Predictive Analytics Midterm:

For all questions 1-4, please provide the answer and 2-3 sentences around the rationale of why that is the answer. For answers 5-8, you will need to provide a separate document of the code.

1). What would be the rationale for using median over mean when trying to clean up missing values in a dataset?

2). At what statistical value would you consider a value to be an outlier. Walk me through a scenario where you would keep the outlier and one here you would delete it (each scenario should be 2-3 sentences)?

3). In order from most important to least important when interpreting a model, how would you order the following: Coefficients, AIC, and P-Value?

4). If you are looking at a confusion matrix and are trying to solve the classification problem of approving a loan for someone or not, which would you want to minimize (assuming 1 that they get approved for a loan and 0 that they don’t), False Positives or False Negatives?

The next 4 questions are based on the dataset attached in the Module. Your goal is to predict *price* which is a binary variable on how high the price is (1 being high range, 0 being low range)?

5). In 2-3 sentences state what the business problem is and what a successful outcome of this model would be if you are helping a company like Verizon understand why this data could be useful?

6). Which columns have missing or unclean data? How would you handle cleaning those up. Discuss in 2-3 sentences and have the code in a separate document.

7). Model the dataset to predict *price*. Which three variables have the lowest p-value (closest to 0)? If there are more than three (tie) rank the top three by highest coefficient (by absolute value)? For this response, please provide the variables, p values and coeficicents (in absolute and non absolute form).

8). What was the overall accuracy of the model? Additionally, if you are a phone manufacturer using this model to understand what differentiates a high value phone vs low value, which square in the confusion matrix would you want to optimize to reduce and why? For this response, provide the following:

1. Confusion Matrix 2. Accuracy Score and 3. 2-3 sentence on which square to optimize and why.