Creating new variables

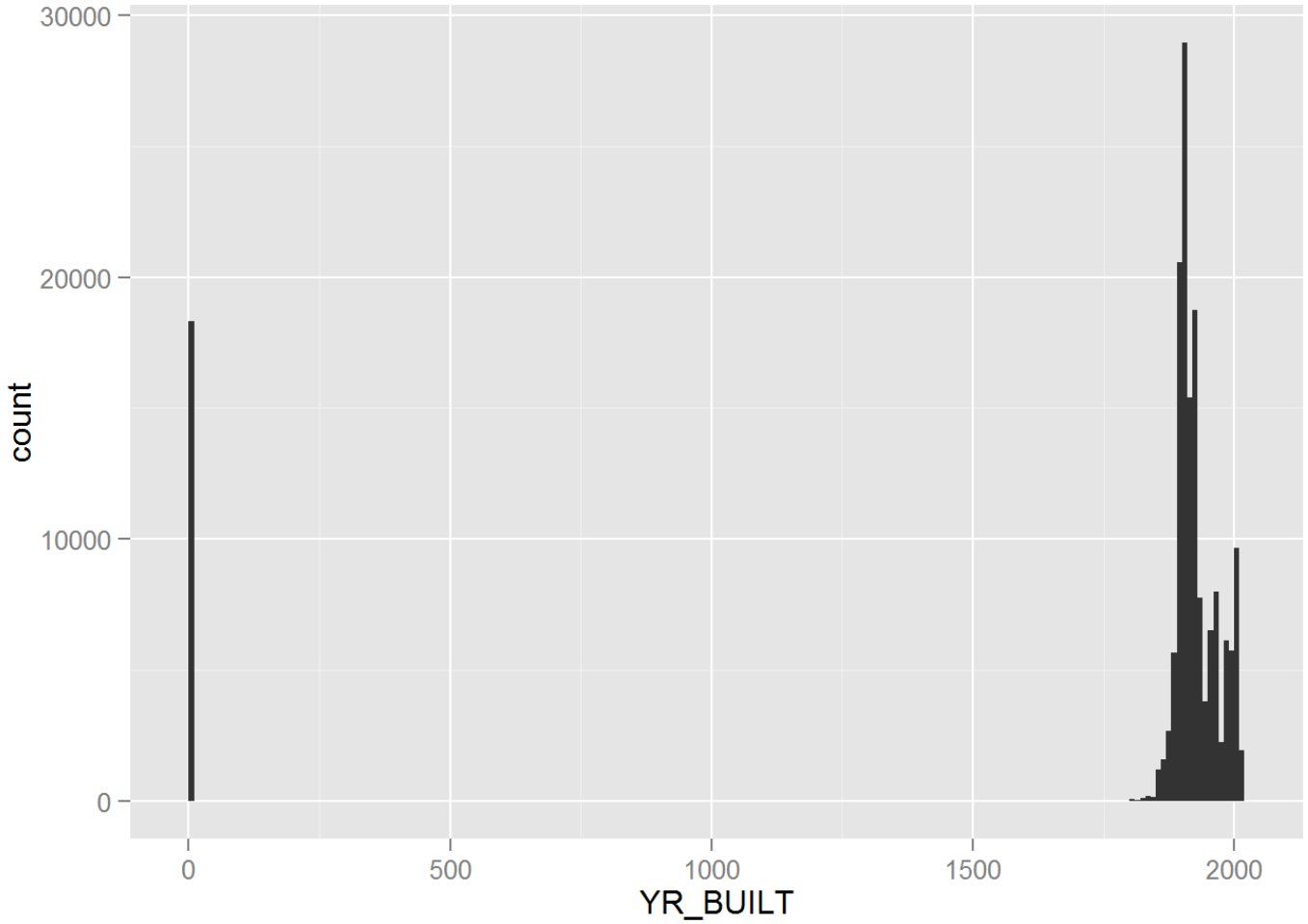
Hector Antonio Vazquez Brust

October 9, 2015

Creating new variables to fix or expand our data

As we've seen before, our dataset holds many properties whose year of construction is listed as zero:

```
ggplot(TAdata, aes(x = YR_BUILT)) + geom_histogram(binwidth = 10)
```

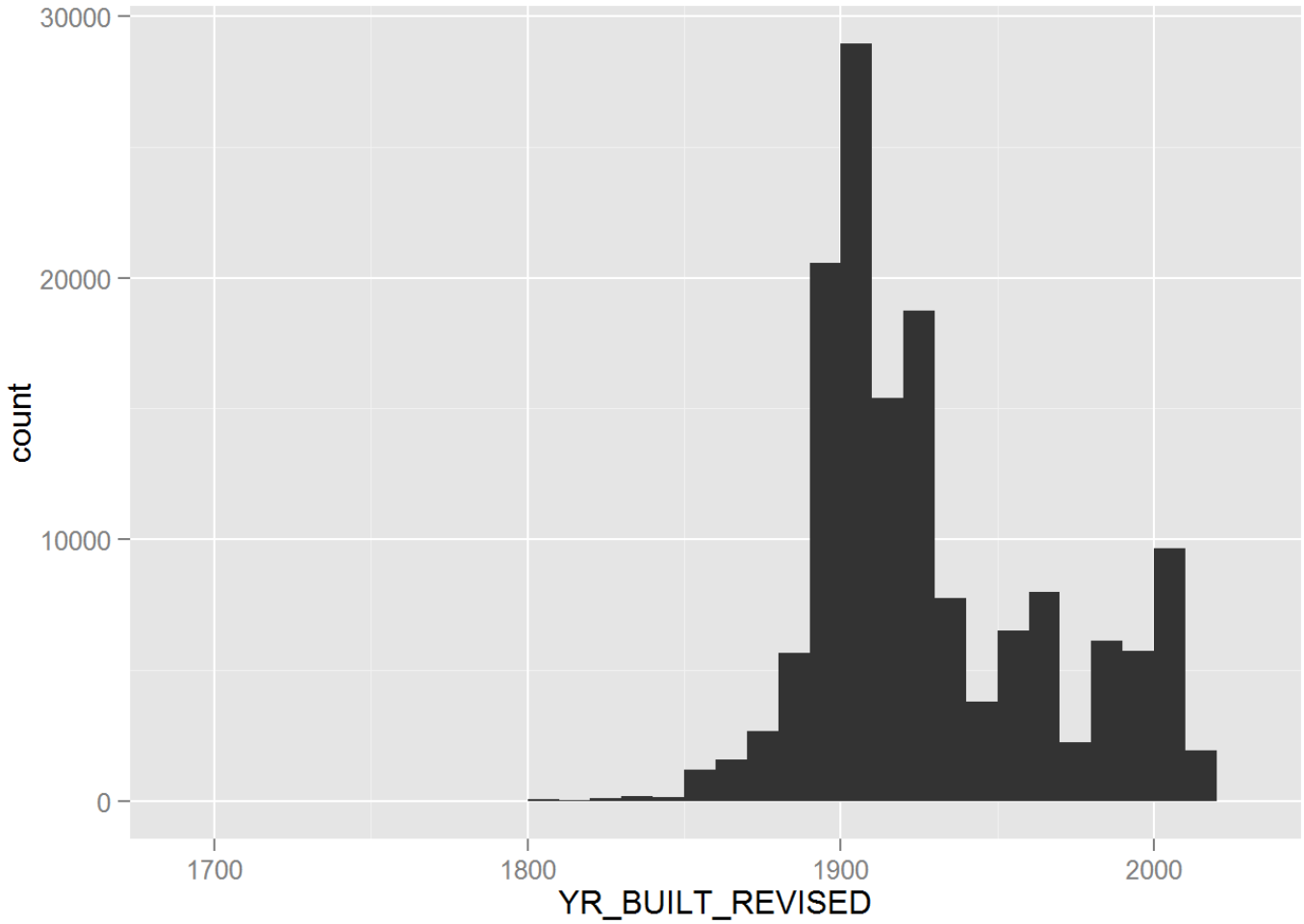


We will fix that by creating a "revised" variable, which will hold a "NA" where the YR_BUILT column lists a "0", and otherwise it will reflect the construction year:

```
TAdata <- transform(TAdata, YR_BUILT_REVISED = ifelse(YR_BUILT == 0, NA, YR_BUILT))
```

Let's check how it looks now:

```
ggplot(TAdata, aes(x = YR_BUILT_REVISED)) + geom_histogram(binwidth = 10)
```



Good; the “0” dates are gone.

Next, we will create two additional values, that will help us provide a more nuanced analysis of assesed land and building values in Boston, by giving us values normalized by total land area or gross area.

“AV_LAND_PER_SF” will tell us the asessed value per square foot of a parcel lot:

```
TAdata <- transform(TAdata, AV_LAND_PER_SF = AV_LAND / LAND_SF)
```

And “AV_BLDG_PER_SF” will give us the asessed value per square foot of a building:

```
TAdata <- transform(TAdata, AV_BLDG_PER_SF = AV_BLDG / GROSS_AREA)
```

How can we use these new variables?

Well, if we compare the total amount of assessed value per neighborhood, we find the downtown area on top, and Hyde Park at the bottom:

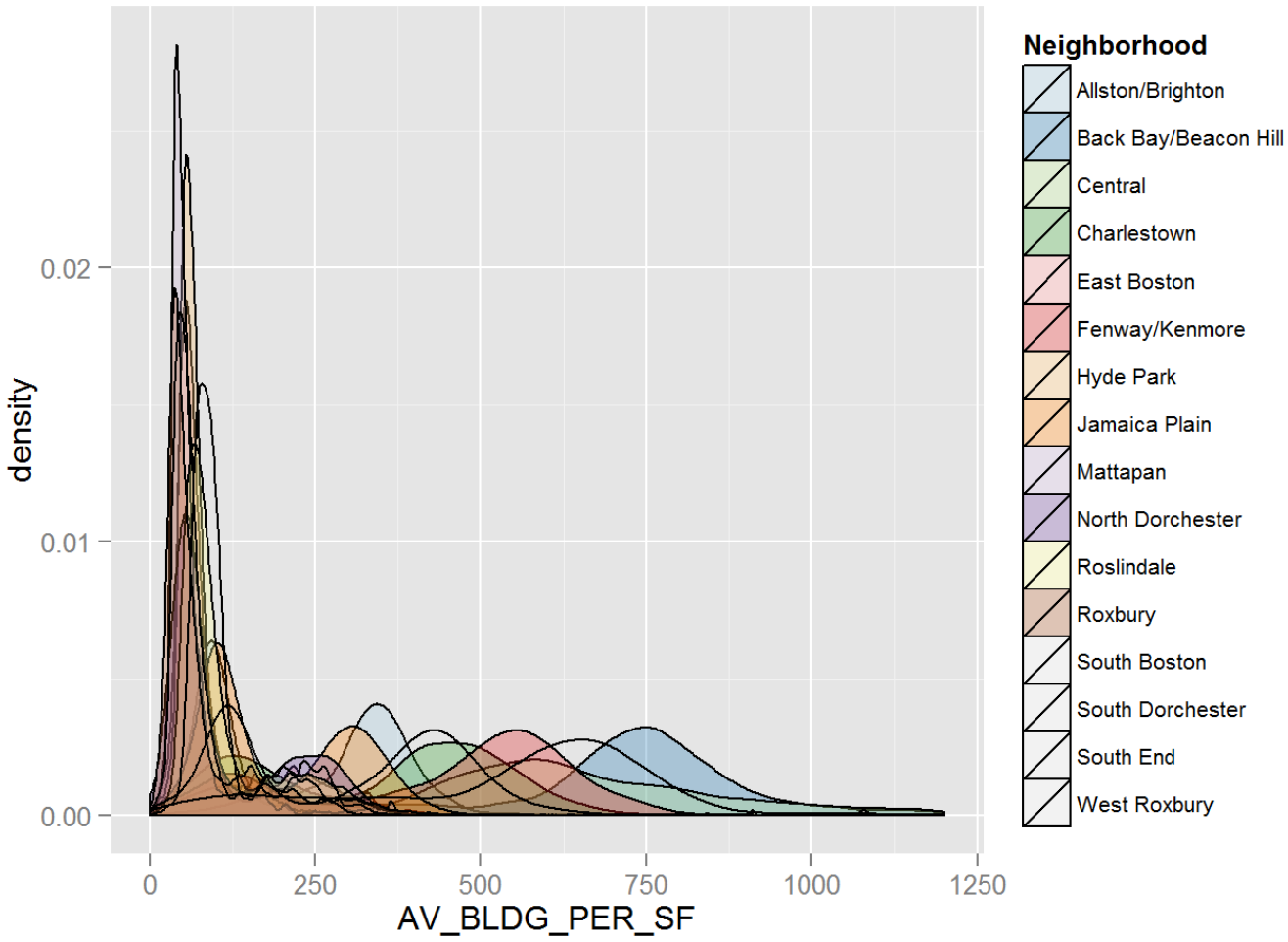
```
BRA_PD.value <- ddply(TAdata[(!is.na(TAdata$BRA_PD) & !is.na(TAdata$AV_BLDG)), ], .(BRA_P
D), summarise, bldg.value = sum(as.numeric(AV_BLDG)/1000000000), land.value = sum(as.nume
ric(AV_LAND)/1000000000))

arrange(BRA_PD.value, desc(bldg.value + land.value))
```

```
##                  BRA_PD bldg.value land.value
## 1               Central 44.9286173  12.322881
## 2        Fenway/Kenmore 33.1777086  11.052647
## 3           East Boston 19.5678924  22.795224
## 4      Allston/Brighton 27.4675421  14.464064
## 5  Back Bay/Beacon Hill 19.5349884   5.433032
## 6           South Boston 14.6486152   8.748036
## 7         Jamaica Plain 15.6877432   4.582938
## 8             South End 11.1079671   3.176731
## 9      North Dorchester  5.9516680   3.821066
## 10              Mattapan  5.0786037   3.088273
## 11          Charlestown  5.6468987   2.356018
## 12         West Roxbury  4.6110990   2.537452
## 13              Roxbury  3.8052696   1.901329
## 14     South Dorchester  3.5989092   1.854839
## 15            Roslindale  2.9080772   1.532717
## 16            Hyde Park  2.2990903   1.971368
## 17                       0.0049947   0.000000
```

(figures expressed in billions of USD)

But if we take a look at the density function for property value by sqaure meter, a different picture emerges:

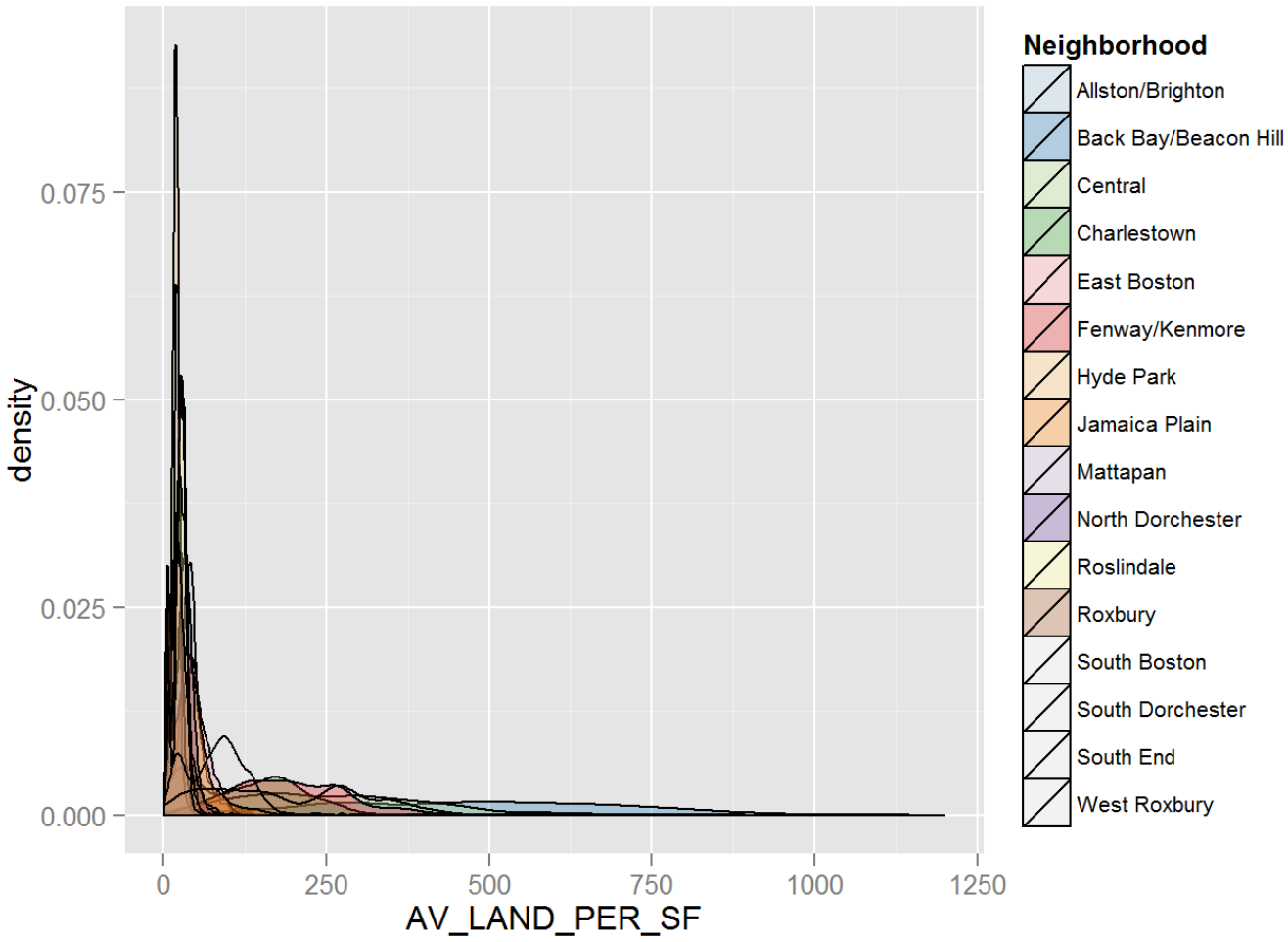
```
ggplot(TAdata[TAdata$BRA_PD != '' & TAdata$AV_BLDG_PER_SF != 0,], aes(AV_BLDG_PER_SF, fil
l = BRA_PD)) + geom_density(alpha = 0.3)  + xlim(1,1200) + scale_fill_brewer(name="Neighb
orhood", palette="Paired") + theme(legend.text = element_text(size = 8))
```



Back Bay/Beacon Hill is the area where the building value tends to be higher, as normalized by area; Mattapan comes last.

And what about land value?

```
ggplot(TAdata[TAdata$BRA_PD != '' & TAdata$AV_LAND_PER_SF != 0,], aes(AV_LAND_PER_SF, fil
l = BRA_PD)) + geom_density(alpha = 0.3) + xlim(1,1200) + scale_fill_brewer(name="Neighbo
rhood", palette="Paired") + theme(legend.text = element_text(size = 8))
```



Again, Back Bay/Beacon Hill is the area where values are higher than in the rest of the city, and Mattapan is where the values tend to be lower.

This illustrates how adding additional variables that combine the information of others in useful ways, can help us arrive at a better understanding of of our data.