Assessing the Relationship between Software Programmers' Job Satisfaction, Work Week Hours, and Compensation

Adrianna Hohil, Bora Mutluoglu, Errick-Michael Santos, Will Hwong

Team 25

Background:

The dataset chosen for this project is a combination of the 2019 and 2020 Stack Overflow software developer self-report surveys. Every year, Stack Overflow maps their demographic of users through self-report surveys, and forms a data frame containing information about each user's personal background, background as a developer, income, job, and career satisfaction. However, between the 2019 and 2020 datasets, some variables differed so in our data cleaning process, we ensured to focus on common variables between both and ones that were relevant to our research interests. We focused on featured variables such as country, education level, job satisfaction, type of compensation, age, number of programming languages, number of years coded, and number of hours worked in a week

2019 Dataset Information: The 1st dataset consists of 88,873 developers surveyed from 170 countries.
2020 Dataset Information: The 2nd dataset consists of 64,461 developers surveyed from 170 countries.

Research Question: Is there a relationship between the reported job satisfaction of a software programmer relative to the amount of hours worked in a week and their reported compensation (standardized to U.S. Dollars)?

Methods: Multivariate Regression and Cross Validation

The primary analysis technique chosen for this project was multivariate regression, and our follow-up analysis technique was cross-validation. Multivariate regression was chosen because our team was interested in comparing a software programmer's reported job satisfaction, work week hours, and compensation to identify a relationship between the variables based on our team's hypothesis.

Hypothesis: Reported Work Week Hours (WWH) and reported compensation will be the strongest predictors of reported job satisfaction relative to other factors found in the survey such as country of origin, years of experience, or age.

If our hypothesis holds true, we predicted that the regression model will assign a higher weight to the compensation feature as common logic would suggest that programmers who are compensated highly are more likely to deal with long hours. Cross validation methods will be utilized in conjunction with the multivariate regression in an effort to have a more concrete conclusion to our analysis of the models we develop. Due to the voluntary nature of the dataset and potential unreliability, there may be some assumptions taken by the project team although it will be clearly noted and explained where intentional manipulations of the data set were made.

Results: 3 Models

Multivariate Linear Regression Model 1

Work Satisfaction Index = w0 + w1 x Compensation + w2 x WorkWeekHours

Multivariate Linear Regression Model 2

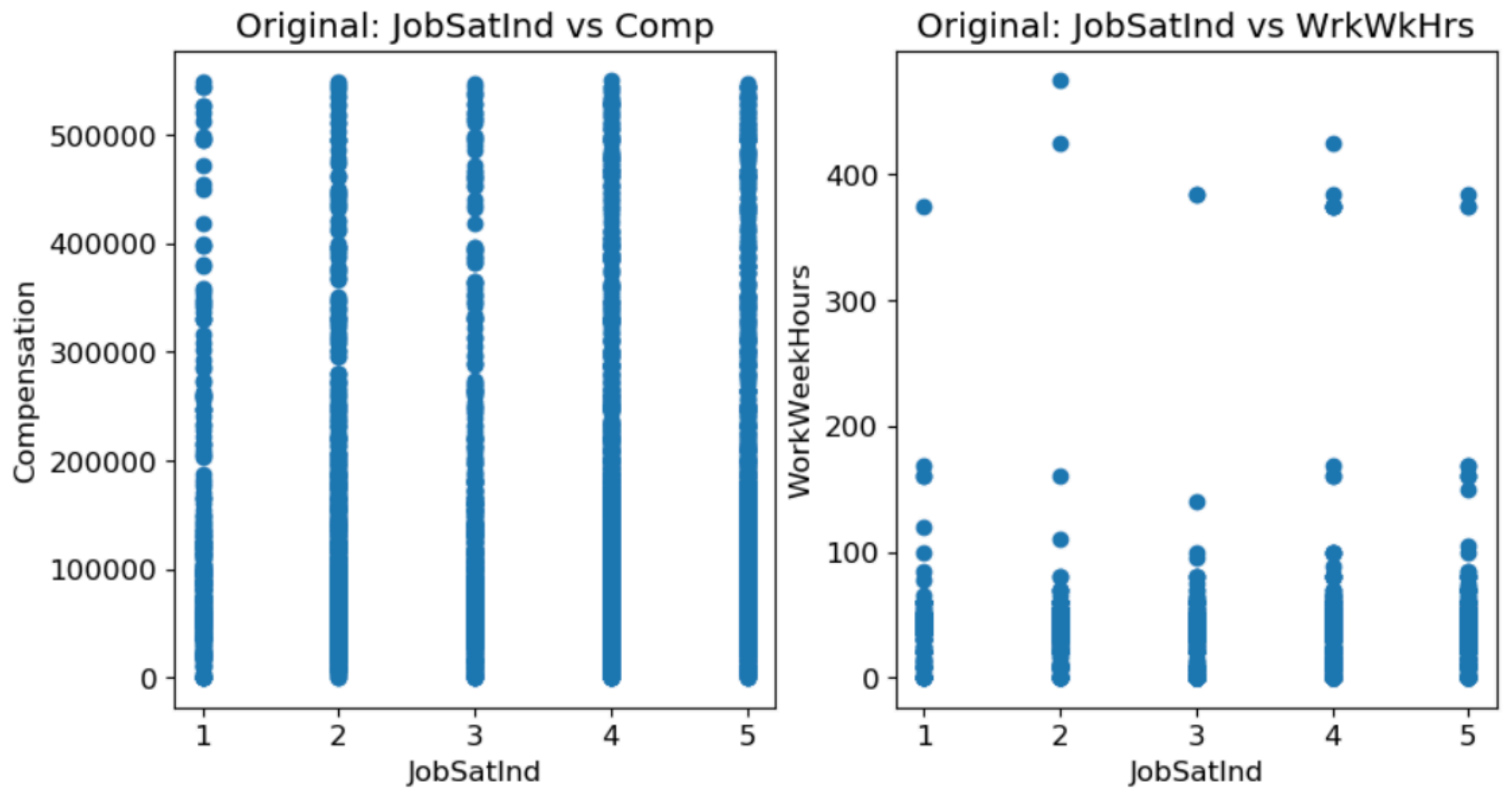
Work Satisfaction Index = w0 + w1 x Compensation + w2 x WorkWeekHoursˆ(1.2)

- How does compensation and WorkWeekHours predict Job Satisfaction levels? Might want to use Multivariate Interaction terms?
- Using a 80% (training) and 20% (testing) data split for Cross Validation

Multivariate Linear Regression Model 3

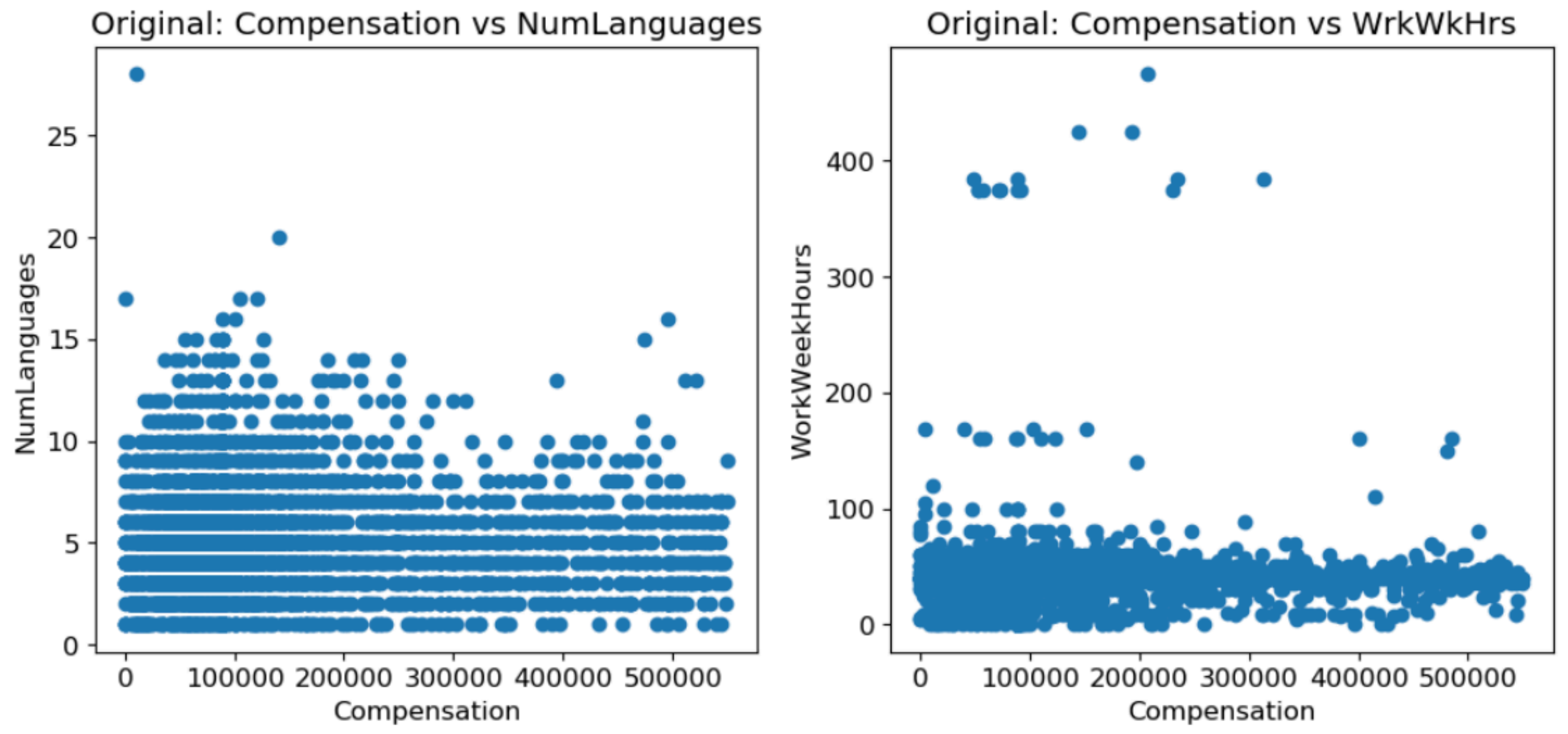
Compensation = 105,487 + 648 * WorkWeekHrs + (-486) * NumLanguages

Graphing of Job Satisfaction Index in Regards to Yearly Compensation and Work Week Hours


Original: JobSatInd vs Comp

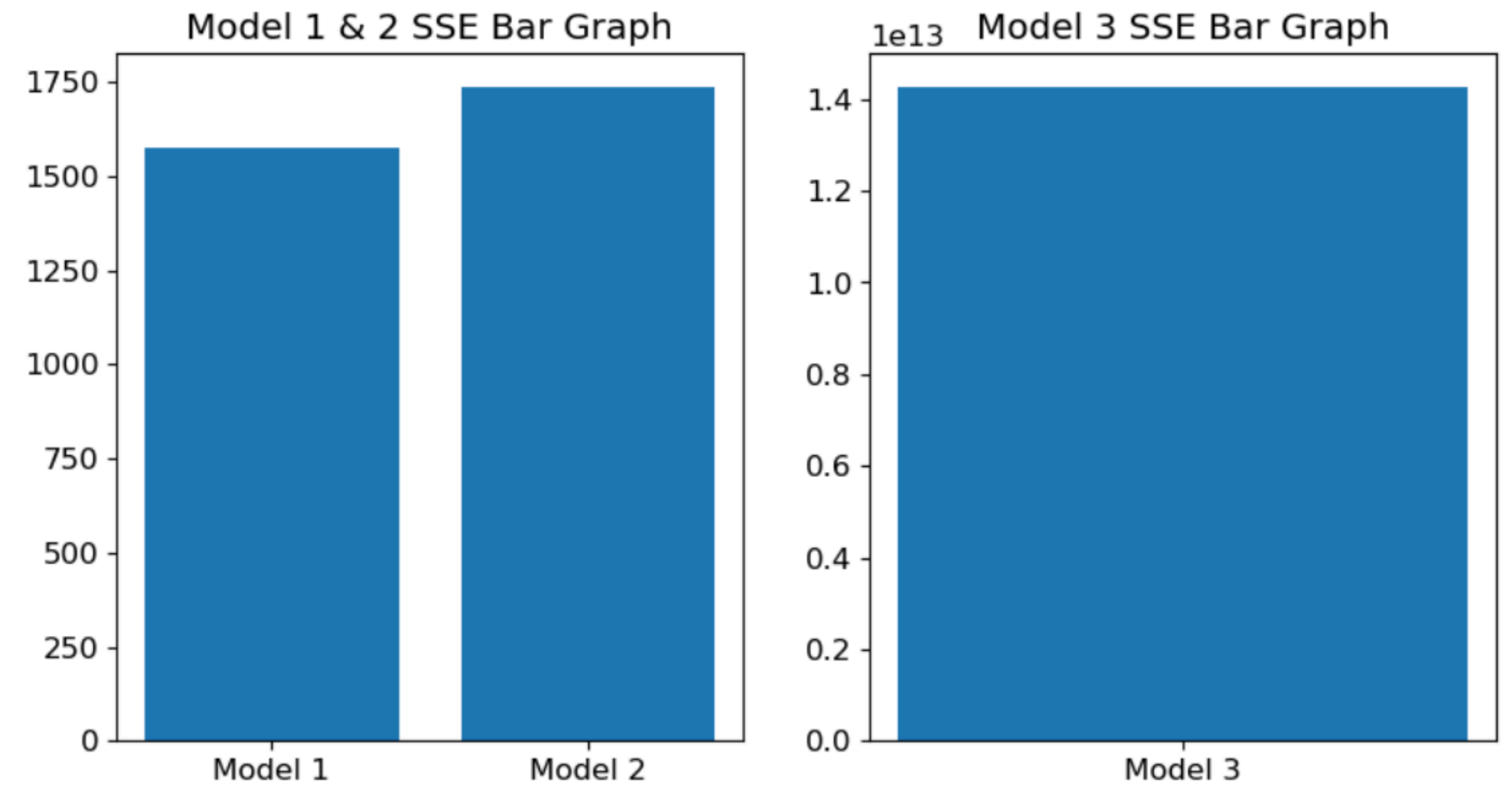
Original: JobSatInd vs WrkWkHrs

Graphing of Compensation in Regards to Work Week Hours and Number of Computer Languages


Original: Compensation vs NumLanguages

Original: Compensation vs WrkWkHrs

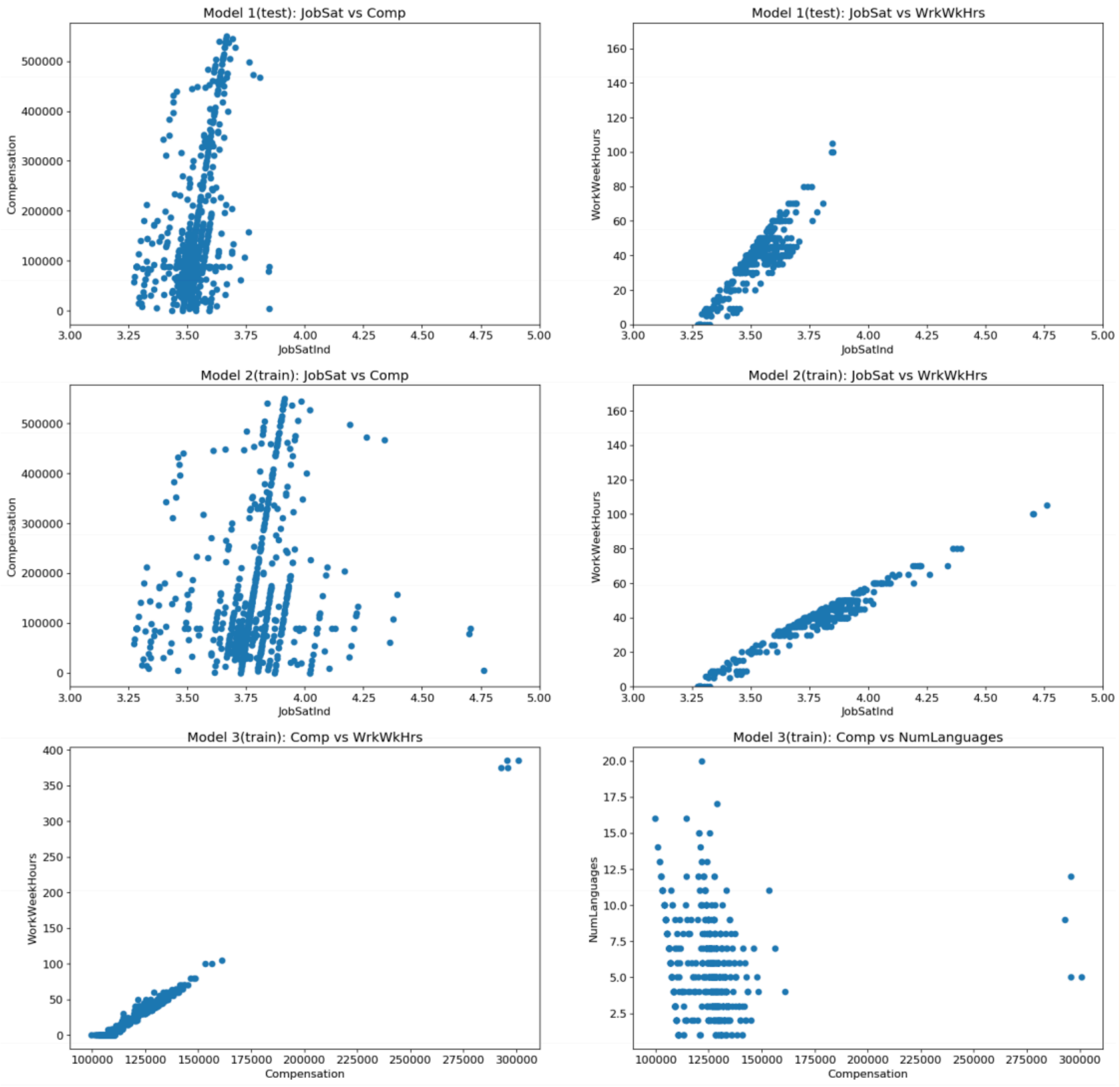
Calculations and Plotting of SSE (Models 1, 2, & 3)


Model 1 & 2 SSE Bar Graph

Model 3 SSE Bar Graph

Results:

Graphing of Testing Data of All Models and Their Respective Data


Model 1(test): JobSat vs Comp

Model 1(test): JobSat vs WrkWkHrs

Model 2(train): JobSat vs Comp

Model 2(train): JobSat vs WrkWkHrs

Model 3(train): Comp vs WrkWkHrs

Model 3(train): Comp vs NumLanguages

Discussion:

Our research question was answered, however, it doesn't seem that WWH and compensation have a direct relationship with job satisfaction based on our 2 models testing job satisfaction versus WWH and compensation. We strongly believe the data analysis technique we used was a good choice given our research question, as we were deciding between regression and clustering, thus the issue instead lies with our data. There was such a varying range of survey results for each comparison that no clear trends or correlations could be seen in our graphs. For example, in the category of job satisfaction and salary, there was a mix of programmers making tons of money and disliking their jobs to programmers making tons of money and loving their jobs. And in the category of job satisfaction and long hours, there was a mix of programmers working long hours and loving their job to programmers working long hours and hating their jobs. Because of this, no clear predictions could be made via this analyzation technique. There is always the possibility that a different technique could have been a slightly better fit, but overall our team thinks this dataset had too much variability and randomness for a good, clear analysis to be made.