# Planning Your Air Travel - a statistical analysis of airlines in U.S.

Sahar Tariq

# Introduction

**Goal:**

1-Rank American domestic airlines based on arrival delays

2- Effects and causes of delays

**Scope:**

- Timeframe: 1 year
- 8 most popular airlines

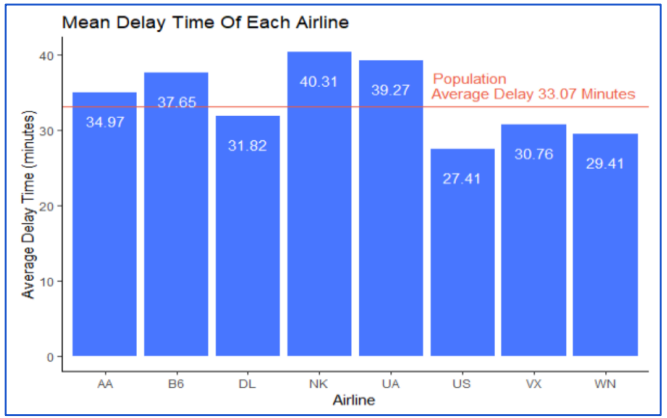
**Datasets used:**

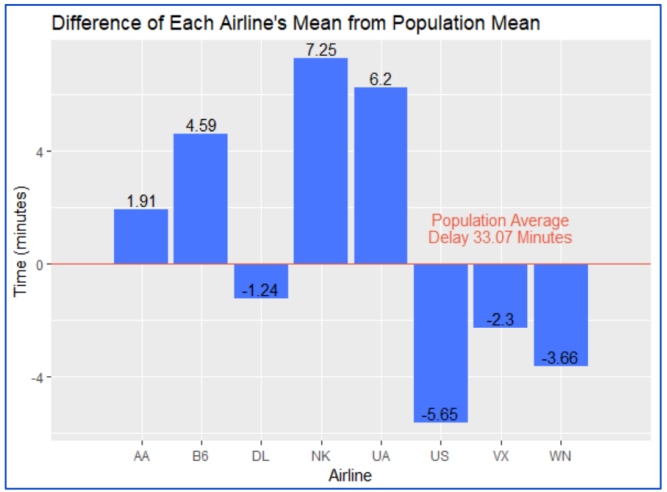
- airplane delays
- airline ticket price
- daily weather
- sample survey

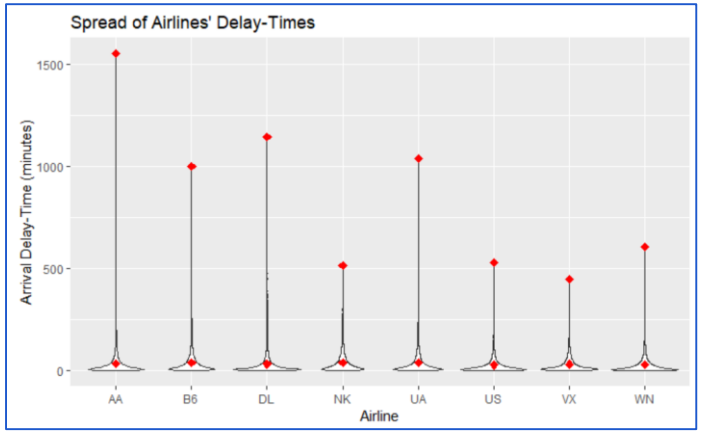
| AIRLINE CODE | AIRLINE |
|---|---|
| **AA** | American Airlines |
| **B6** | JetBlue |
| **DL** | Delta |
| **NK** | Spirit Airlines |
| **UA** | United Airlines |
| **US** | US Airways |
| **VX** | Virgin Airlines |
| **WN** | Southwest Airlines |

# Compare Airlines by Expected Delay Time

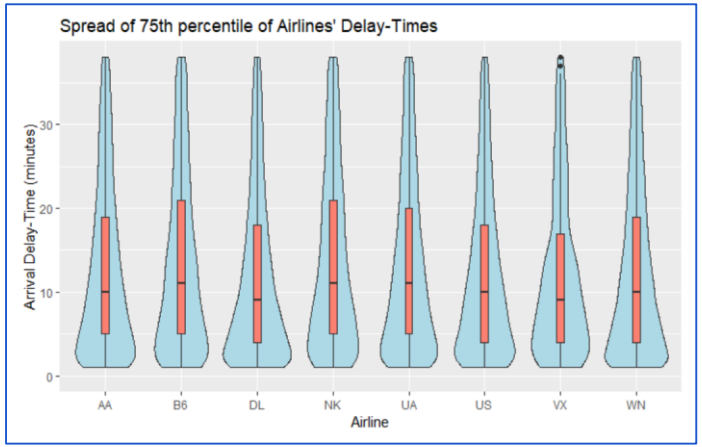
Concepts:
Mean, min, max, variance, quartiles, boxplot, histogram




*Spread: View max, average, and distribution.*

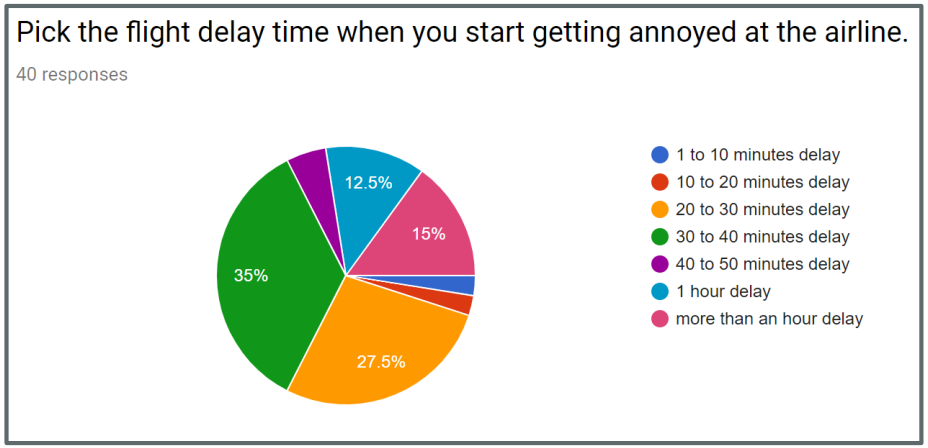



*Spread: mean is skewed to the right due to max*

# Customer Study – Sampling

Concepts: Sampling, population, confidence interval

Results of survey:




Simple random sample
size =40

**62.5% responded delay time between 20-40 minutes**

Is our sample size good view of population ?

- The normal approximation of X-bar is generally good if n>=30, according to Central Limit Theorem.
- Mean of x bar remains same for any sample size

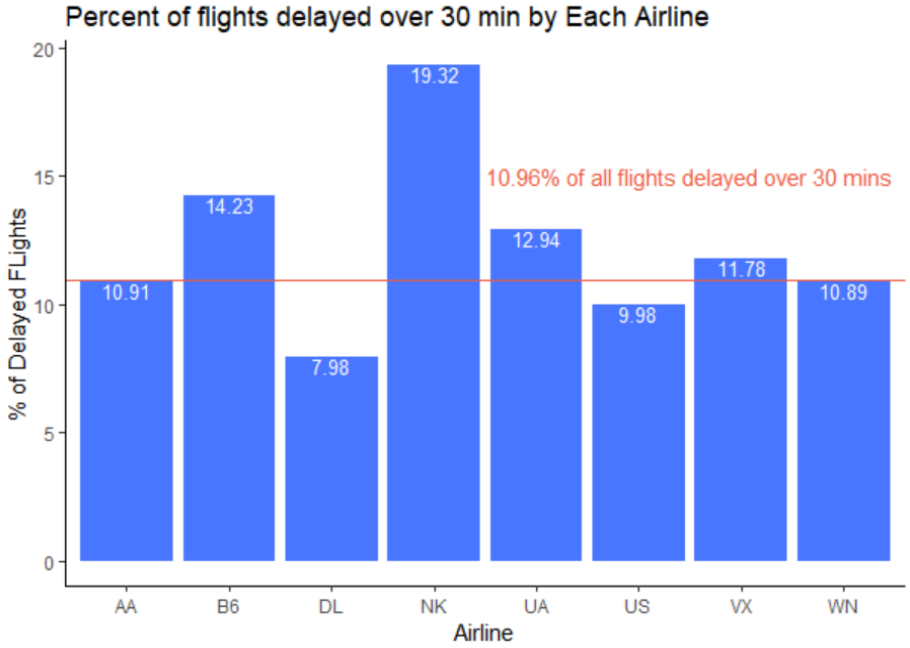
**95% confident that 62.5% ± 15% of the population has delay time threshold of between 20-40 minutes. (using what method?)**

Would need a sample of 90 people to have a tighter error of 10%, and 360 people for error 5%.

*Midpoint of 20-40 minute range = 30 minutes*

# Compare by Percent of Flights Delayed

Concepts:
PMF, Probability, test of hypothesis




Spirit Airlines is worst performer with highest percent of delayed flights, and highest mean delay time.

```
        Welch Two Sample t-test

data:  DL30$ARRIVAL_DELAY and NK30$ARRIVAL_DELAY
t = 2.6403, df = 6819, p-value = 0.008302
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.003572 6.789731
sample estimates:
mean of x mean of y
 88.18921  84.29256
```

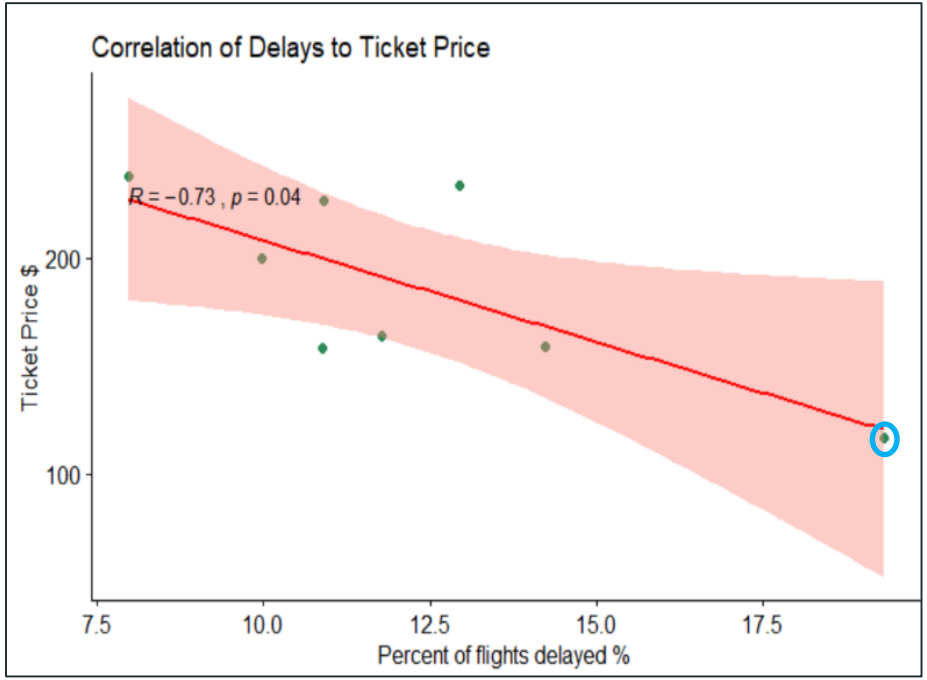
H0: $\mu_1-\mu_2=0$
H1: $\mu_1-\mu_2\neq0$

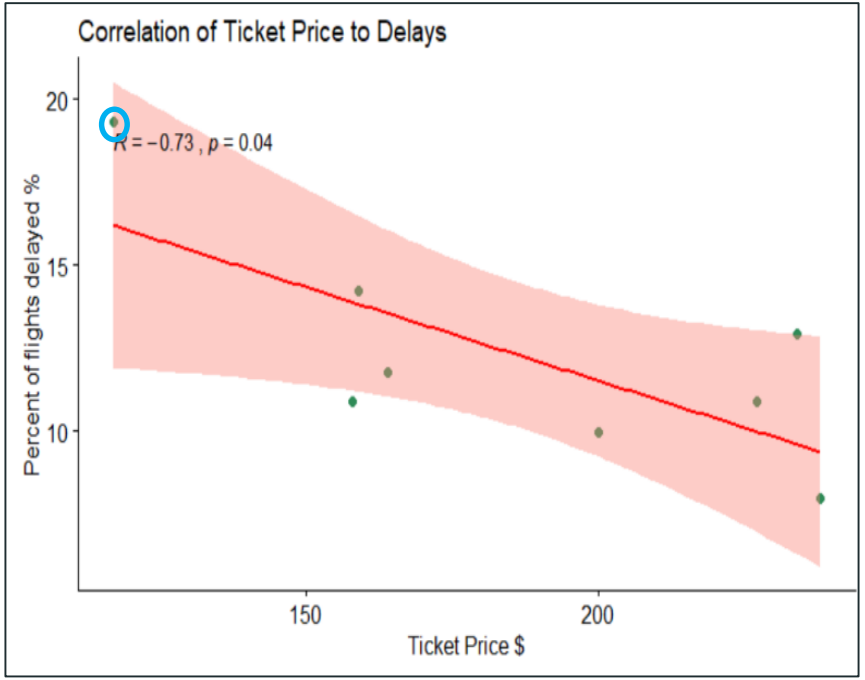
P-value less than 0.05 = reject the null hypothesis.

Significant difference between best and worst performer.

# Correlation of Airline Delays to Price:

Concepts: Correlation coefficient, trendline



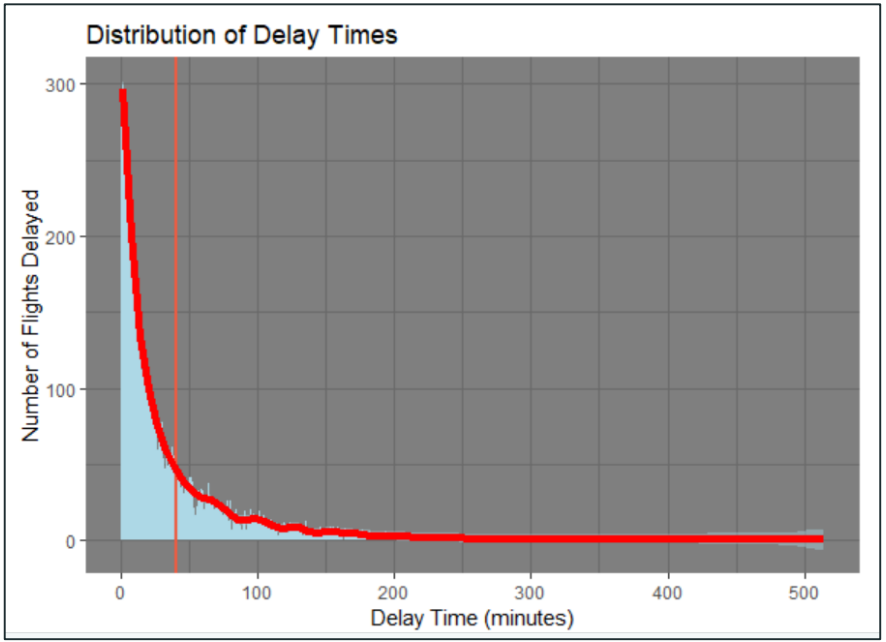

***P value : 0.04***
***Pearson Coefficient : -0.728***
***Kendall Coefficient: 0.428***

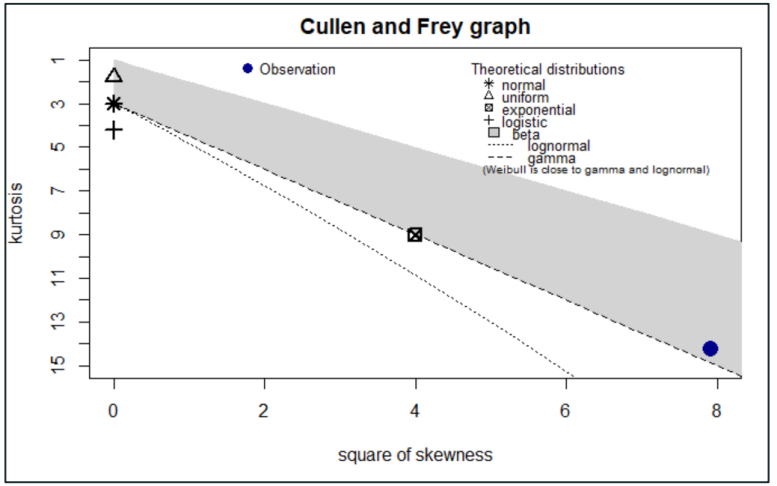
The relationships are dependent and statistically significant :

- p value is below 0.05
- kendall coefficient is not near zero
- good linear relation since R value is more than 0.7
- Inversely proportional relation since R value is negative

# Spotlight : Airline with Worst Delays

Concepts: Distribution, best fit




```
summary statistics
------
min:  1   max:  514
median:  21
mean:  40.31225
estimated sd:  52.85971
estimated skewness:  2.812434
estimated kurtosis:  14.23926
```

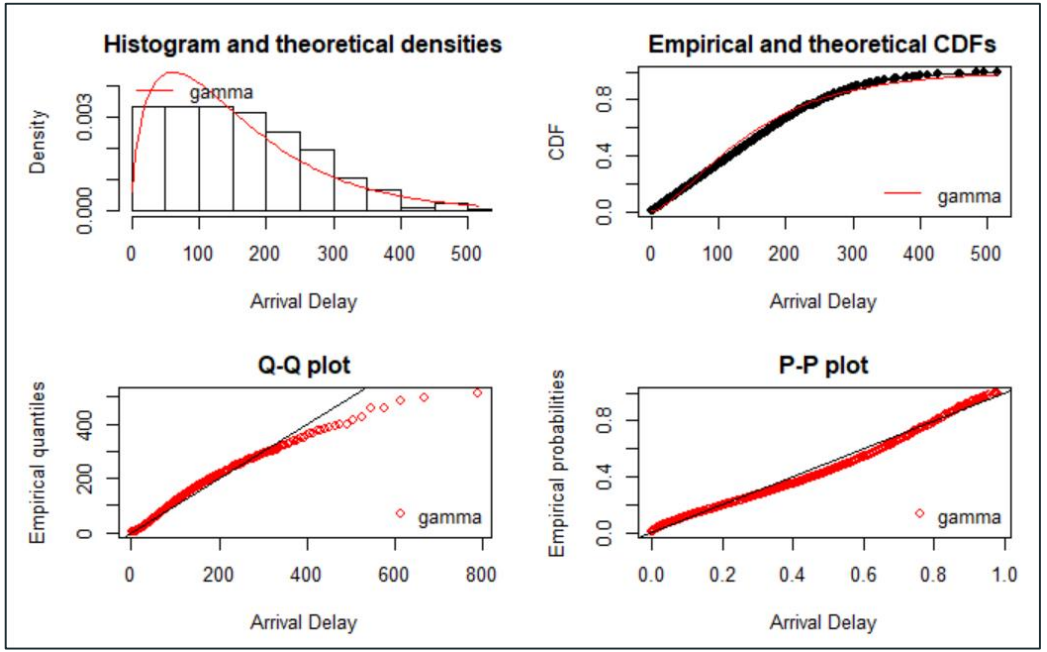
Positive kurtosis and positive skewness = values away from the mean occur more than those to the left of the mean.

*Statistically equivalent to lognormal or gamma distributions*

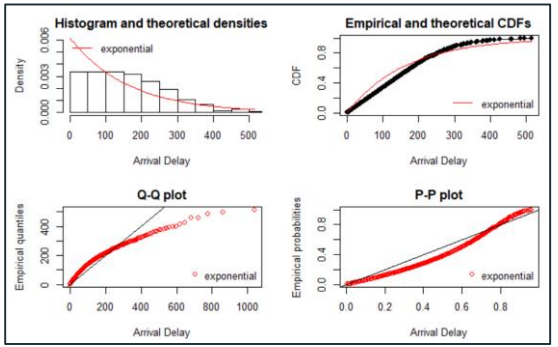
# Best Fit

Concepts:
Best fit

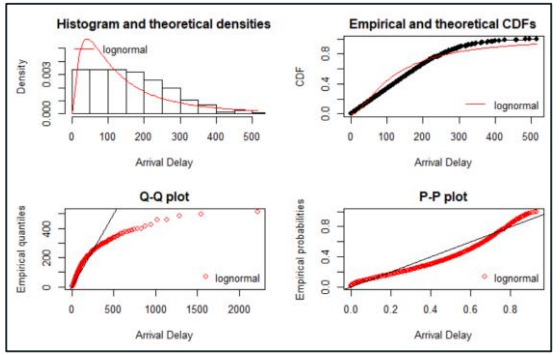
Criteria of best fit: highest loglikelihood, lowest AIC & BIC




| | |
|---|---|
| loglik | -1790.884 |
| aic | 3585.767 |
| bic | 3593.155 |




| | |
|---|---|
| loglik | -1808.817 |
| aic | 3619.635 |
| bic | 3623.329 |




| | |
|---|---|
| loglik | -1834.278 |
| aic | 3672.556 |
| bic | 3679.943 |

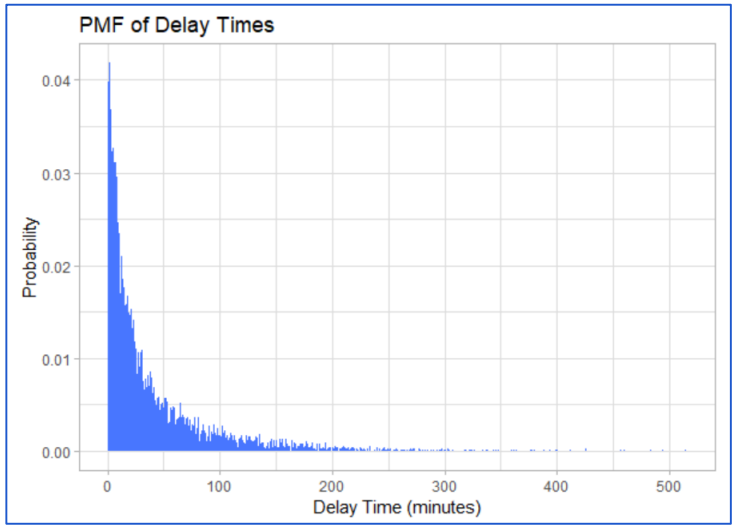
# Probabilities

Concepts:
PMF, probability distributions

## Gamma Distribution

$$x \sim \gamma(\alpha, \beta)$$
$$mean = \alpha \times \beta \qquad variance = \alpha \times \beta^2$$
$$P(x > k) = 1 - pgamma(q = k, shape = \alpha, scale = \beta)$$

The probability of a flight being delayed more than 30 minutes: 40.74%
The probability of a being delayed more than 1 hour: 22.53%




## Poisson Distribution

average rate of 19.36% delays per day
X~p(x;19.36)

Probability of more than each % of flights being delayed over 30 mins per day:

| | % of Flights Delayed | Poisson Probability |
|---|---|---|
| 1 | 5 | 9.998830e+01 |
| 2 | 10 | 9.848095e+01 |
| 3 | 15 | 8.077371e+01 |
| 4 | 20 | 3.842494e+01 |
| 5 | 25 | 8.592409e+00 |
| 6 | 30 | 8.926705e-01 |
| 7 | 35 | 4.571476e-02 |
| 8 | 40 | 1.237296e-03 |

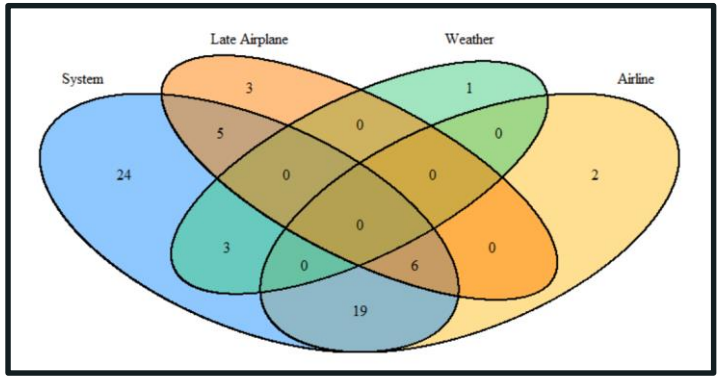
# Delay Reasons in Boston

Concepts: Venn, conditional probability, proportion

## By Number of Flights:

63 Spirit flights from Boston delayed over 30 mins

### Venn Diagram of Delay Reasons




- System delay = 57 flights
- Airline delay= 27 flights
- Late Aircraft delay=14 flights
- Weather delay=4 flights

- System ∩ Airline = 25
- System ∩ Late Airplane =11
- System ∩ Weather = 3
- System ∩ Airline ∩ Late Aircraft = 6

Conditional Probabilities

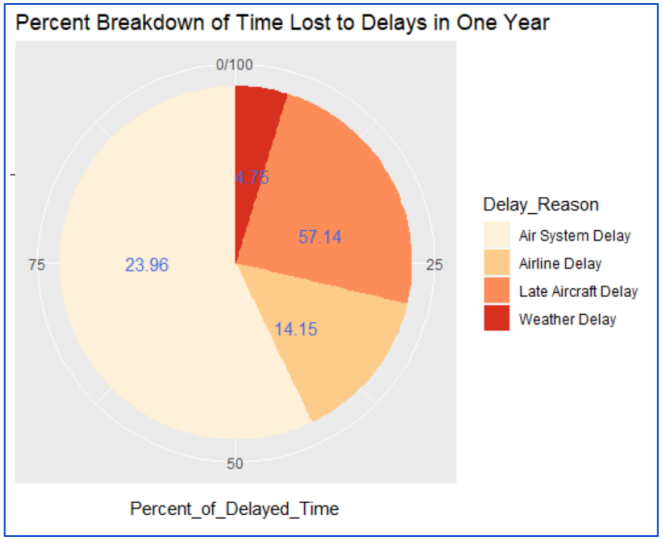
A= *event that flight delayed due to weather*

B= *event that flight delayed due to system*

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = 0.75$$

## By Delayed Hours:

80.6 hours of arrival delays in one year




| Dates of weather delay | Temperature on that day (degree F) |
|---|---|
| **January 1st** | 16 |
| **February 12th** | 11 |
| **February 14th** | 17 |
| **December 28th** | 28 |

# Conclusions

1. Spirit Airlines has highest number of flights delayed
2. You get what you pay for : statistically negative linear correlation between ticket price and delays
3. Delay times follows a gamma distribution, with decreasing number of flights having each delay time
4. 41% probability of flight being delayed over 30 mins
5. 8.6% probability that one-fourth of the flights in a day will be delayed over 30 mins
6. 57% of all delayed flights are due to air system delay

# QUESTIONS?


REFERENCES

http://airlines.org/wp-content/uploads/2016/04/2016Survey.pdf

http://money.com/money/4695479/best-us-airlines-2017/
https://www.statista.com/statistics/540661/us-low-cost-carriers-average-ticket-price/

http://money.com/money/4695479/best-us-airlines-2017/

https://docs.google.com/forms/d/e/1FAIpQLSfu3DcENgb-i-R4ycEWbNclkATkIdaRBHJApI6mg_4KIoelaA/viewform

https://www.kaggle.com/usdot/flight-delays#flights.csv

https://www.kaggle.com/naveenpandianv/weather-data-boston-jul-2012-aug-2015