IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Akshay Harshad Lad
9th August 2024

Outline

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

Github Link: https://github.com/Akshay-Lad31/IBM-Data-Science-Capstone

Executive Summary

- **Summary of methodologies**
  - Data Collection:
    - Using SpaceX API
    - Using Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis:
    - with SQL
    - with Data Visualization
  - Interactive Visual Analytics using Dash App & with Folium
  - Machine Learning Prediction

- **Summary of all results**
  - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive Analytics result

Github Link: https://github.com/Akshay-Lad31/IBM-Data-Science-Capstone

Introduction

- **Project background and context**

SpaceX prominently advertises its Falcon 9 rocket launches at a competitive cost of $62 million, in stark contrast to other providers, whose launch services can exceed $165 million. This significant cost advantage primarily stems from SpaceX's ability to reuse the first stage of the rocket. Therefore, accurately predicting whether the first stage will successfully land is essential for assessing the overall launch cost. This information is particularly valuable for any alternate company seeking to bid against SpaceX for rocket launch services. The objective of this project is to develop a robust machine learning pipeline capable of predicting the successful landing of the Falcon 9's first stage.

- **Problems you want to find answers**
  - What factors determine successful landing of rockets?
  - The interplay among various features that influences the success rate of a successful landing.
  - What operating conditions need to be in place for ensuring success in the landing program.

Github Link: https://github.com/Akshay-Lad31/IBM-Data-Science-Capstone

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - **Data of Space X was obtained from two sources:**
    - Space X API (https://api.spacexdata.com/v4/rockets/)
    - Webscraping Wikipedia page (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_)
- Perform data wrangling
  - For the collected data, we summarized the results and analysed the features and the outcome data was labelled by creating landing outcome labels.

Github Link: https://github.com/Akshay-Lad31/IBM-Data-Science-Capstone

Methodology

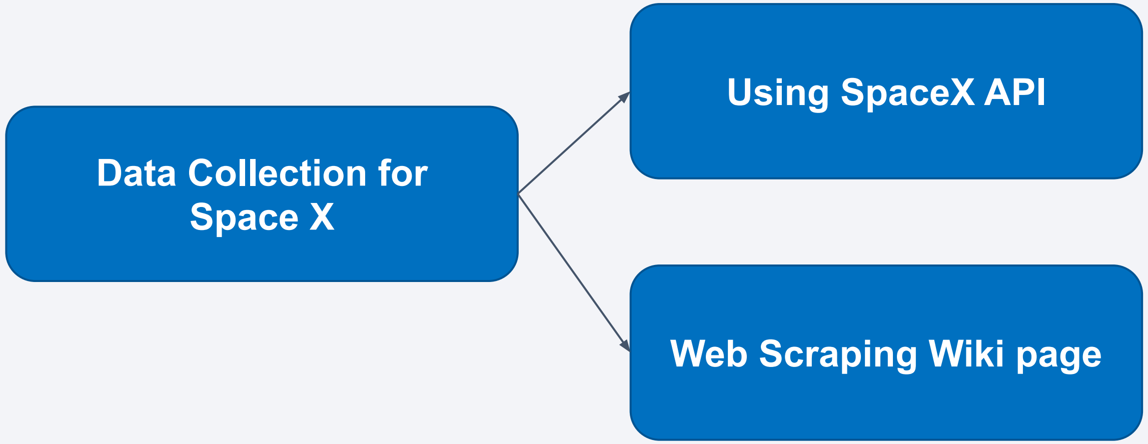
Executive Summary

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Input data was collected, normalised, then divided in test and training data sets.
  - Then the data was evaluated using 4 different classification models on training data and validated on test data
  - Finally, for each model, accuracy and CM was plotted to find the best working model with highest accuracy.

Github Link: https://github.com/Akshay-Lad31/IBM-Data-Science-Capstone

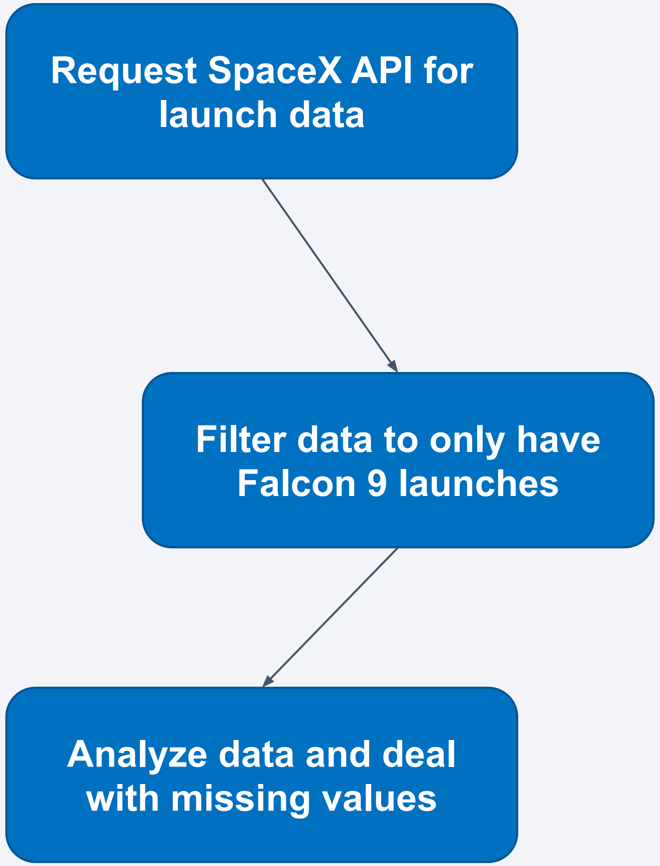
Data Collection

- Data sets were collected from Space X API (https://api.spacexdata.com/v4/rockets/ 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 offers a public API from where data can be obtained and then used.
- Fetching of data using the API and processing done on it is explained in the flowchart
- **Github Notebook URL: https://github.com/Akshay-Lad31/IBM-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb**

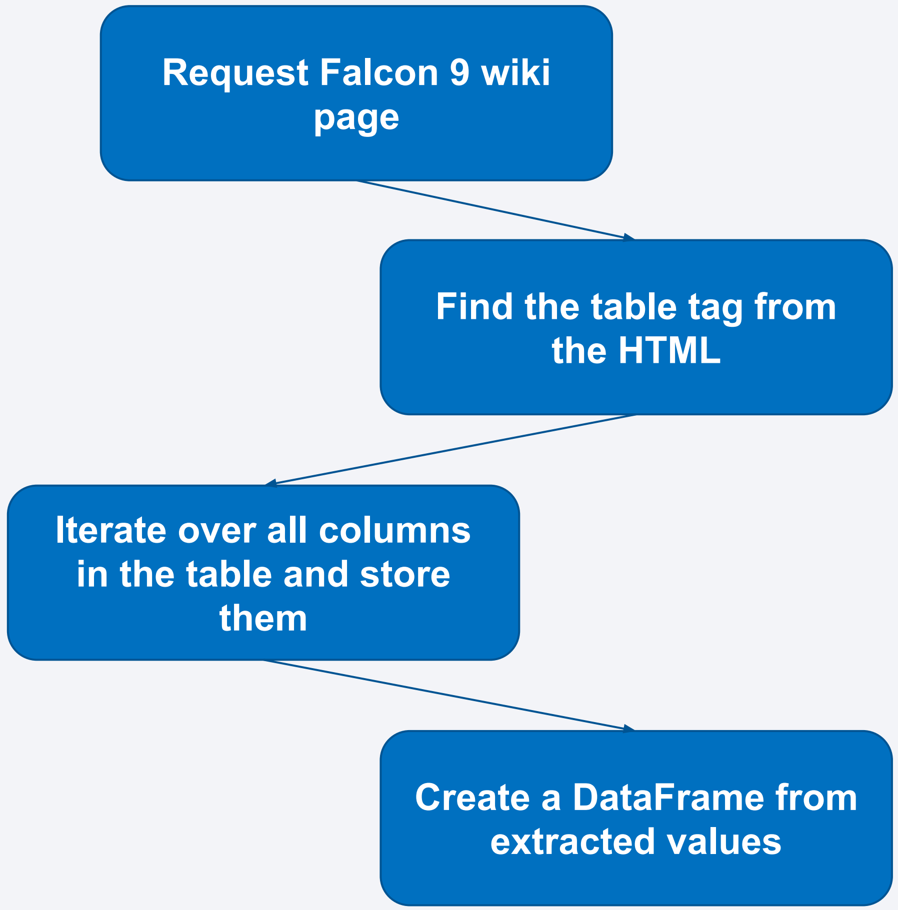


Data Collection - Scraping

- SpaceX data can also be obtained from Wikipedia page
- We used beautifulSoup library to perform web scraping

- Web scraping of Falcon 9 wiki page is explained in the flowchart

- **Github Notebook URL: https://github.com/Akshay-Lad31/IBM-Data-Science-Capstone/blob/main/jupyter-labs-webscraping.ipynb**



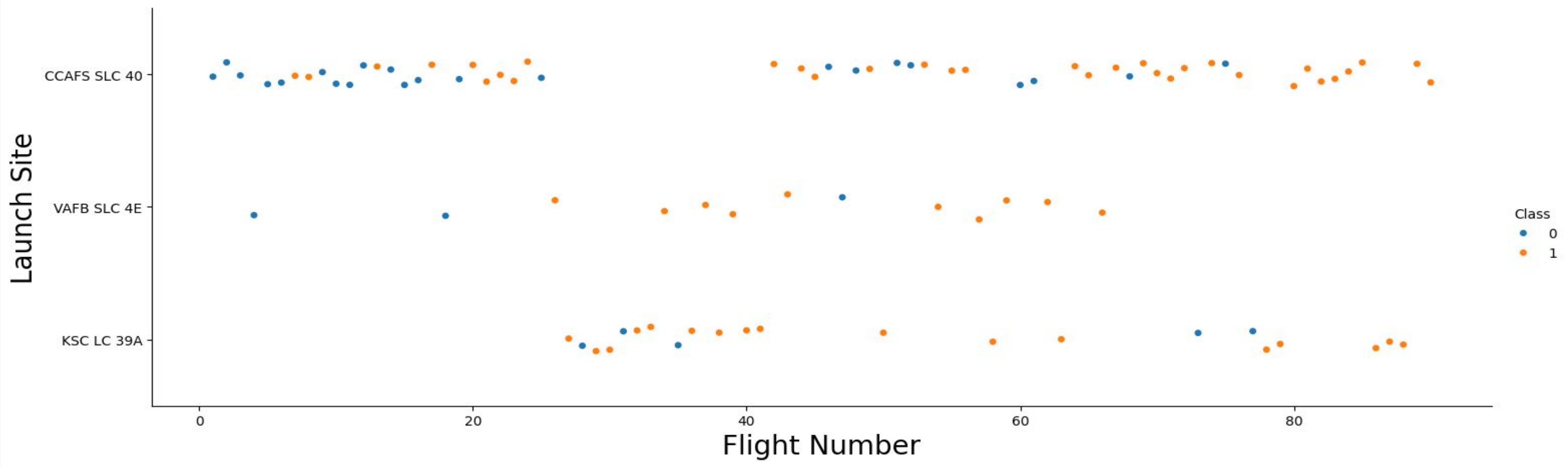
Data Wrangling

- Initially, we performed Exploratory Data Analysis (EDA) on the given dataset.
- We summaries the data by:
  - Launches per site
  - Occurrences of each orbit
  - Occurrence of missing data per orbit type
- Post this, that landing outcome labels were created from the Outcome column

Github Notebook link:
https://github.com/Akshay-Lad31/IBM-Data-Science-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- For EDA, we plotted data using scatter plots and bar plots to visualize the relationship between pair of features:
    - Payload Mass Vs Flight Number, Launch Site Vs Flight Number, Launch Site Vs Payload Mass, Orbit and Flight Number & Payload and Orbit



Github Notebook Link:
https://github.com/Akshay-Lad31/IBM-Data-Science-Capstone/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- Following SQL query were performed:
  - Names of the unique launch sites in the space mission;
  - Top 5 launch sites whose name begin with the string 'CCA’;
  - Total payload mass carried by boosters launched by NASA (CRS);
  - Average payload mass carried by booster version F9 v1.1;
  - Date when the first successful landing outcome in ground pad was achieved;
  - Names of the 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;
  - Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015;
  - Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010 06 04 and 2017 03 20.

GitHub URL:
https://github.com/Akshay-Lad31/IBM-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

We added Markers, circles, lines and marker clusters to the Folium Maps

- **Markers** indicate points like launch sites
- **Circles** indicate highlighted areas around specific coordinates, like NASA Johnson Space Center;
- **Marker clusters** indicates groups of events in each coordinate, like launches in a launch site;
- **Lines** are used to indicate distances between two coordinates.
- We also add **Mouse Position** interaction to get coordinate of location pointed by the mouse.

Github URL:
https://github.com/Akshay-Lad31/IBM-Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location.ipynb

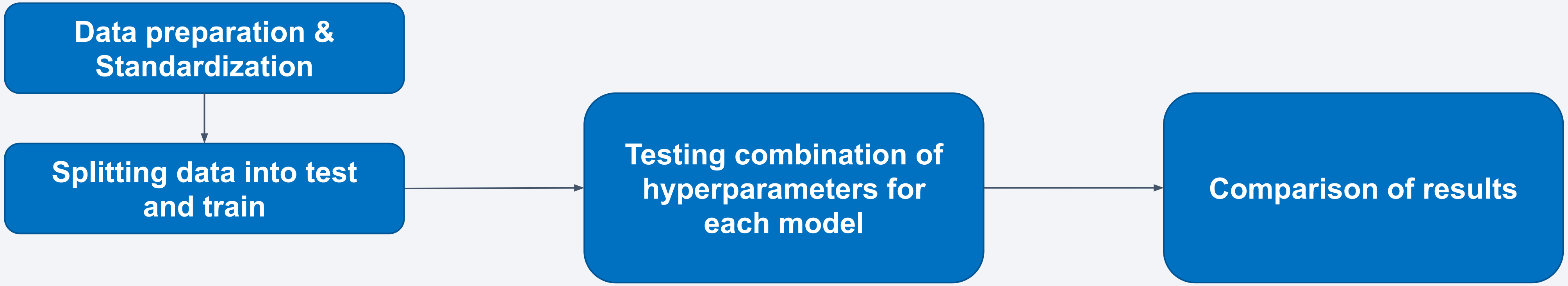
Build a Dashboard with Plotly Dash

- Following graphs, plots and interactions were used in dashboard:
    - Dropdown to select site (with default as ALL sites)
    - Slider range to select payload range
    - Pie chart showing percentage of launches by sites
    - Pie chart showing outcome of launch per site
    - Scatter plot of Payload Vs. Launch Outcome
- All these plots allowed to quickly analyze relation between payload, launch sites and find best place to launch according to the payload.

Github URL: https://github.com/Akshay-Lad31/IBM-Data-Science-Capstone/blob/main/spaceX_Dash_app.py

Predictive Analysis (Classification)

- Four classification models were used: Logistic Regression, Support Vector Machine, Decision Tree and K Nearest Neighbors.
- Process of training and evaluation of models is shown in flowchart:



Github URL:
https://github.com/Akshay-Lad31/IBM-Data-Science-Capstone/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

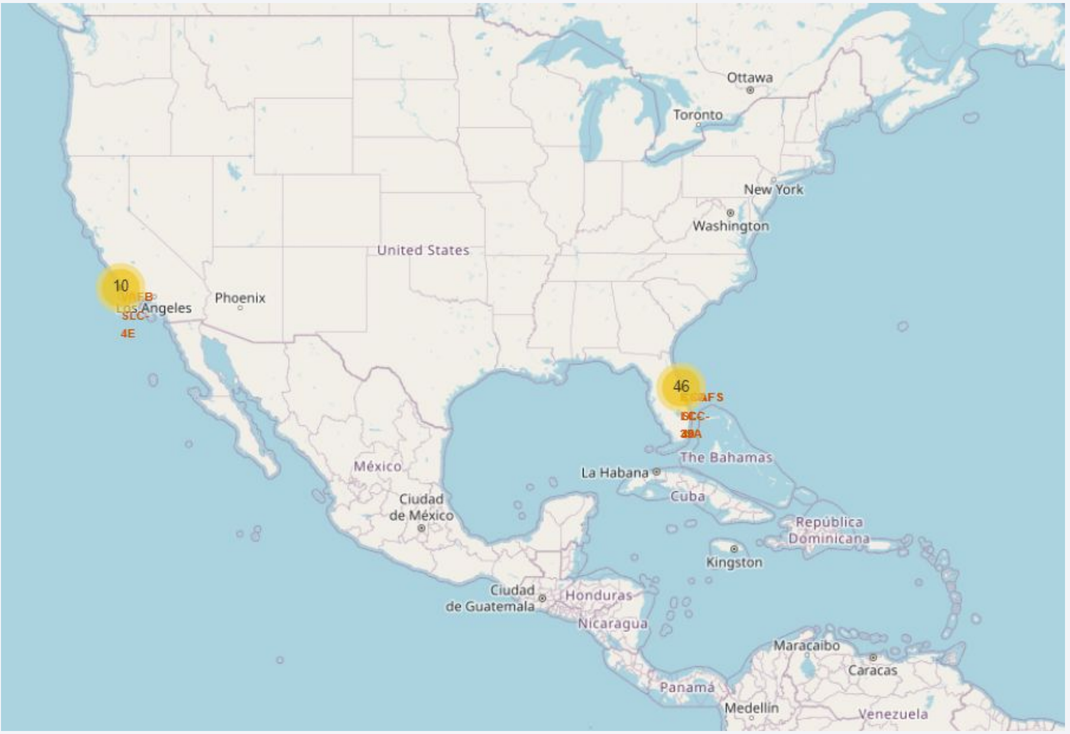
Exploratory data analysis results:

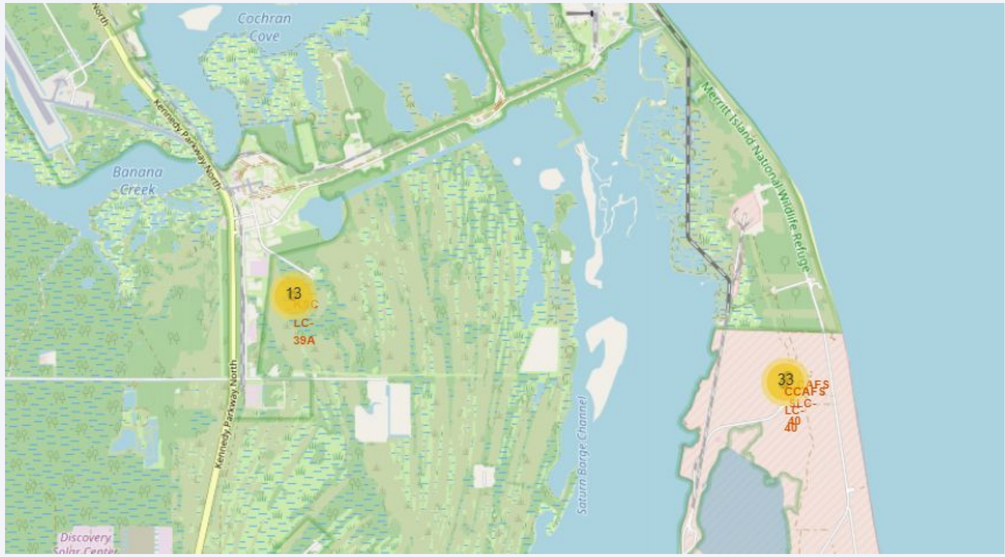
- SpaceX uses 4 different launch sites;
- The first launches were done to Space X itself and NASA;
- The average payload of F9 v1.1 booster is 2,928 kg;
- The first success landing outcome happened in 2015 fiver year after the first launch;
- Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average;
- Almost 100% of mission outcomes were successful;
- Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015;
- The number of landing outcomes became as better as years passed.

Results

Interactive Maps results:

- Using interactive analytics was possible to identify that launch sites use to be in safety places, near sea, for example and have a good logistic infrastructure around.
- Most launches happens at east coast launch sites:



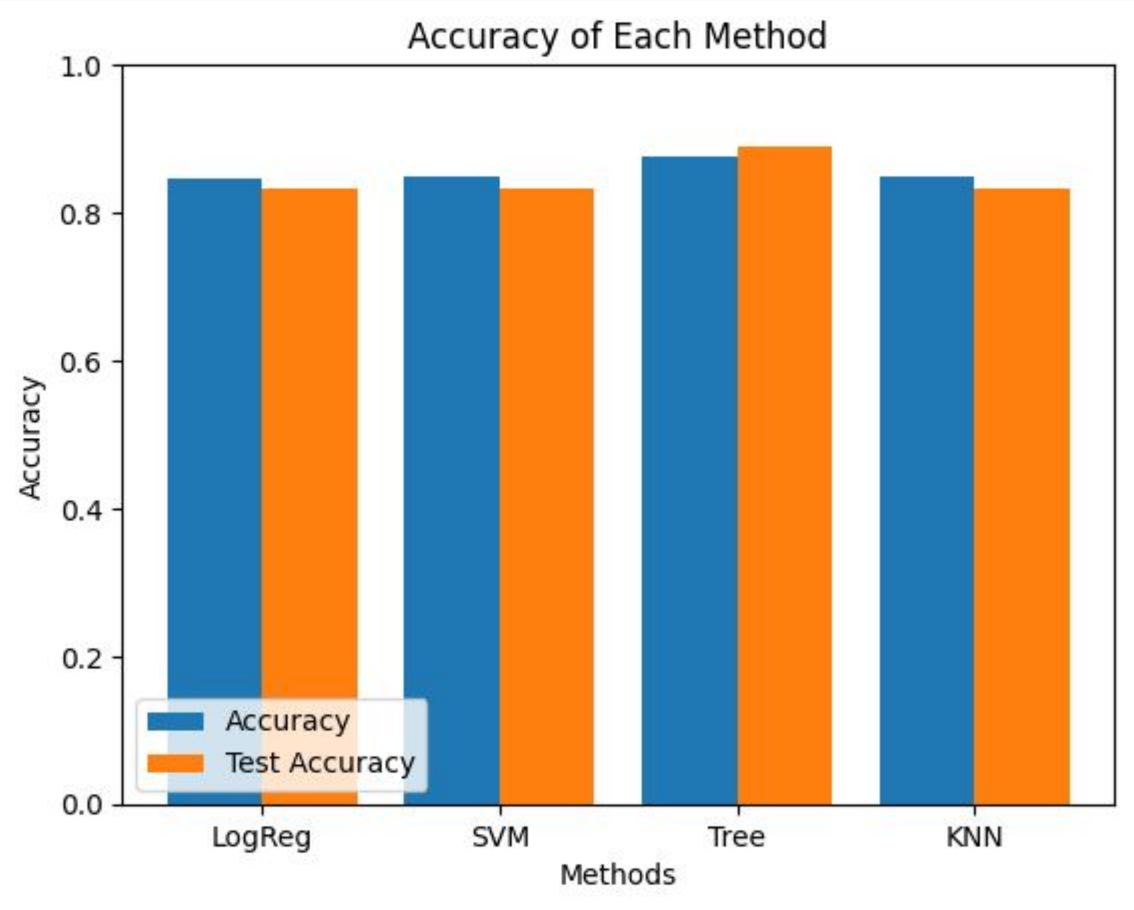


46 out of 56 launches happened on East Coast

Results

Predictive Analysis results:

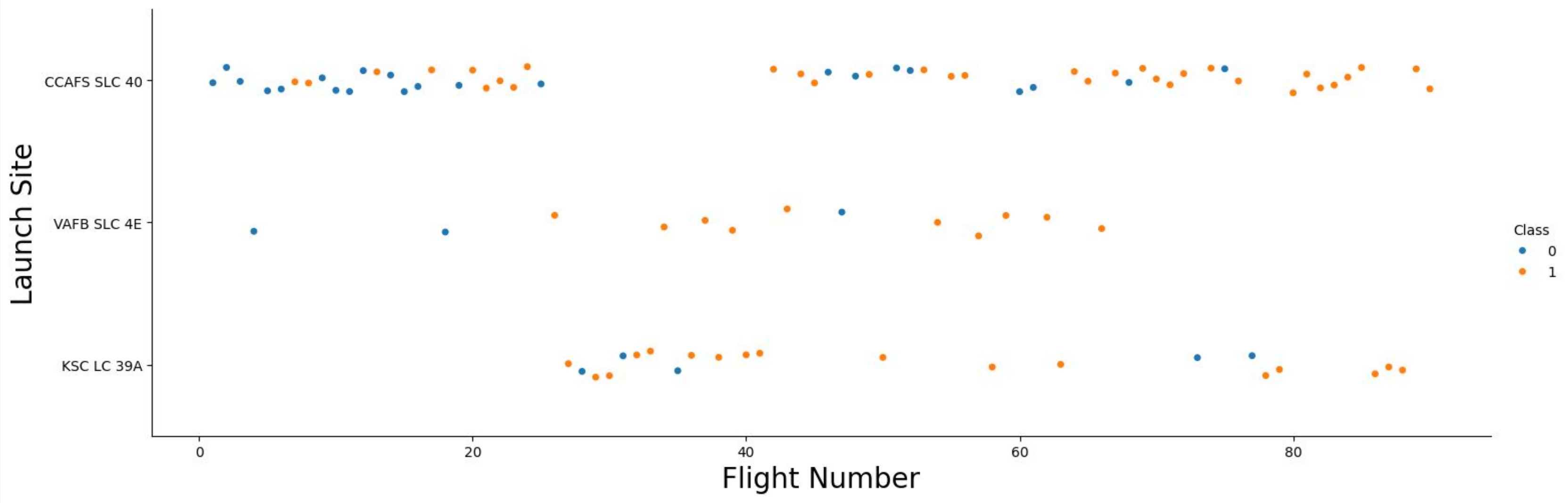
- Out of the 4 classifier, Decision Tree performed the best with accuracy of 88% over test data and 85% over train data



Section 2

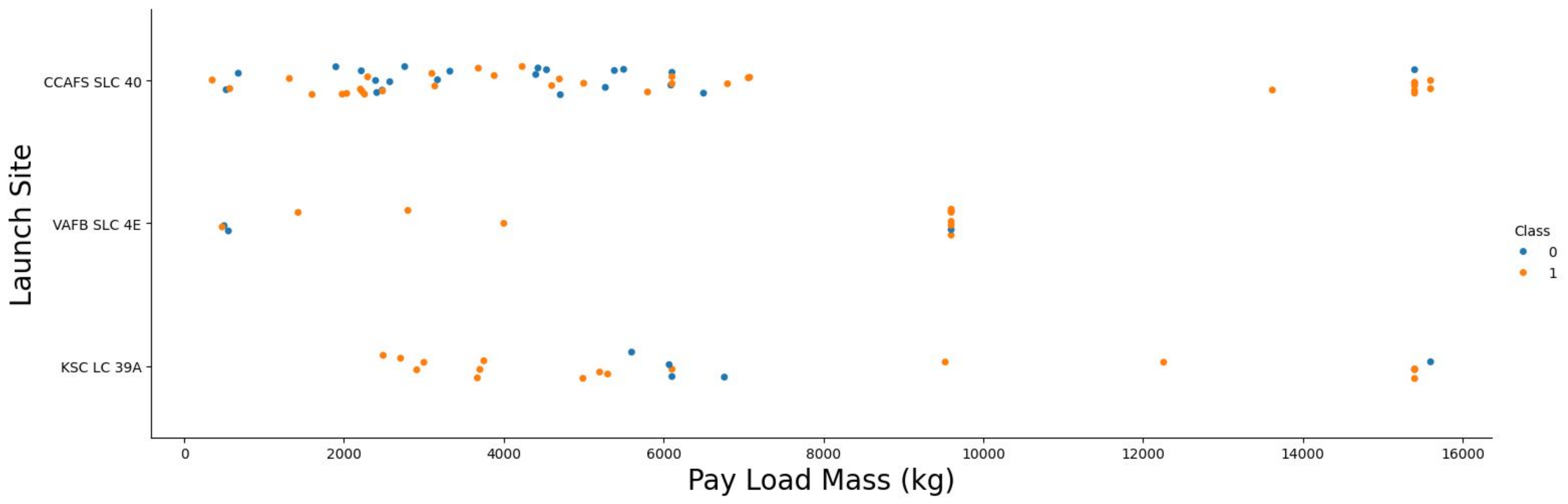
Insights drawn from EDA

Flight Number vs. Launch Site



- In recent time, best launch site is CCAF5 SLC 40, where most recent launches are successful.
- In second place we have VAFB SLC 4E & third we have KSC LC 39A.
- The general success rate of a site improves over time.

Payload vs. Launch Site

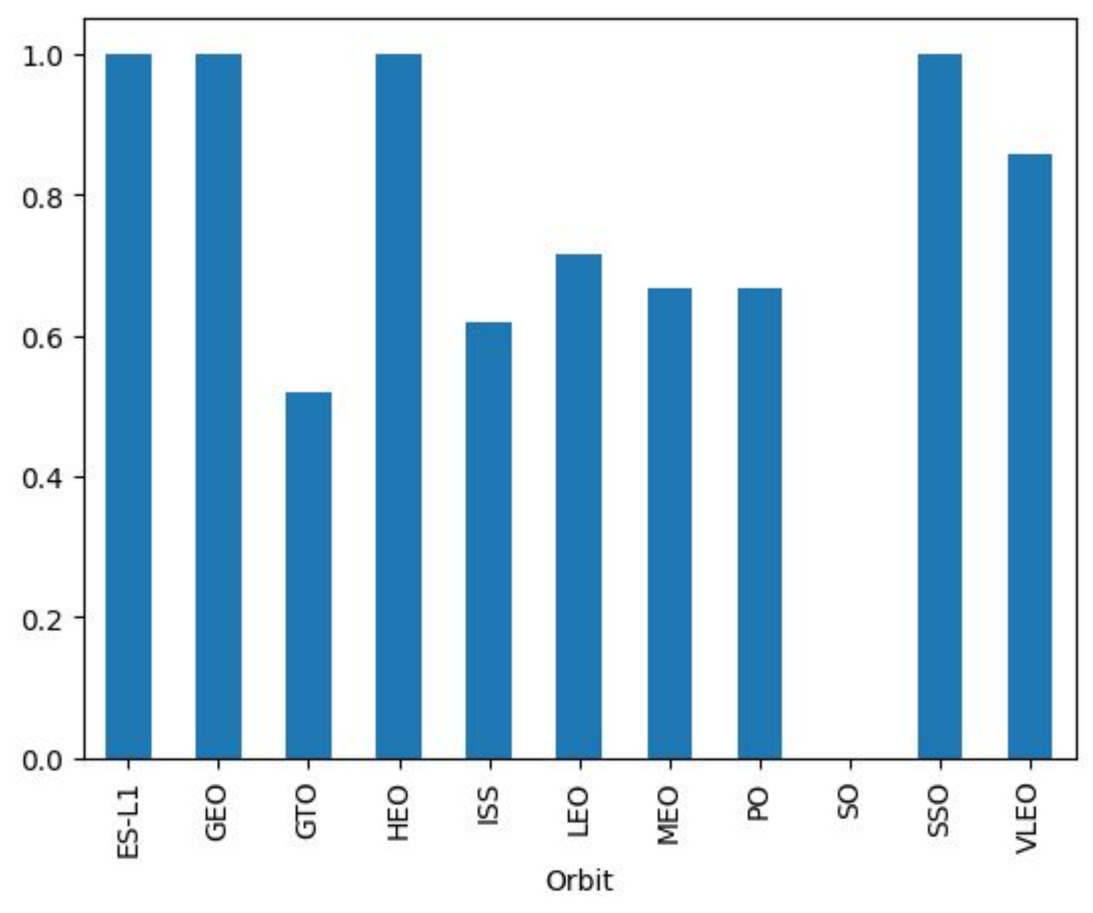


- Payloads over 9,000kg (about the weight of a school bus) have excellent success rate;
- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.

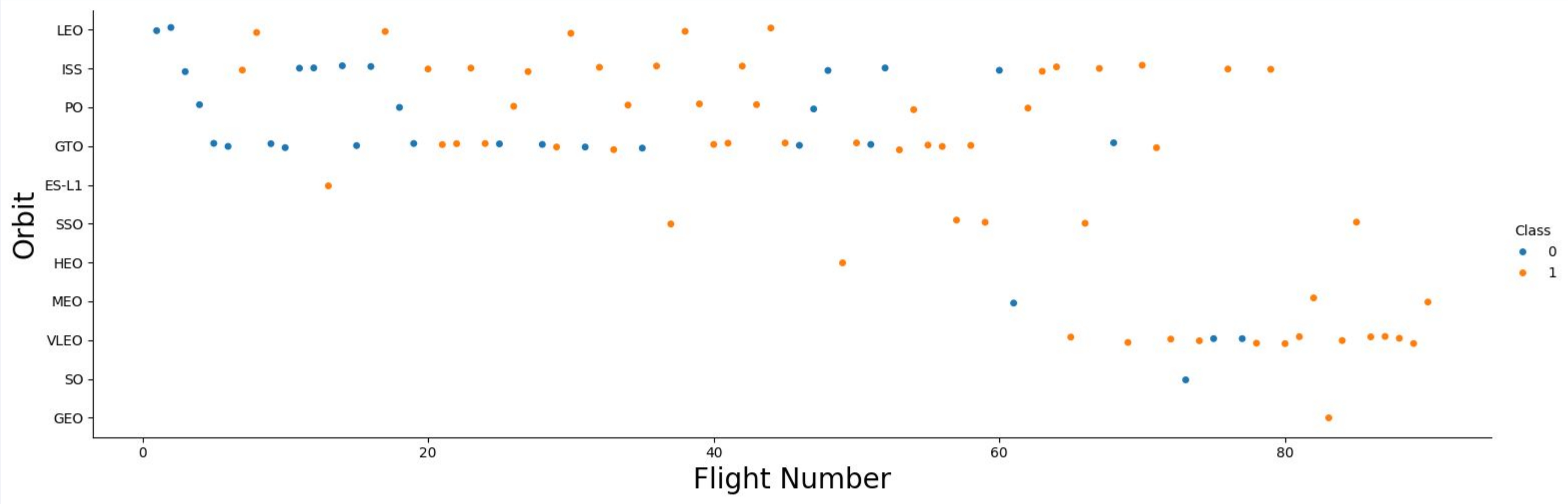
Success Rate vs. Orbit Type

- Highest success rate orbits are:
  - ES-L1
  - GEO
  - HEO
  - SSO

- Followed by: VLEO (~80%) & LFO (> 70%)

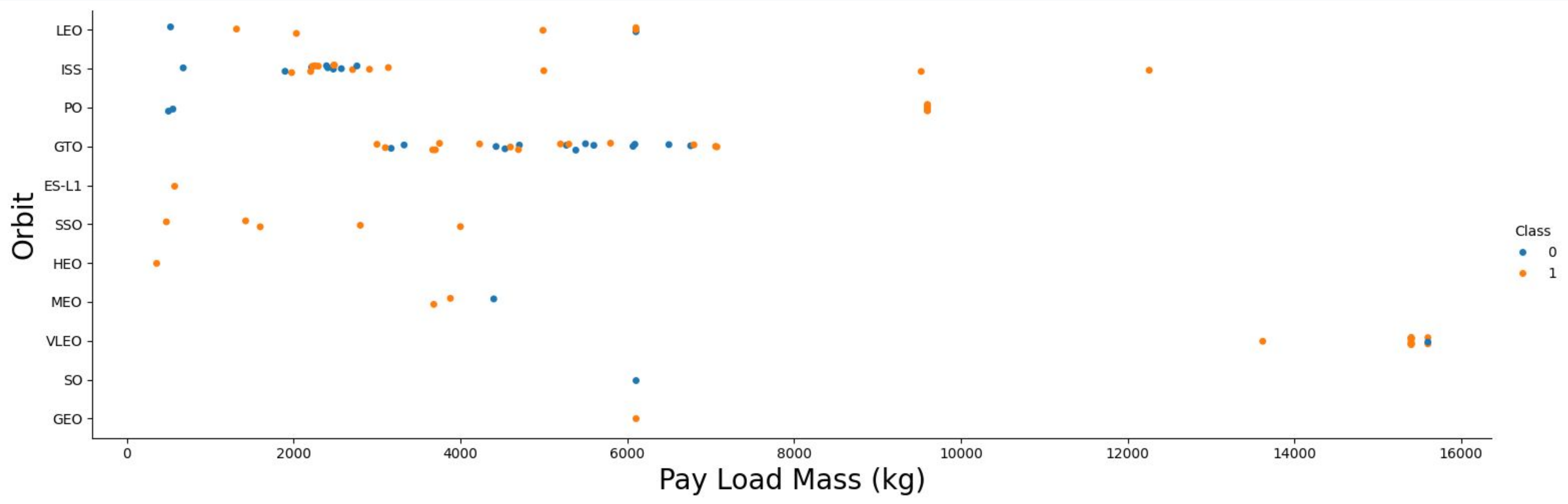


Flight Number vs. Orbit Type



- Success rate improve over time in all orbits.
- VLEO orbit seems to be new due to recent increase in its frequency.

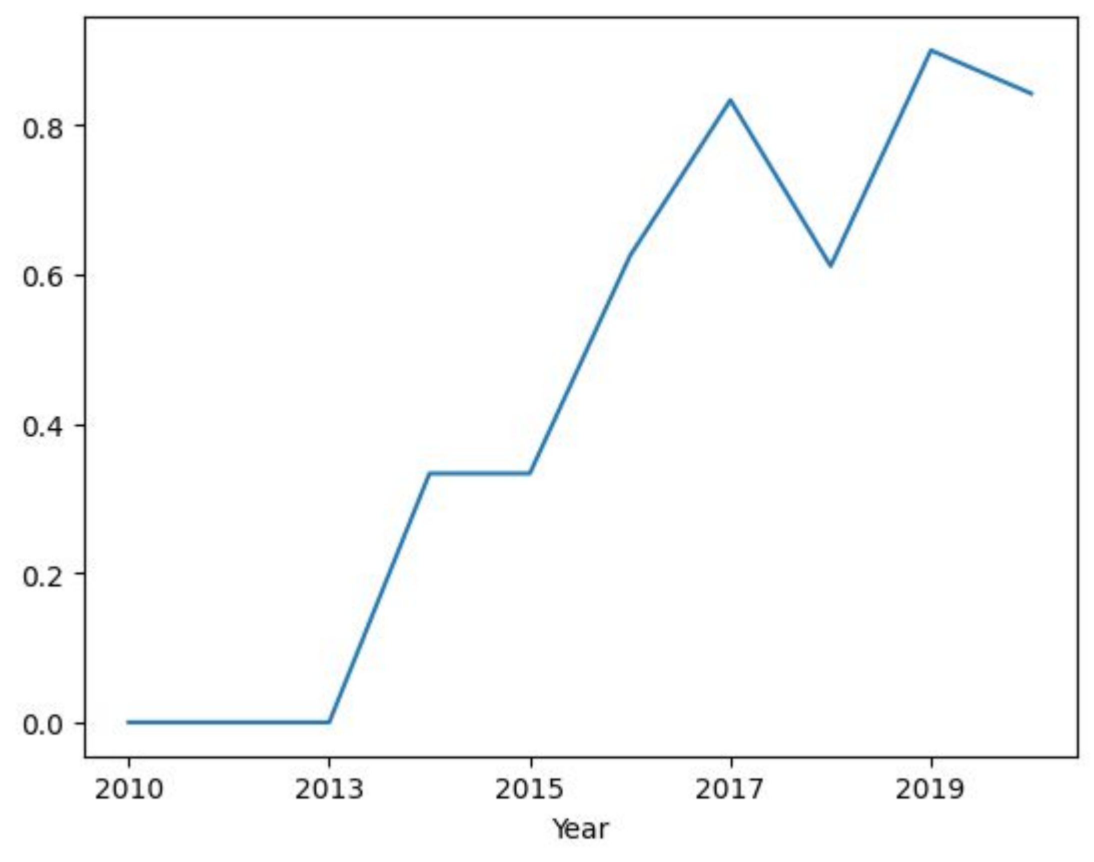
Payload vs. Orbit Type



- No relation between payload & success rate to GTO orbit.
- ISS orbit has widest range of payload with good rate of success
- There are fewer launches to orbit SO & GEO as compared to others.

Launch Success Yearly Trend

- Success rate started to increase in 2013 and kept until 2020
- There was a small dip in 2018
- The first 3 years (2010-2013) where period of improvement and setup of technology.



All Launch Site Names

- As per the data, there are four launch site:

```
%sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;
```

[9]

* sqlite:///my_data1.db
Done.

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

- There were abstained using “DISTINCT” record from the *launch_site* values from the dataset

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`

```
%sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

Python

```
 * 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

- To get 5 records of Cape Canaveral launches were used LIMIT 5

Total Payload Mass

- Total payload carried by boosters from NASA:

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE PAYLOAD  LIKE '%CRS%';
```

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

TOTAL_PAYLOAD
111268

- We summed all the payload values who codes contained ‘CRS’ which corresponds to NASA

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1:

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

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

AVG_PAYLOAD
2928.4

- Filtering data by the booster version above and calculating the average payload mass we obtained the value of 2,928 kg.

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad:

```
%sql SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (ground pad)';
```

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

FIRST_SUCCESS_GP
2015-12-22

- By filtering data by successful landing outcome on ground pad and getting the minimum value for date it's possible to identify the first occurrence, that happened on 2015-12-22.

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

```
%sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 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

- Selected **DISTINCT** booster versions according to filter above on payload mass

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes:

```
%sql SELECT MISSION_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL GROUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME;
```

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

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

- Used Grouping using **GROUP BY** on mission outcomes & then counting records for each group.

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass:

```
%sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL) ORDER BY BOOSTER_VERSION;
```

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

Booster_Version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

- List of booster which carried max payload as per dataset

2015 Launch Records

- The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015:

```
%sql SELECT substr(Date, 6, 2) AS Month, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE LANDING
```

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

Month	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

- There are only two occurrences of failed landing outcomes in drone ships. both have same booster versions and launch site

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

- The count of landing outcomes between the date 2010-06-04 and 2017-03-20:

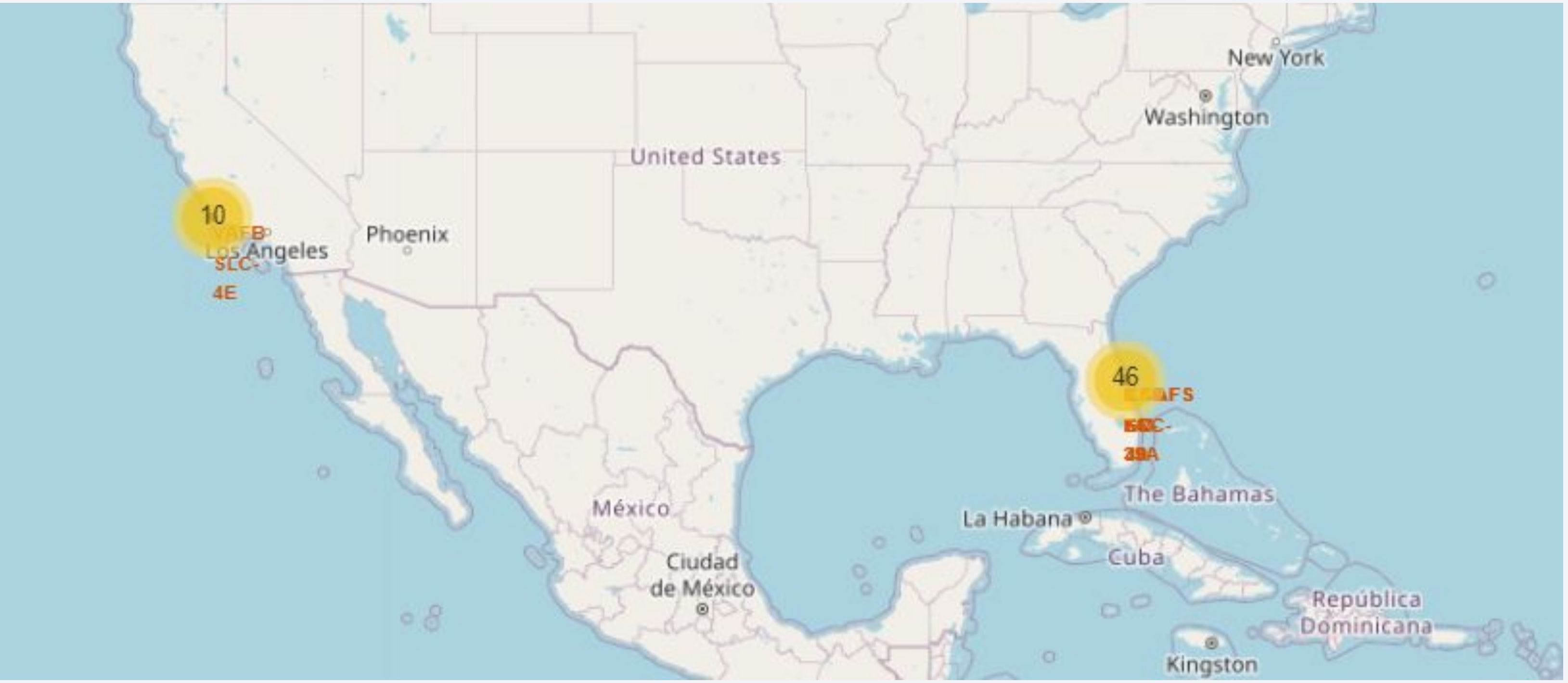
Landing_Outcome	QTY
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

- This data shows us that “No attempt” should be taken into account.

Section 3

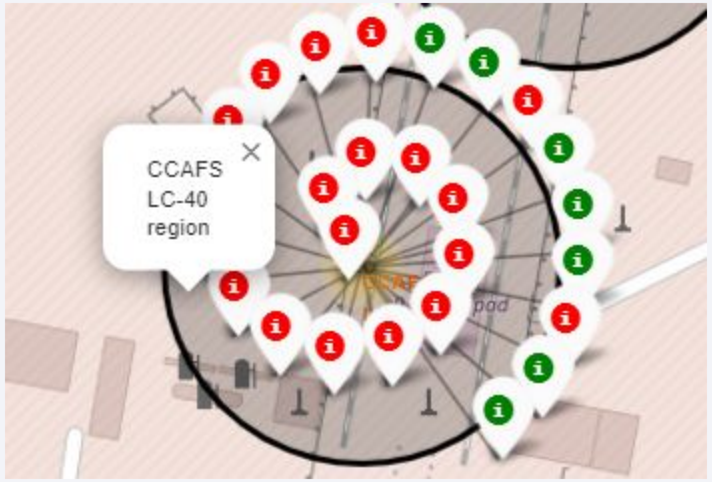
Launch Sites Proximities Analysis

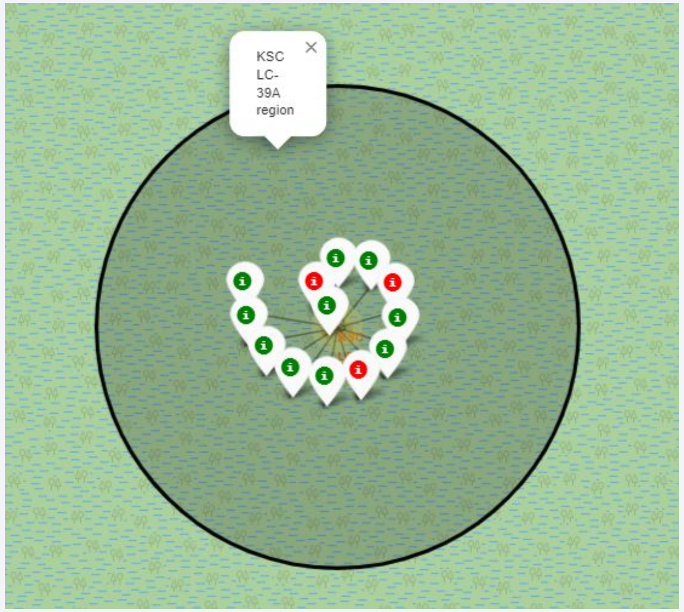
All launch sites



- Launch sites are near the coast (sea), this is probably for safety reasons.

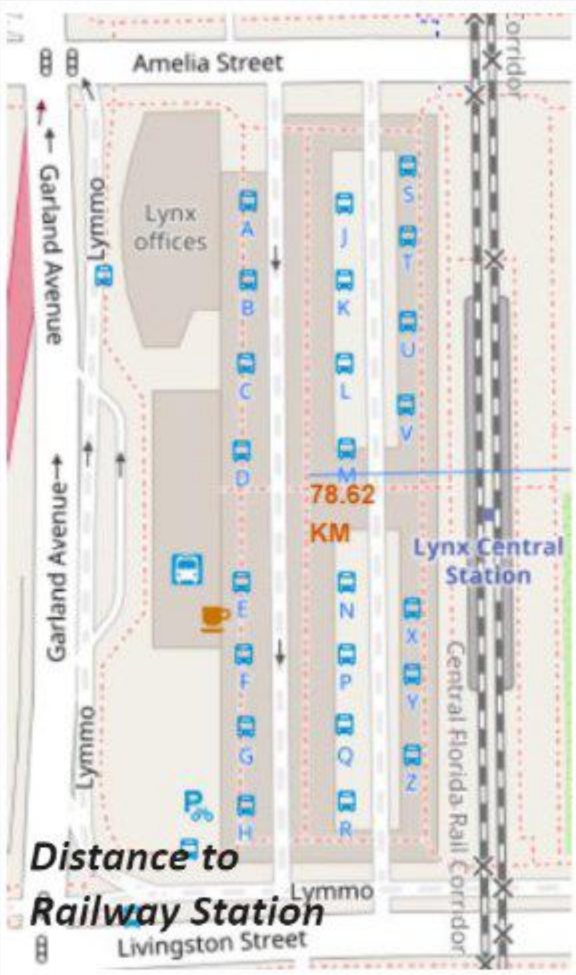
Launch Outcomes by Site (using Markers)

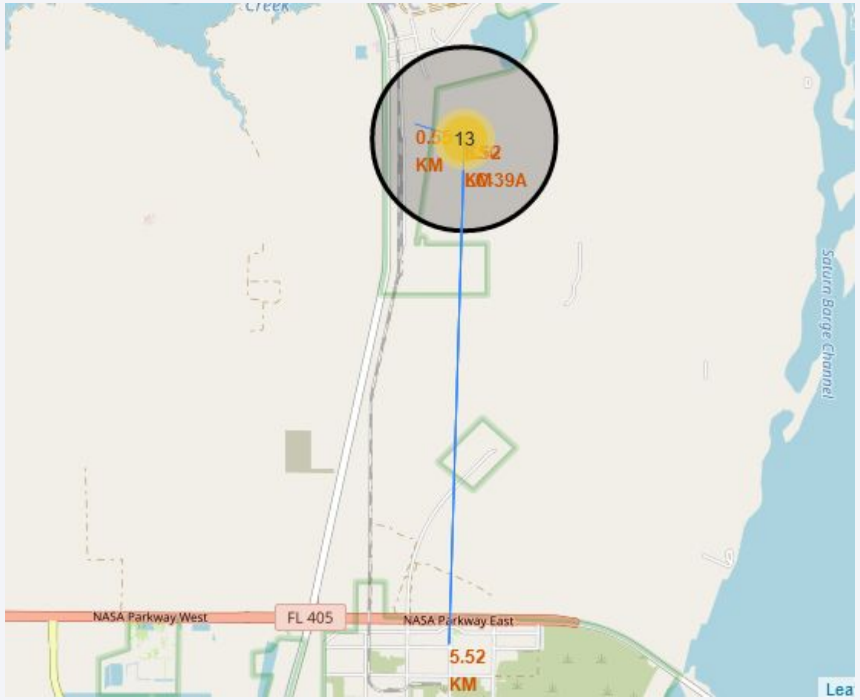




- Red label markers show failed launches and green ones show successful launches.

Launch site distance from Landmarks





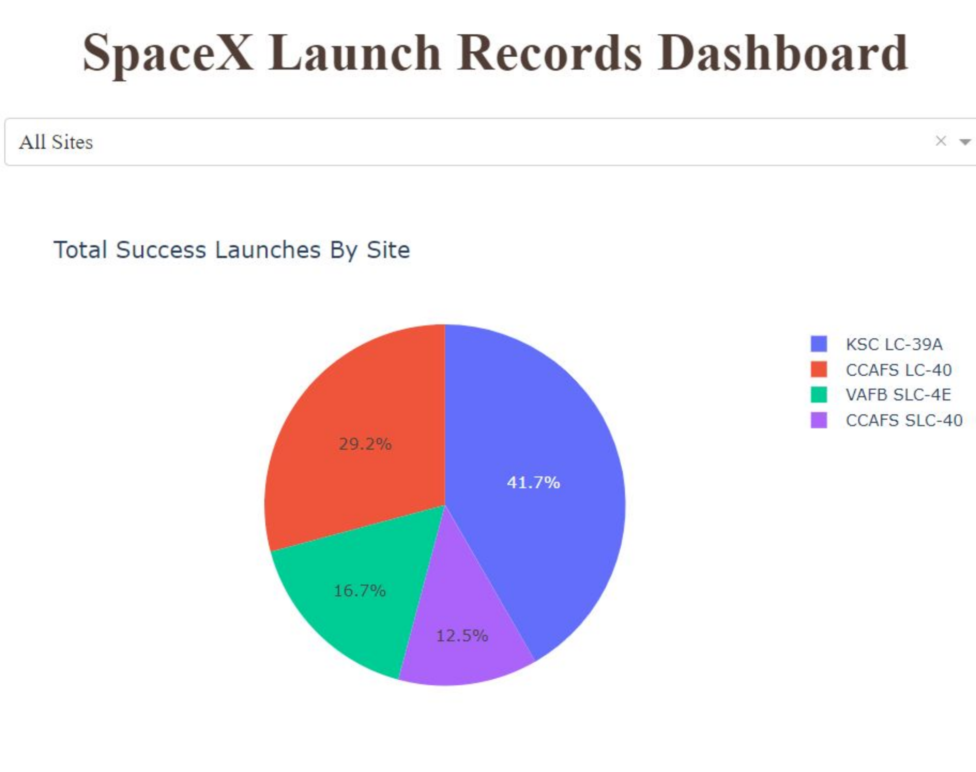
- All sites are close to coastline
- Few launch sites are close proximity to railroads, not true with all.
- Sites are away from major cities and railways station or highways.

Section 4

Build a Dashboard with Plotly Dash

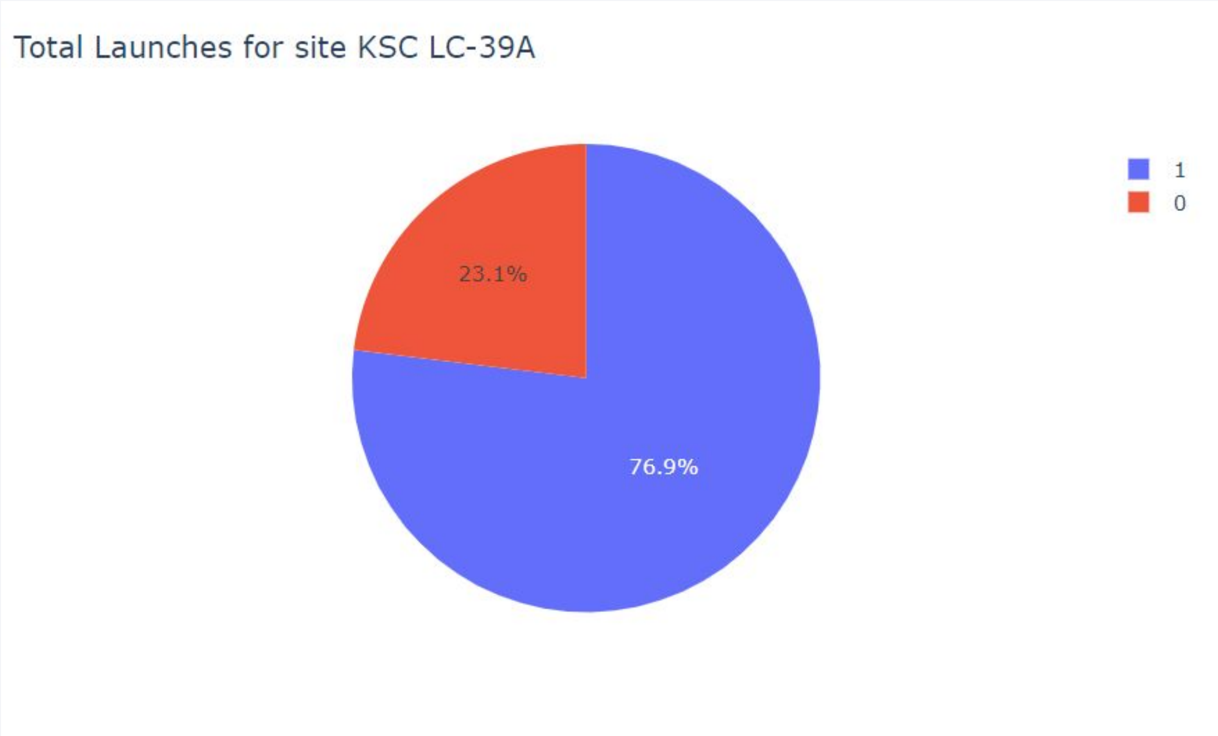
Successful Launches by Sites

The place from where launches are done seems to be a very important factor of success of missions.



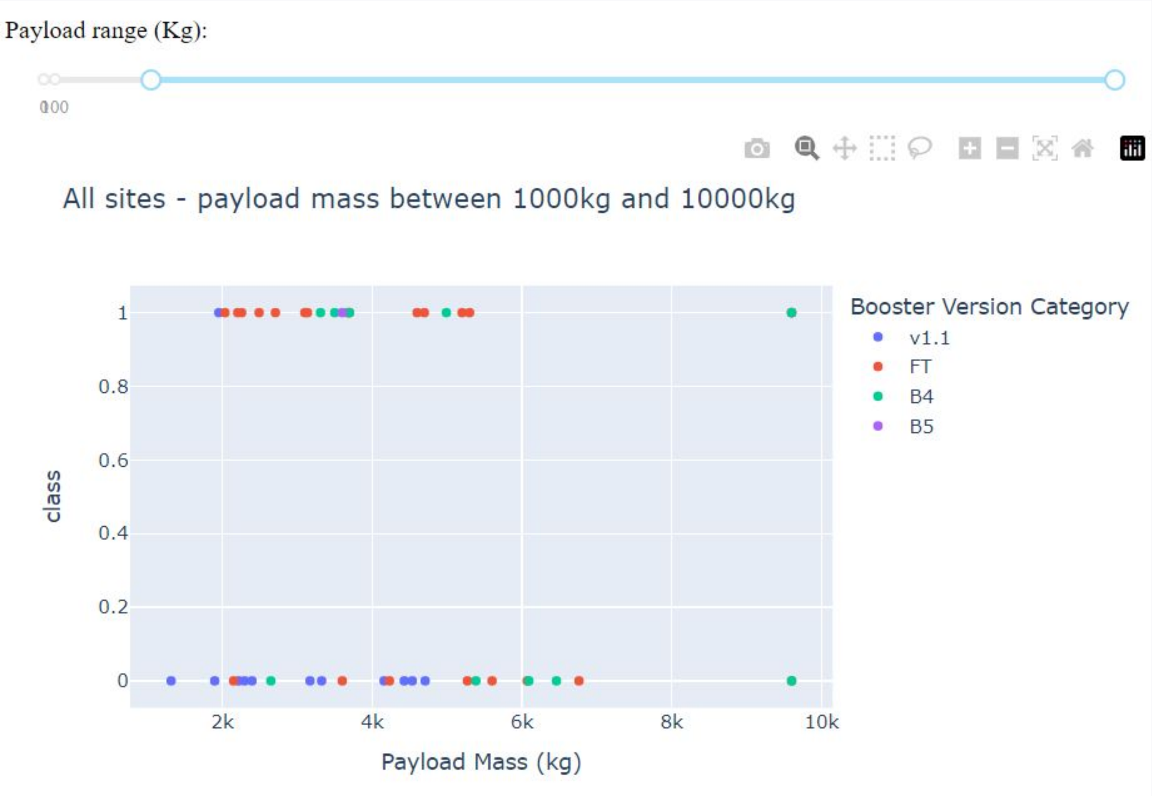
Launch Success ratio of KSC LC-39A

- KSC LC-39A has the highest launch success ratio of 76.9%



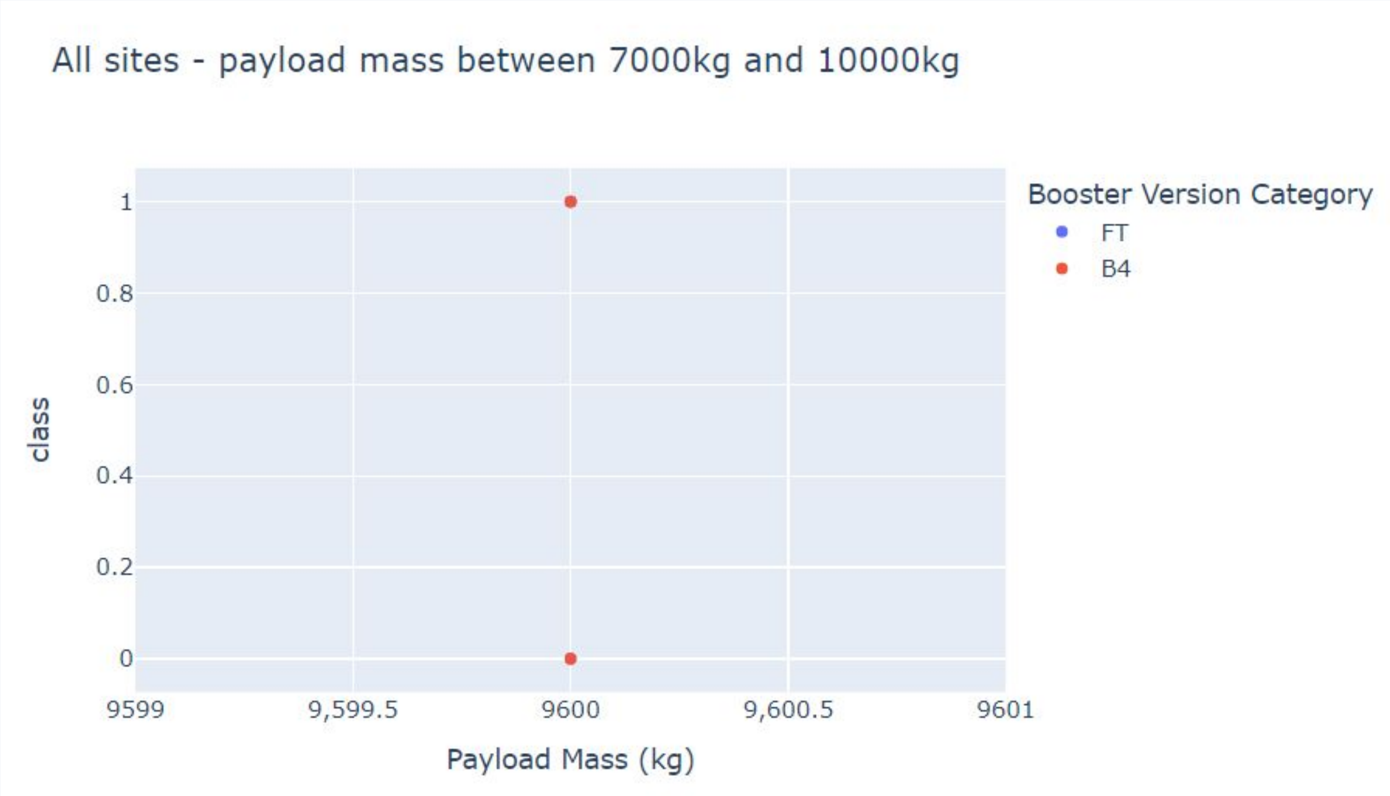
Payload Vs. Launch Outcome

- Payload range under 6,000kgs and FT boosters have most successful combination.



Payload Vs. Launch Outcome

There is no enough data to
assess the risk to launch
with payload over 7,000kgs.



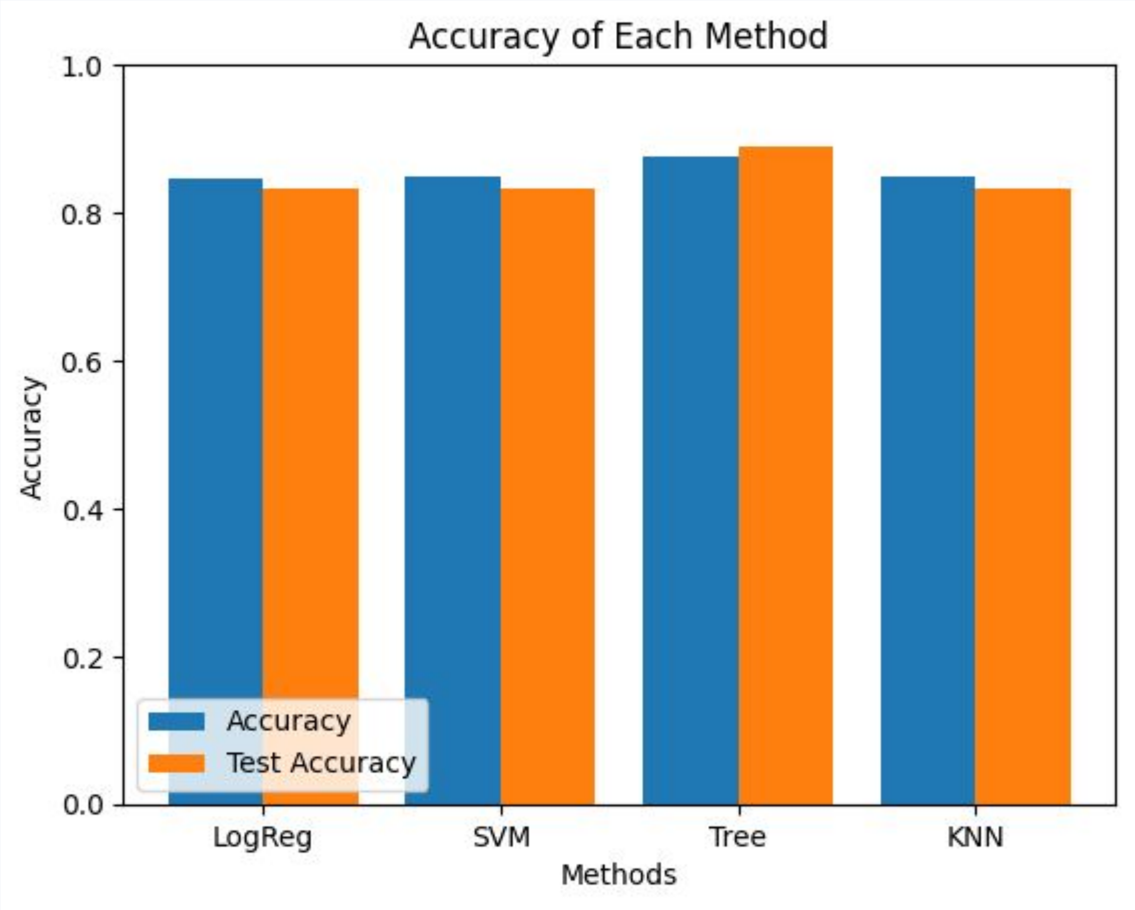
Section 5

Predictive Analysis (Classification)

Classification Accuracy

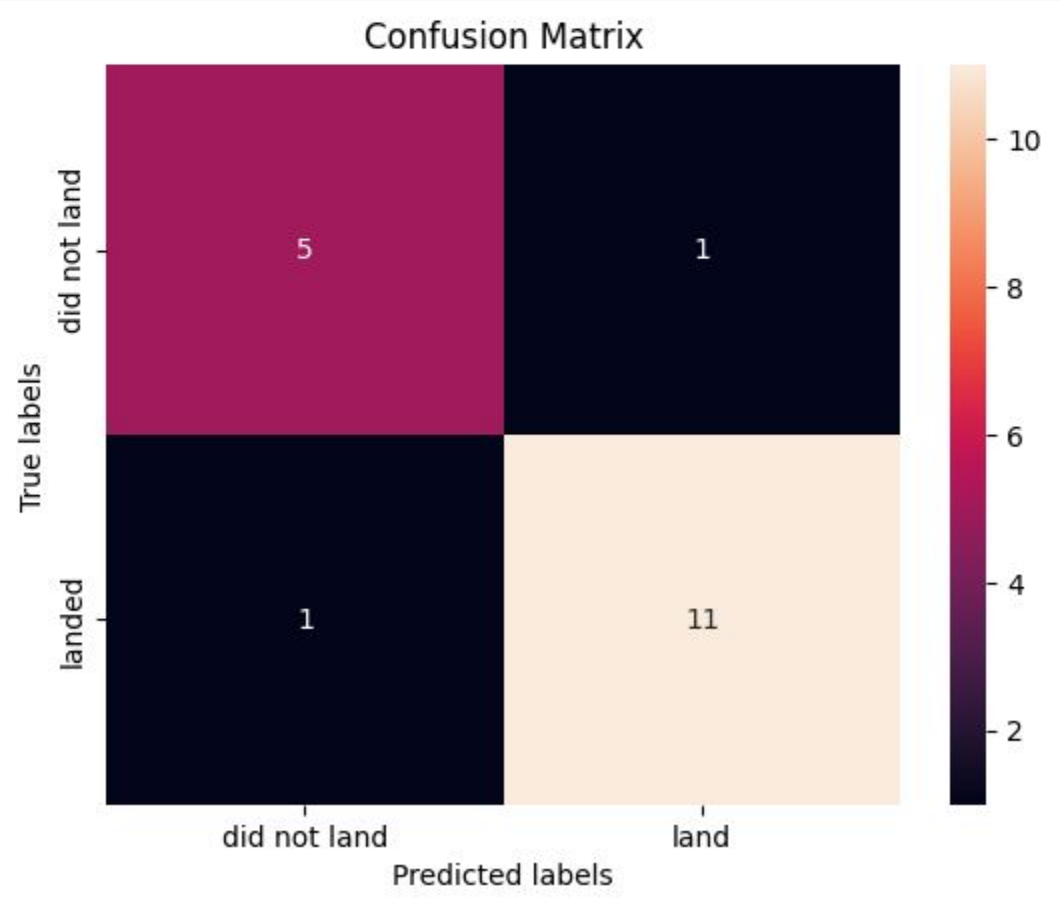
- Four Classifier were tested and their accuracy on test and train data were plotted on bar plot.

- The model with highest classification accuracy is **Decision Tree Classifier**.
- Accuracy of Decision Tree was **88%**



Confusion Matrix

- Confusion matrix of Decision Tree Classifier provided high accuracy by having higher number of True Positive (TP) and True Negative (TN) results as compared to False ones.



Conclusions

- Various data points we analyzed, refining to conclusions along the process
- The best launch site based on successful launches is **KSC LC-39A**
- Launches with Payload Mass above 7,000 kg are less risky.
- Success of landing outcomes seems to improve over time, this could be according to in improvement and advancement in processes.
- Decision Tree classifier gave the hughes accuracy in predicting successful landings and increasing profits.

Appendix

- Screenshots of Folium maps were taken. As Github didn't show them.

Thank you!