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.

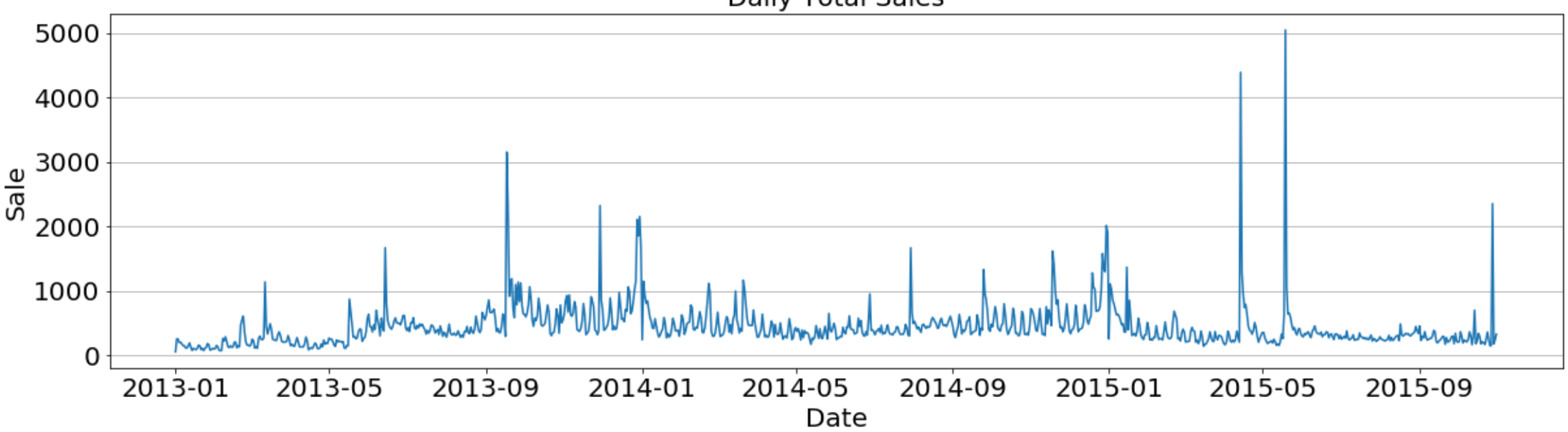
DATA ACQUISITION

- 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.



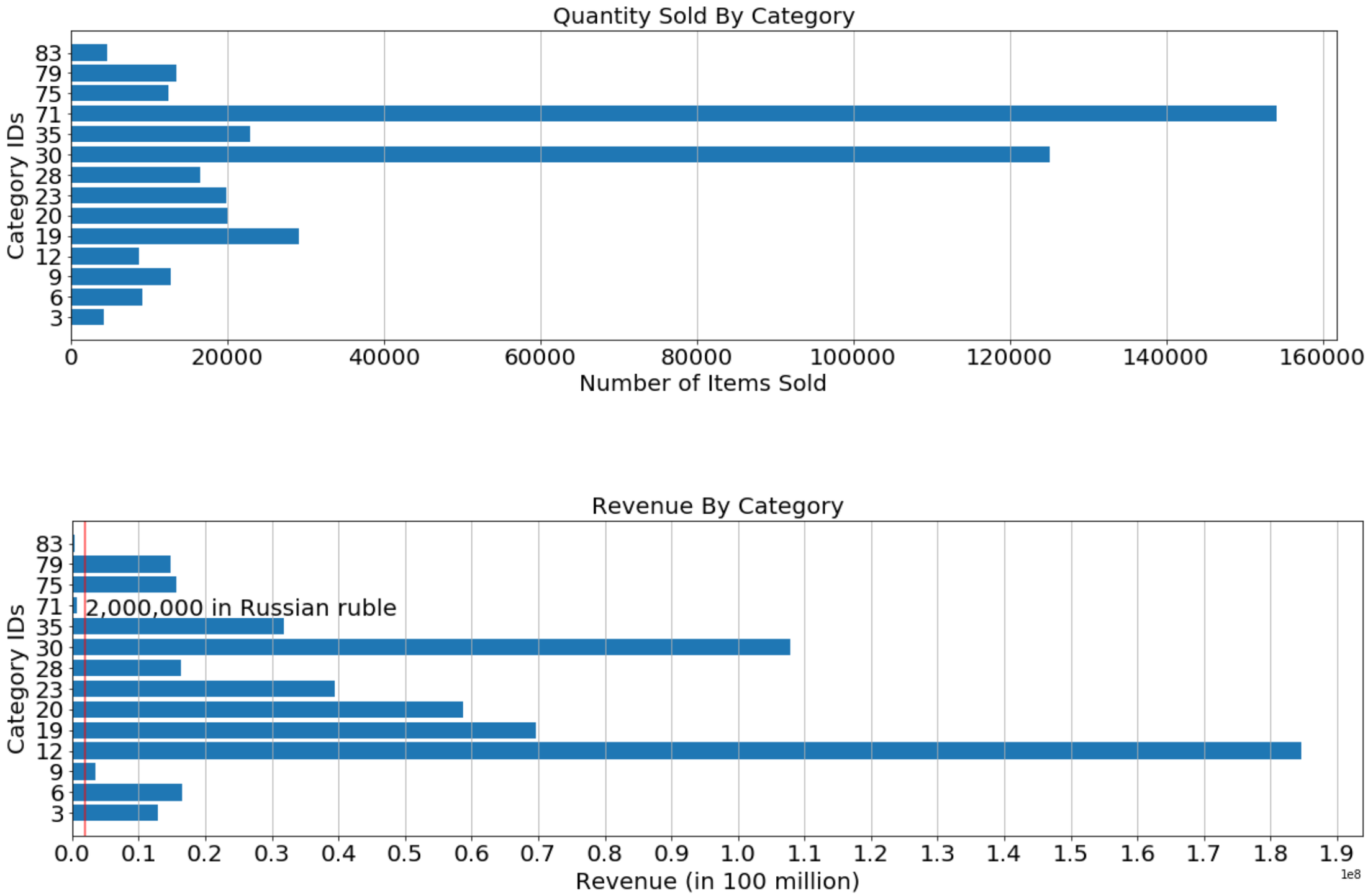
NEGATIVE SALES

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

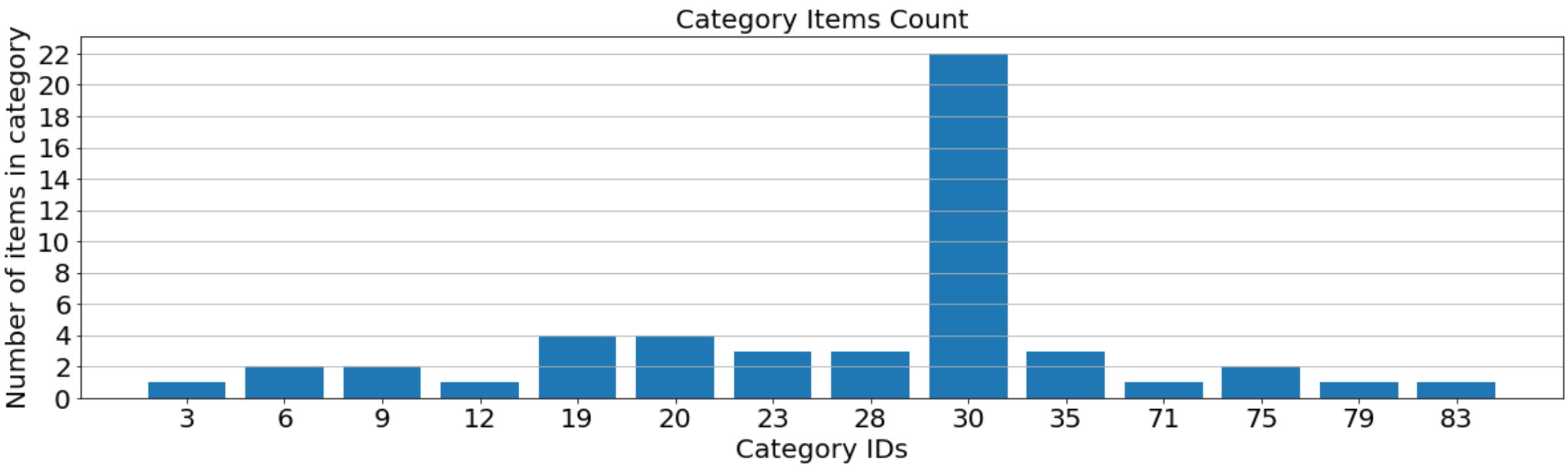
Number of items have negative sales in a given month: 25

	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

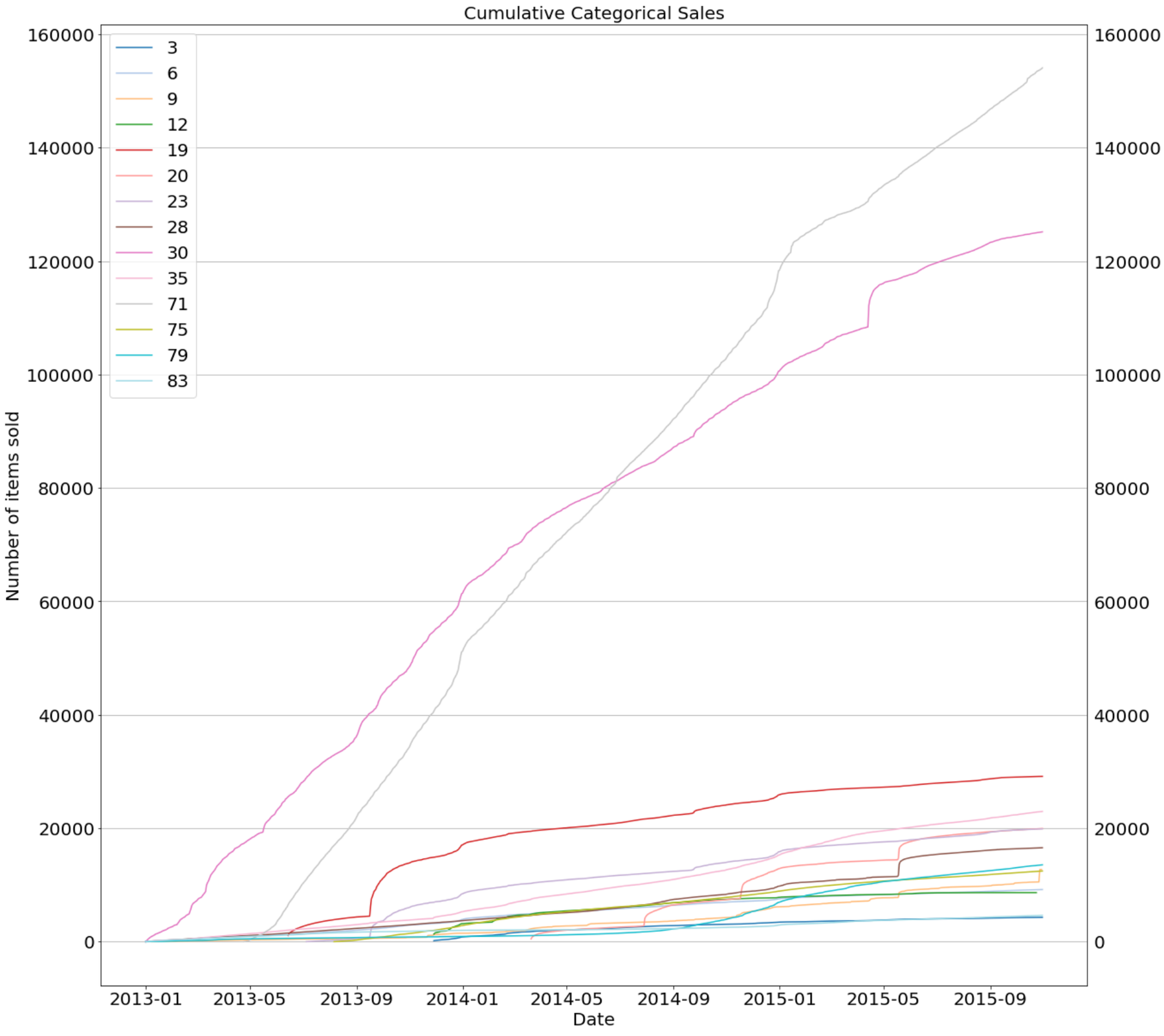


EDA - CATEGORY

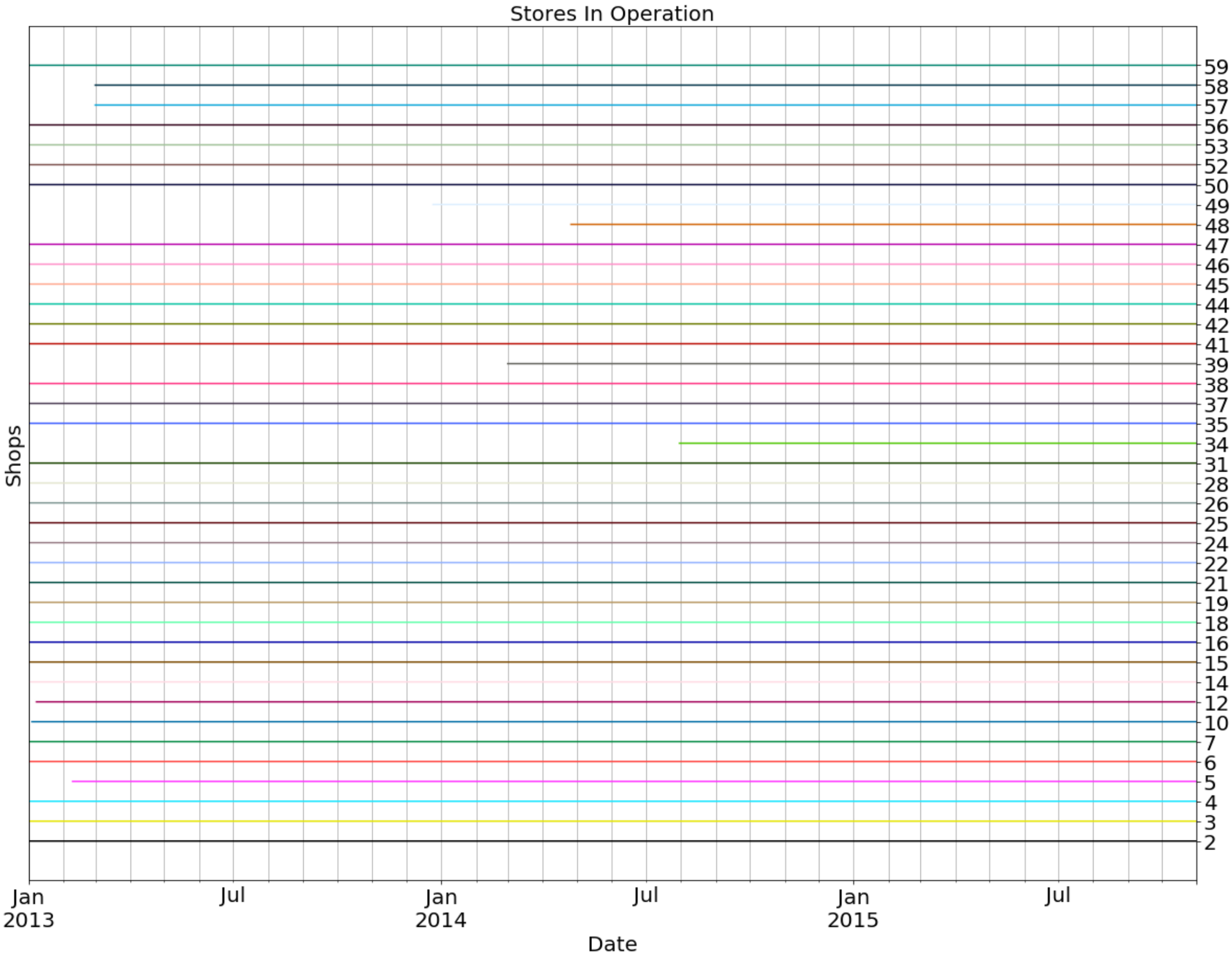


- Category 71 has only one item ID and has the most units sold than other categories
- But the revenue from category 71 is extremely small

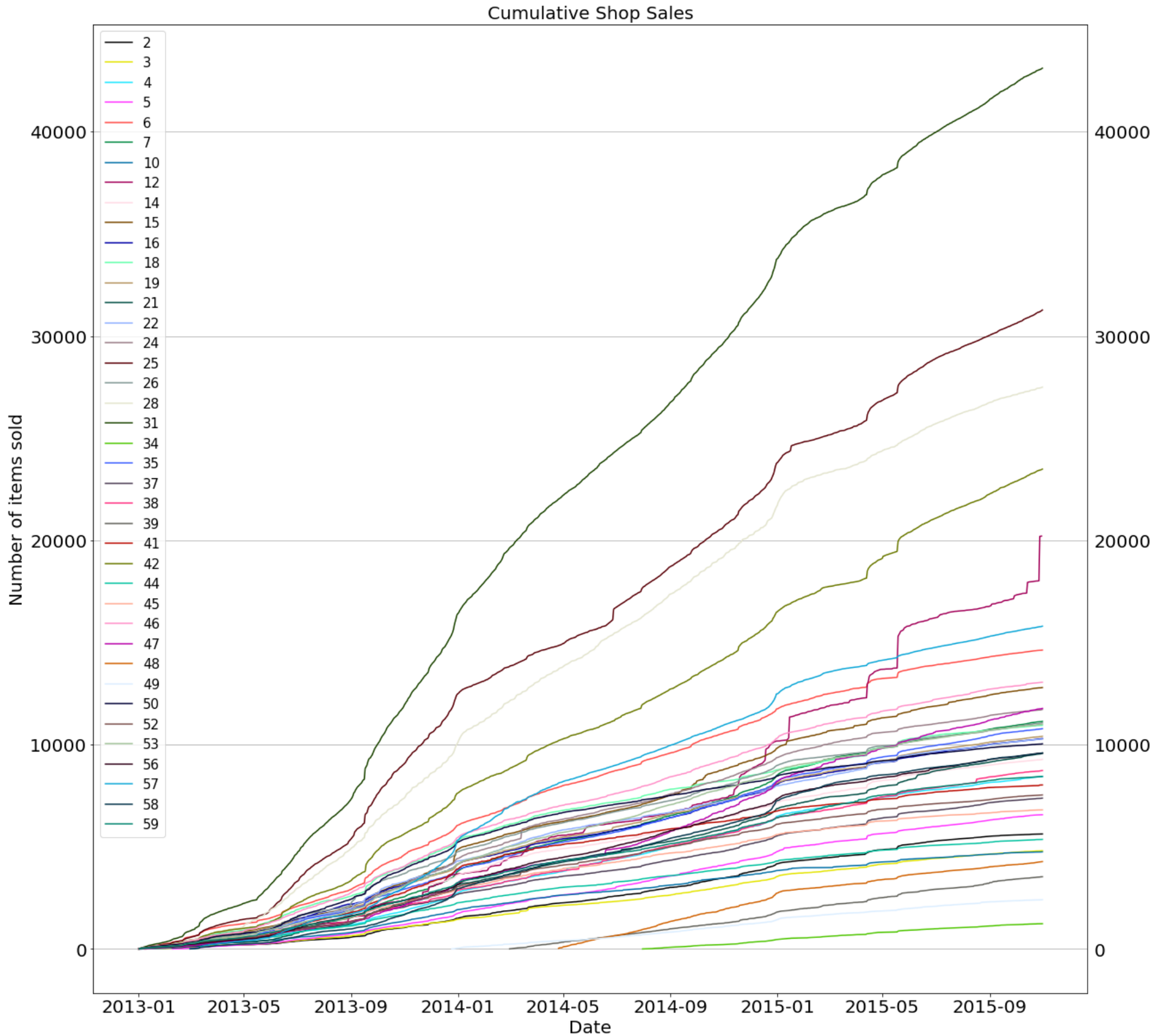
EDA - CATEGORY



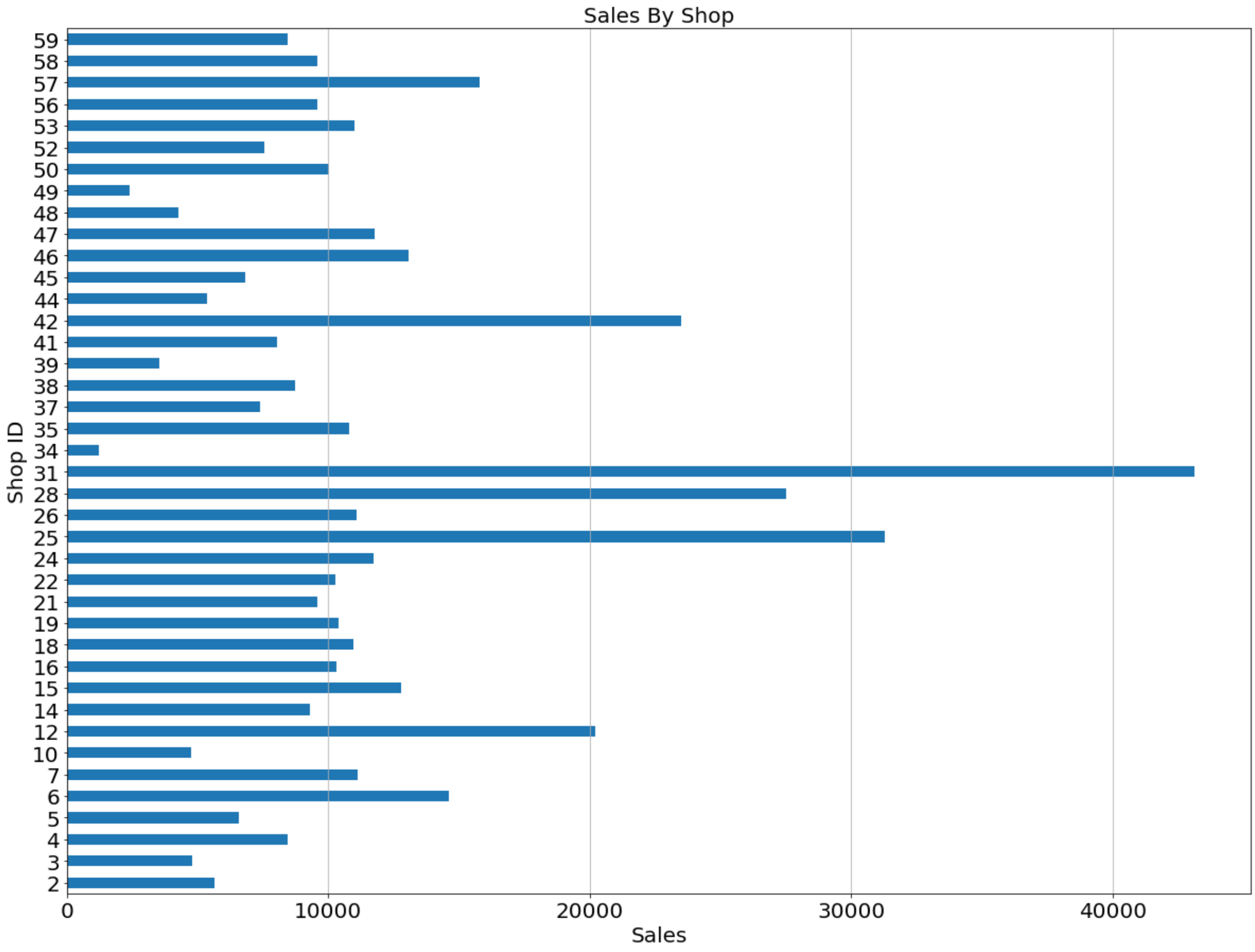
EDA - STORE


Stores In Operation
Shops
59
58
57
56
53
52
50
49
48
47
46
45
44
42
41
39
38
37
35
34
31
28
26
25
24
22
21
19
18
16
15
14
12
10
7
6
5
4
3
2
Jan
2013
Jul
Jan
2014
Jul
Jan
2015
Jul
Date

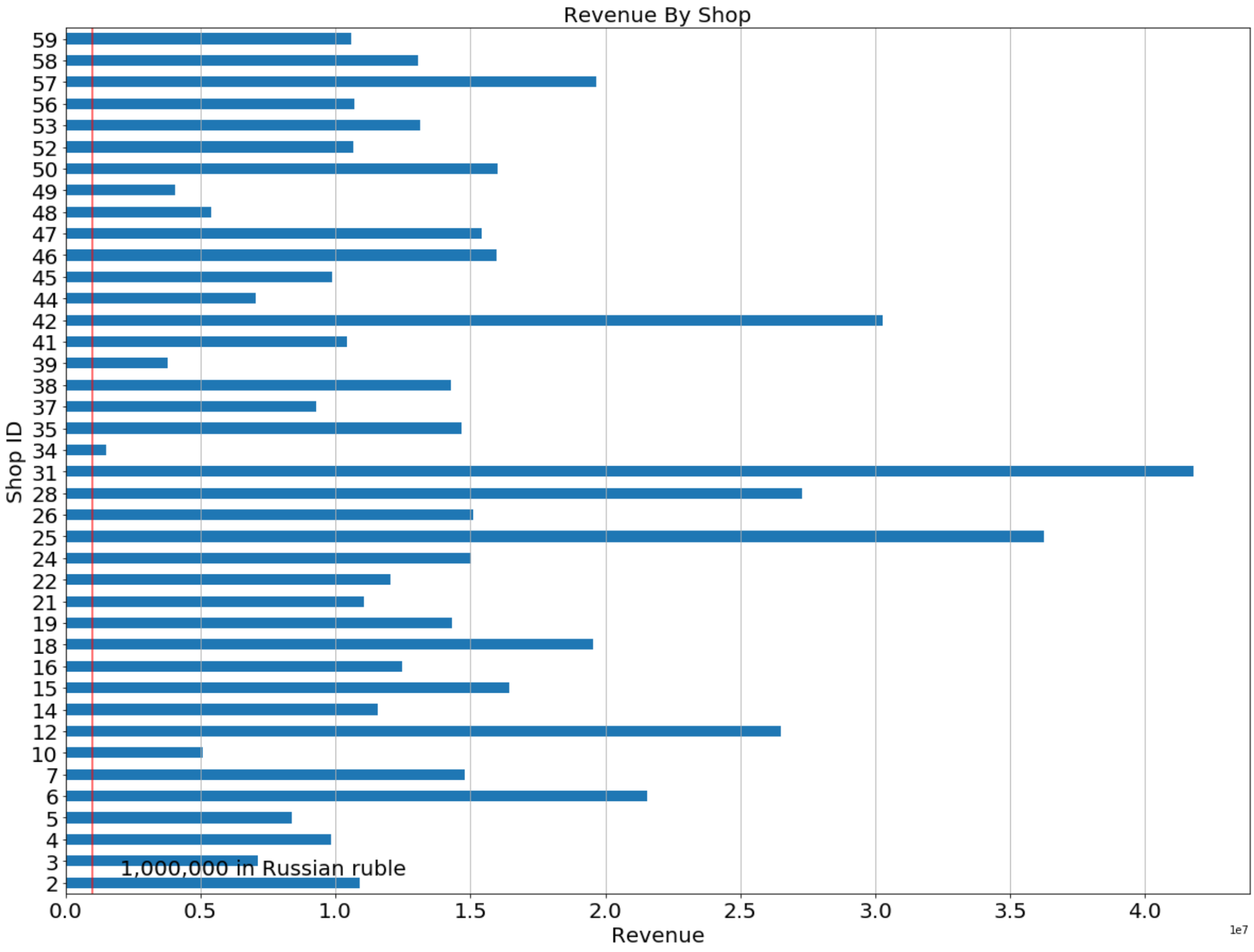
EDA - STORE


Cumulative Shop Sales
Number of items sold
Date
0
10000
20000
30000
40000
2013-01
2013-05
2013-09
2014-01
2014-05
2014-09
2015-01
2015-05
2015-09
2
3
4
5
6
7
10
12
14
15
16
18
19
21
22
24
25
26
28
31
34
35
37
38
39
41
42
44
45
46
47
48
49
50
52
53
56
57
58
59

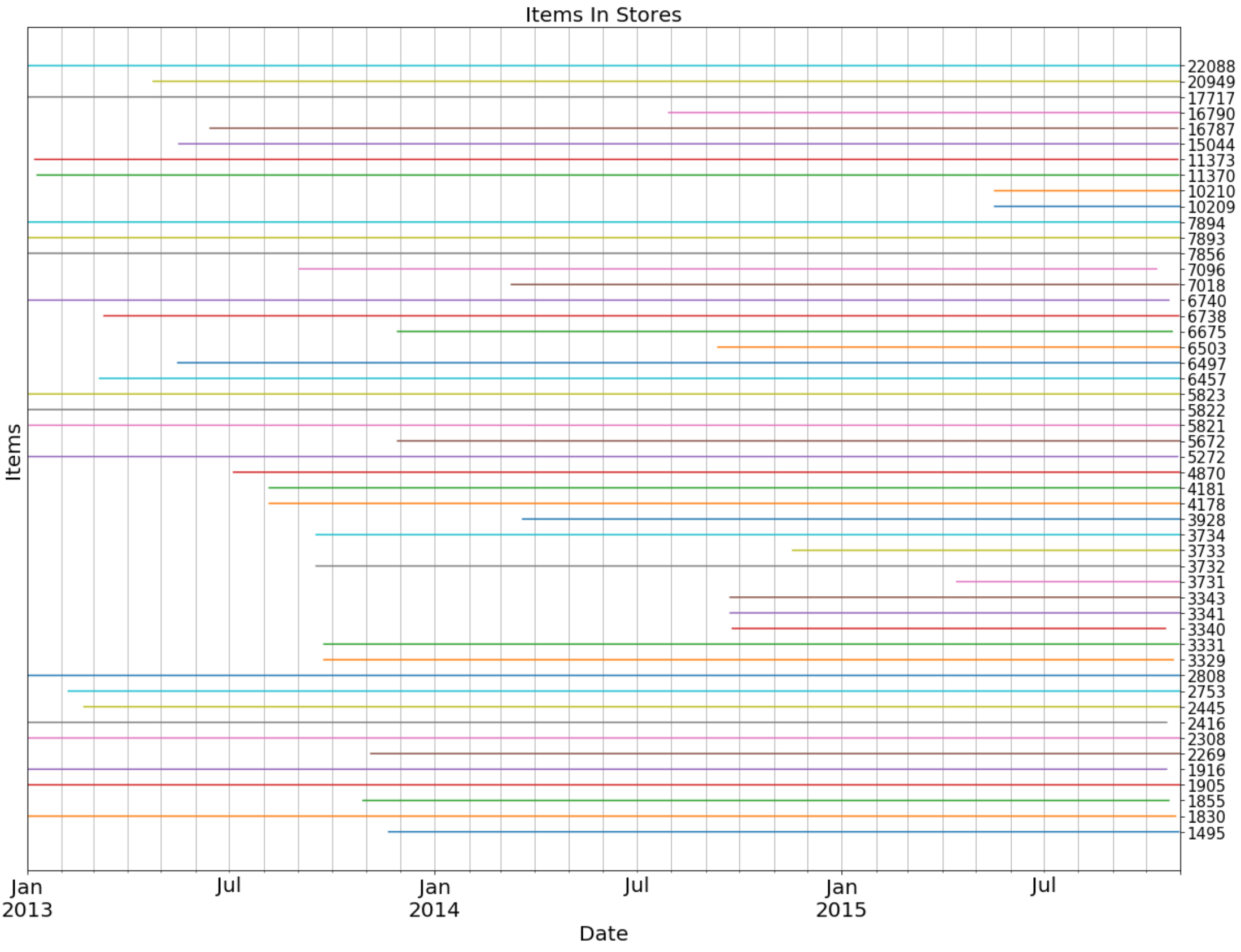
EDA - STORE


Sales By Shop
Shop ID
59
58
57
56
53
52
50
49
48
47
46
45
44
42
41
39
38
37
35
34
31
28
26
25
24
22
21
19
18
16
15
14
12
10
7
6
5
4
3
2
0
10000
20000
30000
40000
Sales

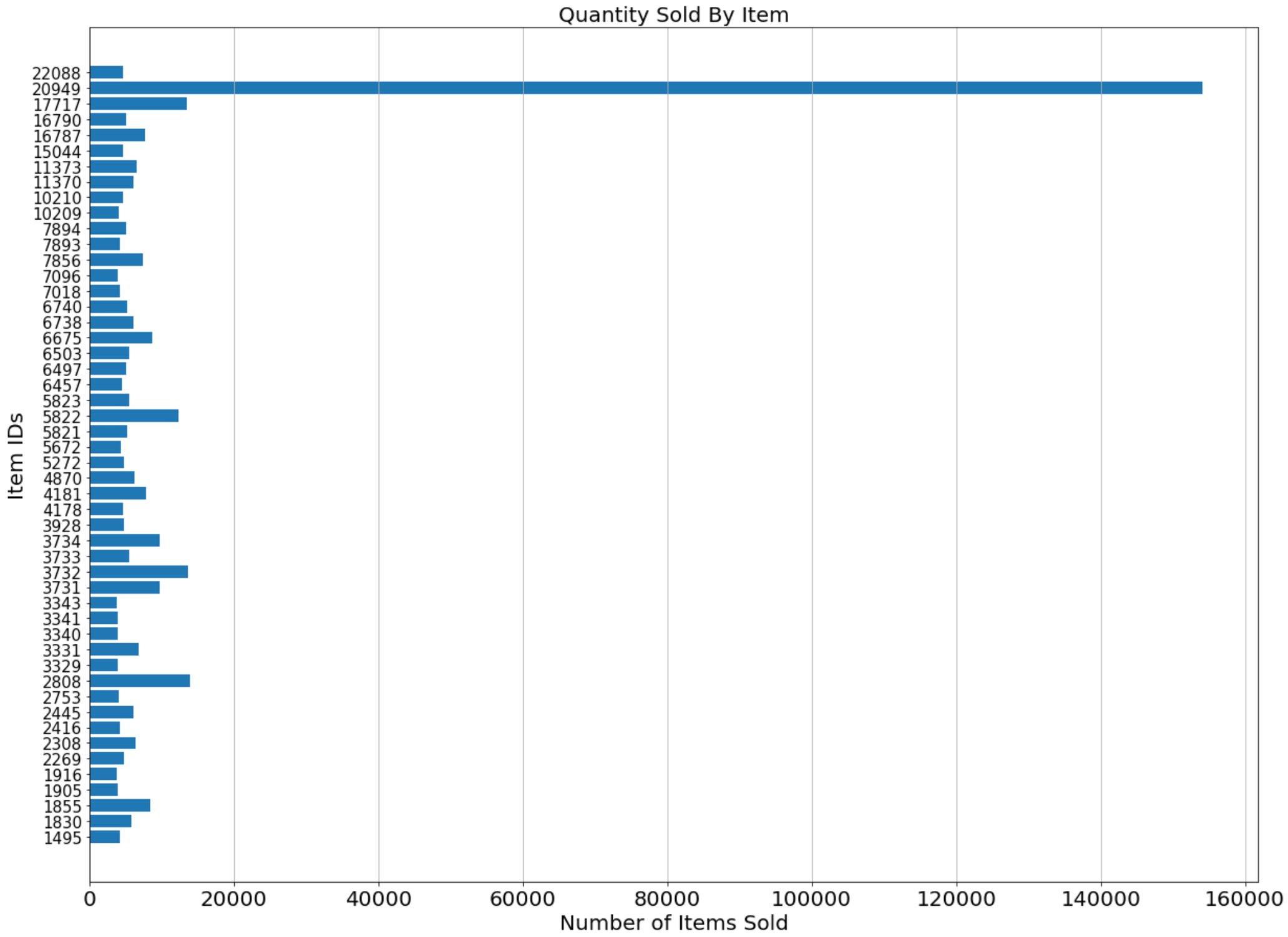
EDA - STORE


Revenue By Shop
Shop ID
59
58
57
56
53
52
50
49
48
47
46
45
44
42
41
39
38
37
35
34
31
28
26
25
24
22
21
19
18
16
15
14
12
10
7
6
5
4
3
2
1,000,000 in Russian ruble
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Revenue
1e7

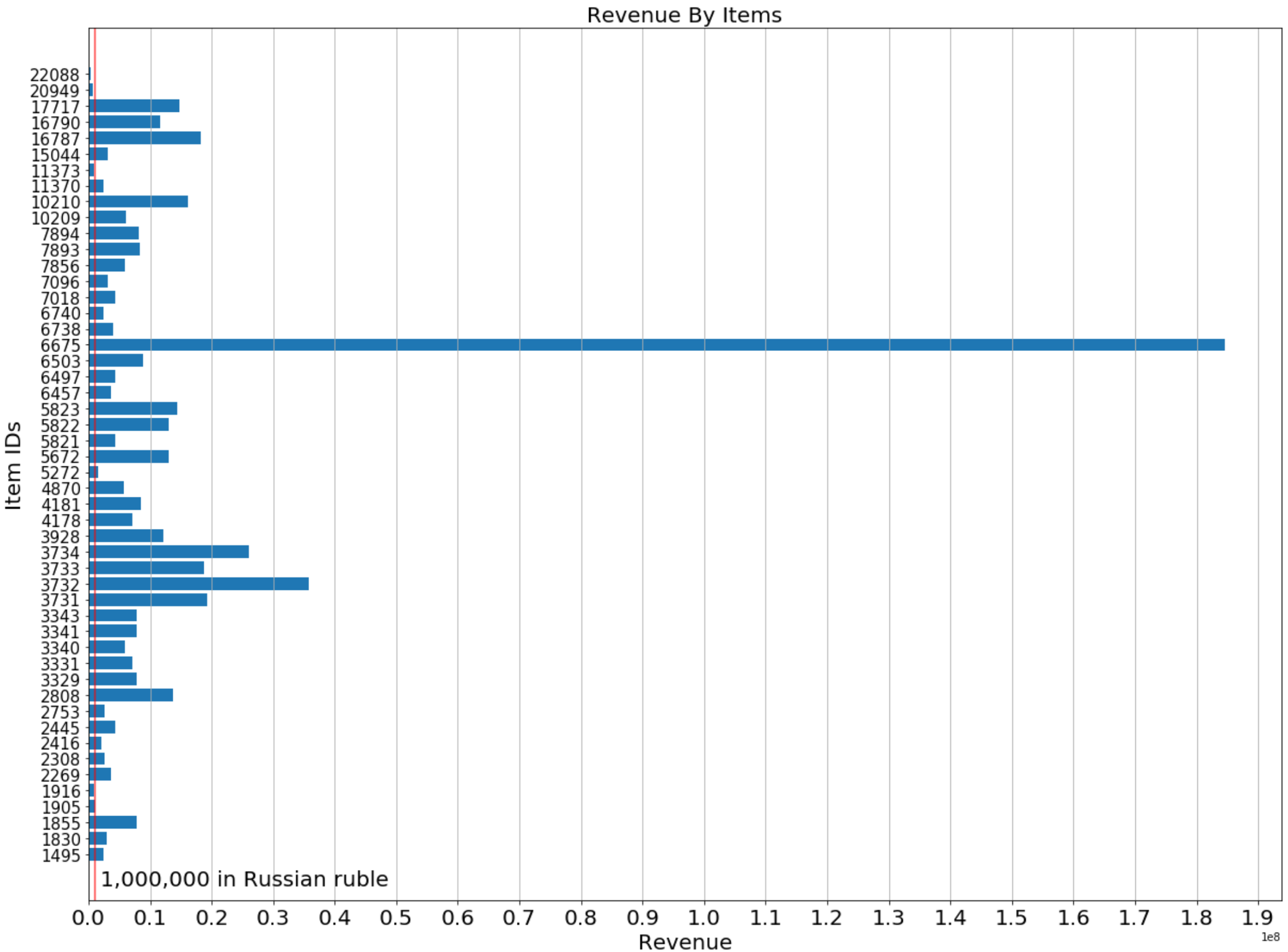
EDA - ITEM


Items In Stores
Items
Date
Jan
2013
Jul
Jan
2014
Jul
Jan
2015
Jul
22088
20949
17717
16790
16787
15044
11373
11370
10210
10209
7894
7893
7856
7096
7018
6740
6738
6675
6503
6497
6457
5823
5822
5821
5672
5272
4870
4181
4178
3928
3734
3733
3732
3731
3343
3341
3340
3331
3329
2808
2753
2445
2416
2308
2269
1916
1905
1855
1830
1495

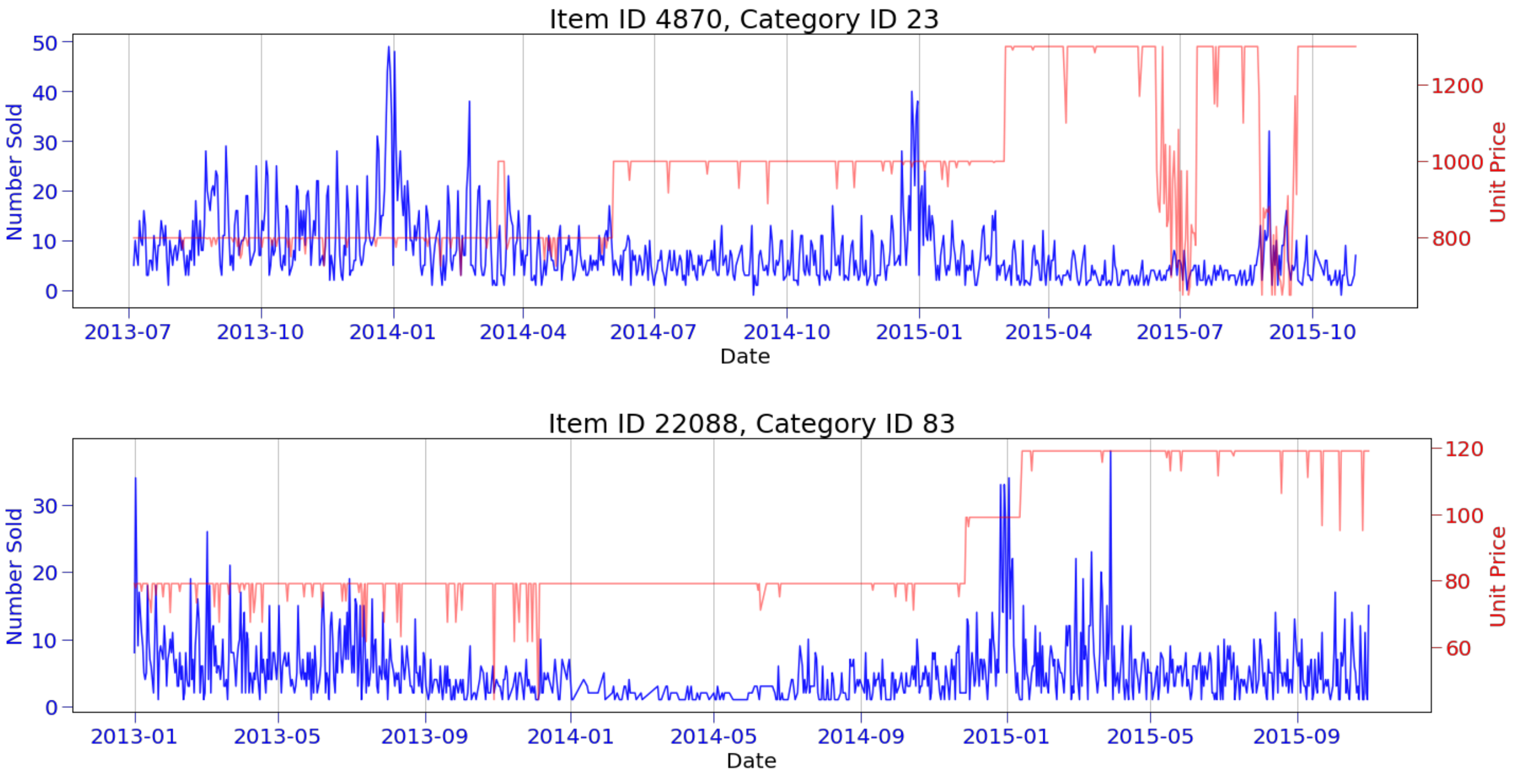
EDA - ITEM


Quantity Sold By Item
Item IDs
22088
20949
17717
16790
16787
15044
11373
11370
10210
10209
7894
7893
7856
7096
7018
6740
6738
6675
6503
6497
6457
5823
5822
5821
5672
5272
4870
4181
4178
3928
3734
3733
3732
3731
3343
3341
3340
3331
3329
2808
2753
2445
2416
2308
2269
1916
1905
1855
1830
1495
0
20000
40000
60000
80000
100000
120000
140000
160000
Number of Items Sold

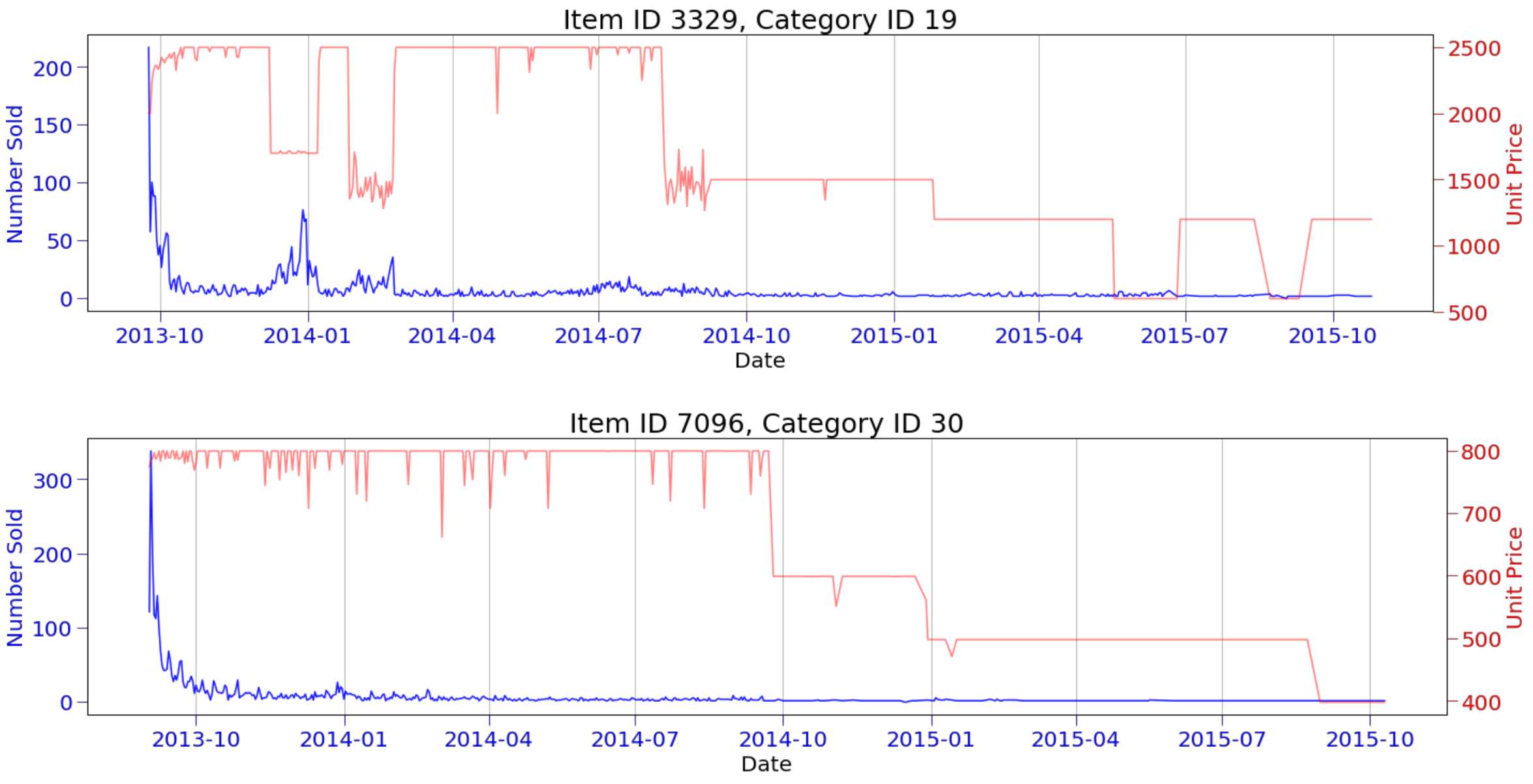
EDA - ITEM


Revenue By Items
Item IDs
22088
20949
17717
16790
16787
15044
11373
11370
10210
10209
7894
7893
7856
7096
7018
6740
6738
6675
6503
6497
6457
5823
5822
5821
5672
5272
4870
4181
4178
3928
3734
3733
3732
3731
3343
3341
3340
3331
3329
2808
2753
2445
2416
2308
2269
1916
1905
1855
1830
1495
1,000,000 in Russian ruble
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8
1.9
1e8
Revenue

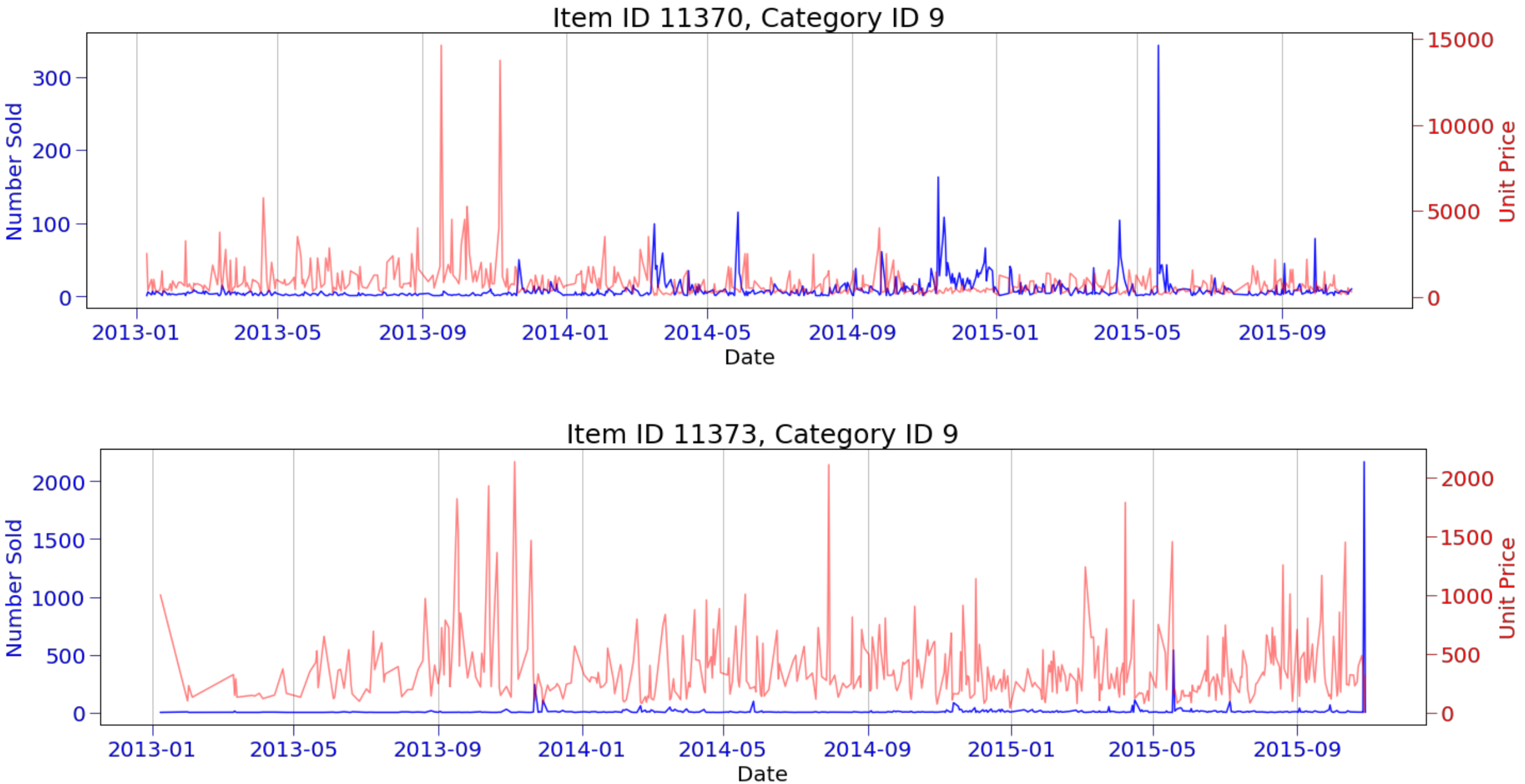
PRICE OF ITEMS INCREASE OVER TIME



PRICE OF ITEMS DECREASE OVER TIME



PRICE OF ITEMS ARE NOT FIXED



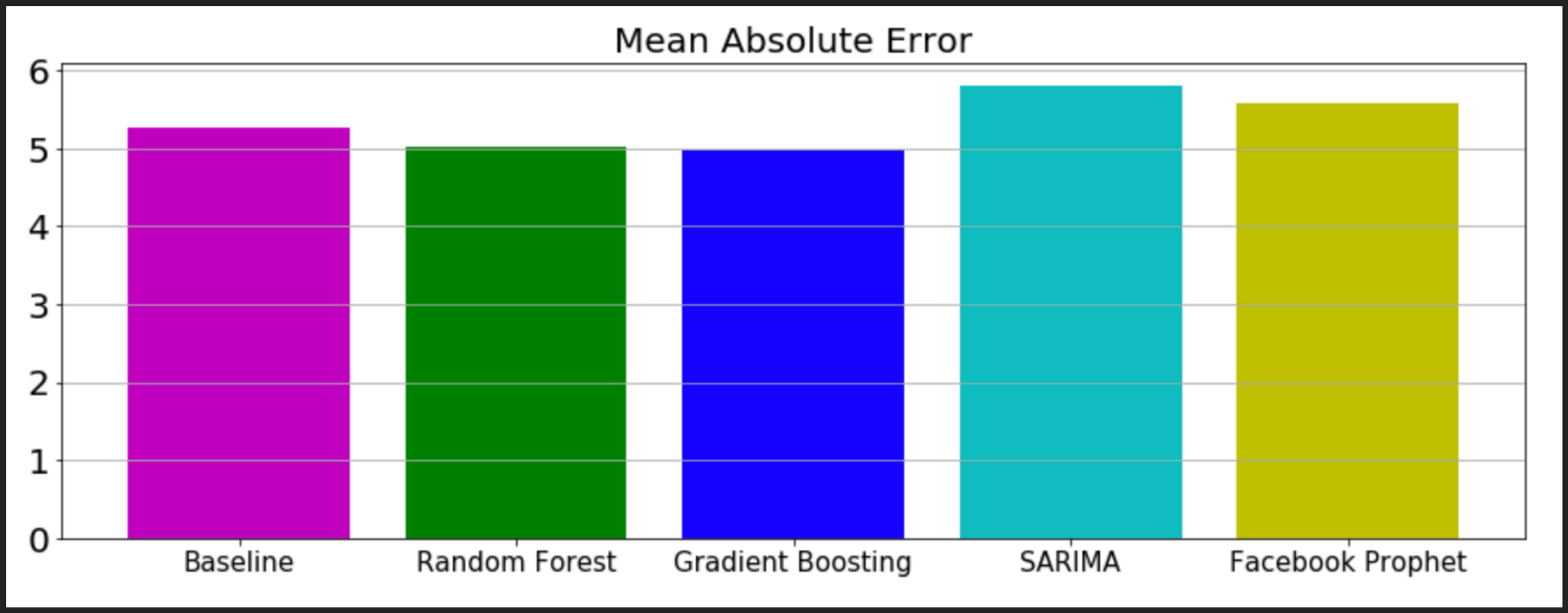
PREDICTIVE MODELING

APPROACH

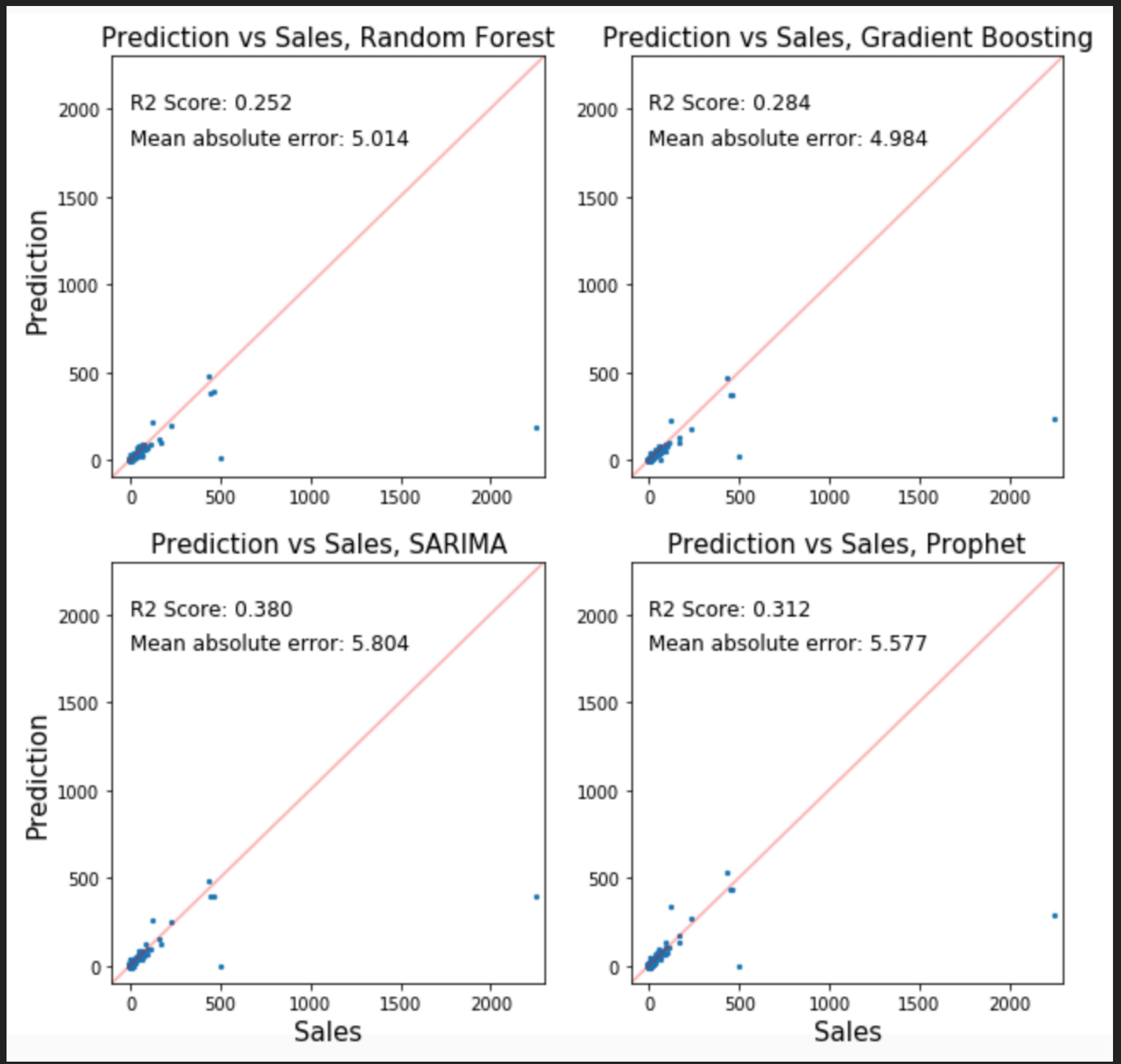
- For methods that can only forecast one time-series at a time, the project will adopt middle out approach using weights of shops to determine proportional sales in each shop
- Forecasting over 2000 items in all shops can be computationally unfeasible with some models for this project, therefore the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

MODEL COMPARISON

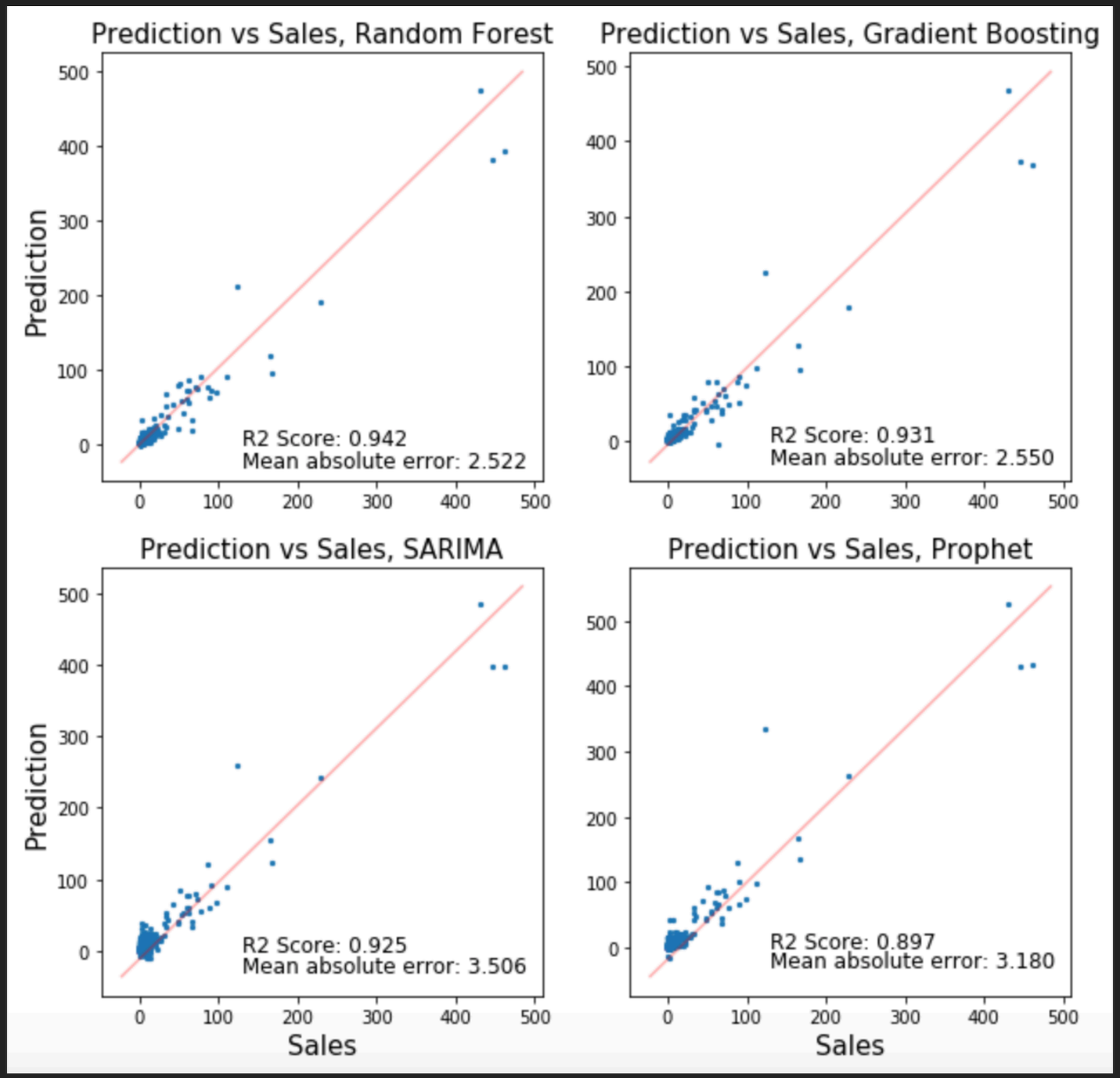
- Gradient boosting regressor yields the lowest mean absolute error at 4.984



MODEL COMPARISON



MODEL COMPARISON AFTER REMOVING ABNORMAL SALES DATA



CONCLUSION

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 FOR IMPROVEMENT

- Include record of special events or national holidays that boost sales, similar to Black Friday or Cyber Monday
- Include data of item price reduction and duration rather than extrapolate from sales data
- Use entire sales data from the source