DATA ACQUISITION AND PROCESSING SYSTEMS ELEC0136 20/21 REPORT

SN: 19162034

ABSTRACT

This report intends to estimate the closing stock price for American Airlines for the month of May 2020. A machine learning algorithm will be trained using time series distribution as well as three external variables: stock price of competitors, fuel prices and number of passengers per day in the US. An additional seasonality factor has been introduced to account for the effect of the COVID-19 pandemic. These four datasets will be preprocessed and the data in them explored to find relationships between them. Two models will be compared: one based only on time series information from the American Airline stock, and one on the time series plus external variables. The second model performs better in all metrics.

1. INTRODUCTION

Stock price estimation is a useful tool to predict whether a stock price will go up or down. This is a challenging area, as little information is available that could predict a stock price in the future, basing itself in speculation and assumptions. Perhaps the most utilized method is forecasting by time series [1], where the historical stock price is analyzed to find patterns and then estimated in the future. Additional variables can better these results, but the challenge is finding variables that are relevant to the original dataset that the prediction will be based on and extrapolating them in the future.

This report shows the working and decision behind designing an algorithm to predict the closing stock price of American Airlines in a specific time period, comparing results to real data. Data will be collected, preprocessed and analyzed to find how it relates to the dataset. Three additional datasets will be acquired: stock price of American Airline's competitors, the fuel prices and the number of passengers. The end goal of this project is not only to estimate the prices and obtain as good results as possible, but also to showcase different methods for data preprocessing and exploration. An additional seasonality factor is also introduced to account for the effect of the COVID-19 pandemic [2].

2. DATA DESCRIPTION

The closing stock price for American Airlines will be the main dataset. It is composed of a table with an entry for every day and multiple columns, such as opening price, closing price, max price, volume… etc. However, the only column useful for the analysis to be done is the closing price, as this is what the algorithm will try and predict. The ultimate objective is to compare the inference of data based on solely this dataset and when combined with other factors that could contribute to the change in stock price. These factors are described below, along with the format of the data.

- The first factor taken into account was the most direct that could be extracted: the stock price of the competition. To correctly evaluate what the competition for American Airlines was, a comparison of their local and international flights had to be done. Choosing one of those factors, say the number of passengers in national and international flights per year, it is clear that the main focus for American Airlines is local traffic. Therefore, the passenger data for airlines operating mainly in the U.S. was collected, and seeing the table below, three of them can be seen as the biggest competition: Delta, Southwest and United Airlines [3]. The jump in numbers from the first four to Alaska Airlines is too big, and therefore these will be the only ones included in this dataset.

	Airline	**2019**	**2018**	**2017**
1	American Airlines	215	204	200
2	Delta Airlines	204	192	186
3	Southwest Airlines	162	163	157
4	United Airlines	162	158	148
5	Alaska Airlines	46	45	44

Table 1: Millions of passengers per airline

  A change in their stock prices could indicate a change to come in the stock price for American Airlines. Same as the American Airlines stock price, the format of these will be a table with an entry for each day and many columns, while only mattering the closing price.
- The second factor was cost-based. To do this, the cost that is most variable should be chosen. Personnel costs, maintenance… etc. should not vary much, and they could be approximated to a fixed cost. Therefore, the cost that was chosen is the jet fuel cost, which is around 10% of operating expenses [4]. The full historical prices are only available per month, and not per day. Therefore, this factor could be used to estimate long term changes in the stock prices, and not momentary spikes or valleys.
- The third one was the number of passengers. The more passengers travelling, the more the stock price should go up (holidays, summer). The dataset will be formed from two sources (see section 3), but the final dataset to

be formed will be only two columns: date and passengers. A better metric would have been load factor, that is, capacity utilization of Aircrafts, but this data proved to be too challenging to obtain [5].

3. DATA ACQUISITION

The data for stock prices was obtained from Yahoo! Finance. The tables were downloaded from 01-04-2017 to 31-05-2020. While the prices will be estimated on 05-2020, it is useful to have the information of the real stock prices to examine the results of the algorithm. The library used was yfinance, for downloading historical prices from Yahoo! Finance. [6]

For the jet fuel cost, index mundi [7] provides a historical price table from December of 2015 until November of 2020. The HTML code of this table was obtained and read with the pandas function *read_HTML*. This dataset was stored.

The number of passengers is provided by the TSA (Travel Security Administration) [8]. The table containing this data was created to compare current travelling numbers (through the COVID-19 pandemic) with numbers the year before. Therefore, there is no information before March of 2019. The HTML code of the table was obtained and the dataset was stored. Some preprocessing was needed to store the data in a tidy format, as the original looked like in Table 2.

Date	Passengers 2021	Passengers 2020	Passengers 2019
01/03/2020	X	Y	Z
02/03/2020	X1	Y1	Z1

Table 2: Format of passenger dataset

To solve the problem, columns were sampled and copied into another dataset where the dates matched the actual dates.

Another data source is used from the Bureau of Transportation Statistics (BTS) [9], which has monthly data from 2002. The same method was followed. These two datasets will be combined, as explained in section 5.4.

4. DATA STORAGE

The data was stored locally in csv format. The data is after imported into the Jupyter notebook, where it is preprocessed. After this step is finished and the final dataset that will be used on data exploration and inference is created, this dataset is saved again to avoid preprocessing the original dataset every time the kernel is run. A final dataset is then created and saved in a csv file, where every column corresponds to a variable and every row to a date.

5. DATA PREPROCESSING

Every dataset collected was preprocessed. How this was done is explained in the section corresponding to it.

5.1. American Airlines stock price.

The main set of data that will be used in the inference presented some problems. The first and most significant was the missing data. The stock market is closed on weekends and some major holidays (1st of January), and therefore these dates were not on the csv file extracted from Yahoo! Finance. To deal with this a strategy was devised that would also be used in other datasets. A vector of dates is created with the function "date_range" in pandas ranging from the 1st of April 2017 to the 31st of May 2020. For every date of the table not in this vector, a new row is created. The values that this row will have will be filled via linear interpolation, that is, sampling the linear regression curve formed by the previous and next valid values. The first two days (1st and 2nd of April 2017) are a weekend, and since there is no previous valid value to interpolate, the value given to them is of the next valid date. This data is stored in a csv file, which will be the one imported from now on.

Apart from this, the original format of the date that was imported was as an object. To be able to plot it, it had to be converted to "datetime64" format.

The boxplot for this dataset can be found in Figure 1.

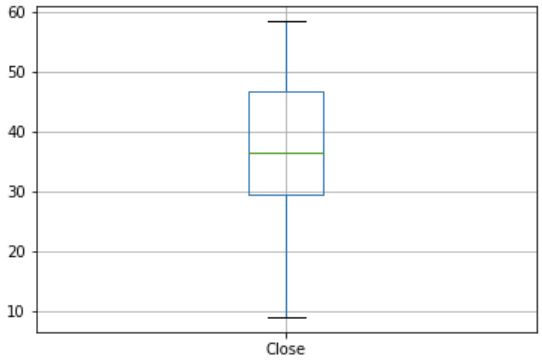

Figure 1: Boxplot for American Airlines closing stock price

As it can be seen, there are no outliers detected by this method. However, this is checked only to find any outliers in the upper end of the dataset. If outliers were detected below the minimum, further information would have been needed in order to take action, as the stocks fell very predictably due to the COVID-19 pandemic.

5.2. Competitor stock prices

The three companies selected to form this dataset, as said before, were Delta, United and Southwestern Airlines. Perhaps a concern could be raised that inserting an additional 3 datasets to the algorithm might be unnecessary. To address this, instead of using every stock price as an independent variable, the weighted mean of all three will be calculated for each day and this will be the only variable used in the algorithm. As seen in Table 1, Delta has similar passenger numbers as AA, so its weight will be double. The other two will have unitary weight. To calculate the mean, the missing date values were accounted for, as explained in 5.1. For every day, the mean was calculated and then input into a vector. As before, the date format had to be changed from object to

"datetime64". The boxplots for every company and the mean can be found in Figure 2.

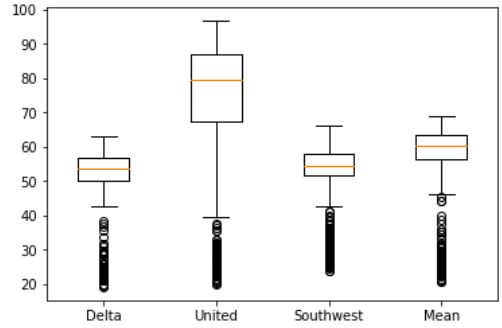

Figure 2: Boxplots for competitor's closing stock prices and mean

As it can be seen, United has much more variance than the other two. A quick look at their stock prices shows that this is due to the fact that United had much higher stock prices before the COVID pandemic, but after it their stock prices were almost equal. Again, many outliers are detected. Considering the effect of the COVID-19 pandemic, these will not be dealt with, since they are true values that indicate how the industry was hit by the pandemic.

5.3. Jet fuel prices

A more limited dataset, in which data is only available once a month instead of every day. Therefore, more useful for detecting long term changes and not instant spikes [10]. The main objective of preprocessing this dataset is to match the length of it to the length of the stock prices. To do this, the data is interpolated between every two values and is sampled every day in between them. The boxplot obtained can be found in Figure 3

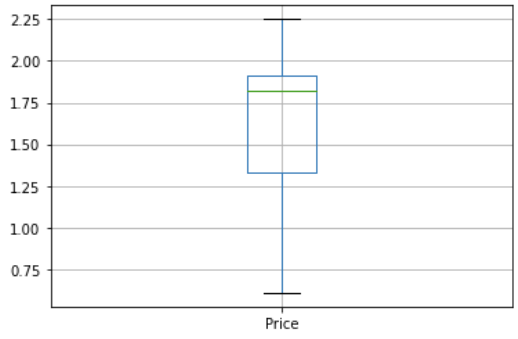

Figure 3: Boxplot for jet fuel prices

No outliers of any kind are detected.

5.4. Number of passengers

This dataset needed the most preprocessing of all, due to the lack of data to begin with. Travel reports inside the US are only released every month with a total number of passengers, which was considered to not be accurate enough. An alternative data source was found were data was available per day but only for 2020 and 2019. However, the dataset was curious, since the starting date was the 1st of March, and the most recent date is the day that it is. Therefore, if the dataset was seen on the 11th of January 2021, it would look like Figure 4

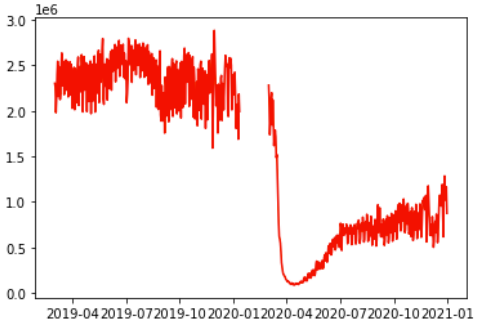

Figure 4: Passenger numbers from TSA

As it can be seen, a chunk of data is missing. This is due to the fact that, as said, the most current date is the date at which the data is collected. Therefore, if this data were collected on the 28th of February 2021, it would be complete. As peculiar as this might be, a workaround has to be found. The chosen strategy to replace the missing data is to use Facebook Prophet to estimate the time series and fill in missing values. For this, the training dataset is from 1/3/2019 to the last day before the missing data begins. An added feature was to account for holidays, as these massive impact the number if travelers per day. The time series was also considered to have weekly seasonality, since there are days of the week when people travel more. Both these hypothesies are corroborated by these graphs in Figure 5, where the increase or decrease in passengers is plotted.

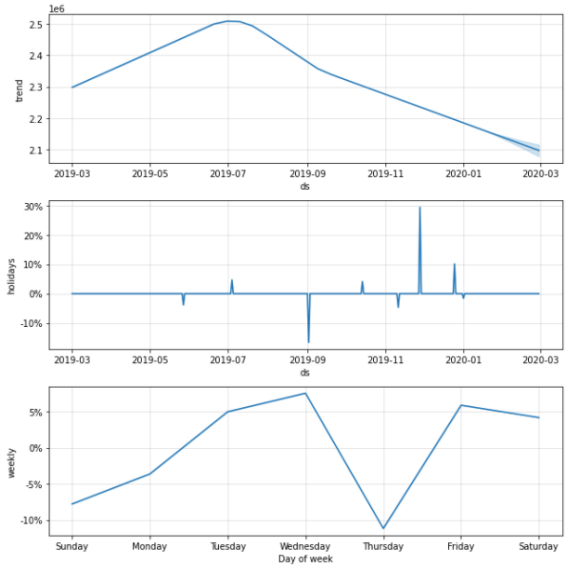

Figure 5: Trend and seasonality of passenger numbers

With this, estimating the missing values the final series is shown in Figure 6.

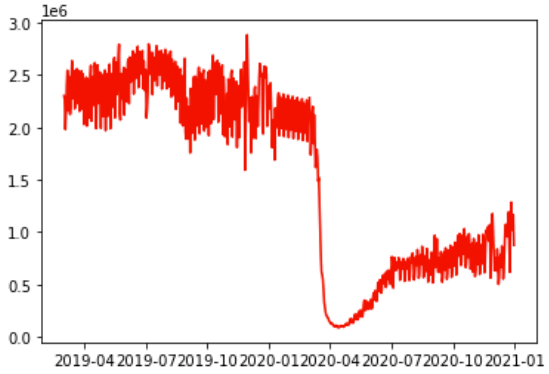

Figure 6: Completed passenger numbers by inference

However, the dataset must have the same size as the dataset with the stock price. An alternate source was chosen to collect data for previous dates (BTS), but since only monthly totals were available, the number of passengers for each month was considered constant every day. This data also should be cropped at the end of May 2020, since this is the end of the period to be studied. The final result is shown in Figure 7.

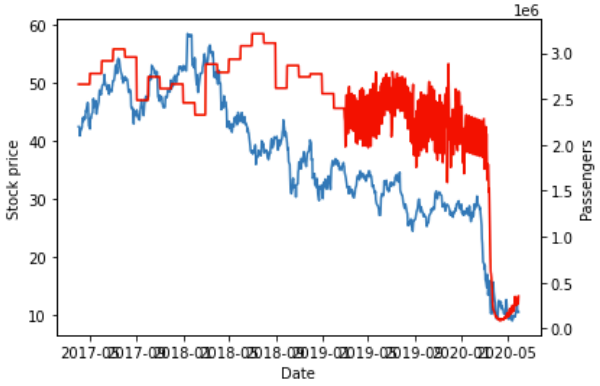

Figure 7: Complete passenger data from TSA, inference and BTS

The boxplot for this dataset can be found in Figure 8.

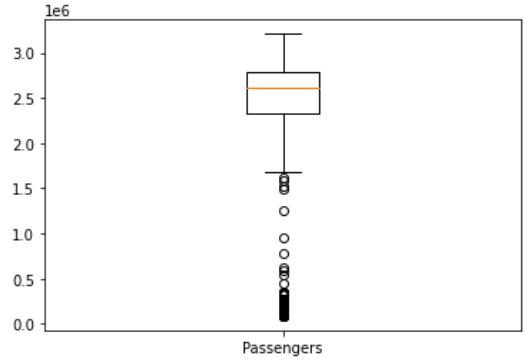

Figure 8: Boxplot for passenger data

Again, many outliers caused by the COVID-19 pandemic.

6. DATA EXPLORATION

The main bulk of the project is to analyze how these variables are related and estimate their values in the month of May. The seasonality of the main dataset will also be examined. To estimate the values of the variables for the month of May, forecasting by time series will be used, analyzing their seasonality and introducing two additional seasonalities that have to do with the drop that all variables took after the COVID-19 pandemic. The forecasting will be done using Facebook Prophet.

6.1. American Airlines

Figure 9 shows the historical stock prices for American Airlines until the 30th of April 2020. This dataset will be explored and hypotheses tested.

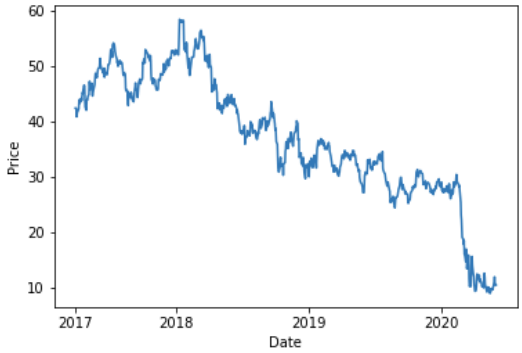

Figure 9: American Airlines closing stock price

First, examine the trend and seasonality of the original dataset of stock price for American Airlines. Three hypotheses will be tested: there is yearly seasonality, there is weekly seasonality and the holidays affect the stock price. To test this hypothesis, Facebook Prophet is used, using the entire dataset as a training set. Figure 10 shows this.

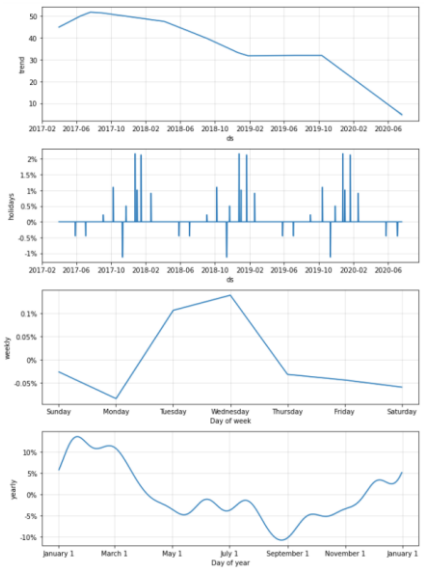

Figure 10: Trend and seasonality of American Airlines closing stock price

From these graphs, three results can be extracted:

1. There is yearly seasonality, since every year the same months have a variation in stock price.
2. There is no weekly seasonality, since the maximum spike per day is 0.1%, too small to signify anything.
3. The holidays have an effect. Every year the same days have very similar spikes, so these will be taken into account.

6.2. Competition stock prices

Figure 11 shows the historical mean price of the three competing stocks (red) against the American Airlines price (blue).

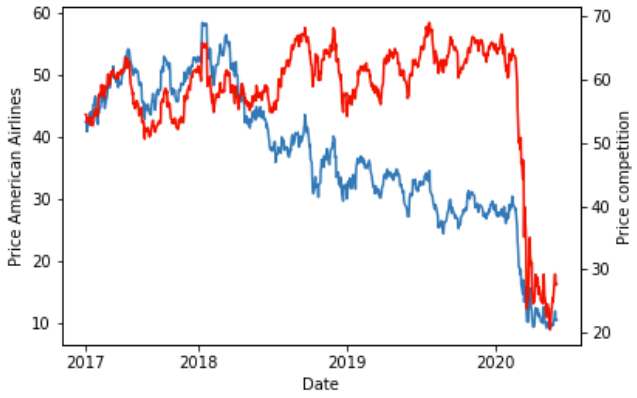

Figure 11: Competition closing stock prices vs American Airlines

An initial assessment is that the stock prices seem to be correlated. In the beginning they are similar, drifting further apart as time goes on, but the peaks and valleys appear in the same dates, so it is reasonable to think that they are correlated. Therefore, the hypothesis is presented: the American Airlines stock price and the price of their competition are independent. To assess this, many methods are used:

1. Comparing seasonality.
   With the same tool used in 6.1. the competition stocks are tested for yearly seasonality. The result can be found in Figure 12.

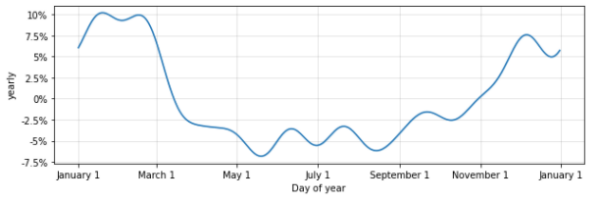

Figure 12: Seasonality of competition stock prices

   The results are quite similar to the American Airlines price, with increments and decrements in the same time periods. This strengthens the argument that they are correlated.
2. Scatter plot.
   Plotting the scatter plot of both variables should give an insight into their correlation [11]. Figure 13 shows that there seems to be a direct correlation.

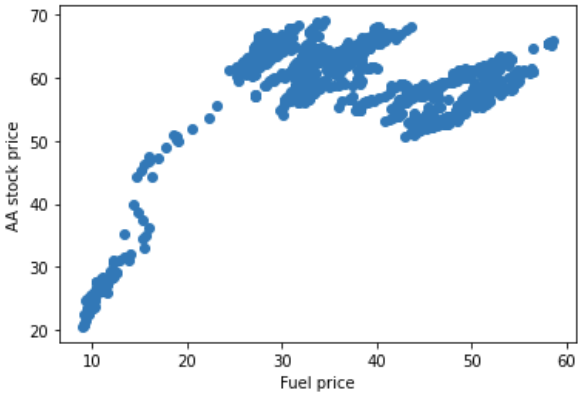

Figure 13: Scatter plot of competition vs American Airlines

3. Pearson and Spearman correlation coefficient [12]:
   These methods tally the correlation of two variables. Pearson does it linearly and Spearman non-linearly. These methods have two parts: the p-value and the correlation coefficient. The p value shows if there is correlation or not, while the coefficient tallies how much correlation there is. The correlation is measured from -1 to 1, showing if it is direct or inverse. The further from 0, the more correlated. If the p-value is higher than 0.05, the initial hypothesis will not be rejected, and the two variables probably won't have a correlation. The value obtained however is much smaller, therefore rejecting the initial hypothesis and concluding that there is a correlation between the competition's stock price and that of American Airlines. Since the coefficient is positive (0.45), this correlation will be directly linear.

To estimate the values for May 2020, Facebook Prophet was used, including seasonality parameters to account for the rapid change in March. The holidays were also taken into account. The result of the estimates for May against the real values can be seen in Figure 14.

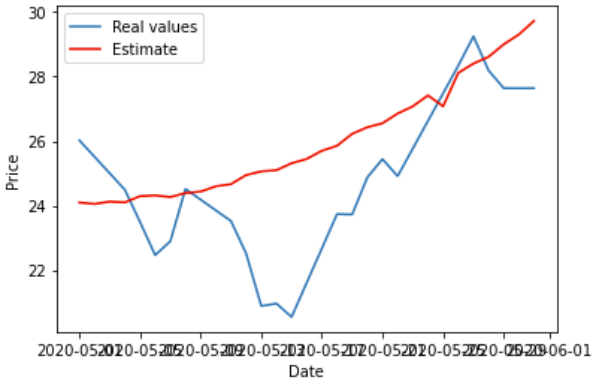

Figure 14: Inferred competition stock prices for May 2020

6.3. Fuel prices

For the next variable, the oil prices are plotted against the stock price in Figure 15.

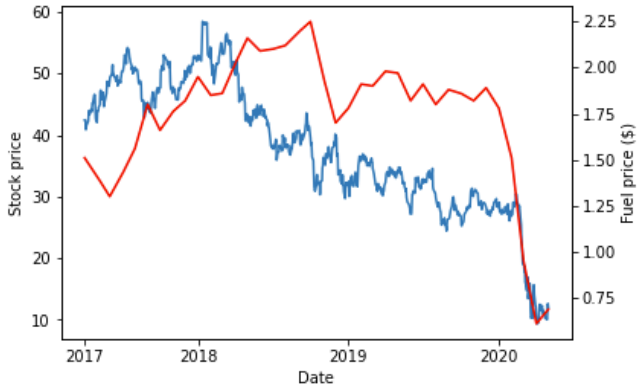

Figure 15: Fuel prices vs American Airlines stock

This plot shows contradictive results, since up until 2020 the two plots seem to be inversely dependent. When the fuel price rises, the stock price falls. However, when 2020 arrives, both plots fall drastically. Two of the previous methods will be used. The scatter plot is found in Figure 16.

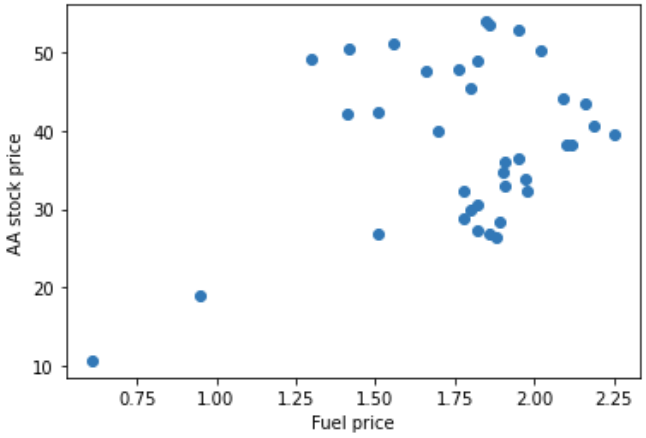

Figure 16: Scatter plot of fuel price vs AA stock price

This scatter plot seems lacking, there is perhaps not enough data to detect a pattern. The p-value is smaller than 0.05, and therefore the initial hypothesis can be rejected, and the two datasets are correlated. However, the coefficient is quite low (0.14), and therefore the correlation is very weak.

As mentioned in the data preprocessing step, the values for oil prices were very far apart, and could only indicate long term changes for stock prices. To do this, the missing data should be interpolated in order to account for every day. The problem was that the dataset was too small, and a pattern could not be detected. If the scatterplot is drawn for the dataset with the missing values, the results are much clearer (Figure 17).

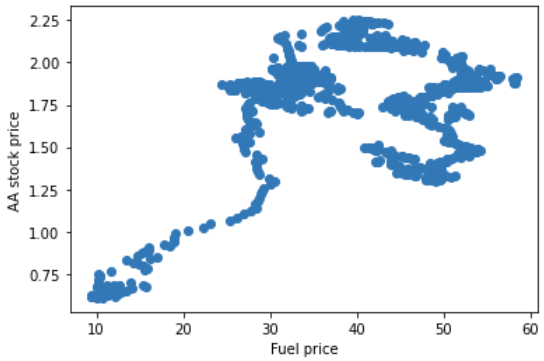

Figure 17: Scatter plot of interpolated fuel price vs AA stock

This looks to have a clearer correlation. The p-values agree with this observation, giving a coefficient much lower than 0.05, rejecting the initial hypothesis and proving that the two variables are correlated. The amount of correlation is 0.37, much higher than before. This is an example of a dataset which was originally not very valuable to the project but through preprocessing could greatly indicate trends in the target dataset. To estimate the values for May 2020, as before Facebook Prophet was used, with seasonalities stemming from the COVID-19 pandemic. The results can be seen in Figure 18.

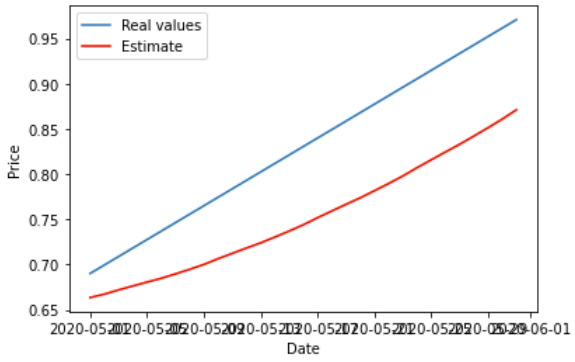

Figure 18: Inferred fuel price for May 2020

6.4. Number of passengers

The relation between the two variables can be seen in Figure 19.

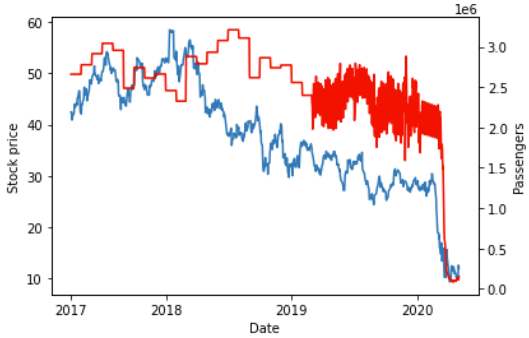

Figure 19: Passengers vs AA stock price

The same hypothesis will be tested: there is no correlation. The scatter plot can be found in Figure 20.

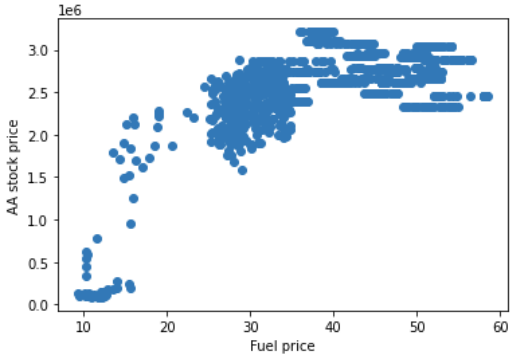

Figure 20: Scatter plot of Passengers vs AA stock

This shows a direct correlation is quite likely. The p-values are much below 0.05, and therefore the hypothesis is rejected to confirm that there is a correlation, with a coefficient of 0.67 (1 is the maximum). Even inferring values in the preprocessing step, the results are very good for this dataset. Again, the values for May will be estimated based on the previous time series. The seasonality taken into account was the effect of the pandemic, and a weekly seasonality found in section 5.4. The results can be seen in Figure 21.

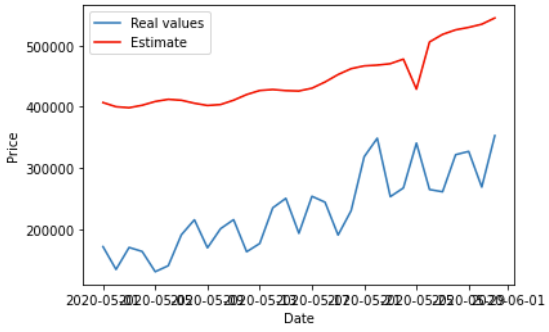

Figure 21: Inferred passenger numbers for May 2020

7. DATA INFERENCE

A single model will be implemented and trained to predict the closing price for American Airlines on May 2020, using only the available data until the end of April 2020. Therefore, the training dataset will be April 2017-April 2020 and the testing set will be May 2020. This will be done using only the company's data and the results will be compared if the other variables are taken into account. Their results and metrics will be compared.

7.1. Model only with company data.

A simple time series prediction will be made, with the data available until April 2020. The results can be seen in Figure 22.

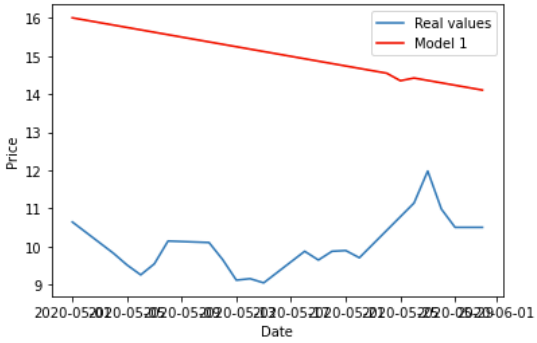

Figure 22: Inferred AA stock price for May 2020 without external variables

The results are clearly not very good. The model has no source of information except for past prices and the fact that there is yearly seasonality. Therefore, it simply compares this time period to other years and estimates the values. This method clearly does not work, since without extra information the estimates can't be accurate.
The trend continues being the same, as there are no external stimuli to change it. Therefore, the model keeps performing badly.

7.2. Model with external data

The time series will now have 3 additional regressors: the competition's stock prices, the fuel prices and the number of passengers. Additionally, seasonality due to the COVID-19 pandemic has been introduced. The results can be seen in Figure 23.

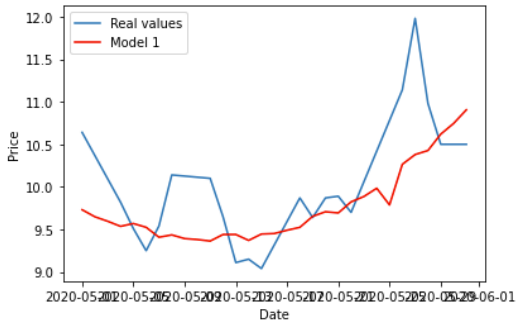

Figure 23: Inferred AA stock price for May 2020 with external variables

A simple look shows much better results. The model offers values which are not much higher, and the trend seems similar. However, this is based in a specific moment in time. If the test period was longer, the data would continue to increment exponentially until an external stimulus took effect. This is very unlikely, since a big change would not happen taking into account its time series. Therefore, it is possible to conclude that the model would not behave well when testing for values far into the future. Then again, this is reasonable, since there is much uncertainty.

7.3. Comparison

To compare the results quantitatively, metrics will be used [13]. They can be found in Table 2.

	Metric	Model 1	Model 2
1	Mean Squared Error	26	0.3
2	Mean Absolute Error	5	0.4
3	R-squared	-62	0.25

Table 3: Metrics for inferred models

The reason for tallying the mean squared error and mean absolute is that the mean squared error punishes more big errors in estimation. As expected, there are very big errors in the first model, mainly shown by the r-squared, which has a negative value. This means that a simple straight line would have provided better results than the model. For model 2, the r-square, although small, is positive, which means that 25% of the model's variance can be explained by the variables.

As for the residual distributions, Table 3 shows their mean, median and skewness [14].

	Metric	Model 1	Model 2
1	Mean	-5	0.28
2	Median	-5.35	0.19
3	Skewness	0.81	0.6

Table 4: Metrics for residual distributions of models

As it can be seen, the mean and median for the first model are very high, while in the second they are not as small as they could be, but a very good results is achieved. As for the skewness, model 1 is heavily skewed towards the left, as the bigger differences in values are at the beginning, while the second model is also skewed to the left, its performance improving as time goes on.

To study uncertainty, the confidence bands for every model are shown in Figure 24. It can be seen that the second model has less uncertainty than the first one, which is the desired result.

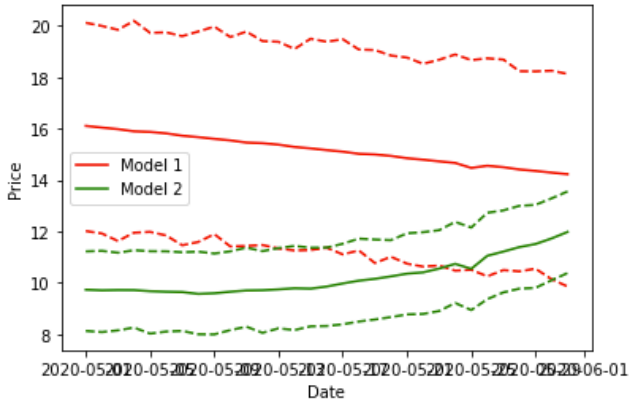

Figure 24: Confidence bands of models

All in all, there is no doubt that the second model greatly improves on the results from the first one. Figure 25 shows both models and the real values.

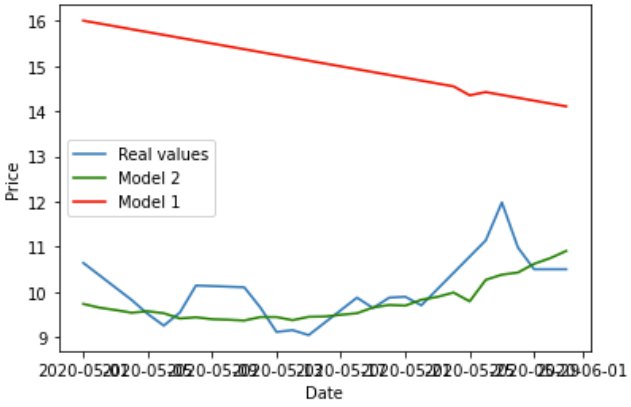

Figure 25: Performance of both inferred models

8. CONCLUSION

Preprocessing the data collected initially is vital to be able to use it in any capability, but understanding it is much more important. The data exploration is perhaps the most significant part of this project, since it shows how the data relates to itself and others. Only with this is it possible to make an informed prediction of the closing stock prices. This project shows that additional sources of data and variables are very important to the performance of an algorithm, and using these provides much better results. The most challenging part was perhaps matching the data from one source to another, as well as inferring the values of the variables for the month of May. Additional improvements could be presented in the form of more variables that could add to the performance of the algorithm, or fine tuning of the variables used. It is clear that, since the data for the variables itself is being inferred, the uncertainty of the algorithm is very high. Finding sources of data to help with this inference could improve the confidence in this data. For example, examine which parameters influence the price of oil and make an educated guess on what its price will be, which would later be introduced in the algorithm.

12. REFERENCES

[1] Adhikari, Ratnadip & Agrawal, R.. (2013). An Introductory Study on Time series Modeling and Forecasting. 10.13140/2.1.2771.8084.

[2] Maneenop, S., & Kotcharin, S. (2020). The impacts of COVID-19 on the global airline industry: An event study approach. Journal of air transport management, 89, 101920. https://doi.org/10.1016/j.jairtraman.2020.101920

[3] Cook, G. N. (1996). A Review of History, Structure, and Competition in the U.S. Airline Industry. Journal of Aviation/Aerospace Education & Research, 7(1).

[4] Camilleri, Mark. (2018). Aircraft Operating Costs and Profitability. 10.1007/978-3-319-49849-2_12.

[5] A. Marzuoli, P. Monmousseau and E. Feron, "Passenger-Centric Metrics for Air Transportation Leveraging Mobile Phone and Twitter Data," 2018 IEEE International Conference on Data Mining Workshops (ICDMW), Singapore, Singapore, 2018, pp. 588-595, doi: 10.1109/ICDMW.2018.00091.

[6] Uk.finance.yahoo.com. 2021. Yahoo! Finance. [online] Available at: <https://uk.finance.yahoo.com/> [Accessed 22 January 2021].

[7] Indexmundi.com. 2021. Jet Fuel - Daily Price - Commodity Prices - Price Charts, Data, And News - Indexmundi. [online] Available at: <https://www.indexmundi.com/commodities/?commodity=jet-fuel&months=60> [Accessed 22 January 2021].

[8] 2021. [online] Available at: <https://www.tsa.gov/coronavirus/passenger-throughput?page=1> [Accessed 22 January 2021].

[9] Transtats.bts.gov. 2021. Data Elements. [online] Available at: <https://www.transtats.bts.gov/Data_Elements.aspx?Data=1> [Accessed 22 January 2021].

[10] Gaudenzi, Barbara & Bucciol, Alessandro. (2016). Jet fuel price variations and market value: a focus on low-cost and regular airline companies. Journal of Business Economics and Management. 17. 977-991. 10.3846/16111699.2016.1209784.

[11] Friendly, M., & Denis, D. (2005). The early origins and development of the scatterplot. Journal of the history of the behavioral sciences, 41(2), 103–130. https://doi.org/10.1002/jhbs.20078

[12] Garcia Asuero, Agustin & Sayago, Ana & González, Gustavo. (2006). The Correlation Coefficient: An Overview. Critical Reviews in Analytical Chemistry - CRIT REV ANAL CHEM. 36. 41-59. 10.1080/10408340500526766.

[13] Hossin, Mohammad & M.N, Sulaiman. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. International Journal of Data Mining & Knowledge Management Process. 5. 01-11. 10.5121/ijdkp.2015.5201.

[14] Feng, C., Li, L. & Sadeghpour, A. A comparison of residual diagnosis tools for diagnosing regression models for count data. BMC Med Res Methodol 20, 175 (2020). https://doi.org/10.1186/s12874-020-01055-2