Group 6 Final Project Draft

Anika Miner, Chris McCabe, & Luke Fredrickson

4/9/2021

I. Introduction

What is the source of the data? Where and when was it created?

The data is sourced from the City of Burlington. In January 2020, an executive order was enacted, establishing the Open Data Policy. Part of our data comes from Burlington's Open Data Portal, in the Rental Property Certificates of Compliance dataset, which contains data on code compliance for every rental property in Burlington, but no information on which landlord owns that property. The other part of our data is scraped from the Burlington Property Database. The Property Database is not downloadable, but it is easy to scrape with a simple python script, and contains more granular detail on property values and landlords. Our final dataset combines both of these datasets to provide a rich view of Burlington rental property data. Both datasets are continuously updated, but our final dataset is built from data accessed on March 26th, 2021. The final dataset can be downloaded here.

If it is a sample, from what population was it drawn, and how was the sample selected?

The data is comprehensive across the entire city of Burlington, and contains data for every single rental property within the city limits.

Do you suspect any sampling bias?

Because the dataset is comprehensive, and not sampled, there is no sampling bias.

Was it an experiment or an observational study?

The data is observational, as it surveys all properties across Burlington, and is collected for city recordkeeping purposes.

How were measurements taken, or questions asked?

Measurements were collected from official city documents, including tax records and official city inspection data. No questions were asked.

Do you suspect any bias in the questions or measurements?

It is possible there is some bias in the assessment of property condition or value due to racial discrimination or other factors, but we don't believe that bias would cause enough variance to be obstructive to our analysis. All other measurements are objective (landlord name, address, coordinates, certificate issue date, etc).

Why is this data of interest to you, and why should the class find it interesting?

On its own, the Rental Property Certificates of Compliance dataset isn't particularly interesting, because it doesn't contain information on property value or who actually owns the property. However, when we combine the CoC data with the Burlington Property Database data, we can analyze which landlords own the most or least property, what neighborhoods they are in, and how compliant they are with Burlington city codes.

What kind of data cleaning was necessary (R code for this must show…)

First, we needed to scrape the data from the Burlington Property Database. We used the python 'Requests' library to query the database, and the 'Beautiful Soup' library to parse through the HTML and grab the data we wanted for each property (tax parcel ID, owner, address, SPAN number, property value, and property taxes). We wrote this data out to

a CSV file with the python ‘CSV’ library. The CoC dataset is separated with semicolons instead of commas, so we needed to read the csv like this:

```
coc <- read.csv("rental-property-certificate-of-compliance.csv", sep=";")
```

We were then able to read in the scraped data csv like normal, and join the two like this:

```
properties <- left_join(properties_scraped, coc, by="TaxParcelId")
```

The property value and property tax data was character data, with $’s and commas, so we needed to convert those columns to numeric data like this:

```
properties$PropertyValue <- as.numeric(gsub('\\$|,', '', properties$PropertyValue))
properties$PropertyTaxes <- as.numeric(gsub('\\$|,', '', properties$PropertyTaxes))
```

The latitude and longitude data was stored in one column, ‘geopoint’, so we split those out into two numeric columns like this:

```
properties <- properties %>%
  mutate(
      lat = as.numeric(unlist(strsplit(as.character(properties$geopoint), ","))[1]),
      long = as.numeric(unlist(strsplit(as.character(properties$geopoint), ","))[2])
  )
```

There were some duplicate columns from joining, and some columns that were not necessary for analysis, so we dropped those columns like this:

```
properties <- properties %>% select(!c(Address, Span, UniqueId, UnitNumber, GISPIN, Update
Date, geopoint))
```

The Burlington properties database was inconsistent in its naming for property owners. There were many instances where a missing comma or an extra period would cause a single landlord to be counted as multiple landlords (“DOE, JOHN M” vs “DOE JOHN M” vs “DOE, JOHN M.” vs “DOE JOHN M.”). This made it seem like there were more single-property landlords than there actually were. To partially fix this, we removed all commas and periods from the owners column like this:

```
properties$Owner <- gsub('\\.|,', '', properties$Owner)
```

This still left several instances where owners were counted incorrectly — discrepancies with middle names were particularly common (“DOE JOHN” vs “DOE JOHN M” vs “DOE JOHN MIDDLE”). There were also instances where LLCs were very similarly named or had typos. There wasn’t an easy way to fix all of these errors in code as the solution for each error was highly contextual, so we had to manually go through the dataset and fix these errors where we saw them.

II. Data Visualizations

```
library(tidyverse)
library(ggmap)
library(viridis)

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

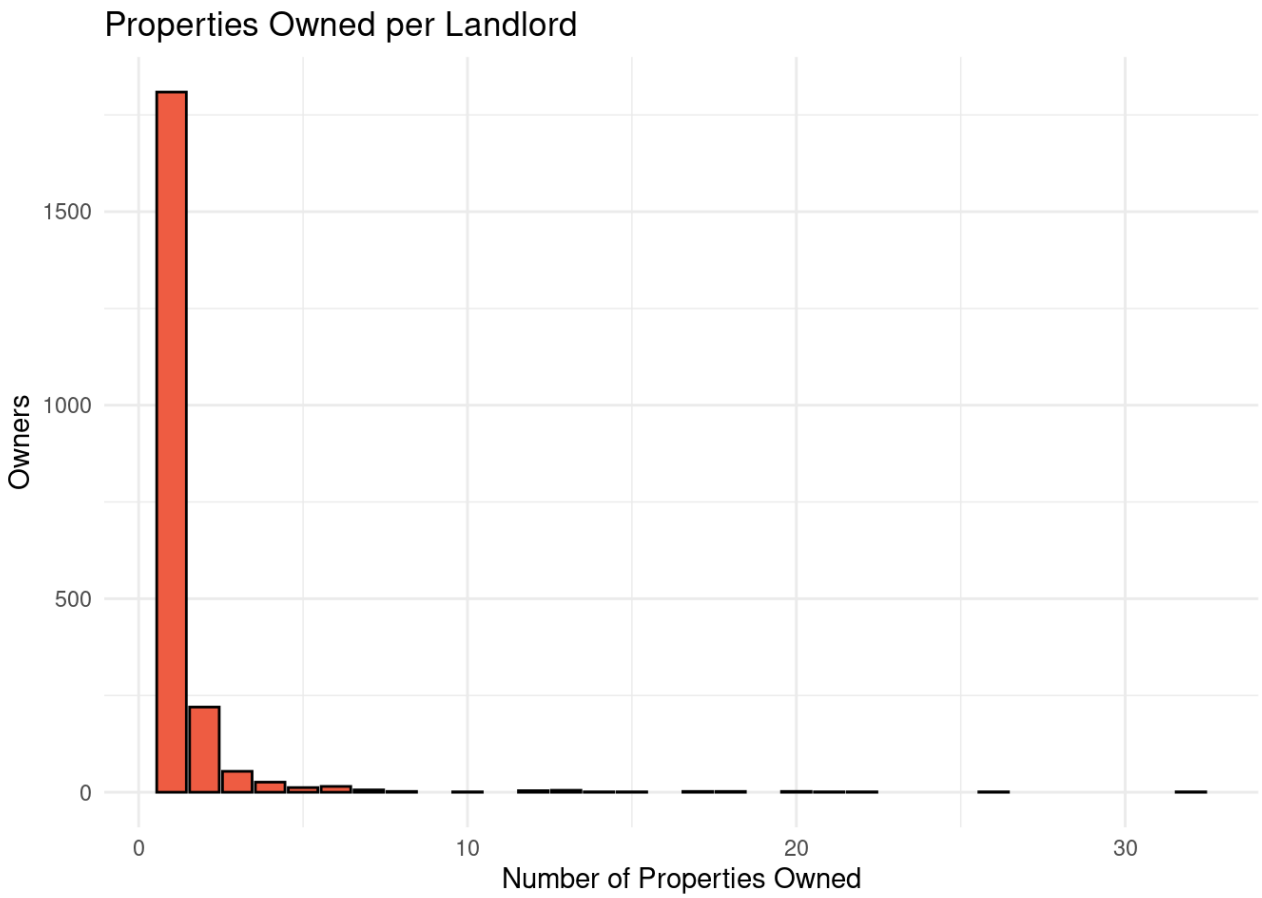
Graph 1: Number of Properties

Out of the 3086 entries in the rental property data we obtained, 2121 of them are owned by owners who own less than 5 properties. Most of that 2121 is made up of the 1809 properties whose owners only own that one property. So to take a deeper look at some of the major landlords and property management companies in Burlington, we filtered the data to include properties whose owners owned at least 5 properties. Shown in the second graph, you can see that a majority of the owners that own more than four properties own 5-7 properties. The outlier in the data set is Diemer Properties who owns a total of 32 properties.

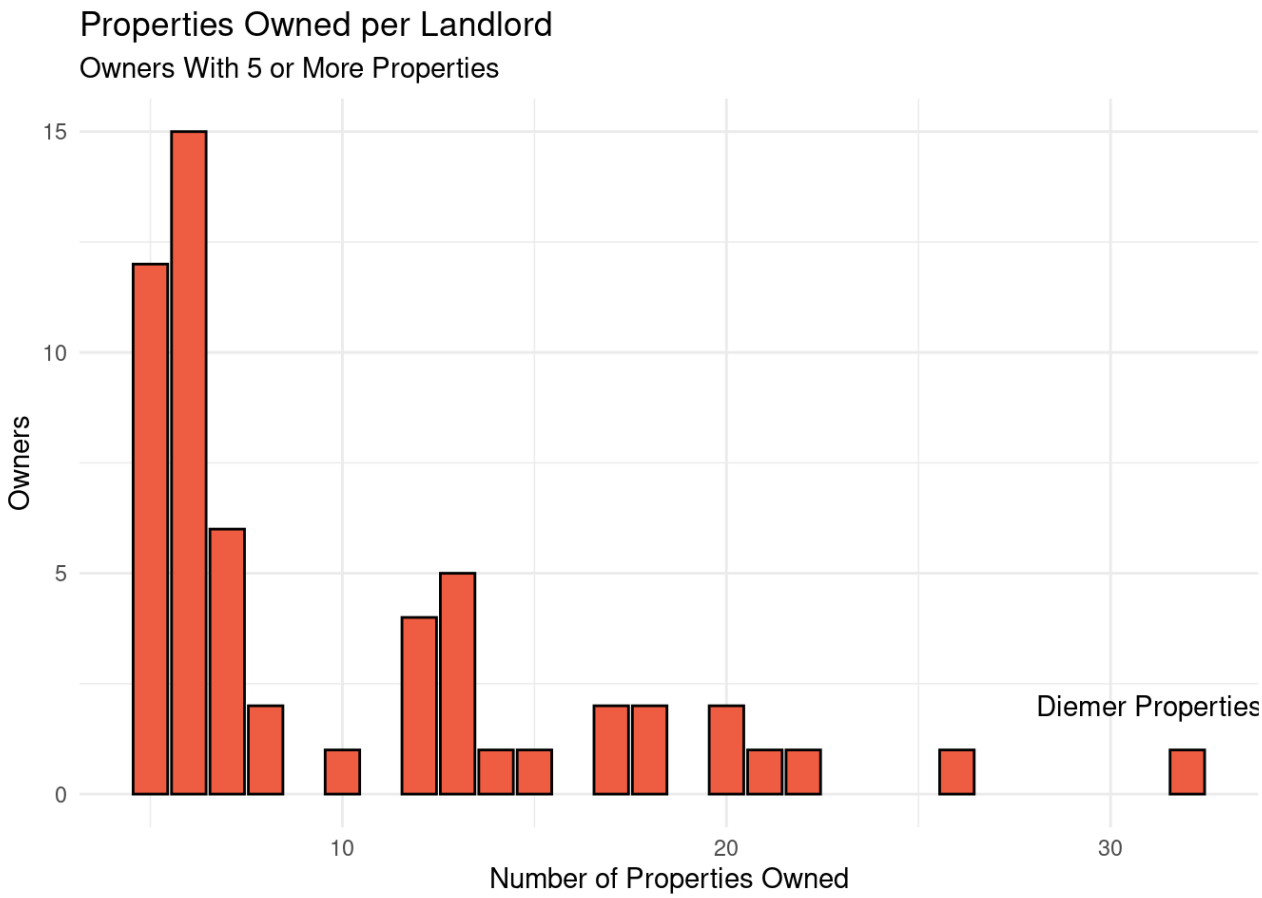
```
NumProperties <- properties %>% group_by(Owner) %>% summarize(nProperties = n())

summary(NumProperties$nProperties)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   1.000   1.425   1.000  32.000
```

```
ggplot(data = NumProperties,
       mapping = aes(nProperties)) +
        geom_bar(color = "black",fill="tomato2")+
        labs(title = "Properties Owned per Landlord",x="Number of Properties Owned",
             y = "Owners")+
        theme_minimal()
```



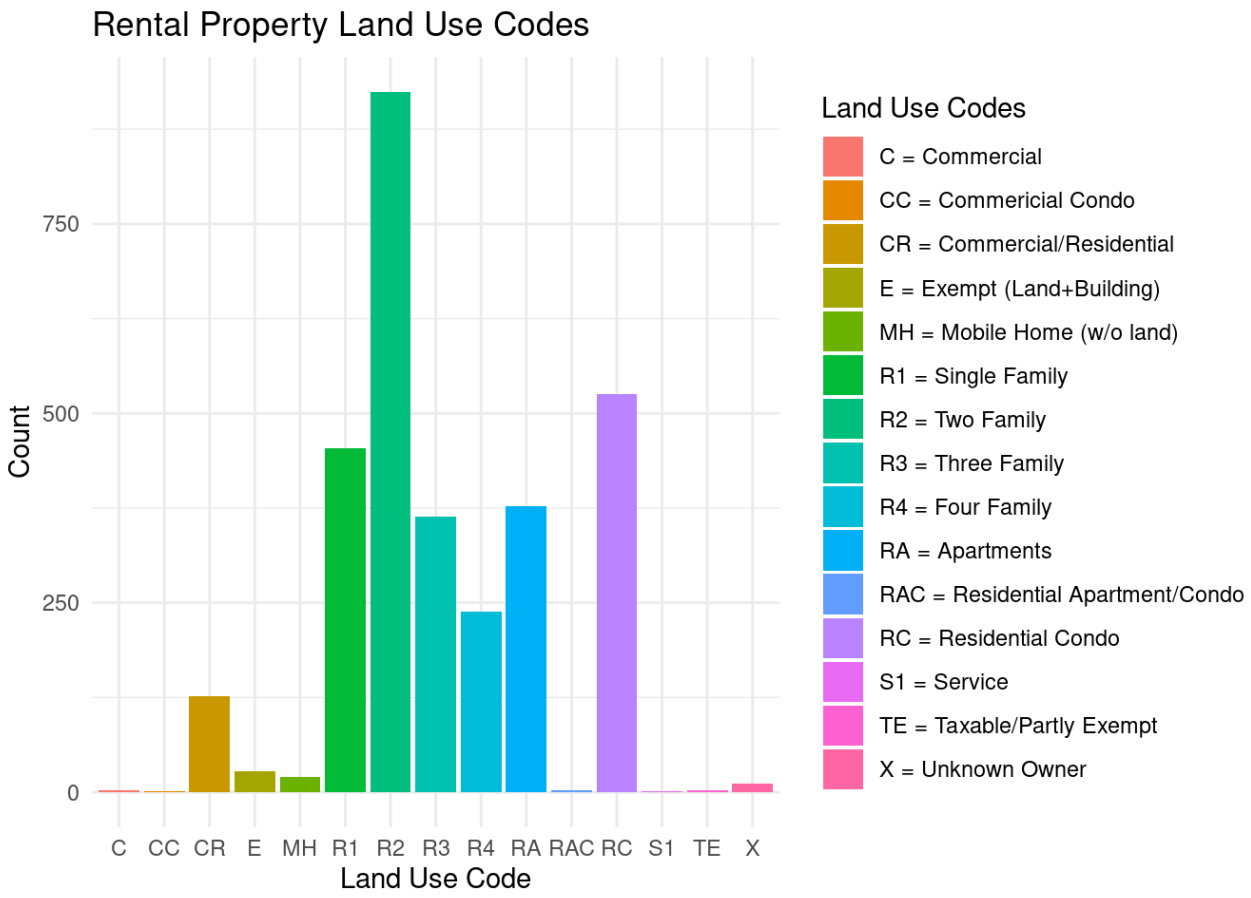

```
ggplot(data = NumProperties %>% filter(nProperties >= 5),mapping = aes(nProperties)) +
        geom_bar(color = "black",fill="tomato2")+
        labs(title = "Properties Owned per Landlord",x="Number of Properties Owned",
             y = "Owners",subtitle = "Owners With 5 or More Properties")+
         annotate( geom = 'text', x = 31, y = 2, label = "Diemer Properties")+
        theme_minimal()
```



Graph 2: Types of Rental Properites

There are 15 active land use codes in our dataset for the rental properties in Burlington. The most common is a two family rental, with one to 5 family apartments, and residential condos taking up the very large majority of properties. The commerical/residential properties (residential above commercial properties e.g. Church Street) are the next most common.

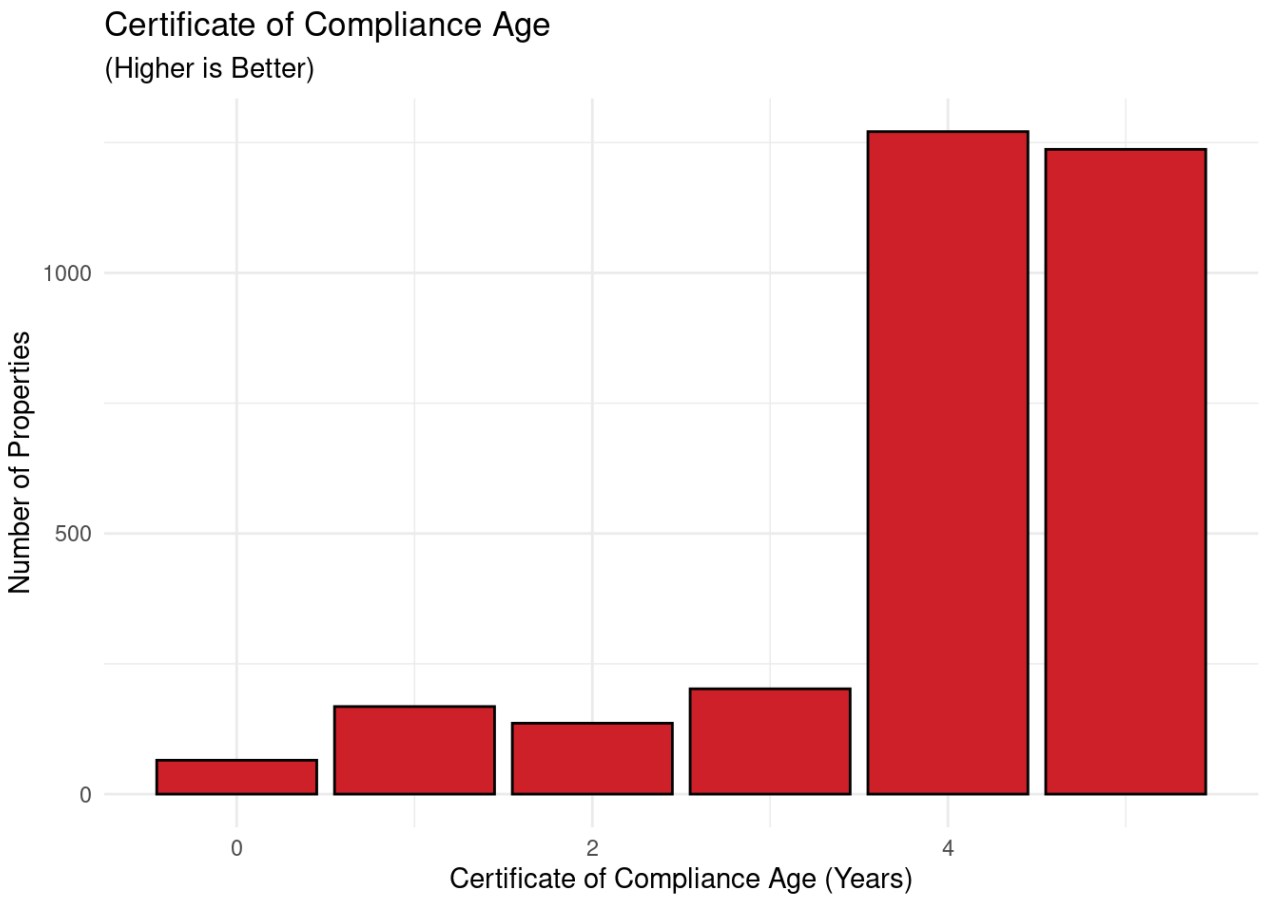
```
ggplot(data = properties %>% filter(!is.na(LandUseCode)),mapping = aes(x = LandUseCode,fil
l = LandUseCode)) +
        geom_bar()+
        labs(title = "Rental Property Land Use Codes",
             x = "Land Use Code",
             y = "Count",
             fill = "Land Use Codes")+
        theme_minimal()+
        scale_fill_discrete(labels = c("C = Commercial","CC = Commericial Condo",
                                       "CR = Commercial/Residential","E = Exempt (Land+Bui
lding)","MH = Mobile Home (w/o land)",
                                       "R1 = Single Family","R2 = Two Family","R3 = Three
Family","R4 = Four Family","RA = Apartments",
                                       "RAC = Residential Apartment/Condo","RC = Residenti
al Condo","S1 = Service","TE = Taxable/Partly Exempt",
                                       "X = Unknown Owner"))
```



Graph 3: Certificate of Compliance

The certificate of compliance is based off house inspections and how many years in a row the property has complied with Burlington's building code. If the property has complied with the code for 5 years, they will have a 5 year Certificate of Compliance. The majority of properties have 4 or 5 year certificates. Properties with a 0 year certificate either failed inspections in the most recent inspection or have not been inspected due to COVID-19.

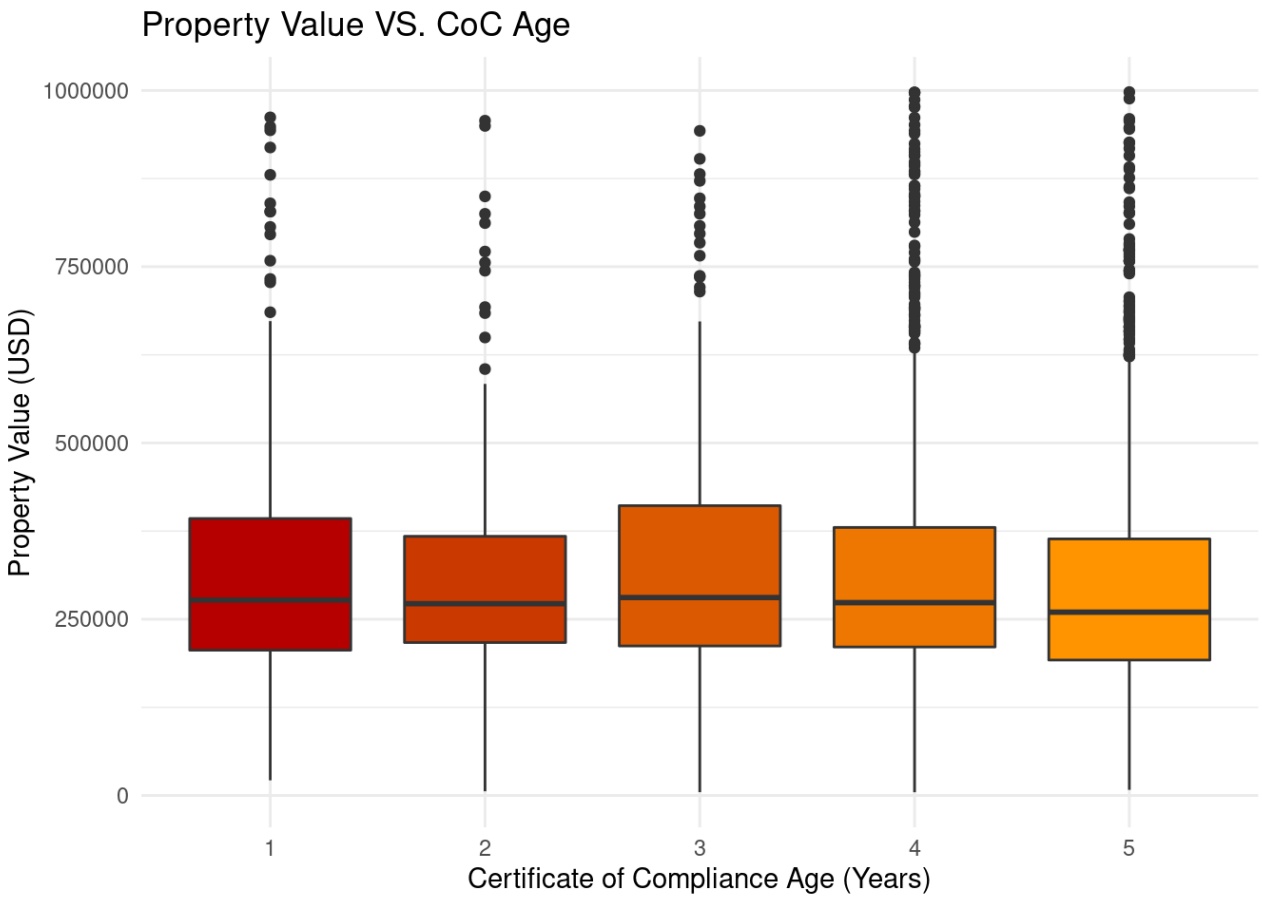
```
ggplot(data = properties %>% filter(!is.na(CoCYears)), mapping = aes(x=CoCYears)) +
        geom_bar(color = "black",fill = "#CE2029")+
        labs(title = "Certificate of Compliance Age",
             subtitle = "(Higher is Better)",
             x="Certificate of Compliance Age (Years)",
             y = "Number of Properties")+
        theme_minimal()
```



Graph 4: Does CoC Affect Property Value

The original hypothesis was that the properties with a less desirable CoC would have a lower property value. However, from the boxplots below we learned that the property value across all year values for the CoC was very similar. The median property value for all CoC years was ~$275,000.

```
ggplot(data = properties %>%
               filter(!is.na(CoCYears) & CoCYears != 0 & !is.na(LastMhInspectionDate) & Pr
opertyValue < 1000000),
       mapping = aes(x=factor(CoCYears), y=PropertyValue,fill = CoCYears)) +
        geom_boxplot()+
        labs(title = "Property Value VS. CoC Age",
             x = "Certificate of Compliance Age (Years)",
             y = "Property Value (USD)")+
        scale_fill_gradient(high = "#ff9300",low = "#b60000")+
        guides(fill=FALSE)+
        theme_minimal()
```



Graph 5: Top 20 Landlords in Burlington

To give some context and greater detail on the landlords in Burlington, here are the top 20 landlords, their total properties owned, and the total value of those combined properties. There are a few interesting outliers in terms of total value – notably, Claire Pointe Owners Association controls 18 properties, but, combined, the properties it does own are 6 times the value of all the Diemer Properties properties combined, even though Diemer properties owns almost twice as many properties.

```
library(knitr)

kable(properties %>%
  group_by(Owner) %>%
  summarize(
    NumberOfProperties = n(),
    TotalValue = sum(PropertyValue)
  ) %>%
  arrange(desc(NumberOfProperties)) %>%
  head(20, NumberOfProperties), caption="Top 20 Landlords in Burlington by Number of Prope
rties Owned")
```

Top 20 Landlords in Burlington by Number of Properties Owned

Owner	NumberOfProperties	TotalValue
DIEMER PROPERTIES	32	5695800

Owner	NumberOfProperties	TotalValue
MCGOWAN JOHN STUART	26	8411300
CHAMPLAIN HOUSING TRUST INC	22	6491200
BPJS MANAGEMENT LLC	21	8839200
J & S LLC	20	9104400
ROONEY RICHARD A	20	3280900
CLAIRE POINTE OWNERS ASSOCIATION	18	33130800
PBGC LLC	18	9997500
BURLINGTON REALTY ASSOCIATES	17	2463700
SISTERS & BROTHERS INVESTMENT GROUP LLP	17	13985100
LARKIN JOHN INC	15	3132900
OFFENHARTZ INC	14	6359300
BOYDEN DOUGLAS G	13	5318800
BURLINGTON HOUSING AUTHORITY	13	22981220
PHE INC	13	2634700
RIELEY PROPERTIES LLC	13	7468500
SWB LLC	13	5180000
BURNS CHARLES C	12	3937100
KHAMNEI CHRIS C	12	5832500
LAFAYETTE MELISSA B	12	3812600

IV. Conclusions

Write some overall conclusions – an overall summary of what you learned from your visualizations and analysis. Summarize in one paragraph.

The dataset we used included data from a little over 3000 properties in Burlington, VT, and goes into their certificates of compliance. Our visualizations of this dataset, which we combined with data from the Burlington property database, are quite helpful in analysis. Our graph of the number of properties owned per landlord (Graph 1) showed that it is much more common for landlords to own only a few properties, and the data is very skewed right. Graph 2 shows the different rental types: what's popular and what is not. The type 'R2' or 2 Family, was found to be most common, with 'RC', or Residential Condo, close behind. Graph 3 showed us that certificates of compliance by years had a left skew. Most of them last four or five years, with a significantly lower number of any less years. Graph 4 displays the similarities between certificate of compliance years and property value, and how regardless of years certificates of compliance were in place, property value stayed more or less the same.

V. Limitations/Recommendations

Write a paragraph describing some of the limitations that are inherent in your study. Also discuss ideas for

future research that might build on the work you did in this project. Summarize in one paragraph.

The major limitations of this dataset stem from the highly contextual nature of property ownership records. A landlord can be listed as the owner of a property under their legal name or an LLC, and a single person can own multiple LLCs. If a landlord owns a large number of properties but wishes to obfuscate that fact, they can distribute ownership of those properties throughout a handful of different LLCs. Similarly, a family can control a large number of properties collectively — take the Bissonette family, for example. Their LLC, “BPJS MANAGEMENT LLC”, is listed as the owner of 21 different properties in the database, but the family collectively owns 6 additional properties under their respective legal names, and may potentially control even more properties via LLCs. Further research could attempt to rectify this problem via deep analysis of who actually owns each individual LLC, as those documents are likely public. This would be a very time-consuming endeavor, but would give a much more granular and accurate view of who controls property in Burlington.