Analysis of Net Gain across Different Factors

Report By: Samikshya Pandey

Executive summary

The average net gain is different for flights is affected by flights that departed late. We can statistically confirm that average net gain is different from flights that departed late compared to those that did not. The test was significant for both test where the flights departed late and for flights that departed late by more than 30 minutes.

The top 5 destinations for flights from New York as IAH, ORD, SFO, ALX and DEN respectively. While the individual average net gain and 95% confidence interval for the average net gain for flights for these 5 destinations are discussed later, we can claim that that average net gain for the top 5 flights range from 6.5 to 8.7 minutes.

The average net gain per hour also statistically differs for flights that departed late. This is true for both cases of flights delay (if they were delayed by 30 minutes or by more than a minute). The average net gain per hour (calculated by dividing net gain over air time in hours) has different average value than flights that did not depart late compared to flights that did not depart late. Similarly, from the t test, we could not statistically claim that the average net gain per hour was same for flights that departed late by 30 minutes compared to those that did not depart late by 30 minutes.

The average net gain per hour differs for short and long flight where short flight is defined as flight length of less than 3.5 hours and long flights are flight longer than 3.5 hours. from the T test difference in mean, we could not statistically claim that the average net gain per hour was same for shorter and longer flights.

Data Preparation

Dataset description

The dataset includes the flight information of New York flights departing from airport like EWR, JFK and LGA to various destinations in the United States for the year 2013. It also includes carries like United Airlines, Delta Airlines, Alaska Airlines among others. This project evaluates only the relationship between United Airlines alias UA.

Data variables used and relationships evaluated

The following paper investigates the relationship between net gain and factors such as delay length: departure delay > 0 (indicated by variable late) and departure delay > 30 (indicated by variable very_late). It also analyzes the top 5 destinations of flights from New York and their distribution of average net gain and their confidence interval. It also investigates the distribution of net gain per hour (calculated by dividing net gain by air time in hours) and their relationship between late and very late variables as well as the impact of net gain per hour with the length of flights.

Variables calculated outside of those in the dataset:

Late: Where departure delay > 0

Very_late: Departure delay > 30

Net gain: dep_delay – arr_delay

Hours_flight : air_time/hour

Gain_hour: net_gain/hours_flight

Flight_length:

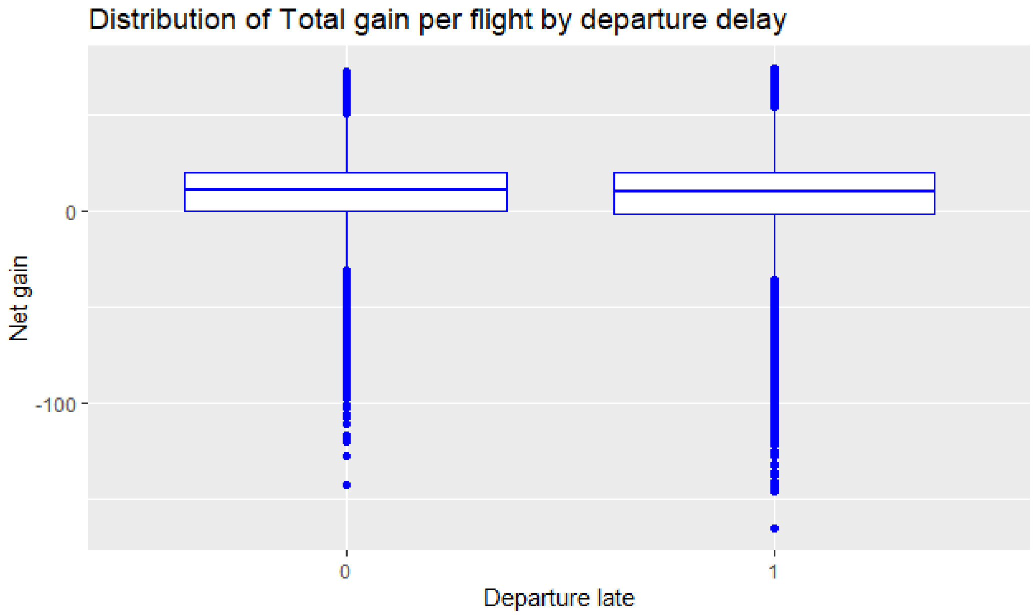
Short: hours_flight <3.5 hours

Long: hours_flight >3.5 hours

Average gain for Flights that departed late:

For flights who departed late (dep_delay >0)

First let’s visualize the distribution of net_gain based on if the flights were late (departed late) or not:



From the figure above, the distribution of the net gain seems similar regardless of the flights delay. Both box plots have values concentrated in 0-25 scale. We can observe for flights(regardless of whether there was a departure delay or not), there are outliers for the net gain value. We can, therefore, infer that state of departure of flight may not impact the net gain of flight.

To confirm the relationship between net gain and departure delay, lets conduct a hypothesis test that tests if the average net gain differs by whether the flight was delayed or not.

Using hypothesis test:

Null hypothesis: The average net gain is same regardless of flight departure delay.

Mean net gain (for departure delay > 0) = mean net gain(for dep_day <0)

Alternative Hypothesis: The average net gain is different for flights depending upon whether they were delayed or not.

Mean net gain (for departure delay > 0) != mean net gain(for dep_day <0)

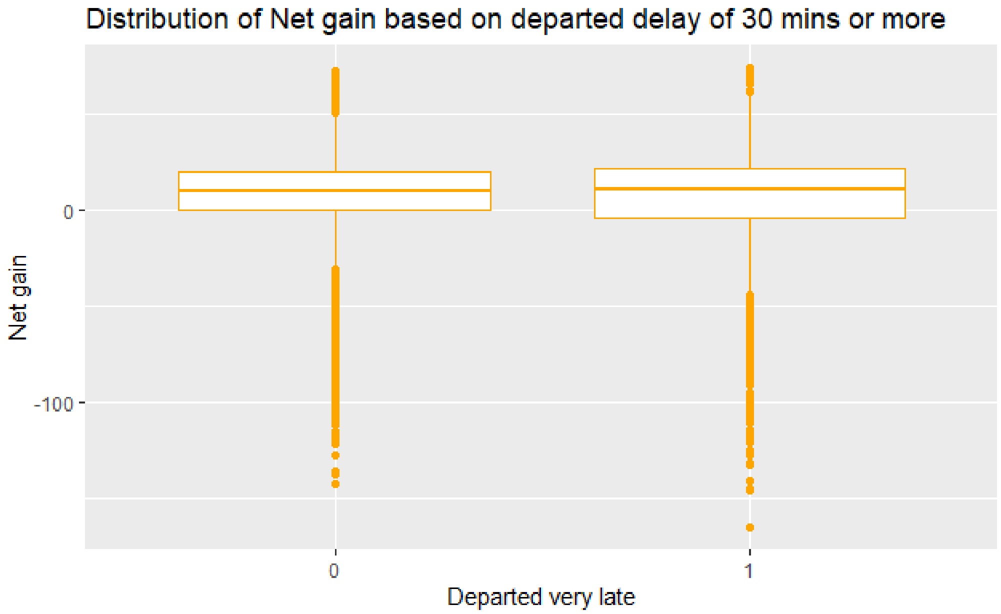
Analysis:

The T test conducted for two sample test shows that the t score id 10.749 with52833 degrees of freedom.

As the p value is really small(<0.05), we can reject the null hypothesis with statistical confidence in favor of alternative hypothesis. We can, therefore, infer that average net gain is affected by whether the flight was delayed or not.

For flights who departed very late (dep_delay > 30)

First, let's visualize the distribution of the net gain based on if the flights were delayed by 30 minutes or more.



From the box plot above, we can see the flights that departed late by more than 30 mins had more distribution for net gain than those that were not late by 30 minutes. It may seem like the distribution of the net gain is impacted by whether the flights were delayed by 30 minutes or more.

We can conduct statistical test to determine the relationship between average net gain for flights that departed 30 minutes late.

Using hypothesis test:

Null hypothesis: The average net gain is same regardless of flight departure delay.

Mean net gain (for departure delay > 30) = mean net gain(for dep_day <30

Alternative Hypothesis: The average net gain is different for flights depending upon whether they were delayed or not.

Mean net gain (for departure delay > 30) != mean net gain(for dep_day <30)

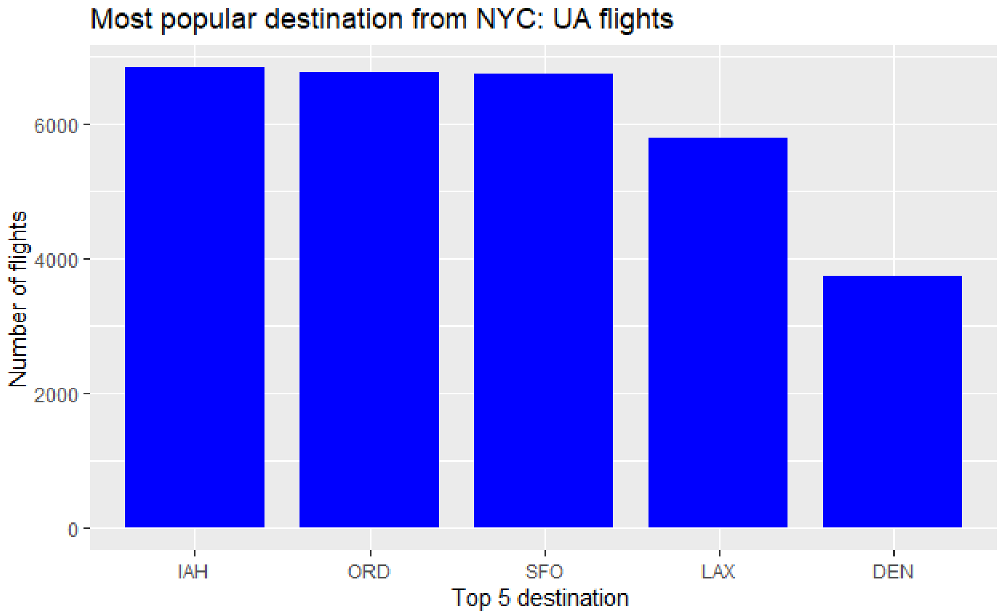
Analysis: It a two-tail t test. From the calculations, we can infer that the average net gain is different for flights that departed more than 30 minutes late compared to those that did not. With the p value of < 0.05, we can reject the null hypothesis that the average net gain is same for flights regardless of whether they were delayed by 30 minutes or not.

Most popular destinations from New York and their distribution of average gain

The 5 most common destination airports for United Airlines flights from New York are:

Destination Airports	No. of flights to destination from NY	% of flight by total flight
Houston, Texas (IAH)	6814	11.79%
Chicago O'Hare, Illinois (ORD)	6744	11.67%
San Francisco, California (SFO)	6728	11.64%
Los Angeles, California (LAX)	5770	9.98%
Denver, Colorado (DEN)	3737	6.47%

Therefore, the top 5 destinations listed above account for 51.55% of total flights from NY for UA flights for the year 2013.



The bar chart above showcases the most popular destination from NYC for UA flights.

Let's look at the distribution and the average gain for each of these five airports. To do so, we need to calculate the confidence interval for these destinations. The table below lists the average net gain for flights from New York to the top 5 popular destination. The confidence interval row tells the range within which we can find the true net gain for each of these destinations with 95% confidence. For example, in the first row, the average net gain for flights to IAH is 6.86 minutes. We can also state with 95% confidence that the average net gain lies within the range of 6.42 minutes to 7.29 minutes.

Airport	Average net gain	Confidence interval
IAH	6.86	6.42 – 7.29
ORD	7.78	7.32-8.23
SFO	8.69	8.16-9.23
LAX	7.83	7.26-8.39
DEN	7.30	6.66-7.95

While the individual average net gain is listed on the table, we can say that for the top 5 flights destinations the average net gain ranges from 6.5 to8.7 minutes.

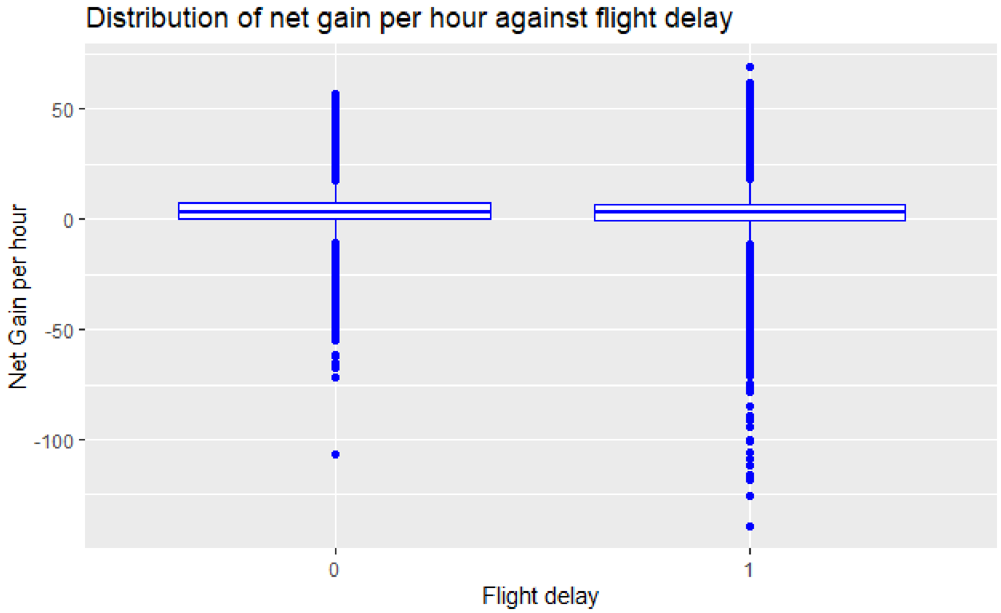
Total gain relative to duration of flight and their relationship with flights that departed late

For flights who departed late (dep_delay > 0)

To analyze this, we need to first create two variables. First is the duration of flight in hour (hours_flight) This is calculated by dividing total airtime in minutes by 60. i.e. (air_time/60).

Next, the gain_hour which calculates the total net gain per hour, so the value is net_gain/hours_flight.

Visualizing: How goes gain per hour differ for flights that departed late



Values for gain_hour seems to be concentrated near 0 and they are not widely dispersed, The q1:q3, max min are all close. Outliers exists in both cases; and there seems to be more outliers in the negative gain hour. The distribution of gain hour does not seem to be highly affected by whether there was a departure delay or not.

Using T test to statistically test the relationship between net gain per hour and flights delay:

Null Hypothesis: There is no difference in net gain per hour regardless of whether there was a departure delay.

> mean of net gain per hour when departure delay < 0 = mean of the net gain per hour when departure delay > 0

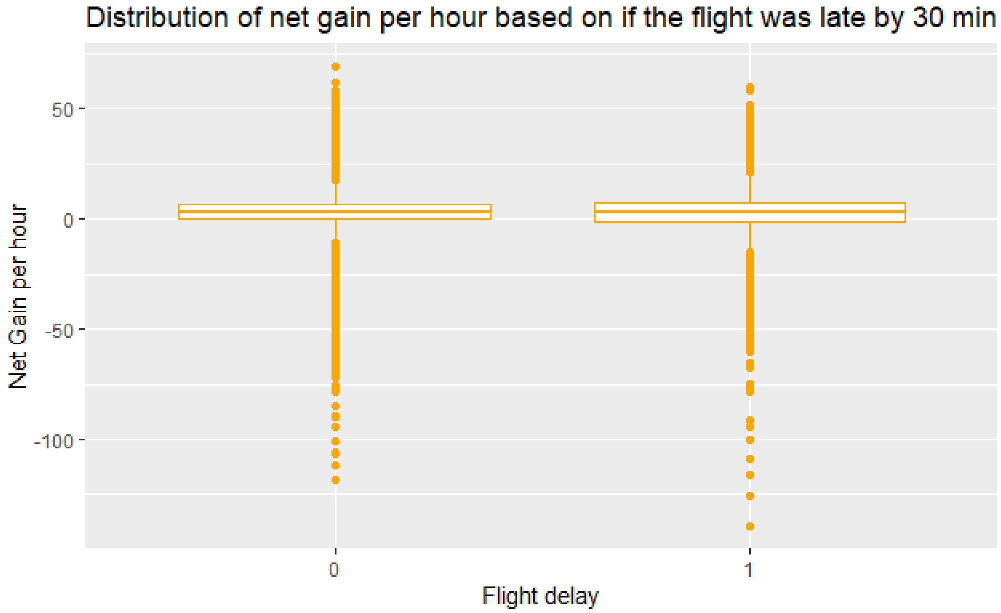
Alternative hypothesis: The mean of the net gain per hour is different based on whether there was a departure delay in flights.

Mean of net gain per hour when dep_delay > 0 <= != mean of net gain per hour when dep_delay < 0

Analysis

From the two-sided t test, we get the p value of less than 0.05. This means that we can statistically claim that there is true difference in mean between two groups. If the null hypothesis was true, then the probability of observing the mean difference of 0.662 (Average net gain when dep delay >0 and average net gai when dep delay is less than 0) is less than 0.05. So, we can reject the null and state that, we do not have enough statistical evidence to claim that there is departure delay has no impact on mean net gain per hour.

For flights who departed very late (dep_delay > 30)



Similar to the late case, even for flights that were delayed by more than 30 minutes, there are observable differences in the distribution of the net gain per hour. The graph does appear to show that there are more outliers in cases where flight departure delay was less than 30 minutes, i.e where value of flight delay = 0 and the width of the boxes differ indicating difference in the distribution of net gain per hour value. We can, therefore, hypothesize that the average net gain per hour may differ for flights that were delayed by 30 mins or more.

Hypothesis test using two tail t-test through hypothesis test: where the null and alternative hypothesis are described below:

Null hypothesis: There is no difference in net gain per hour regardless of whether there was a departure delay of 30 minutes

mean of net gain per hour when departure delay < 30 is same as the mean of the net gain per hour when departure delay > 30

Alternative hypothesis : The mean of the net gain per hour is different based on whether there was a departure delay in flights.

> Mean of net gain per hour when dep_delay >30 <= != mean of net gain per hour when dep_delay < 30.

The t test gives the p value of less than 0.05. Similar to previous conclusion, as the p value is really small, we can reject the null hypothesis. We do not have enough statistical evidence to conclude that the mean net gain per hour is same regardless of whether the flight was delayed by more than 30 minutes.

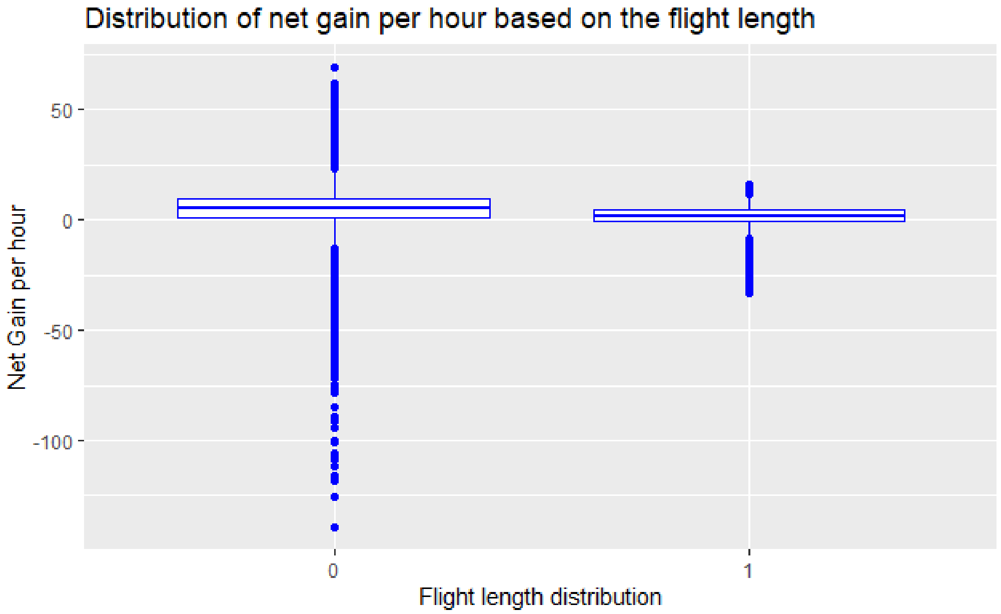
Average gain per hour for shorter and longer flights

Min.	1st Qu	Median	Mean	3rd Qu	Max
0.3833	2.2500	3.2833	3.5299	5.2167	11.5833

The table above outlines that distribution of flight length in hours. So, from the range we can see that the lowest flight value is 0.383 hours, and the mean flight length is 3.53 hours. The longest flight is 11.58 and the 3rd quartile is at 5.2. So, we can divide it into two groups at mean.
Therefore, lets create a new variable where

Short flights are flights with flight length of less than 3.5 (New variable “hours_flight”will take value 0)
Longer flights are flights that are > 3.5. ((New variable “hours_flight” will take value 1)

Visually representing the distribution of flight length distribution across net gain per hour.



From the box plot above, we can see that flight length that have are shorter than 3.5 have more spread and distribution as well as outliers.
<u>**Hypothesis test:**</u> Conducting a t test to see if the difference in net gain per hour differ for short and long flight.
Null hypothesis: The average net gain per hour is same regardless of flight length.

The average net gain per hour for flight length of 3.5 = average net gain per hour for flight length > 3.5

Alternative hypothesis: The average net gain per hour is different for shorter and longer flight.

The average net gain per hour for flight length of 3.5 != average net gain per hour for flight length > 3.5

Analysis

From the t-test, the p value is less than 0.05. So we can reject the null in favor of alternative. This means that we cannot statistically claim that the average gain per hour is same for both shorter and longer flights.