**Level 1 Data Assessment - General Quality and Tidiness Assessment**

The dataset has 3 years worth of daily sales data extending from January 2013 to October 2015. Each data point represents the sale of an item in a shop on a particular date. There are two million nine hundred thirty-five thousand eight hundred forty-nine observations in 6 features.

There are 21,807 items sold in 60 shops. The items are divided into 84 item categories.

The sale in different shops is certainly not uniform. Some shops sell better than others (Shops with ids from 20 to 35 do very well. The shop\_ids less than 20 does worse than shop\_ids more than 35).

Looking at the sales per date-block, we can see that the sales have decreased over the years from 2013 to 2015 with a couple of spikes in between. The year 2013 begins with 200K+ items sold and at the end of 2015 we see it almost halved with 100K+ items sold.

The requirement is to predict monthly sales of each item in each shop.

During the first level assessment, we have identified a few quality issues which need to be rectified. Along with this, we also have some transformation ideas as well that will derive more information from the dataset.

Training Set

Quality Issues

1. Round all float values in the dataframe to two decimal places
2. Remove the row with item\_id = 0
3. Remove 6 duplicate rows
4. There is a row with index 484683 where the item\_price is negative
5. Downcast dataframe to save memory
6. Remove rows where items have been returned (7356 rows). These are rows with negative values
7. There are about 27411 rows which is 0.93% of the rows where item\_cnt\_day is greater than or equal to 6. Remove rows where item\_cnt\_day >= 6. (review row numbers)
8. There are about 253150 rows which is 8.62% of the rows where item\_price is greater than or equal to 2124. Remove rows where item\_cnt\_day >= 2124. (review row numbers)

Tidiness Issues

1. Combine sales\_train and item\_categories with item\_id

Add New Features

1. Item\_price\_class - Divide item prices into 4 classes
2. Split date column - day, year, month, weekday/weekend, day of the week

**Level 2 Data - Assessment - Relationships between variables**

Add New Tables

1. Sales frequency tables
   1. Shop
   2. Item
   3. Item\_category
   4. Day of the week
   5. is\_weekend

Data Types of Features

1. Date - Interval - Discrete
2. Date\_block\_num - Interval - Discrete
3. Shop\_id - Nominal - Discrete
4. Item\_id - Nominal - Discrete
5. Item\_categories\_id - Nominal - Discrete
6. Item\_price - Ratio - Continuous
7. Item\_cnt\_day - Ordinal - Discrete
8. Item\_price\_class - Ordinal - Discrete
9. Day - Interval - Discrete
10. Month - Interval - Discrete
11. Year - Interval - Discrete
12. Day\_of\_week - Nominal - Discrete
13. Is\_weekend - Nominal - Discrete

Part 1 - Questions and observations

1. Item Categories
   1. How many items are there in items\_category?
      1. The number of items in each of the 84 categories is extremely varied
   2. What does the sale of items in each price class look like?
      1. Overall, the number of sale in each categories looks somewhat uniform with number of sales ranging from 0 to 50,000 units
      2. In some categories, there are more items in the cheaper price class than those in the expensive classes.
         1. Item category 40 has around 300,000 items in class 1, 250,000 items in class 2
         2. Similarly item category 30, 40 and 55 have more number of items in items in class 1 and class 2
         3. Analyze IDs -
            1. 40 - Class 1, Class 2
            2. 55 - Class 1, Class 2
            3. 37 - Class 1, Class 2
            4. 30 - Class 1, Class 2
      3. The categories that have sold more expensive items are
         1. 18 - Class 3, Class 4
         2. 20 - Class 3
         3. 22 - Class 3, Class 4
         4. 28 - Class 3
         5. 37 - Class 3
         6. 30 - Class 3
      4. Find trends in the pricing of categories
         1. The variation in median prices is highest in the expensive classes and the lowest in the cheaper classes
            1. Class 1 - 100 to 300
            2. Class 2 - 300 to 600
            3. Class 3 - 750 to 1000
            4. Class 4 - 1250 to 1500
2. Shops
   1. What does the sale of items in each price class look like?
      1. Shops of interest listed with decreasing order in number of sales
         1. 31
         2. 25
         3. 54
         4. 57
         5. 28
         6. 42
         7. 51
         8. 56
         9. 58
   2. Find trends in the pricing in shops
      1. Most shops sells items in the range of 200 to 400 median price
      2. The number of shops in that sell the most expensive items are comparatively less and the median price of items are also highly varied
3. Price Class
   1. How are prices distributed in the 4 price classes?
      1. The prices of items in class 1 and class 2 seem to be uniformly distributed
      2. When it comes to class 3 and class 4, they exhibit higher variation
4. Date Block Num
   1. The sale of items in the final year, 2014 to 2015 shows a higher amount of variation than the rest of the year
   2. The sales across classes started decreasing from the beginning of 2014.
      1. Compared to all other classes, sales in Class 3 seems to be doing better
   3. The number of sales in December is better for 2013 and 2014. There is no data for 2015 December
5. Month
   1. The highest sales are during the months:
      1. 12 - 3, 2, 1, 4
      2. 2 - 1, 3, 2, 4
      3. 3 - 1, 3, 2, 4
      4. 8 - 3, 1, 2, 4
   2. The lowest sales is during the months:
      1. 11 - 1, 3, 2, 4
      2. 10 - 1, 3, 2, 4
      3. 9 - 1, 3, 2, 4
      4. 4 - 1, 3, 2, 4
      5. 5 - 1, 3, 2, 4
   3. The items expensive price class is sold the least in all the months
   4. The variation of prices is highest in the month of December
      1. Class 3 items are sold the most during this month
6. Day of the week
   1. The highest sales is on
      1. Saturdays - 1, 3
      2. Sundays - 1, 3
7. Has the number of shops decreased over time?
   1. After the first year, the number of shops increases
   2. The number of contributing shops have decreased slightly in the last year but it does not have 2 months of data. There is a decrease in the number of shops in the data that is present
8. Has the number of shops decreased over time?
   1. Yes, the number of items placed in the shops decreased from close to 8000 products to close to 6000 products over the date blocks
9. Has there been a decrease in price over the years?
   1. Over the years, different items have varied in prices
10. Do different shops sell the same items for different prices?
    1. Yes
11. For each item, find the average price, its average sales per day and check if these are correlated
    1. There isn’t a correlation between price and the average number of items sold
12. For each shop, find the average price of the items it sells and the average number of sales per day

Insights:

1. By the end of 2015, the number of sales had decreased to less than half of 2013.
2. The item prices show more variation in the months of 2015 than any other year
3. Price class 3 and 1 have the most popular items
4. Most items are sold on day 6 (close to 3 times compared to day 0)
5. Items in Price class 4 are the least bought
6. Every shop does not sell the same items with the same price. Over time, different items have been sold under different price ranges by different shops
7. The variety of items have reduced since the end of 2013 and has been steadily decreasing

Second Level Cleaning

1. Combine item\_cnt\_day to and date\_block\_num to get 'item\_cnt\_month'