Deliverable 3: Introduction

This document summarizes the work done for Deliverable 3, for the BU Spark! project, Spring 2021, MAPC Broadband Digital Equity in MA. This document is split into two parts:

1. This section, which acts as an overview to this deliverable
2. The section titled "Deliverable 3: Final Report Draft," which assumes the role of an initial draft for our final report in this project.

Date: 04/09/21

Student Team

This project has two different teams, denoted MAPC team 1 and MAPC team 2. We represent team 2. There are five students, and one project manager for team 2:

- Adam Streich
- Jenny Li
- Nathan Lauer
- Yutong Shen
- Zhixing Zhao

The project manager is Kamran Arif.

Contact

- Ryan Kelly, RKelly@mapc.org Digital Services lead at the MAPC
- Matt Zagaja, mzagaja@mapc.org , Lead civic web developer at the MAPC

Organization

The Metropolitan Area Planning Council - MAPC

Purpose

In this deliverable, we present our continuing work with the MLAB and Ookla datasets. Notably, we also introduce a new dataset -- 2014-2018 census income data -- and correlate broadbands speeds against this data. As with previous deliverables, there are two primary outcomes of this deliverable.

The first outcome is a continuing analysis of the datasets, with a particular focus on the following:

- Broadband speed measurement density in the Ookla dataset
- Understanding download and upload speeds per municipality in the Ookla dataset
- Correlating MLAB broadband speed against median household income, per municipality
- Correlating Ookla download speed against median household income, per municipality.

The second outcome of this deliverable is an initial draft of our final report for this project. As such, the remaining sections will flow slightly differently than in previous deliverables; the following are a draft of our initial report. Note that we have included work from both previous deliverables and this deliverable in the upcoming sections. Each section is demarcated with the deliverable the associated work was completed in, so our progress towards this deliverable should be clear.

Deliverable 3: Final Draft Report

We present the work completed towards the BU, Spark! project, Spring 2021, MAPC Broadband Digital Equity in MA.

Student Team

This project has two different teams, denoted MAPC team 1 and MAPC team 2. We represent team 2. There are five students, and one project manager for team 2: Adam Streich, Jenny Li, Nathan Lauer, Yutong Shen, and Zhixing Zhao. The project manager is Kamran Arif.

Abstract

In this work, we construct datasets of measured internet speeds from two different organizations, MLAB and Ookla, in the year 2020. MLAB measures speed when someone quires google along the lines of "how fast is my internet," and measures a simulated network request as if it propagated across a significant portion of the larger internet network. Ookla, on the other hand, measures speed at speedtest.net, and presents a measerument of someone's local ISP server's speed. We further analyze this data on a per municipality basis, and the MLAB data on a per provider basis [note: some of this work remains to be done for deliverable 4]. We present descriptive statistics of internet speeds during 2020, and maps of upload and download speeds for each municipality in the state of Massachusetts. We also present this data as correlated against household income data from the 2014-2018 census. Finally, we discuss a number of key findings in the analysis of these datasets. First, there is a significant difference between measured Ookla speeds and measured MLAB speeds. Second, there exists an upwards correlation between MLAB broadband speeds and median household income; as median household income increases, average broadband speed increases as well. Third, the vast majority of municipalities are significantly under the desired 100/100 download/upload speeds in mega-bits-per-second (Mbps), with many not even reaching 50 Mbps. Fourth, there exists significant disparity in broadband coverage across the state. [note: some of this work remains to be done for deliverable 4]

Motivation

The Metropolitan Area Planning Council (MAPC) provides planning capacity to municipalities within Massachusetts in a number of different capicities. Recently, they have turned their attention towards broadband, by trying to help municipalities better understand the available internet broadband within their region. Our team joined the MAPC in this endeavor, to collect and analyze broadband data from MLAB and Ookla.

While it may seem trivial, it is actually not so easy to answer questions about internet broadband such as:

- How fast is my internet?
- Is my internet fast enough?
- What providers are available to me, and are there differences in their broadband speeds?

For example, measuring internet speed might amount to a simple measure of the speed of a local ISP server, or be as complex as measuring a real-time observed speed of a fetch request from some geographically distant server. Further, while many ISPs may claim to be able to provide certain speeds, it may not be so clear that the observed speed match the marketed speeds.

Thus, we are working with the MAPC to try and build a dataset that can answer some of these questions, and provide a basis for answering questions that may inform public policy, such as:

- Is there a correlation between median household income and broadband speed?
- Are there municipalities where the available broadband options do not meet requirements for nominal modern internet usage -- say for example, with large zoom calls being nearly ubiquitous for remote schooling and work -- and how limited are they?
- Can we observe differences in broadband access among different providers in different areas of the state?

To answer these questions, we built datasets from Ookla and MLAB of internet speed measurements throughout the 2020 calendar year. We also pulled household income data from the 2014-2018 census. Aside from the aggregation of data, we also provide analyses of the data, which are sufficient starting points for informing policy decisions.

Ookla Dataset

Ookla provides a service at speedtest.net, where clients can test their internet speed. Broadly speaking, the measurement taken is a measure of the speed of a client's connection with their internet provider or ISP. This differs somewhat from the MLAB data, where Ookla is nominally measuring the broadband provided by an ISP, as opposed to simulating a nominal request across the entire network.

Initial Steps

We started by reading the documentation provided by Ookla, hosted on their public facing GitHub page, at https://github.com/teamookla/ookla-open-data. Here, they provide instructions for accessing their open data, which is the portal we are using to analyze their data in Massachusetts. They provide three levels of access: directly via AWS S3, a few download links, and CLI tool for downloads to the terminal. Additionally, data files are provided in two formats: shapefiles and parquet files.

We started by downloading one of the example shapefile links, but found this to be quite confusing. We were not quite able to find a manner to view the data, or begin to understand it for analysis purposes. We then tried the parquet format, which was more accessible. Parquet is a columnar storage system provided via the Apache Organization, and can be used with any tool in the Hadoop archictecture. It is also easy to integrate with python; we managed to find simple "parquet-to-json" and "parquet-to-csv" tools through python Pandas.

AWS and Programmatic Access

For more regular access, we created our own AWS account. We created a root user, and each of us set up our own IAM accounts. For this, we installed the AWS CLI tools, so that we could each access AWS S3 from our local terminals, and for programmatic access.

Date Pre-Processing

In the subsequent data pre-processing procedure, we discovered that the data could be accessed and read directly from the API Endpoints of Ookla with the aids of “GeoPandas” package. The package will read in the shapefiles as a Pandas DataFrame, which makes the data easier to clean. Thus, using the AWS S3 urls provided by Ookla, we used GeoPandas to access the data for each quarter of 2020.

That data, however, was not limited to Massachusetts. In order to downsample the data to just Massachusetts, we also used GeoPandas, on a list of the boundaries for each county in MA obtained from the Census Bureau. Then, since the data is provided by Ookla in shapefile format, we were able to run a joins operation on these two data sets. This yielded just the subset of data that is within a county boundary of some county in Massachusetts. Further, we were then able to label each data point with its associated county.

Labeling by Municipality

After deliverable 1, the Ookla data was labeled with county information, as this information information is publicly available and easy to obtain. Unfortunately, labeling this data by county does not correlate well with the previous work done by MAPC; as much of MAPC's work is grouped by municipality -- a finer grain resolution than by county -- it was necessary to further granularize the data points in Ookla, by labeling each data point with a specific municipality.

Fortunately, MAPC already had a dataset with a geographically defined area for each polygon. That data can be found here: https://datacommon.mapc.org/browser/datasets/390. Using this dataset, we were able to label each row in the Ookla data with municipality information.

Data Schema

This section descibes the schema of the data, and we provide an example data point for reference.

Columns:

- quadkey: a key that identifies the tile
- avg_d_kbps: the average download speed in kilobits per second within the tile
- avg_u_kbps: the average upload speed in kilobits per second within the tile
- avg_lat_ms: the average latency of the tests in this tile.
- tests: The number of tests that contributed to the other values in this tile.
- devices: The number of unique devices that contributed to the data in this tile.
- geometry: list of latitude/longitude pairs, that collectively form the polygonal shape of this tile.
- index_right: joined column index number
- objectid: identifier for object from joined table
- muni_id: numeric identifier for the municipality.

- municipal: name of the municipality
- shape: list of latitude/longitude pairs, that collectively form the polygonal shape of this municipality.

Example data point:

0302332123102031,240473,108651,9,5,3,"POLYGON ((-71.1749267578125 42.252917783302, -71.16943359375 42.252917783302, -71.16943359375 42.2488517007209, -71.1749267578125 42.2488517007209, -71.1749267578125 42.252917783302))",216,228,73,Dedham,8E0800006601000008001000 6A690000B621070001000000A5BAB4CEB2159EC09FB9A411E8F109A0EB0AE896CE019CF7D403C0F702A0BF08FC970488CC09ECD6068CE306ECE90888BF05D0A530ACF90BE4E96B8CA523FCBC4BA4EF1EE8B027A8DA19D0E313D0B024C094AD05F8B0C009FCB97EF4CAEC01F4820AE4F414F0C5

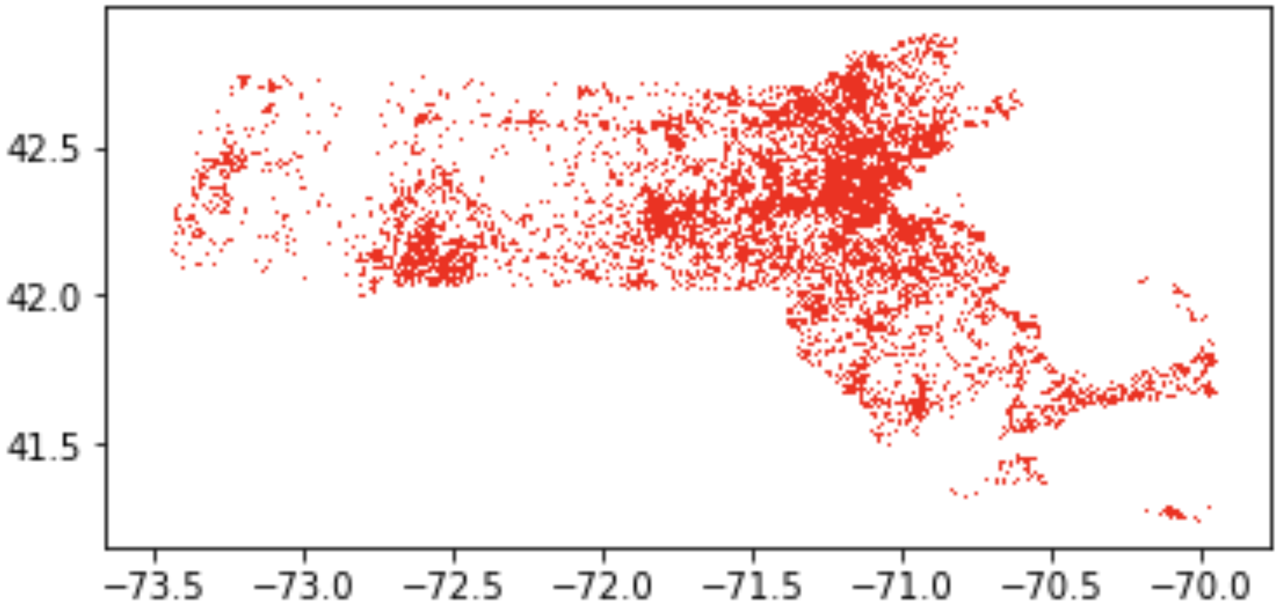
Data Size

Because Ookla provides aggregated information, and we have currently limited our time range to just 2020, the Ookla data is relatively small. We provide the file sizes here

- quarter 1: 588 KB
- quarter 2: 9.2 MB
- quarter 3: 9.5 MB
- Quarter 4: 8 MB

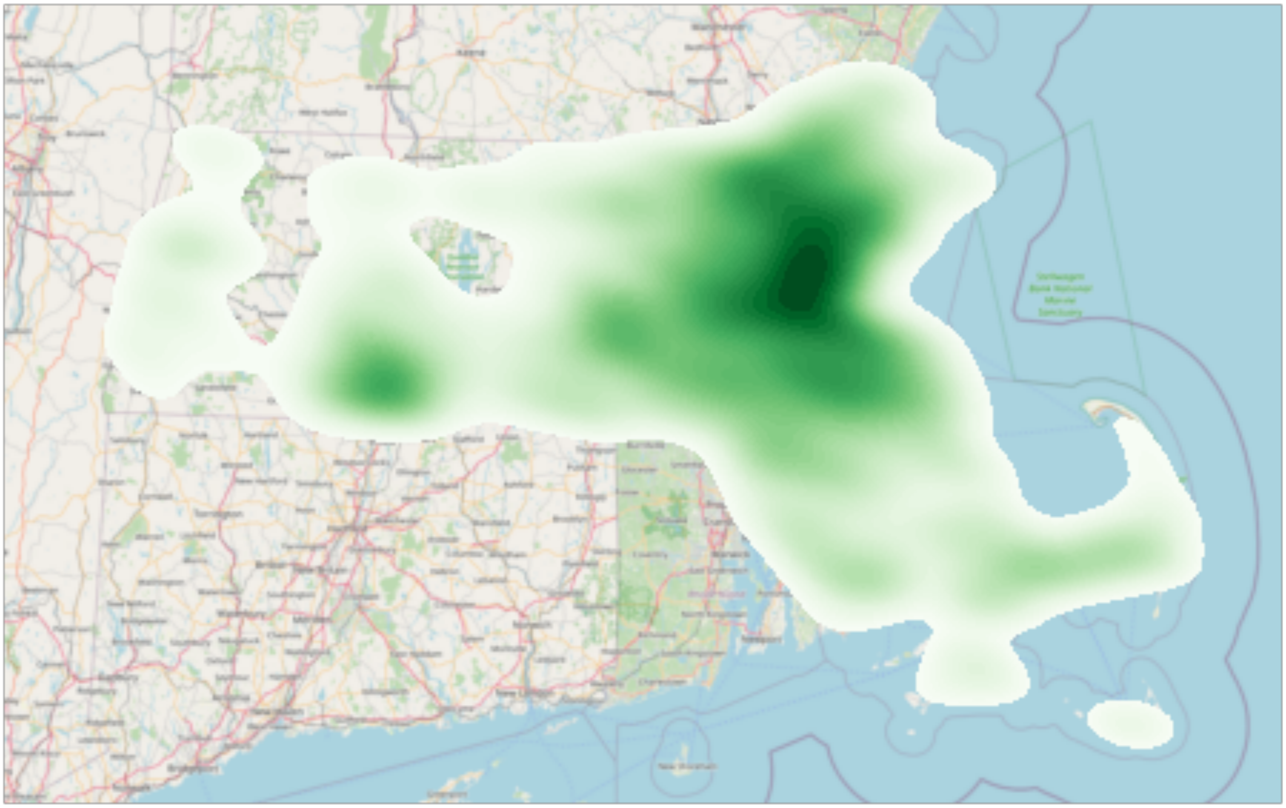
Ookla Geographical Density

Since Ookla data is constructed with tiles labeled with geographical information, we were able to produce density maps showing where in the state the measurements were collected. Here is that map:



The vertical axis is latitude, and the horizontal axis is longitude.

We also present this data as a heatmap:



From the scatter plot of the location of each data point, and the heatmap, we can see that the data around Boston area and Springfield area are denser than average, and there are only a few data points in rural areas. In fact, there are many areas within the state that either don't have coverage at all, or have only minimal amounts of broadband coverage.

Ookla Basic Statistics

Top 25 Cities with Fastest Average Download Speed

City	Average Download Speed, Kbps
West Bridgewater	205943.28276
Montgomery	205546.3
Bridgewater	199698.68388
Gardner	197069.1943
East Bridgewater	191234.58481
Whitman	187003.88316
Ware	185472.76952

Erving	184774.20513
Everett	184351.24286
Dracut	183783.50973
Palmer	182919.57853
Ashby	181227.69079
Revere	180672.46409
Peabody	179239.06
Haverhill	178558.80971
Lowell	177470.17117
Norton	176841.28748
Chelsea	175945.49351
Fall River	175309.47971
Monson	174890.07237
Milton	174573.90667
Groveland	174141.37981
Westminster	173648.61179
Boston	173572.17082
Middleborough	172813.69739

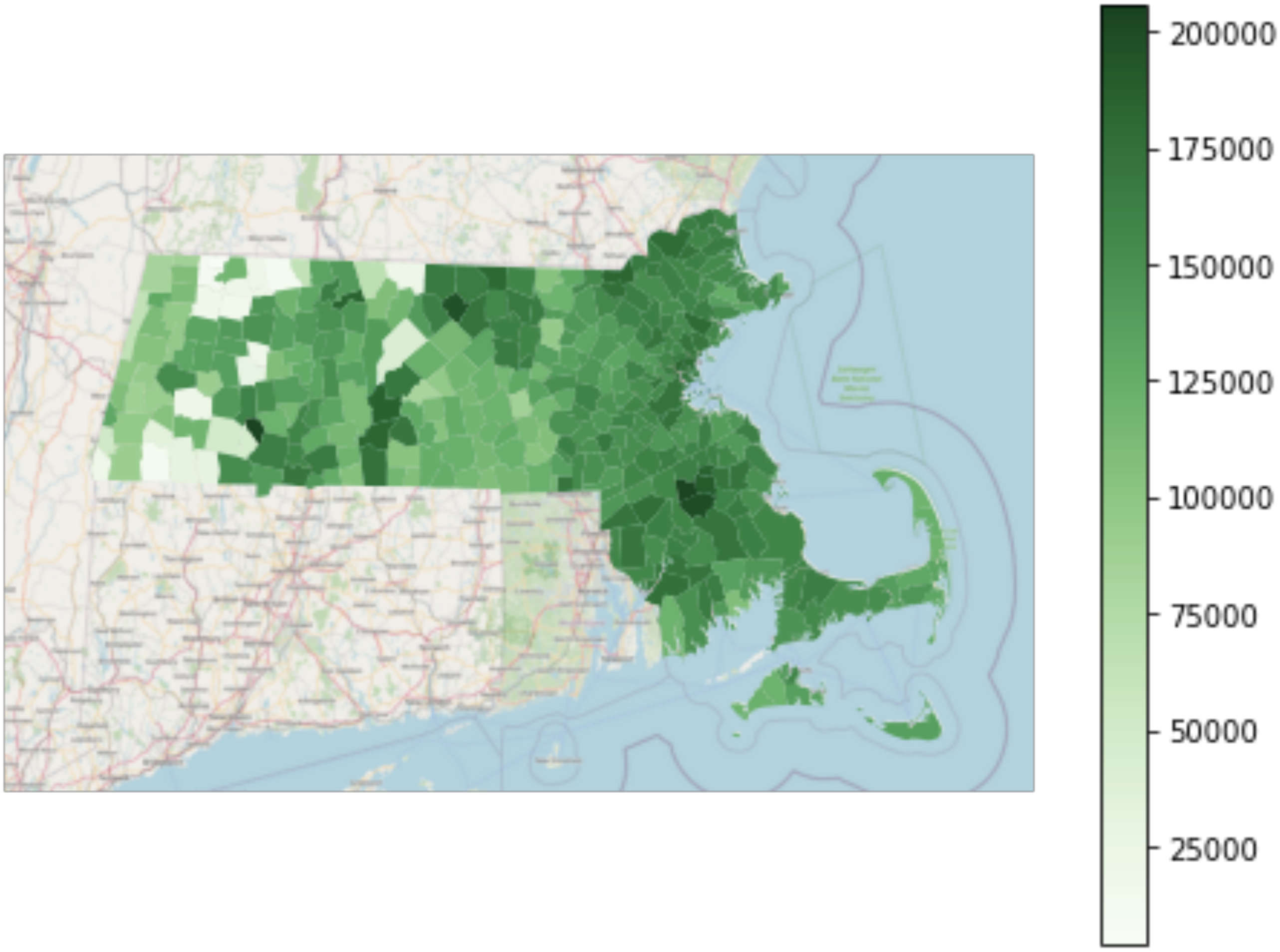
Cities with the Most Average Number of Devices

City	Average Number of Devices
Somerville	24.791
Cambridge	21.39189
Brookline	18.10905
Boston	16.2856
Malden	15.30633
Everett	15.03333

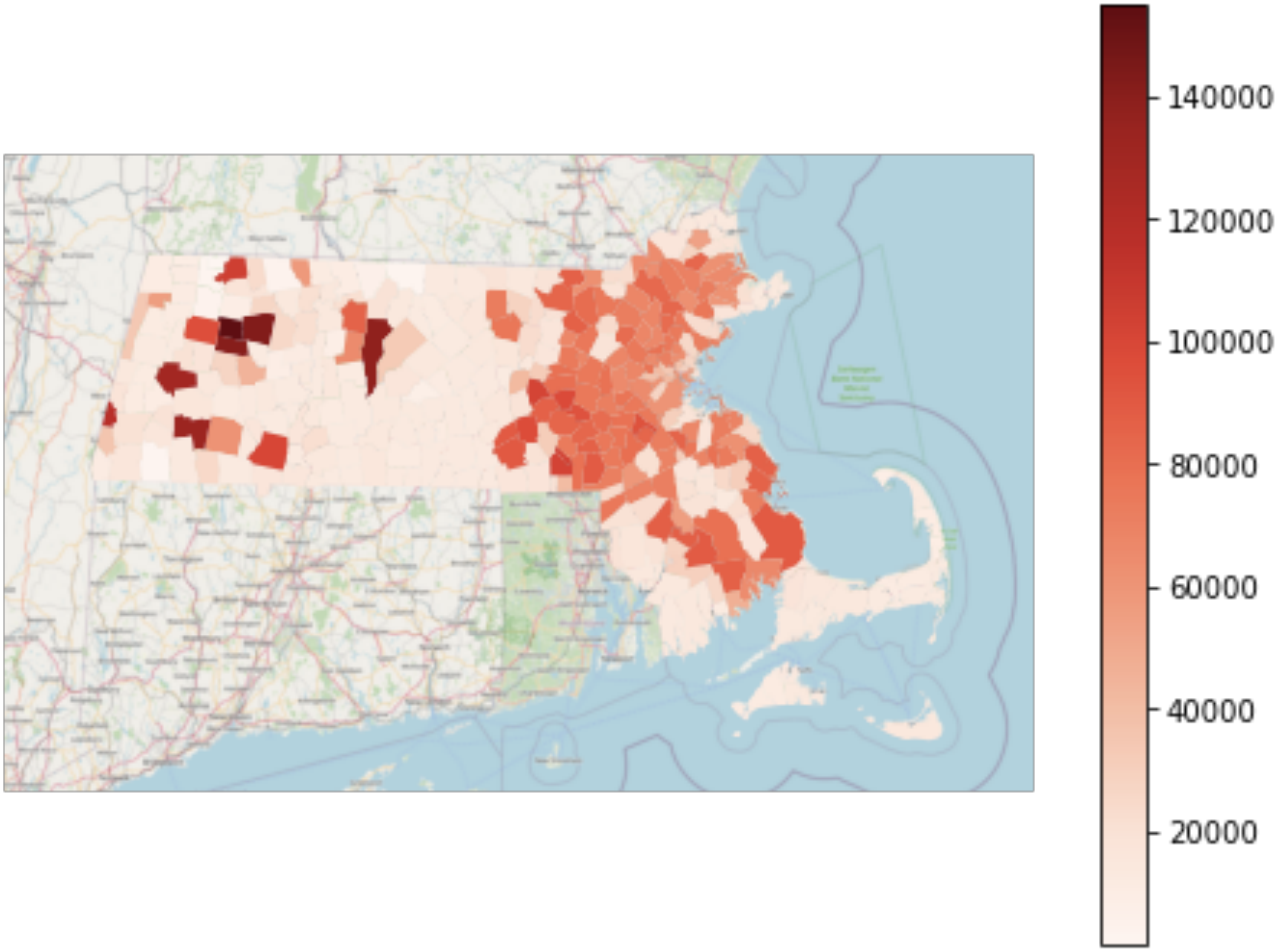
Chelsea	14.59091
Watertown	13.70169
Medford	11.64208
Arlington	11.12895
Revere	10.8232
Quincy	10.56224
Belmont	10.22945
Newton	9.78877
Lawrence	9.05793
Lowell	7.88031
Provincetown	7.71429
Waltham	7.36449
Melrose	7.34375
Winthrop	7.25373
Lynn	7.14741
Winchester	7.04709
Swampscott	6.73604
Salem	6.61962
Worcester	6.52941

Ookla Upload and Download Speeds

Once we labeled the Ookla data on a per municipality basis, we were able to plot internet coverage for the entire state, and generated the following plots:



Here, we chart the download speeds in kilo-bits-per-second (Kbps), per municipality. Each city is colored by the average download speed across 2020, where darker green means faster speeds (up to 200,000 Kbps), and lighter green/white means slower download speeds. As can be seen, there is a significant difference in observed speeds for the various municipalities.



Here, we see the upload speeds, per municipality in MA. There are a number of items to note:

- There are large discrepancies between download and upload speed, in many cities.
- In general, upload speeds are much slower than download speeds. The max download speed was near ~200,000 Kbps, while the max upload speed is closer to ~140,000 Kbps, nearly 60Mpbs slower!
- There is an entire section in the middle of the state where there is effectively no upload broadband available. Within this region, many of the municipalities have respectable download speeds, despite the lack of upload abilities.

The MAPC was concerned that there has existed an imbalance between download and upload speeds for some time, because people often care more about download speed than upload speed. Unfortunately, this problem is particularly important during the present time, since the due to the pandemic, there is an extensive amount of video communication ongoing for both business and education.

MLAB Dataset

The Measurement Lab is an open source project, and aims to advance internet research by providing useful information to anyone about their internet performance. Notably, if you type something along the lines of "how fast is my internet" into Google, MLAB will execute the operations necessary to measure the speed. Unlike Ookla, MLAB attempts to give a more realistic sense of internet speeds, within the context of the larger internet newtork, and not limited to the speeds provided by a specific ISP.

MLAB is particularly useful, because all of their data is accessible for free, and their tools are entirely open source.

Data Access and DownSampling

The MLAB data is provided primarily through SQL query access, via Google BigQuery. Notably, the size of the data is massive. The client provided us with some example queries; we executed a modified version of these for all of MA in the year 2019, and some ~680 GB of data was returned.

The MLAB data is quite large, and thus it was necessary to downsample the data, so that we could work on it effectively. To do so, we sampled a small, but representative portion of the data: we aggregated all measurements in 2020 that occurred within a time range of 8:00am-8:30am, 12:00pm - 12:30pm, 3:30pm-4:00pm, and 8:00pm-8:30pm. The idea here is that these are normal times when people are accessing the internet, and importantly, during school hours when students may be expected to access the classroom via Zoom, or some other digital streaming interactive service. We argue that this downsampling approach is reasonable, because:

- The sampled times are during critical hours of the day
- The amount of data should be siginificantly smaller than the entire 2020 data set.
- It seems a fair assumption that these times are representative of normal to heavy broadband use, which is the subset of data that is most important.

MLAB Automated Scripts

Google BigQuery provides a webportal where SQL queries can be run, and they also provide client libraries where queries can be executed against BigQuery from within an external script. It was a bit tricky to set up, but we managed to obtain BigQuery access through python, using the Google Cloud SDK.

With this, we set up the following query:

```
SELECT
  a.TestTime AS TestTime,
  NET.SAFE_IP_FROM_STRING (client.IP) AS IP,
  a.MeanThroughputMbps AS MeanThroughputMbps,
  a.MinRTT AS MinRTT,
  client.Geo.city AS City,
  client.Geo.Latitude AS Latitude,
  client.Geo.Longitude AS Longitude,
  client.Network.ASNumber AS ProviderNumber,
FROM
  `measurement-lab.ndt.unified_uploads`
WHERE
  client.geo.CountryCode = "US"
  AND client.Geo.region = "MA"
  AND date BETWEEN "{year}-{month_num}-{day_num}"
  AND "{year}-{month_num}-{day_num}"
  AND (
    (
     a.TestTime BETWEEN TIMESTAMP("{year}-{month_num}-{day_num} {first_hour}:00:00.000", "UTC")
     AND TIMESTAMP("{year}-{month_num}-{day_num} {first_hour}:30:00.000", "UTC")
    )
    OR (
     a.TestTime BETWEEN TIMESTAMP("{year}-{month_num}-{day_num} {second_hour}:00:00.000", "UTC")
     AND TIMESTAMP("{year}-{month_num}-{day_num} {second_hour}:30:00.000", "UTC")
    )
    OR (
     a.TestTime BETWEEN TIMESTAMP("{year}-{month_num}-{day_num} {third_hour}:00:00.000", "UTC")
     AND TIMESTAMP("{year}-{month_num}-{day_num} {third_hour}:30:00.000", "UTC")
    )
    OR (
     a.TestTime BETWEEN TIMESTAMP("{year}-{month_num}-{day_num} {fourth_hour}:30:00.000", "UTC")
     AND TIMESTAMP("{year}-{month_num}-{day_num} {fifth_hour}:00:00.000", "UTC")
    )
  )
```

Note: it was necessary to filter by both date and TestTime, as it appears that BigQuery uses the date field as a method of distributing work across servers. Without this filter, an error was returned.

Data Schema

This section descibes the schema of the data, and we provide an example data point for reference.

Unlike the Ookla data, this data is not aggregated, and each data point represents an individual measurement. This likely means that even with the data provided as is, there is some useful cleanup and preprocessing that can be done, such as aggregating measurements from the same device within the timeframe, and other such things.

Columns:

- TestTime: the time at which the test took place, includes both a date and a time.
- IP: the IP address of the device being tested
- MeanThroughputMbps: average broadband throughput in megabits per second. Note that this is

different than the download and upload speeds of Ookla, since they are measuring slightly different things.
- MinRTT: Minimum round trip time, the time it takes to send a signal or data packet and receive back the corresponding acknowledgment
- City: the city in which the test occured
- Latitude: latitude location of the test
- Longitude: longitude location of the test
- ProviderNumber: the autonomous system number of the nearest server system for the test.

Example data point:

2020-10-01 12:01:22.316165 UTC, TBPhGA==, 5.924981493972454, 21.283, Wellfleet, 41.9289, -70.0186, 7922

Data Size

Since we have limited the MLAB data to 2020, and within just 4 time slots (8:00am-8:30am, 12:00pm-12:30pm, 3:30pm-4:00pm, and 8:00pm-8:30pm), then csv file for MLAB is large, but not unreasonably large. It contains 437432 rows of data, and is approximately 49MB is size.

Labeling MLAB Data

Unlike the Ookla data, the MLAB data was already labeled with municipality, and thus the step of labeling each data point by municipality was unnecessary. The MLAB data was also labeled with "ASNumber," which refers to the number assigned to the Autonomous System which controlled the newtork via which each speed test was conducted. Unfortunately, the data did not come with the name of the organization that operates each Autonomous System, and therefore it became necessary to map each of these numbers to a well defined organization.

To do so, we used a publicly available listing of Autonomous Systems, found here: http://www.bgplookingglass.com/list-of-autonomous-system-numbers

With this information, we added a new column to the MLAB schema, containing the name of the Provider that runs each of the various Autonomous Systems. Unfortunately, this type of information is not obtainable with the Ookla data, and therefore this labeling was particularly important, in order to be able to run analyses on a per-provider level.

MLAB Descriptive Statistics

With the MLAB data labeled with providers, we computed a number of decsriptive statistics over the entirety of the dataset, to get a better understanding of the data contained within. In particular, we produced the following metrics:

Basic Statistics

- average MeanThroughputMbps: 43.7
- median MeanThroughputMbps: 12.6
- mode MeanThroughputMbps: 0   11.8
- Standard Deviation MeanThroughputMbps: 91.66

Top 5 Fastest Providers on Average, with at least 1,000 Measurements

Provider Name	Average Mbps
HGE-NET - Holyoke Gas & Electric Department	192.128319
UUNET - MCI Communications Services, Inc. d/b/a Verizon Business	134.443643
LIGHTOWER Lightower Fiber Networks (LIGHT-141)	82.131921
ASN-QWEST-US NOVARTIS-DMZ-US	49.514844
ALKERMES - ALKERMES INCORPORATED	23.363405

Botton 5 Slowest Providers on Average, with at least 1,000 Measurements

Provider Name	Average Mbps
SPCS - Sprint Personal Communications Systems	2.376271
CELLCO - Cellco Partnership DBA Verizon Wireless	5.019919
ASN-SHREWS - Shrewsbury Electric and Cable Operations	6.070376
T-MOBILE-AS21928 - T-Mobile USA, Inc.	8.656454
RR-NYSREGION-ASN-01 - Time Warner Cable Internet LLC	9.678461

Counts of Measurements per Provider, with at least 1,000 Measurements

ProviderName	Measurement Count
COMCAST-7922 - Comcast Cable Communications, Inc.	182906
UUNET - MCI Communications Services, Inc. d/b/a Verizon Business	89356
ALKERMES - ALKERMES INCORPORATED	68790
CHARTER-NET-HKY-NC - Charter Communications	21136
RCN-AS - RCN	15222

ASN-QWEST-US NOVARTIS-DMZ-US	12355
CELLCO - Cellco Partnership DBA Verizon Wireless	6878
T-MOBILE-AS21928 - T-Mobile USA, Inc.	6453
RR-NYSREGION-ASN-01 - Time Warner Cable Interne...	6096
SPCS - Sprint Personal Communications Systems	3502
ASN-SHREWS - Shrewsbury Electric and Cable Oper...	2579
LIGHTOWER Lightower Fiber Networks (LIGHT-141)	1369
HGE-NET - Holyoke Gas & Electric Department	1130

Top 25 Cities by Measurement Count

City	Measurement Count
Needham	34697
Boston	23807
Somerville	15488
Bedford	14515
Ashland	13688
Cambridge	11789
Devens	9150
Worcester	7709
Springfield	7146
Acton	6645
Dorchester	6479
Watertown	6094
Brighton	5325
Arlington	5212
Brookline	5135
Concord	5022

Newton Center	4791
Wellesley Hills	4639
Lexington	4532
Lowell	4415
Quincy	4402
Waltham	4274
Milton	4024
Framingham	3870
Andover	3795

Splitting MLAB Data into Sub-Datasets by Municipality

Since the desired granularity for the data is on the municipal level, we produced a series of csv files each limited to the MLAB data that was collected only within the relevant municipality. This would allow for easier analysis within each of the municipalities, since the overall data set is quite large. Thus, should some analysis focus on a particular municipality, or a short list of municipalities, these files would come in handy.

There are 101 municipalities within MAPC purview, and therefore we produced one output csv file for each of these. They each have the exact same schema as the overall MAPC data, but with data limited to the relevant municipality. For example, there is a single for Acton MA, with a total of 6645 data points contained.

Computing the Average Broadband Speed Per Provider Per Municipality

Here, we produced a csv file with the average broadband speed of each provider, in each municipality. For example, here is an excerpt from that file:

...

Abington,*BIGLEAF - Bigleaf Networks LLC,*1.9385489829867468,*44.9945*

Abington,*CELLCO - Cellco Partnership DBA Verizon Wireless,*14.1583484092067,*53.14081818181819*

Abington,*"COMCAST-7922 - Comcast Cable Communications, Inc.",*10.609942074630164,*25.27878453038673*

Abington,*LIGHTOWER Lightower Fiber Networks (LIGHT-141),*46.193328643945755,*22.193*

Abington,*"UUNET - MCI Communications Services, Inc. d/b/a Verizon Business"*,175.27244063910456,*14.569966666666653*

...

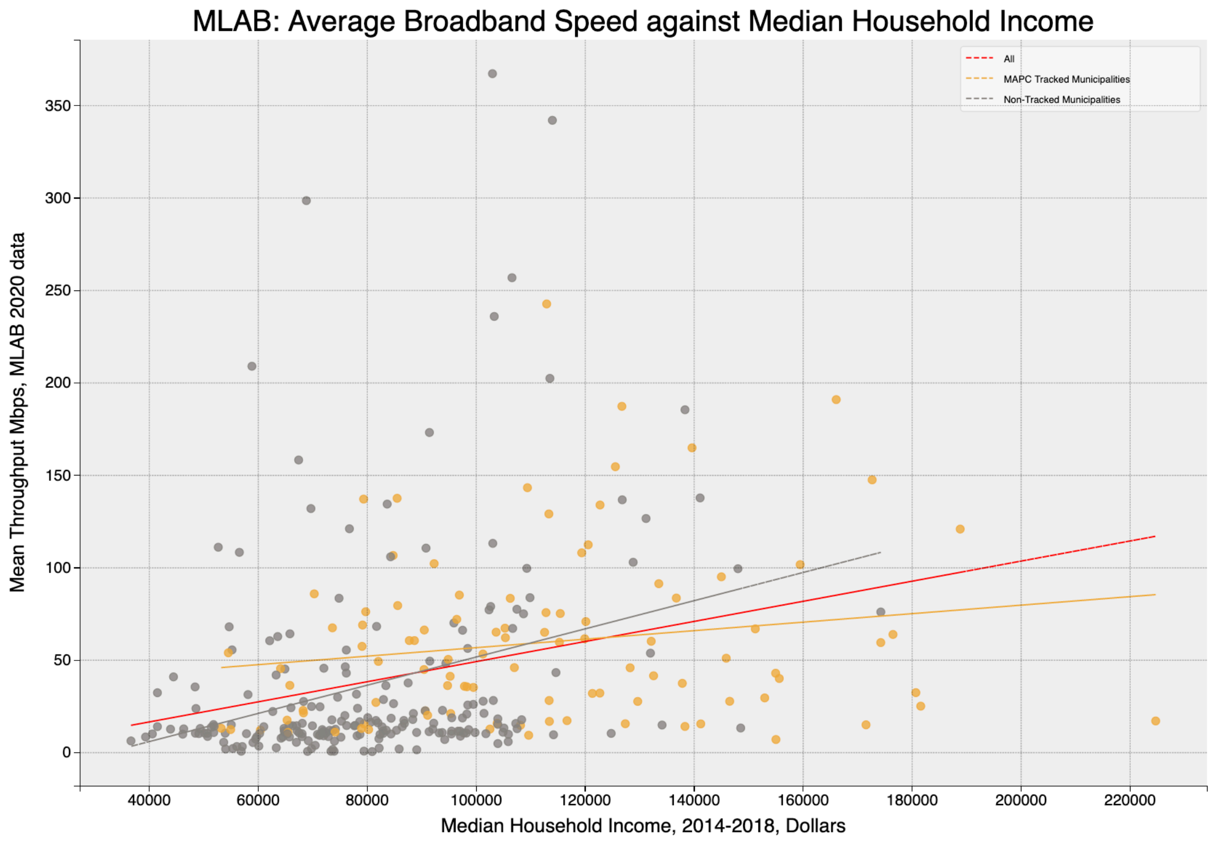
The first column here is the name of the municipality; in this case, we are focused on Abington. The second column is the name of the Provider, or more specifically, the name of the organization that runs the Autonomous System via which a given test was conducted. The third column is the average MeanThroughputMbps, and the fourth column is the average MinRTT, which stands for Minimum Round Trip Time.

As can be seen here, Lightower had an average MeanThroughputMbps of 46.2 megabits-per-second, while Verizon was considerably faster, at 175.27 megabits-per-second.

Results

[For now, as of deliverable 3, we leave this section generally blank, as there is still more work to be done in obtaining results. We have initial results for both Ookla and MLAB, in terms of statewide statistics, upload/download speed maps, and scatter plots of broadband speed per municipality against median household income. However, as there are still more results to obtain, we hold off on formally writing about them here. Nonetheless, we do present two scatter plots, which are the primary outcomes of this deliverable]

MLAB Data and Median Household Income

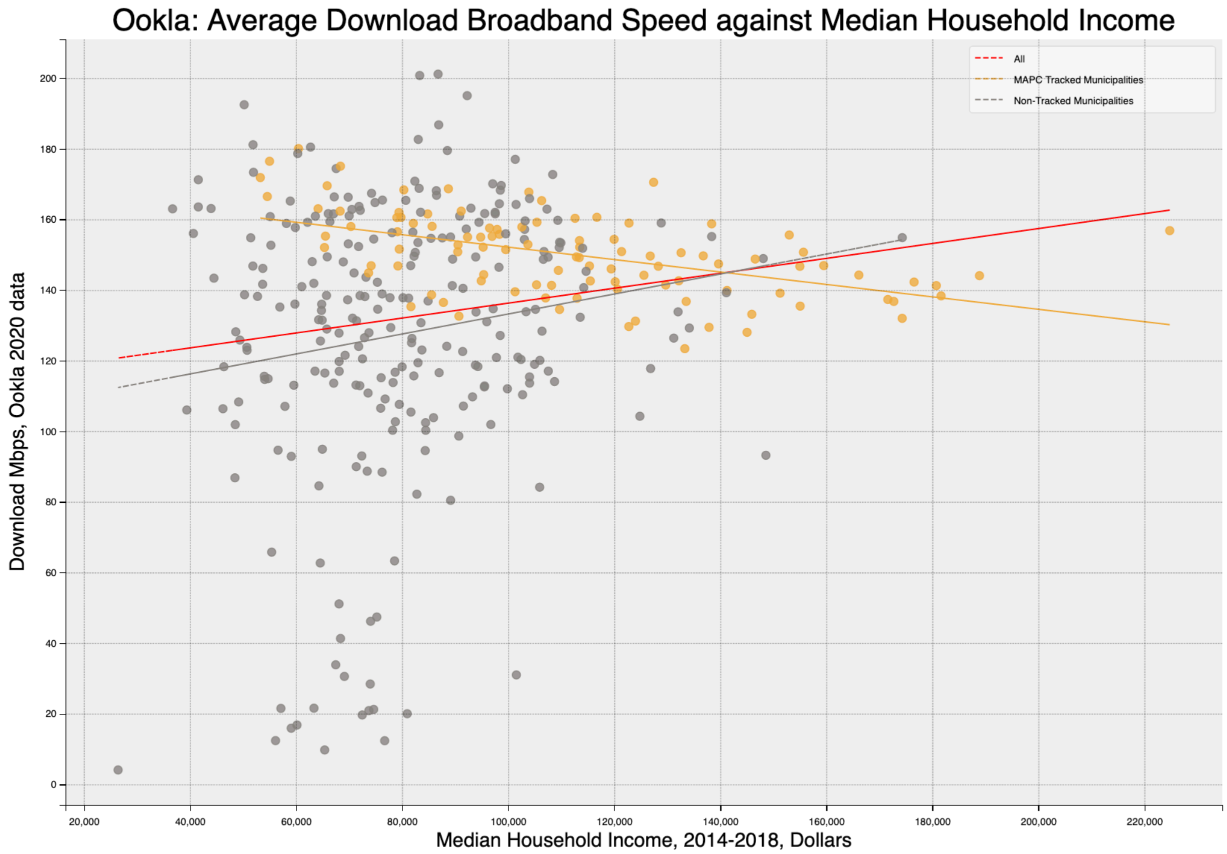


Here, we present a scatter plot of the MLAB data averaged per provider against median household income from the 2014-2018 census. Note that we used data from this census, because the 2020 census data is not yet available. In the chart, each data point is a municiaplity in MA; the data points colored in orange are municipalities within the MAPC's purview, and the grey data points are other cities.

There are a number of key points here:

1. The vast majority of municipalities are beneath the 100 Mbps target broadband speed, many significantly so.
2. There does exist a slight correlation between median household income and throughput Mbps
3. There numerous outliers; some low income municipalities with high throughput Mpbs, and many high income low throughput Mbbs municipalities.
4. It appears that there is a fairly dense clustering of municipalities towards the "bottom left" of the chart; interestingly, there doesn't seem to be much of a difference in terms of broadband speeds for municipalities with median incomes less than $120,000 per year.

Ookla Data and Median Household Income



We also present a similar chart for the Ookla data. Here, there are anumber of key points:

- The biggest difference is in the apparent measured speeds - measuring internet speed against a local ISP server as opposed to a larger network which includes autonomous systems yields much higher measurements.

- Among MAPC municipalities, there actually appears to be a downward trend as compared to median household income, while the opposite relationship exists for other municipalities.

Conclusion and Discussion

[As with the results section, we hold off on writing any conclusions or discussions, as there is still a fair bit of data cleaning, filtering, and further scatter plots to generate.]

Towards Deliverable 4

There is still a decent amount of work left to do for the next and final deliverable, and we discuss that here.

- **Outlier filtering**: it appears there are a number of outliers, and we'd like to filter them out
- **Provider cleanup**: there are many providers within the dataset, and only some of them offer residential or business services. We'd like to remove otehr providers from the data
- **Scatter Plots per Provider**: generate similar scatter plots for each of the primary providers, as we did similarly here with the entire MLAB and Ookla data sets.
- **Time Histories**: given time, generate further such plots for years prior to 2020.
- **Municipality Coverage**: statewide map that includes all relevant information for each municipality: median household income, available providers, overall average speed, per provider average speed.

Summary

In this deliverable, we continue our analysis of the datasets, and present scatter plots of broadband speed against median household income. We also wrote this as an initial draft of the final report.

Checklist:

1. All data is collected: we have collected all the required data. Note: it's possible that if we build a time history, we'll need more data, but we have scripts to do so if necessary.
2. Refine the preliminary analysis of the data performed in PD1&2: as documented above
3. Answer another key question: We examined broadband coverage across the state, and presented maps of download and upload speeds per municipality.
4. Attempt to answer overarching project question: we generated scatter plots of broadband speed against median household income.
5. Create a draft of your final report: as presented.
6. Refine project scope and list of limitations with data and potential risks of achieving project goal: Listed above.
7. Submit a PR with the above report and modifications to original proposal