Brief Analysis of Airline Data Q1 - 2017

By: Kevin Butkovich
1-21-2018
https://github.com/git-kbutkovich/Airline_Study

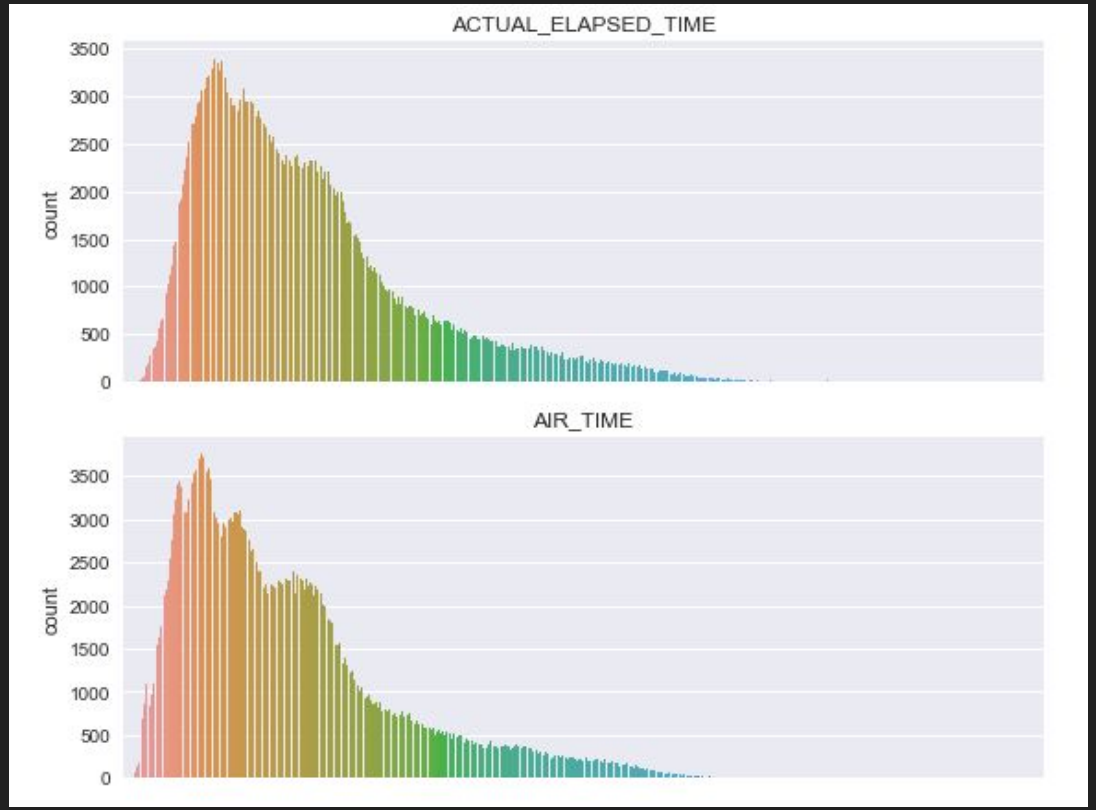
Looking at travel times

```
round(df['ACTUAL_ELAPSED_TIME'].mean(),2),\
df['ACTUAL_ELAPSED_TIME'].median(),\
df['ACTUAL_ELAPSED_TIME'].mode(),
```

```
(143.17, 125.0, 0    81.0
 dtype: float64)
```

We see a fairly normal distribution of travel times, so further statistical analysis could be performed on this feature.

Let's look a bit further into delayed departures.



Preliminary Statistics

Average number of flights delayed at least 15 minutes:

21%

Even with 79% of flights reporting no delayed departure, the average delay is 15 minutes.

```
stats_df = df.describe(include=[float]).T

skew_values = []
for i in stats_df.index:

    col_skew = stats.skew(df[i])
    skew_values.append(col_skew)

stats_df['Skew'] = skew_values
stats_df
```

	count	mean	std	min	25%	50%	75%	max	Skew
DEP_DELAY_NEW	450017.0	14.773015	45.800003	0.0	0.0	0.0	9.0	2755.0	9.747164
DEP_DEL15	441476.0	0.210820	0.407891	0.0	0.0	0.0	0.0	1.0	NaN
ARR_DELAY_NEW	450017.0	15.065384	45.291288	0.0	0.0	0.0	11.0	1944.0	9.306289
ARR_DEL15	439645.0	0.222222	0.415740	0.0	0.0	0.0	0.0	1.0	NaN
CANCELLED	450017.0	0.019746	0.139126	0.0	0.0	0.0	0.0	1.0	6.903877
CRS_ELAPSED_TIME	450013.0	147.936311	77.132200	21.0	90.0	130.0	181.0	712.0	NaN
ACTUAL_ELAPSED_TIME	439645.0	143.172273	76.701789	16.0	86.0	125.0	176.0	734.0	NaN
AIR_TIME	439645.0	118.427863	74.529136	7.0	63.0	100.0	150.0	704.0	NaN
CARRIER_DELAY	97699.0	20.308836	60.649157	0.0	0.0	0.0	17.0	1934.0	NaN
WEATHER_DELAY	97699.0	3.368857	27.809756	0.0	0.0	0.0	0.0	1934.0	NaN
NAS_DELAY	97699.0	15.742495	32.328181	0.0	0.0	4.0	20.0	1457.0	NaN
SECURITY_DELAY	97699.0	0.079745	2.896287	0.0	0.0	0.0	0.0	653.0	NaN
LATE_AIRCRAFT_DELAY	97699.0	24.947492	47.850860	0.0	0.0	3.0	31.0	1392.0	NaN
Unnamed: 29	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Determining problem statement

Q: What is the aspect of airline travel that passengers are most affected by?

A: Delays (in my opinion; not referencing a scientific study)

Looking at delayed flights - Trying to normalize our binomial situation

Isolate our DataFrame to only include flights with delay of > 15 minutes.

Skew : Very High.

Remove outliers with Tukey's method (189min).

Skew : More acceptable

```
delay_stats
```

	count	mean	std	min	25%	50%	75%	max	Skew
DEP_DELAY_NEW	93072.0	66.523326	82.035480	15.0	24.0	41.0	78.0	2755.0	6.077014
DEP_DEL15	93072.0	1.000000	0.000000	1.0	1.0	1.0	1.0	1.0	0.000000

```
no_outliers_stats
```

	count	mean	std	min	25%	50%	75%	max	Skew
DEP_DELAY_NEW	88184.0	52.627404	38.809445	15.0	24.0	39.0	69.0	189.0	1.416098
DEP_DEL15	88184.0	1.000000	0.000000	1.0	1.0	1.0	1.0	1.0	0.000000

Finding confidence intervals (before & after outlier removal)

Notice how the confidence interval becomes more precise with outliers removed.

```
def clean_data(data_list):
    clean_ls = []
    for i in data_list:
        if type(i) == int or type(i) == float:
            clean_ls.append(i)
    return np.array(clean_ls)

def confidence_interval_of_means(distribution, confidence):
    clean_distribution = clean_data(distribution)
    dist_mean = np.mean(clean_distribution)
    alpha = 1-confidence
    z = stats.norm.ppf(alpha/2)
    stderr = np.std(clean_distribution) / np.sqrt(len(clean_distribution))
    lower_bound = dist_mean + z*stderr
    upper_bound = dist_mean - z*stderr
    return lower_bound, upper_bound
```

Before: Δ = 1.054 min

Before outliers removed:

```
confidence_interval = confidence_interval_of_means(delay_test, 0.95)
print('The 0.95 confidence interval of departure delay (in minutes) is {}'.format(confidence_interval))
```

```
The 0.95 confidence interval of departure delay (in minutes) is (65.996292845519818, 67.050359208803727)
```

After: Δ = 0.513 min

After outliers removed:

```
confidence_interval = confidence_interval_of_means(outliers_test, confidence=0.95)
print('The 0.95 confidence interval of departure delay (in minutes) is {}'.format(confidence_interval))
```

```
The 0.95 confidence interval of departure delay (in minutes) is (52.37125771997102, 52.883550408487658)
```

Final Inferences

- As an airline passenger, you only have a 21% chance of your flight being delayed.
- IF your flight is delayed:
  - Your most likely delay time (mode): 15 minutes
  - Your median delay time: 52.63 minutes
- With a large enough sample size, we can say, with 95% confidence, that the median delay time will be between 52.37 and 52.88 minutes.
- These numbers are derived with outliers removed. The outlier threshold for this exercise is 189 minutes (3 hours, 9 minutes)
- Next step: Download Q2-2017 to determine mean of delay time.
- Further Analysis: What are the biggest causes related to delays?