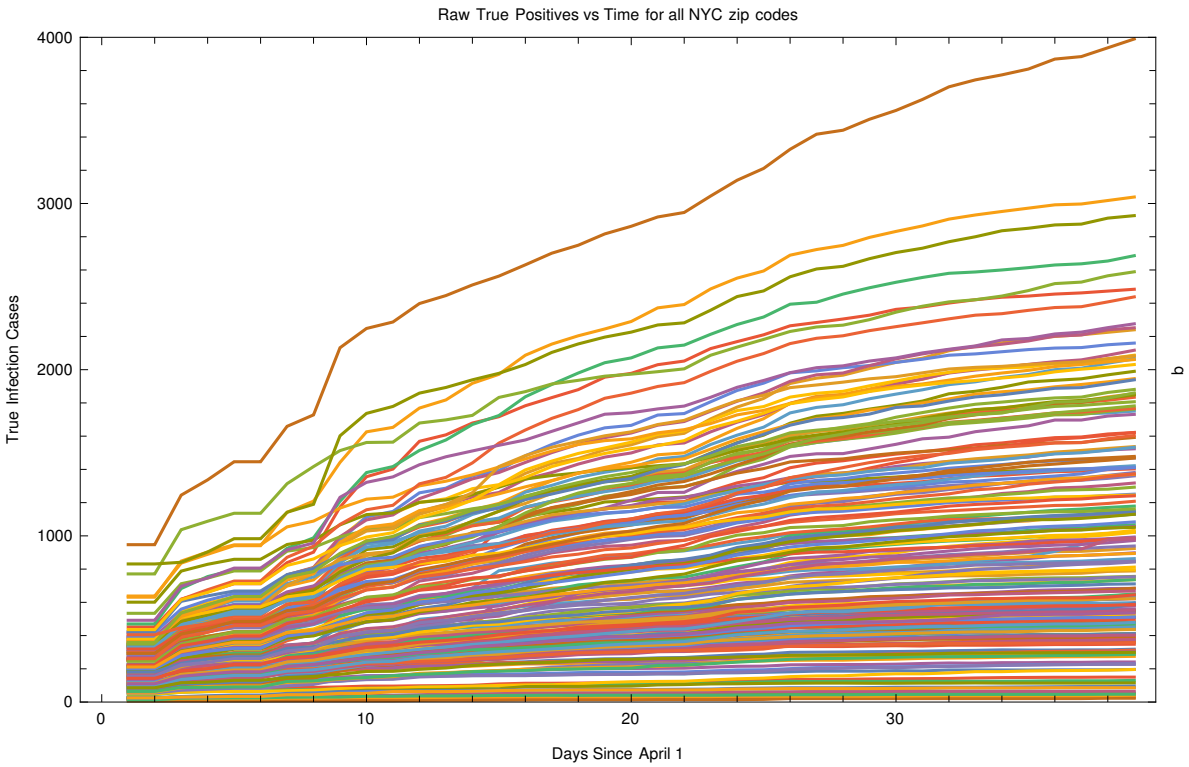
Analysis flow for calculation of Infection Rates

The analysis flow starts from the NYCHealth GITHUB site: https://github.com/nychealth/coronavirus-data.
For this analysis we obtained a copy of all the commits so that we could reconstruct the time series of the true infections, effectively from April 1
through May 10 by zipcode.
Next we obtained the US census American Community Survey (ACS) from Google Big Query. This provided us with 229 variables organized by zipcode. The
sql script is shown below.

```
SELECT
*
FROM `bigquery-public-data.census_bureau_acs.zip_codes_2018_5yr`
WHERE geo_id in (
“10044”,”10065”,”10069”,”10075”,”10128”,”11009”,”11201”,”11385”,”11697”,”10280”,”10282”,”10312”,”10314”,”11004”,”11109”)
OR ( CAST(geo_id as INT64) >= 10001 AND CAST(geo_id as INT64) <= 10007 )
OR ( CAST(geo_id as INT64) >= 10009 AND CAST(geo_id as INT64) <= 10014 )
OR ( CAST(geo_id as INT64) >= 10016 AND CAST(geo_id as INT64) <= 10019 )
OR ( CAST(geo_id as INT64) >= 10021 AND CAST(geo_id as INT64) <= 10040 )
OR ( CAST(geo_id as INT64) >= 10301 AND CAST(geo_id as INT64) <= 10310 )
OR ( CAST(geo_id as INT64) >= 10451 AND CAST(geo_id as INT64) <= 10475 )
OR ( CAST(geo_id as INT64) >= 11101 AND CAST(geo_id as INT64) <= 11106 )
OR ( CAST(geo_id as INT64) >= 11203 AND CAST(geo_id as INT64) <= 11226 )
OR ( CAST(geo_id as INT64) >= 11228 AND CAST(geo_id as INT64) <= 11239 )
OR ( CAST(geo_id as INT64) >= 11354 AND CAST(geo_id as INT64) <= 11358 )
OR ( CAST(geo_id as INT64) >= 11360 AND CAST(geo_id as INT64) <= 11370 )
OR ( CAST(geo_id as INT64) >= 11372 AND CAST(geo_id as INT64) <= 11375 )
OR ( CAST(geo_id as INT64) >= 11377 AND CAST(geo_id as INT64) <= 11379 )
OR ( CAST(geo_id as INT64) >= 11411 AND CAST(geo_id as INT64) <= 11423 )
OR ( CAST(geo_id as INT64) >= 11426 AND CAST(geo_id as INT64) <= 11429 )
OR ( CAST(geo_id as INT64) >= 11432 AND CAST(geo_id as INT64) <= 11436 )
OR ( CAST(geo_id as INT64) >= 11691 AND CAST(geo_id as INT64) <= 11694 )
ORDER BY  CAST(geo_id as INT64) asc
```

Lastly some Information per zipcode (such as the zip code area) was obtained from https://www.zip-codes.com/.

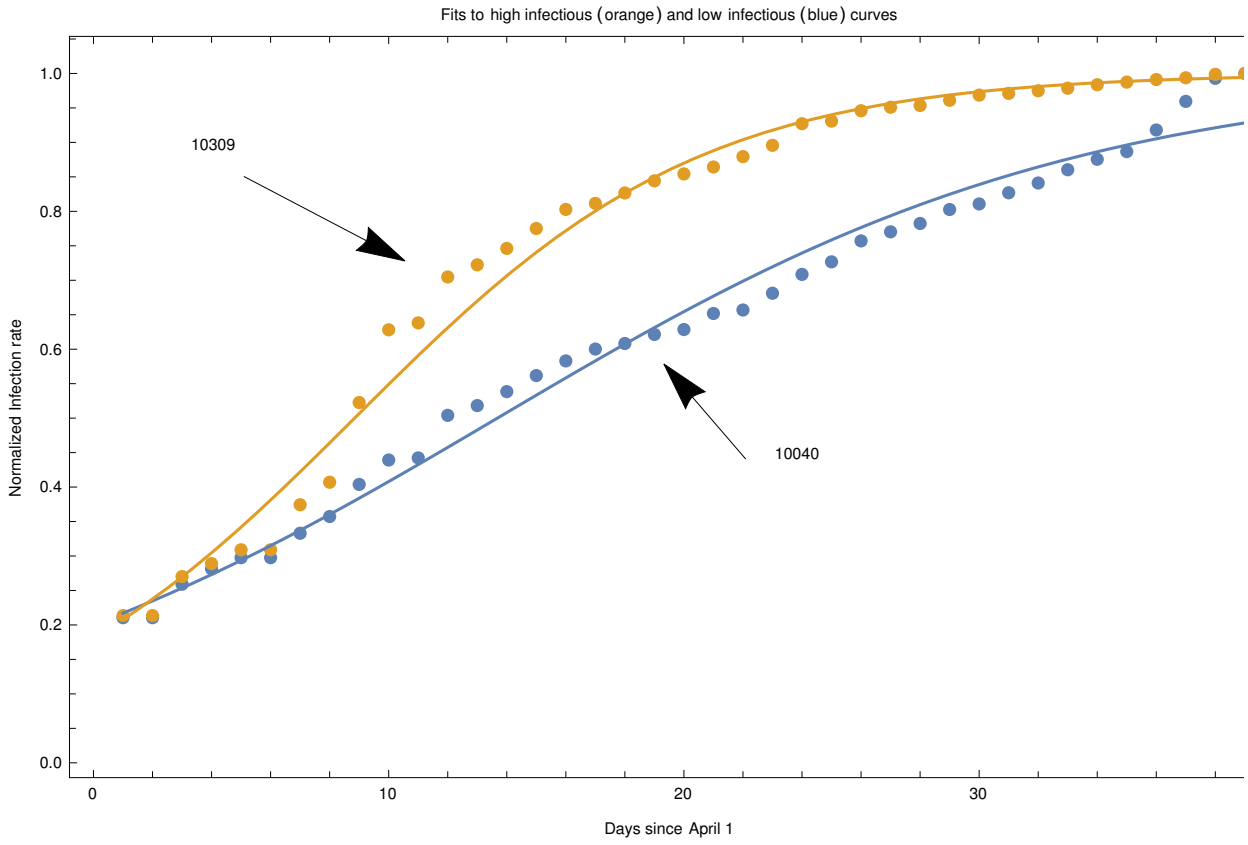
The data set shows a wide variation due to widely different populations and testing rates per zip code.



One approach is to normalize the curves, by dividing by the maximum rates of each time series. This enables all curves to be fit by a logistic regression by a function of the form $f(x)=\frac{1}{1+e^{(a-b*x)}}$, where a and b are fit parameters. The b parameter is used to indenty the infection rate, a higher number leasds to steeper curves.

To insure good fits, time series under consideration were required to have at least 625 counts. Shown below are examples of

a high infection rate (orange) and low infection (blue) rate time series.



To perform the analysis, a set of 30 zip codes with the highest infection rates and 30 zip codes with lowest infection rates are selected from all the fits performed. In addition,
two randomly selected sets of 30 zip codes each are used as the control set. The random set allows us to assess the effects of a null feature. The high and low infection
rate zipcodes are:

High infection rate zip codes.

```
{10454, 10467, 10472, 11429, 10029, 10460, 10035, 11235, 10463, 10473,
 11414, 11236, 10025, 11433, 10455, 10451, 10465, 10469, 11354, 10306,
 10301, 11234, 10305, 10310, 10475, 10466, 10314, 10312, 10303, 10309}
```

Low infection Rate zip codes

{10040,11211,11219,11367,11205,11230,11204,11368,11374,11423,11237,11216,11220,11238,10032,11206,11385,11213,10033,11432,11369,11218,11418,10002,11379,11434,11372,11207,11416,11221}

Random Selected low rate set
{11370,10032,11212,10458,11234,10033,11203,11224,10457,10475,11374,11414,11238,11694,10460,10459,11418,11355,11420,11204,11379,11208,11412,10303,11357,11367,11434,11354,10455,10467}

Random Selected high rate set.
{10463,11211,11433,10466,10452,10453,11373,10009,11218,10310,11369,11421,11233,11221,11385,11207,10035,10461,10305,11422,10465,11226,10314,10027,10472,10469,11210,10309,11372,10304}

Feature selection on the United States census ACS.

The American community survey is a detailed addition to the standard census. It contains over 200 features partitioned by zip code. The approximate distribution is as follows:

Category	Percent
Demographics	30
Financial	10
Housing	25
Commute	10
Education	10
Employment	10

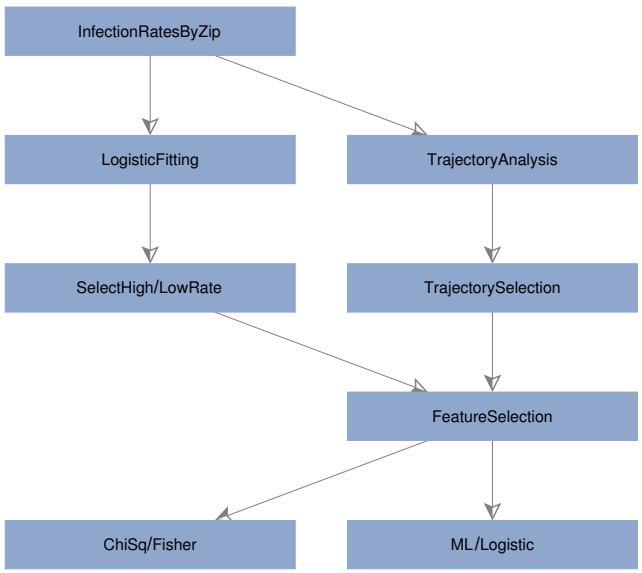
Initially all 229 pairs were compared against each other using the PearsonChiSquared test (this essentially computes a p-value for each feature). The infection set produced consistently higher sets of contrasting features between the two sets than the random selected features. There were 11 vs 5 identified features for 0.05 p value cutoff and 24 vs 8 for p-value 0.1 cutoff.For the $p < 0.1$ set, this variation between infectioon and random cannot be explained by chance. The ChiSquared analysis suggests the following variables are worth exploring

Demographic	Housing	Commute	Employment
geo_id	renter_occupied_housing_units_paying_cash_median_gross_rent	commute_35_44_mins	employed_wholesale_trade
male_under_5	owner_occupied_housing_units_median_value	walked_to_work	occupation_management_arts
male_65_to_66	owner_occupied_housing_units_upper_value_quartile	worked_at_home	management_business_sci_arts_employed
male_70_to_74	vacant_housing_units_for_sale	no_car	
female_80_to_84	dwellings_3_to_4_units	no_cars	
asian_male_55_64	dwellings_5_to_9_units		
	median_year_structure_built		
	male_male_households		
	median_rent		
	million_dollar_housing_units		

In[27]:=

In[26]:=
```
LayeredGraphPlot[{"InfectionRatesByZip" → "LogisticFitting",
  "InfectionRatesByZip" → "TrajectoryAnalysis",
  "LogisticFitting" → "SelectHigh/LowRate", "SelectHigh/LowRate" → "FeatureSelection",
  "FeatureSelection" → "ChiSq/Fisher", "FeatureSelection" → "ML/Logistic",
  "TrajectoryAnalysis" -> "TrajectorySelection",
  "TrajectorySelection" -> "FeatureSelection"},
 PlotTheme → "DiagramBlue", VertexSize → {.4, .1}, AspectRatio → 1.5]
```

Out[26]=



LayeredGraphPlot: LayeredGraphPlot called with 4 arguments; 1 or 2 arguments are expected.