ETM 58D Term Project – Nasibe Deniz & Merve Esin

We were tasked with forecasting the sales quantity of 8 different items on Trendyol.com during the period between 15th of June and 5th of July 2020. These 8 items were a face wash by La Roche Posay, a box of wet wipes by Sleepy, wireless ear buds by Xiaomi, a vacuum cleaner by Fakir, women's leggings by Trendyolmilla, an electric toothbrush by Oral-B, a bikini top by Trendyolmilla and men's jacket by Koton. We were provided with the past price and sales data, alongside various other data, for these diverse set of products. The dataset went as far back to 13/05/2019.

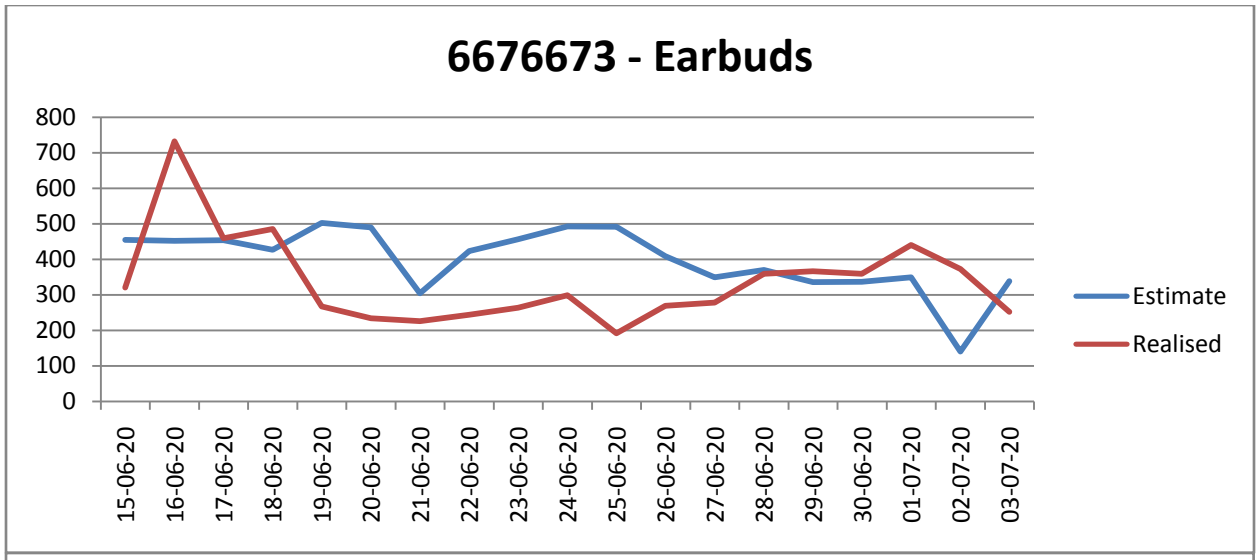
The dataset was complex as each item was sold at different prices during the past year. There were special Trendyol sales like the black Friday sale or just the sellers increasing/lowering their prices which drastically changes the number of items sold for particular periods. There were also the affects of corona virus on peoples purchasing behaviors. Taking everything into account would create a very complex model. When it comes to predictions we believe that the simplest model would work best so we decided to adopt a simple model.

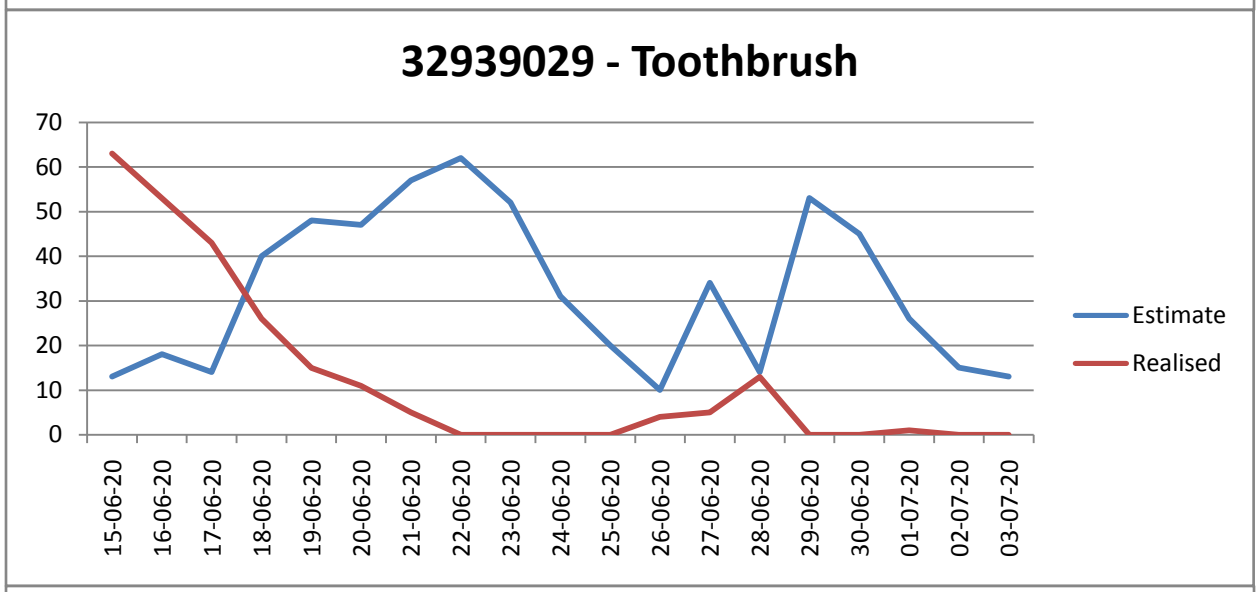
As stated above we adopted a simple approach. We decided to throw out any data before 1st of March 2020 so that we would end up with data from when the corona virus pandemic started and people were in lockdown. It is our belief that that the self isolation and curfews that came during this period significantly changed peoples shopping habits. Therefore we decided the use this period because for the most part people are still not 100% normalizing their behavior even though most of the restrictions are lifted.

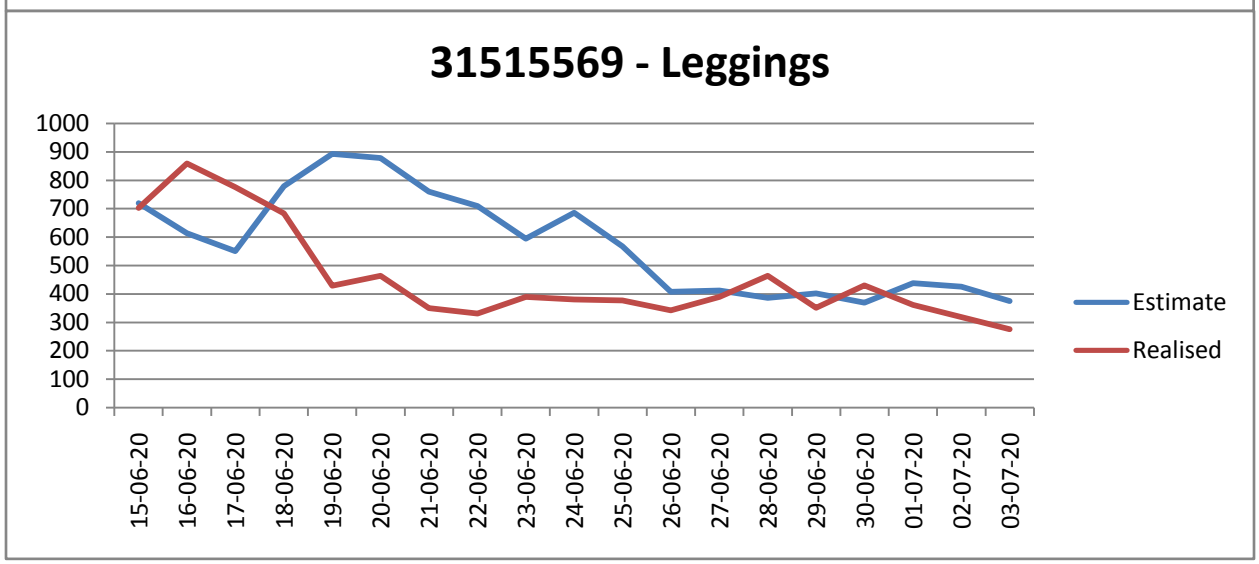
Another thing that we saw while going through the raw data was the fact that the number of products bought diminishes on Thursdays for most products. While we have no concrete evidence, we theorized that this maybe due to people not wanting their items delivered next week, since most items are delivered in 2 days. Also we observed that people tend to buy fewer items on weekends. Because of this we decided to create a model that would work using the data from the same day of the week. For example, our model predicts a Tuesday's sales numbers using the data from the past Tuesdays. This also had the side benefit of us running our code once a week to determine the predictions for the entire next week. Every day, we checked the stock of these items and their prices. Unless there was a significant change in stock and price, we went with this method.

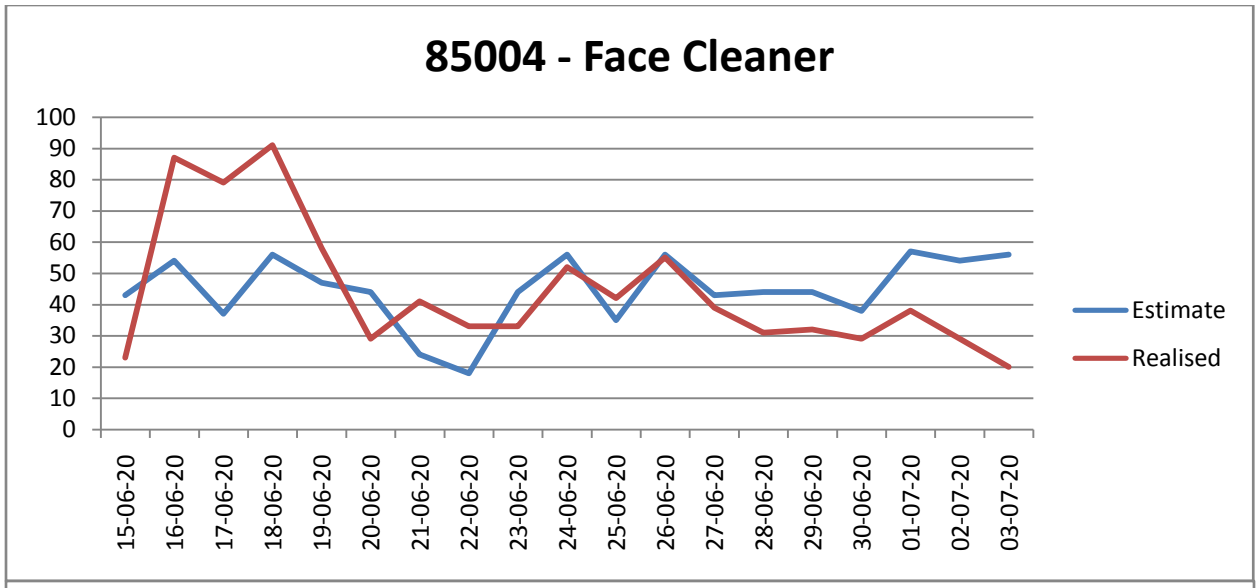
We ran the same model for all the products with 2 exceptions. The men's jacket by Koton was out of stock for the entirety of the time period. So we assigned a zero to it automatically. The bikini top by Trendyolmilla had only size 34 during our time period. Even though it still sold one or two units some of the days, we decided that this was not feasible to predict and we assigned a zero to it as well.

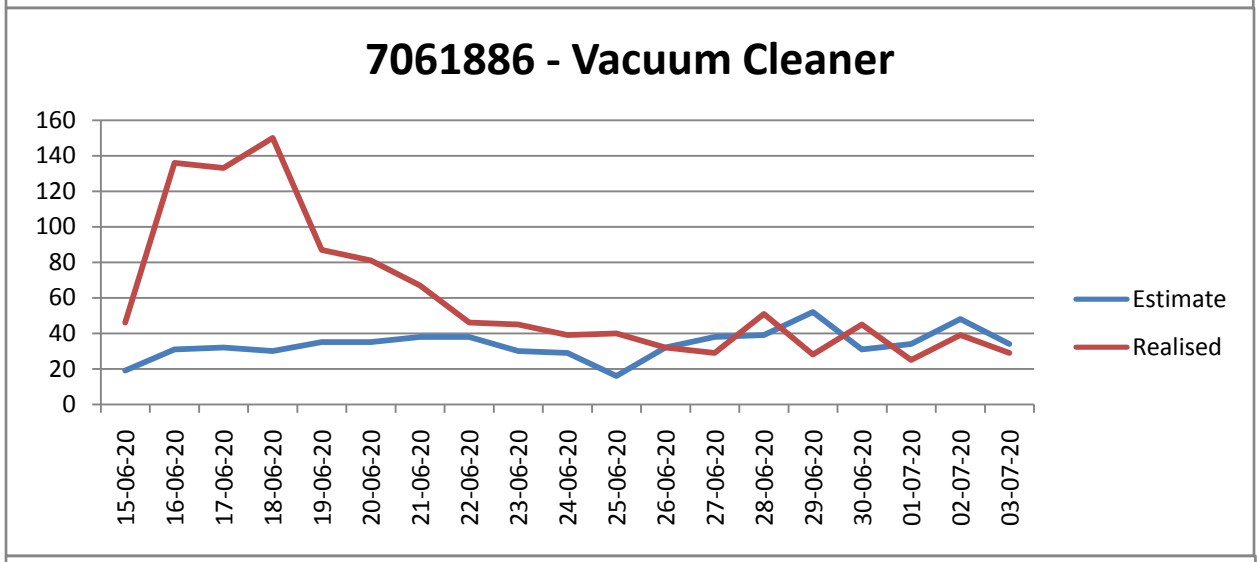
Our model is a linear model which uses data from 1st of march and only the days of the week that are same as the one we are trying to predict, as stated above. The model has 5 coefficients. Those are the price, visit count, basket count, favored count and the intercept. With this model we achieved a median r-squared value of around 0.75. We also tried to run the model without an intercept but that significantly lowered the r-squared value so we decided to keep the intercept in the model.

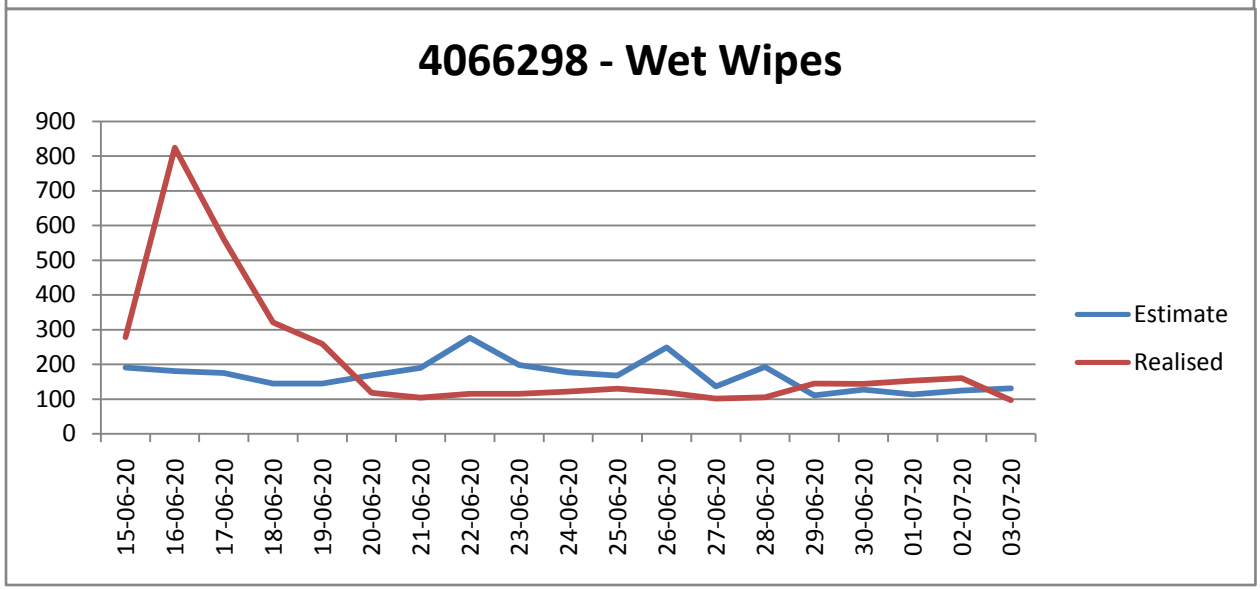

6676673 - Earbuds
800
700
600
500
400
300
200
100
0
15-06-20
16-06-20
17-06-20
18-06-20
19-06-20
20-06-20
21-06-20
22-06-20
23-06-20
24-06-20
25-06-20
26-06-20
27-06-20
28-06-20
29-06-20
30-06-20
01-07-20
02-07-20
03-07-20
Estimate
Realised


32939029 - Toothbrush
70
60
50
40
30
20
10
0
15-06-20
16-06-20
17-06-20
18-06-20
19-06-20
20-06-20
21-06-20
22-06-20
23-06-20
24-06-20
25-06-20
26-06-20
27-06-20
28-06-20
29-06-20
30-06-20
01-07-20
02-07-20
03-07-20
Estimate
Realised


31515569 - Leggings
1000
900
800
700
600
500
400
300
200
100
0
15-06-20
16-06-20
17-06-20
18-06-20
19-06-20
20-06-20
21-06-20
22-06-20
23-06-20
24-06-20
25-06-20
26-06-20
27-06-20
28-06-20
29-06-20
30-06-20
01-07-20
02-07-20
03-07-20
Estimate
Realised


85004 - Face Cleaner
100
90
80
70
60
50
40
30
20
10
0
15-06-20
16-06-20
17-06-20
18-06-20
19-06-20
20-06-20
21-06-20
22-06-20
23-06-20
24-06-20
25-06-20
26-06-20
27-06-20
28-06-20
29-06-20
30-06-20
01-07-20
02-07-20
03-07-20
Estimate
Realised


7061886 - Vacuum Cleaner
160
140
120
100
80
60
40
20
0
15-06-20
16-06-20
17-06-20
18-06-20
19-06-20
20-06-20
21-06-20
22-06-20
23-06-20
24-06-20
25-06-20
26-06-20
27-06-20
28-06-20
29-06-20
30-06-20
01-07-20
02-07-20
03-07-20
Estimate
Realised


4066298 - Wet Wipes
900
800
700
600
500
400
300
200
100
0
15-06-20
16-06-20
17-06-20
18-06-20
19-06-20
20-06-20
21-06-20
22-06-20
23-06-20
24-06-20
25-06-20
26-06-20
27-06-20
28-06-20
29-06-20
30-06-20
01-07-20
02-07-20
03-07-20
Estimate
Realised

Our results are shown in the graphs above. As the graphics show, our predictions do a relatively good job of estimating the trend of the data. Also, our predictions and the realized sales had similar ranges. However, we were way off on our estimates on couple of occasions and we were not really close in general. In addition, with hindsight, we can see that our estimates for the toothbrush sales were all over the place and had no relation to the realized sales.

Product	MAPE
6676673	47.53
31515569	44.78
85004	45.82
7061886	43.99
4066298	54.91

We used the same MAPE formula which is used to evaluate this assignment to properly compare our results. Instead of doing it on a daily basis we calculated MAPE of each product. Results are shown on the table above. One thing to note here is that the toothbrush dataset is omitted from this list. Since some of the days the toothbrush sales were zero, and since the MAPE formula requires us to divide by the actual sales, we couldn't get a result. Also, omitting those days where the actual sales numbers were zero resulted in us finding a MAPE value of 156.07 so we decided to ignore that product.

Our MAPE values are concentrated in the 44% - 55% range. This is not a good result since we need a MAPE value below 20% to call it a good score. Being almost 50% off the actual value is not at all good, but since this our first attempt at predicting such data it can be considered decent.

All in all, our model was a very simple one with very few variables and very few data. With these in mind, our model did decent. We believe that using days of the week was a good idea but we might have had better results if we were to use the whole data set going back to spring 2019. Also, adding another variable which shows the sales from the last 3 days should help this model catch those stretches of days where the sales have changed drastically. This way, the 4 day stretch where the toothbrush sales were zero for example, could have been better predicted. This would also turn our model into a daily ran model. Which would help us make last minute changes (if necessary) to the model.

The R codes that we have used can be seen at our GitHub pages.

https://github.com/ETM-58D/spring20-nsbdeniz

https://github.com/ETM-58D/spring20-mmerveesin