ASSIGNMENT: DATA ANALYSIS AND TRANSFORMATION WITH ALTERYX

STUDENT: ANASTASIA HASWANI

COURSE: CEBD 1150

INSTRUCTOR: ADRIÀN GONSÀLEZ SÀNCHEZ

PART 2
ANALYZING, TRANSFORMING AND PREDICTING DATA WITH ALTERYX

CONTENTS

INTRO

Data used in this section is taken from the Humanitarian Data Exchange. https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases. Novel Corona Virus (COVID-19) epidemiological data since 22 January 2020.

Fields available in the data include Province/State, Country/Region, Last Update, Confirmed, Suspected, Recovered, Deaths. On 23/03/2020, a new data structure was released. The current resources for the latest time series data are:

- time_series_covid19_confirmed_global_narrow.csv
- time_series_covid19_deaths_global_narrow.csv
- time_series_covid19_recovered_global_narrow.csv

Each of the three datasets contains the number of cases for every day in each country. Below are the fields:

- Province/State
- Country/Region
- Lat
- Long
- Date
- Value
- ISO 3166-1 Alpha 3-Codes
- Region Code
- Sub-region Code Intermediate Region Cod

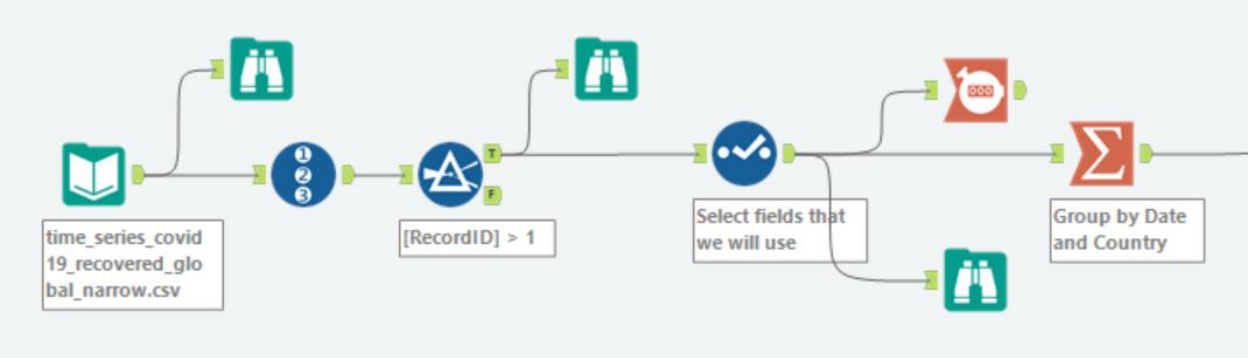
In the first assignment we used Python to do some basic EDA analysis. In this assignment we will concentrate of preparing the data for visualization and prediction.

Our goal is to join the three datasets and create few kinds of visualizations of COVID dynamics. Then we will try to use Alteryx prediction tools.

In this section we will not be describing every tool one-by-one like on the first section. Instead, we will be describing the process in general using screenshots and the results obtained.

LOAD AND PREPARE THE DATA

For each on the three tables we will use the same schema of loading and transforming the data:

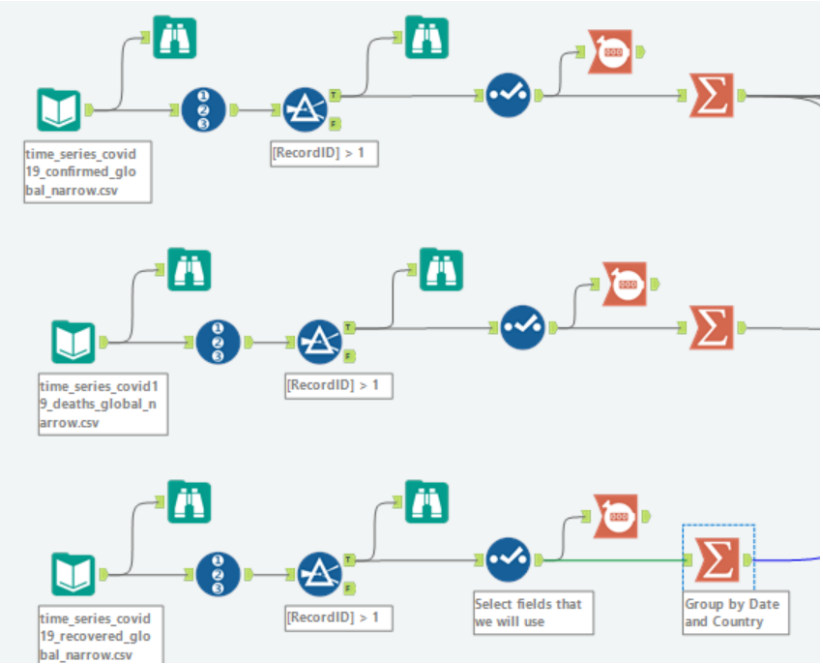


1. Load the dataset.
2. Add the record number to then be able to eliminate the first row. (If we use the *Start Data Import on Line* option, it wont save the name of the fields).
3. Select the fields we want to use.
4. Group the data by date and country.

Here is what we obtained for recovered dataset:

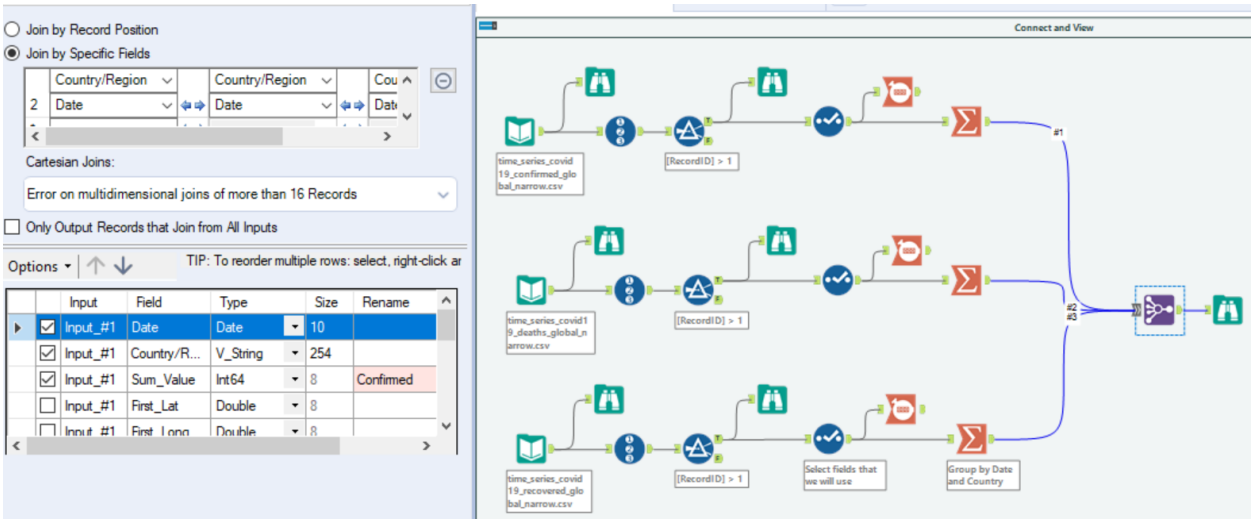
Record	Date	Country/Region	Sum_Value	First_Lat	First_Long
15442	2020-04-13	Bulgaria	71	42.7339	25.4858
15443	2020-04-13	Burkina Faso	161	12.2383	-1.5616
15444	2020-04-13	Burma	2	21.9162	95.956
15445	2020-04-13	Burundi	0	-3.3731	29.9189
15446	2020-04-13	Cabo Verde	1	16.5388	-23.0418
15447	2020-04-13	Cambodia	77	11.55	104.9167
15448	2020-04-13	Cameroon	98	3.848	11.5021
15449	2020-04-13	Canada	7758	56.1304	-106.3468
15450	2020-04-13	Central African Republic	3	6.6111	20.9394
15451	2020-04-13	Chad	2	15.4542	18.7322
15452	2020-04-13	Chile	2367	-35.6751	-71.543
15453	2020-04-13	China	78039	31.8257	117.2264
15454	2020-04-13	Colombia	319	4.5709	-74.2973
15455	2020-04-13	Comoros	0	-11.6455	43.3333

We will repeat this schema for two other datasets:



JOINING AND VIEWING THE DATA

Next step is to gather all the data in one place. We will use **Join Multiple** tool and join all the three tables on *Country* and *Date*:

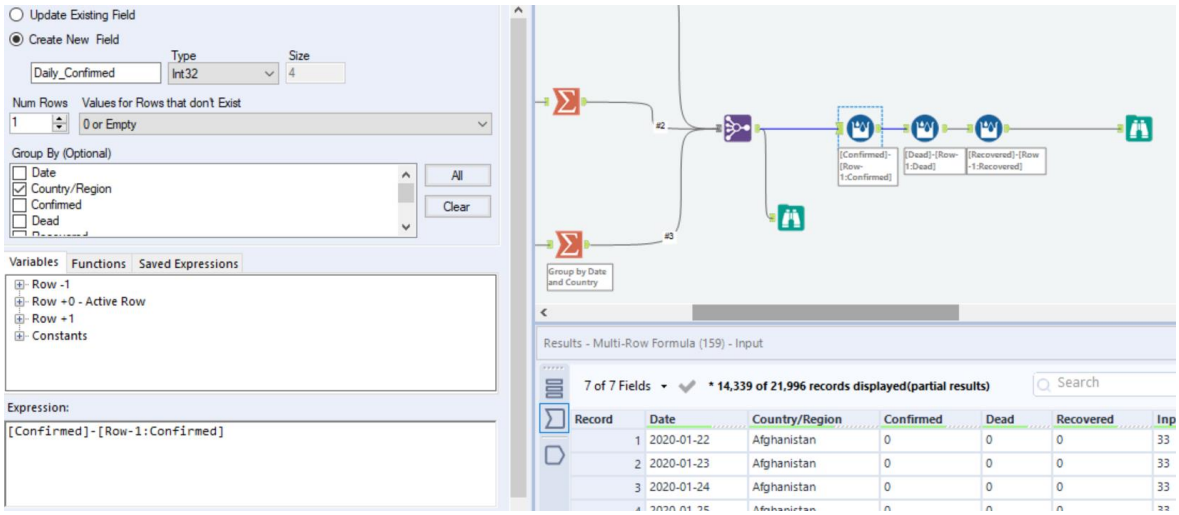


Here is what we got:

Record	Date	Country/Region	Confirmed	Dead	Recovered	Input_#3_First_Lat	Input_#3_First_Long
1725	2020-04-17	Barbados	75	5	15	13.1939	-59.5432
1726	2020-04-18	Barbados	75	5	17	13.1939	-59.5432
1727	2020-04-19	Barbados	75	5	17	13.1939	-59.5432
1728	2020-04-20	Barbados	75	5	19	13.1939	-59.5432
1729	2020-04-21	Barbados	75	5	25	13.1939	-59.5432
1730	2020-04-22	Barbados	75	5	25	13.1939	-59.5432
1731	2020-04-23	Barbados	76	6	30	13.1939	-59.5432
1732	2020-04-24	Barbados	77	6	31	13.1939	-59.5432
1733	2020-04-25	Barbados	79	6	31	13.1939	-59.5432
1734	2020-04-26	Barbados	79	6	39	13.1939	-59.5432
1735	2020-04-27	Barbados	80	6	39	13.1939	-59.5432
1736	2020-04-28	Barbados	80	6	39	13.1939	-59.5432
1737	2020-04-29	Barbados	80	7	39	13.1939	-59.5432
1738	2020-04-30	Barbados	81	7	39	13.1939	-59.5432
1739	2020-05-01	Barbados	81	7	39	13.1939	-59.5432

Next step is to add 3 new fields: daily confirmed, deaths and recovered numbers.

For each one of them we will use **Multi Row Tool.** We will calculate the difference between the current and previous rows and group the data by country to start at 0 for each country. We will write the results into the new column:



Let us check the results:

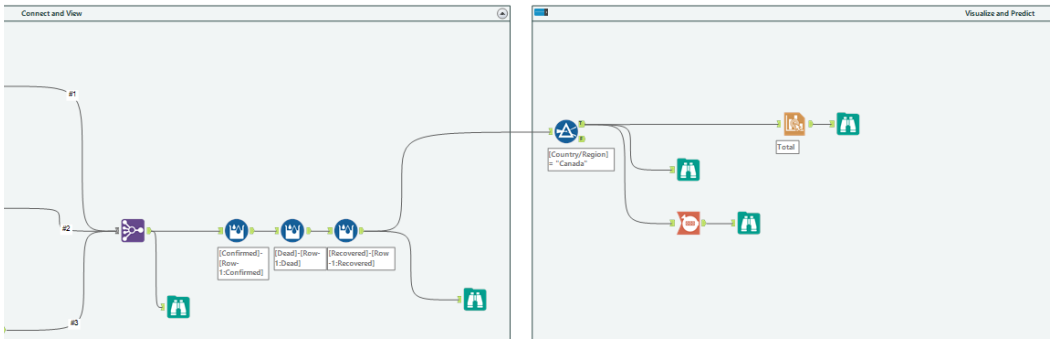
	Date	Country/Region	Confirmed	Dead	Recovered	Input_#3_First_Lat	Input_#3_First_Long	Daily_Confirmed	Daily_Deaths	Daily_Recovered
811	2020-05-09	Argentina	5776	300	1728	-38.4161	-63.6167	165	7	69
812	2020-05-10	Argentina	6034	305	1757	-38.4161	-63.6167	258	5	29
813	2020-05-11	Argentina	6278	314	1837	-38.4161	-63.6167	244	9	80
814	2020-05-12	Argentina	6563	319	1862	-38.4161	-63.6167	285	5	25
815	2020-05-13	Argentina	6879	329	2266	-38.4161	-63.6167	316	10	404
816	2020-05-14	Argentina	7134	353	2385	-38.4161	-63.6167	255	24	119
817	2020-05-15	Argentina	7479	356	2497	-38.4161	-63.6167	345	3	112
818	2020-05-16	Argentina	7805	363	2534	-38.4161	-63.6167	326	7	37
819	2020-05-17	Argentina	8068	373	2569	-38.4161	-63.6167	263	10	35
820	2020-01-22	Armenia	0	0	0	40.0691	45.0382	0	0	0
821	2020-01-23	Armenia	0	0	0	40.0691	45.0382	0	0	0
822	2020-01-24	Armenia	0	0	0	40.0691	45.0382	0	0	0
823	2020-01-25	Armenia	0	0	0	40.0691	45.0382	0	0	0
824	2020-01-26	Armenia	0	0	0	40.0691	45.0382	0	0	0
825	2020-01-27	Armenia	0	0	0	40.0691	45.0382	0	0	0

We see that the daily numbers start at 0 for every country. And the numbers look right.

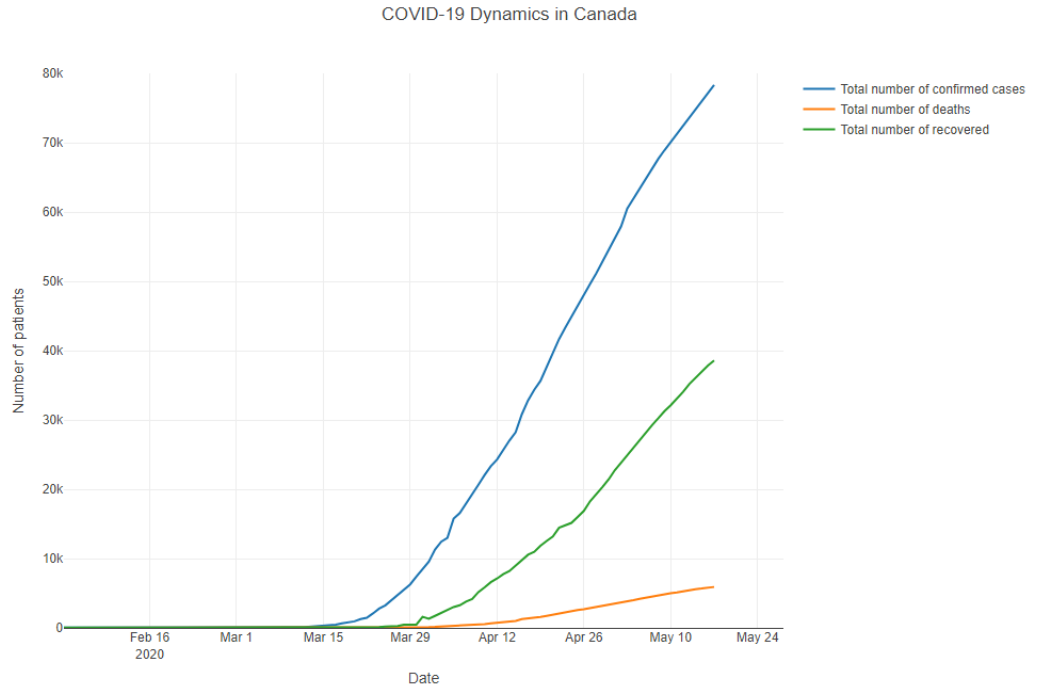
VISUALIZING TRENDS

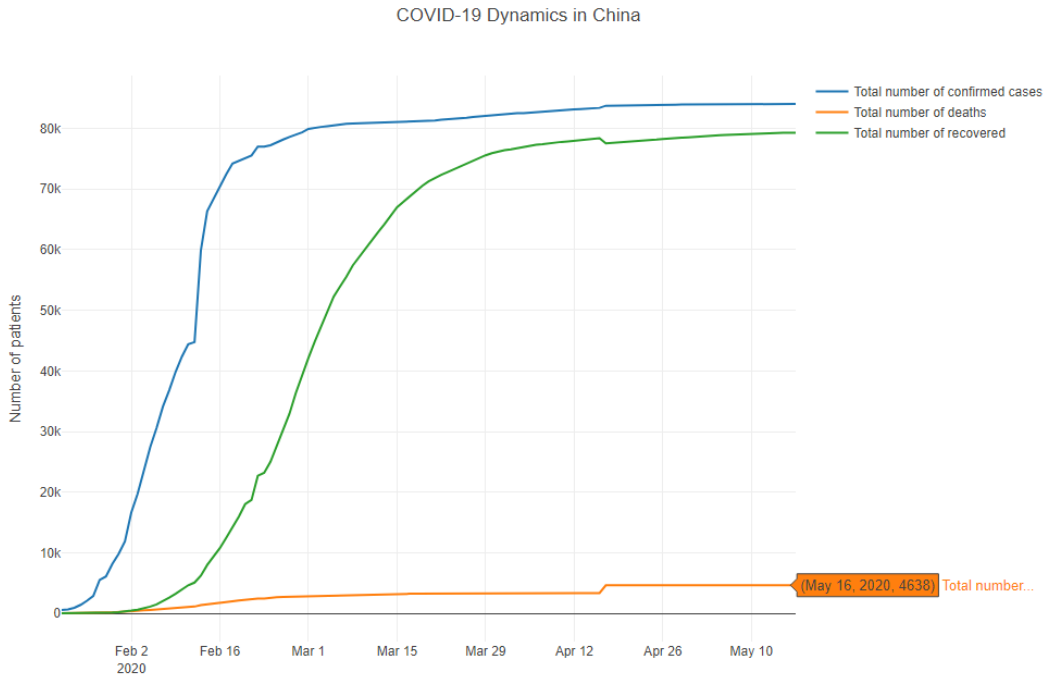
Firstly, we will create a filter where we will be able to choose a country. Then we can start building the charts. For that we will get an **Interactive Chart Tool** and set it up this way:

- **Layer 1:** *X: Date, Y: Confirmed cases*
- **Layer 2:** *X: Date, Y: Deaths cases*
- **Layer 3:** *X: Date, Y: Recovered cases*

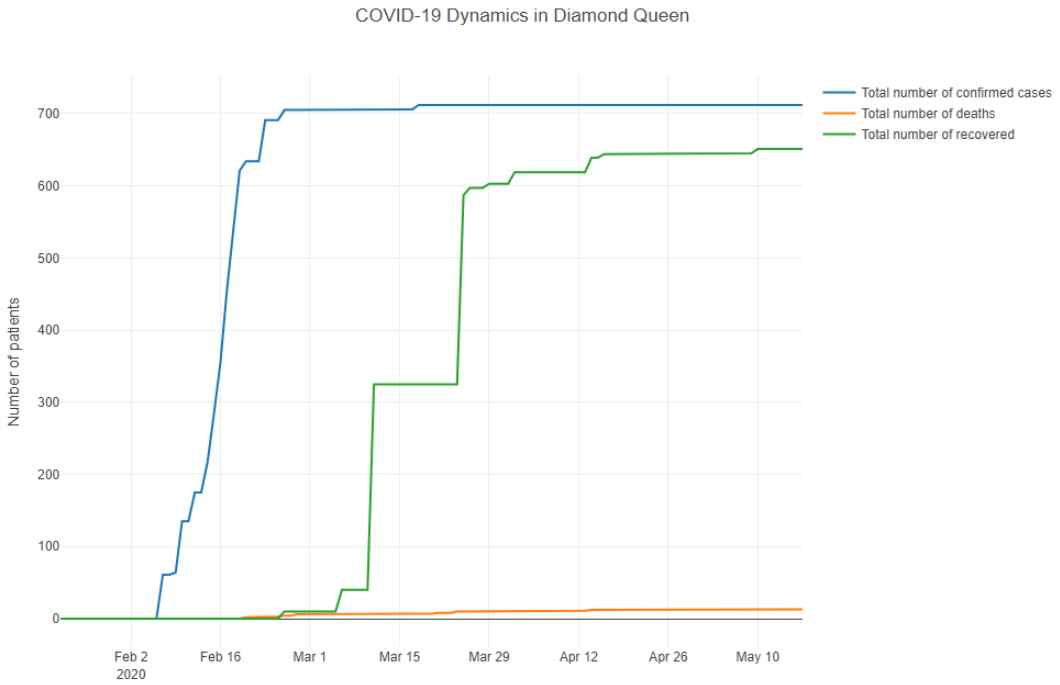


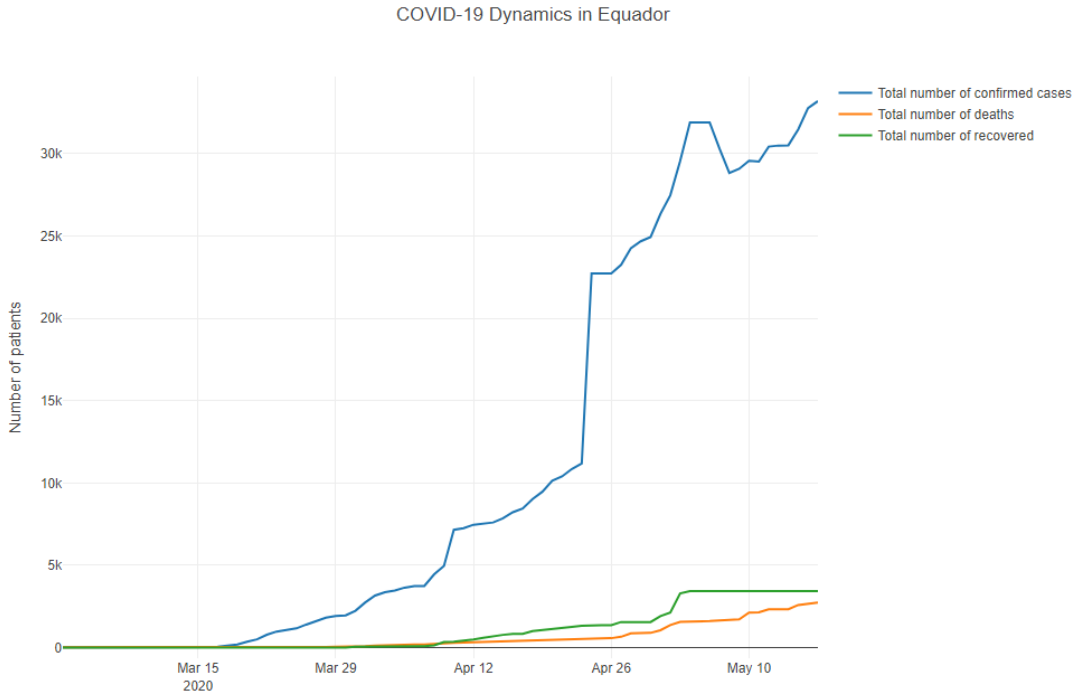
Here are few charts of total number of cases for few countries:





We can see how the three lines become horizontal, when the epidemy is finished. Canada needs more time to get over this, although its efforts of self-confinement have affected the line: Canada got to the 80k point in 9 weeks compared to China's 4 weeks. On another hand, we have a higher number of cases that are still growing.





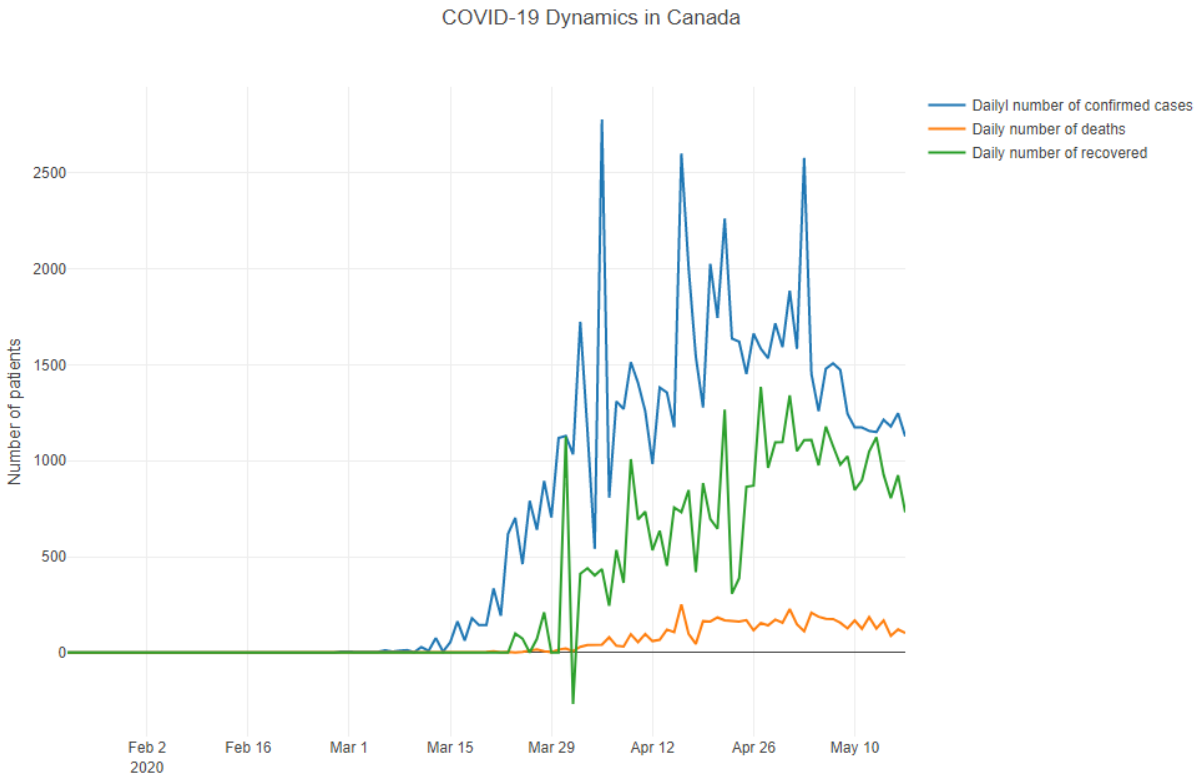
The last two graphs look interesting. In the case of Diamond Queen, since the cruise liner is a small area without major changes in terms of arrival or departure of people, we can notice how the trend tends to be changing quickly or staying still for a while. The local epidemy took a shorter time: the first confirmed case was registered February 7th and the last one March 19.

As for Ecuador, the April 24 confirmed cases jump from 11k to 22k, and the unchanged recovered line from May 4 till May 17 led us to assume that the data there is uneven, and quite possibly far from reality.

To have a better picture of the current dynamics we want to build the daily trends. For this we will create another **Interactive Chart Tool** and set it up this way:

- **Layer 1:** *X: Date, Y: Daily_Confirmed cases*
- **Layer 2:** *X: Date, Y: Daily_Deaths cases*
- **Layer 3:** *X: Date, Y: Daily_Recovered cases*

Here are the results for Canada:



I would say that it is hard to analyze this kind of graph. Let us try to divide our data into 20 groups (we get about a week for each group) and calculate the average values for each group. Then the trend should become clearer.

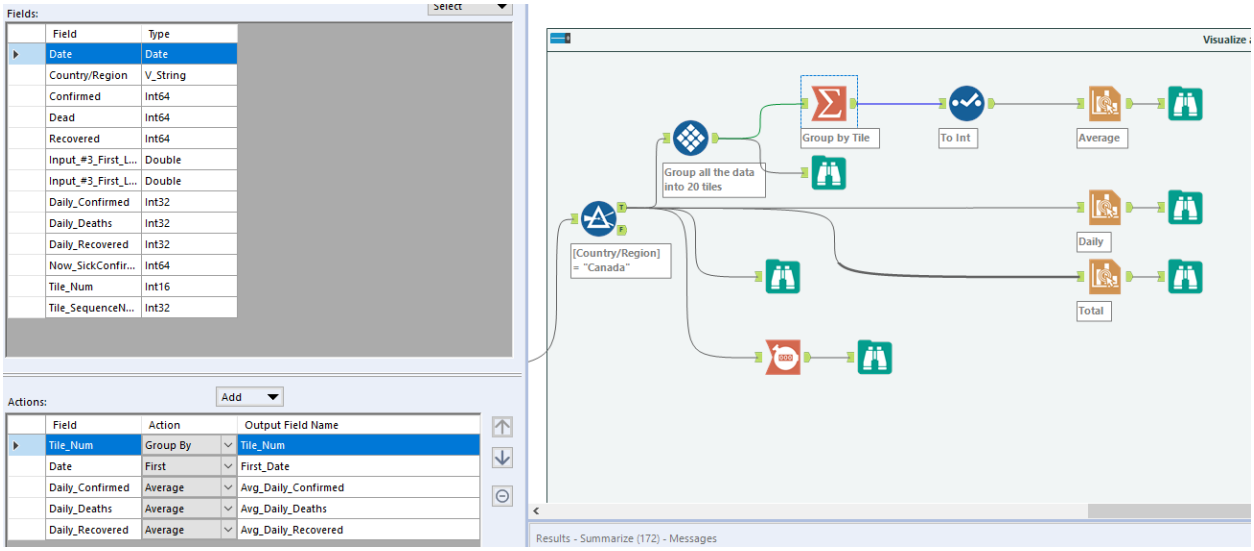
To divide the data into the groups I will use **Tile Tool** and set it up this way:

- *Tile Method: Equal Records*
- *Number of Tiles: 20*

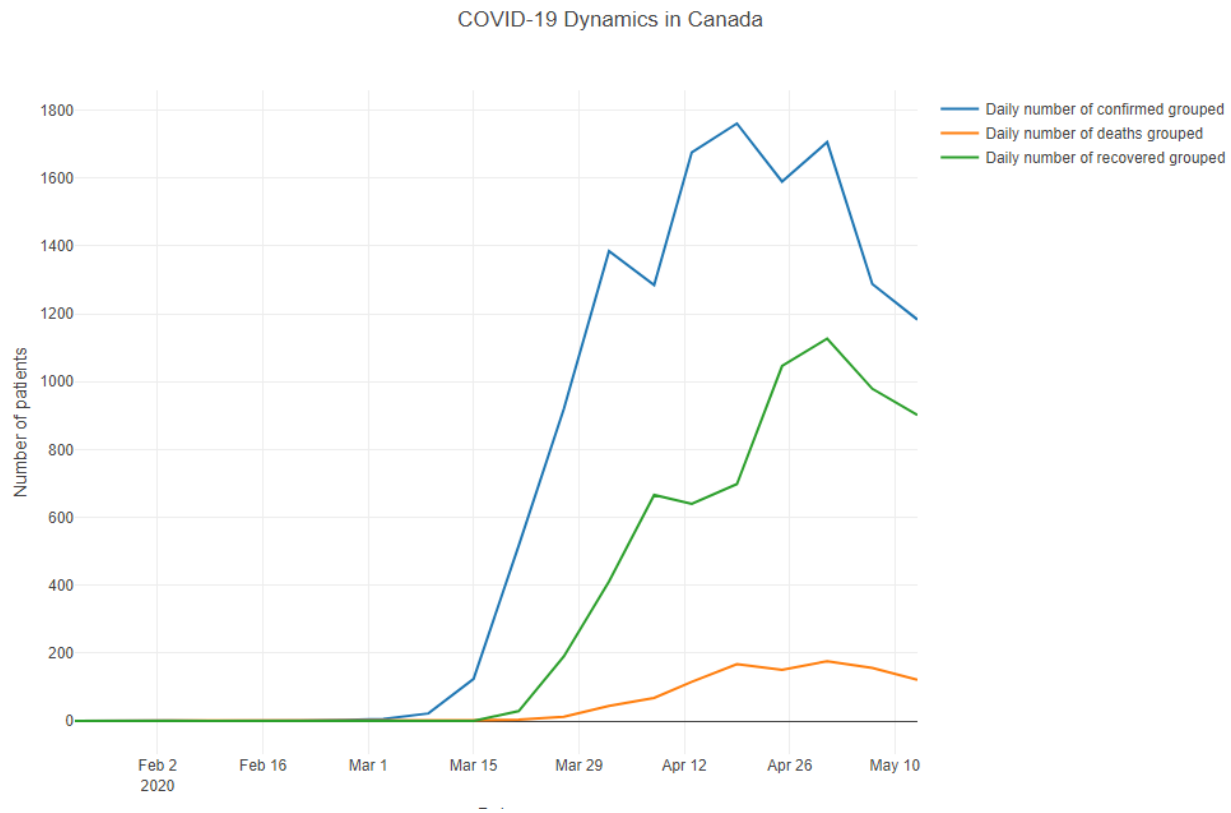
After we will use **Summarize Tool** to calculate the average for each group. We will group the data by Tile number. (See the settings for the Summarize in the image below).

After we will use Select Tool just to change the type of the Average values to Integer in order to round up people.

This is how our branch will look:

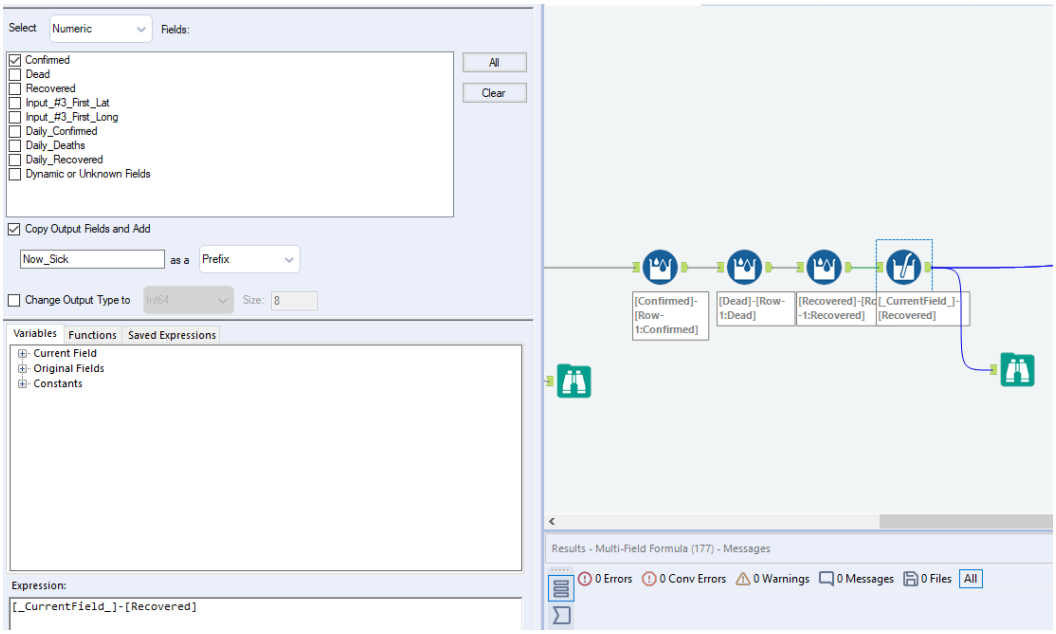


This is the “softer” graph that we obtained with the method of finding averages:

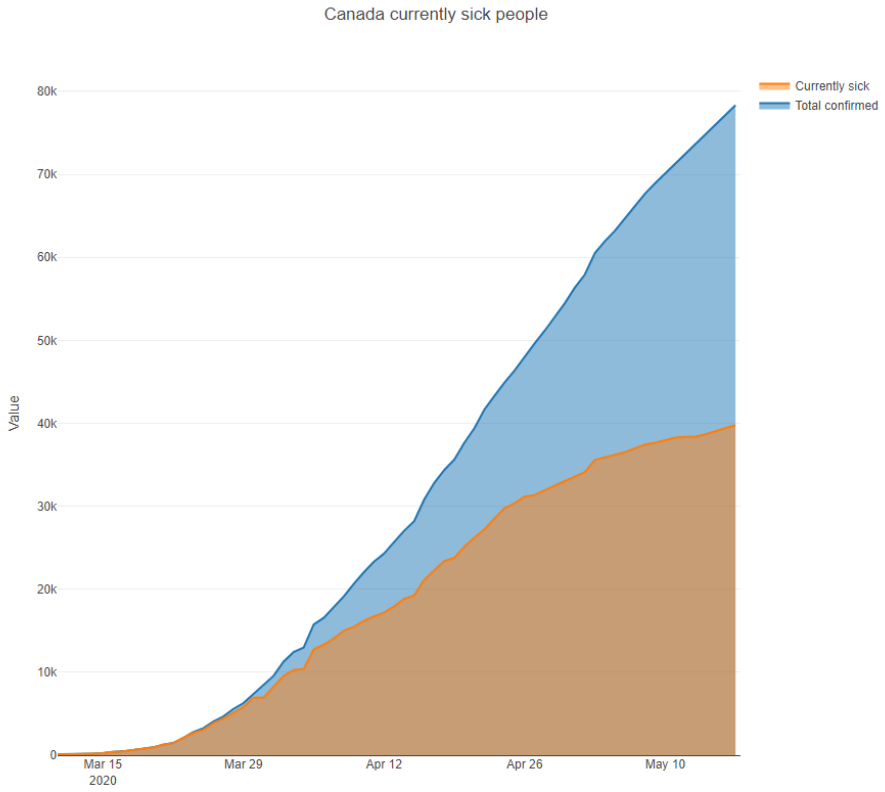


Now we can say that the situation in Canada is getting better.

I would like to see another type of graph: the trend of currently sick people. For this I will have to add one field using Multi-Field Formula Tool, where I will calculate currently ill people by finding the difference between Confirmed and Recovered people:



Bringing two layers to the chart (total confirmed and currently ill) we will have this output:

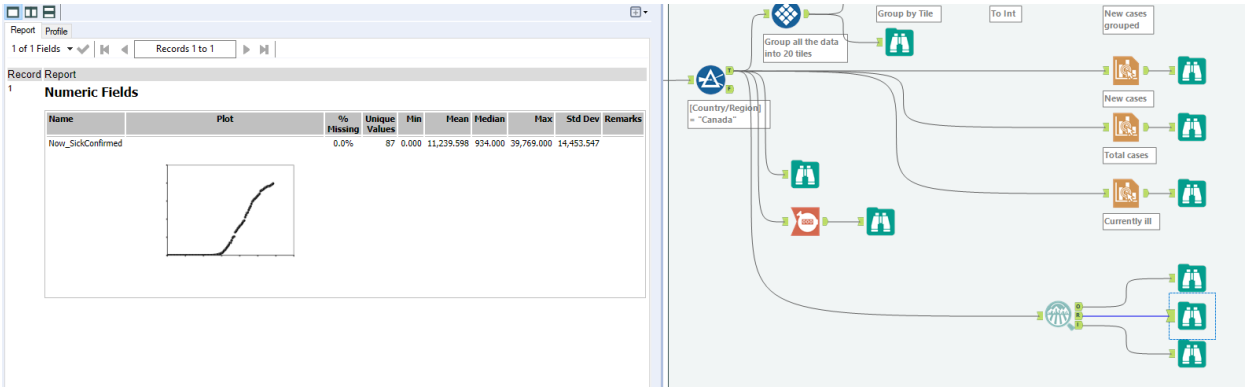


This trend shows that the curve is stabilizing and soon should go down, meaning that there will be more cured people and less new cases.

After looking at all of the charts we would like to build the prediction of the *Currently_Sick* field.

TIME SERIES

1. We will start from using **Field Summary Tool** to have a general look at the field, its chart and to see whether we have any values missing.



Looks like we have no values missing. We will pass to the next step.

2. For the time series we've chose two different tools: **ARIMA** and **ETS**. We will first compare their performance and then choose one of them.

3. To compare them we need to divide the data into train and test groups. We have 117 records for Canada. Let's divide them 100 to 17.

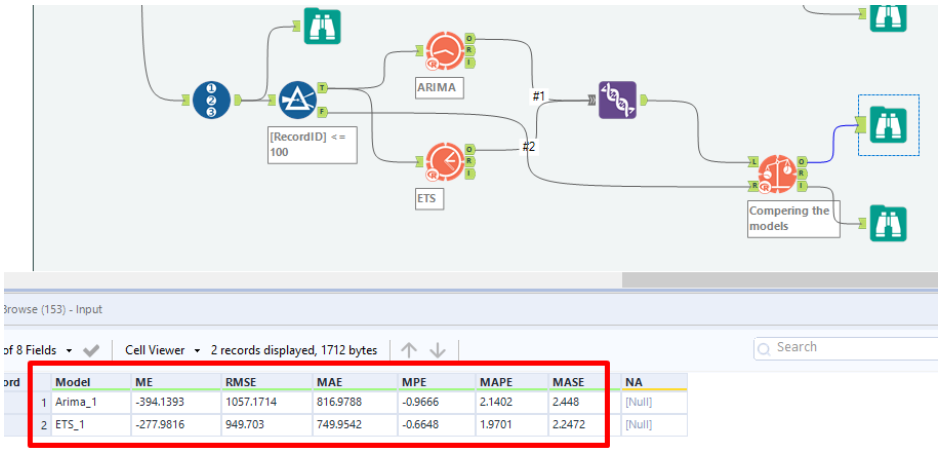
Using **Record ID Tool** we assign ID to each record. Then we will filter out 100 records and send them to both models (the configuration for both is similar):

- *Select the target field: Now_SickConfirmed*
- *Target field frequency: Daily*
- *Other Options tab. The number of periods: 17*

4. Then we direct the output of both models into **Union Tool.**

5. Next tool – **TS Compare.** There are two inputs:
    a. Output of the Union Tool
    b. False output of Filter tool (the 17 records that we left to be our test group)

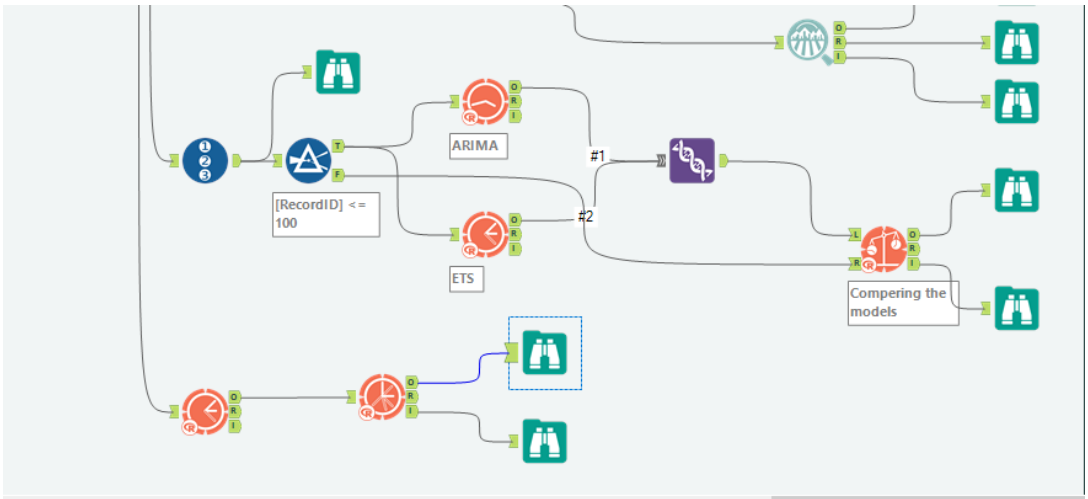
The results show that ETS model that uses exponential smoothing is slightly better then ARIMA that uses autoregressive moving average:



6. Now we will perform the actual forecast and swill start to build a model. We will bring the **ETS Tool** and feed our data to it. We will set it like we did previously with the *Number of periods to include* set to *60.*

7. Next step is to get a **TS Forecast Tool** and connect it to ETS model tool. In the settings we will change *Number of periods into the future to forecast* to *60.*

8. The results are in a table with the values considering different scenarios:

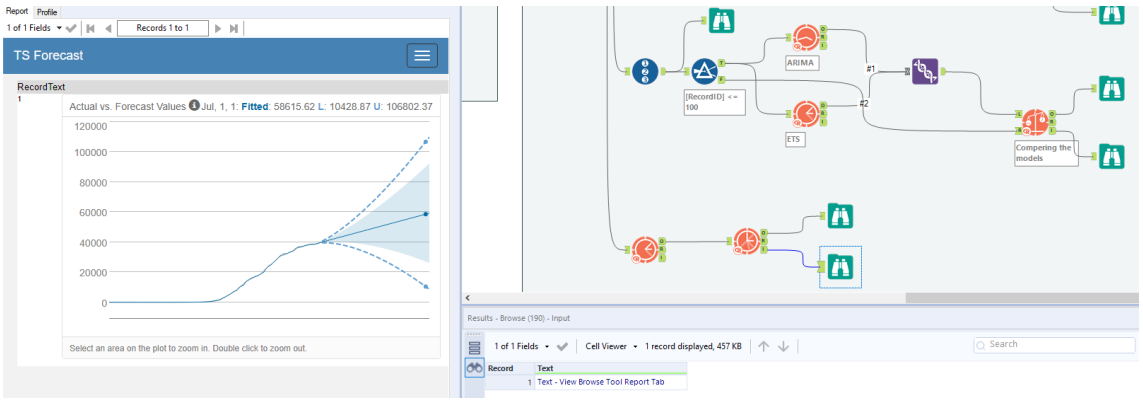


lts - Browse (189) - Input

7 of 7 Fields | Cell Viewer | 60 records displayed, 4497 bytes | Search

Record	Period	Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
1	17	6	40056.745351	40658.038039	40449.909507	39663.581195	39455.452663
2	17	7	40382.339606	41251.3744	40950.570915	39814.108297	39513.304812
3	18	1	40707.933861	41893.911705	41483.403091	39932.464631	39521.956018
4	18	2	41033.528116	42576.26658	42042.270577	40024.785655	39490.789653
5	18	3	41359.122372	43293.09243	42623.677422	40094.567321	39425.152313
6	18	4	41684.716627	44040.91298	43225.350612	40144.082641	39328.520274
7	18	5	42010.310882	44817.235947	43845.660531	40174.961232	39203.385817
8	18	6	42335.905137	45620.150034	44483.357449	40188.452826	39051.660241
9	18	7	42661.499392	46448.119661	45137.437303	40185.561481	38874.879124
10	19	1	42987.093647	47299.869671	45807.066314	40167.12098	38674.317624
11	19	2	43312.687903	48174.315161	46491.535102	40133.840703	38451.060644

And the graphical interpretation:



It possible to perform forecast for the daily number of new cases and for the number of deaths.