BANA 7038
February 20, 2019

Dong, Juntao (M10453134)
Frankenhoff, Brenda (M01775924)

Homework 5

Write a proposal for the final project and make sure your proposal answer the following questions.

1. What kind of data set are you using?

We are using data from Stack Overflow from their 2018 survey asking the developer community about everything from their favorite technologies to their job preferences. We will be exploring the responses from developers in the data analysis and data science areas. Over 100,000 developers took the 30-minute survey in January 2018.

How many variables?

We will be using a subset of 12 fields from the original 129 possible variables. We only chose developers from the US. There were 184 countries represented in the original survey responses. Our final sample has approximately 2500 respondents/responses.

What is the response variable?

We are choosing Converted Salary (annualized salary in USD) as the response variable. We limited salary ranges from $20,000 USD to $250,000 USD as that is more reflective of industry reporting. We will be evaluating the following for our covariants:

- *Education*
- *Major*
- *DevType (filtered for our criteria)*
- *Years professional coding*
- *Job Satisfaction? (We might put this as a response variable)*
- *Languages*
- *Platform/Framework*
- *Daily hours on computer*
- *Gender*
- *Age*
- *Race/Ethnicity*
- *Company Size*

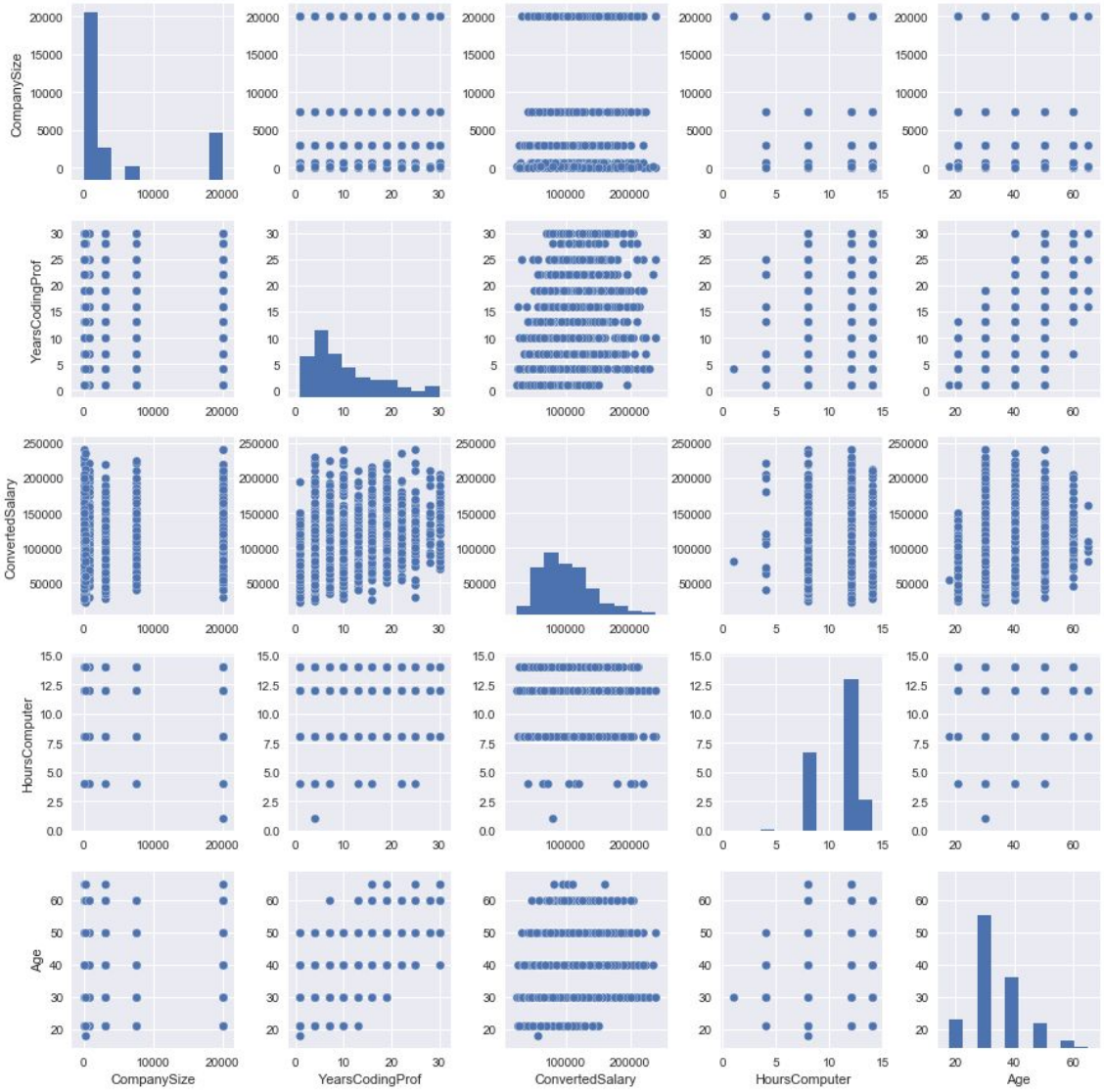
From the above fields, we consolidated values (grouped) for Gender, Age, Race/Ethnicity, Company Size, YearsCodingProf, and HoursComputer to reduce the number of possible values and make them numeric where possible.
For Languages, Platforms and Frameworks, we broke out into separate fields the various values with a 0/1 to indicate experience in that specific item. The final columns we will be evaluating include:

FormalEducation
UndergradMajor
CompanySize
YearsCodingProf
JobSatisfaction
ConvertedSalary
HoursComputer
Gender
RaceEthnicity
Age
Python
Scala
Matlab
SQL
Julia
Java
Salesforce
IBM.Cloud.or.Watson
AWS
Azure
Google.Cloud.Platform.App.Engine
TensorFlow
Torch.PyTorch
Spark
Hadoop
R

Dong, Juntao (M10453134)
Frankenhoff, Brenda (M01775924)

2. Exploratory data analysis of the chosen data set, i.e., figures and tables.

Pair plot of the numerical variables in the cleaned dataset.



3. What kind of models are you using?

We will start with multinomial logistic regression and also keep exploring other models to see which one works best.

4. What are the potential problems in your model? For example, skewness, normality, nonlinearity, multicollinearity, heteroscedasticity, dummy variables, outliers and so on.

The potential problems in the model are skewness (ex: several numerical features have skewed distribution - ConvertedSalary, YearsCodingProf and Age), nonlinearity (ex: several features may not linearly affect salary - Age appears to have a polynomial relationship), multicollinearity (ex. Age and YearsCodingProf), heteroscedasticity (ex:

the variability of the response variable, ConvertedSalary, is unequal across the range of values of several variables that predicts it - YearsCodingProf, HoursComputer and Age), too many dummy variables and outliers.

5. What kind of remedies are you proposing to use to solve the potential issues?

We will use careful feature selection by reviewing initial analysis before validating our model.
We will consider using data transformation where possible or required.
We will use Ridge Regression if we should run into multicollinearity.
We will explore the use of machine learning techniques to see if they will benefit our project.