Predicting the Spread of COVID-19 using R

Brianna MacDonald

Prof. Natallia V. Katenka

STA/DSP 441

Abstract

COVID-19 is a very prevalent topic in today's world, especially as we now near one year of having it in our daily lives. Thanks to scientists and leading researchers (such as John Hopkins University), accessing data about COVID-19 is easier than ever before. A very important question is weighing on many people's minds: when will this pandemic end? This report will try to help answer this question, while many questions are still left in its wake. This report will explain how we can use the number of test cases, as well as the number of positive cases to determine if the spread of COVID-19 will continue over time. The goal of this report is to see if we can predict positive cases based on the current trends of test cases. This will ultimately help to project a possible ending to the pandemic, if these positive trends continue.

Introduction, Problem Formulation, and Literature Review

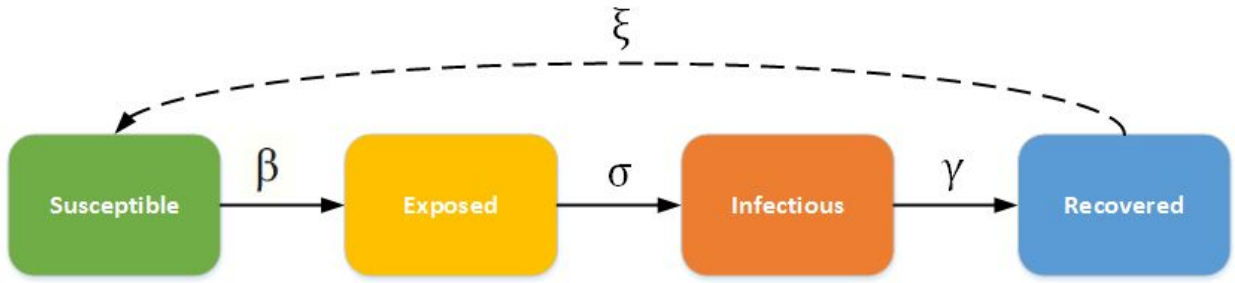
The main question for this project is if the relevant factors above (total number of cases, deaths, and tests done) can be useful when predicting new positive cases of COVID-19. In order to discover the answer to this question, I will perform different types of regression analysis (polynomial, logistic, and linear regression) onto our model and fine-tune the model so it has the best output and the best metrics. To understand this report, I will first give some necessary background information:

COVID-19 and Modeling

There have been many modeling attempts created at both visualizing and modeling COVID-19. These efforts have been done by using Neural Networks, SEIR modeling, and other methods of artificial intelligence and machine learning that we have briefly touched upon throughout class this semester. Throughout my research, I have read countless articles on the different types of modeling that have been researched and used, as well as exploring many different types of metrics and different modeling hyperparameters used in different studies, which I will explore below.

SEIR Modeling

The idea behind SEIR Modeling is that the duration and spread of an epidemic can be modeled by the susceptibility of the epidemic, the number of people exposed from the epidemic, the number of infectious people resulting from that epidemic, and the amount recovered. This is illustrated in the image below.



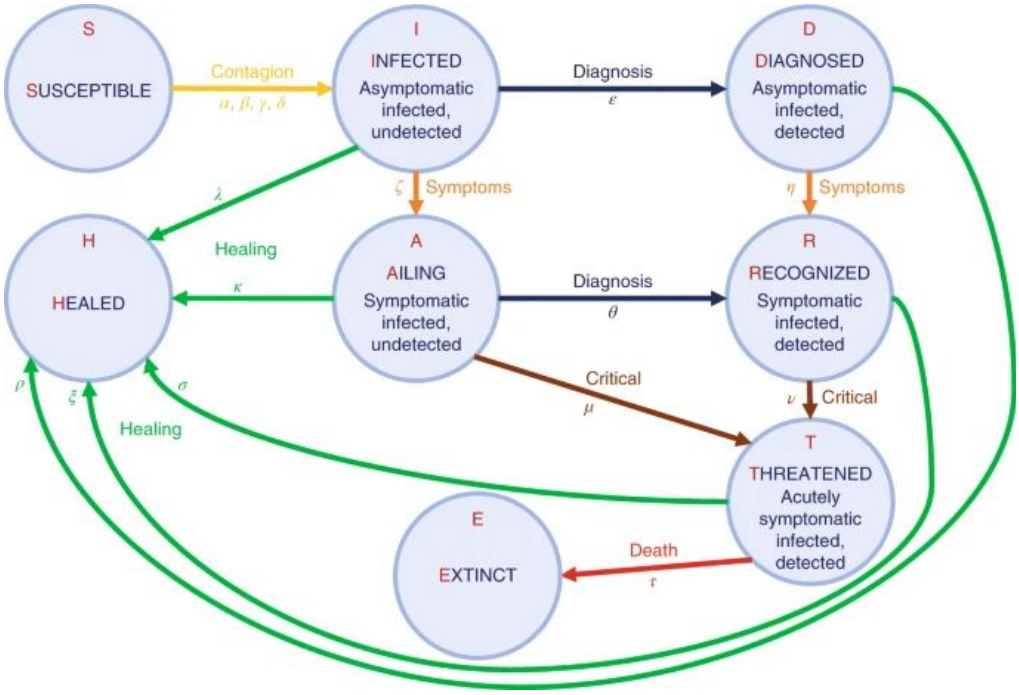
The infectious rate, β, controls the rate of spread which represents the probability of transmitting disease between a susceptible and an infectious individual through different types of transmission (i.e it is not specific to COVID-19, it can be applicable to different types of epidemics depending on rate of transmission and transmission). The incubation rate, σ, is the rate of individuals becoming infectious. Recovery rate, γ = 1/D, is determined by the average duration, D, of infection. This duration can be easily obtained by looking at the different times an individual is symptomatic, or the time in which they become infected, depending on the type of modeling applied. For the SEIR model, ξ is the rate which recovered individuals return to the susceptible state due to loss of immunity. Another variable, infection fatality rate (IFR) is based on all the population that has been infected, i.e., including the undetected individuals and asymptomatic. The github repository located here has a great SHINY app that illustrates the different factors in the SEIR model, with applications to the current COVID-19 epidemic. This research helped me greatly in determining which features I would want my model to be trained on, as well as what features I wanted to use for my predictions.

SIARTHE Modeling

Similar to SEIR modeling, SIARTHE modeling takes into account the following factors:

- S, susceptible (uninfected)
- I, infected (asymptomatic or pauci-symptomatic infected, undetected)
- D, diagnosed (asymptomatic infected, detected)
- A, ailing (symptomatic infected, undetected)
- R, recognized (symptomatic infected, detected)
- T, threatened (infected with life-threatening symptoms, detected)
- H, healed (recovered)
- E, extinct (dead)

These factors and the modeling process, as well as other hyperparameters are highlighted in the following image:



These factors can also be used hand-in-hand with other socially-affected factors, such as quarantine implementations, social distancing practices, and frequency mask use. These variables are very hard to find in an organized format, so for this report I opted to align my modeling more with the SEIR modeling techniques.

Data Characterization, Descriptive Analysis, and Visualization

For this project, I used the following datasets listed below. For each of these datasets, I will provide the source of the data (how/when/by whom data was collected), as well as introducing all the variables that I have used, with the type and units of the variables. For *all* of these datasets, there were little to no outliers present. Because of this ,and because of the importance of all data present in order to make different analysis and predictions, I decided to keep these outliers in.

John Hopkins COVID-19 Data

The following datasets come from John Hopkins public GIT repository, available here. This data is updated almost daily, with the most recent update (as I'm writing this report) 13 hours ago, on December 12th, 2020. Their GitHub lists all aggregated data sources in list format, which is also available at the link listed previously.

Data Names:

- time_series_covid_19_confirmed.csv
- time_series_covid_19_recovered.csv
- time_series_covid_19_deaths.csv

Variables Used/Types:

- Province/State: Character, Province, state or dependency name.
- Country/Region: Character, Country, region or sovereignty name. The names of locations included on the Website correspond with the official designations used by the U.S. Department of State.
- Ts_total: Numeric
- Ts_confirmed: Numeric
- Confirmed: Numeric, Counts include confirmed and probable (where reported).
- Recovered: Numeric, Recovered cases are estimates based on local media reports, and state and local reporting when available, and therefore may be substantially lower than the true number.
- Deaths: Numeric, Counts include confirmed and probable (where reported).
- Date: Character

US State Population Data

This data was used to create a map-style graph indicating different counts of COVID-19 statewide. The daya, along with other Census related data, is available here. This website contains a variety of different values for Census data, including Census data ranging from 2000 to 2019, and mostly recently 2020. This data was used to visualize the spread of COVID-19 using the total number of confirmed cases per capita for each state visualized in a map of the USA. These visualizations will be explored in the next section.

Data Name: US_State_Population.csv

Variables Used/Types:

- Province: Character, State provinces if applicable
- State: Character, State name if applicable, if not it defaults to Province if name is not available

Chicago COVID-19 Data/Illinois COVID-19 Data

This data was collected from Illinois's State Department of Health site, located here. Both of these datasets contain a variety of information, with Chicago containing different variables for that area only, and Illinois containing a more larger-scope view of the spread of COVID-19. These datasets were both fairly large, but I only used a few variables that I determined necessary for my analysis. There had to be some preprocessing done to this data, as I had to remove some commas in the larger numbers, as well as account for some missing variables for specific dates. Because this data was provided by a state/government source, there were not a lot of outliers. This data was used to create multiple different types of regression models. The primary goal of these models were to predict the number of positive test cases based on the number of tests done on that day, as well as predicting the number of deaths for that day (and future days) depending on the number of cases.

Data Names:

- COVID-19_Daily_Testing.csv
- latest_IL.csv
- COVID-19_Daily_cases_Deaths_Hospitalizations (Illinois)

Variables Used/Types:

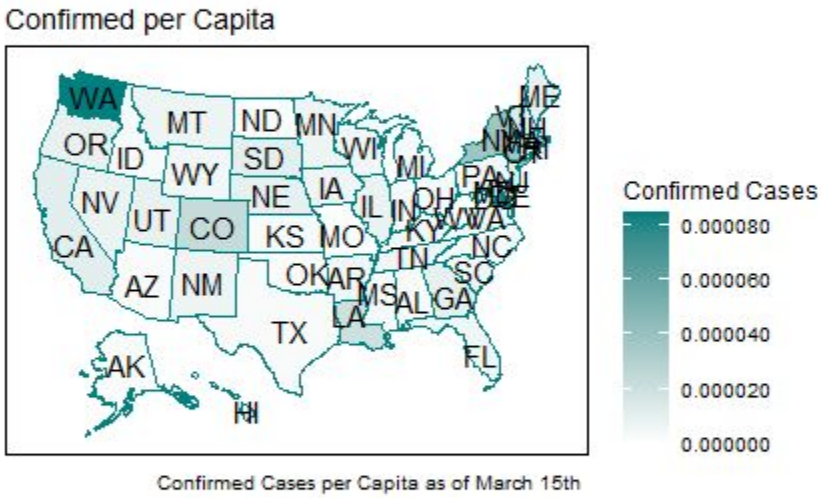
- Date, Character (Date and Time)
- County: Character, County if applicable
- State: Character, State name if applicable, if not it defaults to County if name is not available (because it is for Illinois, all state values are IL).
- Total_cases: Numeric, Number of total cases in the day.
- Total_deaths: Numeric, Number of total deaths in the day.
- Tests: Numeric, Number of tests in the day.
- Cases: Numeric, Number of cases in the day.
- Deaths: Numeric, Number of total deaths
- Hospitalizations: Numeric, Number of total hospitalizations

Brief Visualization - COVID-19 Cases in the US

The following visualizations are created using the data from the US Census regarding population, as well as the number of confirmed cases for that state at specific dates. For these visualizations, I chose the following dates:

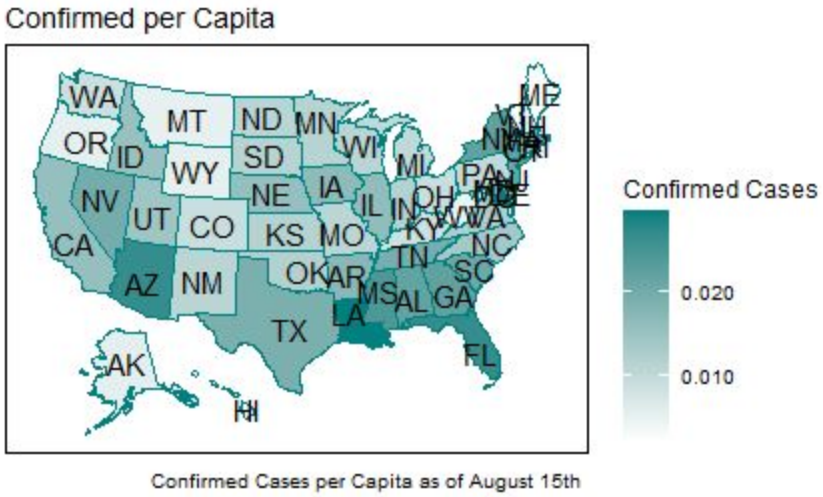
- **March 15th, 2020**: most states had implemented a quarantine or a State of Emergency at this time.
- **August 15th, 2020**: most states had most of their hospitals full to capacity, and had started to implement stricter rules regarding restaurant and recreational facility capacity.
- **November 15th, 2020**: ~6 months since the start of quarantine

March 15th, 2020



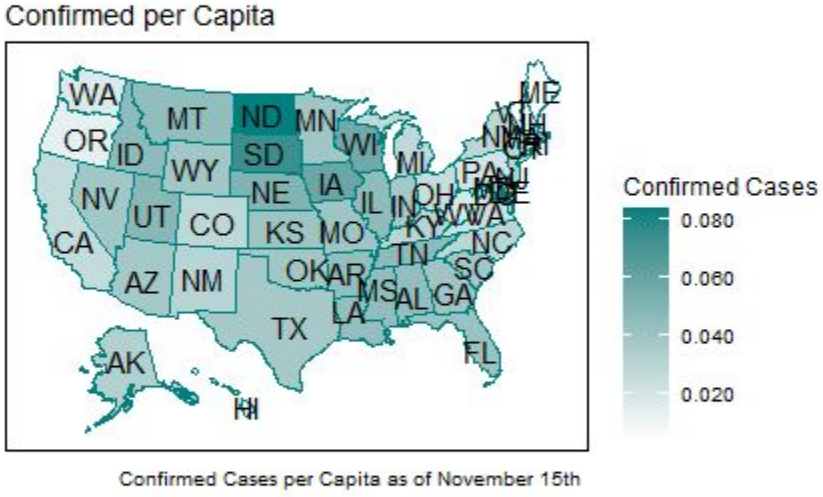
Confirmed Cases per Capita as of March 15th

August 15th, 2020



Confirmed Cases per Capita as of August 15th

November 15th, 2020



Confirmed Cases per Capita as of November 15th

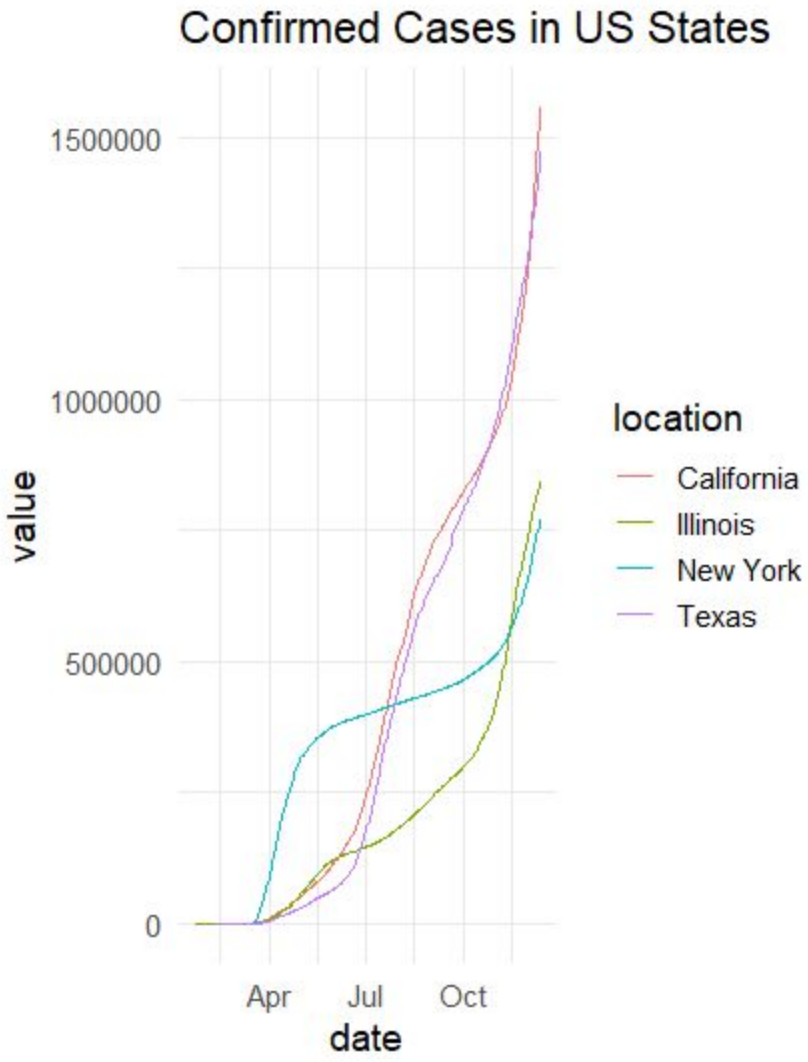
More Visualization - Target States and Confirmed Cases and Deaths

For these visualizations, I decided to look at the following states of interest: New York, Illinois, Texas, and California for the number of confirmed deaths before and after quarantine. The following lists illustrate the different times the above states implemented either a statewide quarantine, or a statewide emergency:

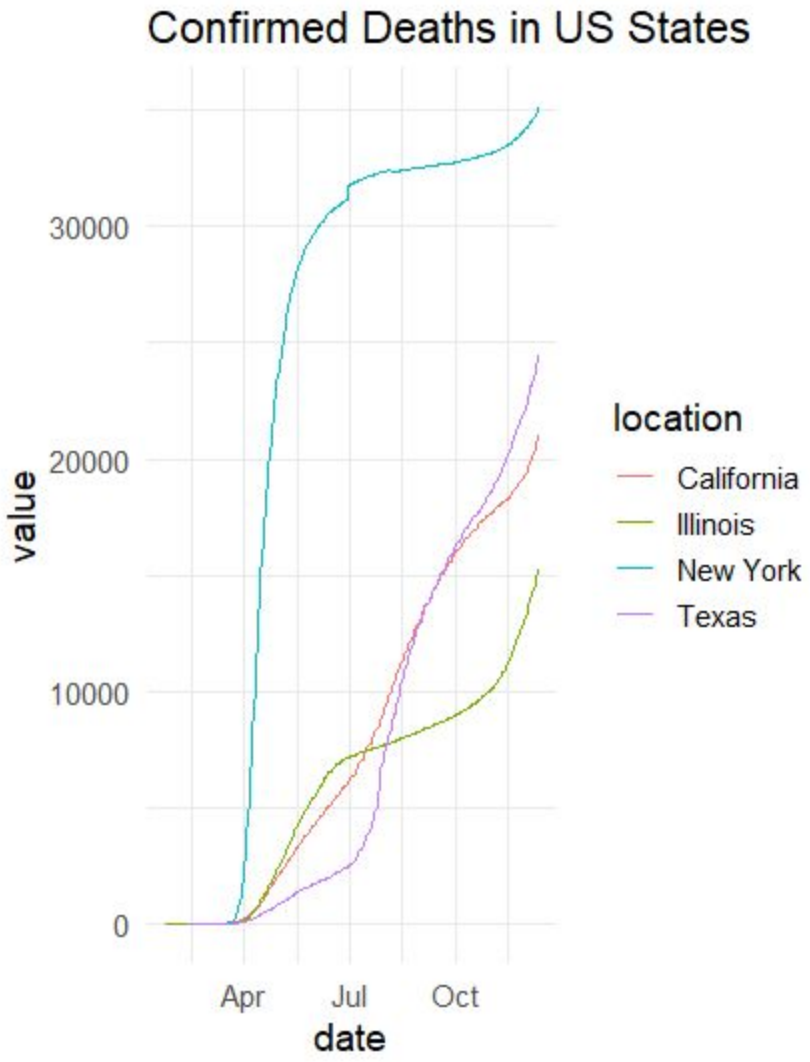
- **New York:**
  Implemented social distancing restrictions and a quarantine on **March 10th, 2020**.
- **Illinois:**
  Implemented social distancing restrictions and a quarantine on **March 21st, 2020**.
- **California:**
  Implemented social distancing restrictions and a quarantine on **March 4th, 2020**.
- **Texas:**
  Implemented social distancing restrictions and a quarantine on **March 13th, 2020**.

The next few pages of this report will have the visualizations for the above.

Confirmed Cases in States of Interest:



Confirmed Deaths in States of Interest:



From these graphs, we can see that New York has the most number of confirmed deaths, which is to be expected as much of New England suffered through the very first rounds of COVID-19 in the United States. California closely follows New York, with a huge spike of Confirmed Cases spanning around July-October, most likely due to summer travel. Texas is very similar to California in this respect, with the number of confirmed cases of the two states mirroring each other directly. Illinois falls directly in the middle of all of the states, which is why I chose it as the perfect predictor for test numbers and the spread of COVID-19. This is because its number can represent the vast majority of other similar states (especially in the middle of the United States). Illinois also has a very diverse population, as well as a variety of different types of towns (urban, suburban, and cities). I believe that the data and predictions made from this data can be very

applicable to other similar states. Because of how diverse Illinois is in terms of location, it can be applicable to nearby states or states of similar composition.

Methods

For this project, I had two main goals. These goals were:

1. To apply and analyze linear regression and polynomial regression on new cases of COVID-19 as well as new tests.
2. To predict and apply regression trees on different predictors of COVID-19 to predict new deaths and non-positive cases (as there are many false positives).

These goals relate to my problem statement because they are directly related to helping predict and analyze the spread and prediction of COVID-19.

Regression is very useful in determining if the spread of COVID-19, as well as the rise of new cases/tests has a linear relationship, or a polynomial relationship. Understanding how cases regress on tests can provide new insight on the two variables, and how they affect each other. Understanding how the different variables in these datasets affect each other is vital in the growing research around COVID-19, and is focal in understanding how we can predict the end of COVID-19 with more research in the future. For my regression model, I wanted to have features on which the case count can have some dependency. We also know that there are many countries where there is inadequate testing done, thus showing less number of cases. Thus, I focused on the datasets where I can get tests and case features to figure out if there is a huge correlation between the two. The Chicago and Illinois datasets both provided excellent data, and were very easy to navigate as they were already cleaned. These datasets were directly available from Illinois's State Department of Health, as listed previously. Regression trees performed very well on predicting new non-positive cases as well as new fatalities for both the Chicago and Illinois datasets. I will go more in-depth about the process of this model as I continue with this report.

Because of the vast amount of data available in the Chicago dataset (such as different ages, genders, and races of people who tested positive, negative, etc). I wanted to perform some sort of analysis on whether or not these predictors could predict if someone was more susceptible to contracting COVID-19. However, the method I applied to this data (RPART Decision Tree), didn't yield especially valuable results. It had a hard time predicting on more than one gender, race, or age. In the future, and as explained in the *Suggestions* portion of this report, I plan to use a different type of model to perform these predictions.

Data Analysis and Main Results

Chicago Dataset (Analysis) - Linear

Variable(s) of Interest: chicago$Cases (numerical), chicago$Tests (numerical)

Explanatory Variable: chicago$Tests

Response Variable: chicago$Cases

Name of Method: Linear Regression

Hypotheses:

A linear regression model will be a good fit for the Chicago dataset when evaluating Tests regressing on Cases. The target (Response) variable in this regression is Cases, with the Explanatory variable being Tests.

Data Conditions:

I needed to preprocess/clean this data so that the numbers were purely numerical. To do this, I ran the following code:

```
chicago$Tests <- gsub(',', '', chicago$Tests)
chicago$Cases <- gsub(',', '', chicago$Cases)
chicago$Tests <- as.numeric(chicago$Tests)
chicago$Cases <- as.numeric(chicago$Cases)
```

This was because a lot of the numbers that were in the thousands had commas in them, and that was presenting some results from the evaluation portion of this analysis to be calculated correctly.

Model Validation/Analysis:

R2: 0.7768795

RMSE: 171.7977

MAE: 138.6275

P-value: 1e-05

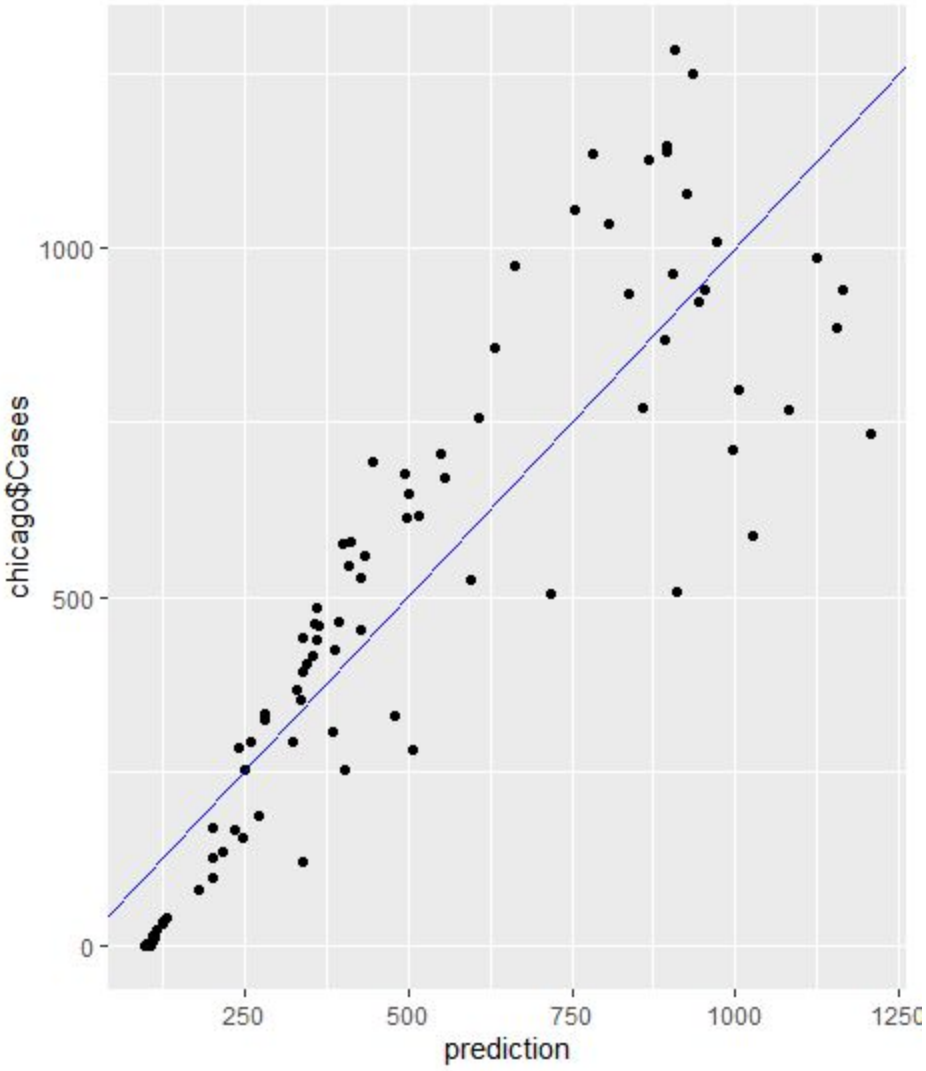
These values turned out much better than expected, considering that there are some outliers that don't follow the linear regression line in the graph at the end of this section. I was very impressed with the RMSE value, as it was lower than I also anticipated. I chose to also explore the MAE (mean absolute error) of this regression because it will allow us to see the difference in terms of how close our predictions are to the actual value.

Tuning Parameters: Not applicable.

Summary:

In summary, while linear regression on the Chicago model using Tests and Cases was *decent*, I wanted to see if there were any other forms of regression that would yield better results. To do this, I will perform polynomial regression on this same data.

Graph of Model: Chicago Linear Regression



Illinois Dataset (Analysis) - Linear

Variable(s) of Interest: illinois$total_cases (numerical), illinois$total_deaths (numerical)

Explanatory Variable: illinois$total_deaths

Response Variable: illinois$total_cases

Name of Method: Linear Regression

Hypotheses:

A linear regression model will be a good fit for the Illinois dataset when evaluating total_deaths regressing on total_cases. The target (Response) variable in this regression is total_cases, with the Explanatory variable being total_deaths. This is different from the previous linear regression model from the Chicago dataset, as now we are regressing cases on deaths, not tests as done previously.

Data Conditions: Data did not need to be cleaned or preprocessed.

Model Validation/Analysis:

R2: 0.89976

RMSE: 4241.519

MAE: 771.3299

P-value: 1e-05

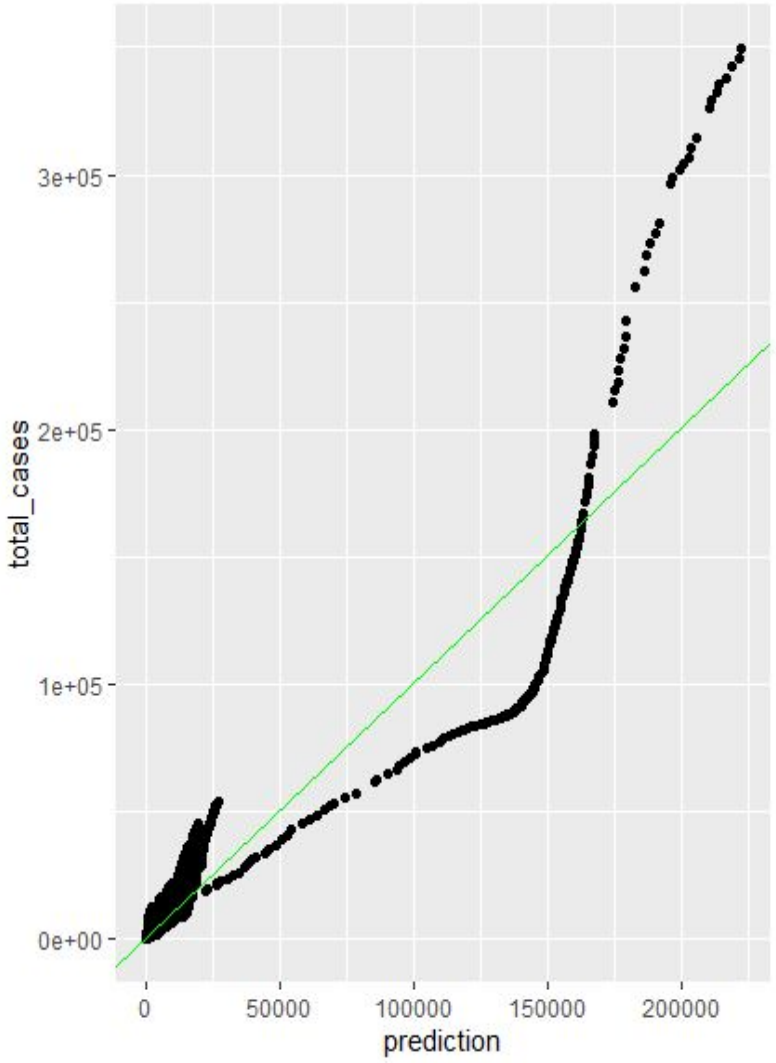
While our R2 score has gone up from the Chicago linear regression, our other values seem to have gone up. The reason is that this model is much more polynomial than the Chicago model, so I will use polynomial regression to develop a new model with the same predictors to see if it yields better results

Tuning Parameters: Not applicable.

Summary:

In summary, while linear regression on the Illinois model using total_deaths and total_cases did provide a better R2 score than it had performed earlier on the Chicago dataset, I still believe that a polynomial regression model will suit this data better.

Graph of Model: Illinois Linear Regression



Chicago Dataset (Analysis) - Polynomial

Variable(s) of Interest: chicago$Cases (numerical), chicago$Tests (numerical)

Explanatory Variable: chicago$Tests

Response Variable: chicago$Cases

Name of Method: Polynomial Regression

Hypotheses:

After reviewing the results of the linear regression model, I believe that a polynomial regression model will be a good fit for the Chicago dataset when evaluating Tests regressing on Cases. The target (Response) variable in this regression is Cases, with the Explanatory variable being Tests.

Data Conditions: Already cleaned from the first regression analysis

Model Validation/Analysis:

R2: 0.8701158

RMSE: 131.0768

MAE: 86.75418

P-value: 1e-05

These values are all around much better than the values achieved using linear regression on the same data with the same regressors. It is safe to say that our hypothesis was correct, and that a polynomial regression model was a better fit for the Chicago dataset than a linear regression model.

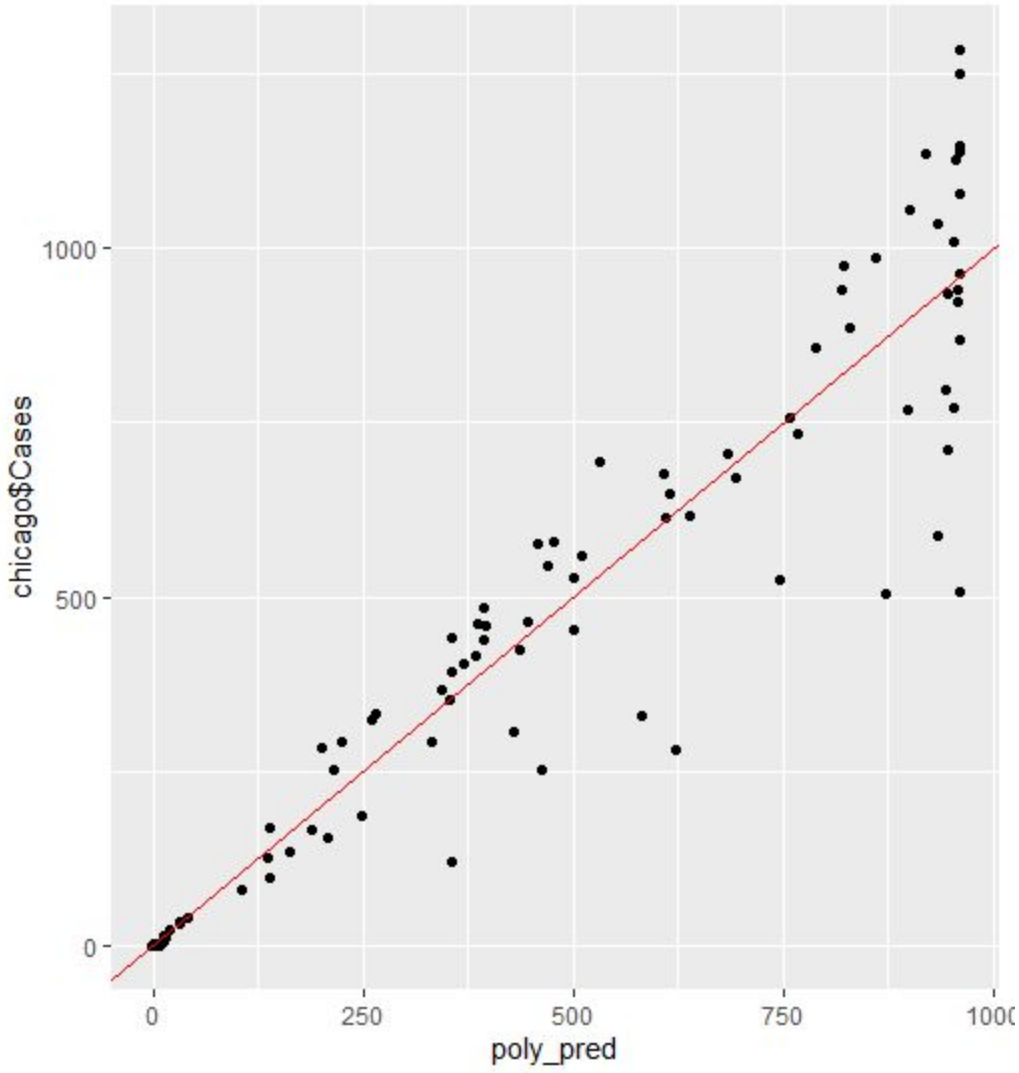
Tuning Parameters: Not applicable.

Summary:

In summary, our polynomial regression model for the Chicago dataset performed much better than our previous linear regression model. This is seen in the following model evaluation below:

```
> # First model
> RMSE(chicago$Cases, prediction)
[1] 171.7977
> R2(chicago$Cases, prediction)
[1] 0.7768795
> mae(chicago$Cases, prediction)
[1] 138.6275
> # Second model
> RMSE(chicago$Cases, poly_pred)
[1] 131.0768
> R2(chicago$Cases, poly_pred)
[1] 0.8701158
> mae(chicago$Cases, poly_pred)
[1] 86.75418
> wrapFTest(model_4)
[1] "F Test summary: (R2=0.8701, F(4,84)=140.7, p<1e-05)."
```

Graph of Model: Chicago Polynomial Regression



Illinois Dataset (Analysis) - Polynomial

Variable(s) of Interest: illinois$total_cases (numerical), illinois$total_deaths (numerical)

Explanatory Variable: illinois$total_deaths

Response Variable: illinois$total_cases

Name of Method: Polynomial Regression

Hypotheses:

As seen from the results previously mentioned regarding how the Illinois dataset performed with a linear regression model, I concluded that a polynomial model would be a much better fit. This is especially true considering how much the Chicago model improved after applying a polynomial regression model

Data Conditions: Data did not need to be cleaned or preprocessed.

Model Validation/Analysis:

R2: 0.9804

RMSE: 1875.233

MAE: 522.2051

P-value: 2.2e-16

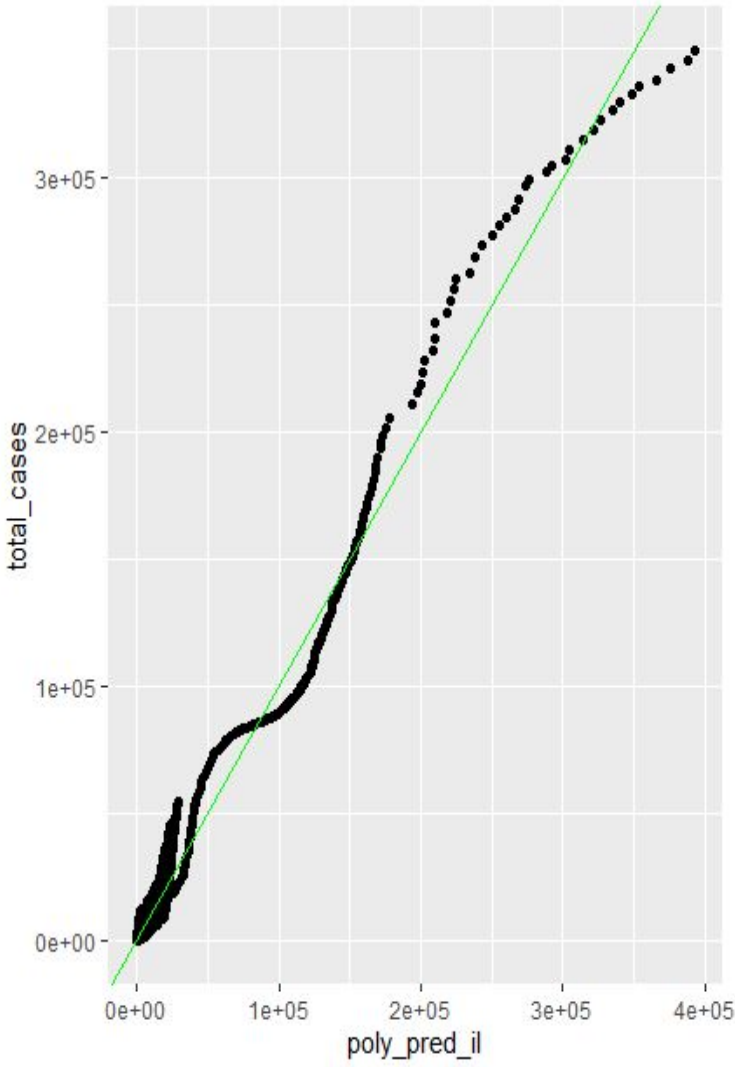
Tuning Parameters: Not applicable.

Summary:

Our R2 score has gone up by about ~0.1 in comparison to our linear regression model, with our RMSE going down by almost ~3000. Additionally, our MAe has gone down by about ~200. This model is much more suited to our dataset than the linear regression model, and this is further proven with the following data below.

```
> # Evaluating model
> # Linear model
> RMSE(total_cases, prediction_il)
[1] 4241.519
> R2(total_cases, prediction_il)
[1] 0.8997683
> mae(total_cases, prediction_il)
[1] 771.3299
> # Polynomial Model
> RMSE(total_cases, poly_pred_il)
[1] 1875.233
> R2(total_cases, poly_pred_il)
[1] 0.9804083
> mae(total_cases, poly_pred_il)
[1] 522.2051
```

Graph of Model: Illinois Polynomial Regression



Illinois Hospitalizations and Deaths/Cases Dataset (Modeling)
This is a model that can predict fatalities (Deaths...Total) based on the total number of cases (Cases...Total) and the total number of hospitalizations (Hospitalizations...Total).
Variable(s) of Interest: Deaths...Total, Cases...Total, Hospitalizations...Total (all numerical)
Explanatory Variable: Cases...Total, Hospitalizations...Total
Response Variable: Deaths...Total
Name of Method: RPART CART Model
Hypotheses:
I believe that an accurate model for predicting fatalities can be made predicting with the total number of cases (Cases...Total) and the total number of hospitalizations (Hospitalizations...Total).
Data Conditions: Data is already preprocessed and cleaned.
Model Validation/Analysis:
R2: 0.932791
RMSE: 43.99774
MAE: 11.44228
P-value: N/A
Tuning Parameters:

- Training size of 65%, testing of 35%
- Number of folds for cross-validation: 5
- CP value of 0

Summary:
This model, as assumed, made very accurate predictions using the number of total cases and hospitalizations to predict new fatalities. I'm glad that I chose to add total hospitalizations as a predictor, because before I was only predicting using cases (which led to a very low R2 score). Overall in this case we can accept our hypothesis. Overall, this model had an exceptionally high R2 score, with a low RMSE and MAE score, which indicates a good fit and not too much variation between random data and selected data.

No visualization, code is available in the Appendix section.

Illinois Dataset (Modeling)

This is a model that can predict non-positive cases(People.Not.Positive...Total) based on the total number of Tests done (Tests). The goal of this model is to illustrate the need for further testing so that the false-positive test is not high.

Variable(s) of Interest: People.Not.Positive...Total, Tests (all numerical)

Explanatory Variable: Cases...Total, Hospitalizations...Total

Response Variable: People.Not.Positive...Total

Name of Method: RPART CART Model

Hypotheses:

I believe that an accurate model for predicting non-positive cases can be made by predicting with the total number of tests.

Data Conditions:

Had to make variables strictly numeric, as there was an error that occurred.

```
il_1$People.Not.Positive...Total <-
as.numeric(il_1$People.Not.Positive...Tota
l)
il_1$Tests <- as.numeric(il_1$Tests)
```

Model Validation/Analysis:

R2: 0.880542

RMSE: 184.071

MAE: 170.4197

P-value: N/A

Tuning Parameters:

- Training size of 80%, testing of 20%
- Number of folds for cross-validation: 5
- CP value of 0

Summary:

This model did fairly well predicting non-positive cases, as I assumed it would. It would be interesting if I were to shuffle the data prior to loading it into the model, but I will explore that in another project. Our R2 score is fairly high, with a low MAE and RMSE, indicating a good fit for the model

No visualization, code is available in the Appendix section.

Chicago Dataset (Modeling)

This is a model that can predict if a woman, aged 40-49, will be tested positive. This was a very exploratory model, and I wish I could've also explored different age groups (teens, young adults, and elderly people) and see if that also provided good results.

Variable(s) of Interest: Cases, People.Tested...Age.40.49, People.Tested...Female (all numerical)

Explanatory Variable: People.Tested...Age.40.49, People.Tested...Female

Response Variable: Cases

Name of Method: Regression Tree

Hypotheses:

I believe that there is definitely a relationship between age (not specifically gender, but this was used in the model to provide another interesting predictor) in whether or not that person has a positive COVID-19 case.

Data Conditions: Data is already preprocessed and cleaned.

Model Validation/Analysis:

R2: 0.85746

RMSE: 132.91554

MAE: 123.92989

P-value: N/A

Tuning Parameters:

- Training size of 65%, testing of 35%
- Number of folds for cross-validation: 5
- CP value of 0

Summary:

While my hypothesis predicted that this model would be accurate (which it was), I did not expect to have the predictors of this model be so important and imperative in predicting whether a case is positive or not. I will definitely continue research into other predictors (race, gender, age) in future projects. Our R2 score is 0.85746, which is exceptionally good considering our model. Our RMSE and MAE also fall within our parameters for our guidelines, so overall this model proves our hypothesis.

No visualization, code is available in the Appendix section.

Conclusions and Research Directions

This project has given me so much insight about the different factors that affect the spread of COVID-19, along with the different predictors that go into account for predicting new deaths of COVID-19, as well as positive and non-positive cases. To conclude, the polynomial model performed much better than the linear regression model in both the Illinois and Chicago datasets, as described in the previous section.There is a fairly polynomial relationship with number of cases and number of tests predicted, as seen in the evaluation of our models in the previous section. We also saw how new fatalities could be predicted using the total number of positive cases and the total number of hospitalizations. It would be even more interesting to predict new deaths on a day-to-day basis. This prediction could help to highlight spikes in the near future so hospitals would know when they would reach capacity. It was also really interesting to see how variables such as age and gender could predict if someone were to contract a positive case of COVID-19. In the future, as I briefly explained before, I plan on looking at the different correlations between age, gender, and race with if they are more likely than not to have a positive COVID-19 case. In the future, I also want to add more factors to calculating/predicting fatality, such as symptoms and pre-existing conditions. Another interesting approach would be to add predictors of mask use, social distancing guidelines, and quarantines to help predict the spread of COVID-19 in states. Additionally, I would like to examine the effect of COVID-19 on different markets, such as the entertainment industry

Using our predictors for our models in this paper, we can ultimately accept our hypothesis that factors such as total cases, deaths, tests, and hospitalizations can accurately model different aspects of COVID-19.

Appendix - R Code

First Linear Regression Model: Chicago

```
## Predicting new cases: CHICAGO
```{r predicting_covid1}
set.seed(143)
chicago <- read.csv("COVID-19_Daily_Testing.csv")
head(chicago, n=10)
chicago$Tests <- gsub(',', '', chicago$Tests)
chicago$Cases <- gsub(',', '', chicago$Cases)
chicago$Tests <- as.numeric(chicago$Tests)
chicago$Cases <- as.numeric(chicago$Cases)
print(chicago$Cases)

# Splitting the data into test and train, 80%
split/20
row.number <- sample(1:nrow(chicago),
0.8*nrow(chicago))
train = chicago[row.number,]
test = chicago[-row.number,]
dim(train)
dim(test)
# Regression
lmchicago <- lm(chicago$Cases~chicago$Tests, data =
train)
wrapFTest(lmchicago)
```
```

Making Predictions for the above Model

```
## Predicting values for our Regression Model
```{r pred_lr}
prediction <- predict(lmchicago)
ggplot(lmchicago, aes(x = prediction, y =
chicago$Cases)) +
  geom_point() +
  geom_abline(color = "blue")
```
```

Logistic Regression Model for Chicago (Not Used)

```
## Logistic Linear Regression
```{r log}
glm.fit <- glm(chicago$Cases~chicago$Tests, data =
train, family = binomial)
summary(glm.fit)
glm.probs <- predict(glm.fit)
glm.probs[1:5]
```
```

Polynomial Regression Model for Chicago

```
## Polynomial Regression
```{r poly}
plot(chicago$Cases, chicago$Tests)

model_1 <- lm(chicago$Cases ~ poly(chicago$Tests,
1))

# Summary statistics for model_1
summary(model_1)

model_2 <- lm(chicago$Cases ~ poly(chicago$Tests,
2))

# Summary statistics for model_2
summary(model_2)

model_3 <- lm(chicago$Cases ~ poly(chicago$Tests,
3))

# Summary statistics for model_3
summary(model_3)

model_4 <- lm(chicago$Cases ~ poly(chicago$Tests,
4))

# Summary statistics for model_3, R2 of 87.01,
2.2e-16 p-value
summary(model_4)
# Predictions for this model
poly_pred <- predict(model_4)
```
```

Plotting Polynomial Model, Evaluating Both Models, and Comparing

```
## Examining best model *model_4*
```{r model_4}
ggplot(model_4, aes(x = poly_pred, y =
chicago$Cases)) +
  geom_point() +
  geom_abline(color = "red")
```
```

```
## Evaluating Model
```{r eval}
# First model
RMSE(chicago$Cases, prediction)
R2(chicago$Cases, prediction)
mae(chicago$Cases, prediction)
# Second model
RMSE(chicago$Cases, poly_pred)
R2(chicago$Cases, poly_pred)
mae(chicago$Cases, poly_pred)
wrapFTest(model_4)
```
```

```
## Comparing models
```{r compare}
anova(lmchicago, model_4)
```
```

Linear Regression Model for Illinois

```
## Exploring more about Illinois
```{r illinois}
library(ggplot2)
illinois <- read.csv("latest_IL.csv")
head(illinois, n=10)
total_cases <- illinois$total_cases
total_deaths <- illinois$total_deaths
dates <- illinois$date
ggplot(aes(x = dates, y = total_cases), data =
illinois) +
  geom_line() +
  theme(axis.text.x = element_text(angle = 45, vjust
= 1, hjust=1)) +
  scale_y_continuous() +
  ggtitle("Confirmed Deaths in Illinois")
```
Linear Regresssion for Future Illinois Cases
```{r illinois_reg}
set.seed(153)

# Regression
lmillinois <- lm(total_cases~total_deaths, data =
illinois)
wrapFTest(lmillinois)
```
```

Predicting Values/Plotting for Illinois & Evaluation + Polynomial Regression

```
## Predicting values for our Regression Model
```{r pred_lr}
prediction_il <- predict(lmillinois)
ggplot(lmillinois, aes(x = prediction_il, y =
total_cases)) +
  geom_point() +
  geom_abline(color = "green")
```
Evaluating Model
```{r eval}
# Evaluating model
RMSE(total_cases, prediction_il)
R2(total_cases, prediction_il)
mae(total_cases, prediction_il)
```
Polynomial Regression Illinois
```{r poly_il}
model_5 <- lm(total_cases ~ poly(total_deaths, 4))
summary(model_5)
# Predictions for this model
poly_pred_il <- predict(model_5)
```
```

Plotting & Evaluating Polynomial Regression for Illinois

```
## Plotting poly illinois
```{r plot_poly_il}
ggplot(model_5, aes(x = poly_pred_il, y =
total_cases)) +
  geom_point() +
  geom_abline(color = "green")
```
Evaluating poly illinois model
```{r eval}
# Evaluating model
RMSE(total_cases, poly_pred_il)
R2(total_cases, poly_pred_il)
mae(total_cases, poly_pred_il)
```
```

CART Model for Illinois for Non-Positive Cases

```
## Training/Predicting with Regression Trees
```{r regr_trees}
# Splitting the data into test and train, 80%
split/20
row.number <- sample(1:nrow(illinois),
0.8*nrow(illinois))
train = illinois[row.number,]
test = illinois[-row.number,]
summary(train)
```
Building model for predicting confirmed
cases
```{r building_model_cases}
num_folds <- trainControl(method = "cv", number
= 5) # Specify 5-fold cross-validation.
parameter_grid <- expand.grid(.cp = 0)
# Predicting Confirmed Cases
cases_model <- train(
  total_cases ~ total_deaths,
  data = train,
  method = "rpart", # CART algorithm
  trControl = num_folds,
  tuneGrid = parameter_grid
)

print(cases_model)
```
```

CART Model for Illinois for # of Fatalities

```
## Building model for predicting fatalities
```{r building_model_deaths}
# Predicting new deaths
il_2 <- read.csv("COVID-19_Daily_Cases__Deaths_
_and_Hospitalizations.csv")
row.number <- sample(1:nrow(il_2),
0.65*nrow(il_2))
train_il = il_2[row.number,]
test_il = il_2[-row.number,]
# Making model
death_model <- train(
  Deaths...Total ~ Cases...Total +
Hospitalizations...Total,
  data = train,
  method = "rpart", # CART algorithm
  trControl = num_folds,
  tuneGrid = parameter_grid,
  na.action=na.exclude
)

print(death_model)
```
```

Model for Predicting Positive Case Based on Specific Age and Gender

```
## Predicting Positive Cases in Females Aged
40-49 using Regression Trees
```{r 18_24}
# Splitting the data into test and train, 65%
split/35
summary(chicago)
split <- initial_split(chicago, prop = .65)
train = chicago[split,]
test = chicago[-split,]
summary(train)
# Predicting new deaths
cases_18 <- train(
  Cases ~ People.Tested...Age.40.49 +
  People.Tested...Female,
  data = train,
  method = "rpart", # CART algorithm
  trControl = num_folds,
  tuneGrid = parameter_grid
)
print(cases_18)

```
```

Code for Visualization of US Maps

```
## US Cases by State, November
```{r us_cases_by_state_nov}
# November 15th
us_map <- us %>% left_join(pop, by = "Province") %>%
    filter(Date == anydate("11/15/2020"))
# Mapping
us_map <- us_map %>%
    mutate(confirmed_per_capita = Confirmed/pop2019)

plot_usmap(data = us_map, values = "Confirmed",
color = teal, labels = TRUE) +
    scale_fill_continuous(low = "white", high =
teal, name = "Confirmed Cases", label =
scales::comma) +
    theme(legend.position = "right") +
    theme(panel.background = element_rect(colour =
"black")) +
    labs(title = "Confirmed", caption = "Confirmed
Cases as of 15 November ")

plot_usmap(data = us_map, values =
"confirmed_per_capita", color = teal, labels = TRUE)
+
    scale_fill_continuous(low = "white", high =
teal, name = "Confirmed Cases", label =
scales::comma) +
    theme(legend.position = "right") +
    theme(panel.background = element_rect(colour =
"black")) +
    labs(title = "Confirmed per Capita", caption =
"Confirmed Cases per Capita as of November 15th ")
```
```

Code for Correlation Plot

```
## Correlations

```{r case_corr}
# Mortality rate and recovery rate
cases_total_date <- cases_total_date %>%
  group_by(Date, Confirmed) %>%
  mutate(Mortality_rate = Deaths / Confirmed,
         Recovery_rate = Recovered / Confirmed) %>%
  ungroup()

# Correlations
cases_total_date %>%
  select(-Date) %>%
  na.omit() %>%
  cor(use = "pairwise.complete.obs") %>%
  corrplot.mixed(tl.col = "black", tl.pos = "d",
tl.cex = 0.7, cl.cex = 0.7,
                 number.cex = 0.7)

```
```

Sample Data: Chicago

Date	Day	Tests	Cases	People Ne	People Te	People Te	People Te	People Te	People Te	People Te	People Te	People Te	People Te	People Te	People Te	People Te	People Te	People Te	People Te
3/1/2020	Sunday	1	0	1	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0
5/2/2020	Saturday	2,431	705	1,726	129	470	458	458	412	281	128	94	1	1,238	1,158	35	569	39	376
########	Thursday	4,098	772	3,326	260	805	833	685	604	471	253	171	16	2,049	1,786	263	848	55	581
3/5/2020	Thursday	17	1	16	4	2	0	4	3	2	0	2	0	8	9	0	2	0	5
3/6/2020	Friday	18	3	15	1	5	1	3	3	2	2	0	1	8	10	0	2	0	1
3/7/2020	Saturday	13	3	10	2	1	2	5	1	0	2	0	0	11	2	0	1	0	4
5/4/2020	Monday	4,141	1,127	3,014	246	790	858	716	674	487	229	135	6	2,116	1,839	186	924	63	634
########	Wednesd	4,283	870	3,413	279	864	840	729	686	449	273	154	9	2,217	1,963	103	967	64	602

Sample Data: Illinois

3/20/2020	17001	Adams	IL	1	0	0.142857	0	11.5	7.66667	17.25	2.51809	1.67873	3.77713	19.8068	13.2046	29.7103	0.000303	0.000202	0.000454
3/21/2020	17001	Adams	IL	1	0	0.142857	0	11	7.33333	16.5	3.65415	2.4361	5.48122	21.9253	14.6169	32.8879	0.000335	0.000223	0.000503
3/22/2020	17001	Adams	IL	1	0	0.142857	0	10.5	7	15.75	3.2702	2.18013	4.90529	23.6955	15.797	35.5432	0.000362	0.000241	0.000543
3/23/2020	17001	Adams	IL	1	0	0.142857	0	10	6.66667	15	6.84775	4.56517	10.2716	29.0789	19.386	43.6184	0.000444	0.000296	0.000667
3/24/2020	17001	Adams	IL	1	0	0.142857	0	9.5	6.33333	14.25	10.6969	7.13124	16.0453	38.3274	25.5516	57.4911	0.000586	0.00039	0.000879

Sample Data: US

UID	iso2	iso3	code3	FIPS	Admin2	Province_	Country_F	Lat	Long_	Combinec	########	########	########	########	########	########	########	########
84001001	US	USA	840	1001	Autauga	Alabama	US	32.53953	-86.6441	Autauga, A	0	0	0	0	0	0	0	0
84001003	US	USA	840	1003	Baldwin	Alabama	US	30.72775	-87.7221	Baldwin, A	0	0	0	0	0	0	0	0
84001005	US	USA	840	1005	Barbour	Alabama	US	31.86826	-85.3871	Barbour, A	0	0	0	0	0	0	0	0
84001007	US	USA	840	1007	Bibb	Alabama	US	32.99642	-87.1251	Bibb, Alab	0	0	0	0	0	0	0	0
84001009	US	USA	840	1009	Blount	Alabama	US	33.98211	-86.5679	Blount, Al	0	0	0	0	0	0	0	0
84001011	US	USA	840	1011	Bullock	Alabama	US	32.10031	-85.7127	Bullock, A	0	0	0	0	0	0	0	0
84001013	US	USA	840	1013	Butler	Alabama	US	31.753	-86.6806	Butler, Ala	0	0	0	0	0	0	0	0
84001015	US	USA	840	1015	Calhoun	Alabama	US	33.77484	-85.8263	Calhoun, A	0	0	0	0	0	0	0	0
84001017	US	USA	840	1017	Chambers	Alabama	US	32.9136	-85.3907	Chambers	0	0	0	0	0	0	0	0
84001019	US	USA	840	1019	Cherokee	Alabama	US	34.17806	-85.6064	Cherokee	0	0	0	0	0	0	0	0
84001021	US	USA	840	1021	Chilton	Alabama	US	32.85044	-86.7173	Chilton, A	0	0	0	0	0	0	0	0

Sample Data: Illinois Hospitalizations

Date	Cases - To	Deaths - T	Hospitaliz	Cases - Ag	Cases - Ag	Cases - Ag	Cases - Ag	Cases - Ag	Cases - Ag	Cases - Ag	Cases - Ag	Cases - Ag
3/29/2020	282	20	130	4	29	48	54	50	50	30	17	0
3/1/2020	0	0	2	0	0	0	0	0	0	0	0	0
3/2/2020	0	0	2	0	0	0	0	0	0	0	0	0
3/3/2020	0	0	3	0	0	0	0	0	0	0	0	0
3/4/2020	0	0	4	0	0	0	0	0	0	0	0	0
3/5/2020	1	0	6	0	0	0	0	1	0	0	0	0
5/9/2020	521	52	100	34	104	100	98	90	54	28	13	0
5/23/2020	329	30	90	31	64	58	54	54	36	18	14	0