REU-DEIM Statistical analysis of Hispanic voters' practices in Florida, Progress Report

Kamila Soto-Ortiz, Dr. Mihhail Berezovski
Embry-Riddle Aeronautical University

I.INTRODUCTION

The Hispanic population is considered to be those that identify with a "Spanish-speaking background and trace their origin or descent from… Spanish-speaking countries" [9]. They are composed of 21 countries and 489 million native speakers, second only to Mandarin [7]. Hispanics are currently about a quarter of Florida's population [12] and they are expected increase in population to 33% by 2045 [13].

Understanding how Hispanics vote then becomes a very crucial task for future elections. For starters, certain social factors such as "religion, region, and social class appear to be the characteristics that have most closely related to voting" [10]. Additionally, past elections have hinted that Hispanic groups do not vote the same [1][2][8], so if understanding their voting practices is important, knowing the group they identify with is paramount. The knowledge of an individual's Hispanic subgroup paired with their voting preferences would trace voting trends amongst the groups, if any, and potentially capture if certain groups have better voting turnout.

An individual's vote is private, but their voter file is public information. One of the elements in the voter file is the party affiliation, which would hint at how a person votes. Yet, the voter file only classifies Hispanic voters under the broad term "Hispanic", without further identification of the group they identify with. Due to this, a classification process must be developed to sort each self-identified Hispanic as belonging to a particular group, then look for party affiliation trends amongst those groups.

From the voter file, the voter's full name is extracted on the basis of Hispanic naming tradition and family names, which are matched to public Census data that provides the most popular and prevalent names on the group's country of origin. With this information, the voter's most probable Hispanic subgroup is calculated. Additionally, the individual's zip code (extracted from the voter file) is also matched to acquired public Census data that details the percentages of population of each group at the given zip code. This is done to ensure the number of registered voters does not exceed the group's population in that location. Once all the voters are classified, the process of looking for trends with the voter's party affiliation and voter activity can begin.

As for additional social factors, the first to be explored is the population density of a zip code [11], income salary and generation of Hispanic voters. Those along with the classified subgroup can be examined with party affiliation and voter activity for a better understanding on their voter practices.

II.DATA: VOTER FILE AND CENSUS DATA

The data is divided in the given and the acquired data. The given data is the Florida voter file [14], which is used to compare with Census data [5][6][15]. This is necessary because the voter file only includes "Hispanic" with no further detailing of the subgroups. In this manner, the Census data can be used for further analysis and as a constraint.
Furthermore, some of the other factors to explore also require additional census data, such as the population density, gross income and voter activity.

A. Voter File

From the Florida voter file [13], the following are extracted: County Code, Voter ID, Name Last, Name First, Name Middle, Residence Zip Code, Gender, Race, Party Affiliation and Date of Birth. There are three different motives for the use of these particular data points. County Code, Voter ID and Race are used for identification purposes; Gender, Party Affiliation and Date of Birth are used for demographic purposes; and Name Last, Name First, Name Middle and Residence Zip Code are used for data analysis.

B. Zip Code Census Data

The first of the acquired data, it describes the different percentages of subgroups per zip code [6]. The zip code data is very valuable because, not only does it indicate the percentages of each group from the total Hispanic population, but it also serves as a constraint. Naturally, if the population of a subgroup is such in each zip code, the number of registered voters should not exceed that value.

The voter's zip code is matched with the zip code census data to aid in the process of subgroup classification. The Hispanic groups in the dataset are Mexicans, Puerto Rican, Cubans, Dominican Costa Rican, Guatemalan, Honduran, Nicaraguan, Panamanian, Salvadoran, Argentinean, Bolivian, Chilean, Colombian, Ecuadorian, Paraguayan, Peruvian, Uruguayan and Venezuelan. This groups are the ones each voter is classified in.

However valuable, because the zip code Census data does not provide any information on which voter belongs to what group, additional data is required for the classification process.

C. Name Census Data

The name data set is divided between forename and surname [5]. It provides with a name's popularity in the country of origin. This was searched to match the Hispanic groups in the zip code Census data. Since surnames are passed down by families, and Hispanics have traditional forename practices, data on these names can give the remaining information we need for voter analysis.

D. Population Density

On the other hand, the population density data provides zip code, population and population density. The zip code from this data set can be matched with the zip code of each voter to determine the voter's population density.

E. Income Salary

The data used for the income salary provided an extensive amount of information regarding tax returns, all broken down by zip code. The adjusted gross income presents how many individuals fall within a range of incomes. The ranges are "$1 under $25,000", "$25,000 under $50,000", "$50,000 under $75,000", "$75,000 under $100,000", "$100,000 under $200,000" and "$200,000 or more".

F. Voter Activity

The voter activity data contained the voter ID along with the type of election and the manner of the voting process. Voters would naturally appear more than once, for which a single voter ID would be extracted along with how many times they had voted.

III.MODEL

Given the wide range of topics being explored, the model is a work in progress. It is divided in the subgroup classification and the exploration of the social factors and their impacts.

A. Subgroup Classification

Once the given data is extracted, the acquired Census data is used to calculate the most probable subgroup is calculated.

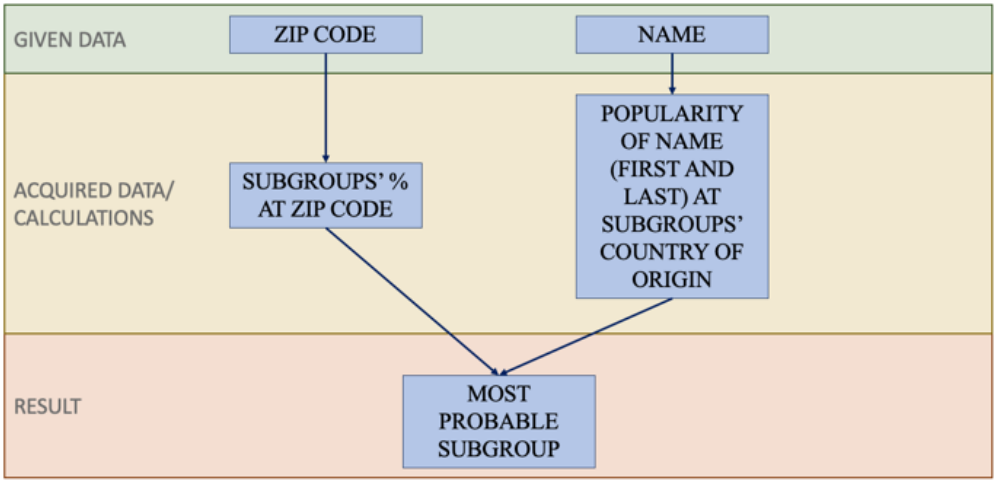

Figure 1 Visualization of Model. The extracted data from the voter file [12] is matched to the acquired census data [4][5]. The calculations are preformed using Equation 1, with the resulting probabilities being the result.

1. Forenames and Surnames

For each subgroup, the first thousand most popular names are taken along with their popularity at the country of origin. The names of the voter are matched with the Census data and the information is extracted for analysis.

The calculations performed on the data to compute the most probable subgroup follow the probability of independent events theory, which proposes that one event occurring does not affect the other taking place as well. In this case, the probability of both occurring is the multiplication of their probabilities. In the case of a voter's names, a person's forename being more popular in a subgroup's country does not affect their surname being more or less popular; the events are independent of each other. The probability of both the forename and the surname belonging to a particular subgroup is their multiplication, hence the ratio of surnames and forenames are multiplied together in a controlled sum.

The controlled sum is performed because of the limitation in data. Due to the name Census data having only the first thousand most popular names in a country, voters' names may not appear in the data. A controlled sum would reveal this information, for which if there is no data for a particular voter in any of

the subgroups, the voter is defaulted to the biggest group in the zip code, with a low confidence in their most probable subgroup.

$$IF(LN1="Y",SG_{LN1},1))*(IF(FN="Y",SG_{FN},1))*(IF(AND(LN1="N",FN="N"),(IF(MN="Y",SG_{MN},1)),1))*(IF(AND(LN1="N",LN2="Y"),SG_{LN2},1)) \quad (1)$$

where LN1=Last Name 1, FN=First Name, MN=Middle Name, LN2=Last Name 2 and "Y" or "N" signifies whether the controlled sum presented any data or not; SG_{LN1}=Last Name 1 data on a particular subgroup, SG_{FN}=First Name data on a particular subgroup, SG_{MN}=Middle Name data on a particular subgroup and SG_{LN2}= Last Name 2 data on a particular subgroup.

The formula first verifies if there is data for the first and last name and multiplies those together. If there is no data for either, the middle name is used instead. Finally, if the last name has no data, then the second last name is inspected for use.

a. Name Data Analysis, False Positives and False Negatives

Because the name data set has only the 1,000 most popular names some fewer common names may not appear. In this case, voters identifying as Hispanic won't have any data that can classify them; these are referred to as false negatives. On the opposite line, some of the world's most popular names with no Hispanic origin may make themselves through the data, cause non-Hispanic voters to have data available without being Hispanic; these are referred to as false positives.

While checking with the voter file if the voter identifies as Hispanic can rule out false positives, false negatives currently have no other method to aid in their classification. These are limitations withing the process.

2. Zip Code

While the zip code data used is from 2010 and not the most updated numbers (Hispanics make only 22% of the Florida population), the trend of the groups remains, which can be appreciated in Figure 2.

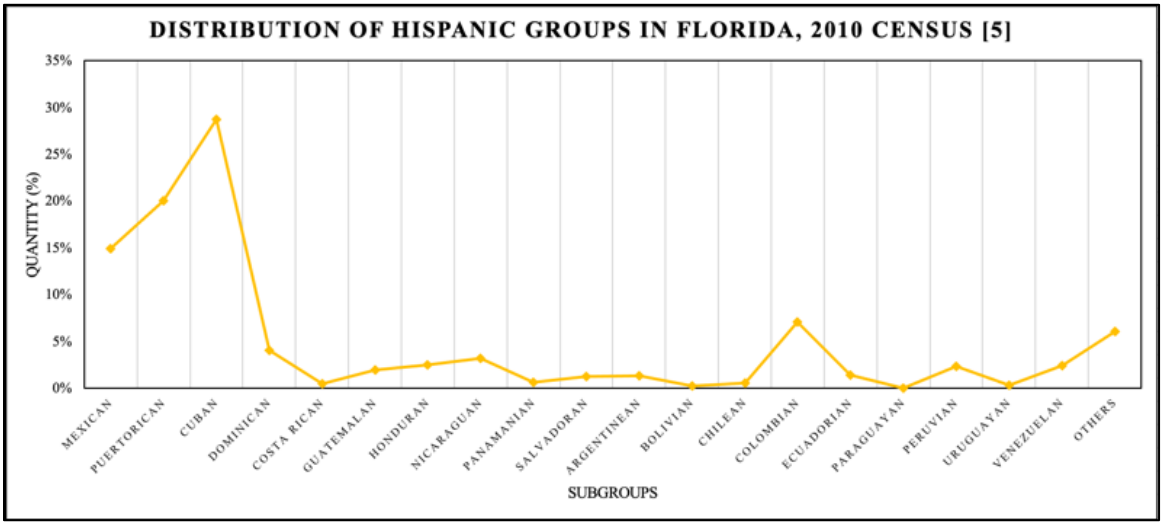

Figure 2 Hispanic groups distribution in Florida from the 2010 Census. While the data is from 2010 and not the most updated numbers, the trend of the groups remains.

For each zip code, the percentages of each subgroup at that location are taken. These percentages are multiplied by Equation 1 to ensure the groups maintain themselves constrained to their population at that zip code.

Equation 1 is then modified to the following:

$$\text{IF(LN1="Y",SG}_{LN1}\text{,1))*(IF(FN="Y",SG}_{FN}\text{,1))*(IF(AND(LN1="N",FN="N"),(IF(MN="Y",SG}_{MN}\text{,1)),1))*}$$
$$\text{(IF(AND(LN1="N",LN2="Y"),SG}_{LN2}\text{,1))* SG}_{ZC} \quad (2)$$

where SG_{ZC}=zip code data on a particular subgroup.

a. Zip Code Data Analysis

The process of Equation 2 may not be necessary for all zip codes though, as some of them have a group dominating the location, thus allowing for the assumption that all voters in that location belong to that group.

B. Social Factors

To fully comprehend the behavior of Hispanic voters in Florida, other factors that influence voting practices may be of use [10]. The ones included in the research are population density, age, income and Hispanic group paired with their activity and party affiliation.

1. Hispanic Group

The Hispanic Group is calculated from the subgroup classification process. In it the full name and zip code from the voter are taken for the calculation using Equation 2.

2. Generations

The date of birth is one of the elements included in the Florida voter file. Each voter is classified in their corresponding generation based on the year they were born in [3].

3. Population Density

The population density of each voter is matched with their respective zip code.

4. Income Salary

The adjusted gross income is extracted, along with the ranges of income and how many individuals fall in each range. These are paired with the voter's zip code and placed on the range with the most people.

5. Party Affiliation and Voter Activity

The two constant variables of analysis throughout the social factors, are paired together to determine which party is the most active.

IV.TEST COUNTY: PALM BEACH

A. Demographics

Testing the method with the entirety of Florida can be quite overwhelming, which is why testing began with a single county, Palm Beach. Out of all Hispanic voters in this county, 31% are considered passive voters having registered but never voted, while 69% are considered active voters. Palm Beach County has 118,926 Active Hispanic Voters (AHV) which is 12.06% of the total Active Voters. It comprises 52 different zip codes, in addition to another shared with Hendry County; approximately 67% of AHV are concentrated in coastal areas.

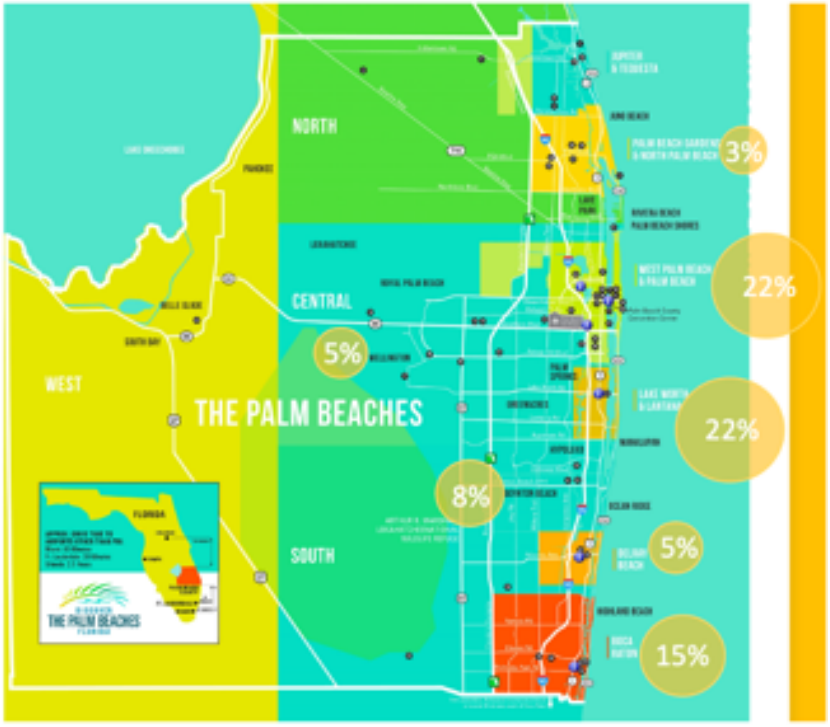

Figure 3 Hispanic population by location; high concentration in coastal areas.

There is a 1.26 female to male ratio. Of the AHV, 42.48% are affiliated with the Florida Democratic Party, 37.33% have No Party Affiliation, 19.13% are affiliated with the Republican Party of Florida and 1.06% are affiliated with other parties.

B. Results

The results are composed of two connected analysis processes: the subgroup classification and the social factors analysis, which uses data from the subgroup classification process. Because the Hispanic group classification is needed for the social factors, that will be discussed first.

1. Subgroup Classification

The subgroup classification process requires two main components to classify a voter on their most probable subgroup. These are their full name and their zip code, which are combined using Equation 2.

An example for a single voter from zip code 33444, Marta Lilliam García shows all datasets that influence her resulting group.

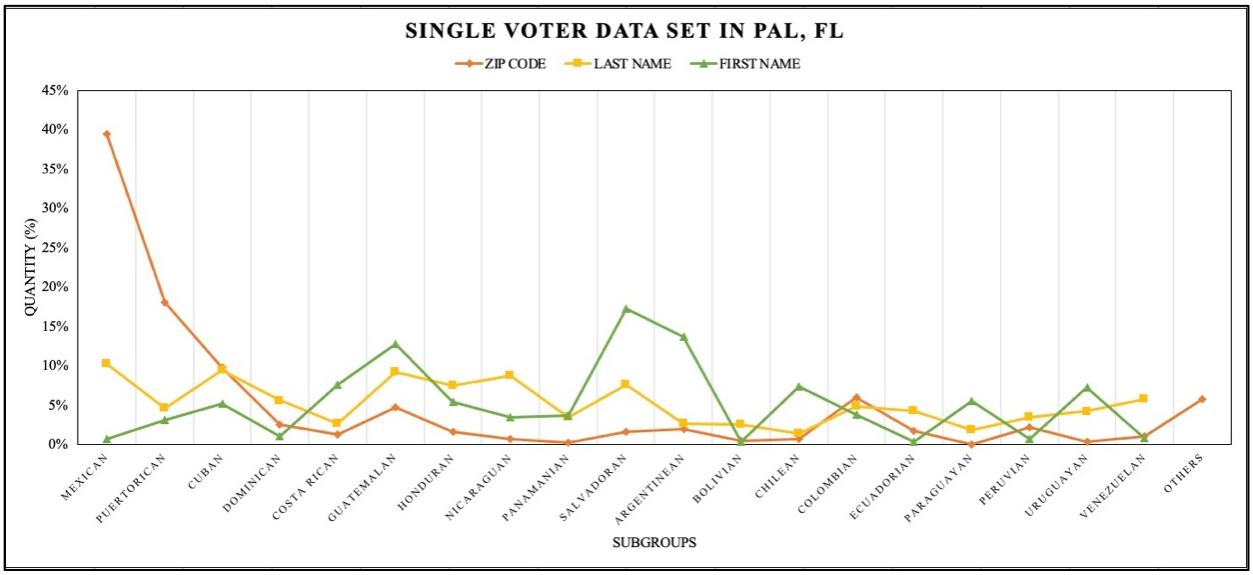

Figure 4 Different categories such as zip code, last name and first name for Marta García.

The biggest peaks for Marta are Cuban, Guatemalan, Salvadoran and Colombian. Because there are many Cubans in the zip code, this will most likely bring up her likelihood although the peaks for first and last name are not as high compared to others. There are enough Guatemalans in the zip code that her last name and first name peaks will likely be the most noticed. Although Salvadoran has the highest peaks, the zip code has too few Salvadorans. Finally, Colombian has somewhat noticeable peaks in first and last name along with enough Colombians.

All probabilities and ratios are brought together with Equation 2.

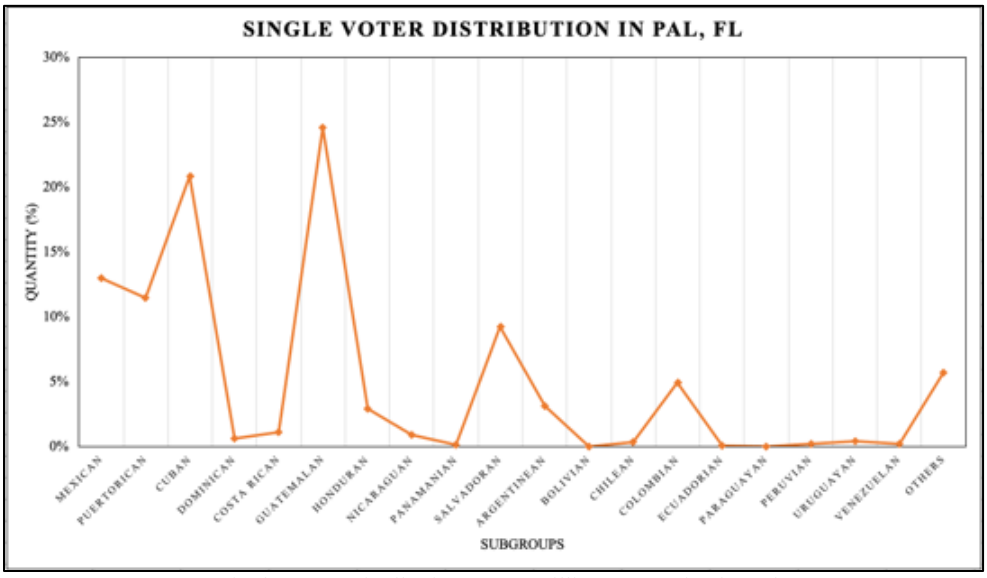

Figure 5 Single Voter Distribution, Marta Lilliam García in zip code 33444.

Her Most Probable Subgroup (the highest peak) is Guatemalan. Because Marta's most probable subgroup has a low confidence in its calculation (25%), Marta is not considered in the sample for the county. The filtering of low-confidence voters leaves the sample with 18,298 AHV, 15% of the original AHV in Palm Beach. This is a limitation of the model discussed in better detail in its corresponding section.

After doing the same calculation for all voters that fall in the sample, the distribution for the county is plotted.

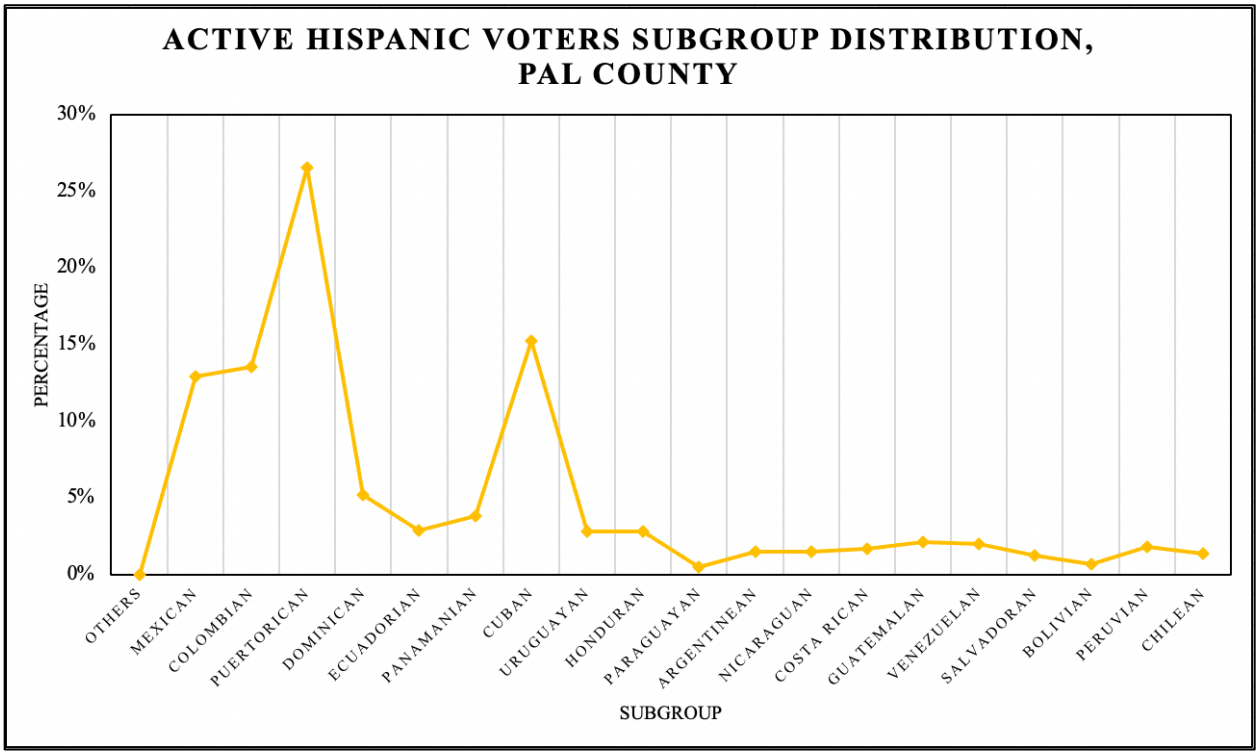

Figure 6 General Distribution of AHV in zip code Palm Beach County.

The largest group of high-confidence AHV in Palm Beach are Puerto Ricans at 27%, followed by Cubans with 15% and Colombians with 14%.

Following the model for the subgroup classification process, 97% out of all AHV in Palm Beach have name data available for the classification.

2. Social Factors

The results for the social factors that influence voting compares the Hispanic groups, ages, population density and income of our voters with their voter activity and their party affiliation.

a. Hispanic Groups

Once the group the voter identifies with is calculated, their voter activity is observed alongside. Out of one group, how many voters are passive versus active? Out of the active voters, how many times each voter has voted? Finally, are there any groups more affiliated with a particular party?

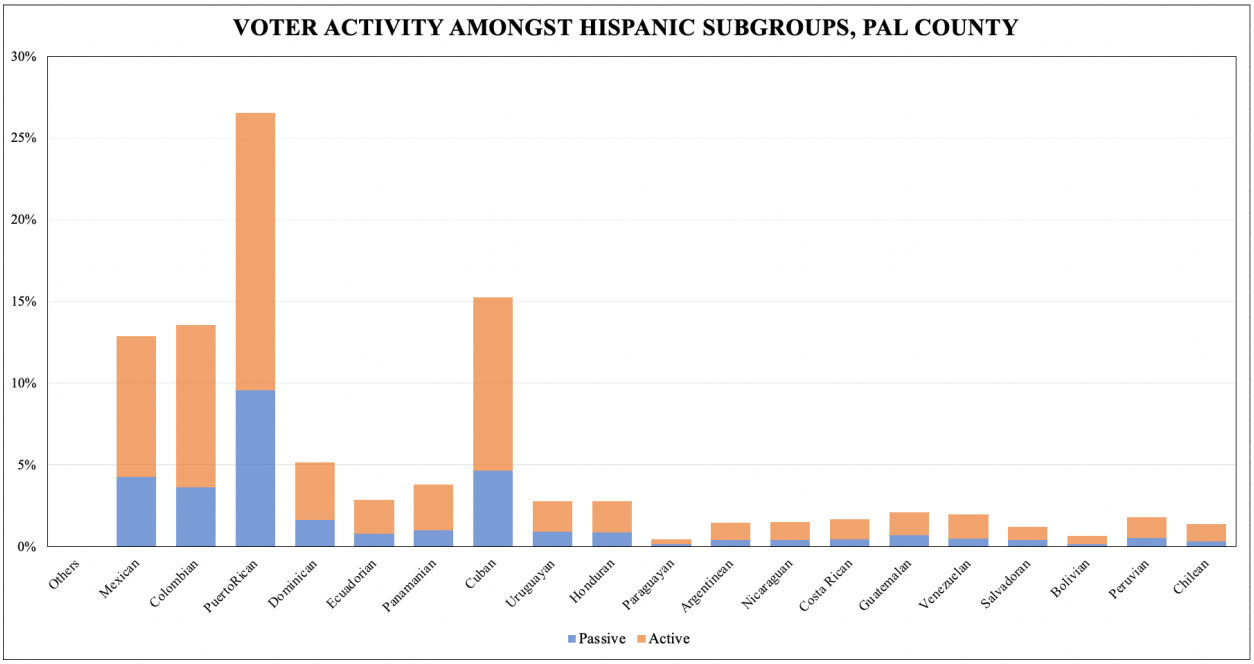

Figure 9 Voter activity amongst the sample of voters' Hispanic groups in Palm Beach.

The most active groups of voters overall are Puerto Ricans, Cubans, Colombians and Mexicans, respectively. The most active groups within their own group population are Bolivians and Chileans. The most active groups within the largest groups of voters are Cubans and Colombians respectively.

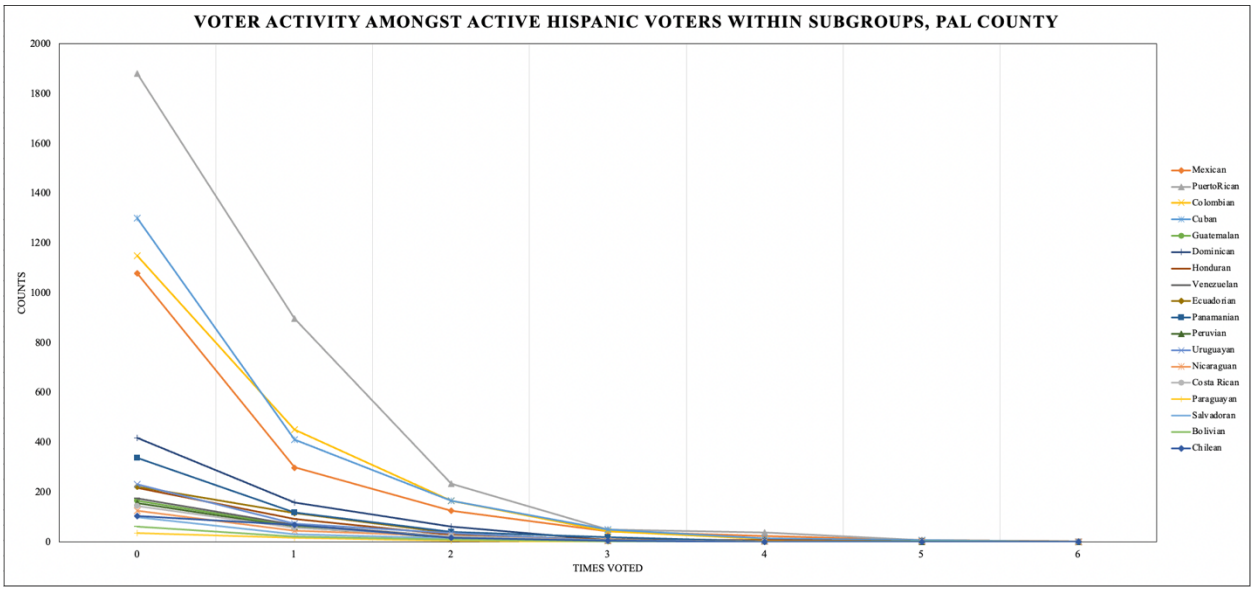

Figure 10 Voter activity amongst the active voters in Hispanic groups at Palm Beach.

After the data has been normalized across the different groups, the count of voters that have voted a number of times can be appreciated in Figure 10. Because of the normalization process, the x-axis values represent a range of numbers that fit each group. Hispanic groups such as Puerto Ricans, Cubans,

Colombians and Mexicans consistently have the highest count of voters throughout the graph due to their high populations yet flatten out with the rest of the groups at high values of the times voted.

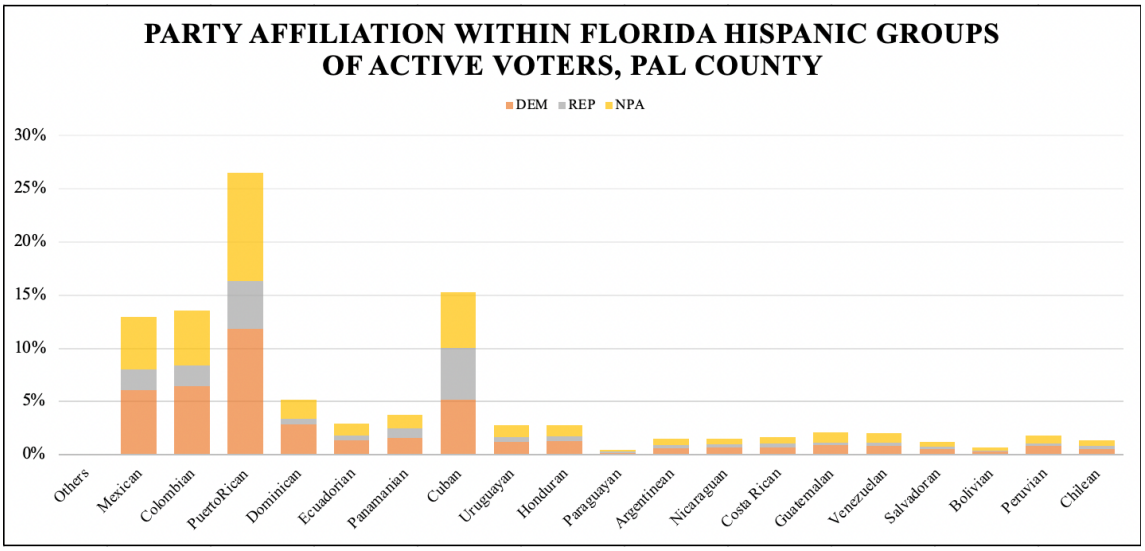

Figure 11 Party affiliation amongst the active voters in Hispanic groups at Palm Beach.

In Figure 11, a consistent trend can be observed amongst most groups: a high percentage of Democrat voters, a near equal amount of Non-Party affiliated voters and a small percentage of Republican voters, all relative to their population. However, certain groups deviate from the trend. Dominicans, for example, have the lowest percentage of Republicans and the highest percentage of Democrats within their population. Cubans have near equal percentages of Democrats, Republicans and Non-Party affiliated. The also make a quarter of Republican Hispanics.

b. Generations

Each generation is then compared with each other for their voter activity and party affiliation preferences.

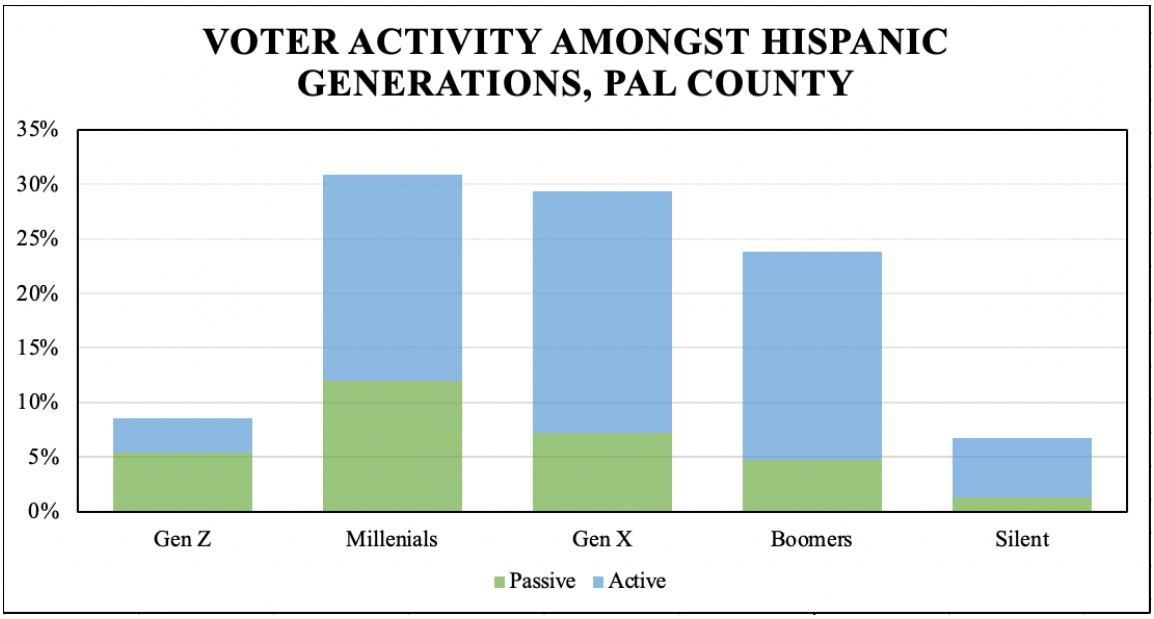

Figure 12 Voter activity amongst Hispanic generations in Palm Beach.

The most active generations of voters overall are Gen X, Boomers and Millennials, respectively. It is no surprise, given Gen Z are an emerging group and the Silent generation is an exiting group. The most active groups within their own population are the Silent generation and Boomers, respectively. Unsurprisingly, since these groups have been around for longer.

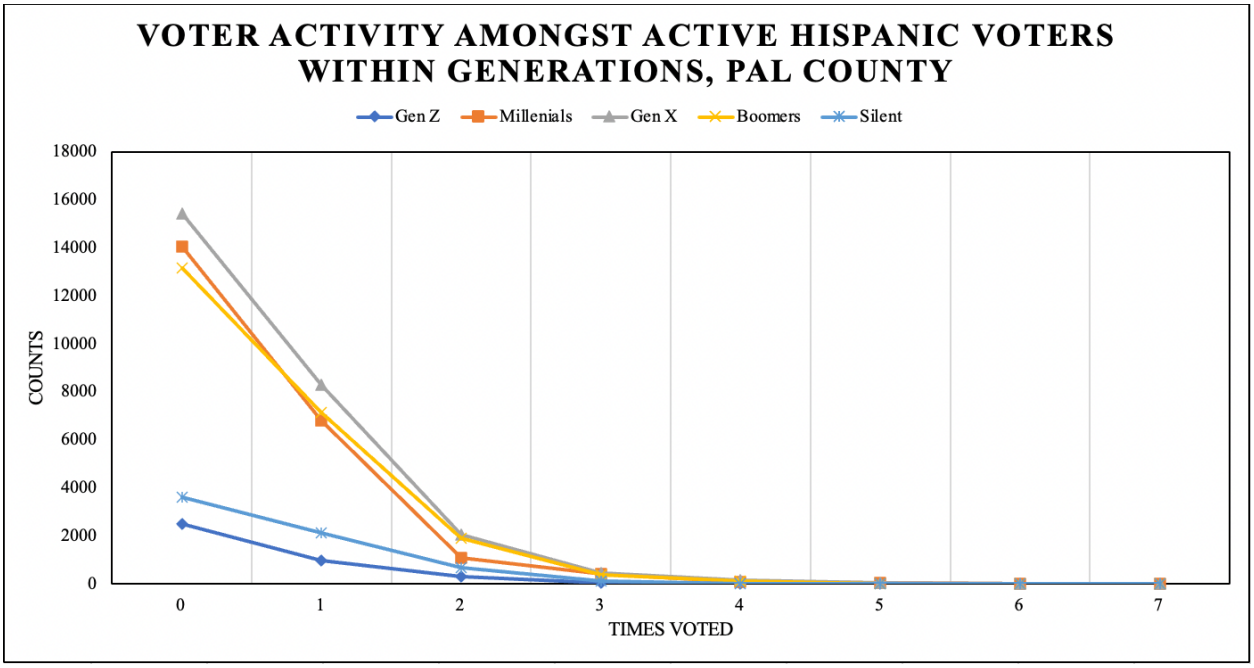

Figure 13 Voter activity amongst the active voters in Hispanic generations at Palm Beach.

Once normalized, Hispanic generations such as Gen X, Boomers and Millennials consistently have the highest count of voters throughout the graph due to their prime voting age yet flatten out with the rest of the generations at high values of the times voted. This was also noted in Figure 10. Before flattening, though, Boomers clearly gain advantage over both Gen X and Millennials.

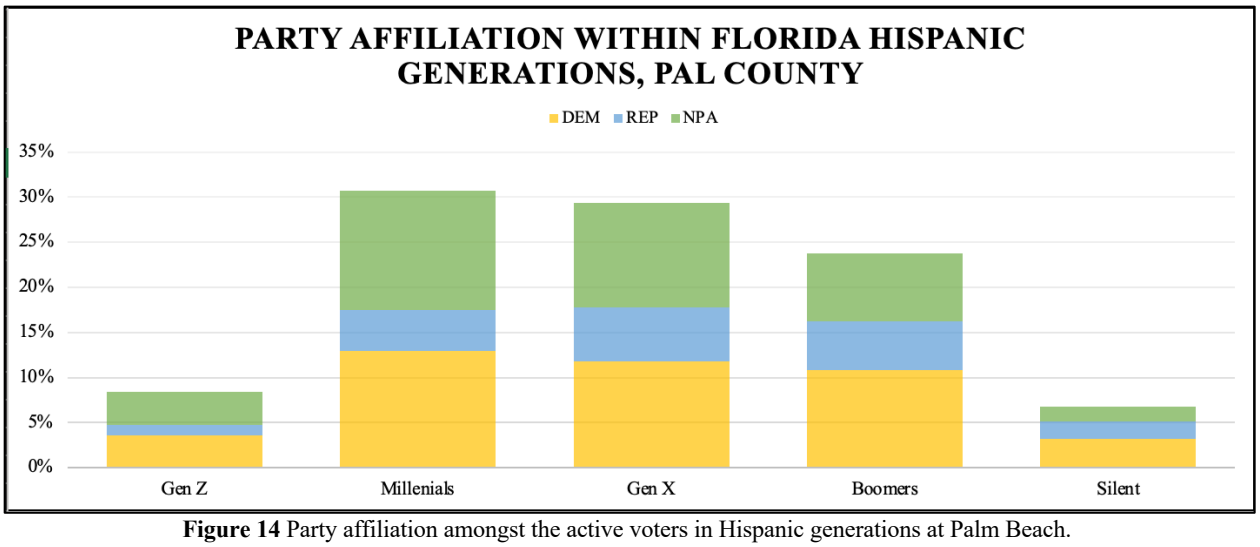

Figure 14 Party affiliation amongst the active voters in Hispanic generations at Palm Beach.

In Figure 14, a consistent trend can be observed amongst most generations: a high percentage of Democrat voters, a near equal amount of Non-Party affiliated voters and a small percentage of Republican

voters, all relative to their population. However, certain generations deviate from the trend. Gen X made near a third of all Republican voters. Millennials, on the other hand, also made near a third of Democrat voters. In the Silent generation, almost half of their voters were Democrats. Although, the Silent generation also had the highest percentage of Republicans within their own generation. Naturally they had a low percentage of Non-Party affiliated voters, which saw a steady increase with more recent generations.

c. Population Density

As previously mentioned, the population density is matched with each voter's zip code to determine their population density, and thus if they live in an urban or rural area. Out of the 69% AHV out of all Hispanic voters, how many people voted within a range of times regarding their population density is plotted.

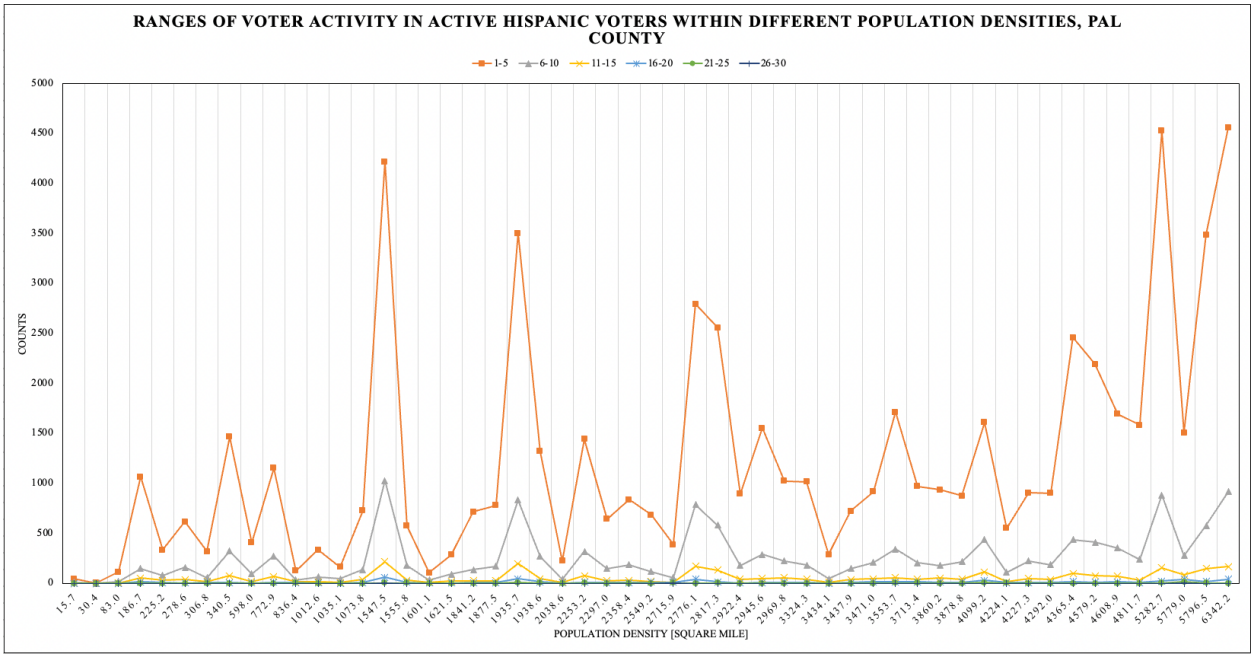

Figure 15 Voter activity amongst the population densities of Hispanic voters at Palm Beach.

Each population density corresponds to a zip code in Palm Beach. The random spikes appreciated in all ranges regardless of the population density could simply imply certain zip codes are simply more active than others or have higher populations of Hispanics than others. Overall, higher population densities lean towards higher activity, which is expected, although the action is not fully consistent, given high spikes at smaller population densities.

The range with the most voters is "1-5", which seems to continue the trend of voters that vote fewer times than not.

On the other hand, when plotting population density with party affiliation possible trends were discovered.

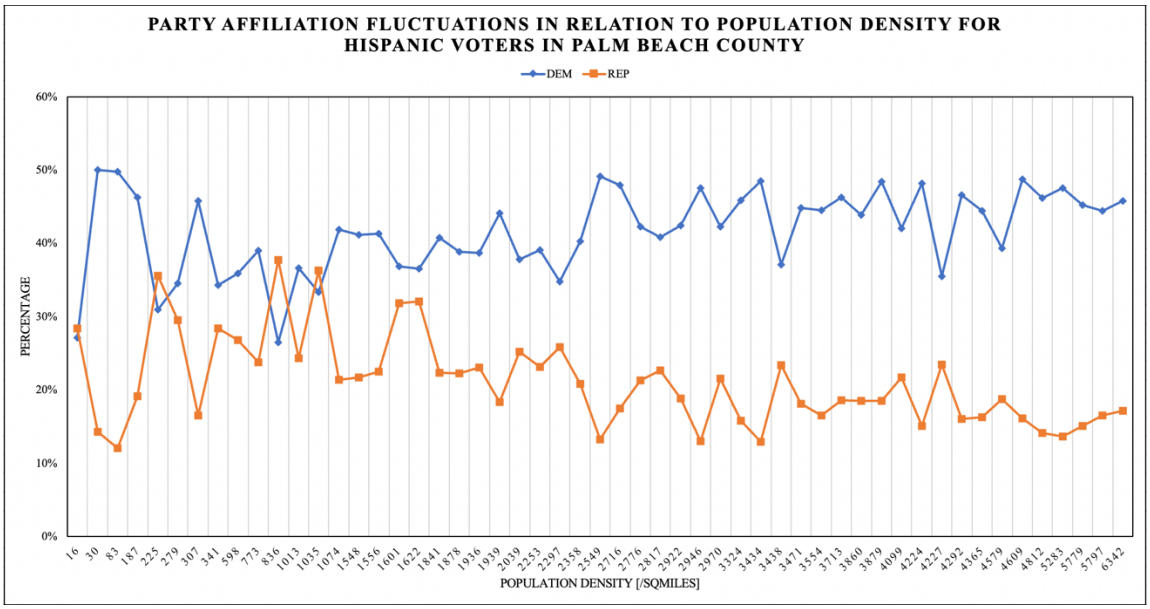

Figure 16 Population density relationship with party affiliation in Hispanic voters.

It was observed that in less densely populated areas, Democrats were still twice as likely to be ahead, yet it was the only time Republicans had the lead. After a crossover point at approximately 1,000 people per square mile, Democrats were consistently in the lead.

d. Income Salary

Once the data is extracted and matched with the voters' zip codes, the majoritarian income for that zip code is calculated as the most like income for that voter.

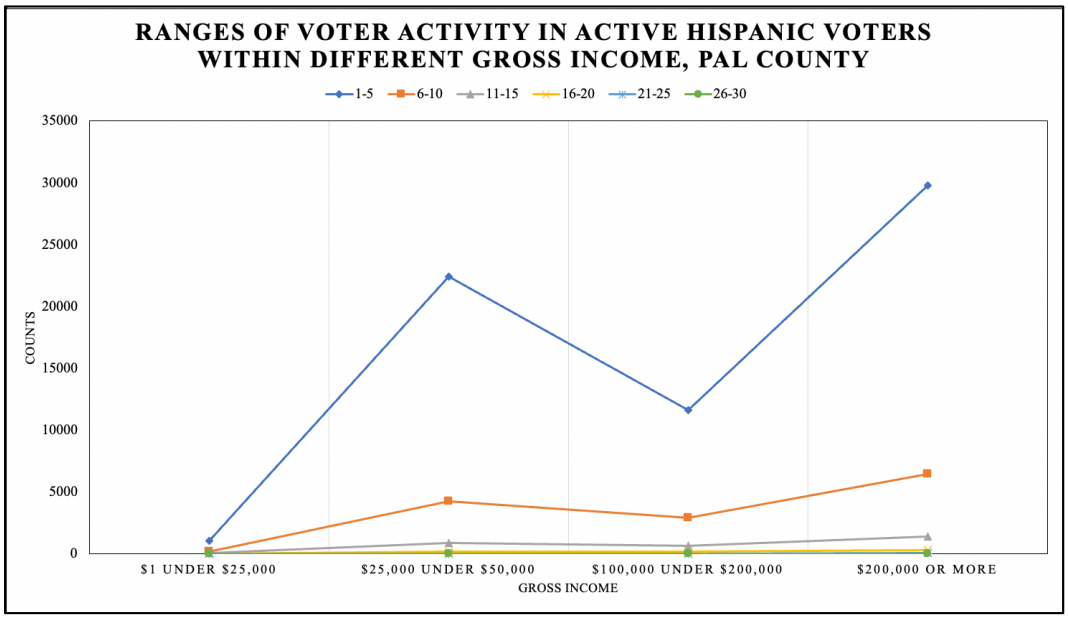

Figure 17 Voter activity amongst the income salary of Hispanic voters in Palm Beach.

Despite the spikes at the range of "$25,000 under $50,000" and the seemingly lack of voters between $50,000 and $100,000, the general trend is that higher income voters have higher voter activity.

As for income salary and party affiliation, the majoritarian party affiliation and income range per zip code is calculated and matched together. The number of zip codes with most voter affiliated in a party that have the majority income range are plotted.

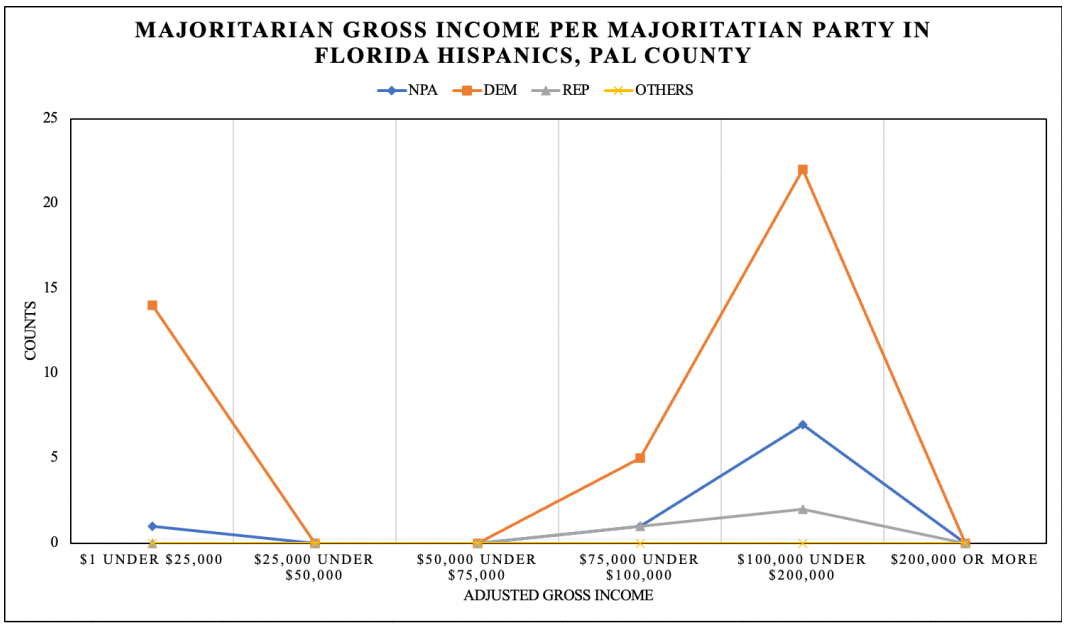

Figure 18 Income salary and party affiliation in Florida Hispanics per zip code in Palm Beach.

It was observed that most zip codes in Palm Beach are mainly affiliated with the Florida Democratic Party, and most of the voters in those zip codes are at extremes when it comes to gross income. The highest spikes are at the ranges "$1 under $25,000" and "$100,000 under $200,000".

e. Party Affiliation and Voter Activity

The two comparison variables amongst the social factors, voter activity and party affiliation are compared to determine any relationships.

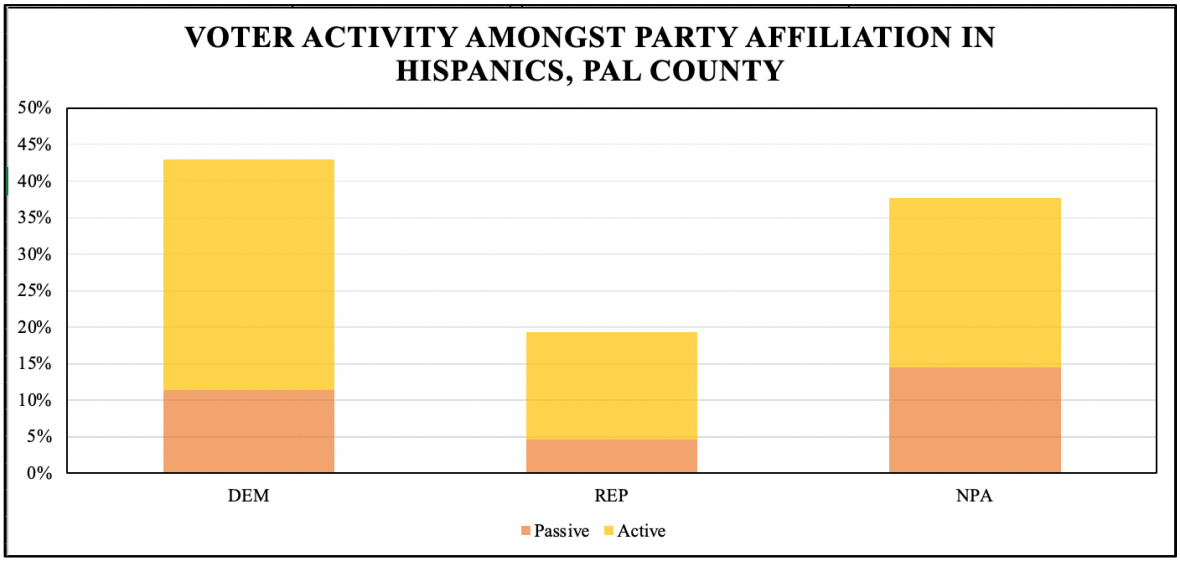

Figure 19 Voter activity within party affiliation of Hispanic voters in Palm Beach

Hispanic voters that were Non-Party affiliated had the highest percentage of passive voters within their own population. On the other hand, Republicans had the highest percentage of active voters within their own voters.

After the data for active voters is normalized, the number of people that voted a number of times is plotted by party affiliation.

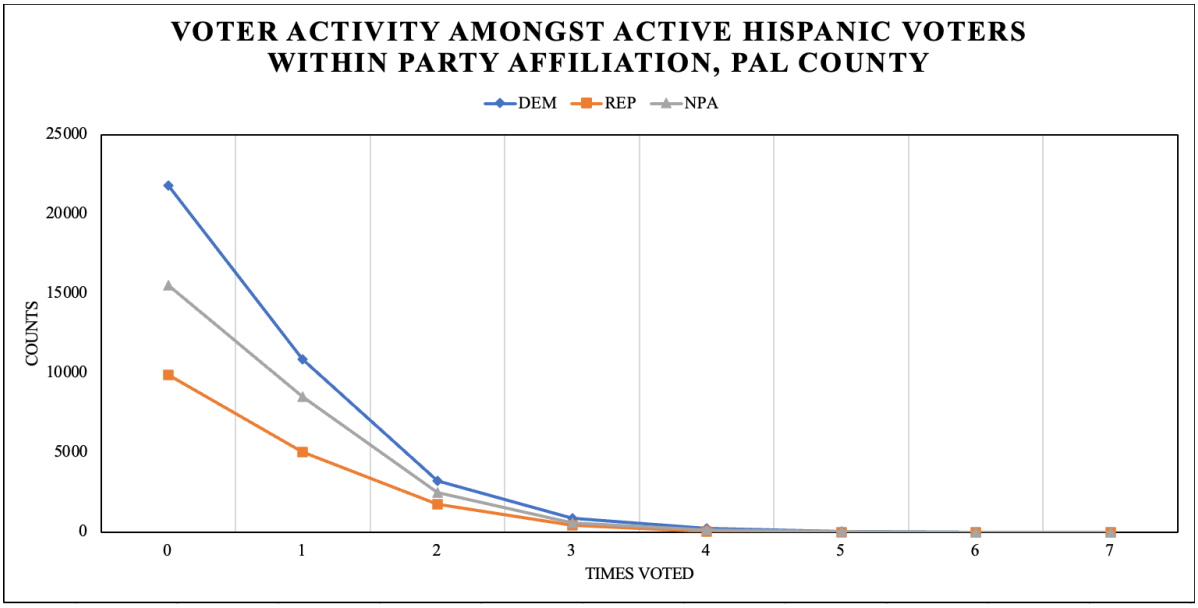

Figure 20 Voter activity within party affiliation of AHV in Palm Beach

All voters follow a similar trend of voting a few times, then flattening once the numbers increase.

V.LIMITATIONS & FUTURE IDEAS

The main limitations of the project are the massive filtering of the low-confidence Hispanic groups and the income salary dataset. The filtering of low-confidence Hispanic voters in the subgroup classification process left the project with 15% of the original AHV. The sample was still large and very useful in analysis, yet a more representative sample would be even better. As for the income salary dataset, since its data only represents an entire zip code, the analysis may prove itself far too generalized for an accurate representation. Increasing confidence for our classification process and a more detailed income salary data set would improve the investigation.

IX.REFERENCES

1. Boryga, A. (2020, October 24). Democrats push Puerto Rican voters To outmuscle CUBAN Republicans in Florida. Retrieved April 18, 2021, from https://www.sun-sentinel.com/news/politics/fl-ne-democrat-push-puerto-rican-voters-florida-20201024-6m6rwvuf6vbg5jl5mnq2cqaoxa-story.html

2. Busette, C., & Shiro, A. (2020, November 06). The importance of Understanding Latino voters in battleground states. Retrieved April 18, 2021, from https://www.brookings.edu/blog/how-we-rise/2020/11/03/the-importance-of-understanding-latino-voters-in-battleground-states/

3. Dimock, M. (2021, May 29). *Defining generations: Where Millennials end and Generation Z begins*. Pew Research Center. https://www.pewresearch.org/fact-tank/2019/01/17/where-millennials-end-and-generation-z-begins/.

4. Florida's Hispanic electorate 2008-2018. (2020, October 14). Retrieved March 03, 2021, from https://www.pewresearch.org/fact-tank/2020/10/19/latinos-make-up-record-17-of-florida-registered-voters-in-2020/ft_2020-10-14_hispanicvoters_05a/

5. Forebears. (2012, June 20). Retrieved April 1, 2021, from https://forebears.io

6. Hispanic Latino population by specific origin by county. (2010) Retrieved February 26, 2021, from http://proximityone.com/hispanic_origin.htm

7. Instituto Cervantes. (2020). El Español: Una Lengua Viva: Informe 2020. Retrieved April 17, 2021, from https://cvc.cervantes.es/lengua/espanol_lengua_viva/pdf/espanol_lengua_viva_2020.pdf

8. Krogstad, J. (2020, October 02). Most Cuban American voters identify as Republican in 2020. Retrieved April 18, 2021, from https://www.pewresearch.org/fact-tank/2020/10/02/most-cuban-american-voters-identify-as-republican-in-2020/

9. Lopez, M., Krogstad, J., & Passel, J. (2020, September 22). Who is hispanic? Retrieved April 18, 2021, from https://www.pewresearch.org/fact-tank/2020/09/15/who-is-hispanic/

10. Prysby, C., & Scavo, C. (n.d.). Voting Behavior. https://www.icpsr.umich.edu/web/pages/instructors/setups/voting.html.

11. Troy, D. (2016, August 22). *Is Population Density the Key to Understanding Voting Behavior?* Medium. https://davetroy.medium.com/is-population-density-the-key-to-understanding-voting-behavior-191acc302a2b.

12. U.S. census bureau QUICKFACTS: Florida. (n.d.). Retrieved February 26, 2021, from https://www.census.gov/quickfacts/FL

13. Vogel, M. (2020, August 26). Florida's Hispanic population boom. Retrieved February 26, 2021, from https://www.floridatrend.com/article/29770/floridas-hispanic-population-boom

14. Voter File…

15. World Media Group. (n.d.). *Florida Population Density Zip Code Rank*. USA.com. http://www.usa.com/rank/florida-state--population-density--zip-code-rank.htm.