$", prettyNum(seq(0, 5000, by = 500), big.mark = ",")))
```



7. Relationships between nativity and race/ethnicity

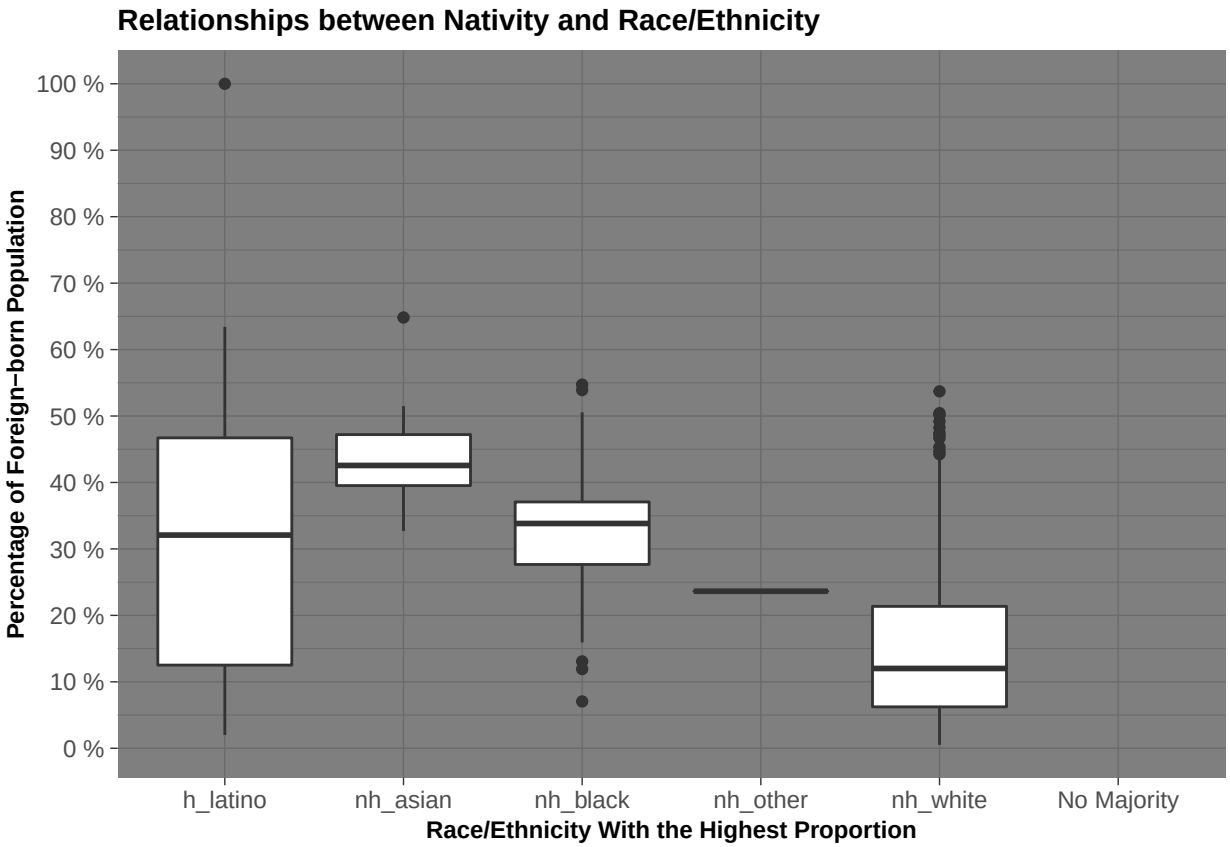
```
anova7 <- aov (pct_foreign ~ maj_race, data = tractMA2018)
summary(anova7)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## maj_race       4  5.385  1.3461   111.3 <2e-16 ***
## Residuals   1459 17.647  0.0121
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 14 observations deleted due to missingness
```

There is a significant relationship between percentage of foreign-born population and race/ethnicity with the highest proportion.

```
ggplot(tractMA2018, aes(x = maj_race, y = pct_foreign)) +
  geom_boxplot() +
  theme_dark() +
  ggtitle("Relationships between Nativity and Race/Ethnicity") +
  theme(title = element_text(size=9, face = "bold")) +
  scale_x_discrete(name = "Race/Ethnicity With the Highest Proportion") +
  scale_y_continuous(name = "Percentage of Foreign-born Population",
                     breaks = seq(0, 1, by = 0.1),
                     labels = paste(prettyNum(seq(0, 100, by = 10)), "%"))
```

```
## Warning: Removed 14 rows containing non-finite values (stat_boxplot).
```



8. Relationships between percentage tenant population and race/ethnicity

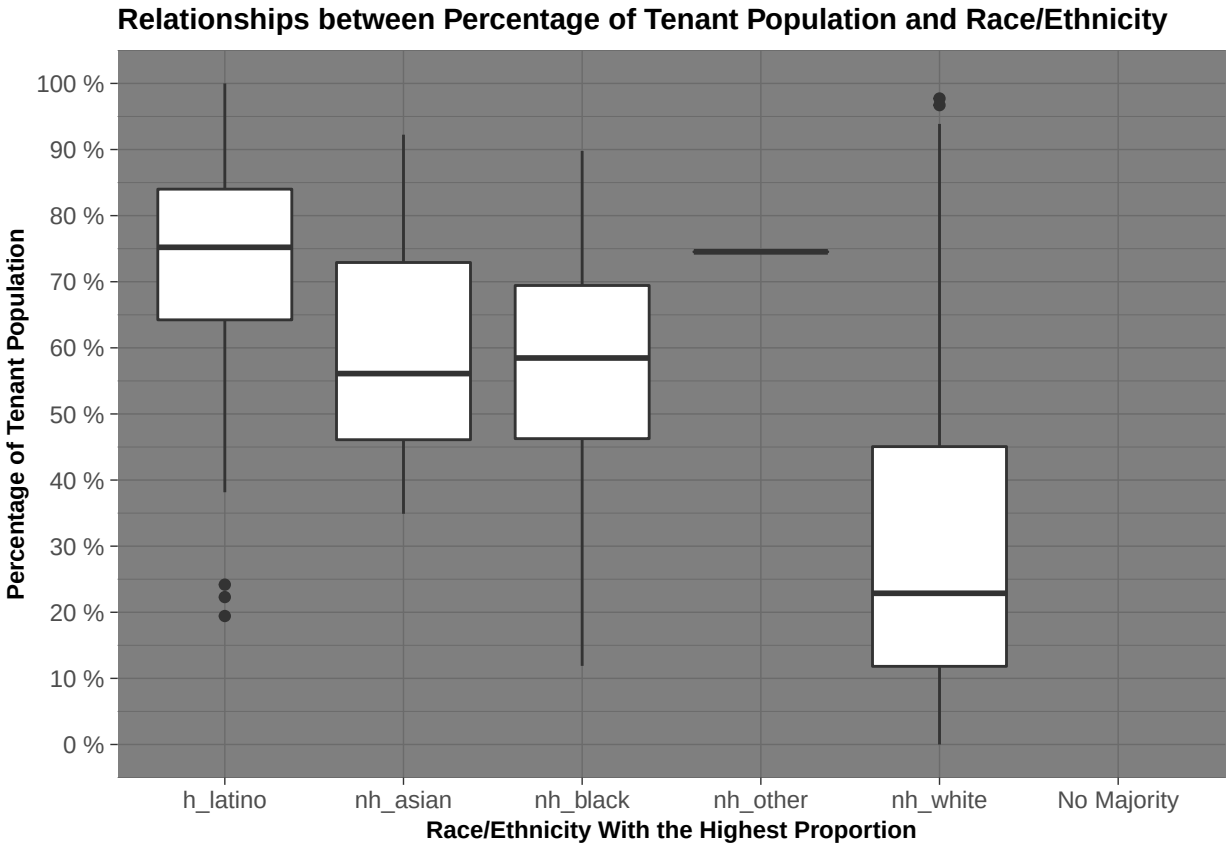
```
anova8 <- aov (pct_ten ~ maj_race, data = tractMA2018)
summary(anova8)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## maj_race       4  23.38   5.845   129.2 <2e-16 ***
## Residuals   1459  66.01   0.045
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 14 observations deleted due to missingness
```

There is a significant relationship between percentage of tenant population and race/ethnicity with the highest proportion.

```
ggplot(tractMA2018, aes(x = maj_race, y = pct_ten)) +
  geom_boxplot() +
  theme_dark() +
  ggtitle("Relationships between Percentage of Tenant Population and Race/Ethnicity") +
  theme(title = element_text(size=9, face = "bold")) +
  scale_x_discrete(name = "Race/Ethnicity With the Highest Proportion") +
  scale_y_continuous(name = "Percentage of Tenant Population",
                     breaks = seq(0, 1, by = 0.1),
                     labels = paste(prettyNum(seq(0, 100, by = 10)), "%"))
```

```
## Warning: Removed 14 rows containing non-finite values (stat_boxplot).
```



This graph is intriguing in that the percentage of tenant population is relatively low in the tracts where non-hispanic white is the race/ethnicity with the highest ratio.

Relationships between Two Categorical Variables (A Chi-Square Test)

1. Relationships between race/ethnicity and transportation mode choice

```
chi_sq1 <- chisq.test(tractMA2018$maj_race, tractMA2018$maj_tra == "Car")
```

```
## Warning in chisq.test(tractMA2018$maj_race, tractMA2018$maj_tra == "Car"): Chi-
## squared approximation may be incorrect
```

```
chi_sq1
```

```
##
##  Pearson's Chi-squared test
##
## data:  tractMA2018$maj_race and tractMA2018$maj_tra == "Car"
## X-squared = 149.44, df = 5, p-value < 2.2e-16
```

```
chi_sq1$observed
```

```
##                      tractMA2018$maj_tra == "Car"
## tractMA2018$maj_race FALSE TRUE
##          h_latino       15   95
##          nh_asian        2   11
##          nh_black        7   55
##          nh_other        0    1
##          nh_white       95 1183
##          No Majority    14    0
```

Here are the values I observe when I assign the value "Car" for the variable "transport mode with the highest modal share". I see that the non-hispanic other and no majority categories violate the chi-square observed table, as they have values less than 5. However, with a p-value so much less than 0.05, we can assume a statistically significant relationship between race/ethnicity and transportation mode choice at a 95% confidence level despite this possible inaccuracy.

```
chi_sq1$expected
```

```
##                      tractMA2018$maj_tra == "Car"
## tractMA2018$maj_race        FALSE         TRUE
##          h_latino      9.89851150  100.1014885
##          nh_asian      1.16982409   11.8301759
##          nh_black      5.57916103   56.4208390
##          nh_other      0.08998647    0.9100135
##          nh_white    115.00270636 1162.9972936
##          No Majority   1.25981055   12.7401894
```

2. Relationships between race/ethnicity and nativity

```
chi_sq2 <- chisq.test(tractMA2018$maj_race, tractMA2018$maj_foreign == TRUE)
```

```
## Warning in chisq.test(tractMA2018$maj_race, tractMA2018$maj_foreign == TRUE):
## Chi-squared approximation may be incorrect
```

```
chi_sq2
```

```
##
##  Pearson's Chi-squared test
##
## data:  tractMA2018$maj_race and tractMA2018$maj_foreign == TRUE
## X-squared = 174.43, df = 5, p-value < 2.2e-16
```

```
chi_sq2$observed
```

```
##
## tractMA2018$maj_race FALSE TRUE
##          h_latino       93   17
```

```
##          nh_asian       10    3
##          nh_black       59    3
##          nh_other        1    0
##          nh_white     1275    3
##          No Majority    14    0
```

Here are the values I observe when I assign the value "TRUE" for the variable "Whether the majority of population is foreign-born or not". I see that the non-hispanic other violates the chi-square observed table, as it has a value less than 5. However, with a p-value so much less than 0.05, we can assume a statistically significant relationship between race/ethnicity and nativity at a 95% confidence level despite this possible inaccuracy.

```
chi_sq2$expected
```

```
##
## tractMA2018$maj_race        FALSE        TRUE
##          h_latino     108.0649526  1.93504736
##          nh_asian      12.7713126  0.22868742
##          nh_black      60.9093369  1.09066306
##          nh_other       0.9824087  0.01759134
##          nh_white    1255.5182679 22.48173207
##          No Majority   13.7537212  0.24627876
```

3. Relationships between nativity and transportation mode choice

```
chi_sq3 <- chisq.test(tractMA2018$maj_foreign, tractMA2018$maj_tra == "Car")
```

```
## Warning in chisq.test(tractMA2018$maj_foreign, tractMA2018$maj_tra == "Car"):
## Chi-squared approximation may be incorrect
```

```
chi_sq3
```

```
##
##  Pearson's Chi-squared test
##
## data:  tractMA2018$maj_foreign and tractMA2018$maj_tra == "Car"
## X-squared = 165.61, df = 2, p-value < 2.2e-16
```

```
chi_sq3$observed
```

```
##                       tractMA2018$maj_tra == "Car"
## tractMA2018$maj_foreign FALSE TRUE
##             FALSE         110 1328
##             No Majority    14    0
##             TRUE            9   17
```

With a p-value so much less than 0.05, we can assume a statistically significant relationship between nativity and transportation mode choice at a 95% confidence level.

```
chi_sq3$expected
```

```
##                          tractMA2018$maj_tra == "Car"
## tractMA2018$maj_foreign      FALSE       TRUE
##             FALSE       129.400541 1308.59946
##             No Majority   1.259811   12.74019
##             TRUE          2.339648   23.66035
```