The challenge is based on a dataset containing information about flights, including details of whether flights were on-time, early, or late. Your goal is to explore the data to identify features that might be predictive of how many minutes late or early a flight will be.

In this challenge, you will work with some built-in sample datasets in Azure Machine Learning. The datasets contain details of flights, including an indication of how many minutes late or early the flight arrived at its destination.

In this first part of the challenge, you will use Azure Machine Learning to clean the data and prepare it for exploration.

How many rows are in the dataset? Columns:29

Initially 7009728

After duplicates 7009724

Remove columns 'CancellationCode','CarrierDelay','WeatherDelay','NASDelay','SecurityDelay','LateAircraftDelay' because a lot of data is missing (see excel)

6855020 afterwards

Display statistical data for the columns. After examining data cancelled and diverted also eliminated

Your eventual goal is to build a model that predicts the **ArrDelay** value. A useful starting point is to understand the range and distribution of values for this value.

To further explore the range and distribution of values in the **ArrDelay** column, you must now create a plot that shows a box plot and a histogram of this value. The histogram should display the values in 30 bins.

Based on the data visualization, which three of the following statements accurately reflect the distribution of **ArrDelay** values?

The median, first quartile, and third quartile are all fairly close to 0, indicating that most flights arrive close to their scheduled time.

The range of arrival times ranges extensively, with some flights arriving as much as 1500 minutes late.

The distribution is right-skewed, so there is a higher range of values for late flights than for early flights.

The flights dataset includes a number of numeric features (for example **DepDelay**, which indicates the number of minutes late a flight departed) or psuedo-numeric features (for example **CRSDepTime**, which indicates the scheduled departure time as a whole number in 24 hour clock format). To explore how these values might be related to arrival delay, you will plot histograms conditioned by the **ArrDel15** column, which is a binary column indicating whether a flight arrived 15 or more minutes late.

Write code to generate conditioned histograms for the following columns, conditioned by the **ArrDel15** column:

* **DepDelay**
* **CRSArrTime**
* CRSDepTime
* DayofMonth
* DayOfWeek
* Month

Based on the conditioned histograms, which three of the following statements are true?

There are significantly more flights that are less than 15 minutes late than there are flights that are 15 minutes late or more.

Flights that are 15 minutes or more late tend to have a higher **DepDelay** value than flights that are on-time.

Late flights tend to occur more frequently for flights with a **CRSArrTime** that is later in the day, the highest volume of delayed flights scheduled to arrive between 3pm (1500 hours) and 8pm (2000 hours)

Late flights tend to occur more frequently at the end of the month.

Taxis

TScatter plots are another useful way to compare two numeric values, and can be conditioned on one or more variables using colors and shapes.

Write code to generate conditioned scatter plots for the following columns, conditioned by the **ArrDel15** column using different colors for values of 0 and 1:

* DepDelay
* CRSArrTime
* CRSDepTime
* DayofMonth
* DayOfWeek
* Month

Based on the conditioned scatter plots, which two of the following statements are true?

There is a near-linear relatonship between **DepDelay** and **ArrDelay** for late flights. As departure delay increases, so does arrival delay.

There is an apparent relationship between **ArrDelay** and **CRSDepTime**. Flights that depart early in the morning are typically less delayed than flights that are scheduled to depart after around 5am (0500 hours), at which time delays tend to get significantly longer. Delays then gradually get shorter as the day progresses.

<http://aspmhelp.faa.gov/index.php/Types_of_Delay>

<http://prudata.webfactional.com/wiki/index.php/Additional_taxi-out_time>