Student Name: Jesseca Johnson

Course: DSC 501: Introduction to Data Science

Dataset Descriptive Statistics with Alteryx

Research Question and Variables

Research question: *How early was COVID-19 spreading in three different locations based on Google Trends data?*

The independent variable is *time* (specifically September 2018-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	Weeks	Google Trends

Project variables information

Updated Scope of Project

The scope of the project was updated to include only searches for the topic "diarrhea" since it yielded the most data by a large margin.

First Confirmed Cases of COVID-19 in Relevant Locations

United States

The first confirmed case of COVID-19 in Seattle (as well as the United States overall) was identified on January 21, 2020 (CDC, 2020), but it is possible that it was already spreading before that time due to the discovery of antibodies present in two patients who were sick in December (Kamb, 2020).

France

The first confirmed cases of COVID-19 in France (as well as in Europe overall) – two in Paris and one in Bourdeaux – were identified on January 24, 2020 (Stoecklin et al., 2020). However, a subsequent test found a patient who was positive for COVID-19 in late December 2019 (Deslandes et al., 2020), and there is some evidence the virus may have already been spreading in November 2019 (Carrat et al., 2021).

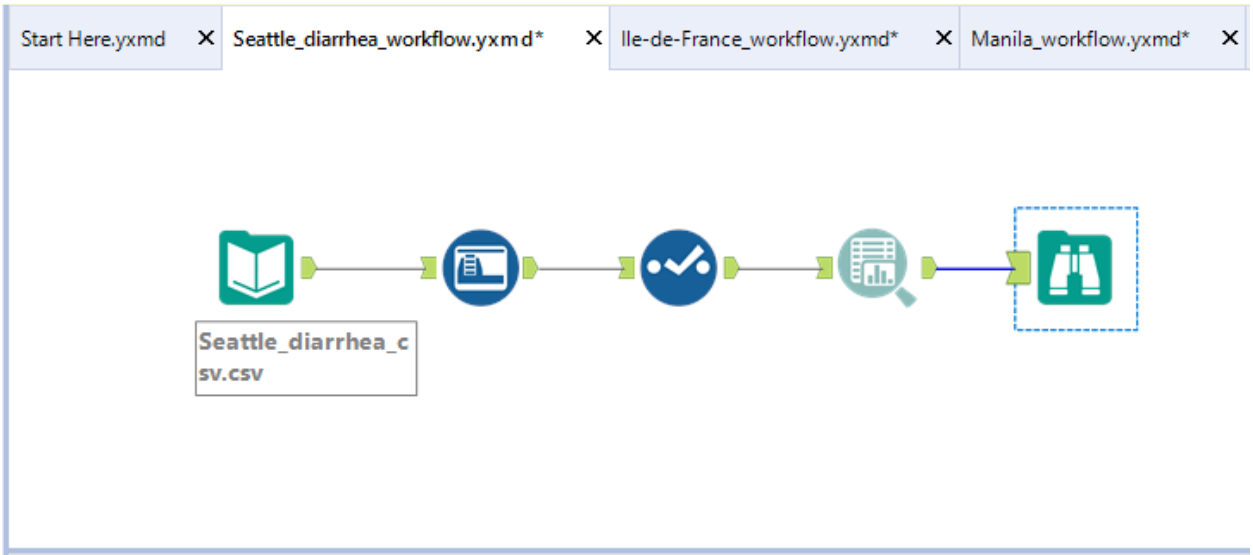
Philippines

The first case of COVID-19 in the Philippines was confirmed on January 22, 2020 in Manila; 633 cases were suspected by March 1, 2020 (Edrade et al., 2020).

Descriptive Statistics Using Alteryx

All basic descriptive statistics in Alteryx match those found in Excel as expected. They are presented in Tables 1-6.

Table 1: Seattle-Tacoma Google search indices for "diarrhea," 2018-19 with Alteryx workflow

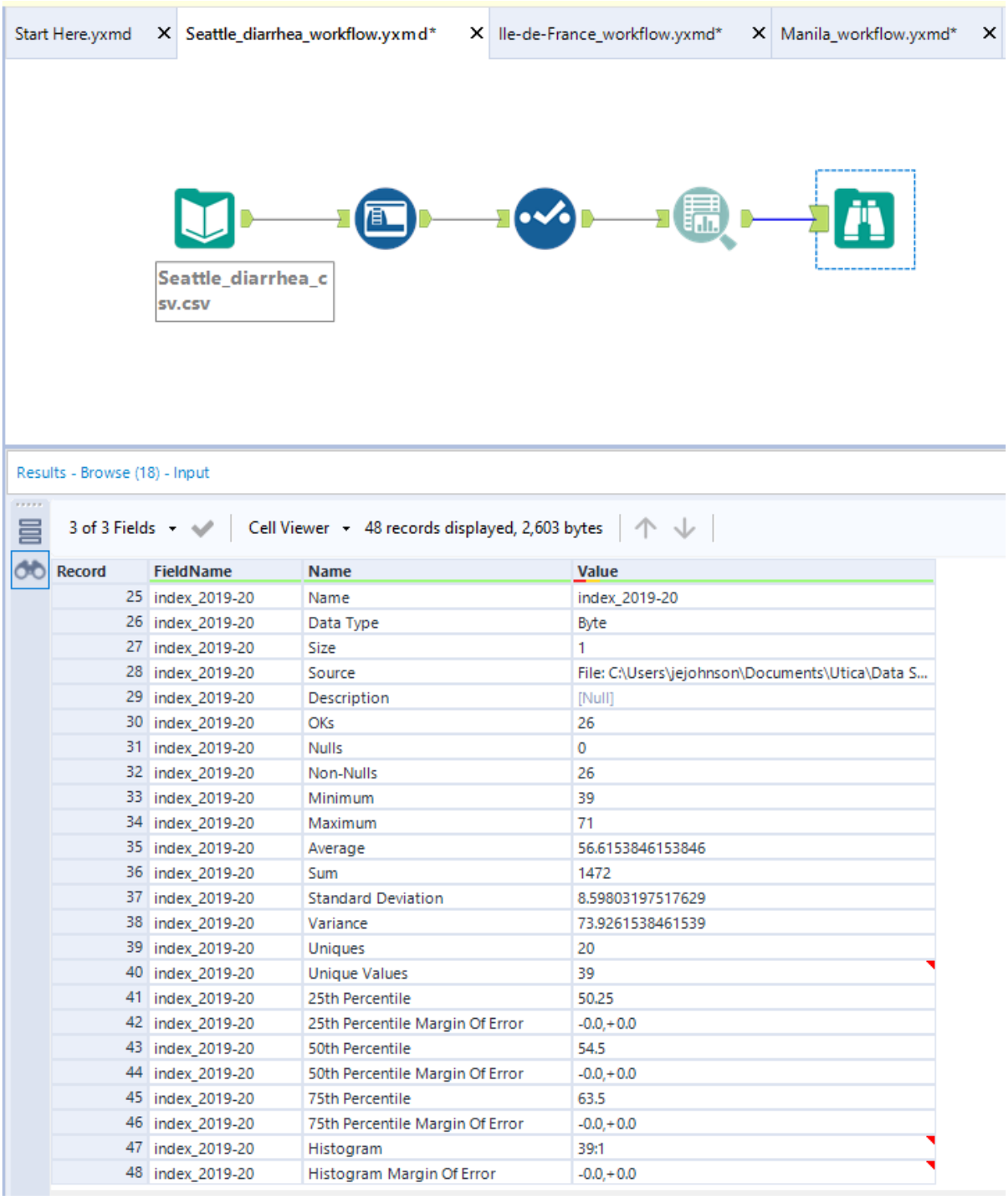


Results - Browse (18) - Input

3 of 3 Fields | Cell Viewer | 48 records displayed, 2,603 bytes

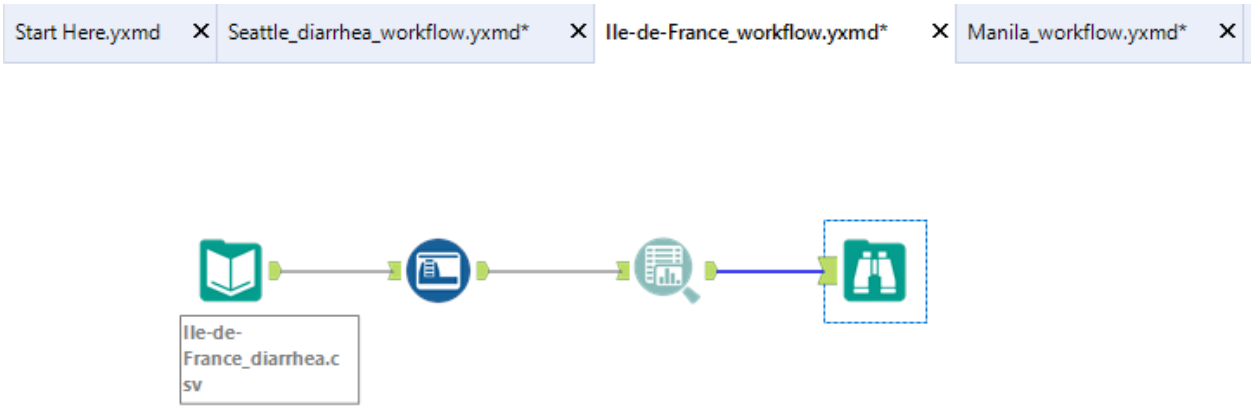
Record	FieldName	Name	Value
1	index_2018-19	Name	index_2018-19
2	index_2018-19	Data Type	Byte
3	index_2018-19	Size	1
4	index_2018-19	Source	File: C:\Users\jejohnson\Documents\Utica\Data S...
5	index_2018-19	Description	[Null]
6	index_2018-19	OKs	26
7	index_2018-19	Nulls	0
8	index_2018-19	Non-Nulls	26
9	index_2018-19	Minimum	34
10	index_2018-19	Maximum	83
11	index_2018-19	Average	52
12	index_2018-19	Sum	1352
13	index_2018-19	Standard Deviation	10.9105453575887
14	index_2018-19	Variance	119.04
15	index_2018-19	Uniques	19
16	index_2018-19	Unique Values	34
17	index_2018-19	25th Percentile	46
18	index_2018-19	25th Percentile Margin Of Error	-0.0,+0.0
19	index_2018-19	50th Percentile	50.5
20	index_2018-19	50th Percentile Margin Of Error	-0.0,+0.0
21	index_2018-19	75th Percentile	55.75
22	index_2018-19	75th Percentile Margin Of Error	-0.0,+0.0
23	index_2018-19	Histogram	34:3
24	index_2018-19	Histogram Margin Of Error	-0.0,+0.0

Table 2: Seattle-Tacoma Google search indices for "diarrhea," 2019-20 with Alteryx workflow



Record	FieldName	Name	Value
25	index_2019-20	Name	index_2019-20
26	index_2019-20	Data Type	Byte
27	index_2019-20	Size	1
28	index_2019-20	Source	File: C:\Users\jejohnson\Documents\Utica\Data S...
29	index_2019-20	Description	[Null]
30	index_2019-20	OKs	26
31	index_2019-20	Nulls	0
32	index_2019-20	Non-Nulls	26
33	index_2019-20	Minimum	39
34	index_2019-20	Maximum	71
35	index_2019-20	Average	56.6153846153846
36	index_2019-20	Sum	1472
37	index_2019-20	Standard Deviation	8.59803197517629
38	index_2019-20	Variance	73.9261538461539
39	index_2019-20	Uniques	20
40	index_2019-20	Unique Values	39
41	index_2019-20	25th Percentile	50.25
42	index_2019-20	25th Percentile Margin Of Error	-0.0,+0.0
43	index_2019-20	50th Percentile	54.5
44	index_2019-20	50th Percentile Margin Of Error	-0.0,+0.0
45	index_2019-20	75th Percentile	63.5
46	index_2019-20	75th Percentile Margin Of Error	-0.0,+0.0
47	index_2019-20	Histogram	39:1
48	index_2019-20	Histogram Margin Of Error	-0.0,+0.0

Table 3: Île-de-France Google search indices for "diarrhea," 2018-19 with Alteryx workflow

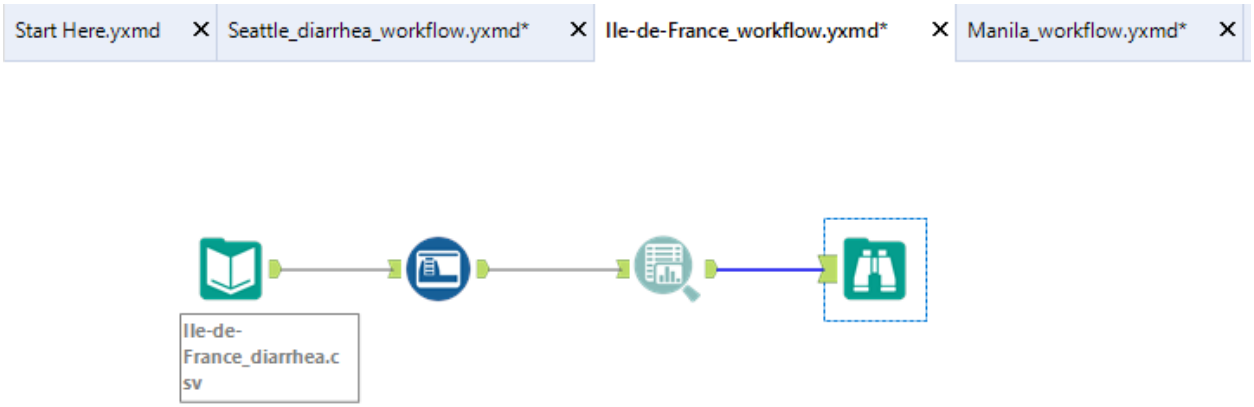


Results - Browse (4) - Input

3 of 3 Fields | Cell Viewer | 114 records displayed, 4,019 bytes

Record	FieldName	Name	Value
67	index_2018-19	Name	index_2018-19
68	index_2018-19	Data Type	Byte
69	index_2018-19	Size	1
70	index_2018-19	Source	File: C:\Users\jejohnson\Documents\Utica\Data S...
71	index_2018-19	Description	[Null]
72	index_2018-19	OKs	26
73	index_2018-19	Nulls	0
74	index_2018-19	Non-Nulls	26
75	index_2018-19	Minimum	43
76	index_2018-19	Maximum	65
77	index_2018-19	Average	54.1538461538462
78	index_2018-19	Sum	1408
79	index_2018-19	Standard Deviation	5.64051279720068
80	index_2018-19	Variance	31.8153846153846
81	index_2018-19	Uniques	17
82	index_2018-19	Unique Values	43
83	index_2018-19	25th Percentile	50.25
84	index_2018-19	25th Percentile Margin Of Error	-0.0,+0.0
85	index_2018-19	50th Percentile	54.5
86	index_2018-19	50th Percentile Margin Of Error	-0.0,+0.0
87	index_2018-19	75th Percentile	57.75
88	index_2018-19	75th Percentile Margin Of Error	-0.0,+0.0
89	index_2018-19	Histogram	43:3
90	index_2018-19	Histogram Margin Of Error	-0.0,+0.0

Table 4: Île-de-France Google search indices for "diarrhea," 2019-20 with Alteryx workflow

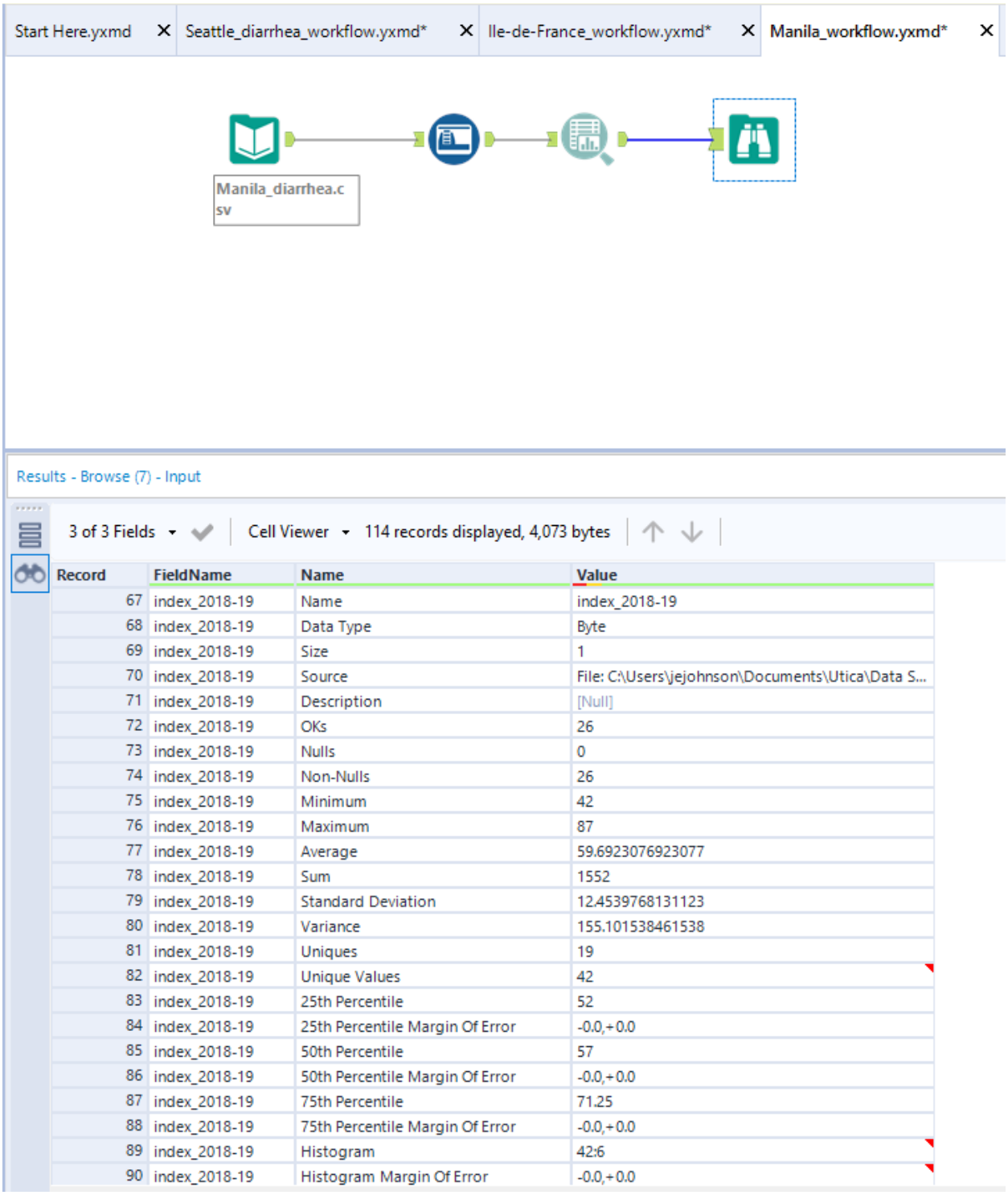


Results - Browse (4) - Input

3 of 3 Fields | Cell Viewer | 114 records displayed, 4,019 bytes

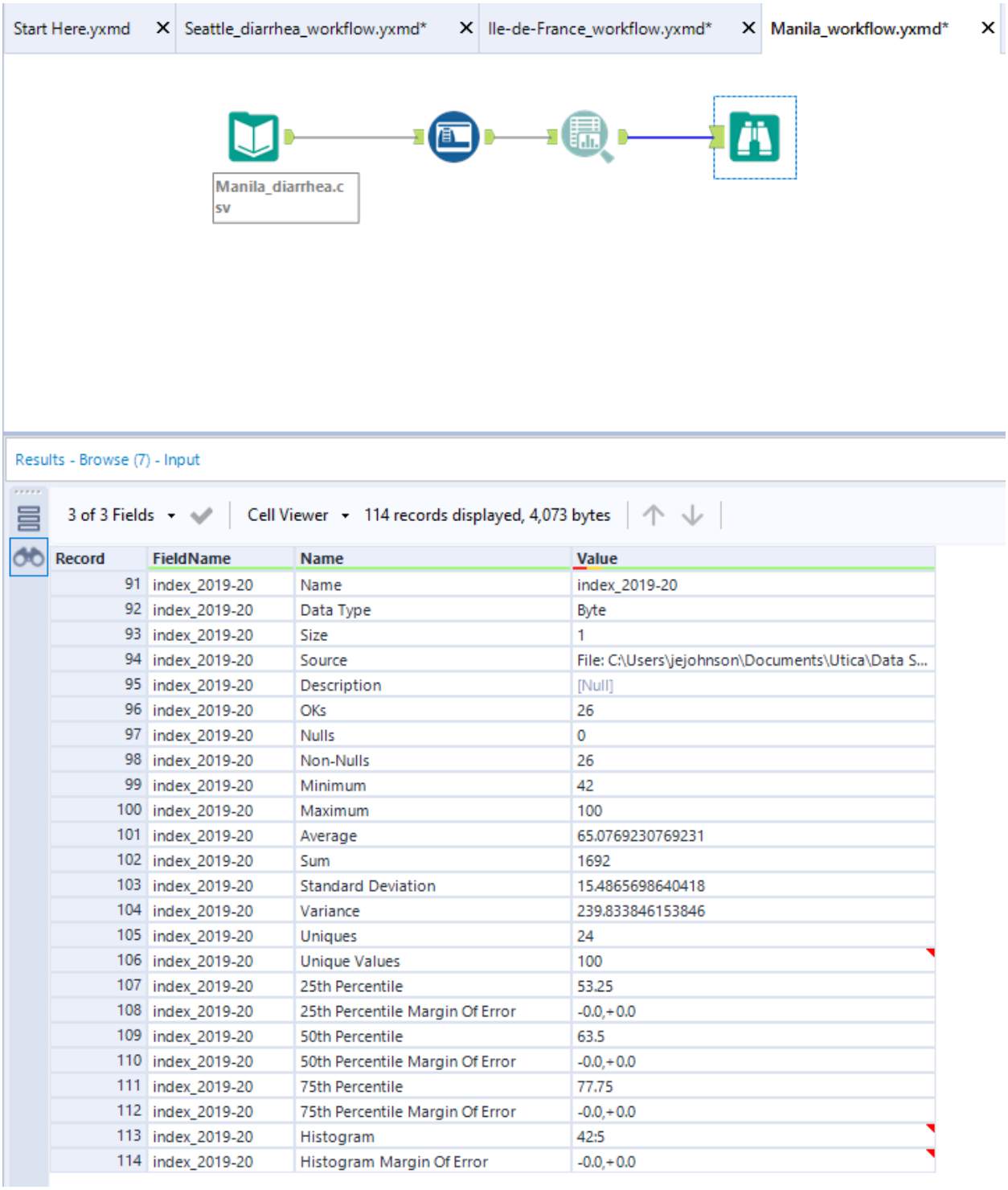
Record	FieldName	Name	Value
91	index_2019-20	Name	index_2019-20
92	index_2019-20	Data Type	Byte
93	index_2019-20	Size	1
94	index_2019-20	Source	File: C:\Users\jejohnson\Documents\Utica\Data S...
95	index_2019-20	Description	[Null]
96	index_2019-20	OKs	26
97	index_2019-20	Nulls	0
98	index_2019-20	Non-Nulls	26
99	index_2019-20	Minimum	39
100	index_2019-20	Maximum	91
101	index_2019-20	Average	52
102	index_2019-20	Sum	1352
103	index_2019-20	Standard Deviation	12.0432553738597
104	index_2019-20	Variance	145.04
105	index_2019-20	Uniques	18
106	index_2019-20	Unique Values	39
107	index_2019-20	25th Percentile	44.25
108	index_2019-20	25th Percentile Margin Of Error	-0.0,+0.0
109	index_2019-20	50th Percentile	48
110	index_2019-20	50th Percentile Margin Of Error	-0.0,+0.0
111	index_2019-20	75th Percentile	54
112	index_2019-20	75th Percentile Margin Of Error	-0.0,+0.0
113	index_2019-20	Histogram	39:13
114	index_2019-20	Histogram Margin Of Error	-0.0,+0.0

Table 5:Metro Manila Google search indices for "diarrhea," 2018-19 with Alteryx workflow



Record	FieldName	Name	Value
67	index_2018-19	Name	index_2018-19
68	index_2018-19	Data Type	Byte
69	index_2018-19	Size	1
70	index_2018-19	Source	File: C:\Users\jejohnson\Documents\Utica\Data S...
71	index_2018-19	Description	[Null]
72	index_2018-19	OKs	26
73	index_2018-19	Nulls	0
74	index_2018-19	Non-Nulls	26
75	index_2018-19	Minimum	42
76	index_2018-19	Maximum	87
77	index_2018-19	Average	59.6923076923077
78	index_2018-19	Sum	1552
79	index_2018-19	Standard Deviation	12.4539768131123
80	index_2018-19	Variance	155.101538461538
81	index_2018-19	Uniques	19
82	index_2018-19	Unique Values	42
83	index_2018-19	25th Percentile	52
84	index_2018-19	25th Percentile Margin Of Error	-0.0,+0.0
85	index_2018-19	50th Percentile	57
86	index_2018-19	50th Percentile Margin Of Error	-0.0,+0.0
87	index_2018-19	75th Percentile	71.25
88	index_2018-19	75th Percentile Margin Of Error	-0.0,+0.0
89	index_2018-19	Histogram	42:6
90	index_2018-19	Histogram Margin Of Error	-0.0,+0.0

Table 6: Metro Manila Google search indices for "diarrhea," 2019-20 with Alteryx workflow



Record	FieldName	Name	Value
91	index_2019-20	Name	index_2019-20
92	index_2019-20	Data Type	Byte
93	index_2019-20	Size	1
94	index_2019-20	Source	File: C:\Users\jejohnson\Documents\Utica\Data S...
95	index_2019-20	Description	[Null]
96	index_2019-20	OKs	26
97	index_2019-20	Nulls	0
98	index_2019-20	Non-Nulls	26
99	index_2019-20	Minimum	42
100	index_2019-20	Maximum	100
101	index_2019-20	Average	65.0769230769231
102	index_2019-20	Sum	1692
103	index_2019-20	Standard Deviation	15.4865698640418
104	index_2019-20	Variance	239.833846153846
105	index_2019-20	Uniques	24
106	index_2019-20	Unique Values	100
107	index_2019-20	25th Percentile	53.25
108	index_2019-20	25th Percentile Margin Of Error	-0.0,+0.0
109	index_2019-20	50th Percentile	63.5
110	index_2019-20	50th Percentile Margin Of Error	-0.0,+0.0
111	index_2019-20	75th Percentile	77.75
112	index_2019-20	75th Percentile Margin Of Error	-0.0,+0.0
113	index_2019-20	Histogram	42:5
114	index_2019-20	Histogram Margin Of Error	-0.0,+0.0

ARIMA Findings

ARIMA stands for Autoregressive Integrated Moving Average. Its primary purpose is forecasting (EricWe, 2020). Several options exist to customize each ARIMA model (start date, time intervals, etc.). The same ARIMA workflow template was applied to all data sets, with one ARIMA icon for the 2018-19 data and one for the 2019-20 data:

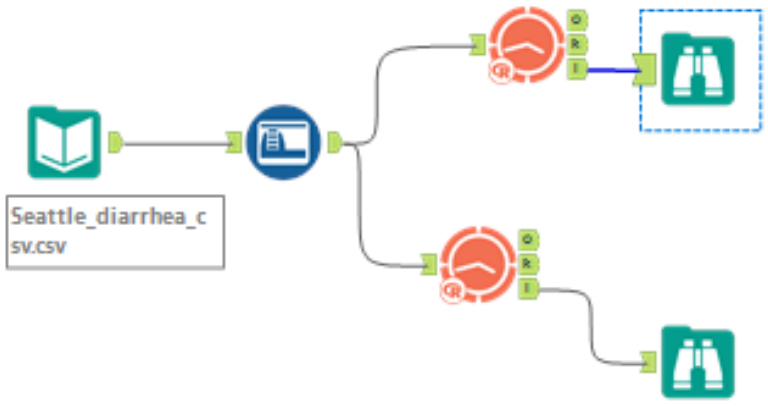


Figure 1: Alteryx ARIMA workflow applied to all data sets

The ARIMA tool has three anchors – O, R, and I – that appear in Figure 1 above on the right side of both orange ARIMA tool (clock) icons. Boopalan (2019) explains the uses of each. While the O anchor is intended to show the ARIMA model object to be used in future forecasts, it did not appear to produce any meaningful information for this study. The R anchor provides "a statistical summary, autocorrelation diagnostic plots and forecast plots" (Boopalan, 2019) partly based on the Ljung-Box Test, including the *p* value (provided for each data set in Table 7 below). The I anchor produces an interactive dashboard which includes forecasts and the differentials between expected and actual values. The ARIMA I dashboard for each data set has been provided in Figures 2-7.

In conclusion, judging by the *p* values produced by the ARIMA models all being greater than .05, it appears that all of the 2019-20 data can be explained by random chance rather than statistical significance. The fact that two out of the three 2018-19 data sets were also unpredictable lends itself to this doubt as well. It is interesting to note the differences between the model's expected values and the actual values (for examples, in Figure 5, where the model's prediction for the beginning of January 2020

was approximately 66, but the actual value was approximately 90. It may have been useful to have gone further back in time for comparison.

Table 7: p values for each relevant data set

Location	Data Set	*p* value
Seattle	2018-19	0.52364
	2019-20	0.624398
Île-de-France	2018-19	0.046089
	2019-20	0.818548
Manila	2018-19	0.214129
	2019-20	0.672012

ARIMA Interactive Dashboards

Seattle

Figure 1: ARIMA model for Seattle Google searches for "diarrhea," 2018-19

Figure 2:ARIMA model for Seattle Google searches for "diarrhea," 2019-20

Île-de-France

Figure 3: ARIMA model for Île-de-France Google searches for "diarrhea," 2018-19

Figure 4: ARIMA model for Île-de-France Google searches for "diarrhea," 2019-20

Manila

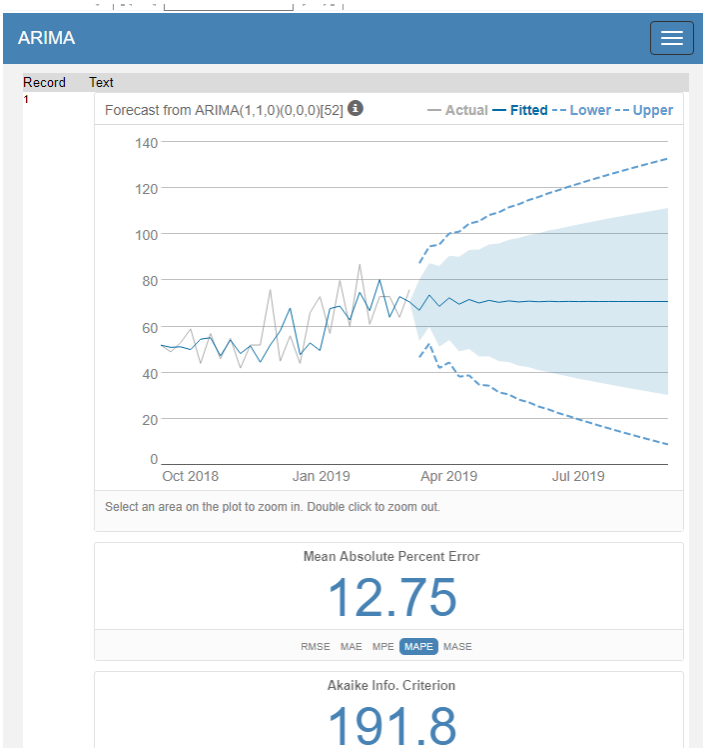


Figure 5:ARIMA model for metro Manila Google searches for "diarrhea," 2018-19

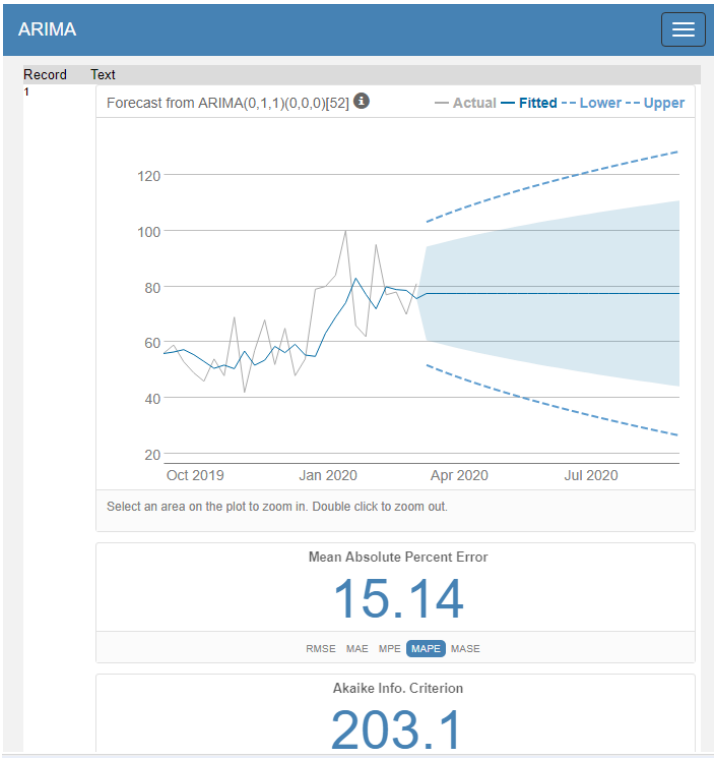


Figure 6: ARIMA model for metro Manila Google searches for "diarrhea," 2019-20

Alteryx vs. Excel

For the purpose of this project, Excel was a great choice for initial exploratory data analysis. The ability to see descriptive statistics and line chart comparisons almost effortlessly was certainly helpful in gleaning an initial, if superficial, understanding the data.

Although the classic Excel has been the premiere all-purpose business data apparatus since the mid-1990s and does have some helpful data functions, it is no match for today's big data needs with billions of rows and the ubiquitous desire for ever-more complex modeling as data becomes the primary driver for more, and more important, business decisions. Enter Alteryx, the data analysis powerhouse first released in 2005 (SRC, n. d.). There is a learning curve with Alteryx that is likely not as drastic as with Excel – if only because Excel has been so prominent for so many years – and some puzzling quirks, such as the need to use the "Auto field" tool to transform data types into numerical form, even when the data is already specified as numeric in Excel. But Alteryx is the clear winner in sophistication. The ARIMA models used for this project took a matter of minutes to set up; it provided many statistics and forecasting, and it is but one of many tools available on the platform. It seems inevitable that Alteryx or

equivalent software will be seen as essential as more businesses attempt to utilize their data to their advantage in more sophisticated ways. Table 8 below lists comparisons between the two applications.

Table 8: Excel vs. Alteryx comparison

Function	Excel	Alteryx
Ease of use	• Market dominance for over 20 years; most users will already have some experience and/or familiarity • Not intended as a robust data analysis tool • Data tools difficult to find; buried in multiple menus • Intuitive process for field selection (simple highlighting) • No drag-and-drop • No automatic zip package option	• Relatively new to market • Robust and specific to data analysis • Data tools divided into logical, color-coded sections and alphabetized • Drag-and-drop data • Interprets all data as v_string even when clearly numeric in Excel, must run "Auto field" tool to fix; sometimes must be followed up with "Select" tool • Automatic zip package allows user to save workflows and data in one zip file
Capacity	• Row maximum: one million	• Row maximum: two billion-plus (Alteryx Documentation, 2020)
Descriptive statistics	• Less information generated by default • Fewer clicks to display table • Better readability • Standard deviation included by default • Ability to compare multiple datasets on same visual	• More information generated by default • Must add "Browse" function in workflow to view • Readability reduced due to larger number of rows • Standard deviation not included by default
Line charts	• Able to create line charts and plot multiple data sets on same graph	• Able to create sophisticated line chart forecasting models, but plotting two or more data sets on same graph requires several complex steps
Other statistical tools	• Fewer tools available (for example, no time series analysis tools)	• Many more tools available (for example, time series analysis, machine learning, text mining)

Works Cited

Alteryx Documentation. (2020, November 20). *Alteryx Database File Format.* https://help.alteryx.com/current/designer/alteryx-database-file-format

Boopalan, P. (2019, September 23). *Forecasting by Arima Model in Alteryx*. Visual BI. https://visualbi.com/blogs/self-service-bi/alteryx/forecasting-arima-model-alteryx/

Carrat, F., Figoni, J., Henny, J., Desenclos, J., Kab, S., de Lamballerie, X., & Zins, M. (2021, February 6). Evidence of early circulation of SARS-CoV-2 in France: findings from the population-based "CONSTANCES" cohort. *European Journal of Epidemiology 36*, 219-222. https://doi.org/10.1007/s10654-020-00716-2

Centers for Disease Control and Prevention (CDC). (2020, January 21). *First Travel-related Case of 2019 Novel Coronavirus Detected in United States*. https://www.cdc.gov/media/releases/2020/p0121-novel-coronavirus-travel-case.html

Deslandes, A., Berti, V., Tandjaoui-Lambotte, Y., Alloui, C., Zahar, J. R., Brichler, S., & Cohen, Y. (2020, June). SARS-CoV-2 was already spreading in France in late December 2019. *International Journal of Antimicrobial Agents 55*(6). https://doi.org/10.1016/j.ijantimicag.2020.106006

Edrada, E. M., Lopez, E. B., Villarama, J. B., Villarama, E. P. S., Dagoc, B. F., Smith, C., Sayo, A. R., Verona, J. A., Trifalgar-Arches, J., Lazaro, J., Balinas, E. G. M., Telan, E. F. O., Roy, L., Galon, M., Florida, C. H. N., Ukawa, T., Villanueva, A. M. G., Saito, N., Nepomuceno, J. R.,... Solante, R. M. (2020, April 14). First COVID-19 infections in the Philippines: a case report. *Tropical Medicine and Health 48*(21). https://doi.org/10.1186/s41182-020-00203-0

EricWe. (2020, December 4). *How to use the ARIMA tool*. Alteryx Designer Knowledge Base. https://community.alteryx.com/t5/Alteryx-Designer-Knowledge-Base/How-to-use-the-ARIMA-tool/ta-p/549669

Kamb, L. (2020, May 14). *When did coronavirus really hit Washington? 2 Snohomish County residents with antibodies were ill in December*. The Seattle Times. https://www.seattletimes.com/seattle-news/antibody-test-results-of-2-snohomish-county-residents-throw-into-question-timeline-of-coronaviruss-u-s-arrival/

SRC. (n. d.). *SRC Corporate Overview*. Retrieved June 13, 2021 from https://web.archive.org/web/20141024024821/http://src.mediaroomshowcase.com/file.php/2441/SRC+backgrounder.pdf

Stoecklin, S. B., Rolland, P., Silue, Y., Mailles, A., Campese, C., Simondon, A., Mechain, M., Meurice, L., Nguyen, M., Bassi, C., Yamani, E., Behillil, S., Ismael, S., Nguyen, D., Malvy, D., Lescure, F. X., Georges, S., Lazarus, C., Tabaï, A.,... Levy-Bruhl, D. (2020, February 13). First cases of coronavirus disease 2019 (COVID-19) in France: surveillance, investigations and control measures, January 2020. *Eurosurveillance 25*(6). https://dx.doi.org/10.2807%2F1560-7917.ES.2020.25.6.2000094