Understanding

COVID-19 Incident and Google Mobility

Data Relationship

using Data Science

Martha Dunne

Regis University

MSDS 692 Data Science Practicum 1

Prof. Paul Andrus

August 23, 2020

Abstract

Many companies provide cell phone mobility data in the location data business. On April 2020 Google released 'COVID-19 Community Mobility Reports'. These reports use aggregated, anonymized global data to chart movement trends over time by geography, across 6 high-level categories, recording an increase or decrease in visits with relation to a baseline.

The objective of this introductory data science project is to merge the Google Mobility data with COVID-19 Cases, and determine if the broad category mobility data would contain enough detail to show relationships or make predictions. The global COVID-19 data is sourced from John Hopkins.

Predicting COVIDd-19 Cases using Google Mobility Data

My goal for my initial data science project was to determine if there is a relationship between the publicly available data sources for COVID-19 cases and Google anonymized mobility data.

The NULL hypothesis is the following:

> H0: There is not a relationship between publicly available mobility data and COVID-19 case count data.

The Alternate hypothesis is the following:

> Ha: There is a relationship between publicly available mobility data and COVID-19 case count data.

On January 30, 2020, the COVID-19 outbreak was declared a Public Health Emergency of International Concern by the World Health Organization[1]. Countries have responded with various forms of social distancing and non-essential travel lockdowns to slow down the spread of the epidemic, such as closing bars, restaurants, universities, workplaces. These are referred to as Non-Pharmaceutical Interventions (NPIs). Tracking changes in an individual's mobility can be useful for tracing individual have been exposed, as well as for monitoring current traffic, and preparing to set policies as lockdowns are relaxed.

The COVID-19 data is sourced from John Hopkins, with global data for confirmed cases, deaths and recovered cases. The Mobility data is sourced from Google 'COVID-19 Community Mobility Reports', with global data to chart movement trends over time by geography, across 6

---

[1] https://www.nature.com/articles/s41597-020-00575-2

high-level categories, recording a percentage point increase or decrease in visits with relation to a baseline.

This project was performed using Excel (Microsoft Office 365 ProPlus v2002), Anaconda Navigator, Jupyter notebook (v6.0.1), Python 3 (v3.8) and Python libraries. Since the data sources are csv files and I was more familiar with Excel, I began my initial analysis of the individual datasets in Excel. Then I prepared a Jupyter notebook and began analysis, visualizations and modeling in python.

Data Sources Description

COVID-19

The John Hopkins COVID-19 repository of time series summary tables is located on github[2]. The files are updated with current information daily around 23:59(UTC). I selected the following three global files for this project.

- time_series_covid19_confirmed_global.csv
- time_series_covid19_deaths_global.csv
- time_series_covid19_recovered_global.csv

COVID-19 File Structure

The three files all have the same wide format, with their associated confirmed, deaths and recovered counts as integers in each days column. The data begins January 22nd, and is current to yesterdays reporting.

[2] https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series.

Here are the raw data fields, including data examples. Green fields will be retained for merge, after cleanup.

Field	Type	Examples
Province/State	string	Bonaire, Sint Eustatius and Saba
Country/Region	string	Saint Vincent and the Grenadines
Lat	float	12.1784
Long	float	-68.2385
1/22/2020	integer	0
daily to current date	integer	9999

COVID-19 File Data

On August 17, 2020 there were 187 countries represented in this file.

The following cases file subset example demonstrates that the countries like Australia which report at the province/state level will need to be aggregated. This file does not contain an aggregated row for these countries. Australia, Canada, China, Netherlands, the UK, France and Denmark report under the province/state level. The US and other countries report at the country level.

Province/State	Country/Region	Lat	Long	1/22/2020	1/23/2020
	Afghanistan	33	65	0	0
	Albania	41.1533	20.1683	0	0
	Algeria	28.0339	1.6596	0	0
	Andorra	42.5063	1.5218	0	0
	Angola	-11.2027	17.8739	0	0
	Antigua and Barbuda	17.0608	-61.7964	0	0
	Argentina	-38.4161	-63.6167	0	0
	Armenia	40.0691	45.0382	0	0
Australian Capital Territory	Australia	-35.4735	149.0124	0	0
New South Wales	Australia	-33.8688	151.2093	0	0
Northern Territory	Australia	-12.4634	130.8456	0	0
Queensland	Australia	-28.0167	153.4	0	0
South Australia	Australia	-34.9285	138.6007	0	0
Tasmania	Australia	-41.4545	145.9707	0	0
Victoria	Australia	-37.8136	144.9631	0	0
Western Australia	Australia	-31.9505	115.8605	0	0

The three files need to be converted to long format, and their value columns need to be renamed (cases, deaths, recovered) before they can be merged into one long format COVID-19 file.

Google Community Mobility Reports

On April 3, 2020, Google released their COVID-19 Community Mobility Reports[3], initially with 131 countries. These reports track movement trends over time by across six high-level categories, and record a percentage point increase or decrease in visits with relation to a baseline. They track visits and length of stay changes compared to a baseline.

For each region-category, the baseline is seven individual values for each day of the week. Each day's baseline is the median day value during the five-week baseline period from January 3 to Feb 6, 2020. Changes for each day are compared to a baseline value for that day of the week.

Mobility 6 High Level Categories

- See Appendix A

Mobility File Structure

The data begins on February 15, 2020 and is current to approximately 3 days ago, the time it takes to produce the datasets. Here are the raw data fields, including data examples.

Green fields will be retained in the merged file, after cleanup.

[3] https://www.google.com/covid19/mobility/

Field	Type	Example
country_region_code	string	US
country_region	string	United States
sub_region_1	string	Null for country
sub_region_2	string	Null for country
metro_area*	string	
iso_3166_2_code	string	Null for country
census_fips_code	string	Null for country
date	date	2/16/2020
retail_and_recreation_percent_change_from_baseline	integer	7
grocery_and_pharmacy_percent_change_from_baseline	integer	1
parks_percent_change_from_baseline	integer	16
transit_stations_percent_change_from_baseline	integer	2
workplaces_percent_change_from_baseline	integer	0
residential_percent_change_from_baseline	integer	-1

Mobility File Data

On August 17, 2020 there were 135 countries represented in this file, a much smaller set than in the COVID-19 files.

This file contains the data aggregated to the various levels. To retrieve only the records pre-aggregated at the country level requires filtering where sub-regions, iso code and fips code is null. Also note the columns are not static. In the time since I started this project, the column metro_area was added to the file.

Below is an example of the data.

Field	Example
country_region_code	US
country_region	United States
sub_region_1	*Where null*
sub_region_2	*Where null*
iso_3166_2_code	*Where null*
census_fips_code	*Where null*
date	2/15/2020
retail_and_recreation_percent_change_from_baseline	6
grocery_and_pharmacy_percent_change_from_baseline	2
parks_percent_change_from_baseline	15

transit_stations_percent_change_from_baseline	3
workplaces_percent_change_from_baseline	2
residential_percent_change_from_baseline	-1

Google recommendations:

- Residential is measured in duration units (all other categories measure in visitors) so do not compare this change with other categories. Also people already spend a lot of time at home, so changes in Residential are likely to be smaller.
- Parks are highly influenced by weather and holidays, expect larger spikes in this category.
- Gaps occur when the quantity of data is too low to meet data quality and anonymity standards, don't interpret this as zero change in visitors.
- Do not infer that larger changes mean more visitors or smaller changes mean less visitors.
- Avoid comparing day-to-day changes. Especially weekends with weekdays.
- Avoid comparing levels across countries or regions. Regions can have local differences in the data which might be misleading (e.g. rural versus urban areas)

Merging COVID-19 and Google Community Mobility Reports

There were multiple iterations of cleanup, merging, and analysis performed as the investigation proceeded.

Here are the final steps required to merge and prepare the file PracticumMergedData.csv.

COVID-19 global file preparation issues:

- three wide format csv files available to merge; cases, deaths, recovered
- 188 countries – not row for row match with Mobility

    - Rename country to match Mobility: Taiwan* to Taiwan
    - Interesting politics file did not contain Hong Kong, N. Korea (does contain several cruise ships from early outbreaks)
- Data starts January 22, 2020, and is current to previous day
- Convert wide format to long format for each file
- Rename each files value field appropriately: cases, deaths, recovered
- Merge three 3 long format files either before or after next step
- Aggregate data rows to handle countries reporting as non-aggregated

Mobility global file prepare issues:

- one long format csv file
- 135 countries – not row for row match with COVID, and smaller countries missing
    - Renames to match COVID-19: United States to US
    - Interesting politics file did not contain China, Korea's, Iran, Syria
- Data starts February 15, 2020, same format as COVID-19 and is current to approximately 3 days ago, because it takes several days to prepare.
- Select country level rows data rows only, filter out non-country level rows
- Retain (rename) fields: country, date, retailrec, grocerrx, parks, transit, work

  Note that residence has different unit of measurement

For report of country mismatch, refer to mergefile_country_compare.csv.

Merging file issues:

- Inner join merge on country_region (string) and date (datetime) fields.
- Merged file is limited by mobility data dates, as well as country_regions. For report of country mismatch, refer to mergefile_country_compare.csv.

- Generate delta column to expose the daily change in cases and deaths; ie. casesdelta and deathsdelta. (Because there is a two week incubation period for the virus, I also investigated a 14 day shift casesdelta column to compare the correlations against. See Appendix B.)

Mobility global file final structure:

Field	**Type**
country_region	string
date	date
cases	integer
deaths	integer
casesdelta	integer
deathsdelta	integer
retailrec	integer
grocerrx	integer
parks	integer
transit	integer
work	integer
reside	integer

Analysis and Reality

Over the course of eight months the best and worst list of countries at managing the COVID-19 pandemic has been dynamic. This is for a variety of reasons, not just related to lockdowns and mobility data. The best countries are those who appear to have handled the pandemic well and kept their case counts low. I selected sample companies depending on the data that existed, but also after referring to various sources[4]. Taiwan has low cases, but it did not participate in a typical lockdown with restricted mobility. They wore masks, closed borders,

[4] Bremmer, Ian (June 12, 2020) 'The Best Global Responses to COVID-19 Pandemic', https://time.com/5851633/best-global-responses-covid-19/

enforced quarantines, followed medical officials and kept their businesses and retail open. Up until a few days ago, New Zealand had maintained zero new cases for several months. Then there were some large countries of interest (China and the Koreas) that were not represented in the mobility data.

In the table below, I display the August 14, 2020 daily and total cases and deaths counts from the merged file for the countries I selected from the dataframe to investigate in more detail. United States, Brazil and India, currently are the worst countries with the largest counts in daily and total cases and deaths. The Global row contains a separate dataframe query only for reference.

8/14/2020	2019 population	cases	deaths	casesdelta	deathsdelta
Taiwan	23,773,876*	481	7	0	0
New Zealand	4,917,000	1609	22	7	0
Australia	25,364,310	23,035	379	293	4
Singapore	5,703,570	55,580	27	83	0
Canada	37,589,260	123,605	9,068	425	5
India	1,366,417,750	2,525,922	49,036	64,732	996
Brazil	211,049,530	3,275,520	106,523	50,644	1,060
US	328,239,520	5,313,252	168,452	64,294	1,342
Global	n/a	20,420,761	734,275	297,341	9,878

I did not consider country population or demographic data sources when I ran this project. For a relative population reference, I am including a column in the table for these countries populations as of 2019, per worldbank[5] and *worldometers[6]. I am not including a global population because the two merged datasets would not contain the entire global population.

[5] https://data.worldbank.org/indicator/SP.POP.TOTL
[6] https://www.worldometers.info/

Descriptive Statistics

Below are the statistics for casesdelta for individual countries; Taiwan, New Zealand, Australia, Singapore, Canada, United States, India, Brazil, and the merged file total.

Cases delta	TA	NZ	AU	SI	CA	US	IN	BR	Merged file total
count	182	182	182	182	182	182	182	182	22797
mean	2.54	8.84	126.48	305.02	679.11	29193.62	13878.68	17997.36	894.45
std	5.26	20.36	166.76	281.98	595.36	20982.73	18934.84	18853.25	4574.82
min	-2	-1	0	0	0	0	0	0	-10034
25%	0	0	10	55.75	242.75	18209.25	176.75	830.5	0
50%	0	1	25.5	237.5	469.5	25594.5	4491.5	11805	22
75%	2	3	218.5	464.5	1128.25	45179.5	19384.25	30901.25	289
max	27	89	716	1426	2778	77255	66999	69074	77255

The distributions have multiple modes, so the python statistics mode function returns an error.

Exploratory Data Analysis

Merged File Correlations

The chart below displays the casesdelta Pearson correlations for the countries of interest.

Correlation Pearson	**TA**	**NZ**	**AU**	**SI**	**CA**	**US**	**BR**	**IN**
cases	-0.290	-0.202	0.621	0.188	-0.022	0.862	0.857	0.989
deaths	-0.408	-0.475	0.486	0.378	-0.117	0.805	0.894	0.992
casesdelta	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
deathsdelta	0.147	0.068	0.658	0.167	0.781	0.367	0.900	0.885
retailrec	-0.221	**-0.571**	-0.080	**-0.789**	**-0.782**	-0.175	0.021	-0.083
grocerrx	-0.103	**-0.555**	-0.131	**-0.587**	**-0.683**	-0.208	0.485	0.280
transit	-0.273	**-0.526**	-0.237	**-0.798**	**-0.730**	-0.330	-0.052	-0.011
work	0.058	**-0.615**	-0.115	**-0.737**	**-0.640**	-0.425	0.046	0.063
parks	-0.003	**-0.537**	-0.039	**-0.769**	-0.369	**0.550**	-0.015	-0.289
reside	0.182	**0.614**	0.123	**0.744**	**0.721**	0.221	0.132	-0.078

Taiwan was included here to confirm their non-shutdown approach and expected mild mobility changes, which were also confirmed during visualizations. I expected the

United States, Brazil and India would have poor correlations. Singapore and Canada have relatively strong positive correlations in reside. For strong negative correlations, Singapore is strongest in transit (-0.798), Singapore and Canada are both strong in retailrec (-0.78). New Zealand had moderate correlations in mobility.

The heatmaps below display the Pearson correlations.

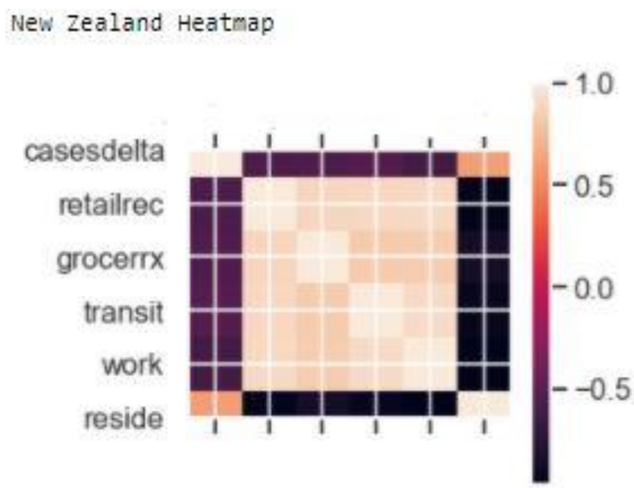


Figure 1: New Zealand Correlation Heatmap

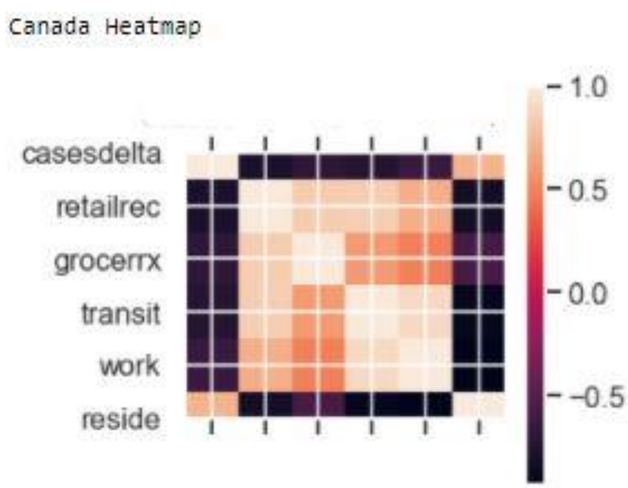


Figure 22: Canada Correlation Heatmap

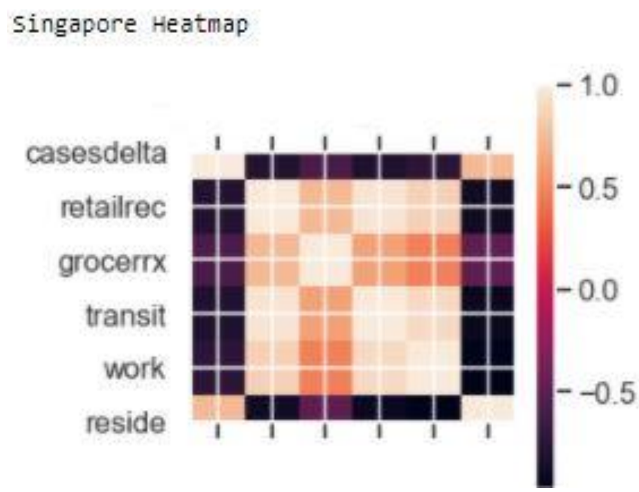


Figure 3: Singapore Correlation Heatmap

The chart below displays the Kendal Tau and Spearman correlations for New Zealand, Singapore and Canada.

	NZ kendall	SI kendall	CA kendall	NZ spearman	SI spearman	CA spearman
retailrec	-0.3759	-0.6847	-0.6034	-0.4991	-0.8636	-0.794922
grocerrx	-0.2981	-0.3454	-0.4837	-0.3856	-0.5256	-0.679335
transit	-0.39	-0.689	-0.6565	-0.5227	-0.8543	-0.837621
work	-0.4403	-0.6227	-0.5789	-0.57	-0.8099	-0.729183
parks	-0.3934	-0.6483	-0.2209	-0.5164	-0.838	-0.289791
reside	0.41839	0.64874	0.58815	0.5512	0.82251	0.767781

The Spearman correlation assumes that there are two ordinal variables or two variables that are related in some way, but not linearly. It is usually larger than the Kendall’s Tau, as it is here. It is only smaller when the deviations are huge among the observations of your data. Below are the statistics for casesdelta for reference.

Cases delta	NZ	SI	CA
count	182	182	182
mean	8.84	305.02	679.11
std	**20.36**	**281.98**	**595.36**
min	-1	0	0
25%	0	55.75	242.75
50%	1	237.5	469.5
75%	3	464.5	1128.25
max	89	1426	2778

Merged File Plots

Plots for low and high COVID-19 count countries are displayed separately because the scale makes them impossible to meaningfully display altogether.

Below are plots of COVID-19 and Mobility Total and Daily data for countries which managed for low cases (New Zealand, Australia, Singapore and Canada).

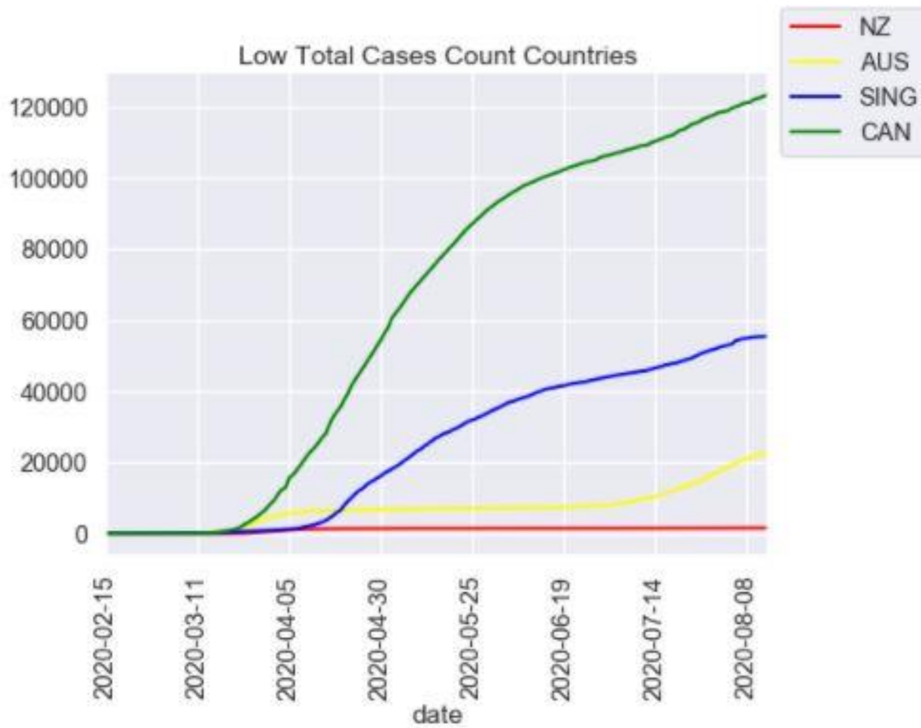


Figure 4: Total Cases Counts for Low Cases Countries

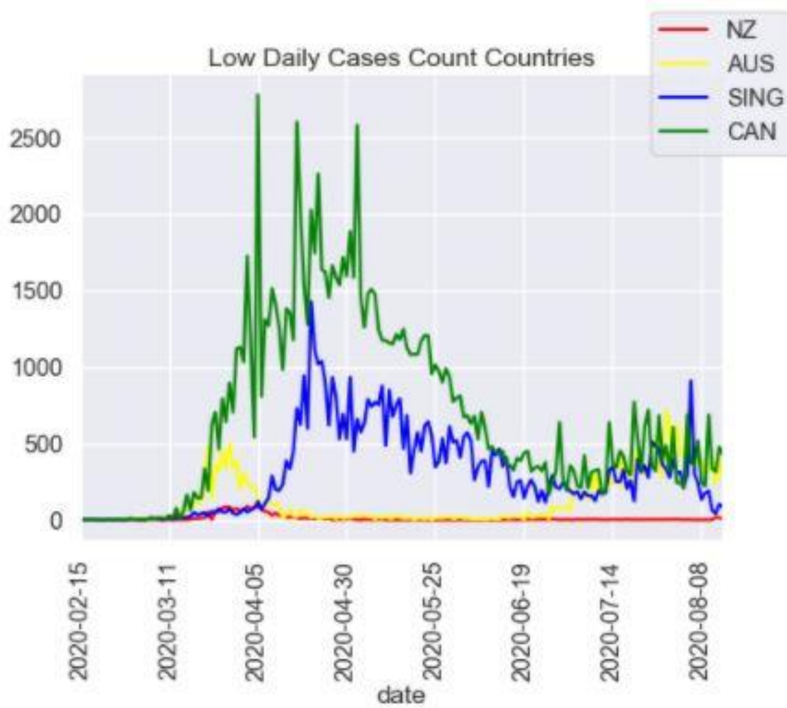


Figure 5: Daily Cases Counts for Low Cases Countries

These are plots of the Singapore data.

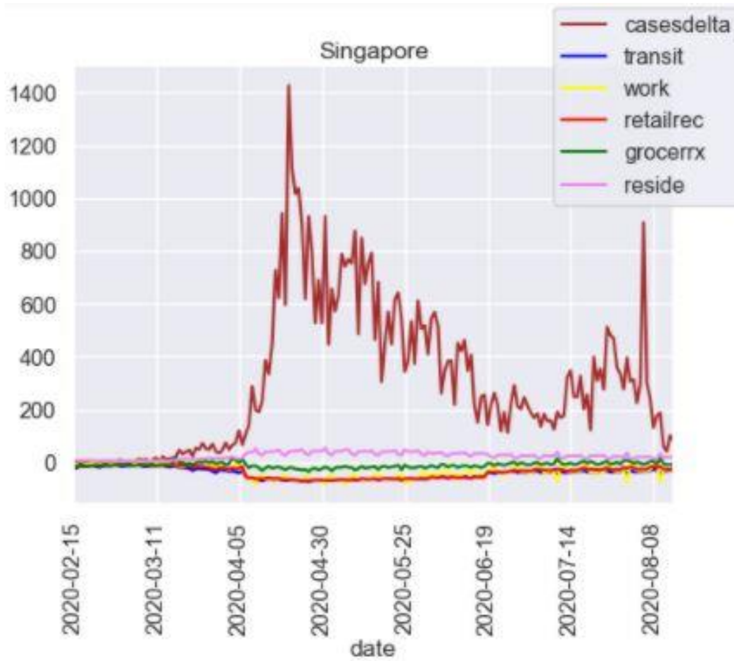


Figure 6: Singapore Covid and Mobility data

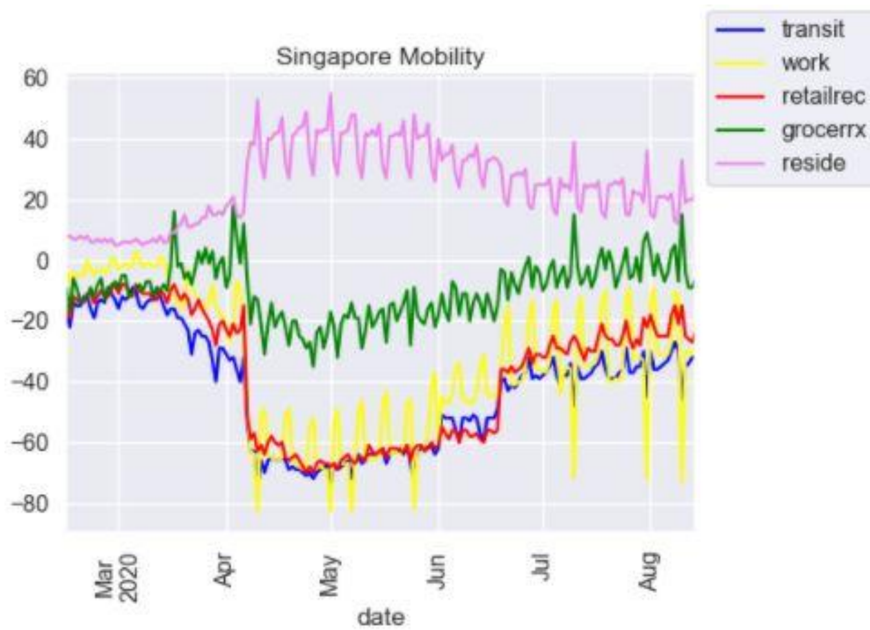


Figure 7: Singapore Mobility data – detail

These are plots of the Canada data.

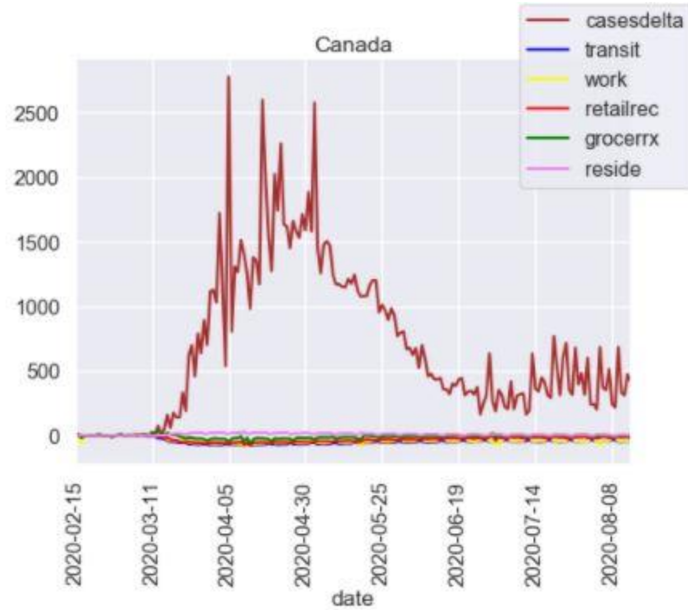


Figure 8: Canada Covid and Mobility data

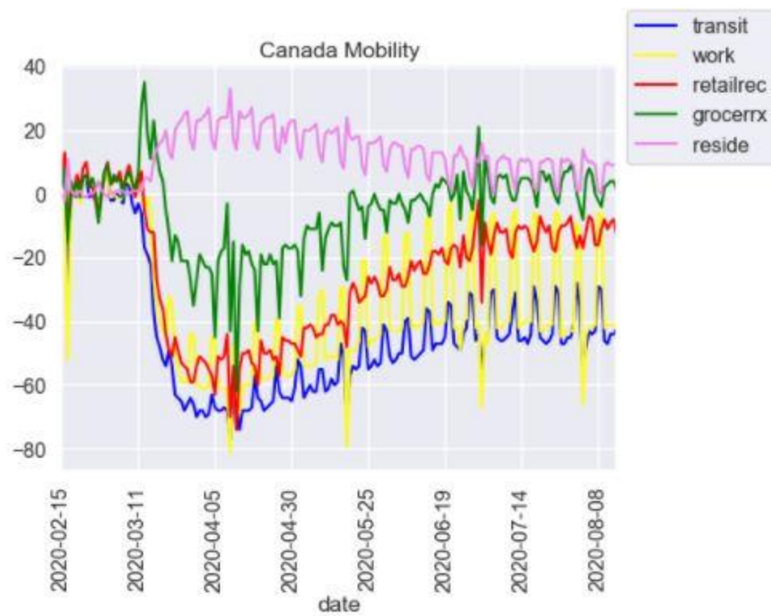


Figure 9: Canada Mobility data – detail

Below are plots of of COVID-19 and Mobility Total and Daily data for three countries with the highest cases (United States, Brazil and India).

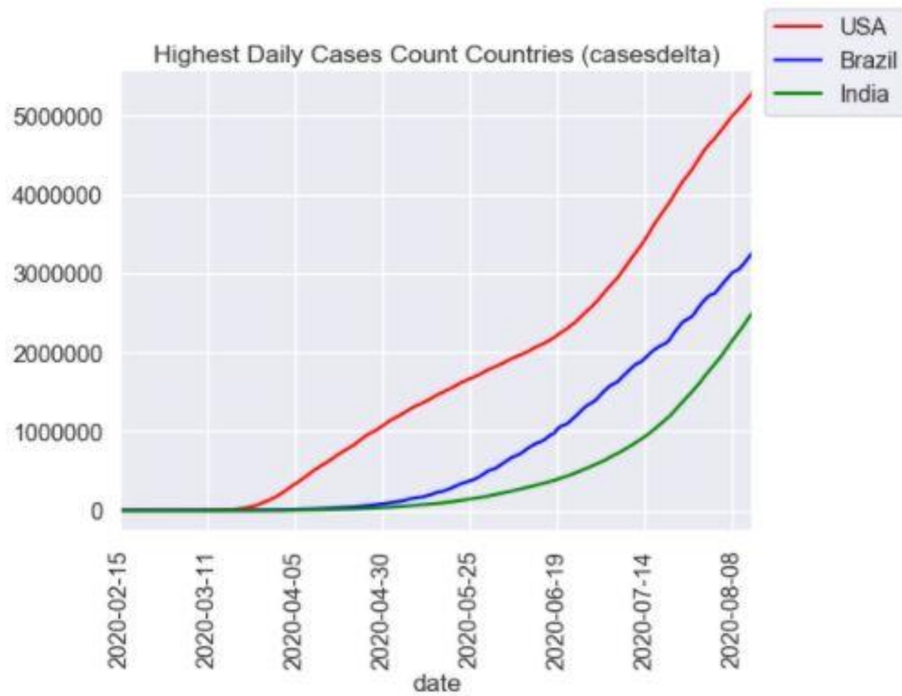


Figure 10: Total Cases Counts for High Cases Countries

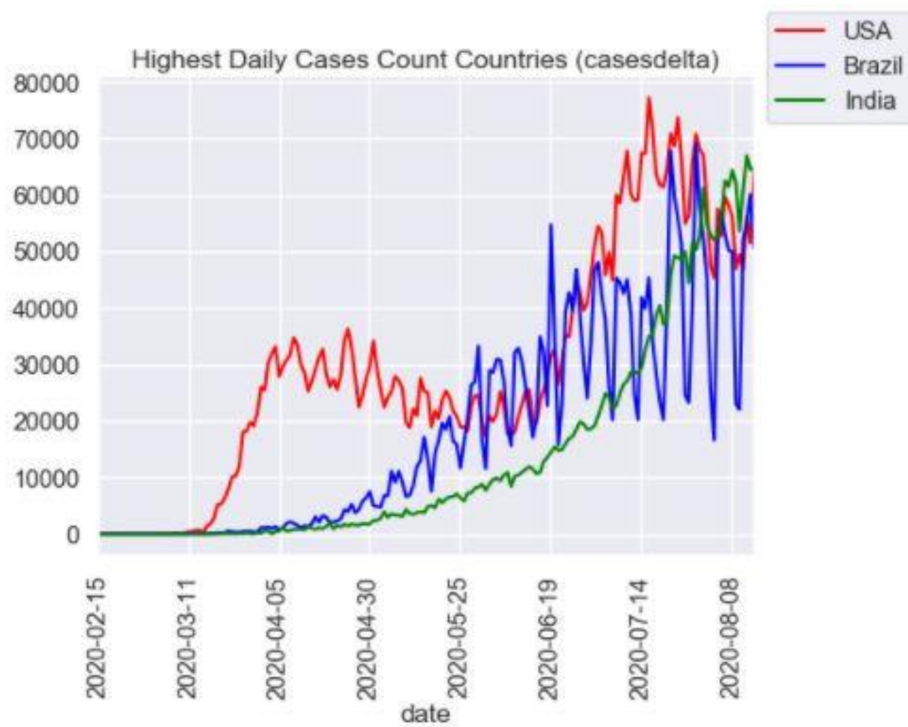


Figure 4: Daily Cases Counts for High Cases Countries

These are plots of the United States and Brazil Mobility data.

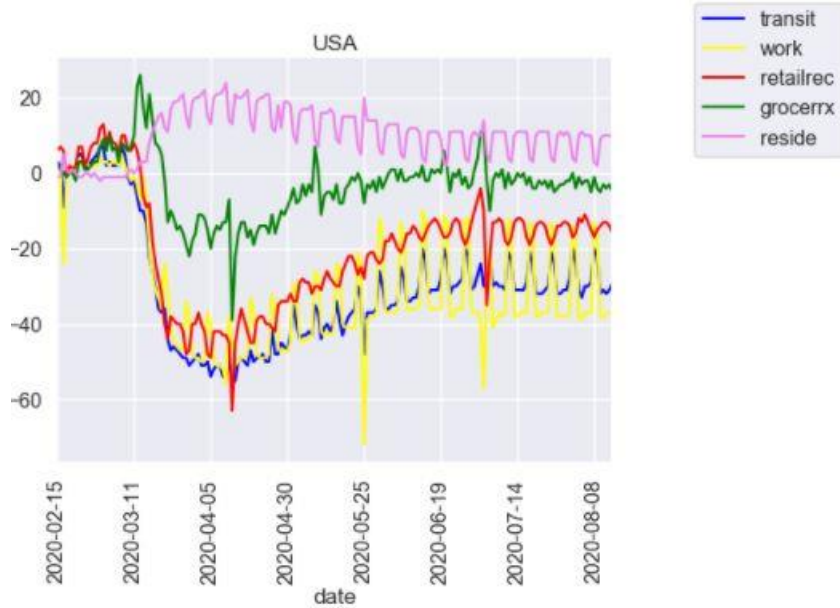


Figure 5: United States Mobility data - detail

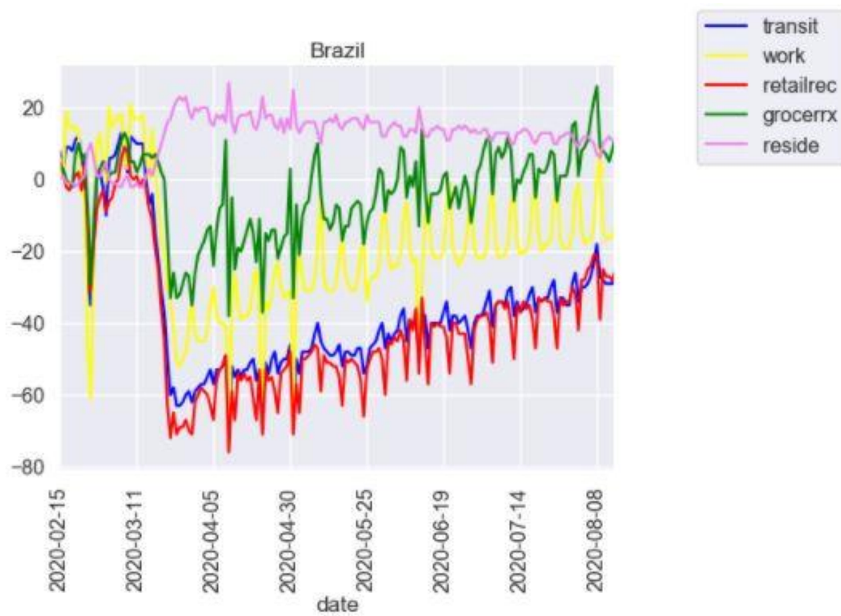


Figure 63: Brazil Mobility data – detail

This is a plot of the India Mobility data.

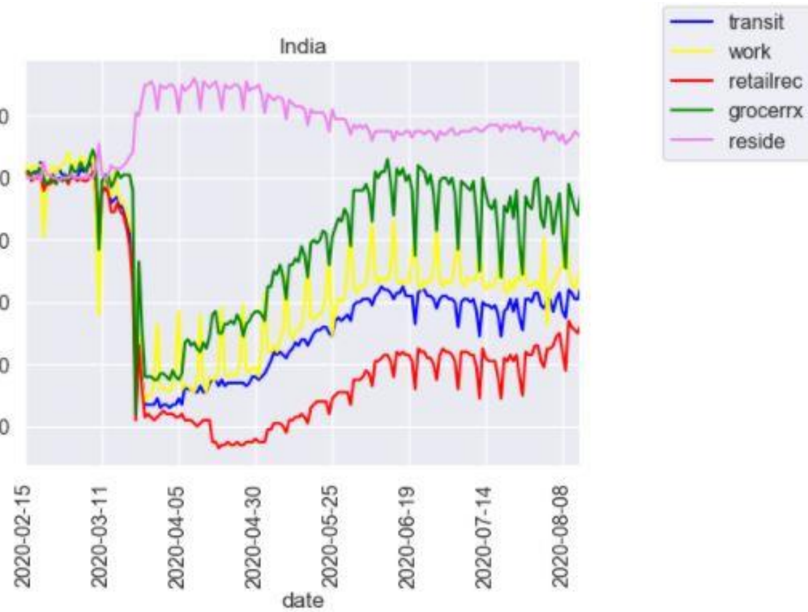


Figure 14: India Mobility data - detail

Modeling

I started modeling by investigating linear relationships using Linear Regression.

Initially I ran regression models against New Zealand data, using a 70/30 split of training and test sets using train_test_split. The outcome(dependant) variable was casesdelta14, the 14 day time lagged daily cases field discussed in Appendix B. The predictor(independent) variable was retailrec. The mean absolute percent errors were around 40%, the statistical measure of how accurate a forecast system is. I determined that I needed to investigate countries further and hopefully locate a better training country.

Regression Model	CVrmse	30%rmse	Gradient	R Value	Rsquared	Rsquared adj	MAPE
Simple Linear	0.312684	0.443399	1.0132	0.7809	0.6098	0.5448	42.5016
Lasso	0.311764	0.442803	1.0183	0.7824	0.6121	0.5474	42.4198
Random Forest	0.325851	0.411909	1.0069	0.8199	0.6722	0.6176	36.4441
SVR (scaling)	0.303993	0.429322	0.9556	0.8187	0.6702	0.6153	40.697

MAPE - mean absolute percent error

A regression model performs better for normally distributed data. For my three best case countries with low COVID-19 case count. The Canada and Singapore data was less skewed, had better correlations, and those countries also had larger populations.

NZ Skewness = 2.7277093749925165
NZ Kurtosis = 6.400003898979181
SI Skewness = 1.0719894512257258
SI Kurtosis = 0.9086860716304752
CA Skewness = 1.008871156124238
CA Kurtosis = 0.7124617760906453

Below is the distribution and probability plot for New Zealand and Singapore.

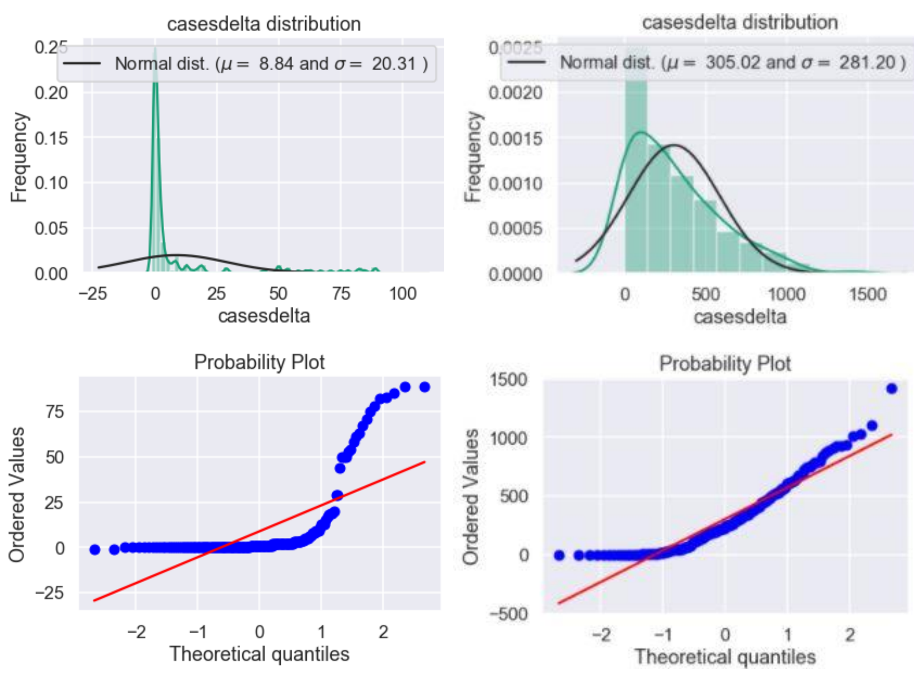


Figure 7: Distribution & Probability plot, New Zealand (left), Singapore (right)

Below are the Seaborn pairplots for the Canada and Singapore, which show the histograms and scatterplots for all the COVID-19 and mobility fields. The mobility columns in general display a strong a strong positive correlation.

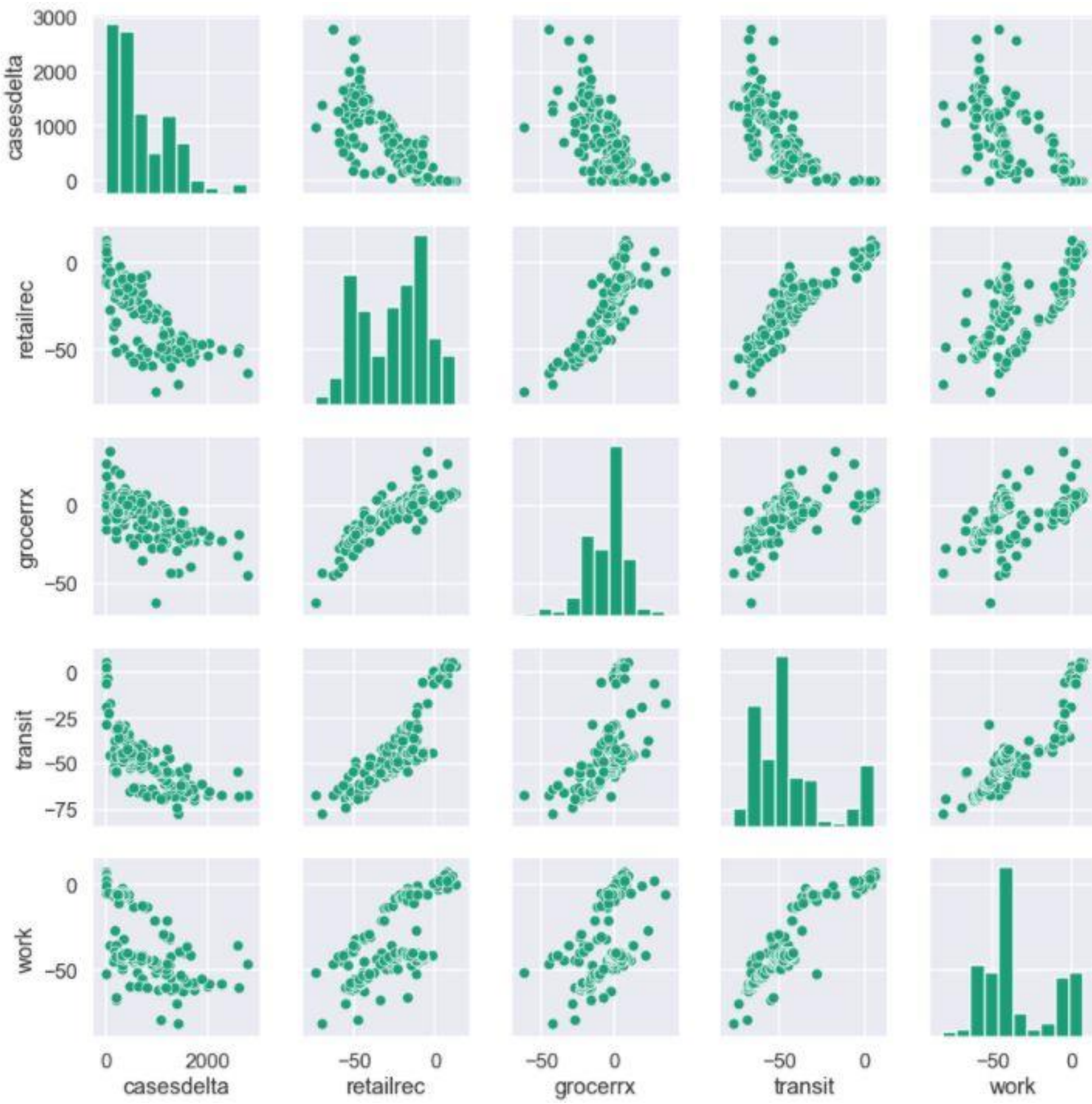


Figure 85: Canada Seaborn pairplot

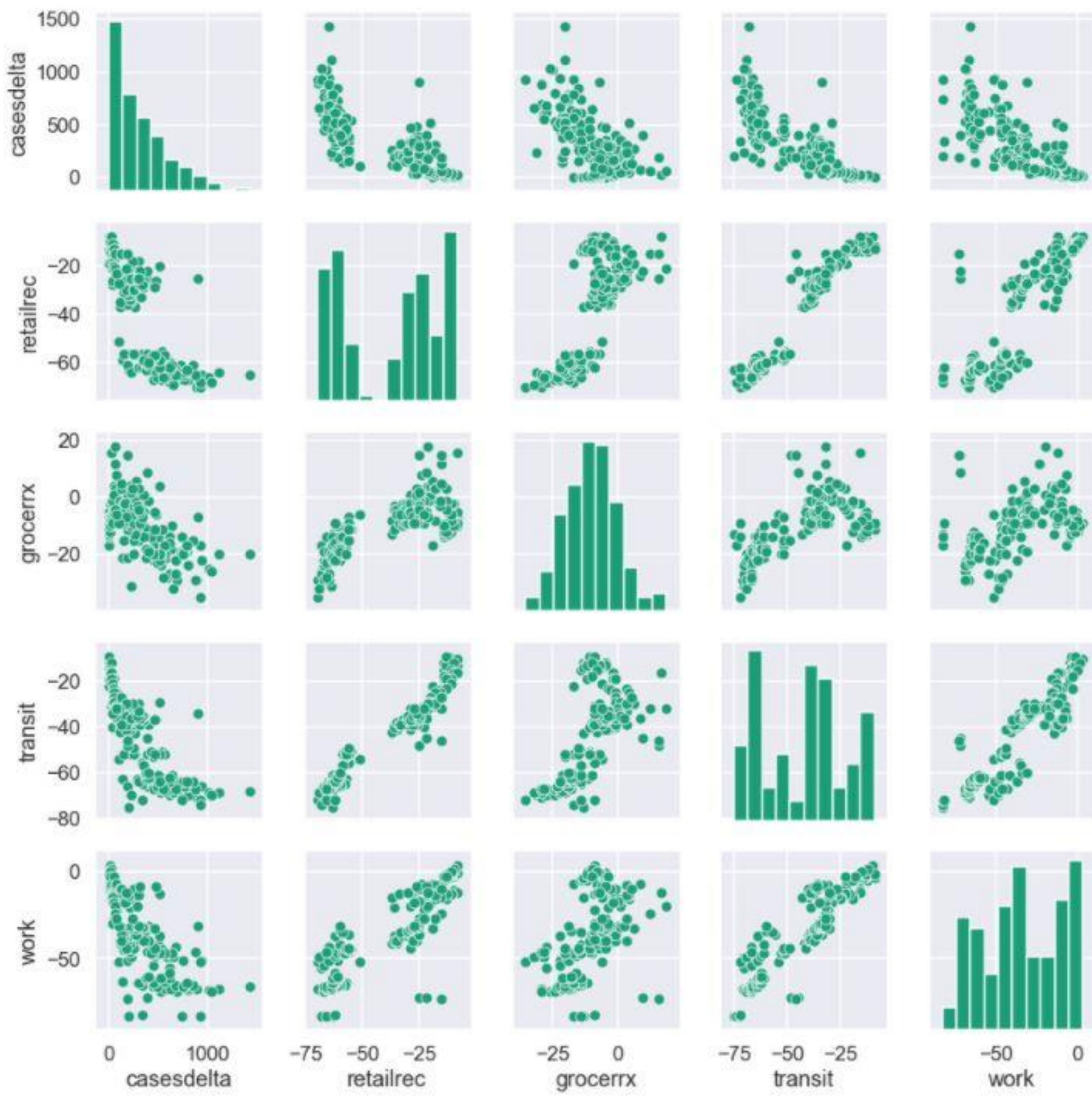


Figure 16: Singapore seaborn pairplot

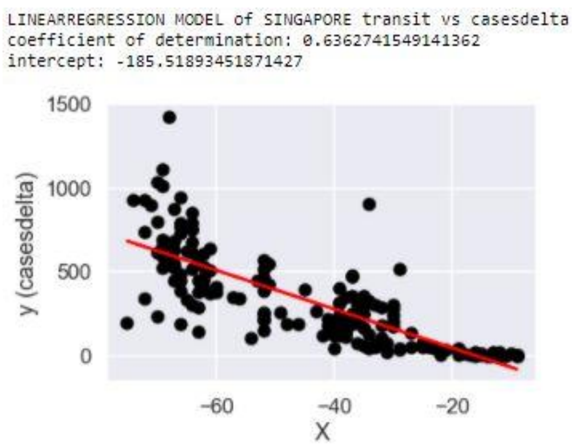
Modeling Linear Regression - use k-folds cross-validation (k=3)

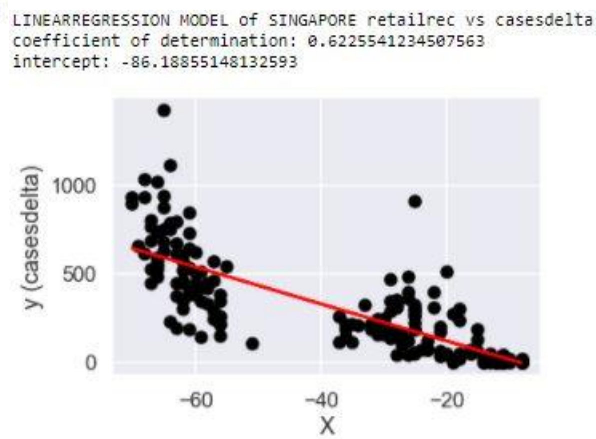
I proceeded to use Linear Regression with k-folds cross-validation (k=3) to assess the performance of the model. The output (predictor) is casesdelta or Daily cases. The mobility inputs (regressors) vary. Sklearn returns the R^2 score which is the coefficient of determination.

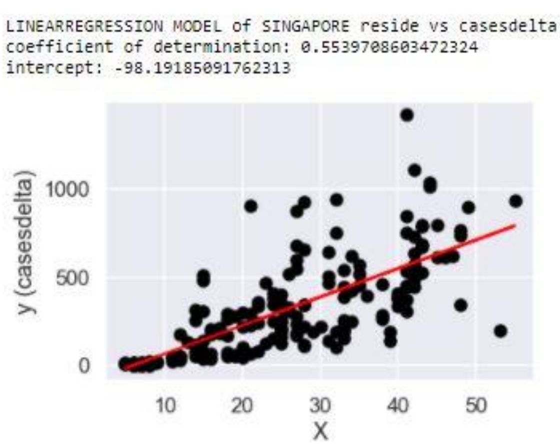
	k1 score	k2 score	k3 score	R^2	intercept
retailrec	0.498752188	0.741077868	0.668348998	0.668348998	-92.46516179
transit	0.513867713	0.729024465	0.68726865	0.68726865	-184.2955449
work	0.447919134	0.578967896	0.598229986	0.598229986	0.027562
reside	0.465248083	0.631974778	0.58525867	0.58525867	-96.99773542

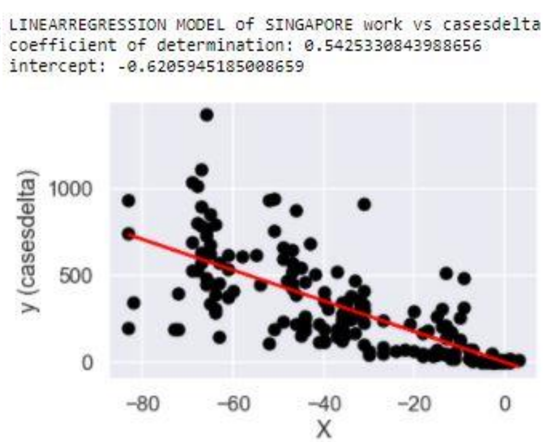
Modeling Linear Regression – Visualization

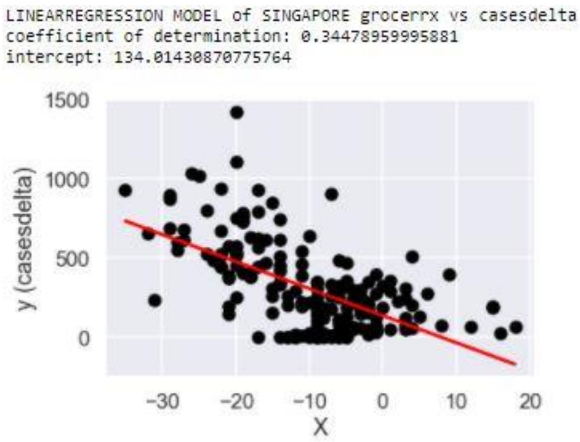
This section displays the model prediction line, as well as the R2 coefficient of determination, for Singapore, for the mobility fields transit, retailrec, transit, reside, work, grocerrx.











Modeling Linear Regression – train_test_split() creating train error

The project ended before I could repeat the original set of New Zealand regression models for Singapore or Canada. After creating the training and test data

```
X_train, X_test, y_train, y_test =
            train_test_split(X, y, test_size=0.3, random_state=0)
```

I ran into a common error

```
ValueError: Input contains NaN, infinity or a value too large for dtype('floa
t64').
```

while creating the training and testing set using

```
reg.fit(X_train, y_train)
```

Conclusions

I was not able to prove or disprove my hypothesis during the timeframe of this project. I made several iterations through from analysis to linear regression, each time discovering a new data wrangling or modeling preparation step. I encountered gaps in my python library knowledge. I also encountered gaps in my understanding of regression and modeling, and I was not able to perform time series regression or null hypothesis testing.

Visualization plots of COVID-19 and mobility data for the best, ‘good’ countries like New Zealand and Singapore does confirm the theory that strict mobility lockdowns at the start of COVID-19 case increases, which are maintained for at least 2 weeks AFTER daily case counts are stabilized a 0, does result in maintaining some control of the spread or the virus. However even these countries are now experiencing a second wave of cases, although it is small. Inversely, countries which neglected to restrict mobility, or removed restrictions at the first sign of daily counts decreasing, are seeing a much higher incidence of spread.

Issues I would consider at the end of this project are

1. Increasing my knowledge of python and the libraries I explored in this project; numpy, statistics, sklearn. A comment was made concerning ‘tool creep’ during the class discussions, and I found that to be an issues.
2. Understanding smoothing and moving averages, and handling testing 2 week incubation time lags.
3. Incorporating population and per capita data.
4. Learning how to handle multiple variable regression, and time series regression.

This live, real-time data raised some interesting questions.

1) Cell phone ownership demographics. India currently has outbreaks in extremely poor, densely packed slums. Cell phone location data is likely not reporting these impoverished and high risk locations.

2) Global differences in data collection and reporting. Even within the United States, even within the state of Colorado, there were differences in how cases, deaths and recovered cases were reported.

3) Indications are that the infections are much higher than those recorded. Antibody testing in Delhi, Mumbai and Pune India indicate the 23% of the populations had the antibodies so had been infected at one point. [AlJazeera]

4) There are other factors than mobility within a country. Taiwan is the example.[Time]. Taiwan has low cases, but it did not participate in a typical lockdown with restricted internal mobility. They wore masks, closed borders, enforced quarantines, followed the advise of their medical officials and kept their businesses and retail open.

When I began this project eight weeks ago, I knew this would be a complicated supervised learning regression task and very challenging with my introductory data science skills. I was motivated to investigate location data and cell phone mobility data. I learned what my weaknesses are, and what I need to pursue. I deliberately decided to tackle the project with the data science skills I had. There is a large community of shared projects available for me to pursue this further.

References

"COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University". https://github.com/CSSEGISandData/COVID-19

https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/

time_series_covid19_confirmed_global.csv

time_series_covid19_deaths_global.csv

time_series_covid19_recovered_global.csv

Google LLC *"Google COVID-19 Community Mobility Reports"*. https://www.google.com/covid19/mobility/, Accessed: August 17, 2020 https://www.gstatic.com/covid19/mobility/Global_Mobility_Report.csv?cachebust=9e64e67cbad15dab

https://www.gstatic.com/covid19/mobility/Global_Mobility_Report.csv

[AlJazeera] https://www.aljazeera.com/news/2020/08/millions-infected-coronavirus-delhi-survey-200821102647796.html

[Time] Bremmer, Ian (June 12, 2020) 'The Best Global Responses to COVID-19 Pandemic', https://time.com/5851633/best-global-responses-covid-19/

Agarrwal, Abhishek, November 11, 2019 "Time Series Analysis in Python | Time Series Forecasting Project [Complete] | Python Data Science"

Data Science Tutorials, https://www.youtube.com/watch?v=MmC4b7gPY0Q

https://benalexkeen.com/correlation-in-python/

https://blog.google/technology/health/covid-19-community-mobility-reports/

https://datascience.stackexchange.com/questions/68291/python-and-gridsearchcv-how-to-eliminate-input-contains-nan-error-when-using-cro

https://ourworldindata.org/covid-mobility-trends

https://pythonbasics.org/seaborn-pairplot/

https://realpython.com/linear-regression-in-python/#simple-linear-regression

https://scikit-learn.org/stable/modules/cross_validation.html#cross-validation

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html#sklearn.linear_model.LinearRegression.score

https://stats.stackexchange.com/questions/58391/mean-absolute-percentage-error-mape-in-scikit-learn

https://towardsdatascience.com/cleaning-financial-time-series-data-with-python-f30a3ed580b7

https://towardsdatascience.com/visualizing-data-with-pair-plots-in-python-f228cf529166

https://www.aljazeera.com/news/2020/08/millions-infected-coronavirus-delhi-survey-200821102647796.html

https://www.forbes.com/sites/johnkoetsier/2020/06/05/the-100-safest-countries-in-the-world-for-covid-19/#795f154f68c5

https://www.kaggle.com/c/house-prices-advanced-regression-techniques#evaluation

https://www.kaggle.com/nickelkumawat/linear-regression-house-prices

https://www.ritchieng.com/pandas-variability/

Appendices

Appendix A. Mobility 6 High Level Categories

- Grocery & pharmacy - places like grocery markets, food warehouses, farmers markets, specialty food shops, drug stores, and pharmacies. In many countries, this category was defined as essential and remained accessible during a lockdown.
- Transit stations - places like public transport hubs such as subway, bus, and train stations. May include seaports, taxi stands, highway rest stops, car rental agencies.
- Retail & recreation - places like restaurants, cafes, shopping centers, theme parks, museums, libraries, and movie theaters.
- Workplaces - places of work. A region's demographic data will explain the range of jobs. For example, does your region contain workplaces that don’t allow mobile devices such as government buildings or military bases
- Residential - places of residence. This shows a change in duration, where the other categories measure a change in total visitors. Because people already spend much of the day at places of residence the capacity for change isn’t so large. Do not compare the change in the Residential category with other categories because they have different units of measurement.
- Parks - places like local parks, national parks, public beaches, marinas, dog parks, plazas, and public gardens. May include public garden, castles, national forests, campgrounds, observation decks. Park visits are heavily influenced by the weather and normally very variable, so it will provide dramatic changes.

Appendix B. 14 Day Shift CasesDelta Column

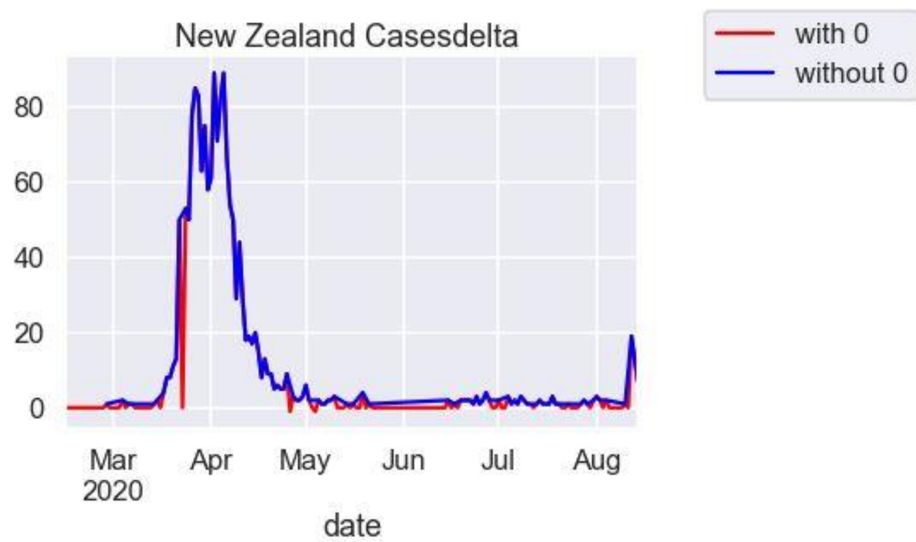
Because there is a two week incubation period for the virus, I also generated a 14 day shift on the casesdelta column to compare the correlations against.

For New Zealand the column casesdelta14 did show a slight increase the inverse correlation compared to casesdelta. And as expected deathsdelta14 also increased slightly more than deathsdelta. I did not retain these fields for the remaining investigations.

New Zealand	Correlation			
	casesdelta	**casesdelta14**	**deathsdelta**	**deathsdelta14**
cases	-0.202229	0.143246	0.100808	0.142596
deaths	-0.475436	-0.220855	-0.055873	0.127605
casesdelta	1	0.297623	0.068413	-0.07011
deathsdelta	0.068413	0.471681	1	0.153376
casesdelta14	0.297623	1	0.471681	0.071348
deathsdelta14	-0.07011	0.071348	0.153376	1
retailrec	-0.571267	-0.655763	-0.405455	-0.335203
grocerrx	-0.555312	-0.625061	-0.365498	-0.215792
transit	-0.525784	-0.590206	-0.360879	-0.272926
work	-0.615183	-0.673962	-0.402097	-0.275587
parks	-0.536743	-0.59014	-0.368937	-0.204813
reside	0.614441	0.680643	0.428117	0.291261

Appendix C. Effect of filtering 0 casesdelta

In the best case scenario, a country daily case count would contain many 0 values. New Zealand was the only country in my initial set which had this behavior. I plotted the difference in New Zealand casesdelta.



I also compared correlations for New Zealand, Singapore, and Canada with and without 0 values.

Corr(pearson)	NZ	NZ no 0	SI	SI no 0	CA	CA no 0
casesdelta	1.000	1.000	1.000	1.000	1.000	1.000
retailrec	-0.571	-0.553	-0.789	-0.776	-0.782	-0.769
grocerrx	-0.555	-0.538	-0.587	-0.602	-0.683	-0.675
transit	-0.526	-0.552	-0.798	-0.790	-0.730	-0.718
work	-0.615	-0.601	-0.737	-0.719	-0.640	-0.610

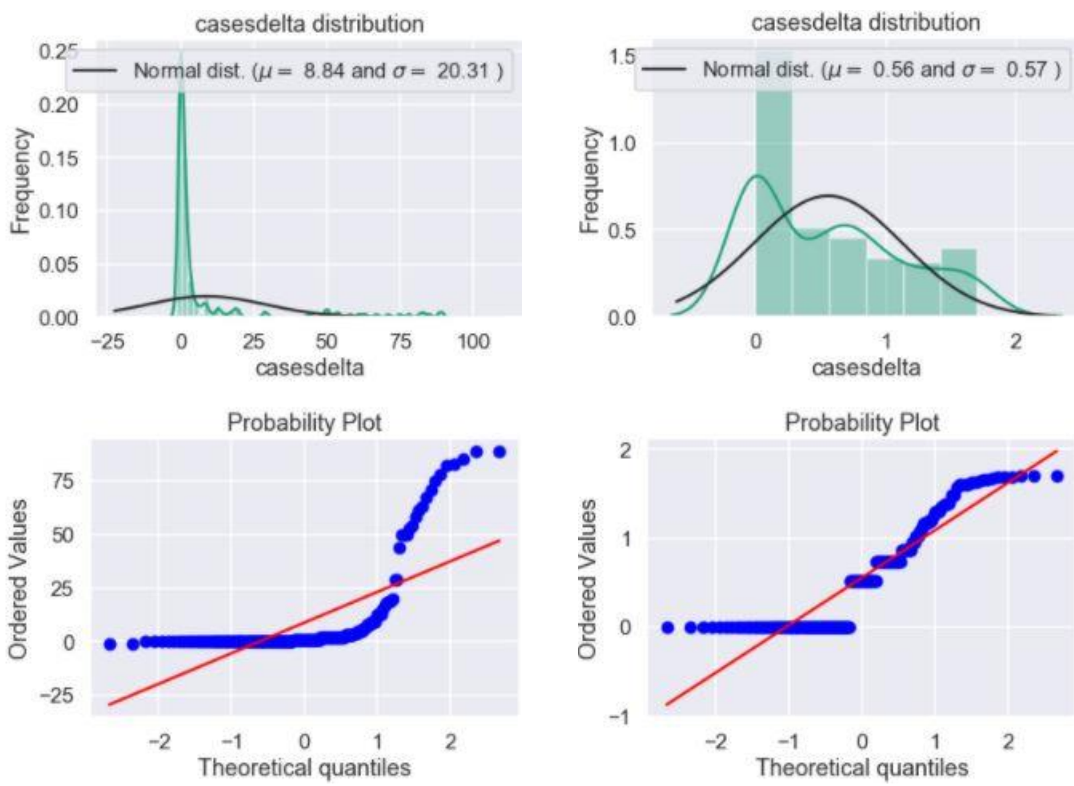
When I was wrapping up my report, I observed that many COVID-19 visualization sites start using a countries datasets from a minimum baseline of total or daily cases. This is especially true for projects where they are comparing data across countries, and want the data to be relative to a certain caseload start point and independent of a calendar date .

Appendix D. Effect of log transform on distribution

A regression model performs better with normally distributed data. The casesdelta column distributions were right skewed. A variable transform can diminish this difference and transform the data closer to a normal distribution. I tried applying a log transform, and observed the distribution changes. But how would the data transform alter the predictions? I was not able to finish this test before the deadline. Code was influenced by

https://www.kaggle.com/nickelkumawat/linear-regression-house-prices

BEFORE AFTER



Scripts, see github

Prac1.pdf	Presentation paper
DunneM-Practicum1-CovidMobility.ppt	presentation powerpoint
Mergefile_compare_countries.csv	compare of covid vs mobility country coverage
Practicum_Clean_Merge.ipynb	covid and mobility files data wrangling and merge
	Saved into clean file *PracticumMergedData.csv*
PracticumMergedData0817.csv	clean file used for presentation
Practicum_Analysis.ipynb	merged file analysis code notebook
Practicum_Model.ipynb	merged file modeling code notebook
Practicum_Asides.ipynb	analysis side testing notebook