IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

by Reuben M Dlamini
Date: 18 July 2024

Outline

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

Executive Summary

The aim of the project was develop a predictive model to determine the likelihood of a stage 1 successful landing as part of an investigation by SpaceY. The initial step included retrieving the data through an API and web scraping to dataframes which were subsequently cleaning including filing in missing data where possible. As part of the data wrangling, a new class column was created to characterize successful launches as 1 and 0 if unsuccessful. The was ensued by a exploratory data analysis through SQL queries, visual charts and dynamic dashboard. Thereafter the maps of the launch sites were evaluated to determine the best location of launch sites. Thereafter a machine learning models were evaluated to determine the best model to predict success rate.

Success rate increased with increasing number of flights and with the passage of time.Pay load of masses between 2000 kg and 6000 kg were determined to be the most ideal. The east coast was the best location for successful launches. The decision tree was the best predictive model with an accuracy of 93%.

Introduction

- Project background and context

The aim of this project is to establish the conditions that enable SpaceX to be able land their first stage rockets and be able to reuse the it. and to apply those conditions in a new company Space Y that offers the same services. This is because SpaceX charges significantly less fees ($62mn) for space trips than competitors ($165m) on account of being able to reuse the first stage component of the rocket. Furthermore this will enable more accurate charging if the landing outcome is known.

A machine learning model is used to determine to predict when Space X will be able to reuse the first stage.

- Problems to be answered
  - Which variables have a high correlation to a safe landing of the first stage?
  - What conditions within those parameters are likely to contribute to the successful landing of the first stage?
  - Which model best predicts the outcome?
  - Can the data and model be used to predict successful landings?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
- Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

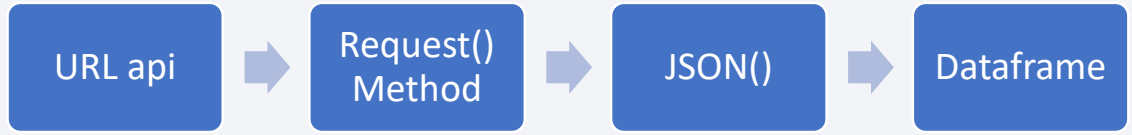
Data Collection

- Data was collected through an Api using the request method. In addition web scrapping was applied to determine
- These are described in more detail in the slides that follow.

Data Collection – SpaceX API

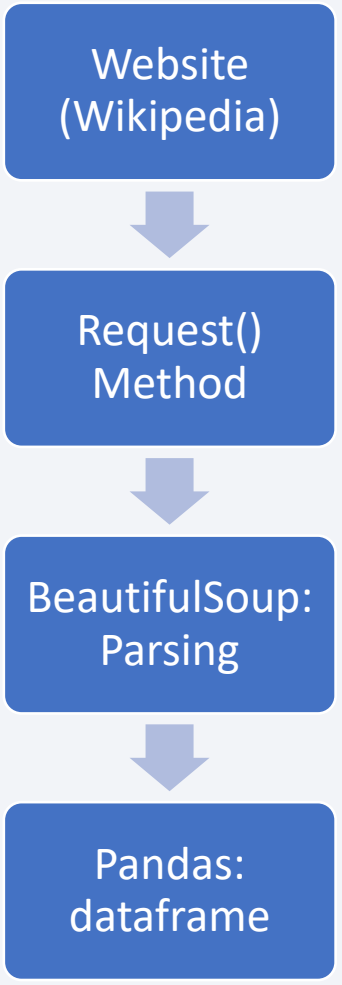
- The SpaceX Rest API with endpoint api.spacexdata.com/v4/launches/past is used to collect data from SpaceX.
- Using a get () request method a Json () file is created and normalised.
- Thereafter, the data was converted to dataframe was used method, the data will be converted to data frame.
- Notebook URL Link:

https://github.com/Reuben-D/SpaceY/blob/main/jupyter-labs-spacex-data-collection-api_RDv1.ipynb



Data Collection - Scraping

- Falcon 9 Data was also collected from Wikipedia Link.
- The requests method was used to obtain the data, then it was parsed with BeautifulSoup. Falcon 9 data was then parsed from the HTLM data.
- Notebook Link: https://github.com/Reuben-D/SpaceY/blob/main/jupyter-labs-webscraping_RDv1.ipynb



Data Wrangling

- After data was collected, the goal was to create successful outcomes with label 1 and label 0 for unsuccessful landings. This was for the machine learning aspect later in the project.
- Value_counts() was used to count the number launches on each site, number and occurrence of mission outcomes on the basis of orbits.
- Based on the landing outcome, the result of was appended on a new column ‘class’ from the ‘Outcome’ of the dataframe. This was achieved using if and for statements to create a list “landing class”.
- Notebook Link: https://github.com/Reuben-D/SpaceY/blob/main/labs-jupyter-spacex-Data%20wrangling%20RD.ipynb

EDA with Data Visualization

- The following charts were generated:
  - Flight Number vs Launch Site: To ascertain launch and Flight Landing had a relationship.
  - Pay load Mass vs. Launch site: To ascertain if pay load mass and launch site had a relationship..
  - Success rate vs. orbit type: Effect of orbit type on landing success rate.
  - Flight Number vs. Orbit Type: Establish if there was relationship between flight number and orbit type.
  - Payload vs. Orbit Type: Determine if there is a relationship between Payload and Orbit Type
  - Launch Success Yearly Trend: Determine if Launch success trend changed over time.
- Notebook Link: https://github.com/Reuben-D/SpaceY/blob/main/edadataviz_RD.ipynb

EDA with SQL

- As part of EDA the following SQL queries were executed to determine:
  - Unique launch site names
  - 5 records where launch sites begin with `CCA`
  - Total payload carried by boosters from NASA
  - Average payload mass carried by booster version F9 v1.1
  - Date of first successful ground pad landing.
  - Boosters which have success in drone ship and have payload mass between 4000 and 6000 kg
  - Total number of successful and failure mission outcomes
  - Names of the booster versions which have carried the maximum payload mass
  - Month names, failure landing outcomes in drone ship ,booster versions, launch site in year 2015
  - Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20
- Notebook Link: https://github.com/Reuben-D/SpaceY/blob/main/jupyter-labs-eda-sql-coursera_sqllite_RD.ipynb

Build an Interactive Map with Folium

- The following maps are shown
  - Circle markers showing launch sites: Determine where most sites are located.
  - Success count markers at launch sites: Determine if location has an effect on the success rate.
  - Proximity to launch sites with lines: Determine ideal location of site.
- Notebook Link: https://github.com/Reuben-D/SpaceY/blob/main/lab_jupyter_launch_site_locationRDv1.ipynb

Build a Dashboard with Plotly Dash

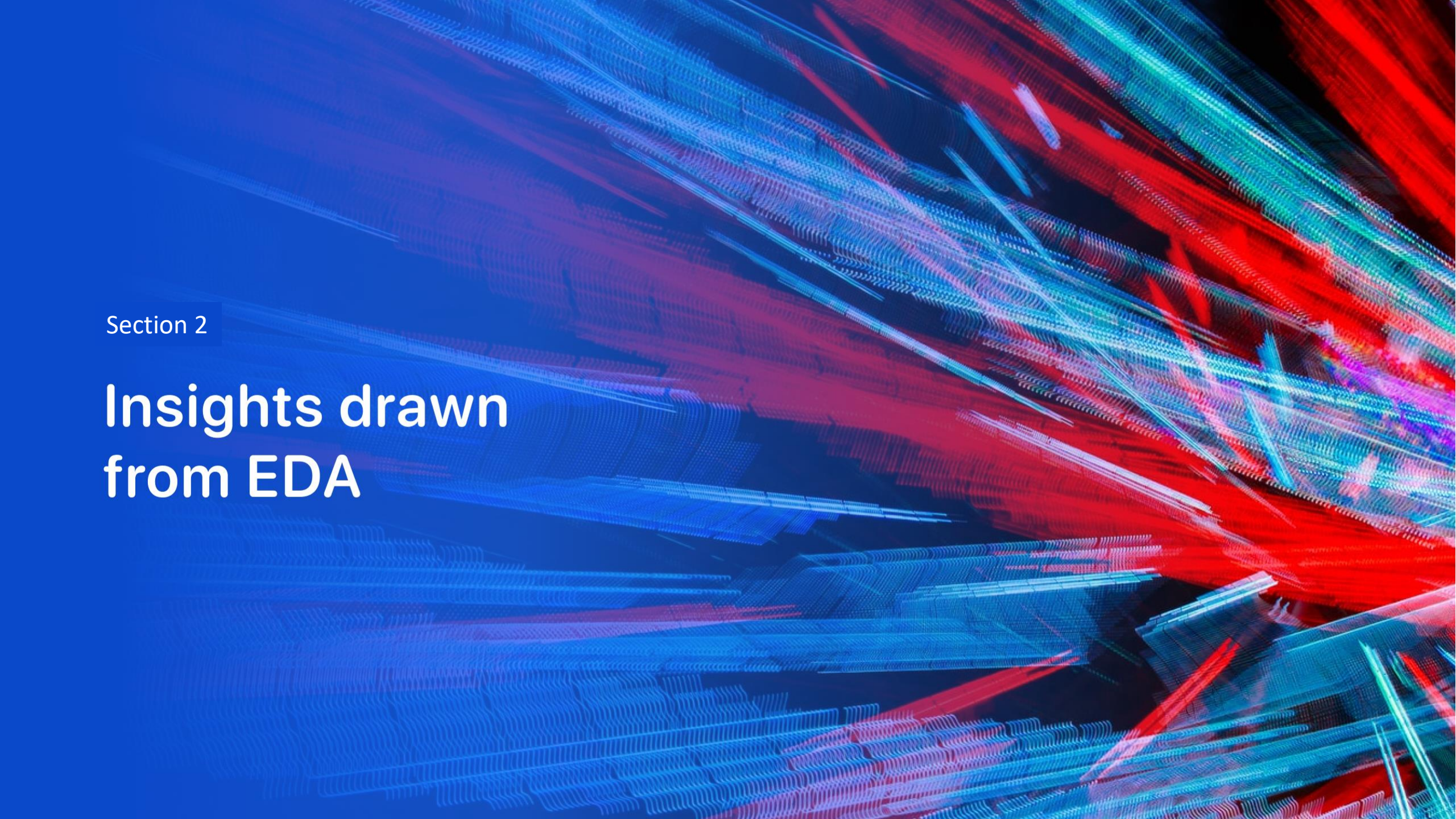
- The following were the charts presented in the Dashboard
  - Successful launches by site if all sites were selected: To determine the site with the most successful launches
  - Success rate by site if an individual site was selected: To establish the success rate of each site
  - Payload vs class for a given launch site and pay load range: To investigate the relationship between payload and launch site
- Notebook Link: https://github.com/Reuben-D/SpaceY/blob/main/spacex_dash_appRD.py

Predictive Analysis (Classification)

- Predictive analyses followed the following sequence:
- The cleaned data was split into independent (X) and dependent (Y) variables prior to normalising the X data.
- The data was further split into test and training data with 0,2 test size and random state being 2.
- A GridSearchCV() analyses with a given set of parameters and10 folds for each analysis for each model was used.
- The following machine learning models were used in the grid search
    - Logistic regression
    - SVM: Support Vector Machines
    - Decision tree
    - KNN: K-nearest Neighbors
- The models were assessed on the various parameters to evaluate which provided the best results.
- A confusion matrix was further used to analyse the model.
- Notebook Link: https://github.com/ReubenD/SpaceY/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5_RDv1.ipynb

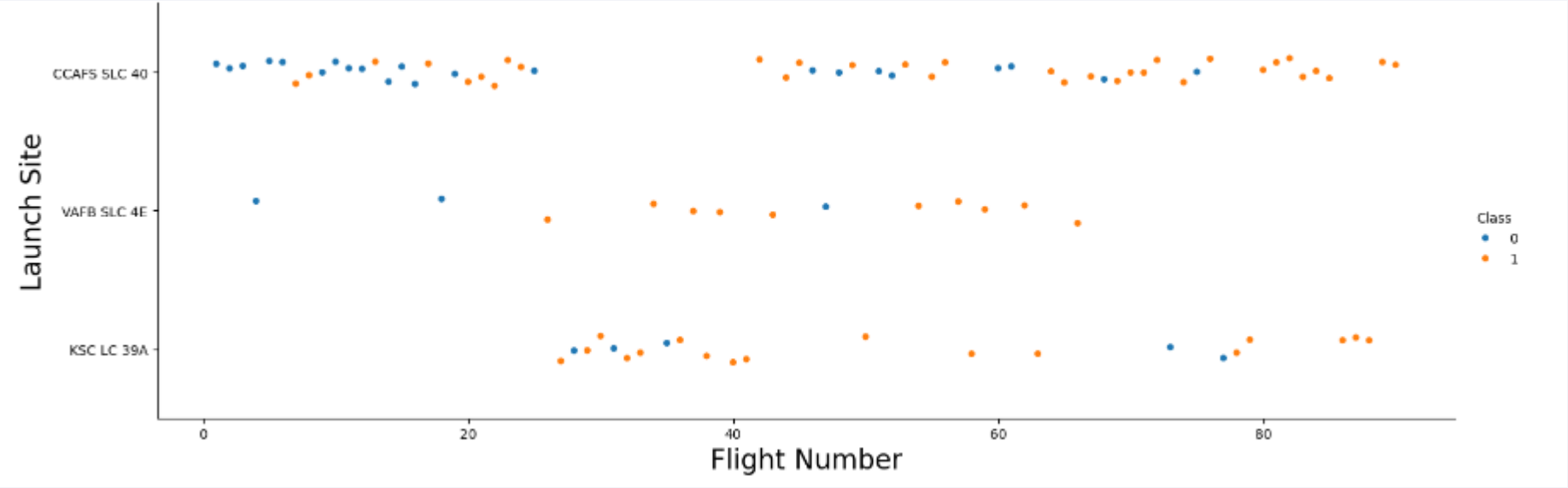
Results

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


Section 2
Insights drawn from EDA

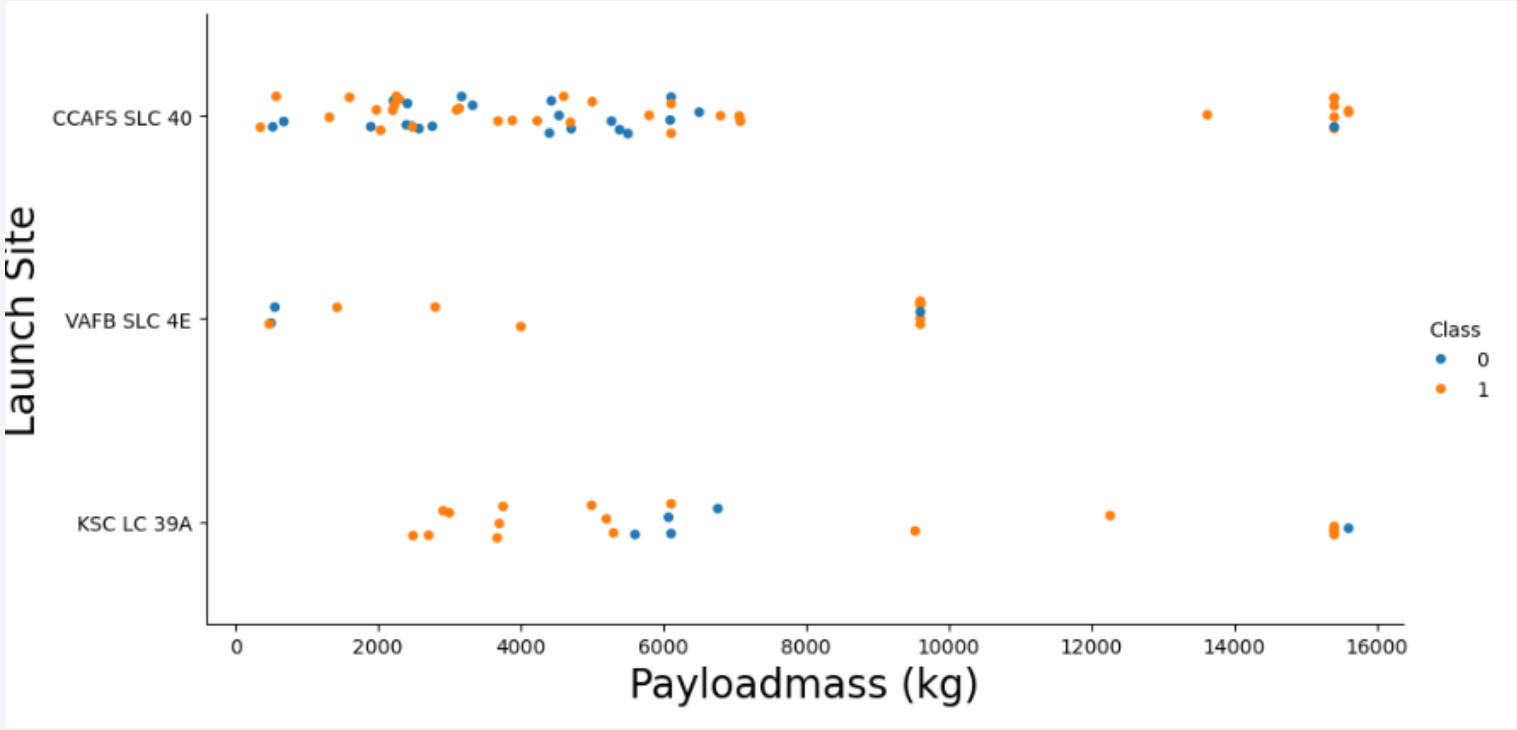
Flight Number vs. Launch Site

- The a is a noticeable trend of the a higher chances of success as the the payload mass increases as evidenced by a higher concentration of orange dots with increasing pay load mass. In addition, as with more iterations, success seems more likely.



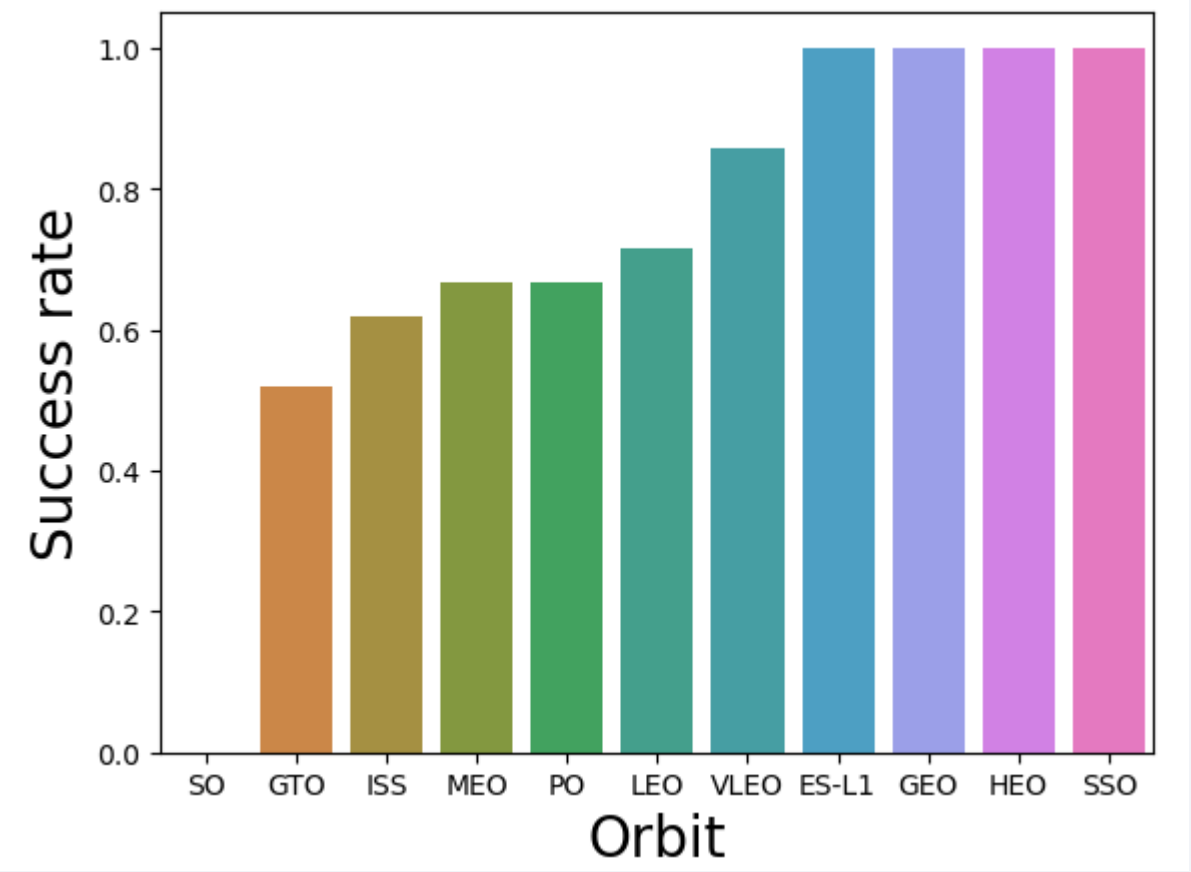
Payload vs. Launch Site

- As the pay load increases across all sites, the higher the chances of successful landings. In addition launch site VAFB SLC 4E and KSC LC 39A had higher success rates in comparison to CCAFS SLC 40.



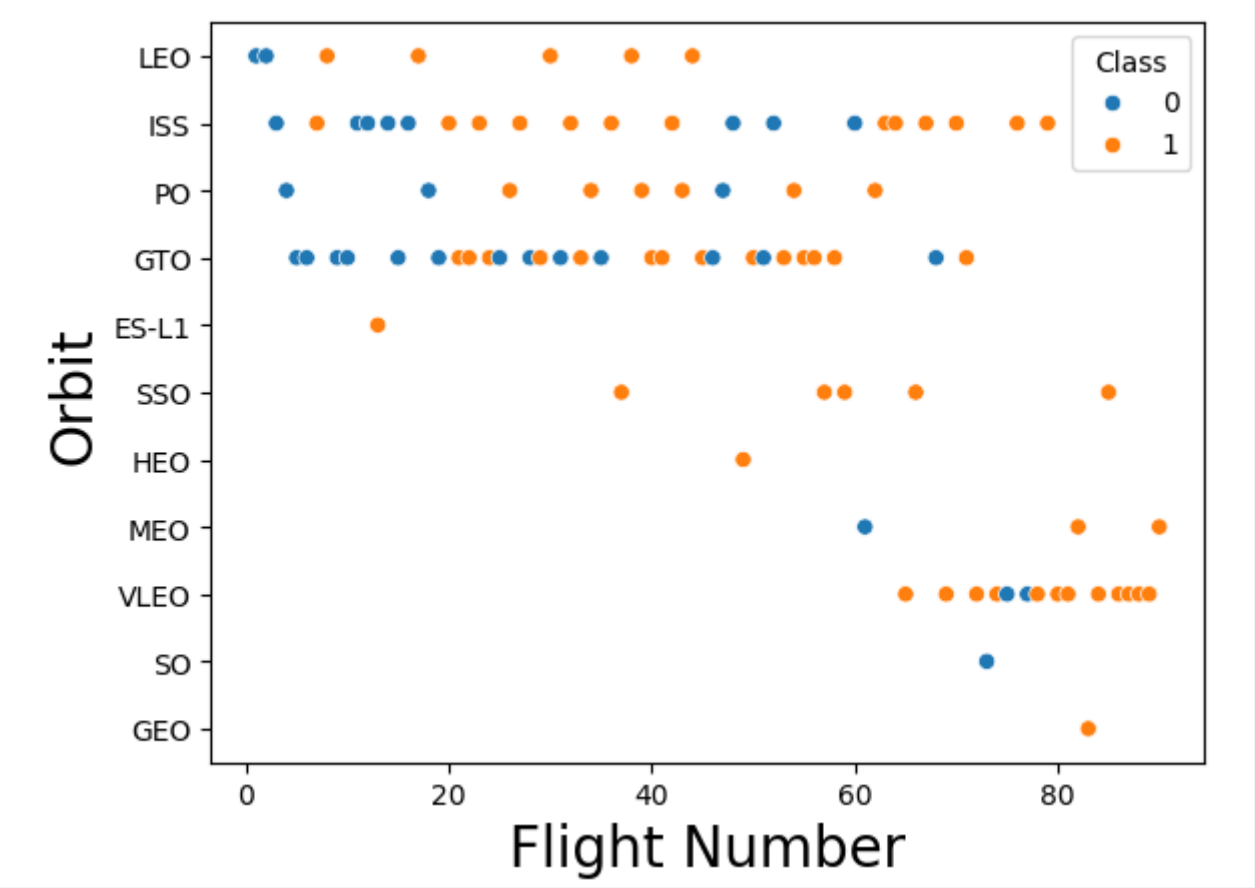
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO and SSO had the highest success rates.



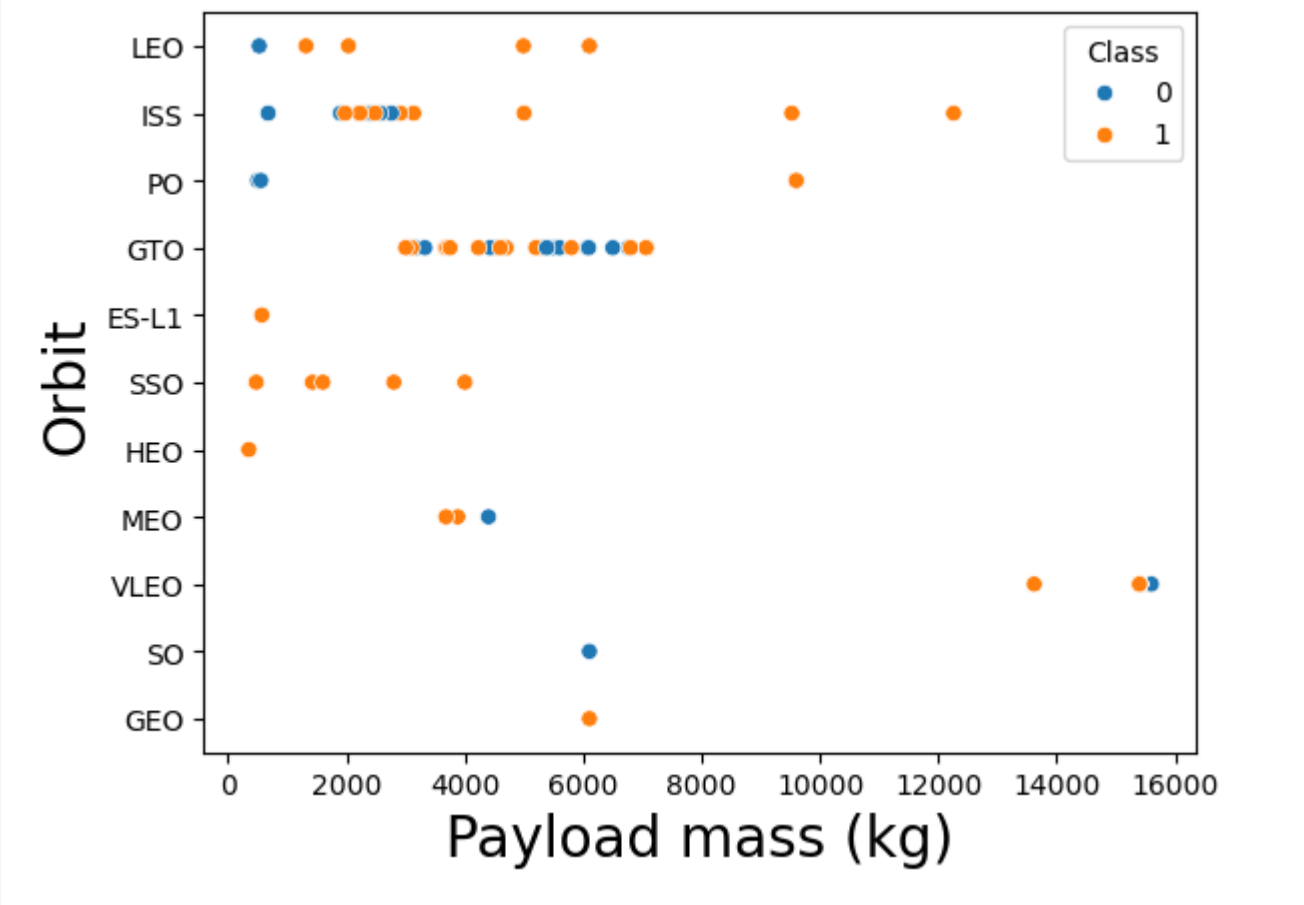
Flight Number vs. Orbit Type

- The success rate seemed to increase with higher flight number.



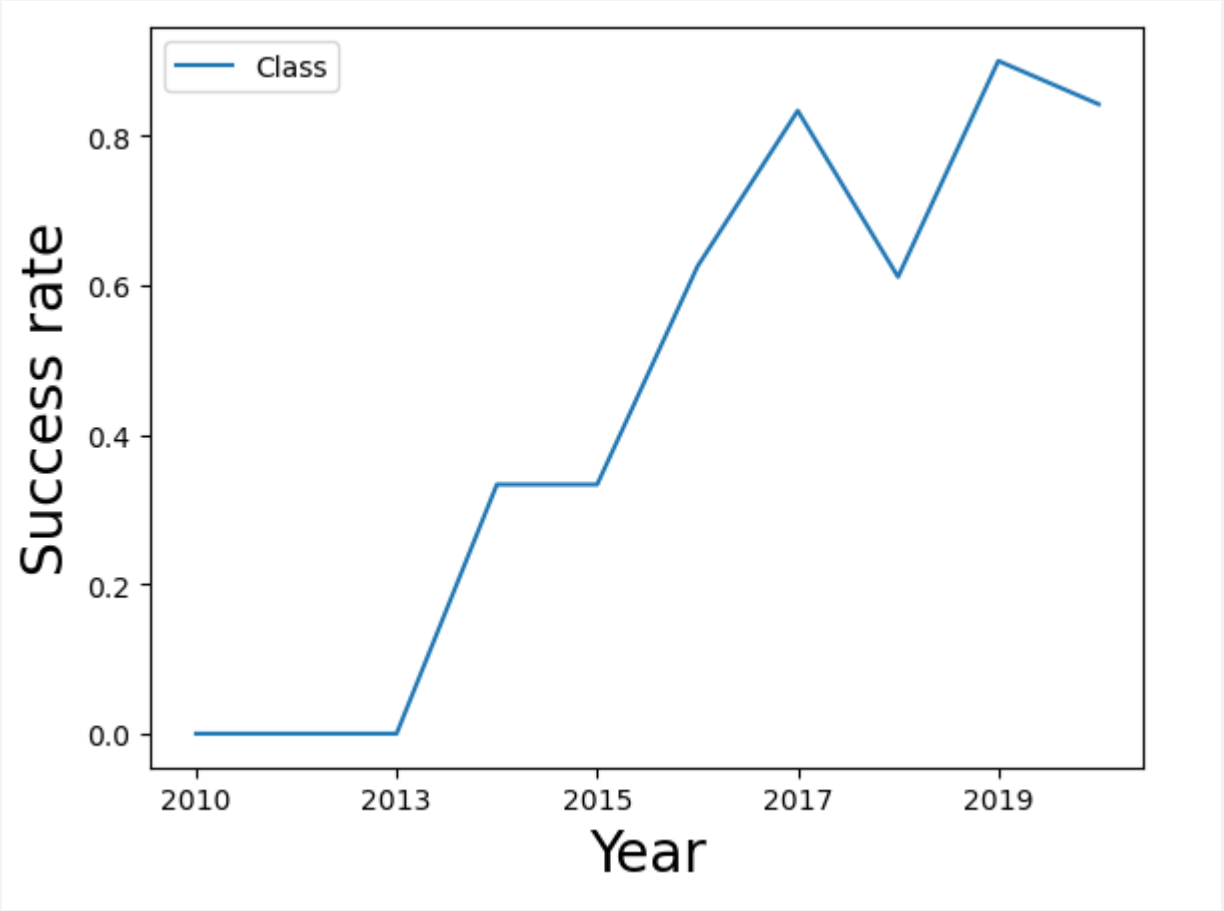
Payload vs. Orbit Type

- It is unclear if there is a relationship between payload mass and orbit type GTO and SSO.
- Orbit Polar, LEO and ISS have higher chances of successful landing with heavier pay loads.



Launch Success Yearly Trend

- The launch success rate improved as time elapsed



All Launch Site Names

- The Unique Launch sites are shown below.
- This was determined using SQL query and the distinct function.

```
%sql Select Distinct("Launch_Site") From SPACEXTABLE
```

```
 * 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'

- 5 records where launch sites begin with `CCA` are also shown.
- This was accomplished using a query with the ‘like’ function.

```
%sql Select * From SPACEXTABLE 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

- Total payload carried by boosters from NASA was 45 596 kg
- The ‘where’ function was used to filter NASA records and the added with the sum function.

```
%%sql
Select sum(PAYLOAD_MASS__KG_) as  "Total Payload Mass in kg", Customer
From SPACEXTABLE
where Customer = 'NASA (CRS)'
limit 5
```

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

Total Payload Mass in kg	Customer
45596	NASA (CRS)

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1 was 2 928.4 kg
- The ‘avg’ was used to calculate the average value and ‘where’ function to filter the booster version

```
%%sql
Select avg(PAYLOAD_MASS__KG_) as  "AVG Payload Mass in kg", "booster_version"
From SPACEXTABLE
where "booster_version" = 'F9 v1.1'
```

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

AVG Payload Mass in kg	Booster_Version
2928.4	F9 v1.1

First Successful Ground Landing Date

- Date of the first successful landing outcome on ground pad was 2015-12-22
- The earliest data was found by using the ‘min’ function on the Data column where the outcome was successful on the ground pad.

```
%%sql
Select min(Date) as "Earliest Date", "Landing_Outcome"
From SPACEXTABLE
where "Landing_Outcome" = 'Success (ground pad)'
```

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

Earliest Date	Landing_Outcome
2015-12-22	Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 kg are shown below.
- The ‘where’ along with the ‘and’ functions were used to filter two conditions.

```
%%sql
Select Booster_Version, Landing_Outcome, PAYLOAD_MASS__KG_
From SPACEXTABLE
where (PAYLOAD_MASS__KG_ between 4000 and 6000) and (Landing_Outcome = "Success (drone ship)")
```

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

Booster_Version	Landing_Outcome	PAYLOAD_MASS__KG_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

Total Number of Successful and Failure Mission Outcomes

- The total number of successful mission outcomes were 100 whilst failure mission outcomes 1. This is a success rate of almost 100%.
- The ‘count’ function was used to count the total records and the ‘group by’ function to group the according to outcome records.

```
%%sql
Select count(*), Mission_Outcome
From SPACEXTABLE
Group by Mission_Outcome
```

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

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

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass of 15600 kg are listed on the right.
- A nested query to determine the maximum payload and then used to filter the records with the ‘where’ function.

```
 %%sql
Select Distinct(Booster_Version), PAYLOAD_MASS__KG_
From SPACEXTABLE
Where PAYLOAD_MASS__KG_ =(Select Max(PAYLOAD_MASS__KG_) From SPACEXTABLE)
```

```
 * 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

- The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015 are shown on the right.
- The ‘substr’ on the Date column was used to select the dates whilst the ‘where’ function was to filter the other conditions.

```
 %%sql
Select  substr(Date,0,5) as Year,
substr(Date, 6,2) as Month,
Booster_Version, Launch_Site,
Landing_Outcome

From SPACEXTABLE

Where Date like '2015%' and (Landing_Outcome ="Failure (drone ship)")

Limit 10
```

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

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

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

- The Rank of the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order.

- A combination of the ‘group by’, and ‘order by’ function were used to filter the results.

```
 %%sql
Select count(*) AS 'Count',
Landing_Outcome

From SPACEXTABLE

Where Date between "2010-06-04" and "2017-03-20"

Group by Landing_Outcome

Order by Count DESC
```

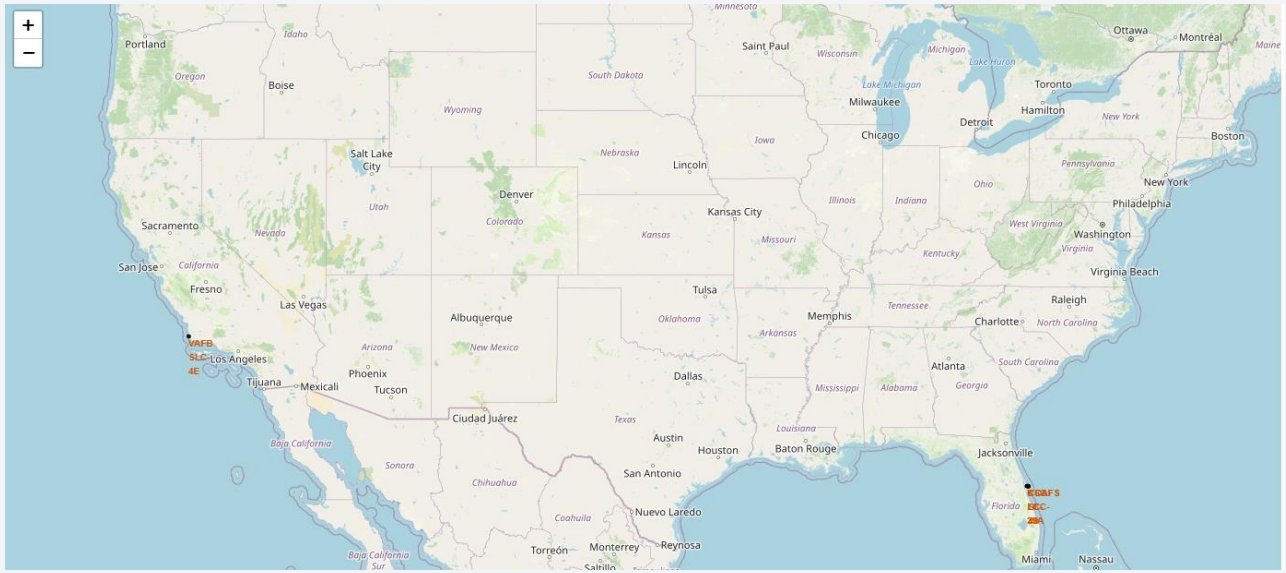
```
 * sqlite:///my_data1.db
Done.
```

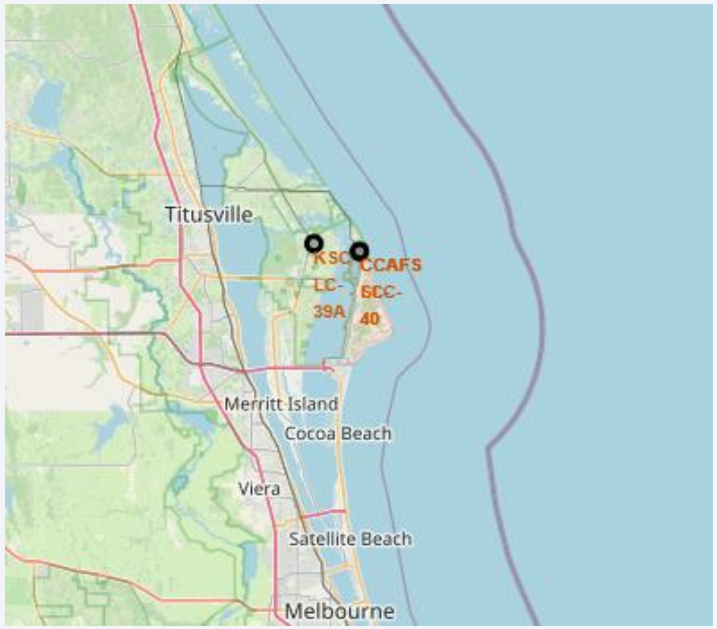
Count	Landing_Outcome
10	No attempt
5	Success (drone ship)
5	Failure (drone ship)
3	Success (ground pad)
3	Controlled (ocean)
2	Uncontrolled (ocean)
2	Failure (parachute)
1	Precluded (drone ship)

Section 3

Launch Sites Proximities Analysis

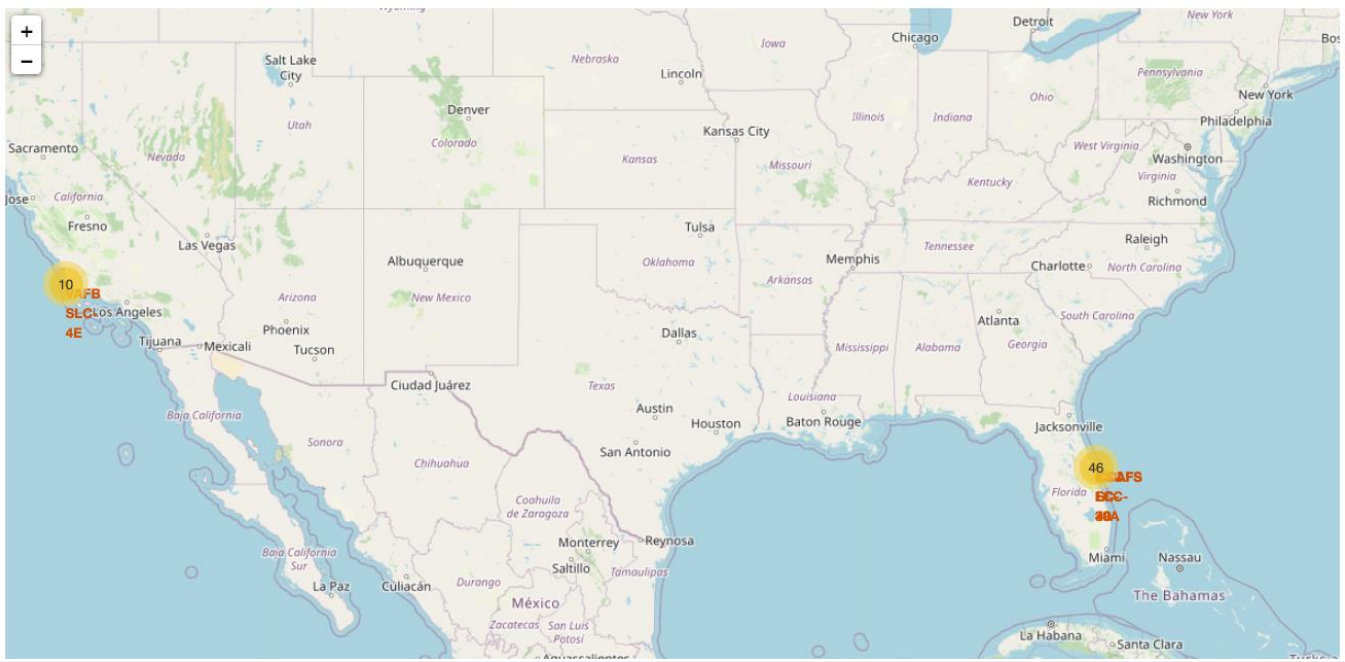
Launch site on globe map

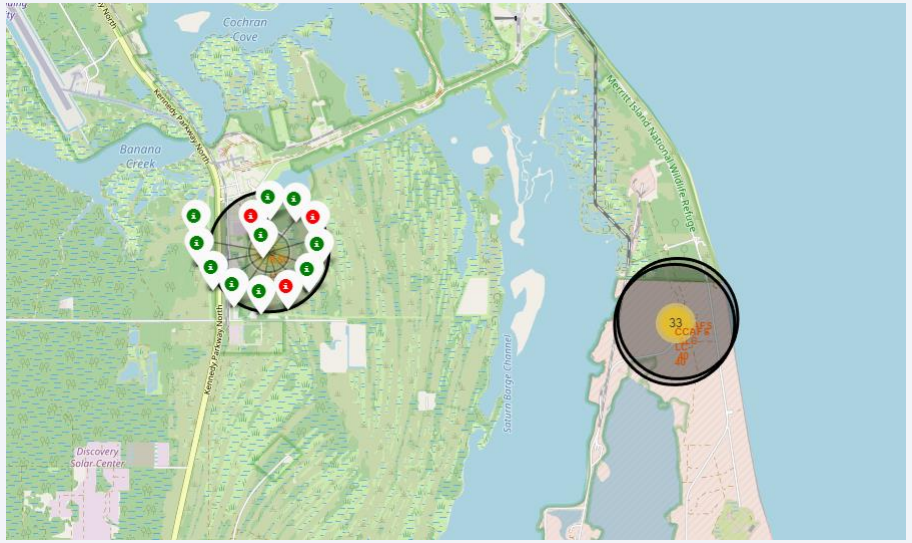




The locations of the launch sites are shown on the maps above. On the right is a blown up image of the 3 launch sites on the ease which are not easily visible on the main map.

Successful launches map

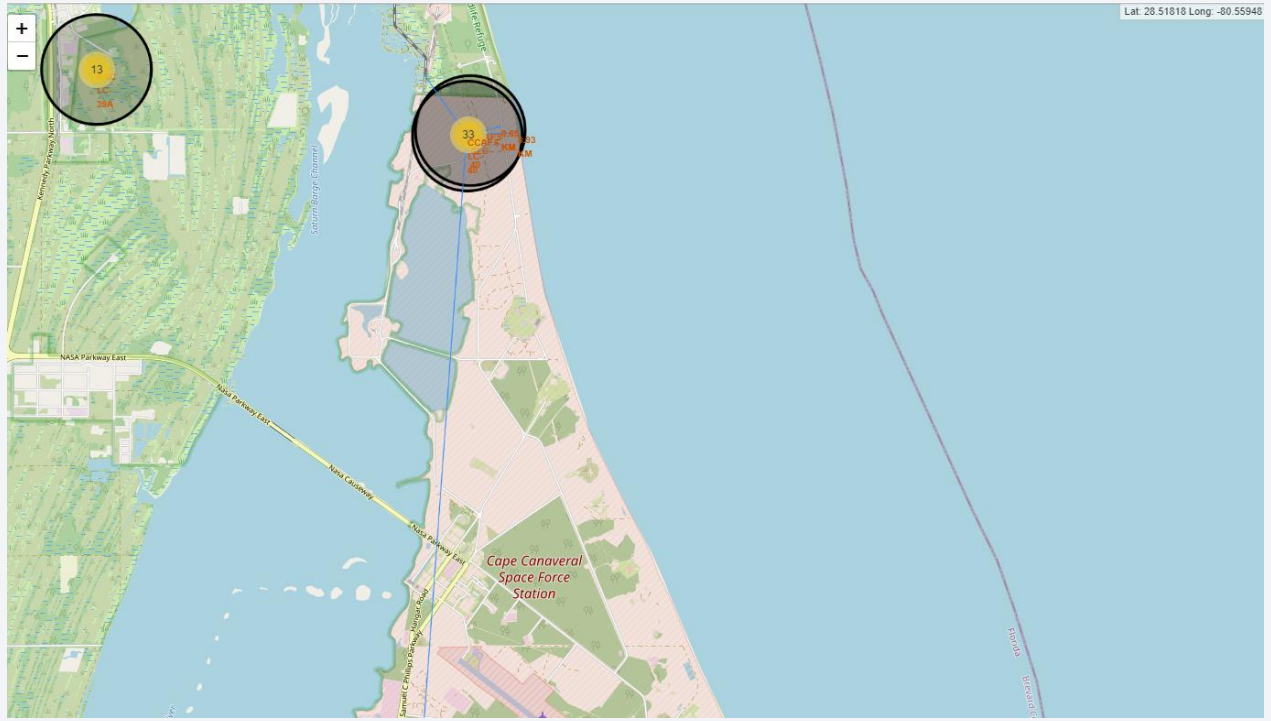


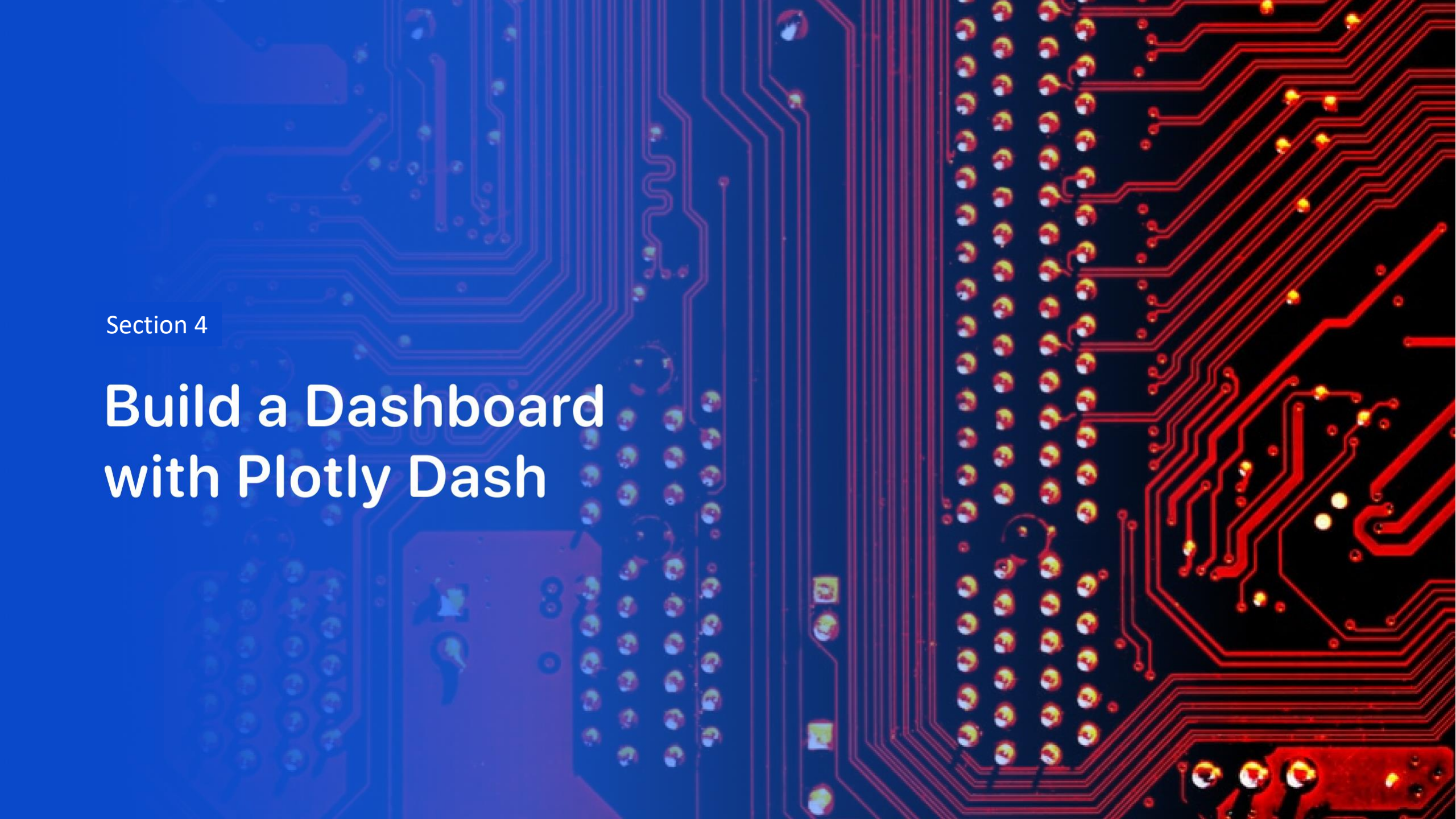


The site with the most successful launch site is KSC LC-39A which is also on the east coast.

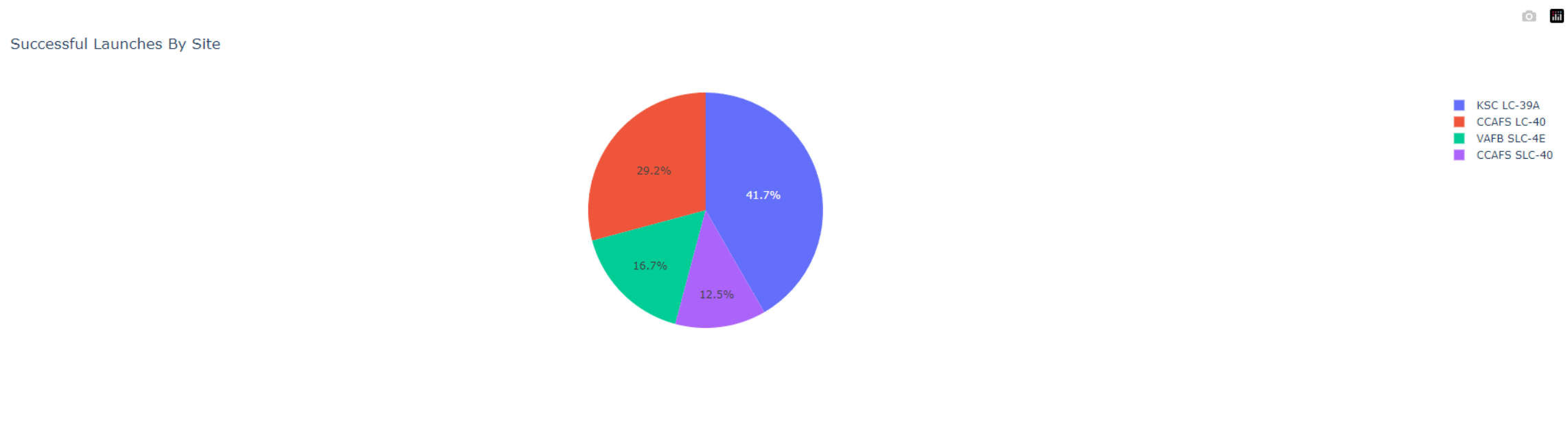
Proximity to key points

- Distance to
  - Railway: 1.3 km
  - Highway: 0.65 km
  - Coastline: 0.93 km
  - City: 56.6 km
- The launch is closest to transport routes and farthest from the city



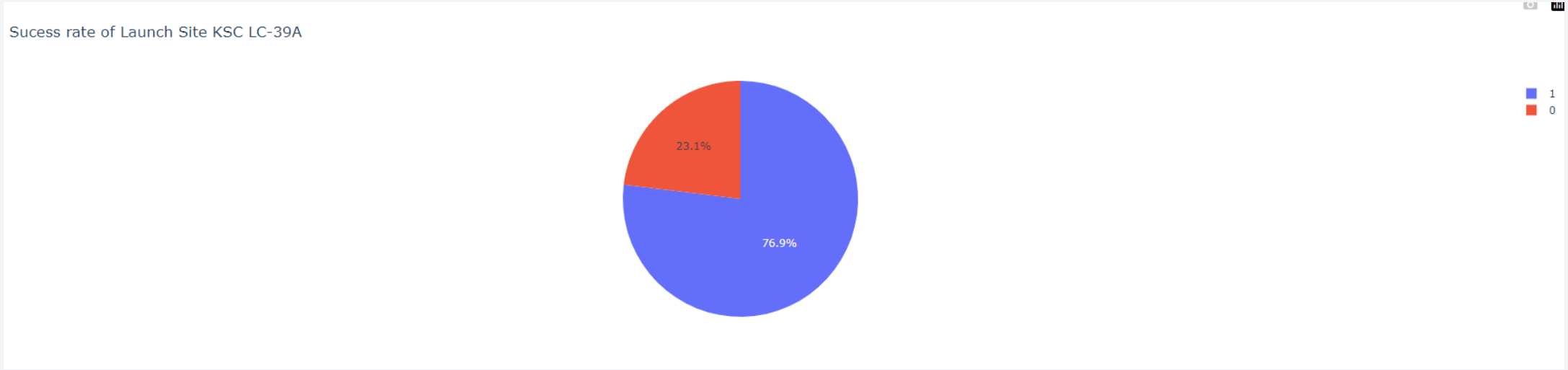

Section 4
Build a Dashboard with Plotly Dash

Successful Launches by site



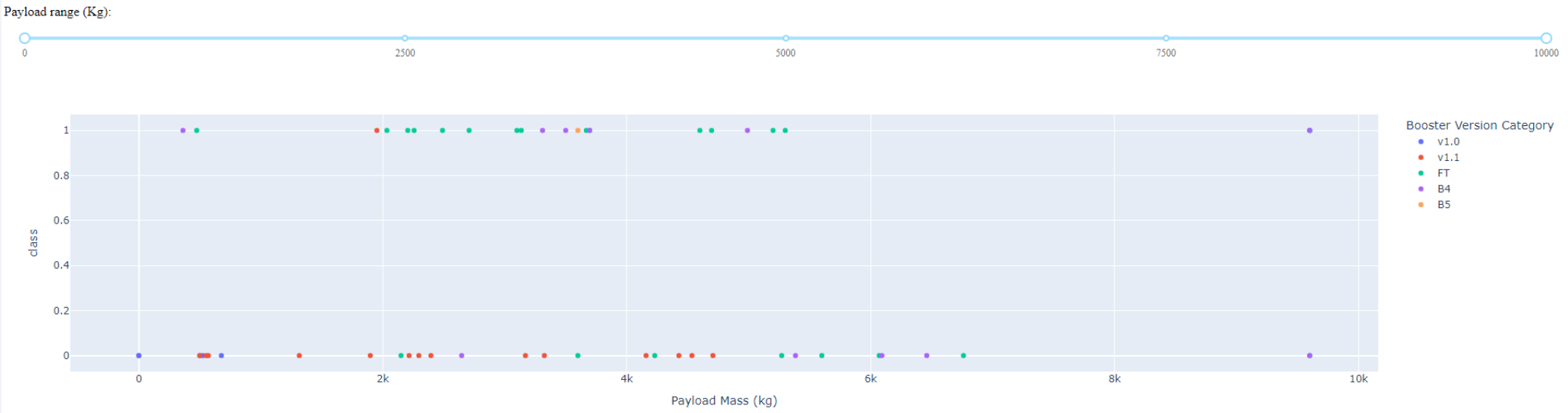
The pie chart shows that by count which site had the most successful launches. For instance it can be observed that the KSC LC-39A site had the highest count.

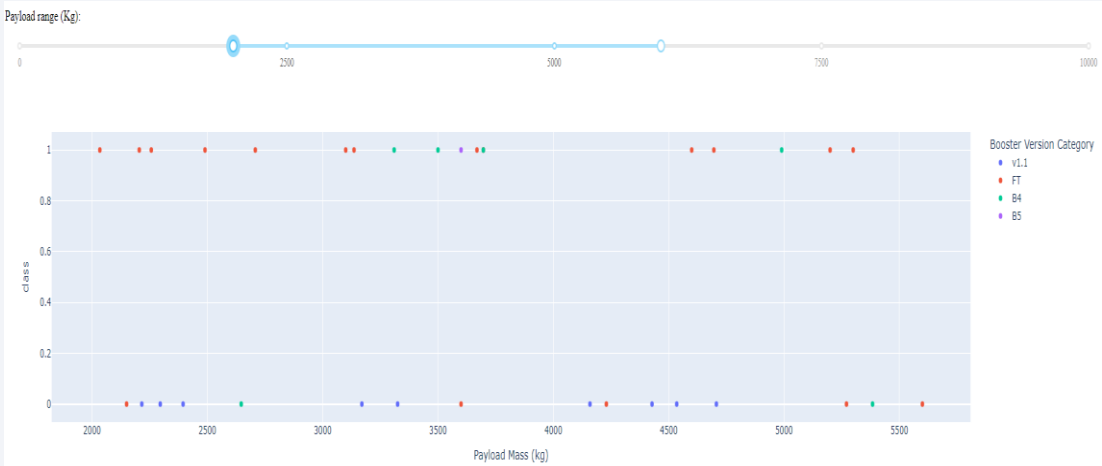
Success rate of launch site KSC LC-39A



The site with a the largest count KSC LC-39A had a successful launch rate of 76,9%.

Payload vs Launch Outcome scatter





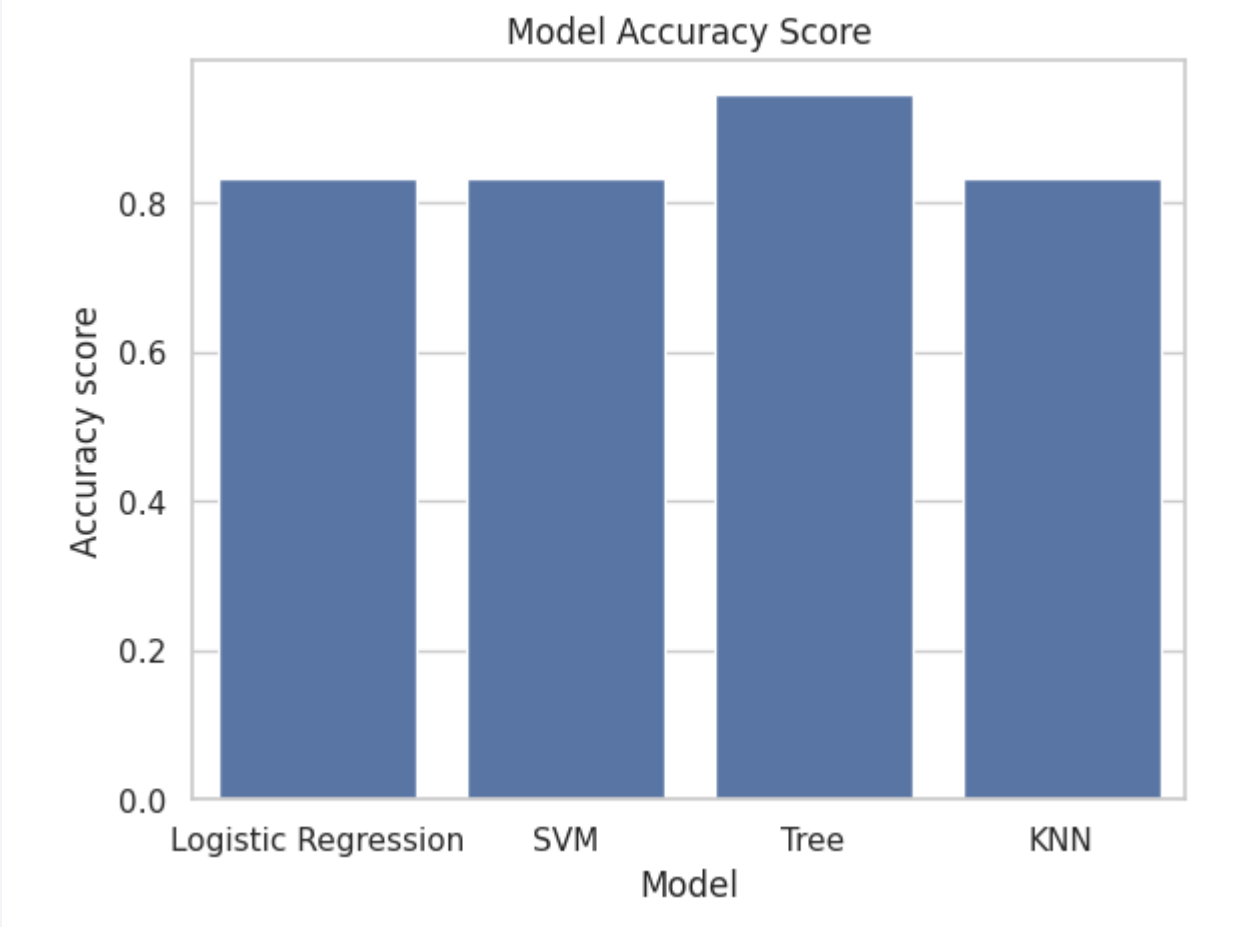
- The 2000 to 7000 kg had the highest success rate with across all the booster versions.
- More specifically the FT Booster version had the highest success.

Section 5

Predictive Analysis (Classification)

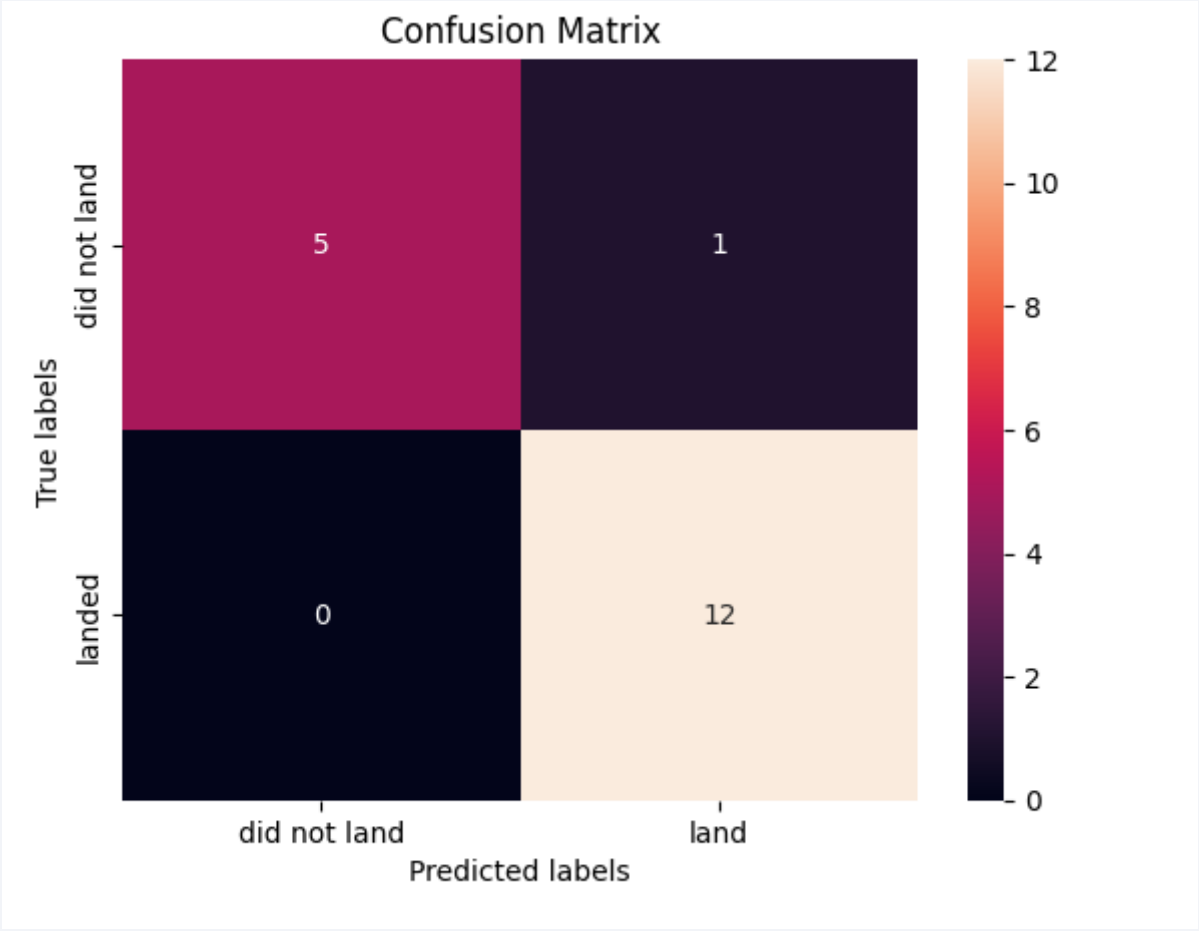
Classification Accuracy

- The model with the highest accuracy was the Decision Tree model.



Confusion Matrix

- The confusion matrix for the decision model is shown on the right.



Conclusions

- Success rate increased with increasing number of flights and with the passage of time.
- Pay load had a significant impact on success rate and was optimal between 2000 kg and 6000 kg.
- The best location is in the east coast.
- The decision tree predictive model had the best accuracy
- The model can predict the success rate with 93 accuracy.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!