IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Karthikeyan S

25-May-2025

Outline

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

Executive Summary

- Summary of methodologies
    - Data Collection through Web scrapping
    - Data Preprocessing (Data Wrangling)
    - Exploring Data to identify independent input variables and correlation across them
    - Visual exploration using Maps and Dashboards
    - Prediction of Launch success & failure using different ML models, and identifying the best fit ML model
- Summary of all results
    - Key input variables are: ‘Payload Mass’, ‘Orbit’, ‘Booster Version’, and Launch sites. Year of launch
    - Decision Tree ML algorithm helped to reach an accuracy level of 0.88, with further accuracy improvement being possible using Grid Search Hyper parameter tuning. KNN had a higher accuracy (at 0.92) post Hyper Tuning

Introduction

- Project background and context
  - Rocket launching has become a good commercial activity due to multiple requirements of placing satellites into different height Orbits, such as weather monitoring and surveillance
  - Much of the margin is derived by reusing 'Stage One' of rocket being launched
  - Safe landing of Stage One depends on many parameters such as Payload, Demography of Landing Site, etc.
  - Use publicly available SpaceX information to build a model, for prediction of Landing
  - This can decide on the price point and hence the margin. This information is very useful for any new entrant that competes with the companies such as SpaceX
- Problems you want to find answers
  - What are the publicly available information to build a landing outcome Model?
  - What are the input variables that influences the landing outcome prediction?
  - Are there any physical proximity factors to be considered in landing sites and the outcome impact?
  - Which Data models are best to predict the landing outcomes, using the best corelating input variables?

Section 1

Methodology

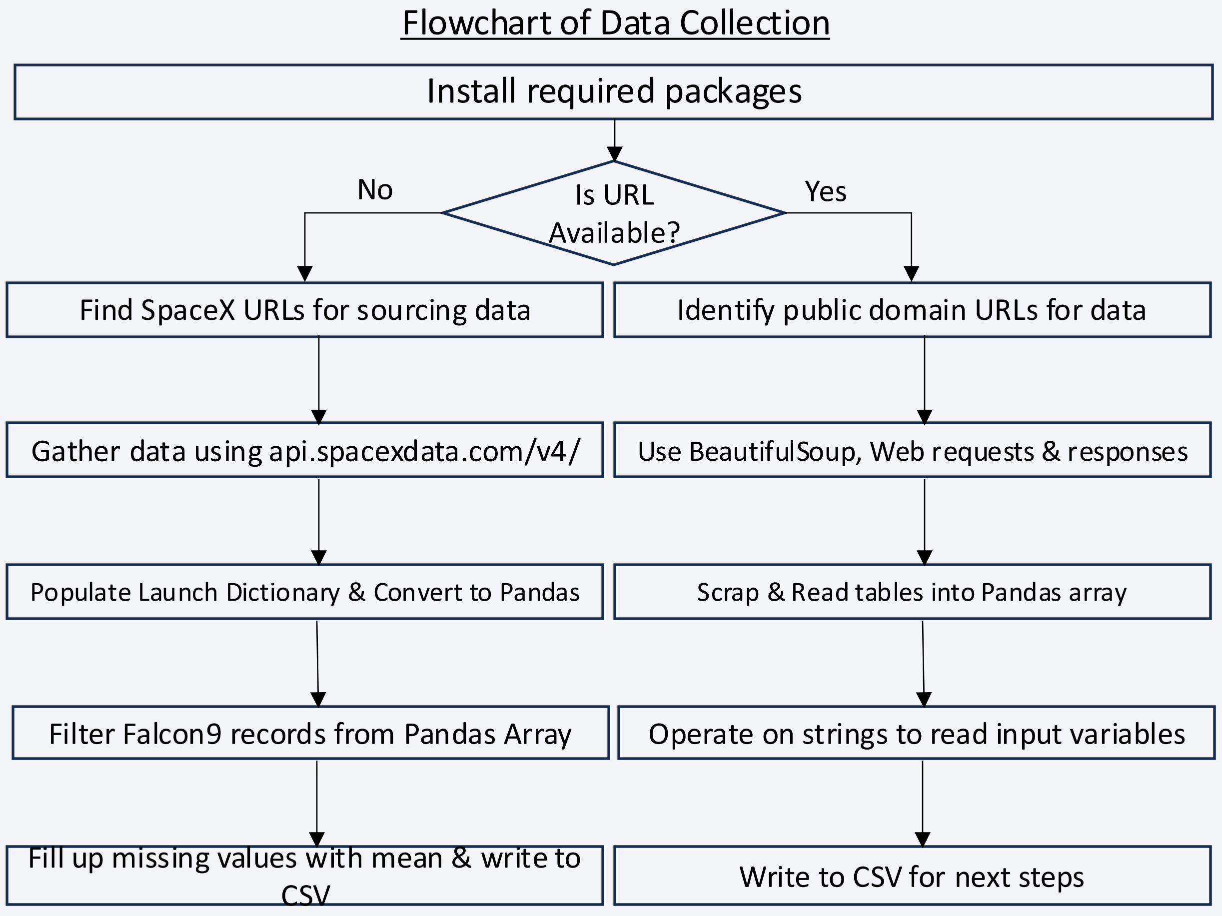
Methodology

Executive Summary

S#	Stage	Activities
1	Data collection methodology	Falcon9 Data was collected from web scrapping of SpaceX website and also using SpaceX RESTful APIs
2	Perform data wrangling	• Missing values were filled up with average value of the columns • Categorical variable columns were converted into Numerical with Dummy or Ordinal encoding • Standard scaler operation was done to normalize each Numerical column
3	Perform exploratory data analysis (EDA) using visualization and SQL	• To identify the correlation between different input variables and clustering with the outcomes. This is done using Scatter plots. Trending is done with a line chart. • With SQL queries: ranges, average, count of different input variables are identified
4	Perform interactive visual analytics using Folium and Plotly Dash	• Explore launch sites over a Map for the Geographic details such as nearest roads, cities and seashore • Explore different Launch sites for success ratio, Booster versions
5	Perform predictive analysis using classification models	• Split data into training and testing sets • Assess different ML models (Logistic Regression, Decision Tree, SVM, KNN.

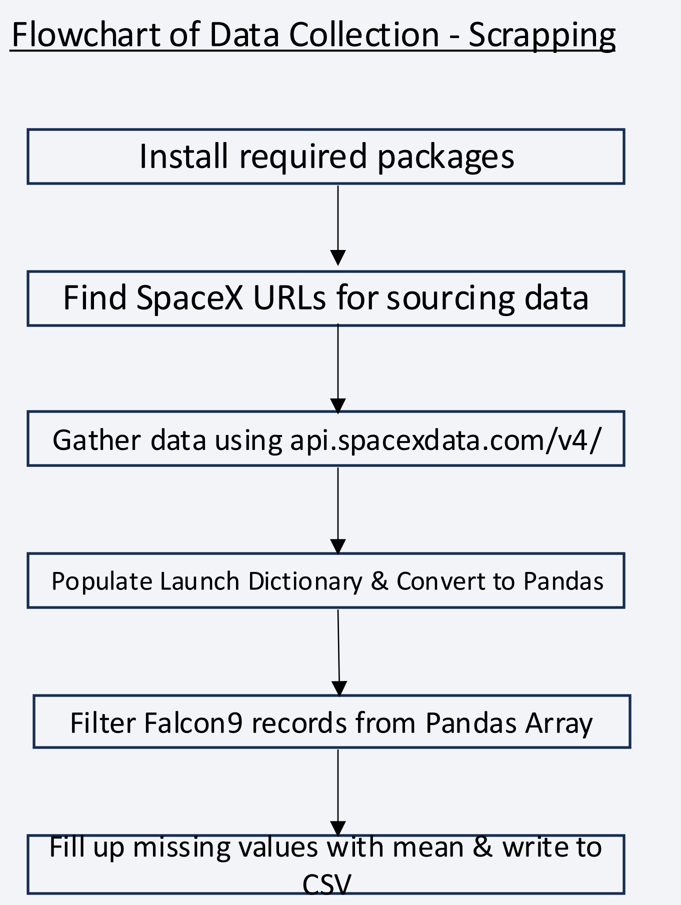
Data Collection

- Describe how data sets were collected
  - RESTful free API (api.spacexdata.com/v4/) calls as provided by SpaceX
  - Web Scrapping of SpaceX website
- Data collection process use key phrases and flowcharts
  - Read the gathered data (using API or web scrapping) into Pandas array
  - Filter only for Falcon9 records
  - Fill missing values with the column mean
  - Write to CSV file for the next stage



Data Collection – SpaceX API

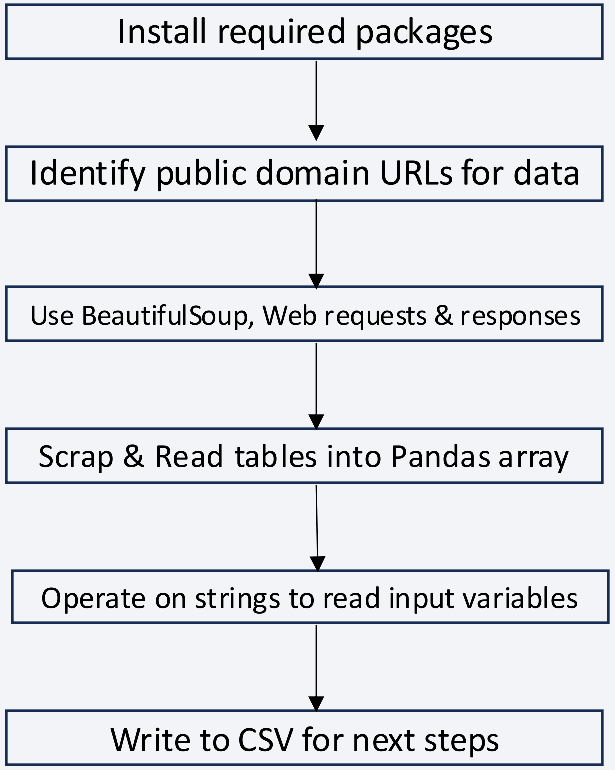
- Present your data collection with SpaceX REST calls using key phrases and flowcharts
    - Installing required packages
    - Find SpaceX URLs for sourcing data, and the free API (api.spacexdata.com/v4/) supported
    - Gather Launch Dictionary data and convert it to Pandas
    - Filter only for Falcon9 records
    - Identify missing values, and fill them with the mean values of the column
    - Write the tabular format into CSV file
- GitHub URL of the completed SpaceX API calls notebook: https://github.com/karthaxkarthax/DS/blob/main/Spacex-data-collection-api.ipynb



Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts
    - Installing required packages
    - Identify the public domain URLs that has the Falcon launches data
    - Use Beautiful Soup (Or Selenium), Web requests & Response objects, and their methods
    - Scrap the HTML content, locate tables columns using 'th' HTML tag, and values with 'tr' HTML tag, and read values
    - If some input variables are embedded as part of strings, retrieve the values onto specific input values
    - Write the tabular format into CSV file
- GitHub URL of the completed web scraping notebook:
  https://github.com/karthaxkarthax/DS/blob/main/SpaceX-webscraping.ipynb

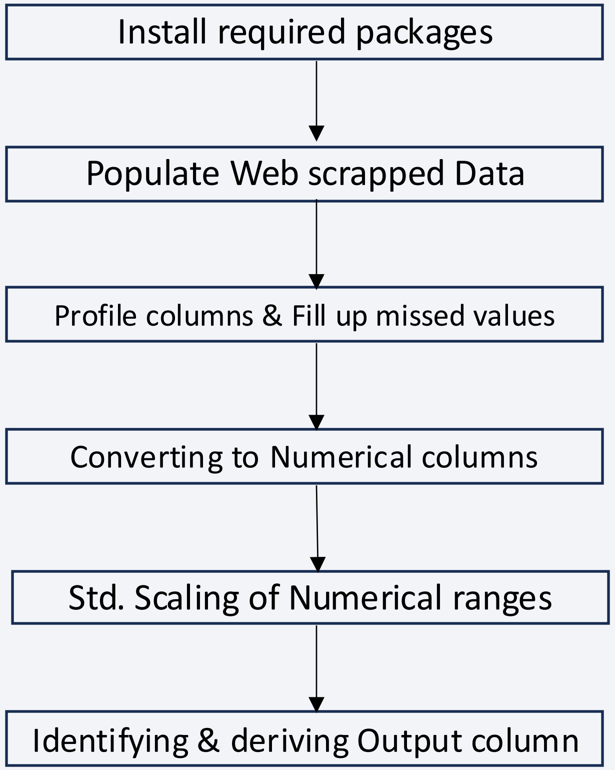
Flowchart of Data Collection - Scrapping



Data Wrangling

- Describe how data were processed
  - Installing required packages
  - Reading the Web scrapped data file (CSV) into Pandas Array
  - Identifying the missing values, and deciding on one of the following for dealing with them: Remove the records | Fill them up with median values
  - Listing the data types of each columns, and converting the categorical columns to numerical columns with encoding (Dummy or Ordinal)
  - Use standard scaler across numerical columns for normalizing each within their range
  - Identifying the column for the outcome (output variable) and the independent input variables
- GitHub URL of completed Data Wrangling:
  https://github.com/karthaxkarthax/DS/blob/main/Spacex-Data%20wrangling.ipynb

Flowchart of Data Wrangling



EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts

S#	Chart type selection	Reasoning
1	Scatter Plot with hue color coding for the following: • Flight Number vs. Launch Site • Payload vs. Launch Site • Success Rate vs. Orbit Type • Flight Number vs. Orbit Type • Payload vs. Orbit Type	• To visualize the relationship between 2 independent input variables • To find correlation patterns over outcome (using hue colors) • To identify clusters and outliers
2	Line chart for the following: • Launch Success Yearly Trend	To check on the success rate trending

- GitHub URL of completed EDA with Data Visualization:
  https://github.com/karthaxkarthax/DS/blob/main/SpaceXDataVisualization.py

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed

S#	SQL Query	Observations possible
1	SELECT * from SPACEXTBL WHERE Launch_site LIKE "CCA%"	Identifying Launch site names starting with CCA
2	SELECT SUM(PAYLOAD_MASS__KG_) from SPACEXTBL where customer = 'NASA (CRS)';	Payload, Launch Sites, Success details of NASA
3	SELECT AVG(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version = 'F9 v1.1';	Payload details for a given Booster version
4	SELECT MIN(Date) from SPACEXTBL where Landing_Outcome = 'Success (ground pad)';	Date of the first successful Ground pad landing
5	SELECT Booster_Version from SPACEXTBL where Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;	Names of boosters thaaat have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
6	SELECT COUNT(*) from SPACEXTBL where Mission_Outcome LIKE 'Fail%'; SELECT COUNT(*) from SPACEXTBL where Mission_Outcome LIKE 'Success%';	Total number of successful and failure mission outcomes
7	SELECT Booster_Version from SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);	Names of the booster which have carried the maximum payload mass
8	SELECT Landing_Outcome, COUNT(Landing_Outcome), Booster_Version, Launch_Site from SPACEXTBL WHERE Date > '2014-12-31' AND Date < '2016-01-01' AND Landing_Outcome = 'Failure (drone ship)'	Failed landings in drone ship, their booster versions, and launch site names for year 2015
9	SELECT Landing_Outcome, COUNT(Landing_Outcome) from SPACEXTBL WHERE Date > '2010-06-04' AND Date < '2017-03-20' GROUP BY (Landing_Outcome) ORDER BY COUNT (Landing_Outcome) DESC ;	Ranking landing outcomes for a given duration

- GitHub URL of completed EDA with SQL notebook: https://github.com/karthaxkarthax/DS/blob/main/SpaceX_queries.rtf

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map

S#	Objects added	Benefits of the Objects	Observations supported
1	Circle	To highlight the point of interest in the Map	Bird's eye view of all Launch sites
2	Marker	To call out the specific point of interest in terms of Site name, address, etc.	Getting the details of a given Launch site
3	Marker Cluster	Arrange multiple markers in one Circle	Looking at the details on all Launch sites in a given Circle (point of interest)
4	Mouse Position	Experimenting the map for the nearest road, city, seashore	Exploring the map for nearest road, city and seashore
5	Poly Line	Building a line from the mouse position to the launch site for displaying distance measurement	Finding the distance between the Launch site and the landmarks being explored such as Road, City, Seashore

- GitHub URL of completed interactive Folium map : https://github.com/karthaxkarthax/DS/blob/main/SpaceX_%20launch_site_locationAnalysis.ipynb

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs & interactions of the dashboard, and the reasons

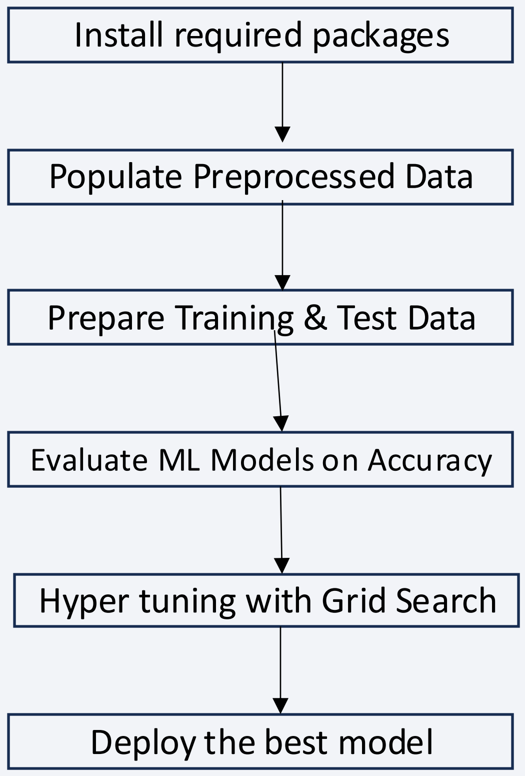
S#	Dash Theme	Interactions	Benefits
1	Launch count check across sites	Select 'ALL sites' option in the drop down	Identifying the highest launch# site, and highest success rate site, and understanding the success ratio distribution across site
2	Highest Success case site	Select a specific site option in the drop down	Identifying the success rate for the given site
3	Payload vs. Launch Outcome	Select range of payload using slider	Identifying Booster version performance within a given Payload range

- GitHub URL of completed Plotly Dash: https://github.com/karthaxkarthax/DS/blob/main/SpaceXDashAnalysis.py

Predictive Analysis (Classification)

- Steps to evaluate, improve and find the best performing classification model
  - Installing required packages
  - Moving the preprocessed data into numpy array
  - Designating the Class column as Y (output variable) and the rest as X (input variables)
  - Converting the categorical variables into numbers using encoding
  - Converting variable ranges to standard scaler scale for a fair weightage analysis of input variables
  - Splitting the dataset into Training and Testing
  - Instantiating from sklearn library, and fitting various data models with the training data
  - Testing the Model with test data and reporting the accuracy
  - Using GridSearch for fine-tuning the hyper parameters, and identifying the best mode
- GitHub URL of completed predictive analysis: https://github.com/karthaxkarthax/DS/blob/5332320fecbbf9c98c858f3c1f9fce46496bbb52/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Flowchart of predictive analysis



Results

- Exploratory data analysis results

S#	Exploration Theme	Findings & Conclusion
1	Flight Number vs. Launch Site	90 Launches are reported in this period | ~61% of the launches are from CCSFS SLC 40 | 60% of launches from CCSFS SLC 40 has been with successful outcome | Remaining 2 Launch sites have >77% of success rate | **Conclusion: Launch Site is an important input variable for deciding the landings outcome**
2	Payload vs. Launch Site	'CCAFS SLC 40' site has 60% of launches with <5000 kg payloads | 68% of the launches from 'KSC LC 39A' site has > 5000 kg payload | 'VAFB SLC 4E' site has the highest success rate at 60% of heavy payloads (>5000 kg) | **Conclusion: There is a correlation between Payload Vs. Launch Site over the landing outcome**
3	Success Rate vs. Orbit Type	Orbits of ES-L1, GEO, HEO, SSO had 8 launches together, and they were all successful | GTO had the highest launch # of 27 and had a success rate of 52% | Second highest launches (#21) were done from ISS with a success rate of 62% | **Conclusion: Different Orbits seems to have different payload requirements**
4	Flight Number vs. Orbit Type	All 14 launches made for VLEO has payloads of >13500 kg | ES-L1, GEO, SO, VLEO Orbits did not have any payloads <=5000 kg on their launches | ISS had 18 launches of <5000 kg payload and ~44% of them have failed | PO had all <5000kg payloads failing, and 86% of >5000 kg payloads making it successful | **Conclusion: Some Orbits seems to address different payload ranges**
5	Payload vs. Orbit Type	Payloads range from 350 kg to 15600 kg | Every 2 launches out of 3 have been successful VLEO had all their launches above > 13500 kg, with a success rate of >85% | For 6100kg payload GEO orbit had successful launch while SO had failed launch | **Conclusion: Some Orbits seems to have a positive landing outcome on certain payload ranges**
6	Launch Success Yearly Trend	90 Launches have been analyzed | There has been an increasing trend of successful launches in 2010 to 2020 | 2019 has the highest success ratio of 90% in the yearly launches | **Conclusion: Risks of failed landing outcome seems going down over the years**

Results

- Interactive analytics demo in screenshots

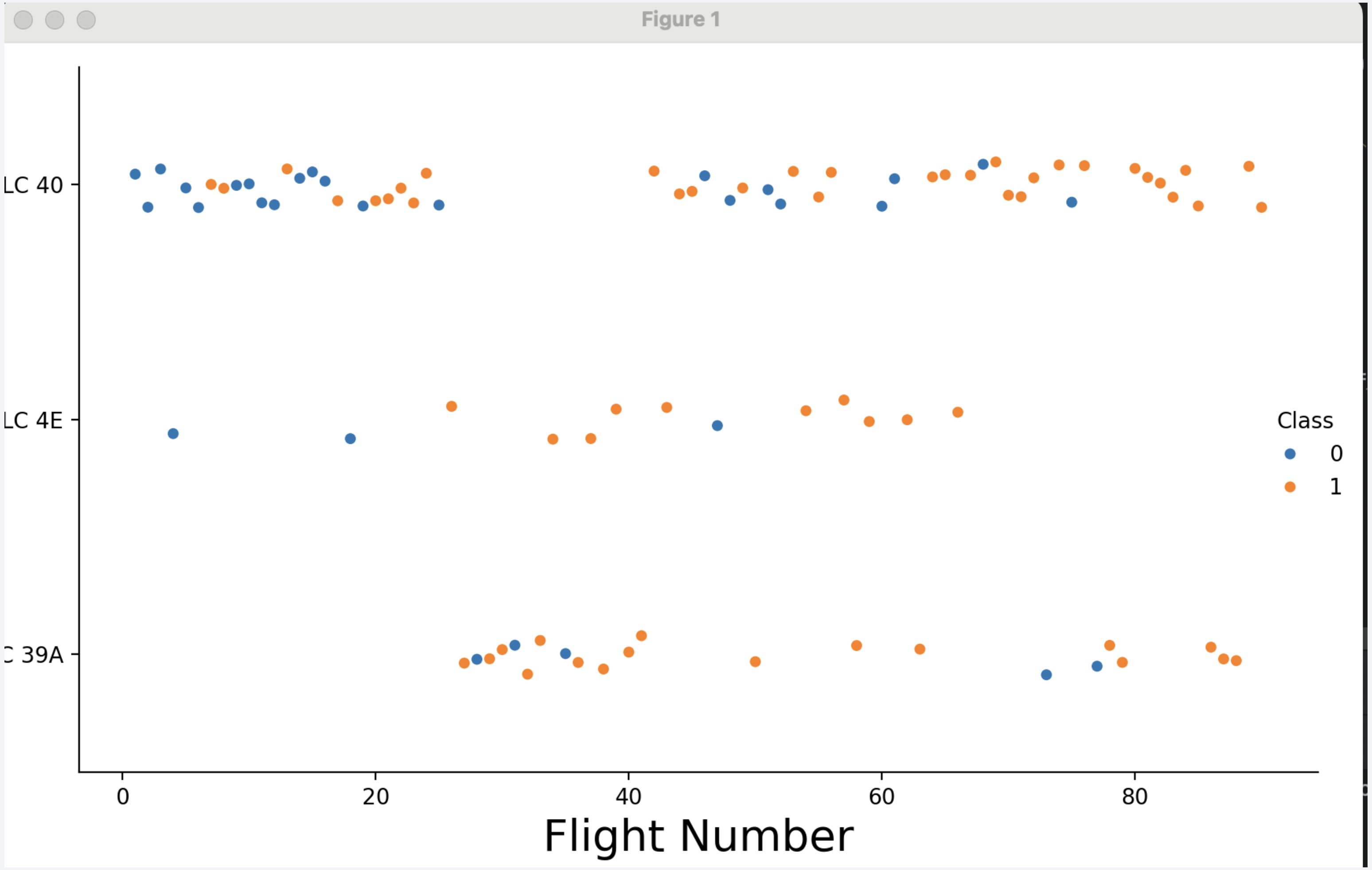
S#	Exploration Theme	Findings & Conclusion
1	Launch count check across sites	Out of all sites, CCAFS LC-40 has made the highest number of launches (#26) | CCAFS LC-40 has the lowest success rate (27%) though | KSC LC-39A has launched only half the number of launches (#13) in comparison to CCAFS LC-40, but has ~77% success rate | **Conclusion: KSC LC-39A seems to have a Geographic advantage over the landing outcome**
2	Highest Success case site	KSC LC-39A has launched 13 launches and has ~77% success rate | Each of the launches had unique boosters version | 1 out 3 Booster Versions only had failed launches | **Conclusion: Booster versions seems to influence the landing result**
3	Payload vs. Launch Outcome	52 Booster versions has been used for 56 Launches analyzed | ~43% of the Booster versions had successful launches 'F9v1.1' had 5 launches with all of them resulting in a failed outcome | **Conclusion: Booster versions seems to influence the landing result**

- Predictive analysis results
  - Decision Tree ML algorithm helped to reach an accuracy level of 0.88 in this problem statement. A GridSearch logic could further improve the ranking, through hyper tuning of ML parameters
  - Key input variables are: 'Payload Mass', 'Orbit', 'Booster Version', and Launch sites. Year of launch influences as well, possibly due to technological improvements

Section 2

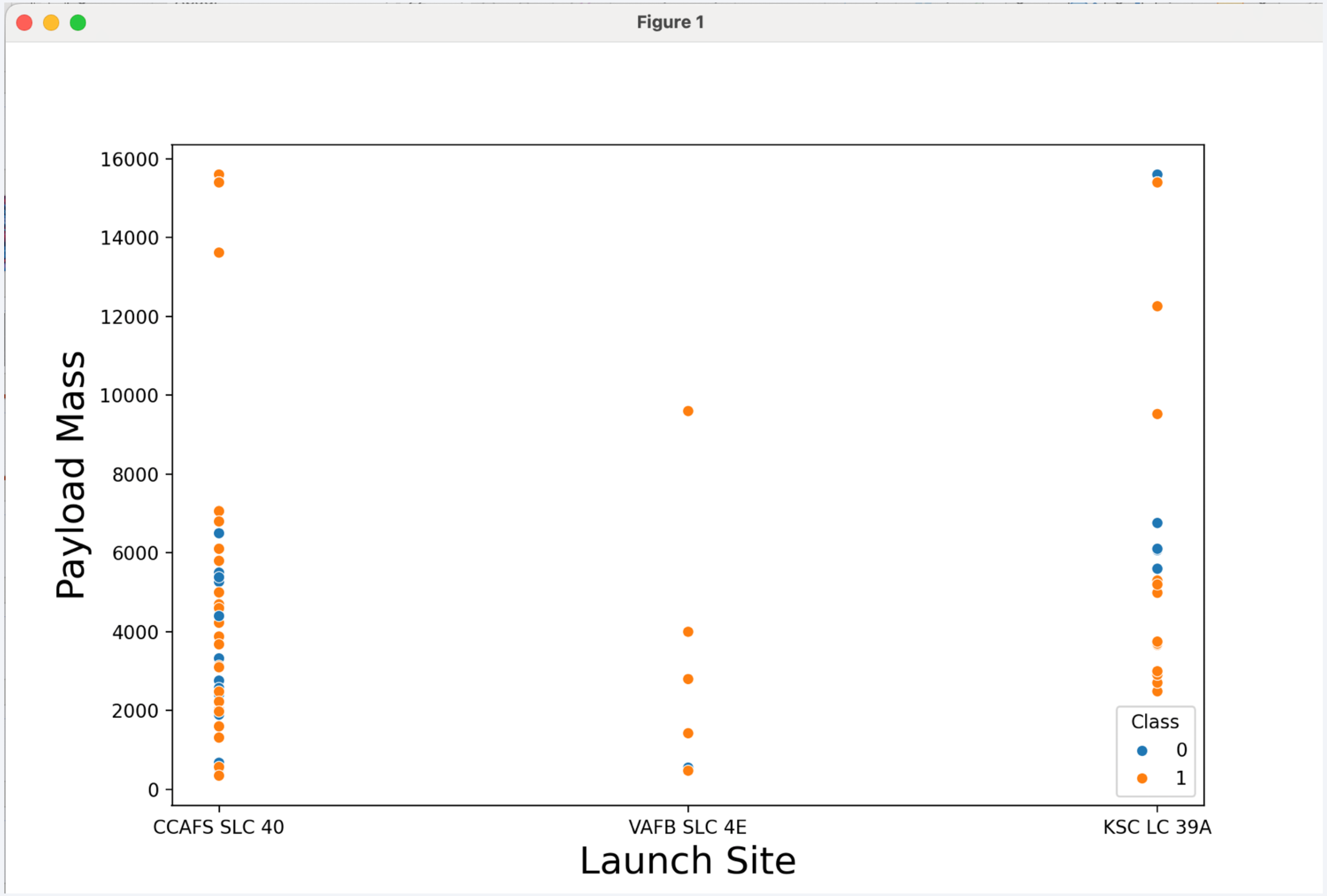
Insights drawn from EDA

Flight Number vs. Launch Site



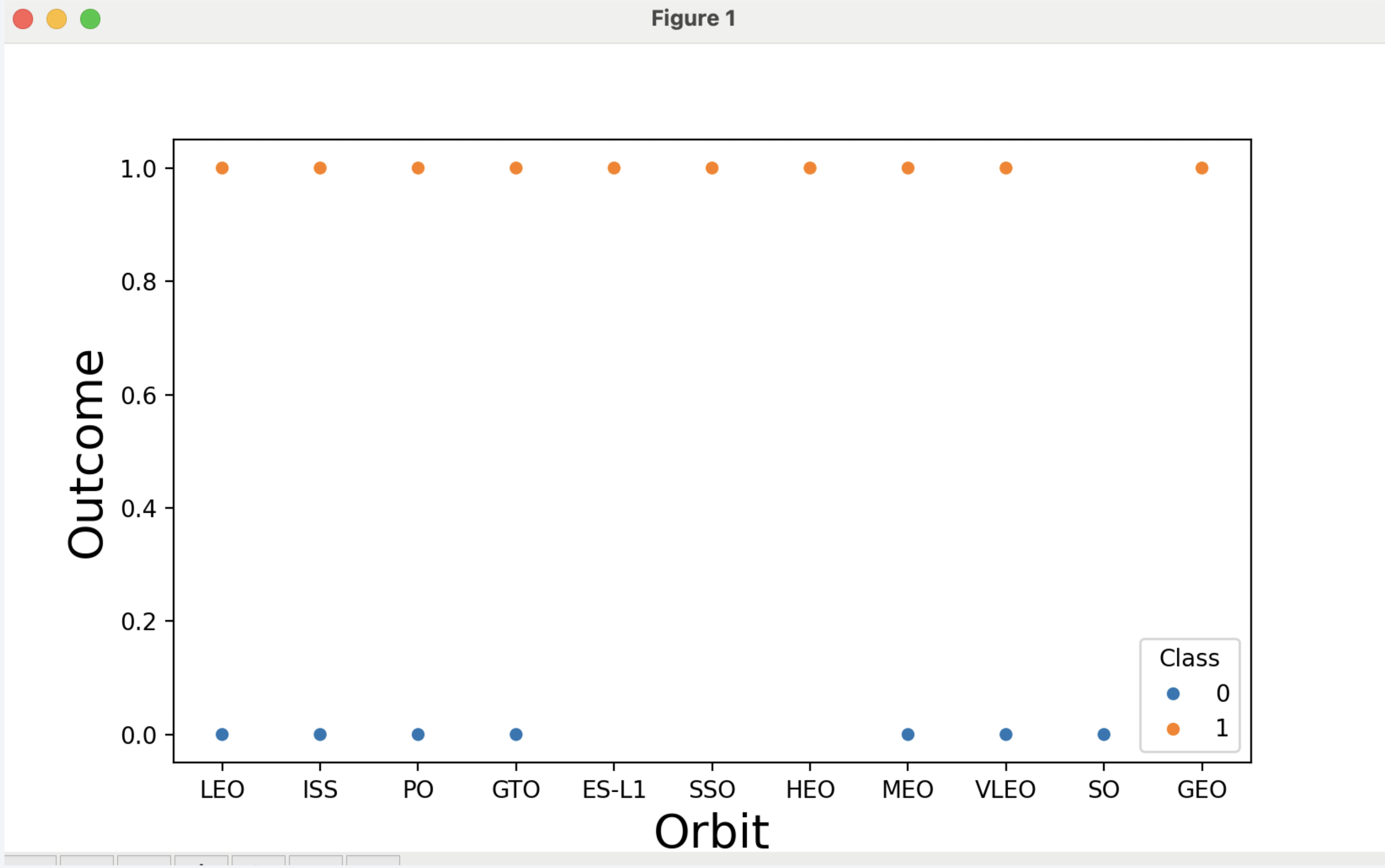
- 90 Launches are reported in this period
- ~61% of the launches are from CCSFS SLC 40
- 60% of launches from CCSFS SLC 40 has been with successful outcome
- Remaining 2 Launch sites have >77% of success rate

Payload vs. Launch Site



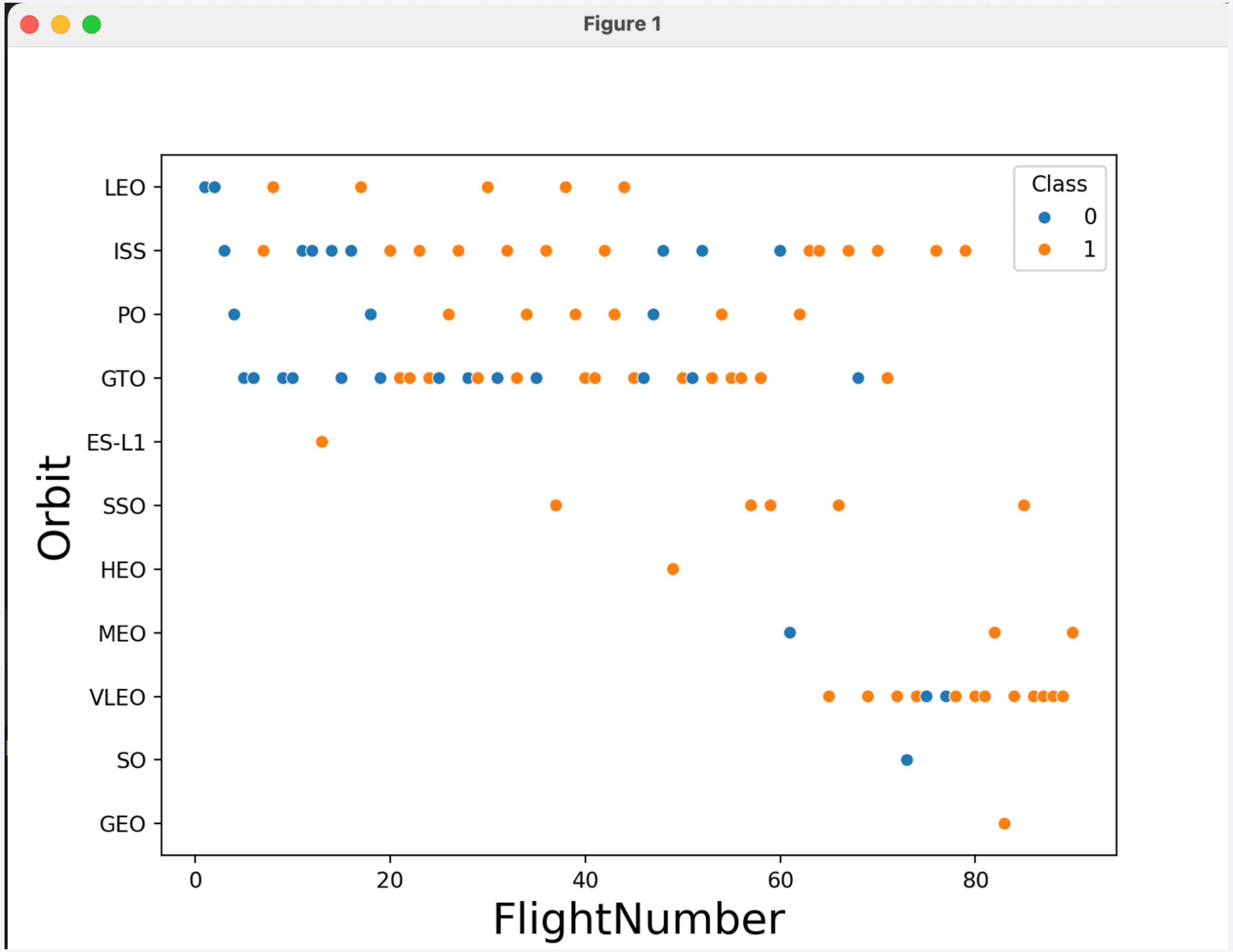
- ‘CCAFS SLC 40’ site has 60% of launches with <5000 kg payloads
- 68% of the launches from ‘KSC LC 39A’ site has > 5000 kg payload
- ‘VAFB SLC 4E’ site has the highest success rate at 60% of heavy payloads (>5000 kg)

Success Rate vs. Orbit Type



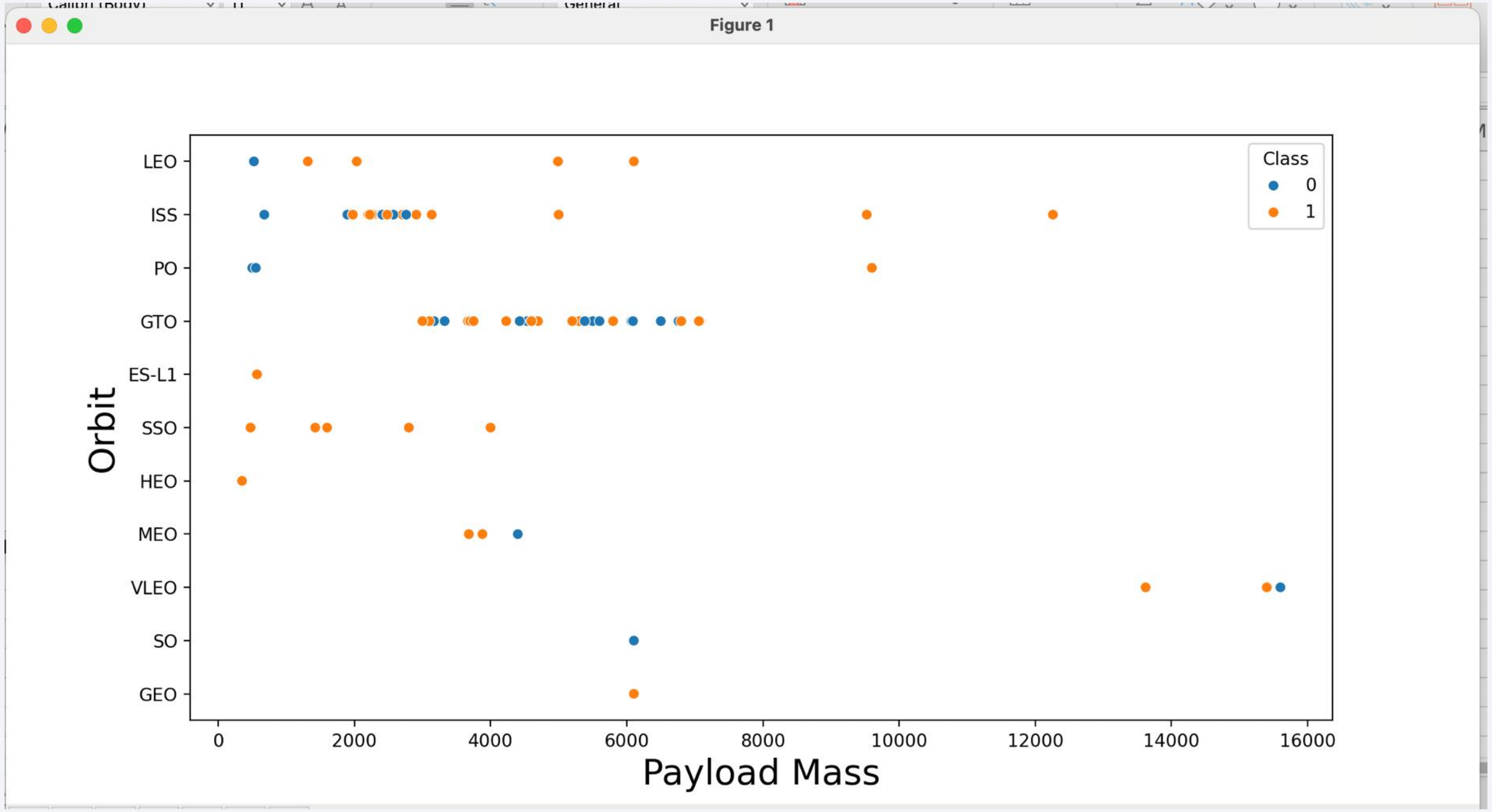
- Orbits of ES-L1, GEO, HEO, SSO had 8 launches together, and they were all successful
- GTO had the highest launch # of 27 and had a success rate of 52%
- Second highest launches (#21) were done from ISS with a success rate of 62%

Flight Number vs. Orbit Type



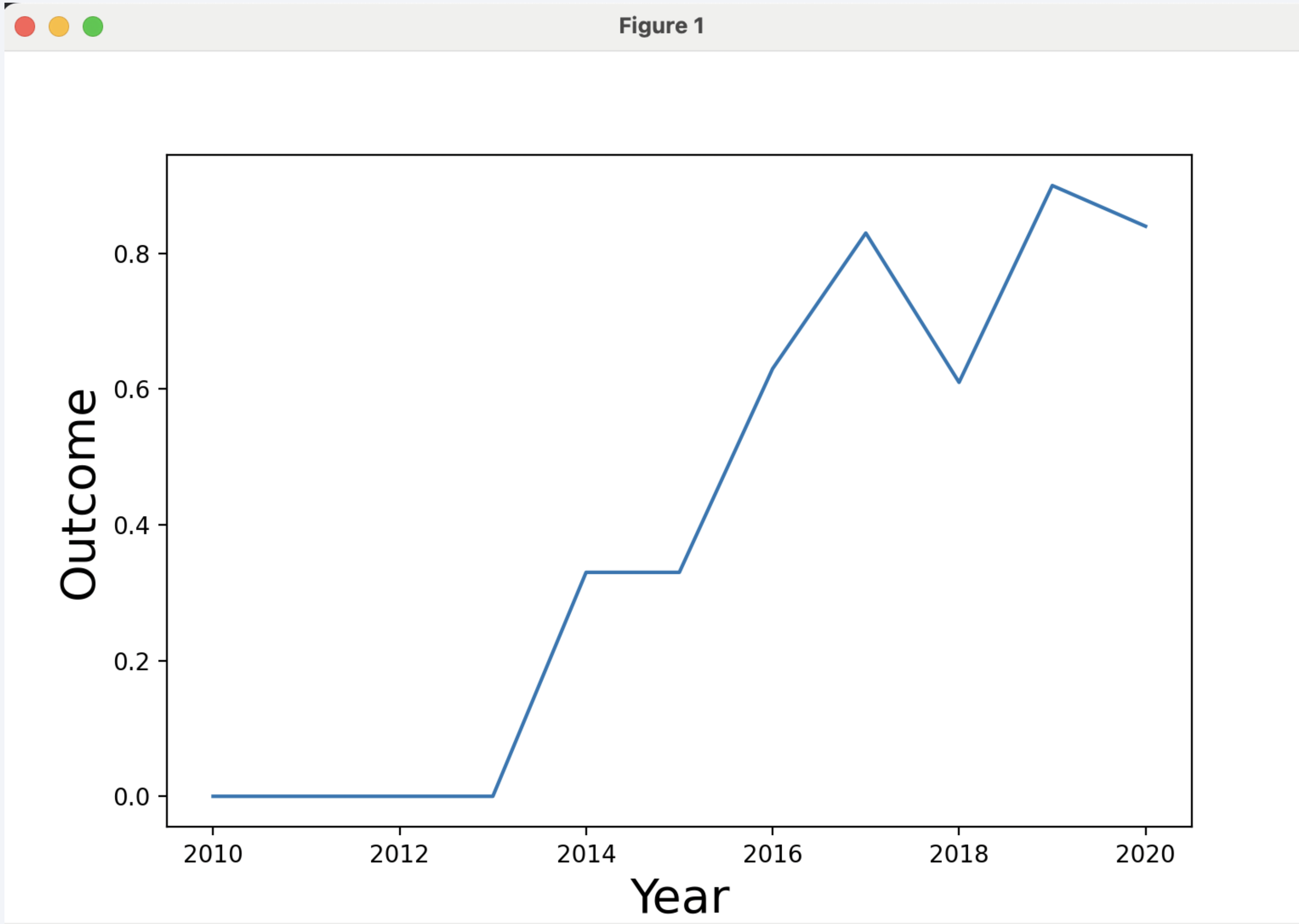
- All 14 launches made for VLEO has payloads of >13500 kg
- ES-L1, GEO, SO, VLEO Orbits did not have any payloads <=5000 kg on their launches
- ISS had 18 launches of <5000 kg payload and ~44% of them have failed
- PO had all <5000kg payloads failing, and 86% of >5000 kg payloads making it successful

Payload vs. Orbit Type



- Payloads range from 350 kg to 15600 kg
- Every 2 launches out of 3 have been successful
- VLEO had all their launches above > 13500 kg, with a success rate of >85%
- For 6100kg payload GEO orbit had successful launch while SO had failed launch

Launch Success Yearly Trend



- 90 Launches have been analysed
- There has been an increasing trend of successful launches in 2010 to 2020
- 2019 has the highest success ratio of 90% in the yearly launches

All Launch Site Names

1. CCAFS SLC 40:
    - 55 launches in total
    - 60% success rate
2. KSC LC 39A
    - 22 launches in total
    - 77% success rate
3. VAFB SLC 4E
    - 13 launches in total
    - 77% success rate

Launch Site Names Begin with 'CCA'

```
%sql SELECT * from SPACEXTBL WHERE Launch_site LIKE "CCA%"
```

- CCAFS LC-40
- CCAFS SLC-40

Total Payload Mass

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) from SPACEXTBL where customer = 'NASA (CRS)';
```

- Calculate the total payload carried by boosters from NASA:
  - 37620 kg payload
- There has been 20 payloads in a range of 0-3136 kg, and an average of 1881 kgs
- There has been only 1 out of 20 launches that failed in mission outcome
- 8 out of 20 had a successful landing outcome

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version = 'F9 v1.1';
```

- Calculate the average payload mass carried by booster version F9 v1.1
  - There has been 5 payloads in this category
  - 2928.4 kg is the average payload
  - All of them had a successful mission outcome, though their landing outcome failed
  - Launch site has been CCAFS LC-40 for all of them

First Successful Ground Landing Date

```
%sql SELECT MIN(Date) from SPACEXTBL where Landing_Outcome = 'Success (ground pad)';
```

- Find the dates of the first successful landing outcome on ground pad
  - 22/12/2015 is the first successful date of ground pad landing
  - CCAFS LC-40 is the first launch site to witness this successful ground pad landing
  - Orbit related it LEO
  - There has been 9 such landings between 2015 and 2018 with Orbits of LEO or LEO (ISS)

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT Booster_Version from SPACEXTBL where Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;
```

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
  - There are 24 of them, listed as:
  - F9 v1.1, F9 v1.1 B1011, F9 v1.1 B1014, F9 v1.1 B1016, 9 FT B1020, F9 FT B1022, F9 FT B1026, F9 FT B1030, F9 FT B1021.2, F9 FT B1032.1, F9 B4 B1040.1, F9 FT B1031.2, F9 B4 B1043.1, F9 FT B1032.2, F9 B4 B1040.2, F9 B5 B1046.2, F9 B5 B1047.2, F9 B5 B1046.3, F9 B5B1054,F9 B5 B1048.3, F9 B5 B1051.2 , F9 B5B1060.1, F9 B5 B1058.2 , F9 B5B1062.1
  - They all had a successful outcome and 62.5% of landing success

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT COUNT(*) from SPACEXTBL where Mission_Outcome LIKE 'Fail%';

%sql SELECT COUNT(*) from SPACEXTBL where Mission_Outcome LIKE ‘Success%';
```

- Calculate the total number of successful and failure mission outcomes
  - 100 out of 101 had successful outcome and 1 failed
  - Single failed launch was done from CCAFS LC-40
    - Booster version used is. F9 v1.1 B1018
    - Payload is 1952 kg

Boosters Carried Maximum Payload

```
%sql SELECT Booster_Version from SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT
MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

- List the names of the booster which have carried the maximum payload mass
  - 12 Booster Versions carried a maximum payload of 13600 kgs. They are listed as:
  - 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
  - They all were targeted for an Orbit of LEO
  - They all were launched for the Internal customer: SPACEX

2015 Launch Records

```
%sql SELECT Landing_Outcome, COUNT(Landing_Outcome), Booster_Version, Launch_Site from SPACEXTBL
WHERE Date > ‘2014-12-31' AND Date < ‘2016-01-01’ AND Landing_Outcome = ‘Failure (drone ship)’
```

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - There were only 2 records with a landing outcome of ‘Failure (drone ship)’ in 2015
  - CCAFS LC-40 is the only Launch site involved
  - Launches were done over LEO (ISS) Orbit
  - Payloads were 1898 and 2395 kgs respectively
  - Both launches are considered mission successful, though their landing outcome were failures

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

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

```
%sql SELECT Landing_Outcome, COUNT(Landing_Outcome) from SPACEXTBL WHERE Date > '2010-06-04'
AND Date < '2017-03-20' GROUP BY ( Landing_Outcome) ORDER BY COUNT (Landing_Outcome) DESC ;
```

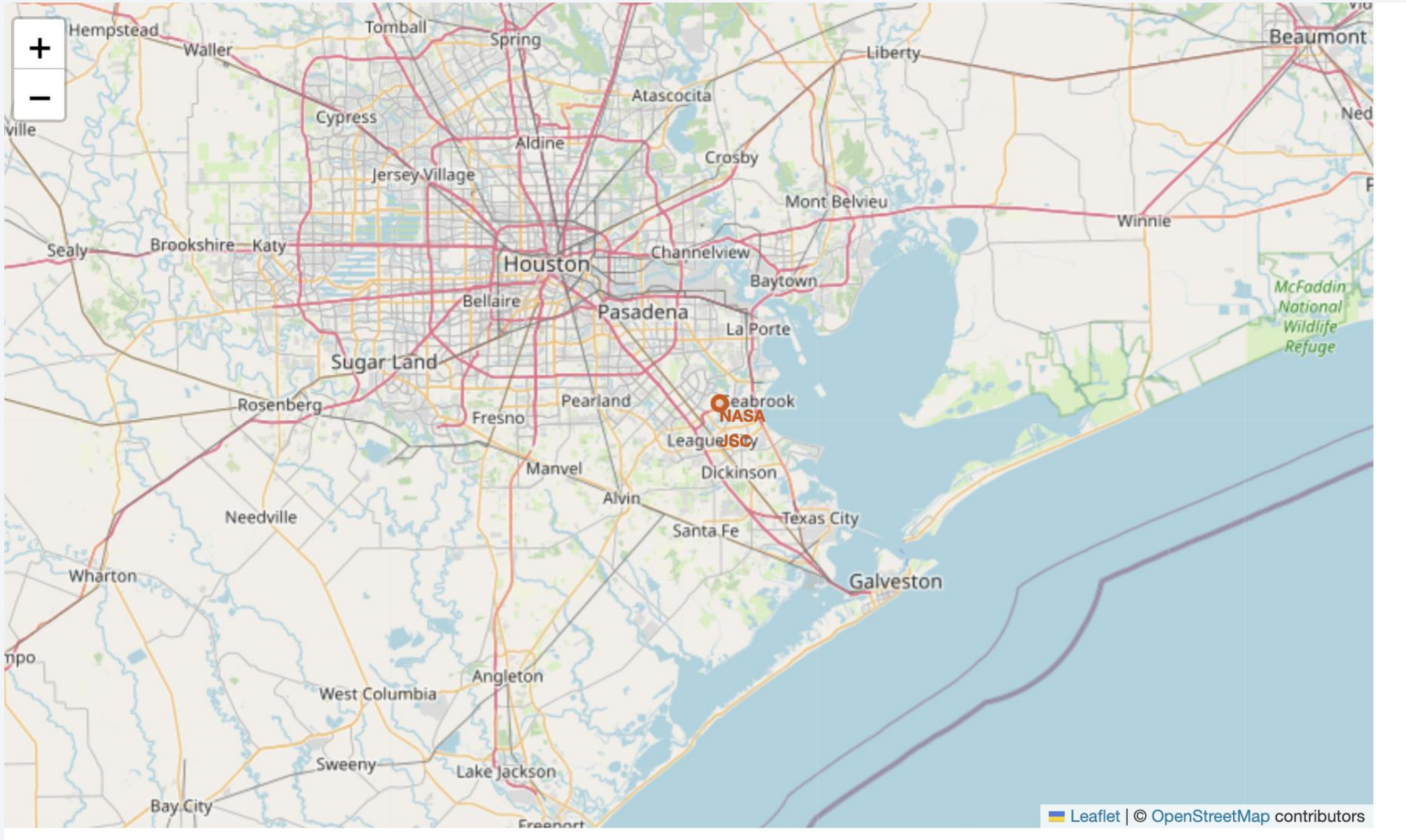
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order
  - 31 Records are found in this duration
  - 26% corresponds to successful landing
  - 35% landings were attempted in Drone ship with almost half of it failing
  - 16% landings were attempted to fall in ocean

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

Section 3

Launch Sites Proximities Analysis

Folium Map Screenshot 1: NASA Launch site



- Marking is done in Red

<Folium Map Screenshot 2>

- Replace <Folium map screenshot 2> title with an appropriate title
- Explore the folium map and make a proper screenshot to show the color-labeled launch outcomes on the map
- Explain the important elements and findings on the screenshot

<Folium Map Screenshot 3>

- Replace <Folium map screenshot 3> title with an appropriate title

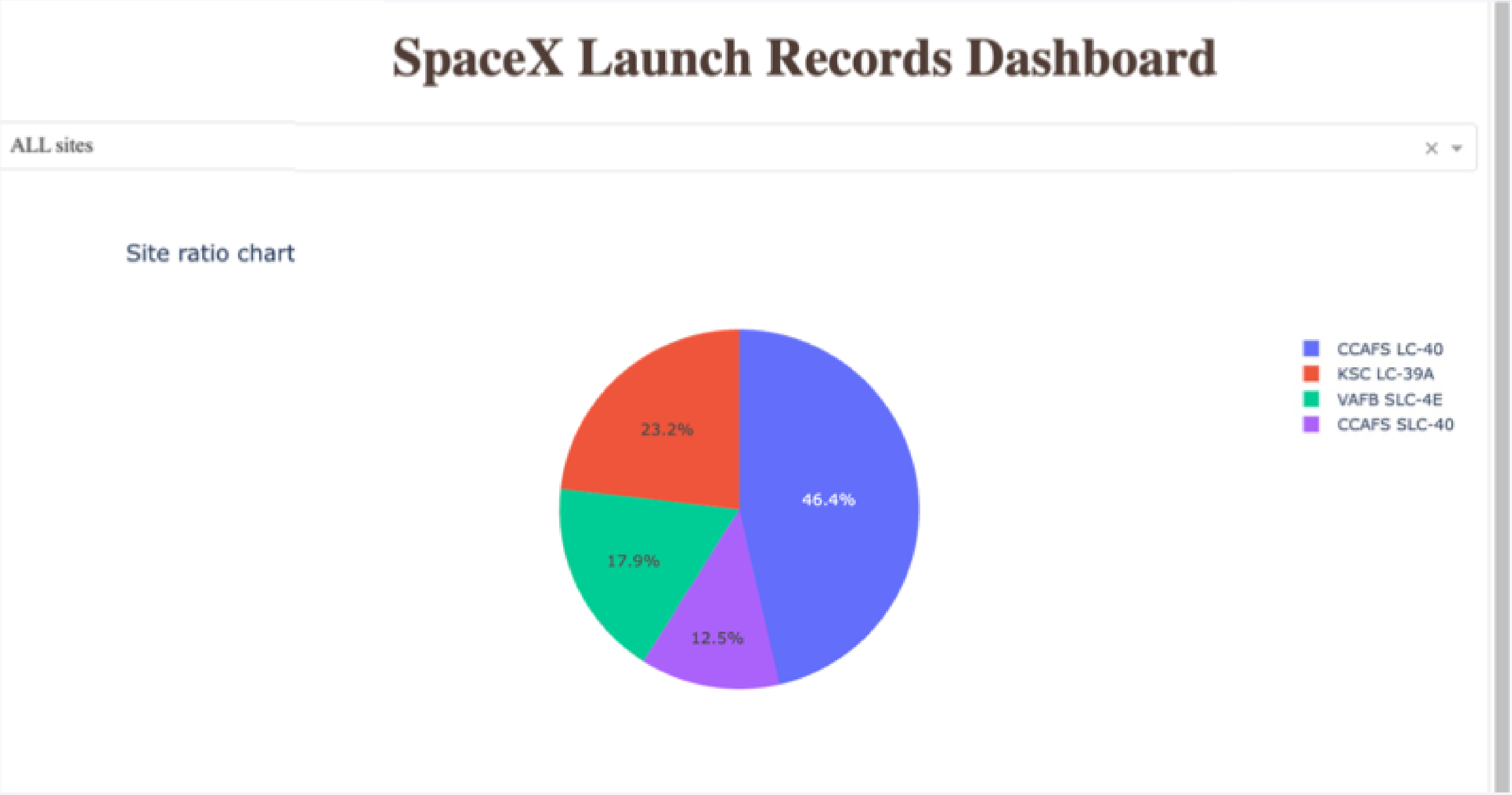
- Explore the generated folium map and show the screenshot of a selected launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed

- Explain the important elements and findings on the screenshot

Section 4

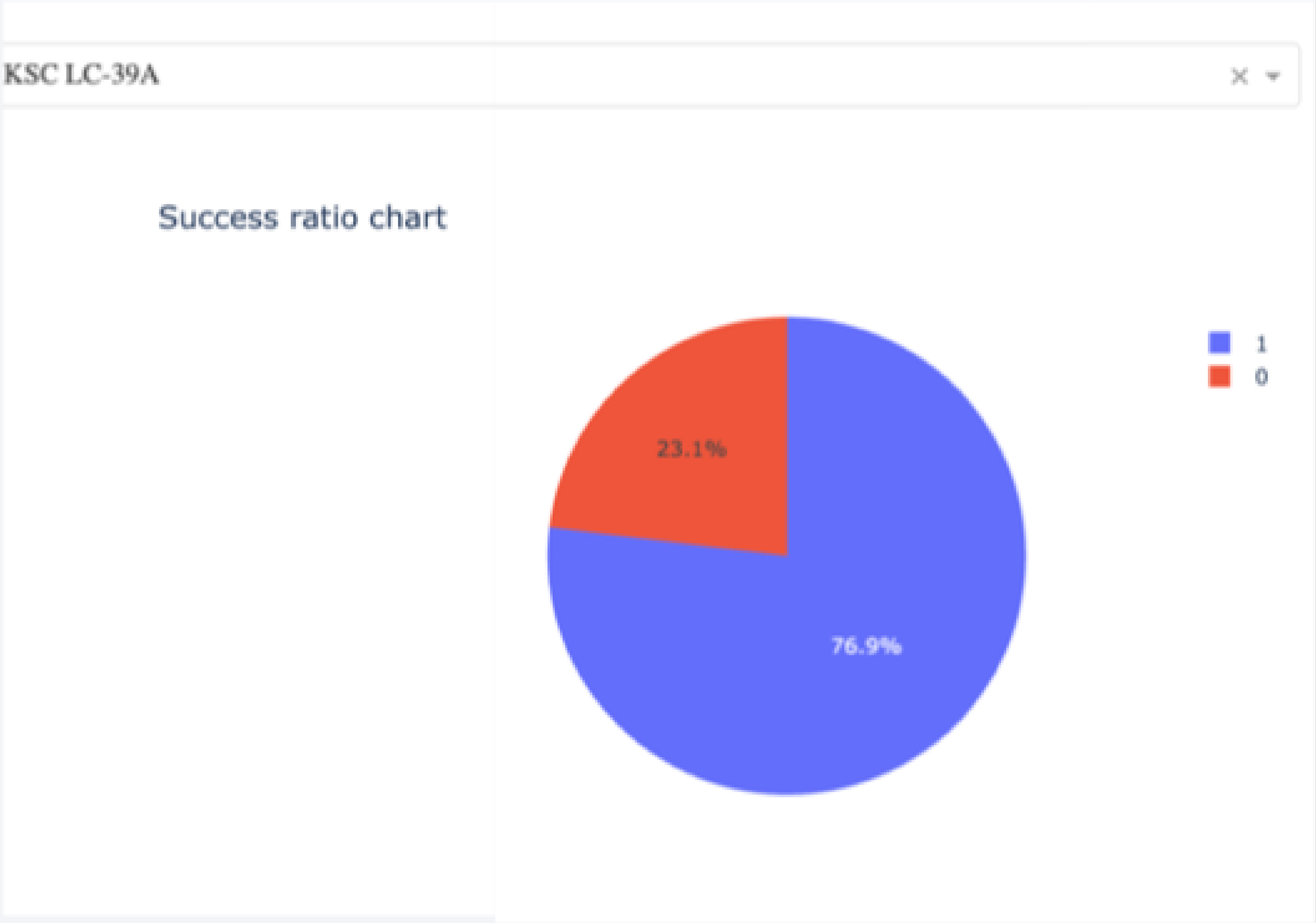
Build a Dashboard with Plotly Dash

SpaceX Launch Records Dashboard: Launch count check across sites



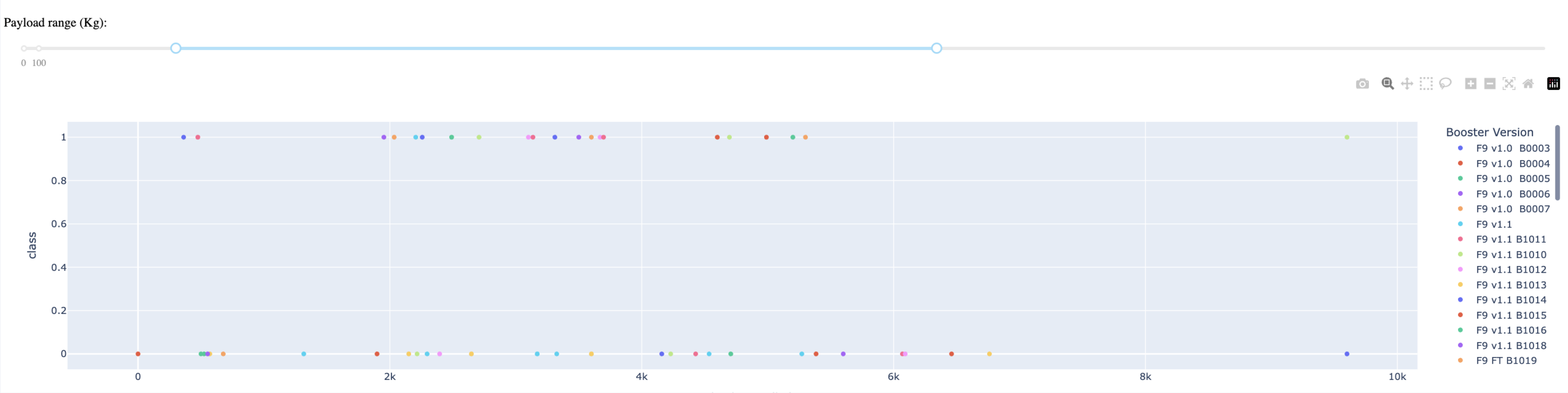
- Out of all sites, CCAFS LC-40 has made the highest number of launches (#26)
- CCAFS LC-40 has the lowest success rate (27%) though
- KSC LC-39A has launched only half the number of launches (#13) in comparison to CCAFS LC-40, but has ~77% success rate.

SpaceX Launch Records Dashboard: Highest Success case site



- KSC LC-39A has launched 13 launches and has ~77% success rate
- Each of the launches had an unique booster version
- 1 out 3 Booster Versions only had failed launches

SpaceX Launch Records Dashboard: Payload vs. Launch Outcome

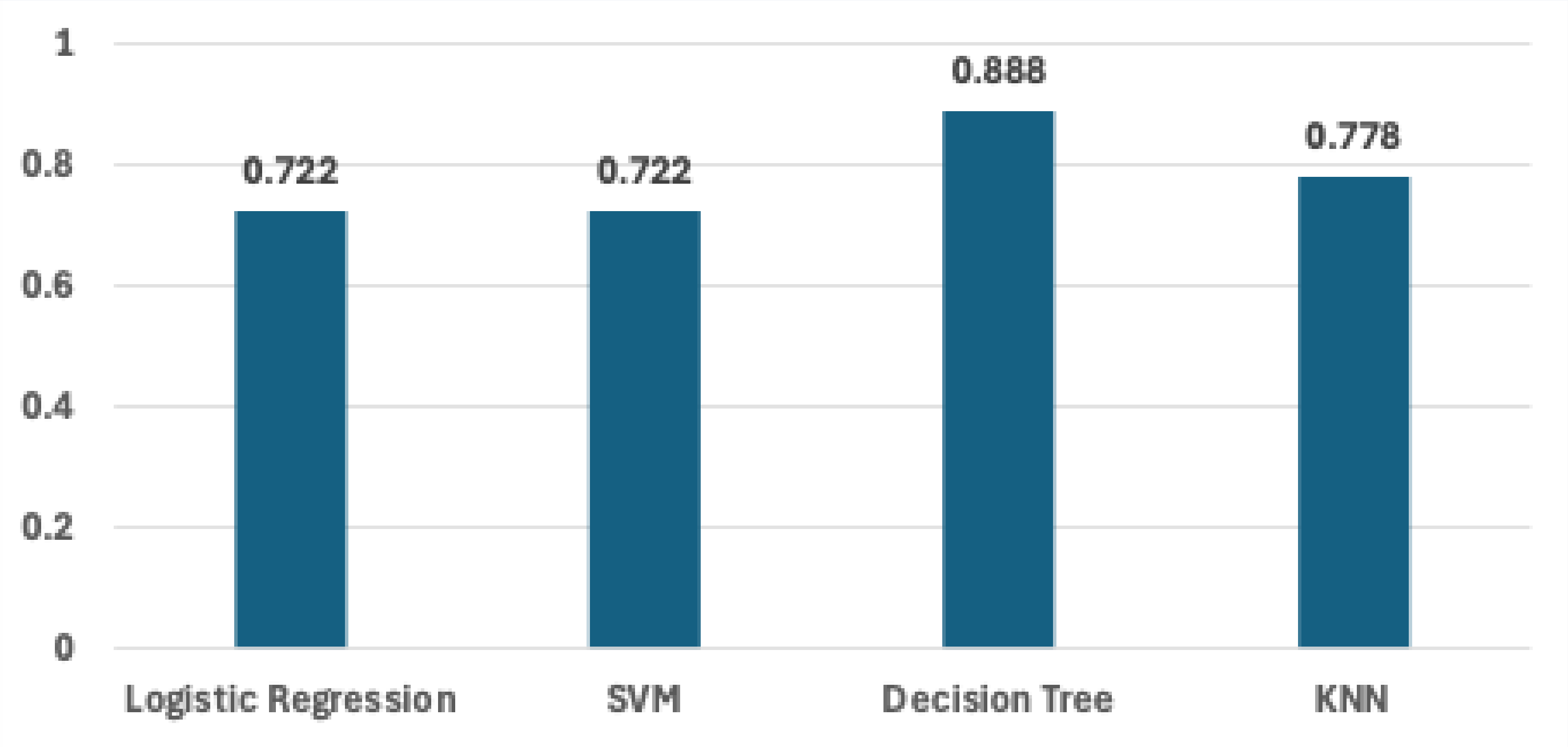


- 52 Booster versions has been used for 56 Launches analyzed
- ~43% of the Booster versions had successful launches
- 'F9v1.1' had 5 launches with all of them resulting in a failed outcome

Section 5

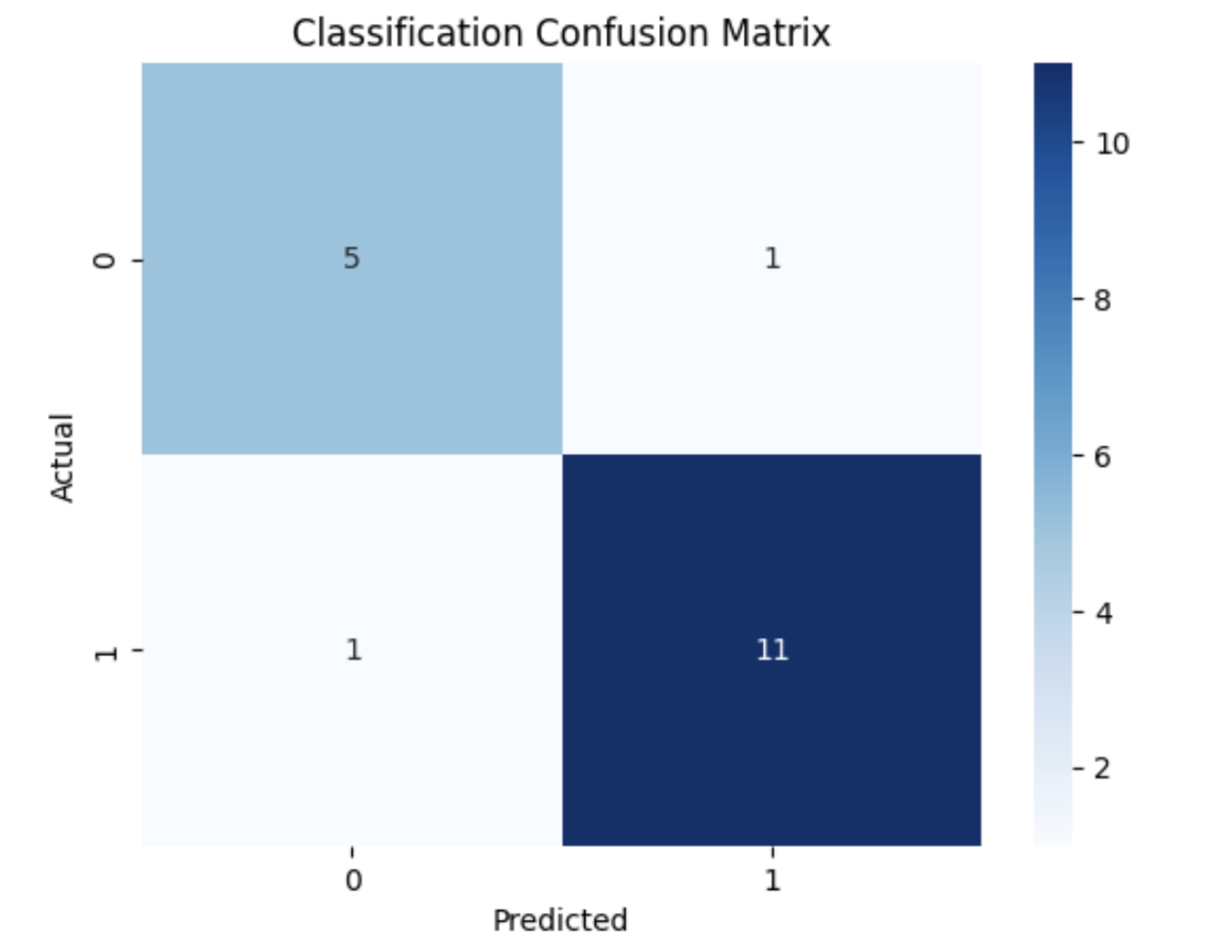
Predictive Analysis (Classification)

Classification Accuracy



- Decision Tree has an acceptable level of accuracy by being > 0.8
- Logistic Regression and SVM have < 0.75 as the classification accuracy

Confusion Matrix



- Decision Tree has the highest accuracy of 0.88
- Out of 18 entries tested, 1 false positive, and 1 false negative observed

Conclusions

1. Logistic regression, SVM, KNN have < 0.8 accuracy
2. Their accuracy levels are increasing as we finetune their Hyper parameters using GridSearchCV
3. Decision Tree seems an optimal model, to be used for Launch outcome prediction
4. Key input identified variables are: ‘Payload Mass’, ‘Orbit’, ‘Booster Version’, and Launch sites. Additional independent features and more records could improve the accuracy further
5. Year of launch influences as well, possibly due to technological improvements

Thank you!