# Risk Determination of Exposure to COVID-19

**2020 City of LA and RMDS COVID-19 Computational Challenge**

**Team: Risk Detectives**

**Binlun Feng, Leon Lu, Lei Wang, You Wang, Jennifer Ware, Xiangbo Wen**

**Mentored by:**

**Dr. Yuanjie (Ed) He, Dr. Honggang Wang**

## Introduction

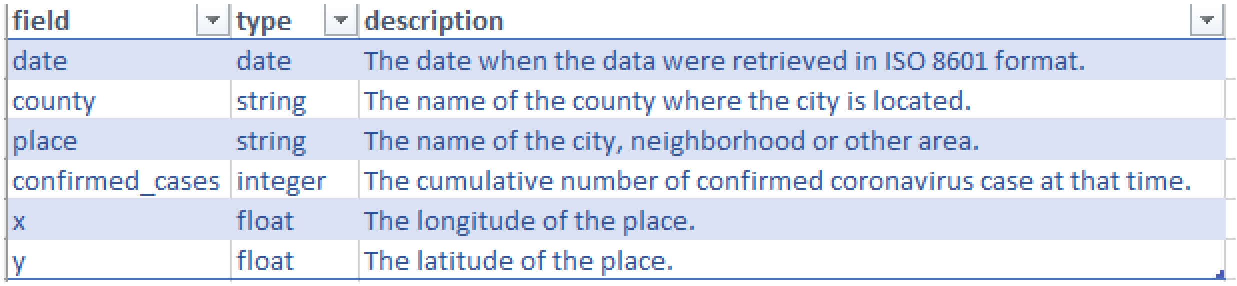
Through literacy research and logical analysis, the risk score for a specific area is determined as the potential increased COVID-19 cases per population density, by the formula: $risk\ score\ = \frac{Number\ of\ New\ Daily\ Cases*\ 10^{6}}{Population\ /\ Area\ size}$.

## Data

### 1. Data sources

Data is collected at the city or community level within Los Angeles County. This includes neighborhoods and communities within the city of Los Angeles and nearby cities inside Los Angeles County. Our data comes from four different sources:

- **Latimes-place-totals.csv**
  LA Times keeps track of COVID-19 in California from multiple health and city agencies. Updated cases are updated on a daily basis. The github repository collects data from LA Times beginning on March 16, 2020 to Present. This is the only source we found that has the information of COVID-19 at city and neighborhood level. From this dataset, we selected these features for our study:



| field | type | description |
|---|---|---|
| date | date | The date when the data were retrieved in ISO 8601 format. |
| county | string | The name of the county where the city is located. |
| place | string | The name of the city, neighborhood or other area. |
| confirmed_cases | integer | The cumulative number of confirmed coronavirus case at that time. |
| x | float | The longitude of the place. |
| y | float | The latitude of the place. |

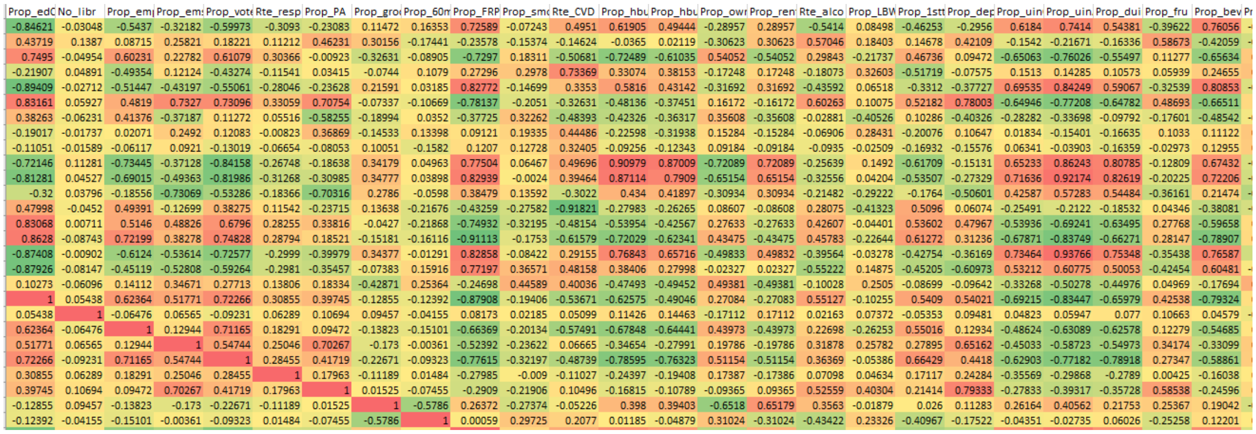
*GitHub link*

**Data handling:**

1. Filter out the records which are not in the Los Angeles County
2. Calculate 'new_cases' by subtracting 'confirmed_cases' on two consecutive days in a certain place
3. We believe that the risk level in one city is not isolated, meaning that it will be affected by the risk levels in the nearby cities. Therefore, we believe that the nearby new cases are relevant to this study. In order to get the 'nearby_cases,' we summed up 'x' and 'y' which gives us a general idea of the location. We then sorted 'x+y' and summed up five consecutive 'new cases' and assigned this sum value to 'nearby_cases'.

4. We determined to use the next day's new case for our risk score. We created a new column, ‘tmr_new_cases’ to store this value.

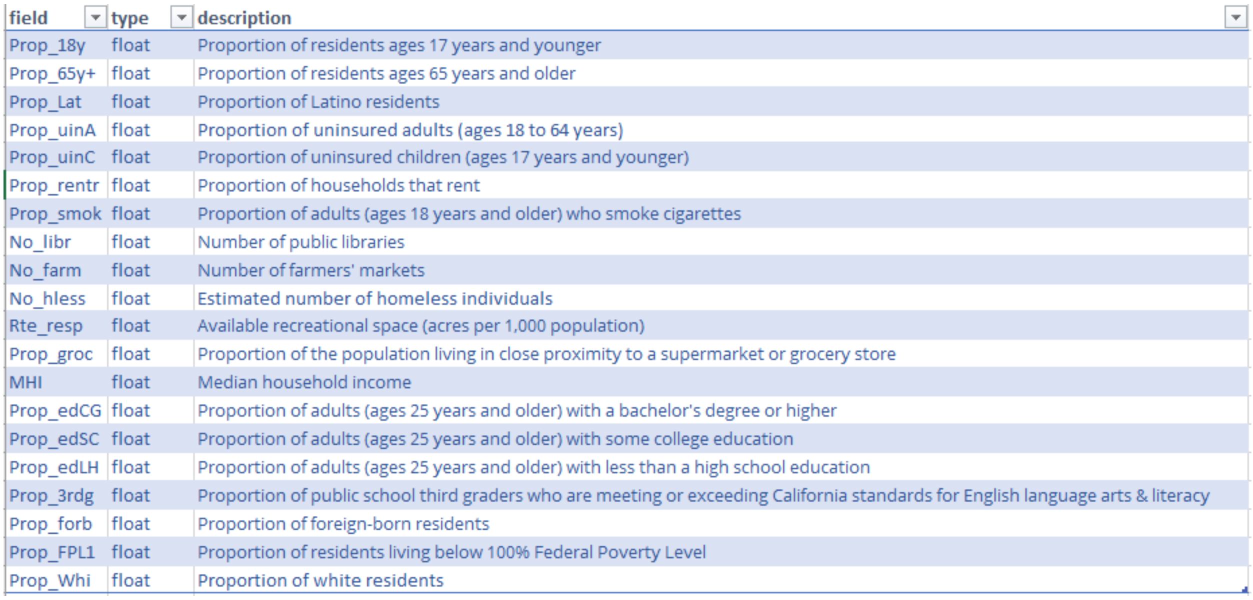
- **Los Angeles County City and Community Health Profiles 2018**
  The city and health profile is the next dataset we found accurate to the city and neighborhood level. This dataset is provided by the Los Angeles County Public Health Department and contains demographic, social, and health information. In the beginning, the dataset had over 80 columns. The heatmap listed below shows an overview of the features in this dataset that are highly correlated. In order to reduce redundancy and limitation of the data size, we did not use all the features provided by the dataset. Instead, we manually selected the features that we believed to be more relevant to our study.



| Prop_edC | No_libr | Prop_em | Prop_em: | Prop_vote | Rte_resp | Prop_PA | Prop_gro | Prop_60m | Prop_FRP | Prop_smo | Rte_CVD | Prop_hbu | Prop_hbu | Prop_owr | Prop_ren | Rte_alco | Prop_LBW | Prop_1stt | Prop_dep | Prop_uin | Prop_uin. | Prop_dui | Prop_fru | Prop_bev |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -0.84621 | -0.03048 | -0.5437 | -0.32182 | -0.59973 | -0.3093 | -0.23083 | 0.11472 | 0.16353 | 0.72589 | -0.07243 | 0.4951 | 0.61905 | 0.49444 | -0.28957 | 0.28957 | -0.5414 | 0.08498 | -0.46253 | -0.2956 | 0.6184 | 0.7414 | 0.54381 | -0.39622 | 0.76056 |
| 0.43719 | 0.1387 | 0.08715 | 0.25821 | 0.18221 | 0.11212 | 0.46231 | 0.30156 | -0.17441 | -0.23578 | -0.15374 | -0.14624 | -0.0365 | 0.02119 | -0.30623 | 0.30623 | 0.57046 | 0.18403 | 0.14678 | 0.42109 | -0.1542 | -0.21671 | -0.16336 | 0.58673 | -0.42059 |
| 0.7495 | -0.04954 | 0.60231 | 0.22782 | 0.61079 | 0.30366 | -0.00923 | -0.32631 | -0.08905 | -0.7297 | 0.18311 | -0.50681 | -0.72489 | -0.61035 | 0.54052 | -0.54052 | 0.29843 | -0.21737 | 0.46736 | 0.09472 | -0.65063 | -0.76026 | -0.55497 | 0.11277 | -0.65634 |
| -0.21907 | 0.04891 | -0.49354 | 0.12124 | -0.43274 | -0.11541 | 0.03415 | -0.0744 | 0.1079 | 0.27296 | 0.2978 | 0.73369 | 0.33074 | 0.38153 | -0.17248 | 0.17248 | -0.18073 | 0.32603 | -0.51719 | -0.07575 | 0.1513 | 0.14285 | 0.10573 | 0.05939 | 0.24655 |
| -0.89409 | -0.02712 | -0.51447 | -0.43197 | -0.55061 | -0.28046 | -0.23628 | 0.21591 | 0.03185 | 0.82772 | -0.14699 | 0.3353 | 0.5816 | 0.43142 | -0.31692 | 0.31692 | -0.43592 | 0.06518 | -0.3312 | -0.37727 | 0.69535 | 0.84249 | 0.59067 | -0.32539 | 0.80853 |
| 0.83161 | 0.05927 | 0.4819 | 0.7327 | 0.73096 | 0.33059 | 0.70754 | -0.07337 | -0.10669 | -0.78137 | -0.2051 | -0.32631 | -0.48136 | -0.37451 | 0.16172 | -0.16172 | 0.60263 | 0.10075 | 0.52182 | 0.78003 | -0.64946 | -0.77208 | -0.64782 | 0.48693 | -0.66511 |
| 0.38263 | -0.06231 | 0.41376 | -0.37187 | 0.11272 | 0.05516 | -0.58255 | -0.18994 | 0.0352 | -0.37725 | 0.32262 | -0.48393 | -0.42326 | -0.36317 | 0.35608 | -0.35608 | -0.02881 | -0.40526 | 0.10286 | -0.40326 | -0.28282 | -0.33698 | -0.09792 | -0.17601 | -0.48542 |
| -0.19017 | -0.01737 | 0.02071 | 0.2492 | 0.12083 | -0.00823 | 0.36869 | -0.14533 | 0.13398 | 0.09121 | 0.19335 | 0.44486 | -0.22598 | -0.31938 | 0.15284 | -0.15284 | -0.06906 | 0.28431 | -0.20076 | 0.10647 | 0.01834 | -0.15401 | -0.16635 | 0.1033 | 0.11122 |
| -0.11051 | -0.01589 | -0.06117 | 0.0921 | -0.13019 | -0.06654 | -0.08053 | 0.10051 | -0.1582 | 0.1207 | 0.12728 | 0.32405 | -0.09256 | -0.12343 | 0.09184 | -0.09184 | -0.0935 | -0.02509 | -0.16932 | -0.15576 | 0.06341 | -0.03903 | -0.16359 | -0.02973 | 0.12955 |
| -0.72146 | 0.11281 | -0.73445 | -0.37128 | -0.84158 | -0.26748 | -0.18638 | 0.34179 | 0.04963 | 0.77504 | 0.06467 | 0.49696 | 0.90979 | 0.87009 | -0.72089 | 0.72089 | -0.25639 | 0.1492 | -0.61709 | -0.15131 | 0.65233 | 0.86243 | 0.80785 | -0.12809 | 0.67432 |
| -0.81281 | 0.04527 | -0.69015 | -0.49363 | -0.81986 | -0.31268 | -0.30985 | 0.34777 | 0.03898 | 0.82939 | -0.0024 | 0.39464 | 0.87114 | 0.7909 | -0.65154 | 0.65154 | -0.32556 | 0.04204 | -0.53507 | -0.27329 | 0.71636 | 0.92174 | 0.82619 | -0.20225 | 0.72206 |
| -0.32 | 0.03796 | -0.18556 | -0.73069 | -0.53286 | -0.18366 | -0.70316 | 0.2786 | -0.0598 | 0.38479 | 0.13592 | -0.3022 | 0.434 | 0.41897 | -0.30934 | 0.30934 | -0.21482 | -0.29222 | -0.1764 | -0.50601 | 0.42587 | 0.57283 | 0.54484 | -0.36161 | 0.21474 |
| 0.47998 | -0.0452 | 0.49391 | -0.12699 | 0.38275 | 0.11542 | -0.23715 | 0.13638 | -0.21676 | -0.43259 | -0.27582 | -0.91821 | -0.27983 | -0.26265 | 0.08607 | -0.08608 | 0.28075 | -0.41323 | 0.5096 | 0.06074 | -0.25491 | -0.2122 | -0.18532 | 0.04346 | -0.38081 |
| 0.83068 | 0.00711 | 0.5146 | 0.48826 | 0.6796 | 0.28255 | 0.33816 | -0.0427 | -0.21868 | -0.74932 | -0.32195 | -0.48154 | -0.53954 | -0.42567 | 0.27633 | -0.27633 | 0.42607 | -0.04401 | 0.53602 | 0.47967 | -0.53936 | -0.69241 | -0.63495 | 0.27768 | -0.59658 |
| 0.8628 | -0.08743 | 0.72199 | 0.38278 | 0.74828 | 0.28794 | 0.18521 | -0.15181 | -0.16116 | -0.91113 | -0.1753 | -0.61579 | -0.72029 | -0.62341 | 0.43475 | -0.43475 | 0.45783 | -0.22644 | 0.61272 | 0.31236 | -0.67871 | -0.83749 | -0.66271 | 0.28147 | -0.78907 |
| -0.87408 | -0.00902 | -0.6124 | -0.53614 | -0.72577 | -0.2999 | -0.39979 | 0.34377 | -0.01291 | 0.82858 | -0.08422 | 0.29155 | 0.76843 | 0.65716 | -0.49833 | 0.49832 | -0.39564 | -0.03278 | -0.42754 | -0.36169 | 0.73464 | 0.93766 | 0.75348 | -0.35438 | 0.76587 |
| -0.87926 | -0.08147 | -0.45119 | -0.52808 | -0.59264 | -0.2981 | -0.35457 | -0.07383 | 0.15916 | 0.77197 | 0.36571 | 0.48158 | 0.38406 | 0.27998 | -0.02327 | 0.02327 | -0.55222 | 0.14875 | -0.45205 | -0.60973 | 0.53212 | 0.60775 | 0.50053 | -0.42454 | 0.60481 |
| 0.10273 | -0.06096 | 0.14112 | 0.34671 | 0.27713 | 0.13806 | 0.18334 | -0.42871 | 0.25364 | -0.24698 | 0.44589 | 0.40036 | -0.47493 | -0.49452 | 0.49381 | -0.49381 | -0.10028 | 0.2505 | -0.08699 | -0.09642 | -0.33268 | -0.50278 | -0.44976 | 0.04969 | -0.17694 |
| 1 | 0.05438 | 0.62364 | 0.51771 | 0.72266 | 0.30855 | 0.39745 | -0.12855 | -0.12392 | -0.87908 | -0.19406 | -0.53671 | -0.62575 | -0.49046 | 0.27084 | -0.27083 | 0.55127 | -0.10255 | 0.5409 | 0.54021 | -0.69215 | -0.83447 | -0.65979 | 0.42538 | -0.79324 |
| 0.05438 | 1 | -0.06476 | 0.06565 | -0.09231 | 0.06289 | 0.10694 | 0.09457 | -0.04155 | 0.08173 | 0.02185 | 0.05099 | 0.11426 | 0.14463 | -0.17112 | 0.17112 | 0.02163 | 0.07372 | -0.05353 | 0.09481 | 0.04823 | 0.05947 | 0.077 | 0.10663 | 0.04579 |
| 0.62364 | -0.06476 | 1 | 0.12944 | 0.71165 | 0.18291 | 0.09472 | -0.13823 | -0.15101 | -0.66369 | -0.20134 | -0.57491 | -0.67848 | -0.64441 | 0.43973 | -0.43973 | 0.22698 | -0.26253 | 0.55016 | 0.12934 | -0.48624 | -0.63089 | -0.62578 | 0.12279 | -0.54685 |
| 0.51771 | 0.06565 | 0.12944 | 1 | 0.54744 | 0.25046 | 0.70267 | -0.173 | -0.00361 | -0.52392 | -0.23622 | 0.06665 | -0.34654 | -0.27991 | 0.19786 | -0.19786 | 0.31878 | 0.25782 | 0.27895 | 0.65162 | -0.45033 | -0.58723 | -0.54973 | 0.34174 | -0.33099 |
| 0.72266 | -0.09231 | 0.71165 | 0.54744 | 1 | 0.28455 | 0.41719 | -0.22671 | -0.09323 | -0.77615 | -0.32197 | -0.48739 | -0.78595 | -0.76323 | 0.51154 | -0.51154 | 0.36369 | -0.05386 | 0.66429 | 0.4418 | -0.62903 | -0.77182 | -0.78918 | 0.27347 | -0.58861 |
| 0.30855 | 0.06289 | 0.18291 | 0.25046 | 0.28455 | 1 | 0.17963 | -0.11189 | 0.01484 | -0.27985 | -0.009 | -0.11027 | -0.24397 | -0.19408 | 0.17387 | -0.17386 | 0.07098 | 0.04634 | 0.17117 | 0.24284 | -0.35569 | -0.29868 | -0.2789 | 0.00425 | -0.16038 |
| 0.39745 | 0.10694 | 0.09472 | 0.70267 | 0.41719 | 0.17963 | 1 | 0.01525 | -0.07455 | -0.2909 | -0.21906 | 0.10496 | -0.16815 | -0.10789 | -0.09365 | 0.09365 | 0.52559 | 0.40304 | 0.21414 | 0.79333 | -0.27833 | -0.39317 | -0.35728 | 0.58538 | -0.24596 |
| -0.12855 | 0.09457 | -0.13823 | -0.173 | -0.22671 | -0.11189 | 0.01525 | 1 | -0.5786 | 0.26372 | -0.27374 | -0.05226 | 0.398 | 0.39403 | -0.6518 | 0.65179 | 0.3563 | -0.01879 | 0.026 | 0.11283 | 0.26164 | 0.40562 | 0.21753 | 0.25367 | 0.19042 |
| -0.12392 | -0.04155 | -0.15101 | -0.00361 | -0.09323 | 0.01484 | -0.07455 | -0.5786 | 1 | 0.00059 | 0.29725 | 0.2077 | 0.01185 | -0.04879 | 0.31024 | -0.31024 | -0.43422 | 0.23326 | -0.40967 | -0.17522 | -0.04351 | -0.02735 | 0.06026 | -0.25258 | 0.12201 |

*--Segment of Health Profile correlation heatmap*

The features we manually selected for this model are listed below:



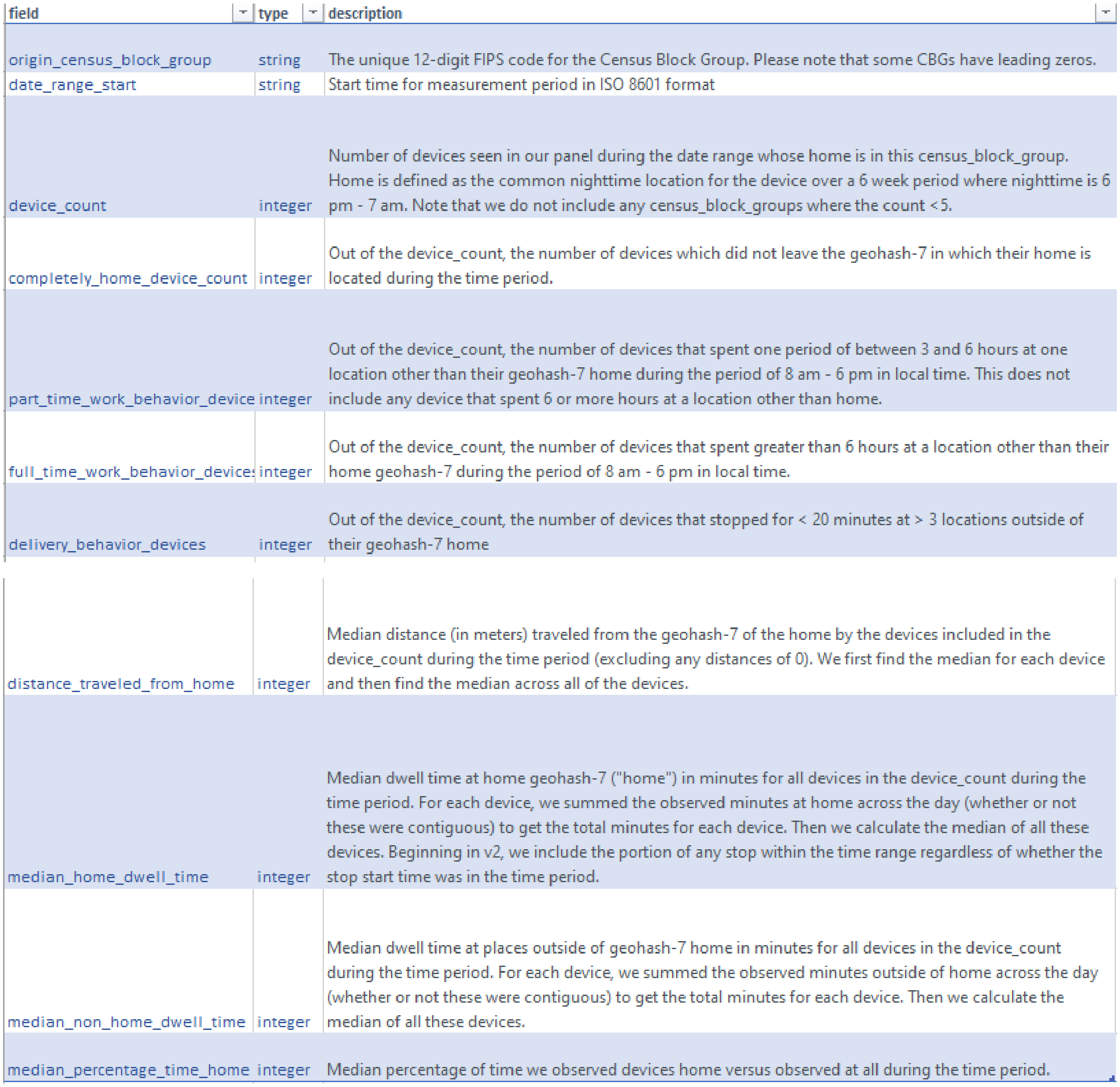
| field | type | description |
|---|---|---|
| Prop_18y | float | Proportion of residents ages 17 years and younger |
| Prop_65y+ | float | Proportion of residents ages 65 years and older |
| Prop_Lat | float | Proportion of Latino residents |
| Prop_uinA | float | Proportion of uninsured adults (ages 18 to 64 years) |
| Prop_uinC | float | Proportion of uninsured children (ages 17 years and younger) |
| Prop_rentr | float | Proportion of households that rent |
| Prop_smok | float | Proportion of adults (ages 18 years and older) who smoke cigarettes |
| No_libr | float | Number of public libraries |
| No_farm | float | Number of farmers' markets |
| No_hless | float | Estimated number of homeless individuals |
| Rte_resp | float | Available recreational space (acres per 1,000 population) |
| Prop_groc | float | Proportion of the population living in close proximity to a supermarket or grocery store |
| MHI | float | Median household income |
| Prop_edCG | float | Proportion of adults (ages 25 years and older) with a bachelor's degree or higher |
| Prop_edSC | float | Proportion of adults (ages 25 years and older) with some college education |
| Prop_edLH | float | Proportion of adults (ages 25 years and older) with less than a high school education |
| Prop_3rdg | float | Proportion of public school third graders who are meeting or exceeding California standards for English language arts & literacy |
| Prop_forb | float | Proportion of foreign-born residents |
| Prop_FPL1 | float | Proportion of residents living below 100% Federal Poverty Level |
| Prop_Whi | float | Proportion of white residents |

County of Los Angeles Open Data

- **Social Distancing Metrics**

  SafeGraph provides useful foot traffic data, which we used in our study. Specifically, SafeGraph uses panels to detect GPS pings from anonymous mobile devices in different census block groups during the quarantine period. Due to the large size of the data and limited computing capacity, we decided to use the datasets from May 2020. Based on the scope of our study, we filtered out the locations outside LA County as well as removed the columns that display the hourly level.

  The features we selected from the SafeGraph dataset are listed below:



| field | type | description |
|---|---|---|
| origin_census_block_group | string | The unique 12-digit FIPS code for the Census Block Group. Please note that some CBGs have leading zeros. |
| date_range_start | string | Start time for measurement period in ISO 8601 format |
| device_count | integer | Number of devices seen in our panel during the date range whose home is in this census_block_group. Home is defined as the common nighttime location for the device over a 6 week period where nighttime is 6 pm - 7 am. Note that we do not include any census_block_groups where the count <5. |
| completely_home_device_count | integer | Out of the device_count, the number of devices which did not leave the geohash-7 in which their home is located during the time period. |
| part_time_work_behavior_device | integer | Out of the device_count, the number of devices that spent one period of between 3 and 6 hours at one location other than their geohash-7 home during the period of 8 am - 6 pm in local time. This does not include any device that spent 6 or more hours at a location other than home. |
| full_time_work_behavior_device: | integer | Out of the device_count, the number of devices that spent greater than 6 hours at a location other than their home geohash-7 during the period of 8 am - 6 pm in local time. |
| delivery_behavior_devices | integer | Out of the device_count, the number of devices that stopped for < 20 minutes at > 3 locations outside of their geohash-7 home |
| distance_traveled_from_home | integer | Median distance (in meters) traveled from the geohash-7 of the home by the devices included in the device_count during the time period (excluding any distances of 0). We first find the median for each device and then find the median across all of the devices. |
| median_home_dwell_time | integer | Median dwell time at home geohash-7 ("home") in minutes for all devices in the device_count during the time period. For each device, we summed the observed minutes at home across the day (whether or not these were contiguous) to get the total minutes for each device. Then we calculate the median of all these devices. Beginning in v2, we include the portion of any stop within the time range regardless of whether the stop start time was in the time period. |
| median_non_home_dwell_time | integer | Median dwell time at places outside of geohash-7 home in minutes for all devices in the device_count during the time period. For each device, we summed the observed minutes outside of home across the day (whether or not these were contiguous) to get the total minutes for each device. Then we calculate the median of all these devices. |
| median_percentage_time_home | integer | Median percentage of time we observed devices home versus observed at all during the time period. |

*social distancing metric documentation*

**Data handling:**

1. In order to merge this data to the other sources, we used Census Tract Locations (LA) from USC data socrata to decrypt the fips code in the columns of 'origin_census_block_group' to a city and neighborhood's name.
2. Since one city and neighborhood may have multiple fips codes, we aggregate different columns into city and neighborhood levels.
3. Since the data provided by SafeGraph uses data sampling, the real numbers are not very meaningful. Therefore, we calculated the percentage values of 'completely home device count,' 'part time work behavior devics,' 'full time

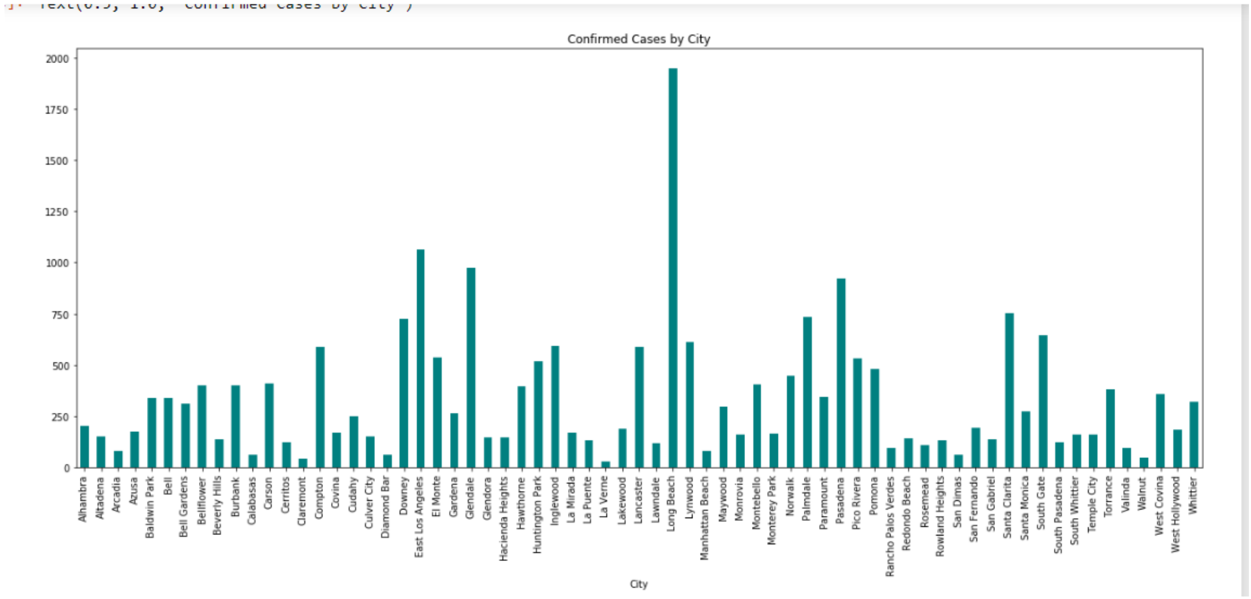
work behavior devices,' and 'delivery behavior devices' by dividing the 'device count.'

- **Los Angeles Neighborhood Map**
  Based on our exploration, we found the number of confirmed cases are greatly affected by the population size and size of the location. We think the number of cases decided by population density might be a better indicator to show the severity of COVID-19. From the 2018 health profile, we found the total population ('Tot_Pop') for different geographic areas. Although we are missing the area sizes for the different cities and neighborhoods, we were able to find the 'sqmi,' which indicates the square miles of a certain city/neighborhood, from this website . We used this feature to calculate our risk score.

  After merging the above datasets, our final dataset contained 1626 rows.

## 2. Data Exploratory

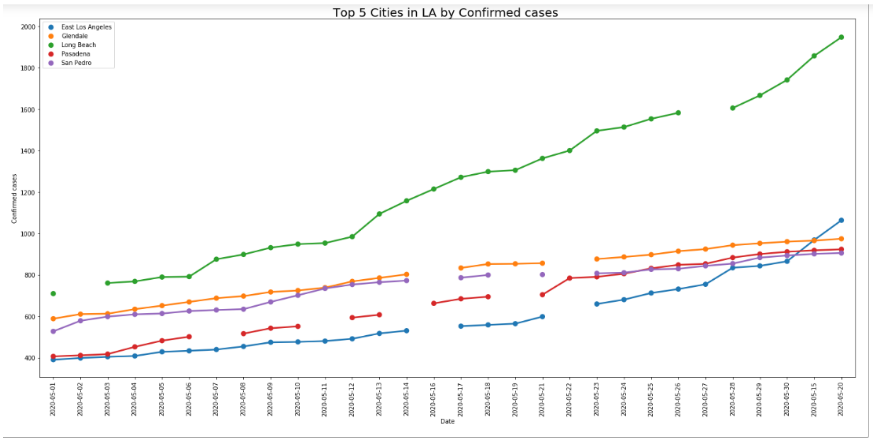
**Number of Confirmed Cases by City**



```
Text(0.5, 1.0, 'Confirmed Cases by City')
```

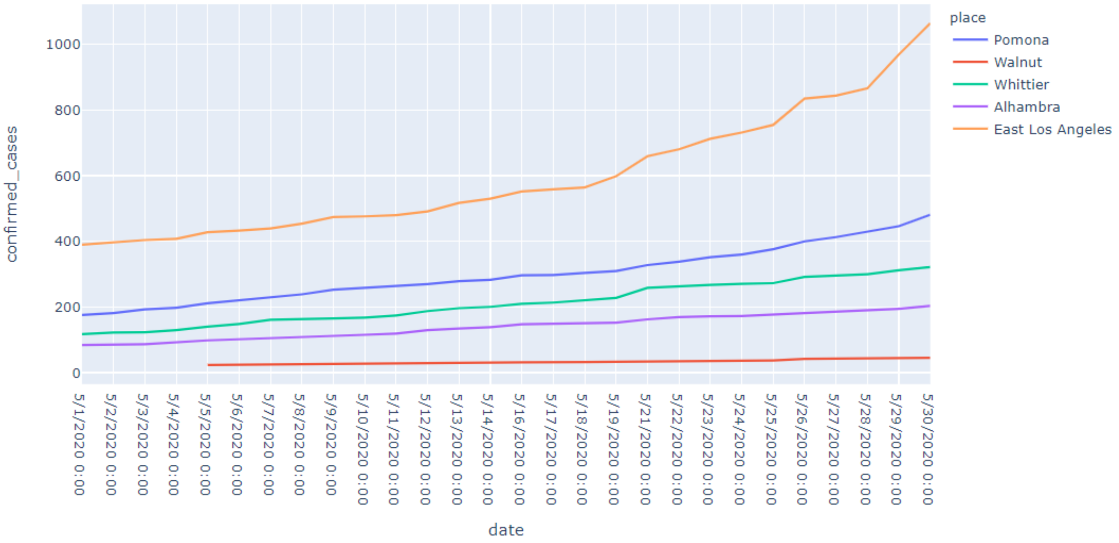
In the beginning of our study, we used the data to see which cities contained the most and the least amount of confirmed cases in order to help us analyze and compute an appropriate risk score. Based on this graph, we found that the city of Long Beach has the most confirmed cases, being approximately 2000 confirmed cases.

### Top Five Cities in Los Angeles by Highest Number of Confirmed Cases



We believe cities with a higher number of confirmed cases should be considered “high risk” for contracting COVID-19. From the data, we found that the top five cities in LA county, by the number of confirmed cases, are East Los Angeles, Glendale, Long Beach, Pasadena, and San Pedro. Because of the few missing data points in the dataset, the lines do not completely connect together, as seen in the graph above. For this discrepancy, we input the missing data by using mean imputation.

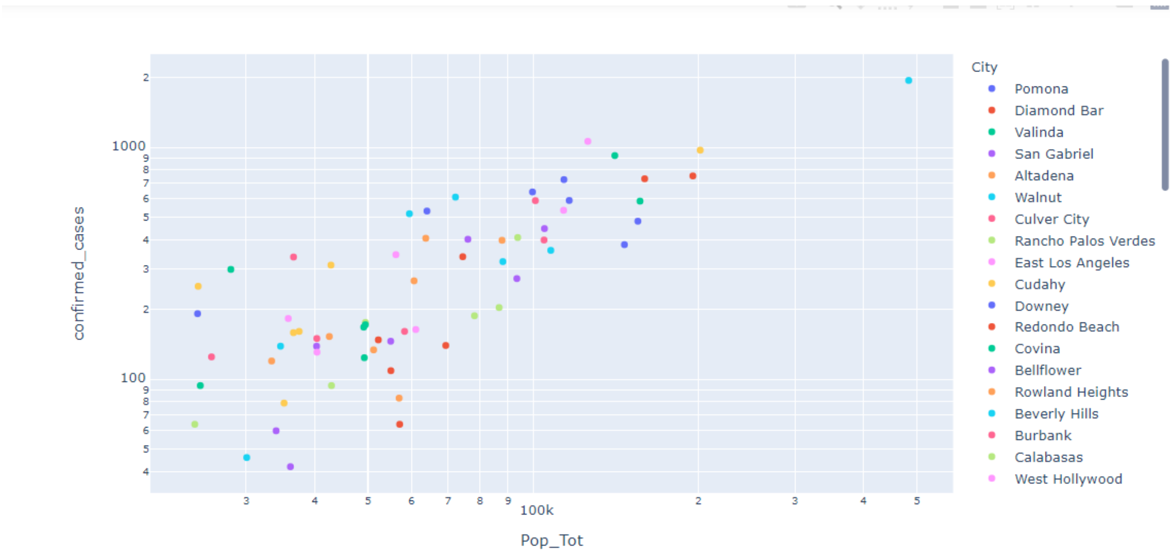
### Trend of Five Selected Cities in Adjacent Proximity to Cities with Highest Case Count



We selected five cities and neighborhoods surrounding the city of Los Angeles to analyze the trend of total infected cases between 5/1/2020 and 5/30/2020. We wanted to see whether the number of total confirmed cases of a particular city affects the surrounding cities. The data shows that Pomona has the most confirmed cases out of the five selected cities and this number continues to increase. The dataset also demonstrates that the number of confirmed cases in

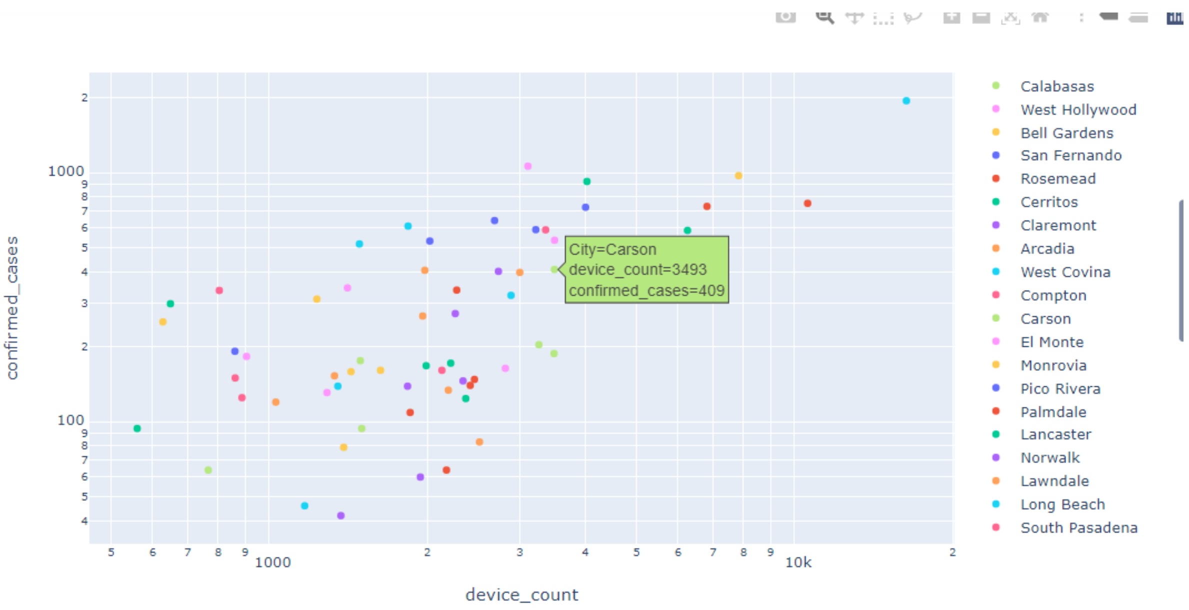
Walnut did not increase a lot as shown by the somewhat horizontal line in the graph visual. This line demonstrates that Walnut has a very few numbers of confirmed cases. As a result, it can be concluded that the confirmed case count of a city may not significantly affect the case count of nearby surrounding cities.

### Number of Confirmed Cases to Population Total by City



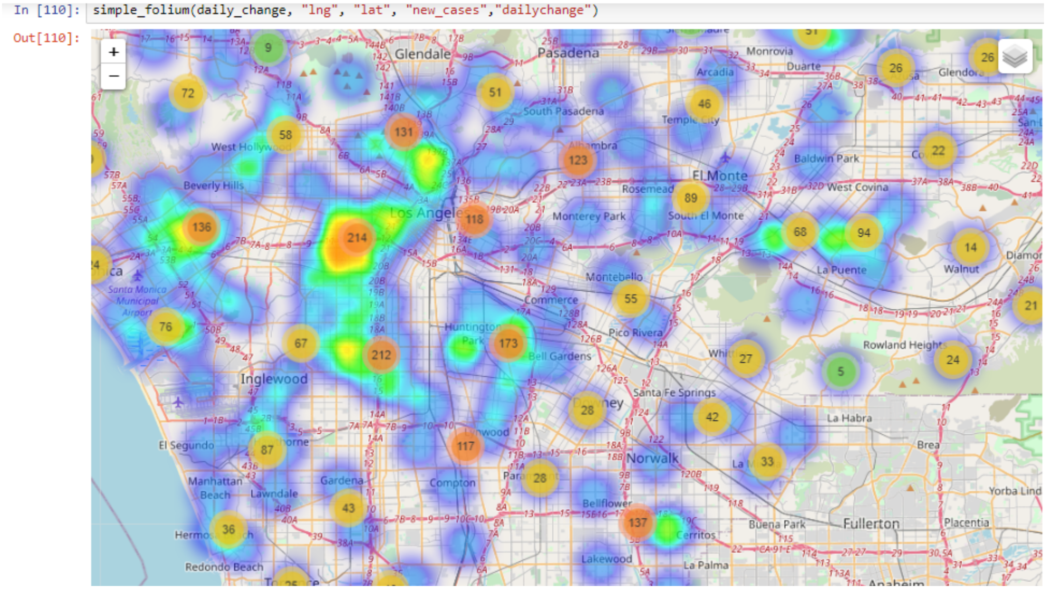
Here, we analyzed the relationship between confirmed cases and the population count for each city in order to see if the total population number could affect the number of confirmed cases. The graph above shows that there is a relationship between confirmed cases and total population for each city. In other words, cities with a lower population show to have a lower number of confirmed cases while cities with a higher population show to have a higher number of confirmed cases. However, it is important to note that in addition to the total population number, the mobility data for individuals, who travel from other cities, is another important feature to incorporate into the study.

### Number of Confirmed Cases to Device Count by City



This visual demonstrates the relationship between the number of confirmed cases and the device count for each city. ‘Device_count’ refers to the number of active technological devices within a given city. While the government recommends that people stay at home, there are still essential workers, including service or delivery workers, who must travel outside of his or her city for work. Because of this, we decided to see whether mobility would affect confirmed case count within a particular city in Los Angeles. As shown in the graph above, the data shows that although there is a relationship between confirmed cases and device count for each city, the relationship is not very strong. By way of illustration, some cities with a lower device count still show very high confirmed cases while cities with a high device count also show to have more confirmed cases.

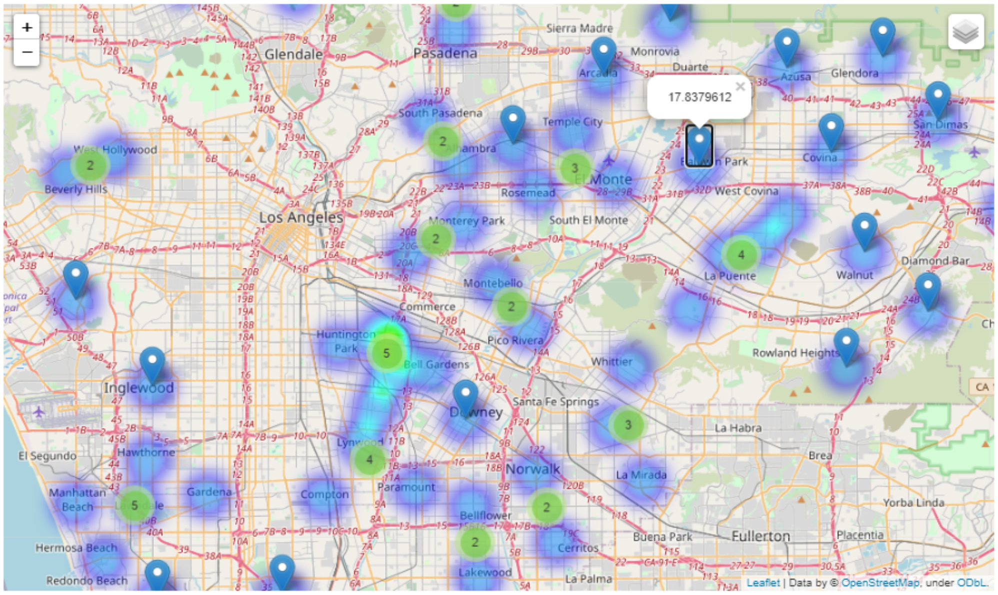
### Heatmap: Number of Confirmed Cases Throughout Los Angeles



```
In [110]: simple_folium(daily_change, "lng", "lat", "new_cases","dailychange")
Out[110]:
```

This visual demonstrates a heatmap by number of confirmed cases throughout Los Angeles. For this heatmap, we summed the daily new cases from 5/1/2020 to 5/31/2020 for each city and then incorporated its latitude and longitude. From this heatmap, we can see that the most confirmed cases are in the Downtown Los Angeles area.

### Heatmap: Risk Score for Each City



This visual demonstrates a heatmap of the risk score for each city throughout Los Angeles, on a scale of 1-5, with 5 being the highest risk. In order to compute the risk score for this heatmap, the latest available data must be used. The risk score prediction for the following day can serve as an important indicator to show what cities people should avoid traveling on that given day.

## Methodology

For our predictive models, we incorporated  features that we selected from the pool and ran three regression models in order to compare each of their performance. The three regression models we used are Linear Regression Model, Random Forest Regression Model, and XGBoost Regression Model.

Before building the models, we normalized the risk score to scale all the scores between 0 and 1. Also, since the tree-based model works better with discretization data, we use K means based discretization to modify our features into five bins. Figure 1 shows the comparison of before and after discretization.

**Figure 1:  Before (left) and After (right) Discretization**

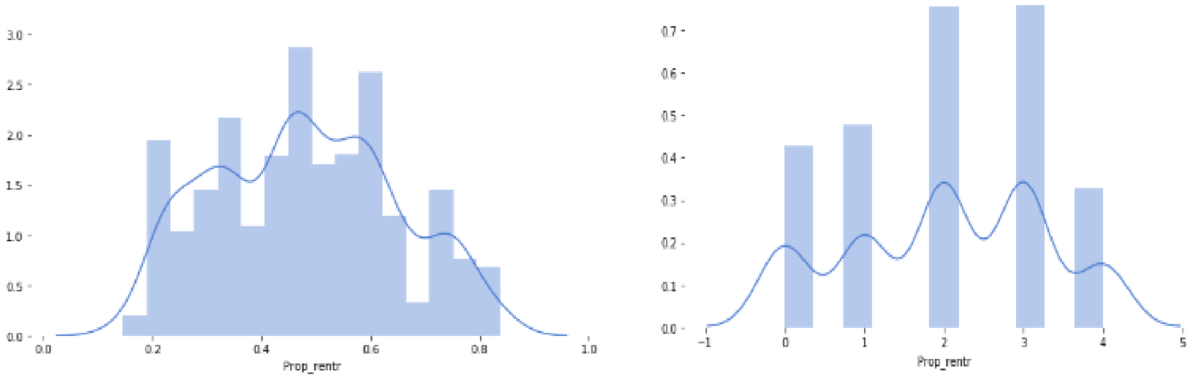


**Figure 2: Risk Score Distribution**

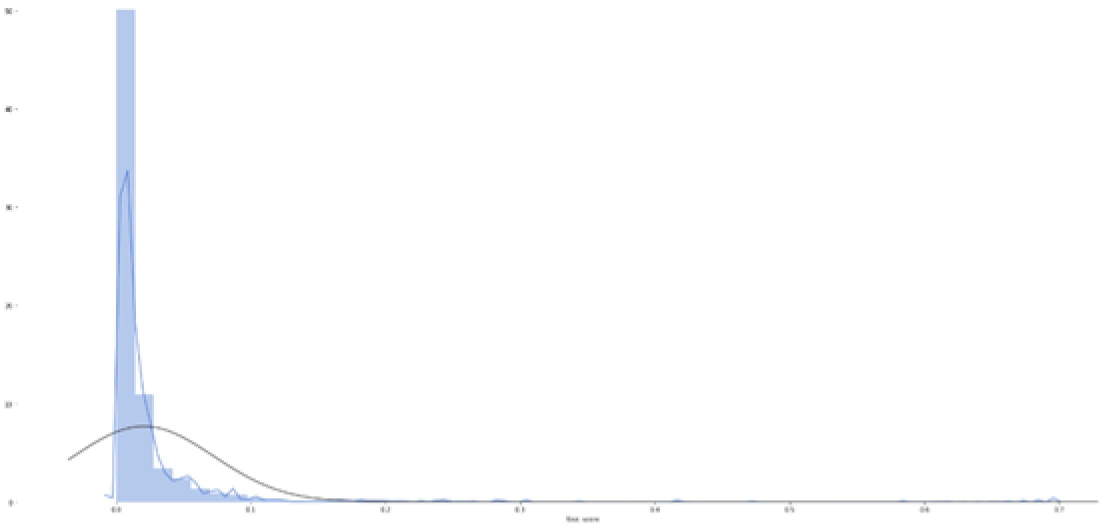
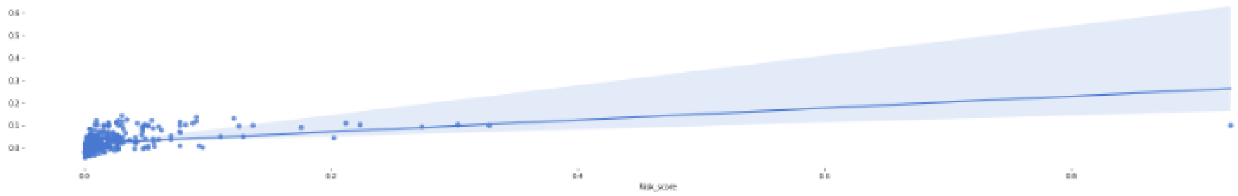
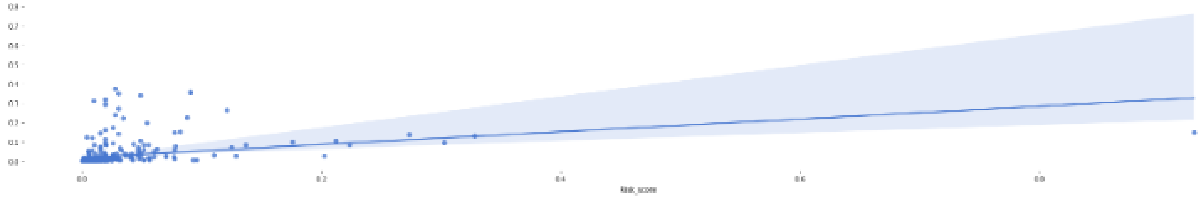


**Figure 3: Linear Regression**



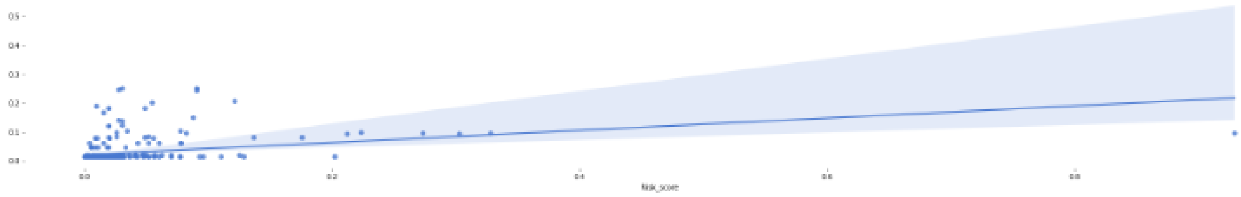
Since the risk score is not normally distributed as the Figure 2 shows, the Linear Regression model does not have a very good predictive ability. However, it still gives the best predictive ability over three models since it gives a better prediction on the risk score between 0 to 0.2, which contains most of the value. As Figure 3 shows the testing set and prediction, the model has Mean Squared Error (MSE) of 0.0026.

**Figure 4: Random Forest Regression**



The Random Forest Regression model has an advantage of dealing with unnormalized distributed value, however, the model can be overfitted easily. So, even though we have an MSE of 0.002 for the training model, the prediction model has an MSE of 0.0039, which means the model is overfitting.

**Figure 5: XGBoost Regression**

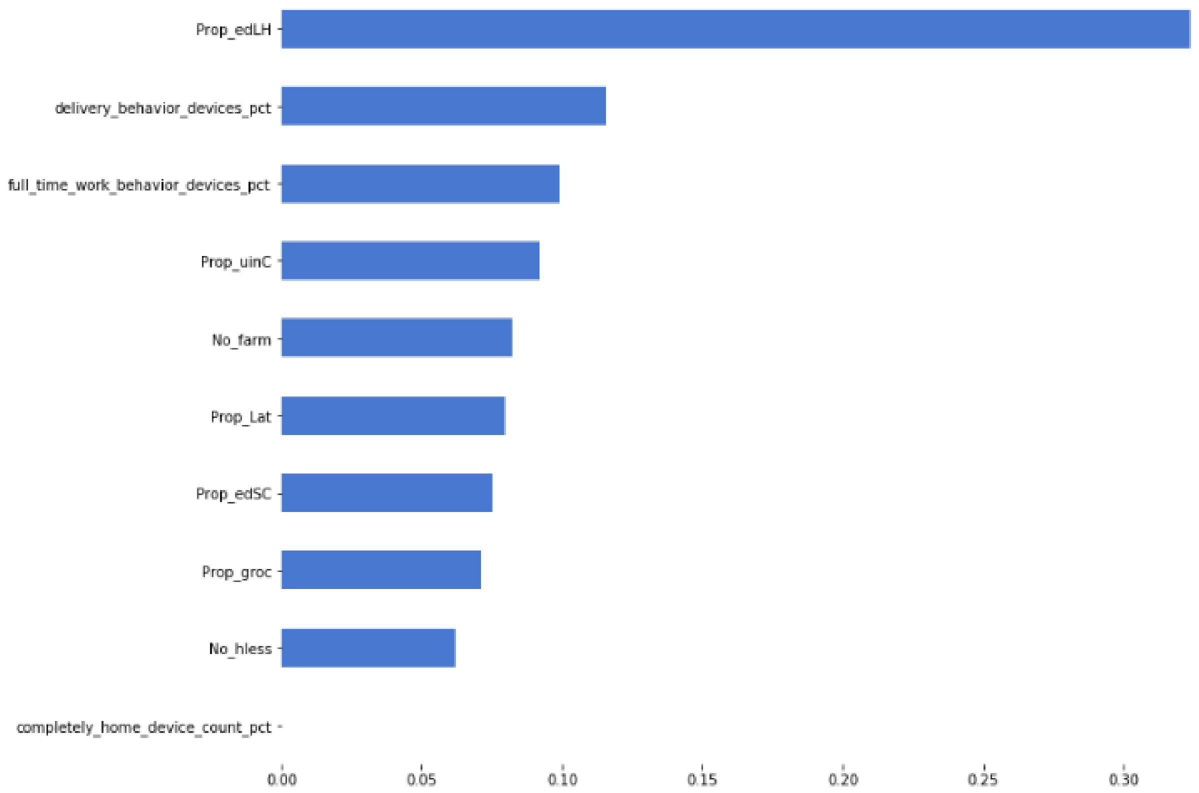


The XGBoost Regression has the ability to avoid overfitting, high flexibility, and continue on Existing Model. As a result, this model gives a better outcome than the Random Forest Regression model with MSE of 0.0029

Based on the observation by these three models, we can conclude that the Linear Regression Model has a low variance but higher bias and the XGBoost Regression Model has high variance but less bias. Because of this we formed a simple weight averaging ensemble method to further reduce variance, which aggregates 70% of the linear model prediction and 30% XGBoost model prediction for a better outcome of 0.0025 MSE.

## Results

**Figure 6: Feature Importance**



Based on Figure 6, features are ranked by importance given the coefficient associated with the risk score. The feature that influences our model the most is the Prop_edLH (Low education proportion) feature with 0.3 coefficient. The following features and their coefficients are full_time_work_behavior_devices_pct, delivery_behavior_devices_pct, and 'Prop_uniC' (uninsurance child proportion), which has a coefficient between  0.1 to 0.15, as well as, 'No_farm'(Number of farmers market), Prop_Lat (Latino), Prop_edSC (some college education), Prop_groc (living near grocery store), and No_hles (Number of Homeless), which are features that have a coefficient between 0.05 to 0.1.

Based on this graph, cities with a higher number of residents with low education, uninsured children, homeless people, farms market, people living closer to stores, and people that are highly mobile will tend to have a higher risk score.

## Risk Mitigation Recommendations:

1. One  short-term recommendation would be to keep a more watchful eye out for cities with high density and high mobility in order to keep the virus transmission in control, since these cities are more likely to have higher risk scores. This includes enforcing strict government ordinances to deter travel from individuals emerging high-risk areas.
2. Since low socioeconomic areas show to have a high confirmed case count, which can be caused by a plethora of reasons, such as not having the economic resources to stay home or get food delivered, another short-term recommendation would be to enact government programs to keep citizens at home while stimulating local economies. An example of a possible government program to enact would be subsidizing free grocery delivery to low socioeconomic areas to contain the spread of COVID-19.
3. Based on the model and the prominent important features used to compute the risk score, a long-term recommendation would be to focus on informing cities with lower educational levels about the dangers and proper safety protocols regarding COVID-19.
4. Another long-term recommendation would be to make more of an effort to ensure that citizens are not experiencing difficulty accessing healthcare and that those who have access to healthcare, yet choose to remain uninsured, are taking the proper safety protocols especially during times of pandemic.

## Acknowledgement

The advice provided by Dr. He and Dr. Wang has been extremely helpful throughout the journey of this challenge. Their knowledge and enriched experience in business and data analytics helped guide us through this open-ended puzzle with their useful insights.

## References

Los Angeles Neighborhood Map
https://usc.data.socrata.com/dataset/Los-Angeles-Neighborhood-Map/r8qd-yxsr

Los Angeles County City and Community Health Profiles 2018
https://data.lacounty.gov/Health/Los-Angeles-County-City-and-Community-Health-Profi/capb-kusk

Latimes-place-totals.csv
https://github.com/datadesk/california-coronavirus-data#los-angeles-countywide-statistical-areasjson

Social distance metrics

https://docs.safegraph.com/docs/social-distancing-metrics