WINNING THE SPACE RACE WITH DATA SCIENCE

Daniel J Hunt
06 December 2024

Outline

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

Executive Summary

What we know:

- Launching Rockets into Space is expensive and SpaceX has reduced cost per launch , by $100M, through recovering and re-using the first Stage Booster BUT
- Recovery of the first Stage Booster does not happen every launch therefore the actual cost is dependent on the success rate.

What we learnt:

- Success is based on launch location, orbit and payload.
- The success rate has climbed and is steady around 80%.
- We have a model that has an accuracy rating of 0.88 for launch success prediction.

How we learnt it:

- **Collecting** and **Wrangling** data from websites and open source information to determine simple ‘yes’ or ‘no’ for success.
- **Exploring** and **Analysing** the data, both visually and with statistics, reviewing factors such as payload, launch sites and flights numbers against the success/failure results.
- Using a variety of **Machine Learning** models to determine future success rates.

Introduction

Background:
The commercial space age is here and companies are making space travel affordable for everyone. There are a number of companies doing this but perhaps the most successful is SpaceX.

SpaceX advertises launches on its website with a cost of $62M; other providers cost upwards of $165M each: much of the savings is because SpaceX recovers and re-uses the first stage.

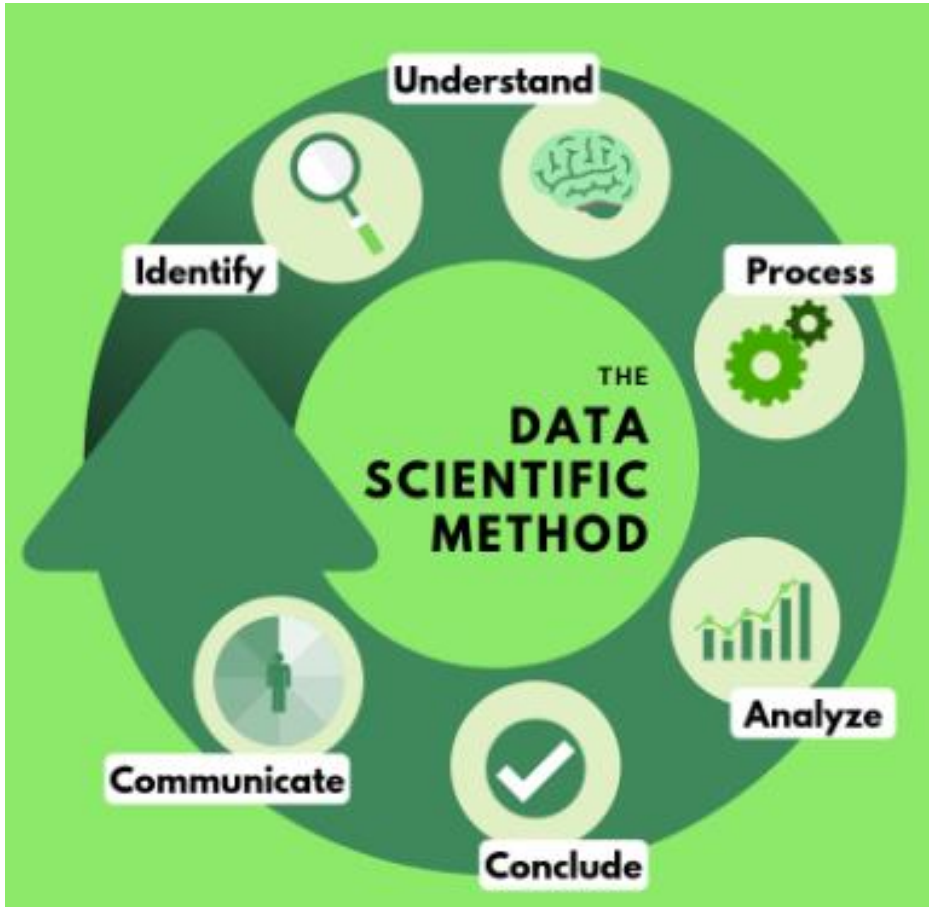
But this doesn't happen on every mission. Therefore, if we can determine future likely success, we can determine the actual cost of a launch.

Questions:
How successful are SpaceX in recovering the first stage and what might that success rate look like in the future?

What determines success?

Section 1

METHODOLOGY



Methodology: Executive Summary

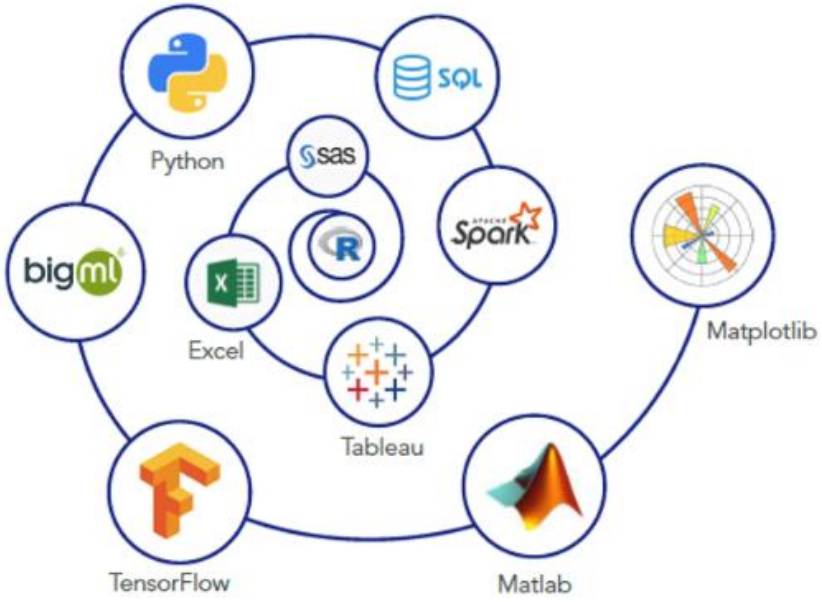
Data collection methodology:
Utilise SpaceX REST API and webscraping to obtain open source data.

Perform data wrangling
Filter the data, address missing values, apply one-hot encoding and preparing for analysis and modelling to capture the data is a single dataframe dictionary.

Perform exploratory data analysis (EDA) using visualization and SQL
Utilise tools such as SQLlite, Pandas and Matplotlib to produce tables and charts

Perform interactive visual analytics using Folium and Plotly Dash
Develop Maps and dashboards to help better understand the data.

Perform predictive analysis using classification models
Build and tune classification models using Scikit learn and evaluative tools such as Jaccard and F1 scores



Data Collection – SpaceX API

Process:

Use an Application Programming Interface (API) to pull the raw data on rocket launches from the Space X website.

Manipulate the data so it is able to be transformed in to data you can interrogate and learn from.

This includes:

- Normalising the data
- Formatting the data into dictionary form
- Using the dictionary form to produce a dataframe
- Cleaning the data by selecting only the data you need and dealing with any missing data

This provides us with an appropriately robust set of data to use for exploration.

Github URL: https://github.com/djhunt25/IBM-Data-Science-Capstone/blob/main/DJ_lab1_Spacex_Data_Collection_API.ipynb

Data Collection – Scraping

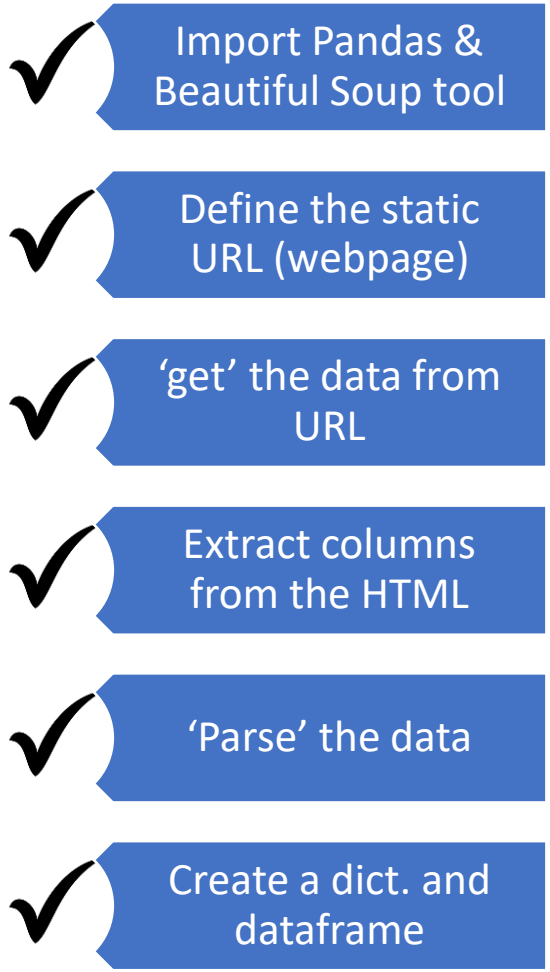
Process:

Extracting data from an open source, in this case Wikipedia, on Falcon 9 rocket launches.

This HTML based data is then manipulated to form a dataset.

This involved:

- Converting the data into a BeautifulSoup object
- Extracting the column names
- Parsing the HTML tables
- Formatting the data into dictionary form
- Using the dictionary form to produce a dataframe
- Saving the data as a comma separated values)CSV) file for later use.



Github URL: https://github.com/djhunt25/IBM-Data-Science-Capstone/blob/main/DJ_lab2_jupyter_webscraping.ipynb

Data Wrangling

Process:

The data wrangling involves taking the data we have collected and beginning to analyse it.

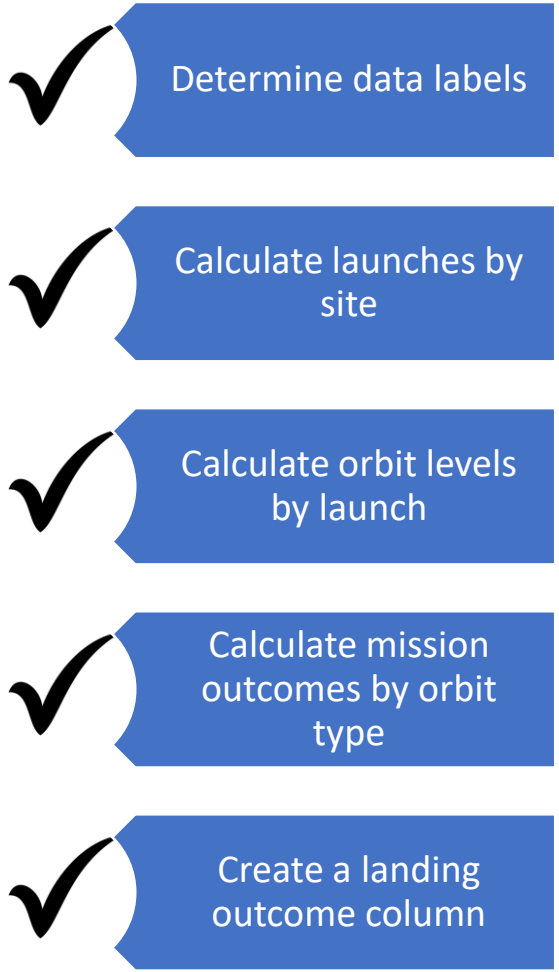
By using different methods to view the data we are able to form insights starters. These starters then help shape our next step.

This involved:

- Determining the categorical and numerical columns
- Extracting cumulative numbers on launches by the four different sites.
- Extracting the cumulative number of launches per each orbit level (generically the height and position the launch was delivering payload to).

This gave insights into the relationship between launches, sites and orbit whilst allowing us to filter the data base don launch success.

Github URL: https://github.com/djhunt25/IBM-Data-Science-Capstone/blob/main/DJ_lab3_jupyter_Data_wrangling.ipynb



EDA with SQL

Process:

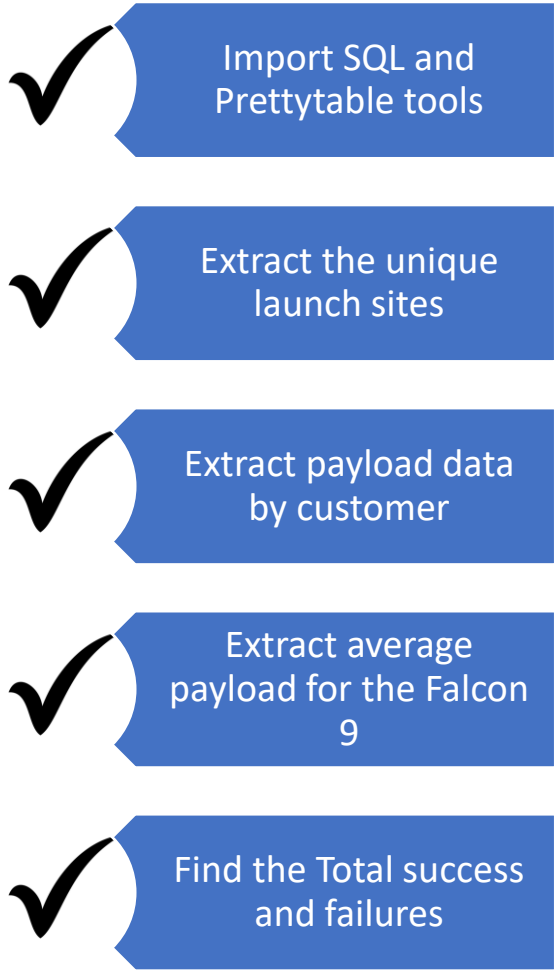
Exploratory Data Analysis (EDA), using tools such as SQL and Prettytable, allows us to extract Data in a table format to allow us to obtain insights.

This process involved:

- Determining the unique launch sites used (four)
- Extract the payload by specific customers, in this case NASA (45596 kg)
- Obtain the average payload per Falcon 9 launch (2928 kg)
- When the first successful landing occurred on a landing pad (22 Dec 2015)
- Combinations of tabulated data based on booster versions, launch sites and landing outcomes.

This provided us initial insights that helped guide the visualisation phase of EDA.

Github URL: https://github.com/djhunt25/IBM-Data-Science-Capstone/blob/main/DJ_lab4_EDA_SQL_sqllite.ipynb



EDA with Data Visualization

Process:

Data visualisation enables us to get better insights into the dataset (in this case 90 flights).

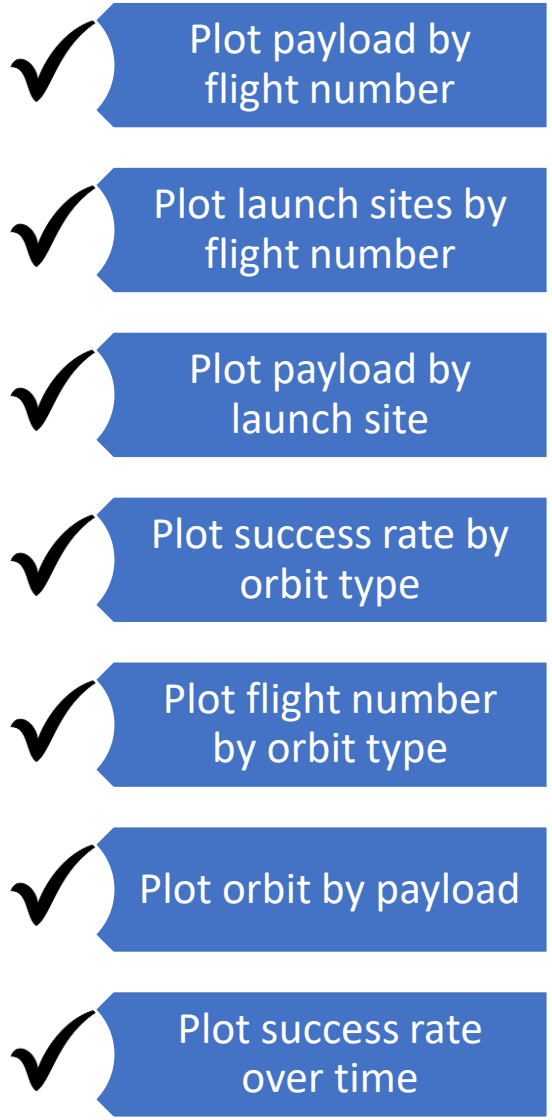
Using tools such as Matpltlib and Seaborn we were able to construct scatterplots, bar charts and line charts.

This focussed on:

- Total flights numbers
- Payload mass
- Orbit type
- Launch success

Combining these on different charts we begun to see if there are any relationships.

Github URL: https://github.com/djhunt25/IBM-Data-Science-Capstone/blob/main/DJ_lab5_jupyter_EDA_data_viz.ipynb



Build an Interactive Map with Folium

Process:

Plotting locations and incidents on a Map allows us to visualise and access data easily. In this case we used the Folium tool.

This included:

- Plotting the launch sites (all coastal, three on east coast (Florida) and one on west coast (California and near the Equator)
- Plot markers for each successful (green) and failed (red) launch and its location to see if there is any relationship.
- Plot objects to discern distances from the coast, highways, railways and cities to visualise.
    - These are important factors, for logistics, access, workforce and launch.

This allowed us to see how launch site location my impact launch success and flight rates.

Github URL: https://github.com/djhunt25/IBM-Data-Science-Capstone/blob/main/DJ_lab6_jupyter_launch_site_location.ipynb

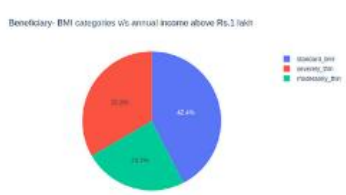
Build a Dashboard with Plotly Dash

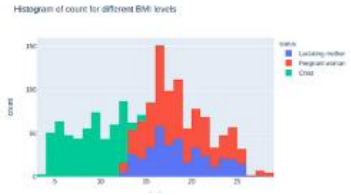
Process:

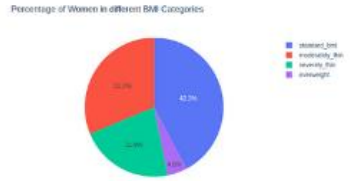
A dashboard allows data to be accessed easily, through a URL, by users other than the data team. In this case w e used the Plotly Dash tool to create an interactive dashboard.

This included:

- A drop down menu to look at different launch site statistics.
- A Pie charts to visualise launch success rates.
- A scatter chart showing payload by flight, together with a slider to allow different payload sizes to be viewed.
- Callback functions to update pie charts and slider as data became available.







Github URL: https://github.com/djhunt25/IBM-Data-Science-Capstone/blob/main/DJ_lab7_SpaceX_dash_app.py

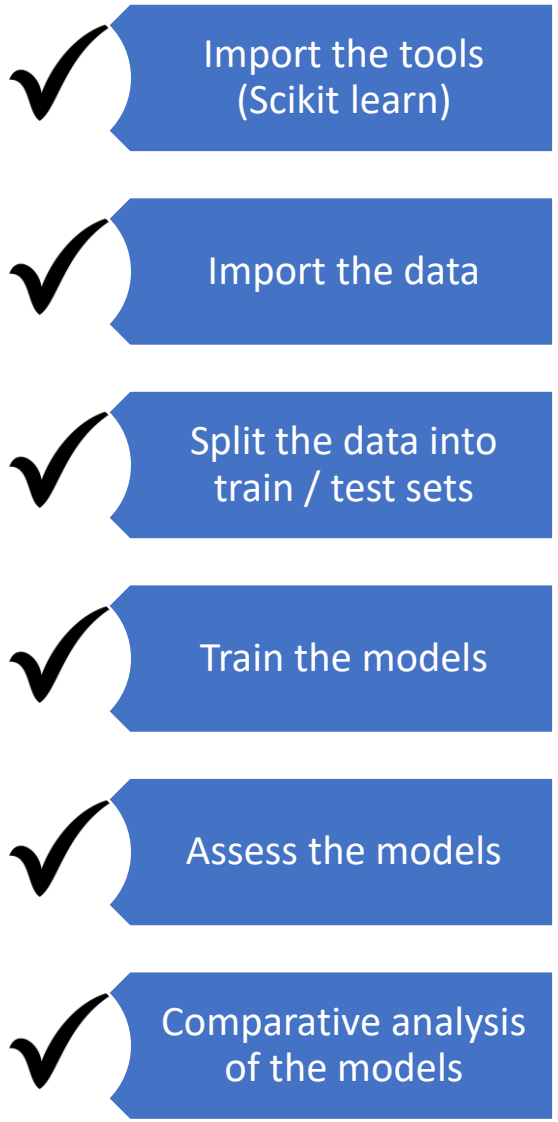
Predictive Analysis (Classification)

Process:

Use of machine learning models to develop a model to predict launch success based on a range of variables.

This included:

- A logistic regression model.
- A support vector machine model
- A tree model
- A KNN model
- An evaluation to compare the models accuracy
- This involved tuning and confusion matric analysis.



Github URL: https://github.com/djhunt25/IBM-Data-Science-Capstone/blob/main/DJ_lab8_jupyter_SpaceX_Machine%20Learning%20Prediction.ipynb

Results

Cape Canaveral is the most prolific launch site for the Falcon 9 based on both flight numbers and payload.

Success rates are highest for ES-L1, GEO, HEO, SSL orbits.

The success rate has grown steadily since 2013 and now sits at an 80% success rate.

Launch sites are close to the coast and have good access to railways, highways and people to support the operations.

The average payload for the Falcon 9 is 2928 kg.

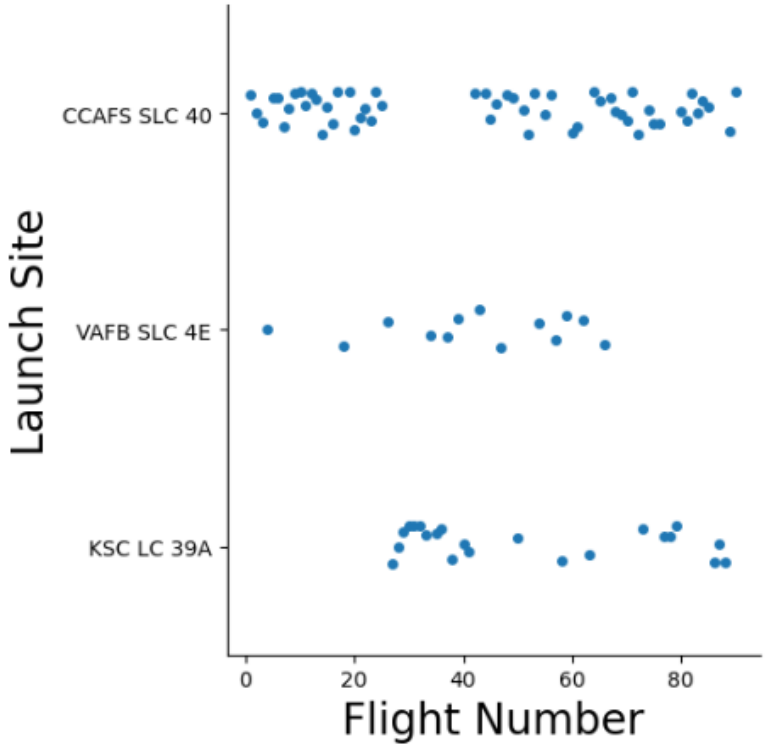
The Kennedy Space Centre is the most successful launch site.

A prediction SVM model with an accuracy rating of 0.87 was able to developed.

Section 2

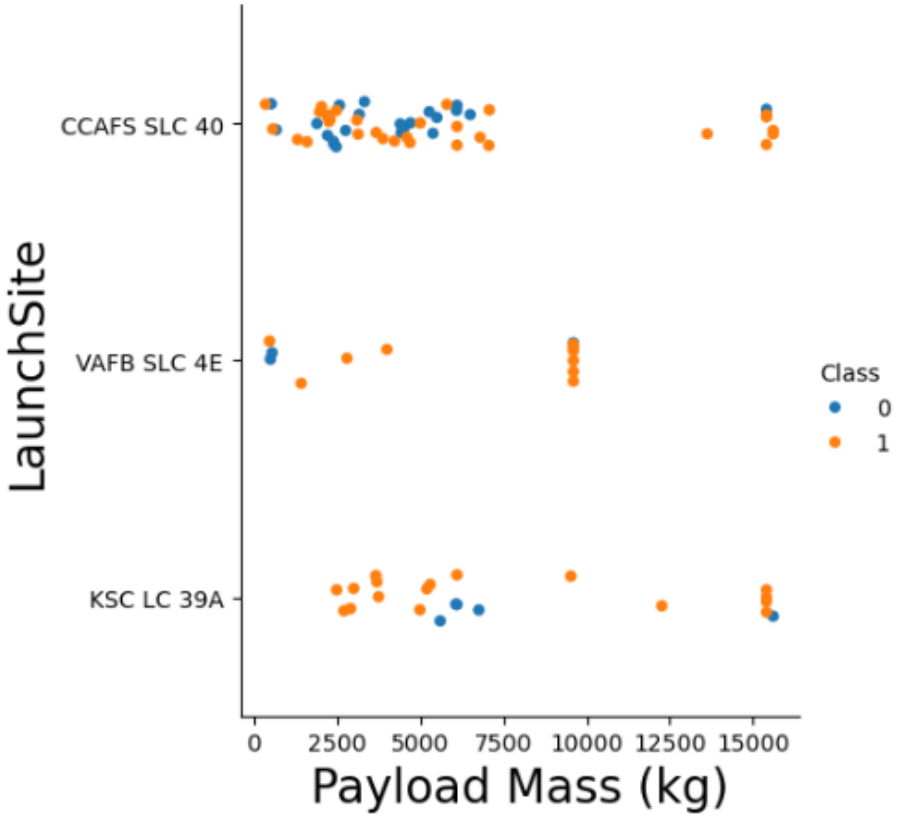
INSIGHTS DRAWN FROM EXPLORATORY DATA ANALYSIS (EDA)

Flight Number vs. Launch Site



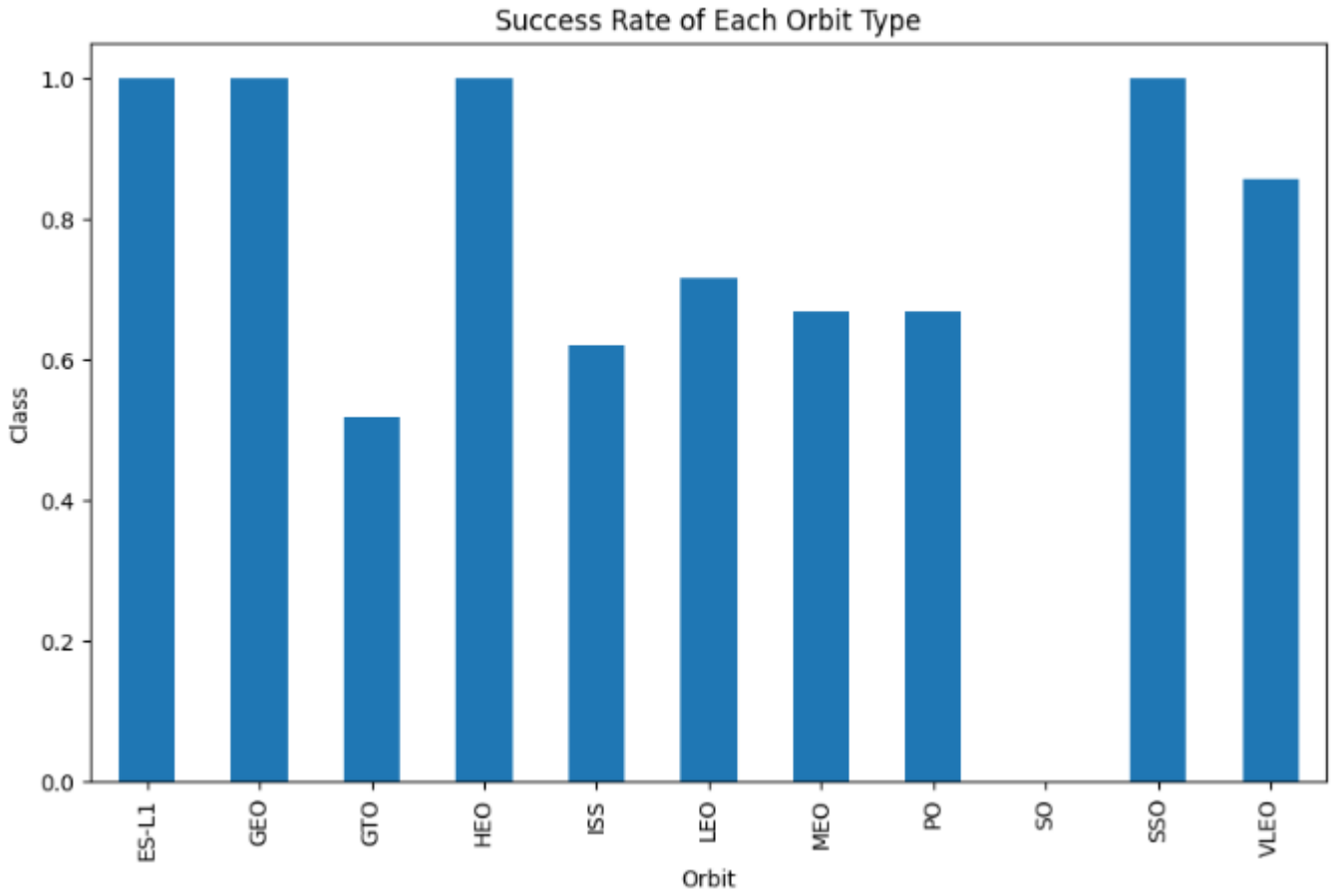

```
Now try to explain the patterns you found in the Flight Number vs. Launch Site scatter point plots:
The launch site CCAFS SLC40 has been the most prolific launch site and was used lammost exclusively for the first 20 flights.
Flight 20-40 then was predominantly launched out of KSC LC 39A before launches returnd, in the main, to CAFS SLC40.
This tends to indicate that SpaceX have learnt that CCAFS SLC40 is the 'best' launch site for the Falcon 9.
```

Payload vs. Launch Site



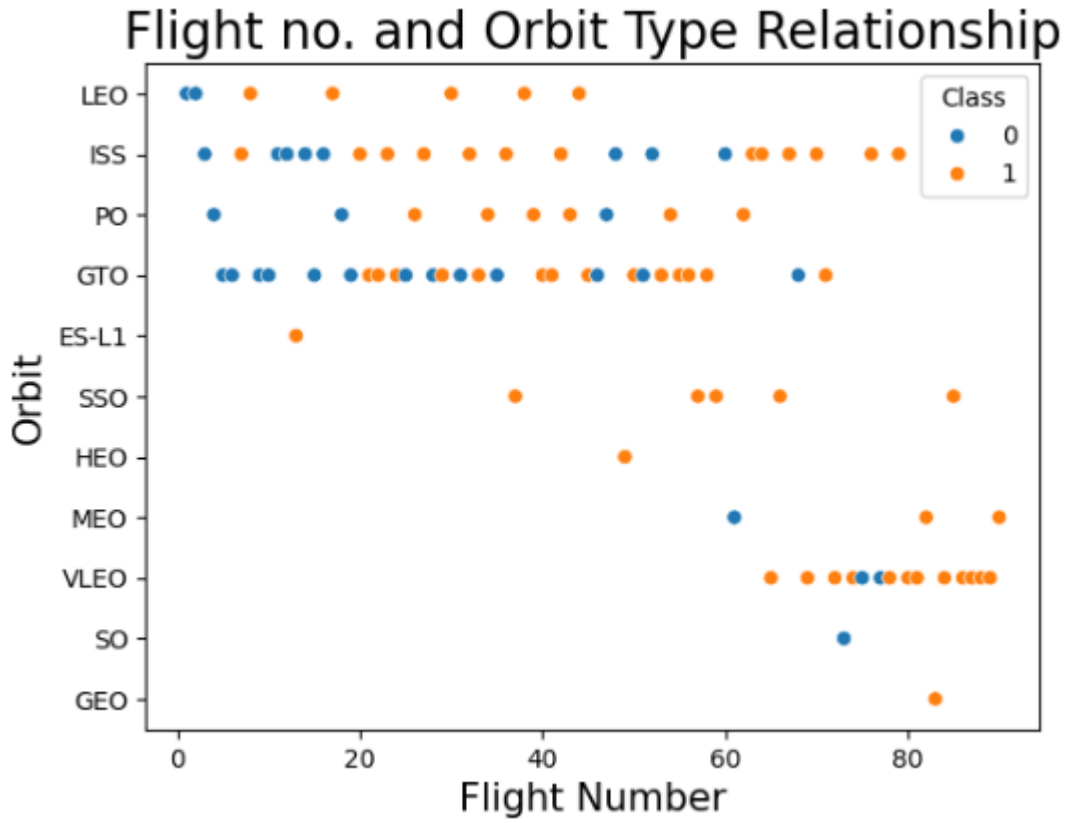
Now if you observe Payload Mass Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type



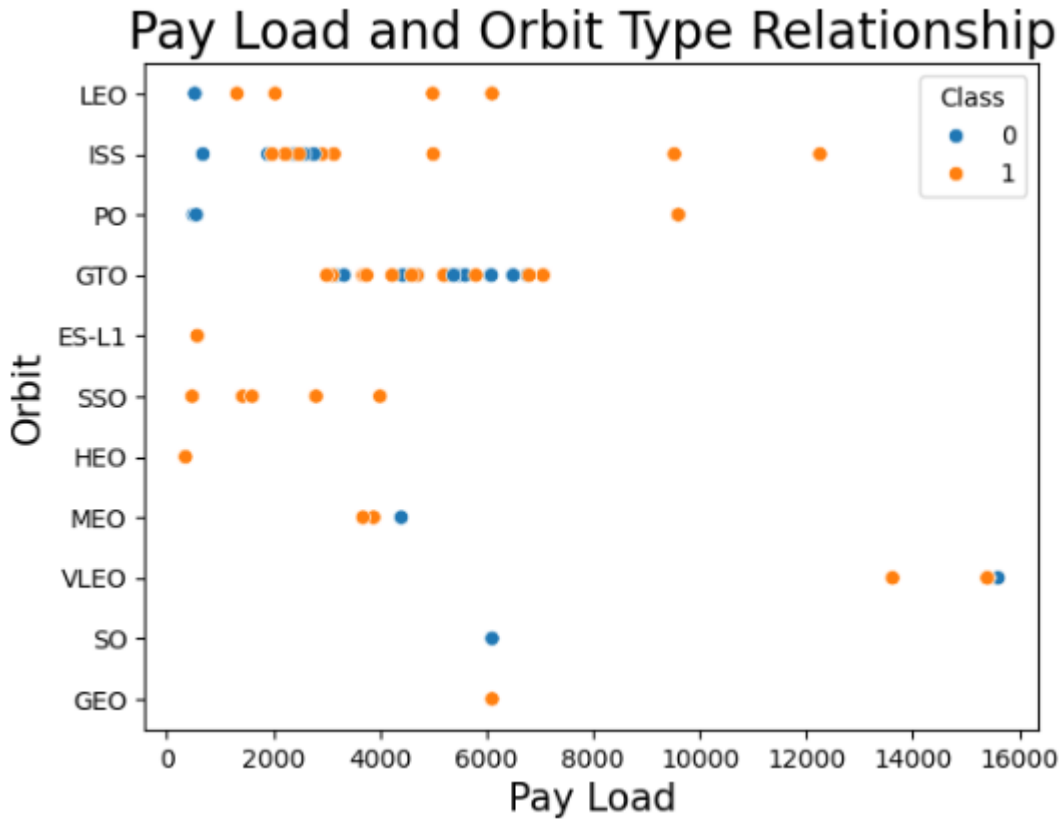

```
Analyze the plotted bar chart to identify which orbits have the highest success rates.
ES-11, GEO, HEO and SSO have the highest succes rate
```

Flight Number vs. Orbit Type



You can observe that in the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.

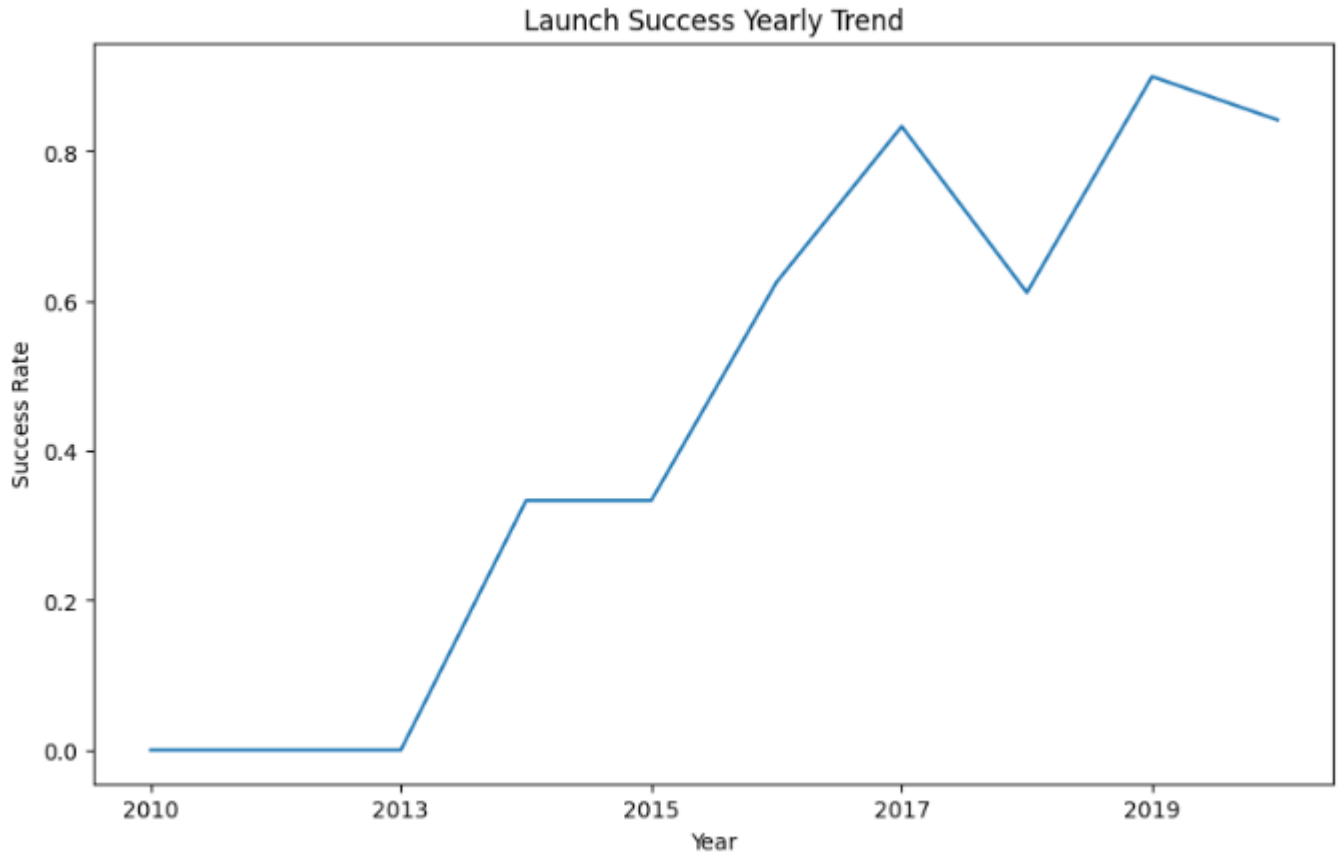
Payload vs. Orbit Type



With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend



You can observe that the sucess rate since 2013 kept increasing till 2020

All Launch Site Names

Task 1

Display the names of the unique launch sites in the space mission

[11]:
```
%%sql

select distinct Launch_Site from spacextbl
```

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

Done.
```

[11]:

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

There are four unique launch sites - two at Cape Canveral, one at the Kennedy Space Centre and one at Vandenburg Air Force Base (the only West Coast site).

Launch Site Names Begin with 'CCA'

Display 5 records where launch sites begin with the string 'CCA'

[12]:
```
%%sql

select * from spacextbl where Launch_Site LIKE 'CCA%' limit 5;
```

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

Done.
```

[12]:

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

Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

[13]:
```
%%sql

select sum(PAYLOAD_MASS__KG_) from spacextbl where Customer = 'NASA (CRS)'
```

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

Done.
```

[13]:

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

Task 4

Display average payload mass carried by booster version F9 v1.1

[14]:
```
%%sql

select avg(PAYLOAD_MASS__KG_) from spacextbl where Booster_Version LIKE 'F9 v1.1';
```

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

Done.
```

[14]:

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
%%sql

select min(Date) as 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

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%%sql

select Booster_Version from spacextbl where (PAYLOAD_MASS__KG_> 4000 and PAYLOAD_MASS__KG_ < 6000)
and (Landing_Outcome = 'Success (drone ship)');
```

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

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

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

```
%%sql

select Mission_Outcome, count(Mission_Outcome) as counts from spacextbl group by Mission_Outcome;
```

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

Mission_Outcome	counts
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

Task 8

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%%sql

select Booster_Version, PAYLOAD_MASS__KG_ from spacextbl
    where PAYLOAD_MASS__KG_ = (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

Task 9

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
%%sql

select Landing_Outcome, Booster_Version, Launch_Site, Date 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	Date
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	2015-01-10
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	2015-04-14

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

Task 10

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.

```
%%sql

select Landing_Outcome, count(*) as LandingCounts from spacextbl where Date between '2010-06-04' and '2017-03-20'
group by Landing_Outcome
order by count(*) desc;
```

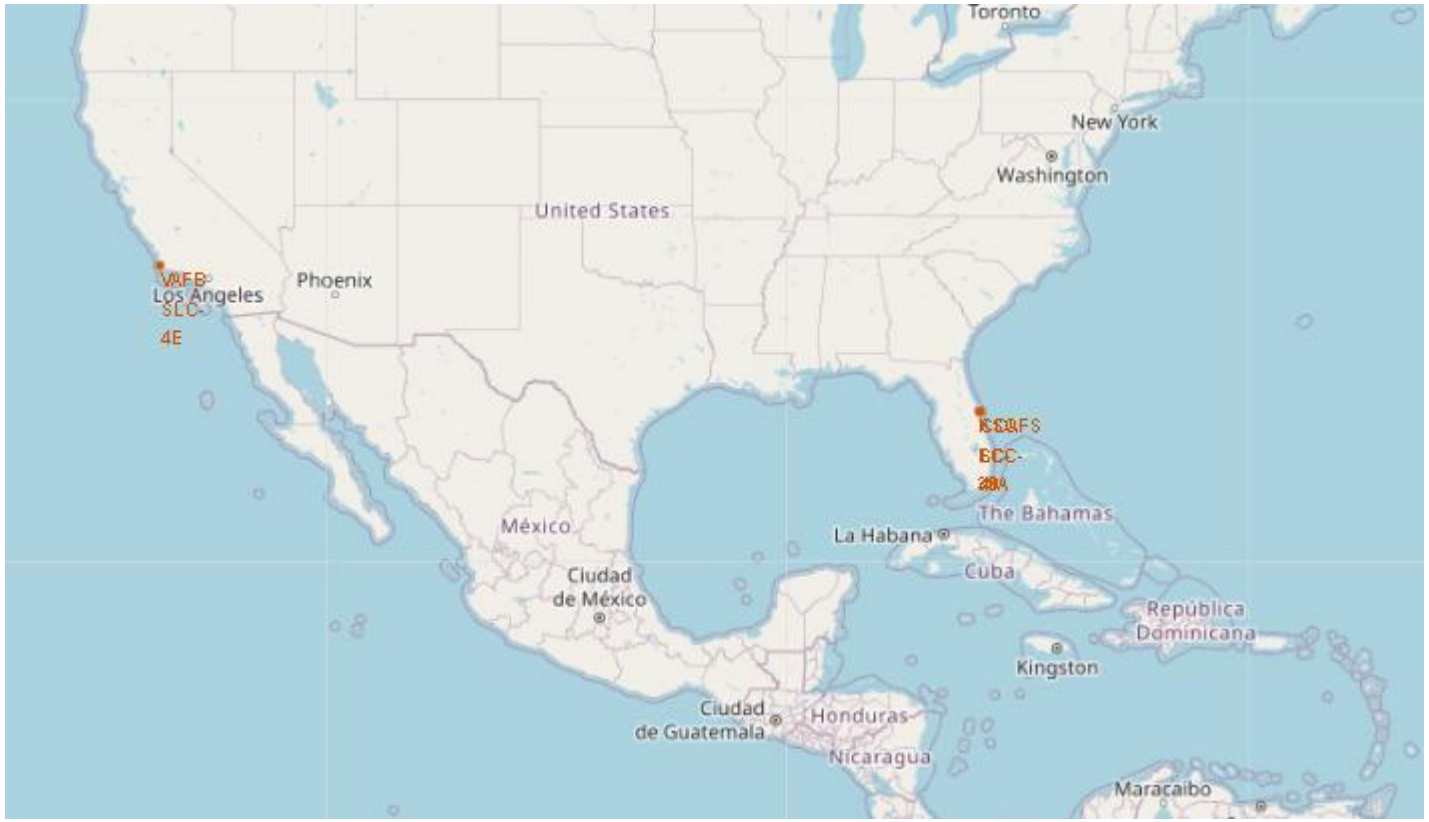
```
 * sqlite:///my_data1.db

Done.
```

Landing_Outcome	LandingCounts
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

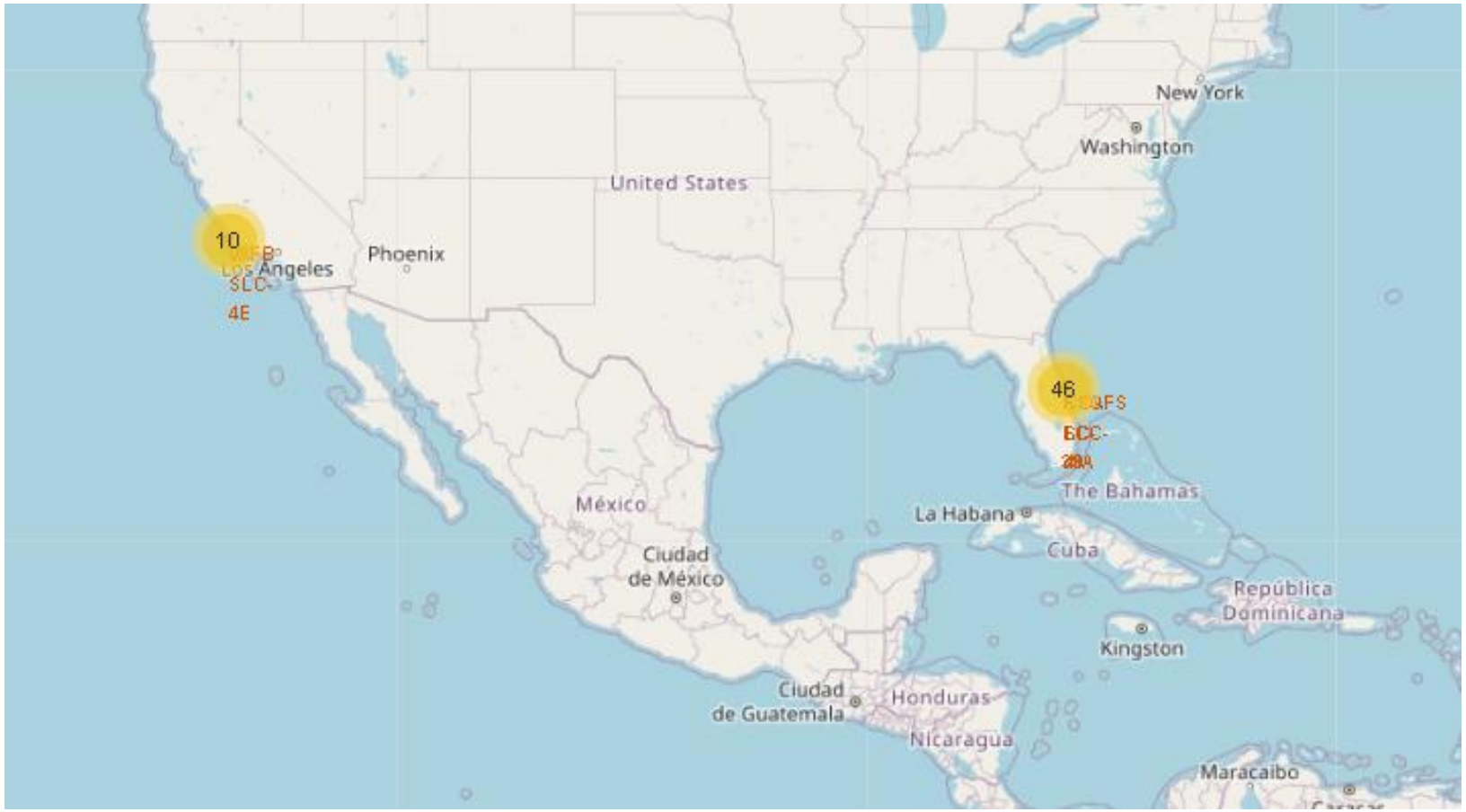
LAUNCH SITE PROXIMITIES ANALYSIS

SpaceX Launch Sites for the Falcon 9



- This shows the four launch sites in continental USA – three in Florida and one in California

Successful Launches by Site



- This shows the successful launches, by site, clustered within the West Coast and East Coast sites.

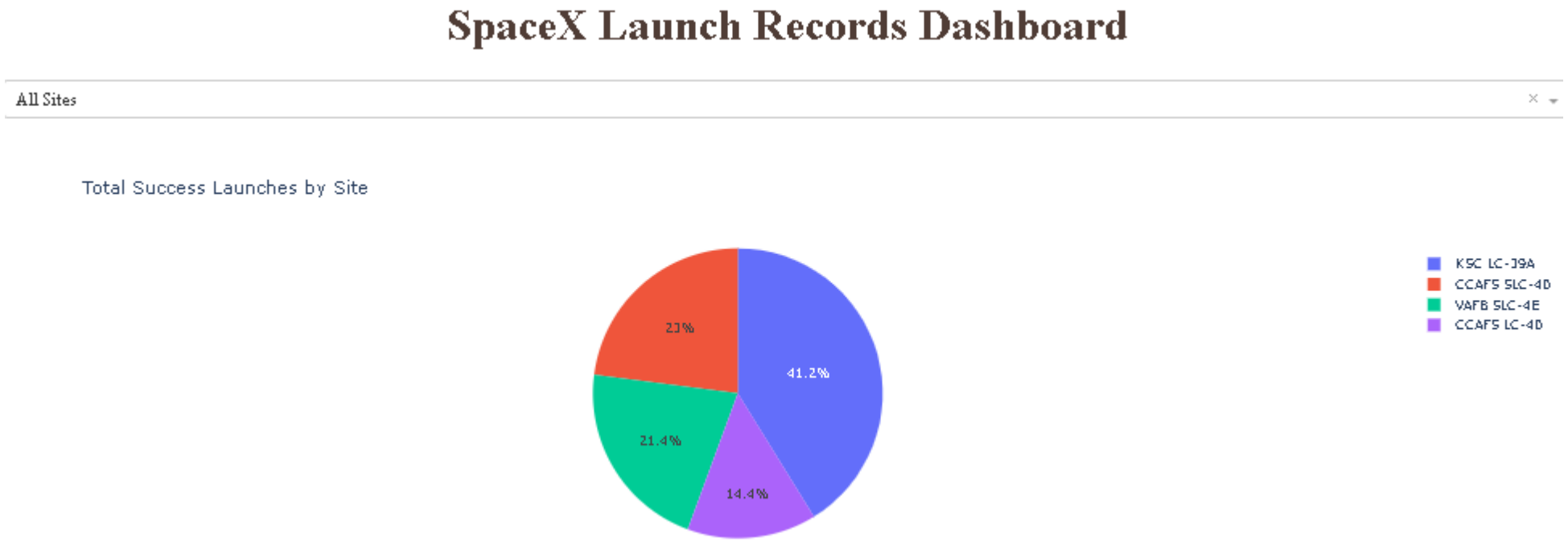
Distance to the closest Coastline

- This shows the closest coastline to the launch site at a distance of 0.86 km.

Section 4

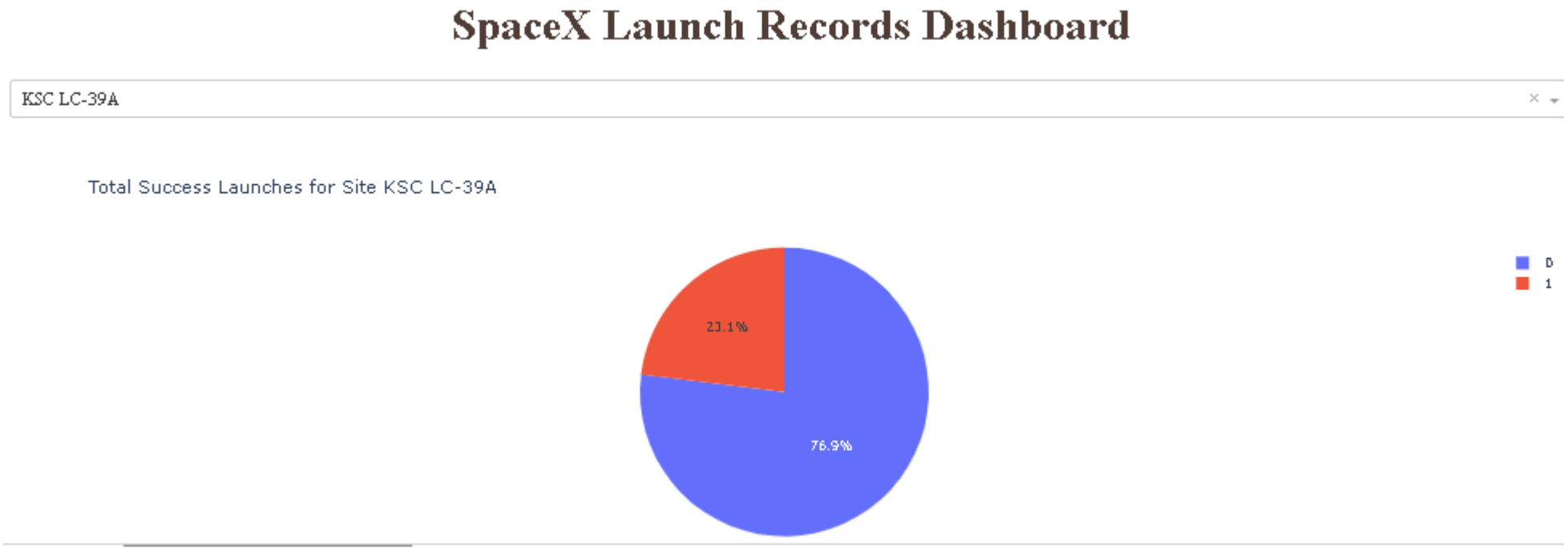
PLOTLY DASHBOARDS

Successful Launches by Site



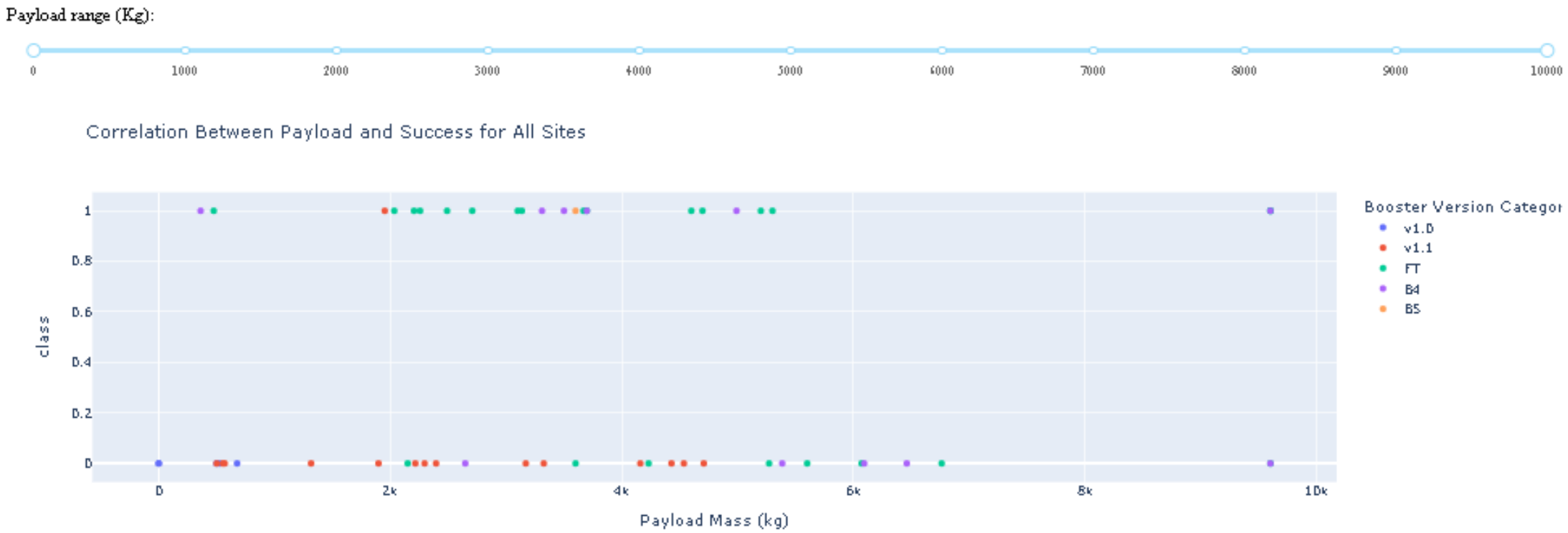
- This shows the Kennedy Space Centre has the best success rate.

Site with Highest Launch Success Rate



- This allows the user to view launch success rates for specific locations.

Payload versus Success

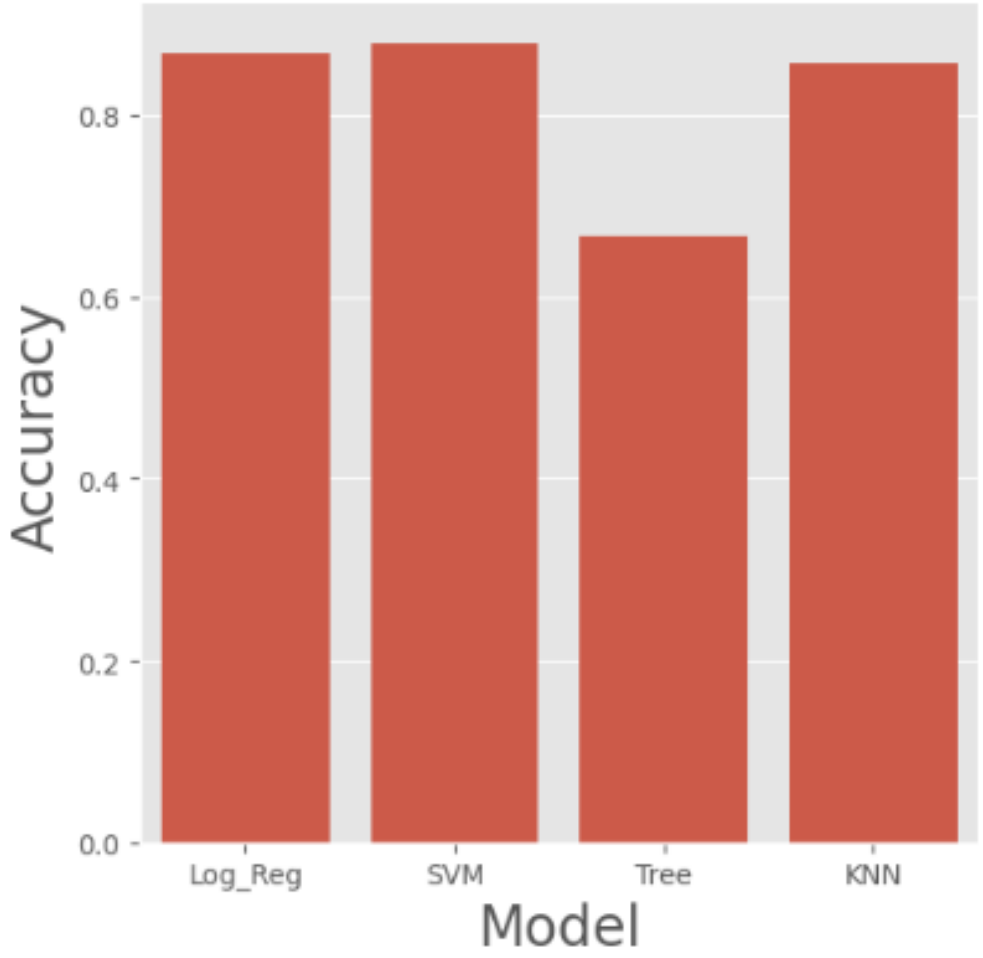


- Payloads between 2000 and 5000 kg have the best success rate.

Section 5

PREDICTIVE ANALYSIS (Classification)

Classification Accuracy

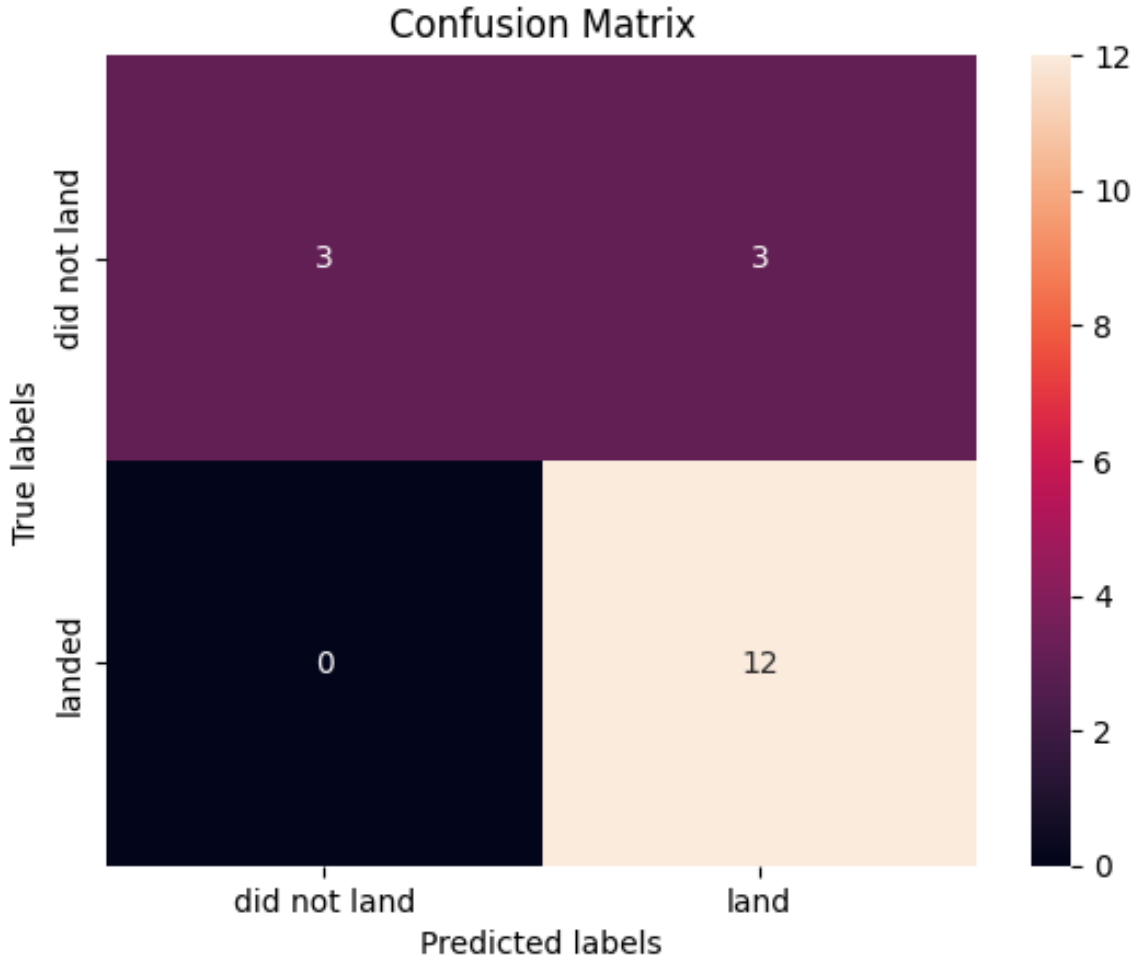


	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.666667	0.819444
F1_Score	0.909091	0.916031	0.800000	0.900763
Accuracy	0.866667	0.877778	0.666667	0.855556

Confusion Matrix

The SVM model confusion matrix shows the model was tested 18 times.

- 12 time it predicted success and a success occurred.
- 3 time it predicted failure and failure occurred.
- 3 times it predicted an outcome (either success or failure) and the opposite outcome occurred.



Conclusions

Success in recovering the Falcon-9 for reuse occurs:

- When launching from the East Coast (Florida)
- With a payload of between 2000-6000 kg
- For an orbit of either ES-L1, GEO, HEO or SSO
- We have a model that has an 88% accuracy for predicting successful recoveries based on these variables.

If we wish to compete with SpaceX these are the factors to consider

THANK YOU