**Space X Falcon 9 First Stage Landing Prediction**

**Lab 2: Data wrangling**

Estimated time needed: **60** minutes

In this lab, we will perform some Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models.

In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground pad.True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship.

In this lab we will mainly convert those outcomes into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful.

Falcon 9 first stage will land successfully



Several examples of an unsuccessful landing are shown here:

**Objectives**

Perform exploratory Data Analysis and determine Training Labels

* Exploratory Data Analysis
* Determine Training Labels

**Import Libraries and Define Auxiliary Functions**

We will import the following libraries.

In [1]:

*# Pandas is a software library written for the Python programming language for data manipulation and analysis.*

**import** pandas **as** pd

*# NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays*

**import** numpy **as** np

**Data Analysis**

Load Space X dataset, from last section.

In [2]:

df**=**pd**.**read\_csv("https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/dataset\_part\_1.csv")

df**.**head(10)

Out[2]:

|  | **FlightNumber** | **Date** | **BoosterVersion** | **PayloadMass** | **Orbit** | **LaunchSite** | **Outcome** | **Flights** | **GridFins** | **Reused** | **Legs** | **LandingPad** | **Block** | **ReusedCount** | **Serial** | **Longitude** | **Latitude** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 1 | 2010-06-04 | Falcon 9 | 6104.959412 | LEO | CCAFS SLC 40 | None None | 1 | False | False | False | NaN | 1.0 | 0 | B0003 | -80.577366 | 28.561857 |
| **1** | 2 | 2012-05-22 | Falcon 9 | 525.000000 | LEO | CCAFS SLC 40 | None None | 1 | False | False | False | NaN | 1.0 | 0 | B0005 | -80.577366 | 28.561857 |
| **2** | 3 | 2013-03-01 | Falcon 9 | 677.000000 | ISS | CCAFS SLC 40 | None None | 1 | False | False | False | NaN | 1.0 | 0 | B0007 | -80.577366 | 28.561857 |
| **3** | 4 | 2013-09-29 | Falcon 9 | 500.000000 | PO | VAFB SLC 4E | False Ocean | 1 | False | False | False | NaN | 1.0 | 0 | B1003 | -120.610829 | 34.632093 |
| **4** | 5 | 2013-12-03 | Falcon 9 | 3170.000000 | GTO | CCAFS SLC 40 | None None | 1 | False | False | False | NaN | 1.0 | 0 | B1004 | -80.577366 | 28.561857 |
| **5** | 6 | 2014-01-06 | Falcon 9 | 3325.000000 | GTO | CCAFS SLC 40 | None None | 1 | False | False | False | NaN | 1.0 | 0 | B1005 | -80.577366 | 28.561857 |
| **6** | 7 | 2014-04-18 | Falcon 9 | 2296.000000 | ISS | CCAFS SLC 40 | True Ocean | 1 | False | False | True | NaN | 1.0 | 0 | B1006 | -80.577366 | 28.561857 |
| **7** | 8 | 2014-07-14 | Falcon 9 | 1316.000000 | LEO | CCAFS SLC 40 | True Ocean | 1 | False | False | True | NaN | 1.0 | 0 | B1007 | -80.577366 | 28.561857 |
| **8** | 9 | 2014-08-05 | Falcon 9 | 4535.000000 | GTO | CCAFS SLC 40 | None None | 1 | False | False | False | NaN | 1.0 | 0 | B1008 | -80.577366 | 28.561857 |
| **9** | 10 | 2014-09-07 | Falcon 9 | 4428.000000 | GTO | CCAFS SLC 40 | None None | 1 | False | False | False | NaN | 1.0 | 0 | B1011 | -80.577366 | 28.561857 |

Identify and calculate the percentage of the missing values in each attribute

In [3]:

df**.**isnull()**.**sum()**/**df**.**count()**\***100

Out[3]:

FlightNumber 0.000

Date 0.000

BoosterVersion 0.000

PayloadMass 0.000

Orbit 0.000

LaunchSite 0.000

Outcome 0.000

Flights 0.000

GridFins 0.000

Reused 0.000

Legs 0.000

LandingPad 40.625

Block 0.000

ReusedCount 0.000

Serial 0.000

Longitude 0.000

Latitude 0.000

dtype: float64

Identify which columns are numerical and categorical:

In [4]:

df**.**dtypes

Out[4]:

FlightNumber int64

Date object

BoosterVersion object

PayloadMass float64

Orbit object

LaunchSite object

Outcome object

Flights int64

GridFins bool

Reused bool

Legs bool

LandingPad object

Block float64

ReusedCount int64

Serial object

Longitude float64

Latitude float64

dtype: object