Winning Space Race with Data Science

<Wei Shien Chang>
<27/01/2024>

Outline

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

Executive Summary

- Machine Learning is used to predict the success rate of the landing of the first stage of Falcon 9 by SpaceX.
- Data is collected from historical data from Falcon 9 launches via REST API calls and Webscrapping from Wikipedia. The collected data is cleaned and normalised for processing.
- Data Visualisation tools including scatter plots, line plots, bar charts are used to visualise relationship between parameters including payload mass, orbit, launchsite, landing pad and outcome. For orbit PO, ISS and LEO, higher payload suggests a higher chance of successful landing. Over the years, the success rate shows an upward trends.
- Machine Learning algorithms including logistic regression, support vector machine, decision tree and k-nearest neighbours are used to predict the success rate of landing of 1st stage of Falcon 9. Decision tree performed the best, with an accuracy of 88.88%.

Introduction

- SpaceX's Falcon 9 can recover the first stage, making it the relatively inexpensive in launching rockets in the space industry.
- There is a total 4 launch sites and 11 orbits, with different models of rocket carrying different payload.
- In this project, the success rate of whether the first stage will land after launching is predicted. This is to determine whether the first stage can be re-used, which is then being taken into consideration to determine the price of each launch.

Section 1

Methodology

Methodology

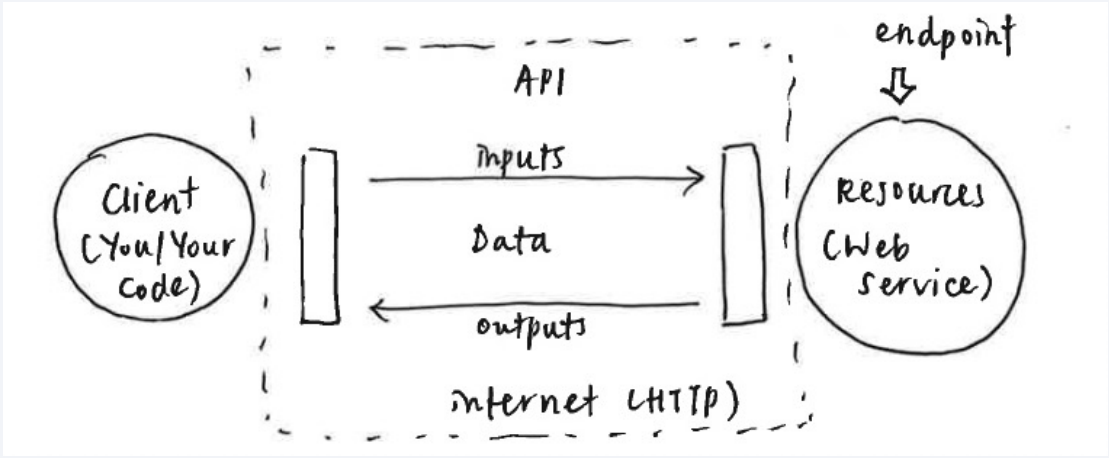
- Data collection methodology:
  - Historical data of Falcon 9 launches is collected via REST API calls and Webscrapping of Wikipedia
- Perform data wrangling
  - First a label for landing outcome is created, converting to numerical data
  - Dummy variables are created for other variables including orbit, launch sites, landing pad and serial number using 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
  - Logistic Regression, Support Vector Machine, Decision Tree, K-Nearest Neighbours models are used.

Data Collection

- Historical data is collected via REST API calls and Webscrapping from Wikipedia.

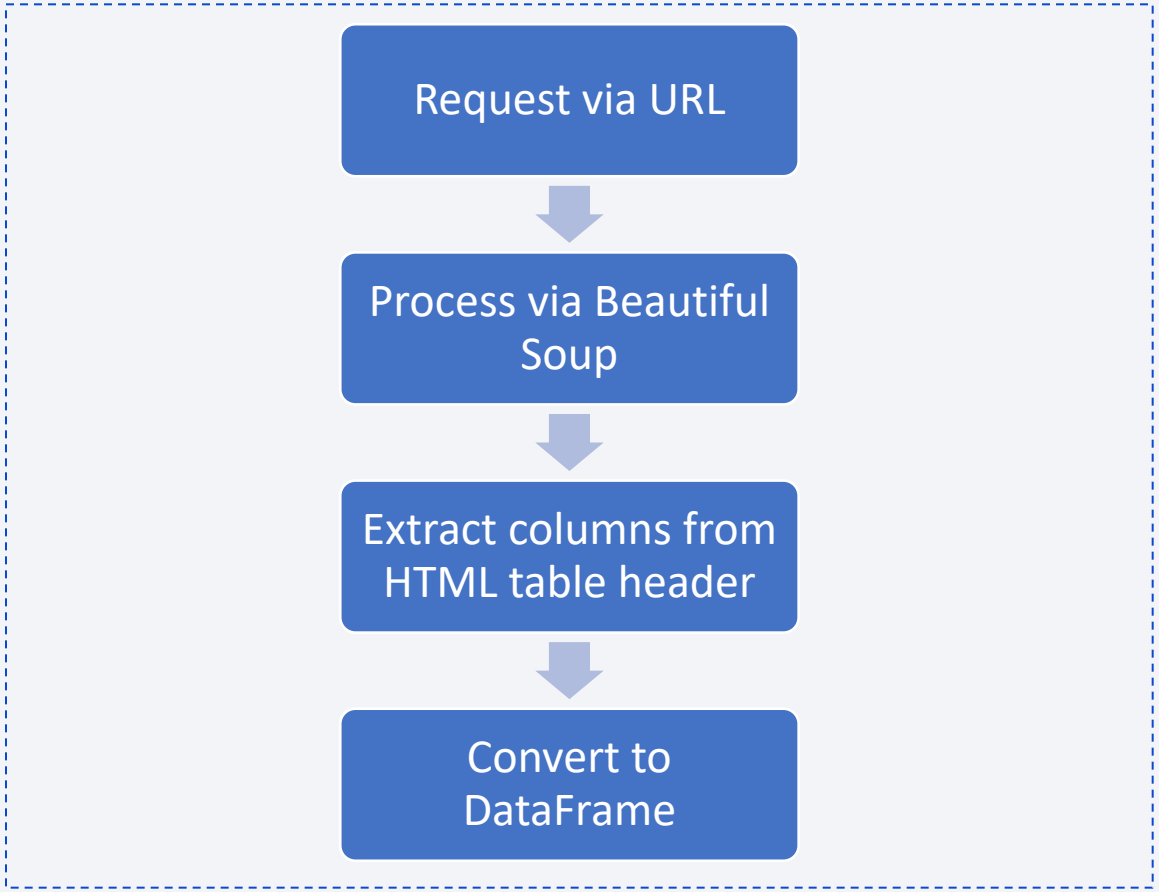
Data Collection – SpaceX API

- Sends a request via an HTTP message containing JSON file
- Transmitted to the webservice (spacexdata) via internet
- Webservice performs the operation
- Webservice returns a response via an HTTLP message where information is returned via a JSON file.
- GitHub URL



Data Collection - Scraping

- Request Falcon9 Launch historical data from Wikipedia URL using Beautiful Soup
- Extract all columns/variable names from the HTML table header
- Create a DataFrame by parsing the extracted HTML table
- GitHub URL



Data Wrangling

- Categorical variables are converted into numerical
  - Landing outcome is assigned labels: 1 for successful landings, 0 for failed landings
  - Other features including orbit, launch sites, landing pad and serial number of rockets are converted to numerical variables by creating dummy variables via one hot encoding.

- GitHub URL

EDA with Data Visualization

- Scatter plot between Payload Mass and Flight Number to determine their relationship.
- Scatter plot between Launch Site and Flight Number to determine their relationship.
- Scatter plot between Launch Site and Payload to determine their relationship.
- Bar plot of Average success rate and orbit. This is suitable as orbit is categorical variable.
- Scatter plot between orbit type and flight number to determine their relationship.
- Scatter plot between orbit and payload mass to determine their relationship.
- Line chart for the landing success rate over the years to see the trend
- GitHub URL

EDA with SQL

- 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
- List the records which will display the month names, failure landing outcomes in drone ship, booster version and launch sites for the months in year 2015
- Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20 in descending order
- GitHub URL

Build an Interactive Map with Folium

- Marked all launch sites on a map using a blue circle marker and red label, based on their coordinates.
- Created a marker cluster to record the numbers of successful and failed landing at each launch sites.
- Calculated the distance between a launch site and its proximities to find any correlation to successful landing.
- GitHub URL

Build a Dashboard with Plotly Dash

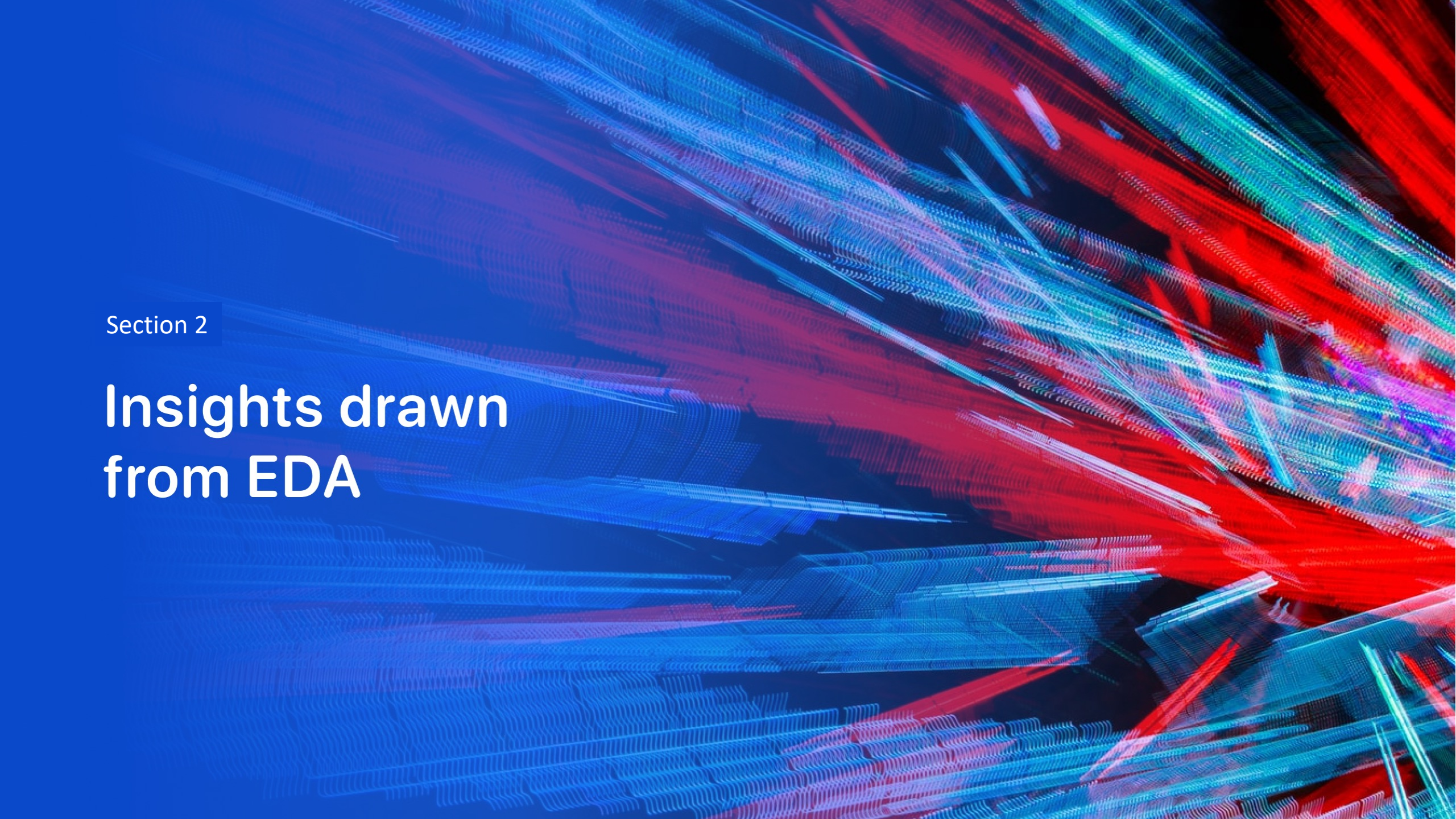
- An interactive pie chart is created to visualise the successful landing based on launch site.
- User can view the proportion of success launches for all sites or specific site.
- An interactive slider is created for user to select minimum and maximum payload mass.
- This information is used to plot the scatter plot between success rate and payload mass to examine the correlation for all sites or specific site.
- GitHub URL

Predictive Analysis (Classification)

- A separate variable (NumPy array) is created for the target, in this case the landing outcome.
- The data is transformed and normalised using Standard Scaler.
- The data set is split into training and testing set with training size of 20%.
- 4 models are tested to find the most suitable model. They are logistic regression, support vector machine, decision tree and k-nearest neighbours.
- Grid Search is applied to each model to find the best parameter to be used for prediction.
- R2 is calculated to determine the accuracy of the prediction. Confusion matrix is created to determine the accuracy of the model as well.
- GitHub URL

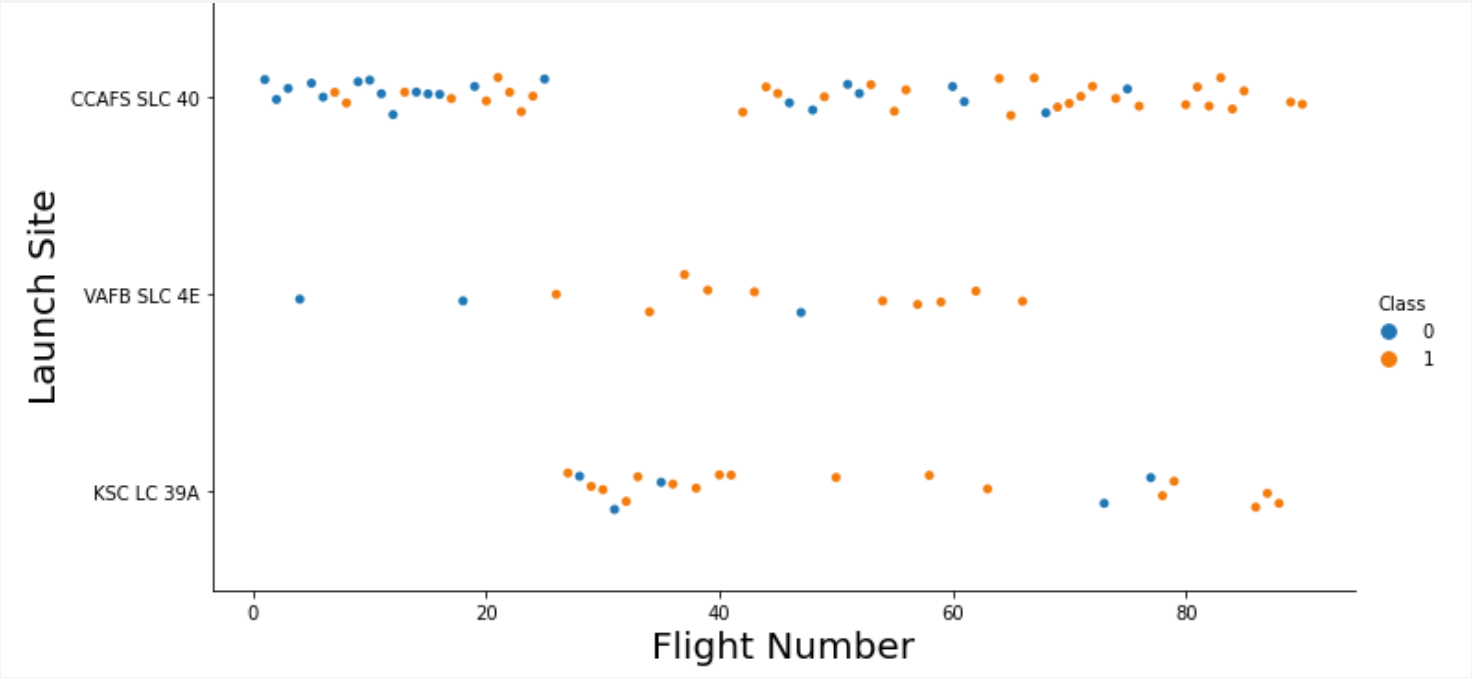
Results

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


Section 2
Insights drawn from EDA

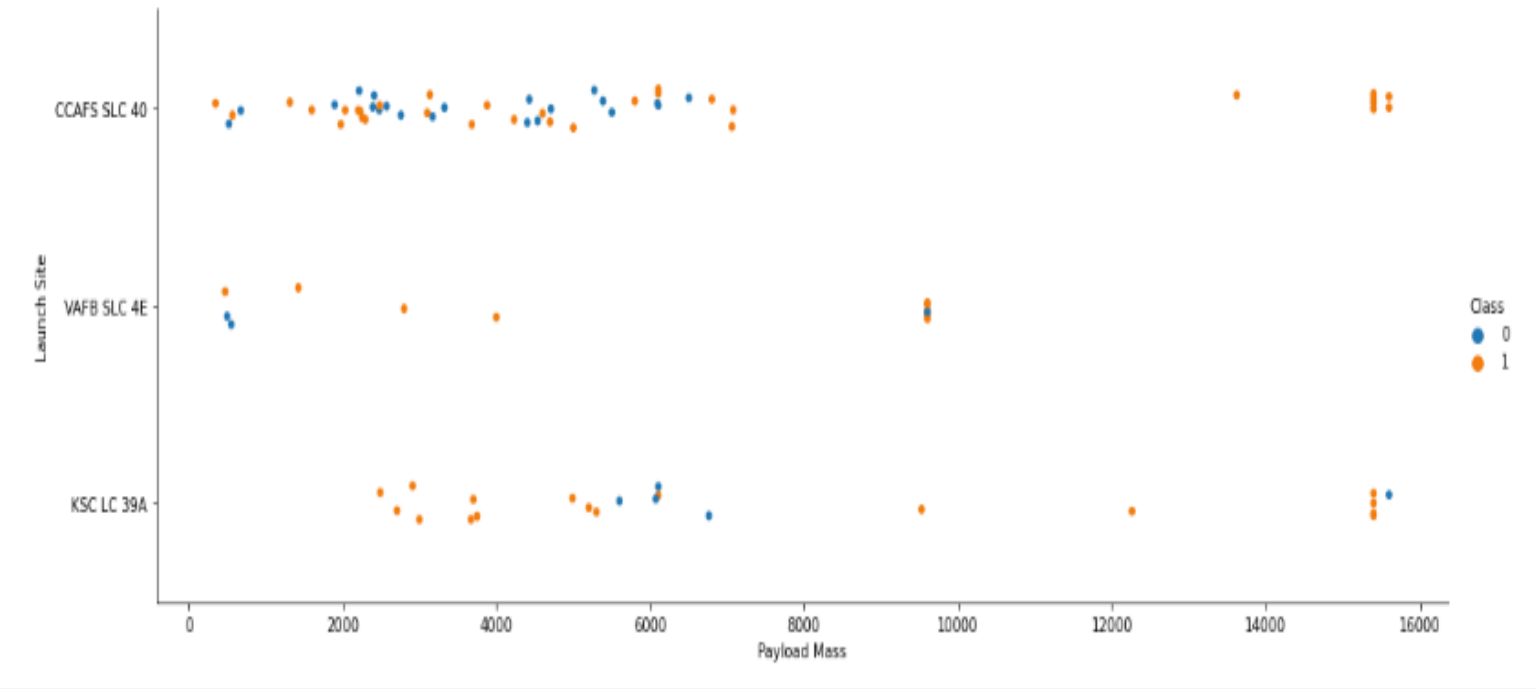
Flight Number vs. Launch Site

- Scatter plot between Launch Site and Flight Number to determine their relationship on successful landing.
- Result shows that recent landing at all launch sites are mostly successful.



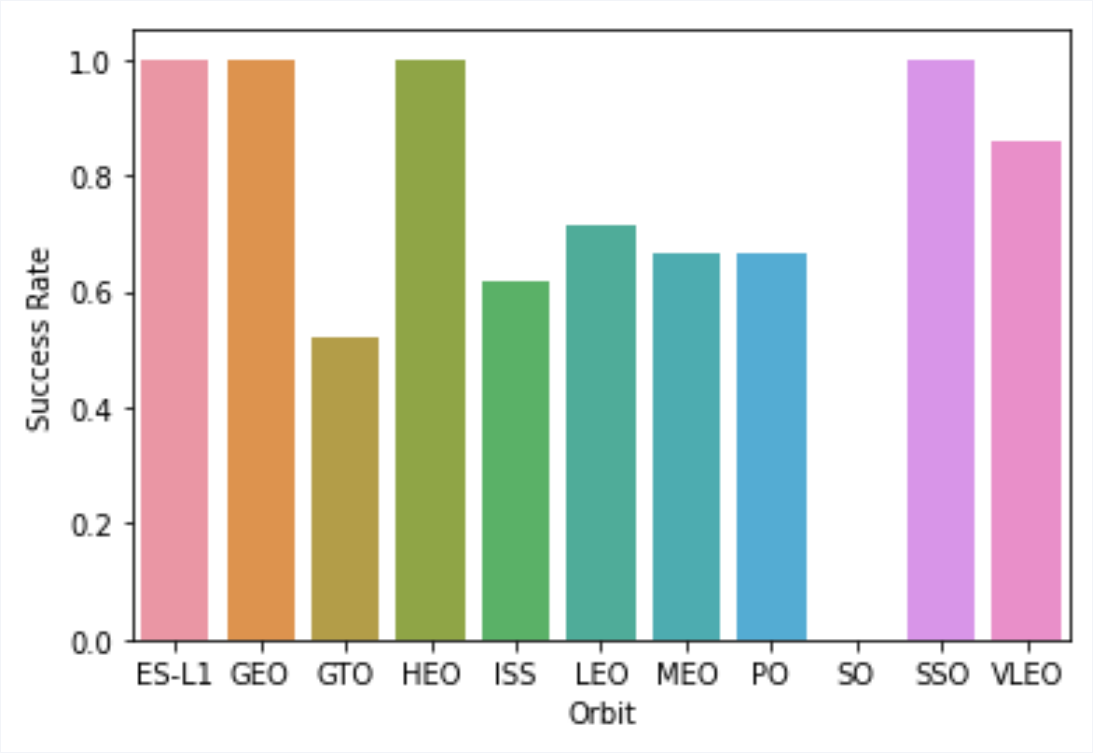
Payload vs. Launch Site

- Scatter plot between Launch Site and Payload to determine their relationship.
- Result shows that higher payload (above 7000kg) has higher success rate launching at CCAFS SLC-40, whereas payload ranging between 1000kg and 5000kg has higher success rate launching at VAFB SLC-4E and KSC LC 39A.



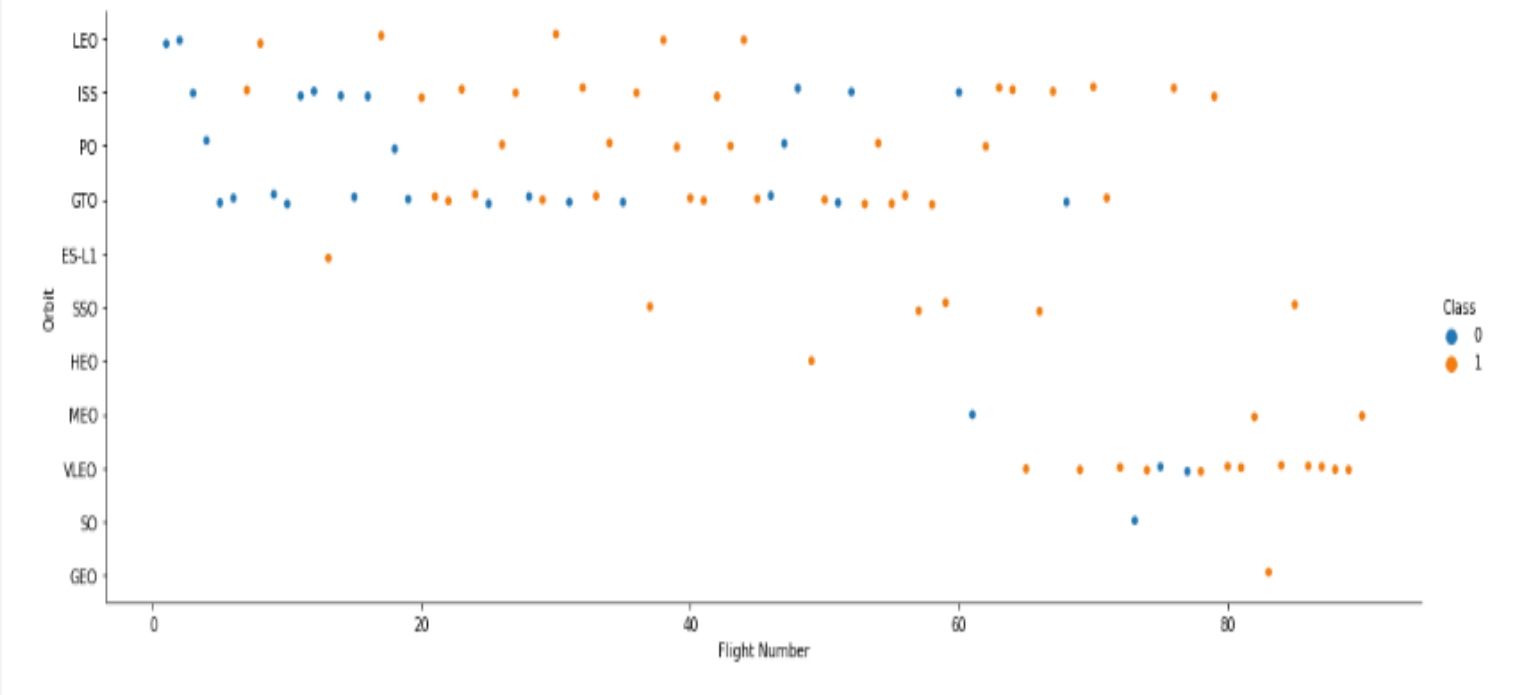
Success Rate vs. Orbit Type

- Bar plot of Average success rate and orbit. This is suitable because orbit as the independent variable is categorical.
- ES-I1, GEO, HEO and SSO show 100% success rate.



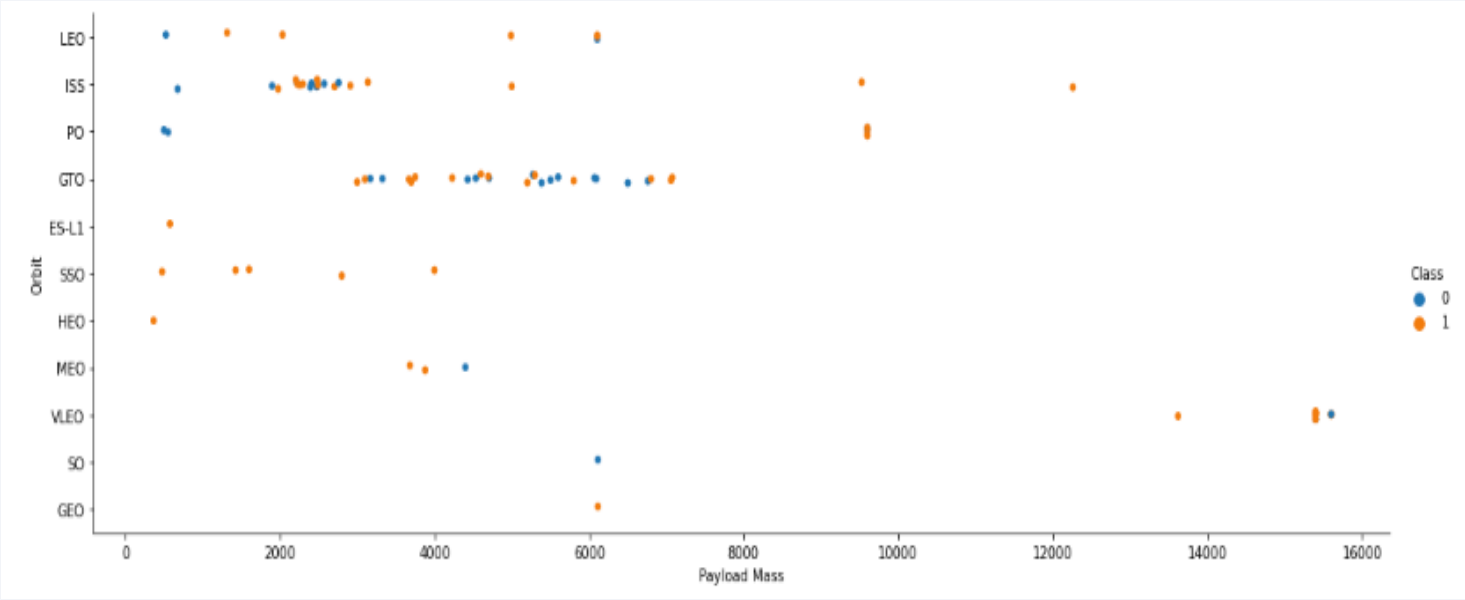
Flight Number vs. Orbit Type

- Scatter plot between orbit type and flight number to determine their relationship.
- Results suggest a relationship between LEO orbit and flight number, with more recent flight records successful landing.
- However, no apparent relationship for other orbits with flight number.



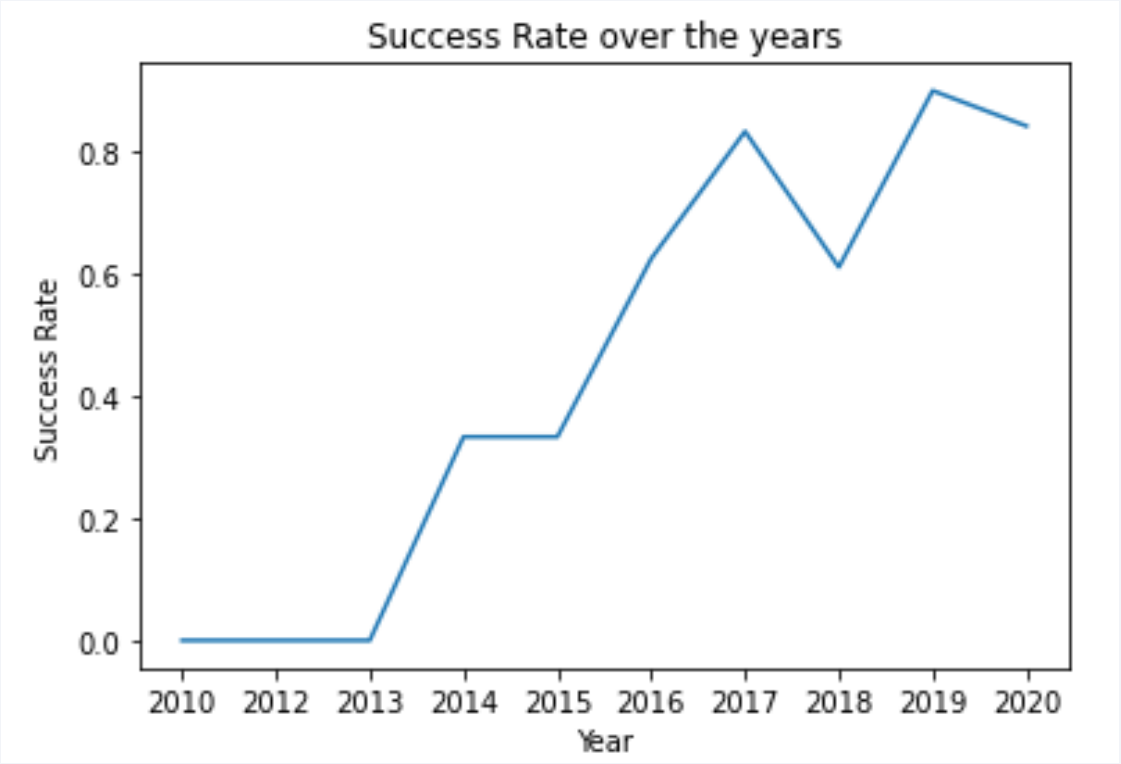
Payload vs. Orbit Type

- Scatter plot between orbit and payload mass to determine their relationship.
- With heavy payloads, the successful landing or positive landing rate are more for Polar, LEO and ISIS.



Launch Success Yearly Trend

- Line chart for the landing success rate over the years to see the trend
- From 2013, the success rate records a steady increase until 2017, with a dip to 2018, before rising until 2019 and records a slight dip in 2020.



All Launch Site Names

- Launch Sites are CCAFS LC-40, VAFB SLC-4E, KSC LC-39A and CCAFS SLC-40

In [8]:
```
%%sql

SELECT DISTINCT Launch_Site
FROM SPACEXTABLE
;
```

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

Out[8]:

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

Launch Site Names Begin with 'CCA'

In [11]:

```
%%sql

SELECT *
FROM SPACEXTABLE
WHERE Launch_Site LIKE 'CCA%'
LIMIT 5
;
```

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

Out[11]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outco
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Succ
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	Succ
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Succ
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Succ
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Succ

Total Payload Mass

- Total payload mass is 45596 kg.

In [18]:
```
%%sql

SELECT SUM(PAYLOAD_MASS__KG_)
AS Total_Payload_Mass_KG
FROM SPACEXTABLE
WHERE Customer = 'NASA (CRS)'
;
```

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

Out[18]:

Total_Payload_Mass_KG
45596

Average Payload Mass by F9 v1.1

- Average payload mass is 2928.4 kg.

In [20]:
```
%%sql

SELECT AVG(PAYLOAD_MASS__KG_)
AS Average_Payload_Mass_KG
FROM SPACEXTABLE
WHERE Booster_Version = 'F9 v1.1'
;
```

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

Out[20]:

Average_Payload_Mass_KG
2928.4

First Successful Ground Landing Date

- First successful ground late is on 2015-12-22.

In [23]:
```
%%sql

SELECT MIN(Date)
AS Date_First_Successful_Landing_in_Ground_Pad
FROM SPACEXTABLE
WHERE Landing_Outcome = 'Success (ground pad)'
;
```

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

Out[23]:

Date_First_Successful_Landing_in_Ground_Pad
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

In [25]:

```
%%sql

SELECT Booster_Version
FROM SPACEXTABLE
WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
AND Landing_Outcome = 'Success (drone ship)'
;
```

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

Out[25]:

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

Total Number of Successful and Failure Mission Outcomes

In [27]:

```
%%sql

SELECT Mission_Outcome, COUNT(Mission_Outcome)
FROM SPACEXTABLE
GROUP BY Mission_Outcome
;
```

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

Out[27]:

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

Boosters Carried Maximum Payload

In [29]:

```
%%sql

SELECT Booster_Version
FROM SPACEXTABLE
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
;
```

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

Out[29]:

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

In [30]:

```
%%sql

SELECT substr(Date,6,2) AS Month_Names, Landing_Outcome, Booster_Version, Launch_Site
FROM SPACEXTABLE
WHERE Landing_Outcome = 'Failure (drone ship)'
AND substr(Date,0,5) = '2015'
;
```

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

Out[30]:

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

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

In [32]:

```
%%sql

SELECT Landing_Outcome, COUNT(Landing_Outcome)
FROM SPACEXTABLE
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.
```

Out[32]:

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

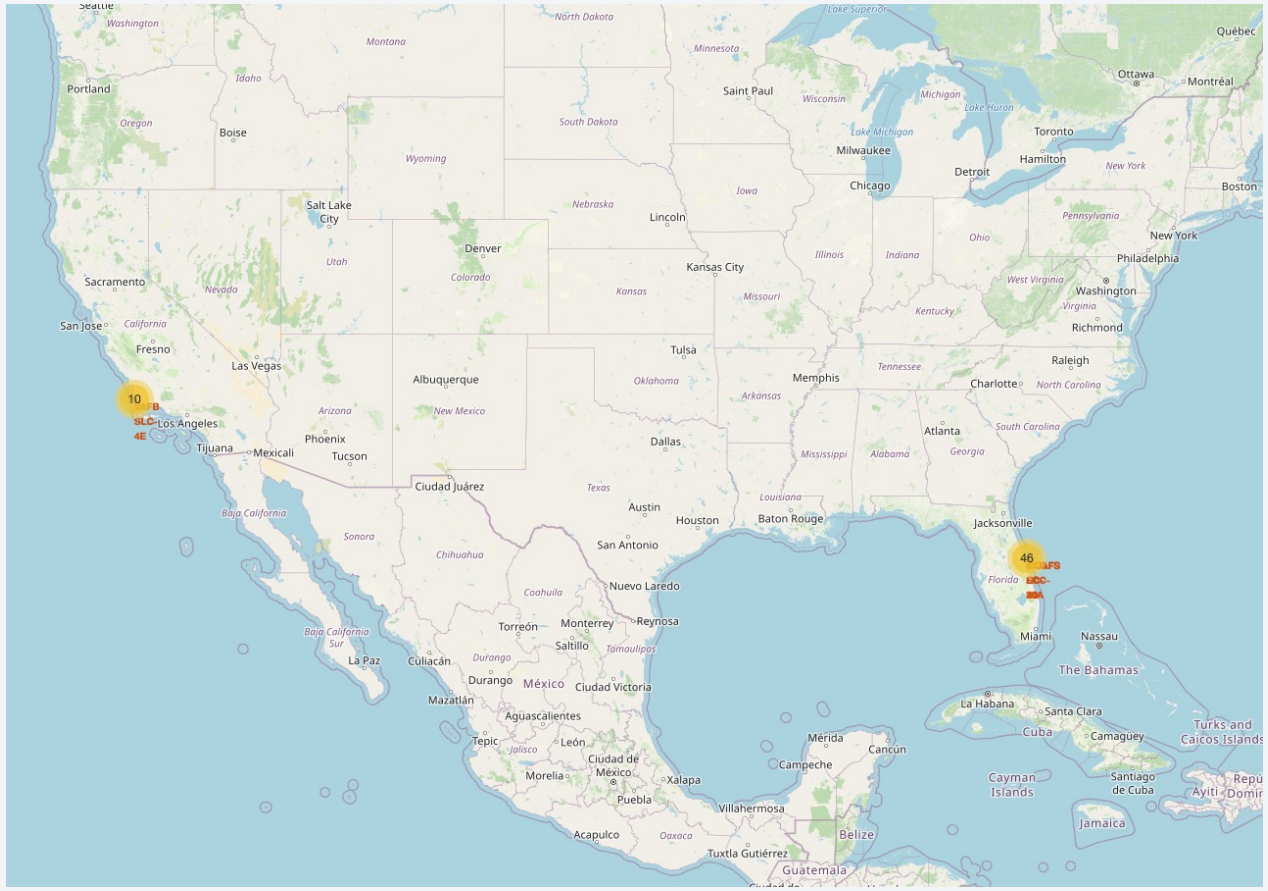

Section 3
Launch Sites
Proximities Analysis

Map of Launch Sites

- All launch sites are located near coastline.

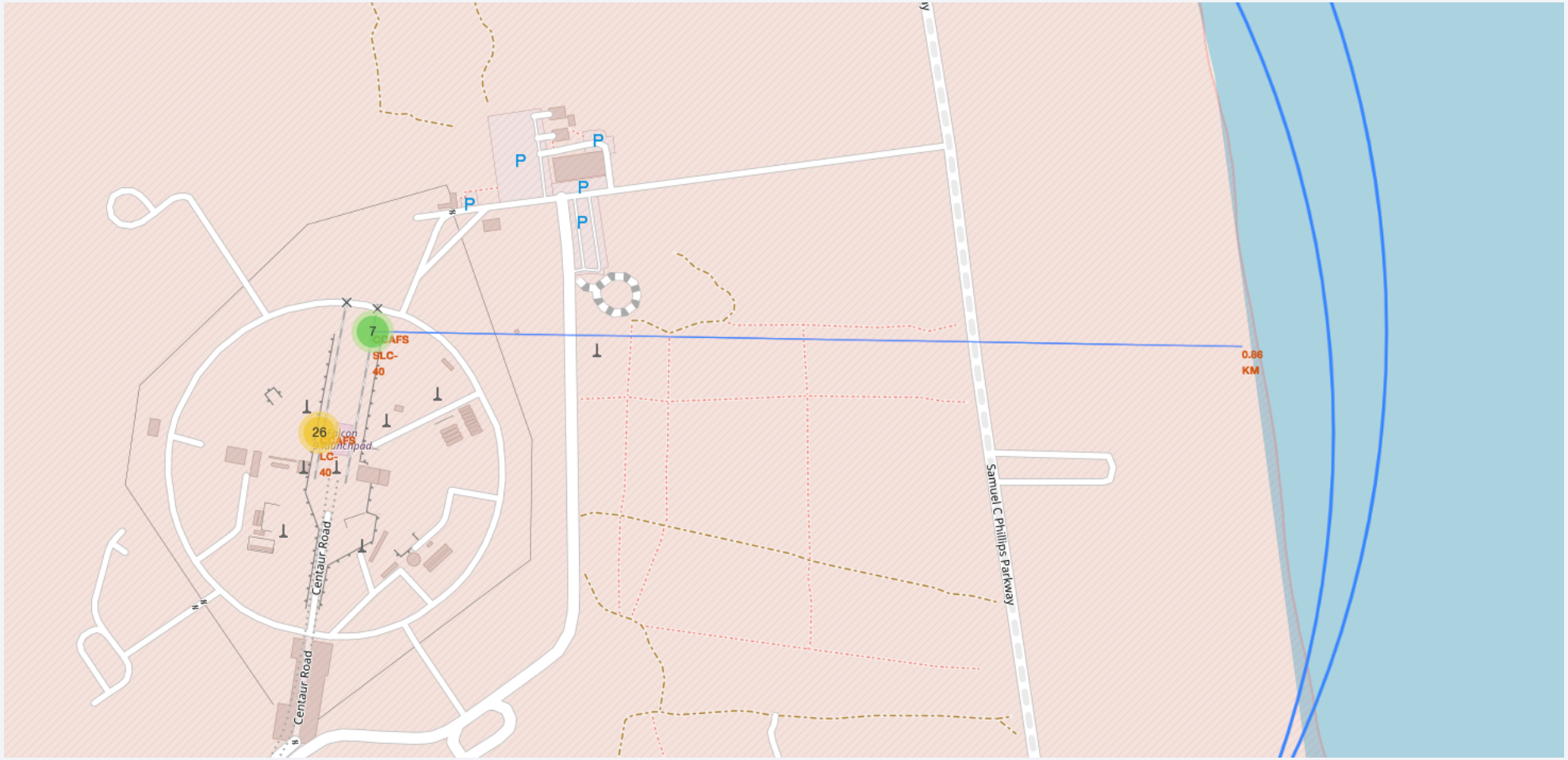
Launches and Success Rate at Launch Sites

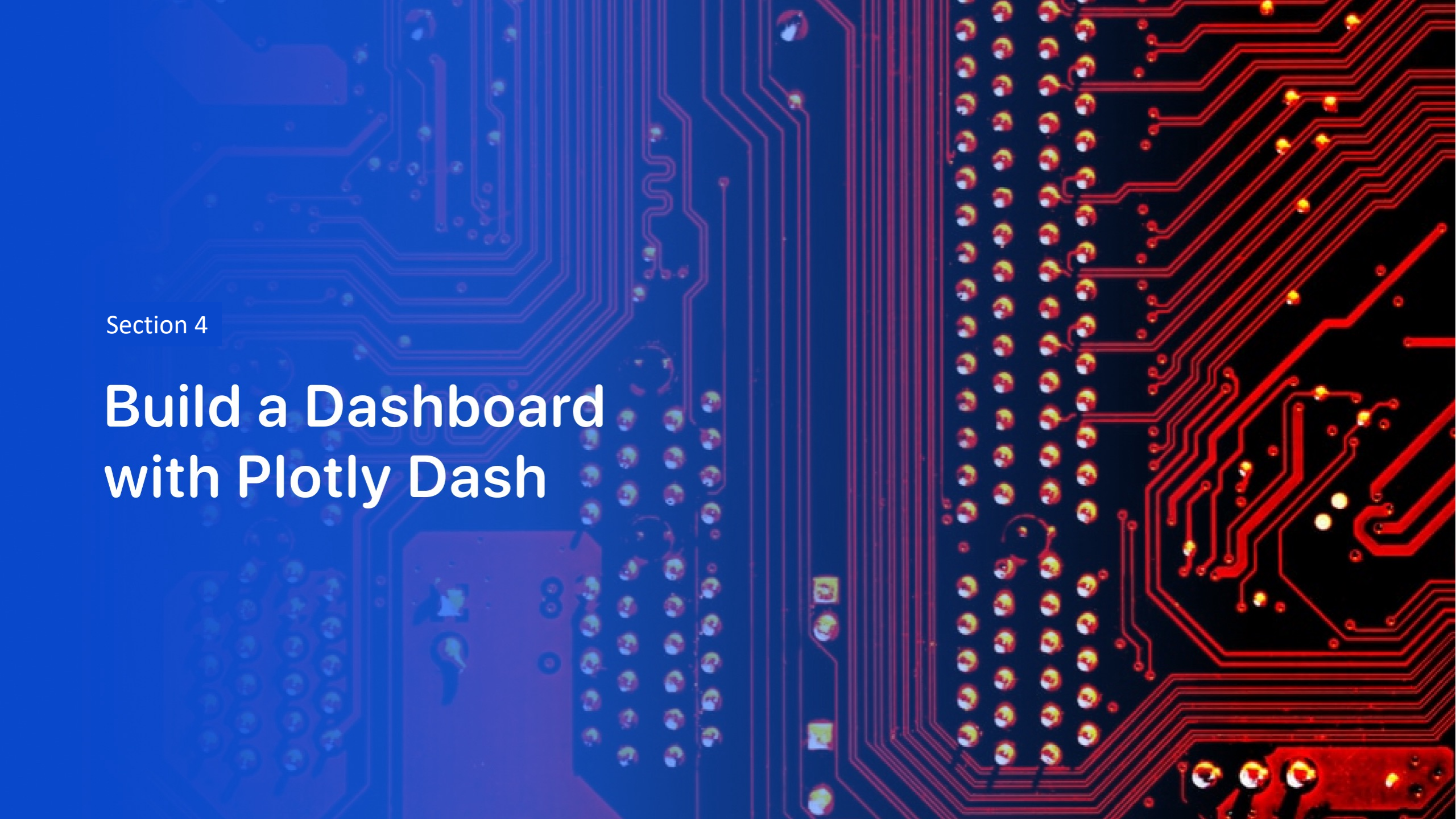
- Highest success rate recorded at KSC LC-39A



Launch Sites Proximities

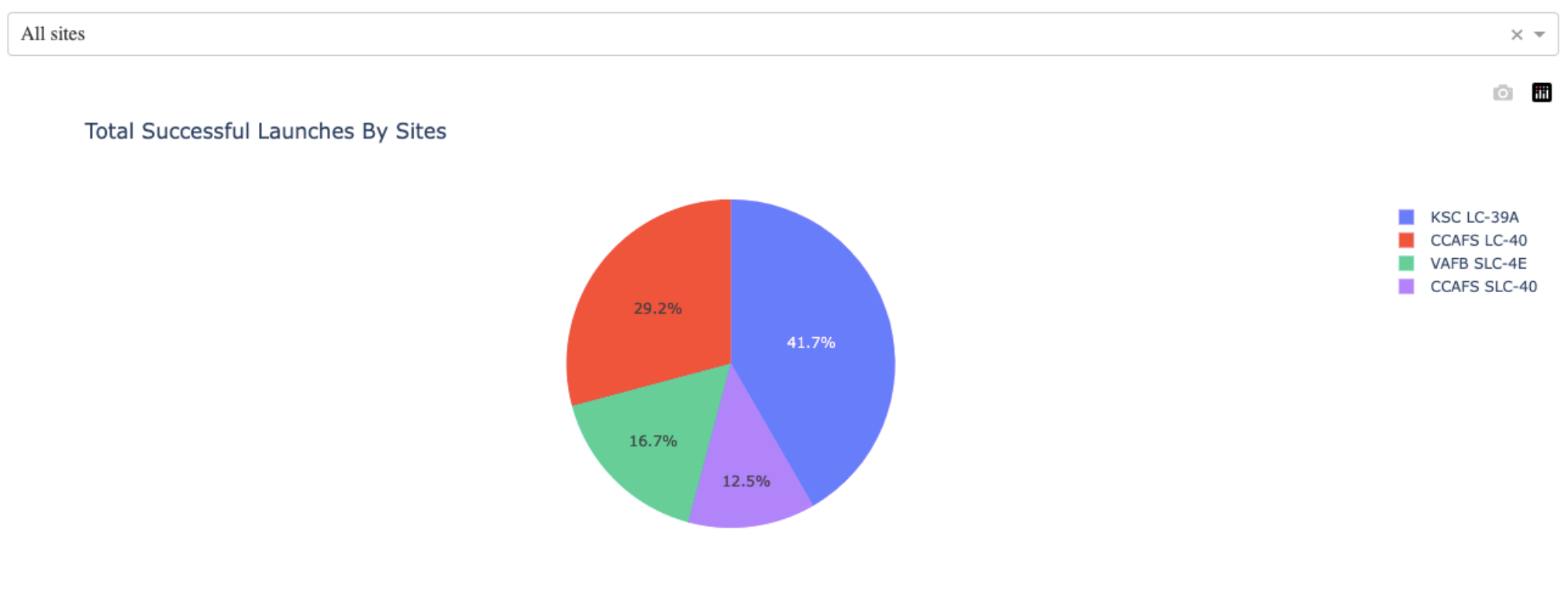
- Launch sites are usually close to coastlines, around 1km away.




Section 4
Build a Dashboard with Plotly Dash

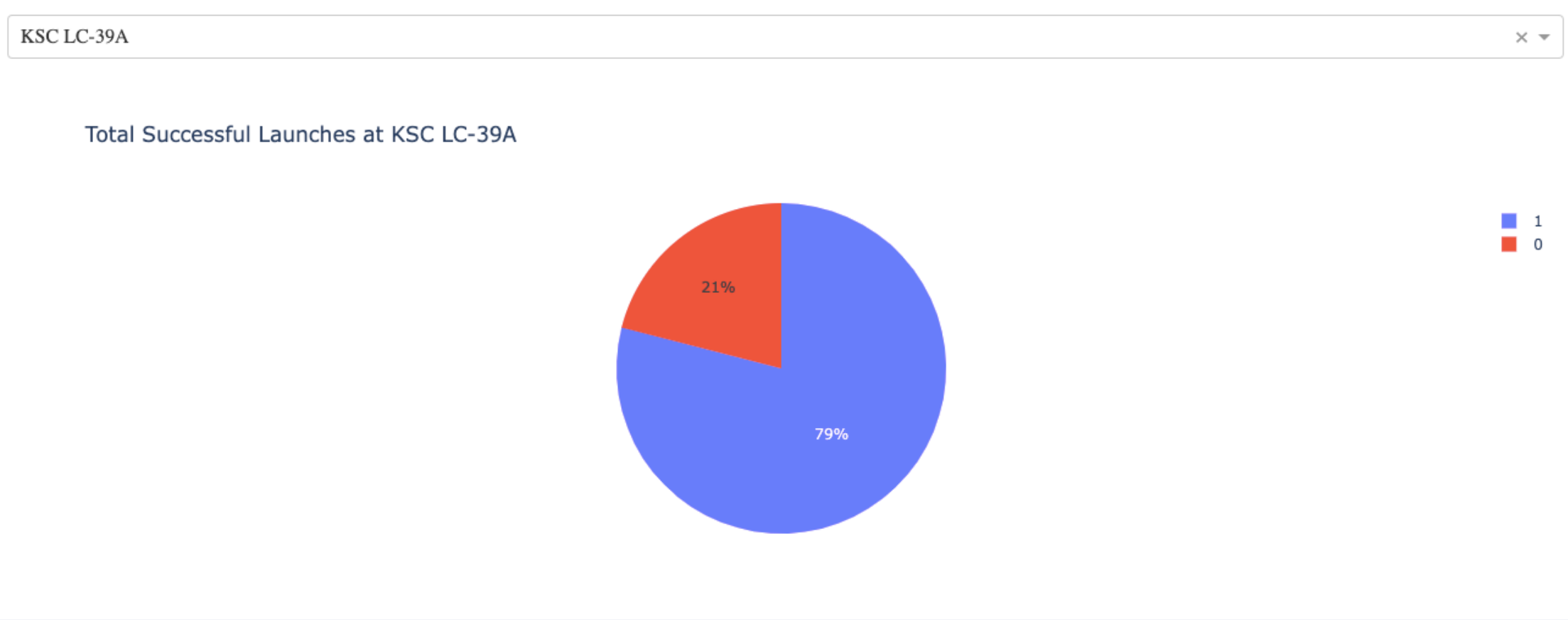
Total Successful Launches at All Sites

- Highest proportions of successful launches are recorded at KSC LC-39A, followed by CCAFS LC-40, VAFB SLC-4E and CCAFS SLC-40.



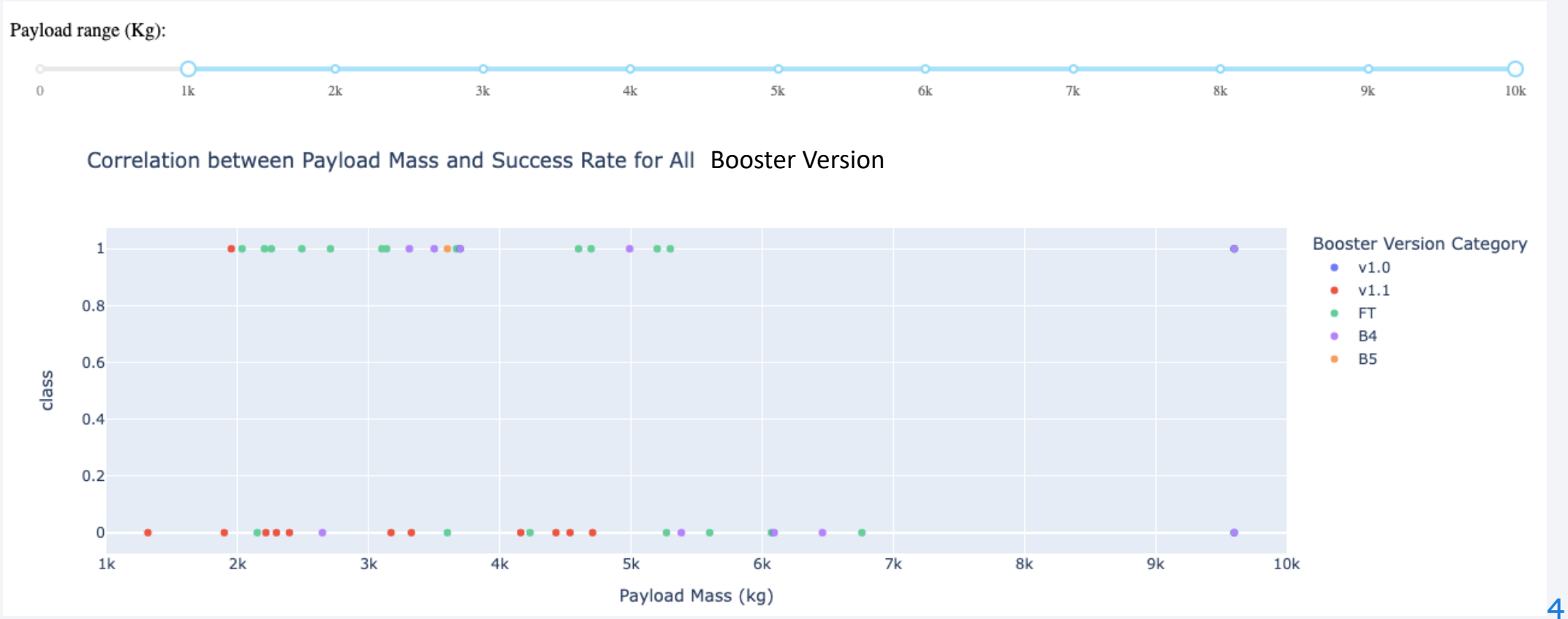
<Dashboard Screenshot 2>

- At KSC LC-39A, the success rate is 79%.



Correlation between Booster Version and Payload Mass

- User can select the payload range from the slider to narrow down to an interested range of payload to determine the success rate of different booster version.



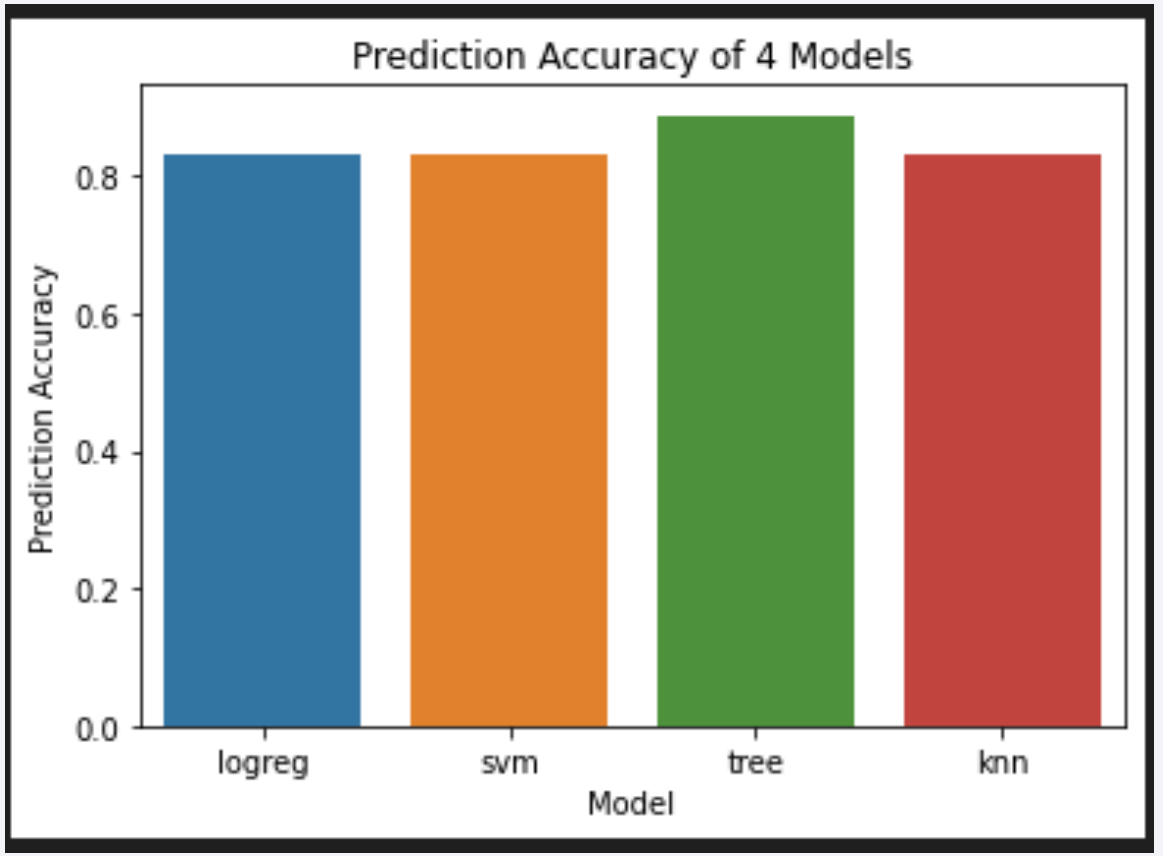
Section 5

Predictive Analysis (Classification)

Classification Accuracy

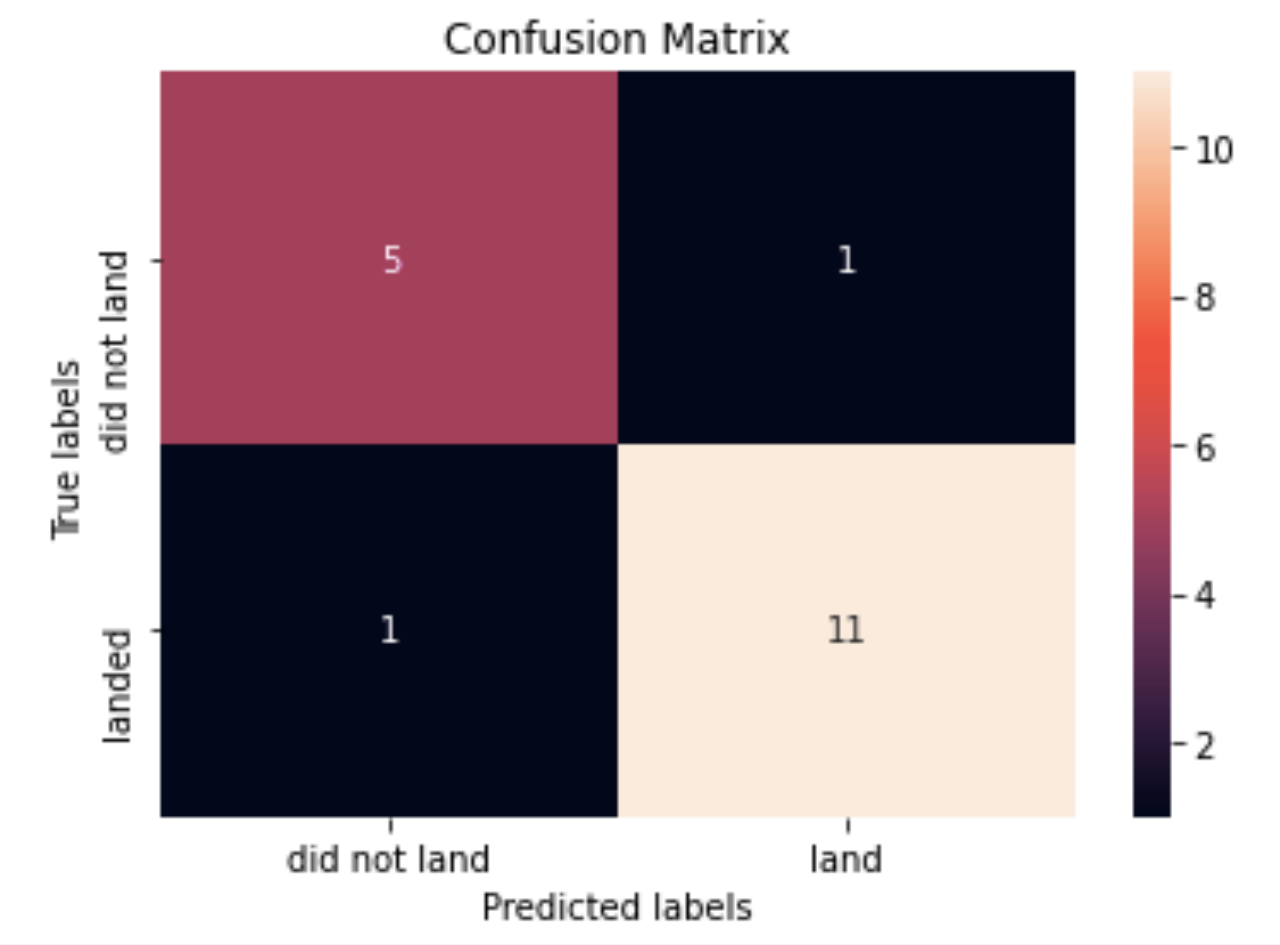
- Best accuracy is determined by using a bar chart.

- Decision tree has the highest accuracy among all 4 models.



Confusion Matrix

- Decision Tree has the highest accuracy among all models. According to the confusion matrix, only 1 false positive and 1 false negative is reported.



Conclusions

- In terms of orbit, with heavy payloads, the successful landing or positive landing rate are more for Polar, LEO and ISIS.
- In terms of launch sites, higher payload (above 7000kg) has higher success rate launching at CCAFS SLC-40, whereas payload ranging between 1000kg and 5000kg has higher success rate launching at VAFB SLC-4E and KSC LC 39A.
- KSC LC-39A records the most number of successful launches at 41.7%, which its success rate of 79%.
- Decision Tree is the most suitable model to be used for prediction of whether 1st stage of Falcon 9 will land successfully.

Thank you!