HBRP PUBLICATION

Journal of Network Security and Data Mining
Volume 4 Issue 2

Analysis of Factors Influencing Airfare of Domestic Airlines: Data from Local Ticket Booking Agency

Tohfa Niraula[1*], Pratistha Palikhel[2], Dr. Sushil Shrestha[3]
[1,2,3] Department of Computer Science and Engineering, Kathmandu University, Dhulikhel, Kavre, Nepal.

****Corresponding Author***
E-Mail Id:- *tohfa.niraula1@gmail.com*

ABSTRACT
Airline ticket prices can vary dramatically for flights in the same sector and from the same airlines depending on the festive seasons. Last year, during festival season, to facilitate festive commuters, domestic airline companies had increased flights to many different destinations. In October of 2019, air transport operators said that though the flights from Kathmandu to other destinations are booked to 100 per cent capacity, flights return to Kathmandu at less than 20 percent occupancy. Nearly 2.5 million people were expected to leave Kathmandu during Dashain.[1] The difference in the price of flight tickets is affected by other factors like the length of the flight route and the difference in booking date and flight date. This paper aims to study the correlation and interactions among different attributes of an airfare price using data mining and machine learning techniques and answer some questions related to airfare price using statistical analysis based hypothesis testing methods.

Keywords:-*Airfare price, data mining, machine learning, statistical analysis, hypothesis testing.*

INTRODUCTION

Airfare prices are calculated by applying complex methods [2]. Airlines take into account various factors: social, economic and commercial that impact airfares. The fluctuation in airfare is difficult to understand due to the complex strategies used to design the pricing models. For this reason, many have used machine learning techniques to predict airfare prices. Groves and Gini [2] applied a regression model to optimize airline ticket purchase, with 75.3% accuracy (acc.). Ren, Yang and Yuan [3] studied the performance of Linear Regression (77.06% acc.), Naïve Bayes (73.06% acc.), and SVM (80.6% acc. for two bins) models in predicting air ticket prices. K. Tziridis, Th. Kalampokas and G.A. Papakostas [2] compared the accuracy of SVM, Linear Regression, Multilayer Perceptron (MLP) and Random Forest along with other machine learning techniques.

The above-mentioned works use machine learning techniques to predict the sales price. While the purpose of this paper is to identify the factors that impact air ticket prices and test different propositions using statistical analysis based hypothesis testing.

PROBLEM STATEMENT

Flight cost changes dynamically, and airline corporations use complex methods to assign airfare prices and consider several financial, marketing, commercial and social factors. Understanding how factors affect change in flight cost, identifying the factors that impact the air ticket prices and testing different propositions using statistical analysis

based hypothesis testing is a complex but valuable study.

RESEARCH QUESTIONS

The purpose of this study is to answer some questions related to airfare price: (i)Will the price of flight tickets be higher on the week before and after the festival days compared to non-festival days? (ii)Will the price of the flight ticket be higher on the festival days compared to non-festival days? (iii) Does the distance between the airports have a relation with price per km between those airports? (iv)Does the difference between booking date and flight date affect the price of the flight ticket?

RELATED WORKS

Many have used machine learning techniques to predict airfare prices. Some of the research topics are "Airline ticket price and demand prediction: A survey"[4] by Juhar Ahmed Abdellaa Nazar Zakib Khaled and Shuaiba Fahad Khan uses the concept of social media data for ticket/demand prediction. "Airfare Prices Prediction Using Machine Learning Techniques"[2] by K.Tziridis, Th. Kalampokas, G.A. Papakostas compares the accuracy of different machine learning techniques in predicting the airfare price. "Factors influencing online flight ticket purchasing"[5] by Tae-Hong Ahh and Timothy Jeong Yeol Lee investigates the factors that influence online ticket purchasing through a survey of Internet consumers and "Airline Pricing under Different Market Conditions: evidence from European Low-Cost Carriers"[6] by Volodymyr Bilotkach, Alberto A. Gaggero, Claudio A. Piga uses the presence of reductions in proposed fares over time as an indicator of an active yield management intervention by two main European Low-Cost Carriers. The paper 'A study of Factors Influencing Purchase Decision of Thai Passenger in Bangkok'[7] by Miss Piyanart Chaisorn determines factors that influence purchasing behaviours of Thai passengers in Bangkok.

As in the works described above, vast areas like customer preference, ticket price prediction, booking rate prediction, ticket distribution strategy, effects of flight delay in the ticket price, price movements of competing airlines, etc. have been researched previously. This research paper aims to find the factors that impact the price of flight tickets and answer some questions associated with airfare by proving hypotheses related to attributes: sales, festive, airport distance.

METHODOLOGY

This section describes each step involved in the process. Subsections define the stages of the analysis.

Data Source

The source of data was a local ticket booking company 'Aurgia Destination Managers', located in Golfutar, Kathmandu. The data of two festive months Ashwin and Bhadra were taken. The two months were crucial to be included in the data collection as the two biggest festivals of Nepal: Dashain and Tihar fall in these months and those are crucial for the study and proof of three of the hypotheses which are based on festivals. Extensive research has been performed in the dataset of the flight booking company.

Data Preprocessing

Since the data was entered manually in the google sheet, there was some mistake in the data which was later corrected using some techniques. Besides that, there were missing values in the original data, most of them were filled using assumption techniques and others were ignored.

Data Cleaning

1.Handling Missing Data: Some of the missing values were predicted using

assumption. For example, if the flight price was missing in the tuple, corresponding Airlines and routes would be considered and the mod on their sales will be filled in place of the missing value. When the flight date was missing we could not assume the flight day of that customer. So, we had to remove that tuple.

2. Handling Noisy Data: Noisy values were treated using the binning method. The flight time of customers was divided into the early morning, late morning and early afternoon.

3. Data Transformation: Normalization of airport distance and cost per km attributes were performed using Min-max normalization. The formula for min-max normalization is:

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

4. Data Reduction: In data reduction, irrelevant attributes are removed. To prove the hypothesis, we needed attributes related to sales and festive days. So, attributes were selected using correlated feature selection. After data preprocessing, we had 179 tuples and 10 attributes.

Exploratory Data Analysis (EDA)

Exploratory data analysis (EDA) is used to analyze the data sets to summarize their main characteristics. The exploratory data analysis of a dataset from Auriga Destinations are: Pie Chart, Scatter Plot, Histogram and Double Bar Chart.

A. Pie Chart

The below pie chart shows the preference of airlines by the customers of Auriga Destination Managers for the month of September, October and November of 2019.

From Figure 4.1 it can be seen that the first preference of the customers is Budhha airlines with 35.49%. Similarly, Yeti airlines hold 30.3%, RA airlines hold 20.2%, Saurya airlines hold 2.2% and Shree airlines hold 11.8%

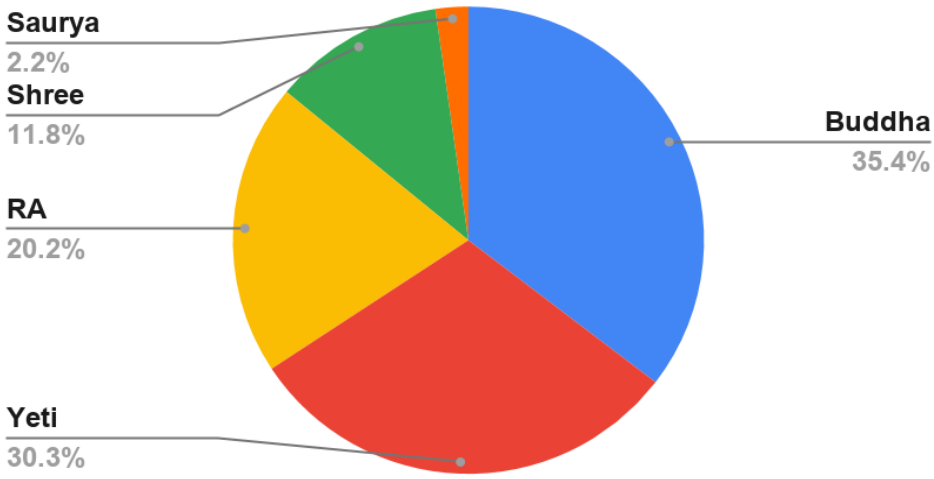

Fig.1:-*Count of airlines of Aurgia Destination Managers for the month Bhadra and Aswin of 2076*

B. Scatter Plot

A scatter plot uses dots to represent values and observe relationships between two variables. The position of each dot on the axes indicates values for an individual data point. Figure 4.2 shows that the sales of flight tickets are both high and more frequent when the difference in flight date from booking date is less than 5 days.

Auriga Destination Managers for the month of Bhadra and Aswin of 2076. It shows that the customers in their mid-forties were the most flight booking customers for those months

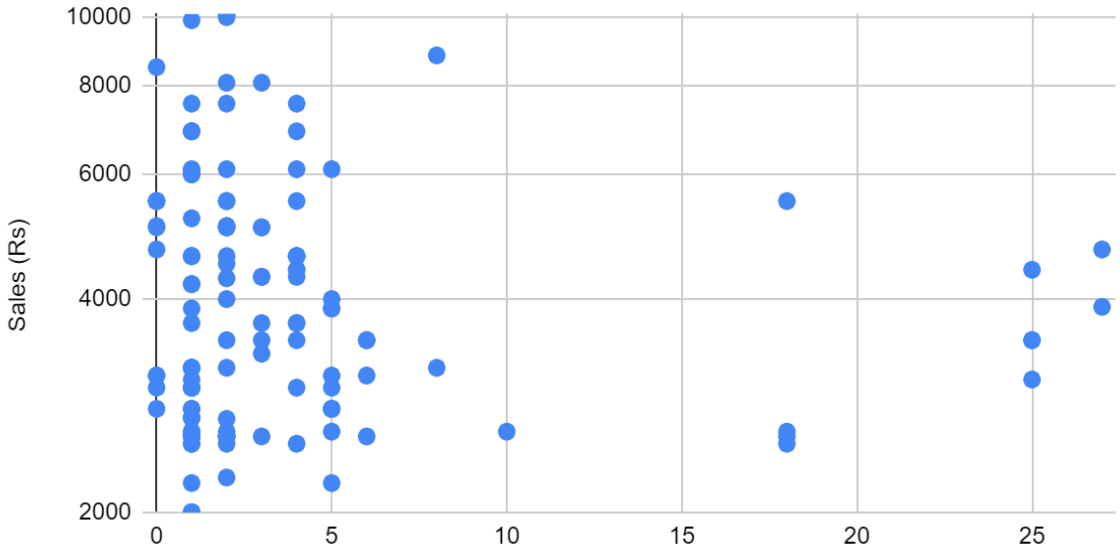

Fig.2:-*Correlation between the difference between flight date and booking date, and sales*

C. Histogram

A histogram displays the shape and spread of continuous sample data using bars of different heights. Figure 4.3 shows the age distribution of the customers of Auriga for the month of Bhadra and Aswin of 2076.

D. Double Bar Chart

Double bar chart, as seen in figure 4.4, is used to compare minimum and maximum sales made by each of the airlines in Auriga for the month of Bhadra and Aswin of 2076. Here, the difference in range of Airfare from each airline for the same sector can be seen.

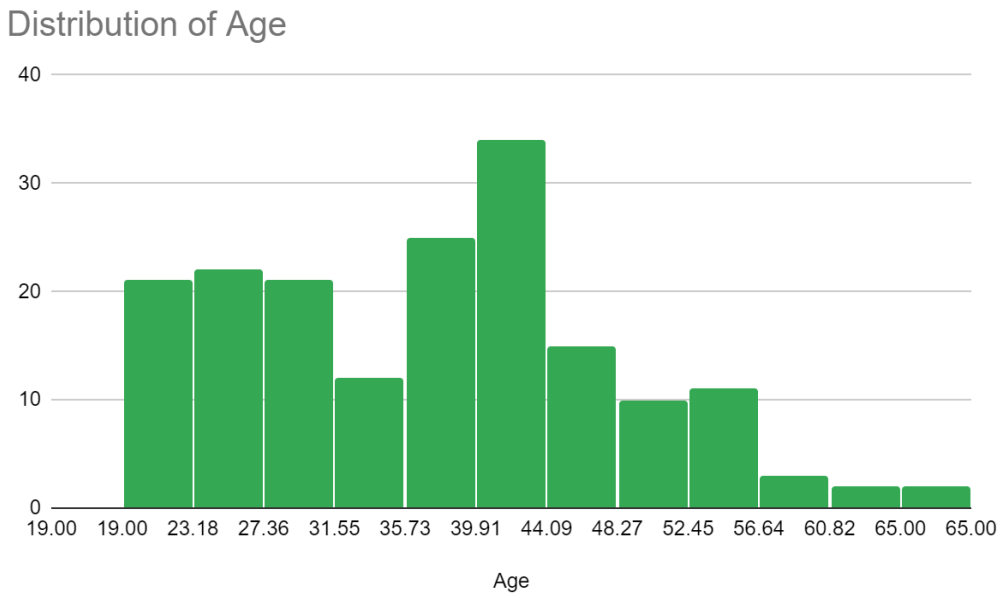

Fig.3:-*Age Distribution of Passengers*

From Auriga for the month of Bhadra and Aswin of 2076.

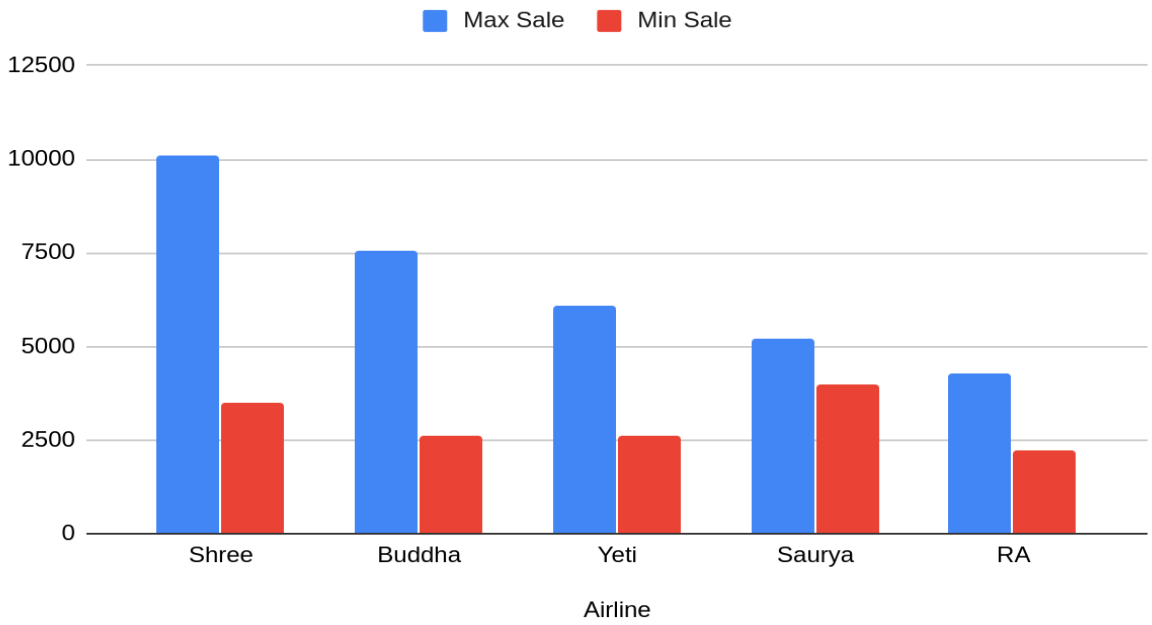

***Fig.4:**-Minimum and Maximum Sale made from each Airline by Auriga for the month of Bhadra and Ashwin, 2076*

RESEARCH METHODOLOGY

Four hypotheses are tested in this study (i) Will the price of flight tickets be higher on the week before and after the festival days compared to non-festival days? (ii)Will the price of the flight ticket be higher on the festival days compared to non-festival days? (iii) Does the distance between the airports have a relation with price per km between those airports? (iv)Does the difference between booking date and flight date affect the price of the flight ticket?

RESEARCH HYPOTHESIS

Based on research questions we have listed following hypotheses:

Hypothesis 1: "The price of flights increases 7 days before and after the festival days"

The following are the null and alternative hypotheses need to be tested:

Null Hypothesis (H0): The price of flights decreases 7 days before and after the festival days

Alternate Hypothesis (H1): The price of flights increases 7 days before and after the festival days.

- H0: $\mu 1 \leq \mu 2$
- H1: $\mu 1 > \mu 2$

Here, $\mu 1$ = mean of sales on 7 days before and after the festival days

$\mu 2$ = mean of sales on regular days

This corresponds to a right-tailed test, for which a t-test for two population means, with known population standard deviations, will be used.

The mean($\bar{x}$), standard deviation(σ) and sample size(n) of the two groups are given below:

Sales on	$\bar{x}$	σ	n
7 days off festivals	6289.318182	1954.62625	13
Regular days	5003.903021	1818.64213	13

The hypothesis was tested using t-test and the values computed by the test is given below:

A	df	t_c	t	p
0.05	24	1.711	1.736	0.0477

Since it is observed that t = 1.736 > tc = 1.711, it is then concluded that the null hypothesis is rejected. Using the P-value approach: The p-value is p = 0.0477, and since p = 0.0477 < 0.05, it is concluded

that the null hypothesis is rejected.

Hypothesis 2: "The price of flights increases on the day of the festival."

The following null and alternative hypotheses need to be tested:

Null Hypothesis (H0): The price of flights decreases on the festival days

Alternate Hypothesis (H1): The price of flights increases on the festival days.

- H0: $\mu 1 \leq \mu 2$
- H1: $\mu 1 > \mu 2$

Here, $\mu 1$ = mean of sales on the festival days

$\mu 2$ = mean of sales on regular days

This corresponds to a right-tailed test, for which a t-test for two population means, with known population standard deviations, will be used.

The mean($\bar{x}$), standard deviation(σ) and sample size(n) of the two groups are given below:

Sales on	$\bar{x}$	σ	n
Festival days	7207.092593	1901.138379	9
Regular days	5373.561568	2199.223367	9

The hypothesis was tested using t-test and the values computed by the test is given below:

α	t_c	t	*df*	p
0.05	1.746	0.1.892	16	0.0384

Since it is observed that t = 0.1.892 > t_c = 1.746, it is then concluded that the null hypothesis is rejected. Using the P-value approach: The p-value is p = 0.0384, and since p = 0.0384 < 0.05, it is concluded that the null hypothesis is rejected.

Hypothesis 3: "Distance of a route has a positive effect on the cost per km of the route."

The following null and alternative hypotheses need to be tested:

Null Hypothesis (H_0): Distance of a route has no effect on the cost per km of the route.

Alternate Hypothesis (H_1): Distance of a route has a positive effect on the cost per km of the route.

- H_0: $\sigma_1^2 \neq \sigma_2^2$
- H_1: $\sigma_1^2 > \sigma_2^2$

Here, σ_1^2 = variance of airport distance

σ_2^2 = variance of sales per km

This corresponds to a right-tailed test, for which a F-test for two population variances will be used. The variance(σ^2) and sample size(n) of the two groups are given below:

Groups	σ^2	n
Airport Distance(km)	14958.97846	26
Cost per km (Rs/km)	98.06153846	26

The hypothesis was tested using f-test and the values computed by the test is given below:

α	f_v	f
0.05	1.955	152.547

Since it is observed that f= 152.547 > f_v = 1.955, it is then concluded that the null hypothesis is rejected.

Hypothesis 4: "Difference of Booking date and Flight date affects the ticket price"

The following null and alternative hypotheses need to be tested:

Null Hypothesis (H_0): Difference of Booking and Flight date has no effect on the ticket price.

Alternate Hypothesis (H_1): Difference of Booking and Flight date has affected the on ticket price.

- H_0: $\mu_1 \neq \mu_2$
- H_1: $\mu_1 = \mu_2$

Here, μ_1 = mean of the difference between flight date and booking date

μ_2 = mean of the ticket price

This corresponds to a χ^2-test for two population means, with the known observed and expected value, will be used.

The values computed by the χ^2 test is given below:

α	χ^2_c	χ^2	P
0.05	392.501	363.503	0.273

Using the P-value approach: The p-value is p =0.273, and since p = 0.273 $\geq$ 0.05, it is concluded that the null hypothesis is not rejected.

RESULTS

In hypothesis 1 with a 5% level of significance, the calculated p-value is 0.0477 in data and since 0.0477 < 0.05 there is sufficient evidence to reject the null hypothesis. So, the price of flights does increase 7 days before and after the festival days. In figure 7.2, showing the result of hypothesis 2 with a 5% level of significance, the calculated p-value is 0.0389 in data and since 0.0384 < 0.05 there is sufficient evidence to reject the null hypothesis. So, the price of flights does increase on festival days.

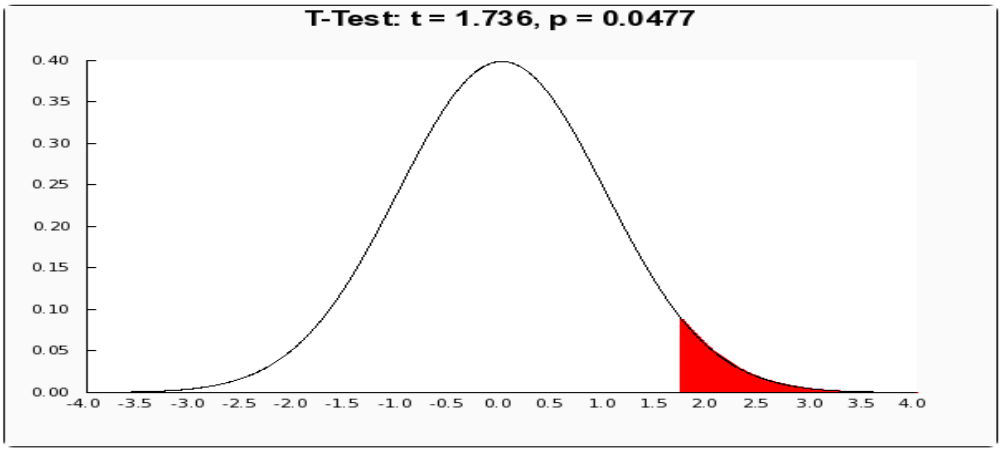

Fig.5:-*A graph showing t-distribution for p = 0.0477 < 0.05*

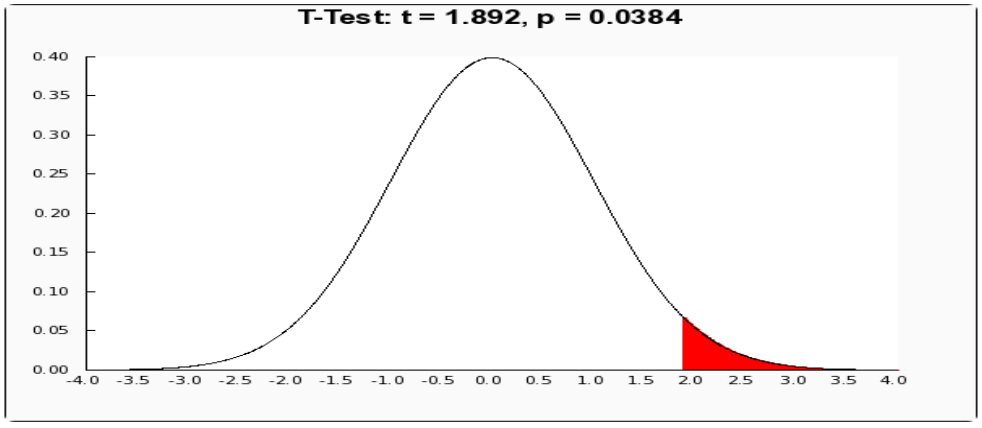

Fig.6:-*A graph showing t-distribution for p = 0.0384 < α =0.05*

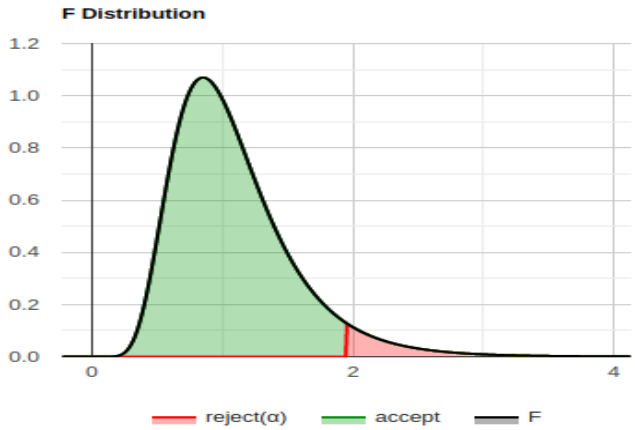

Fig.7:- *A graph showing F distribution for f= 152.547 > f_v = 1.955*

In Figure 7 showing the result of hypothesis 3, with a 5% level of significance, from figure 7.3, it is observed that F= 152.547 > F_v = 1.955, it is then concluded that the null hypothesis is rejected. So, it is concluded that Distance of a route has a positive effect on the cost per km of the route.

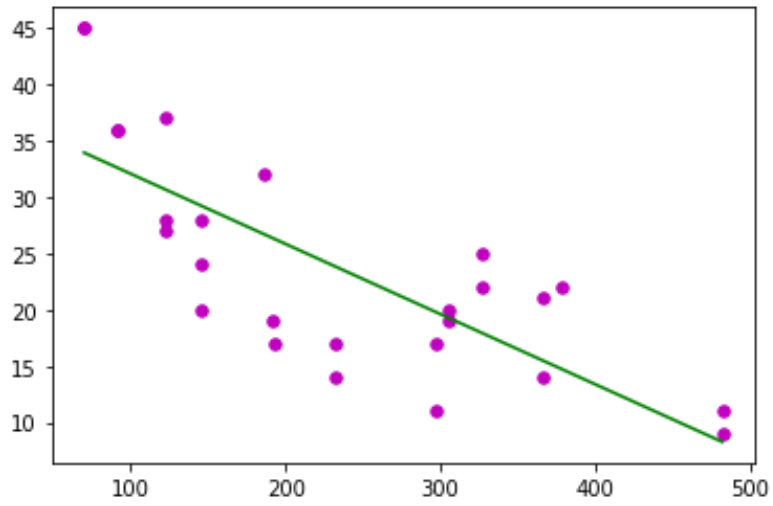

Fig.8:-*Linear regression of Cost per km and Airport Distance*.

As seen in Figure 8, the linear regression equation for cost per km and Airport distance is: Cost per km = -0.0623 * Airport Distance + 38.3079. In hypothesis 4, with a 5% level of significance, the critical value is 392.501. The H_o is rejected if $\chi^2 \geq 392.501$. But calculated χ^2 = 363.503 so, the difference in Booking date and Flight date does not affect the ticket price.

CONCLUSION

Through different hypothesis testing methods, we found that the sales rate doesn't increase a week before and after a week of festival days, the sales don't increase during the festival days. Also, the ticket price doesn't depend on the flight booking date for Aurgia Destination Managers. But through hypothesis testing and linear regression, it is seen that the distance between airports is inversely proportional to cost per km of travelling between those airports.

FUTURE SCOPE

This research has presented how some factors, mainly festive season, time of flight, and distance affect change in flight prices. There are many factors like time of ticket purchase, competition, price of oil and more that may have a greater influence on airfare and are yet to be explored. For further research more factors can be taken into account to analyze their influence on the constantly shifting price of airline tickets.

REFERENCES

1. Airlines adding flights to ease festive commute.(2020). *The Himalayan Times*. Available:https://thehimalayantimes.com/business/airlines-adding-flights-to-ease-festive-commute/. [Accessed: 14-- 2020].
2. Tziridis, K., Kalampokas, T., Papakostas, G. A., & Diamantaras, K. I. (2017, August). Airfare prices prediction using machine learning techniques. In *2017 25th European Signal Processing Conference (EUSIPCO)* (pp. 1036-1039). IEEE.
3. Ren, R., Yang, Y., & Yuan, S. (2014). Prediction of airline ticket price. *University of Stanford*.

4. Abdella, J. A., Zaki, N. M., Shuaib, K., & Khan, F. (2021). Airline ticket price and demand prediction: A survey. *Journal of King Saud University-Computer and Information Sciences*, *33*(4), 375-391.
5. Ahn, T. H., & Lee, T. J. (2011). Research note: Factors influencing online flight ticket purchasing. *Tourism Economics*, *17*(5), 1152-1160.
6. Bilotkach, V., Gaggero, A. A., & Piga, C. A. (2015). Airline pricing under different market conditions: Evidence from European Low-Cost Carriers. *Tourism Management*, *47*, 152-163.
7. *Ethesisarchive.library.tu.ac.th*, 2020. [Online]. Available: http://ethesisarchive.library.tu.ac.th/thesis/2016/TU_2016_5802040880_6036_4588.pdf. [Accessed: 24- May- 2020].