

**Worcester Polytechnic Institute**

# Project Report - Data Science II

# 

**Authors**

Megan Sin

Adina Palayoor

Molly Sunray

Meena Meyyappan

**Date**

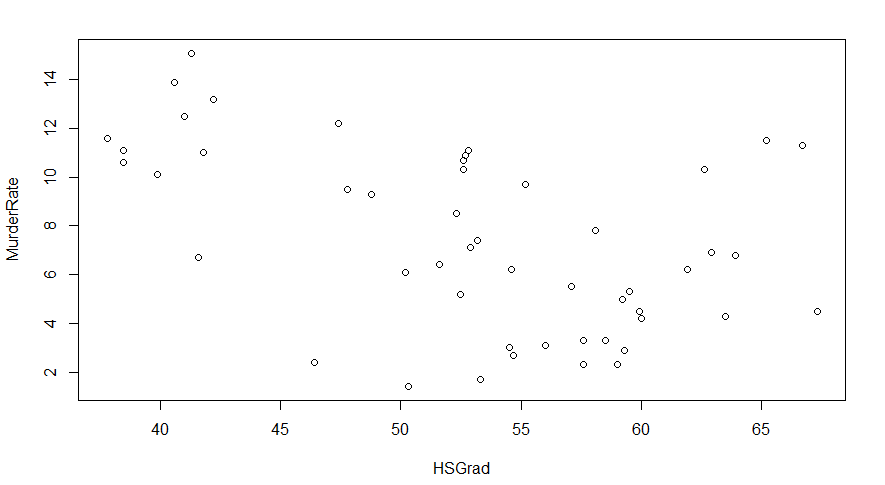
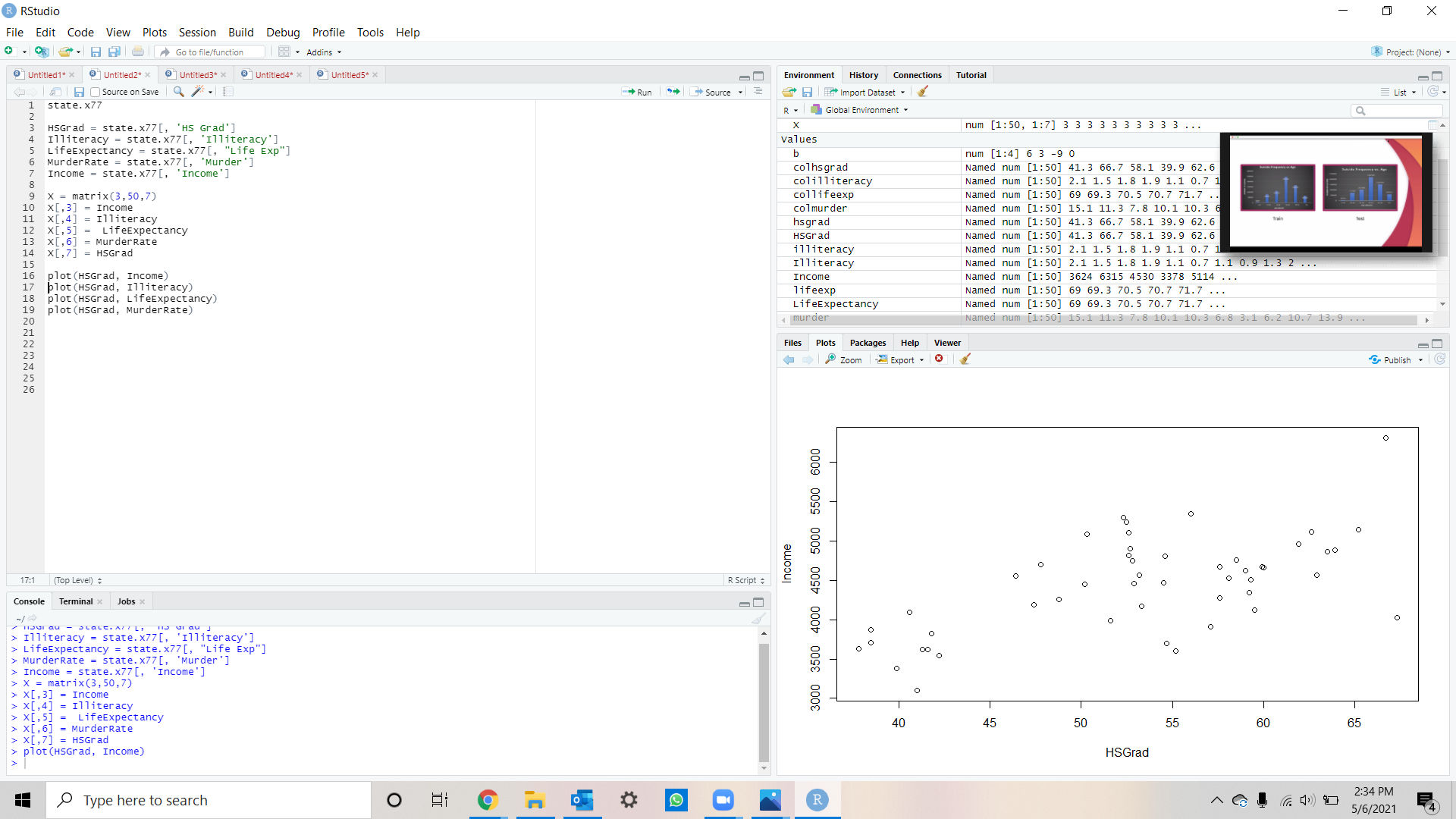
5/12/2021

**Report Submitted to**

Professor Mangoubi

We studied how the high school graduation rate in each state affected factors within that state, such as illiteracy, income, murder rate, and life expectancy in 1977. We hoped to gain an understanding of which of these other variables correlated with the rate of high school graduation. For example, does a lower graduation rate imply lower income or higher murder rates? All of these factors have the potential to be related, and we hoped to gain an understanding of the relationships by the end of the project.

We used one of the datasets within R, state.x77. This data set contained 50 rows, one respective to each state, with 8 columns showing the data collected from 8 different variables. The 8 variables were Population (an estimation of the population), Income (per capita), Illiteracy (percent of the population classified as illiterate), Life expectancy (in number of years), Murder (the rate of murder and non-negligent manslaughter per 100,000 people in the population), High School graduates (percent of high school graduates), Frost (average number of days with minimum temperatures below freezing in the capital or a large city), and Area (the area of land in square miles). We decided to eliminate the Frost and Area variables during our analysis as these don’t tell us anything useful. The source used to collect this data is from the US Department of Commerce, Bureau of the Census in 1977.

We first wanted a general understanding of the effects High School Graduation had on other variables: Murder, Income, Life Expectancy, and Illiteracy. As we compared and analyzed scatter plots, it was important to ensure we understood the general trends of each graph: were the trends positive or negative? Were they linear or polynomial? By looking at the trends the graphs presented us, we could create conclusions, or at least hypotheses, about each variable. For example, we found a possible negative polynomial relationship between High School Graduation and Murder. Alongside that, there was a positive linear relationship when High School Graduation was compared to both Income and Life Expectancy rates. Finally, there was a negative linear trend for Illiteracy. 

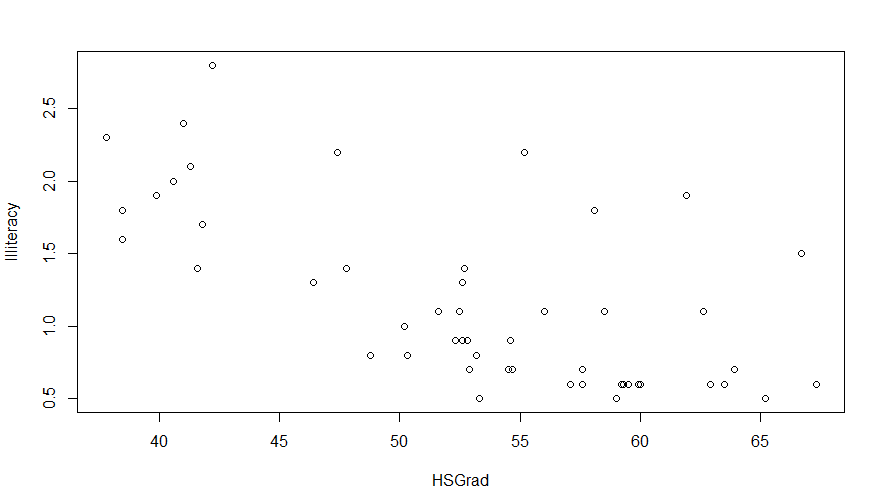
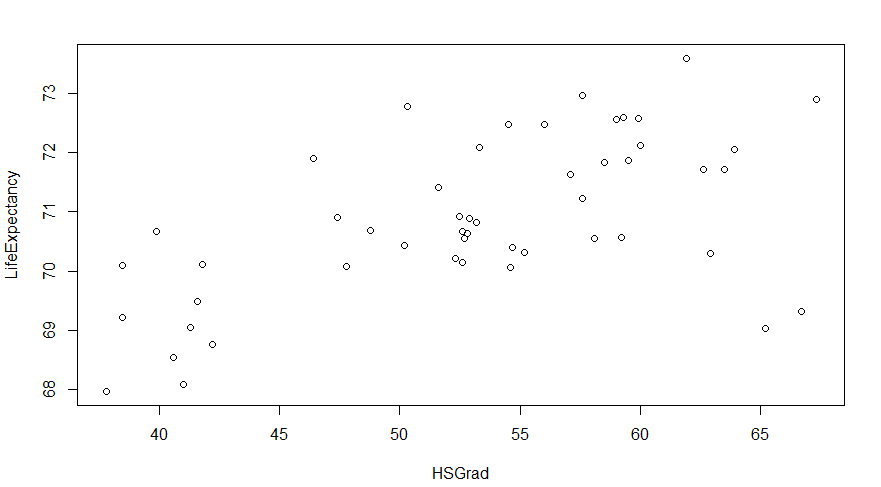


Figure 1: Scatterplots corresponding with each variable

Pink: Murder, Blue: Income, Green: Life Expectancy, Yellow: Illiteracy

We then performed the Pearson correlation test and looked at the corresponding p-values between HSGraduation and all the other variables. There is evidence suggesting a correlation between HSGraduation and Income, Illiteracy, LifeExpectancy, and Murder as their p-values are all less than 1%. However, the HSGraduation and Population correlation test results showed a significant p-value of 0.8093, which is well above the 1% alpha level.

| Relationship (x, y) | P-value | Correlation Coefficient |
| --- | --- | --- |
| HSGrad, Population | 0.8093 | -0.04230587 |
| HSGrad, Income | 1.422e-05 | 0.6630467 |
| HSGrad, Illiteracy | 1.846e-05 | -0.6565166 |
| HSGrad, LifeExp | 7.73e-05 | 0.617608 |
| HSGrad, Murder | 0.000576 | -0.5526576 |

Figure 2: Table of p-values and correlation coefficients from all the Pearson correlation tests performed.

We then split up the data into training and test data by randomly assigning 70% of the states to be the training dataset and the other 30% to be the testing dataset. We used regression to study the dataset and tried out three techniques to determine the most representative model for our data set: linear regression, polynomial regression, and linear regression with interaction terms.

The process for performing linear regression began with first determining which variables may have linear relationships with HSGraduation. Next, we created linear models for HSGraduation v. Income, Illiteracy, Life Expectancy, and Murder Rate using the training data. Lastly, we used the testing data to compute the testing error. The scatterplots for Income and Life Expectancy showed positive linear trends, while the scatterplots for Illiteracy and Murder Rate showed negative linear trends. When computing the testing error, we found that linear regression worked well for certain variables and that the models were representative of the data. The testing error for all variables being studied resulted in small mean squared errors, except for Income, which had a mean squared error of 361388.6. The computed errors for Illiteracy, Life Expectancy, and Murder Rate were 0.2149918, 1.493826, and 14.36564, respectively. Based on these values, it is clear that the relationship between HSGraduation and Illiteracy produced the best model due to the small error.

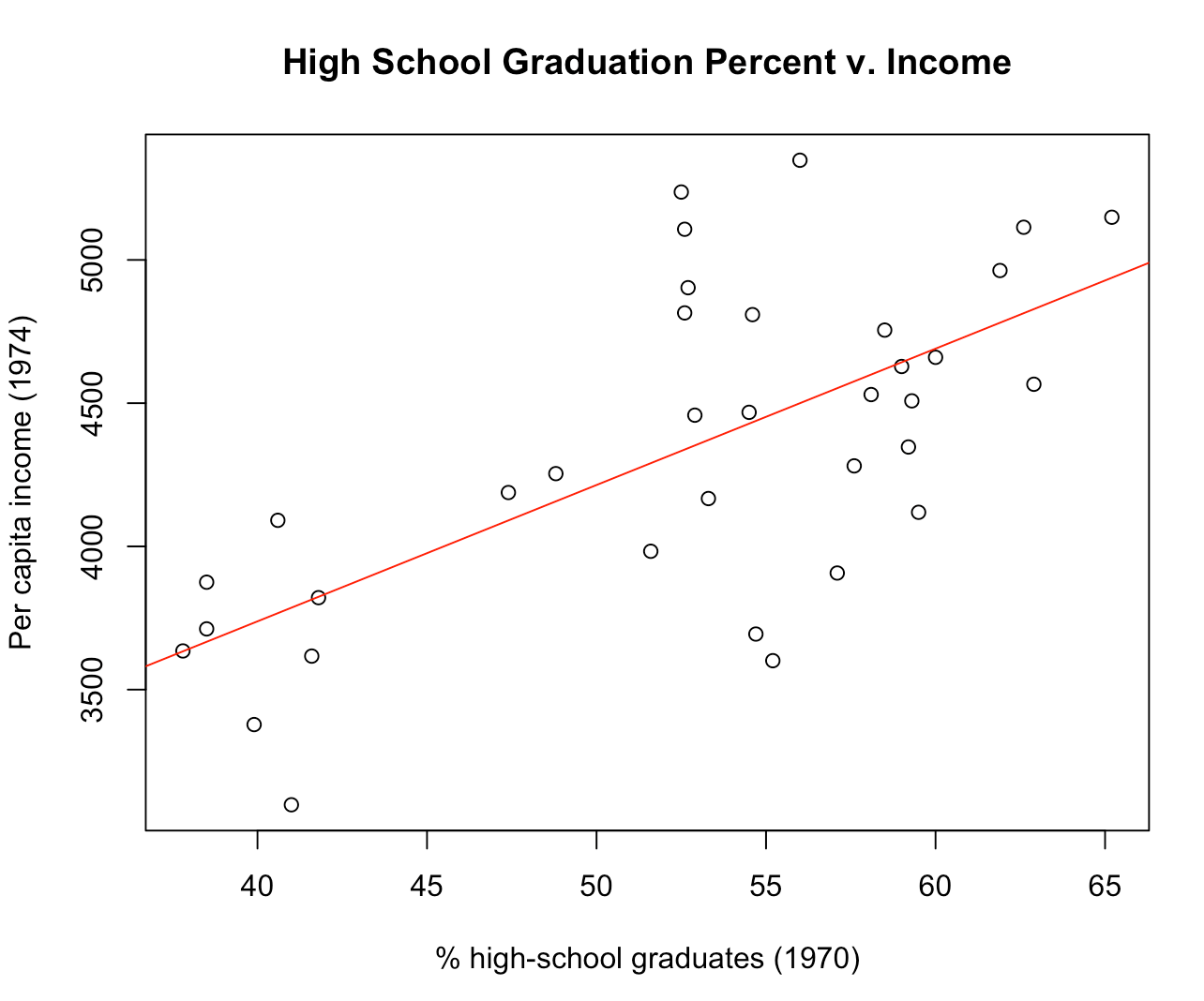


Figure 3: Scatterplot of High School Graduation Percent v. Income with fitted linear regression model.

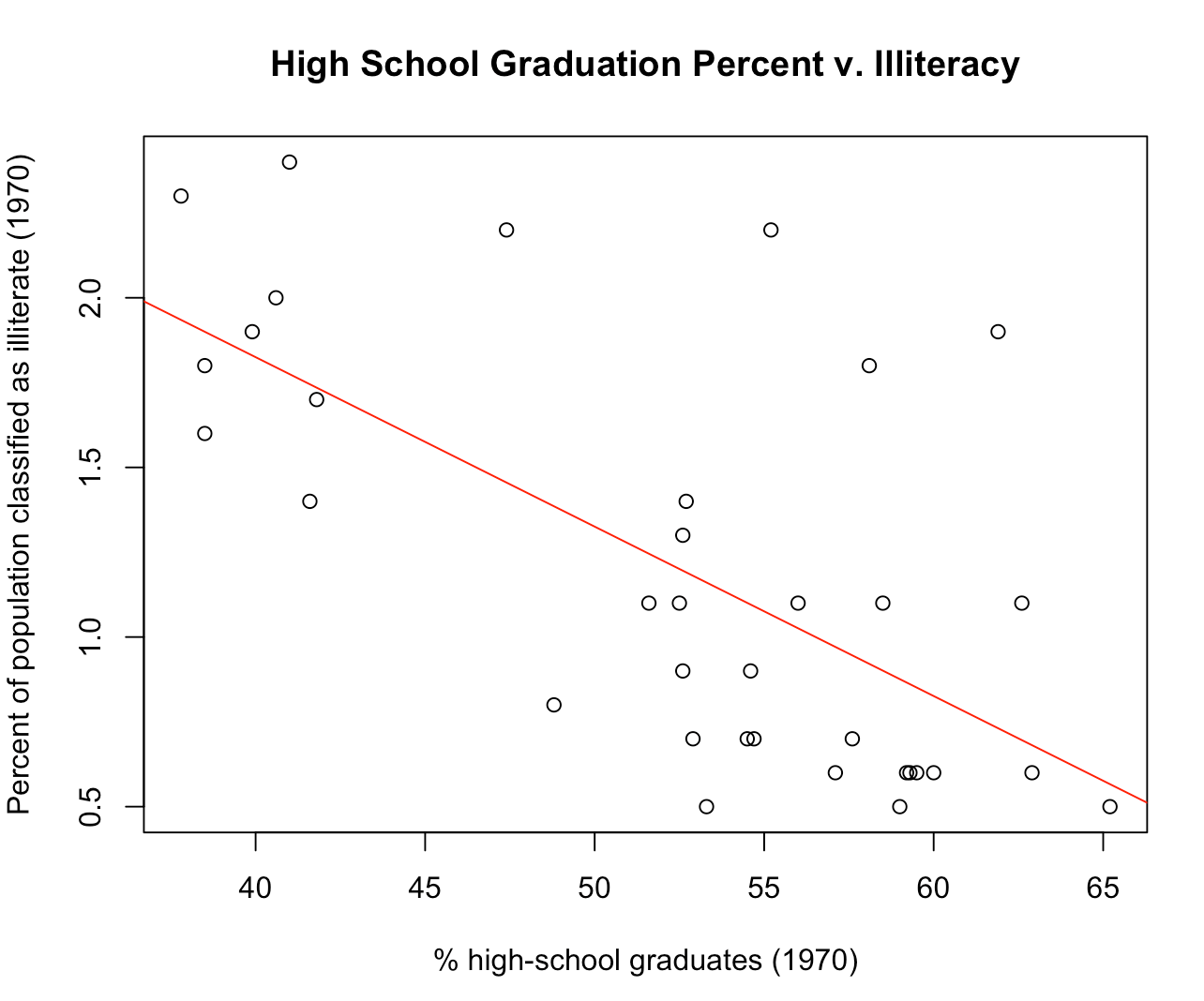


Figure 4: Scatterplot of High School Graduation Percent v. Illiteracy with fitted linear regression model.

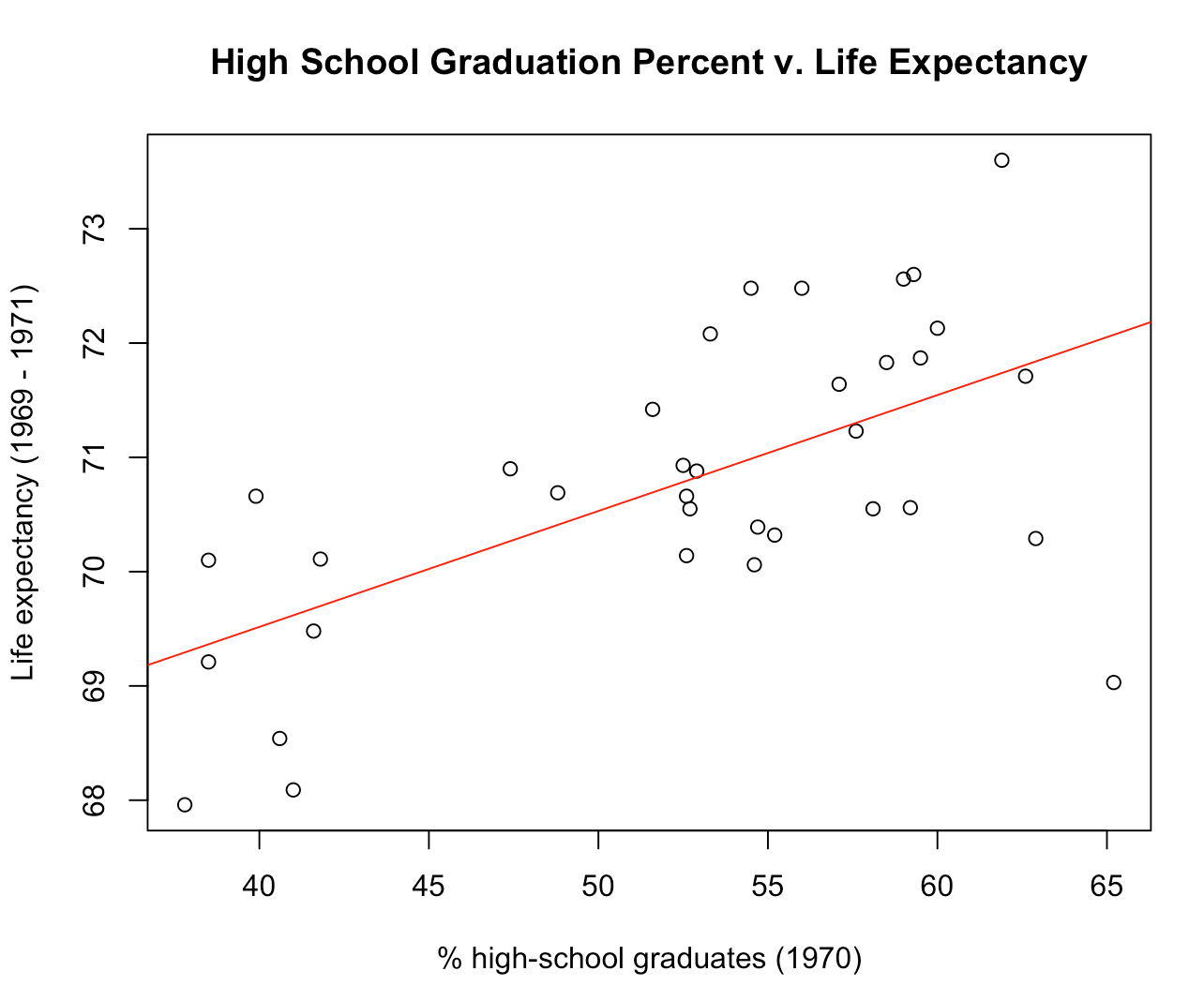


Figure 5: Scatterplot of High School Graduation Percent v. Life Expectancy with fitted linear regression model.

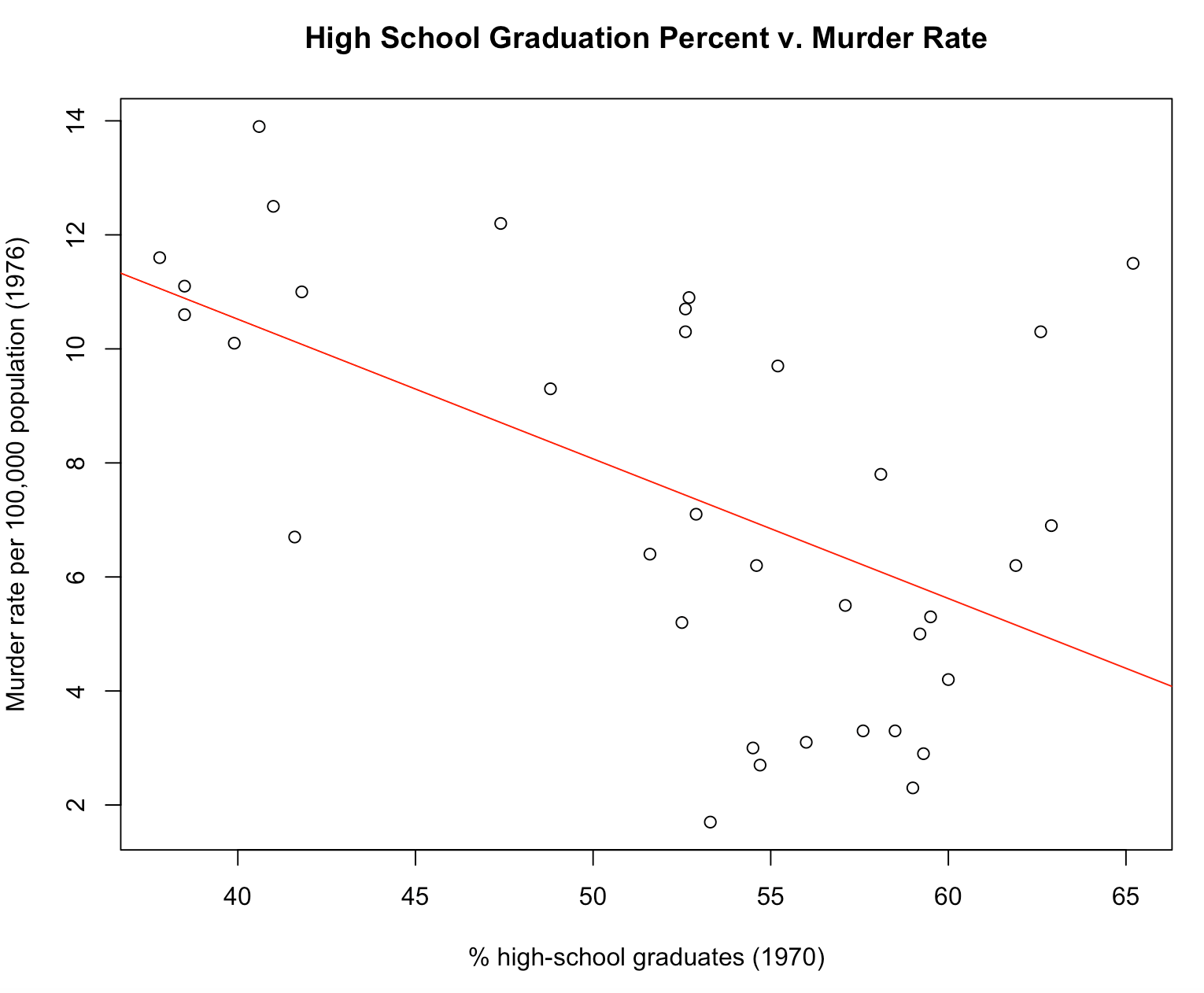


Figure 6: Scatterplot of High School Graduation Percent v. Murder Rate with fitted linear regression model.

The steps taken to try the polynomial regression method included first making a guess based on the scatterplots and taking note of which relationships may be polynomial. Then, creating polynomial models for HSGraduation v. Income, Illiteracy, LifeExpectancy, and Murder up to the power of 6. Next, creating ANOVA tables for each of these polynomial models and then comparing the p-values for each degree in each model. And finally, computing the testing error for the polynomial models, given they are an appropriate model. Only the HSGraduation v. Murder scatterplot looked like a polynomial model might be suitable from solely looking at the scatterplots. Following this initial observation and the creation of polynomial models up to the power of 6 their respective ANOVA tables, the p-values for each degree for each model were analyzed. For all the p-values in the ANOVA tables for HSGraduation v. Income, Illiteracy, and LifeExpectancy, the p-values were greater than 5%. However, for the ANOVA table for HSGraduation v. Murder, the p-values for the second and third power were less than 1%; p-value = 0.022335 and p-value = 0.002193, respectively. Following this analysis, we only continued further investigation for these two models: HSGraduation v. Murder with HSGraduation to the second power and HSGraduation v. Murder with HSGraduation to the third power. The calculated testing error, which was the mean squared error, for the model to the second power was 11.48441, and the model to the third power was 22.46635. The model to the second power was a better model for the relationship between high school graduation rate and murder rate, likely because it was a simpler model than the model to the third power. As complexity increases, the likelihood of overfitting increases, especially if there are not many data points; this was evident in our investigation. Simpler models are ideal for a dataset where m is small and are less likely to overfit.

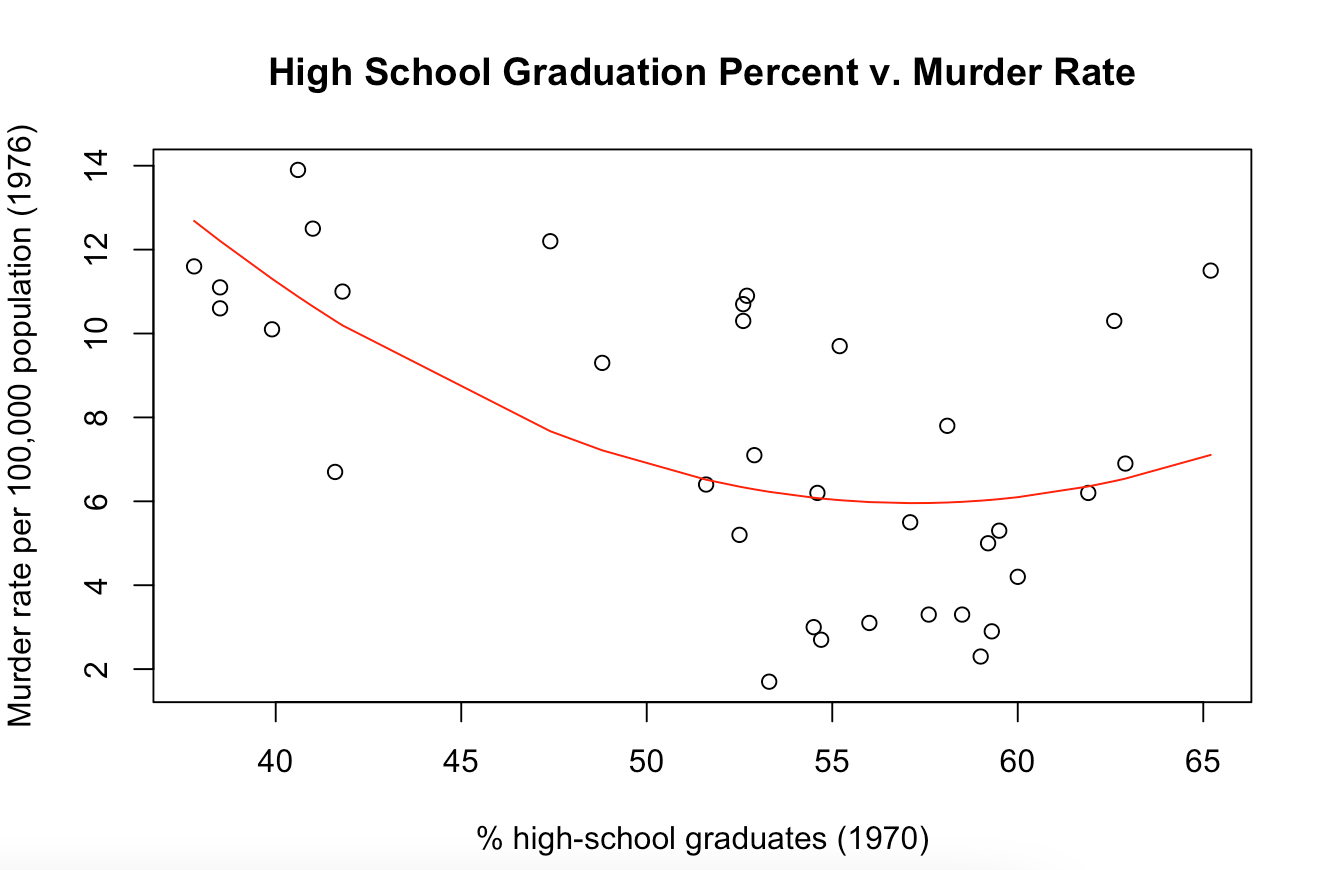


Figure 7: Scatterplot of High School Graduation Percent v. Murder Rate with fitted quadratic regression model.

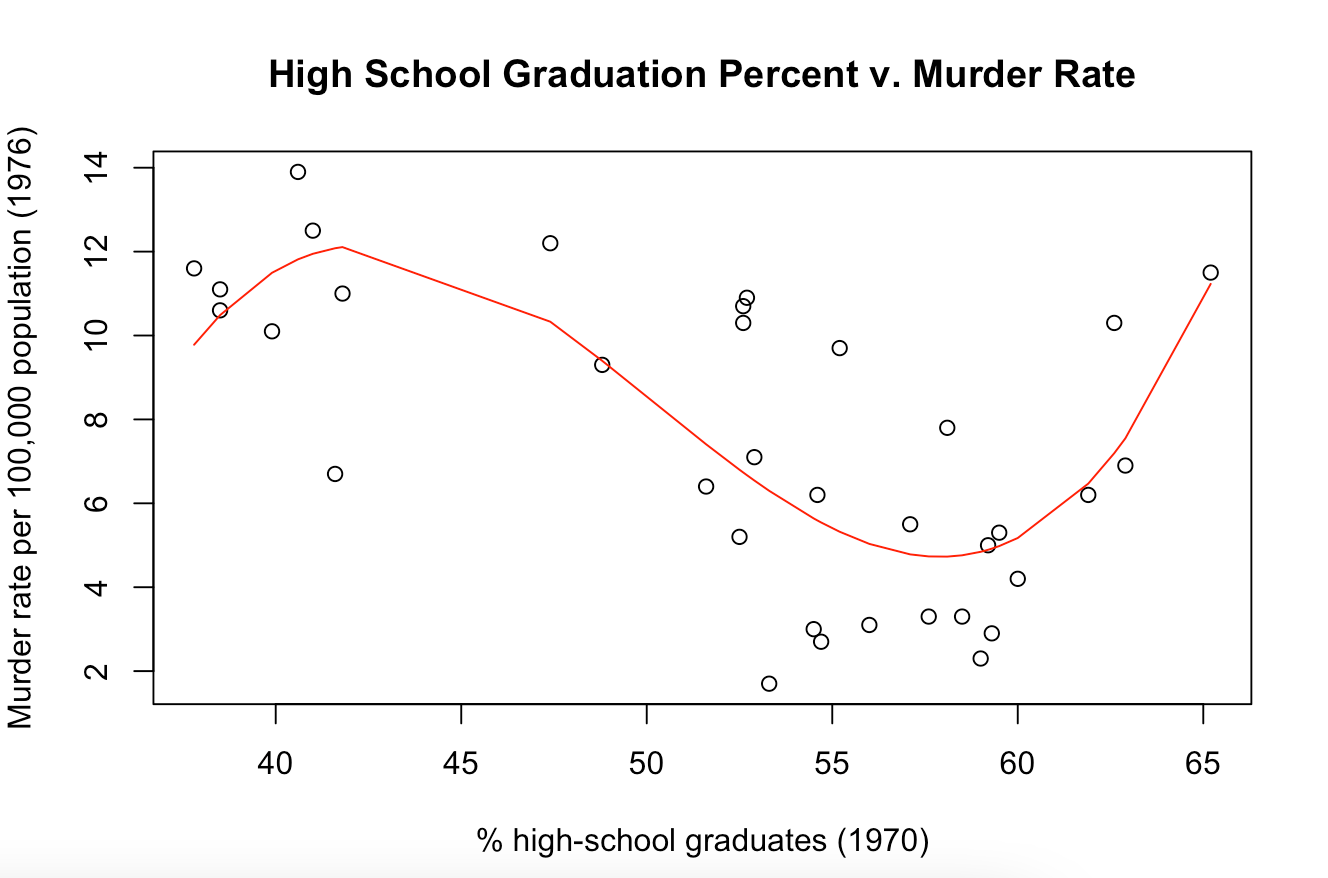


Figure 8: Scatterplot of High School Graduation Percent v. Murder Rate with fitted cubic regression model.

We also attempted to use linear regression with interaction terms to see if this would be a good fit for our model. We started by determining which variables we wanted to use as the interaction terms. We tried to pick variables that we believed could, in combination, be more representative of our data. We then created the plot and trendline of the model with our selected interaction terms. As we tried out variable combinations, it was increasingly apparent that none of the combinations were great fits for the model due to a low R-squared. We also tried running the testing data through our model, and all variable combinations resulted in a high MSE value. For example, when we used Murder rate and Illiteracy, the MSE was 78.59. This MSE was significantly higher than the MSE within both the polynomial and linear models. All other variable combinations had worse MSE values, with Income and Illiteracy resulting in a MSE of 4214909. Linear Regression with Interaction terms did not produce a very representative model for our data, and other methods proved to be much more explanatory.

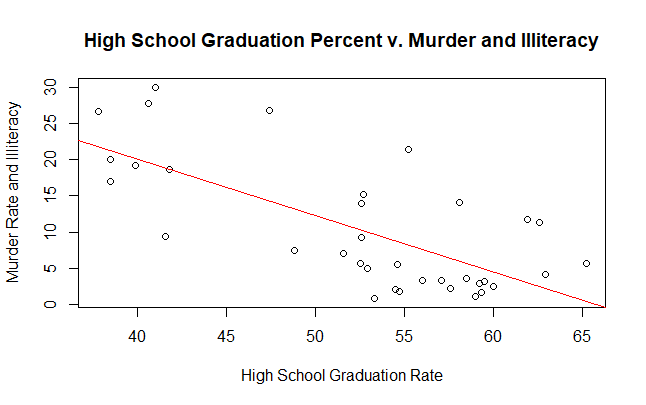


Figure 9: Scatterplot of High School Graduation rate v. Murder and Illiteracy Rate with fitted linear regression with interaction terms model.

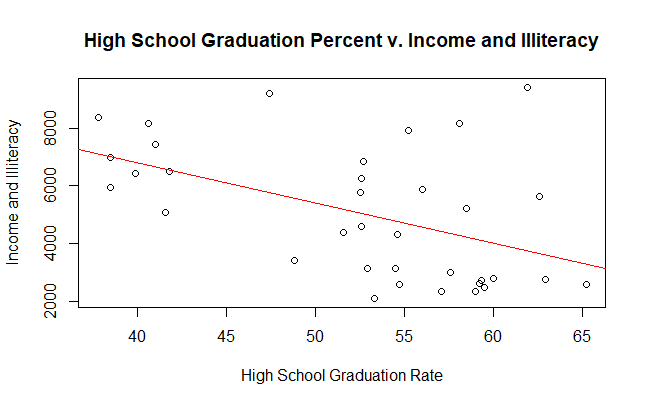


Figure 10: Scatterplot of High School Graduation rate v. Murder and Illiteracy Rate with fitted linear regression with interaction terms model.

We discovered that as high school graduation rate generally increased, income and life expectancy increased linearly. On the other hand, the illiteracy rate and murder rate decreased linearly. While earlier we stated that polynomial regression created a better model for HS Graduation Rate v. Murder Rate, the MSE was only slightly better than the linear model (MSE = 14.36564 for the linear model and MSE = 11.48441 for the quadratic model). More data points would be necessary to determine which type of relationship would best fit the high school graduation rate and the murder rate. We also learned that there was no direct relationship between the high school graduation rate and the state’s population from the initial correlation test performed. Overall, for most of the variables, linear regression was the best approach used as it was simple to use and understand, didn’t overfit the data, and had relatively low MSE.