**HW #2**

Team:

1. First name: Joanna Last name: Ge
2. First name: Betty Last name: Sun
3. First name: Xinhui Last name: Yu
4. First name: Stella Last name: Zhou

**Instructions:** submit your 1) this word file, please put your non-code answers below each question, and 2) python code. Each student should submit both word and python code files. Students in one team can submit identical files. Pleas list your teammates in both word file and python code file.

**Case 1: Covid impact on Air Passenger Demand**

**Motivation:**

The air transport industry relies heavily on forecasting air passenger demand for supporting management decisions on a variety of application areas: operation of new routes, closing routes, ticket price policies, management of mileage programs, acquisition of aircrafts, staff training programs, participation on alliances of strategic cooperation between airline companies, sufficiency of current infrastructure, the opportunity of building new terminals, etc.

**Task:**

You are hired as a business consultant to evaluate the impact of COVID the air transport industry.

**COVID period = March 2020 – March 2021**

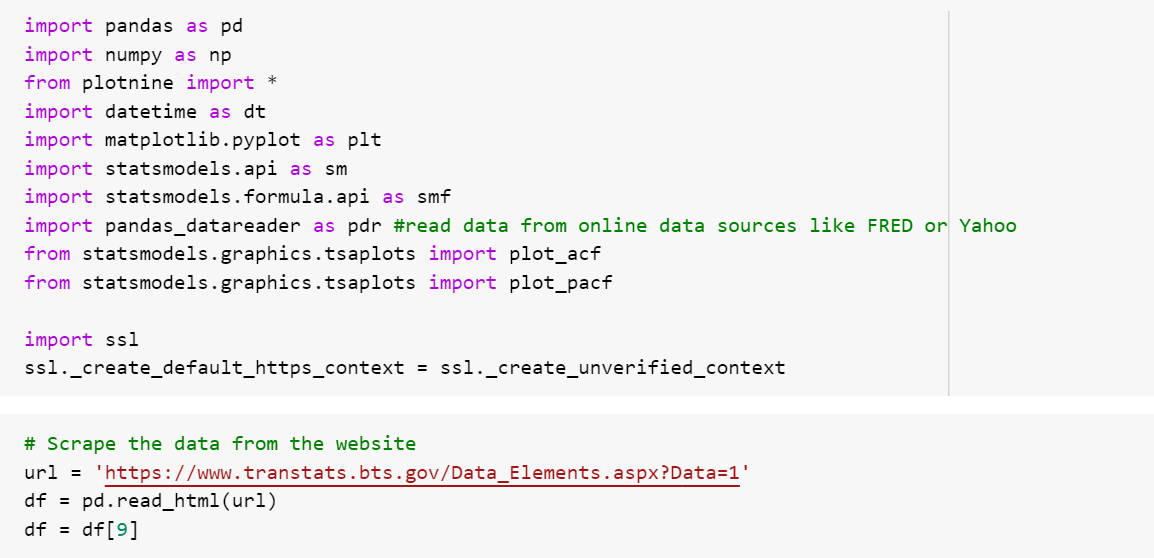
**Data:**

U.S. DOT’s Bureau of Transportation Statistics (BTS) maintains the passenger databases. Information on the passenger databases can be obtained from the BTS Office of Airline Information website at

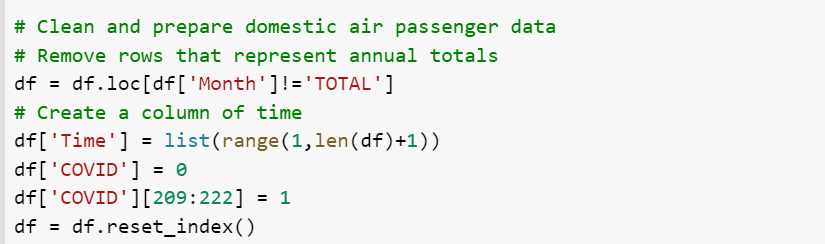
[**https://www.transtats.bts.gov/Data\_Elements.aspx?Data=1**](https://www.transtats.bts.gov/Data_Elements.aspx?Data=1)

**Data Preparation:**

1. Gathering data: write code to scrape the data from the website.

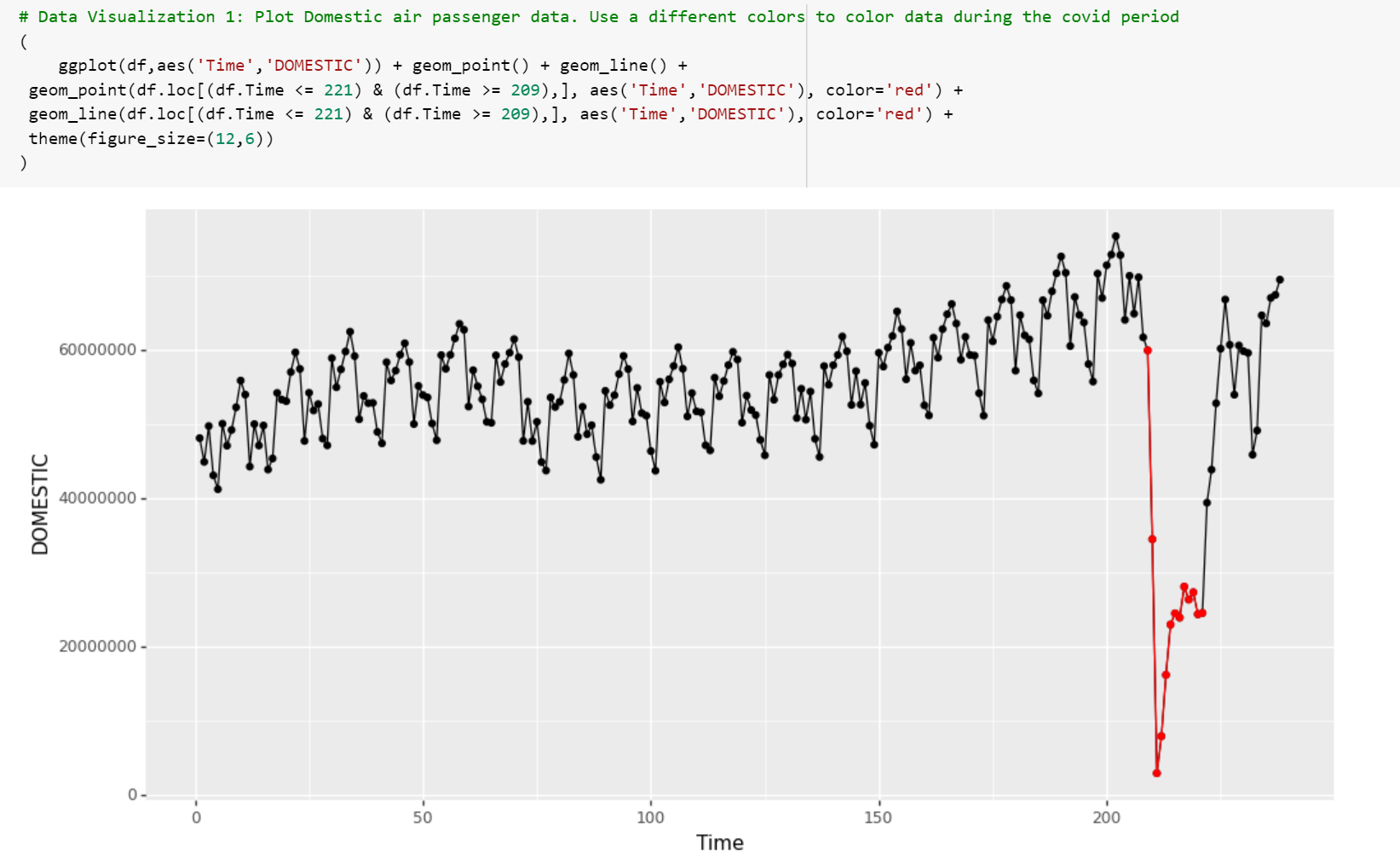


1. Cleaning and preparing DOMESTIC air passenger data for the analysis. For example: you may need to remove commas separating thousands, remove rows that represent annual totals.



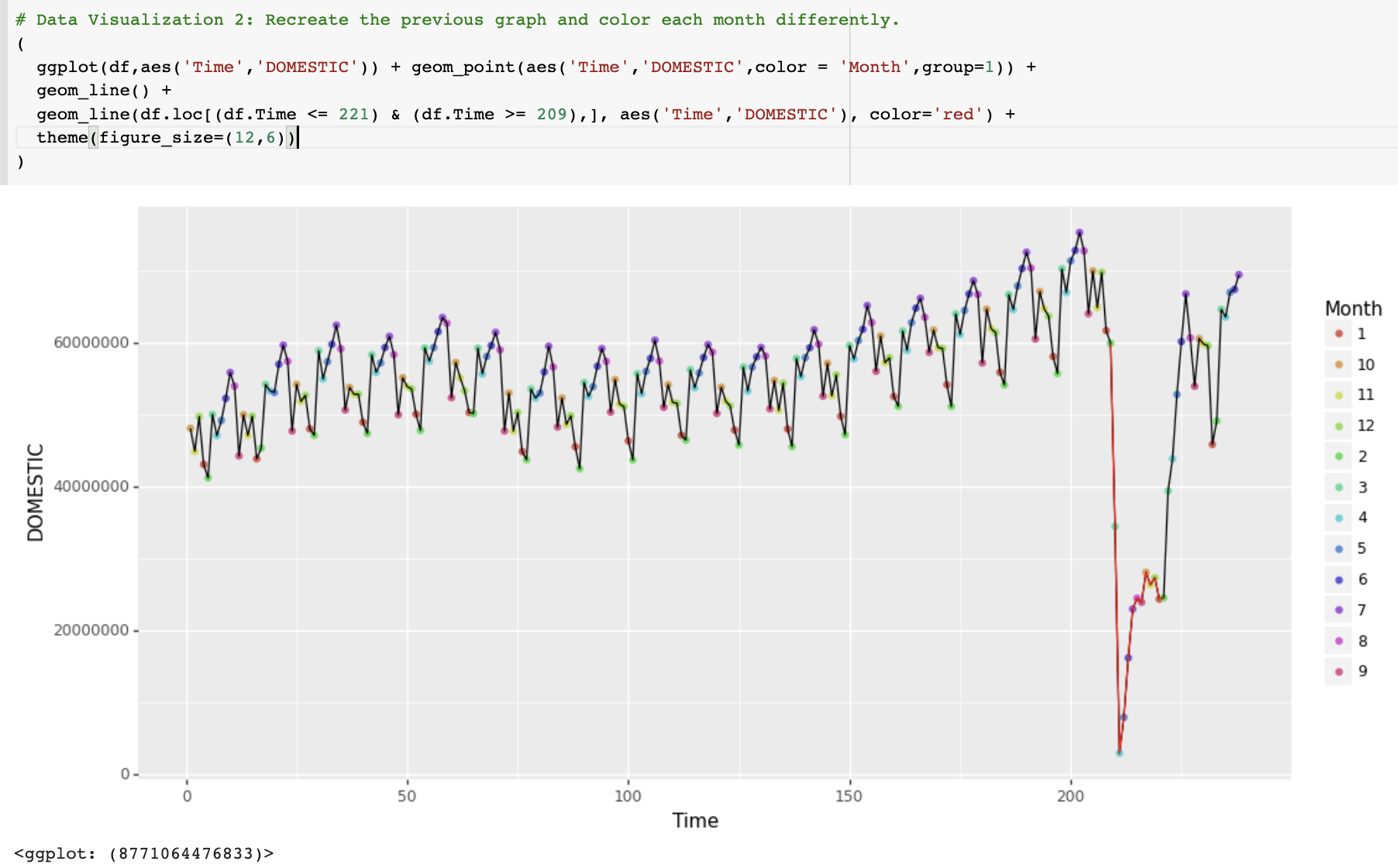
**Data Understanding:**

1. Data Visualization 1: Plot DOMESTIC air passenger data. Identify and interpret the patterns in data. Use a different colors to color data during the covid period.



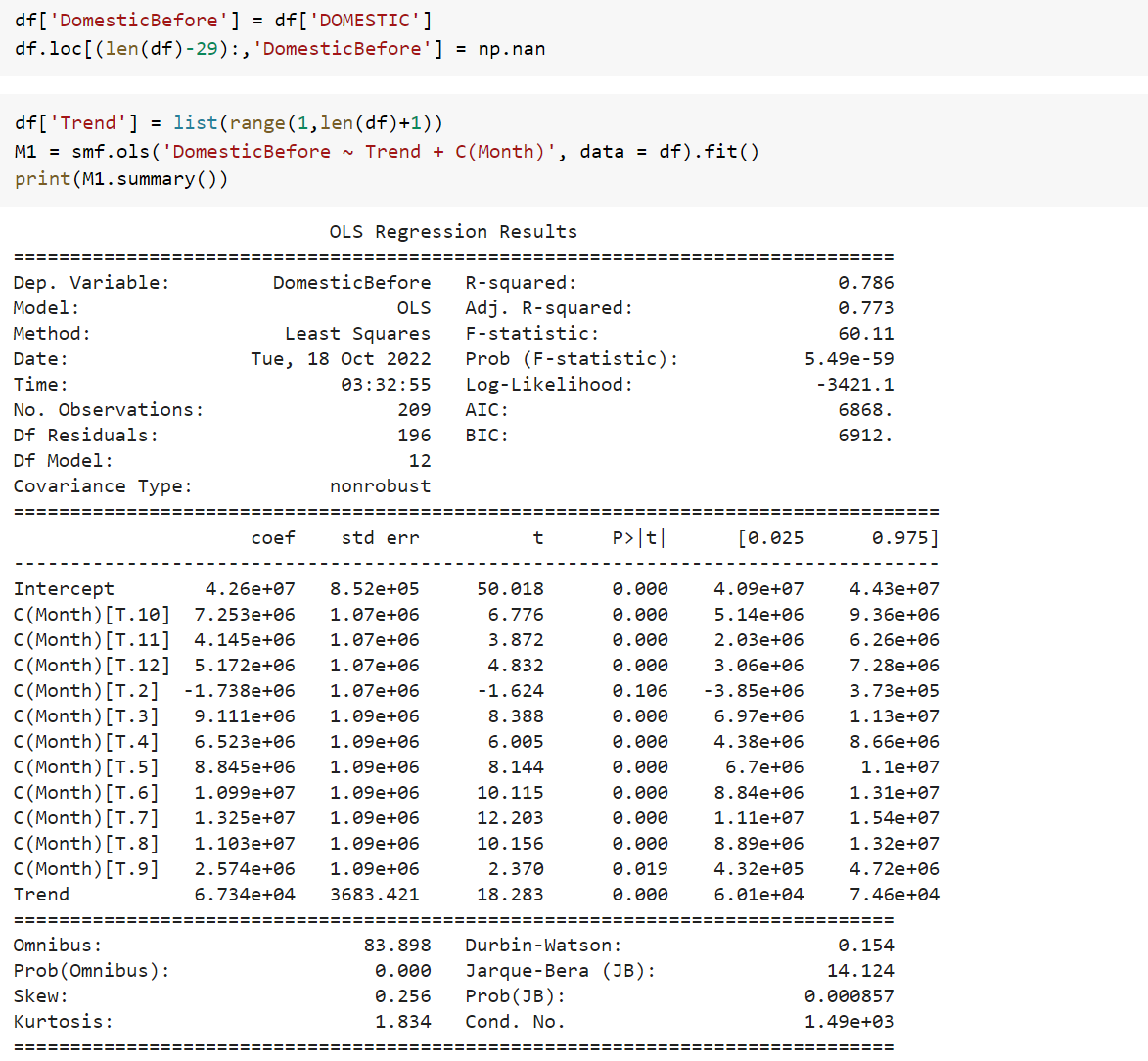
There was a sharp drop in the number of domestic air passengers during the covid, especially during the first 3 months of the pandemic. Then the number of domestic air passengers gradually increased with some small fluctuations, but still had not recovered to the pre-covid level.

1. Data Visualization 2: recreate the previous graph and color each month differently.



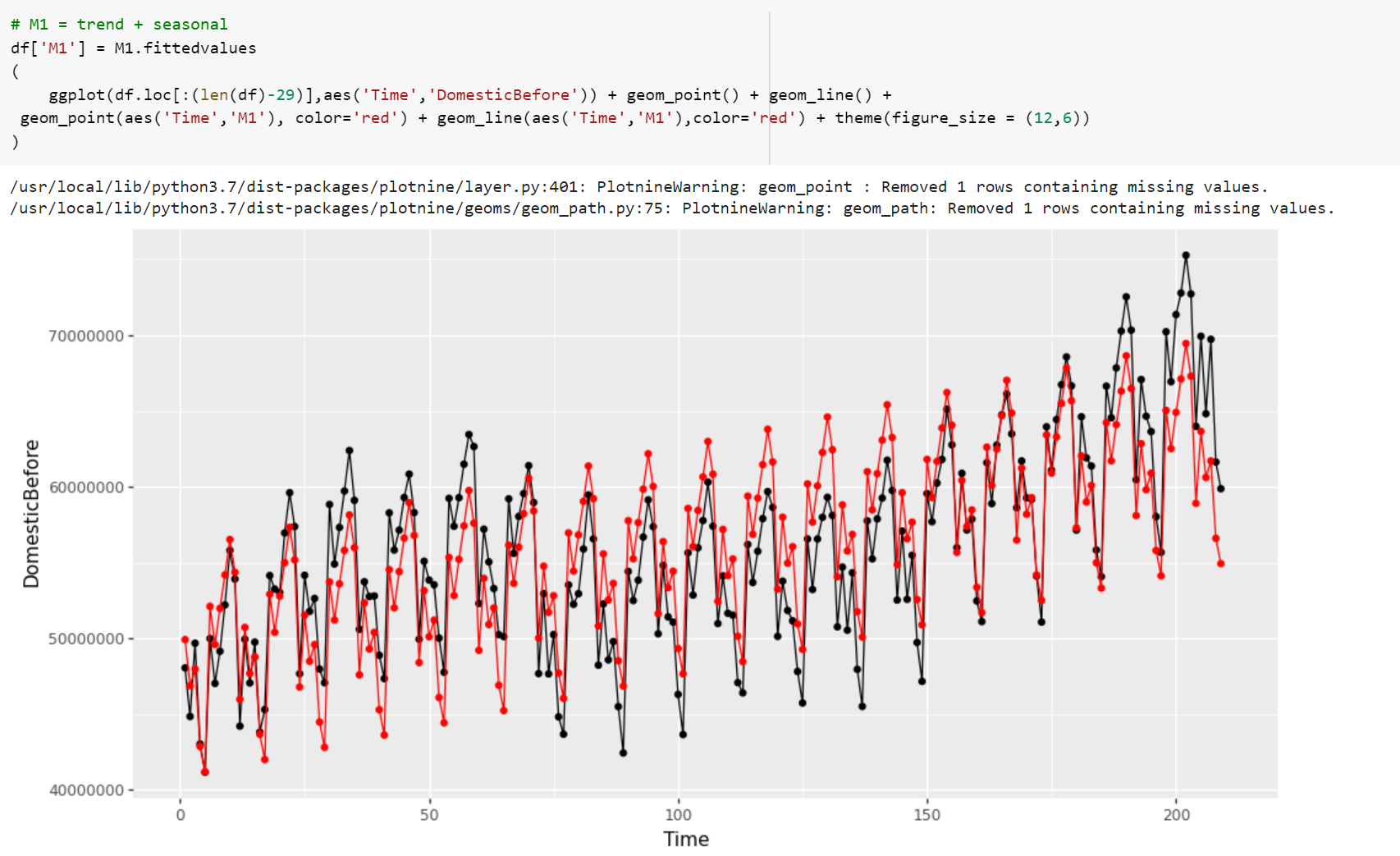
**Modeling:**

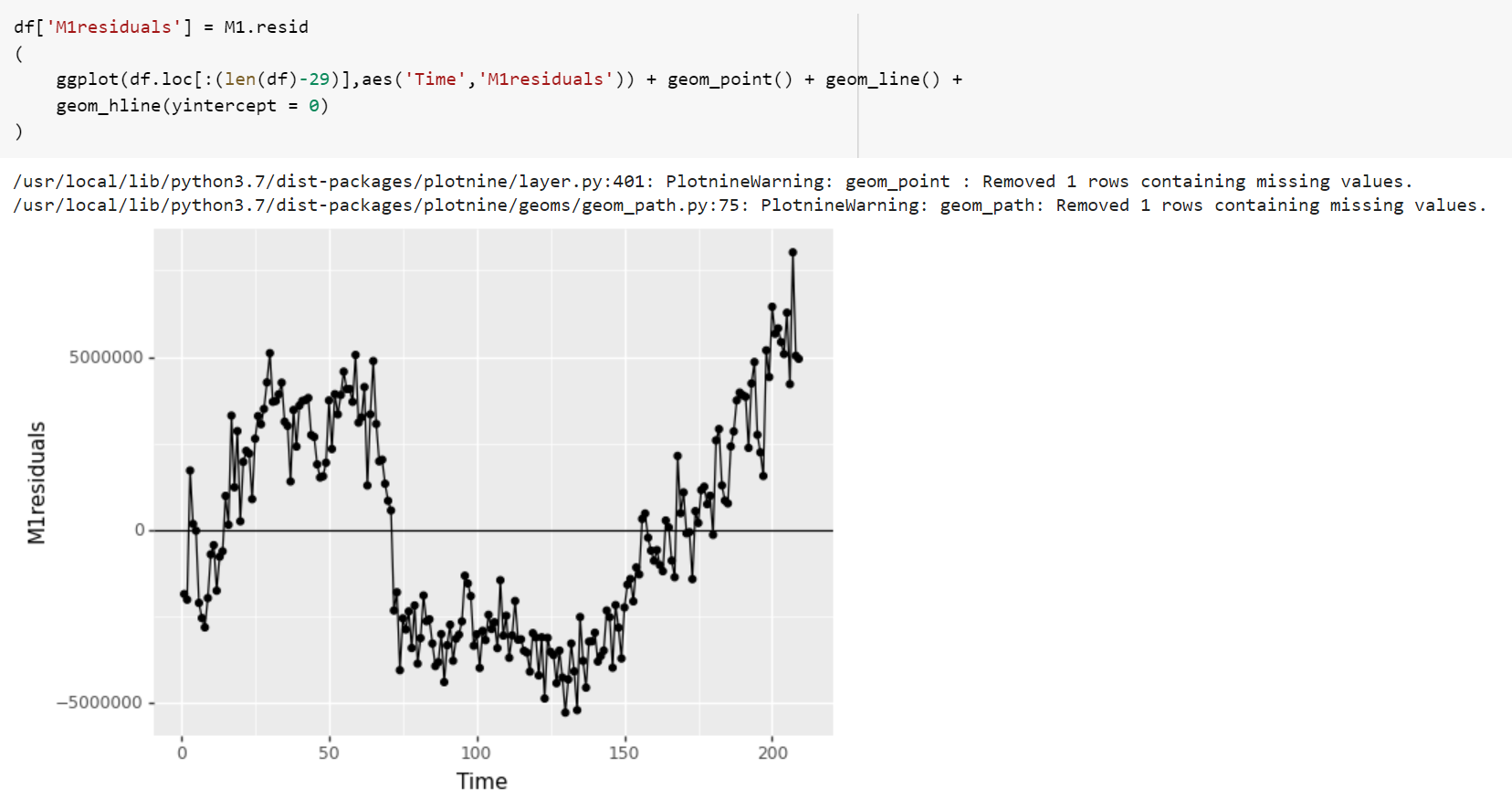
1. Build a model to estimate the impact of recession on DOMESTIC air passenger data. Present the model building steps and interpretations of the key outcomes of all of the important steps that help understand how you built a model (something similar to what we did for car sales).



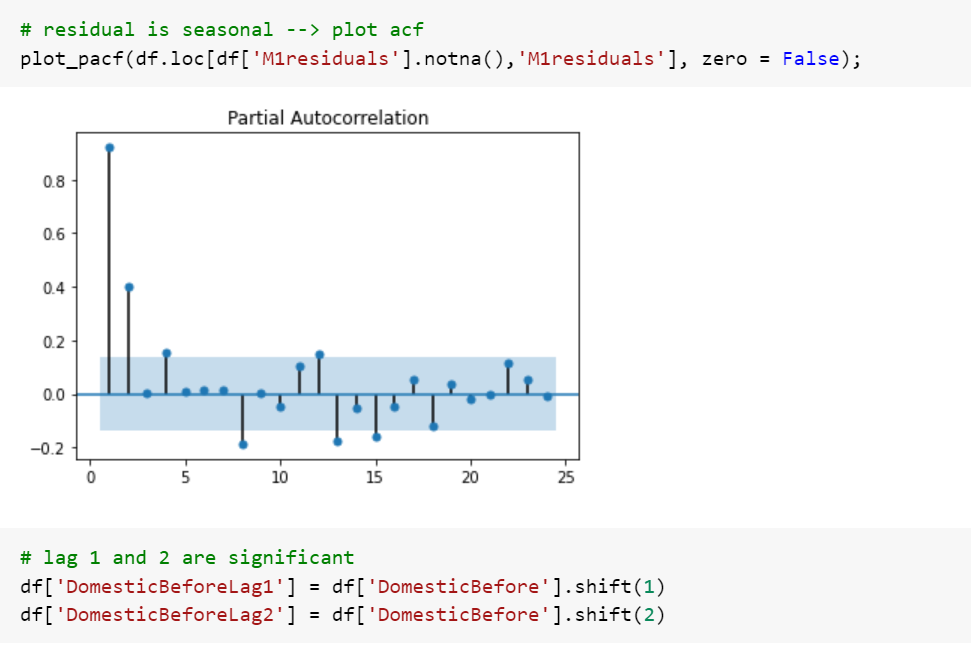
Interpretation of the regression coefficients:

* Intercept = 4.26e+07 = average passengers in M1 (M1 does not have a dummy)
* C(Month)[T.2] = -1.738e+06. On average passengers in M2 are smaller by 1.738e+06 than average passengers in month 1 -> average passengers in M2 = 4.26e+07 - 1.738e+06 = 4.086e+07
* C(Month)[T.3] = 9.111e+06. On average passengers in M3 are larger by 8.666e+05 than average passengers in month 1 -> average passengers in M3 = 4.26e+07 + 9.111e+06 = 5.171e+07
* ……
* Trend = 6.734e+04. On average passengers increase by 6.734e+04 per month.

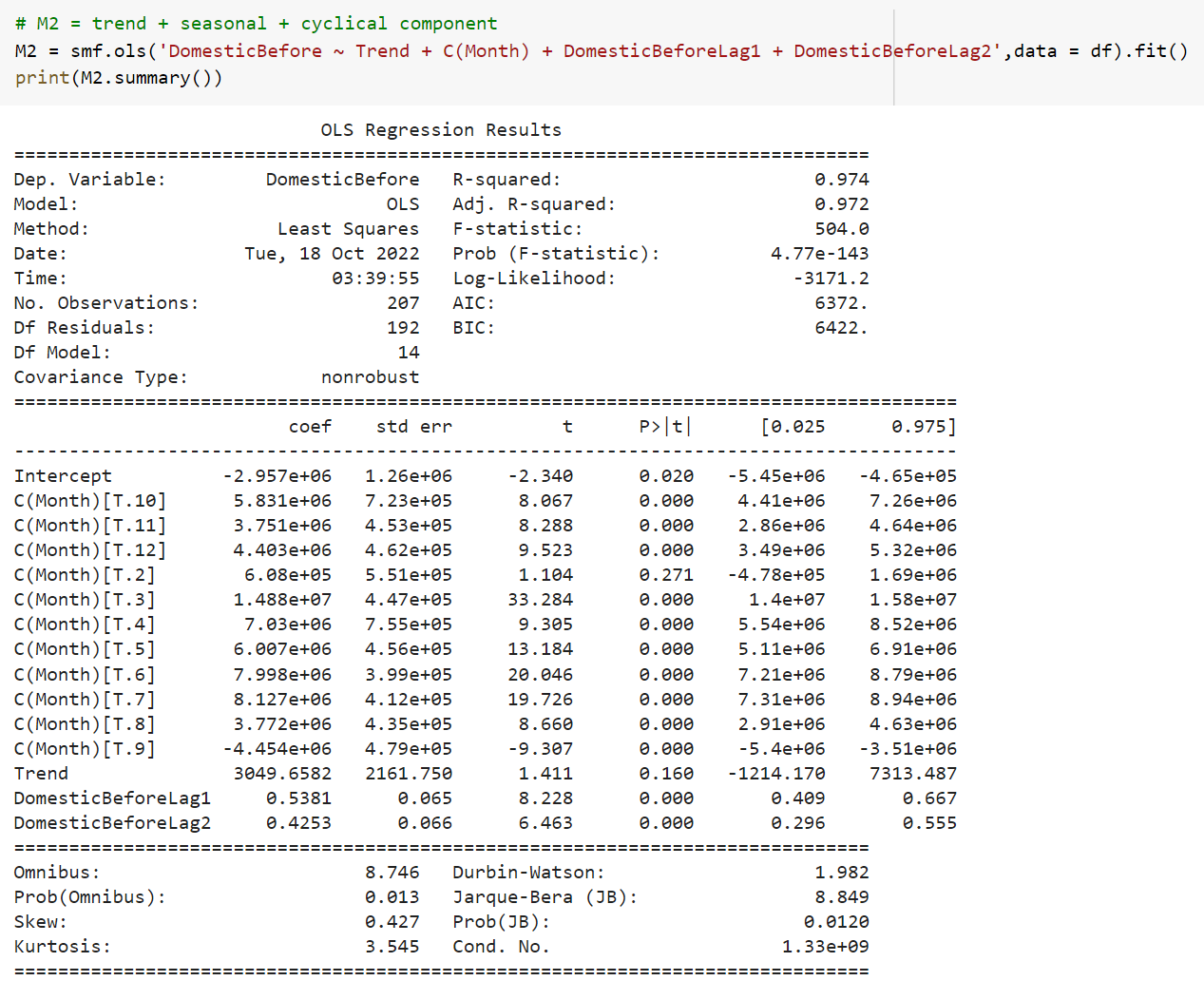




The graph of the residual is still not completely random noise and it seems to have a cyclical pattern. We need to use plot\_pacf to confirm if the residual is cyclical and if so, we need to create lags to include cyclical patterns into our predictions.

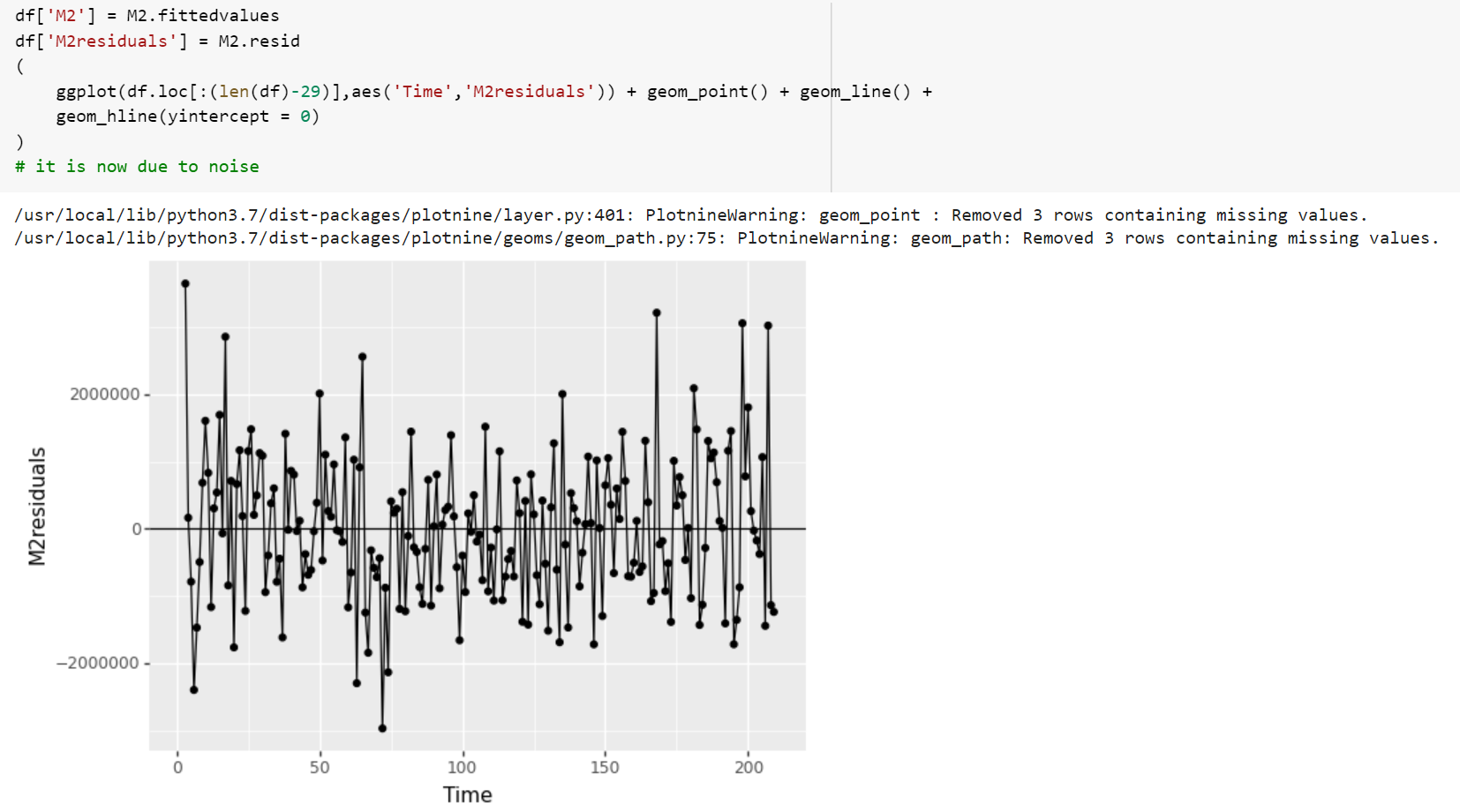


From the plot above, we can see that the first and the second lags are significant as they fall outside the 5% confidence interval. Thus, we will include the two lags in our prediction as cyclical components.

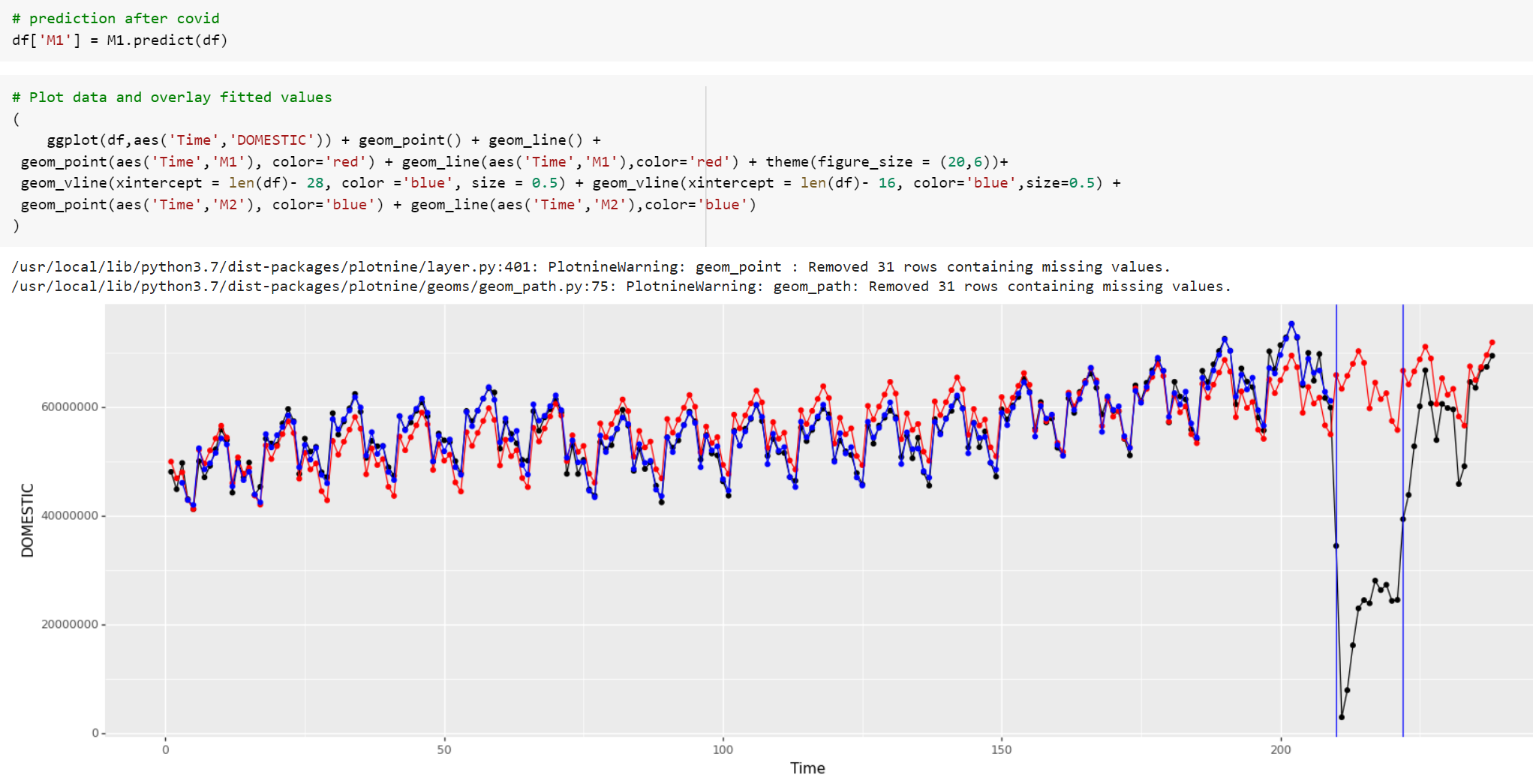


Interpretation of the regression coefficients:

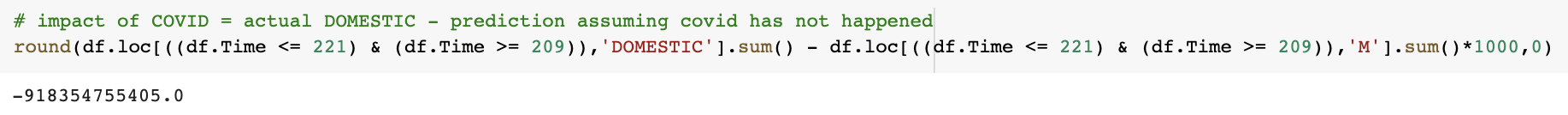
* Intercept = -2.957e+06 = average passengers in M1 (M1 does not have a dummy)
* C(Month)[T.2] = -6.08e+05. On average passengers in M2 are smaller by 6.08e+05 than average passengers in month 1 -> average passengers in M2 = -2.957e+06 - 6.08e+05 = 3.565e+06
* C(Month)[T.3] = 1.488e+07. On average passengers in M3 are larger by 8.666e+05 than average passengers in month 1 -> average passengers in M3 = 1.488e+07 - 2.957e+06 = 1.1923e+07
* ……
* Trend = 3049.6582. On average passengers increase by 3049.6582 per month.



Now the graph of M2residuals looks like random noise after we exclude the cyclical component, which indicates that we have built a good model.

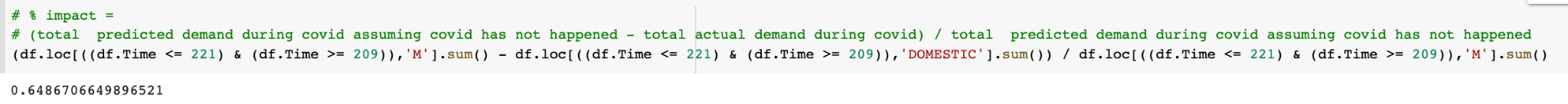


1. What is the (total) impact of COVID on DOMESTIC air passenger data?



During the pandemic, the number of domestic air passengers dropped by 918354755405.

1. Represent your estimated impact on the percentage scale.



On the percentage scale, covid made the number of domestic air passengers dropped by 64.9%.

1. Has the industry recovered? If yes, when?

The industry has partially recovered but nowhere near its full potential if covid hadn’t happened. Starting from approximately April 2021, the number of domestic air passengers have been climbing drastically until around July or August 2021 where it reaches the point from which the seasonal patterns begins to resume. However, in term of the overall trend, or the absolute number of passengers, the industry has not recovered to the pre-covid level since the number of passengers in the post-covid period is still considerably less than that of our forecast.

1. Using the following scenarios: assess whether your model that you built overfits the data? (cross-validation for time series). Briefly explain your approach. RMSE and MAPE may be helpful.   
   Are all scenarios appropriate for assessing whether regression overfits the data.   
   **Hint 1**: one way to assess overfitting is by comparing model’s performance on training and testing sets: if the error on the testing set is much larger than on the training set – that’s an indication of overfitting.   
   **Hint 2**: for each scenario built a new model: build a model using training set only!

**Scenario 1:**

Training set: start = 10/2002, end=12/2019

Testing set: start = 01/2020, end=12/2020

**Scenario 2:**

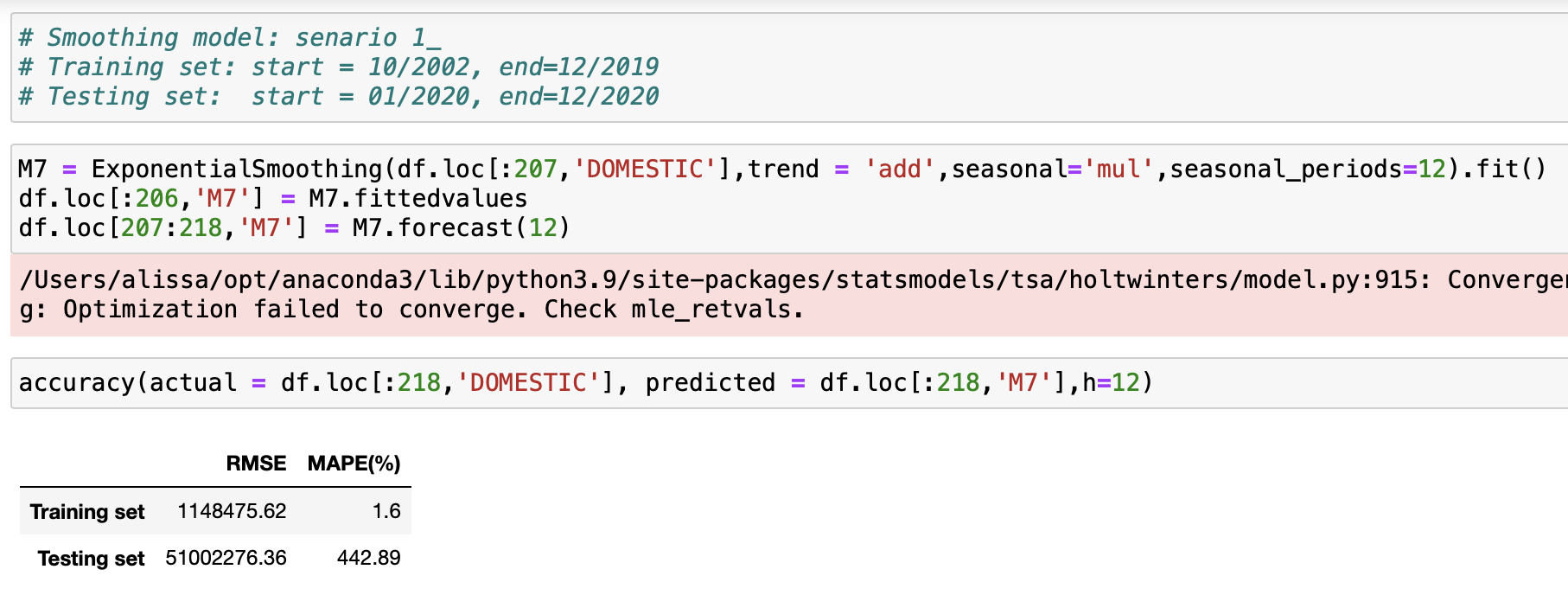
Training set: start = 10/2002, end=12/2018

Testing set: start = 01/2019, end=12/2019

**Scenario 3:**

Training set: start = 10/2002, end=12/2017

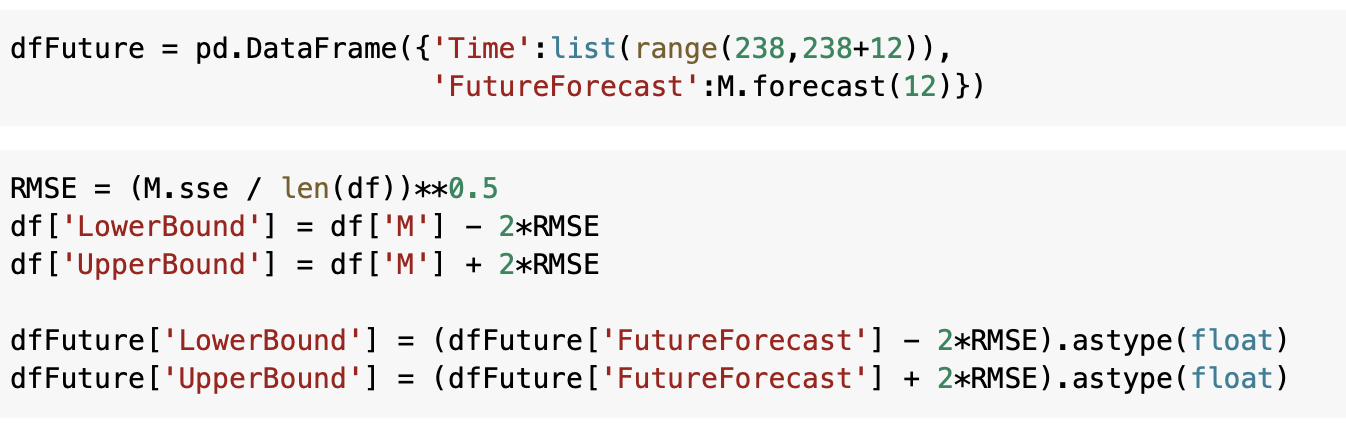
Testing set: start = 01/2018, end=12/2018

****

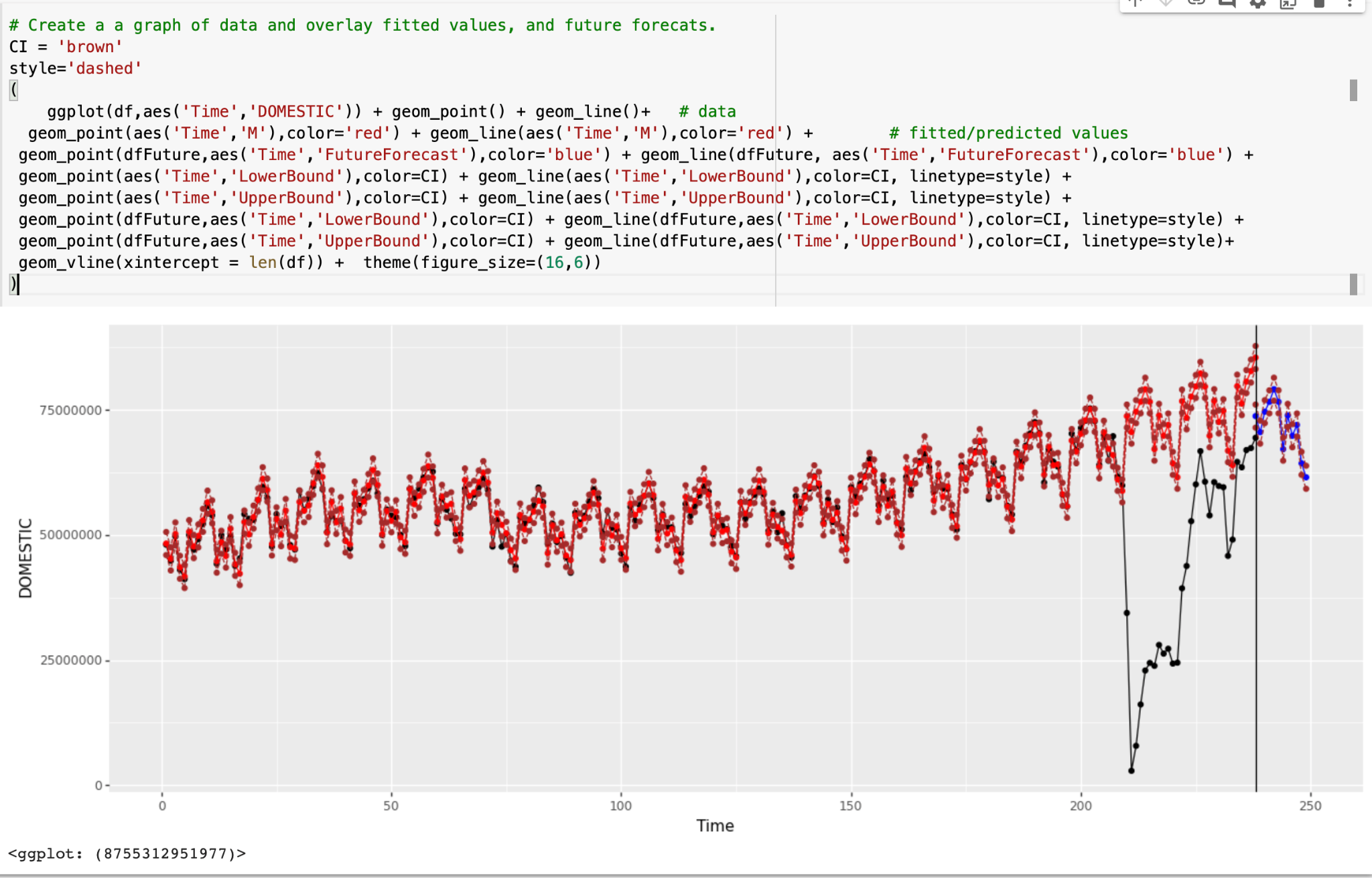
****

We built a smoothing model. Our analysis reflects that scenario 3 most effectively avoids the overfitting problem. In the smoothing model of scenario 3, both RMSE and MAPE for the testing set are less than those of the training set, which indicates that the model does not overfit the data. Therefore, the comparative analysis shows that scenario 3 best avoids the overfitting problem.

1. Build a model using all available data and predict next 12 months ofDOMESTIC air passenger demand.  
   **Hint**: you will need to use all available data to build a model.



1. Plot data and overlay fitted values, future forecast and confidence/prediction intervals on one graph.   
   **Hint**: use .conf\_int()   
   https://www.statsmodels.org/dev/generated/statsmodels.regression.linear\_model.PredictionResults.conf\_int.html



**Evaluation:**

1. Do you consider that your modeling approach presents an accurate picture of current and future demand for air travel? Does your model need to be improved? Briefly mention what situations could potentially distort predictions, cause challenges in finding reliable and accurate predictive models to forecast air passenger demand and discuss the ways to overcome these challenges. Exploring airlines’ websites might be helpful.

We find that the modeling does not necessarily present an accurate picture especially of the future demand for air travel in the post-covid period. The model forecasts the future trend purly from a statistic standpoint, where it depends its future value on the most recent previous values (due to the two most recent lags that we created). However, this could distort the predicted data immediately following the covid period as they are affected by the data during covid, which are abnormally low. As the data during covid is not part of a sustainable trend but rather only the result of a disruptive economic period, the weight we assign to the data during covid should be evaluated to be different from that of other non-covid periods. Moreover, a better modeling should take into consideration the social factors and economic incentives as well. For example, a large number of airline companies are taking an aggressive move in offering discounts and mileage benefits in order to encourage more travellers. Data has also shown that people tend to intentionally seek more travel after covid because of the confinement during covid. Those are the social factors that our model also needs to incorporate to improve its practicability.