Data608_Hw2

Jason Joseph

February 19, 2017

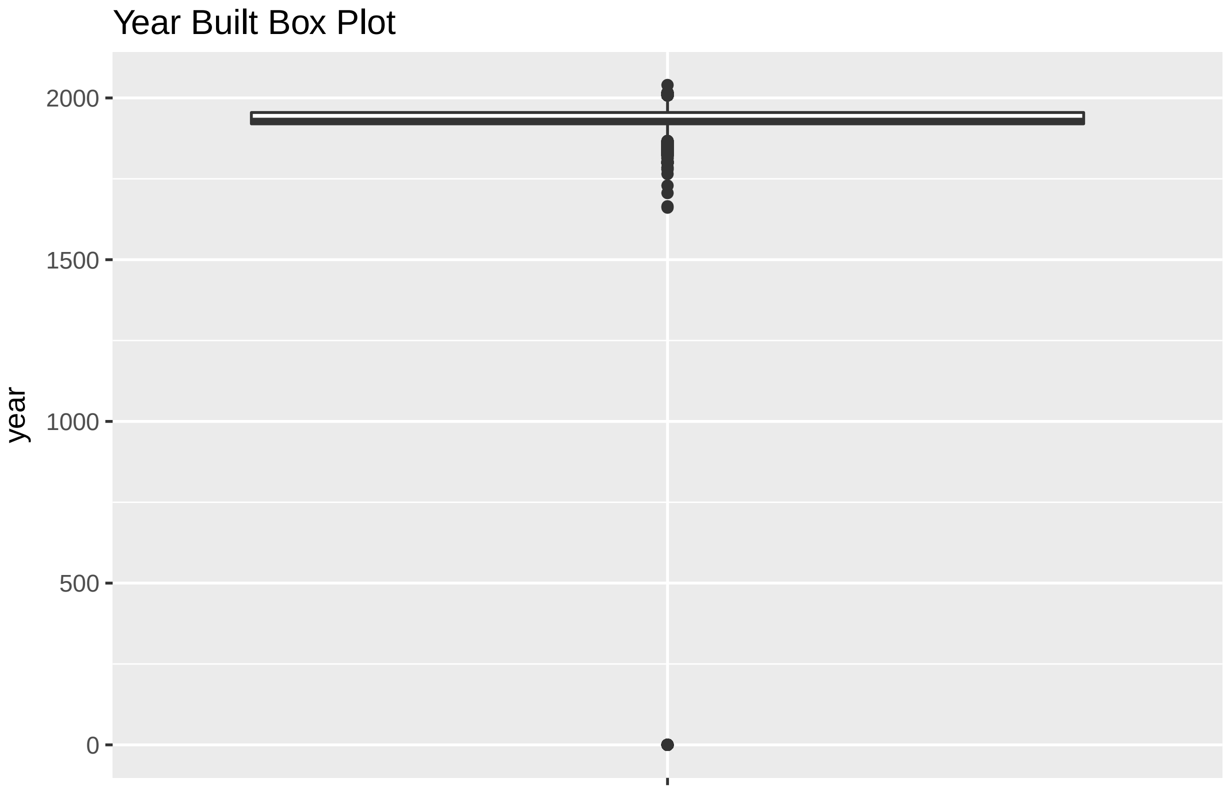
```r
# Load Data.
data <- read.csv("combined.csv")
```

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)

```r
# Closer look at Year Built
summary(data$YearBuilt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0    1920    1930    1841    1955    2040
```

```r
yearBuild.df <- data.frame( year = data$YearBuilt, cat = " ")
ggplot(yearBuild.df, aes(cat, year)) + geom_boxplot() + labs(title = "Year Built Box Plot") + xlab("")
```



The Summary data shows that the Year built range is between 0 to 2040. 0 and 2040 aren’t valid years so the data for those years will be omitted.

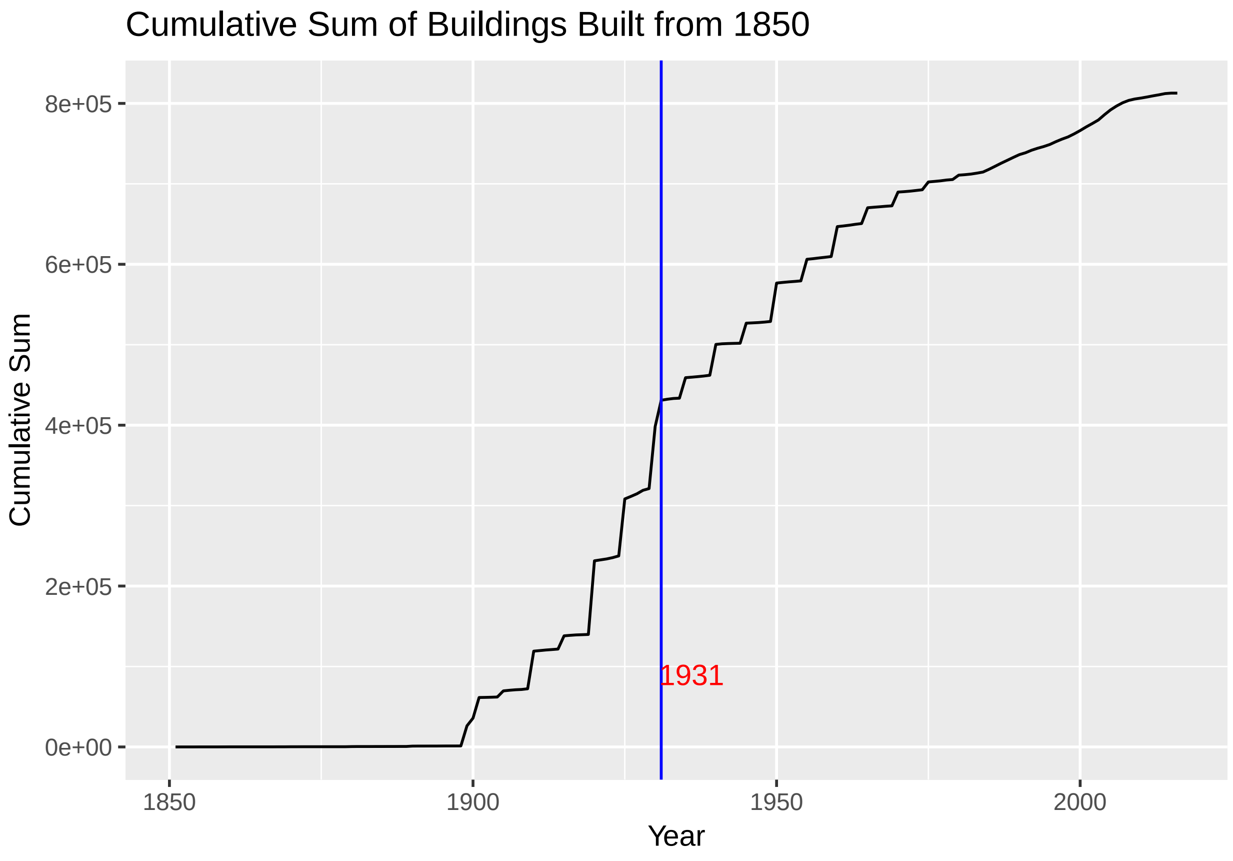
```
# remove data with Year Built 0 and 2040 && filter for buildings
# built adter 1850 and ensure there is a building
yearBuild.df.filtered <- data %>% group_by(YearBuilt) %>% filter(YearBuilt > 0, YearBuilt < 2040, YearBu
```

To find the “cut-off”" date before most city were constructed. I will find the year when the number of buildings built surpassed half the total amount of buildings built.

```
# finding half of buildings built
halfPoint <- sum(yearBuild.df.filtered$total) / 2
# find cumulative sum
yearBuild.df.filtered$cumFreq <-  cumsum(yearBuild.df.filtered$total)
# find Half built Year
halfPointYear <- yearBuild.df.filtered %>% filter(cumFreq > halfPoint) %>% slice(1) %>% select(YearBuil
halfPointYear <- halfPointYear[[1]]
halfPointYear
```

```
## [1] 1931
```

```
# graph of buildins built since 1850
ggplot(data=yearBuild.df.filtered, aes(x=YearBuilt, y=cumFreq, group =1)) +
  geom_line() +
  labs(title="Cumulative Sum of Buildings Built from 1850", x = "Year", y= "Cumulative Sum") +
  geom_vline(aes(xintercept = halfPointYear), colour="blue") +
  annotate("text", x = 1936, y = 90000, label = halfPointYear, colour="red")
```

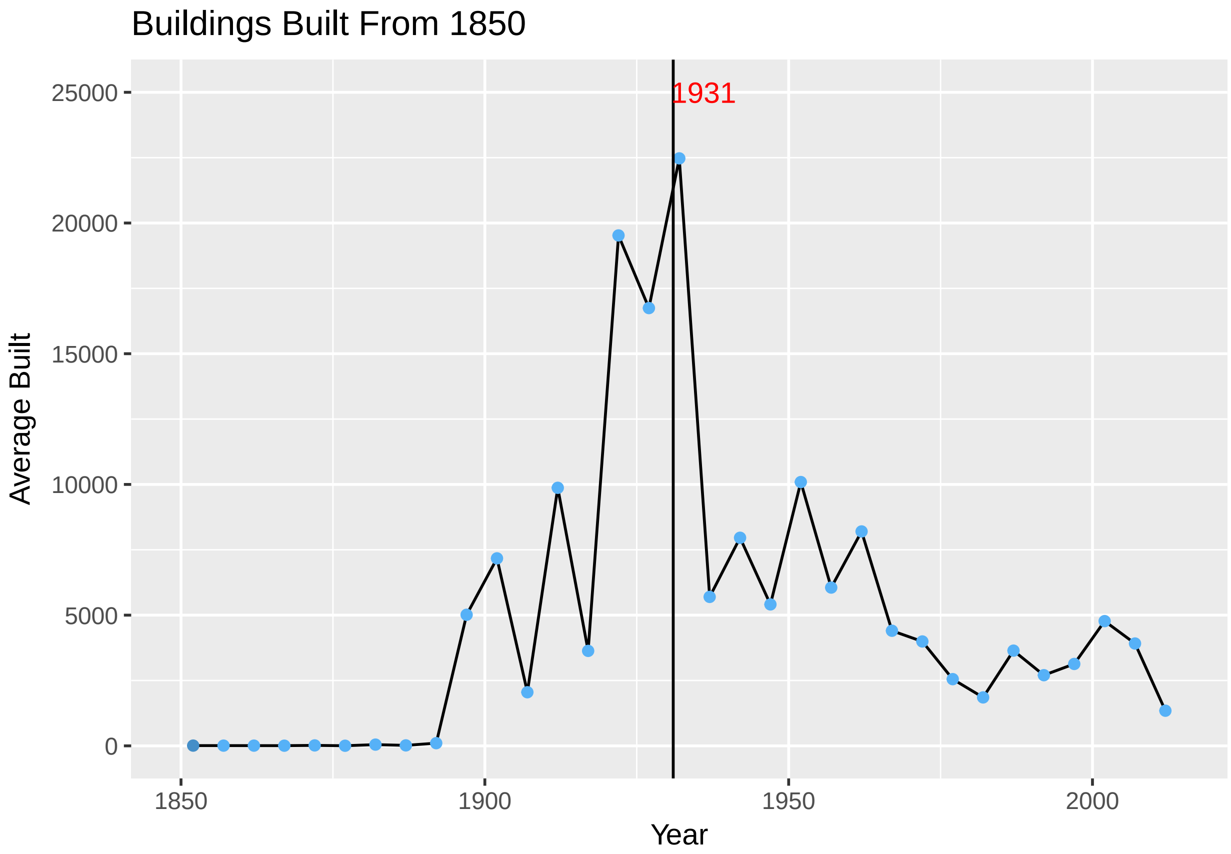


```
ggsave("q1.1.png", width = 8, height=7)
```

```
# Plotting 5 year averages
yr <-  with(yearBuild.df.filtered, condense(bin(YearBuilt, 5), z=total))
```

```
## Summarising with mean
```

```
autoplot(yr) + xlim(1850, 2014) + geom_vline(aes(xintercept = halfPointYear)) +
  labs(title="Buildings Built From 1850",  x = "Year", y= "Average Built")+
  annotate("text", x = 1936, y = 25000, label = halfPointYear, color="red") +
  theme(legend.position="none")
```



```
ggsave("q1.2.png", width = 8, height=7)
```

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.

```
# create dataframe of number of floors built in each year.

buildingFloors <- data %>% filter(YearBuilt > 0, YearBuilt < 2040, YearBuilt > 1850, NumFloors !=0, Num
  select(NumFloors,YearBuilt) %>% group_by(NumFloors,YearBuilt) %>% summarise(total = n())
```

Creating rounding column that will round NumFloors to the nearest tens place. Floors less than 10 will be rounded to 0

```
buildingFloors$roundedFloors <- round(buildingFloors$NumFloors,digits = -1)
head(buildingFloors)
```

```
## Source: local data frame [6 x 4]
## Groups: NumFloors [2]
##
##   NumFloors YearBuilt total roundedFloors
##       <dbl>     <int> <int>         <dbl>
## 1       0.5      1920     1             0
## 2       1.0      1855     2             0
## 3       1.0      1860     1             0
## 4       1.0      1866     1             0
## 5       1.0      1870     1             0
## 6       1.0      1874     1             0
```

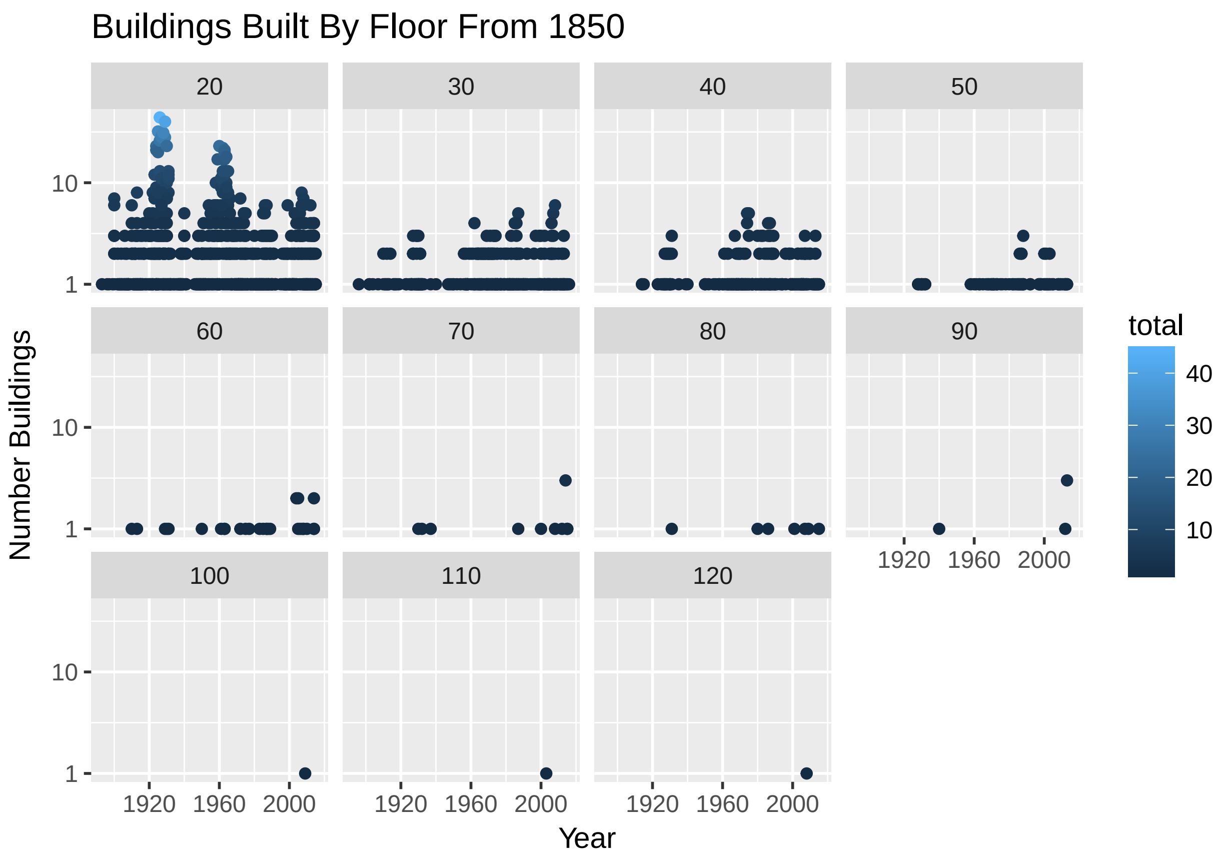
Removing records what have rounded floors above 10 and count per year that is more than zero

```
buildingFloorsFinal <- buildingFloors %>% filter(roundedFloors > 10,total > 0)
head(buildingFloorsFinal)
```

```
## Source: local data frame [6 x 4]
## Groups: NumFloors [1]
##
##   NumFloors YearBuilt total roundedFloors
##       <dbl>     <int> <int>         <dbl>
## 1        15      1900     7            20
## 2        15      1904     1            20
## 3        15      1905     1            20
## 4        15      1906     3            20
## 5        15      1907     1            20
## 6        15      1908     1            20
```

Create Plot

```
ggplot(buildingFloorsFinal, aes(x=YearBuilt, y=total, color=total))+
   geom_point() + scale_y_log10() + scale_color_continuous() +
  facet_wrap(~ roundedFloors)+
  labs(title="Buildings Built By Floor From 1850",  x = "Year", y= "Number Buildings")
```



```r
ggsave("q2.png", width = 8, height=7)
```

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.

```r
# calculating the average assessed value per floor for each year
# Since the war period spans 5 years, I will limit my post and Pre War data rage to 5 years before and

yearBuildVal <- data %>% group_by(YearBuilt) %>% filter(YearBuilt > 0, YearBuilt < 2040, YearBuilt > 18
summarise(total = n(), assedVal = mean(sum(AssessTot - AssessLand)/sum(NumFloors)))
yearBuildVal <- yearBuildVal %>% filter(YearBuilt >= 1936, YearBuilt <= 1951)
head(yearBuildVal)
```

```
## # A tibble: 6 × 3
##   YearBuilt total  assedVal
##       <int> <int>     <dbl>
## 1      1936   641 316940.44
## 2      1937   666 529239.86
## 3      1938   763 340981.64
## 4      1939   966 473216.64
## 5      1940 38357  38003.55
## 6      1941   745 337905.82
```

Adding World War II category Information

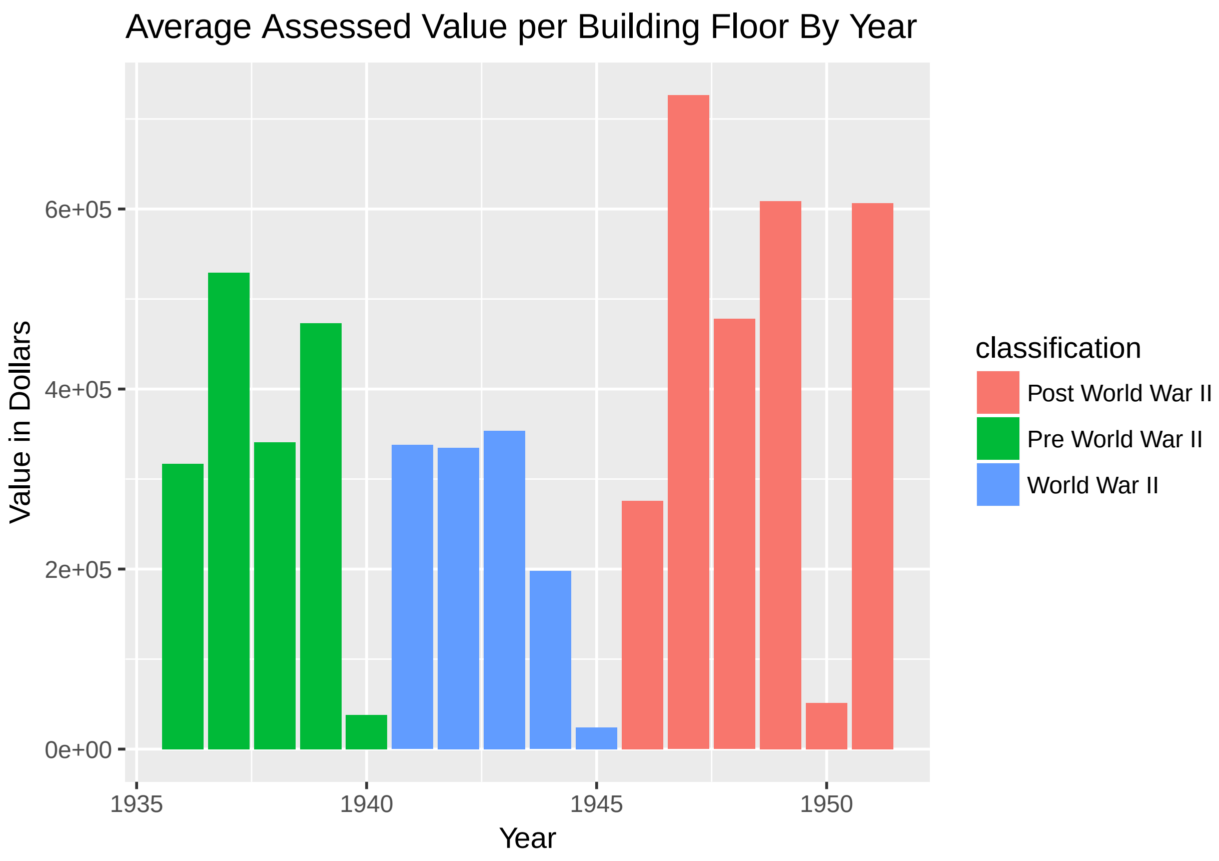
```
yearClassification <- function(year)
{
  if(year < 1941)
  {
    return ("Pre World War II")
  }
  else if(year >= 1941 && year <= 1945)
  {
    return ("World War II")
  }
  else if(year > 1945)
  {
    return ("Post World War II")
  }

}

yearBuildVal$classification <- mapply(yearClassification,yearBuildVal$YearBuilt)
head(yearBuildVal)
```

```
## # A tibble: 6 × 4
##   YearBuilt total  assedVal   classification
##       <int> <int>     <dbl>            <chr>
## 1      1936   641 316940.44 Pre World War II
## 2      1937   666 529239.86 Pre World War II
## 3      1938   763 340981.64 Pre World War II
## 4      1939   966 473216.64 Pre World War II
## 5      1940 38357  38003.55 Pre World War II
## 6      1941   745 337905.82     World War II
```

```
#plotting Data
 ggplot(yearBuildVal, aes(x=YearBuilt, y=assedVal, fill=classification)) + geom_bar(stat="identity")+
  labs(title="Average Assessed Value per Building Floor By Year", x= "Year", y= "Value in Dollars")
```



```r
ggsave("q3.png", width = 8, height=7)
```