Quant Assignment 3

Kristy Henrich

9/18/2020

Setup & load libraries

For this assignment, I took advantage of Carole's tutorial and Cat's helpful assignment example. I appreciate how she annotates her file clearly, which is helping me learn as I am working on quant assignments.

Relationships between continuous variables

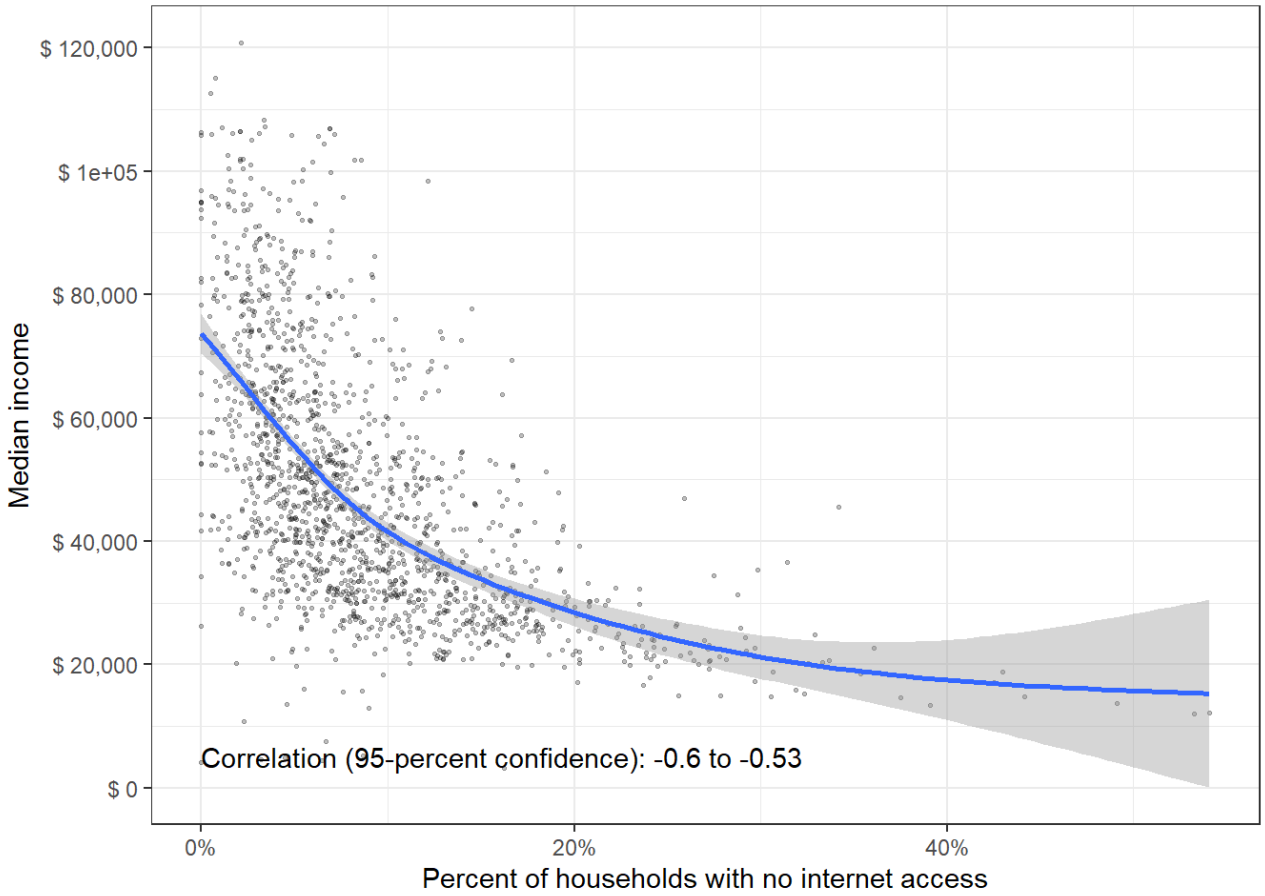
Run correlation tests to test the relationships between three continuous variables:
1. Percent of population that has no internet access : no_internet
2. Median income : med_incomeE
3. Average household size : avg_hh_sizeE

1. Percent no internet & median income

```
correlation1 <- cor.test(tract_data$pct_no_internet, tract_data$med_incomeE)
correlation1
```

```
ggplot(tract_data, aes(x = pct_no_internet, y = med_incomeE)) +
  geom_point(size = 0.5, alpha = 0.25, position = "jitter") +
  geom_smooth() +
  theme_bw() +
  scale_x_continuous(name = "Percent of households with no internet access",
                     labels = scales::percent) +
  scale_y_continuous(name = "Median income",
                     breaks = seq(0, 120000, by = 20000),
                     labels = paste("$",
                                    prettyNum(seq(0, 120000, by = 20000),
                                              big.mark = ","))) +
  annotate(geom = "text", x = 0, y = 5000,
           label = paste("Correlation (95-percent confidence):",
                         prettyNum(correlation1$conf.int[1], digits = 2),
                         "to",
                         prettyNum(correlation1$conf.int[2], digits = 2)),
           hjust = 0)
```

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



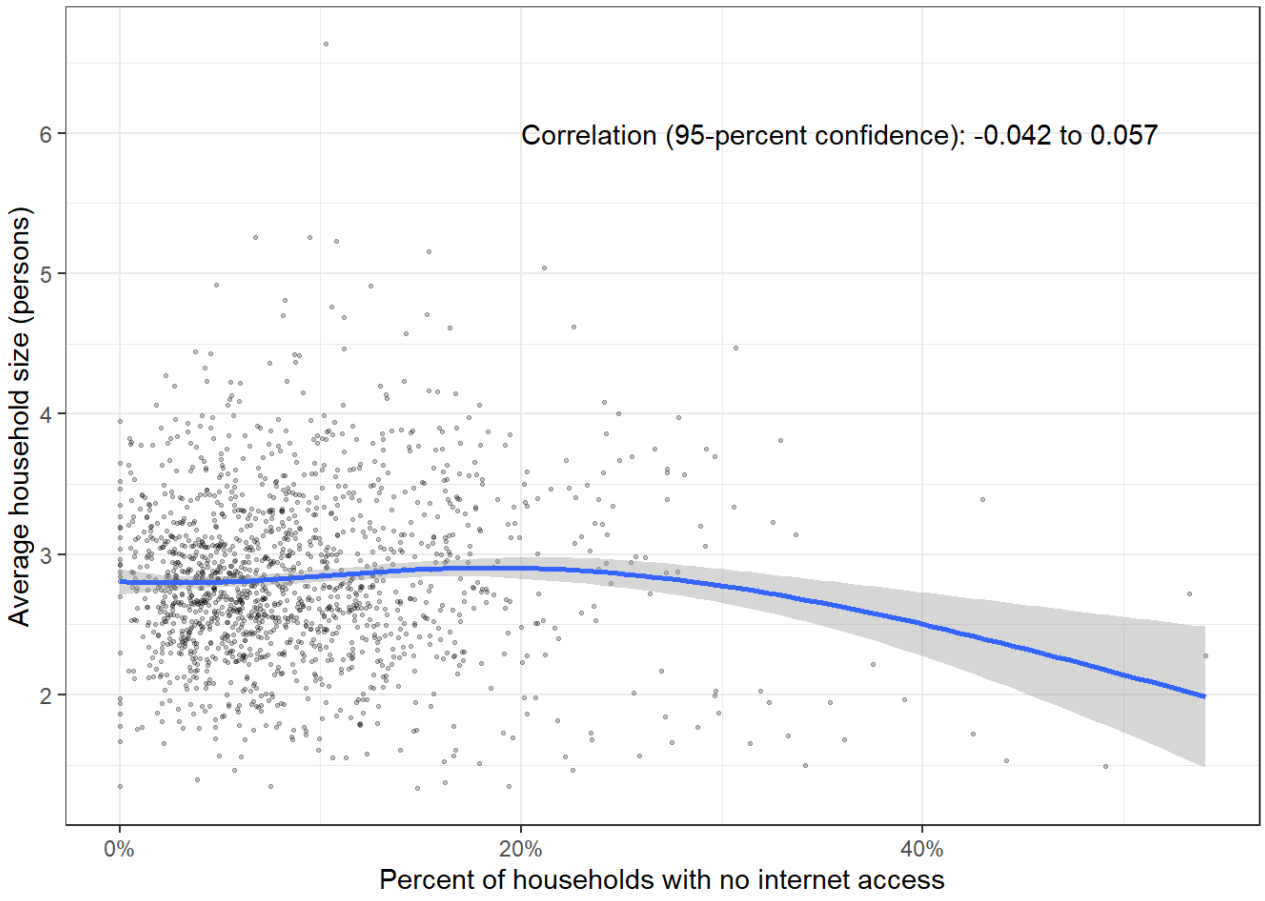
The correlation test found the following relationship between percent of households with no internet access & median income: a moderate, but statistically signficant, negative correlation at a 95% confidence level. This relationship is meaningful in the real world as we think about questions of internet access and economic opportunity

2. Percent no internet & average household size

```
correlation2 <- cor.test(tract_data$pct_no_internet, tract_data$avg_hh_sizeE)
correlation2
```

```
ggplot(tract_data, aes(x = pct_no_internet, y = avg_hh_sizeE)) +
  geom_point(size = 0.5, alpha = 0.25, position = "jitter") +
  geom_smooth() +
  theme_bw() +
  scale_x_continuous(name = "Percent of households with no internet access",
                     labels = scales::percent) +
  scale_y_continuous(name = "Average household size (persons)",
                     breaks = seq(0, 10, by = 1)) +
  annotate(geom = "text", x = .2, y = 6,
           label = paste("Correlation (95-percent confidence):",
                         prettyNum(correlation2$conf.int[1], digits = 2),
                         "to",
                         prettyNum(correlation2$conf.int[2], digits = 2)),
           hjust = 0)
```

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



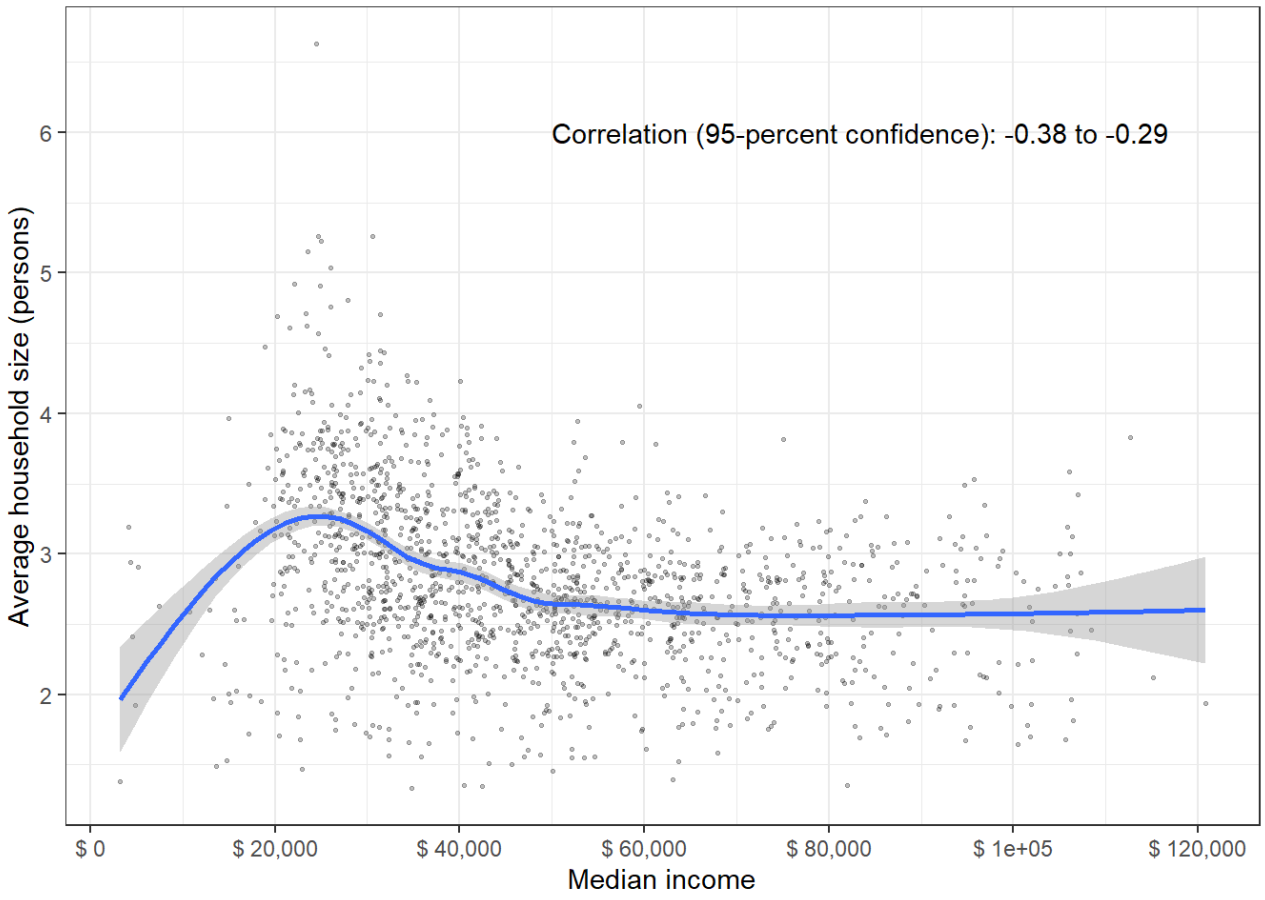
The correlation test found the following relationship between percent of households with no internet access & household size: no statistically significant relationship

3. Median income & average household size

```
correlation3 <- cor.test(tract_data$med_incomeE, tract_data$avg_hh_sizeE)
correlation3
```

```
ggplot(tract_data, aes(x = med_incomeE, y = avg_hh_sizeE)) +
  geom_point(size = 0.5, alpha = 0.25, position = "jitter") +
  geom_smooth() +
  theme_bw() +
  scale_x_continuous(name = "Median income",
                     breaks = seq(0, 120000, by = 20000),
                     labels = paste("$",
                                    prettyNum(seq(0, 120000, by = 20000),
                                              big.mark = ","))) +
  scale_y_continuous(name = "Average household size (persons)",
                     breaks = seq(0, 10, by = 1)) +
  annotate(geom = "text", x = 50000, y = 6,
           label = paste("Correlation (95-percent confidence):",
                         prettyNum(correlation3$conf.int[1], digits = 2),
                         "to",
                         prettyNum(correlation3$conf.int[2], digits = 2)),
           hjust = 0)
```

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



The correlation test found the following relationship between median income & houeshold size: a weak, but statistically significant, negative correlation at a 95% confidence level. This relationship appears to be somewhat important for remaining analyses, though not in a huge manner

Relationships between categorical variables

Run chi-square tests to test the relationships between three categorical variables:
1. Bay Area county : county
2. Majority race / ethnicity of the population : maj_re
3. Whether the majority of the units are occupied by renters vs. owners : maj_units_rented

4. County & majority race/ethnicity

```
chi_sq_county_race <- chisq.test(tract_data$county, tract_data$maj_re)
chi_sq_county_race
chi_sq_county_race$observed
chi_sq_county_race$expected
```

Several of my categories violate the chi-square observed table. However, the p-value is so much less than 0.05, so I have assumed a statistically significant relationship between county and majority race/ethnicity.

5. County & majority units rented

```
chi_sq_county_rent <- chisq.test(tract_data$county, tract_data$maj_units_rented)
chi_sq_county_rent
chi_sq_county_rent$observed
chi_sq_county_rent$expected
```

None of my categories violate the chi-square observed table, and the p-value is much less than 0.05. There is a statistically significant relationship between county and whether a majority of units are rented (TRUE) vs. owned (FALSE), and the real-world implications on renting vs. ownership are quite interesting

6. Majority race/ethnicity & majority units rented

```
chi_sq_race_rent <- chisq.test(tract_data$maj_re, tract_data$maj_units_rented)
chi_sq_race_rent
chi_sq_race_rent$observed
chi_sq_race_rent$expected
```

One data point (counties that are majority Black and do not have a majority of units rented); however, the p-value is much less than 0.05. I will conclude that there is a statistically significant relationship between majority race/ethnicity and whether a majority of units are rented (TRUE) vs. owned (FALSE). Additionally, this conclusion does lead to some interesting real world follow up questions such as are these tracts with rent vs. own by race/ethnicity clustered in certain counties?

Relationship between continuous variables and rent (two-category categorical variable)

Run two-sample t-tests to test the relationships between whether a majority of the units are rented and three continuous variables:

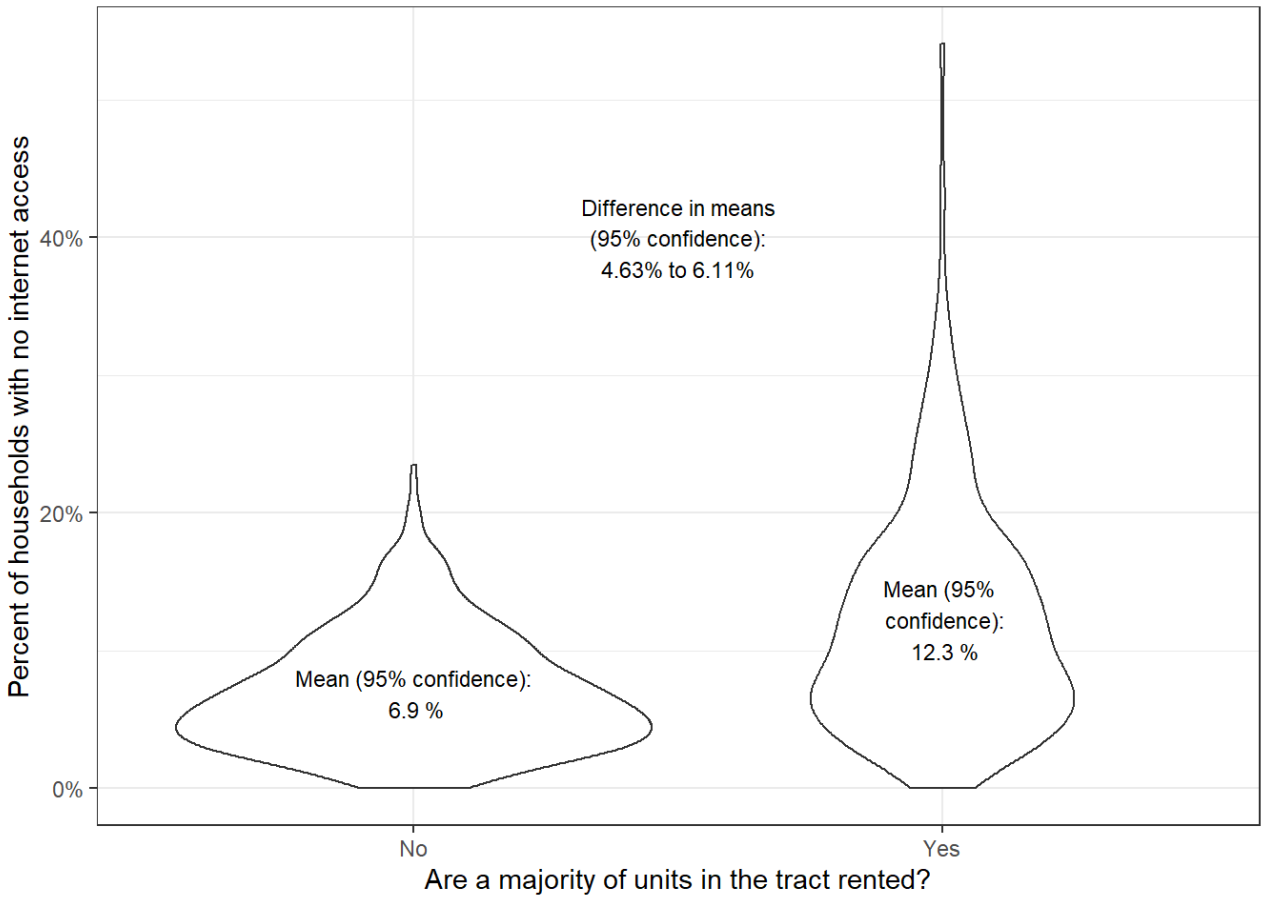
1. Percent of population that has no internet access : no_internet
2. Median income : med_income
3. Average household size : avg_hh_size

7. Percent of households without access to internet & whether a majority of the units are rented

```
difference1 = t.test(pct_no_internet ~ maj_units_rented == "TRUE",
                     data = tract_data)
difference1
```

```
##
##  Welch Two Sample t-test
##
## data:  pct_no_internet by maj_units_rented == "TRUE"
## t = -14.315, df = 753.12, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.06108018 -0.04634814
## sample estimates:
## mean in group FALSE  mean in group TRUE
##          0.06895253          0.12266669
```

```
ggplot(tract_data, aes(x = maj_units_rented, y = pct_no_internet)) +
  geom_violin() +
  theme_bw() +
  scale_x_discrete(name = "Are a majority of units in the tract rented?",
                   labels = c("No", "Yes")) +
  scale_y_continuous(name = "Percent of households with no internet access",
                     labels = scales::percent) +
  annotate(geom = "text", x = 1.5, y = .4, size = 3,
           label = paste("Difference in means\n(95% confidence):\n",
                         prettyNum(abs(difference1$conf.int[2])*100,
                                   digits = 3), "% to ",
                         prettyNum(abs(difference1$conf.int[1])*100,
                                   digits = 3), "%",
                         sep="")) +
  annotate(geom = "text", x = 1, y = difference1$estimate[1], size = 3,
           label = paste("Mean (95% confidence):\n",
                         prettyNum(difference1$estimate[1]*100, digits =3), "%")) +
  annotate(geom = "text", x = 2, y = difference1$estimate[2], size = 3,
           label = paste("Mean (95% \n confidence):\n",
                         prettyNum(difference1$estimate[2]*100, digits =3), "%"))
```



Zero is not included in the 95 percent confidence interval and the p-value is much less than 0.05, so I can conclude there is a statistically significant relationship between majority units rented & percent without internet: tracts with a majority of units rented have a greater percent of households with no internet access. The percent of majority rent-based tracts without internet access is an impactful amount larger

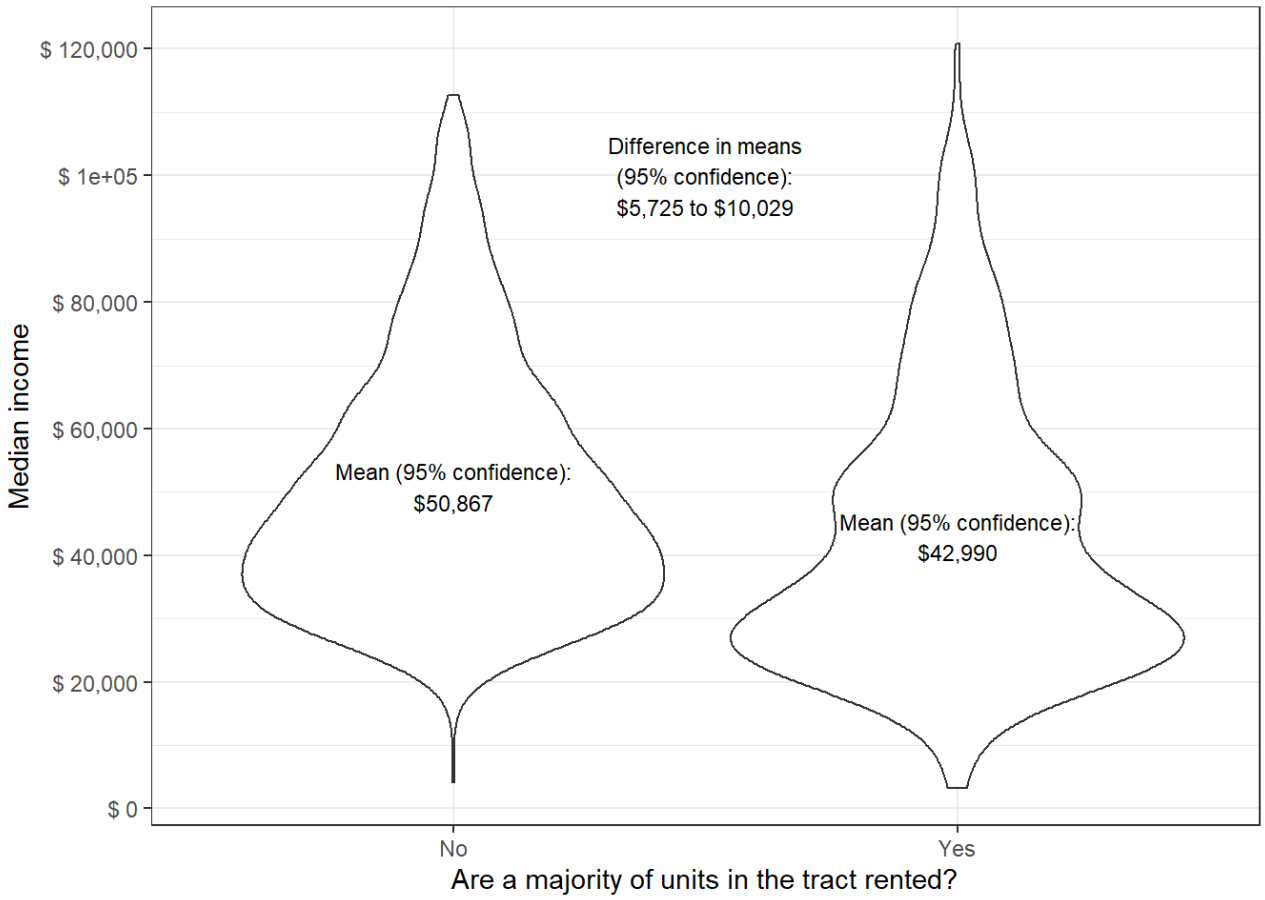
8. Median income & whether a majority of the units are rented

```
difference2 = t.test(med_incomeE ~ maj_units_rented == "TRUE",
                     data = tract_data)
difference2
```

```
##
##  Welch Two Sample t-test
##
## data:  med_incomeE by maj_units_rented == "TRUE"
## t = 7.1827, df = 1099.8, p-value = 1.257e-12
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   5725.331 10028.975
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            50867.17            42990.02
```

```
ggplot(tract_data, aes(x = maj_units_rented, y = med_incomeE)) +
  geom_violin() +
  theme_bw() +
  scale_x_discrete(name = "Are a majority of units in the tract rented?",
                   labels = c("No", "Yes")) +
  scale_y_continuous(name = "Median income",
                     breaks = seq(0, 120000, by = 20000),
                     labels = paste("$",
                                    prettyNum(seq(0, 120000, by = 20000),
                                              big.mark = ","))) +
  annotate(geom = "text", x = 1.5, y = 100000, size = 3,
           label = paste("Difference in means\n(95% confidence):\n$",
                         prettyNum(abs(difference2$conf.int[1]),
                                   digits = 0, big.mark = ","), " to $",
                         prettyNum(abs(difference2$conf.int[2]),
                                   digits = 5, big.mark = ","),
                         sep="")) +
  annotate(geom = "text", x = 1, y = difference2$estimate[1], size = 3,
           label = paste("Mean (95% confidence):\n$",
                         prettyNum(difference2$estimate[1], digits = 5, big.mark = ","
),
                         sep="")) +
  annotate(geom = "text", x = 2, y = difference2$estimate[2], size = 3,
           label = paste("Mean (95% confidence):\n$",
                         prettyNum(difference2$estimate[2], digits = 5, big.mark = ","
),
                         sep=""))
```

```
## Warning: Removed 1 rows containing non-finite values (stat_ydensity).
```



Zero is not included in the 95 percent confidence interval and the p-value is much less than 0.05, so I can conclude there is a statistically significant relationship between majority units rented & median income: tracts with a majority of units rented have a lower median income. This income difference is quite meaningful in its real-world implications

9. Average household size & whether a majority of the units are rented

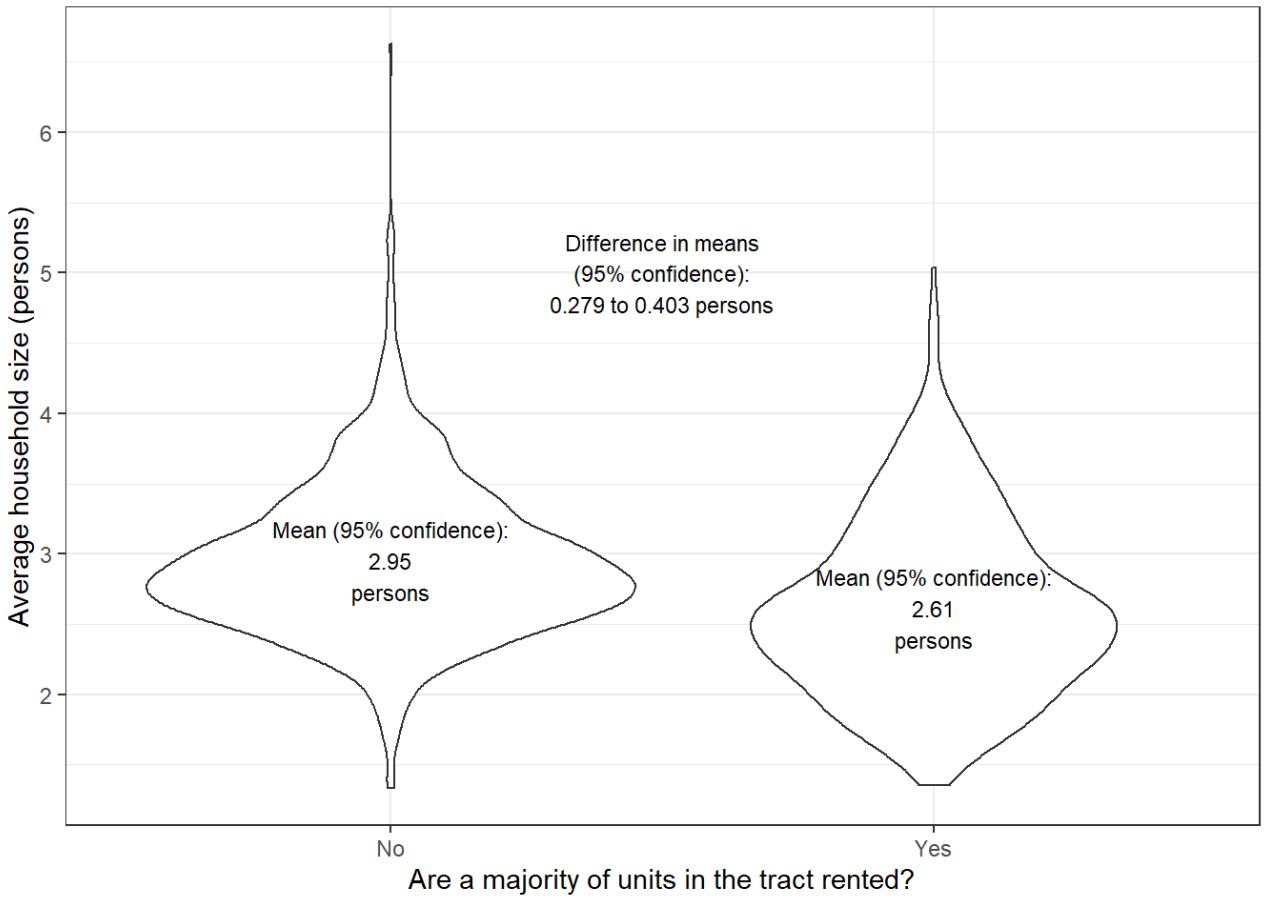
```
difference3 = t.test(avg_hh_sizeE ~ maj_units_rented == "TRUE",
                     data = tract_data)
difference3
```

```
##
##  Welch Two Sample t-test
##
## data:  avg_hh_sizeE by maj_units_rented == "TRUE"
## t = 10.806, df = 1073.8, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.2788540 0.4025868
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            2.954589            2.613869
```

```
ggplot(tract_data, aes(x = maj_units_rented, y = avg_hh_sizeE)) +
  geom_violin() +
  theme_bw() +
  scale_x_discrete(name = "Are a majority of units in the tract rented?",
                   labels = c("No", "Yes")) +
  scale_y_continuous(name = "Average household size (persons)",
                     breaks = seq(0, 10, by = 1)) +
  annotate(geom = "text", x = 1.5, y = 5, size = 3,
           label = paste("Difference in means\n(95% confidence):\n",
                         prettyNum(abs(difference3$conf.int[1]), digits = 3), " to ",
                         prettyNum(abs(difference3$conf.int[2]), digits = 3), " person
s",
                         sep="")) +
  annotate(geom = "text", x = 1, y = difference3$estimate[1], size = 3,
           label = paste("Mean (95% confidence):\n",
                         prettyNum(difference3$estimate[1], digits = 3), "\npersons"),
                         sep="") +
  annotate(geom = "text", x = 2, y = difference3$estimate[2], size = 3,
           label = paste("Mean (95% confidence):\n",
                         prettyNum(difference3$estimate[2], digits = 3), "\npersons"),
                         sep="")
```

```
## Warning: Ignoring unknown parameters: sep

## Warning: Ignoring unknown parameters: sep
```



Zero is not included in the 95 percent confidence interval and the p-value is much less than 0.05, so I can conclude there is a statistically significant relationship between majority units rented & household size: tracts with a majority of units rented have a smaller household size (though not by much in real-life terms)

Relationship between continuous variables and two categorical variables (county and race/ethnicity)

I use the anova test to find the relationship between county and race alone with each of my continuous variables: age, unemployment, and educational attainment. I followed each with a Tukey HSD test to take a closer look

10. County & percent of households with no internet access

```
anova1 <- aov (pct_no_internet ~ county, data = tract_data)
summary(anova1)
```

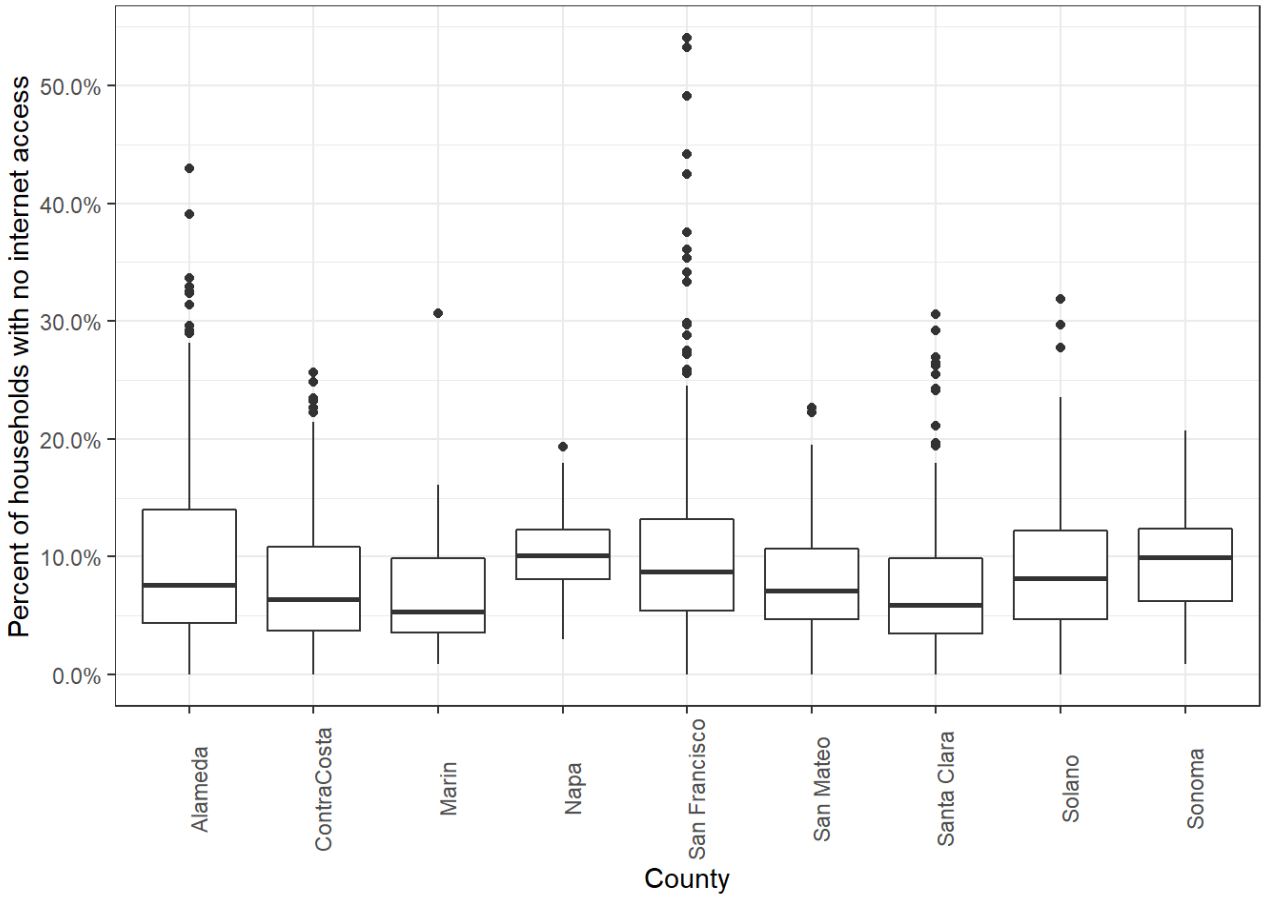
```
##               Df Sum Sq Mean Sq F value Pr(>F)
## county         8  0.299 0.03742   8.792  8e-12 ***
## Residuals   1568  6.674 0.00426
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
differences <- TukeyHSD(anova1)

as_tibble(cbind(pair = row.names(differences$county),
                differences$county))
```

```
## # A tibble: 36 x 5
##    pair            diff            lwr            upr            `p adj`
##    <chr>           <chr>           <chr>          <chr>          <chr>
##  1 ContraCosta-Al~ -0.01995513832~ -0.0376308985~ -0.0022793781~ 0.013778723125~
##  2 Marin-Alameda   -0.02872431508~ -0.0582959295~ 0.00084729937~ 0.064724063727~
##  3 Napa-Alameda    0.004589354273~ -0.0291838094~ 0.03836251794~ 0.999972959996~
##  4 San Francisco-~ 0.012351162660~ -0.0056666028~ 0.03036892817~ 0.453537636869~
##  5 San Mateo-Alam~ -0.01840478228~ -0.0378285867~ 0.00101902212~ 0.079554677688~
##  6 Santa Clara-Al~ -0.02677539073~ -0.0417569085~ -0.0117938729~ 1.187639349642~
##  7 Solano-Alameda  -0.00778037667~ -0.0312516010~ 0.01569084765~ 0.983042116581~
##  8 Sonoma-Alameda  -0.00454613692~ -0.0275425706~ 0.01845029675~ 0.999537637656~
##  9 Marin-ContraCo~ -0.00876917675~ -0.0397334717~ 0.02219511822~ 0.994000057685~
## 10 Napa-ContraCos~ 0.024544492600~ -0.0104545612~ 0.05954354640~ 0.420439302867~
## # ... with 26 more rows
```

```
ggplot(tract_data, aes(x = county, y = pct_no_internet)) +
  geom_boxplot() +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_x_discrete(name = "County") +
  scale_y_continuous(name = "Percent of households with no internet access",
                     breaks = seq(0, 1, by = .1),
                     labels = scales::percent)
```



There is a statistically significant association at the 95 percent confidence level between county & percent of households with no internet access. This plot shows some interesting results with Alameda county showing the greatest interquartile range, but some other counties also showing fairly large interquartile ranges

11. County & median income

```
anova2 <- aov (med_incomeE ~ county, data = tract_data)
summary(anova2)
```

```
##               Df    Sum Sq   Mean Sq F value Pr(>F)
## county         8 5.984e+10 7.480e+09   19.02 <2e-16 ***
## Residuals   1567 6.162e+11 3.932e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 1 observation deleted due to missingness
```

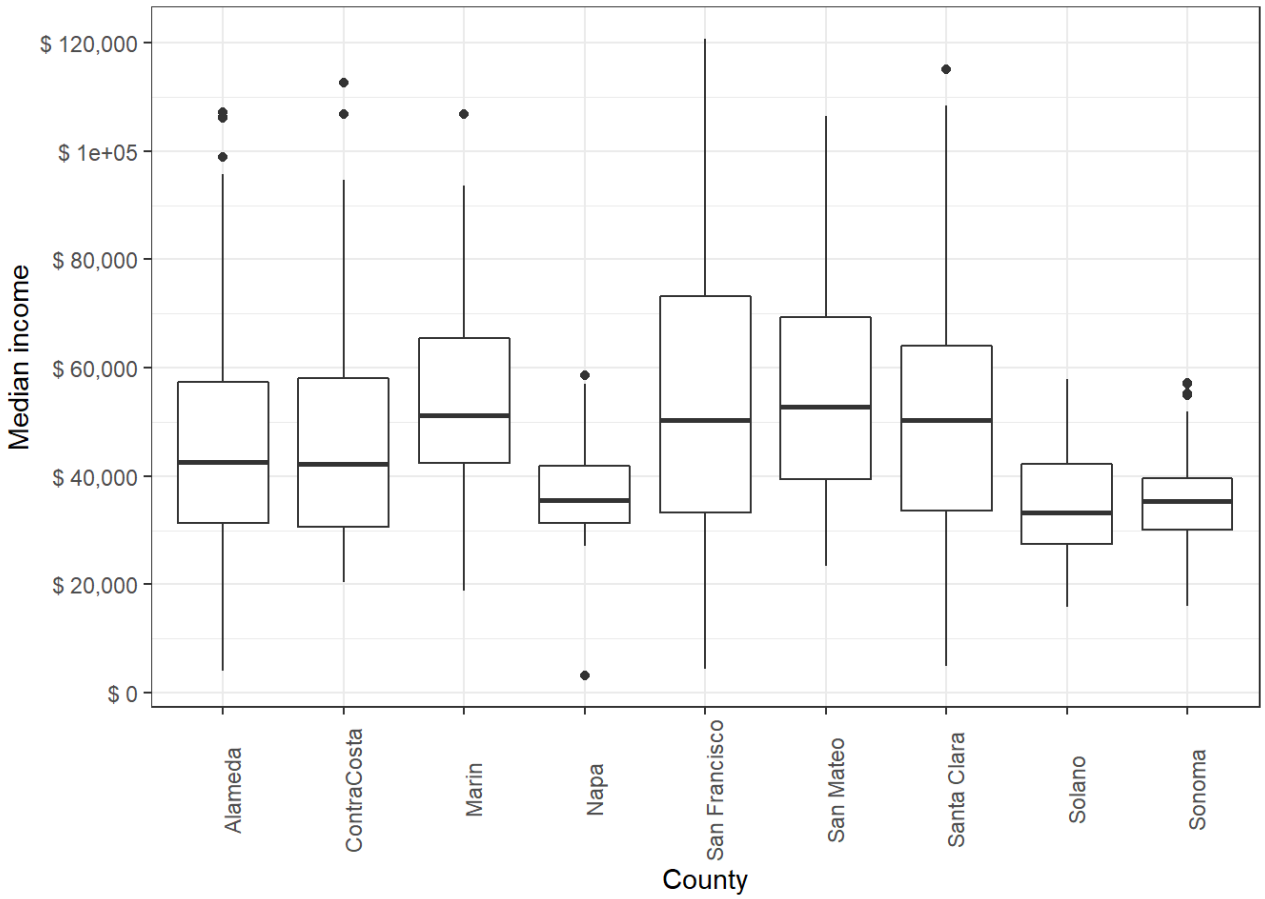
```
differences <- TukeyHSD(anova2)

as_tibble(cbind(pair = row.names(differences$county),
                differences$county))
```

```
## # A tibble: 36 x 5
##    pair              diff           lwr           upr           `p adj`
##    <chr>             <chr>          <chr>         <chr>         <chr>
##  1 ContraCosta-Alam~ 198.562741377~ -5176.847960~ 5573.9734430~ 0.9999999990779~
##  2 Marin-Alameda     9438.18270917~ 448.03650398~ 18428.328914~ 0.0311285276283~
##  3 Napa-Alameda      -8604.2413649~ -18871.27679~ 1662.7940698~ 0.1859737478288~
##  4 San Francisco-Al~ 8509.51376330~ 3030.1998088~ 13988.827717~ 5.4088421382880~
##  5 San Mateo-Alameda 9762.83940432~ 3856.3432698~ 15669.335538~ 1.1297370438034~
##  6 Santa Clara-Alam~ 6061.60863509~ 1504.6422259~ 10618.575044~ 0.0012561886796~
##  7 Solano-Alameda    -9922.9764712~ -17059.28771~ -2786.665230~ 0.0005632551333~
##  8 Sonoma-Alameda    -9598.6792436~ -16590.71676~ -2606.641718~ 0.0007154285724~
##  9 Marin-ContraCosta 9239.61996779~ -172.2091623~ 18651.449097~ 0.0590319003214~
## 10 Napa-ContraCosta  -8802.8041062~ -19441.02836~ 1835.4201535~ 0.2000663903619~
## # ... with 26 more rows
```

```
ggplot(tract_data, aes(x = county, y = med_incomeE)) +
  geom_boxplot() +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_x_discrete(name = "County") +
  scale_y_continuous(name = "Median income",
                     breaks = seq(0, 120000, by = 20000),
                     labels = paste("$",
                                    prettyNum(seq(0, 120000, by = 20000),
                                              big.mark = ",")))
```

```
## Warning: Removed 1 rows containing non-finite values (stat_boxplot).
```



There is a statistically significant association at the 95 percent confidence level between county & median income. The interquartile range of median incomes by tract in each county beg some interesting follow-up questions and analysis (e.g. San Francisco's range is quite large compared to Napa or Sonoma's - is this due possibly to a larger variety of industries and any trends that might suggest there is a greater variety of socioeconomic statuses outside of industry type?)

12. County & average household size

```
anova3 <- aov (avg_hh_sizeE ~ county, data = tract_data)
summary(anova3)
```

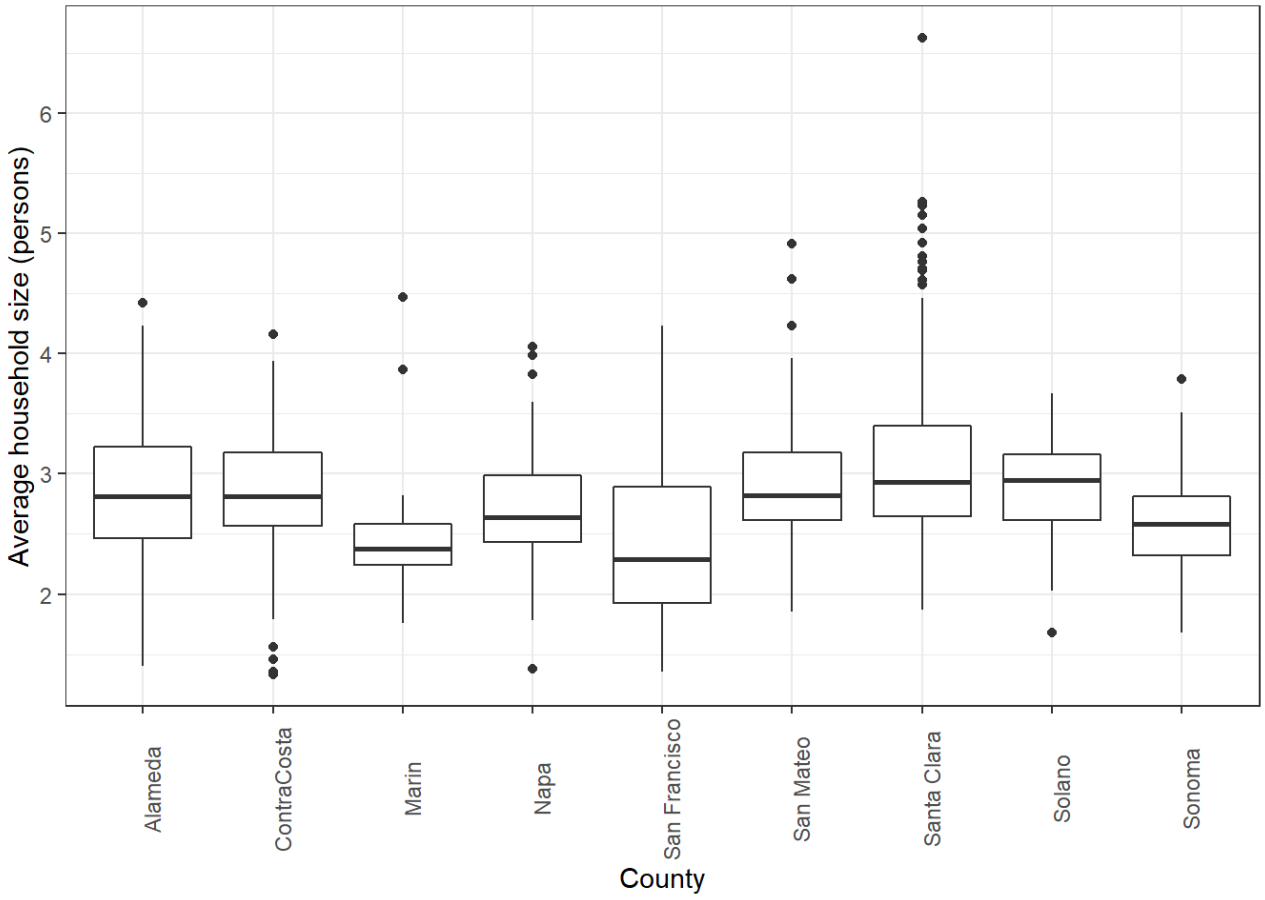
```
##               Df Sum Sq Mean Sq F value Pr(>F)
## county         8   67.1   8.394   25.93 <2e-16 ***
## Residuals   1568  507.5   0.324
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
differences <- TukeyHSD(anova3)

as_tibble(cbind(pair = row.names(differences$county),
                differences$county))
```

```
## # A tibble: 36 x 5
##    pair             diff           lwr            upr            `p adj`
##    <chr>            <chr>          <chr>          <chr>          <chr>
##  1 ContraCosta-Ala~ 0.04212439613~ -0.1120089561~ 0.19625774841~ 0.995296224345~
##  2 Marin-Alameda    -0.4183425925~ -0.6762082805~ -0.1604769046~ 1.842787777051~
##  3 Napa-Alameda     -0.0895555555~ -0.3840589245~ 0.20494781347~ 0.990326256706~
##  4 San Francisco-A~ -0.3723824786~ -0.5294981310~ -0.2152668261~ 1.007716132761~
##  5 San Mateo-Alame~ 0.08439957264~ -0.0849767965~ 0.25377594185~ 0.832591783740~
##  6 Santa Clara-Ala~ 0.24745788530~ 0.11681843692~ 0.37809733368~ 1.756614674830~
##  7 Solano-Alameda   0.06167316784~ -0.1429968686~ 0.26634320433~ 0.9908942857907
##  8 Sonoma-Alameda   -0.2404974747~ -0.4410273173~ -0.0399676321~ 0.006275915028~
##  9 Marin-ContraCos~ -0.4604669887~ -0.7304769078~ -0.1904570695~ 4.779616062156~
## 10 Napa-ContraCosta -0.1316799516~ -0.4368731327~ 0.17351322939~ 0.919124567781~
## # ... with 26 more rows
```

```
ggplot(tract_data, aes(x = county, y = avg_hh_sizeE)) +
  geom_boxplot() +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_x_discrete(name = "County") +
  scale_y_continuous(name = "Average household size (persons)",
                     breaks = seq(0, 10, by = 1))
```



There is a statistically significant association at the 95 percent confidence level between county & average household size. For many counties, these differences between average household size seems to be quite small practically

All of the p-values were much less than 0.05 for each of the three tests above with county as the categorical variable

13. Majority race/ethnicity & percent of households with no internet access

```
anova4 <- aov (pct_no_internet ~ maj_re, data = tract_data)
summary(anova4)
```

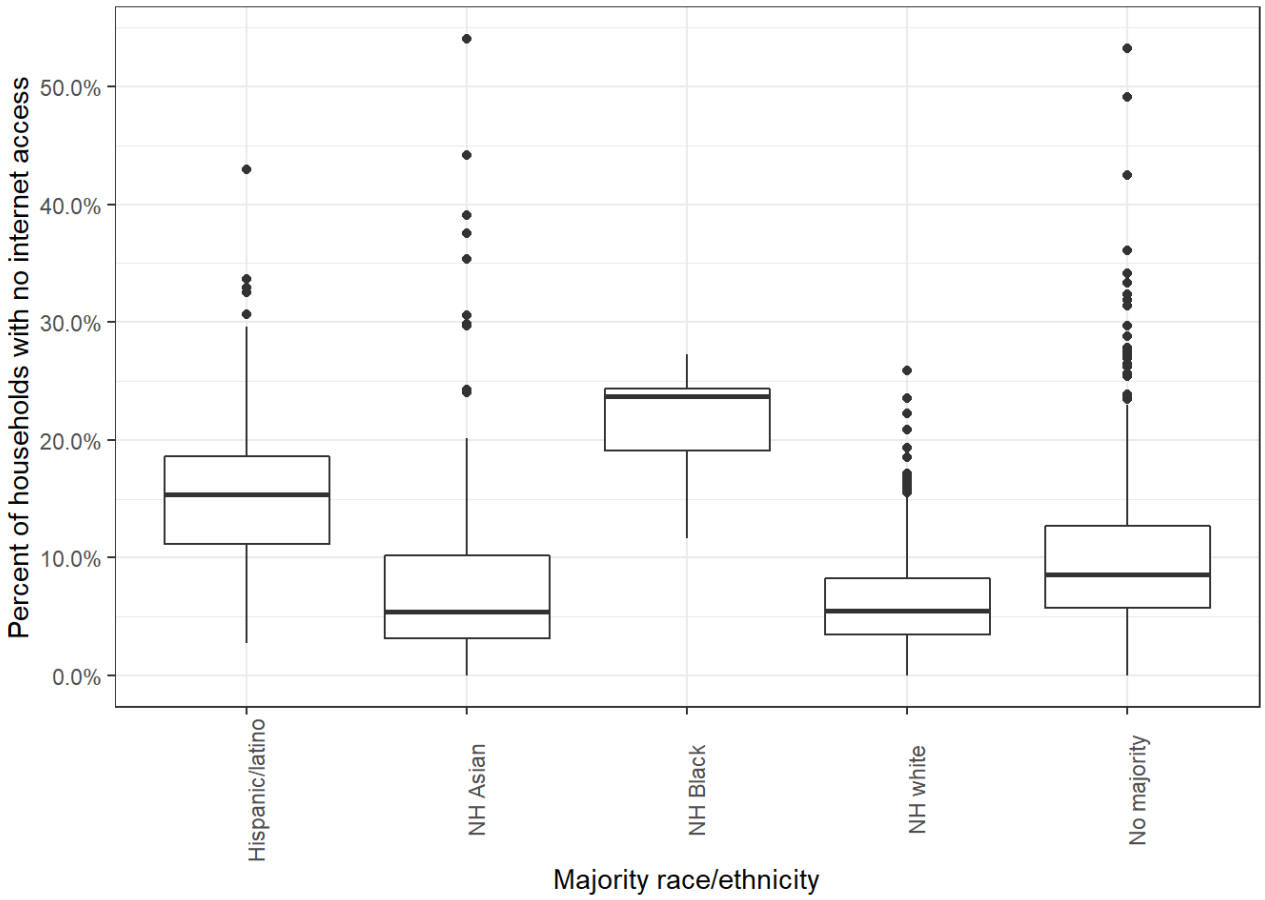
```
##               Df Sum Sq Mean Sq F value Pr(>F)
## maj_re         4  1.321  0.3303   91.87 <2e-16 ***
## Residuals   1572  5.652  0.0036
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
differences <- TukeyHSD(anova4)

as_tibble(cbind(pair = row.names(differences$maj_re),
                differences$maj_re))
```

```
## # A tibble: 10 x 5
##    pair             diff           lwr            upr            `p adj`
##    <chr>            <chr>          <chr>          <chr>          <chr>
##  1 nh_asian-hisp_l~ -0.0817239465~ -0.0997493012~ -0.0636985919~ 0
##  2 nh_black-hisp_l~ 0.05358633446~ -0.0098229329~ 0.11699560187~ 0.142819677325~
##  3 nh_white-hisp_l~ -0.0966975254~ -0.1119847442~ -0.0814103066~ 0
##  4 no_maj_re-hisp_~ -0.0601557088~ -0.0754475644~ -0.0448638532~ 0
##  5 nh_black-nh_asi~ 0.13531028104~ 0.07233958318~ 0.19828097890~ 5.352519827361~
##  6 nh_white-nh_asi~ -0.0149735788~ -0.0283254571~ -0.0016217006~ 0.018928778899~
##  7 no_maj_re-nh_as~ 0.02156823771~ 0.00821105080~ 0.03492542463~ 0.000107824418~
##  8 nh_white-nh_bla~ -0.1502838599~ -0.2125260838~ -0.0880416360~ 5.818125980994~
##  9 no_maj_re-nh_bl~ -0.1137420433~ -0.1759854062~ -0.0514986804~ 6.630106332838~
## 10 no_maj_re-nh_wh~ 0.03654181660~ 0.02720371771~ 0.04587991548~ 0
```

```
ggplot(tract_data, aes(x = maj_re, y = pct_no_internet)) +
  geom_boxplot() +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_x_discrete(name = "Majority race/ethnicity",
                   labels = c("Hispanic/latino",
                              "NH Asian",
                              "NH Black",
                              "NH white",
                              "No majority")) +
  scale_y_continuous(name = "Percent of households with no internet access",
                     breaks = seq(0, 1, by = .1),
                     labels = scales::percent)
```



There is a statistically significant association at the 95 percent confidence level between majority race/ethnicity & percent of households with no internet access. This relationship is quite evident from the box plot visualization and leads to some interesting follow-up questions on societal implications (e.g. the majority NH Black tracts appear to have quite a bit greater percent of households without access to internet)

14. Majority race/ethnicity & median income

```
anova5 <- aov (med_incomeE ~ maj_re, data = tract_data)
summary(anova5)
```

```
##               Df    Sum Sq   Mean Sq F value Pr(>F)
## maj_re         4 1.652e+11 4.129e+10     127 <2e-16 ***
## Residuals   1571 5.109e+11 3.252e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 1 observation deleted due to missingness
```

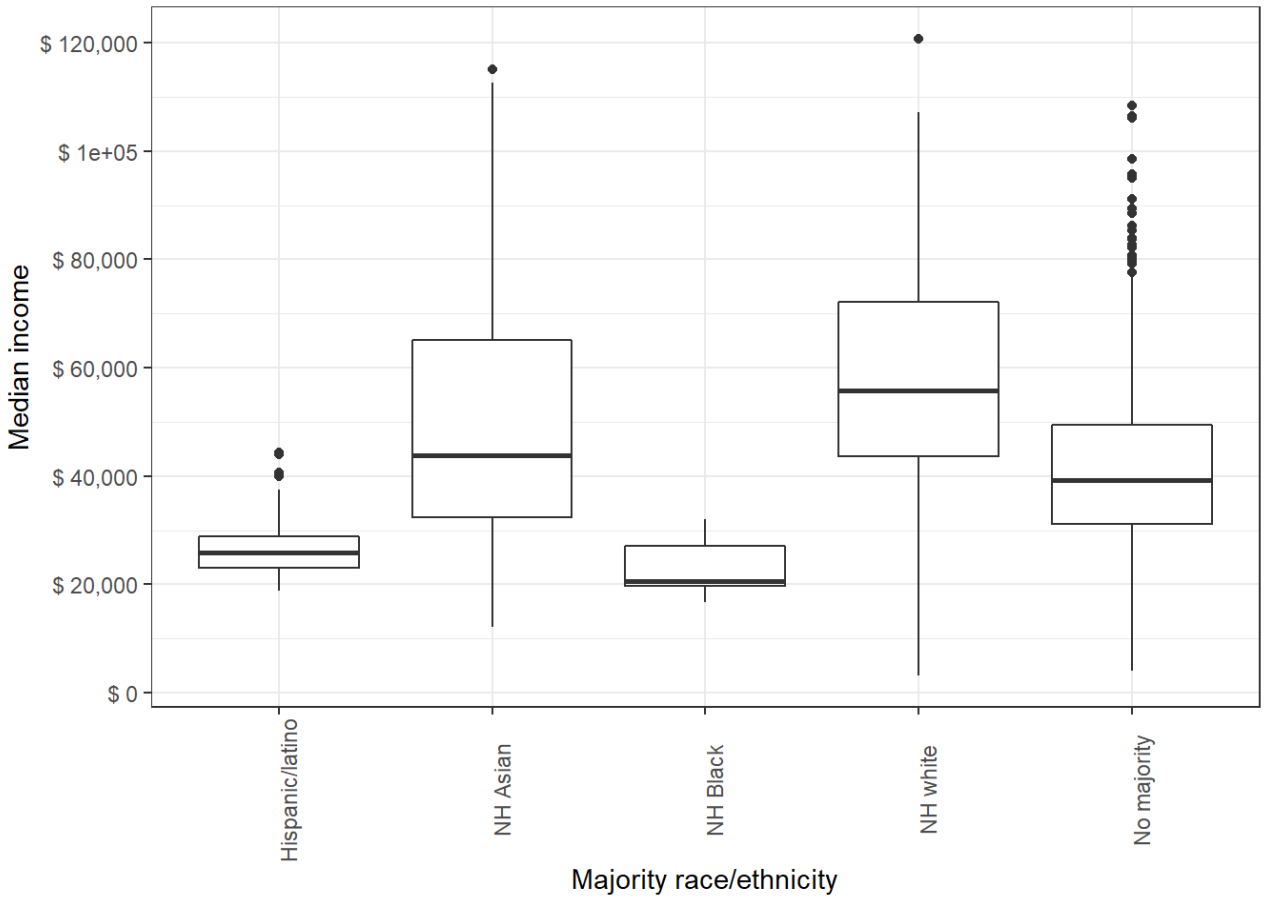
```
differences <- TukeyHSD(anova5)

as_tibble(cbind(pair = row.names(differences$maj_re),
                differences$maj_re))
```

```
## # A tibble: 10 x 5
##    pair              diff           lwr           upr           `p adj`
##    <chr>             <chr>          <chr>         <chr>         <chr>
##  1 nh_asian-hisp_la~ 24733.0417691~ 19311.972330~ 30154.111207~ 0
##  2 nh_black-hisp_la~ -3075.3049645~ -22145.44523~ 15994.835307~ 0.9922016451282~
##  3 nh_white-hisp_la~ 32163.9291818~ 27565.649427~ 36762.208936~ 0
##  4 no_maj_re-hisp_l~ 15280.7064361~ 10681.728350~ 19879.684521~ 0
##  5 nh_black-nh_asian -27808.346733~ -46746.58857~ -8870.104893~ 0.0006072495764~
##  6 nh_white-nh_asian 7430.88741267~ 3414.5544225~ 11447.220402~ 4.8242775654294~
##  7 no_maj_re-nh_asi~ -9452.3353330~ -13469.46782~ -5435.202845~ 1.7307064670291~
##  8 nh_white-nh_black 35239.2341463~ 16519.907586~ 53958.560705~ 3.0531259482824~
##  9 no_maj_re-nh_bla~ 18356.0114006~ -363.4867116~ 37075.509512~ 0.0577101309940~
## 10 no_maj_re-nh_whi~ -16883.222745~ -19692.76652~ -14073.67896~ 0
```

```
ggplot(tract_data, aes(x = maj_re, y = med_incomeE)) +
  geom_boxplot() +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_x_discrete(name = "Majority race/ethnicity",
                   labels = c("Hispanic/latino",
                              "NH Asian",
                              "NH Black",
                              "NH white",
                              "No majority")) +
  scale_y_continuous(name = "Median income",
                     breaks = seq(0, 120000, by = 20000),
                     labels = paste("$",
                                    prettyNum(seq(0, 120000, by = 20000),
                                              big.mark = ",")))
```

```
## Warning: Removed 1 rows containing non-finite values (stat_boxplot).
```



There is a statistically significant association at the 95 percent confidence level between majority race/ethnicity & median income. This relationship could be quite meaningful in the real world as these income differences appear by majority race/ethnicity of this set of 9 clustered counties

15. Majority race/ethnicity & average household size

```
anova6 <- aov (avg_hh_sizeE ~ maj_re, data = tract_data)
summary(anova6)
```

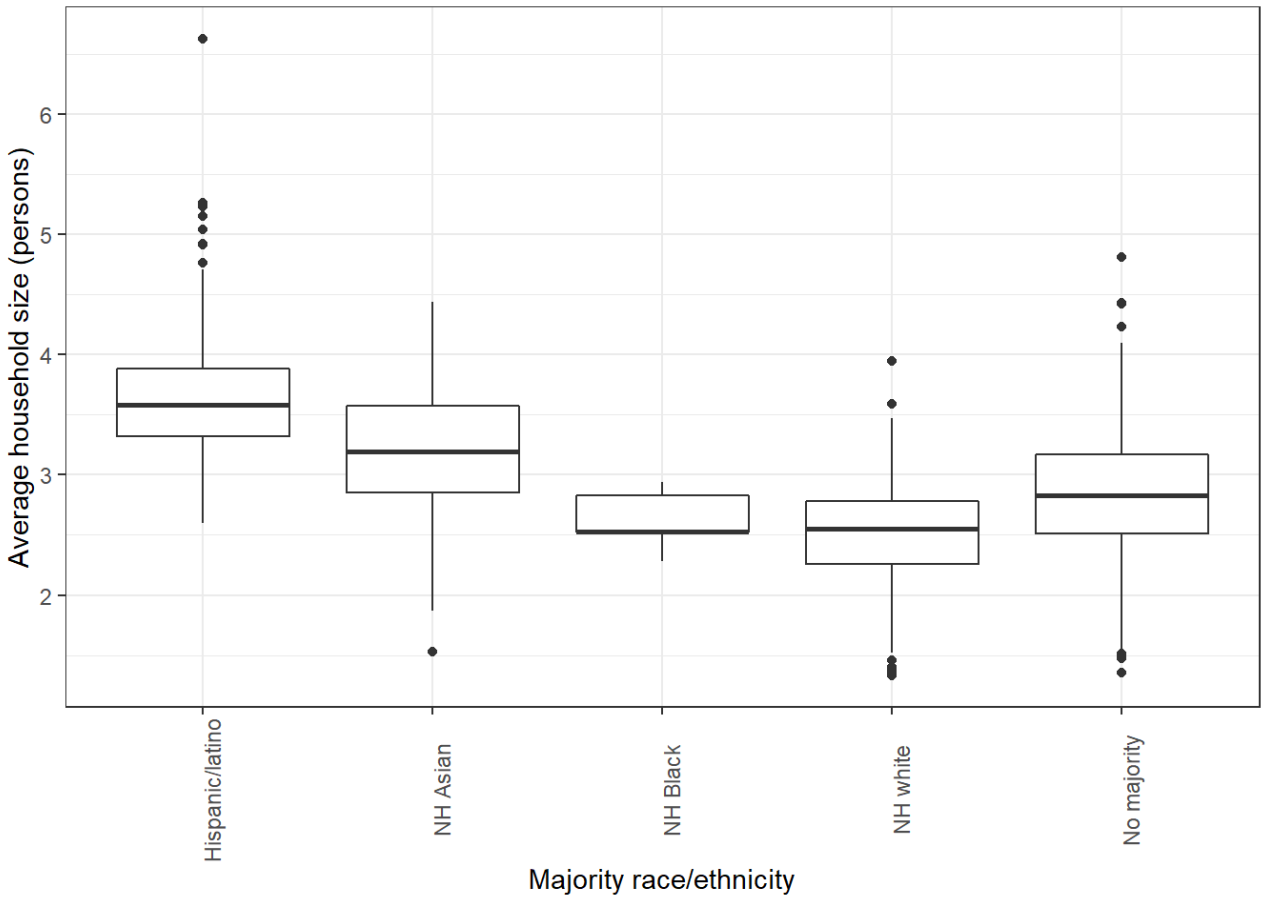
```
##               Df Sum Sq Mean Sq F value Pr(>F)
## maj_re         4  198.1   49.52   206.8 <2e-16 ***
## Residuals   1572  376.5    0.24
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
differences <- TukeyHSD(anova6)

as_tibble(cbind(pair = row.names(differences$maj_re),
                differences$maj_re))
```

```
## # A tibble: 10 x 5
##    pair             diff           lwr            upr            `p adj`
##    <chr>            <chr>          <chr>          <chr>          <chr>
##  1 nh_asian-hisp_l~ -0.4935863715~ -0.6407097411~ -0.3464630020~ 0
##  2 nh_black-hisp_l~ -1.0601621073~ -1.5777100426~ -0.5426141721~ 2.602493104353~
##  3 nh_white-hisp_l~ -1.1895452242~ -1.3143198644~ -1.0647705840~ 0
##  4 no_maj_re-hisp_~ -0.8634124563~ -0.9882249422~ -0.7385999705~ 0
##  5 nh_black-nh_asi~ -0.5665757358~ -1.0805440559~ -0.0526074156~ 0.022238247106~
##  6 nh_white-nh_asi~ -0.6959588527~ -0.8049371979~ -0.5869805075~ 0
##  7 no_maj_re-nh_as~ -0.3698260848~ -0.4788477592~ -0.2608044103~ 0
##  8 nh_white-nh_bla~ -0.1293831168~ -0.6374056155~ 0.37863938173~ 0.957494893400~
##  9 no_maj_re-nh_bl~ 0.19674965100~ -0.3112821441~ 0.70478144614~ 0.828123607177~
## 10 no_maj_re-nh_wh~ 0.32613276788~ 0.24991498470~ 0.40235055105~ 0
```

```
ggplot(tract_data, aes(x = maj_re, y = avg_hh_sizeE)) +
  geom_boxplot() +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_x_discrete(name = "Majority race/ethnicity",
                   labels = c("Hispanic/latino",
                              "NH Asian",
                              "NH Black",
                              "NH white",
                              "No majority")) +
  scale_y_continuous(name = "Average household size (persons)",
                     breaks = seq(0, 10, by = 1))
```



There is a statistically significant association at the 95 percent confidence level between majority race/ethnicity & average household size. This is also a fairly clear relationship to discern in the box plot visualization

All of the p-values were much less than 0.05 for each of the three tests above with majority race/ethnicity as the categorical variable