

SPRINGBOARD CAPSTONE PROJECT

---

**PREDICTING FUTURE SALES**

# RETAIL BUSINESS MODEL

- ▶ Stock stores or warehouse with merchandise for customers to make purchases in store or online.
- ▶ Make forecast of future demand to make more accurate stocking.
- ▶ Conventionally, companies use historical data to make forecast.



# THE PROBLEM

- ▶ Under stocking can lose out on potential revenue.
- ▶ Overstock can result in too much merchandize in stock and fail to turn goods into revenue.
- ▶ For electronic stores, demand is constantly shifting. Therefore, historical data is not always reliable.
- ▶ Changing demand can be hard to track and predict.

# DATA ACQUISITION

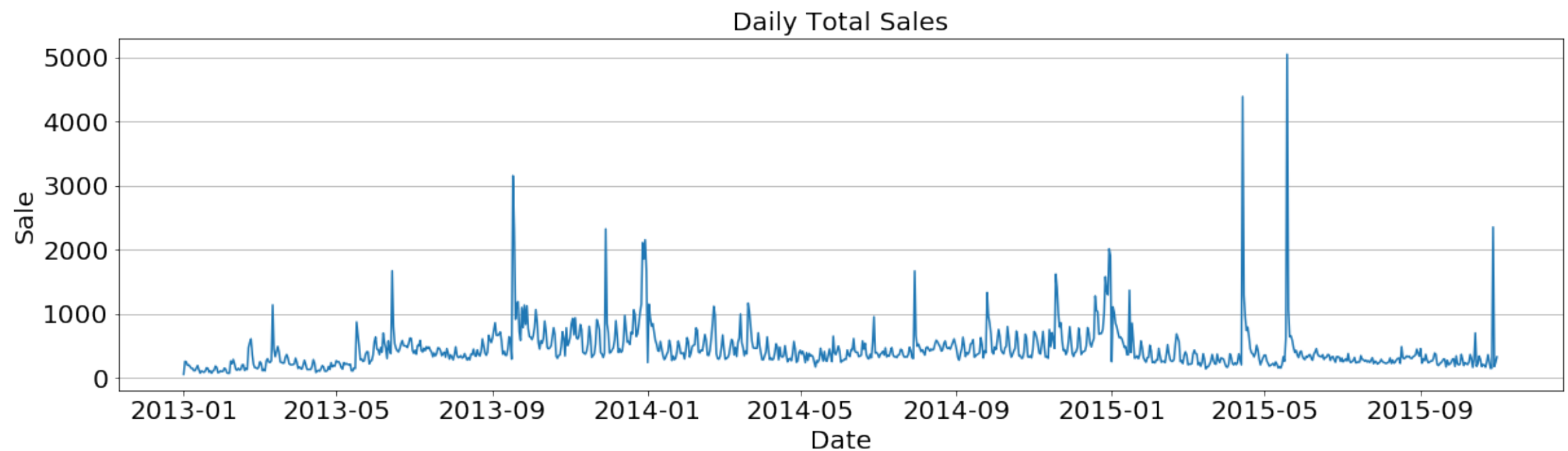


- ▶ Dataset used in this project is featured in a Kaggle challenge: Predict Future Sales - Final project for "How to win a data science competition" Coursera course.
- ▶ Over 2,935,849 entries of daily sales
- ▶ Over 20,000 unique item IDs (SKUs) in 50 stores
- ▶ Data span from January 2013 to October 2015
- ▶ Each data point contains date, store ID, item ID, item price, number of units sold

# EXPLORATORY DATA ANALYSIS

# TOTAL SALES

- ▶ Shows higher daily sale volumes between September and February each year.



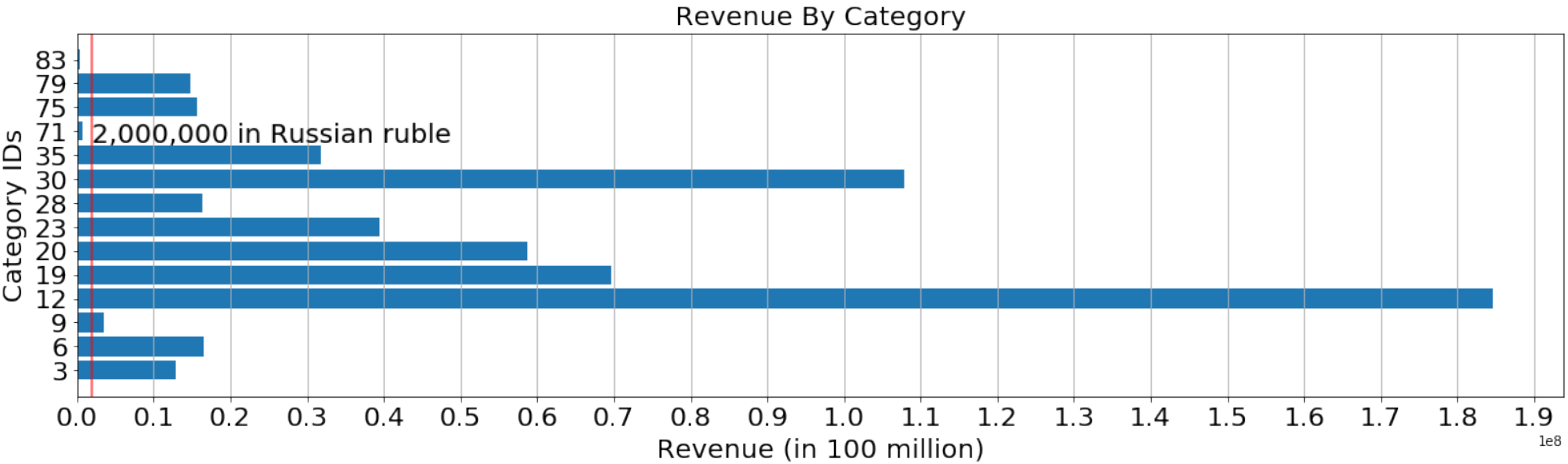
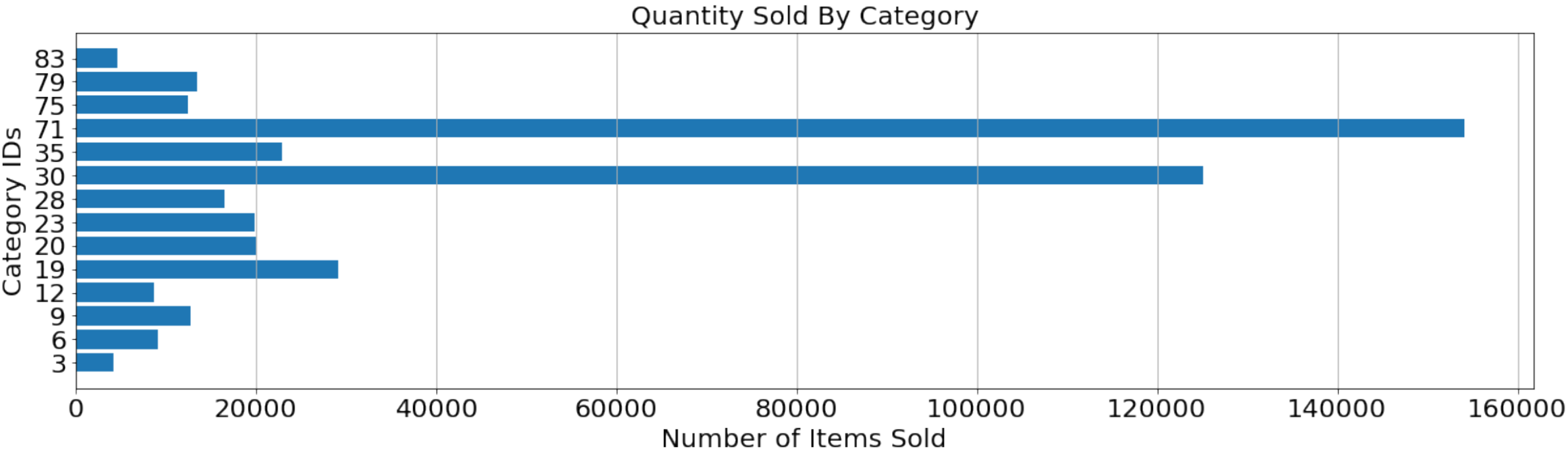
## NEGATIVE SALES

- ▶ Negative sales in some daily and monthly totals, assuming returned items

Number of items have negative sales in a given month: 25

	date_block_num	shop_id	item_id	item_cnt_day
4604	7	2	6457	-1.0
4751	7	12	1830	-1.0
4754	7	12	2753	-1.0
4758	7	12	5822	-1.0
4760	7	12	6740	-2.0
4762	7	12	7894	-1.0
4765	7	12	15044	-1.0
6694	9	12	1830	-1.0
7786	10	12	3329	-1.0
7788	10	12	3732	-4.0
7789	10	12	3734	-2.0
11083	12	56	6675	-1.0
13864	15	6	2753	-1.0
16517	17	12	1905	-1.0
18098	18	24	4870	-1.0
20806	20	21	1830	-1.0
26566	24	12	3733	-1.0
28218	25	19	7018	-1.0
31785	28	4	3733	-1.0
35087	30	41	5672	-1.0
35613	31	4	6675	-1.0
35771	31	12	6497	-1.0
36900	32	5	6738	-1.0
36995	32	12	2808	-1.0
37900	32	57	6675	-1.0

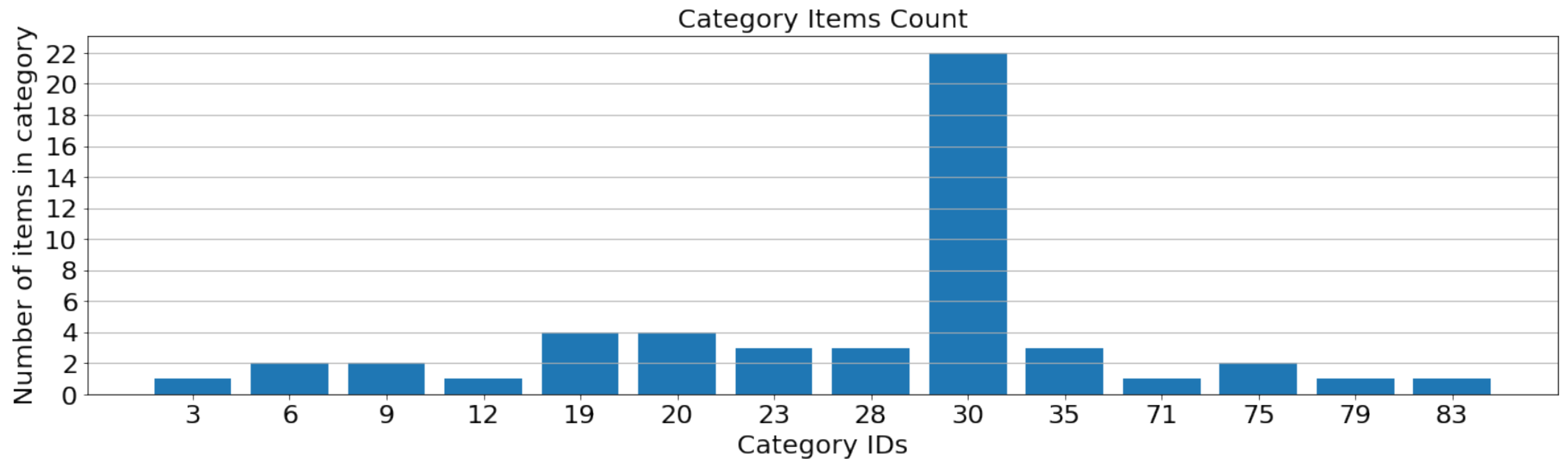
# EDA - CATEGORY





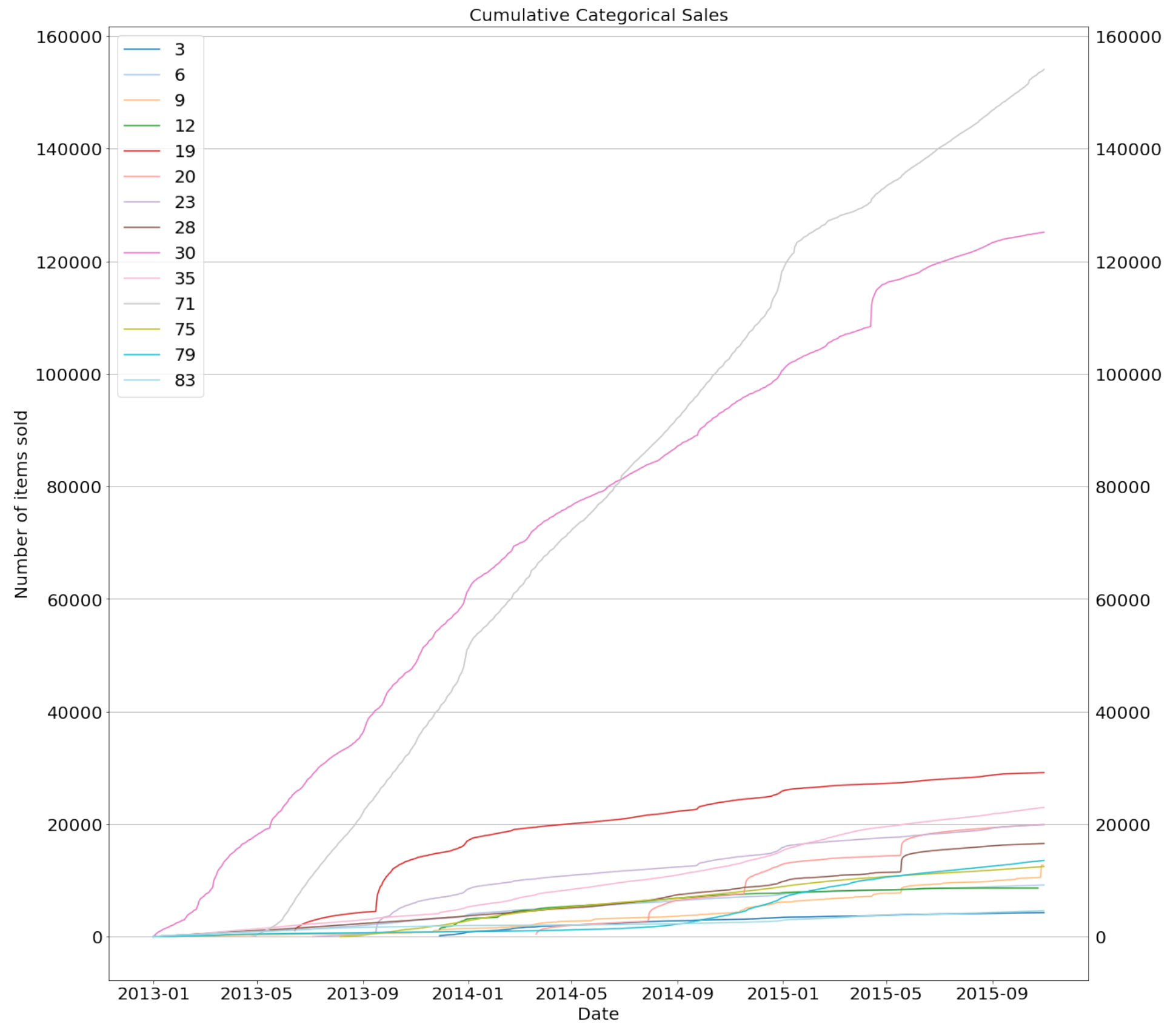
## EDA - CATEGORY

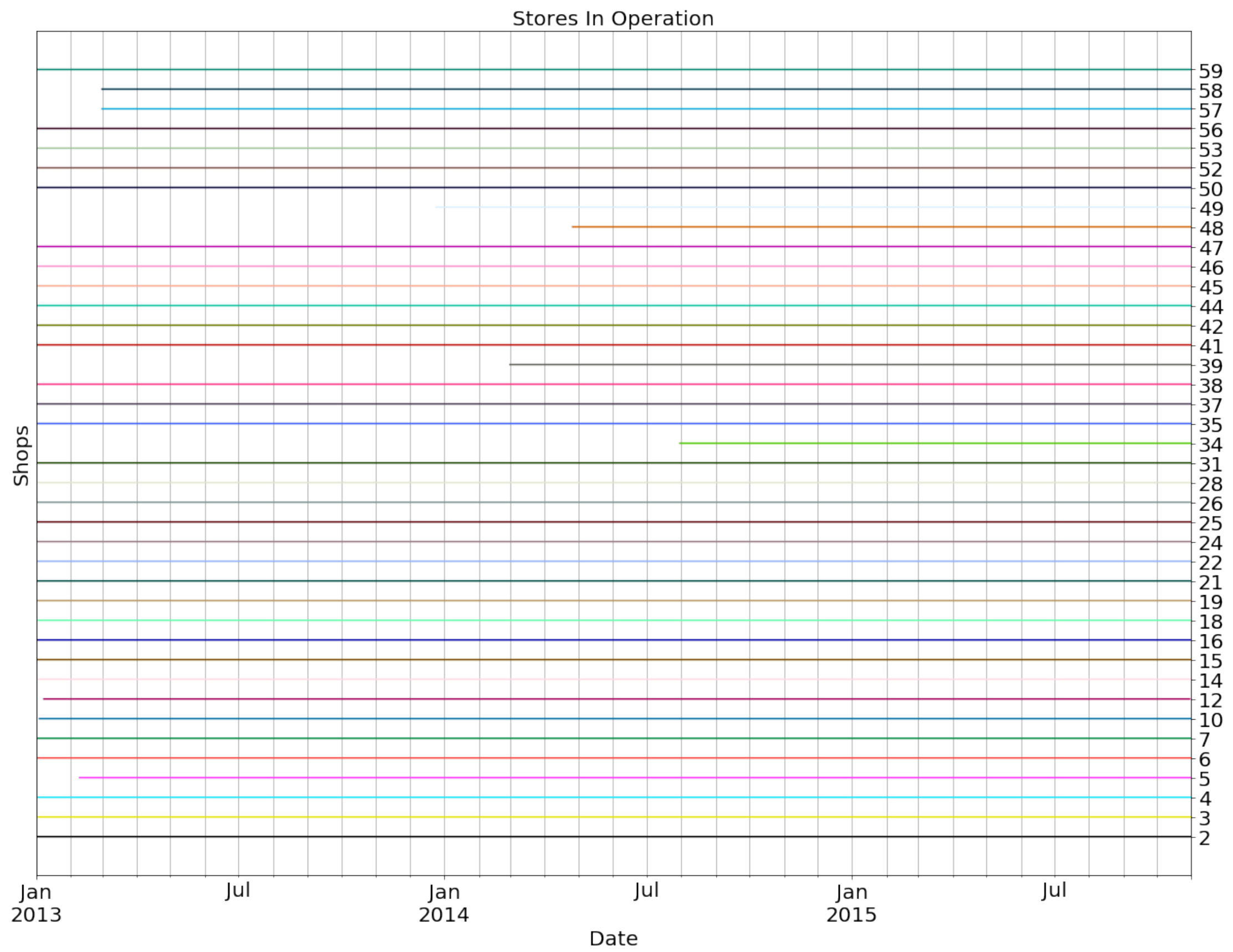
---

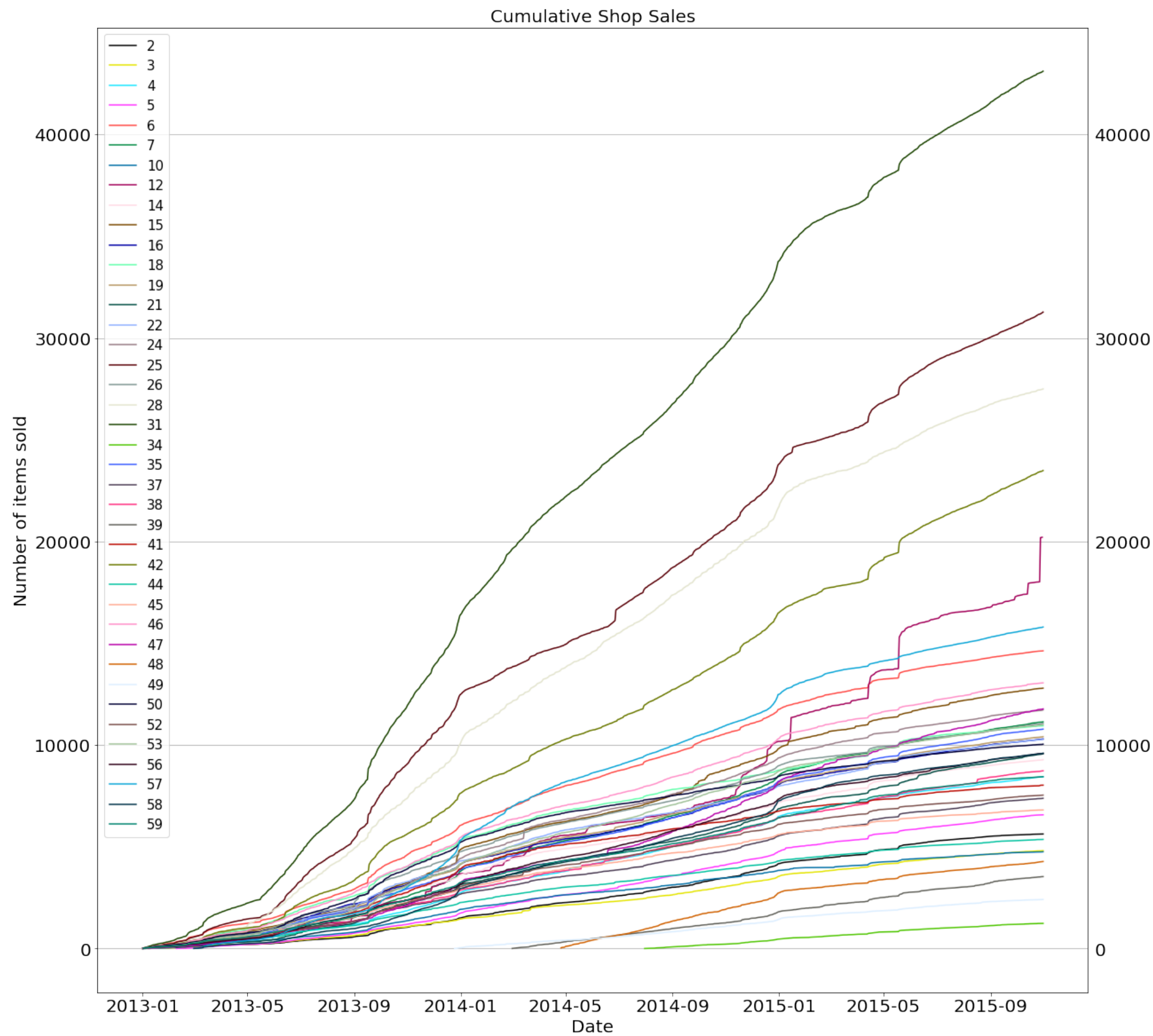


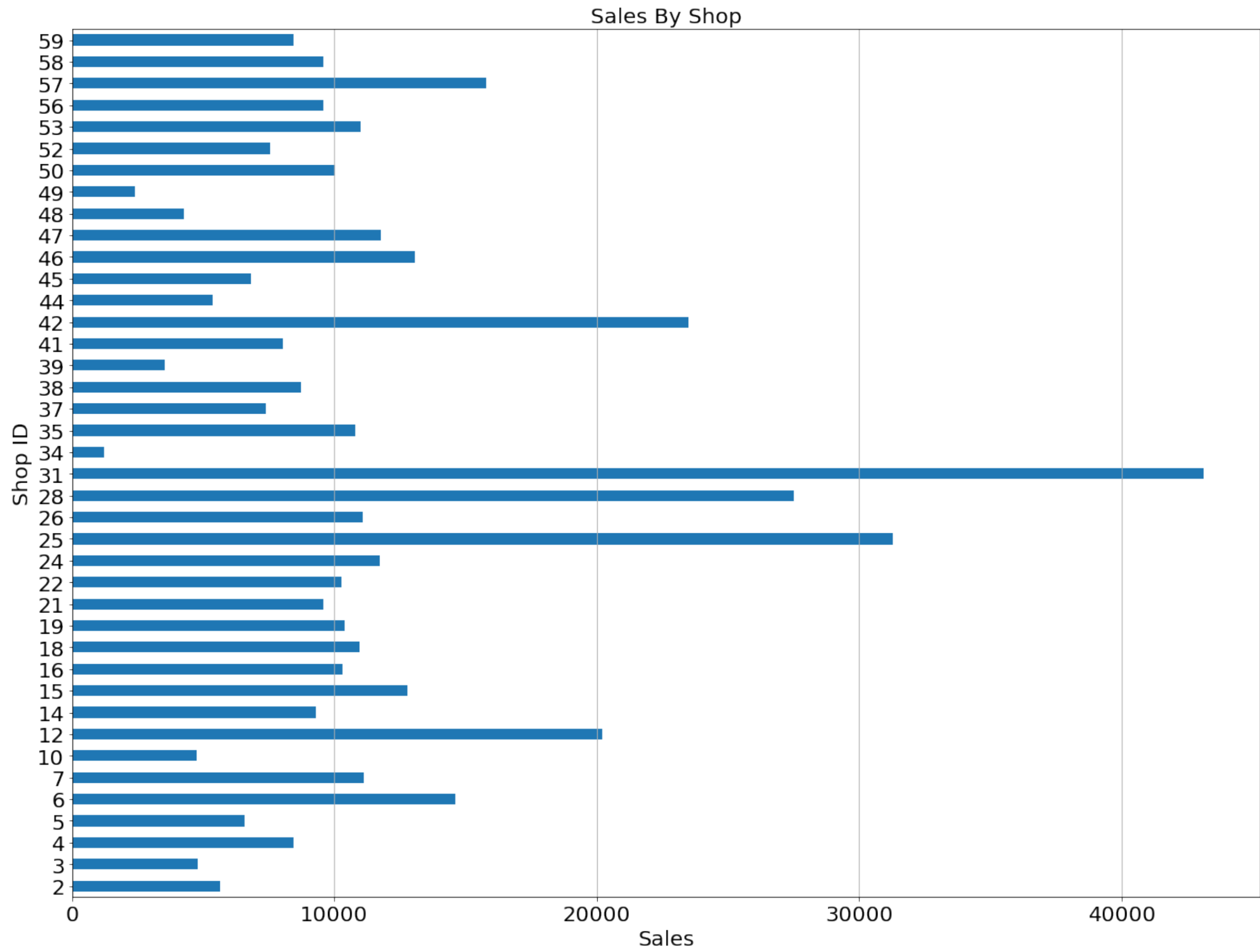
- ▶ Category 71 has only one item ID and has the most units sold than other categories
- ▶ But the revenue from category 71 is extremely small

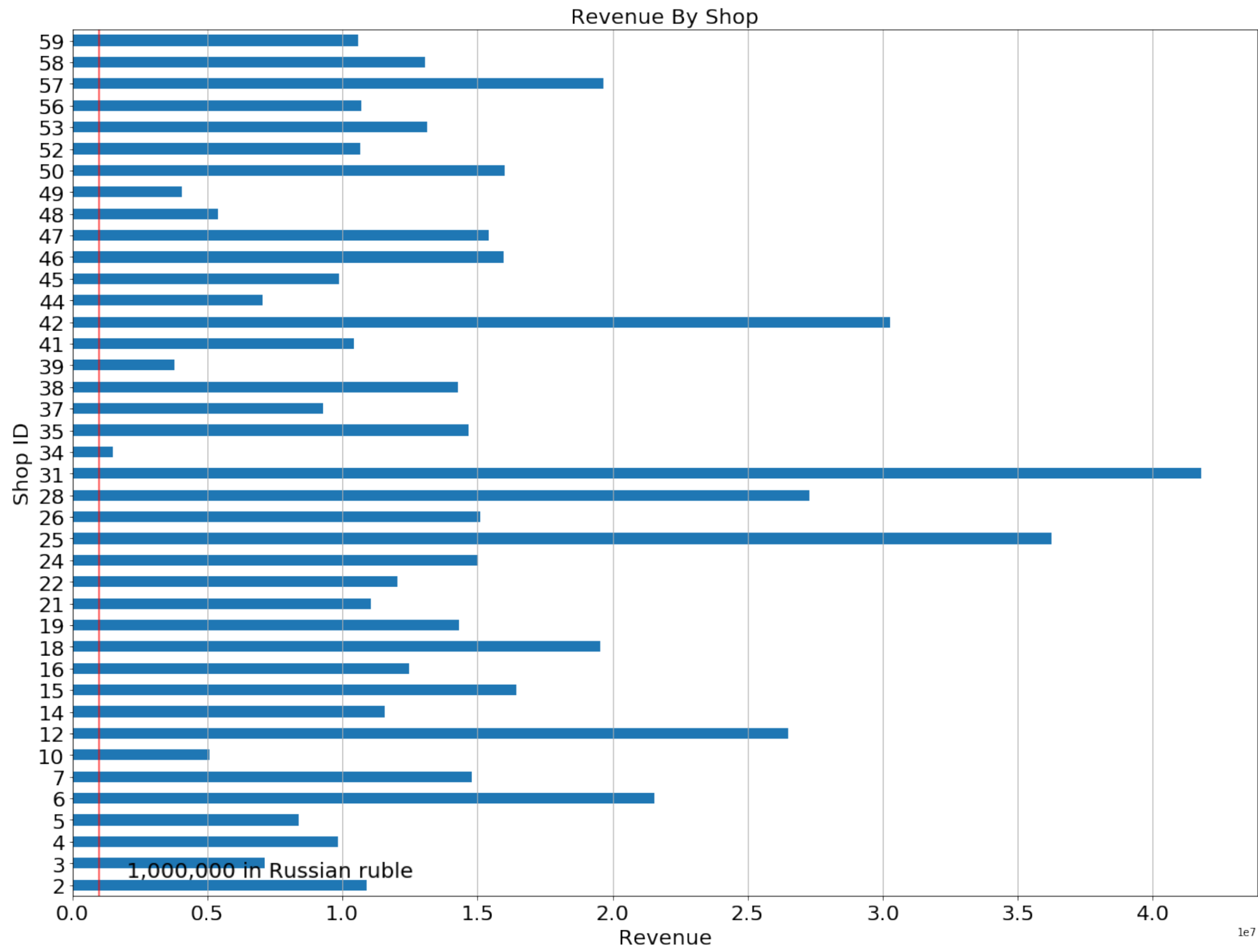
# EDA - CATEGORY

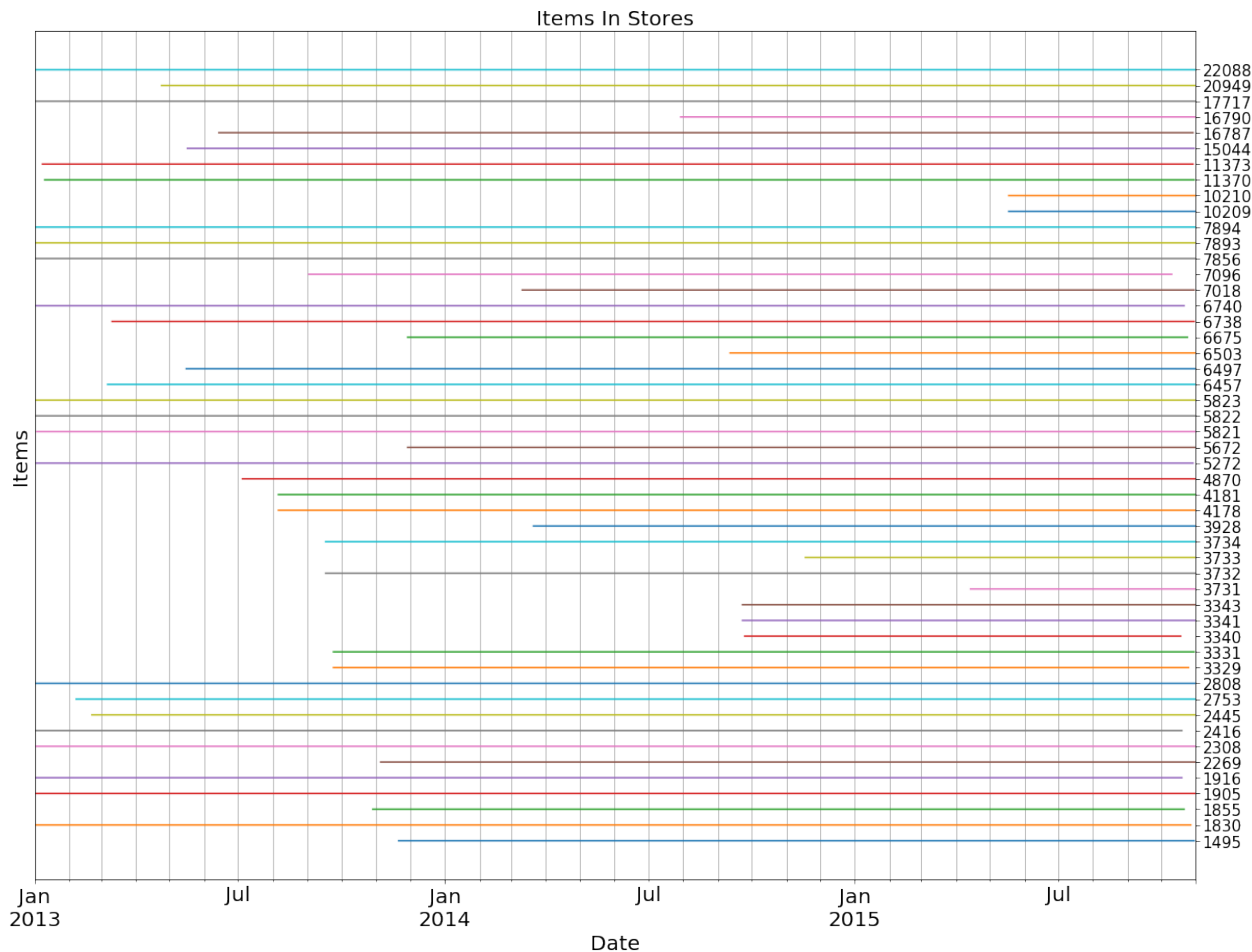


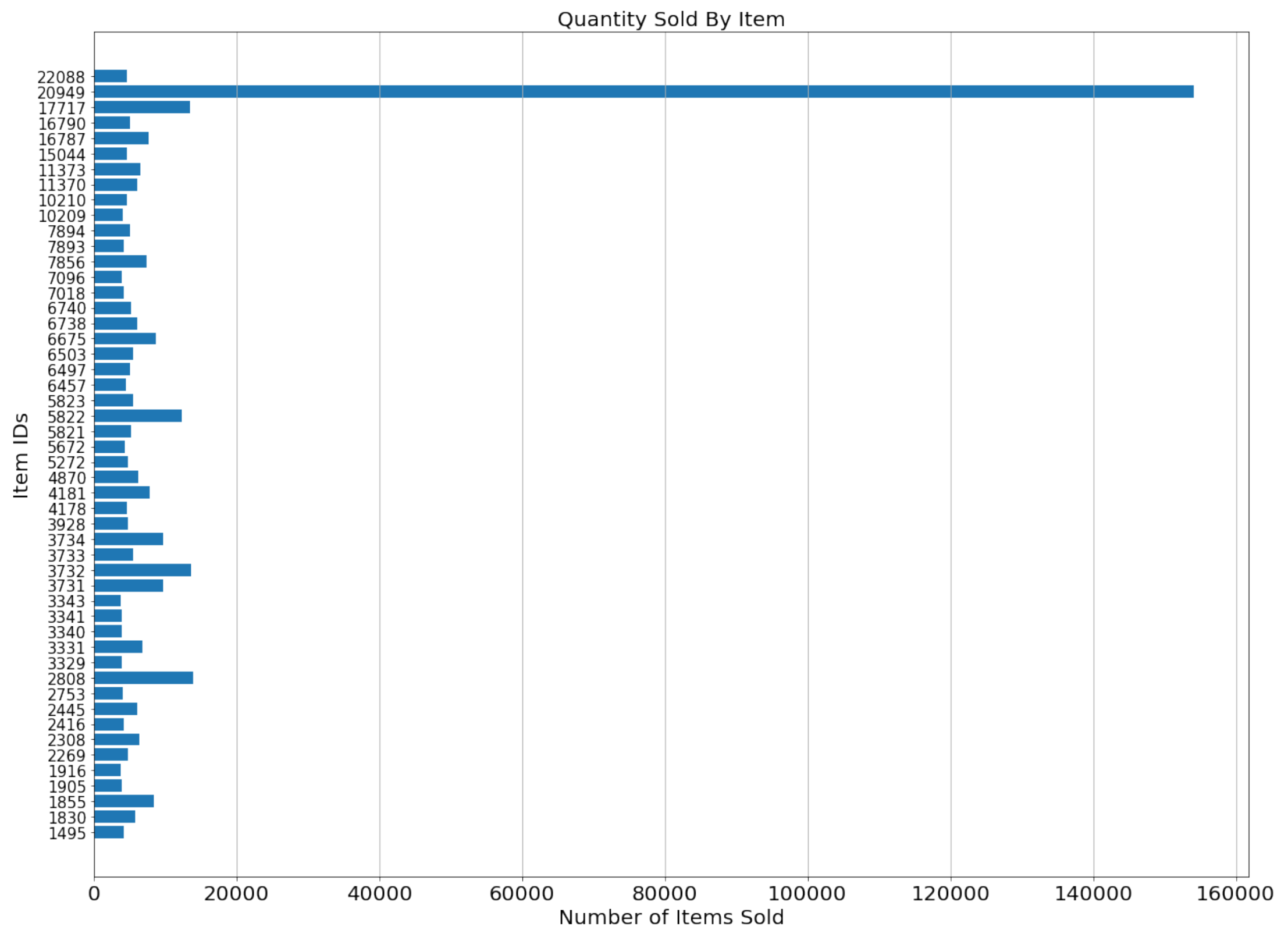




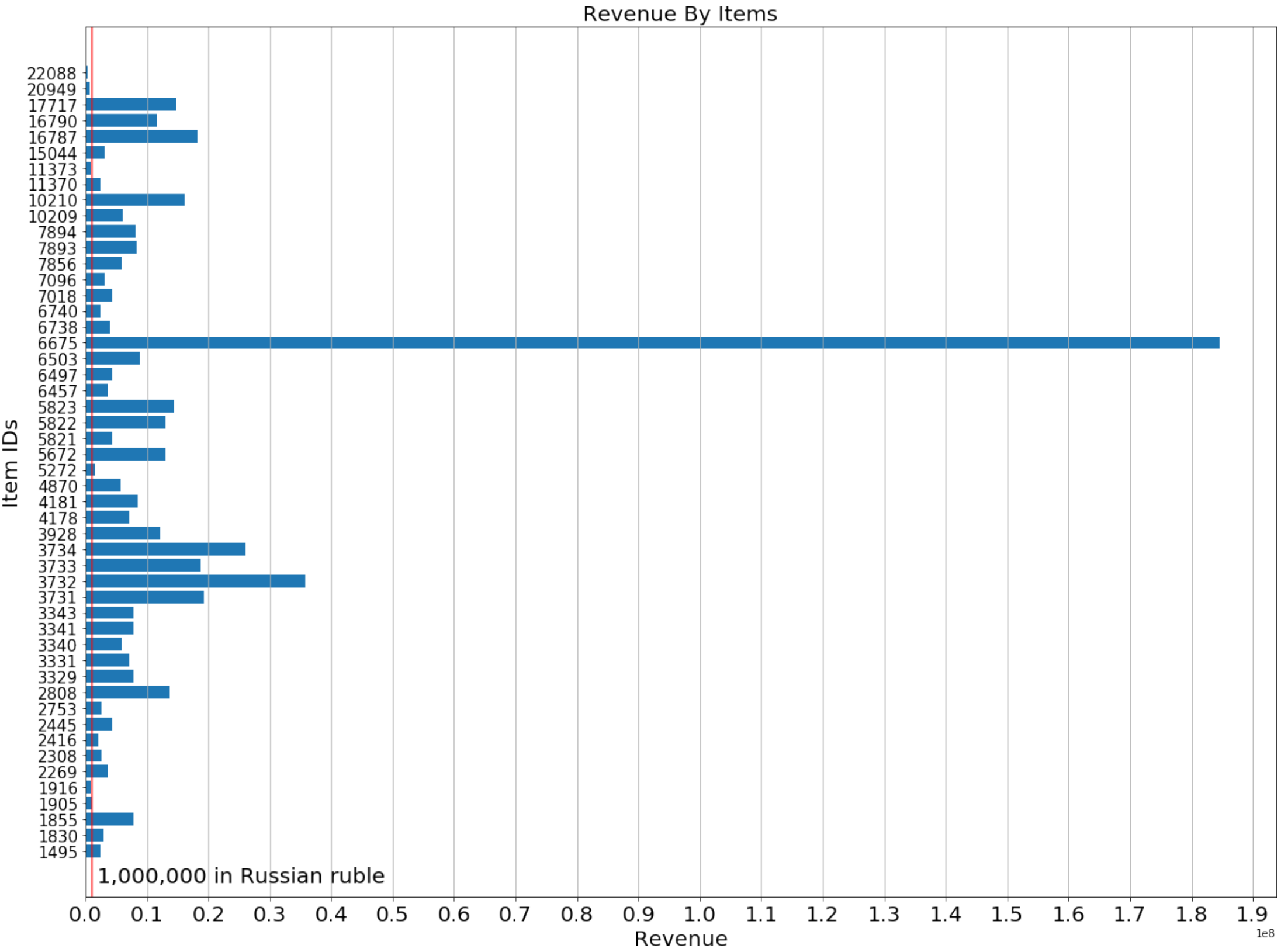




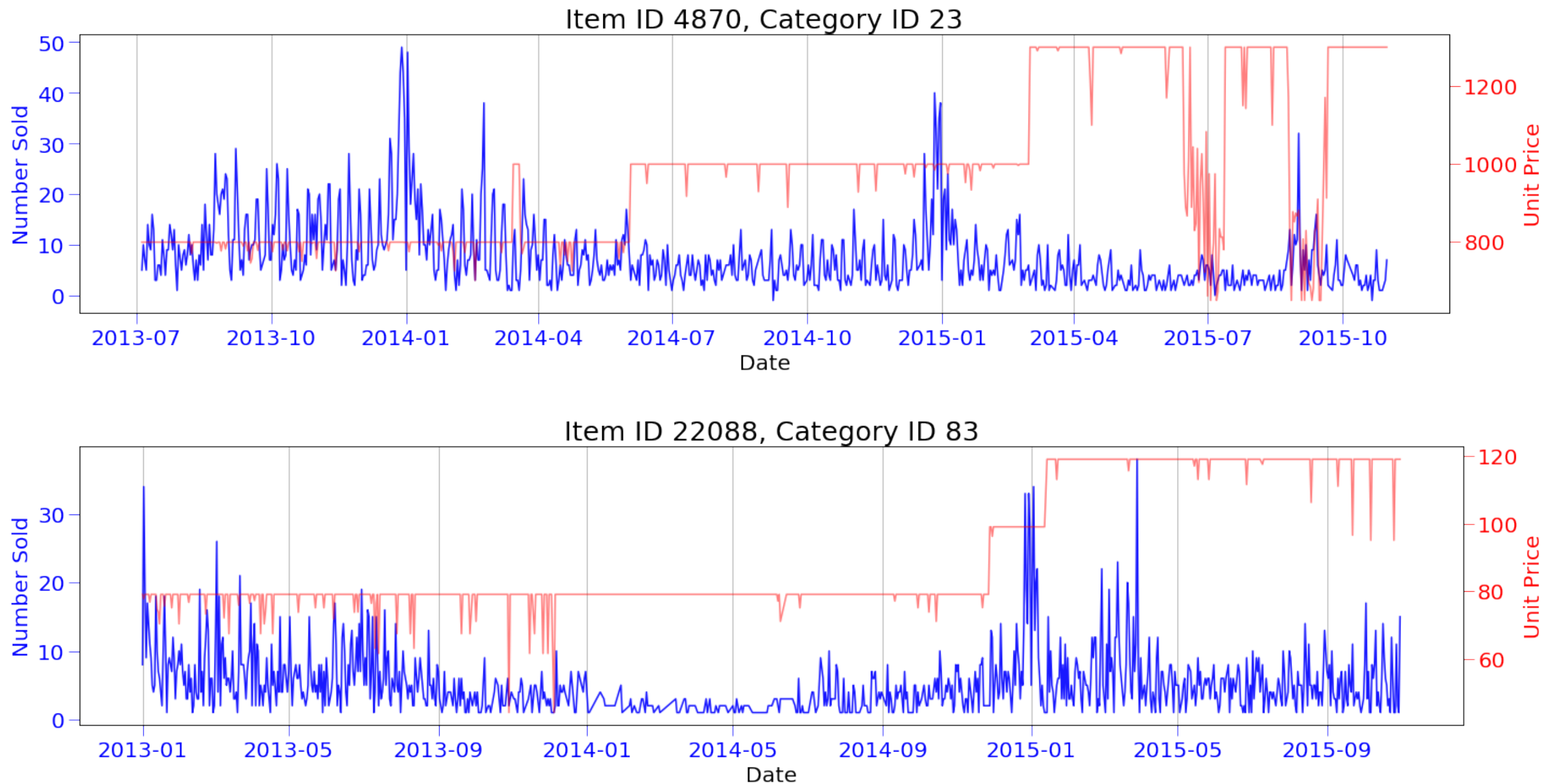




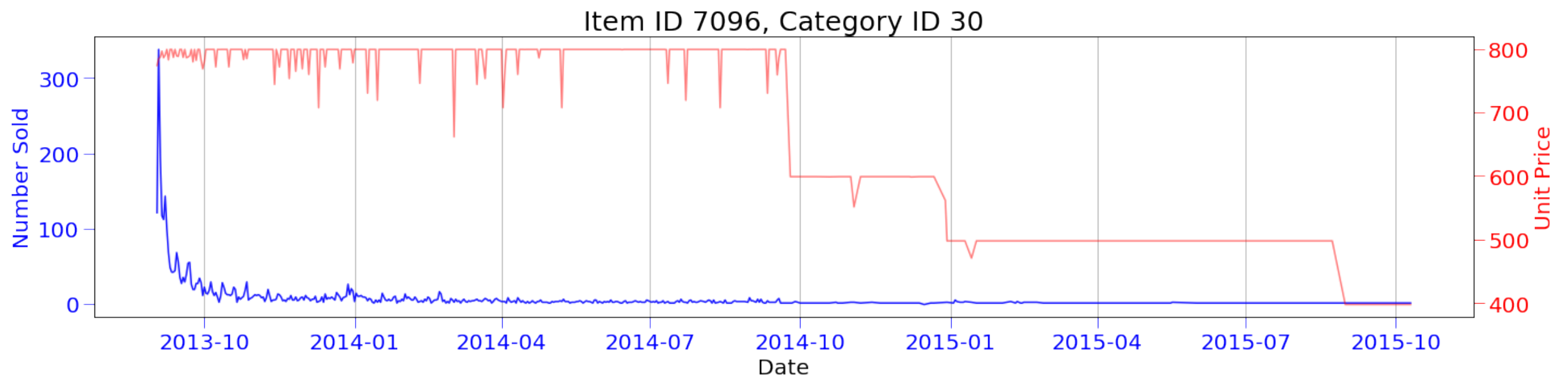
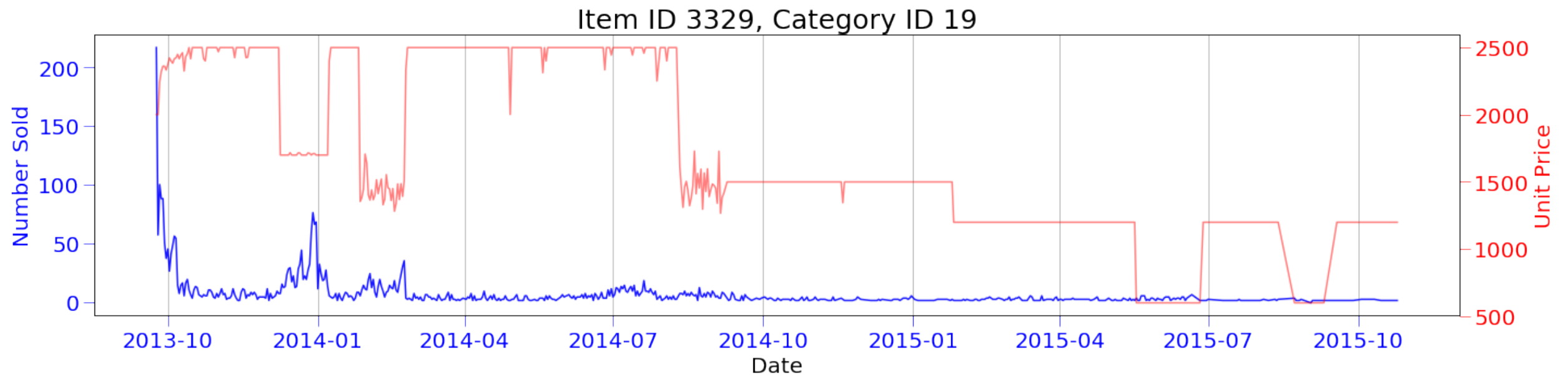




# PRICE OF ITEMS INCREASE OVER TIME

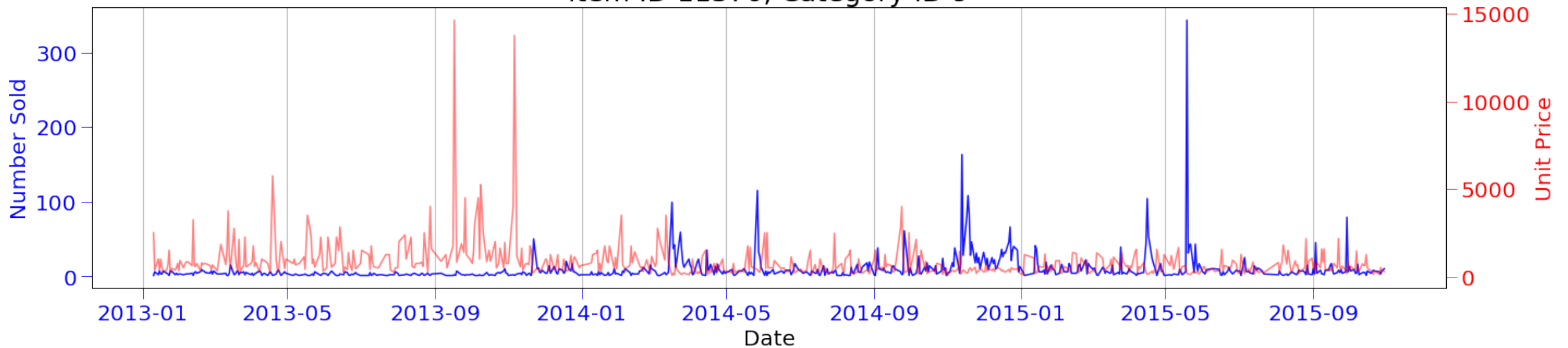


# PRICE OF ITEMS DECREASE OVER TIME

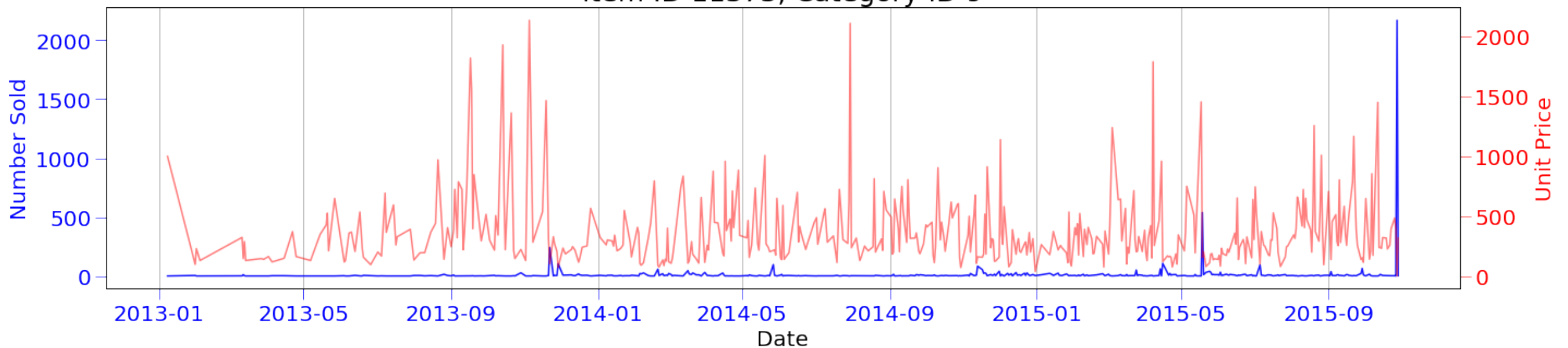


# PRICE OF ITEMS ARE NOT FIXED

Item ID 11370, Category ID 9



Item ID 11373, Category ID 9



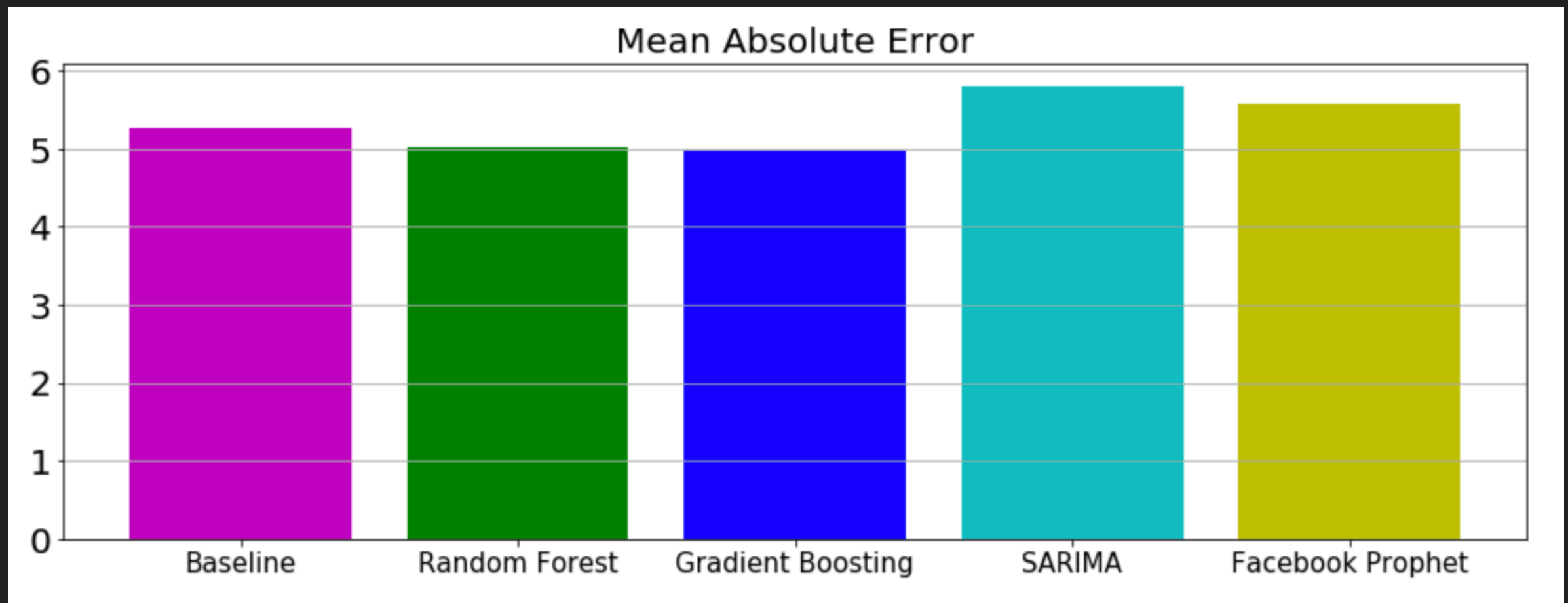
# PREDICTIVE MODELING

# APPROACH

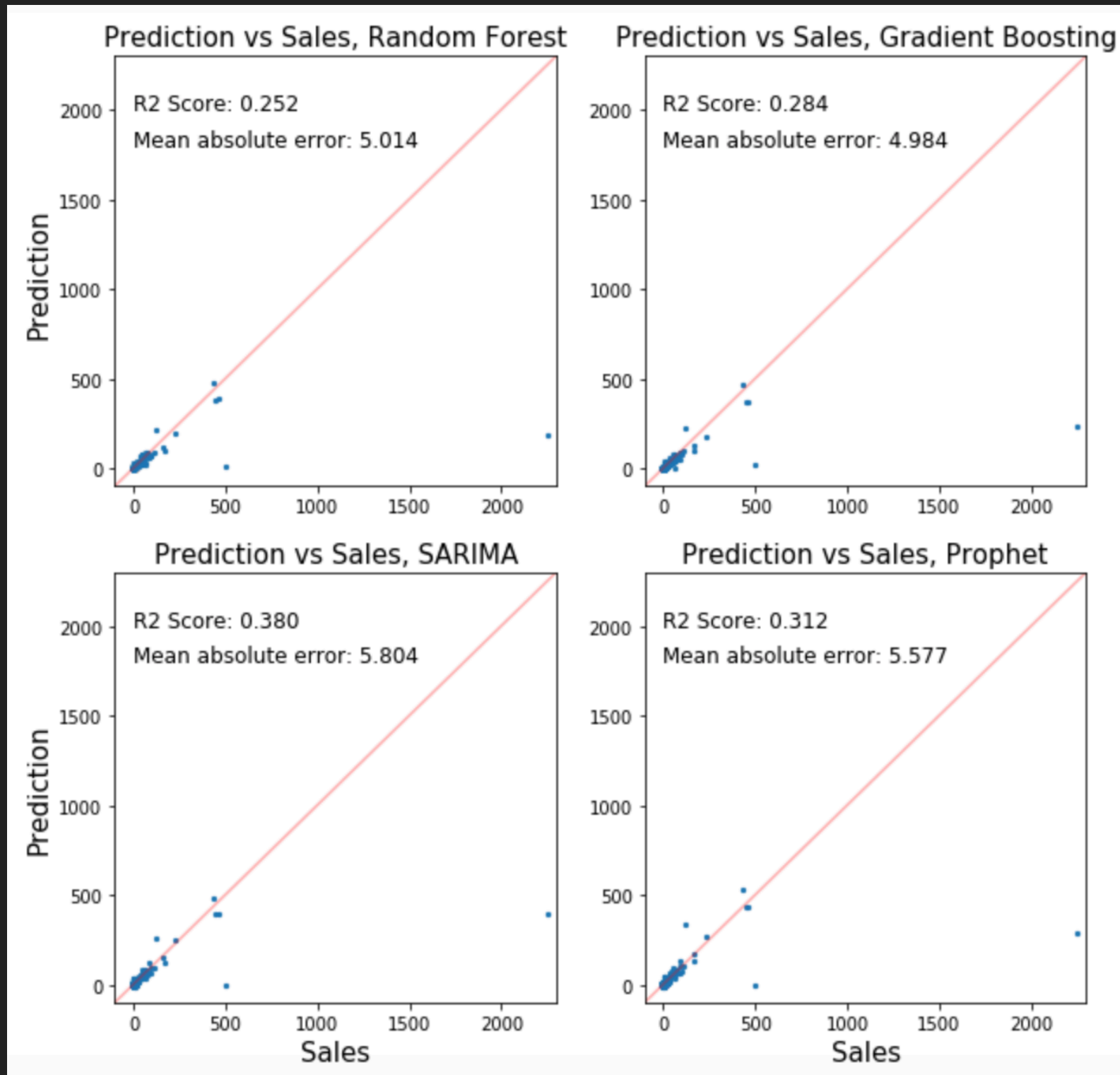
- ▶ For methods that can only forecast one time-series at a time, the project will adopt middle out approach using weights of shops to determine proportional sales in each shop
- ▶ Forecasting over 2000 items in all shops can be computationally unfeasible with some models for this project, therefore the data will be narrowed down to 50 top selling items
- ▶ Use baseline model to measure model performance

# MODEL COMPARISON

- ▶ Gradient boosting regressor yields the lowest mean absolute error at 4.984

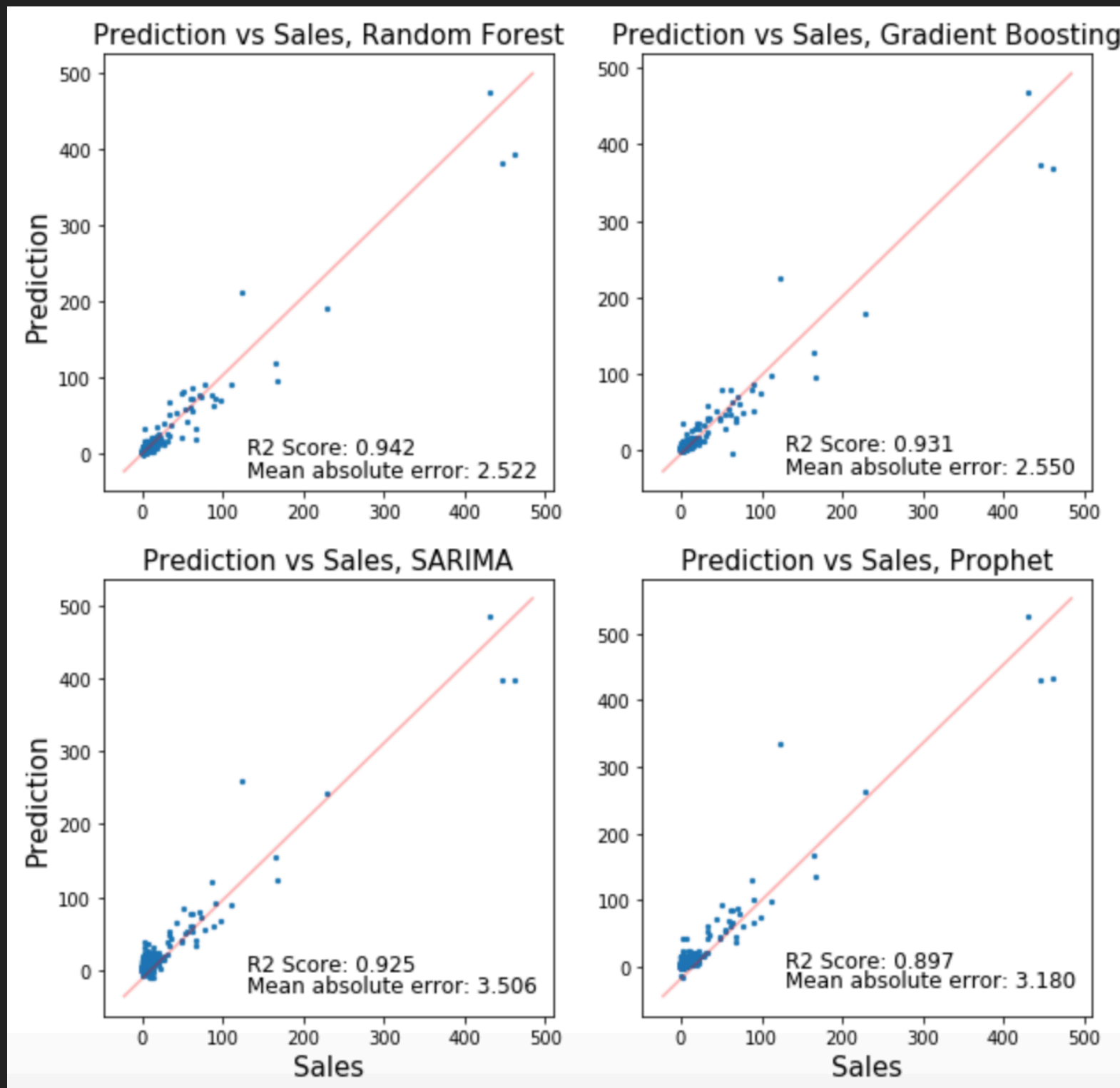


## MODEL COMPARISON





## MODEL COMPARISON AFTER REMOVING ABNORMAL SALES DATA



**CONCLUSION**

## CONCLUSION

### MODEL CONCLUSION

- ▶ Gradient boosting reduced error by 0.289, which is 5.5% error reduction.
- ▶ By total amount of items sold in Oct 2015, it is equivalent to 300 units of monthly overstocking and under stocking combined.



# ASSUMPTION & LIMITATION

- ▶ Sales are not independent
- ▶ Negative units sold are assumed as returned item and is built into models

## RECOMMENDATION & IMPROVEMENT

- ▶ Companies can use the model by feeding historical sales data to reduce both overstocking and under-stocking beyond using historical sales data as par stock
- ▶ Include record of special events or national holidays that boost sales, similar to Black Friday or Cyber Monday
- ▶ Use entire sales data from the source