IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Usman Arif
12-Oct-2022

Outline

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

Executive Summary

Summary of methodologies

- Data was Collected from SpaceX API & Wikipedia Page
- Data wrangling was performed by dealing with null values & One Hot Encoding
- EDA was performed using SQL
- Data Visualization was created for EDA
- Interactive Visualization & Dashboard were created
- Predictive models were created for classification

Summary of all results

- Visualizations helped in understanding Data
- Optimal Predictive Model was obtained

Introduction

Project background and context

- Reusage of first stage of Falcon 9 rocket launches help SpaceX to launch rocket at a reduced cost of 62 million dollars each, compared to 165 million dollars price tag of conventional launch.

Problems you want to find answers

- Project task is to predict whether the first stage of SpaceX Falcon 9 rocket will land successfully or not.

Section 1

Methodology

Methodology

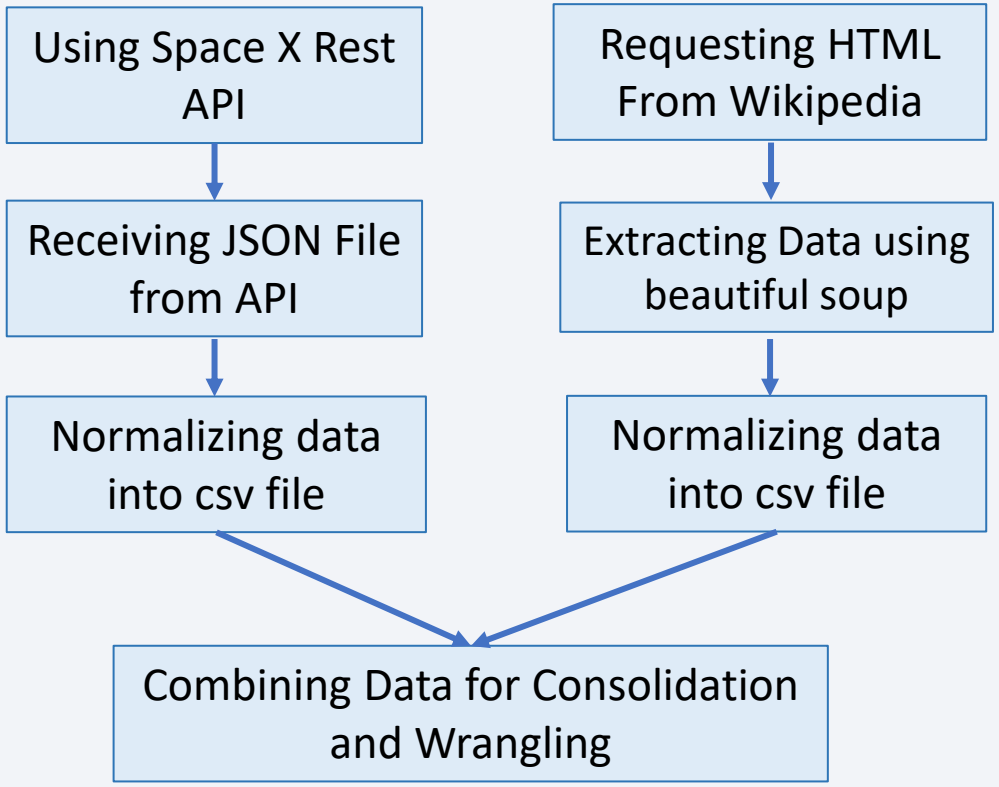
Executive Summary

- Data collection methodology:
  - Using SpaceX Rest API & Web Scrapping of Wikipedia
- Perform data wrangling
  - Cleaning of Null values & Irrelevant columns & One Hot Encoding of data fields
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Different classification models were built using optimum parameters found by Grid Search to find optimal model for classification.

Data Collection

Describe how data sets were collected.

- Data was collected using SpaceX API and Web scrapping. The process flow was as following



Data Collection – SpaceX API

1. Requesting Data from SpaceX API
2. Normalizing JSON File to Pandas Data Frame
3. Cleaning Data using Custom Functions
4. Extracting record for only Falcon 9 Rocket
5. Dealing with Null values & Exporting data As csv files

https://github.com/usmanarif1/Data-Science-Capstone-Project/blob/bbcf0ba7415607ce1e0c0d87fee40228dbb7b8b4/Space%20X%20data%20collection%20notebook.ipynb

Data Collection – SpaceX API

1. Requesting Data from SpaceX API
2. Normalizing JSON File to Pandas Data Frame
3. Cleaning Data using Custom Functions
4. Extracting record for only Falcon 9 Rocket
5. Dealing with Null values & Exporting data As csv files

https://github.com/usmanarif1/Data-Science-Capstone-Project/blob/bbcf0ba7415607ce1e0c0d87fee40228dbb7b8b4/Space%20X%20data%20collection%20notebook.ipynb

Data Wrangling

1. Calculating % of null values in each column
2. Calculating Number of launches on each site
3. Calculating number of occurrence of each orbit
4. Calculate number and occurrence of mission outcomes per orbit type
5. Create a landing outcome label

https://github.com/usmanarif1/Data-Science-Capstone-Project/blob/bbcf0ba7415607ce1e0c0d87fee40228dbb7b8b4/Space%20X%20data%20wrangling%20notebook.ipynb

EDA with Data Visualization

1. Categorial Plots to see relationship between Fligh Number, Payload Mass, Orbit & Launch Site
2. Bar Plot to observe relationship between orbits and success percentage
3. Scatter Plot to observe relationship between flight number and orbit
4. Line Plot to observe relationship between year and success Rate

https://github.com/usmanarif1/Data-Science-Capstone-Project/blob/bbcf0ba7415607ce1e0c0d87fee40228dbb7b8b4/EDA%20with%20Data%20Visualization.ipynb

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. Use a subquery
- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for 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

https://github.com/usmanarif1/Data-Science-Capstone-Project/blob/bbcf0ba7415607ce1e0c0d87fee40228dbb7b8b4/Exploratory%20Data%20Analysis%20using%20SQL.ipynb

Build an Interactive Map with Folium

- Marker, Circle, Icon, Polynomial Line and markercluster were placed on Map.
- These were used to mark launch sites on map, mark the success/ failed launches for each site on map and to calculate distances between launch site to its proximities
- https://github.com/usmanarif1/Data-Science-Capstone-Project/blob/bbcf0ba7415607ce1e0c0d87fee40228dbb7b8b4/Data%20visualization%20with%20folium.ipynb

Build a Dashboard with Plotly Dash

- **Two interactions were added:**
  Dropdown interaction for launch site & Slider interaction for Payload Mass
- **Two plots were added:**

1. A success pie Chart linked to dropdown interaction
2. A success payload scatter Chart linked to both interactions

- These plots and interactions were added to visualize relationship and to drive insights from payload mass and launch site data with success of the mission
- https://github.com/usmanarif1/Data-Science-Capstone-Project/blob/bbcf0ba7415607ce1e0c0d87fee40228dbb7b8b4/spacex_dash_app.py

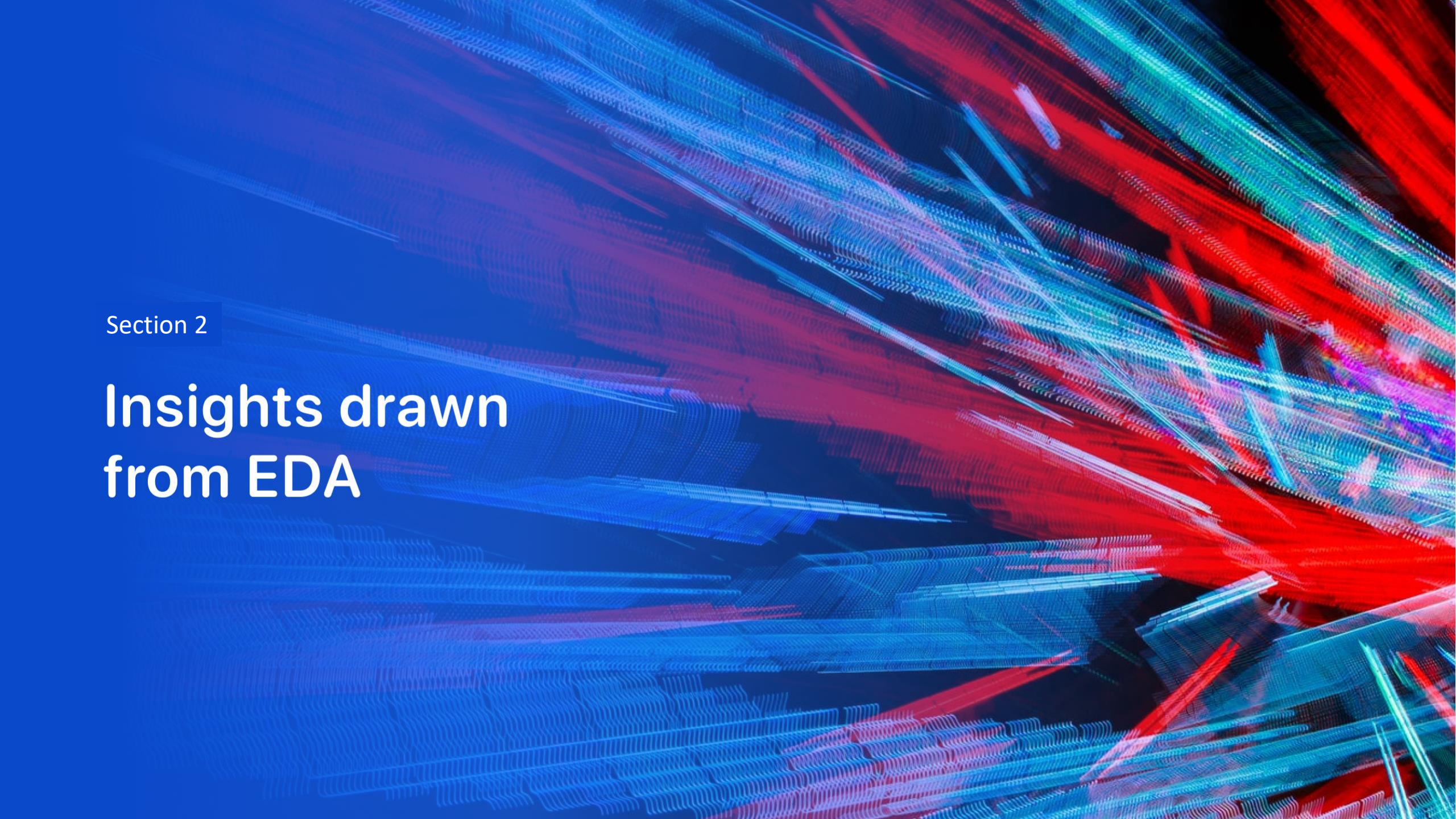
Predictive Analysis (Classification)

1. Created an array from the variable to be predicted ‘Class’.
2. Standardized the data set to be used for modeling
3. Train Test split of Data Set
4. Tuned Hyper Parameters for logistic regression, Decision tree, Support Vector Machines and K nearest Neighbor using Grid CV to achieve their maximum accuracy.
5. Predicted output variable using tuned models.
6. Plotted confusion matrix for the predicted values of different models.

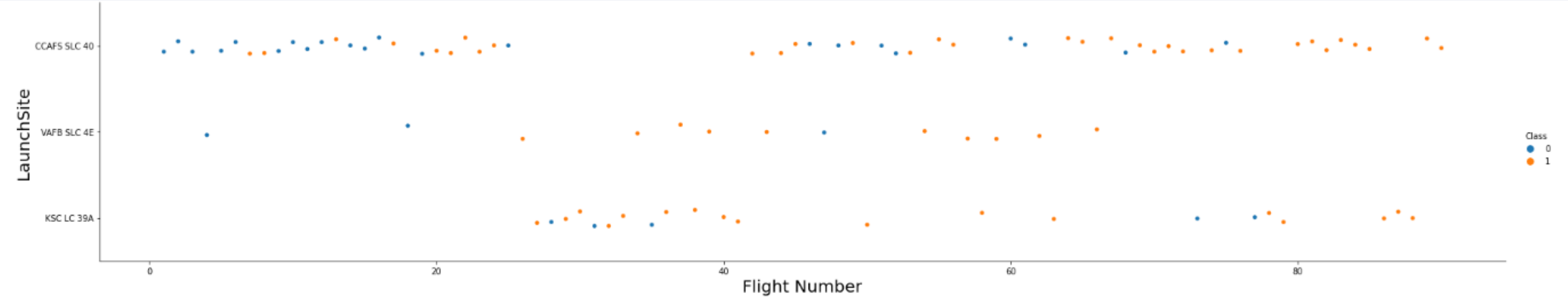
https://github.com/usmanarif1/Data-Science-Capstone-Project/blob/bbcf0ba7415607ce1e0c0d87fee40228dbb7b8b4/Machine%20Learning.ipynb

Results

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

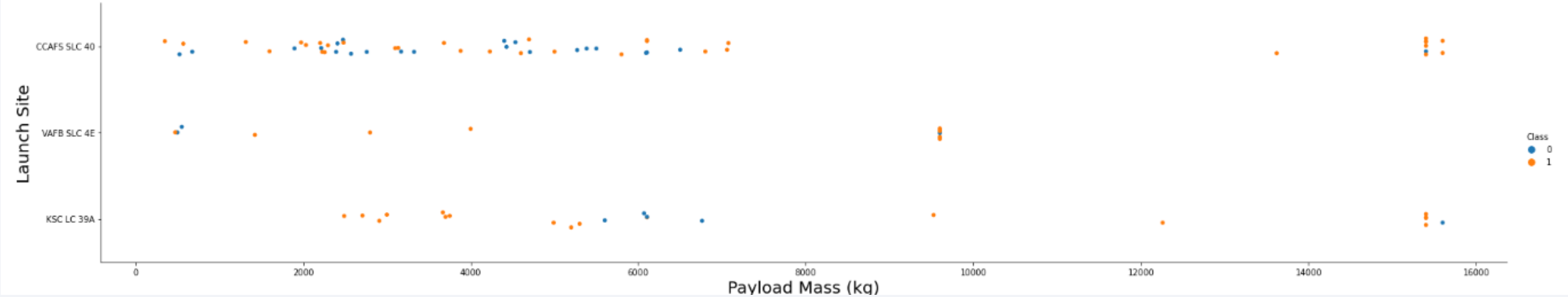

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



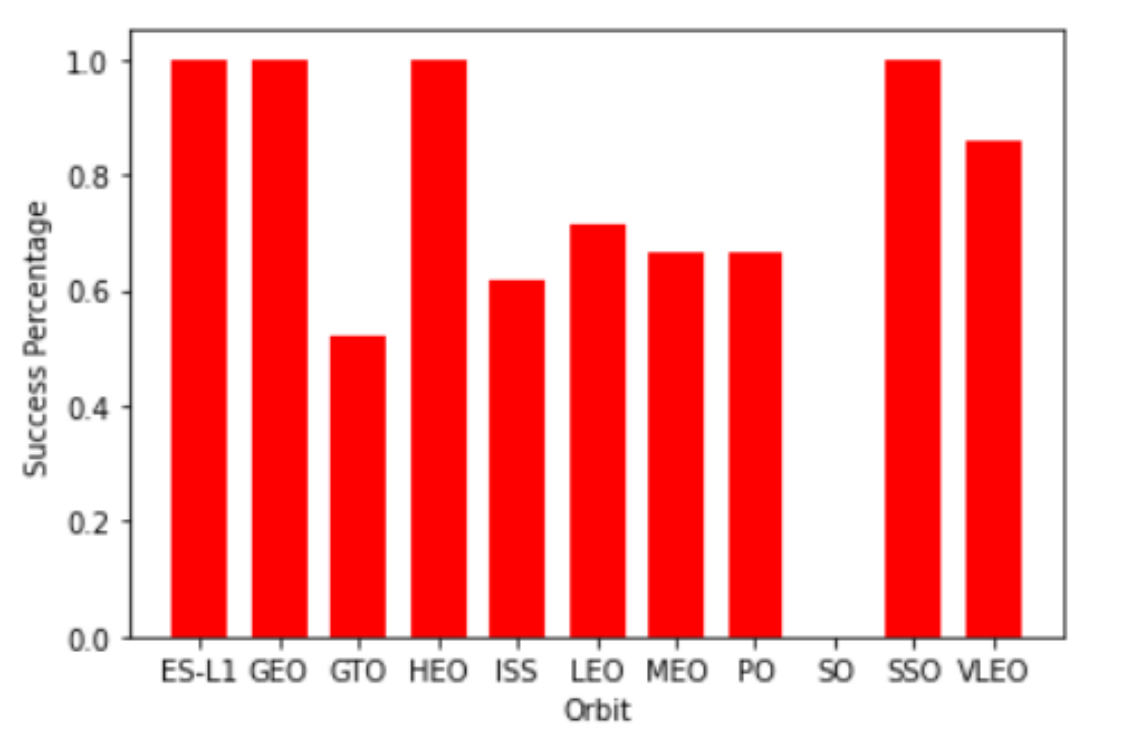
- This scatter plot shows that most of the flights happened from launch site of CCAFS SLC 40

Payload vs. Launch Site



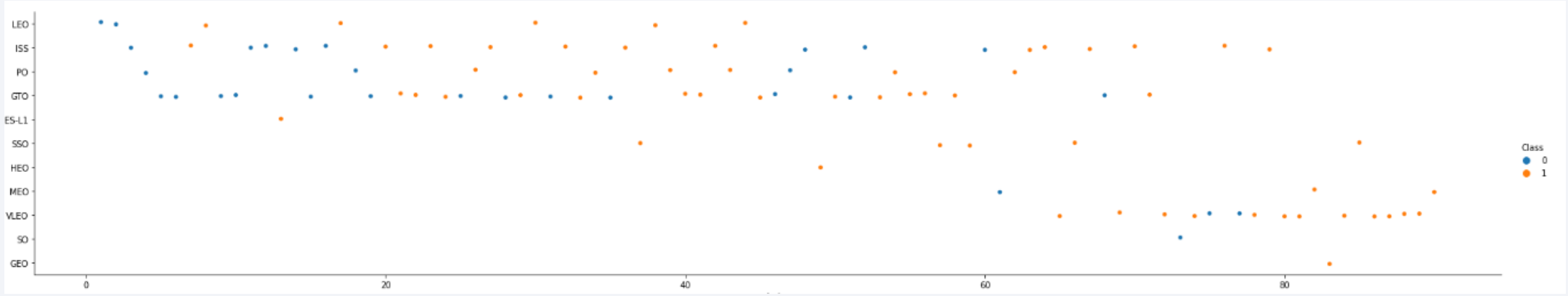
- This Scatterplot shows that for payload greater than 10000kg VAFB SLC 4E launch site is not used.

Success Rate vs. Orbit Type



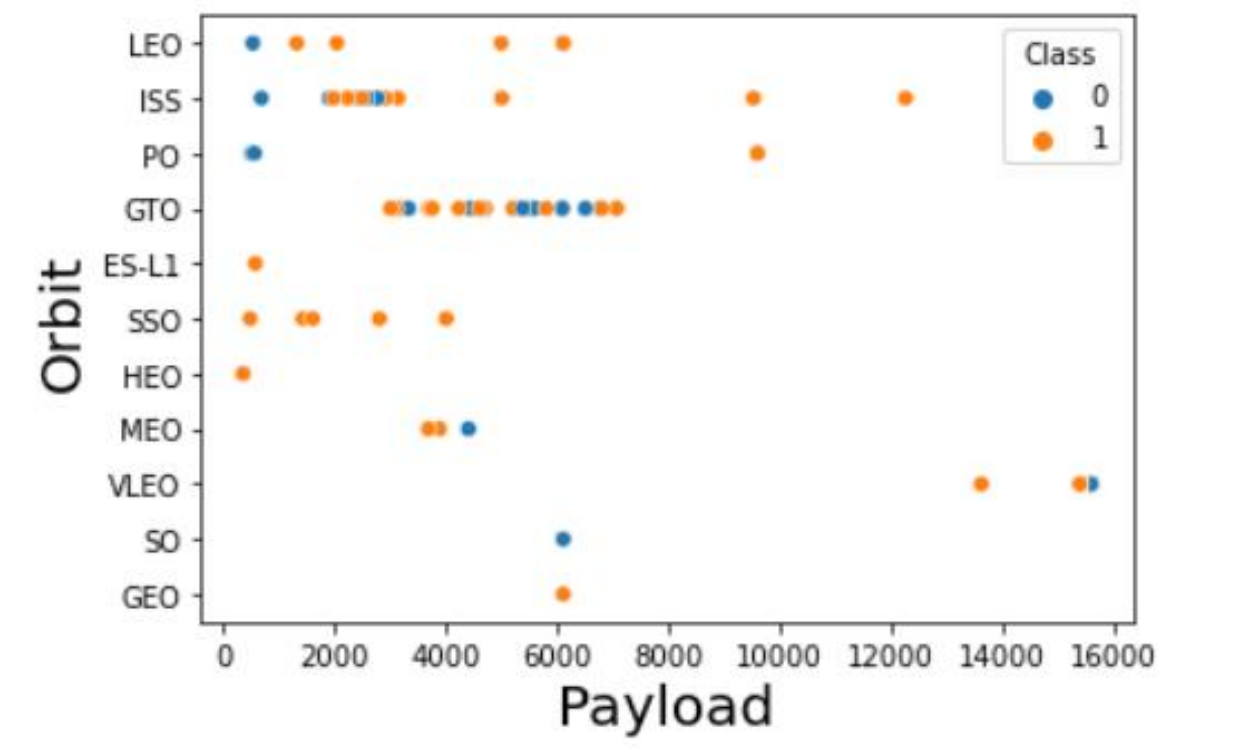
- This bar chart shows that for orbits: ES-L1, GEO, HEO & SSO success percentage of 100%

Flight Number vs. Orbit Type



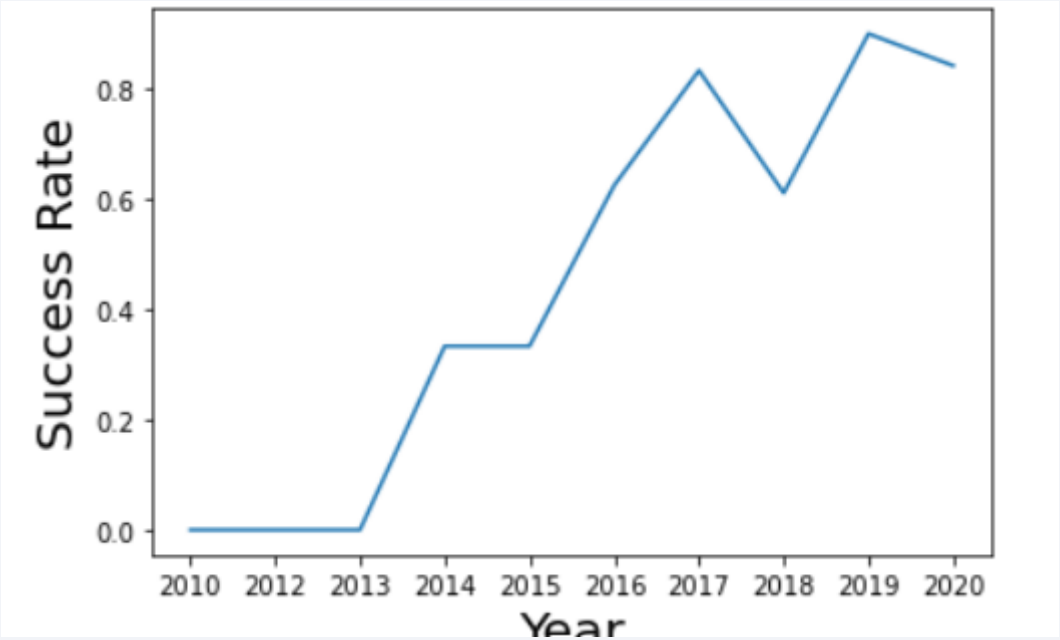
- Shows that later flight numbers are usually successfully launched in VLEO Orbit

Payload vs. Orbit Type



It shows that for higher payloads ISS Orbit or VLEO orbit are preffered

Launch Success Yearly Trend



- This line plot shows that the trend of successful landing of booster is rising every year

All Launch Site Names

```
In [8]: %sql SELECT Distinct LAUNCH_SITE FROM SPACEXTBL
```

Out[8]:

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

Launch Site Names Begin with 'CCA'

In [9]:
```
%sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5
```

Out[9]:

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	None	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	None	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	None	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	None	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	None	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

```
In [10]: %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE CUSTOMER='NASA (CRS)'
```

Out[10]:

1
45596

Average Payload Mass by F9 v1.1

Display average payload mass carried by booster version F9 v1.1

In [11]:
```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE BOOSTER_VERSION='F9 v1.1'
```

Out[11]:

1
2928.400000

First Successful Ground Landing Date

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

In [12]:
```
%sql SELECT min(DATE) FROM SPACEXTBL WHERE LANDING__OUTCOME='Success (ground pad)'
```

Out[12]:

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

In [13]:
```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ between 4000 and 6000 AND LANDING__OUTCOME='Success (drone ship)'
```

Out[13]:

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

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

In [14]:
```
%sql SELECT COUNT(*) FROM SPACEXTBL WHERE MISSION_OUTCOME LIKE '%Success%' OR MISSION_OUTCOME LIKE '%Failure%'
```

Out[14]:

1
101

Boosters Carried Maximum Payload

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

In [15]:
```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

Out[15]:

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

2015 Launch Records

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

In [16]:

```
%sql SELECT TO_CHAR(TO_DATE(MONTH("DATE"), 'MM'), 'MONTH') AS MONTH_NAME, \
    LANDING__OUTCOME AS LANDING__OUTCOME, \
    BOOSTER_VERSION AS BOOSTER_VERSION, \
    LAUNCH_SITE AS LAUNCH_SITE \
    FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Failure (drone ship)' AND "DATE" LIKE '%2015%'
```

Out[16]:

month_name	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

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

In [17]:

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

Out[17]:

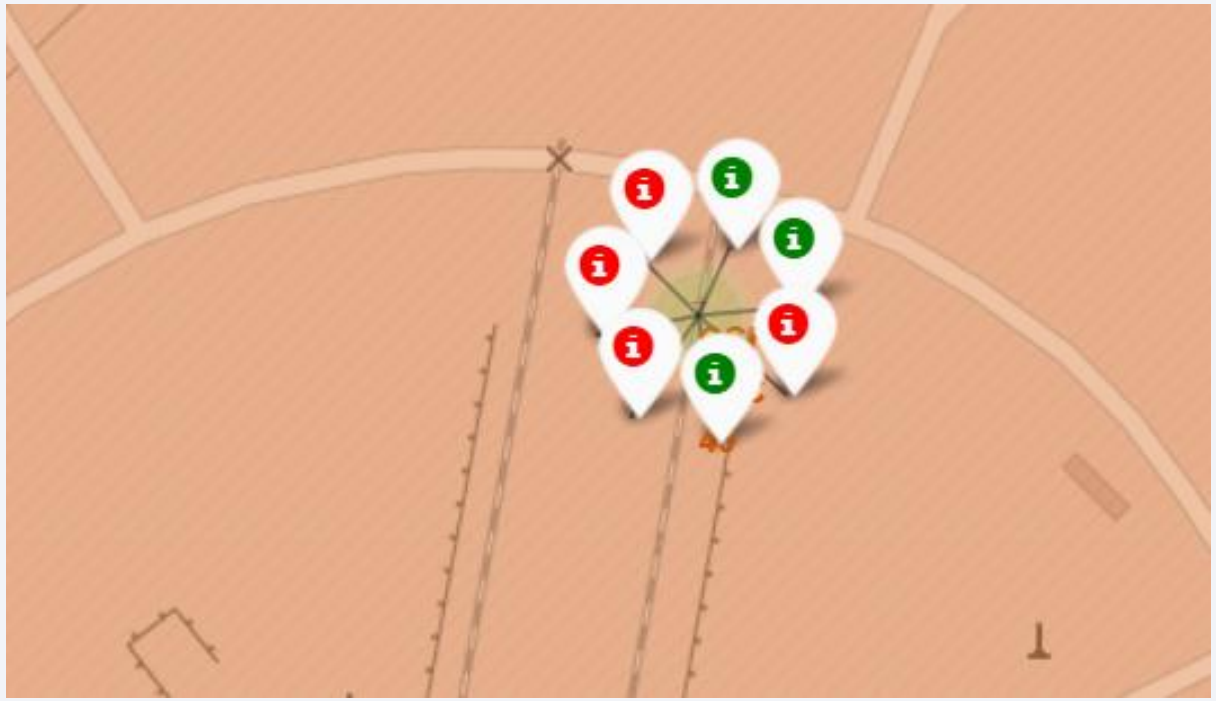
DATE	COUNT
2015-12-22	1
2016-04-08	1
2016-05-06	1
2016-05-27	1
2016-07-18	1
2016-08-14	1
2017-01-14	1
2017-02-19	1

Section 3

Launch Sites Proximities Analysis

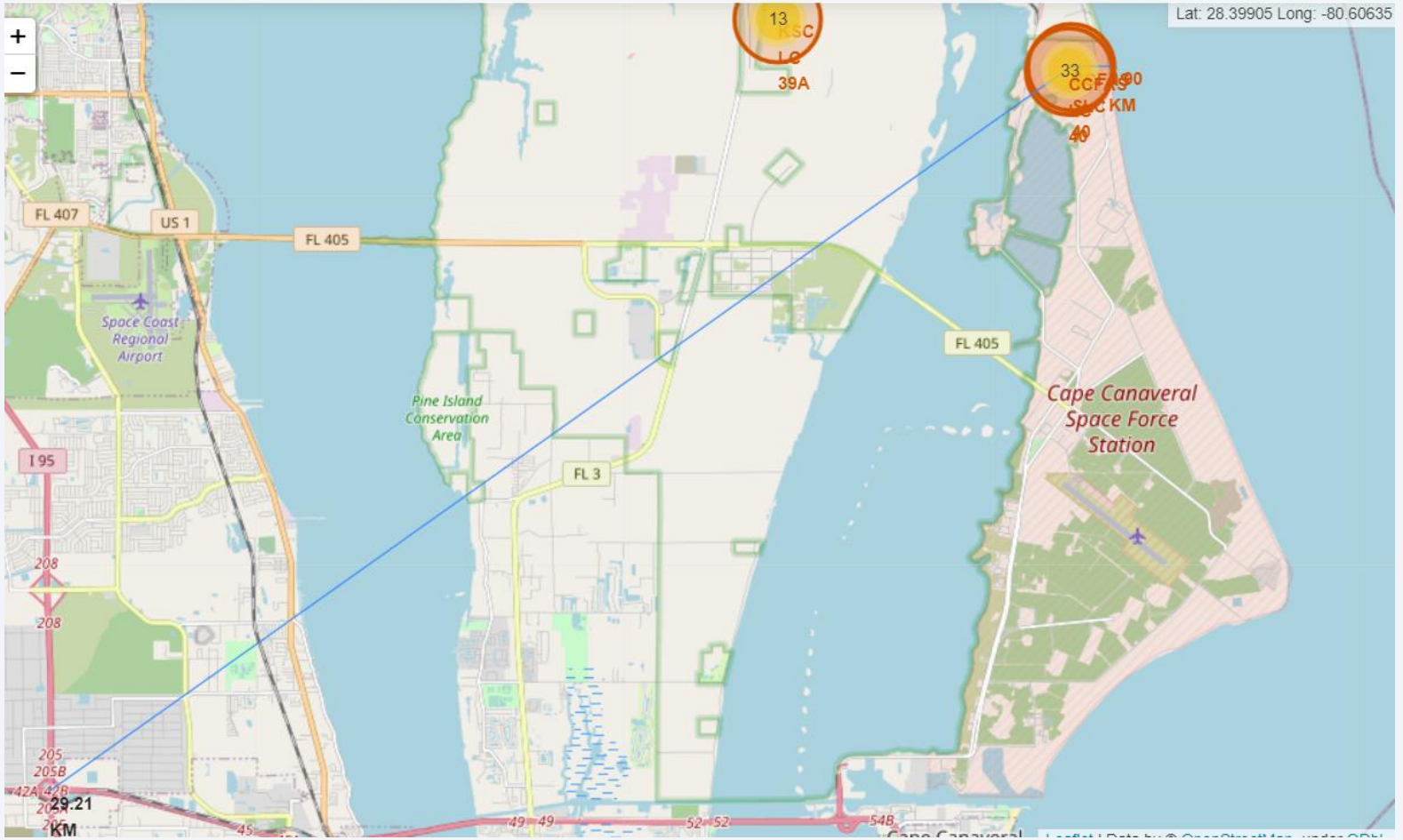
Launch Sites on Map

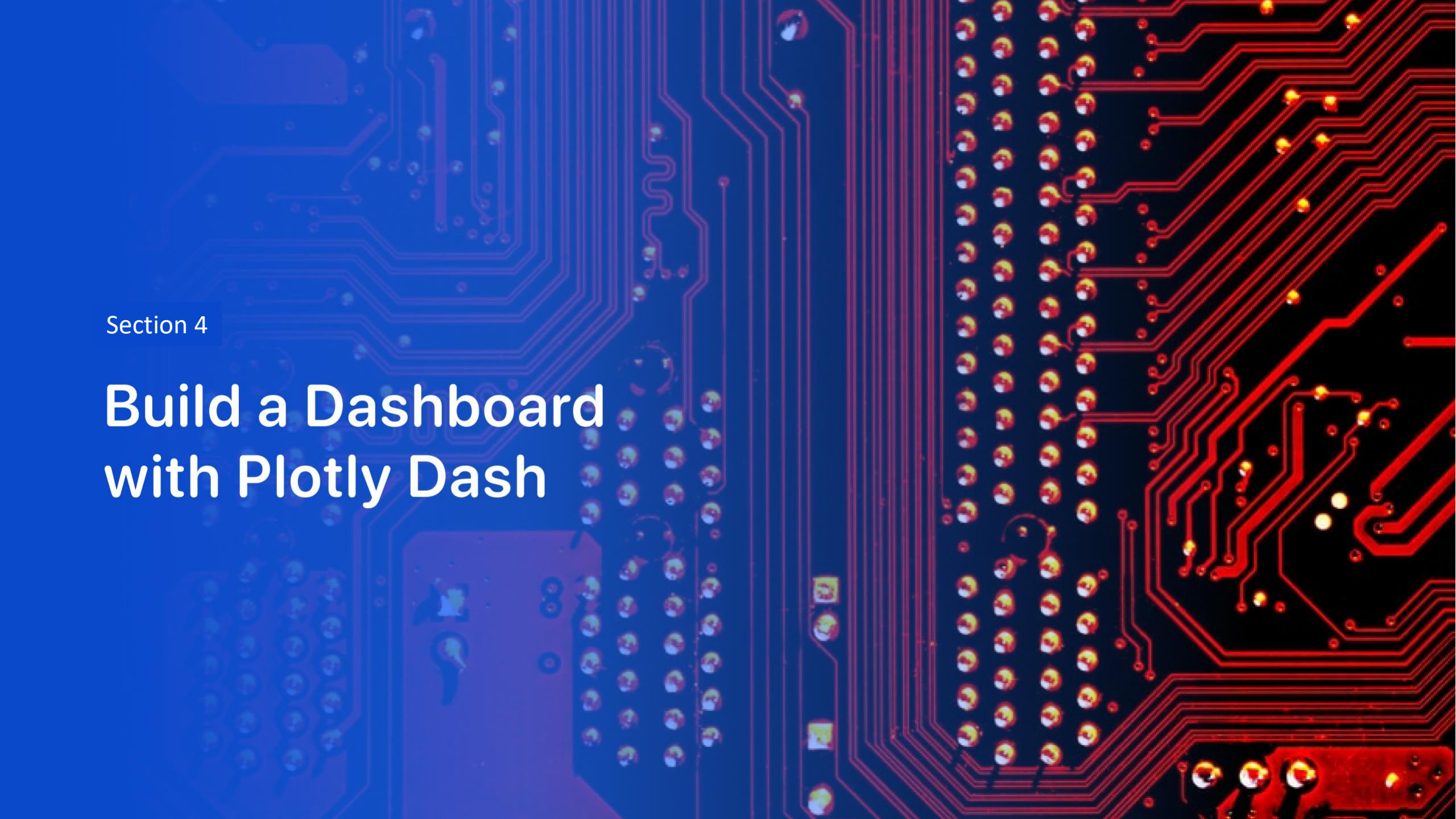
Outcomes from launch site



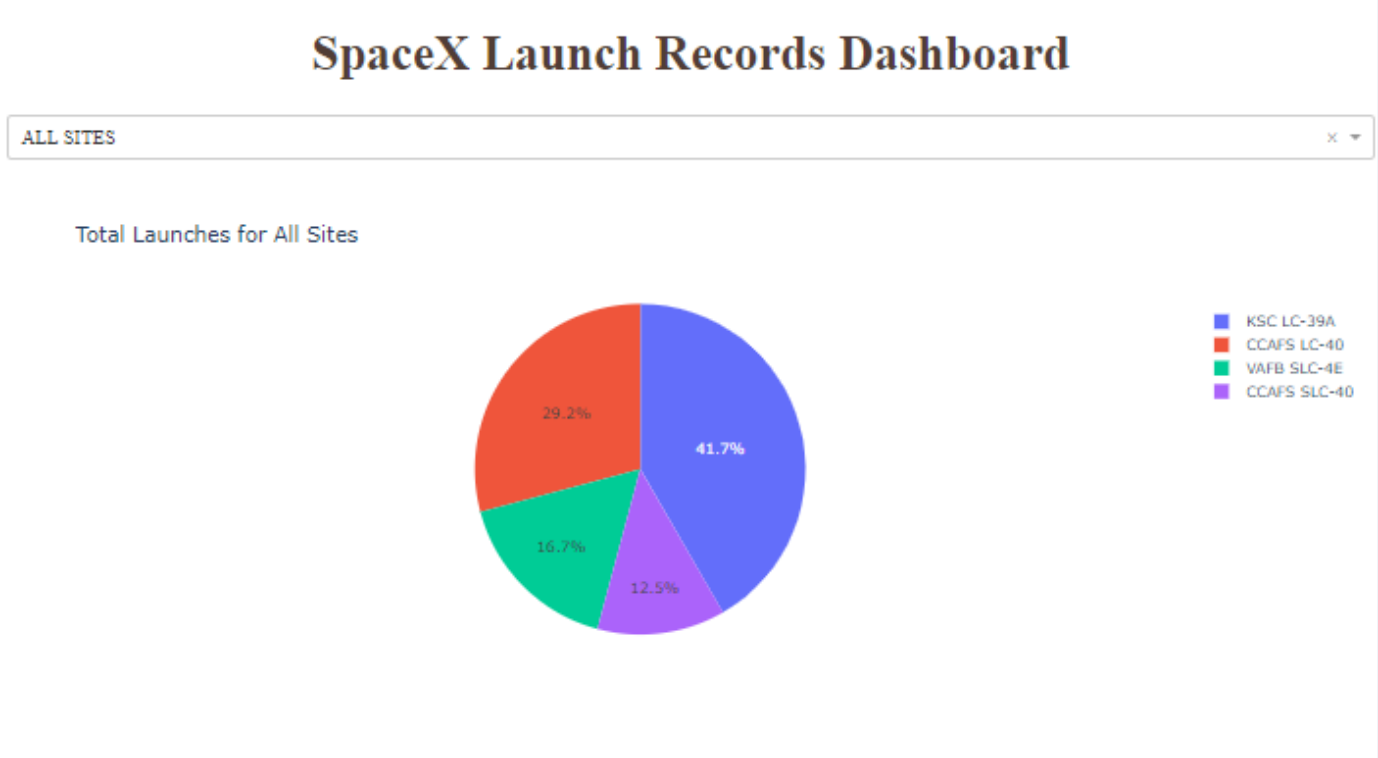
- Green shows successful outcomes

Proximity Distance from launch site



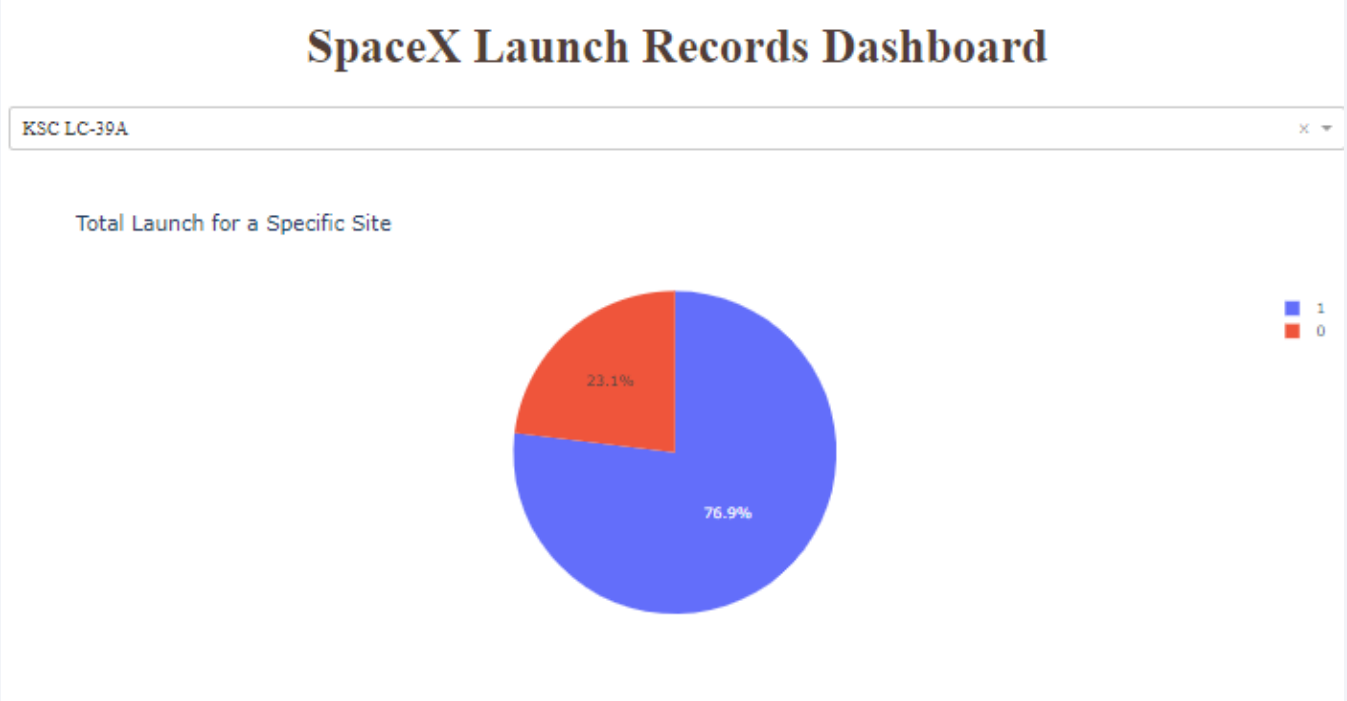

Section 4
Build a Dashboard with Plotly Dash

All Sites Pie Chart



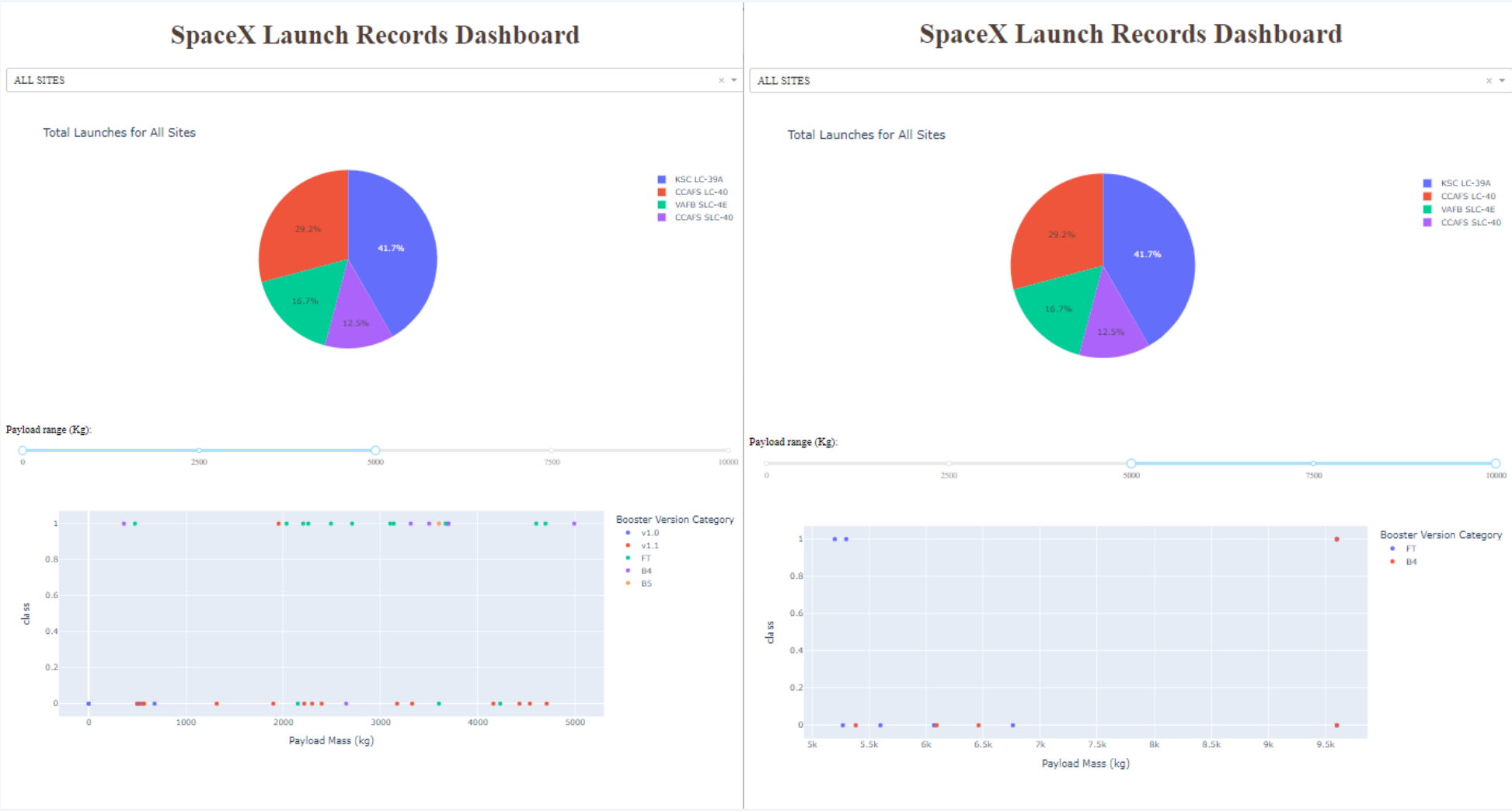
- Pie Chart shows percentage of launches from each launch site

Highest launch success ratio



- KSC LC-39A has highest launch success ratio with 76.9% as shown in pie chart

Payload vs Launch Outcome

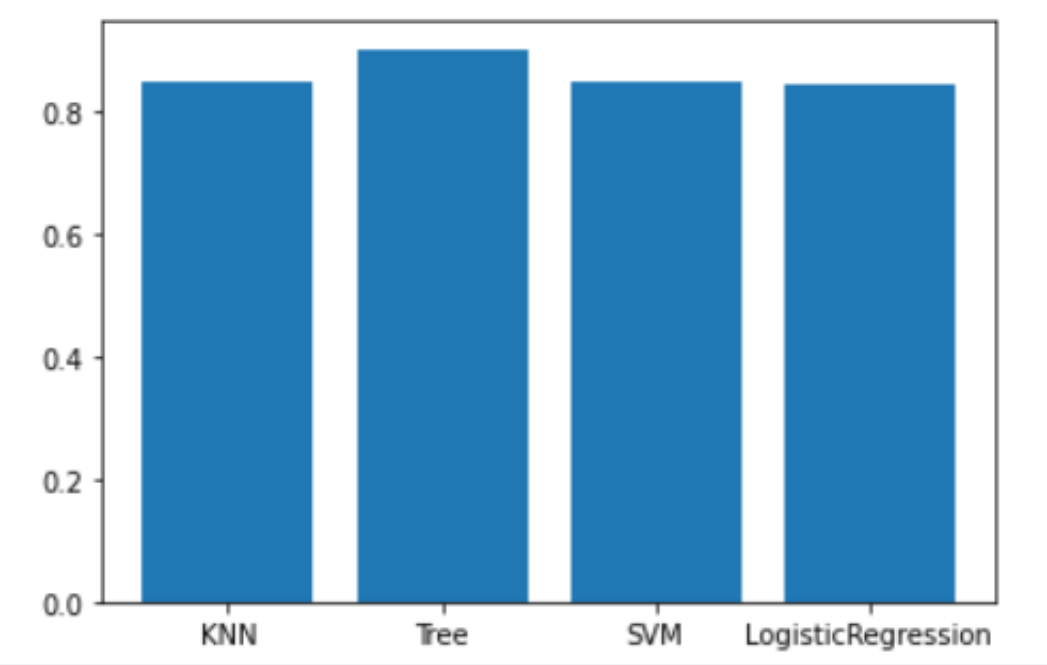


Section 5

Predictive Analysis (Classification)

Classification Accuracy

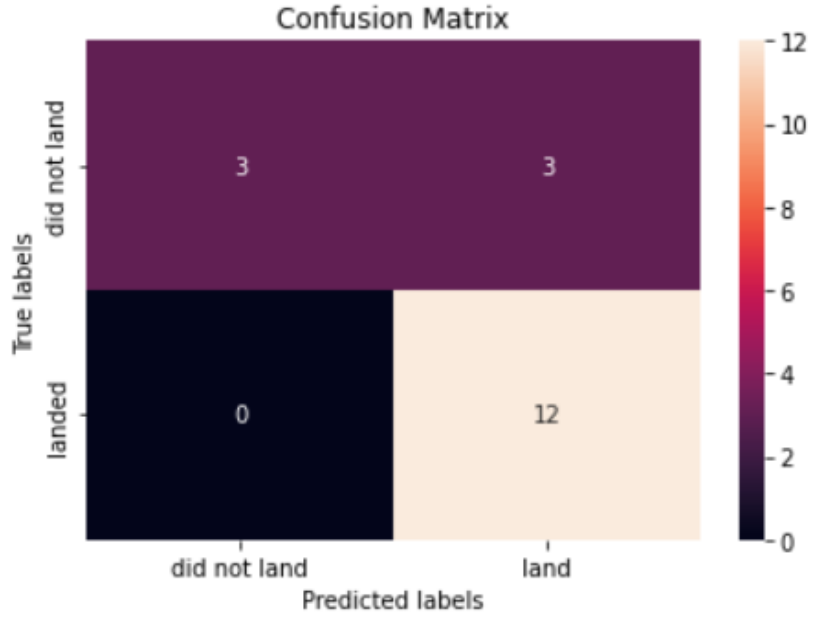
- The decision tree has the highest accuracy



Confusion Matrix

- Confusion matrix for decision tree

```
In [32]: yhat = tree_cv.predict(X_test)
         plot_confusion_matrix(Y_test,yhat)
```



Conclusions

- Decision Tree model has the highest accuracy of 0.90 approx.
- Lighter payloads perform better than the heavier payloads
- KSC LC-39A has highest success in launch
- Success rate is increasing annually which will eventually lead to perfection
- GEO, HEO , SSO & ES L1 orbits have highest accuracy.
- Distance of any point from launch site can be calculated using interactive map

Appendix

- Complete project link Github:
- https://github.com/usmanarif1/Data-Science-Capstone-Project.git

Thank you!