Applied Data Science Capstone Project - Paycheck Protection Program loans

Introduction

This project will seek to determine if political persuasion influences utilisation of government assistance.

We will investigate the uptake of loans granted under the Small Business Administration Paycheck Protection Program ("SBA-PPP") in the United States.

The SBA-PPP was set up by the US Treasury to provide funding to businesses impacted by the COVID-19 pandemic.

The outcome of this analysis will be of interest to social scientists studying the impact of political persuasion on behaviours of businesses in the United States.

The idea for this project came after reading a number of articles pointing out inconsistencies in an organisations beliefs and actions in relation to government assistance. These two articles (In sign of the times, Ayn Rand Institute approved for PPP loan, and Vocal Opponents Of Federal Spending Took PPP Loans), highlight what the authors see as hypocritical behaviour.

These examples represent individual data points, and in all likelihood are called out for their ability to generate headlines.

A more thorough analysis will use granular data, that covering a broad population of firms, to answer the question posed above. Posing that question in a slightly different way. Are those who we would expect to be opposed to government intervention, less likely to accept government assistance?

Data

Determining if political persuasion influences uptake of loans requires us to do three things:

1. Quantify loan uptake
2. Assess political persuasion
3. Identify and hold other characteristics constant

The data sources that will allow us to do this are outlined below.

Loan uptake

Data on loans granted under the Small Business Administration Paycheck Protection Program ("SBA-PPP") has been made available at the U.S Treasury website. This data set contains records for each loan recipient along with attributes such as geographical location, industry membership and business type.

Political persuasion

We will assess political persuasion based on electoral results data. Our source for this data is the MIT Election Data and Science lab. This organisation has published numerous election results datasets to Github.

Confounding characteristics

The characteristics held constant will be a combination of demographic and industry attributes. Demographic data will be sourced from US census data via the uszipcode python library. This library aggregates geographic, demographic, employment and education data.

Foursquare data will be used to assess the type of region or neighbourhood. This will be done using the quantity and type of businesses in specific geographical regions.

Methodology

Our aim is to assess the impact of political persuasion or ideology on the use of government financial assistance. To this end we will narrow our focus to two states at opposing ends of the political spectrum.

This website contains visualisations of state level electoral results over varying time periods. Missouri and Illinois have voted Republican and Democrat respectively, on a consistent basis over the past 20 years. Missouri and Illinois are neighbours and therefore are expected to have similar industries and economies. This is important to our analysis as these are the type of characteristics we wish to control for.

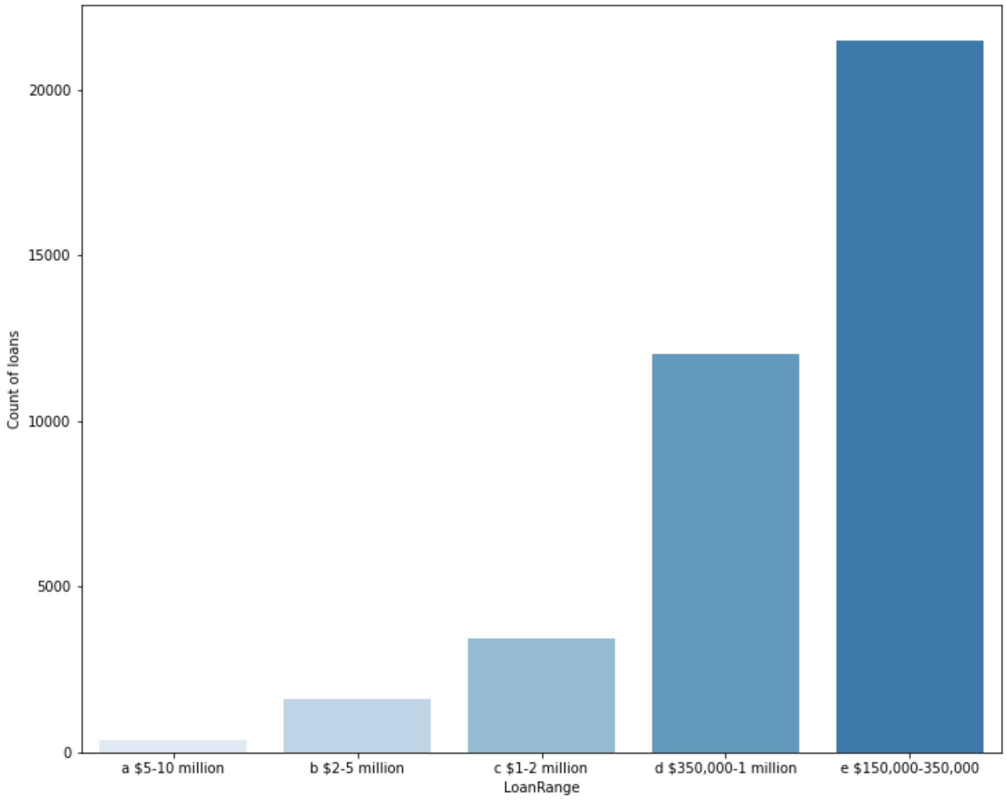
For Missouri and Illinois, we will gather characteristics at zipcode level, and group zipcodes into similar regions. These regions will be agnostic with respect to political persuasion and loan uptake.

For example, we may find zipcodes in both states that are low population density, high manufacturing industry density, medium income and high density of leisure venues. This would constitute one group. We can then assess the political persuasion of each zipcode in this group from electoral data, and finally determine if this results in differing level of loan uptake.

In summary, the rates of loan uptake will be assessed by comparing the count and value of loans issued by geographical regions known to be of differing political persuasion. In performing this assessment, confounding characteristics will be isolated and held constant.

We start by exploring the SBA-PPP loan data

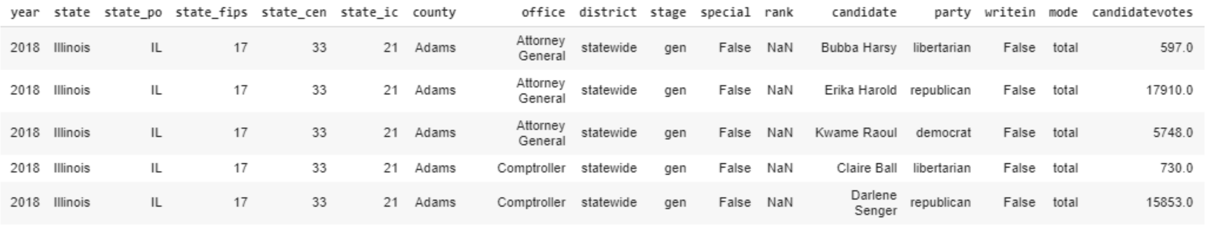
This data is held at individual loan level and contains zipcode along with a range of loan amounts. The table below shows the count of loans in each band.



Loans in the range of 150 to 350 thousand dollars are most frequent. The amount of loans increase inversely with there value. Nothing surprising there.

Moving to the MIT Election Data

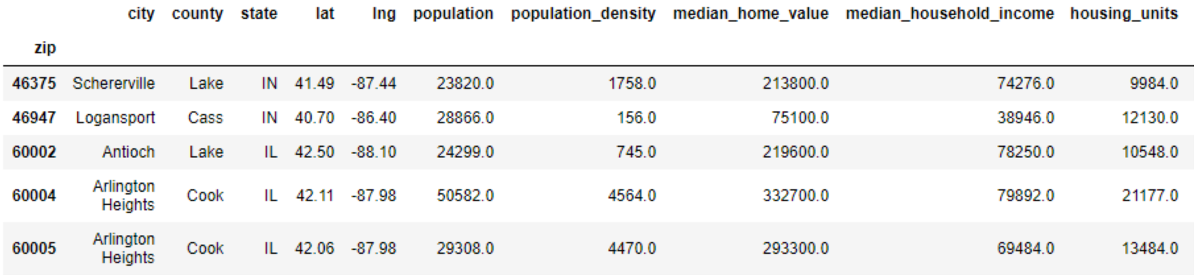
This data is held at county level. A sample of this data is shown below.



	year	state	state_po	state_fips	state_cen	state_ic	county	office	district	stage	special	rank	candidate	party	writein	mode	candidatevotes
	2018	Illinois	IL	17	33	21	Adams	Attorney General	statewide	gen	False	NaN	Bubba Harsy	libertarian	False	total	597.0
	2018	Illinois	IL	17	33	21	Adams	Attorney General	statewide	gen	False	NaN	Erika Harold	republican	False	total	17910.0
	2018	Illinois	IL	17	33	21	Adams	Attorney General	statewide	gen	False	NaN	Kwame Raoul	democrat	False	total	5748.0
	2018	Illinois	IL	17	33	21	Adams	Comptroller	statewide	gen	False	NaN	Claire Ball	libertarian	False	total	730.0
	2018	Illinois	IL	17	33	21	Adams	Comptroller	statewide	gen	False	NaN	Darlene Senger	republican	False	total	15853.0

The demographic data

This data is also held at county level.



	city	county	state	lat	lng	population	population_density	median_home_value	median_household_income	housing_units
zip										
46375	Schererville	Lake	IN	41.49	-87.44	23820.0	1758.0	213800.0	74276.0	9984.0
46947	Logansport	Cass	IN	40.70	-86.40	28866.0	156.0	75100.0	38946.0	12130.0
60002	Antioch	Lake	IL	42.50	-88.10	24299.0	745.0	219600.0	78250.0	10548.0
60004	Arlington Heights	Cook	IL	42.11	-87.98	50582.0	4564.0	332700.0	79892.0	21177.0
60005	Arlington Heights	Cook	IL	42.06	-87.98	29308.0	4470.0	293300.0	69484.0	13484.0

Re-shaping

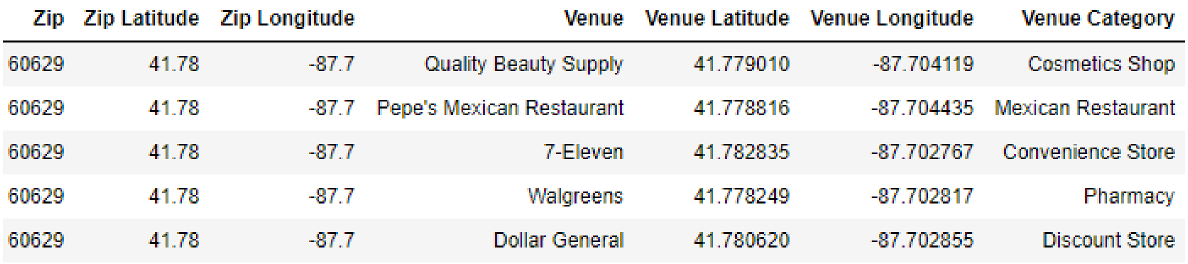
The three raw data sets are joined and re-shaped:

1. Creating a binary indicator for political persuasion
2. Reming columns not required
3. Remove rows with NaN's

After this process, the resultant data is sampled, selecting the top 20 zipcodes by state and political party by population. This sampling is required in order to reduce the requests to the Foursquare api.

Foursquare data

Foursquare data is retrieved using the api inteface. The data looks as follows.



Zip	Zip Latitude	Zip Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
60629	41.78	-87.7	Quality Beauty Supply	41.779010	-87.704119	Cosmetics Shop
60629	41.78	-87.7	Pepe's Mexican Restaurant	41.778816	-87.704435	Mexican Restaurant
60629	41.78	-87.7	7-Eleven	41.782835	-87.702767	Convenience Store
60629	41.78	-87.7	Walgreens	41.778249	-87.702817	Pharmacy
60629	41.78	-87.7	Dollar General	41.780620	-87.702855	Discount Store

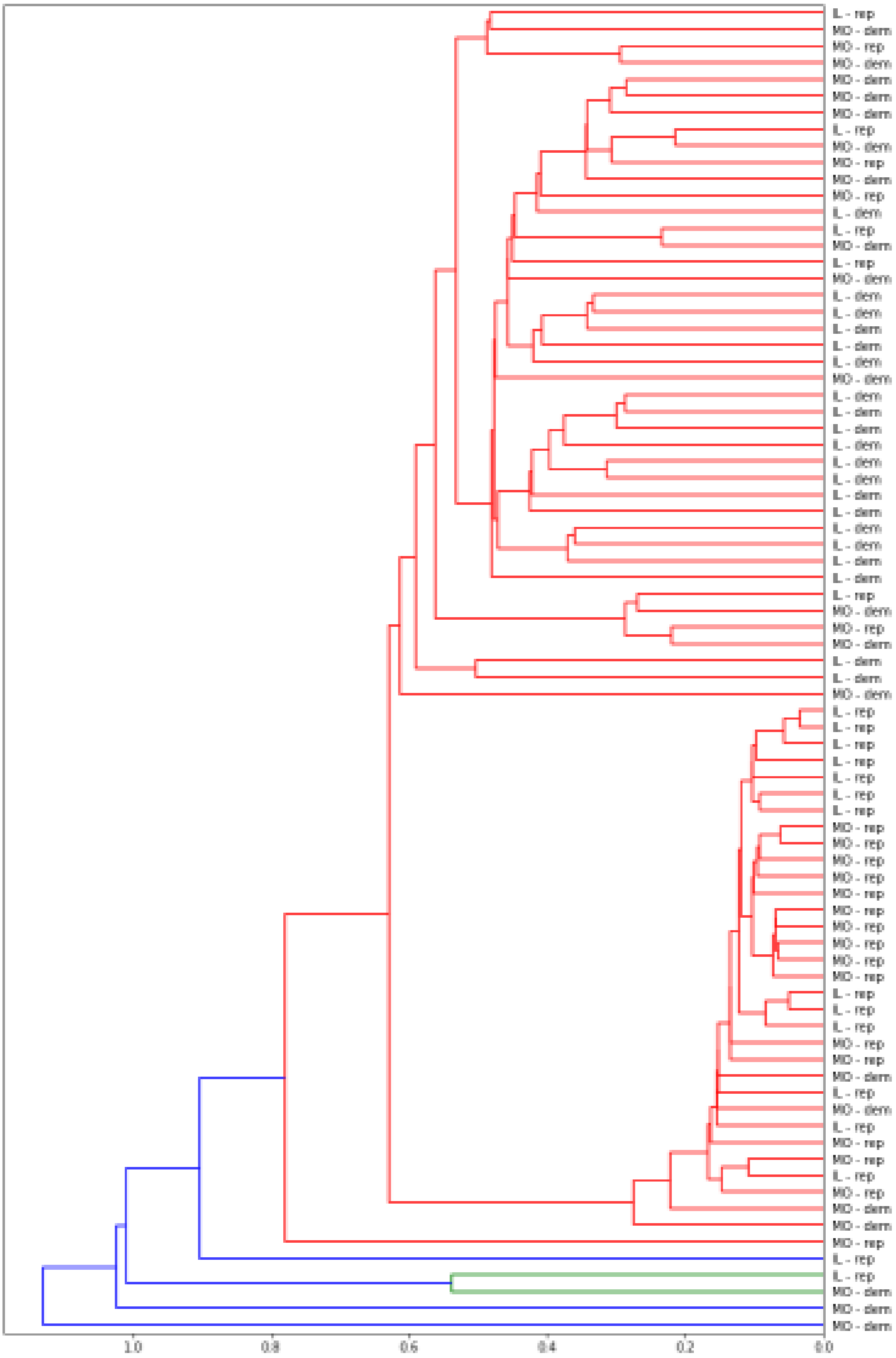
This is joined to the created loan, demographic and electoral data referred to above.

It should be noted that venue categories have been imputed for the Foursquare data where the api has not retrieved information for the zipcode requested. This has been performed filling missing data with the most frequent category, the mode, by zipcode.

Hierarchical clustering

As stated in the methodology section *“we will gather characteristics at zipcode level, and group zipcodes into similar regions. These regions will be agnostic with respect to political persuasion and loan uptake”*

We will apply a hierarchical clustering algorithm to accomplish this. A dendrogram representation of this is shown below.



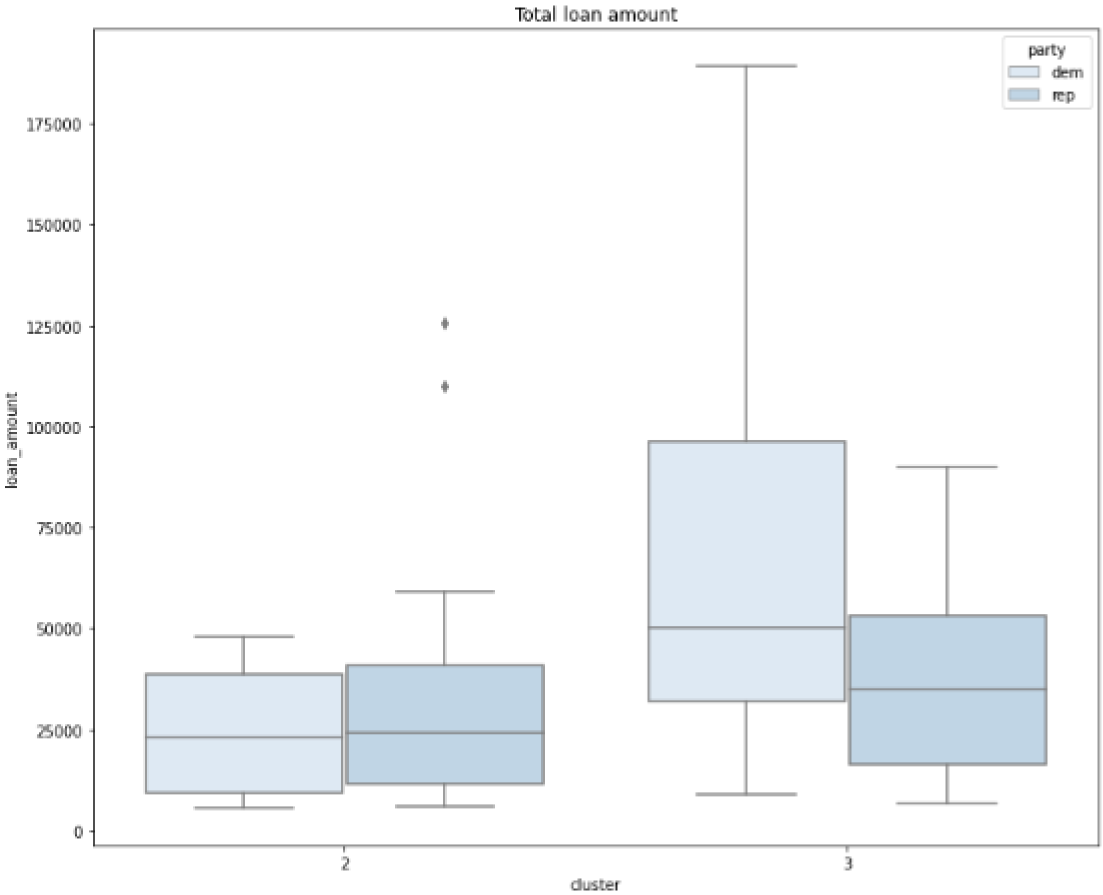
We slice the dendrogram to return 8 clusters. We then assign this cluster membership information back to zipcodes and the related loan and political persuasion information.

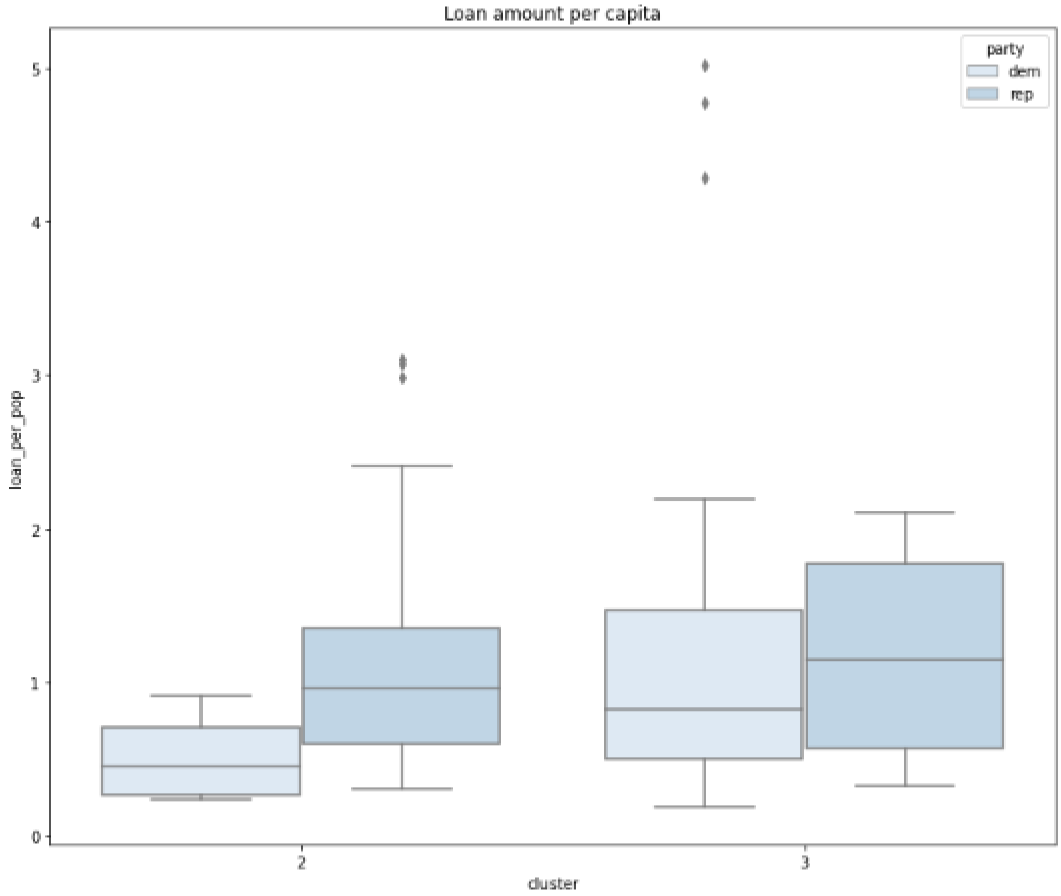
It should be noted that each of the clusters identified by the hierarchical clustering algorithm represent a group of similar zipcodes for which we will assess differing loan uptake by political persuasion. Each of the clusters are effectively controls, meant to take account of the confounding characteristics, i.e. to hold them held constant.

It turns out two of the eight clusters represent 73 of the 80 sampled zipcodes. The results section will therefore focus on these two clusters.

Results

We will now inspect the distribution of loan values for the two clusters returning the vast majority of zipcodes.





The box plots above represent the distribution of loan values (top) and value of loans per capita (bottom) for all zipcodes in each of the categories labelled 2 and 3. The categories are clusters determined with reference to demographic and neighbourhood characteristics.

These plots do not suggest a significant difference in loan uptake between Democrat (light) and Republican (dark) zipcodes.

Discussion

A number of unrelated points on methodology and data (with the benefit of hindsight).

Ideally we should formally test for differences in means of the data presented in the box plots above. The results of an analysis of variance test would be informative in this regard. Unfortunately deadlines loom (this report is being prepared for a Data Science course), and as a result of that we will rely only on the visual inspection referred to above.

An interesting point to consider in light of the methodology used above, is the fact that electoral results may be sufficiently close at zipcode to mask any underlying variation in the loan data.

Reconsidering the question posed may have us regressing the count of loans over the count of all business establishments, on the percent of democrat or republican vote by zipcode. This is beyond the scope of this analysis due to the lack of total business establishments data.

Conclusion

This report set out to determine if political persuasion influences utilisation of government assistance, namely the uptake of loans issued under the Small Business Administration Paycheck Protection Program in the US.

Ultimately the results were inconclusive.

A number of things could have impacted the results. Changing any of the items listed below may result in different and more robust results:

1. Greater coverage of Foursquare data
2. A larger coverage of zipcodes
3. Using electoral data at a more granular level
4. Having access to total business establishment data

These items should be considered for further analysis.