IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

James Toner
29 December, 2023

Outline

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

Executive Summary

- Methodologies
  - SpaceX launch data was collected with API Calls and Webscraping techniques.
  - The data was then cleaned and formatted into workable forms.
  - Exploratory data analysis was performed on the dataset with SQL queries and visualization techniques in order to get a sense of which variables affected launch outcomes in meaningful ways.
  - An interactive map and dashboard were created in order to explore certain relationships with ease.
  - Four different classification models were developed, tuned, and evaluated to find the best predictive model.
- Results
  - Successful launch outcomes appear to be influenced by orbit type, payload mass, site location, booster type, and flight number.
  - The most effective classification model was found to be a Decision Tree Classifier, with an R2 score of 0.944 on the testing dataset.

Introduction

- SpaceX offers much cheaper rocket launches than its competitors due to its ability to reuse the first stage rocket boosters.
- We want to be able to predict if a first stage booster will land successfully, as this is the main driver for the cheaper price offered by SpaceX.
- This information could potentially be used by a competitor company wanting to bid against SpaceX for a launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API Call
  - Webscraping
- Perform data wrangling
  - Replacing null values and creating new data columns with more workable data forms.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building, tuning, and evaluating four different classification models.

Data Collection – SpaceX API

API Call

- `response = requests.get(spacex_url)`

Decode

- Decode response as JSON and store in temporary DataFrame
- `data = pd.json_normalize(response.json())`

Dictionary

- `Filter and extract key variables and store in dictionary.`

DataFrame

- Read dictionary into final DataFrame
- `df = pd.DataFrame({key:pd.Series(value) for key, value in launch_dict.items()})`

API notebook: https://github.com/tjapple/IBM_proj/blob/main/spacex_data_collection_api.ipynb

Data Collection - Scraping

Request

- HTTP GET Request
- `response = requests.get(static_url).text`

BeautifulSoup

- Create BeautifulSoup Object
- `soup = BeautifulSoup(response)`

Dictionary

- Extract text data from soup and store in dictionary

DataFrame

- Read dictionary into pandas DataFrame
- `df= pd.DataFrame({ key:pd.Series(value) for key, value in launch_dict.items() })`

Webscraping notebook: https://github.com/tjapple/IBM_proj/blob/main/spacex_webscraping.ipynb

Data Wrangling

Identify

- Identify null values
- data_falcon9.isnull().sum()

Replace

- Replace null values with appropriate metric.
- In this case, we replace NaN with mean mass.

Identify

- Identify failed landing outcomes
- `landing_outcomes = df.value_counts('Outcome')`

Create

- Create new column for "Class" of landing outcome
- Failed outcome types are given a value of 0, successful outcomes are given a value of 1.

Data wrangling notebook: https://github.com/tjapple/IBM_proj/blob/main/data_wrangling.ipynb

EDA with Data Visualization

- Flight Number vs Launch Site and Payload Mass vs Launch Site
  - Explore effects of flight number and payload mass on launch outcome at each site.
- Orbit Type vs Success Rate
  - Explore how the type of orbit affects rate of success.
- Orbit Type vs Payload Mass and Orbit Type vs Flight Number
  - Explore effects of payload mass and flight number on launch outcome for each orbit type.
- Launch Outcome Success Rate over Time
  - Visualize the trend of success rate over the years.

- EDA with visualization notebook: https://github.com/tjapple/IBM_proj/blob/main/wk2_visualization.ipynb

EDA with SQL

- Launch site location queries.
- Payload mass queries and which boosters carried the maximum.
- Successful landings within certain payload mass ranges.
- First successful landing dates for certain types of landings.
- Landing outcomes for certain date ranges.

- EDA with SQL notebook: https://github.com/tjapple/IBM_proj/blob/main/wk2_SQL.ipynb

Interactive Map with Folium

- Launch site locations with color-coded launch outcomes and lines indicating distances to nearest railways, coastlines, and highways.
- Exploring important variables between launch locations and their possible influence on launch outcomes.

- Folium notebook: https://github.com/tjapple/IBM_proj/blob/main/folium_launch_site_location.ipynb

Interactive Dashboard with Plotly

- Total successful landings by site location.
  - Pie chart to visualize locations with the most amount of successful landings.
- Launch outcomes per site location.
  - Pie chart to visualize success rates depending on location.
- Launch outcome vs payload mass per booster category.
  - Scatter plot with interactive payload mass slider.
  - Analyze success rates for booster categories, depending on payload mass.

- Plotly Dash code file: https://github.com/tjapple/IBM_proj/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

Split data into train/test sets → Create the 4 models on the training data → Optimize hyper-parameters with grid search → Use optimized models on test data → Visualize model results

- Test data was a 0.2 split.
- 4 models: Logistic Regression, Support Vector Machine, Decision Tree Classifier, K-Nearest Neighbors.
- Predictive Analysis notebook: https://github.com/tjapple/IBM_proj/blob/main/Machine_Learning_Prediction.ipynb

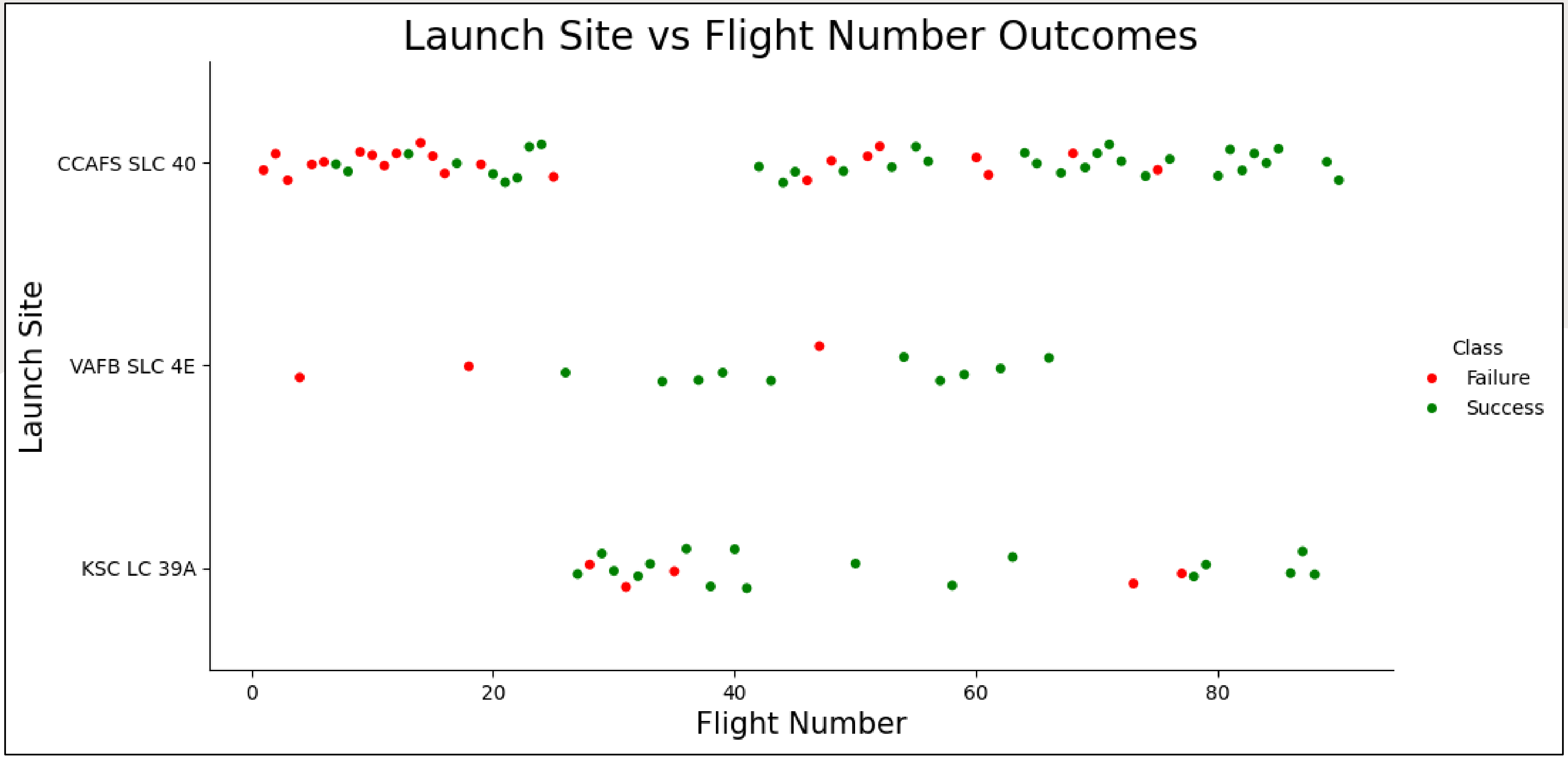
Results

- Successful launch outcomes appear to be influenced by orbit type, payload mass, site location, booster type, and flight number.
- Interactive Folium map displaying launch locations and proximities to landmarks.
- Interactive Plotly dashboard exploring success rates at different launch site locations and the effect of payload mass on the success rate of the different booster types.
- The most effective classification model was found to be a Decision Tree Classifier, with an R2 score of 0.944 on the testing dataset.

Section 2

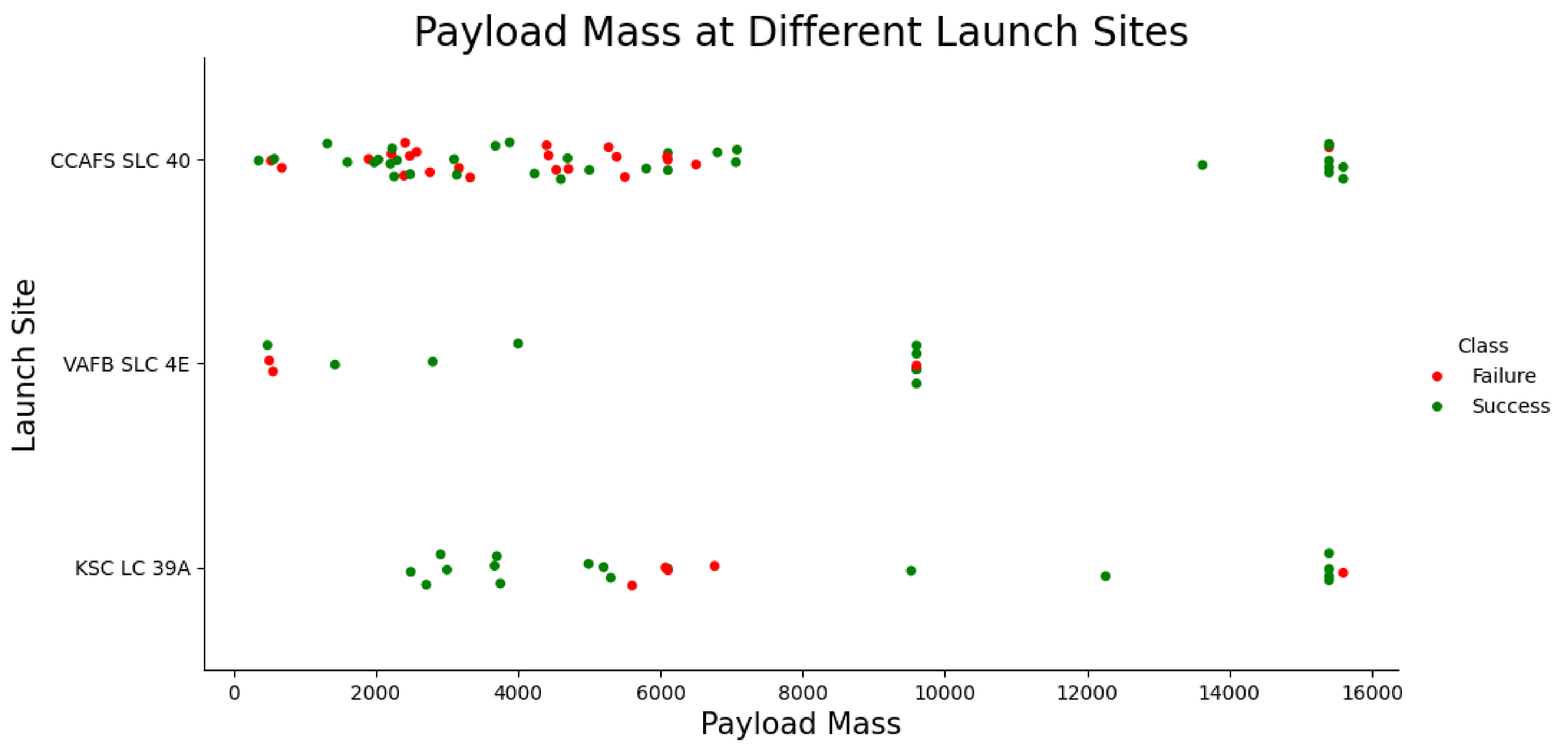
Insights drawn from EDA

Flight Number vs. Launch Site



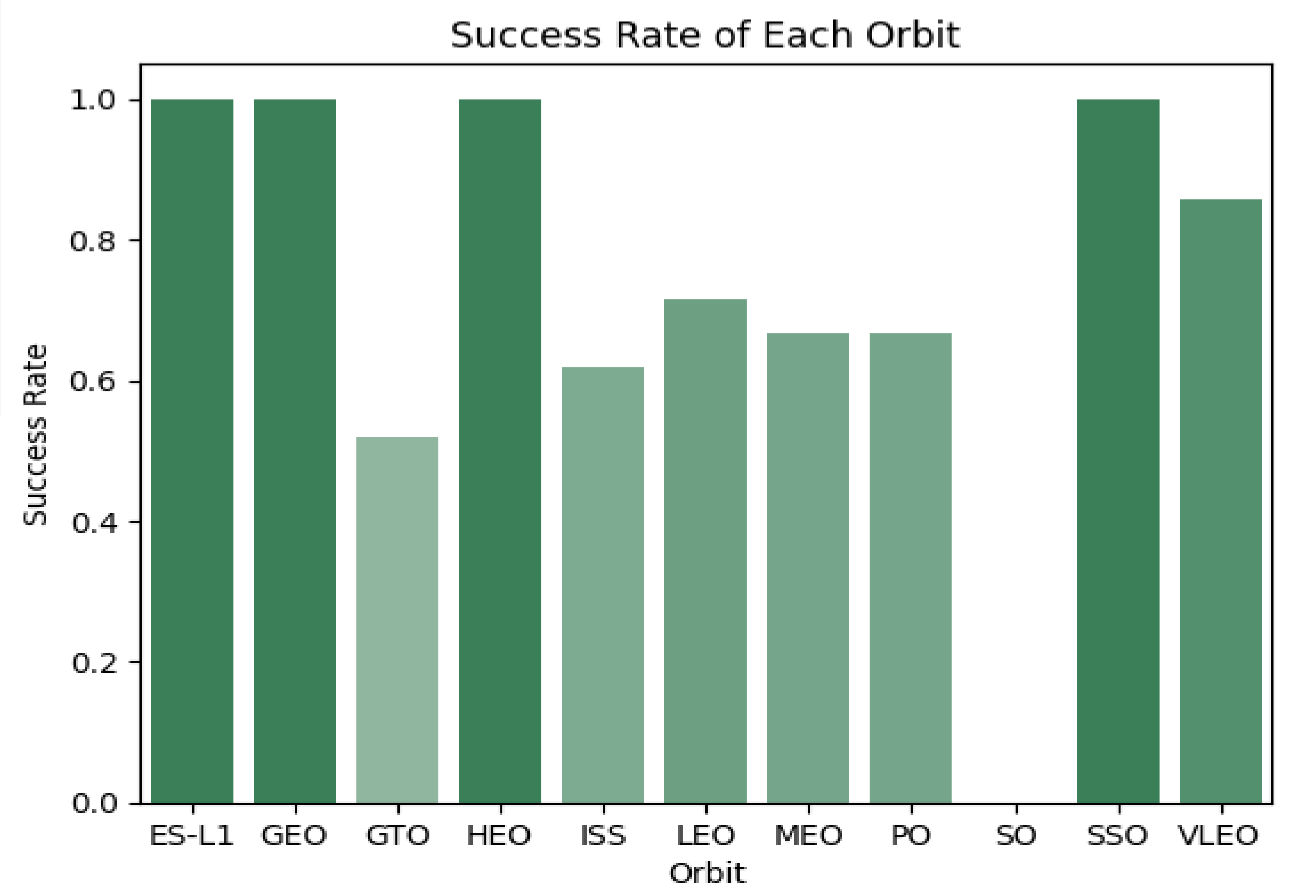
- Launch Site KSC LC 39A was not used for the first few dozen launches.
- Lanch Site VAFB SLC 4E was not used in the last few dozen launches.
- Success rate increases as the flight number increases.

Payload Mass vs. Launch Site



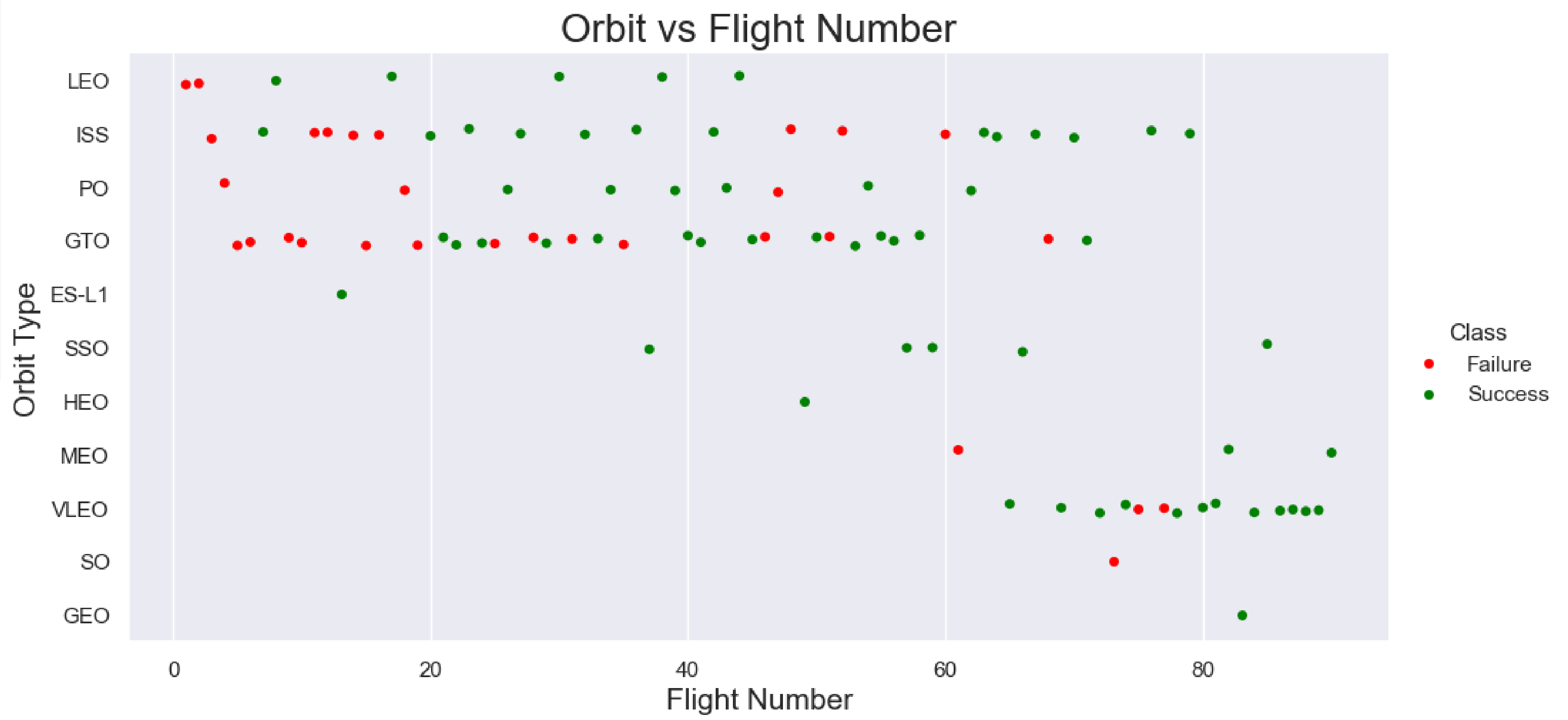
- No rockets are launched at VAFB SLC 4E with a payload greater than 10000 kg.

Success Rate vs Orbit Type



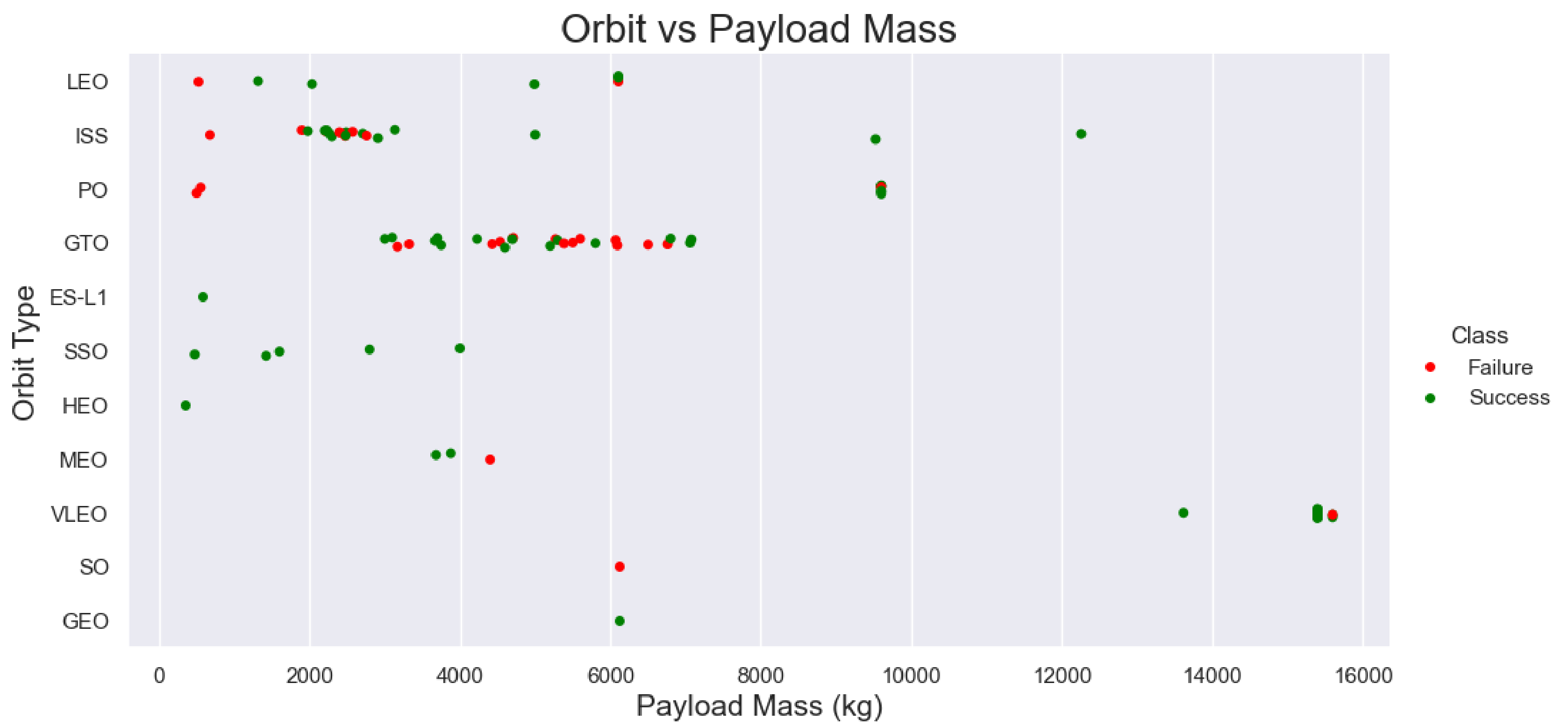
- ES-L1, GEO, HEO, SSO, and VLEO all have success rates over 0.8
- SO has the lowest success rate at 0

Orbit vs Flight Number



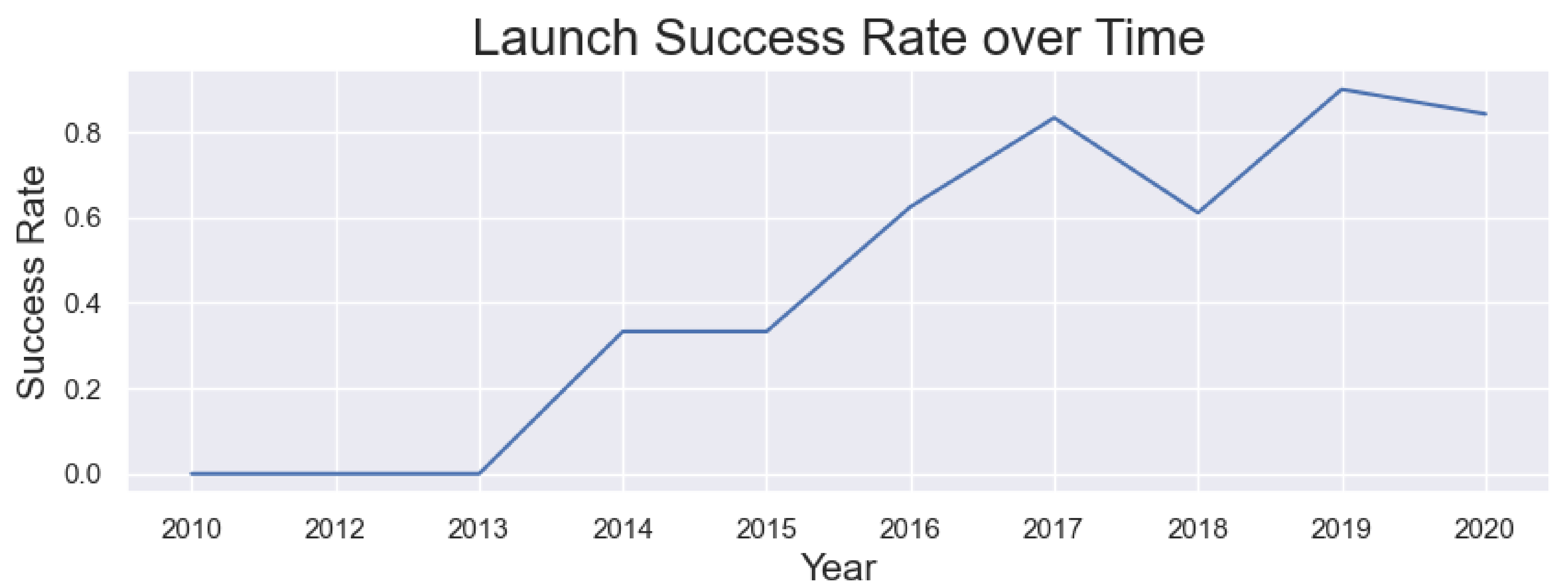
- With LEO orbit, success appears to be related to flight number.
- Strong correlations elsewhere cannot be found.

Orbit vs Payload Mass



- ISS and PO appear to have better success rates with heavier payloads.
  - Strong correlations cannot be found elsewhere

Launch Success Rate over Time



- Success rates have steadily increased since the year 2013.

SQL: All Launch Site Names

```
1  pd.read_sql('SELECT launch_site, COUNT(launch_site) FROM spacextable GROUP BY launch_site', con)
✓ 0.0s
```

	Launch_Site	COUNT(launch_site)
0	CCAFS LC-40	26
1	CCAFS SLC-40	34
2	KSC LC-39A	25
3	VAFB SLC-4E	16

- Unique launch cites with the number of launches from each.

SQL: Launch Site Names Beginning with 'CCA'

```
pd.read_sql_query('SELECT * FROM spacextable WHERE Launch_Site LIKE "CCA%" limit 5', con)
```

0s Python

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- First 5 records where launch cite names begin with 'CCA'.

SQL: Total Payload Mass

```
pd.read_sql('''SELECT SUM(PAYLOAD_MASS__KG_) AS "Total Payload (kg)"
            FROM spacextable WHERE Customer LIKE "NASA (CRS)"''', con)
```

✓ 0.0s

	Total Payload (kg)
0	45596

- NASA boosters carried a total payload of 45,596 kilograms.

SQL: Average Payload Mass by F9 v1.1

```
pd.read_sql('''SELECT AVG(PAYLOAD_MASS__KG_) AS "Average Payload (kg)"
                    FROM spacextable
                    WHERE Booster_Version LIKE "F9 v1.1%"''', con)
```

✓ 0.0s

	Average Payload (kg)
0	2534.666667

- The booster version F9 v1.1 carried an average payload of 2,535 kilograms.

SQL: First Successful Ground Landing Date

```
pd.read_sql('''SELECT Date AS "First Date" FROM spacextable
                WHERE Landing_Outcome LIKE "Success (ground pad)"
                ORDER BY Date LIMIT 1''', con)
```

0.0s

First Date
2015-12-22

- The date of the first successful ground landing was December 22, 2015

SQL: Successful Drone Ship Landings with Payloads between 4000 and 6000

```python
pd.read_sql('''SELECT Booster_Version FROM spacextable
            WHERE Landing_Outcome LIKE "Success (drone ship)"
            AND PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000''', con)
```

✓ 0.0s

	Booster_Version
0	F9 FT B1022
1	F9 FT B1026
2	F9 FT B1021.2
3	F9 FT B1031.2

- This query retrieves unique booster versions that have successful drone ship landings while carrying payloads between 4,000 and 6,000 kilograms.

SQL: Value Counts of Mission Outcomes

```python
pd.read_sql('''SELECT Mission_Outcome, COUNT(Mission_Outcome) AS "Total"
                FROM spacextable GROUP BY Mission_Outcome''', con)
```

✓ 0.0s

	Mission_Outcome	Total
0	Failure (in flight)	1
1	Success	98
2	Success	1
3	Success (payload status unclear)	1

- There has been only one failed mission.
- Nearly all of the missions were successful.

SQL: Boosters That Carried Maximum Payload

```
pd.read_sql('''SELECT Booster_Version FROM spacextable
            WHERE PAYLOAD_MASS__KG_ =
                (SELECT MAX(PAYLOAD_MASS__KG_) FROM spacextable)''', con)

```

✓ 0.0s

	Booster_Version
0	F9 B5 B1048.4
1	F9 B5 B1049.4
2	F9 B5 B1051.3
3	F9 B5 B1056.4
4	F9 B5 B1048.5
5	F9 B5 B1051.4
6	F9 B5 B1049.5
7	F9 B5 B1060.2
8	F9 B5 B1058.3
9	F9 B5 B1051.6
10	F9 B5 B1060.3
11	F9 B5 B1049.7

- List of boosters that have all carried the maximum payload mass.

SQL: 2015 Launch Records

```
pd.read_sql('''SELECT Date, Booster_Version, Launch_Site FROM spacextable
            WHERE Landing_Outcome = "Failure (drone ship)"
            AND substr(Date, 0, 5) = "2015"''', con)
```

✓ 0.0s

	Date	Booster_Version	Launch_Site
0	2015-01-10	F9 v1.1 B1012	CCAFS LC-40
1	2015-04-14	F9 v1.1 B1015	CCAFS LC-40

- Launches in 2015 that resulted in failed landings on drone ships.

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

```
pd.read_sql('''SELECT Date, Landing_Outcome, COUNT(Landing_Outcome) AS "Count"
                FROM spacextable
                WHERE Date BETWEEN "2010-06-04" and "2017-03-20"
                GROUP BY Landing_Outcome
                ORDER BY Count DESC''', con)
```

✓ 0.0s

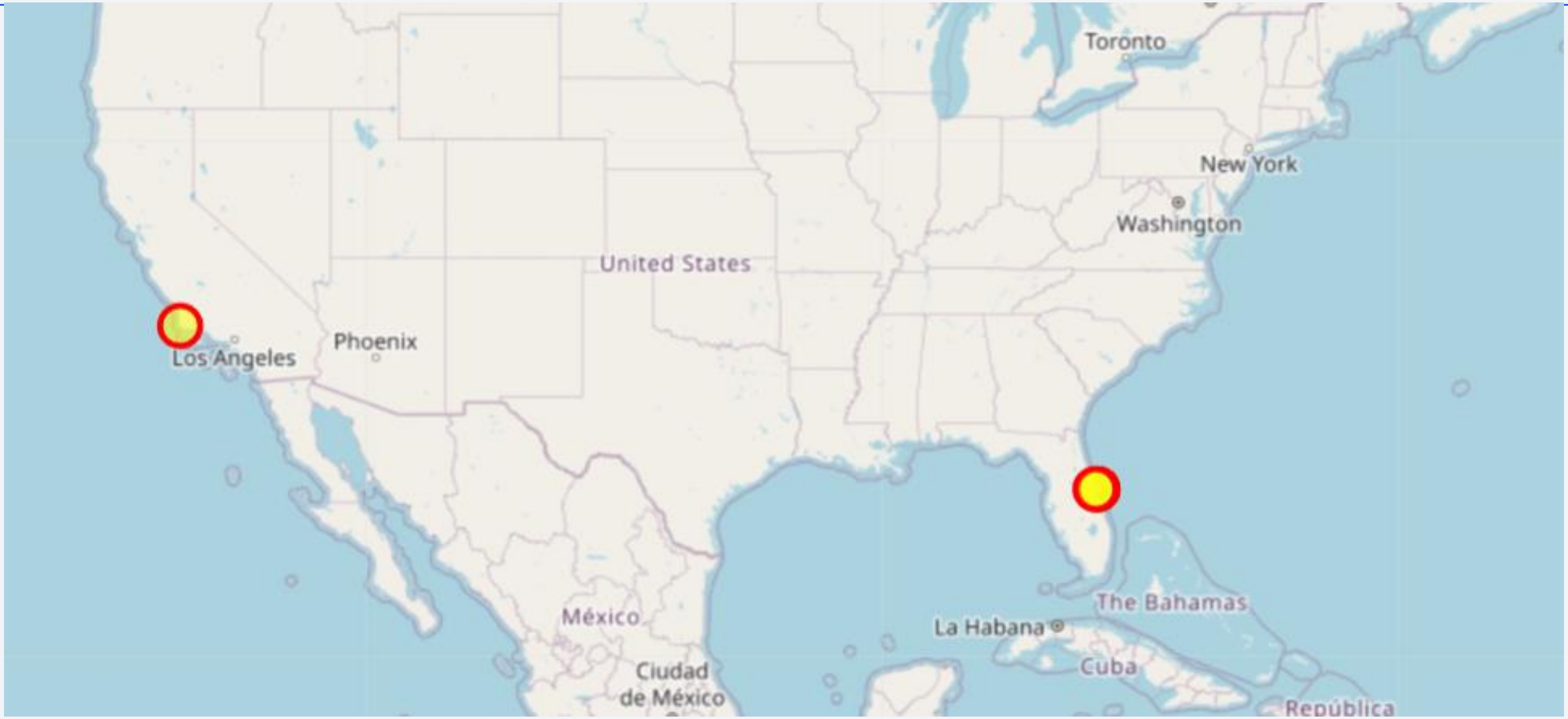
	Date	Landing_Outcome	Count
0	2012-05-22	No attempt	10
1	2016-04-08	Success (drone ship)	5
2	2015-01-10	Failure (drone ship)	5
3	2015-12-22	Success (ground pad)	3
4	2014-04-18	Controlled (ocean)	3
5	2013-09-29	Uncontrolled (ocean)	2
6	2010-06-04	Failure (parachute)	2
7	2015-06-28	Precluded (drone ship)	1

- Landing outcomes ranked for missions between June 4, 2010 and March 20, 2017.

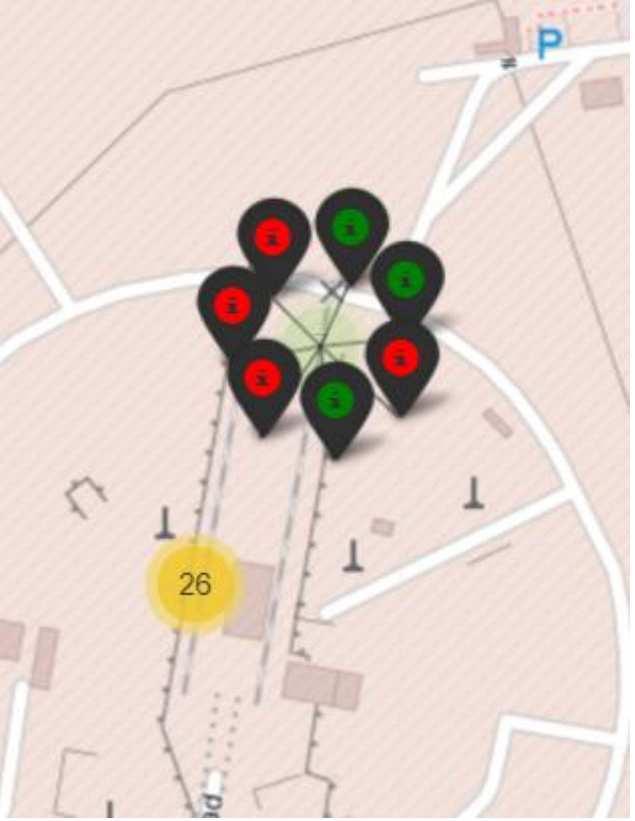
Section 3

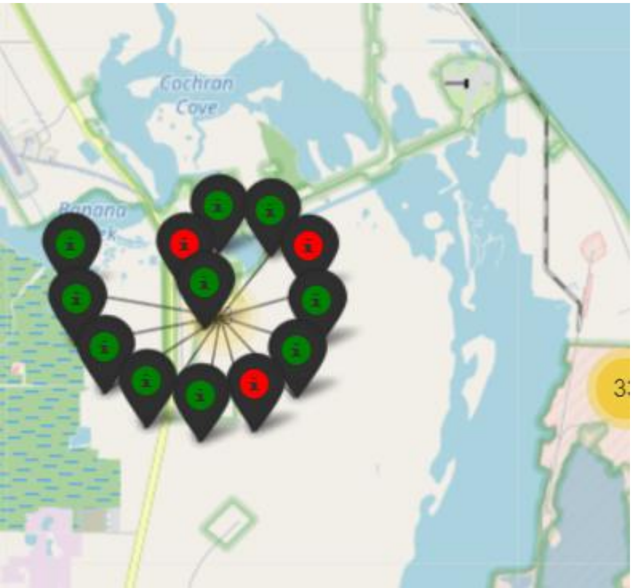
Launch Sites Proximities Analysis

SpaceX Launch Locations



- There are three locations on the east coast of Florida, and one location on the coast near Los Angeles.
- Proximity to coastline may be important.





Launch Outcomes

- Site locations from left to right and top to bottom: CCAFS SLC-40, CCAFS LC-40, KSC LC-39A, VAFB SLC-4E
- Highest success rate is found at site KSC LC-39A

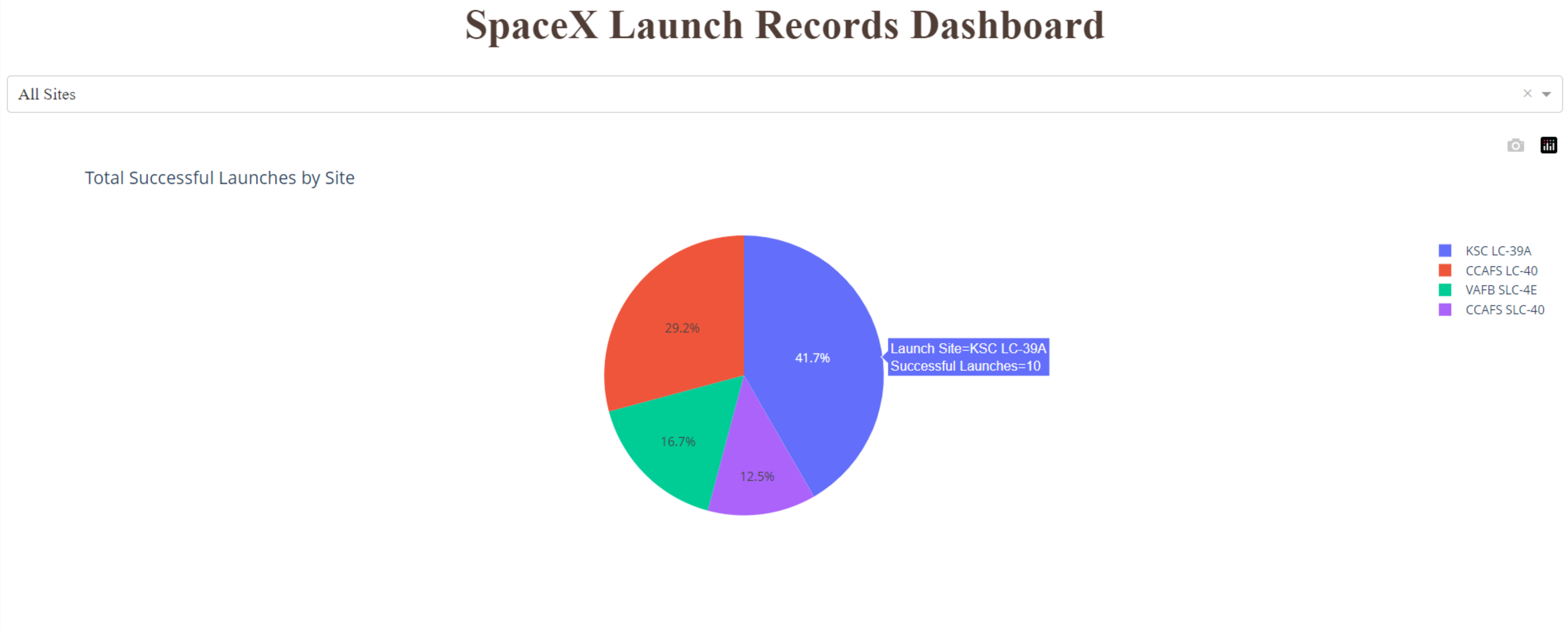
CCAFS SLC-40 Proximities

- CCAFS SLC-40 and CCAFS LC-40 are both within 1 kilometer to a coastline, highway, and railway.

Section 4

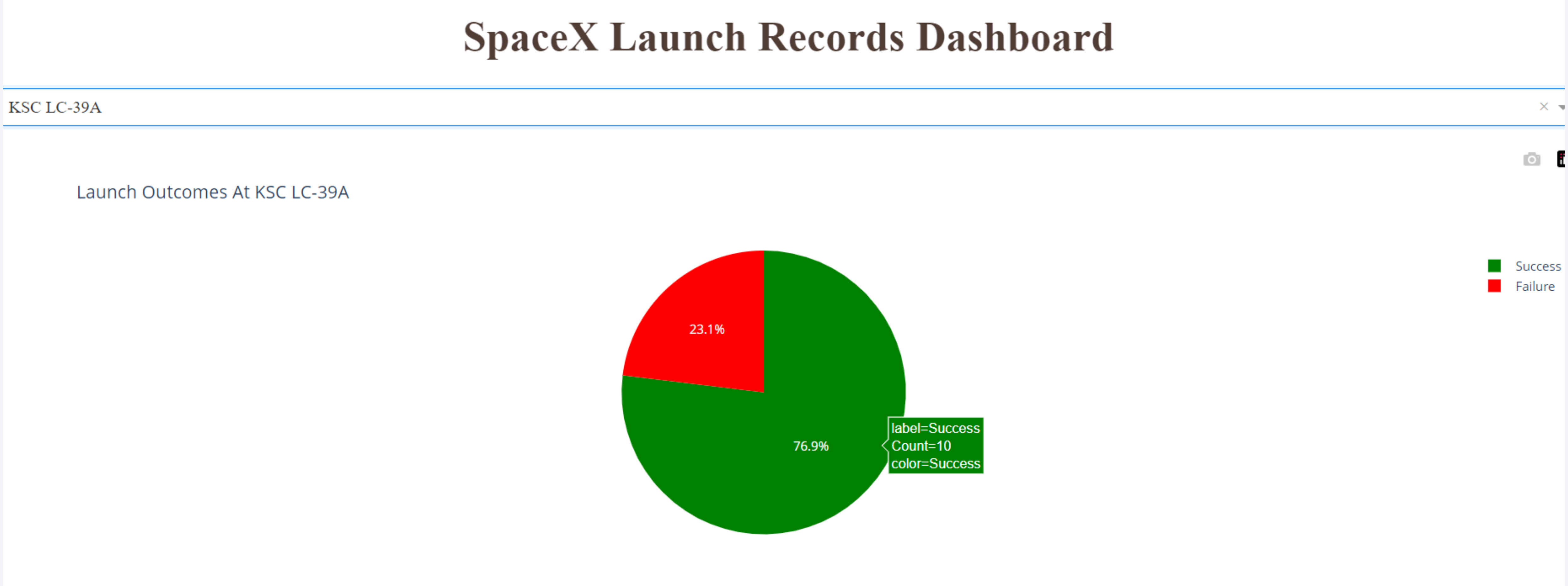
Build a Dashboard with Plotly Dash

Launch Successes by Site



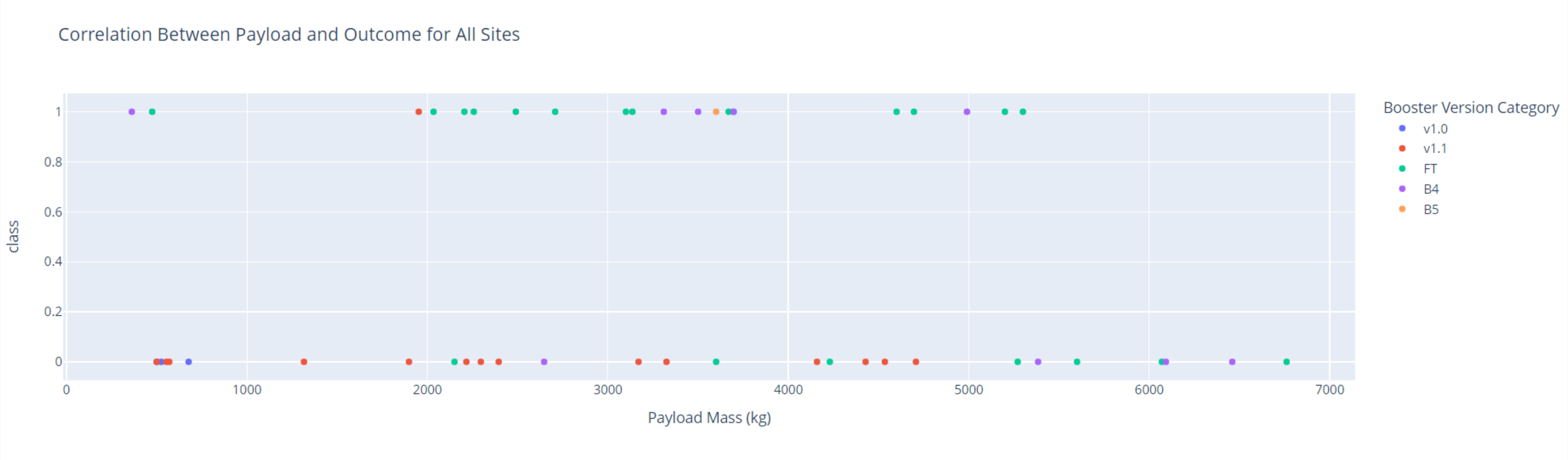
- The majority of successful launches were carried out at KSC LC-39A

Launch Outcomes for KSC LC-39A



- 76.9% of launches ended in success at KSC LC-39A

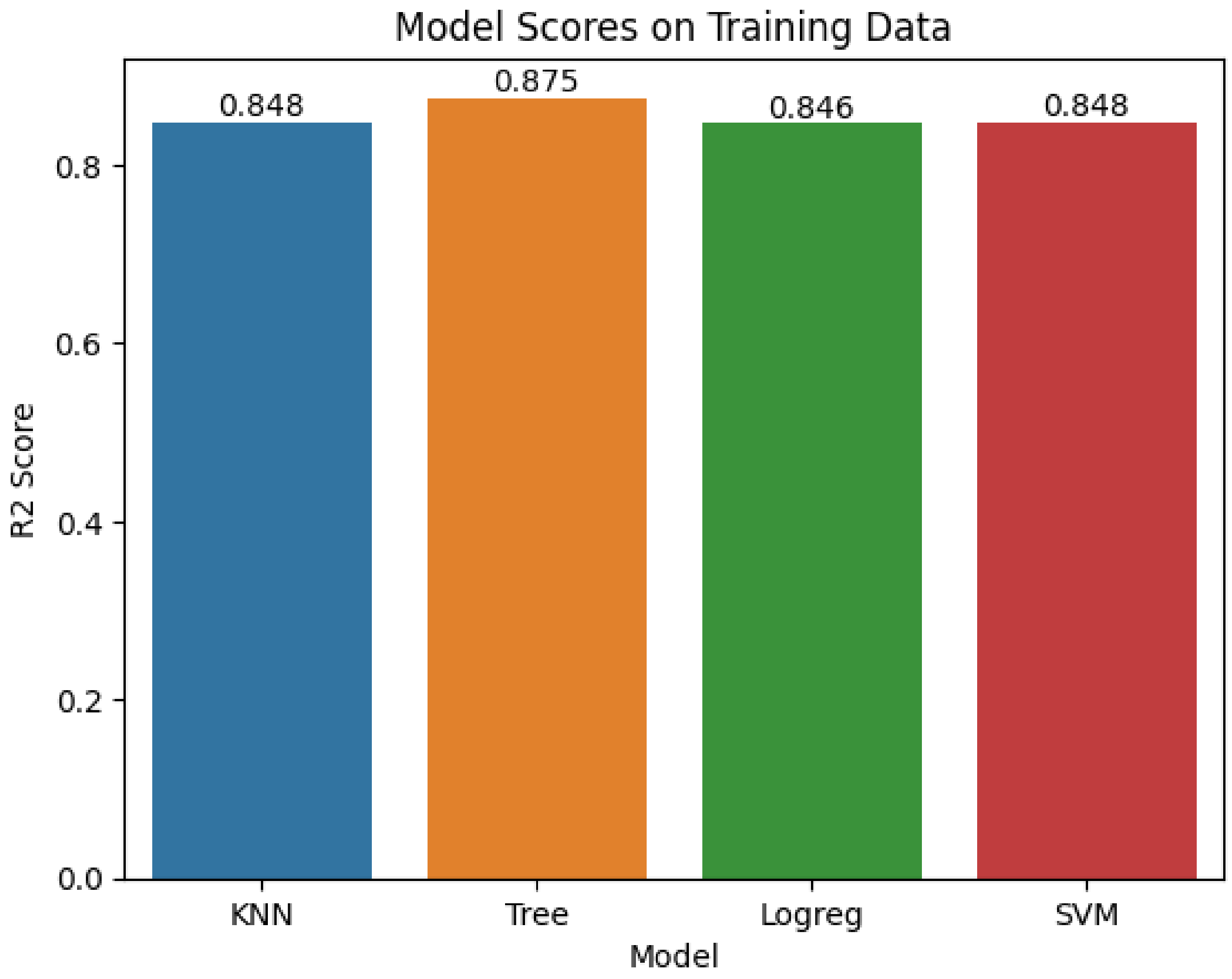
Payload vs Launch Outcome with Booster Categories



- The v1.1 boosters only have one success, while the FT boosters have the highest success rate.
- Both FT and B4 boosters have a higher success rate with payloads less than 5500 kg.

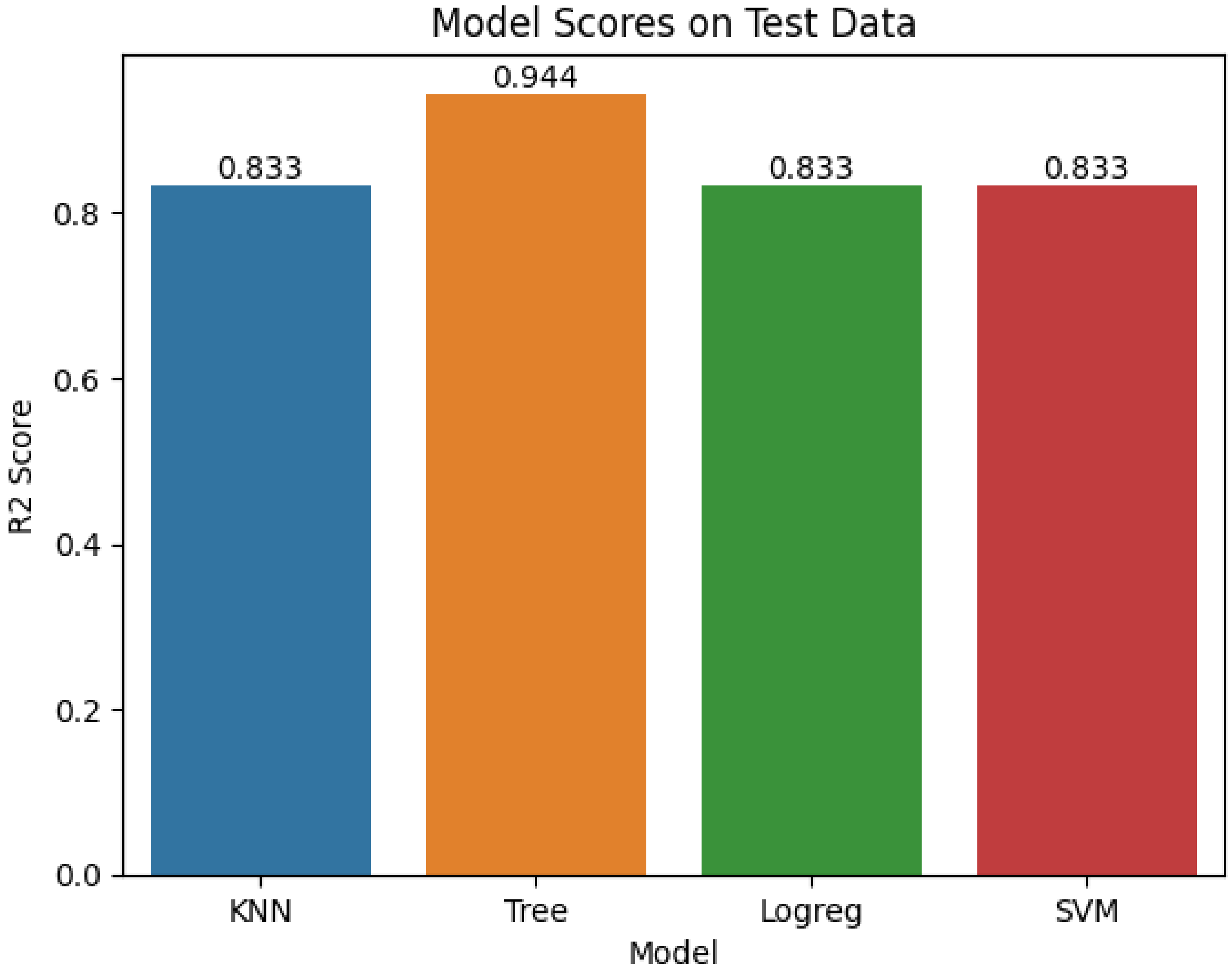
Section 5

Predictive Analysis (Classification)



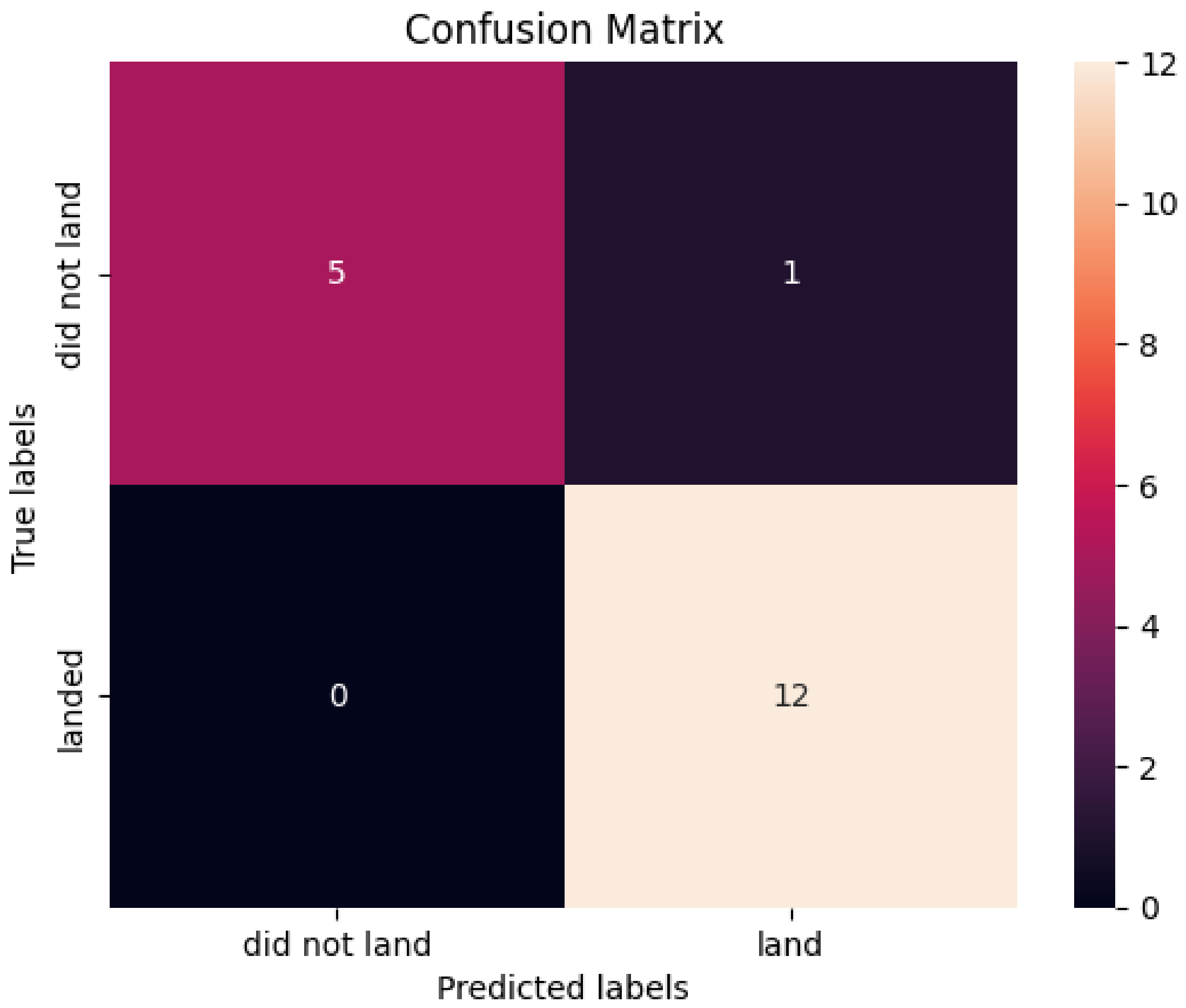
Classification Accuracy on Training Data

- The Decision Tree Classifier resulted in the highest R2 score on the training dataset.



Classification Accuracy on Test Data

- The Decision Tree Classifier remains the highest scoring model on the testing data as well.



Decision Tree Confusion Matrix

- The model distinguishes between the two classes very well, with only one false positive and zero false negatives.

Conclusions

- Successful launch outcomes appear to be influenced by orbit type, payload mass, site location, booster type, and flight number.
- The most effective classification model was found to be a Decision Tree Classifier, with an R2 score of 0.944 on the testing dataset.

Appendix

- URL used for API call: https://api.spacexdata.com/v4/launches/past
- URL used for webscraping: "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922"

Thank you!