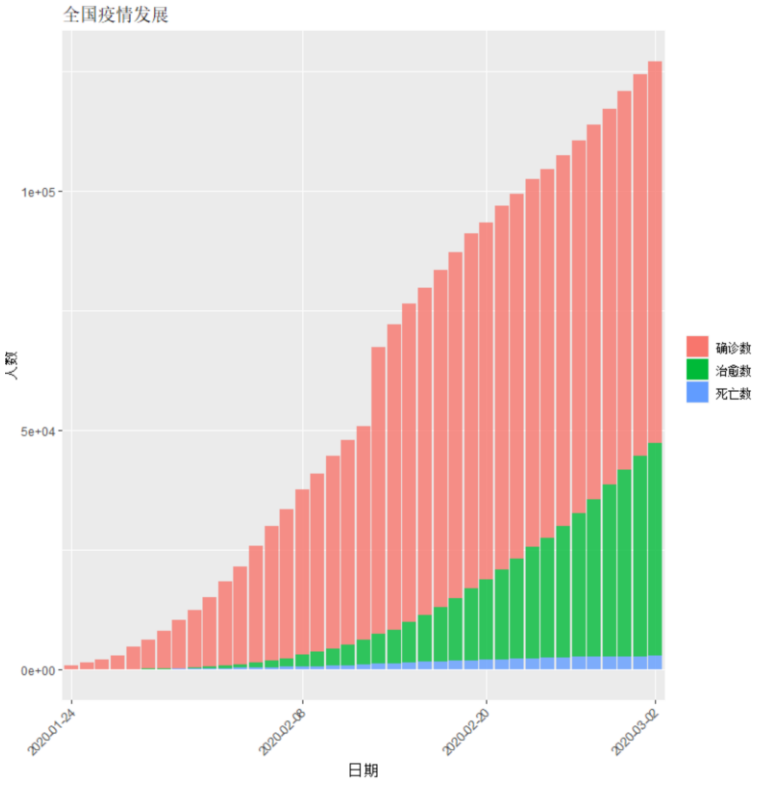
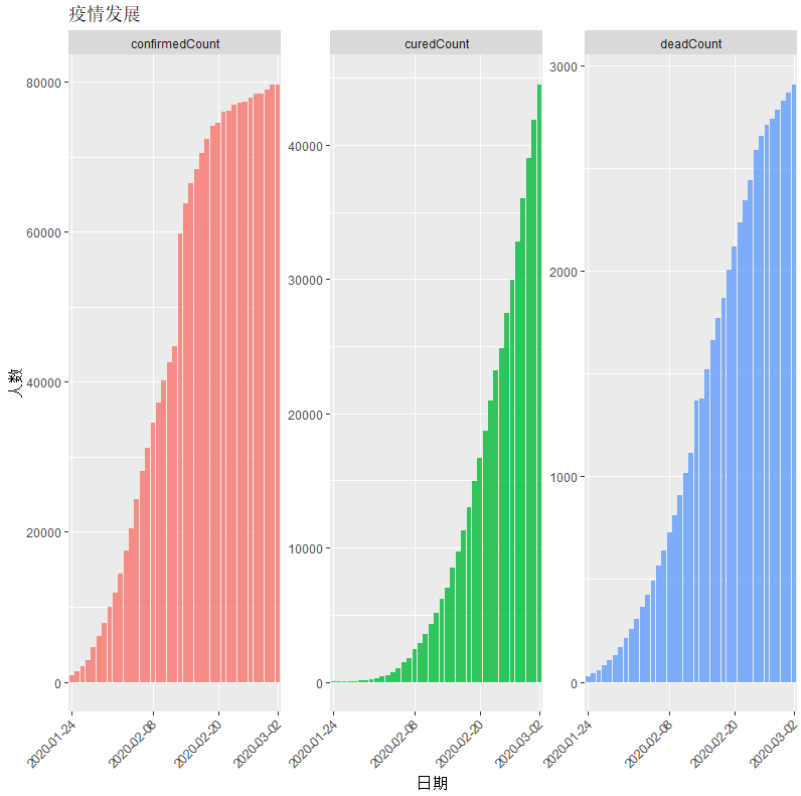
该 project 做的是新冠的可视化分析及时间序列预测
首先，数据来源于 github：
https://github.com/BlankerL/DXY-COVID-19-Data/releases/tag/2020.12.22
数据并不完整，做这个是想挑战一下自己
第一部分：数据整理
首先载入数据，并转 tibble 表
对数据执行一系列操作
从全球数据筛选中国数据，筛选确诊，治愈，死亡数据
因为每天更新数次，每天有好几个数据，没有每天单独的数据
于是通过切片得到日期和时分秒，再以以日期省市为分组，筛选最大时分秒，以此得到每日最新数据
再剔除国内没有数据的日期
再以日期为组，将各省市数据加起来得到全国数据
不过数据不完整
只有前 39 行较准确，故取前 39 行

再挑选数据完整的 2020.2.23.以省份为组，将各市数据加起来得到各省当天数据
省份数据中缺失台湾，香港，故手动添加台湾，香港数据

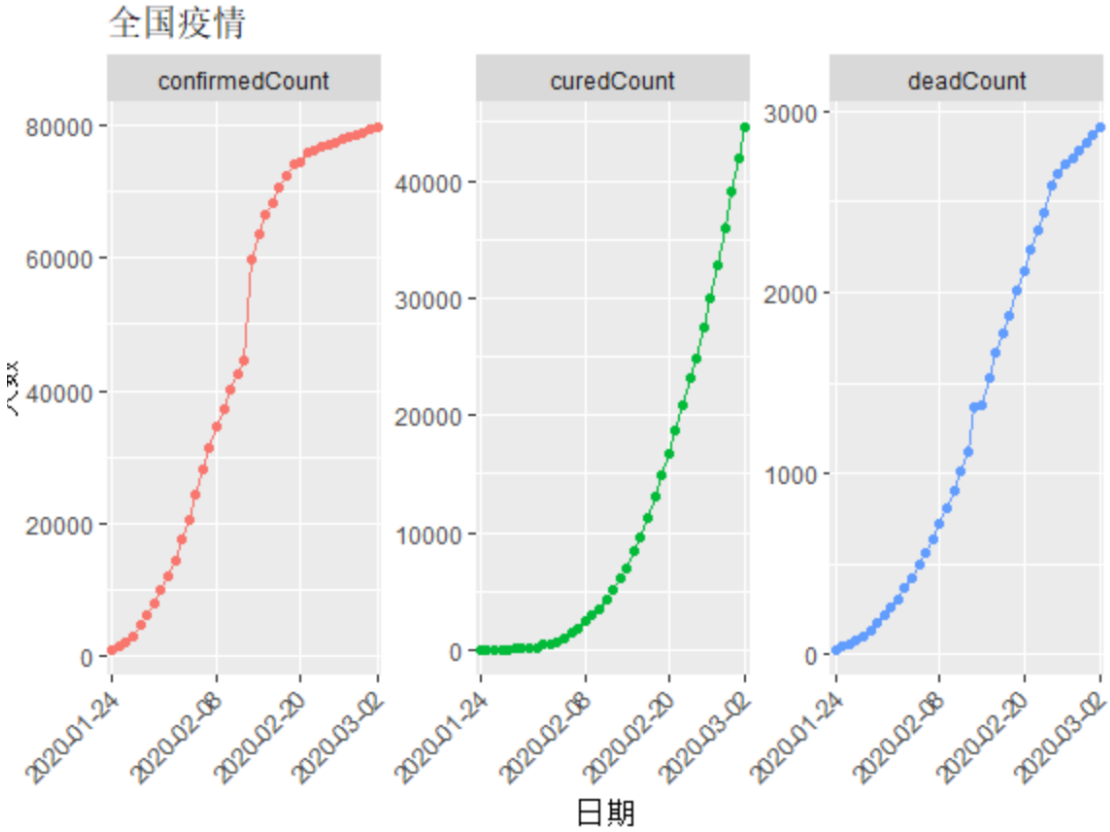
再为各省确诊人数设置级别，各级别转因子
第二部分：可视化
全国确诊情况图



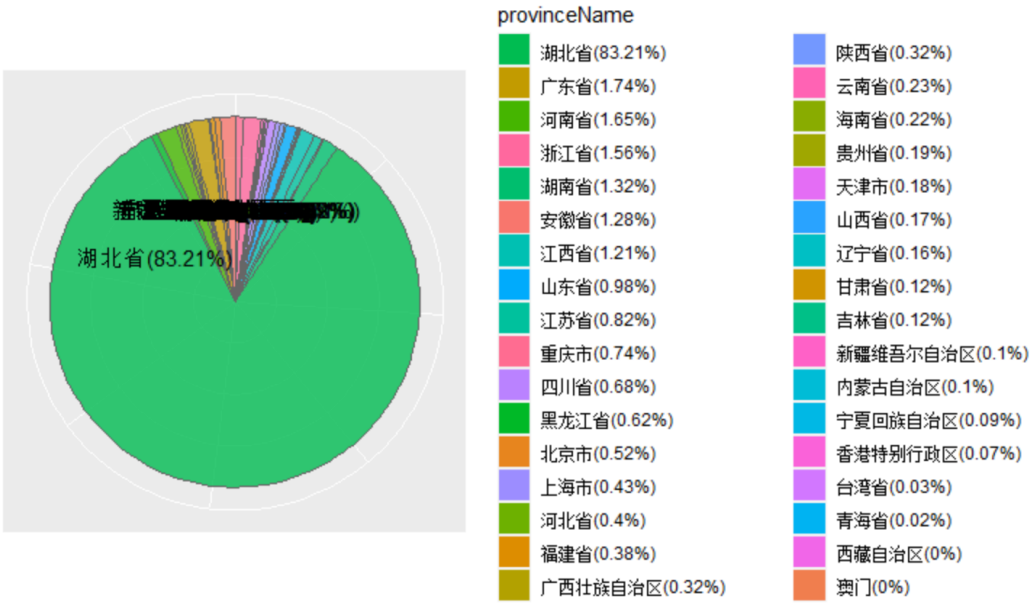
facet_wrap 分开的图



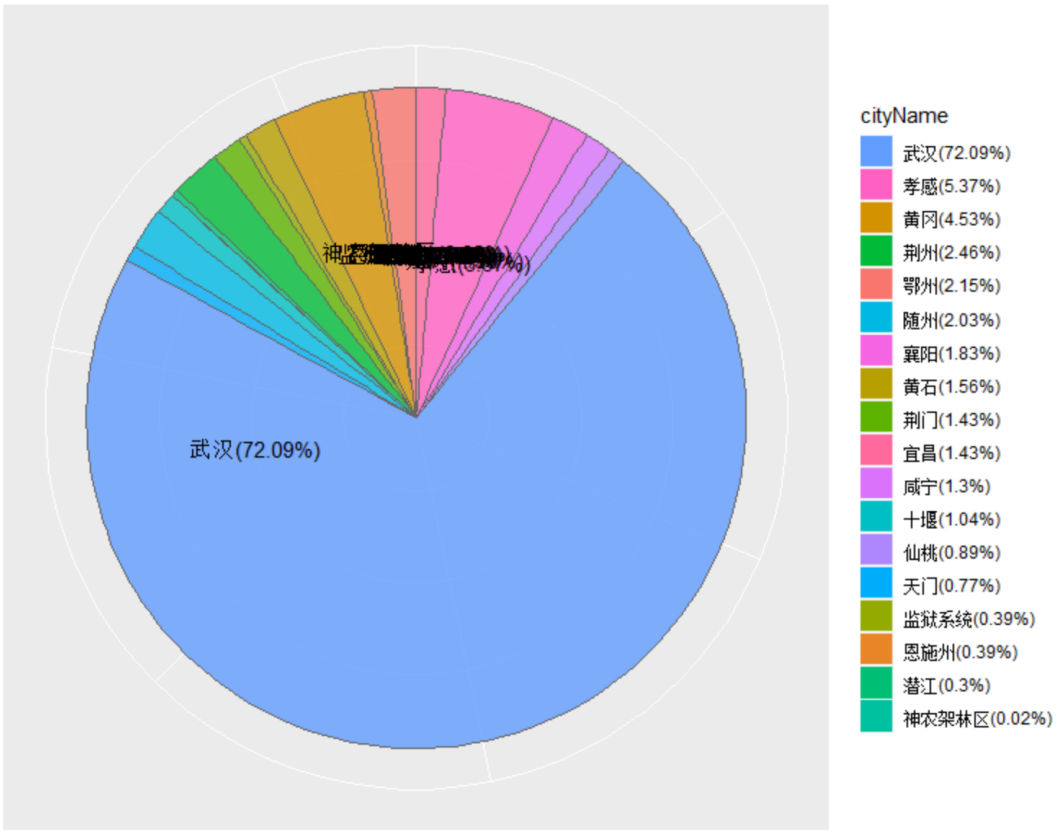
全国疫情数据折线图



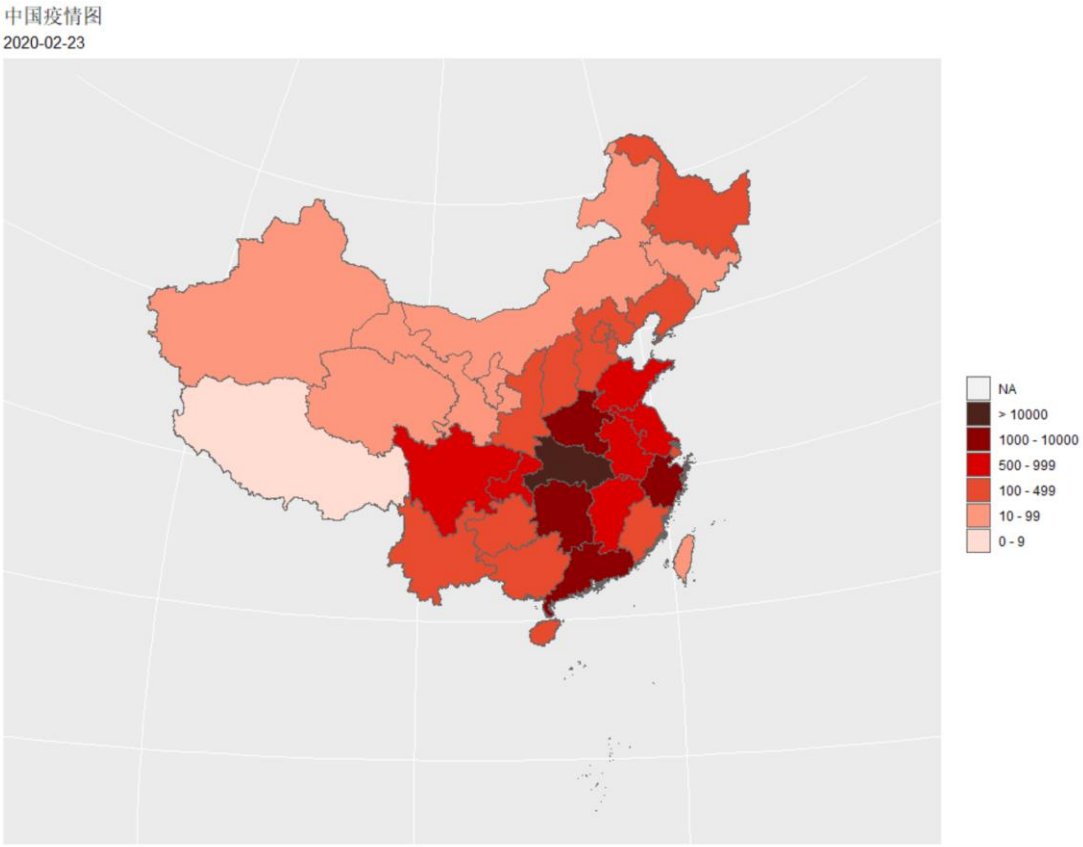
从上边的图中可以看出确诊数被控制了下来，死亡数也有所趋缓，治愈数极速增加



2.23 日各省市确诊数占比，可以看出 83%的确诊案例在湖北省，湖北的疫情严重
再画出 2.23 日湖北各市确诊数占比，可以看出湖北省 72%的确诊案例在武汉市，武汉市的疫情最严重



再导入地图文件将疫情数据与地图数据关联整合，得到 2 月 23 号的中国疫情地图



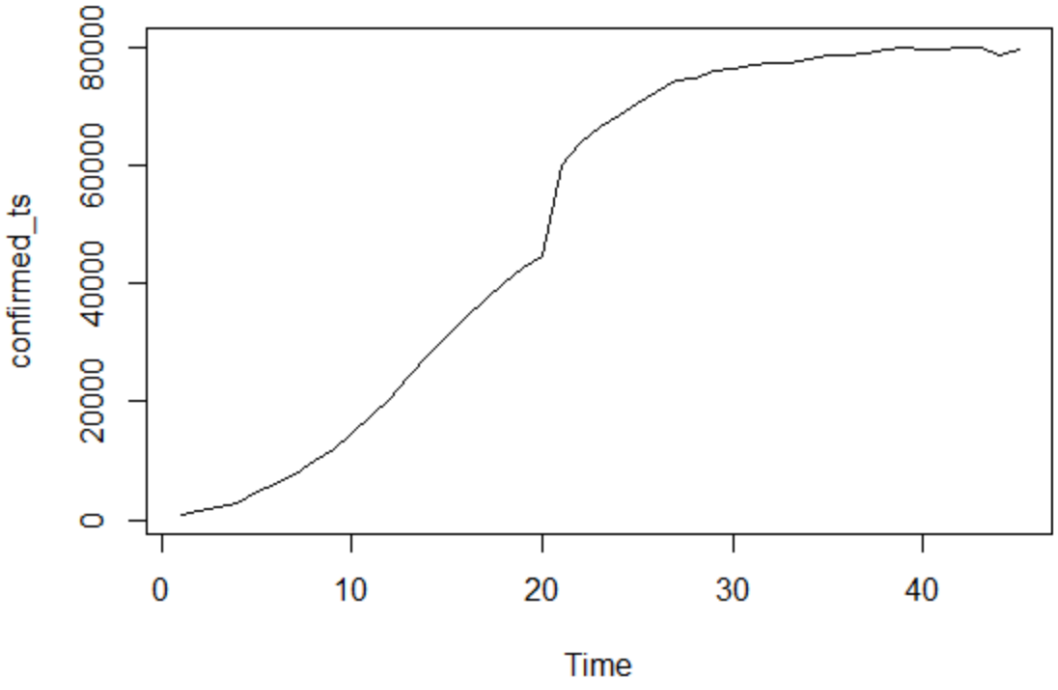
从图中可以看出，湖北疫情最严重，其次是湖南，河南，广东，江苏，34 个省份中西藏状况最佳

第三部分：

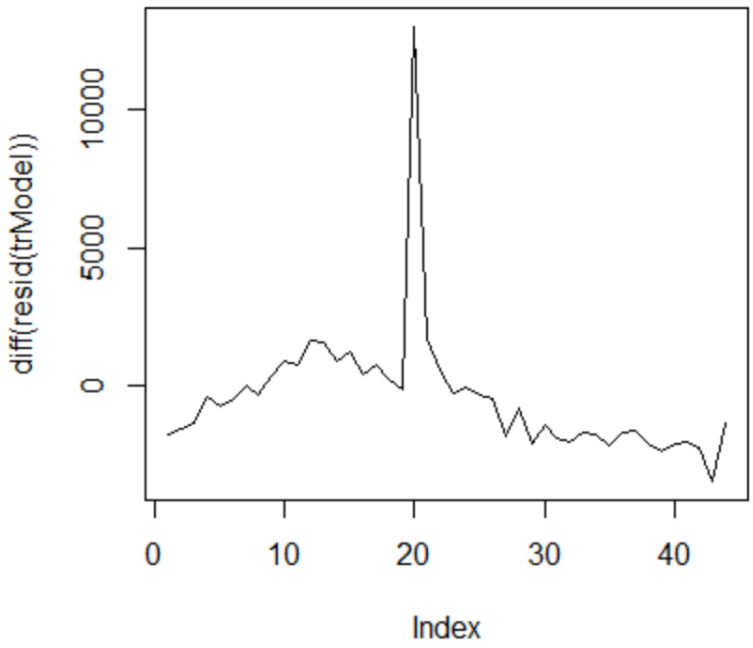
挑选数据较准确的 45 天

将确诊人数转化成时间序列

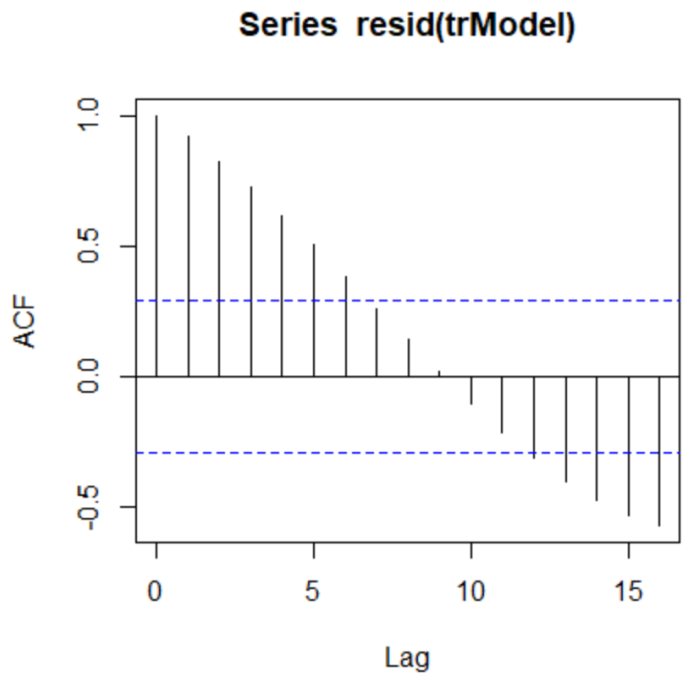
画出前 45 天确诊人数图



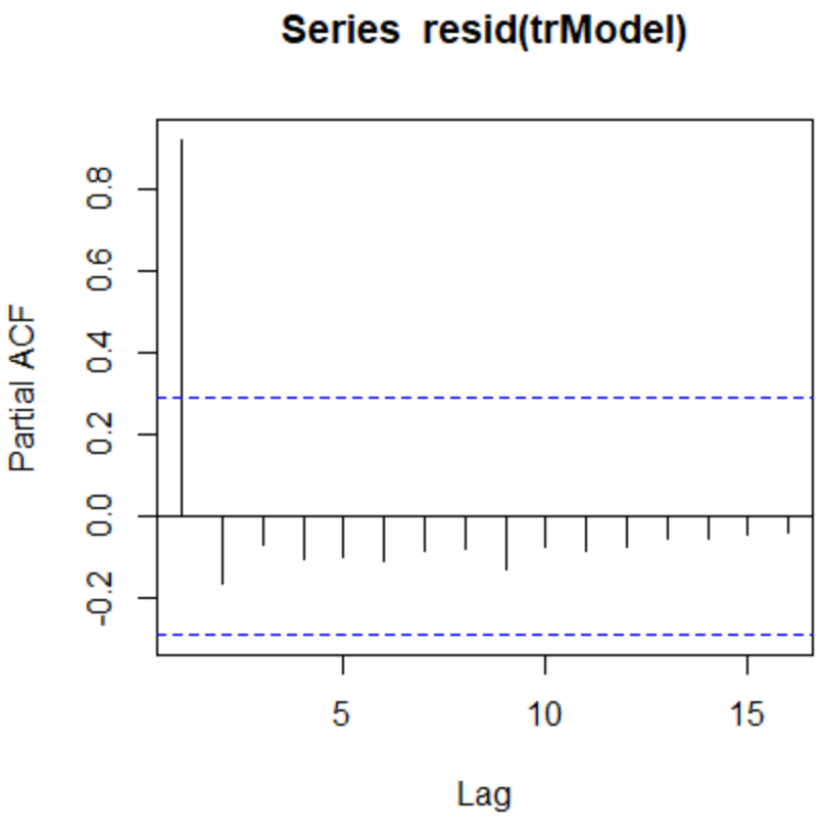
差分


diff(resid(trModel))
10000
5000
0
0
10
20
30
40
Index

ACF


Series resid(trModel)
ACF
1.0
0.5
0.0
-0.5
0
5
10
15
Lag

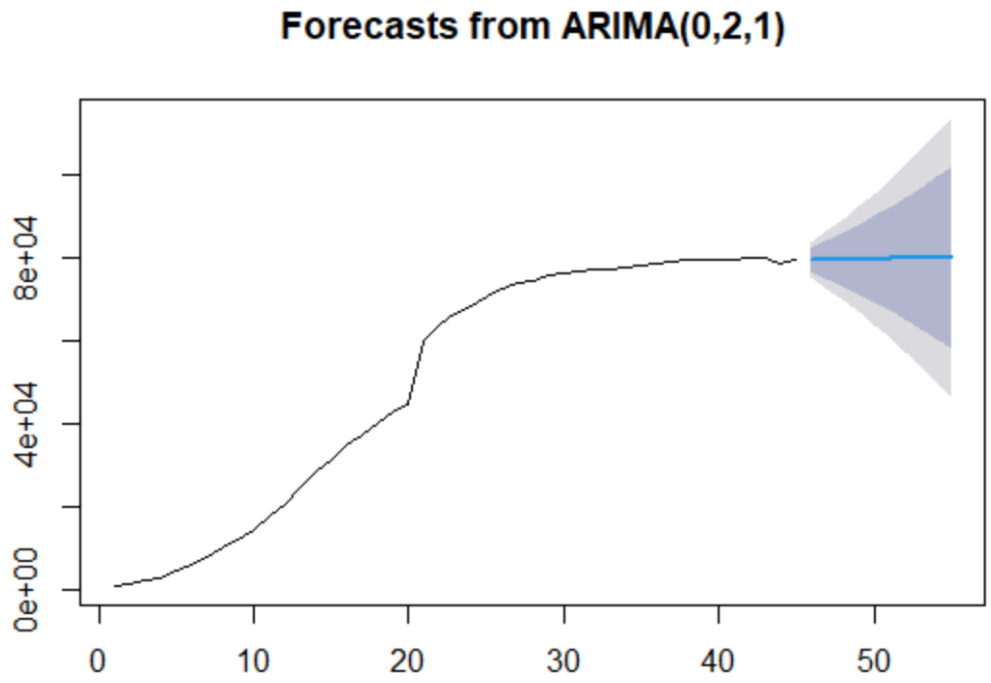
PACF



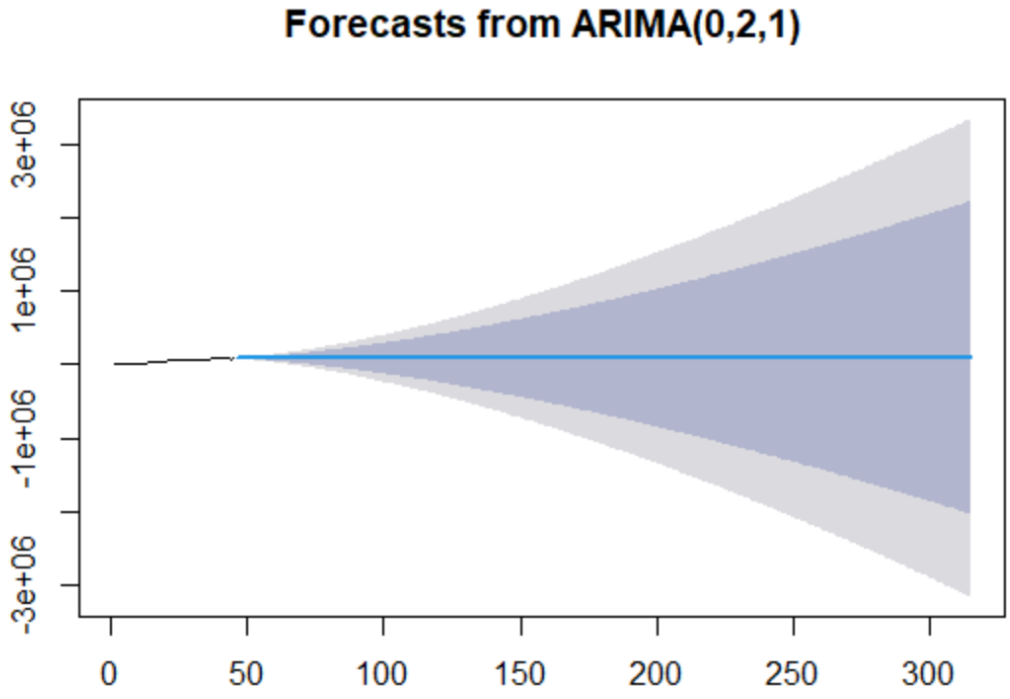
建模并总览模型:

```
> summary(model)
Series: confirmed_ts
ARIMA(0,2,1)

Coefficients:
          ma1
      -0.7129
s.e.   0.1081

sigma^2 estimated as 4944624:  log likelihood=-392.26
AIC=788.52   AICc=788.82   BIC=792.04

Training set error measures:
                    ME     RMSE      MAE       MPE     MAPE      MASE         ACF1
Training set -46.58912 2148.251 1052.958 1.676043 3.294831 0.5694465 -0.007100158
```

预测 10 天后



预测 270 天后:



10 天后的置信区间

```
> forecast_10$lower[10,2]
      95%
46467.45
> forecast_10$upper[10,2]
      95%
113600.9
```

可以从结果看出，10 和 270 天的预测基本水平，即控制住了疫情，但是置信区间特别大所以需要有更多的数据更准确的数据才能完成较准确的预测（数据越多越准确是数据分析，预测的王道）