IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ben Taber
July 30, 2025

Outline

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

https://commons.wikimedia.org/wiki/File:Falcon_9_Booster_CRS-11_Landing_at_LZ-1.jpg

https://commons.wikimedia.org/wiki/File:Launch_of_Falcon_9_carrying_DSCOVR_%2816511407248%29.jpg

Executive Summary

- Summary of methodologies
  - The investigation of SpaceX Falcon 9 stage 1 landings employs a structured data science workflow starting with data cleaning and preprocessing to handle missing values and inconsistencies. Exploratory data analysis using visualizations helps uncover patterns and relationships in the data. Feature engineering and selection refine the dataset for improved model performance. Multiple machine learning algorithms are trained and evaluated using cross-validation and relevant metrics, with hyperparameter tuning applied to optimize results. Standardization and class imbalance handling techniques further enhance model accuracy and robustness.
- Summary of all results
  - SpaceX Falcon 9 first stage landing outcomes have shown improvement over time, with launches from Florida achieving higher success rates than those from California. Factors such as lower payload mass and lower delta-v orbits contribute positively to first stage recovery. Despite limited data, four machine learning models predicted landing outcomes with over 80% accuracy, though additional data is needed to refine these models and identify the optimal approach.

Introduction

- Reusing launch vehicles is critical to reducing the cost of space launches, and by understanding the factors that lead to successful booster recovery by SpaceX will help the company decrease costs
- By analyzing launch data, we aim to answer
  - What are the factors that determine if a rocket will land successfully?
  - Which machine learning model has the best prediction for whether a Falcon 9 first stage is recovered?
  - How likely is a future Falcon 9 first stage landing to be successful given its context?

https://en.wikipedia.org/wiki/File:CRS-8_(26239020092).jpg

Section 1

Methodology

Methodology

Executive Summary

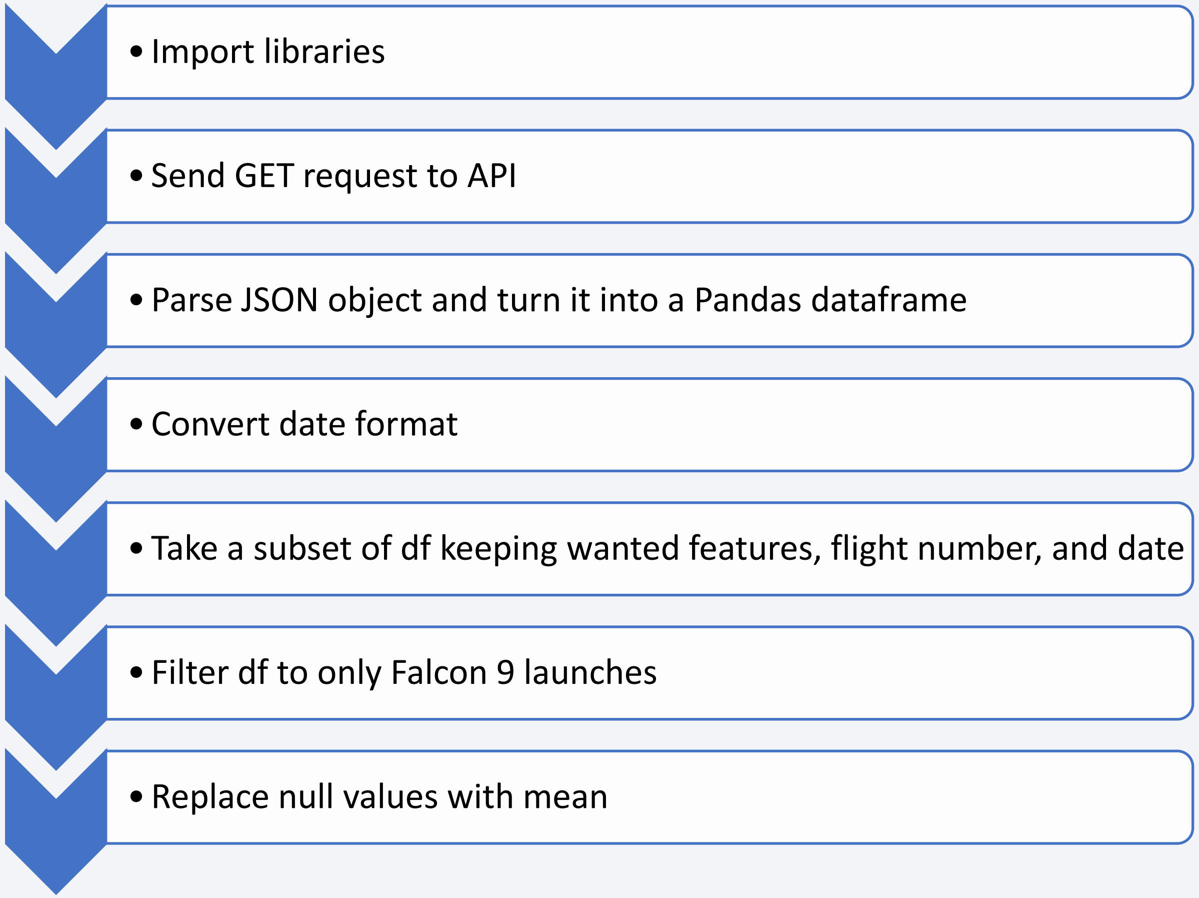
- Data was collected via the SpaceX API and through web scraping the Wikipedia list of Falcon 9 and Falcon Heavy Launches
- Collected data was initially in the form of a JSON object an HTML tables, which was then parsed into a Pandas dataframe
- Exploratory data analysis (EDA) performed using visualization and SQL
- Folium and Plotly Dash employed for interactive visual analytics
- Several machine learning classification models used to perform predictive analysis using several machine learning classification models

Data Collection

- Data sets were gathered from:
    - A publicly-available API with launch data as a JSON object
    - A Wikipedia page with launch data in an HTML table
    - CSV files provided by the IBM Skills Network

Data Collection – SpaceX API

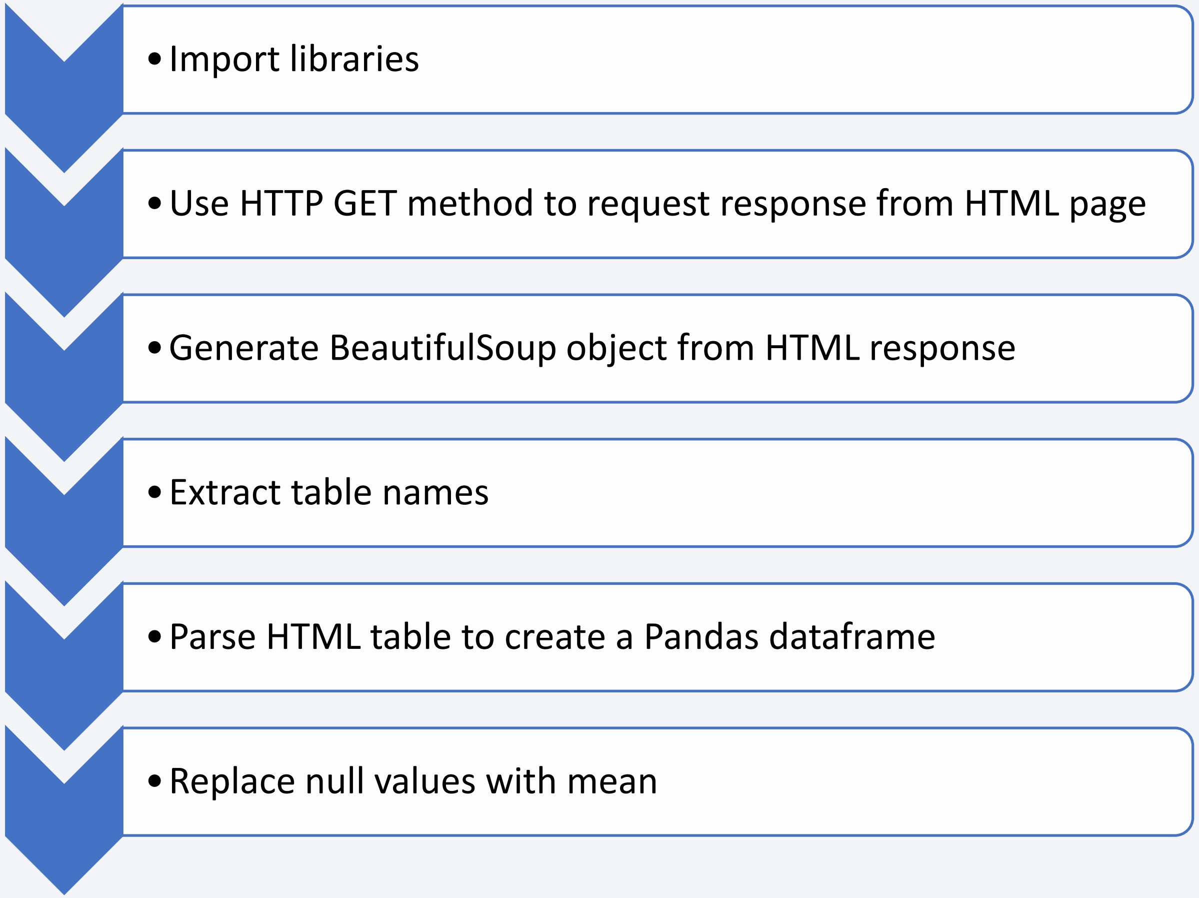
- SpaceX data is available at: https://api.spacexdata.com/v4/launches/past
- Data extracted from the response was loaded into a Pandas dataframe for data wrangling and analysis
- The notebook is available on GitHub



- Import libraries
- Send GET request to API
- Parse JSON object and turn it into a Pandas dataframe
- Convert date format
- Take a subset of df keeping wanted features, flight number, and date
- Filter df to only Falcon 9 launches
- Replace null values with mean

Data Collection - Scraping

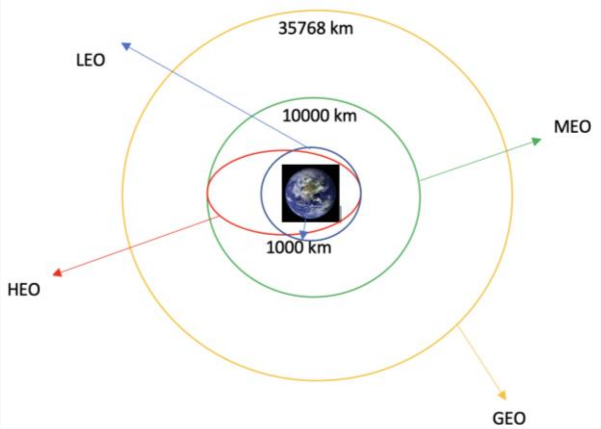
- SpaceX Falcon 9 launch data is updated June 9th, 2021 is available at: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- Data scraped from the Wikipedia page was loaded into a Pandas dataframe for data wrangling and analysis
- The notebook is available on GitHub

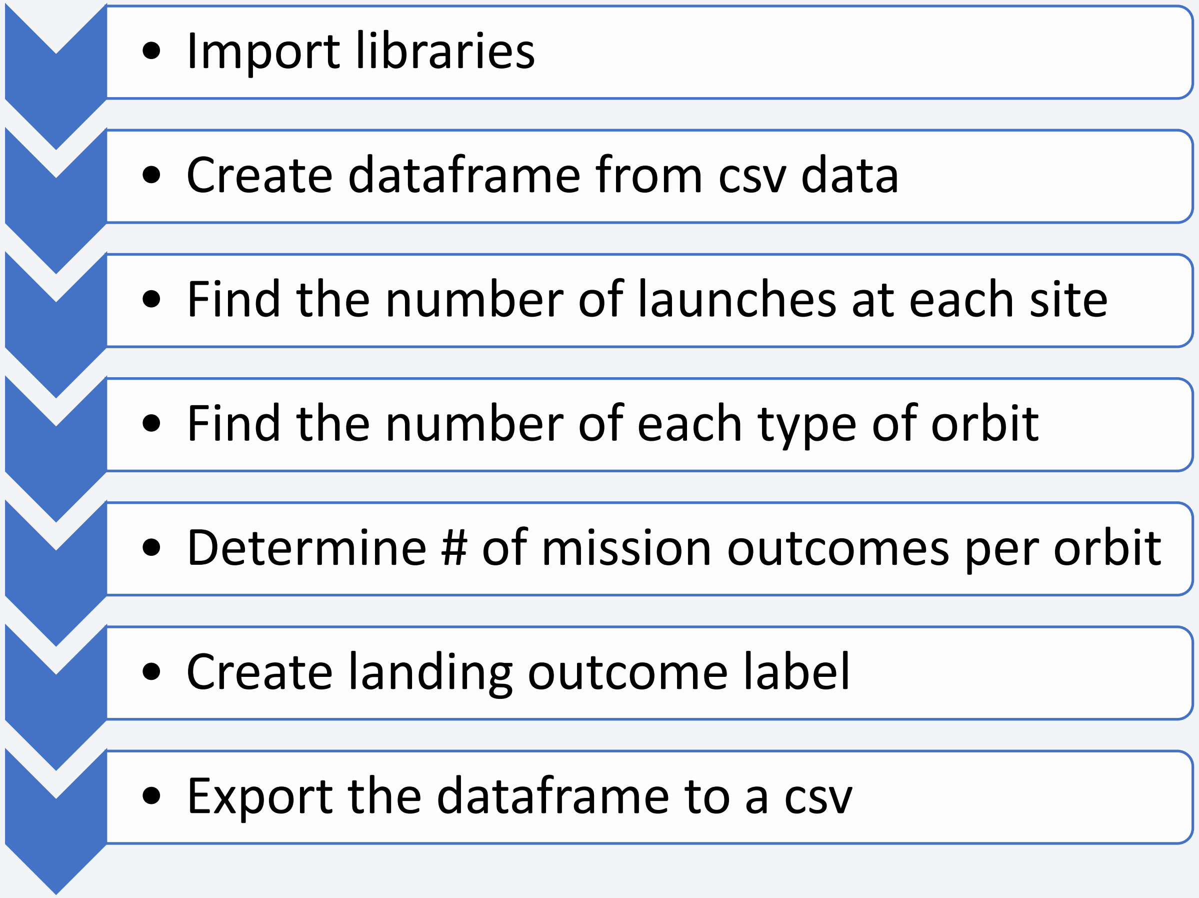


- Import libraries
- Use HTTP GET method to request response from HTML page
- Generate BeautifulSoup object from HTML response
- Extract table names
- Parse HTML table to create a Pandas dataframe
- Replace null values with mean

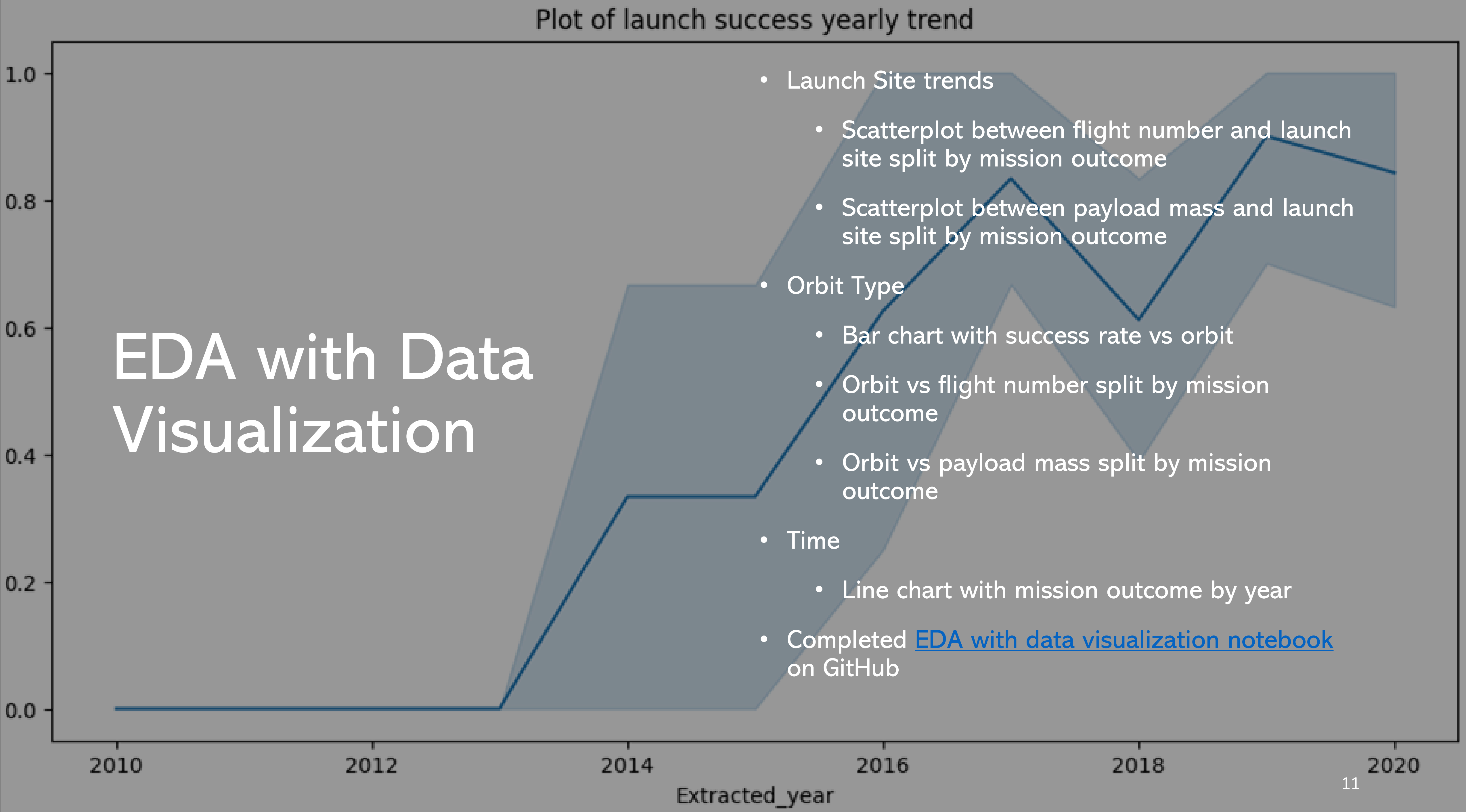
Data Wrangling

- Data came from a csv file representing the expected outcome from calling the API
- The launch sites, orbit types, and mission outcomes were processed and formatted
    - Outcomes changed to a 1 or 0 with one-hot encoding
- The [notebook is available on GitHub]





- Import libraries
- Create dataframe from csv data
- Find the number of launches at each site
- Find the number of each type of orbit
- Determine # of mission outcomes per orbit
- Create landing outcome label
- Export the dataframe to a csv



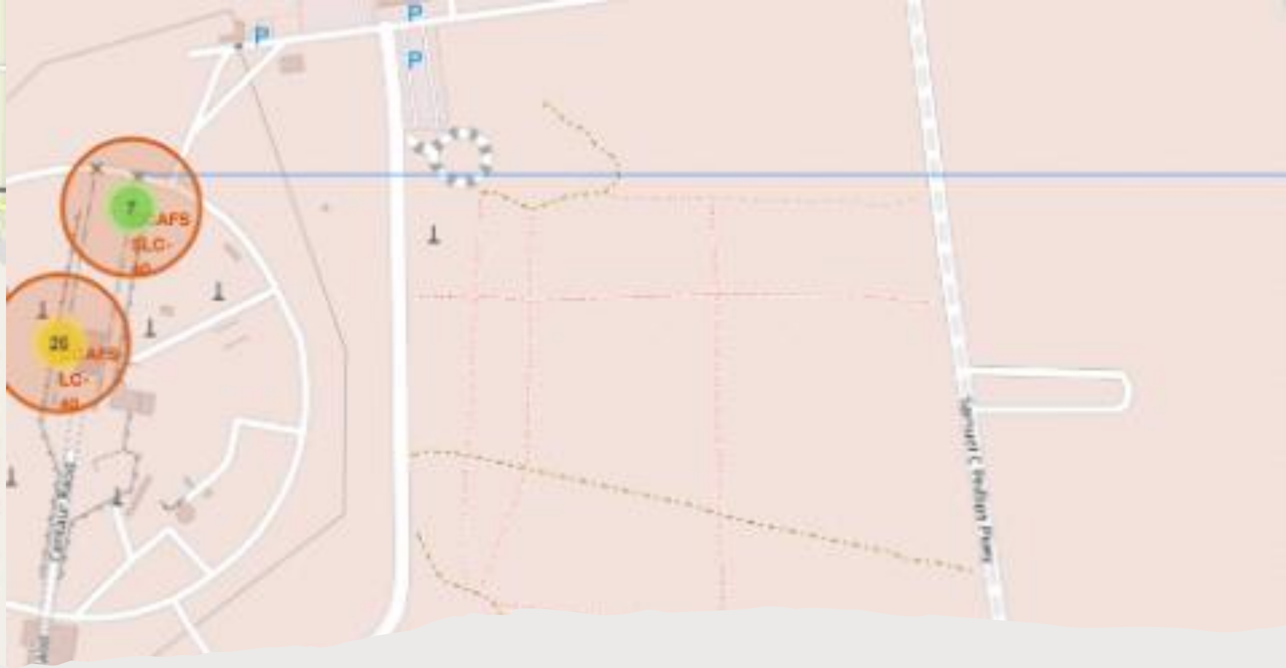
EDA with Data Visualization

- Launch Site trends
  - Scatterplot between flight number and launch site split by mission outcome
  - Scatterplot between payload mass and launch site split by mission outcome
- Orbit Type
  - Bar chart with success rate vs orbit
  - Orbit vs flight number split by mission outcome
  - Orbit vs payload mass split by mission outcome
- Time
  - Line chart with mission outcome by year
- Completed EDA with data visualization notebook on GitHub

EDA with SQL

- Used SQL queries to
  - Determine the launch site locations
  - Five records of KSC LC-39A launches
  - Total Payload mass carries by NASA-launched boosters
  - Average payload mass carried by booster version F9 v1.1
  - Date of successful landing on drone ship
  - Names of success boosters with payload mass between 4000 and 6000 kg
  - Total number of successful and failed mission outcomes
  - The names of booster version with the max payload mass
  - Landing records from 2017
  - Number of landings between 2010-06-04 and 2017-03-20
- Completed EDA with SQL notebook on GitHub

+
−
Vandenberg State Marine Reserve
VAFB SLC 4E



Build an Interactive Map with Folium

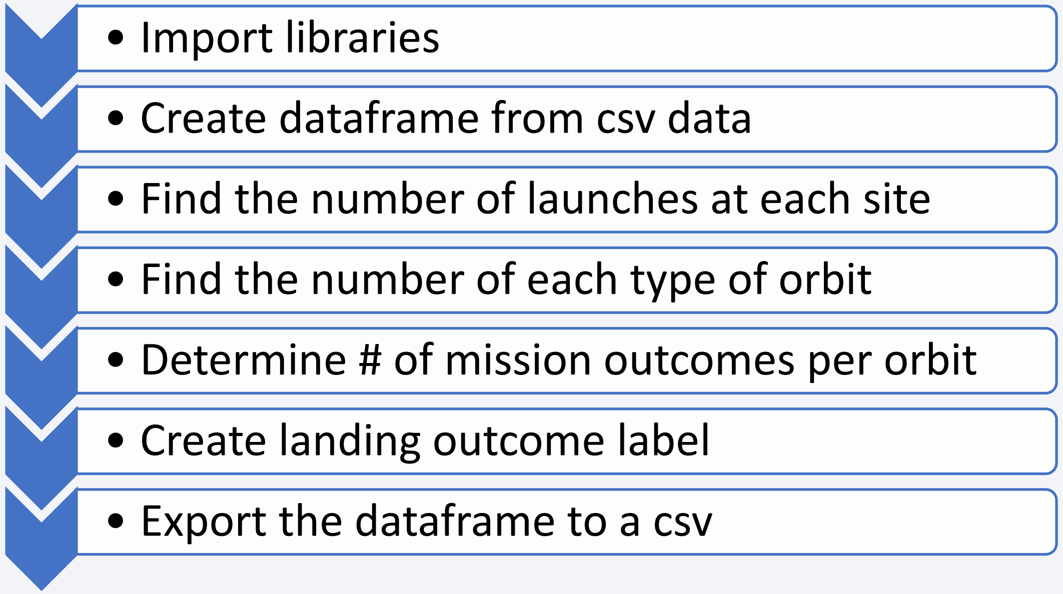
- Created and added map objects to the Folium map
  - Markers for all launch sites and the NASA Johnson Space Cente
  - Circles to designate and label launch sites
  - Lines to calculate and show the distances between a launch site and nearby points of interest
- By adding these objects, it becomes easier to identify which launch sites have relatively high success rates
- [Interactive map with Folium on GitHub]()

Build a Dashboard with Plotly Dash

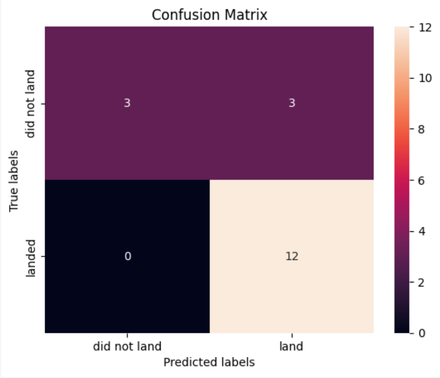
- By constructing an dashboard with Plotly Dash, we can interactively
  - Plot pie charts showing the total launches by site
  - Generate scatter plots showing the relationship between payload mass and outcome for different boosters
  - Display the distribution of the Falcon 9 first stage landings split out by payload mass, mission outcome, and booster version
- Through the dashboard, we can visualize relationships in real time, thereby enabling the finding of insights
- Plotly Dash lab on GitLab

Predictive Analysis (Classification)

- The dataset was split into training and test sets
- Four ML models were trained:
  - Logistic regression
  - Support vector machine
  - Decision tree
  - K-nearest neighbors
- Hyperparameters were evaluated the selected
- The four models with their best hyperparameters were scored using the test set
- [Predictive analysis lab on GitHub]()



- Import libraries
- Create dataframe from csv data
- Find the number of launches at each site
- Find the number of each type of orbit
- Determine # of mission outcomes per orbit
- Create landing outcome label
- Export the dataframe to a csv

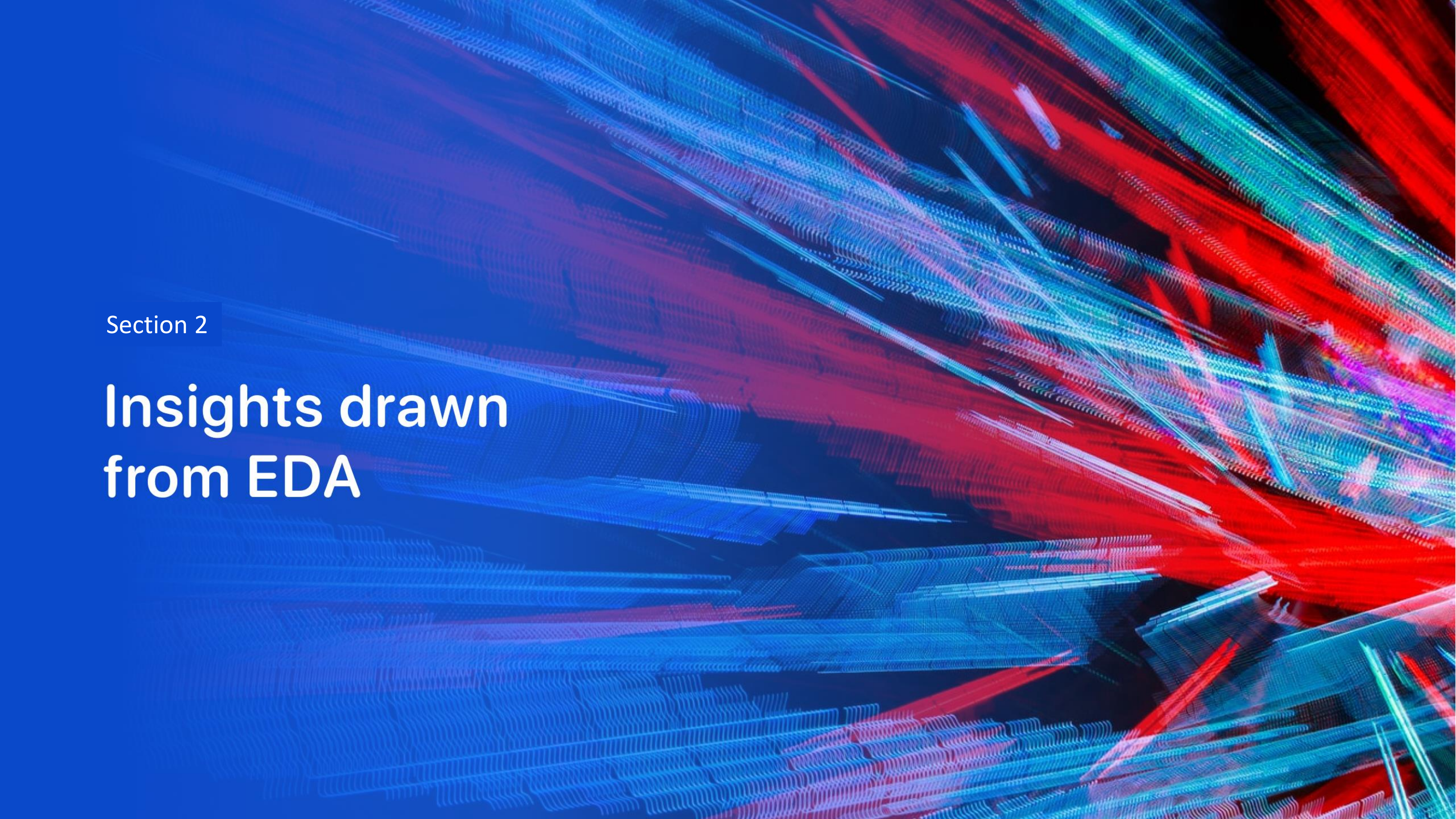


Results

EXPLORATORY DATA ANALYSIS RESULTS

INTERACTIVE ANALYTICS DEMO IN SCREENSHOTS

PREDICTIVE ANALYSIS RESULTS

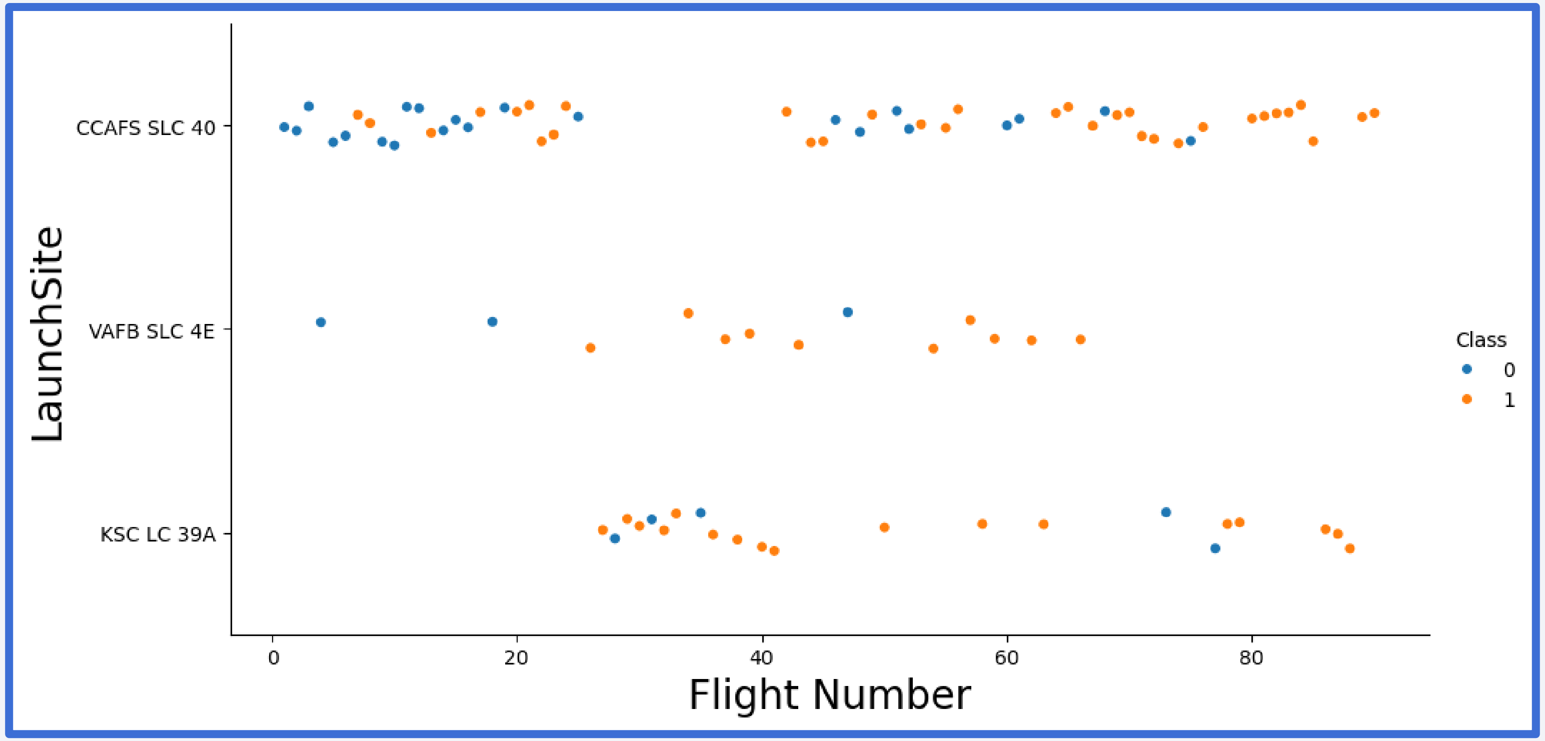


Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

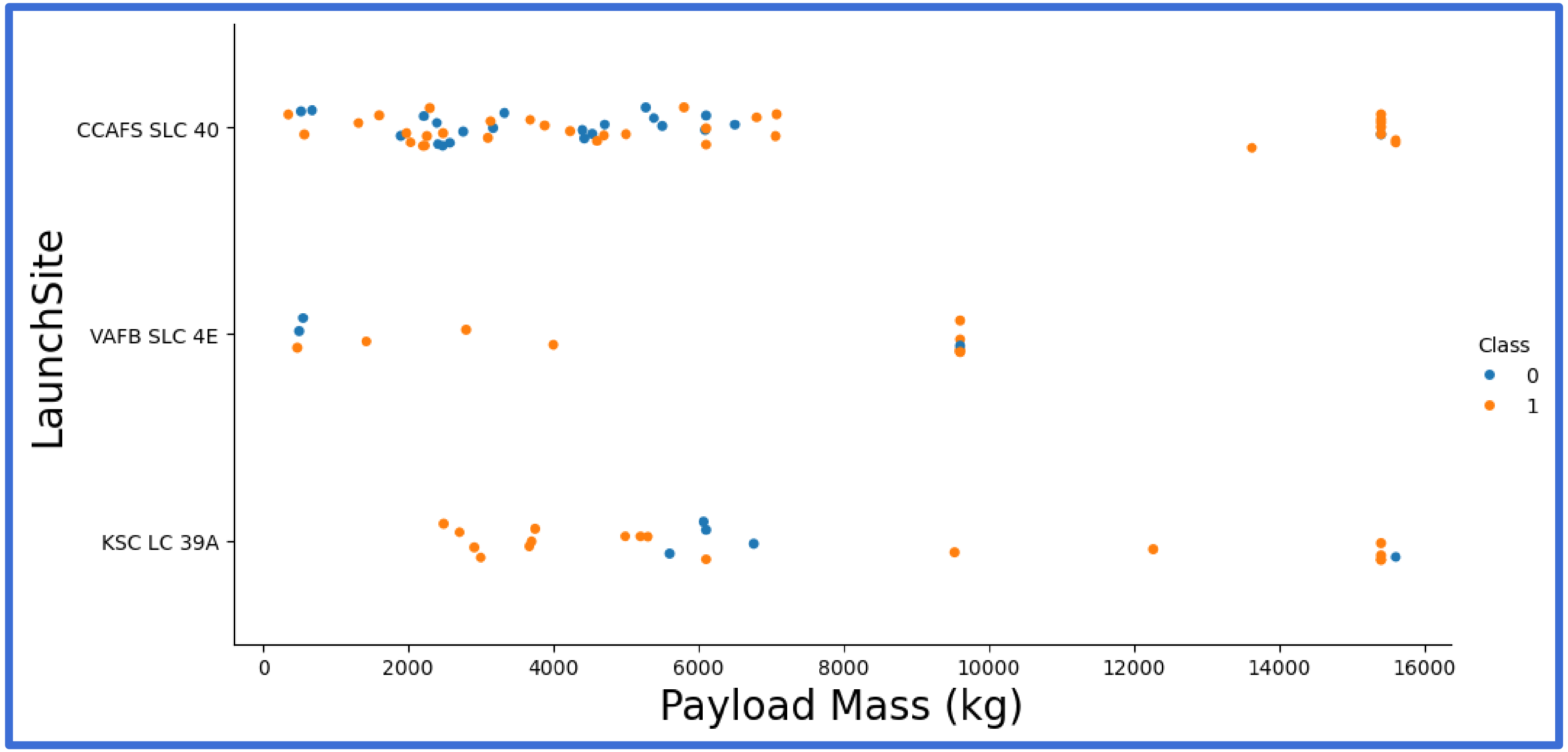
- There was significant variation in success rate between launch sites
- Successful first stage landings become more common over time as flight number increases



Falcon 9 first stage failed and successful landings

Payload vs. Launch Site

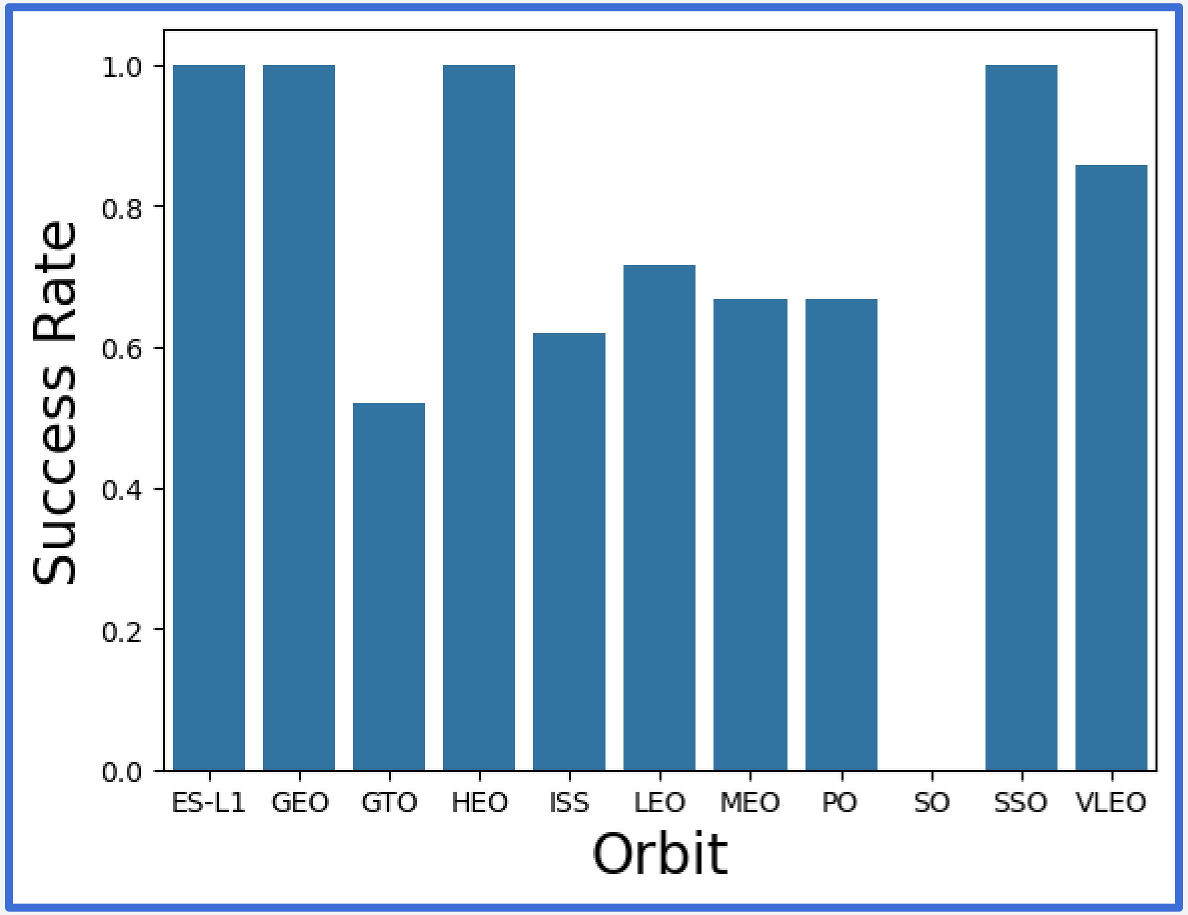
- Payload mass and landing outcome appear to be correlated
  - Especially for CCAFS SLC 40
- In general, more mass decreases success rate
- KSC LC 39A failures are clustered around 6000 kg



Falcon 9 first stage failed and successful landings

Success Rate vs. Orbit Type

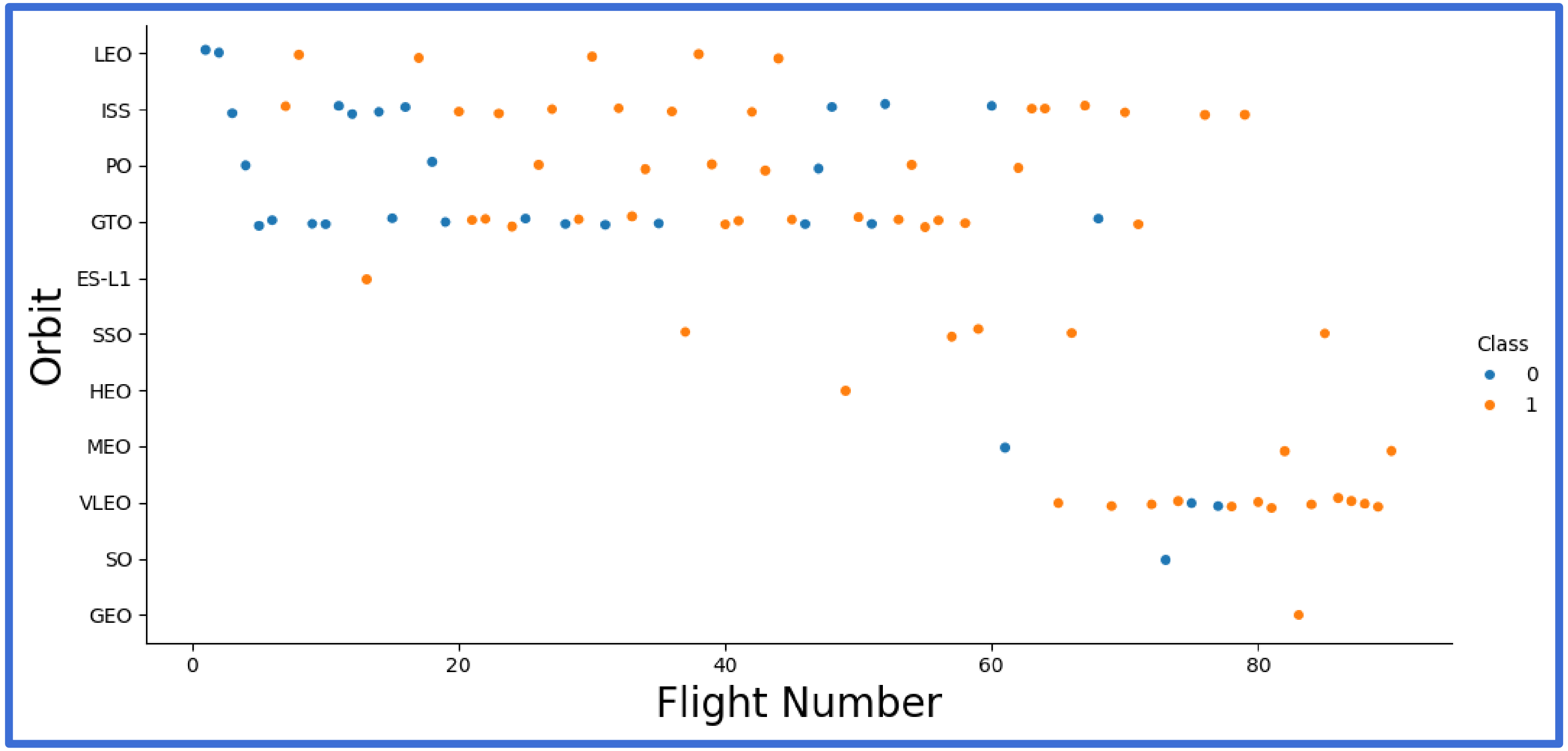
- Lagrange point 1 (ES-L1), and highly elliptical (HEO) orbit insertions were 100% successful in first stage recoveries
- Geosynchronous (GTO+GEO) had the lowest success rate of near 50%
    - Most attempts at 27+1



Note that one failed sun-synchronous orbit is labeled as SO, while the five successful ones are SSO

Flight Number vs. Orbit Type

- Lagrange point 1 (ES-L1), and highly elliptical (HEO) orbit insertions were 100% successful in first stage recoveries
- Geosynchronous (GTO+GEO) had the lowest success rate of near 50%
  - Most attempts at 27+1

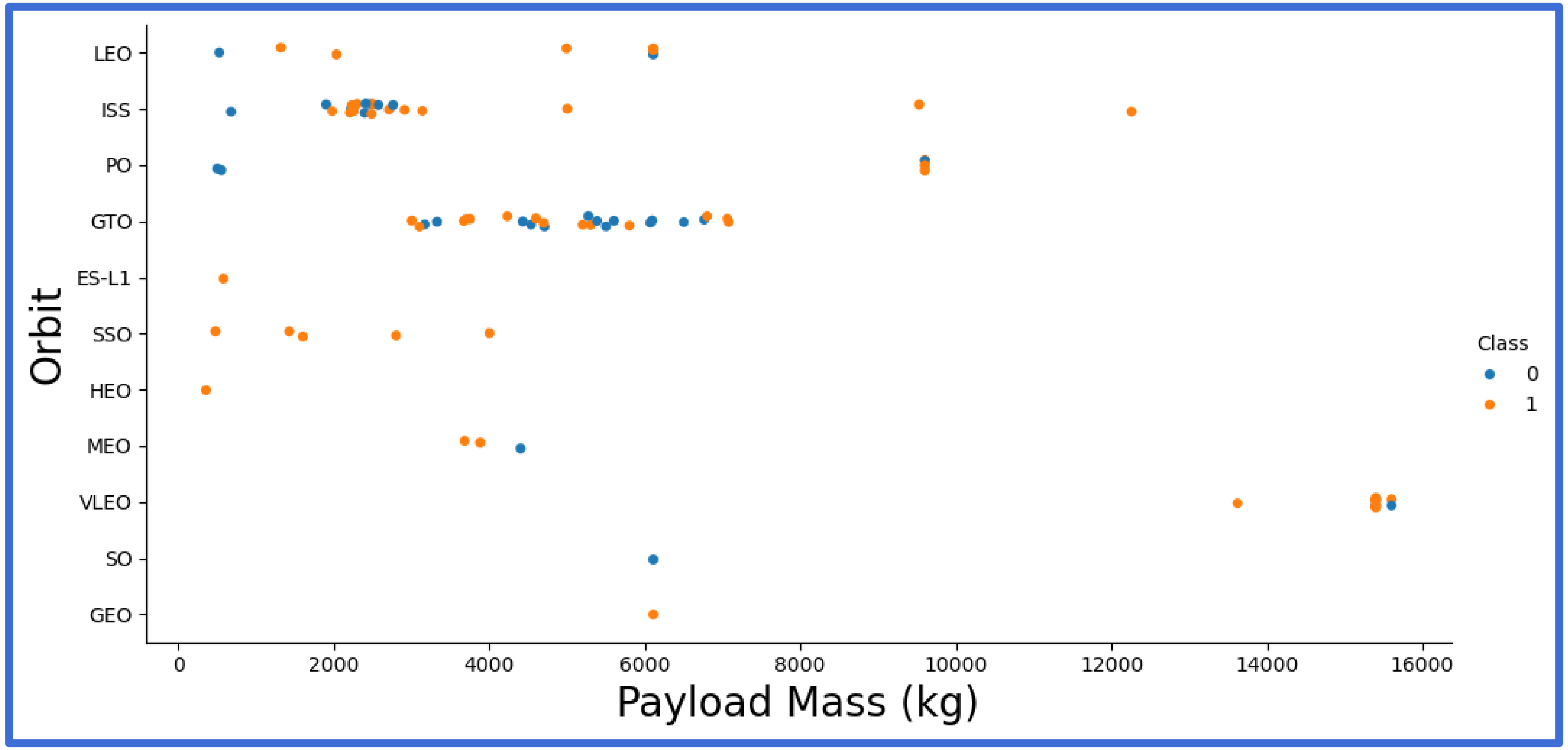


Falcon 9 first stage failed and successful landings

Note that one failed sun-synchronous orbit is labeled as SO, while the five successful ones are SSO

Payload vs. Orbit Type

- Some orbits (ex GTO) had tighter payload mass clustering
- All 13000 kg+ payloads were inserted into VLEO
- Orbital insertions requiring the most delta-v (ES-L1 and HEO) had the lowest payload mass

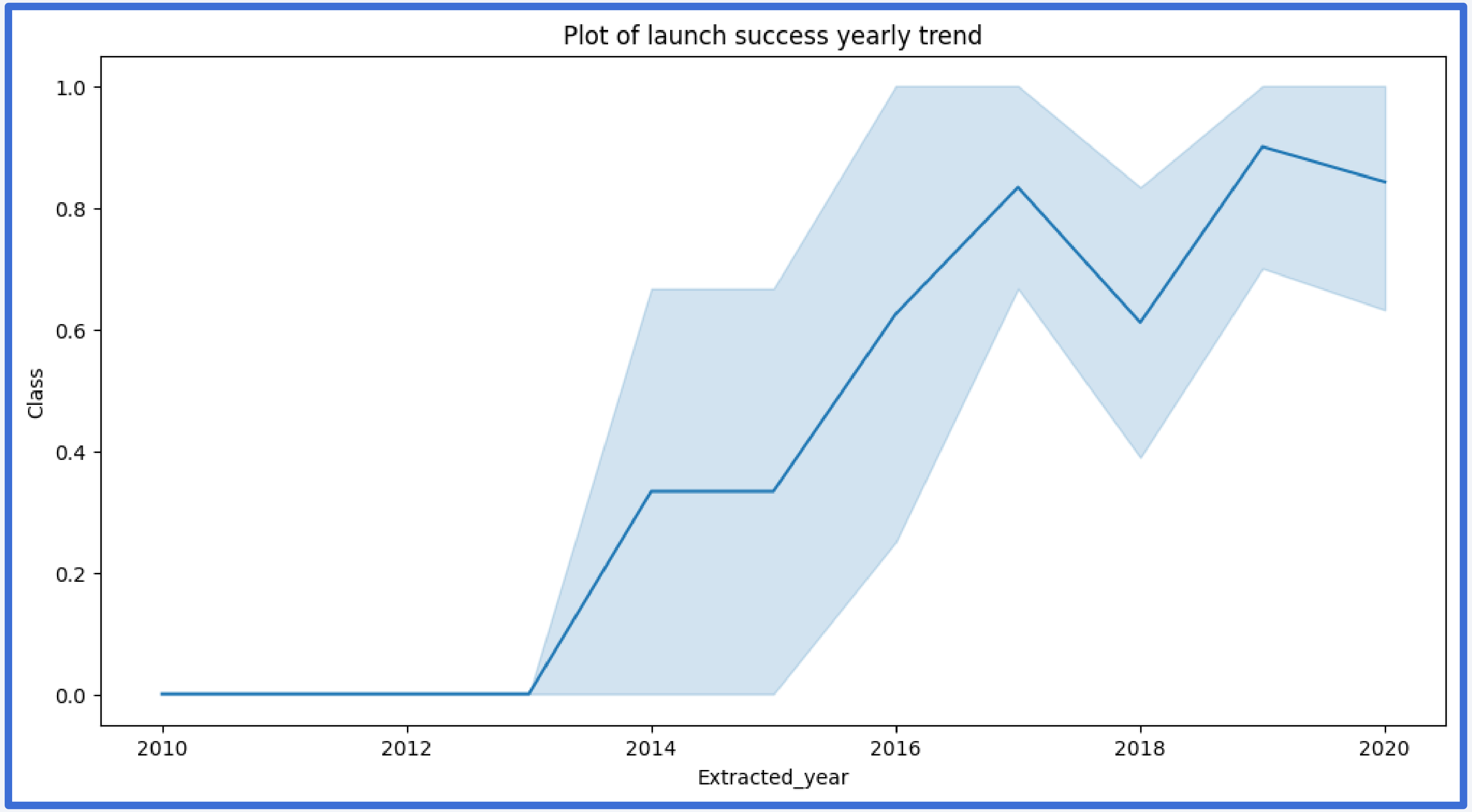


Falcon 9 first stage failed and successful landings

Note that one failed sun-synchronous orbit is labeled as SO, while the five successful ones are SSO. GTO and GEO are also both geosynchronous.

Launch Success Yearly Trend

- In general, launch success rate increased year over year
  - Except 2018 was down



All Launch Site Names

- Query:

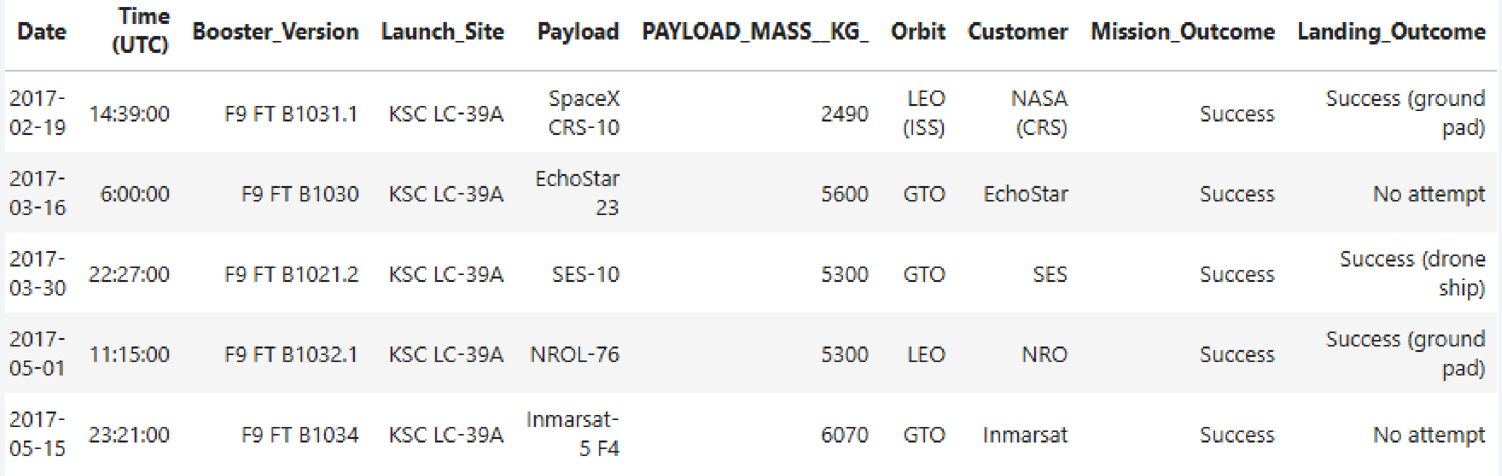
```
%sql select distinct "Launch_Site" from SPACEXTABLE2;
```

- Result:
  - CCAFS LC-40
  - VAFB SLC-4E
  - KSC LC-39A
  - CCAFS SLC-40
- Explanation: There are four different launch sites.

Launch Site Names Begin with 'KSC'

- Query: `%sql SELECT * FROM SPACEXTABLE2 WHERE "Launch_Site" LIKE 'KSC%' LIMIT 5`

- Result:



Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-03-16	6:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-05-01	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
2017-05-15	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

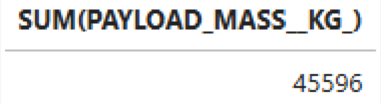
- Explanation: This sampling approach is a way to gain basic understanding of the data in the dataframe

Total Payload Mass

Query:

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE2 WHERE CUSTOMER='NASA (CRS)'
```

- Result:



SUM(PAYLOAD_MASS__KG_)
45596

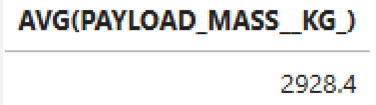
- Explanation: The total payload mass for NASA is 45,596 kg

Average Payload Mass by F9 v1.1

- Query:

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE2 WHERE
BOOSTER_VERSION='F9 v1.1’
```

- Result:



AVG(PAYLOAD_MASS__KG_)
2928.4

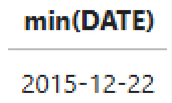
- Explanation: The average payload mass for the Falcon 9 v1.1 booster is 2,928.4 kg

First Successful Ground Landing Date

- Query:

```
%sql SELECT min(DATE) FROM SPACEXTABLE2 WHERE "Landing_Outcome"='Success
(ground pad)’
```

- Result:



min(DATE)
2015-12-22

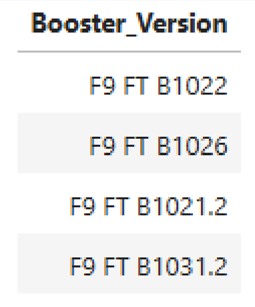
- Explanation: The first successful ground pad landing was on December 22, 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

- Query:

```
%sql SELECT BOOSTER_VERSION FROM SPACEXTABLE2 WHERE PAYLOAD_MASS__KG_
between 4000 and 6000 AND "Landing_Outcome"='Success (drone ship)';
```

- Result:



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

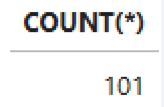
- Explanation: Four booster versions with launch payloads from 4000 to 6000 kg have had successful landings on a drone ship.

Total Number of Successful and Failure Mission Outcomes

- Query:

```
%sql SELECT COUNT(*) FROM SPACEXTABLE2 WHERE MISSION_OUTCOME LIKE
'%Success%' OR MISSION_OUTCOME LIKE '%Failure%’
```

- Result:



COUNT(*)
101

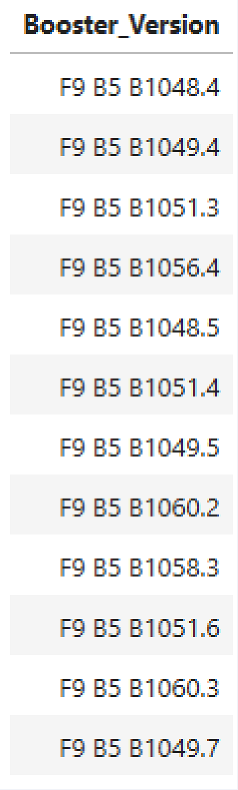
- Explanation: There are 101 total missions in the dataset that have a mission outcome.

Boosters Carried Maximum Payload

- Query:

```
%sql SELECT BOOSTER_VERSION FROM SPACEXTABLE2 WHERE PAYLOAD_MASS__KG_ =
(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

- Result:



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

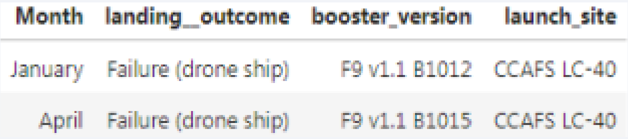
Explanation: Twelve different Falcon 9 booster versions had the maximum payload mass of 15,600 kg in the dataset.

2015 Launch Records

- Query:

```
%sql SELECT MONTHNAME(`DATE`) AS ‘Month’, `landing__outcome`,
`booster_version`, `launch_site` FROM `SPACEXTABLE2` WHERE
`landing__outcome` = 'Failure (drone ship)' AND YEAR(`DATE`) = 2015;
```

- Result:



Month	landing__outcome	booster_version	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

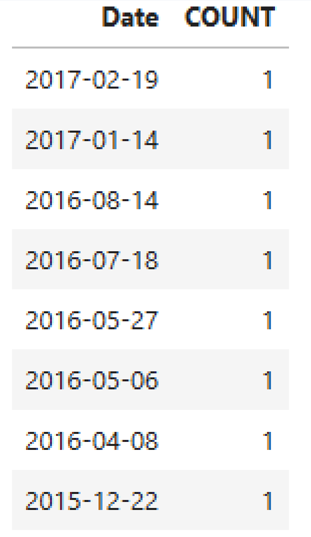
- Explanation: There were two drone ship failures in 2015

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

- Query:

```
%sql SELECT "DATE", COUNT("Landing_Outcome") as COUNT FROM SPACEXTABLE2
WHERE "DATE" BETWEEN '2010-06-04' and '2017-03-20' AND "Landing_Outcome"
LIKE '%Success%' GROUP BY "DATE` ORDER BY COUNT("Landing_Outcome") DESC
```

- Result:



Date	COUNT
2017-02-19	1
2017-01-14	1
2016-08-14	1
2016-07-18	1
2016-05-27	1
2016-05-06	1
2016-04-08	1
2015-12-22	1

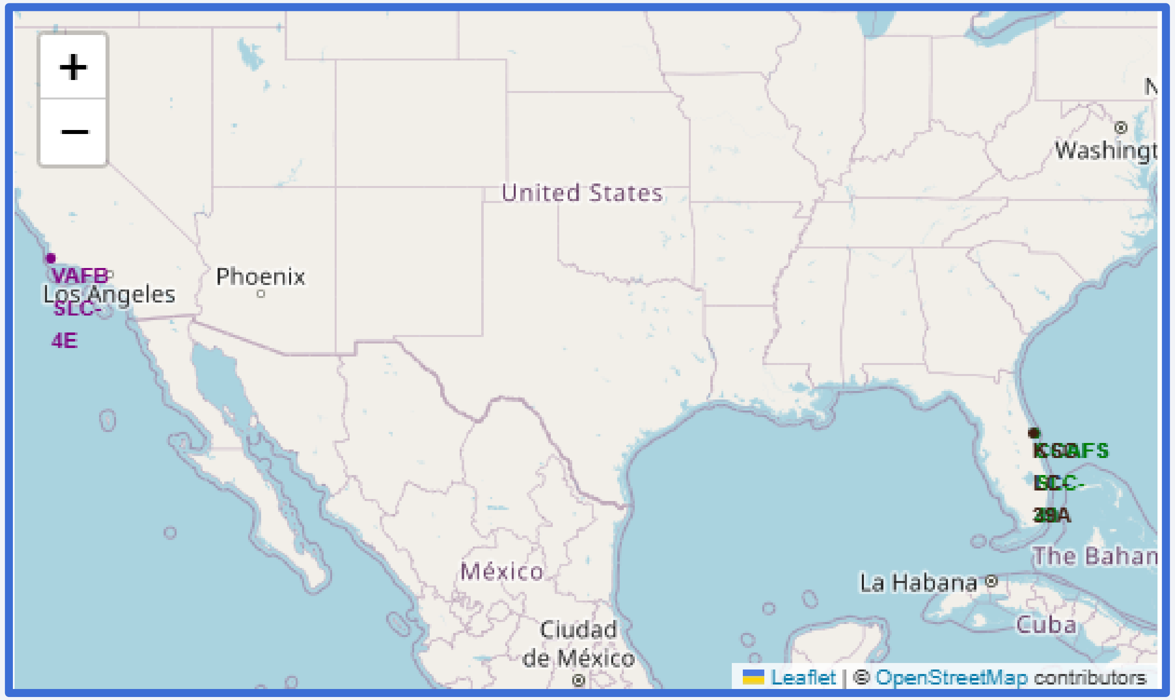
- Explanation: There were eight successful landings in the given date range.

Section 3

Launch Sites Proximities Analysis

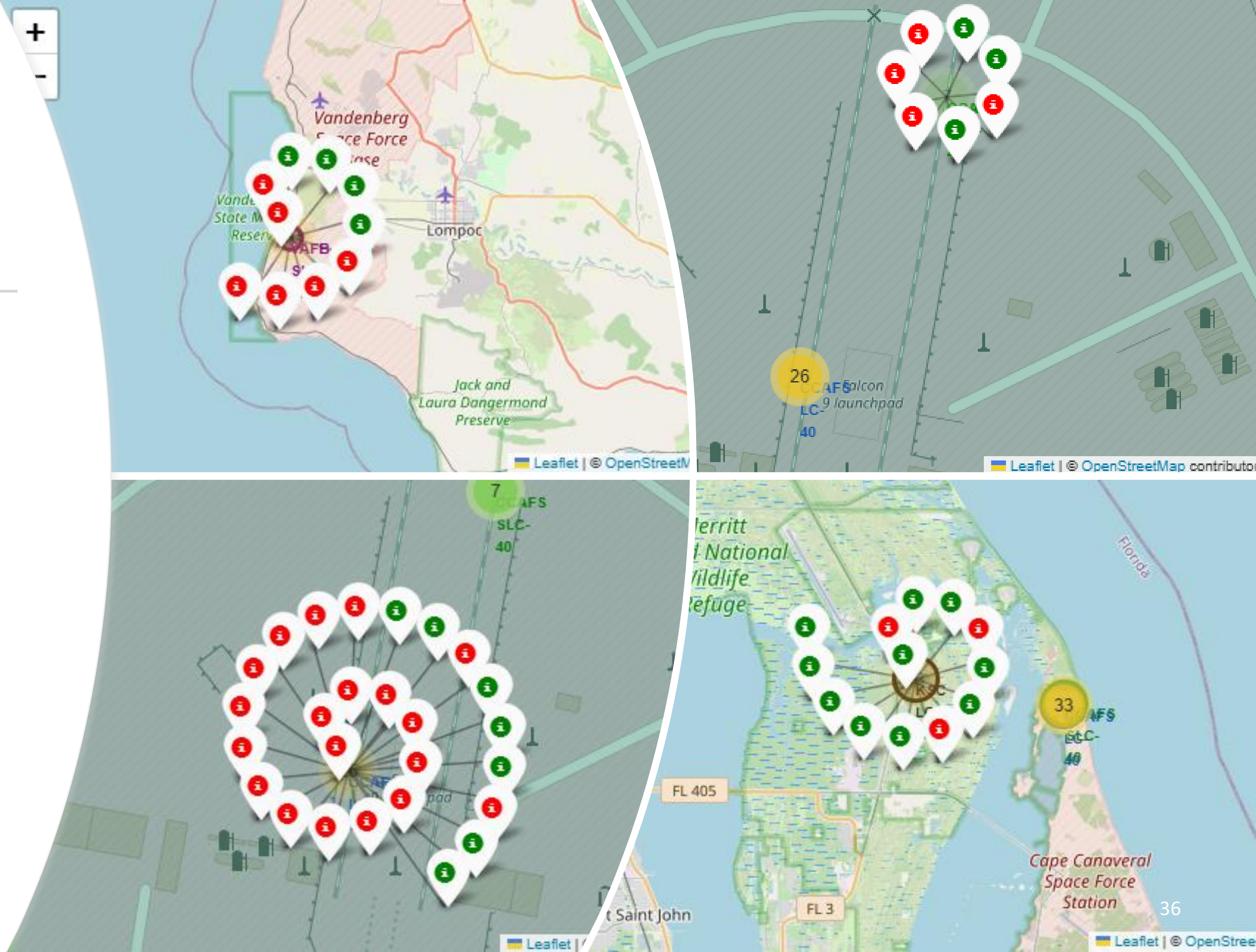
SpaceX Falcon 9 Launch Site Locations

- All but one (Vandenberg AFB) launch site is located in Florida
- All launch sites are near the ocean
- All launch sites are in the southern portion of the US



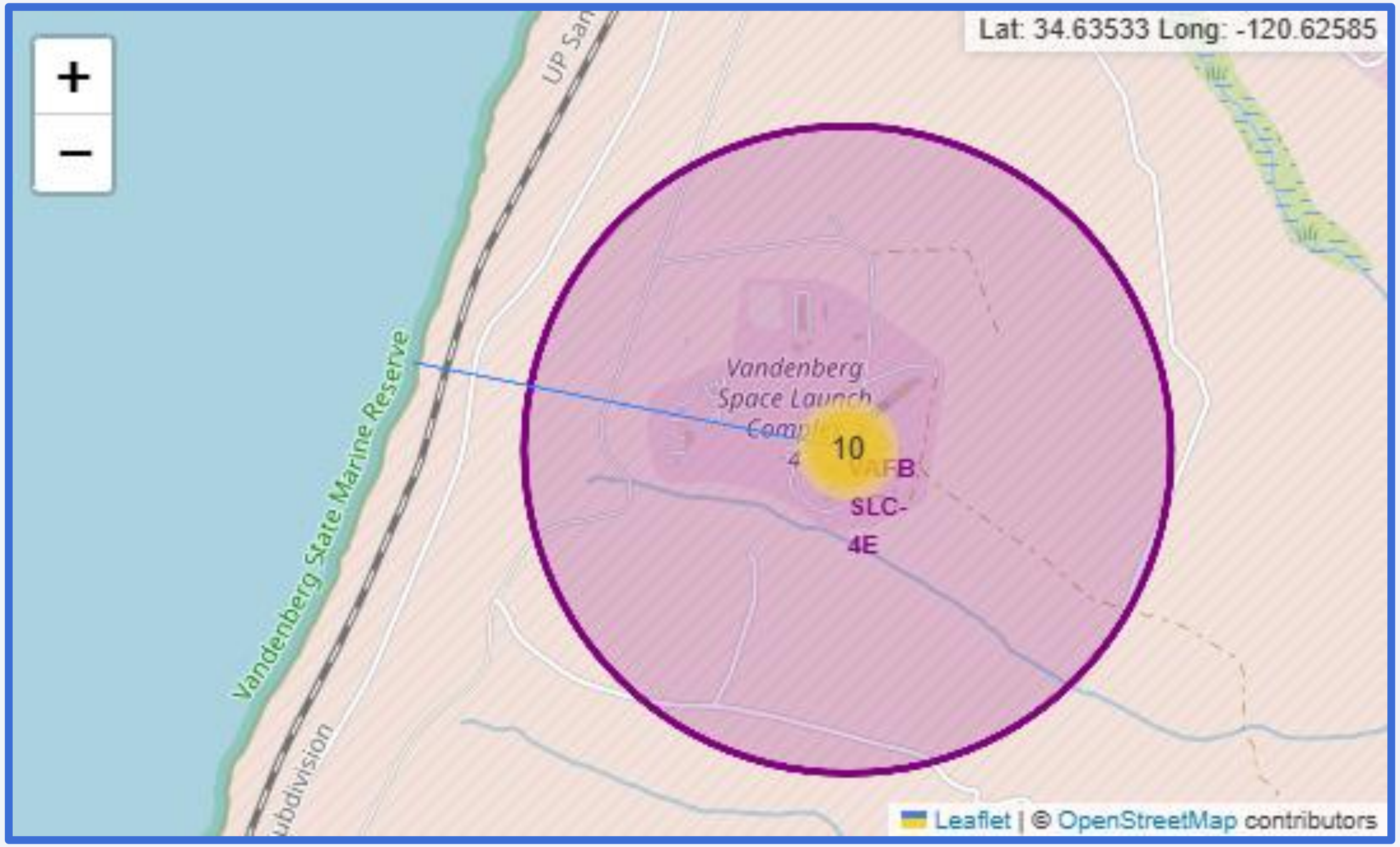
Launch Site Successes and Failures

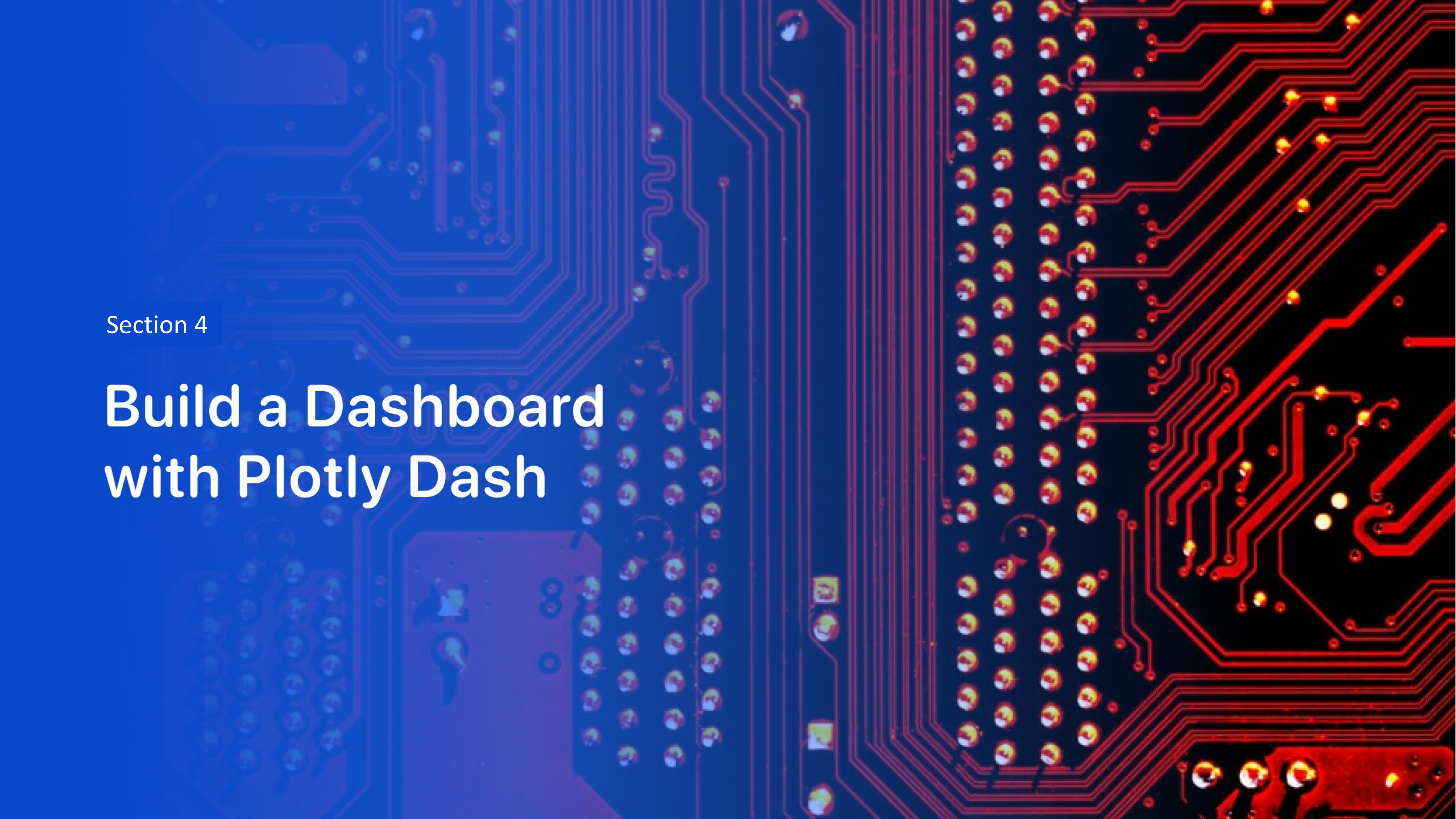
- All launch sites have successes and failures
- Earlier launches were more likely to have failed stage 1 booster recoveries
- Cape Canaveral had a higher success rate



Vandenberg

- This launch site is close to the Pacific Ocean


Lat: 34.63533 Long: -120.62585
+
−
UP San
Vandenberg State Marine Reserve
Subdivision
Vandenberg Space Launch Complex 4
10
VAFB SLC-4E
Leaflet | © OpenStreetMap contributors

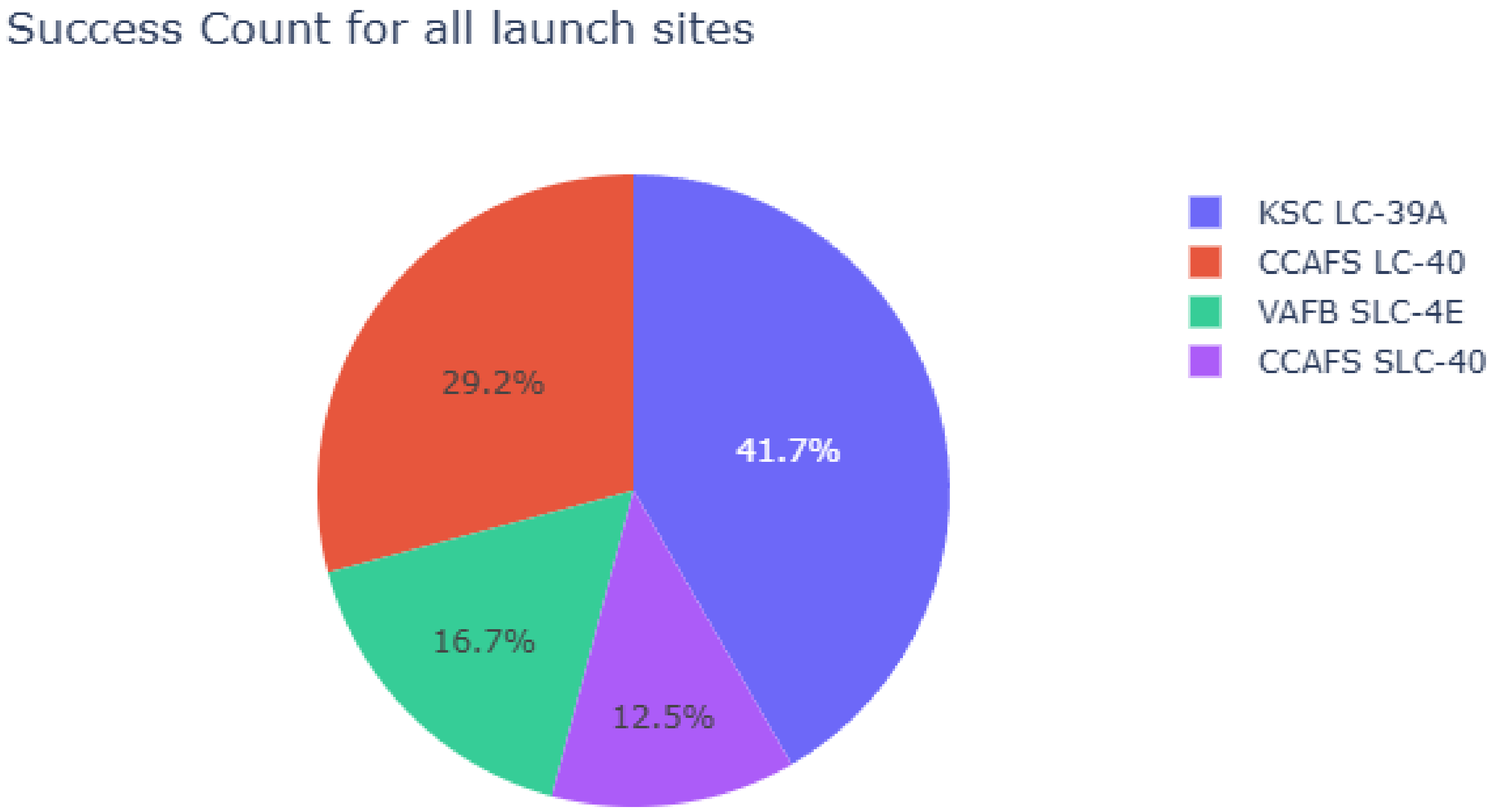


Section 4

Build a Dashboard with Plotly Dash

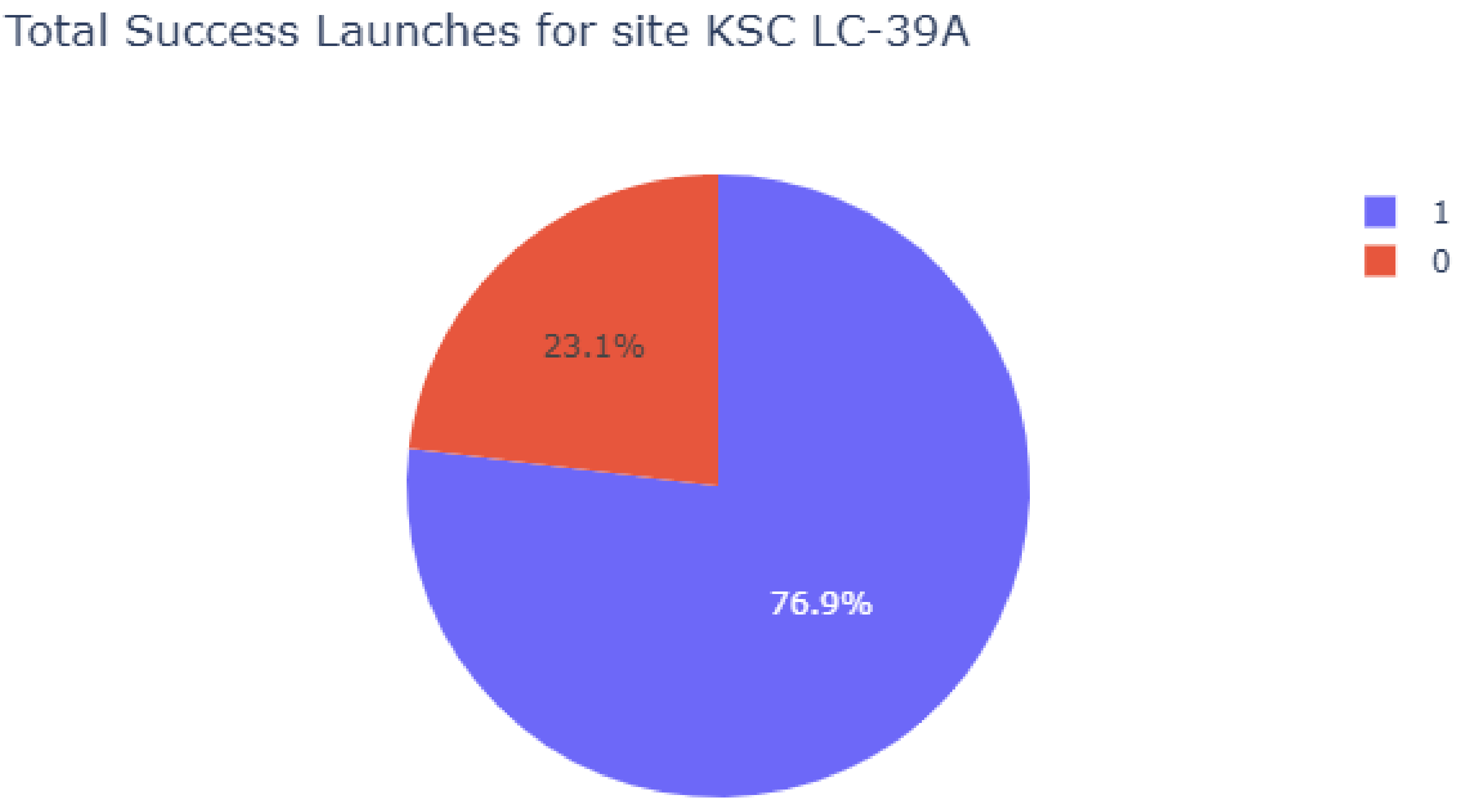
Successful Falcon 1 First Stage Recoveries

- Most successful landings were launched in Florida
  - VAFB only accounted for 16.7% of successful first stage landings
- The plurality (41.7%) of successful first stage recoveries were launched at KSC LC-39A



Successes at KSC LC-39A

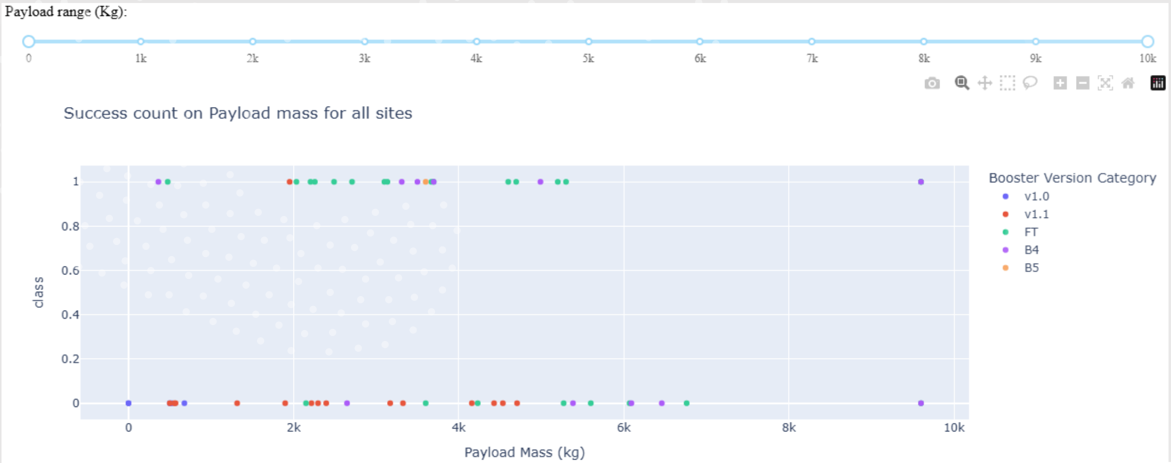
- Of the Falcon 9 launches at Kennedy Space Center LC-39A
  - 76.9% resulted in successful first stage recoveries
  - 23.1% did not successfully recover the firs stage

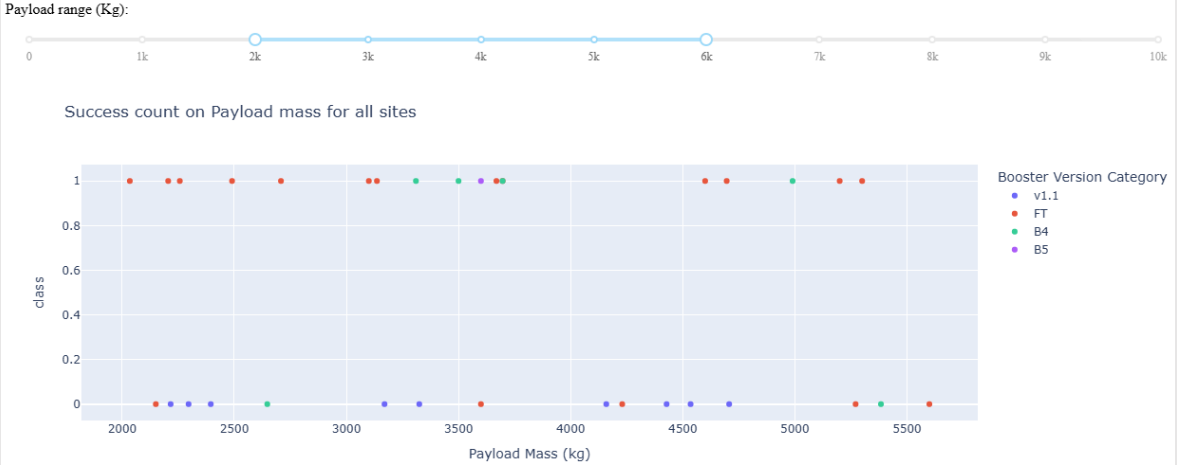


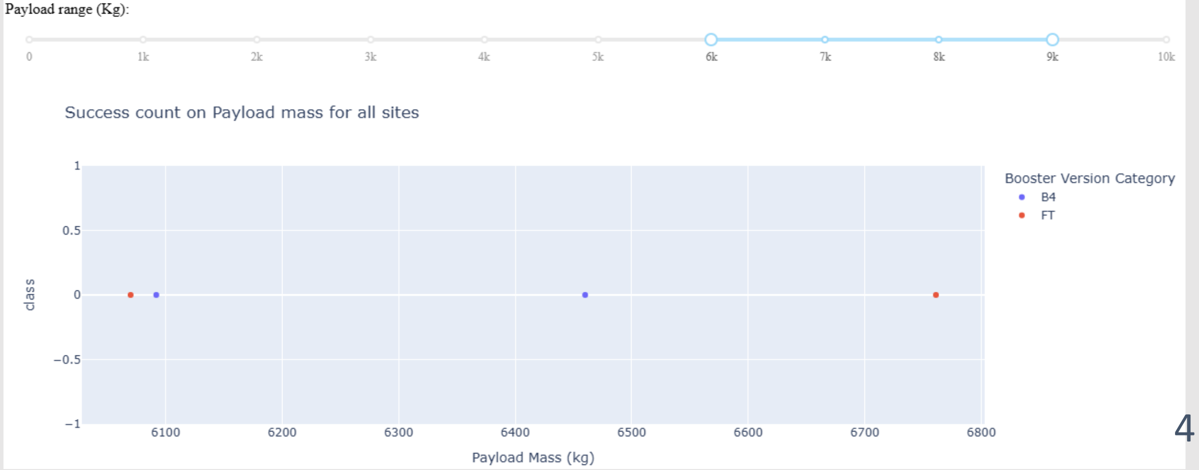
Total Success Launches for site KSC LC-39A

Success vs Payload Mass

- Most successful payloads under 10,000 kg clustered in the 2,000 to 5,000 kg range
- There were no successful Falcon 9 recoveries from 6,000 to 7,000 kg payload masses
- The FT booster version had success across a wide payload mass range







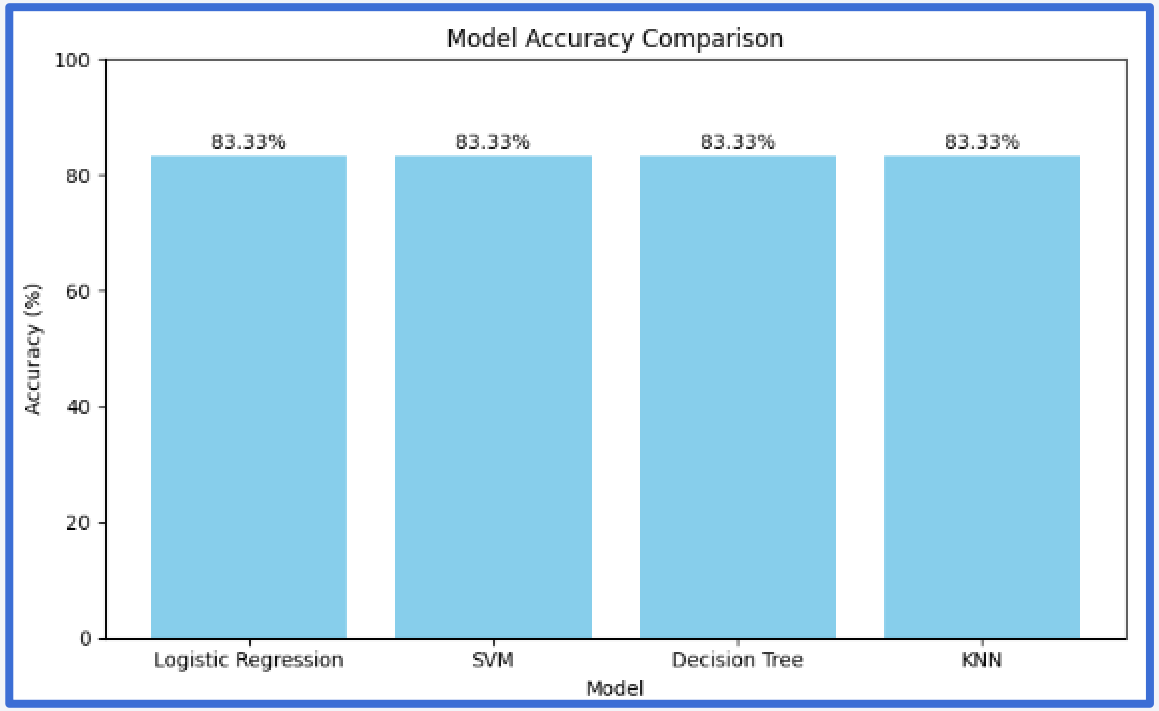
Section 5

Predictive Analysis (Classification)

Classification Accuracy

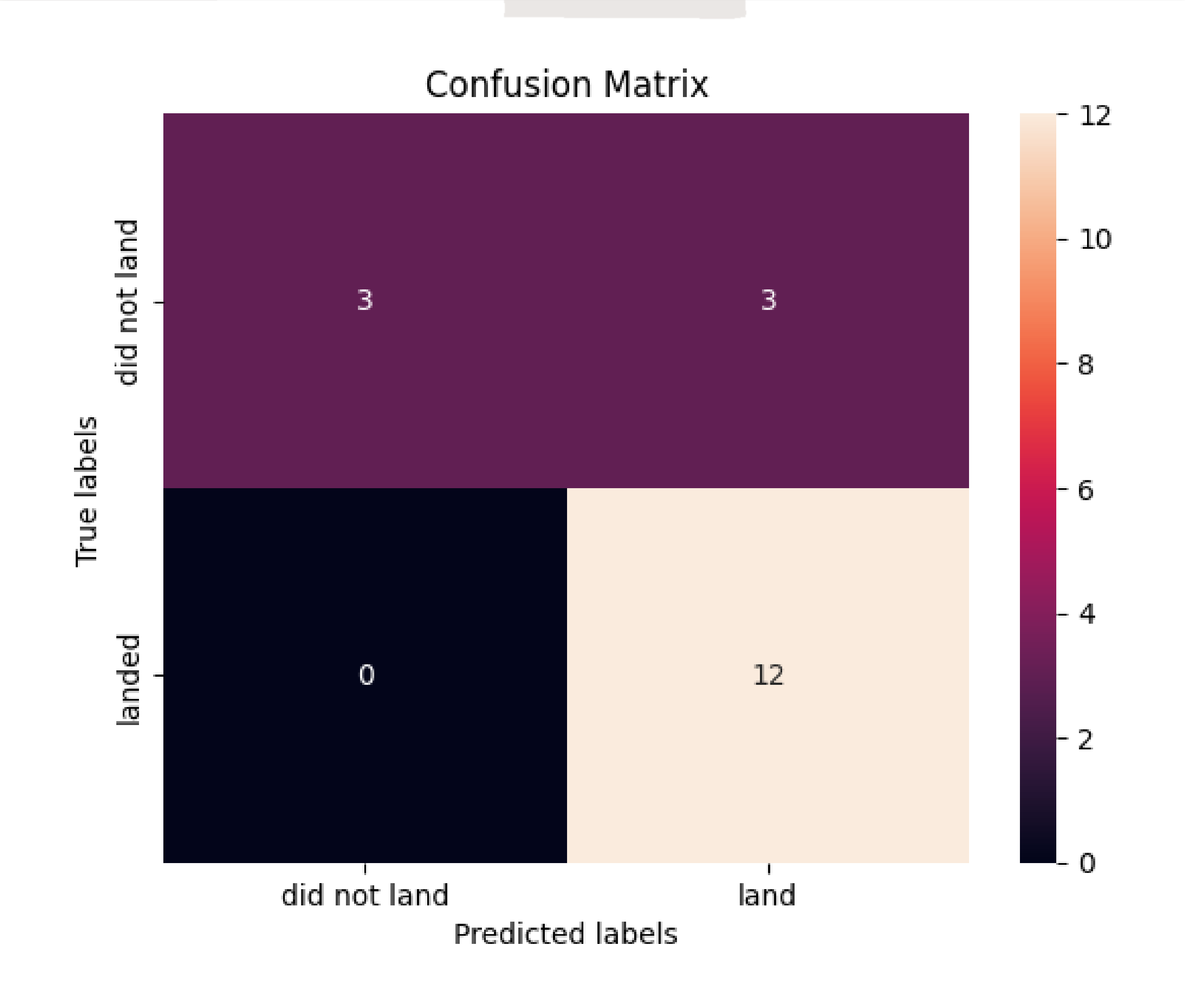
- All models resulted in an 83.33% accuracy against the Test data set

- Testing models against additional (post-2021) data needed to determine best model
  - And allow for further refinement



Confusion Matrix

- The best performing models successfully identified
  - Twelve true positives
  - Three true negatives
  - No false negatives
- The best performing models failed with
  - Three false positives, where the model predicted success, but the landing actually failed



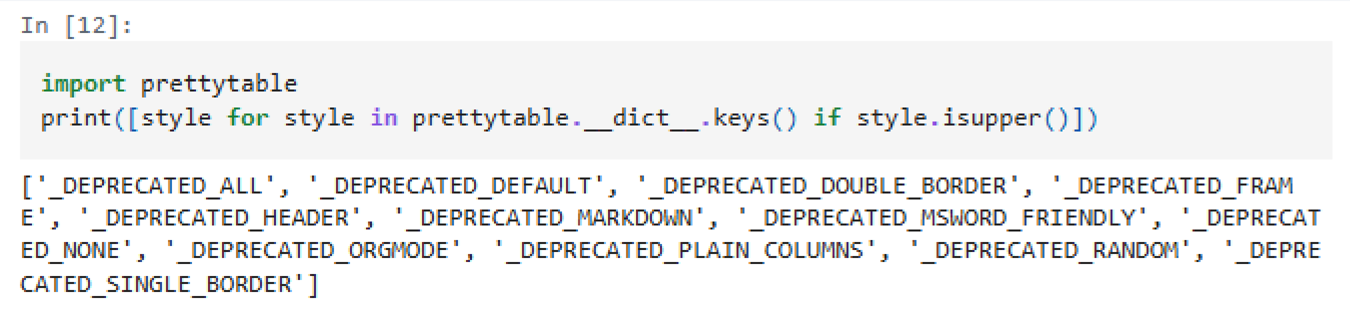
Conclusions

- Over time, SpaceX Falcon 9 first stage landing outcomes have improved
- Florida-based launches have been more successful than California-based
- Lower payload mass and lower Δv orbits improve first stage recovery outcomes
- With even the limited given dataset, all four machine learning models predicted SpaceX Falcon9 first stage learning outcomes with >80% accuracy
  - More data needed to refine models and determine "best" modeling approach

https://commons.wikimedia.org/wiki/File:SpaceX,_Falcon_9,_Amazonas_Nexus,_02-06-2023_%2852673414272%29.jpg

Appendix

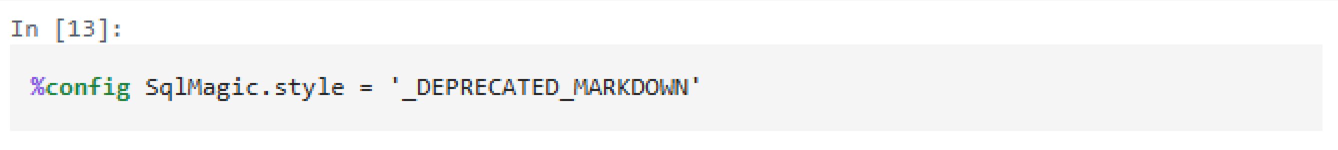
- In the SQL lab (lab 4), I needed to import prettytable and set %sql style as ‘_DEPRECATED_MARKDOWN’ in order to display results of queries



In [12]:

```python
import prettytable
print([style for style in prettytable.__dict__.keys() if style.isupper()])
```

```
['_DEPRECATED_ALL', '_DEPRECATED_DEFAULT', '_DEPRECATED_DOUBLE_BORDER', '_DEPRECATED_FRAM
E', '_DEPRECATED_HEADER', '_DEPRECATED_MARKDOWN', '_DEPRECATED_MSWORD_FRIENDLY', '_DEPRECAT
ED_NONE', '_DEPRECATED_ORGMODE', '_DEPRECATED_PLAIN_COLUMNS', '_DEPRECATED_RANDOM', '_DEPRE
CATED_SINGLE_BORDER']
```



In [13]:

```
%config SqlMagic.style = '_DEPRECATED_MARKDOWN'
```

Appendix

- Python code snippet using matplotlib to generate a bar chart of the model results

```
import matplotlib.pyplot as plt

# Prepare the data
models = ['Logistic Regression', 'SVM', 'Decision Tree', 'KNN']
accuracies = [83.33, 83.33, 83.33, 83.33]

# Create the bar chart
plt.figure(figsize=(8, 5))
bars = plt.bar(models, accuracies, color='skyblue')

# Add accuracy labels on top of bars
for bar in bars:
    height = bar.get_height()
    plt.text(bar.get_x() + bar.get_width()/2, height + 0.5, f'{height:.2f}%',
ha='center', va='bottom')

# Set title and labels
plt.title('Model Accuracy Comparison')
plt.ylabel('Accuracy (%)')
plt.ylim(0, 100)
plt.xlabel('Model')

# Show plot
plt.tight_layout()
plt.show()
```

Thank you!