



IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sourabh Khatri
17-08-2023



Outline

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

Executive Summary

Various methodologies were employed for the data analysis:

- Data was acquired through web scraping and utilizing the SpaceX API.
- Exploratory Data Analysis (EDA) techniques were applied, encompassing tasks such as data cleaning, visual representation, and interactive visual assessments.
- Machine Learning techniques were employed for predictive purposes.

A comprehensive summary of the outcomes is as follows:

- Significant and relevant data was successfully gathered from publicly accessible sources.
- The EDA phase effectively pinpointed the most influential features for predicting the success of launch missions.
- Through Machine Learning Prediction, a robust model was identified that successfully discerns the crucial factors for optimizing launch success. This model employs the complete dataset compiled during the data collection phase.

Introduction

The aim is to assess the feasibility of the emerging company, Space Y, in establishing itself as a competitive force in contrast to Space X.

Preferred insights include:

- Determining the optimal approach for estimating the overall launch costs, achieved through predicting the favorable outcomes of initial rocket stage landings.
- Identifying the prime launch location for maximizing operational efficiency.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

Data from Space X was obtained from 2 sources:

- Space X API (<https://api.spacexdata.com/v4/rockets/>)
- Web-Scrapping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)

- Perform data wrangling

- Collected data was enriched by creating a landing outcome label based on outcome data after summarizing and analyzing features

- Perform exploratory data analysis (EDA) using visualization and SQL

- Perform interactive visual analytics using Folium and Plotly Dash

Methodology

- Perform predictive analysis using classification models
 - Data that was collected until this step were normalized, divided in training and test data sets and evaluated by four different classification models, being the accuracy of each model evaluated using different combinations of parameters

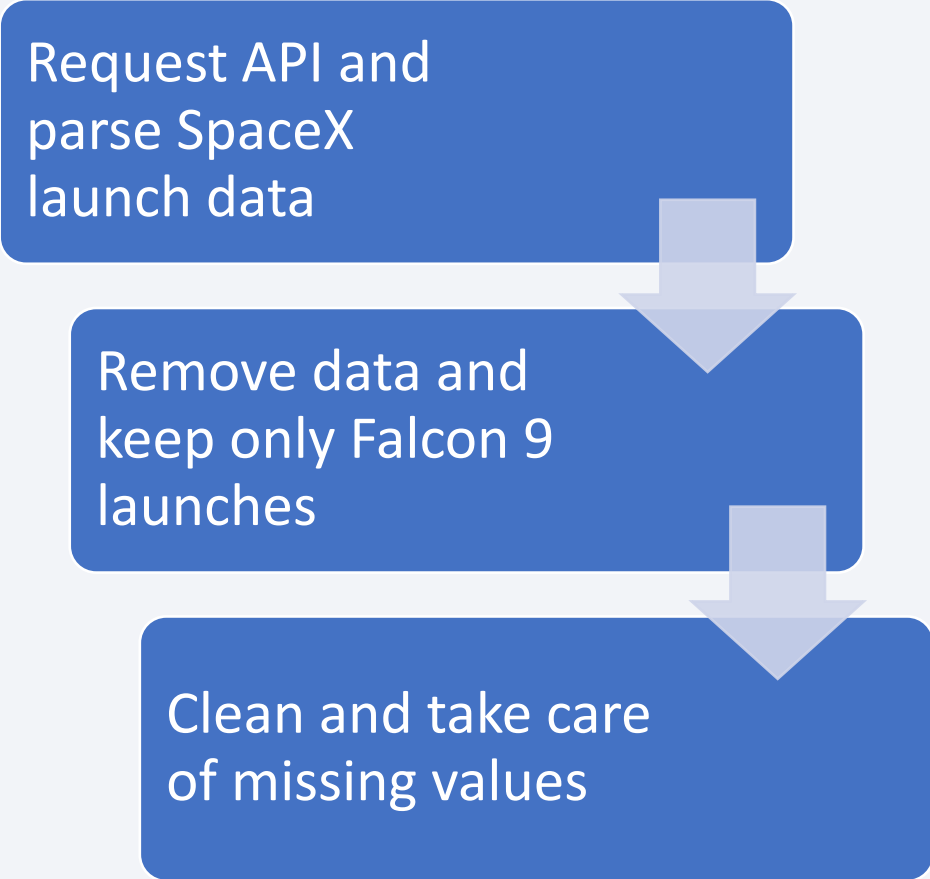
Data Collection

- Data sets were collected from Space X API (<https://api.spacexdata.com/v4/rockets/>) and from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches), using web scraping technics

Data Collection – SpaceX API

- SpaceX provides an accessible public API that facilitates the retrieval of relevant data for subsequent utilization.
- The outlined flowchart was followed to interact with this API, and the acquired data was then stored for further processing.
- GitHub URL: <https://github.com/Sourabh-khatrii/Applied-Data-Science-Capstone/blob/main/W1.1%20jupyter-labs-spacex-data-collection-api.ipynb>

Request API and
parse SpaceX
launch data



Remove data and
keep only Falcon 9
launches

Clean and take care
of missing values

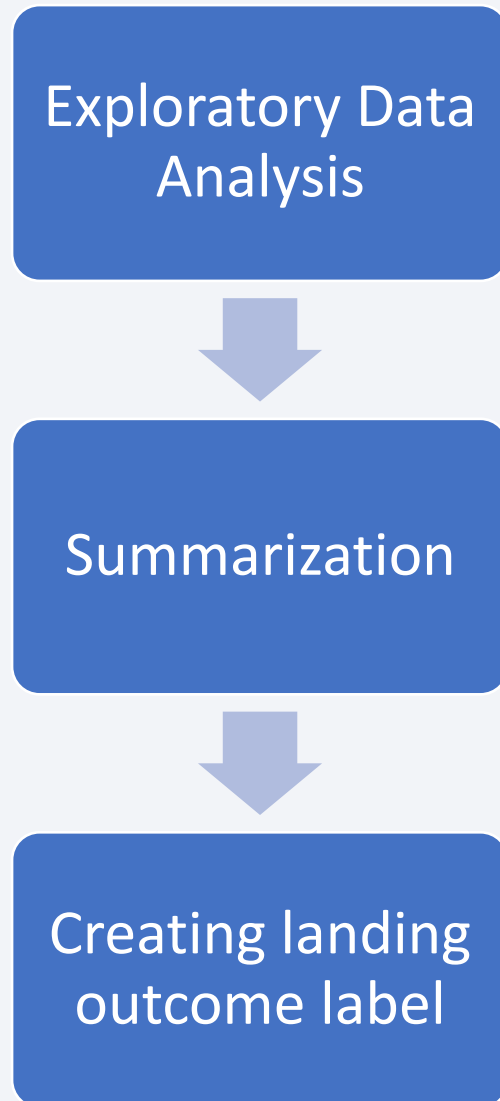
Data Collection - Scraping

- Information regarding SpaceX launches is additionally attainable through Wikipedia sources.
- The process involves downloading data from Wikipedia, adhering to the provided flowchart, and subsequently storing the acquired data for future reference.
- GITHUB URL: <https://github.com/Sourabh-khatrii/Applied-Data-Science-Capstone/blob/main/W1.2%20jupyter-labs-webscraping.ipynb>



Data Wrangling

- At the outset, an exploratory analysis of the dataset was conducted.
- Subsequently, calculations were carried out to determine the count of launches per launch site, the frequency of each orbit type, and the distribution of mission outcomes based on orbit type.
- Lastly, the labeling of landing outcomes was established by deriving this information from the "Outcome" column.
- GitHub URL: https://github.com/Sourabh-khatrii/Applied-Data-Science-Capstone/blob/main/W1.3%20labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb



EDA with Data Visualization

- Scatterplots and bar plots were used to visualize the relationship between pairs of features to explore data, as they allow to compare two unrelated variables and easily find the relationship between them:
 - Payload Mass X Flight Number
 - Launch Site X Flight Number
 - Launch Site X Payload Mass
 - Orbit X Flight Number
 - Payload mass X Orbit
- GitHub URL: <https://github.com/Sourabh-khatrii/Applied-Data-Science-Capstone/blob/main/W2.2%20jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb>

EDA with SQL

The subsequent SQL queries were executed:

- Identification of distinct launch sites engaged in space missions.
- Retrieval of the top 5 launch sites whose names commence with the sequence 'CCA'.
- Calculation of the cumulative payload mass transported by boosters launched under NASA's CRS initiative.
- Computation of the average payload mass carried by booster version F9 v1.1.
- Determination of the date when the initial successful landing on a ground pad was accomplished.
- Listing of booster names that achieved success on drone ships while carrying payload masses ranging from 4000 to 6000 kg.
- Compilation of the total count of successful and unsuccessful mission outcomes.
- Identification of booster versions responsible for transporting the highest payload masses.
- Gathering of data about failed landing outcomes on drone ships, including corresponding booster versions and launch site names, specifically for the year 2015.
- Ranking of landing outcome counts (e.g., Failure on drone ship or Success on ground pad) within the timeframe spanning from June 4, 2010, to March 20, 2017.

GitHub URL: https://github.com/Sourabh-khatrii/Applied-Data-Science-Capstone/blob/main/W2.1%20jupyter-labs-eda-sql-coursera_sqlite.ipynb

Build an Interactive Map with Folium

Folium Maps were utilized with a range of visual elements, including markers, circles, lines, and marker clusters.

- Markers were employed to denote significant locations, such as launch sites.
- Circles served to emphasize specific regions around given coordinates, such as the NASA Johnson Space Center.
- Marker clusters were utilized to group events at individual coordinates, such as aggregating launches at a particular launch site.
- Lines were employed to visually represent distances between pairs of coordinates.

GitHub URL: https://github.com/Sourabh-khatrii/Applied-Data-Science-Capstone/blob/main/W3.1%20lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

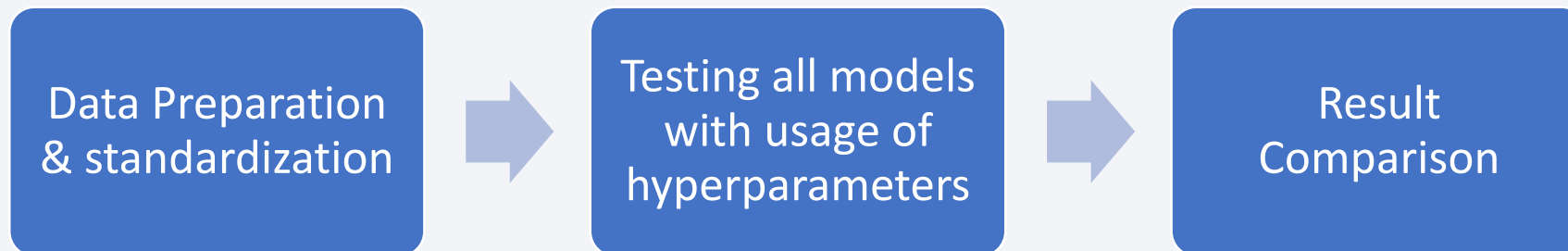
An array of graphs and visual representations was employed to depict the data effectively:

- Percentage of launches by site
- Payload range.
- This combination of visualizations facilitated swift analysis of the connection between payloads and launch sites, aiding in the identification of optimal launch locations based on payload considerations.
- GitHub URL: [https://github.com/Sourabh-khatrii/Applied-Data-Science-Capstone/blob/main/W3.2%20spacex dash app.py](https://github.com/Sourabh-khatrii/Applied-Data-Science-Capstone/blob/main/W3.2%20spacex%20dash%20app.py)

Predictive Analysis (Classification)

Four classification models were used and compared: Logistic regression, support vector machine, decision tree and k nearest neighbors.

GitHub URL: [https://github.com/Sourabh-khatrii/Applied-Data-Science-Capstone/blob/main/W4.1%20SpaceX Machine Learning Prediction Part 5.jupyterlite.ipynb](https://github.com/Sourabh-khatrii/Applied-Data-Science-Capstone/blob/main/W4.1%20SpaceX%20Machine%20Learning%20Prediction%20Part%205.ipynb)



Results

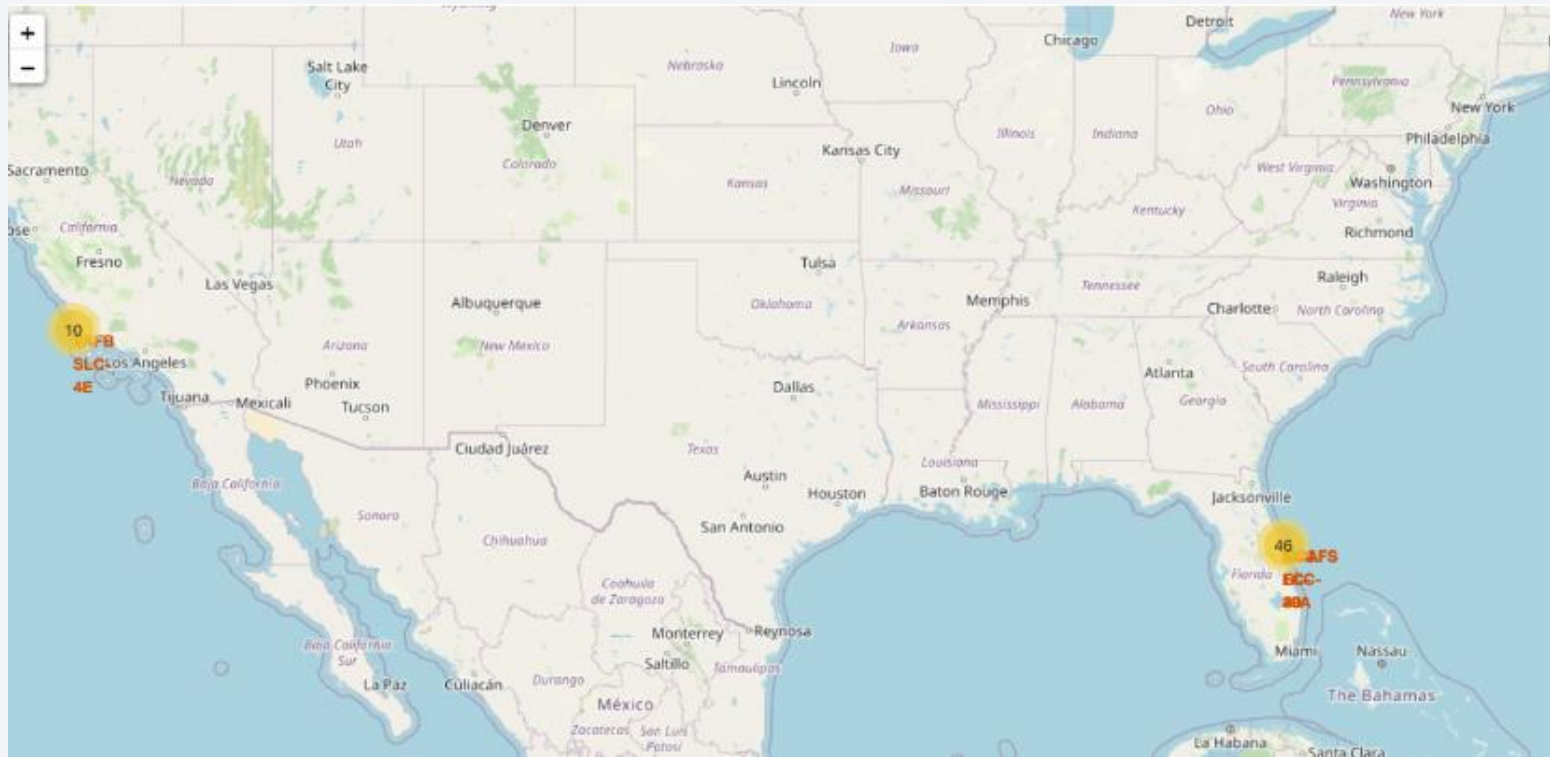
Key findings from the exploratory data analysis are as follows:

- Space X has utilized four distinct launch sites.
- Initial launches were directed towards both Space X's own initiatives and NASA projects.
- On average, the F9 v1.1 booster carried a payload of 2,928 kg.
- The first successful landing outcome was achieved in 2015, five years after the inaugural launch.
- Numerous iterations of the Falcon 9 booster achieved successful landings on drone ships with payloads surpassing the established average.
- Nearly a complete success rate was observed for mission outcomes.
- In 2015, two specific booster versions, namely F9 v1.1 B1012 and F9 v1.1 B1015, experienced failed landing attempts on drone ships.
- The quality of landing outcomes demonstrated a consistent improvement as the years progressed.

Results

Through the utilization of interactive analytics, it was feasible to ascertain that launch sites are typically situated in secure locations, often in proximity to bodies of water like the sea, ensuring safety. Furthermore, these sites tend to possess robust logistical infrastructure in their vicinity.

- The majority of launch events are concentrated at launch sites located along the eastern coastline.



Results

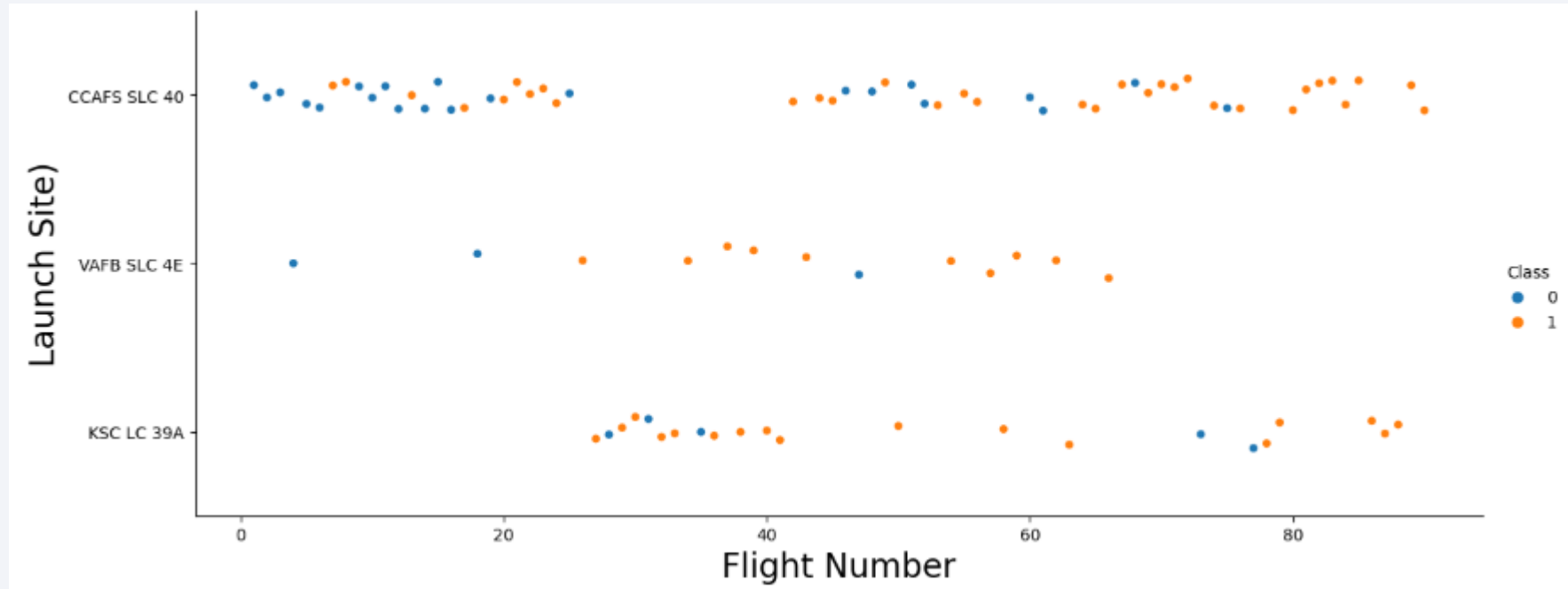
Based on the results of predictive analysis, the support vector machine emerged as the optimal model for forecasting successful landings. This model exhibited an accuracy exceeding 83.33% for overall predictions.

The background of the slide is an abstract composition. It features a dark blue base color. Overlaid on this are numerous diagonal streaks in shades of red and cyan. A faint, light blue grid pattern is also visible, particularly in the lower half of the image. The overall effect is dynamic and technological.

Section 2

Insights drawn from EDA

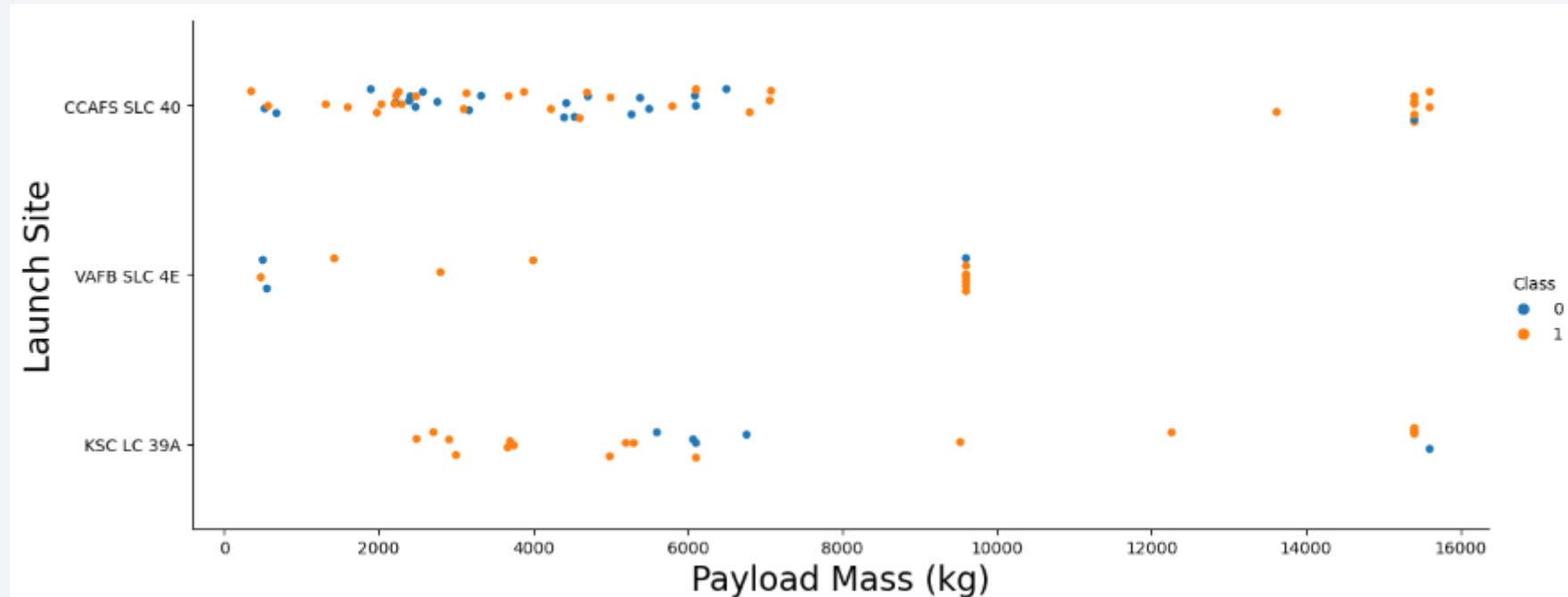
Flight Number vs. Launch Site



Referring to the depicted plot, it becomes evident that the current prime launch site is CCAFS SLC 40, evidenced by a notable proportion of recent launches yielding successful outcomes.

- The second-highest performing site is VAFB SLC 4E, followed by KSC LC 39A in third position.
- Noteworthy as well is the progressive enhancement in the overall success rate over the course of time.

Payload vs. Launch Site



- Payloads exceeding 9,000 kg exhibit a notably high rate of success.
- Payloads surpassing the 12,000 kg threshold appear feasible primarily at the CCAFS SLC 40 and KSC LC 39A launch sites.

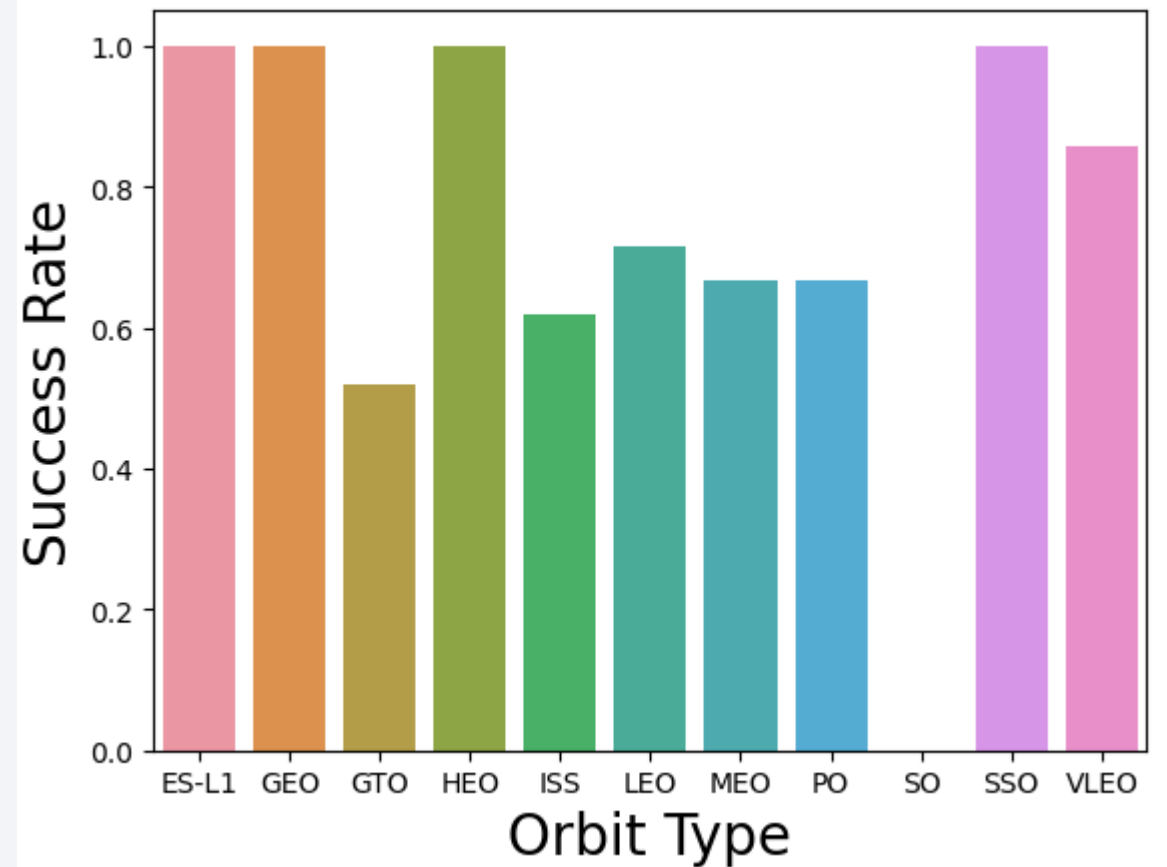
Success Rate vs. Orbit Type

The biggest success rates happens to orbits:

- ES-L1
- GEO
- HEO
- SSO

They are followed by:

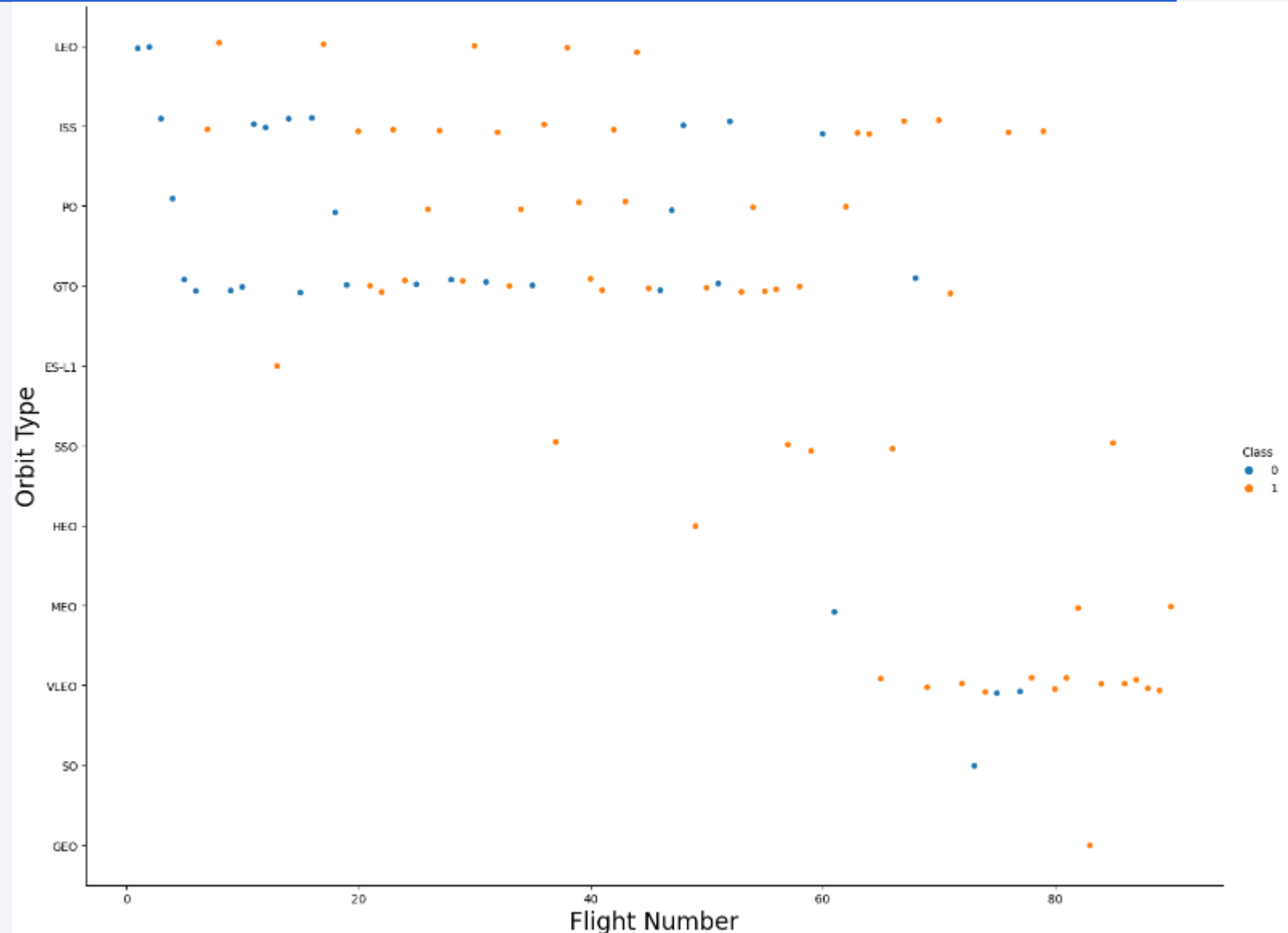
- VLEO (above 80%) and
- LFO (above 70%).



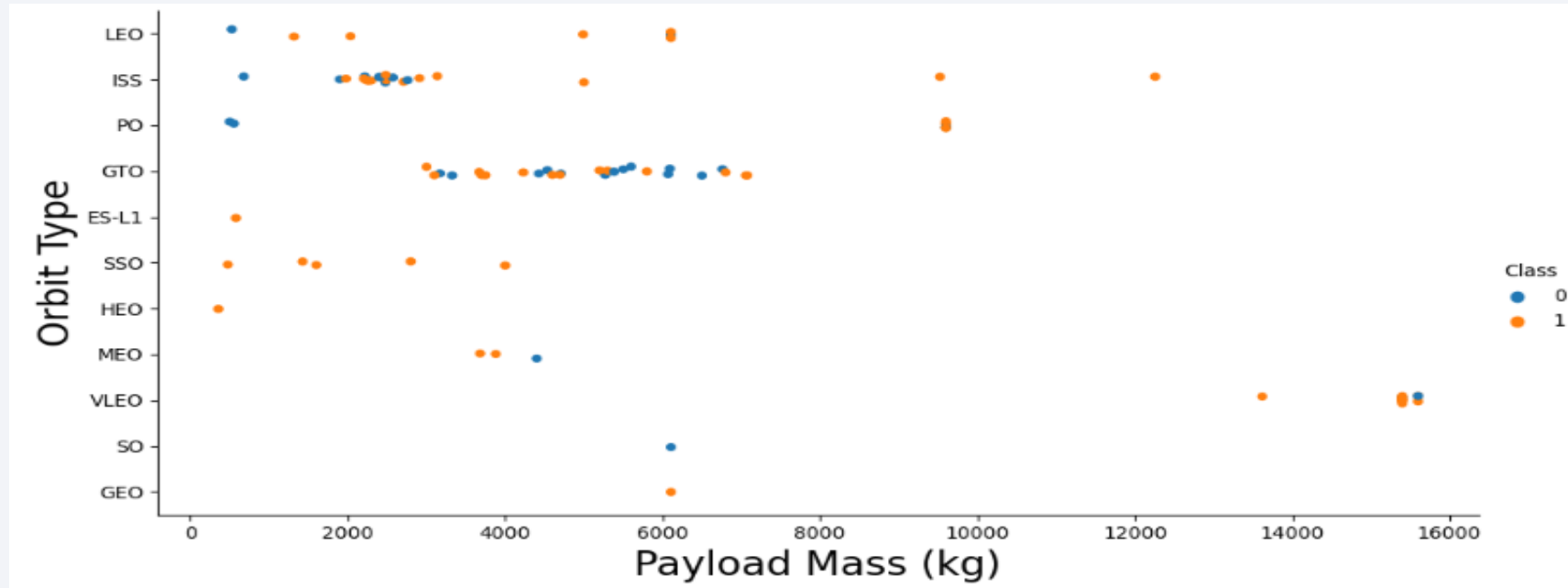
Flight Number vs. Orbit Type

Evidently, there has been a consistent improvement in success rates across all orbital categories over the passage of time.

- Notably, the Very Low Earth Orbit (VLEO) presents a promising emerging business prospect, as indicated by its recent surge in launch frequency.



Payload vs. Orbit Type

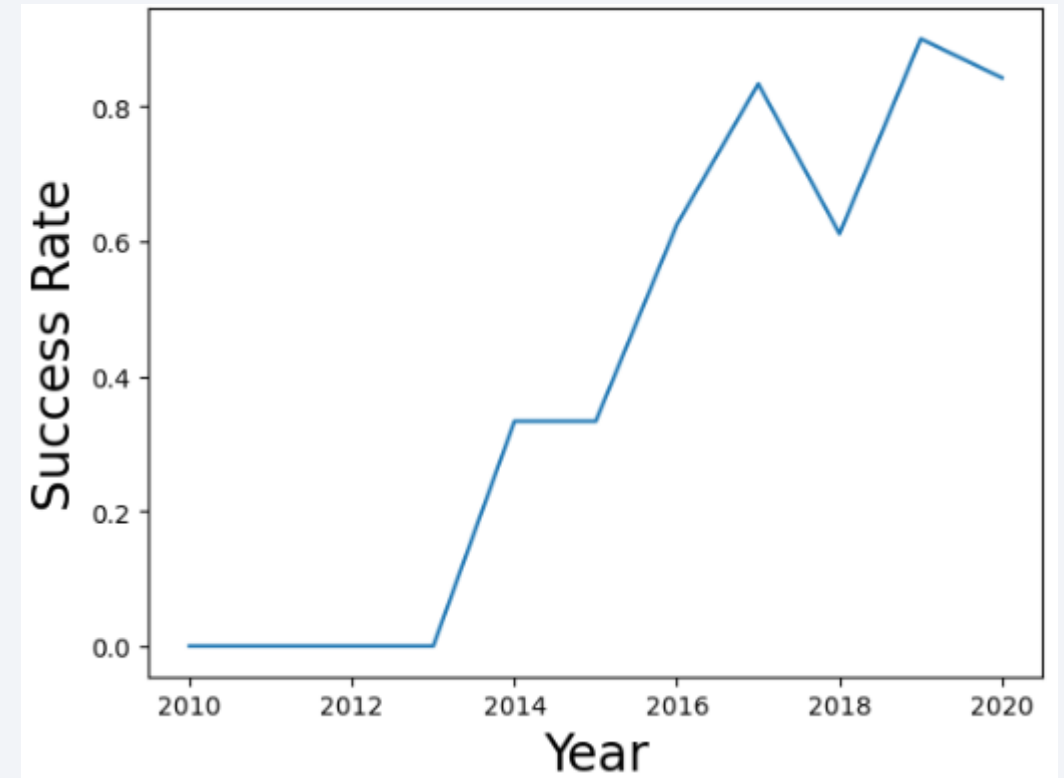


Observationally, there seems to be no discernible correlation between payload size and the success rate specifically for orbits categorized as Geostationary Transfer Orbit (GTO).

- The International Space Station (ISS) orbit demonstrates the broadest spectrum of payload capacities, coupled with a commendable success rate.
- Notably, launches targeting orbits categorized as Sun-Synchronous Orbit (SO) and Geostationary Orbit (GEO) appear to be comparatively infrequent.

Launch Success Yearly Trend

- Commencing from 2013, a discernible upward trend in success rates became evident, extending through the year 2020.
- The initial three years appeared to constitute a phase of calibration and technological refinement, signifying a period marked by adjustments and advancements in the realm of technology.



All Launch Site Names

From the data, we can see there are 4 launch sites:

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

They are extracted by selecting distinct occurrences of “launch_site” values from the available data.

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

In [8]: %sql SELECT distinct(LAUNCH_SITE) FROM SPACEXTBL;

* 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'

- We can see the 5 records available for launch sites beginning with “CCA”:

```
In [9]: %sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

* sqlite:///my_data1.db
Done.

```
Out[9]:
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS_KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

The total payload carried by boosters is extracted as follows:

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

In [10]: %sql SELECT SUM(PAYLOAD_MASS_KG_) FROM SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)';

* sqlite:///my_data1.db
Done.
Out[10]: SUM(PAYLOAD_MASS_KG_)
         45596
```

It is extracted by summing all payloads.

Average Payload Mass by F9 v1.1

Average payload mass carried by F9 v1.1 is:

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

In [11]:

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

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

```
Done.
```

Out[11]:

```
AVG(PAYLOAD_MASS_KG_)
```

```
2928.4
```

We have filtered the data by version F9 v1.1 and taken average of it.

First Successful Ground Landing Date

- The first successful landing outcome was on December 22, 2015.
- It is extracted as follows by taking minimum of data.

```
In [12]: %sql SELECT MIN(DATE) AS First_Successful_Landing_Date FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (ground pad)';

* sqlite:///my_data1.db
Done.

Out[12]: First_Successful_Landing_Date
          2015-12-22
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- The successful drone ship landing with payload between 4000 and 6000 are:

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

In [13]: %sql SELECT PAYLOAD FROM SPACEXTBL WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000

* sqlite:///my_data1.db
Done.

Out[13]:
```

Payload
JCSAT-14
JCSAT-16
SES-10
SES-11 / EchoStar 105

- They are obtained by selecting distinct booster versions and putting required filters

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- Present your que

```
In [14]: %sql SELECT MISSION_OUTCOME, COUNT(*) as total_number FROM SPACEXTBL GROUP BY MISSION_OUTCOME;
```

* sqlite:///my_data1.db
Done.

```
Out[14]:
```

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

Boosters Carried Maximum Payload

- Boosters which carried maximum loads are:

```
In [15]: %sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);

* sqlite:///my_data1.db
Done.

Out[15]: 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

- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

Booster Version	Launch Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

- The list above has the only two occurrences.

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

- All landing outcomes between date 2010-06-04 and 2017-03-20 are:

Landing Outcome	Occurrences
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

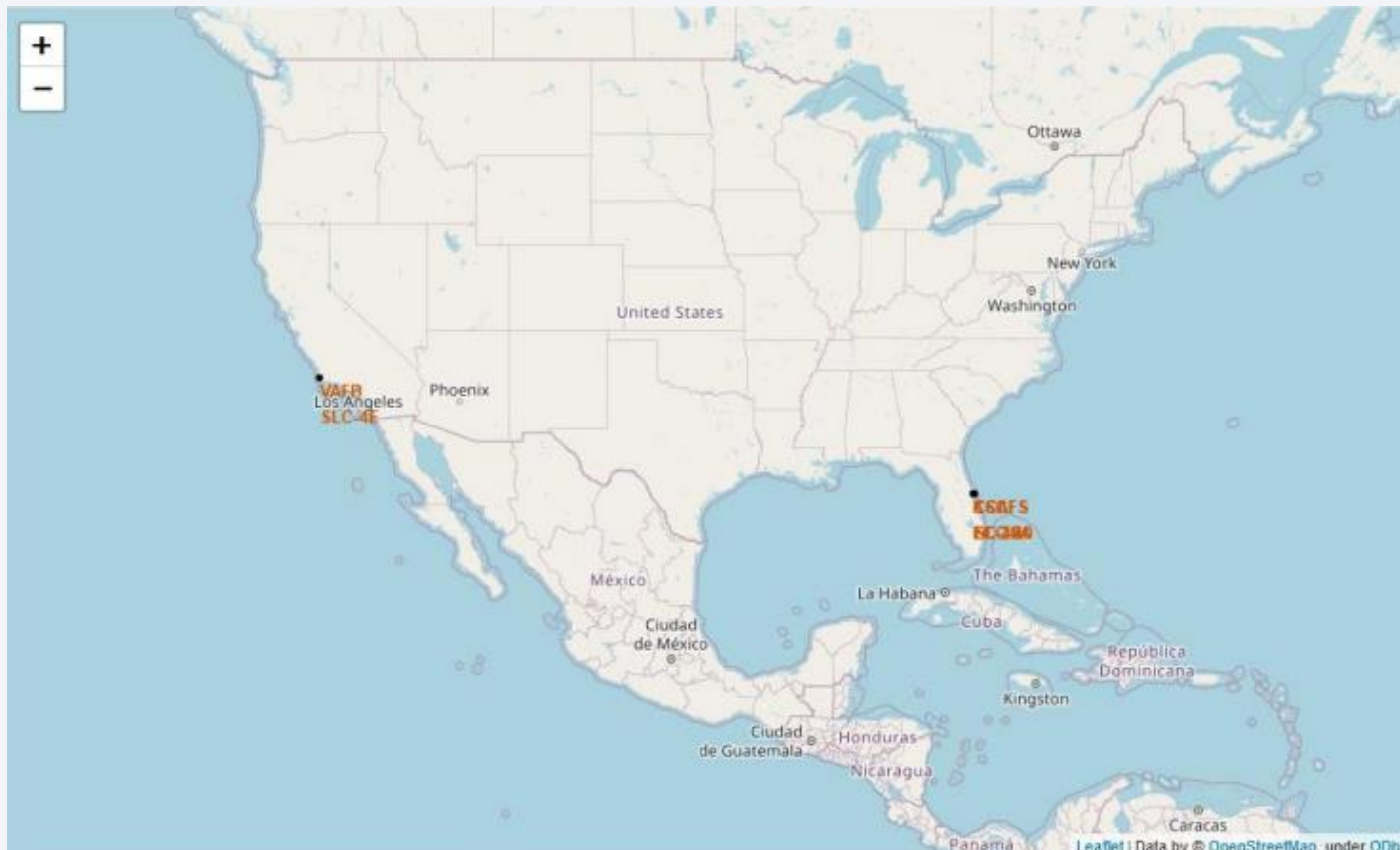
A satellite view of Earth from space, showing the curvature of the planet and city lights at night. The background is a deep blue gradient.

Section 3

Launch Sites Proximities Analysis

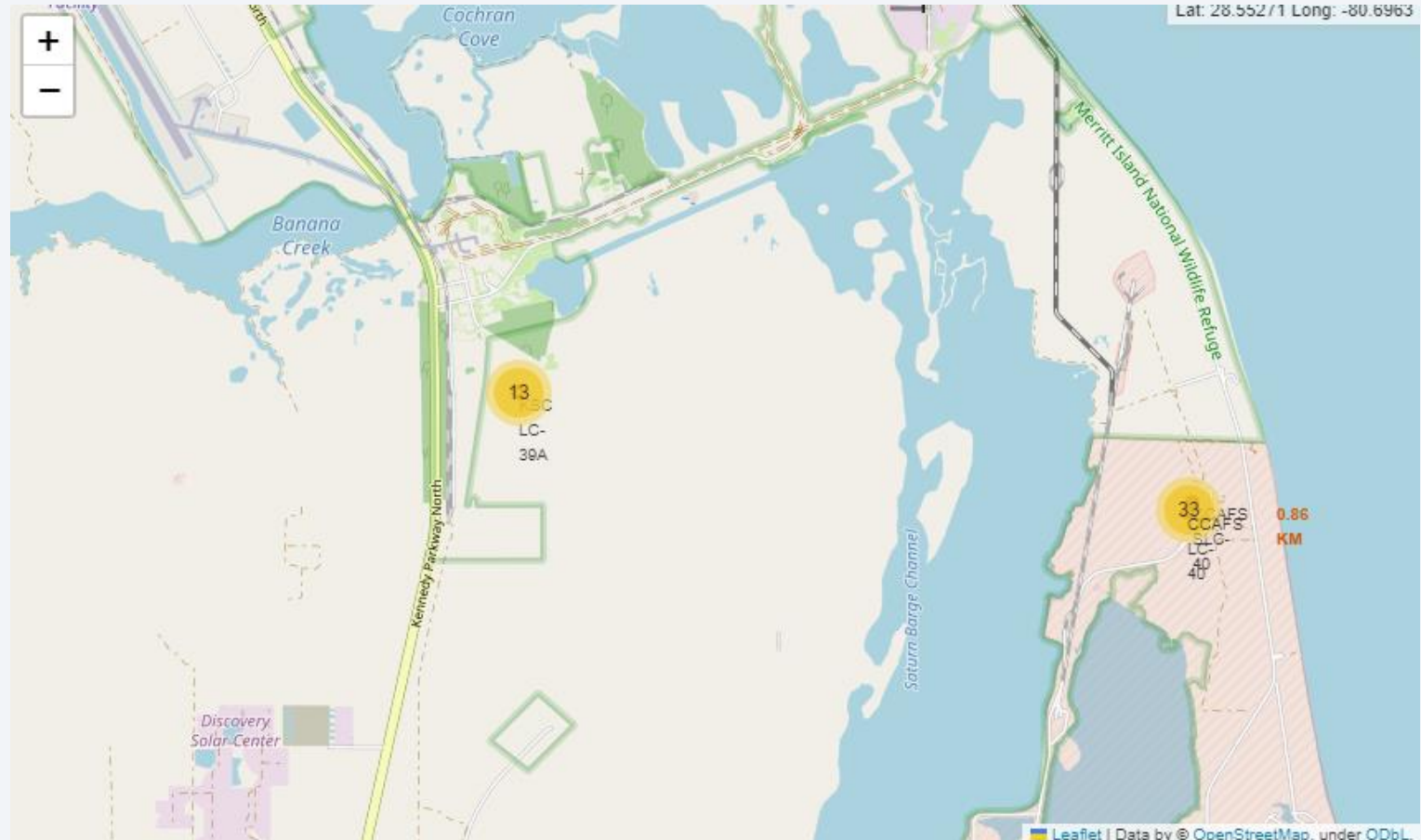
All Launch Sites

- It can be observed that launch sites are near sea, probably for safety causes, but not too far from roads and railroads



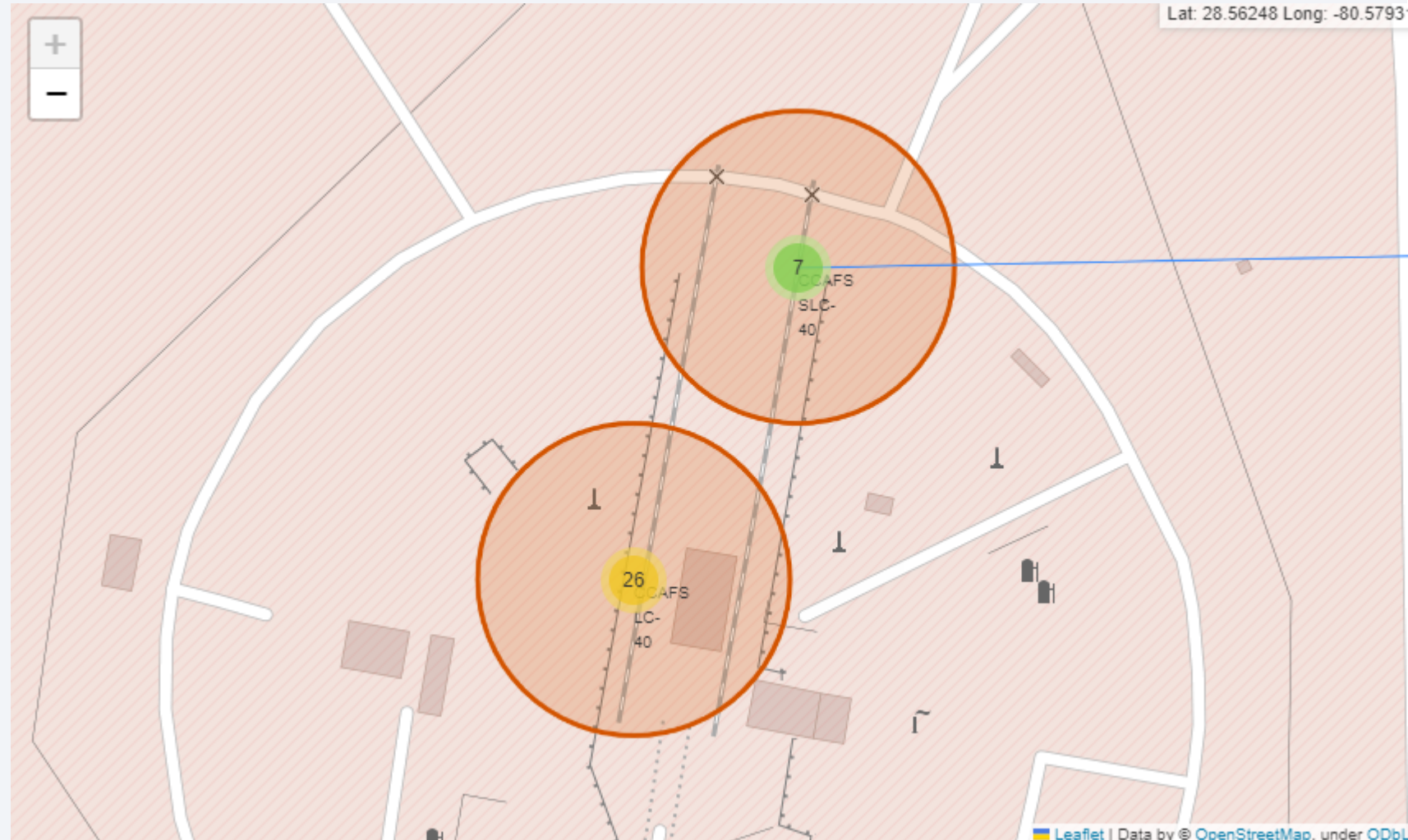
Launch Outcomes by Site

- After browsing a particular launch site, we can see that there are 33 and 13 different launches from the site.



Proximity of Launch Sites

- The launch sites have good logistics aspects as it has good connectivity to road and railroads.

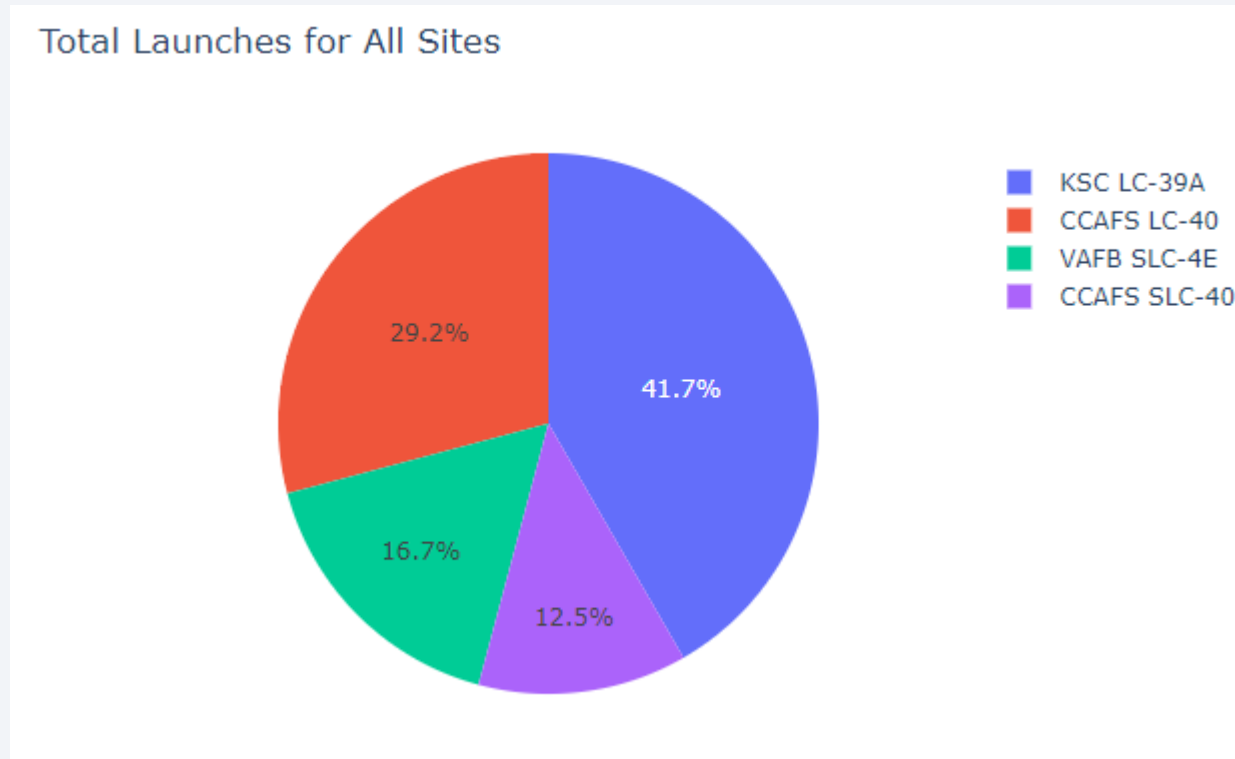




Section 4

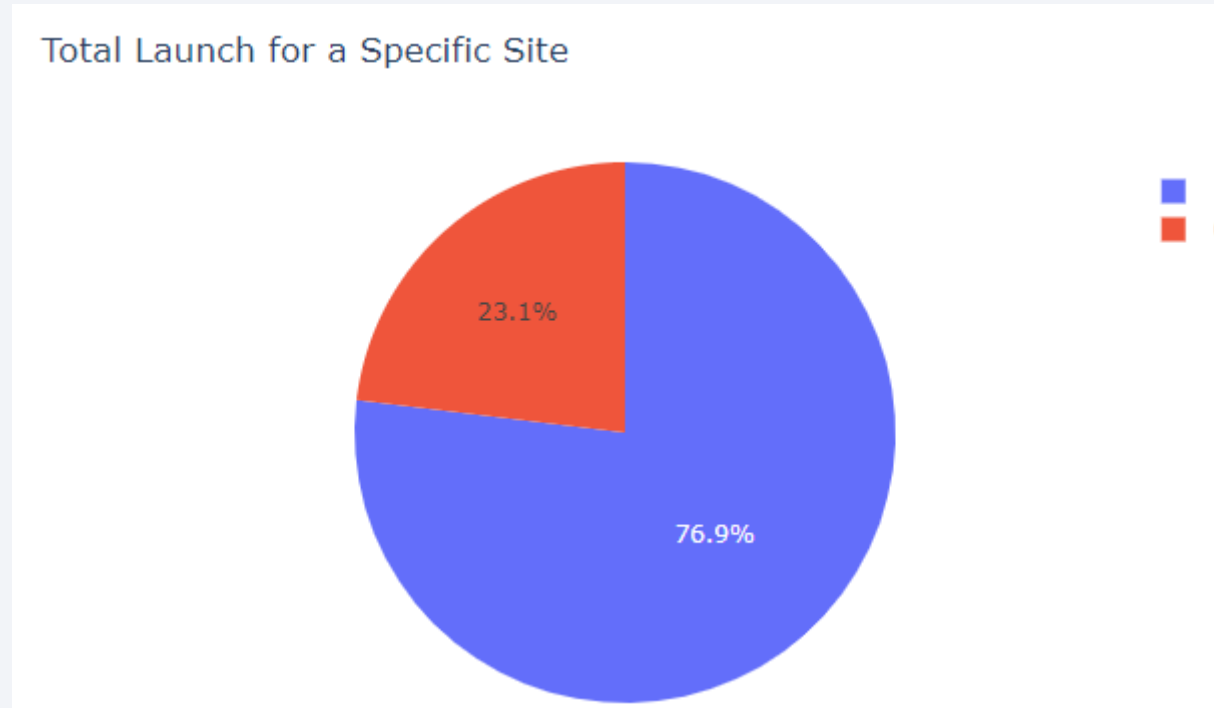
Build a Dashboard with Plotly Dash

Successful Launches by Launch Sites



- It can be seen that the launch sites are an important factor for the success of missions.

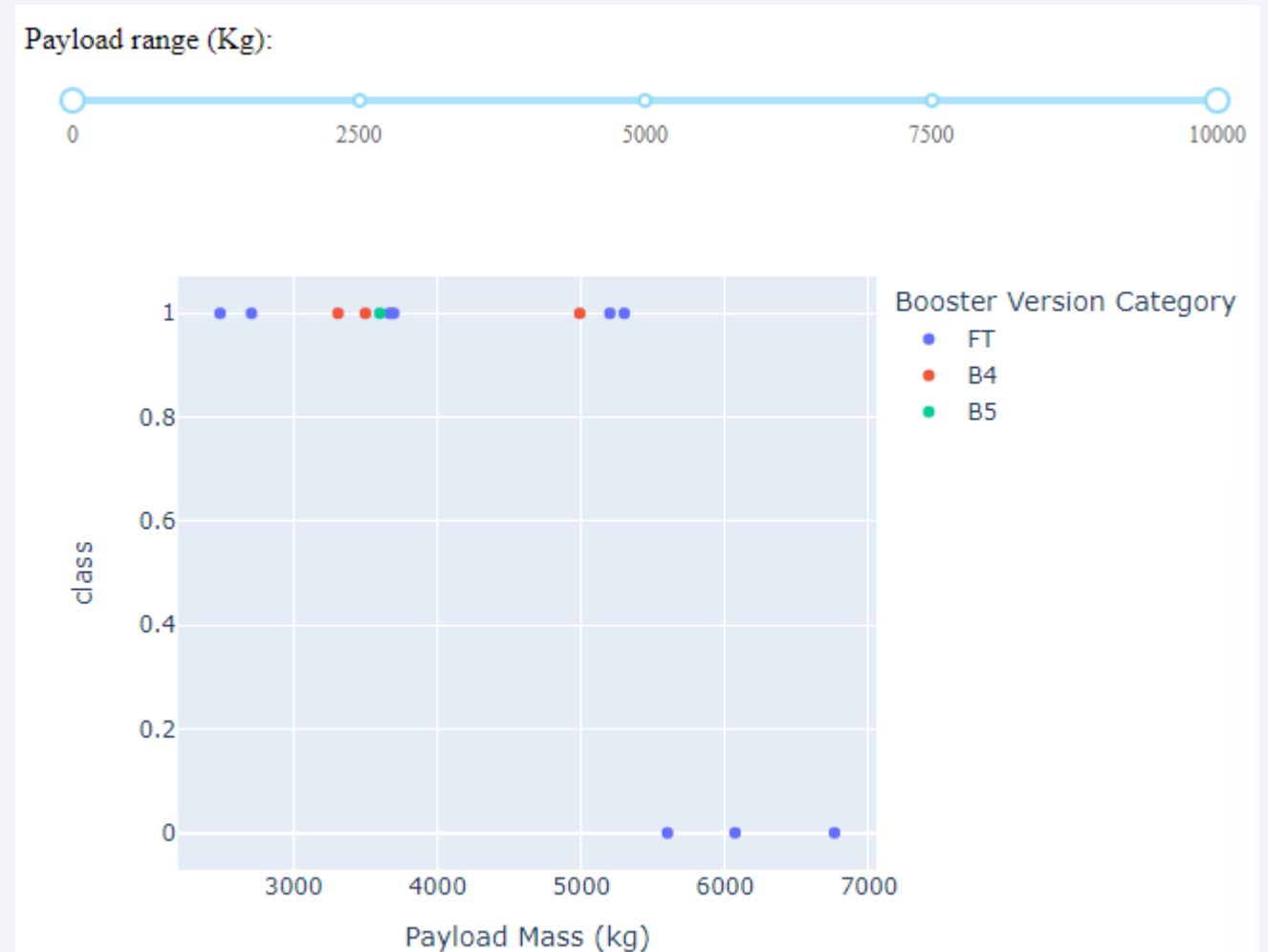
Launch Success Ratio for KSC LC-39A



The KSC LC-39A has the highest success ration of 76.9% missions.

Payload Vs. Launch Outcome

- Payloads under 5500 are generally successful
- FT boosters used for payloads above 5500 is generally not successful.

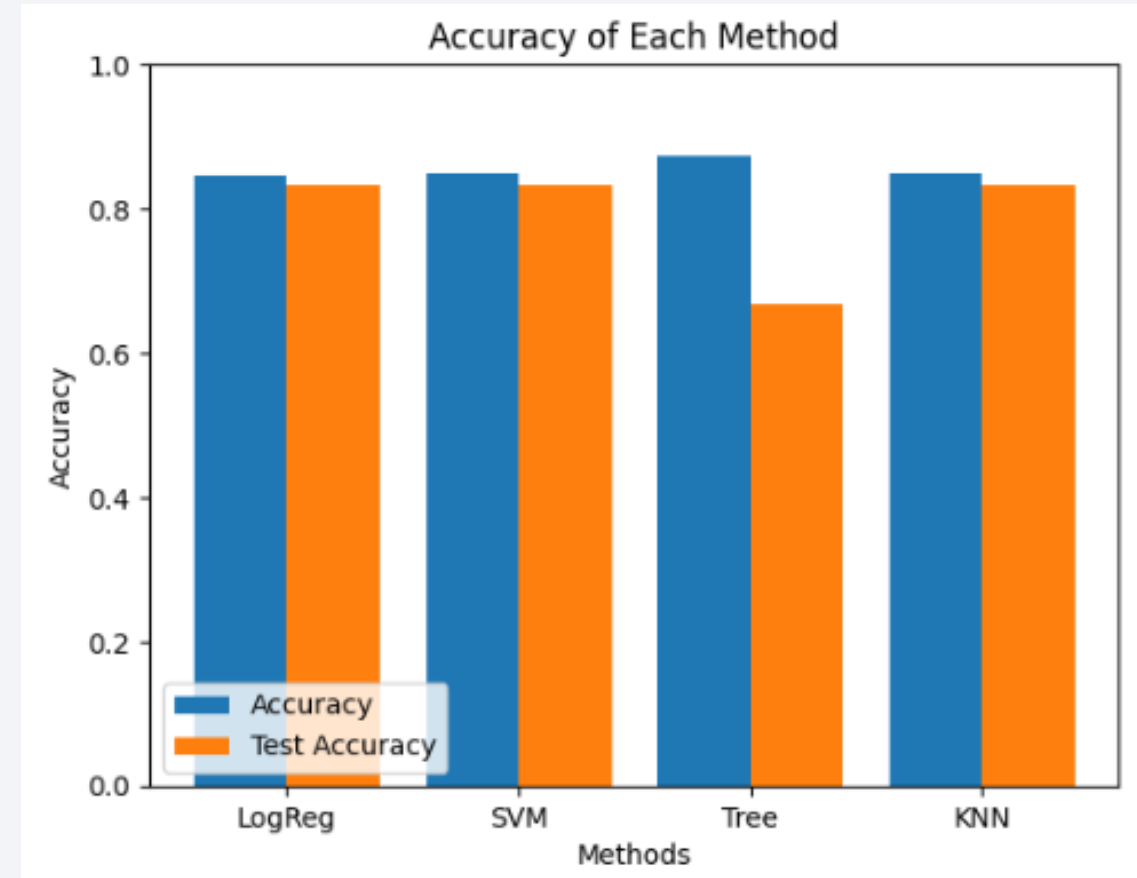


Section 5

Predictive Analysis (Classification)

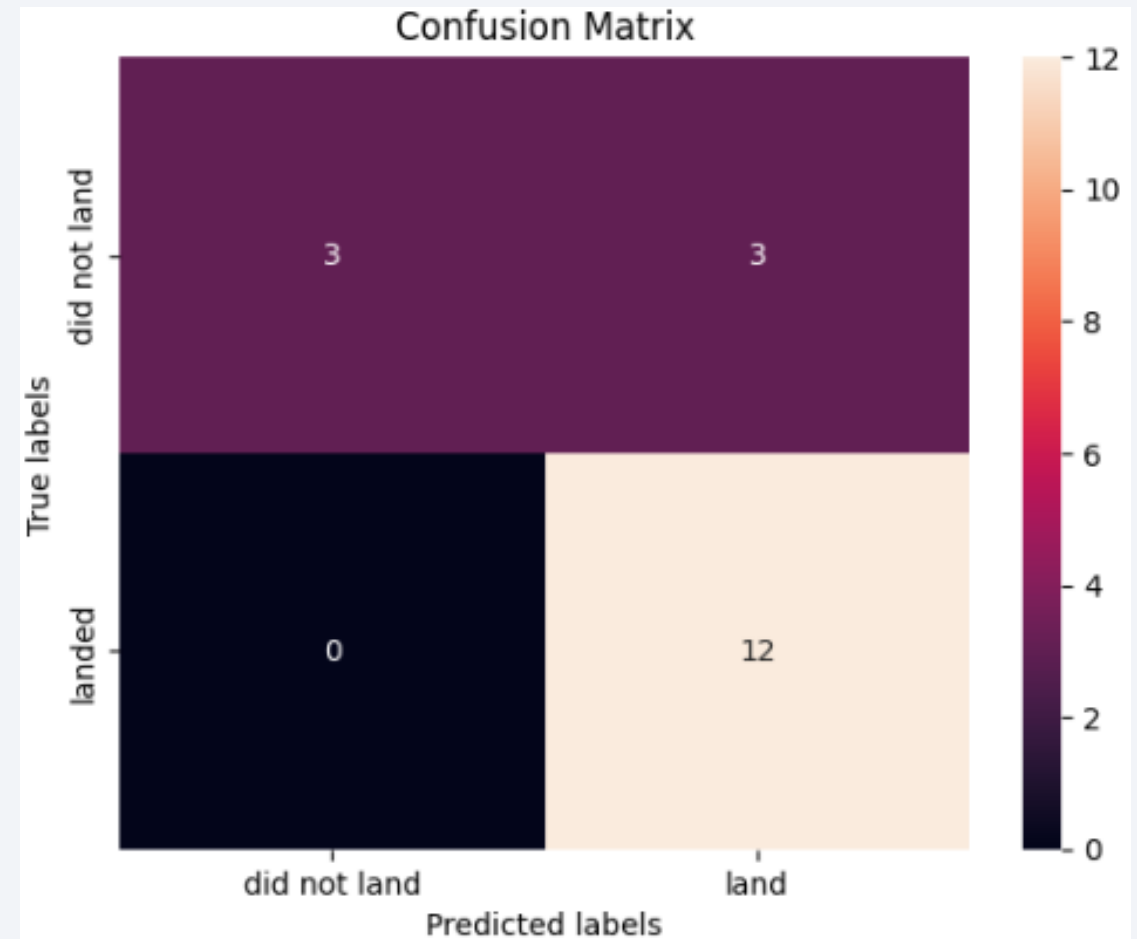
Classification Accuracy

- We tested 4 classification models and their training and test accuracies are plotted beside:
- We have two models with highest accuracies (SVM and KNN).



Confusion Matrix (SVM & KNN)

- Confusion matrix of SVM and KNN are similar. It proves the accuracy by showing the big numbers of true positives and true negatives compared to the false positives and false negatives.



Conclusions

Multiple data sources underwent scrutiny, progressively honing the derived conclusions.

- The preeminent launch site identified is KSC LC-39A.
- Launch missions with payloads surpassing 7,000 kg appear to carry lower inherent risks.
- Despite a predominant prevalence of successful mission outcomes, the trajectory of successful landing results exhibits an observable augmentation over time, a phenomenon potentially aligned with advancements in processes and rocket technology.
- The SVM and KNN Classifier model holds potential utility in forecasting successful landings, subsequently fostering enhanced profitability.

Appendix

- GitHub URL for complete codes: <https://github.com/Sourabh-khatrii/Applied-Data-Science-Capstone>

Thank you!

