Final Project

Erin Poole

6/17/2021

Big Question

The big question that I had for this data set: is there a correlation between community,race makeup, and violent crime rates?

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.
3.1 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.1     v dplyr   1.0.6
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflict
s() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(tidyr)
library(dplyr)
library(Rmisc)
```

```
## Loading required package: lattice
```

```
## Loading required package: plyr
```

```
## ------------------------------------------------------------------------
----
```

```
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, th
en dplyr:
## library(plyr); library(dplyr)
```

```
## ------------------------------------------------------------------------
----

##
## Attaching package: 'plyr'

## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## The following object is masked from 'package:purrr':
##
##     compact

library(tree)

## Registered S3 method overwritten by 'tree':
##   method     from
##   print.tree cli

library(ISLR)
library(factoextra)

## Welcome! Want to learn more? See two factoextra-related books at https://g
oo.gl/ve3WBa

library(stats)
library(devtools)

## Loading required package: usethis

library(ggbiplot)

## Loading required package: scales

##
## Attaching package: 'scales'

## The following object is masked from 'package:purrr':
##
##     discard

## The following object is masked from 'package:readr':
##
##     col_factor

## Loading required package: grid
```

Data Sources

The data set that I picked is communities and crime. The reason I picked this data set is because I'm the most interested in crime. The data source I'm using is multivariate, where

the attribute characteristics are real, and the associated task is regression. The data source included attributes that were picked if there wase any plausible connection to crime, plus the attribe to be predicted, which was per capita violent crimes. ## Processing Problems ### Problem 1 The first problem I had was figuring out how to clean up the data and how to name the columns. I didn't realize that the column names came with the data, so that was the first hurdle. I figured out how to associate the name of the column to the appropriate spot. The next hurdle was trying to figure out how to fill in the spaces that are emptied. I first turned the zeros into NA. There were many attempts at turning NA into a number. I finally found the formula that would turn NA into a mean by using the surrounding numbers to help fill in the gaps.

```
finalp<-read.csv("./CrimesandCommunitiesdatafile.csv")
names(finalp) <- c(
  'state',
  'county',
  'community',
  'communityname',
  'fold',
  'population',
  'householdsize',
  'racepctblack',
  'racePctWhite',
  'racePctAsian',
  'racePctHisp',
  'agePct12t21',
  'agePct12t29',
  'agePct16t24',
  'agePct65up',
  'numbUrban',
  'pctUrban',
  'medIncome',
  'pctWWage',
  'pctWFarmSelf',
  'pctWInvInc',
  'pctWSocSec',
  'pctWPubAsst',
  'pctWRetire',
  'medFamInc',
  'perCapInc',
  'whitePerCap',
  'blackPerCap',
  'indianPerCap',
  'AsianPerCap',
  'OtherPerCap',
  'HispPerCap',
  'NumUnderPov',
  'PctPopUnderPov',
  'PctLess9thGrade',
  'PctNotHSGrad',
```

```
  'PctBSorMore',
  'PctUnemployed',
  'PctEmploy',
  'PctEmplManu',
  'PctEmplProfServ',
  'PctOccupManu',
  'PctOccupMgmtProf',
  'MalePctDivorce',
  'MalePctNevMarr',
  'FemalePctDiv',
  'TotalPctDiv',
  'PersPerFam',
  'PctFam2Par',
  'PctKids2Par',
  'PctYoungKids2Par',
  'PctTeen2Par',
  'PctWorkMomYoungKids',
  'PctWorkMom',
  'NumIlleg',
  'PctIlleg',
  'NumImmig',
  'PctImmigRecent',
  'PctImmigRec5',
  'PctImmigRec8',
  'PctImmigRec10',
  'PctRecentImmig',
  'PctRecImmig5',
  'PctRecImmig8',
  'PctRecImmig10',
  'PctSpeakEnglOnly',
  'PctNotSpeakEnglWell',
  'PctLargHouseFam',
  'PctLargHouseOccup',
  'PersPerOccupHous',
  'PersPerOwnOccHous',
  'PersPerRentOccHous',
  'PctPersOwnOccup',
  'PctPersDenseHous',
  'PctHousLess3BR',
  'MedNumBR',
  'HousVacant',
  'PctHousOccup',
  'PctHousOwnOcc',
  'PctVacantBoarded',
  'PctVacMore6Mos',
  'MedYrHousBuilt',
  'PctHousNoPhone',
  'PctWOFullPlumb',
  'OwnOccLowQuart',
  'OwnOccMedVal',
```

```
  'OwnOccHiQuart',
  'RentLowQ',
  'RentMedian',
  'RentHighQ',
  'MedRent',
  'MedRentPctHousInc',
  'MedOwnCostPctInc',
  'MedOwnCostPctIncNoMtg',
  'NumInShelters',
  'NumStreet',
  'PctForeignBorn',
  'PctBornSameState',
  'PctSameHouse85',
  'PctSameCity85',
  'PctSameState85',
  'LemasSwornFT',
  'LemasSwFTPerPop',
  'LemasSwFTFieldOps',
  'LemasSwFTFieldPerPop',
  'LemasTotalReq',
  'LemasTotReqPerPop',
  'PolicReqPerOffic',
  'PolicPerPop',
  'RacialMatchCommPol',
  'PctPolicWhite',
  'PctPolicBlack',
  'PctPolicHisp',
  'PctPolicAsian',
  'PctPolicMinor',
  'OfficAssgnDrugUnits',
  'NumKindsDrugsSeiz',
  'PolicAveOTWorked',
  'LandArea',
  'PopDens',
  'PctUsePubTrans',
  'PolicCars',
  'PolicOperBudg',
  'LemasPctPolicOnPatr',
  'LemasGangUnitDeploy',
  'LemasPctOfficDrugUn',
  'PolicBudgPerPop',
  'ViolentCrimesPerPop') # Changed column names
finalp[finalp=="?"] <- NA # Changed ? to NA
finalp$county <- as.numeric(finalp$county)
finalp$county[is.na(finalp$county)]<-mean(finalp$county, na.rm=TRUE)
finalp$community<-as.numeric(finalp$community)
finalp$community[is.na(finalp$community)]<-mean(finalp$community, na.rm=TRUE)
```

Problem 2

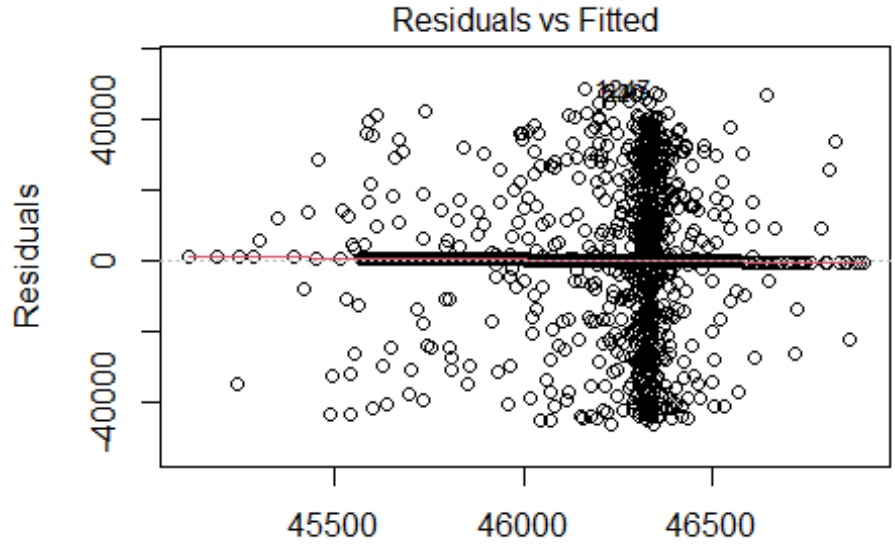
The second main problem I had was figuring out which graph would work best for the information I wanted to pull. However, I couldn't for a while figure out how to get the graphs to show up. After looking up different resources, I found the libraries that I needed to load into the script in order to run the graphs that I needed.

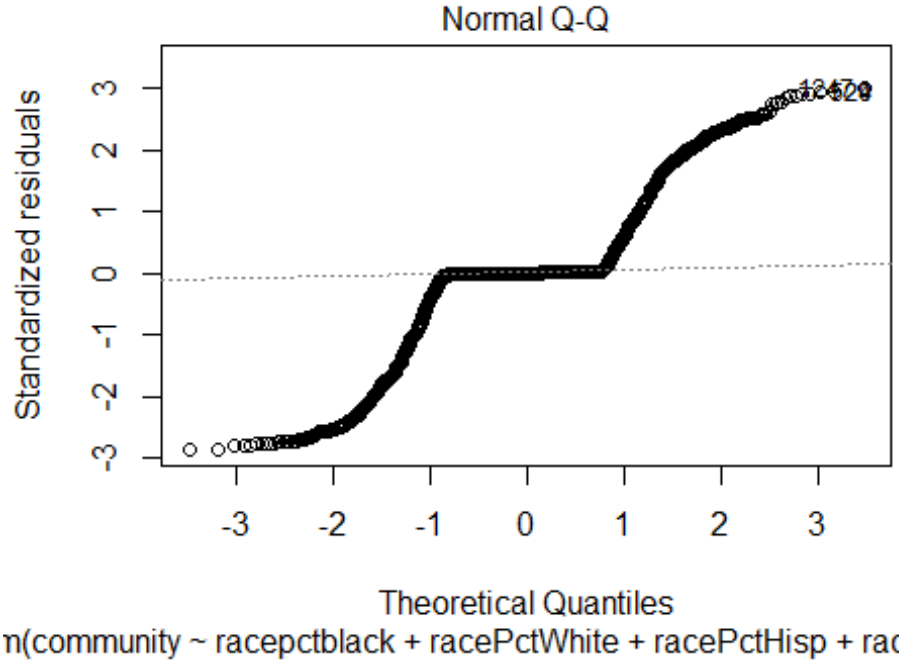
Problem 3

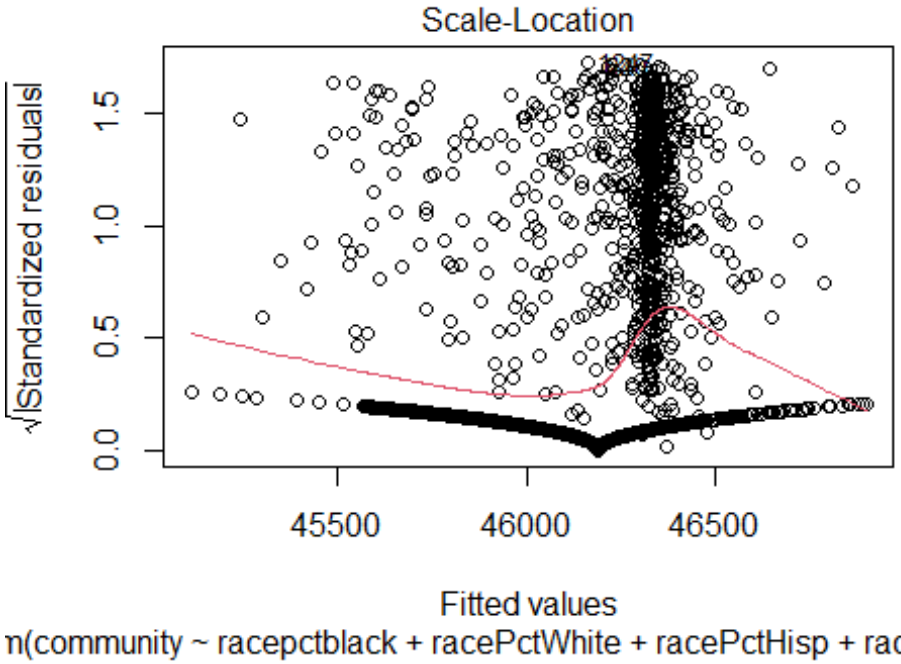
It was hard to figure out what question I wanted to ask. The column names were hard to understand and I couldn't figure out what they meant until I had to look them up. When I realized that the data wasn't organized by what crimes were committed, but just by population and breaking that down into smaller pieces, I had to change my question. After reading the data source description and going through the columns, I had to re-frame my question.

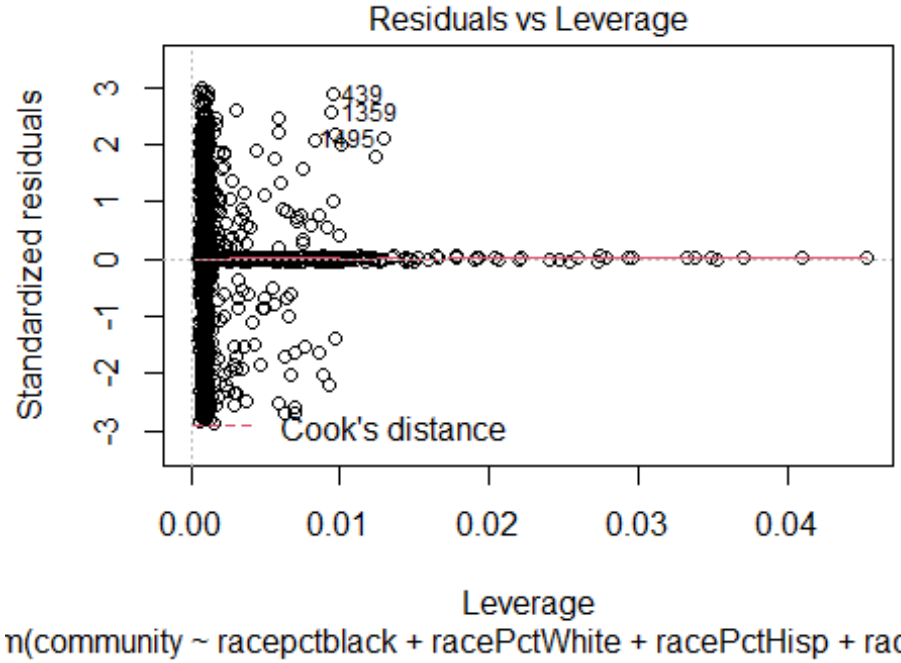
Visuializations

```
model1<-lm(community~racepctblack+racePctWhite+racePctHisp+racePctAsian,data=
finalp)
summary(model1)

##
## Call:
## lm(formula = community ~ racepctblack + racePctWhite + racePctHisp +
##     racePctAsian, data = finalp)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -46240   -231    -14    535  48435
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  46311.71    5894.06   7.857 6.37e-15 ***
## racepctblack  -736.38    4963.75  -0.148    0.882
## racePctWhite    10.15    5806.73   0.002    0.999
## racePctHisp   -666.09    3109.92  -0.214    0.830
## racePctAsian   632.35    2456.93   0.257    0.797
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16200 on 1989 degrees of freedom
## Multiple R-squared:  0.0002582,  Adjusted R-squared:  -0.001752
## F-statistic: 0.1284 on 4 and 1989 DF,  p-value: 0.9721

plot(model1)
```


Residuals vs Fitted
Residuals
40000
0
-40000
45500
46000
46500
Fitted values
lm(community ~ racepctblack + racePctWhite + racePctHisp + racePct

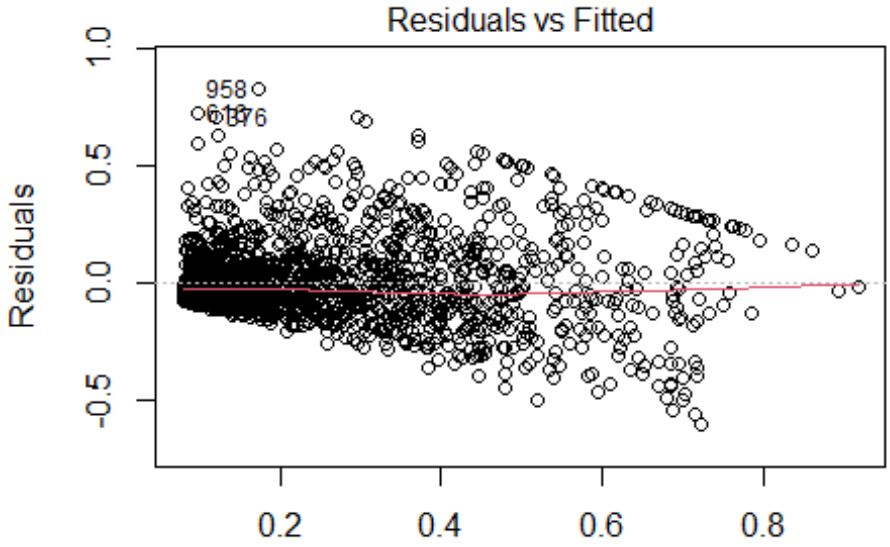

Normal Q-Q
Standardized residuals
3
2
1
0
-1
-2
-3
Theoretical Quantiles
lm(community ~ racepctblack + racePctWhite + racePctHisp + racePct



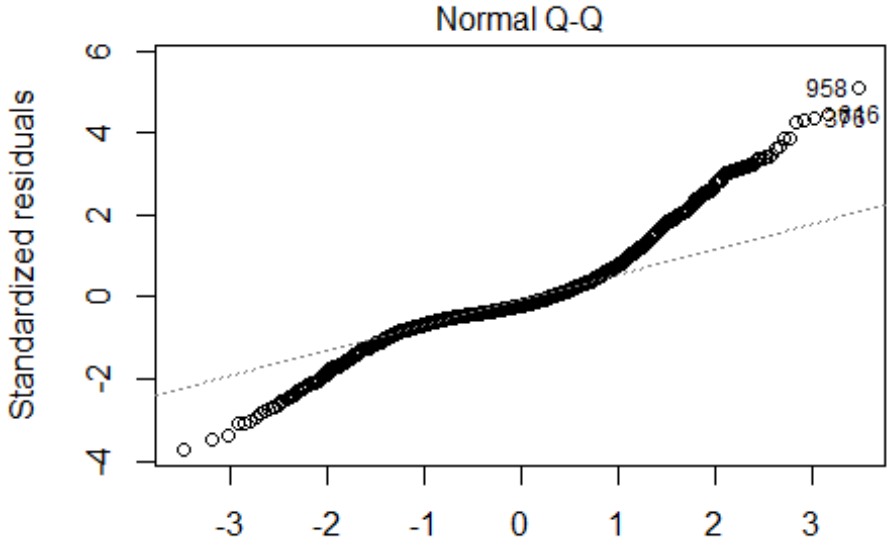



```
model2<-lm(ViolentCrimesPerPop~racepctblack+racePctWhite+racePctHisp+racePctA
sian,data=finalp)
summary(model2)
```

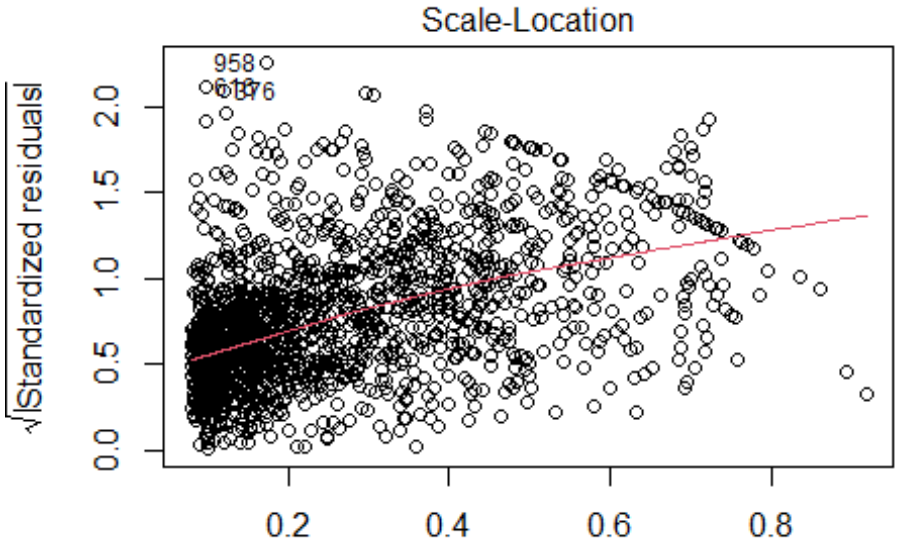
```
##
## Call:
## lm(formula = ViolentCrimesPerPop ~ racepctblack + racePctWhite +
##     racePctHisp + racePctAsian, data = finalp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.60241 -0.07988 -0.03394  0.05589  0.82612
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.25058    0.05916   4.235 2.39e-05 ***
## racepctblack  0.46328    0.04982   9.298  < 2e-16 ***
## racePctWhite -0.17018    0.05829  -2.920  0.00354 **
## racePctHisp   0.25450    0.03122   8.153 6.21e-16 ***
## racePctAsian -0.02734    0.02466  -1.108  0.26781
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1626 on 1989 degrees of freedom
## Multiple R-squared:  0.5137, Adjusted R-squared:  0.5127
## F-statistic: 525.3 on 4 and 1989 DF,  p-value: < 2.2e-16

plot(model2)
```

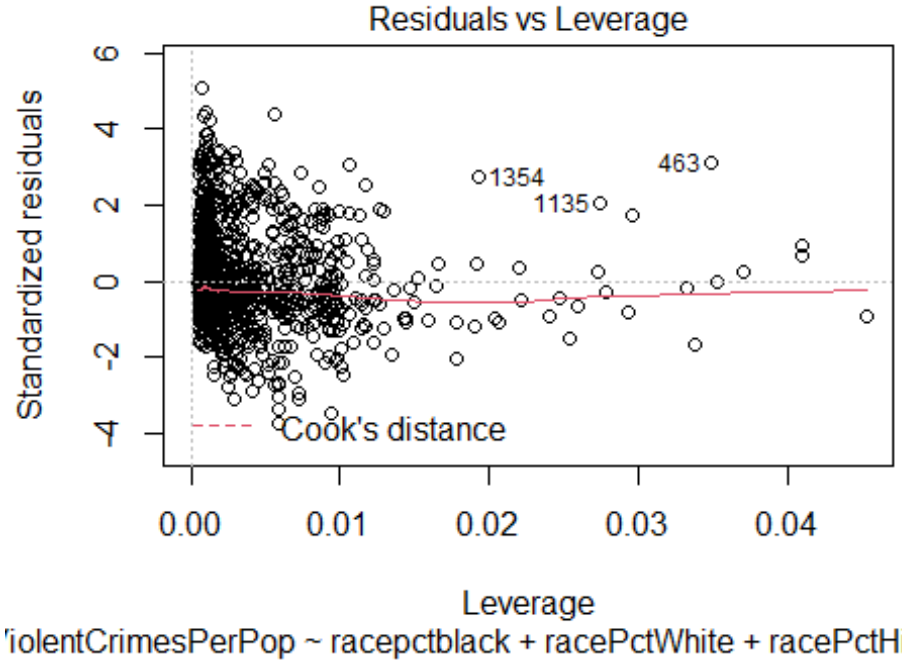


'iolentCrimesPerPop ~ racepctblack + racePctWhite + racePctHisp +



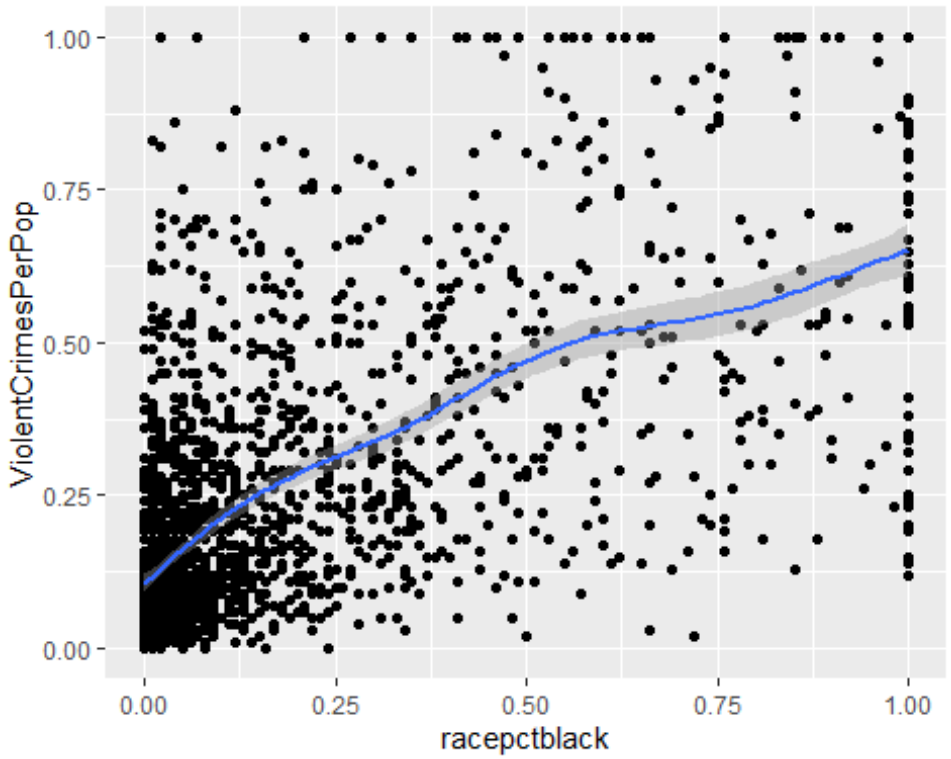
'iolentCrimesPerPop ~ racepctblack + racePctWhite + racePctHisp +



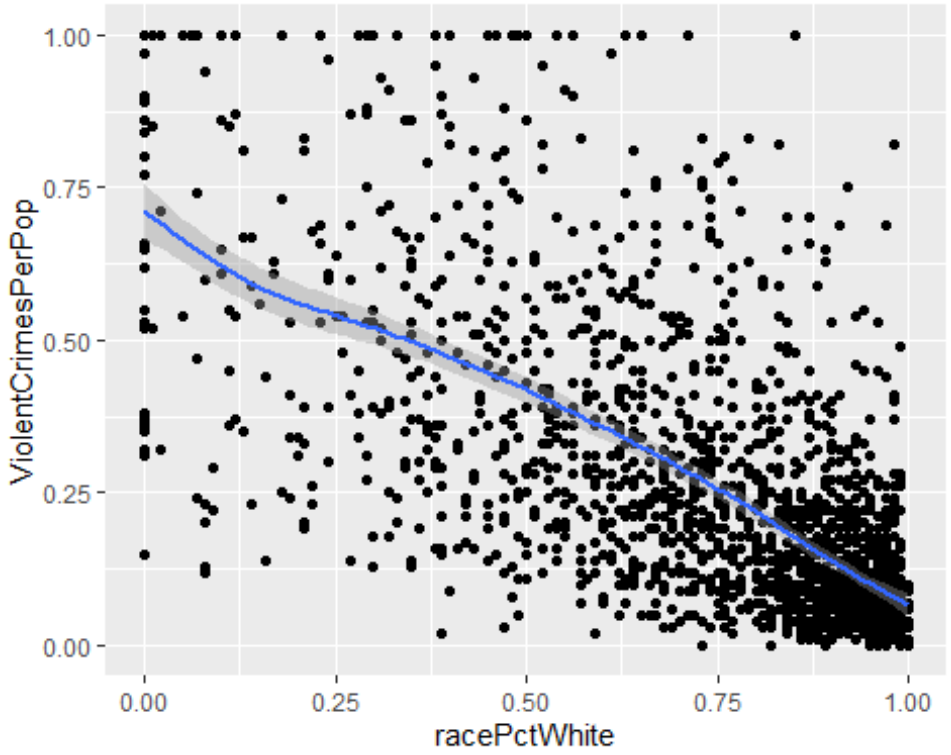



```
ggplot(finalp,aes(x=racepctblack,y=ViolentCrimesPerPop))+
  geom_point()+
  geom_smooth()
```

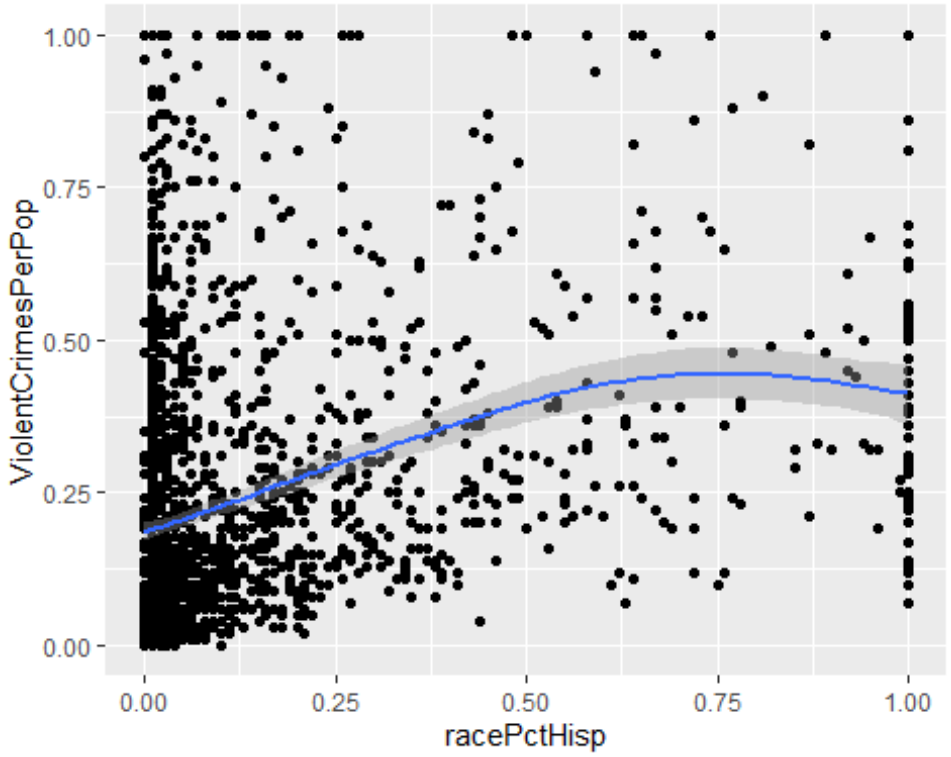
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



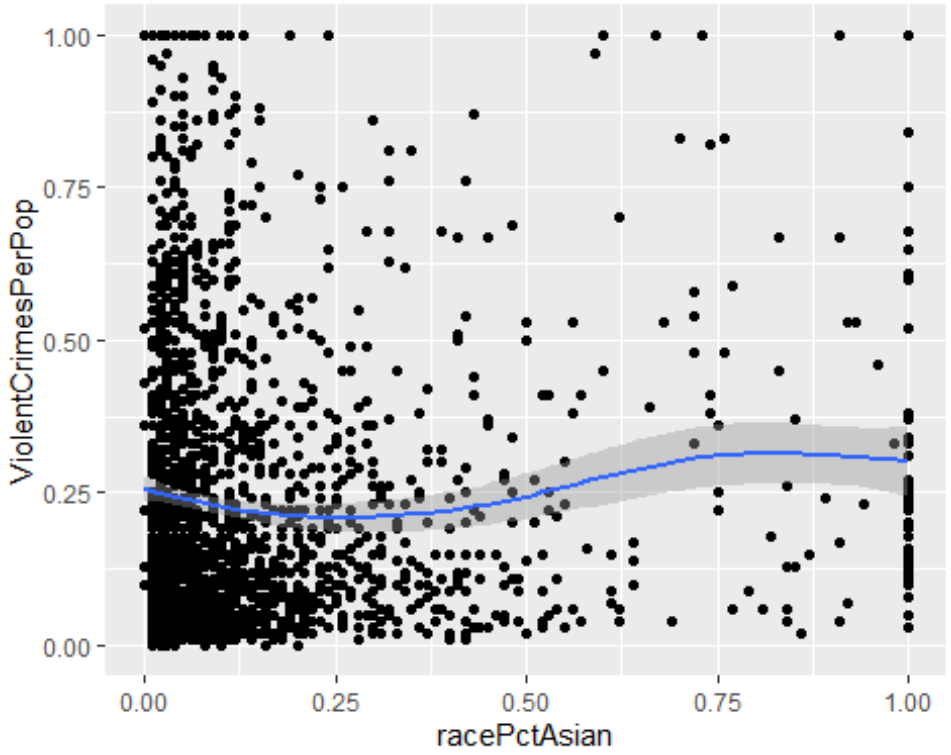

```
ggplot(finalp,aes(x=racePctWhite,y=ViolentCrimesPerPop))+
  geom_point()+
  geom_smooth()

## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




```
ggplot(finalp,aes(x=racePctHisp,y=ViolentCrimesPerPop))+
  geom_point()+
  geom_smooth()

## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




```
ggplot(finalp,aes(x=racePctAsian,y=ViolentCrimesPerPop))+
  geom_point()+
  geom_smooth()

## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



Analysis

For the Residuals vs. Fitted graph, since the data seems to follow a pattern, there might not be a linear relationship between the predictor and the outcome variables. On the Normal Q-Q, the points approximately fall along the reference line, so it is safe to assume normality. On the Scale-Location graph, the residuals are not spread equally along the ranges of predictors. They are all clustered to the right side.

According to the linear graphs that are based on race, there is an apparent correlation between higher value of white population and the lower value of crime.

Conclusions

According to the data as shown above, there are fewer violent crimes committed in predominantly white communities. The other analyses I would have liked to carry out is which violent crimes were the highest in which communities. It was would have been easier if there weren't any empty values, that way there wouldn't have been so much cleaning that needed to be done with the data.