Technical Appendix

Chase Henley, Harrison Marick, Joe Feldman (Group D)

4/29/2018

First, we read in our data obtained from College Scorecard:

```
prelim_data <- read.csv("https://awagaman.people.amherst.edu/stat225/datasetsS2018/groupDdata.csv")
```

After reading in the initial data, we need to do some data wrangling to select for variables of interest, remove observations (i.e. institutions) that are repetitive or have missing data, and make the data set grouped by insitution.

```
# Must start with a data-read in command from Prof. Wagaman
prelim_data2<-prelim_data[,c(-6, -18, -19, -(27:34), -(42:89))]
prelim_data2<-prelim_data2[,c(-2,-3,-5)] %>%
  as.data.frame()

#change some columns to type "character"
prelim_data2[,3:27]<-sapply(prelim_data2[,3:27], as.character)

#change some columns to type "numeric"
prelim_data2[,3:27]<-sapply(prelim_data2[,3:27], as.numeric)
prelim_data3 <-prelim_data2 %>%
  group_by(INSTNM) %>% #organize by institution name
  mutate(rank=min(UNITID), n=n()) %>%
  filter(n<2) %>% #ensure there are no repeated observations in data set
  as.data.frame()

#remove any last undesirable variables
data<-prelim_data3[,c(-1, -28, -29)]
```

Our final data frame, "data", is the data set we will proceed with in the ensuing analysis. Let's now take a look at our variables.

```
names(data)
```

```
##  [1] "INSTNM"                    "PCT_WHITE"
##  [3] "PCT_BLACK"                 "PCT_ASIAN"
##  [5] "PCT_HISPANIC"              "PCT_BA"
##  [7] "PCT_GRAD_PROF"             "PCT_BORN_US"
##  [9] "MEDIAN_HH_INC"             "POVERTY_RATE"
## [11] "UNEMP_RATE"                "LN_MEDIAN_HH_INC"
## [13] "MN_EARN_WNE_P10"           "MD_EARN_WNE_P10"
## [15] "PCT10_EARN_WNE_P10"        "PCT25_EARN_WNE_P10"
## [17] "PCT75_EARN_WNE_P10"        "PCT90_EARN_WNE_P10"
## [19] "SD_EARN_WNE_P10"           "GT_25K_P10"
## [21] "MN_EARN_WNE_INC1_P10"      "MN_EARN_WNE_INC2_P10"
## [23] "MN_EARN_WNE_INC3_P10"      "MN_EARN_WNE_INDEP0_INC1_P10"
## [25] "MN_EARN_WNE_INDEP0_P10"    "MN_EARN_WNE_INDEP1_P10"
```

Many of these variables are possible predictors of economic success, and there are plenty that are indicators of economic success. However, we decided to base our project on the following:

`MD_EARN_WNE_P10`: Median income 10 years after entry of students working and not enrolled in school;

quantitative. A measure of economic success.

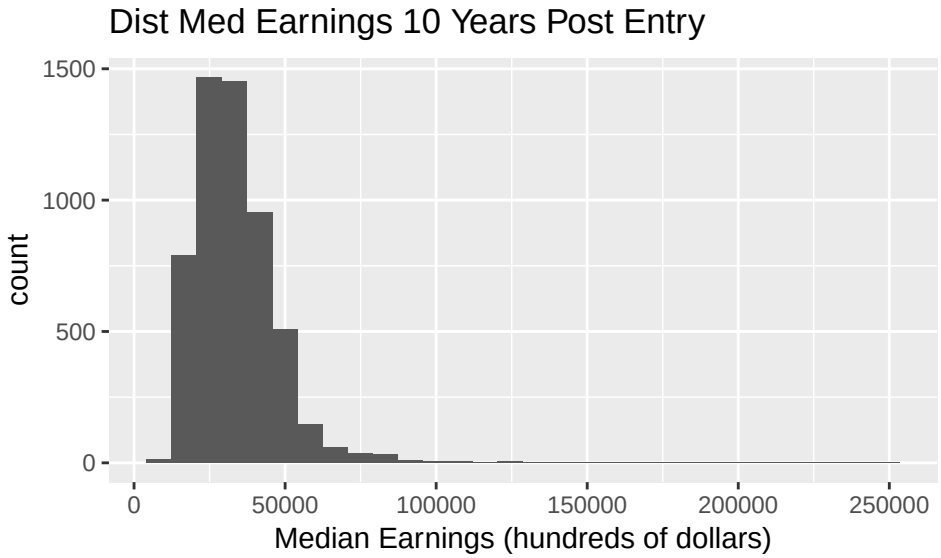
`PCT_WHITE`: Percent of the population from students' zip codes that is White, via Census data; quantitative. A possible predictor of economic success.

Let's first look at our response variable, `MD_EARN_WNE_P10`:

```
favstats(data$MD_EARN_WNE_P10)
```

```
##   min    Q1 median    Q3    max     mean       sd    n missing
##  9100 24100  31200 39750 250000 33625.62 15551.22 5499    1813
```

```
ggplot(data,aes(as.numeric(MD_EARN_WNE_P10)))+geom_histogram()+
  ggtitle("Dist Med Earnings 10 Years Post Entry")+
  xlab("Median Earnings (hundreds of dollars)")
```



The distribution of median earnings of students working and not enrolled 10 years after entry has a median of 31100 dollars and a mean of 33500.42 dollars, and so we see the distribution is skewed right. The range of the distribution is 9,100 to 250,000 dollars, and the standard deviation is 15,444.58 dollars which indicates a relatively large spread.

This variable will serve as a response variable for later in our analysis.

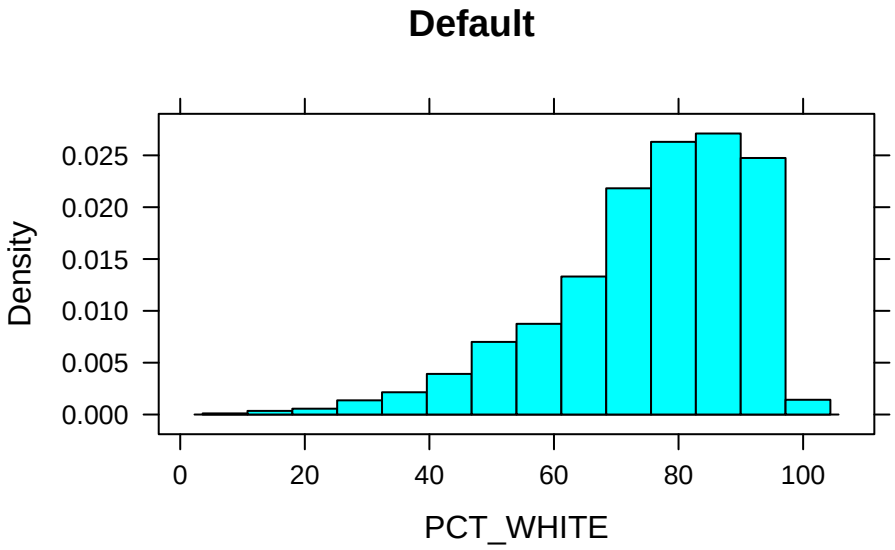
Let's now get a sense of the distribution of the `PCT_WHITE` variable, a possible predictor of post-graduation economic success:

```
#Univariate Analysis
favstats(data$PCT_WHITE)
```

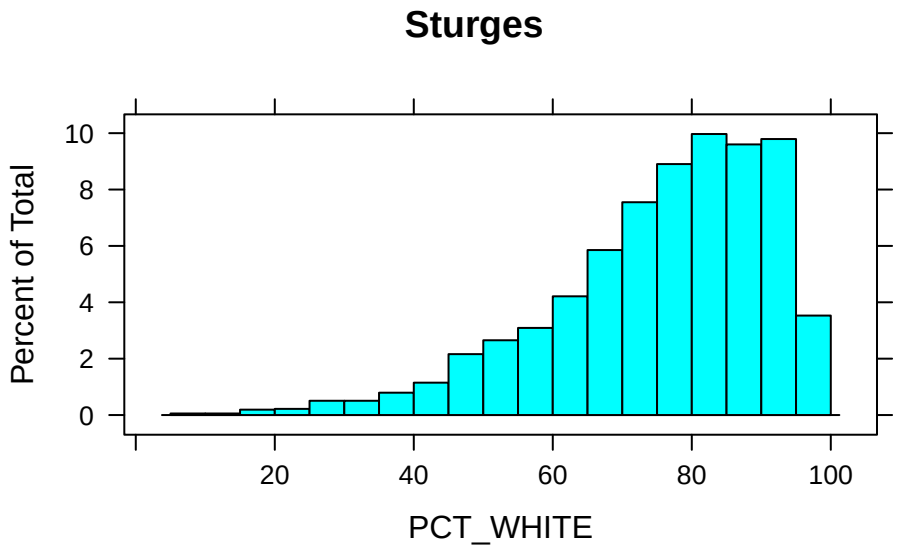
```
##   min    Q1 median    Q3   max     mean       sd    n missing
##  5.34 67.09 78.655 87.83 98.92 75.48166 15.86593 5176    2136
```

The mean percentage is 75.509 while the median is 78.685. The data ranges from from 5.34 to 98.98 with a standard deviation of 15.879. These numbers suggest the distribution is skewed left but let's see if this holds true visually:

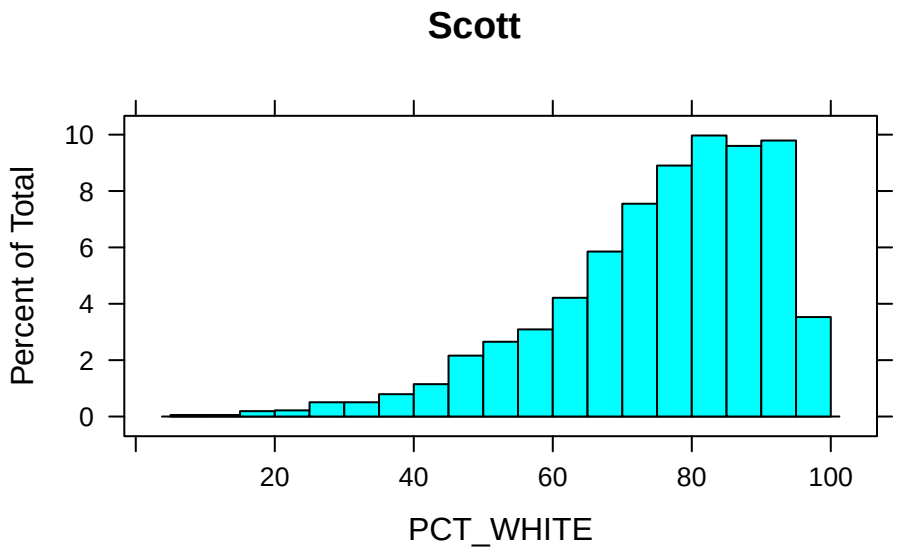
```
histogram(~PCT_WHITE,data=data,main="Default")
```



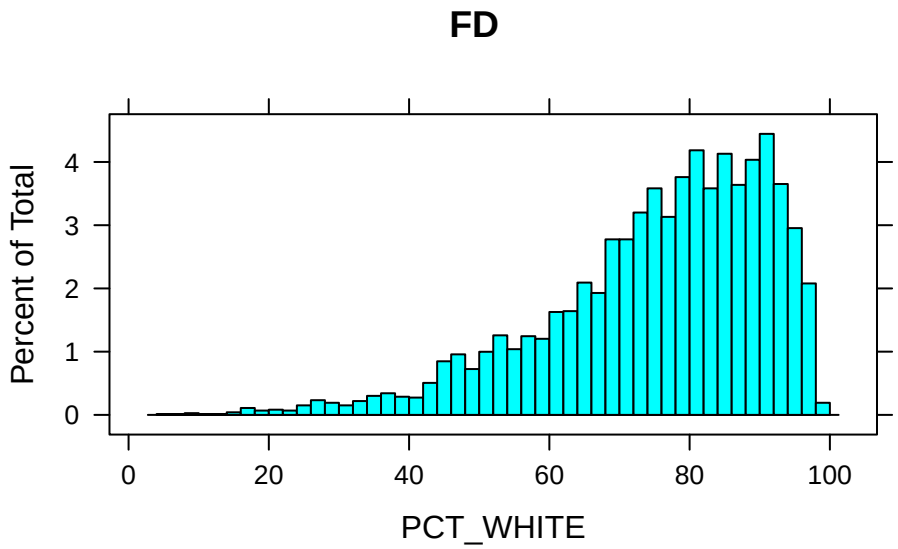

```
histogram(~PCT_WHITE,data=data,breaks="Sturges",main="Sturges")
```



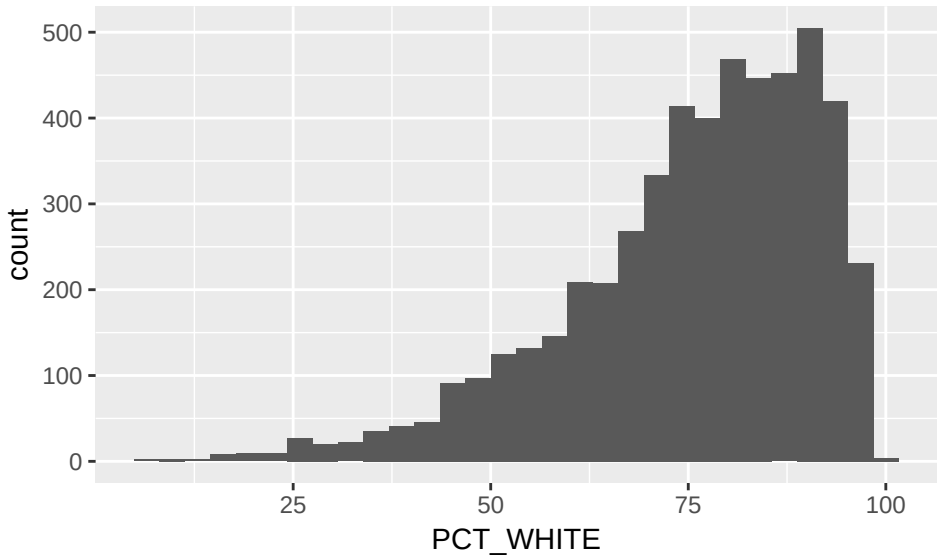

```
histogram(~PCT_WHITE,data=data,breaks="Scott",main="Scott")
```




```
histogram(~PCT_WHITE,data=data,breaks="FD",main="FD")
```




```
ggplot(data, aes(PCT_WHITE))+geom_histogram() #better graphics
```




```
#favstats(~PCT_WHITE, data=data4)
#FDh<-2*(87.83-67.09)*(5176^(-1/3)); FDh
#Scotth<-3.5*15.86593*(5176^(-1/3));Scotth
#Sturgesm<-log2(5176) +1; Sturgesm
#maxmin<-98.92+5.34
#FDm<-maxmin/FDh;FDm
#Scottm<-maxmin/Scotth;Scottm
```

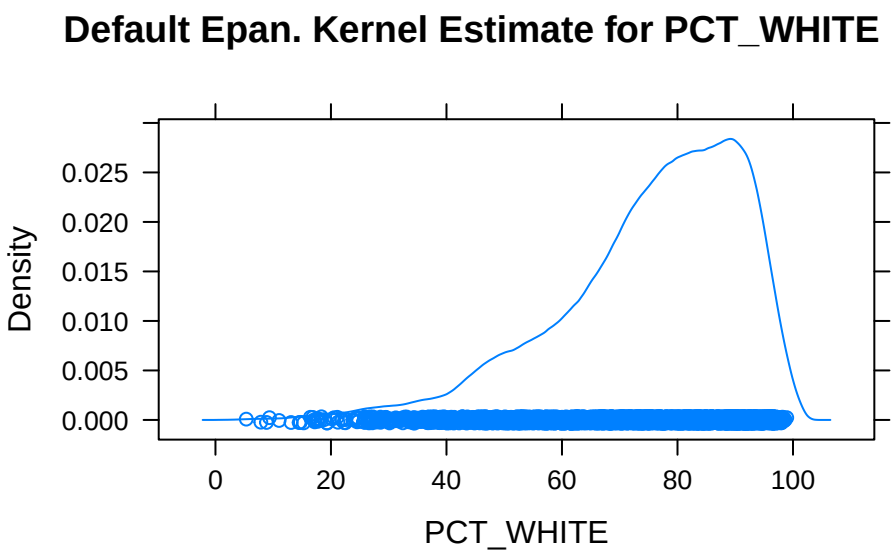
We do in fact see a strong skew left.

Friedman's produces a histogram with many bins and this makes it too blocky since it doesn't have enough observations per bin. The default histogram contains fewer bins and isn't a bad choice.

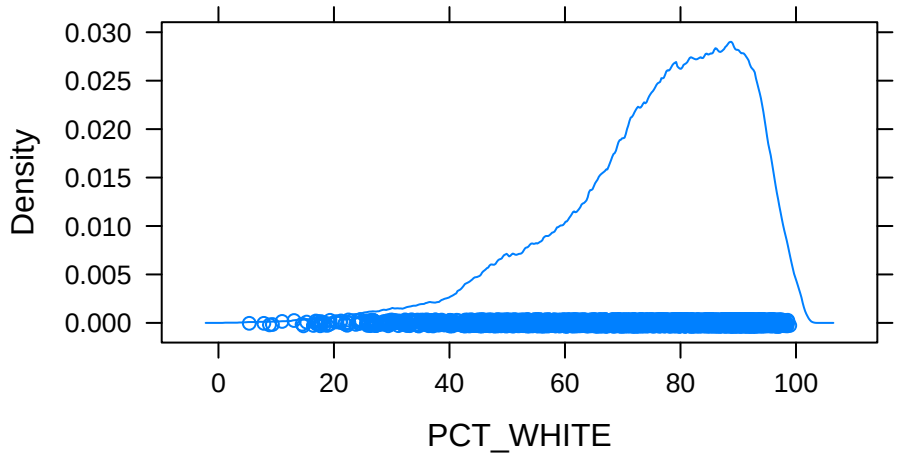
However, Sturges and Scott produce similar histograms and are our preferred choices since they portray the shape of the distribution best.

Since all of the histograms indicate a strong left skew of the percent white variable, so now we will do some density plot kernel comparisons to verify this:

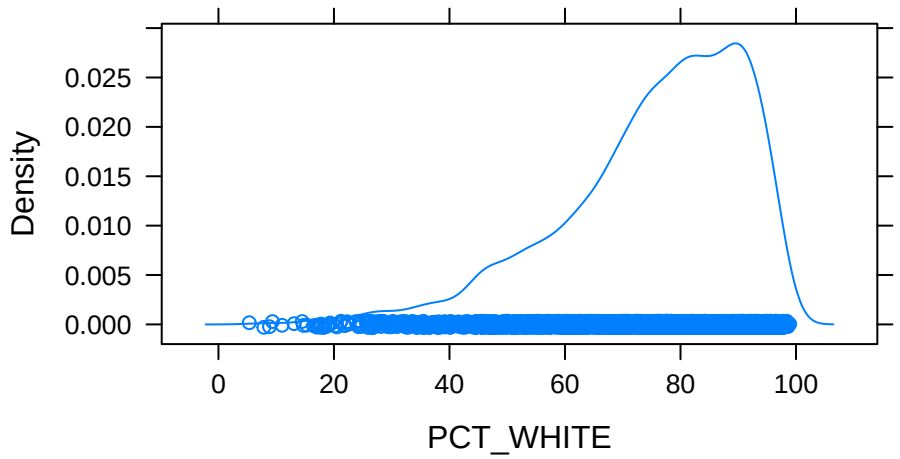
```
densityplot(~PCT_WHITE,data=data,kernel="e",main="Default Epan. Kernel Estimate for PCT_WHITE")
```



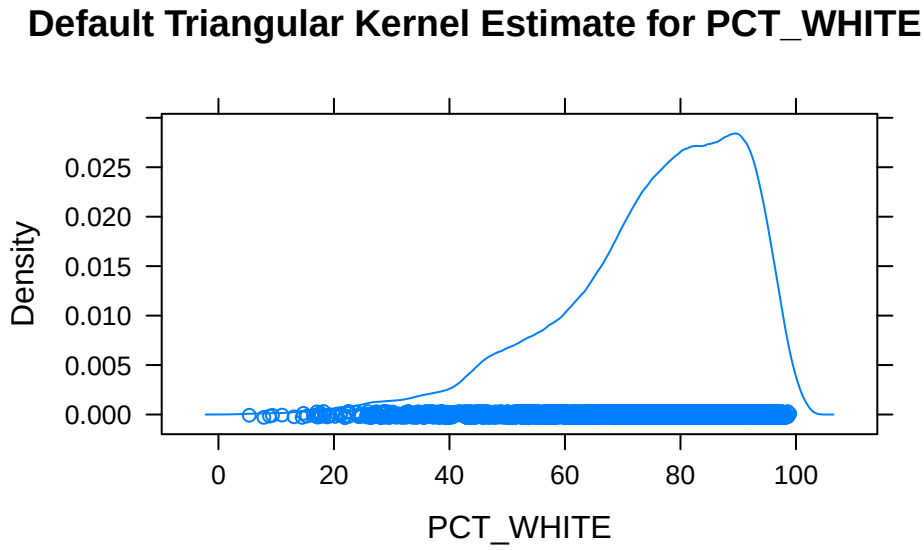
```
densityplot(~PCT_WHITE,data=data,kernel="r",main="Default Box Kernel Estimate for PCT_WHITE")
```




```
densityplot(~PCT_WHITE,data=data,kernel="g",main="Default Normal Kernel Estimate for PCT_WHITE")
```




```
densityplot(~PCT_WHITE,data=data,kernel="t",main="Default Triangular Kernel Estimate for PCT_WHITE")
```



All the density plots confirm a strong left skew.

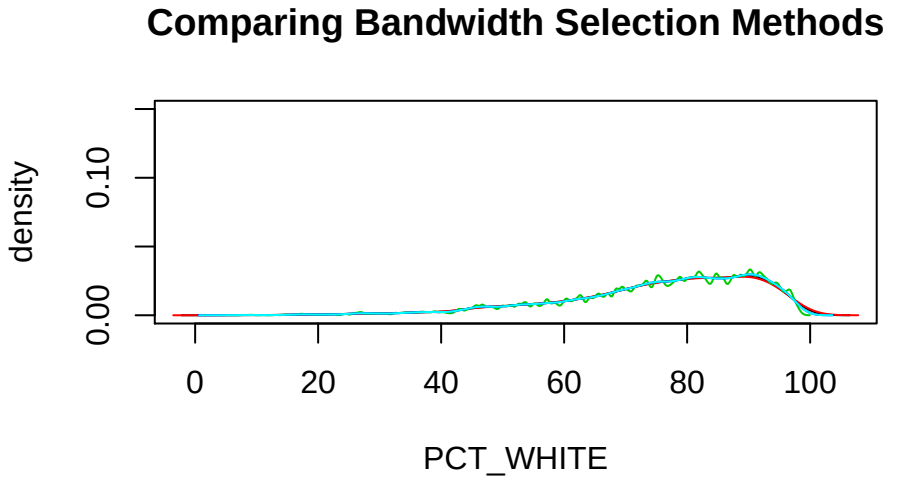
The density plot created using the box, or uniform kernel, is the least smooth out of the options. Plots constructed with the epanechnikov, triangular, and normal kernels are all fairly smooth, but the plot using the normal kernel stands out as the most smooth.

Therefore, the normal, or gaussian, kernel is the one we will prefer.

Finally, some bandwidth comparisons:

```
norm<-density(data$PCT_WHITE, kernel="g",bw="nrd0",na.rm=TRUE) #saved normal kernel, default h
norm2<-density(data$PCT_WHITE, kernel="g",bw="nrd",na.rm=TRUE)
norm3<-density(data$PCT_WHITE, kernel="g",bw="ucv",na.rm=TRUE)
norm4<-density(data$PCT_WHITE, kernel="g",bw="bcv",na.rm=TRUE)
norm5<-density(data$PCT_WHITE,kernel="g",bw="SJ",na.rm=TRUE)

plot(norm$x,norm$y,ylim=c(0,0.15),type="l",xlab="PCT_WHITE", ylab="density", main="Comparing Bandwidth S
lines(norm2$x,norm2$y,col=2) #red
lines(norm3$x,norm3$y,col=3) #green
lines(norm4$x,norm4$y,col=4) #blue
lines(norm5$x,norm5$y,col=5) #light blue
```



We see that the green and light blue lines are the least smooth—these were produced using the ucv and SJ

bandwidth selection methods, respectively. These are both data-driven cross-validation (cv) methods.

The black, red, and darker blue lines are all fairly similar in terms of smoothness, and these were produced with the nrd0, nrd, and ucv bandwidth selection methods. The nrd0 and nrd methods are both normal-based and work ideally with data sets that are normally distributed.‘ The ucv method is a cross-validation method as its name implies.

Thus, considering our data is not normally distributed, we prefer the ucv method in this setting.

Intuitively, the skewed distribution makes sense. White people have a traditionally stronger representation in the American college system, and this distribution supports that claim as there is more density around colleges with a high-percentage white student body.

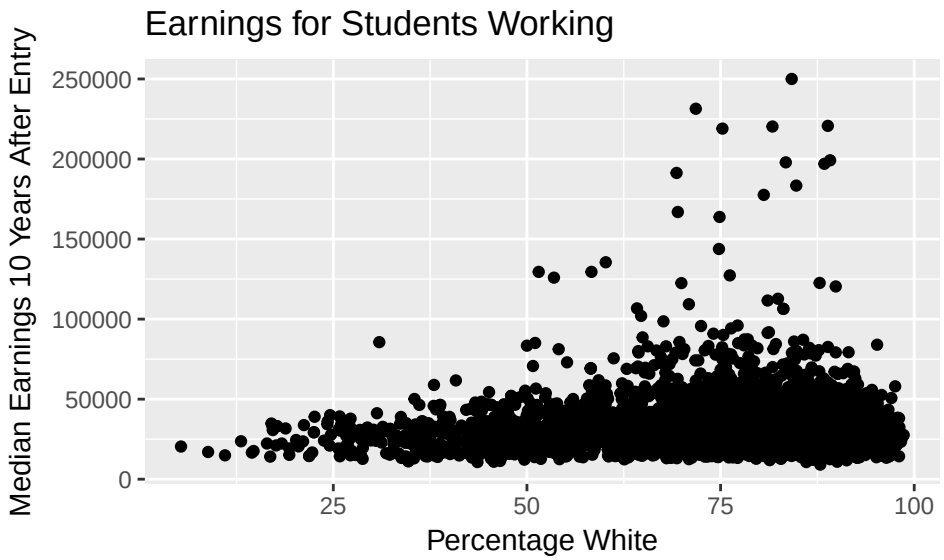
We decide to add another variable to distinguish institutions as being either predominantly white, or non-predominantly white.

```
data<-mutate(data, "PredominantlyWhite" = ifelse(PCT_WHITE> 67.09,1, 0))
```

The “cut-off”" we use is the value of the first quartile of the distribution of ‘PCT_WHITE’, meaning that all institutions with a value for this variable greater than Q1 are distinguished as predominantly white.

Let’s evaluate the relationship between ‘PCT_WHITE’ and ‘MD_EARN_WNE_P10’:

```
#remove missing observations
data<-filter(data, !is.na(PCT_WHITE) & !is.na(MD_EARN_WNE_P10))
ggplot(data, aes(PCT_WHITE, MD_EARN_WNE_P10)) + geom_point() + xlab("Percentage White") + ylab("Median
```



There appears to be a weak positive linear correlation between percentage of student body that is white and median earnings 10 years after college.

Let’s look at the difference in distributions between predominantly white and non-predominantly white institutions in terms of median income 10 years post-enrollment:

Joe, Insert overlayed densityplot here

We think these distributions are similar enough in shape and spread to assume a shift-based model.

Let’s use a t-test to see if there is a difference in distributions between predominantly white and non-predominantly white institutions in terms of median income 10 years post-enrollment:

```
t.test(MD_EARN_WNE_P10~PredominantlyWhite, data=data)
```

```
## MD_EARN_WNE_P10 ~ PredominantlyWhite
```

```
##
##  Welch Two Sample t-test
##
## data:  MD_EARN_WNE_P10 by PredominantlyWhite
## t = -10.377, df = 2623.3, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5875.991 -4008.179
## sample estimates:
## mean in group 0 mean in group 1
##        30545.85        35487.93
```

Our t-statistic is -10.377 and the p-value is close to 0 so we have sufficient evidence to reject the null hypothesis that the median income between the two groups were not equal, and conclude that students coming from institutions we distinguished as predominantly white had a higher median income ten years after enrollment.

Since we appear to be working with distributions that are non-normal, let's compare our t-test results with a nonparametric rank-sum test:

```
with(data=data, wilcox.test(PredominantlyWhite,MD_EARN_WNE_P10))
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  PredominantlyWhite and MD_EARN_WNE_P10
## W = 0, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

W=0 and we obtain a p-value close to 0, so the nonparametric procedure agrees with the results from the t-test: we can reject the null hypothesis and conclude that students coming from institutions we distinguished as predominantly white had a higher median income ten years after enrollment.

Now let's fit a simple linear model using percent white to predict median earnings.

```
summary(rfit(MD_EARN_WNE_P10~PCT_WHITE, data=data),overall.test="drop")
```

```
## Call:
## rfit.default(formula = MD_EARN_WNE_P10 ~ PCT_WHITE, data = data)
##
## Coefficients:
##              Estimate Std. Error t.value   p.value
## (Intercept) 23257.840    795.038  29.254 < 2.2e-16 ***
## PCT_WHITE     116.008     10.319  11.242 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Multiple R-squared (Robust): 0.02728063
## Reduction in Dispersion Test: 130.4127 p-value: 0
```

The p-value for model utility is very close to 0, indicating that there is a significant linear relationship between these two variables. This relationship is one we will continue to investigate as our project progresses.

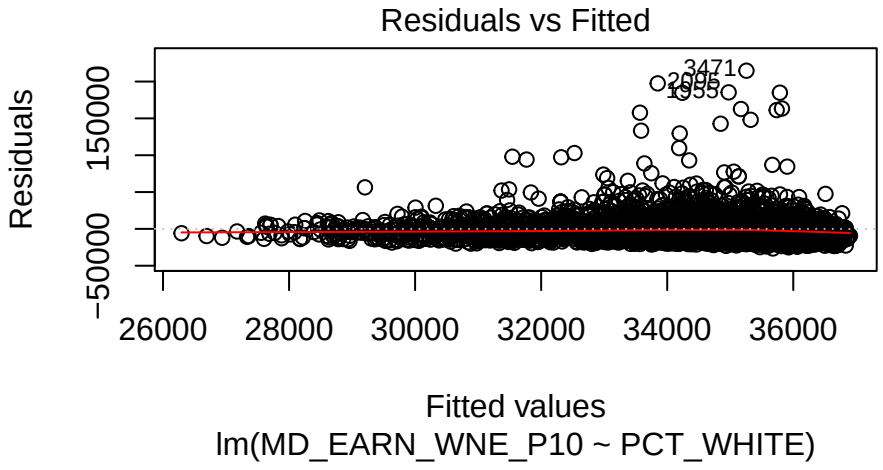
Let's fit a simple linear model using percent white to predict median earnings:

```
lin<-lm(MD_EARN_WNE_P10~PCT_WHITE, data=data)
summary(lin)
```

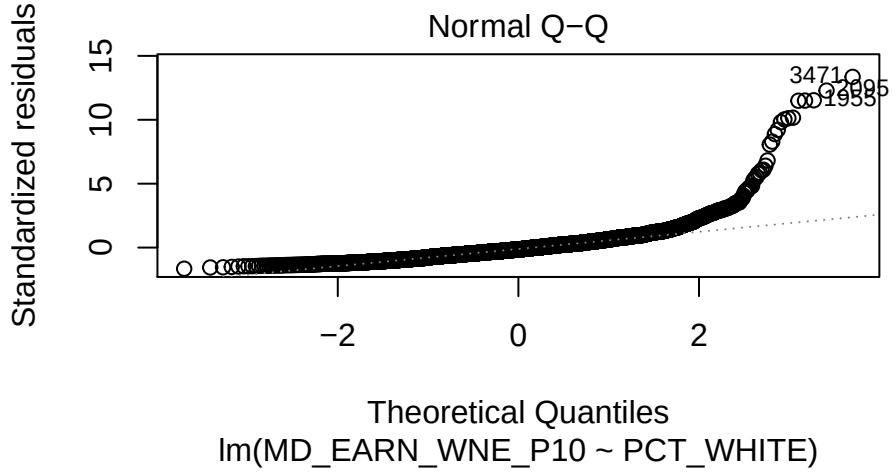
```
##
## Call:
```

```
## lm(formula = MD_EARN_WNE_P10 ~ PCT_WHITE, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -26576  -9009  -2248   5558 214743
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25686.44    1143.28  22.467  < 2e-16 ***
## PCT_WHITE     113.69      14.88   7.641  2.6e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16080 on 4650 degrees of freedom
## Multiple R-squared:  0.0124, Adjusted R-squared:  0.01219
## F-statistic: 58.38 on 1 and 4650 DF,  p-value: 2.603e-14
```

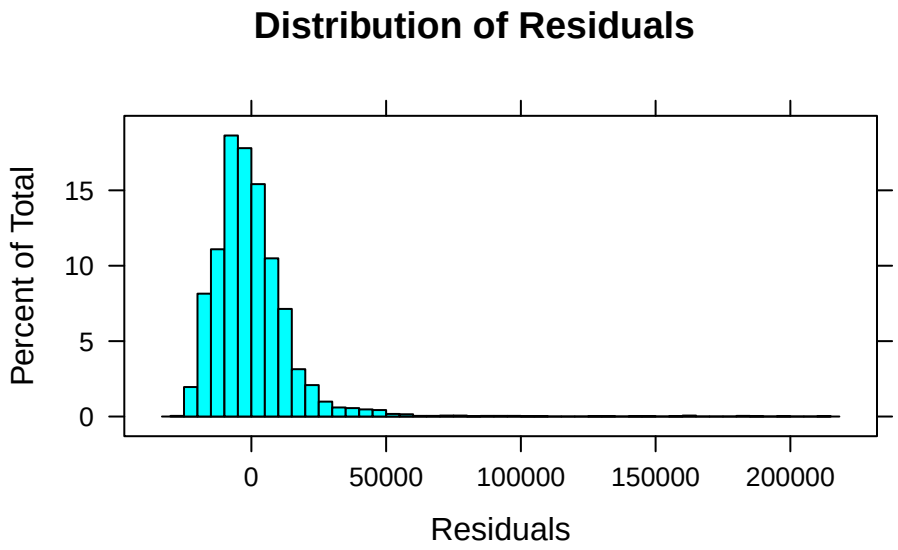
```
plot(lin, which=1)
```




```
plot(lin, which=2)
```




```
histogram(lin$residuals, breaks="Scott",xlab="Residuals",main="Distribution of Residuals") #giant right
```



Our F-statistic of 58.38 on 1 and 4650 DF gives us a p-value close to 0, so our overall model is useful for predicting median earnings.

However, we see the conditions for our parametric SLR do not appear to check out. The residuals vs fitted plot shows unequal variance, the normal qq plot has a major deviation in the tail, and we see from the distribution of residuals that the residuals are in fact skewed right heavily.

Since our conditions don't check out, let's fit a nonparametric linear model using percent white to predict median earnings.

```
summary(rfit(MD_EARN_WNE_P10~PCT_WHITE, data=data),overall.test="drop")
```

```
## Call:
## rfit.default(formula = MD_EARN_WNE_P10 ~ PCT_WHITE, data = data)
##
## Coefficients:
##              Estimate Std. Error t.value   p.value
## (Intercept) 23257.840    795.038  29.254 < 2.2e-16 ***
## PCT_WHITE     116.008     10.319  11.242 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Multiple R-squared (Robust): 0.02728063
## Reduction in Dispersion Test: 130.4127 p-value: 0
```

The p-value for model utility is very close to 0, indicating that there is a significant linear relationship between these two variables.

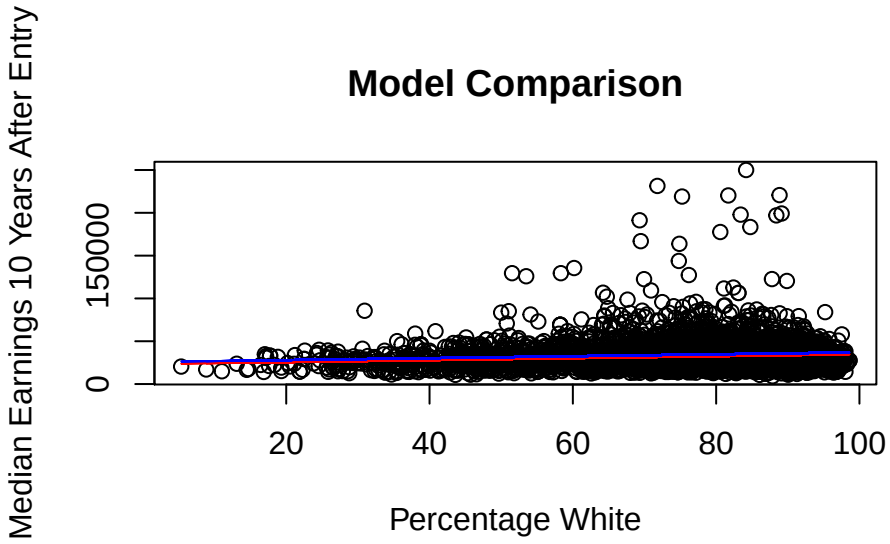
Let's see how the parametric and nonparametric models compare:

```
nonP<-rfit(MD_EARN_WNE_P10~PCT_WHITE, data=data)
summary(nonP, overall.test='drop')
```

```
## Call:
## rfit.default(formula = MD_EARN_WNE_P10 ~ PCT_WHITE, data = data)
##
## Coefficients:
##              Estimate Std. Error t.value   p.value
## (Intercept) 23257.840    795.038  29.254 < 2.2e-16 ***
## PCT_WHITE     116.008     10.319  11.242 < 2.2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Multiple R-squared (Robust): 0.02728063
## Reduction in Dispersion Test: 130.4127 p-value: 0
```

```r
plot(data$PCT_WHITE, data$MD_EARN_WNE_P10, xlab="Percentage White", ylab="Median Earnings 10 Years Afte
lines((nonP$x)[,2], nonP$fitted, col=2)
lines(data$PCT_WHITE, lin$fitted, col=4)
```



The parametric estimated slope is 113.69 which is very close to the nonparametric slope of 116.01.

We see that these slopes produce lines that are very similar and are right on top of each other (intercepts are close to each other), and they both indicate a slight positive relationship between Percentage white and Median Earnings 10 Years Post-Entry.