Data-Driven Research: Worry about Climate Change in the Texan Borderlands

Minerva University

NS125: Research Methods

Prof. Zoogman

October 27, 2022

930 Words

Data-Driven Research: Worry about Climate Change in the Texan Borderlands

Exploratory Data Analysis

For this investigation of climate change perceptions, I began with a dataset from Howe et al. (2015). They conducted a US-wide survey assessing geographic variation in climate change perceptions and collecting county level data on a variety of variables pertaining to climate change perception. For this analysis, I focus on worry about climate change, which Howe et al. quantified as the proportion of people in a county who stated that they were either somewhat or very worried about global warming. Find the raw data from Howe et al. here.

Figure 1a shows the Z-scores of worry by county in the United States (mean national worry $\bar{x}_{\text{USA}} = 52.03\%$). Z-scores are a more interesting metric to focus on than the simple proportion of surveyed individuals, because they highlight outlier counties. Outliers, because they deviate from the norm, can be interesting to investigate as we can investigate the mechanisms that make them an outlier. The code used to clean the Howe data, and map the Z scores, is in Listing 1 in the Appendix.

I wondered what variables this might correlate with. One that I fell upon was voting patterns. Many counties in the Northeast, Southwest and West Coast of the United States, where people are more worried about global warming (Figure 1a), also vote Democrat. Figure 1b, using data originally published by Reuters and assembled by Kaggle user Raphael Fontes, shows this clearly: counties above average in worry about global warming tended to vote for Joe Biden, while Donald Trump was popular in less worried counties. [1]

Figure 2 shows the distribution of worry for Biden and Trump counties. A t-test reveals that the 11.06% difference in mean worry between Biden and Trump counties is statistically significant at the 5% level (code in Listing 2 in the Appendix). This supports the visual

---

[1] #differences: Though I was able to produce an interesting visualisation of geographic trends in climate change worry, it was difficult to evaluate the plausibility of different possibly correlating variables to explore. So, I asked an American classmate if there were aspects of Figure 1a that surprised her, and if she noted any geographic patterns. One such pattern was that the worry levels appeared to track with political lines. This is not a trend that I would have easily picked up on.

correlation I first noticed in the maps (Figure 1. [2], 95% confidence interval [10.61, 11.50] [3] See Listing 2 in the Appendix for the code used to join, visualise and analyse the voting data.

Considering that Republican counties are generally less concerned about climate change than Democratic counties, regions that are shown as Republican in Figure 1b, but as more worried about climate change like those along the Texas-Mexico border in Figure 1a, are anomalous and interesting to further explore. Counties along the border are considerably more worried than the national average about global warming, yet the majority of the state, both by population and by county, is less worried than average about global warming. This is an interesting result, as it suggests that there may be features specific to this region that prompt this difference. [4]

A Texan classmate suggested that there could be a correlation with the proportion of Hispanic people in the county. Using data from the US Census Bureau (2019), I explored this correlation. The correlation is clear in Figure 3. There is a statistically significant, strong, positive correlation between the proportion of Hispanic residents in Texan counties and the proportion of residents who are worried about climate change ($r = 0.81$, $p < 0.05$). Though this does not imply a causal relationship, it does prompt further investigation into the opinions of worry in these

---

[2]#significance: Used a parametric t-test to assess whether the difference in means between Trump and Biden voting counties was statistically significant. The assumptions for the t-test were met:

1. Simple random sample: Howe et al.'s survey methodology used random sampling to gather survey results.
2. Sample distribution of sample means is approximately normal. This is accomplished as both sample sizes are much greater than 30, so by the Central Limit Theorem, the sample distribution of sample means will be approximately normal.

Used these results to justify controlling for political affiliation for the remainder of the analysis.

[3]#confidenceintervals: The 11.06% difference in mean worry between Biden and Trump counties is only a point estimate. This is especially relevant considering that only a subset of Biden counties were considered for the analysis. Using a confidence interval provides insight into the range of values that the difference could feasibly cover: 95% of the time, the true difference between these groups will be on the interval [10.61, 11.50]. The confidence interval was calculated from the same t-distribution used to perform the t-test. The assumptions for using this distribution were verified [2].

[4]#heuristics: As we discussed in Session 12, creativity is often important for producing intriguing angles of exploration that are both novel and fruitful. To find such angles for this project, I applied the creative heuristic of accounting for deviations from the natural trend: I aimed to explain how Republican counties are generally less worried about climate change than the national average (Figure 2. In an exploratory context, unusual results provide chances to uncover new patterns and causal mechanisms in follow-up research.

counties, as it seems they may differ ethnographically.[5]

Further Research

In the exploratory analysis, I identified that counties in the Texan borderlands with high proportions of Hispanics are unusually worried about climate change. Many of these counties also tend to vote red. A follow-up study that explores the attitudes of residents within these Republican but worried counties would shed light onto the worries regarding climate change that are specific to this group. Such a study is important as Hispanic people, being a minority group in the United States, may be more vulnerable to the effects of climate change (Hansen et al., 2013). By understanding this perspective, we begin to build a picture of how climate change, and climate policy, fits into the marginalising structures encumbering Hispanic people.

A qualitative approach is well-suited to understanding the Hispanic perspective on global warming along the Texas-Mexico border. Such an approach allows for new themes and ideas to be elucidated and explored; this is something that a quantitative approach does not do as effectively. Considering the findings of the exploratory data analysis, the exploration would seek to identify the climate-caused threats identified by the residents of these counties, both Hispanic and not, local climate policy satisfaction and how they felt that dissatisfaction was understood. To this end, semi-structured interviews are a useful approach. They allow consistency in the themes explored, but still allow for participant responses to guide the conversation. A structured approach would be inappropriate, as it limits the interviewer's ability to explore surprising topics, while an unstructured approach means that specific issues to explore may not be adequately addressed.

Given that this is an exploratory study, a convenience sample that strives to collect a diversity of opinion would be appropriate. Participants could be recruited from a variety of community programs: leisure centers, schools and community centers are all places where participants could be sought. Importantly, these are places where participants of a range of

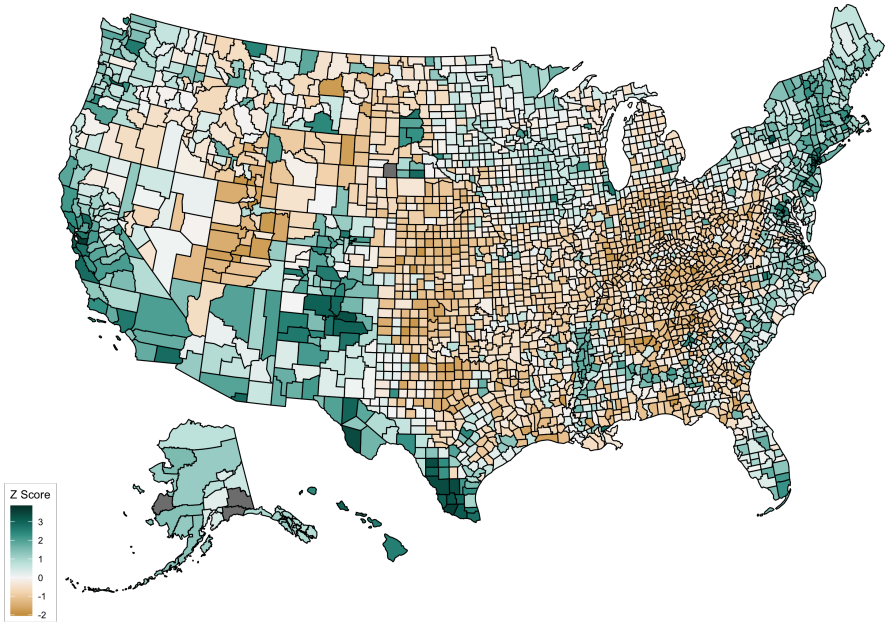
---

[5]#correlation: I describe the positive correlation between the proportion of Hispanic residents and worry, and interpret the meaning of Pearson's r. I justify why this correlation is interesting in spite of the *post hoc ergo propter hoc* fallacy.

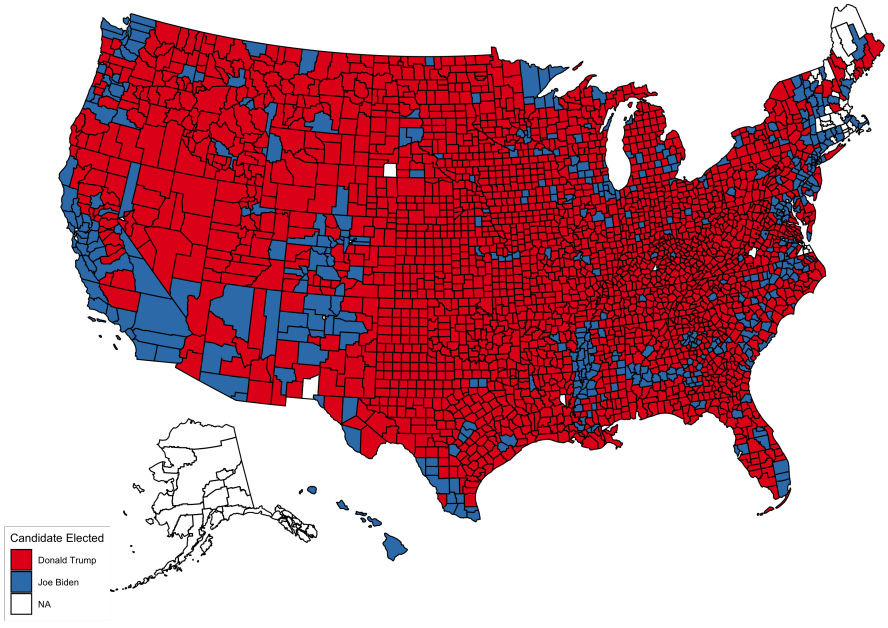
demographics visit, decreasing inherent biases in the sample. Nonetheless, non-identifying demographic data would be collected to assess for confounds retrospectively.

Important to the processing and interpretation of results will be community involvement and participant verification. Having participants play an active role in the coding, interpretation and discussion of results will ensure that the interpretations made are accurate and reflect those held by the participants. This especially important when the researcher is not a member of the group being surveyed. After conducting the interviews, theme extraction could then be used to understand the high-level trends in opinion and their differences with other groups in Texas.

930 Words



(a) *Z scores by county for degree of worry about climate change ($\bar{x}_{USA} = 52.03\%$, $s_{USA} = 5.970\%$). Data from Howe et al. (2015)*



(b) *Candidate elected in each county. Data for Alaska, Massachusetts and Maine unavailable as these states do not publish electoral results by county. Data from Fontes (n.d.)*

Figure 1

Worry about climate change (2016) (panel 1a) and candidate elected by county (2020) (panel 1b). Note the similarity in the two maps: more concerned counties appear to generally vote for Joe Biden, while less concerned counties tend to vote for Donald Trump.

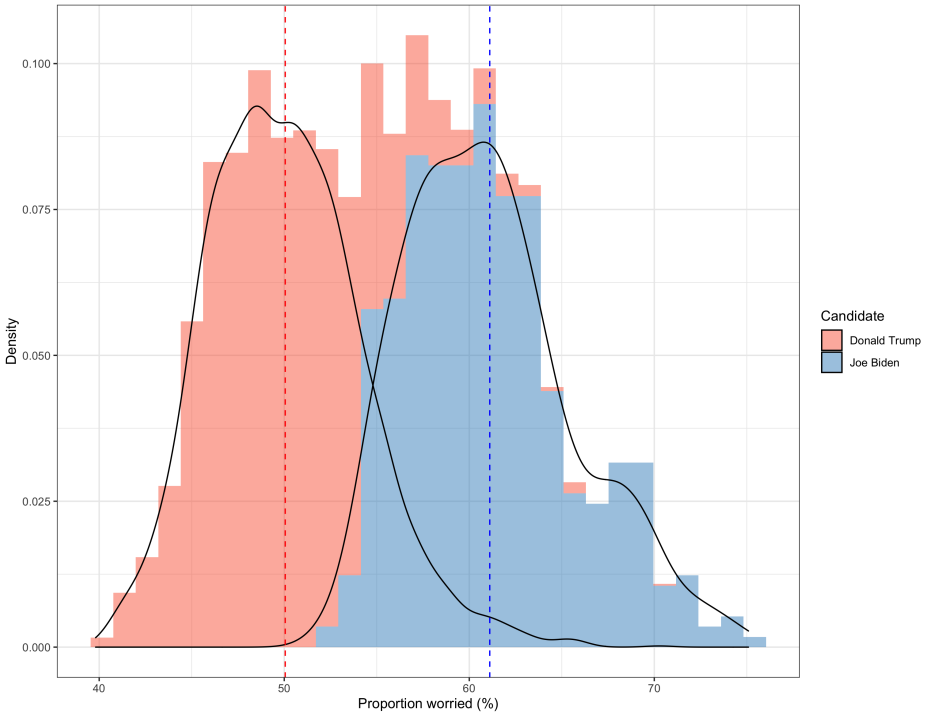


Figure 2

Histogram showing density of worry in counties that voted for Donald Trump (red, $n = 2561$) and Joe Biden (blue $n = 468$). Mean worry about climate change for Trump counties is significantly lower than for Biden counties ($\bar{x}_{Trump} = 50.05\%$, $\bar{x}_{Biden=61.11\%}$). T-test reveals $p < 0.05$, CI = [10.61, 11.50]. Data combined from Fontes (n.d.) and Howe et al. (2015). Worry data for Alaska, Maine and Massachusetts excluded as voting data was unavailable for these states.

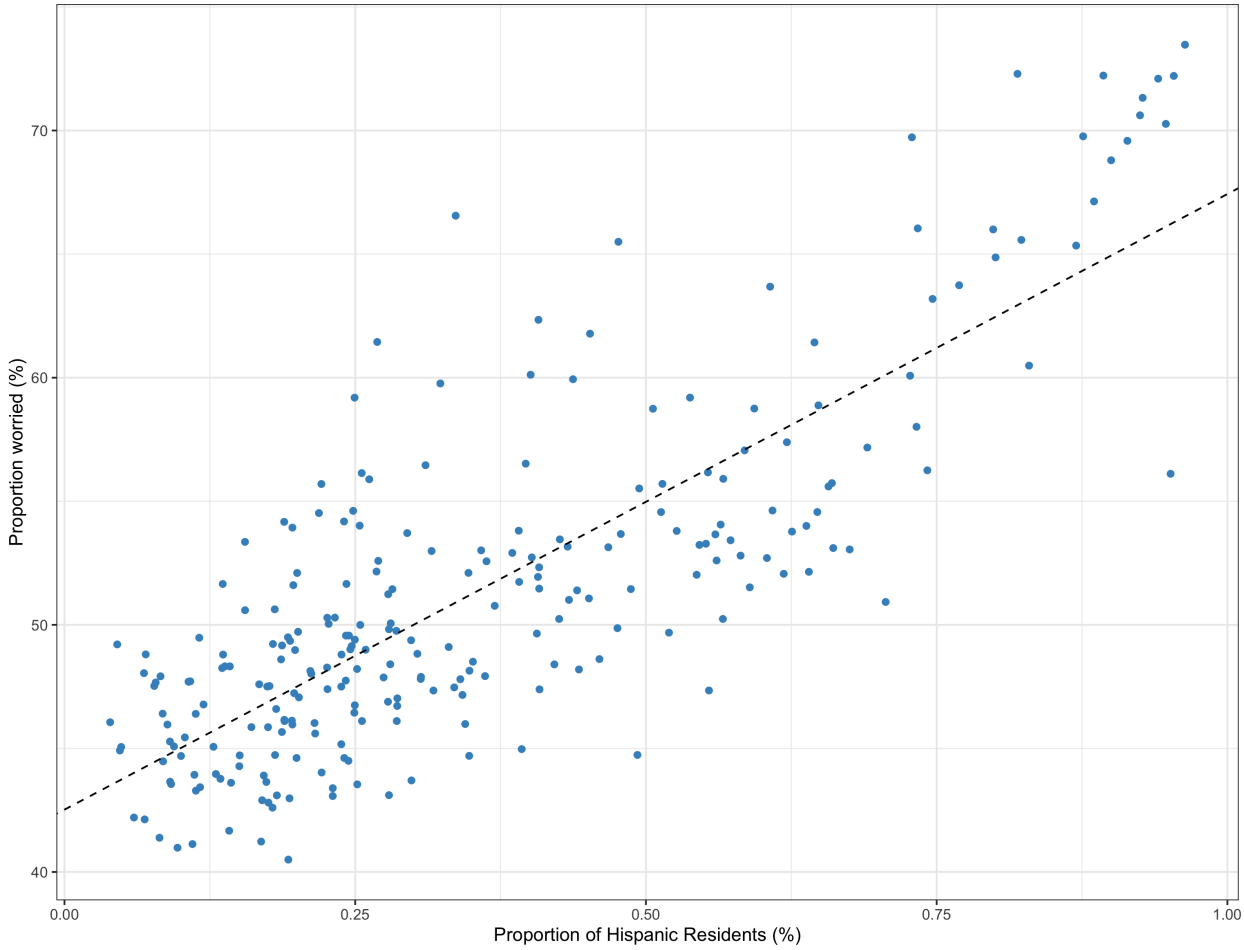


Figure 3

Scatter plot showing the proportion of Hispanic residents and degree of climate change worry in Republican Texan counties. There is a statistically significant correlation between the two variables ($r = 0.81, p < 0.05$). *Data from US Census Bureau (2019) and Howe et al. (2015).*

References

Fontes, R. (n.d.). US Election 2020. Retrieved October 22, 2022, from https://www.kaggle.com/datasets/unanimad/us-election-2020

Hansen, A., Bi, L., Saniotis, A., & Nitschke, M. (2013). Vulnerability to extreme heat and climate change: Is ethnicity a factor? *Global Health Action*, *6*(1), 21364. https://doi.org/10.3402/gha.v6i0.21364

Howe, P. D., Mildenberger, M., Marlon, J. R., & Leiserowitz, A. (2015). Geographic variation in opinions on climate change at state and local scales in the USA. *Nature Climate Change*, *5*(6), 596–603. https://doi.org/10.1038/nclimate2583

US Census Bureau. (2019). County Population by Characteristics: 2010-2019 [Section: Government]. Retrieved October 22, 2022, from https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-detail.html

Appendix

R Code Snippets

Listing 1: Manipulating Howe data; Z score map

```
#cleaning the Howe data
#the data here is surprisingly clean! some NA values in the mediaweekly and
    mediaweeklyOppose columns, but that's all

#start by breaking the data up into state, county and CBSA level
state_df <- df %>% filter(GeoType == 'State')
county_df <- df %>% filter(GeoType == 'County')
CBSA_df <- df %>% filter(GeoType == 'CBSA')

#keeping only the non-oppose columns
oppose_col_vec <-
    c('discussOppose','CO2limitsOppose','trustclimsciSSTOppose','regulateOppose',
              'supportRPSOppose','fundrenewablesOppose','mediaweeklyOppose','happeningOppos
              'humanOppose','consensusOppose','worriedOppose','personalOppose','harmUSOppos
              'devharmOppose','futuregenOppose','harmplantsOppose','timingOppose','mediawe
county_df_no_oppose <- county_df %>%
    select(-oppose_col_vec)

#I'll be using the plot_usmap function to produce maps of the US. This
    package requires a fips column that contains the unique FIPS code of a
    county. This is conveniently in the GEOID column, which just needs to be
    renamed so that plot_usmap recognises it as the FIPS code column.

county_df_no_oppose <- county_df_no_oppose %>%
    rename('fips' = 'GEOID')

#making the Z score plot
```

```
    #only using numeric columns (as colMeans only accepts numeric columns)
    county_df_no_oppose_numeric <- select(county_df_no_oppose,
        -c('GeoType','GeoName'))
    #calculating the means of each column
    means <- colMeans(county_df_no_oppose_numeric)
    #computing the difference between each value and the mean
    diff_means_df <-
        county_df_no_oppose_numeric[]-means[col(county_df_no_oppose_numeric[])]

    #plotting the Z score map
    counties_deviation_df <- diff_means_df %>%
        select(c('worried','fips'))

    counties_deviation_df$z_worried <-
        counties_deviation_df$worried/sd(counties_deviation_df$worried)

    library('colorspace') #needed for customising the midpoint of the colour scale

    plot_usmap(regions = 'counties', data = counties_deviation_df, values =
        'z_worried')+
      scale_fill_continuous_divergingx(palette = 'BrBG', mid = 0, name = 'Z
          Score')
```

Listing 2: Joining voting data; producing the histogram; t-test

```
president_df <- read.csv('president_county_candidate.csv')
#the FIPS csv was obtained by using the Python code found at the end of this
    Kaggle workbook:
    https://www.kaggle.com/code/tygrha/add-fips-to-2020-election-data/notebook
fips_df <- read.csv('FIPS.csv')
```

```
#Goal is to compare voting patterns and worry on the county level, so I will
    want to join the datasets based on county. The FIPS code is a unique county
    identifier and is appropriate for joining the data. Fips codes and the
    equivalent counties are given in FIPS.csv and stored under fips_df.

#adding FIPS labels for easy joining
#first, we need to drop the word "county" from the county names, as otherwise
    the join will not work
president_df$county <- sub(' County', '', president_df$county)

#producing the histogram comparing worry in Trump and Biden voting counties
#first we want a dataframe that shows worry by county as well as whether the
    county voted for Biden or Trump.
#The left join means that only counties where voting data was available will be
    included. This means that Maine, Alaska and Massachussetts is excluded, as
    county level voting data is unavailable.
joined_df_ungroup <- president_df %>%
  left_join(fips_df) %>%
  right_join(county_df_no_oppose)%>%
  select(-c('GeoType','GeoName')) %>% #filter out the duplicate columns
  filter((candidate == 'Donald Trump' & won == 'True') | candidate == 'Joe
      Biden' & won == 'True' )

#producing the plot
ggplot(data = joined_df_ungroup, mapping = aes(x = worried, fill = candidate))+
  #using density for the histogram as sample sizes differ greatly: density
      allows normalises the histograms meaning we can more easily examine them
      'side by side'
```

```
  geom_histogram(aes(y = ..density..), alpha = 0.5)+
  scale_fill_manual(values = c('#fb6a4a','#4292C6'), name = 'Candidate')+
  #adding in the density lines makes it easier to see the overall, approximate
     shape of the distributions. Useful for checking significance assumptions.
  geom_density(alpha = 0.0)+
  geom_vline(xintercept = joined_df$worried_av[1], linetype = 'dashed',
             colour = 'red', name = 'Mean Worry Trump Counties')+
  geom_vline(xintercept =joined_df$worried_av[2], linetype = 'dashed',
              colour = 'blue', name = 'Mean Worry Biden Counties')+
  xlab('Proportion worried (%)')+
  ylab('Density')+
  theme_bw()

  #difference of means t-test
  trump <- joined_df_ungroup %>%
  filter(candidate == 'Donald Trump') %>%
  select(c('worried'))
joe <- joined_df_ungroup %>%
  filter(candidate == 'Joe Biden')%>%
  select(c('worried'))

t.test(x = trump, y = joe) #statistically significant result.
```

Listing 3: Joining race demographic data; producing scatter plot; correlation test

```
#the texas race df shows the percentage of each county that is non-white.
texas_race_df <- read.csv('cc-est2019-alldata-48.csv') %>%
  #we only care about the 2019 estimate for all ages
  filter(YEAR == 12, AGEGRP == 0)
#creating a column that represents the proportion of Hispanic residents, PCT_HS
texas_race_df$PCT_HS <- (texas_race_df$H_MALE +
```

```
    texas_race_df$H_FEMALE)/texas_race_df$TOT_POP

#again, the county level data here should be joined to the worried data frame to
    ease analysis
#first need to remove " County" so that we can do the join along the county
    names columns
texas_race_df$county <- sub(' County', '', texas_race_df$CTYNAME)
texas_race_df %>%
  rename('state' = 'STNAME')#this lets us additionally join along the state
      column in fips_df, so that we are only getting the FIPS code of Texan
      states.

#performing the joins and selecting only relevant columns
texas_race_df <- texas_race_df %>%
  left_join(filter(fips_df, state == 'Texas')) %>% #adding fips so we can join
      with fips codes
  left_join(county_df_no_oppose) %>% #joining in the climate change perceptions
      data on the fips codes
  select(c('fips', 'county','worried','PCT_NWA', 'PCT_HS'))#selecting only
      columns of interest

#producing the plot:

#linear regression for the line
linreg <- lm(texas_race_df$worried~texas_race_df$PCT_HS)
ggplot(data = texas_race_df, mapping = aes(x = PCT_HS, y = worried))+
  geom_point(colour = '#4292C6')+
  geom_abline(intercept = linreg$coefficients[1], slope =
      linreg$coefficients[2], linetype = 2)+
  xlab('Proportion of Hispanic Residents (%)')+
```

```
  ylab('Proportion worried (%)')+
  theme_bw()


#correlation
cor.test(texas_race_df$PCT_HS, texas_race_df$worried)
```