Lab 2 Report

Alexandra Drossos, Andres de la Rosa, Linh Nguyen

August 3rd, 2021

1. An Introduction

As we all know, the COVID-19 pandemic has caused tourism to our city of Portlandia to wane. We need a marketing strategy to attract more visitors for the latter half of 2021 and end this current period of economic stagnation. Over the past month, we have been working on behalf of the head of Portlandia's tourism agency to provide data-driven insights into which counties within the US we should target most of our marketing efforts towards. The broad question we're looking for answers to is: What characteristics of a county's population have an impact on their willingness to travel and go on vacations this summer? If we can answer this question, we can specifically target those counties with favorable attributes for our tourism campaigns. However, firstly, the concepts within this question have to be operationalized to enable data collection. To measure a county's "willingness" to travel, we'll be looking at the number of trips their residents have taken per capita in June 2021 that are farther than 100 miles away. The second concept we need to operationalize for data collection and analysis is what characteristics of a county's population we'll be looking at. The first county metric we'll be using as an independent variable is county vaccination rate. This variable is appropriate for the intended purposes of our analysis because vaccinated individuals may feel more comfortable traveling and therefore more inclined to vacation. However, these are just initial expectations and may not be reflected in the data. As we iterate on the original model, we'll add in additional covariates as further explanatory variables. However, our primary research question asks: How does a county's vaccination rate as of June 1st affect the number of trips their residents have taken per capita that are farther than 100 miles away in the month of June 2021? These variable specifications will be rationalized as we go through the model building process.

2. A Model Building Process

As described in the introduction, our goal is to explain how certain characteristics of a county's population affects their willingness to travel. This information will be used for Portlandia's tourism agency to enable more accurate audience targeting for their summer tourism promotions. To help answer this research question, we will create a series of explanatory models and analyze their results.

The output variable in each of the models will represent the concept of a county population's willingness to travel throughout this summer. To measure this we will use data from the U.S. Department of Transportation Bureau of Transportation Statistics. This dataset provides us with the number of trips taken each day at different aggregation levels (i.e. National, State, County). We chose to use the county aggregation level, because using state level data would not produce enough data points to allow us to use a large sample model. Given this, we filtered the data on the county aggregation rows. In addition, we also only want to look at trips between June 1st and 31st 2021. This is simply due to the fact that our objective is to measure summer traveling willingness, so other months of trips data would likely have varying distributions that are not relevant. We will also filter on these date conditions. In addition, we will only be counting trips that are greater than 100 miles away from the traveler's home. This is because Portlandia lacks major neighboring cities, so our data needs to represent only individuals that are willing to fly to their vacation destination. This is represented in the data by separate columns listing the number of trips taken between certain intervals of distance: <25, 25-50, 50-100, 100-250, 250-500, and >500.

To create the output variable we need, we summed the counts in the columns representing Number of trips 100-

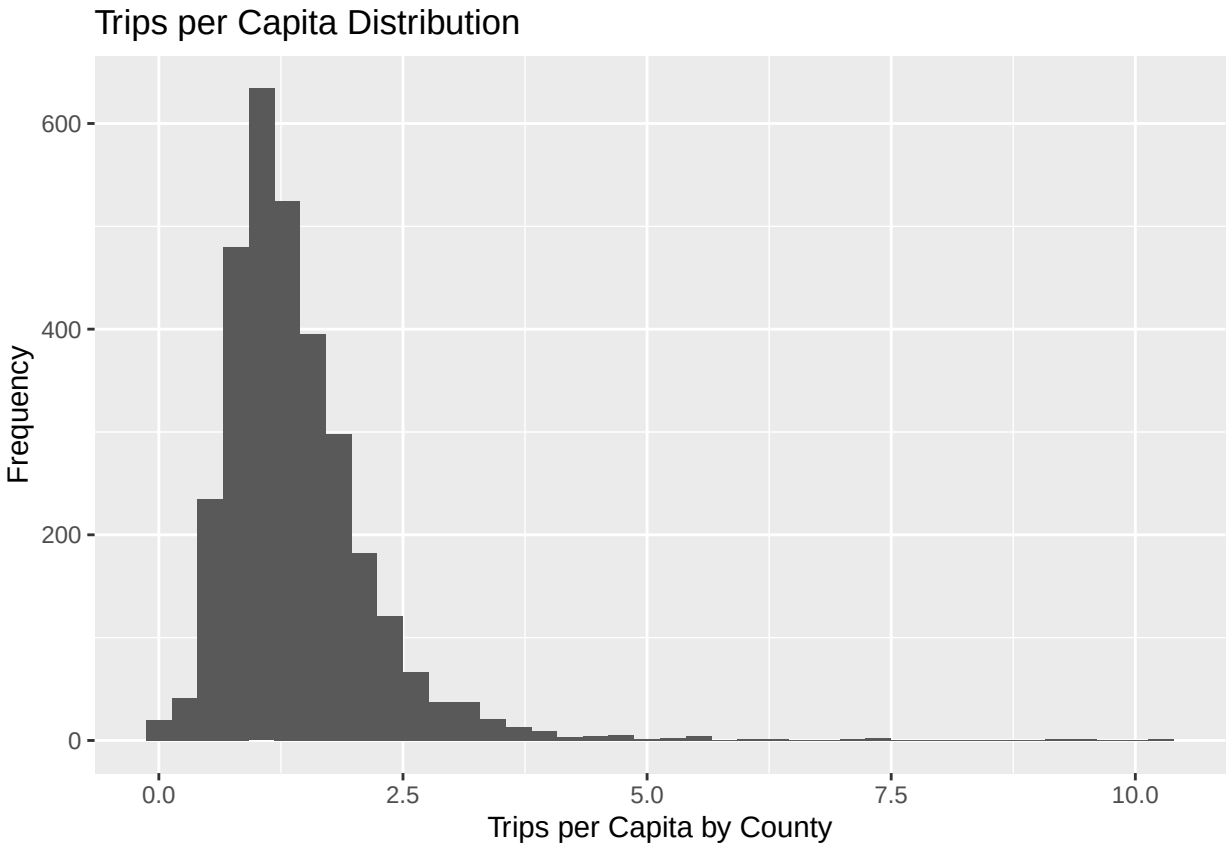
250, 250-500, and >500. The last transformation we need to make is to normalize the variable. As it originally comes, we're only seeing the number of trips within a certain county, so we need to divide each of these county values by their respective population for the outcome variable to become number of trips *per capita*. Thankfully this population data was included within the dataset as two columns: Population.Staying.at.Home and Population.Not.Staying.at.Home. To create the per capita figure we grouped the dataset by county and summed up the total number of trips in June 2021 divided by the total population of the county. Some summary statistics of this data are shown below:

Table 1: Trips per Capita EDA

Min	Mean	Median	Max
0	1.407144	1.26253	10.26435

One thing to note here is the "FIPS" column. Federal Information Processing Standards (FIPS), now known as Federal Information Processing Series, are numeric codes assigned by the National Institute of Standards and Technology (NIST) to uniquely identify geographic areas. In this case, it's referring to county FIPS code. This will be the column we use to join together our variables for all three models we will produce.

From the sample output and EDA tables, you can see that the value of sum_of_trip ranges from 0 to ~10. Because there is such a big difference between the mean of 1.4 and the maximum value of the column, we can extrapolate that there are some outliers in the data. To get a more clear picture of what this variable really looks like, we'll have to look at the underlying distribution by using a histogram.



The distribution of our output variable, shown above, looks to be relatively normal. As predicted, there are some outlying counties, as shown by the maximum value of trips per capita figure being around 10 while the mean is 1.4. However, the majority of the values lie between 0 and 2.5 with a relatively normal distribution.

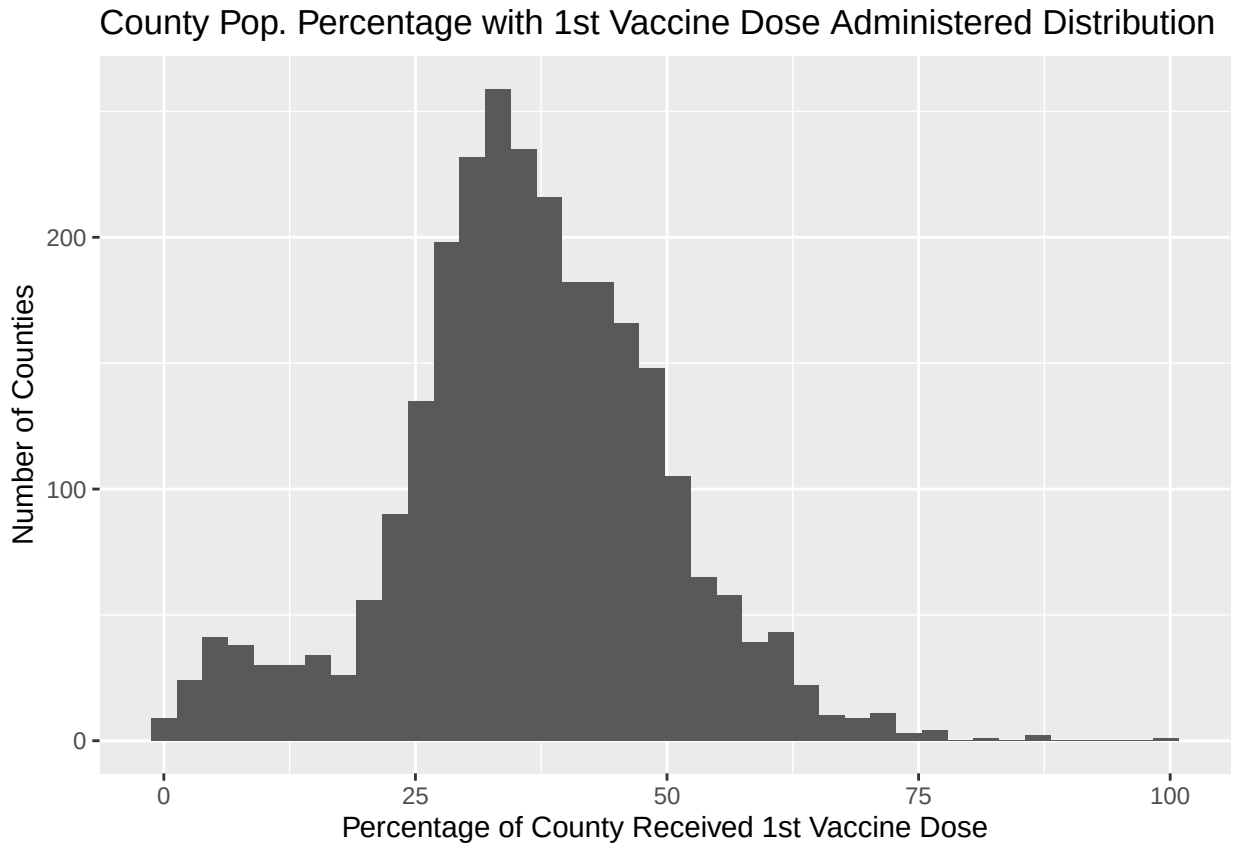
While we do already have more than enough data points to use the large sample model, this underlying normality of the variable is helpful for any assumptions sake.

Now that we have described what we're going to be measuring, we can begin the process of creating our models.

1. **Limited Model**: Our first model, the limited model, will include only the key variables we want to measure, the minimum number of covariates needed to provide some insights on what effects our output variable. For this model, our only input variable will be county vaccination rate.

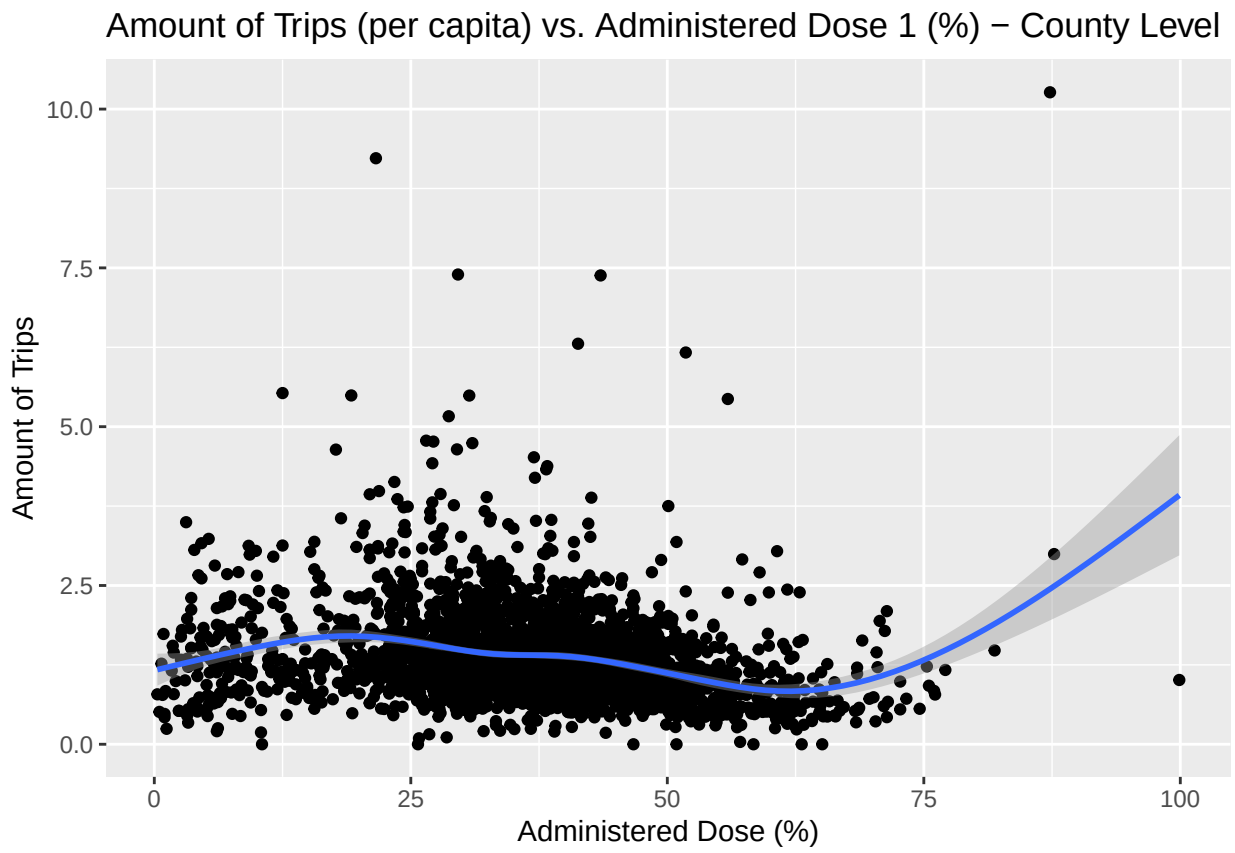
The CDC was our trusted source for this variable; a time series dataset tracking each county's progress in vaccination rate by date since December 13th, 2021. For this research, because we're only looking at trips taken between June 1st and June 31st 2021, we're concerned with a county's vaccination rate as of June 1st, 2021 to see the full casual impact of the vaccination rate on the number of trips taken after the fact. The only other specification we proposed was to look at the 1st dose administered rate rather than both boses administered. This is due to the fact that counties across the country has varying availbility of the vaccine that was preventing some from getting it. By June 1st, it was available to anyone 18 years or older, but some may still have been waiting for their second dose. No real transformation needs to be applied to this data, because its format as is in the dataset is sufficient for our modeling purposes. We just have to filter the dataset on vaccination rate by county as of June 1st. From there we can look at the range, mean, and median of the data for some sanity checks.

From this table we can see that the minimum value for the percentage of a county's population that's received their first vaccine dose is 0. After looking through the CSV file, the 0 values are associated with counties they don't have adequate information on, so these need to be filtered out. After doing so, the new minimum value is 0.3% and the maximum is still 99.9%. With a mean value of 31%, this means that the data is certainly not centered, but we can assess that closer by producing a histogram of the cleaned data's histogram.



The cluster of counties with population percentages between 0-20% is producing the skew in this distribution.

However, apart from that and some of the outlier counties to the far right of the distribution, it is relatively normal. For our limited model creation, we need to look at the relationship between the vaccination rates as of June 1st 2021 and the number of trips taken in June for each county in the US. To accomplish this, we had to join the datasets together on the FIPS code and then scatter plot the variables against eachother.



From eyeing this, the sum of trips figure remains fairly constant across the values of vaccinated population percentage until you get to around 60%. Performing a log transformation on both variables did not create a more clear linear relationship between them, so we did not have the justification to make any transformations to either varialbe in the limited model. Although this scatter plot didn't produce a clear picture of the nature of the relationship between the variables, we'll have to run linear regression and see the output of the model to determine if there's any significance.

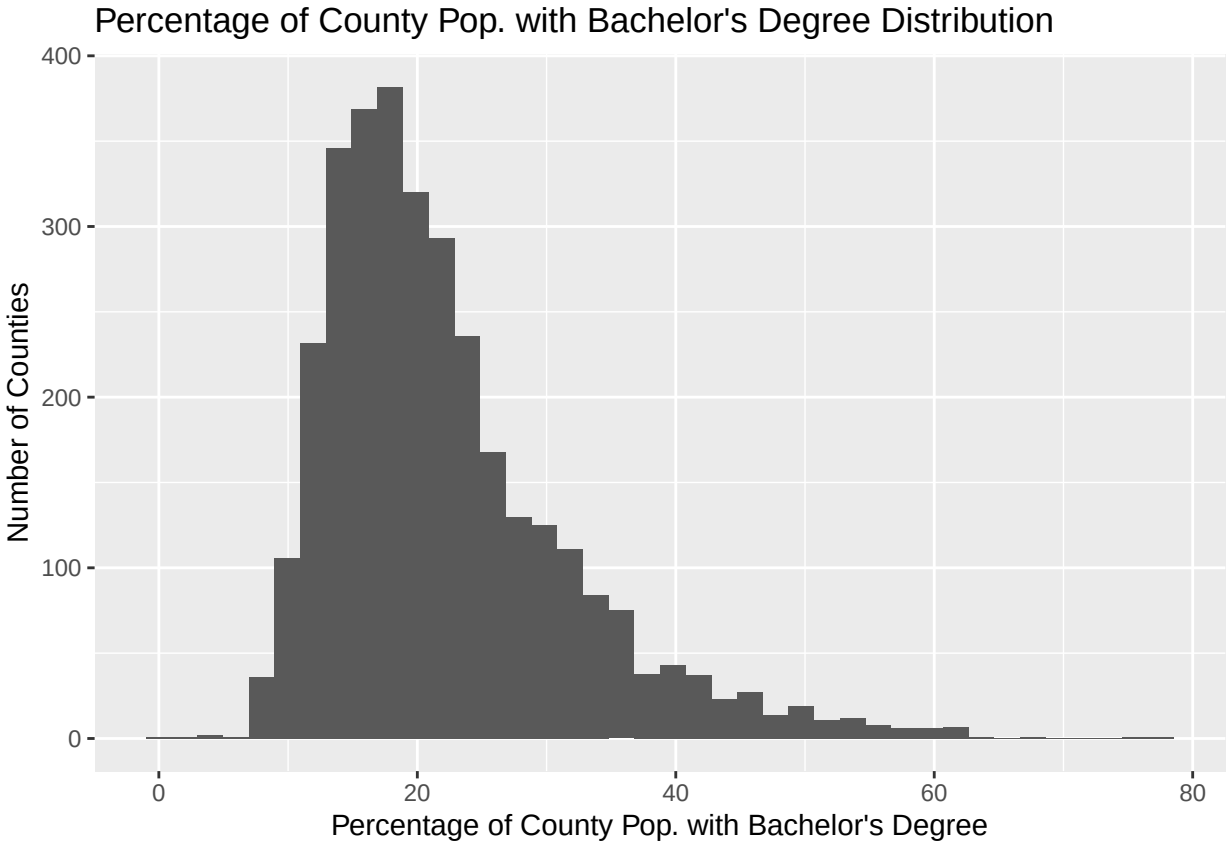
After running this linear regression model, the resulting pval we got was 2.2 * 10ˆ-6, which means that the vaccinate rate variable we used has a significant effect on the total trips per capita taken in that respective county. While this result is significant, there are other relevant covariates that we will want to add to help us reach our modeling goals.

2. **Model Two**: For our second model, we will include some additional covariates to help explain the output variable, total trips per capita for a specific county. The covariates of interest for this model are the percent of adults with a bachelor's degree or higher, median household income, and change in unemployment rate between 2019 and 2020. These covariates will add to our understanding of the current demographics of a county after COVID effects have settled.

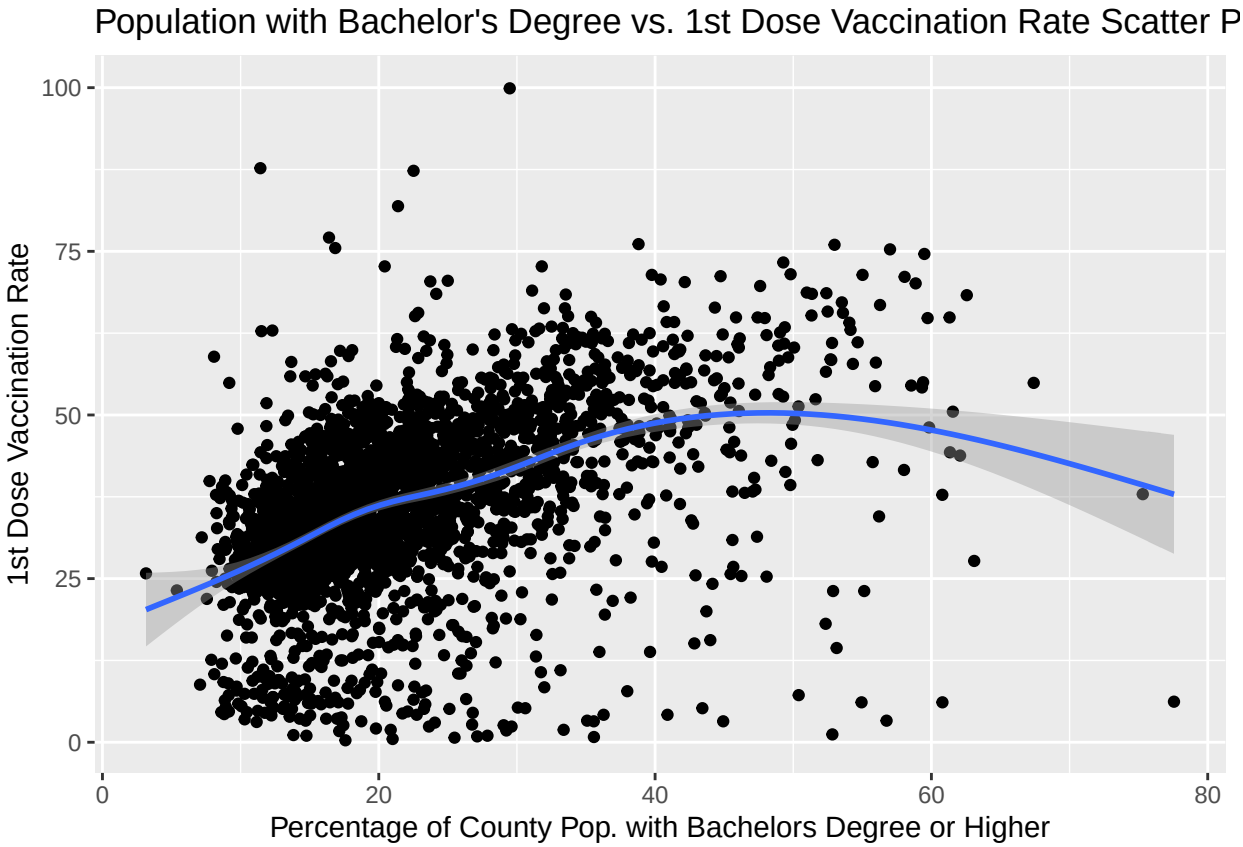
The first additional input variable we'll explore is education. More specifically, we added the percentage of adults in a county with a bachelor's degree or higher as of 2019. This specification is relevant because of the correlation between higher education and wage, which would impact someone's ability to be able to go on vacations. We first must perform some basic sanity checks on the data to ensure good quality.

Table 2: Percentage of County Pop. with Bachelor's EDA

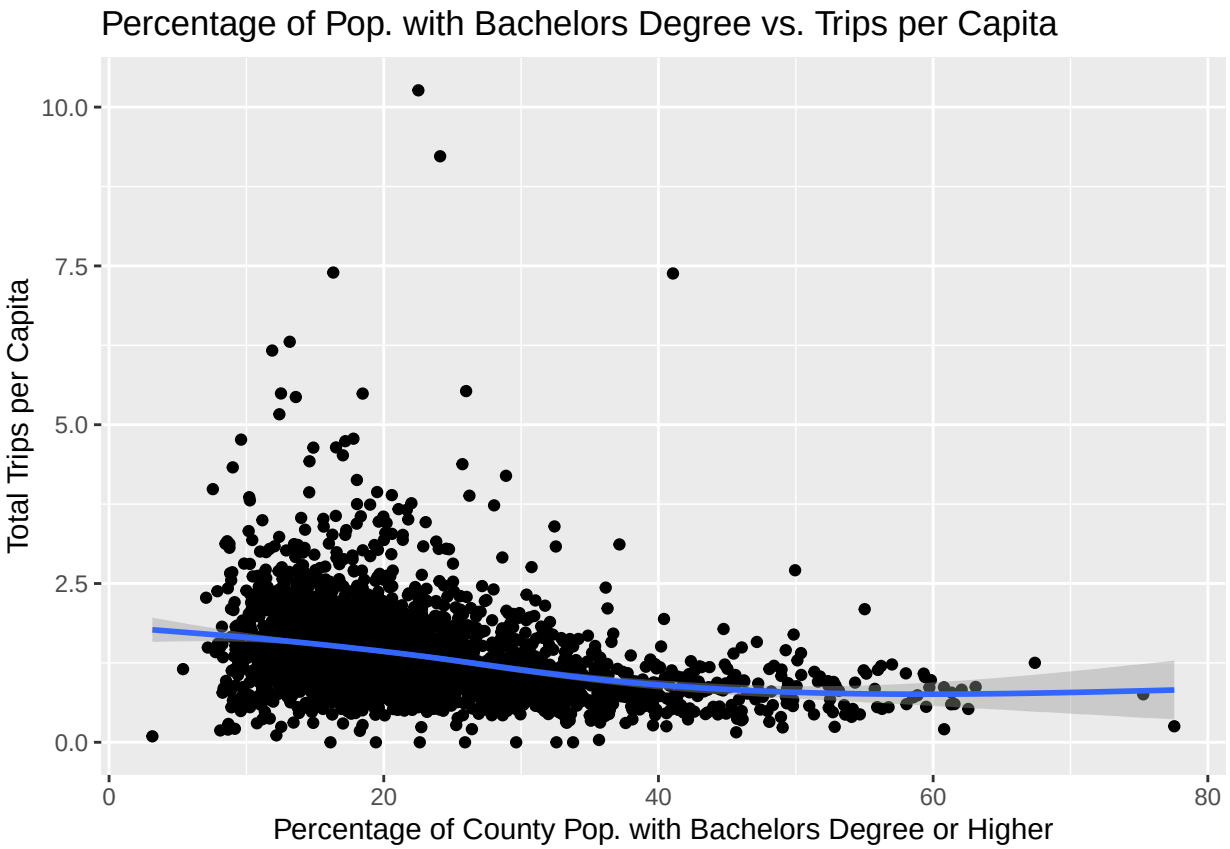
Min	Mean	Median	Max
0	22.12556	19.77686	77.55741



This distribution of the percentage of county population with a bachelor's degree or higher has a wide range with right skew. The mean value is around 22%, which is in itself an interesting finding for the country's education distribution by county. Relative to forming this model though, the underlying distribution is normal enough for us to feel comfortable including it in the model. In addition, after looking at the scatter plot below between the percentage of population with bachelor's degree vs. 1st dose vaccination rate by county, we can see that there isn't a strict linear relationship between the two, so add this new variable into the model will not introduce too much multicollinearity.



Lastly, we need to observe the relationship between our new input variable, Percentage of Population with Bachelor's Degree or Higher, and the output variable, total trips per capita by County. To do this we can run a scatter plot and geom_smooth as we did above with the vaccination data.



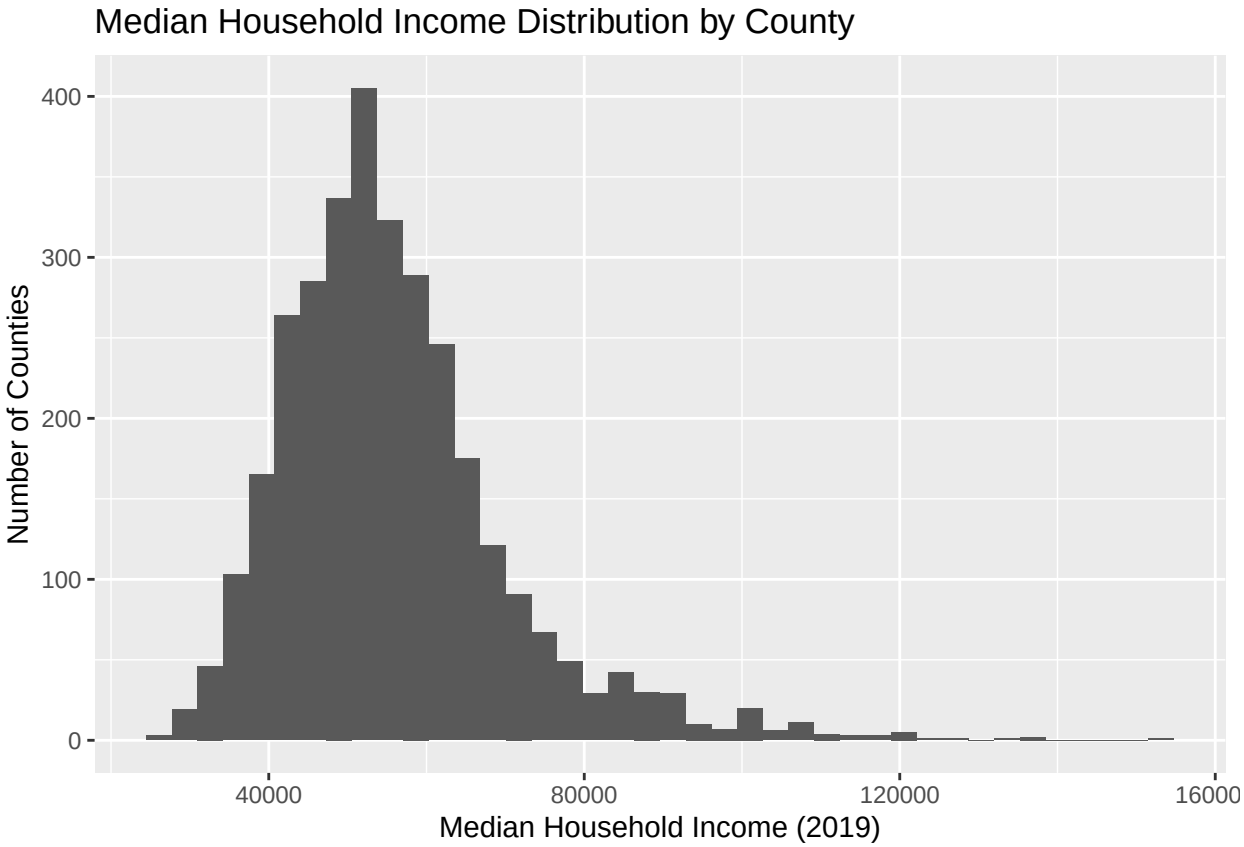
We can see that there appears to be a fairly linear relationship between these variables, with the number of trips per capita decreasing as the percentage of a population with bachelor's degree grows. We will confirm that this variable is significant once we run linear regression on the models and check the resulting pvalues, but for now we can feel comfortable including it. Following these plots, we did some research into the counties on both sides of the spectrum. We saw there was one county with 99.9% of people with a Bachelor's degree or higher, and that was Falls Church, VA. Interestingly, we found this information on Wikipedia: The median income for a household in the city was $120,000, with 4% of the population below the poverty line, the lowest level of poverty of any independent city or county in the United States. There seems to be, in the case of this county atleast, a correlation between education level and poverty.

We will have to perform the same EDA to be sure, but we decided from this to then include median household income as the next input variable in our model 2. This data was pulled from the Economic Research Service as part of the U.S. Department of Agriculture. The most recent year that it provided data on was 2019, which is sufficient for our purposes as we'll explain later on as we continue to build out the model.

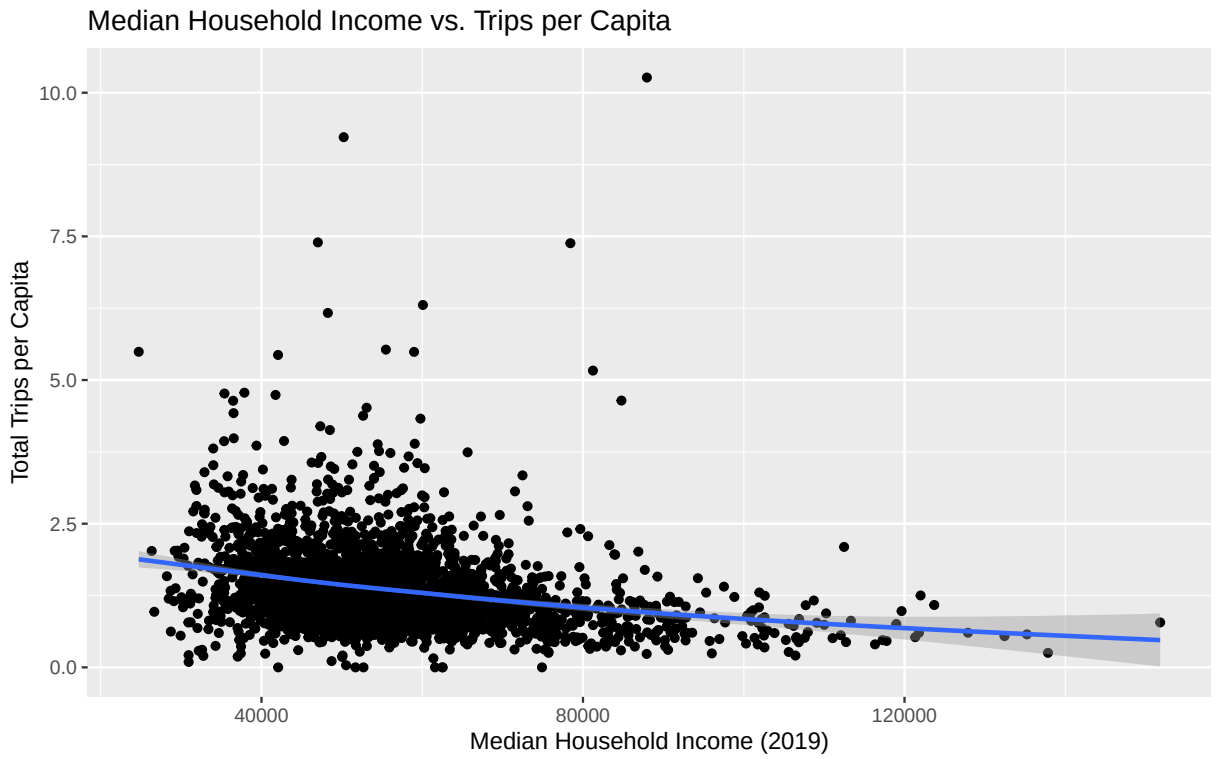
Table 3: Median Household Income EDA

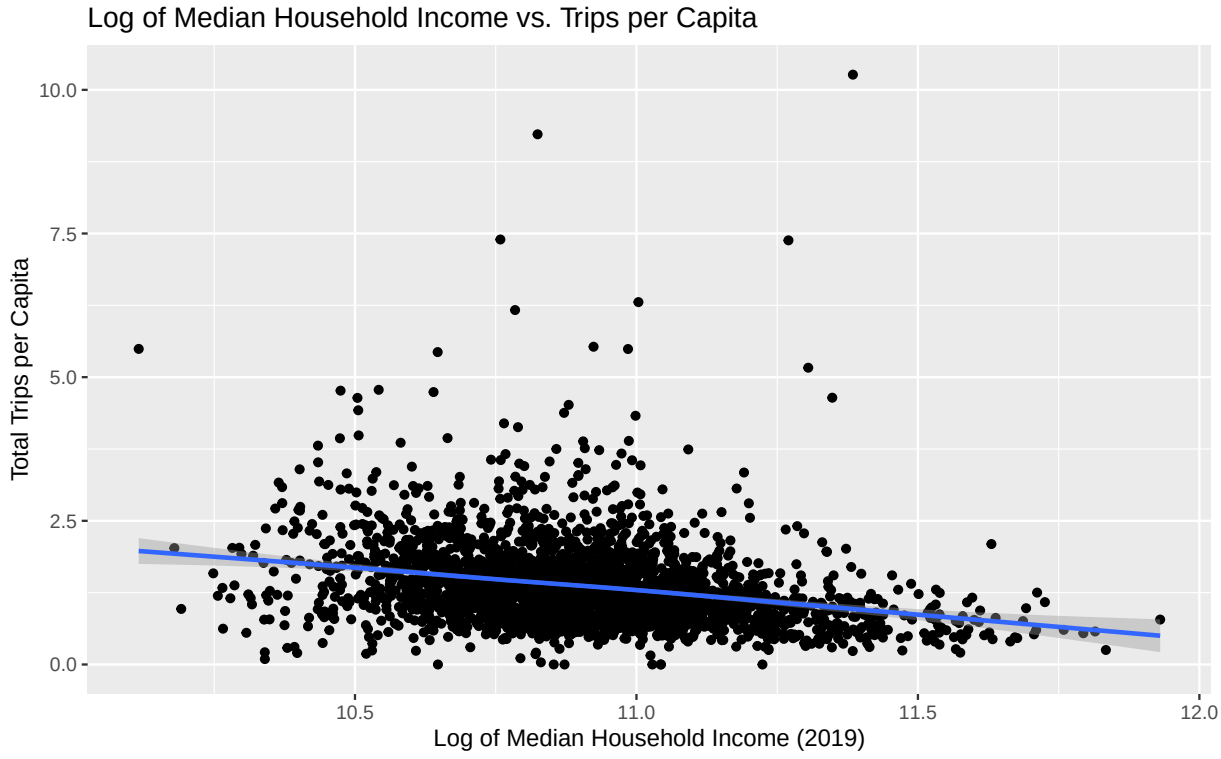
Min	Mean	Median	Max
24732	55874.76	53505	151806

From pulling these summary statistics from the median household income data, there don't look to be any data entry errors, so we can go ahead with plotting the distribution.

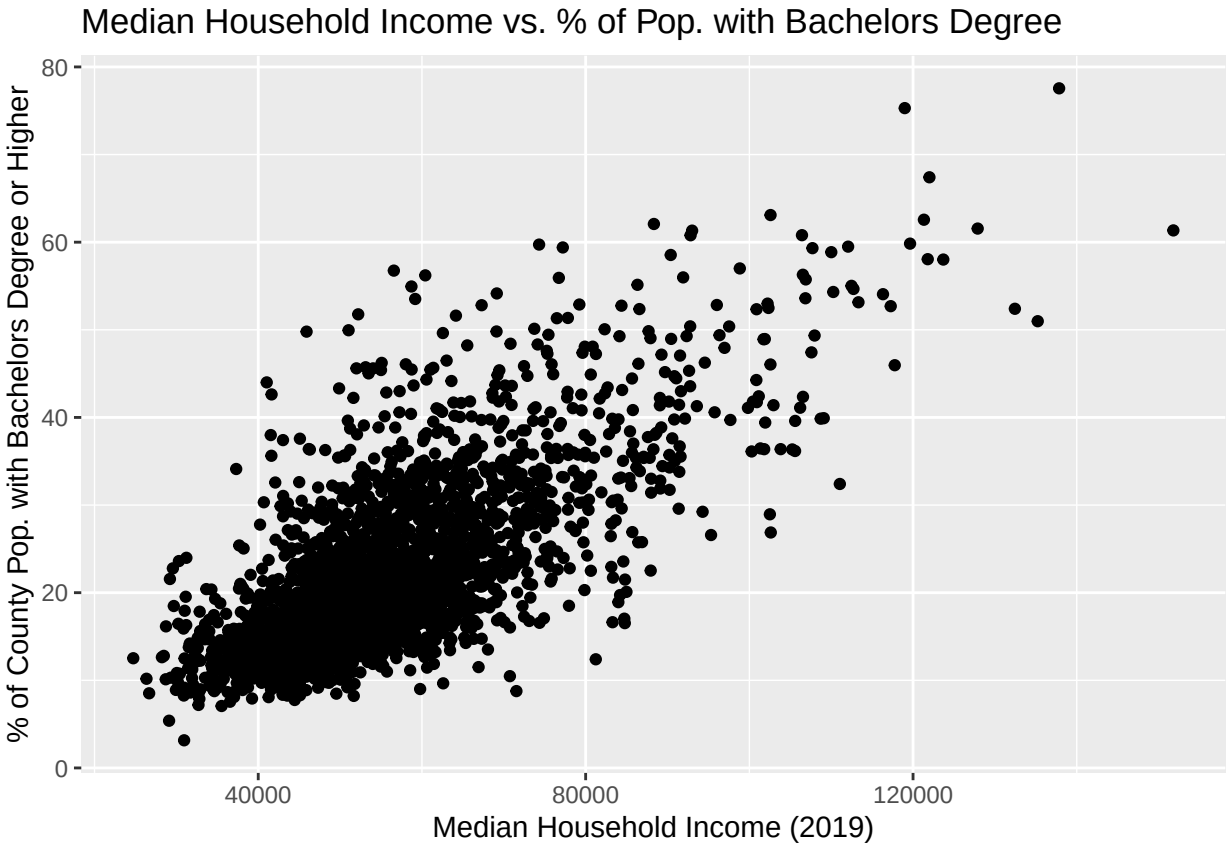


From plotting the histogram above of Median Household Income by county, we can see that the distribution of this variable is positively skewed. Apart from that, it is normally distributed. To account for this positive skew, we may have to take the log transformation of this variable when including it in our model. To confirm this, we will have to plot both against our output variable below.





You can see from the two plots shown here that taking the log transform of the median household income centered the scatter plot distribution more and controlled for the outliers in the underlying distribution. This relationship looks to be linear, so the other only other condition we have to check for is perfect collinearity between this median household income and the education variable explored just previously.



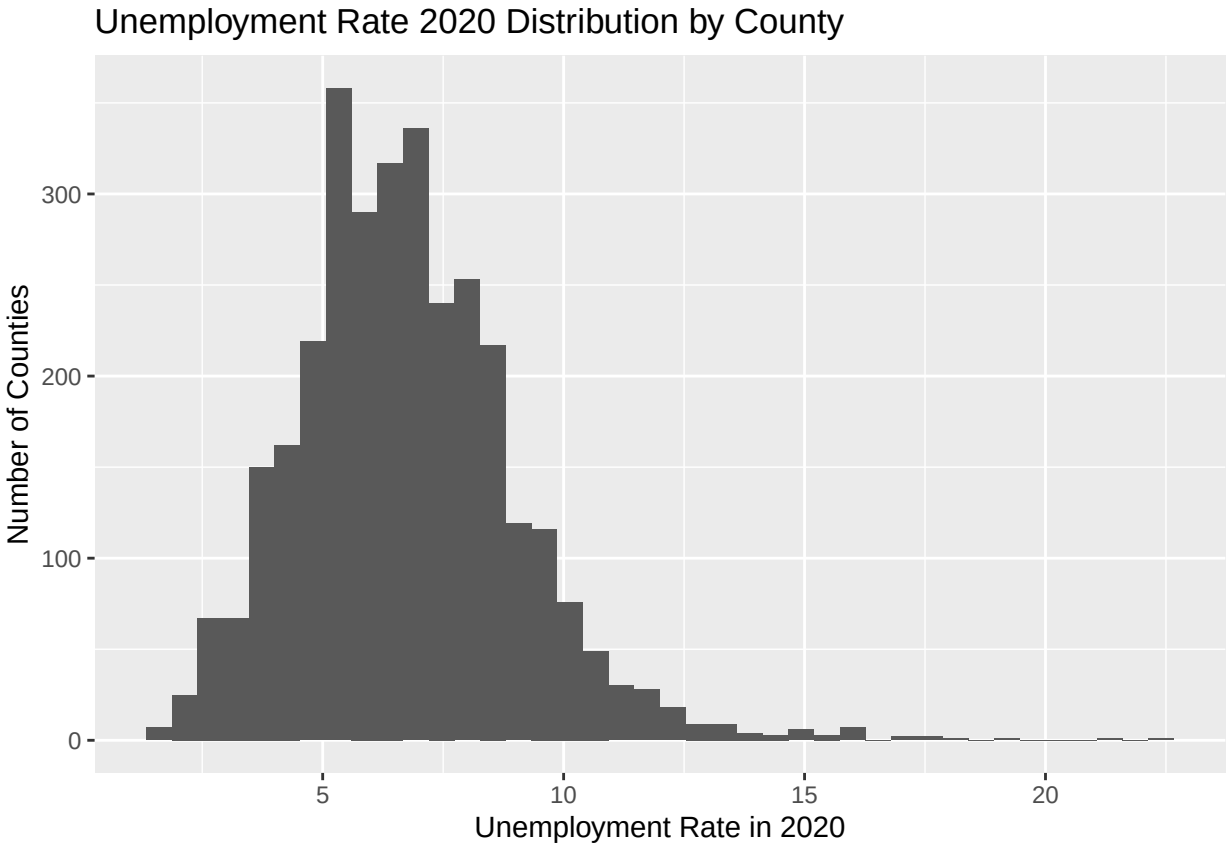
We can see from this scatter plot here that there is possibly high collinearity between the median household income variable and the education variable that we chose. This aligns with our existing assumptions and understanding of that correlation in real life. However, it's not perfect collinearity, so we're still safe to include it in this model.

The last variable we want to include is the unemployment rate in 2020. We're planning to include this to account for the changes in county demographics through the COVID pandemic. This variable in combination with the Median Household Income of 2019 essentially represents a baseline + a delta change from 2019 and 2020. This data was also pulled from the Economic Research Service as part of the U.S. Department of Agriculture, with data from both 2019 and 2020. This explanatory variable is relevant in the sense that it will represent the change county prosperity over the last year, which will definitely impact if people within that county are planning to take trips this summer. We need to do some initial explanatory data analysis to ensure this variable is viable to include and therefore finalize our model 2.

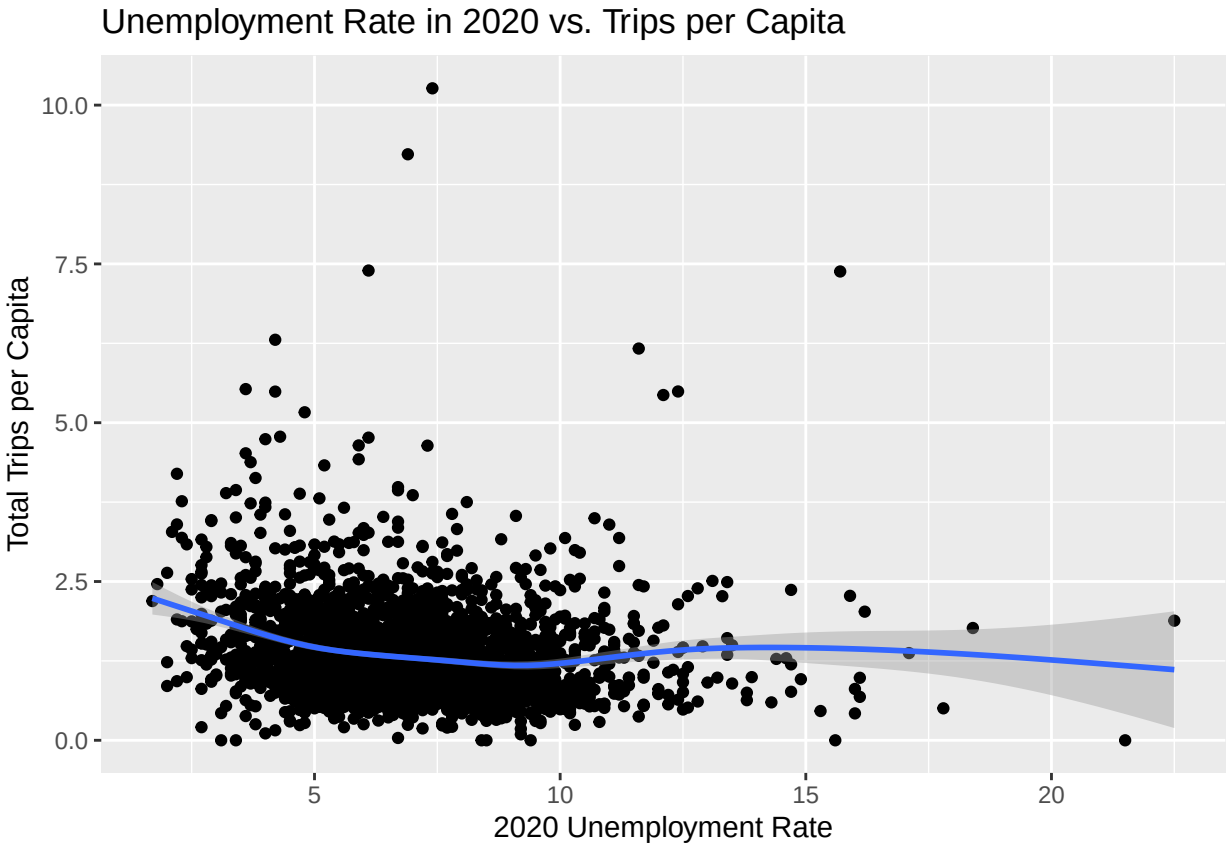
Table 4: Unemployment Rate 2020 EDA

Min	Mean	Median	Max
1.7	6.753728	6.5	22.5

You can see from the summary figures above that the maximum value of around 22% is fairly distanced from the mean value of around 6%, so we will have to produce a histogram to observe any skew.

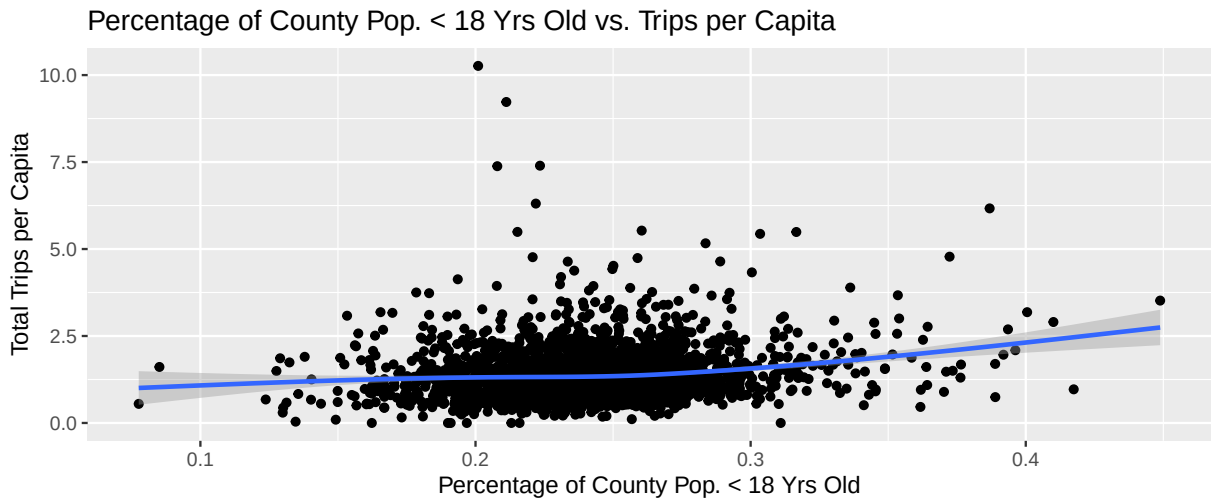


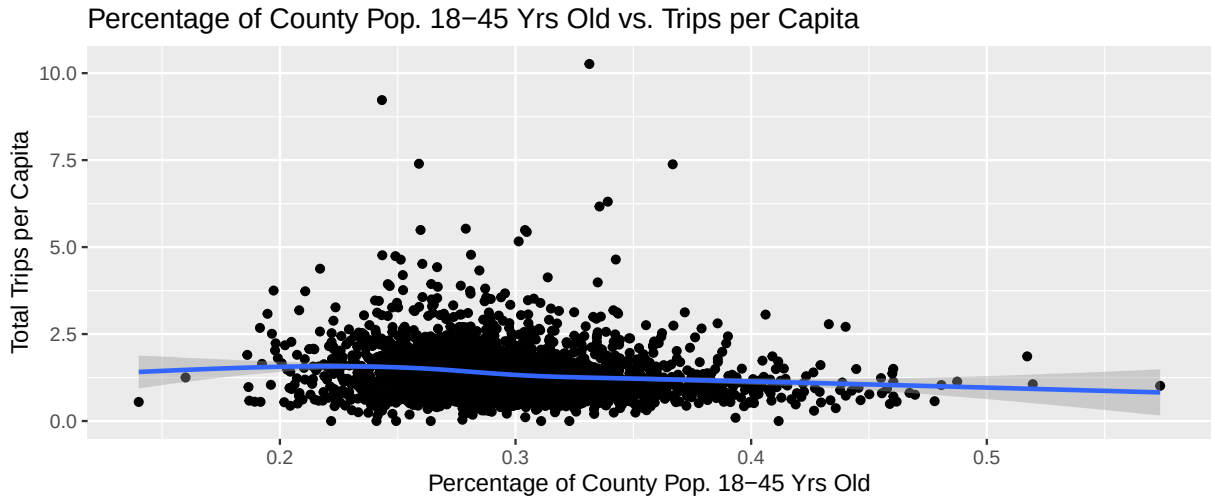
There are some outlying counties present in this distribution, however it's not enough of a skew to require any transformations as of yet. We produced a scatter plot with the trips variable below to continue our exploratory data analysis of this variable.

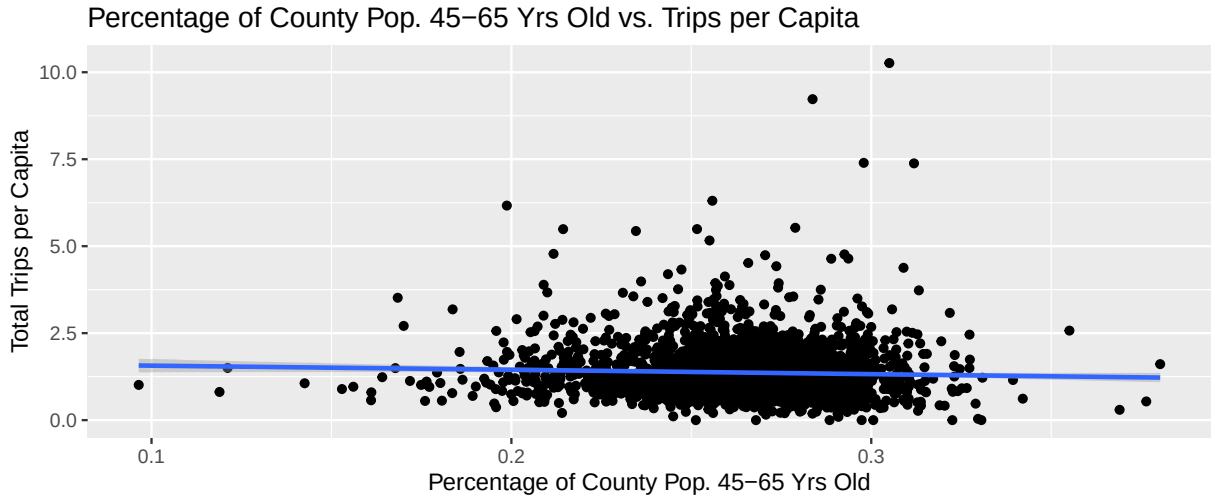


There is a relatively linear relationship between this new input variable and our output variable between the x values of 0 to 10% unemployment. We can feel comfortable moving ahead with including this variable into our model, and will do additional statistical tests to determine its significance. This concludes the exploratory data analysis for the additional covariates we plan to add into our second iteration of our model.

3. **Model Three**: For our third and final model, we will include all other previously defined covariates plus some additional ones. Our purpose for this model is to control for age of county population. This control variable is relevant because different age demographics within a population will certainly affect how many trips families or individuals will take this summer. This age data was pulled in from the US census, and we transformed it to be described within 3 variables, percentage of population between less than 18, between 18-45, and 45-65. We will include each of these three into our model.







These scatter plots produced linear fit lines, particularly the graph representing percentage of the population between 45 and 65 years old. However, we will have to run additional statistical tests to explain the relationship between these variables more clearly and see if there is a significant effect.

To conclude this section, our final model specifications are as follows:

Limited Model

$$Total\ Trips\ Per\ Capita\ in\ June\ 2021 = \beta_0 + \beta_1\ Percent\ of\ Pop.\ 1st\ Vac\ Dose\ Admin\ by\ June\ 1st\ 2021 \quad (1)$$

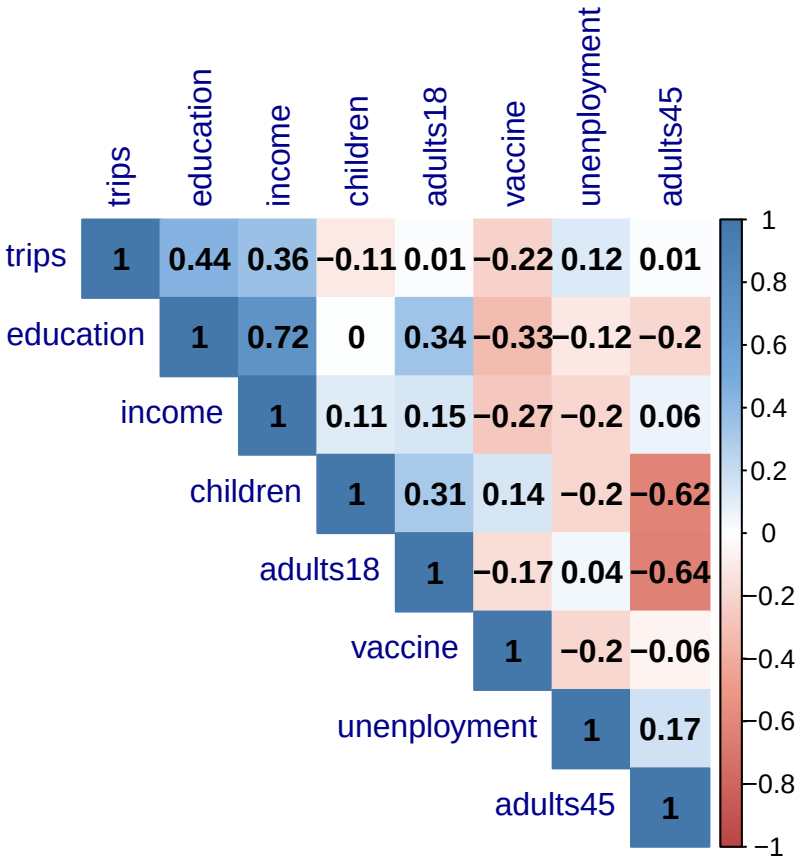
Model 2

$$Total\ Trips\ Per\ Capita\ in\ June\ 2021 = \beta_0 + \beta_1\ Percent\ of\ Pop.\ 1st\ Vac\ Dose\ Admin\ by\ June\ 1st\ 2021 + \beta_2\ Percent\ college\ or\ Higher + \beta_3\ Median\ Household\ Income + \beta_4\ Change\ in\ Unemployment\ Rate \quad (2)$$

Model 3

$$Total\ Trips\ Per\ Capita\ in\ June\ 2021 = \beta_0 + \beta_1\ Percent\ of\ Pop.\ 1st\ Vac\ Dose\ Admin\ by\ June\ 1st\ 2021 + \beta_2\ Percent\ college\ or\ Higher + \beta_3\ Median\ Household\ Income + \beta_4\ Change\ in\ Unemployment\ Rate + \beta_5\ pct_children + \beta_6\ pct_adult18 + \beta_7\ pct_adult45 \quad (3)$$

The plot below shows the crosswise correlations values between each variables. We can see that all the covariates, except for education and income have a negative linear relationship with the number of trips per capita that citizens are taking. This negative relationship will pass on to the regression coefficients.



3. A Regression Table

Model 1 is a limited model with only our primary variable included, the percentage of administered dose one vaccination by county, Model 2 is the model that includes sociodemographics features such as education, income, and unemployment, and in Model 3, we included the age groups covariates.

From our regression table, we can infer that, at a county level, when we increase the administered dose 1 of covid vaccine percentages, the number of trips per capita decreases 0.002 points. This is a statistically significant coefficient, at 5% level of significance. This coefficient is meaningful only when we include the age groups (model 3), and its the only covariate of the model. As expected, in counties with a higher percentage of people with a bachelor's degree or higher, the number of trips decreases, statistically significant at a 1% level of significance. The county's median income has a negative effect on the number of trips that the population is taking, meaning that a % increase in income results in a 0.007 increase in the number of trips per capita. As expected, in counties with high changes in unemployment, the amount of trips per capita decreases. The only covariate that has a positive effect is the percentage of children that a county has.

The adjusted R squared in our third model increased compared to the R squared of model one. Suggesting that this model is explaining more variability in the number of trips per capita. Also, all the coefficients are statistically significant.

```
stargazer(model_1, model_2, model_3, title="Regression Results", type= "text",scalebox='0.7')
```

```
##
## Regression Results
## ================================================================================================
##                                                Dependent variable:
##                              ---------------------------------------------------------------------
##                                                    sum_of_trips
##                                     (1)                  (2)                    (3)
## ------------------------------------------------------------------------------------------------
## Administered_Dose1_Pop_Pct       -0.011***             -0.001                 -0.002*
##                                   (0.001)              (0.001)                (0.001)
##
## pct_college_higher                                     -0.017***              -0.013***
##                                                        (0.002)                (0.002)
##
## log(MEDHHINC_2019)                                     -0.471***              -0.450***
##                                                        (0.076)                (0.086)
##
## Unemployment_rate_2020                                 -0.087***              -0.069***
##                                                        (0.006)                (0.007)
##
## pct_children                                                                  1.729***
##                                                                               (0.519)
##
## pct_adults18                                                                  -2.937***
##                                                                               (0.418)
##
## pct_adults45                                                                  -3.239***
##                                                                               (0.912)
##
## Constant                          1.766***             7.495***               8.404***
##                                   (0.039)              (0.805)                (0.815)
##
## ------------------------------------------------------------------------------------------------
## Observations                       2,421                2,421                  2,421
## R2                                 0.047                0.182                  0.208
## Adjusted R2                        0.047                0.180                  0.205
## Residual Std. Error          0.667 (df = 2419)    0.619 (df = 2416)      0.609 (df = 2413)
## F Statistic                120.052*** (df = 1; 2419) 134.025*** (df = 4; 2416) 90.286*** (df = 7; 241
## ================================================================================================
```

```
## Note:                                                                  *p<0.1; **p<0.05; ***p<0
##
## Regression Results
## ===
## 0.7
## ---
```

4. Model Limitations

We have evaluated the assumptions of the classical linear model and identified a few points of concern where CLM model assumptions may have been violated. It is worth noting that since we are conducting an analysis with US county-level granularity, we have over 2,000 data points at our disposal and can consequently invoke the assumptions of the large sample model (IID data and a unique best linear predictor exists), rather than those of the classical linear model. However, examining our model's performance against the classical linear model limitations can help us better understand our model limitations, especially with regards to potential biases in our estimates and standard errors.

Independent and Identically Distributed Data (IID): Our work studies the effects of vaccination rates on travel in the US. Although our analysis focuses on examining counties, we ultimately wish to understand the factors that influence why people, in general, go on trips. Since we want to glean insights on the tendencies of the American population, the population of interest for this analysis is all the residents of the United States. Our model, however, is based on data on vaccination rates and trips taken per county, rather than a random selection of people in the USA. From this observation alone, we recognize that there is a disconnect between our sampling methods and the population we wish to generalize our findings to. Ideally, an experimental design in which people were randomly sampled from the entire American population would have yielded a truly independent and identically distributed sample that is more reflective of our analysis goals. However, we were limited by the availability of good datasets with more granularity than the county level.

Some additional aspects of our sample pose problems for the assumption of identically distributed data. Our sample is composed of all the counties in the United States, except for the ones that did not report COVID vaccination rate data. (We assume that this lack of reporting occurs randomly, but in reality could very much be dependent on systematic factors like county/state resources.) Counties are defined in a non-random manner, with a myriad of dependencies on various geographical, political, social, and economic factors that could also affect resident travel and COVID-19 vaccination rates. For example, geographical clustering could cause neighboring counties to have similarities in culture and politics that influence our predictor and outcome variables. Furthermore, the COVID-19 virus as well as COVID-19 vaccination availability reached different counties at different points in time that are dependent on political and economic factors (that could also have an effect on travel).

Counties are also not identically distributed in population. Some counties have much higher populations or higher population densities than others, which leads to underrepresentation of their effects in our analysis if the population we are interested in is indeed the entire US population. Though our variables are normalized by population, due to the non-random geopolitical divisions of counties, the effects of factors such as population density on vaccination rate and travel are not captured.

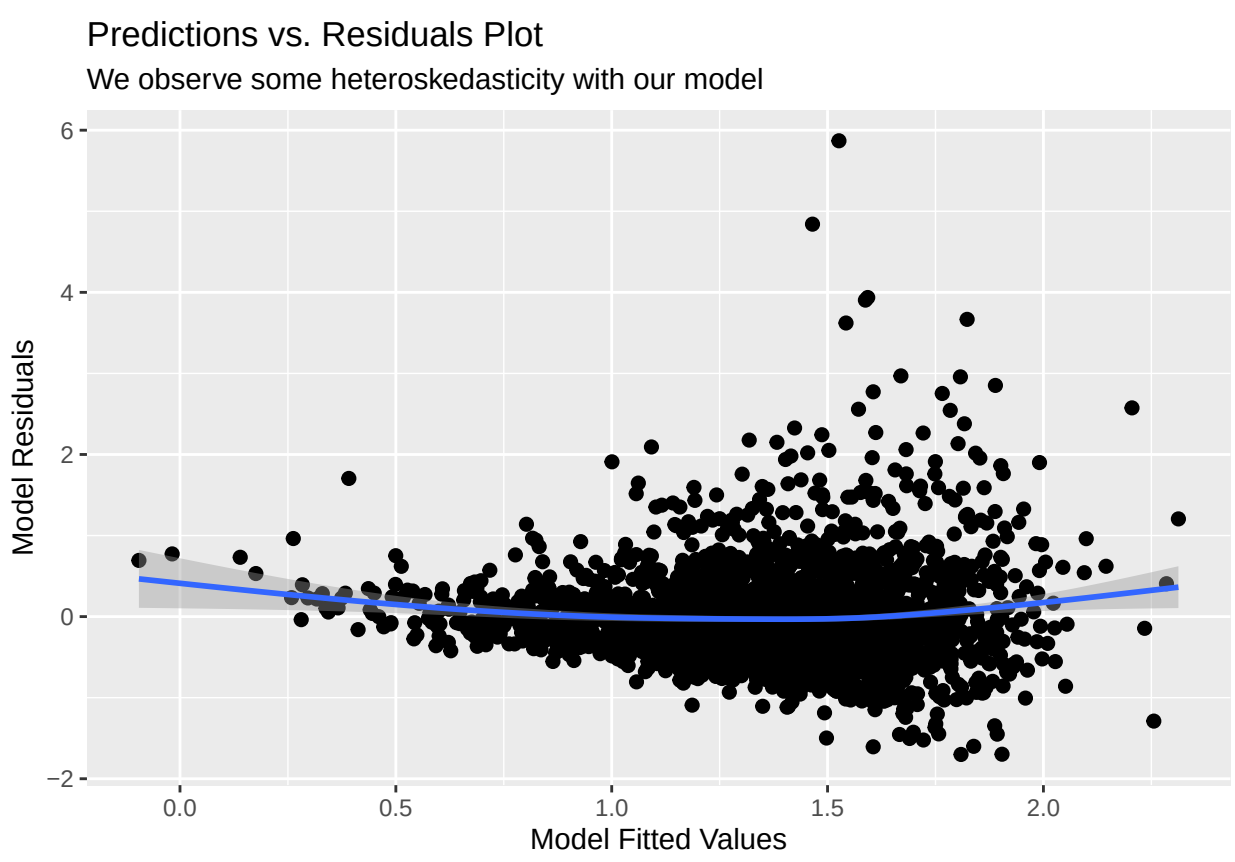
These concerns about violations to the IID assumption suggest that we must be very careful with the interpretation of our results. Since our model was constructed under the large-sample model, which requires just 2 assumptions - IID data and the existence of a unique best linear predictor, having non-IID data threatens the validity of our model and our results. That said, we still did our best to design our study to mitigate the effects of clustering brought about by the geopolitical nature of political and administrative divisions, given the limitations of the available datasets. By choosing to leverage county data rather than state data, we added more granularity to the data that is expected to reduce some unwanted effects of clustering. For example, the tendencies of a very urban county within a mostly rural state would have barely been captured if using state-level data.

Homoscedasticity (homogeneity of variance): To assess the variation of the residuals of our model, we can examine a plot of the model predicted values vs. the residuals. From this plot, we we observe no major

issues with the linear conditional expectation assumption. (The line fitting the residuals slightly deviates from the horizontal residuals = 0 line, but is not too concerning.) However, we do observe heteroskedasticity, meaning that the variance of the residuals is not constant. We see that as the fitted values get higher, the width of the deviation in residuals increases. This suggests some systematic variation of residuals as the number of trips variable increases that could be explained by additional features in our model. It also suggests that, although our estimator is still unbiased, the ordinary least squares estimator may be inefficient.

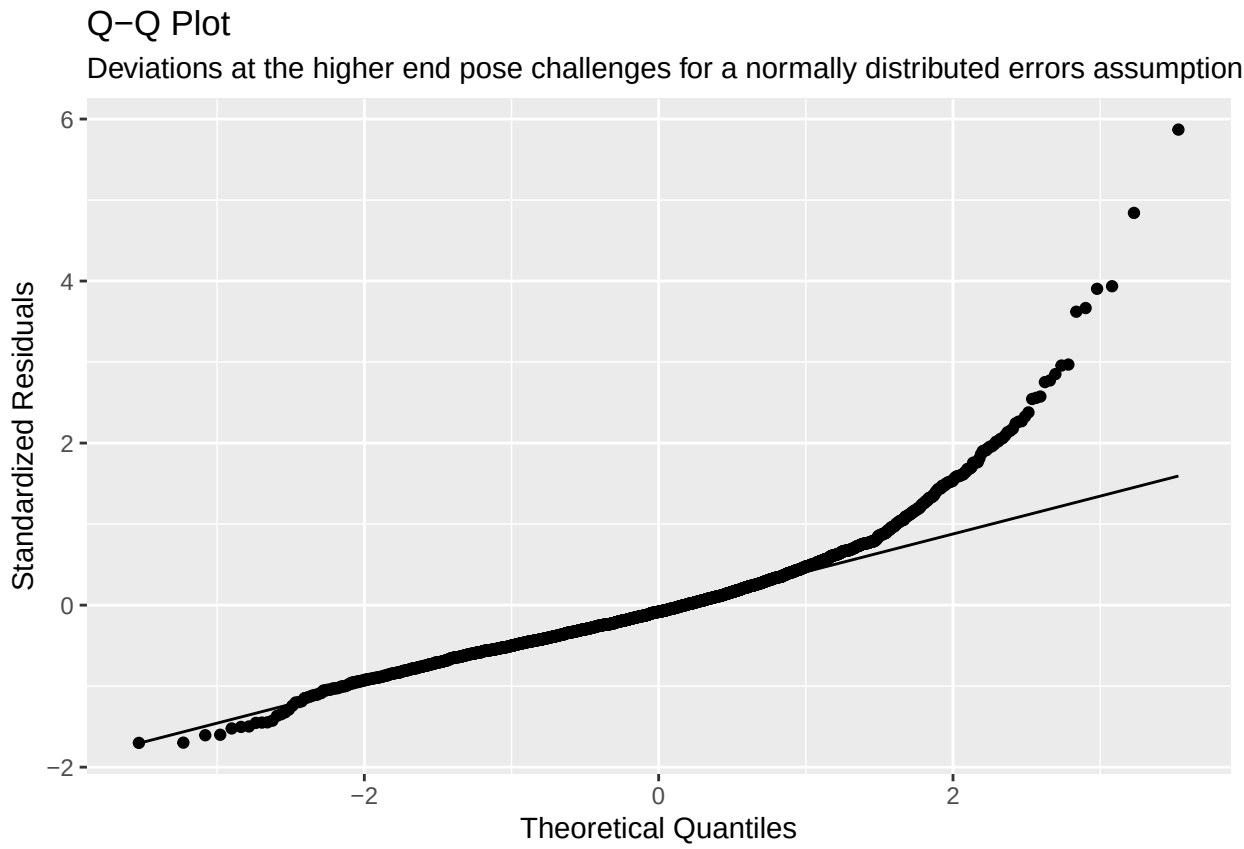
```
augmented_data <- broom::augment(model_3)

predictions_residuals <- augmented_data %>%
  ggplot() +
  aes(x = fitted(model_3), y = residuals(model_3)) +
  geom_point(size=2, shape=19) +
  labs(
    title = "Predictions vs. Residuals Plot",
    subtitle = "We observe some heteroskedasticity with our model",
    x = "Model Fitted Values",
    y = "Model Residuals"
  ) +
  geom_smooth()
predictions_residuals
```



Normally distributed errors: We can assess the distribution of the model errors using a Q-Q plot to compare against a theoretical normal distribution of residuals. Although the line we get is mostly straight, we do see deviation towards the higher end of the plot that poses some concern for the validity of a normally distributed errors assumption. This might be due to the presence of outliers and relative sparseness of data in the higher range of values for number of trips and vaccination rate.

```r
qq <- augmented_data %>%
  ggplot(aes(sample = residuals(model_3))) +
  stat_qq() +
  stat_qq_line() +
  labs(
    title = "Q-Q Plot",
    subtitle = "Deviations at the higher end pose challenges for a normally distributed errors assumptio
    x = "Theoretical Quantiles",
    y = "Standardized Residuals"
  )
qq
```



High Collinearity: We do not observe any perfect collinearity in across our model factors. However, there are some highly correlated variables that might split the effect of interest across variables. In particular, from the correlation plot provided in the "Exploratory Data Analysis" section, we see that the "adults18" and "adults65over" variables have a correlation of -0.82. Still, we are not too concerned about the multicollinearity here because these variables are control variables, rather than the primary predictor and outcome variables of interest.

5. Discussion of Omitted Variables

Our analysis primarily investigates the relationship between COVID-19 vaccination rates and the number of trips that Americans take this summer. There are many social and economic factors that influence whether someone decides to go on a somewhat long trip (100+ miles), implying a fairly complex underlying causal model. Although we include several demographic variables to capture these factors (e.g. income, age, employment), our linear model simply cannot capture all the features that may influence our response.

Rather, we examine a few factors that potentially help explain travel tendencies and the direction of their corresponding omitted variable biases.

"Travel policy"-related indicator variables could be used to reflect the degree to which stringent local- and state-level travel policies were enacted in a region and its surrounding areas. Whether there were stay-at-home orders, quarantine requirements, or COVID testing requirements may generate a barrier to travel that would reduce the number of trips people take. Our current model does not have any variables that can proxy this omitted variable, but from this cursory understanding of how these travel policies restrict travel, we would expect the omitted variable bias to be negative (the estimate for the effect of vaccination is lower than the true relationship). Given that the parameter estimate for the effect of vaccination rate on the number of trips is negative, we would expect the direction of this omitted variable bias to be towards zero.

Additionally, an urban-rural classification variable (as defined by the US Census Bureau) could be used to capture the effect that this classification of county type has on travel. Since we expect residents of urban counties to travel more (Czepkiewicz 2020) and to be more likely to get vaccinated against COVID, we expect a positive bias for an omitted urban county indicator variable. We would therefore expect the direction of this omitted variable bias to be away from zero.

We do not consider the reverse causality effect to be a particularly large concern (people deciding to get vaccinated because they want to take trips), though this effect could still exist to a minor extent and generate a negative feedback loop for the size of our parameter estimate.

6. Conclusion

This study has yielded interesting insights for the tourism agency of Portlandia. We saw from our generation of three models that our key input variable, the percentage of a county's population that has received their 1st vaccine dose, has a small effect on the total number of trips taken in that county per capita. While it is small, the effect is still there. However, the other demographic variables that we brought into the model, such as education, median household income, and unemployment, have more of an effect. We recommend focusing on those characteristics of a county when considering where to focus our marketing efforts towards. Lastly, we also advise cautious with the interpretation of these results. Had we had more time and funding to complete this study, we would have formulated a sampling strategy rather than including data points from all counties across the country. Generating a random sample to model the population would have allowed the IID assumption to have been met.

References

Centers for Disease Control and Prevention. "COVID-19 Vaccinations in the United States, Jurisdiction." *CDC Data*, accessed July 18, 2021. https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-Jurisdi/unsk-b7fc

Czepkiewicz, M., et al. "Who travels more, and why? A mixed-method study of urban dwellers' leisure travel." *Travel Behavior and Society*, Volume 19: Apr 2020. p. 67-81. https://doi.org/10.1016/j.tbs.2019.12.001

U.S. Department of Agriculture. "County-Level Data Sets." *Economic Research Service*, accessed July 21, 2021. https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/

U.S. Census Bureau. "County Population by Characteristics: 2010-2019." *Annual County and Resident Population Estimates by Selected Age Groups and Sex*, accessed July 21, 2021. https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-detail.html

U.S. Department of Transportation. "Trips by Distance." *Bureau of Transportation Statistics*, accessed July 18, 2021. https://data.bts.gov/Research-and-Statistics/Trips-by-Distance/w96p-f2qv/data