Final Project: Hypothesis Generation

Mawada Elmahgoub

Last compiled on: 07 May, 2020

Contents

Part 1: Exploring correlations between the continuous variables

After the EDA, I wanted to take a further look at the correlation coefficient between the continuous variables in the Data Set.

```
corlist<-c()
for(x in colnames(vacancydatax)){
  for(y in colnames(vacancydatax)){
    if (x==y) next
    if (is.numeric(vacancydatax[[x]]) && is.numeric(vacancydatax[[y]])){
      corlist[[paste(x,'&',y)]]<- cor(vacancydatax[x],vacancydatax[y])
    }
  }
}
#corlist
sigdata<- stack(corlist)
df<- sigdata[(sigdata$values>0.6 & sigdata$values<0.99),]
df1<- df[order(df$values), c(1,2)][c(1,3,5,7,9,11,13),]
df1$'Correlation Coefficient'<- df1$values
df1$values<- NULL

df1$'Variables'<- df1$ind
df1$ind<-NULL

rownames(df1)<- NULL

kable(df1, caption="Correlation Coefficients between Continuous Variables", align = "c")
```

Table 1: Correlation Coefficients between Continuous Variables

Correlation Coefficient	Variables
0.6074642	CURRENT_LAND_VALUE & Shape___Length
0.6407240	STATEDAREA & CURRENT_LAND_VALUE
0.6453008	CURRENT_LAND_VALUE & Shape___Area
0.6453056	CURRENT_LAND_VALUE & SHAPEACRES
0.7938359	SHAPEACRES & Shape___Length
0.7938373	Shape___Area & Shape___Length
0.8084508	STATEDAREA & Shape___Length

Key Findings:

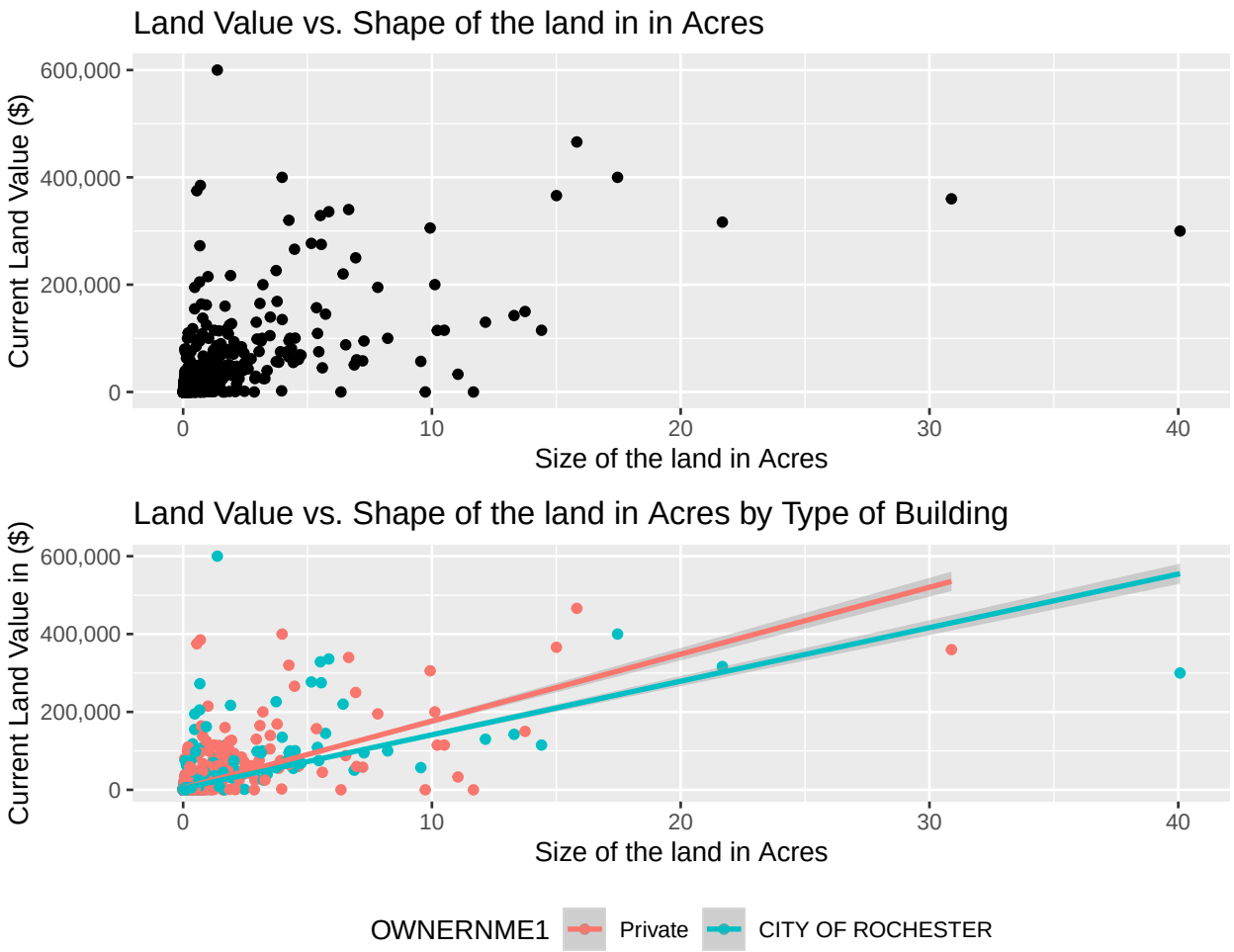
- Many of the higher correlations are from dimension attributes which is not exactly helpful
- `CURRENT_LAND_VALUE` increases with land dimension attributes
- Highest correlation is between `CURRENT_LAND_VALUE` and `SHAPEACRES`

Exploring Land Value vs. Shape of the land in Acres by Type of Building

```
p1<- ggplot(vacancydatax, aes(x=SHAPEACRES, y=CURRENT_LAND_VALUE)) +
  geom_point() +
  scale_y_continuous(labels=comma) +
  labs( title = "Land Value vs. Shape of the land in in Acres" ) +
  ylab("Current Land Value ($)") +
  xlab("Size of the land in Acres")

p2<- ggplot(vacancydatax, aes(x=SHAPEACRES, y=CURRENT_LAND_VALUE, color=OWNERNME1)) +
  geom_point(aes(fill=OWNERNME1)) +
  geom_smooth(method="lm") +
  scale_y_continuous(labels=comma) +
  labs( title = "Land Value vs. Shape of the land in Acres by Type of Building" ) +
  ylab("Current Land Value in ($)") +
  xlab("Size of the land in Acres") +
  theme(legend.position = "bottom")

grid.arrange(p1,p2)
```



Testing for difference

```
t.test(vacancydatax$CURRENT_LAND_VALUE~vacancydatax$OWNERNME1)
```

```
##
##  Welch Two Sample t-test
##
## data:  vacancydatax$CURRENT_LAND_VALUE by vacancydatax$OWNERNME1
## t = 3.1679, df = 4262.2, p-value = 0.001546
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1002.106 4256.464
## sample estimates:
##           mean in group Private mean in group CITY OF ROCHESTER
##                        9736.216                        7106.931
```

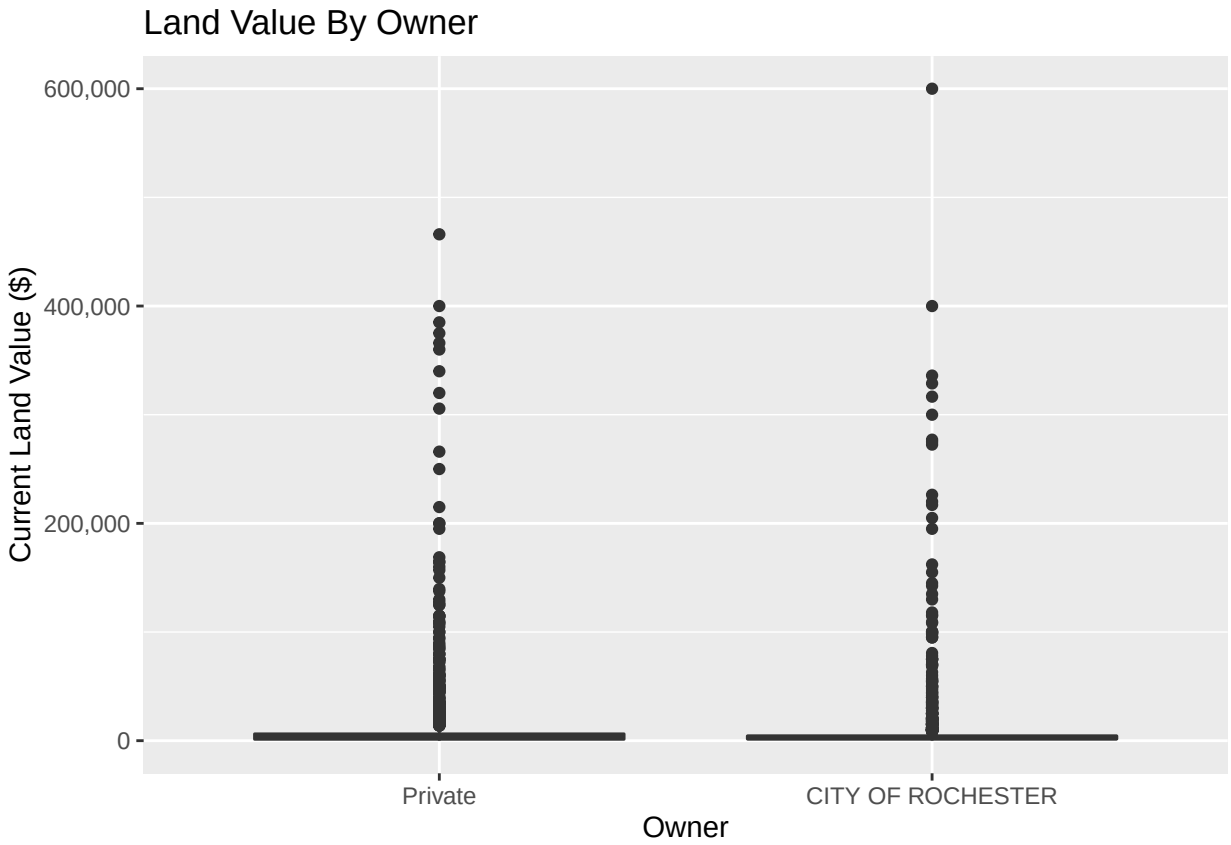
```
t.test(vacancydatax$SHAPEACRES~vacancydatax$OWNERNME1)
```

```
##
##  Welch Two Sample t-test
##
## data:  vacancydatax$SHAPEACRES by vacancydatax$OWNERNME1
## t = 1.2237, df = 4639.8, p-value = 0.2211
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.02535882  0.10959746
## sample estimates:
##           mean in group Private mean in group CITY OF ROCHESTER
##                       0.2899285                       0.2478091
```

At the alpha = 0.05 level, there is infact a difference in the means for Land Value by Owner type. However, there is no difference in the means for Shape in Acres by Owner Type. Therefore it may be interesting to further look at the relationship between land value and owner type alone.

Exploring Land Value by Type of Building

```
ggplot(vacancydatax, aes(y=CURRENT_LAND_VALUE, x=OWNERNME1 )) +
  geom_boxplot(aes(fill=OWNERNME1),  show.legend = FALSE) +
  scale_y_continuous(labels = comma) +
  labs( title = "Land Value By Owner") +
  ylab("Current Land Value ($)") +
  xlab("Owner")
```

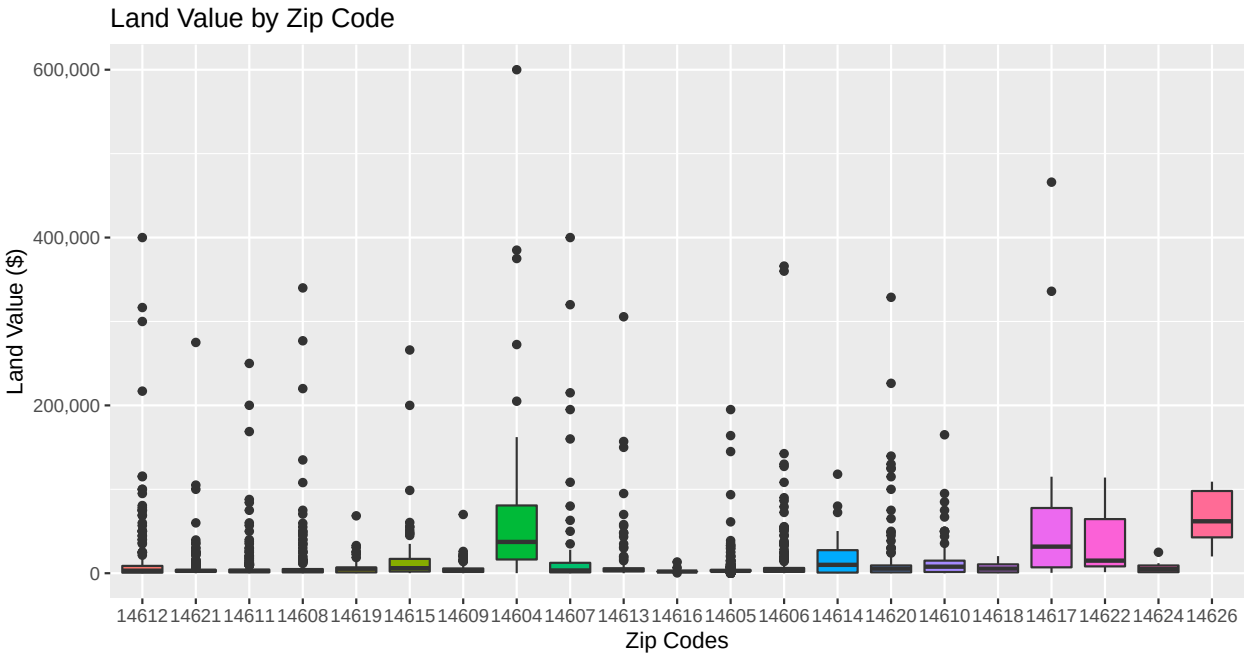


The boxplot shows that for both owner types, most of the land value is below $50,000. Generally what this is telling us is that most of the data owned both privately and by the City of Rochester has very low value, perhaps if improvements were made to this vacant land, it's value would substantially increase.

Exploring Land Value by Zip Codes

Something else I wanted to explore was Land Value by Zip Codes. This could probably better help understand location and value.

```
ggplot(vacancydatax, aes(y=CURRENT_LAND_VALUE, x=ZIP5)) +
  geom_boxplot(aes(fill=ZIP5), show.legend = FALSE) +
  scale_y_continuous(labels=comma) +
  labs( title = "Land Value by Zip Code" ) +
  ylab("Land Value ($)") +
  xlab("Zip Codes")
```

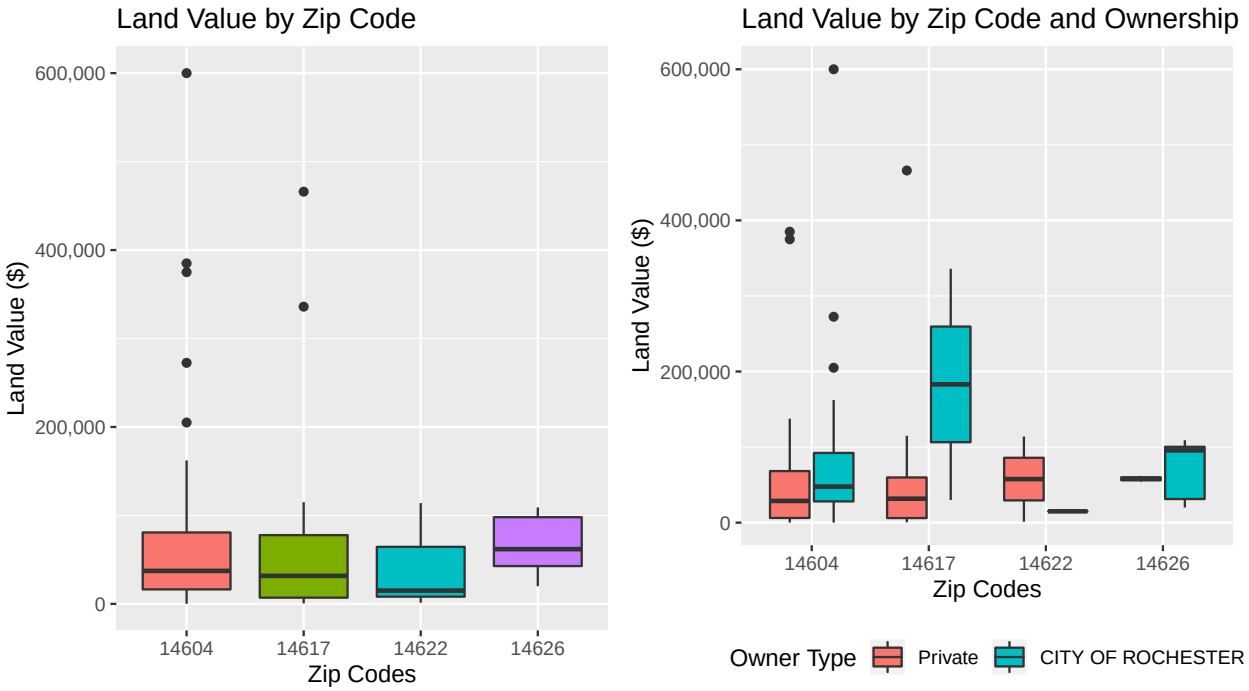


Based on the chart, it can be see that the zipcodes 14604 (Rochester), 14617 (West Irondequoit), 14622 (East Irondequoit) and 14626 (Greece) have a very interesting range. Below, I filter the data set to only look at data from these zip codes to get a better view and also plot the relationship by Ownership.

```
zipsal<- filter(vacancydatax, ZIP5=="14604" | vacancydatax$ZIP5=="14617"|
                  vacancydatax$ZIP5=="14622" | vacancydatax$ZIP5=="14626")
pi<-  ggplot(zipsal, aes(y=CURRENT_LAND_VALUE, x=ZIP5)) +
  geom_boxplot(aes(fill=ZIP5), show.legend = FALSE) +
  scale_y_continuous(labels=comma) +
  labs( title = "Land Value by Zip Code" ) +
  ylab("Land Value ($)") +
  xlab("Zip Codes")

pp<- ggplot(zipsal, aes(y=CURRENT_LAND_VALUE, x=ZIP5)) +
  geom_boxplot(aes(fill=OWNERNME1)) +
  scale_y_continuous(labels=comma) +
  labs( title = "Land Value by Zip Code and Ownership", fill="Owner Type" ) +
  ylab("Land Value ($)") +
  xlab("Zip Codes") +
  theme(legend.position = "bottom")
```

```
grid.arrange(pi,pp,nrow=1)
```



From the chart, the most prominent finding is that in the 14617 (West Irondequoit) zip code, the mean land value that is owned by the city is the highest. In the 14622 (East Irondequoit) zip code, the Land Value is more distributed for the Privately owned buildings and in the 14626 (Greece) zip code, the Land Value is more distributed for the City Owned buildings. This suggests that there may even be a relationship between zip codes and ownership which is explored below.

```
chisq.test(vacancydatax$ZIP5, vacancydatax$OWNERNME1)
```

```
## Warning in chisq.test(vacancydatax$ZIP5, vacancydatax$OWNERNME1): Chi-
## squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  vacancydatax$ZIP5 and vacancydatax$OWNERNME1
## X-squared = 533.1, df = 20, p-value < 2.2e-16
```

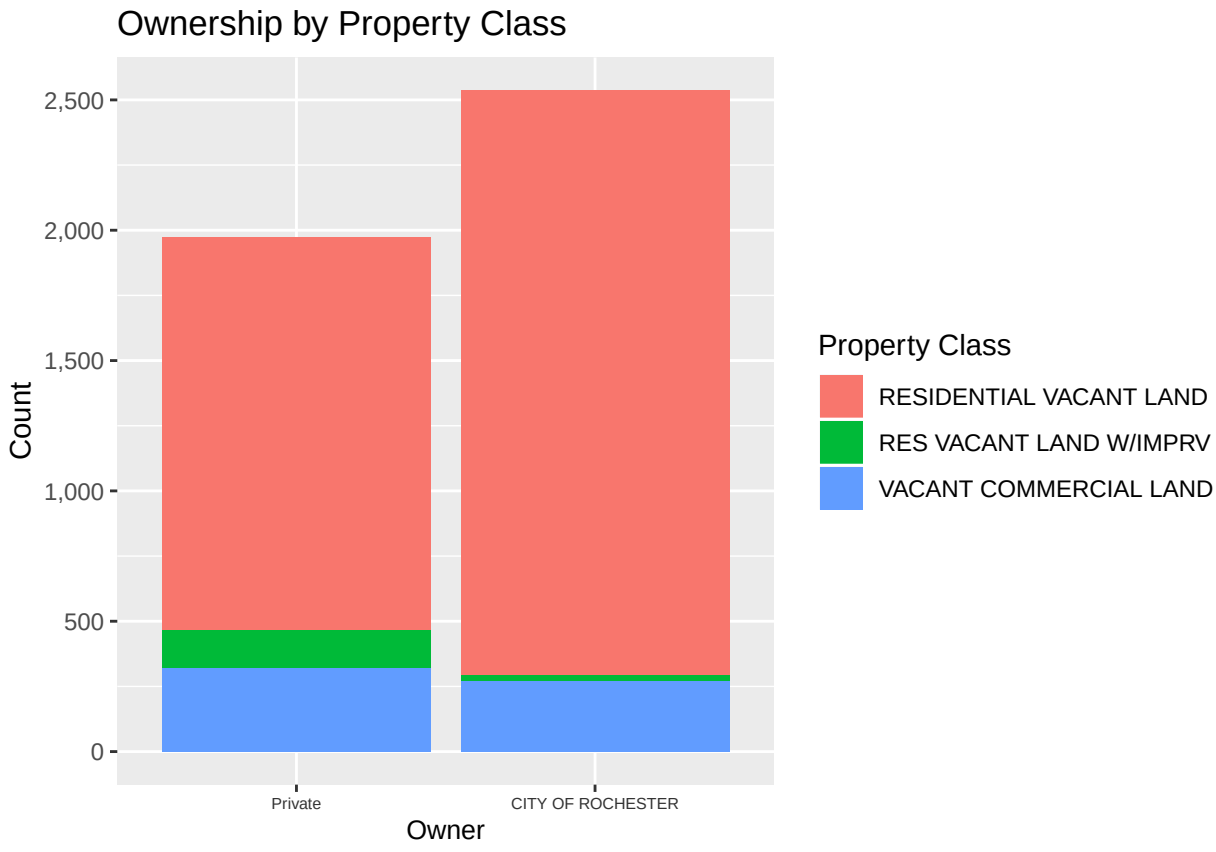
This result from the Chi Square test is statistically significant, since the p-value is significant at the alpha = 0.01 level, it can be concluded that there is an association between Zip Code and Ownership. This confirms what can be seen in the graphs above. More information is needed on the specific areas in Rochester to understand this relationship especially as it relates to the land value.

Part 2: Ownership by Property Class

I was curious to further explore the relationship between owner type and type of property. This will provide a better idea of the type of land that is owned, and look at it relative to each owner. For this, I filtered the data only to look at the top 3 property types in the data set.

```
classdata<- filter(vacancydatax,
                   vacancydatax$CLASSDSCRP=="RESIDENTIAL VACANT LAND"|
                     vacancydatax$CLASSDSCRP=="RES VACANT LAND W/IMPRV" |
                     vacancydatax$CLASSDSCRP=="VACANT COMMERCIAL LAND")

ggplot(classdata, mapping=aes(fill=CLASSDSCRP, OWNERNME1)) +
  geom_bar() +
  theme(axis.text.x = element_text(color = "grey20", size = 6),
  axis.title.x = element_text(size = 10)) +
  scale_y_continuous(labels=comma) +
  labs(title = "Ownership by Property Class", fill="Property Class" ) +
  xlab("Owner")  +
  ylab("Count")
```



Some things to notice are that the City of Rochester owns a lot of residential land compared to other private owners. Furthermore, private owners have more residential vacant land that has been under some form of improvement. This begs the question, what exactly is being done with the residential vacant land and what are the plans that the city has for it? To confirm that these variables are infact good indicators of one another, a test for association is necessary.

Testing for Association

```
chisq.test(vacancydatax$CLASSDSCRP,vacancydatax$OWNERNME1)
```

```
## Warning in chisq.test(vacancydatax$CLASSDSCRP, vacancydatax$OWNERNME1):
## Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  vacancydatax$CLASSDSCRP and vacancydatax$OWNERNME1
## X-squared = 276.71, df = 8, p-value < 2.2e-16
```

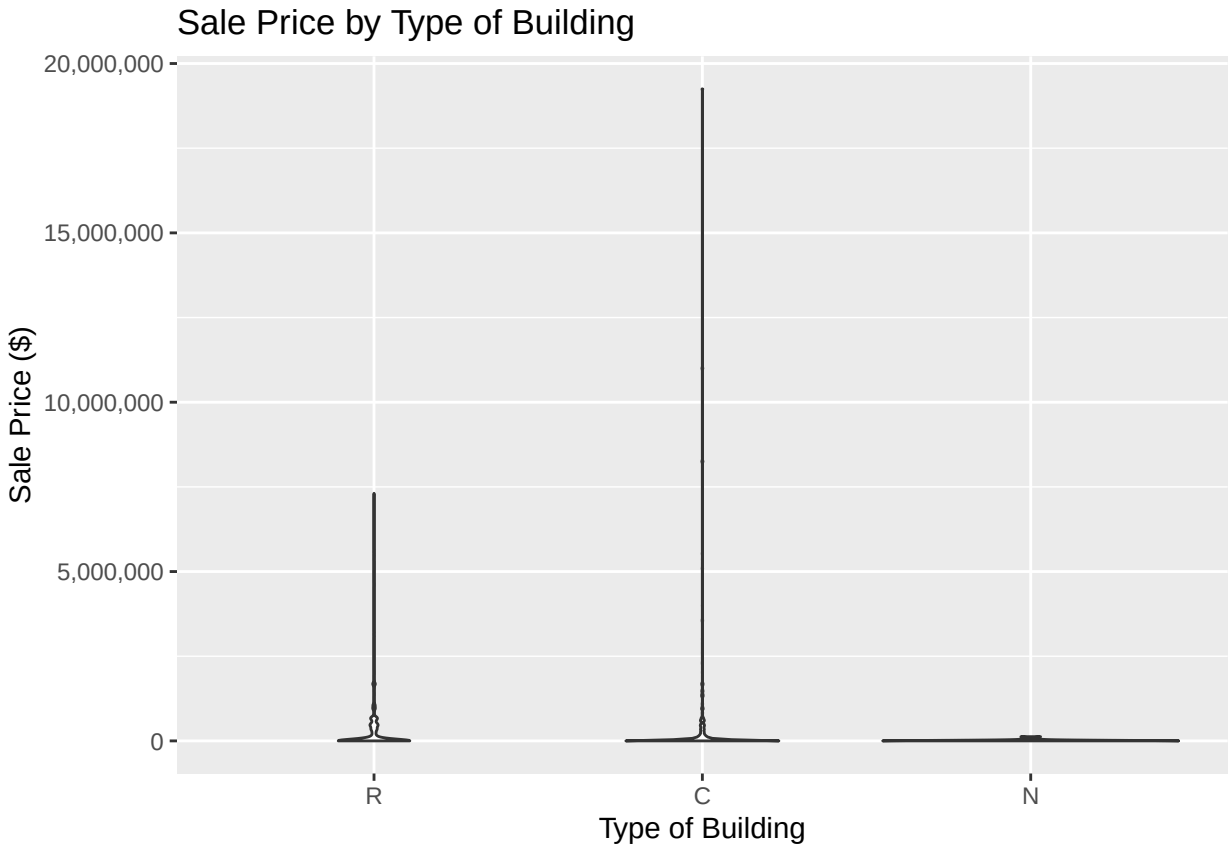
This result from the Chi Square test is statistically significant, since the p-value is significant at the alpha = 0.01 level, it can be concluded that there is an association between Property Class and Ownership.

Part 3: Sale Price and Land Type

I also wanted to explore the relationship between Sale Price and the type of land. Whether it is residential, commerical or other. I thought that the sale price would be higher for commercial buildings than the others.

```
vacancydataxi<- vacancydatax[(!vacancydatax$SALE_PRICE==0),]

ggplot(vacancydatax, aes(y=SALE_PRICE, x=RESCOM)) +
  geom_violin() +
  scale_y_continuous(labels=comma) +
  labs( title = "Sale Price by Type of Building" ) +
  ylab("Sale Price ($)") +
  xlab("Type of Building")
```



In the violin plot above, you can see that the spread of the values for all land types are mostly close to zero. This is because not all of the plots have been sold. Furthermore, you can also see that commercial land has the largest spread and the sale price extends past $15,000,000. To see if the mean sale price differs significantly among the levels of the land type, an Analysis of Variance (anova) is used.

Testing for Difference

```
a1<- aov(vacancydatax$SALE_PRICE~vacancydatax$RESCOM)
anova(a1)
```

```
## Analysis of Variance Table
##
## Response: vacancydatax$SALE_PRICE
##                         Df     Sum Sq    Mean Sq F value    Pr(>F)
## vacancydatax$RESCOM    2 1.4093e+13 7.0465e+12  10.413 3.073e-05 ***
## Residuals           4803 3.2503e+15 6.7672e+11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Through this it can seen that the P value is significant at the alpha=0.01 level, and thus there is a difference between groups. To further explore this, by looking at the summary table below it can be seen that the means differ among the groups. Commercial buildings have the highest mean sale price, while Non-residential and Non-Commerial buildings have the lowest mean sale price.

```
vacancydatax%>% group_by(RESCOM) %>%
  summarise(`Mean`=mean(SALE_PRICE), `Std. Deviation`= sd(SALE_PRICE),
            Frequency=n()) %>% arrange(desc(Mean)) %>%
  kable(caption="Summary of Sale Price by Type of Building")
```

Table 2: Summary of Sale Price by Type of Building

RESCOM	Mean	Std. Deviation	Frequency
C	368249.083	1615872.10	821
R	229539.179	528784.66	3968
N	9793.176	31340.38	17