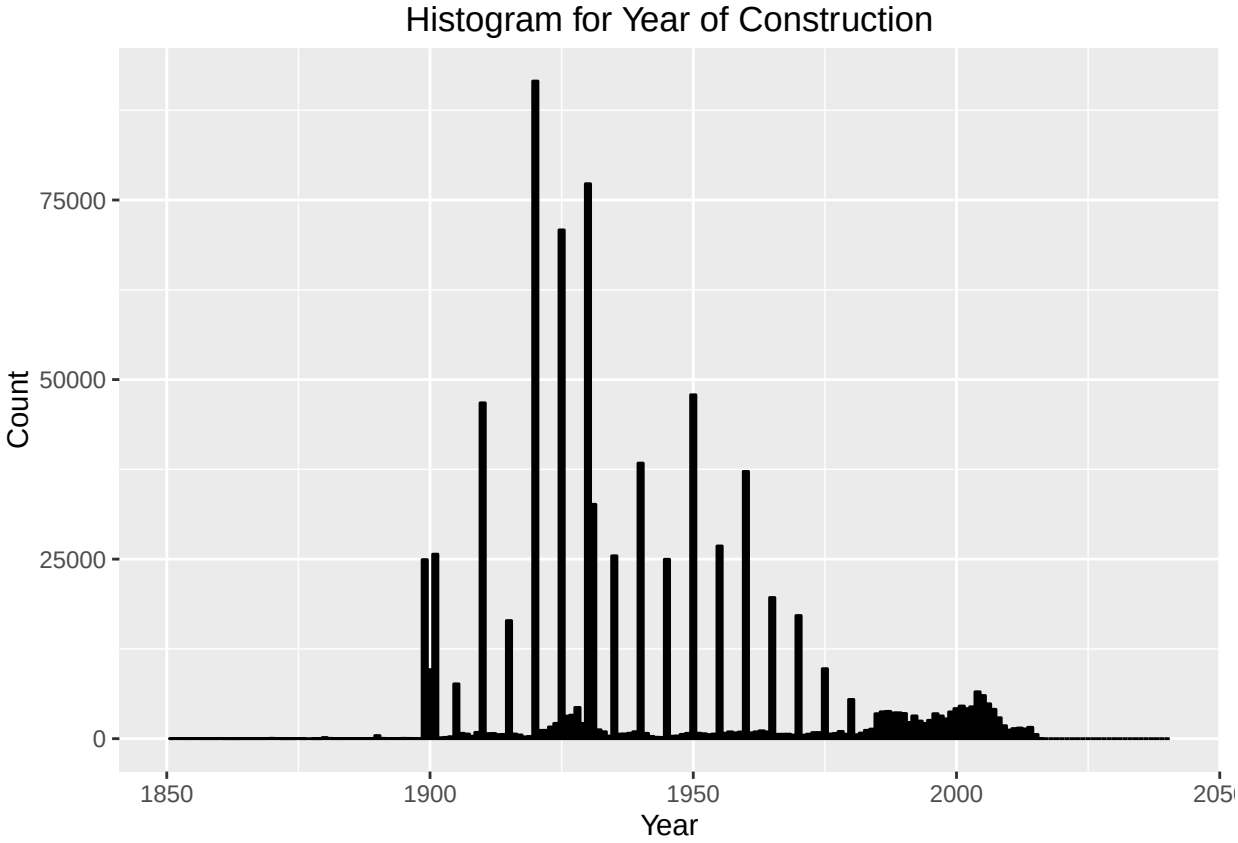
Homework 2

Aadi Kalloo

February 26, 2017

1.

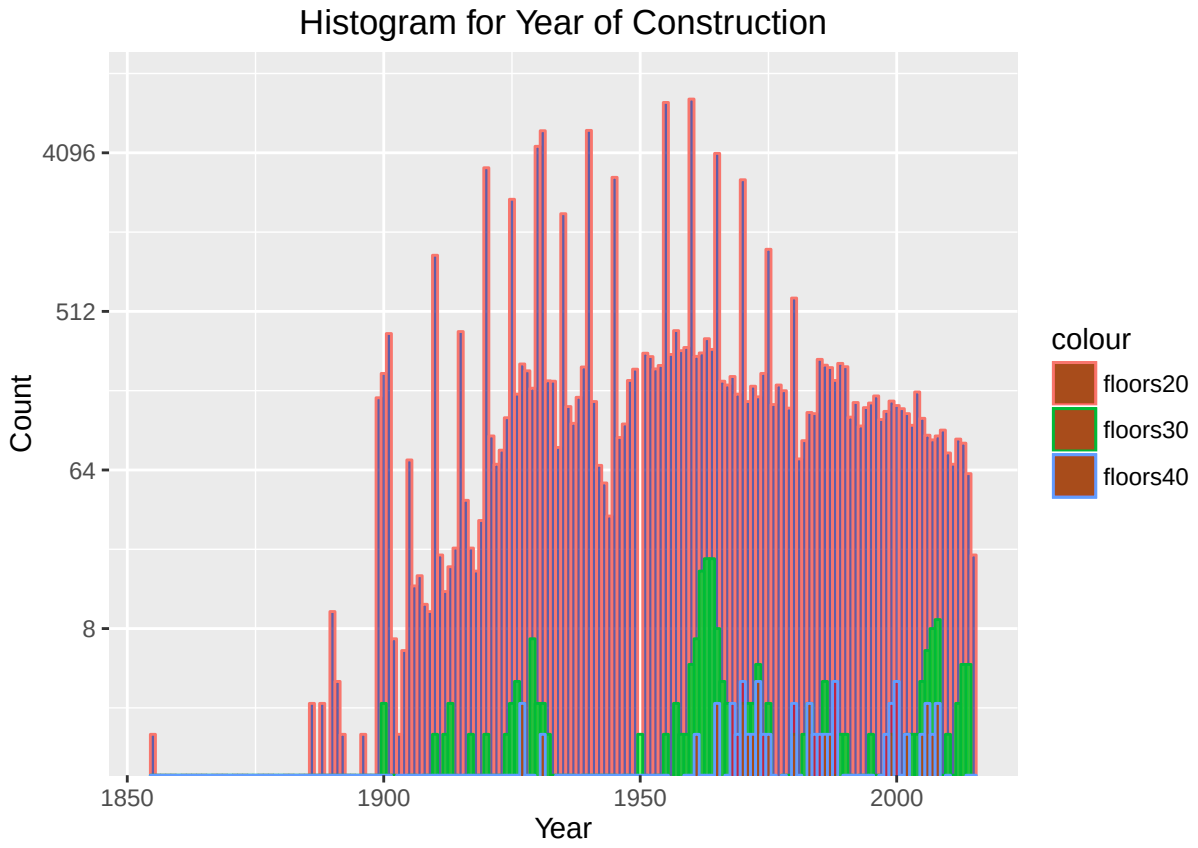
```
year_built_valid = subset(all_pluto_data, all_pluto_data$YearBuilt!=0 & all_pluto_data$YearBuilt>1850)
#hist(year_built_valid$YearBuilt, breaks=length(unique(year_built_valid$YearBuilt)))

ggplot(data=year_built_valid, aes(year_built_valid$YearBuilt)) +
  geom_histogram(binwidth=1, colour="black", fill="black") +
  labs(title="Histogram for Year of Construction") +
  labs(x="Year", y="Count")
```



The data here seems to be "clustered". It seems to me that it was easier when recording construction dates to round to the nearest 5- or 10-year interval. If this is indeed the case, the data is obviously not very accurate but I sitll think it provides a good representation of the distribution and big picture.

2.

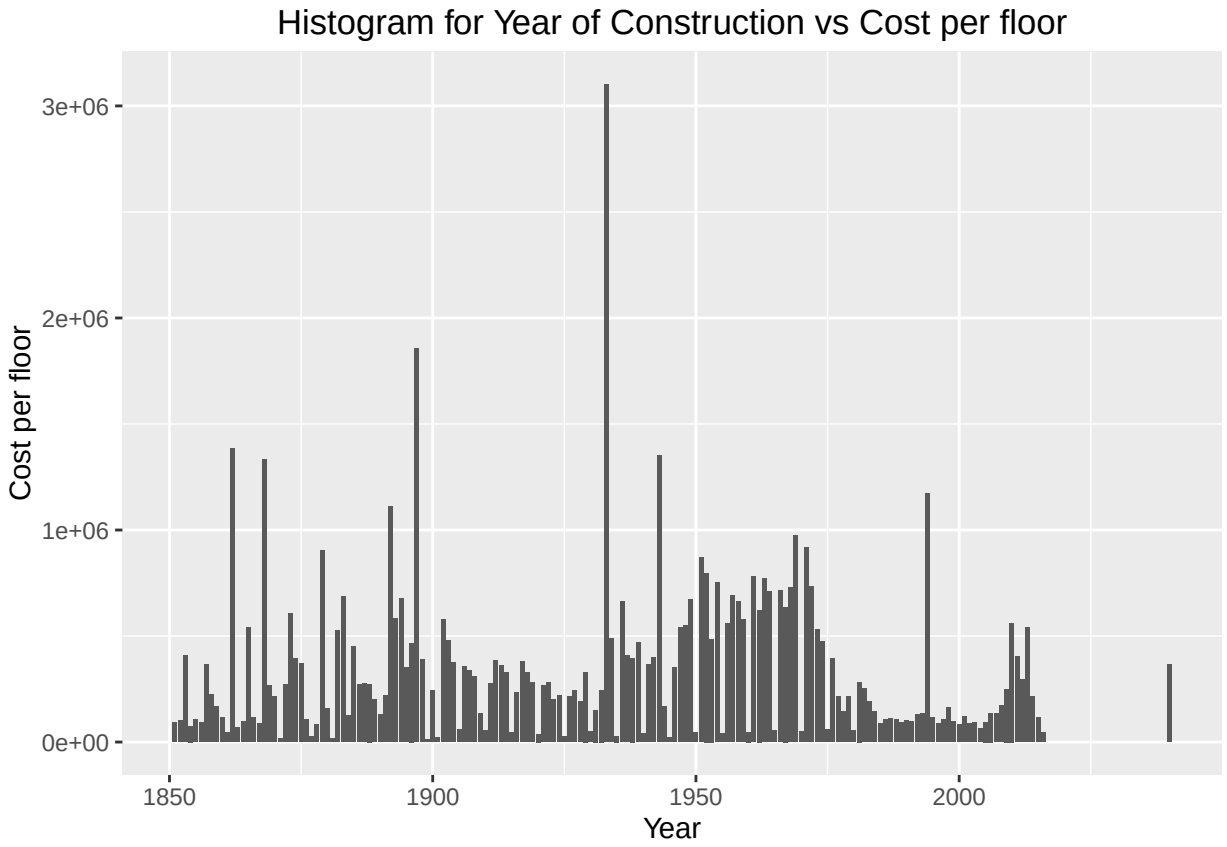
```
num_floors_valid = year_built_valid #subset(all_pluto_data, all_pluto_data$NumFloors!=0)
floors_20 = subset(num_floors_valid, num_floors_valid$NumFloors %in% c(1, 20))
floors_20$floorgroup = 20
floors_30 = subset(num_floors_valid, num_floors_valid$NumFloors %in% c(21, 30))
floors_30$floorgroup = 30
floors_40 = subset(num_floors_valid, num_floors_valid$NumFloors %in% c(31, 40))
floors_40$floorgroup = 40

allfloorgroups = rbind(floors_20, floors_30, floors_40)

ggplot(data=floors_20, aes(floors_20$YearBuilt, colour="floors20")) +
  geom_histogram(binwidth=1, fill="navy", alpha=0.6) +
  geom_histogram(data=floors_30, aes(floors_30$YearBuilt, colour="floors30"), binwidth=1, fill="green",
  geom_histogram(data=floors_40, aes(floors_40$YearBuilt, colour="floors40"), binwidth=1, fill="red", al
  scale_y_continuous(trans="log2", limits=c(1.8,10000), na.value=0) +
  labs(title="Histogram for Year of Construction") +
  labs(x="Year", y="Count") +
  scale_fill_manual(name="Line Color", values=c(floors20="navy", floors30="green", floors40="red"))
```



3.

```
num_floors_valid$PPF = num_floors_valid$AssessTot/num_floors_valid$NumFloors
num_floors_valid = num_floors_valid[is.finite(num_floors_valid$PPF),]
f = ddply(num_floors_valid, .(YearBuilt), summarize,  Average=mean(PPF))
```

```
ggplot(data=f, aes(x=YearBuilt, y=Average)) +
  geom_bar(stat="identity") +
  labs(title="Histogram for Year of Construction vs Cost per floor") +
  labs(x="Year", y="Cost per floor")
```



It looks like she could be right but at the same time might be wrong. The years immediately prior to 1950 seem to be a bit higher than the late 1930s, and then there are a couple of vivid outliers. I don't have much experience in real estate, but cost per floor may not be the best indicator as cost of materials and services could have fluctuated wildly due to the ongoing events.