Building the Model:

We built a simple supervised linear regression model of our data using 4 variables: item IDs, shop IDs, year, and month. We experimented with a variation of the item IDs and shop IDs, along with manually inputted values for the year and month variables (reflecting November 2015). Prior to model experimentation, we built our model using grouped category IDs and grouped shop IDs by grouping similarly performing unique IDs. We validated our groupings by calculating statistical significance between these groups using two-sample t-tests.

After careful experimentation, we realized that the shop and item category groupings yielded very poor results. In addition, we realized that it was computationally feasible to completely remove or increase the number of groupings. As a result, we attempted multiple different approaches including removing both groupings and grouping similarly performing item ids instead. We generated 24 groups for similarly performing item IDs.

Preprocessing:

Since we're working with four categorical variables, we generated dummy variables to use in both our training and testing. However, there was a choice to use the date as a continuous (time-series) variable. We chose to use categorical variables instead because we believed that, as a time-series, the date poorly reflected the yearly cyclical trends that we were observing. We may choose to return to this point in the future to retrain our model using the data as a time-series.

To appropriately train and test the model, we split the data into train and test data sets. Twenty percent of the data was allocated to the test data set with the randomly chosen random_state 21 to ensure reproducibility. An additional stratify argument along the month variable to ensure that the months were equally represented during the training.

Variable Experimentation & Accuracy Measurements:

Thus far, we have attempted a total of 5 variations of our model variables. The first model used the generated shop and item category groups. Using the double groupings, we observed that many of the predicted values, in both training and test sets, were right skewed. In addition, the plotted histograms show that there is an irregular amount of negative predictions. Using seaborn's regression plot, we also observed that the model underestimates the values and has a cutoff value for the max predicted values which is not reflected in the actual values (Figure 1). Root Mean Squared Error was used as the accuracy metric to reflect the Kaggle competitions method for determining score. Calculating the score with our split test data, we achieved a RMSE score of 426.39, whereas Kaggle scored the predictions as 1198.72.

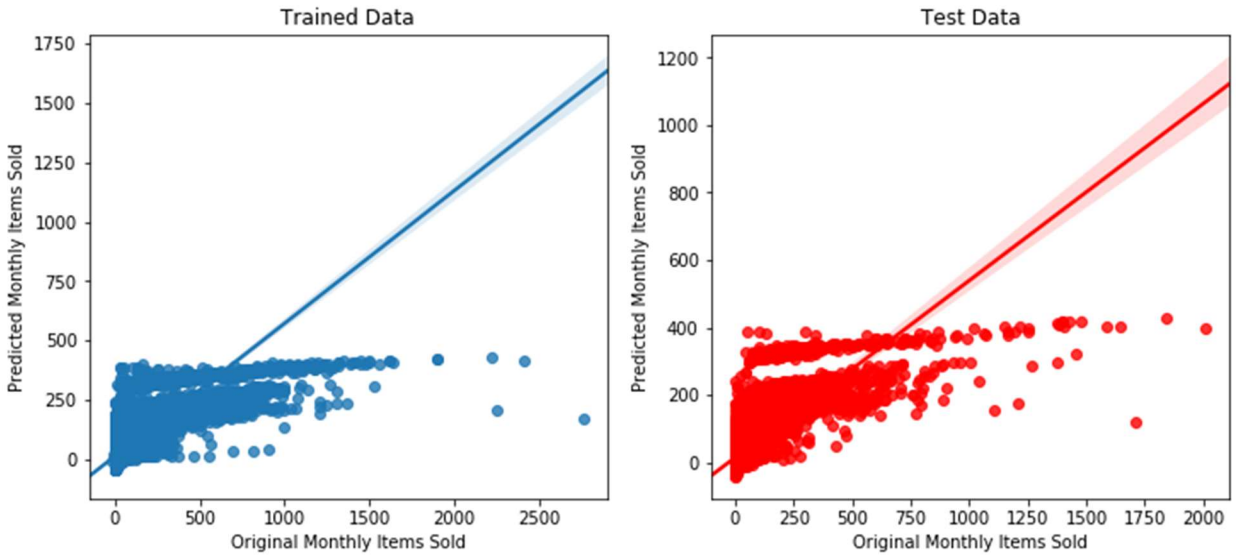

Figure 1. Regression plot of model 1. The model appears to consistently underestimate the values. There are also unexpected max value cutoffs in the predictions that are not in the original.

The second model scrapped the grouped shop IDs, but continued to use the item category groups. The predicted values continue to have a right skew along with a couple of negative predictions. From the regression plots, we observed a less drastic max cutoff value, but similar underestimated predicted values. Calculating the score with our split test data, we achieved a RMSE score of 154.30, whereas Kaggle scored the predictions as 243.86.

The third model scrapped both groups and used the individual shop and item category IDs. The plots follow similar trends to the two previous models (underestimated values along with a right skew). However, the RMSE score for the model was unexpected as the generated RMSE was 90.17 (an improvement from 154.30 from the second model) while the Kaggle scored the prediction as 260.18 (a drop from 243.86).

The fourth model and fifth models used grouped item IDs to train the model. The data for the fifth model was averaged across the time span of the data (the three years were combined). Both scored better than the previous three models, but was far more than the desired score (Figure 3).

Model:	Changed Variable:	Generated RMSE:	Kaggle RMSE:
1	Grouped Shop ID and Item Category ID	426.39	1198.72
2	Scrapped Shop ID	154.30	243.86
3	Scrapped Item Category ID	90.17	260.18
4	Shop IDs and Grouped Item IDs	60.45	152.59
5	Averaged Data and removed year variable	47.46	161.70

Figure 2. Table describing the changes made between each consecutive model. Scores from both Kaggle and split test data are provided as an accuracy measurement.

Noteworthy:

Only the second model had a regression plot that did not have an irregular max cutoff value seen in all the other models (Figure 1 vs. Figure 3).

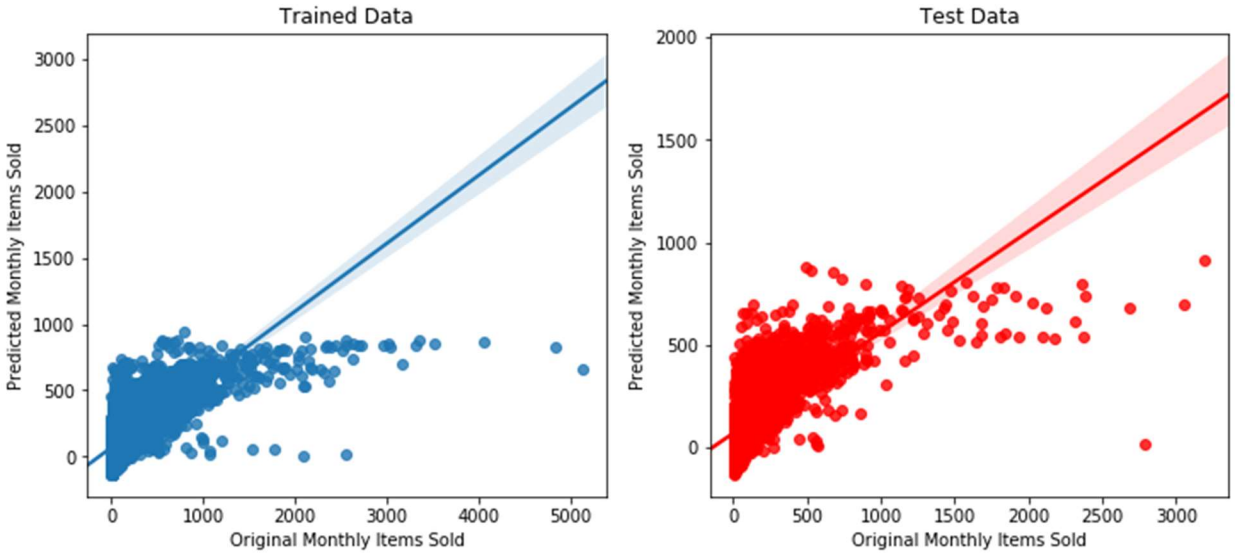

Figure 3. Regression plot of the second model. This is the only model that did not have that irregular y-axis cutoff seen in all the other plots.

Future Direction:

Provided time, we would first attempt to add preprocessing steps from the sklearn.preprocessing module to further clean up the data for our model. In addition, we would explore different classifier or combinations of classifiers to better reflect the model. A possible classifier is the Support Vector Regressor with a polynomial kernel which might help address some of the non-linear aspects of our data. It's also important to optimize the parameters of the classifiers that we choose.

Next, we'd go back to the beginning and attempt different methods of cleaning the data. Especially where we rejected all entries with a negative item price or even left all the negative item counts. The negative values for these variables could have a meaning that was not described by the provider. There are other approaches, but we will not be attempting them at this time.