IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Walter Chang
February 19th 2025

Outline

- Methodology Overview
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Methodology Overview

Data Collection:

- Data was gathered using the SpaceX API and web scraping techniques. This included historical information on Falcon 9 launches, such as payload details, launch sites, and outcomes.
- The collected data was cleaned and preprocessed to handle missing values, standardize formats, and encode categorical variables for machine learning models.

Methodology Overview

Exploratory Data Analysis (EDA):

- EDA focused on identifying patterns, relationships, and anomalies within the data. Key steps included:
  - Visualizing trends over time (e.g., success rates by year).
  - Analyzing the impact of features like payload mass, orbit type, and launch site on mission outcomes.
  - Identifying key variables that influence launch success through statistical and visual methods such as scatter plots and correlation matrices.
- Launch sites like KSC LC-39A were found to have the highest success rates, while certain orbits (e.g., GEO, SSO) exhibited 100% success rates.

Methodology Overview

Machine Learning Prediction:

- Models such as Logistic Regression, Support Vector Machines (SVM), Decision Trees, and Random Forests were trained to predict the likelihood of a successful landing.
- The dataset was split into training and testing subsets to evaluate model performance. Metrics like accuracy and precision were used for assessment.
- The Decision Tree model slightly outperformed others in predicting outcomes, highlighting critical factors such as payload mass and launch site conditions

Methodology Overview

Summary of Results:

- The analysis revealed that:
  - Launch success rates have improved over time.
  - Proximity to the equator and coastal locations enhances success due to reduced fuel requirements.
  - Payload mass significantly influences launch outcomes, with lighter payloads often linked to higher success rates.
- Machine learning models provided insights into which features most strongly predict successful landings, aiding in decision-making for future launches

Introduction

- Objective:
  - Assess the feasibility of Space Y entering the market as a competitor to SpaceX.
- Key Questions:
  - What is the optimal method to estimate total launch costs by predicting the likelihood of successful first-stage rocket landings?
  - Where are the most advantageous locations for launch sites to maximize efficiency and success rates?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data from Space X was obtained from :
    - https://api.spacexdata.com/v4/rockets/
    - https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches
- 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
  - The collected data was normalized, split into training and test sets, and evaluated using four classification models with accuracy assessed across various parameter combinations.

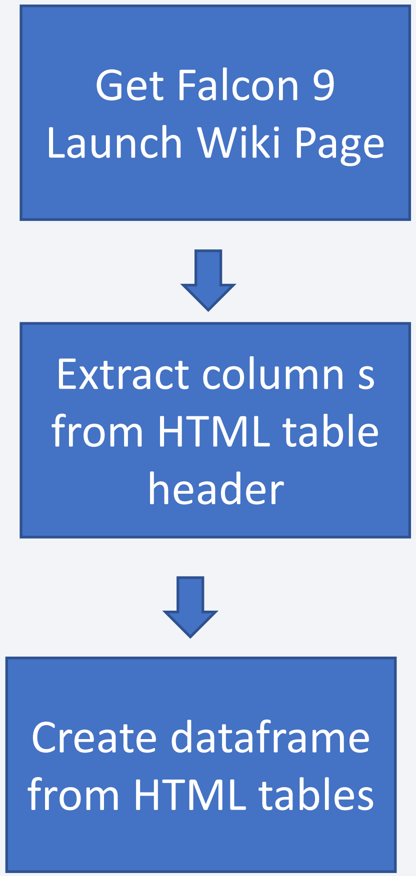
Data Collection – SpaceX API

https://github.com/wxcuop/AppliedDataScienceCapstone/blob/main/Module1/jupyter-labs-spacex-data-collection-api-v2.ipynb

Data Collection - Scraping

- Use python requests library to get SpaceX launches from Wikipedia
- Source:

https://github.com/wxcuop/AppliedDataScienceCapstone/blob/main/Module1/jupyter-labs-webscraping.ipynb



Data Wrangling

- Simple Exploratory Data Analysis (EDA) was performed
- Summaries of launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type were calculated

- Source https://github.com/wxcuop/AppliedDataScienceCapstone/blob/main/Module1/labs-jupyter-spacex-Data%20wrangling-v2.ipynb

EDA with Data Visualization

- Scatterplots and Bar Plots were used to visualize the relationship between pair of features:

Payload Mass X Flight Number
Launch Site X Flight Number
Launch Site X Payload Mass
Orbit and Flight Number
Payload and Orbit

Source: https://github.com/wxcuop/AppliedDataScienceCapstone/blob/main/Module2/jupyter-labs-eda-dataviz-v2.ipynb

EDA with SQL

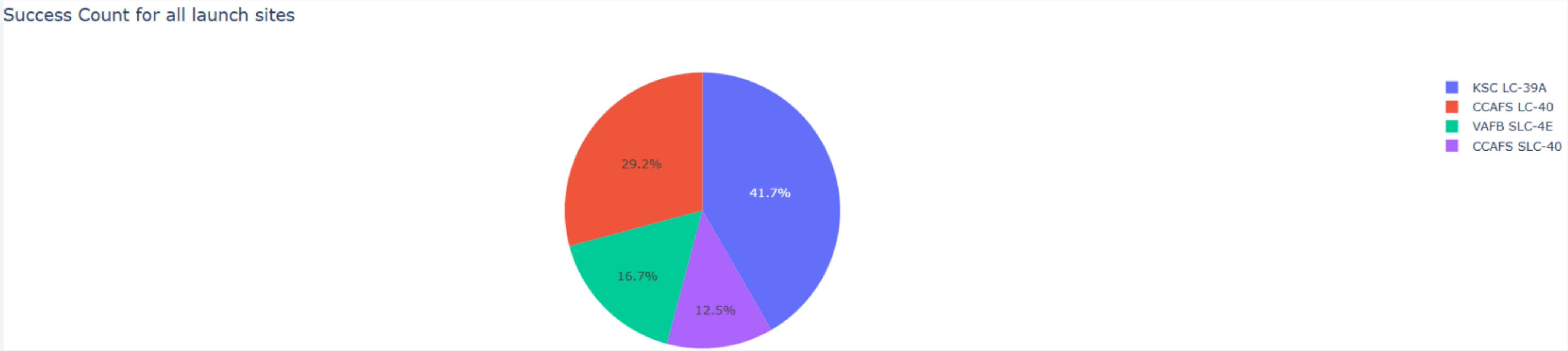
- SQL queries performed:
  - Names of the unique launch sites in the space mission
  - Top 5 launch sites whose name begin with the string 'CCA'
  - Total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1
  - Date when the first successful 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.
- Source: https://github.com/wxcuop/AppliedDataScienceCapstone/blob/main/Module2/jupyter-labs-eda-dataviz-v2.ipynb

Build an Interactive Map with Folium

- 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.
- Source: https://github.com/wxcuop/AppliedDataScienceCapstone/blob/main/Module3/lab-jupyter-launch-site-location-v2.ipynb

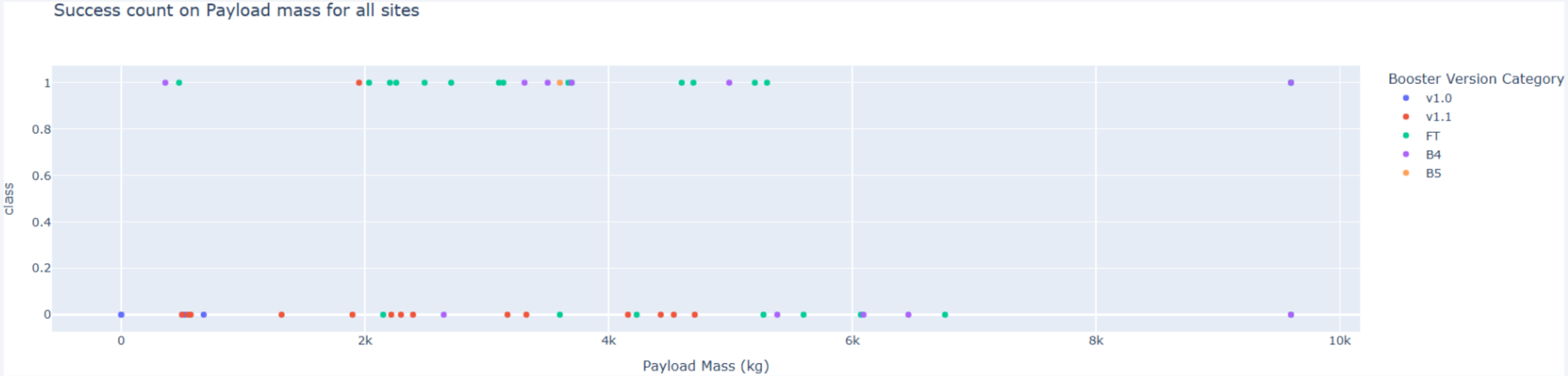
Build a Dashboard with Plotly Dash

- Source code: https://github.com/wxcuop/AppliedDataScienceCapstone/blob/main/Module3/lab-jupyter-launch-site-location-v2.ipynb



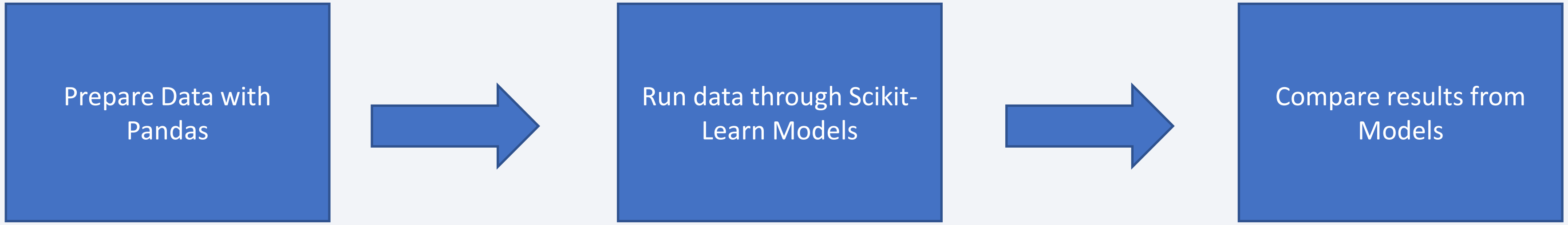
Build a Dashboard with Plotly Dash

- Source code: https://github.com/wxcuop/AppliedDataScienceCapstone/blob/main/Module3/lab-jupyter-launch-site-location-v2.ipynb



Predictive Analysis (Classification)

- Source Code: https://github.com/wxcuop/AppliedDataScienceCapstone/blob/main/Module4/SpaceX-Machine-Learning-Prediction-Part-5-v1.ipynb



Results

	Algorithm	Accuracy Score	Best Score
0	Logistic Regression	0.833333	0.846429
1	Support Vector Machine	0.833333	0.848214
2	Decision Tree	0.666667	0.885714
3	K Nearest Neighbours	0.833333	0.848214

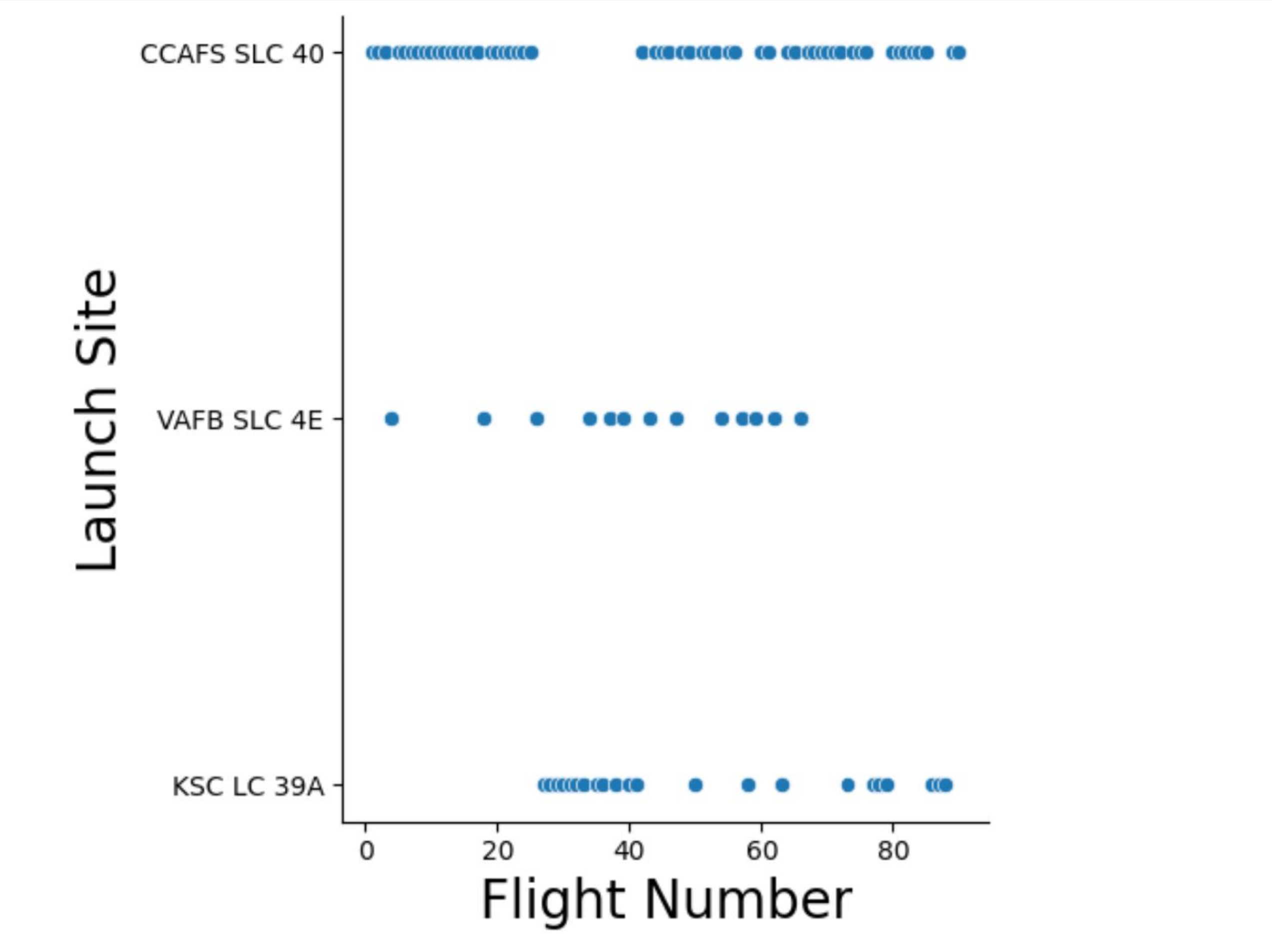
Results showed that while Decision Tree had the lowest accuracy, it had the best score over test data at 88.6%

Section 2

Insights drawn from EDA

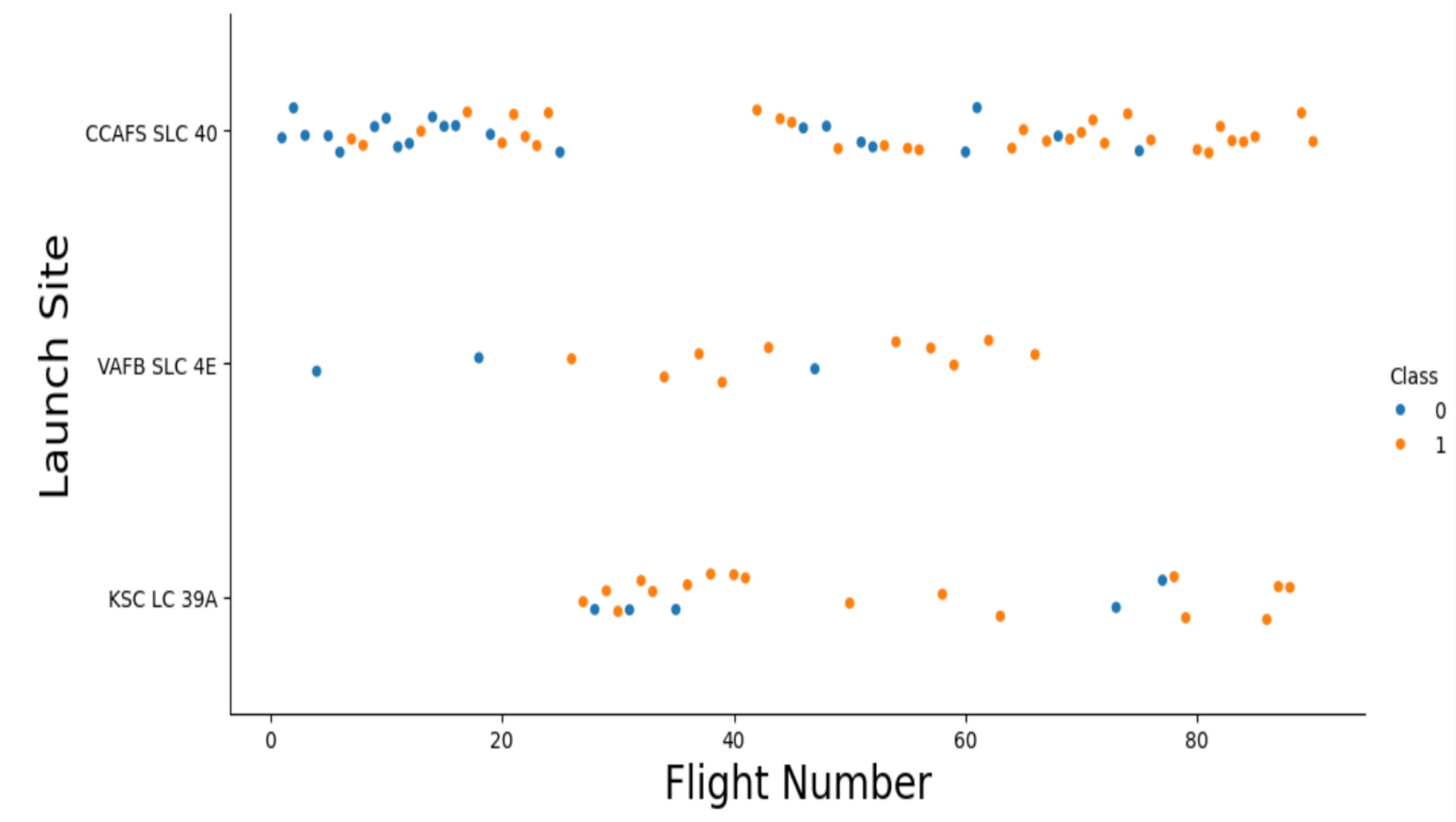
Flight Number vs. Launch Site

- From the plot, you can see that CCAFS SLC 40 is where most of the launches were sucessful



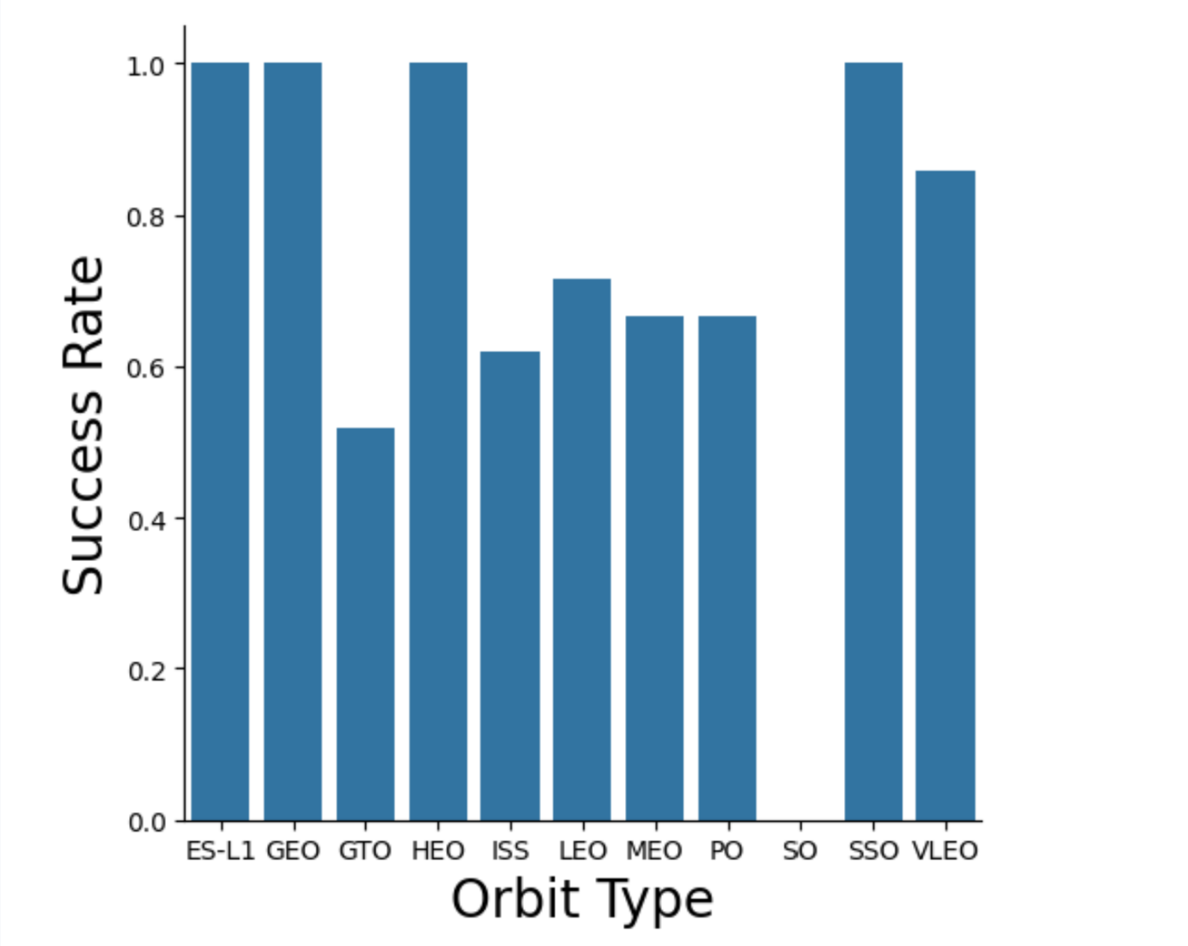
Payload vs. Launch Site

- if you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000)



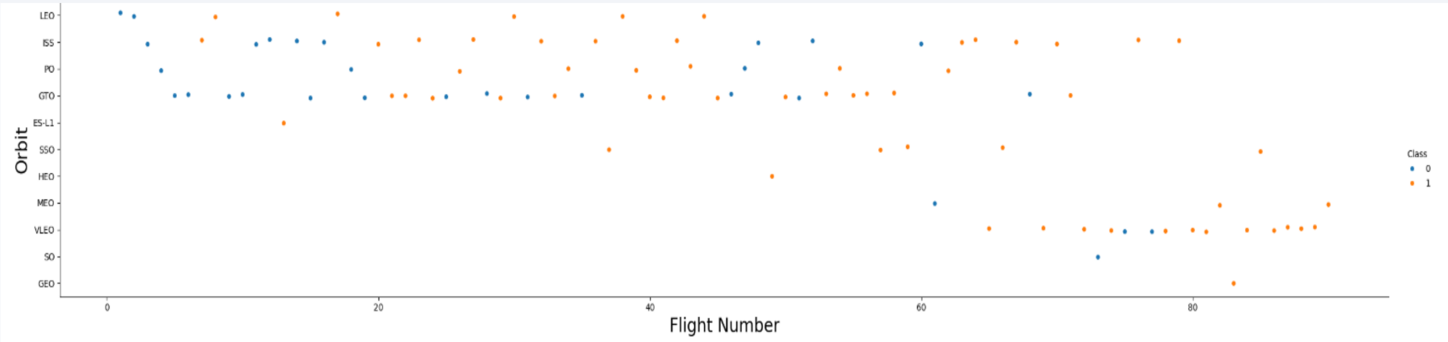
Success Rate vs. Orbit Type

- In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.



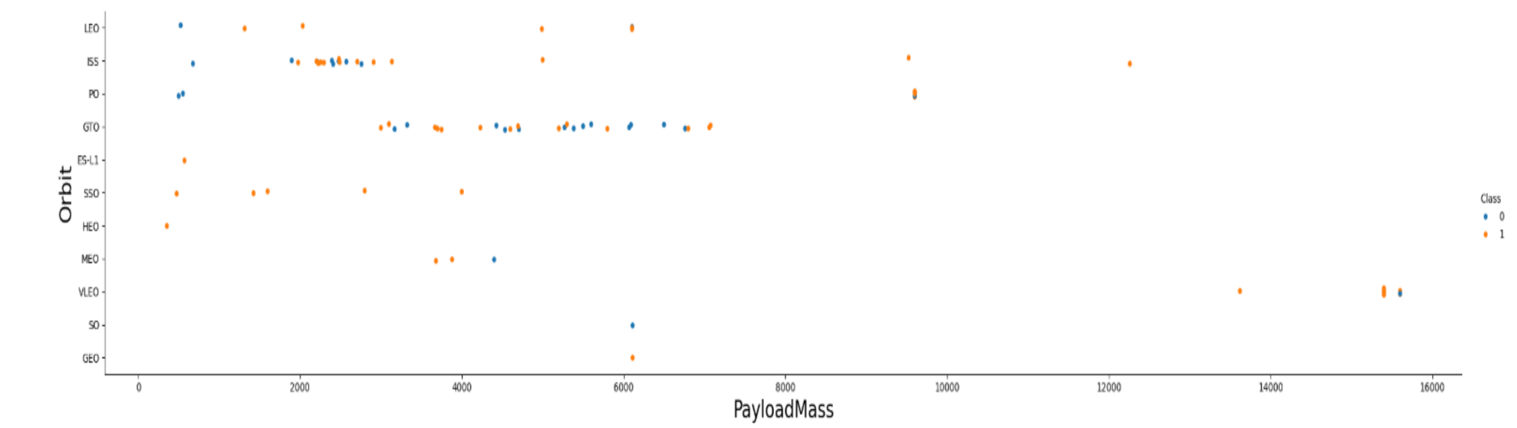
Flight Number vs. Orbit Type

- In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.



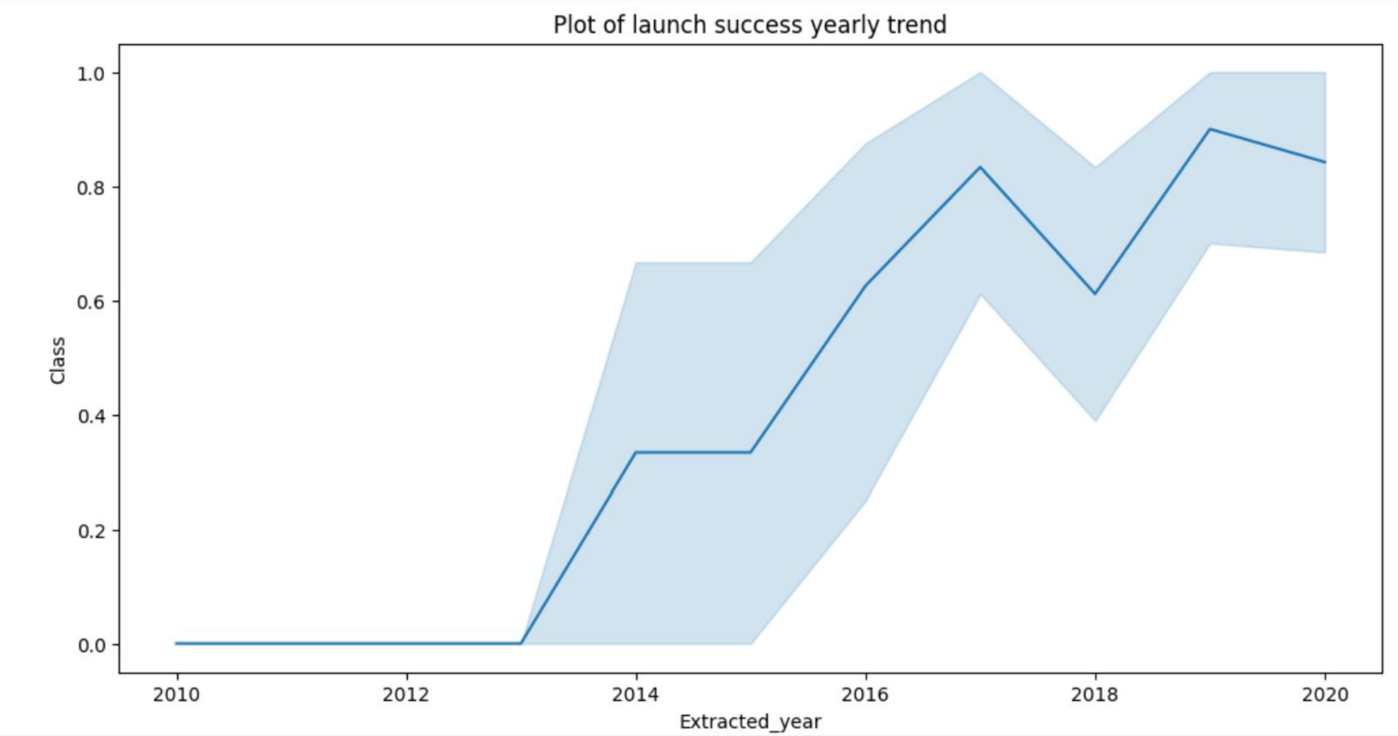
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.



Launch Success Yearly Trend

- You can observe that the success rate since 2013 kept increasing till 2017 (stable in 2014) and after 2015 it started increasing.



All Launch Site Names

FlightNumber	PayloadMass	Orbit	LaunchSite	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	
0	1	6104.959412	LEO	CCAFS SLC 40	1	False	False	False	NaN	1.0	0	B0003
1	2	525.000000	LEO	CCAFS SLC 40	1	False	False	False	NaN	1.0	0	B0005
2	3	677.000000	ISS	CCAFS SLC 40	1	False	False	False	NaN	1.0	0	B0007
3	4	500.000000	PO	VAFB SLC 4E	1	False	False	False	NaN	1.0	0	B1003
4	5	3170.000000	GTO	CCAFS SLC 40	1	False	False	False	NaN	1.0	0	B1004

Launch Site Names Begin with 'CCA'

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

- The total payload carried by boosters from NASA is

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

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

```
%sql SELECT MIN(DATE) FROM SPACEXTBL WHERE LANDING_OUTCOME="Success";
```

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

MIN(DATE)
2018-07-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List 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 Landing_Outcome FROM SPACEXTBL;
```

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

Landing_Outcome
Failure (parachute)
No attempt
Uncontrolled (ocean)
Controlled (ocean)
Failure (drone ship)
Precluded (drone ship)
Success (ground pad)
Success (drone ship)
Success
Failure
No attempt

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

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

Boosters Carried Maximum Payload

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

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

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

month	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

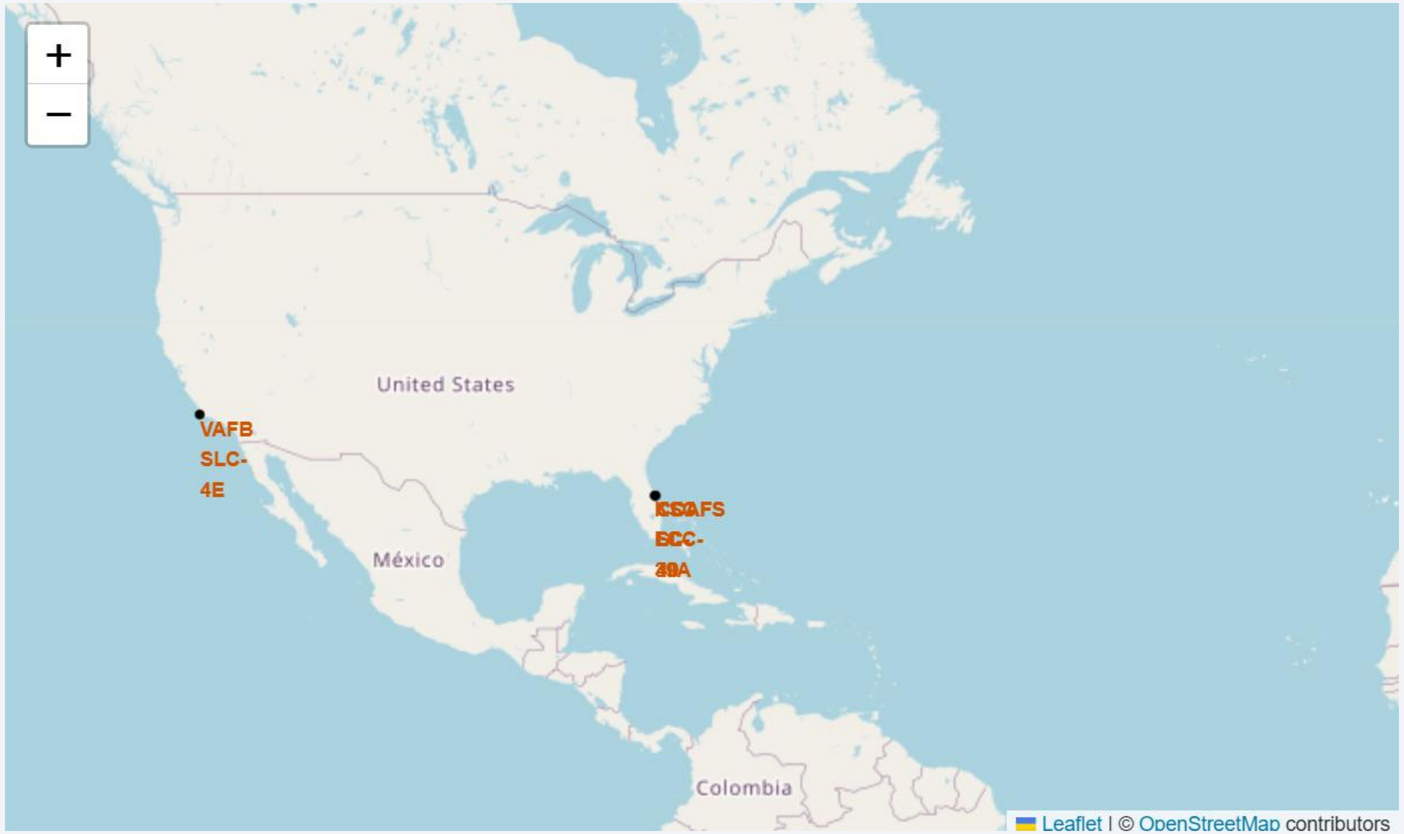
Landing_Outcome	COUNT(*)
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

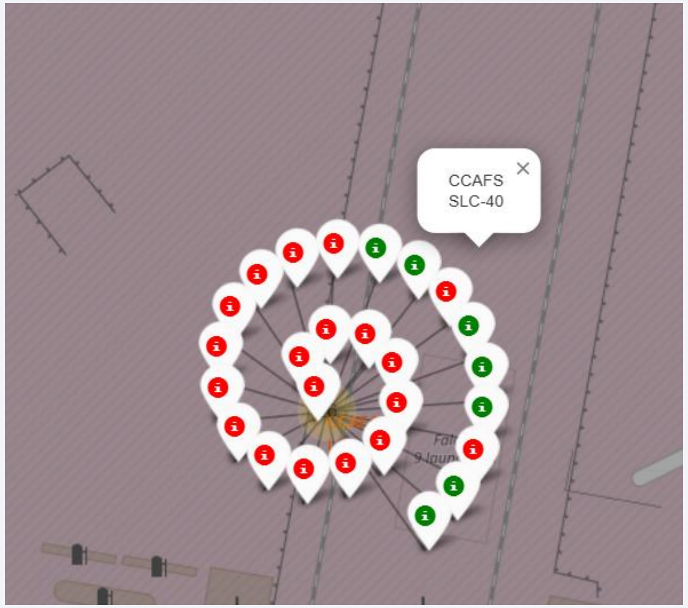
Launch Sites Proximities Analysis

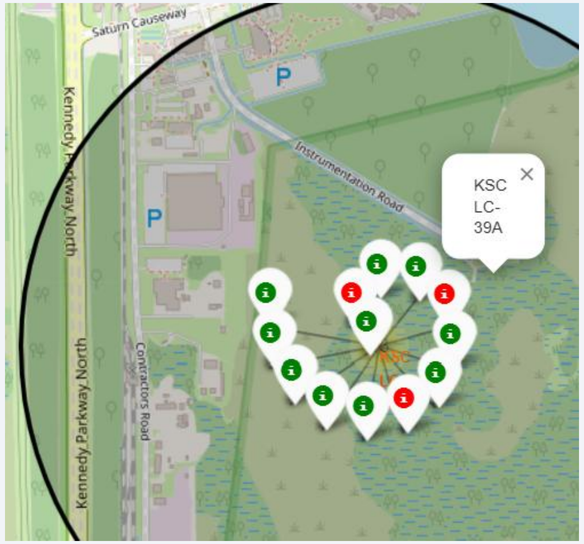
All launch sites on a map

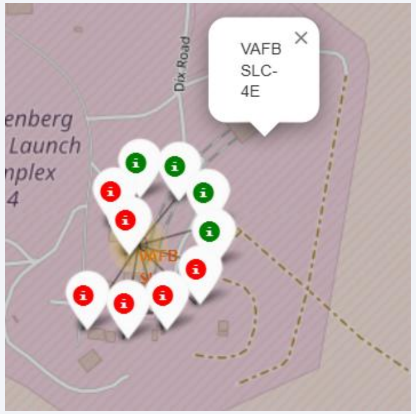
- Launch sites are near the oceans, likely due to safety concerns

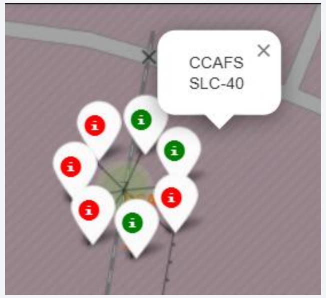


Success/Failed launches for each site on the map





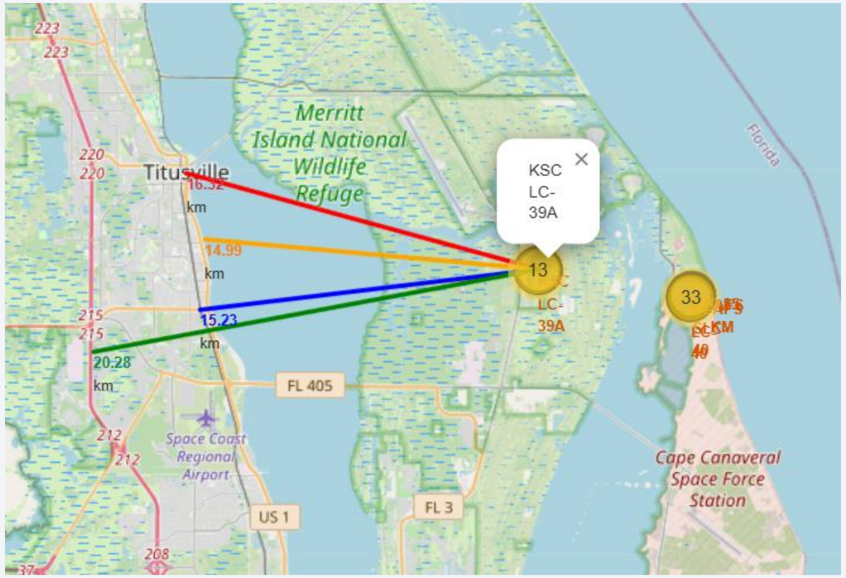




Green markers indicate successful and red ones indicate failure.

Visual analysis of the launch site KSC LC-39A

- From the visual analysis of the launch site KSC LC-39A it is:
- close to railway (15.23 km)
- close to highway (20.28 km)
- close to coastline (14.99 km)
- Also the launch site KSC LC-39A is close to its closest city Titusville (16.32 km).
- There is moderate risk to local population



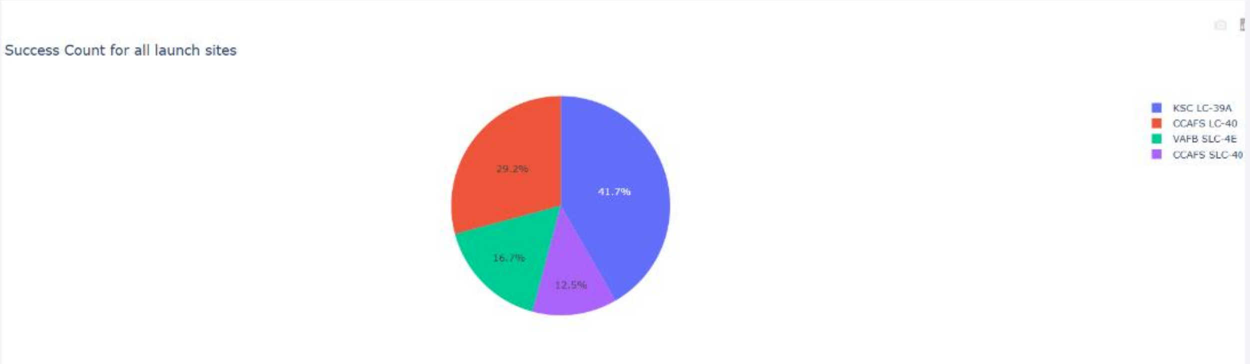
Section 4

Build a Dashboard with Plotly Dash

Successful Launches by Site

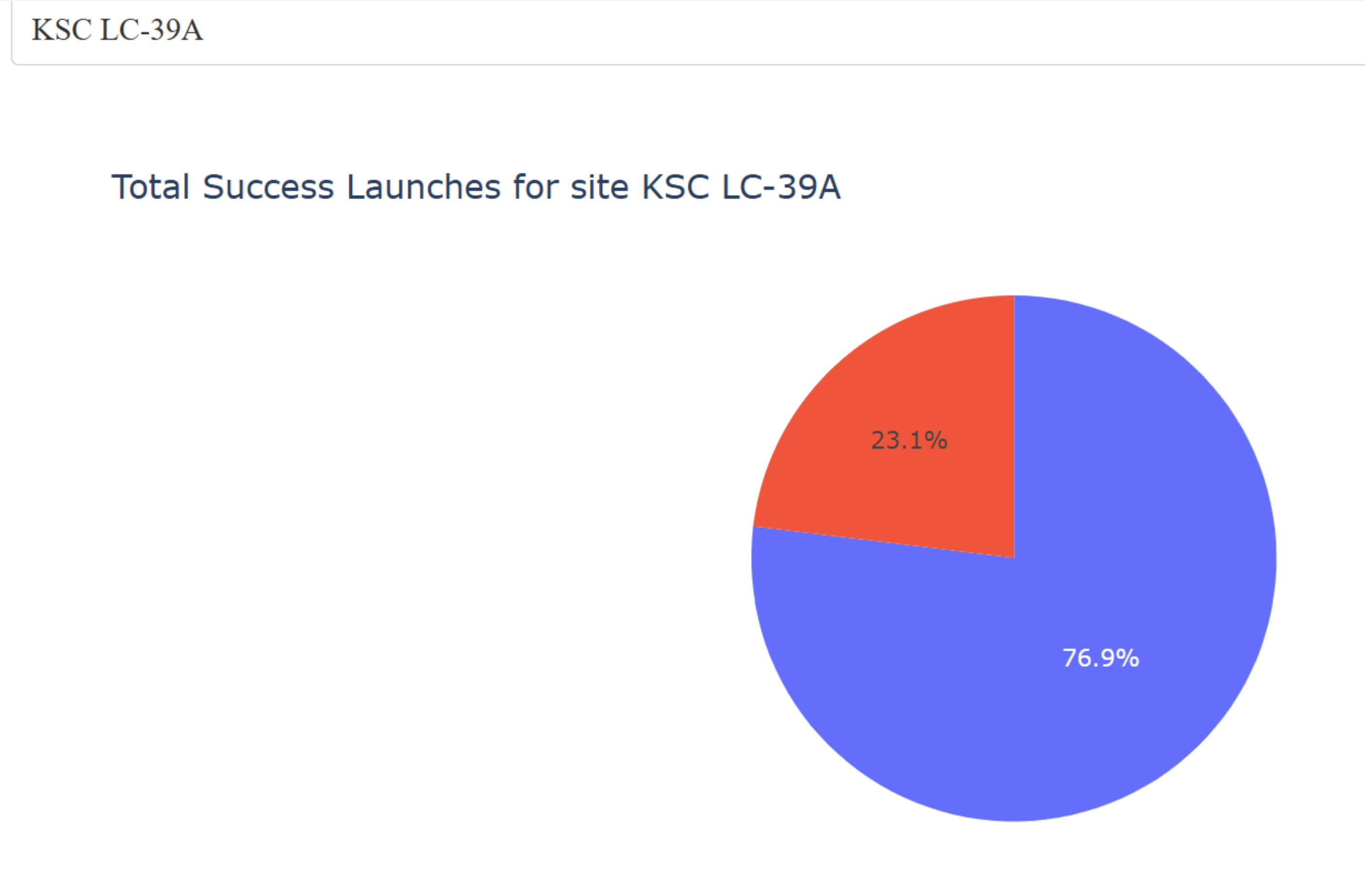
- KSC LC-30 has the highest percentage with 41.7%

- Location of launch site is statistically significant, as shown by the pie chart



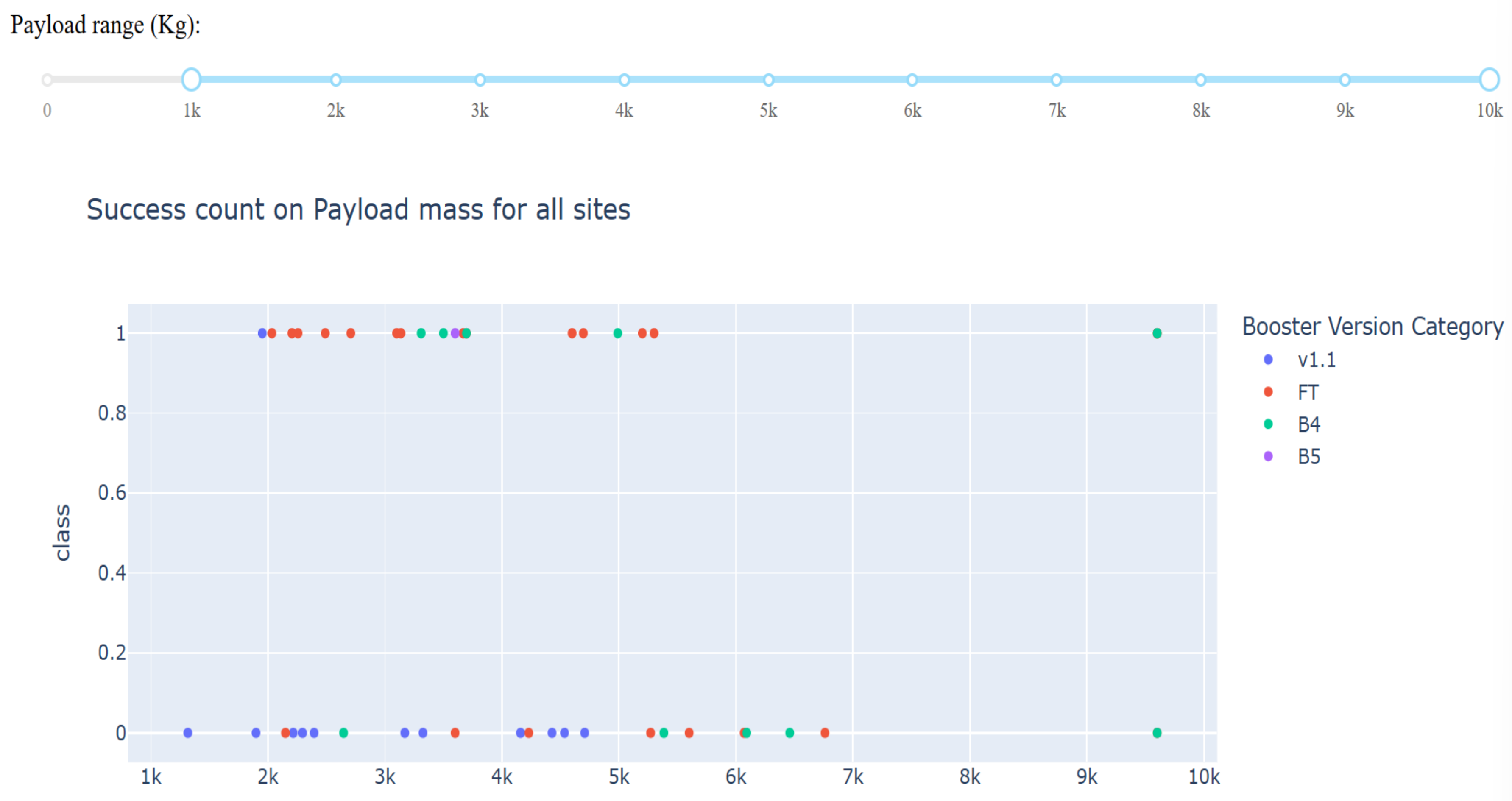
KSC LC-39A

- 76.9% of launches are successful in site KSC LC-39A



Payload vs. Launch

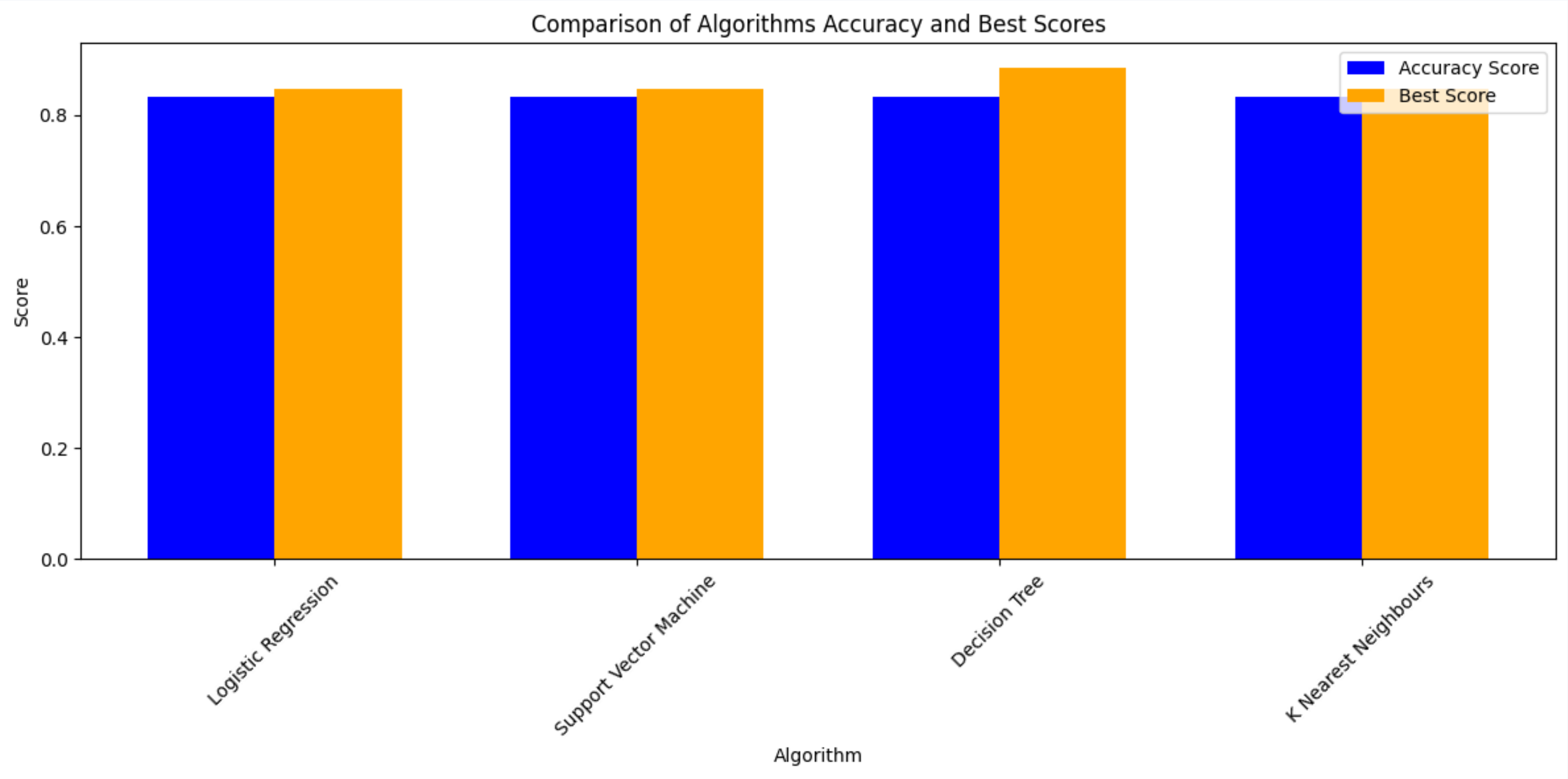
- Payloads under 7,000kg and FT boosters are the most successful combination.



Section 5

Predictive Analysis (Classification)

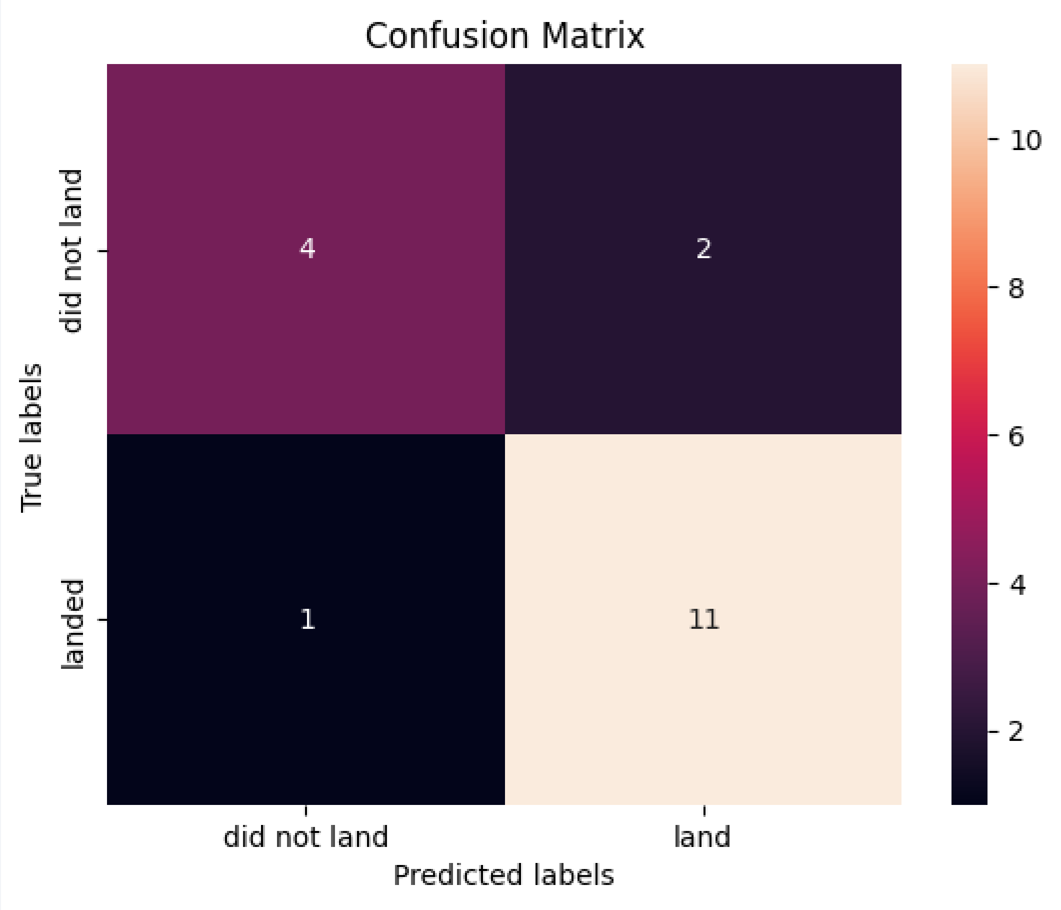
Classification Accuracy



- Decision Tree had best classification accuracy with accuracy over 88%

Confusion Matrix

- Confusion matrix of Decision Tree Classifier proves its accuracy by showing the big numbers of true positive and true negative compared to the false ones.



Conclusions

- Decision Tree Model is the best algorithm for this dataset.
- Launches with a low payload mass show better results than launches with a larger payload mass.
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- The success rate of launches increases over the years.
- KSC LC-39A has the highest success rate of the launches from all the sites.

Appendix

- In the decision tree classifier, the ‘auto’ in ‘max_features” is no longer valid: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

Thank you!