"BATH BEACH" "BAY RIDGE" "BEDFORD STUYVESANT"

"BENSONHURST" "BERGEN BEACH" "BOERUM HILL"

"BOROUGH PARK" "BRIGHTON BEACH" "BROOKLYN HEIGHTS"

"BROWNSVILLE" "BUSHWICK" "CANARSIE"

"CARROLL GARDENS" "CLINTON HILL" "COBBLE HILL"

"COBBLE HILL-WEST" "CONEY ISLAND" "CROWN HEIGHTS"

"CYPRESS HILLS" "DOWNTOWN-FULTON MALL" "DYKER HEIGHTS"

"EAST NEW YORK" "FLATBUSH-CENTRAL" "FLATBUSH-EAST"

"FLATBUSH-LEFFERTS GARDEN" "FLATBUSH-NORTH" "FLATLANDS"

"FORT GREENE" "GERRITSEN BEACH" "GOWANUS"

"GRAVESEND" "GREENPOINT" "KENSINGTON"

"MADISON" "MANHATTAN BEACH" "MARINE PARK"

"MIDWOOD" "MILL BASIN" "NAVY YARD"

"OCEAN HILL" "OCEAN PARKWAY-NORTH" "OCEAN PARKWAY-SOUTH"

"OLD MILL BASIN" "PARK SLOPE" "PARK SLOPE SOUTH"

"PROSPECT HEIGHTS" "RED HOOK" "SEAGATE"

"SHEEPSHEAD BAY" "SUNSET PARK" "WILLIAMSBURG-EAST"

"WILLIAMSBURG-NORTH" "WILLIAMSBURG-SOUTH" "WINDSOR TERRACE"

"WYCKOFF HEIGHTS" "SPRING CREEK" "BUSH TERMINAL"

"DOWNTOWN-FULTON FERRY" "DOWNTOWN-METROTECH" "WILLIAMSBURG-CENTRAL"

Categories:

Williamsburg

1. WILLIAMSBURG-CENTRAL
2. WILLIAMSBURG-EAST
3. WILLIAMSBURG-NORTH
4. WILLIAMSBURG-SOUTH

Park Slope

1. PARK SLOPE
2. PARK SLOPE SOUTH

Cobble Hill

1. COBBLE HILL
2. COBBLE HILL-WEST

Downtown

1. DOWNTOWN-FULTON FERRY
2. DOWNTOWN-METROTECH
3. DOWNTOWN-FULTON MALL

Beach

1. BERGEN BEACH
2. BRIGHTON BEACH
3. GERRITSEN BEACH
4. MANHATTAN BEACH
5. BATH BEACH

Mill Basin

1. MILL BASIN
2. OLD MILL BASIN

Ocean Parkway

1. OCEAN PARKWAY-NORTH
2. OCEAN PARKWAY-SOUTH

Flatbush

1. FLATBUSH-CENTRAL
2. FLATBUSH-EAST
3. FLATBUSH-LEFFERTS GARDEN
4. FLATBUSH-NORTH

Central Brooklyn

1. CROWN HEIGHTS
2. KENSINGTON
3. WINDSOR TERRACE

Eastern Brooklyn

1. BROWNSVILLE
2. CANARSIE
3. CYPRESS HILLS
4. EAST NEW YORK
5. SPRING CREEK

Northern Brooklyn

1. BEDFORD STUYVESANT
2. BUSHWICK
3. GREENPOINT
4. OCEAN HILL
5. WYCKOFF HEIGHTS

Southern Brooklyn

1. CONEY ISLAND
2. FLATLANDS
3. GRAVESEND
4. MADISON
5. MARINE PARK
6. MIDWOOD
7. SEAGATE
8. SHEEPSHEAD BAY
9. BUSH TERMINAL

Northwestern Brooklyn

1. BOERUM HILL
2. BROOKLYN HEIGHTS
3. CARROLL GARDENS
4. CLINTON HILL
5. FORT GREENE
6. GOWANUS
7. NAVY YARD
8. PROSPECT HEIGHTS
9. RED HOOK

Southwestern Brooklyn

1. BAY RIDGE
2. BENSONHURST
3. BOROUGH PARK
4. DYKER HEIGHTS
5. SUNSET PARK

<https://en.wikipedia.org/wiki/List_of_Brooklyn_neighborhoods>

Williamsburg

1. WILLIAMSBURG-CENTRAL
2. WILLIAMSBURG-EAST
3. WILLIAMSBURG-NORTH
4. WILLIAMSBURG-SOUTH

Park Slope

1. PARK SLOPE
2. PARK SLOPE SOUTH

Cobble Hill

1. COBBLE HILL
2. COBBLE HILL-WEST

Downtown

1. DOWNTOWN-FULTON FERRY
2. DOWNTOWN-METROTECH
3. DOWNTOWN-FULTON MALL
4. CARROLL GARDENS
5. NAVY YARD
6. FORT GREENE
7. BOERUM HILL

Beach

1. BERGEN BEACH
2. BRIGHTON BEACH
3. GERRITSEN BEACH
4. MANHATTAN BEACH
5. BATH BEACH

Mill Basin

1. MILL BASIN
2. OLD MILL BASIN

Ocean Parkway

1. OCEAN PARKWAY-NORTH
2. OCEAN PARKWAY-SOUTH

Flatbush

1. FLATBUSH-CENTRAL
2. FLATBUSH-EAST
3. FLATBUSH-LEFFERTS GARDEN
4. FLATBUSH-NORTH

* CROWN HEIGHTS
* KENSINGTON
* WINDSOR TERRACE
* BROWNSVILLE
* CANARSIE
* CYPRESS HILLS
* EAST NEW YORK
* SPRING CREEK
* BEDFORD STUYVESANT
* BUSHWICK
* GREENPOINT
* OCEAN HILL
* WYCKOFF HEIGHTS
* CONEY ISLAND
* FLATLANDS
* GRAVESEND
* MADISON
* MARINE PARK
* MIDWOOD
* SEAGATE
* SHEEPSHEAD BAY
* BUSH TERMINAL
* BROOKLYN HEIGHTS
* CLINTON HILL
* GOWANUS
* PROSPECT HEIGHTS
* RED HOOK
* BAY RIDGE
* BENSONHURST
* BOROUGH PARK
* DYKER HEIGHTS
* SUNSET PARK

# Final Assignment 1 - Statistical Analysis

library(dplyr)

library(tidyverse)

library(lubridate)

library(stringr)

library(ggplot2)

standard\_cols <- c('borough','neighborhood','bldclasscat','taxclasscurr','block','lot','easement','bldclasscurr','address','aptnum','zip','resunits','comunits','totunits','landsqft','grosssqft','yrbuilt','taxclasssale','bldclasssale','price','date')

data2016 <- read.csv("D://University of Chicago//Courses///ADSP 31007 IP06 - Statistical Analysis//Final Assignment//2016\_brooklyn.csv", skip = 4, header = TRUE, col.names = standard\_cols)

data2017 <- read.csv("D://University of Chicago//Courses///ADSP 31007 IP06 - Statistical Analysis//Final Assignment//2017\_brooklyn.csv", skip = 4, header = TRUE, col.names = standard\_cols)

data2018 <- read.csv("D://University of Chicago//Courses///ADSP 31007 IP06 - Statistical Analysis//Final Assignment//2018\_brooklyn.csv", skip = 4, header = TRUE, col.names = standard\_cols)

data2019 <- read.csv("D://University of Chicago//Courses///ADSP 31007 IP06 - Statistical Analysis//Final Assignment//2019\_brooklyn.csv", skip = 4, header = TRUE, col.names = standard\_cols)

data2020 <- read.csv("D://University of Chicago//Courses///ADSP 31007 IP06 - Statistical Analysis//Final Assignment//2020\_brooklyn.csv", skip = 6, header = TRUE, col.names = standard\_cols)

#Combining all the Data into one dataset

combined\_data <- rbind(data2016, data2017, data2018, data2019, data2020)

combined\_data$date <- mdy(combined\_data$date)

#Checking variables

combined\_data %>% group\_by(landsqft) %>% count()

combined\_data %>% group\_by(grosssqft) %>% count()

combined\_data %>% group\_by(yrbuilt) %>% count()

combined\_data %>% group\_by(price) %>% count()

combined\_data %>% group\_by(date) %>% count()

combined\_data %>% group\_by(comunits) %>% count()

combined\_data %>% group\_by(resunits) %>% count()

combined\_data %>% group\_by(totunits) %>% count()

#Cleaning Columns after combining data [removing white spaces, replacing "-" with NA and converting datatype to numeric]

combined\_data$neighborhood <- trimws(combined\_data$neighborhood)

combined\_data$resunits <- trimws(combined\_data$resunits)

combined\_data$resunits[combined\_data$resunits == "-"] <- NA

combined\_data$resunits <- as.numeric(combined\_data$resunits)

combined\_data$comunits <- trimws(combined\_data$comunits)

combined\_data$comunits[combined\_data$comunits == "-"] <- NA

combined\_data$comunits <- as.numeric(combined\_data$comunits)

combined\_data$totunits <- trimws(combined\_data$totunits)

combined\_data$totunits[combined\_data$totunits == "-"] <- NA

combined\_data$totunits <- as.numeric(combined\_data$totunits)

combined\_data$landsqft <- trimws(combined\_data$landsqft)

combined\_data$landsqft[combined\_data$landsqft == "-"] <- NA

combined\_data$landsqft <- as.numeric(gsub("[,]","", combined\_data$landsqft))

combined\_data$grosssqft <- trimws(combined\_data$grosssqft)

combined\_data$grosssqft[combined\_data$grosssqft == "-"] <- NA

combined\_data$grosssqft <- as.numeric(gsub("[,]","", combined\_data$grosssqft))

combined\_data$price <- trimws(combined\_data$price)

combined\_data$bldclasssale <- trimws(combined\_data$bldclasssale)

#Filtering Data to select only single family homes, Condos, Totunits = 1, Resunits = 1, grosssqft > 0 and no considering NA price values

filtered\_data <- combined\_data %>%

filter(str\_detect(bldclasssale, "[AR]"),

totunits == 1,

resunits == 1,

grosssqft > 0,

!is.na(price)

)

#Cleaning Price column to remove $ Sign and ,

filtered\_data$price[filtered\_data$price == '-'] <- NA

filtered\_data$price <- gsub("[$,]","", filtered\_data$price)

filtered\_data$price <- as.numeric(filtered\_data$price)

na\_count <- sum(is.na(filtered\_data$price))

filtered\_data <- filtered\_data[!is.na(filtered\_data$price), ]

# Counting occurrences of 0 values in the 'price' column

zero\_count <- sum(filtered\_data$price == 0, na.rm = TRUE)

filtered\_data <- filtered\_data[filtered\_data$price != 0, ]

filtered\_data$year <- year(filtered\_data$date)

filtered\_data$BuildingAgeatSale <- year(filtered\_data$date) - filtered\_data$yrbuilt

#Box plot for price

ggplot(filtered\_data, aes(y = filtered\_data$price)) +

geom\_boxplot() +

ylab("Price") +

ggtitle("Boxplot of Price")

summary(filtered\_data$price)

#Removing the outliers on extreme ends

filtered\_data <- filtered\_data %>%filter(price >= 50000 & price <= 9000000)

nrow(filtered\_data)

#unique(filtered\_data$neighborhood)

#

# filtered\_data <- filtered\_data %>%

# mutate(Neighborhood\_Category = case\_when(

# neighborhood %in% c("CROWN HEIGHTS", "KENSINGTON", "OCEAN PARKWAY-NORTH", "OCEAN PARKWAY-SOUTH", "WINDSOR TERRACE") ~ "Central Brooklyn",

# neighborhood %in% c("FLATBUSH-CENTRAL", "FLATBUSH-EAST", "FLATBUSH-LEFFERTS GARDEN", "FLATBUSH-NORTH") ~ "Flatbush",

# neighborhood %in% c("BROWNSVILLE", "CANARSIE", "CYPRESS HILLS", "EAST NEW YORK", "SPRING CREEK") ~ "Eastern Brooklyn",

# neighborhood %in% c("BEDFORD STUYVESANT", "BUSHWICK", "GREENPOINT", "OCEAN HILL", "WYCKOFF HEIGHTS") ~ "Northern Brooklyn",

# neighborhood %in% c("WILLIAMSBURG-CENTRAL", "WILLIAMSBURG-EAST", "WILLIAMSBURG-NORTH", "WILLIAMSBURG-SOUTH") ~ "Williamsburg",

# neighborhood %in% c("BOERUM HILL", "BROOKLYN HEIGHTS", "CARROLL GARDENS", "CLINTON HILL", "FORT GREENE", "GOWANUS", "NAVY YARD", "PROSPECT HEIGHTS", "RED HOOK") ~ "Northwestern Brooklyn",

# neighborhood %in% c("PARK SLOPE", "PARK SLOPE SOUTH") ~ "Park Slope",

# neighborhood %in% c("COBBLE HILL", "COBBLE HILL-WEST") ~ "Cobble Hill",

# neighborhood %in% c("DOWNTOWN-FULTON FERRY", "DOWNTOWN-METROTECH", "DOWNTOWN-FULTON MALL") ~ "Downtown",

# neighborhood %in% c("CONEY ISLAND", "FLATLANDS", "GRAVESEND", "MADISON", "MARINE PARK", "MIDWOOD", "SEAGATE", "SHEEPSHEAD BAY", "BUSH TERMINAL") ~ "Southern Brooklyn",

# neighborhood %in% c("MILL BASIN", "OLD MILL BASIN") ~ "Mill Basin",

# neighborhood %in% c("BAY RIDGE", "BENSONHURST", "BOROUGH PARK", "DYKER HEIGHTS", "SUNSET PARK") ~ "Southwestern Brooklyn",

# neighborhood %in% c("BERGEN BEACH", "BRIGHTON BEACH", "GERRITSEN BEACH", "MANHATTAN BEACH", "BATH BEACH") ~ "Beach",

# TRUE ~ "Other"

# ))

#Filtering neighborhood based on similarity and proximity

filtered\_data <- filtered\_data %>%

mutate(Neighborhood\_Category = case\_when(

neighborhood %in% c("WILLIAMSBURG-CENTRAL", "WILLIAMSBURG-EAST", "WILLIAMSBURG-NORTH", "WILLIAMSBURG-SOUTH") ~ "Williamsburg",

neighborhood %in% c("PARK SLOPE", "PARK SLOPE SOUTH") ~ "Park Slope",

neighborhood %in% c("COBBLE HILL", "COBBLE HILL-WEST") ~ "Cobble Hill",

neighborhood %in% c("DOWNTOWN-FULTON FERRY", "DOWNTOWN-METROTECH", "DOWNTOWN-FULTON MALL", "CARROLL GARDENS","NAVY YARD","FORT GREENE","BOERUM HILL") ~ "Downtown",

neighborhood %in% c("BERGEN BEACH", "BRIGHTON BEACH", "GERRITSEN BEACH", "MANHATTAN BEACH", "BATH BEACH") ~ "Beach",

neighborhood %in% c("MILL BASIN", "OLD MILL BASIN") ~ "Mill Basin",

neighborhood %in% c("OCEAN PARKWAY-NORTH", "OCEAN PARKWAY-SOUTH") ~ "Ocean Parkway",

neighborhood %in% c("CROWN HEIGHTS") ~ "CROWN HEIGHTS",

neighborhood %in% c("KENSINGTON") ~ "KENSINGTON",

neighborhood %in% c("WINDSOR TERRACE") ~ "WINDSOR TERRACE",

neighborhood %in% c("BROWNSVILLE") ~ "BROWNSVILLE",

neighborhood %in% c("CANARSIE") ~ "CANARSIE",

neighborhood %in% c("CYPRESS HILLS") ~ "CYPRESS HILLS",

neighborhood %in% c("EAST NEW YORK") ~ "EAST NEW YORK",

neighborhood %in% c("SPRING CREEK") ~ "SPRING CREEK",

neighborhood %in% c("BEDFORD STUYVESANT") ~ "BEDFORD STUYVESANT",

neighborhood %in% c("BUSHWICK") ~ "BUSHWICK",

neighborhood %in% c("GREENPOINT") ~ "GREENPOINT",

neighborhood %in% c("OCEAN HILL") ~ "OCEAN HILL",

neighborhood %in% c("WYCKOFF HEIGHTS") ~ "WYCKOFF HEIGHTS",

neighborhood %in% c("CONEY ISLAND") ~ "CONEY ISLAND",

neighborhood %in% c("FLATLANDS") ~ "FLATLANDS",

neighborhood %in% c("GRAVESEND") ~ "GRAVESEND",

neighborhood %in% c("MADISON") ~ "MADISON",

neighborhood %in% c("MARINE PARK") ~ "MARINE PARK",

neighborhood %in% c("MIDWOOD") ~ "MIDWOOD",

neighborhood %in% c("SEAGATE") ~ "SEAGATE",

neighborhood %in% c("SHEEPSHEAD BAY") ~ "SHEEPSHEAD BAY",

neighborhood %in% c("BUSH TERMINAL") ~ "BUSH TERMINAL",

neighborhood %in% c("BROOKLYN HEIGHTS") ~ "BROOKLYN HEIGHTS",

neighborhood %in% c("CLINTON HILL") ~ "CLINTON HILL",

neighborhood %in% c("GOWANUS") ~ "GOWANUS",

neighborhood %in% c("PROSPECT HEIGHTS") ~ "PROSPECT HEIGHTS",

neighborhood %in% c("RED HOOK") ~ "RED HOOK",

neighborhood %in% c("BAY RIDGE") ~ "BAY RIDGE",

TRUE ~ "Other"

))

# Model Execution

model <- lm(price ~ (grosssqft-landsqft) + (block\*lot) + BuildingAgeatSale + Neighborhood\_Category, data = filtered\_data)

summary(model)

sqrt(mean(model$residuals^2))

###########################

write.csv(df\_model, "D://University of Chicago//Courses//ADSP 31007 IP06 - Statistical Analysis//Final Assignment", row.names=FALSE)

<https://www.cityneighborhoods.nyc/bergen-beach>