Visualizing Education Across Boston Neighborhoods using ggplot2

```
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
wd <- getwd()
setwd(wd)
ACS_1216_TRACT_CSV <- paste(wd, '/ACS_1216_TRACT.csv', sep = '')
acs1216 <- read.csv(ACS_1216_TRACT_CSV)
colnames(acs1216)[colnames(acs1216)=="ï..CT_ID_10"] <- "CT_ID_10"
```

```
Tracts_Boston_2010_BARI_CSV <- paste(wd, '/Tracts_Boston_2010_BARI.csv', sep = '')
bos_tracts <- read.csv(Tracts_Boston_2010_BARI_CSV)
```

```
acs1216_bos<-merge(acs1216,bos_tracts,by='CT_ID_10',all.y=TRUE)
names(acs1216_bos)
```

```
##  [1] "CT_ID_10"         "TotalPop"         "PopDen"
##  [4] "SexRatio"         "AgeU18"           "Age1834"
##  [7] "Age3564"          "AgeO65"           "ForBorn"
## [10] "White"            "Black"            "Asian"
## [13] "Hispanic"         "TwoOrMore"        "EthHet"
## [16] "MedHouseIncome"   "PubAssist"        "GINI"
## [19] "FamPovPer"        "UnempRate"        "TotalHouseH"
## [22] "FamHousePer"      "FemHeadPer"       "SameSexCoupPer"
## [25] "GrandHeadPer"     "LessThanHS"       "HSGrad"
## [28] "SomeColl"         "Bach"             "Master"
## [31] "Prof"             "Doc"              "CommuteLess10"
## [34] "Commute1030"      "Commute3060"      "Commute6090"
## [37] "CommuteOver90"    "ByAuto"           "ByPubTrans"
## [40] "ByBike"           "ByWalk"           "TotalHouseUnits"
## [43] "VacantUnitPer"    "RentersPer"       "HomeOwnPer"
## [46] "MedGrossRent"     "MedHomeVal"       "MedYrBuiltRaw"
## [49] "MedYrBuilt"       "MedYrMovedInraw"  "MedYrRentMovedIn"
## [52] "MedYrOwnMovedIn"  "GEOID10"          "ALAND10"
## [55] "AWATER10"         "POP100"           "HU100"
## [58] "Type"             "Res"              "BRA_PD_ID"
## [61] "BRA_PD"           "City_Counc"       "WARD"
## [64] "ISD_Nbhd"         "Police_Dis"       "Fire_Distr"
## [67] "PWD"
```

I decided to take a look at education levels across the various Boston neighborhoods, and how much of an effect median household income has on education.

```
# Looking at the summary of the different levels of education to see if
# there are any obviously skewed variables
summary(acs1216_bos$LessThanHS, na.rm=TRUE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## 0.00000 0.04703 0.12039 0.14660 0.20130 1.00000       4
```

```
summary(acs1216_bos$HSGrad, na.rm=TRUE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  0.1102  0.2070  0.2039  0.3063  0.4615       4
```

```
summary(acs1216_bos$SomeColl, na.rm=TRUE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  0.1191  0.1689  0.1809  0.2413  0.5227       4
```

```
summary(acs1216_bos$Bach, na.rm=TRUE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  0.1331  0.2545  0.2562  0.3470  0.6030       4
```

```
summary(acs1216_bos$Master, na.rm=TRUE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## 0.00000 0.05246 0.13083 0.13356 0.20350 0.40227       4
```

```
summary(acs1216_bos$Doc, na.rm=TRUE)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
## 0.000000 0.005952 0.020710 0.033176 0.049918 0.363636        4
```

```
# Defining histograms for each of the levels of education
hist_LessThanHS<-ggplot(aes(x=LessThanHS), data=acs1216_bos) + geom_histogram()
hist_HSGrad<-ggplot(aes(x=HSGrad), data=acs1216_bos) + geom_histogram()
hist_SomeColl<-ggplot(aes(x=SomeColl), data=acs1216_bos) + geom_histogram()
hist_Bach<-ggplot(aes(x=Bach), data=acs1216_bos) + geom_histogram()
hist_Master<-ggplot(aes(x=Master), data=acs1216_bos) + geom_histogram()
hist_Doc<-ggplot(aes(x=Doc), data=acs1216_bos) + geom_histogram()
```

I decided to remove data relating to Doctorate degrees as the majority of people do not reach this level of education (<10% of people).

Here I wanted to see if there are any obvious correlations between a specific neighborhood and education level.

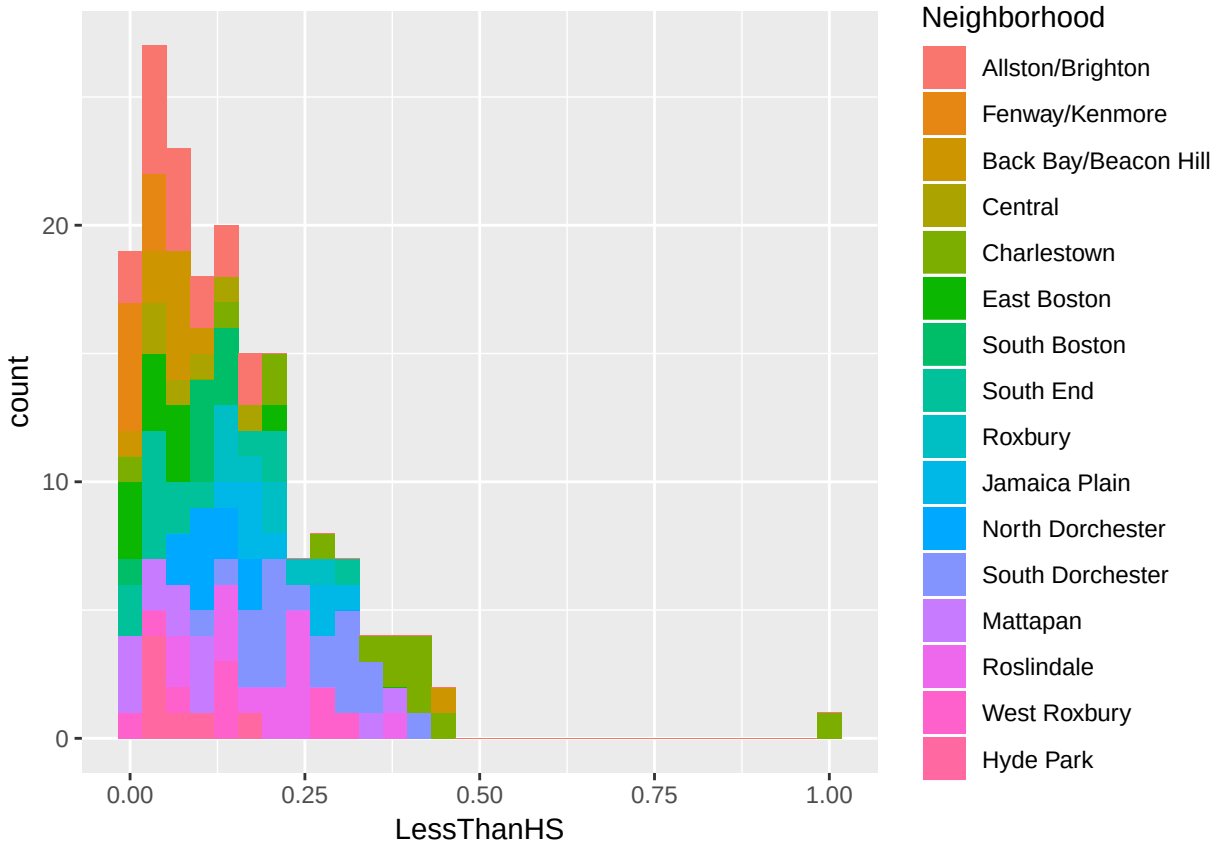
```
# Make vector of the neighborhoods
unique_neighborhoods <- unique(acs1216_bos$BRA_PD)

# Show stacked histograms for select education levels across neighborhoods
stack_hist_LessThanHS<-hist_LessThanHS + geom_histogram(aes(fill=BRA_PD)) + scale_fill_hue(name="Neighb
stack_hist_LessThanHS
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 4 rows containing non-finite values (stat_bin).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 4 rows containing non-finite values (stat_bin).
```



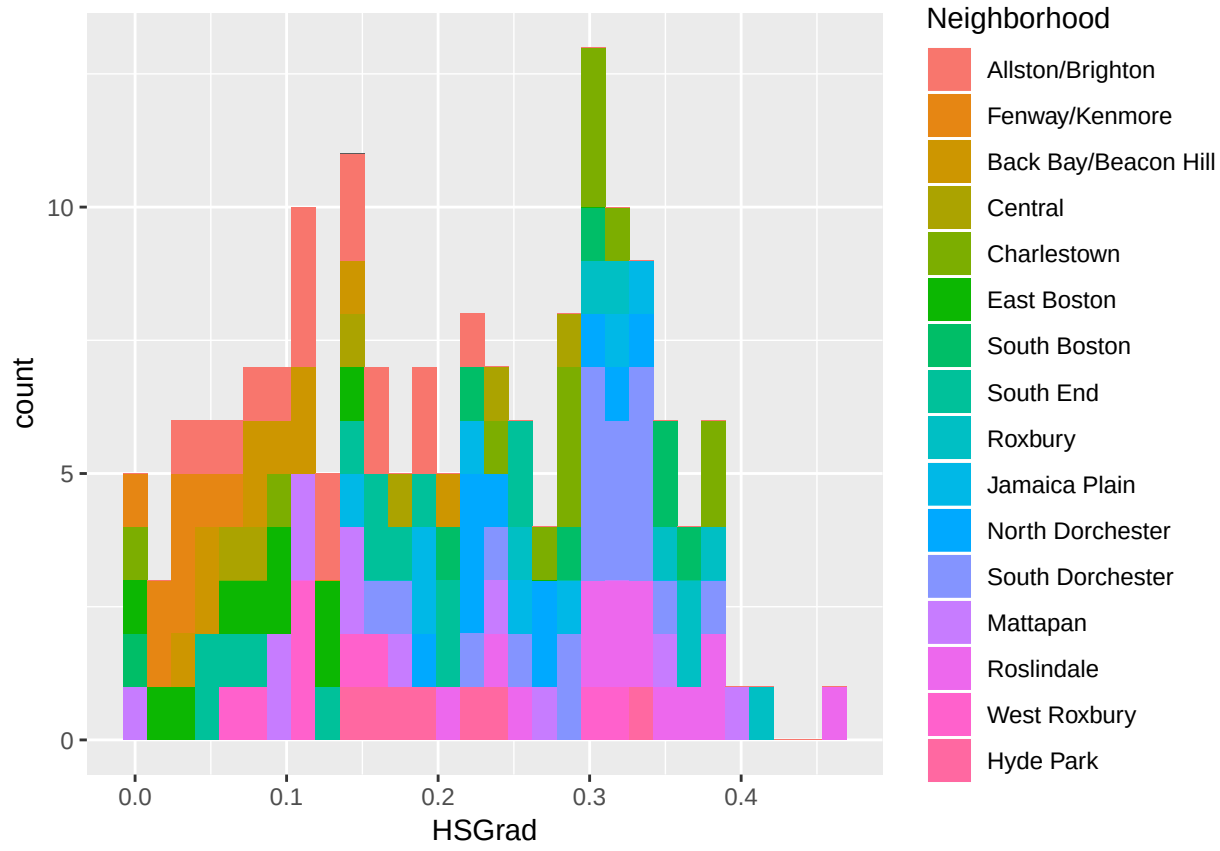

```
stack_hist_HSGrad<-hist_HSGrad + geom_histogram(aes(fill=BRA_PD)) + scale_fill_hue(name="Neighborhood",
stack_hist_HSGrad
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 4 rows containing non-finite values (stat_bin).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 4 rows containing non-finite values (stat_bin).
```



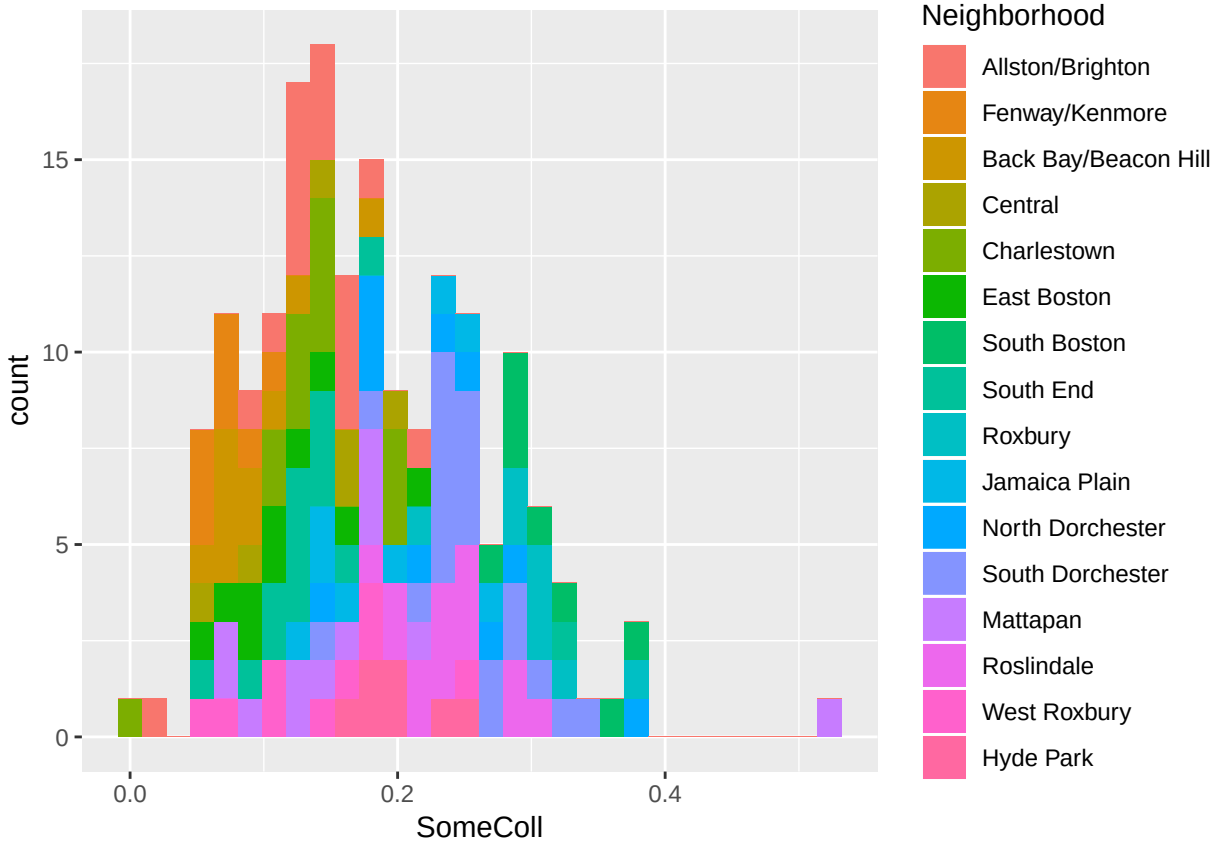

```
stack_hist_SomeColl<-hist_SomeColl + geom_histogram(aes(fill=BRA_PD)) + scale_fill_hue(name="Neighborho
stack_hist_SomeColl
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 4 rows containing non-finite values (stat_bin).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 4 rows containing non-finite values (stat_bin).
```



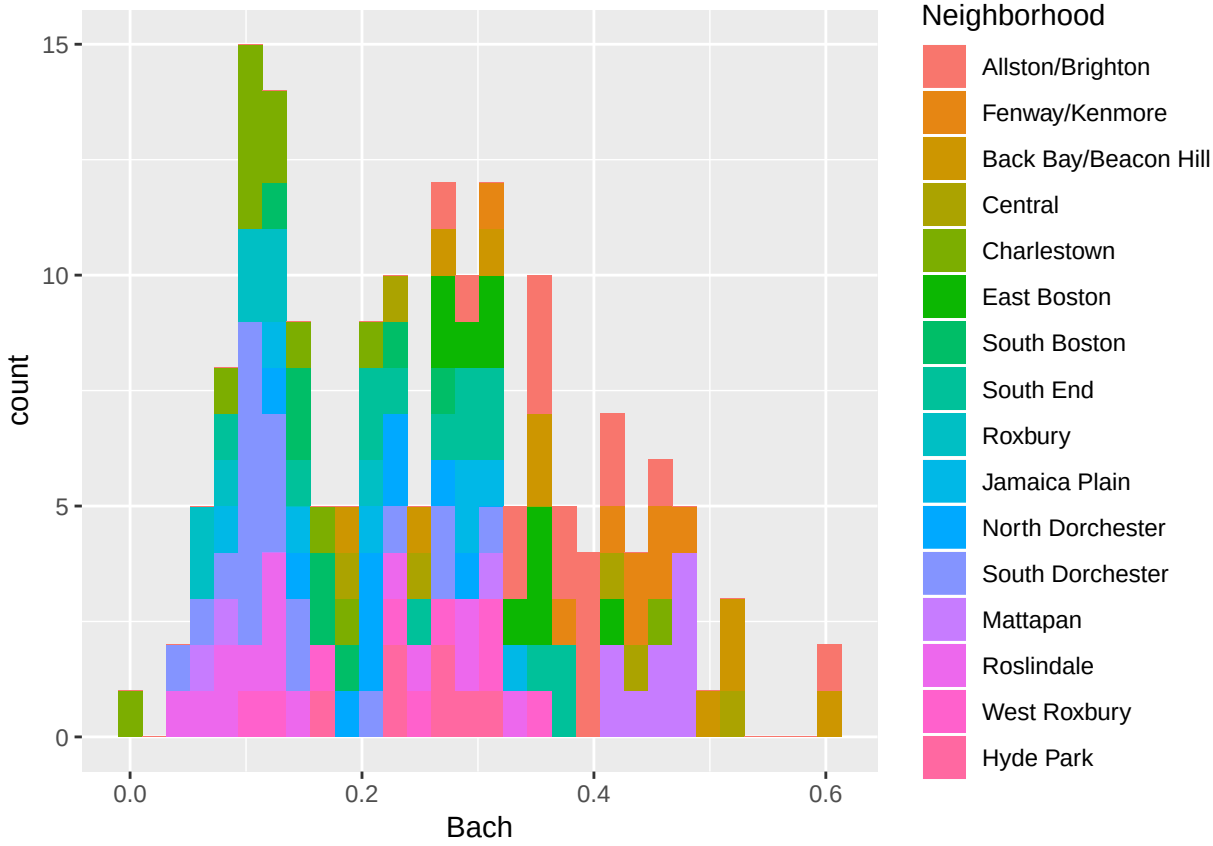

```
stack_hist_Bach<-hist_Bach + geom_histogram(aes(fill=BRA_PD)) + scale_fill_hue(name="Neighborhood", labe
stack_hist_Bach
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 4 rows containing non-finite values (stat_bin).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 4 rows containing non-finite values (stat_bin).
```



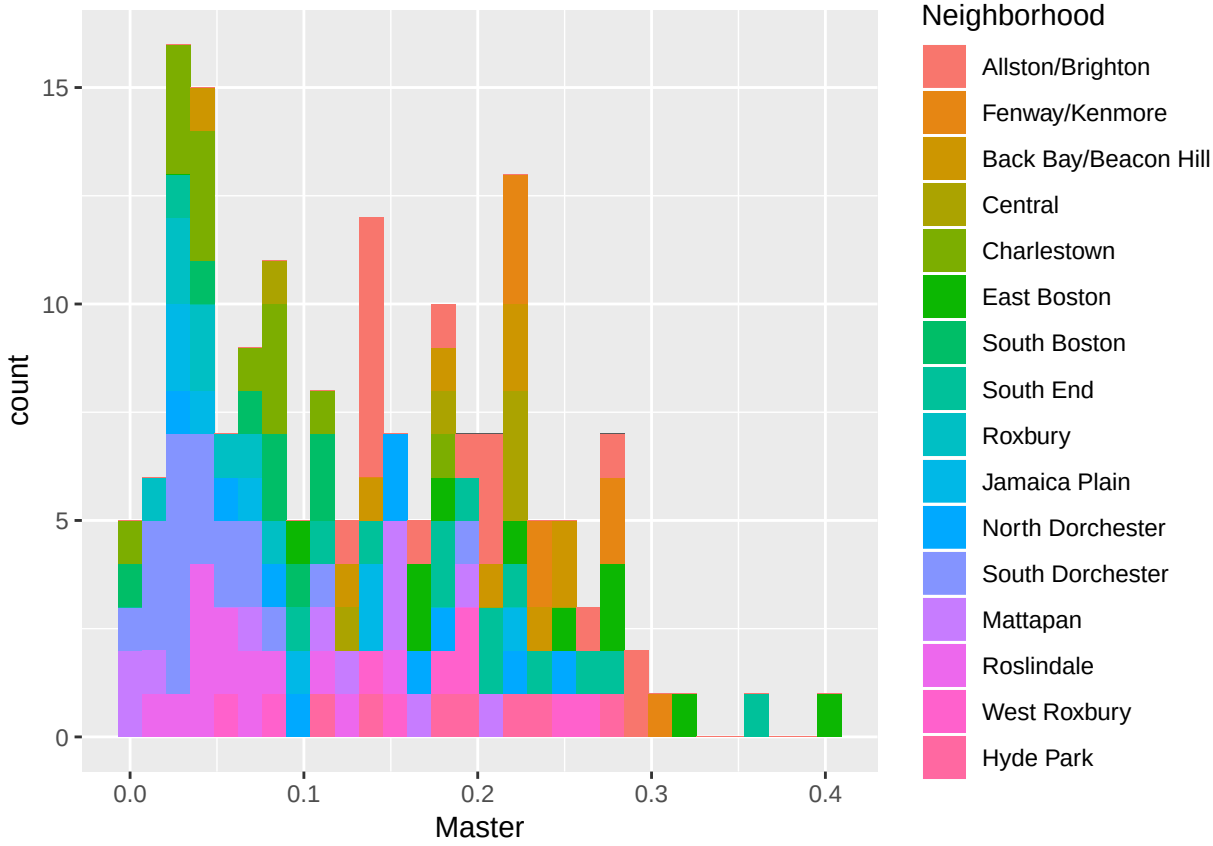

```
stack_hist_Master<-hist_Master + geom_histogram(aes(fill=BRA_PD)) + scale_fill_hue(name="Neighborhood",
stack_hist_Master
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 4 rows containing non-finite values (stat_bin).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 4 rows containing non-finite values (stat_bin).
```



It seems to be that Allston/Brighton, Fenway/Kenmore, and Back Bay/Beacon Hill have the lowest percentage of people with less than high school education. The nieghborhood with the highest percentage of people with less than high school education is Charlestown and South Dorchester.

On the other end of the spectrum, those same neighborhoods (Allston/Brighton, Fenway/Kenmore, and Back Bay/Beacon) seem to hold the largest percentage count of people with college graduates (bachelors and masters degrees). The neighborhood with the lowest percentage of college graduates seems to be Charlestown and South Dorchester.

I was also curious to see what the median household income looked like for the different neighborhoods, to see if this could potentially have an effect on education level.

```
# Histogram of median household income across neighborhoods
stack_hist_income<-ggplot(aes(x=MedHouseIncome), data=acs1216_bos) + geom_histogram(aes(fill=BRA_PD)) +
stack_hist_income + facet_wrap(~BRA_PD)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 8 rows containing non-finite values (stat_bin).
```

It is not clear from this chart of Median Household Income by neighborhood which neighborhood has the highest overall median household income. Central, Charlestown, and Back Bay look to be some of the highest, but overall it is hard to distginguish. This does show however that Roxbury seems to have the largest counts of low household incomes.

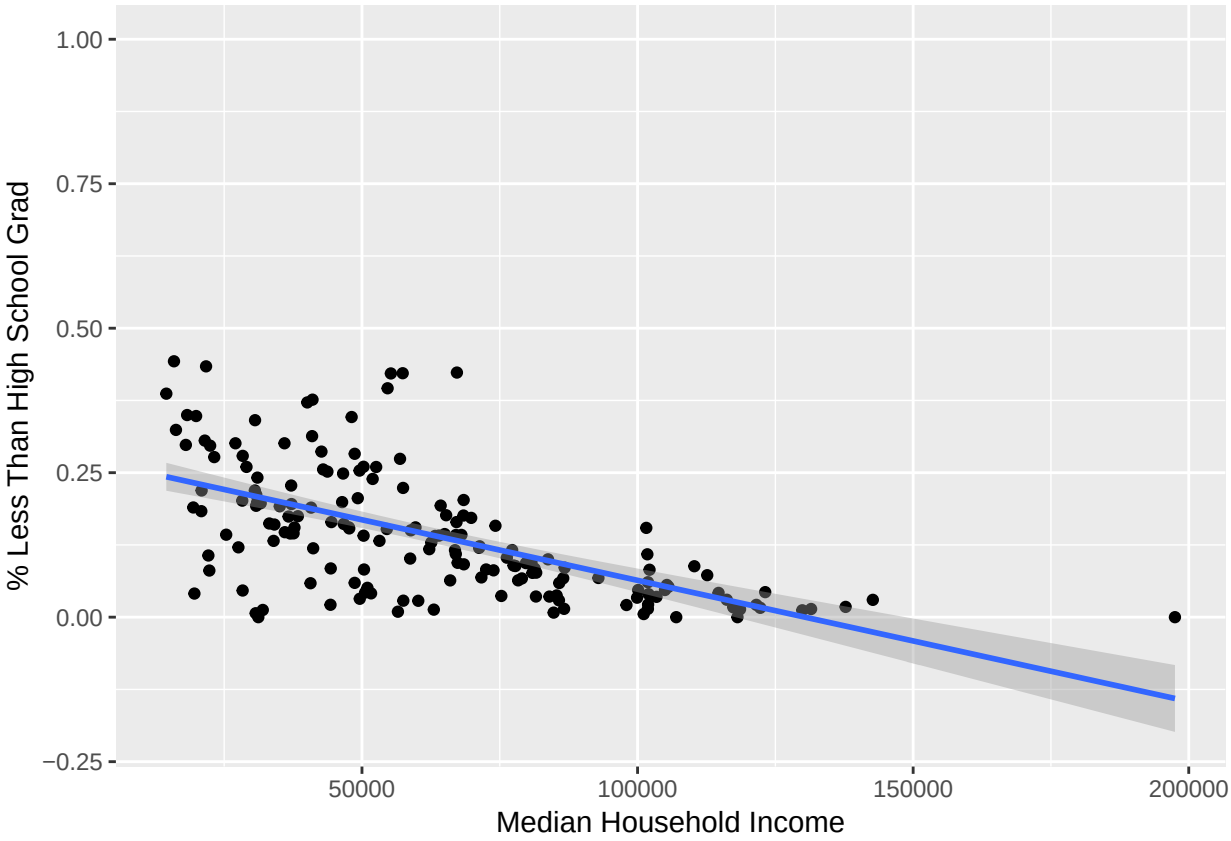
We now know which neighborhoods tend to have more people with higher education (Allston/Brighton, Fenway/Kenmore, and Back Bay/Beacon), and which neighborhoods tend to have more people with less education (Charlestown, South Dorchester).

I want to investigate more to see if median household income has anything to do with education level.

```
# Plot median household income against percentage of people who have less than high school education, w
plot_income_LessThanHS<-ggplot(data=acs1216_bos, aes(x=MedHouseIncome, y=LessThanHS)) + geom_point() + s
plot_income_LessThanHS
```

```
## Warning: Removed 8 rows containing non-finite values (stat_smooth).
```

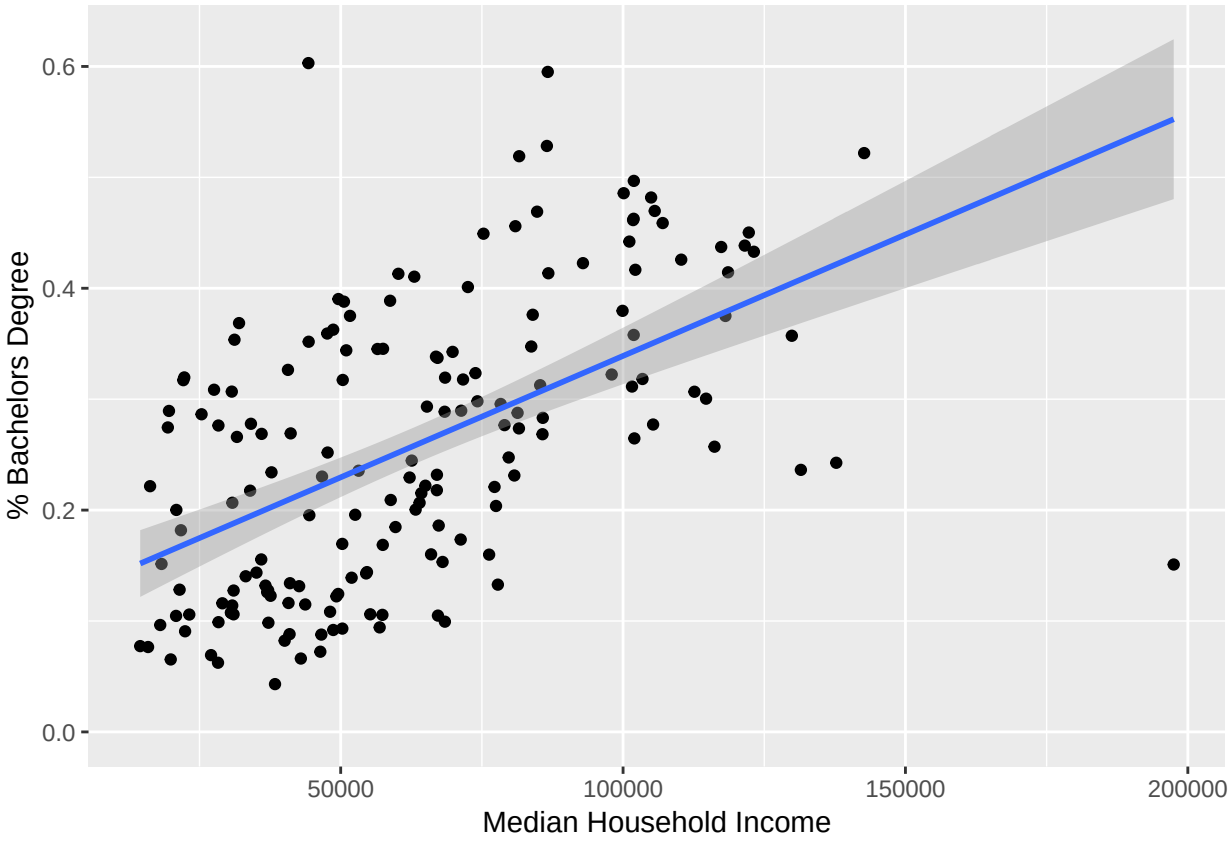
```
## Warning: Removed 8 rows containing missing values (geom_point).
```




```
# Plot median household income against percentage of people who have a bachelors degree, with a line of
plot_income_Bach<-ggplot(data=acs1216_bos, aes(x=MedHouseIncome, y=Bach)) + geom_point() + xlab("Median
plot_income_Bach
```

```
## Warning: Removed 8 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 8 rows containing missing values (geom_point).
```



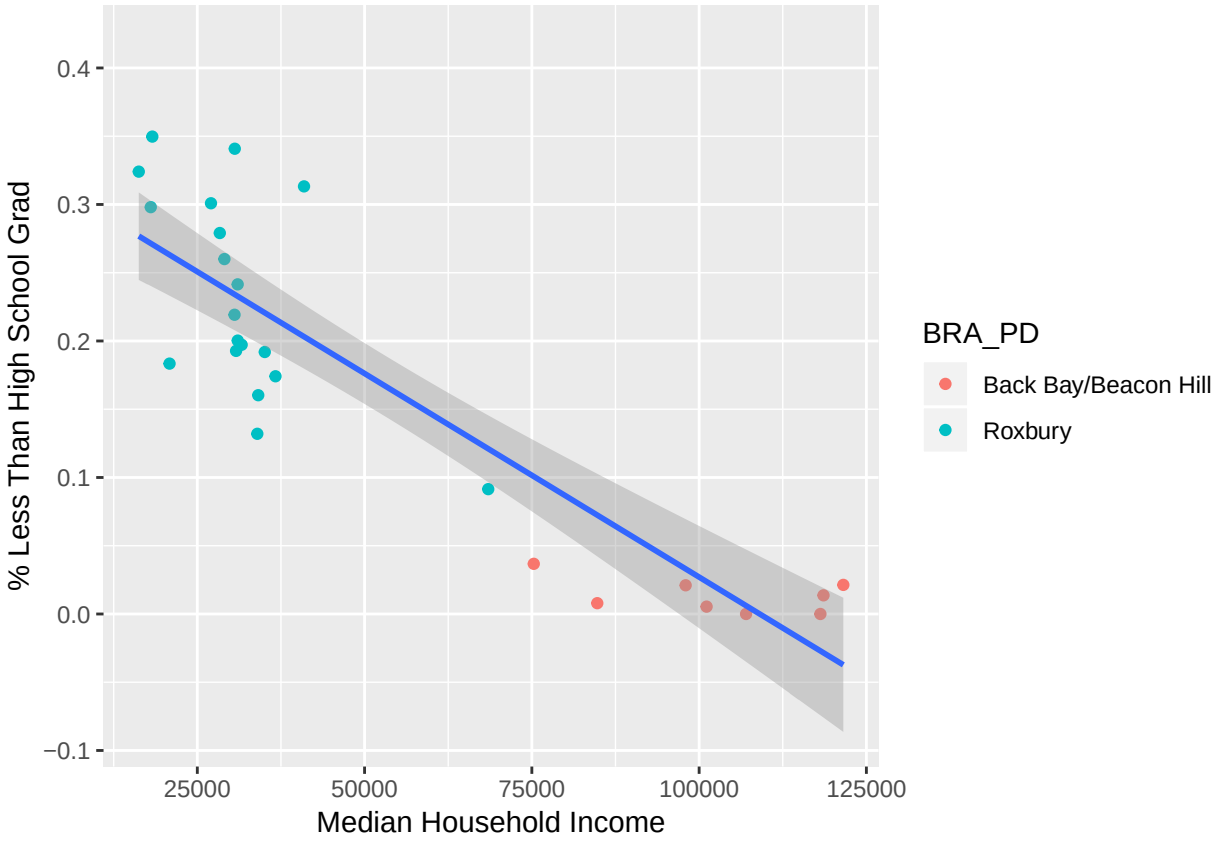
In these two plots, we can see a negative correlation between median household income and percentage of people not completing high school, and a slightly stronger positive correlation between median household income and percentage of people graduating college with a bachelors degree.

Finally, lets compare the education levels of two neighborhoods in Boston on opposite ends of the economic spectrum, Roxbury and Back Bay/Beacon Hill.

```
# Get the subset of the data only including back bay and beacon hill
data_roxbury_backbay_beacon <- subset(acs1216_bos, acs1216_bos$BRA_PD == "Roxbury" | acs1216_bos$BRA_PD

# Plot median household income against percentage of people who have less than high school education, w
plot_income_LessThanHS<-ggplot(data=data_roxbury_backbay_beacon, aes(x=MedHouseIncome, y=LessThanHS)) +
plot_income_LessThanHS
```

```
## Warning: Removed 2 rows containing non-finite values (stat_smooth).
```

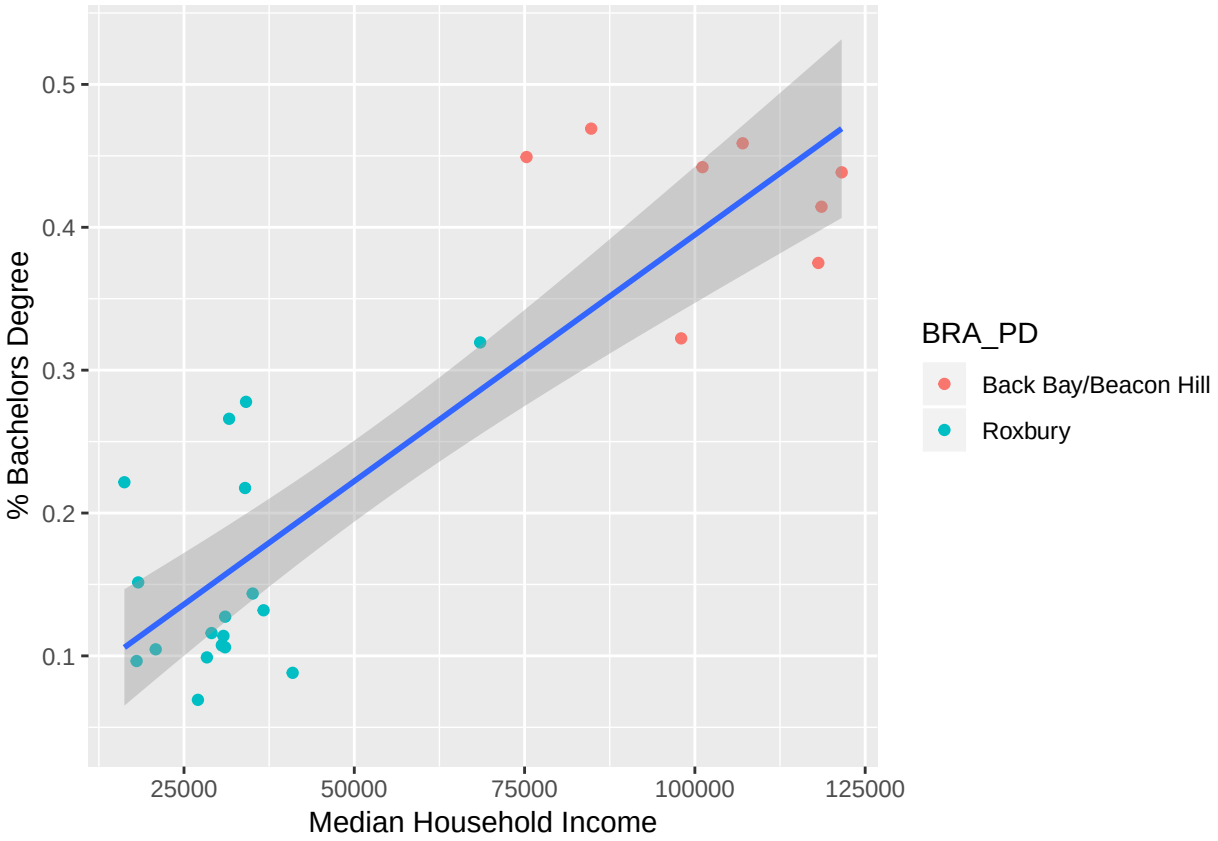
```
## Warning: Removed 2 rows containing missing values (geom_point).
```




```
# Plot median household income against percentage of people who have a bachelors degree, with a line of
plot_income_Bach<-ggplot(data=data_roxbury_backbay_beacon, aes(x=MedHouseIncome, y=Bach)) + geom_point(a
plot_income_Bach
```

```
## Warning: Removed 2 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```



To conclude, we can see that as median household income increases, the chances of higher education level increases. Although this is a fairly basic concept, college tuition prices being as high as they are, the idea is crystal clear when the data is visualized.

Overall, there were not huge disparities in the data, most neighborhoods had fairly evenly dispersed levels of education and income. If there was a particularly high-income neighborhood, chances are they can afford higher education. If there was a particularly low-income neighborhood, its less likely they can afford higher education. Neighorhoods in between, it could go either way.