Student Name: Jesseca Johnson

Course: DSC 501: Introduction to Data Science

Assignment: Week 7 – 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 January 1, 2019 – March 31, 2020).

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

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 and the timeline was updated to January 1, 2019 – March 31, 2020 (456 total data points/periods).

Time Series Models

Because the data were so erratic, with 324 out of 456 instances being zero (Figure 1), Alteryx was not able to create an actual model for this data using ARIMA or ETS. ARIMA produced a fixed average of 13.06 (Figure 2); ETS produced a nearly fixed average of between 13.01 and 13.1 (Figure 3).

Record	date	score
350	2019-12-16	0
351	2019-12-17	0
352	2019-12-18	0
353	2019-12-19	0
354	2019-12-20	0
355	2019-12-21	0
356	2019-12-22	0
357	2019-12-23	0
358	2019-12-24	0
359	2019-12-25	0
360	2019-12-26	26
361	2019-12-27	0
362	2019-12-28	0
363	2019-12-29	0
364	2019-12-30	0
365	2019-12-31	0
366	2020-01-01	64
367	2020-01-02	51

Figure 1: Majority of Google index scores in the last part of 2019 were zero; this is typical for the whole data set

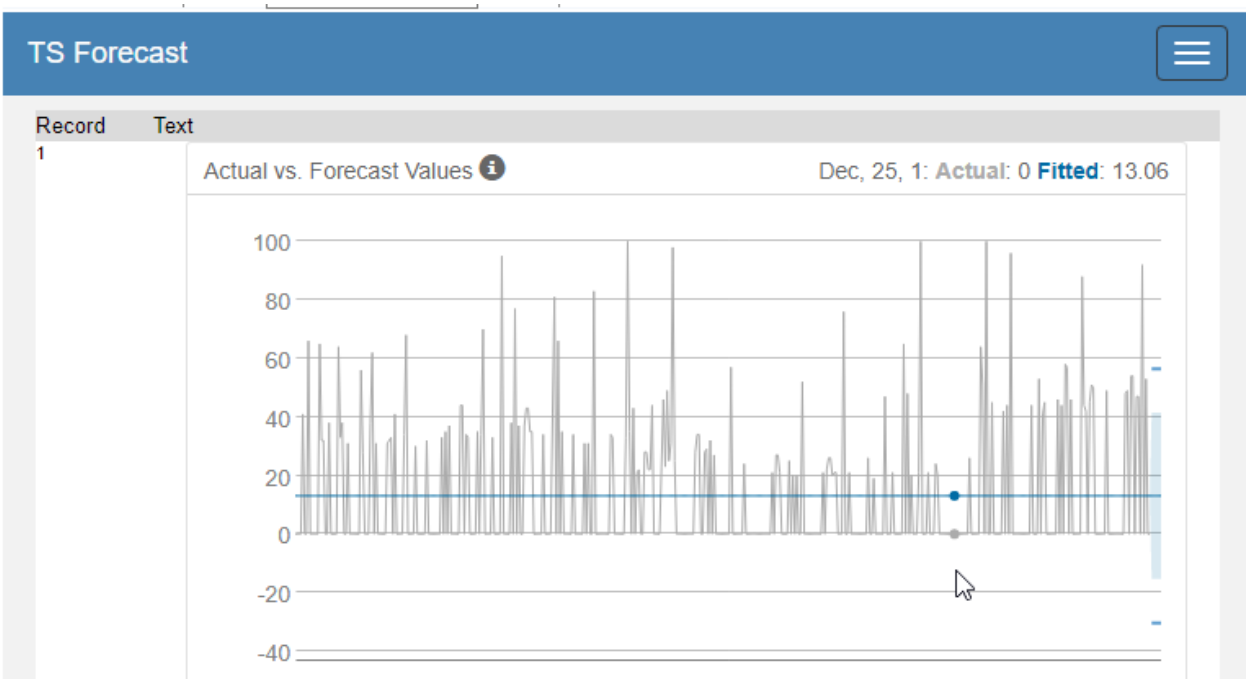
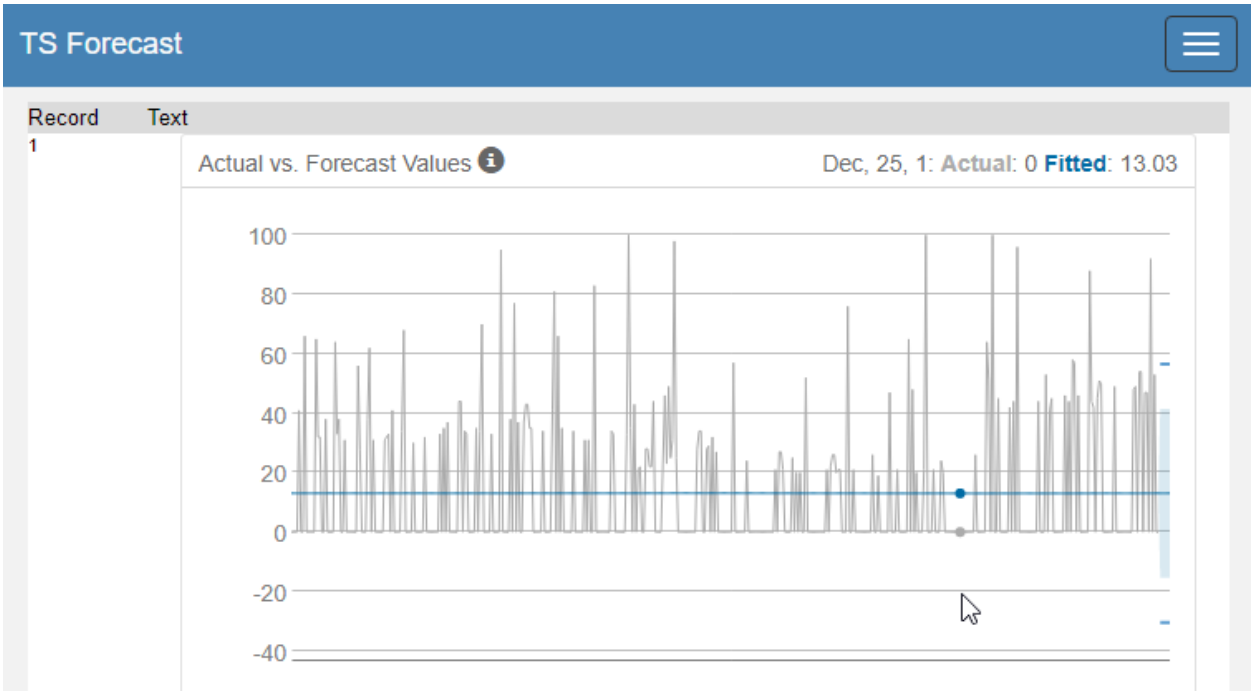
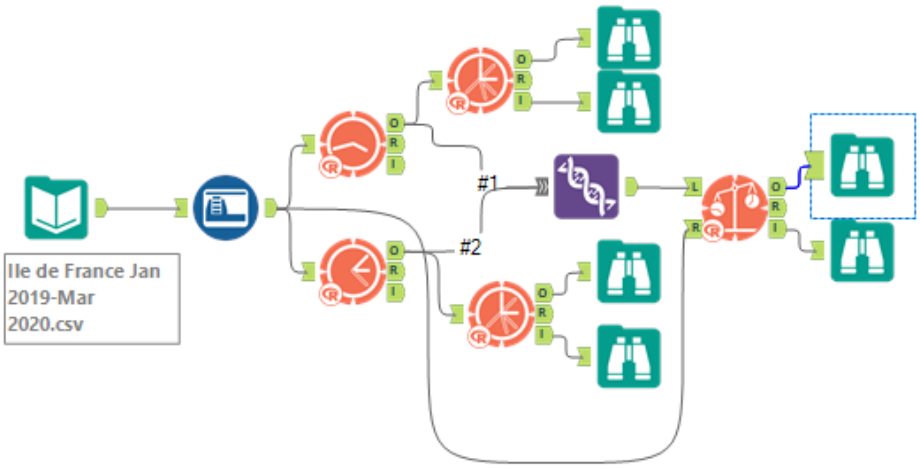


Figure 2: ARIMA forecast model with fixed average of 13.06. The x-axis represents the time period of January 1, 2019 through March 31, 2020



3: ETS forecast model with average between 13.01 and 13.1. The x-axis represents the time period of January 1, 2019 through March 31, 2020

ARIMA and ETS accuracy were virtually the same (Figure 4). 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. 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) – but at .97, just barely. The naïve method "simply states that we forecast that this period will be the same as the previous period" (Avercast, n. d.). The root mean square error (RMSE) is the square root of the square of the residuals (errors in prediction) (RMSE: Root Mean Square Error, n. d.). Given that the range of scores was 0-100, the RMSE of 22.1407 for both models seems extremely high, although there is no fixed rule to which we can refer. In choosing one or the other, ARIMA would make more sense at it is both slightly more accurate and the most commonly used forecasting tool (*ARIMA Tool*, 2021).



Results - Browse (8) - Input

8 of 8 Fields | Cell Viewer | 2 records displayed, 1,653 bytes | Search

Record	Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
1	ARIMA	0	22.1407	17.9308	-1.#INF	1.#INF	0.9706	[Null]
2	ETS	-0.0014	22.1407	17.9313	-1.#INF	1.#INF	0.9706	[Null]

Figure 4: Time series comparison (ARIMA vs. ETS)

Conclusion: Challenges of Using Google Trends in Research

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 2019 or the first three months of 2020 based on this specific data.

Although initially promising, the use of big data in scientific research has been proven fraught with peril. In The Parable of Google Flu: Traps in Big Data Analysis (2014), Lazer et al. describe how the

original Google Trends algorithm completely missed the nonseasonal 2009 influenza A–H1N1 pandemic in 2009; after updating it that same year, it severely overestimated flu cases until it was eventually retired in 2015 (O'Connor, 2015). Lazer et al. also point out that Google Trends data in health care is not completely without use as "greater value can be obtained by combining [Google Flu Trends] with other near–real time health data." In the future, studies such as this one should make use of multiple search queries along with other data.

Works Cited

ARIMA Tool. (2021, February 22). Alteryx Documentation. Retrieved June 26, 2021 from https://help.alteryx.com/current/designer/arima-tool

Lazer, D., R. Kennedy, G. King, and A. Vespignani. 2014. The Parable of Google Flu: Traps in Big Data Analysis. *Science 343*(6176) (March 14): 1203–1205. 10.1126/science.1248506

O'Conner, Fred. (2015, August 20). *Google Flu Trends calls out sick, indefinitely*. PCWorld. Retrieved June 27, 2021 from https://www.pcworld.com/article/2974153/google-flu-trends-calls-out-sick-indefinitely.html

RMSE: Root Mean Square Error. (n. d.). Statistics How-To. Retrieved June 26, 2021 from https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/

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/