A Logistic Regression Model of Voter Turnout for the 2024 Presidential Election in Hidalgo County, Texas

John Knight
The University of Texas Rio Grande Valley
MATH-6364 Statistical Methods

December 9, 2024

Abstract

Hidalgo County is a predominantly Hispanic county in the Rio Grande Valley area of Texas with historically below-average democratic participation. The study aimed to analyze factors associated with higher or lower voter turnout in Hidalgo County in the 2024 presidential election. Demographic variables, including age, income, education level, and ethnicity were obtained for each census tract from the most recent American Community Survey estimates and matched to publicly available voter records using a geocoding API.

In a univariate linear correlation analysis, age, income, and education level were positively associated with voter turnout, while the proportion of Hispanic residents showed a negative correlation. A binomial logistic regression model was constructed using the number of votes cast as "successes" and non-voters as "failures" for each tract. In the multiple logistic regression model, average age, income, education level, and Hispanic resident proportion were all significantly associated with increased voter turnout ($p < 0.05$).

However, when pairwise interaction terms were introduced to the model, the main effect terms all became negative, with a mixture of positive and negative interaction terms. The study shows that although areas with a high Hispanic population may have a lower voter turnout, this may be driven by the fact Hispanic populations are often lower in educational and income measures. The significant interaction terms underscore the nuanced relationships between demographic characteristics and voter turnout.

1 Introduction

1.1 Background

The 2024 United States presidential election took place on November 5, 2024. Along with choosing the next president, voters could elect members of congress and other public offices, as well as voting on a range of propositions. In Hidalgo County, home of The University of Texas Rio Grande Valley, early voting was also available from October 21 to November 1.

Hidalgo County is a predominantly Hispanic area with low income and education levels, and has historically had a lower turnout for presidential elections than the national average, as illustrated in Figure 1. Previous research has found that demographic factors such as age and education level can be predictive of voter turnout.[1] The objective of this study was to use statistical analysis to identify which demographic variables were associated with higher or lower voter turnout in Hidalgo County.

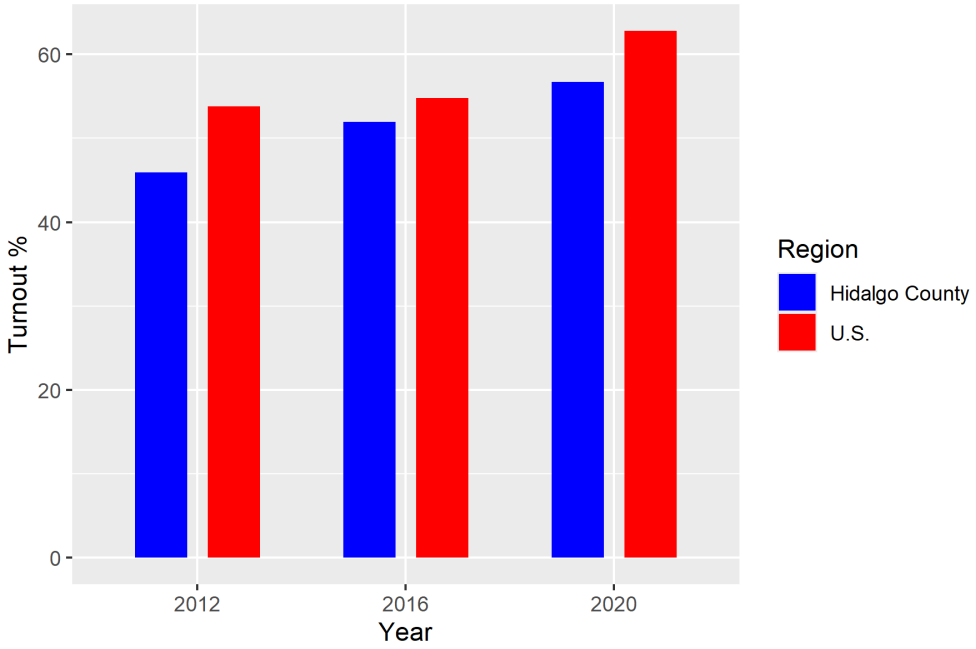

Figure 1: Hidalgo County turnout versus national average in 2012-2020 presidential elections.

1.2 Data

By Texas law, counties must release the names and addresses of everyone who has voted, and this data is freely available to download from the Hidalgo County website.[2] Who each person voted for remains confidential. According to this data at time of reporting, a total of 159,324 people voted in early voting, and a further 52,780 voted on election day.

While the characteristics of each individual are not known, census data can be used to associate each voter with the demographic data of the census tract they live in. The R library `tidycensus` provides a free API that allows users to download data from the most recent American Community

Survey (ACS) to get census estimates for geographic areas such as census tracts, which are small, relatively permanent statistical subdivisions of a county, averaging about 4,000 inhabitants each.[3]

After gathering the voter addresses and census tract data, each voter was mapped to the census tract in which they live. This was done by first geocoding their address into latitude and longitude coordinates with the Google geocoding API, using the `ggmap` library in R. A small number of addresses (0.1%) were unable to be geocoded, which was unlikely to be enough to have a significant effect on the results. Then, using the `tigris` package, these coordinates were linked to the appropriate census tract. The resulting dataset had 211 rows (one for each census tract in Hidalgo County) and 8 variables, as listed in Table 1.

Table 1: Summary of variables and their descriptions.

Variable	**Description**
`name`	The name of the census tract.
`citizens_over_18`	Estimated number of U.S. citizens aged 18 and over.
`median_age`	Estimated median age.
`median_income`	Estimated median income.
`bachelors_degree_pct`	Estimated % of residents over 25 with at least a bachelor's degree.
`hispanic_pct`	Estimated % of residents who are Hispanic.
`voter_count`	Number of residents who voted in the election.
`voter_pct`	`voter_count` / `citizens_over_18` $*$ 100.

2 Methodology

First, the data were inspected for outliers. As a result, three census tracts were removed from the dataset, as they were considered outliers that may skew the results. Census tract 204.09, circled in red in Figure 2, has a much higher average age and a much lower proportion of Hispanic residents than the other tracts. This area is populated by "winter Texans" (residents who live in northern states during the summer, then migrate south for the winter) who may not be reflective of the local population.

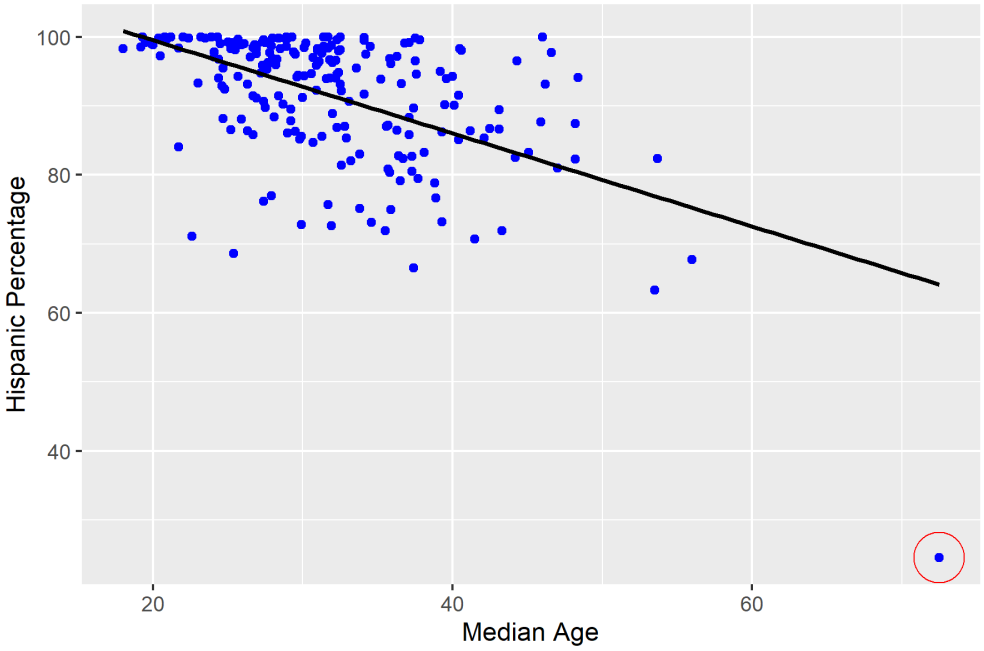

Figure 2: Scatter plot comparing `hispanic_pct` and `median_age` for each census tract in Hidalgo County. Census tract 204.09 highlighted in red circle.

Additionally, the ACS population estimate for tract 241.31 was too low due to the fast development of the Tres Lagos master-planned community in that area, meaning the values for `voter_pct` would be artificially high. And finally, tract 235.30 contains a large prison; since prisoners appear on the census but are unable to vote, results from this tract would not be meaningful. The final dataset had 208 rows (each representing a census tract) and 8 variables. A heat map, correlation matrix, and descriptive statistics table were produced for this final dataset.

Because the response variable (`voter_pct`) was bounded by the interval [0, 100], linear regression was not deemed appropriate for this model, since it could result in predictions less than 0% or greater than 100%. Instead, the number of votes cast and the number of non-voters for each census tract were passed as ”successes” and ”failures” into a binomial logistic regression model.

Missing values were found in 5 rows of `median_income`. These values appeared to be missing at random, and were imputed using the MICE method (Multiple Imputation by Chained Equation) which estimates the value of median income based on the other features in the dataset.

After fitting the logistic regression model, variables were inspected for potential multicollinearity via their Variance Inflation Factor (VIF). The model was also assessed for goodness of fit by running a Hosmer-Lemeshow test, and a binned residual plot was created to highlight any unusual distribution of the residuals. A second logistic regression model was then fit using additional pairwise interaction terms, and the model's performance was compared to the simpler model using the Akaike Information Criterion (AIC). A similar round of diagnostic checks was performed on the second model to ensure its appropriateness.

A significance level of $\alpha = 0.05$ was used for all statistical tests. R software version 4.2.0 was

used for all analysis.

3 Data Analysis

Descriptive statistics can be seen in Table 2. Compared to the national average, Hidalgo County residents are younger (mean age 31.8; national average 38.5), with a lower median income (mean income $52,435; national average $75,149), less likely to have a bachelor's degree (mean 19.92%; national average 34.3%), and more likely to be Hispanic (mean 91.81%; national average 18.7%).

Table 2: Descriptive statistics for Hidalgo County census tracts.

Variable	n	Mean	SD	Min	Max
median_age	208	31.80	7.13	18.00	56.00
median_income	203	52435.22	21713.85	12808.00	145129.00
bachelors_degree_pct	208	19.92	13.71	0.00	59.93
hispanic_pct	208	91.81	8.40	63.30	100.00
voter_pct	208	47.53	13.55	18.33	95.96

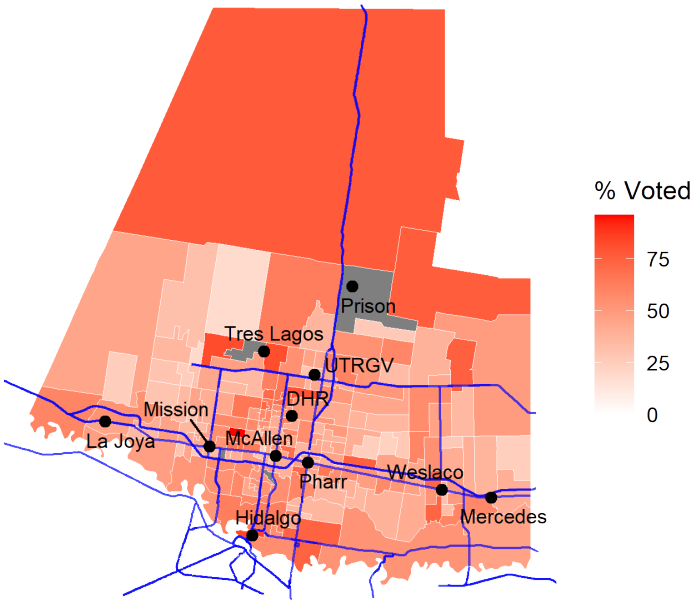

Figure 3: Heat map showing `voter_pct` for each census tract. Outlier tracts that were removed from the data are show in gray.

A histogram showing the distribution of `voter_pct` for all census tracts can be found in the appendix. This is also visualized in the heat map in Figure 3. The highest turnout was 96.0% in census tract 203.03, an affluent area of Mission in the Sharyland school district. The lowest turnout was 18.3% in census tract 241.20, a poor, rural area north of the city of Alton.

The correlation matrix (Figure 4) shows a strong positive correlation (0.69) between `bachelors_degree_pct` and `median_income`. There is a moderate negative correlation between `hispanic_pct` and both `median_age` (-0.46) and `bachelors_degree_pct` (-0.49). There is also a moderate positive correlation between `voter_pct` and both `median_income` (0.46) and `bachelors_degree_pct` (0.44).

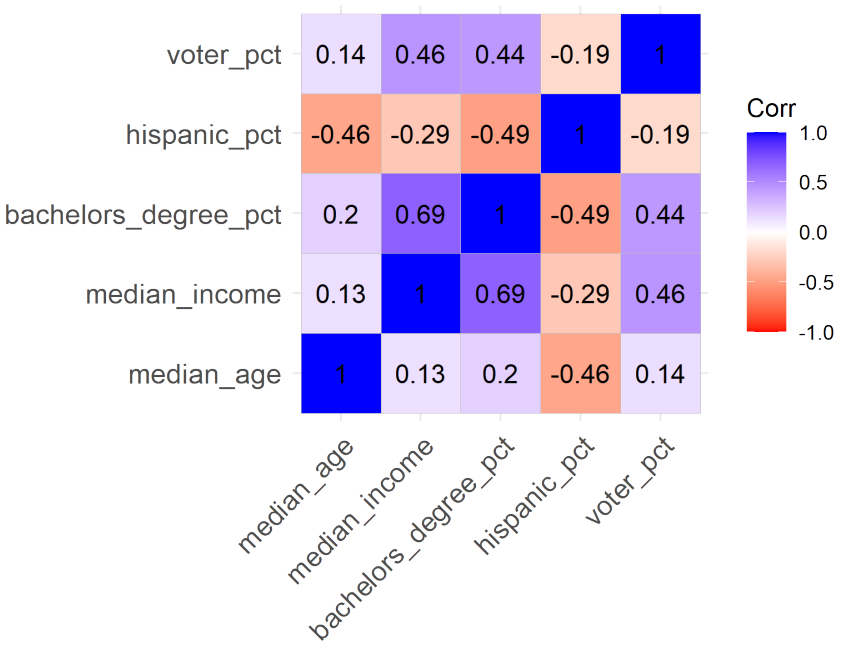

Figure 4: Correlation matrix.

The results of the simple logistic regression model can be seen in Table 3. `median_income`, `median_age`, `bachelors_degree_pct`, and `hispanic_pct` all had positive coefficients that were found to be statistically significant. This means that an increase in any of these four variables, holding other variables constant, would be expected to increase voter percentage. The Akaike Information Criterion score for this model was 22,962.

Table 3: Results of logistic regression.

Variable	Coefficient	p-value
median_income	7.783 e-06	< 0.001
median_age	1.036 e-02	< 0.001
bachelors_degree_pct	9.375 e-03	< 0.001
hispanic_pct	4.285 e-03	< 0.001

A variance inflation factor (VIF) of 5 or greater is a concern for multicollinearity, but all VIFs were below this range. The binned residual plot showed residuals to be evenly scattered around the mean with no obvious pattern. The table of VIFs and the binned residual plot can be found in the

appendix. A Hosmer-Lemeshow goodness of fit test returned $\chi^2 = 0.3109$ on 8 degrees of freedom with a p-value close to 1, indicating that the model was a good fit for the data.

Table 4 displays the results of the more complex logistic regression model involving pairwise interactions. The coefficients for `median_income`, `median_age`, `bachelors_degree_pct`, and `hispanic_pct`, which were all positive in the simple model, are all negative in this model. All are statistically significant with the exception of `median_income` (p = 0.440).

The pairwise interaction terms are all statistically significant, with a mixture of positive and negative coefficients. For example, `median_income` and `median_age` have a positive interaction, meaning as both variables increase simultaneously, voter turnout increases more than would be expected by the increase in the two variables independently. However, `median_age` and `bachelors_degree_pct` have a negative coefficient, meaning as both variables increase simultaneously, voter turnout is lower than expected.

Although VIF is not appropriate for a model involving interaction terms, a binned residual plot showed no signs of concern and can be seen in the appendix. A Hosmer-Lemeshow test returned $\chi^2 = 0.6259$ on 8 degrees of freedom with a p-value close to 1, again indicating that the model was a good fit for the data. The AIC for this model was 22,131, suggesting that it is an improvement on the simpler model.

Table 4: Results of logistic regression with interaction terms.

Variable	Coefficient	p-value
median_income	-2.401 e-06	0.440
median_age	-5.202 e-02	< 0.001
bachelors_degree_pct	-4.239 e-02	< 0.001
hispanic_pct	-2.535 e-02	< 0.001
median_income:median_age	4.406 e-07	< 0.001
median_income:bachelors_degree_pct	1.579 e-07	< 0.001
median_income:hispanic_pct	-8.925 e-08	0.002
median_age:bachelors_degree_pct	-6.626 e-04	< 0.001
median_age:hispanic_pct	5.548 e-04	< 0.001
bachelors_degree_pct:hispanic_pct	7.000 e-04	< 0.001

4 Conclusions

The superior performance of the model with interactions demonstrates that the relationship between independent variables can be nuanced. The coefficients of a simple regression model are interpreted as ”while holding other variables constant” - and yet, in reality, other variables do not remain constant.

As an illustrative example shown in Table 5, the voting percentage for four hypothetical census tracts A, B, C, and D were predicted using the model with interactions. Tract A, representative of the typical census tract in Hidalgo County, was predicted a turnout of 47.4%. Tract B, with a lower

median income, age and education level but a higher Hispanic population, was predicted to have a turnout of 38.7%. Tract C, with a higher median income, age and education level but a lower Hispanic population, was associated with a turnout of 55.1%. Tract D, an area representative of young, educated Hispanic professionals, has the highest prediction of all at 64.1%.

Table 5: Predictions for hypothetical census tracts using the model with interaction terms.

Tract	Median Income	Median Age	Degree %	Hispanic %	**Prediction**
A	52000	32	20	92	**47.4**%
B	30000	25	5	98	**38.7**%
C	100000	45	60	70	**55.1**%
D	80000	30	60	90	**64.1**%

It remains apparent that U.S. citizens in certain demographic areas are predictably less likely to participate in the democratic process. However, while a univariate analysis shows lower voter turnout in areas with a higher Hispanic population, this relationship is not true once demographic factors like income and education are accounted for. Rather than Hispanic Americans being inherently less interested in the democratic process, it may simply be the case that they are, on average, more poorly educated and with a less wealthy, especially in areas such as Hidalgo County.

A possible weakness of this study is that, since demographics are aggregated to the census tract level, the results are not necessarily reflective of individual voters. It should also be noted that postal ballots were not included in this study, because the identities of postal voters were not published at the time of the study. This could potentially bias the results against demographics with a higher tendency to use postal ballots. One potential area of future interest would be to compare population demographics with the tendency to vote for Donald Trump or Kamala Harris, although this would be more difficult as voting precincts do not correlate with census tracts.

References

[1] John G. Matsusaka and Filip Palda. "Voter Turnout: How Much Can We Explain?" In: *Public Choice* 98 (1999). Accepted 11 March 1997, pp. 431–446.

[2] Hidalgo County, Texas. *Unofficial Early Voting Totals and Rosters*. 2024. URL: https://www.hidalgocounty.us/3339/Unofficial-Early-Voting-Totals-Rosters.

[3] U.S. Census Bureau. *A Guide to State and Local Census Geography: Census Tracts*. 2024. URL: https://www2.census.gov/geo/pdfs/education/CensusTracts.pdf.

Appendix

.1 Additional Figures and Tables

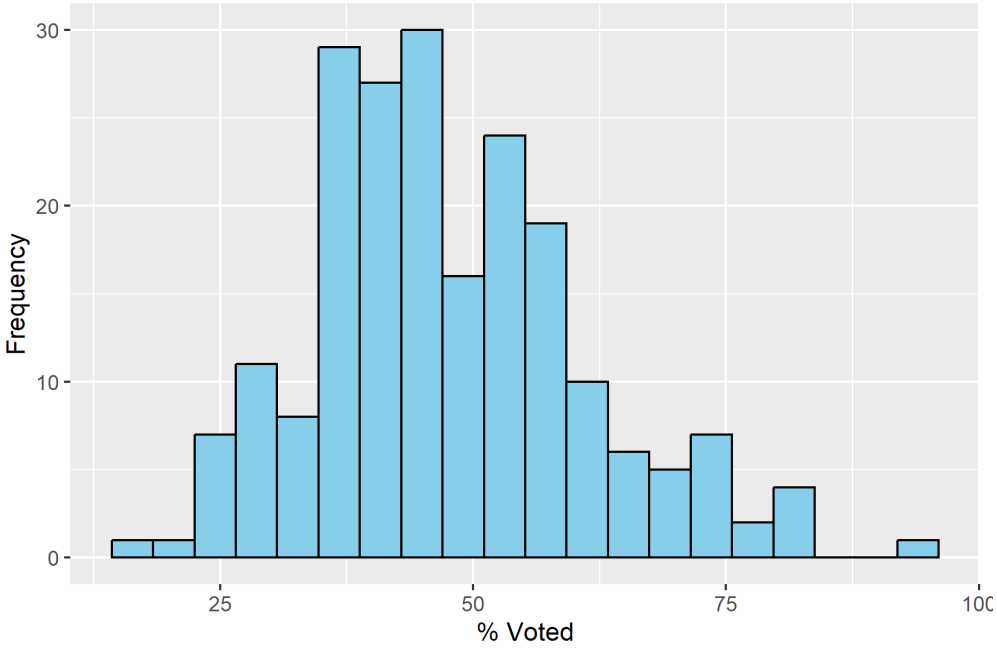

Figure 5: Histogram of voter_pct for Hidalgo County census tracts.

```
Call:
glm(formula = cbind(votes_cast, non_voters) ~ median_income +
    median_age + bachelors_degree_pct + hispanic_pct, family = binomial,
    data = hidalgo_main)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -1.448e+00  5.429e-02 -26.680   <2e-16 ***
median_income        7.783e-06  1.982e-07  39.273   <2e-16 ***
median_age           1.036e-02  5.020e-04  20.637   <2e-16 ***
bachelors_degree_pct 9.375e-03  3.366e-04  27.854   <2e-16 ***
hispanic_pct         4.285e-03  4.767e-04   8.988   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 29236  on 207  degrees of freedom
Residual deviance: 21300  on 203  degrees of freedom
AIC: 22962

Number of Fisher Scoring iterations: 4
```

Figure 6: Results of logistic regression model.

Table 6: Variance inflation factors for independent variables.

Variable	VIF
median_income	1.817
median_age	1.177
bachelors_degree_pct	2.339
hispanic_pct	1.631

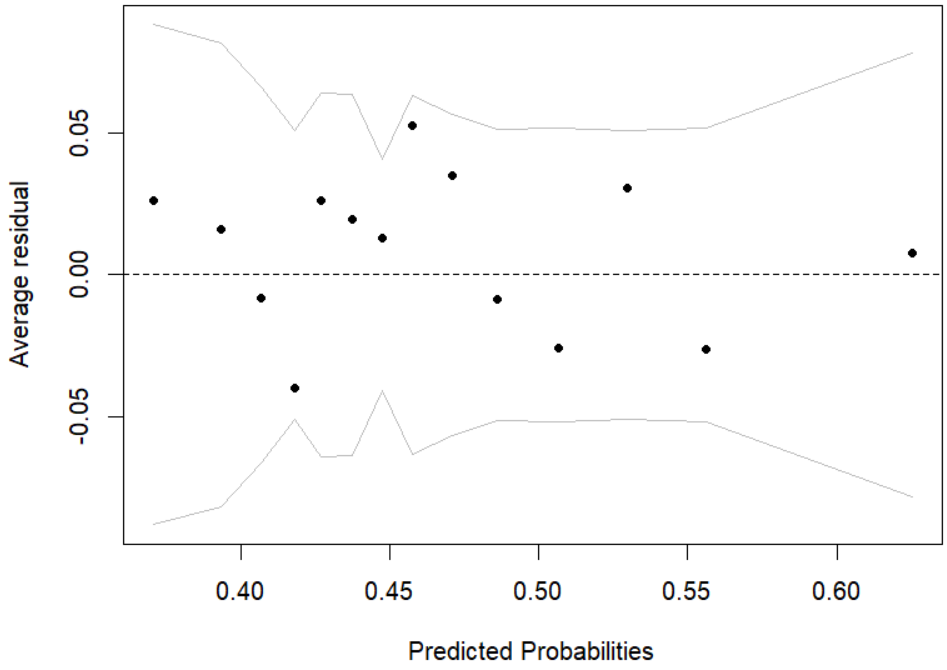

Figure 7: Binned residual plot for logistic regression model.

```
Call:
glm(formula = cbind(votes_cast, non_voters) ~ (median_income +
    median_age + bachelors_degree_pct + hispanic_pct)^2, family = binomial,
    data = hidalgo_main)

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                      1.898e+00  2.335e-01   8.131 4.26e-16 ***
median_income                   -2.401e-06  3.109e-06  -0.772  0.43991
median_age                      -5.202e-02  5.039e-03 -10.323  < 2e-16 ***
bachelors_degree_pct            -4.239e-02  4.738e-03  -8.948  < 2e-16 ***
hispanic_pct                    -2.535e-02  2.338e-03 -10.844  < 2e-16 ***
median_income:median_age         4.406e-07  3.631e-08  12.135  < 2e-16 ***
median_income:bachelors_degree_pct  1.579e-07  1.332e-08  11.854  < 2e-16 ***
median_income:hispanic_pct      -8.925e-08  2.871e-08  -3.108  0.00188 **
median_age:bachelors_degree_pct -6.626e-04  5.642e-05 -11.744  < 2e-16 ***
median_age:hispanic_pct          5.548e-04  5.275e-05  10.517  < 2e-16 ***
bachelors_degree_pct:hispanic_pct  7.000e-04  4.166e-05  16.801  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 29236  on 207  degrees of freedom
Residual deviance: 20457  on 197  degrees of freedom
AIC: 22131
```

Figure 8: Logistic regression model including pairwise interaction terms.

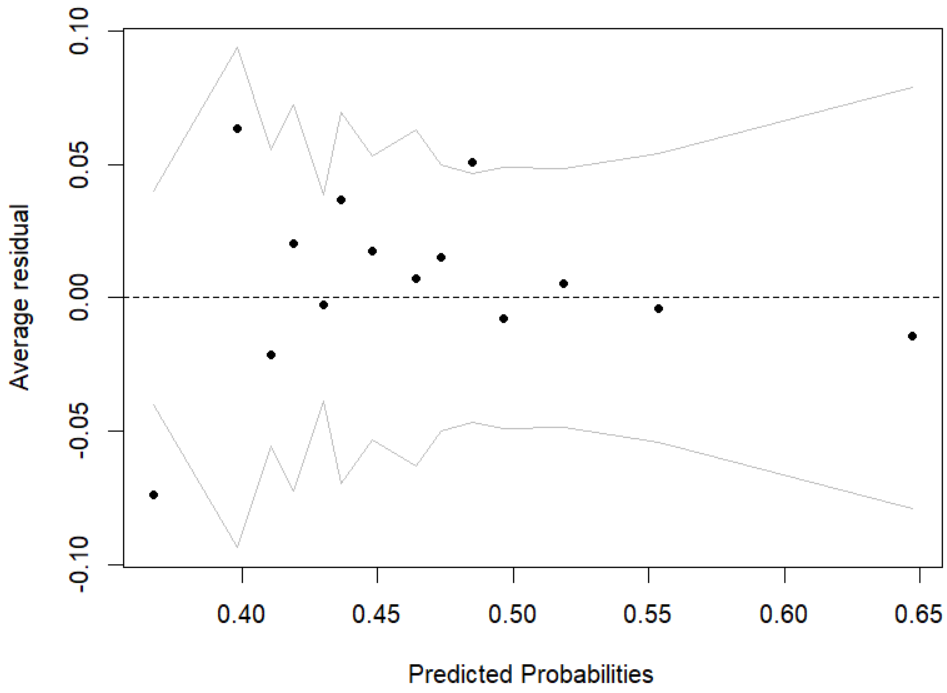

Figure 9: Binned residual plot for model with interaction terms.

.2 R Code

```
################################
### Final Project for MATH 6364##
#########John Knight ############
################################
## Hidalgo County Voter Turnout##

# This project calculates the % of over-18s who
# voted in the 2024 US Election in each census
# tract in Hidalgo County, Texas, and compares
# with various demographic measures to draw
# statistical insights.

### Load libraries

library(tidycensus)
library(tidyr)
library(dplyr)
library(tigris)
library(osmdata)
library(ggplot2)
library(sf)
library(tidygeocoder)
library(furrr)
library(ggmap)
library(ggrepel)
library(knitr)
library(kableExtra)
library(ggcorrplot)
library(mice)
library(car)
library(arm)
library(ResourceSelection)

setwd("set your working directory here")

# census_api_key(census_api_key, install = TRUE)
# This is a free API key available from US Census.
# You will also need a google API key if you don't already have one.
# It is a subscription but you get $200 free per month (enough for this
    project).

#########################################################
### Load the census tract data for Hidalgo County#########
### (This data is free, via tidycensus library)###########
#########################################################
```

```r
# Get list of ACS variables (can use these to explore variables)
variables_acs <- load_variables(2022, "acs5", cache = TRUE)

# Define variables of interest, including all education categories
variables_of_interest <- c(
  population_total = "B01003_001",
  population_male_under_5 = "B01001_003",
  population_male_5_to_9 = "B01001_004",
  population_male_10_to_14 = "B01001_005",
  population_male_15_to_17 = "B01001_006",
  population_female_under_5 = "B01001_027",
  population_female_5_to_9 = "B01001_028",
  population_female_10_to_14 = "B01001_029",
  population_female_15_to_17 = "B01001_030",
  population_25_plus = "B15003_001",
  median_income = "B19013_001",
  no_schooling = "B15003_002",
  nursery_school = "B15003_003",
  kindergarten = "B15003_004",
  grade_1 = "B15003_005",
  grade_2 = "B15003_006",
  grade_3 = "B15003_007",
  grade_4 = "B15003_008",
  grade_5 = "B15003_009",
  grade_6 = "B15003_010",
  grade_7 = "B15003_011",
  grade_8 = "B15003_012",
  grade_9 = "B15003_013",
  grade_10 = "B15003_014",
  grade_11 = "B15003_015",
  grade_12 = "B15003_016",
  high_school = "B15003_017",
  ged = "B15003_018",
  college_less_than_one_year = "B15003_019",
  college_more_than_one_year = "B15003_020",
  associate_degree = "B15003_021",
  bachelors_degree = "B15003_022",
  masters_degree = "B15003_023",
  professional_degree = "B15003_024",
  doctorate_degree = "B15003_025",
  race_white_alone = "B02001_002",
  race_black_alone = "B02001_003",
  race_native_american_alone = "B02001_004",
  race_asian_alone = "B02001_005",
  race_pacific_islander_alone = "B02001_006",
  race_other_alone = "B02001_007",
  race_two_or_more = "B02001_008",
```

```
  non_hispanic_white = "B03002_003",
  hispanic_total = "B03002_012",
  median_age = "B01002_001",
  male_over_18_not_citizen = "B05003_012",
  female_over_18_not_citizen = "B05003_023"
)

# Retrieve ACS data for Hidalgo County census tracts
hidalgo <- get_acs(
  geography = "tract",
  variables = variables_of_interest,
  state = "TX",
  county = "Hidalgo",
  year = 2022,
  survey = "acs5"
)

# Clean and reshape the data
hidalgo <- hidalgo %>%
  dplyr::select(-moe) %>%  # Remove the margin of error (moe) column
  pivot_wider(names_from = variable, values_from = estimate) # Reshape
     to have variables as columns

hidalgo$population_18_plus <- hidalgo$population_total - (hidalgo$
    population_male_under_5
                                + hidalgo$population_male_5_to_9 +
                                   hidalgo$population_male_10_to_14
                                + hidalgo$population_male_15_to_17 +
                                   hidalgo$population_female_under_5
                                + hidalgo$population_female_5_to_9 +
                                   hidalgo$population_female_10_to_14
                                + hidalgo$population_female_15_to_17)
hidalgo$bachelors_degree_pct <- (hidalgo$bachelors_degree + hidalgo$
    masters_degree
                                + hidalgo$professional_degree
                                + hidalgo$doctorate_degree) / hidalgo$
                                   population_25_plus * 100
hidalgo$hispanic_pct <- hidalgo$hispanic_total / hidalgo$population_
    total * 100
hidalgo$citizens_over_18 <- hidalgo$population_18_plus - hidalgo$male_
    over_18_not_citizen - hidalgo$female_over_18_not_citizen

# Retrieve ACS data for entire United States
nationwide <- get_acs(
  geography = "us",
  variables = variables_of_interest,
  year = 2022,
```

```
  survey = "acs5"
)

# Clean and reshape the data
nationwide <- nationwide %>%
  dplyr::select(-moe) %>%  # Remove the margin of error (moe) column
  pivot_wider(names_from = variable, values_from = estimate) # Reshape
     to have variables as columns

nationwide$population_18_plus <- nationwide$population_total - (
   nationwide$population_male_under_5
                                                                + nationwide$
                                                                   population
                                                                   _male_5_to
                                                                   _9 +
                                                                   nationwide
                                                                   $
                                                                   population
                                                                   _male_10_
                                                                   to_14
                                                                + nationwide$
                                                                   population
                                                                   _male_15_
                                                                   to_17 +
                                                                   nationwide
                                                                   $
                                                                   population
                                                                   _female_
                                                                   under_5
                                                                + nationwide$
                                                                   population
                                                                   _female_5_
                                                                   to_9 +
                                                                   nationwide
                                                                   $
                                                                   population
                                                                   _female_10
                                                                   _to_14
                                                                + nationwide$
                                                                   population
                                                                   _female_15
                                                                   _to_17)
nationwide$bachelors_degree_pct <- (nationwide$bachelors_degree +
   nationwide$masters_degree
                                    + nationwide$professional_degree
                                    + nationwide$doctorate_degree) /
                                       nationwide$population_25_plus * 100
```

```
nationwide$hispanic_pct <- nationwide$hispanic_total / nationwide$
    population_total * 100
nationwide$citizens_over_18 <- nationwide$population_18_plus -
    nationwide$male_over_18_not_citizen - nationwide$female_over_18_not_
    citizen


##############################################
##############################################

##############################################
### Geocode the voter data using Google API##
##############################################

### NOTE: First you will have to get the latest voting
### data from https://www.hidalgocounty.us/3339/Unofficial-Early-Voting
    -Totals-Rosters
### and rename as hidalgo_voters_early.csv and hidalgo_voters_election_
    day.csv

hidalgo_voters_early <- read.csv("hidalgo_voters_early.csv")
hidalgo_voters_election_day <- read.csv("hidalgo_voters_election_day.
    csv")
hidalgo_voters <- rbind(hidalgo_voters_early, hidalgo_voters_election_
    day)

# register_google(key = google_api_key)

# Load the latest list of voters and ensure ADDRESS is a character
hidalgo_voters <- hidalgo_voters %>%
  mutate(ADDRESS = as.character(ADDRESS))

# Create or load `addresses` data frame with distinct addresses
if (file.exists("addresses.csv")) {
  addresses <- read.csv("addresses.csv", stringsAsFactors = FALSE)
} else {
  # If addresses.csv does not exist, create a distinct addresses data
      frame
  addresses <- hidalgo_voters %>%
    select(ADDRESS) %>%
    distinct() %>%
    mutate(latitude = NA, longitude = NA)
}


# Check for new addresses in `hidalgo_voters` and add them to `
    addresses`
```

```r
new_addresses <- hidalgo_voters %>%
  dplyr::select(ADDRESS) %>%
  distinct() %>%
  filter(!ADDRESS %in% addresses$ADDRESS)

# Append any new addresses to the addresses data frame
if (nrow(new_addresses) > 0) {
  new_addresses <- new_addresses %>%
    mutate(latitude = NA, longitude = NA)

  addresses <- bind_rows(addresses, new_addresses)
  cat("Added", nrow(new_addresses), "new addresses to geocode.\n")
}

# Filter for addresses that still need geocoding (i.e., NA in latitude
   or longitude)
ungeocoded_addresses <- addresses %>%
  filter(is.na(latitude) | is.na(longitude))

# Set total addresses to be geocoded
total_addresses <- nrow(ungeocoded_addresses)

# Loop through ungeocoded addresses and geocode
for (i in seq_len(total_addresses)) {
  # Get the current address
  current_address <- ungeocoded_addresses$ADDRESS[i]

  # Geocode the current address
  result <- tryCatch(
    geocode(current_address, source = "google"),
    error = function(e) NULL
  )

  # Check if geocoding was successful
  if (!is.null(result)) {
    # Find the row in `addresses` corresponding to the current address
    address_index <- which(addresses$ADDRESS == current_address)

    # Update latitude and longitude in `addresses`
    addresses$latitude[address_index] <- result$lat
    addresses$longitude[address_index] <- result$lon
  }

  # Print progress
  cat("Geocoded", i, "of", total_addresses, "ungeocoded addresses\n")

  # Save progress every 1000 addresses or on the last iteration
```

```
  if (i %% 1000 == 0 || i == total_addresses) {
    write.csv(addresses, file = "addresses.csv", row.names = FALSE)
    cat("Progress saved to 'addresses.csv' at address", i, "\n")
  }
}

# Final save to ensure all updates are written
write.csv(addresses, file = "addresses.csv", row.names = FALSE)

# Join the geocoded addresses back to `hidalgo_voters` as needed
hidalgo_voters <- hidalgo_voters %>%
  left_join(addresses, by = "ADDRESS")

write.csv(hidalgo_voters, file = "hidalgo_voters_geocoded.csv", row.
   names = FALSE)

##############################################
##############################################

##############################################
### Combine Demographic and Voter Data #####
##############################################

# Need to get the census tract data from above.

# Assuming voter addresses are already geocoded, can
# load from here:

hidalgo_voters <- read.csv("hidalgo_voters_geocoded.csv")

# Convert `hidalgo_voters` to an sf object
hidalgo_voters_sf <- hidalgo_voters %>%
  filter(!is.na(latitude) & !is.na(longitude)) %>%  # Only keep rows
     with valid coordinates
  st_as_sf(coords = c("longitude", "latitude"), crs = 4326)  # Set CRS
     to WGS84 (EPSG:4326)

# Load Hidalgo County census tracts
hidalgo_tracts <- tracts(state = "TX", county = "Hidalgo", cb = TRUE,
   class = "sf")

# Ensure both are in the same CRS (project to match if necessary)
hidalgo_voters_sf <- st_transform(hidalgo_voters_sf, crs = st_crs(
   hidalgo_tracts))

# Perform a spatial join to add census tract information to each voter
```

```
hidalgo_voters_with_tracts <- st_join(hidalgo_voters_sf, hidalgo_tracts
    , join = st_within)

# Extract the GEOID column from census tracts as the tract identifier
hidalgo_voters_with_tracts <- hidalgo_voters_with_tracts %>%
  mutate(tract_id = GEOID) %>%  # Rename GEOID to tract_id if preferred
  dplyr::select(-geometry)  # Drop the geometry column if you don't
     need it

# Count occurrences of each GEOID in hidalgo_voters_with_tracts
tract_counts <- hidalgo_voters_with_tracts %>%
  group_by(GEOID) %>%
  summarise(voter_count = n())  # `voter_count` will be the count of
     rows per GEOID

# Join the counts back to hidalgo based on GEOID
hidalgo <- hidalgo %>%
  left_join(tract_counts, by = "GEOID")

# Replace NA values in `voter_count` with 0 for tracts with no voters
hidalgo$voter_count[is.na(hidalgo$voter_count)] <- 0

# Convert to a pct based on over 18 eligible voting population
hidalgo$voter_pct = hidalgo$voter_count / hidalgo$citizens_over_18 *
    100

# Create a new DF with just the interesting columns
hidalgo_main <- hidalgo[,c("GEOID", "NAME","citizens_over_18", "median_
    age", "median_income", "bachelors_degree_pct", "hispanic_pct", "
    voter_count", "voter_pct")]

# Remove row with GEOID = 48215980000 as it has no population or
    statistics
remove_row <- which(hidalgo_main$GEOID=="48215980000")
hidalgo_main <- hidalgo_main[-remove_row,]


################################################
################################################

################################################
### Deal with Outliers ########################
################################################

# Remove Tres Lagos tract because the population is way off and voter_
    pct > 100
remove_row <- which(hidalgo_main$voter_pct > 100)
```

```r
hidalgo_main <- hidalgo_main[-remove_row,]

# Scatter plot showing winter Texans tract as outlier
outlier_median_age <- 72.5
outlier_hispanic_pct <- 24.5

outlier_scatterplot <- ggplot(hidalgo_main, aes(x = median_age, y =
    hispanic_pct)) +
  geom_point(color = "blue") +
  labs(x = "Median Age", y = "Hispanic Percentage") +
  geom_smooth(method = "lm", color = "black", se = FALSE) +
  geom_point(aes(x = outlier_median_age, y = outlier_hispanic_pct),
             color = "red", size = 10, shape = 1)

ggsave("outlier_scatterplot.png", plot = outlier_scatterplot, width =
    6, height = 4, dpi = 300)

# Remove winter Texans tract because it is a huge outlier and some may
    vote up North
remove_row <- which(hidalgo_main$hispanic_pct < 50)
hidalgo_main <- hidalgo_main[-remove_row,]

# Remove the prison tract
remove_row <- which(hidalgo_main$NAME == "Census Tract 235.30; Hidalgo
    County; Texas")
hidalgo_main <- hidalgo_main[-remove_row,]

################################################
################################################

################################################
#### Create Heatmap ############################
################################################

### Need to ensure hidalgo_main is already loaded

hidalgo_tracts <- tracts(state = "TX", county = "Hidalgo", cb = TRUE,
    class = "sf")

heatmap_data <- hidalgo_tracts %>%
  left_join(hidalgo_main, by = "GEOID")

# Use osmdata to get highways in Hidalgo County
hidalgo_highways <- opq(bbox = st_bbox(heatmap_data)) %>%
  add_osm_feature(key = "highway", value = c("motorway", "trunk", "
      primary")) %>%
  osmdata_sf()
```

```
# Calculate a slightly expanded bounding box
bbox <- st_bbox(hidalgo_tracts)
expand_ratio <- 0.05  # Adjust this ratio as needed

xlim <- c(bbox["xmin"] - expand_ratio * (bbox["xmax"] - bbox["xmin"]),
          bbox["xmax"] + expand_ratio * (bbox["xmax"] - bbox["xmin"]))
ylim <- c(bbox["ymin"] - expand_ratio * (bbox["ymax"] - bbox["ymin"]),
          bbox["ymax"] + expand_ratio * (bbox["ymax"] - bbox["ymin"]))

# Load places.csv and separate latitude and longitude
places <- read.csv("places.csv", stringsAsFactors = FALSE)

# Separate coordinates into latitude and longitude
places <- places %>%
  separate(Coordinates, into = c("latitude", "longitude"), sep = ", ")
     %>%
  mutate(
    latitude = as.numeric(latitude),
    longitude = as.numeric(longitude)
  )

# Convert places data frame to an sf object
places_sf <- st_as_sf(places, coords = c("longitude", "latitude"), crs
   = 4326)

# Add landmarks to the plot
my_heatmap <- ggplot() +
  # Base layer: Census tracts with values
  geom_sf(data = heatmap_data, aes(fill = voter_pct), color = "white",
     lwd = 0.2) +
  scale_fill_gradient(low = "white", high = "red", limits = c(0, NA)) +

  # Overlay highways
  geom_sf(data = hidalgo_highways$osm_lines, color = "blue", size =
     0.4, alpha = 0.7) +

  # Add landmarks with ggrepel
  geom_sf(data = places_sf, color = "black", size = 2) +  # Marker for
     each place
  geom_text_repel(data = places, aes(x = as.numeric(longitude), y = as.
     numeric(latitude), label = Place),
                  color = "black", size = 3, nudge_x = 0.001) +  #
                     Place name labels with repelling

  # Set expanded coordinate limits
  coord_sf(xlim = xlim, ylim = ylim, expand = FALSE) +
```

```
  # Theme and labels
  theme_minimal() +
  theme(
    axis.title = element_blank(),
    axis.text = element_blank(),
    axis.ticks = element_blank(),
    panel.grid = element_blank()
  ) +
  labs(fill = "% Voted")

# Save the plot
ggsave("heatmap.png", plot = my_heatmap, width = 6, height = 4, dpi =
    300)

################################################
################################################

################################################
### Descriptive Stats ##########################
################################################

### Summary Table

# Define the variables of interest
my_variables <- c("median_age", "median_income", "bachelors_degree_pct"
    ,
                  "hispanic_pct", "voter_pct")

# Initialize an empty data frame to store the results
summary_data <- data.frame(Variable = character(),
                           n = integer(),
                           Mean = numeric(),
                           SD = numeric(),
                           Min = numeric(),
                           Max = numeric(),
                           stringsAsFactors = FALSE)

# Loop through each variable in my_variables and calculate the
    statistics
for (var in my_variables) {
  # Calculate statistics for the current variable
  n <- sum(!is.na(hidalgo_main[[var]]))
  mean_val <- mean(hidalgo_main[[var]], na.rm = TRUE)
  sd_val <- sd(hidalgo_main[[var]], na.rm = TRUE)
  min_val <- min(hidalgo_main[[var]], na.rm = TRUE)
  max_val <- max(hidalgo_main[[var]], na.rm = TRUE)
```

```
  # Add a new row to the summary_data data frame
  summary_data <- rbind(summary_data, data.frame(Variable = var,
                                                 n = n,
                                                 Mean = mean_val,
                                                 SD = sd_val,
                                                 Min = min_val,
                                                 Max = max_val))
}

# Display the table with kable for LaTeX output
kable(summary_data, format = "latex", booktabs = TRUE, digits = 2,
      caption = "Descriptive Statistics for Selected Variables")

### Histograms / Boxplots

my_hist <- ggplot(hidalgo_main, aes(x = median_age)) +
  geom_histogram(bins = 20, fill = "skyblue", color = "black") +
  labs(x = "Median Age of Census Tract",
       y = "Frequency")

ggsave("hist_age.png", plot = my_hist, width = 6, height = 4, dpi =
   300)

my_hist <- ggplot(hidalgo_main, aes(x = median_income)) +
  geom_histogram(bins = 20, fill = "skyblue", color = "black") +
  labs(x = "Median Income of Census Tract",
       y = "Frequency")

ggsave("hist_income.png", plot = my_hist, width = 6, height = 4, dpi =
   300)

my_hist <- ggplot(hidalgo_main, aes(x = bachelors_degree_pct)) +
  geom_histogram(bins = 20, fill = "skyblue", color = "black") +
  labs(x = "% of Residents with Bachelor's Degree",
       y = "Frequency")

ggsave("hist_degree.png", plot = my_hist, width = 6, height = 4, dpi =
   300)

my_hist <- ggplot(hidalgo_main, aes(x = hispanic_pct)) +
  geom_histogram(bins = 20, fill = "skyblue", color = "black") +
  labs(x = "Hispanic Resident %",
       y = "Frequency")

ggsave("hist_hispanic.png", plot = my_hist, width = 6, height = 4, dpi
   = 300)
```

```
my_hist <- ggplot(hidalgo_main, aes(x = voter_pct)) +
  geom_histogram(bins = 20, fill = "skyblue", color = "black") +
  labs(x = "% Voted",
       y = "Frequency")

ggsave("hist_voter_pct.png", plot = my_hist, width = 6, height = 4, dpi
    = 300)

### Correlation Matrix

# Subset the data to only include my_variables
data_subset <- hidalgo_main[my_variables]

# Calculate the correlation matrix
corr_matrix <- cor(data_subset, use = "complete.obs")

# Create the correlation matrix plot
my_corrplot <- ggcorrplot(corr_matrix,
                          method = "square", # Type of plot: "circle",
                             "square", etc.
                          lab = TRUE,        # Show correlation
                             coefficients
                          colors = c("red", "white", "blue"), # Colors
                             for correlation
                          ggtheme = theme_minimal())

ggsave("corr_matrix.png", plot = my_corrplot, width = 6, height = 4,
   dpi = 300)

### Side-by-side barplot of turnout %
# Create the data frame
turnout_data <- data.frame(
  Year = rep(c(2012, 2016, 2020), each = 2),
  Region = rep(c("Hidalgo County", "U.S."), times = 3),
  Turnout = c(45.95, 53.8, 51.97, 54.8, 56.70, 62.8)
)

# Plot the data
turnout_plot <- ggplot(turnout_data, aes(x = factor(Year), y = Turnout,
    fill = Region)) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.7),
     width = 0.5) +
  labs(x = "Year", y = "Turnout %") +
  scale_fill_manual(values = c("Hidalgo County" = "blue", "U.S." = "red
     "))
```

```
ggsave("turnout_plot.png", plot = turnout_plot, width = 6, height = 4,
    dpi = 300)

##############################################
### Missing Data Imputation ##################
##############################################

# 5 tracts have median_income data missing
# Tracts are 202.06, 205.13, 207.35, 241.31, 242.11

# Use the MICE imputation method
impute_data <- hidalgo_main[, c("median_income", "median_age", "
    bachelors_degree_pct", "hispanic_pct")]
imputed_data <- mice(impute_data, method = "pmm", m = 1, maxit = 5,
    seed = 915)
completed_data <- complete(imputed_data)

# Update the original dataset with imputed values
hidalgo_main$median_income <- completed_data$median_income


##############################################
### Logistic Regression ######################
##############################################

### Need to convert voter_pct into a binomial count

# Calculate the number of votes and non-votes
hidalgo_main$votes_cast <- round(hidalgo_main$voter_pct / 100 * hidalgo
    _main$citizens_over_18)
hidalgo_main$non_voters <- hidalgo_main$citizens_over_18 - hidalgo_main
    $votes_cast

# Fit the logistic regression model
model <- glm(cbind(votes_cast, non_voters) ~ median_income + median_age
     + bachelors_degree_pct + hispanic_pct,
              family = binomial, data = hidalgo_main)
summary(model) # AIC 22962

### Diagnostic checks

# Inspect VIF to assess multicollinearity
vif(model)

# Binned Residual Plot
png("binned_residual_plot.png", width = 800, height = 600, res = 120)
par(mar = c(5, 4, 2, 2)) # Adjusts margins: c(bottom, left, top, right)
```

```
binnedplot(fitted(model), residuals(model, type = "response"),
           xlab = "Predicted Probabilities",
           main = "")
dev.off() # Close the device


# Hosmer-Lemeshow Test for goodness of fit
hidalgo_main$voter_prop <- hidalgo_main$voter_pct / 100
predicted_probs <- fitted(model)
hoslem.test(hidalgo_main$voter_prop, predicted_probs, g = 10)


### Try interactions

# Fit the logistic regression model
full_model <- glm(cbind(votes_cast, non_voters) ~ (median_income +
   median_age + bachelors_degree_pct + hispanic_pct)^2,
            family = binomial, data = hidalgo_main)
summary(full_model) # AIC 22131

# Binned residual plot
# Set margins to remove top space
png("binned_residual_plot_interactions.png", width = 800, height = 600,
    res = 120)
par(mar = c(5, 4, 2, 2)) # Adjusts margins: c(bottom, left, top, right)

# Create the plot without a title and no top gap
binnedplot(fitted(full_model), residuals(full_model, type = "response")
   ,
           xlab = "Predicted Probabilities",
           main = "")

dev.off() # Close the device

# Hosmer-Lemeshow Test for goodness of fit
predicted_probs <- fitted(full_model)
hoslem.test(hidalgo_main$voter_prop, predicted_probs, g = 10)

### Predictions

# Make up a few examples
new_data <- data.frame(
  median_income = c(52000, 30000, 100000, 80000),
  median_age = c(32, 25, 45, 30),
  bachelors_degree_pct = c(20, 5, 60, 60),
  hispanic_pct = c(92, 98, 70, 90)
)
```

```
# Make predictions using the model
predictions <- predict(full_model, newdata = new_data, type = "response
    ")

# Display the predictions
predictions
```