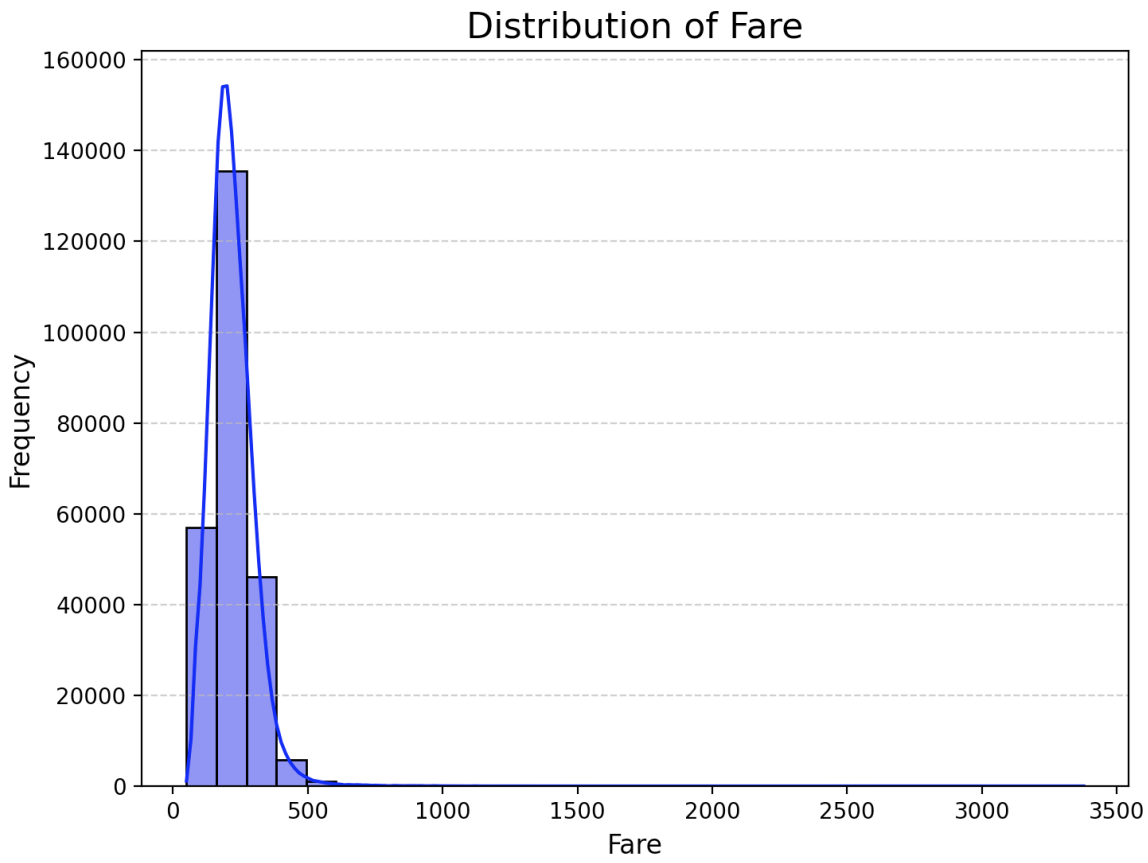
US Air Travel: Data Visualization and Linear Regression Model

Xuchao Zhou

Data Visualization

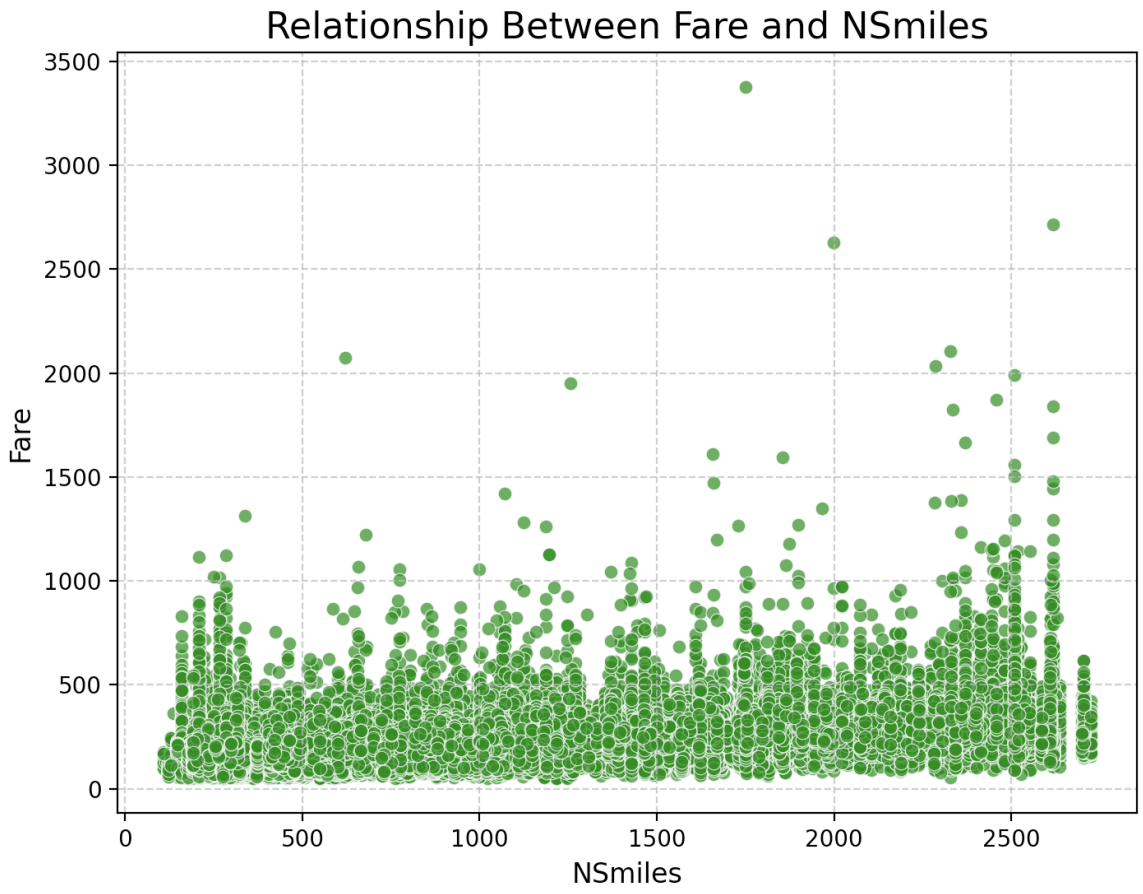
To better explore the dataset and discover any potential relationships that I may be interested in, I made three visualizations that demonstrate several key pieces of data. The first visualization is a histogram that shows the distribution of air travel fare, which is an important indicator of the market and may be chosen as the response variable in our project:



As we can see, the average air ticket price for a trip within the US mostly falls between 0 and 500 dollars, and the price interval with the highest frequency is approximately $200 - $300. With this information, we now move to see if the average fare is affected by the travel distance (which is, by common sense, what should be an important factor when airlines make the prices):

US Air Travel: Data Visualization and Linear Regression Model

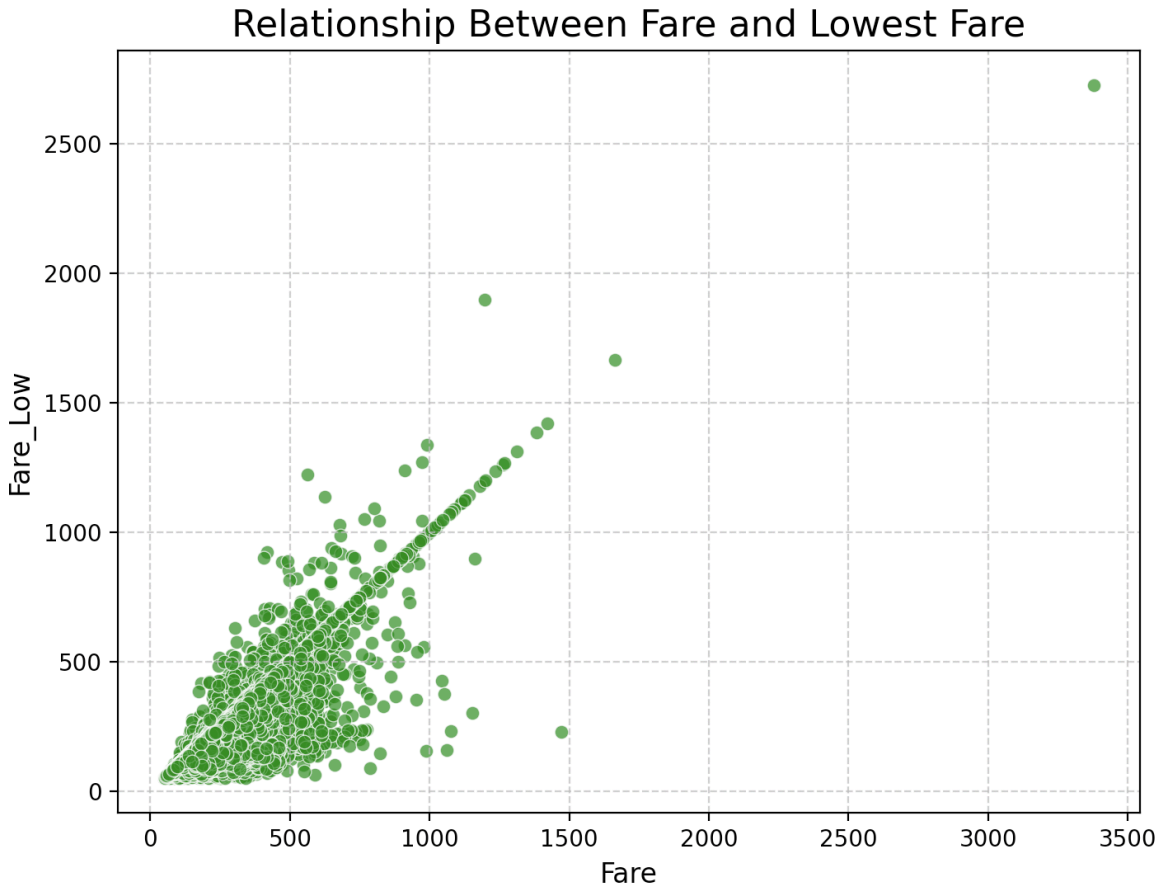
Xuchao Zhou



Unfortunately, the scatterplot shows weak or no relationship between fare and NSmiles (the travel distance variable in the dataset). Fare seems to be equally distributed at any travel distance, and there is no significant pattern that indicates a fare change as the travel distance gets longer. This indicates that the travel distance is not the only factor that contributes to the fare of a trip, and that other factors should be considered when we are training the model that predicts the trip fare. In addition, when buying air tickets, we are also interested in the lowest price (especially for budget travellers), so we go on to look if there is any relationship between the average and lowest fare on a flight:

US Air Travel: Data Visualization and Linear Regression Model

Xuchao Zhou



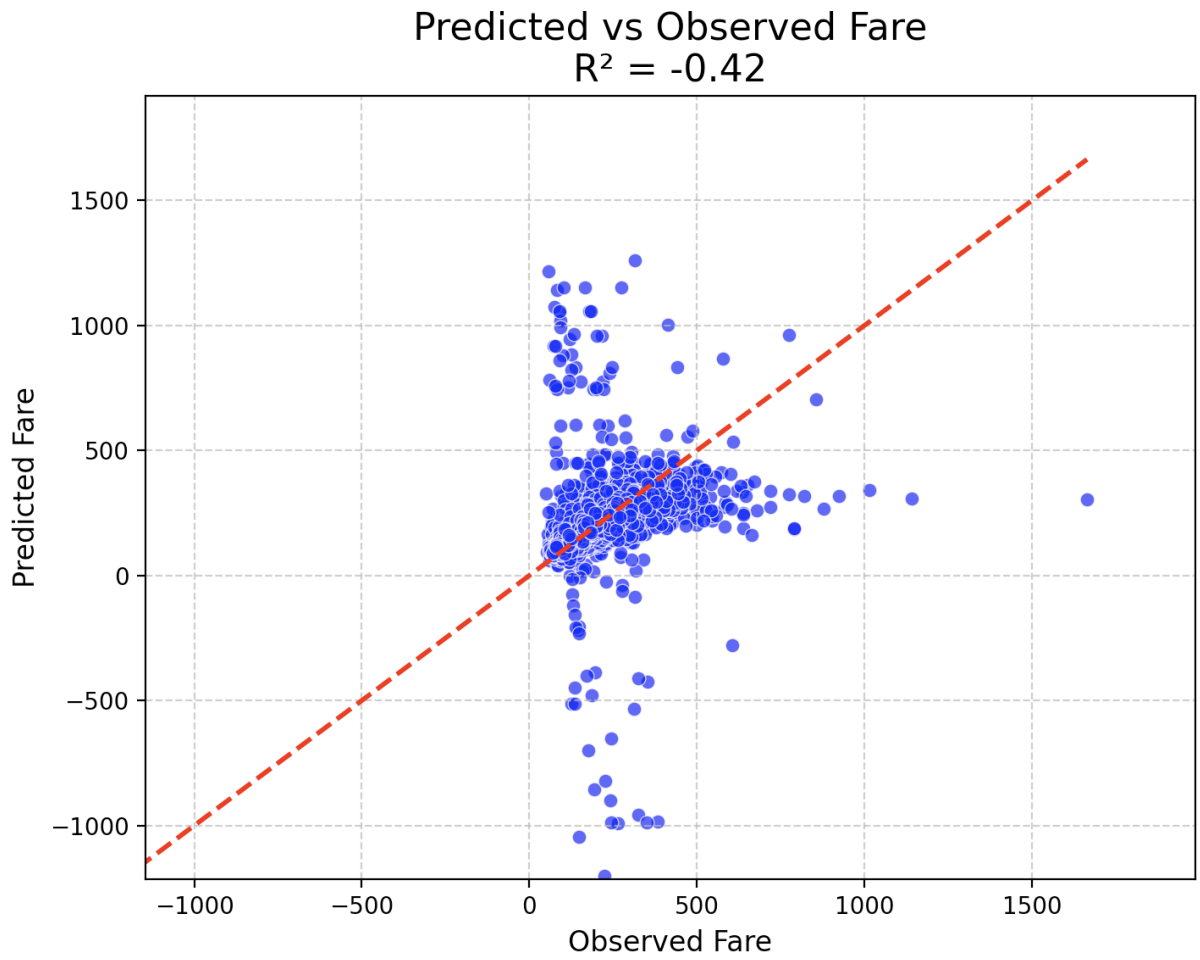
This figure confirms my hypothesis that there is a positive relationship between the average fare and the lowest fare on a flight: points on the graph with higher average fare tend to appear with a higher lowest fare.

Linear Regression

Carry on from what we get from the previous section, we now try to fit a linear regression model on our dataset that has the variable “fare” (average fare on a trip) as the response variable and the variables “nsmiles” (travel distance of a trip), “passengers” (number of passengers on a trip), and “route” (this variable is generated by combining the airport code of the departure airport and the arrival airport) as features (predictor variables). We then generate a linear regression model using scikit-learn in Python and drew a predicted v. actual scatterplot for the data to evaluate the performance of the model:

US Air Travel: Data Visualization and Linear Regression Model

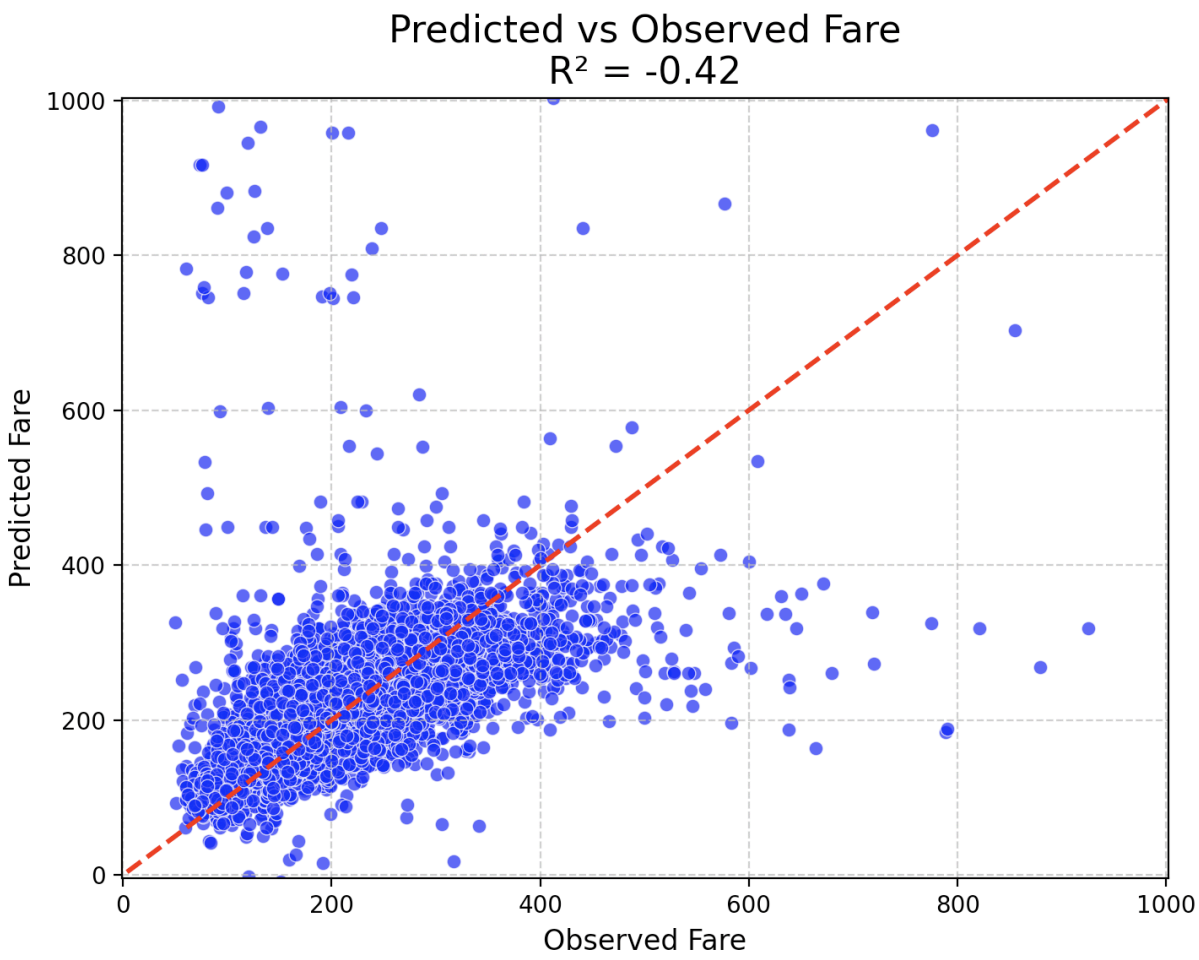
Xuchao Zhou



Both the plot and the R-squared value indicates that this linear regression model's performance is not satisfiable. There are a lot of points that are distant from the perfect-prediction line (the 45 degrees red line) which means that the actual value of fare varies a lot from our prediction; the negative R-squared value also indicates that this model's performance is worse than simply predicting the mean of the target variable. It suggests that we should consider using a different regression model or change the features used in this model. Nevertheless, if we zoom in the central region where the largest cluster of points locate, we can see a slightly different story:

US Air Travel: Data Visualization and Linear Regression Model

Xuchao Zhou



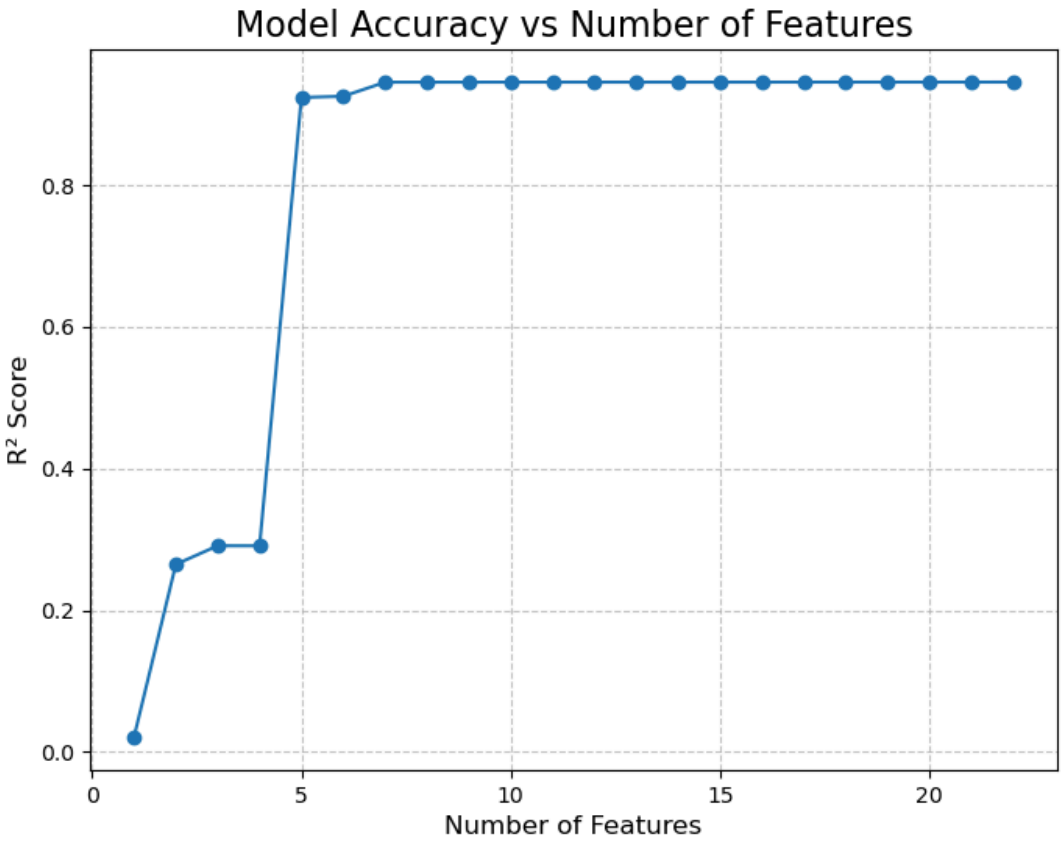
If we exclude those extreme outliers, the remaining part follows the pattern of the 45 degrees line, so we suspect that it is those extreme outliers that cause the bad performance of the model - among those outliers there is even negative fare predicted, which is impossible, and this suggests the necessity of reconsidering the model chosen.

Complexity vs. Accuracy

After applying the linear regression model, we move to explore the relationship between model complexity and accuracy. As we all know, more complexity (more features) does not necessarily mean more accuracy because of the overfitting issue. Here we use the R-squared value as the accuracy measurement of the model and the number of features as the complexity measurement. Here is the data we get:

US Air Travel: Data Visualization and Linear Regression Model

Xuchao Zhou



As we can see from the figure, the R-squared value pops up after we include 5 features, and it did not drop so there is no issue of overfitting. The R-squared value close to 1 indicates that all models with 5 or more features have good accuracy in predicting air ticket price.