

STAT 225 - Prelim Project Look Group D

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

What is the True Value of a College Education?

Perhaps the most prevalent rationale for the pursuance of a college degree is the promise of economic mobility upon graduation. However, our project aims to investigate whether this opportunity is equally available for all students. Specifically, our group will analyze whether the racial and socio-demographic composition of a college affects the proportion of students who go on to earn higher wages after enrolling in the institution.

Read in the data

```
data<-read.csv("Scorecard.csv")
data2<-data[,c(-6, -18, -19, -(27:34), -(42:89))]
data2<-data2[,c(-2,-3,-5)] %>%
  as.data.frame() # Must start with a data-read in command from Prof. Wagaman
data2[,3:27]<-sapply(data2[,3:27], as.character)
data2[,3:27]<-sapply(data2[,3:27], as.numeric)
data3 <-data2 %>%
  group_by(INSTNM) %>%
  mutate(rank=min(UNITID), n=n()) %>%
  filter(n<2) %>%
  as.data.frame()
data4<-data3[,c(-1, -28, -29)]
```

Summary command on data set

```
summary(data4)
```

```
##                               INSTNM          PCT_WHITE
## A & W Healthcare Educators      :    1   Min.      : 5.34
## A T Still University of Health Sciences:    1   1st Qu.:67.09
## Aaniiih Nakoda College         :    1   Median   :78.66
## Aaron's Academy of Beauty      :    1   Mean     :75.48
## ABC Beauty Academy             :    1   3rd Qu.:87.83
## ABC Beauty College Inc         :    1   Max.     :98.92
## (Other)                        :7306   NA's     :2136
##   PCT_BLACK      PCT_ASIAN      PCT_HISPANIC      PCT_BA
## Min.   : 0.020   Min.   : 0.060   Min.   : 0.410   Min.   : 3.00
## 1st Qu.: 4.150   1st Qu.: 0.900   1st Qu.: 2.700   1st Qu.:11.41
## Median : 8.345   Median : 1.785   Median : 5.865   Median :14.08
## Mean   :12.947   Mean   : 3.028   Mean   :12.957   Mean   :14.34
## 3rd Qu.:17.720   3rd Qu.: 3.470   3rd Qu.:15.120   3rd Qu.:16.96
## Max.   :85.820   Max.   :51.720   Max.   :99.350   Max.   :30.70
## NA's   :2136    NA's   :2136    NA's   :2136    NA's   :2173
##   PCT_GRAD_PROF      PCT_BORN_US      MEDIAN_HH_INC      POVERTY_RATE
## Min.   : 0.600   Min.   : 28.92   Min.   : 15429   Min.   : 3.010
## 1st Qu.: 5.880   1st Qu.: 85.11   1st Qu.: 49663   1st Qu.: 6.987
## Median : 7.340   Median : 92.54   Median : 57740   Median : 8.950
## Mean   : 7.863   Mean   : 88.87   Mean   : 57912   Mean   :10.763
## 3rd Qu.: 9.330   3rd Qu.: 96.14   3rd Qu.: 65851   3rd Qu.:12.180
## Max.   :24.360   Max.   :100.00   Max.   :100871   Max.   :55.280
```

```

## NA's :2173      NA's :2173      NA's :2136      NA's :2136
## UNEMP_RATE      LN_MEDIAN_HH_INC MN_EARN_WNE_P10 MD_EARN_WNE_P10
## Min. : 1.850    Min. : 9.64    Min. : 12200   Min. : 9100
## 1st Qu.: 3.100   1st Qu.:10.78   1st Qu.: 27000   1st Qu.: 24100
## Median : 3.550   Median :10.92   Median : 35100   Median : 31200
## Mean : 3.847     Mean :10.90     Mean : 37854     Mean : 33626
## 3rd Qu.: 4.240   3rd Qu.:11.05   3rd Qu.: 44000   3rd Qu.: 39750
## Max. :15.360     Max. :11.49     Max. :250000     Max. :250000
## NA's :2136      NA's :2136      NA's :1813      NA's :1813
## PCT10_EARN_WNE_P10 PCT25_EARN_WNE_P10 PCT75_EARN_WNE_P10
## Min. : 800      Min. : 4600     Min. : 18700
## 1st Qu.: 4800    1st Qu.: 13300   1st Qu.: 37750
## Median : 7000    Median : 17400   Median : 47600
## Mean : 8247      Mean : 19798     Mean : 50487
## 3rd Qu.:10600    3rd Qu.: 24600   3rd Qu.: 57600
## Max. :61200      Max. :173300     Max. :250000
## NA's :2743      NA's :2301      NA's :2301
## PCT90_EARN_WNE_P10 SD_EARN_WNE_P10 GT_25K_P10 MN_EARN_WNE_INC1_P10
## Min. : 25100     Min. : 9600      Min. :0.0900    Min. : 14100
## 1st Qu.: 54500    1st Qu.: 19500   1st Qu.:0.4840   1st Qu.: 26400
## Median : 66300    Median : 25200   Median :0.6160   Median : 32200
## Mean : 69828      Mean : 27826     Mean :0.6029     Mean : 35185
## 3rd Qu.: 79000    3rd Qu.: 31600   3rd Qu.:0.7420   3rd Qu.: 40475
## Max. :250000      Max. :172700     Max. :0.9770     Max. :182900
## NA's :2743      NA's :1813      NA's :1813      NA's :2806
## MN_EARN_WNE_INC2_P10 MN_EARN_WNE_INC3_P10 MN_EARN_WNE_INDEP0_INC1_P10
## Min. : 19900      Min. : 18800     Min. : 17400
## 1st Qu.: 35900     1st Qu.: 39200   1st Qu.: 29100
## Median : 41600     Median : 45900   Median : 34000
## Mean : 43007       Mean : 47298     Mean : 35368
## 3rd Qu.: 47875     3rd Qu.: 52575   3rd Qu.: 39400
## Max. :135900       Max. :155500     Max. :120900
## NA's :3994        NA's :3982      NA's :4800
## MN_EARN_WNE_INDEP0_P10 MN_EARN_WNE_INDEP1_P10
## Min. : 13300      Min. : 14900
## 1st Qu.: 29300     1st Qu.: 28200
## Median : 35900     Median : 35600
## Mean : 37237       Mean : 38627
## 3rd Qu.: 42800     3rd Qu.: 45100
## Max. :130300       Max. :193200
## NA's :3167        NA's :3167

```

Data Codebook

PCT_WHITE: Percent of the population from students' zip codes that is White, via Census data; quantitative

MEDIAN_HH_INC: Median Household Income of Students' Parents; quantitative.

GT_25K_P10: Percentage of students earning at least 25k 10 years after entry; quantitative

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

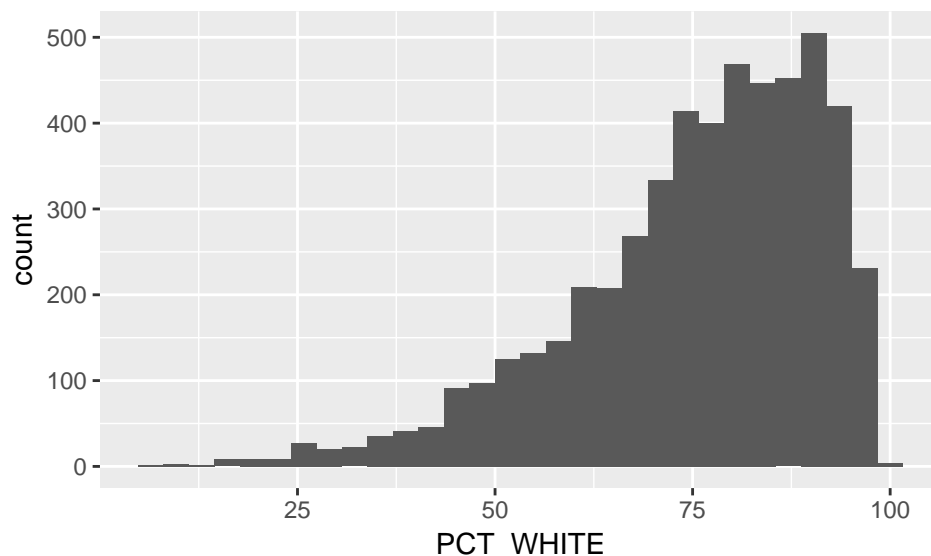
Analysis Plan

We plan to focus our project on explaining the relationships between MD_EARN_WNE_P10 and various predictor variables. Below, we have highlighted a few potential predictors, their distributions, and their relationships with our primary response variable.

Prelim Univariate Analysis

Percentage White

```
#Univariate Analysis
#favstats(data4$PCT_WHITE)
#histogram(data4$PCT_WHITE)
ggplot(data4, aes(PCT_WHITE))+geom_histogram()
```



```
#densityplot(data4$PCT_WHITE)
```

Conducting Univariate Analysis on the distribution of the percentage of students that are white, it is clear that this distribution is quite skewed left. The mean percentage is 75.509 while the median is 78.685. The data ranges from 5.34 to 98.98 with a standard deviation of 15.879.

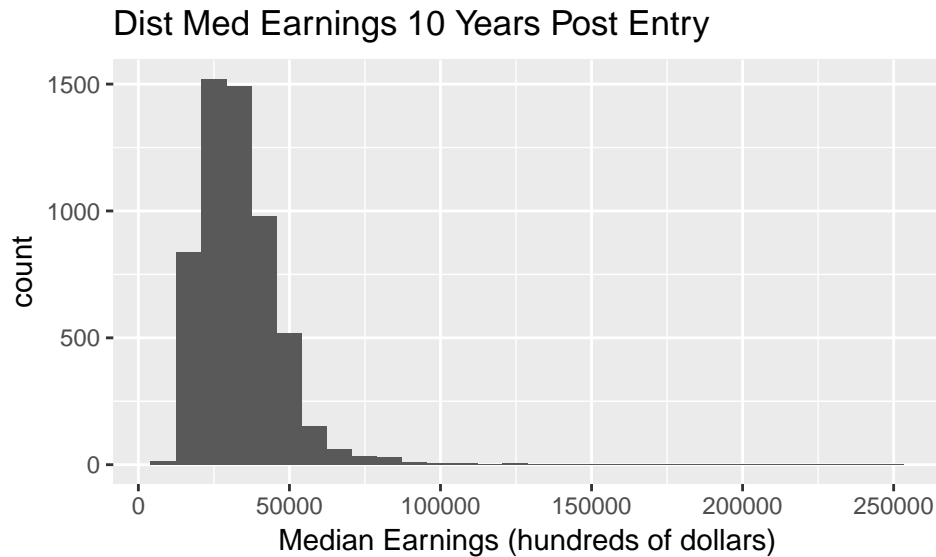
Intuitively, this distribution makes sense. White people have a traditionally stronger representation in the American college system, and this distribution supports that claim as there is more density around colleges with a proportionately high white student body.

Median Earnings 10 Years Post College

```
favstats(data2$MD_EARN_WNE_P10)
```

```
##   min    Q1 median    Q3   max   mean    sd   n missing
##   9100 24100 31100 39500 250000 33500.42 15444.58 5682    1911
```

```
ggplot(data2, aes(as.numeric(MD_EARN_WNE_P10)))+geom_histogram()+
  ggtitle("Dist Med Earnings 10 Years Post Entry")+
  xlab("Median Earnings (hundreds of dollars)")
```



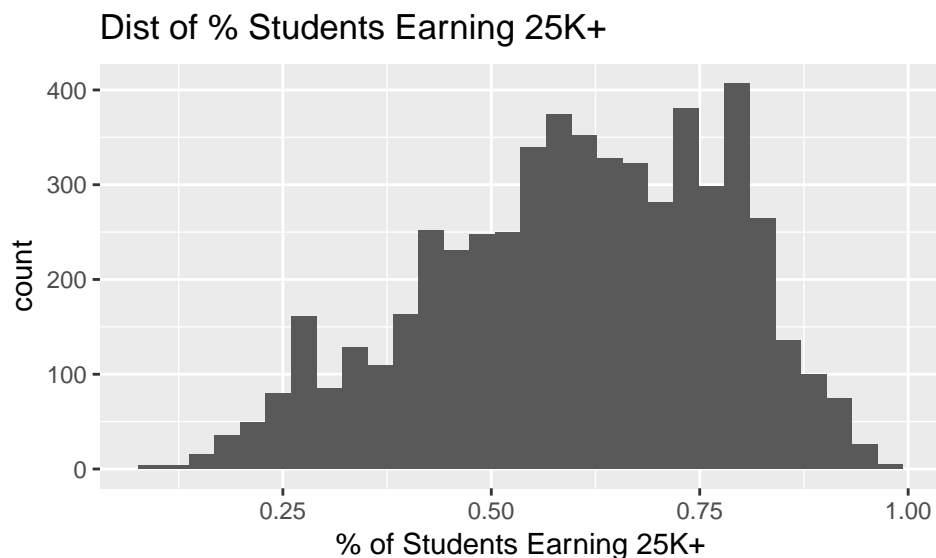
The distribution of median earnings of students working and not enrolled 10 years after entry has a median of 27,800 dollars and a mean of 33,055.8 dollars, and so we see the distribution is skewed right. There is a particularly high concentration (over 1750 observations) with median earnings above 60,000 dollars, while the next highest concentration consists of less about 500 observations—this is interesting and worth exploring further in the future. The range of the distribution is 1,000 to 61,600 dollars, and the standard deviation is 19,470.25 dollars which indicates a relatively large spread.

% of Students Earning At Least 25K

```
favstats(data4$GT_25K_P10)
```

```
##   min    Q1 median    Q3   max    mean      sd    n missing
##  0.09 0.484  0.616 0.742 0.977 0.6029171 0.1748034 5499    1813
```

```
ggplot(data4,aes((GT_25K_P10)))+
  geom_histogram()+ggtitle("Dist of % Students Earning 25K+")+
  xlab("% of Students Earning 25K+")
```



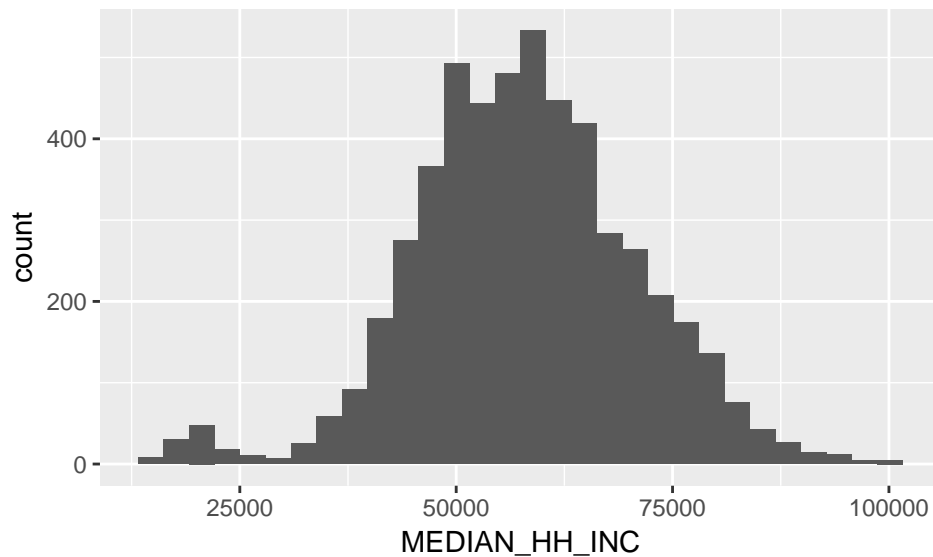
The distribution share of students earning over \$25,000/year (threshold earnings) 10 years after entry has a

median of 0.616 and a mean of 0.602, and we see the distribution is skewed left. The standard deviation is 0.175 which indicates a fairly large spread.

Median Household Income

As expected, the distribution of this variable is primarily bell-shaped. While there are a couple of spikes, there is one primary peak and it is relatively symmetric around a median of 57,739.61. The values for this variable range from 15,429.01 to 100,870.80.

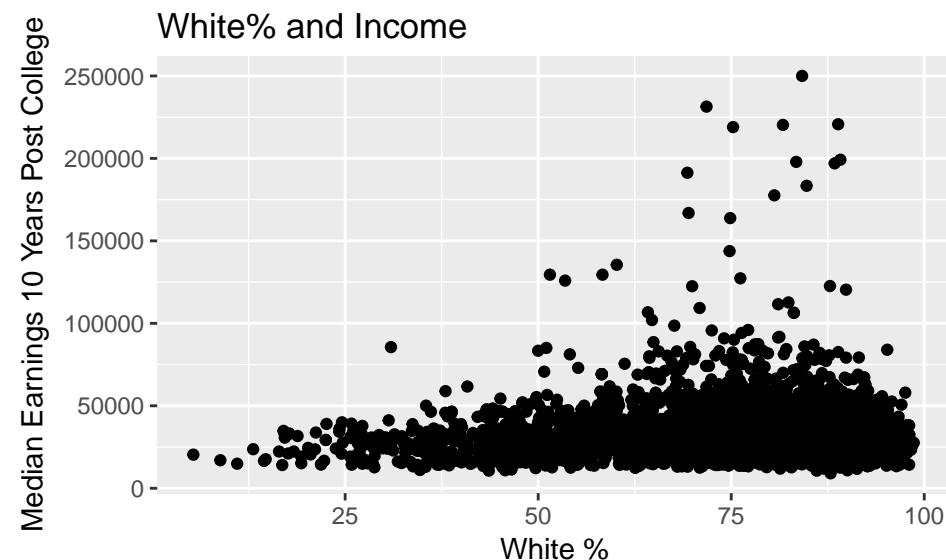
```
ggplot(data4, aes(MEDIAN_HH_INC)) + geom_histogram()
```



Prelim Bivariate Analysis

Percentage White and Median Earnings 10 Years Post College

```
ggplot(data4, aes(PCT_WHITE, MD_EARN_WNE_P10)) + geom_point() + xlab("White %") +  
ylab("Median Earnings 10 Years Post College") + ggtitle("White% and Income")
```



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

```
summary(rfit(MD_EARN_WNE_P10~PCT_WHITE, data=data4),overall.test="drop")
```

```
## Call:
## rfit.default(formula = MD_EARN_WNE_P10 ~ PCT_WHITE, data = data4)
##
## 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
```

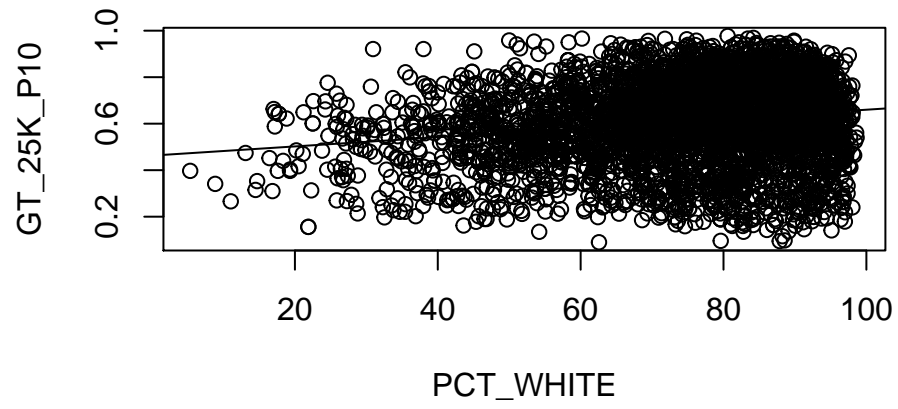
Above, we have fit a simple linear model using percent white to predict median earnings. Most notably, the p-value for model utility is very close to 0, indicating that there is a significant linear relationship between these two variables. This relationship is one we will continue to investigate as our project progresses.

Percentage White and % Student Body Earning >25k

```
mod<-lm(GT_25K_P10~PCT_WHITE, data = data4)
summary(mod)
```

```
##
## Call:
## lm(formula = GT_25K_P10 ~ PCT_WHITE, data = data4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.54106 -0.10740  0.01677  0.13102  0.39726
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.4626911  0.0122237  37.85  <2e-16 ***
## PCT_WHITE    0.0019731  0.0001591  12.40  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1719 on 4650 degrees of freedom
## (2660 observations deleted due to missingness)
## Multiple R-squared:  0.03202,    Adjusted R-squared:  0.03181
## F-statistic: 153.8 on 1 and 4650 DF,  p-value: < 2.2e-16

with(data4,plot(PCT_WHITE,GT_25K_P10))
abline(mod)
```



There seems to be a weak to moderate positive relationship between PCT-WHITE and the proportion of students earning over 25K within 10 years of graduating.