Correlation of San Francisco Precinct-Level Demographic Data with DeLeon's Progressive Voter Index

David Latterman
dclatter@sbcglobal.net

Summary
In this paper, many demographic variables from the 2000 U.S. Census have been collated into San Francisco precincts. These have been correlated with Rich DeLeon's Progressive Voter Index (PVI) – both citywide and by district - to determine which variables consistently have positive or inverse correlations with PVI scores. One multivariate OLS regression model is also provided as an example of this kind of work with these new data.

In the bivariate correlation analysis, higher precinct percentages of Asian/Pacific Islander (API) population, older voters, owned housing units, median household income correlate strongly with lower PVI (politically moderate) scores. Hispanic, black, and younger age brackets correlate with higher (progressive) PVI scores. There is a strong trend of more conservative voting habits as a precinct's population ages. The multivariate model reveals similar conclusions, also showing that there is a more liberal trend towards precincts with higher LGBT populations.

Introduction

Building on the PVI Index that Rich DeLeon[1] created several years ago, I have collated 2000 U.S. Census data by San Francisco Voting Precinct. This paper will be the first of an occasional series examining these data with various San Francisco voting trends. Here, I examine several bivariate correlations with the demographic data vs. PVI index. This is done both citywide and by district, to see at a first pass what demographic characteristics affect the PVI throughout the City and its districts.

These data and analyses are meant to be one form of looking at the San Francisco electorate. It does need to be mentioned, however, that there are several potential issues with using these data, described below. Moreover, this is meant to augment to, not replace, poll data of how people voted or feel about certain issues. This type of work can give tantalizing indications of who votes in certain ways, but it is by no means inclusive of all variables.

In order to truly understand why certain demographics vote the way they do, nothing serves better than attending town meetings and talking to people, working on campaigns, or attending various hearings at City Hall. My goal is not to reduce the electorate to

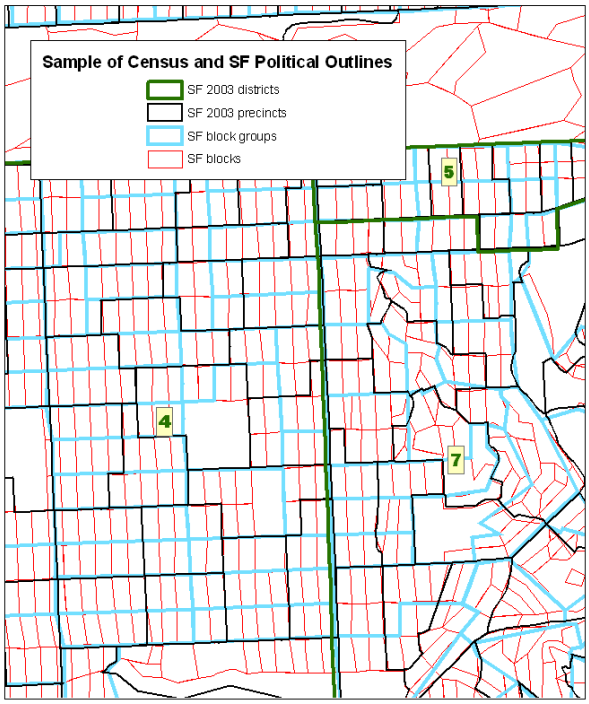
[1] I wish to sincerely thank Rich DeLeon for his critical commentary on the drafts and his quality control work on the demographic data. He was able to mimic similar results he found in his previous work, and offer much advice as to how to improve the analytical work. I look forward to many more collaborations! Also, thanks to Alex Clemens for allowing me to post this work on the Usual Suspects. His site provides a valuable service to the San Francisco electorate.

faceless numbers, it is simply to provide one tool in the effort to understand how democracy works in the most dynamic political environment in the country.

Demographic Methodology

The bulk of this work has been to collect various demographic data from the 2000 U.S. Census, aggregating them into usable values for San Francisco precincts. Because various characteristics have different smallest units of size (i.e., blocks, block groups, and tracts), different techniques had to be employed resulting in various levels of accuracy. Below are the different techniques used to obtain the data, along with an appraisal of how reliable these particular data are. Figure 1 displays all the census units (except tracts) along with SF precincts.

Figure 1: A portion of San Francisco with all geographic outlines except census tracts used in this analysis



Blocks

Blocks are the smallest unit of demographic data in the U.S. Census. In San Francisco, a census block is about 1-4 city blocks. Only some variables are available by block, like race, gender, and certain household data. Because blocks fall entirely within precincts, directly aggregating census blocks to precinct values by summation is seen as the most accurate data in this report.

Census data for block-level data and other groupings were transformed to SF precincts via grouping functions in Excel, Access, and SPSS. Precincts were examined for the number of blocks that comprised them with a nonzero population, as some blocks did have zero population because they are in commercial or industrial portions of the city.[2] It

[2] Around 18% of the nearly 5800 census blocks in San Francisco had zero population.

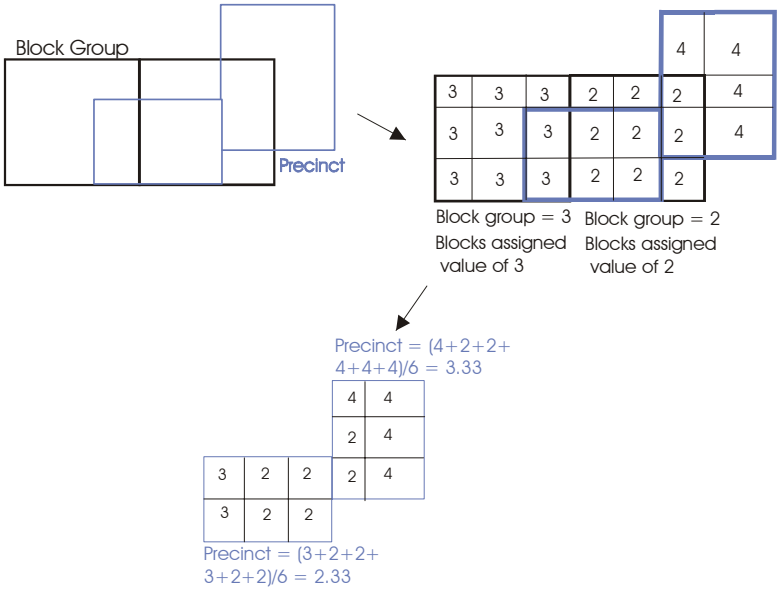
was necessary to determine that precincts had a large enough population to render analyses valid.

Block Groups

Most usable data had its smallest demographic unit as the census block group. Block groups are roughly the same size as a SF precinct, but they geographically do not coincide. To collect usable data from block groups, I broke the block group into its constituent blocks, assigning each block the value for the selected characteristic. I then aggregated the blocks into their respective SF precincts. For example, if a block group had a median income of $50,000, each block within that group was assigned a median income value of $50,000. A precinct took the average value of its component blocks, which may have come from different block groups.

This method of breaking block groups into their component blocks was seen as more reliable than averaging all the block group values that are adjacent to, intersect with, or fall inside a SF precinct. This adjacency method produced values approximately 15-20% less accurate (as compared to the block group value) than using the component block method. The component block method also has the added advantage of weighting block groups that have a larger presence in any given precinct.

Figure 2: Diagram showing the conversion of block group data to precincts



One slight issue with this technique is that it was very difficult to use the block groups from Daly City, immediately south of San Francisco. Considering there are only about 13 precincts that this affects, the effort was aborted. Moreover, the Daly City block groups that are adjacent to the D10 and D11 precincts are demographically similar to the

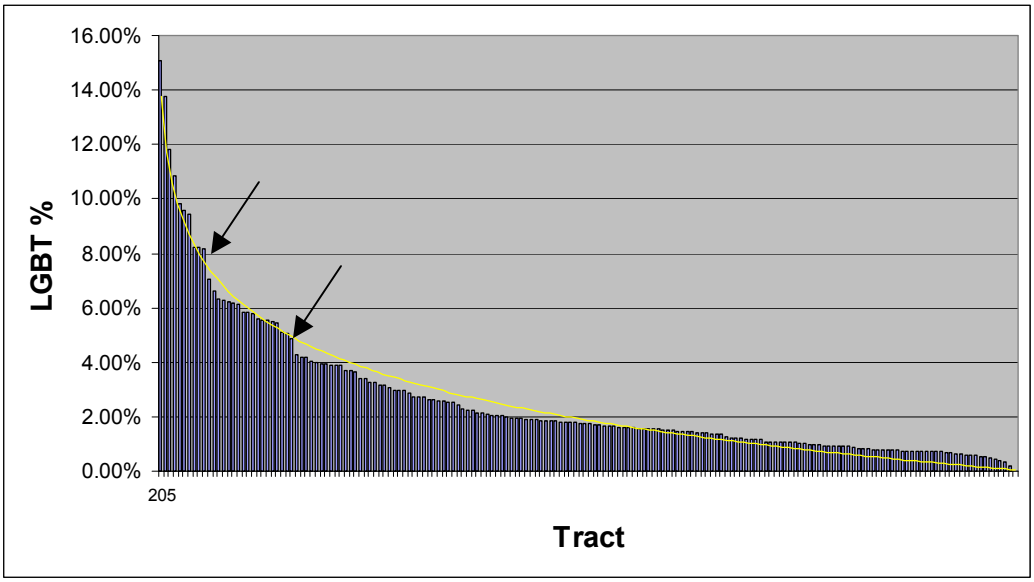
San Francisco populations, so the relative demographic compositions are not particularly compromised.

Tracts

The only characteristic for which tract-level data was the smallest geographic unit was the LGBT index. The index was manually constructed by examining precincts within and intersecting relevant census tracts. The tracts were seen as too big to simply break them into their constituent blocks as above. Because using tract-level data was seen as less reliable as the other method, these characteristics were used in the construction of dummy variables.

For LGBT data, the percentages for the index values were constructed by looking at all the tract values for estimated LGBT householders, and then deciding where the breaks were. The two major breaks came at the 4.5% and 8% household percentages (see Figure 3). The overall tract values were best fit by a natural exponential curve. Precincts that intersected these tracts took on the tract values, and if tracts of different value categories both intersect a precinct, the precinct was assigned the higher value. This is seen as a reliable methodology, as most of the precincts with values in the above categories fell within D8 (mainly The Castro).

Figure 3: Chart showing LGBT% per census tract. Arrows point to the 8% and 4.5% breaks.



Analytical Methodology and Reasoning

For this analysis, 29 demographic variables were compared by precinct, as independent variables, with Rich DeLeon's PVI index by precinct as a dependent variable[3]. Although any number of election or ballot initiative results could be used as a dependent variable,

[3] For a detailed explanation as to the composition and methodology of the PVI index, please see DeLeon's papers located at http://www.sfusualsuspects.com/deleon.html.

DeLeon's index is seen to be the most reliable overall measure of San Francisco voting trends by political leanings. The PVI incorporates many recent ballot initiatives, and correlates remarkably to 2003 elections, so it is an excellent first step to examine the reliability of the demographic data.

Bivariate correlations are seen as a good initial step to analyzing the significance of these kind of data. Mainly, this was to see what variables consistently and strongly correlate with PVI both citywide and locally. Before stronger and more accurate models can be constructed, like using multivariate regression analysis, it is necessary to first see which variables are the most correlative to the PVI. These correlations can be considered an introductory analysis.

Each variable was correlated by precinct to its corresponding PVI value, both citywide and by district. The Pearson correlation, or R value, is displayed here. For those that are more used to seeing R^2 values, R is simply its square root. R is used in order to preserve the correlative sign.

It must be reiterated strongly that *precinct* values are being correlated, not individual voters. That type of information can come from polls, but those data are not readily available publicly. We cannot escape the "ecological fallacy" in which data from individual voters are lost when they are aggregated into precincts; however, these are the most commonly available data and sufficient to make the rather broad interpretations that are attempted here.

A second potential problem with the correlations is non-linearity in the independent variables. Some of the bivariate correlations are neither linear nor random. If the graph of Y on X seems parabolic or clusters in regions, the data correlation is non-linear, and the R value will be close to zero. However, this is not to say there is no relationship between the variables. Non-linearity is suggested in the gender and some of the racial variables (Hispanic, black). These are discussed in further detail below. In these cases, district-level correlations may be more reliable.

Results

Table 1 displays the variables used in this analysis, along with its description, demographic unit available from the census, and relevant notes. These are not all the variables compiled, but only those used in this paper or mentioned in the text. A set of some of these variables is included with this work in the accompanying spreadsheet.

Table 1: List of variables and descriptions used in this report

Variable	Description	Census grouping	Census Notes
2000 Census-derived percentages per precinct			
p_white	Race - white	Block	White alone or self-identified mixed
p_black	Race - black	Block	Black alone or self-identified mixed
p_api	Race – Asian or Pacific Islander	Block	Asian/PI alone or self-identified mixed
p_hisp	Ethnicity – Latino/Hispanic	Block	Identified Latino or Hispanic
p_male	Gender – male	Block	
p_female	Gender - female	Block	

p_m_017	Age by Gender, male, 0-17	Block	
p_m_1824	Age by Gender, male, 18-24	Block	
p_m_2529	Age by Gender, male, 25-29	Block	
p_m_3039	Age by Gender, male, 30-39	Block	
p_m_4049	Age by Gender, male, 40-49	Block	
p_m_5059	Age by Gender, male, 50-59	Block	
p_m_6069	Age by Gender, male, 60-69	Block	
p_m_70	Age by Gender, male, 70 and older	Block	
p_f_017	Age by Gender, female, 0-17	Block	
p_f_1824	Age by Gender, female, 18-24	Block	
p_f_2529	Age by Gender, female, 25-29	Block	
p_f_3039	Age by Gender, female, 30-39	Block	
p_f_4049	Age by Gender, female, 40-49	Block	
p_f_5059	Age by Gender, female, 50-59	Block	
p_f_6069	Age by Gender, female, 60-69	Block	
p_f_70	Age by Gender, female, 70 and older	Block	
p_t_017	Age by Gender, total, 0-17	Block	
p_t_1824	Age by Gender, total, 18-24	Block	
p_t_2529	Age by Gender, total, 25-29	Block	
p_t_3039	Age by Gender, total, 30-39	Block	
p_t_4049	Age by Gender, total, 40-49	Block	
p_t_5059	Age by Gender, total, 50-59	Block	
p_t_6069	Age by Gender, total, 60-69	Block	
p_t_70	Age by Gender, total, 70 and older	Block	
p_own_hu	Owned housing units	Block	
p_p_owhu	Population in owned housing units	Block	
p_native	Population U.S. born	Block Group	
p_foregn	Population not U.S. born	Block Group	
p_ntrlzd	Naturalized immigrant population	Block Group	
p_im_nzd	Naturalized immigrant population as a percentage of immigrants	Block Group	
p25m_nhs	Not graduated from high school - male	Block Group	Population over 25
p25m_hs	High school graduate - male	Block Group	Population over 25
p25m_sc	Some college or Associate's degree - male	Block Group	Population over 25, Summed 'some college' and 'associates degree'
p25m_bac	Bachelor's degree - female	Block Group	Population over 25
p25m_adv	Advanced degree - female	Block Group	Population over 25, Summed MS, PhD, or Professional degree
p25f_nhs	Not graduated from high school - female	Block Group	Population over 25
p25f_hs	High school graduate - female	Block Group	Population over 25
p25f_sc	Some college or Associate's degree - female	Block Group	Population over 25, Summed 'some college' and 'associates degree'
p25f_bac	Bachelor's degree - male	Block Group	Population over 25
p25f_adv	Advanced degree - male	Block Group	Population over 25, Summed MS, PhD, or Professional degree
p25t_nhs	Not graduated from high school - total	Block Group	Population over 25
p25t_hs	High school graduate - total	Block Group	Population over 25
p25t_sc	Some college or Associate's degree - total	Block Group	Population over 25, Summed 'some college' and 'associates degree'
p25t_bac	Bachelor's degree - total	Block Group	Population over 25
p25t_adv	Advanced degree - total	Block Group	Population over 25, Summed MS, PhD, or Professional degree
p16m_wrk	Employed male	Block Group	1999 Population over 16
p16m_nwk	Not employed male	Block Group	1999 Population over 16
p16f_wrk	Employed female	Block Group	1999 Population over 16
p16f_nwk	Not employed female	Block Group	1999 Population over 16
p16t_wrk	Employed total	Block Group	1999 Population over 16
p16t_nwk	Not employed total	Block Group	1999 Population over 16
2000 Census-derived other variables			
med_hh_i	Median household income	Block Group	
families	Number of family units	Block Group	
fam_mar	Number of family units – married couples	Block Group	
fam_chrn	Number of family units – with children	Block Group	
lgbt8	LGBT index – greater than 8% per precinct	Tract	Dummy variable
lgbt4_8	LGBT index – 4.5-8% per precinct	Tract	Dummy variable

Table 2 is the correlations results table; the rows represent the demographic variables and the columns are the geographic area for the correlation. The first column displays citywide results and the subsequent columns the districts. N is provided for each geographical region. Statistical significance here is indicated by color, where black indicates the correlation is significant at the 99% level (most reliable), blue values are significant at the 95% level, and red values indicate the correlation is significant only below the 95% level and therefore not as reliable.

Significance is controlled by the number of values (n) and the strength of the correlation. The larger the number of observations, the less R can be for it still to be considered statistically significant. For the citywide observations (n = 561), R can be as low as 0.1 and still be significant at the 99% level. In that case, we can say convincingly that that variable has little correlation with PVI. For the district correlations, where n = 41 through 65, R must be over 0.2 or higher for a significant correlation. However, if R = 0.2 then the correlation is still very weak. Generally, I consider $R \approx \pm 0.5$ to be a meaningful correlation, which corresponds to an R^2 value of 0.25.

Appendix 1 shows the raw percentages for each variable both citywide and in the districts. This is provided so readers can observe the relative importance of each variable in the districts, and for general demographic purposes.

Table 2: R (Pearson correlation) value of selected variables vs. PVI. Black text indicates values significant at the 99% level. Blue text indicates values significant at the 95% level. Red text indicates values NOT significant at the 95% level.

		District										
	City-wide	1	2	3	4	5	6	7	8	9	10	11
precinct n=	561	47	60	46	46	65	45	55	64	41	50	42
p_white	-0.029	0.183	-0.614	-0.215	0.426	0.145	-0.488	-0.327	-0.036	0.255	0.377	-0.268
p_black	0.269	0.448	0.648	0.416	0.495	0.094	-0.132	0.225	-0.054	-0.172	0.440	0.499
p_api	-0.472	-0.348	0.331	0.071	-0.532	-0.542	-0.243	-0.140	-0.504	-0.884	-0.909	-0.512
p_hisp	0.427	0.492	0.708	0.621	0.400	0.398	0.686	0.575	0.521	0.558	-0.153	-0.022
p_female	-0.469	0.399	-0.112	-0.596	-0.589	-0.692	-0.156	0.254	-0.037	-0.539	0.057	-0.146
p_t_017	-0.143	-0.517	-0.110	-0.009	-0.560	-0.134	0.615	-0.169	-0.296	-0.304	0.164	0.292
p_t_1824	0.268	0.461	0.591	0.726	0.550	0.460	0.312	0.317	0.642	0.241	-0.125	0.369
p_m_2529	0.578	0.328	0.491	0.440	0.821	0.607	0.075	0.638	0.748	0.614	0.190	0.426
p_m_3039	0.455	0.151	0.012	0.246	0.734	0.571	-0.219	0.503	0.359	0.723	0.263	0.025
p_m_4049	0.230	-0.604	-0.027	0.323	0.235	0.304	-0.016	-0.137	-0.168	0.127	0.073	-0.109
p_m_5059	-0.106	-0.436	-0.343	-0.223	-0.060	-0.060	0.071	-0.696	-0.514	-0.212	-0.268	-0.073
p_m_6069	-0.440	-0.597	-0.516	-0.151	-0.345	-0.442	-0.067	-0.574	-0.689	-0.657	-0.766	-0.431
p_m_70	-0.456	-0.610	-0.273	-0.329	-0.760	-0.529	-0.301	-0.662	-0.501	-0.700	-0.675	-0.204
p_f_2529	0.407	0.421	0.375	0.276	0.733	0.422	0.106	0.581	0.820	0.645	0.591	0.110
p_f_3039	0.091	0.079	-0.091	-0.207	0.222	0.076	-0.161	0.550	0.436	0.418	0.429	-0.444
p_f_4049	-0.396	-0.453	-0.105	-0.600	-0.264	-0.202	0.041	-0.180	-0.427	-0.231	0.084	-0.157
p_f_5059	-0.607	-0.425	-0.510	-0.686	-0.255	-0.567	-0.211	-0.607	-0.571	0.309	-0.530	-0.203
p_f_6069	-0.607	-0.606	-0.469	-0.417	-0.478	-0.453	-0.298	-0.639	-0.603	-0.691	-0.788	-0.385
p_f_70	-0.429	-0.712	-0.028	-0.359	-0.795	-0.521	-0.353	-0.373	-0.348	-0.635	-0.650	-0.026
p_own_hu	-0.596	-0.654	-0.460	-0.686	-0.824	-0.290	-0.381	-0.680	-0.824	-0.689	-0.756	-0.254
p_foregn	-0.180	-0.167	0.307	0.167	0.053	-0.489	0.235	0.331	0.307	-0.299	-0.822	-0.424
p_im_nzd	-0.423	-0.287	0.240	-0.032	-0.227	-0.563	-0.132	-0.137	-0.105	-0.854	-0.874	-0.450
p25t_nhs	0.089	-0.251	0.412	0.060	-0.010	-0.264	0.303	0.016	0.448	-0.157	-0.461	0.010
p25t_hs	0.021	0.003	0.372	0.612	-0.343	-0.029	0.453	0.244	0.209	-0.502	-0.125	-0.118
p25t_sc	0.092	-0.290	0.354	0.589	0.087	0.205	0.396	0.409	0.062	-0.399	-0.071	0.113
p25t_bac	-0.018	0.347	0.049	-0.187	0.262	0.243	-0.403	-0.248	0.181	0.420	0.252	-0.068
p25t_adv	-0.160	0.175	-0.627	-0.575	-0.101	-0.199	-0.435	-0.264	-0.599	0.359	0.383	0.023
p16t_nwk	-0.224	-0.233	-0.141	0.026	-0.487	-0.405	-0.144	-0.364	-0.375	-0.632	-0.101	-0.030
med_hh_i	-0.488	-0.292	-0.621	-0.619	-0.230	-0.093	-0.262	-0.618	-0.633	-0.163	-0.142	-0.021

Most of the values in the citywide correlations are statistically significant at the 99% level. However, significant values lessen in the districts as the number of values significantly diminishes. D6 and D11, with n = 45 and 42 respectively, have fewer reliable results, due to the fewer number of precincts and fewer correlative variables with PVI. Notice that D9, with n = 41, has more significant statistically significant values than D6 or D11, due to stronger correlations. This should not be surprising considering the strong progressive identity of D9. D6, with South Beach and the Tenderloin, is quite varied, as is the multi-racial D11.

Correlations indicate whether a variable is positively or negatively correlated with PVI, where a positive correlation value indicates the particular demographic is more likely –

on an aggregate precinct level – to vote progressively. It must be stressed yet again that these values do represent rollups from individual voter data aggregated to their precincts, so these data can only suggest certain trends. Yet, the higher the correlation value, the stronger a conclusion one can make from the correlation.

Table 3 shows the correlations that are greater than 0.5, a general benchmark of a relatively strong correlation. Values are only displayed if they are significant at greater than the 99% level (which is true for any R value this high). The variables listed here can be said to have some particular progressive or moderate leaning, as a group. Blue values indicate the correlation is positive (higher PVI value). Interestingly, more strong correlations are associated with moderate voting trends.

Table 3: Table showing the correlations of demographic variables and precinct PVI scores citywide and district wide greater than 0.5. Black indicates an inverse correlation with PVI (more moderate), while blue indicates more progressive (higher PVI value) precincts.

City-wide	D1	D2	D3	D4	D5
p_m_2529	p_t_017	p_white	p_hisp	p_api	p_api
p_f_5059	p_m_4049	p_black	p_female	p_female	p_f_5059
p_f_6069	p_m_6069	p_hisp	p_t_1824	p_t_017	p_f_6069
p_own_hu	p_m_70	p_m_6069	p_f_4049	p_t_1824	p_f_70
	p_f_6069	p_f_5059	p_f_5059	p_m_2529	p_own_hu
	p_f_70	p25t_adv	p_own_hu	p_m_3039	p_foregn
	p_own_hu	med_hh_i	p25t_hs	p_m_70	p_im_nzd
			p25t_sc	p_f_2529	
			p25t_adv	p_f_70	
			med_hh_i	p_own_hu	

D6	D7	D8	D9	D10	D11
p_hisp	p_hisp	p_api	p_api	p_api	p_api
p_t_017	p_m_2529	p_hisp	p_hisp	p_m_6069	
	p_m_3039	p_t_1824	p_m_2529	p_m_70	
	p_m_5059	p_m_2529	p_m_3039	p_f_2529	
	p_m_6069	p_m_5059	p_m_6069	p_f_5059	
	p_m_70	p_m_6069	p_m_70	p_f_6069	
	p_f_2529	p_m_70	p_f_2529	p_f_70	
	p_f_3039	p_f_2529	p_f_6069	p_own_hu	
	p_f_5059	p_f_5059	p_f_70	p_foregn	
	p_f_6069	p_f_6069	p_own_hu		
	med_hh_i	p_own_hu	p_im_nzd		
		p25t_adv	p25t_hs		
		med_hh_i	p16t_nwk		

Discussion

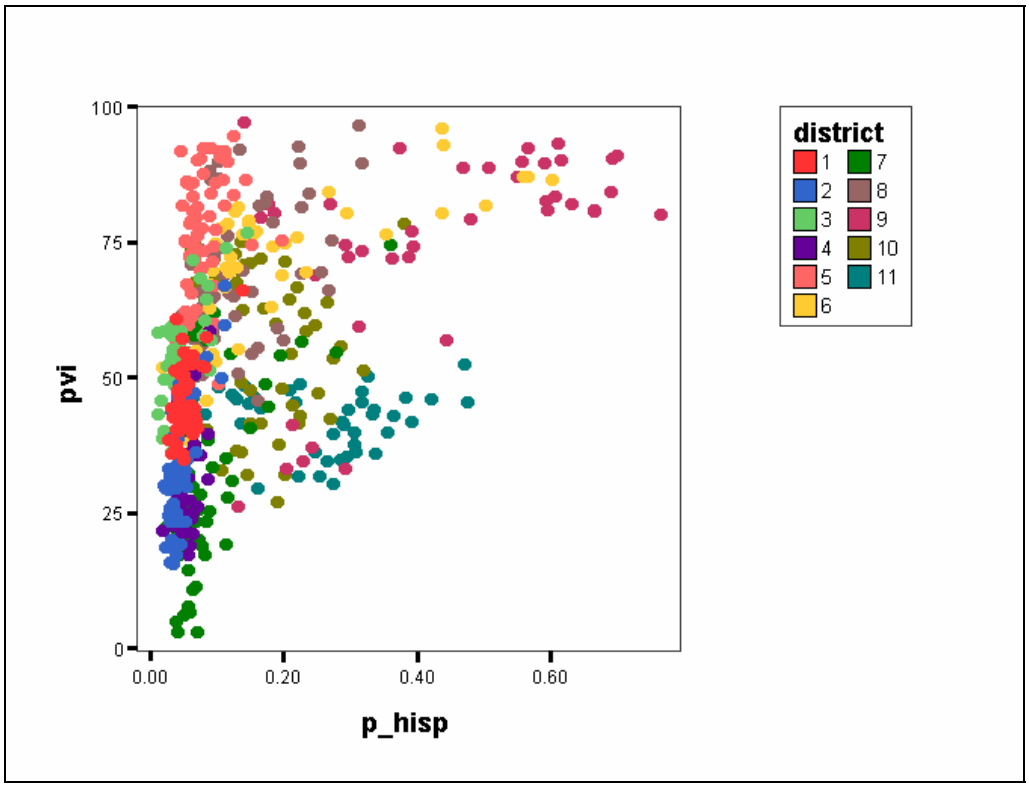
The correlation table is provided in part to allow the reader to draw his or her own conclusions. One should look for consistent trends between citywide and district results. Some brief comments, grouped by general type of demographic trait, are provided here on what struck me as important.

Race/ethnicity

The strongest correlations are high-percentage API precincts to low PVI scores, indicating moderate voting trends. Although these correlations are somewhat strong citywide, they are especially strong in the precincts in which these groups are the minority. For instance, the API correlations to low PVI scores are very high in D9 and D10, with high Hispanic and African-American populations, respectively.

The correlation of percent Hispanic citywide is somewhat nonlinear (Figure 4). It is noteworthy that where the precincts seems to "flare out" are in the progressive districts of 6 and 9. This indicates that areas of higher Hispanic population are more liberal than areas with lower Hispanic population, suggesting that place matters and may carry its own political ideology.

Figure 4: Chart of precinct percentage Hispanic with PVI, by district.

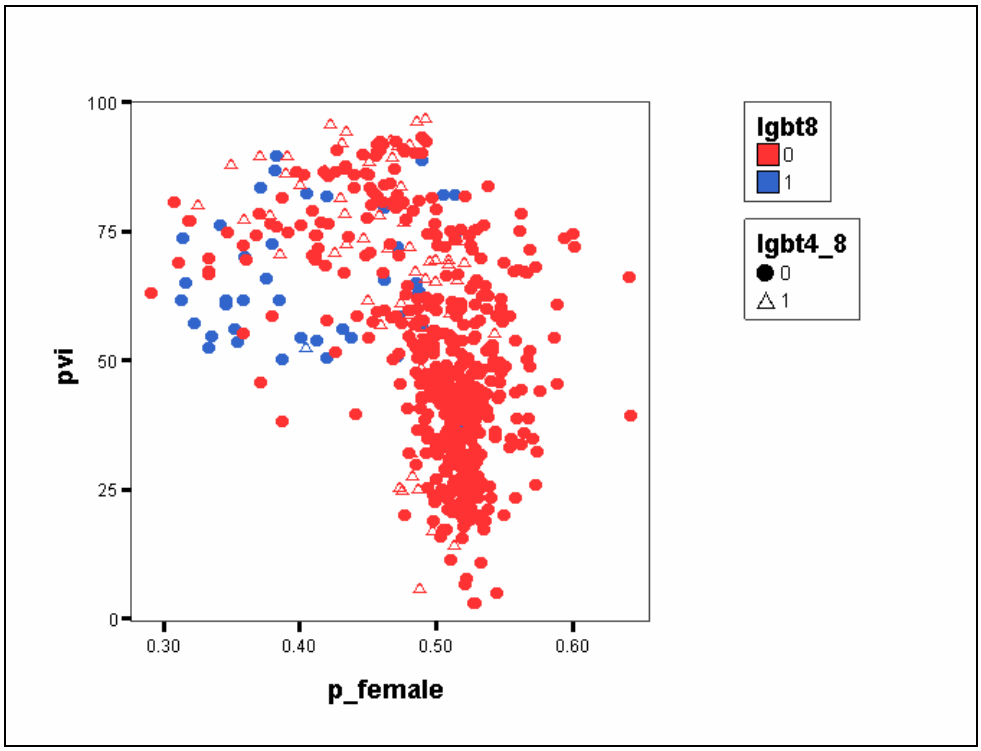


High-percentage white precincts do not correlate strongly citywide with PVI, but that is understandable considering how varied whites are throughout the city. For example, whites comprise most of D7 and D5, which are disparate politically. Black precincts correlate somewhat well with higher PVI scores, though not in all districts. It is interesting that the strongest correlation with black precincts and high PVI scores is in heavily white D2 (Marina), and it is also high in heavily Asian D4 (The Sunset).

Gender

Because of the non-linearity of gender with PVI, at least on a citywide scale, overall trends are difficult to discern. Figure 5 shows the percent female per precinct versus PVI score. The highlighted markers are the precincts that have LGBT indices set to one, meaning they have high LGBT populations. Interestingly, but perhaps not surprisingly, the precincts with the lower percentage of females correlate to the high LGBT precincts of the Castro – and these precincts correlate with higher PVI scores[4].

Figure 5: Correlation of precinct female percentage with PVI. Blue and triangle markers are precincts with LGBT indices of 1.



Age

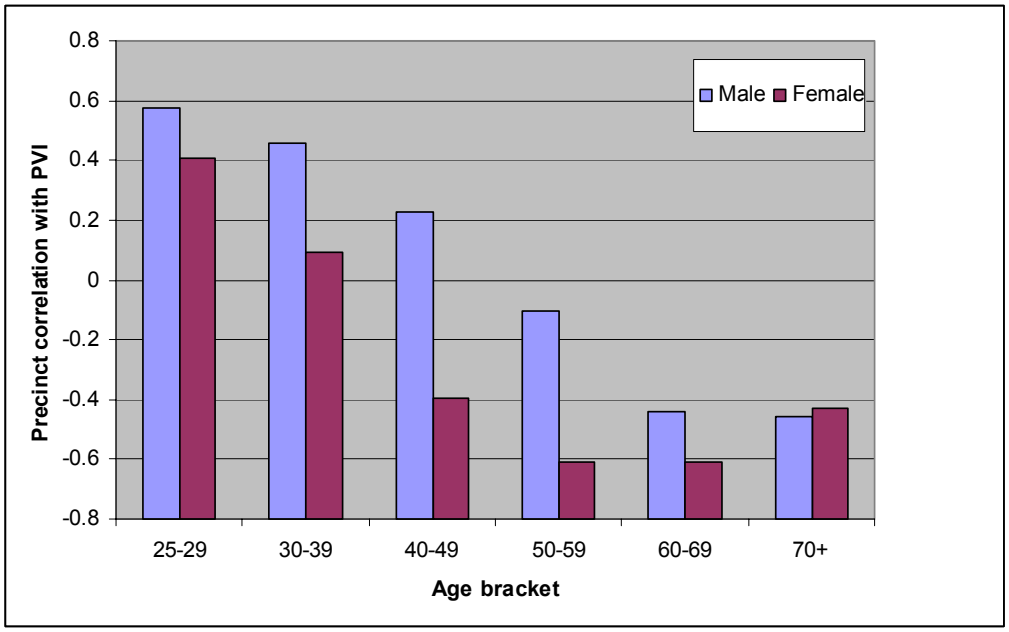
Age correlates remarkably well with PVI scores in that the older the precinct population, the lower the PVI scores to which it correlates. Looking citywide - and this trend is visible in most districts - the precinct correlations start strong and high positive for both men and women at the 25–29 age bracket, and diminish until the correlations become inverse in the 50-59 age bracket for men and 40-49 age bracket for women. The trend seems to reverse a bit once the precincts have more people over 70. This relationship also reinforces the trend of a more conservative female population. These data suggest analytically that people really do become more conservative as they get older (see Figure 6).

The age brackets for 0-17 age were combined, as were the 18-24 age brackets. The 0-17 populations indicate children, and precincts with more children correlate slightly, but not

[4] Thanks to Rich for pointing this out.

too significantly, with lower PVI scores. Although family data were collected for this research, it was not examined for this paper. The 18-24 age bracket indicates students and people just out of college (if they went). Not surprisingly, these precincts correlate somewhat positively with PVI, especially in D3, D4, and D8.

Figure 6: Chart showing the precinct PVI correlation progression with age. Note the steady move toward lower PVI scores as the precinct population ages.



Immigration status

Generally, immigration status itself does not correlate well with PVI. The citywide correlation for precincts with high foreign-born populations is low inverse, while for most of the districts it does not correlate significantly. Notable exceptions are D5 and D10, otherwise relatively liberal Districts, but with high African-American populations.

Precincts with large numbers of naturalized immigrants, however, correlate much better with low PVI scores, especially in D9 and D10. This trend may have implications for the current debate in San Francisco as to whether non-citizens can vote in local elections. There is no voting record for non-citizens, but precincts with high percentages of naturalized immigrant populations appear to vote somewhat conservatively.

Education

For the most part, this does not correlate too well to PVI scores. Citywide, only precincts with a high number of people with advanced degrees correlate somewhat to lower PVI scores. This trend is more pronounced in D2, D3, and D8 – relatively affluent districts. Notably, the one correlation with advanced degrees and high PVI scores is in D10.

Other education correlations are spotty, with some strong positive correlations and some stronger inverse correlations. It is likely that these variables need to be interacted with

other variables in more advanced models in order to discern the true effect of education on San Francisco voting patterns.

Housing status, work status, and income

There is some correlation between populations with lower employment and lower PVI scores. All statistically significant correlations are inverse, and surprisingly stronger in D9 with its strong immigrant workforce.

As may be expected, some of the strongest inverse correlations are with wealthy precincts and precincts with high home ownership rates. All statistically significant correlations for median household income are inverse, and all correlations for home ownership are inverse, yielding some of the strongest correlations in the analysis.

It is interesting to note where the exceptions lie, and this seems to be D11 and D5 income for housing, though the correlations are not significant at the 95% level. The fact D5 somewhat bucks the housing trend is not so surprising, but Inner Siberia, with its mixed demographics, from old labor whites to Asian and Latino immigrants, seems to do its own thing.

A multivariate model

Given the breath of the available demographic data available for conversion into SF geographic precincts, there are many possibilities for analytical studies with these data. PVI does not necessarily have to be one of the correlated variables in a bivariate model. We can also examine various elections and ballot initiatives against demographic variables. And, we can employ multivariate correlations instead of multiple bivariate correlations. Detailed ordinary least square (OLS) models with several variables can provide greater insight into the relative importance of certain variables.

Table 4 shows one such model, generated citywide for all San Francisco precincts. PVI is the dependent variable, while several variables that were used in the bivariate analysis are independent variables. Variables were chosen that had a clear correlative linear relationship with PVI. These are listed in the tables below.

The only age variable used is a rolled up value for precinct percentages of people over 50 years old (P_T_50). P_DEGREE represents all people with either a bachelor's or an advanced degree. Two LGBT dummy variables were added, indicating precincts with 4.5- 8% same sex households and greater than 8% same sex households. Also, dummy variables for D2 and D5 were added to the model, because these precincts are both at the far ends of the right and left spectrum, respectively. Furthermore, according to DeLeon, these districts are more politically extreme than they should be based on residual analysis (not shown) of an OLS model without these dummies.

Table 4: OLS model of several demographic variables vs. PVI, with Model and ANOVA tables (n=559)

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
Model	B	Std. Error	Beta		
(Constant)	110.493	1.983		55.722	.000
P_API	-40.182	3.271	-.358	-12.284	.000
P_OWN_HU	-28.028	2.604	-.348	-10.764	.000
LGBT8	10.825	1.754	.137	6.171	.000
LGBT4_8	11.926	1.405	.183	8.488	.000
P_DEGREE	-23.421	3.886	-.215	-6.026	.000
MED_HH_I	-1.730E-04	.000	-.192	-4.690	.000
P_T_50	-51.378	5.175	-.224	-9.928	.000
DIS_2	-14.954	1.739	-.220	-8.601	.000
DIS_5	9.472	1.483	.144	6.386	.000

Dependent Variable: PVI

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	201713.363	9	22412.596	266.521	.000
Residual	46251.265	550	84.093		
Total	247964.628	559			

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.902	.813	.810	9.17

To make sure the model meets the OLS assumptions, I include the P-P plot (Figure 7), and the residual plot (Figure 8) testing for equal variances. Looking at the plots below, the assumption of normal distribution and equal variances appears to be met[5]. Moreover, looking at a colinearity table of all the independent variables (not shown), no correlations of any two variables with each other are greater than 0.8 (after which colinearity becomes a concern).

[5] Several residuals fall outside the ±2 SD range, but the overall shape shows relatively consistent variances on either side of 0 SD.

Figure 7: P-P plot testing for assumption of normal distribution

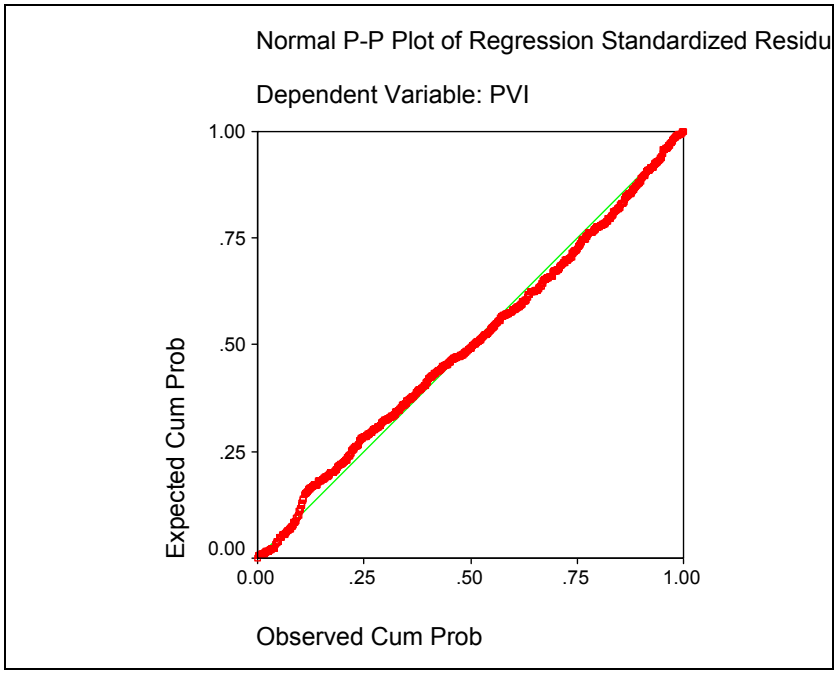
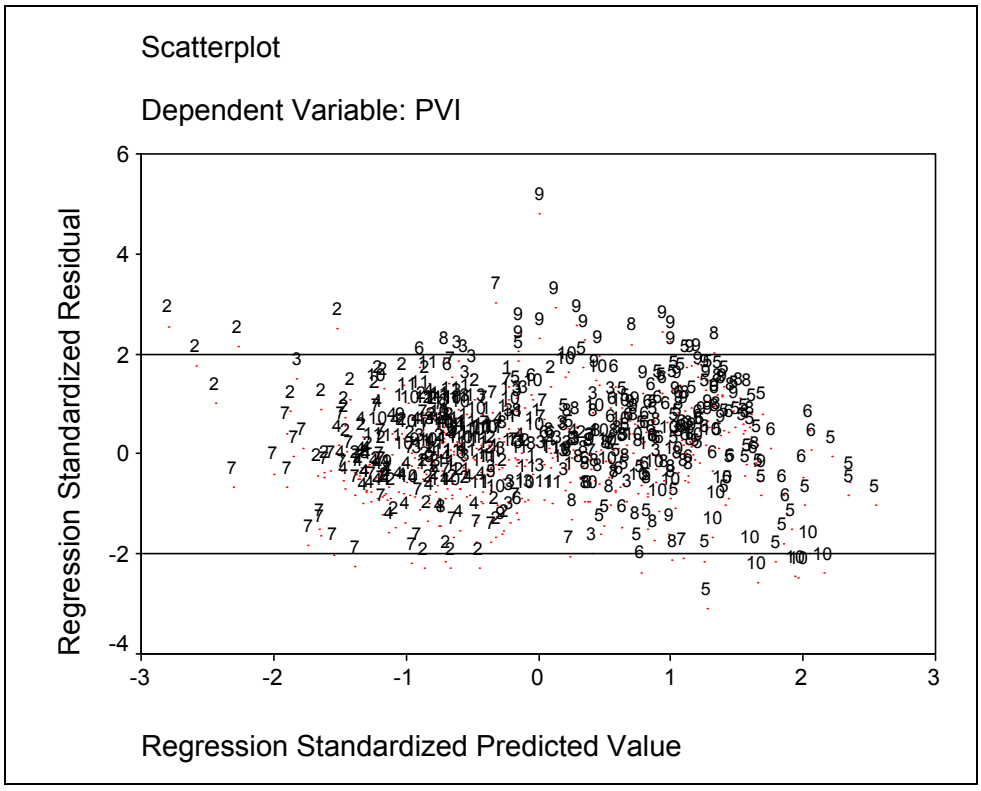


Figure 8: Residual plot testing for equal variance assumption



Looking at the above model result, it is more important to look at the signs than the absolute coefficients. All B coefficients are significant at the 95% level. The R^2 is 0.813, so the above variables account for 81% of the PVI variation – a pretty good fit.

All signs on the coefficients are negative, meaning these variables correlate with lower PVI scores where they have higher populations of that particular characteristic. The only variables in this model that correlate positively with PVI are the dummy variables for precincts with higher LGBT populations, and the dummy for district 5.[6]

As seen in the bivariate analysis, the precincts with higher proportions of Asians, people with advanced degrees, people over 50, and owned housing units are correlated with lower PVI scores. The one variable that is difficult to explain is income, which does not correlate as strongly with PVI in this model as it does when directly correlated with PVI. Looking at its standardized beta, income can be considered only a moderately important variable.

The way to read this is, by plugging in some numbers, that if the median household income of a precinct rises by $10,000, its PVI score will decrease by 1.73. This variable has less of an effect than, for example, raising the over 50 percentage of a precinct by 1%, which would decrease the precinct's PVI by 0.51.

This is just one example of an OLS model, and there are no doubt better ones that can be gleaned from the data. I only include this as a starting point to consider future work.

Conclusions

These data are only meant to suggest certain trends gleaned from the U.S. Census and voting patterns from the past several years. All of this work is predicated on the reliability of the U.S. Census, my extraction methodology, and the PVI Index. I feel pretty confident in the first and third, but I welcome criticism on how creating the precinct demographics can be improved.

I also only present findings and point out some notable patterns or values. The data is presented for all, and I make no attempt at sweeping generalizations of various demographics and their voting patterns based on these data (the 'whys' of the trends we see). Workers within the neighborhoods and in the political world can no doubt draw deeper meanings than I can based on experience.

One thing that particularly interests me is the idea that "place matters" in San Francisco. People with similar demographic characteristics can have markedly different politics depending on where they live in the City. There are some clear citywide trends, but there is a tremendous amount of regional variability in such a small city. Do people in San Francisco self-select to live in areas that match their politics, or does a place impart its views on people? Further work may concentrate on this phenomenon.

[6] This model without the D2 and D5 dummies had an R^2 of 0.752.

Appendix 1: Chart showing the percentage values at the District level for all the variables used in this analysis. Coupled with the correlations shown above, we can infer much about who lives in various parts of the City and how they vote (at least at the precinct level).

DISTRICT	Citywide	1	2	3	4	5	6	7	8	9	10	11
P_WHITE	43.7%	46.2%	78.4%	43.9%	38.4%	57.7%	39.1%	53.0%	70.6%	28.2%	16.1%	16.7%
P_BLACK	8.2%	2.2%	1.9%	2.0%	1.4%	16.3%	10.2%	4.7%	4.0%	4.6%	31.5%	8.9%
P_API	31.4%	44.5%	13.8%	47.7%	53.2%	15.9%	26.5%	32.3%	10.3%	20.9%	33.6%	45.4%
P_HISP	14.6%	5.3%	4.5%	4.8%	5.2%	7.6%	20.8%	8.2%	13.0%	44.4%	17.1%	27.3%
P_FEMALE	49.3%	52.7%	53.0%	49.5%	51.5%	49.0%	39.5%	51.8%	43.2%	48.9%	51.3%	51.0%
P_T_017	13.9%	13.5%	9.1%	8.5%	16.4%	9.5%	9.6%	16.8%	8.5%	19.1%	25.4%	20.8%
P_T_1824	8.5%	10.4%	5.5%	8.1%	8.5%	10.3%	11.1%	7.8%	5.0%	10.0%	9.2%	9.5%
P_M_2529	6.0%	5.7%	6.9%	6.4%	4.7%	9.5%	7.8%	3.9%	6.8%	6.4%	4.0%	4.4%
P_M_3039	11.4%	9.5%	14.2%	10.6%	8.8%	14.3%	14.1%	8.4%	18.1%	11.2%	8.6%	8.4%
P_M_4049	8.1%	7.3%	6.6%	7.5%	7.8%	7.5%	10.9%	7.8%	12.0%	7.6%	7.2%	7.3%
P_M_5059	5.6%	5.2%	5.1%	6.2%	5.6%	4.7%	7.1%	6.1%	7.3%	4.6%	4.6%	5.4%
P_M_6069	3.5%	3.5%	3.3%	4.9%	4.1%	2.5%	3.9%	4.0%	3.1%	2.8%	3.1%	3.7%
P_M_70	4.1%	4.5%	4.2%	6.1%	4.9%	3.0%	4.6%	5.6%	2.7%	2.8%	2.9%	3.9%
P_F_2529	5.7%	5.9%	8.3%	6.3%	4.4%	9.2%	5.1%	4.0%	6.4%	5.4%	4.2%	4.0%
P_F_3039	9.3%	9.2%	14.0%	8.8%	8.3%	10.8%	7.4%	7.8%	11.2%	9.3%	8.5%	7.6%
P_F_4049	7.0%	7.7%	6.8%	6.4%	8.1%	5.8%	5.6%	7.6%	6.9%	7.2%	7.5%	7.5%
P_F_5059	5.5%	5.8%	6.0%	5.9%	6.1%	4.3%	4.0%	6.4%	5.2%	5.0%	5.4%	5.9%
P_F_6069	3.9%	4.4%	3.7%	4.9%	4.9%	2.7%	3.0%	4.5%	2.6%	3.5%	3.6%	4.8%
P_F_70	6.2%	6.9%	6.6%	8.7%	7.6%	5.5%	4.7%	9.0%	3.9%	4.3%	4.4%	6.4%
P_OWN_HU	39.3%	34.4%	28.7%	13.3%	60.8%	18.7%	9.8%	59.6%	37.2%	41.6%	53.0%	69.4%
P_FOREGN	33.4%	40.5%	17.7%	43.4%	48.3%	20.9%	38.4%	30.4%	18.0%	43.2%	34.1%	51.1%
P_IM_NZD	19.3%	26.5%	9.7%	23.2%	35.2%	11.1%	16.5%	19.9%	9.4%	20.1%	19.6%	31.6%
P25T_NHS	17.1%	15.4%	4.3%	26.4%	18.8%	9.5%	23.2%	8.7%	6.7%	28.5%	30.3%	29.1%
P25T_HS	13.5%	12.6%	6.8%	10.8%	17.4%	10.9%	15.3%	12.3%	8.5%	17.2%	20.7%	21.1%
P25T_SC	22.7%	22.7%	16.5%	18.7%	24.7%	22.8%	24.7%	25.0%	21.2%	21.1%	26.2%	27.0%
P25T_BAC	29.3%	31.6%	43.0%	28.8%	26.6%	36.1%	25.4%	30.5%	37.2%	22.0%	14.6%	17.2%
P25T_ADV	17.4%	17.7%	29.3%	15.3%	12.6%	20.8%	11.3%	23.6%	26.3%	11.1%	8.2%	5.6%
P16T_NWK	27.4%	28.1%	21.0%	30.8%	30.8%	22.1%	34.3%	30.8%	17.4%	27.0%	32.6%	32.8%
MED_HH_I	62030	59250	91127	45940	63188	54152	35213	82363	72156	56767	51040	57286