IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Douglas Heatherly
April 24, 2025

Outline

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

Executive Summary

- Methodologies used in this study of SpaceX.
  - Data was collected from a public API source of SpaceX using web scraping.
  - Data cleaning and wrangling methods were used to extract landing outcomes to serve as input for machine learning.
  - Exploratory Data Analysis (EDA) was conducted using SQL queries and data visualizations (static plots, interactive maps and interactive dashboards) to discover insights about the data.
  - Predictive analysis was performed using machine learning models including (Logistic Regression, Support Vector Machine (SVM), Decision Tree and k-Nearest Neighbors (kNN).
- Results Summary
  - The data collected for the SpaceX Falcon 9 included key features that could be used to predict successful launches in the future including (launch date, payload mass, orbit type, launch site).
  - Logistic Regression, KNN and SVM performed equally well in predicting mission outcome.

Introduction

- This company wants to make predictions about the success of SpaceX Falcon 9 rocket first stage landings in competition with SpaceX. If the available data can be used to predict mission outcomes, then this coming can focus efforts in areas that are more likely to generate successful mission outcomes and compete more effectively with SpaceX.
- Problems to explore in this study include:
  - What is the nature, extent and availability of public data available on SpaceX Falcon 9 first stage landings?
  - Which machine learning model works best for predicting mission outcomes?
  - Can a predictive model created that can be used for forecasting future successful landings?

Section 1

Methodology

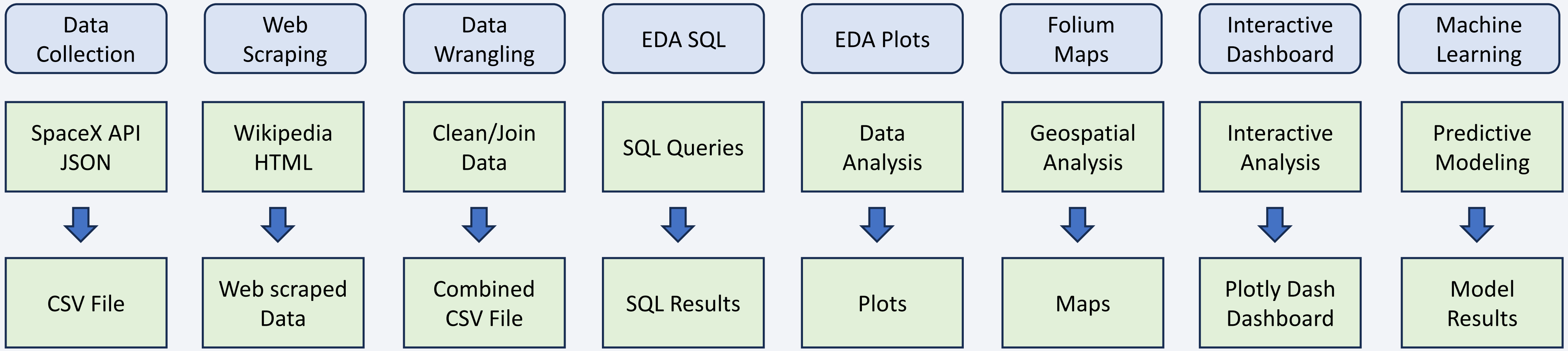
Methodology

Executive Summary

- Data collection methodology:
  - Data was extracted from a public API and web scraped from a Wikipedia article. Data sets were saved in CSV file format for use in the study.
- Perform data wrangling
  - Data was wrangled and cleaned for use with visualizations, queries and machine learning models.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Several classification and machine learning models were used to model the data.

Data Collection

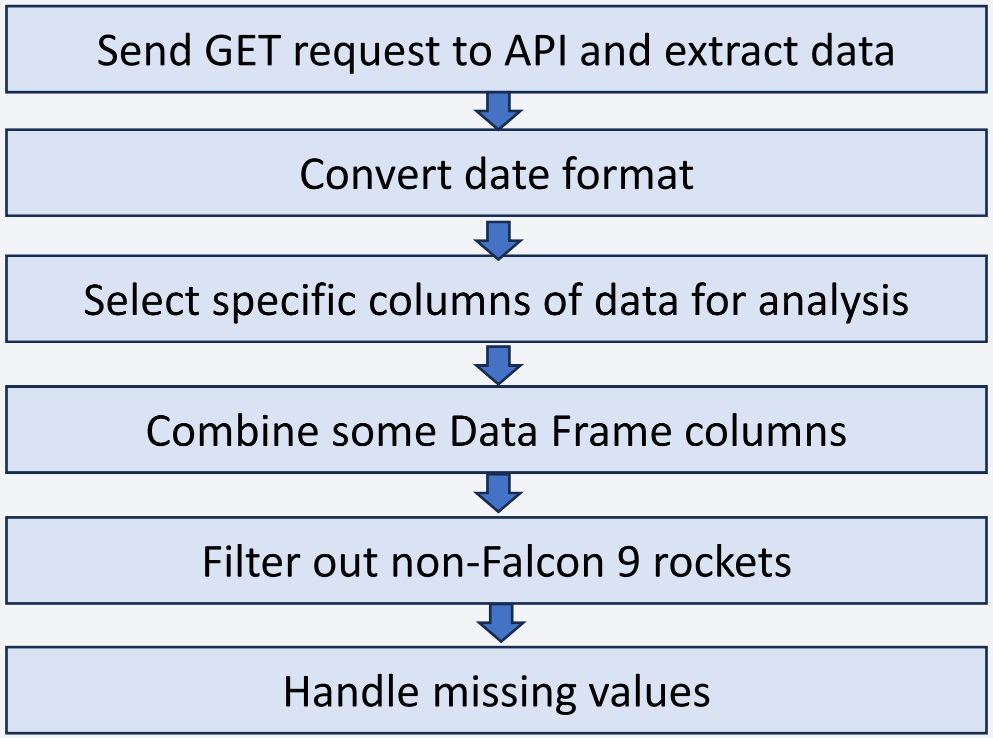
- The data was collected from the following sources:
  - A copy of a response from IBM from a public API with launch data in JSON format.
  - A Wikipedia page with launch data in HTML tables (web scraped)
  - Additional data provided as CSV files by the study sponsors.
- Data collection process key phrases and flowchart



Data Collection – SpaceX API

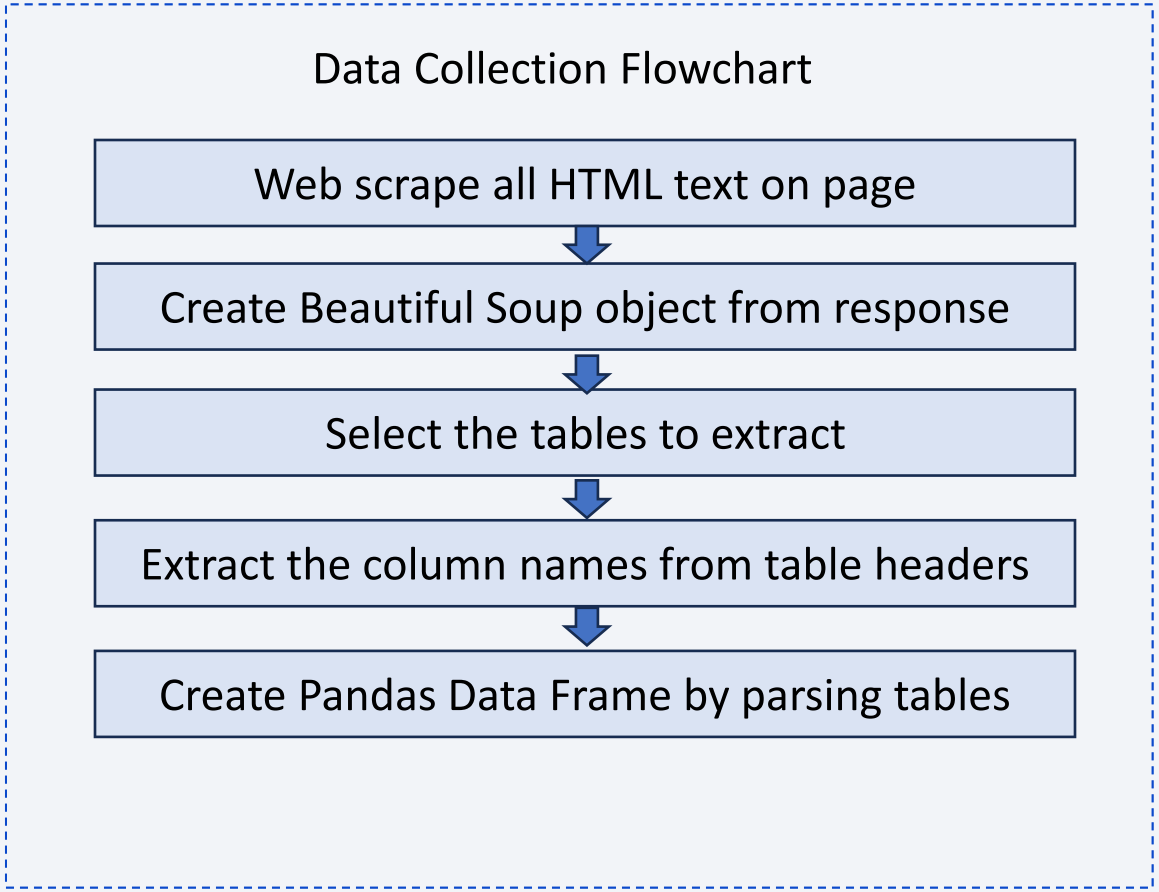
- SpaceX data was obtained from API endpoint: https://api.spacexdata.com/
- Data was extracted from API and loaded into a Panda data frame
- GitHub URL for Jupyter Notebook Data Collection: https://github.com/dheatherly/Data-Science/blob/main/jupyter-labs-spacex-data-collection-api-v2.ipynb

Data Collection Flowchart



Data Collection - Scraping

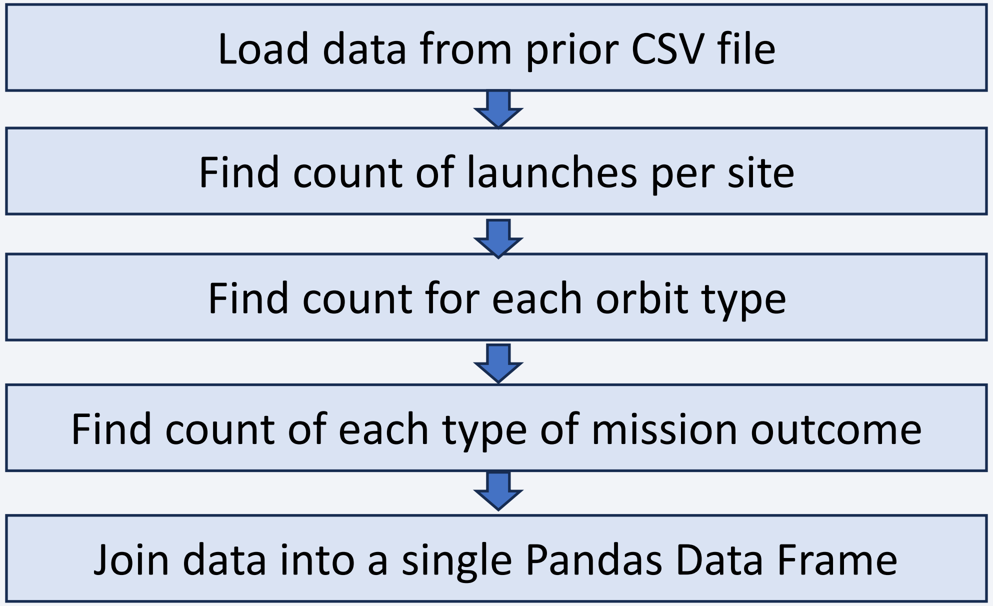
- SpaceX launch data was scraped from SpaceX Wikipedia webpage HTML tables and loaded into Pandas Data Frame for analysis.
- GitHub URL for Jupyter Notebook (web scraping): https://github.com/dheatherly/Data-Science/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

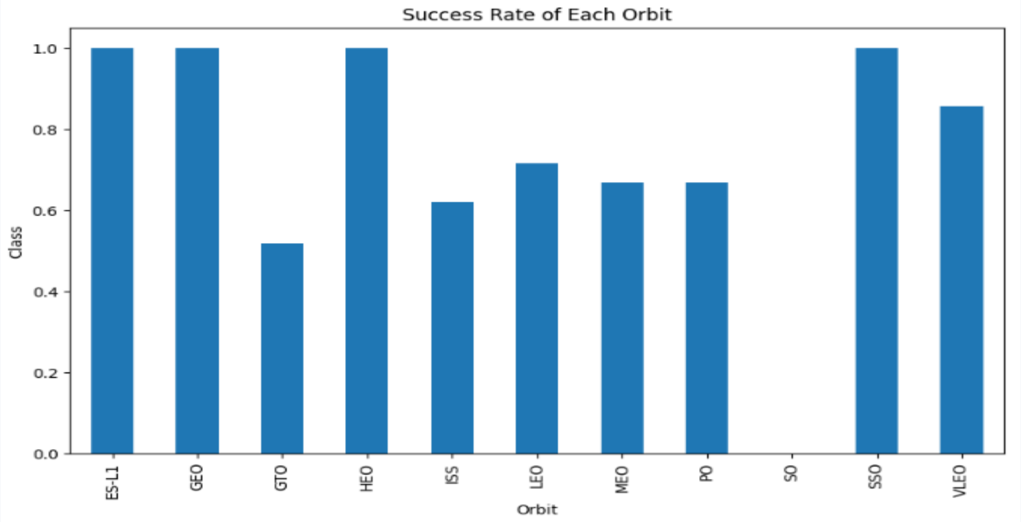
- Reformat and organize orbit type, launch sites and mission outcomes.
- Convert mission outcomes to binary using one-hot encoding technique (1 being mission success and 0 being mission failure) and add the new outcome variable to the Data Frame.
- GitHub URL for Jupyter Notebook (data wrangling): https://github.com/dheatherly/Data-Science/blob/main/labs-jupyter-spacex-Data%20wrangling-v2.ipynb

Data Wrangling Flowchart



EDA with Data Visualization

- Scatterplots of launch site trends were created showing mission outcome by launch site and flight number and launch site and payload.
- Scatterplots of orbit types were created for launch site trends showing mission outcome relationship by orbit type and flight number and orbit type and payload.
- GitHub URL for Jupyter Notebook (data visualization): https://github.com/dheatherly/Data-Science/blob/main/jupyter-labs-eda-dataviz-v2.ipynb

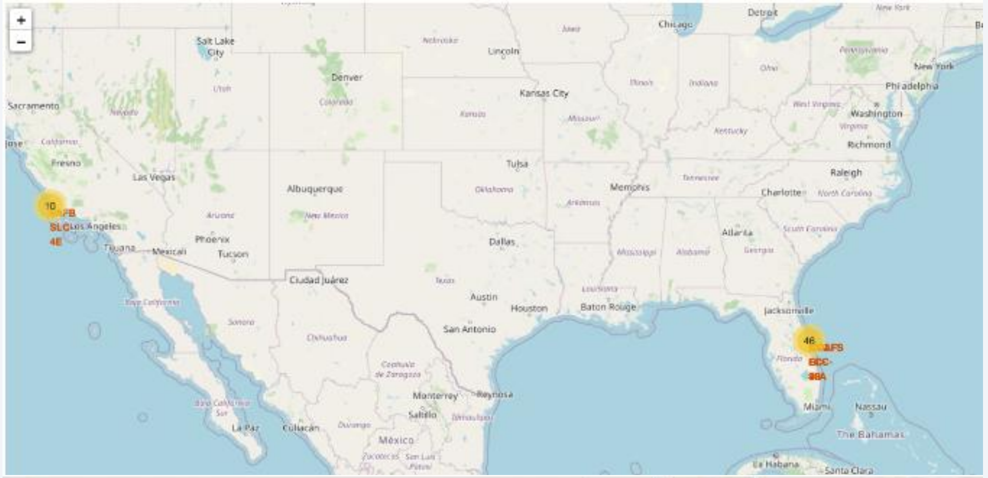


EDA with SQL

- SQL Queries Performed
  - Display the names of 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 data 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, launch site 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 for Jupyter Notebook (SQL queries): https://github.com/dheatherly/Data-Science/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Found patterns in the geographical data using map of the launch sites:
  - All launch sites
  - Successful and Failed Launches
  - Distances between launch sites and proximal landmarks
- GitHub URL for Jupyter Notebook (Maps): https://github.com/dheatherly/Data-Science/blob/main/lab-jupyter-launch-site-location-v2.ipynb



Build a Dashboard with Plotly Dash

- Built an interactive dashboard using Plotly Dash
- Plotted pie charts showing total launches for specific sites
- Created scatterplots showing relationship between mission outcome and payload mass for different booster versions.
- GitHub URL for python program (Plotly Dashboard): https://github.com/dheatherly/Data-Science/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Data split into training and testing sets for modeling.
- Machine learning models include:
  - Logistic Regression
  - Support Vector Machine (SVM)
  - Decision Tree
  - K-Nearest Neighbor (kNN)
- Hyper parameters assessed using GridSearchCV
- Each model was scored for accuracy
- GitHub URL for Jupyter Notebook (Machine Learning): https://github.com/dheatherly/Data-Science/blob/main/SpaceX-Machine-Learning-Prediction-Part-5-v1.ipynb

Predictive Analysis Flowchart

Cleaned data loaded into Pandas Data Frame

↓

Data split into training and test datasets

↓

Four models were trained as specified

↓

Each model evaluated with test dataset

↓

Models compared using accuracy scores

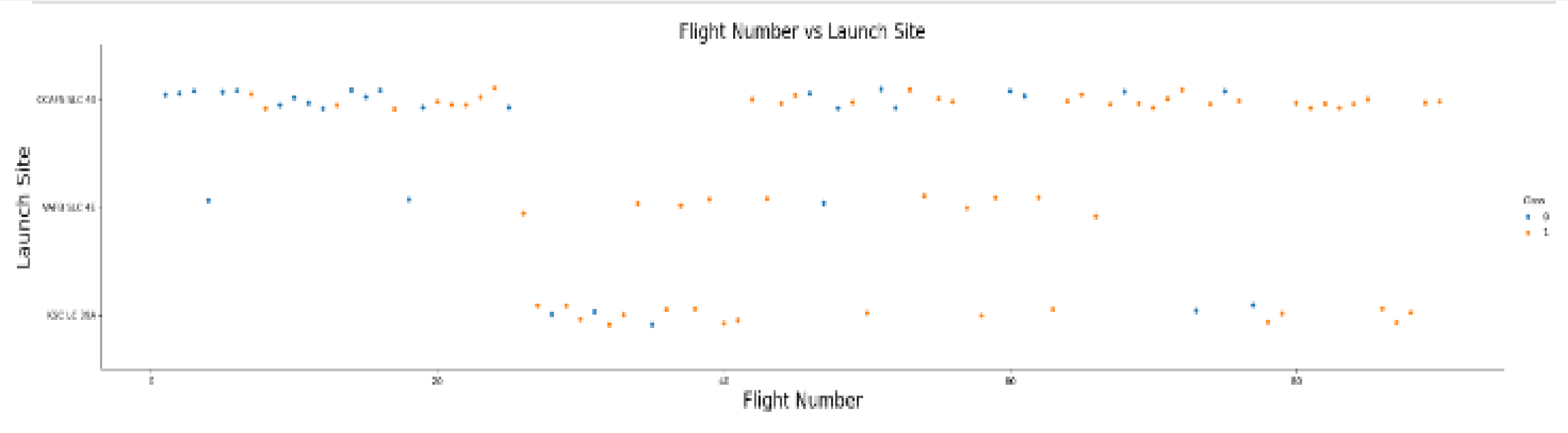
Results

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

Section 2

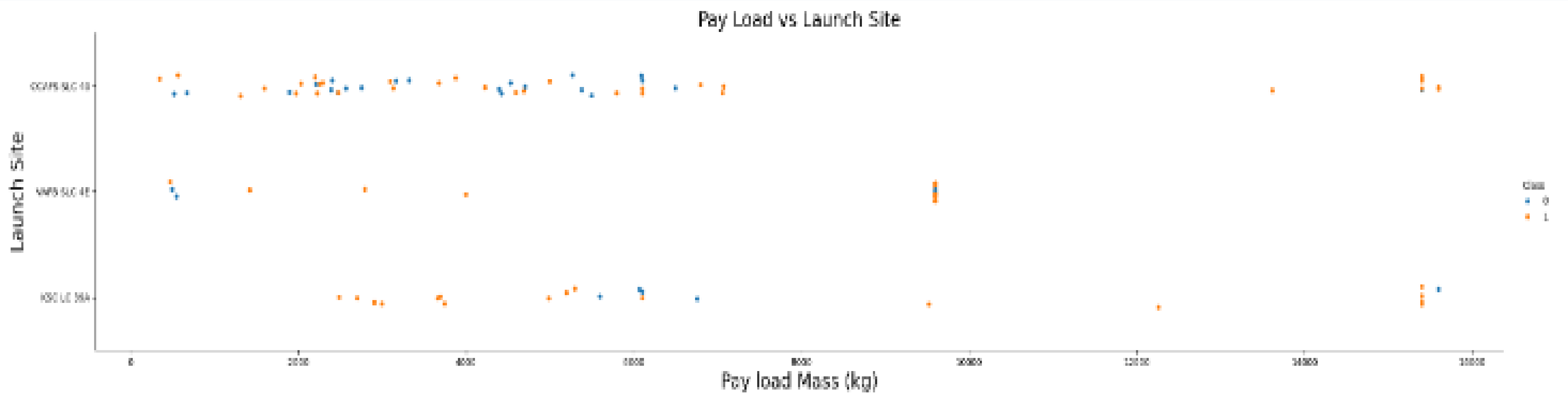
Insights drawn from EDA

Flight Number vs. Launch Site



- Mission success varies by launch site
- Falcon 9 first stage landings appear more successful as flight number increases.

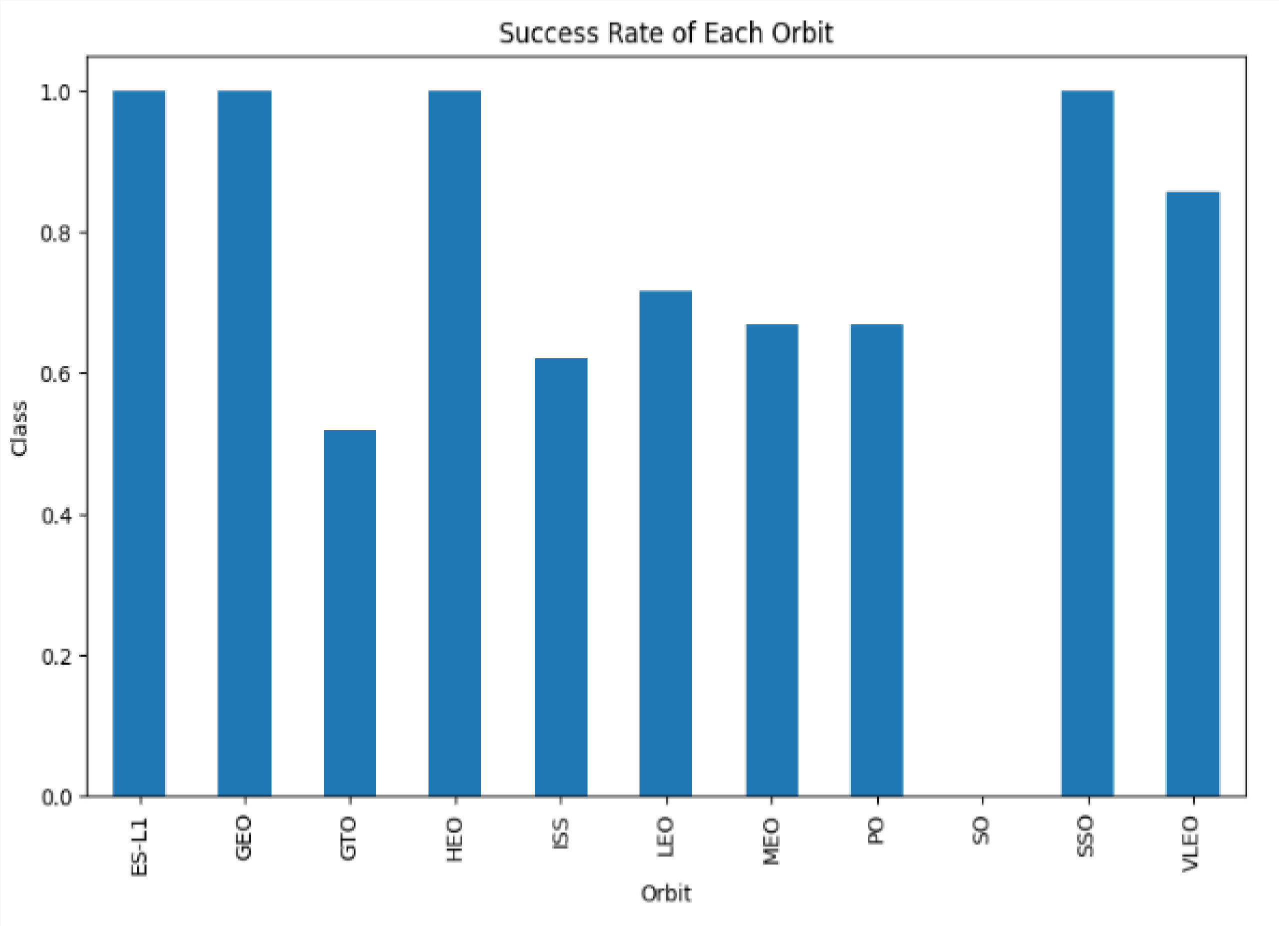
Payload vs. Launch Site



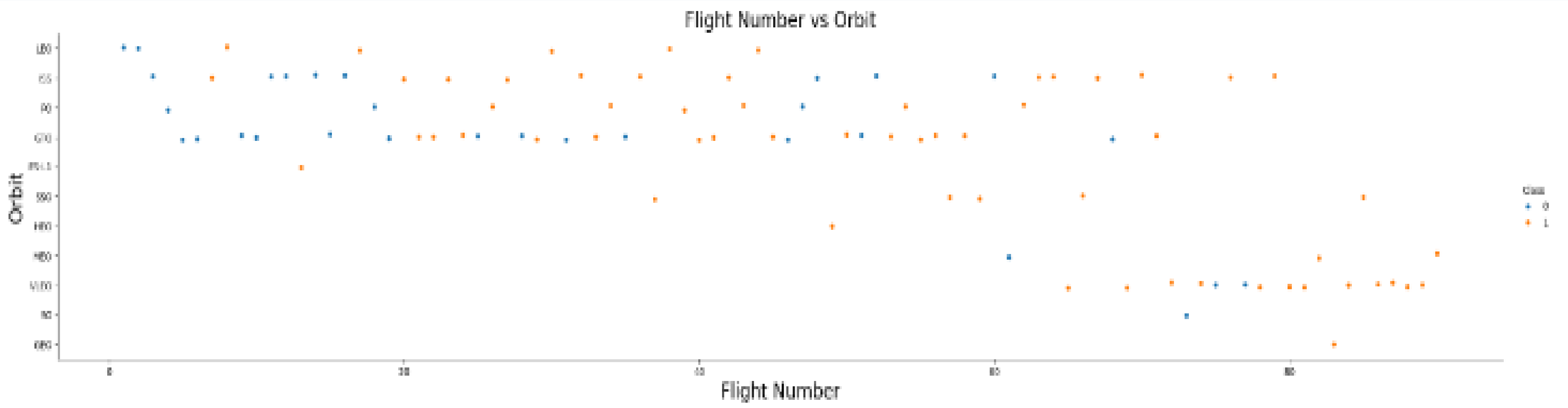
- Payloads greater than 9000 kg have excellent success rate for all three launch sites
- Payloads greater than 12000 kg seem to be limited to only two launch sites (CCAFS SLC40 and KSC LC).

Success Rate vs. Orbit Type

- ES-L1, GEO, HEO and SSO orbits all have 100% successful first stage landings.
- SO orbit has no successful first stage landings.

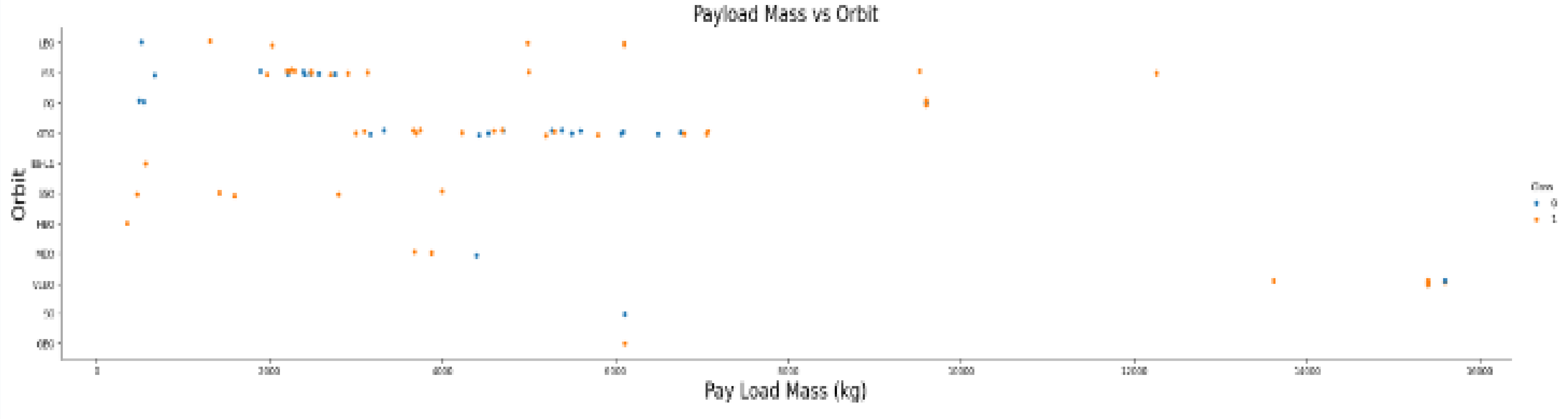


Flight Number vs. Orbit Type



- Success rates improve over time as reflected in the higher success rates with higher flight numbers.
- GTO and ISS orbits have a higher number of launches with the lowest success rate.

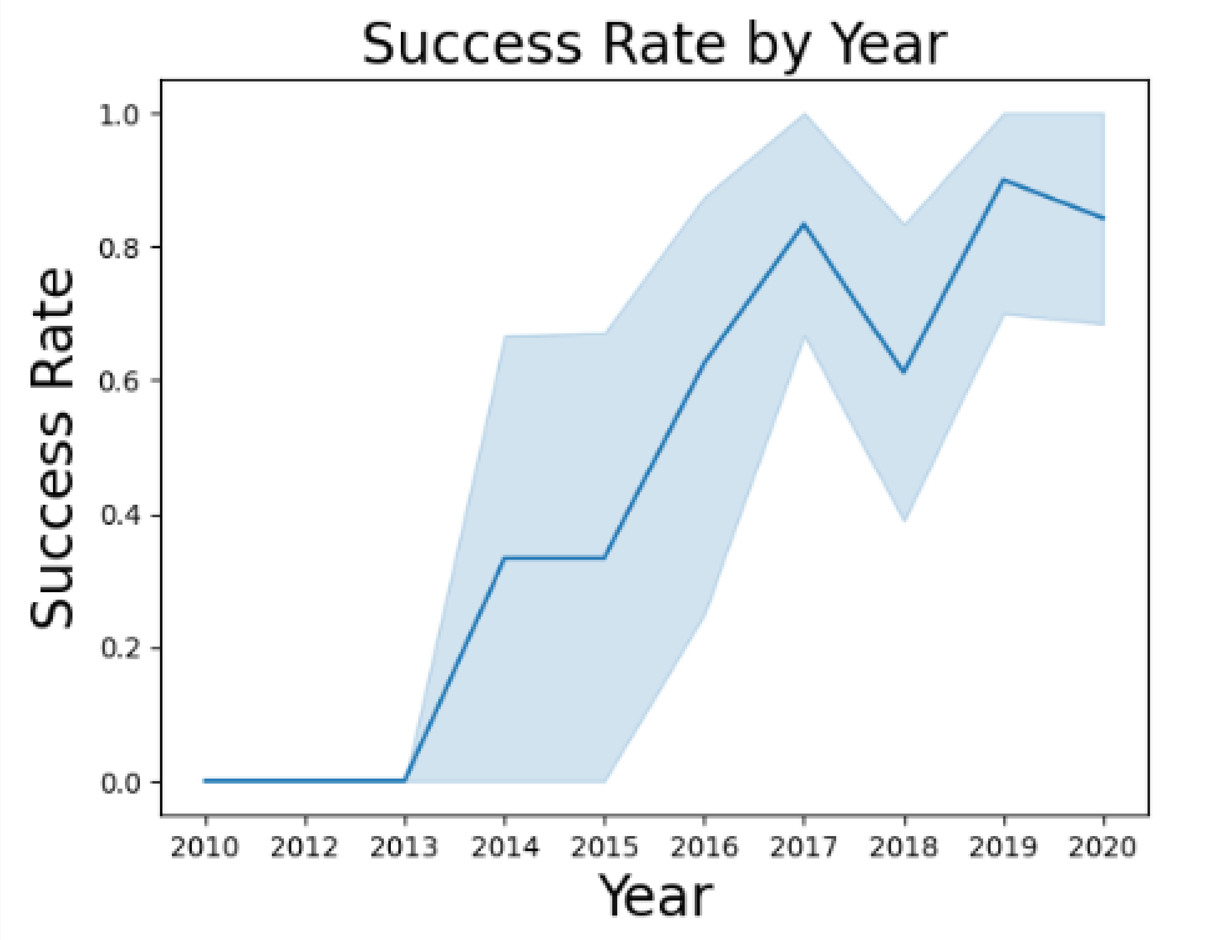
Payload vs. Orbit Type



- With heavy payloads, the successful landing or positive landing rate are more for Polar, LEO and ISS.
- GTO has an equal mixture of successful and failed landings.

Launch Success Yearly Trend

- The success rate since 2013 kept increasing until 2017 (stable in 2014) and after 2015 it started increasing again.



All Launch Site Names

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

```
%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'

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

```
%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

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

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)'
```

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

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

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1'
```

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

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

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

Hint:Use min function

```
%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

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 LANDING_OUTCOME = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND 
```

```
 * 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

List the total number of successful and failure mission outcomes

```
%sql SELECT count(MISSION_OUTCOME) FROM SPACEXTBL WHERE MISSION_OUTCOME LIKE 'Success%' OR MISSION_OUTCOME LIKE 'Failure%'
```

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

count(MISSION_OUTCOME)
101

Boosters Carried Maximum Payload

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

```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT max(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

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

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

```
%sql SELECT substr(DATE,6,2) as MONTH, LANDING_OUTCOME, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE LANDING_OUTCOME =
```

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

MONTH	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

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(LANDING_OUTCOME) from SPACEXTBL WHERE DATE between '2010-06-04' and '2017-03-20' GROUP B
```

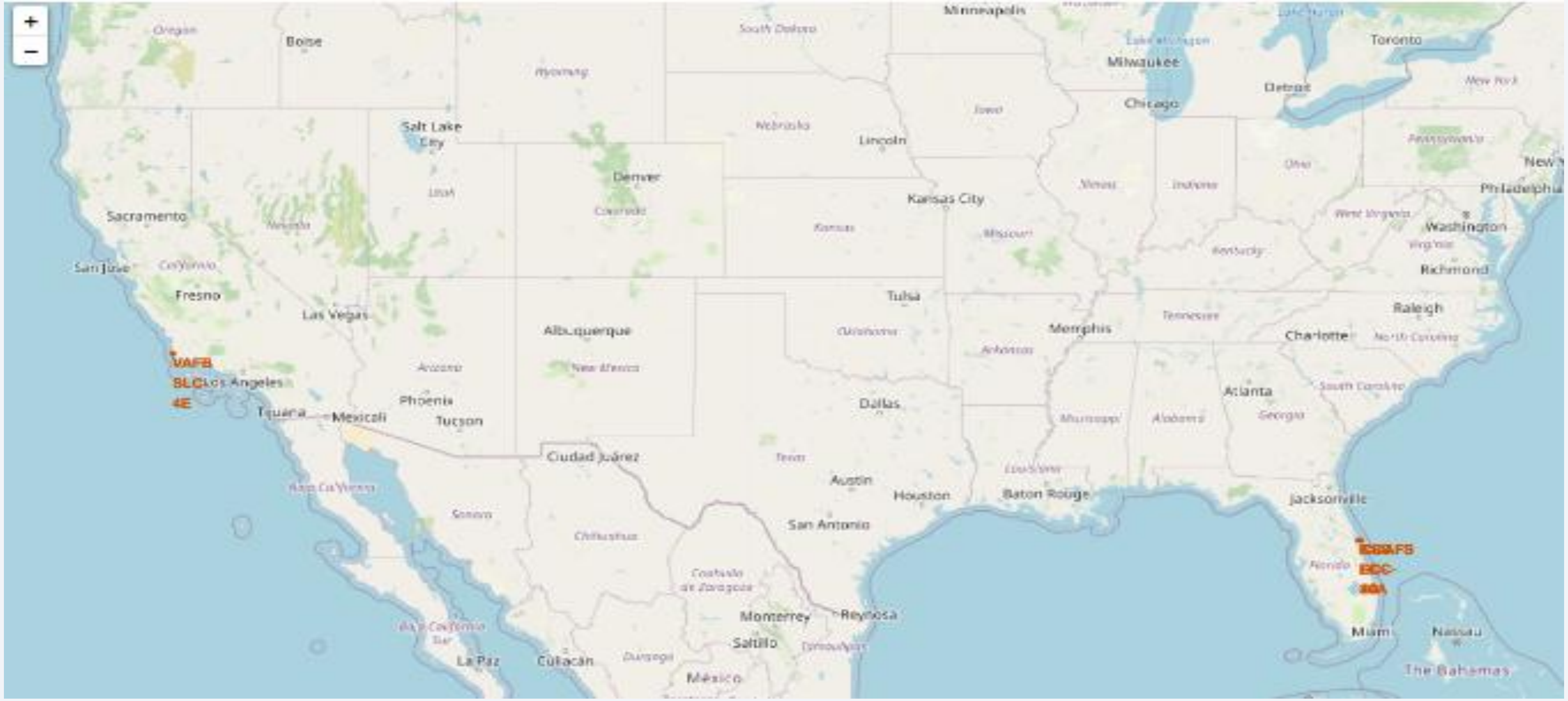
```
 * sqlite:///my_data1.db
Done.
```

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

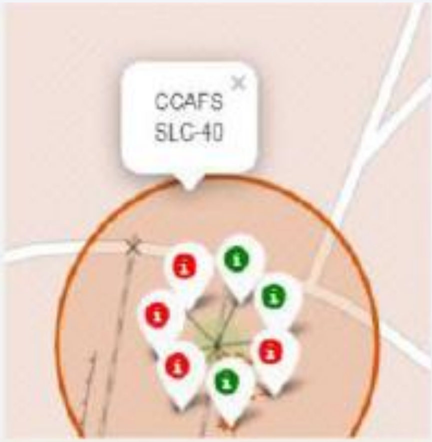
Launch Sites Proximities Analysis

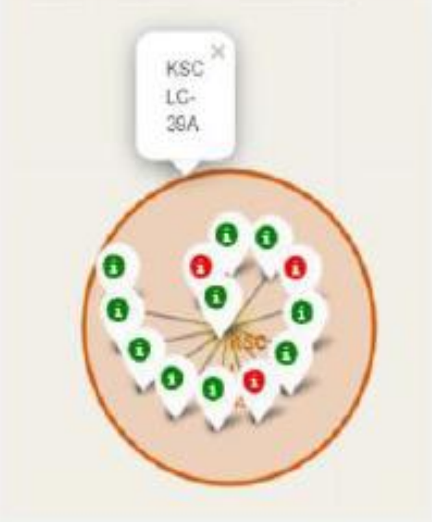
Launch Site Locations

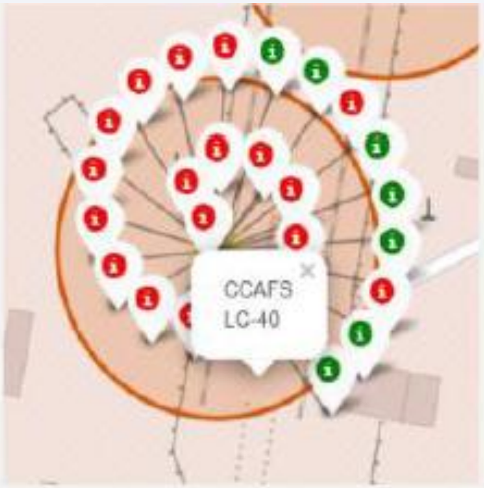


- All launch sites are in the US near the equator
- All launch sites are near the coast.

Launch site locations with success/failed markers





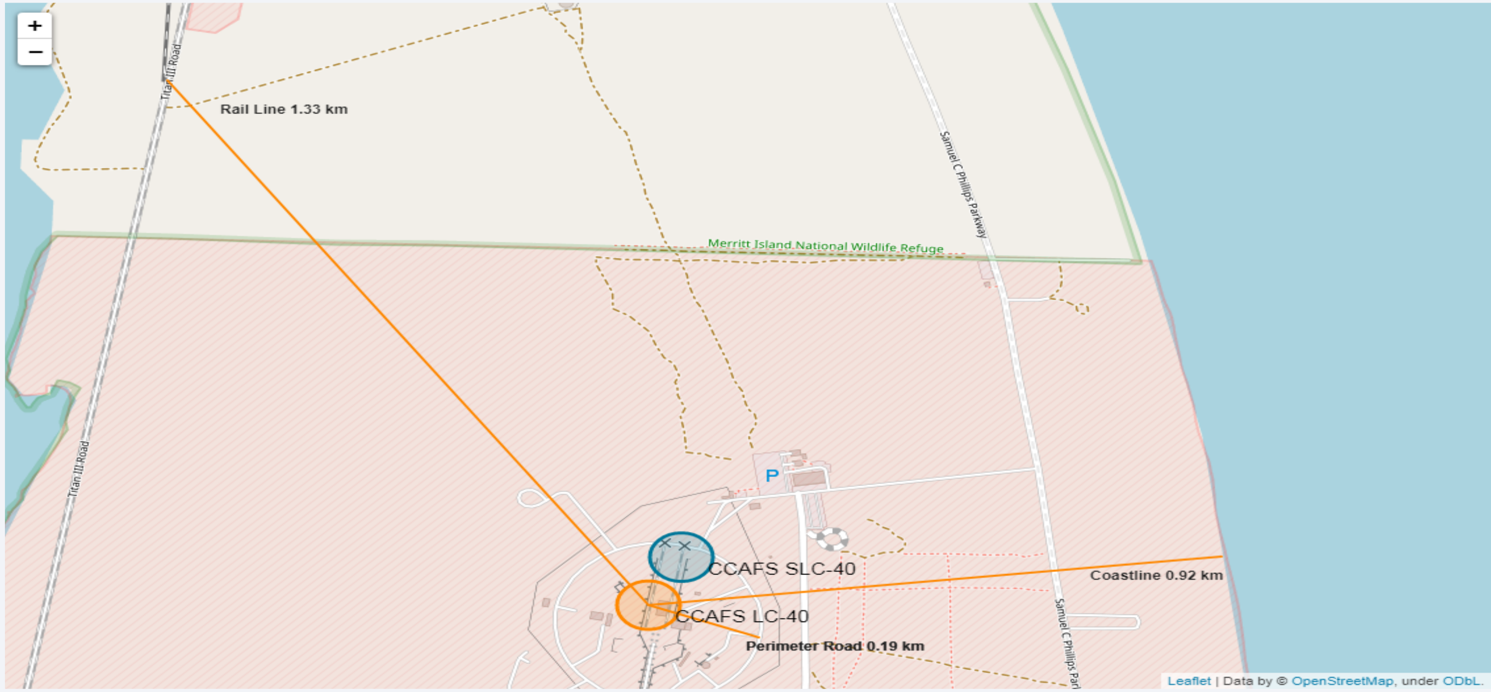


California Launch Site

Florida Launch Sites

Green Marker shows successful Launches and Red Marker shows Failures

<Folium Map Screenshot 3>



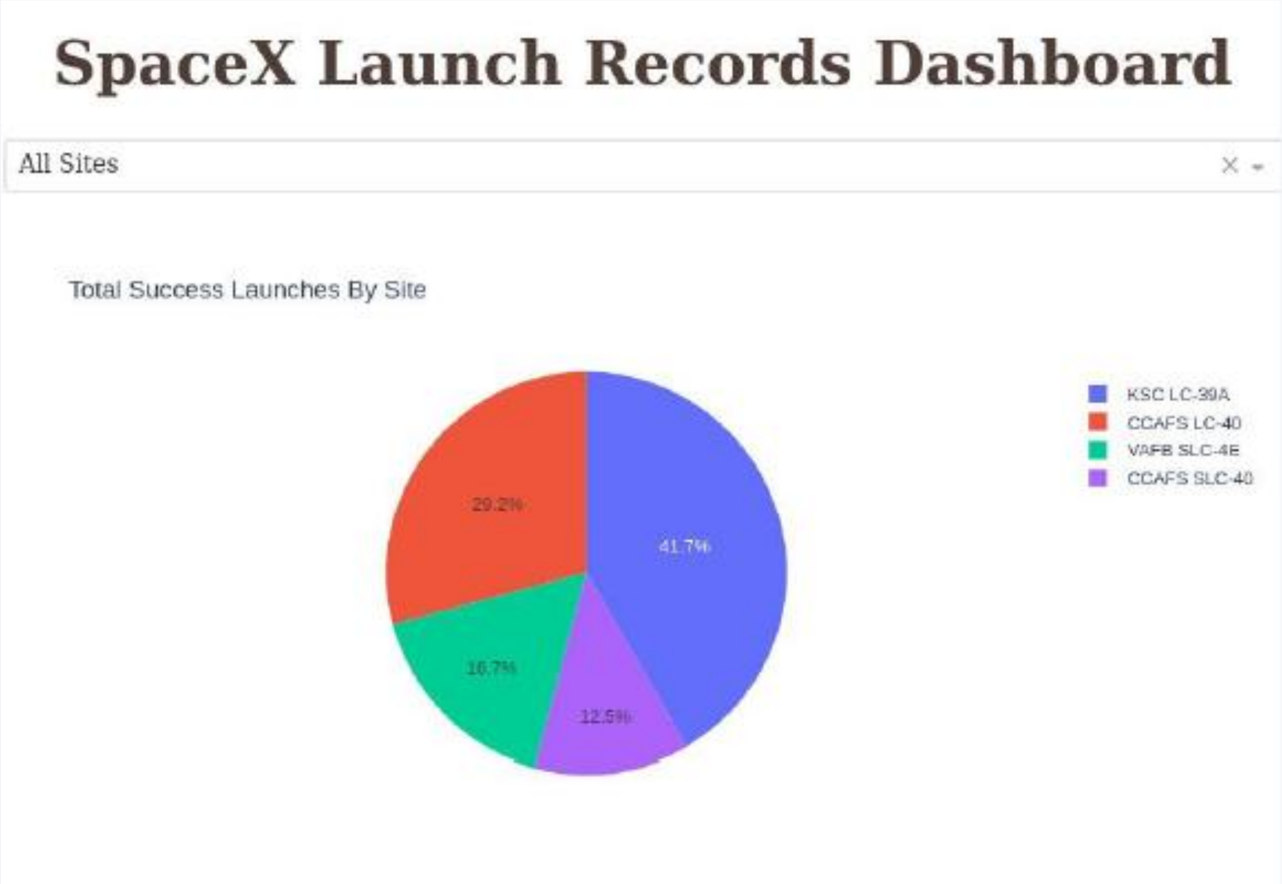
- Launch site is optimally located near railroad and roads and away from populated areas.

Section 4

Build a Dashboard with Plotly Dash

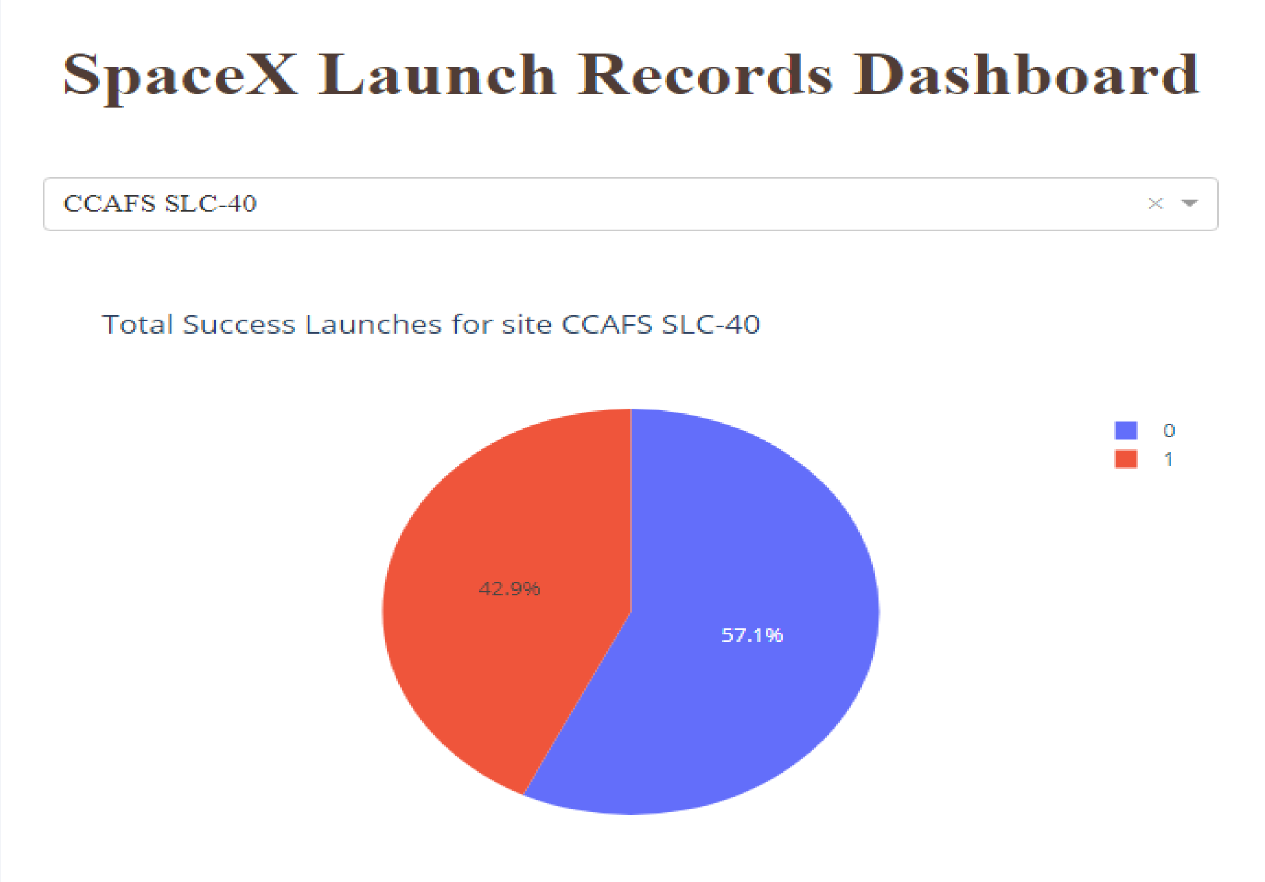
Total Successful Launches by Site (all sites)

- The greatest successful first stage landing occurred at KSC LC-39A and 41.7%.

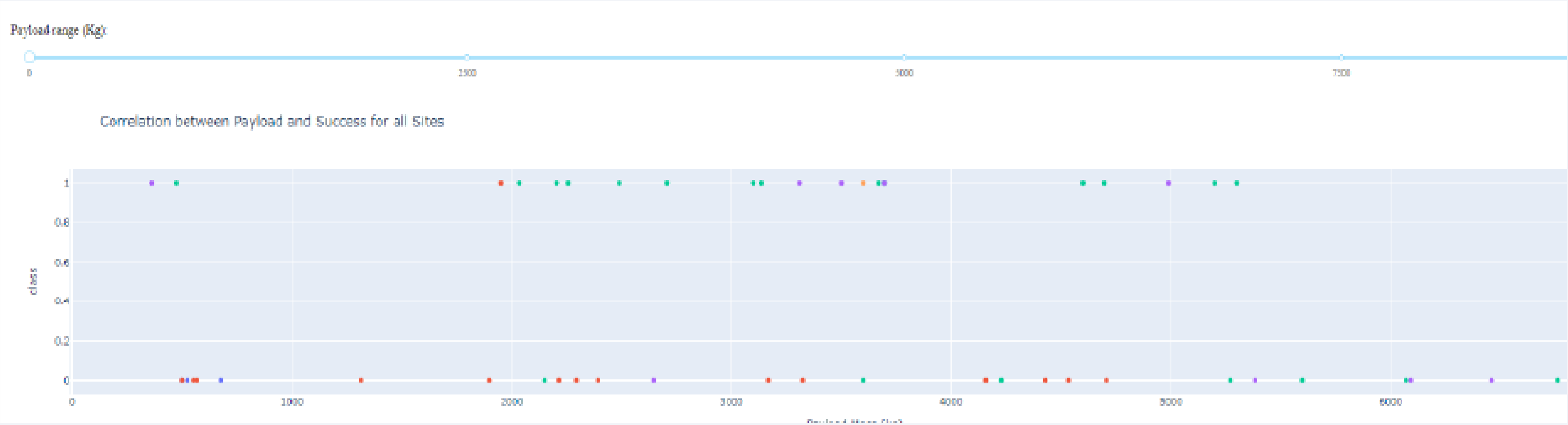


Launch Site with Highest Success Ratio

- CCAFS SLC-40 has the highest first stage landing success rate at 42.9%.



Payload vs Launch Outcome



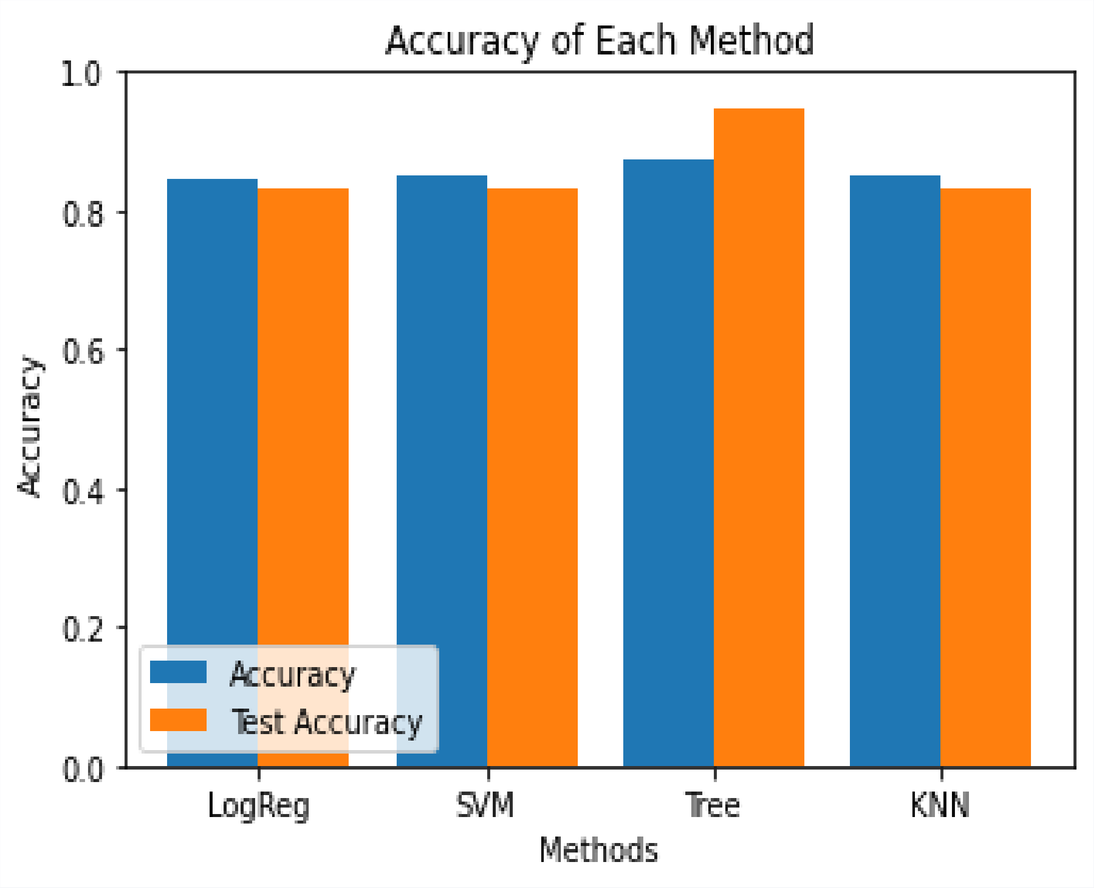
- In general, success rates are higher for lower weight payloads than higher weight payloads.
- Payloads under 5000kg and FT boosters have the most successful combination.

Section 5

Predictive Analysis (Classification)

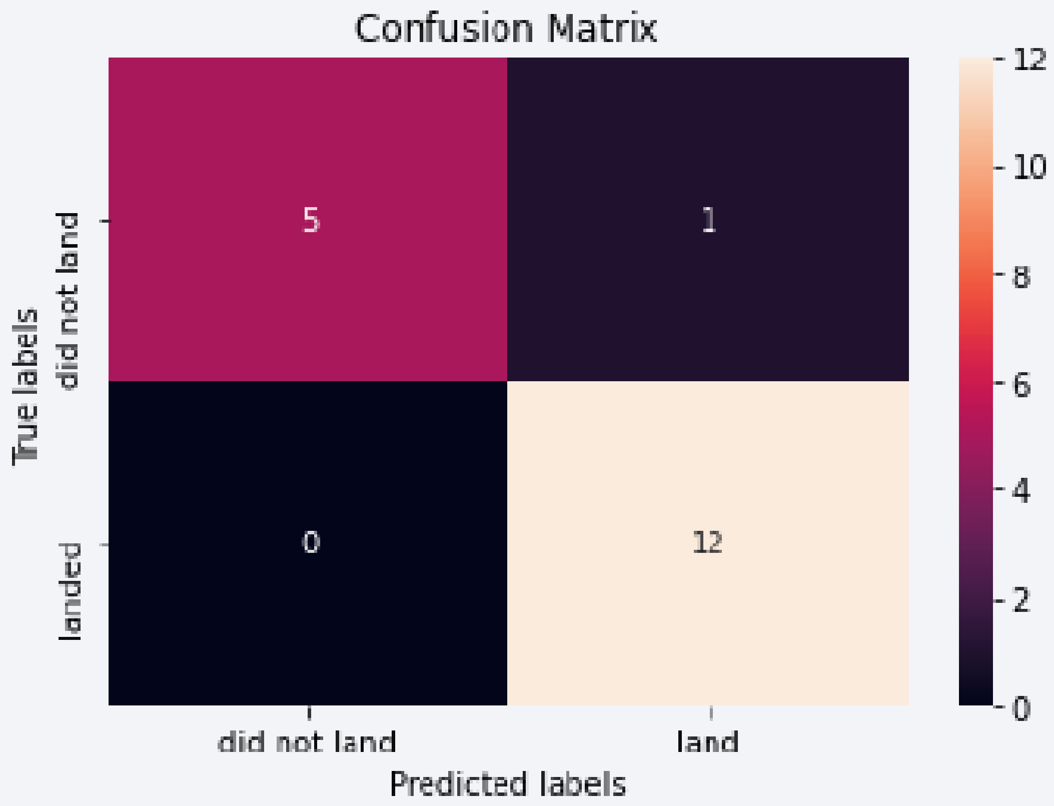
Classification Accuracy

- The Decision Tree Classifier shows best results with the highest accuracy compared to Logistic Regression, Support Vector Model and k-Nearest Neighbors



Confusion Matrix

- The Decision Tree Classifier shows best results with high predicted vs. actual labels for true positive and true negative.



Conclusions

- First stage landings success rates improve as more flights are launched from sites.
- Payloads greater than 9000 kg tend to have excellent success rate for all sites
- In general, success rates appear to improve year over year.
- The greatest successful first stage landing occurred at KSC LC-39A
- The Decision tree classifier is the best machine learning model used for this study.

Thank you!