Effect of Marginal Difference in Flight Duration on Ticket Price: LAX to NRT

Koko Xu

New York University

Courant Institute of Mathematical Sciences

December 25th, 2021

Introduction

In this research paper, I look to explore the relationship between the marginal differences in airplane flight time and the prices of airplane tickets.

This research topic is inspired by Venus Aerospace, a Hypersonic Transportation startup promising to deliver cost-effective global transports in one hour. According to Bloomberg Business, Venus Aerospace founders Sarah and Andrew Duggleby left Virgin Orbit to start the company because their Tokyo to Los Angeles flight time caused them to miss Sarah's grandmother's 95th birthday.

Although the reported flight is from Narita International Airport (NRT) to Los Angeles International Airport (LAX), given my residence in the United States, along with my hypothesis that airfares in either direction are comparable, I decided to explore the flight route of LAX to NRT instead.

It is known from Economics that Consumer Utility is the direct driver of Consumer Willingness to Pay. I wish to explore how much of Consumer Utility in the abovementioned flight route is due to flight time, and how said Utility Factor influence WTP and therefore price. Ultimately, I aim to use my findings to develop a basic pricing model for Venus Aerospace's one-hour transportation solution.

Methodology

Data Collection

The data used in this study are drawn from the flight bookings platform kayak.com. I drew data from the website pertaining to all flights from LAX to NRT, from December 26th, 2021 to January 1st, 2022. This period spans seven consecutive days, eliminating day-of-week bias. Three data points are collected: Airline, Flight Duration, and Price. The data is compiled into a CSV file.

Data Handling

I used Python and Pandas to organize and clean the data. I uploaded the CSV file as a dataframe and printed the first and last lines, as well as some basic information about the data for reference.

1. *Data Cleaning*

First, I converted the Flight Duration column from the String format of "##h ##m" into the Int format of "####" with minutes as the unit. Next, I converted the Price column from the String format of "$#,###.00" into the Int format of "####" with USD as the unit.

2. *Data Organization*

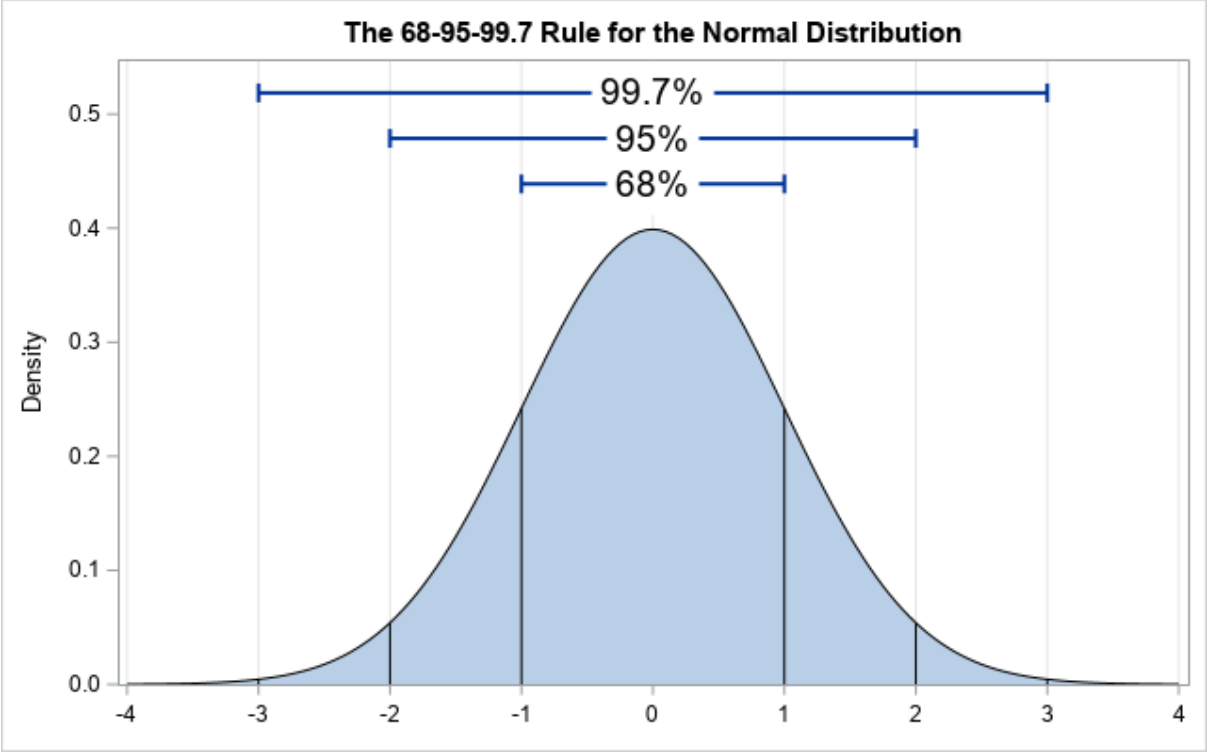
I organized the data by Airline and created a new dataframe for each airline containing said airline's data entries. This is because different airlines provide different utilities to their customers - some are known for good customer service and more comfortable seats; others are known for cheap prices and long waiting times - and thus, conducting separate analyses on each airline eliminates quality bias and highlights flight duration as the main driver of price.

3. *Data Visualization*

I used Seaborn and Matplotlib to plot the data with Time on the x-axis and Price on the y-axis. I also ran a linear regression between the two variables to produce a line of best fit. Note that this LoBF does not take into account the difference in service of different airlines, it is produced merely for visual demonstration.

4. *Data Selection*

Kayak.com assumes browsers are searching for the cheapest possible offerings, and therefore the majority of ticket listings show prices for Economy Class tickets. However, on occasions when the Economy Class is sold out, the next cheapest class ticket is shown. In order to eliminate outliers resulting from data describing Business Class and First Class ticket prices, I created a new column of data called "Ratio" where entries are calculated by Price divided by Time of the flight. I assumed that the distribution of this ratio for the flight LAX to NRT follows the Normal Distribution. Under this assumption, I eliminated any data points with ratios above one standard deviation of the mean to preserve the bottom 84% of the distribution.



5. *Data Analysis*

I used Pandas and Numpy to analyze the cohorts of data. I calculated the Overall Correlation of the dataframe for reference. I found the slope of the line of best fit with Time as the independent variable and Price as the dependent variable for each Airline. I then calculated the weight of the slope for the airline, defined by the number of data entries from the airline after eliminating outliers divided by the total number of entries after eliminating outliers. I multiplied the slope by the weight to arrive at the weighted slope of the airline, and summed the weighted slope of all airlines to arrive at Weighted Average Slope for all airlines. I repeated this process to find the Weighted Average Intercept for all airlines.

Results

Data Collection

Airline	Time	Price
Japan Airlines	11h 55m	$582.00
ANA	12h 00m	$642.00
United Airlines	14h 15m	$610.00
Air Canada	15h 15m	$448.00
United Airlines	16h 30m	$610.00
Alaska Airlines	17h 00m	$1,808.00
Alaska Airlines	17h 10m	$1,940.00
Air Canada	29h 40m	$425.00
Singapore Airlines	11h 45m	$471.00
Japan Airlines	11h 55m	$582.00
ANA	12h 00m	$617.00
ANA	14h 05m	$647.00
United Airlines	14h 15m	$610.00
Air Canada	15h 15m	$500.00
ANA	16h 20m	$647.00
Alaska Airlines	16h 20m	$5,541.00
United Airlines	16h 30m	$610.00
Air Canada	32h 08m	$425.00
Alaska Airlines	35h 05m	$1,790.00
United Airlines	11h 50m	$362.00
ZIPAIR	11h 55m	$362.00
Japan Airlines	11h 55m	$392.00

American Airlines	11h 55m	$610.00
ANA	12h 00m	$390.00
American Airlines	14h 15m	$366.00
United Airlines	14h 15m	$3,782.00
Alaska Airlines	15h 00m	$808.00
Air Canada	15h 15m	$500.00
United Airlines	16h 30m	$3,782.00
Alaska Airlines	16h 30m	$11,584.00
Alaska Airlines	17h 00m	$730.00
JetBlue	17h 40m	$1,210.00
Alaska Airlines	17h 45m	$587.00
Philippine Airlines	22h 20m	$2,315.00
American Airlines	25h 42m	$372.00
Alaska Airlines	28h 00m	$701.00
Singapore Airlines	11h 45m	$520.00
Japan Airlines	11h 55m	$561.00
ANA	12h 00m	$634.00
United Airlines	14h 15m	$610.00
American Airlines	14h 15m	$765.00
Alaska Airlines	15h 00m	$792.00
Air Canada	15h 45m	$3,778.00
United Airlines	16h 30m	$610.00
Alaska Airlines	16h 30m	$5,541.00
Alaska Airlines	17h 00m	$808.00
JetBlue	17h 40m	$1,210.00

Alaska Airlines	17h 45m	$682.00
United Airlines	23h 55m	$366.00
United Airlines	11h 50m	$362.00
ZIPAIR	11h 55m	$363.00
Japan Airlines	11h 55m	$392.00
ANA	12h 00m	$395.00
United Airlines	14h 05m	$629.00
United Airlines	14h 15m	$366.00
Air Canada	15h 15m	$425.00
United Airlines	16h 20m	$629.00
United Airlines	16h 30m	$366.00
Alaska Airlines	17h 10m	$1,418.00
Philippine Airlines	22h 20m	$2,203.00
American Airlines	25h 42m	$372.00
Singapore Airlines	11h 45m	$569.00
Japan Airlines	11h 55m	$561.00
ANA	12h 00m	$606.00
American Airlines	14h 15m	$614.00
United Airlines	14h 15m	$660.00
American Airlines	15h 00m	$664.00
Alaska Airlines	15h 00m	$792.00
Air Canada	15h 15m	$425.00
Air Canada	15h 45m	$610.00
United Airlines	16h 30m	$660.00
Alaska Airlines	17h 00m	$736.00

Alaska Airlines	17h 10m	$817.00
JetBlue	17h 40m	$2,258.00
Alaska Airlines	17h 45m	$730.00
United Airlines	23h 53m	$366.00
ZIPAIR	11h 45m	$363.00
Japan Airlines	11h 45m	$392.00
Singapore Airlines	11h 45m	$516.00
United Airlines	11h 50m	$362.00
ANA	12h 00m	$377.00
ANA	14h 05m	$588.00
United Airlines	14h 15m	$366.00
Air Canada	15h 15m	$448.00
Japan Airlines	15h 45m	$356.00
Alaska Airlines	26h 10m	$1,110.00

The table shows 86 data entries for flights from LAX to NRT from the week of 12/26/2021 to 1/1/2022 recording each flight's Airline provider, duration of flight ("Time"), and Price of ticket. The data is collected from kayak.com.

Data Handling

Below find the Python code I wrote to process the data and arrive at the Weighted Average Slope for the dataset.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

```
#General Information
df = pd.read_csv("~/Desktop/Python/Venus Aero/LAX-_NRT_ Price vs. Time - Sheet1.csv")
print(df.head(3))
print(df.tail(3))
print(df.info())
print(df.describe())


#Data Cleaning
k = 0
while k < len(df.index):
    i = df["Time"][k].split(" ")
    hours = int(i[0][:2])
    minutes = int(i[1][:2])
    minutes += hours * 60
    df["Time"][k] = minutes

    price = df["Price"][k].strip("$")
    price = price.replace(',', '')
    price = price.replace('.00', '')
    df["Price"][k] = int(price)

    name = df["Airline"][k].replace(" ", "")
    df["Airline"][k] = name
    k+=1


#Data Organization
df["Ratio"] = df["Price"] / df["Time"]
airlines = list(df.Airline.unique())
for airline in airlines:
    globals()[airline] = df.copy()
    globals()[airline].set_index("Airline", inplace=True)
    airlines.remove(airline)
    globals()[airline].drop(airlines, inplace=True)
    globals()[airline] = globals()[airline].reset_index()
    airlines = list(df.Airline.unique())
ANA = df.copy()
ANA.set_index("Airline", inplace=True)
airlines.remove("ANA")
ANA.drop(airlines, inplace=True)
ANA = ANA.reset_index()
```

```
#Data Visualization
x = pd.Series(df["Time"], name="Time", dtype='float64')
y = pd.Series(df["Price"], name="Price", dtype='float64')
ax = sns.regplot(x=x, y=y)
plt.show()

#Data Selection
def remove_outliers(dataframe):
   dataframe["Z-Score"] = (dataframe["Ratio"] -
dataframe["Ratio"].mean())/dataframe["Ratio"].std(ddof=0)
   k=0
   while k < len(dataframe.index):
      if dataframe["Z-Score"][k] > 1:
         dataframe.drop(k, inplace=True)
      k+=1

for airline in airlines:
   remove_outliers(globals()[airline])
remove_outliers(ANA)


#Data Analysis
df_x = pd.Series(df["Time"], dtype='float64')
df_y = pd.Series(df["Price"], dtype='float64')
corr = df_x.corr(df_y)
print("Overall Correlation:", corr)

def get_slope(airline):
   time = pd.Series(airline["Time"], dtype='float64')
   price = pd.Series(airline["Price"], dtype='float64')
   m, b = np.polyfit(time, price, 1)

   total_entries = len(JapanAirlines) +len(ANA) +len(UnitedAirlines) +len(AirCanada)
+len(AlaskaAirlines) +len(AmericanAirlines) +len(PhilippineAirlines) +len(JetBlue) +len(ZIPAIR)
+len(SingaporeAirlines)
   weight = len(airline) / total_entries

   slope = m * weight
   return slope

def get_intercept(airline):
   time = pd.Series(airline["Time"], dtype='float64')
```

```
    price = pd.Series(airline["Price"], dtype='float64')
    m, b = np.polyfit(time, price, 1)

    total_entries = len(JapanAirlines) +len(ANA) +len(UnitedAirlines) +len(AirCanada)
+len(AlaskaAirlines) +len(AmericanAirlines) +len(PhilippineAirlines) +len(JetBlue) +len(ZIPAIR)
+len(SingaporeAirlines)
    weight = len(airline) / total_entries

    intercept = b * weight
    return intercept

weighted_slope = 0
weighted_intercept = 0
for airline in airlines:
    weighted_slope += get_slope(globals()[airline])
    weighted_intercept += get_intercept(globals()[airline])
    print(airline + ": " + str(get_slope(globals()[airline])) + ", " + str(get_intercept(globals()[airline]))
)
weighted_slope += get_slope(ANA)
weighted_intercept += get_intercept(ANA)
print("ANA:", get_slope(ANA), get_intercept(ANA))
print("Weighted Average Slope:", weighted_slope)
print("Weighted Average Intercept:", weighted_intercept)
```

Data Visualization

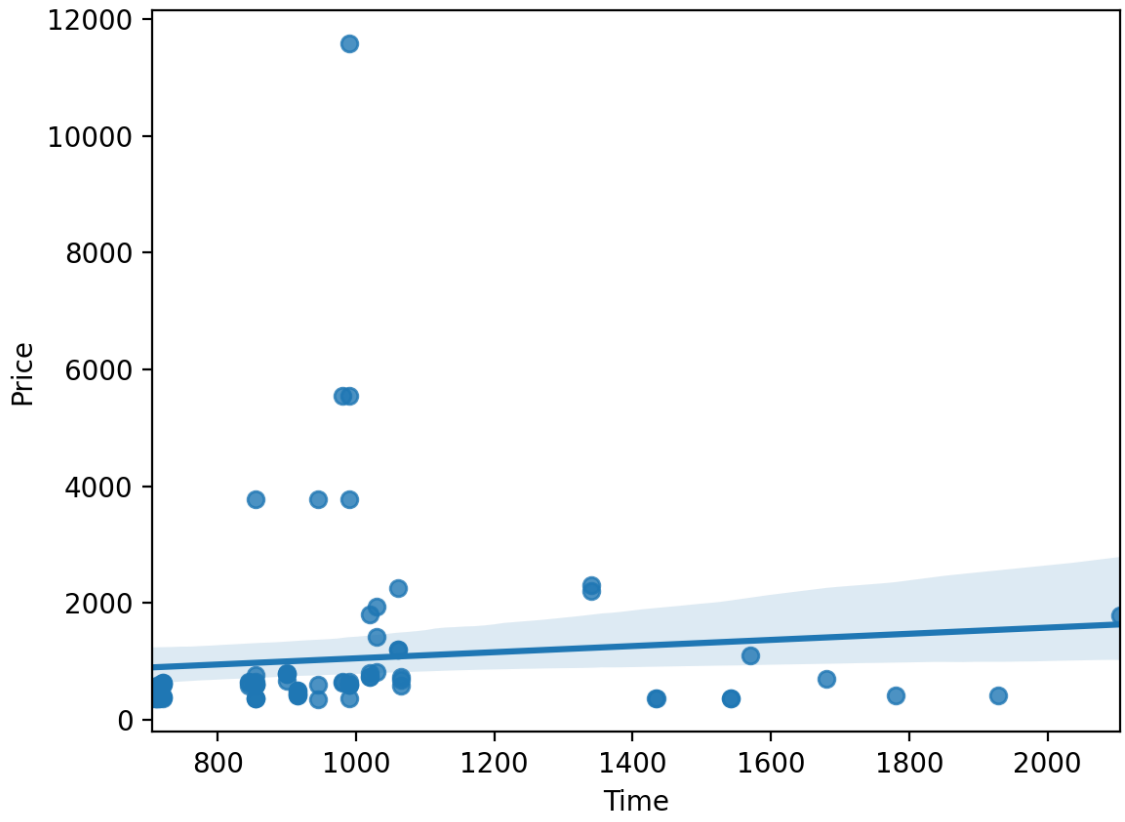


Figure above shows the distribution of all data entries for flights from LAX to NRT from the week of 12/26/2021 to 1/1/2022. The line represents the line of best fit for the data before eliminating outliers.

Data Analysis

Overall Correlation: 0.09854667721438881

Airline	Weighted Slope	Weighted Intercept
Japan Airlines	-0.036398211225811786	64.73960379487602
United Airlines	-0.016063495347135526	144.01409201215714

Air Canada	-0.006955935957443411	67.07877644375706
Alaska Airlines	0.09702220366641995	117.16234295861155
Singapore Airlines	0.015053441214663864	10.612676056338028
ZIPAIR	0.007140746577366297	5.105633802816899
American Airlines	-0.017721485670349004	53.815030698294834
JetBlue	0.01607759766144034	17.042253521126753
Philippine Airlines	0.02374395627496322	31.81690140845071
ANA	0.08590089623855827	-7.2262358776742746
Total Weighted Average	**0.16779971343267222**	**504.1610748187546**

The above table shows each airline's Weighted Slope value as taken from the line of best fit of each airline's Time and Price series. The table also shows the Weighted Average Slope of the whole dataset, approximately *0.16780*. This value means that for each one unit marginal increase in Time of flight from LAX to NRT, the Price of said flight increases by 0.16780 units. Since the units used for Time and Price throughout the analysis are minutes and USD respectively, we can interpret this as the following:

> *When the duration of a flight from LAX to NRT during the week from 12/26/2021 to 1/1/2022 increases by 1 minute, the price of said flight also increases by about $0.17, and vice versa.*

This result is significant because it can help us to extrapolate a hypothesized ticket price for a LAX to NRT flight that is one-hour in duration.

Conclusion

Using the Weighted Average Slope and Weighted Average Intercept found, we can conclude the following formula:

*Price = 0.16779971343267222 * Time + 504.1610748187546*

Using this formula, we can extrapolate that the price for a 60-minute flight from LAX to NRT is $514.23.

From common sense, we know that duration of a flight should be negatively correlated to the price of the ticket. This means that people should be willing to pay more for shorter flights, and less for longer flights. However, in this paper we found the slope to be a positive number, which is an unreasonable result. This can be due to several reasons. One, the sample size is too small. The dataset only contains 86 entries, and after eliminating outliers only 71 entries remained. These 71 entries are representing 10 airlines, which means on average each airline is only represented by 7.1 data entries. Since both the Weighted Average Slope and Weighted Average Intercept are calculated on an airline-specific basis, this small sample size can inflict significant skew to the data. Second, the sample is taken during the week starting from the day after Christmas. Christmas is a major holiday and consumer behaviors around the holiday might be different than the average consumer behavior. Lastly, the Overall Correlation of the dataset is about 0.09855. We can calculate R^2 by squaring the Overall Correlation, arriving at the value 0.00971. This means that only 0.971% of the price of a LAX to NRT flight ticket from 12/26/2021 to 1/1/2022 is due to its duration, a very weak correlation.

Work Cited

kayak.com. “Lax to NRT, 1/1.” *KAYAK*, 2021, https://www.kayak.com/flights/LAX-NRT/2022-01-01?fs=cfc&sort=duration_a.