Predict Future Sales

The main objective is to build a model that can predict the sales of an item in the shop for the given month. The test data has two categorical columns: shop_id and item_id.

Dependent Variables

Sales data, which is Item unit times prices

Independent Variables

Time period and the dummy variables(on item categories, shops)

The item categories and shops dataset contains categorical data.

The item categories dataset contains 84 unique item category values. They can be further interpreted as category type with 22 and category subtype with 66 unique values.

The shops dataset contains 57 unique shop values. They can be further interpreted with location as city with 31 unique values.

Statistical Tests

In the training data set the average sales price was: 932.1405, total items sold: 3646036 and total sales value: 3398617900.

Single sample t-test: Average price per sale is equal to 800

- Null Hypothesis H0 - Average price per sale is equal to 800
- Alternative Hypothesis H1 - Average price per sale is not equal to 800

Interpretation: Since p-value is 0, H0 is rejected. We conclude that Average price per sale is not equal to 800

Independent sample t-test: Average price per sale of the cities Москва & Н.Новгород are compared

- Null Hypothesis H0 - Average price per sale is the same for the two cities - Москва & Н.Новгород
- Alternative Hypothesis H1 - Average price per sale is not the same for the two cities - Москва & Н.Новгород

Interpretation: Since p-value is nearly 0, H0 is rejected. We conclude that Average price per sale of the two cities are different.

ANOVA: Average sales is different for 12 months

- Null Hypothesis H0 - Average sales for 12 months are same
- Alternative Hypothesis H1 - Average sales is different for atleast 1 month

Interpretation: Since p-value is 0, H0 is rejected. We conclude that Average sales is different for atleast 1 month.

Chi square: Number of total Item sales per month is Uniformly distributed or not

- Null Hypothesis H0 - Number of total Item sales per month is Uniformly distributed
- Alternative Hypothesis H1 - Number of total Item sales per month is not Uniformly distributed

Interpretation: Since p-value is 0, H0 is rejected. We conclude that Number of total Item sales per month is not Uniformly distributed

Linear Regression:

Predicting Average sales price per sales from month(date_block_num).

The estimated regression line is

avg_sales_per_block = 809.3849 + (22.6911 * date_block_num)

The regression model is significant , F = 28.07, P = 0.000

Regression equation is used to predict the next two month blocks - 34 & 35

Predicted avg_sales_per_block for the block 34 is 1580.8823 and 35 is 1603.5734