Winning Space Race with Data Science

Samson Akula
March 2023

Outline

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

Executive Summary

- Summary of methodologies:
  - Data Collection
  - Data Wrangling
  - Exploratory Data Analysis with Data Visualization
  - Exploratory Data Analysis with SQL
  - Building an Interactive map with Folium
  - Building a Dashboard with Plotly
- Summary of all results:
  - Exploratory Data Analysis Results
  - Interactive Analysis
  - Predictive Analysis

Determining the possibility of the Space Y program sustaining its advantage over similar companies/organizations. The promising Falcon9 and similar spaceships have reusable parts that could save millions of dollars. Utilizing available data and various techniques to determine which aspects of Space Y could lead to the discovery of reusable ships and which implies a more promising/successful option.

Introduction

- The commercial space age is here, companies are making space travel affordable for everyone
- Virgin Galactic, division of Virgin Atlantic, is providing suborbital spaceflights; Rocket Labs is small satellite provider; Blue Origin manufactures sub-orbital and orbital reusable rockets
- SpaceX is sending ships to the International Space Station
    - Launched Starlink, a satellite internet constellation providing Internet access and sending manned missions into Space
- SpaceX is the most promising in retaining reusable parts and relatively inexpensive
    - SpaceX's Falcon 9 spaceship costs $62 million while the competitors cost $165 million
    - Falcon 9 is more successful because of reusable parts
- Purpose of this Capstone is to take on the role of a data scientist working for a new company: SpaceY founded by Allon Tusk, would like to compete with SpaceX
- Question: Determine the price of each launch
    - Gathering information about SpaceX, creating dashboards for analysis, training a Machine Learning model, use available information to predict if SpaceY will reuse spaceship parts

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Used SpaceX API
  - Extracted the Falcon9 launch records HTML table from Wikipedia
- Perform data wrangling
  - Converted outcomes into training labels: 1 means successful booster landing, 0 means unsuccessful booster landing
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Created a Machine Learning pipeline to predict if the first stage will land given the data
  - Tuned and evaluated different classification models

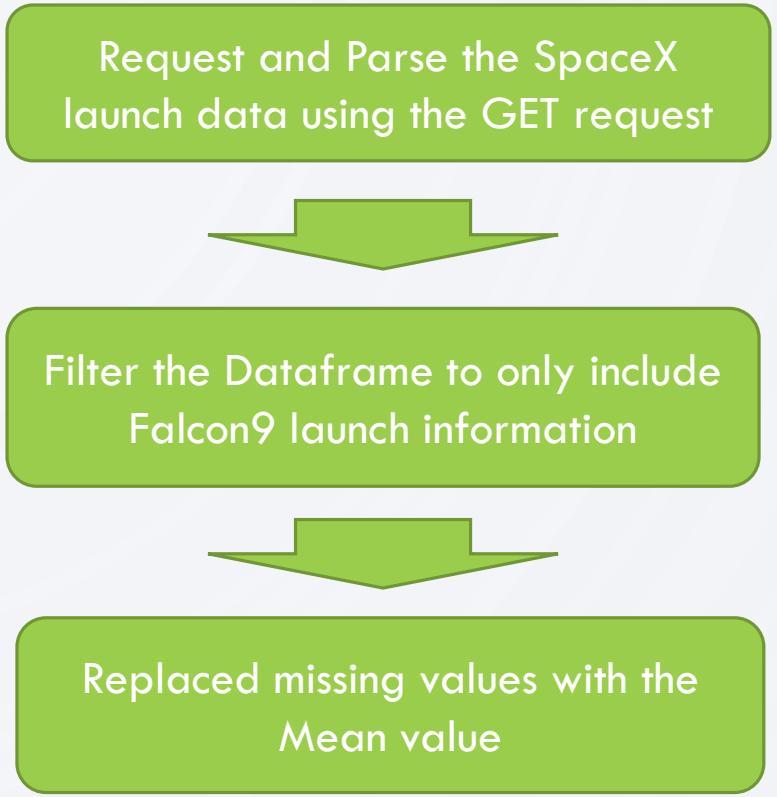
Data Collection

- SpaceX REST API: an unofficial open source REST API for SpaceX launch, rocket, core, capsule, starlink, Launchpad, and landing pad data
- Web Scraping: collecting Falcon9 historical launch records from Wikipedia using Beautiful Soup
  - BeautifulSoup is a python library package for parsing HTML and XML documents; creates a parse tree and parsed pages that can be used to extract data from HTML

Data Collection – SpaceX API

- Used a GET request to the SpaceX API: GET a JSON file of the data, JSON file was then converted into a DataFrame

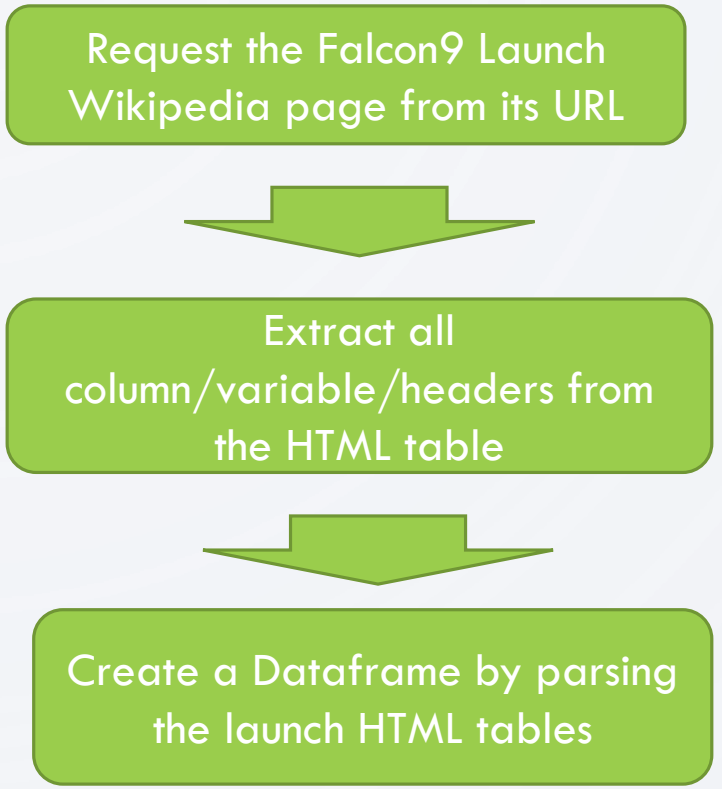
- GitHub: https://github.com/TheSamsonKnight/Data-Science-Capstone/blob/main/DATA%20COLLECTION%20API.ipynb



Data Collection - Scraping

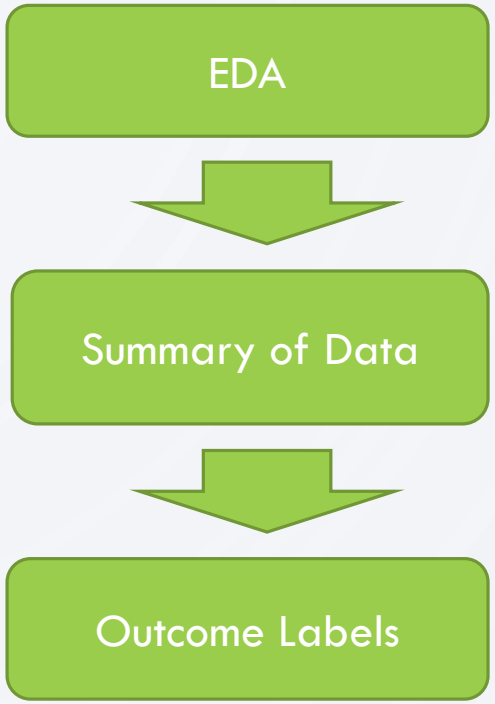
- Web Scrape Falcon9 launch records with BeautifulSoup

- GitHub: https://github.com/TheSamsonKnight/Data-Science-Capstone/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb



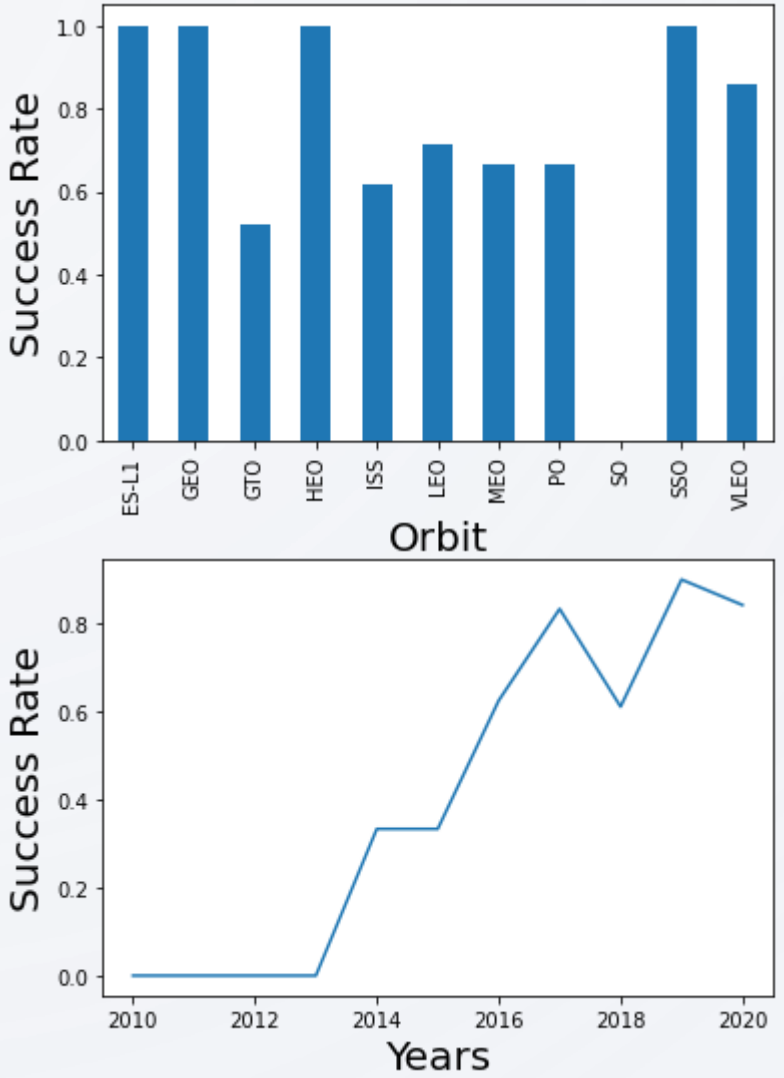
Data Wrangling

- Initially, Exploratory Data Analysis (EDA) was performed on the dataset
- Then the summaries launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type were calculated
- Finally, the landing outcome label was created from Outcome column
- GitHub: https://github.com/TheSamsonKnight/Data-Science-Capstone/blob/main/Data%20Wrangling.ipynb



EDA with Data Visualization

- To explore data, scatterplots, bar and line charts were used to visualize the relationship between pair of features:
  - Flight Number and Launch Site
  - Payload and Launch Site
  - Success Rate and Orbit Type
  - Flight Number and Orbit Type
  - Payload and Orbit Type
  - Launch Success by Year
- GitHub: https://github.com/TheSamsonKnight/Data-Science-Capstone/blob/main/EDA%20with%20Pandas%20and%20Matplotlib.ipynb

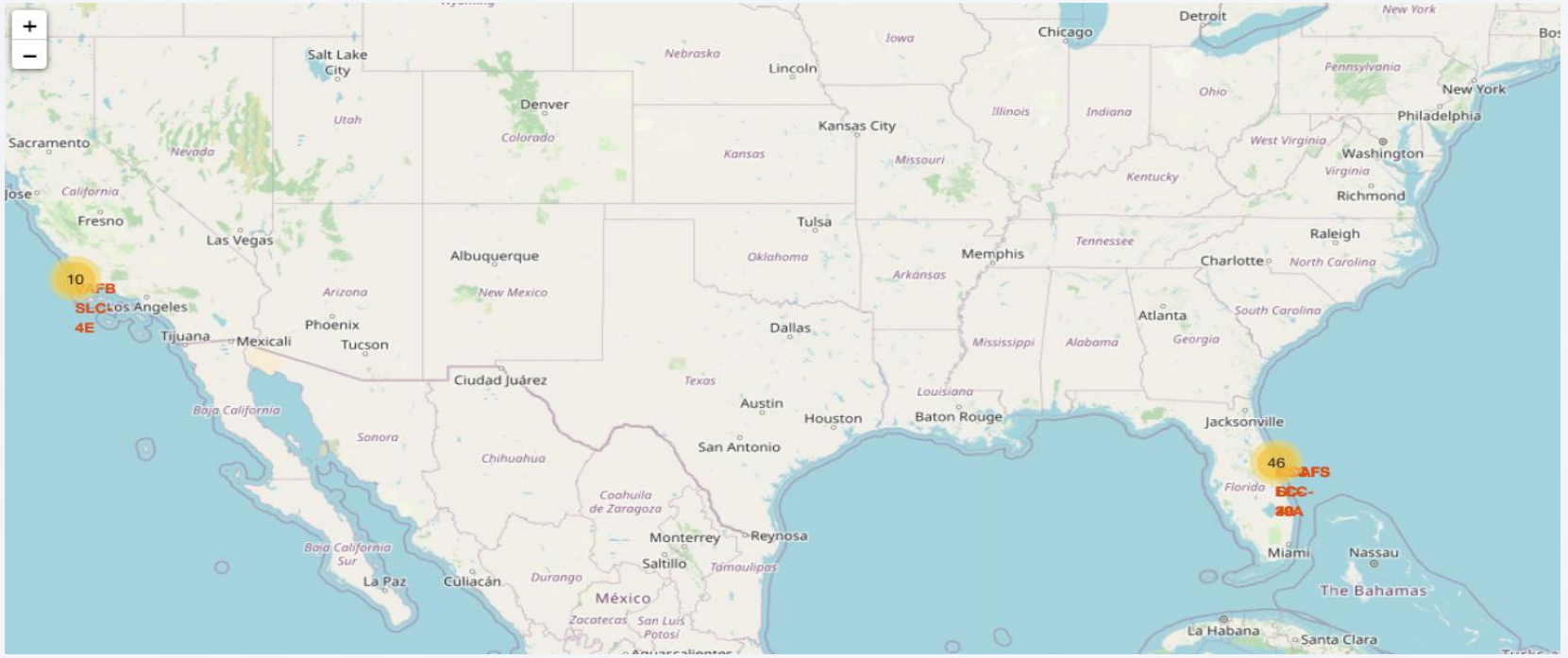


EDA with SQL

- The following SQL queries were performed:
  - Display the names of UNIQUE launch sites in the space mission
  - Display the 5 records where launch sites begin with the string 'CCA'
  - Display the Total Payload Mass carried by Boosters launched by NASA (CRS)
  - Display the Average Payload Mass carried by Booster Version Falcon9 v1.1
  - List the Date when the first Successful Landing outcome in Ground Pad was achieved
  - List the Names of the Boosters which have Success in Drone ship and have Payload Mass greater than 4000 but less than 6000
  - List the Total number of Successful and Failure mission outcomes
  - List the Names of the Booster Version which have carried the Maximum Payload Mass
  - List the records which will Display the Month names, Failure Landing Outcomes in Drone ship, Booster Versions, Launch site for the months in 2015
  - Rank the Count of Successful Landing Outcomes between the Date 6/4/2010 and 3/20/2017 in Descending order
- GitHub: https://github.com/TheSamsonKnight/Data-Science-Capstone/blob/main/EDA%20and%20SQL.ipynb

Build an Interactive Map with Folium

- The following markers were added on the map:
  - All launch sites: visualizing locations by pinning them on map
  - Successful/Failed launches for each site on the map: visualize which sites high Success rate
  - Calculate the distances between a launch site and Railways, Cities, Coastlines, Highways for safety purposes
- GitHub: https://github.com/TheSamsonKnight/Data-Science-Capstone/blob/main/Folium%20Lab.ipynb

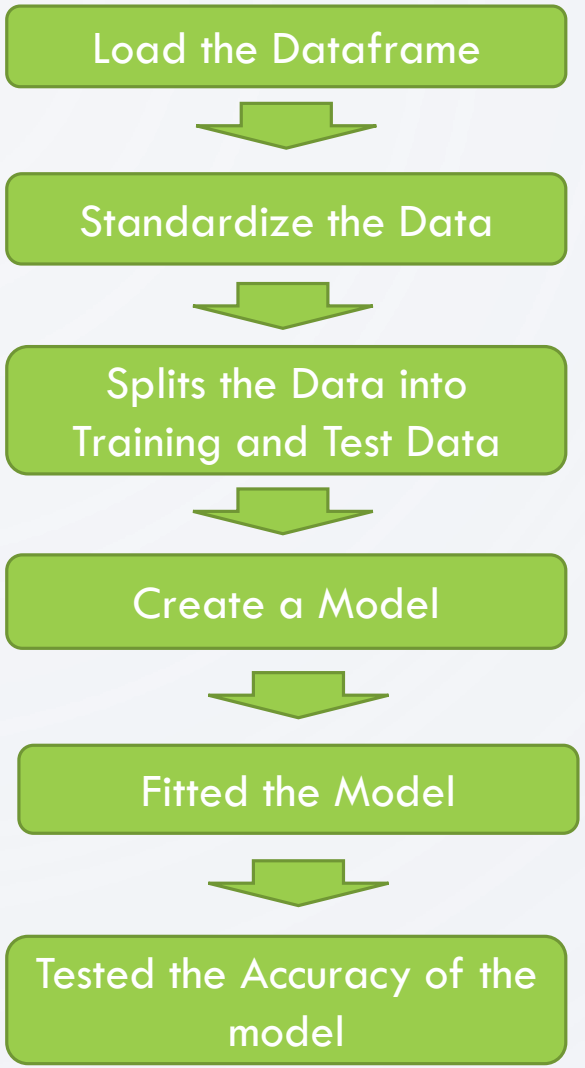


Build a Dashboard with Plotly Dash

- The following graphs and plots were used to visualize data:
  - Percentage of Launches by Site
  - Payload Range
- This combination allowed to quickly analyze the relation between Payloads and Launch sites; helping to identify where the best place to launch according to Payloads
- GitHub: https://github.com/TheSamsonKnight/Data-Science-Capstone/blob/main/EDA%20with%20Pandas%20and%20Matplotlib.ipynb

Predictive Analysis (Classification)

- Four different models were used on the data:
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree
  - K Nearest Neighbor
- GitHub: https://github.com/TheSamsonKnight/Data-Science-Capstone/blob/main/Machine%20Learning%20Predictive%20Analysis.ipynb



Results

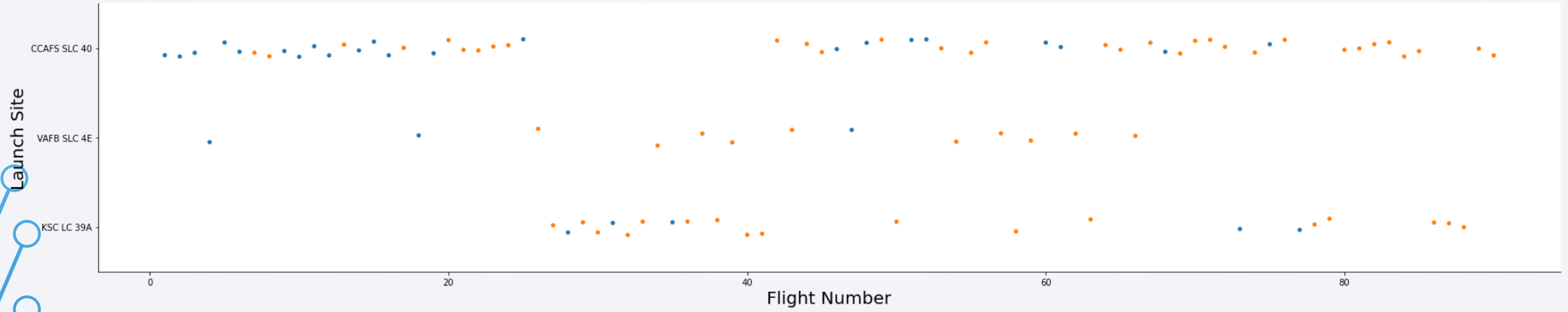
- EDA results:
  - SpaceX uses 4 different launch sites
  - Average Payload of Falcon9v1.1 Booster is 2,938 kg
  - The first Successful Landing outcome happened in 2015
  - Many Falcon9 Booster Versions were Successful at landing in Drone ships having the Payload above the Average
  - Almost 100% of Mission outcomes were Successful
  - Two Booster Versions failed at Landing in Drone ships in 2015: Falcon9 v1.1 B1012 and Falcon9 v1.1 B1015
  - The number of Landing Outcomes improved over time
  - Launches took place mainly on the east coast of USA
  - Decision Tree Classifier is the best model to predict Successful landing, having an accuracy over 87% and the 94% accuracy for Test data

Section 2

Insights drawn from EDA

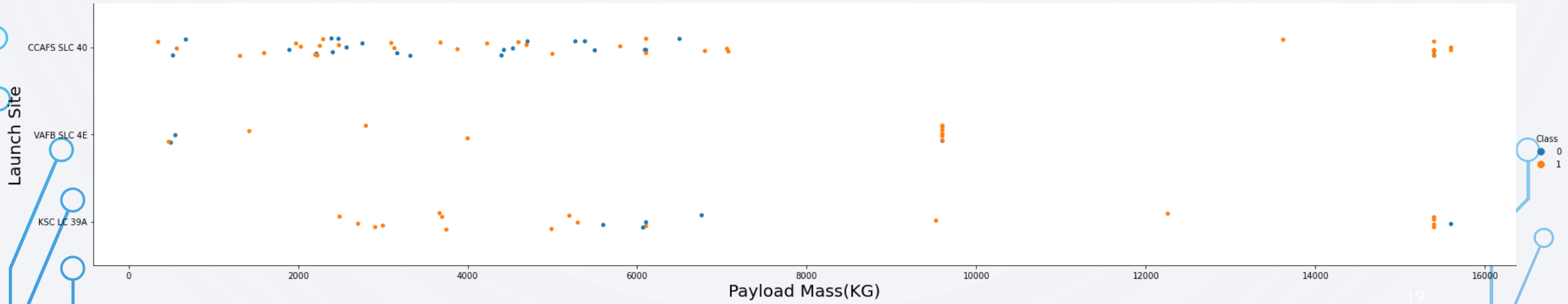
Flight Number vs. Launch Site

- According to the plot below:
  - CCAF5 SLC 40 has been the longest used Launch Site
  - CCAF5 SLC 40 has been used the most recently
  - CCAF5 SLC 4E has been used less frequently and has not been used recently
  - KSC LC 39A is the newest addition, the last 5 launched have been Successful



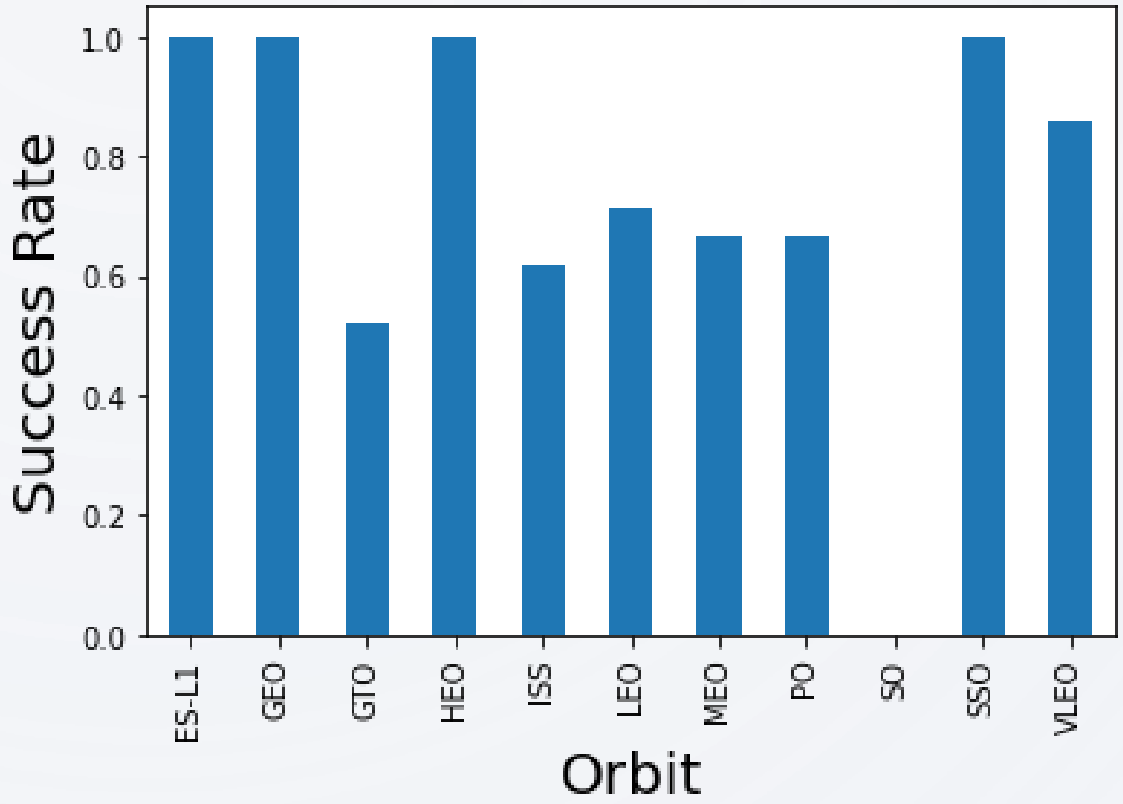
Payload vs. Launch Site

- CCAF5 SLC 40 has been used to test smaller Payload sizes with varying results
- VAFB SKC 4E has a maximum payload size of 9000kg
- KSC LC 39A has done well with smaller and larger Payload sizes but has Failed more often in the 5000kg-7000kg Payload size range
- Heavier Payloads have been more Successful overall



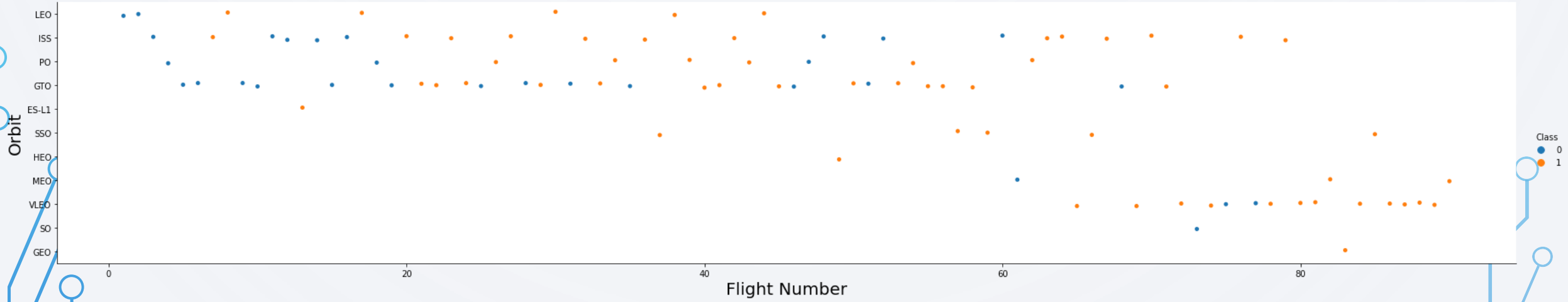
Success Rate vs. Orbit Type

- The following Orbits have been 100% Succesful:
  - ES-L1
  - Geo
  - Heo
  - Sso
- The following Orbits have been 70% or higher Succesful:
  - VLEO
  - LFO



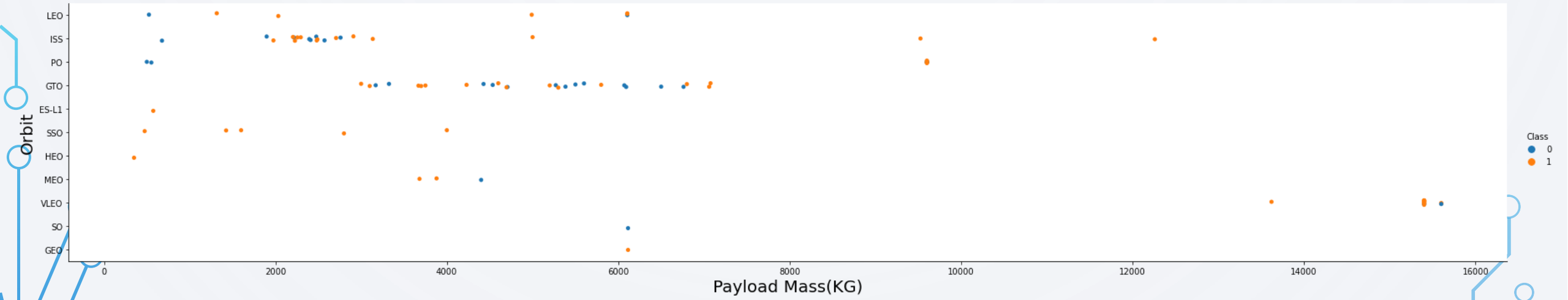
Flight Number vs. Orbit Type

- VLEO has the best Success rate
- GTO and ISS have the worst Success rate
- LEO, ISS, PO, GTO were the first Orbit types used with low Success rates initially
- More recently, WEO has been the most used orbit with a high Success rate



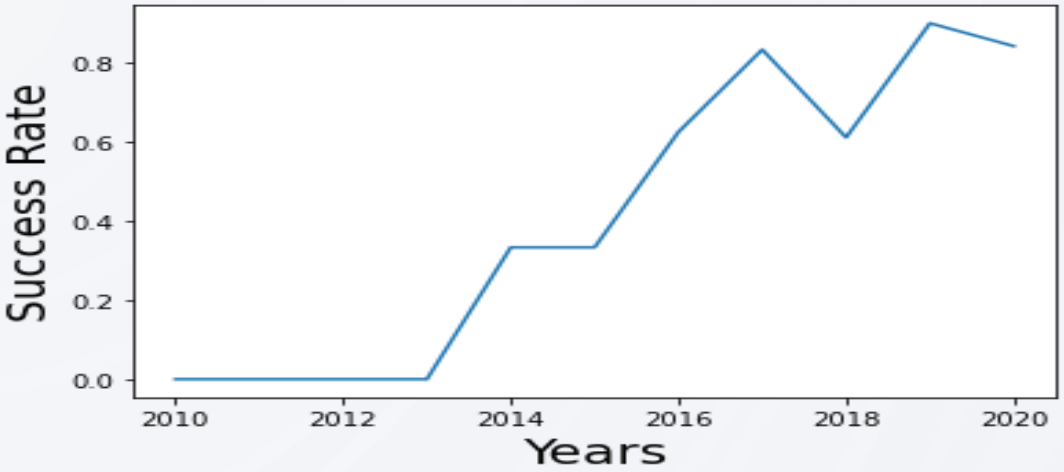
Payload vs. Orbit Type

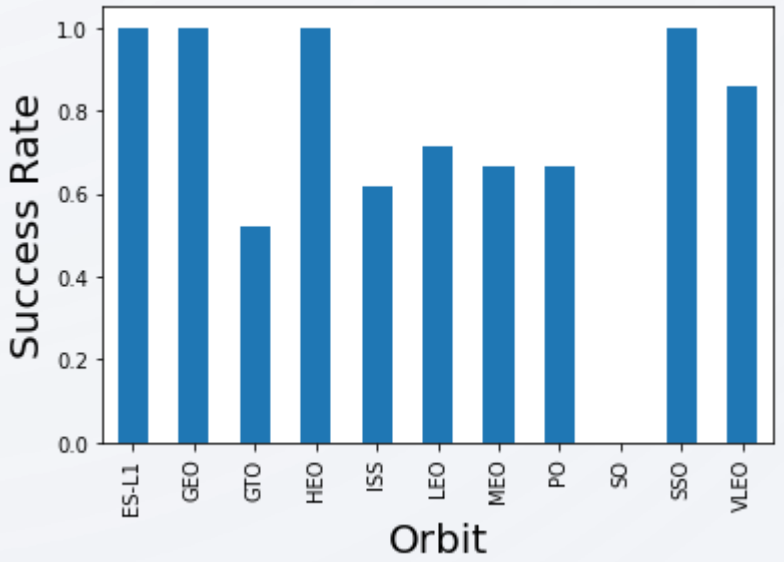
- VLEO is used exclusively for large Payloads
- GTO caters for Payloads between 3000kg and 7000kg
- ISS has the widest range of Payload weights



Launch Success Yearly Trend

- Success rates have been steadily improving
- No ‘perfect years’
- 2010-2013 had Zero Success rate (possible testing phase)
- 2018 was below the trend line
- Future launches should have high success rates if trend continues





All Launch Site Names

- Use Distinct() to find unique values

```
In [11]: %sql select distinct(LAUNCH_SITE) from SPACEXTBL;
```

Out[11]:

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

Launch Site Names Begin with 'CCA'

- Use LIKE ‘%String%’ to search records which contain a certain string

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Use SUM() to find the total a specific column

```
13]: %sql SELECT SUM(payload_mass__kg_) FROM SPACEXTBL WHERE CUSTOMER LIKE '%CRS%';
```

13]:

SUM(payload_mass__kg_)
48213

Average Payload Mass by F9 v1.1

- Use AVG() to find the average/mean of the specific column

```
In [14]: %sql SELECT AVG(payload_mass__kg_) FROM SPACEXTBL WHERE booster_version LIKE '%F9 v1.1%';

Out[14]: AVG(payload_mass__kg_)
         2534.6666666666665
```

First Successful Ground Landing Date

- Use MIN() to find the smallest value in a specific column

```
In [15]: %sql SELECT MIN(DATE) FROM SPACEXTBL WHERE "Landing_Outcome" LIKE '%Success (ground pad)%';

Out[15]: MIN(DATE)
         01-05-2017
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- Use BETWEEN to find values in a certain range

In [17]:
```
%sql select BOOSTER_VERSION from SPACEXTBL where "Landing_Outcome"='Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000;
```

Out[17]:

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- Use COUNT() to count the number of occurrences in a column
- Use GROUPBY() to arrange identical data into groups

List the total number of successful and failure mission outcomes

In [20]:
```
%sql select MISSION_OUTCOME, count(MISSION_OUTCOME) as Mission_Outcomes from SPACEXTBL GROUP BY MISSION_OUTCOME;
```

Out[20]:

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

Boosters Carried Maximum Payload

- Use a subquery first use MAX() to find the largest value
  - Then SELECT the Boosters with the max values

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

In [21]:
```
%sql select BOOSTER_VERSION as boosterversion from SPACEXTBL where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTBL);
```

Out[21]:

boosterversion
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

- Use LIKE ‘%2015%’ to find dates which have 2015 in them

```
In [25]: %sql SELECT DATE, "Landing_Outcome", booster_version, launch_site FROM SPACEXTBL WHERE "Landing_Outcome" LIKE '%Failure (drone ship)%' AND DATE LIKE '%
```

Out[25]:

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

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

- Use ORDER BY FIELD DESC to arrange outcomes highest to lowest

In [32]:
```
%sql SELECT "Landing_Outcome", COUNT("Landing_Outcome") AS count FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017
```

Out[32]:

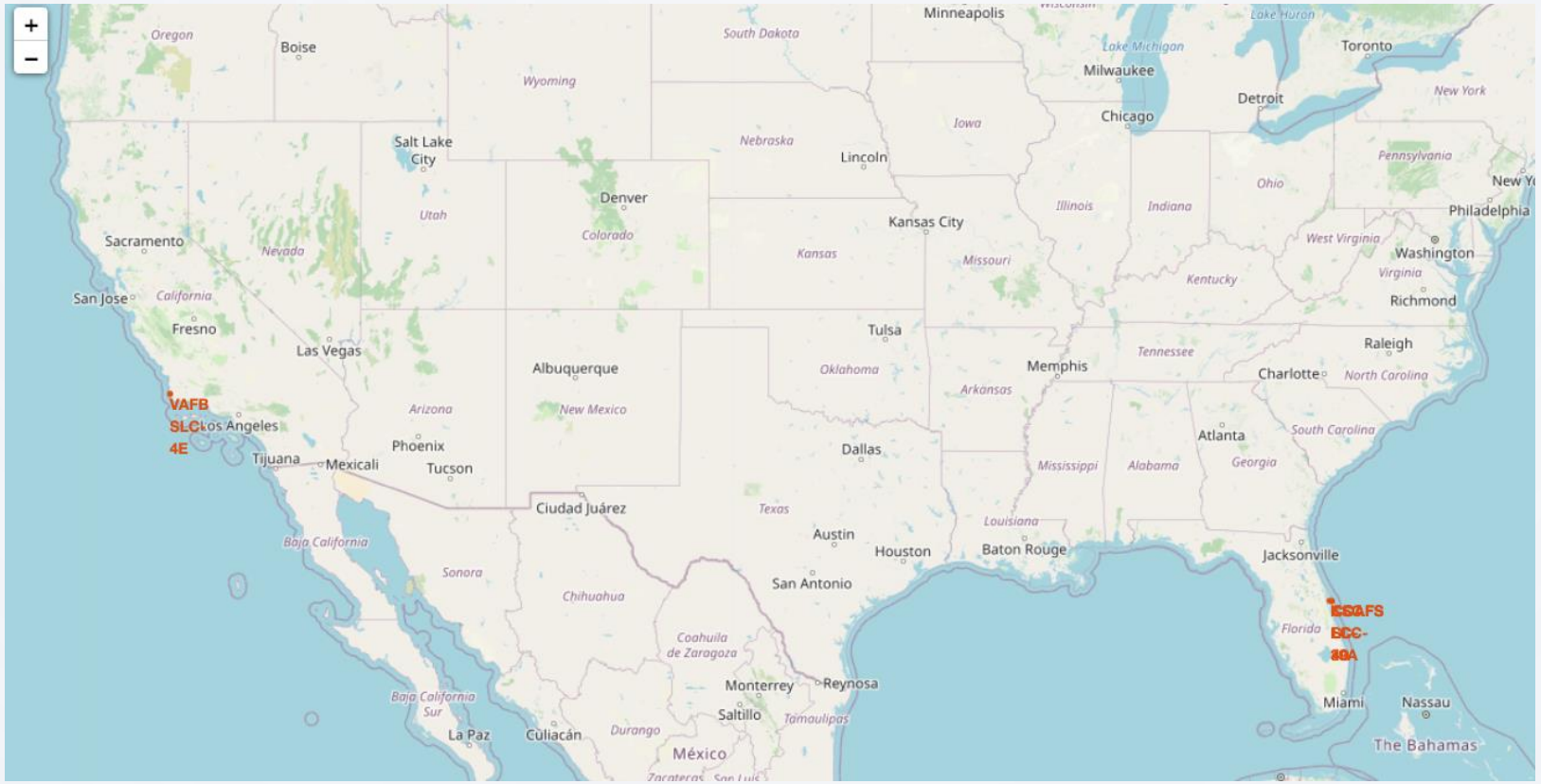
Landing Outcome	Occurrences
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

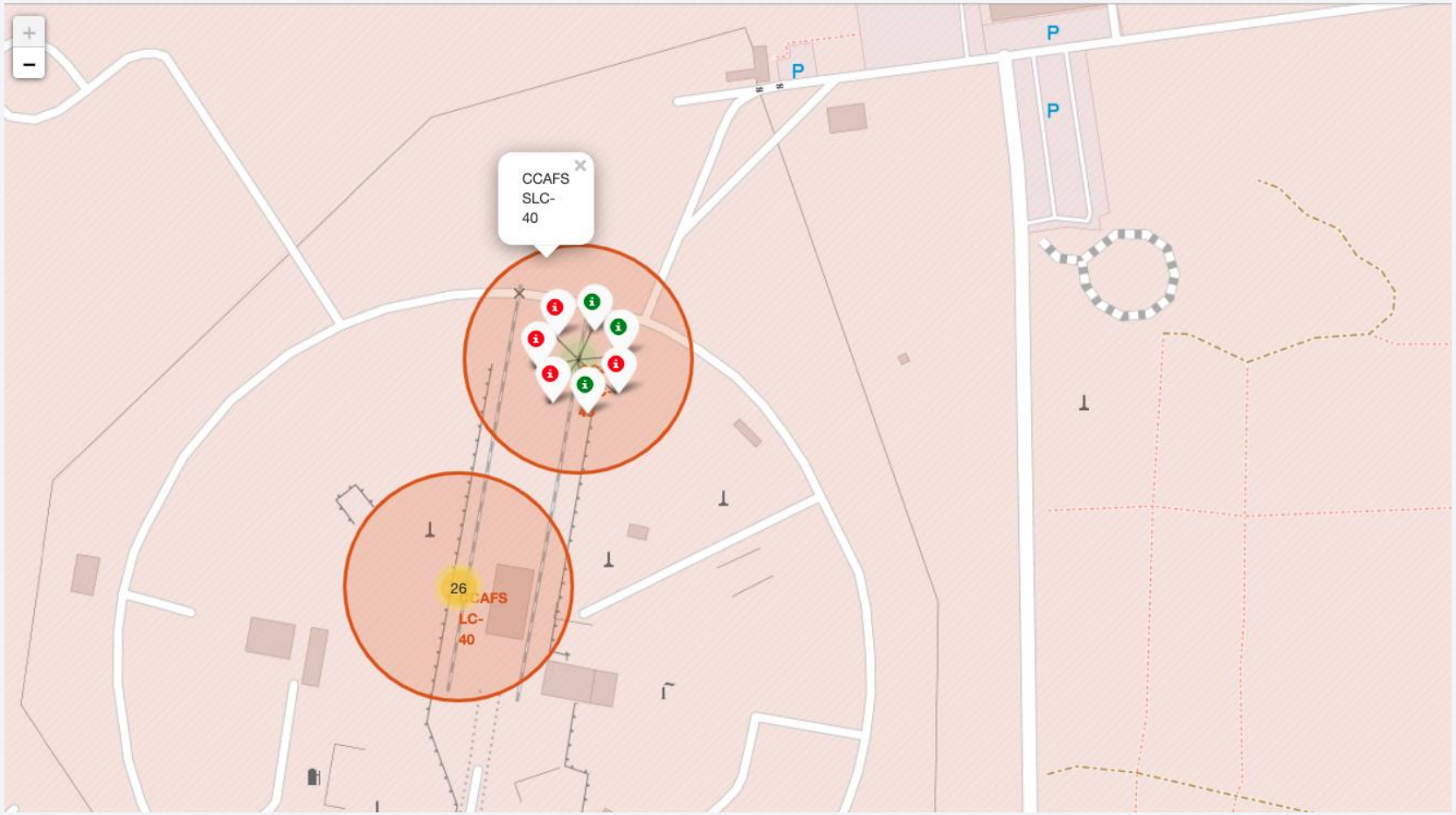
All Launch Sites

- Launch sites are to the coastline



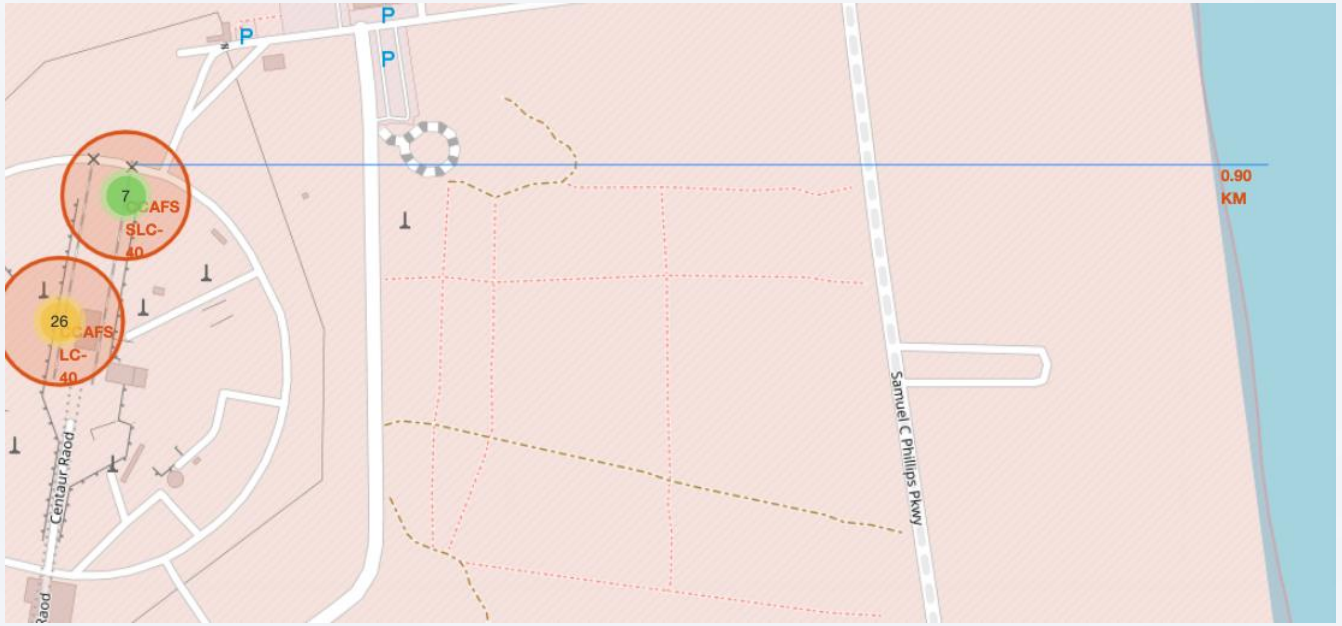
Successful and Failed Launches

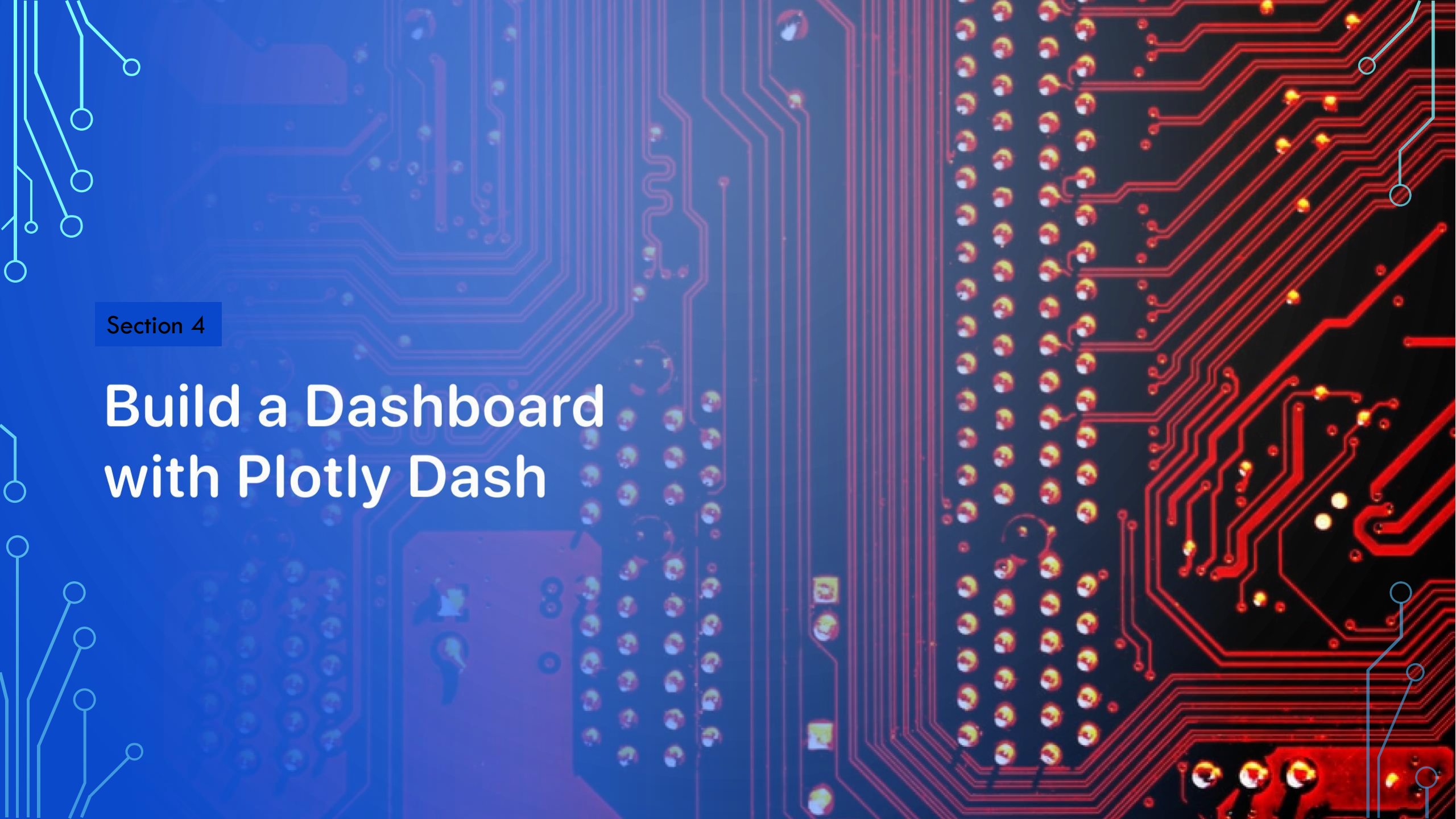
- Red indicates Failed launches, Green indicates Successful launches



Launch Site Proximity

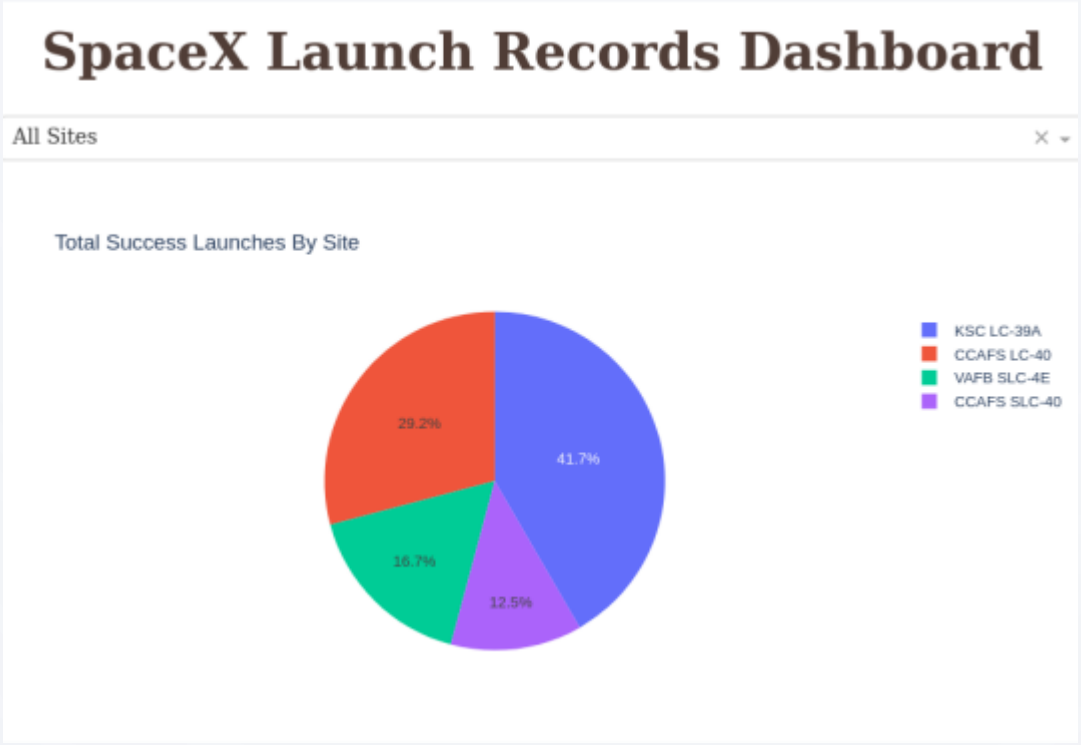
- Launch sites are close to the coast and ocean/sea for safety reasons of avoiding collateral damage such as people and cities
- Relatively close to railway lines
  - Trains used for transporting parts or ships




Section 4
Build a Dashboard with Plotly Dash

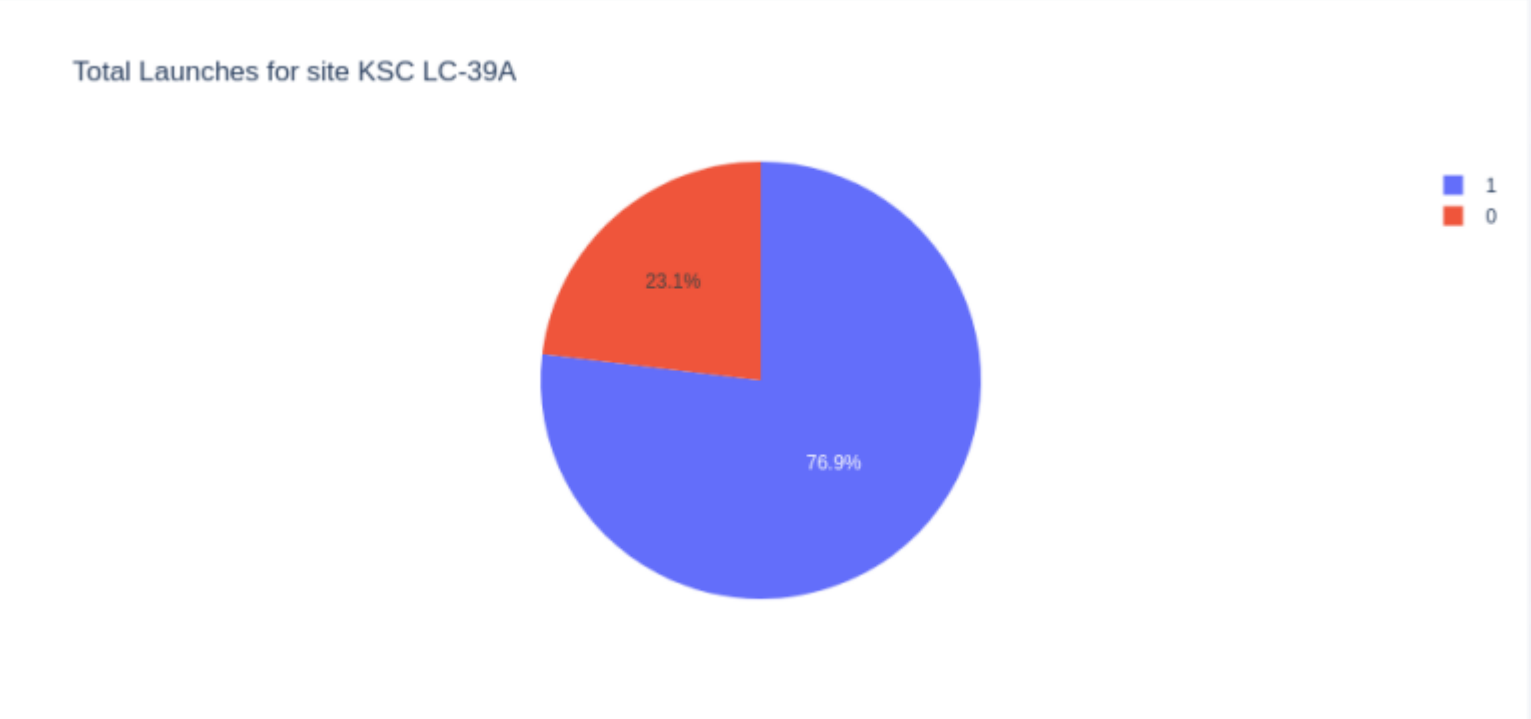
Successful Launches by Site

- Location of launch sites implies a correlation/key factor in Success rate



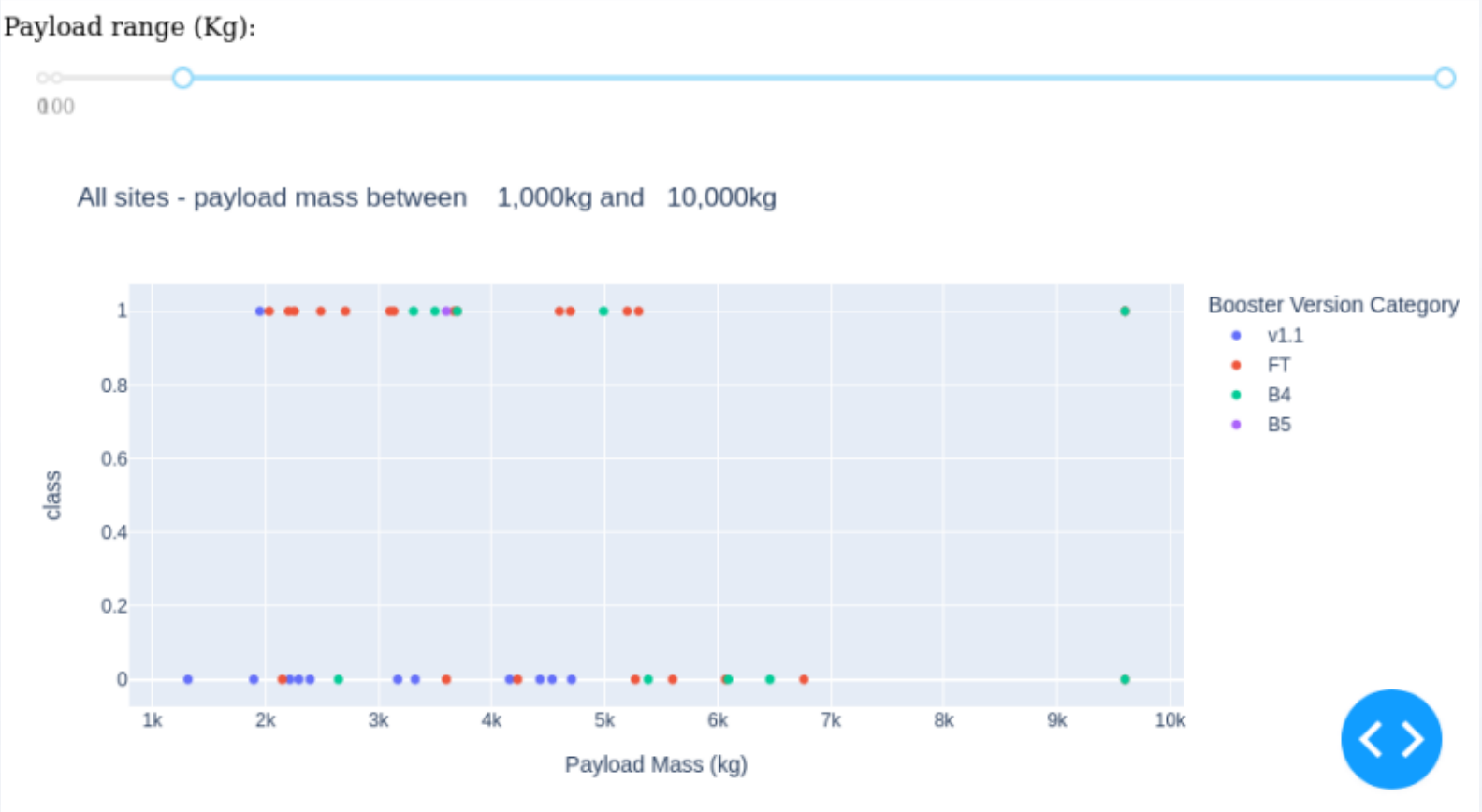
Launch Success Ration for KSC LC 39A

- Significant amount of Launches are Successful at this site: 76%



Payload vs. Launch Outcome

- Payloads below 6000kg are Successful
  - FT Boosters are successful with the Payload

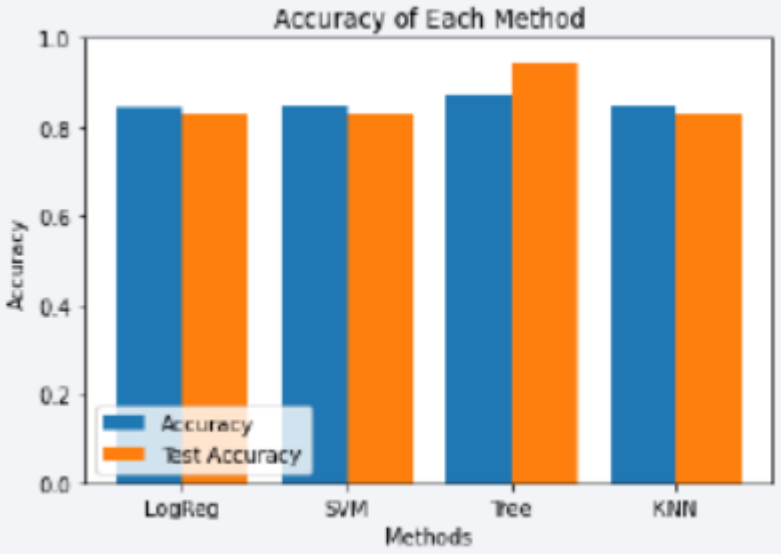


Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Four classification models were tested
- Model that represents the highest accuracy is Decision Tree Classifier
  - More that 87% accurate



Confusion Matrix

- Decision Tree represented the best result with no False-negatives and only 3 False-positives

Conclusion

- Different data sources were analyzed, refining conclusions along the process
- Best Launch site is KSC LC 39A
- Launches about 7000kg are more promising
- Most mission outcomes are Successful; Successful landing outcomes improved over time due to the evolution of technology, engineering, and aircrafts
- Decision Tree Classifier indicated Successful landings
  - increase profits/reusable parts

Appendix

- Space Race began in 1955 between the USA and Soviet Union, to achieve superior air and space capability
  - Sputnik 1(Soviet Union) was launched in 1957 igniting the spark of dominating the final frontier
  - Explorer 1(USA) was launched in 1958
  - Apollo 11(USA) was 1st to land on the moon in 1969
- Space X was founded in 2002
- Virgin Atlantic started Virgin Galactic in 2004
  - Won a contract for orbital/sub-orbital commercial transportation

Thank you!