DATA608 HW2

Bin Lin

2017-9-23

```
#install.packages("devtools")

devtools::install_github("hadley/bigvis")
```

```
Skipping install of 'bigvis' from a github remote, the SHA1 (9cce2405) has not changed since las
t install.
  Use `force = TRUE` to force installation
```

```
suppressWarnings(library(ggplot2))
suppressWarnings(library(dplyr))
```

```

Attaching package: 'dplyr'
```

```
The following objects are masked from 'package:stats':

    filter, lag
```

```
The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
suppressWarnings(library(bigvis))
```

```
Loading required package: Rcpp
```

```

Attaching package: 'bigvis'
```

```
The following object is masked from 'package:stats':

    smooth
```

1. After a few building collapses, the City of New York is going to begin investigating older buildings for safety. However, the city has a limited number of inspectors, and wants to find a 'cut-off' date before most city buildings were constructed. Build a graph to help the city determine when most buildings were constructed. Is there anything in the results that causes you to question the accuracy of the data? (note: only look at buildings built since 1850)

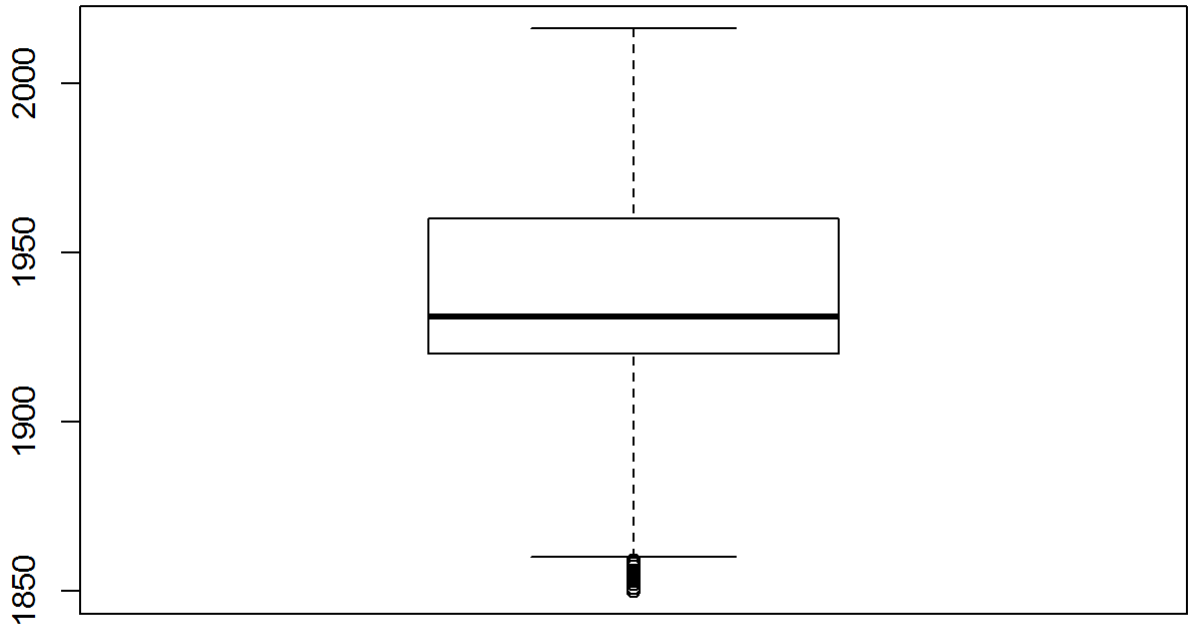
I think the data is zero inflated because there are more than 44000 data entries have the YearBuilt of zero. This really affects the validity of the data, since when I try to create graphs, I would have to eliminate those data entries.

```
BK <- read.csv("BORO_zip_files_csv/BK.csv")
BX <- read.csv("BORO_zip_files_csv/BX.csv")
MN <- read.csv("BORO_zip_files_csv/MN.csv")
QN <- read.csv("BORO_zip_files_csv/QN.csv")
SI <- read.csv("BORO_zip_files_csv/SI.csv")

all_PLUTO_data <- rbind(BK, BX, MN, QN, SI)
sub1 <- all_PLUTO_data %>%
  filter(YearBuilt >= 1850 & YearBuilt <= 2017)
```

```
## Warning: package 'bindrcpp' was built under R version 3.3.3
```

```
#The boxplot shows around 50% of houses were built between 1920 and 1960s with a peak around 192
0 to 1930, according to the range of first quantile.
boxplot(sub1$YearBuilt)
```




```
quantile(sub1$YearBuilt, 0.25)
```

```
##  25%
## 1920
```

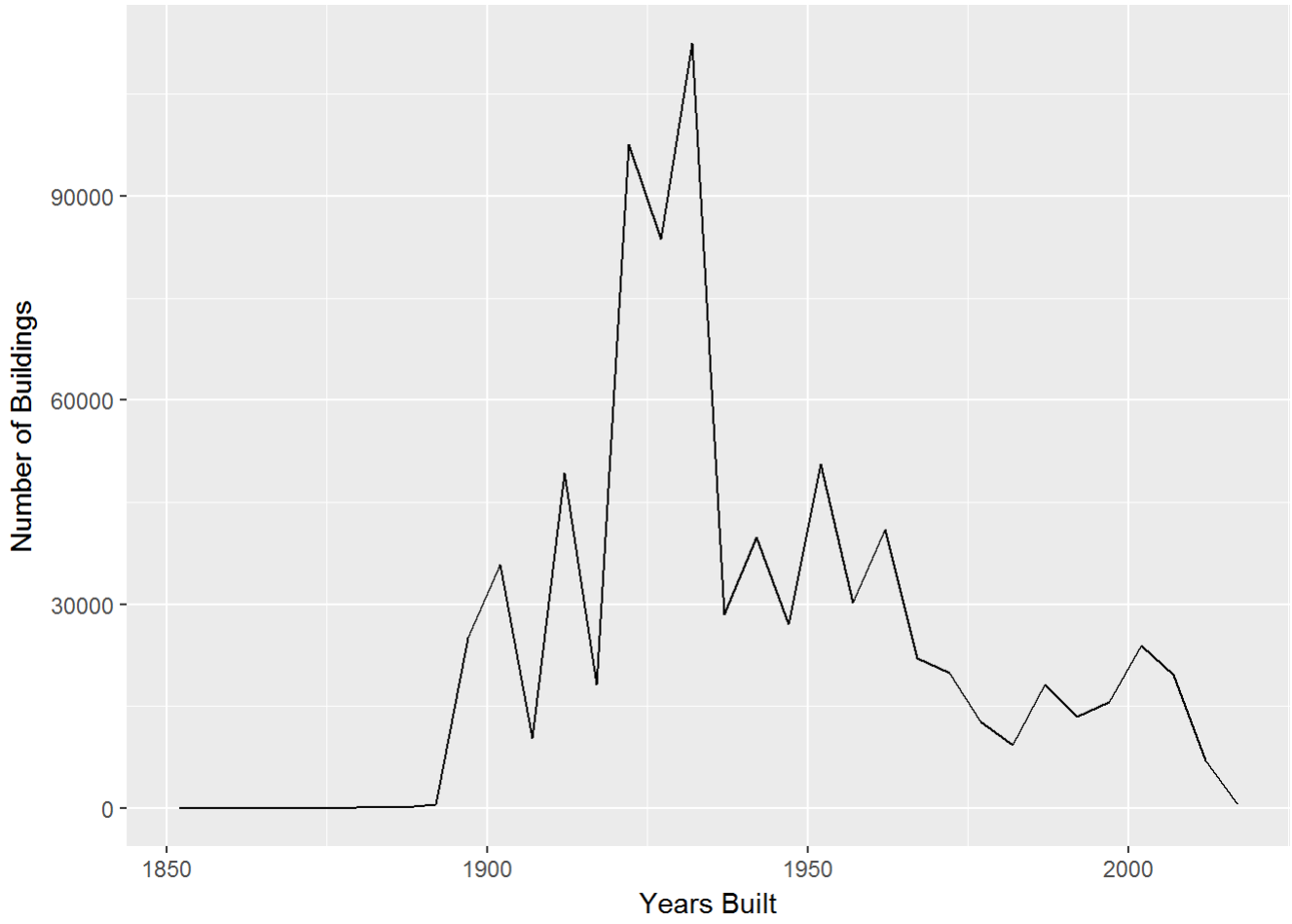
```
quantile(sub1$YearBuilt, 0.5)
```

```
##  50%
## 1931
```

```
#The condensed data based on count shows a peak around 1920 and 1930 which is pretty much the sa
me as what is shown in the boxplot.
a <- condense(bin(sub1$YearBuilt, 5))
```

```
## Summarising with count
```

```
autoplot(a) + labs(x = "Years Built", y = "Number of Buildings")
```




```
ggsave("Figure1.jpg")
```

```
## Saving 7 x 5 in image
```

2. The city is particularly worried about buildings that were unusually tall when they were built, since best-practices for safety hadn't yet been determined. Create a graph that shows how many buildings of a certain number of floors were built in each year (note: you may want to use a log scale for the number of buildings).

It should be clear when 20-story buildings, 30-story buildings, and 40-story buildings were first built in large numbers.

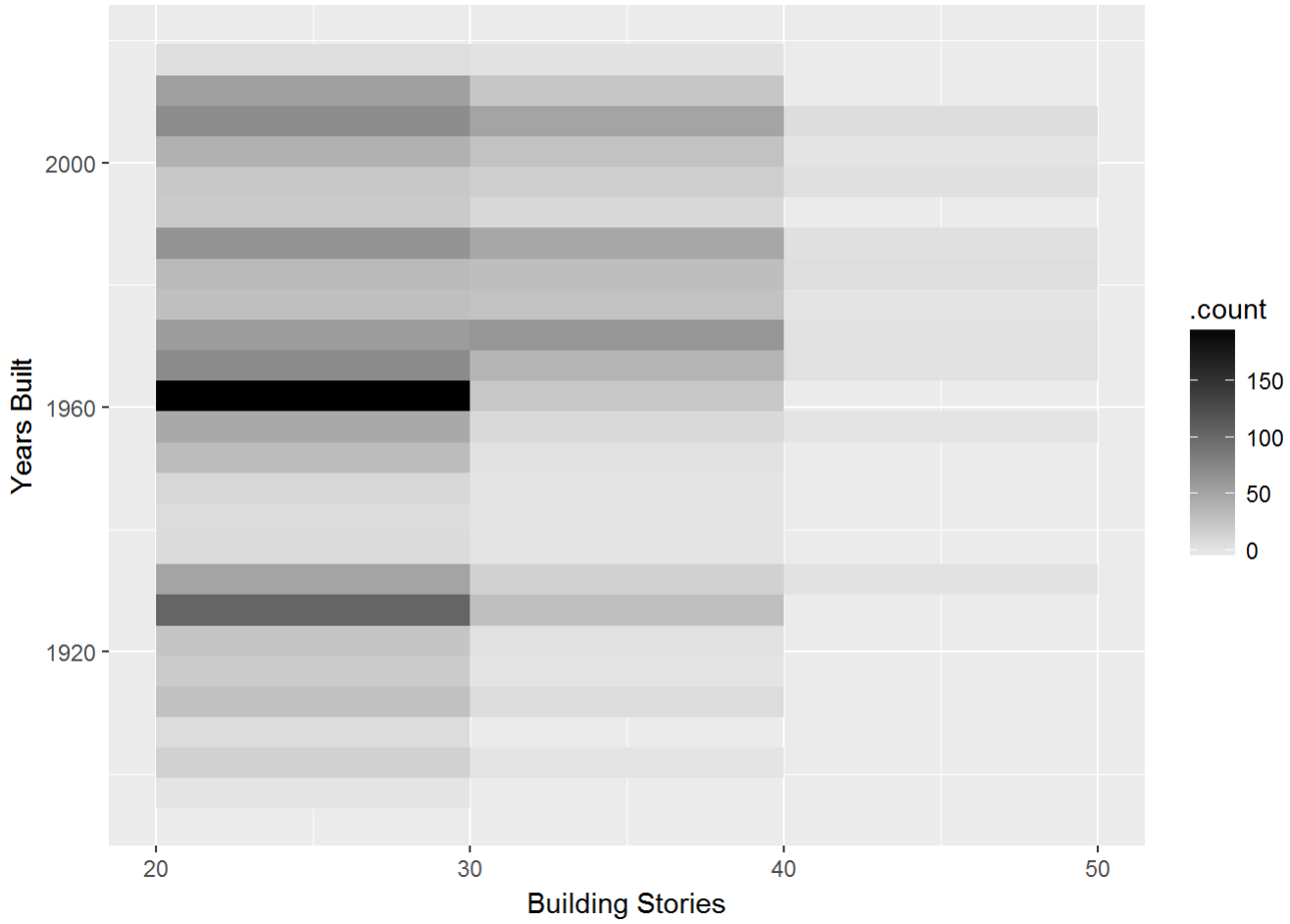
```
sub2<- sub1%>%
  filter(NumFloors >= 20, NumFloors <= 40)%>%
  select(YearBuilt, NumFloors)


#The condensed data showed most tall buildings were built during 1960s. During 1930s, there is a
 peak but it quickly faded and after 2000, US again start build tall buildings.

b <- condense(bin(sub2$NumFloors, 10), bin(sub2$YearBuilt, 5))
```

```
## Summarising with count
```

```
autoplot(b) + labs(x = "Building Stories", y = "Years Built")
```




```
ggsave("Figure2.jpg")
```

```
## Saving 7 x 5 in image
```

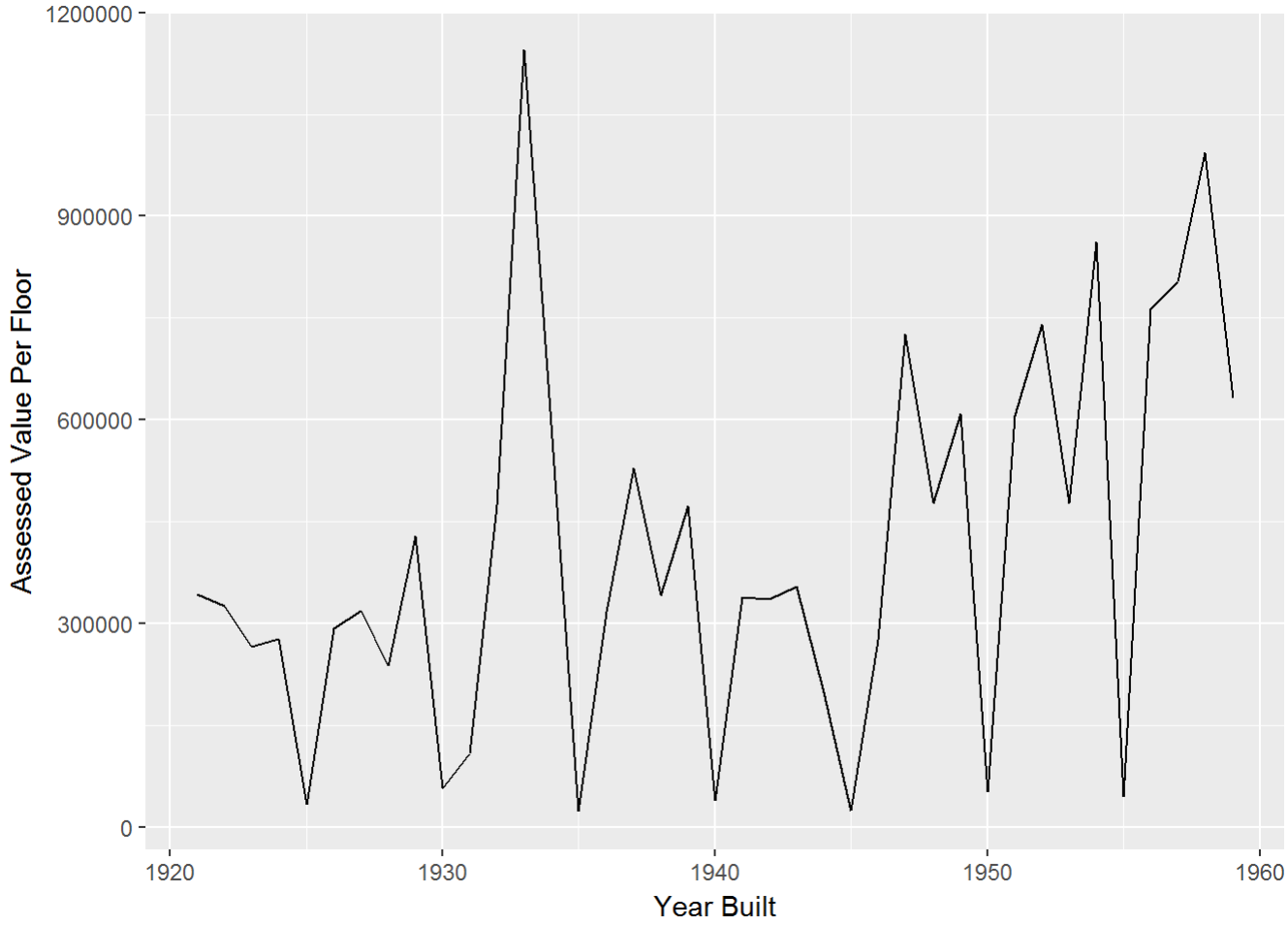
3. Your boss suspects that buildings constructed during the US's involvement in World War II (1941-1945) are more poorly constructed than those before and after the way due to the high cost of materials during those

years. She thinks that, if you calculate assessed value per floor, you will see lower values for buildings at that time vs before or after. Construct a chart/graph to see if she's right.

```
sub3 <- all_PLUTO_data %>%
  filter(NumFloors > 0, AssessLand > 0, AssessTot > 0, YearBuilt > 1920, YearBuilt < 1960)%>%
  select(YearBuilt, NumFloors, AssessLand, AssessTot)%>%
  group_by(YearBuilt)%>%
  mutate(AssessValuePerFloor = (sum(AssessTot) - sum(AssessLand)) / sum(NumFloors))

#The following line chart clearly shows during WWII, the assess value per floor is relatively lo
w compare to those houses that were built later. However, if we compare the value against those
 that were built before WWII, the differences were not that obvious. This tells us that the war
 is probably not the reason why we have lower assessed value per floor for those buildings, but
 The Great Depression was the main reason, because The Great Depression lasted longer compare to
 the wartime.

ggplot() + geom_line(data = sub3, aes(x = YearBuilt, y = AssessValuePerFloor)) + labs(x = "Year
 Built", y = "Assessed Value Per Floor")
```




```
ggsave("Figure3.jpg")
```

```
## Saving 7 x 5 in image
```