IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Mrunal Dodhiwala
Feb 10, 2024

# Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

# Executive Summary

## Summary of methodologies

- Data Collection via API, Web Scrapping
- Data Wrangling
- Exploratory Data Analysis with SQL
- EDA with Data Visualization
- Interactive Map with Folium
- Building dashboard with Plotly Dash
- Predictive Analysis

## Summary of all results

- Exploratory Data Analysis results
- Interactive maps
- Dashboard
- Results of Predictive analysis - Classification

# Introduction

## Project background and context

SpaceX is the most successful private company in commercial space industry. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch. The aim of this project is to predict if the Falcon 9 first stage will land successfully.

## Problems you want to find answers

- Which variables affect the landing success?
- How do the variables impact the success rate?
- Which conditions will enable SpaceX to achieve the best landing rate?

Section 1

# Methodology

# Methodology

## Executive Summary

- Data collection methodology:
  - Using SpaceX REST API
  - Using Web Scrapping from Wikipedia
- Perform data wrangling
  - Handling missing values
  - One Hot Encoding
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

# Data Collection

## Data collection involved:

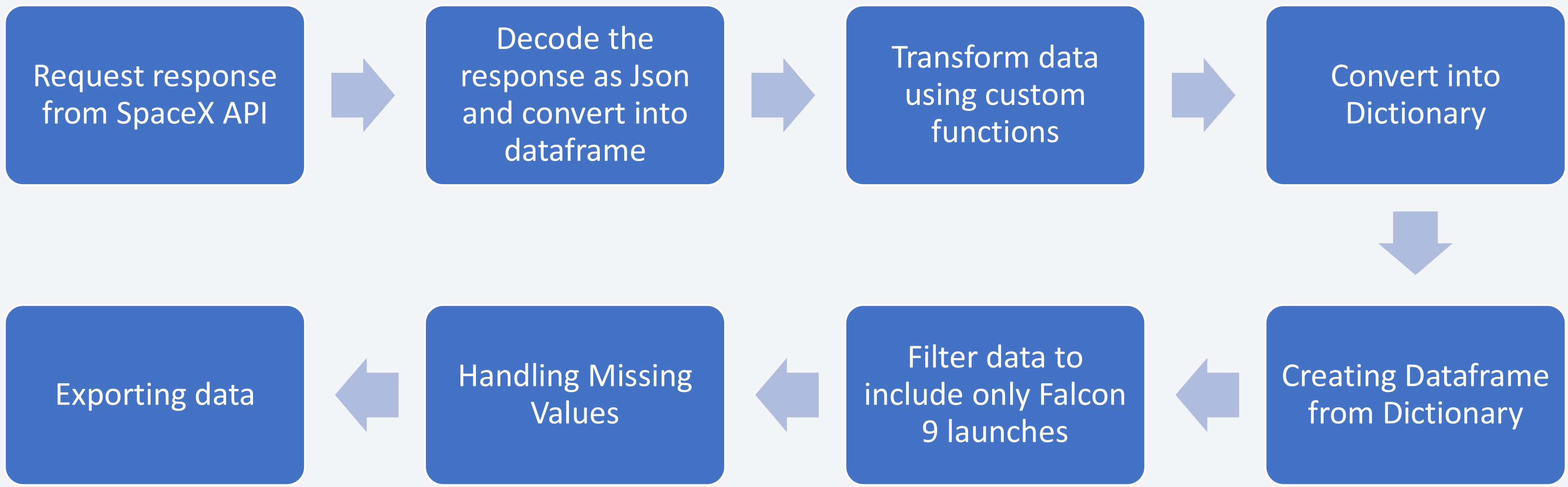
### 1. API request from SpaceX REST API

- Information collected: Flight Number, Date, Booster Version, Payload Mass, Orbit, Launch Site, Outcome, Flights, GridFins, Reused, Legs, Landing Pad, Block, Reused Count, Serial, Longitude, Latitude


### 2. Webscrapping Wikipedia

- Information collected: Flight Number, Launch site, Payload, Payload Mass, Orbit, Customer, Launch Outcome, Booster Version, Booster Landing, Date, Time

# Data Collection – SpaceX API




URL: [SpaceX API calls notebook]

# Data Collection - Scraping

1. Request response from Wikipedia
2. Create BeautifulSoup Object
3. Extracting column names from HTML table header
4. Creating Dictionary by parsing HTML tables
5. Creating Dataframe from Dictionary
6. Exporting data

URL: [Webscrapping notebook]

# Data Wrangling

- There are several cases where the booster did not land successfully.
  - True Ocean, True RTLS, True ASDS represent successful landing
  - False Ocean, False RTLS, False ASDS, None ASDS, None represent unsuccessful landing
- We converted these outcomes such that ‘1’ represents successful landing and ‘0’ represents unsuccessful landing

Calculate number of launches for each site → Calculate number & occurrence of each orbit → Calculate number & occurrence of mission outcome of the orbits → Create landing outcome label from Outcome column → Export data

URL: [Data Wrangling notebook]

# EDA with Data Visualization

### Scatter Plots

- Flight Number vs Payload Mass
- Flight Number vs Launch Site
- Payload vs Launch Site
- Flight Number vs Orbit
- Payload Mass vs Orbit

Scatter plots show the relationship between variables

### Bar chart

- Success rate of each orbit

Bar charts show comparisons among categorical variables

### Line chart

- Yearly trend of launch success

Line charts show trends in data over time

URL: [EDA with Data Visualization notebook]

# EDA with SQL

I performed the following SQL queries to explore the dataset:

- Displaying the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string 'CCA'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in ground pad was achieved
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which have carried the maximum payload mass
- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
- Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20 in descending order

URL: [EDA with SQL notebook]

# Build an Interactive Map with Folium

- Added the following markers:
  - Added marker with circle, popup label and text label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.
  - Added markers with circle, popup label and text label of all launch sites using their latitude and longitude coordinates to show their geographical locations and proximity to equator and coasts.
  - Added colored markers of success (green) and failed (red) launches using Marker Cluster to identify which launch sites have relatively high success rates.
  - Added colored lines to show distances between launch site and key locations like railway, highway, coastline and closest city.
- These objects are created in order to understand better the problem and the data. We can show easily

URL: Interactive map with Folium map

# Build a Dashboard with Plotly Dash

- Following plots/graphs and interactions are added to dashboard:
  - Dropdown - allows a user to choose the launch site or all launch sites
  - Pie chart - shows the total success and the total failure for the launch site chosen with the dropdown component
  - Rangeslider - allows a user to select a payload mass in a fixed range
  - Scatter plot components - shows the relationship between two variables, in particular Success vs Payload Mass

URL: Plotly Dash lab

# Predictive Analysis (Classification)

Creating a NumPy array from column 'Class' in data

Standardizing the data with StandardScaler, and transforming it

Splitting into training and test sets with train_test_split function

Creating a GridSearchCV object with cv=10 to find the best parameters

Applying GridSearchCV on LogReg, SVM, Decision Tree, and KNN models

Calculating the accuracy of all models using the method .score()

Examining the confusion matrix for all models

Finding the best performing method

URL: [Predictive Analysis lab]

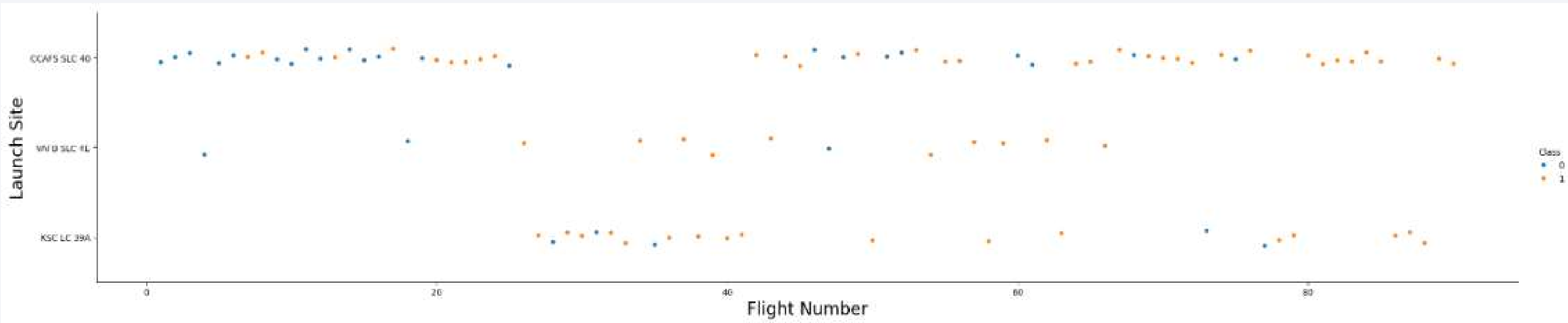
# Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

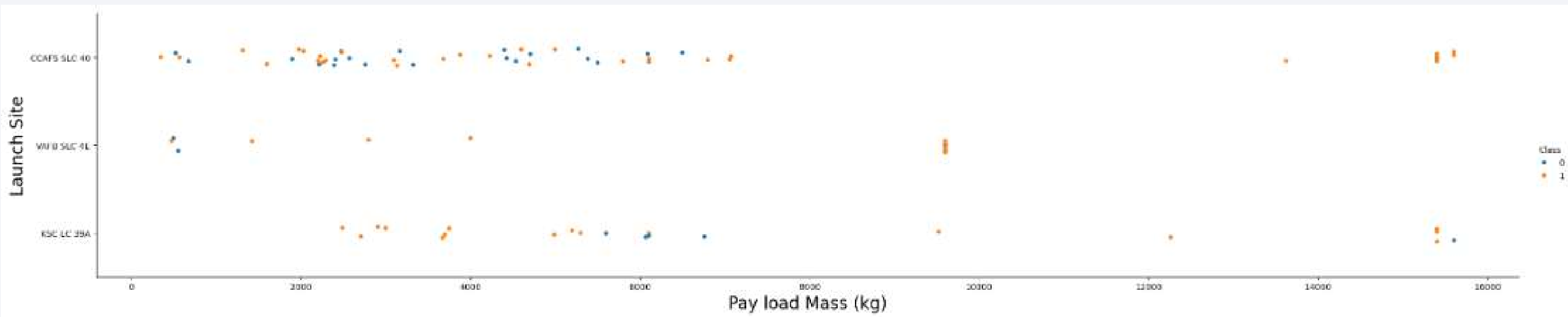
# Insights drawn from EDA

# Flight Number vs. Launch Site




- The early launches have higher failures while later launches have higher success.
- Maximum launches are from CCAFS SLC 40 launch site.
- Launch sites VAFB SLC 4E and KSC LC 39A have higher success rates.

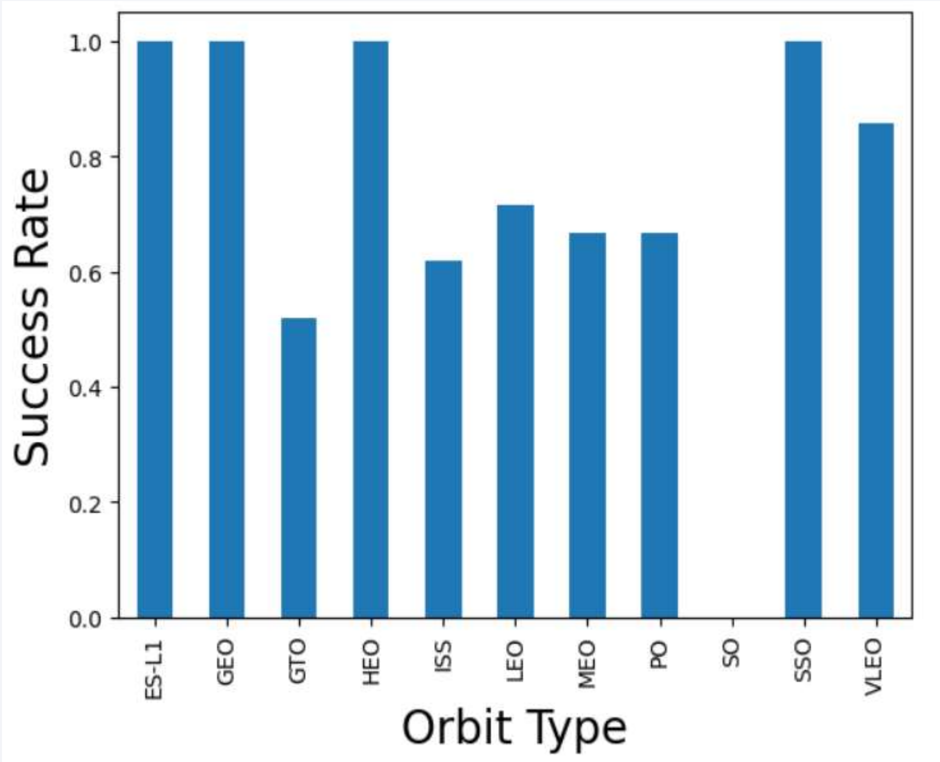
# Payload vs. Launch Site



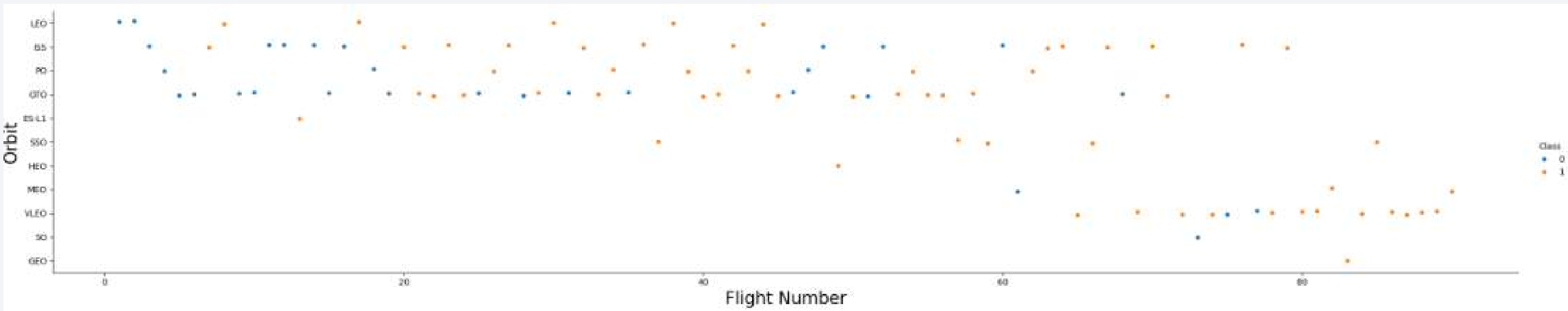

- Very high success rates for launches with payload mass over 7000 kg.
- For every launch site, the success rates are higher for higher payload mass.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg.

# Success Rate vs. Orbit Type

- Orbits ES-L1, GEO, HEO, SSO have a 100% success rate.
- Orbit SO has a 0% success rate
- The remaining orbits GTO, ISS, LEO, MEO, PO have success rate between 50% and 85%

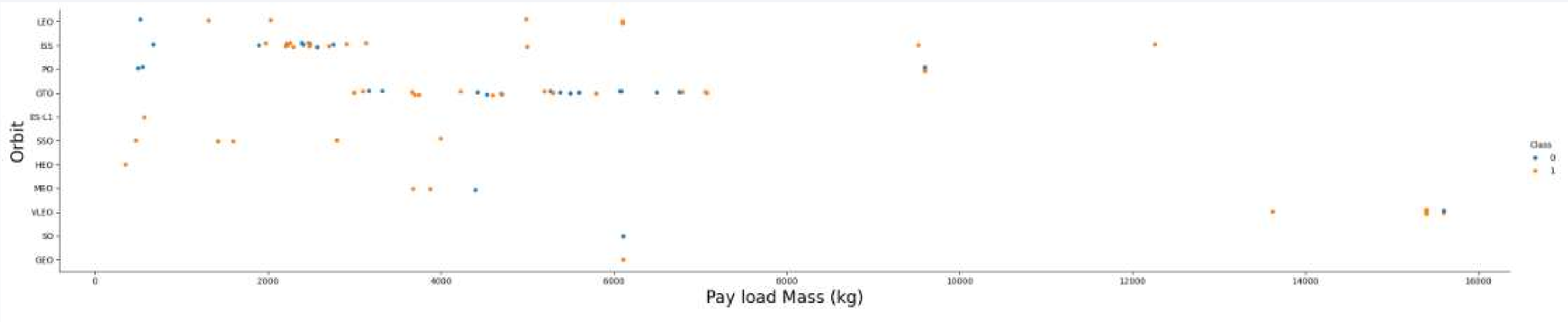


# Flight Number vs. Orbit Type




- In the LEO orbit the success rate increases with the number of flights
- In GTO orbit, there seems to be no relationship between number of flights and the success rate

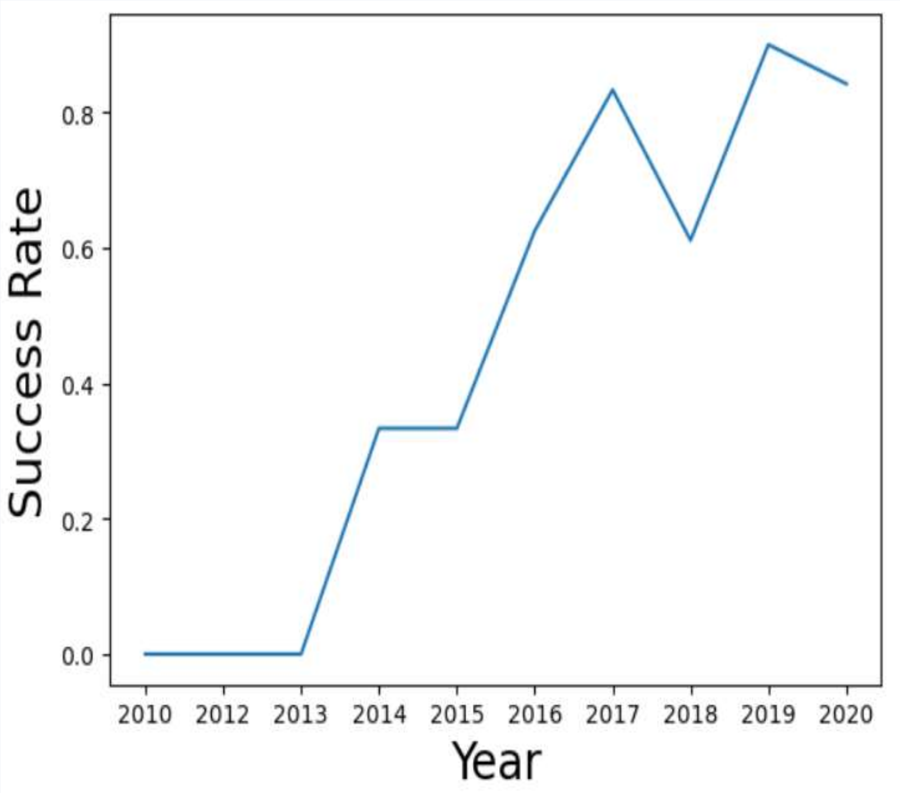
# Payload vs. Orbit Type




- For LEO orbit, heavier payload mass increases the success rate.
- For GTO orbit, lighter payload mass increases the success rate.

# Launch Success Yearly Trend

- SpaceX experienced first success in 2013.
- Since then, the success rate has shown an increasing trend.



# All Launch Site Names

- Displaying the names of unique launch sites.
- The keyword ‘distinct’ removes the duplicate entries from the result.

```
%sql select distinct launch_site from SPACEXTBL;
```

```
 * sqlite:///my_data1.db
Done.
```

| Launch_Site |
| --- |
| CCAFS LC-40 |
| VAFB SLC-4E |
| KSC LC-39A |
| CCAFS SLC-40 |

# Launch Site Names Begin with 'CCA'

- The keyword ‘like’ in the WHERE clause filters only the launch sites that contain the substring ‘CCA’.
- The keyword ‘limit 5’ shows the top 5 records.

```
%sql select * from SPACEXTBL where launch_site like 'CCA%' limit 5;
```

```
 * sqlite:///my_data1.db
Done.
```

| Date | Time (UTC) | Booster_Version | Launch_Site | Payload | PAYLOAD_MASS__KG_ | Orbit | Customer | Mission_Outcome | Landing_Outcome |
|---|---|---|---|---|---|---|---|---|---|
| 2010-06-04 | 18:45:00 | F9 v1.0 B0003 | CCAFS LC-40 | Dragon Spacecraft Qualification Unit | 0 | LEO | SpaceX | Success | Failure (parachute) |
| 2010-12-08 | 15:43:00 | F9 v1.0 B0004 | CCAFS LC-40 | Dragon demo flight C1, two CubeSats, barrel of Brouere cheese | 0 | LEO (ISS) | NASA (COTS) NRO | Success | Failure (parachute) |
| 2012-05-22 | 7:44:00 | F9 v1.0 B0005 | CCAFS LC-40 | Dragon demo flight C2 | 525 | LEO (ISS) | NASA (COTS) | Success | No attempt |
| 2012-10-08 | 0:35:00 | F9 v1.0 B0006 | CCAFS LC-40 | SpaceX CRS-1 | 500 | LEO (ISS) | NASA (CRS) | Success | No attempt |
| 2013-03-01 | 15:10:00 | F9 v1.0 B0007 | CCAFS LC-40 | SpaceX CRS-2 | 677 | LEO (ISS) | NASA (CRS) | Success | No attempt |

# Total Payload Mass

- The aggregate function ‘sum’ totals the column figure represented by (‘PAYLOAD_MASS__KG_’) for all rows where customer is ‘NASA (CRS)’.

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer in ('NASA (CRS)');
```

```
 * sqlite:///my_data1.db
Done.
```

| sum(PAYLOAD_MASS__KG_) |
|---|
| 45596 |

# Average Payload Mass by F9 v1.1

- This query returns the average of all payload mass for rows in which the ‘Booster_Version’ contains the substring ‘F9 v1.1’.
- The aggregate function ‘avg’ returns the average.

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version like 'F9 v1.1%';
```

```
 * sqlite:///my_data1.db
Done.
```

| avg(PAYLOAD_MASS__KG_) |
|---|
| 2534.6666666666665 |

# First Successful Ground Landing Date

- The function ‘min’ returns the earliest date where the landing outcome on ground pad was successful.

```
%sql select min(Date) from SPACEXTBL where Landing_Outcome = 'Success (ground pad)';
```

```
 * sqlite:///my_data1.db
Done.
```

| min(Date) |
|---|
| 2015-12-22 |

# Successful Drone Ship Landing with Payload between 4000 and 6000

- The ‘where’ clause & keyword ‘and’ filter the conditions and return the booster version where landing was successful on drone ship and payload mass is between 4000 kg and 6000 kg.

```
%%sql
select Booster_Version
from SPACEXTBL
where Landing_Outcome = 'Success (drone ship)' and
    4000 < PAYLOAD_MASS__KG_ < 6000;
```

```
 * sqlite:///my_data1.db
Done.
```

| Booster_Version |
| --- |
| F9 FT B1021.1 |
| F9 FT B1022 |
| F9 FT B1023.1 |
| F9 FT B1026 |
| F9 FT B1029.1 |
| F9 FT B1021.2 |
| F9 FT B1029.2 |
| F9 FT B1036.1 |
| F9 FT B1038.1 |
| F9 B4 B1041.1 |
| F9 FT B1031.2 |
| F9 B4 B1042.1 |
| F9 B4 B1045.1 |
| F9 B5 B1046.1 |

# Total Number of Successful and Failure Mission Outcomes

- Listing the total number of successful and failure mission outcomes.
- The clause 'group by' groups the resultant counts based on the mission outcome.

```
%%sql
select Mission_Outcome, count(Mission_Outcome)
from SPACEXTBL
group by Mission_Outcome;
```

```
 * sqlite:///my_data1.db
Done.
```

| Mission_Outcome | count(Mission_Outcome) |
|---|---|
| Failure (in flight) | 1 |
| Success | 98 |
| Success | 1 |
| Success (payload status unclear) | 1 |

# Boosters Carried Maximum Payload

- Listing the names of the booster versions which have carried the maximum payload mass.
- The subquery filters data by returning only the heaviest payload mass with MAX function. The main query uses subquery results and returns booster version with the heaviest payload mass.

```
%%sql
select Booster_Version, PAYLOAD_MASS__KG_
from SPACEXTBL
where PAYLOAD_MASS__KG_ in (
    select max(PAYLOAD_MASS__KG_) from SPACEXTBL);
```

```
 * sqlite:///my_data1.db
Done.
```

| Booster_Version | PAYLOAD_MASS__KG_ |
|---|---|
| F9 B5 B1048.4 | 15600 |
| F9 B5 B1049.4 | 15600 |
| F9 B5 B1051.3 | 15600 |
| F9 B5 B1056.4 | 15600 |
| F9 B5 B1048.5 | 15600 |
| F9 B5 B1051.4 | 15600 |
| F9 B5 B1049.5 | 15600 |
| F9 B5 B1060.2 | 15600 |
| F9 B5 B1058.3 | 15600 |
| F9 B5 B1051.6 | 15600 |
| F9 B5 B1060.3 | 15600 |
| F9 B5 B1049.7 | 15600 |

# 2015 Launch Records

```
%%sql
SELECT Landing_Outcome, Booster_Version, Launch_Site, substr(Date, 6,2) as month, substr(Date,0,5) as year
FROM SPACEXTBL
WHERE Landing_Outcome = 'Failure (drone ship)'
    AND substr(Date,0,5) = '2015';
```

```
 * sqlite:///my_data1.db
Done.
```

| Landing_Outcome | Booster_Version | Launch_Site | month | year |
|---|---|---|---|---|
| Failure (drone ship) | F9 v1.1 B1012 | CCAFS LC-40 | 01 | 2015 |
| Failure (drone ship) | F9 v1.1 B1015 | CCAFS LC-40 | 04 | 2015 |

- This query returns landing outcomes, booster version, launch site, month, and year where there was a failed landing in drone ship in the year 2015
- Substr(Date, 6, 2) fetches month and substr(Date, 0, 5) fetches year.

# Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.
- The 'group by' clause groups results by landing outcome and 'order by … desc' sorts the results in descending order.

```
%%sql
SELECT Landing_Outcome, count(Landing_Outcome)
FROM SPACEXTBL
WHERE DATE between '2010-06-04' and '2017-03-20'
group by Landing_Outcome
order by count(Landing_Outcome) desc;
```

```
 * sqlite:///my_data1.db
Done.
```

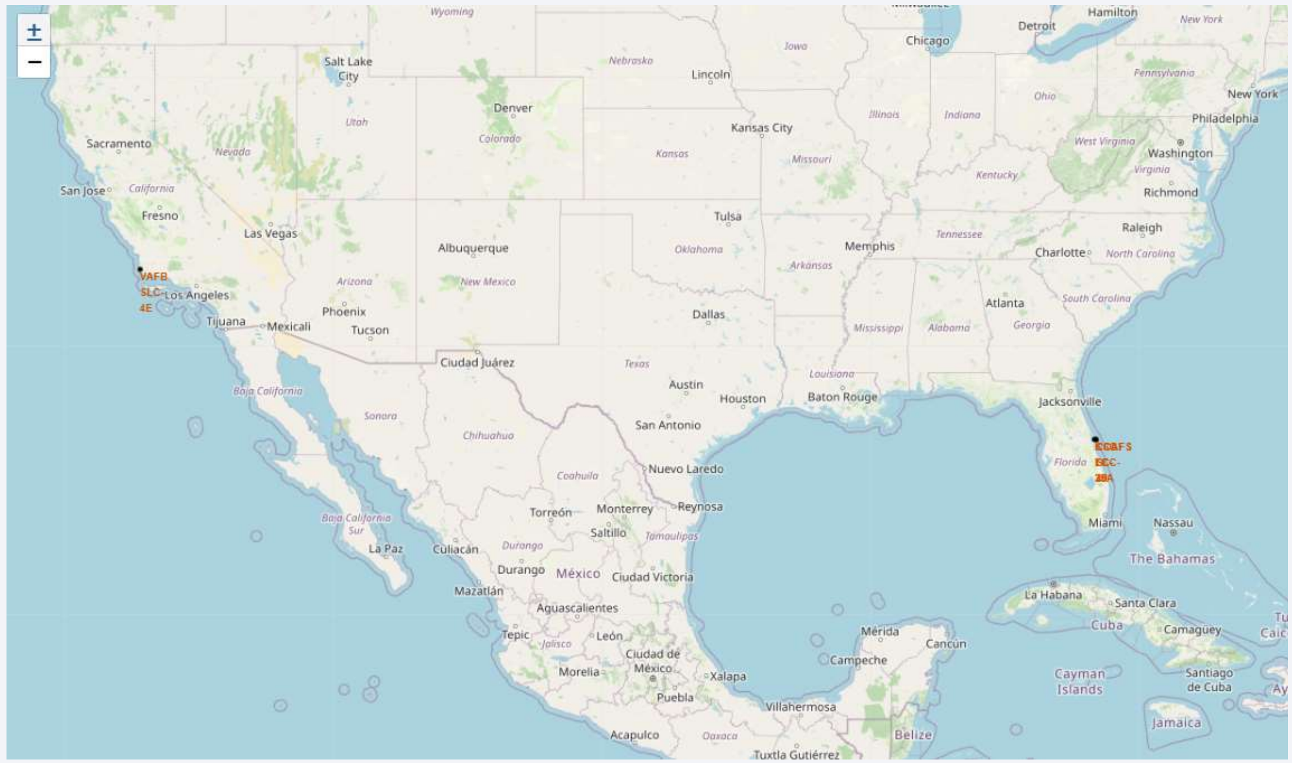
| Landing_Outcome | count(Landing_Outcome) |
|---|---|
| No attempt | 10 |
| Success (drone ship) | 5 |
| Failure (drone ship) | 5 |
| Success (ground pad) | 3 |
| Controlled (ocean) | 3 |
| Uncontrolled (ocean) | 2 |
| Failure (parachute) | 2 |
| Precluded (drone ship) | 1 |

Section 3

# Launch Sites Proximities Analysis

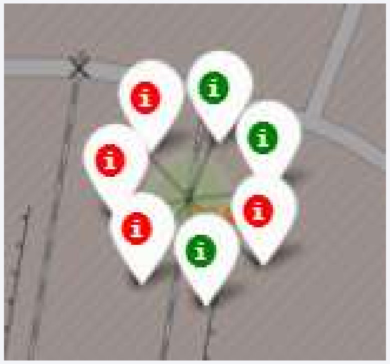
# Location Marking of all launch sites on Folium map

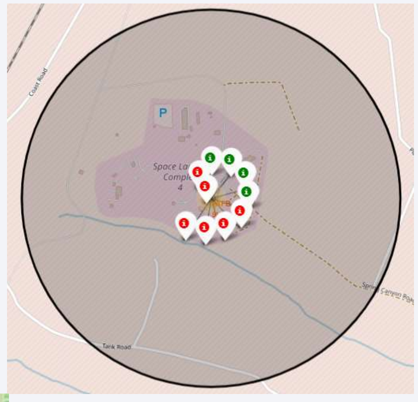
All the Space X launch sites are located near the coastline and close to the equator.

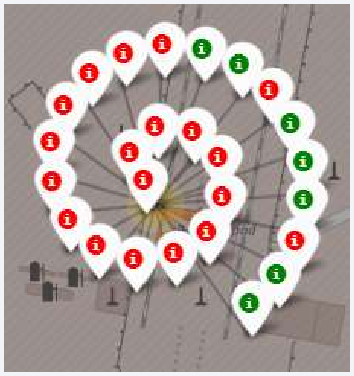


# Color-labeled success/failed launches

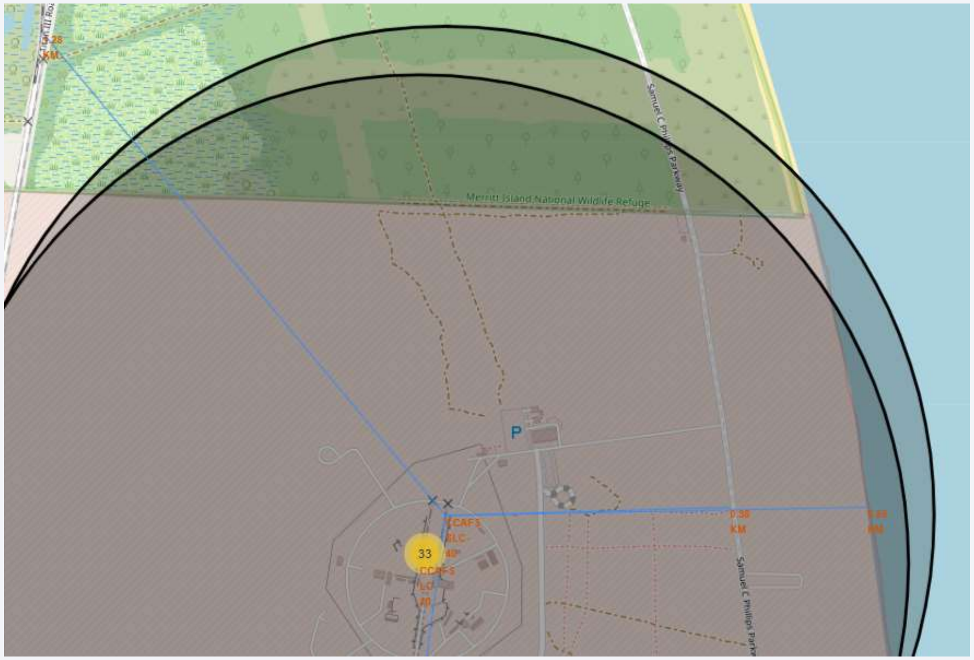
- Explain the important elements and findings on the screenshot







# Distance between launch site and its proximities




- CCAFS SLC-40 is in close proximity to railways, highway, coastline, but is pretty far from cities.

Section 4

# Build a Dashboard with Plotly Dash

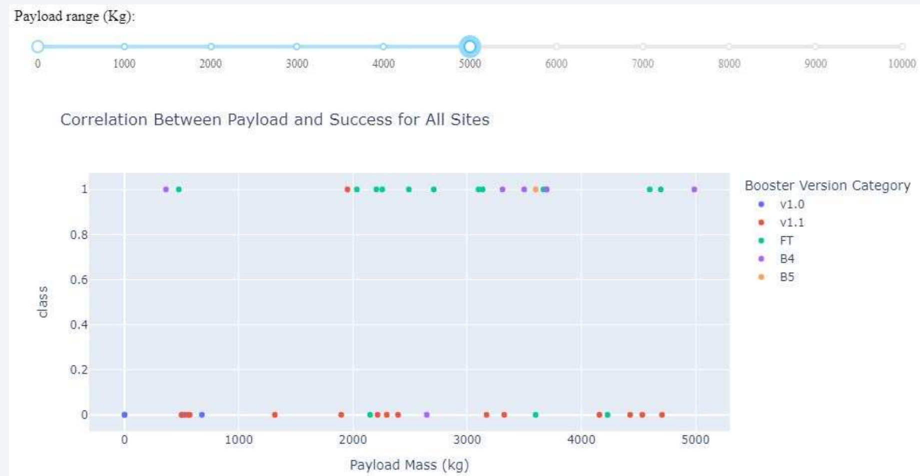
# Dashboard – Site-wise launch success

- From the pie-chart, it is evident that KSC LC-39A has the most successful launches.

# Success ratio of launch site KSC LC-39A

- The pie-chart shows that KSC LC-39A has the highest launch success rate (76.9%)

# Payload mass vs Outcome for all sites

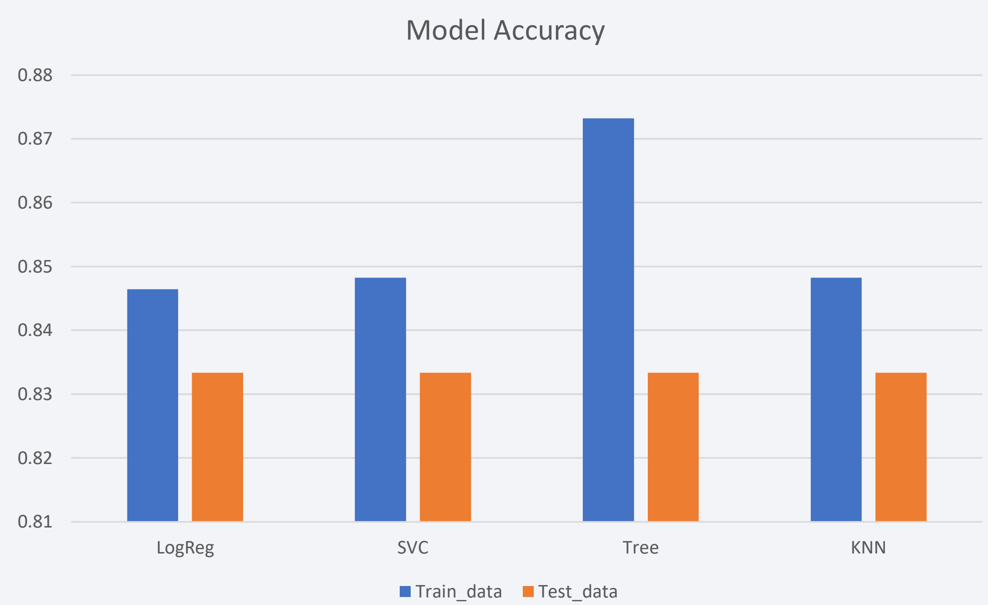

- Launches with low payload mass have a better success rate than the heavy payload mass.

Section 5

# Predictive Analysis (Classification)

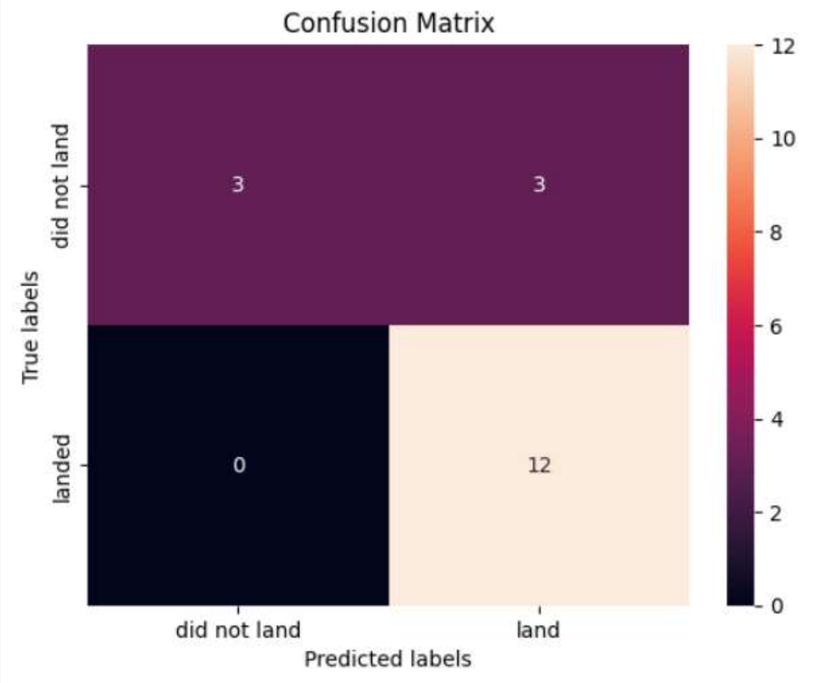
# Classification Accuracy

- The accuracy on the test data is same for all the 4 classification models: Logistic Regression, Support Vector Machine, Decision Tree, K-Nearest Neighbor.
- This may be due to the small sample size of the test data.
- Hence comparing on the train data, Decision Tree model has the highest classification accuracy.



# Confusion Matrix

- All the four models produced the same confusion matrix, which is displayed here.
- The main issue in predicting successful launches is 'False Positive'.
- False Positive represent those cases where the model predicted the launch as successful, whereas it was a failure in reality.



# Conclusions

- The company experienced a learning curve, contributing to higher success over time.
- The launch sites are near the equator and close to coastline, highways and railway; although they are far from big cities.
- Launches with low payload mass have better success rates than those with higher payload mass.
- All 4 models had the same test accuracy, but when checked for the entire dataset, Decision Tree model gave the best results.
- If the models produce different confusion matrix, we should select the model with least 'False Positive (FP)' cases. FP cases cost time, effort and money to the company.


Thank you!