

covid tree analysis

Cross sectional Analysis of Covid-19 Spread in US Counties

Covid-19 or corona virus started its spread from China at the end of Year 2020. By February 2020, cases start to appear in the United States and spread rapidly across all the states. The growth in the number of cases has been exponential like any other case of infectious disease. Scientists and researchers around the world are actively looking at the data to understand this virus. In uncertain times like these, it is important to know the factors that affect the spread of this virus to make plausible predictions about the future.

Since the first case of covid-19 was discovered in the US in Snohomish county in the Washington State, the virus has now spread to other places and in some cities like New York and Chicago, the spread has been faster, making them new hotspots for the virus even though cities like Seattle & Los Angeles reported their first cases before these cities. A simple line graphs for some counties that saw their first case of infection around the same time shows the difference in the rate of virus spread.

```
## Warning: package 'ggplot2' was built under R version 3.6.2
```

```
## Warning: package 'ggrepel' was built under R version 3.6.2
```

This poses a question that; what are the different factors that are leading to differences in the pace of covid-19 spread across US counties? Can a cross sectional comparison of these counties help us identify any socio-economic or geographical feature that has any effect on virus spread and can we use these features to predict the number of infections for other counties that are behind in trend. We carry out a cross sectional analysis of counties and their features in comparison to the number of covid-19 cases reported by using statistical and data mining tools. The purpose of the report is to identify any trend or characteristics of counties that are connected to the pace of virus spread.

Data

- **Cases & Deaths:** We collect the data of total number of covid-19 infection cases and deaths at county level as on April 30, 2020. The data has been obtained from NYTimes Github repository.

We selected 1462 counties which had non-zero number of deaths reported. This was done because ACS data was only available for counties that have any death reported for covid-19.

We divided the total number of cases by population to have cases per hundred thousand of population. This gives a good comparison of cases across counties with different population size. From now on we will refer to cases as cases per hundred thousand of population. The density graph shows that a lot of counties have lower number of cases and a very skewed distribution of cases.

- **Days since first Case:** We calculated the total number of days since 1st infection for all counties as on April 30, 2020. This variable measures the time period for spread of virus. A scatter plot of days since first infection and number of cases per 100 thousand people shows that there is a general exponential trend for but some counties have been able to keep their number of cases down even though they got the virus before other counties.

```
ggplot(covid, mapping = aes(days_since_first_infection1,cases1))+geom_point()+labs(title= "Scatter Plot
```

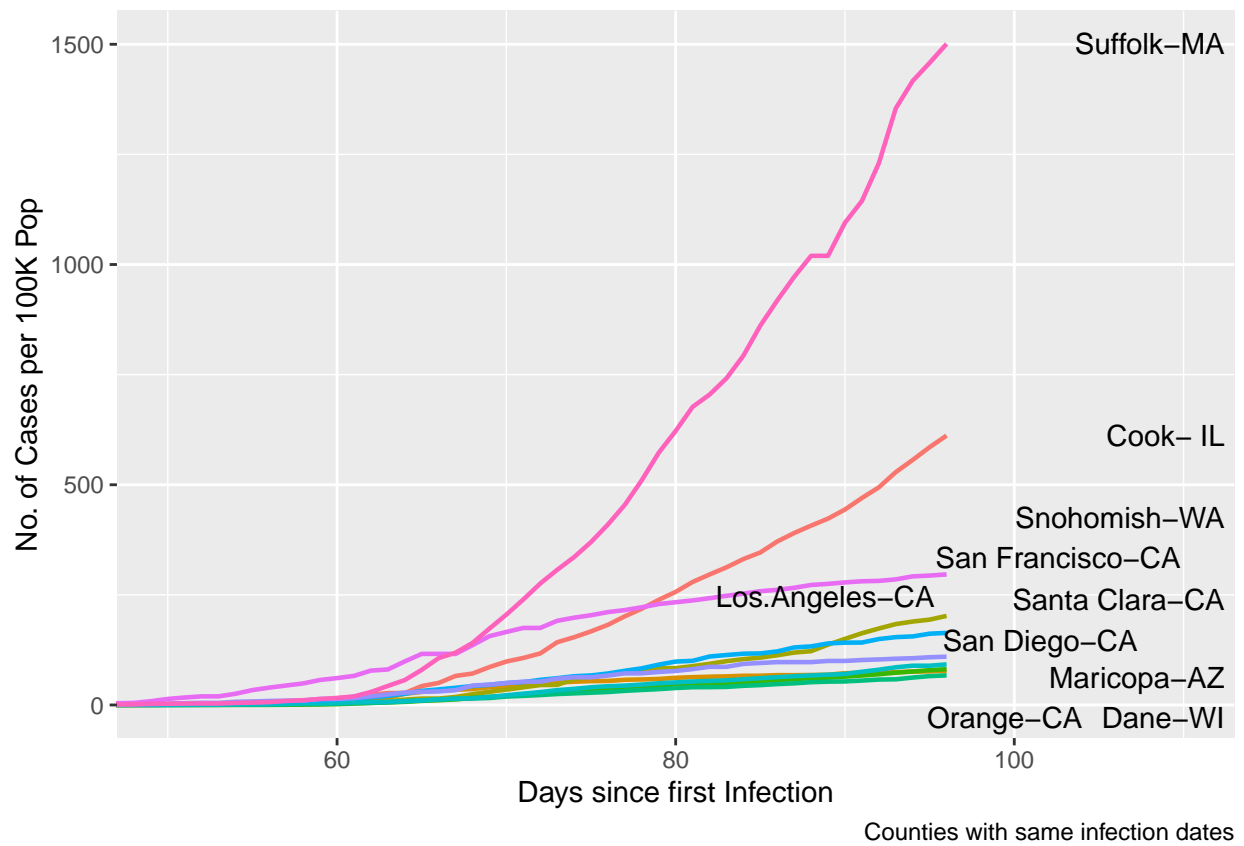


Figure 1: Trend Line Graph for Counties

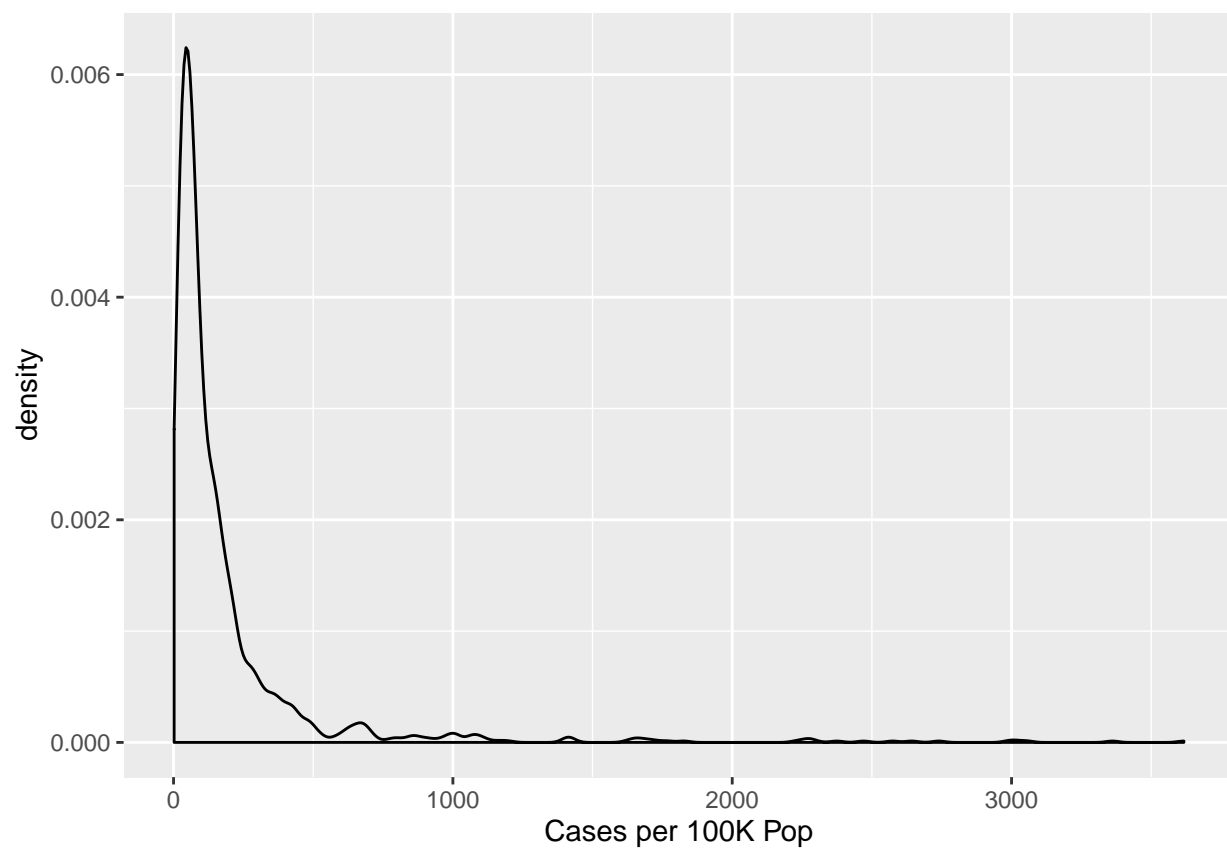
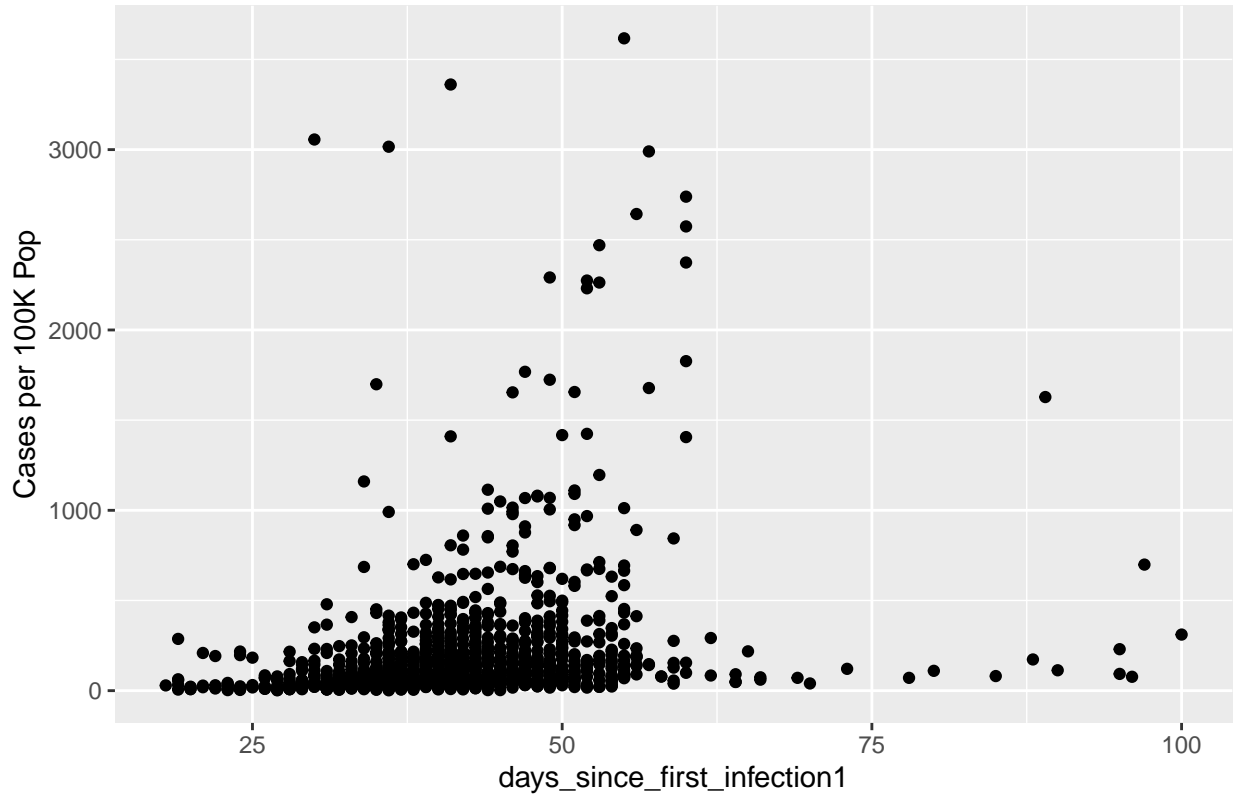


Figure 2: Density Graph for Number of Cases

Scatter Plot of Time and Cases



- **Days since Stay at home orders:** Similarly, we calculated number of days since stay at home orders were issued by State and County authorities as on April 30, 2020. This gives us the time period for interventions taken by authorities. Along with this we also obtained Google's movement analytics for each county which gives a percentage decline in people's movement from the baseline average.
- **Socio-Economic Variables:** We obtained socio economic variables on the population of counties from American Community Survey reports. Main variables include: Unemployment rate, % of White, % of Households in Poverty, % of population employed in different industries, median income and % of population with a bachelor's degree.
- **Demographic Variables:** % of male, % of age between 18 and 64, % of White, Median age, % of total households as family etc. The complete list is attached in the appendix.
- **Geographical variables:** We obtained variables like population density, average of last 8 years highest summer temperature and humidity rate. We wanted to obtain the recent monthly averages for counties but could not find the latest data, so we used these variables.

There were many socio-economic variables that were very highly correlated, so we dropped some variables and finalized 30 variables for this analysis. The complete list of these variables is provided in the appendix.

Note : Before we start with our analysis, it is important that we acknowledge that number of cases reported depends heavily on the number of tests conducted. Unfortunately, testing policy has not been uniformed across US and that can lead to a measurement error in cases. Secondly, there are reports that the given time frame for start of covid-19 is not accurate and it is presumed that covid-19 reached many major cities of the US way before the first case was reported.

However, due to data limitation, we must assume that the official reporting of cases represents somewhat actual position for the start of virus infection. For testing, we tried to get data on total tests conducted on county level but only find the data at state level. So, we calculated total tests conducted per hundred thousand people for each state and then use it for each county in respective state. This is not perfect, but we

are assuming that the rate of testing is uniform across counties in each state and will help in differentiating counties of different states based on testing rates.

Methods

Principal Component Analysis

We use principal component analysis to see if we can look at the variations in the counties and evaluate if those variations are related to number of cases in these counties. After scaling the data, we see that 90% of variation in 30 explanatory variables can be explained by 16 principal components, pointing towards correlation between these variables.

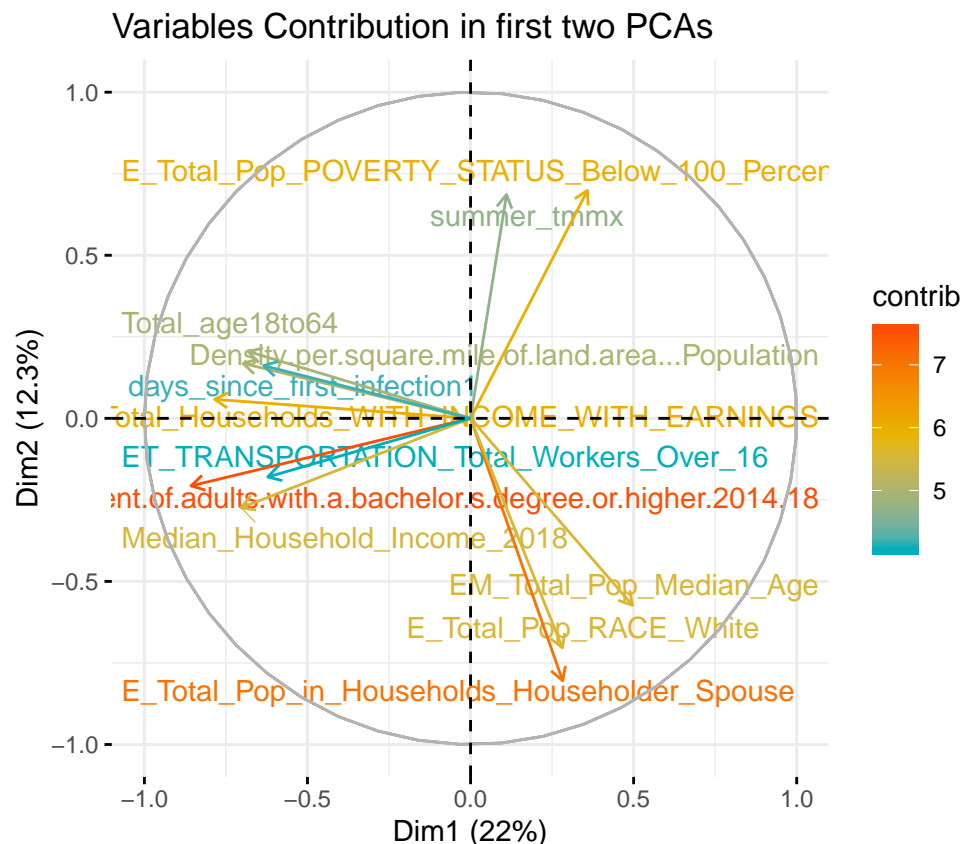
The biplot of main variables that contribute to variance in first two PCs shows that PC1 captures variation in population density, time since first case, proportion of population between 18 and 64 and change in the movement of people to their workplaces.

```
library(factoextra)

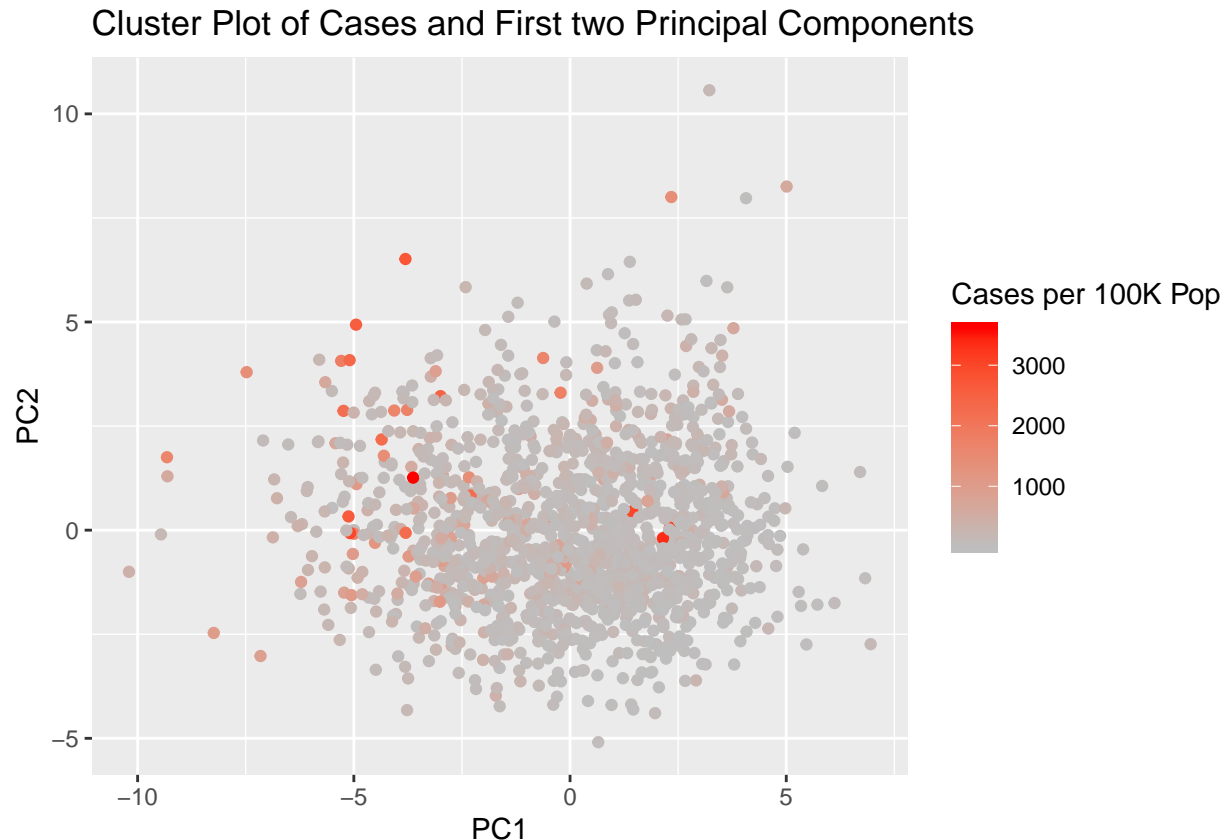
## Warning: package 'factoextra' was built under R version 3.6.3

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

covid = read.csv("covid2zargham.csv")
X= covid[,-c(1:5, 35,37)]
PCA = prcomp(X, scale. = TRUE)
fviz_pca_var(PCA, col.var="contrib",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),select.var = list(contrib=12),
             repel = TRUE, title = "Variables Contribution in first two PCAs ")
```



```
scores = PCA$x
scores= cbind(covid$cases1, scores)
scores = as.data.frame(scores)
qplot(PC1,PC2, data=scores, color= V1)+scale_color_gradient(low = "grey", high= 'red')+ggtitle("Cluster
```



Plotting the scatter plot with first two components with color scaling the number of cases, we see that there is no specific clustering of high cases counties but generally, we see that there are more red points in the upper left quadrant of the graph. Which means that variables mentioned above are correlated with higher number of covid-19 cases in counties. Similarly, lower right quadrant has less number of high case counties and from the biplot we saw that this is associated with higher proportion of white population, higher median age and higher proportion of spouse households in total households. Interestingly, we cannot say much about the effect of temperature on virus spread but there does seem to be a slight negative correlation between temperature and virus spread.

Negative Binomial Logistic Model

Since our outcome is count of cases with overdispersion, we fit a negative binomial logistic model on our data. As we have only those counties that reported the any cases, we use zero truncated negative binomial logistic model. This model has been used by Wu, Nethery, Sabath (2020) in their study of exposure to air pollution and its effect on covid death rate in US counties. We include all the explanatory variables in our model to which of them are correlated with cases count. The summary of the model is provided in the appendix.

```
## Warning: package 'VGAM' was built under R version 3.6.3
## Loading required package: stats4
## Loading required package: splines
```

```
##
## Call:
## vglm(formula = cases1 ~ ., family = pospoisson(), data = covid1)
##
## Pearson residuals:
##           Min       1Q Median       3Q      Max
## loglink(lambda) -27.12 -6.795 -3.276  2.381 248.3
##
## Coefficients:
##                                     Estimate
## (Intercept)                      -3.219e+01
## days_since_first_infection1       7.335e-03
## Tests.per.100thousand             2.861e-04
## stayhome                        -1.311e-02
## Percent.of.adults.with.a.bachelor.s.degree.or.higher.2014.18 -4.185e+00
## Unemployment_rate_2018           -3.766e+00
## Total_age18to64                  -4.171e+00
## E_Total_Pop_SEX_Male              1.083e+01
## E_Total_Pop_RACE_One_Race         1.315e+01
## E_Total_Pop_RACE_White            -2.034e+00
## E_Total_Pop_in_Households_Householder_Spouse -1.128e+00
## E_Total_Pop_in_Households_Householder_Parent  6.089e+00
## E_Total_Pop_Over_15_MARITAL_STATUS_Divorced -6.925e+00
## E_INDUSTRY_Administrative_Waste_Management -1.159e+01
## E_INDUSTRY_Educational             1.678e+00
## E_INDUSTRY_Healthcare_Social_Assistance  6.860e+00
## E_INDUSTRY_Arts_Entertainment_Recreation -5.160e+00
## E_INDUSTRY_Accommodation_Food_Services -1.913e+01
## E_INDUSTRY_Other_Services          -2.508e+01
## E_INDUSTRY_Public_Administration    -3.197e+00
## ET_TRANSPORTATION_Total_Workers_Over_16 -1.166e+00
## E_TRANSPORTATION_Total_Workers_Over_16_Drove_Alone  1.863e+00
## E_Total_Households_TYPE_Family     -3.567e+00
## E_Total_Households_WITH_INCOME_WITH_EARNINGS  6.603e+00
## E_Total_Pop_POVERTY_STATUS_Below_100_Percent  1.880e+00
## summer_tmmx                      -1.090e-02
## summer_rmax                      -1.244e-03
## Density.per.square.mile.of.land.area...Population  2.244e-01
## EM_Total_Pop_Median_Age            2.702e+00
## Median_Household_Income_2018       1.214e+00
## workplaces_percent_change_from_baseline -5.393e+00
##                                     Std. Error
## (Intercept)                      5.269e-01
## days_since_first_infection1       2.793e-04
## Tests.per.100thousand             1.889e-06
## stayhome                        2.799e-04
## Percent.of.adults.with.a.bachelor.s.degree.or.higher.2014.18 5.270e-02
## Unemployment_rate_2018           2.551e-01
## Total_age18to64                  1.281e-01
## E_Total_Pop_SEX_Male              2.050e-01
## E_Total_Pop_RACE_One_Race         2.080e-01
## E_Total_Pop_RACE_White            1.839e-02
## E_Total_Pop_in_Households_Householder_Spouse 4.926e-02
## E_Total_Pop_in_Households_Householder_Parent 6.085e-01
```

```

## E_Total_Pop_Over_15_MARITAL_STATUS_Divorced 2.058e-01
## E_INDUSTY_Administrative_Waste_Management 5.913e-01
## E_INDUSTY_Educational 2.249e-01
## E_INDUSTY_Healthcare_Social_Assistance 1.831e-01
## E_INDUSTY_Arts_Entertainment_Recreation 5.069e-01
## E_INDUSTY_Accommodation_Food_Services 3.385e-01
## E_INDUSTY_Other_Services 6.790e-01
## E_INDUSTY_Public_Administration 2.027e-01
## ET_TRANSPORTATION_Total_Workers_Over_16 5.720e-02
## E_TRANSPORTATION_Total_Workers_Over_16_Drove_Alone 6.203e-02
## E_Total_Households_TYPE_Family 7.406e-02
## E_Total_Households_WITH_INCOME_WITH_EARNINGS 8.664e-02
## E_Total_Pop_POVERTY_STATUS_Below_100_Percent 2.931e-02
## summer_tmmx 1.025e-03
## summer_rmax 2.945e-04
## Density.per.square.mile.of.land.area...Population 2.834e-03
## EM_Total_Pop_Median_Age 4.885e-02
## Median_Household_Income_2018 2.517e-02
## workplaces_percent_change_from_baseline 4.572e-02
## z value Pr(>|z|)
## (Intercept) -61.099 < 2e-16
## days_since_first_infection1 26.264 < 2e-16
## Tests.per.100thousand 151.442 < 2e-16
## stayhome -46.830 < 2e-16
## Percent.of.adults.with.a.bachelor.s.degree.or.higher.2014.18 -79.416 < 2e-16
## Unemployment_rate_2018 -14.762 < 2e-16
## Total_age18to64 -32.565 < 2e-16
## E_Total_Pop_SEX_Male 52.818 < 2e-16
## E_Total_Pop_RACE_One_Race 63.235 < 2e-16
## E_Total_Pop_RACE_White -110.595 < 2e-16
## E_Total_Pop_in_Households_Householder_Spouse -22.901 < 2e-16
## E_Total_Pop_in_Households_Householder_Parent 10.007 < 2e-16
## E_Total_Pop_Over_15_MARITAL_STATUS_Divorced -33.647 < 2e-16
## E_INDUSTY_Administrative_Waste_Management -19.608 < 2e-16
## E_INDUSTY_Educational 7.458 8.78e-14
## E_INDUSTY_Healthcare_Social_Assistance 37.461 < 2e-16
## E_INDUSTY_Arts_Entertainment_Recreation -10.178 < 2e-16
## E_INDUSTY_Accommodation_Food_Services -56.514 < 2e-16
## E_INDUSTY_Other_Services -36.932 < 2e-16
## E_INDUSTY_Public_Administration -15.774 < 2e-16
## ET_TRANSPORTATION_Total_Workers_Over_16 -20.381 < 2e-16
## E_TRANSPORTATION_Total_Workers_Over_16_Drove_Alone 30.036 < 2e-16
## E_Total_Households_TYPE_Family -48.158 < 2e-16
## E_Total_Households_WITH_INCOME_WITH_EARNINGS 76.216 < 2e-16
## E_Total_Pop_POVERTY_STATUS_Below_100_Percent 64.134 < 2e-16
## summer_tmmx -10.636 < 2e-16
## summer_rmax -4.224 2.40e-05
## Density.per.square.mile.of.land.area...Population 79.181 < 2e-16
## EM_Total_Pop_Median_Age 55.304 < 2e-16
## Median_Household_Income_2018 48.239 < 2e-16
## workplaces_percent_change_from_baseline -117.952 < 2e-16
##
## (Intercept) ***
## days_since_first_infection1 ***

```



```

## Tests.per.100thousand ***
## stayhome ***
## Percent.of.adults.with.a.bachelor.s.degree.or.higher.2014.18 ***
## Unemployment_rate_2018 ***
## Total_age18to64 ***
## E_Total_Pop_SEX_Male ***
## E_Total_Pop_RACE_One_Race ***
## E_Total_Pop_RACE_White ***
## E_Total_Pop_in_Households_Householder_Spouse ***
## E_Total_Pop_in_Households_Householder_Parent ***
## E_Total_Pop_Over_15_MARITAL_STATUS_Divorced ***
## E_INDUSTY_Administrative_Waste_Management ***
## E_INDUSTY_Educational ***
## E_INDUSTY_Healthcare_Social_Assistance ***
## E_INDUSTY_Arts_Entertainment_Recreation ***
## E_INDUSTY_Accommodation_Food_Services ***
## E_INDUSTY_Other_Services ***
## E_INDUSTY_Public_Administration ***
## ET_TRANSPORTATION_Total_Workers_Over_16 ***
## E_TRANSPORTATION_Total_Workers_Over_16_Drove_Alone ***
## E_Total_Households_TYPE_Family ***
## E_Total_Households_WITH_INCOME_WITH_EARNINGS ***
## E_Total_Pop_POVERTY_STATUS_Below_100_Percent ***
## summer_tmmx ***
## summer_rmax ***
## Density.per.square.mile.of.land.area...Population ***
## EM_Total_Pop_Median_Age ***
## Median_Household_Income_2018 ***
## workplaces_percent_change_from_baseline ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Name of linear predictor: loglink(lambda)
##
## Log-likelihood: -97019 on 1411 degrees of freedom
##
## Number of Fisher scoring iterations: 10
##
## Warning: Hauck-Donner effect detected in the following estimate(s):
## '(Intercept)', 'Total_age18to64', 'E_Total_Households_TYPE_Family', 'summer_tmmx'

```

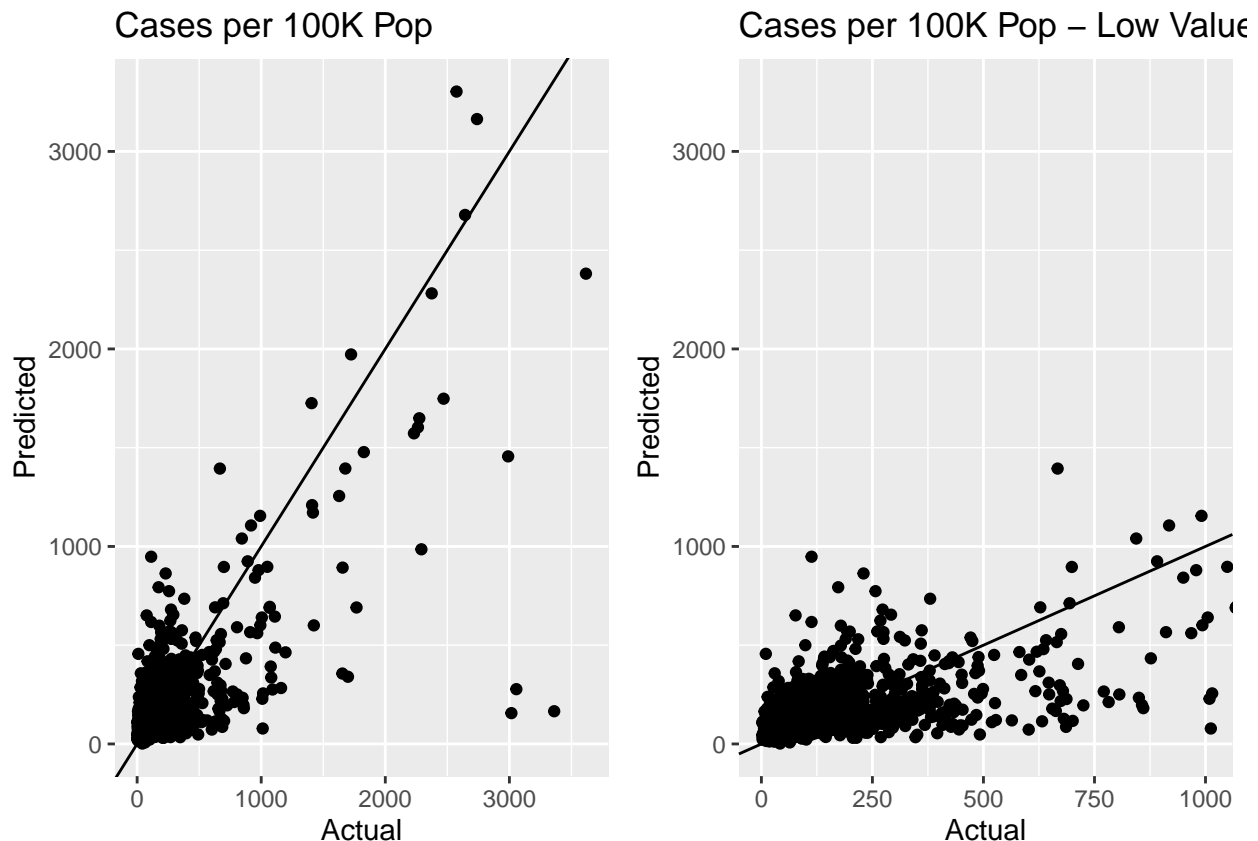
We can clearly see that infection start date is positively correlated with the log count of cases. Along with it, stay at home orders are linked negatively to cases growth which is also theoretically correct. Another important variable is population density per square mile area which means that congested counties see higher number of cases. Counties with higher percentage of educated people and people with race as white also observe lower cases. Higher proportion of people employed in educational and health services and lower proportion of people in arts, recreational, accommodation and public administration industries are correlated with higher cases of virus. The signs for these co-efficient are consistent with our PCA analysis.

We have a problem of small sample and it would have been better if the sample size could have been bigger. We have Hauck-Donner effect in two variables: % of population age 18 to 64 and E_Total_Households_TYPE_Family. The co-efficient magnitude cannot be considered as casual effect. However, the signs of co-efficient points towards some plausible correlation between number of cases and different variables. To check the predictive power of our model, we do 300 test/ train splits of our sample and run the model. We get a RMSE of around 240 cases per hundred thousand people. Plotting the predicted

values against the actual values, we see that the model on average underpredicts the number of cases for counties with high number of cases and overpredicts for counties with lower number of cases. Overall, we do not think this model does a good job at prediction.

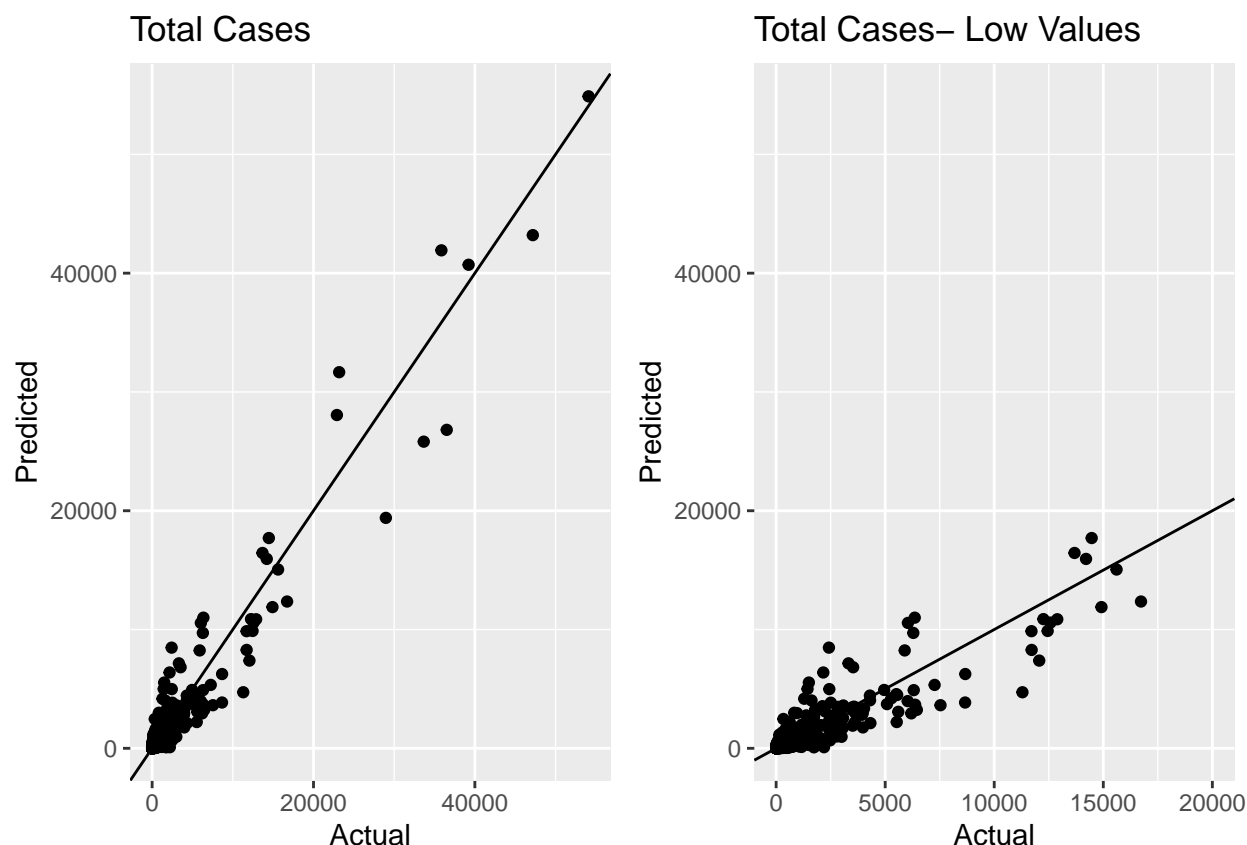
```
## [1] 242.3946
```

```
yhat1 = predict(m12, type = 'response')
modelfit1= as.data.frame(cbind(yhat1, covid$cases1))
b1=ggplot(modelfit1, mapping = aes(V2, V1))+geom_point()+geom_abline(slope = 1,intercept = 0)+ labs(x="Actual", y="Predicted")
b2= ggplot(modelfit1, mapping = aes(V2, V1))+geom_point()+geom_abline(slope = 1,intercept = 0)+ labs(x="Actual", y="Predicted")
grid.arrange(b1,b2,ncol = 2)
```



We try the same model with actual number of cases rather than number of cases per hundred thousand people and include log of population in the model to control for community size. The RMSE of 300 test and train split comes out to be around 1500 and the prediction graph is more balanced as compared to the model with cases per hundred thousand population. Therefore, we prefer using total number of cases in the model for prediction purpose.

```
## [1] 1550.263
```



Random Forest and Boosting Trees

In this section we perform an analysis over the county level data of Covid-19 reports. We know that the tree models are characterized to be flexible fitters that can capture non-linearity and interactions between the variables. Tree models also are useful when we have information in different magnitudes as this case, where we have variables that are values while others are proportions. Specifically, we will work with with a Random Forest model and a Boosting Tree mode. The first, model mainly fit many large trees to bootstrap-resampled versions of the training data by relevance, while the second fits several trees to reweighted versions of the training dataset and then classifies by weighted majority relevance.

Our goal in this section will be make predictions over cases per 100k population. As in the previous section, our data set includes a set of socioeconomic, demographic and geographical variables for many US states. The data set includes information about the Covid crisis such as tests per 100k population and policies issued to fight the pandemic. By the use of this sort of models we can get information about variable relevance, that can shed light about what variables make more likely the presence of the disease in a community and to improve the understanding of this pandemic. To run our models, we first splitted our sample in a training and test subsamples in order to confirm the model performance. After this, We created 1000 trees by random forest including a minimum of 5 features per bucket, and later we compute the out of sample RMSE.

Graph XX summarizes the variable importance information derived from the random forest model (RF). We can notice that as expected the population density per square mile plays an importante role as expected. It suggest that Counties more dense must expect a higher number of people infected, as happened in New York state. The RF model also suggets that the number of test performed in county matters, this goes in the direction of people that support the lower initial of reported cases in the US was due to the lack of testing. The RF model shows that the variable that measures the decrease in number of people at work places explains is relevant, which suggest two things, one that places that started the lockdown earlier could have more cases or the lockdown did stop the virus expansion, since variable importance here does not gives us the impact

sign. From the variable importance chart we can also see covariates related with the population that have had more exposition the the virus, such as occupation. We can see that the proportion of people working at food services in a county explains the number of cases, which make sense since this kind of jobs are likely to have contact with many people, increasing the probability to get sick. 239.036862

Population_density_per_sqrm
 Tests_per_100k
 Total_Pop_Over_15_Divorced
 Total_Pop_15_Households_Householder_Spouse
 Total_Pop_RACE_White
 workplaces_percent_change_from_baseline
 days_first_case
 Occ_Accommodation_Food_Services
 Total_Pop_in_Households_Householder_Parent
 Stay_home_order
 Occ_Other_Services
 summer_rmax
 Total_Pop_SEX_Male
 Median_Household_Income
 Total_Workers_Over_16_Drove_Alone
 Population_age_18to64
 Total_Households_TYPE_Family
 Occ_Healthcare
 Occ_Educational
 summer_rmax
 Occ_Administrative_Waste_Management
 Households_with_incomes
 Total_Workers_Over_16
 Total_poverty
 Adults_Bachelor_degree_or_higher
 Occ_Arts_Entertainment_Recreation
 Unemployment_rate_2018
 Median_Age
 Total_Pop_RACE_One_Race
 Occ_Public_Administration

0.

Graph XX. Variable Importance (Random Forest)

After the RF analysis, we fit a Boosting Tree on cases per 100k.