Baystate Banner: LatinX Republican Support (Spring 2021)	
Contact	**Senior Editor: Yawu Miller** yawu@bannerpub.com **Project Manager: Lingyan Jiang, lingyan@bu.edu** **Staff Lead: Steve Backman, sbackma1@bu.edu** **Team Members: Ngozi Omatu (nomatu@bu.edu), Song Xie (sxie2@bu.edu), Matan Ziegel (mziegel@bu.edu), Gil Lotzky (glotzky@bu.edu), Anna Xie (annaxyw@bu.edu)**
Organization	Baystate Banner
Organization Description	The Bay State Banner is an African American-owned newsweekly that reports on the political, economic, social, and cultural issues that are of interest to African American and English speaking Latinos in Boston and throughout New England.
Project Type	Data Science
Project Description	The client would like to understand the components of support for Republicans over the time period of 2014-2020 including Presidential elections and the Governor's races. The goal of the project is to find whether or not there is a significant difference in the voting pattern of the LatinX community from 2016 to 2020. We will be collecting data from cities with a majority LatinX population and non-LatinX populations. We will then compare the two populations to see whether there is a significant difference in voting patterns between the LatinX community and the control group (which has voted Republican consistently in the past). If there is a significant difference between the communities, we will analyze the sub groups within the LatinX population to identify the cause. The main goal is to conclude which LatinX voters changed their votes and which voters stayed consistent.
Data Sets	2014 & 2018 Mass. Demographic dataset+Election Results Scattorpoints of the shift in Democrats and Republican votes for Presidential Bar Data for the Government Demographic Cleaned Datasets

Suggested Steps	• **Step one:** Collect government election results & demographic data of majority LatinX towns in Mass. • **Step two:** Use pandas to read through csv files and convert to data frames. Only keep key attributes for demographic data (Tract, estimated total pop, LatinX sub-group pop, and voting age pop (18+)) and election data (city zip codes, city names, total votes per candidate, total votes). Created a map of the Massachusetts counties using the tracts and geopanda • **Step three:** Relate city election data to city demographic data (either via each city's precincts). Analyze different kinds of how did their support change from 2014 to 2018 • **Step four:** Complete analysis on correlations between how these towns voted • **Step five:** Complete data visualizations by using the election results from each major city to highlight them on the geopanda map.
Strategic questions	1. How has support for Trump shifted across the LatinX population from 2016 to 2020? Was there a significant shift? 2. What is the breakdown of LatinX sub-groups in their support for Democratic vs. Republican candidates? 3. Which LatinX groups exhibited changes in their votes and which groups remained the same?
Additional Information	**Tools & Methods** *Data pre-processing:* Pandas to convert csv files into data frames, NumPy for processing and organizing the pre-processed data *Data Visualization:* Matplotlib, Seaborn, Tableau for all kinds of interactive visualizations, Geopanda, Microsoft Excel **Weekly Meeting Schedule:** **Wednesdays 11am - 12:30pm EST**

GitHub repo with all python/csv files: https://github.com/glotzky/LatinXBaystate.git

Refine the preliminary analysis of the data performed in PD1&2:

Tract Data:

Cleaning of Data: The data used in this deliverable consisted of the population data for all tracts in Massachusetts for 2014,2016,2018, and 2019. Some basic deletion of unnecessary rows was performed in excel. The data was read into dataframes using pandas read_csv function. The necessary population data attributes specific to LatinX population, general population, and tract

number were kept. The columns were then renamed for simpler understanding and converted the Tracts from float values to string values, as they are identifiers not numerical values. Then the 2014 and 2018 and the 2016 and 2019 data frames were merged together for further analysis to be relevant to the presidential and governor elections. New columns to describe the percentage changes in the various populations were created. Additionally, new latitude and longitude columns with arbitrary values were created in order to convert the dataframe to a geopandas dataframe to combine with the geospatial information obtained from the census website. We then read in the Massachusetts Tract shapefiles using geopandas read file function and convert the tracts to floats then strings in order to directly match the values from the population data with the shape files. This is very necessary as the dataframes will not merge correctly if it does not find an exact match for its respective tracts.
Merging Data: Using geopandas merge, we are able to combine our population data with our geospatial tract data. This is very important in order to correctly map any of the populations or for further merging with election data. We then drop all unnecessary attributes, and rename columns for further processing with geopandas. This completes the necessary processing for the Tract and population data.

Precinct Data:

Cleaning Data: The presidential election data for 2016, 2020 and the governor election data for 2014 and 2018 are read into dataframes using read_csv. The precinct shape file for Massachusetts was read into a geopandas dataframe. We then rename columns to be consistent with categorical names(Democratic and Republican). We drop all empty cells using dropna. An error had appeared when combining the data frames together, so it was important to diagnose the problem. No precincts with missing wards were appearing after merging the data so we attempted to look closely at those cases. After setting boolean statements for 5 cells it was showing the Wards were False when set equal to each other. After printing out a sample from each set, the presidential data for 2020 had an empty space in front of it that was impossible to notice from printing the data set. The spaces were stripped and the was ready to merge with the geospatial data frame for precinct data.
Merging Data: The matching presidential data and governor data were successfully merged together into a pandas dataframe. Arbitrary latitude and longitude columns were appended to the merged dataframe in order to convert it to a geopandas dataframe. We drop the unnecessary columns from the geopandas dataframe and set the city/town values to uppercase in order to exactly match the cases with the geospatial geopandas dataframe for the precincts. We replace the None values with '-' in order to create exact match cases. We then successfully merged all the presidential election and governor election data with their respective precinct geospatial geopandas dataframes. Further cleaning was done through renaming the attributes and dropping unnecessary information.

Final Dataset:

Now that we have two geopandas dataframes containing the presidential voting information with their respective precincts and geospatial information, the governor voting information with their respective precincts and geospatial information, and the population information with their respective tracts, we can now merge all the dataframes together using a spatial join. A spatial join was performed on points of intersection. The data sets were then converted into csv files where further analysis can be performed. No further issues with the dataframes were present.

```
In [63]: test16['Ward']==test20['Ward']

Out[63]: 0    False
         1    False
         2    False
         3    False
         4    False
         Name: Ward, dtype: bool

In [64]: print(test20['Ward'][0])
         print(test16['Ward'][0])

          –
         –

In [65]: pres2016['Ward'] = pres2016['Ward']#.str.replace('–','0')
         pres2020['Ward'] = pres2020['Ward'].str.strip(' ') #.str.replace('–',)
```

Debugging the Ward issue

```
type(mergedP)
```

```
pandas.core.frame.DataFrame
```

```
mergedP['Latitude'] = -40.266666
mergedP['Longitude'] = 72.3452
```

```
mergedPG = gpd.GeoDataFrame(
   mergedP, geometry=gpd.points_from_xy(mergedP.Longitude, mergedP.Latitude))
```

```
mergedPG = mergedPG.drop(columns = ["Latitude","Longitude"])
```

```
type(mergedPG)
```

```
geopandas.geodataframe.GeoDataFrame
```

Converting pandas dataframes to geopandas dataframes with arbitrary latitude and longitude columns

```
MA_t = gpd.read_file("CENSUS2010_BLK_BG_TRCT_SHP/CENSUS2010TRACTS_POLY.shp")
```

```
MA_p = gpd.read_file("wardsprecincts_poly/WARDSPRECINCTS_POLY.shp")
```

Reading in shapefiles

```
FinalData.columns
```

```
Index(['STATEFP10', 'COUNTYFP10', 'TRACTCE10', 'GEOID10', 'NAME10',
       'NAMELSAD10', 'MTFCC10', 'ALAND10', 'AWATER10', 'INTPTLAT10',
       'INTPTLON10', 'AREA_SQFT', 'AREA_ACRES_left', 'POP100_RE', 'HU100_RE',
       'LOGPL94171', 'LOGSF1', 'LOGACS0610', 'LOGSF1C', 'SHAPE_AREA_left',
       'SHAPE_LEN_left', 'geometry', 'Tract', '% Point Change in LatinX Pop.',
       '% Point Change in Total Pop.', '% Point Puerto Rican Change',
       '% Point Mexican Change', '% Point Cuban Change',
       '% Point Other LatinX Change', 'Total Population 2016',
       'LatinX Population 2016', 'Mexican 2016', 'Puerto Rican 2016',
       'Cuban 2016', 'Other LatinX 2016', 'Total Population 2019',
       'LatinX Population 2019', 'Mexican 2019', 'Puerto Rican 2019',
       'Cuban 2019', 'Other LatinX 2019', 'index_right', 'WP_NAME', 'WARD',
       'PRECINCT', 'DISTRICT', 'POP_2010', 'TOWN', 'TOWN_ID', 'AREA_SQMI',
       'AREA_ACRES_right', 'YEAR', 'SHAPE_AREA_right', 'SHAPE_LEN_right',
       'City/Town', 'Pct', 'Ward', 'Democratic 2016', 'Republican 2016',
       'Total Votes Cast', 'Democratic 2020', 'Republican 2020',
       'Total Votes Cast 2020', '% Point Change in Democratic Votes',
       '% Change in Total Votes', '% Point Change in Republican Votes'],
      dtype='object')
```

```
FinalData.shape
```

```
(10567, 66)
```

Successful Merged Dataset with all columns necessary for geospatial mapping and statistical analyses.

Regression Cleaning:
We read the .CSV output from Final Dataset to turn it into a dataframe for further processing. Each column in the dataframe is currently a string which allows us to remove anything that prevents use from converting each column's numbers to floats, such as unnecessary spaces, NaN, and commas. After cleaning, we convert the values of the necessary columns (any column containing voting data and demographic data) to floats using .to_numeric(). This allows us to use columns to conduct calculations such as division.

Regression:
We calculated % point change over a 4 year interval via:

% Point Change in Democratic Support:
(Dem votes 2020 / total votes 2020) - (Dem votes 2016 / total votes 2016)

% Point Change in Republican Support:
(Rep votes 2020 / total votes 2020) - (Rep votes 2016 / total votes 2016)

% Point Change in subgroup:
(Subgroup pop 2020 / total pop 2020) - (subgroup pop 2016 / total pop 2016)

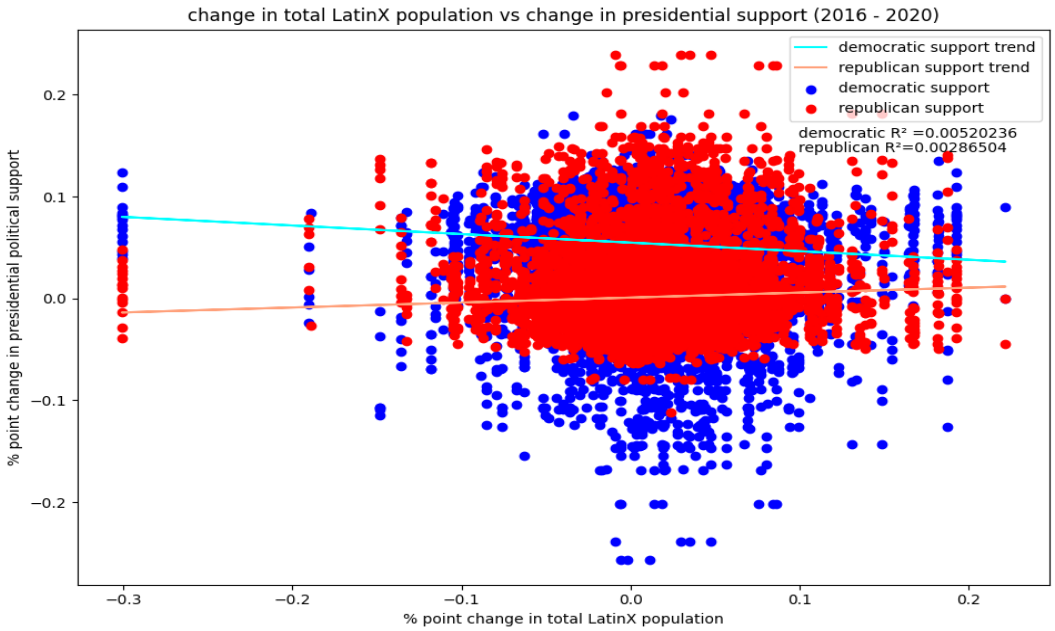
% Point Change in Total LatinX Pop:
(Sum of subgroup pop 2020 / total pop 2020) - (Sum of subgroup pop 2016 / total pop 2016)

We then run linear regression on these new dataframe objects which allows us to plot the trend based on the % point change of political support vs. % point change in LatinX population.

Strategic Question 1:

How has support for Trump shifted across the LatinX population from 2016 to 2020? Was there a significant shift?

Graph 1: LatinX Population vs. Political Support (Presidential Election) Changes (2016-2020)

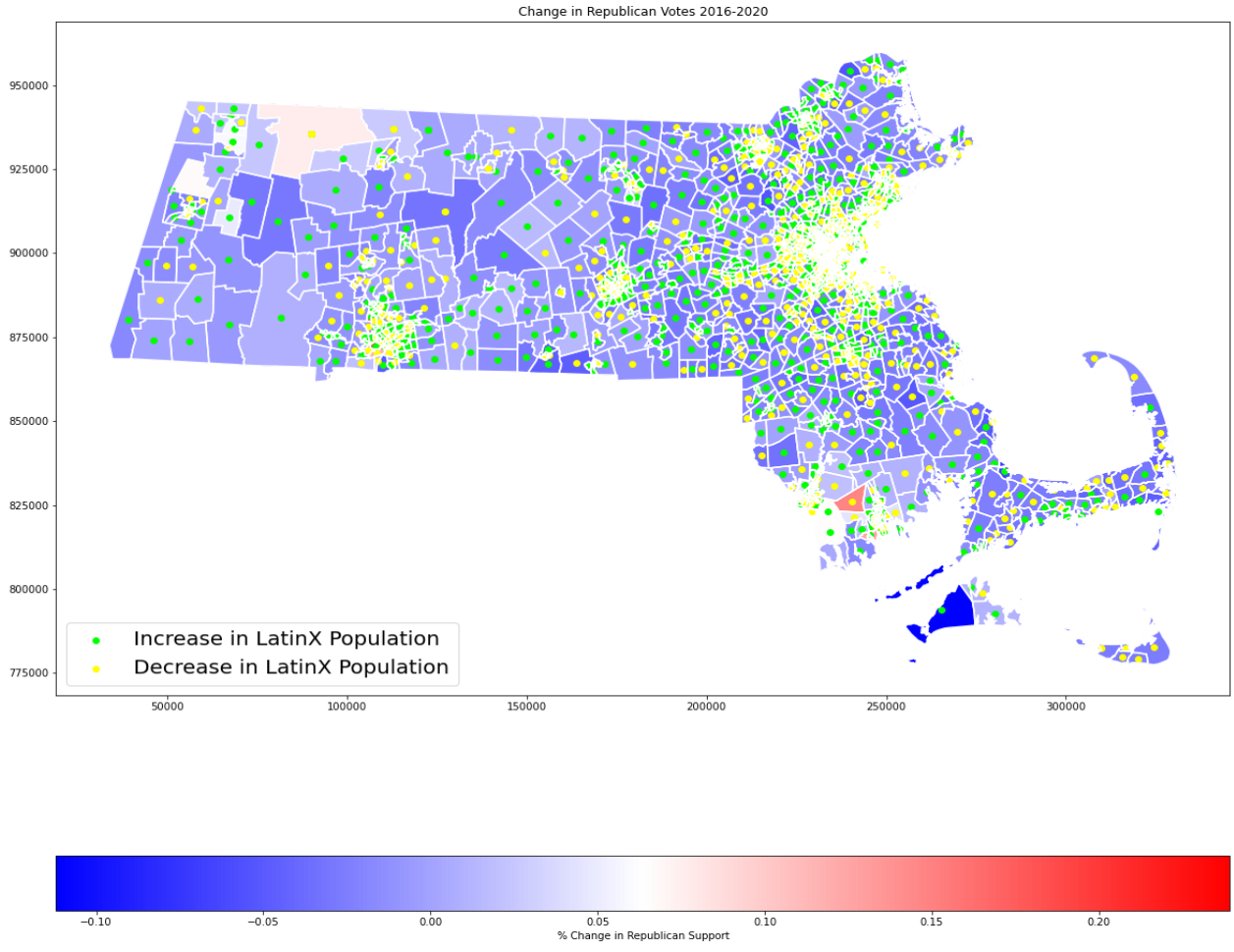


Our team analyzed changes in both the LatinX population and presidential election support within each Tract in Massachusetts. We analyzed how changes in the total LatinX population correlates with changes in presidential election results from 2016 to 2020. We used a linear regression model to identify whether there was a correlation.

For the changes in Republican support relative to the changes in total LatinX population, the trend line is almost flat which indicates that changes in LatinX population had little to no effect on the changes in Republican support. This is also supported by the trend line's R-Squared value of 0.0029 which is extremely close to zero. This means that almost none of the

variance in the change in Republican support can be explained by the changes in the LatinX population. Therefore, we cannot conclude that changes in the total LatinX population have an effect on Republican support from 2016 to 2020.

The geospatial mapping of changes in Republican votes and changes in LatinX Population shows that 23.85% of tracts in Massachusetts that experienced an increase in LatinX population also saw an increase in Republican Support. In addition, 36.60% of tracts in Massachusetts experienced an increase in LatinX population and a decrease in Republican Votes, 24.16% of tracts experienced a decrease in LatinX population and a decrease in Republican Votes, and 14.57% of tracts in Massachusetts experienced a decrease in LatinX population and an increase in Republican Support. This also supports our inference that changes in LatinX population has no direct effect on changes in Republican support.



Strategic Question 2:

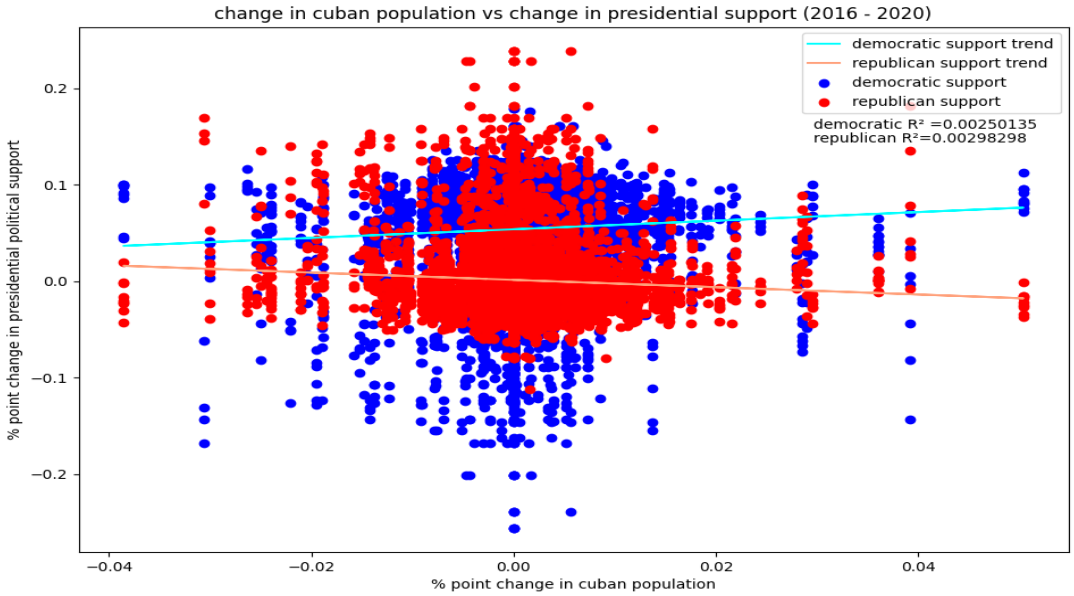
What is the breakdown of LatinX sub-groups in their support for Democratic vs. Republican candidates?

None of the LatinX subgroups in Massachusetts show significant support towards one specific presidential party. The linear regression models show that the variation in presidential support from 2016 to 2020 is not explained by the change in any of the subgroup populations. Since there is no data on specific subgroups and how each subgroup had voted, our conclusion is derived from the correlation between the changes in LatinX subgroup populations and changes in Democratic and Republican support. Even if there appears to be a correlation between changes in LatinX subgroup population and changes in political support for a specific political party, we must recognize that correlation does not imply causation.

**The regression models for each LatinX Subgroup are shown below. These models additionally help to answer the question which LatinX voters changed their votes and which subgroups if any stayed the same.*

Attempt to Answer Overarching Project Question:

Graph 2: Presidential Support v. Cuban Population Change (2016-2020)



Democratic Support:

R-squared: 0.0025

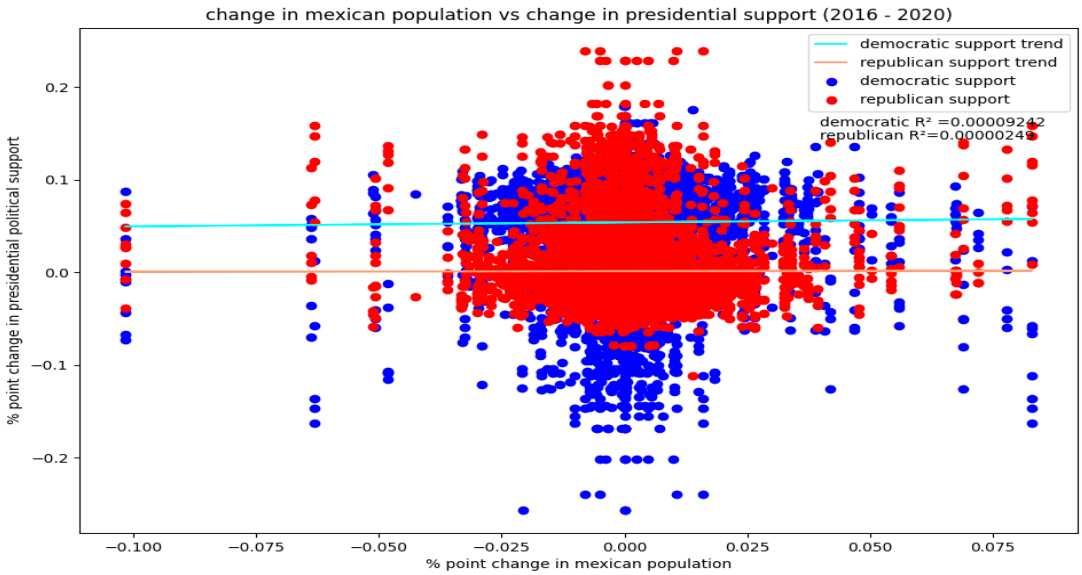
Conclusion: R-squared is ~0, therefore, we cannot conclude that the changes in Democratic support from 2016 to 2020 can be explained by the changes in the Cuban population.

Republican Support:

R-squared: 0.0029

Conclusion: R-squared is ~0, therefore, we cannot conclude that the changes in Republican support from 2016 to 2020 can be explained by the changes in the Cuban population.

Graph 3: Presidential Support v. Mexican Population Change (2016-2020)



Democratic Support:

R-squared: 0.00009

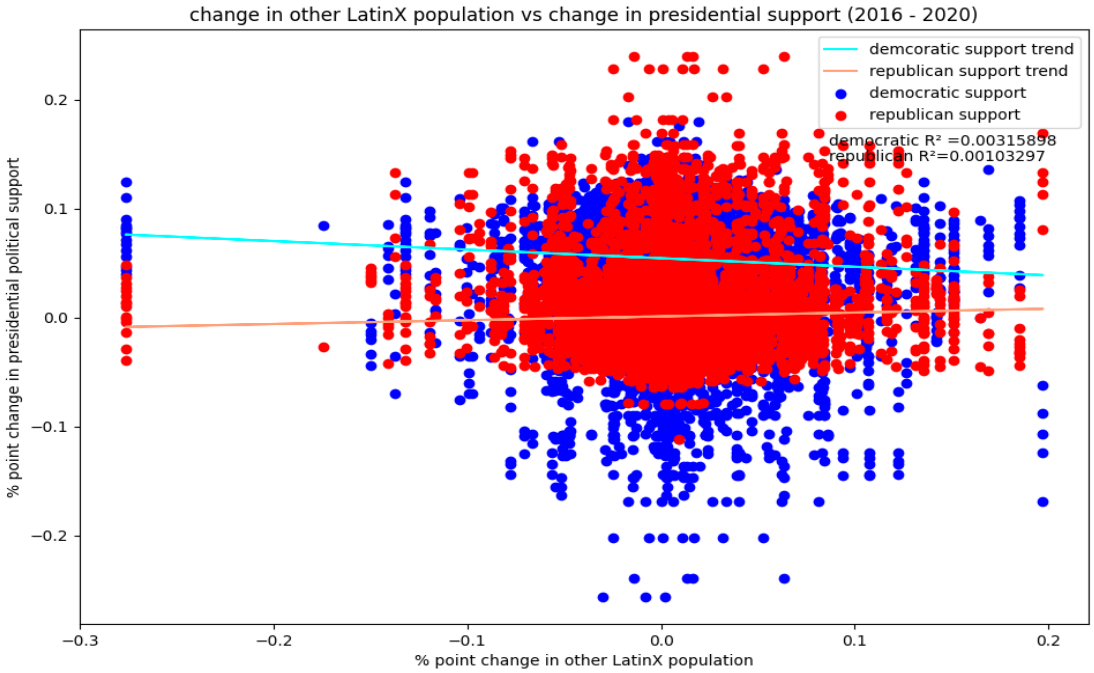
Conclusion: R-squared is ~0, therefore, we cannot conclude that the changes in Democratic support from 2016 to 2020 can be explained by the changes in the Mexican population.

Republican Support:

R-squared: 0.000002

Conclusion: R-squared is ~0, therefore, we cannot conclude that the changes in Republican support from 2016 to 2020 can be explained by the changes in the Mexican population.

Graph 4: Presidential Support v. other LatinX Population Change (2016-2020)



Democratic Support:

R-squared: 0.003

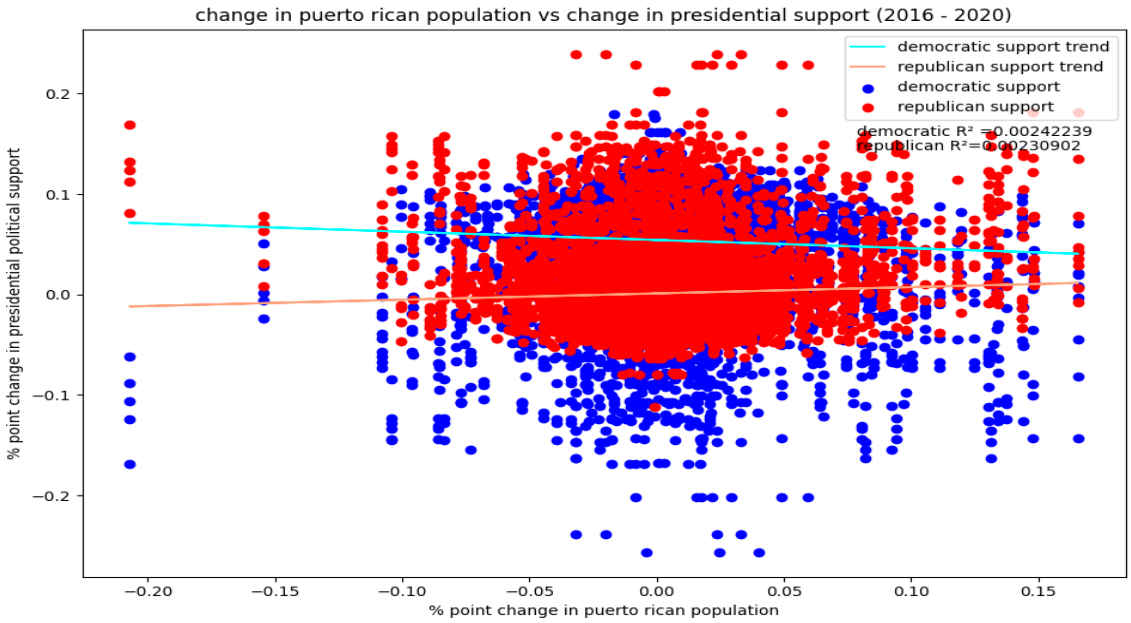
Conclusion: R-squared is ~0, therefore, we cannot conclude that the changes in Democratic support from 2016 to 2020 can be explained by the changes in Other LatinX population.

Republican Support:

R-squared: 0.001

Conclusion: R-squared is ~0, therefore, we cannot conclude that the changes in Republican support from 2016 to 2020 can be explained by the changes in the Other LatinX population.

Graph 5: Presidential Support v. Puerto Rican Population Change (2016-2020)



Democratic Support:

R-squared: 0.0024

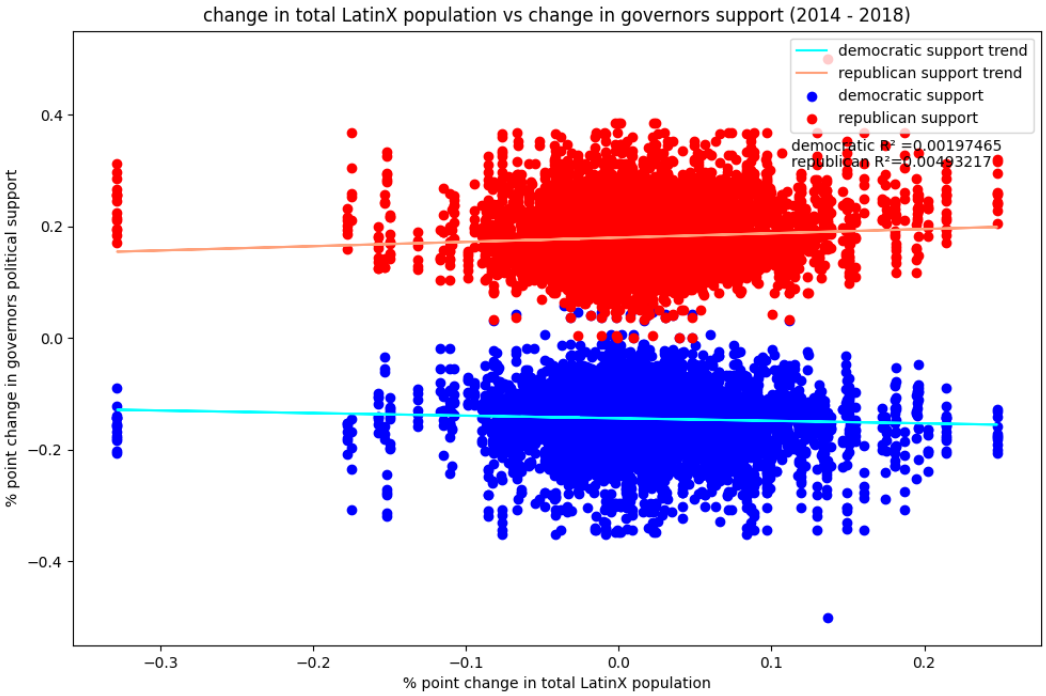
Conclusion: R-squared is ~0, therefore, we cannot conclude that the changes in Democratic support from 2016 to 2020 can be explained by the changes in the Puerto Rican population.

Republican Support:

R-squared: 0.0023

Conclusion: R-squared is ~0, therefore, we cannot conclude that the changes in Republican support from 2016 to 2020 can be explained by the changes in the Puerto Rican population.

LatinX Population vs. Political Support (Governors Election) Changes (2014-2018)



Democratic Support:

R-squared: 0.0020

Conclusion: R-squared is ~0, therefore, we cannot conclude that the changes in Democratic support from 2014 to 2018 can be explained by the changes in the LatinX population.

Republican Support:

R-squared: 0.0049

Conclusion: R-squared is ~0, therefore, we cannot conclude that the changes in Republican support from 2014 to 2018 can be explained by the changes in the LatinX population.

Answer Another Key Question:

Strategic Question 3:

Which LatinX groups exhibited changes in their votes and which groups remained the same?

Our team had a difficulty answering this question because we were not able to generate the necessary data to do so. In order to see how LatinX subgroup voting patterns changed, we needed data regarding how these subgroups had voted (i.e. who they voted for). Since this data was not available, we were limited in our ability to analyze changes in LatinX voting patterns and how they may have changed from 2016 to 2020.

The data that we analyzed included changes in the population of LatinX subgroups in Massachusetts as well as changes in overall election results. It would not be practical to infer changes in LatinX voting patterns purely based on changes in the LatinX demographic and election results. Correlation between changes in LatinX subgroups and changes in election results does not yield any insight into the actual voting patterns of those subgroups. Therefore, our team was limited in being able to analyze the change in the voting patterns of LatinX subgroups.

Potential Limitations and Risks:

A major limitation of this project is that our data does not go into detail about the demographic breakdown of votes, specifically within the LatinX population. The lack of connection between city election results and LatinX demographic data has made it difficult to directly answer the key questions for this project and infer a direct cause and effect relationship between shifts in demographic data and shifts in election results. Therefore, we had to resort to analyzing LatinX voting patterns by comparing changes in the LatinX population to changes in election results. This method is risky because we cannot assume that changes in the LatinX demographic have any relation to election results. There are a number of external factors that can cause shifts in election results that are not at all related to changes in the LatinX population. Any correlation between these two variables does not guarantee causation between them.

Refined Project Scope:

Our focus from the last deliverable was to encompass both the shifts in the overall LatinX population and shifts within each sub-group and compare them to the shifts in the overall election results for both presidential (2016 - 2020) and governor (2014 - 2018) elections. We

increased our sample size and generated data for all cities in Massachusetts in order to help us better infer a correlation between demographic changes and election results. We'll try to come to a more concrete conclusion for the data we found and the questions that our client wanted answers. We will also be focusing on writing a detailed report on the project and the algorithms used and refine the data visualizations.

Additional Visualizations:

Fig 1.

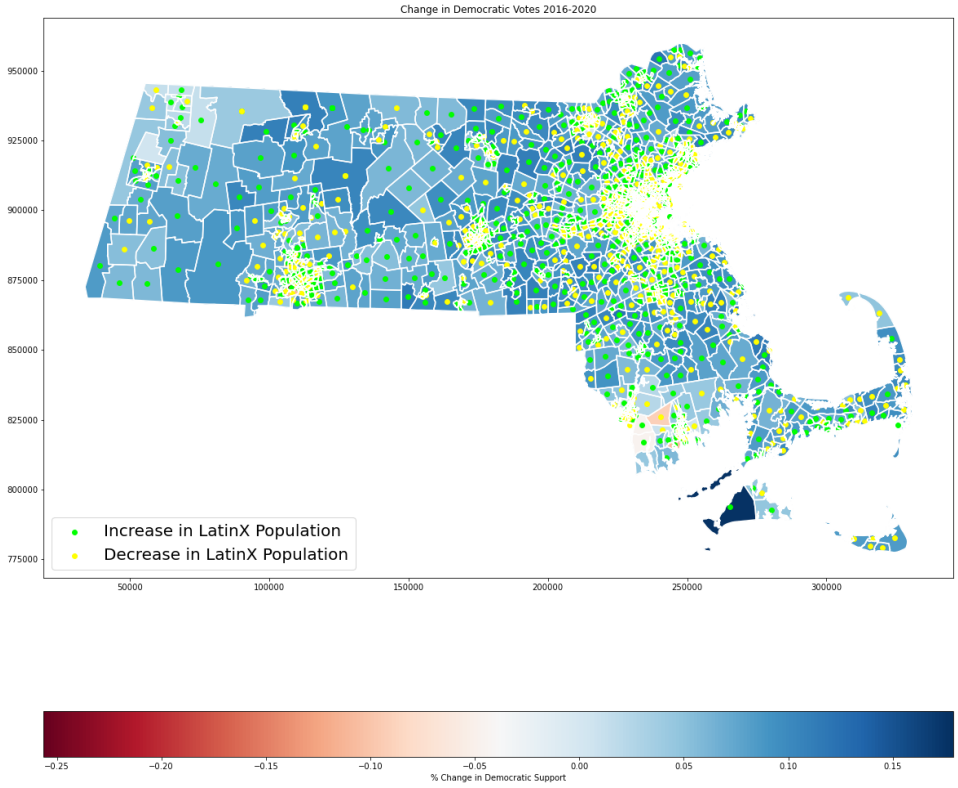


Fig 1 description: The visualization shows a heatmap of the changes in Democratic support for the 2016 to 2020 Massachusetts presidential election. Coloration of red indicates a decrease in democratic support. A coloration of blue indicates a positive change in Democratic support. Additionally, the dots depict changes in the LatinX populations within each tract. A yellow dot shows a decrease in LatinX population. A lime dot shows an increase in LatinX population.

Fig 2:

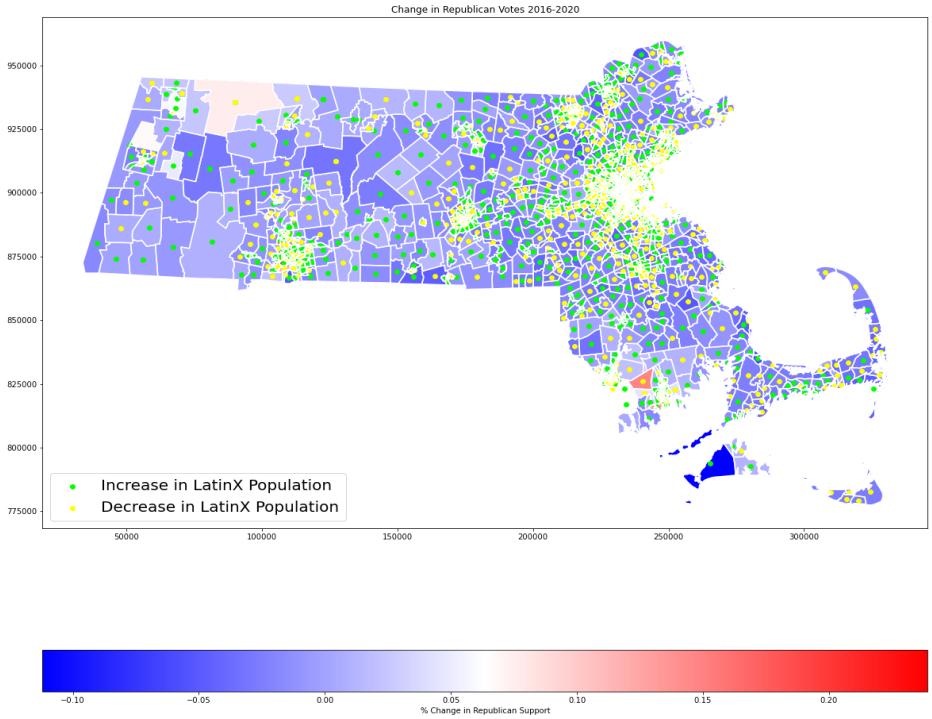


Fig 2 description: The visualization shows a heatmap of the changes in Republican support for the 2016 to 2020 Massachusetts presidential election. A coloration of red indicates a decrease in Republican support. A coloration of blue indicates a positive change in Republican support. Additionally, the dots depict changes in the LatinX populations within each tract. A yellow dot shows a decrease in LatinX population. A lime dot shows an increase in LatinX population.

Fig 3:

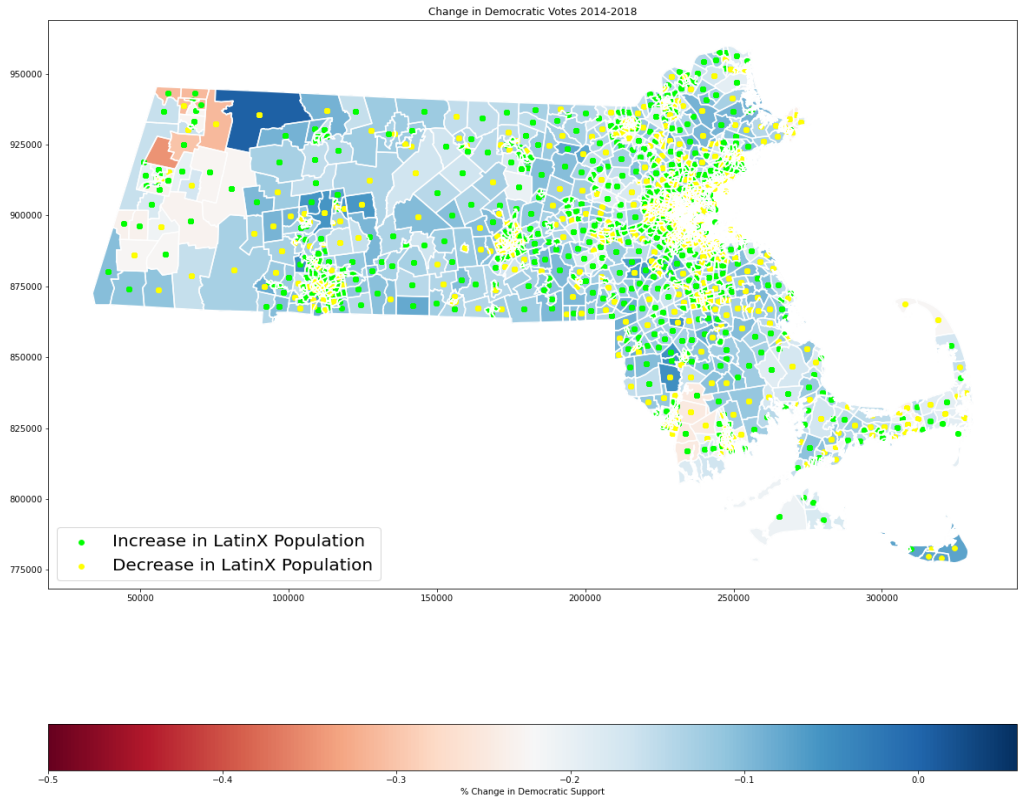


Fig 3 description: The visualization shows a heatmap of the changes in Democratic support for the 2014 to 2018 Massachusetts governor election. A coloration of red indicates a decrease in Republican support. A coloration of blue indicates a positive change in Democratic support. Additionally, the dots depict changes in the LatinX populations within each tract. A yellow dot shows a decrease in LatinX population. A lime dot shows an increase in LatinX population.

Fig 4:

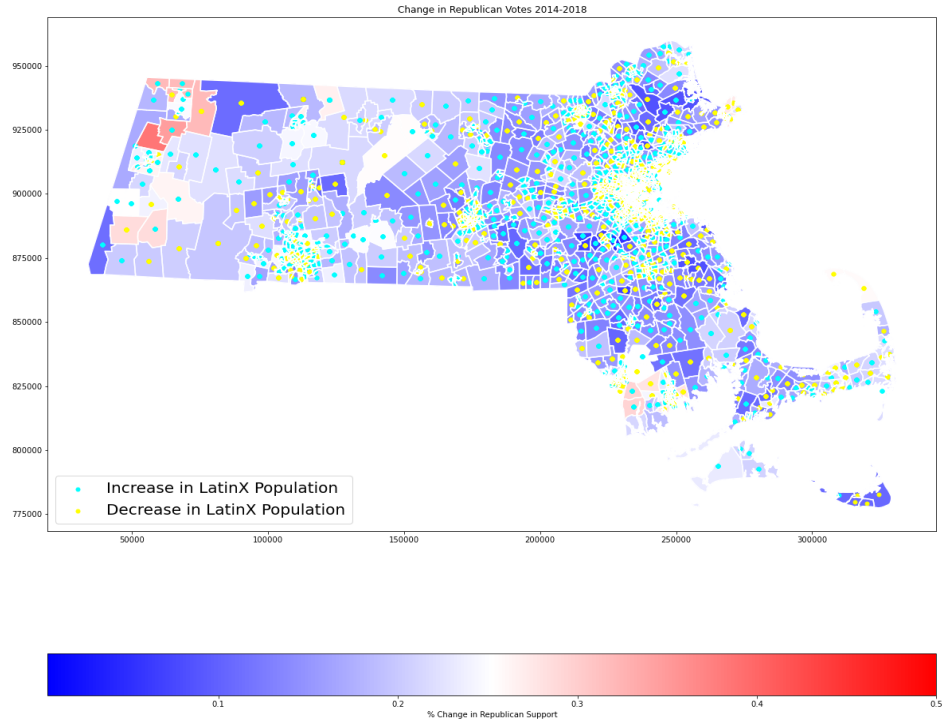


Fig 4 description: The visualization shows a heatmap of the changes in Republican support for the 2014 to 2018 Massachusetts governor election. A coloration of red indicates a decrease in Republican support. A coloration of blue indicates a positive change in Republican support. Additionally, the dots depict changes in the LatinX populations within each tract. A yellow dot shows a decrease in LatinX population. A lime dot shows an increase in LatinX population.

Draft of Final Report:

For the final report, we will mention the challenges we encountered when learning and using shapefiles and Geopandas. Additionally, we also ran into trouble with gathering the appropriate datasets for the project (The results of the presidential and governor elections based on LatinX sub-group) and how we could answer the key questions without those datasets. With these challenges, we also gained insight on how to pre-process data and create visualizations that were not learned from our courses. We received first-hand experiences on how data scientists conduct a project with their clients and how to handles challenges that may result in having to change the original goals for the project. We will also go into more detail about how we were able to get all the ward and precincts (both manually and by geoPandas), merge the data with the LatinX demographic dataset, and create the graphs and heatmaps for it by taking the

difference in percentage points of the LatinX subgroups and elections results. This includes the regression algorithms we used to create the trendline in the graphs.