Submission of Assignment for the position of “Data Scientist”

Jorge Orozco

NEWYORKER

Agenda

- Introduction and remarks
- Task 1: Pricing
- Task 2: Regression

Introduction and remarks

- All working code is in the jupyter notebook files. Feel free to execute them to reproduce the results.
- A summary of the findings and main points will be presented in this document.
- To see a more in-depth of the code, please refer to the notebooks.

Task 1: Pricing

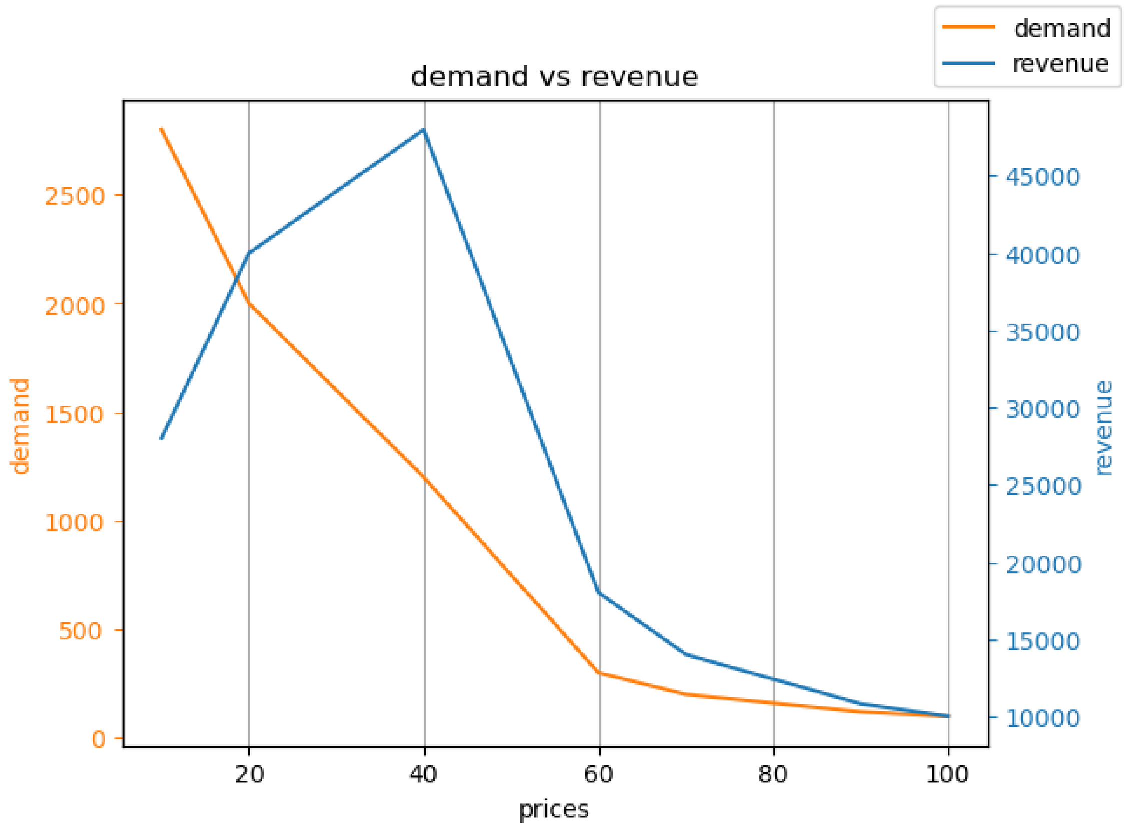
Our marketing manager was surveying the willingness to buy one of our fashion items at a certain price. They discovered the demand quantities, i.e. the number of people willing to buy, at various price levels. We have them below in two arrays.

Please find out the optimal price that maximizes revenue.

Task 1: Pricing. Overview

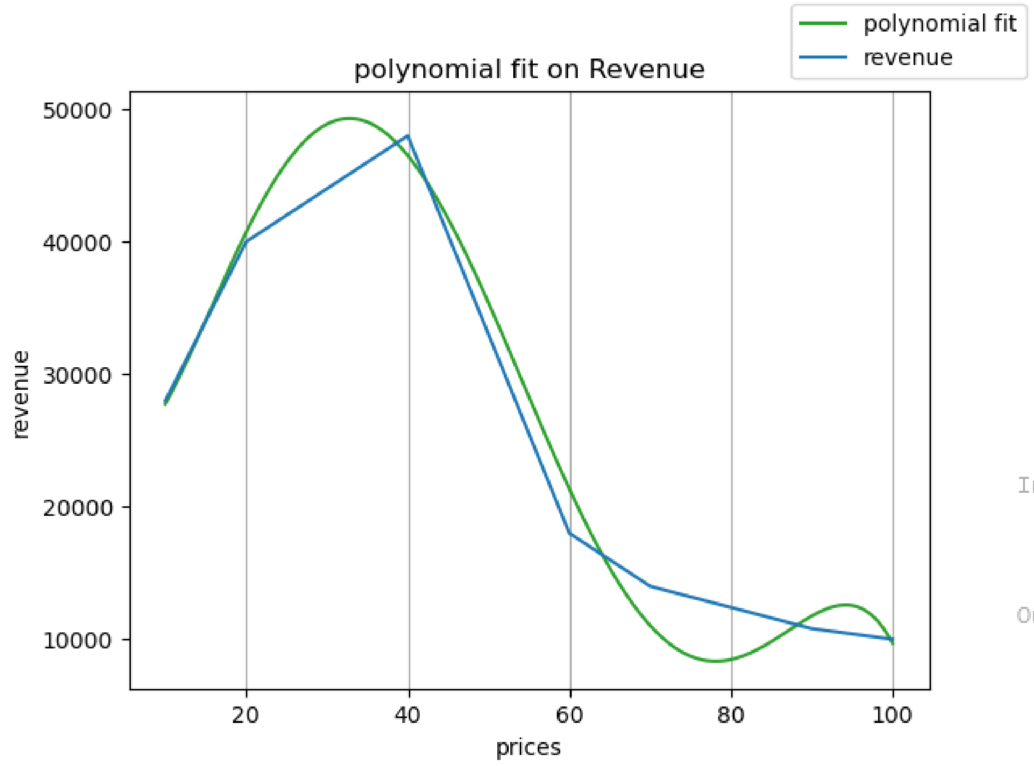
Out[6]:

	prices	demand	revenue
0	100	100	10000
1	90	120	10800
2	70	200	14000
3	60	300	18000
4	40	1200	48000
5	20	2000	40000
6	10	2800	28000



- Revenue column immediately obtained through a simple equation:
  - Revenue = prices * demand
- General observations on the plot:
  - The lower demand, the higher the price
  - The higher the price, the lower the demand
  - A simple answer to the general problem can be **40**, as it maximizes the revenue for the values given.
  - (Blue line) There is a “crest” missing to the left side of the 40. If we fit a curve there, can we fit the optimal value?
  - Two methods proposed: fitting a polynomial, and interpolate.

Task 1: Pricing. Method 1



- Using numpy.polyfit to play around and find a good enough polynomial.
- The polynomial of degree 5 that best fitted the data is

$$-0.0002794\,x^5 + 0.07301\,x^4 - 6.438\,x^3 + 204.9\,x^2 - 1352\,x + 2.652e{+}04$$

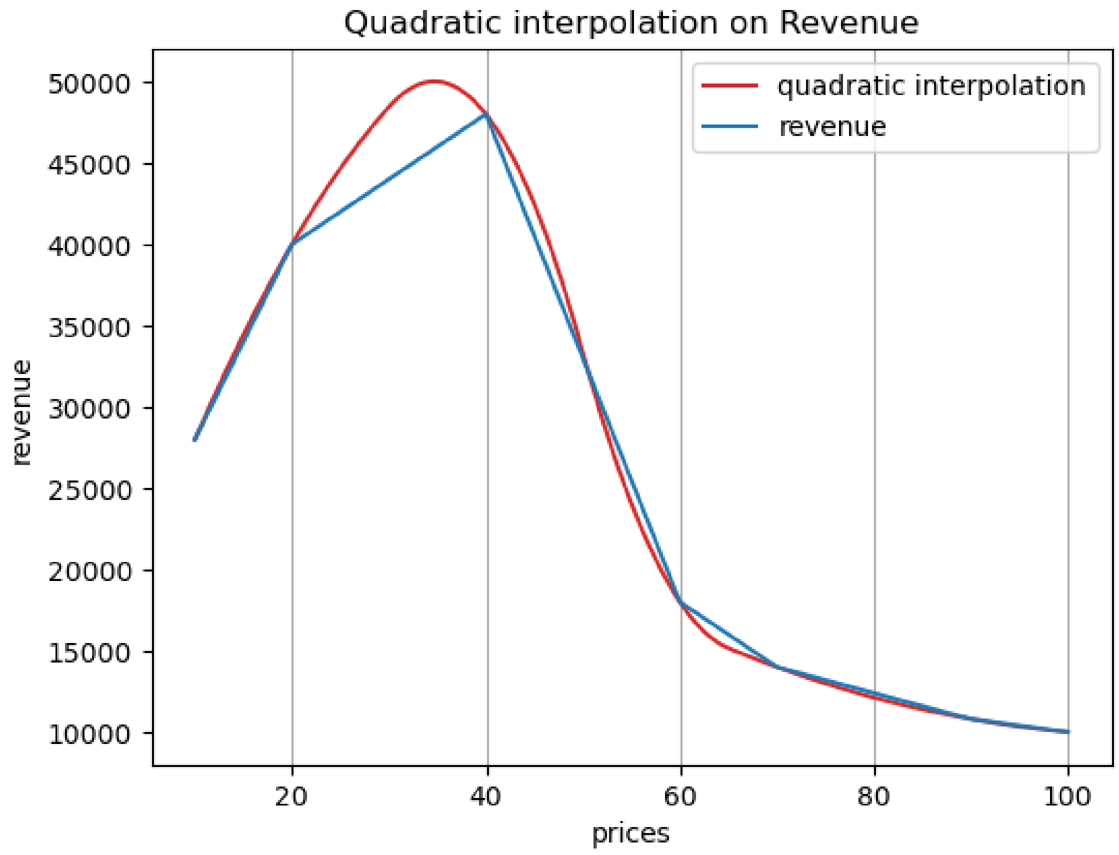
- We can then treat it as a bounded maximization problem, and can be treated with the scipy minimize function

In [12]:
```
res = minimize(-p, 20, bounds=Bounds(10,100))
res
```

Out[12]:
```
  message: CONVERGENCE: REL_REDUCTION_OF_F_<=_FACTR*EPSMCH
  success: True
   status: 0
      fun: -49309.28450823987
        x: [ 3.278e+01]
      nit: 5
      jac: [-1.455e-03]
     nfev: 16
     njev: 8
 hess_inv: <1x1 LbfgsInvHessProduct with dtype=float64>
```

Optimal price found: 32.78

Task 1: Pricing. Method 2



- Second approach: use scipy interp1d to find the interpolated value over a smoothed curve.
- Used the "quadratic" as interpolation method.
- Crest of the data is well approximated.

Optimal price found: 34.71

Task 1: Pricing. Further observations

- A single function called *revenue_maximizing_price* was given as requirement in the assignment but should not be used to generalize data optimization. This was part of an ad-hoc investigation and other data with other structures might require different exploration and methods.
- The two results of the exercise gave 32.78 and 34.71. I would recommend the marketing manager to select **33.75**, the average of both.
- This exploration only is used to maximize revenue. No assumptions on cost were considered.

Task 2: Regression

In the attached file sales.csv there is weekly sales of individual product types.

1. What can be said about the overall trend and seasonality of sales? What of the individual product type?
2. Are there correlations between sales of some product types, and if so, which?
3. Select a single product type and make forecast about its sales for 5 time periods (weeks) from the last observed data point. Note: Please make sure that we can reproduce your results, and feel free to ask questions if needed.

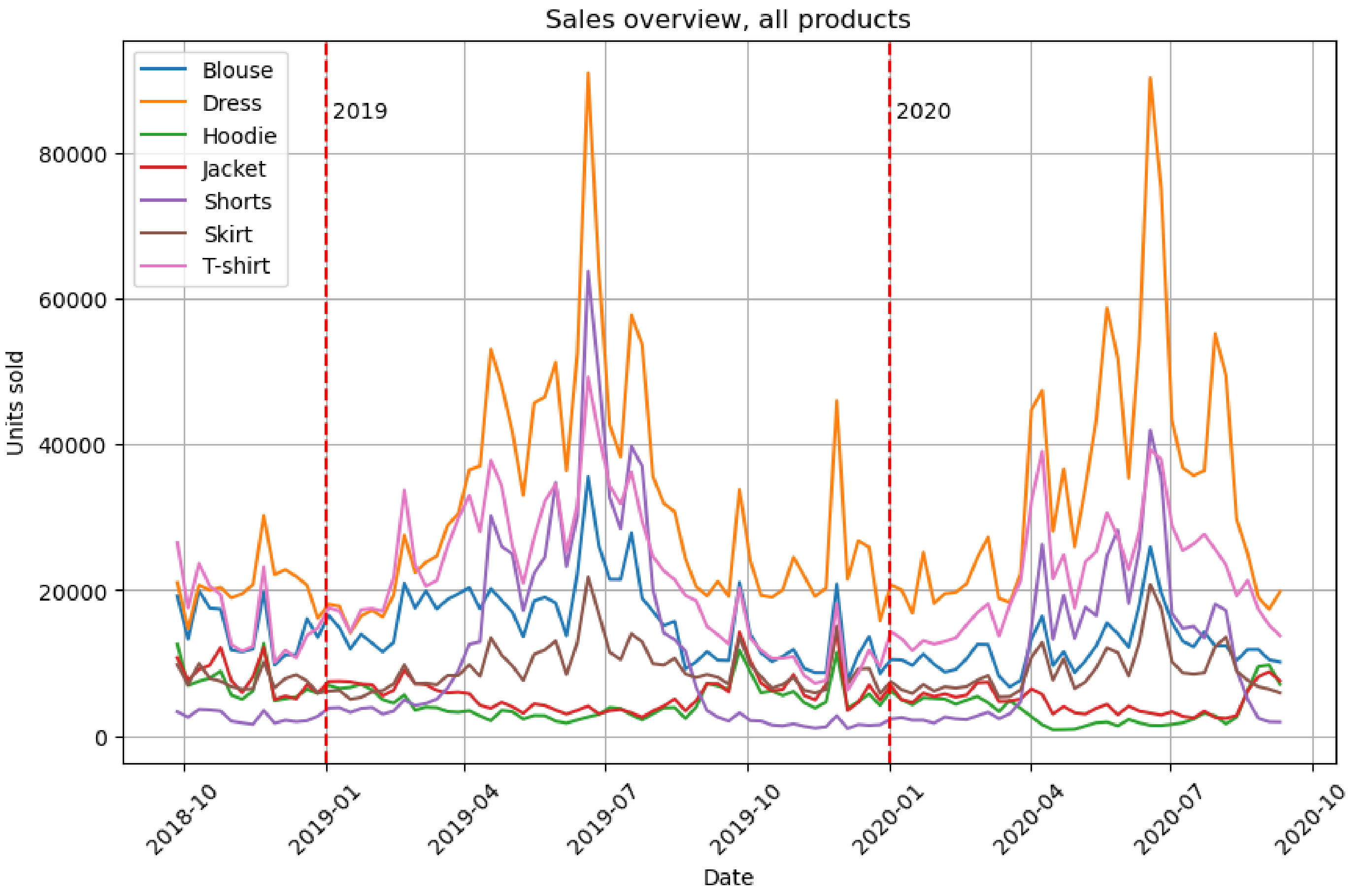
Task 2: Regression. Overview

- **Assumption**: data is in units sold.
- Span of 3 years, 2018-09-27 to 2020-09-10.
- Almost two years or weekly data.
- Products have a similar selling behavior.
- Dress and T-Shirt were items with top items sold on a given period:

```
max_seller
Dress      92
T-shirt    11
```

- Hoodies, Jackets and Shorts were the worst selling items:

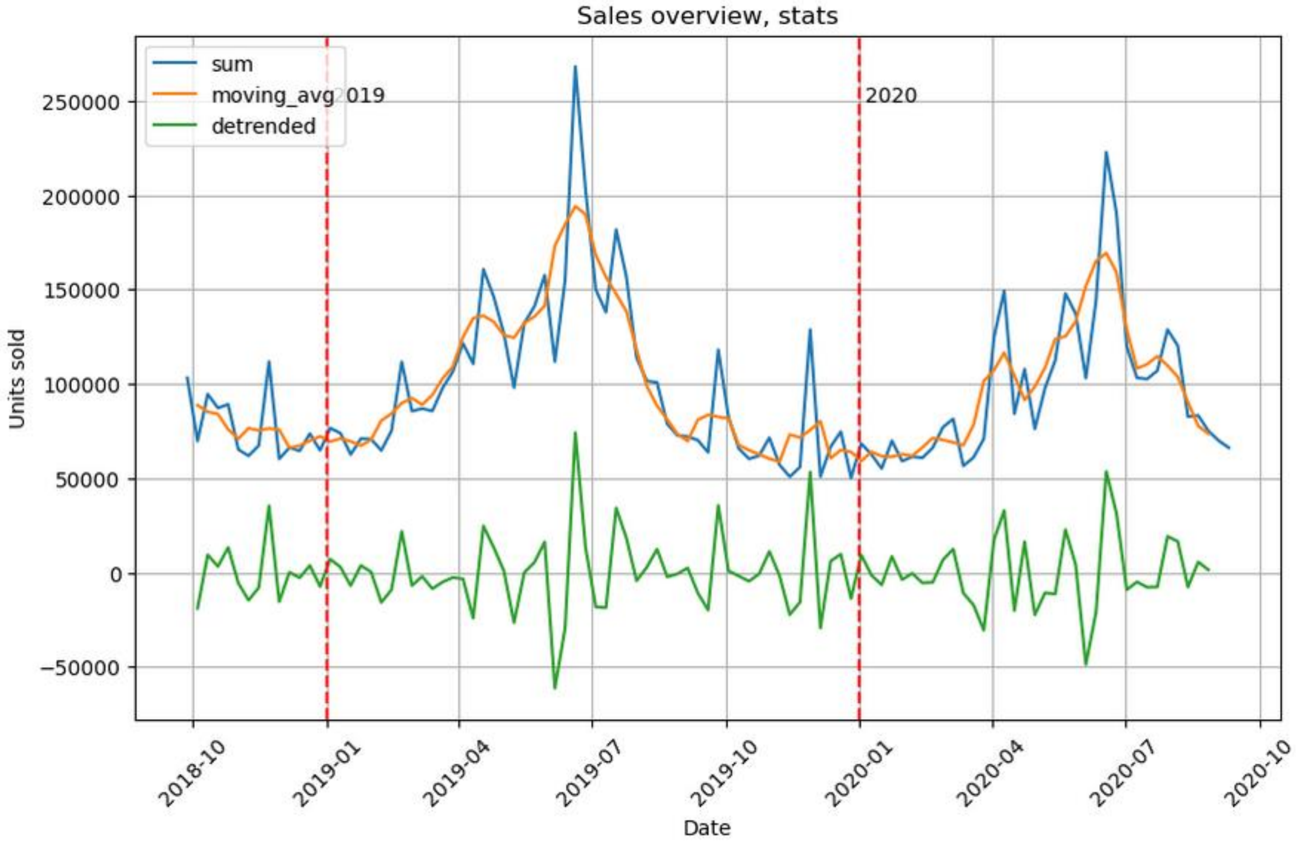
```
min_seller
Hoodie    45
Jacket     3
Shorts    55
```



	Blouse	Dress	Hoodie	Jacket	Shorts	Skirt	T-shirt
best_period	2019-06-20 00:00:00	2019-06-20 00:00:00	2018-11-22 00:00:00	2019-11-28 00:00:00	2019-06-20 00:00:00	2019-06-20 00:00:00	2019-06-20 00:00:00
worst_period	2020-03-19 00:00:00	2019-01-17 00:00:00	2020-04-16 00:00:00	2020-08-06 00:00:00	2019-12-05 00:00:00	2019-01-17 00:00:00	2019-12-05 00:00:00
total_items_sold	1480410	3204910	469219	584533	1148128	922041	2173075

Task 2: Regression. Question 1

What can be said about the overall trend and seasonality of sales?



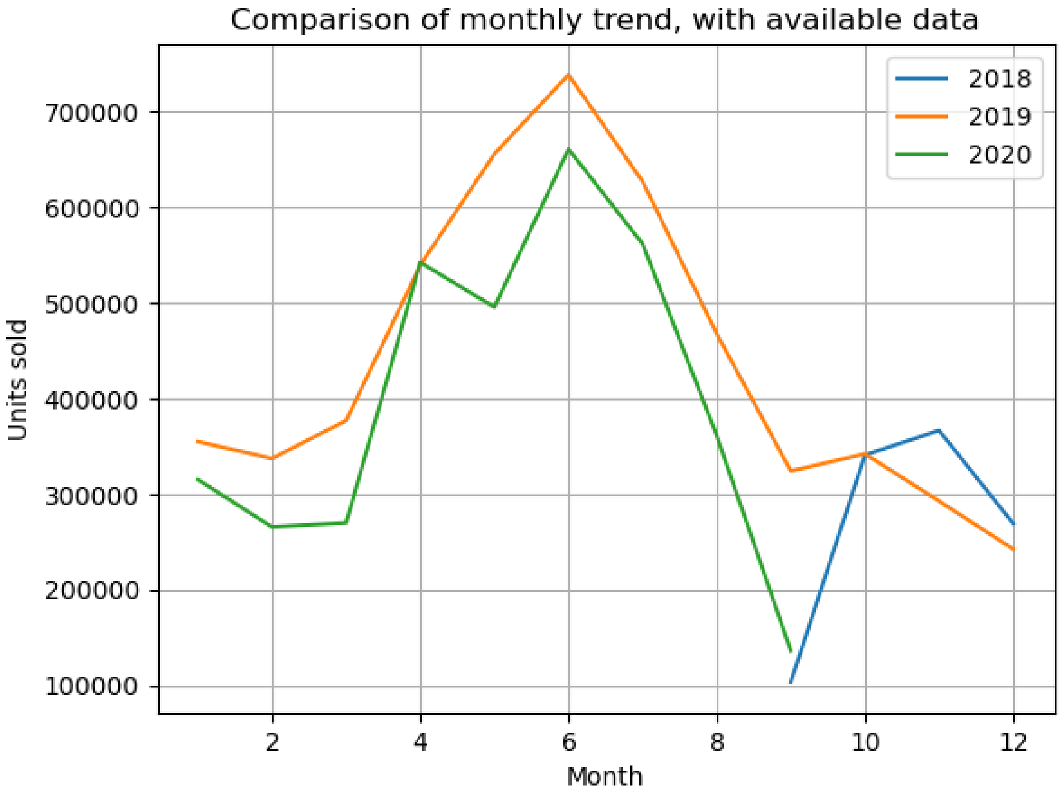
- Used overall sum of products sold on each period (Blue).
- Moving average with m=4 (a month) to obtain the **trend** of the sales. Biggest sales period is around summer. Some spikes over the year can be seen.

- Subtracted **trend** to the original values to obtain the de-trended values. Big spikes of data can be seen. A closer look can give marketing a good idea on events where sales would happen.

Task 2: Regression. Question 1

What can be said about the overall trend and seasonality of sales?

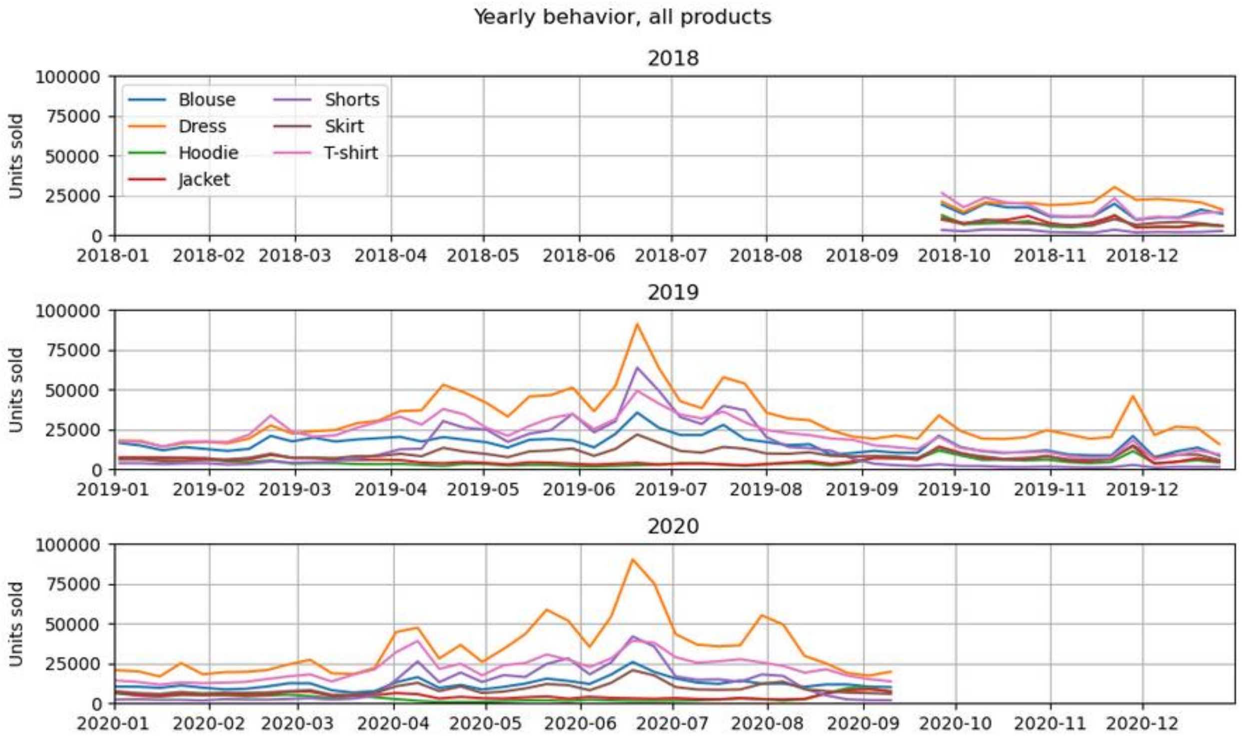
- When grouped monthly, the sales of 2019 were greater than of 2020
- A regression model should account for that.



Task 2: Regression. Question 1

What of the individual product type?

- Products follow a similar pattern with each other (more to come on that).
- No comments can be made over a year vs product comparison, because 2019 is the only year where we have complete data.



year	Blouse	Dress	Hoodie	Jacket	Shorts	Skirt	T-shirt
2018	204061	290105	100735	111537	35979	108196	228617
2019	826534	1647917	238258	299462	666709	482704	1134713
2020	449815	1266888	130226	173534	445440	331141	809745

Task 2: Regression. Question 2

Are there correlations between sales of some product types, and if so, which?

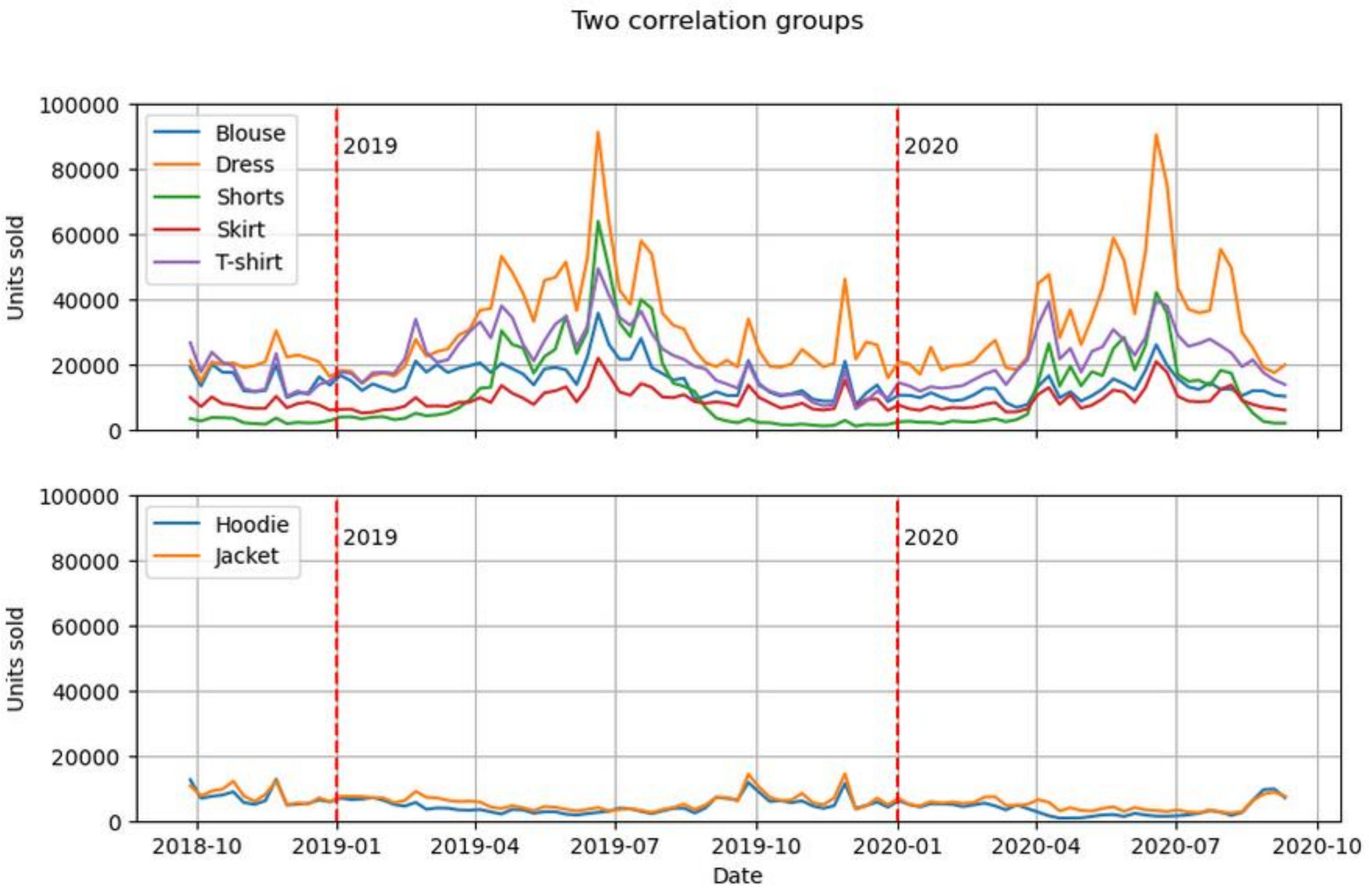
- There are two main groups: {Blouse, Dress, Shorts, Skirt, T-shirt} and {Jacket, Hoodie}.

- Jackets and Hoodies are only bought together.
- The best periods for Hoodie and Jacket was 2018-11-22, and 2019-11-28 respectively. People only buy them to prepare for the cold weather.

- Blouse, Dress, Shorts, Skirt, and T-shirt are correlated with each other.

Task 2: Regression. Question 2

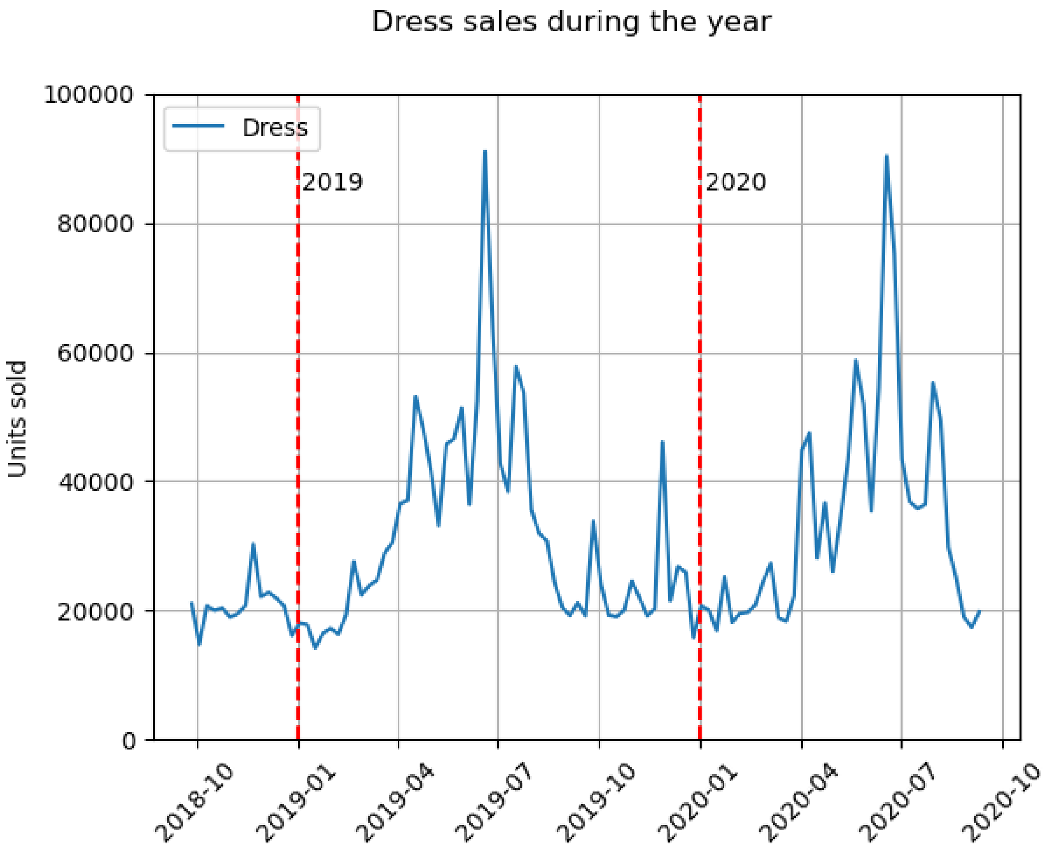
Are there correlations between sales of some product types, and if so, which?



- The seasonality on both groups van be observed once data has been separated

Task 2: Regression. Question 3

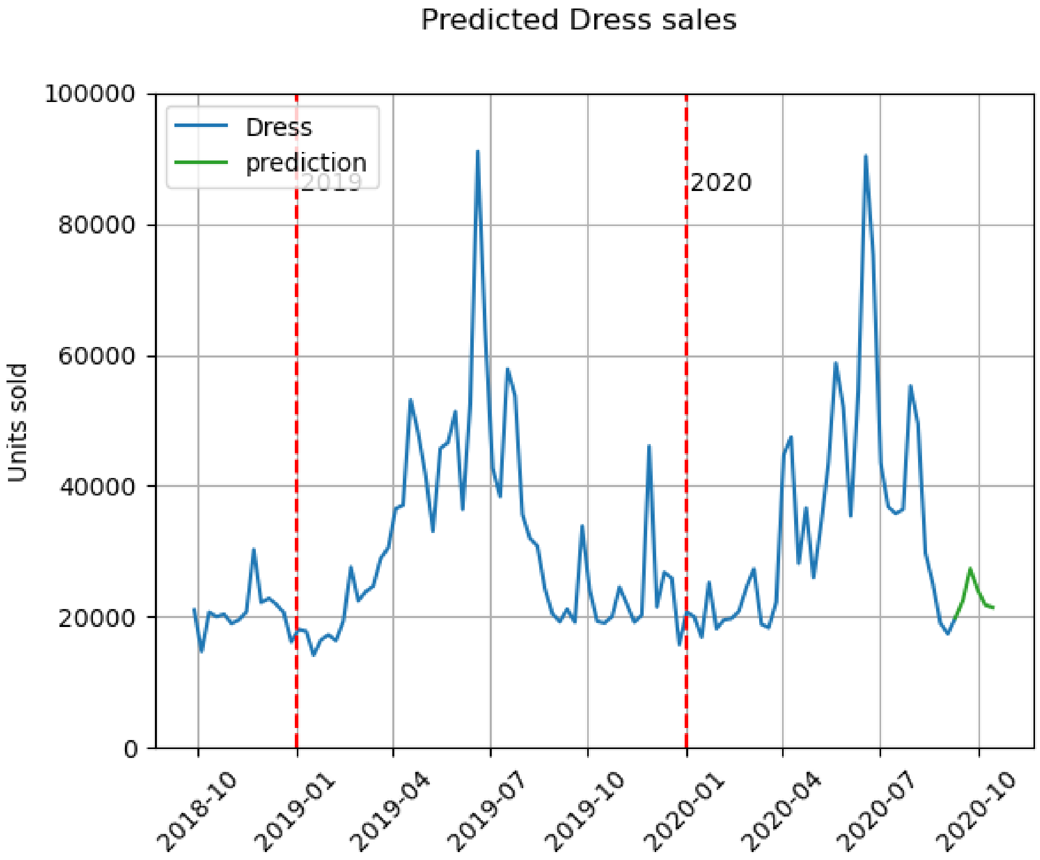
Select a single product type and make forecast about its sales for 5 time periods (weeks) from the last observed data point.



- Dresses are the most sold item, forecasting its sales was considered the most important

Task 2: Regression. Question 3

Select a single product type and make forecast about its sales for 5 time periods (weeks) from the last observed data point.



- 28 features, including 25 lag features in total:

```
Index(['date', 'Dress', 'month', 'week_no', 't-25', 't-24', 't-23', 't-22',
       't-21', 't-20', 't-19', 't-18', 't-17', 't-16', 't-15', 't-14', 't-13',
       't-12', 't-11', 't-10', 't-9', 't-8', 't-7', 't-6', 't-5', 't-4', 't-3',
       't-2', 't-1'],
      dtype='object')
```

- Loss of data in the beginning due to lag features
- Random forest chosen as model
- 78 training, 10 testing datapoints
- R2 score: 0.408
- Predictions (green) do predict the smaller spike that is seen in previous year.

Further observations

- More models can be explored, more feature engineering, hyperparameter selection.
- A comparison of the sales of the year 2020 and 2019 could be done if taken only up to the month where data is available.
- Feel free to run the code, or contact me if there are any doubts.

Thank you!