Problem 1:

I used basic functions like shape, dtypes, describe and isnull from pandas to explore the data. The data contains 205 samples and each sample has 26 attributes. There are three types of data, which are integer, object, and float. Also, no null values are existing in the data.

Problem 2:

In order to get an appropriate matrix to do linear regression, I decide to clean data first including converting strings to float and encoding categorical data.

For the ‘?’ values, I replaced them with mean values on columns with 5 - 10 ‘?’s. Although mean imputation reduces variance, but it is very fast and in our case variance does not vary too much. As the ‘?’ values in normalized-losses column is over 20%, and these null values are randomly distributed. I drop this column at first for simplicity. And I used linear regression to predict these values at last of the notebook.

For categorical data encoding, I applied one-hot encoding on columns with categories less than 4, label encoding on less than 7, and binary encoding on else. The main reason is to reduce dimensional space. Also, because there is no relation or order between these classes, so the algorithm can consider them as some order.

Problem 3:

The train\_test\_split function from sklearn was used to split data.

Problem 4:

Since the matrix contains multiple features, I decided to use multiple linear regression. For the regression plot, I found the most important feature among all features to draw a plot vs price. As the data is not standardized, the coefficient value changes greatly while the importance of the variable remains constant. So, larger coefficients don’t necessarily identify more important predictor variables.

Problem 5:

LR with l2 regularization (Ridge) was applied. The score is better than the LR without L2 regularizer. Using grid search to find a best value of alpha for the data.

Problem 6:

MinMaxScaler from sklearn was used to normalize data. the most dominated feature fit the target value pretty good after normalization.