Startups Funding

Michael Audie, Karen Thornton, Hannah Schofield

2023-04-28

```
#load packages
library(lubridate)
library(ggplot2)
library(forecast)
library(tseries)
library(smooth)
library(dplyr)
library(kableExtra)
```

```
## Error: package or namespace load failed for 'kableExtra':
##  .onLoad failed in loadNamespace() for 'kableExtra', details:
##   call: !is.null(rmarkdown::metadata$output) && rmarkdown::metadata$output %in%
##   error: 'length = 2' in coercion to 'logical(1)'
```

Background +Clean Tech startups are on the rise +There are two stages of funding stages: early stage and late stage. Early stage can be considered seed funding and late stage can be something like series C funding. Late stage funding is often more money for fewer companies +Companies that make it to late stage funding have a higher chance of being sucessfull. +Greater late stage funding investment suggests a more maturation of the Clean Tech industry.

Research Questions 1. Are investments in clean tech startups expected to continue to increase?

2. As the clean tech industry matures will investments in late-stage startups outpace investments in earlier-stage startups?

Data The data was collected from the IEA. The data is broken up into early and late stage investments. The data original went to 2022, but with Covid, the War in Ukraine, and a recession, there were dips in investment in both stages. Since we believe this doesn't reflect the trend of the industry, we decided to remove the year 2022.

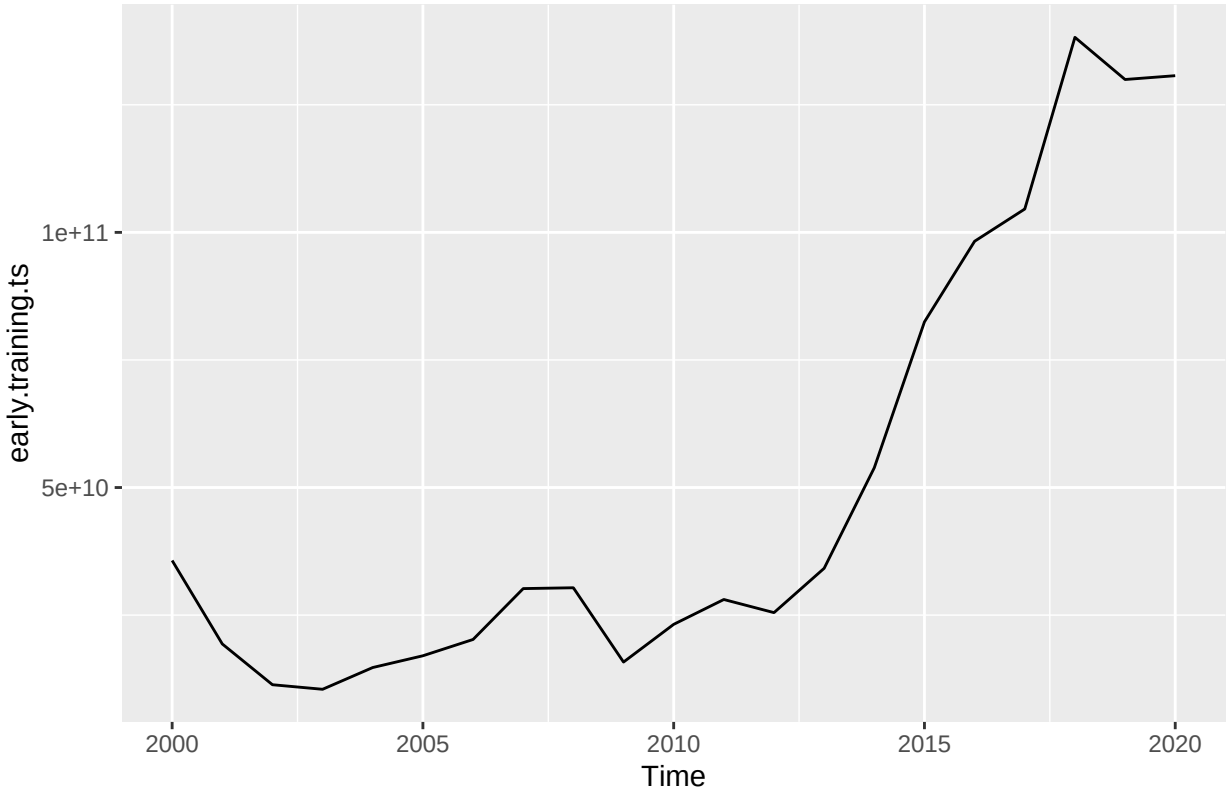
Limitations The dataset had some limitations that made it harder to predict future years. These limitations include, but are not limited to: 1. The data was yearly, as opposed to daily or monthly. This made it difficult to show seasonality. It also made the data less granular and will less data points it was harder to see a trend to model. 2. As mentioned above, covid, and other factors like the war in Ukraine, and overall recession could be factors in the model that cannot be predicted or anticipated.

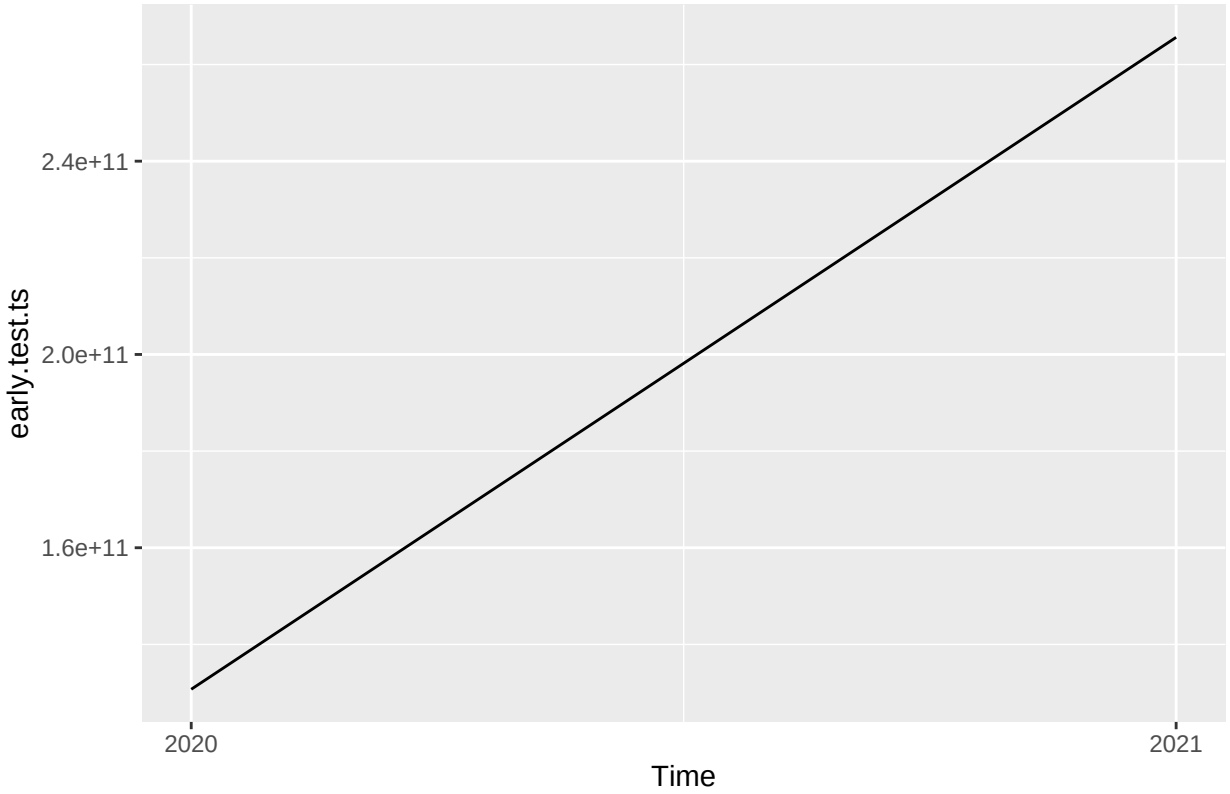
Data Processing Steps #Data Wrangling - Full sets

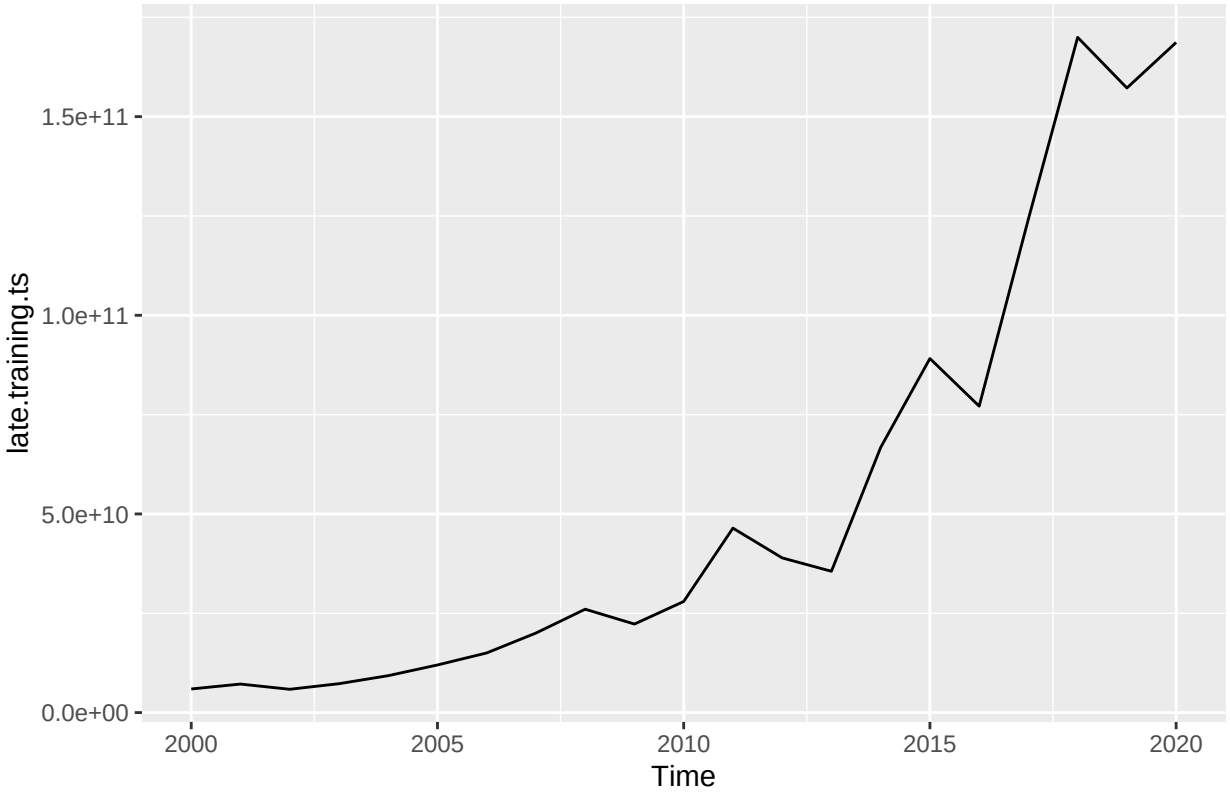
#Data Wrangling - Training Sets

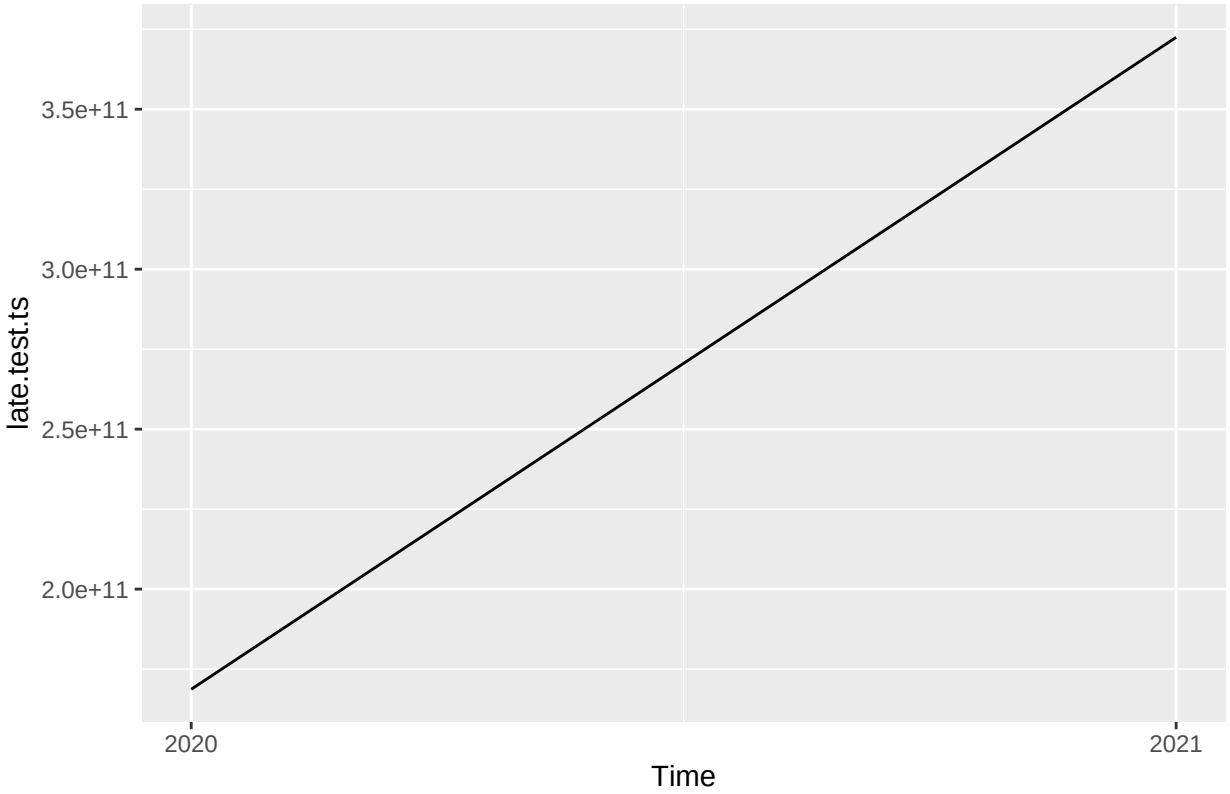
#Creating Time Series

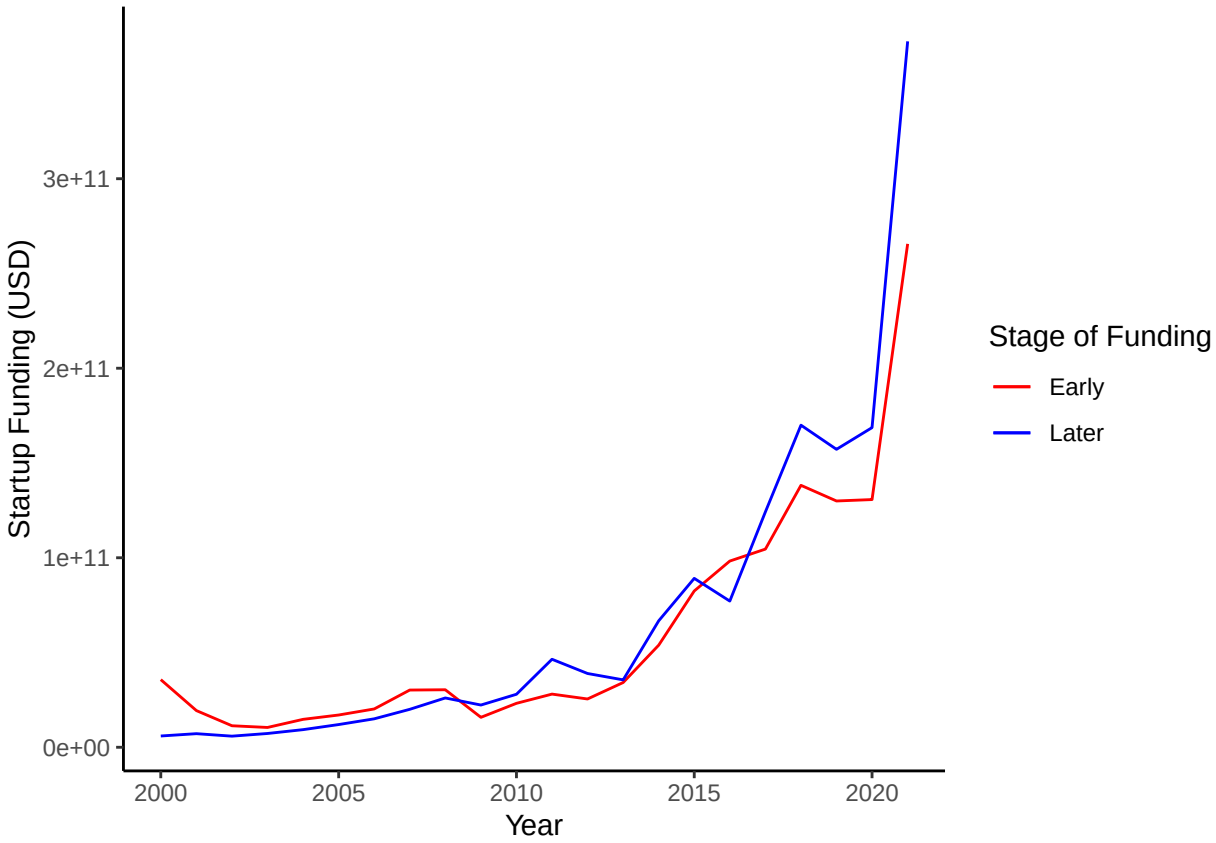
#Data Wrangling - Testing Sets


early.training.ts
1e+11
5e+10
2000
2005
2010
2015
2020
Time


2.4e+11
2.0e+11
1.6e+11
early.test.ts
2020
2021
Time


late.training.ts
1.5e+11
1.0e+11
5.0e+10
0.0e+00
2000
2005
2010
2015
2020
Time

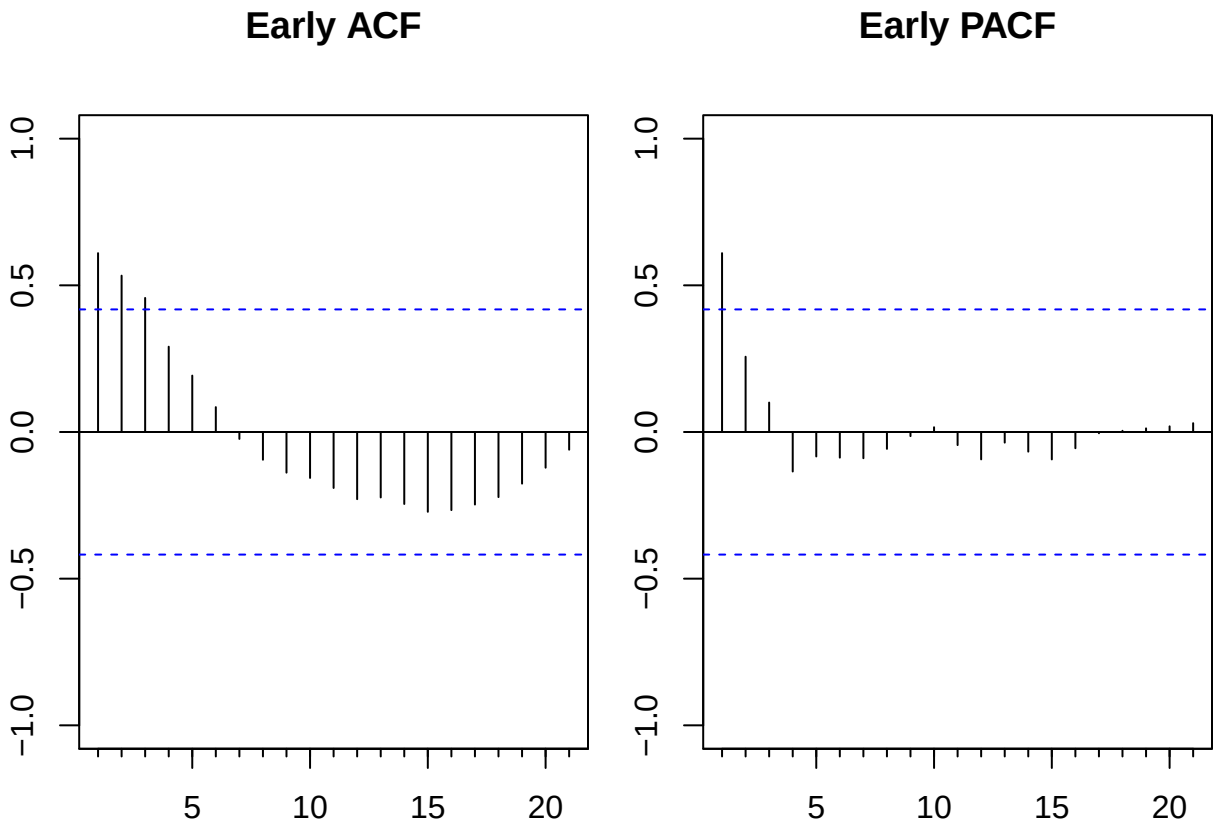

late.test.ts
3.5e+11
3.0e+11
2.5e+11
2.0e+11
2020
2021
Time

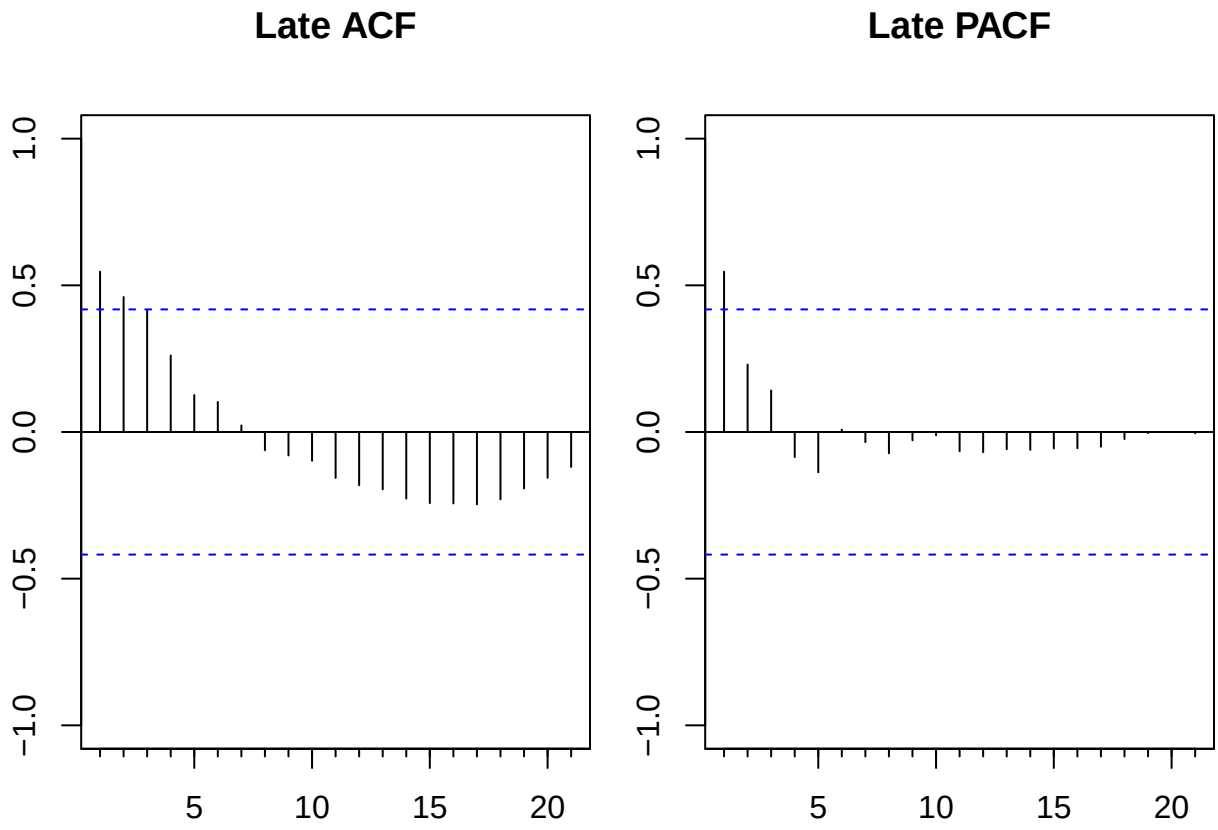


#Visualization

As you can see, both early and late stage funding has increased drastically over the years. Late stage passed early stage around 2016 and has been greater than early stage ever since.

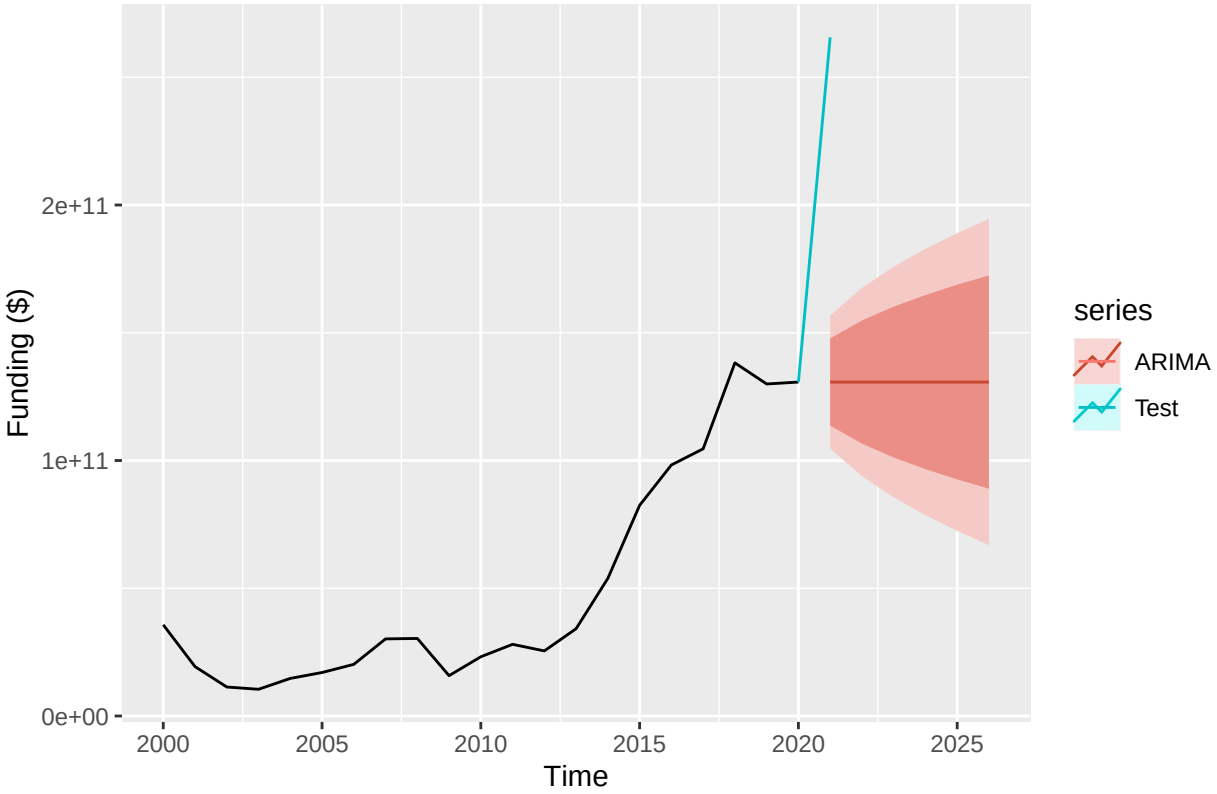
#Create ACFs and PACFs


Early ACF
Early PACF
1.0
0.5
0.0
−0.5
−1.0
5
10
15
20

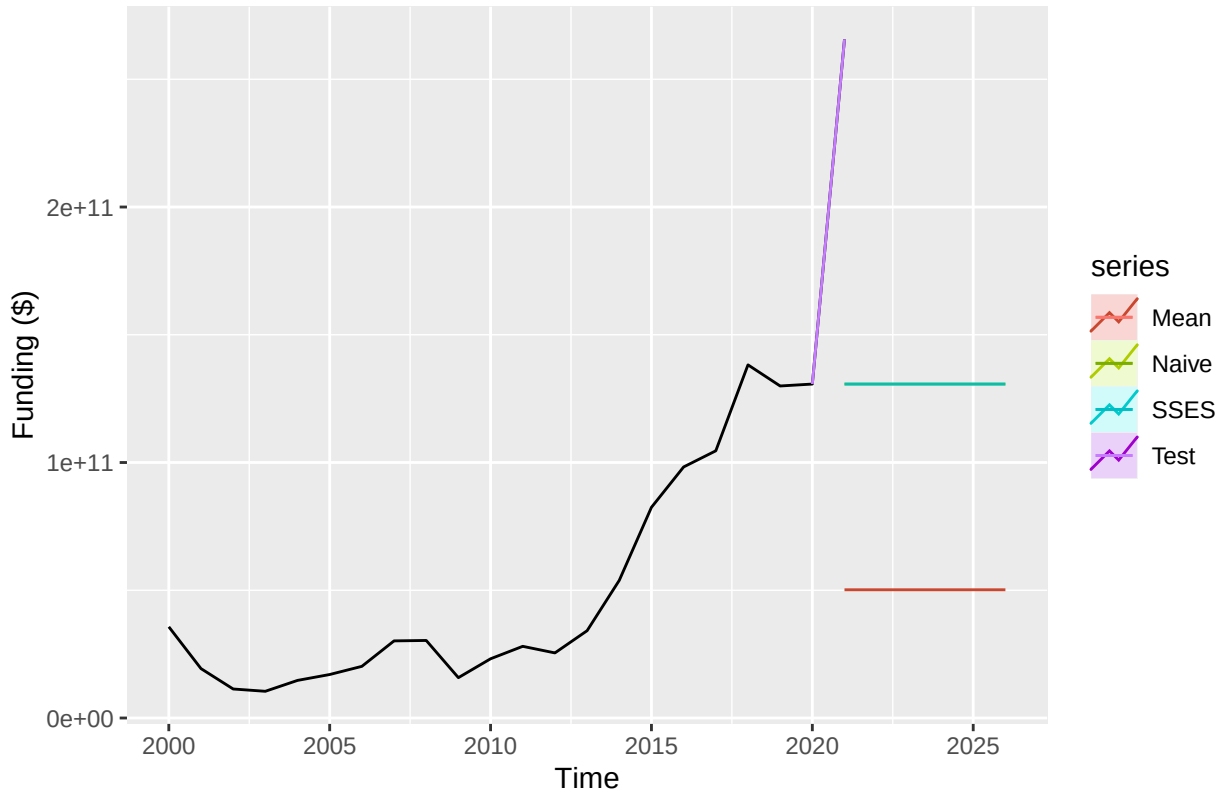


#Model Fit - Arima (Early)

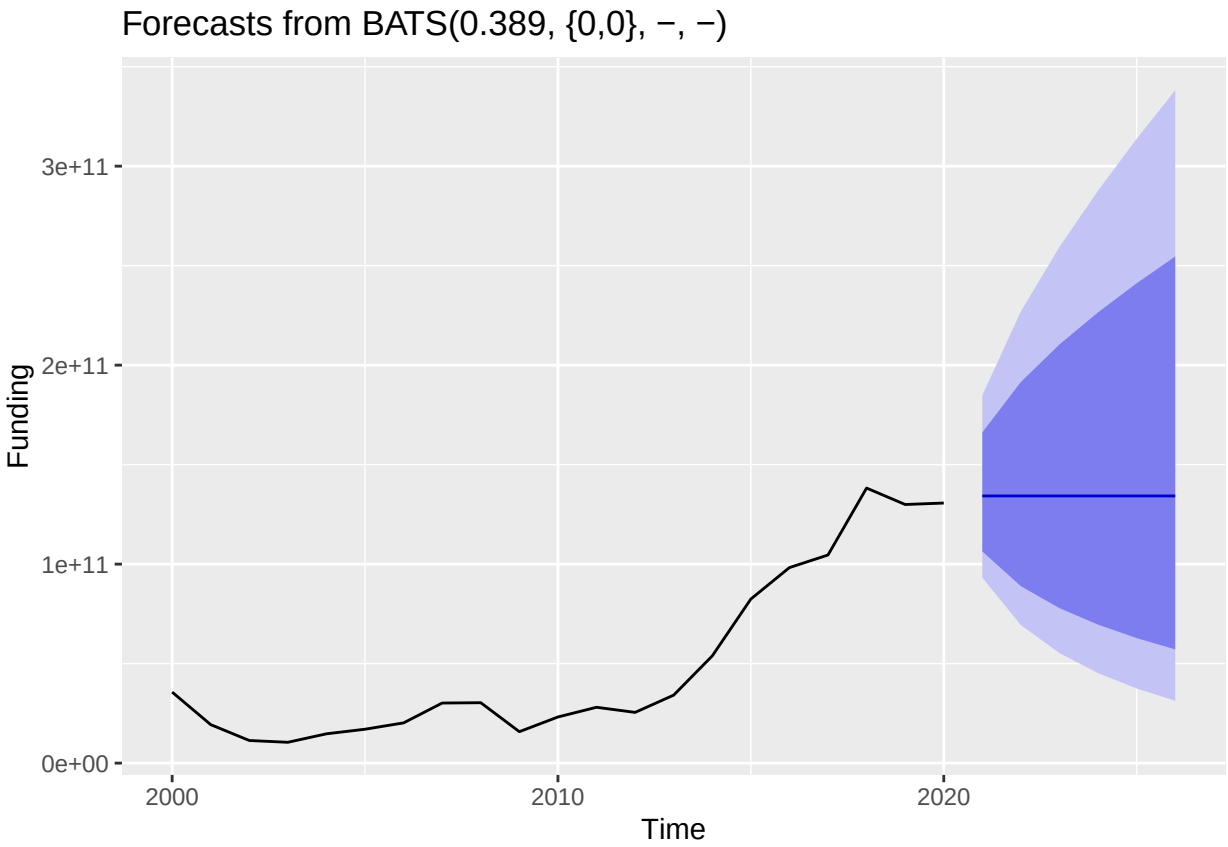
```
## Series: early.training.ts
## ARIMA(0,1,0)
##
## sigma^2 = 1.767e+20:  log likelihood = -494.59
## AIC=991.17   AICc=991.4   BIC=992.17
##
## Training set error measures:
##                       ME        RMSE        MAE     MPE     MAPE      MASE
## Training set 4525751369 12971505571 9348854382 0.60679 26.52674 0.9525542
##                    ACF1
## Training set 0.2714486
```

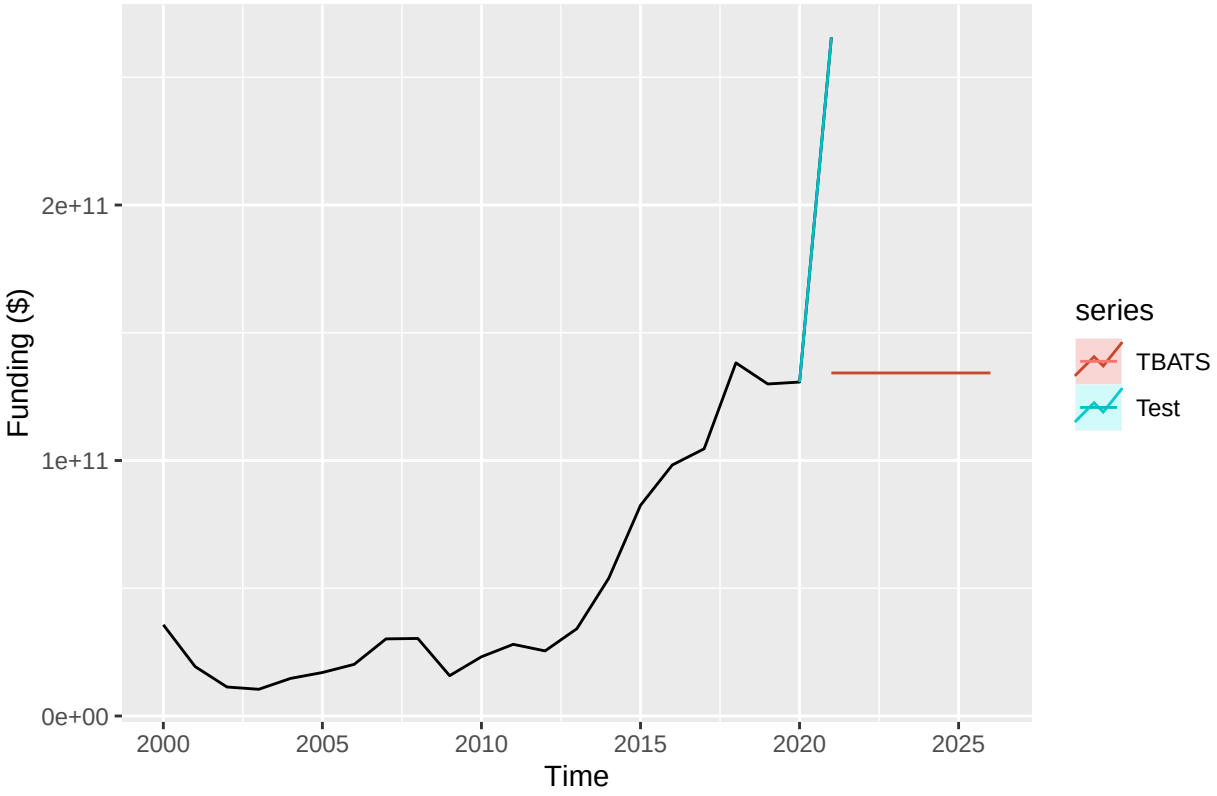


#Model Fit - Mean, Naive, SSES (Early)

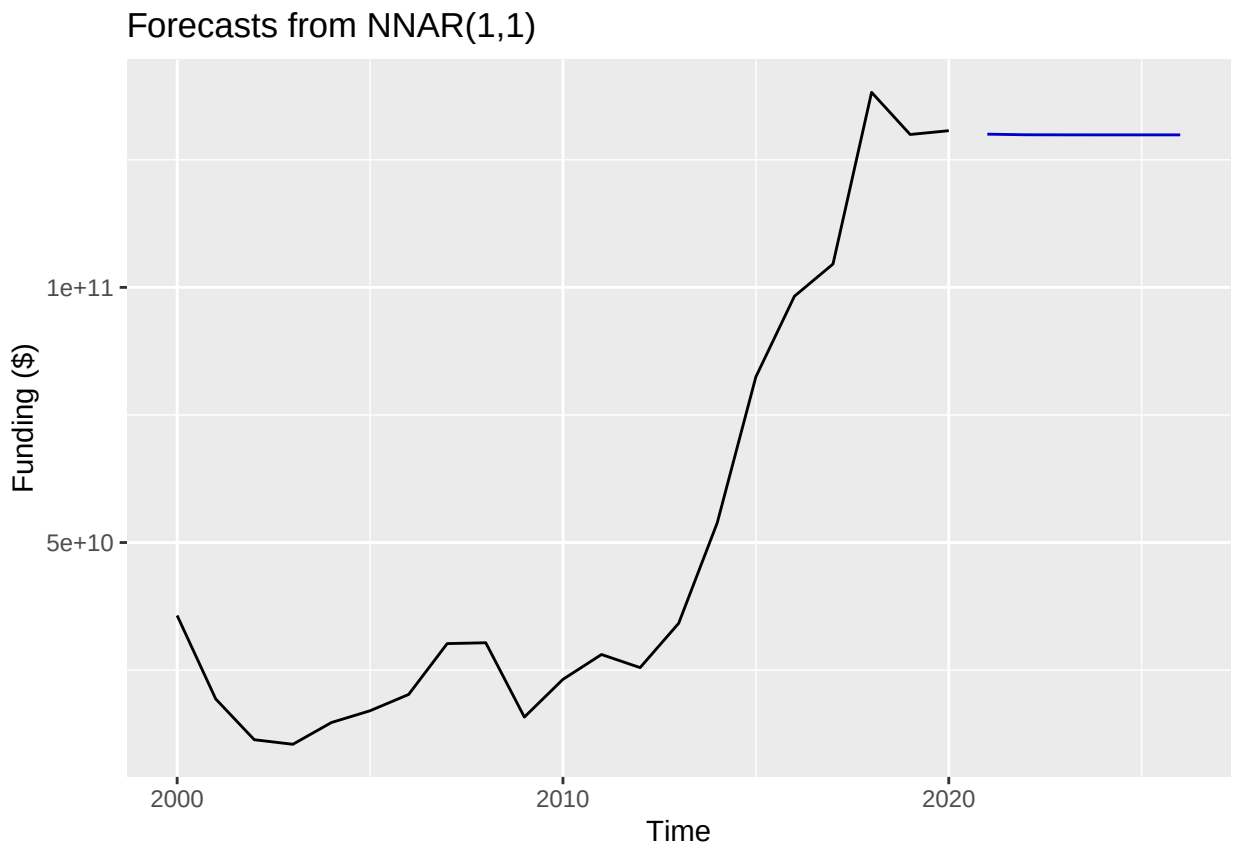


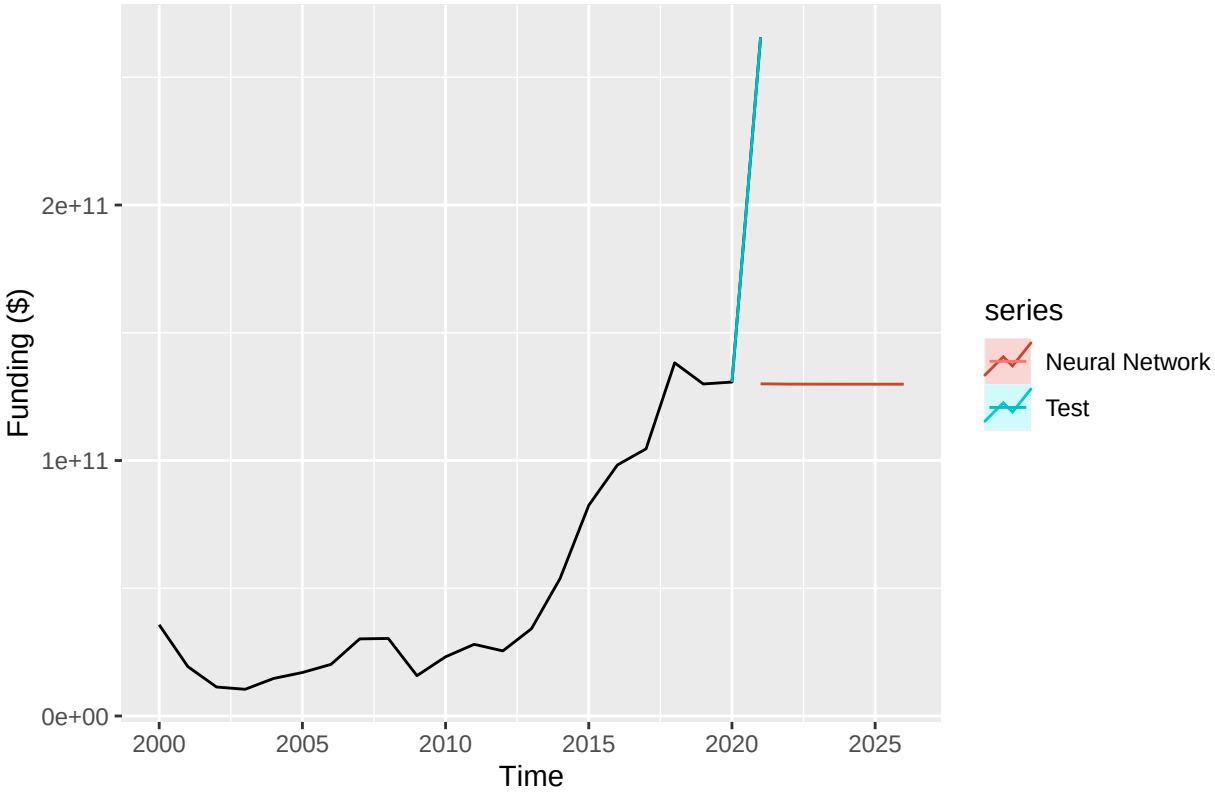
#Model Fit - TBATS (Early)

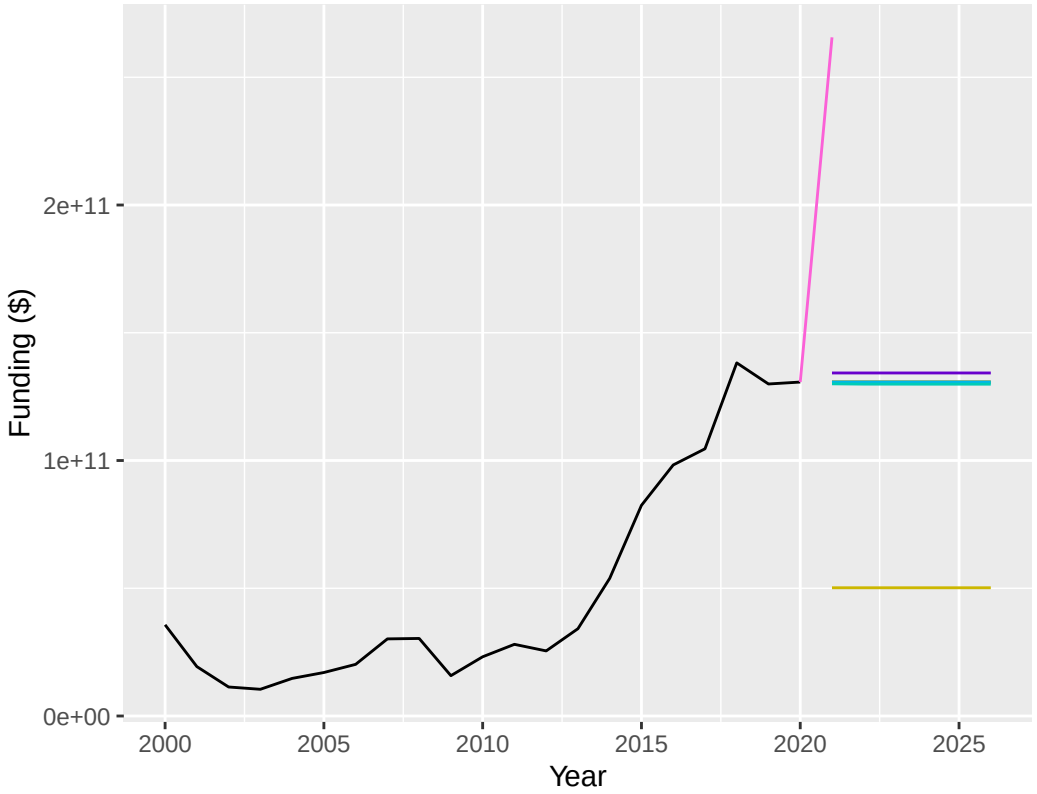

Forecasts from BATS(0.389, {0,0}, −, −)
Funding
3e+11
2e+11
1e+11
0e+00
2000
2010
2020
Time



#Model Fit - NN (Early)


Forecasts from NNAR(1,1)
Funding ($)
1e+11
5e+10
2000
2010
2020
Time


Funding ($)
2e+11
1e+11
0e+00
2000
2005
2010
2015
2020
2025
Time
series
Neural Network
Test



#Model Plotting (Early)

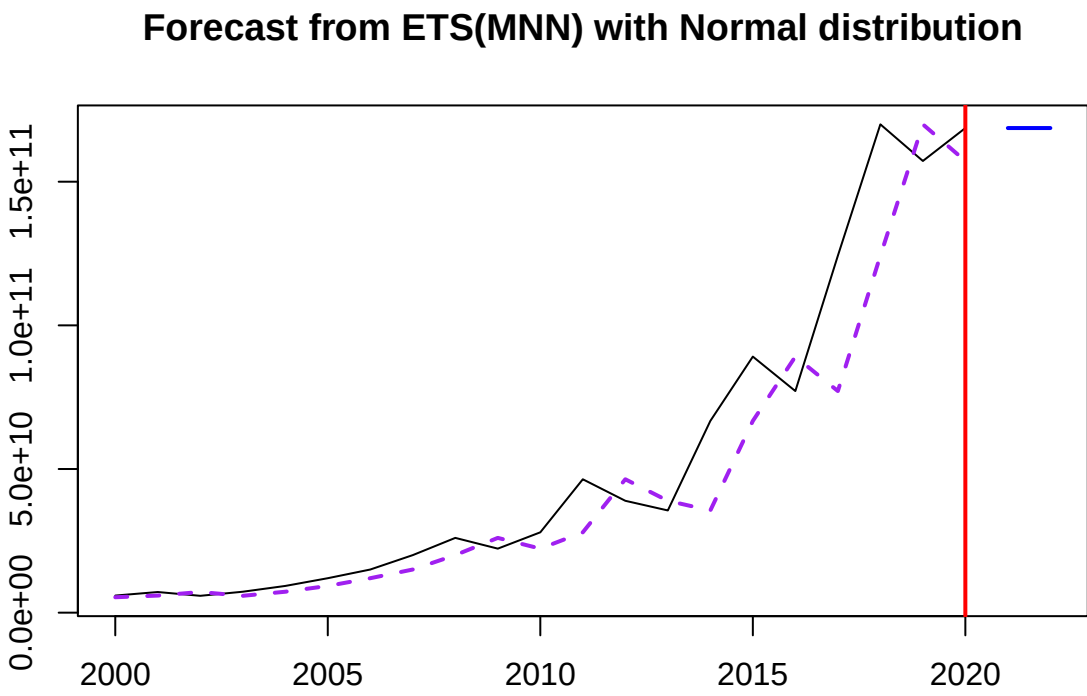
#Model Scoring (Early)

```
##                          ME         RMSE          MAE      MPE     MAPE
## ARIMA           134920037676 134920037676 134920037676 50.79304 50.79304
## Mean            215430828331 215430828331 215430828331 81.10275 81.10275
## Naive           134920037676 134920037676 134920037676 50.79304 50.79304
## SSES            134920037676 134920037676 134920037676 50.79304 50.79304
## TBATS           131357180623 131357180623 131357180623 49.45174 49.45174
## Neural Network  135605199642 135605199642 135605199642 51.05098 51.05098
```

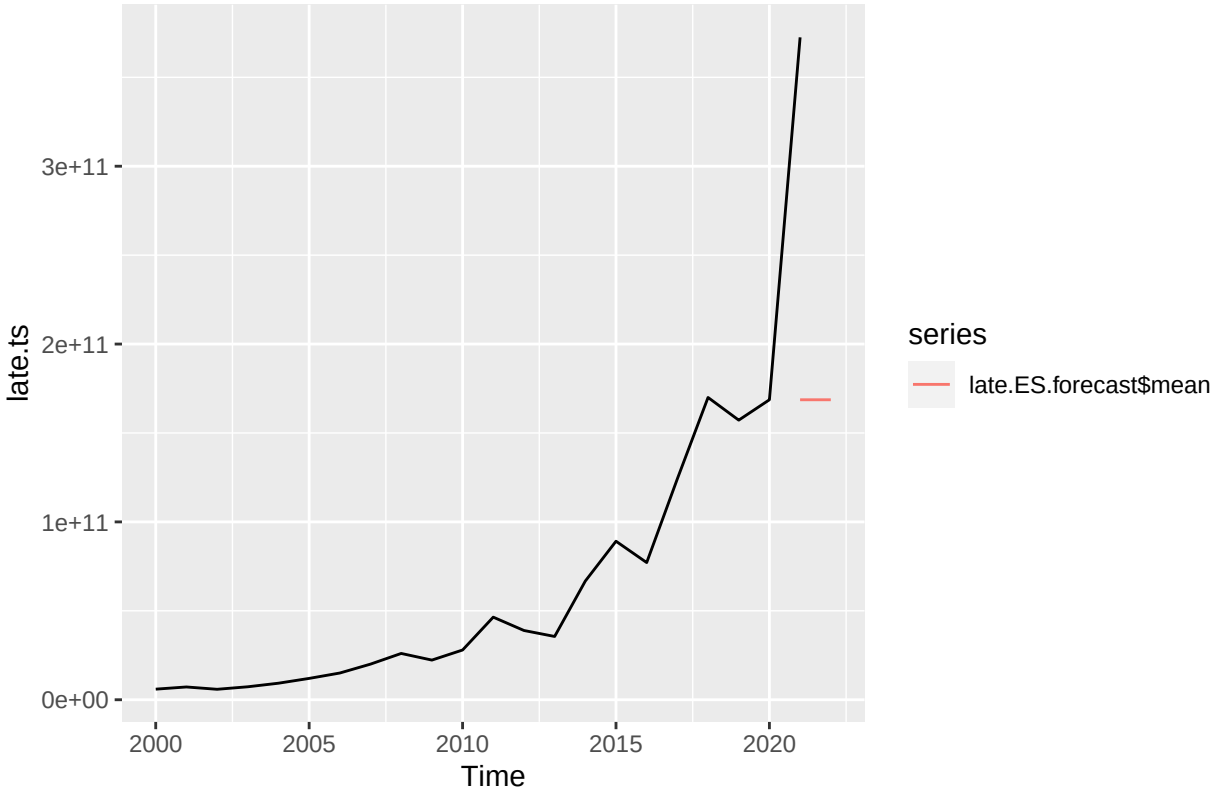
```
## The best model by RMSE is: TBATS
```

#RMSE - Early

```
## Error in kable_styling(., latex_options = "striped", stripe_index = which.min(early.scores[, : could
```

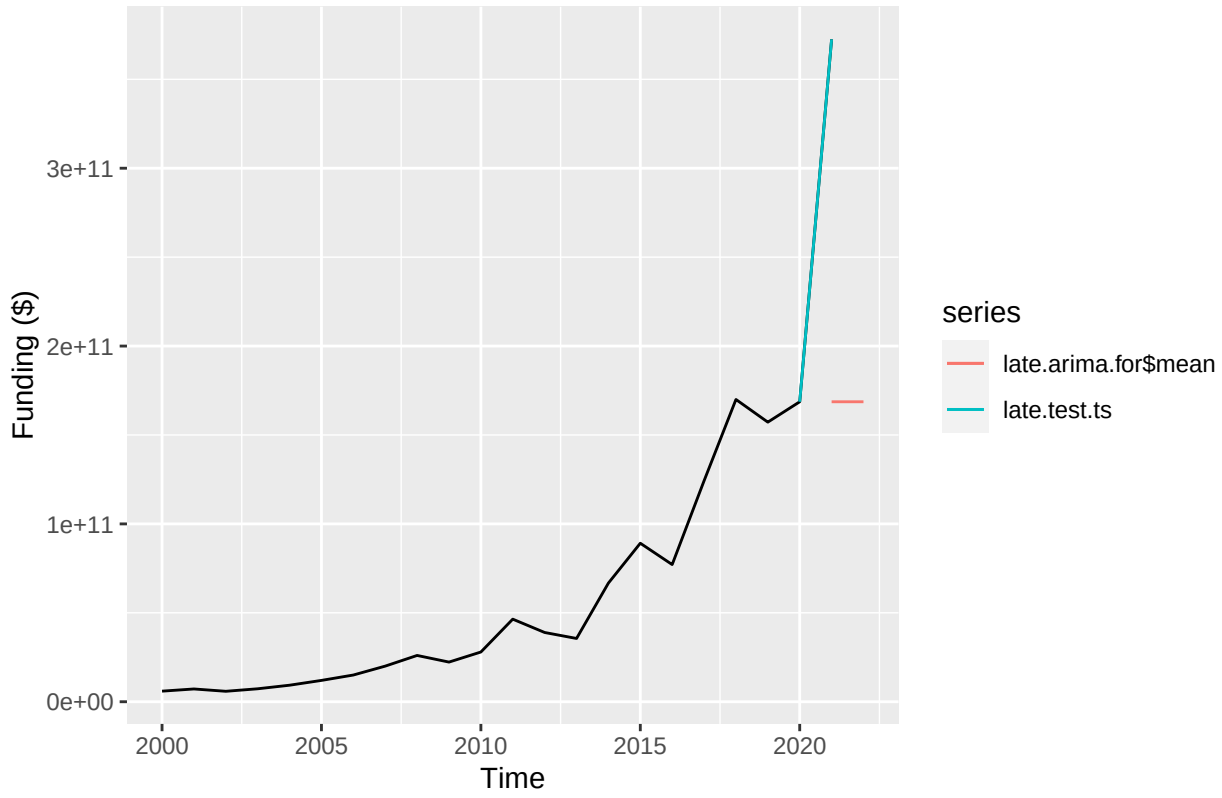


#Model Fit - ES (Late)

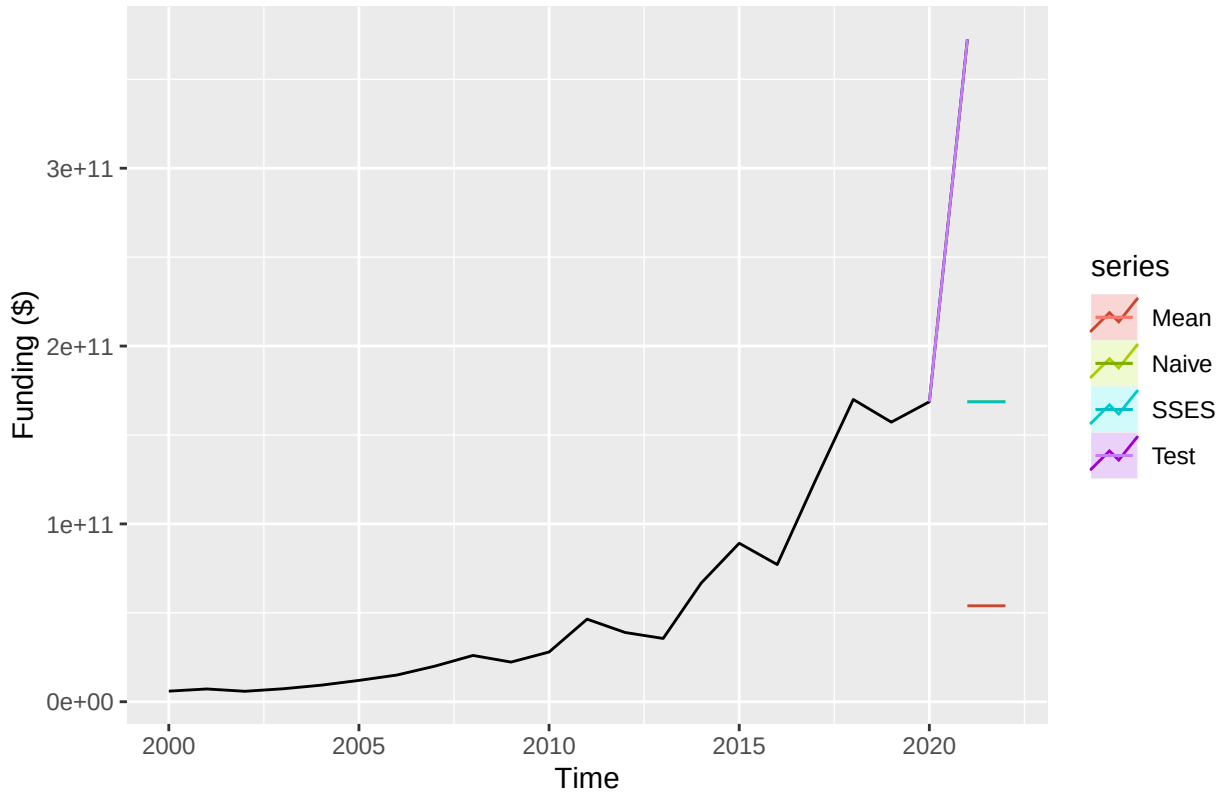


#Model Fit - Arima (Late)

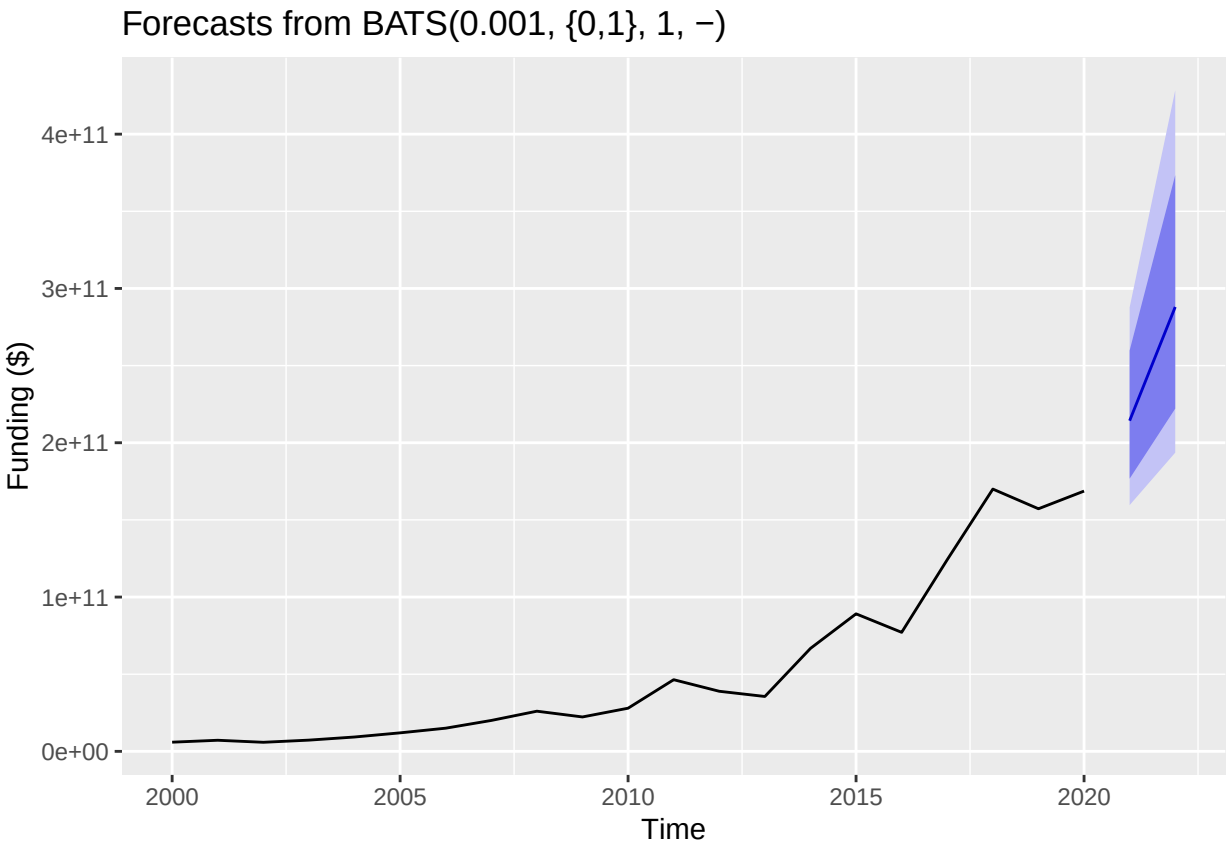
```
## Series: late.training.ts
## ARIMA(0,1,0)
##
## sigma^2 = 3.38e+20:  log likelihood = -501.08
## AIC=1004.15   AICc=1004.37   BIC=1005.15
##
## Training set error measures:
##                      ME        RMSE         MAE      MPE     MAPE      MASE
## Training set 7748821973 17942595684 11619982072 12.42824 21.15022 0.9524042
##                      ACF1
## Training set -0.007499492
```

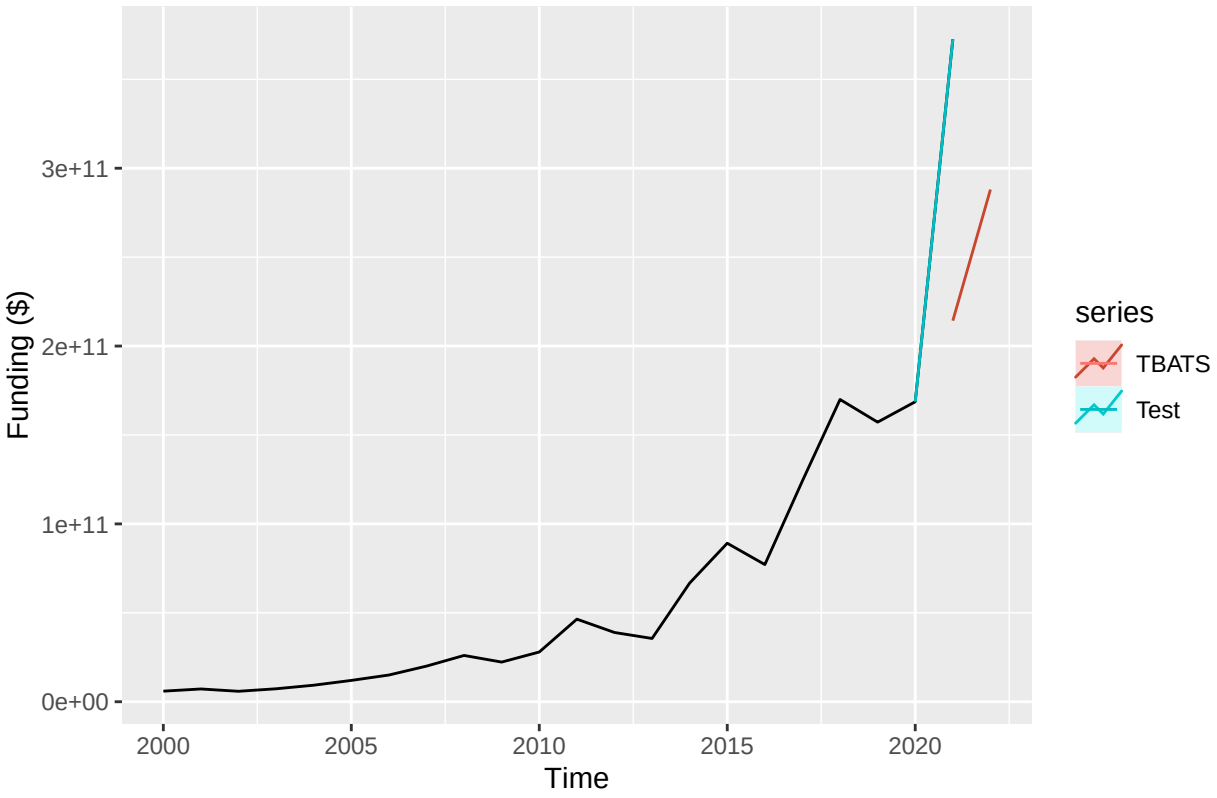


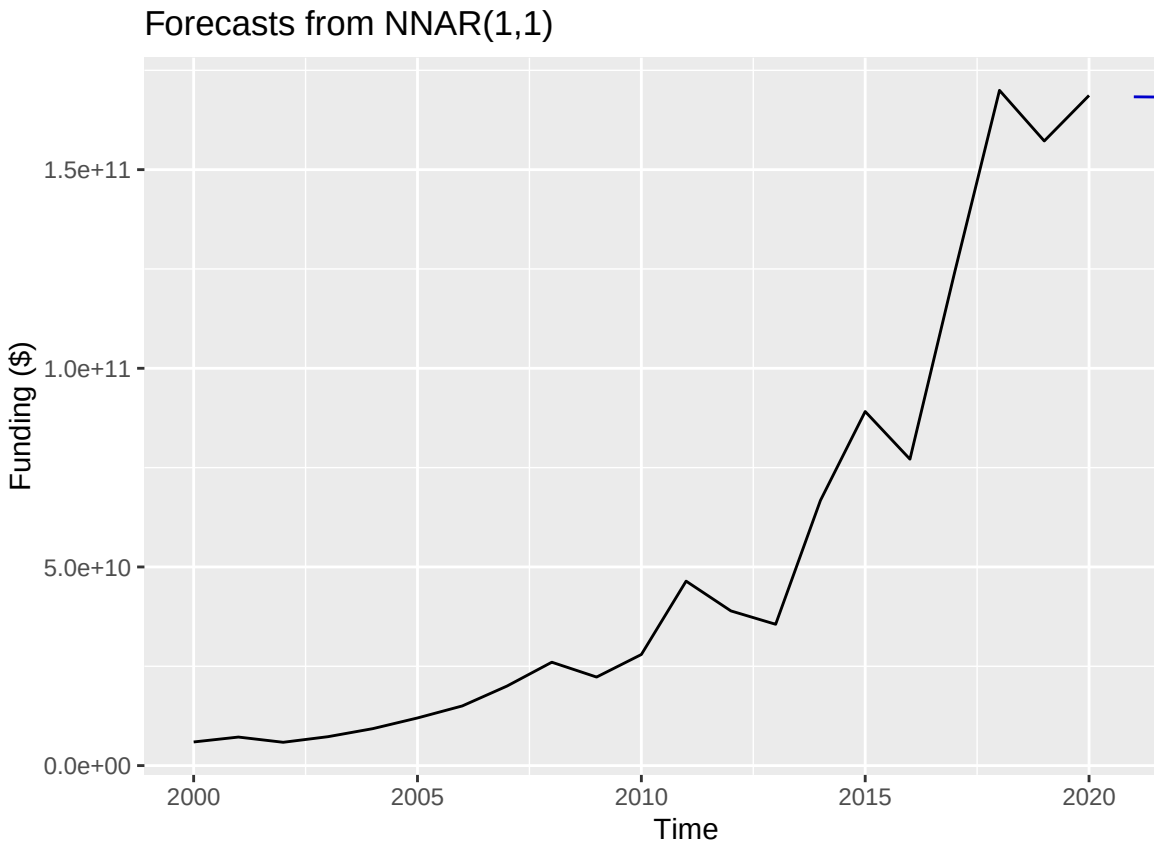
#Model Fit - Mean, Naive, SSES (Late)



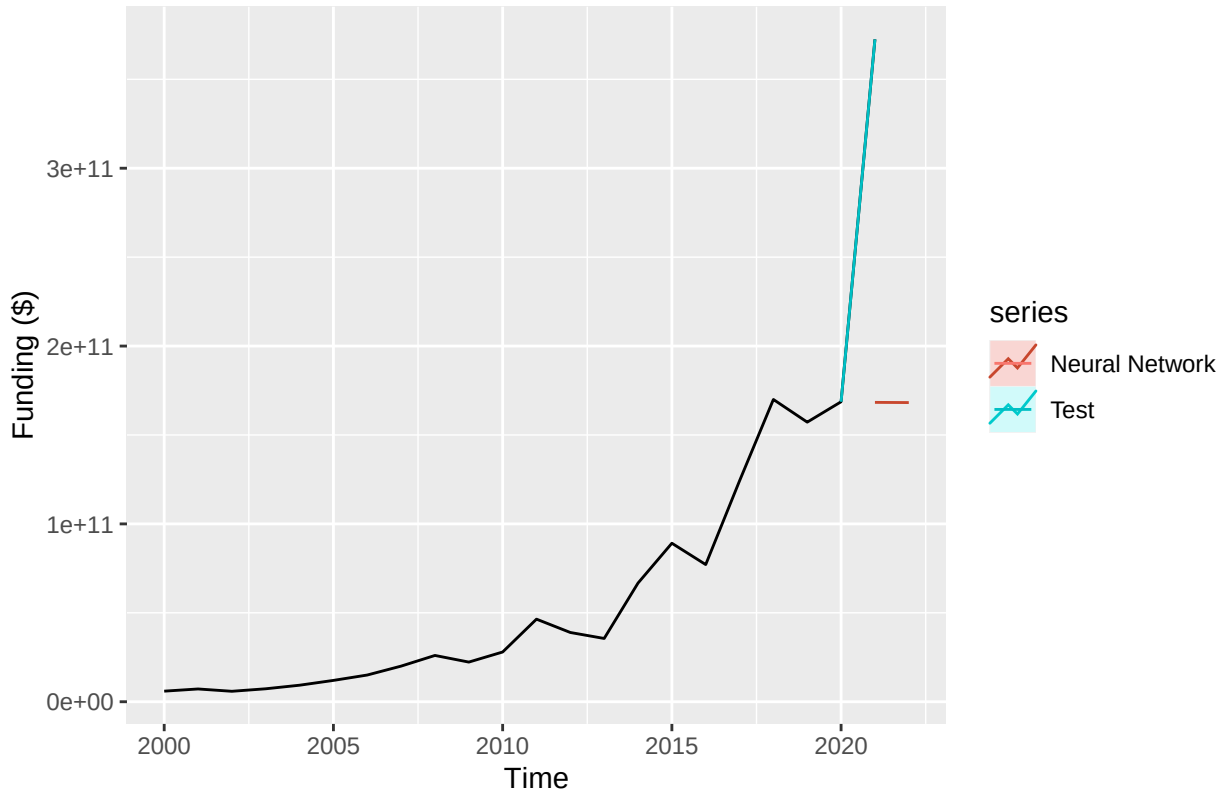
#Model Fit - TBATS (Late)

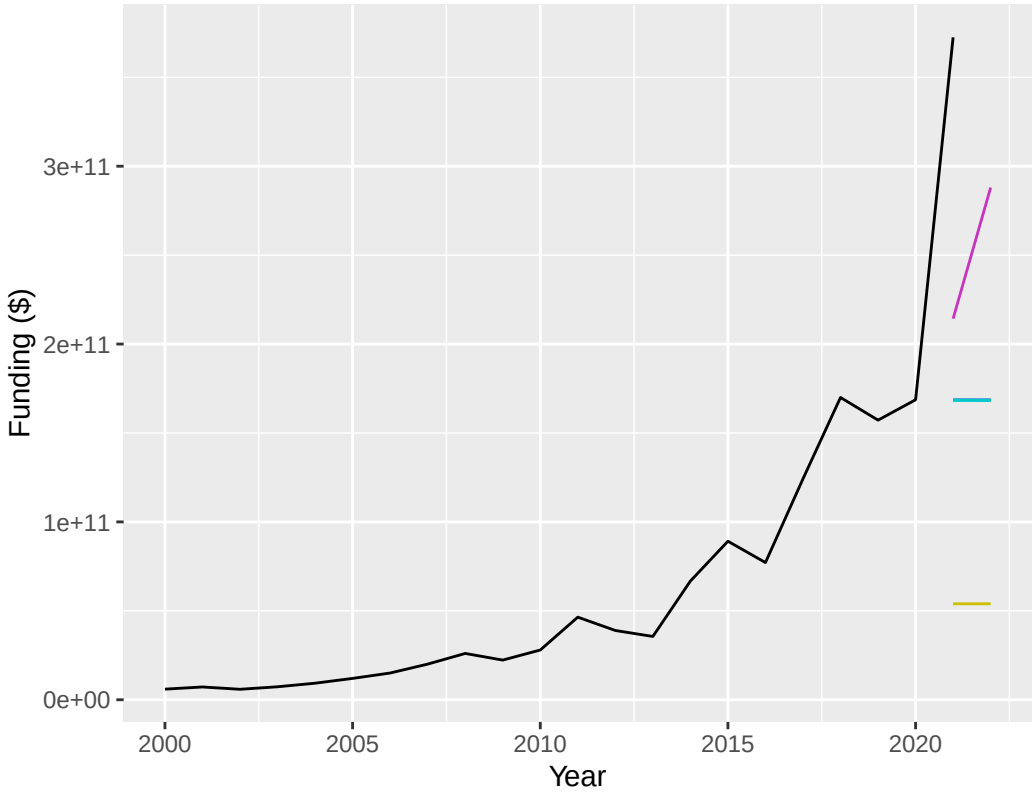

Forecasts from BATS(0.001, {0,1}, 1, −)
Funding ($)
4e+11
3e+11
2e+11
1e+11
0e+00
2000
2005
2010
2015
2020
Time


Funding ($)
3e+11
2e+11
1e+11
0e+00
2000
2005
2010
2015
2020
Time
series
TBATS
Test



#Model Fit - NN (Late)


Funding ($)
3e+11
2e+11
1e+11
0e+00
2000
2005
2010
2015
2020
Time
series
Neural Network
Test



#Model Plotting (Late)

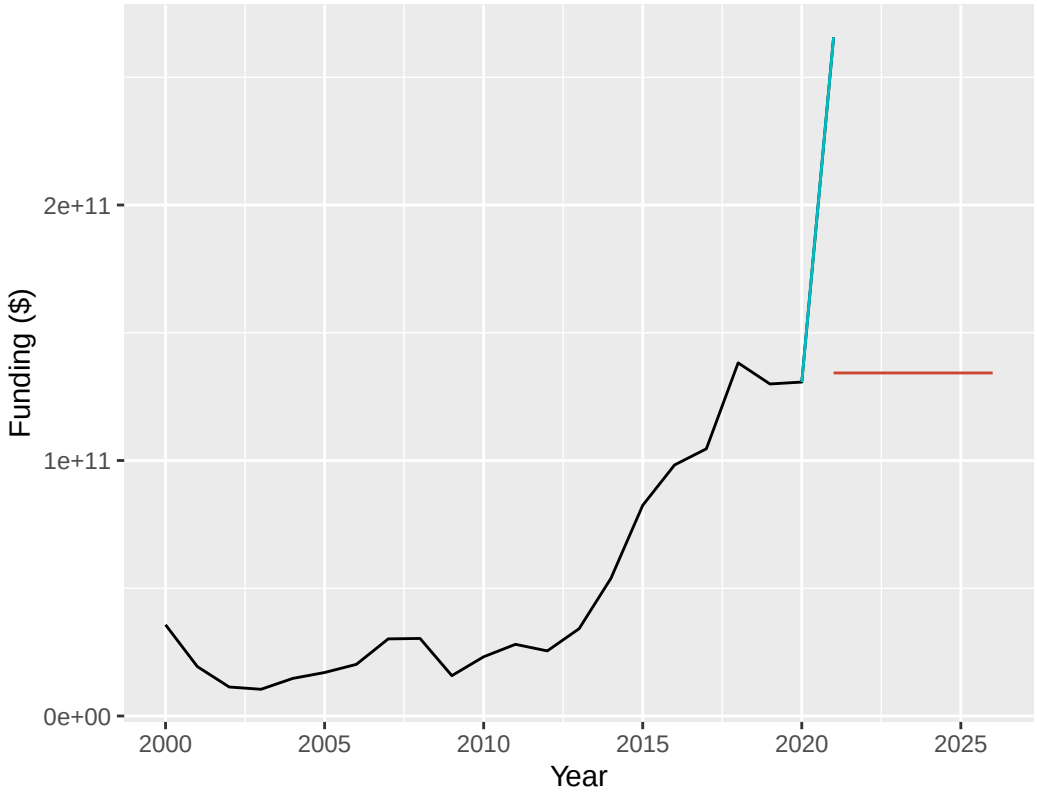
#Model Scoring (Late)

```
##                          ME         RMSE          MAE      MPE     MAPE
## ARIMA           203812130292 203812130292 203812130292 54.71748 54.71748
## Mean            318535727367 318535727367 318535727367 85.51734 85.51734
## Naive           203812130292 203812130292 203812130292 54.71748 54.71748
## SSES            203812130292 203812130292 203812130292 54.71748 54.71748
## TBATS           158210031971 158210031971 158210031971 42.47468 42.47468
## Neural Network  204156118479 204156118479 204156118479 54.80983 54.80983
```

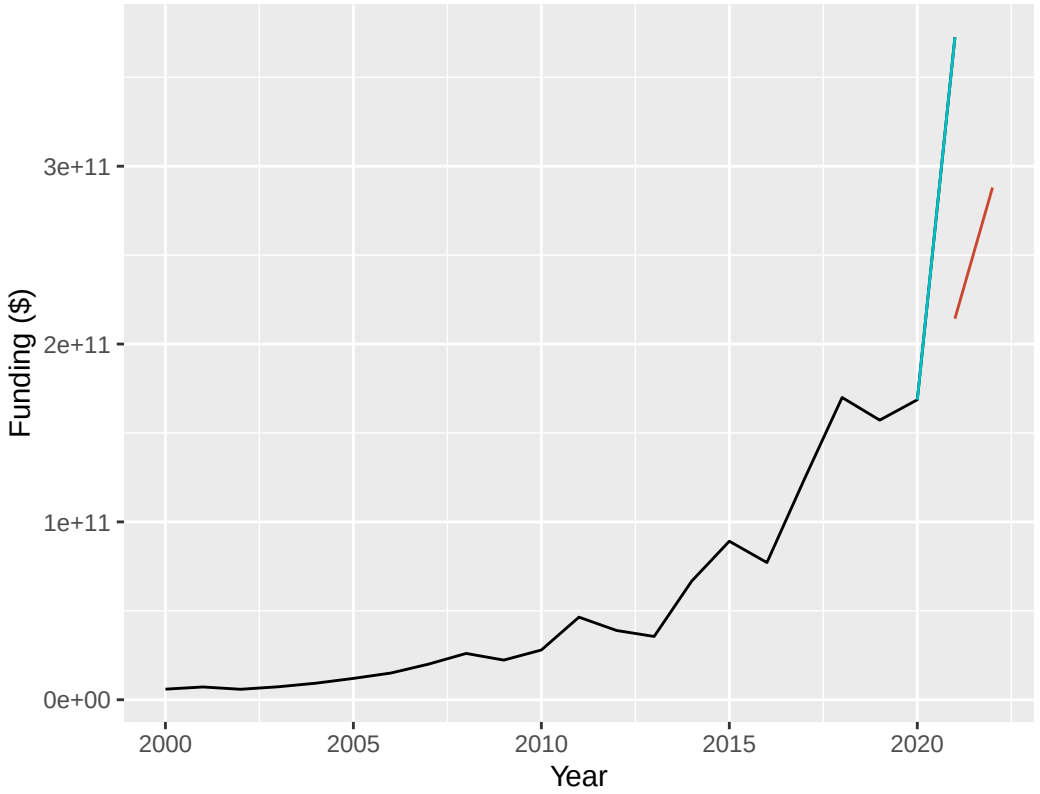
```
## The best model by RMSE is: TBATS
```

#RMSE - Late

```
## Error in kable_styling(., latex_options = "striped", stripe_index = which.min(late.scores[, : could n
```



#Plot the Best Model (Early)



#Plot the Best Model (Late)

Conclusions There are some limitations to this dataset which have made it difficult to predict. These limitations made predictions difficult. It is unclear whether there will continue to be an increase in clean tech startup investment in the future. Looking at the "best" model, the forecast for early stage investment seems to stay constant (a straight line), where late stage investment has an increase.This shows that, with the limited data available, there will be an overall increase in clean tech investment and the late stage investment will continue to be greater than early stage investment.

Possible Future Steps Analyze the breakdown of early vs. late-stage investment by sector (energy, non-energy, grid, fossil fuel, energy efficiency, etc.)