**HW #3**

**Instructions: submit your R or python code. Include your answers in your R or python code. One submission per team, please include the names of your teammates.**

**Business Understanding | Data Understanding | Data Preparation | Modeling | Evaluation | Deployment**

**Case 1: Forecasting Air Passenger Demand**

**Motivation:** Why air passenger demand forecast is important?

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 build a predictive model to forecast future passenger demand for **next 12 months.**

**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.
2. Cleaning and preparing data for analysis: write code to convert columns DOMESTIC and INTERNATIONAL to numeric format.

**Data Understanding:**

1. Data Visualization 1: Plot both columns on separate graphs. Identify and interpret the patterns in data.
2. Data Visualization 2: recreate previous graphs and use 12 colors representing 12 months to color all points.
3. Data Visualization 3 (Recreate the graph): Plot both trajectories on the same graph: use 2 vertical axis: the left one for DOMESTIC and the right one for INTERNATIONAL. Note that that both axes are in mln (you can divide data by 1 mln). Also add a range selector at the bottom of the graph.

Instructions: to recreate this graph you will need: 1. a new time series visualization library called **dygraphs** and the manual contains examples: <https://rstudio.github.io/dygraphs/index.html>

**Hint:** start with creating a time series object that contains not one, but two time series (similar to lungDeaths in the dygraphs manual):



Graphical user interface, application

Description automatically generated

**Modeling**

1. To ensure that your modeling and decision making process and recommendations based on your models are robust and credible in a range of possible future outcomes, you will apply the validation approach three times (see Scenario 1 - Scenario 3 described below) and evaluate models’ predictive accuracy using validation MAPE and RMSE for both scenarios. Then you will average MAPE’s from both scenarios for each model and obtain the average MAPE for each model.

**Scenario 1:**

Training set: start = 10/2002, end=12/2007

Testing set: start = 01/2008, end=12/2008

**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=12/2019

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

After you build all models create a table in R that contains the following information:



**Clarification:** You can create as many models (regression(required for this HW), auto.arima, etc) as you want. Simple or not accurate models are not bad! They can serve as baseline models and can help you understand better more accurate models and can help you quantify the improvement. Feel free to automate your code.

**Evaluation:**

1. For each scenario and each model: 1. create a graph of training and testing data and overlay fitted/predicted values and forecast on testing set; 2. the graph include the graph of residuals and 3. The Acf graph of residuals. Briefly comment whether each model is adequate or no.

If you want to challenge yourself you may think how to create a graph of residuals below the data so that you can better see the time periods in which the model does not capture the patterns data accurately. The graph below illustrates the idea. It is fine if you create separate graphs.

Chart, line chart

Description automatically generated

**Deployment:**

1. The model that has smallest average MAPE is the champion model. You will need to use the champion model to forecast next 12 months air passenger demand. Whatever forecasting approach is used to estimate future patterns of demand for air travel, the results will inevitably be subject to significant uncertainty. An understanding of how this uncertainty could affect any decisions is therefore about as important as the forecast itself. Thus report both future forecast and corresponding 95% prediction interval in Table 2 in the answer sheet. (Do not include decimal digits!)

If you are using predict function, you can get by including interval = "prediction” in predict:

predict(Model, newdata, interval = "prediction)

Create and present the results the following table in R:



**Evaluation:**

1. For champion model (Q7): plot all historical data, future forecasts with prediction intervals, and fitted values. **Save your graph in a pdf or jpeg format and submit it on blackboard.**
2. 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? Exploring airline’s website (history of a company, important events, etc) might be helpful.
3. 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.

**Evaluation (Recession impact):**

1. The Covid-19 pandemic has caused the worst financial crisis in the history of the airline industry. Using this data quantify the impact of recession on the airline industry. Based on data, has the passenger traffic recovered yet? Assume covid started in 2020.

**Case 2: Covid impact on Adidas and Nike**

**Data**: **Adidas.xlsx** and **Nike.xlsx** contain historical quarterly income statement information for two popular sportswear companies: Nike (US) and Adidas (German). The data was downloaded from S& P Capital IQ (available for free of charge for USC faculty and staff).

Evaluate covid impact on Adidas and Nike gross profit or Normalized Diluted EPS (see income statement) (**pick** whether you work with profit or EPS). If you are not familiar with Earnings per share, please glance over the following article:

<https://www.investopedia.com/terms/d/diluted-normalized-earnings-per-share.asp>

In your code file briefly describe your modeling approach and the results.

Clarification: you can load your data from specific areas of excel files using read\_xlsx from readxl package and specify range.