Visualizing the Mobility Gradient Across Different Demographic Groups

Justin Snider
New York University
New York, USA
js10853@nyu.edu

Anchit Srivastava
New York University
New York, USA
as14022@nyu.edu

Diksha Chouhan
New York University
New York, USA
dc4454@nyu.edu

ABSTRACT

We propose to study the relative mobility patterns of different demographic groups in 2019 and 2020. The SafeGraph Social Distancing Data allows us to observe daily behaviors such as full-time work, part-time work, and full-time stay at home generated from mobile phone records. The daily statistics are aggregated by Census Block Group (CBG). Also, CBG demographic data is available from the American Community Survey (2016) 5-year estimate. Our study focuses on the metropolitan areas of New York, Los Angeles, and Chicago. We will compare the mobility of Census Block Groups with high, mean, and low percentages of poverty. In addition, we will compare the mobility of Census Block Groups with high, mean, and low percentages of minority populations. By analyzing the mobility behaviour we hope to better visualize and understand who has been able to effectively decrease their mobility. Understanding who has not been able to decrease their mobility can help us understand who is most at risk and who is most in need of assistance to reduce their exposure to COVID-19.

CCS CONCEPTS

• **Human-centered computing → Visualization**.

KEYWORDS

COVID-19, mobility

1 INTRODUCTION

Many recent papers, including "Mobility network models of COVID-19 explain inequities and inform reopening" [1] and "Coronavirus infections and deaths by poverty status: The effects of social distancing" [2] have established an essential link between mobility and COVID-19 infection rates and deaths. We propose to study the daily change in mobility of the general population during 2019 and 2020 in the three most populous cities in the US: New York, Los Angeles, and Chicago. Using the SafeGraph Social Distancing Metrics data[5], we will generate a mobility index at the resolution of the census block group level. We will use attributes such as the number of devices that remain at home, the count of devices with full-time work behavior, the median distance traveled by devices, and other vital actions relevant to mobility. We will then visualize each city's mobility based on the mobility index using graphs and maps produced with the GeoJSON file provided by SafeGraph of the Census Blocks.

We will compare with visualizations the mobility index of census blocks with the largest and smallest percentage of people in poverty based on the Census American Survey data provided by SafeGraph[4]. Also, we will compare with visualizations the mobility index of census blocks with the largest and smallest percentage of minority residents based on the Census American Survey data provided by SafeGraph. Our objective is to better understand and visualize to what degree different groups have adopted limited mobility to mitigate the documented threat of COVID-19 exposure created by higher mobility.

Several months have passed since the publication of these articles. All the 2020 data is now available from SafeGraph. There is now an opportunity to evaluate visualize the relationships between mobility, income, race, and the impact of COVID-19 in our communities.[3]

All code used to extract, clean, integrate, and visualize the data is available on the project Github repository https://github.com/chouhandiksha/bigdataproject. All the steps used for acquiring, extracting, integrating, and cleaning the data used for the project are described in sections 3 through 6.

2 RELATED WORK

In the *Nature* article "Mobility network models of COVID-19 explain inequities and inform reopening"[1] Chang et al. produce a model that predicts the number of anticipated SARS-CoV-2 infections. The model takes as an input mobility data derived from mobile phone data. The model tracks the predicted number of susceptible, exposed, infectious, and removed(SEIR) people anticipated to be in the population. The model is able to predict real case trajectories on held out data. The model correctly predicts higher infection rates among disadvantaged racial and socioeconomic groups solely as the result of differences in mobility.

"Coronavirus infections and deaths by poverty status: The effects of social distancing"[2] have shown the highest initial number of cases are in both the wealthiest and poorest countries. However, there is a great difference between the wealthy and poor countries when stay at home policies are put in place. The wealthiest countries have been able to show much greater reductions in mobility. As a result the wealthiest countries are able to curb the infection rates more effectively than poorer countries.

, ,

3 DATA SOURCES

We propose to use the SafeGraph COVID-19 Response Social Distancing Metrics[5], The SafeGraph Demographic American Survey Census Data[4], and The New York Times COVID-19 Data. [6] We have verified all three datasets are available to researchers, secured access to, and downloaded the datasets.

The SafeGraph COVID-19 Response Social Distancing Metrics [5] schema can be viewed online at https://docs.safegraph.com/docs. The dataset is organized by Census Block and includes mobility statistics generated from cell phone data such as median distance travel from home, completely at-home device count, full-time work behavior, and many more mobility aggregate statistics.

The SafeGraph Demographic American Survey Census [4] schema can be viewed online at https://docs.safegraph.com/docs/open-census-data. The dataset is organized by Census Block and contains demographic information, including statistics about race, household income, and poverty.

The New York Times COVID-19 Data [6] can be accessed online at https://github.com/nytimes/covid-19-data. The dataset contains reported daily COVID cases and deaths organized at the Country, State, and county levels.

4 DATA ACQUISITION

4.1 New York Times COVID-19 Data

The New York Times COVID-19 data can easily be downloaded directly from the Github repository at https://github.com/nytimes/covid-19-data. We are using the us-counties.csv file for county level daily case and death counts. The CSV file was about 46 MB, a manageable size.

4.2 SafeGraph Social Distancing Metrics and Census Block Group Data

The SafeGraph Social Distancing metrics include daily aggregate data for each CBG for all of 2019, 2020, and the beginning of 2021.

The SafeGraph Census Block Group Data include the American Community Survey (2016) 5-year estimate on the Census Block Group level.

You can find comprehensive instructions on how to download both the SafeGraph Social Distancing Metrics and Census Block Group Data in the code repository for the project: https://github.com/chouhandiksha/bigdataproject.

Step By Step Instructions:

- Step 1: Create an account at https://www.safegraph.com/covid-19-data-consortium. Those with academic NYU email address can register for free.
- Step 2: Follow the instructions and sign the legal document received in the email mentioning we should not release the data to the public and use it only for research/project purposes.
- Step 3: Open the links provided for the various datasets mentioned above including the Social Distancing Metrics and Census Block Group Data.
- Step 4: Download and Install AWS CLI using this link: https://docs.aws.amazon.com/cli/latest/userguide/install-cliv2.html
- Step 5: Click the "Reveal Access Key" button present on the SafeGraph data screen.
- Step 6: Complete the setup displayed on the screen after revealing the access key.
- Step 7: Go back to the dataset page and select the menu option "CLI".
- Step 8: Run the command with the local directory to complete the downloading process.

5 DATA EXTRACTION

A FIPS code is a number that uniquely identify a specific geographic area. Each Census Block Group (CBG) is represented by a 12 digit FIPS code. The first two digits of a FIPS code represents the state. The next three digits represent the county. Then, six digits give the Census Tract. Our final digit is the block group. Knowing the FIPS codes for the counties in a given metropolitan area allows us to extract the rows we need and leave the other behind.

5.1 New York Times COVID-19 Data

The New York Times COVID-19 file us-counties.csv has a FIPS column. After loading the file into a dataframe using Pandas we filter out just the rows with FIPS matching the counties in the metro areas. The result allowed us to get all the relevant row for the counties of Los Angeles and Chicago.

Filtering for the New York FIPS counties returned no rows. Upon investigation we discovered ironically the New York Times does not maintain county level information for the counties of New York City. They just group all 5 counties of New York City together. Filtering for New York City in the county name column we were able to secure the city level cases and deaths for New York City.

5.2 SafeGraph Census Block Group Data

SafeGraph provides a download of the 2016 American Community Survey (ACS) 5-year estimate on the Census Block Group level. We where able to filter out just the counties of the New York, Los Angeles, and Chicago Metro areas from the dataset one CSV file at a time using the FIPS code. We simply filter out CBGs where the first five digits of the CBG FIPS code matches a county from one of our 3 metro areas. Each row includes a Census Block Group that the row attributes describe. There are so many attributes in the ACS that the data is broken up into separate CSV files. The name of the file tells you want attributes are included in the file. Inside each file there is one row for each CBG.

We have extracted the following columns for all CBG in the three metro areas:

- **Column B01003e1:** the total estimate population
- **Column B02001e2** the white only estimated population
- **Column C17002e1** the total population for whom poverty status is determined
- **Column C17002e2** the population for whom the ratio of income to poverty level in the past 12 month is under 0.5 from the population for whom poverty status is determined
- **Column C17002e3** the population for whom the ratio of income to poverty level in the past 12 months is between 0.5 to 0.99 inclusive from the population for whom poverty status is determined

To better make comparisons between CBGs we calculated the white only and poverty percentages for each CBG from the given population values. The CBG white only percentage is calculated by $\frac{w}{t} \cdot 100$ where w is the white only population and t is the total population. The CBG percentage in poverty is calculated by $(a + b)/d * 100$ where $a + b$ is the total population with income below the poverty level and d is the total population for whom the poverty status is determined.

5.3 SafeGraph Social Distancing Metrics

The SafeGraph Social Distancing Metrics data is downloaded from AWS S3 storage to a local directory. The file structure provides the labels for the many separate gzip csv files. The folder structure pattern is: **YYYY/MM/DD.gz**. For instance, we have 12 folders representing each month i.e. from 01 to 12 in the "2019" folder, and in each of the folders a zipped file representing each day of the month were present. This was massive data that took us a significant amount of time to download.

There is one file for each day of 2019 and 2020. There are also daily files for the first few months of 2021. The data set is still being updated every day with about a 1 week lag time between when data is captured and when it is added to the data set.

We looped through all the files in the data set reading them into memory and keeping just the CBG rows from the 3 metro areas using the FIPS code column. For convenience once selected we save the filtered rows for each day to a csv file. Our new files follow a different format. The name of the file gives the year, month, and date by using the format **YYY-MM-DD-social-distancing.csv** and are all stored in a single directory for the entire year. For analysis we just read any needed daily data into memory by looping through the new filtered data files.

In addition to filtering the metro area rows, we have calculated several critical percentage values: the percentage of devices exhibiting part-time work behavior, percentage of devices exhibiting full-time work behavior, and percentage of devices exhibiting completely home status. Each of these percentages was taken using $b/d * 100$ where b is the population exhibiting the behavior in the CBG and d is the total number of devices in the CBG.

Finally, for each of these three percentages we have also calculated the normalized value by using $(v - \mu)/\sigma$ where v is the value to normalize, μ is the mean, and σ is the standard deviation.

6 DATA CLEANING

6.1 New York Times COVID-19 Data

In order to clean the data we have used Python, Pandas, and OpenRefine to take the following steps:

- Verify the dates fall between January 24th, 2020 and March 18th, 2021 using Pandas min and max function.
- We have verified the extracted rows only include the desired rows from the 3 metro areas using Pandas.
- We have verified all unique date and CBG combos have only one row by comparing the total number of rows to the number of unique combos. We found the count matches for all extracted data.
- Verify the cases and deaths are integer values as expected using the Pandas min and max function.
- We have verified there are no empty, Null or NaN values in the extracted data we are using.

6.2 SafeGraph Census Block Group Data

In order to clean the ACS data we have used Python, Pandas, and OpenRefine to take the following steps:

- We have checked that each CBG is unique and is not duplicated.
- We have checked to make sure all percentages fall within the valid range from zero to a hundred using the Pandas min and max function.
- We have produced histograms for each attribute to verify the shape of the data and range of values are valid.
- We have visualized the CBG values on maps to give an intuition into the contents of the data and look for outliers.
- Using Pandas '.isna' function we verified there are no empty, NaN, or Null entries in any of the Social Distancing CSV files.

6.3 SafeGraph Social Distancing Metrics

In order to clean Social Distancing data we have used Python, Pandas, and OpenRefine to take the following steps:

- OpenRefine we have checked for duplicate CBG on a subset of the hundreds of csv files. We found no duplicates in any files we checked.
- Using Pandas we checked the minimum and maximum date values to ensure all the dates fall within the expected time frame. We found no values outside the expected time frame for all Social Distancing CSV files.
- Using the Pandas max and min function we checked that all the percentage values fall within the expected range for all percentage columns. We found no values outside the expected range for all Social Distancing CSV files.
- Using the Pandas max and min function we checked that all integer data types fall within the expected range, which includes no negative values. We found no values outside the expected range for all Social Distancing CSV files.
- Using Pandas '.isna' function we verified there are no empty, NaN, or Null entries in any of the Social Distancing CSV files.
- For each city we visualized the mean daily and mean monthly values for all three percentage values **percentage_completely_home**, **percentage_part_time_work**, and **percentage_full_time_work** in both 2019 and 2020. All three cities show an increase in completely home individuals during 2020. There is also a visible decrease in part-time and full-time work behaviour in 2020. Given the documented decrease in mobility during 2020 these are the trends we expect to see.

7 ISSUES FACED

The engineers at SafeGraph advocate transparency in data so that their customers face minimal issues in analyzing their different datasets. However, due to large dataset sizes and encoding of the columns, we faced several challenges while extracting the valuable

information from the whole set. These challenges are discussed as below:

- The social distancing metrics file size was 81 GB at the time of our download. It took more than 4 to 6 hours to download. Further, such a big file suffered download failed multiple times.
- The Census Block Group Data was 10 GB at the time of our download. It took 3 hours to download.
- The folder structure for the social distancing metrics dataset for a year is divided based on months and further based on days. Hence, for the years 2019 and 2020, we found 12 sub-folders (each representing a month) and archive files inside these sub-folders for each day of the month. It caused difficulty in integrating the dataset in a single data frame. Therefore, we ran a 'loop' for every month and day to extract the data rows and combine them into one data frame to perform analysis and visualization.
- We tried to write the combined dataset for 2020 in one CSV file, which caused memory overload and crashed our local machine.
- Each column in open census data is a particular census attribute estimated by the US Government. It includes 7500 *table_ids* (columns) for the 220,000 + census block groups in the USA. it was quite overwhelming and took time to wrap our heads around. After following the specific guidelines (and cheat sheet) from the safegraph, we identified the type of population represented by each *table_id*.

8 CONCLUSIONS

The SafeGraph data is of high quality. As a result our comprehensive data quality study found no major issues. The Census Block Group data was unique, all the values were found to fall within the expected range, and there were only a few empty or null values.

The datasets are huge and require a lot of computational time. The data extraction process can be run on a modest machine, but very slowly. Especially the very large SafeGraph Social Distancing dataset, which can take two hours to extract the data for just one city. However, after the extraction process loading data for one year of one city into memory only takes 15 minutes. Once in memory exploration and analysis can be preformed quickly.

Using Google Colab with Google Drive the data can be extracted from the raw data set. However, occasionally the Google API call limit is reached. The issue only occurred when extracting the very large SafeGraph Social Distancing dataset. When the API limit is reached we found that closing all notebooks and waiting 24 hours allowed allowed us to finish running the extraction notebooks.

The semantics of working with the Census data is not very intuitive. You must understand the FIPS code system, which is the primary method used for encoding the geological entities throughout the United States. The ACS columns naming system is not intuitive. In addition, the column descriptions in the documentation are not always easily understood either. Fortunately, the ACS documentation is very rigorous and there are many resources online to provide additional instruction on how to work with Census data. In addition, once you figure out the semantics and structure of the data there is very little cleaning to do because the quality of data is very good.

We anticipated a spike in the number of individuals staying at home starting in March due to many factors, including the issuing of stay at home orders. Inversely, we expected to see the number of people exhibiting part-time and full-time work behavior decline in March. We visualized the mean percentage values for each of these three attributes for each of the three cities in our study.14 The line graphs conform to our expectations, which gives us confidence in the data. In comparison, the line graphs of 2019 show gradual change.13

9 CITATIONS AND BIBLIOGRAPHIES

REFERENCES

[1] Serina Chang, Emma Pierson, Pang Wei Koh, Jaline Gerardin, Beth Redbird, David Grusky, and Jure Leskovec. 2021. Mobility network models of COVID-19 explain inequities and inform reopening. *Nature* 589, 7840 (01 Jan 2021), 82–87. https://doi.org/10.1038/s41586-020-2923-3

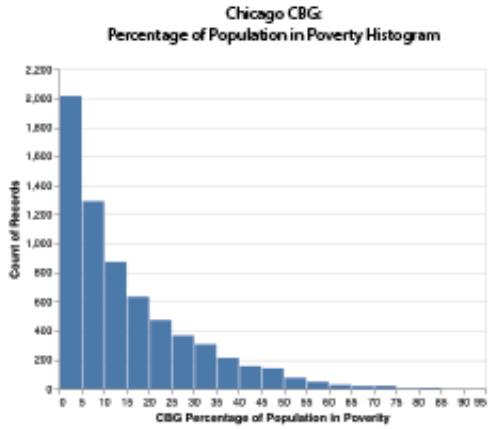
[2] Juergen Jung, James Manley, and Vinish Shrestha. 2021. Coronavirus infections and deaths by poverty status: The effects of social distancing. *Journal of Economic Behavior & Organization* 182 (2021), 311–330. https://doi.org/10.1016/j.jebo.2020.12.019

[3] Richard V. Reeves and Jonathan Rothwell. 2020. Class and COVID: How the Less Affluent face Double Risks. (2020). https://www.brookings.edu/blog/up-front/2020/03/27/class-and-covid-how-the-less-affluent-face-double-risks

[4] SafeGraph. 2016. Census Block Group Data. https://docs.safegraph.com/docs/open-census-data (All data from 2016 American Community Survey by Census Block Group).

[5] SafeGraph. 2021. Social Distancing Metrics. https://docs.safegraph.com/docs/social-distancing-metrics

[6] The New York Times. 2021. Coronavirus (COVID-19) Data in the United States. https://github.com/nytimes/covid-19-data



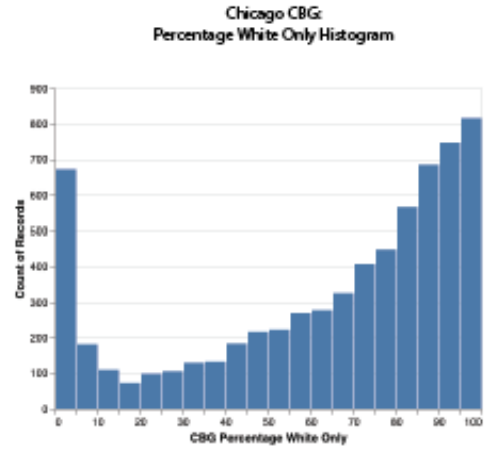
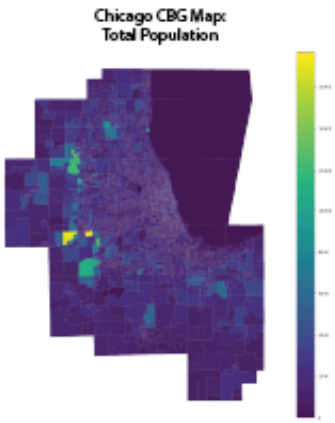
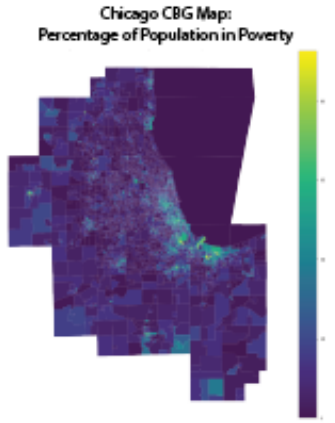


Figure 1: Chicago Metro Area CBG histograms for the total population, percentage below the poverty level, and percentage white only from the 2016 American Community Survey 5-year Estimate on the Census Block Group level.





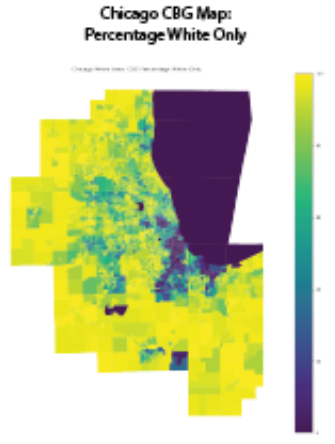
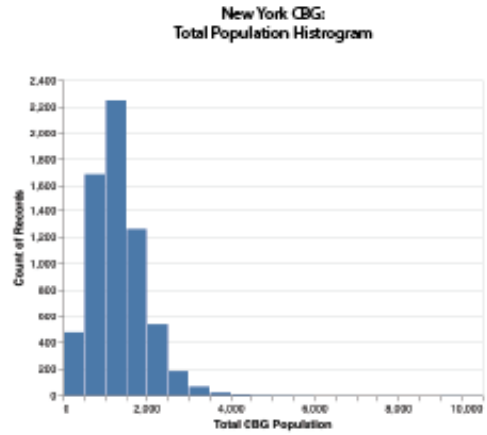
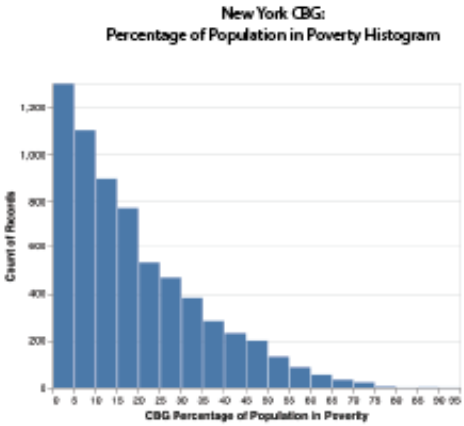


Figure 2: Chicago Metro Area CBG maps for the total population, percentage below the poverty level, and percentage white only from the 2016 American Community Survey 5-year Estimate on the Census Block Group level.





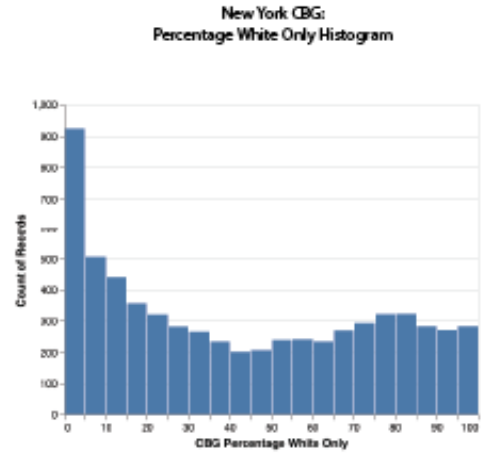
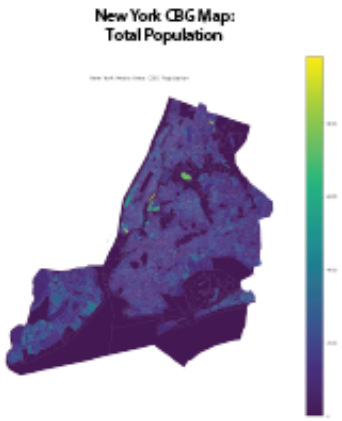
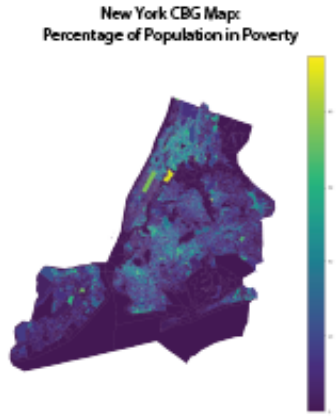


Figure 3: New York Metro Area CBG histograms for the total population, percentage below the poverty level, and percentage white only from the 2016 American Community Survey 5-year Estimate on the Census Block Group level.





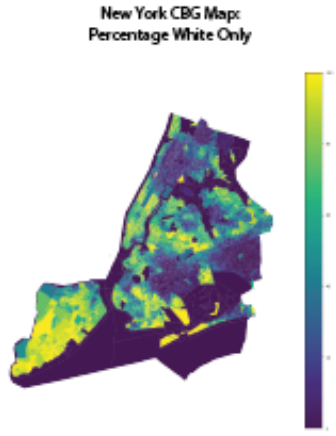
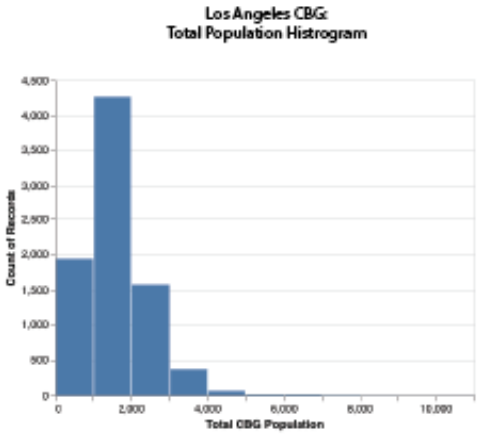
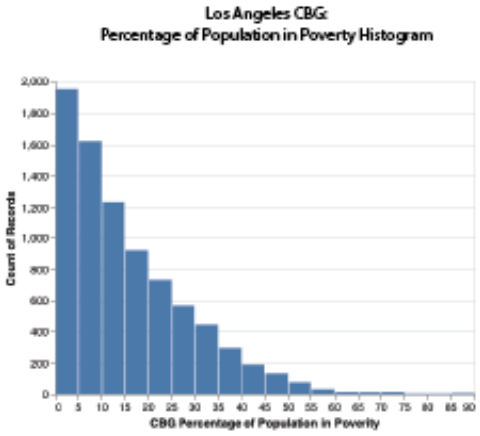


Figure 4: New York Metro Area CBG maps for the total population, percentage below the poverty level, and percentage white only from the 2016 American Community Survey 5-year Estimate on the Census Block Group level.





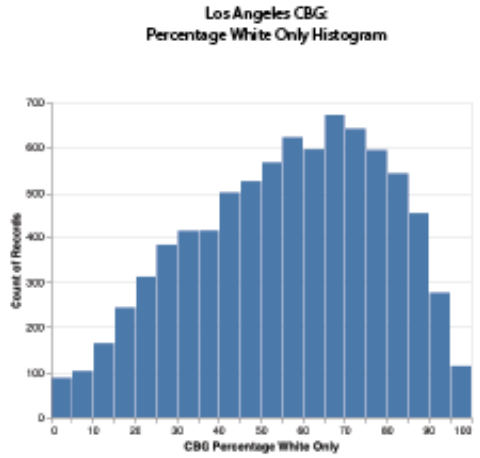
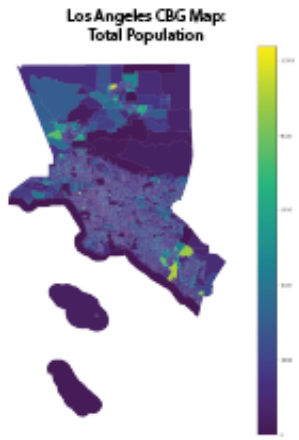
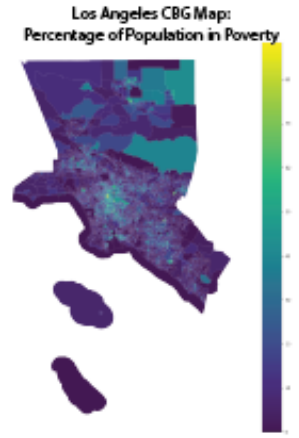


Figure 5: Los Angeles Metro Area CBG histograms for the total population, percentage below the poverty level, and percentage white only from the 2016 American Community Survey 5-year Estimate on the Census Block Group level.





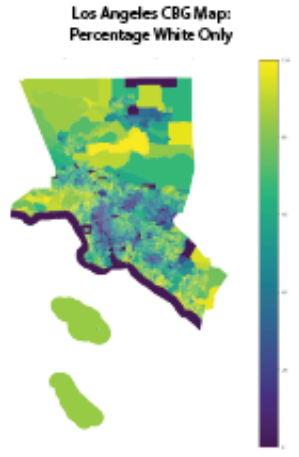


Figure 6: Los Angeles Metro Area CBG maps for the total population, percentage below the poverty level, and percentage white only from the 2016 American Community Survey 5-year Estimate on the Census Block Group level.

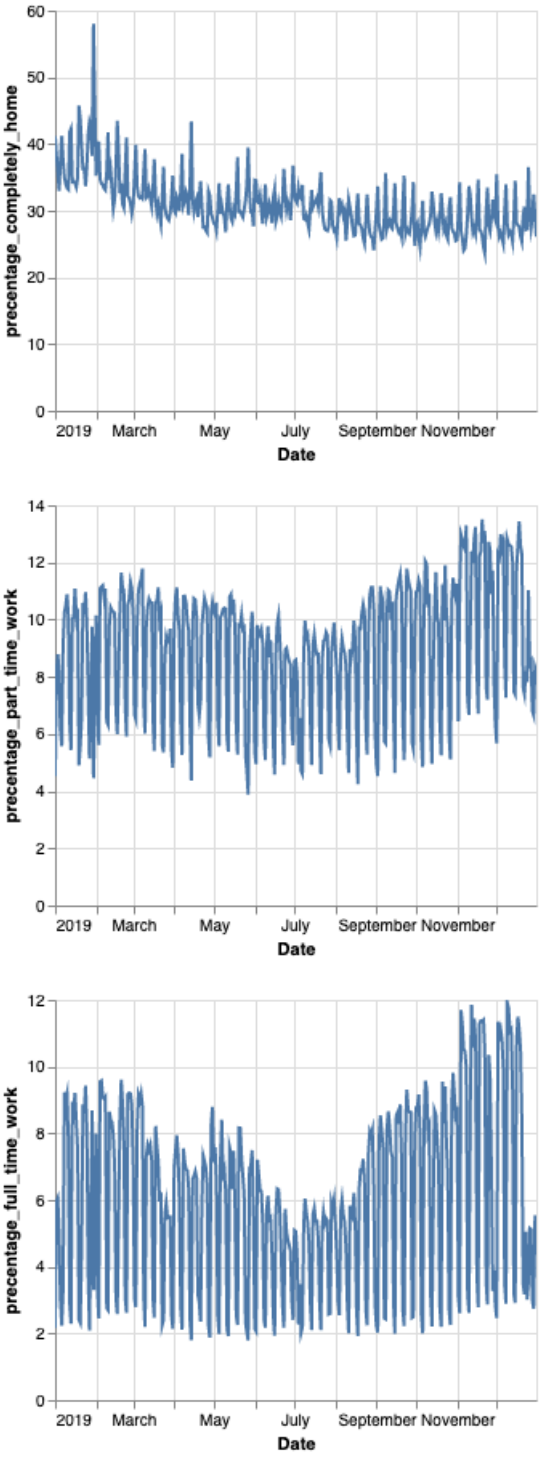


Figure 7: 2019 Chicago Metro Area daily mean percentages calculated using the SafeGraph Social Distancing data set.

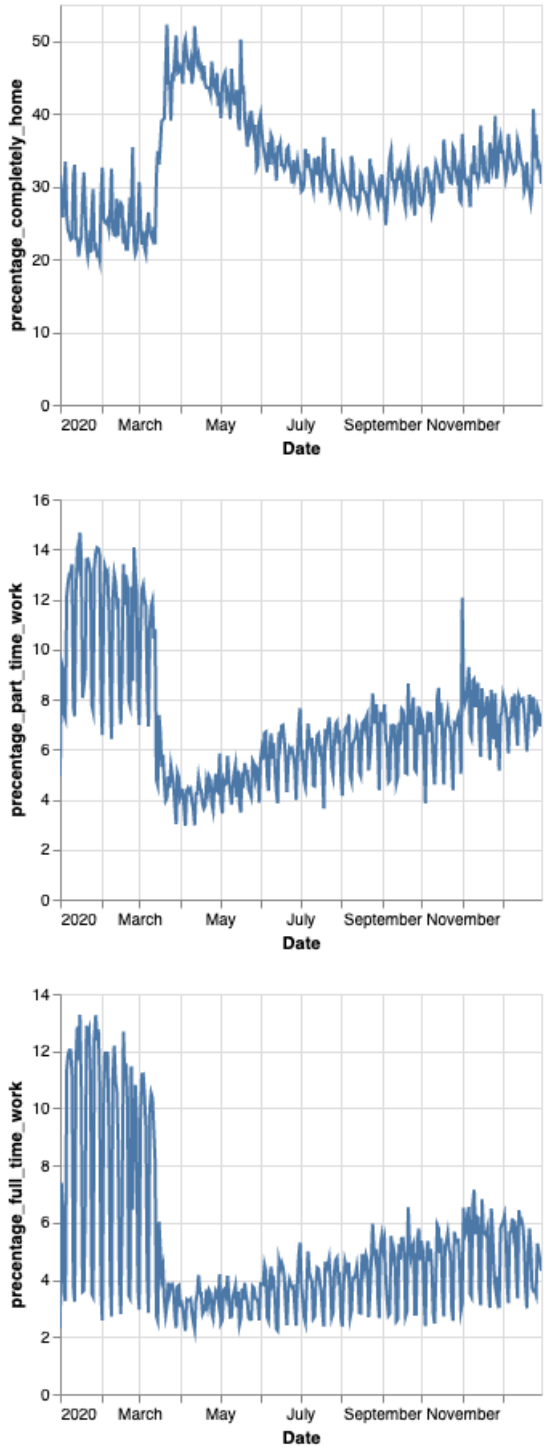
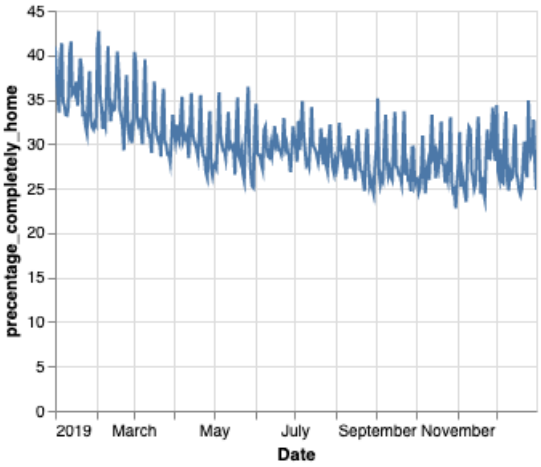
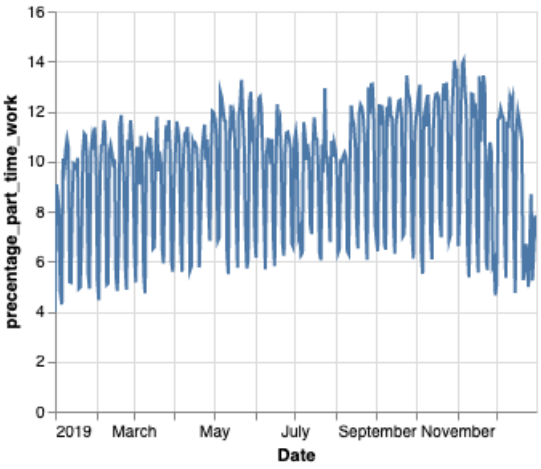


Figure 8: 2020 Chicago Metro Area daily mean percentages calculated using the SafeGraph Social Distancing data set.





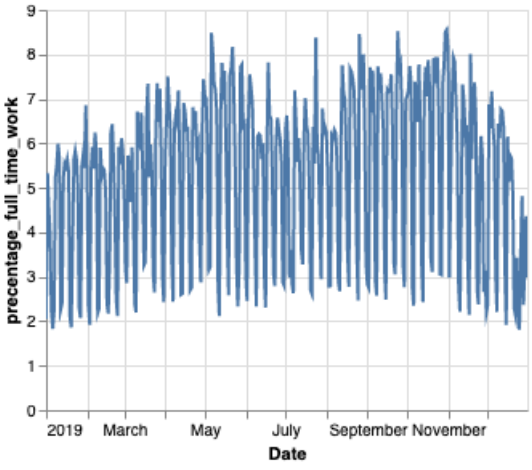
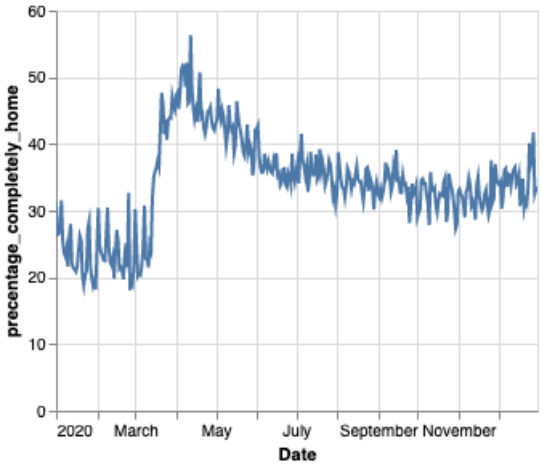
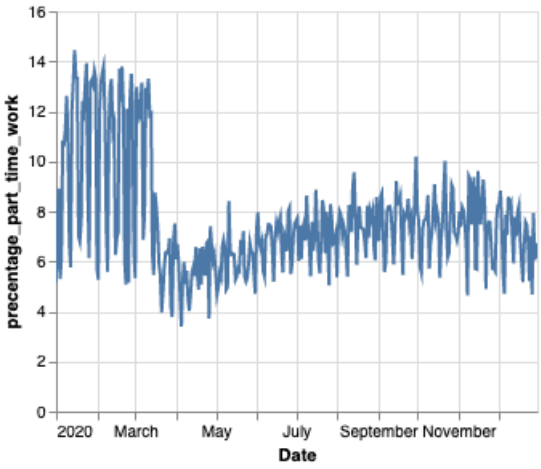


Figure 9: 2019 Los Angeles Metro Area daily mean percentages calculated using the SafeGraph Social Distancing data set.





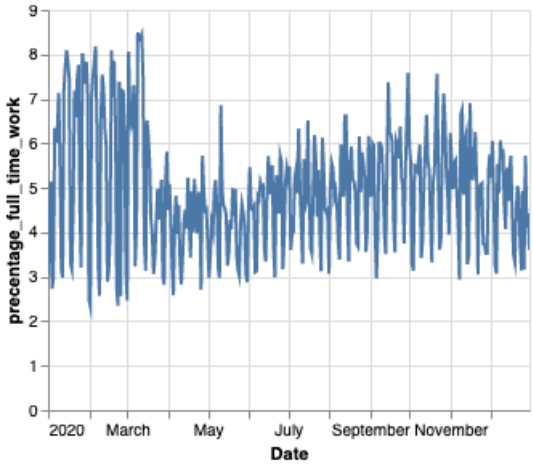
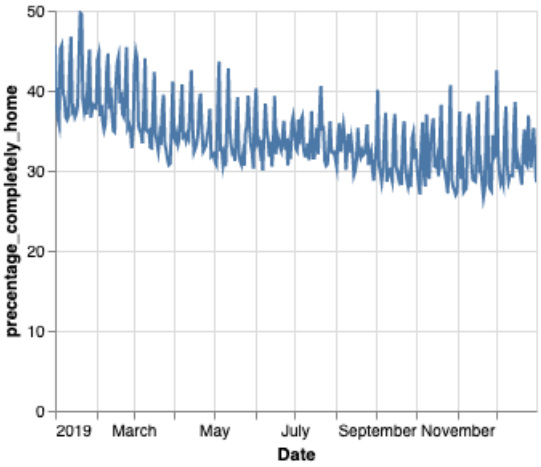
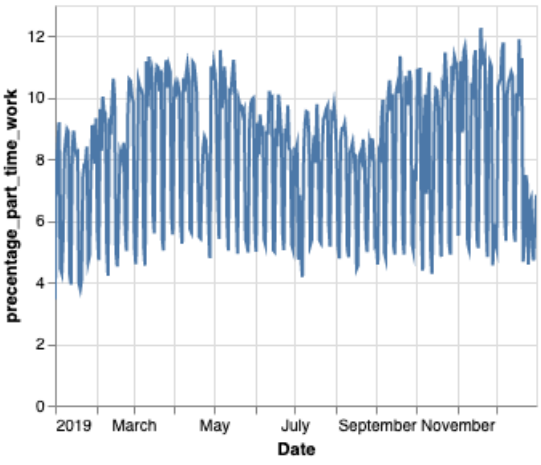


Figure 10: 2020 Los Angeles Metro Area daily mean percentages calculated using the SafeGraph Social Distancing data set.





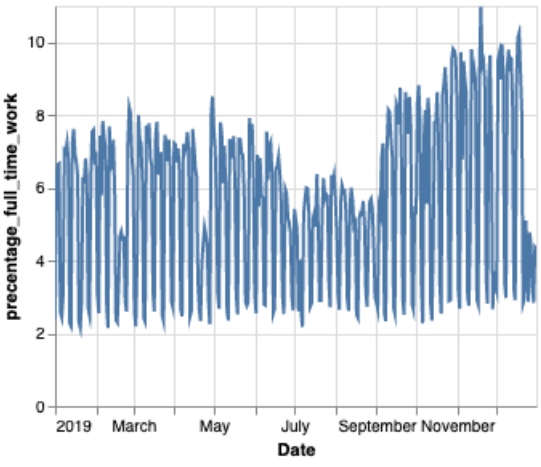
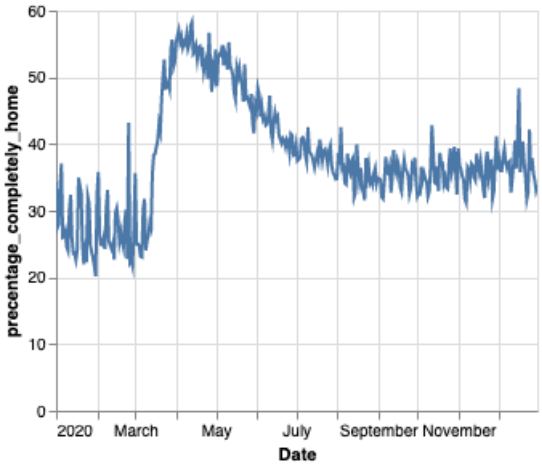
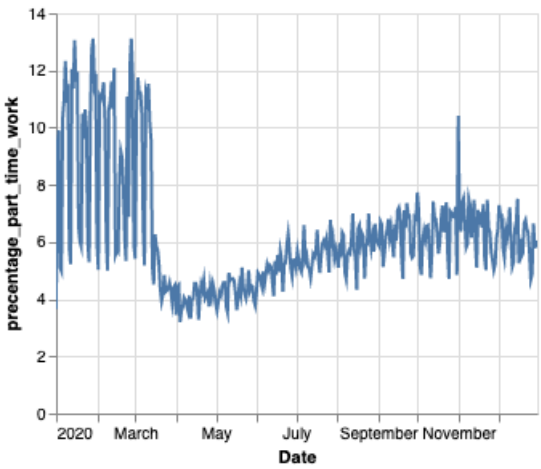


Figure 11: 2019 New York Metro Area daily mean percentages calculated using the SafeGraph Social Distancing data set.





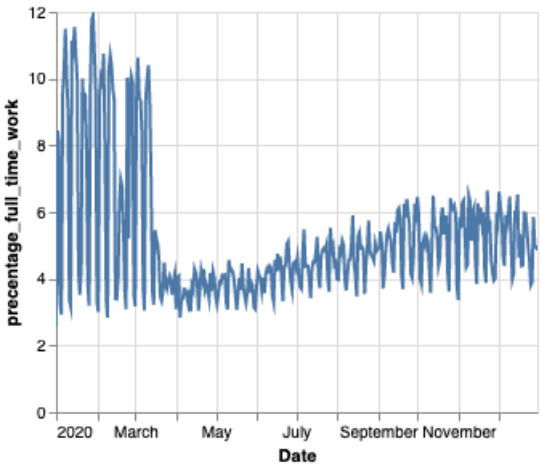
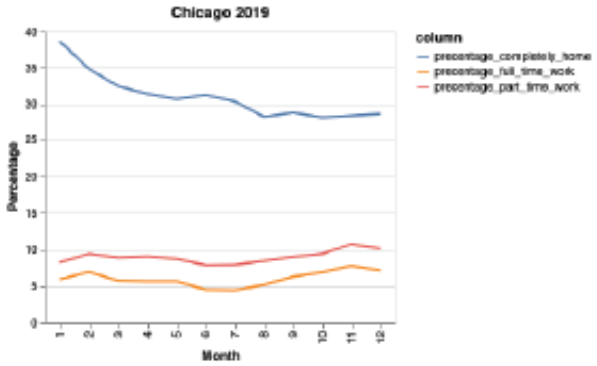
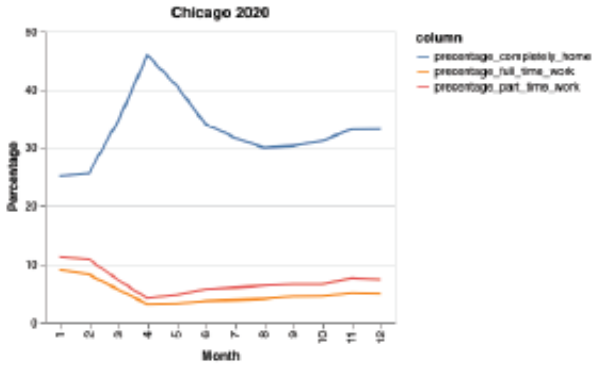
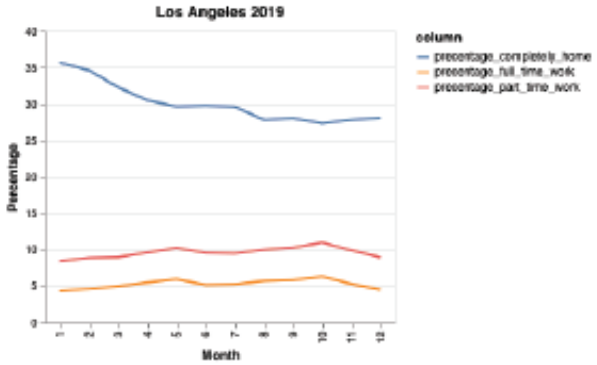
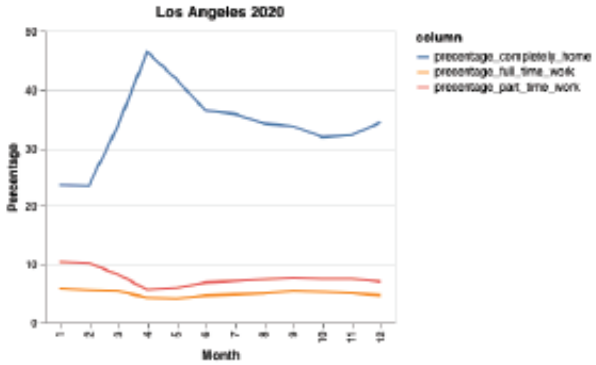


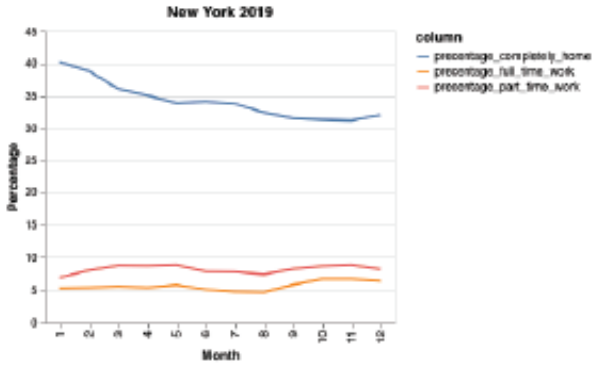
Figure 12: 2020 New York Metro Area daily mean percentages calculated using the SafeGraph Social Distancing data set.











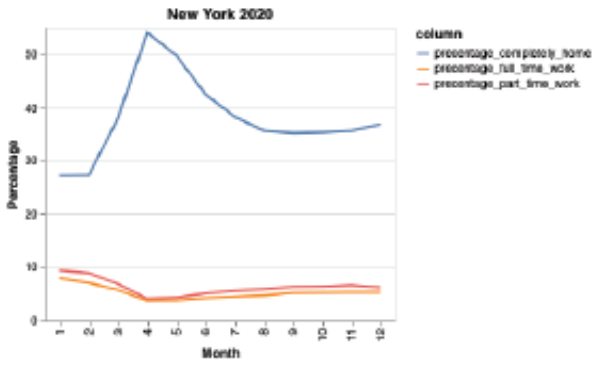


Figure 13: 2019 New York Metro Area monthly mean percentages calculated using the SafeGraph Social Distancing data set.

Figure 14: 2020 New York Metro Area monthly mean percentages calculated using the SafeGraph Social Distancing data set.