IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Griff Servo
May 20, 2023

Outline

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

Executive Summary

- Methodologies used:
    - Data collection via API calls and webscraping
    - Data wrangling and Exploratory Data Analysis using Python: numpy, pandas libraries
    - Data visualization via Folium
    - Dashboard construction via Plotly Dash
    - Machine learning predicting using Python: seaborn, sklearn libraries
- Summary of all results:
    - SpaceX is successful in landing 66.66% of its launches since the company's inception
    - Success rates have been trending upwards since its initial launch
    - Launches performed at Kennedy Space Center Launch Complex 39 have a 76.9% success rate
    - Accuracy rate of all predictive classification models were identical

Introduction

- Project background and context
  - SpaceY is looking to determine the price of each of SpaceX's launches, to prepare competing bids
  - Launch costs are drastically decreased upon successful landing of a rocket's first stage
  - Publicly available SpaceX launch data was used for this project

- Problems you want to find answers
  - What launch conditions/variables are required to maximize the chance of a successful first-stage rocket landing?
  - Which controllable factors contribute to minimizing launch costs?
  - Can existing SpaceX data confidently predict the results of future launches?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data for this project was collected using an API to extract from spacexdata.com
  - Helper functions were utilized to construct a dataset appropriate for analysis
- Perform data wrangling
  - Data was cleaned of missing values via replacement with the mean of that variable
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Sklearn library used to split data for classification model training and testing
  - Seaborn library used to visualize confusion matrix for classification model review

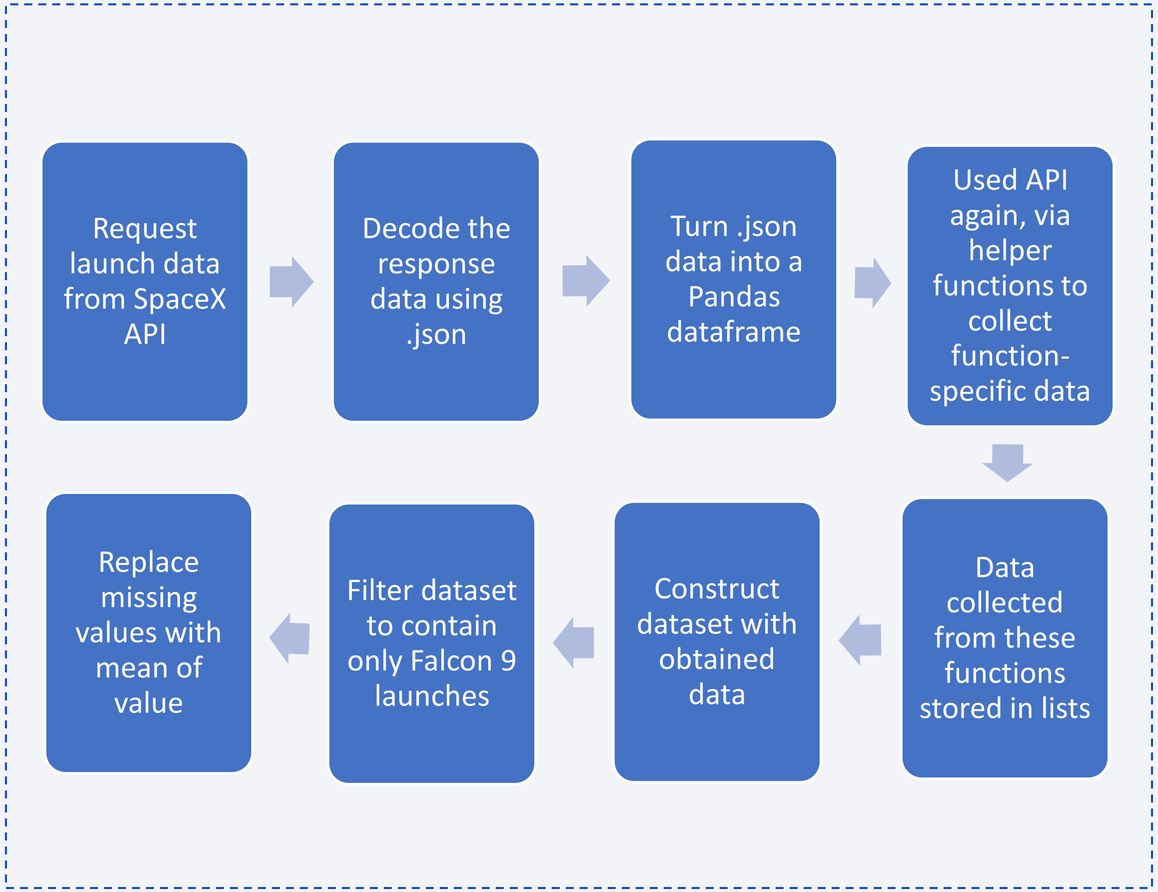
Data Collection

- Data sets were collected by REST API calls
- Using identification numbers in the raw launch data, helper functions were used with the API to extract further information such as:
  - Booster name
  - Launch site (including longitude and latitude)
  - Payload (including payload mass and orbit type)
  - Landing outcomes and types of landing
  - Various launch core-related data (Gridfins, core reuse rate, serial number of core, etc.)

Data Collection – SpaceX API

- A dataset was constructed via all the data collected.
- Data collected from API calls was cleaned and filtered. This process includes:
  - Filtering out Falcon 1 launches
  - Replacing missing values with the dataset mean of that value field

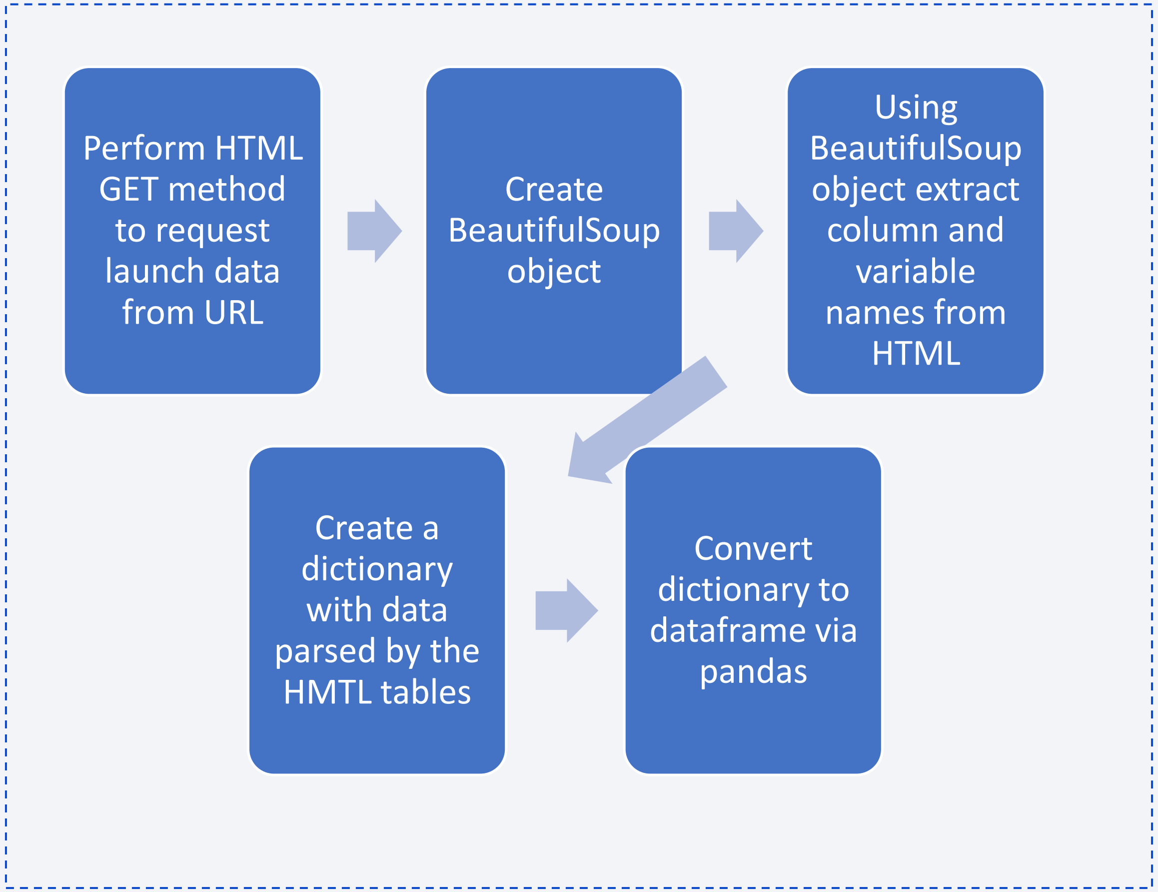
- https://github.com/bswervo/capstone/blob/65a4a59bdbe5e354761d42652dd534024a8883ef/WK1-LAB1-jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- Python libraries utilized:
  - Pandas
  - BeautifulSoup4
- Static URL data was scraped from:
  - https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

- https://github.com/bswervo/capstone/blob/65a4a59bdbe5e354761d42652dd534024a8883ef/WK1-LAB2-jupyter-labs-webscraping.ipynb



Data Wrangling

SpaceX dataset loaded → Exploratory data analysis performed → Training labels determined → Landing outcome label created → Success rate calculated via landing outcome

- Exploratory data analysis includes:
  - Calculating number of launches on each site
  - Calculating number and occurrence of each orbit type
  - Calculating number and occurrence of mission outcome per orbit type

- https://github.com/bswervo/capstone/blob/65a4a59bdbe5e354761d42652dd534024a8883ef/WK1-LAB3-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

EDA with Data Visualization

- Charts plotted and reasoning:
    - Flight Number vs. Launch Site – To plot a chronological trend of locations used
    - Payload vs. Launch Site – To predict best fitting launch site with specific payload
    - Success Rate vs. Orbit Type – To predict optimal orbit type for any future launches
    - Flight Number vs. Orbit Type – To plot a chronological trend of orbit types used
    - Payload vs. Orbit Type – To plot a correlation between these factors
    - Launch Success Yearly Trend – To re-assure stakeholder's over concerns of launch failures
- https://github.com/bswervo/capstone/blob/65a4a59bdbe5e354761d42652dd534024a8883ef/WK2-LAB2-jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- A summary of SQL queries performed:
  - All Launch Site Names
  - Launch Site Names Beginning with 'CCA'
  - Total payload mass launched for NASA (CRS)
  - Average payload mass of booster version F9 v1.1
  - First successful ground landing date
  - Successful drone ship landing with payload between 4000 and 6000KG
  - Total number of successful and failure mission outcomes
  - Booster versions having carried maximum payload
  - 2015 launch records – drone ship landing failures
  - Landing outcomes ranked - between 2010-06-04 and 2017-03-20

- https://github.com/bswervo/capstone/blob/65a4a59bdbe5e354761d42652dd534024a8883ef/WK2-LAB1-jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

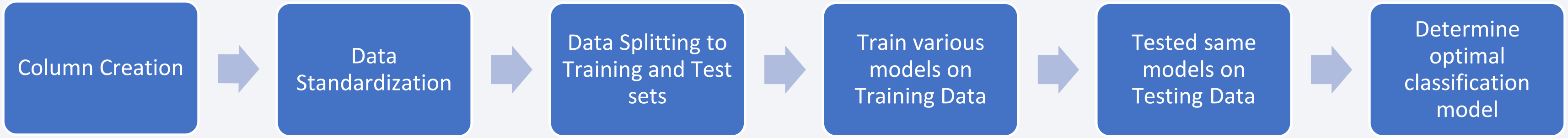
- Map Objects Added:
  - Markers depicting SpaceX launch site locations in the United States
    - To give a broad overview of SpaceX's chosen areas of operations
  - Cluster markers indicating successful and unsuccessful launches at each launch site
    - To aid in determining most successful launch site selection
  - Line markers indicating distance from launch sites to nearby areas such as:
    - Railways – To aid in logistics and payload transportation planning
    - Ocean coasts – To ensure to stakeholders no risk of damage to environment or civilian populations
- https://github.com/bswervo/capstone/blob/bf6e923a57afcfb37e245ad3f268a80f1ab6632e/WK3-LAB1-jupyter_launch_site_location.jupyterlite.ipynb

Build A Dashboard with Plotly Dash

- Visuals included:
    - Launch Success Counts – All Sites
        - To briefly visualize the success rate per launch site, to determine which site to further investigate
    - Highest Success Count – Specific Launch Site
        - To identify outlier launch site success for further analysis
    - Payload vs Launch Outcome – All Sites
        - To determine and visualize any potential correlations between the success of a launch and the mix of launch site, payload size, and booster type
- https://github.com/bswervo/capstone/blob/bf6e923a57afcfb37e245ad3f268a80f1ab6632e/WK3-LAB2-spacex_dash_app.py

Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model
  - Created a column for class – (successful/unsuccessful launch)
  - Standardized data
  - Split data into training data and test data
  - Models are trained with training data
  - Models are tested with testing data
  - Optimal model is selected



- https://github.com/bswervo/capstone/blob/bf6e923a57afcfb37e245ad3f268a80f1ab6632e/WK4-LAB1-SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

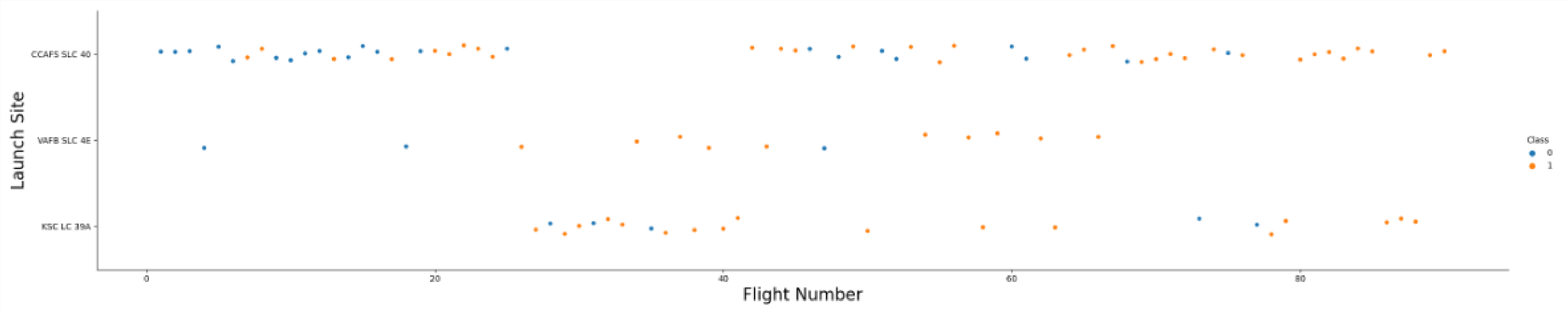
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

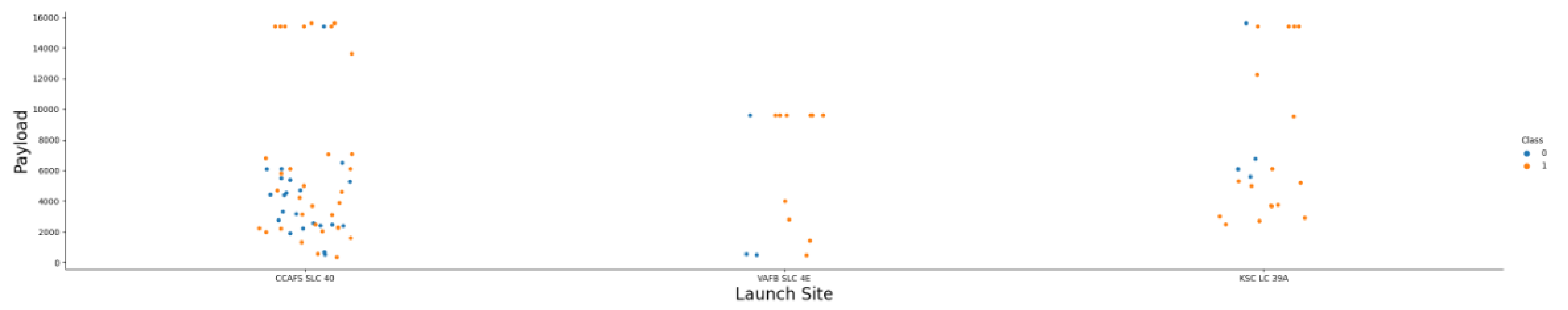
Flight Number vs. Launch Site

- Orange icons indicate successful launches, blue icons indicate unsuccessful ones.
- CCAPS SLC 40 and KSC LC 39A have consistently been used since SpaceX's initial launches.
- CCAPS SLC 40 is the company's most-used launch site.



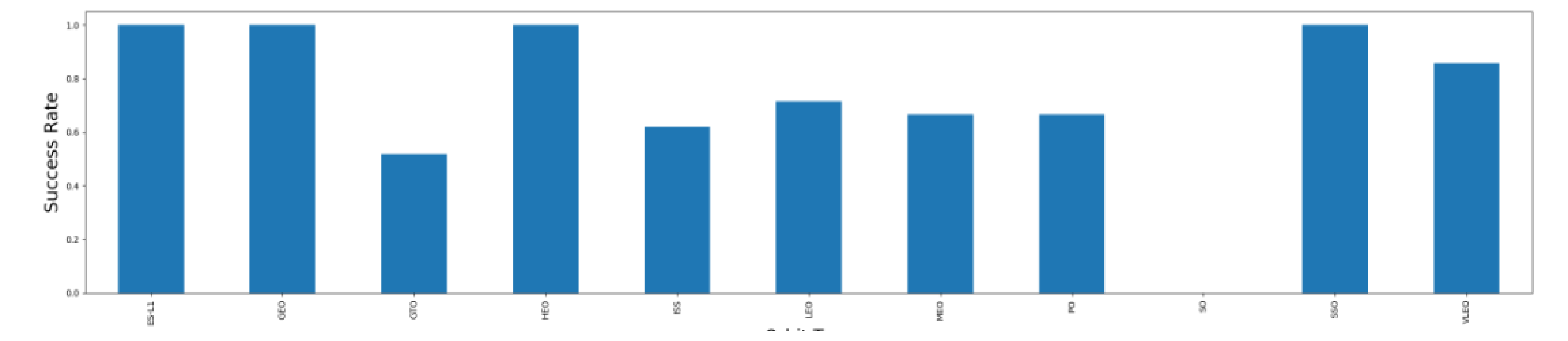
Payload vs. Launch Site

- Heaviest payloads have been launched from CCAPS SLC 40 and KSC LC 39A
- VAFB SLC 4E has had high-rates of success but a low volume of launches.



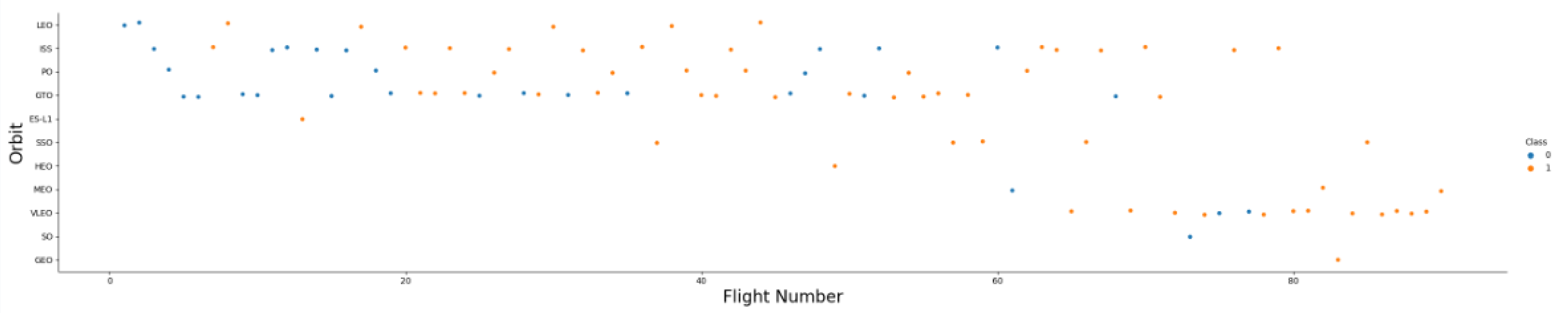
Success Rate vs. Orbit Type

- Y-axis is Success rate, X-axis is Orbit type



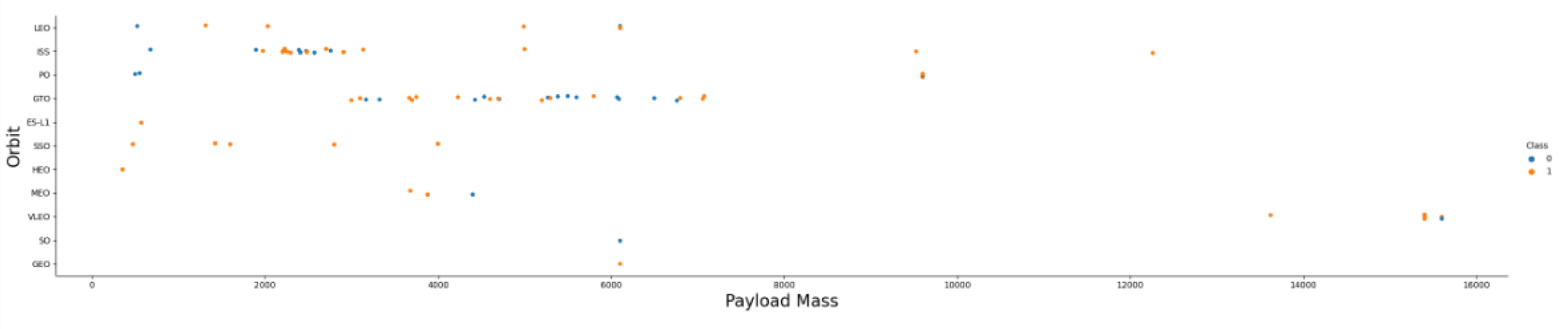
Flight Number vs. Orbit Type

- As flight numbers increase the orbit type mix has changed



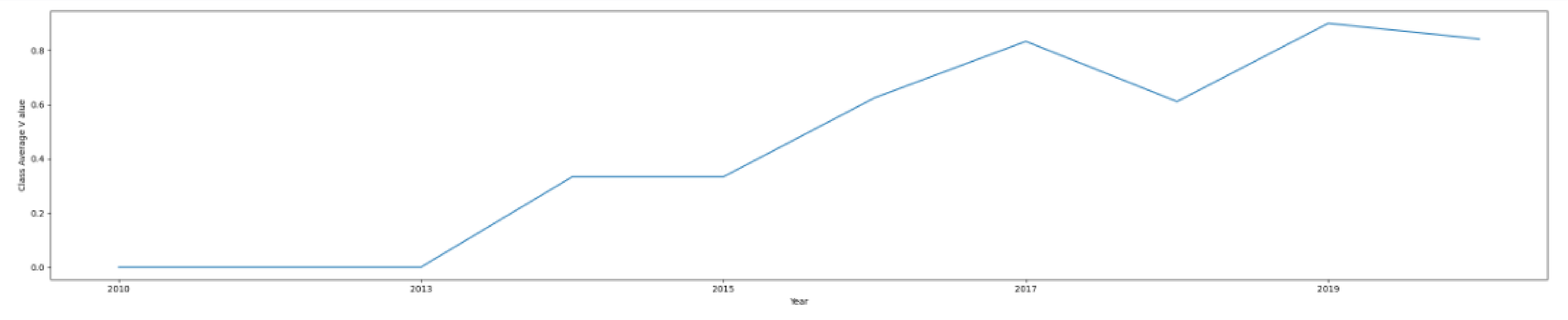
Payload vs. Orbit Type

- Large payload launches are associated with VLEO orbit types
- Small payload launches are associated with ISS and GTO orbit types



Launch Success Yearly Trend

- Success rates have been increasing since 2010



All Launch Site Names

- From the data collected we can determine SpaceX has four unique launch sites

```
[14]: %sql SELECT DISTINCT Launch_Site FROM SPACEXTBL
 * sqlite:///my_data1.db
Done.
```

[14]:

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

Launch Site Names Beginning with 'CCA'

- As results are in chronological order, we see that the first five results return CCAFS LC-40, CCAFS SLC-40 launches are not reflected

```
%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
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- For the data range collected SpaceX had launched over 45,000KG of payload mass for NASA

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

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

SUM(PAYLOAD_MASS__KG_)	Customer
45596.0	NASA (CRS)

Average Payload Mass by F9 v1.1

- This query shows the average payload mass of the F9 booster series is 2534KG

```
%sql SELECT AVG(PAYLOAD_MASS__KG_), Booster_Version AS AveragePayloadMass FROM SPACEXTBL WHERE Booster_Version LIKE 'F9 v1.1%'

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

AVG(PAYLOAD_MASS__KG_)	AveragePayloadMass
2534.6666666666665	F9 v1.1 B1003

First Successful Ground Landing Date

- The first successful landing on a ground pad was 01/08/2018, the payload status was unclear

```
%sql SELECT MIN(Date), Mission_Outcome FROM SPACEXTBL WHERE Landing_Outcome = 'Success (ground pad)'
```

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

MIN(Date)	Mission_Outcome
01/08/2018	Success (payload status unclear)

Successful Drone Ship Landing with Payload between 4000 and 6000KG

- Four versions of the F9 booster have successfully landed to a drone ship with a payload between 4000 and 6000kg

```
%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

Total Number of Successful and Failure Mission Outcomes

- Mission Outcomes are overwhelmingly successful

```
%sql SELECT COUNT(Mission_Outcome), Mission_Outcome FROM SPACEXTBL GROUP BY Mission_Outcome
```

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

COUNT(Mission_Outcome)	Mission_Outcome
0	None
1	Failure (in flight)
98	Success
1	Success
1	Success (payload status unclear)

Boosters Carried Maximum Payload

- We see that the F9 B5 booster-series carries the largest payload

```
%sql SELECT DISTINCT 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

- We see two failures to land to the drone ship in 2015

```
%sql SELECT substr(Date,4,2) as month, booster_version, launch_site, landing_outcome, Date FROM SPACEXTBL
WHERE landing_outcome = 'Failure (drone ship)' and substr(Date,7,4)='2015'
```

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

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

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

- We see that successful landings from this time period are an equal mix of ground pad landings and drone ship landings

```
%sql SELECT Landing_Outcome, COUNT(Landing_Outcome) AS LANDING_OUTCOME_COUNT FROM SPACEXTBL
WHERE substr(Date,7,4) || substr(Date,4,2) || substr(Date,1,2) between '20100604' AND '20170320'
GROUP BY Landing_Outcome ORDER BY COUNT(Landing_Outcome) DESC
```

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

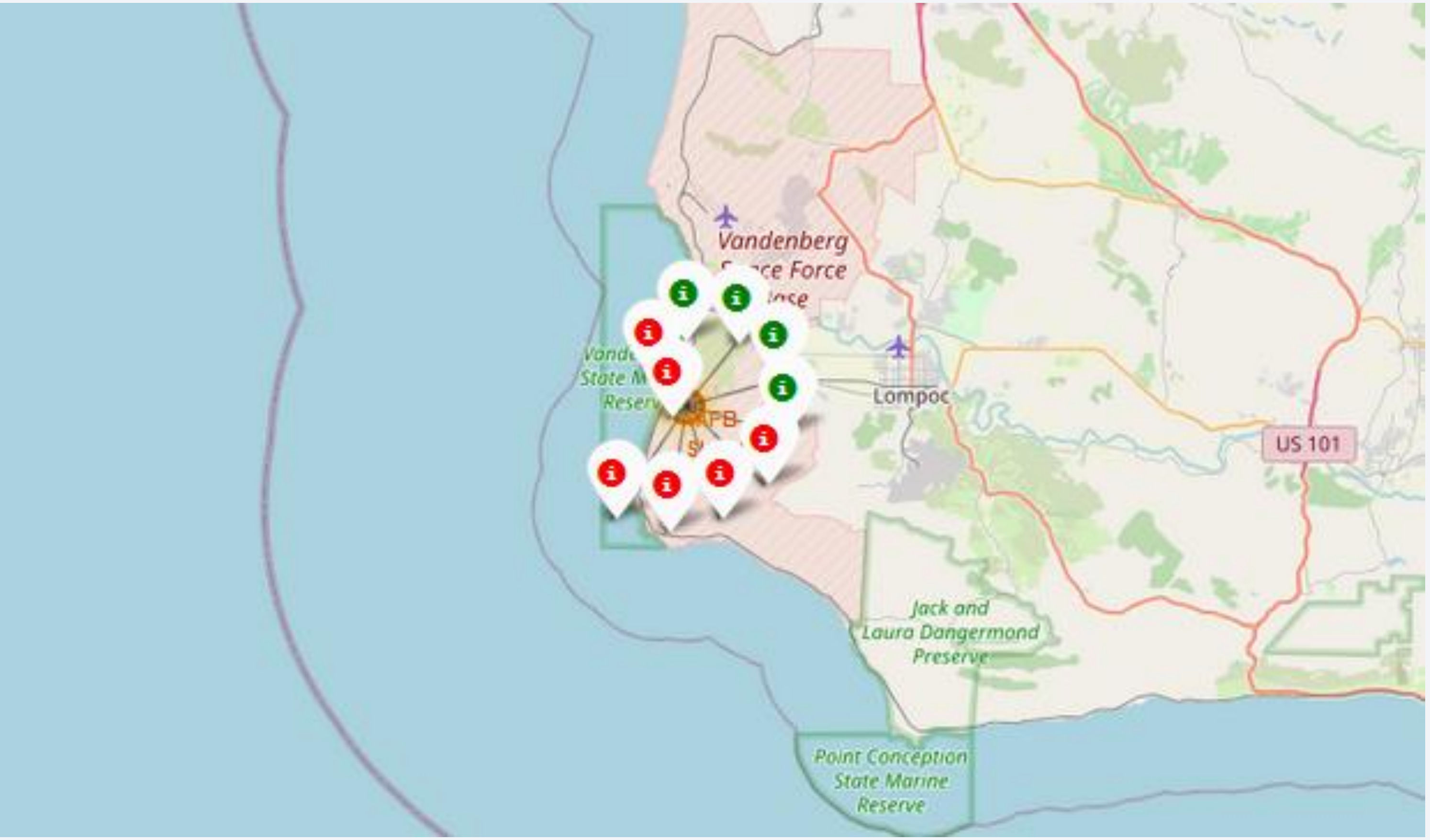
Landing_Outcome	LANDING_OUTCOME_COUNT
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

Section 3

Launch Sites Proximities Analysis

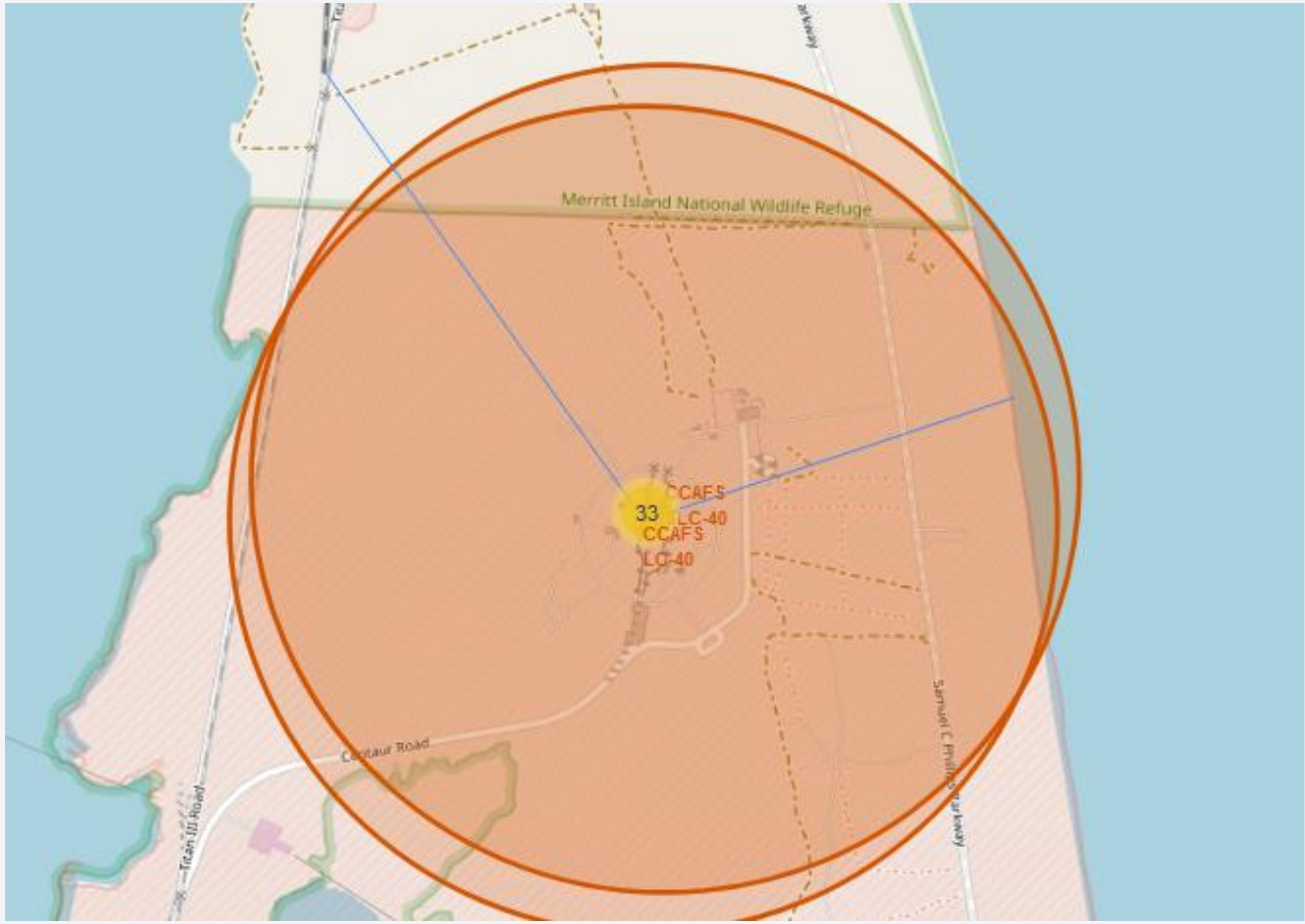
Color-Labeled Launch Outcomes - Geographic

- Green labeled items reflect successful launches, red labeled items reflect unsuccessful ones



Proximity of Launch Site CCAFS LC-40 with Coast and Railroad

- Blue lines indicates distance from launch site to coast and nearest railway

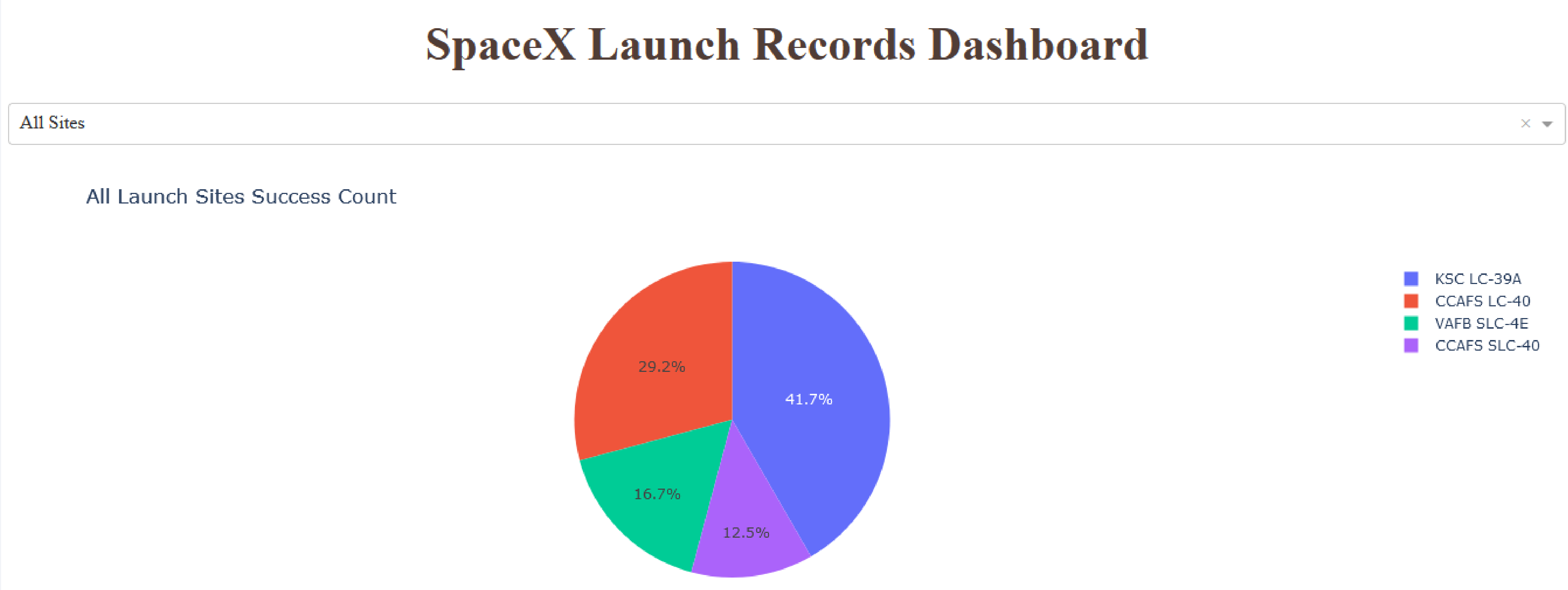


Section 4

Build a Dashboard with Plotly Dash

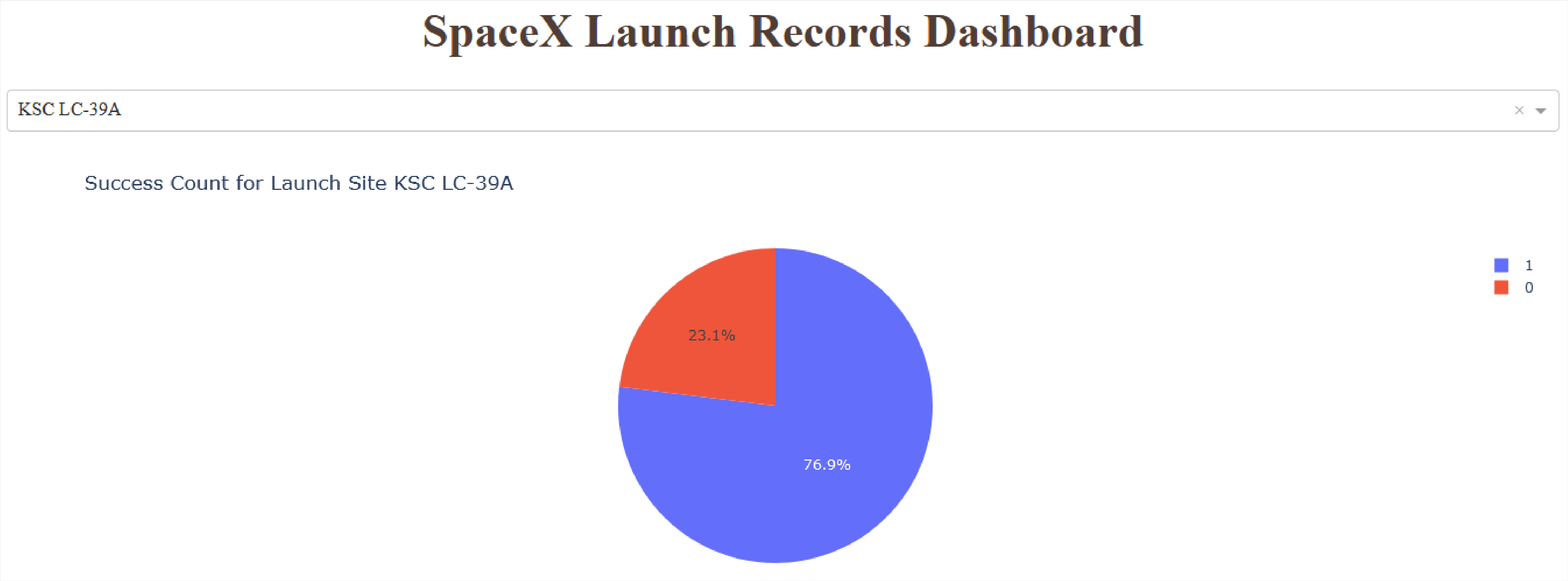
Dashboard - Launch Success Counts – All Sites

- Pie chart indicates launches from site KSC LC-39A make up the majority of the company’s successful launches



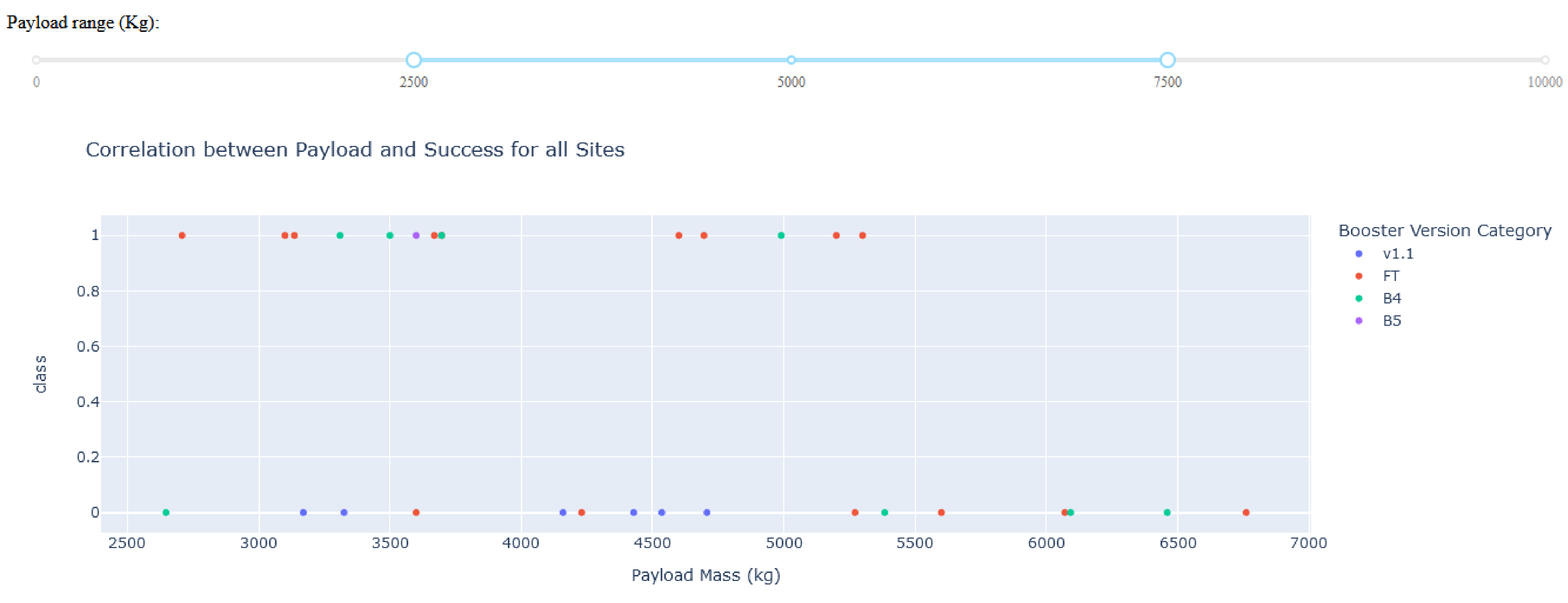
Dashboard - Highest Success Count – Launch Site

- Drilling into the success count of LSC LC-39A, we see that it has a 76.9% successful launch rate



Dashboard - Payload vs Launch Outcome

- With the selected payload range of 2500KG to 7500KG, we see that the FT Booster Version was the most successful across the entire payload range

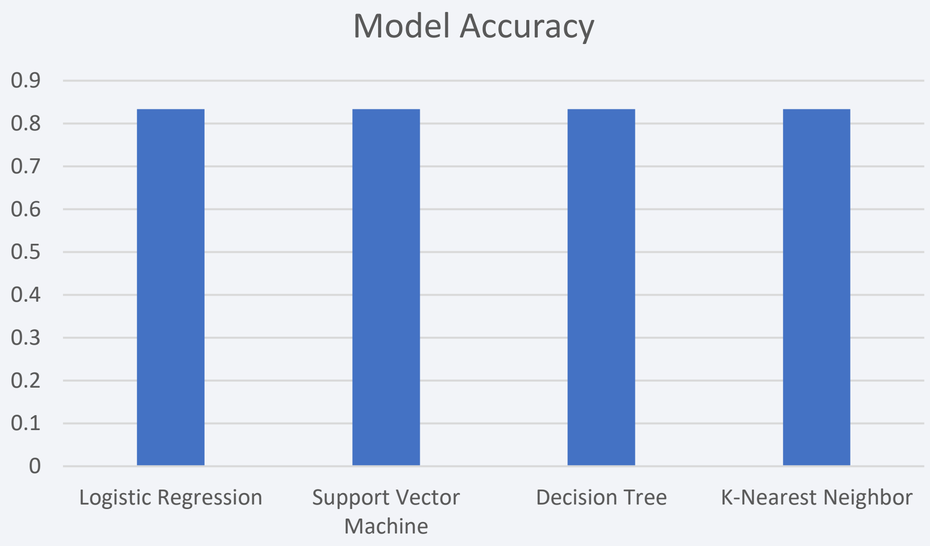


Section 5

Predictive Analysis (Classification)

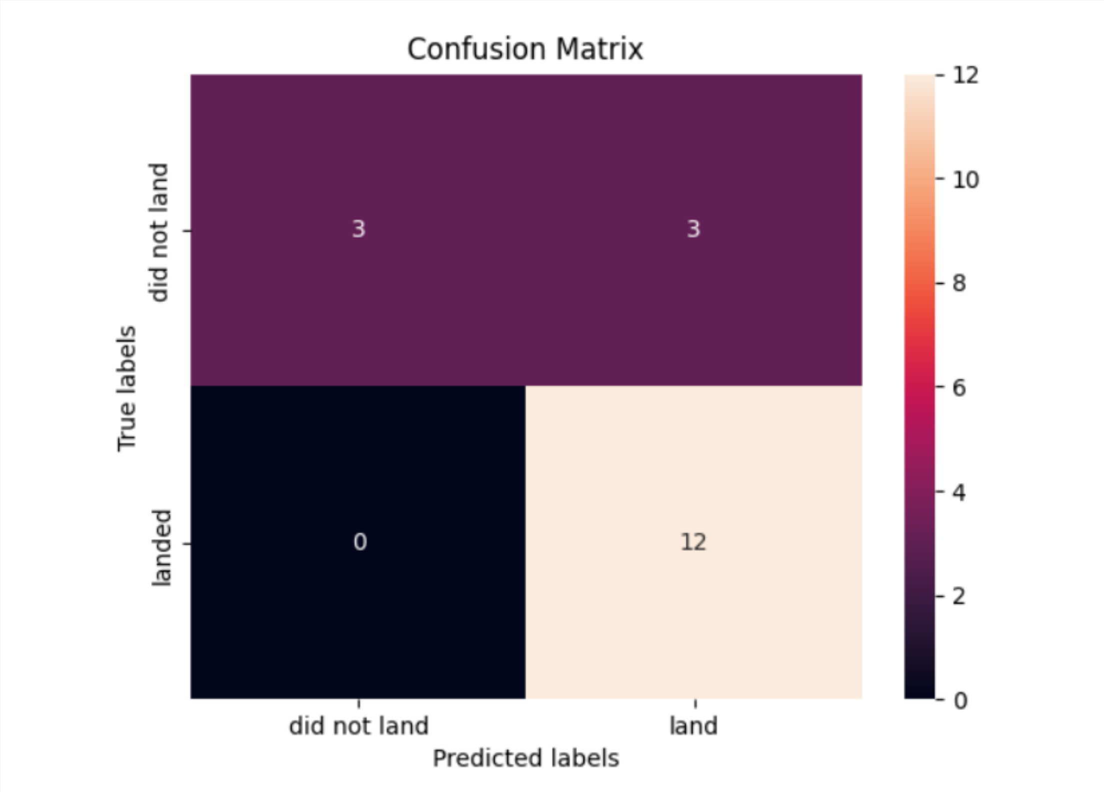
Classification Accuracy

- Models tested:
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree
  - K-Nearest Neighbor
- All models were virtually identical in accuracy
- Given the small size of a dataset, a simpler model is better as it is less prone to overfitting



Confusion Matrix – Logistical Regression Model

- The Confusion Matrix was identical for all models trained and tested
- There are 3 false positive results in the attached confusion matrix
- Were the dataset to expand exponentially a review of classification models may be necessary



Conclusions

- SpaceX has a 66.66% success rate in landings, and this rate is increasing year-by-year, indicating SpaceY needs to be aggressive in obtaining market share
- SpaceX's first successful launch was 2014, indicating SpaceY will require years of cash runway before achieving a successful launch
- Launches performed at Kennedy Space Center Launch Complex 39 have a 76.9% success rate, further analysis may be required as to why
- Given the consistent improvements in successful landings on SpaceX's part, I suspect the relationship between launch features and landing outcomes to be linear
- Logistic Regression is a suitable model for predicting future launch successes

Appendix/Acknowledgement

- I'd like to thank the all the instructors and discussion board moderators involved in the teaching of this certificate
- I'd also like to thank my fellow students for their contributions to the discussion boards which were invaluable for my completion of this report
- Please review github repository for all files used to contribute to this report:
    - https://github.com/bswervo/capstone

Thank you!