Winning Space Race with Data Science

Divya Subramanian
July 2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive Analytics result

Introduction

Project background and context

Task A: Determine the price of each launch
Method: By gathering data about Space X + creating dashboards

Task B: Determine if SpaceX will reuse the first stage
Method: Train a machine learning model

Task C: Determine if the first stage will land successfully
Method: Use public information to predict if SpaceX will reuse the first stage

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX API and web scraping from Wikipedia.
- Perform data wrangling
  - One-hot encoding was applied to categorical features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

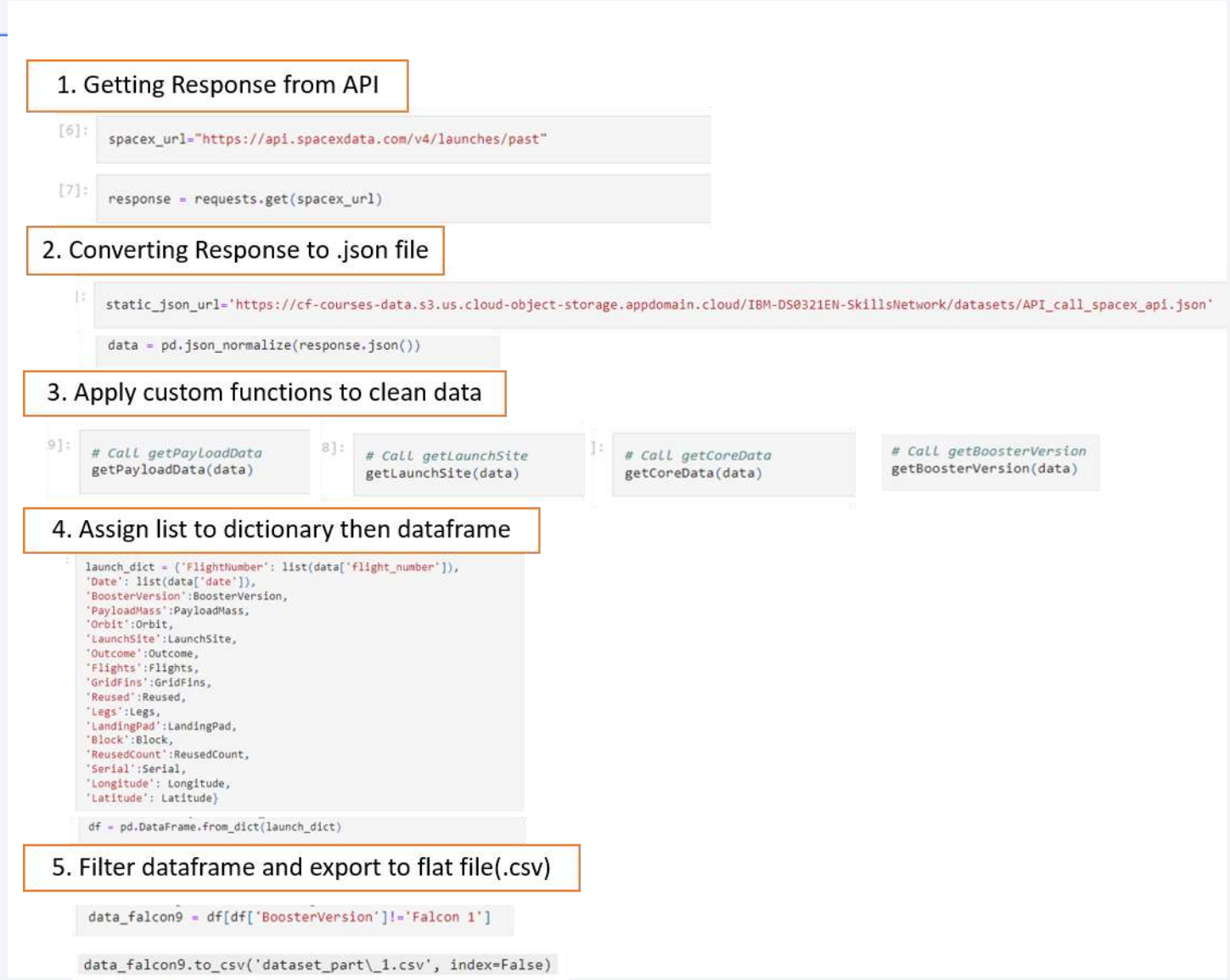
Data Collection

1. Data collected using **Get** request from SpaceX API
2. Response content decoded as a json using **.json()** function call
3. Normalized data turned into pandas dataframe using .json_normalize()
4. Data cleaned, checked for missing values
5. **Web scraping** for Falcon 9 launch data using **BeautifulSoup** from Wikipedia
6. Launch records extracted as HTML table, **parsed** and converted to a pandas dataframe for data analysis

Data Collection – SpaceX API

- 'Get Request' used to collect data from SpaceX API
- Basic data wrangling and formatting done
- Github link click here.

https://github.com/DS-Github-DS/DSIBMCAPSTONE/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Scraping

- Data collected from Wikipedia source
- BeautifulSoup used to web-scrape Falcon 9 launch data
- Table parsed and converted into pandas dataframe.
- Github link click here

https://github.com/DS-Github-DS/DSIBMCAPSTONE/blob/main/jupyter-labs-webscraping.ipynb

1. HTTP Get method to request Falcon 9 data

```
static_url = "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922"
data_falcon9 = requests.get(static_url).text
data_falcon9
```

2. Creating BeautifulSoup object from HTML response

```
soup = BeautifulSoup(data_falcon9, "html.parser")
soup.title
```

```
<title>List of Falcon 9 and Falcon Heavy launches - Wikipedia</title>
```

3. Extracting column data from HTML tables

```
html_tables = soup.find_all('table')

column_names = []

for row in first_launch_table.find_all('th'):
    name = extract_column_from_header(row)
    if name != None and len(name) > 0:
        column_names.append(name)
```

4. Creating a dataframe by parsing HTML tables

```
launch_dict= dict.fromkeys(column_names)
del launch_dict['Date and time ( )']
launch_dict['Flight No.'] = []
launch_dict['Launch site'] = []
launch_dict['Payload'] = []
launch_dict['Payload mass'] = []
launch_dict['Orbit'] = []
launch_dict['Customer'] = []
launch_dict['Launch outcome'] = []
launch_dict['Version Booster']=[]
launch_dict['Booster landing']=[]
launch_dict['Date']=[]
launch_dict['Time']=[]
df= pd.DataFrame({ key:pd.Series(value) for key, value in launch_dict.items() })
```

5. Export data to file(.csv)

```
df.to_csv('spacex_web_scraped.csv', index=False)
```

Data Wrangling

- EDA: Exploratory Data Analysis performed using data collected
- Data summarization of launches per site, orbit occurrences & mission outcomes per orbit type calculated
- Landing outcome label created from outcome column
- GitHub link click here

https://github.com/DS-Github-DS/DSIBMCAPSTONE/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

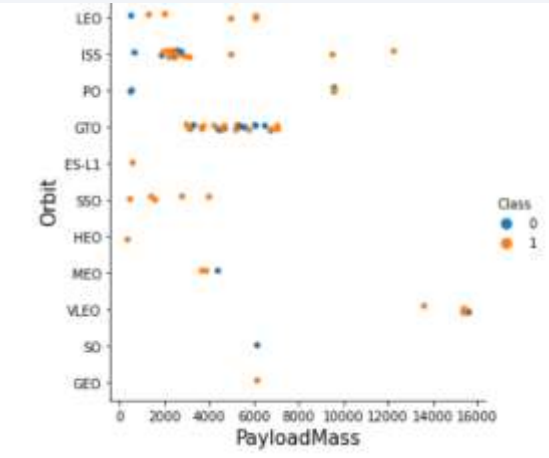
EDA with Data Visualization

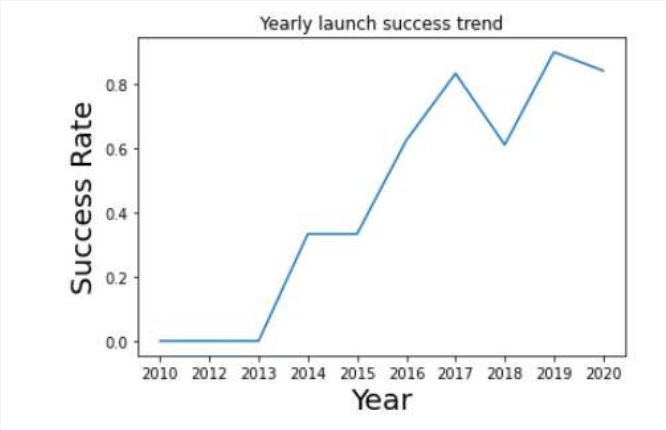
- Scatterplot charts used to identify impact between feature pairs as:

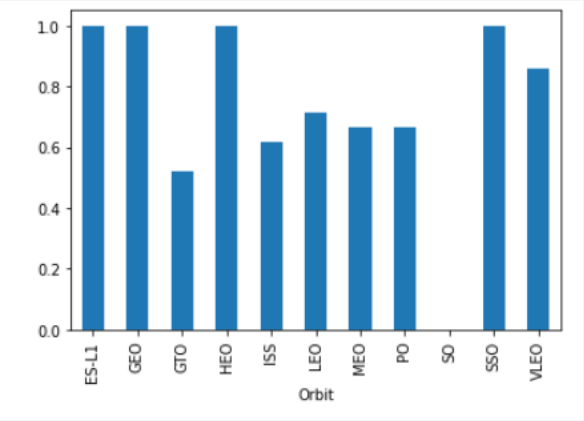
  Flight Number vs. Launch Site, Payload Vs. Launch Site, Flight Number vs. Orbit, Payload vs. Orbit type

- Line chart used to visualize the launch success yearly trend
- Barplot used to study the relationship between success rate and orbit type
- GitHub link click here

https://github.com/DS-Github-DS/DSIBMCAPSTONE/blob/main/jupyter-labs-eda-dataviz.ipynb







EDA with SQL

Following SQL queries performed:

- Display names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List date when first successful landing outcome in ground pad was achieved
- List names of boosters with success in drone ship and payload mass greater than 4000 but less than 6000
- List total number of successful and failure mission outcomes
- List names of the booster versions which have carried the maximum payload mass. Use a subquery
- List records of months, drone ship failure landing outcomes, booster versions, launch site for months in 2015
- Rank count of successful landing outcomes between 04-06-2010 and 20-03-2017 in descending order.

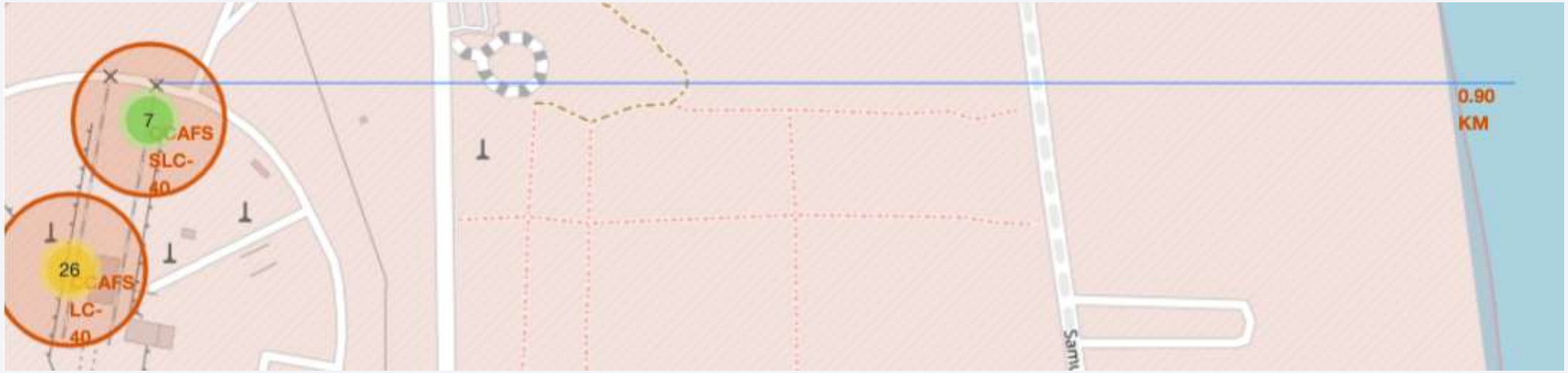
- GitHub link click here

https://github.com/DS-Github-DS/DSIBMCAPSTONE/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Map objects of circle and markers created to indicate each launch site on folium map
- PolyLine used to measure distance between a launch site to the selected point
- Commands folium.Circle, folium.Marker, folium.PolyLine used to create map objects
- GitHub URL click here

https://github.com/DS-Github-DS/DSIBMCAPSTONE/blob/main/lab_jupyter_launch_site_location.ipynb



Build a Dashboard with Plotly Dash

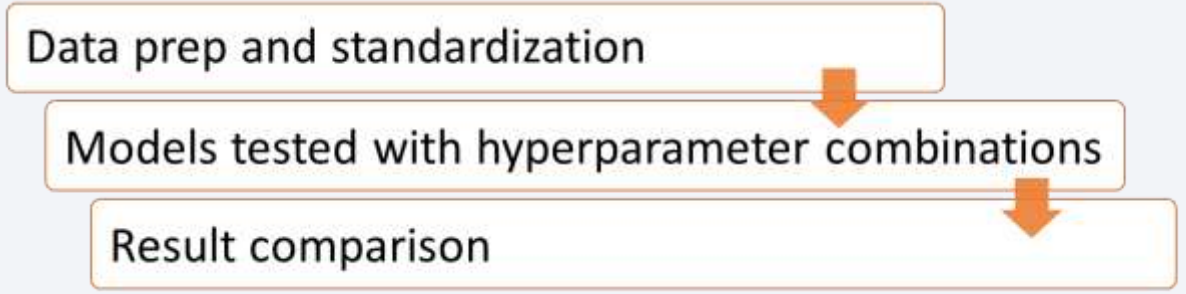
- Interactive dashboard created with Plotly dash
- Pie charts showing percentage of launches by launch site created
- Scatter plot of Outcome vs. Payload Mass (Kg) for different booster version made
- Pie and scatter chart used to easily visualize impact of payload on launch site and identify best launch site
- GitHub URL click here

https://github.com/DS-Github-DS/DSIBMCAPSTONE/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Data loaded using numpy & pandas, transformed, split into training and testing set
- Machine learning models created to test different hyperparameters using GridSearchCV
- Improved ML model accuracy using feature engineering and algorithm tuning
- Logistic regression, support vector machine, decision tree & k-nearest neighbor models compared
- GitHub URL click here

https://github.com/DS-Github-DS/DSIBMCAPSTONE/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



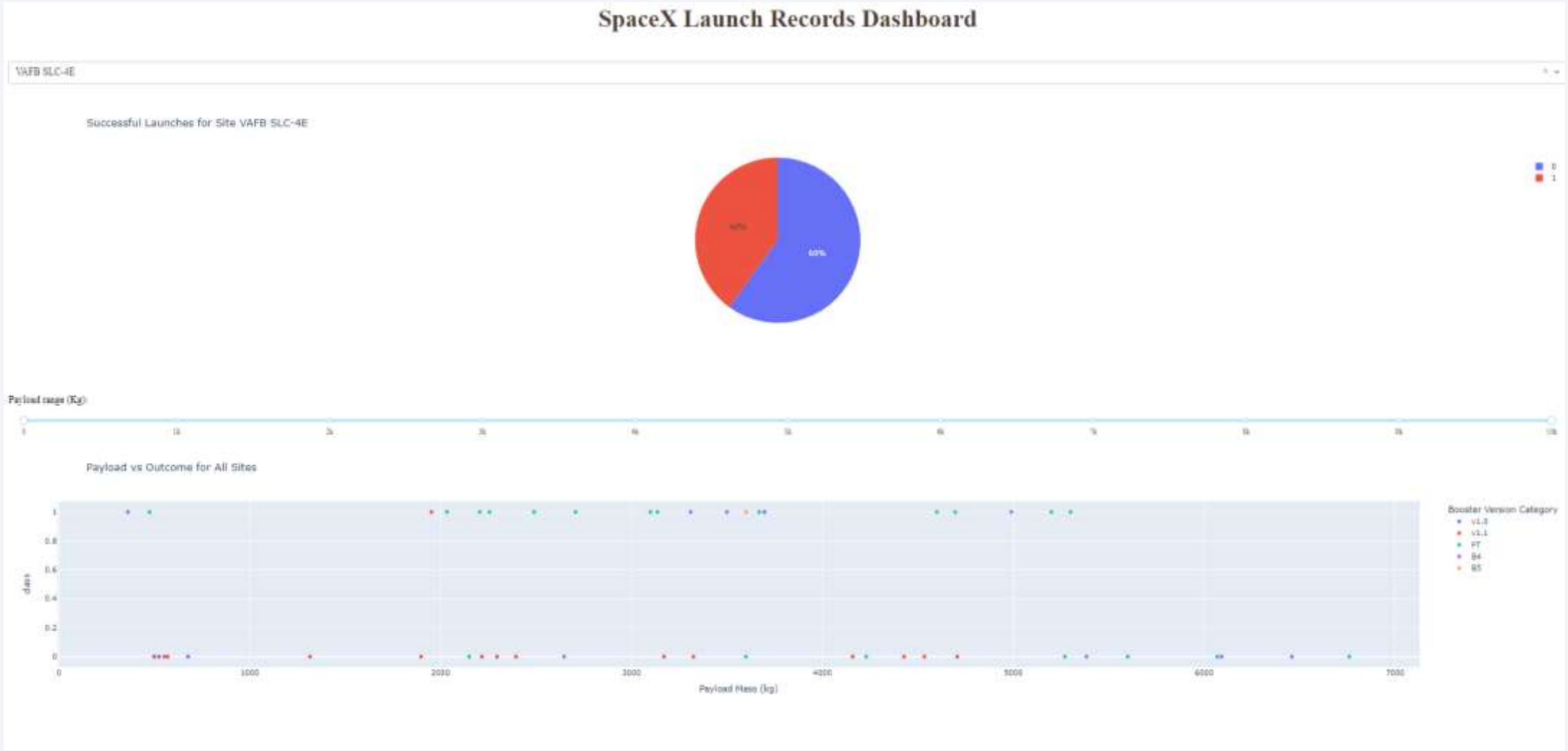
Results - Exploratory data analysis

- SpaceX uses 4 different launch sites
- NASA was the first launch site
- Average payload F9 v1.1 booster is 2928kg
- 2015 was the first successful launch year after 5 years
- Falcon 9 booster was the most successful in landing drone ships
- Failure for booster versions' launch in 2015: F9 v1.1 B1012 & F9 v1.1 B1015
- Landing outcomes improved with the years passing

Results - Interactive analytics demo

- With interactive analytics launch sites observed to be in close proximity to coastline as a safety measure at certain distance away from cities

Results - Interactive analytics demo



Results - Predictive analysis

- 4 classification models tested
- Decision tree classifier model showed highest accuracy
- Test and model accuracy plotted as in table

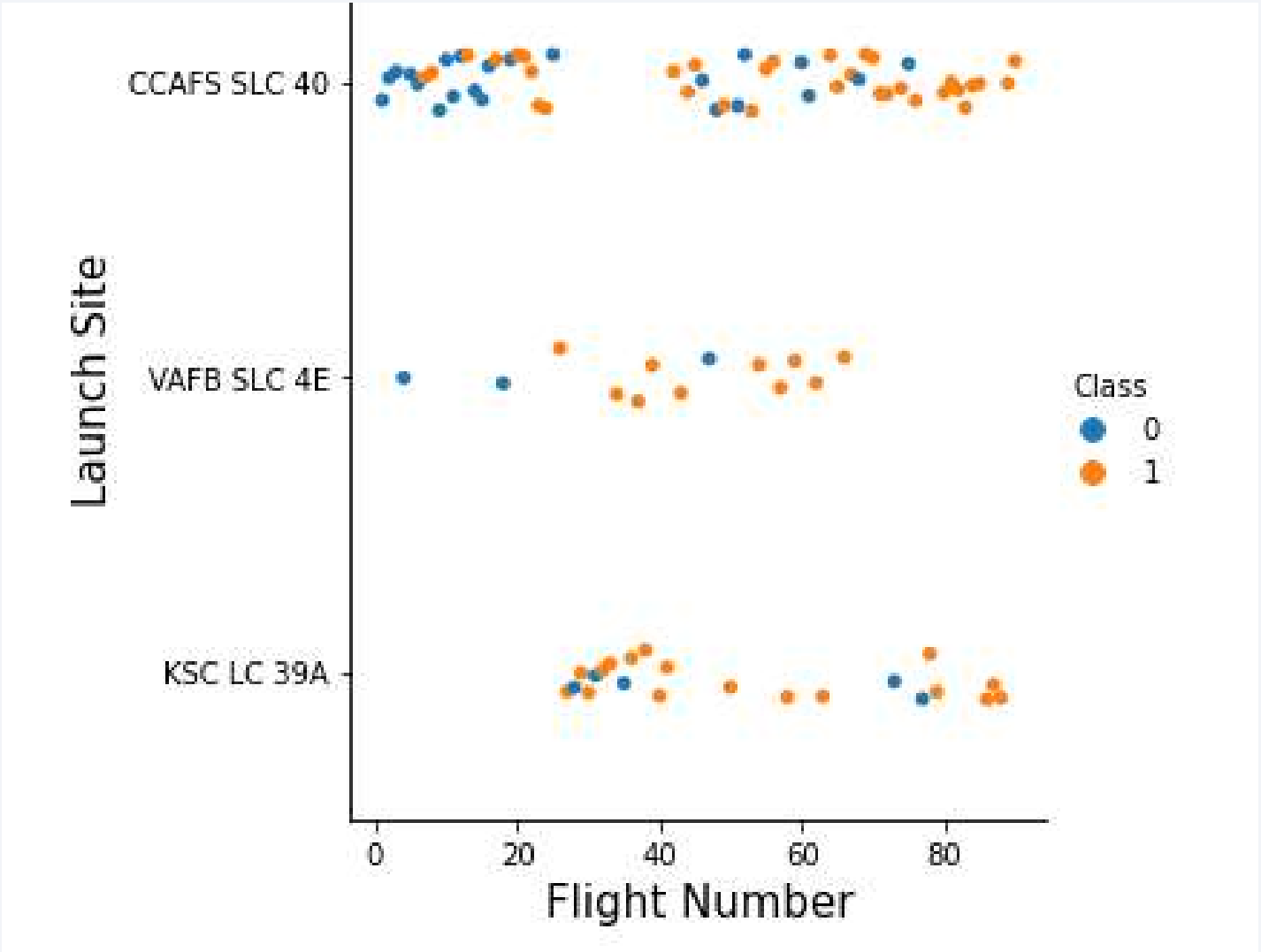
Method	Model Accuracy	Test Accuracy
Logistic Regression	0.84722	0.83333
SVM	0.83334	0.83333
Tree	0.88888	0.88888
KNN	0.84722	0.83333

Section 2

Insights drawn from EDA

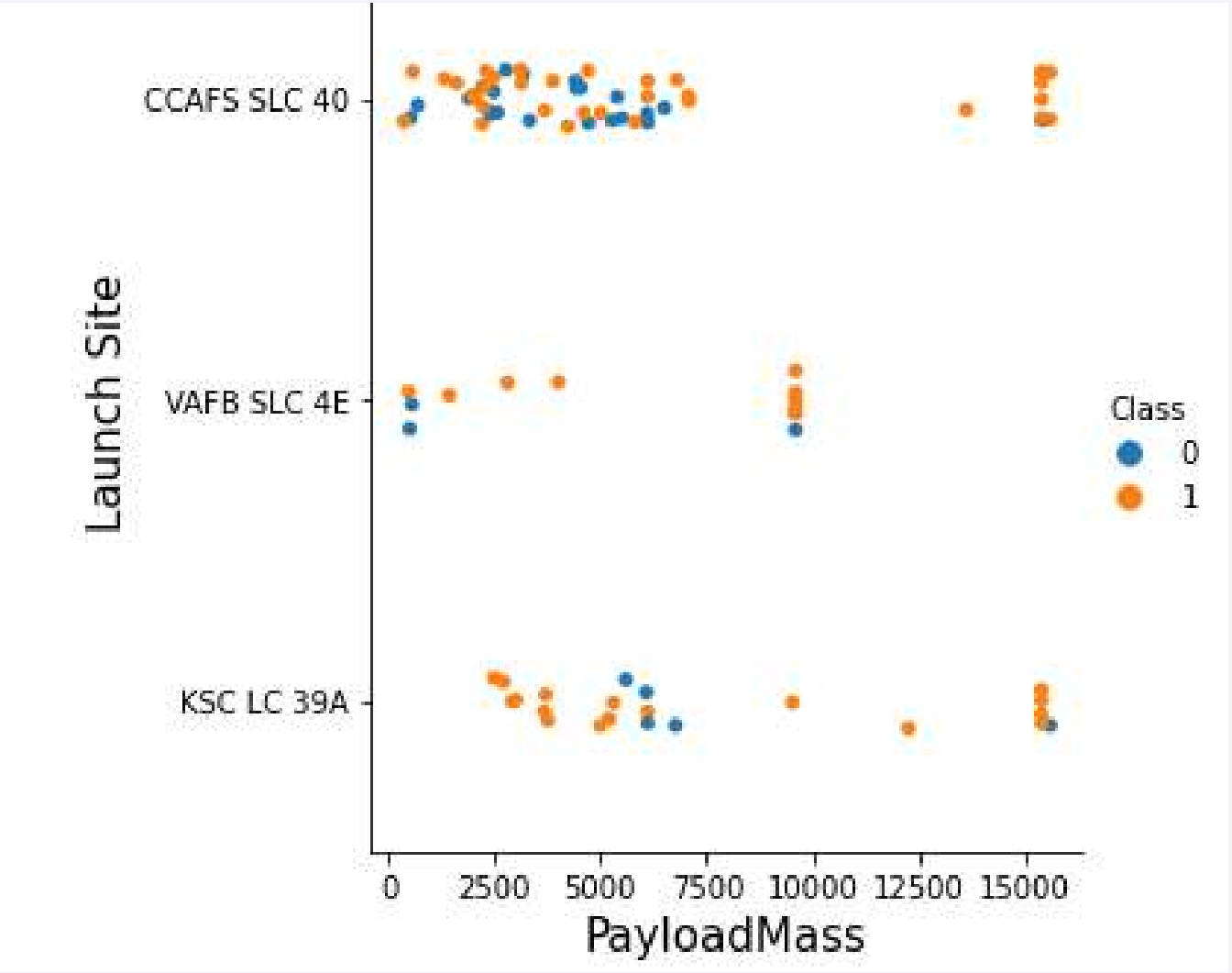
Flight Number vs. Launch Site

- CCAFS SLC 40 and VAFB SLC 4E show higher success rate with increase in flight number
- KSC LC 39A initial flight numbers started with 20+ flight numbers



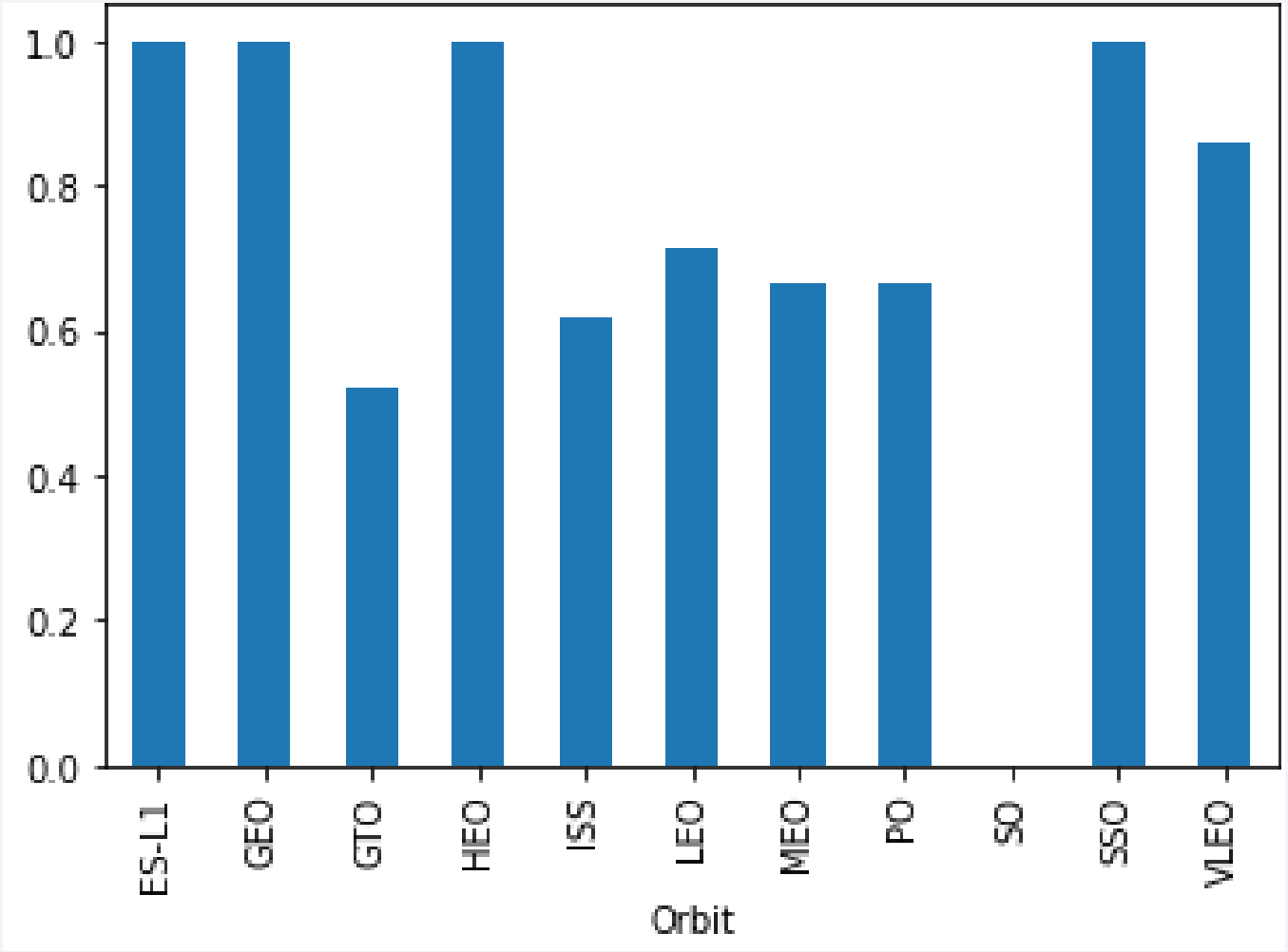
Payload vs. Launch Site

- CCAFS LC-40, has a success rate of 60%,
- But if the mass is above 10,000 kg the success rate is 100%
- VAFB SLC 4E has payload mass capacity within 10000



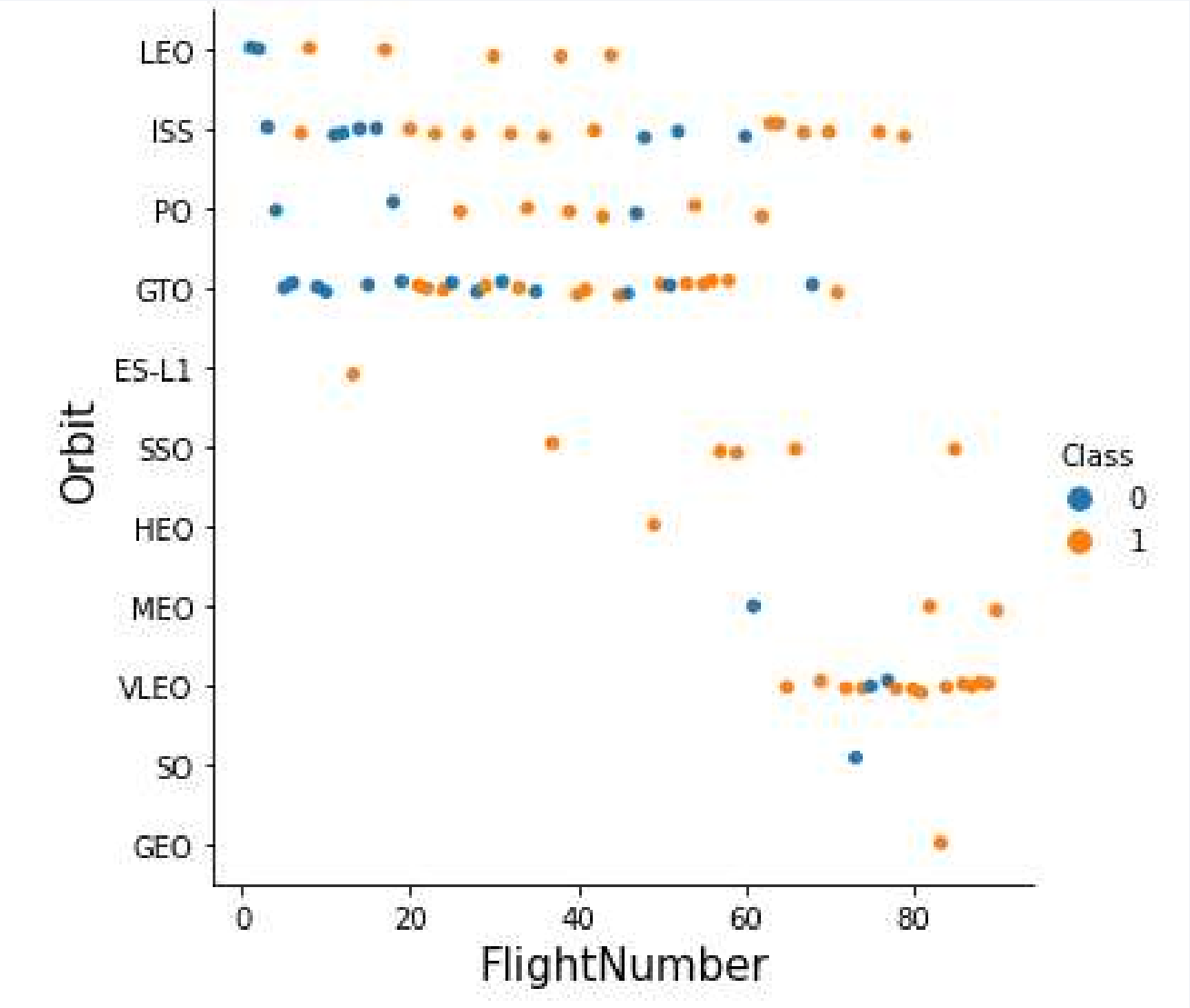
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, SSO show highest success rate
- GTO shows the lowest
- Except SO, all orbits showed success rates 50% and higher



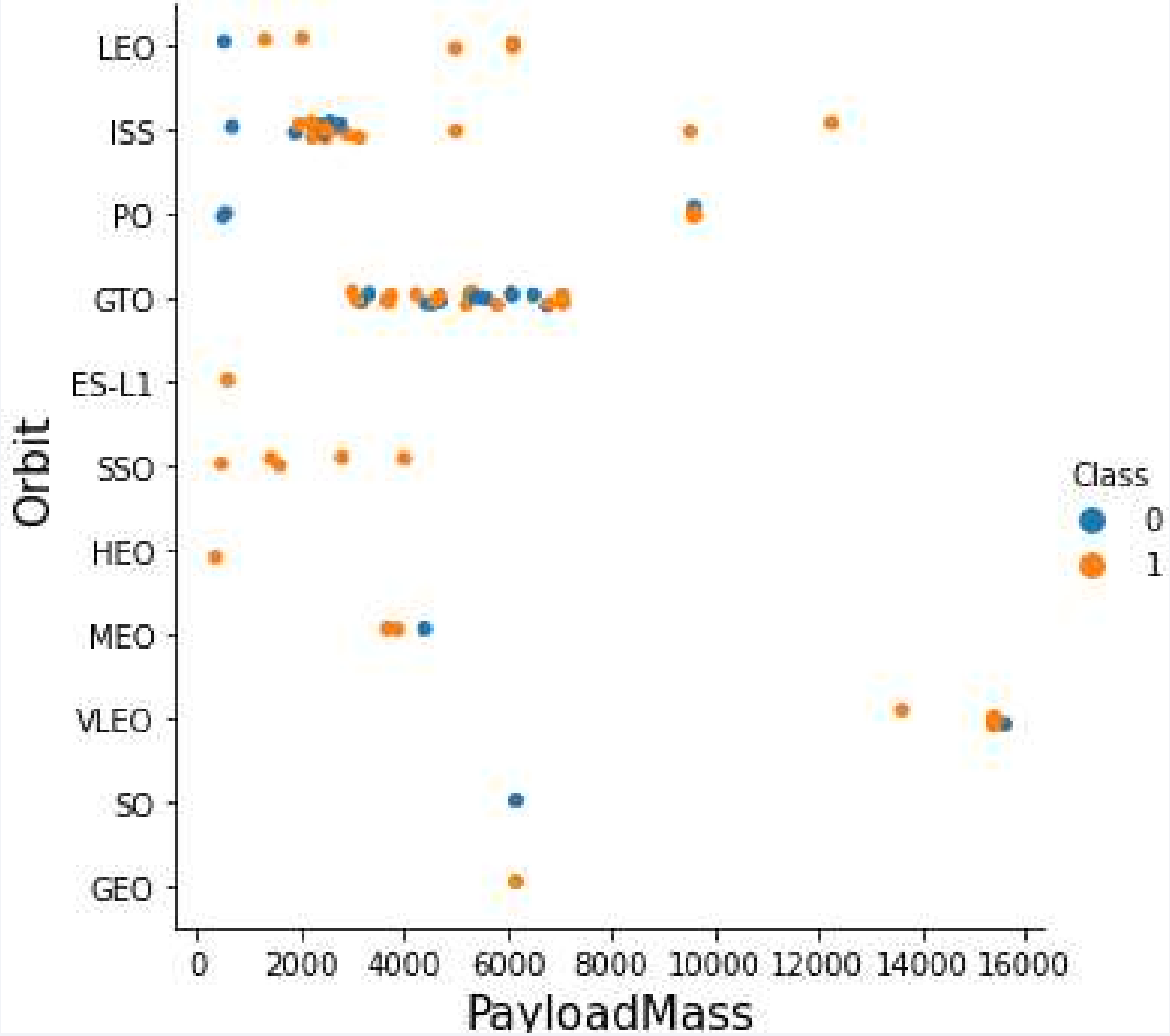
Flight Number vs. Orbit Type

- LEO orbit success appears related to the number of flights
- no relationship between flight number when in GTO orbit seen
- ES L1, HEO, SO and GEO show only singular launch events



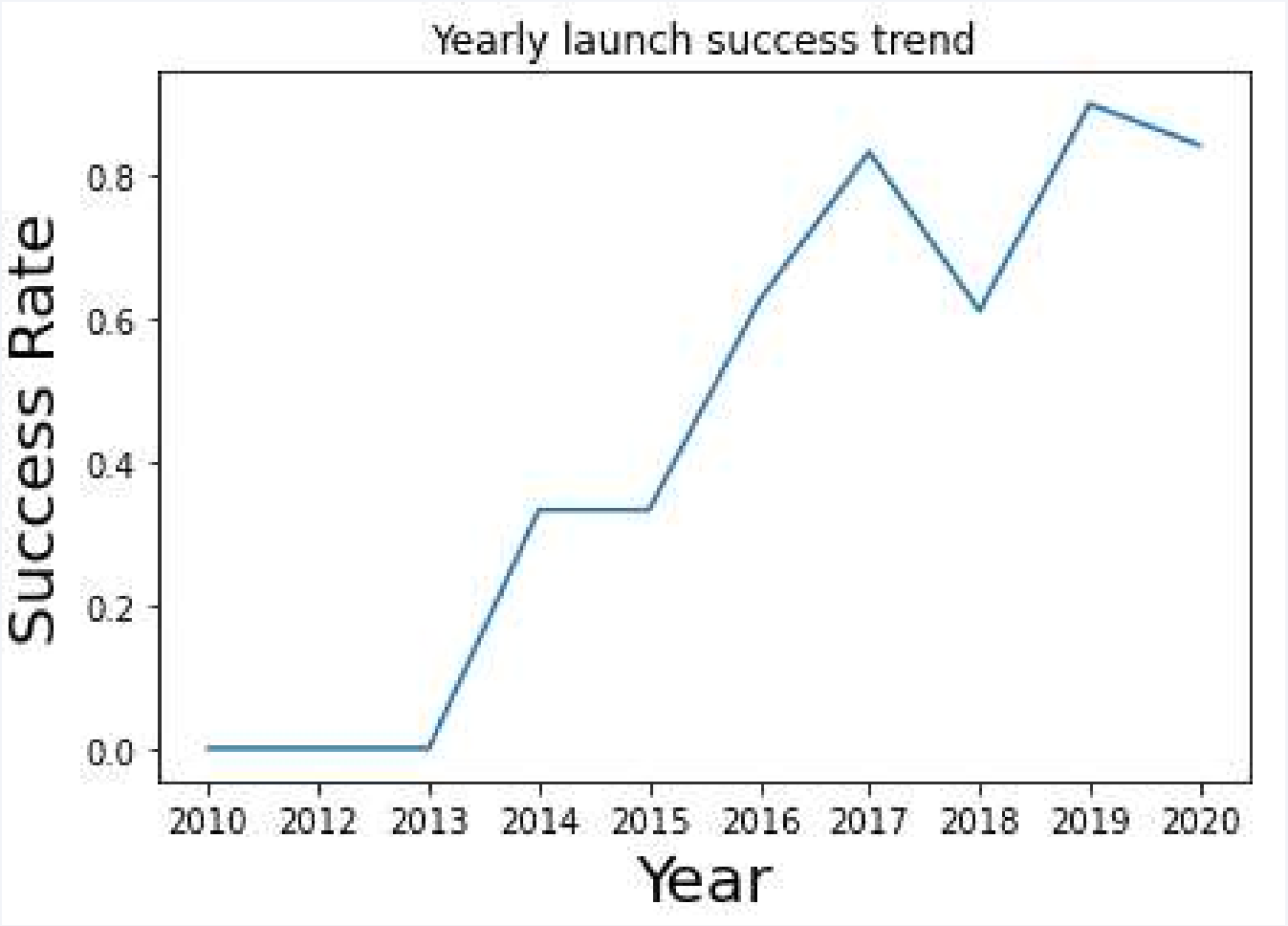
Payload vs. Orbit Type

- Polar, LEO and ISS have more success rate with heavier payloads
- GTO shows both results at most of the payloads
- SSO, HEO, MEO, SO, GEO have launches with lower payloads
- VLEO shows highest payload range



Launch Success Yearly Trend

- As the years advance, the general success trend increases
- 2018 records a significant dip in the success rate after 26 successful falcon 9 launches
- Steepest success increase seen between 2013-2014



All Launch Site Names

- 4 unique launch sites identified using DISTINCT SQL function
- CCAFS LC 40, CCAFS SLC 40, KSC LC 39A and VAFB SLC 4E

```
7]: %%sql  select DISTINCT "Launch_Site"
    from SPACEXTBL

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

7]:

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`
- Like function used to derive the result

[8]:
```
%%sql select * from SPACEXTBL
where "Launch_Site" like 'CCA%' limit 5;
```

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

[8]:

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

- Total payload carried by boosters from NASA as 45596

```
[9]: %%sql SELECT TOTAL(PAYLOAD_MASS__KG_) from SPACEXTBL
         WHERE CUSTOMER = "NASA (CRS)";
 * sqlite:///my_data1.db
Done.
[9]: TOTAL(PAYLOAD_MASS__KG_)
                      45596.0
```

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1 is 2928.4 Kg

```
[10]: %%sql SELECT AVG(PAYLOAD_MASS__KG_) from SPACEXTBL
          WHERE "BOOSTER_VERSION" = "F9 v1.1";
```

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

[10]:

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- First successful landing outcome on ground pad was 1st May 2017

```
[11]: %%sql SELECT MIN("DATE") from SPACEXTBL
              WHERE "Landing _Outcome" = 'Success (ground pad)';
```

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

[11]:

MIN("DATE")
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

Boosters successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are

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

```
%%sql SELECT "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

- Total number of successful and failure mission outcomes are 101

```
%%sql SELECT COUNT(*) from SPACEXTBL
WHERE "Mission_Outcome" LIKE '%Success%' or "Mission_Outcome" LIKE '%Failure%'
```

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

COUNT(*)
101

Boosters Carried Maximum Payload

- Boosters that have carried the maximum payload mass are:

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

- Query used to determine the result:

```
%%sql SELECT "Booster_Version" from SPACEXTBL
where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

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

2015 Launch Records

- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015 are as follows:

```
%%sql
SELECT "Booster_Version", "Launch_Site" , substr(Date, 4, 2) as month FROM SPACEXTBL
WHERE "Landing _Outcome" = 'Failure (drone ship)'

and SUBSTR(Date,7,4)='2015'
```

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

Booster_Version	Launch_Site	month
F9 v1.1 B1012	CCAFS LC-40	01
F9 v1.1 B1015	CCAFS LC-40	04

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

- Count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order:

```
%%sql
SELECT booster_version, launch_site from SPACEXTBL
WHERE "Landing _Outcome" = 'Failure (drone ship)'
and SUBSTR(Date,7,4)='2015'
```

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

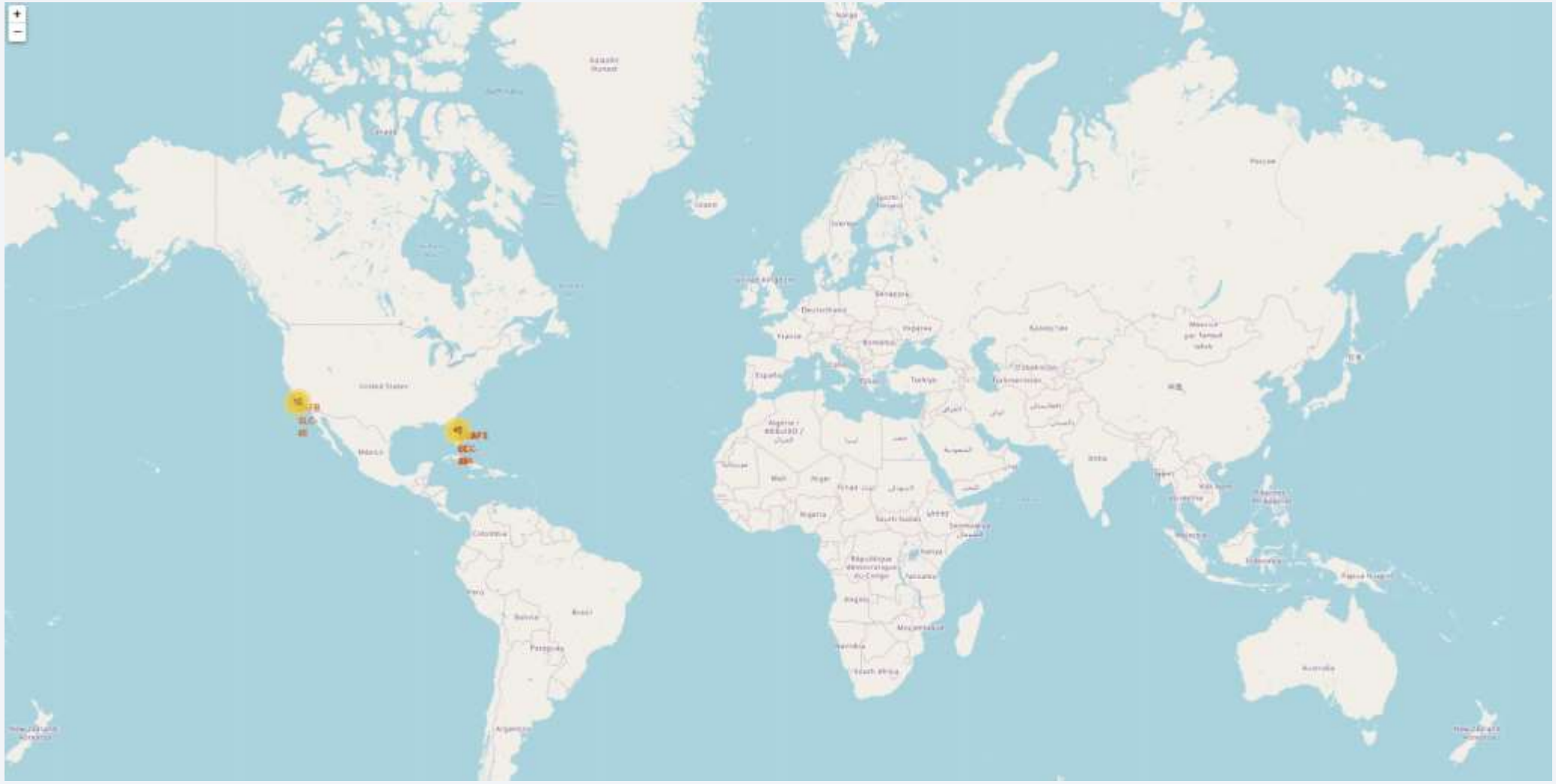
Booster_Version	Launch_Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

Section 3

Launch Sites Proximities Analysis

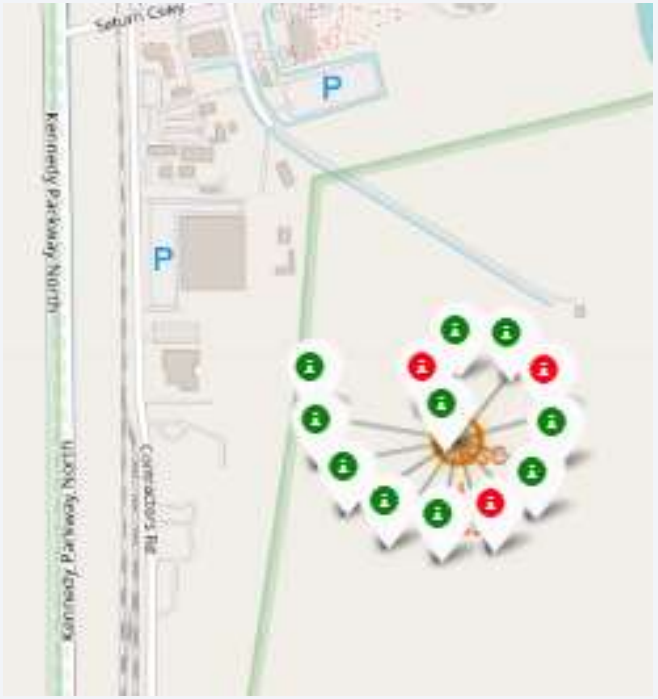
All launch sites location

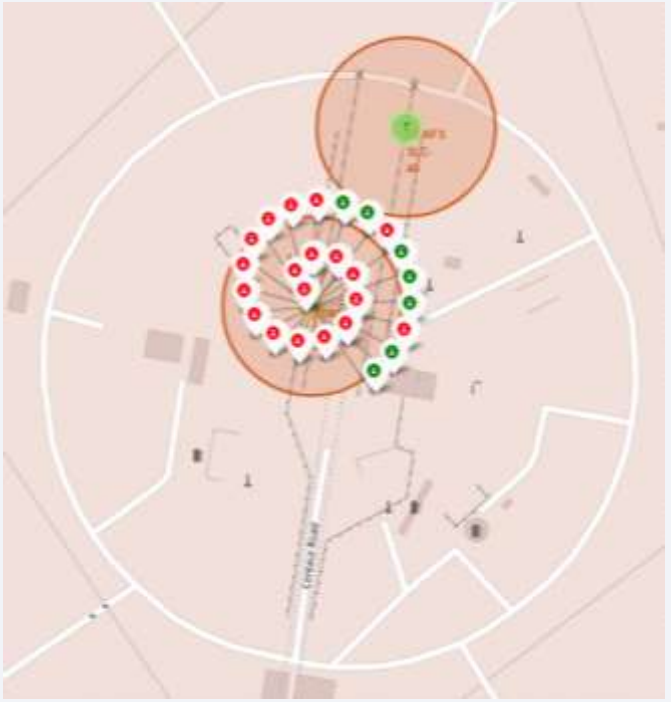
- SpaceX launch sites location are mainly in USA located on the east and west coast



Launch Outcomes

- Green markers show successful launches & Red indicates failed launches

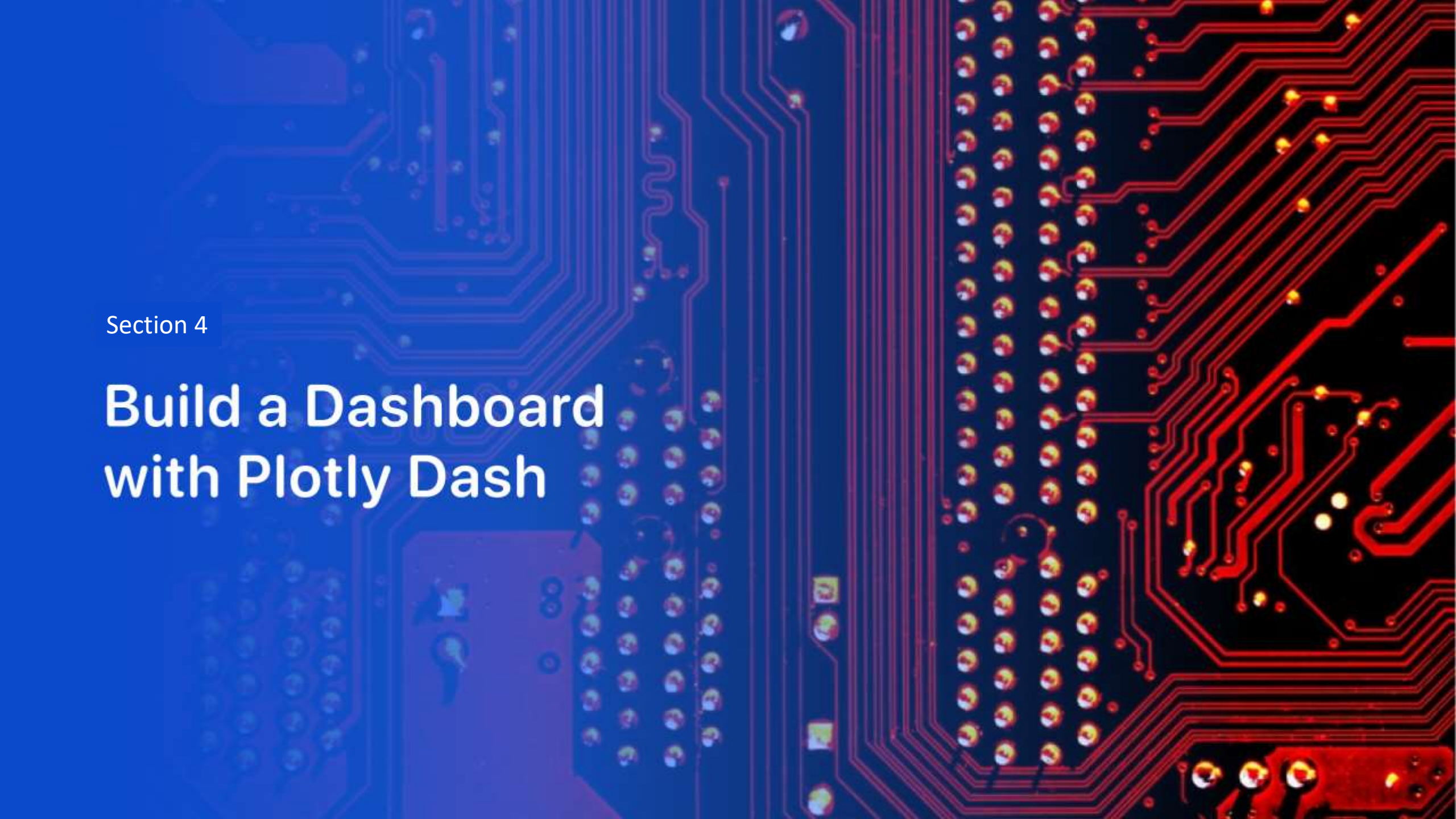




Launch Site Location Features

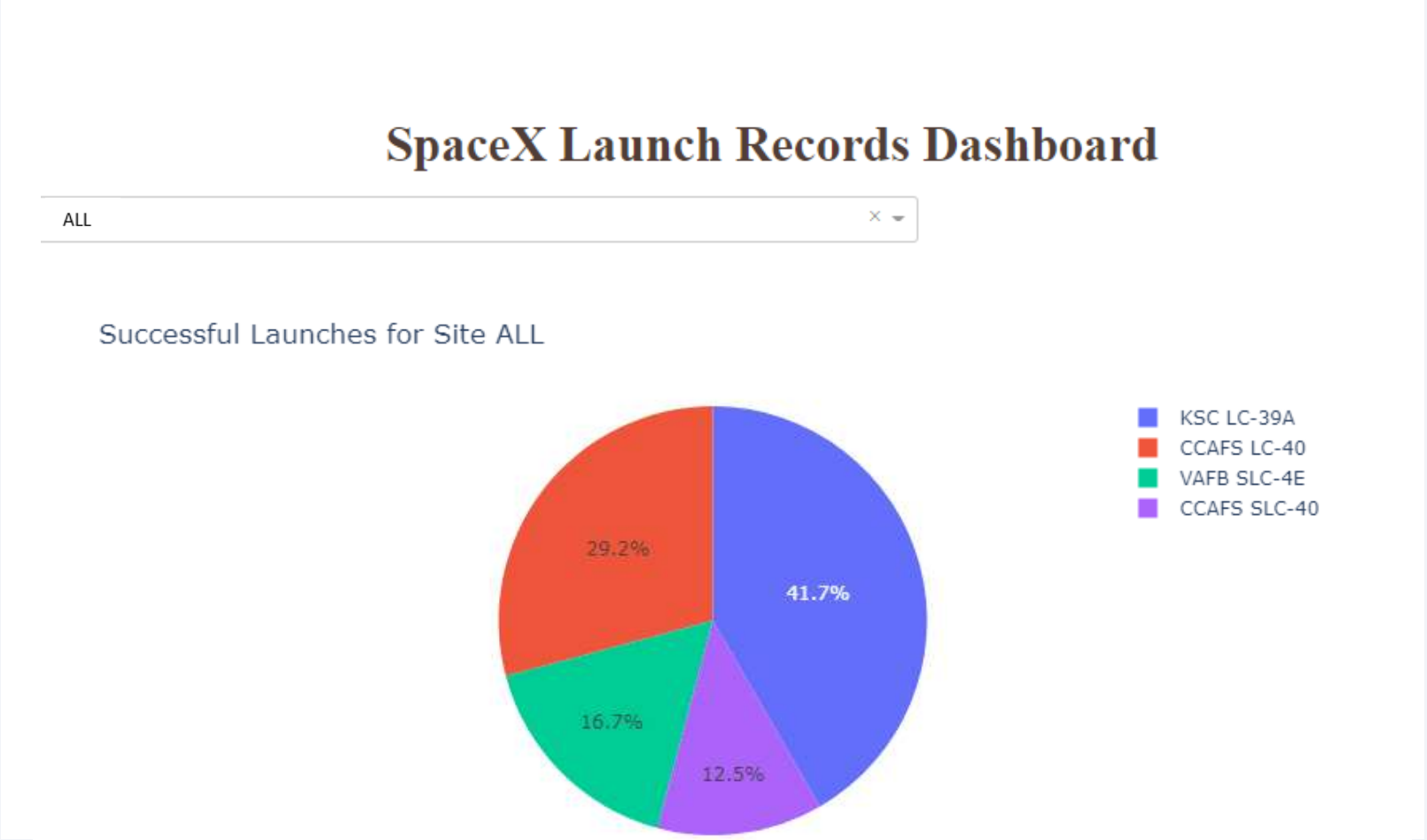
- CCAFS SLC 40 launch site located close to the coastline which is measured using polyline

- KSC LC 39A launch site located close to railways and highways to facilitate in logistics


Section 4
Build a Dashboard with Plotly Dash

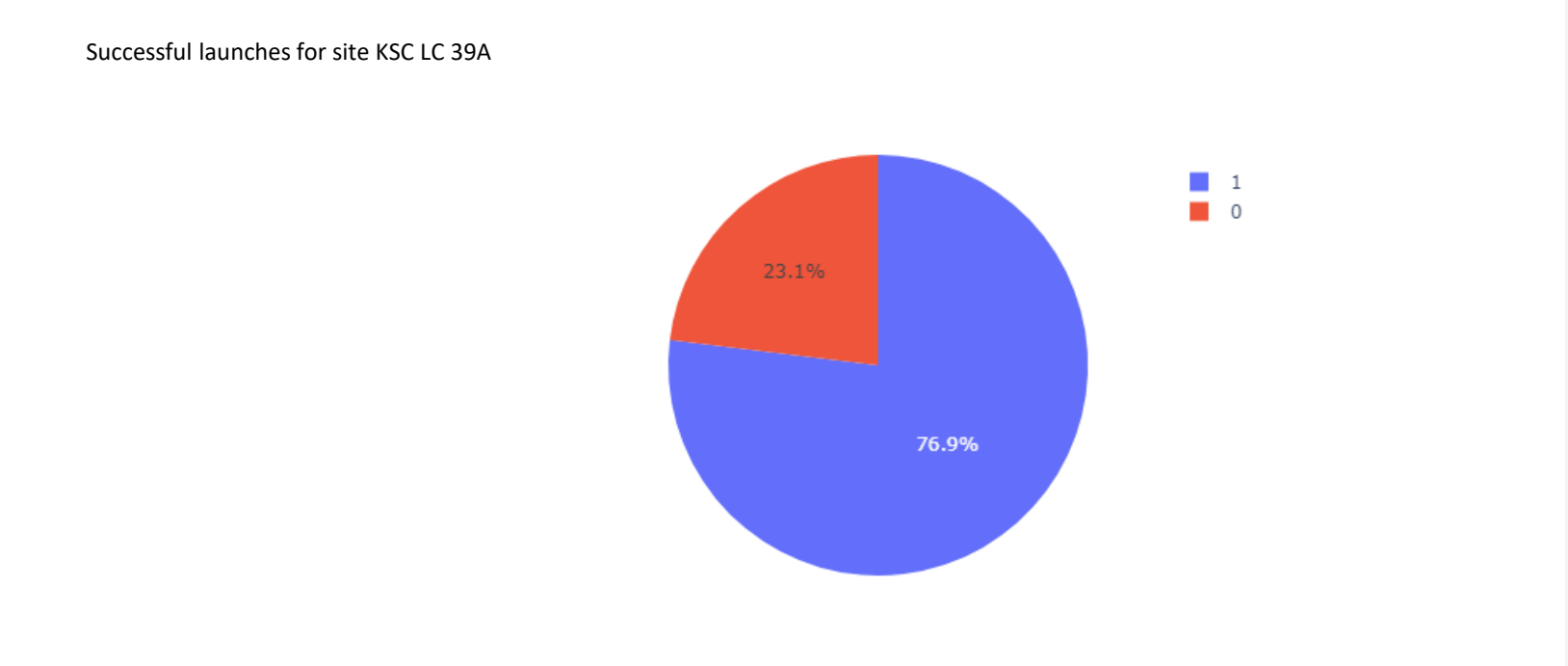
Space X Launch Success by site

- KSC LC 39A shows the highest share of successful launches at 41.7%
- CCAFS SLC 40 shows the lowest share at 12.5%



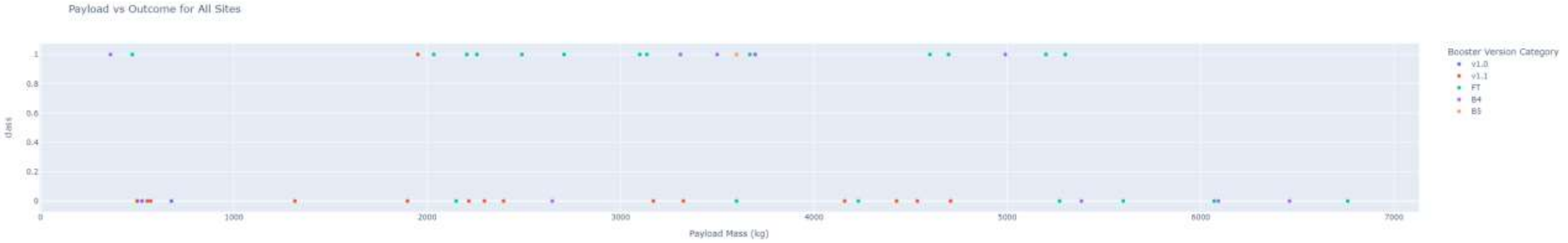
KSC LC 39A Launch Success

- KSC LC 39A has the highest launch success
- 76.9% success vs 23.1% failure for launches



Payload vs. Launch Outcome

- Payloads with lower mass (>6000 kg) have higher success outcomes
- 2000 – 6000 kg payload range has the highest success rate
- Booster version F1 has higher success and v1.1 has lowest while v1.0 has no success reported

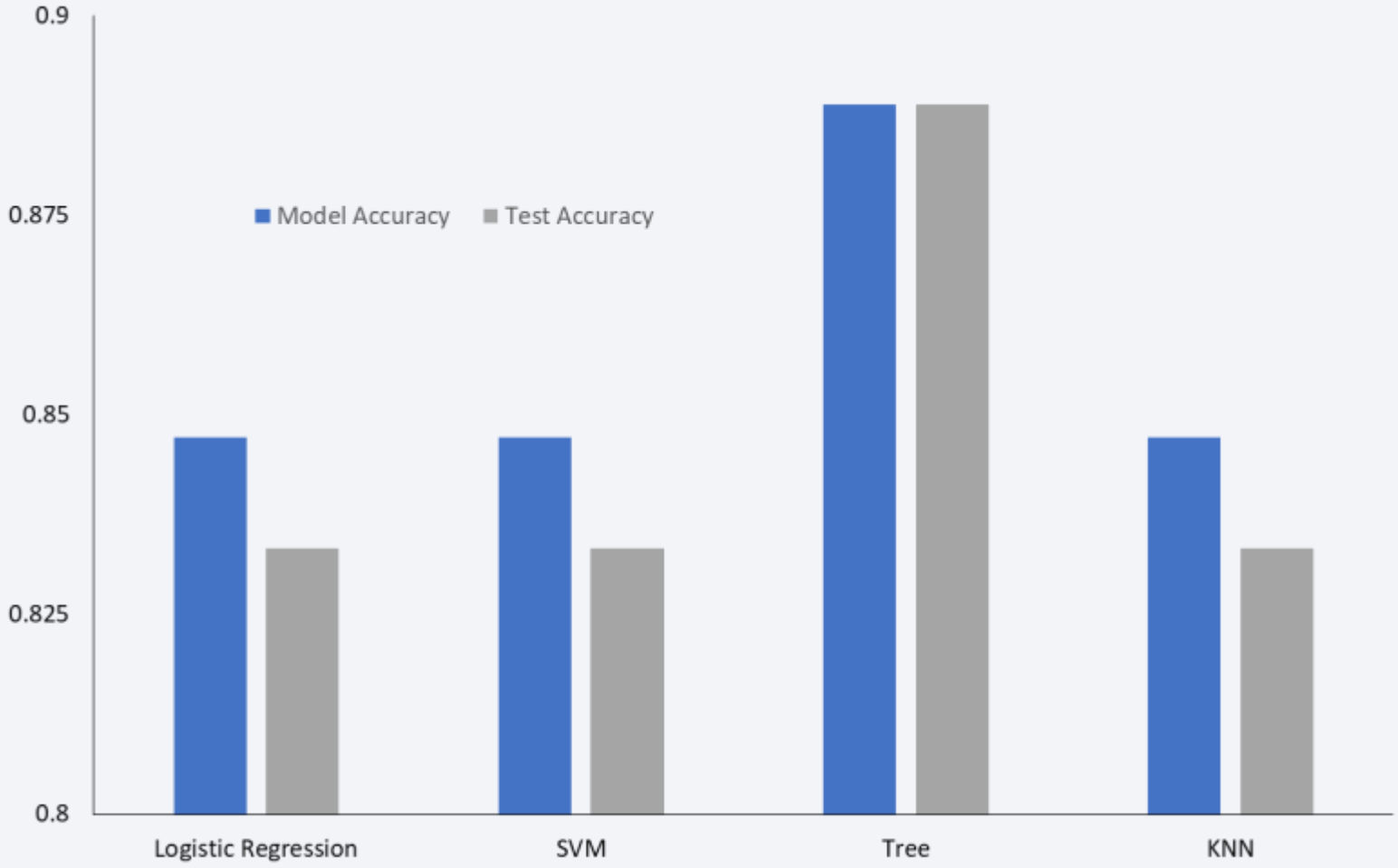


Section 5

Predictive Analysis (Classification)

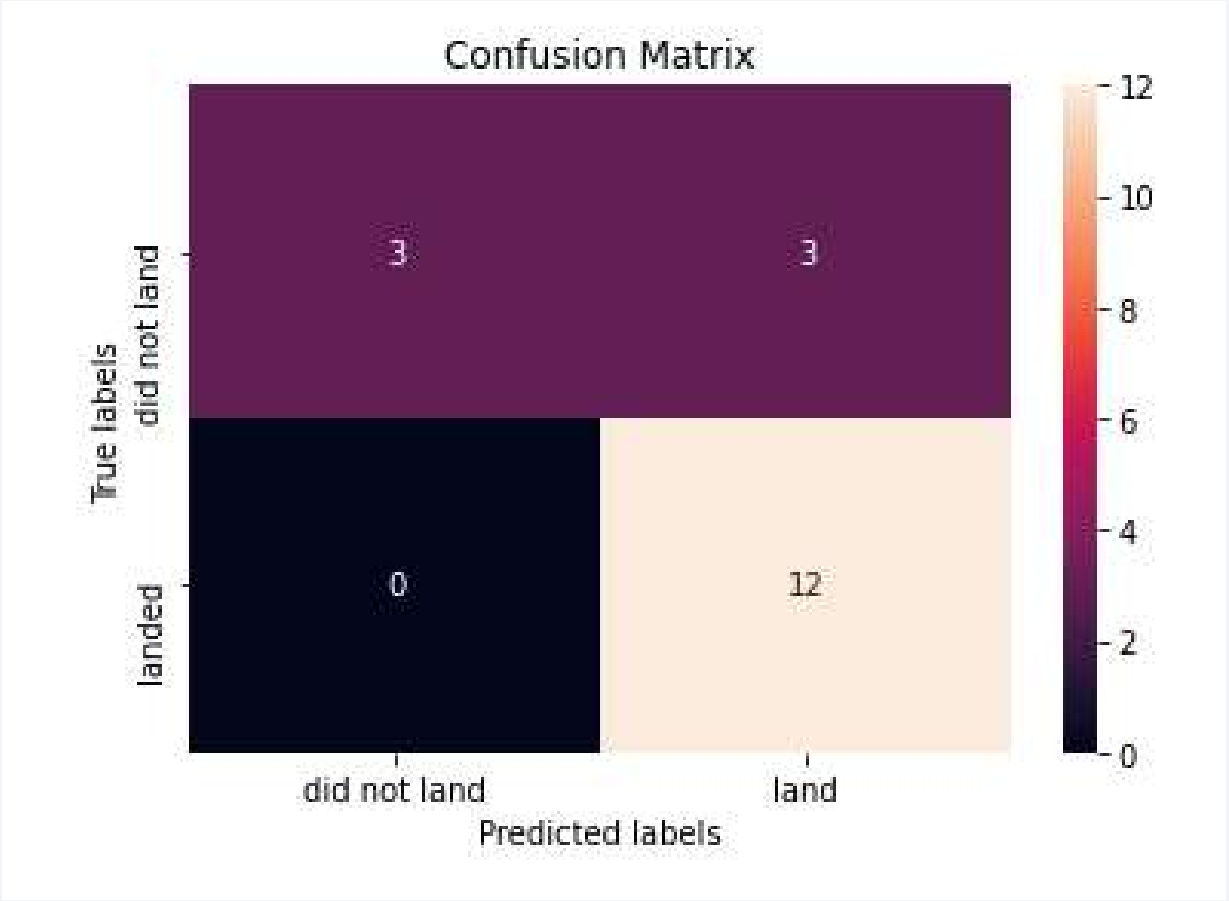
Classification Accuracy

- Bar chart plotted showing the built& test model accuracy for all 4 classification models
- Decision Tree classifier model has the highest classification accuracy



Confusion Matrix

- Confusion matrix of the best performing model is Decision Tree as shown
- The matrix indicates the presence of 3 false negatives and 3 false positives



Conclusions

- KSC LC-39A had the most successful launches
- Decision tree classifier is best suited machine learning algorithm to predict successful landings
- Orbits ES-L1, GEO, HEO, SSO show highest success rate
- Since 2013, the general trend is increasing success rate for launches
- Landing outcomes improved with the years passing
- Payload Mass range 2000 – 6000 kg has the highest success rate for launches

Appendix

- Github link for the project repository click here
- https://github.com/DS-Github-DS/DSIBMCAPSTONE

Thank you!