IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ahmed Mohamed Nagib

March 9th 2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection
  - Data wrangling
  - EDA with data visualization
  - EDA with SQL
  - Building an interactive map with Folium
  - Building a Dashboard with Plotly Dash
  - Predictive analysis (Classification)
- Summary of all results
  - Exploratory data analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

Introduction

- **Project background and context**

The commercial space age is here. And many companies are making space travel cheaper and cheaper every day. Perhaps the most successful is SpaceX, the main factor behind that success is that their rocket launches are relatively inexpensive.

SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.

Therefore, we will predict if the Falcon 9 first stage will land successfully. 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.

- **Problems we want to find answers**
    - The correlations between rocket variable and successful landing rate for each launch
    - The conditions to get the best results and to ensure the best successful landing rate

Section 1

Methodology

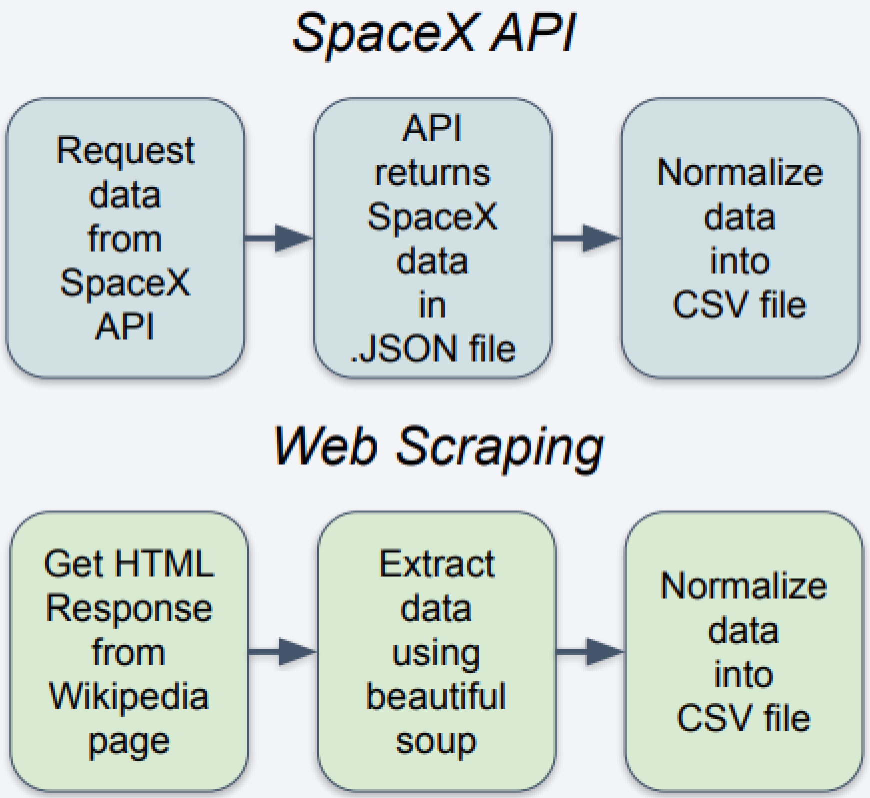
Methodology

Executive Summary

- Data collection methodology:
    - SpaceX API and Web scraping Falcon9 Wikipedia Page
- Perform data wrangling
    - Determining Training Labels and Outcome Label
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Optimizing hyper-parameters for each classification model to find the best model

Data Collection

- During data collection we combined the use of API requests from SpaceX API and web scraping data from tables in the Faclon 9 Wikipedia page.



Data Collection – SpaceX API

- **Collecting SpaceX data using REST calls**

- **Cleaning and Wrangling the data**

SpaceX Data Collection API notebook

Github URL

1- Collecting and Parsing Data

Requesting SpaceX rocket lauch data from SpaceX APIs using GET requests

Parsing and decoding the received data into JSON format

Converting the data into a Pandas dataframe

Given the IDs for each launch site. We use the API again to collect more information

Using the API we collect: Payload, Launchpad, Cores and Rocet Booster Versions

Converting all collected data into a single Pandas Dataframe

2- Data Cleaning and Wrangling

Filtering the dataframe to only include Falcon 9 launches

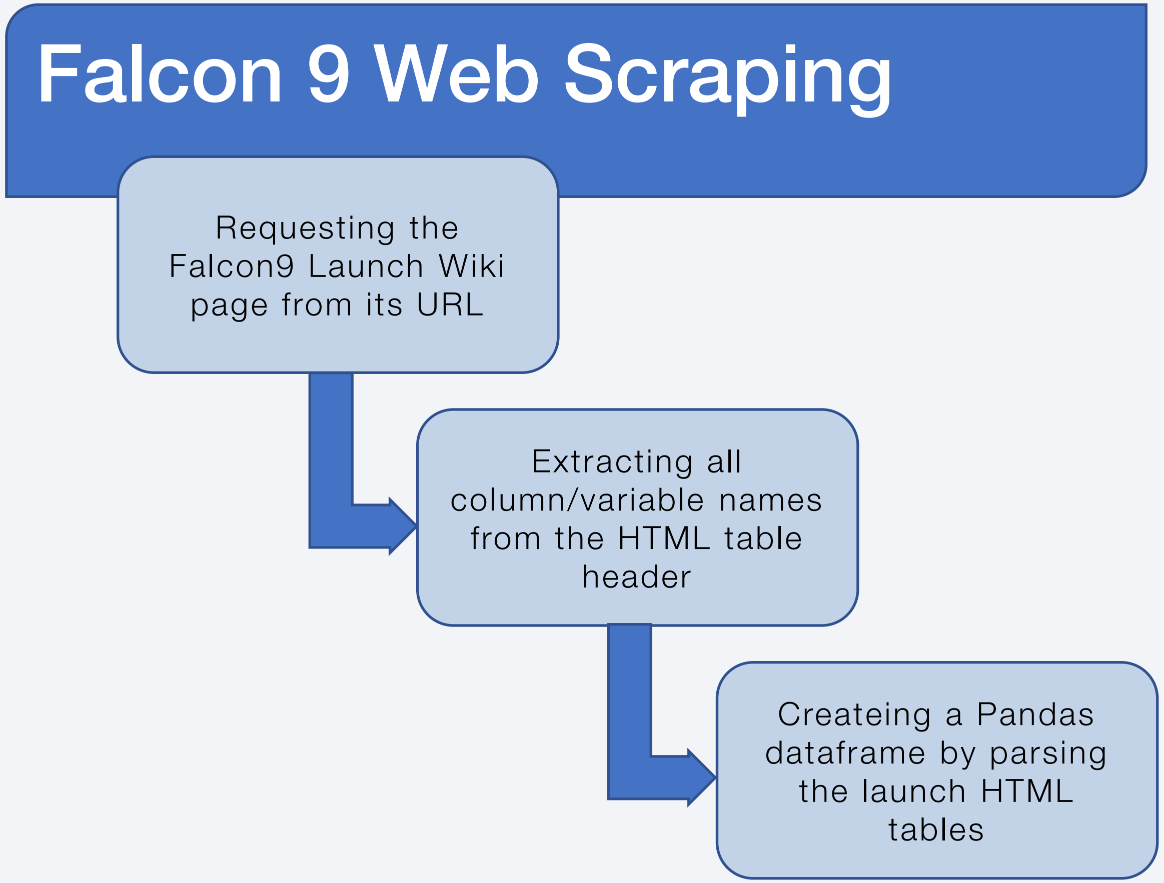
Dealing with Missing Values

Saving the data as a CSV file

Data Collection - Scraping

- Web scraping Falcon 9 and Falcon Heavy Launches Records from Wikipedia using HTTP requests and BeautifulSoup
- Using custom functions to extract data from HTML tables
- Converting the cleaned data into Pandas dataframe and saving the data as a CSV file

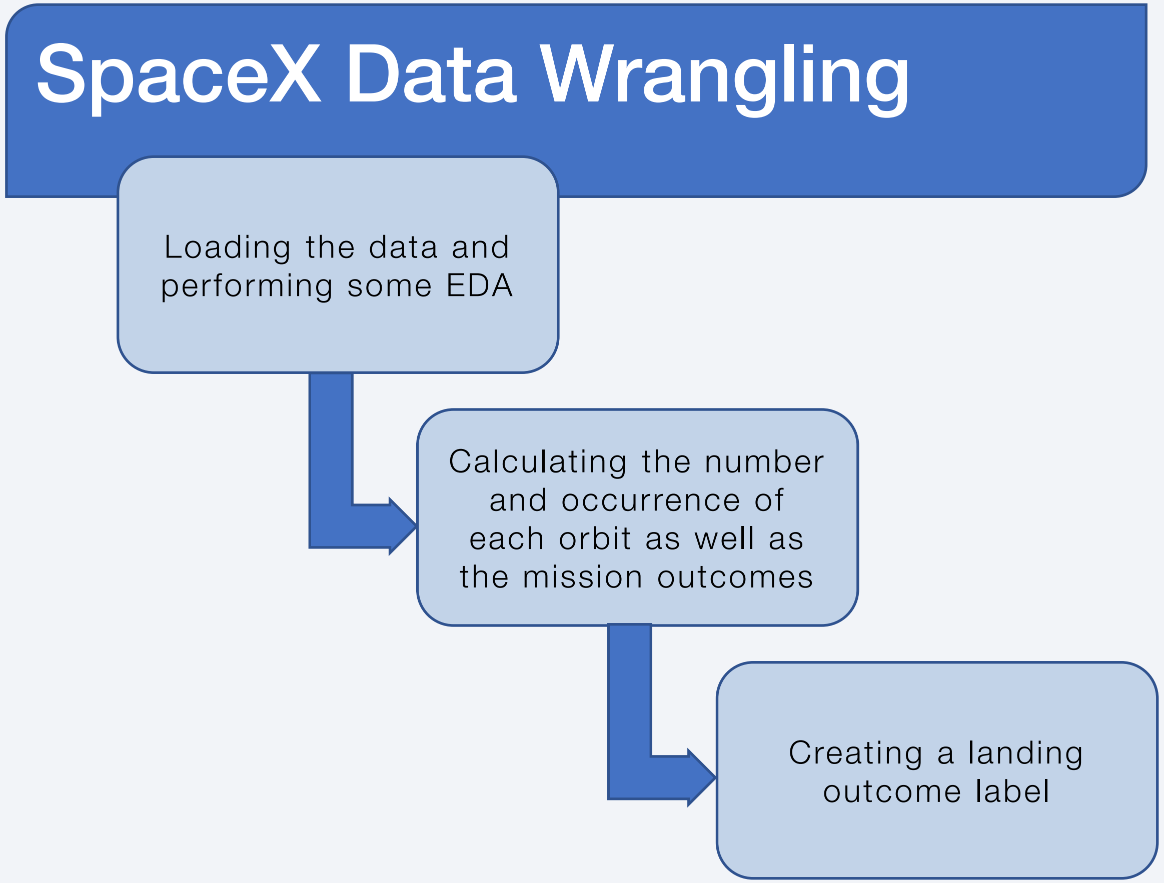
[SpaceX Webscraping notebook Github URL]()



Data Wrangling

- Perform some Exploratory Data Analysis (EDA) to find patterns in the data and determine what would be the label for training supervised models
- Converting those outcomes into Training Labels:
  - 1 = Successful landing
  - 0 = Unsuccessful landing.

[SpaceX Data Wrangling notebook Github URL]



EDA with SQL

Storing the dataset into a Db2 database, and using SQL queries to answer the below questions:

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA‘
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was achieved.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Rank 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

SpaceX EDA with SQL notebook Github URL

EDA with Data Visualization

- **Scatter charts:** Because a scatter chart shows the relationship between two variables (called correlation) and demonstrates how much a variable is affected by another, we used it to show the relationship between:
  - **Flight Number and Launch Site Payload vs. Launch Site**
  - **Payload and Launch Site**
  - **Flight Number and Orbit type**
  - **Payload and Orbit type**
- **Bar charts:** Because a bar chart used to convey relational information quickly as the bars display the quantity for a particular category, we used it to:
  - **Demonstrate the success rate of each orbit type**
- **Line chart:** Because the line chart is a type of chart used to visualize the value of something over time (the trend), we used it to:
  - **Visualize the launch success yearly trend**

[SpaceX EDA with Visualization notebook Github URL]

Build an Interactive Map with Folium

- We created and added the below object to the map:
    - Markers that show all launch sites on the map with a circle and a label
    - Markers that show all success/failed launches for each site on the map using Marker Clusters
    - Lines to show the distances between a launch site and its proximities

- By adding those objects we were able to:
    - Easily see the location for each launch site and how many successful/failed launches each site had
    - We see that each launch site in close proximity to at least one railway, one highway and a coastline
    - We see the each launch site keep certain distance from nearby cities

SpaceX Launch Sites Locations Analysis with Folium notebook Github URL

Build a Dashboard with Plotly Dash

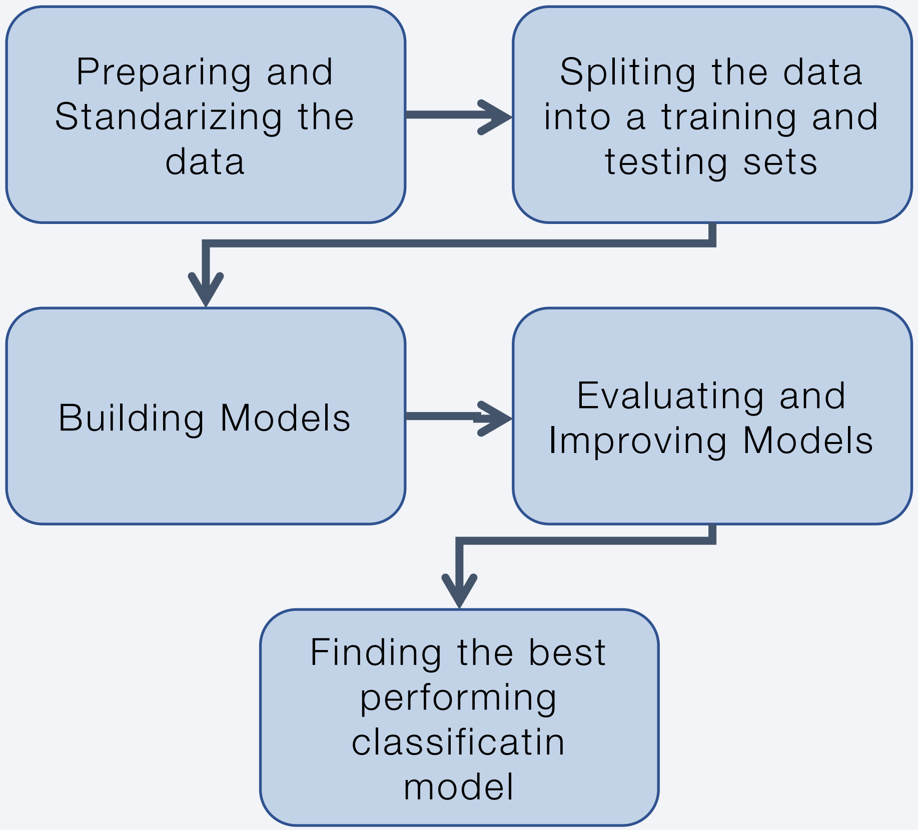
The built dashboard contains a pie chart and a scatter plot chart

- Pie Chart:
  - Showing total success launches for each launch site and as well as combined for all sites
  - It also indicates the success rate for each launch site
- Scatter Chart:
  - Sowing the relationship between Payload mass (kg) and Outcomes, for different boosters
  - **The chart utilizes two inputs:**
    - A dropdown list to select: Certain Site/All Sites
    - A slider to select Payload mass in (kg)

[SpaceX Interactive Dashboard with Ploty Dash notebook Github URL]

Predictive Analysis (Classification)

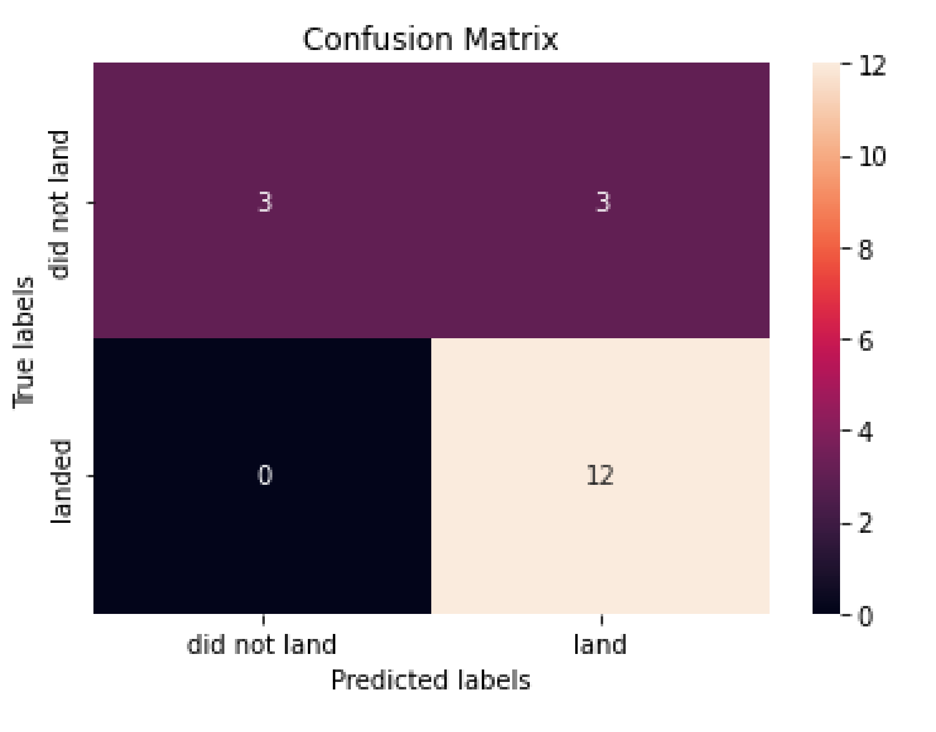
- **Preparing The Data**
  - Selecting independent and dependent variables
  - Standardizing the data
  - Splitting the data into training and testing sets
- **Building Models**
  - Finding the best Hyperparameters for each classification model
  - Training and testing each model
  - Finding the best performing model using the test data

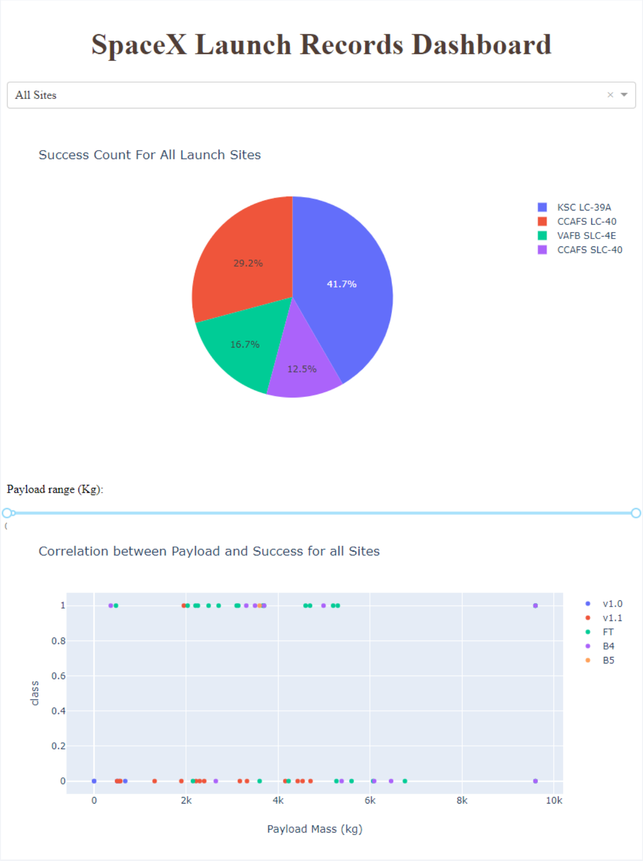


[SpaceX Machine Learning Prediction notebook Github URL]()

Results

- On the right a screenshot for a preview of the Interactive analytics in Plotly Dash
- Also below diagram shows our classification model confusion matrix





Section 2

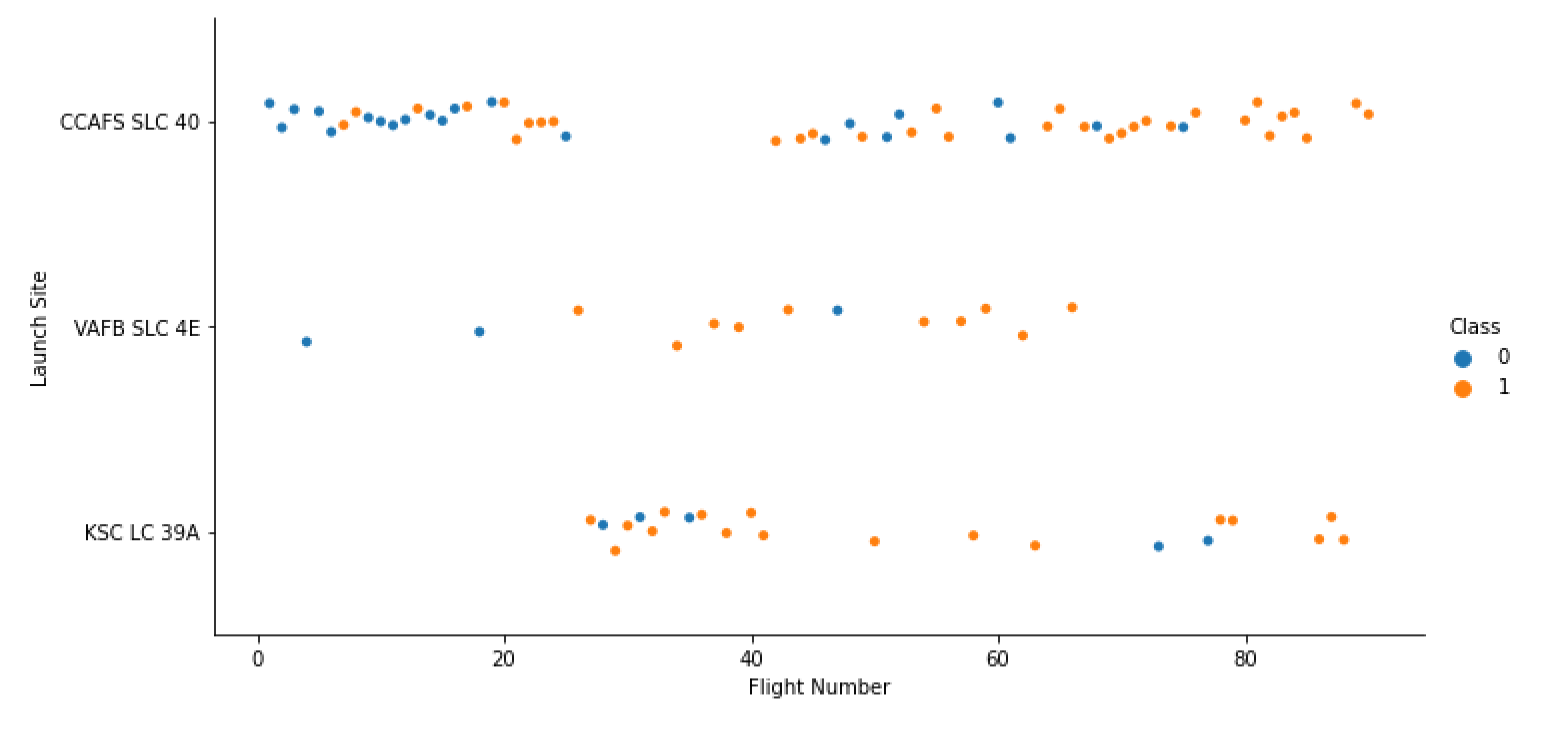
Insights drawn from EDA

Flight Number vs. Launch Site

- The figure shows that **the success rate increased as the number of flights increased.**
- **Success rate significantly enhanced after the 20th flight.** This should go under further analysis and investigation.

Class 0 = Unsuccessful Launch

Class 1 = Successful Launch

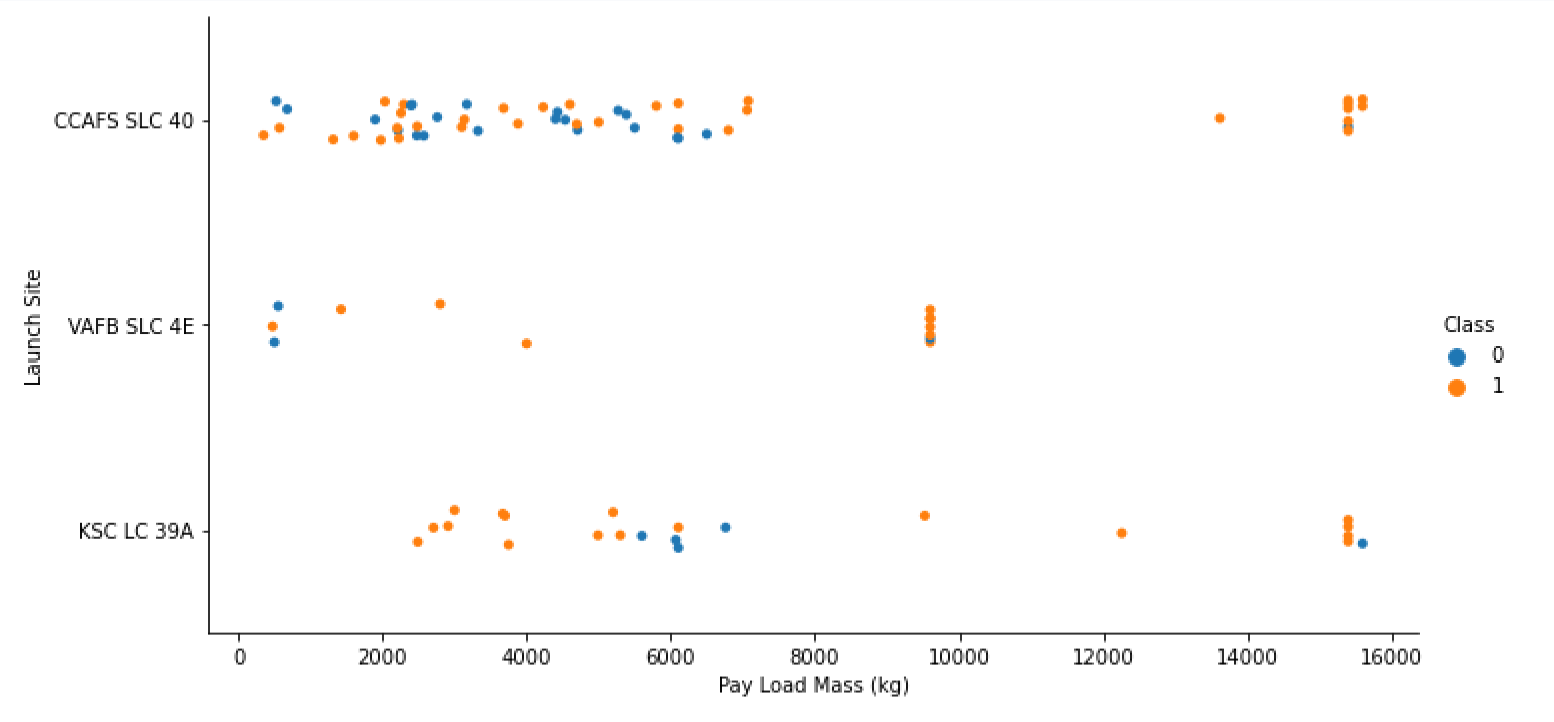


Payload vs. Launch Site

- While the chart shows that **the larger the payload mass, the higher the success rate, it is difficult to confirm as no clear pattern between both variables.**
- It is not very clear if the launch site plays a significant role affecting the success rate or not

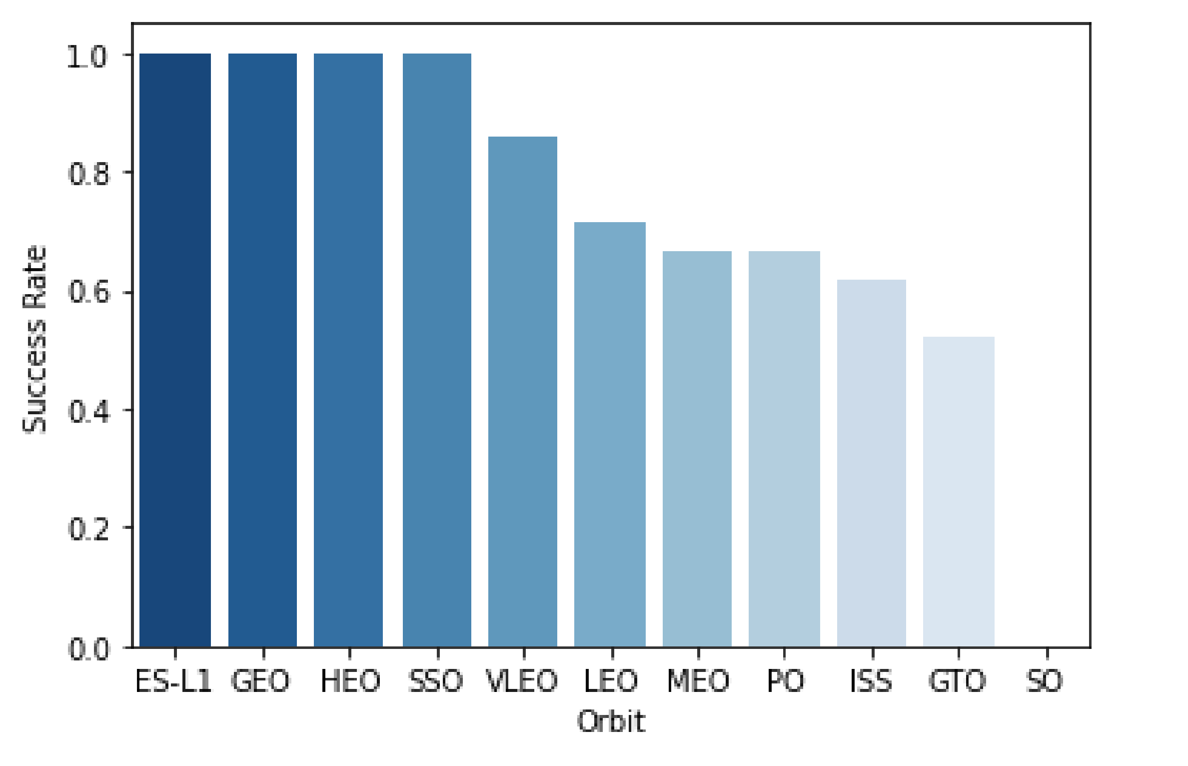
Class 0 = Unsuccessful Launch

Class 1 = Successful Launch



Success Rate vs. Orbit Type

- Orbit types ES-L1, GEO, HEO and SSO have the highest success rates *(100%)*.
- However we should consider the number of flights for each orbit type as well. ES-L1, GEO and HEO orbits had only one flight attempt that was successful, hence the 100% success rate
- On the other hand, SO orbit single attempt failed resulting in 0% success rate

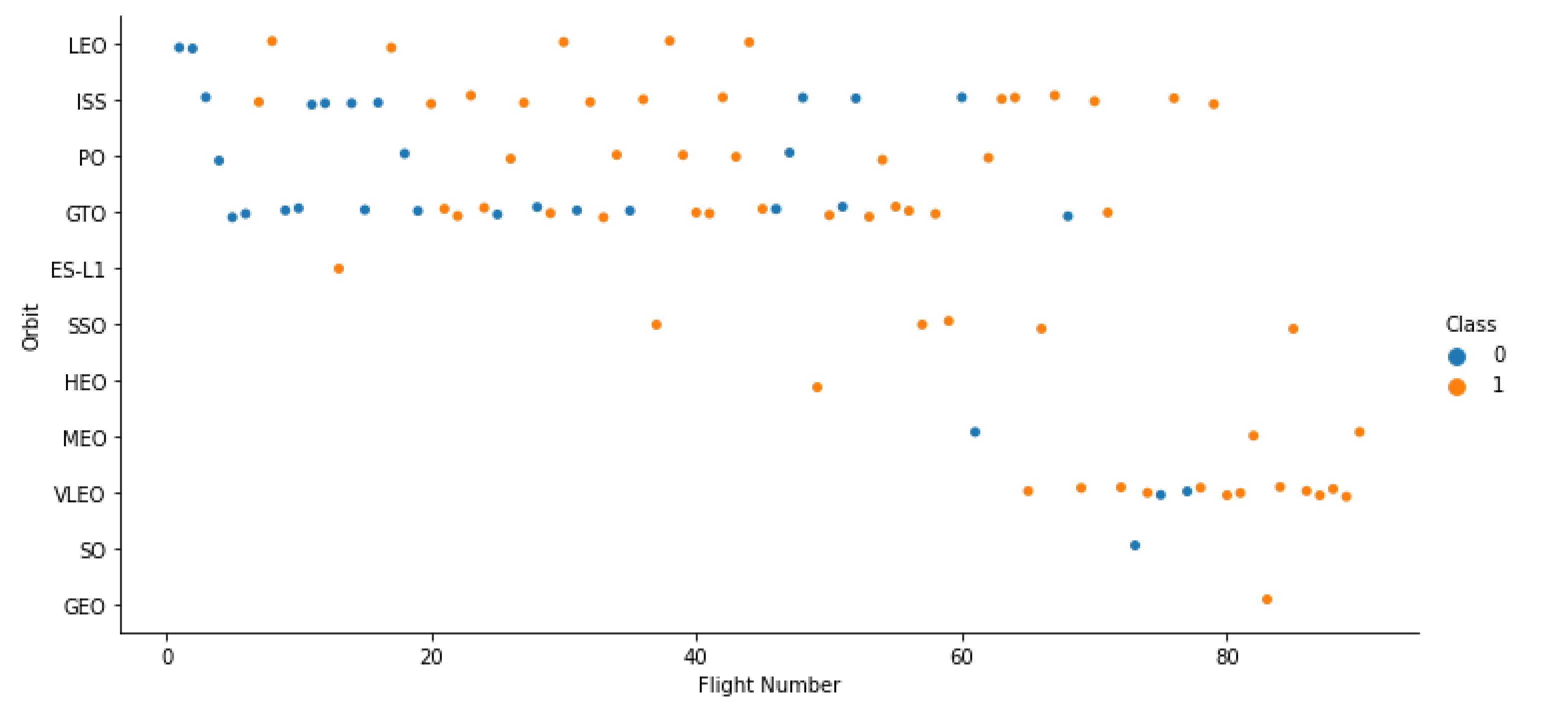


Flight Number vs. Orbit Type

- SSO orbit has the best success rate with more than one launch attempt
- Other than SSO, there is no significant correlation between orbit types and success launch's. It looks like it is more related to flight numbers
- SpaceX started with LEO but recently went to VLEO orbit which shows better success rate.

Class 0 = Unsuccessful Launch

Class 1 = Successful Launch

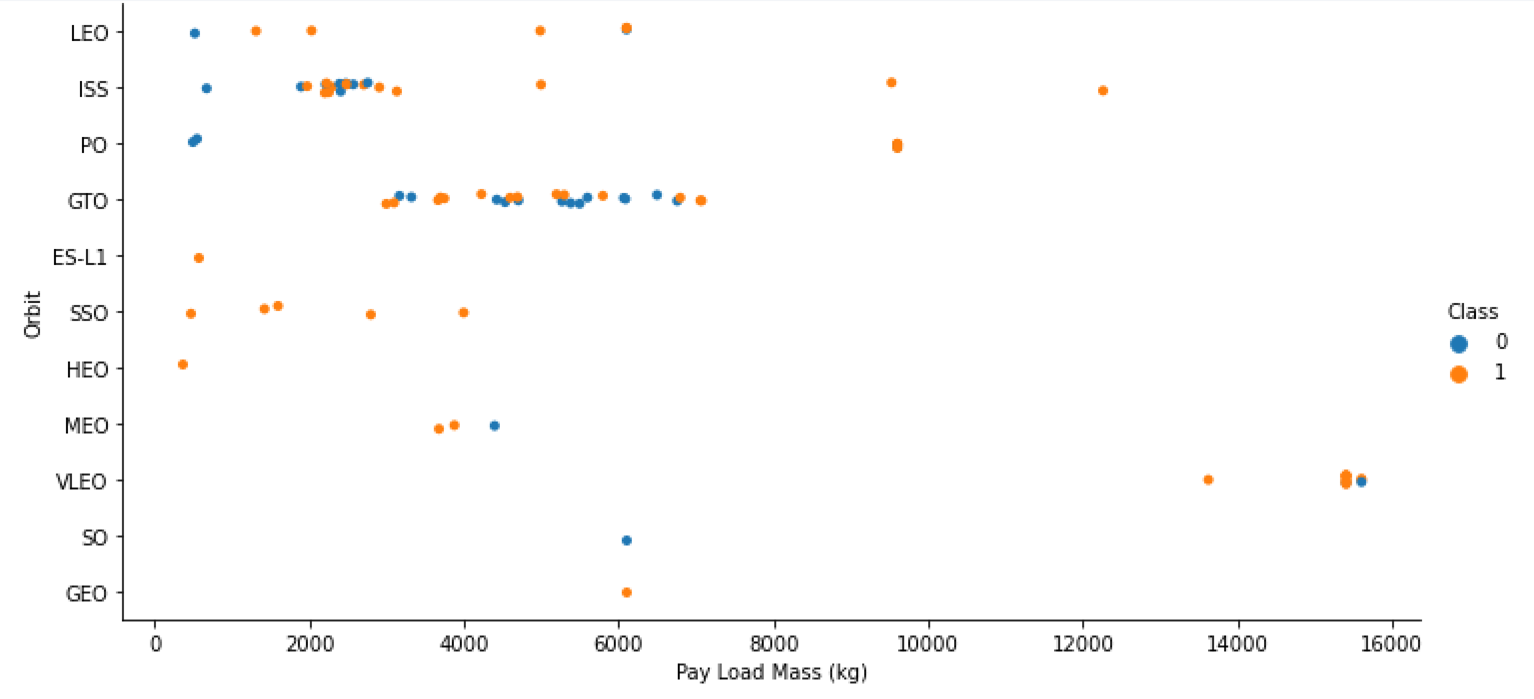


Payload vs. Orbit Type

- Better success rate with heavier payload mass.
- Although we should again notice that heavier payload mass (>8000kg) are much less than less heavy ones

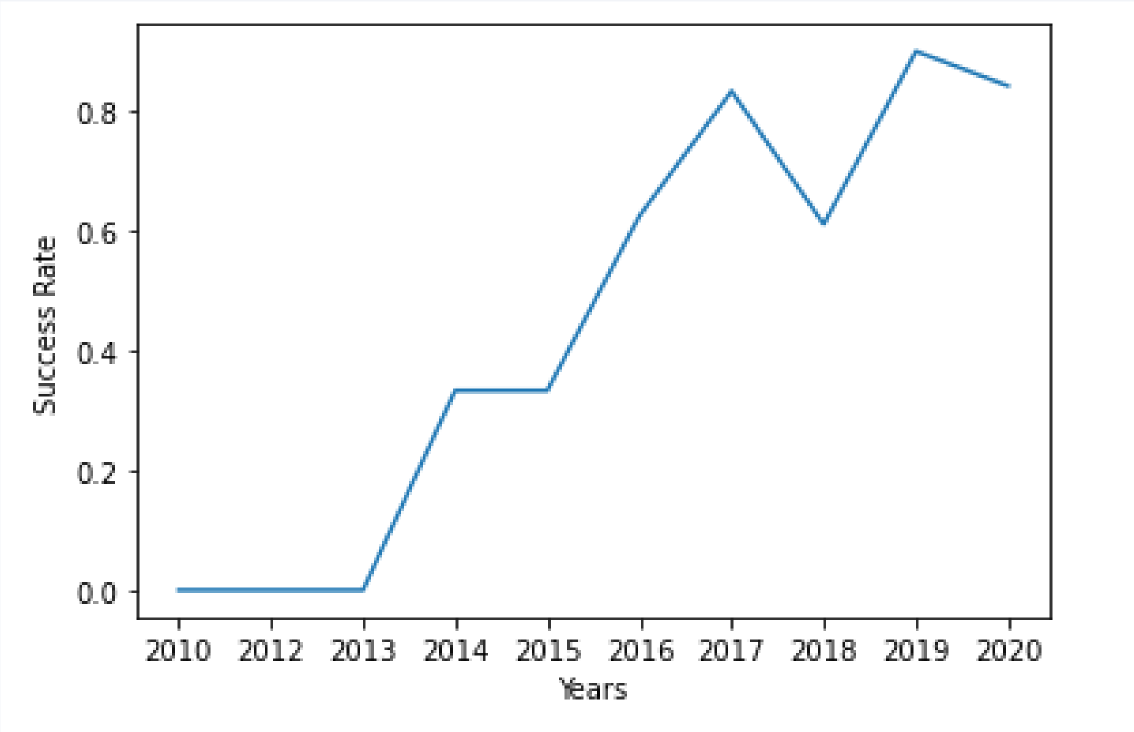
Class 0 = Unsuccessful Launch

Class 1 = Successful Launch



Launch Success Yearly Trend

- The success rate trend increasing since 2013. except for the dip at 2018
- Success rate in the recent years is almost *80%*



All Launch Site Names

- Using DISTINCT Selection we can find the Launch sites names as below:
    - *CCAFS LC-40*
    - *CCAFS SLC-40*
    - *KSC LC-39A*
    - *VAFB SLC-4E*

- Query

```
SELECT DISTINCT LAUNCH_SITE FROM SPACEXDATASET;
```

- Result

launch_site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

Launch Site Names Begin with 'CCA'

- Filtering with WHERE LAUNCH_SITE LIKE 'CCA%' we can find the Launch sites starting with CCA below:

  - *CCAFS LC-40*
  - *CCAFS LC-40*
  - *CCAFS LC-40*
  - *CCAFS LC-40*
  - *CCAFS LC-40*

- Query

```
SELECT LAUNCH_SITE
  FROM SPACEXDATASET
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5;
```

- Result

launch_site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Launch Site Names Begin with 'CCA'

- Filtering with WHERE LAUNCH_SITE LIKE 'CCA%' we can find the Launch sites starting with CCA below:

  - *CCAFS LC-40*
  - *CCAFS LC-40*
  - *CCAFS LC-40*
  - *CCAFS LC-40*
  - *CCAFS LC-40*

- Query

```
SELECT LAUNCH_SITE
  FROM SPACEXDATASET
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5;
```

- Result

launch_site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass for NASA Boosters

- Using Aggregation function SUM() with groupby and filtering we can find the payload carried by boosters from NASA as below:
  - *45596 kg*

- Query

```
SELECT SUM(payload_mass__kg_) AS TOTAL_PAYLOAD_MASS
  FROM SPACEXDATASET
 GROUP BY CUSTOMER
HAVING CUSTOMER = 'NASA (CRS)'
```

- Result

total_payload_mass
45596

Average Payload Mass by F9 v1.1

- Using Aggregation function AVG() with groupby and filtering we can find the average payload for F9 v1.1 as below:

    - *2928 kg*

- Query

```
SELECT AVG(payload_mass__kg_) AVG_PAYLOAD_MASS
  FROM SPACEXDATASET
 GROUP BY BOOSTER_VERSION
HAVING BOOSTER_VERSION = 'F9 v1.1';
```

- Result

avg_payload_mass
2928

First Successful Ground Landing Date

- Using MIN() function with filtering we can find the first successful landing outcome in ground pad was achieved at:
  - *2015-12-22*

- Query

```
SELECT MIN(DATE) FROM SPACEXDATASET
WHERE LANDING__OUTCOME = 'Success (ground pad)'
```

- Result

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Using DISTINCT and multiple filtering conditions we can list the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000 as below:
  - *F9 FT B1021.2*
  - *F9 FT B1031.2*
  - *F9 FT B1022*
  - *F9 FT B1026*

- Query

```
SELECT DISTINCT BOOSTER_VERSION
FROM SPACEXDATASET
WHERE LANDING__OUTCOME = 'Success (drone ship)' and
(PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 )
```

- Result

booster_version
F9 FT B1021.2
F9 FT B1031.2
F9 FT B1022
F9 FT B1026

Total Number of Successful and Failure Mission Outcomes

- Using COUNT and groupby we can find total number of successful and failure mission outcomes as below:
  - *Failure (in flight): 1*
  - *Success: 99*
  - *Success (payload status unclear): 1*

- Query

```
SELECT MISSION_OUTCOME, COUNT(*) NUMBER_OF_OUTCOMES
FROM SPACEXDATASET
GROUP BY MISSION_OUTCOME;
```

- Result

mission_outcome	number_of_outcomes
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- Using subquery we find the result as below:
  - F9 B5 B1048.4
  - F9 B5 B1049.4
  - F9 B5 B1051.3
  - F9 B5 B1056.4
  - F9 B5 B1048.5
  - F9 B5 B1051.4
  - F9 B5 B1049.5
  - F9 B5 B1060.2
  - F9 B5 B1058.3
  - F9 B5 B1051.6
  - F9 B5 B1060.3
  - F9 B5 B1049.7

- Query

```
SELECT BOOSTER_VERSION
FROM SPACEXDATASET
WHERE PAYLOAD_MASS__KG_ =
(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXDATASET)
```

- Result

booster_version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

- Using YEAR(DATE) function we were able to list failed landing outcomes in the year 2015 as below:

DATE	landing__ outcome	booster_ version	time__ utc_
2015-01-10	Failure (drone ship)	F9 v1.1 B1012	09:47:00
2015-04-14	Failure (drone ship)	F9 v1.1 B1015	20:10:00

- Query

```
SELECT DATE, LANDING__OUTCOME, BOOSTER_VERSION, TIME__UTC_
FROM SPACEXDATASET
WHERE (LANDING__OUTCOME = 'Failure (drone ship)') AND (YEAR(DATE) = 2015);
```

- Result

DATE	landing__outcome	booster_version	time__utc_
2015-01-10	Failure (drone ship)	F9 v1.1 B1012	09:47:00
2015-04-14	Failure (drone ship)	F9 v1.1 B1015	20:10:00

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

- Using COUNT, groupby and then ordering we can rank landing outcomes between the given dates as shown in the result screenshot.

- Query

```
SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS TOTAL
FROM SPACEXDATASET
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LANDING__OUTCOME
ORDER BY 2 DESC;
```

- Result

landing__outcome	total
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

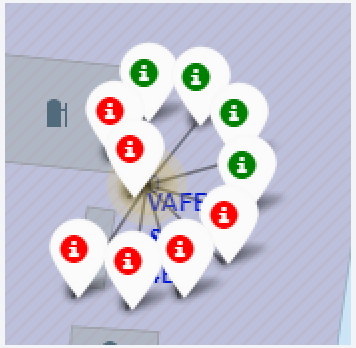
All Launch Sites' Locations

We can see that all launch sites have the below characteristics:

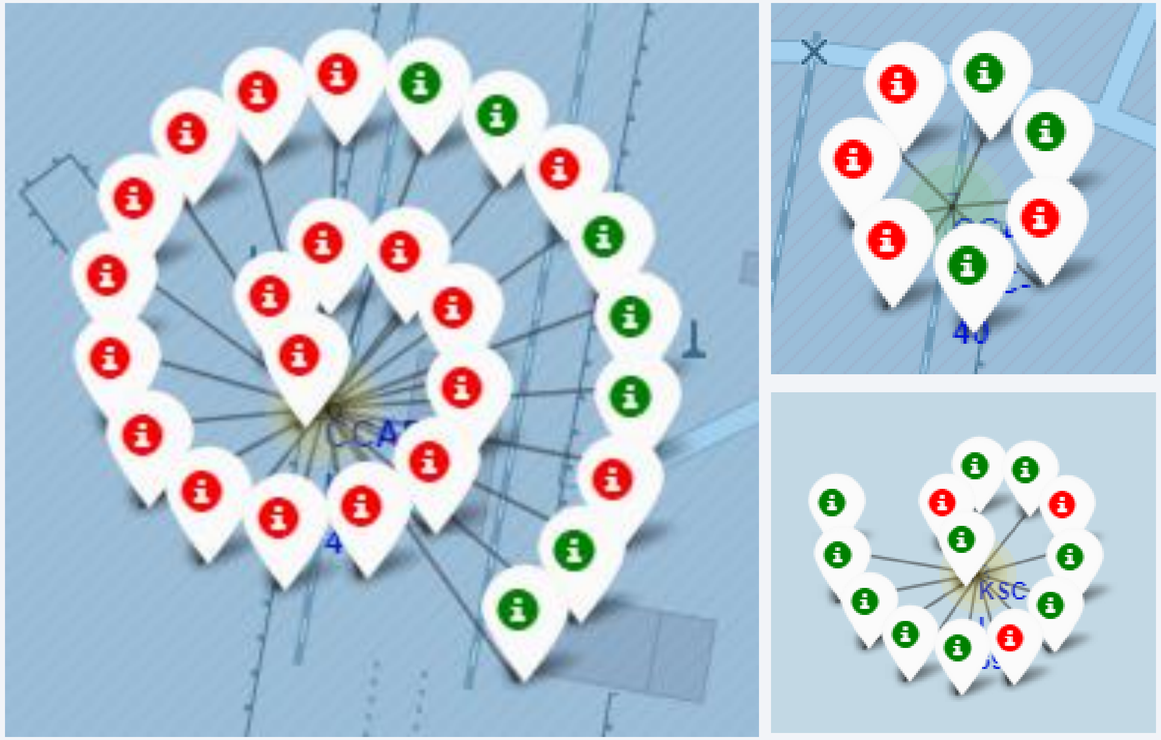
- *All sites located in USA*
- *All sites in proximity to the Equator line*
- *All sites in very close proximity to the coast*

Success/Failed Launch Outcomes Color Coded

Launch Site on the west coast



By clicking on the marker cluster we can see successful landings in (green) and failed landings in (red)

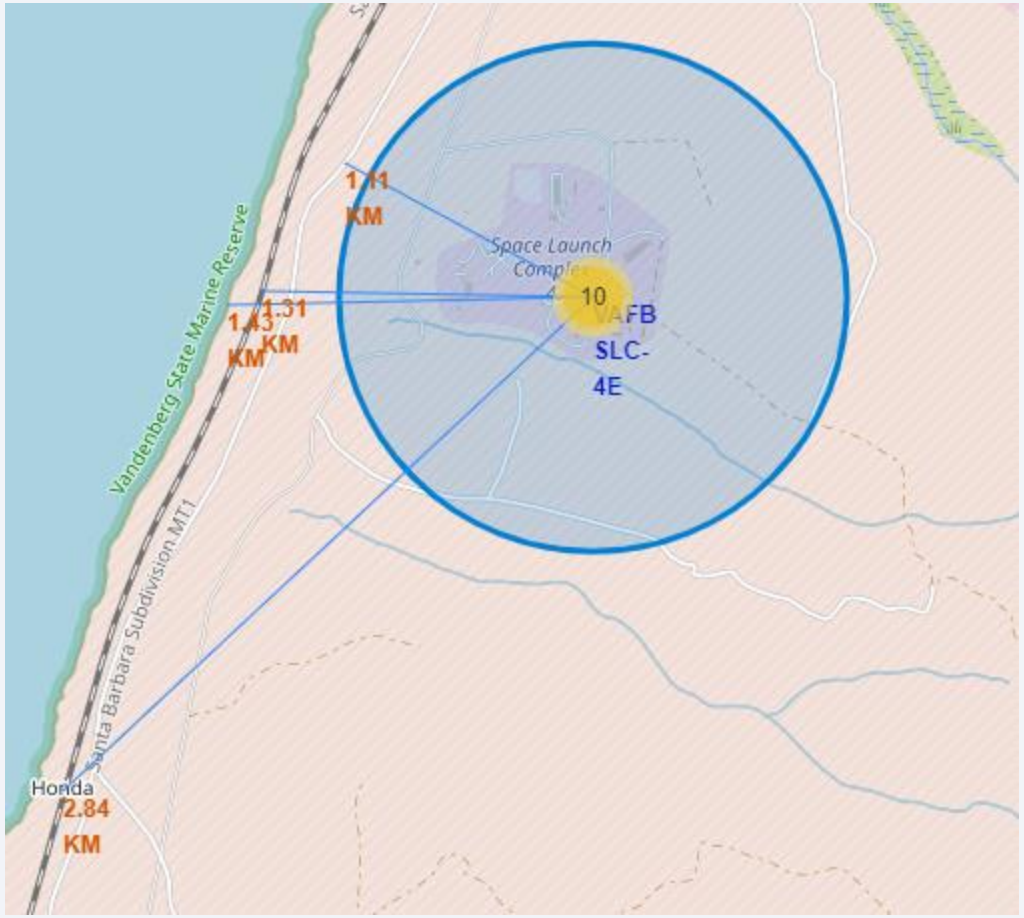


Launch Sites on the east coast

Proximities of Launch Sites

- We can see that the launch site is in close proximity to railways and highways, probably for transportation of equipment and personnel

- On the other hand, the launch site is relatively far from the cities so that launch failure does not pose a threat.



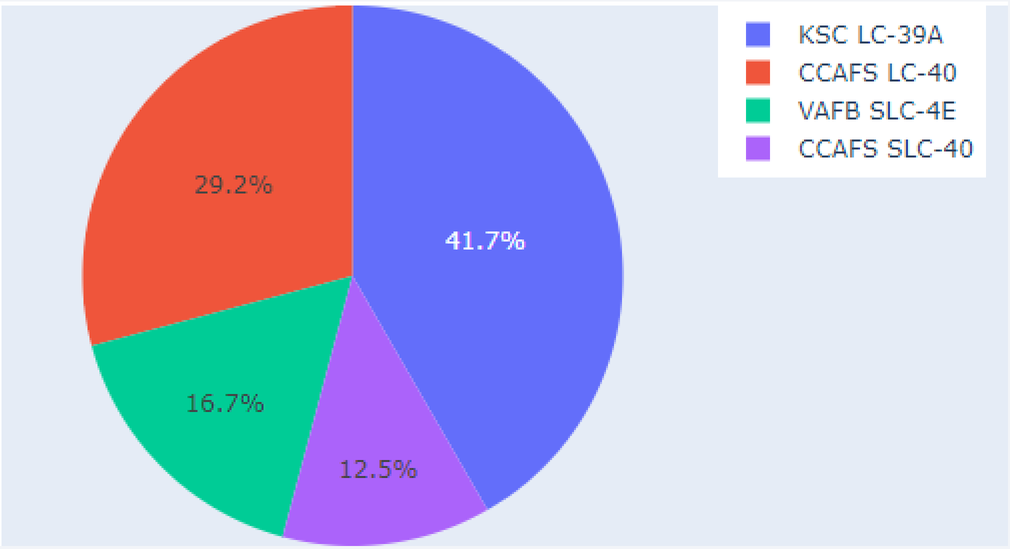
Section 4

Build a Dashboard with Plotly Dash

Total Success Launches By all sites

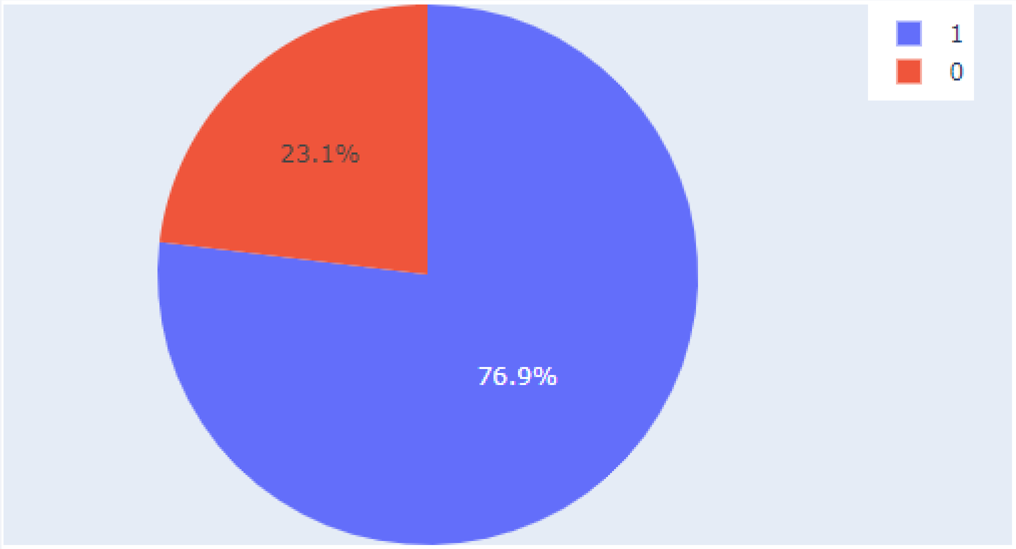
- **KSLC-39A has the most successful launches among all sites**
- **VAFB SLC-4E has the least successful launches among all sites**

Noting that these figures are influenced by the total number of launches per each site.



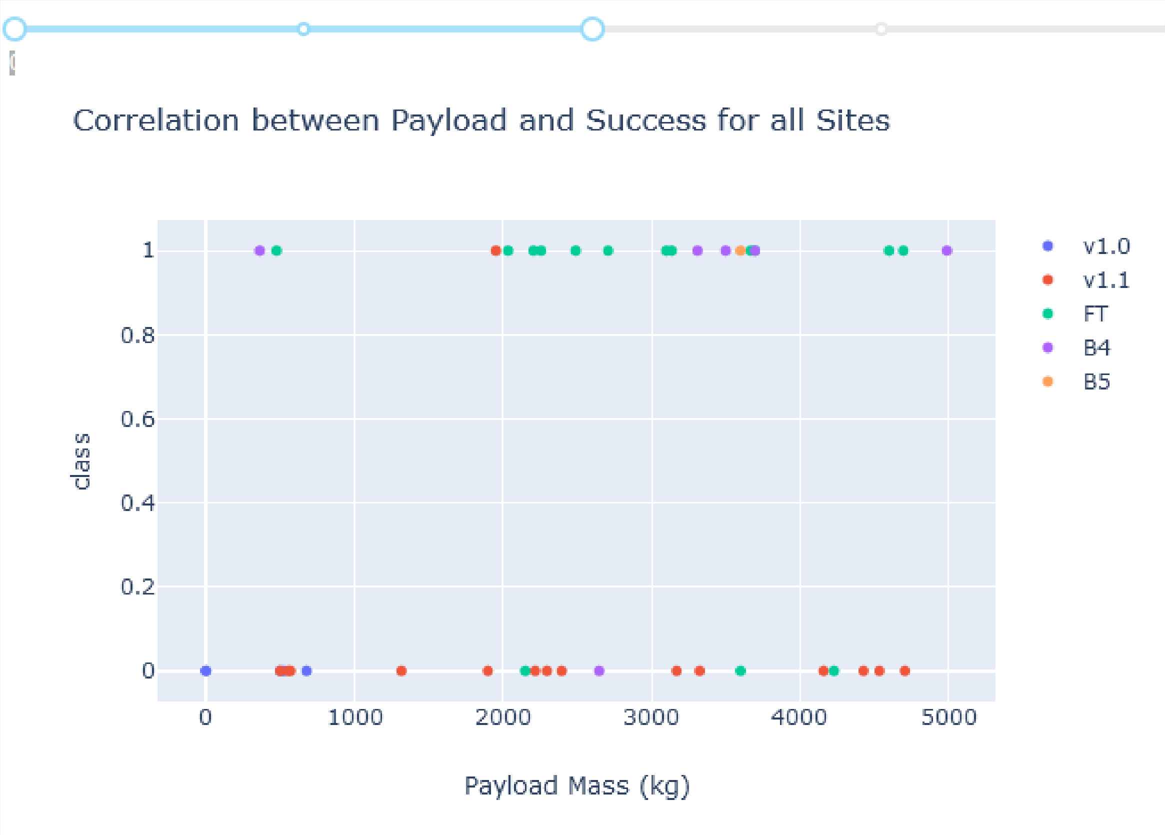
Highest Launch Success Ratio

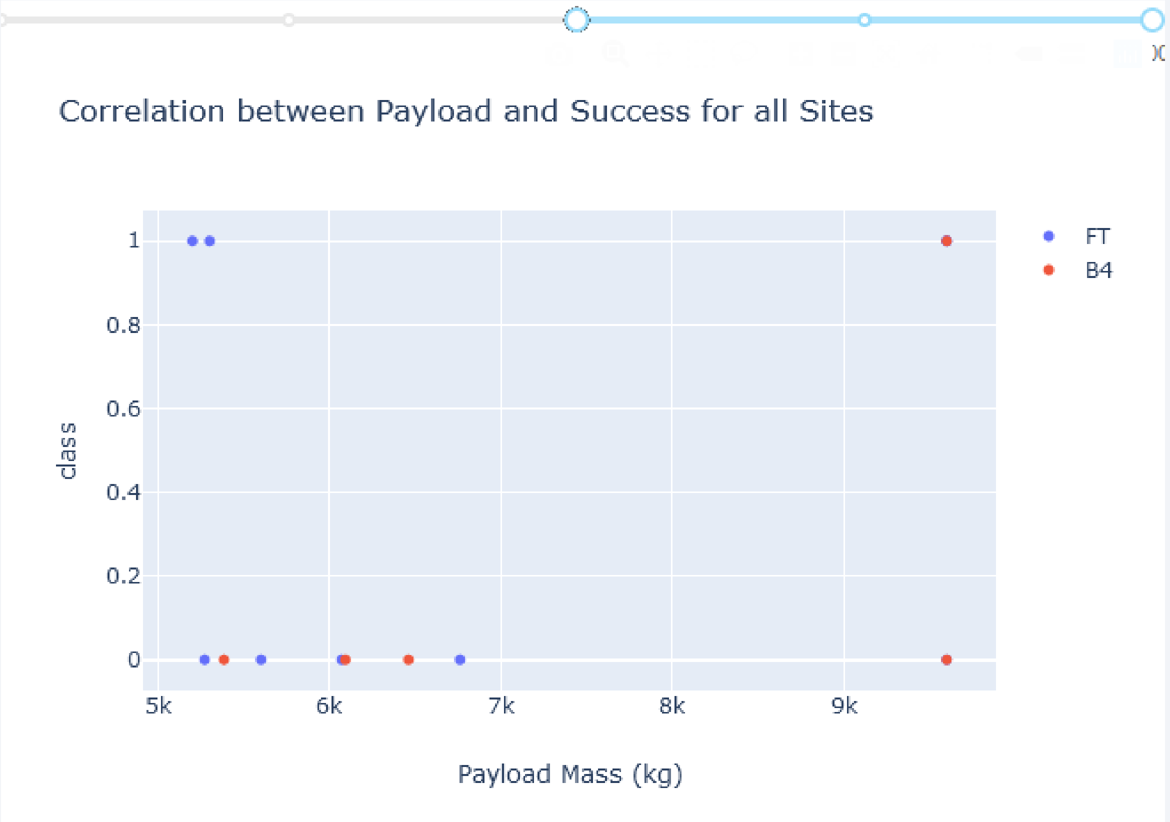
- KSLC-39A has the highest success rate:
  - 10 successful landings *(76.9%)*
  - 3 failed landings *(23.1%)*



Payload vs. Launch Outcome Scatter Plot

These figures show that the launch success rate for lower payloads (*<5000 kg*) is higher than that of heavier payloads(*>5000kg*)



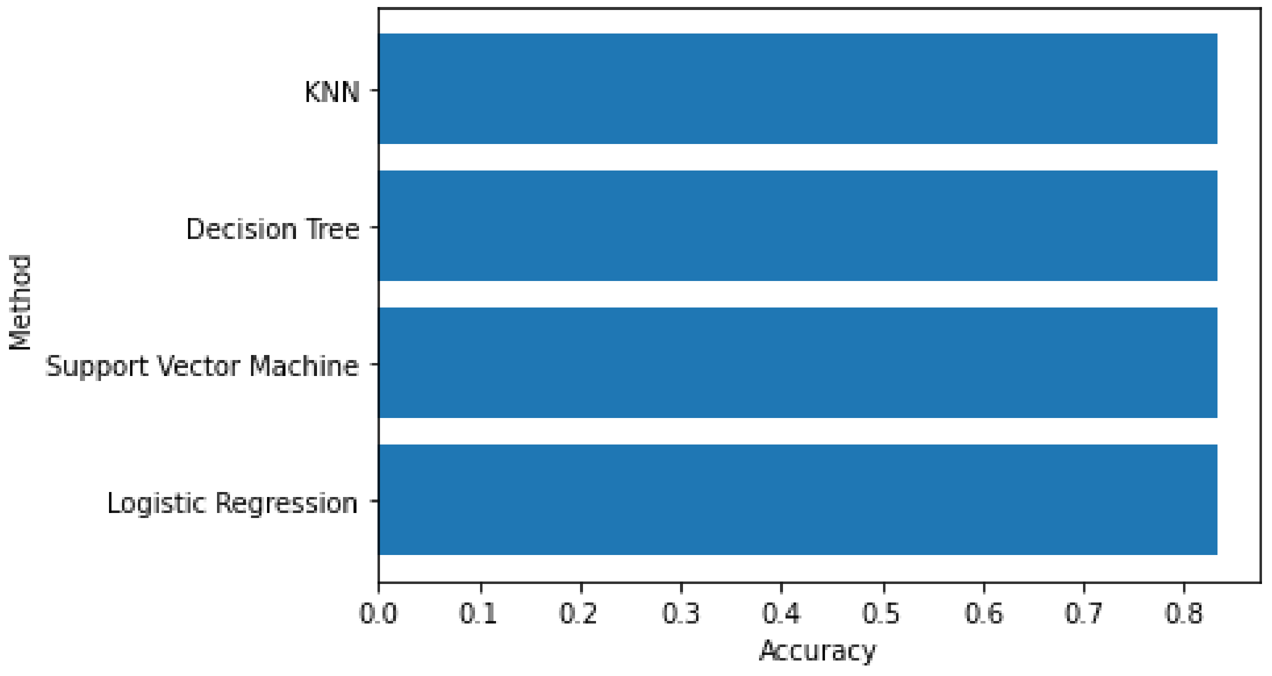


Section 5

Predictive Analysis (Classification)

Classification Accuracy

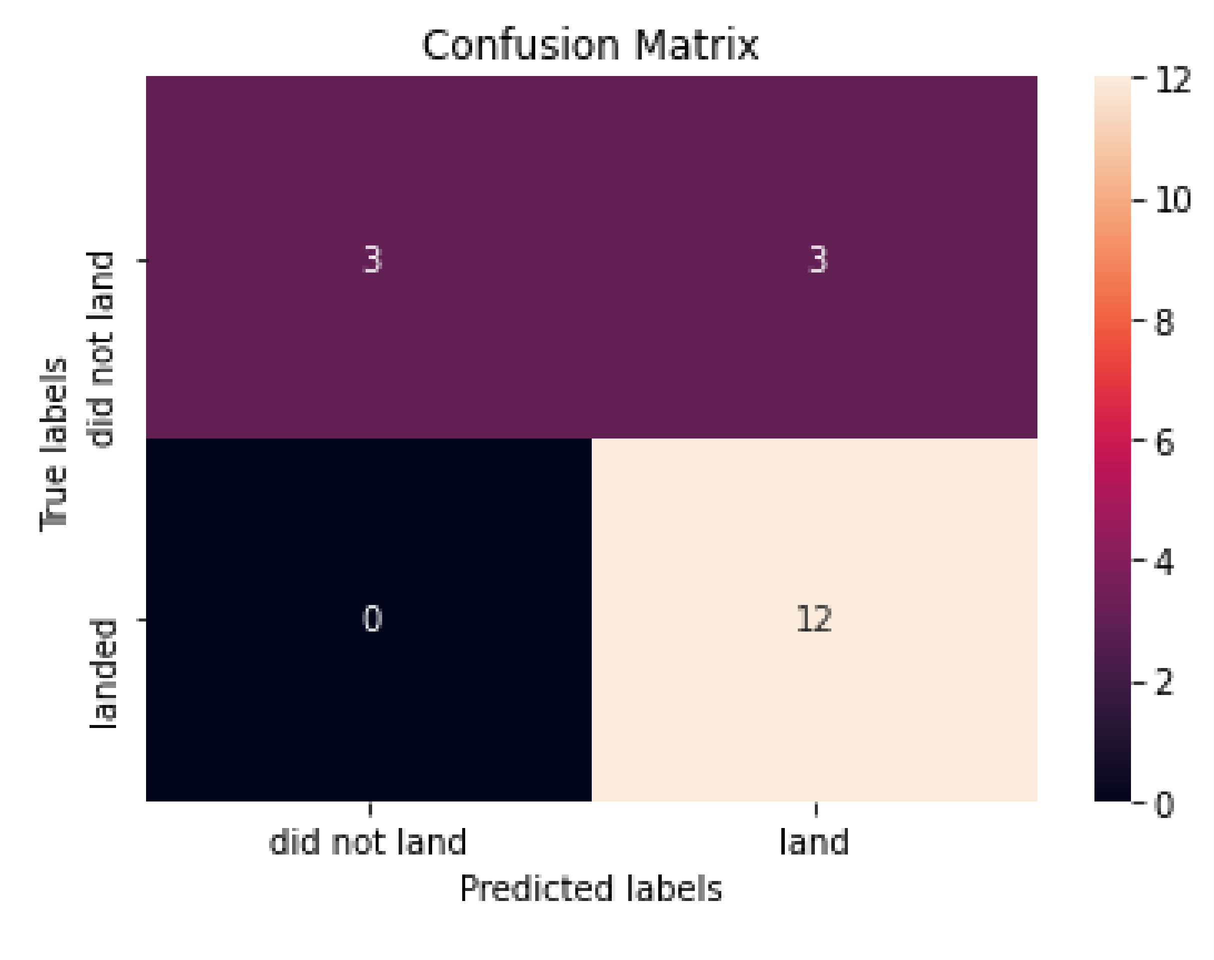
- **All models have the same accuracy (83.3%) for the test set**
- Noting that the test set size was small with only 18 records
- More data is needed to determine the optimal model



	Model	Accuracy
0	Logistic Regression	0.833333
1	Support Vector Machine	0.833333
2	Decision Tree	0.833333
3	KNN	0.833333

Confusion Matrix

- The confusion matrix were the same for all models
- Correctly predicted all successful landing (12) and the (3) failed landings
- 3 False Positive predictions where successful landings predicted while actually it were failures
- Overall, **the models predict successful landings**.



Conclusions

- As the number of flights increased, the success rate increased, and recently it has exceeded 80%.
- Orbital types SSO, HEO, GEO, and ES-L1 have the highest success rate (100%).
- The launch site is close to railways, highways, and coastline, but far from cities.
- KSLC-39A has the highest number of launch successes and the highest success rate among all sites.
- The launch success rate of low weighted payloads is higher than that of heavy weighted payloads.
- In this dataset, all models have the same accuracy (83.33%), but it seems that more data is needed to determine the optimal model due to the small data size.

Appendix

- [Project Github Repository URL]
- [Coursera Applied Data Science Capstone Project]

Thank you!