**IST 718**

**Big Data Analytics**

**Project Title**

New Zealand Domestic Air-Fare

To predict domestic air fare of New Zealand’s airline’s based on the given attributes.

**Group Number**

19

**Group Members**

Jasdeep Monga-**414509580**

Priya Rao-**660997227**

Yunhuai Cao

**Project under guidance of**

*Professo****r Emory Creel***

**MA Physics**

**M.S. Probability and Statistics**

**Major**

**M.S. Applied Data Science**

**at**

***Syracuse University***



**Objective**: Analysis is being done on airline pricing practices and competition for airline carriers in New Zealand. The data set displays domestic airfares for New Zealand airlines on various domestic routes, especially for Air New Zealand, Jetstar, and SoundAir. Based on the above characteristics, the airlines compete with one another to predict domestic fares for various New Zealand carriers.

**Data Set Description:**

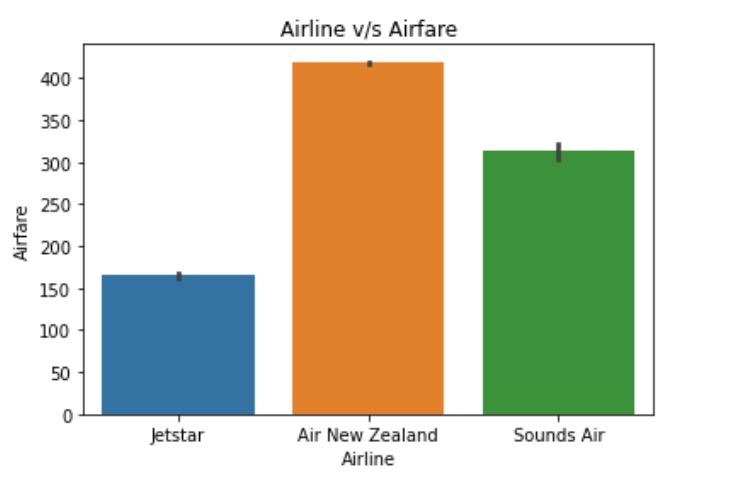
* **Overview:** By analyzing all the independent features with our dependent variable(fare) we will be predicting the price of the fare.
* **Number of rows and cols**- We have **162833 rows** and **11 columns.**
* **Sample predictors: Below is the list of Sample predictors:**
  + **Travel Date**- Day the person is boarding the flight
  + **Departure Airport**: Flight departure airport
  + **Arr. Airport**- Arrival airport
  + **Duration**: Total duration of the flight i.e., time
  + **Direct**- Flight has in between stops or not.
  + **Transit**- Stopover at a particular airport
  + **Baggage**- Is the check-in bag allowed with the fare.
  + **Airline**- Airline chosen by the traveler.
  + **Airfare**-Price of the ticket (Target Variable)
  + **Data Set link**- [NZ\_Airfares](https://drive.google.com/file/d/1LgkUl6bUNst2A-Jl78TtnyNfBuAOC-T5/view?usp=sharing)
  + **Interesting Facts**
    - The data was collected by **Expedia group**, one of the biggest online travel company and it is from **Year 2019**.
    - Major airlines in NZ are ‘**Air New Zealand**’ and ‘**Jetsar**’.
    - The data set ranges from **19th September 2019** to **18th December 2019, a total of 90 days.**
    - **There are 26 airport pairs.**

**Preliminary Data Exploration**

* To begin with we checked the shape of the data set it has **162833 rows** and **11 columns.**
* **After removing all the Null Values we have 162804 rows and 11 columns.**
* Data type of all the columns is ‘**str’** which we changed to their respective data types for e.g., all the data columns were encoded as ‘str’ so we changed it to datetime.
* **Maximum Fare** Price is **$1364** , **Minimum Fare** Price is **$32** and **Average Fare** Price is **$411**.
* The primary airports are: **Auckland (AKL),** **Christchurch(CHC),** **Wellington(WLG)** and **Queenstown(ZQN).** Secondary airport are: **Dunedin (DUD**), **Napier/Hastings(NPE),** **New Plymouth(NPL)**, **Nelson(NSN)** and **Palmerston North(PMR).**
* There are **52 one way routes**, out of which **12 are trunk routes** i.e., between primary airports and **40 secondary routes**.

**Visualization**

* **Airfare v/s Airline**
* Which Airline has the highest fare in NZ can be observed from the bar chart below.

****

**Predictions:**

To predict the Airfare we will take into considerations some actors that could heavily influence the **price** such as:

* **Travel Date**- From the travel date we will extract the **Month** and **Day** when the person was travelling and how did that affect the Airfare.
* **Direct**- Whether having multiple stops has an effect on the Airfare.
* **Baggage**-Flights that allow multiple bags and the ones that don’t allow any bag how is that affecting the price.
* **Transit-** Transit Duration at an airport how is that changing the Airfare.
* **Airline-** Which airline has the lowest airfare and is being considered by most of the passengers.

**Inference:**

The inference that I would draw from my project is as follows:

* We will get to know which is the best **month** to travel i.e., when the airline fare is cheaper and make travel plans accordingly.
* Which airline is the most convenient choice for passengers because of low fare.
* Do we have low fare for airlines that takes fewer stops to reach the destination.

**Non Spark Packages:**

* Non spark packages that we will use in this project will be: **Numpy, pandas, matplotlib, seaborn and plotly**.
* We will use numpy and pandas package to create dataframes.
* The use of **matplotlib, seaborn and plotly** packages will be only limited to creating graphs.
* The entire project will be done in spark itself.