TIME SERIES for FORECASTING

Khobie Maseko
31/12/2021

Project Background and Objective

- This data was obtained from a dataset in Kaggle titled “airlines-passenger-data”.
- Dateset source: https://www.kaggle.com/ternaryrealm/airlines-passenger-data?select=international-airline-passengers.csv.
- The dataset contains a time series data set of international airline passengers; monthly totals in thousands. Jan 1949 – Dec 1960.
- The main **objective** of the project is to build a time series model that will be able to forecast the number of airline passengers for the next 12 months. We will be using the LSTM method to find the most accurate model.
- We have to note that this dataset consists of two features: ‘Month’ and ‘International airline passengers: monthly totals in thousands. Jan 49 ? Dec 60’.

Project Workflow

The following are the steps that we will take to determine which model is the one best suited to meeting our objective for this dataset:

1. Data Overview
2. Data Preprocessing
3. LTSM Model 1
4. LTSM Model 2
5. LTSM Model 3
6. Time Series Analysis (with predicted results)
7. Key Findings and Insights
8. Recommendations
9. Conclusion

Data Overview - Data Attributes

Below are the attributes of the dataset as retrieved from Kaggle.

It has the previously mentioned two columns.

We also determined that the last row contained a ‘NaN’ value.

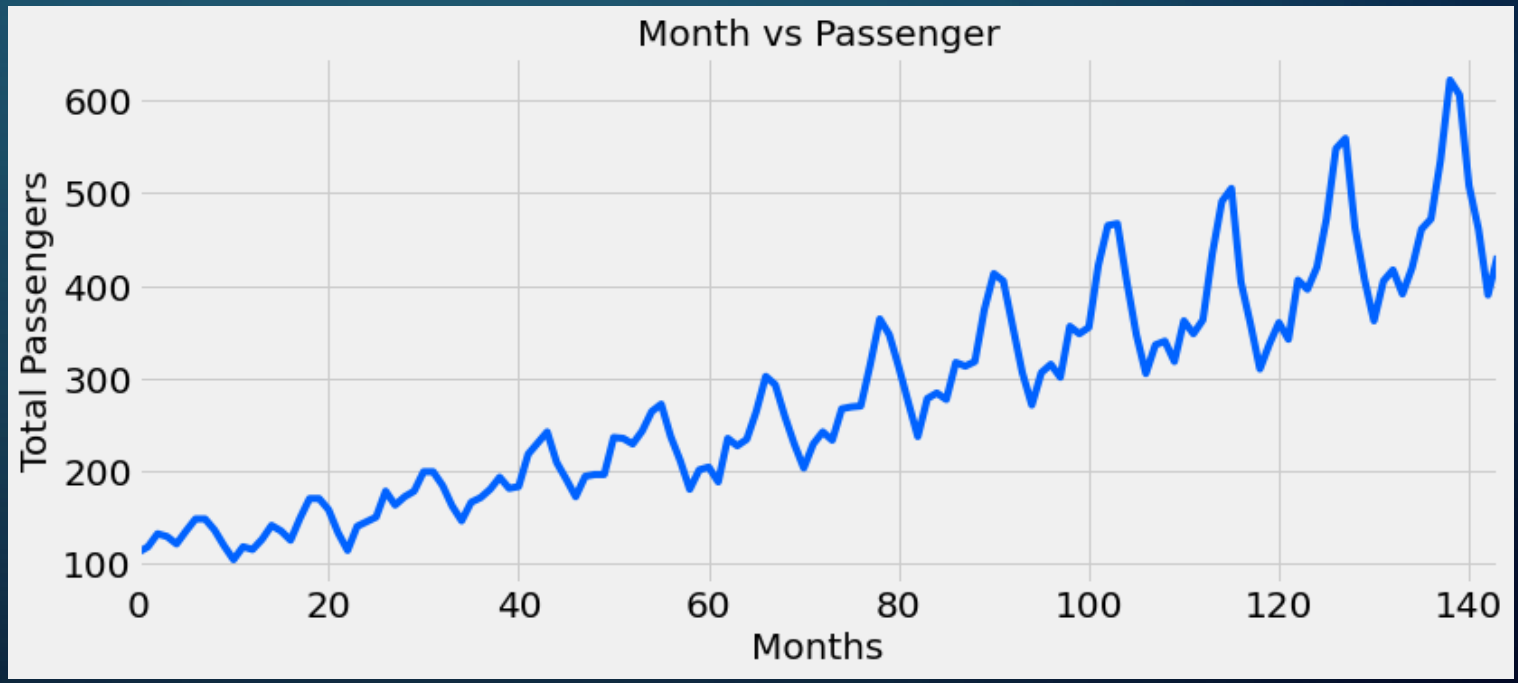
	Month	International airline passengers: monthly totals in thousands. Jan 49 ? Dec 60
0	1949-01	112.0
1	1949-02	118.0
2	1949-03	132.0
3	1949-04	129.0
4	1949-05	121.0

Data Overview - Visualization

We did a preliminary visualization of the data before we worked on it.

The plot shows us that the data has a strong upward trend from year to year.

There is also a seasonality that is explained by the increase of the number of passengers in the summer periods and decrease in the winter periods.



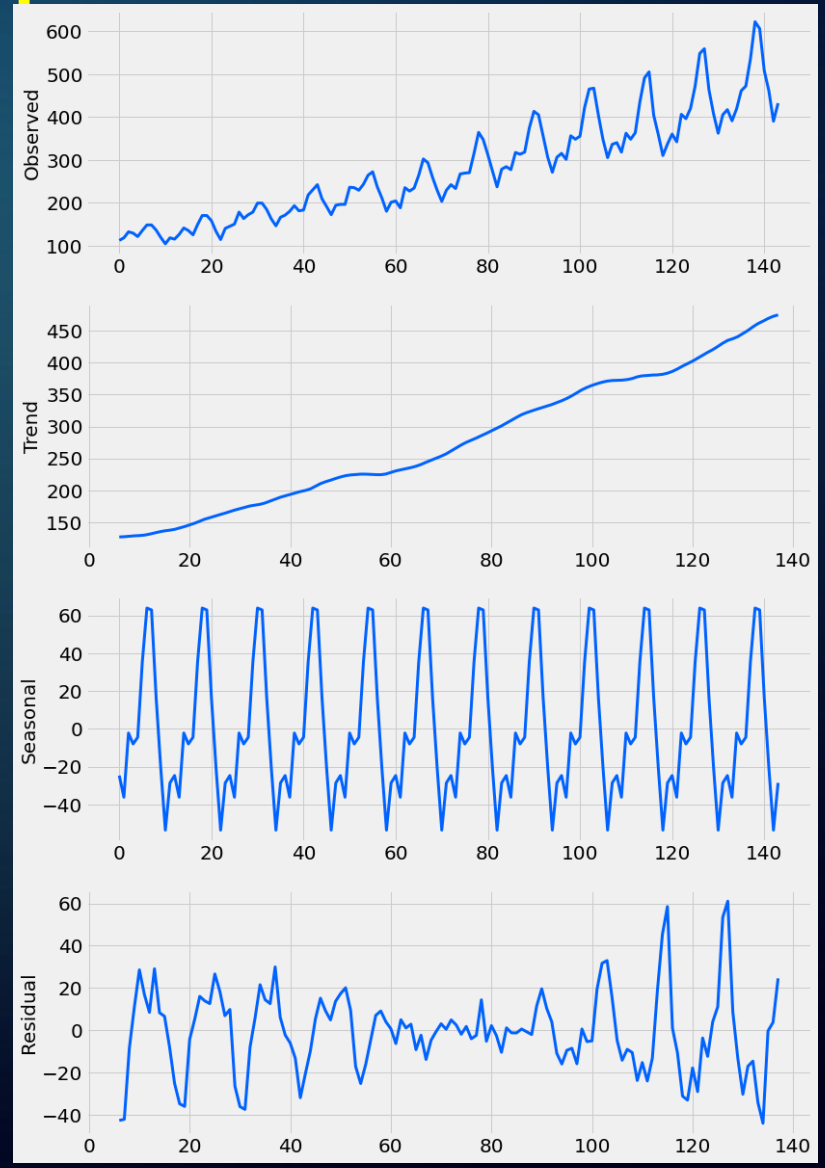
Data Overview - Decomposition

To further preview the data for our model designs, we did the following:

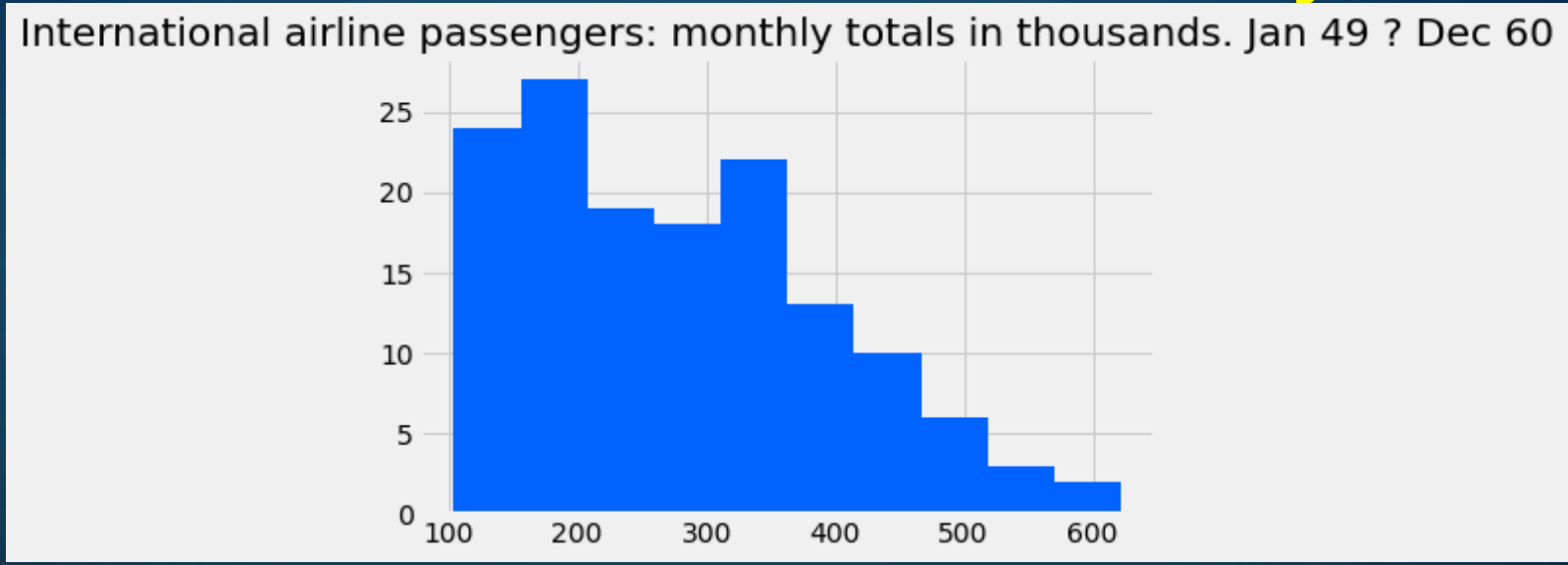
- dropped the row that had a NaN value. We noticed in data description.
- decomposed the data into Observed, Trend, Seasonality, and Residual through time series decomposition for clearification.

We noted that there was seasonality in the data.

We can also see that the number of customers steadily increases on the Trend plot.



Data Overview - Stationarity



We plotted a histogram and noted that the data is abnormally distributed, making it likely that is non-stationary.

This could have a huge impact on our model design and training.

We also calculated the p-value of the data and it was: 0.9918802434376409.

That is simply too high, so that confirmed that our dataset is non-stationary.

Data Preprocessing - Decomposition

To further prepare the data for our model designs, we did the following:

- split data into Train and Test sets for model design.
- performed Min_Max scaling on the data.
- created a create_inout_sequences function that transformed the raw input data into sequence data to fit training.
- modelled the data using a LTSM layer.

```
[(tensor([-0.9648, -0.9385, -0.8769, -0.8901, -0.9253, -0.8637, -0.8066, -0.8066,
          -0.8593, -0.9341, -1.0000, -0.9385]),
  tensor([-0.9516])),
 (tensor([-0.9385, -0.8769, -0.8901, -0.9253, -0.8637, -0.8066, -0.8066, -0.8593,
          -0.9341, -1.0000, -0.9385, -0.9516]),
  tensor([-0.9033])),
 (tensor([-0.8769, -0.8901, -0.9253, -0.8637, -0.8066, -0.8066, -0.8593, -0.9341,
          -1.0000, -0.9385, -0.9516, -0.9033]),
  tensor([-0.8374])),
 (tensor([-0.8901, -0.9253, -0.8637, -0.8066, -0.8066, -0.8593, -0.9341, -1.0000,
          -0.9385, -0.9516, -0.9033, -0.8374]),
  tensor([-0.8637])),
 (tensor([-0.9253, -0.8637, -0.8066, -0.8066, -0.8593, -0.9341, -1.0000, -0.9385,
          -0.9516, -0.9033, -0.8374, -0.8637]),
  tensor([-0.9077]))]
```

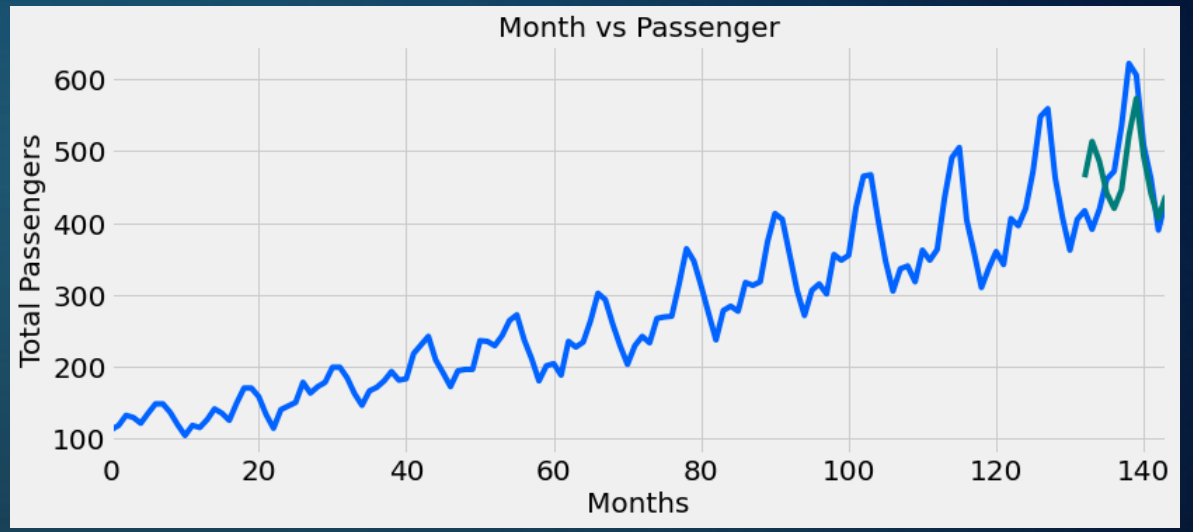
LTSM Model 1

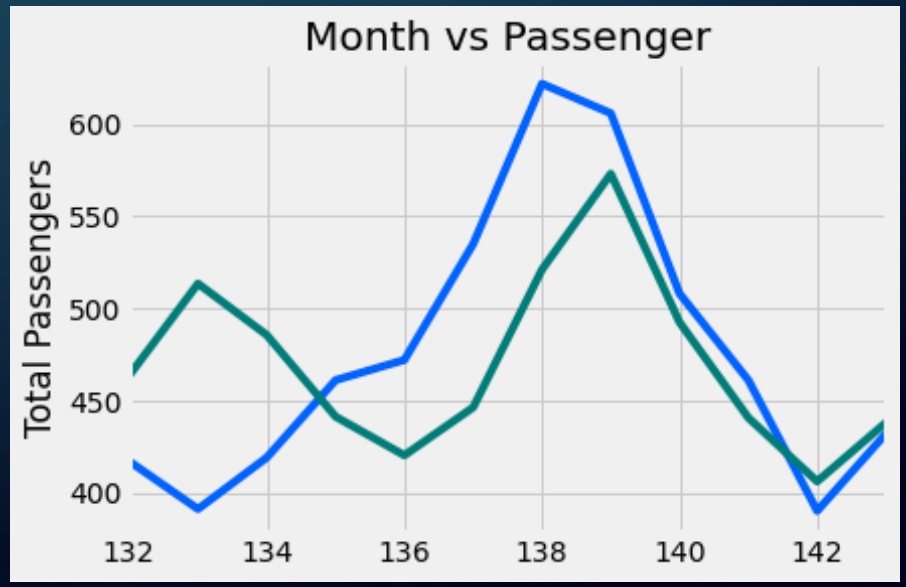
We:

- set the model to run for 500 epochs.
- set the model to have 2 layers.
- set the model to have 128 hidden layers.
- noted that the results are relatively close.

The bottom graph gives us a closer look at the results

It was possible to spot an upward trend based on fluctuations in the total number of passengers traveling over the past 12 months.





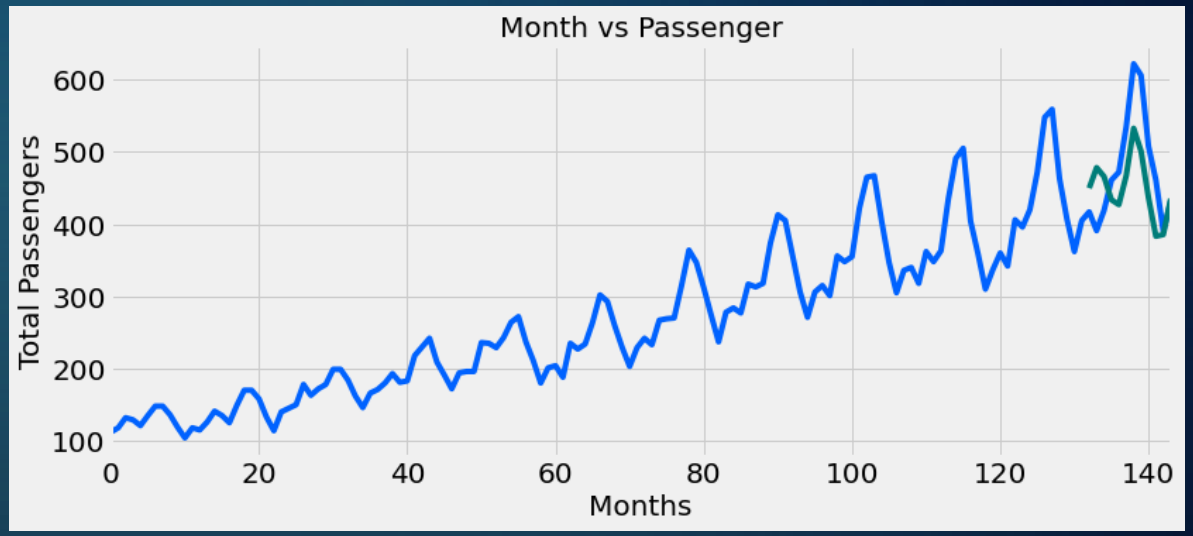
LTSM Model 2

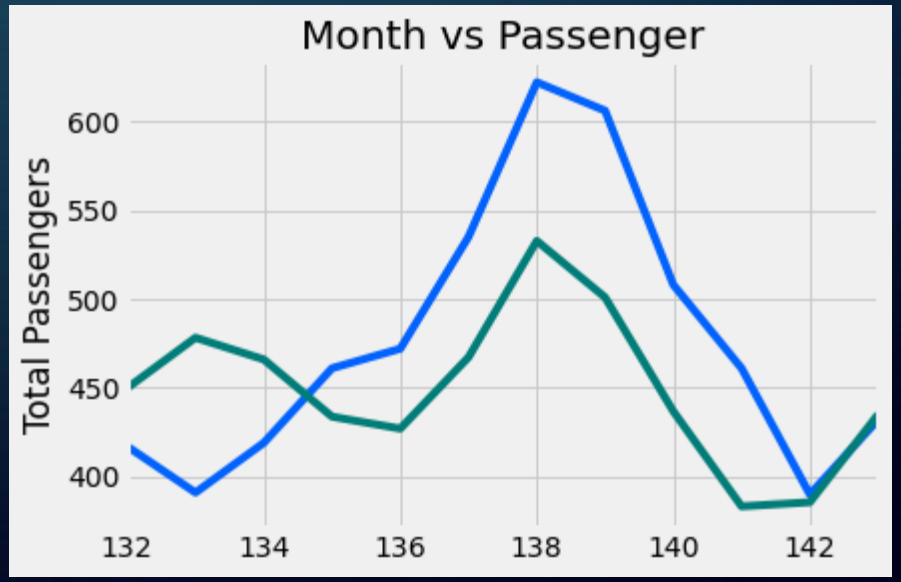
We:

- set the model to run for 750 epochs.
- set the model to have 2 layers.
- set the model to have 128 hidden layers.
- noted that the results are not as close as model 1.

The bottom graph gives us a closer look at the results.

It definitely performed worse than the first model but it wasn't too far off, all things considered.





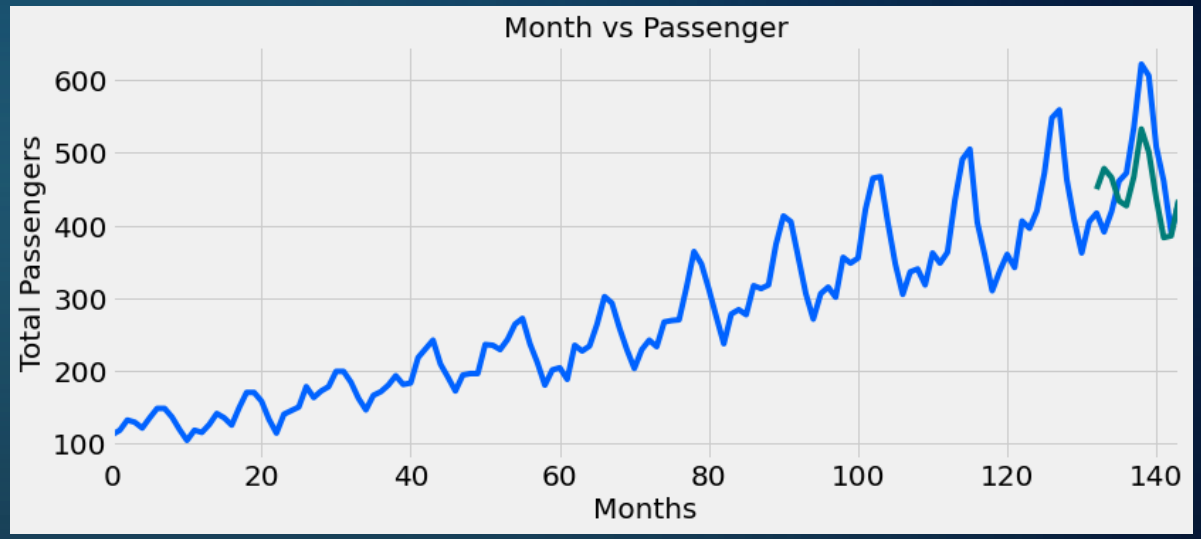
LTSM Model 3

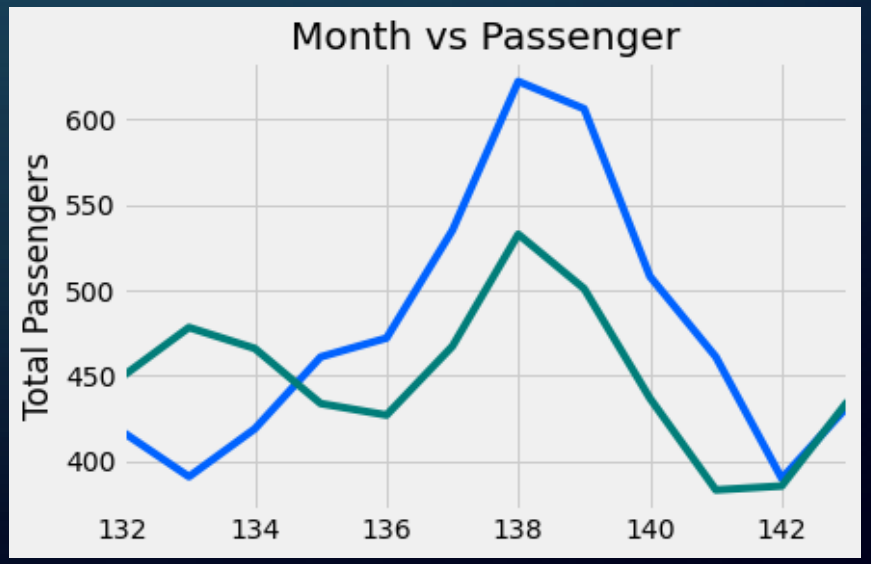
We:

- set the model to run for 1000 epochs.
- set the model to have 2 layers.
- set the model to have 128 hidden layers.
- noted that the results are relatively close.

The bottom graph gives us a closer look at the results

It definitely performed worse than the first two models.



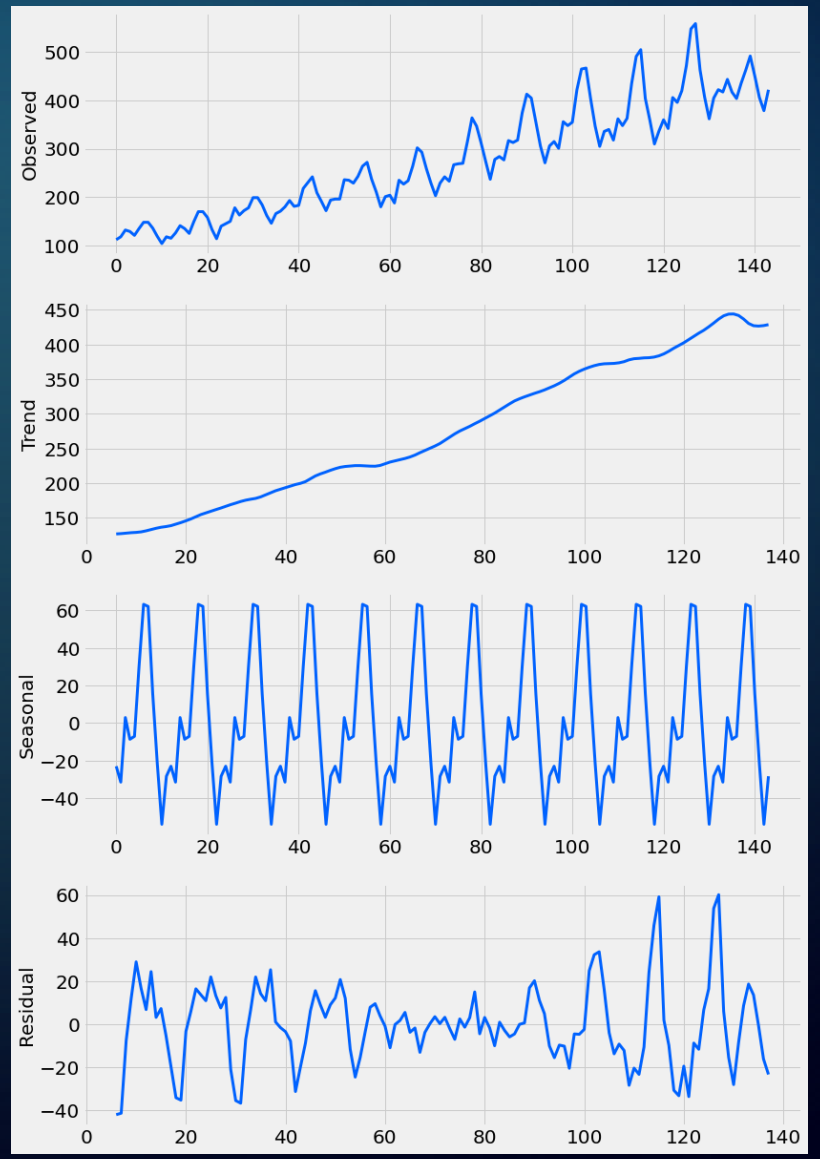


Time Series Analysis (with predicted results)

See plotted decomposition graphs on the right.

We checked whether the predicted results were learned while preserving the trend, seasonality, and residual of the timer series that the original time series had.

We noted that, despite the simple model, it was predicted while well preserving the trend, seasonality, and residual.



Key Findings and Insights

We noted that Model 1 has the least number of epochs undertaken by LSTMs and was the best performer out of the three.

In general, the predictions of all 3 models are not accurate enough to build a business model on in their current forms, but they are good enough to understand airline market trends at least.

Recommendations

I would recommend the following:

- using methods such as smoothing, differencing, etc. to ensure the stationarity of our dataset. Stationary data is best for forecasting purposes.
- additional fine-tuning of the hyperparameters of the models that we will use in future in areas such as hidden layers, epochs and itemizer(s).
- using validation set with EarlyStopping callback to avoid overfitting.

Conclusion

We may conclude that LSTM Model 1 is the best performing one the three that we designed. I believe that if the aforementioned recommendations were to be implemented, we could achieve much better results.