CSE 544 Project, Spring 2025
Due May 20, 2pm via Brightspace

In this project, you will analyse the flight ticket prices data from `https://www.kaggle.com/datasets/dilwong/flightprices/data`. This dataset is a CSV file where each row is a purchasable ticket found on Expedia between 2022-04-16 and 2022-10-05. The data contains detailed information about flights including distances, durations, airlines, airplane types, and fares. Please refer to the link for a full description of the columns.

The data has been randomly split into 20 disjoint subsets which are available at **`https://github.com/PACELab/cse544-sp25-project-data`**. You will do the project in groups and can use the same groups as your assignments or build new groups. Groups should be at most of size 4. Each group will choose one of the subsets (we will maintain a Google spreadsheet **`cse544_sp25__project_groups`** where you will mention the group members and which subset you are using). Please select a different subset from every other group.

The tasks to be completed are listed below; we plan to add more tasks in the coming week as more topics are covered in class.

PART A: Speed Differences Across Airlines, Airplanes, and Routes

In this part, we will try to answer questions like : Are American Airlines flights slower than Delta flights? Are Boeings faster than Airbuses? To do so we will analyze the average speeds of the flights listed in the dataset.

(i) Dataset preprocessing

Consider only non-stop flights for this part. You have to create a new dataset listing the distance, duration, airline, and airplane type for all non-stop flights. Add a new column called "`speed`," which is calculated by dividing the distance by the duration.

Many rows in your new dataset will be exactly identical. What should you do with them? Your goal in this part of the project is to check if there are differences in flight speeds across airplane types, airlines, and routes. Should you count repeated rows only once, or as many times as they appear? What if two rows are almost identical (e.g., two rows with the same airline, airplane, and route but slightly different durations or distances)? Should you use one, both, or their average? Make a choice and explain it.

Since the dataset is very large you should choose at what stage of the analysis you should handle repetitions to keep the analysis tractable.

(iv) Examine the probability distribution of speeds.

Plot a histogram of the speeds. Comment on it.

Filter out the speeds for the airplane type Boeing 737-800. Do a K-S test to check whether it is normal with the sample mean and sample variance as the parameters of the Normal. Choose a threshold value of 0.05. Do the same for the speeds for airline Delta.

Using the 2-sample K-S test and the Permutation test check if the probability distributions of speeds for the Boeing 737-800 and Boeing 737-900 airplanes differ or are same.

Do the same for the Delta vs American Airlines flights.

(v) Compare speeds for airplane types, airlines, and routes.

Consider the following 4 airplane types: Boeing 737-800, Boeing 737-900, Airbus A321, Airbus A320. For each pair in the above (so 6 pairs), perform hypothesis tests to check if the speeds between the two airplane types are different. Compare using means and use two independent sample Wald's, Z-, and T-tests. For Z-test, use the sample variance as the true variance. For Wald's, use sample mean as the estimator. For all tests, perform the tests assuming all required assumptions hold. Then, comment on whether these assumptions actually seem to hold or not. For the critical value to compare against, use appropriate z-scores everywhere (even for the T-test).

Repeat the same tests for the following pairs of airlines : American Airlines and Delta, Delta and Alaska Airlines, JetBlue Airlines and Delta.

Divide your dataset into short-haul flights (distance < 1000), medium-haul flights (distance between 1000 and 2200), and long-haul flights (distance>2200). Perform the same test for all pairs (so 3 pairs) between these three categories.

Part B: How do these airlines decide the fares?

In this part, we will put the fares charged by the airlines under our statistical microscope to determine how they come up with their mysterious numbers. Use the **`baseFares`** column of the data for fares.

(i) Outlier detection and handling

Start by determining outliers. Use outlier detection methods discussed in the class that you find appropriate. Clearly report which methods you use and what you find. Decide what to do with the outliers (remove/modify/keep as is).

(ii) Relationship between fare and distance.

Fares should definitely be influenced by the distance—longer flights should be more expensive. We will check this relation in this part. Focus only on non-stop flights for now.

Plot fares against the distance travelled in the dataset. You may have a hard time discerning a clear trend from this plot directly. Notice that each distance has many fares corresponding to it (because each row of the data is one ticket, and all tickets of the same flight—even those on different dates—have the same distance, but different fares). To better see the effect of distance on fares, find the mean of the fares for each distance. Now plot these means against the distances; you should see something more reasonable. Perform a simple regression analysis of fares on distances and report your results (MAPE and SSE).

(iii) Predicting Ticket Fares Using Multiple Linear Regression

In this part, you will explore how well flight-related attributes can predict the ticket fares. Specifically, you will build a multiple linear regression model to estimate the `totalFare` of a non-stop flight using a set of numerical features. The goal is to identify which features contribute most to fare prediction, and whether a reduced set of features performs nearly as well as the full set.

Use only the following 8 numerical features from the dataset:

- `baseFare`
- `seatsRemaining`
- `elapsedDays`
- `segmentsDurationInSeconds`
- `segmentsDistance`
- `segmentsDepartureTimeEpochSeconds`
- `segmentsArrivalTimeEpochSeconds`

- travelDuration

Begin by preparing your dataset to include just non-stop flights and the above features. Compute the Pearson correlation coefficient between each of these features and totalFare. Based on the correlation values, select a subset of features that you believe are the most informative for predicting fare — explain your reasoning and selection criteria.

To improve the quality of your selected features, perform a similarity-based feature filtering step. Compute pairwise cosine similarity between your chosen features to identify pairs of features that are very similar to each other. For any group of highly similar features, retain only one representative feature and discard the rest. You may decide what similarity threshold to use, but justify your choice.

You will then train two multiple linear regression models:

- One using all 8 features
- Another using your selected subset which provides the lowest error

Evaluate and compare the performance of these models using the **Sum of Squared Errors (SSE)**. You need to evaluate the models on the whole dataset using a reasonable train-test split (e.g., 70:30) and compare both train and test SSEs for the models. You can use a bar plot to visualize the comparison.

Part C: Time series analysis of fares

The dataset has information about flights from around mid-April to mid-November in 2022 which is a period of around 31 weeks. Looking at the daily variation of fares in this time interval reveals interesting patterns that we shall explore in this part.

We will fix two routes to examine: a) JFK to ORD (Chicago), b) LAX (Los Angeles) to Atlanta (ATL). For both of these routes (separately) do the following:

(i) Extract the flightDate and the totalFare for all flights in the full dataset for the route. Find the average fare for each day. Plot the average fare vs date.

(ii) Perform an AR(k) analysis on the average fares with k varying from 3 to 10. Find the MAPE and the SSE metrics for your analyses and plot these as a function of k (k from 3

to 10) in two separate plots. What do you observe? What values of k work well? Why do you think these values of k work well? Note that for making the prediction and for evaluating your prediction, use the following methodology:
Use the first 15 days of the dataset for training, and predict the 16th day's value. Then, for predicting the 17th day's value, use the first 16 days of the dataset. Continue this way until the end of the dataset. If there are x days in the dataset, use the (x-1) days to train your model to make the prediction for the x'th day. When computing SSE and MAPE, compare predicted values with ground truth values as usual.

(iv) Perform EWMA on the same data for ɑ=0.5 and ɑ=0.8. Choose $\widehat{y}_1 = y_1$ as your initial predictor. Again find the MAPE and the SSE metrics. Comment on the performance of the best AR method (the one with the best p among those tried) vs the performance of the best EWMA method (the one with the best ɑ among those tried). Use the same methodology for training/testing as in (iii) above.

(v) Repeat the above analyses for weekly rather than daily averages. The dataset has information for 31 weeks starting from a Sunday (17 Apr 2022). Find the average fare for each week (Sunday to the next Saturday) and perform the AR (use k=2,3,4) and the EWMA analyses (use ɑ=0.5, 0.8). Comment on the results obtained. For training/testing, start by training on the first 5 weeks and predicting for the 6th week. Then train on the first 6 weeks to predict the 7th week, and so on.

(vi) What differences do you notice between the two routes in the performances of the methods. Plot the daily average fare data for both routes in one plot and explain if the performance differences between the two routes make sense according to the trends you observe.