Student Name: Jesseca Johnson

Course: DSC 501: Introduction to Data Science

Assignment: Week 6 – Findings

Research Question and Variables

Research question: *How early was COVID-19 spreading in Île-de-France based on Google Trends data?*

The independent variable is *time* (specifically June 2019-February 2020).

The dependent variables are *Google Trends index scores* (discussed in Excel EDA paper).

Null hypothesis:

> *H_0 = There was no significant increase in Google searches for COVID-19 symptoms in late 2019 or the first two months of 2020.*

Alternative hypothesis:

> *H_a = There was a significant increase in Google searches for at least one COVID-19 symptom in late 2019 or the first two months of 2020.*

Name	Variable Type	Units	Source
Google Trends index score	Dependent	0-100 scale	Google Trends
Time	Independent	Days	Google Trends

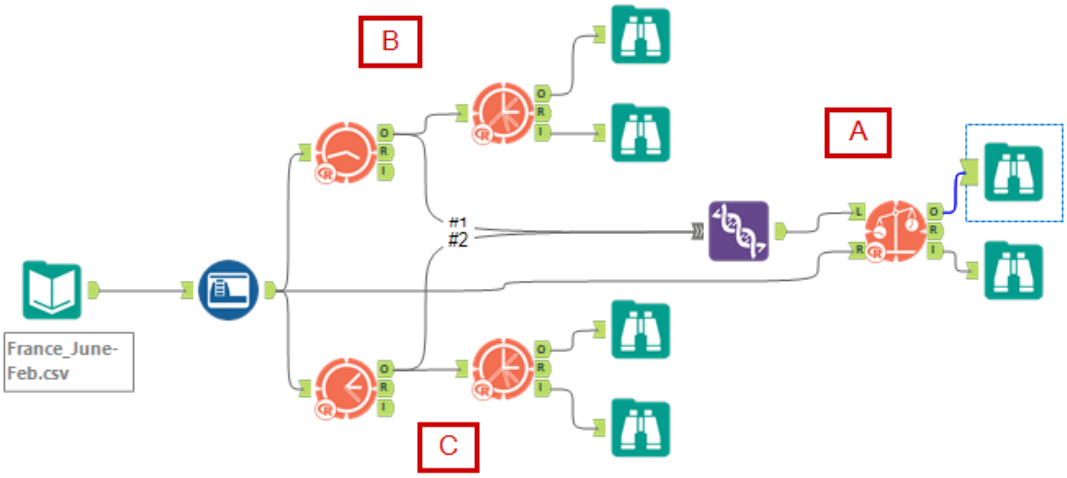
Project variables information

Updated Scope of Project

The scope of the project was updated to searches for "diarrhea" in Île-de-France on a daily rather than a weekly basis (274 total data points/periods).

Time Series Models

A time series comparison of the ARIMA and ETS models was created in Alteryx for daily Google Trends search index scores for "diarrhea" in Île-de-France between June 1, 2019 and February 28, 2020 (Figure 1). The accuracy statistics for each model were virtually identical, which was expected due to the data being univariate. The ARIMA method was used, although similar results could be expected by using the ETS model.



Results - Browse (10) - Input

8 of 8 Fields | Cell Viewer | 2 records displayed, 1,666 bytes

Record	Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
1	ARIMA1	0	23.9249	20.59	-1.#INF	1.#INF	0.882	[Null]
2	ETS	0.0058	23.9249	20.5886	-1.#INF	1.#INF	0.8819	[Null]

Figure 1: Time series comparison (area A) between ARIMA (area B) and ETS (area C) models for Google Trends index scores. The table under "Results" indicates the accuracy statistics for each model.

Hyndman explains forecasting accuracy statistics in his 2006 article “Another Look at Forecast-Accuracy Metrics for Intermittent Demand” and recommends the mean absolute scaled error (MASE) as the best accuracy metric in general because it is without a scale. The mean error (ME), root mean square error (RMSE), and mean absolute error (MAE) are all scaled to the data, so it would be impossible to compare the Google Trends data with the sample data introduced on page 3. The mean percentage error (MPE) and mean absolute percentage error (MAPE) are meaningless when the dataset contains values of zero as in this case. Because both MASE scores are less than 1, they appear to be “better forecasts than the average one-step, naïve forecast computed in-sample” (Hyndman, 2006). The naïve method “simply states that we forecast that this period will be the same as the previous period” (Avercast, n. d.).

The ARIMA model forecast produced a flat fitted average of a 16.44 index score (Figure 2).

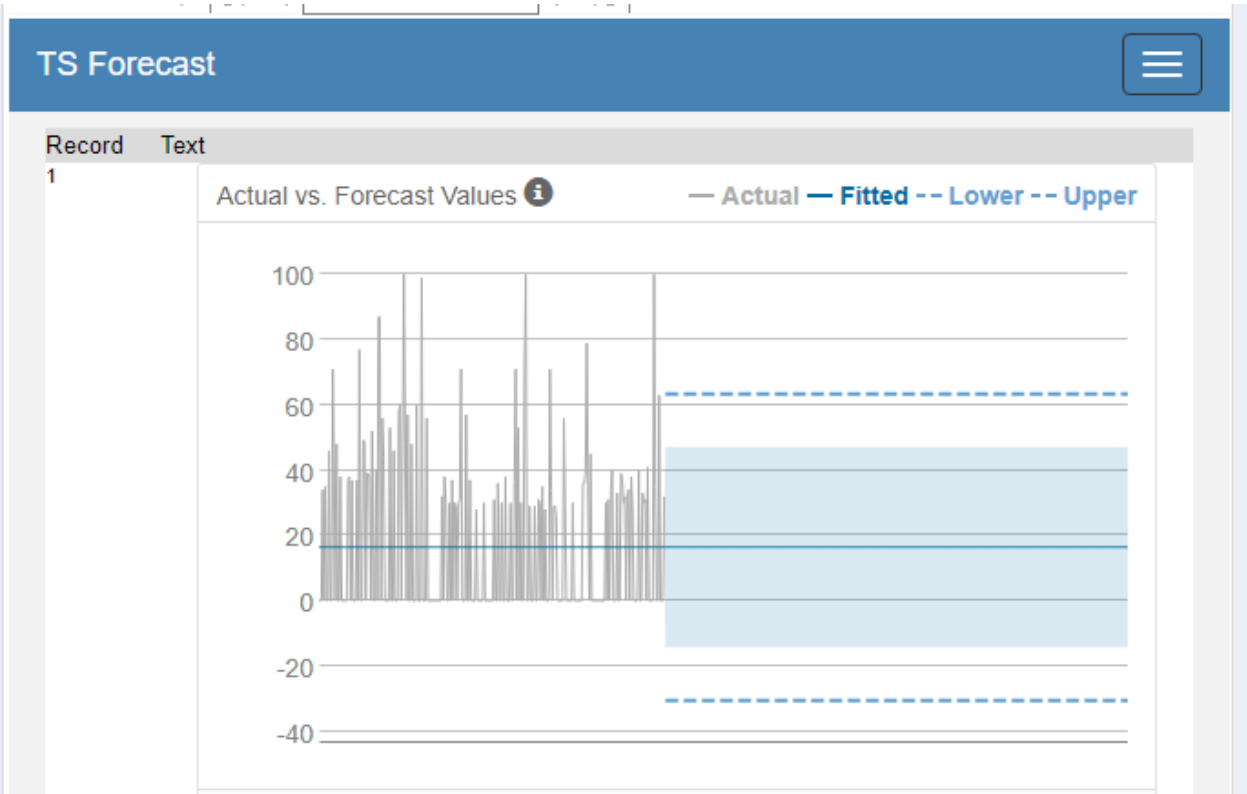
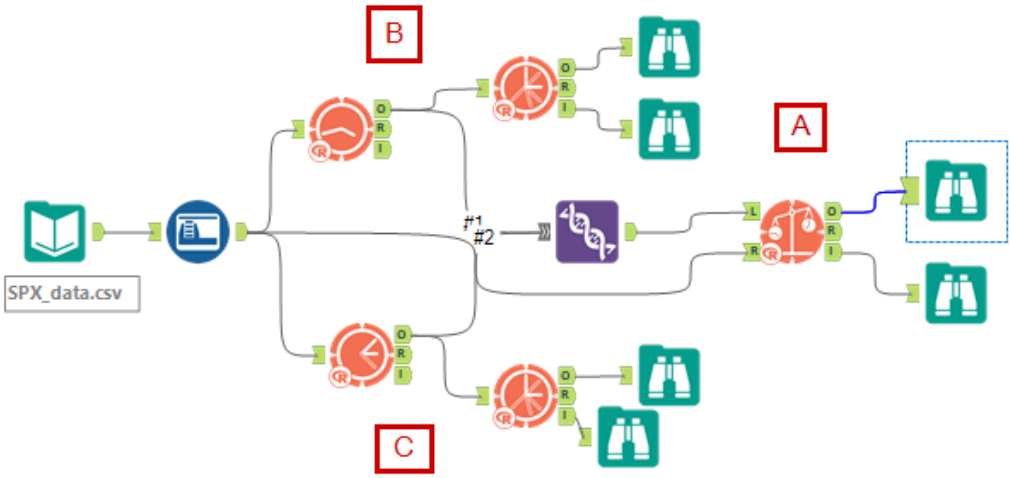


Figure 2: Alteryx ARIMA Google Trends index score for "diarrhea" in Île-de-France from June 2019 to February 2020. The gray area represents actual vs. expected scores and the blue area represents future scores for which data was not provided. The blue line represents the predicted value across the time period of 16.44.

It is notable that the absolute maximum value of 100 was reached three separate times between June 2019 and February 2020 – as well as a fourth instance of 99 – despite a predicted value of 16.44. But that the model was only able to produce a single fixed average across the entire time period, and that the scores were so variable with such a large range (0-100), suggests that the values were unpredictable, and that the average may be essentially meaningless.

Comparison Data

Sample S&P500 (SPX) closing values from April-December 2019 (MarketWatch, 2021) (231 total data points/periods) was used as a baseline comparison against the Google Trends data; an exact replica of the Google Trends data model was applied to the SPX data. Like the Google Trends data, because the sample is univariate, the accuracy statistics for ARIMA and ETS are almost identical (Figure 3). The MASE scores for these SPX data indicate that the models were less accurate than the naïve method would have been.



Results - Browse (7) - Input

8 of 8 Fields | Cell Viewer | 2 records displayed, 1,694 bytes | Search

Record #1, Field Model (V_WString)

ARIMA_SPX

Record	Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
1	ARIMA_SPX	73.039	171.4899	125.8126	2.1639	4.0105	7.0356	[Null]
2	ETS_SPX	73.0365	171.4888	125.8122	2.1639	4.0104	7.0356	[Null]

Figure 3: Time series comparison (area A) between ARIMA (area B) and ETS (area C) models for S&P500 closing values. The table under "Results" indicates the accuracy statistics for each model.

However, unlike the Google Trends index data, the ARIMA model for SPX produced a fitted line with separate predictions for each data point, suggesting a predictable pattern (Figure 4).

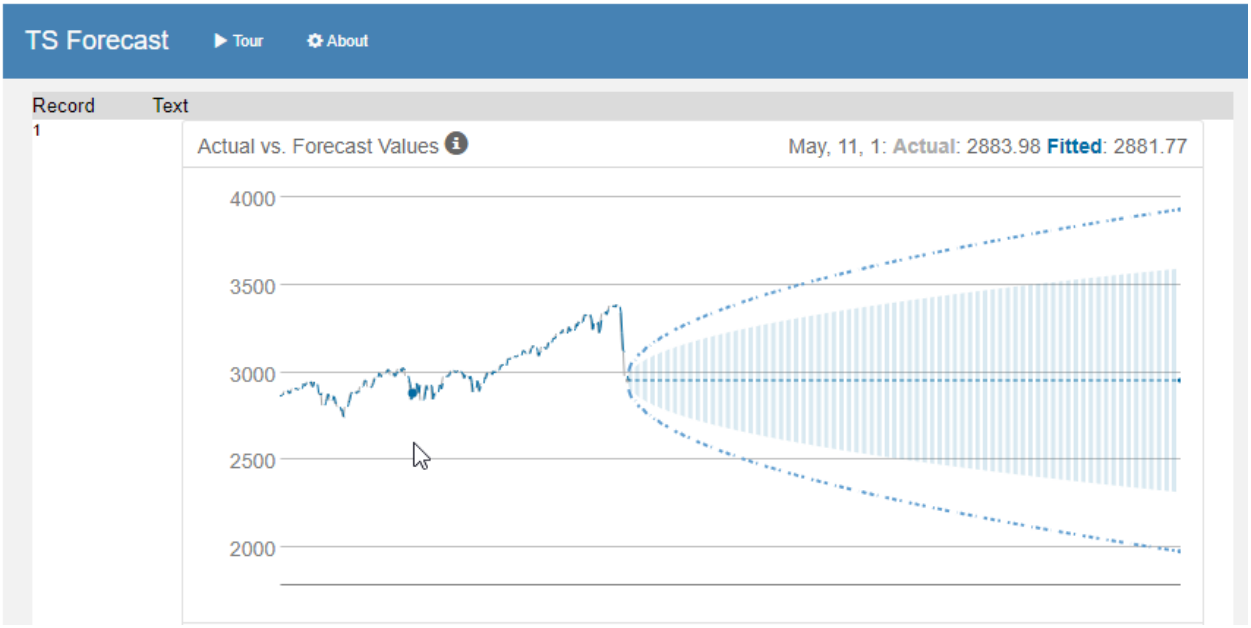


Figure 4: ARIMA model for SPX data, whose predictions appear to be very close to the actual values. Note that the timeline (x-axis) is incorrect due to the inability to specify a custom data starting point in Alteryx when using daily data; the actual date of the selected point on the curve is approximately October 11 rather than May 11.

Challenges of Alteryx

Alteryx presents some challenges on a basic level. First, its recognition of the MM-DD-YYYY format as a date is inconsistent; this pattern was used in both sets of data, but Alteryx read the SPX dates as strings until they were changed to the YYYY-MM-DD format on the CSV. Second, the maximum number of periods in the forecast is limited to 100. Third, Alteryx does not allow the user to specify a series starting period when using days – only when using weeks, months, quarters, or years. The result is that although the data starts on either June 1, 2019 (Google Trends data) or April 1, 2019 (SPX data), the first data point is labeled as January 1, 2019 in Alteryx (Figure 4). Fourth, the SPX MASE score of >7 does not appear to match the reality of the actual ARIMA forecast, which seems to be quite accurate.

Type I and Type II Errors and Conclusion

A Type I error would occur if the null hypothesis were rejected when it was in fact true: in other words, a false positive. A Type II error would occur if the null hypothesis were accepted when it was in fact false: in other words, a false negative (Statistics Solutions, n. d.). In this case, despite some interesting data, there is insufficient evidence to reject the null hypothesis and it cannot be concluded that there was a significant increase in Google searches for at least one COVID-19 symptom in late 2019 or the first two months of 2020 based on this specific data. It is possible that a different conclusion could be reached using more data going further back in time. It could perhaps also be useful to look at

previous data to see whether the maximum score of 100 had been reached prior to the start of this data. If not, that might suggest that by reaching the same maximum at some points in the months prior to the known start of the pandemic (August 2019-February 2020) as what was reached during the known pandemic period (beginning March 2020) when that maximum had not been reached in, say, 2015-2018, COVID-19 was spreading earlier than initially thought.

References

Avercast. (n. d.). *Naïve Forecasting*. Retrieved June 20, 2021 from https://www.avercast.com/post/naive-forecasting

Hyndman, R. J. (2006, June). *Another Look at Forecast-Accuracy Metrics for Intermittent Demand*. Retrieved June 19, 2021 from https://robjhyndman.com/papers/foresight.pdf

MarketWatch. (2021, June 18). *Historical Quotes*. Retrieved June 19, 2021 from https://www.marketwatch.com/investing/index/spx/download-data

Statistics Solutions. (n. d.). *To Err is Human: What are Type I and II Errors*? Retrieved June 20, 2021 from https://www.statisticssolutions.com/to-err-is-human-what-are-type-i-and-ii-errors/