Homework 4C - DATA-312

Jeffrey Williams

01, May 2022

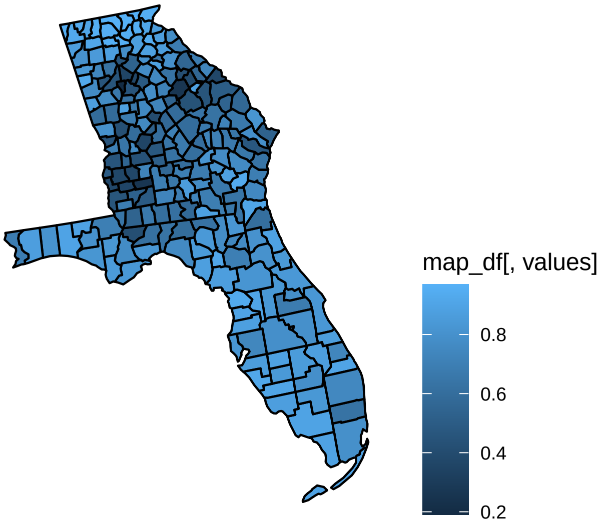
Abstract

This writeup explores a cumulative dataset representative of various demographics per state county in the United States. With such a generous array of data, the objective is to perform a series of analyses that speak to potential influential relationships between a combination of categorical and numerical variables. Particularly, such analyses entail modeling to provide further insight as to whether a trend or correlation exists, and to what degree is such a trend or correlation strong. At the conclusion of each analysis, discussion ensues as to what real-world implications exist as a result of these findings. What do these findings say about a particular county, state, demographic, or perhaps the entire country?

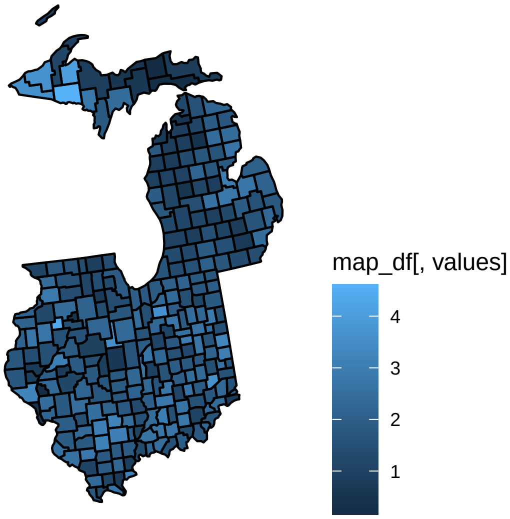
EXPLORING DATA SET

As indicated, this data consists of a rather large array of quantitative variables that speak to various demographic climates for certain states and counties in the United States. The magnitude of this data set incurs large opportunities to gain insight into various trends in demographic climates that can be supplemented by modeling scenarios. At the exploration stage, experimentation will be made with various available variables, with the goal of seeking possible influences.

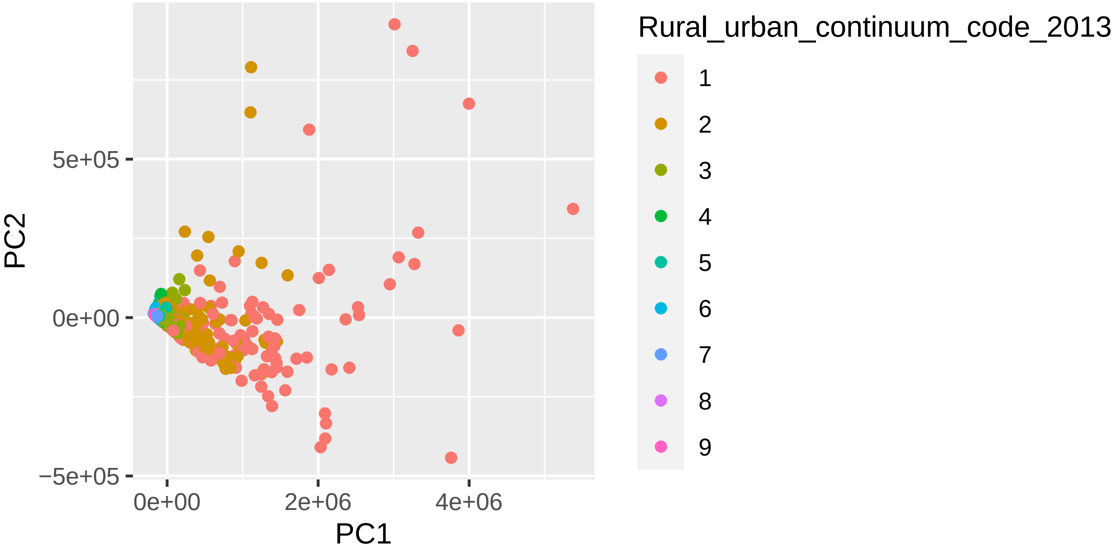
In the midst of the exploration process, certain relationships were implied. When exploring ethnic breakdown in various states with well-known metropolitan jurisdictions, it becomes evident that a more diverse ethnic makeup is consistent in metropolitan areas especially, while rural areas tend to have greater White populations, as demonstrated by the following map, portraying Georgia and Florida. Each locality is colored according to their respective White population percentage. Note the cluster of jurisdictions surrounding Atlanta.



A query was made to determine whether certain outlying localities experience larger counts of COVID deaths. An interesting observation is demonstrated below, where the counties are colored according to their respective COVID-19 death rate per thousand.



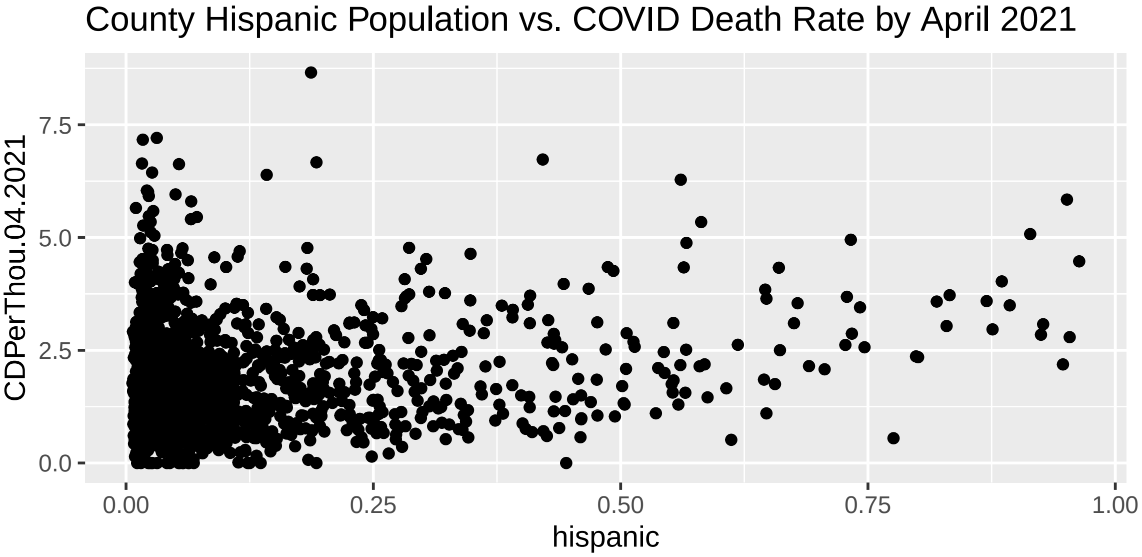
This dataset consists of several different quantitative variables that may influence a variety of other variables. In the midst of the exploration process, principal component analysis (PCA) is used to consolidate these variables, thus condensing the number of columns representing them. The new variables are plotted against each other, colored by a categorical variable indicative of a given county's score on the *Rural Urban Continuum Code*, which speaks to how rural or urban it is, however it is more challenging to make definitive inferences about this particular graph, as it is representative of multiple variables that may not correlate with each other well.



MODELING PROBLEM 1: ANALYZING RELATIONSHIP BETWEEN HISPANIC POPULATION AND COVID-19 DEATH RATE PER THOUSAND IN COUNTIES

In the midst of the exploration process, interest was piqued in regards to the particular relationship between certain ethnic groups and COVID-19 death rates. While there are several quantitative variables representing COVID-19 death rates at certain points in 2020 and 2021, and while there are several quantitative variables representative of different racial backgrounds, we will specifically analyze COVID death rates per thousand for counties with Hispanic populations in the United States as of April 2021 (variables: `hispanic` and `CDPerThou.04.2021`).

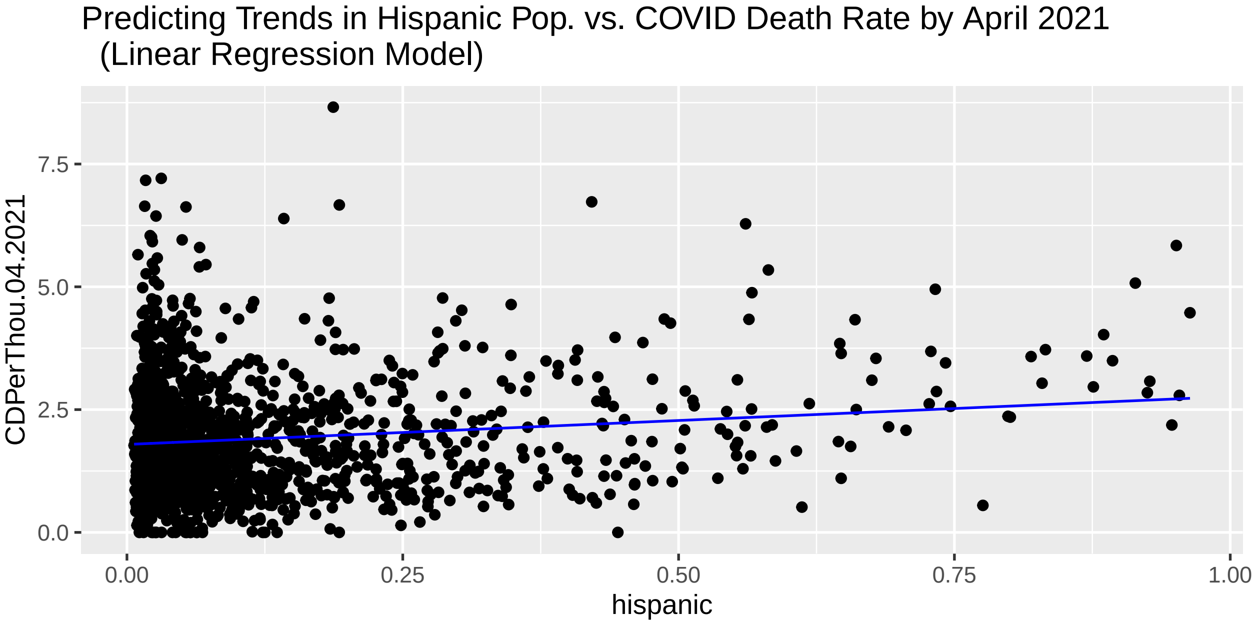
To set the stage for analysis, we visualize the data to determine if there is an existing relationship between a given county's Hispanic population and the COVID-19 death rate per thousand. The following plot provides us with such insight, utilizing data from one of three samples from the original data set that have been extracted for the model prediction series.



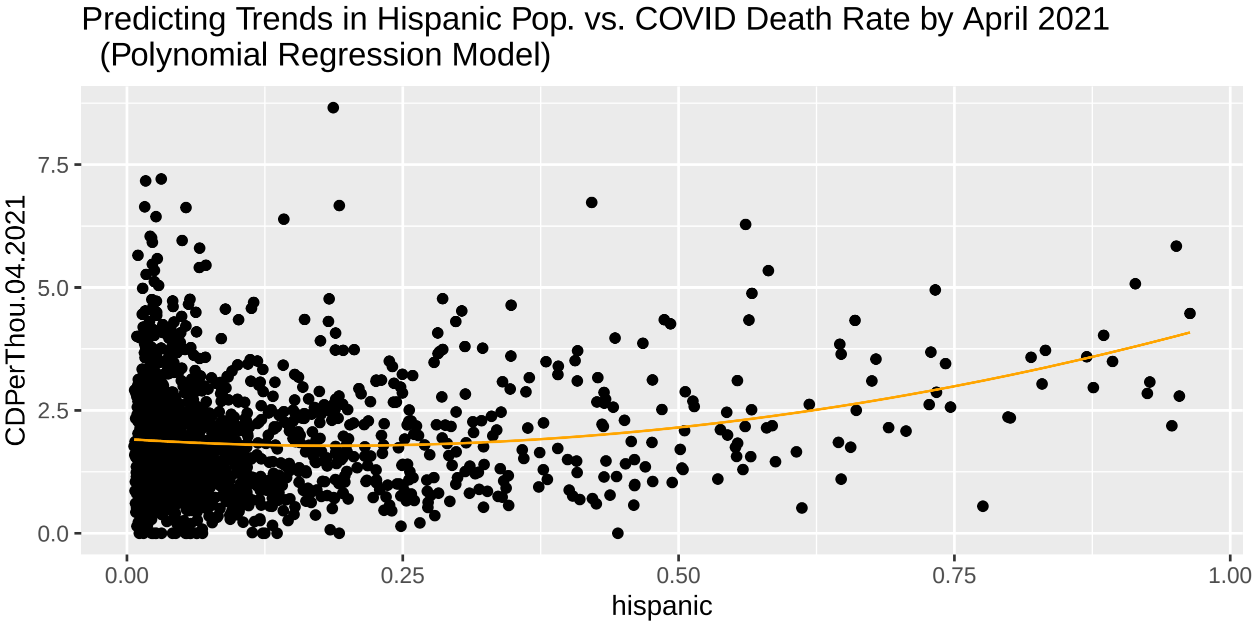
While not necessarily strong, there is an implied correlation between the COVID-19 death rate per thousand and the Hispanic population, such that counties with a larger Hispanic population seem somewhat likely to experience more deaths related to COVID-19. There is a considerably large concentration of counties with smaller Hispanic populations that experience anywhere from less than 1000 to over 3000 COVID-19 deaths, while counties that demonstrate both higher COVID-19 deaths per thousand and a larger Hispanic population occur less frequently. The graph quickly becomes less dense moving to the upper right.

As there appears to be a trend such that counties with larger Hispanic populations may experience higher COVID-19 death rates, a linear regression model is proposed as a potential means to predict COVID-19 death rates given the size of a county's Hispanic population. A polynomial model will be considered as a secondary candidate model, under the belief that any existing non-linear behavior in the graph may be more well-represented. After evaluation of the above graph, it is hypothesized that there is a considerable correlation between the percentage of Hispanics in a given county and the COVID-19 death rate per thousand.

It is acknowledged in advance that such models may not perform very well, considering the general sporadic nature of the data. A trend appears evident, but the data is spread rather liberally, as demonstrated by the above graph. Moreover, the uneven density of the data, such that the majority of observed counties are situated near the middle and bottom left, may contribute further to the limitations of the proposed models.

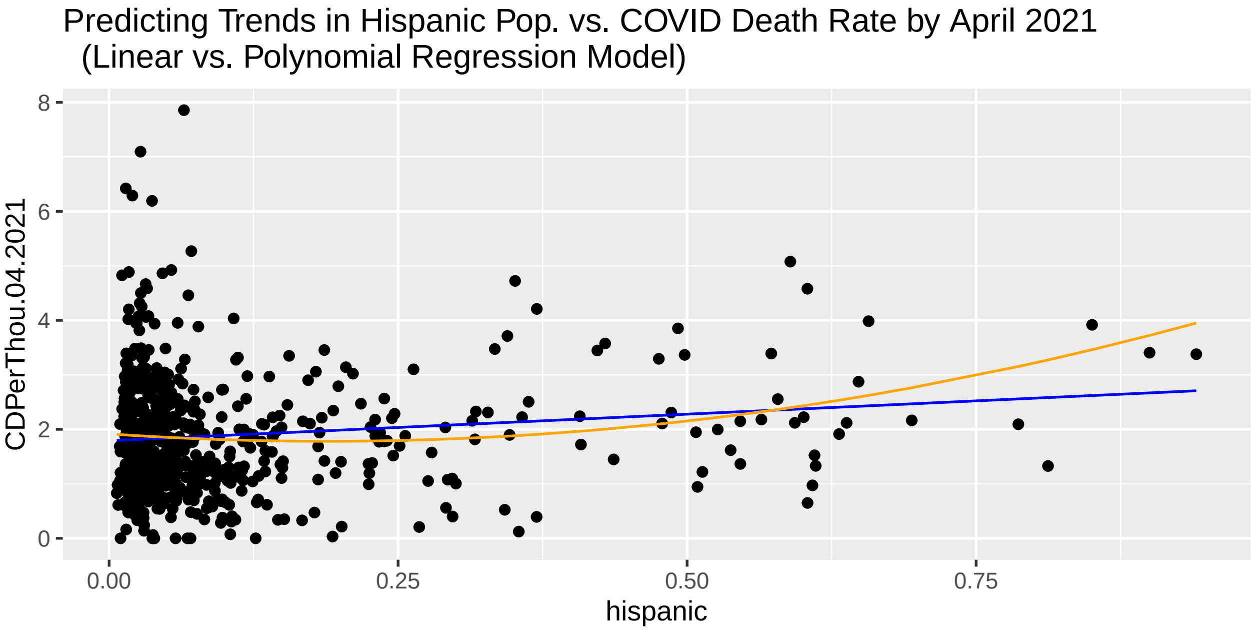


The resulting linear regression line indicates that there is a quite subtle increase in COVID-19 death rates for observed counties with larger Hispanic populations. Notably the line is quite distant from a variety of points on the y avis, serving as further indication that such a model may not be ideal. The line is nearly flat, which may suggest the possibility that this model is not ideal for modeling the correlation, or the correlation may be weak or nonexistent.



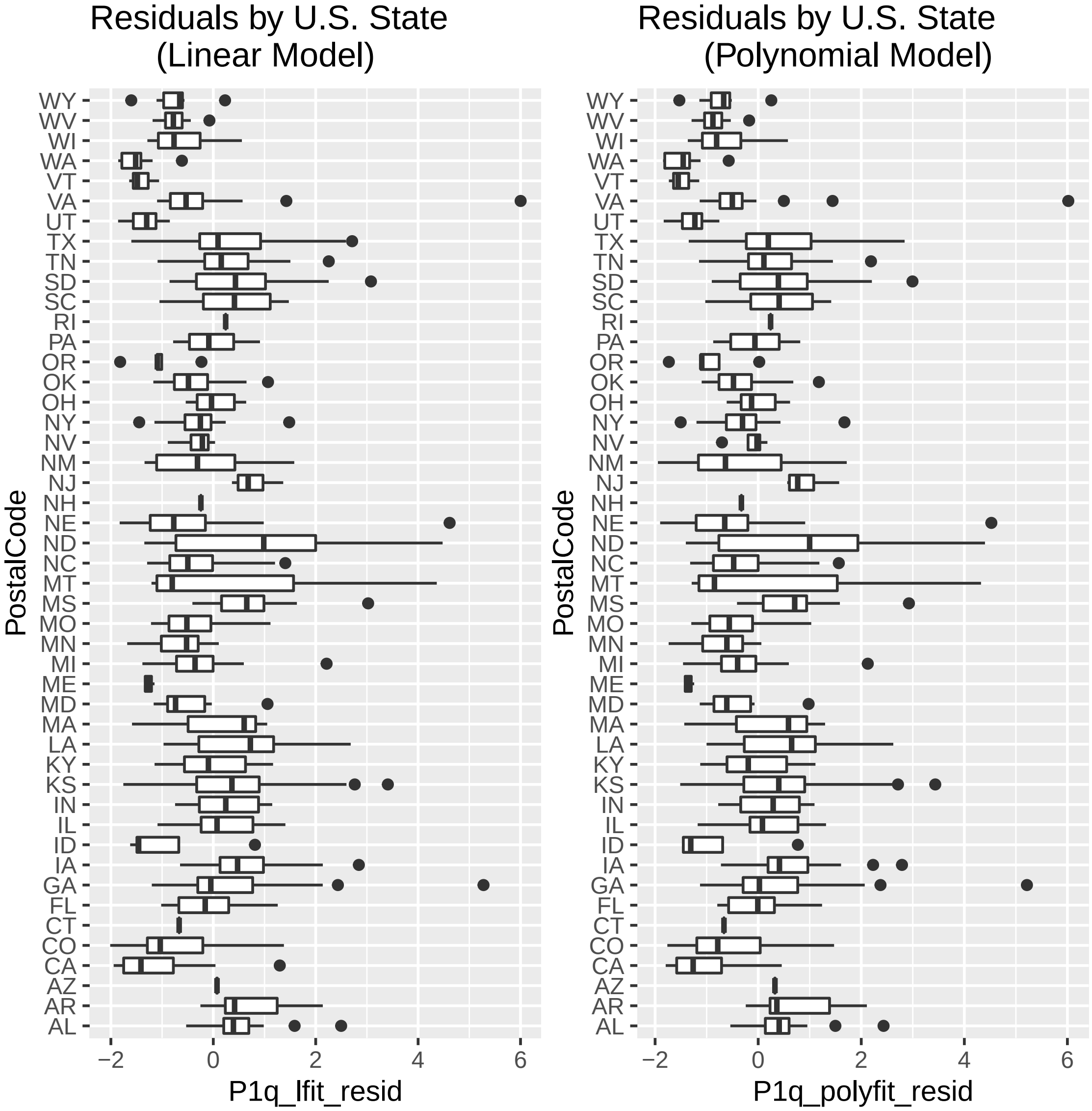
The resulting polynomial regression line suggests, likewise, that there is a subtle correlation between COVID-19 death rates and the size of a county's Hispanic population. The more flexible nature of the polynomial regression line results in its closer distance to certain observations and conforms with the data somewhat more comfortably than the linear regression model. It appears more indicative of a correlation.

Both models were applied to a second sample of data for further investigation. This time, both regression lines are represented in a singular graph for a more convenient and informative visual analysis.


Predicting Trends in Hispanic Pop. vs. COVID Death Rate by April 2021
(Linear vs. Polynomial Regression Model)

It is clearer that both models ultimately predict rather subtle upward trends in the data. With the exception of a significant departure from the linear regression line at circa x = 60, the polynomial regression model does not appear to deviate strongly. Both lines are relatively flat, adding to the sense of doubt that a county's Hispanic populations and COVID death rates are strongly correlated.

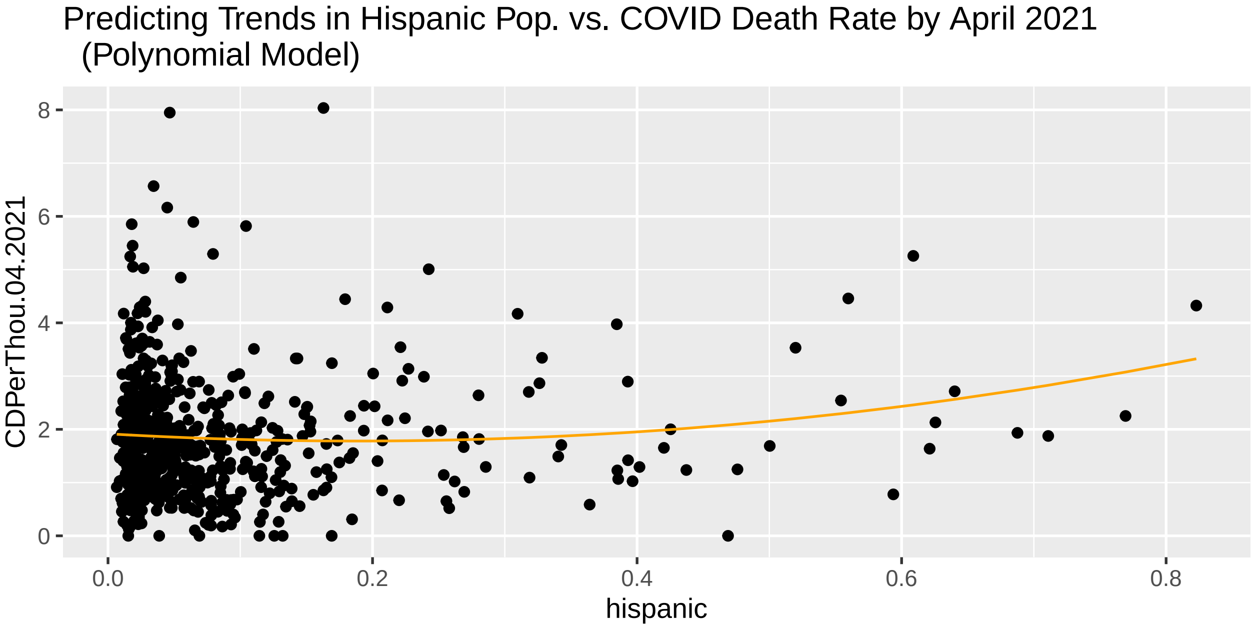
Analyzing the residuals aids in the process of determining which model is superior, or would more accurately predict COVID-19 death rates given a county's Hispanic population. The histograms below demonstrate the degree to which the data deviates from the lines presnted by hte models.



Observing the residuals indicates that there is generally little difference in performance between the linear and polynomial regression models, as distributions per state are relatively similar with regards to the interquartile range, whiskers, and the associated outliers, with few exceptions. Moreover, such similar distributions resemble a rightward skew, indicative that the residuals are not equally distributed and of how both models are not entirely explicative of the trends in the datset. It is now clearer the extent to which both a linear and polynomial regression model present limitations in representing this relationship, as well as the strength of the correlation examined.

Providing further insight is our understanding of the mean and standard deviation. For the linear regression model, the mean is -0.0240626 and the standard deviation is 1.0666524. For the polynomial regression model, the mean is -0.0224292 and the standard deviation is 1.0585083. The difference between the performance of both models is rather negligible in terms of standard deviation from the mean.

In any case, it is observed that in spite of an underwhelming difference, the polynomial regression model undoubtedly performs somewhat better, with a a smaller standard deviation from the mean. For the final sample, we proceed with this model alone. Below is the application of the polynomial regression model to the final sample of the data.



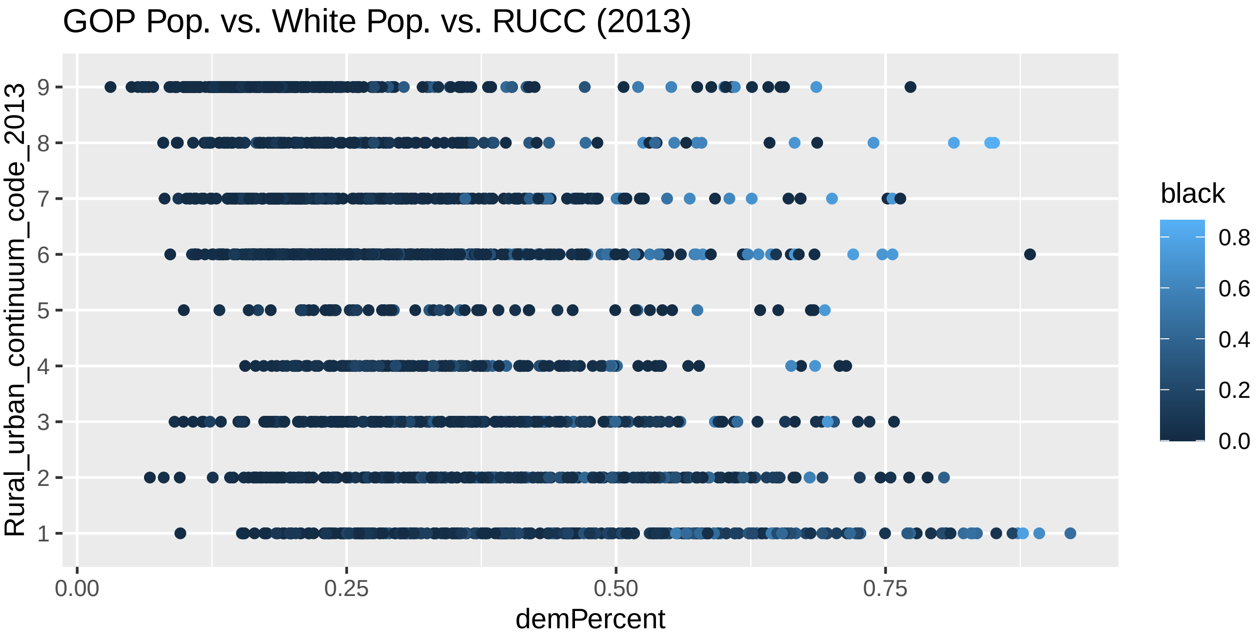
The final model implies that the higher the Hispanic population of a given county, their COVID-19 death rate per one thousand may also be higher. This is a rather weak correlation, which may be argued to be near nonexistent. However, should such a correlation be verified in the future by additional date and means of analysis, stronger findings may be representative of several potential real-life situations, though currently, such observations may be deemed as speculative without further investigation. For instance, a higher Hispanic population in a given county may be particularly indicative that it has a larger general population, which may involve other ethnic groups. A larger COVID-19 death count could be the result of a denser population that may have an effect on the ability to socially distance in a way that prevents transmission of the deadly disease.

Another possibility is that ethnic groups considered minorities are more likely to die from COVID-19 than those who identify as white, which may be explicative of the larger concentration of observations in the graphs near the bottom left, representative of counties with smaller Hispanic populations. An article from KFF.org, *COVID-19 Cases and Deaths by Race/Ethnicity: Current Data and Changes Over Time*, deduces that Hispanic people "represent a larger share of cases relative to their share of the total population (24% vs. 18%), while their share of deaths is more proportionate to their share of the population (17% vs. 18%)" [1].

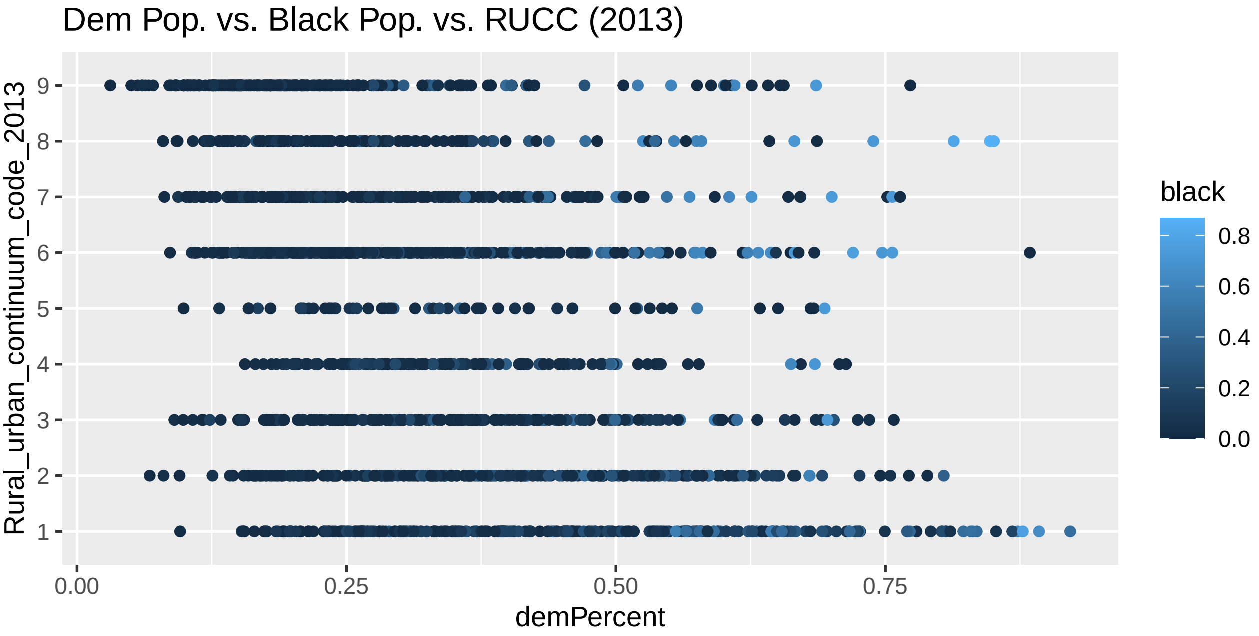
MODELING PROBLEM 2: ANALYZING RELATIONSHIP BETWEEN DEMOGRAPHIC BREAKDOWN AND HOW RURAL/URBAN COUNTIES ARE

With a more fixed amount of categorical variables provided in the data set, interest was also piqued into the relationship between certain demographic information about political and racial alignment and the density of a particular area as implied by the *Rural / Urban Continuum Code*. More succinctly, an inquiry is in order as to whether it can be determined if an observed county is rural or urban to certain extents, based on available data on Democrat/Republican presence and votes, as well as racial groups. Are White Republicans more likely to reside in rural areas? Are Black or Hispanic members of the Democratic Party more likely to reside in urban spaces? Does another combination of such identities have a particular predominance in a certain area?

As with the first modeling scenario, visualization of the data helps in the process of determining a concrete hypothesis. This was achieved by plotting several numerical variables against the *Rural / Urban Continuum Code*, as demonstrated below, with the intention of determining the possibility of correlations.



Relationships between several of these variables are implied by this graph. For instance, it appears likely that there is a larger presence of people who identify as Republicans and are racially classified as White in more rural areas. There are fewer rural counties with a population that is hardly white, though White people appear to compose a large amount of the population in many counties. Urban counties with a smaller Republican population appear to have smaller White populations. The following graph provides further insight.



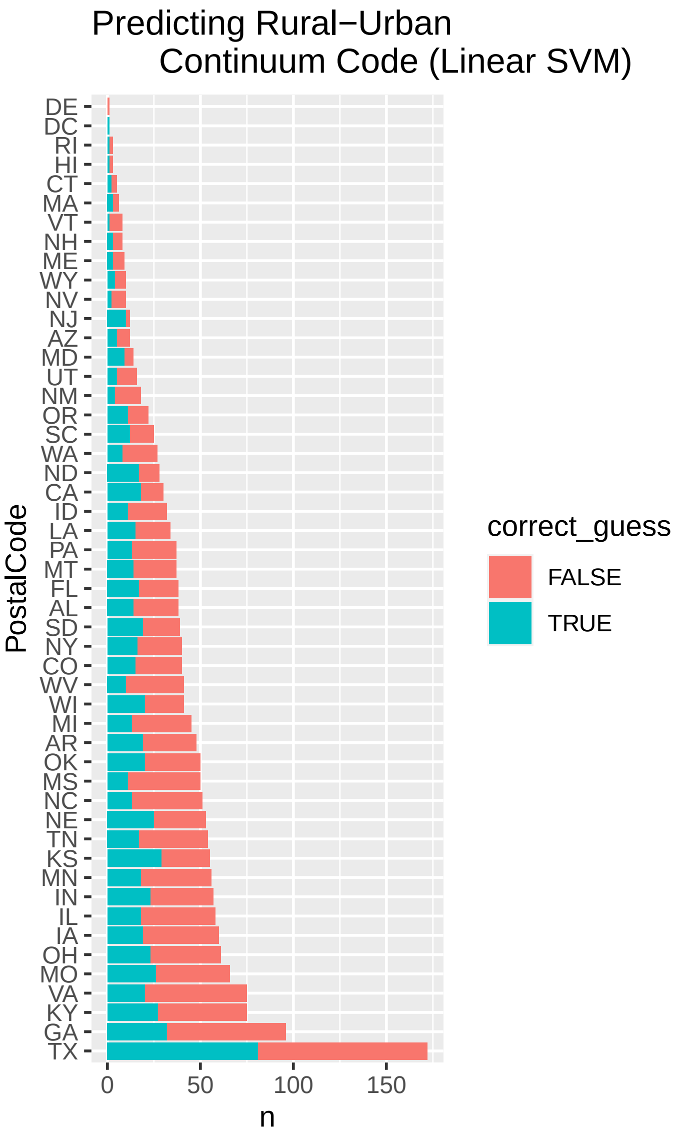
This graph supplements what was previously displayed, incurring the inference that rural areas typically have a lower population of Democrats with a few outliers, while that number subtly increases when moving lower on the index. Many counties, interestingly, have few Black populations as well.

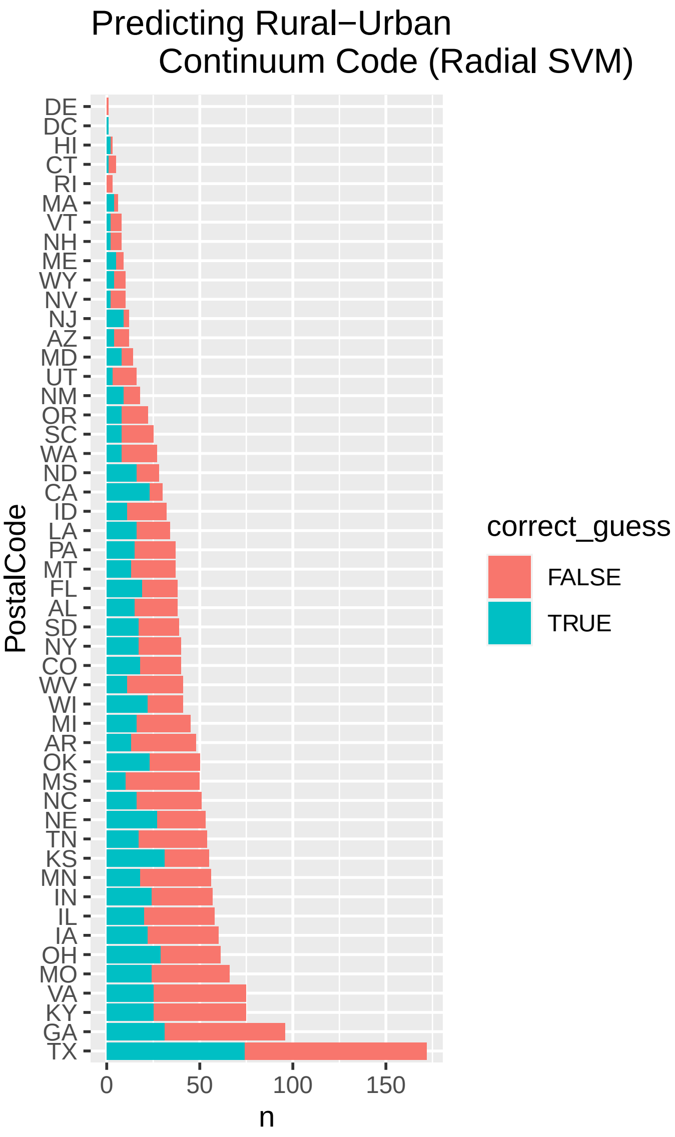
We wish to determine a model suitable for predicting a county's respective value on the *Rural / Urban Continuum Code* to gain further insight into the trends in more rural counties versus more urban counties with respect to various demographic factors. Support Vector Machines (SVM) are the most appropriate means with which to perform such modeling. The above graphs suggest a possible trend in that Democratic presence in more rural counties is less likely and that racial diversity is rarer in more rural counties. We propose a linear SVM model and a radial model which may accommodate the nature of such apparent correlations, in that a larger number in one demographic categorical variable may result in a higher score on the *Rural / Urban Continuum Code*.

The following variables are observed in the testing stage:

- `TOT_POP` - Number of total population for a given county
- `gopPercent` - Percentage of citizens in a given county who identify as GOP/Republican
- `demPercent` - Percentage of citizens in a given county who identify as Democrat
- `black` - Percentage of citizens in a given county who identify as Black
- `white` - Percentage of citizens in a given county who identify as White
- `hispanic` - Percentage of citizens in a given county who identify as Hispanic
- `votes_gop` - Percentage of GOP/Republican votes for a given county in the 2020 Presidential Election
- `votes_dem` - Percentage of Democratic votes for a given county in the 2020 Presidential Election

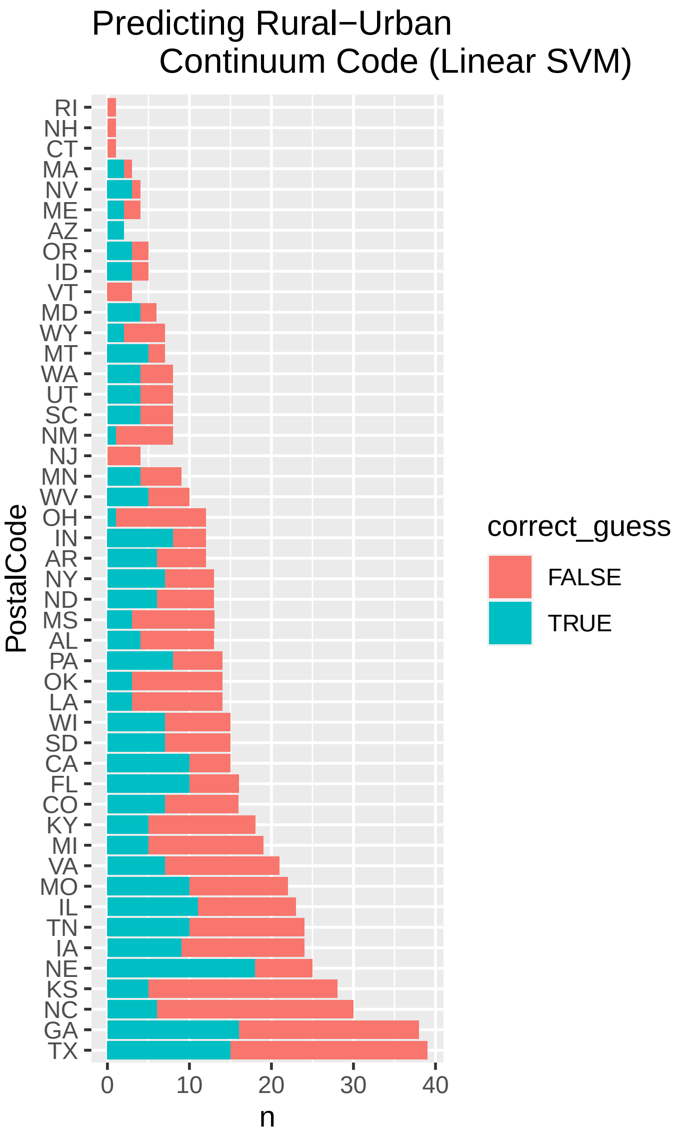
Below is a graph portraying the proportion of predictions made by the polynomial and radial SVM model that are correct, versus which predictions were incorrect.

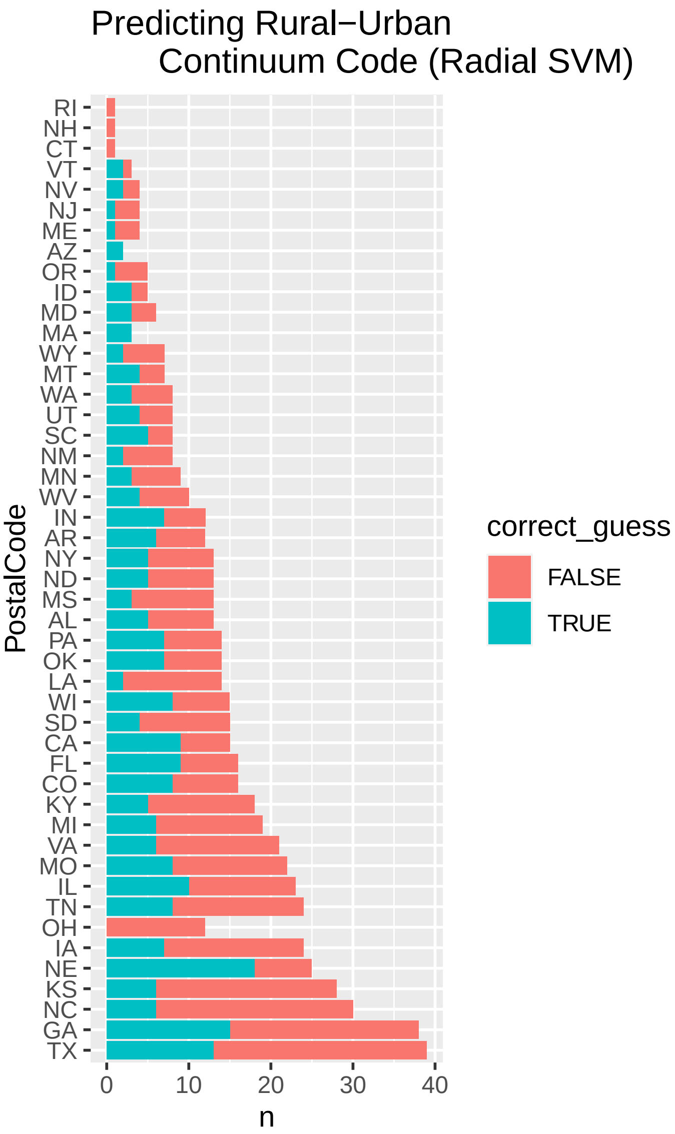




Visually, it is evident that both models performed rather similarly, with immediate differences being that the linear model had signifcantly more correct guesses for the state of Texas than the radial model. The percentage of correct guesses for the linear model is 0.5 and on average, the radial model had a success rate of 0.4948454.

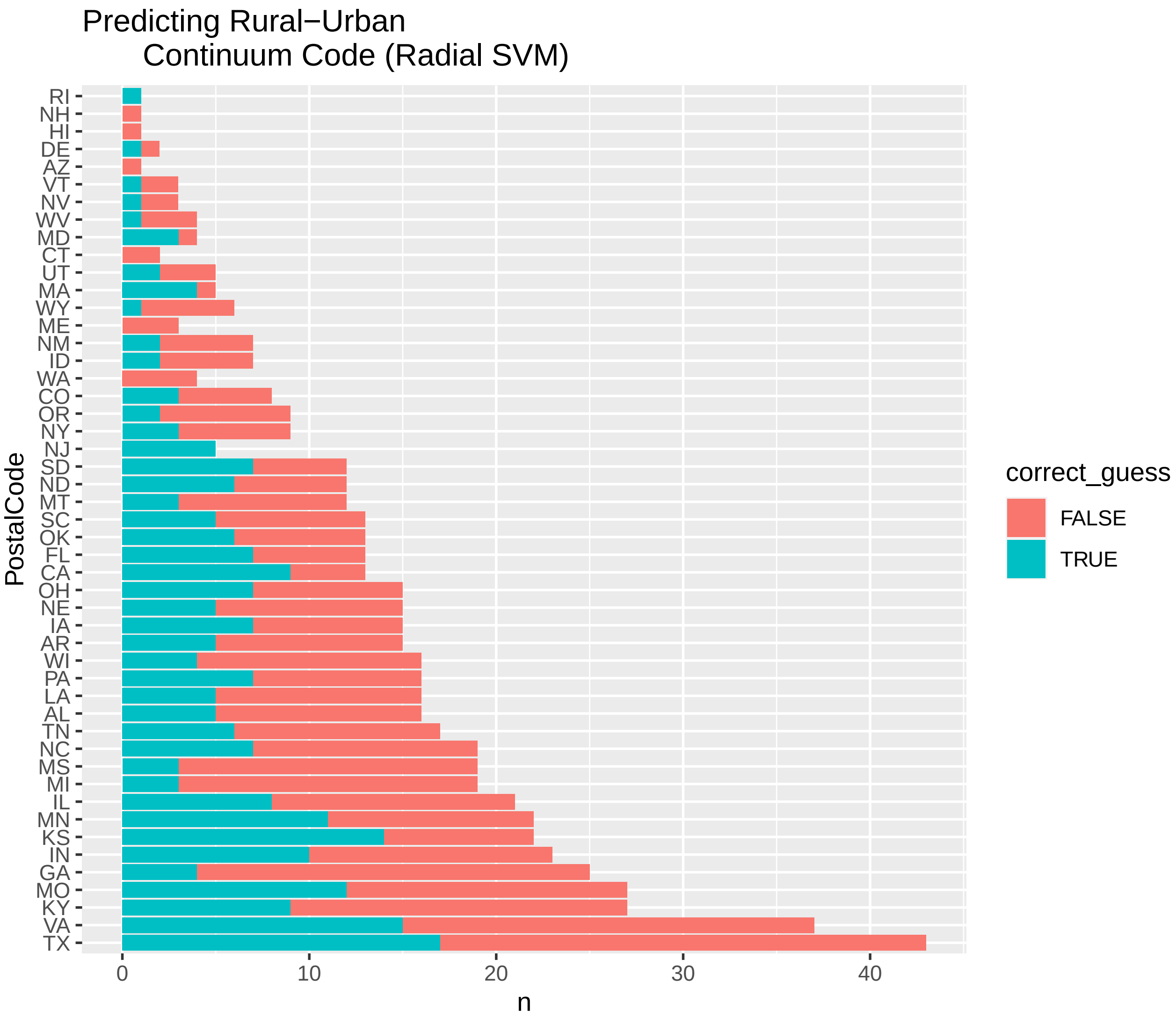
The model is run again, utilizing data from a second sample to observe any potential changes. The results are as follows:





Visually, it remains that both models are difficult to distinguish in terms of performance, with the oustanding difference being the state of Texas, where the linear SVM model accomplishes more correct predictions. It appears that while correct guesses are higher for the linear SVM for some states, they are lower in other states where the linear SVM is at an advantage. There is a sense of inversion between which states are correctly guessed by which model. Notably, both models accomplish 0 correct guesses for at least two states, which is also see in the execution of the model for the previous sample. The percentage of correct guesses for the linear model is 0.4772727 and on average, the radial model had a success rate of 0.4886364. Akin to the difficulty in determining a superior model due to negligible differences in performance in the first modeling problem, we experience an issue in deciding which of these two models to perform. It trends that the radial SVM model has a small edge over the linear SVM model in predicting a county's score on the *Rural / Urban Continuum Code*. In such case, the radial SVM model will be selected for the testing stage.

The results of the radial SVM model's performance at the testing stage are as follows:



The final radial SVM model resembles what was evident in the prior two, in that the number of incorrect observations generally outweighs the number of correct guesses, indicative that this model is not configured in an optimal way, or that the quantitative variables considered in the model may not have a correlation with the *Rural / Urban Continuum Code* score, collectively and/or individually. At best, this modeling scenario provided a rudimentary level of insight, considering that very few instances existed where either model accomplished more correct guesses than incorrect guesses, and the percentage of correct guesses was roughly 50% on average. Further investigation is required to arrive at a more definitive answer as to which quantitative demographic variable is strongly related to how rural or urban a county is considered.

As far as interpretations are concerned, several sources speak to the inference that localities that are mostly rural are more likely to be more highly comprised of people who identify as White and people who identify with the Republican party. Specifically, the United States Department of Agriculture (USDA) states that 2018 findings indicate that Whites comprised 78.2% of the rural population and 57.3% of the urban population, whereas minority ethnic groups like Hispanics and Blacks only made up 8.6% and 7.8% of rural areas respectively, while making up 19.8% and 13.1% of urban areas, respectively [2]. Moreover, the 2020 election results demonstrate higher concentrations of Democratic votes in more urban, metropolitan jurisdictions in comparison to more rural areas, where Republican votes are greater. CityMonitor demonstrates with a linear regression model that more densely-populated areas “swung” to Joe Biden as a candidate of choice for the election, as the urban/rural divide apparently continues to heighten [3].

REFERENCES

[1] KFF, *COVID-19 Cases and Deaths by Race/Ethnicity: Current Data and Changes Over Time.* https://www.kff.org/coronavirus-covid-19/issue-brief/covid-19-cases-and-deaths-by-race-ethnicity-current-data-and-changes-over-time/

[2] U.S. Department of Agriculture, *Racial and ethnic minorities made up about 22 percent of the rural population in 2018, compared to 43 percent in urban areas.* https://www.ers.usda.gov/data-products/chart-gallery/gallery/chart-detail/?chartId=99538#:~:text=In%202018%2C%20Whites%20accounted%20for,19.8%20percent%20of%20urban%20areas

[3] CityMonitor, *The urban-rural divide only deepened in the 2020 US election.* https://citymonitor.ai/government/the-urban-rural-divide-only-deepened-in-the-2020-us-election