IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sum
6th June, 2023

Outline

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

Executive Summary

- Data is collected through API from Space x website and webscraping wiki page.
- Exploratory Data Analysis is done on data and output landing output column is binarized in data wrangling.
- Various plots are plotted to find any correlations among attributes including payload mass, orbit, flight number and landing outcomes in data visualization.
- It was observed that success rate of Falcon 9 kept increasing since 2013.
- The launching sites are plotted on world map using Folium map.
- It was found that launching sites are located near coastline.
- The dash application is created to visualize success rate among launching sites using pie chart.
- It is also created for visualize relation among payload mass and success rate.
- Various machine learning classification models are trained on these dataset.
- It was found that K Nearest Neighbor performs better than other algorithms including Logistic regression, SVM and decision tree.

Introduction

- Space x advertises their Falcon-9 rocket with low cost.
- It’s because of it reuses first stage.
- Hence, if landing outcome can be predicted, the cost of launch can be predicted.
- Our goal is to determine whether launch will be successful or not by examining previous launch data.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data was collected from space x website through REST API’s and webscraping wiki pages.
- Perform data wrangling
  - Data was processed using Python
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Various models were build using data

Data Collection

- Data was collected by two ways.
  1. REST API
  2. Webscraping

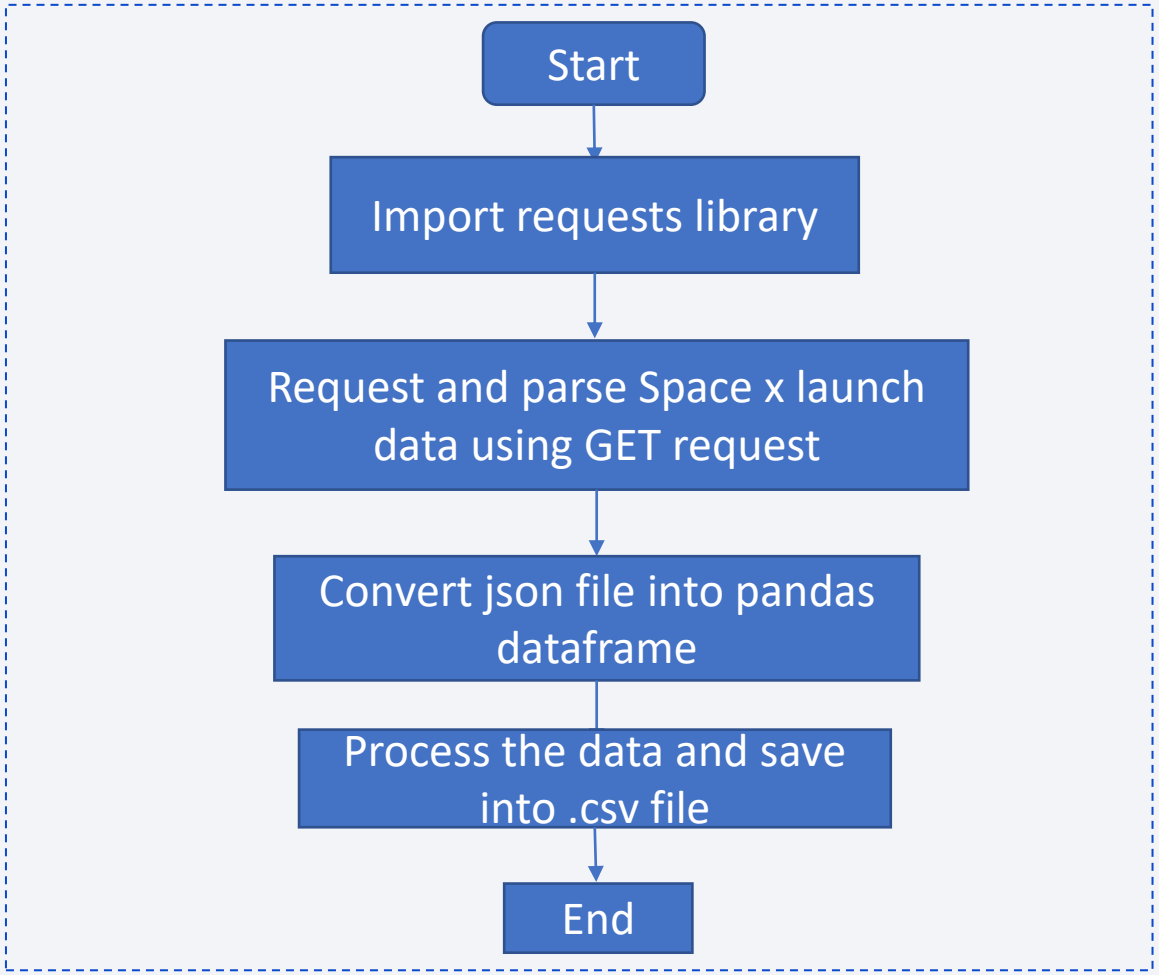
- From space x website, data was collected using REST API.

- Falcon-9 launch data was collected by webscraping wiki pages.

Data Collection – SpaceX API

- Using REST API calls, Space X launch data is obtained.
- The launch data comprises data of Falcon 1 and 9 regarding outcome of their launches.
- This data is further filtered to include only Falcon 9 launches.
- Reference:

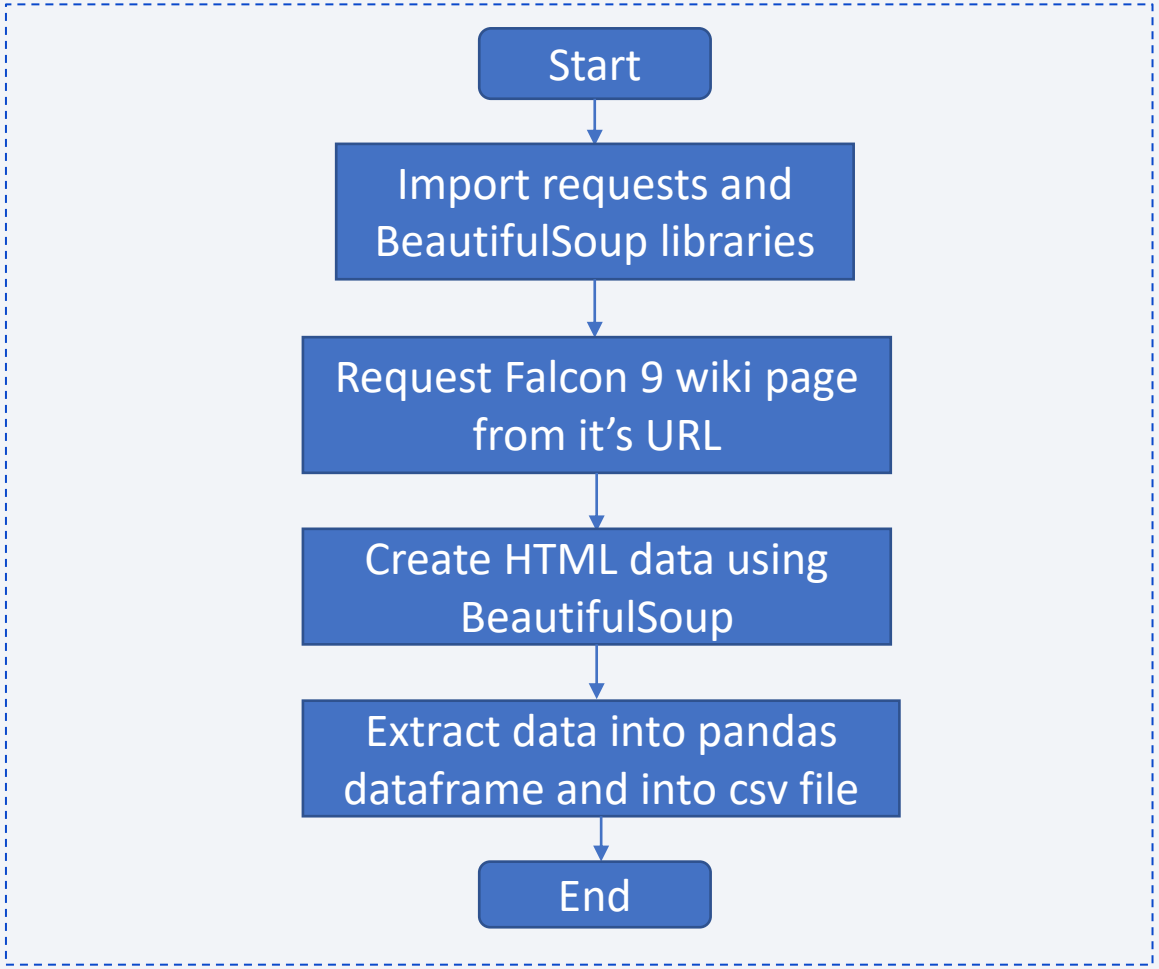
[Data Collection API notebook]



Data Collection - Scraping

- Falcon 9 Launches records obtained by webscraping wiki pages
- Comprises of data regarding Falcon 9 launch outcomes, payload, launch sites and so on.
- Reference:

Data Collection using webscraping



Data Wrangling

- Performed Exploratory Data Analysis (EDA) to find pattern in data sing pandas dataframe along with data preparation

- **Three launching sites** are identified
  - **CCAFS SLC 40** occurred most frequently.

- Number of occurrences of each **orbit** is calculated.
  - Out of them, **GTO** orbit occurred most frequently.

- **Class** column is created to binarize launch outcomes.
- Reference : Data Wrangling Notebook

EDA with Data Visualization

- Scatterplot between Payload and Flight number to identify relationship between Flight number, Payload and Class
- Cat plot between Launch site and Flight number to find pattern in class
- Cat plot between Launch site and Payload to find pattern
- Bar plot of each orbit to find success rate of each orbits
- Line plot to identify trends in success in each year
- Reference : EDA data visualization notebook

EDA with SQL

- For EDA with SQL, sqlite3 is used. Dataset is loaded into the database.
- Distinct **Launch sites** are identified using SQL query.
- Total and average **payload** is calculated.
- **Booster versions** are identified for specific range of payload and successful landing in drones.
- Total counts of success and failures are identified.
- All the successful launch records for year 2015 are identified.
- Reference : EDA using SQL notebook

Build an Interactive Map with Folium

- Visualize geospatial data with Folium
- Marking of all launch sites on map by creating circle and marker objects add adding those objects on map
- Marking success and failed launches for each site on map using MarkerCluster object and adding on map
- Adding distance calculated from launch site to proximities and adding to map
- Reference : Map with folium notebook

Build a Dashboard with Plotly Dash

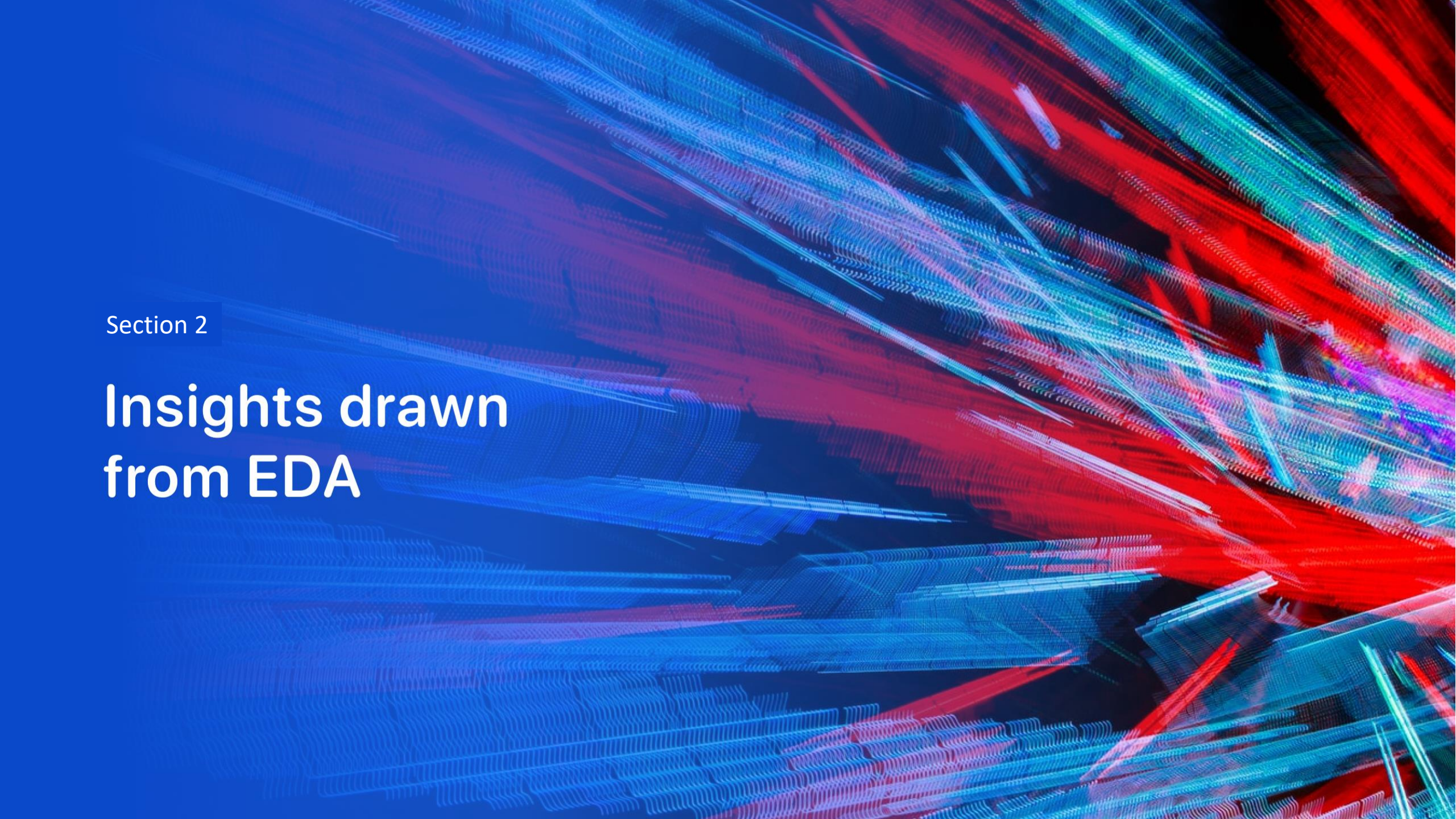
- Plotly Dash application to perform interactive visual analytics in real time
- Addition of dropdown input components to get input feeds from user about launch site and payload
- Addition of pie chart component to visualize success rate of launches for user defined launching site
- Addition of range slider for payload mass range
- Addition of scatter plot between payload and success rates
- Reference : Space X dash application file

Predictive Analysis (Classification)

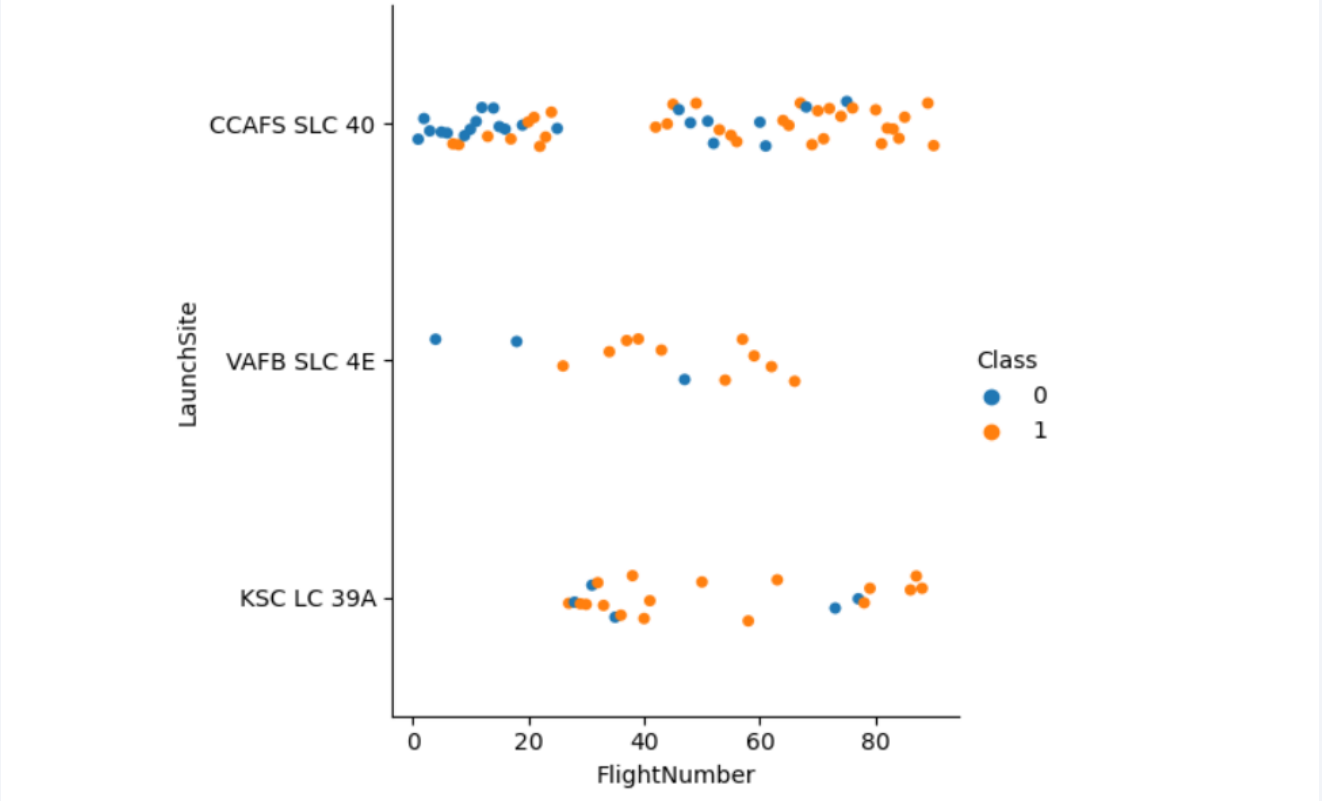
- Predictive analysis and model building for predictions on landing
- Data set divided in **training and test datasets**
- Predictive models including **logistic regression**, **Support Vector Machine**, **Decision tree** and **K Nearest Neighbors** are trained on training set
- These models are evaluated on test set
- Based on results, it was observed that decision tree performs better.
- Reference : Model Building Notebook

Results

- It was seen that different launch sites have different success rates.
- Success rates are increasing since year 2013.
- For orbits LEO, Polar, ISS with heavy payloads success rate is more.
- Among all possible models, decision tree performs well with accuracy around 87 percent.

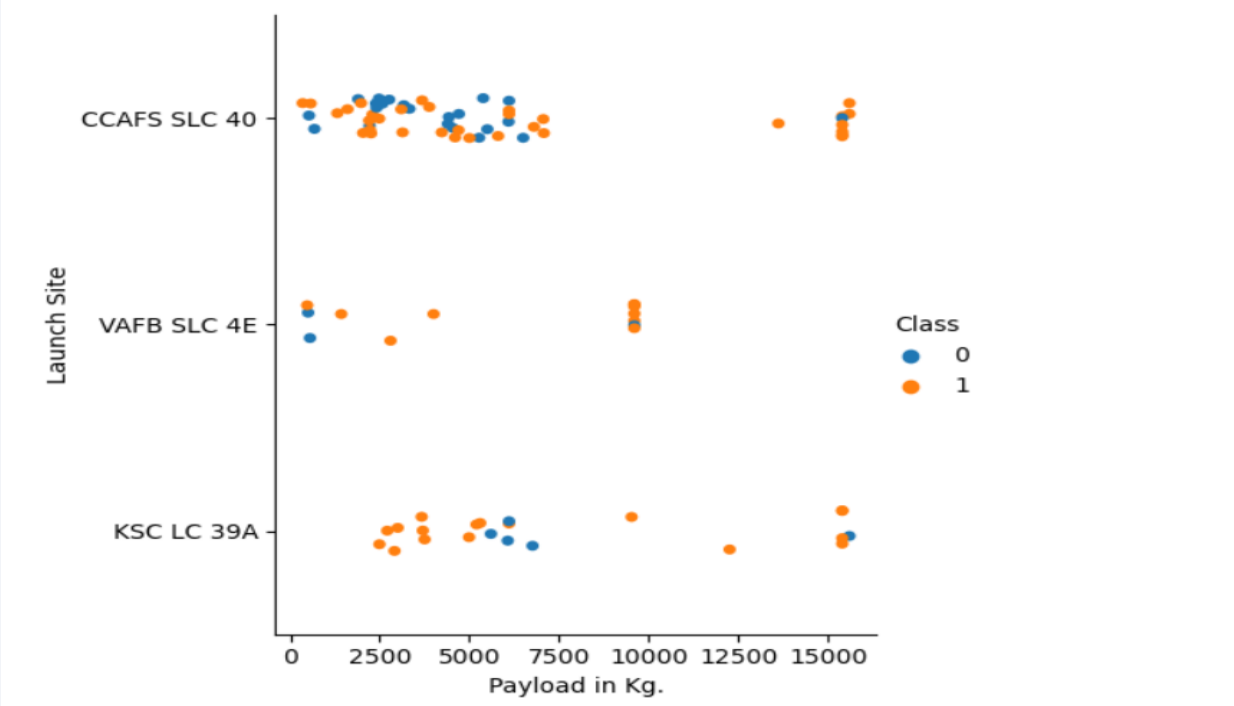

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



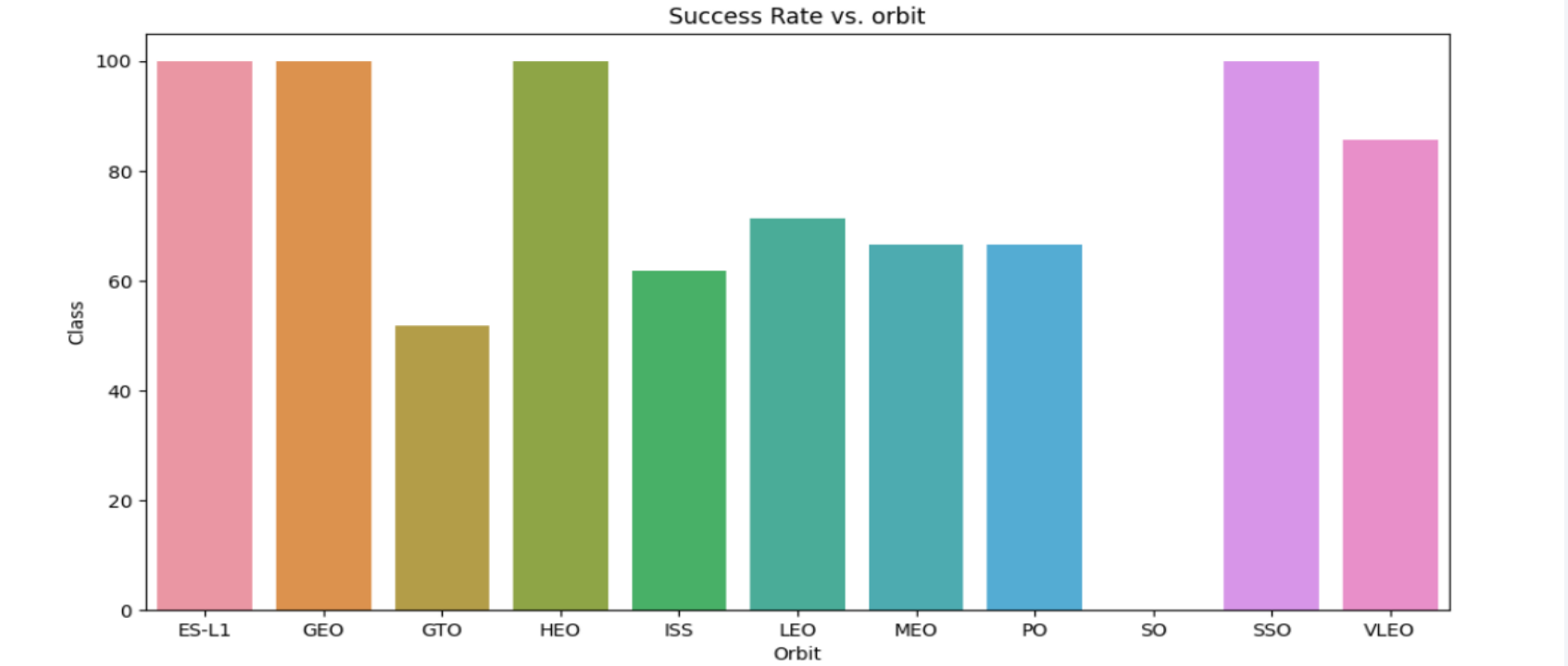
- Different sites have different success rates, KSC LC 39A and VAFB SCL 4E have higher success rates than CCAFS SCL 40 but CCAFS SLC 40 has highest number of launches among all three.

Payload vs. Launch Site



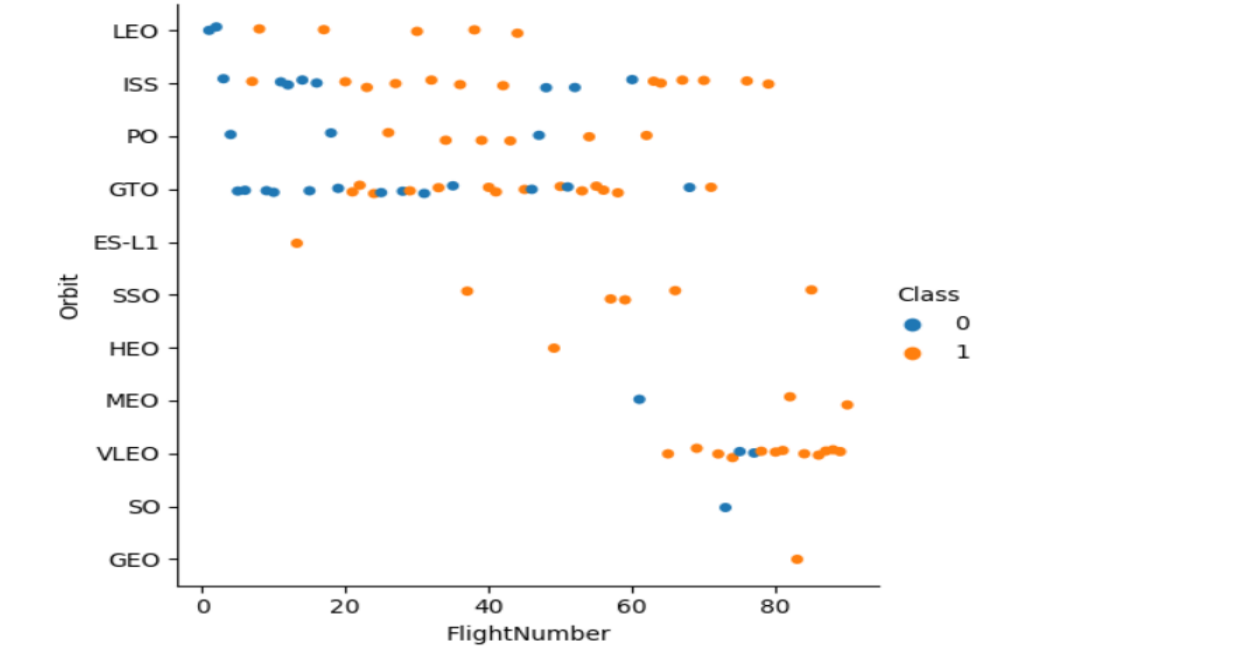
- Different sites have different distribution of payload but VAFB SLC 4E has no higher payload.

Success Rate vs. Orbit Type



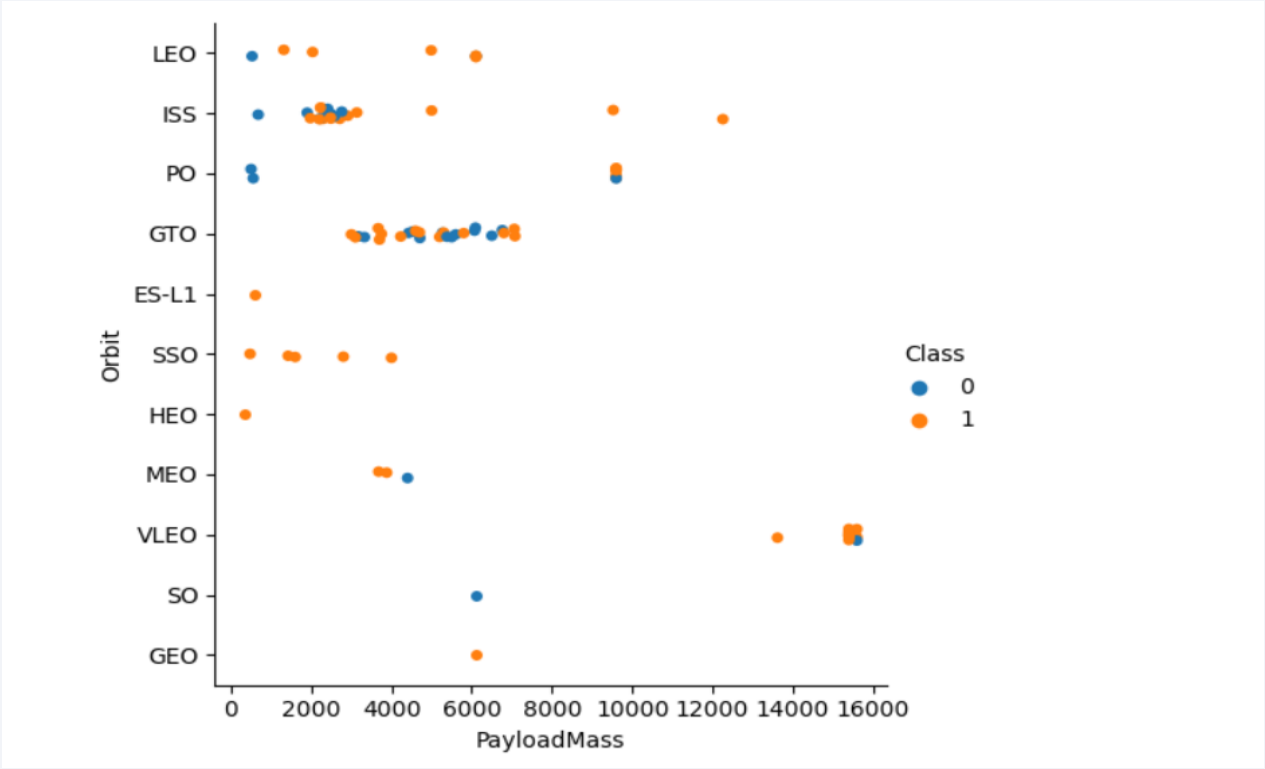
- Orbits ES-L1, GEO, HEO and SSO have 100 percent success rates but orbit SO has zero percent success rate accounting only single launching for orbit SO.

Flight Number vs. Orbit Type



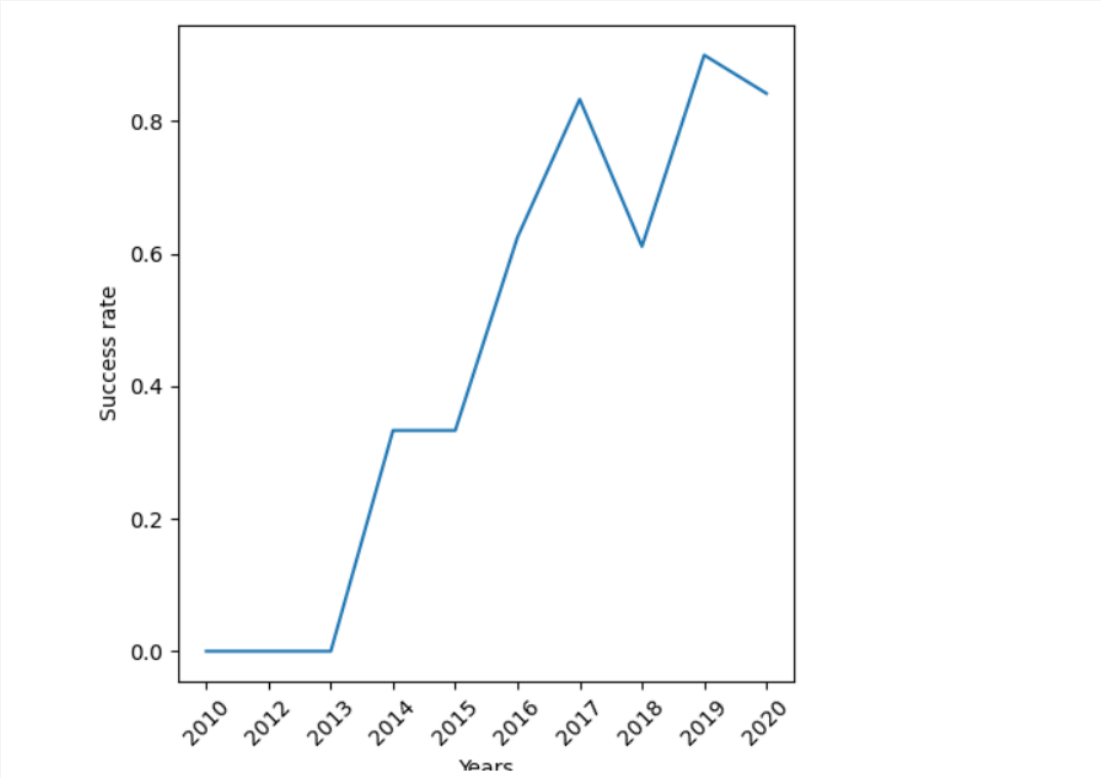
- For LEO orbit, it appears there is relationship between success rate and flight number but for GTO orbit, it doesn’t seem there is any relationship.

Payload vs. Orbit Type



- For heavy payloads, positive success rate are more for LEO, ISS and polar but there is no such relation for GTO.

Launch Success Yearly Trend



- Success rate is low, almost zero for years 2010,2012 and 2013 but since then, it keeps increasing until 2020.

All Launch Site Names

- The unique launch sites are:
  1. CCAFS LC-40
  2. VAFB SLC-40
  3. KSC LC-39A
  4. CCAFS SLC-40

[8]:

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

Launch Site Names Begin with 'CCA'

[10]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

```
[37]: %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer LIKE 'NASA%';
```

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

[37]:

SUM(PAYLOAD_MASS__KG_)
99980.0

Average Payload Mass by F9 v1.1

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

[38]:

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

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

[72]:

MIN(DATE)
01/08/2018

Successful Drone Ship Landing with Payload between 4000 and 6000

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

[48]:

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

Total Number of Successful and Failure Mission Outcomes

Done.

[59]:

FAILURE	SUCCESS
10	989

Boosters Carried Maximum Payload

[62]:

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

[65]:

MONTH	Landing_Outcome	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

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

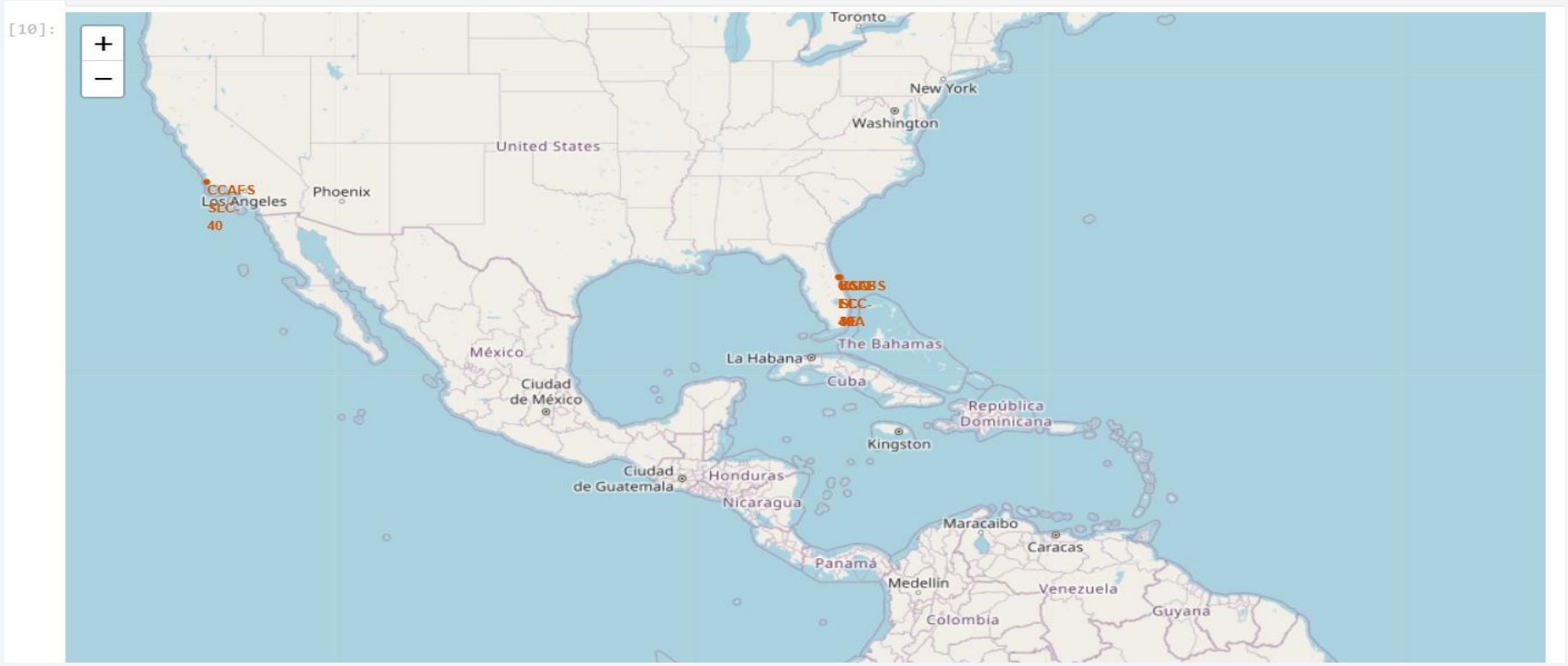
[20]:

Landing_Outcome	SUCCESS
Success (drone ship)	14
Controlled (ocean)	5
Failure	3
Failure (parachute)	2
No attempt	1

- Success rate in drone ship has most counts whereas controlled in ocean landing has relatively low count.


Section 3
Launch Sites
Proximities Analysis

Launch Sites on Map

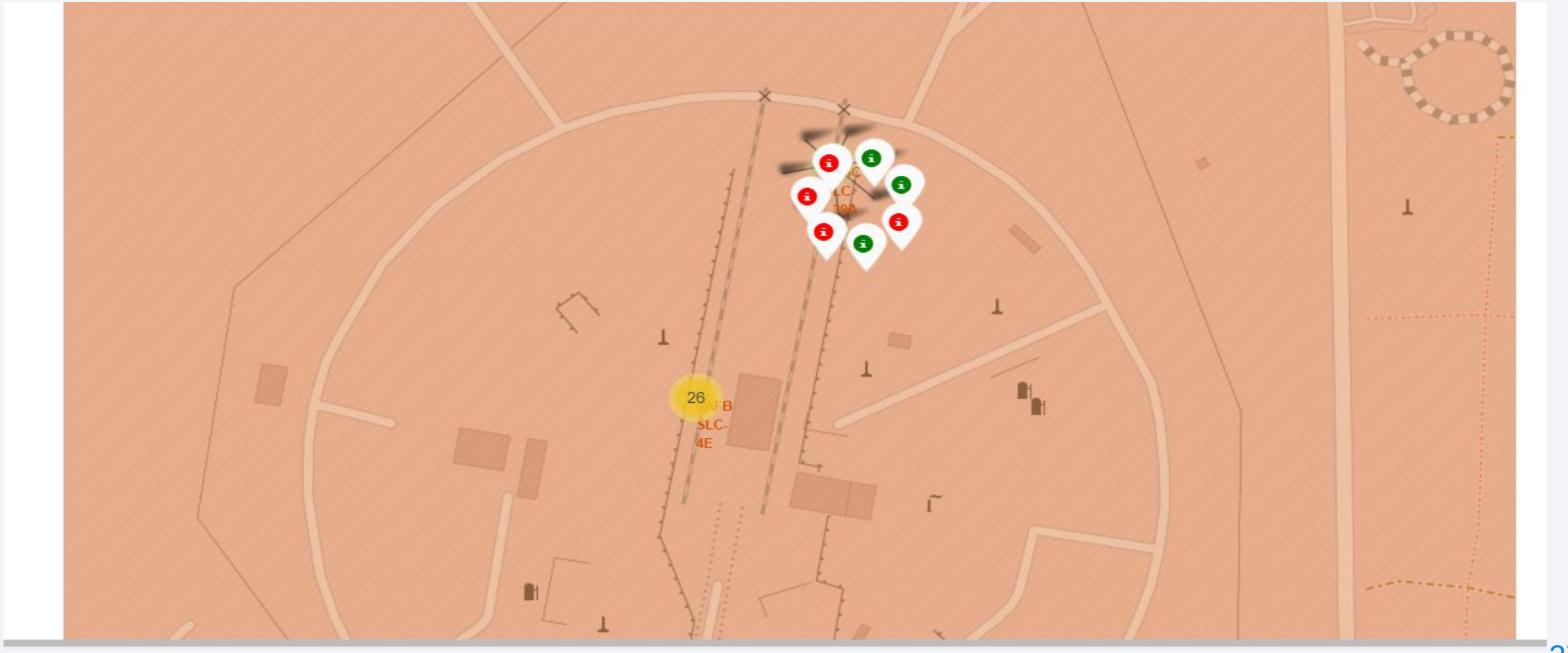


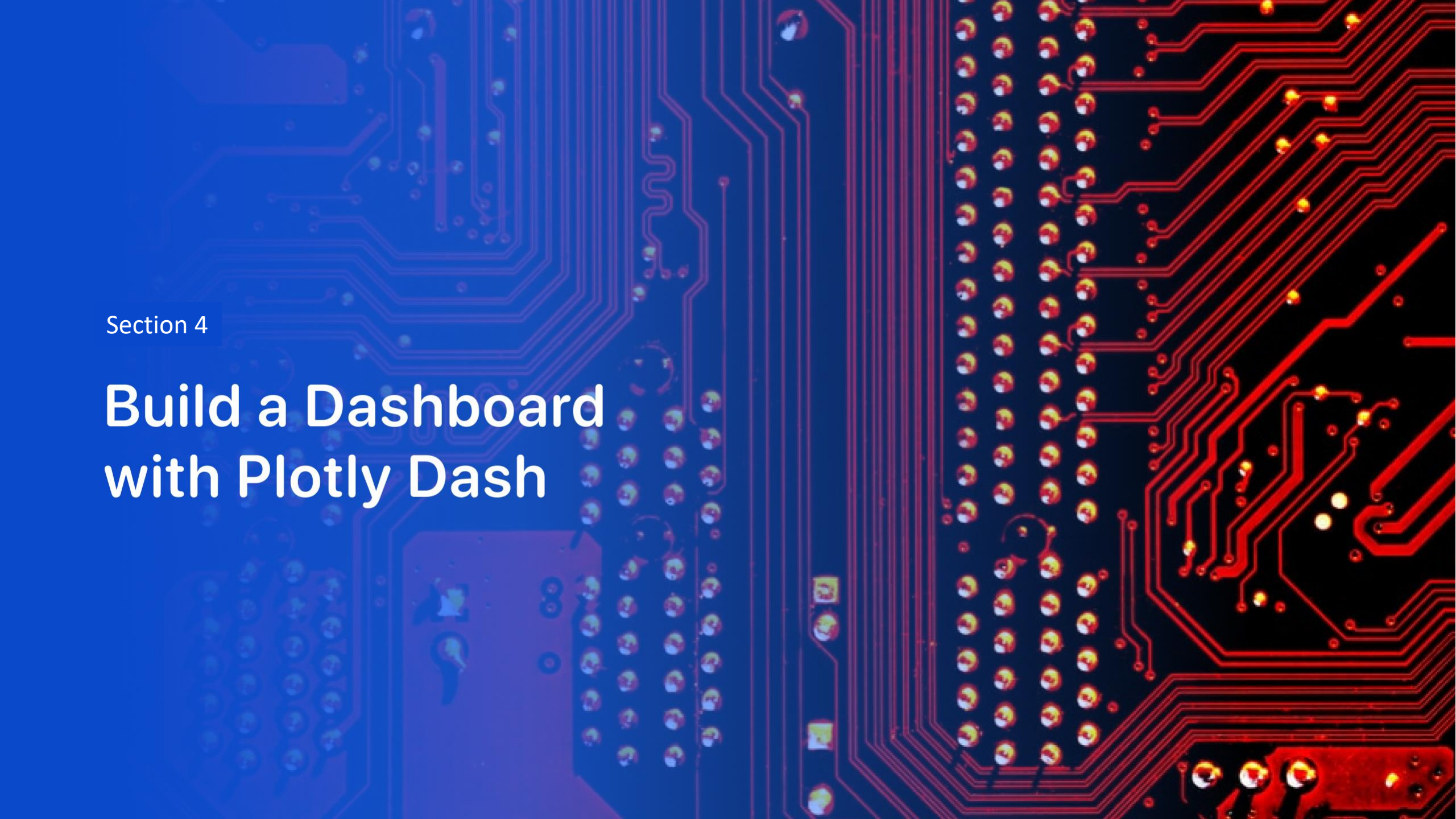
- These launch sites are located near coast lines.

Success and Failed launches on Launch sites

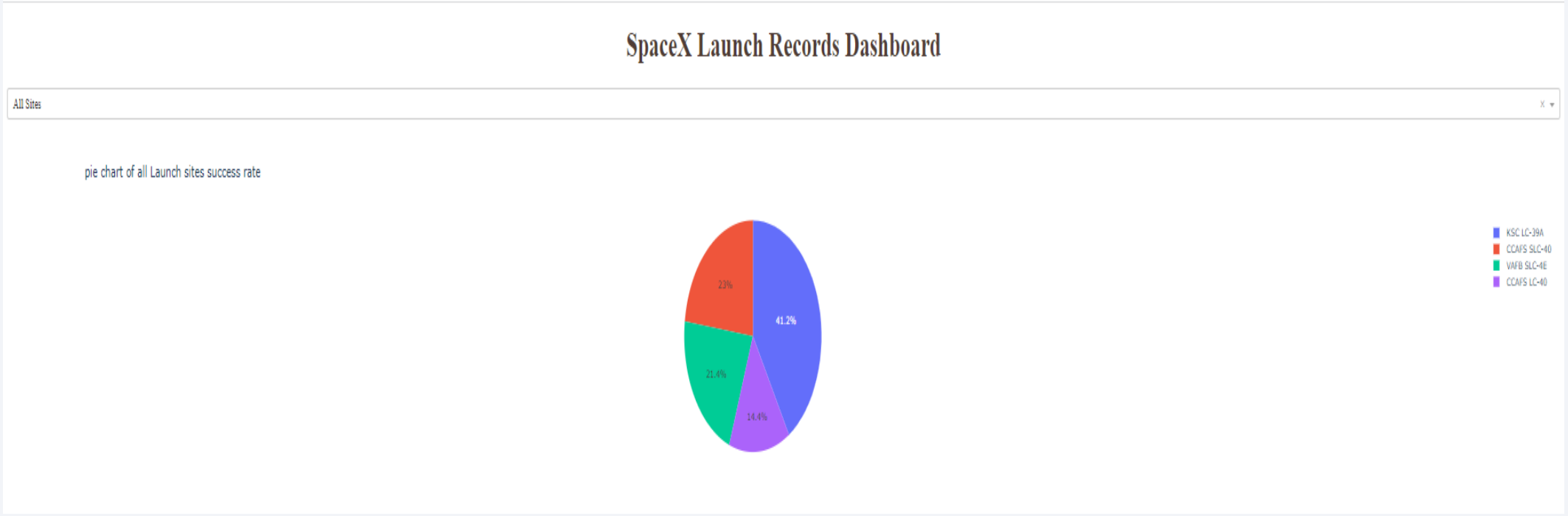
- The red marks show failed launches whereas green shows successful launches.

Distance between launch site and proximity



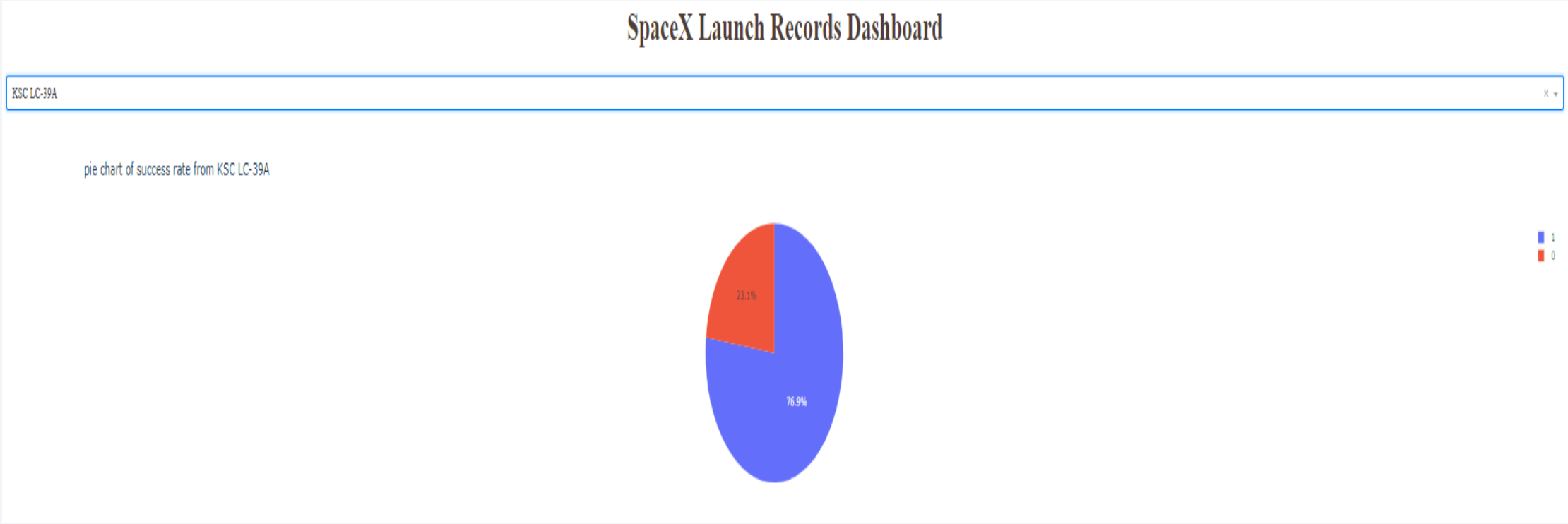

Section 4
Build a Dashboard with Plotly Dash

Dashboard



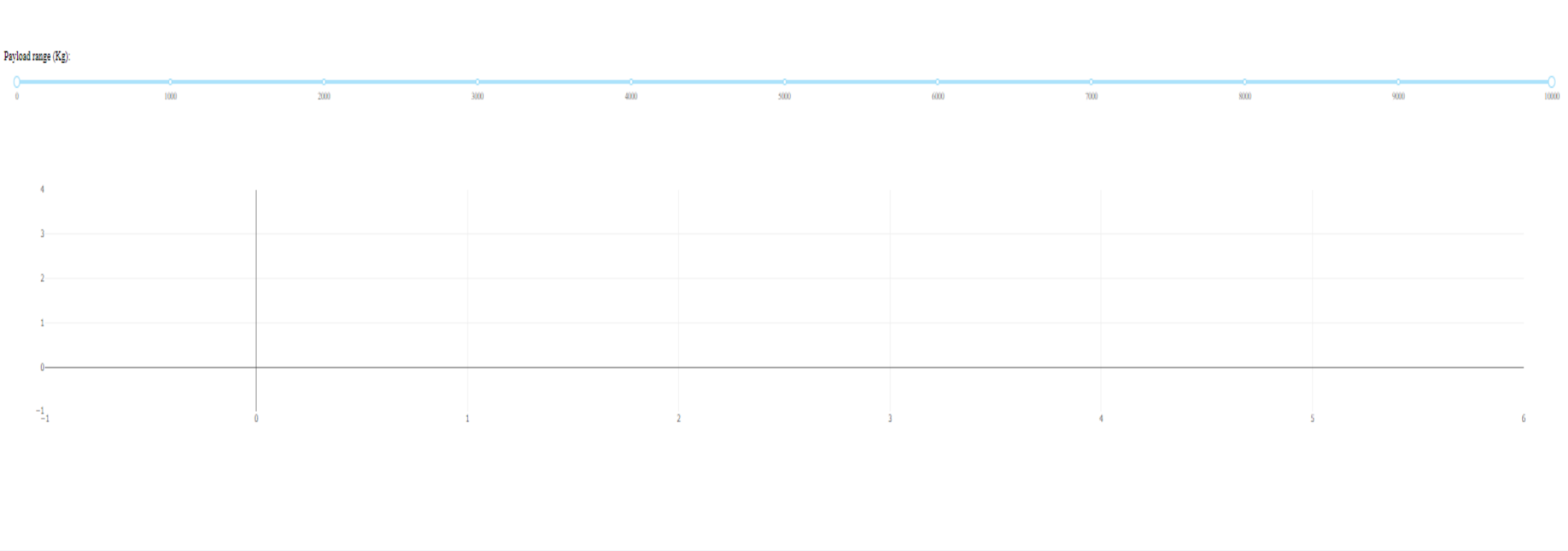
- KSC LC-39A has highest success rate whereas CCAFS LC-40 has lowest success rate.

Success and failure rate of KSC LC-39A



- Success rate for KSC LC-39A is 76.9 percent whereas failure rate is 23.1 percent.

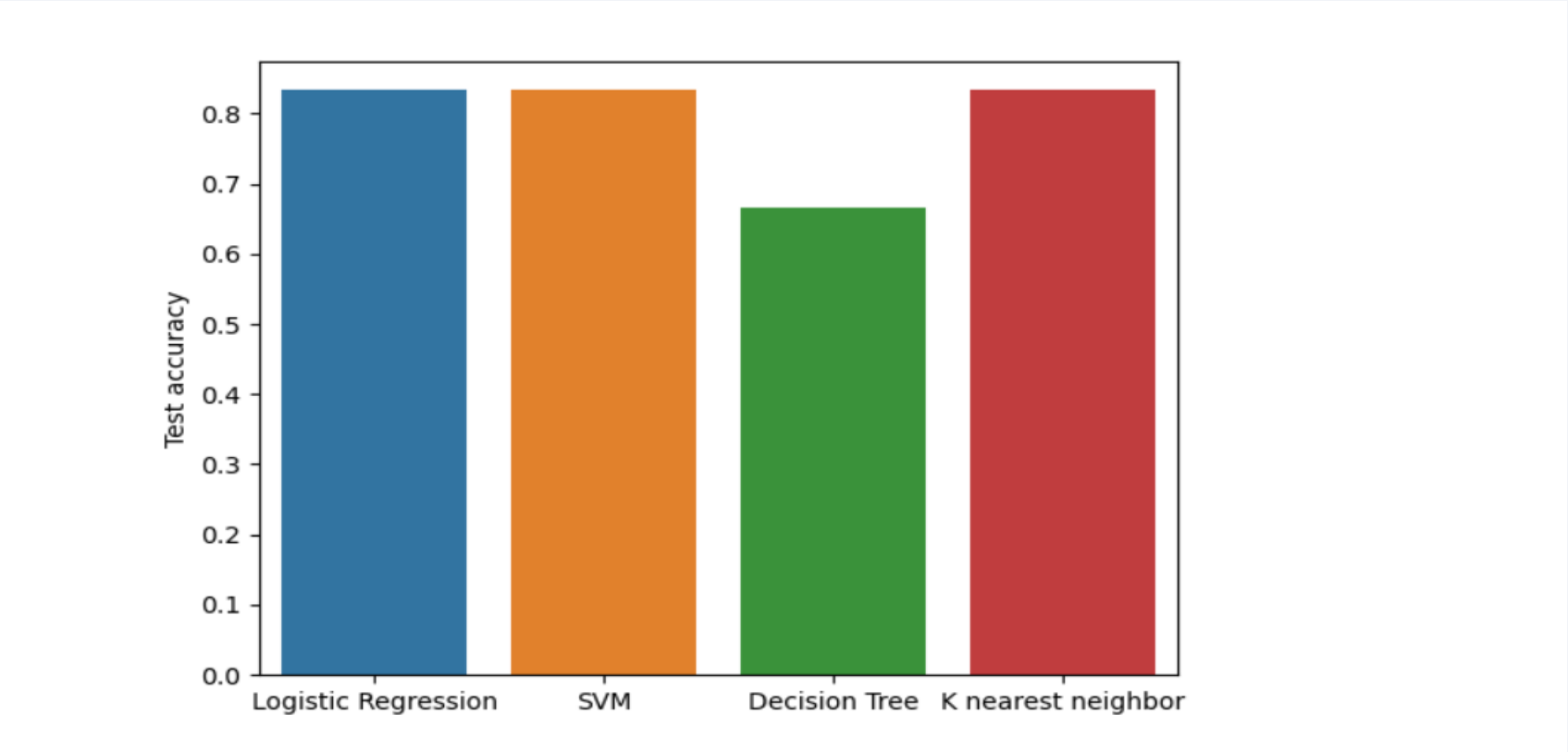
Scatter plot



Section 5

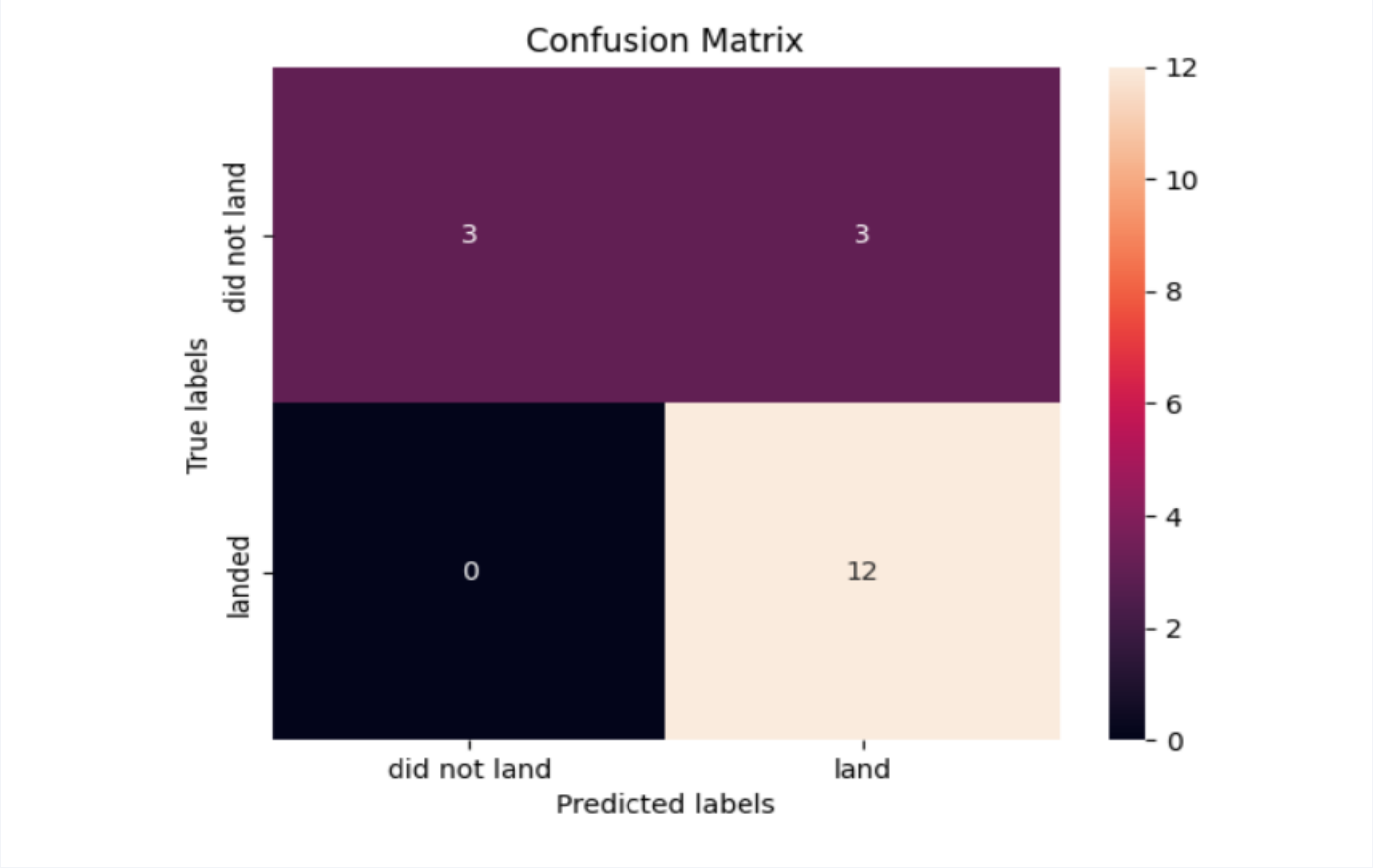
Predictive Analysis (Classification)

Classification Accuracy



- All algorithm has same almost accuracy on training dataset however decision tree performs worst on test set compared to other algorithms.

Confusion Matrix



- KNN algorithm performs well on both test and training datasets.

Conclusions

- Among all algorithms, decision tree performs better on training set.
- However, it performs worst on test data set.
- KNN performs well on both datasets.
- All three algorithms logistic regression, SVM and KNN perform same on test set but KNN performs better on training set.
- There exist correlation between landing outcome and payload mass, flight number and orbit.

Thank you!