IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Roman Bilyy

26th August 2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies
  - Data Collection through API with Requests
  - Data Collection with Web Scraping with BeautifulSoap
  - Data Wrangling with Pandas
  - Exploratory Data Analysis with Sqlite
  - Exploratory Data Analysis with Data Visualization with Seaborn
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction with Sklearn

Summary of all results

- Out of the 90 landings of Falcon 9 rocket between the year of 2010 and 2020, the overall success rate is ≈ 66.67%, showing stable increase since 2013
- Four machine learning models were trained on collected data: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors.
- Decision tree model shows the best prediction performance among the other trained models with training accuracy of 87.32% and testing accuracy of 88.9%.

Introduction

- Project background and context

  Current price of SpaceX Falcon 9 rocket launch is 67 million dollars. Other providers give price like 165 million dollars each. The big difference in prices comes from the ability of SpaceX to reuse their first stage capacities. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch.

  This information can be used if SpaceX competitor wants to bid against SpaceX for a rocket launch.

- Problems you want to find answers

  What factors / operating conditions determine if the rocket will land successfully?

  What model from given test-space will predict Falcon 9 first stage landing success with highest accuracy?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology

  We used RESTful API to get info from SpaceX API service and Web Scraping to collect information from web-site html source.

- Perform data wrangling

  We used Pandas to preprocess data and to do feature engineering to create new column ‘Class’ which will represent landing success. This gave us opportunity to estimate success rate over different sites.

- Perform exploratory data analysis (EDA) using visualization and SQL

  With SQL queries to database we can get optimization of our code by providing DB aggregating and sorting tasks.

Methodology

Executive Summary

- Perform interactive visual analytics using Folium and Plotly Dash

  Interactive visualizations were created using Folium and Plotly Dash libraries.

- Perform predictive analysis using classification models

  Four supervised machine learning techniques were used to build predictive models to predict the outcomes of rocket landing.

Data Collection

- Process of data collection

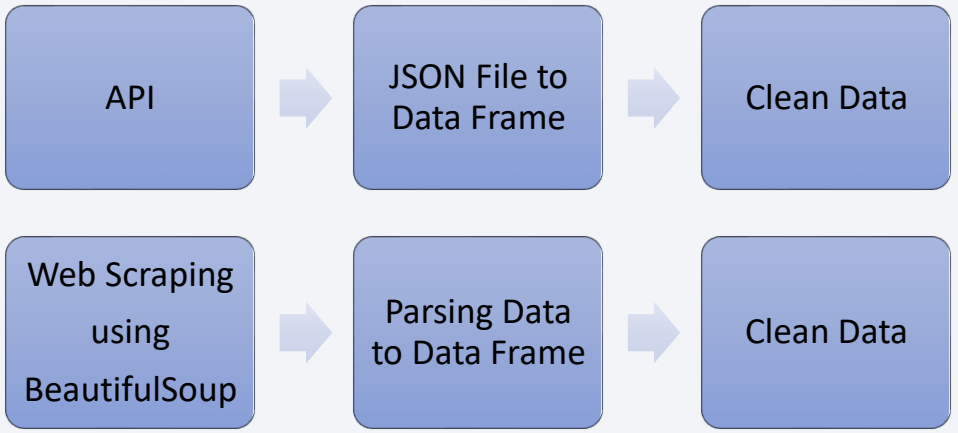
  First, we did get request to the SpaceX API.

  Next, we decoded the response content as a Json using .json() function call and turn it into a pandas dataframe using .json_normalize().

  We then cleaned the data, checked for missing values and fill in missing values where necessary.

  In addition, we performed web scraping from Wikipedia for Falcon 9 launch records with BeautifulSoup.

  The objective was to extract the launch records as HTML table, parse the table and convert it to a pandas dataframe for future analysis.



Data Collection – SpaceX API

1. Call API

```python
spacex_url="https://api.spacexdata.com/v4/launches/past"
```

```python
response = requests.get(spacex_url)
```

2. Check response code

```python
response.status_code
```

```
200
```

3. Normalize and convert to DataFrame

```python
# Use json_normalize meethod to convert the json result into a dataframe
data = pd.json_normalize(response.json())
```

4. Check data

```python
# Get the head of the dataframe
data.head()
```

- The link to the notebook: https://github.com/btvd/SpaceX-Falcon9-First-Stage-Landing-Prediction/blob/master/SpaceX_data-collection-api.ipynb

Data Collection - Scraping

1. Perform an HTTP GET method to request the Falcon9 Launch HTML page, as an HTTP response.

```python
# use requests.get() method with the provided static_url
# assign the response to a object
response = requests.get(static_url)
```

2. Create a `BeautifulSoup` object from the HTML `response`

```python
# Use BeautifulSoup() to create a BeautifulSoup object from a response text content
soup = BeautifulSoup(response.content)
```

3. Extract all column/variable names from the HTML table header

```python
# Use the find_all function in the BeautifulSoup object, with element type `table`
# Assign the result to a list called `html_tables`
html_tables = soap.find_all('table')
```

```python
# Let's print the third table and check its content
first_launch_table = html_tables[2]
print(first_launch_table)
```

5. Parse data

```python
column_names = []

# Apply find_all() function with `th` element on first_launch_table
# Iterate each th element and apply the provided extract_column_from_header() to get a column name
# Append the Non-empty column name (`if name is not None and len(name) > 0`) into a list called column_names
column_names = []
for elem in first_launch_table.find_all('th'):
    name = extract_column_from_header(elem)
    if name is not None and len(name) > 0:
        column_names.append(name)
```

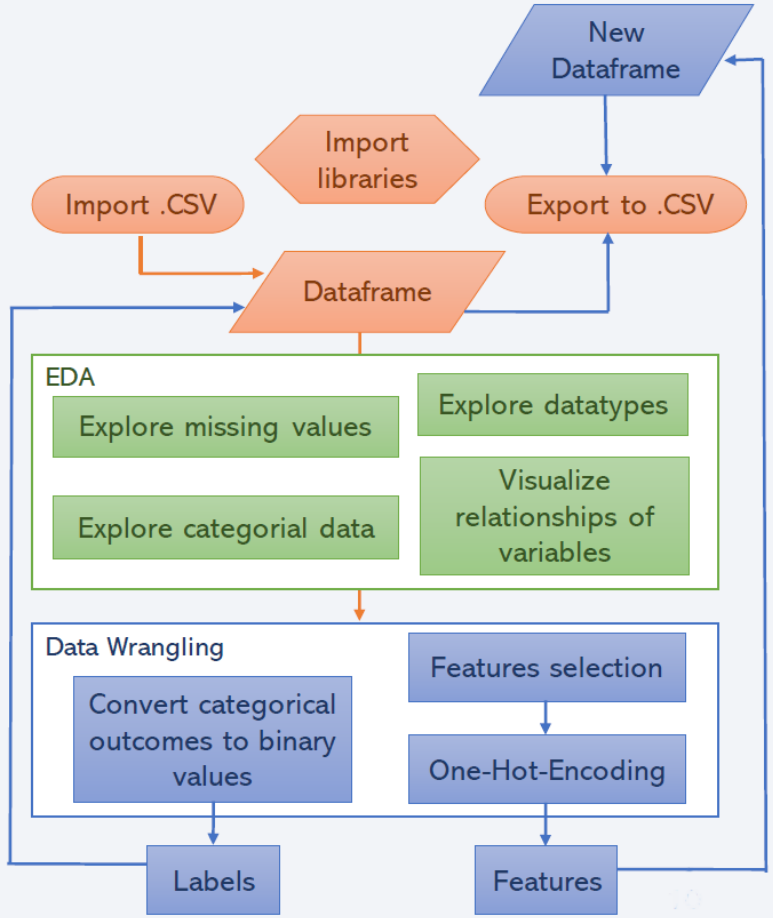
- The link to the notebook: https://github.com/btvd/SpaceX-Falcon9-First-Stage-Landing-Prediction/blob/master/SpaceX_webscraping.ipynb

10

Data Wrangling

1. Exploratory Data Analysis (EDA) was performed to find some patterns in the data.
2. Than we did feature engineering – converted categorical variables into binary datatype and encoded needed variables.
3. We exported the results to csv.

- The link to the notebook: https://github.com/btvd/SpaceX-Falcon9-First-Stage-Landing-Prediction/blob/master/SpaceX_data_wrangling.ipynb



EDA with Data Visualization

Scatterplot charts from package Seaborn were plotted to see what relationships are there between variables and how they will affect on landings outcome.

Bar chart was plotted to compare success rate of different orbit types.

Line chart was plotted to see trends in the yearly success rates.

Scatterplot	Bar chart	Line chart
• FlightNumber vs. PayloadMass • FlightNumber vs LaunchSite • Payload and Launch Site • FlightNumber and Orbit type • Payload and Orbit type	• Success rate of each orbit type	• Landing success yearly trend

- The link to the notebook: https://github.com/btvd/SpaceX-Falcon9-First-Stage-Landing-Prediction/blob/master/SpaceX_eda-dataviz.ipynb

EDA with SQL

The following SQL queries were performed

- Display the names of the unique launch sites in the space mission.
- Display 5 records where launch sites begin with the string 'CCA'.
- Display the total payload mass carried by boosters launched by NASA (CRS).
- Display average payload mass carried by booster version F9 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 payloadmass greater than 4000 but less than 6000.
- List the total number of successful and failure mission outcomes.
- List the names of the booster_versions which have carried the maximum payload mass. Use a subquery.
- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for year 2015.
- Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

- The link to the notebook: https://github.com/btvd/SpaceX-Falcon9-First-Stage-Landing-Prediction/blob/master/SpaceX_eda_sqllite.ipynb

Build an Interactive Map with Folium

Map components such as circles, markers, and polylines were built and added to the folium map.

MarkerClusters were also used to group together many markers with the same coordinate.

A MousePosition was also introduced to acquire the coordinates for a mouse over a map point.

This allows us to understand why launch sites may be located where they are, visualize successful landings relative to location and track distances between sites and key locations: Railway, Highway, Coast and City.

- The link to the notebook: https://github.com/btvd/SpaceX-Falcon9-First-Stage-Landing-Prediction/blob/master/SpaceX_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

Dashboard includes a pie chart and a scatter plot.

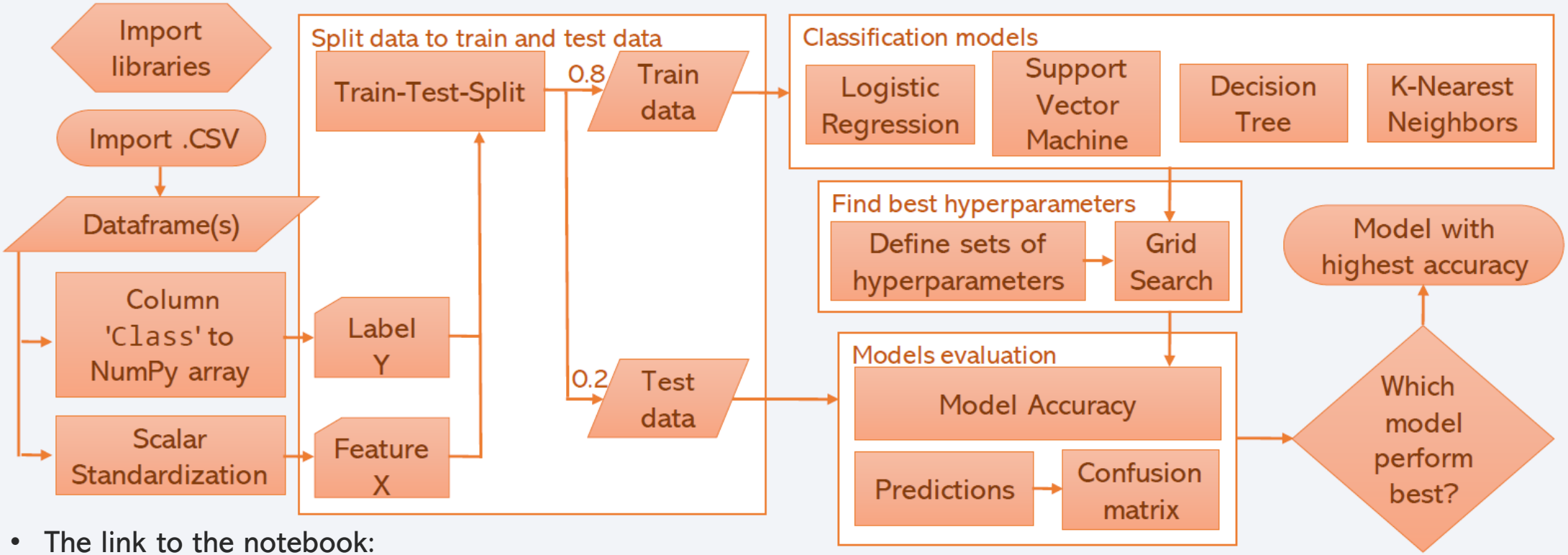
Pie chart can be selected to show the distribution of successful landings across all launch sites and can be selected to show individual launch site success rates.

Scatter plot takes two inputs: All sites or individual sites and payload mass on a slider between 0 and 10000kg. The scatter plot can help us see how success varies across launch sites, payloadmass, and booster version category.

- The link to the notebook: https://github.com/btvd/SpaceX-Falcon9-First-Stage-Landing-Prediction/blob/master/spacex_dash_app.py

Predictive Analysis (Classification)

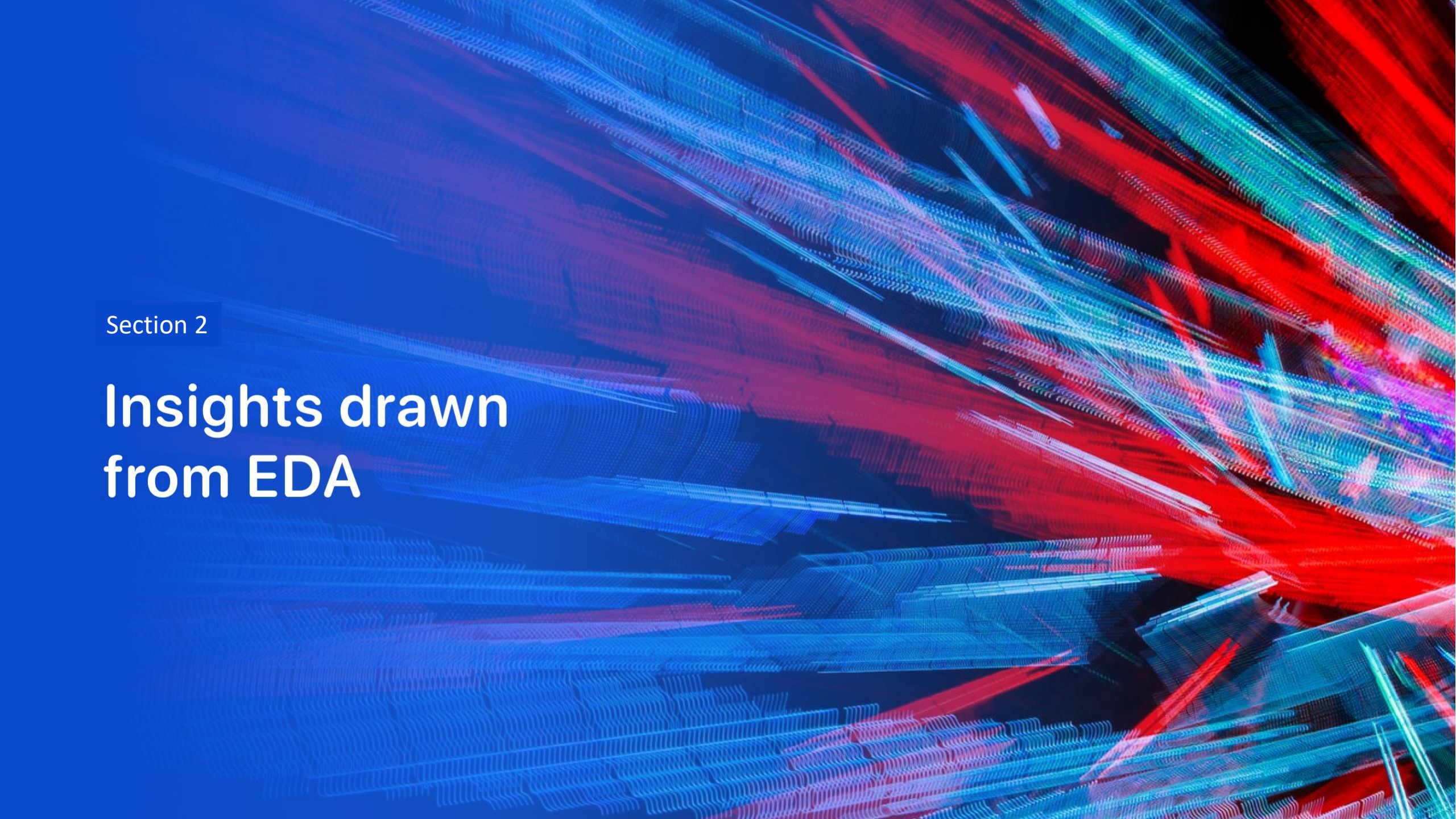
The best performing classification model was built, evaluated, improved, and found as flowchart below:



- The link to the notebook: https://github.com/btvd/SpaceX-Falcon9-First-Stage-Landing-Prediction/blob/master/SpaceX_machine_learning_prediction.ipynb

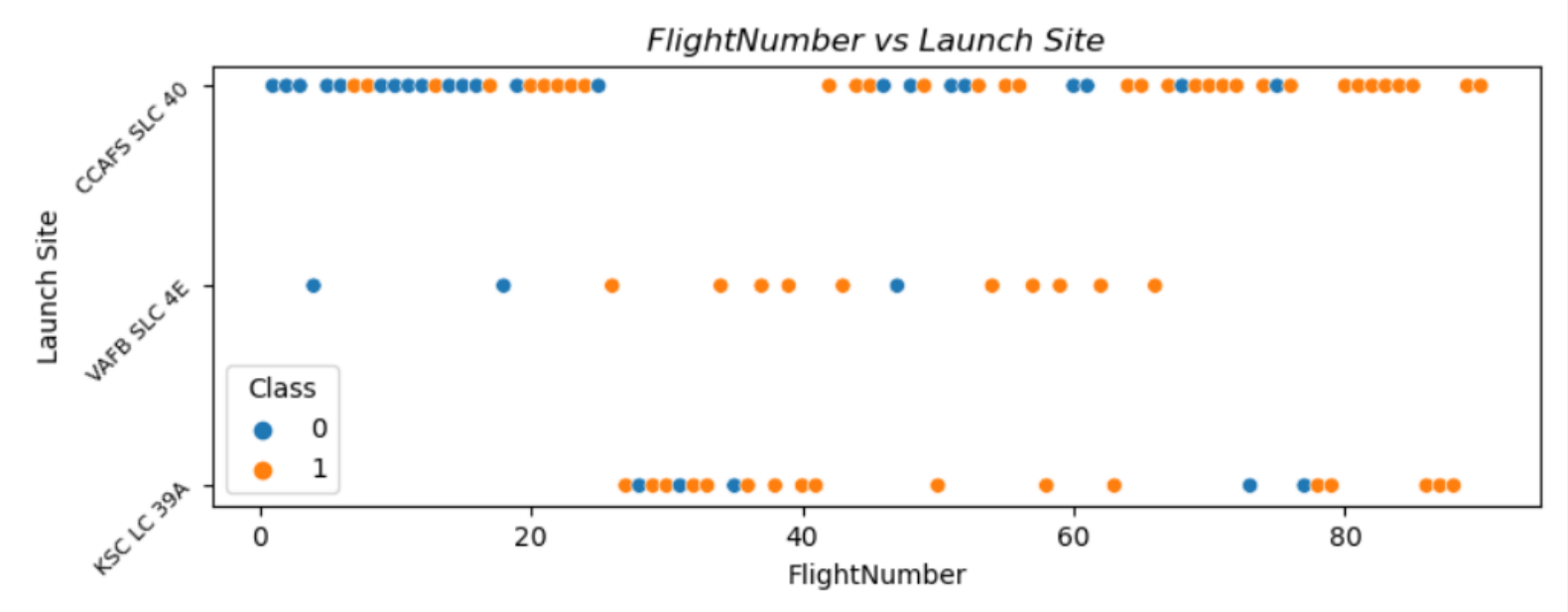
Results

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


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

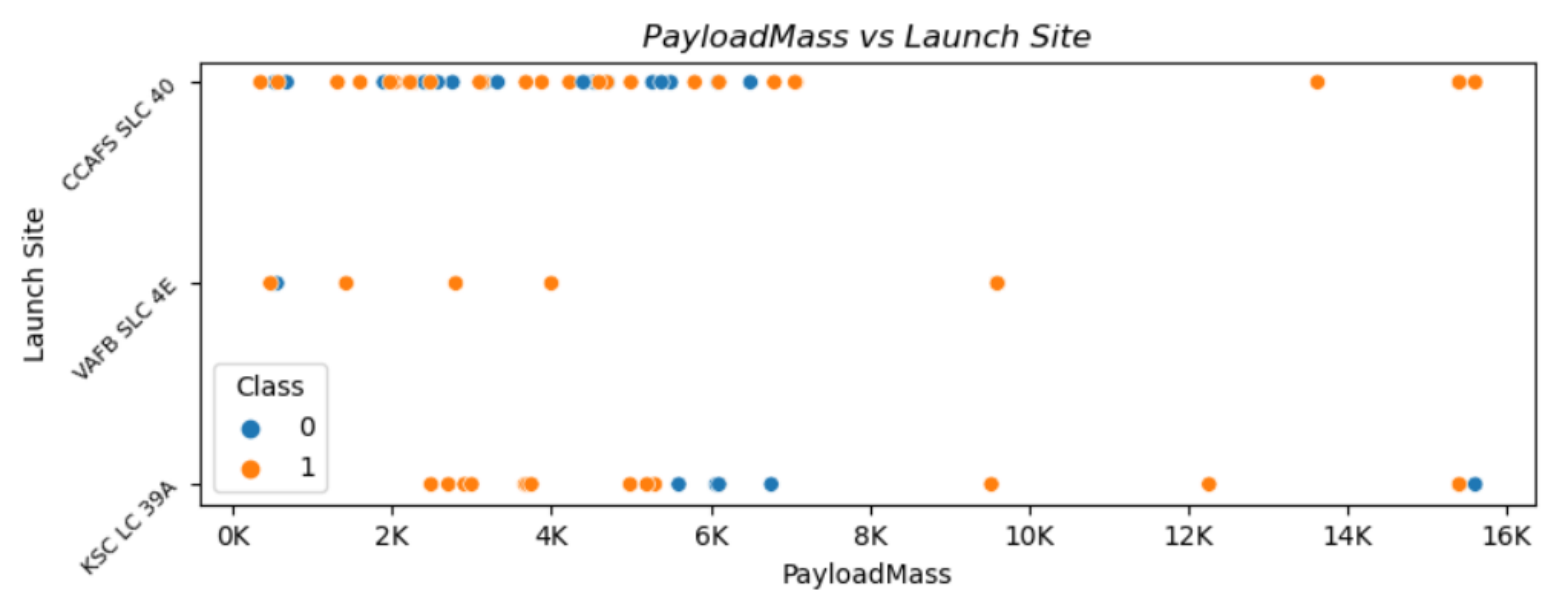
We can see on the plot below that distribution of number of launches between different launch sites are not equal, but we see increase in success rate with the number of flights rising on each site.



Payload vs. Launch Site

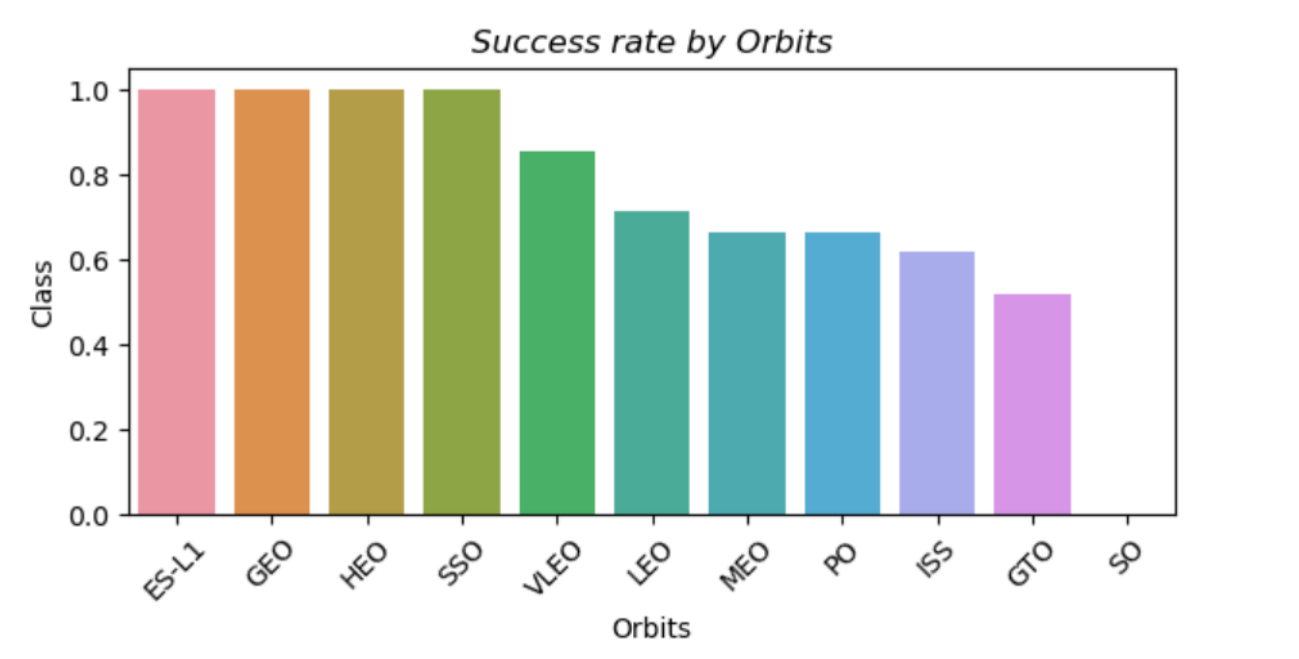
Looking at the figure below we can see, that:

- There are no rockets launched for heavy payload mass (greater than 10,000) at the VAFB SLC 4E launch site
- CCAFS SLC 40 has the least success rate of landing among the other sites.

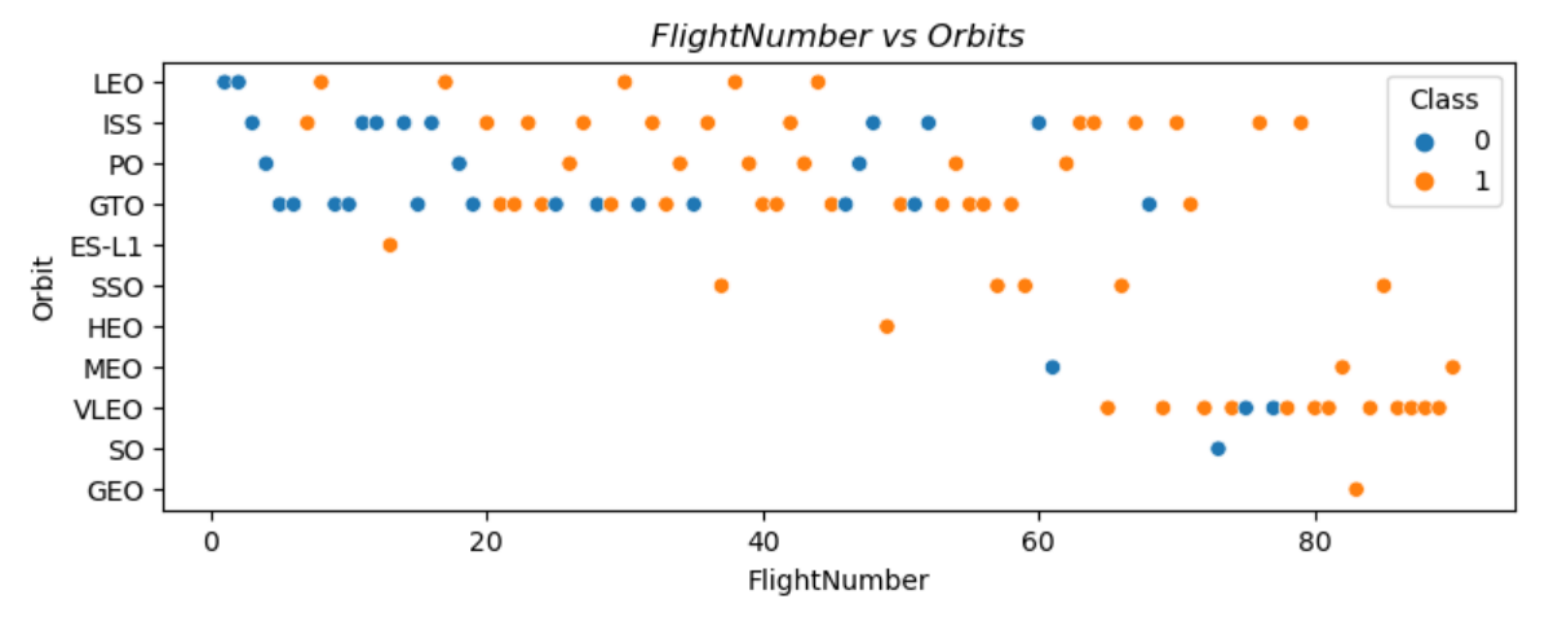


Success Rate vs. Orbit Type

Data shows that orbits ES-L1, GEO, HEO, and SSO all has 100% success rates of landing, whereas orbit SO has no successfully landed rockets.

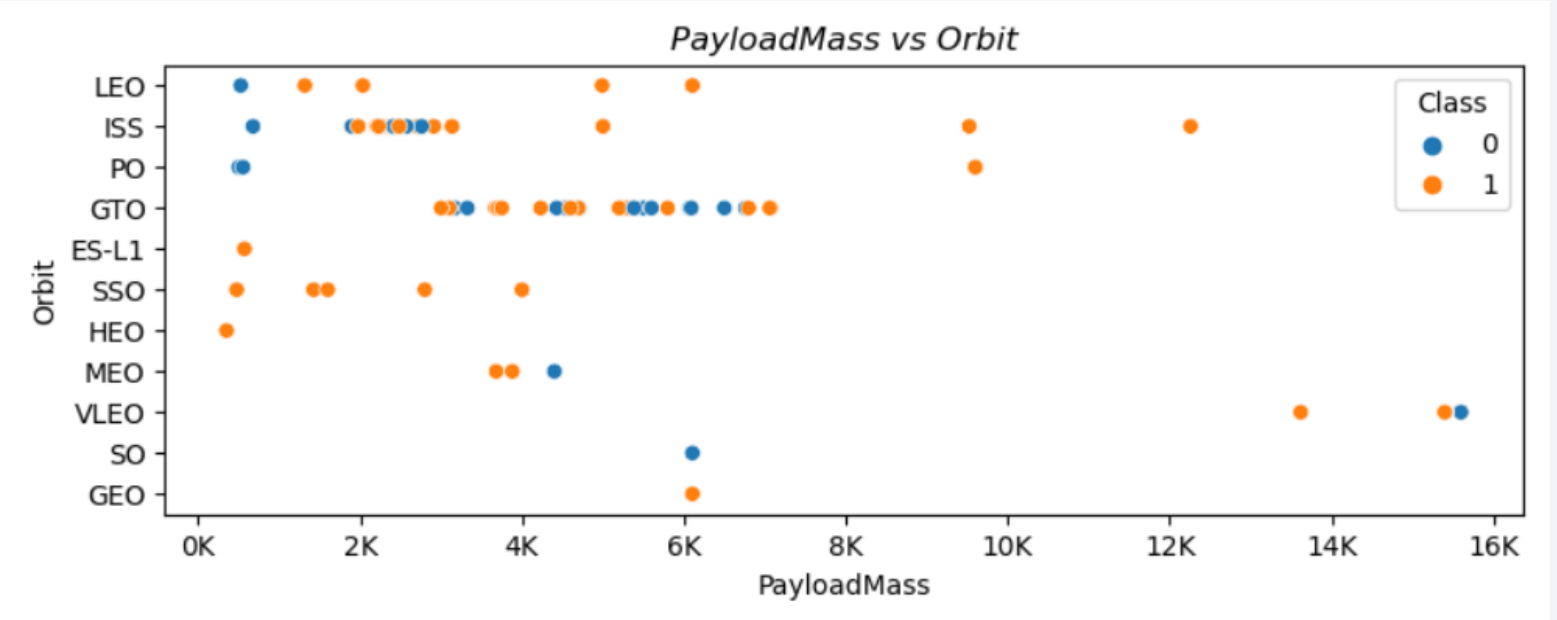


Flight Number vs. Orbit Type



On scatterplot above we can observe that in the LEO , VLEO and MEO orbit, landing success positively correlates to the number of flights. Whereas, in the GTO orbit, there is no relationship between success and the number of flights.

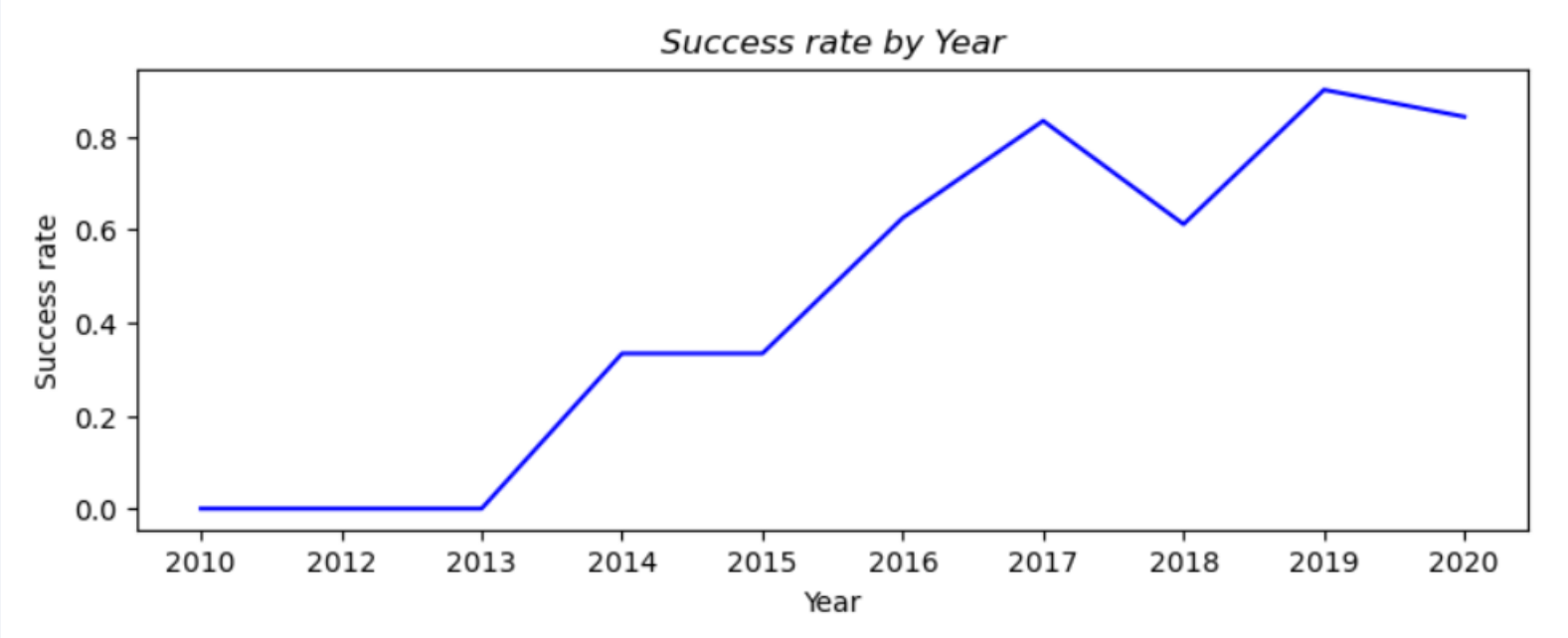
Payload vs. Orbit Type



Looking at plot above we can conclude that:

- The more the payload mass - the better the success rate of landing for the most of the orbits.
- SSO, HEO, ES-L1, GEO orbits has the highest landing success rate.
- GTO shows mixed data

Launch Success Yearly Trend



We can observe on the plot above that since 2013 the success rate of rocket landing started steady increase until 2017. After the drop in rates in 2018, in 2019 rates of landing success reached local maximum.

All Launch Site Names

We used the keyword DISTINCT to show only unique launch sites.

```
1 %sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL;
```

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

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

Launch Site Names Begin with 'CCA'

```
1 %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
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

Combination of WHERE, LIKE and LIMIT is used in this query. We set at what column look at and with what it should start. Than we set limit to output.

Total Payload Mass

```
1 %sql SELECT SUM(payload_mass__kg_) FROM SPACEXTBL WHERE customer = 'NASA (CRS)';
```

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

SUM(payload_mass__kg_)
45596

We figured out that total payload carried by boosters from NASA is 45596. Combination of SUM and WHERE is used in query.

Average Payload Mass by F9 v1.1

```
1 %sql SELECT AVG(payload_mass__kg_) AS F9_avr_payload FROM SPACEXTBL WHERE booster_version = 'F9 v1.1';
```

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

F9_avr_payload
2928.4

We calculated that the average payload mass carried by booster version F9 v1.1 is 2928.4. Combination of AVG and WHERE is used in query.

First Successful Ground Landing Date

```
1 %sql SELECT MIN(DATE) as Earliest_date FROM SPACEXTBL WHERE `Landing _Outcome` = 'Success (ground pad)';
```

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

Earliest_date
01-05-2017

The first successful landing outcome on ground pad happened on **01-05-2017**.

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_ BETWEEN 4000 and 6000;
```

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

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

We got 4 boosters that satisfied the query: F9FTB1022, F9FTB1026, F9FTB1021.2, and F9FTB1031.2.

Total Number of Successful and Failure Mission Outcomes

```
1 %sql SELECT mission_outcome, COUNT(mission_outcome) AS "Total number" FROM SPACEXTBL GROUP BY mission_outcome ;
```

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

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

There is only 1 failure mission outcome and 100 successful missions

Boosters Carried Maximum Payload

```
1 %sql SELECT 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

The result shows a list of 12 booster versions that carried the maximum payload mass of 15600 kg.

2015 Launch Records

```
%%sql SELECT substr(Date, 4, 2) AS "Month",`Landing _Outcome`,BOOSTER_VERSION,LAUNCH_SITE
FROM SPACEXTBL WHERE substr(Date,7,4)='2015' AND `Landing _Outcome` = 'Failure (drone ship)';
```

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

Month	Landing _Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

In 2015 booster version F9 v1.1 B1012 and B1015 have failed landing in drone ship both on launch site CCAFS LC 40.

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

Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

```
%%sql SELECT `Landing _Outcome`,COUNT(`Landing _Outcome`) AS "Success Numb"
FROM SPACEXTBL WHERE `Landing _Outcome` LIKE 'Success%' AND 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;
```

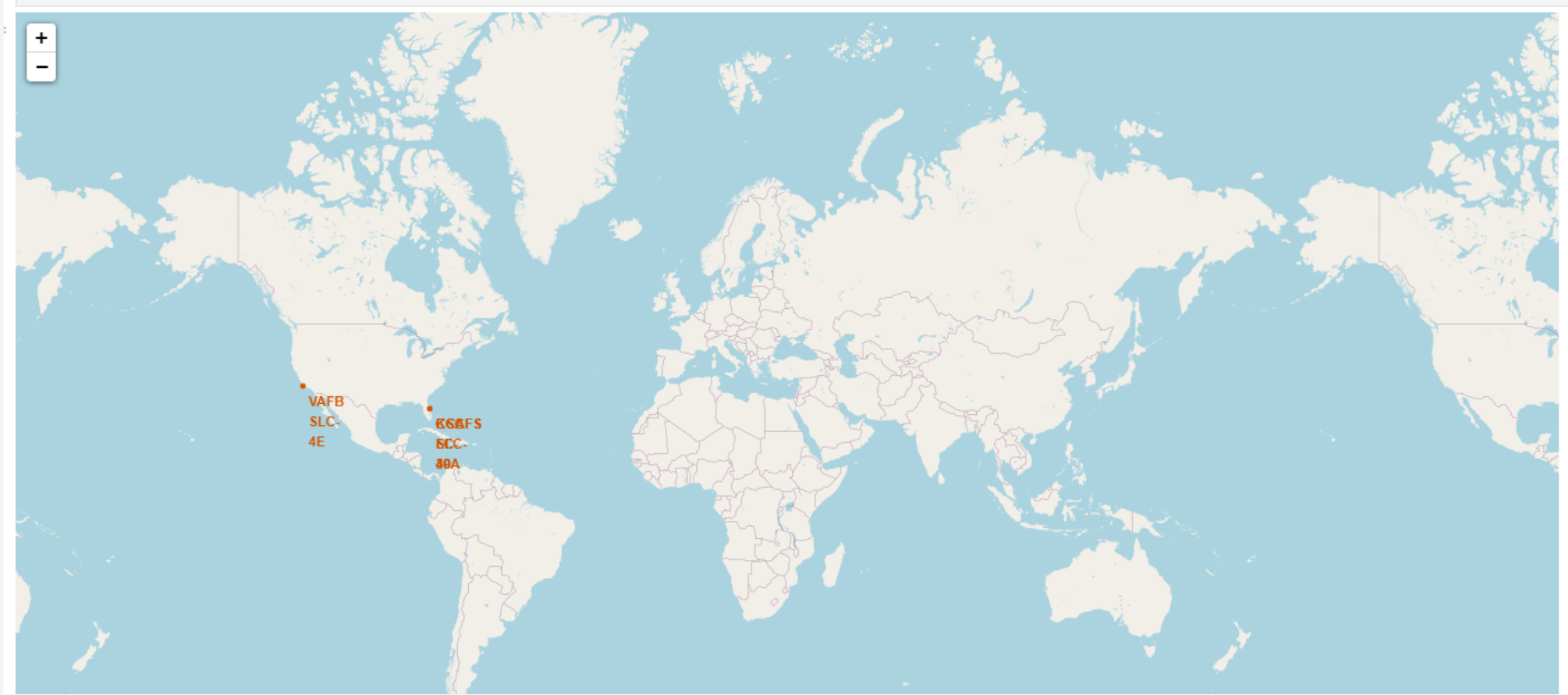
```
 * sqlite:///my_data1.db
Done.
```

Landing _Outcome	Success Numb
Success (drone ship)	5
Success (ground pad)	3

As we see, there were more successful drone ship missions than ground pad missions during given period.

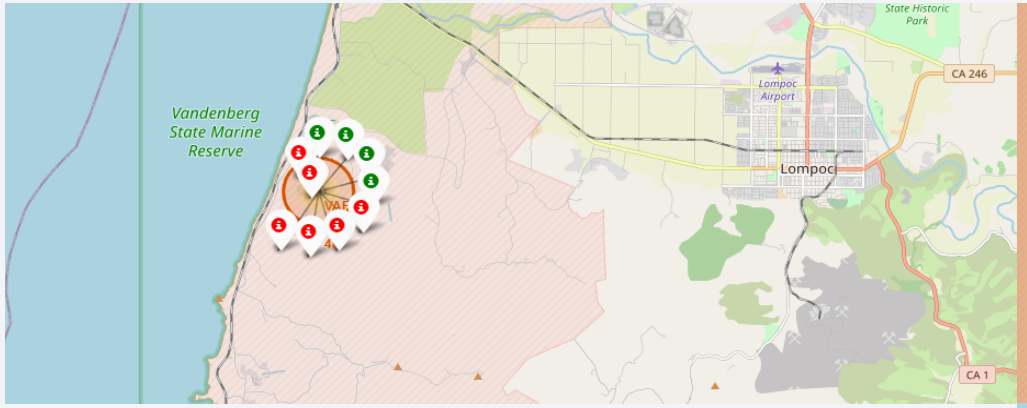

Section 3
Launch Sites
Proximities Analysis

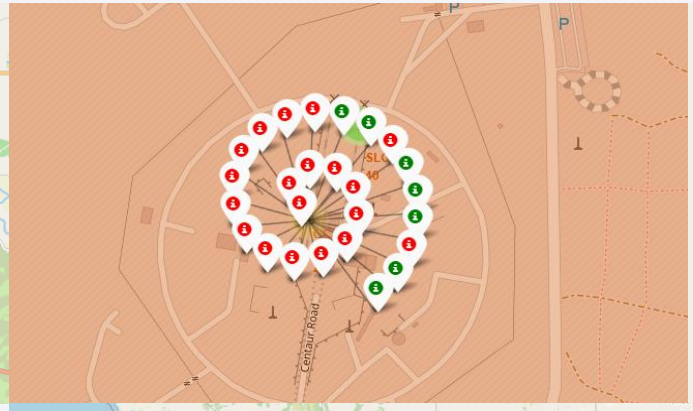
Locations of All Launch Sites

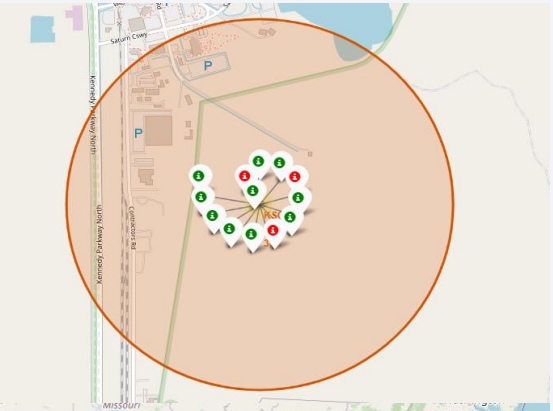


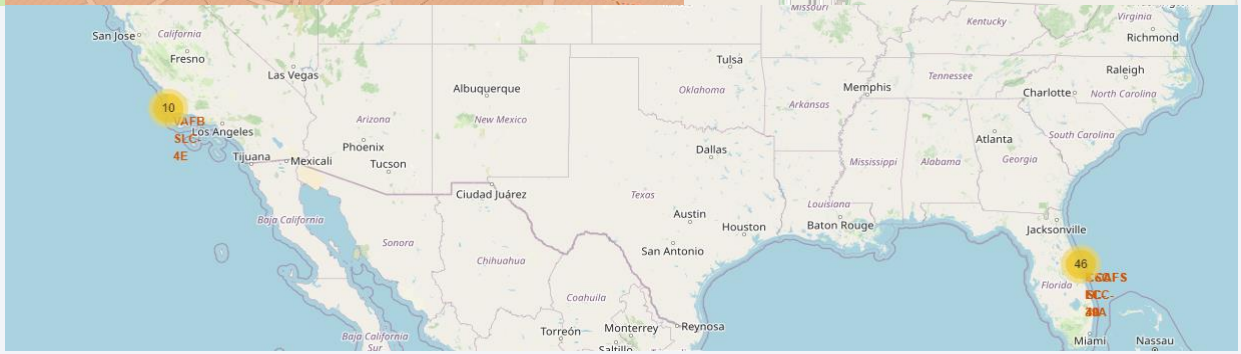
As shown on map, SpaceX launch sites are located in the United States of America at Florida and California coasts.

Success or Failed Launches for each Site







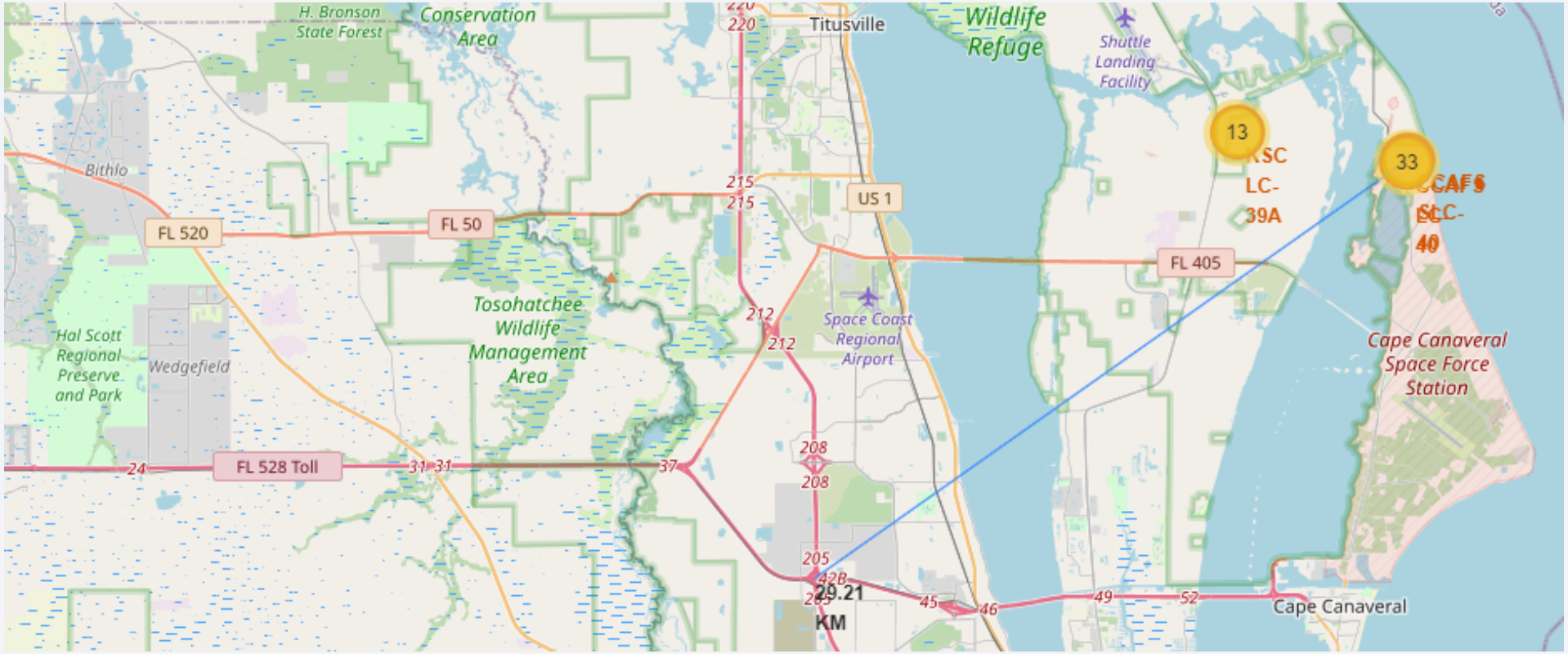


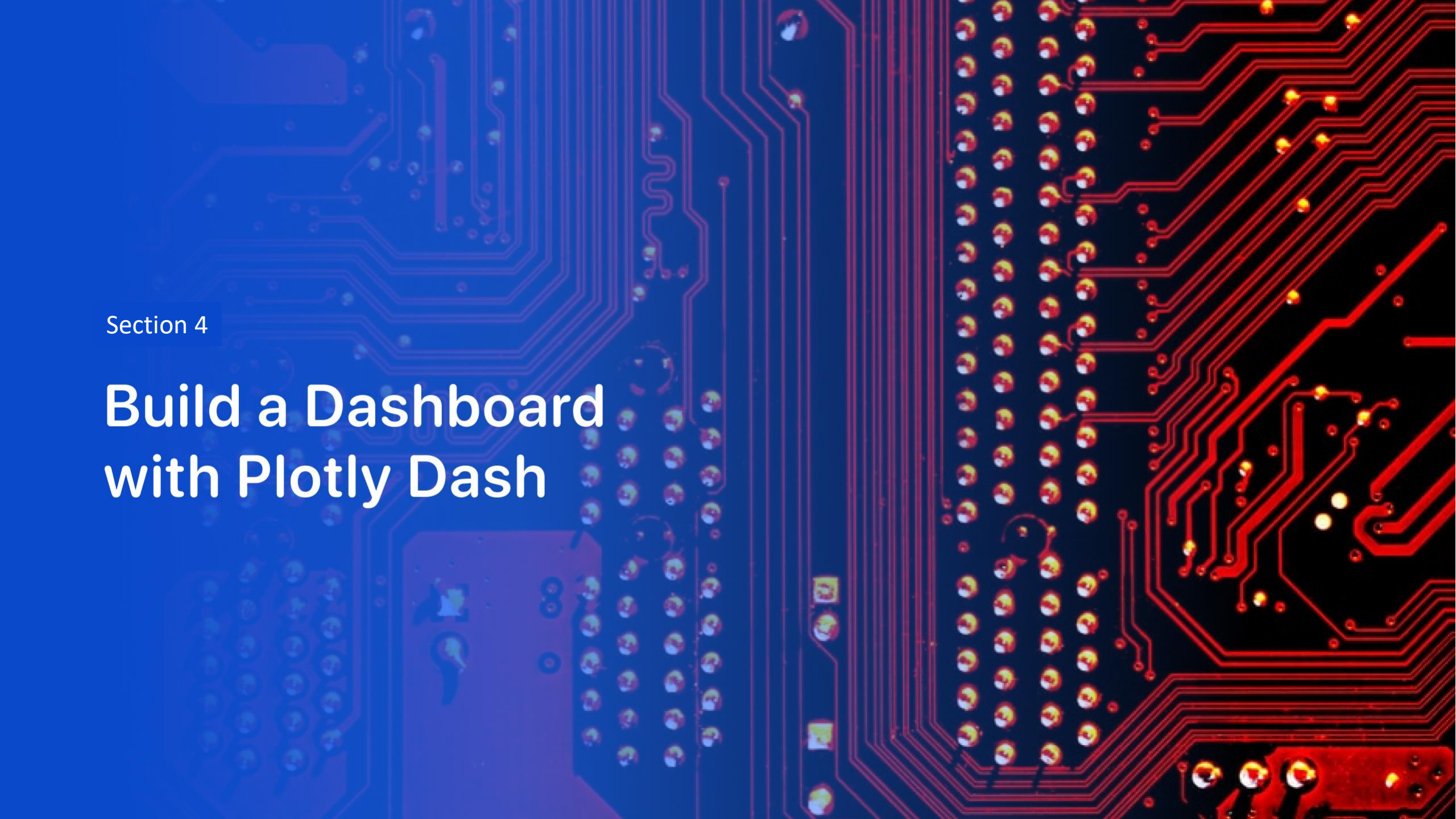
Green markers on plots are meant as successful launches and red markers are failed ones.

Comparing ratio between markers of different colors for different launch sites we can observe that KSC LC-39A has relatively high success rate.

Launch Site distance to landmarks

Here we see distance to nearest Highway.

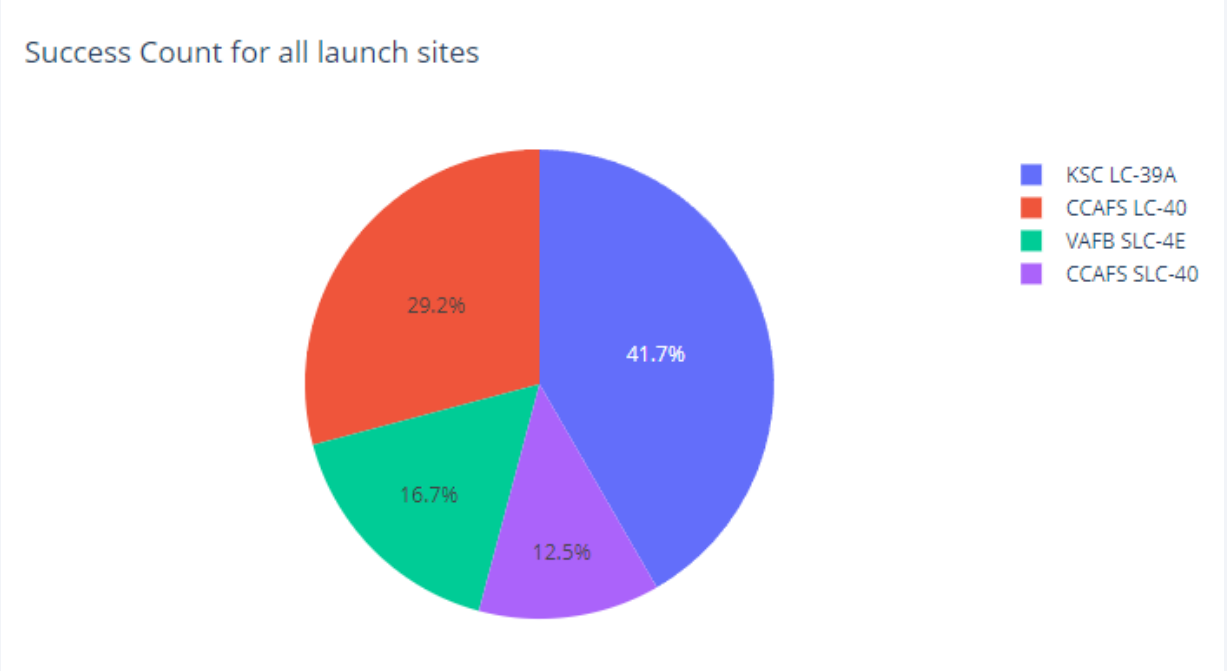



Section 4
Build a Dashboard with Plotly Dash

Launch Success Counts for All Sites

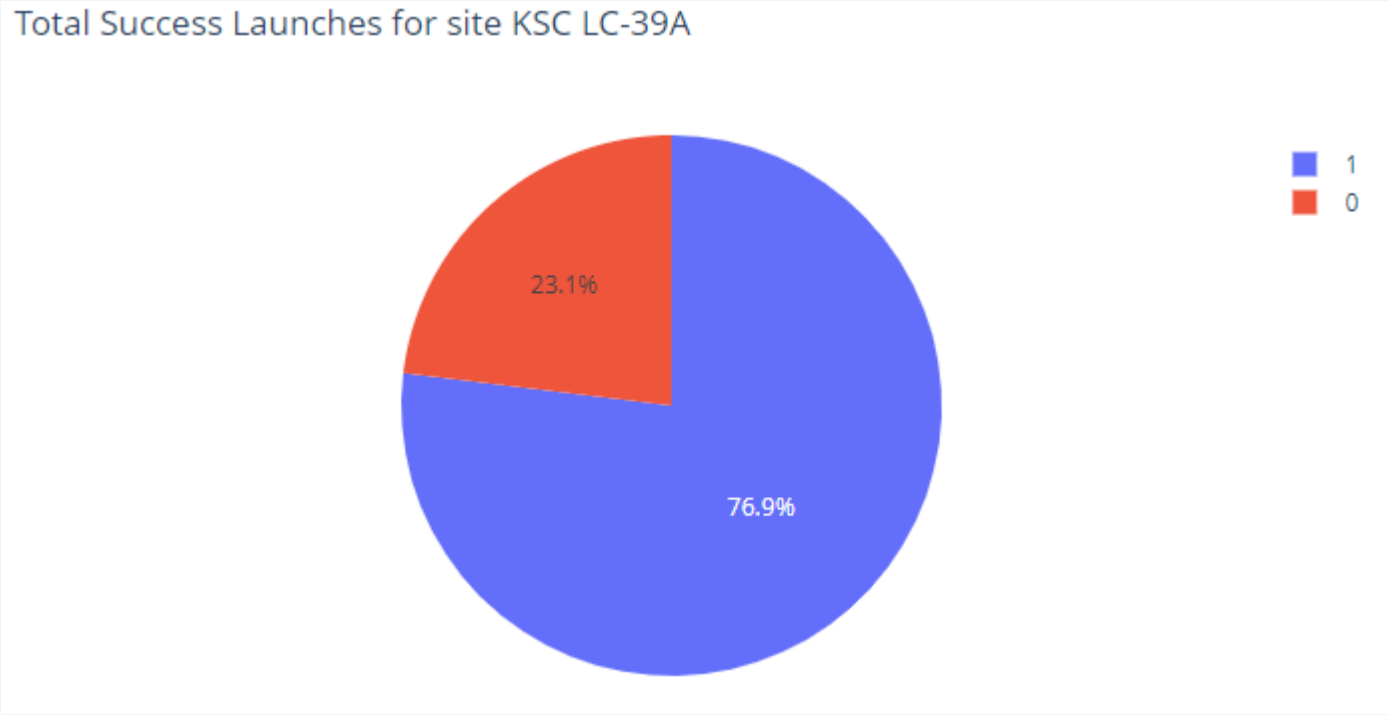
Pie chart on the right displays the total counts of success for all sites.

Launches from KSC LC-39A site have the highest success counts with 41.7% from total.



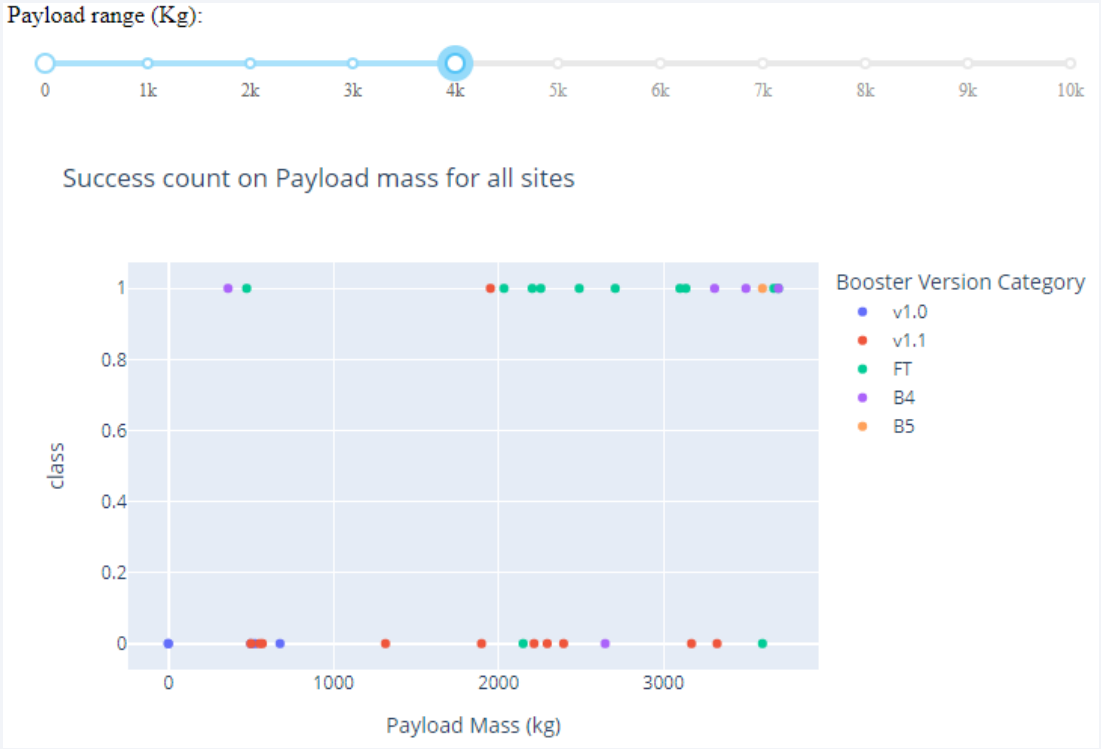
Launch Success Ratio for Launch Site KSC LC-39A

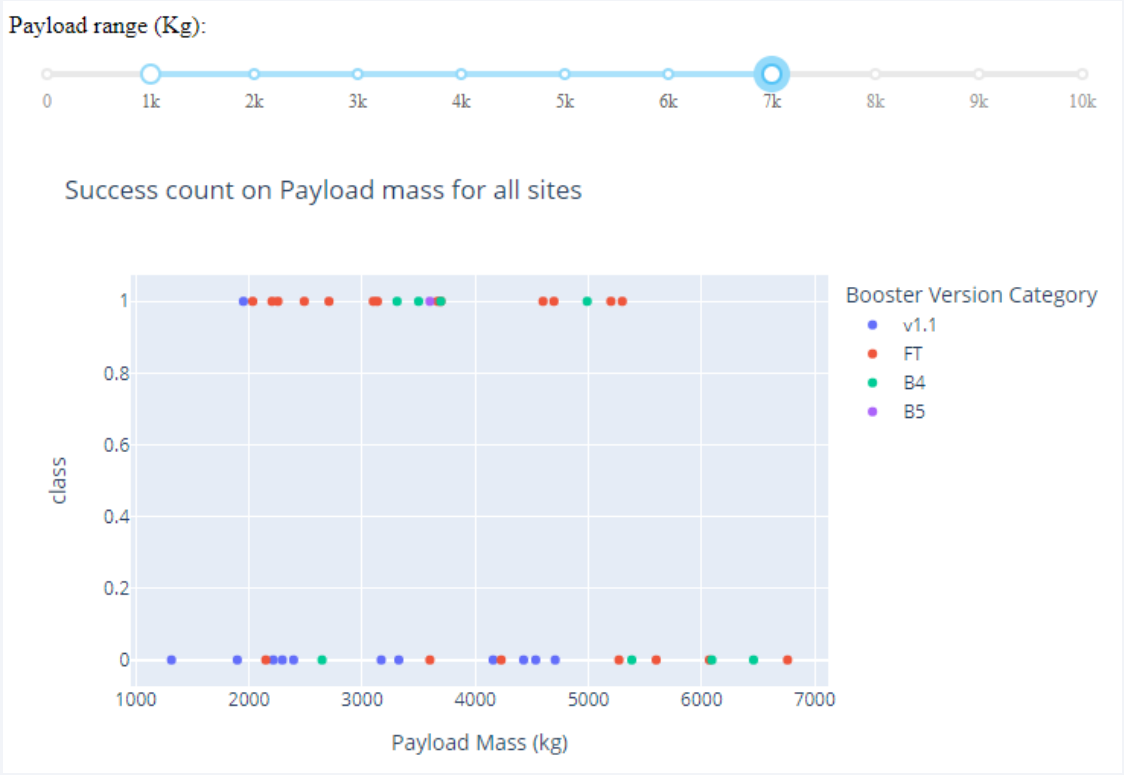
KSC LC-39A site has the highest success ratio of 76.9%.



Payload VS Launch Outcome for All Sites

The scatterplots illustrate the relationships between payload and launch outcome for different booster versions. Overall, greater payload mass will decrease the success rate.



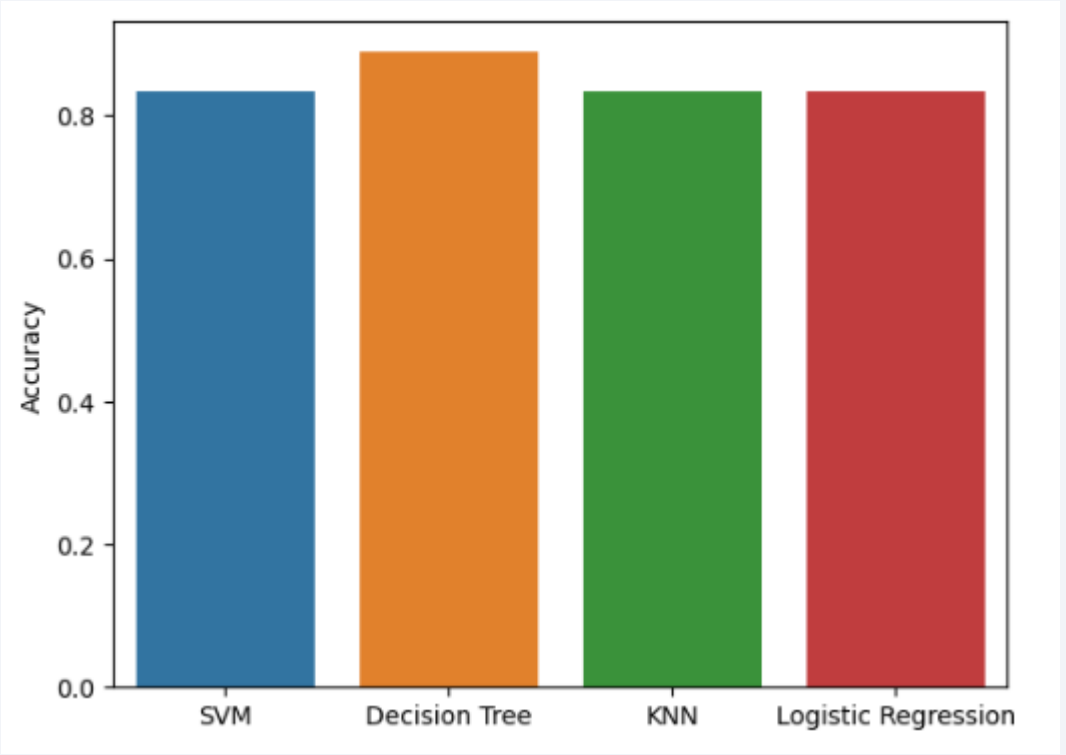


Section 5

Predictive Analysis (Classification)

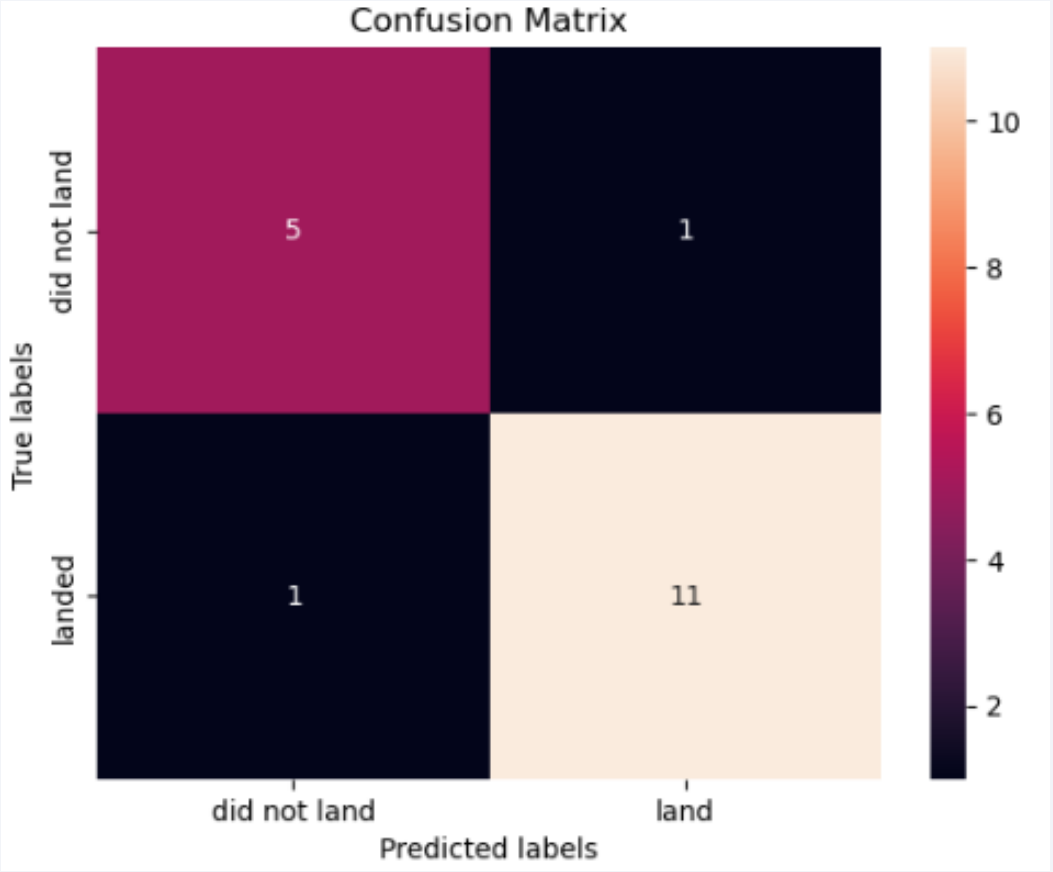
Classification Accuracy

The decision tree classifier is the model with the highest classification accuracy.



Confusion Matrix

As we see from confusion matrix of DecisionTree Classifier, trained model still has Type I and Type II errors, but their amount is low.



Conclusions

- Launches above 7,000kg are less risky.
- The best launch site is KSC LC-39A.
- All four models including Logistic Regression, SVM, Decision Tree, and KNearestNeighbors are suitable models showing model accuracy more than 80%.
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
- The Decision tree classifier showed the best result for this task.

Thank you!