Final Project (Teaching Case)

Emmett Greenberg

2023-12-13

Overview

My analysis is inspired by wanting to better understand the factors associated with building permit values. In particular, I wanted to find out the differences between commercial permits and residential permits. For each, what are the variables that have a positive association with permit values? Are these variables different when looking at commercial vs. residential permit types?

Permit values are an indicator of real estate demand. As such, business owners might have different preferences when it comes to location than homeowners/renters have. In thinking about some of the variables that might relate to such demand, I am considering data related to nearby transportation options, as well as census indicators such as population density, median income, race, home ownership, and more. The skills essential to doing this analysis include aggregation, merging data sets, making maps, and statistical inference using t-tests and multivariate regression.

Data Preparation

In addition to the building permits dataset, I am using following data sets:

- Land Parcels (from BARI’s 2022 Geographical Infrastructure)
- Bluebikes (bike share) Stations (from course material)
- ACS Indicators 2015-19 by Tract (from BARI’s Massachusetts Census Indicators)

Steps after loading the data sets:

- Cleanse the data sets, removing duplicates and entries for which Census Tract (CT_ID_10) is missing.
- New record-level variables
    - Permit_Value: The declared permit valuation per square foot.
    - parcel_comm and parcel_res: Binary (0 or 1) variables that denote whether a parcel is commercial and residential, respectively.
        - Commercial and residential are categorized using the parcel's land use (LU) variable. Commercial includes mixed residential-commercial type (RC) because it is closer in nature to commercial buildings.
        - *Functions*: ifelse
- Join permits and parcels data:
    - Aggregate permits by Land Parcel ID
        - *Variables*: total_permits, avg_permit_value
    - Merge into parcels on Land Parcel ID, keeping all data from parcels to ensure accuracy.
        - *Functions*: right_join
- Aggregate parcels_permits data by Census Tract, the unit I want to use for my analysis.
    - pp_comm_byCT for commercial permits
    - pp_res_byCT for residential permits.
    - New Variables:
        - Total commercial permits per CT (total permits on commercial parcels)
        - Total Residential permits per CT (total permits on res parcels)
        - Total commercial parcels
        - Total residential parcels
        - Average permit values per CT (average of averages per parcel)

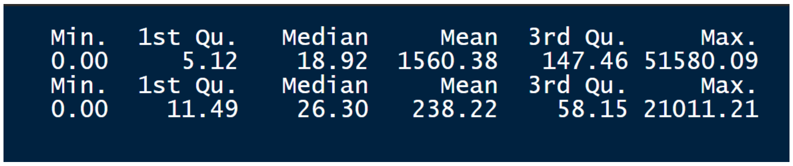
  - *Functions*: group_by, summarize, sum, mean
- Merge **ACS** data into each of comm/res permits data.
- Aggregate **BlueBikes** stations by CT, measuring total stations and total docks.
- Merge **BlueBikes** data into each of comm/res permits data.

Now, I have two data frames, *pp_comm_data* (Commercial permits) and *pp_res_data* (Residential permits), which I will use for my analysis.

Analysis

Summary Statistics for the commercial permits:

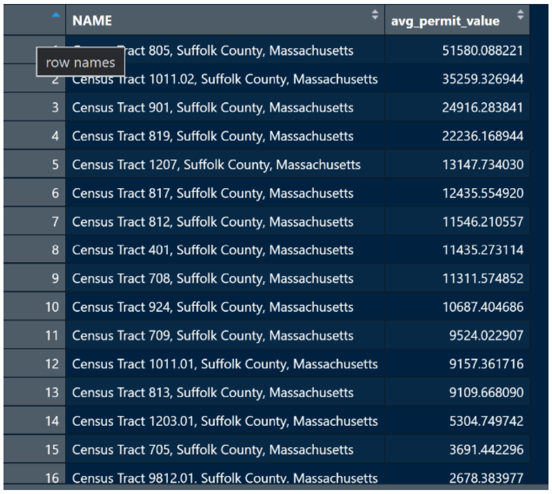
```
summary(pp_comm_data$avg_permit_value)

summary(pp_res_data$avg_permit_value)
```



```
Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.00     5.12    18.92  1560.38   147.46 51580.09
Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.00    11.49    26.30   238.22    58.15 21011.21
```

When looking at these summary statistics, the differences in average permit values stand out. At every quantile, the commercial group's permit values are higher than the residential group's permit values. At the 1st quantile, it is 11.49 vs. 5.12; at the median, 26.30 vs. 18.92, and at the 3rd quantile, 147.46 vs 58.15. This indicates that the lower halves of values are closer to each other than the upper halves are to each other. The mean and max values also indicate that as these respective values increase, they become more dispersed. This is especially evident with each group's mean value being much greater than its 3rd quantile, and with the max values being so much higher than any of the other statistics.

Outliers:



	NAME	avg_permit_value
1	ract 805, Suffolk County, Massachusetts	51580.088221
2	Census Tract 1011.02, Suffolk County, Massachusetts	35259.326944
3	Census Tract 901, Suffolk County, Massachusetts	24916.283841
4	Census Tract 819, Suffolk County, Massachusetts	22236.168944
5	Census Tract 1207, Suffolk County, Massachusetts	13147.734030
6	Census Tract 817, Suffolk County, Massachusetts	12435.554920
7	Census Tract 812, Suffolk County, Massachusetts	11546.210557
8	Census Tract 401, Suffolk County, Massachusetts	11435.273114
9	Census Tract 708, Suffolk County, Massachusetts	11311.574852
10	Census Tract 924, Suffolk County, Massachusetts	10687.404686
11	Census Tract 709, Suffolk County, Massachusetts	9524.022907
12	Census Tract 1011.01, Suffolk County, Massachusetts	9157.361716
13	Census Tract 813, Suffolk County, Massachusetts	9109.668090
14	Census Tract 1203.01, Suffolk County, Massachusetts	5304.749742
15	Census Tract 705, Suffolk County, Massachusetts	3691.442296
16	Census Tract 9812.01, Suffolk County, Massachusetts	2678.383977

There are several very high values here, and they might be due to input mistakes, but I don't see any overwhelming evidence that they should be removed, so I will keep them.

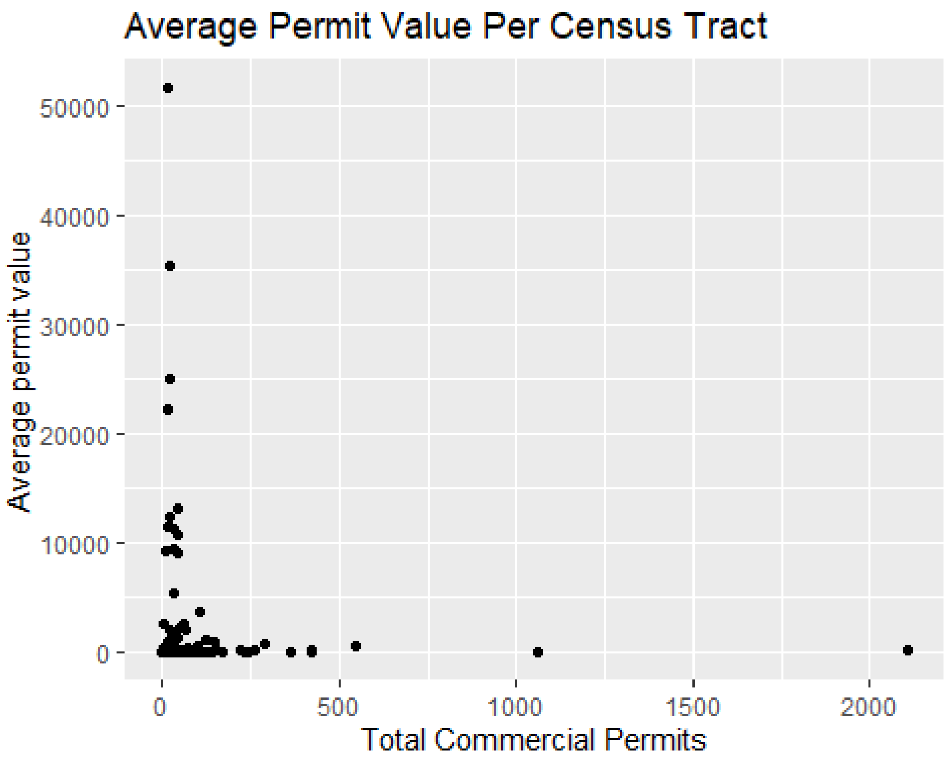
Comparing Residential and Commercial permits:

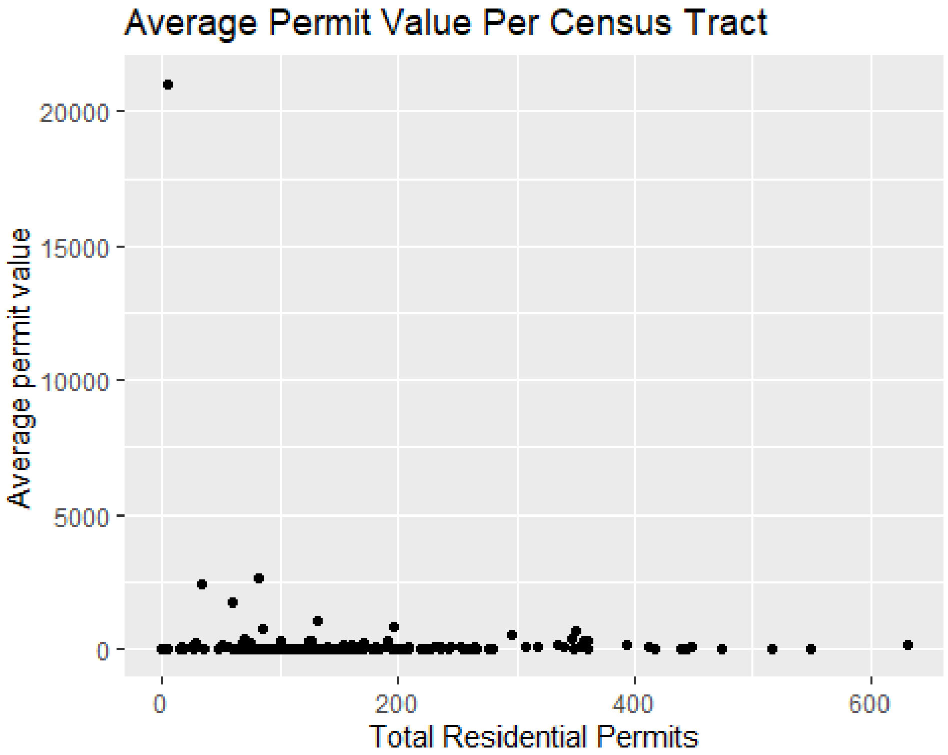
How similar are their permit values, on average? I will use a t-test to find out.

```
t.test(pp_comm_byCT$avg_permit_value, pp_res_byCT$avg_permit_value)

##
##  Welch Two Sample t-test
##
## data:  pp_comm_byCT$avg_permit_value and pp_res_byCT$avg_permit_value
## t = 2.869, df = 198.87, p-value = 0.004563
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   413.3814 2230.9371
## sample estimates:
## mean of x mean of y
## 1560.3777  238.2184
```

These results show that there is a statistically significant difference between average permit values for commercial and residential permits (p < 0.004563). The mean value is much higher for comm permits.

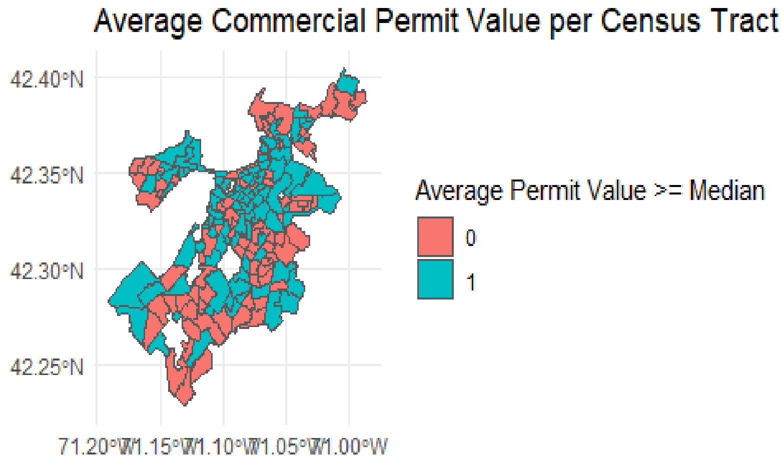
Below are dot plots showing the distributions of these values:

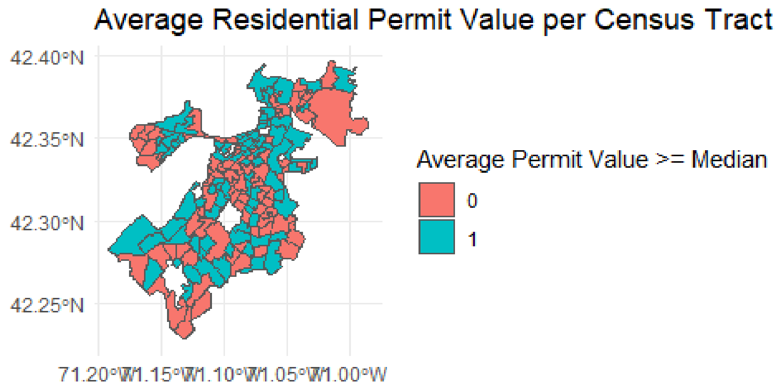

Average Permit Value Per Census Tract
Average permit value
50000
40000
30000
20000
10000
0
0
500
1000
1500
2000
Total Commercial Permits


Average Permit Value Per Census Tract
Average permit value
20000
15000
10000
5000
0
0
200
400
600
Total Residential Permits

Both commercial and permits tend to have higher values in areas where there are not very many of them, which is interesting. It is not surprising that when permit activity is rarer, the associated costs will be higher. There may be a need to perform a more extensive assessment of land on which there are fewer buildings, to assess the quality of the soil and other characteristics that play a role in building construction.

To visualize how permit values manifest throughout the city, I am mapping average permit values onto census tracts to see where the lower and upper halves are. To do this, I am creating a "binary" permit value variable where the value is 0 if it is less than or equal to median, and 1 if it is greater than the median.


Average Commercial Permit Value per Census Tract
42.40°N
42.35°N
42.30°N
42.25°N
71.20°W
71.15°W
71.10°W
71.05°W
71.00°W
Average Permit Value >= Median
0
1


Average Residential Permit Value per Census Tract
42.40°N
42.35°N
42.30°N
42.25°N
71.20°W
71.15°W
71.10°W
71.05°W
71.00°W
Average Permit Value >= Median
0
1

There is a lot of overlap in the areas that have higher values for both commercial and residential permits. Many of these areas are closer to downtown, and the relatively affluent neighborhoods of Fenway-Kenmore and West Roxbury also have greater values for both occupancy types.

Regression

I am running multivariate regressions on the commercial and permits data, with the goal of finding variables associated with average permit value.

From my City Exploration assignment, I learned that there is a statistically significant correlation between higher permit values and more BlueBikes docking stations, so I want to include that in my analysis. From the Census data, I suspect that population density tends to correlate positively as well, because business and retail benefit from being able to serve a lot of people that live in the neighborhood. Additionally, median income and shorter commute times likely have positive associations with permit values.

In building my model, I tried many combinations of variables before settling on these.

```
## Call:
## lm(formula = avg_permit_value ~ PopDen + MedHouseIncome + bbike_docks + 
##     CommuteLess10, data = pp_comm_data)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
##  -4111  -2275  -1314    -77  46891 
## 
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)   
## (Intercept)     3.733e+03  1.359e+03   2.748  0.00668 **
## PopDen          2.717e-02  2.762e-02   0.984  0.32665   
## MedHouseIncome -3.175e-02  1.238e-02  -2.565  0.01122 * 
## bbike_docks     2.196e+01  1.361e+01   1.614  0.10851   
## CommuteLess10  -1.506e+04  7.816e+03  -1.927  0.05571 . 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 5786 on 162 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.05927,    Adjusted R-squared:  0.03604
## F-statistic: 2.552 on 4 and 162 DF,  p-value: 0.04113
```

As shown, median income has a statistically significant association with commercial permit values (p < 0.5), and so does the proportion of residents with a commute of less than 10 minutes. In other versions of this model, white population, ethnic diversity, and BlueBikes stations were significant factors.

```
##
## Call:
## lm(formula = avg_permit_value ~ PopDen + MedHouseIncome + bbike_docks +
##     CommuteLess10, data = pp_res_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1849.8  -355.8  -188.6    96.6 20118.6
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     5.769e+02  3.885e+02   1.485   0.1395
## PopDen          2.942e-03  7.795e-03   0.377   0.7064
## MedHouseIncome -5.989e-03  3.510e-03  -1.706   0.0899 .
## bbike_docks     7.972e+00  3.857e+00   2.067   0.0403 *
## CommuteLess10  -2.376e+03  2.214e+03  -1.073   0.2848
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1638 on 162 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.03818,    Adjusted R-squared:  0.01443
## F-statistic: 1.608 on 4 and 162 DF,  p-value: 0.1748
```

Median income is shown to have a less significant association with residential permit values (p < 0.1). Interestingly, the presence of BlueBikes stations is significant (p < 0.5).

This goes against my findings from previous assignments, where BlueBikes docking stations correlated only with *commercial* permit values. Therefore, this is unlikely to a causal relationship, but rather a by-product of having greater bikeshare access in areas with higher median incomes, which also tend to be closer to downtown, and which tend to have higher commercial *and* residential values.

Conclusion

It is very difficult to pinpoint the root cause of the variance in building permit valuations. However, characteristics of the surrounding population such as income, race/ethnicity, and transportation access tend to be associated with permit activities and expenditures. Addressing the inequities in places with lower permit values is critical to improving the affordability of housing and residential properties, as well as improving the affordability and access to goods and services that commercial properties provide. The skills of analyzing this data set, especially multivariate regression, can be applied to more broadly to addressing the housing and rental affordability crisis in Boston.

Data Dictionary

Record-Level

permits

- *Permit_Value: The declared permit valuation per square foot.*

parcels

- parcel_comm: Denotes whether parcel is commercial (1) or not (0).
- parcel_res: Denotes whether a parcel is commercial (1) or not (0).

Aggregate-Level

permits_byLP: permits aggregated by Land Parcel ID.

- *total_permits: The number of permits in a given parcel.*
- avg_permit_value: The average permit value for a given parcel.

pp_comm_byCT: commercial permits aggregated by 2010 Census Tract (CT_ID_10).

- *total_permits:* Total commercial permits in a given CT.
- *total_comm_parcels*: Total commercial permits in a given CT.
- *avg_permit_value:* The average commercial permit value for a given CT.

pp_res_byCT: residential permits aggregated by 2010 Census Tract (CT_ID_10).

- *total_permits:* Total residential permits in a given CT.
- *total_res_parcels:* Total residential permits in a given CT.
- *avg_permit_value:* The average residential permit value for a given CT.

bbikes_byCT: BlueBikes stations data aggregated by 2010 Census Tract (CT_ID_10).

- *bbike_sta:* Total stations in a given CT.
- *bbike_docks:* Total bike docks in a given CT.

Annotated Syntax

Data Preparation

```
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
## ✔ dplyr     1.1.3     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.4     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.0
## ✔ purrr     1.0.2
## ── Conflicts ──────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors

library(ggplot2)
library(sf)

## Linking to GEOS 3.11.2, GDAL 3.7.2, PROJ 9.3.0; sf_use_s2() is TRUE

# Permits / Parcels
all_permits <-
read.csv("../dataverse_files/Permits.Records.Geocoded.2023.csv")
all_parcels <- read.csv("../dataverse_geo/LandParcels.2022.csv")
tracts_shp <- st_read('../Tracts_Boston_2010_BARI/Tracts_Boston BARI.shp')

## Reading layer `Tracts_Boston BARI' from data source
##   `C:\Users\emmet\OneDrive\Northeastern\MSUI\MSUI Fall 2023\Big Data for
Cities (PPUA 5263)\Building Permits\Tracts_Boston_2010_BARI\Tracts_Boston
BARI.shp'
##   using driver `ESRI Shapefile'
## Simple feature collection with 178 features and 16 fields
## Geometry type: POLYGON
## Dimension:     XY
## Bounding box:  xmin: -71.19115 ymin: 42.22788 xmax: -70.98471 ymax:
42.40493
## Geodetic CRS:  NAD83

# Census indicators
acs_1519_TRACT <- read.csv("../Census_Indicators/ACS_1519_TRACT.csv") # ACS
2015-2019
dec_2010_TRACT <- read.csv("../Census_Indicators/DEC_CENSUS_2010_TRACT.csv")
# BlueBikes Stations
all_bbikes <-
```

```
read.csv("../Blue_Bike_Stations/Bluebikes_Stations_spatial.csv")
bbikes_shp <- st_read("../Blue_Bike_Stations/Blue_Bike_Stations.shp")

## Reading layer `Blue_Bike_Stations' from data source
##   `C:\Users\emmet\OneDrive\Northeastern\MSUI\MSUI Fall 2023\Big Data for
Cities (PPUA 5263)\Building
Permits\Blue_Bike_Stations\Blue_Bike_Stations.shp'
##   using driver `ESRI Shapefile'
## Simple feature collection with 462 features and 8 fields
## Geometry type: POINT
## Dimension:     XY
## Bounding box:  xmin: -71.24776 ymin: 42.2556 xmax: -70.87021 ymax:
42.53467
## Geodetic CRS:  WGS 84

# Removing Cert of Occupancy permits.
# These are essentially duplicates,  representing work that has already been
completed though previous permits.
permits_noCOO <- all_permits %>% filter(permittypedescr != "Certificate of
Occupancy")

# Remove duplicate permits with identical permit number, since permit number
is a unique identifier.
permits_noDup <- permits_noCOO[!duplicated(subset(permits_noCOO, select =
c(PermitNumber))),]

# Removing entries where census tract is missing (NA).
permits_noNACT <- permits_noDup %>% filter(!is.na(CT_ID_10))
parcels_noNACT <- all_parcels %>% filter(!is.na(CT_ID_10))
bbikes_noNACT <- all_bbikes %>% filter(!is.na(CT_ID_10))

# Save the cleansed permits data as "permits", and the cleansed parcels data
as "parcels".
permits <- permits_noNACT
parcels <- parcels_noNACT

# permit values
permits$permit_value <- permits$DECLARED_VALUATION / permits$sq_feet
summary(permits$permit_value)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##    -Inf     Inf     Inf     NaN     Inf     Inf   26891

# remove NAs and infinites
permits <- permits %>% filter(!is.na(permit_value),
!is.infinite(permit_value))
summary(permits$permit_value)

##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
##        0        1        4     1251       25 15000787
```

```
# Using the "LU" (Land Use) variable from the parcels data to denote whether
each parcel is commercial or residential.

# land use categorizations
comm_LUs <- c("CC", "C", "RC")
res_LUs <- c("R1", "R2", "R3", "R4", "A", "CD")

# boolean variables
parcels$parcel_comm <- ifelse(parcels$LU %in% comm_LUs, 1, 0)
parcels$parcel_res <- ifelse(parcels$LU %in% res_LUs, 1, 0)

# 1. Aggregate permits by parcel ID
permits_byLP <- permits %>%
    group_by(Land_Parcel_ID) %>%
    summarise(total_permits=n(), avg_permit_value=mean(permit_value))

# 2. Merge the new data frame with the parcels data.
parcelsAndPermits <- right_join(permits_byLP, parcels, by="Land_Parcel_ID")

# 3. Add in the missing values for new aggregate measures.
# *Where the total permits is NA, replace it with "0".
# *Where average permit value NA, replace it with "0".
parcelsAndPermits<-parcelsAndPermits%>%
    mutate(total_permits= ifelse(is.na(total_permits), 0, total_permits)) %>%
    mutate(avg_permit_value= ifelse(is.na(avg_permit_value), 0,
avg_permit_value))

# commercial
pp_comm_byCT <- parcelsAndPermits %>%
    group_by(CT_ID_10) %>%
    filter(parcel_comm==1) %>%
    summarise(total_comm_permits=sum(total_permits),
              total_comm_parcels=n(),
              avg_permit_value=mean(avg_permit_value))

summary(pp_comm_byCT)

##     CT_ID_10          total_comm_permits total_comm_parcels
avg_permit_value
##  Min.   :2.503e+10   Min.   :   0.00    Min.   :  1.00     Min.   :
0.00
##  1st Qu.:2.503e+10   1st Qu.:  12.00    1st Qu.: 16.00     1st Qu.:
5.12
##  Median :2.503e+10   Median :  27.50    Median : 29.50     Median :
18.92
##  Mean   :2.503e+10   Mean   :  68.94    Mean   : 41.54     Mean   :
1560.38
##  3rd Qu.:2.503e+10   3rd Qu.:  55.00    3rd Qu.: 53.00     3rd Qu.:
147.46
```

```
##  Max.   :2.503e+10   Max.   :2105.00    Max.   :309.00     Max.   
:51580.09

# residential
pp_res_byCT <- parcelsAndPermits %>%
    group_by(CT_ID_10) %>%
    filter(parcel_res==1) %>%
    summarise(total_res_permits=sum(total_permits),
              total_res_parcels=n(),
              avg_permit_value=mean(avg_permit_value))

summary(pp_res_byCT)

##     CT_ID_10         total_res_permits total_res_parcels avg_permit_value
##  Min.   :2.503e+10   Min.   :  0.0     Min.   :   1.0    Min.   :    0.00
##  1st Qu.:2.503e+10   1st Qu.: 89.5     1st Qu.: 186.8    1st Qu.:   11.49
##  Median :2.503e+10   Median :151.5     Median : 408.0    Median :   26.30
##  Mean   :2.503e+10   Mean   :170.4     Mean   : 449.7    Mean   :  238.22
##  3rd Qu.:2.503e+10   3rd Qu.:230.0     3rd Qu.: 611.2    3rd Qu.:   58.15
##  Max.   :2.503e+10   Max.   :631.0     Max.   :1734.0    Max.   :21011.21

# Remove entries from ACS that are not in the same county as Boston, which is
"Suffolk."
acs_BOS <- acs_1519_TRACT %>% filter(str_detect(NAME, "Suffolk County"))

# merge
pp_comm_ACS <- right_join(acs_BOS,pp_comm_byCT, by="CT_ID_10")

# merge
pp_res_ACS <- right_join(acs_BOS, pp_res_byCT, by="CT_ID_10")

bbikes_byCT <- bbikes_noNACT %>%
    group_by(CT_ID_10) %>%
    summarize(bbike_sta=n(), bbike_docks= sum(Total.docks))

summary(bbikes_byCT)

##     CT_ID_10           bbike_sta       bbike_docks
##  Min.   :2.503e+10   Min.   : 1.000   Min.   : 10.00
##  1st Qu.:2.503e+10   1st Qu.: 1.000   1st Qu.: 17.00
##  Median :2.503e+10   Median : 2.000   Median : 29.00
##  Mean   :2.503e+10   Mean   : 2.153   Mean   : 38.43
##  3rd Qu.:2.503e+10   3rd Qu.: 3.000   3rd Qu.: 49.00
##  Max.   :2.503e+10   Max.   :14.000   Max.   :314.00

# keep all rows from pp_comm_ACS
pp_comm_data <- right_join(bbikes_byCT, pp_comm_ACS, by="CT_ID_10")
pp_res_data <- right_join(bbikes_byCT, pp_res_ACS, by="CT_ID_10")

summary(pp_comm_data$bbike_sta)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##   1.000   1.000   2.000   2.165   3.000  14.000      57
```

```
summary(pp_comm_data$bbike_docks)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##   10.00   17.00   29.00   38.69   49.50  314.00      57
```

```
# replace NAs (stations, docks) with 0
columns_to_replace <- c("bbike_sta", "bbike_docks")

pp_comm_data <- pp_comm_data %>%
    mutate_at(vars(columns_to_replace), ~ifelse(is.na(.), 0, .))
```

```
## Warning: Using an external vector in selections was deprecated in
tidyselect 1.1.0.
## ℹ Please use `all_of()` or `any_of()` instead.
##   # Was:
##   data %>% select(columns_to_replace)
## 
##   # Now:
##   data %>% select(all_of(columns_to_replace))
## 
## See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
pp_res_data <- pp_res_data %>%
    mutate_at(vars(columns_to_replace), ~ifelse(is.na(.), 0, .))

summary(pp_comm_data$bbike_sta)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.000   0.000   1.000   1.448   2.000  14.000
```

```
summary(pp_comm_data$bbike_docks)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.00    0.00   17.00   25.87   34.00  314.00
```

Analysis

```
summary(pp_comm_data$avg_permit_value)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
##     0.00     5.12    18.92  1560.38   147.46 51580.09
```

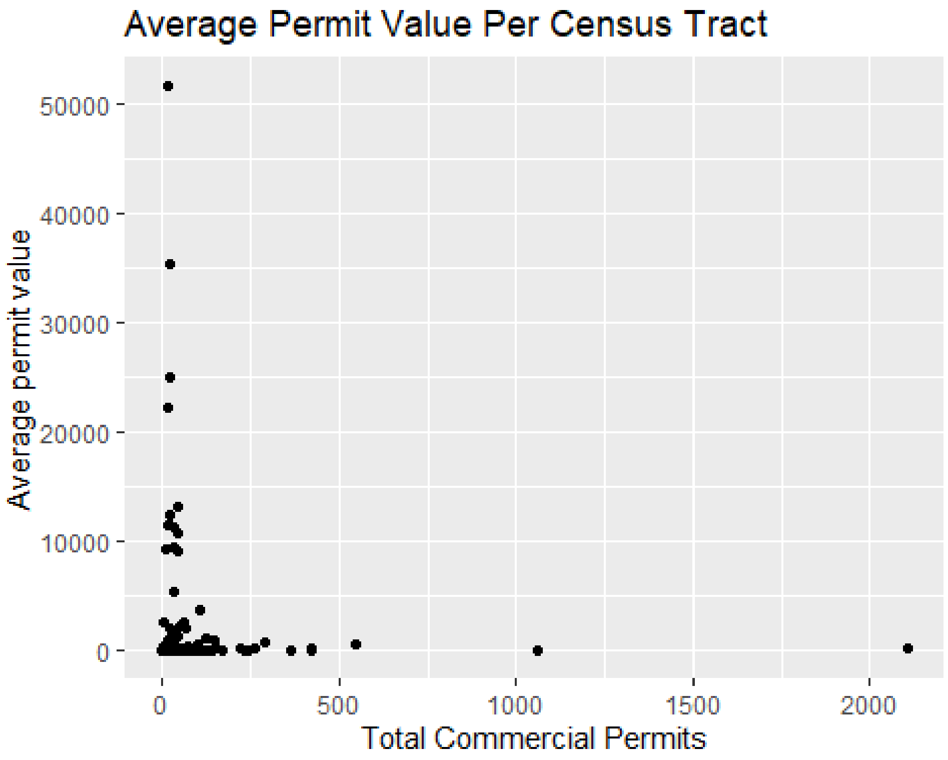
```
summary(pp_res_data$avg_permit_value)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
##     0.00    11.49    26.30   238.22    58.15 21011.21
```

```
high <- pp_comm_data %>% select(NAME, avg_permit_value) %>%
arrange(desc(avg_permit_value))

t.test(pp_comm_byCT$avg_permit_value, pp_res_byCT$avg_permit_value)

##
##  Welch Two Sample t-test
##
## data:  pp_comm_byCT$avg_permit_value and pp_res_byCT$avg_permit_value
## t = 2.869, df = 198.87, p-value = 0.004563
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   413.3814 2230.9371
## sample estimates:
## mean of x mean of y
## 1560.3777  238.2184
```

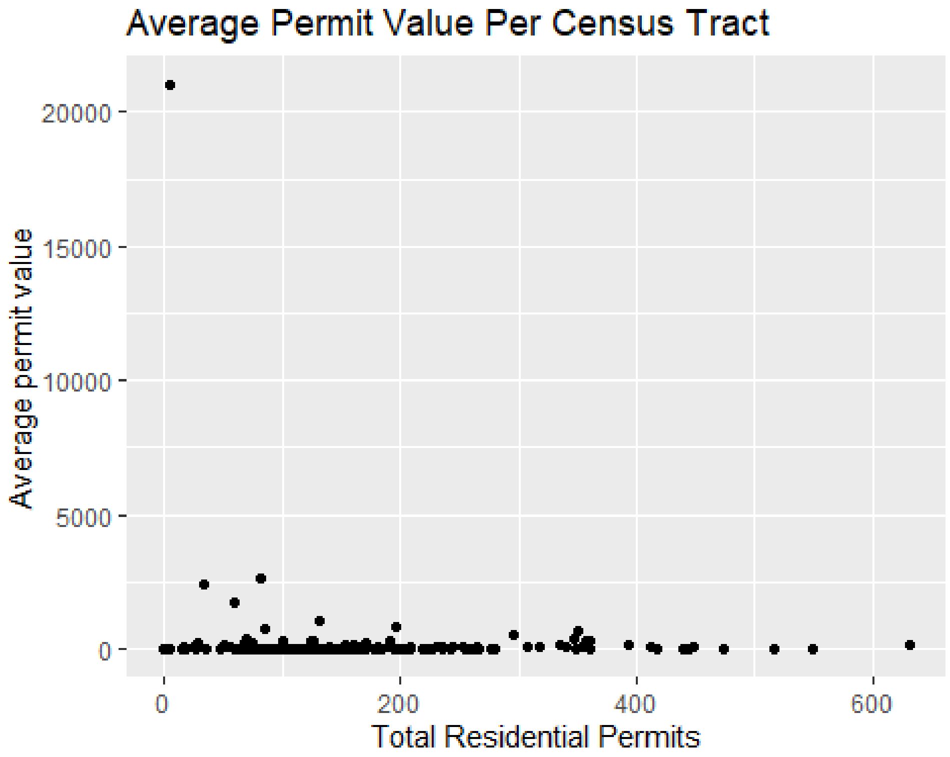
Plot permit values for comm. vs res.

```
commPlot <- ggplot(data=pp_comm_data, aes(x=total_comm_permits,
y=avg_permit_value)) +
    geom_point() + xlab("Total Commercial Permits") +
    ylab("Average permit value") + labs(title="Average Permit Value Per
Census Tract")

commPlot
```

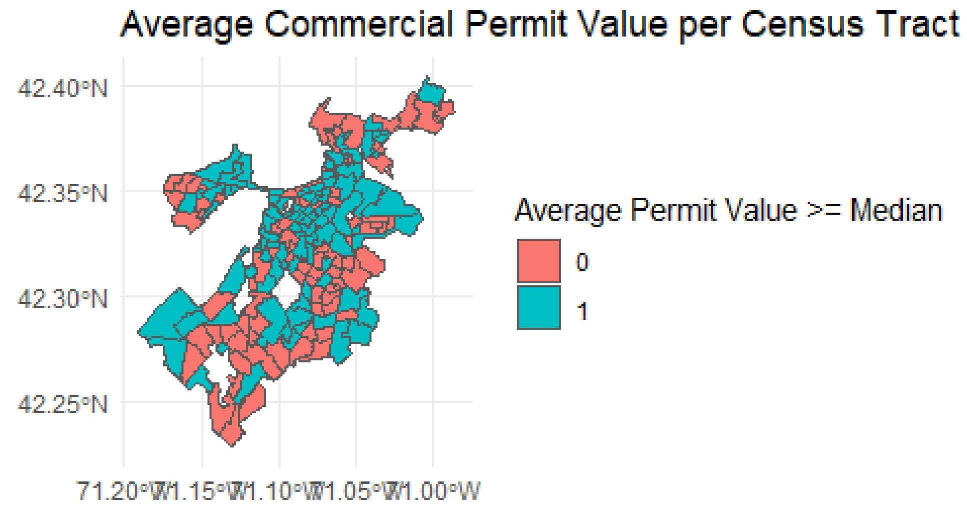


Plot permit values for comm. vs res.

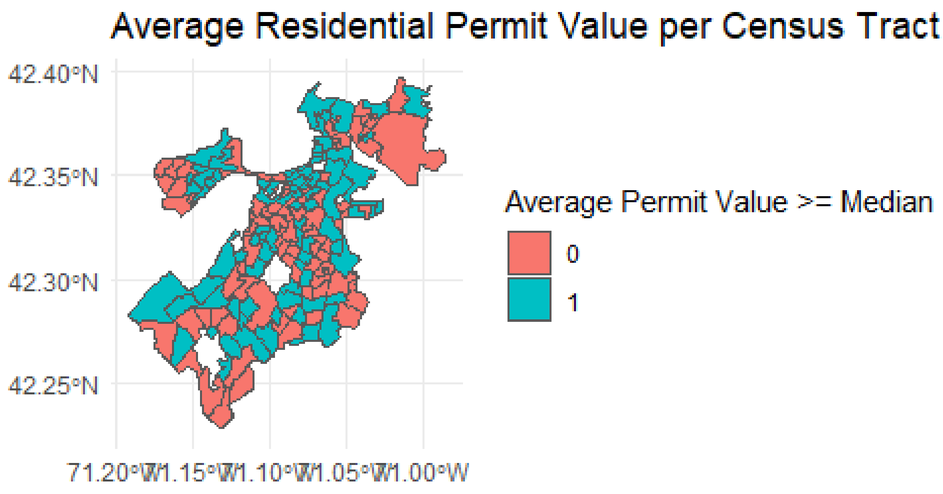
```
resPlot <- ggplot(data=pp_res_data, aes(x=total_res_permits,
y=avg_permit_value)) +
    geom_point() + xlab("Total Residential Permits") +
    ylab("Average permit value") + labs(title="Average Permit Value Per
Census Tract")

resPlot
```



```
# convert CT ID to char for compatibilty
pp_comm_data$CT_ID_10 <- as.character(pp_comm_data$CT_ID_10)
pp_res_data$CT_ID_10 <- as.character(pp_res_data$CT_ID_10)

tracts_data_comm <- right_join(tracts_shp, pp_comm_data, by="CT_ID_10")
tracts_data_res <- right_join(tracts_shp, pp_res_data, by="CT_ID_10")

# create a permit binary
tracts_data_comm$avg_permit_value_binary <-
ifelse(tracts_data_comm$avg_permit_value <
median(tracts_data_comm$avg_permit_value), 0, 1)

tracts_data_comm$avg_permit_value_binary <-
as.factor(tracts_data_comm$avg_permit_value_binary)

ggplot() +
geom_sf(data=tracts_data_comm, aes(fill = avg_permit_value_binary)) +
# scale_fill_gradient(low="lightblue", high="darkblue") +
labs(title = "Average Commercial Permit Value per Census Tract",
fill="Average Permit Value >= Median") +
theme_minimal()
```


Average Commercial Permit Value per Census Tract
42.40°N
42.35°N
42.30°N
42.25°N
71.20°W
71.15°W
71.10°W
71.05°W
71.00°W
Average Permit Value >= Median
0
1

```
# create a permit binary
tracts_data_res$avg_permit_value_binary <-
ifelse(tracts_data_res$avg_permit_value <
median(tracts_data_res$avg_permit_value), 0, 1)

tracts_data_res$avg_permit_value_binary <-
as.factor(tracts_data_res$avg_permit_value_binary)

ggplot() +
geom_sf(data=tracts_data_res, aes(fill = avg_permit_value_binary)) +
# scale_fill_gradient(low="lightblue", high="darkblue") +
labs(title = "Average Residential Permit Value per Census Tract",
fill="Average Permit Value >= Median") +
theme_minimal()
```


Average Residential Permit Value per Census Tract
42.40°N
42.35°N
42.30°N
42.25°N
71.20°W
71.15°W
71.10°W
71.05°W
71.00°W
Average Permit Value >= Median
0
1

```
comm_data_lm <- lm(avg_permit_value ~ PopDen + MedHouseIncome + bbike_docks +
CommuteLess10, data=pp_comm_data)

summary(comm_data_lm)

##
## Call:
## lm(formula = avg_permit_value ~ PopDen + MedHouseIncome + bbike_docks +
##     CommuteLess10, data = pp_comm_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -4111  -2275  -1314    -77  46891
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.733e+03  1.359e+03   2.748  0.00668 **
## PopDen          2.717e-02  2.762e-02   0.984  0.32665
## MedHouseIncome -3.175e-02  1.238e-02  -2.565  0.01122 *
## bbike_docks     2.196e+01  1.361e+01   1.614  0.10851
## CommuteLess10  -1.506e+04  7.816e+03  -1.927  0.05571 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5786 on 162 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.05927,    Adjusted R-squared:  0.03604 
## F-statistic: 2.552 on 4 and 162 DF,  p-value: 0.04113
```

```
res_data_lm <- lm(avg_permit_value ~ PopDen + MedHouseIncome + bbike_docks +
CommuteLess10, data=pp_res_data)

summary(res_data_lm)
```

```
## 
## Call:
## lm(formula = avg_permit_value ~ PopDen + MedHouseIncome + bbike_docks + 
##     CommuteLess10, data = pp_res_data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -1849.8  -355.8  -188.6    96.6 20118.6 
## 
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)  
## (Intercept)     5.769e+02  3.885e+02   1.485   0.1395  
## PopDen          2.942e-03  7.795e-03   0.377   0.7064  
## MedHouseIncome -5.989e-03  3.510e-03  -1.706   0.0899 .
## bbike_docks     7.972e+00  3.857e+00   2.067   0.0403 *
## CommuteLess10  -2.376e+03  2.214e+03  -1.073   0.2848  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1638 on 162 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.03818,    Adjusted R-squared:  0.01443 
## F-statistic: 1.608 on 4 and 162 DF,  p-value: 0.1748
```

```
cor.test(pp_comm_data$MedHouseIncome, pp_comm_data$avg_permit_value)
```

```
## 
##  Pearson's product-moment correlation
## 
## data:  pp_comm_data$MedHouseIncome and pp_comm_data$avg_permit_value
## t = -2.3124, df = 165, p-value = 0.02199
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3204165 -0.0260104
## sample estimates:
##        cor 
## -0.1771742
```

```
cor.test(pp_res_data$MedHouseIncome, pp_res_data$avg_permit_value)
```

```
## 
##  Pearson's product-moment correlation
## 
```

```
## data:  pp_res_data$MedHouseIncome and pp_res_data$avg_permit_value
## t = -1.4117, df = 165, p-value = 0.1599
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.25684668  0.04333835
## sample estimates:
##        cor
## -0.1092442
```