**Research / Study Aim**

The purpose of this research is to study the Airline Data in the state of Ohio and understand various aspects of it. The focus of this study is to find if Delays of the Flights are affected by factors such as Departing Airport or the Airline Service Provider. This study also covers the aspects such as relations between Delays, and other Time factors which are used to study the Travel time of the flight.

**Introduction**

**Data Description**

The Airline data contains all the data related to the flights that travelled in the Month of March 2023. The dataset has 33 Attributes with 616,234 Observations that describe the Journey of the Flight and Time factors related to it.

For the study, the data is focused on Ohio State. Ohio State data consists of 33 Attributes with 11,844 Observations.

**Problem Statement**

Following are the problem statements for the study conducted on the Airline Data of Ohio State:

1. ***Statement 1***: At the time of departure, does the departing airport affect the flight delays?
2. ***Statement 2***: Does an Airline Carriers (Service Providers) impact the delays that occur in the state of Ohio at the time of take-off?
3. ***Statement 3***: Are Reservation System’s Elapsed Time and Actual Elapsed Time related?
4. ***Statement 4***: Does Departure Delay affect Arrival Delay?
5. ***Statement 5***: Is Taxi In time (Travel Time between Landing and Arrival Airport Gate) related to Taxi Out Time (Travel Time between Gate and Take Off)?

By analyzing the above problem statements, we can understand the Relations among the attributes and Factors that cause or affect delays for the flights that depart from various airports in Ohio State

**Exploratory Data Analysis**

A close-up of a computer screen

Description automatically generated

**Hypothesis Testing**

For the above provided problem statements, following are the Null and Alternative Hypotheses.

**Hypothesis 1:**

To study the if Origin Airport affects the Delay at the Time of Departure.

We use Analysis of Variance (ANOVA) to find whether the average delay of all the flights from the Origin Airports are same or different. For ANOVA, we use Departure Delay (DEP\_DELAY) as our Quantitative Variable (Measure of Time) and Origin Airport (ORIGIN) as Categorical Variable to study if Mean of the DEP\_DELAY is same, under each ORIGIN variable or not.

Following are the Null and Alternate Hypothesis for Hypothesis 1

*Null Hypothesis (****H0****)*: *μCAK* = *μ*CLE = *μ*CMH = *μ*CVG = *μ*DAY = *μ*LCK = *μ*TOL i.e., Average Flight Delays in all Airports of Ohio are not Different, where *μ* is Mean of DEP\_DELAY

*Alternate Hypothesis* (***HA*)**: At least 1 Mean is different i.e., *μCAK* ≠ *μ*CLE or *μCAK* ≠ *μ*TOL or so-on, 1 or more Means of the DEP\_DELAY is different from each other under different ORIGIN.

***Procedure:***

Using R and RStudio, we select the columns that are needed for Hypothesis 1 (ORIGIN, ORIGIN\_CITY\_NAME, DEP\_DELAY). Following Graph shows the general outline of Departure Delays based on Origin Airport

A graph of different colored dots

Description automatically generated

Null Values and Outliers are removed from the data as Outliers affect the ANOVA (Data Selection and Cleaning). Following Boxplots represent the data before and after cleaning.

A graph of a graph with black dots

Description automatically generated

A graph with a number of boxes

Description automatically generated

After cleaning data, we perform ANOVA to determine whether our Null Hypothesis is accepted or not. Following is result obtained after performing ANOVA for clean data:

A screenshot of a computer code

Description automatically generated

Above image shows the results obtained for the Analysis of Variance Test in R Programming using aov( ) function and DEP\_DELAY and ORIGIN as parameters.

***Results:***

The above results show that the main effect of ORIGIN on DEP\_DELAY is statistically significant and small (F(6, 9952) = 33.36, *p* < .001; Eta2 = 0.02, 95% Confidence Interval [0.01, 1.00])

As *p-value* is less than 0.001 with is less that the Significance Level (*α*) = 0.05. We reject *H0* and conclude that Mean (DEP\_DELAY) of at least 1 Origin Airport is different and Origin Airport affects the Departure Delay for the Flights that take off from various Airports from Ohio State.

**Hypothesis 2**

To study if there is an impact on Flight Delays at Departure Time by the Airline Carriers (Service Providers).

We use Analysis of Variance (ANOVA) to find whether the average delays of all the flights of the Airline Carriers are same or different. For ANOVA, we use Departure Delay (DEP\_DELAY) as our Quantitative Variable (Measure of Time) and Airline Carrier (MKT\_UNIQUE\_CARRIER) as Categorical Variable to study if Mean of the DEP\_DELAY is same, under all MKT\_UNIQUE\_CARRIER variables or not.

Following are the Null and Alternate Hypothesis for Hypothesis 2

*Null Hypothesis (****H0****)*: *μAA* = *μ*AS = *μ*B6 = *μ*DL = *μ*F9 = *μ*G4 = *μ*NK = *μ*UA = *μ*WN i.e., Average Flight Delays for all Airline Carriers are not Different, where *μ* is Mean of DEP\_DELAY

*Alternate Hypothesis* (***HA*)**: At least 1 Mean is different i.e., *μAA* ≠ *μ*UA or *μWN* ≠ *μ*B6 or so-on, 1 or more Means of the DEP\_DELAY is different from each other for MKT\_UNIQUE\_CARRIER.

***Procedure:***

Using R and RStudio, we select the columns that are needed for Hypothesis 2 (MKT\_UNIQUE\_CARRIER, MKT\_CARRIER\_FL\_NUM, DEP\_DELAY). Following Graph shows the general outline of Departure Delays for all Airline Carriers that Depart from Ohio State

A graph of different colored dots

Description automatically generated

We remove Null Values and Outliers (instead of replacing them as they cover only 9.1% of data) and use the cleaned data for Analysis of Variance. Following Graphs shows the comparison for data with and without outliers.

<Boxplot3>

<Boxplot4>

With the dataset obtained after cleaning the outliers and null values, Analysis of Variance is performed on it to decide whether to accept or reject Null Hypothesis.

<Code Screenshot>

Above imag shows the results of Analysis of Variance performed in R Programming by passing DEP\_DELAY and MKT\_UNIQUE\_CARRIER as the parameters for the function aov().

***Results:***

The results from the above performed analysis confirms that the main effect of MKT\_UNIQUE\_CARRIER on DEP\_DELAY is statistically significant and small (F(8, 9950) = 55.90, *p <* .001; Eta2 = 0.04, 95% Confidence Interval [0.04, 1.00])

The results drawn shows that the *p-value* is less than Level of Significance (*α*) = 0.05. We reject *H0* and conclude that Mean (DEP\_DELAY) of at least 1 Airline Carrier (Service Provider) is different and Airline Carriers affects the Departure Delay for the Flights that take off from various Airports from Ohio State.

**Hypothesis 3**