Predict Future Sales

Data Wrangling

This document describes the various data cleaning and data wrangling methods applied to daily historical sales data. The following sections are based on the

Missing Data

There are no Missing data - NaNs in our train data set

Outliers

There are outliers present in the item_cnt_day and item_price of train data set

Datatype

date column of train data has been converted to pandas datetime format

Duplicates

There were 6 duplicates were found and looks like they were bought twice due to their popularity

Negative item sold count

We see -1 in th count , assuming that the product was returned

Memory

To keep the memory space allocated more than necessary, downcasting the datatypes to minimum requirements. We see that the memory usage of training dataset from 134.4+ MB has come down to 61.6+ MB

General Observations

- There are 22170 items in the catalogue and 60 shops
- The total possible number of combinations 22170*60 = 1330200
- The train set consists of 60 unique shops and 21807 items
- The test set consists of 42 unique shops and 5100 items
- There are 84 different categories in the catalogue
- There are 2,14,200 items altogether in the test file, we need to predict items sold per month for the 60 shops
- item_cnt_month from samples_submission gives us the idea of how many items are sold in a shop per month(this is given in test)¶
- The data is for 34 months , for 60 shops and around 2.9 million(2935849) items
- Train data is Daily Historical data from January 2013 to October 2015