Midterm 2

Evan Tiffany

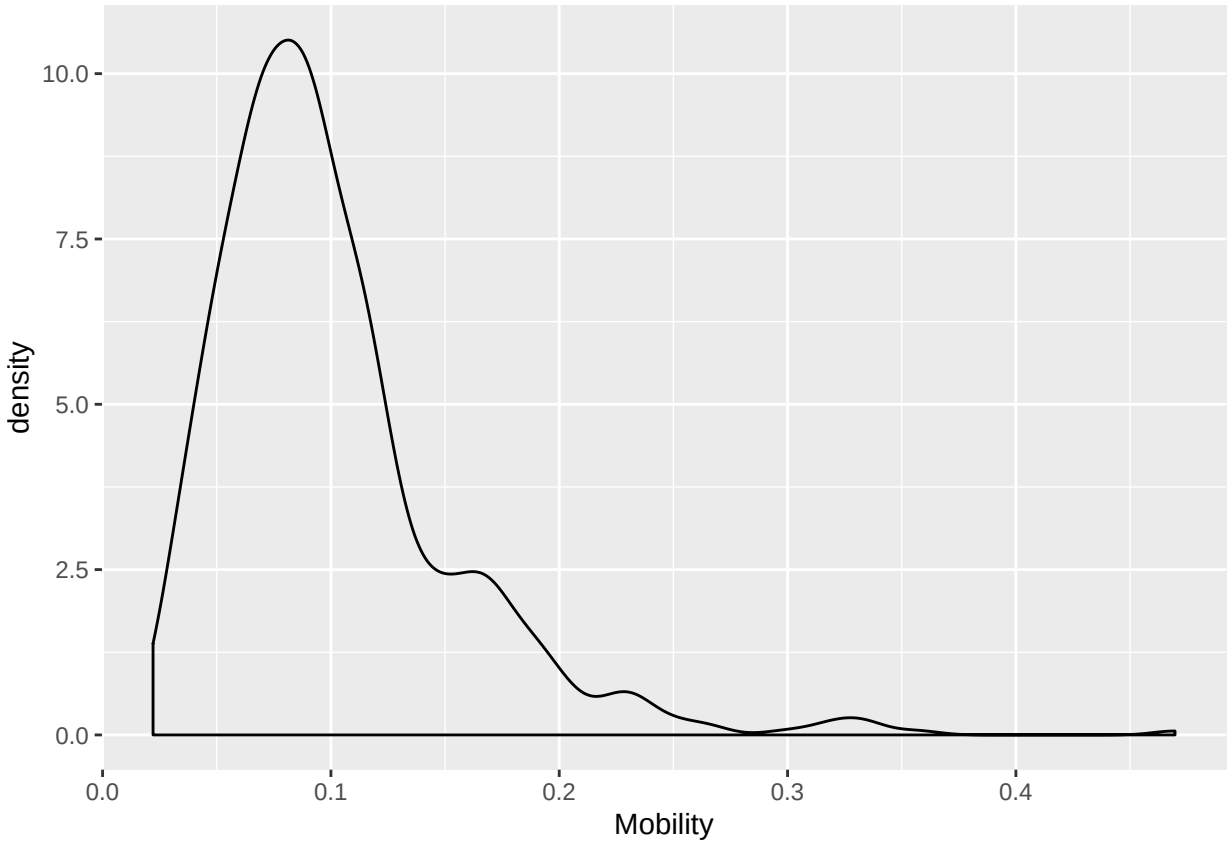
12/5/2019

Introduction

To analyze the economic and geographical effects that hometowns have on the future success of individuals, this report will utilize data obtained by Chetty, Raj, Nathaniel Hendren, Patrick Kline and Emmanuel Saez (2014). This data was originally used in the report "Where is the Landof Opportunity? The Geography of Intergenerational Mobility in the United States." published in the Quarterly Journal of Economics. The dataset contains 43 variables with different types of background information for each of the 741 cities chosen. This includes data on education, race, crime, home life, the economy, and much more. With this data, I aim to create a model that can accurately predict the possibility of success of an individual based on his or her situation growing up. Through this model, it should be evident in which areas of the United States growth is most present as well as what other variables play into an individual's economic growth.

Exploratory Data Analysis

```
## Warning: Removed 10 rows containing non-finite values (stat_density).
```

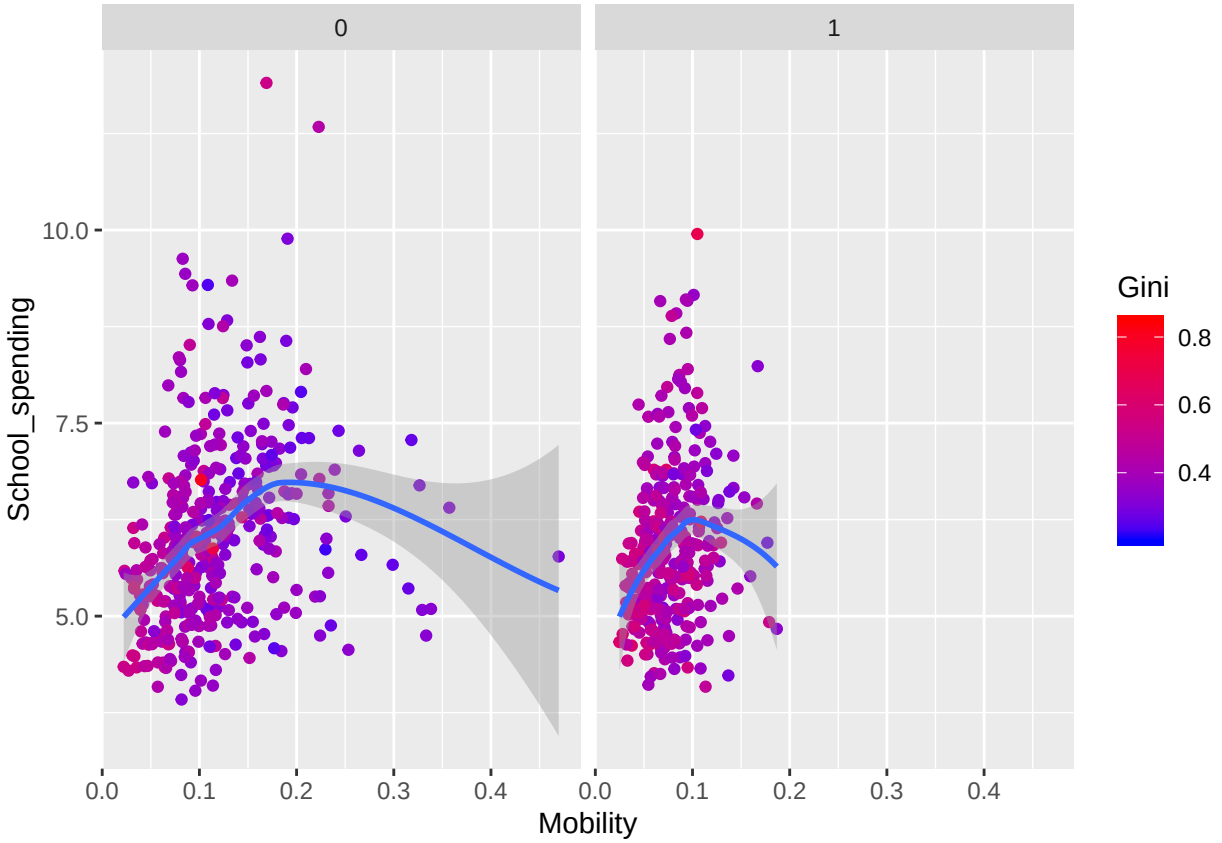


As we see in the density plot above, the data on mobility is right skewed with its peak around 0.1. This shows how unlikely mobility is across the board.

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
## Warning: Removed 13 rows containing non-finite values (stat_smooth).
```

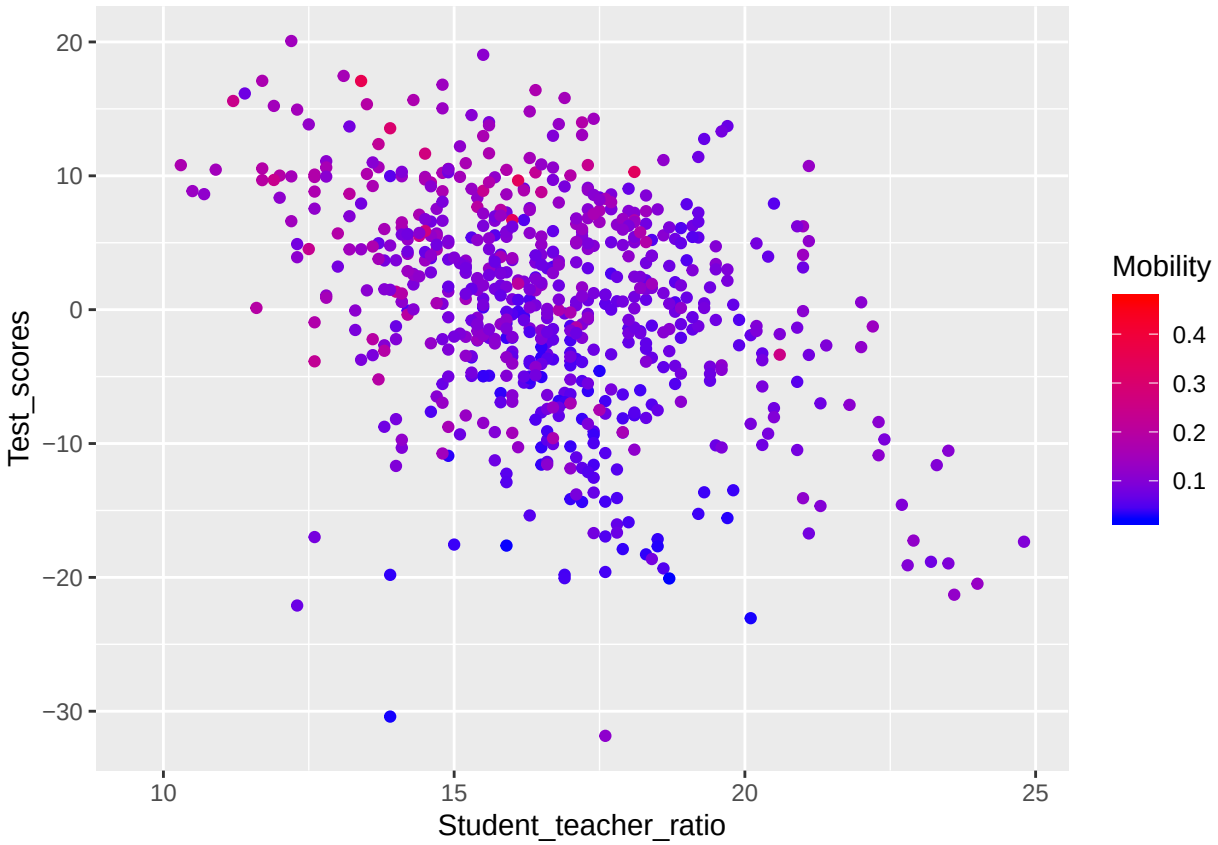
```
## Warning: Removed 13 rows containing missing values (geom_point).
```



In this plot we see not only that school spending has an increasing relationship with mobility, appearing to have diminishing returns which could partially be due to the right skew of the mobility data. Additionally, the inclusion of the Gini index shows a lot lower numbers in the rural data and an overall trend toward lower gini indices when mobility increases.

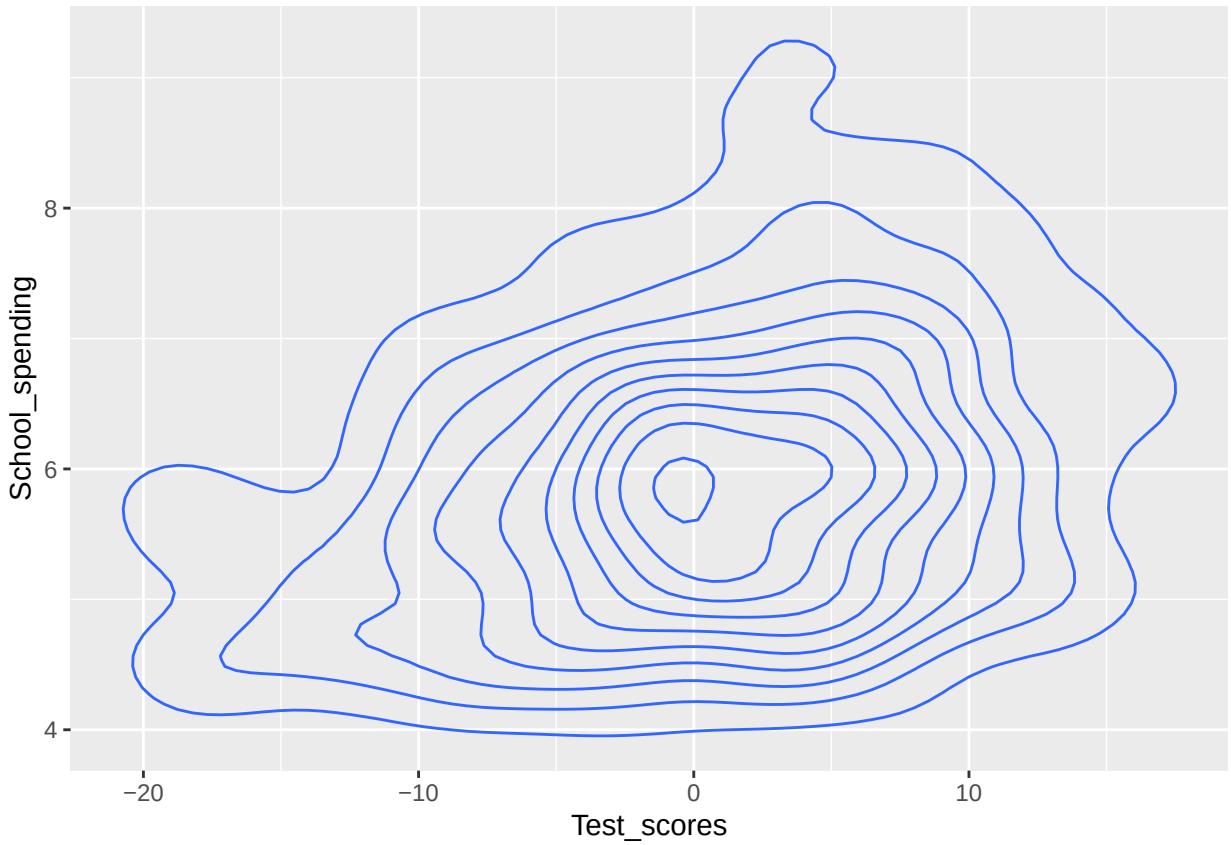
```
ggplot(mob, aes(x = Student_teacher_ratio, y = Test_scores, color = Mobility)) + geom_point() + scale_c
```

```
## Warning: Removed 59 rows containing missing values (geom_point).
```



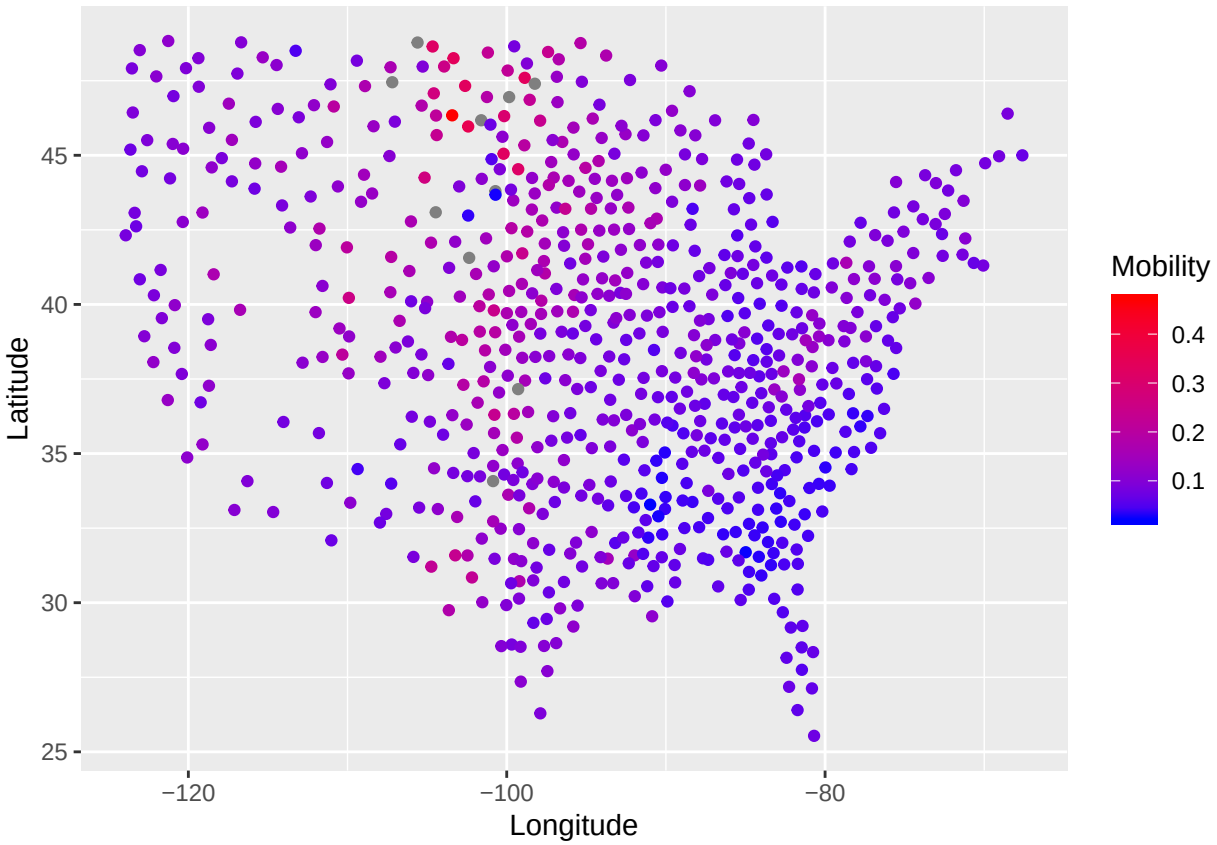
This plot shows that lower student/teacher ratios provide for higher test scores. Additionally the gradient aims to show each explanatory variables effect on mobility. There appears to be very little relationship between student/teacher ratio and mobility but its clear that mobility increases with test scores.

```
## Warning: Removed 33 rows containing non-finite values (stat_density2d).
```



With this density plot we can see the highest density around the average test score, for obvious reasons, but also at a school spending level of 6. Overall these two predictors appear to both have relatively normal distributions.

```
ggplot(mob, aes(x = Longitude, y = Latitude, color = Mobility)) + geom_point() + scale_color_gradient(l
```



This scatterplot shows the highest mobility in the nothern part of the country. The coastal areas seem to have less mobility which makes sense. As shown prior, the urban areas have lower mobility, which would be due to the high population and, going along with the predictors we've seen, a higher student/teacher ratio which leads to lower test scores. The lowest areas are shown to be in the southeast, around Alabama, Florida, and Georgia. These areas are known to have lower achieving schools which would account for the low mobility.

```
## Loading required package: magrittr
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
## [1] "Mobility and Population"
```

```
## [1] "Pearson"
```

```
## [1] -0.134157
```

```
## [1] "Kendall"
```

```
## [1] -0.3272999
```

```
## [1] "Spearman"
```

```
## [1] -0.4794005
```

```
## [1] "Mobility and Income"
```

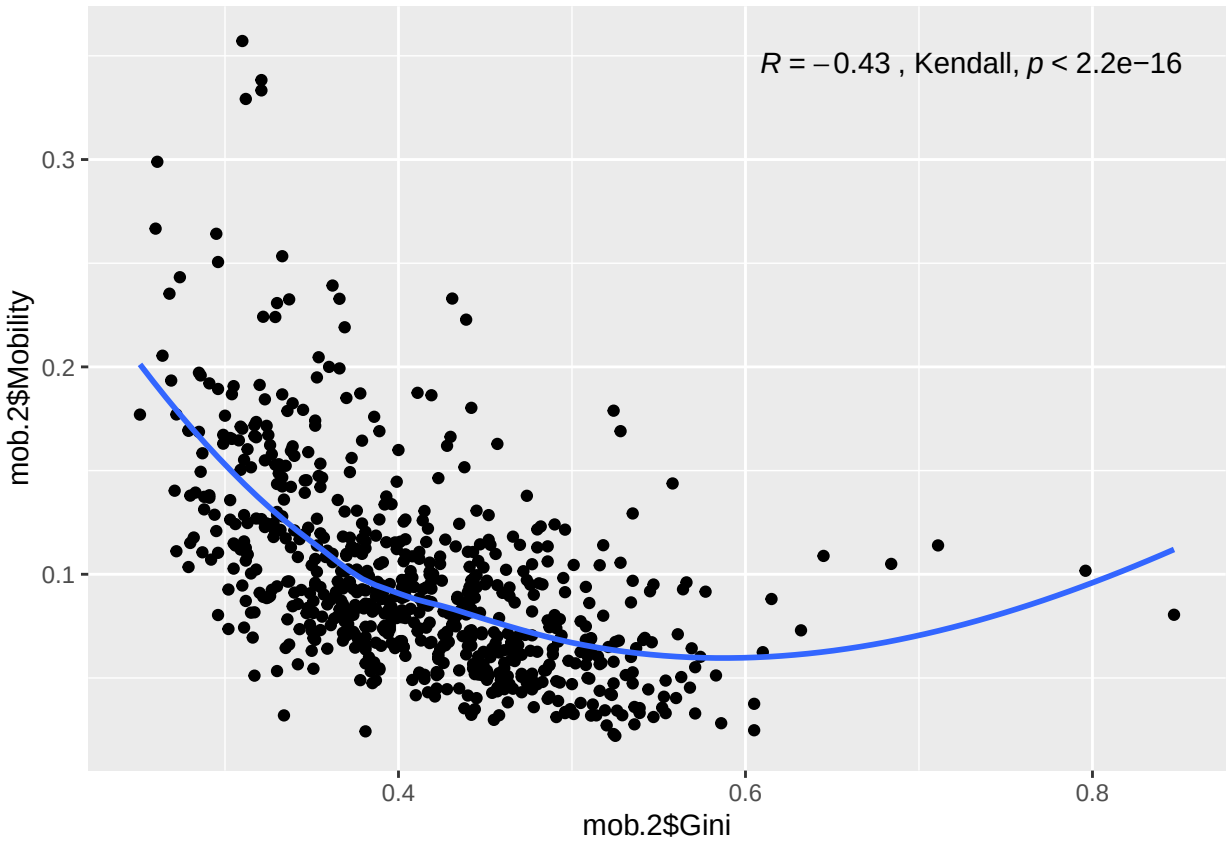
```
## [1] "Pearson"
```

```
## [1] 0.02914856
## [1] "Kendall"
## [1] 0.03659869
## [1] "Spearman"
## [1] 0.05522611
## [1] "Mobility and Gini index"
## [1] "Pearson"
## [1] -0.5252648
## [1] "Kendall"
## [1] -0.4268167
## [1] "Spearman"
## [1] -0.5973715
## [1] "Mobility and School Spending"
## [1] "Pearson"
## [1] 0.2418035
## [1] "Kendall"
## [1] 0.2060139
## [1] "Spearman"
## [1] 0.3056827
## [1] "Mobility and Test Scores"
## [1] "Pearson"
## [1] 0.4715755
## [1] "Kendall"
## [1] 0.358182
## [1] "Spearman"
## [1] 0.505029
```

As shown, the Gini index and test scores were had the highest correlations with mobility. To find this, I ran correlation tests of differing methods: Pearson, Kendall and Spearman. The Spearman tests overall seemed to give the highest correlations. Kendall and Spearman would give the best correlations as the Pearson correlation is for linear relationships, which is unlikely as weve seen the data is skewed. The two highest correlations are shown in the scatterplots below.

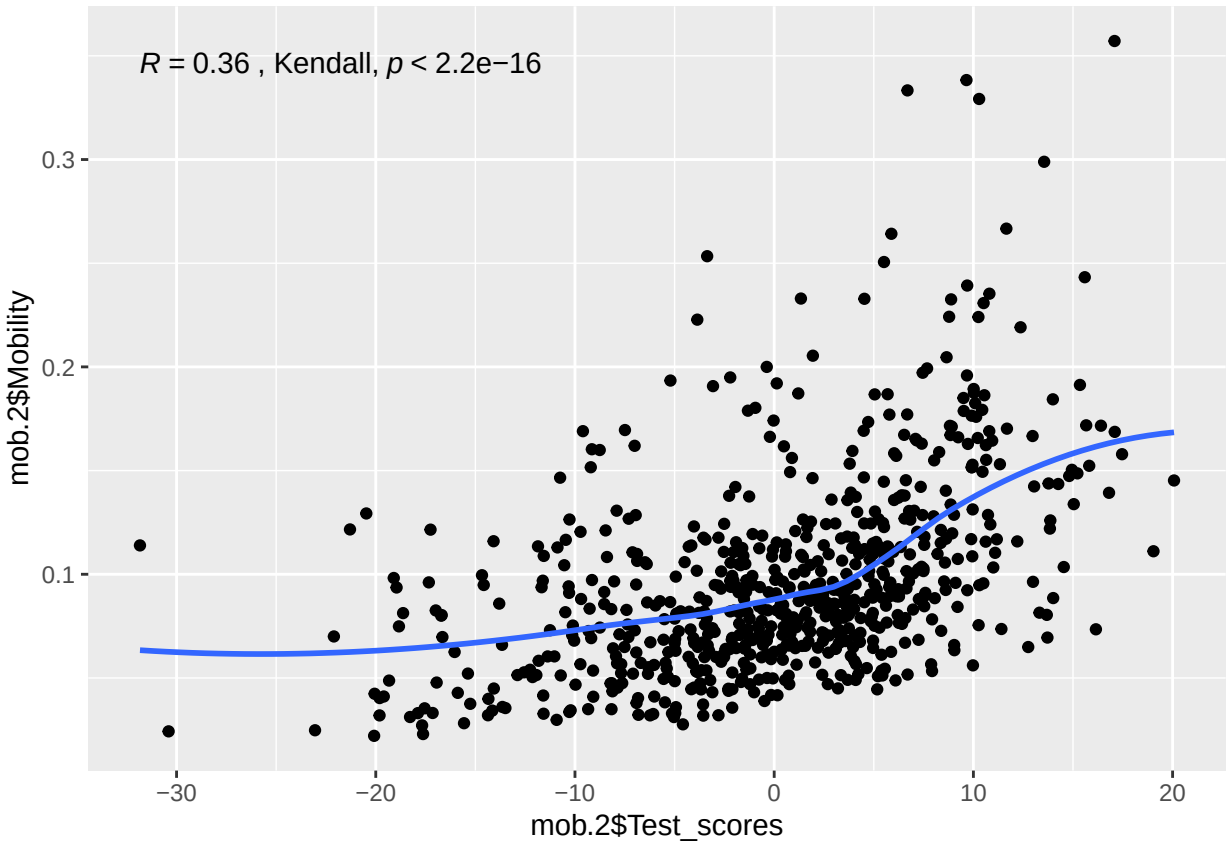
```
ggplot(mob.2, aes(x = mob.2$Gini, y = mob.2$Mobility)) + geom_point() + geom_smooth(se = FALSE) + stat_
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



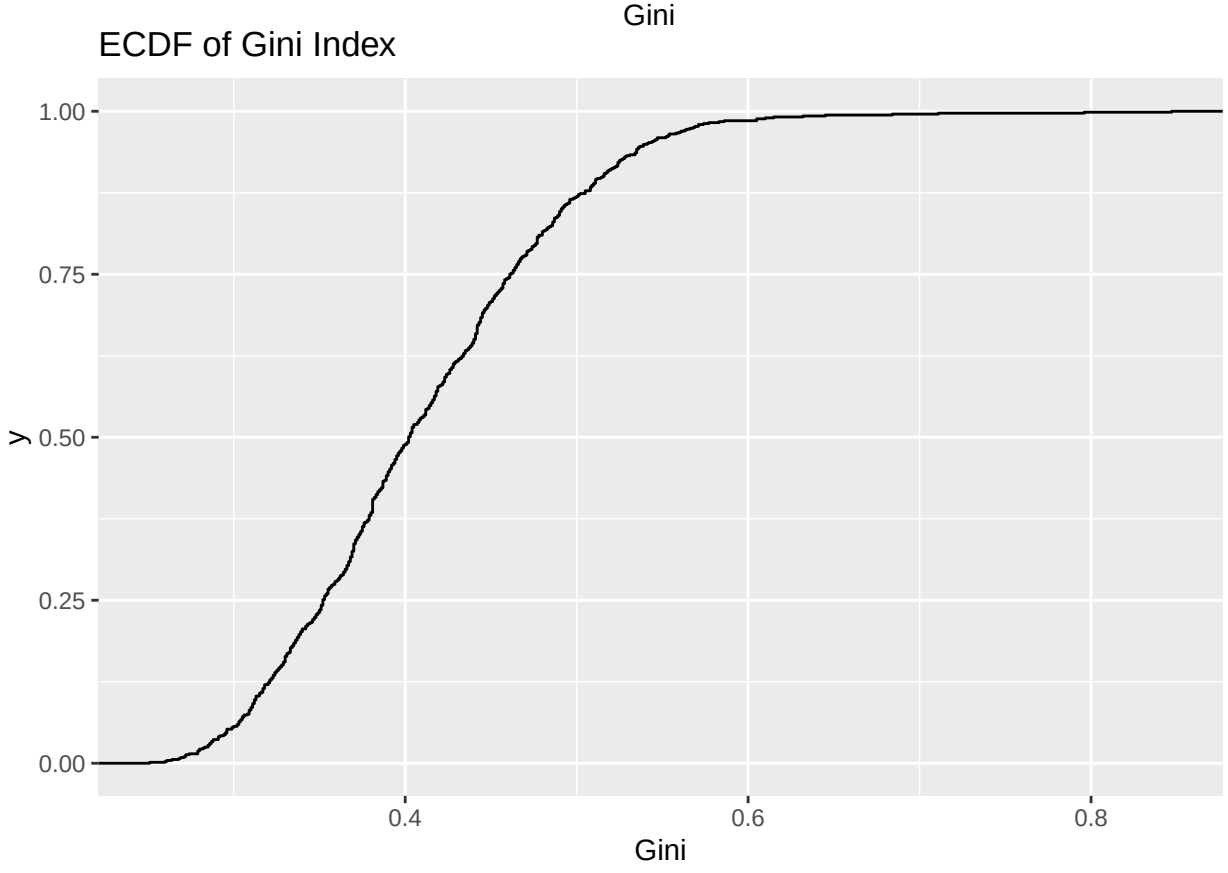

```
ggplot(mob.2, aes(x = mob.2$Test_scores, y = mob.2$Mobility)) + geom_point() + geom_smooth(se = FALSE) +
```

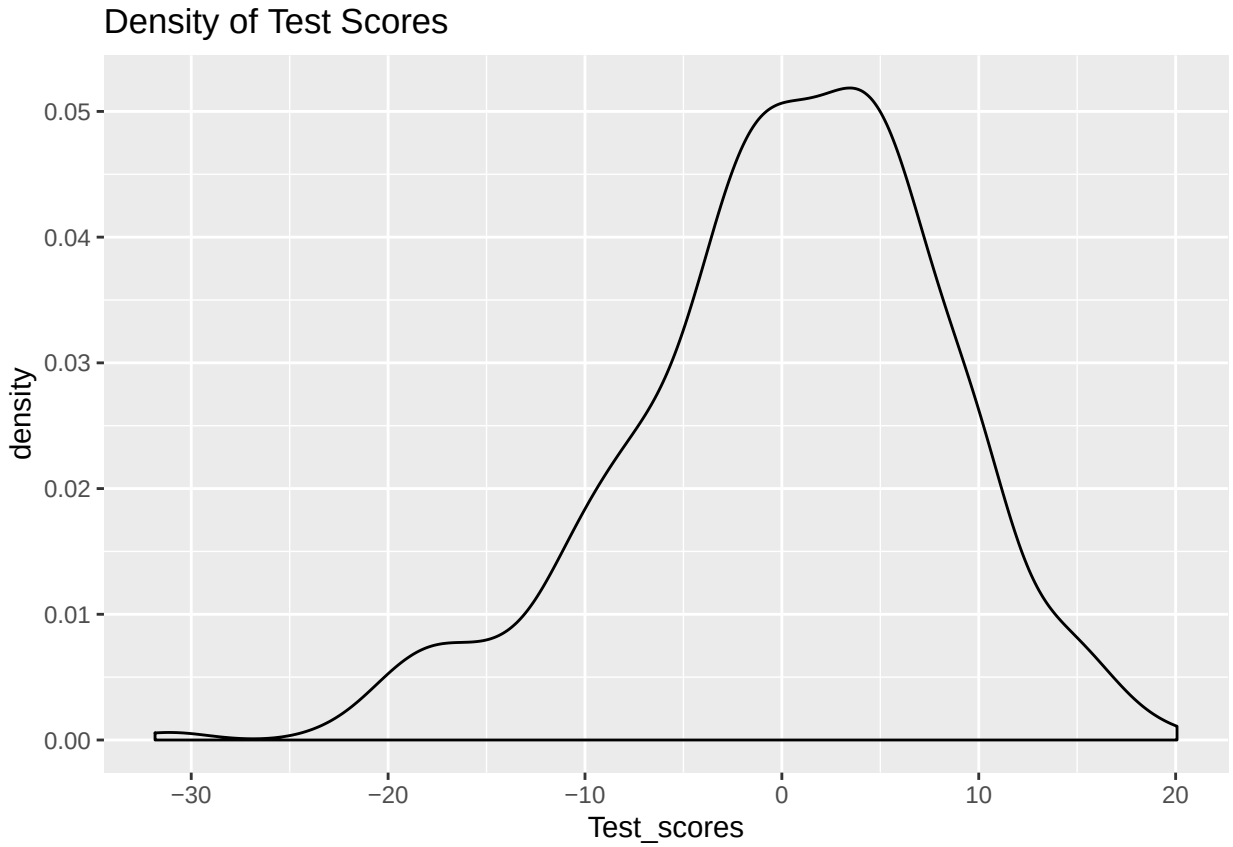
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Again, these make sense as test scores tend to decide the future of students. Additionally, the Gini index, showing income gaps, would be great for predicting the possibility of movement upward in life.

Density and ECDFs





The density plots are both slightly skewed but appear approximately normal other than that. The Gini index

is skewed right while the Test scores are skewed left. The ECDFs also show approximate normality with the s curves. They both look relatively smooth as well.

Density Plots

Conditional

```
library(np)
```

```
## Nonparametric Kernel Methods for Mixed Datatypes (version 0.60-9)
## [vignette("np_faq",package="np") provides answers to frequently asked questions]
## [vignette("np",package="np") an overview]
## [vignette("entropy_np",package="np") an overview of entropy-based methods]
```

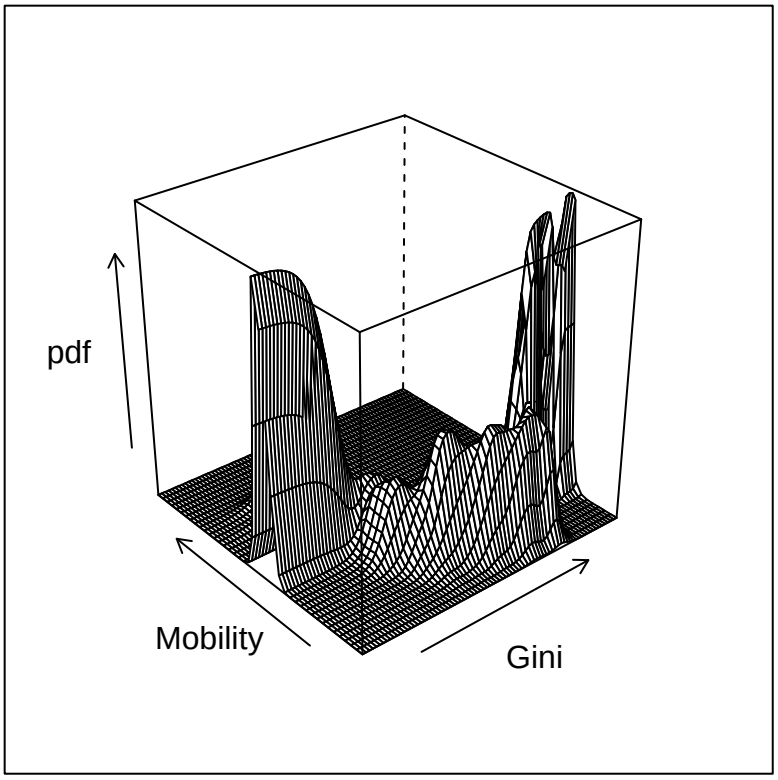
```
print('Gini')
```

```
## [1] "Gini"
```

```
mob.cdens <- npcdens(Mobility~Gini, data = mob.2)
```

```
##
Multistart 1 of 2 |
Multistart 1 of 2 |
Multistart 1 of 2 |
Multistart 1 of 2 /
Multistart 1 of 2 -
Multistart 1 of 2 \
Multistart 1 of 2 |
Multistart 1 of 2 |
Multistart 2 of 2 |
Multistart 2 of 2 |
Multistart 2 of 2 /
Multistart 2 of 2 -
Multistart 2 of 2 |
Multistart 2 of 2 |
Multistart 2 of 2 /
Multistart 2 of 2 -
```

```
grid <- expand.grid(Gini = seq(0, 0.8, 0.01), Mobility = seq(0, 0.4, 0.01))
fhat = predict(mob.cdens, newdata= grid)
wireframe(fhat~grid$Gini*grid$Mobility, xlab = "Gini", ylab = "Mobility", zlab = "pdf")
```



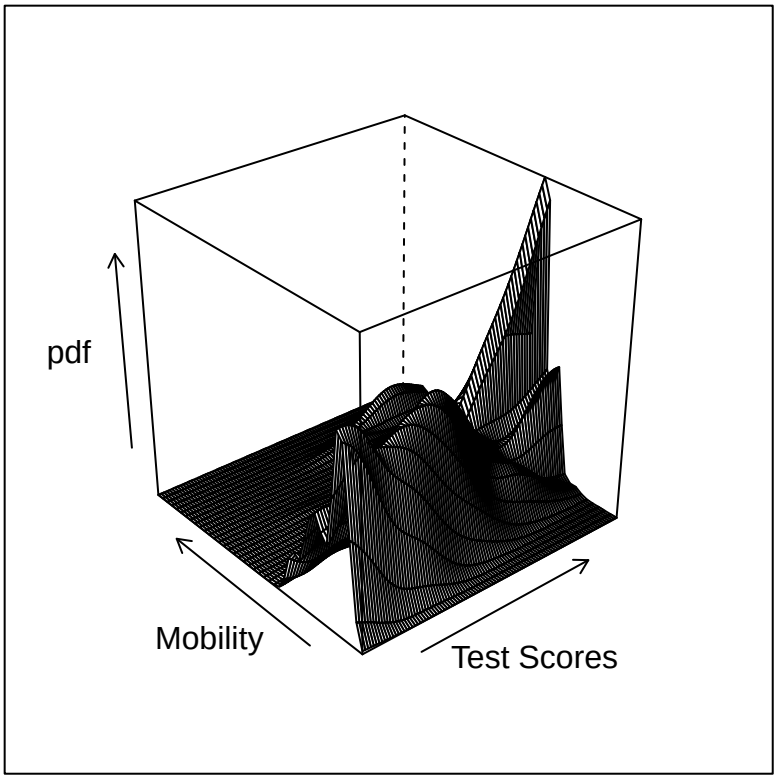

```
print('Test Scores')
```

```
## [1] "Test Scores"
```

```
mob.cdens <- npcdens(Mobility~Test_scores, data = mob.2)
```

```
##
Multistart 1 of 2 |
Multistart 1 of 2 |
Multistart 1 of 2 |
Multistart 1 of 2 /
Multistart 1 of 2 |
Multistart 1 of 2 |
Multistart 1 of 2 /
Multistart 1 of 2 -
Multistart 2 of 2 |
Multistart 2 of 2 |
Multistart 2 of 2 /
Multistart 2 of 2 -
Multistart 2 of 2 |
Multistart 2 of 2 |
Multistart 2 of 2 /
Multistart 2 of 2 -
```

```
grid <- expand.grid(Test_scores = seq(-20, 30, 0.5), Mobility = seq(0, 0.4, 0.01))
fhat = predict(mob.cdens, newdata= grid)
wireframe(fhat~grid$Test_scores*grid$Mobility, xlab = "Test Scores", ylab = "Mobility", zlab = "pdf")
```

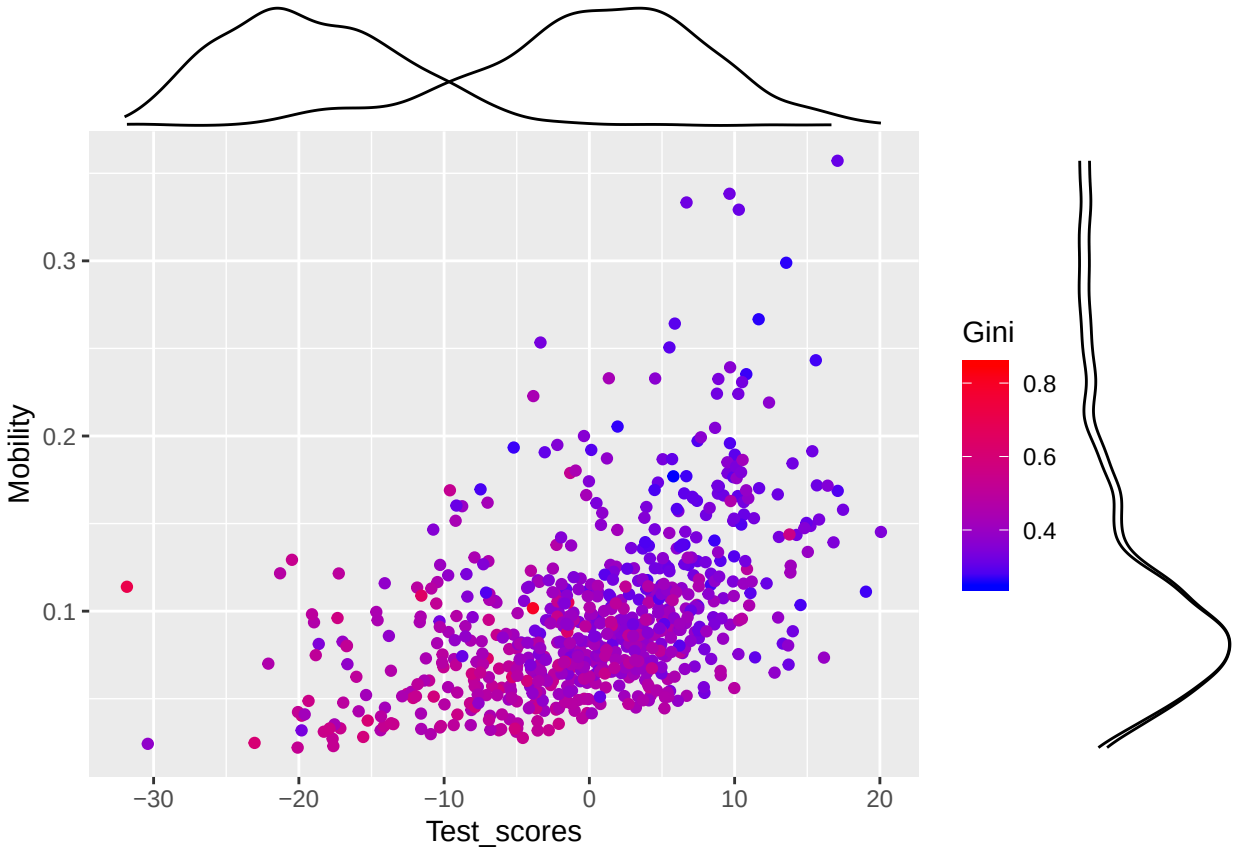


The curved relationship shows in the conditional density plot between Gini index and mobility. The conditional density plot for test scores has a lot more noise, showing less of a direct relationship, however there is a spike at high test scores and high mobility.

Marginal

```
library(ggExtra)
g = ggplot(mob.2, aes(x = Gini, y = Mobility, color = Test_scores)) + geom_point() + scale_color_gradien
ggMarginal(g, type = "density")

g1 = ggplot(mob.2, aes(x = Test_scores, y = Mobility, color = Gini)) + geom_point()+ scale_color_gradien
ggMarginal(g1, method = 'density')
```

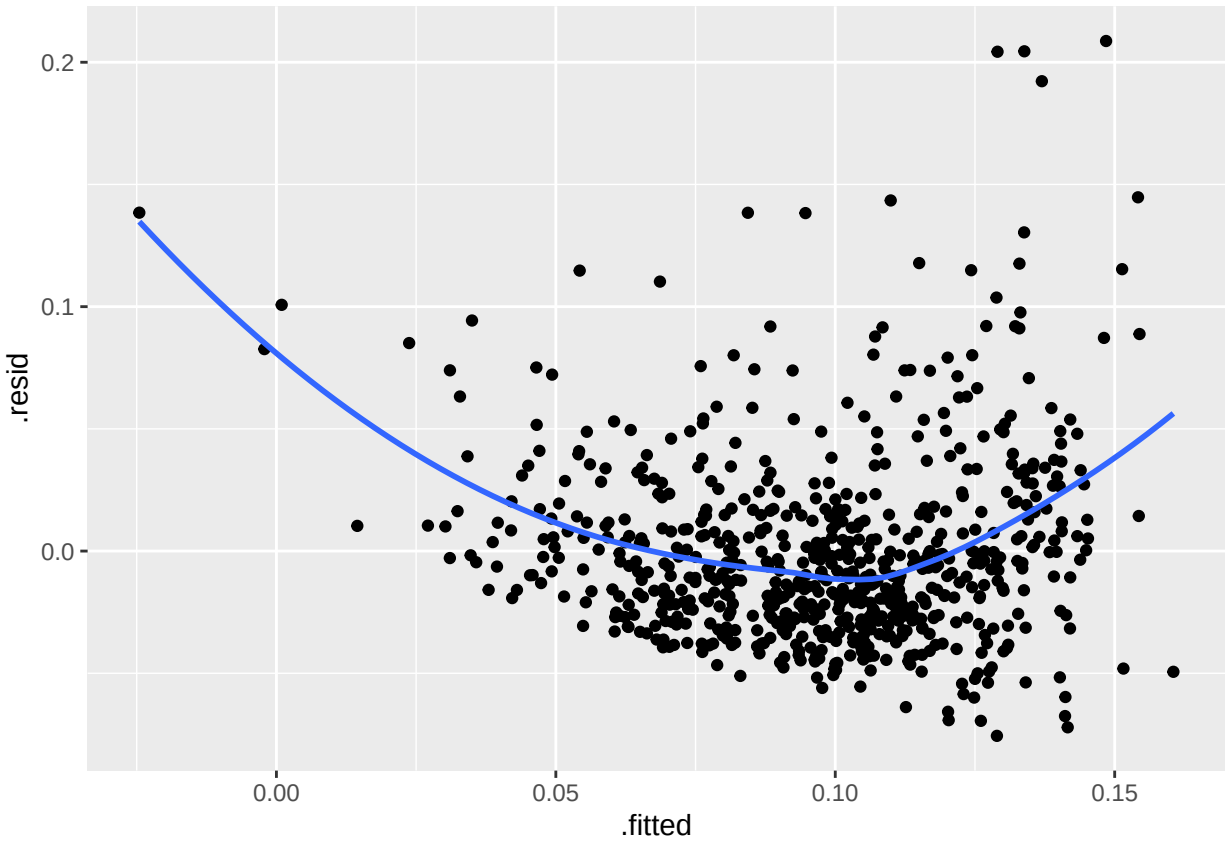
```
## Warning: Ignoring unknown parameters: method

## Warning: Ignoring unknown parameters: method

## Warning: Ignoring unknown parameters: method

## Warning: Ignoring unknown parameters: method
```



These plots show the marginal densities on each axis. We can see mobility and test scores are left skewed while the Gini index is right skewed. Additionally, color was added to each plot to show the interaction of each predictor.

```
ggplot(lm(Mobility ~ Gini + Test_scores, mob.2), aes(x = .fitted, y = .resid)) + geom_point() + geom_sm
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



As shown, we can see the residuals are a bit left skewed and not perfectly aligned on the axis. They are a bit parabolic as well. This wouldn't be a great model, however it still would be decent at prediction.

Conclusions

Overall, this data has shown what all affects the future of a child, from race to location to education. We see how much impact every detail has on someone. My focus was primarily on the socioeconomic level of a person's hometown with education as shown through test scores. I believe that these are two good markings of how someone will turnout. In the beginning I wanted my focus solely on education, but saw that simplifying a model to one factor will not produce good results. To find the best results, a model should probably include more than even two of these predictors, however, that makes it much less interpretable.