Who'd You Vote For?: Predicting Voting Records Using Demographic Data

Ryan Dumas, Aastha Joshi, Lucy Tai, Jaldhir Trivedi

94-832 B1 Business Intelligence and Data Mining with SAS

17 October 2021

PROJECT OVERVIEW

Project Introduction
In determining what data to research, and thus what question to attempt to answer, we were continually drawn back to utilizing local data to answer questions about Pittsburgh. Simply looking back to this past summer, we found a treasure trove of data that we could use to see if the methods we learned in class were perhaps more beneficial/truthful than the methods used by professional journalists. As such, we settled on our ultimate research question: "Can demographic data be used to predict who someone voted for in the 2021 Pittsburgh Mayoral Primary Election?"

Elections
Before diving into our data and methodology, we feel it best to begin with an overview of how Pittsburgh elections work. While the "actual" election for Mayor will be this November, in early summer Pittsburgh held a primary election, in order to select candidates from the major parties who would be on the ballot in November. Pennsylvania rules dictate a closed primary, meaning you can only vote in a primary if you are a registered member of a political party, and then can only vote in that party's primary.[1]

As Pittsburgh is a heavily Democratic city, it is generally assumed that the winner of the Democratic Mayoral Primary will ultimately be elected Mayor. This summer there were two main candidates in the Democratic Primary: Bill Peduto, the two-term/eight-year incumbent, and Ed Gainey, a state Senator and challenger from Wilkinsburg.

The winner of the election is ultimately determined by who gets the most votes in total—regardless of if that vote total is less than 50% of the total population (i.e., it is a "first past the post" election). Votes are typically reported by precinct: the city is divided into wards, and then wards are further subdivided into precincts where people vote. It does not matter to the final results who won each individual precinct—however, precinct-level results are reported, most often for the use of political scientists, statisticians, and data analysts like us.

Ed Gainey ultimately won the Democratic Primary, defeating the incumbent Peduto. Most major news outlets had predicted a Peduto win. Thus, in answering our question, we were not just curious as to whether demography could be a useful predictor for voting patterns, but to see if the resulting analysis would have led us to predict the winner correctly.

Methodology
Our methodology was relatively traditional. Ryan was in charge of initial data collection, and pulled our precinct-level results from the Allegheny County Department of Elections, which collects and stores historic (and verified) election data. He then separated results by precinct, and utilized both Census and ESRI reports to match demographic data to each precinct. Afterwards, Lucy took charge of cleaning the data, and determining the most important variables (in consultation with the team) after performing some key high-level data visualizations in Python.

---

[1] "Voting in Pennsylvania," Ballotpedia, accessed 26 September 2021, https://ballotpedia.org/Voting_in_Pennsylvania.

Afterwards, the team decided on what would be the best types of supervised and unsupervised analyses to run on the data.

Data Overview

Data was split by precinct, leading to 387 individual lines of data. Each line included a Precinct and Ward Number, the percentage of votes for each candidate, the actual number of votes for each candidate, the precinct's zip code, and for each precinct the median age, average household size, Diversity Index, racial majority (and what percentage of the population belonged to said majority race), unemployment rate, median income, average adjusted gross income (or AAGI), the total votes cast in the precincts, and the total population of each precinct.

Data Dictionary

Feature Name	Description
Precinct	The ward and district the observation corresponds to
Edward C. Gainey	Number of votes cast for Gainey in precinct
William Peduto	Number of votes cast for Peduto in precinct
Tony Moreno	Number of votes cast for Moreno in precinct
Michael Thompson	Number of votes cast for Thompson in precinct
Write-in (D)	Proportion of precinct's votes cast for a Democratic write in candidate
Write In (R)	Proportion of precinct's votes cast for a Republican write in candidate
Result (D)	The winning candidate in the precinct (Peduto, Gainey, or Tie)
Zip Code	The zip code corresponding to the precinct
Median Age	Median age of zip code population
Household Size	Average household size of zip code population
Diversity Index	Percentage probability of two randomly selected people from zip code population being two different races (0 to 100)
Population	Number of people living in precinct's zip code
Racial Majority	The largest racial group living in precinct's zip code (Black, White)
Racial Majority - Percentage	The percentage of population that consists of the racial majority
Unemployment Rate	Percentage of population that is unemployed
Median Income	Median income of households in precinct's zip code in USD
AAGI	Average Adjusted Gross Income of households in precinct's zip code in USD
Total Precinct Votes	Total number of votes cast in precinct
Featurized_Majority_percentage	Racial Majority - Percentage with coded positive values to Black majority, negative values to White majority

PYTHON ANALYSIS

Data Combination

Three datasets were ultimately compiled into one for our uses: one dataset showing the winning result and zip code of each precinct, one showing the raw number of votes cast for each candidate in each precinct, and one containing a variety of demographic data by zip code. We joined the first two datasets on precinct and then joined the result to the third data set on zip code. It is worth noting that a single zip code may contain several precincts, which is one limiting factor on our analysis, as demographic data was not available at the precinct granularity.

Exploratory Data Analysis

	count	mean	std	min	25%	50%	75%	max
Write-in (D)	387.0	0.491124	0.009975	0.430769	0.486105	0.494505	0.498615	0.506667
Write In (R)	387.0	0.018706	0.019797	0.000000	0.004386	0.011628	0.028456	0.138462
Median Age	387.0	36.829974	5.810303	24.500000	33.900000	37.400000	39.200000	48.000000
Household Size	387.0	2.124109	0.269996	1.670000	2.060000	2.120000	2.240000	2.560000
Diversity Index	387.0	57.658915	11.133483	33.000000	50.600000	56.300000	67.200000	75.700000
Population	387.0	21504.821705	7680.286339	3294.000000	13851.000000	23350.000000	27895.000000	34580.000000
Racial Majority - Percentage	387.0	70.018863	11.932986	47.940000	61.230000	69.160000	81.670000	92.060000
Unemployment Rate	387.0	5.609561	0.069689	5.600000	5.600000	5.600000	5.600000	6.400000
Median Income	387.0	39129.940568	11227.891523	18422.000000	34290.000000	36724.000000	51079.000000	62959.000000
AAGI	387.0	66585.968992	28421.907805	36160.000000	48780.000000	63550.000000	69710.000000	178290.000000
Total Precinct Votes	387.0	287.793282	148.622200	10.000000	175.000000	270.000000	377.000000	893.000000

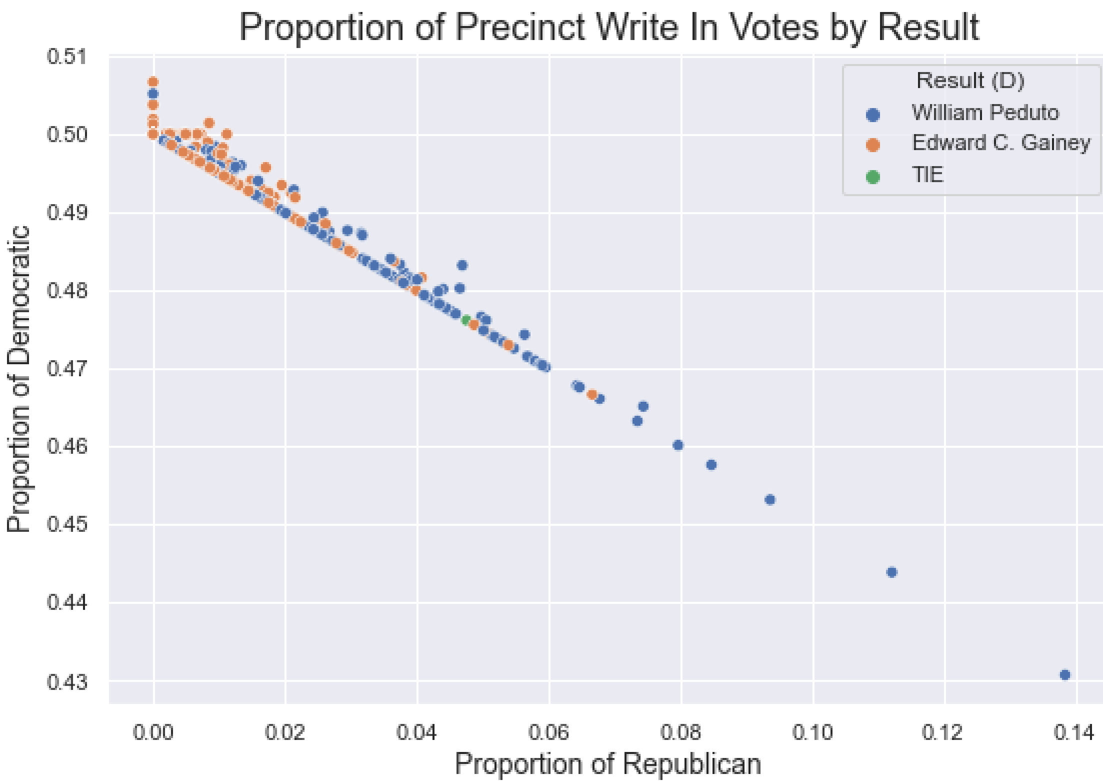
Above is a table containing some summary statistics of numeric features. There were some interesting observations to be made from this, especially combined with some visualizations, which will be discussed subsequently. There were no missing values, and so no values needed to be imputed or dropped. All visualizations included in this section were created in Python with the use of libraries seaborn, matplotlib, plotly, and geojson. A special Python environment needed to be set up to create interactive choropleths. Said choropleths are not included in the code output of the Jupyter notebook files due to the amount of bloat they add to file size, but the code used to generate them is provided. Static versions of the choropleths are included in this report.

Target Variable (Result (D))

Result	Number of Precincts Won	Proportion of Precincts Won
Edward C. Gainey	221	0.571059
William Peduto	164	0.423773
Tie	2	0.005168

The results show that there were only two major contending candidates, Gainey and Peduto. Overall, Gainey won more precincts than any other candidate, 57 more than Peduto who was in second place. The candidates tied in two precincts. Although the precinct ultimately does not matter in determining the winner of the primary election, it is a useful tool in examining if there are regional and demographic differences in voting patterns.

Write In Votes



Scatterplot of Proportion of Precinct Write-In Votes for Each Party, Colored by Winning Candidate.

These were originally the raw number of votes, but were changed to proportions of the total votes cast in a precinct for better comparison between precincts. Generally, about half of the votes from any precinct were write-in votes for an assortment of Democratic candidates. This may indicate widespread, general dissatisfaction throughout Pittsburgh's voting Democrats towards all the major candidates. Additionally, despite this being a Democratic primary, on average about 2% of the votes per precinct were cast for a write-in Republican candidate. From the graph above it is clear that the greater the proportion of Republican write-in votes, the more likely a precinct was to vote for William Peduto. The two proportions also share a very strong linear relationship with a negative correlation— this is somewhat to be expected, as the higher one proportion is, the lower another proportion of the same whole can be. We chose to include these two variables in our models because of this; however, we omitted voting data for specific candidates as those would be strongly collinear with our target variable.

Going forward, we will examine some socioeconomic factors that are known to affect political voting and discuss their importance.

Median Age

Choropleth of the Median Age by Pittsburgh Zip Code.

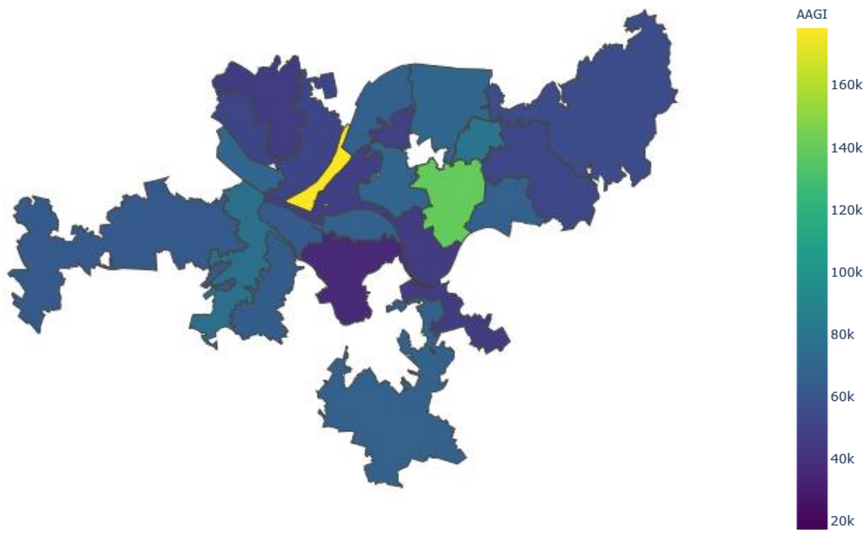
Political typologies are known to differ between generations. Pew research identified several similar voting brackets, two of which were 18-29 and 30-49 (found here). How does this relate to Pittsburgh? The city's population is relatively young, with all precincts having a median age of below 50 years, however it does not differ too much from area to area, with over half of precincts with a median age in the 30s. In the choropleth above, median age is plotted by zip code. There are two areas that have somewhat lower median age in dark purple— this is likely due to a high student population. These zip codes (15213, 15203) correspond to university student-heavy areas that encompass the Carnegie Mellon University and University of Pittsburgh campuses, as well as common districts for off-campus housing. It could be possible that median age factors into precinct voting patterns, but we expect it to be less significant because most precincts fall into the same 30-49 political bracket.

Economic Factors

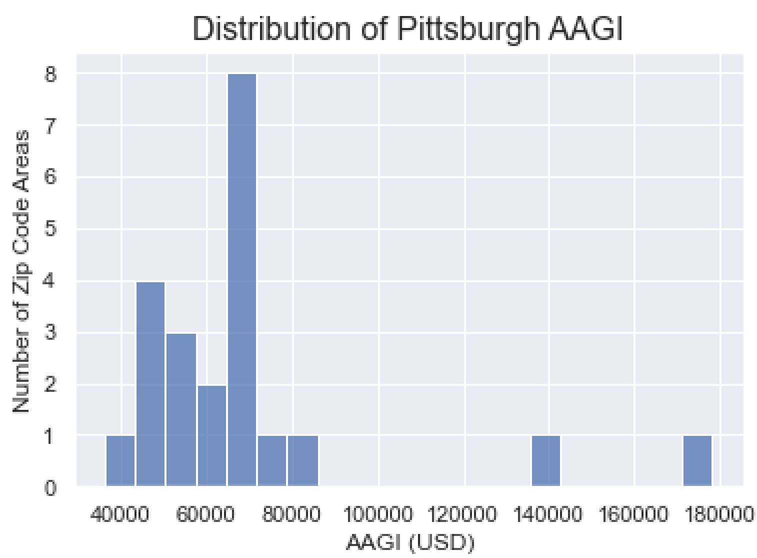
Correlation Heatmap of Unemployment Rate, Median Income, and AAGI.

While we have a variety of economic factors to examine (median income, AAGI, unemployment rate), from the correlation heatmap above it is evident that average adjusted gross income (AAGI) and median income are strongly positively correlated. Given that they both measure income, they were likely to be strongly collinear. We chose to focus on AAGI over median income because it is adjusted for household size, so we felt that it better reflected the economic situation of an area. Additionally, median income was weakly correlated with unemployment rate, so it could be collinear with that as well— whereas AAGI was not correlated with unemployment rate.

AAGI



Choropleth of AAGI (USD) by Pittsburgh Zip Code



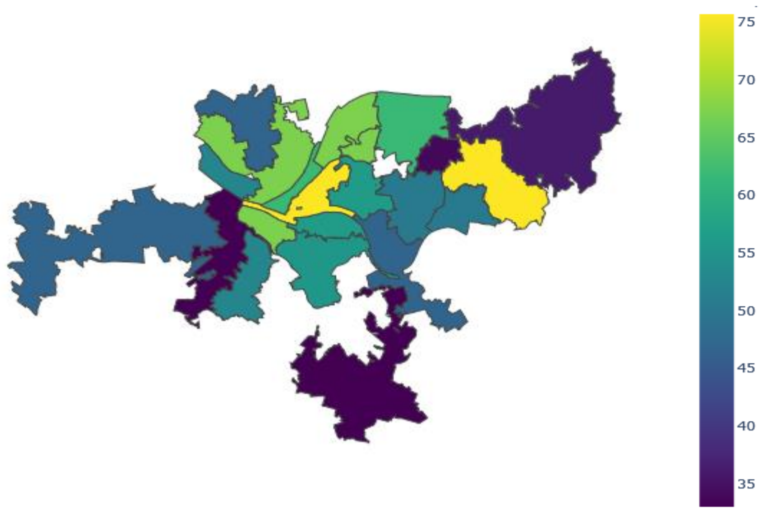
Histogram of AAGI (USD) by Pittsburgh Zip Code.

Overall the distribution of AAGI in Pittsburgh zip codes is very right skewed, with 75% of observations below $70k. Wealth is concentrated in two primary areas, with the highest wealth in Downtown (15222), in which all 3 precincts voted for Peduto. The second most affluent area includes Squirrel Hill (15217) and is less conclusive— 20 precincts voted for Peduto, and 17 for Gainey. The zip code (15210) with lowest AAGI swung more towards Peduto, with 22 Peduto to 14 Gainey precincts. As such, it is difficult to tell how AAGI impacts precinct voting patterns.

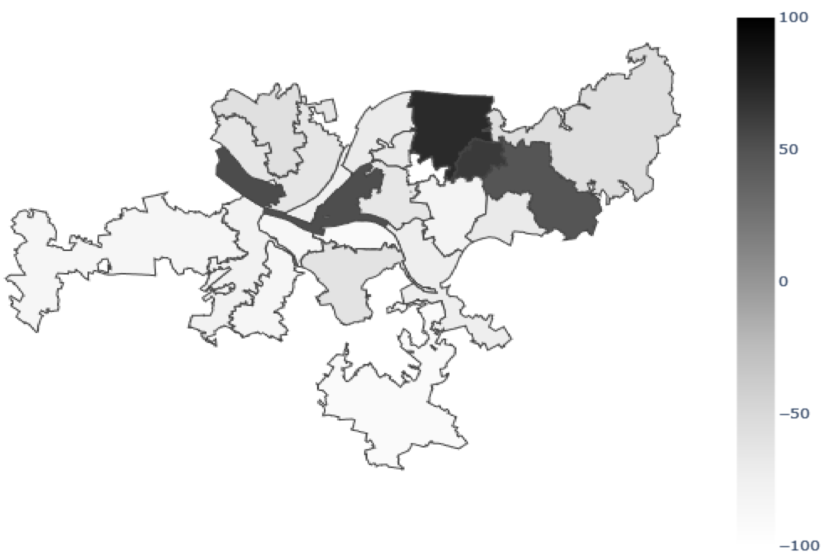
Racial Factors

Our data had three pieces of racial data: diversity index, racial majority, and majority percentage. Similar to other socioeconomic factors, race is also known to affect political leanings. (Some discussion related to race and the 2020 general election found here.) We combined racial majority and majority percentage together so that we had a metric that showed the majority percentage as well as the type of majority it was. As we only had two results for racial majority (White, Black), we arbitrarily coded areas with White majority as negative and Black as positive.

We examine diversity index, the percentage probability of two randomly selected people from a population being two different races, and our featurized racial majority together:

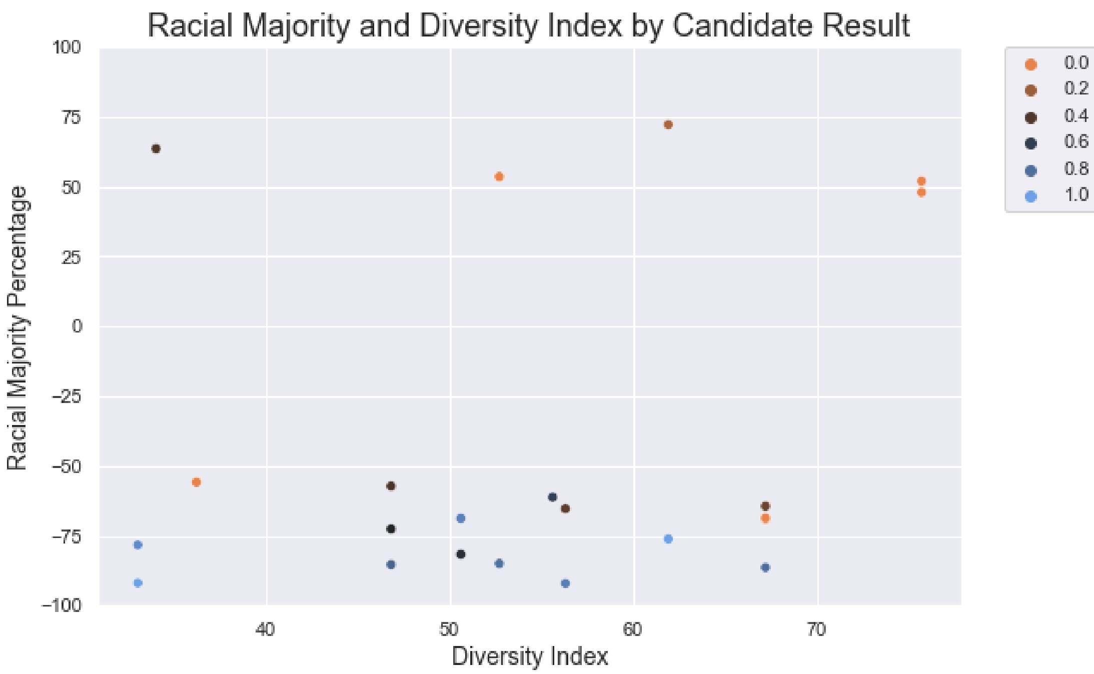


Choropleth of Diversity Index by Pittsburgh Zip Code.



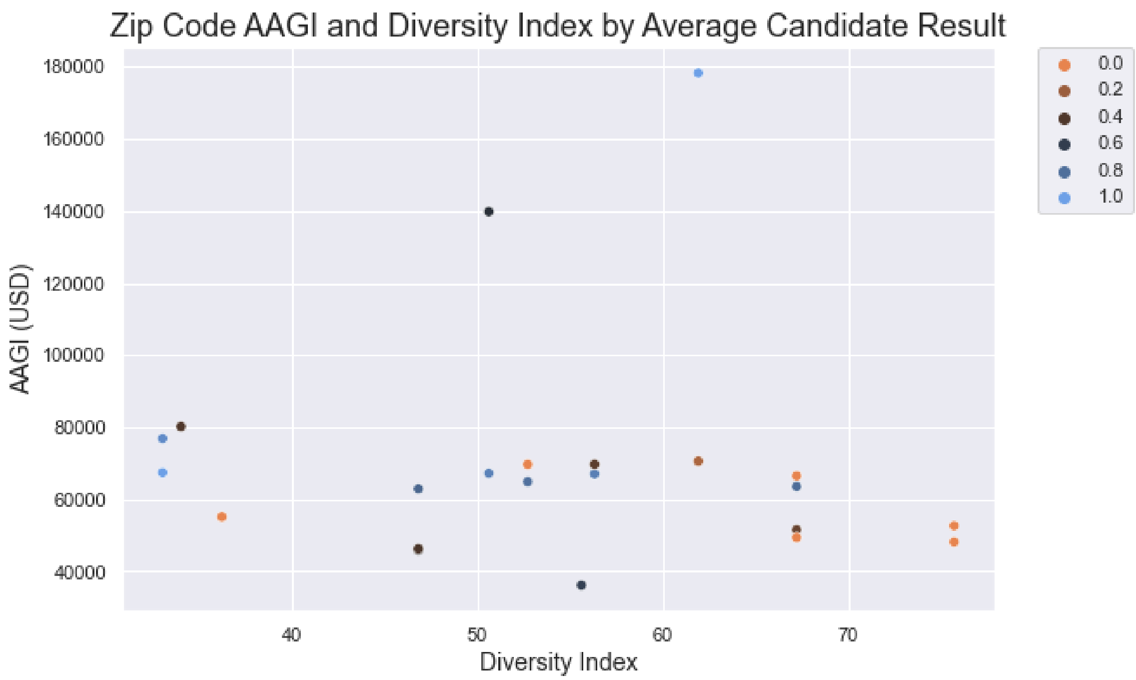
Choropleth of Featurized Racial Majority Percentage by Pittsburgh Zip Code (+ Black, - White).

It's evident that Pittsburgh is primarily White in most zip code districts. Comparing the two choropleths reveals that the more diverse an area, the more likely it is to have a Black racial majority. Similarly, areas that are least diverse tend to be primarily White. The two most affluent districts also both happen to be White majority areas, although they have middling diversity (50-70).



Scatterplot of Zip Code Diversity Index and Racial Majority, Colored by Winning Candidate Scale. (Blue Peduto, Orange Gainey)

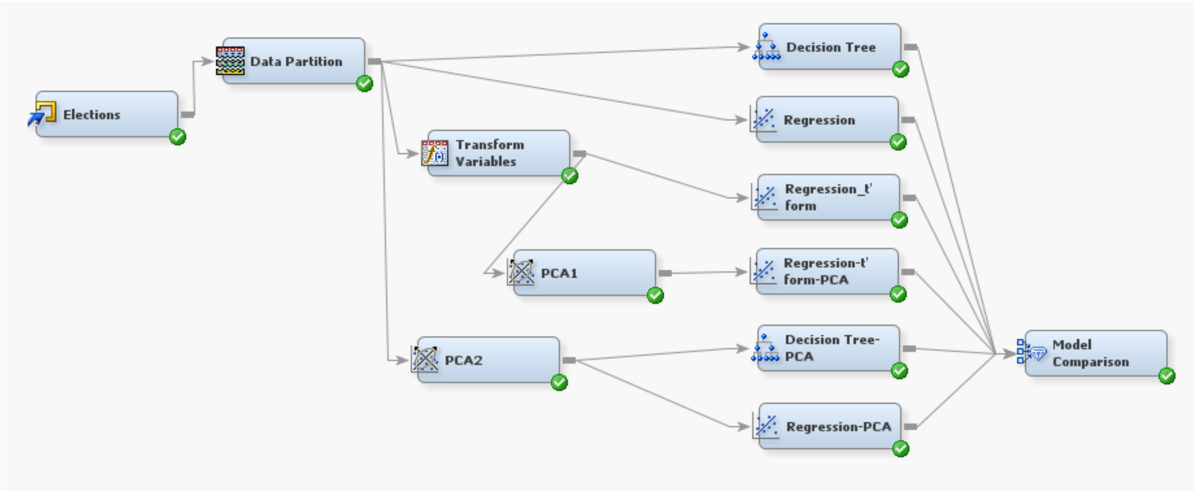
Above, we graph the two features and color on a scale by the average result of a zip code's precincts: 0/blue being Peduto, 1/orange being Gainey, and ties as 0.5/dark. This gives us a little bit of insight on how they affect voting patterns. It seems that areas that have a black majority regardless of diversity index are more likely to have voted for Gainey on average. The opposite is true for white majority areas.



Scatterplot of Zip Code Diversity Index and AAGI, Colored by Winning Candidate Scale. (Blue Peduto, Orange Gainey)

Finally we graph the AAGI against the diversity index and color it by the same candidate scale to see if the combination of this economic and racial feature can give a better idea of their relationship to our target variable. There doesn't seem to be a conclusive relationship between AAGI and diversity in general. However, the graph does suggest that precincts in areas that are better off economically may be more likely to vote for Peduto, and areas that are more diverse are more likely to vote for Gainey. There's a good amount of variability at lower values of AAGI and diversity index, however, but it's probable that these features will be important to our model. After completing our Python EDA, we moved on to SAS for modeling and other analysis.

SAS ANALYSIS



We divide the utility of SAS into 3 parts as shown above diagram.

Data Preparation

Data Import: We import clean data created in Python with no column having no missing values as discussed previously. This means that we will not have to impute values in any feature column. Below figure provides the individual feature details including the type of data and role that they will play in the analysis. We use Result_D as our target label. This column stores who won the precinct among frontrunners (Peduto and Gainey). There have been two precinct which garnered exact same votes to Peduto and Gainey hence they are marked as TIE.

Name	Role	Level
Precinct	ID	Nominal
Racial_Majority	Input	Nominal
Total_Precinct_Votes	Input	Interval
Population	Input	Interval
AAGI	Input	Interval
Write_In__R_	Input	Interval
Write_in__D_	Input	Interval
Unemployment_Rate	Input	Interval
Household_Size	Input	Interval
Diversity_Index	Input	Interval
Featurized_Majority_percentage	Input	Interval
Median_Income	Input	Interval
Median_Age	Input	Interval
Zip_Code	Rejected	Nominal
Edward_C__Gainey	Rejected	Interval
William_Peduto	Rejected	Interval
Michael_Thompson	Rejected	Interval
Racial_Majority___Percentage	Rejected	Interval
Tony_Moreno	Rejected	Interval
Result__D_	Target	Nominal

As is evident from the figure we use 12 features to predict the target variable i.e. Result__D_. Only one of the feature is categorical variable while others are continuous variable. We also use Featurized_Majority_percentage instead of the original Majority_percentage found in the dataset. We derive this by multiplying percentage value by -1 if the majority race is White and keep the same value if the majority race is Black.

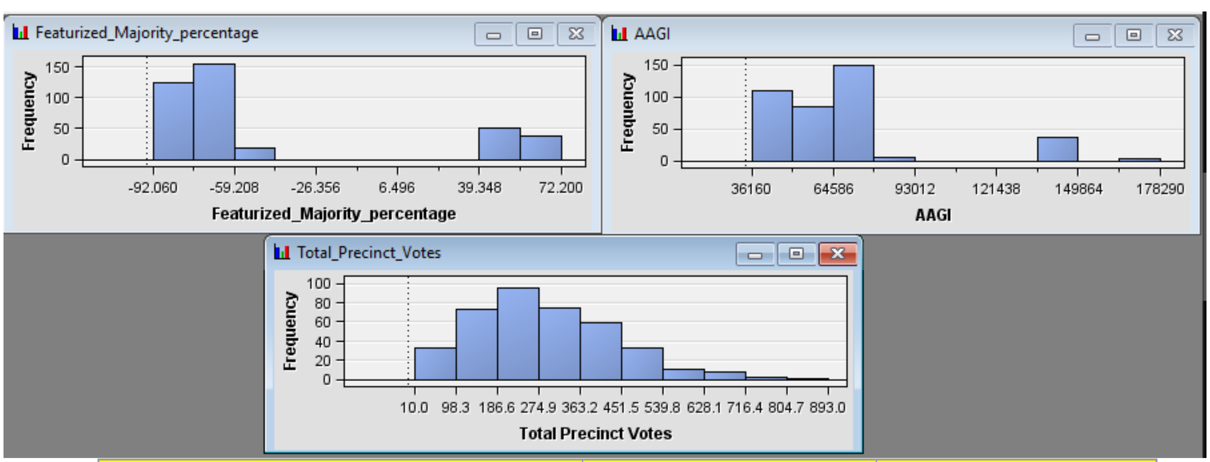
Data Set Allocations	
Training	70.0
Validation	30.0
Test	0.0

Data Partition: We use 70% and 30% partition for creating Training & Validation dataset. Given that we only have around 390 datapoints we decide to not include Test dataset. Furthermore, test dataset is important only when the models require substantial amount of hyperparameter tuning (such as in deep neural networks). Given that going forward we are to only use Decision Trees and Logistic Regression, which don't have too many hyperparameters, we don't expect to experience any major setbacks due to the omission of test dataset.

Pre-processing

For preprocessing we use Transform Variable and Principal Components Analysis cell blocks. However, we create multiple models to understand the individual effect of using or not using these cell blocks.

Transformed Variable: We see that three features have skewness which we can improve using square root transform. This features include Featurized_Majority_percentage, AAGI and Total_Precinct_Votes. The below figure can help us visualize the skewness in the feature distribution. We also provide the skewness measure before and after transformation in the table below.



Variable	Original Skewness	Final Skewness
AAGI	2.02	1.53
Featurized Majority Percentage	1.37	0.98

Total Precinct Votes	**0.51**	**-0.18**

***PCA*:** We also use Principal component analysis to test if we can compress the feature data in an efficient manner in a much smaller dimension than that in the original feature space. PCA is an unsupervised learning method which uses variance maximization to find orthonormal set of principal components. We see that we can compress our 12D data into 7D space of top 7 feature components and still capture 94% of the variance of the original feature space. We understand that there is a tradeoff between compressing the data and incurring loss in classifier models down the line. We would want to check if our models are able to give close enough result without the tradeoff substantially hurting model performance. Below we give you the cumulative variance because of successive addition of principal components:

Component #	Cumulative Variance
1	**37%**
2	**53%**
3	**66%**
4	**76%**
5	**84%**
6	**90%**
7	**94%**

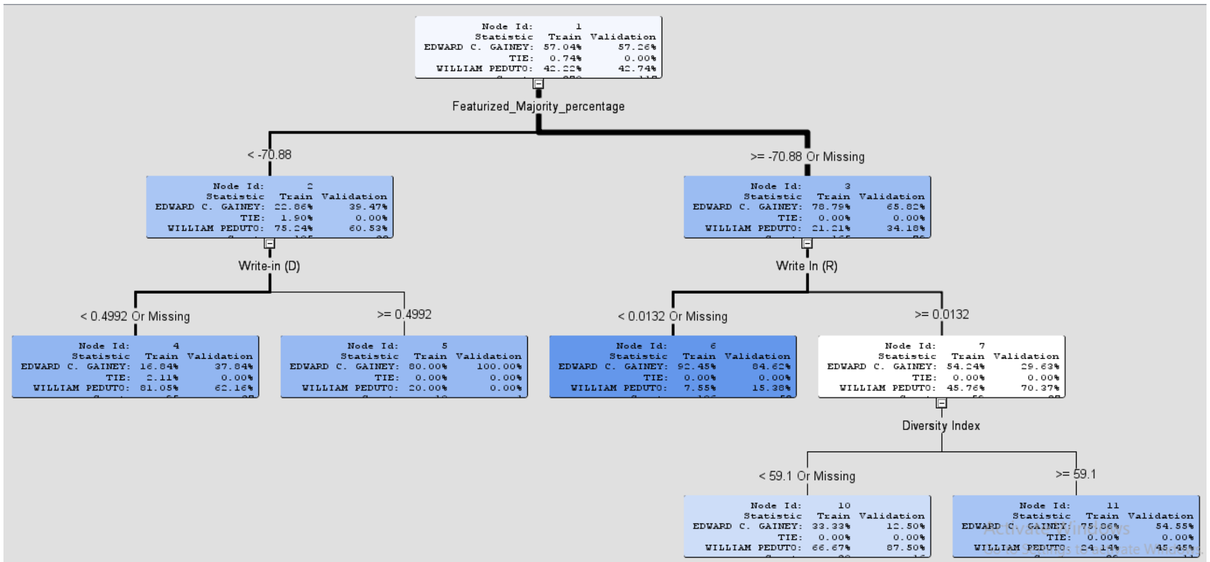
Another downside of using PCA in models such as decision tree is that the model which is otherwise easily interpretable is not so straightforward to infer if Principal components are used instead of original features as model input. This would be detailed later when we discuss Modelling Results.

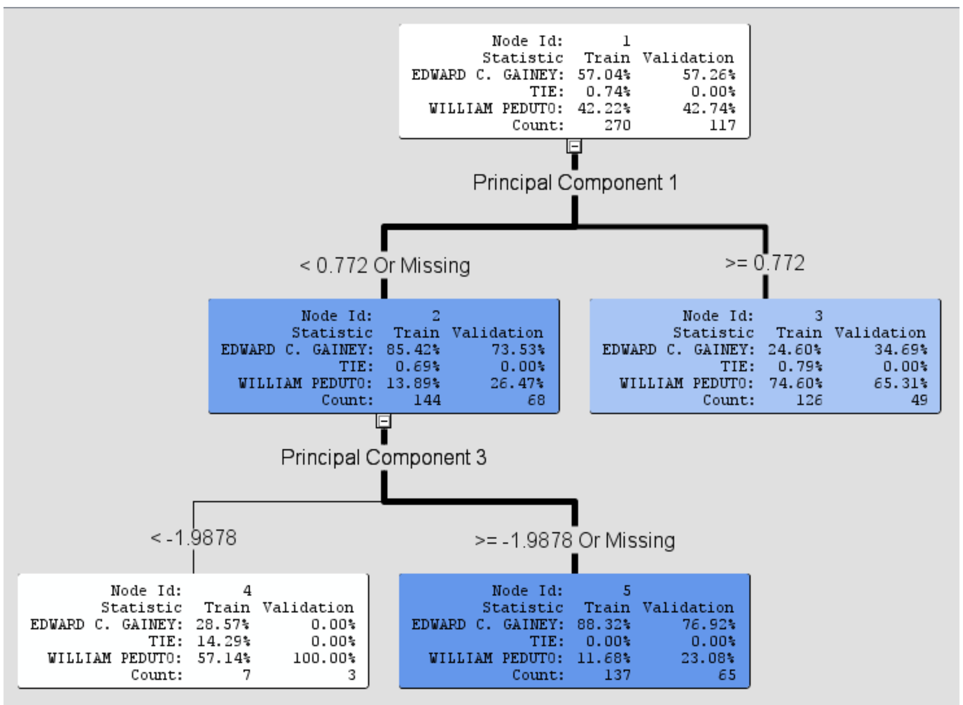
Modelling

We use six different models. These models are either decision tree or logistic regression models. We use 2 decision trees with each accepting original features and top 7 principal components respectively. Furthermore, we use 4 logistic regression models: a) Vanilla Logistic Regression which accepts original untransformed feature space, b) logistic regression with transformed features (the 3 features as discussed in preprocessing), c) logistic regression with principal components of transformed variable and d) logistic regression with principal components on untransformed variables. This setup will allow us to understand 1) which out of decision tree and logistic regression work the best for the problem 2) what are the influence that PCA and transformed variable have on our models.

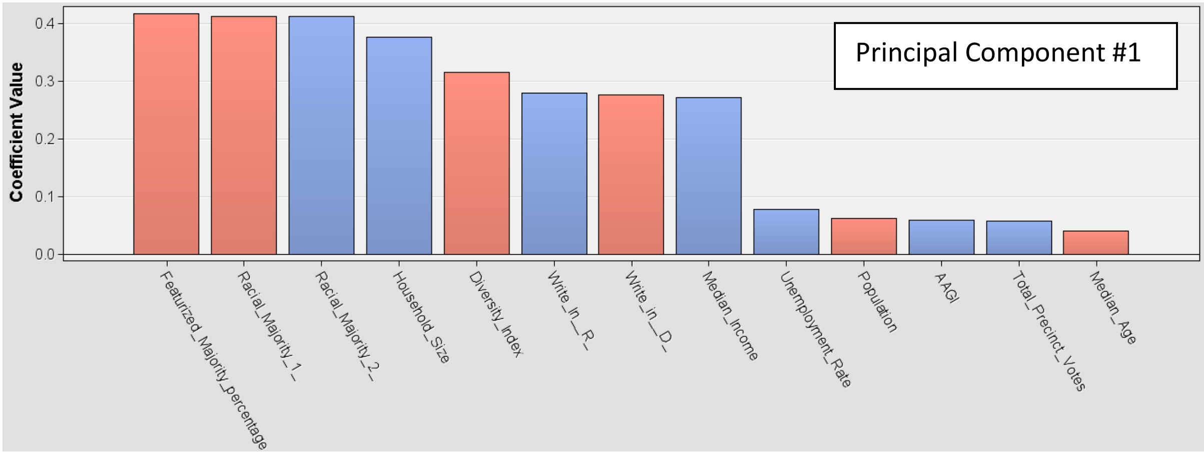
Results

Decision Tree: ###AASTHA, ADD KEY OBSERVATION FOR BOTH DECISION TREES###





As you can see, PCA has made our decision tree simpler in terms of number of leaf nodes and tree depth, but the above decision tree is not readily interpretable because unlike vanilla decision tree the nodes are split not based on original (interpretable) features but based on principal components. To interpret the decision tree, we will first need to understand the weights attached to each of the original feature space which collectively makes up individual Principal components. The below figure gives you the weights to create Principal Component #1. Other principal components would have different set of weights for each of the original feature. Thus, each principal component is linear combination of the original feature space, this ultimately leads to the loss of model interpretability not only in decision tree but also in logistic regression



Logistic Regression: We use model selection to be None thus it will return the results of regression with all variables. We use this to get the best performing model and tackle limiting the usage of unemblematic variables through their diminished weights in Principal components. As discussed earlier we trained four logistic regression model and we discuss the results in the next section:

Result

Model	Validation Error Rate	Train Error Rate
Logistic Regression	23.08%	20.74%
Logistic Regression on PCA	24.79%	19.63%
Decision Tree	24.79%	16.67%
Logistic Regression with Transform	25.64%	20.74%
L. Regression with T'form and PCA	26.50%	22.96%
Decision Tree with PCA	27.35%	18.89%

We use misclassification rate for our error rate calculation. We see that the Logistic regression works best on validation data set with 23% misclassification. The baseline model which predicts same winner (Gainey) for every precinct will have 47% misclassification rate. Thus, we show that our model can harness intelligence to an extent to predict the precinct winner with less than half misclassification rate as compared to baseline model.

We also observe that PCA has worked particularly well with logistic regression. It is to be noted that even after using just 7 input features the model performance does not faulter. The logistic regression with PCA still gives a good error rate of 24.79% as opposed to 23.1% for vanilla logistic regression with increase in error rate of just 1.7%.

We observe that transforming original 3 variable through square root has not helped model accuracy. Logistic regression with transformed variable has 2.6% higher error rate than the vanilla logistic regression. This can be attributed to a couple of reasons:

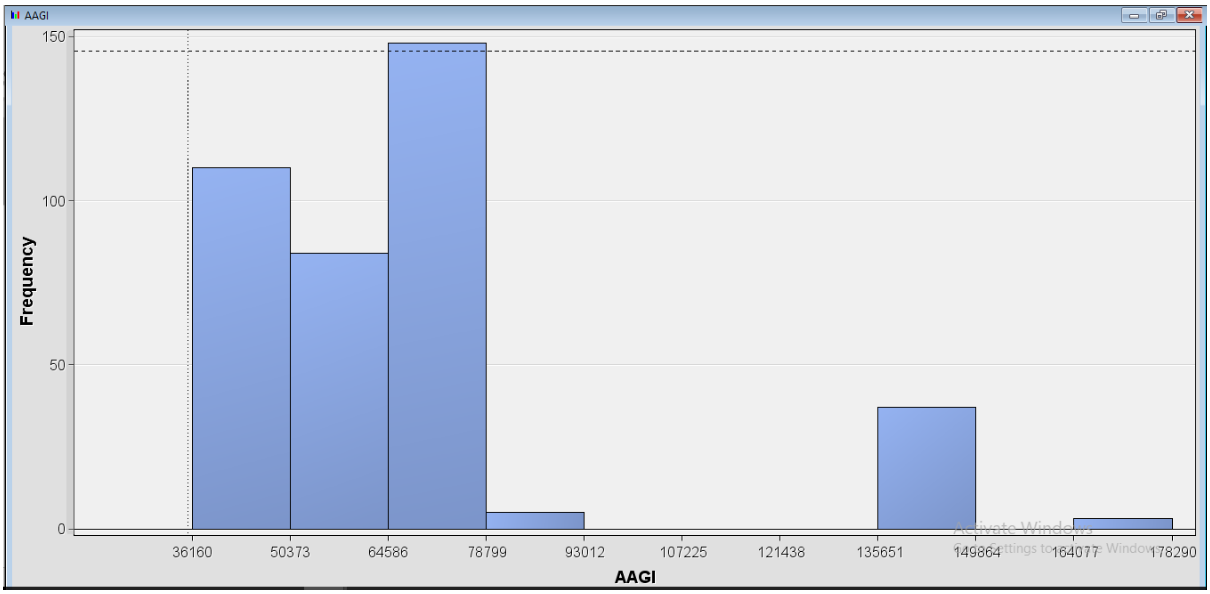
1. The number of datapoints are very low for skewness to be a problem in the distribution

2. The original variables are indeed a better predictor of precinct winner than their square root
3. By taking a square root we decrease the skewness, but we also decrease the variance of the distribution. This is not good as finding correct balance of weights across different features requires even variance across all the features.

SAS Takeaways

AAGI

- Clear clusters of Top 10% and bottom 90%
- 40 Precinct AAGI>113500
- 347 Precinct AAGI<93012



Conclusion: Population Sequestered in precincts based on economic class

The AAGI in Black Majority precinct ranges from $48k to $80k but the same in White majority precinct ranges from $36k to $178k. The average AAGI in White precinct is $68k while the same in Black precinct is $60k.

Racial Majority Percentage: The Majority percentage in Black majority precincts (88 precincts) ranges from 48% to 72% (Avg. 59%) but that in white majority precincts (301 precincts) ranges from 56% to 92% (Avg. 73%).

Conclusion: The black majority precincts are much more diverse precincts while many white majority precincts tend to be much more exclusive and monolithically white precincts.

Additional Observations:

Ed Gainey won the primary by a margin of 3848 votes. Ed Gainey carried 81 out of 88 black majority precincts while he was able to carry 142 out 301 white majority precincts. Even though black majority precincts are fewer in number, they have proved to be key decision precincts as in white majority precincts the margin between Peduto and Gainey was just 414 votes (18203 votes to Gainey Vs 17789 votes to Peduto). The bulk of the remaining margin of 3434 votes comes from Black majority precincts.

There were 54695 write in (Democrat) votes as compared to votes casted to all the 4 candidates combined who managed to garner 54583 votes. Curiously, there are only 19 precincts where write in (democrats) votes are not more than combined votes of Peduto and Gainey. 18 out of the 19 precincts were carried by Gainey.

OVERALL ANALYSIS

Exploratory

Ed Gainey won the 2021 Mayoral Primary by 3,848 votes. When we drill down further, of the 88 Black Majority precincts Gainey won 81 of them, or roughly 92%. Of the 301 White Majority precincts, Peduto won 159, or roughly 53%. While this clearly shows that Peduto won a higher number of White Majority precincts than Gainey did, it is to a lesser degree than Gainey's higher number of wins in Black Majority precincts (i.e. 53% < 92%).

Additionally, of Gainey's 3,848 vote margin of victory, 414 came from White Majority precincts. The rest – 3,434 votes – came from Black Majority precincts.

Summary Findings

1. Most of Gainey's margin of victory of 3,848 votes came from precincts with a majority Black population.
2. Precincts with a higher Diversity Index had a higher likelihood of being won by Gainey than Peduto.
3. Precincts with a higher Average Adjusted Gross Income (AAGI) had a higher likelihood of being won by Peduto.

Conclusion

"Can demographic data be used to predict who someone voted for in the 2021 Pittsburgh Mayoral Primary Election?" Our analysis shows that the demographic parameters closely correlate with the electorate's voting behaviour. However, correlation cannot be mistaken for causation. The features that we used are strong indicators of how a person may vote but it may not be termed as a cause for him to vote for a certain candidate.

We showed that a person who lives in a precinct with a high Diversity Index, low to average AAGI, and a majority Black population was statistically most likely to vote for Ed Gainey. Similarly, a person who lives in a precinct with a low Diversity Index, high AAGI, and a majority White population was most likely to vote for Bill Peduto. Our models reflect this intelligence, it gets validation to an extent through the initial exploratory analysis as well as the actual outcomes of the election itself.

THIS PROJECT: WHAT WE LEARNED

SAS EM is a powerful tool and very convenient in many ways. For instance, it's very easy to transform variables and impute missing values in SAS. However, other functions are much less straightforward. As a group we found it much easier to use Google sheets and Python to perform data combination, featurization and cleaning. It also eliminated the need to log on to the virtual machine and stress about constantly saving our progress.

We also opted to make most of our data visualizations in Python because we were able to create graphics that were much more attractive than SAS EM's. Python also included flexibility and extra features, allowing us to create choropleths without needing latitude/longitude coordinates, and gave us better control over how our visualizations looked and what data they communicated. Because Python is open source, powerful libraries have been created by the community which can be utilized to get great visualization. SAS on the other hand is a proprietary software with a limited number of features developed by a limited number of developers.

SAS has great advantages over conventional ML python libraries and is much more accessible to professionals than python is. The interactive user interface provides greater flexibility in terms of changing models, comparing models & tuning parameters. However, it is difficult to deploy machine learning models which require extraction of model parameters in a structured manner which is not fully integrable in SAS. Simply put, SAS is able to create the machine learning pipeline for the users in a fraction of time. The users are able to train and test their models and see the results, but to deploy models for real life applications may be a little difficult to do in SAS.

Working on a project remotely with other students was also an interesting experience. Our group mainly met through Zoom, and because of this it was easy to reschedule meetings to fit all four of our schedules. It was also easy to stay on the same page schedule and work-wise, because we summarized what we talked about in a digital group chat. Through the project work we were able to work on a real life problem that can be solved through data analysis and machine learning. We saw the challenges behind combining and cleaning a data set, visualization of dataset through multiple tools (both in python and SAS). Finally modelling the dataset using various models and preprocessing techniques.

Bibliography

“Council District Map.” Department of City Planning, City of Pittsburgh. Accessed 28 September 2021. https://apps.pittsburghpa.gov/council/Voting_District_Changes_-_Preliminary_RAC_Map_June_2012.pdf.

“Detail XLS.” ClarityElections.com. Allegheny County, PA. Accessed 26 September 2021. https://results.enr.clarityelections.com//PA/Allegheny/109361/277443/reports/detailxls.zip.

“May 18, 2021 Primary Election.” ClarityElections.com. Allegheny County, PA. Accessed 26 September 2021. https://results.enr.clarityelections.com/PA/Allegheny/109361/web.276935/#/summary.

“Pittsburgh City Council Precincts.” City of Pittsburgh. Accessed 28 September 2021. https://apps.pittsburghpa.gov/redtail/images/13614_2021_Council_Districts___Wards_Revised.pdf.

“Pittsburgh Wards.” Maps.ARCGIS.com. City of Pittsburgh. Accessed 28 September 2021. https://pittsburghpa.maps.arcgis.com/apps/OnePane/basicviewer/index.html?appid=2a57c4fbe92248e38a57220d20f23ae2.

“Primary and General Election Results.” Allegheny County. Accessed 26 September 2021. https://www.alleghenycounty.us/elections/election-results.aspx.

“Tapestry Segmentation.” ESRI. Accessed 28 September 2021. https://www.esri.com/en-us/arcgis/products/data/data-portfolio/tapestry-segmentation.

“QuickFacts: Pittsburgh city, Pennsylvania.” Census.gov. US Census Bureau. Accessed 28 September 2021. https://www.census.gov/quickfacts/pittsburghcitypennsylvania.

“Voting in Pennsylvania.” Ballotpedia. Accessed 26 September 2021. https://ballotpedia.org/Voting_in_Pennsylvania.