Population Distribution of Prices with in Flight Trip Factors

Ahmed Helali[1]

[1]*Faculty of Bio and Computer Sciences*, Applied Mathematics

Introduction

"The best day to buy tickets on is Tuesday"[6] is one of most given advice regarding buying plane tickets. In our research, we investigated the relation between prices and different other factors regarding flight tickets.

There was many research related to flight tickets[6], we note that most of the analysis done was on the united states[6][2] and we could not find a study targeted for Germany. We also noted that many of the analysis was done on a macro scale[2], where the data was analysed without grouping.

We aimed to investigate the different correlations between departure dates, airlines, lay overs with prices and gather deeper insights by checking locations independently. We gathered our data from an different APIs, our time frame for the flights was the month of August and our destinations included six cities; Cairo, New York, Rio de Janeiro, Rome, Sydney and Tokyo, one for each continent. We checked for any significance in the population distributions.

In section 2, we go more into detail regarding data collection methods, organization and features. We also outline our approach for our investigation. In section 3, we show our results and their significance to our problem. In Section 4, we discuss our insights on the results and possible future work and improvements. In the final section we include our references.

Methods

The data was collected From different APIs; we preformed searches of **One Way** tickets from **Berlin** to **6 different cities**, over the **whole month** of **August**, we used a *Python* program to organize the data and compile it in a *csv* file, the *csv* file contained 5069 entry rows, a row represents a flight trip, and 11 columns act as our variables. The table below a brief summary of the variables features of the data.

Variable name	Data type	Description	Example
Airline	string	The name of airline company	Qatar Airways
Departure	string	Name of the departure airport	Berlin (BER)
Arrival	string	Name of the arrival airport	Cairo (CAI)
Destination	string	The city of arrival	Rome
price	float	the ticket price	786
Duration	integer	flight time in mins	455
Stops	integer	no of layovers	2
Departure_time	string	24h time format of departure	19:35
Arrival_Time	string	24h time format of Arrival	08:05
Day_Week	string	the weekday of the date	Sunday
Day_Month	string	the date	20/08/2023

Table 1. The description of our variables. In csv file, they are also the header names.

In our research, our interest lies in finding analytical insights regarding the tickets prices. Our target is to assess the correlation between our variables and the the price variable.

First, we investigate the distributions of our variables and samples. We preformed visual techniques and **Shapiro-Wilk's method**[5] to test the normality of the data. The data failed in the normality test.

We proceeded with non-parametric tests, we used **Kruskal-Wallis test**[4][3] to check whether the population distributions are identical without assuming normality, our null hypothesis was that the population distributions would be the same. We compared the population distribution of price with in week days, airlines, lay overs with respect to all of the data and grouped by city.

We were able to obtain insights on the nature of ticket prices that will be discussed in the results section.

Results

With our now obtained data set as described in the previous section. We start for checking for normality in our data.

We plotted the density function of each variable of our data to check the normality. Examples are in figure 3. We also preform **Shapiro-Wilk's method**[5] to double check our visual results. Our data failed the normality test. In the table, we recorded the p-values.

We would like to understand the nature of prices more within the factors week days, airlines and lay overs. The data type of our price variable is numeric, while our factor variables is ordinal or categorical. And since:

1. our data is not normal.
2. we are comparing two different data types.

we preform **Kruskal-Wallis test**.

Our first null hypothesis was, the **ticket prices** have **identical distributions** in **all week days**:

$$H_0 : m_1 = m_2 = ... = m_n,$$

where m_i is the distribution of the ith day of the week.

Weeks start on *Monday* and end on *Sunday*, this implies that in our m_3 is the distribution of *Wednesday* and so on.

We obtained the p-values 0.004374, so we reject our null hypothesis. We do summary statistic analysis, calculating the mean, standard deviation and IQR.

In figure 1, we perform a box plot[1][4] on the distribution of prices with week days and performed the same box plot again but now grouped by cities.

Although, the plot on the whole data might suggest that the distributions is similar, we know by **Kruskal-Wallis test** that the distributions vary.

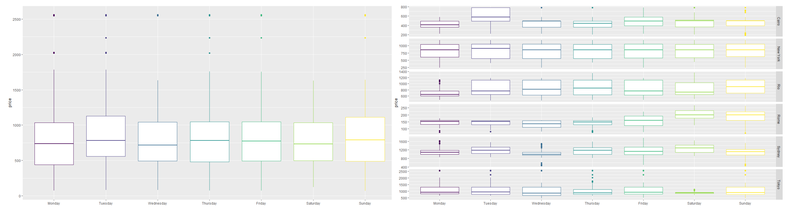

Figure 1. Box plots of price distribution with in week days on all of the data and divided by cities respectively.

From figure 1, there are unique results that we would like to point out:

1. the prices population distributions with in week days, in the whole data.
2. the prices prices population distributions with in week days when grouped by cities except for *New York*.

We would like to note that the results regarding *New York*, we preform **Kruskal-Wallis test** comparing prices and week days again only on *New York*, we obtain the p-value $= 0.8622$, and we are unable to reject our null hypothesis[3].

From our results, we reach the following conclusions.

City	Best Overall	Min	Price	Max	Price
Cairo	Monday	Sunday	189	Friday	784
New York	no diff	Monday	241	Tuesday	1187
Rio de Janeiro	Monday	Tuesday	454	Saturday	1357
Rome	Wednesday	Sunday	69	Saturday	268
Sydney	Wednesday	Sunday	463	Friday	1754
Tokyo	Wednesday	Wednesday	461	Tuesday	2559

Discussion

In our analysis, we

- gathered data from different APIs,
- organized & compiled the data,
- preformed normality tests,
- checked the significance between prices and other factors.

From our results, we conclude the that population densities are significantly different with in week days.

Al thought not included here, we performed also analysis and tests for prices with in other factors.

The data would suggest that the increase of the number of layovers is not always necessary to get a better price in some cities. The data suggests a cut off distance between the need to increase the number of lay overs to get a better price. This is interesting as it gives insights to previous research[2].

The investigation into the cutoff among also other improvements to our data set is intended in the future work. Also, the investigation of price with in airlines significance, questions like "Is it the day or the airline that affect prices the most?" and "Is a certain day cheaper because of the airline or is a certain airline cheaper because of the day?" is included into future studies.

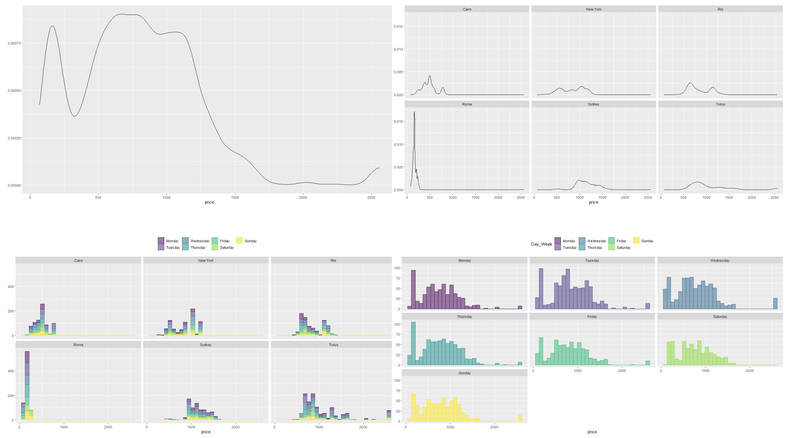

Figure 3. Top: prices density with in all data and divided by Cities. Bottom: price histograms with in cities and week days.

References

[1] Data Carpentry data visualization with ggplot2. https://statistics.laerd.com/spss-tutorials/kruskal-wallis-h-test-using-spss-statistics.php. Accessed: 2023-01-23.

[2] Google google flights. https://blog.google/products/travel/how-to-find-the-best-deal-on-your-next-flight/. Accessed: 2023-01-23.

[3] Laerd Statistics kruskal-wallis test. https://statistics.laerd.com/spss-tutorials/kruskal-wallis-h-test-using-spss-statistics.php. Accessed: 2023-01-23.

[4] R Tutor kruskal-wallis test. https://www.r-tutor.com/elementary-statistics/non-parametric-methods/kruskal-wallis-test. Accessed: 2023-01-23.

[5] STHDA normality test in r. http://www.sthda.com/english/wiki/normality-test-in-r. Accessed: 2023-01-23.

[6] Thrifty Traveler. https://thriftytraveler.com/news/travel/google-flights-data-analysis/. Accessed: 2023-01-23.

ahelali@hs-mittweida.de