IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

K Rao
Data Scientist @ space Y
18-Jan-2024

Outline

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

Executive Summary

- Data is collected using spaceX REST API and Webscrapping
- Data analyzed and interactive dash board is created for further analysis
- First stage landing outcome is dependent on several variables
- Logistic, KNN,Tree, SVM models have been created , tested and validated and all of them shows same accuracy.
- Further tuning can be performed to get a model with better accuracy
- First stage landing outcome can be predicted with 83.33% accuracy

Introduction

Company

SPACE Y

- New company in the space transportation services industry founded by billionaire industrialist Allon Musk.
- Competing with the likes of SpaceX and Blue Origin.
- Provide affordable Space Transportation Services

Mission

Predict the failure or success of landing of SpaceX falcon 9

- Analyze the Falcon 9 launches of SpaceX as it is the only company that costs less than others due to the fact that the first stage is reused.
- Retrieve data using SpaceX API and web scraping Wikipedia to predict whether SpaceX will attempt to land a rocket or not.
- Analyze the data to understand if the payload, orbit, launch location, and other parameters have any impact on the failure or success of the first stage landing.
- Provide detailed dashboards and vision of this project.
- Using various machine learning models and displaying the best model that works for Space Y's needs.

Section 1

Methodology

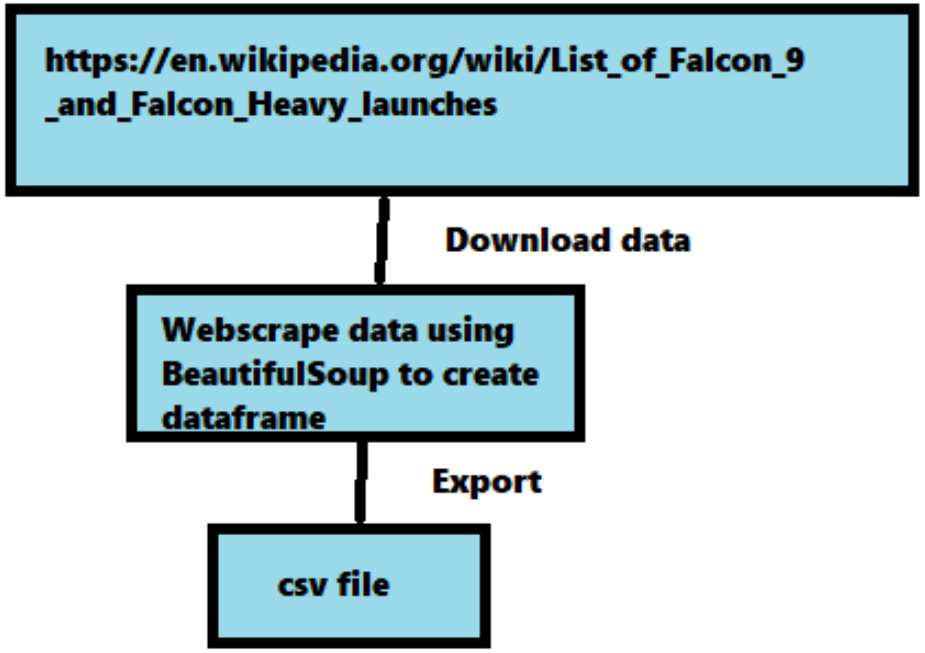
Methodology

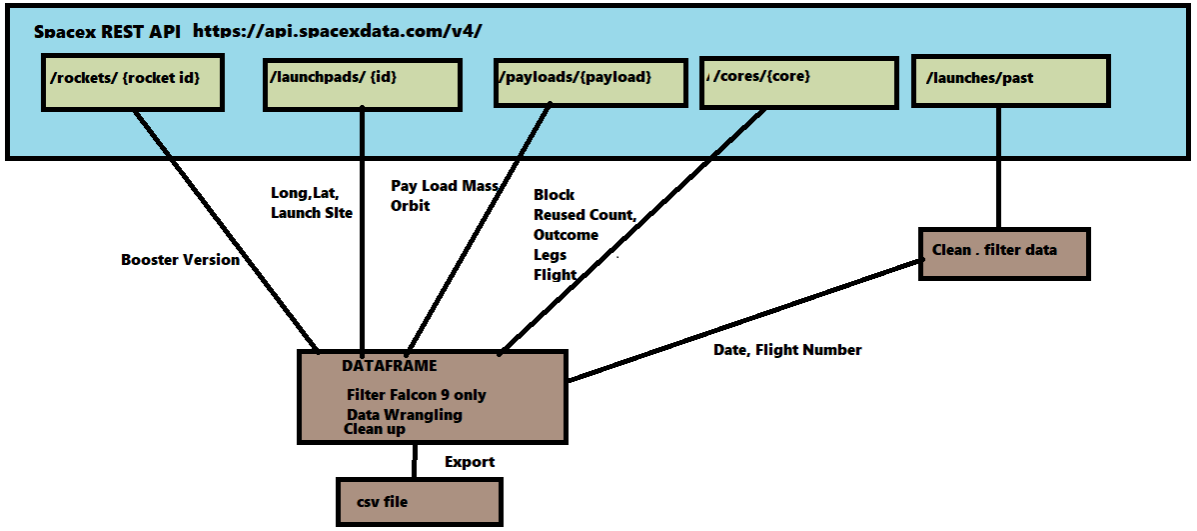
Executive Summary

- **Data collection methodology**:
  - Spacex REST API
  - Web Scraping
- **Perform data wrangling**
  - Data Analysis (Data types, types of labels and counts for categorial field)
  - Data Restructure (Add new columns, update/delete data)
  - Publish (saved as csv)
- **Perform exploratory data analysis (EDA) using visualization and SQL**
- **Perform interactive visual analytics using Folium and Plotly Dash**
- **Perform predictive analysis using classification models**
  - Compare various models using GridSearchCV and select the one with best accuracy

Data Collection

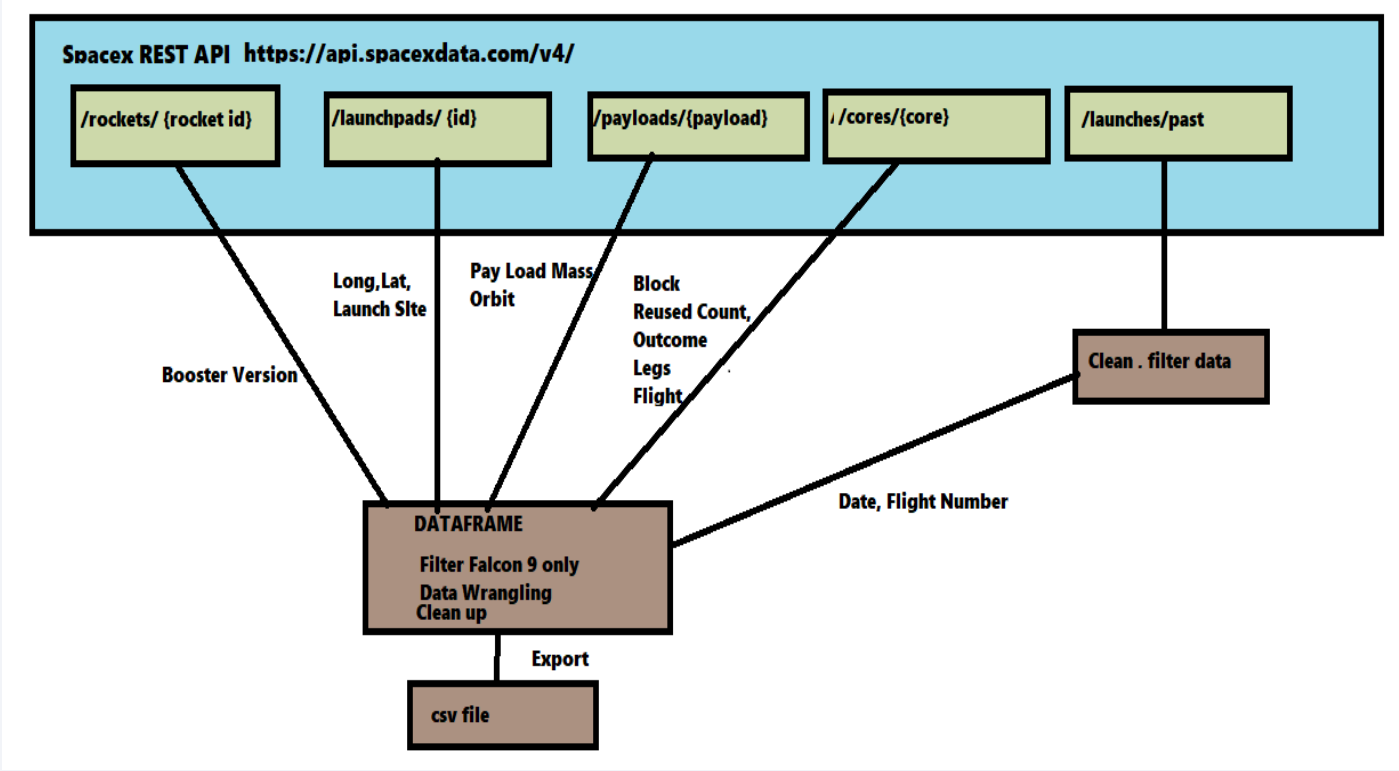
Data is collected by using SpaceX API and Webscraping .





Data Collection – SpaceX API

- Data is retrieved using REST API from https://api.spacexdata.com/v4/
- Download the data using requests object using /launches/past
- Download the payload, launchpad, core data using the appropriate endpoints for each of the record obtained from previous request.
- Merge data to create a data frame
- Cleanup and filter data as required
- Export the dataframe to csv file

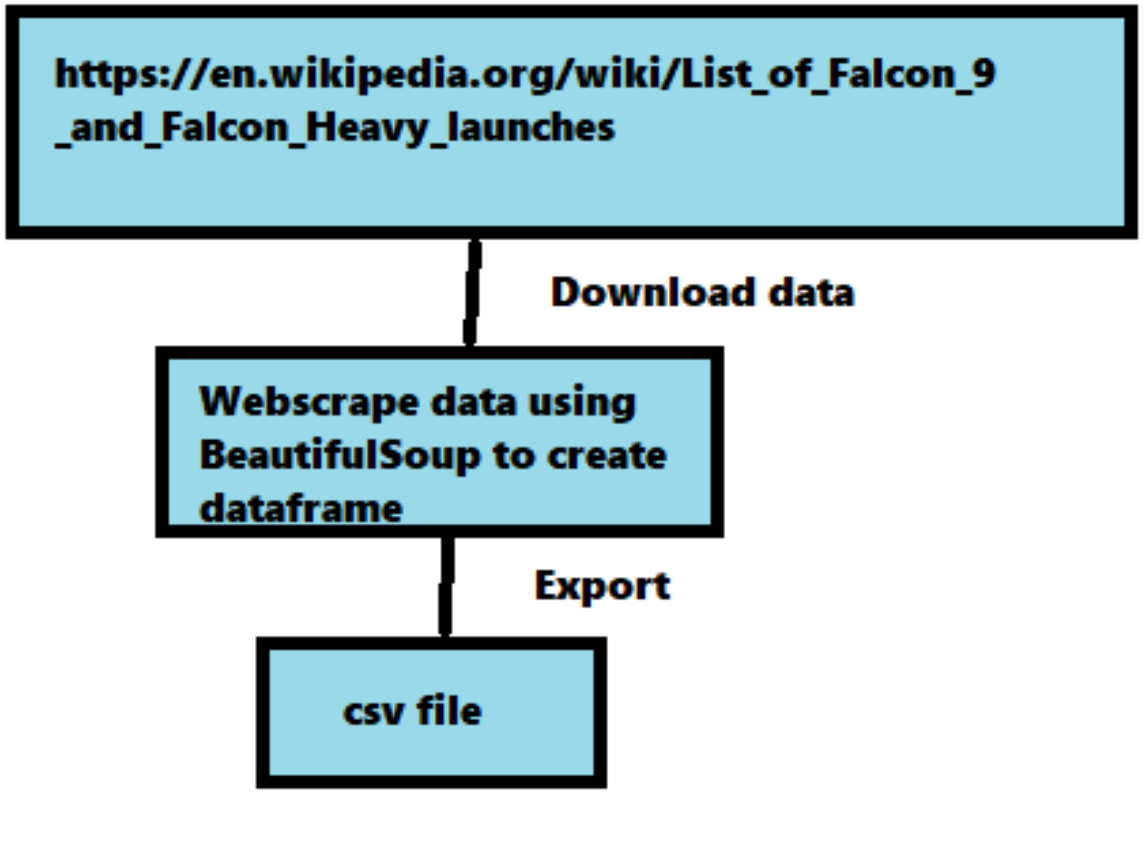


GitHub Jupyter Notebook

Data Collection - Scraping

- Dowload data from Wikepedia page using requests object
- Using BeautifulSoup scrape the data specific to Falcon 9
- Cleanup and filter data
- Export data as CSV

GitHub Jupyter Notebook



Data Wrangling

- Discover
  - Data types
  - Identify Categorical labels (Launch Sites, Outcomes, orbit etc)
  - Occurrences of Categorical data
- Transformation
  - Create class Column with binary values based on Outcome column
- Validation
  - Ensure changes have been applied
- Publish
  - Save the data to a csv file

GitHub Jupyter Notebook

EDA with Data Visualization

- Scatter plot : Outcome is overlayed in each case
  - Flight Number vs Payload Mass
    - Success rate is higher for flights with higher payload
  - Flight Number vs Launch Site
    - Success is related to number of flights at all sites
    - No rockets launched from VAFB-SLC recently
  - Payload vs Launch Site
    - VAFB-SLC launch site did not launch rockets with payload greater than 10000
  - Flight Number vs Orbit type
    - Some orbit types like LEO are related to flight number with higher success rate.
  - Payload vs Orbit Type
    - Success rate is higher for some of the orbit types (Polar ,LEO and ISS) for higher payload
- Line Plot : Success rate increased since 2013 to 2019 with a dip in 2018

GitHub Jupyter Notebook

EDA with SQL

- 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 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_versions 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 year 2015.
- Rank the 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.

GitHub Jupyter Notebook

Build an Interactive Map with Folium

- Added **Circle** and **Marker** on each site to analyze if the location is closer to coastal line or equator line. All launch sites are on costal area. Sites near water bodies are preferred as launch site
- Created **Marker Cluster** object and markers to represent success and failure to identify location with higher success rate. (KSL and VAFB has higher success rate compare to CCAFS)
- **Lines** drawn between launch sites and the closest railway road/ highway /city and distance was calculated. All sites are closer to railway and highway. This helps with transportation. All sites are away from highly populated area reduces the risk to population and property on launch disasters

GitHub Jupyter Notebook

Build a Dashboard with Plotly Dash

- Dropdown allows user to select "All sites" or individual sites.
- A slider allows user to change the range of payload mass
- Pie chart
  - Drop down="All sites": Displays the success rate for various sites.
  - Dropdown = Specific site: Displays success and failure rate for the selected site
- Scatter Chart
  - Selected payload mass range relationship with all or selected launch sites overlayed with Booster Version

GitHub Plotly

Predictive Analysis (Classification)

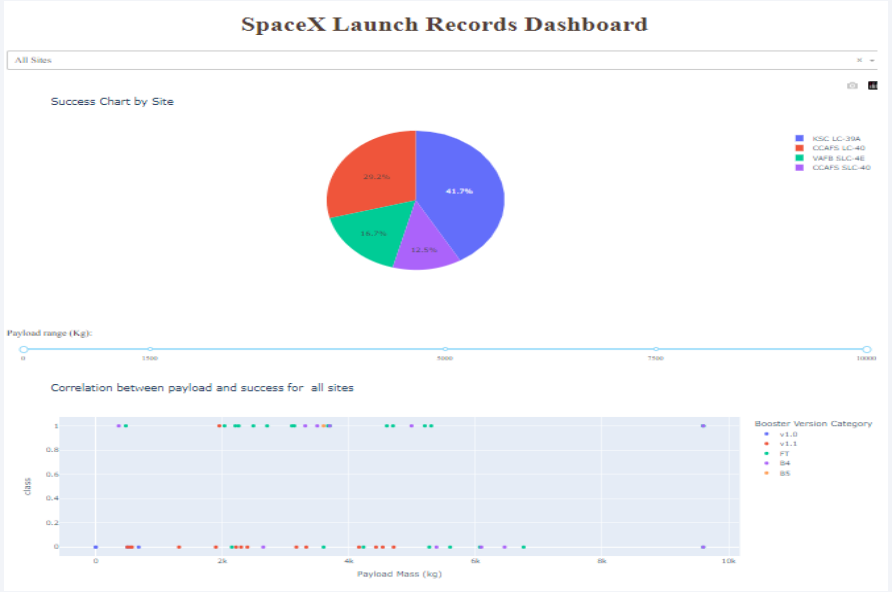
- Data Preparation
  - Created dataframe from the CSV file
  - Created the Outcome
  - inependent variables are standardized using standardscalar
- Define Model Validation Strategy
  - Created test and train set with test size of 20% and randon state=2. 18 test samples were obtained
- Model Development
  - Logistic Regression, SVM, KNN, Decision tree models are created
  - For each of the model
    - GridSearchCv is used for hyperparameter tuning. Cv =10 is used
    - Trained the data and checked the accuracy
    - Validated the data by testing and checking the accuracy
    - Analyzed the confusion matrix
- Model Selection
  - Compared the accuracy of all the models and it is found all of them have same accuracy.

GitHub Jupyter Notebook

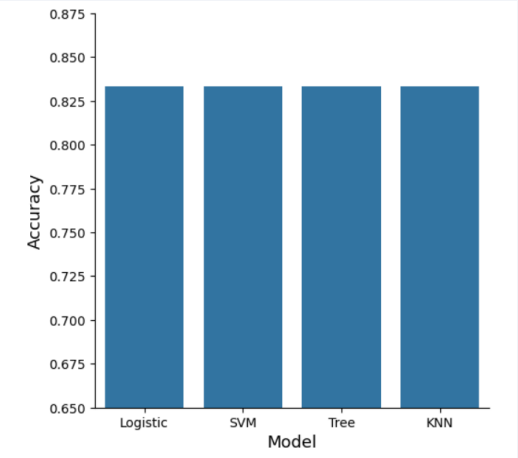
Results

- Exploratory data analysis results
  - As flight number increased the first stage is more likely to land successfully
  - As the payload increases less likely the first stage return
  - Success rate varies by launch site
  - Success rate is higher for some of the orbit type like GEO, ES-L1

Interactive analytics demo in screenshots

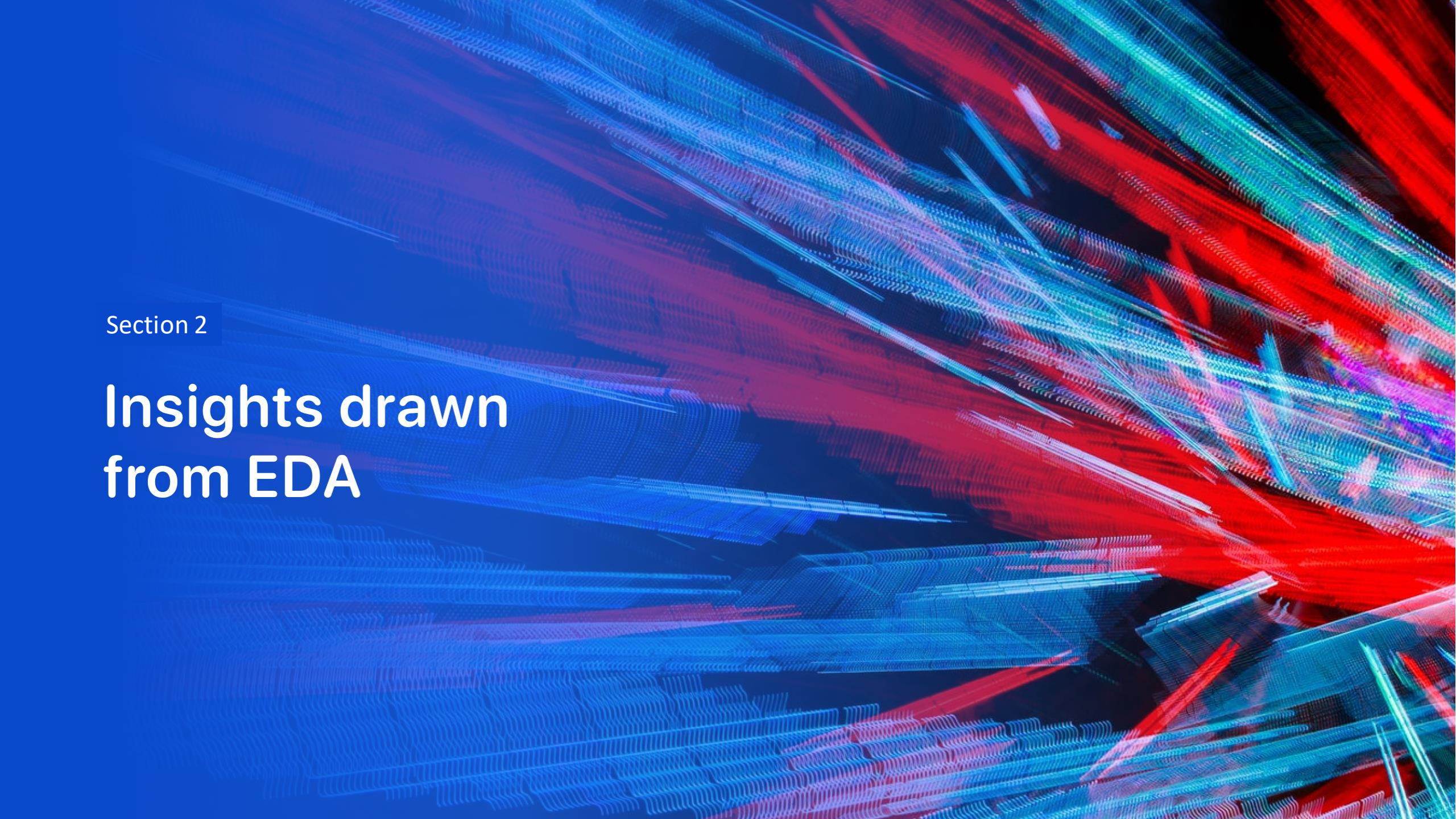


Predictive Analysis results



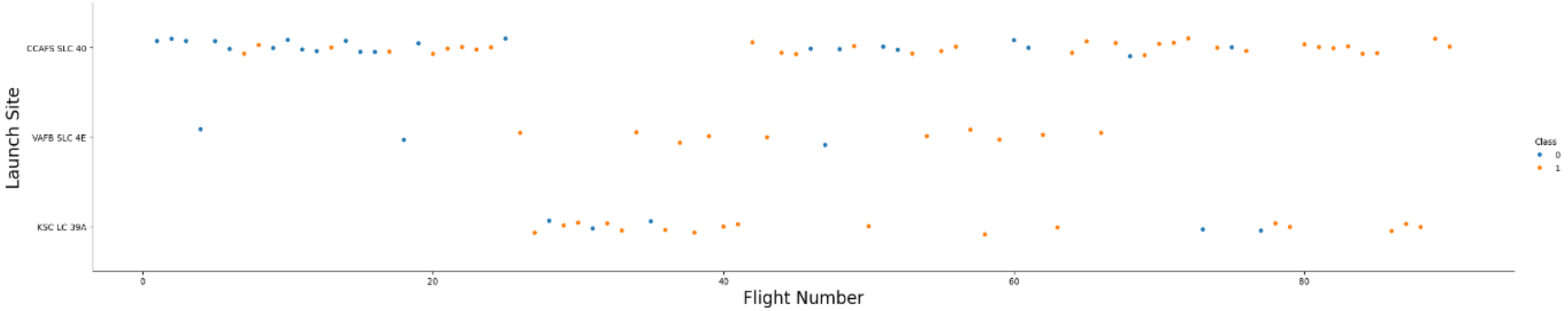
Section 2

Insights drawn from EDA



Flight Number vs. Launch Site

[6]:
```
### TASK 1: Visualize the relationship between Flight Number and Launch Site
sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()
```

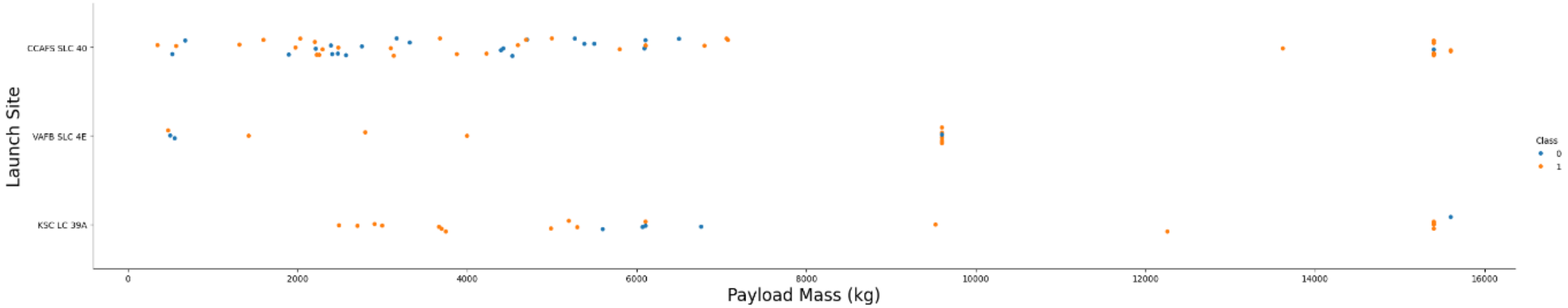


Use the function `catplot` to plot `FlightNumber` vs `LaunchSite`, set the parameter `x` parameter to `FlightNumber`,set the `y` to `Launch Site` and set the parameter `hue` to `'class'`

Success rate is higher for increase in flight number

Payload vs. Launch Site

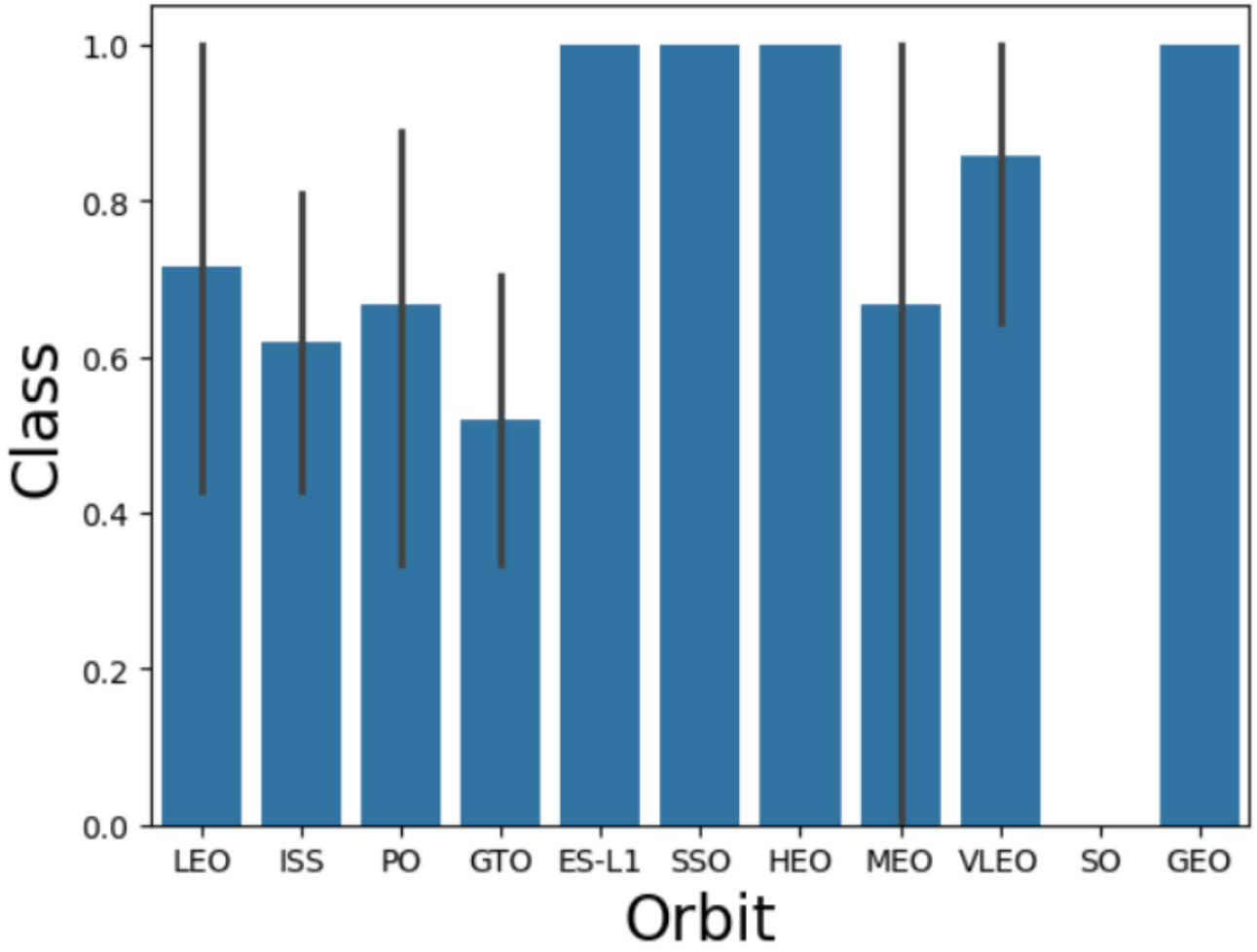
[7]:
```
### TASK 2: Visualize the relationship between Payload and Launch Site
sns.catplot(y="LaunchSite", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("Payload Mass (kg)",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()
```



We also want to observe if there is any relationship between launch sites and their payload mass.

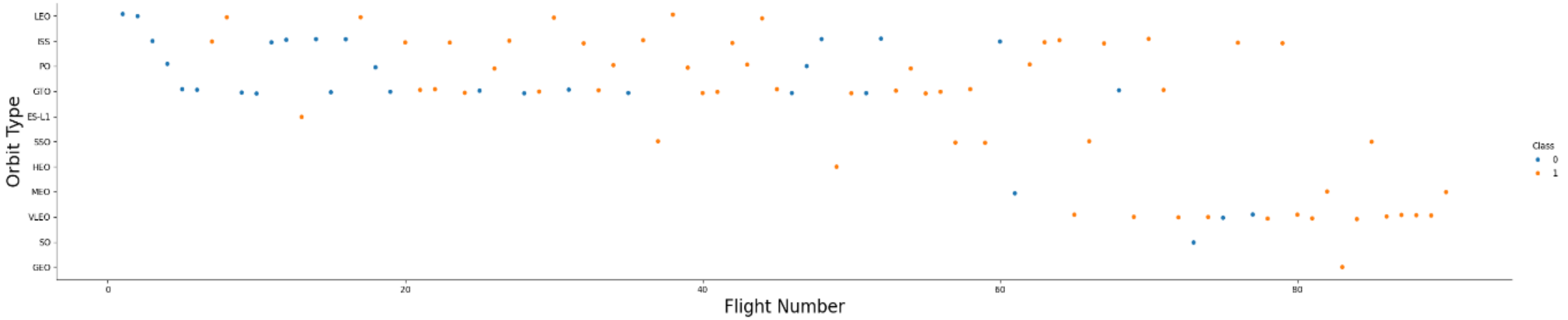
VAFB-SLC launch site has no rockets launched for payload mass more than 10000

Success Rate vs. Orbit Type



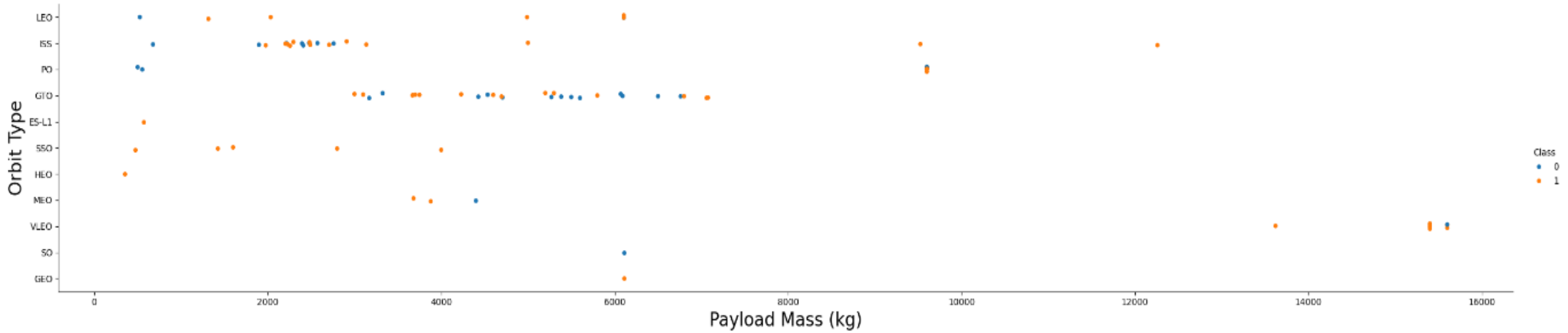
Success rate is higher for orbit types GEO, ES-L1,SSO,HEO

Flight Number vs. Orbit Type



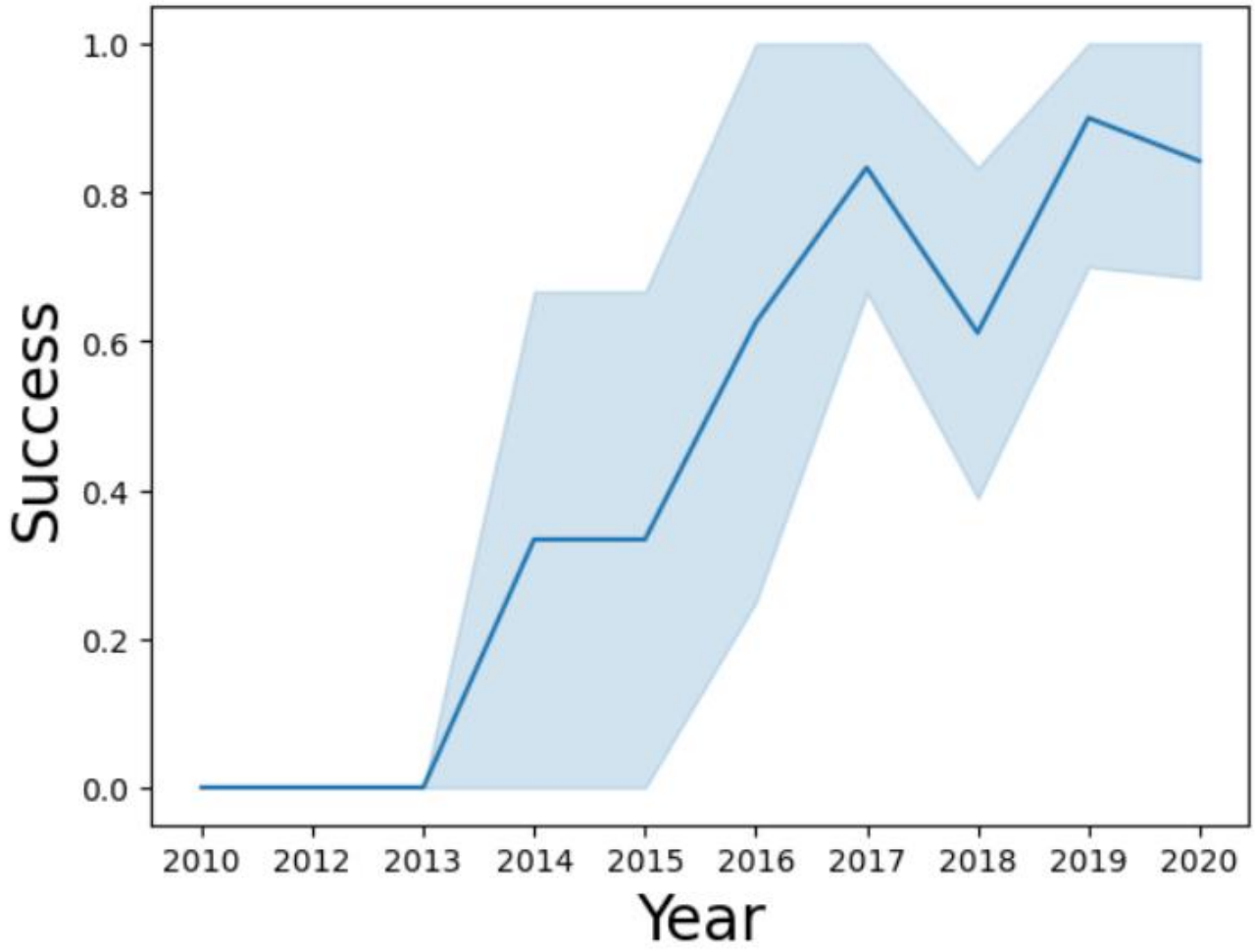
Orbit Types ES, LEO, VLEO has increased success rate with flight number where as others do not show any relationship

Payload vs. Orbit Type



Success rate increased for orbit type Polar,LEO and ISS as the payload increased

Launch Success Yearly Trend



you can observe that the sucess rate since 2013 kept increasing till 2020

All Launch Site Names

There are 4 launch sites

```
[10]: %sql select distinct Launch_Site from SPACEXTBL
 * sqlite:///my_data1.db
Done.
```

[10]:

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

Launch Site Names Begin with 'CCA'

Out of the 5 records 2 of them have failed outcome. 4 out of 5 launches were for NASA

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

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

[13]:

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 Brouere cheese	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

Total Payload Mass

SpaceX has space delivered 45596 kg payload for NASA (CRS)

```
[15]: %sql select sum(PAYLOAD_MASS__KG_) as 'Total Payload Mass' from SPACEXTBL where customer='NASA (CRS)'
 * sqlite:///my_data1.db
Done.
```

[15]:

Total Payload Mass
45596

Average Payload Mass by F9 v1.1

Average payload carried by booster version v1.1 is low to mid range.

```
[16]: %sql select avg(PAYLOAD_MASS__KG_) as 'Avg Payload Mass' from SPACEXTBL where Booster_Version='F9 v1.1'
 * sqlite:///my_data1.db
Done.
```

[16]:

Avg Payload Mass
2928.4

First Successful Ground Landing Date

```
%sql select min(date) as 'First Successful Landing'  from SPACEXTBL where Landing_Outcome = 'Success (ground pad)'
```

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

First Successful Landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

4 boosters were able to successfully land on drone ship with payload between 4000 and 6000

```
%sql select distinct Booster_Version as Boosters   from SPACEXTBL where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000
```

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

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

Total Number of Successful and Failure Mission Outcomes

There are total of 101 successful and failure mission outcomes

```
%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

- There are 9 booster versions that have carried maximum payload

```
%sql select Distinct Booster_Version as 'Booster Versions' from SPACEXTBL where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

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

Booster Versions
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

- In 2015 the first stage failed to land on drone ship in jan and april

```
%sql select  substr(Date, 6,2) as Month from SPACEXTBL where substr(Date,0,5)='2015' and Landing_Outcome = 'Failure (drone ship)'
```

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

Month
01
04

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

8 types of landing outcome has been recorded between the queried date

```
%sql select  Landing_Outcome , count(Landing_Outcome) as counts \
     from  SPACEXTBL  \
     where  Date between '2010-06-04'  and '2017-03-20' \
     group by Landing_Outcome \
     order by counts desc
```

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

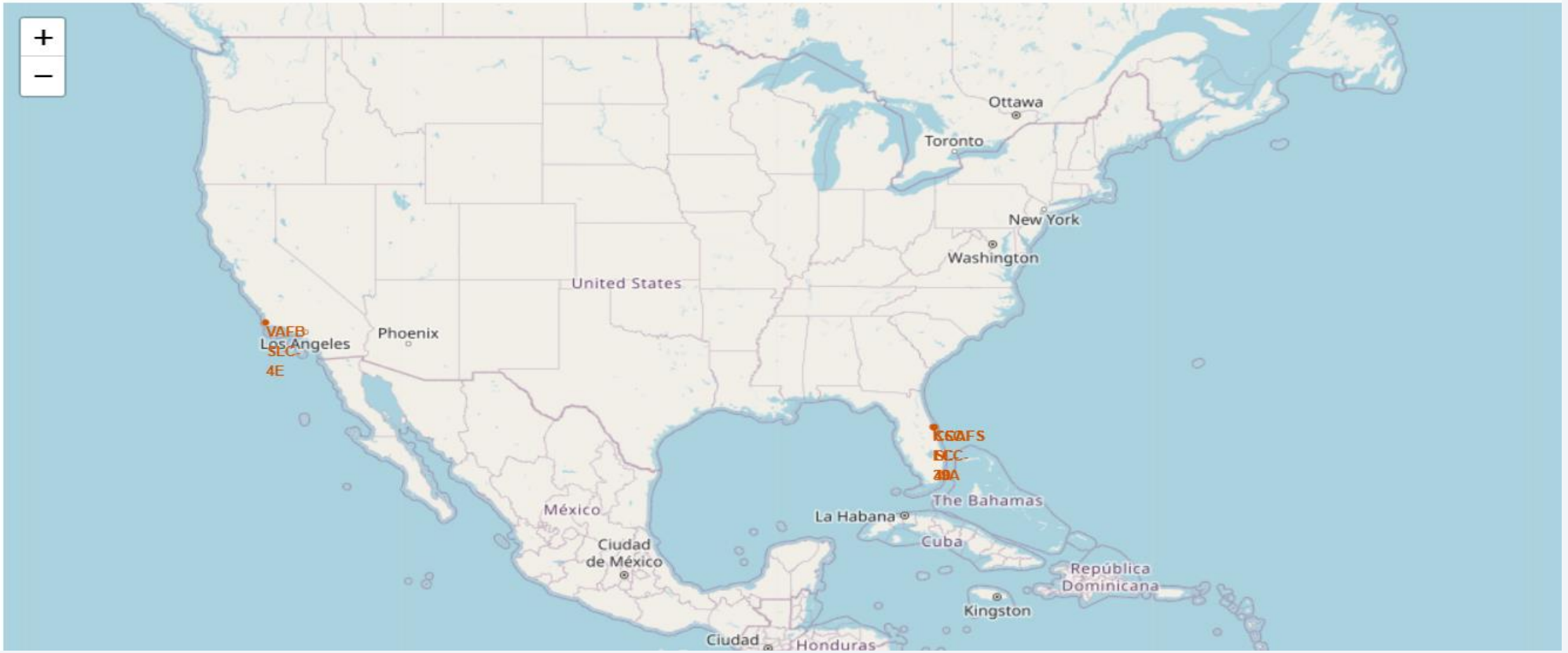
Landing_Outcome	counts
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1


Section 3
Launch Sites
Proximities Analysis

Map with all launch sites

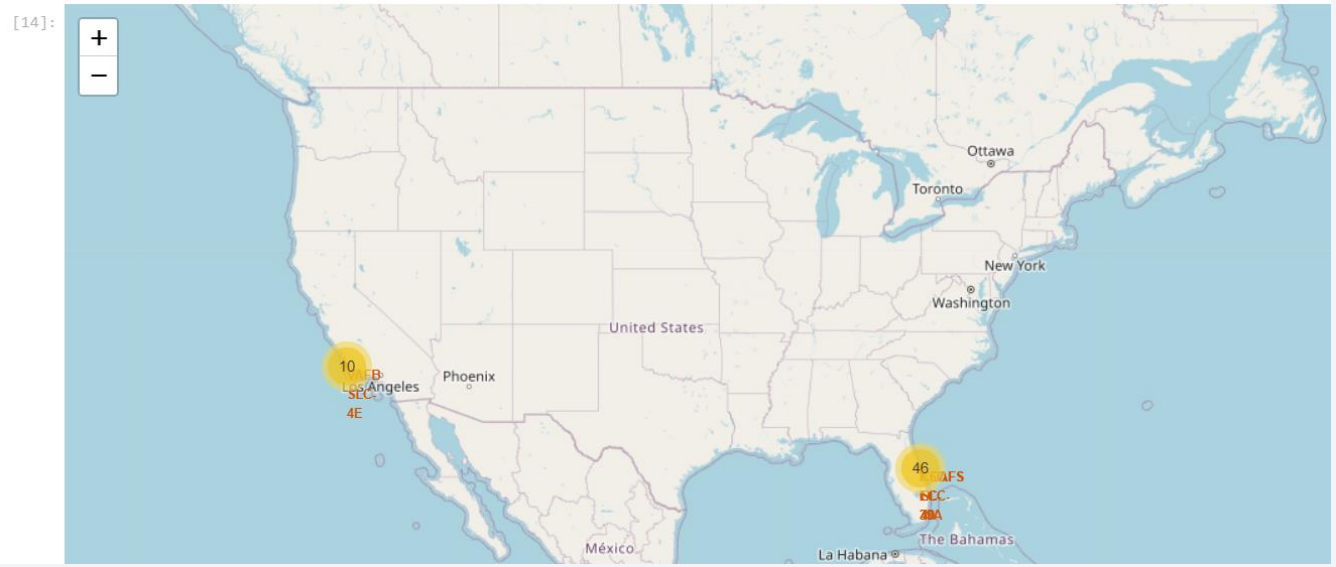
- All the launch sites are on costal line.
- There are 2 CCAFS sites in close proximity

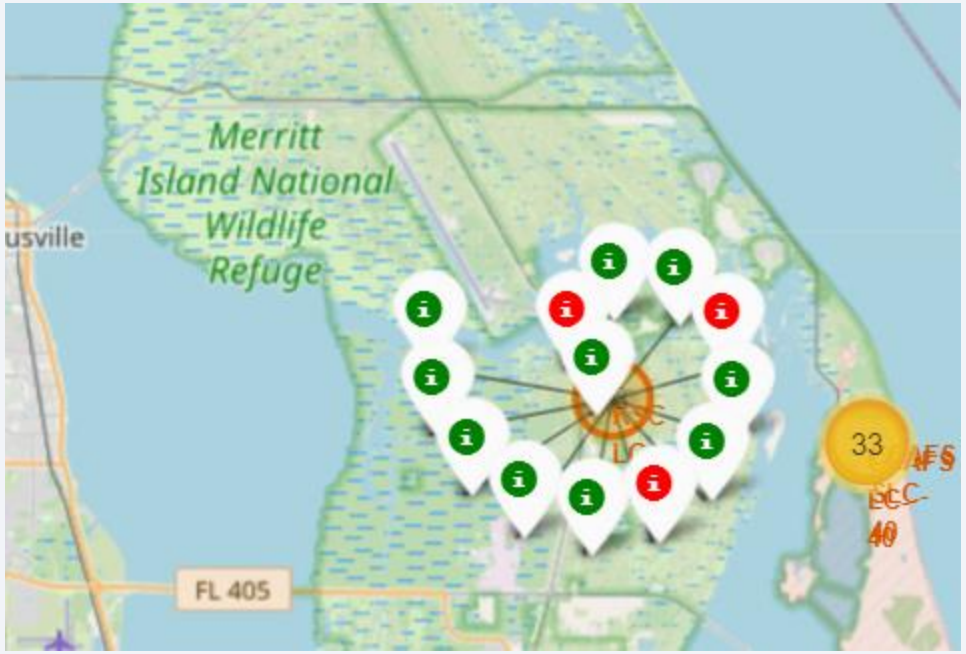
[9]:



Launch Outcome for each site

KSC LC-39A and VAFB SLC 4E has success rate of 77%

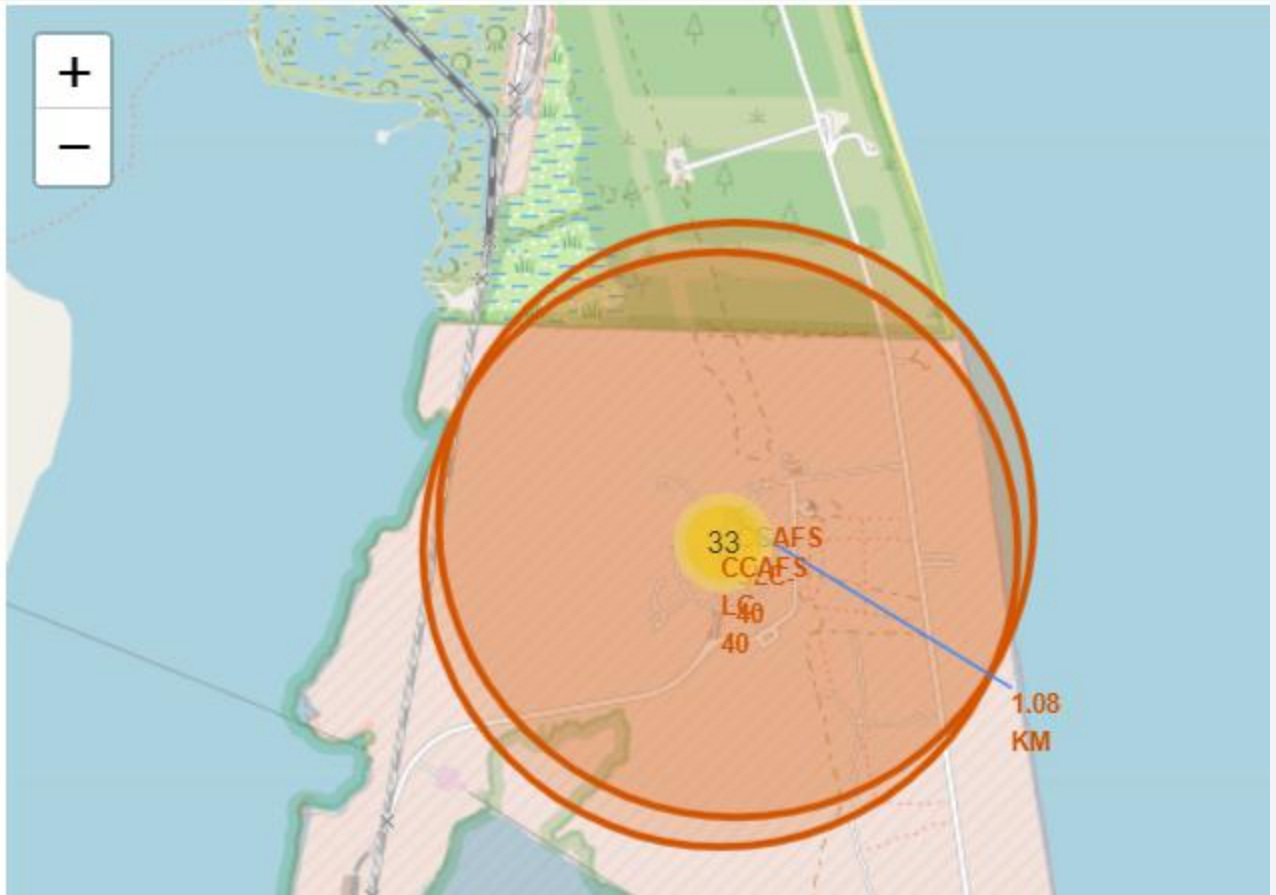


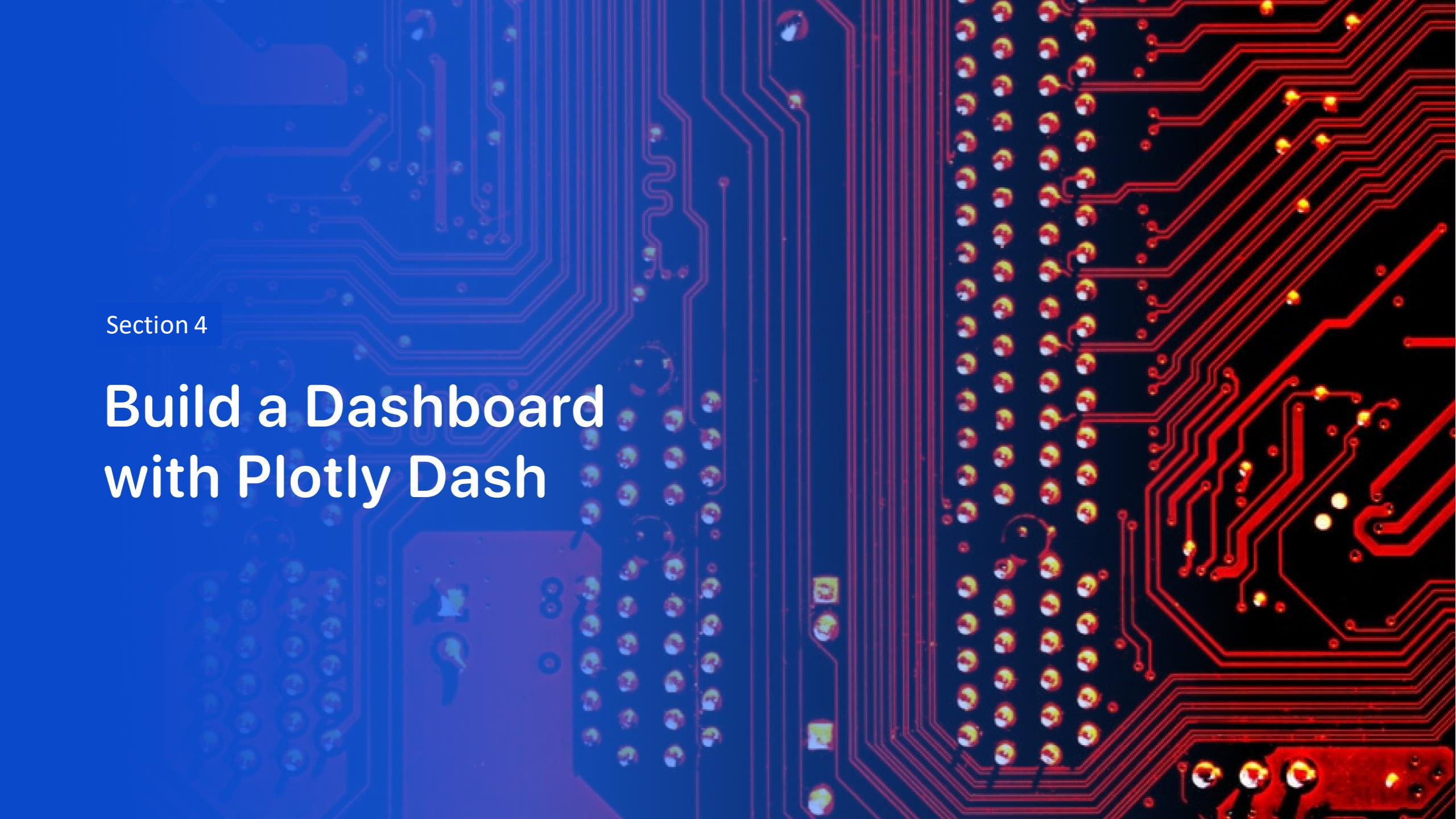


Distance from CCAFS-SLC-40 to the closest coastal line

- Distance from CCAFS-SLC-40 to coastal line is 1.08 KM

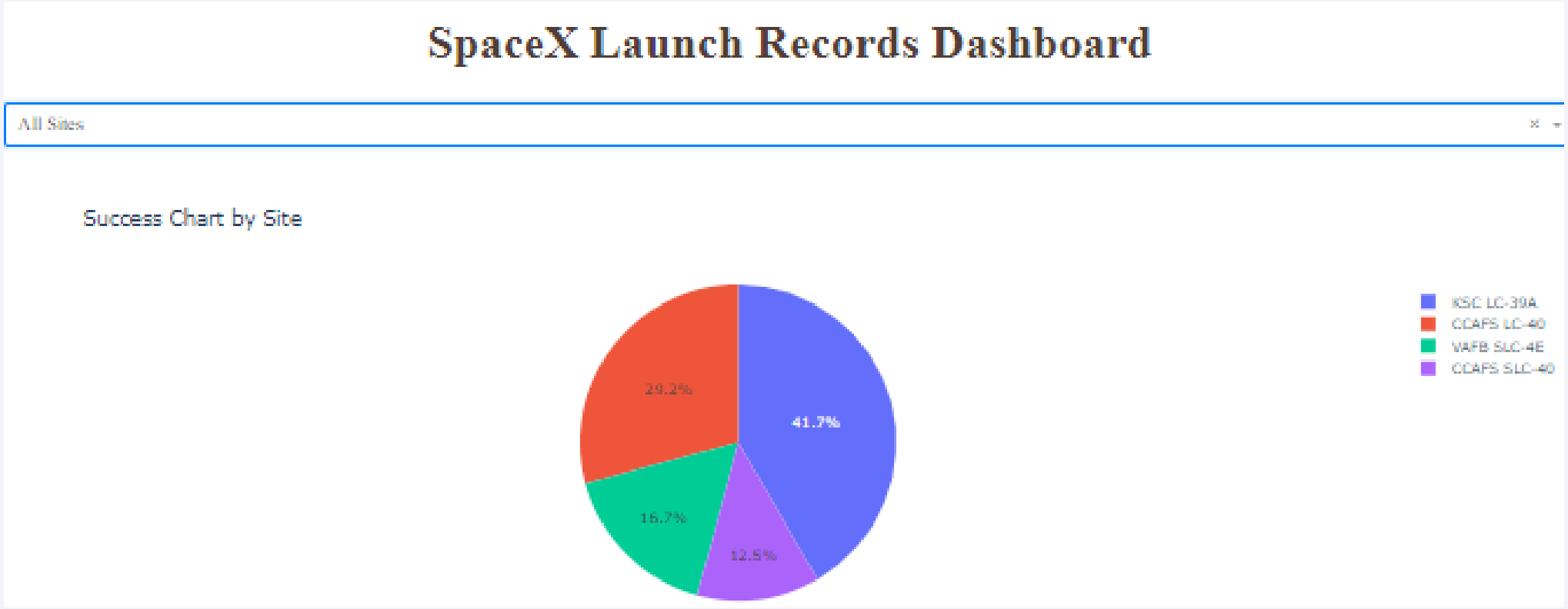
[25]:




Section 4
Build a Dashboard with Plotly Dash

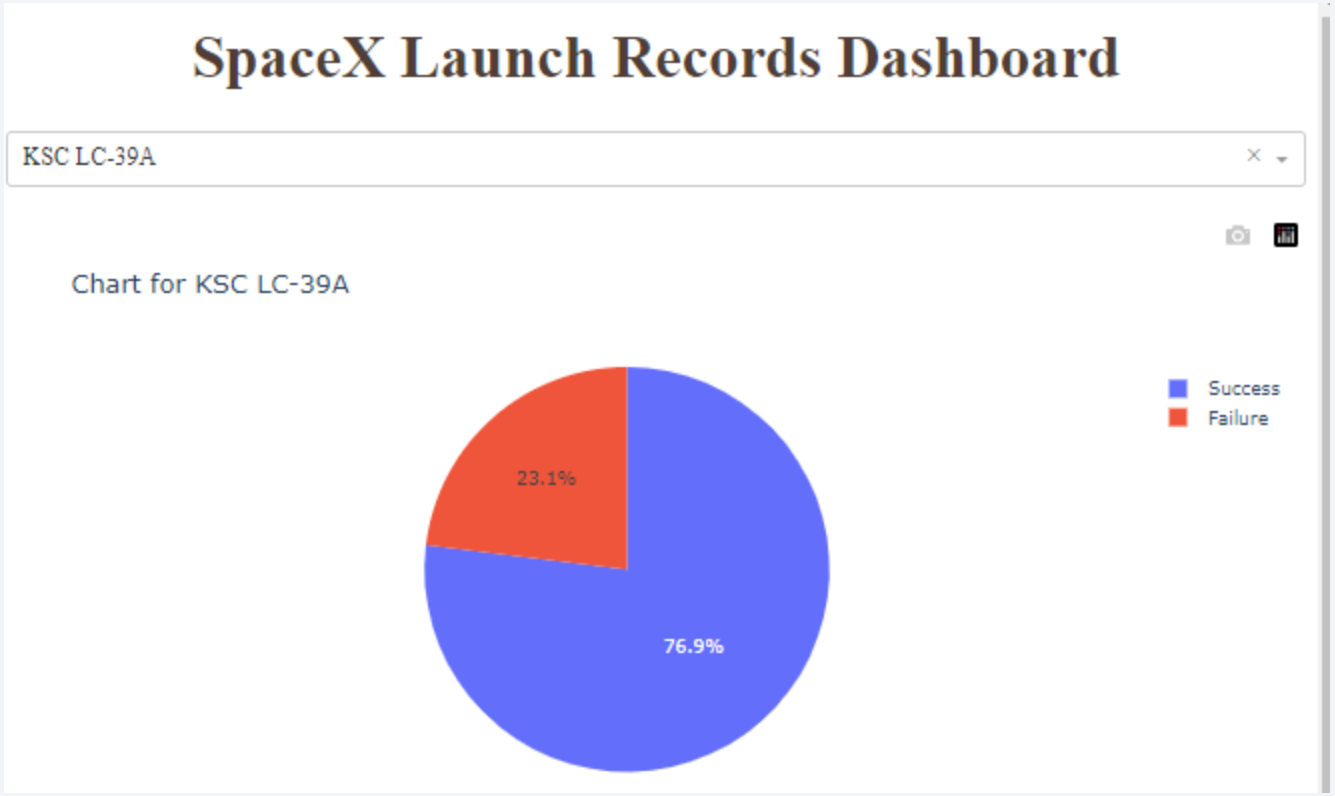
Pie Chart- Launch success for all sites

KSC LC-39A has the highest rate of success



Pie Chart- Launch Site with Highest Success rate

- KSC LC-39A has only 23.1% failure rate



Correlation between Payload and success for all sites

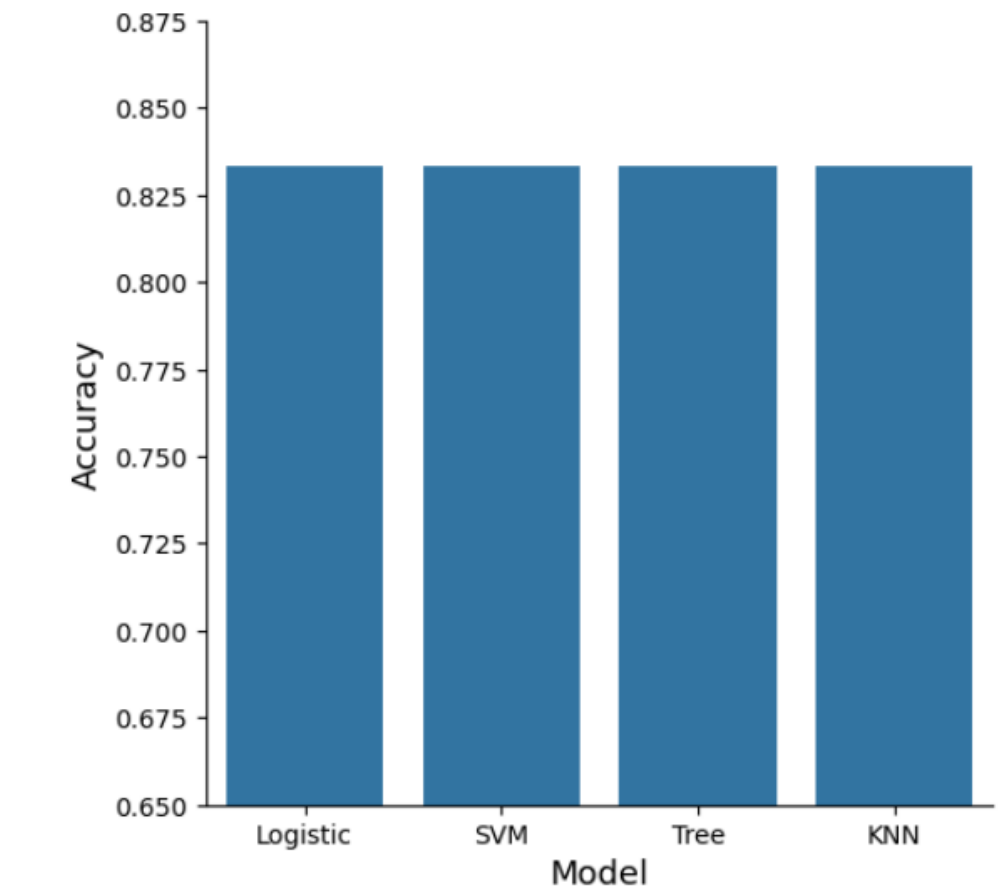
- Booster category V1.1 has higher failure rate and FT has higher success rate for payload between 1500 to 7500

Section 5

Predictive Analysis (Classification)

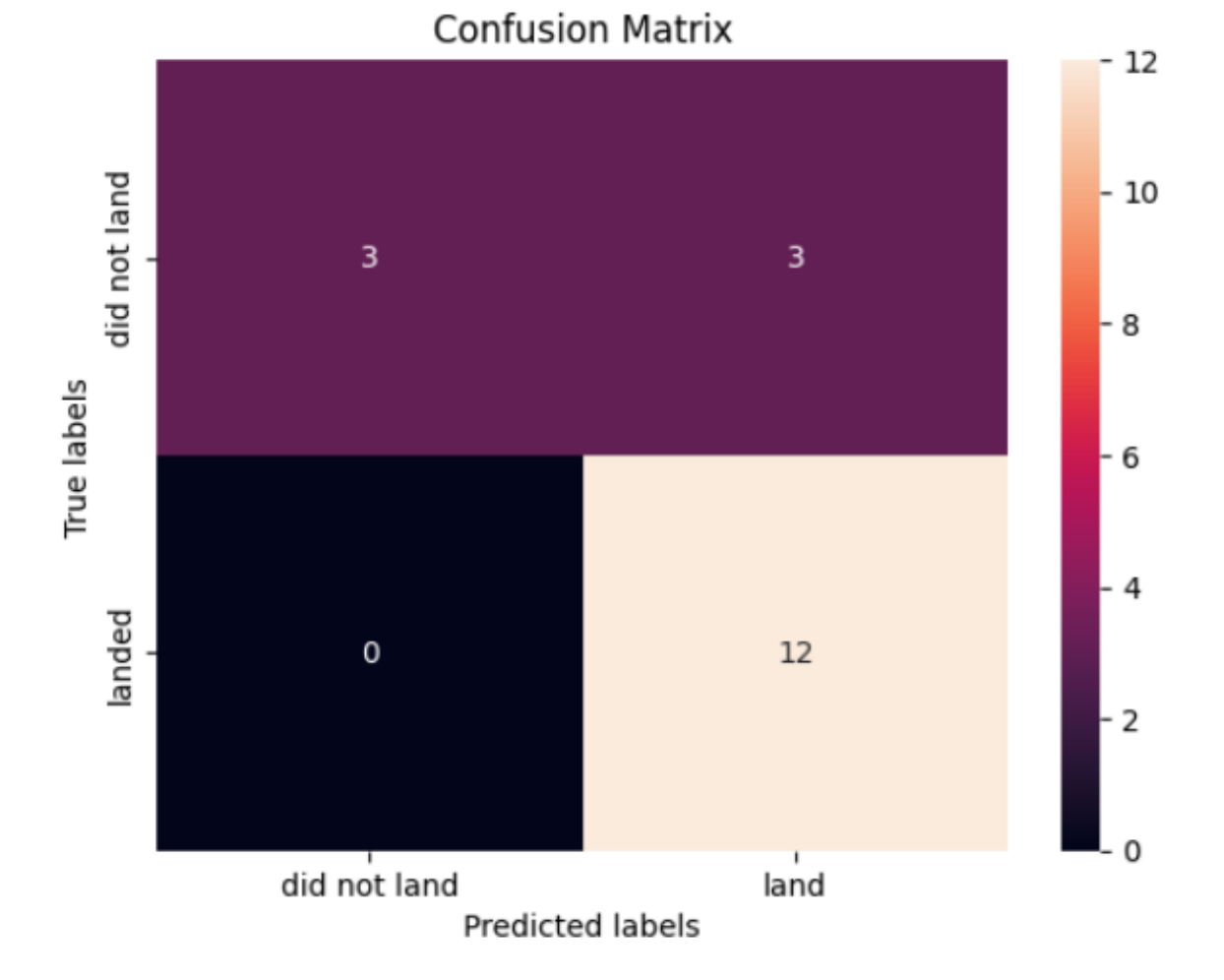
Classification Accuracy

- All the models have same accuracy of 83.33%



Confusion Matrix

- All the models have similar matrix where it can distinguish between different classes but has problem with false positives

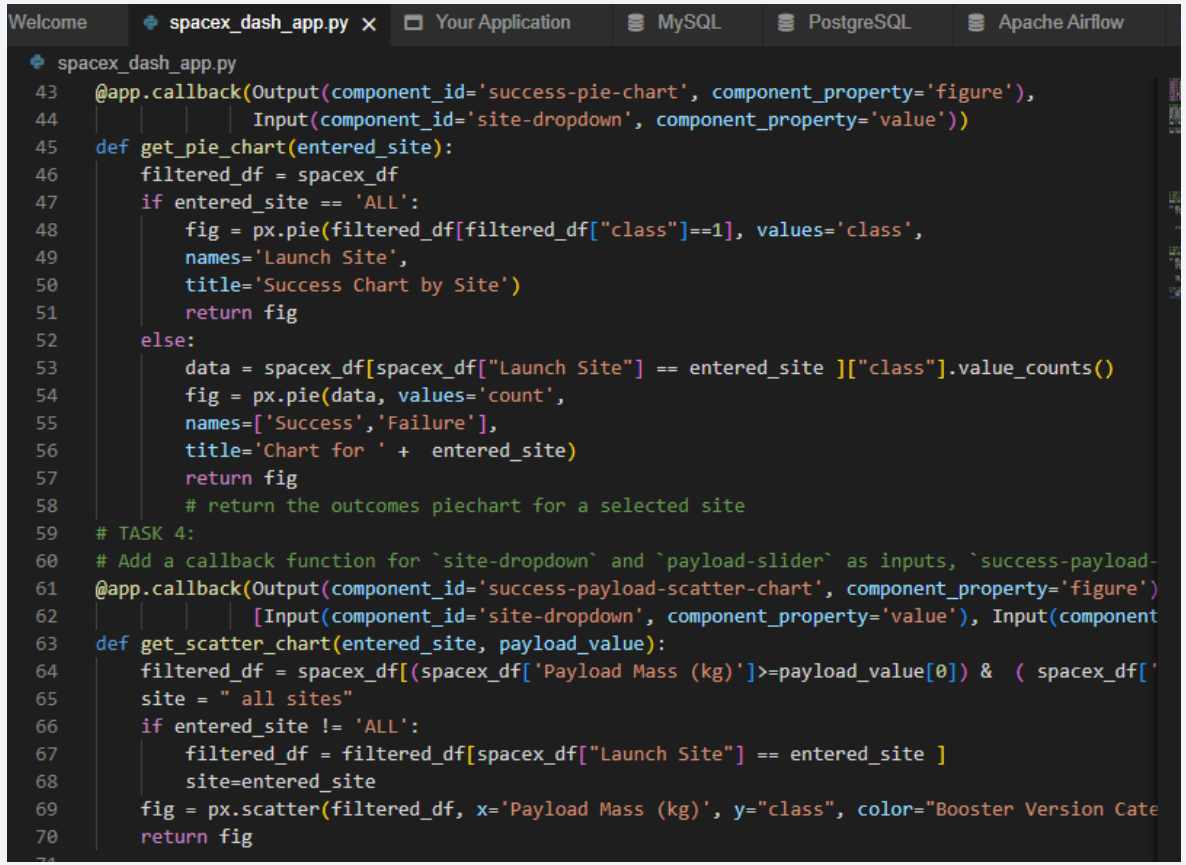


Conclusions

- Data for analysis is available in public and can be accessed easily
- Data analysis shows the first stage landing outcome is dependent on orbit type, flight number, payload mass and several other attributes
- Logistic, SVM, Tree and KNN all performed with same accuracy of 83.33% .
- Further tuning can be done by providing additional parameters to Gridsearchcv , Cv folds, test split percentage to get a model with better accuracy.
- This should be a iterative process starting with data collection to use latest data to model validation and tuning to achieve best prediction.

Appendix

All the project files can be accessed by clicking here



Thank you!