STATS 501 : PROJECT PROPOSAL

Statistical Analysis of Airline Delays

1. Group Members

Aman Singh Thakur - amansinghtha@umass.edu
Astha Baranwal - abaranwal@umass.edu
Samriddhi Raj - samriddhiraj@umass.edu

2. Dataset :

To understand valuable insights, patterns and trends in flight delays and cancellations, as well as relationship between different variables, we'll be performing statistical analysis on the 'Airline On-Time Statistics and Delay Causes dataset' provided by the US Bureau of Transportation.

Link : https://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp

Variable	Description
year	the year in which the flight took place
month	the month in which the flight took place
carrier	the airline carrier code
carrier_name	the name of the airline carrier
airport	the airport code
airport_name	the name of the airport
arr_flights	the total number of flights arriving at the airport
arr_del15	the number of flights arriving at the airport delayed by 15 minutes or more
carrier_ct	the number of delayed flights due to the carrier
weather_ct	the number of delayed flights due to weather conditions
nas_ct	the number of delayed flights due to the National Aviation System (NAS)

security_ct	the number of delayed flights due to security issues
late_aircraft_ct	the number of delayed flights due to late aircrafts
arr_cancelled	the number of canceled flights
arr_diverted	the number of diverted flights
arr_delay	the total number of minutes of delay for all arriving flights
carrier_delay	the total number of minutes of delay due to carrier issues
weather_delay	the total number of minutes of delay due to weather conditions
nas_delay	the total number of minutes of delay due to the National Aviation System (NAS)
security_delay	the total number of minutes of delay due to security issues
late_aircraft_delay	the total number of minutes of delay due to late aircrafts

Table 1 - Dataset Columns and Descriptions

3. Statistical Analysis using R -

a. Summarization of Population -

1. Calculation of **Mean** and **Standard Deviation** of the Numerical Variables : arr_flights, arr_del15, arr_delay, carrier_delay, weather_delay, nas_delay, security_delay, and late_aircraft_delay.

2. Graphical Representation of Data : Use **PieCharts** to summarize categorical data and **BoxPlots** to summarize numerical data.

3. Create a new variable flight_not_delayed from arr_del15=0 to differentiate flights on-time vs delayed.

4. Calculate the **percentage/proportion** of flights that arrived on time = number of flights arrived on time (flight_not_delayed = 1) / total number of flights

b. Perform Hypothesis Test -

We can perform **hypothesis testing** to compare the mean number of delayed flights between any two major US airlines. We plan to choose the 2 major airlines that have similar route networks, size and customer bases. This will help to minimize the confounding factors that can affect the final results.

c. Estimate differences between population means -

To see if there are differences between two population means, We plan to use **Z-test** to compare the proportions of two populations. For instance, we could compare the proportion of on-time flights for two different carriers in the airlines dataset using this test.

d. Check if there is a relationship between two categorical variables

A **chi-square test** of independence can be used to determine the relationship between two categorical variables in the airlines dataset (for example, carrier and airport). This test will enable us to assess if the two variables have a significant correlation.

e. Linear Regression

We plan to perform a **linear regression** on the arr_delay variable (response variable) using the arr_flights, carrier_ct, weather_ct, nas_ct, security_ct, and late_aircraft_ct variables as predictors.

f. Analysis of Variance (ANOVA) test -

ANOVA test is used to compare the means of three or more populations. For our airlines dataset, we can use **ANOVA** to compare the average delay time of flights for three different airports.