The affect of Vaccinations on Long Distance Trip

Anand Patel, Viswanathan Thiagarajan, Vijay Ranganatha

7/26/2021

1. An Introduction

COVID-19 has had a devastating impact on the lives of many people. There were various policies which the government implemented to control the spread of the virus. Such policies as quarantine measures, lockdowns, and travel bans halted tourism travel. After a year-long struggle with the new norms and lockdowns imposed, there appears to be a change in the perception of Covid-19 as vaccines become available to people living in the United States. With vaccine rates increasing in the United States, mask mandates are lifting, people are venturing out to see friends and family, and advertisements are marketing products and activities as “returning” to the the pre-pandemic way of life. We must examine if the data supports this return in the context of tourism travel in the continental west coast states.

With summer of 2021 approaching, we would like to examine if increasing vaccination rates is leading people to taking tourism trips again. This research would benefit the **Tourism Board of Continental West Coast States** by informing this organization if places with high vaccination rate show an increased interest in tourism travel and subsequent demand for travel destinations. If the analysis supports that increasing vaccination rates among West Coast States leads to more tourism travel, then the **Tourism Board of Continental West Coast States** would know to increase staffing, reopening, lodging, and advertising efforts across Continental West Coast for travel destinations since the demand is increasing. If no such increase is shown, then the **Tourism Board of Continental West Coast States** would be informed to decrease or maintain the current operational procedures. Knowing this information, the **Tourism Board of Continental West Coast States** can take cognitive action of either increasing, decreasing or maintaining the current operational state whereby they will able to manage their resources and also would help them in planning their budget.

Research Question

We are presenting to the tourism board of Continental West Coast States (California, Oregon, Washington) and showing **if having a greater % of vaccinated people is leading to an increase in tourism as represented by more long distance trips, 50+ miles from home, taken.**

Operationalization

To perform the analysis, we are examining counties from the Continental West Coast States because residents of these counties are likely to do tourism travel to locations within these states and the vaccine drives here have been substantial. The number of counties in California, Oregon, and Washington total to 135.

Cross-section in Time

For each county, we observe the percentage of people vaccinated by 1-May-2021 since this is close to the summer. To account for possible 2-week vaccination incubation for individuals who were vaccinated exactly on May 1st, we consider the daily tourism trips averaged across 05/14/2021 to 05/21/2021. For covariates, we will also evaluate if the county population, county resident age, if the county leans towards Republican, and wealth of residents play an important role in determining the number of trips taken as well.

Variables

Our outcome variable should measure the number of tourism trips taken. We operationalize a tourism trip as any trip taken that is far enough from home to signify an occasional trip. We consider tourism to be long distance travel away from home, since a long distance trip is likely to be a special or unique occasion. To analyze the impact of our treatment and controls on tourism, we will consider any trips greater than 50+ miles from home as tourism trips. The US Department of Transportation provides a county level dataset on the number of trips taken by distance from home (https://data.bts.gov/Research-and-Statistics/Trips-by-Distance/w96p-f2qv). We sum up the columns for trips taken over 50 miles to get the number of long distance trips taken by county.

Our treatment variable is county vaccination rate, measured as the percentage of county population that is fully vaccinated with 2 doses or one dose from a single-dose vaccine. This data for county level vaccination rate is readily available by the Center for Disease Control as a dataset (https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh). This dataset, for California, does not report vaccination rate for counties with population below 20,000. We lose 10 California counties as a result.

County population would directly impact the number of trips taken since more individuals means more people who can potentially take tourism trips. County population also increases for urban counties versus rural counties, so this variable could also represent urban vs rural in our study which can capture the difference in Covid impact on these environments. Our control for county population can be operationalized by calculating the 2021 county population estimate used in the CDC Vaccine Rate dataset using the columns available.

$$\text{pop\_est} = 100 * \frac{\text{tot\_num\_vaccinated}}{\text{vaccination\_rate}}$$

County resident age can influence how many trips are taken since 65+ individuals might not have a long distance travel oriented lifestyle and tourism trips would likely be low for counties where older residents live. We operationalize county resident age using the county level median age dataset available from 2020 U.S. Census population estimates (https://www.census.gov/programs-surveys/popest/technical-documentation/research/evaluation-estimates/2020-evaluation-estimates/2010s-county-detail.html).

Republican leaning states, like Florida, have shown a resistance to Covid prevention policies and a belief in science for Covid decision making. Republican leaning areas have been associated with neglecting Covid safety protocols due to residents being less concerned with precautionary measures. Adding a control for indicating if a West Coast county leans Republican would capture if people are taking tourism trips mainly based on their political beliefs and their associate attitude towards Covid-19. We might expect Republican leaning counties to have more tourism trips taken regardless of vaccination rate. We operationalize if a county is Republican leaning by determining if the county voted more for a Republican candidate in the 2020 presidential race over the Democratic candidate, computed from the United States General Election Presidential Results (https://github.com/tonmcg/US_County_Level_Election_Results_08-20/blob/master/2020_US_County_Level_Presidential_Results.csv).

We would expect that counties with more wealthy residents have more access to the money and means to travel. We operationalize the wealth of residents using the county level median household income made available in the USDA's Economic Research Service dataset (https://data.ers.usda.gov/reports.aspx?ID=17828). This dataset also provides the county level unemployment percentage and percentage of State Median Household Income, which we evaluate as other potential controls but rule out in our analysis below.

Our final dataset

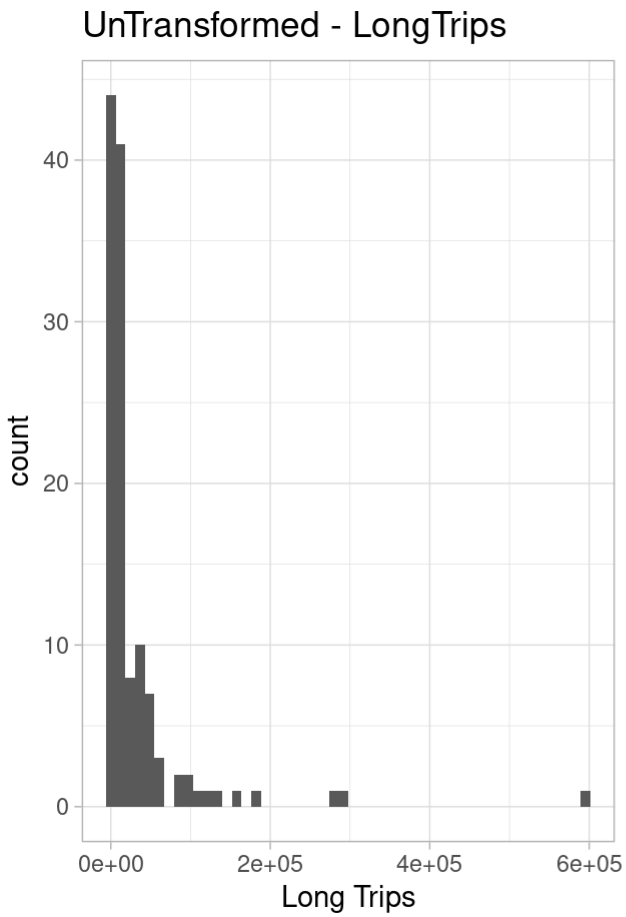
We download, clean, and combine the datasets using the following R Markdown file: `src/Lab2_Data_Wrangling.Rmd` . The final dataset is saved out at the completion of this `Lab2_Data_Wrangling.Rmd` and used for our analysis here.

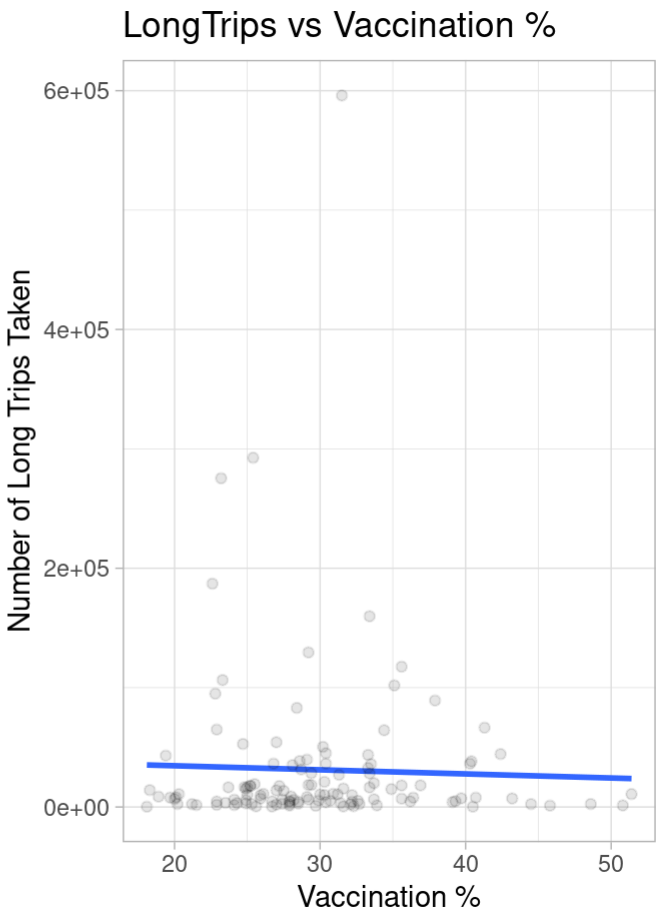
Our final dataset includes 125 counties, after cleaning out counties that are missing trips taken (Alpine County in California) and California counties whose vaccination rate was unreported due to the population being below 20,000 (as outlined in the data collection methodology of the CDC vaccination rate dataset). This may present a small issue since we are under representing small population West Coast counties in our analysis, but dropping only 10 counties from 135 maximum counties would still leave us with a sizeable dataset of 125 which does include a variety of county populations.

2. A Model Building Process

As part of the Research, we are interested in understanding if increasing vaccination rates is leading people to taking tourism trips again. To analyze this we will be using the trips of 50+ miles from home as long distance trips to capture outcome variable. The treatment variable would be the vaccination rate. The other co-variate variables which we believe might have an meaningful impact are also anayzed: County Population, County Resident Age, Does County affiliates more towards Republican Party and County Average Income.

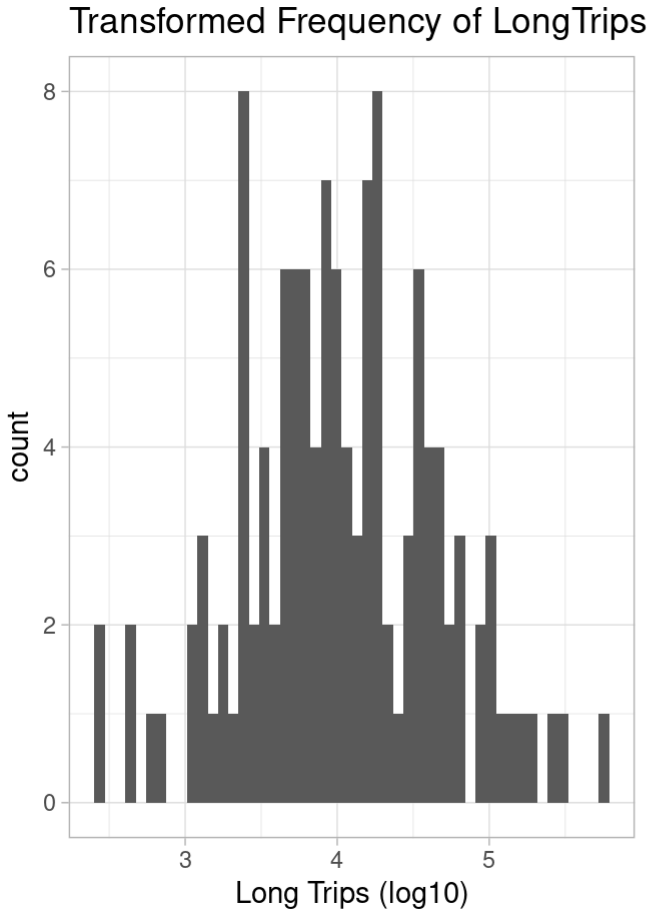
Variables of Interest

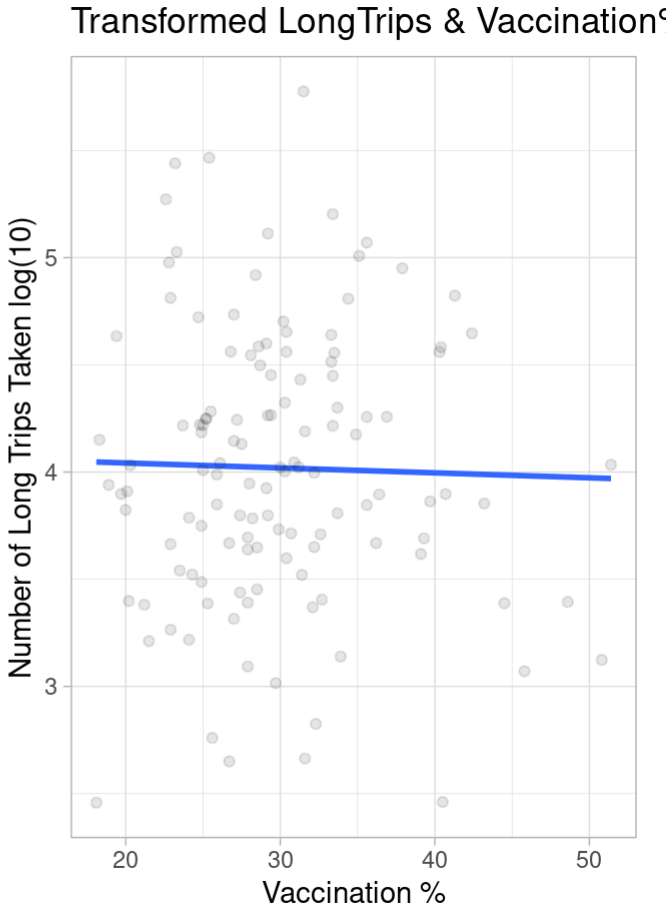




Untransformed Variables

Checking the Long Trips in the histogram shows us that the plot is not normal and the data is very much isolated towards the left. Also, the scatter plot of Vaccination percentage and Number of Long Trips, shows us that the data is very much dispersed for Long Trips. Hence, there appears to be a need to apply transformation to make them normal.

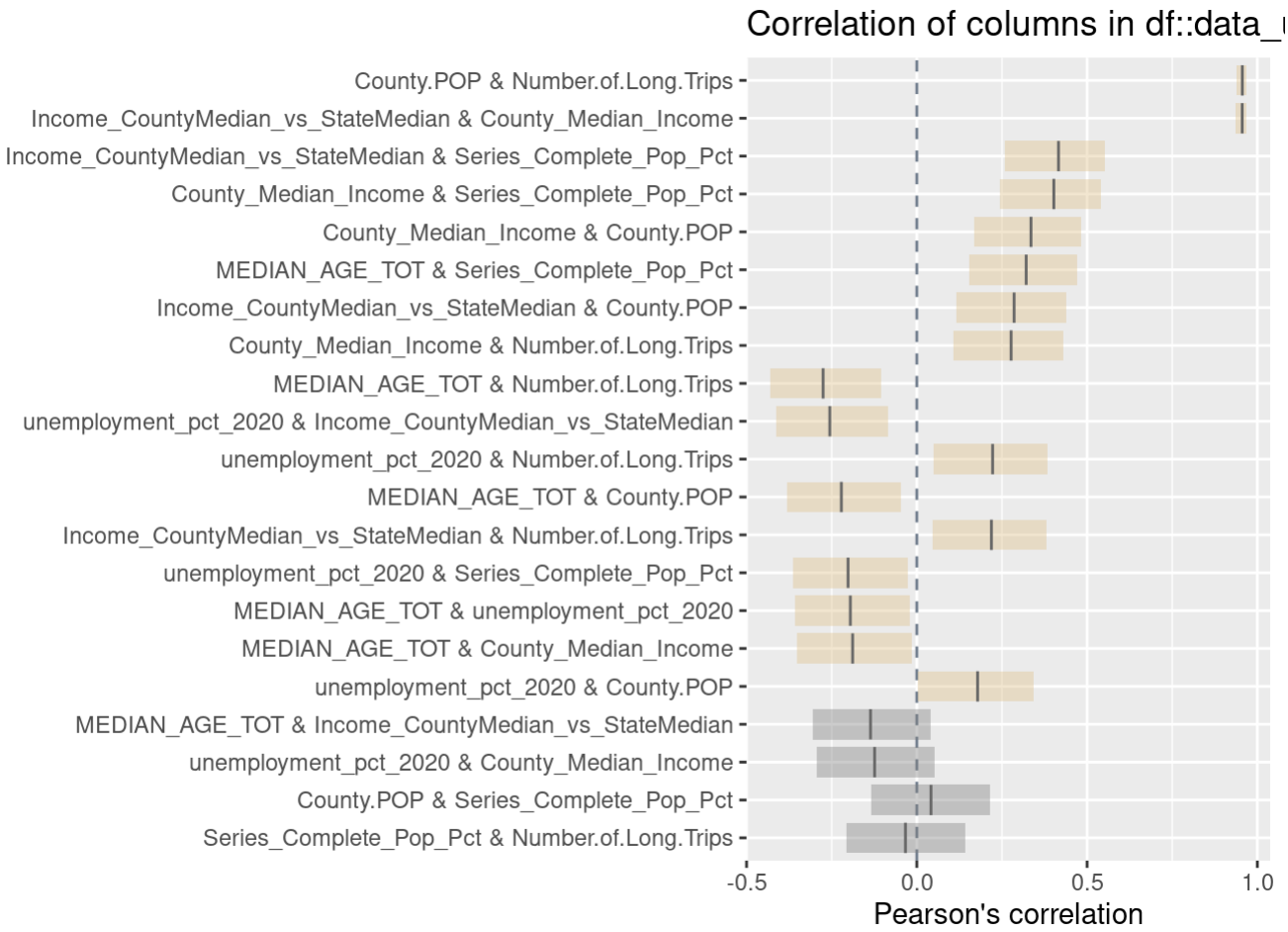




Transformed Variables

As can be noticed after the log10 transformation, the data of Dependent Variable (Long Trips) appears to be a lot normal when compared with the previous plot. Also, the scatter plot of Vaccination percentage and Number of Long Trips, shows us that the data can be analyzed better with log10 transformation.

To understand the correlation between the variables that we are interested to analyze, lets take a look into their relationship. The following plot provides the correlation of the different variables:

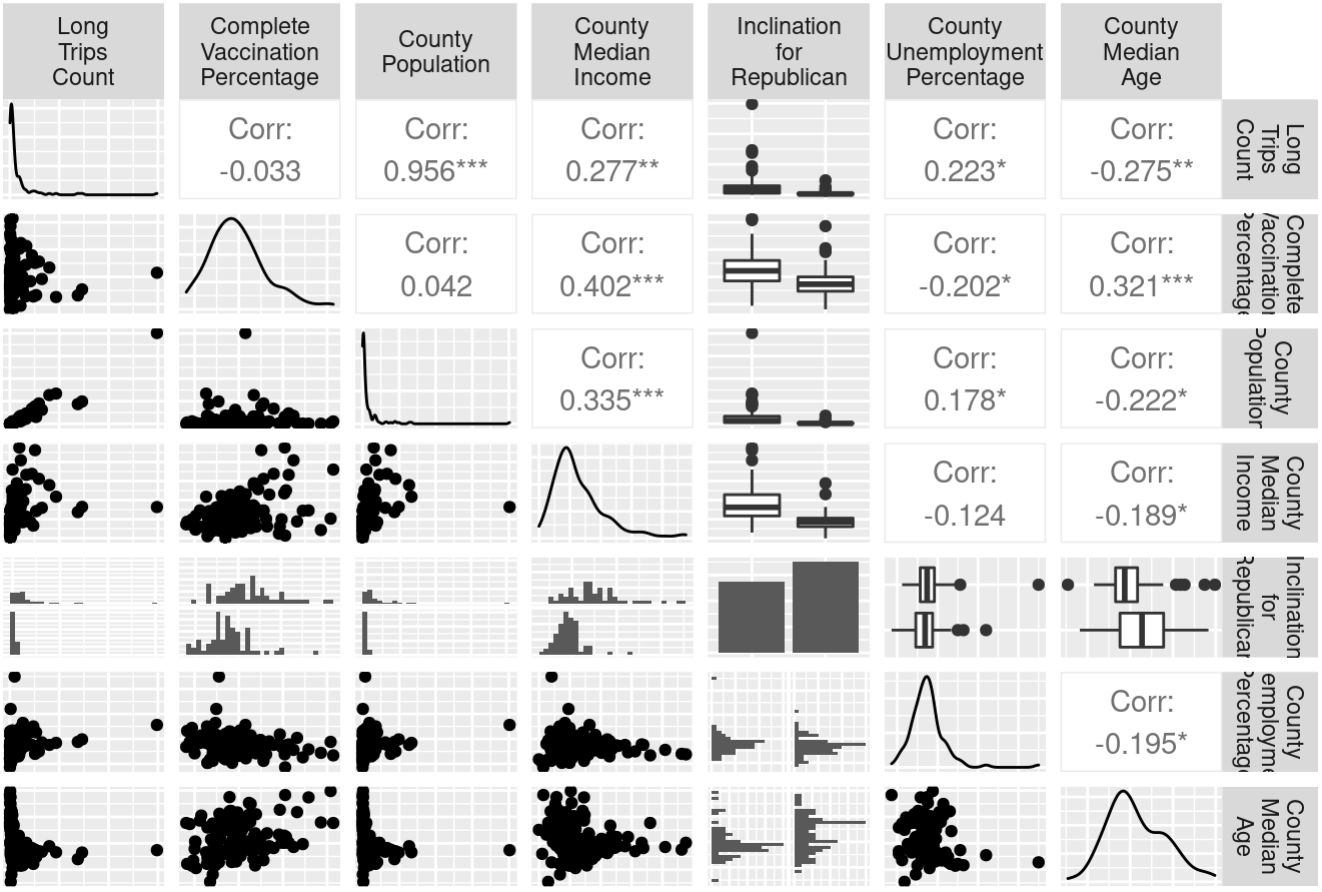


Columns Correlation

The correlation matrix shows us that there is very high positive correlation between County Median Income Compared with State Median Income. Knowing this, we will only use county median income to represent wealth as a control for our analysis. The other variables are not highly correlated with each other and hence we should be good with using them as part of the model building process.

Potential Covariates

Untransformed: Plots for Outcome and Features



Untransformed Pairs Plot

The above plot is for variables where there is no transformation. Following observations can be made about the variables density plot (along the diagonal):

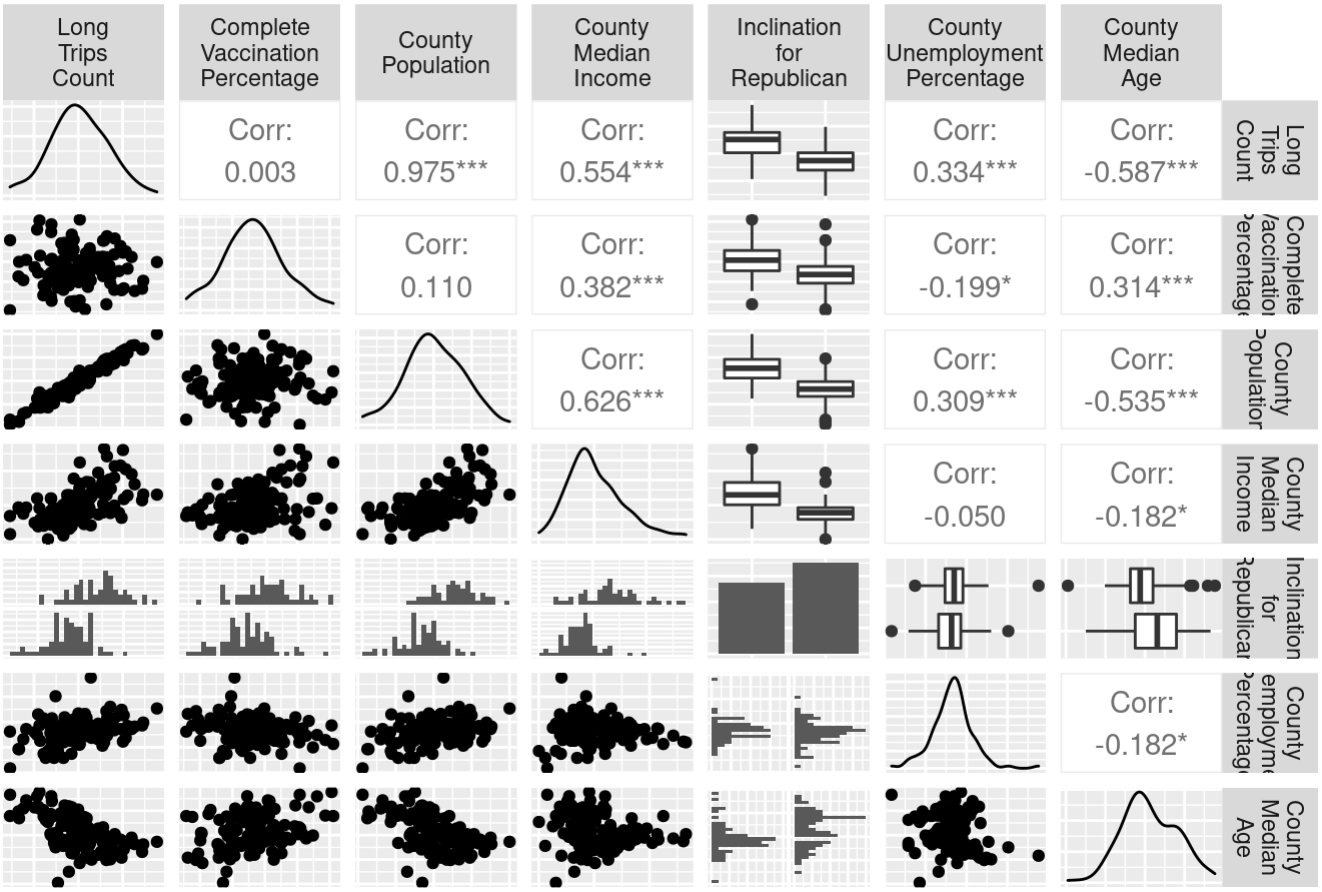
- Long Trips Count (Outcome Variable) - The distribution is not normal and hence transformation needs to be applied
- Complete Vaccination Percentage (Treatment Variable) - The distribution is not normal and is skewed a lot towards the right
- County Population (CoVariant) - The distribution is not normal and hence transformation needs to be applied
- County Median Income - The distribution is not normal and is skewed a lot towards the right
- Inclination for Republican (CoVariant) - This is a binary variable and hence bar plot is provided. We can notice that there is more counties inclining more towards the Republican party.
- County Unemployment Percentage (CoVariant) - The distribution is not normal and is skewed a lot towards the right
- County Median Age (CoVariant) - The distribution is not normal and transformation is required.

For right-skewed distributions, we can transform our variables down the Box-Cox Ladder of Powers. We try transforming all our right-skewed variables using a log base 10 transform, and plotting the GGPairs again to see if we have more normally distributed variables now.

Transformations

We will now transform the variables which did not have a normal distribution from the previous observation. We use a GGPairs to look at the correlation between variables, the distributions of the variables along the plot diagonal, and scatterplots of each variable pair.

Log10 Transformed: Plots for Outcome and Features



Transformed Pairs Plot

After the transformation, we notice that all of the variables distribution look much more normal now. The `County Median Income` still looks a bit right-skewed. We could go one step further down on the box-cox ladder of powers to transform this feature using a reciprocal root, but that comes at the cost of difficulty in interpreting the coefficient for `County Median Income`. We will choose to log base 10 transform all the variables.

From the bar chart for `isRepublican` we can see we have slightly more Republican counties in the dataset than Democratic counties, but the number of counties are almost equal. The Republican counties do have a lower number of long trips taken, lower vaccination rate, lower county population, lower median income, lower unemployment percentage, but higher median age.

Variable Table

Variable.Name	Variable.Short.Form	Variable.Description
Number.of.Long.Trips_log10	County Level Long Trips Taken (Log10)	This is the 50+ mile trips taken by ppl from their home.
Series.Complete.Pop.Pct_log10	County Level Vaccination complete percentage (Log10)	Percentage of people who have taken both doses of vaccine
County.POP_log10	County Population (Log10)	County Population Number

Variable.Name	Variable.Short.Form	Variable.Description
County.Median.Income_log10	County Median Income (Log10)	Median County Population
isRepublican	Binary Value 1 for Republican	Indicates Party affiliation of County
Unemployment.Pct_log10	County Unemployment Percentage (Log10)	Counties Unemployment Percentage
Median.Age_log10	County Median Age (Log10)	Counties Median Age

Akaike Information Criterion

With our transformed variables, we run the Akaike Information Criterion (AIC) to see which models fit the given data the most. For the test, we begin with our limited model and we have our full possible model as including every covariate.

Limited Model:

$$f_1(\text{Long\_Distance\_Trips }) = \beta_0 + \beta_1 f_2(\text{Fully\_Vaccinated\_Pct})$$

Full Model:

$$f_1(\text{Long\_Distance\_Trips }) = \beta_0 + \beta_1 f_2(\text{Fully\_Vaccinated\_Pct}) + \beta_2 f_3(\text{County\_Population}) + \beta_3 f_4(\text{Median\_County\_Income}) + \beta_4 \text{isRepublican} + \beta_5 f_5(\text{Unemployment\_Pct }) + \beta_6 f_7(\text{Median\_Age})$$

Where the transformations are:

$$\begin{aligned} f_1(x) &= \log_{10}(x) \\ f_2(x) &= \log_{10}(x) \\ f_3(x) &= \log_{10}(x) \\ f_4(x) &= \log_{10}(x) \\ f_5(x) &= \log_{10}(x) \\ f_6(x) &= \log_{10}(x) \\ f_7(x) &= \log_{10}(x) \end{aligned}$$

Running the AIC Test:

```
## Start:  AIC=-106.52
## Number.of.Long.Trips_log10 ~ Series.Complete.Pop.Pct_log10
##
##                              Df Sum of Sq    RSS     AIC
## + County.POP_log10            1    49.708  1.926 -515.59
## + Median.Age_log10            1    19.810 31.825 -165.01
## + County.Median.Income_log10  1    18.496 33.139 -159.95
## + isRepublican                1    18.222 33.413 -158.92
## + Unemployment.Pct_log10      1     6.034 45.600 -120.05
## <none>                                   51.635 -106.52
##
## Step:  AIC=-515.59
## Number.of.Long.Trips_log10 ~ Series.Complete.Pop.Pct_log10 +
##     County.POP_log10
##
##                              Df Sum of Sq    RSS     AIC
## + Median.Age_log10            1  0.061865 1.8645 -517.67
## + County.Median.Income_log10  1  0.055888 1.8705 -517.27
## + isRepublican                1  0.045225 1.8811 -516.56
## <none>                                   1.9264 -515.59
## + Unemployment.Pct_log10      1  0.003859 1.9225 -513.84
##
## Step:  AIC=-517.67
## Number.of.Long.Trips_log10 ~ Series.Complete.Pop.Pct_log10 +
##     County.POP_log10 + Median.Age_log10
##
##                              Df Sum of Sq    RSS     AIC
## + isRepublican                1  0.053733 1.8108 -519.32
## + County.Median.Income_log10  1  0.048687 1.8158 -518.97
## <none>                                   1.8645 -517.67
## + Unemployment.Pct_log10      1  0.007764 1.8567 -516.19
##
## Step:  AIC=-519.32
## Number.of.Long.Trips_log10 ~ Series.Complete.Pop.Pct_log10 +
##     County.POP_log10 + Median.Age_log10 + isRepublican
##
##                              Df Sum of Sq    RSS     AIC
## + County.Median.Income_log10  1  0.037326 1.7734 -519.92
## <none>                                   1.8108 -519.32
## + Unemployment.Pct_log10      1  0.008301 1.8025 -517.90
##
## Step:  AIC=-519.92
## Number.of.Long.Trips_log10 ~ Series.Complete.Pop.Pct_log10 +
##     County.POP_log10 + Median.Age_log10 + isRepublican + County.Median.Income_log10
##
##                           Df Sum of Sq    RSS     AIC
## <none>                                  1.7734 -519.92
## + Unemployment.Pct_log10  1 0.0016297 1.7718 -518.04
```

```
##
## Call:
## lm(formula = Number.of.Long.Trips_log10 ~ Series.Complete.Pop.Pct_log10 +
##     County.POP_log10 + Median.Age_log10 + isRepublican + County.Median.Income_log10,
##     data = d_log10)
##
## Coefficients:
##                   (Intercept)  Series.Complete.Pop.Pct_log10
##                       2.13713                       -0.43283
##              County.POP_log10               Median.Age_log10
##                       0.86149                       -0.42401
##                 isRepublican1     County.Median.Income_log10
##                       0.05273                       -0.23730
```

The AIC tests show that the model with the lowest AIC score, which fits the data the best, is actually the model that only excludes the unemployment percentage variable `Unemployment.Pct_log10`. This lowest AIC model will be our model 3. Our model 2 will be the model that includes the vaccination rate, county population, and the median age (`Series.Complete.Pop.Pct_log10`, `County.POP_log10`, `Median.Age_log10`). The AIC of this model 2 is -517.67 and choosing model 3 only decreases the AIC by approximately 2. Model 2 has a limited number of controls, but should capture the data almost as well as our best model 3 and these 3 variables are not going to be multicollinear.

Models

We present the following three models for our report.

Limited Model:

$$f_1(\text{Long\_Distance\_Trips }) = \beta_0 + \beta_1 f_2(\text{Fully\_Vaccinated\_Pct})$$

Model 2:

$$f_1(\text{Long\_Distance\_Trips }) = \beta_0 + \beta_1 f_2(\text{Fully\_Vaccinated\_Pct}) + \beta_2 f_3(\text{County\_Population}) + \beta_3 f_4(\text{Median\_Age})$$

Model 3:

$$f_1(\text{Long\_Distance\_Trips }) = \beta_0 + \beta_1 f_2(\text{Fully\_Vaccinated\_Pct}) + \beta_2 f_3(\text{County\_Population}) + \beta_3 f_4(\text{Median\_Age}) + \beta_4 \text{isRepublican} + \beta_3 f_5(\text{Median\_County\_Income})$$

Our transformations are:

$$\begin{aligned} f_1(\text{Long\_Distance\_Trips }) &= \log_{10}(\text{Long\_Distance\_Trips }) \\ f_2(\text{Fully\_Vaccinated\_Pct}) &= \log_{10}(\text{Fully\_Vaccinated\_Pct}) \\ f_3(\text{County\_Population}) &= \log_{10}(\text{County\_Population}) \\ f_4(\text{Median\_Age}) &= \log_{10}(\text{Median\_Age}) \\ f_5(\text{Median\_County\_Income}) &= \log_{10}(\text{Median\_County\_Income}) \end{aligned}$$

3. A Regression Table

Regression Table using Robust Standard Errors:

```
##
## ================================================================================================
========
##                                                  Dependent variable:
##                                -----------------------------------------------------------
--------
##                                                Number.of.Long.Trips_log10
##                                      (1)                    (2)                       (3)
## ------------------------------------------------------------------------------------------------
--------
## Series.Complete.Pop.Pct_log10       0.017                -0.612***                 -0.433**
##                                    (0.680)               (0.162)                   (0.183)
##
## County.POP_log10                                         0.822***                 0.861***
##                                                          (0.025)                   (0.033)
##
## Median.Age_log10                                          -0.419                    -0.424
##                                                          (0.320)                   (0.313)
##
## isRepublican1                                                                       0.053
##                                                                                    (0.036)
##
## County.Median.Income_log10                                                          -0.237
##                                                                                    (0.147)
##
## Constant                           3.994***              1.478**                   2.137**
##                                    (1.000)               (0.574)                   (0.878)
##
## ------------------------------------------------------------------------------------------------
--------
## Observations                         125                   125                       125
## R2                                 0.00001                0.964                     0.966
## Adjusted R2                        -0.008                 0.963                     0.964
## Residual Std. Error           0.648 (df = 123)       0.124 (df = 121)          0.122 (df =
119)
## F Statistic                  0.001 (df = 1; 123) 1,076.652*** (df = 3; 121) 669.157*** (df =
5; 119)
## ================================================================================================
========
## Note:                                                                *p<0.1; **p<0.05; *
**p<0.01
```

The above table shows the comparison of our Limited Model, Model_2 and Model_3. We note that both Model_2 and Model_3 are statistically significant.

Statistical Significance

Both Models 2 and 3 are statistically significant. The Adjusted R^2 of both the models are 0.96. Also, the standard errors are very near to each other with a value of 0.124 and 0.122 respectively. The β_1 in our models is a negative value. We have to note that both the outcome variable and the treatment variable are both log10 transformed.

From this, we notice that a 10% increase in vaccine rate (Treatment Variable) means a **reduction** of 4.2% (model 3) or 5.6% (model 2) of Long Trips (Outcome Variable).

Practical Significance

Even though we notice a statistical significance, we do not notice a practical significance. This is because a 10% increase in vaccine rate (reasonable increase in the short term) means a 4.2% (model 3) or 5.6% (model 2) decrease in number of tourism trips taken. If Disney was a destination expecting 1000 visitors for in a week for example, then getting 944 visitors instead does not warrant significant action in terms of promotion or switching over to maintenance over operation.

4. Limitations of your Model

> As a team, evaluate all of the CLM assumptions that must hold for your model. However, do not report an exhaustive examination all 5 CLM assumptions in the main report. Instead, bring forward only those assumptions that you or your reader might think pose significant problems for your analysis. For each problem that you identify, describe the statistical consequences. If you are able to identify any strategies to mitigate the consequences, explain these strategies.

Our model shows a decrease in trips taken with increasing vaccination rate. However, if vaccination rates were in the realm of 70% to 90% we would not expect this trend to hold since it means that the United States, or the West Coast States at least, are in a different state of the Pandemic than now. In short, our model does not extend to very large increases in vaccination rate that might put most counties to be almost fully vaccinated. Instead, our model provides causal insight into what small, reasonable increases in vaccination rate would impact tourism trips taken now.

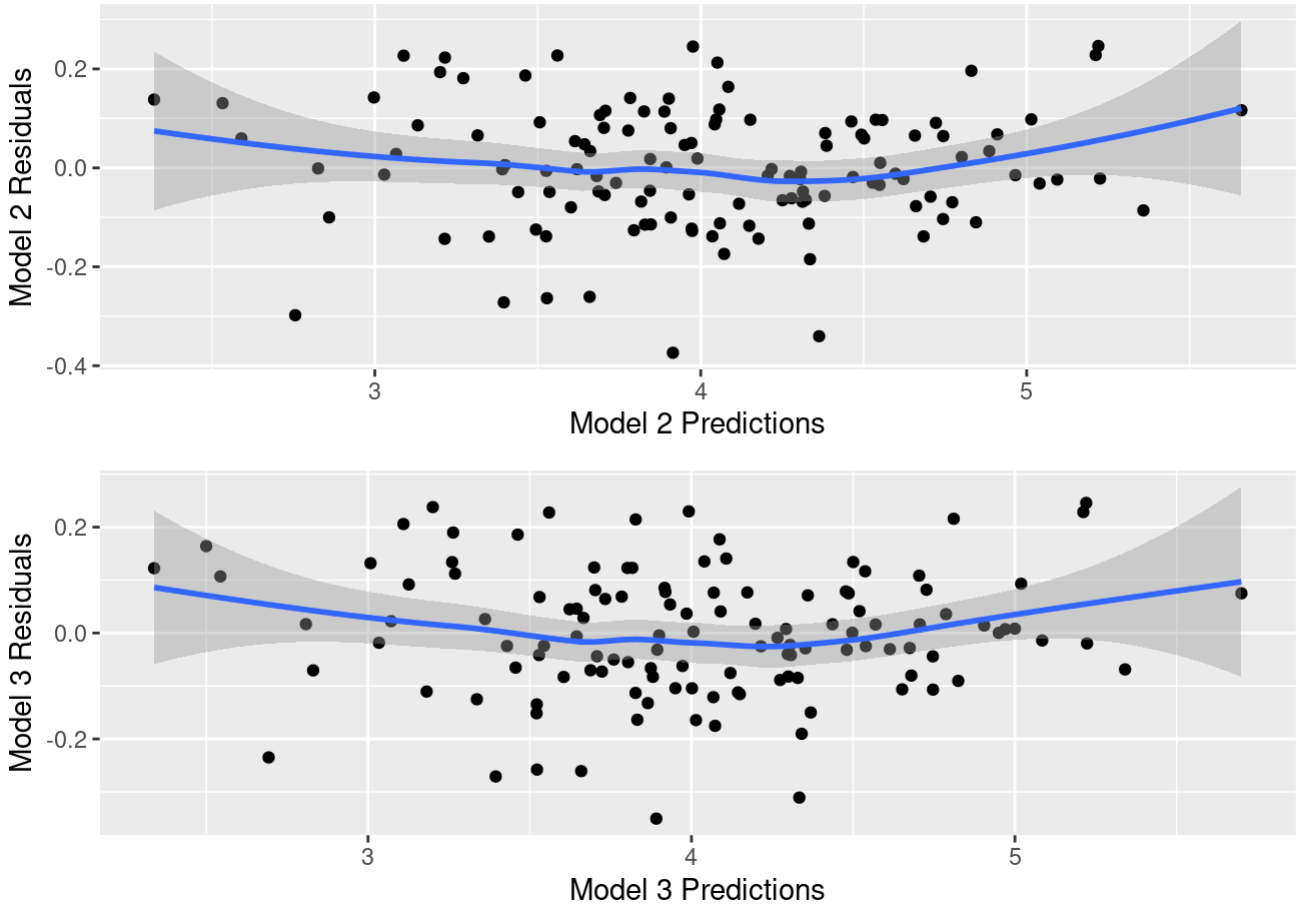
Below we will discuss and check whether our two models (not including the limited model) satisfy the Classical Liner Model assumptions.

CLM Assumption 1: Independent and Identically Distributed (IID)

The observations in our models are limited to all the counties in the continental west coast states of California, Oregon, and Washington. Most of the west coast region counties share similar social, cultural, political, economic, and geographical features. This similarity of features means that the counties in this region are mostly identical except for some deviations like population, median income level, and counties' political leaning (Democrat or a Republican party). Population, median income level, and political leaning can influence the number of long-distance trips taken in a county, in addition to the percentage of vaccinated people. To control the differences due to these additional features, we include population, median income level, and 2020 county-level presidential results (party with the highest percentage of votes) as covariates in our model and assume that the observations are identical. We also acknowledge that some neighboring counties or clusters of counties in this region can influence or cause some dependency between the observations. Hence, the assumption of independence is not met in the sample. Based on the deficiencies in the IID assumption, we need to be wary of interpreting the model results and understand that the actual results may vary more than what the model suggests, i.e., the standard errors in the model are likely lower than actual uncertainty.

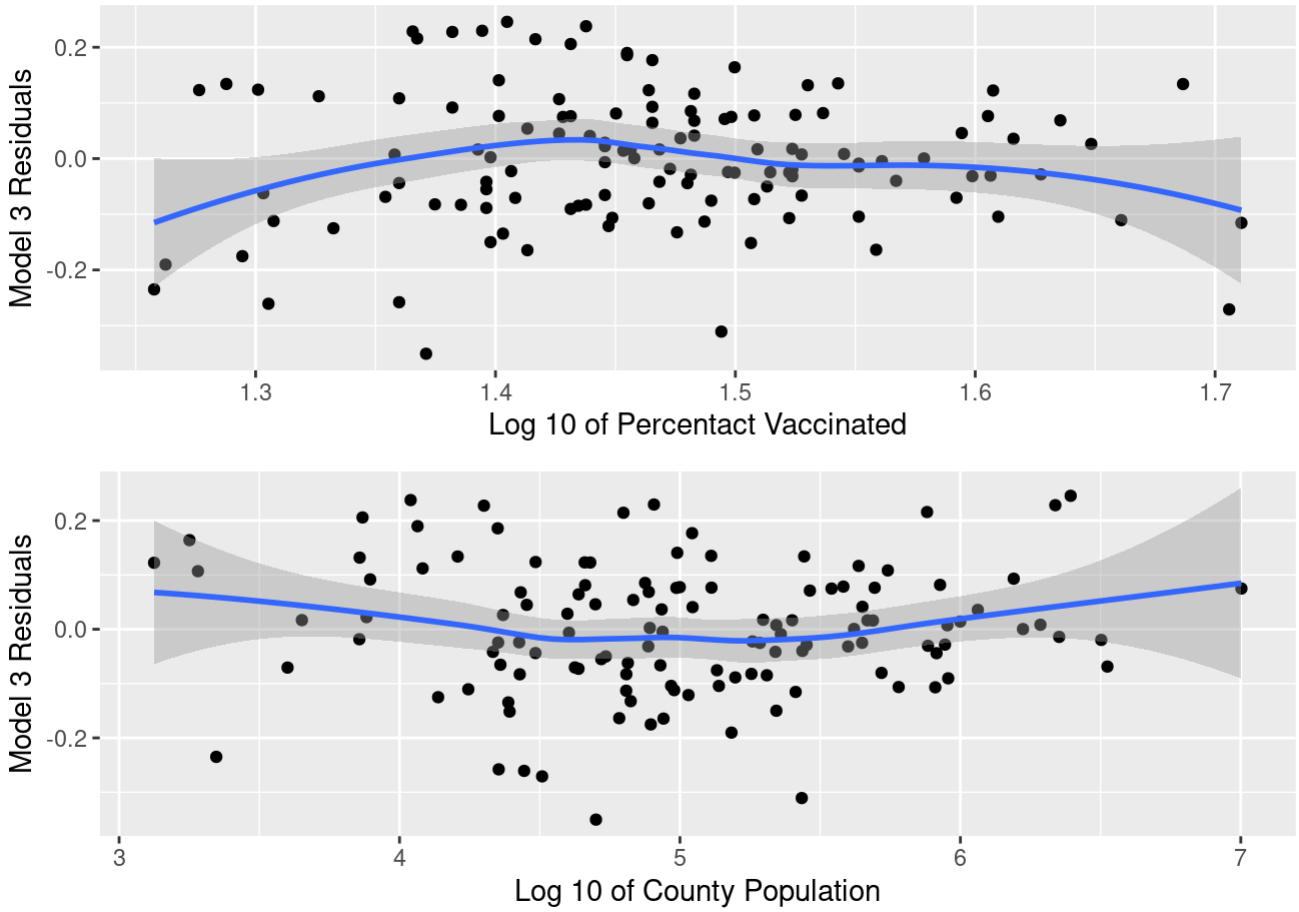
CLM Assumption 2: Linear Conditional Expectation

We validate the linear conditional expectation assumption for our models 2 and 3 by a visual inspection of the plot between the model residuals on the y-axis and the model predicted values on the x-axis. The plots show that the residuals line for both the models 2 and 3 are very close to 0 and only reaching 0.1 at the tails.

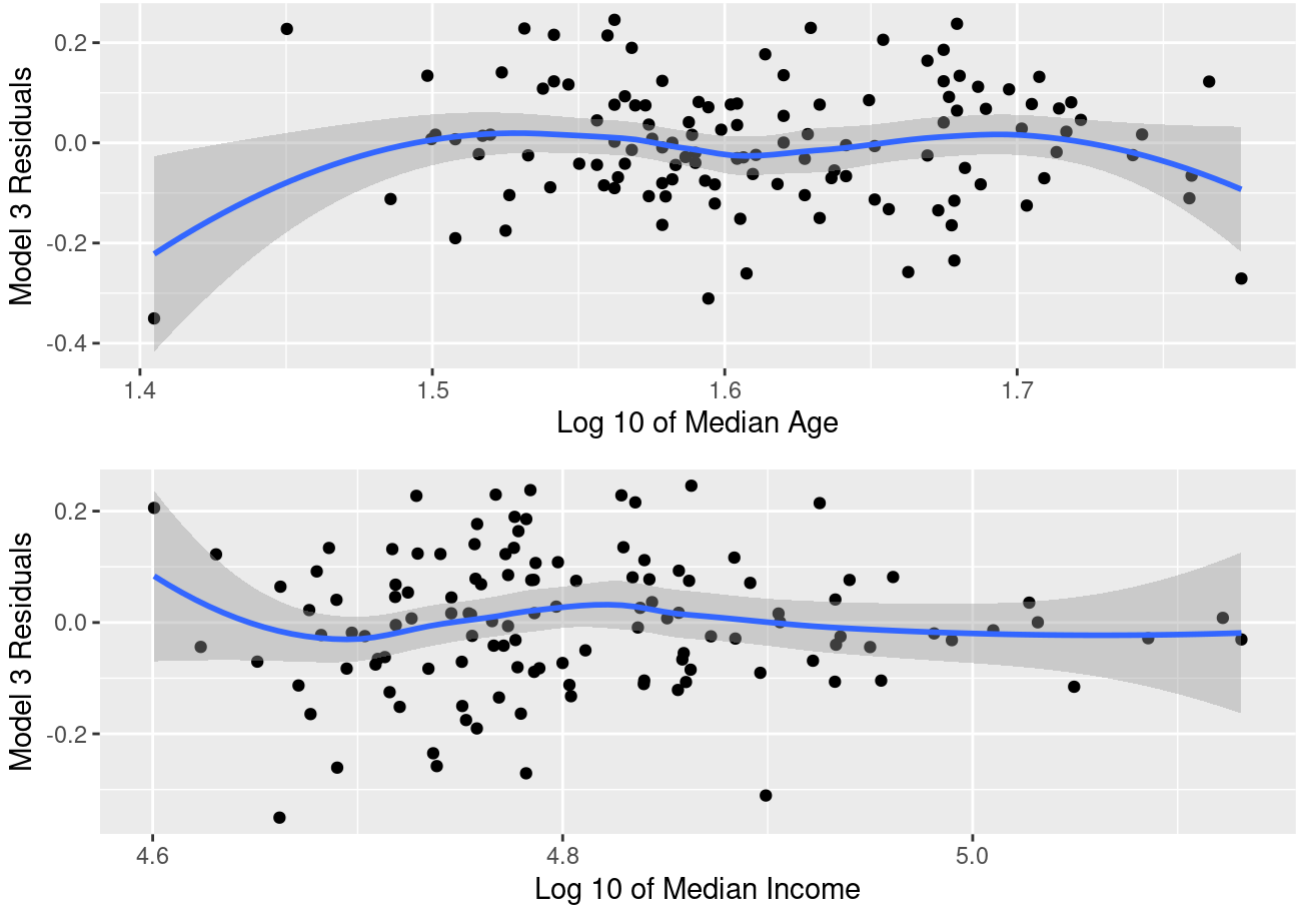


Model Residuals and Predictions

While the models look good in meeting the linear conditional expectation assumption, we check if there are any issues with some of the key covariates included in our model 3 by plotting the model residuals on the y-axis with the covariates on the x-axis. The covariates selected for these plots with the model 3 residuals are the log10 values of the following variables: percentage of population vaccinated in each county, the total population of the county, median age of county residents, and the median income of the county residents. All the plots between the model residuals and the covariates look good and the residual lines are closer to 0 in each of the plots and do not cross 0.2. There seems to be no issues with the covariates included in the model. The linear conditional expectation is met for both models 2 and 3.



Model Residuals and Variables



Model Residuals and Variables

CLM Assumption 3: No perfect collinearity

If one of the covariates can be written as a linear combination of another covariate, it is not possible to solve for the OLS estimator. In such cases we say that there is perfect multicollinearity. We can check for multicollinearity in two different ways 1) By using the Variance Inflation Factor (VIF) of the regression coefficients in the model 2) Calculating the paired correlation values of the covariates in the model.

We checked the VIF of both models 2 and 3. Below are the results

```
## Model 2 VIF
```

```
## Series.Complete.Pop.Pct_log10             County.POP_log10
##                      1.260643                     1.591182
##              Median.Age_log10
##                      1.743492
```

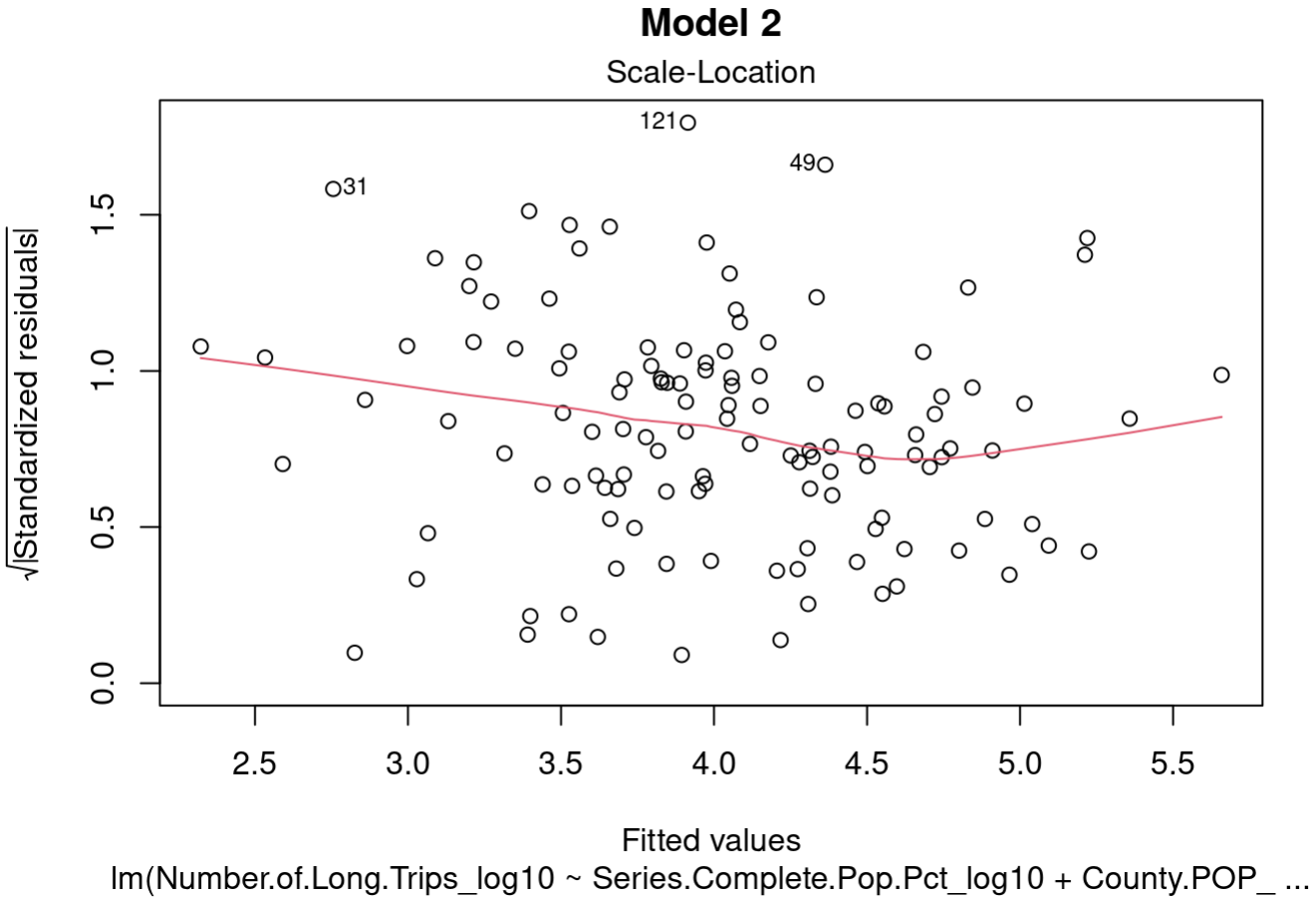
```
##
## Model 3 VIF
```

```
## Series.Complete.Pop.Pct_log10             County.POP_log10
##                       1.651518                    2.794738
##              Median.Age_log10                isRepublican
##                       1.763073                    2.020928
##    County.Median.Income_log10
##                       2.004960
```

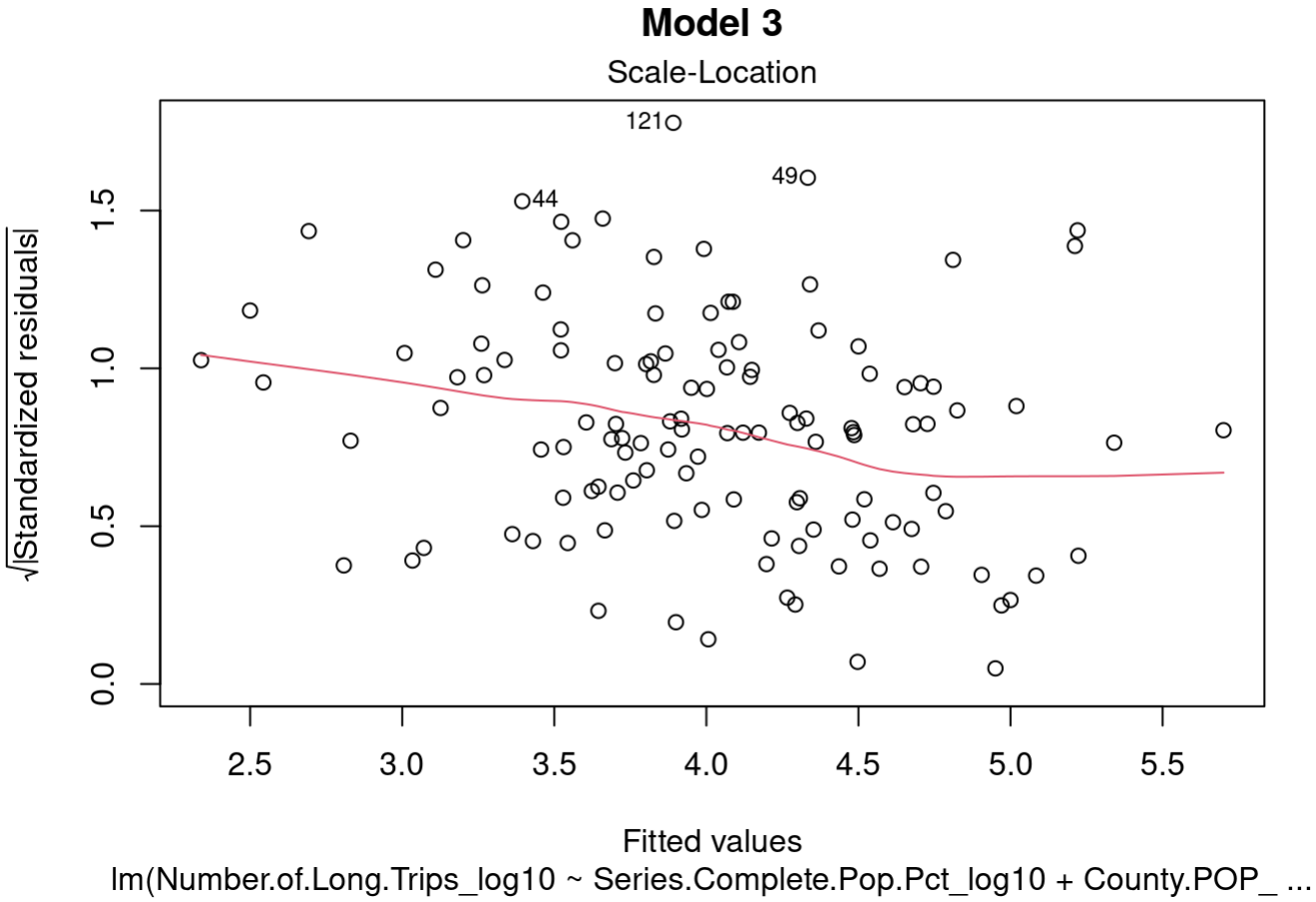
Usually, when VIF of the regression coefficients is less than 4 we say there is no perfect collinearity. We can see that the VIF of the regression coefficients is less than 4 in both model 2 and 3. The assumption of no perfect collinearity is met for both the models.

CLM Assumption 4: Homoskedastic Errors

In the CLM model, we assume that the errors are homoskedastic. This means that the errors terms do not vary much based on the predicted values in the model. We can evaluate homoskedastic assumption by viewing the plot between the standardized residuals and the fitted values or by performing a Breusch–Pagan test on the model. If the errors are homoscedastic, the line in the standardized residuals vs. the fitted values plot will be horizontal. We can see that for both our models 2 and 3, the line is not horizontal and there is a slope to it. There seems to be larger variation associated with the error terms based on different predicted values. This observation can be confirmed by the results of the Breusch–Pagan test on both the models which produce a p-value less than 0.05 that is statistically significant. Based on the p-values, the null hypothesis that the error variances are equal is rejected. Models 2 and 3 do not meet the assumption of homoscedastic errors. We should understand that the estimated errors in these models may not reflect the true uncertainty. Hence, any conclusion arrived at based on the errors and the hypothesis tests on these models may not be reliable.



Standardized Residuals and Fitted Values



Standardized Residuals and Fitted Values

```
##
##  studentized Breusch-Pagan test
##
## data:  model_2
## BP = 11.173, df = 3, p-value = 0.01082
```

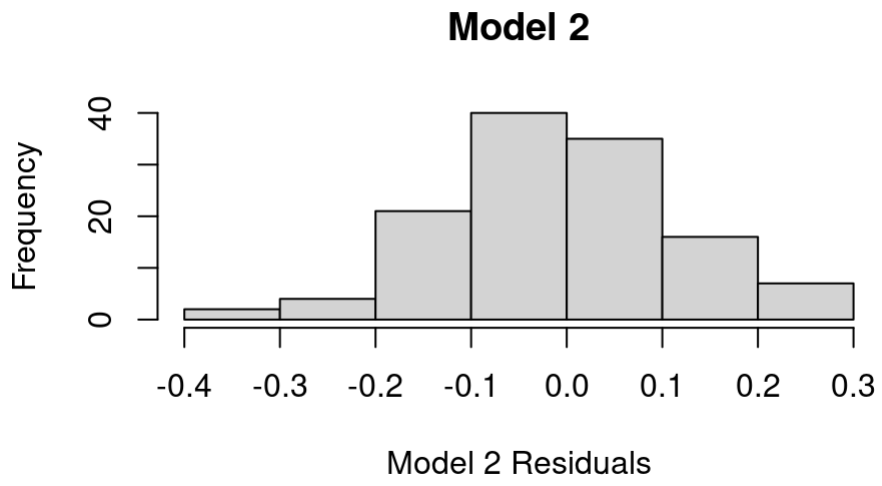
```
##
##  studentized Breusch-Pagan test
##
## data:  model_3
## BP = 19.363, df = 5, p-value = 0.001645
```

The Breusch-Pagan test rejects the null hypothesis for homoskedastic conditional variance for models 2 & 3. We should use robust standard errors in our reporting because we do not have homoskedasticity and our models do not meet this CLM assumption.

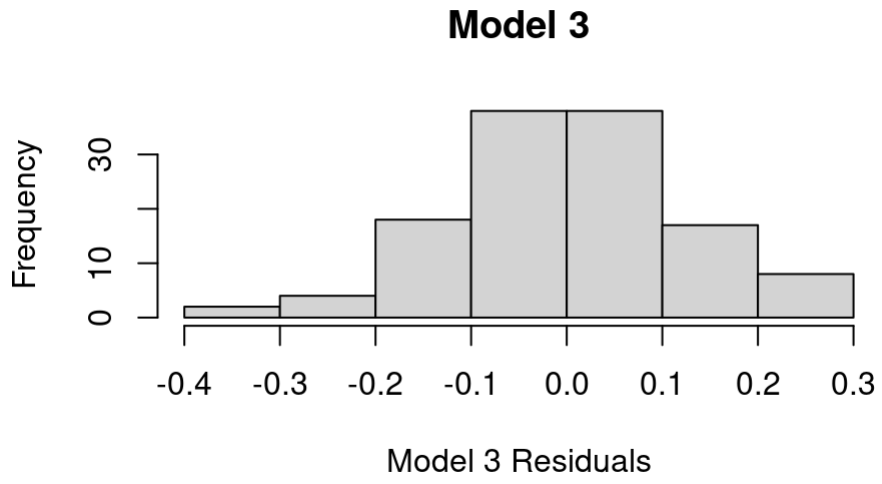
CLM Assumption 5: Normally Distributed Errors

Normality of errors is another key assumption of the classical liner model. We can verify the normal distribution assumption by viewing the histogram of the model residuals or by viewing a Q-Q norm plot between the theoretical values and the standardized residuals. The histogram of the residuals of both models look almost normal. The Q-Q norm plot confirms that the actuals and the expectations fall along the same line for a major portion except for

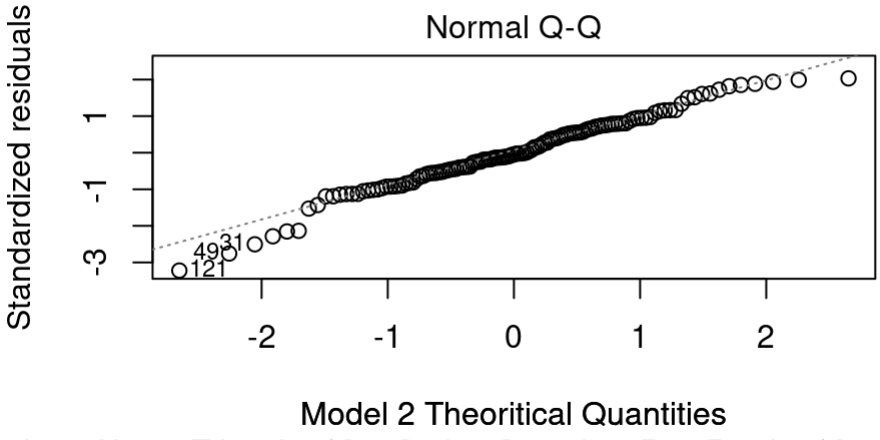
the ends in both the models. The plots are not too bad as the central portion is looking good. Usually, Q-Q norm plots tend to curve at the tails due to lack of data. We can confirm that our models 2 and 3 meet the assumption of normally distributed errors.



Residuals Histogram



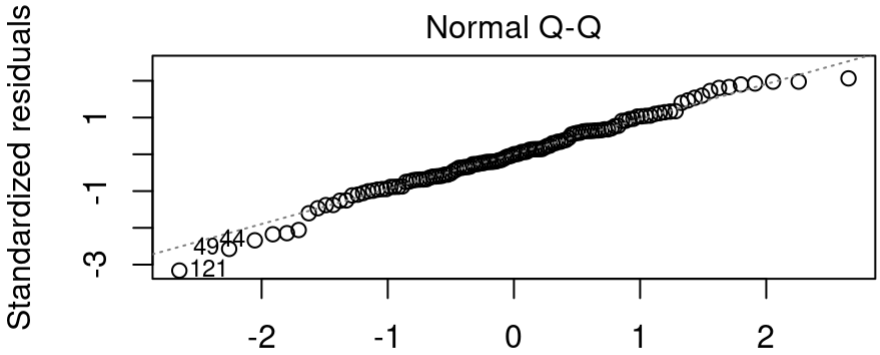
Residuals Histogram



Model 2 Theoritical Quantities

mber.of.Long.Trips_log10 ~ Series.Complete.Pop.Pct_log10 + Cour

Q-Q Norm plots



Model 3 Theoritical Quantities

mber.of.Long.Trips_log10 ~ Series.Complete.Pop.Pct_log10 + Cour

Q-Q Norm plots

```
##
##  Shapiro-Wilk normality test
##
## data:  d_log10$model_2_residuals
## W = 0.98109, p-value = 0.07764
```

```
##
##  Shapiro-Wilk normality test
##
## data:  d_log10$model_3_residuals
## W = 0.98715, p-value = 0.2894
```

The Shapiro-Wilk normality test on the residuals of model 2 and model 3 both result in a p-value exceeding 0.05. This means that we fail to reject the null hypothesis of normally distributed residuals. Our assumption that Model 2 and Model 3 have normally distributed residuals is met.

Note that you may need to change your model specifications in response to violations of the CLM.

5. Discussion of Omitted Variables

If the team has taken up an explanatory (i.e. causal) question to evaluate, then identify what you think are the most important omitted variables that bias results you care about. For each variable you name, you should reason about the direction of bias caused by omitting this variable. If you can argue whether the bias is large or small, that is even better. State whether you have any variables available that may proxy (even imperfectly) for the omitted variable. Pay particular attention to whether each omitted variable bias is towards zero or away from zero. You will use this information to judge whether the effects you find are likely to be real, or whether they might be entirely an artifact of omitted variable bias.

Our model tries to find the causal relationship between the number of long trips undertaken by the residents of the counties in the continental west coast states and the percentage of fully vaccinated people in these counties. While we try to improve our model's fit by including a number of applicable covariates that affect the predictor variable, we also acknowledge that there may be some omitted variables (OV) that we should be aware of. In this section we talk about the omitted variables and their potential impact on the models.

1. OV - Policy of lifting mask ban: We tried to gather the policy data, but unable to find the details at the county level. Hence, we discuss its potential impact of this omitted variable on our model, if such a data was available. Based on the model coefficients calculated, we see that the number of long trips has a negative relationship with the vaccination rate. We anticipate that higher vaccination percentages would more likely lead to lifting a mask ban which is a positive relationship. Lifting a mask ban would could boost the confidence of the county residents that the pandemic is under control and that might have a positive relationship with the number of long trips. With these relationships the estimated coefficient of vaccination percentage is greater than the true coefficient and the direction of bias is towards 0.
2. OV – COVID Fatigue: COVID fatigue measures how tired people are of being in a lockdown after a lockdown is lifted. There is no data available on COVID fatigue. Hence, we discuss its potential impact of this omitted variable on our model, if such a data was available. We discussed above that the number of long trips has a negative relationship with the vaccination rate. We anticipate that more fatigue people have the more likely they will travel which is a positive relationship. COVID fatigue can have a positive relationship with the vaccination rates as people want to carry on with their daily lives worry free. With these relationships the estimated coefficient of vaccination percentage is greater than the true coefficient and the direction of bias is towards 0.
3. OV – Vaccine Hesitancy: This variable measure the percentage of people at the county level who describe themselves as "unsure", "probably not", or "definitely not" going to get a COVID-19 vaccine. There is no vaccine hesitancy data available at the county level. Hence, we discuss its potential impact of this omitted variable on our model, if such a data was available. We discussed above that the number of long trips has a negative relationship with the vaccination rate. Having a higher percentage of vaccine hesitancy or skepticism means that there may be a positive relationship with the number of trips, since these people may

not believe in science. Getting more people vaccinated could reduce the vaccine hesitancy to some extent because it becomes normalized and trusted, so the relationship is negative. With these relationships the estimated coefficient of vaccination percentage is lesser than the true coefficient and the direction of bias away from 0.

6. Conclusion

> Make sure that you end your report with a discussion that distills key insights from your estimates and addresses your research question.

Our analysis shows that tourism demand is slightly decreasing with rising vaccination rates, but not at a practically significant amount. This shows that, even though people are getting vaccinated they are wary about the possibility of uncertainty looming around with different variants and may not be feeling safe to take vacations.

Based on the research conducted above, we can recommend the following to **Tourism Board of Continental West Coast States**:

- It is recommended to exercise caution prior to promoting or increasing any of the expenditure on tourism, because based on the research people are not yet feeling comfortable to take vacations even after being fully vaccinated.
- Reevaluate promoting tourism in the future when the vaccination rates are much higher across counties. The Covid-19 climate in the West Coast might be different than now if all vaccination rates are above the threshold of 80% compared to the current 29.9% average.