BU Spark! Project Proposal

This document contains the project proposal for the BU Spark! project, Spring 2021, for the MAPC Broadband Digital Equity in MA. It's purpose is to formally record the stated goals of this project, as presented on pitch day and as discussed during the kickoff meeting with MAPC.

Date: 02/23/21

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

Organization Description

The Metropolitan Area Planning Council (MAPC) is the regional planning agency serving the people who live and work in the 101 cities and towns of Metropolitan Boston.

MAPC tries to provide capacity for planning projects for municipalities. Matt and Ryan and on the software and data side; most of their jobs are along the lines of trail map, or mass bill, the idea being to provide capacity for municipalities where they don't have it.

Project Type

Data Science

Project Description

Broadband work for MAPC is new; they have not had a group work on this type of infrastucture, especially with broadband. The ultimate goal is for municipalities to better understand the role they play in broadband distribution. During the past few months, MAPC has been trying to understand what data is available, what further data needs to be collected, and what leverage municipalities have to affect broadband investment.

The goal in this project is to be able to answer two questions with regards to digital broadband in the state of MA:

1. How is internet broadband speed distributed?
2. Does the data back up the conventional notion of how we think broadband is distributed?

Successful outcomes include a statewide map, based on speed, inclusion, and coverage. The NDIA has already built similar maps for Cleveland OH and Dallas TX, and the goal here is to explore the above questions with this data backed map. Further, the state wants to be able to examine digital redlining from a determininstic perspective. That is, determine if there is a correlation between internet broadband speeds and socio-economic factors.

Data will be drawn from four different sources (more information in the next section); however, this data is not in a state of readiness where it can be explored from a policy perspective. Thus, the first portion of this project is to compile and clean these datas such that MAPC is in a position to begin examining them for policy purposes.

The second part of this project focuses more on digital redlining, correlating broadband speed with FCC data and census data, to help municipalities understand how their permitting capabilities affect the end user. Is the broadband that we currently have good enough? Are we fast enough? Are there some areas that are fast enough, while other aren't, and is there some correlation there with socio-economic factors? These are the types of questions that will be explored in the second part of this project.

Original Description Given on Pitch Day

The MAPC would like to allocate newly released funds from the CARES act towards increasing broadband access across MA. They will decide how best to allocate this money based on their dataset of historical broadband speeds with hundreds of features such as income, ethnicity, % of uploads/downloads, etc. Time series analysis on internet use by hour and day will also be done to capture broad trends.

The second part of the project will specifically focus on Municipal Digital Divide Planning efforts on Gateway cities and analyze differences amongst provider speeds so they can choose the best provider to expand broadband access.

Data Sets

The dataset will be provided by the client from MLab and Ookla, it is part of a larger dataset containing historical broadband speeds across the world coming in at Terabytes of data. However, we will only be working with MA historical data.

Ookla dataset (Broadband speed data): this data can be downloaded from Amazon warehouse storage. Initially, the goal will be downloading it, and extracting out the parts of this data that are relevant to MA.

MLab dataset (Broadband speed data): this data is very large, and can't be simply downloaded as a csv file. Here, the goal initially is to download this data into postgres tables, extract out the part that is relevant to MA, and then provide csv files of this data that they can then use for various applications.

FCC dataset (Broadband provider coverage data) - Shows how many broadband providers cover each census tract. The census tracts will mostly nest within municipalities. FCC data is available to download as a zip file.

Census county subdivision data for MA

At the kickoff meeting, we decided to split this project between the two teams by dividing the data sets: team 1 will work on the FCC and census data, while team 2 (this team), will work on Ookla and MLAB.

Suggested Steps

Step one: Clean and preprocess the MLab & Ookla data for MA, this will involve some format of SQL queries and Pandas preprocessing in Python after. Duplicate the client's approach in querying the data, the approach will be provided by the client.

From the kickoff meeting: what can be said about the data sets? How big are they? Are there data sets that are missed? Start by exploring the data sets, understanding it, and describing them.

Step two: Overlay the MLab, Ookla, and FCC datasets with the Census municipality (county subdivision) data**.** The initial three datasets should have geographic units down to the municipality level.

Step three: Analyze discrepancy in broadband coverage and speeds across MA municipalities using demographics information with the merged dataset above. Is there a noticeable presence of digital redlining - Communities of color receiving poorer coverage and speeds? Setup and conduct a regression test to predict broadband speeds using demographics, income levels, and housing density as predictors. We want to best understand which of these variables contribute towards faster broadband speeds. A further step could also be clustering similar broadband speeds and analyzing their similarities.

Step three: Focus on the Gateway cities - Revere, Everett, and Quincy and compare differences in provider speeds here. The outcome should be visualizations showing the difference. This step will help with their Digital access plan.

Step four: Conduct a time series analysis for the state by month, day, and hour to study trends of when internet usage is concentrated and find possible explanations.

Step five: Summarize findings using data visualizations for the time series analysis and provider speed differences in the gateway cities.

Limitations and Risks

There are a couple of potential risks associated with this project:

- Unclear how much work will be required to simply get the data into a useful state. This is the first goal of the project, but Ryan said that would be a huge accomplishment.
- Further unclear whether or not this will be possible for the entire state, or perhaps if we may focus on a number of smaller, specific areas.
- Not clear if we'll have time to overlay this data on the geographical data from the census.

With that said, Ryan made clear that a successful outcome is simply to produce a map of internet broadband speen in MA. Once we have the data in a useful state, this will be our primary focus.

Previous Work

No significant projects. This is MAPC's first foray into digital broadband speed. Matt has spent some time with the data in the past few months, but outside of that, there are no other works that this team need review.

Strategic Questions

1. What are the discrepancies in coverage and speeds among MA municipalities? Identify key features
2. Is there presence of digital redlining? Black communities and communities of color receiving poor coverage relative to the rest.
3. How do broadband provider speeds vary in the three gateway cities - Revere, Everett, and Quincy?
4. What are the hours of highest internet activity? Test this as a hypothesis and find possible explanations.
5. What are the leading predictors for higher broadband speeds in MA?

Additional Information

Tools & Methods

Data pre-processing: Pandas, NumPy for processing and cleaning the data. BigQuery to create SQL queries and obtain subsets from the database.

Machine Learning: scikit-learn, pytorch for machine learning and regression tools

Data Visualization: Matplotlib, Seaborn, Tableau for all kinds of interactive visualizations

Related Links & Resources

- Running List of potential report visualizations and reports
- Running List of terms and definitions
- NDIA Cleveland AT&T Digital Redlining Report

- ISLR Video discussion about NDIA Report
- ISLR blog on the topic of broadband
- RWJF thread with link to Brookings Report

Useful links

MAPC has provided a number of useful resources - their repositories, websites, access to internal tools, etc. These are the links that were provided to us:

- https://datacommon.mapc.org/
- https://www.verizon.com/coverage-map/
- https://airtable.com/shrv7Uv7LMWkKDW1b
- https://airtable.com/shrZkjM3DUASjEVmk
- https://airtable.com/shrML6GmsFUwwRQpo
- https://datacommon.mapc.org/calendar/2020/december
- https://datacommon.mapc.org/calendar/2020/april
- https://www.measurementlab.net/data/
- https://www.speedtest.net/insights/blog/announcing-ookla-open-datasets/
- https://www.fcc.gov/document/fcc-annual-broadband-report-shows-digital-divide-rapidly-closing

Client Meetings

We will be meeting with Ryan and Matt every other week, on Friday at 1:30pm. This was the time of the kickoff meeting, and we decided to continue meeting biweekly at this time.

Project Manager Meetings

We will be meeting with Kamran, the project manager, every week, Tuesday afternoon at 4:30pm.