HW11

Emily Logan, Hyeyeon Kim, Veronica Imbert

5/3/2021

Data Splitting

```
# Importing Shapefiles
Census_Geo <- readOGR("Census")
```

```
## OGR data source with driver: ESRI Shapefile
## Source: "/Users/hyeyeonkim/Desktop/UR/Spring 2021/STAT 276W/Group Project/Census", layer: "Census_ACS
## with 12 features
## It has 149 fields
```

```
Real_Estate_Geo <- readOGR("Real_Estate")
```

```
## OGR data source with driver: ESRI Shapefile
## Source: "/Users/hyeyeonkim/Desktop/UR/Spring 2021/STAT 276W/Group Project/Real_Estate", layer: "Recen
## with 5017 features
## It has 50 fields
```

```
#Importing normal .csv files
Real_Estate_csv <- read.csv("Real_Estate.csv")

Census <- read.csv("Census.csv")
```

```
#selecting price, square feet, (and object ID for use in graphics)
re_data <- Real_Estate_csv %>% select(price = v_recentsales2_GIS_SALEPRICE,
            sqft = v_recentsales2_GIS_SQFT, id = OBJECTID,
            bedroom = v_recentsales2_BEDROOMS, year = v_recentsales2_YEARBUILT,
            stories = v_recentsales2_STORIES, bathroom = v_recentsales2_BATHROOMS)


re_data <- re_data %>% mutate(stories = ceiling(stories))


re_data$bedroom <- ifelse(re_data$bedroom >= 4, yes = "More than 3", no = re_data$bedroom)

#Making variable for total number of bathrooms since the original variable was formatted differently
#Original variable was (number of full bath) . (number of half bath)
re_data <- re_data %>% mutate(half_bath = round(10*(bathroom %% 1), digits = 0), full_bath = floor(bathr
```

```
re_data <- re_data %>% mutate(bathroom = half_bath + full_bath)

re_data$full_bath_c <- ifelse(re_data$full_bath >= 4, yes = "More than 3", no = re_data$full_bath)
re_data$half_bath_c <- ifelse(re_data$half_bath >= 4, yes = "More than 3", no = re_data$half_bath)
re_data$bathroom_c <- ifelse(re_data$bathroom >= 4, yes = "More than 3", no = re_data$bathroom)


#Scrubbing price and sqft to not have commas or $ so that they're numeric
re_data <- re_data %>%
  mutate(price = as.numeric(gsub(",", "",
      as.character(sub("$", "", price , fixed=TRUE)))))
re_data <- re_data %>% mutate(sqft = as.numeric(gsub(",", "", sqft)),
                              id = as.integer(id),
         price = as.numeric(price), sqft = as.numeric(sqft))

#Adding in price/sqft variable
re_data <- re_data %>% mutate(ppsqft = price/sqft)


#Creating `own` variable based on ownership of property
help <- Real_Estate_csv %>% select(bill = TaxParcel_PSTLCITY, zip = TaxParcel_ZIP5) %>%
  mutate(ny = "Rochester, NY") %>% mutate(site = paste(ny, zip))

ownership <- ifelse(help$bill == help$site, yes = TRUE, no = FALSE)


re_data <- re_data %>% mutate(own = ownership)

re_data <- re_data %>% mutate(pthou = price/1000)


set.seed(32)
samp <- sample(1:nrow(re_data),floor(.6*nrow(re_data)))

r_train_data_csv <- re_data[samp,] ## This is the 60% chunk
remain40 <- re_data[-samp,]  ## This is used for further bifurcation

samp2 <- sample(1:nrow(remain40),floor(.5*nrow(remain40)))

r_test_data_csv <- remain40[samp2,] ## First chunk of 20%
r_conf_data_csv <- remain40[-samp2,] ## Second Chunk of 20%

Reduce("intersect",list(r_train_data_csv,r_test_data_csv,r_conf_data_csv))


## data frame with 0 columns and 0 rows

Census <- Census%>% select(poverty_level = F99x_poverty_level,
          median_earnings = Median_earnings_in__past_12_mon, id = OBJECTID)

#Linking the shapefile to the normal .csv file
r_spdf_fortified <- tidy(Real_Estate_Geo, region = "OBJECTID")
```

```
r_spdf_fortified <- r_spdf_fortified %>% mutate(id = as.integer(id))

r_spdf_fortified1 <-  r_spdf_fortified %>%
  left_join(. , r_test_data_csv, by=c("id"="id"))

r_test_data <- na.omit(r_spdf_fortified1)

r_spdf_fortified2 <-  r_spdf_fortified %>%
  left_join(. , r_train_data_csv, by=c("id"="id"))

r_train_data <- na.omit(r_spdf_fortified2)

r_spdf_fortified3 <-  r_spdf_fortified %>%
  left_join(. , r_conf_data_csv, by=c("id"="id"))
r_conf_data <- na.omit(r_spdf_fortified3)

rm(help)
rm(r_spdf_fortified)
rm(re_data)
rm(Real_Estate_csv)
rm(Real_Estate_Geo)
rm(remain40)
```

```
#Linking the shapefile to normal .csv file
c_spdf_fortified <- tidy(Census_Geo, region = "OBJECTID")

Census <- Census %>% mutate(OBJECTID = as.integer(id))
c_spdf_fortified <- c_spdf_fortified %>% mutate(id = as.integer(id))

c_data <-  c_spdf_fortified %>%
  left_join(. , Census, by=c("id"="OBJECTID"))
rm(c_spdf_fortified)
rm(Census)

rm(Census_Geo)
```

```
#save(r_train_data,file = "C:/Users/tcbra/OneDrive/Documents/emily/r_train_data.csv")
#save(r_test_data, file = "C:/Users/tcbra/OneDrive/Documents/emily/r_test_data.csv")
#save(r_conf_data, file = "C:/Users/tcbra/OneDrive/Documents/emily/r_conf_data.csv")
#save(c_data, file =  "C:/Users/tcbra/OneDrive/Documents/emily/c_data.csv")

#rm(r_train_data_csv)
#rm(r_test_data_csv)
#rm(r_conf_data_csv)
#rm(r_spdf_fortified1)
#rm(r_spdf_fortified2)
#rm(r_spdf_fortified3)
```

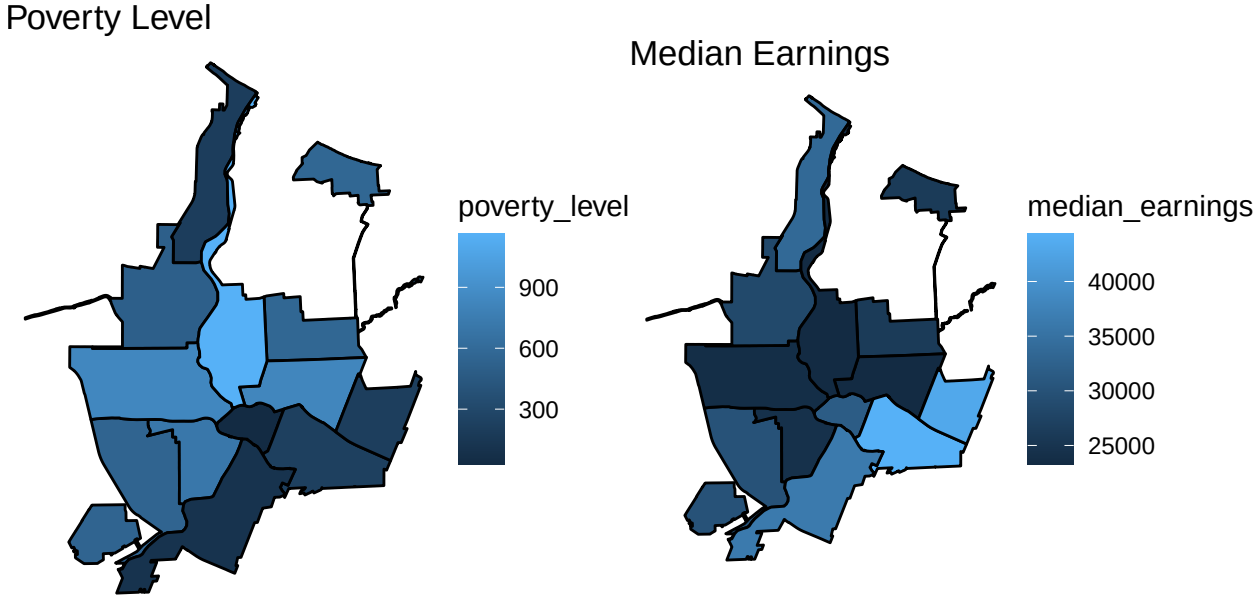
```
c_data <- read.csv("c_data.csv")
r_train_data <- read.csv("r_train_data.csv")
```

EDA

Univariate Graphics

Census

```
c1 <- ggplot() +
  geom_polygon(data = c_data, aes(x = long, y = lat, group = group, fill = poverty_level), col = "black
  theme_void() +
  coord_map() +
  labs(title = "Poverty Level")

c2 <- ggplot() +
  geom_polygon(data = c_data, aes(x = long, y = lat, group = group, fill = median_earnings), col = "bla
  theme_void() +
  coord_map() +
  labs(title = "Median Earnings")

c <- grid.arrange(c1, c2, ncol = 2)
```



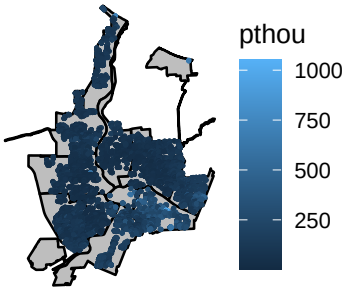
Real Estate

```
r1 <- ggplot() +
  geom_polygon(data = c_data, aes(x = long, y = lat, group = group),
               col = "black", fill = "grey") +
  geom_point(data = r_train_data, aes(col = pthou, x = long, y = lat, group = group), size = .25)+
  theme_void() +
  coord_map() +
  labs(title = "Price in Thousands")
```

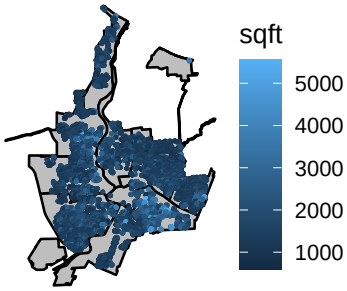
```
r2 <- ggplot() +
  geom_polygon(data = c_data, aes(x = long, y = lat, group = group),
               col = "black", fill = "grey") +
  geom_point(data = r_train_data, aes(col = sqft, x = long, y = lat, group = group), size = .25)+
  theme_void() +
  coord_map() +
  labs(title = "Square Footage")

r3 <- ggplot() +
  geom_polygon(data = c_data, aes(x = long, y = lat, group = group),
               col = "black", fill = "grey") +
  geom_point(data = r_train_data, aes(col = ppsqft, x = long, y = lat, group = group), size = .25)+
  theme_void() +
  coord_map() +
  labs(title = "Price per Square Footage")

r4 <- ggplot() +
  geom_polygon(data = c_data, aes(x = long, y = lat, group = group),
               col = "black", fill = "grey") +
  geom_point(data = r_train_data, aes(col = own, x = long, y = lat, group = group), size = .25)+
  theme_void() +
  coord_map() +
  labs(title = "Ownership")


r5 <- ggplot() +
  geom_polygon(data = c_data, aes(x = long, y = lat, group = group),
               col = "black", fill = "grey") +
  geom_point(data = r_train_data, aes(col = stories, x = long, y = lat, group = group), size = .25)+
  theme_void() +
  coord_map() +
  labs(title = "Number of Stories")


r6 <- ggplot() +
  geom_polygon(data = c_data, aes(x = long, y = lat, group = group),
               col = "black", fill = "grey") +
  geom_point(data = r_train_data, aes(col = bathroom_c, x = long, y = lat, group = group), size = .25)+
  theme_void() +
  coord_map() +
  labs(title = "Total Number of Bathrooms")


r7 <- ggplot() +
  geom_polygon(data = c_data, aes(x = long, y = lat, group = group),
               col = "black", fill = "grey") +
  geom_point(data = r_train_data, aes(col = half_bath_c, x = long, y = lat, group = group), size = .25)+
  theme_void() +
  coord_map() +
  labs(title = "Number of Half Bathrooms")

r8 <- ggplot() +
  geom_polygon(data = c_data, aes(x = long, y = lat, group = group),
```

```
               col = "black", fill = "grey") +
  geom_point(data = r_train_data, aes(col = full_bath_c, x = long, y = lat, group = group), size = .25)+
  theme_void() +
  coord_map() +
  labs(title = "Number of Full Bathrooms")

r9 <- ggplot() +
  geom_polygon(data = c_data, aes(x = long, y = lat, group = group),
               col = "black", fill = "grey") +
  geom_point(data = r_train_data, aes(col = bedroom, x = long, y = lat, group = group), size = .25)+
  theme_void() +
  coord_map() +
  labs(title = "Number of Bedrooms")

r10 <- ggplot() +
  geom_polygon(data = c_data, aes(x = long, y = lat, group = group),
               col = "black", fill = "grey") +
  geom_point(data = r_train_data, aes(col = year, x = long, y = lat, group = group), size = .25)+
  theme_void() +
  coord_map() +
  labs(title = "Year the House was Built")

r <- grid.arrange(r1, r2, r3, r4, r5, r6, r7, r8, r9, r10, ncol = 2)
```

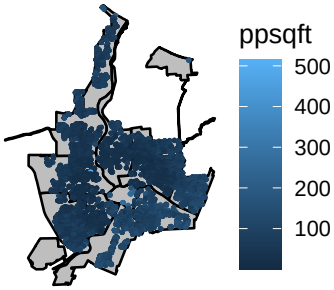
Price in Thousands



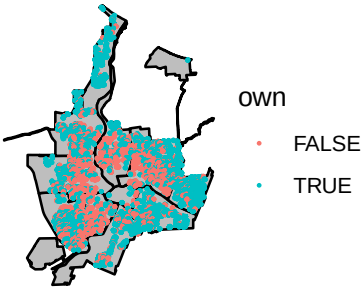
Square Footage



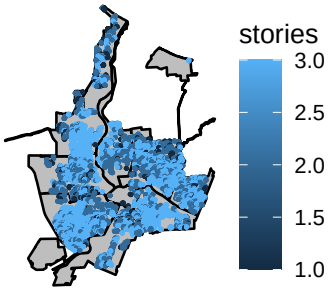
Price per Square Footage



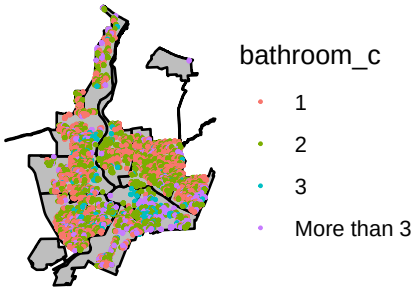
Ownership



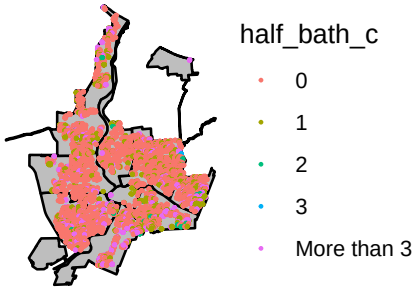
Number of Stories



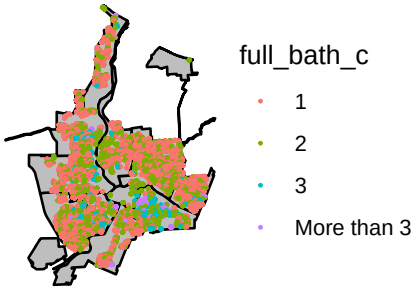
Total Number of Bathrooms



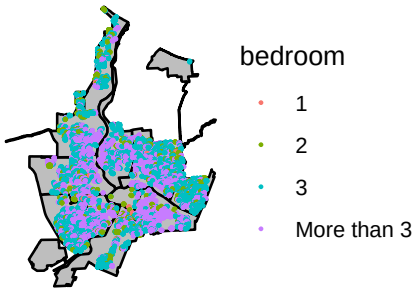
Number of Half Bathrooms



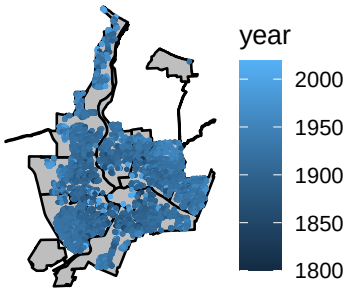
Number of Full Bathrooms



Number of Bedrooms



Year the House was Built



```
re1 <- ggplot(r_train_data, aes(pthou)) +
  geom_histogram() +
  labs(title = "Price in Thousands")

re2 <- ggplot(r_train_data, aes(sqft)) +
  geom_histogram()+
  labs(title = "Square Footage")

re3 <- ggplot(r_train_data, aes(ppsqft)) +
  geom_histogram()+
  labs(title = "Price per Square Footage")

t4 <- knitr::kable(table(r_train_data$stories), caption = "Number of Stories")

re4 <- ggplot(r_train_data, aes(stories)) +
  geom_bar() +
  labs(title = "Number of Stories")

t5 <- knitr::kable(table(r_train_data$own), caption = "Ownership")

re5 <- ggplot(r_train_data, aes(own)) +
  geom_bar() +
  labs(title = "Count of Ownership")

t6 <- knitr::kable(table(r_train_data$bathroom_c), caption = "Total Number of Bathrooms")

re6 <- ggplot(r_train_data, aes(bathroom)) +
  geom_bar()+
  labs(title = "Total Number of Bathrooms")

knitr::kable(table(r_train_data$half_bath_c), caption = "Number of Half Bathrooms")
```

Table 1: Number of Half Bathrooms

Var1	Freq
0	14055
1	2983
2	165
3	5
More than 3	2479

```
re7 <- ggplot(r_train_data, aes(half_bath)) +
  geom_bar()+
  labs(title = "Number of Half Bathrooms")

knitr::kable(table(r_train_data$full_bath_c), caption = "Number of Full Bathrooms")
```

Table 2: Number of Full Bathrooms

Var1	Freq
1	11500
2	6763
3	973
More than 3	451

```
re8 <- ggplot(r_train_data, aes(full_bath)) +
  geom_bar()+
  labs(title = "Total Number of Full Bathrooms")

knitr::kable(table(r_train_data$bedroom), caption = "Number Bedrooms")
```

Table 3: Number Bedrooms

Var1	Freq
1	183
2	2197
3	9988
More than 3	7319

```
re10 <- ggplot(r_train_data, aes(year)) +
  geom_histogram()+
  labs(title = "Year Built")


t4
```

Table 4: Number of Stories

Var1	Freq
1	1542
2	7950
3	10195

```
t5
```

Table 5: Ownership

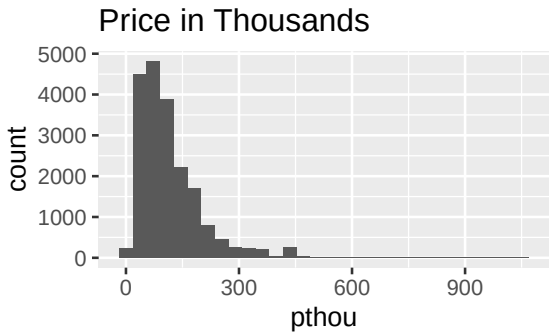
Var1	Freq
FALSE	6870
TRUE	12817

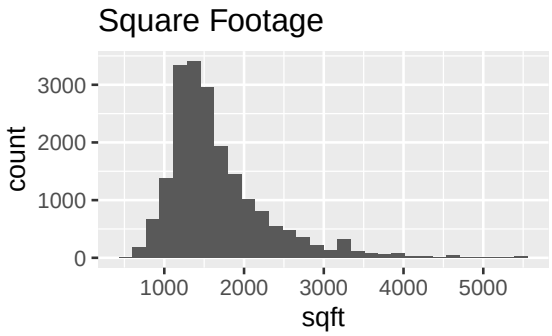
```
t6
```

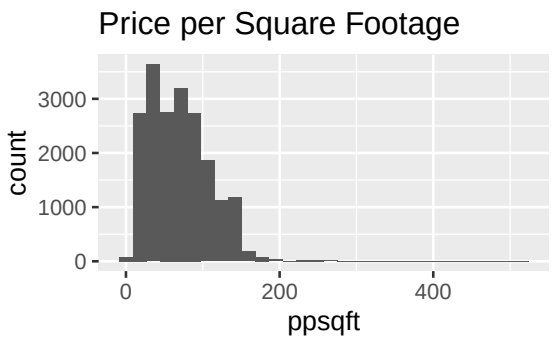
Table 6: Total Number of Bathrooms

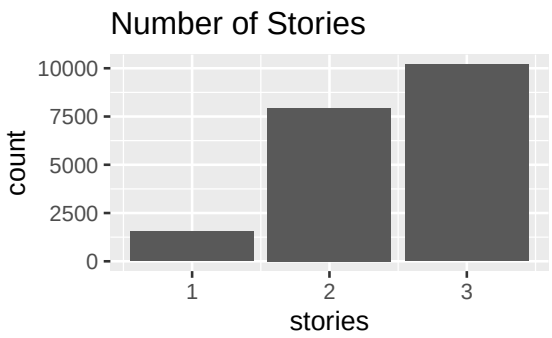
Var1	Freq
1	7330
2	8275
3	1154
More than 3	2928

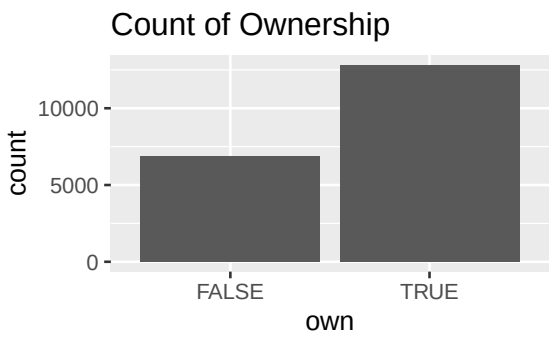
```
grid.arrange(re1, re2, re3, re4, re5, re6, re7, re8, re10, ncol = 2)
```

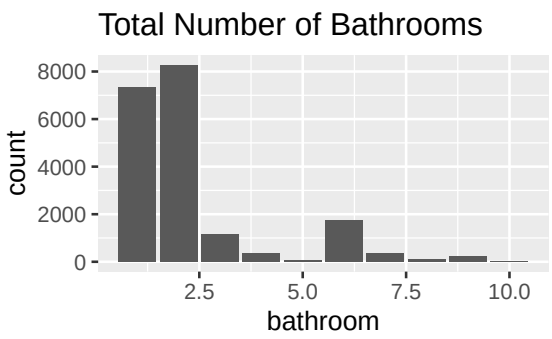

Price in Thousands
count
pthou
0
300
600
900
0
1000
2000
3000
4000
5000

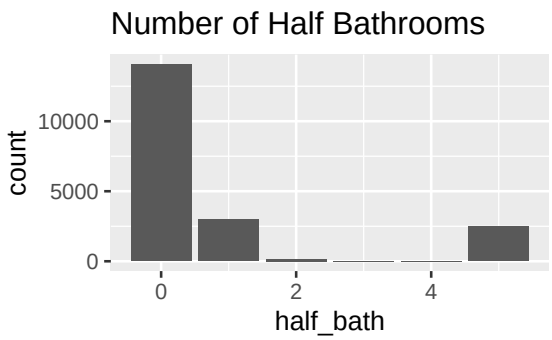

Square Footage
count
sqft
1000
2000
3000
4000
5000
0
1000
2000
3000

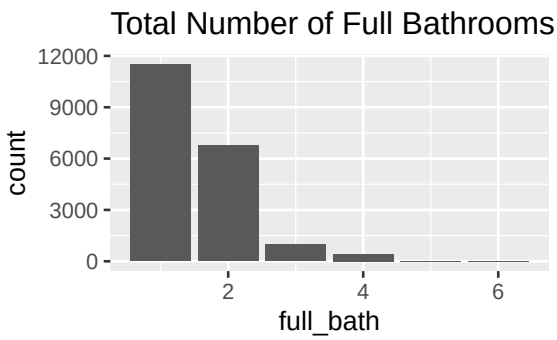

Price per Square Footage
count
ppsqft
0
200
400
0
1000
2000
3000

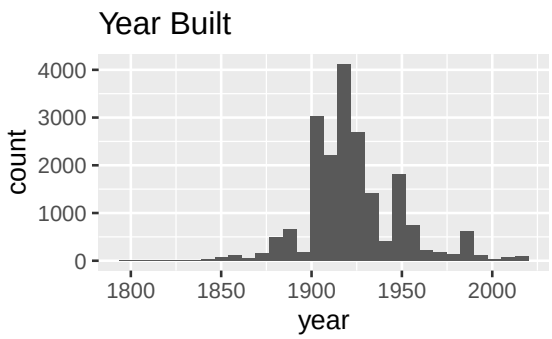

Number of Stories
count
stories
1
2
3
0
2500
5000
7500
10000


Count of Ownership
count
own
FALSE
TRUE
0
5000
10000


Total Number of Bathrooms
count
bathroom
2.5
5.0
7.5
10.0
0
2000
4000
6000
8000

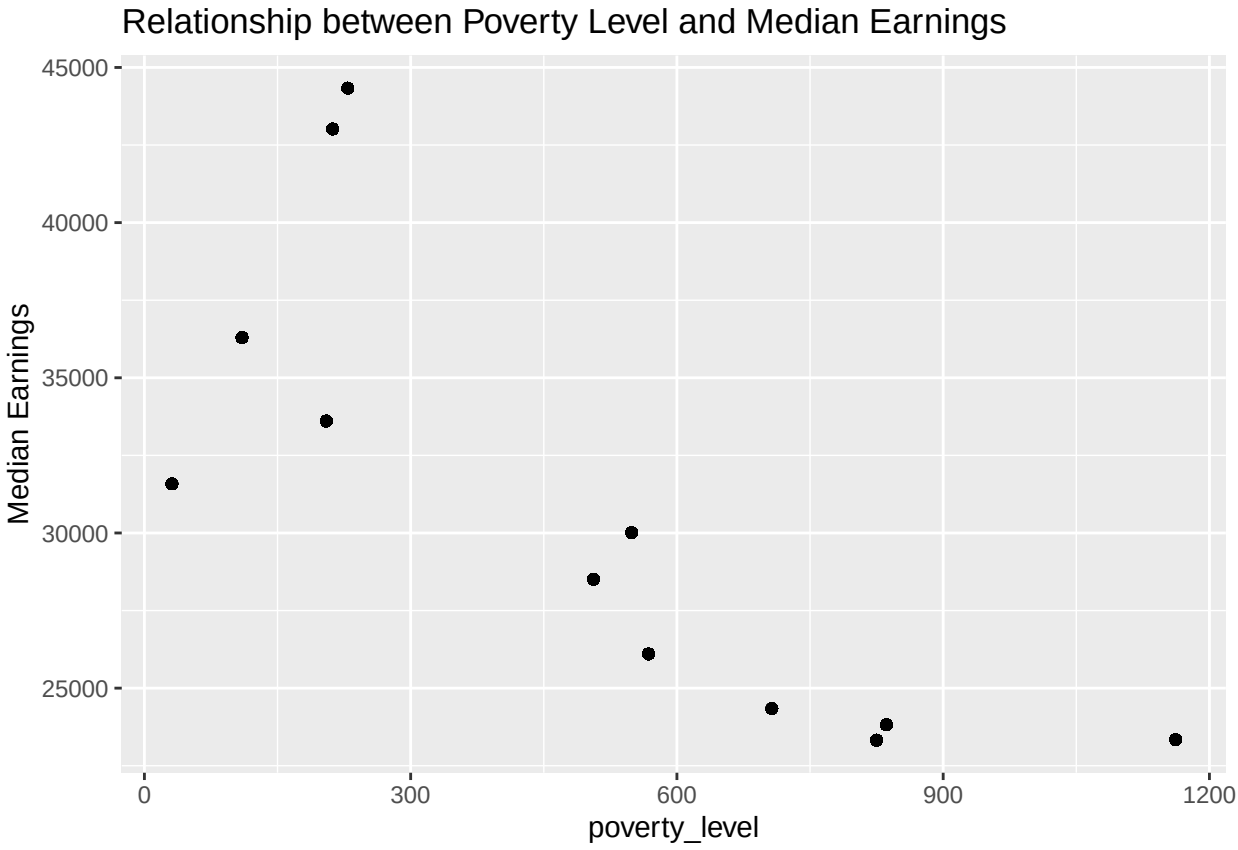

Number of Half Bathrooms
count
half_bath
0
2
4
0
5000
10000


Total Number of Full Bathrooms
count
full_bath
2
4
6
0
3000
6000
9000
12000


Year Built
count
year
1800
1850
1900
1950
2000
0
1000
2000
3000
4000

Hypothesis Generation

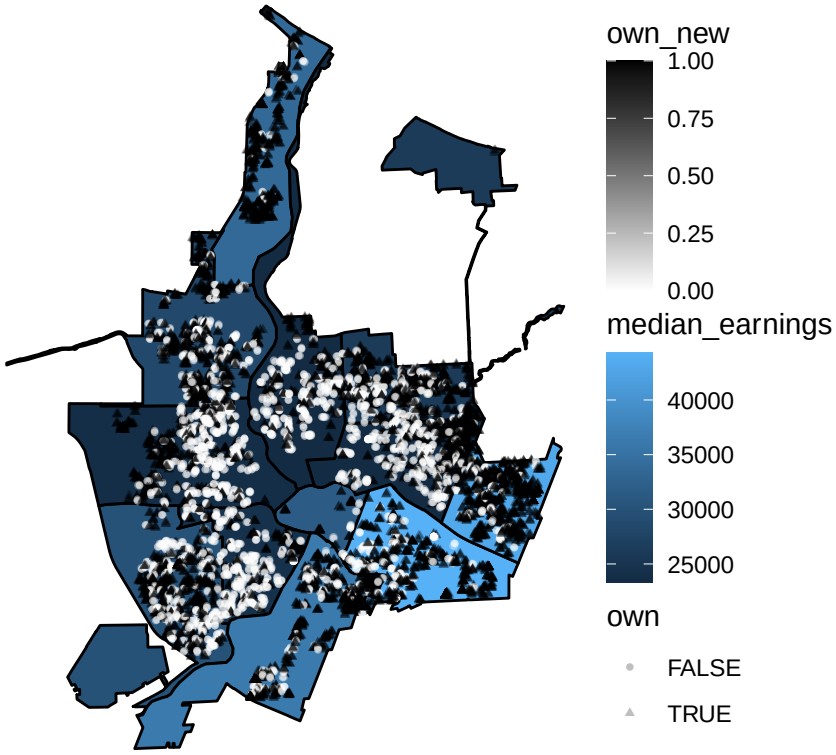
```
r_train_data$own_new <- as.integer(r_train_data$own)

# Relationship between the median earnings and poverty level
ggplot(data = c_data) +
  geom_point(mapping = aes(x = poverty_level, y = median_earnings)) +
  ylab("Median Earnings") +
  labs(title = "Relationship between Poverty Level and Median Earnings")
```



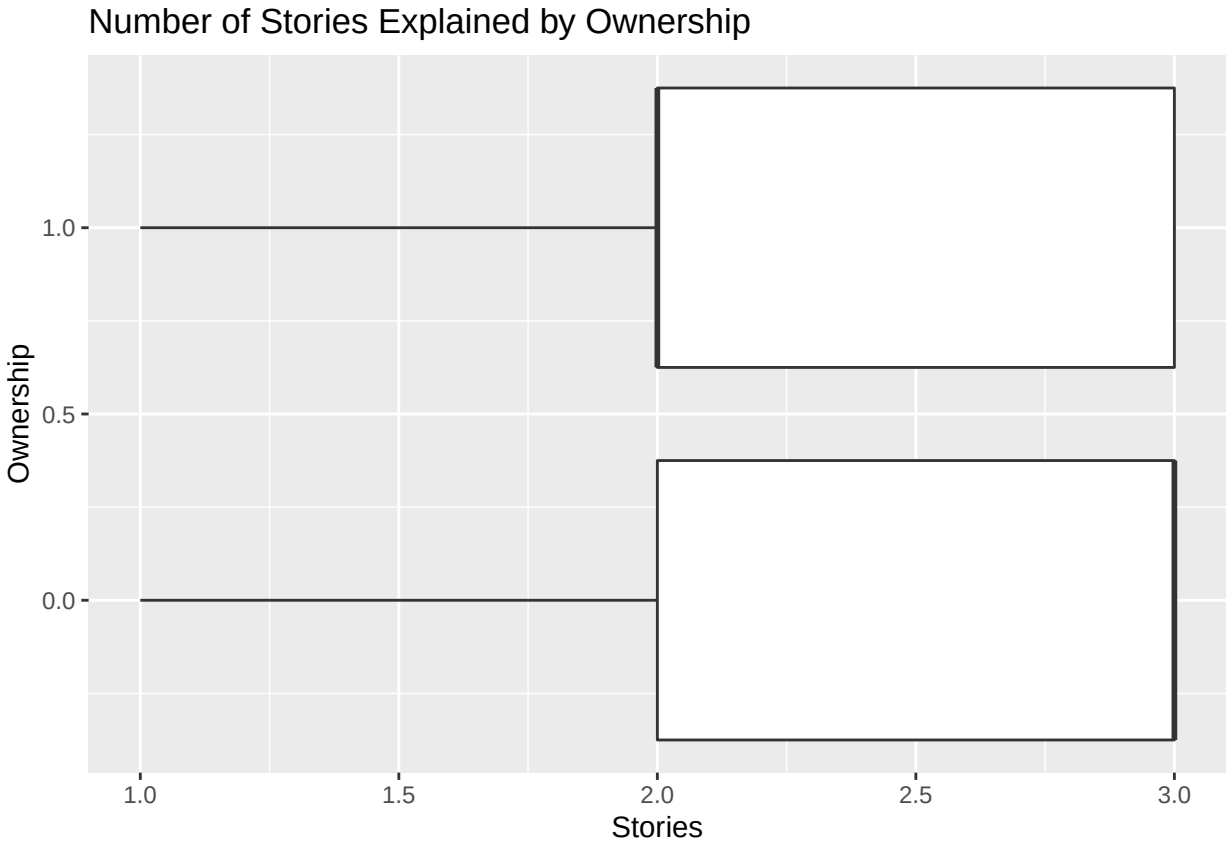
Hannah

```
# Compare the ownership of the house by median earnings
ggplot() +
  geom_polygon(data = c_data, aes(fill = median_earnings, x = long, y = lat, group = group),
               col = "black") +
  geom_point(data = r_train_data, aes(col = own_new, x = long, y = lat, group = group, shape = own), si
  scale_color_gradient(low="white", high = "black") +
  theme_void() +
  coord_map() +
  labs(title = "Ownership of House Explained by Median Earnings")
```

Ownership of House Explained by Median Earnings




```
# Compare the ownership of the house and stories of the house
ggplot(r_train_data) +
  geom_boxplot(aes(x = stories, y = own_new, group = own_new)) +
  ylab("Ownership") +
  xlab("Stories") +
  labs(title = "Number of Stories Explained by Ownership")
```



Introduction

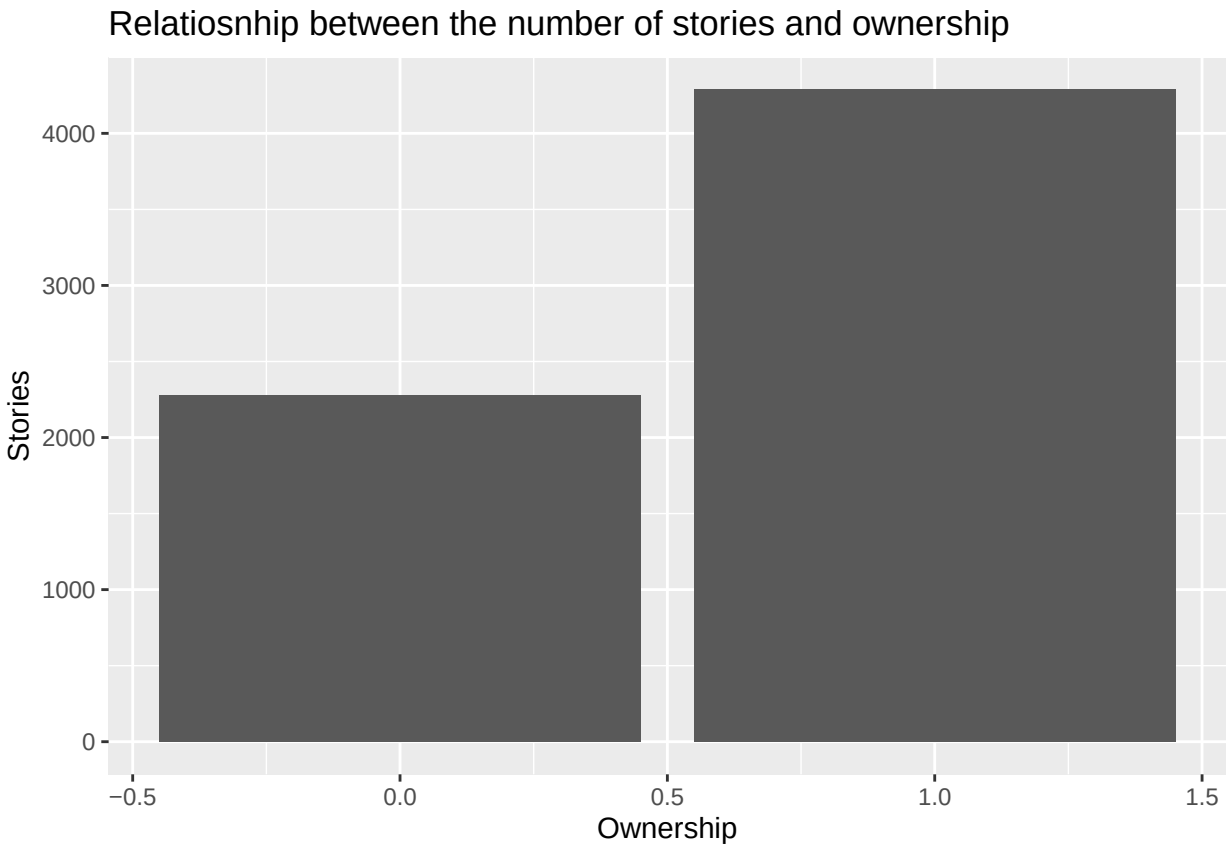
The testing data set has 1003 observations of 21 variables

Testing Data

```
r_test_data <- read.csv("r_test_data.csv")
```

Section 4: Own vs. Stories I was inspired by this graph here:

```
r_test_data$own_new <- as.integer(r_test_data$own)

ggplot(r_test_data) +
  geom_bar(aes(x=own_new, group = stories)) +
  ylab("Stories") +
  xlab("Ownership") +
  labs(title = "Relatiosnhip between the number of stories and ownership")
```



Ownership = 0 means it is not owned by tenants and ownership = 1 means the house is owned.

I did linear regression between two variables, the number of stories and ownership of house.

```
lm_4 <- lm(data=r_train_data, stories ~ own_new)
summary(lm_4)
```

```
##
## Call:
## lm(formula = stories ~ own_new, data = r_train_data)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5387 -0.3864  0.4613  0.6136  0.6136
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.538719   0.007609  333.64   <2e-16 ***
## own_new     -0.152357   0.009430  -16.16   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6307 on 19685 degrees of freedom
## Multiple R-squared:  0.01309,    Adjusted R-squared:  0.01304
## F-statistic:   261 on 1 and 19685 DF,  p-value: < 2.2e-16
```

Observations

For the own vs. stories, we can conclude that there is no significant relationship between the number of stories and the ownership of the house even if the graph shows higher stories' houses are more owned by people.