Exploratory Data Analysis and Inferential Statistics:

The purpose of this project is to build a model that can predict the performance of a specific item in a specific shop. The test data provides us two categorical columns: shop_id and item_id. After exploring the datasets, there are a couple of things we can note. The most relevant is that there are nearly 22,000 unique item ids. Considering that we need to use dummy variables to apply categorical data into our linear regression model, it is impractical and too computationally intensive to work with that many values. However, these item ids are also separated into item categories, which have 84 unique ids. At this point, we must acknowledge that by using item category ids to replace individual item ids, we are allowing our model to assume that all items in each category do not have any statistical differences from each other. This assumption can be fatal as performance can vary heavily in each category.

Independent Variables:

The main variables we will be working with to build the model are item category ids and shop ids. Both these variables are categorical data and contain 84 and 60 unique IDs respectively. To reduce the number of dummy variables we need to create and work with, we created groups for categories and shops that performed at approximately the same level. The result for the item categories was 6 groupings and 11 standalone categories to give us a total of 17 values to work with. For the shops, there were 5 groupings and 8 standalone shops to give us a total of 13 values to work with.

Statistical Tests:

To validate the groupings for both variables, we utilized two-sample t-tests. In our t-tests, the null hypothesis was that there was no statistical difference between our groupings, and alternative hypothesis being there was statistical differences. Using a significance threshold of 0.05, we were able to validate our groupings. Due to a t-score strongly supporting statistical differences, I did not feel it was necessary to obtain further support via the frequentist approach.