Winning Space Race with Data Science

Marawan Kamal Bahgat Abouelsaad
09.05.2023

Table of Contents

Introduction

- Project background and context

  Space MB is a new company in the field of space rocket and the company recognized an opportunity to bid against its market rival Spacex.

  Whileas, SpaceX advertises Falcon 9 rocket launches on its website with a cost of $62M, other providers cost upward of $165M.

- Problems you want to find answers

  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 Space MB wants to bid against SpaceX.

Executive Summary

- Summary of methodologies
  - Data Collection & wrangling.
  - EDA with data visualization and with SQL
  - Building interactive map with Folium and dashboard with Dash
  - Machine learning prediction
- Summary of all results
  - EDA results
  - Interactive analysis
  - Predictive analysis

Section 1

Methodology

Methodology

Executive Summary

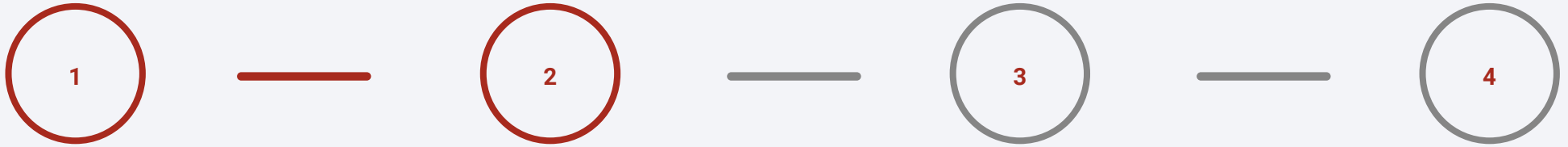
- Data collection methodology:
  - Space X REST API
  - Web scraping from Wikipedia
- Perform data wrangling
  - Determine and encode labels for training ML supervised models.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - LR, KNN, SVM and DT have been built and evaluated

Data Collection

Data Collection process has to phase

- SpaceX API
- Web Scraping for Falcon9 Wikipedia page

Data Collection – SpaceX API



Preparation

Import Libraries and Define Auxiliary Functions

Request Data

Request and parse the SpaceX launch data using the GET request

Filtering Data

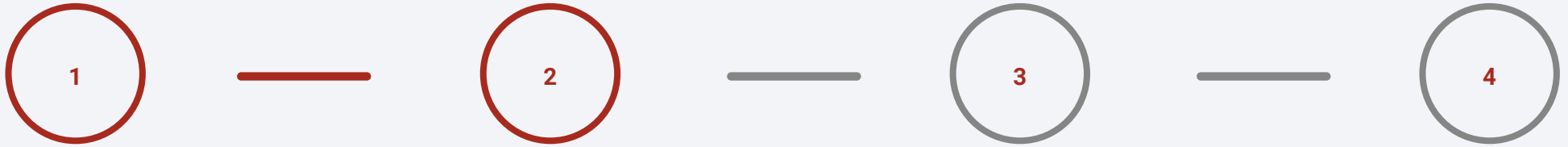
Filter the dataframe to only include Falcon 9 launches

Data Wrangling

Dealing with Missing Values of PayloadMass by using its mean

For more details follow the link below
https://github.com/marwanbahgat/Applied-Data-Science-Capstone/blob/9d889c6102ef791ab451e0a1923bf8d9597e75ec/01%20Hands-on%20Lab%20Complete%20the%20Data%20Collection%20API%20Lab.ipynb

Data Collection - Scraping



Preparation

import required packages

Request Data

Request the Falcon9 Launch Wiki page from its URL

Data Processing

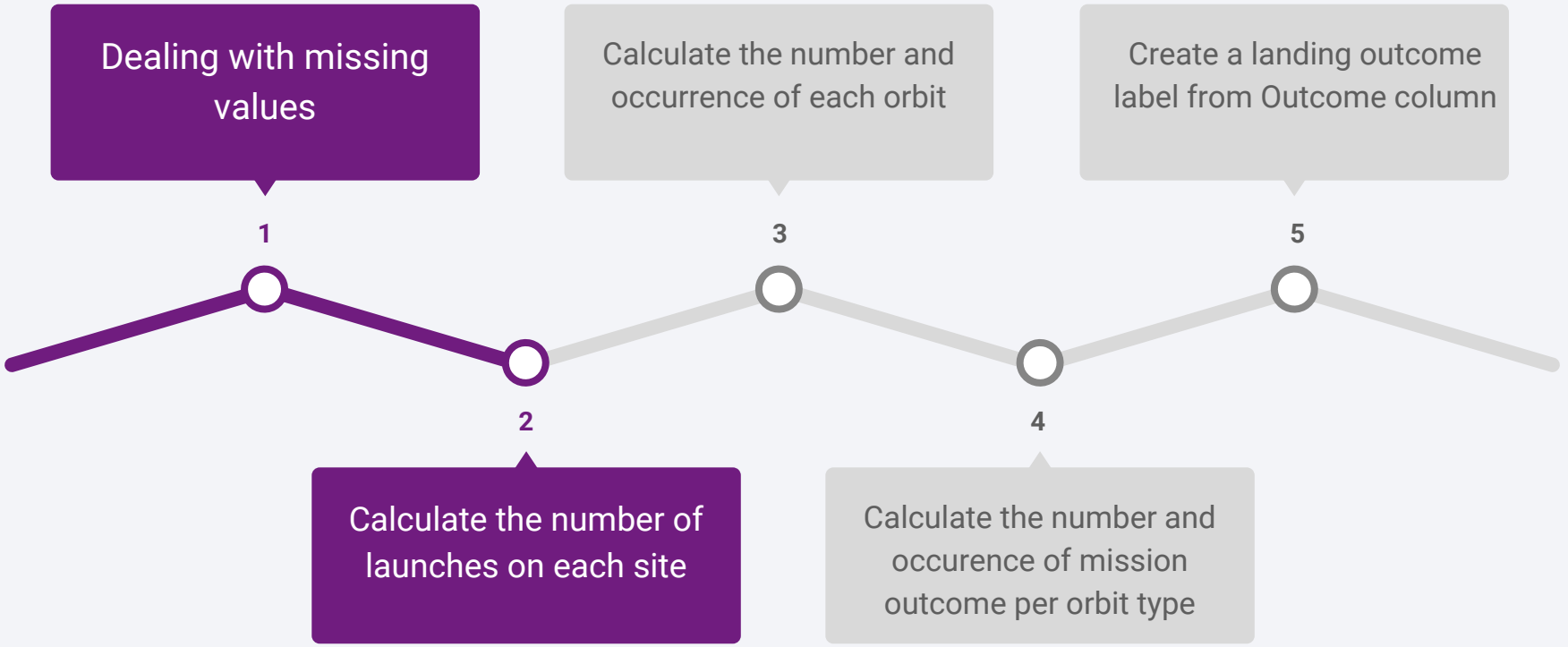
Extract all column/variable names from the HTML table header

Data Parsing

Create a data frame by parsing the launch HTML tables

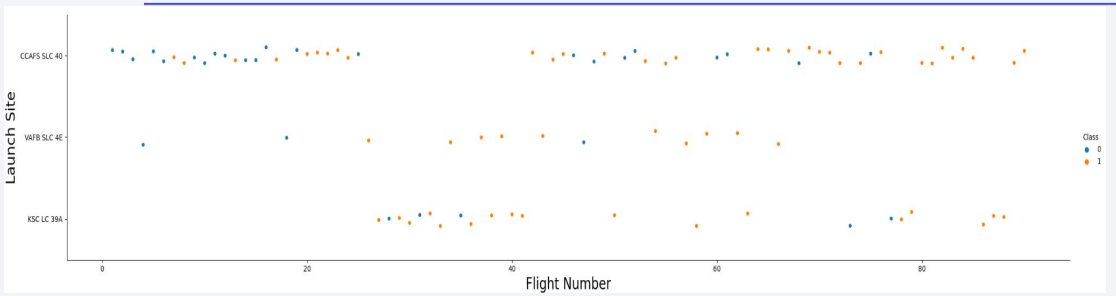
For more details follow the link below
https://github.com/marwanbahgat/Applied-Data-Science-Capstone/blob/9d889c6102ef791ab451e0a1923bf8d9597e75ec/02%20Hands-on%20Lab%20Complete%20the%20Data%20Collection%20with%20Web%20Scraping%20lab.ipynb

Data Wrangling

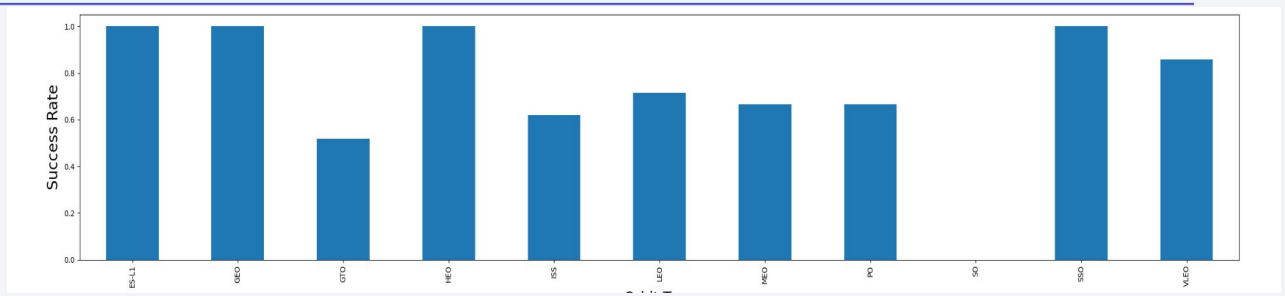


For more details follow the link below
https://github.com/marwanbahgat/Applied-Data-Science-Capstone/blob/9d889c6102ef791ab451e0a1923bf8d9597e75ec/03%20Hands-on%20Lab%20Data%20Wrangling.ipynb

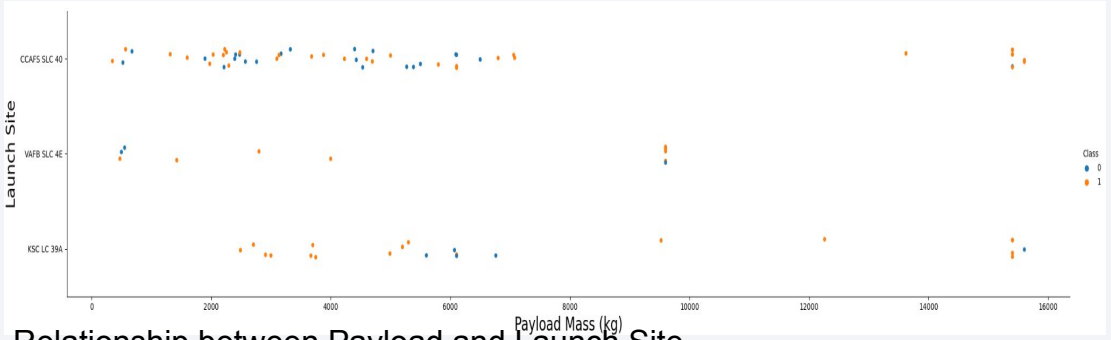
EDA with Data Visualization



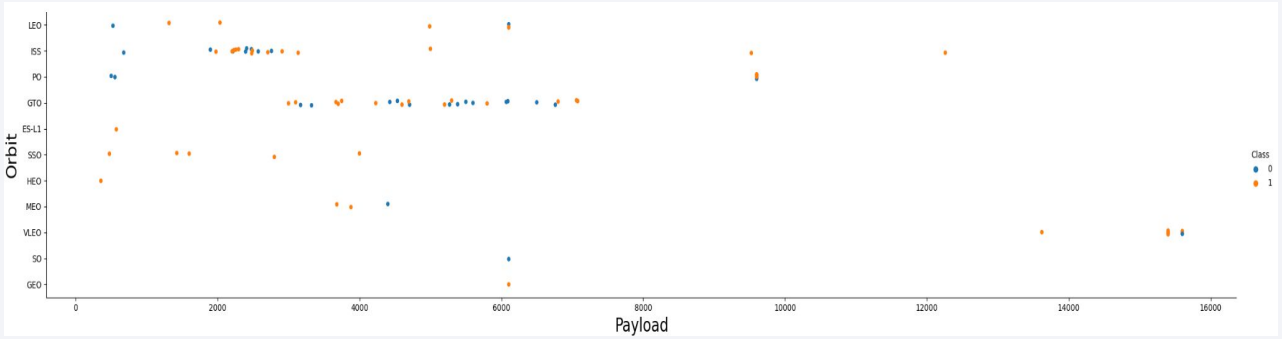
Relationship between Flight Number and Launch Site



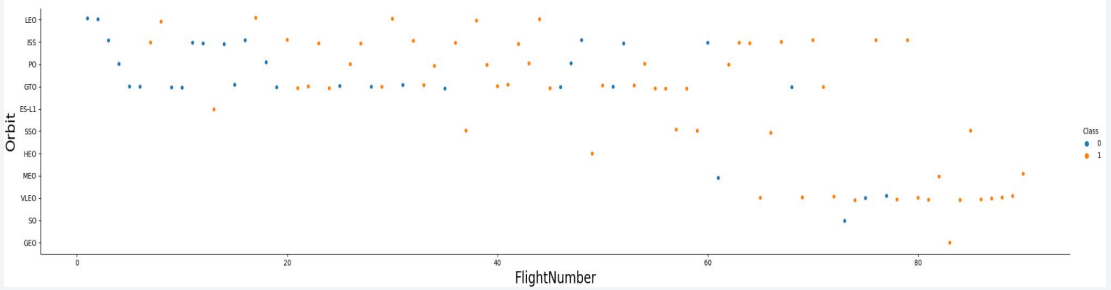
Relationship between success rate of each orbit type



Relationship between Payload and Launch Site

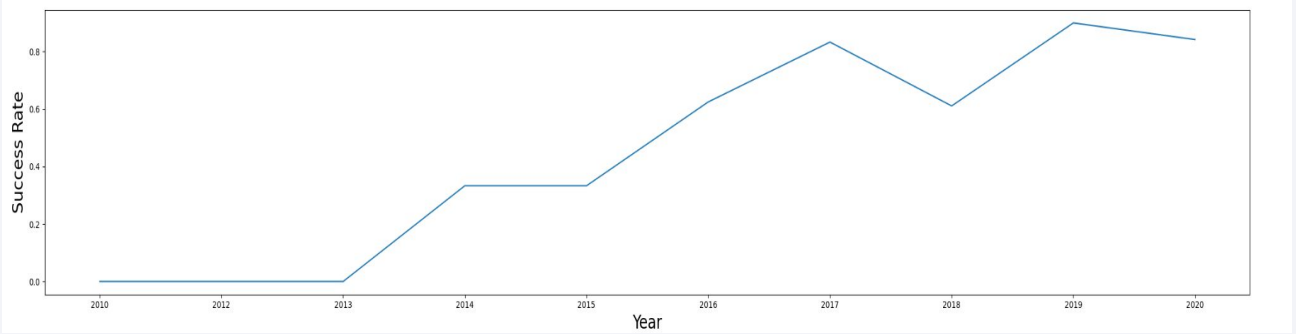


Relationship between Payload and Orbit type



Relationship between FlightNumber and Orbit type

https://github.com/marwanbahgat/Applied-Data-Science-Capstone/blob/9d889c6102ef791ab451e0a1923bf8d9597e75ec/05%20EDA%20with%20Visualization%20Lab.ipynb



Visualize the launch success yearly trend

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 the 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

For more details follow the link below
https://github.com/marwanbahgat/Applied-Data-Science-Capstone/blob/9d889c6102ef791ab451e0a1923bf8d9597e75ec/04%20Hands-on%20Lab%20Complete%20the%20EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- Mark all launch sites on a map
  - add a highlighted circle area with a text label on a specific coordinate.
- Mark the success/failed launches for each site on the map
  - create a MarkerCluster object.
  - For each launch result in spacex_df data frame, add a folium.Marker to marker_cluster
- Calculate the distances between a launch site to its proximities
  - a MousePosition added on the map to get coordinate for a mouse over a point on the map.
  - Mark down a point on the closest coastline using MousePosition and calculate the distance between the coastline point and the launch site.
  - Create and add a folium.Marker on your selected closest coastline point on the map

object added to find some geographic patterns about the launch site

For more details follow the link below
https://github.com/marwanbahgat/Applied-Data-Science-Capstone/blob/9d889c6102ef791ab451e0a1923bf8d9597e75ec/05%20Hands-on%20Lab%20Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

- Used Python interactive dashboarding library Plotly Dash to enable stakeholders to explore and manipulate data in an interactive and real-time way
- Pie chart showing success rate
  - Color coded by launch site
- Scatter chart showing payload mass vs. landing outcome
  - Color coded by booster version
  - With range slider for limiting payload amount
- Drop-down menu to choose between all sites and individual launch sites
- Explain why you added those plots and interactions

For more details follow the link below

https://github.com/marwanbahgat/Applied-Data-Science-Capstone/blob/9d889c6102ef791ab451e0a1923bf8d9597e75ec/06%20space%20x%20dash.py

Predictive Analysis (Classification)

SPACE MB

Preparing Data	Standardized the data	Splitting Data	Fit training data	Vestibulum congue tempus	Evaluation
Created a column for our training label 'Class' created during data wrangling	Standardize the data in X then re-assign it to the variable X using the transform	Split the data into training data and test data	• Logistic Regression • Support Vector Machine • Decision Tree Classifier • K Nearest Neighbors Classifier	Used a cross-validated grid-search over a variety of hyperparameters to select the best ones for each model	Evaluated accuracy of each model using test data to select the best model

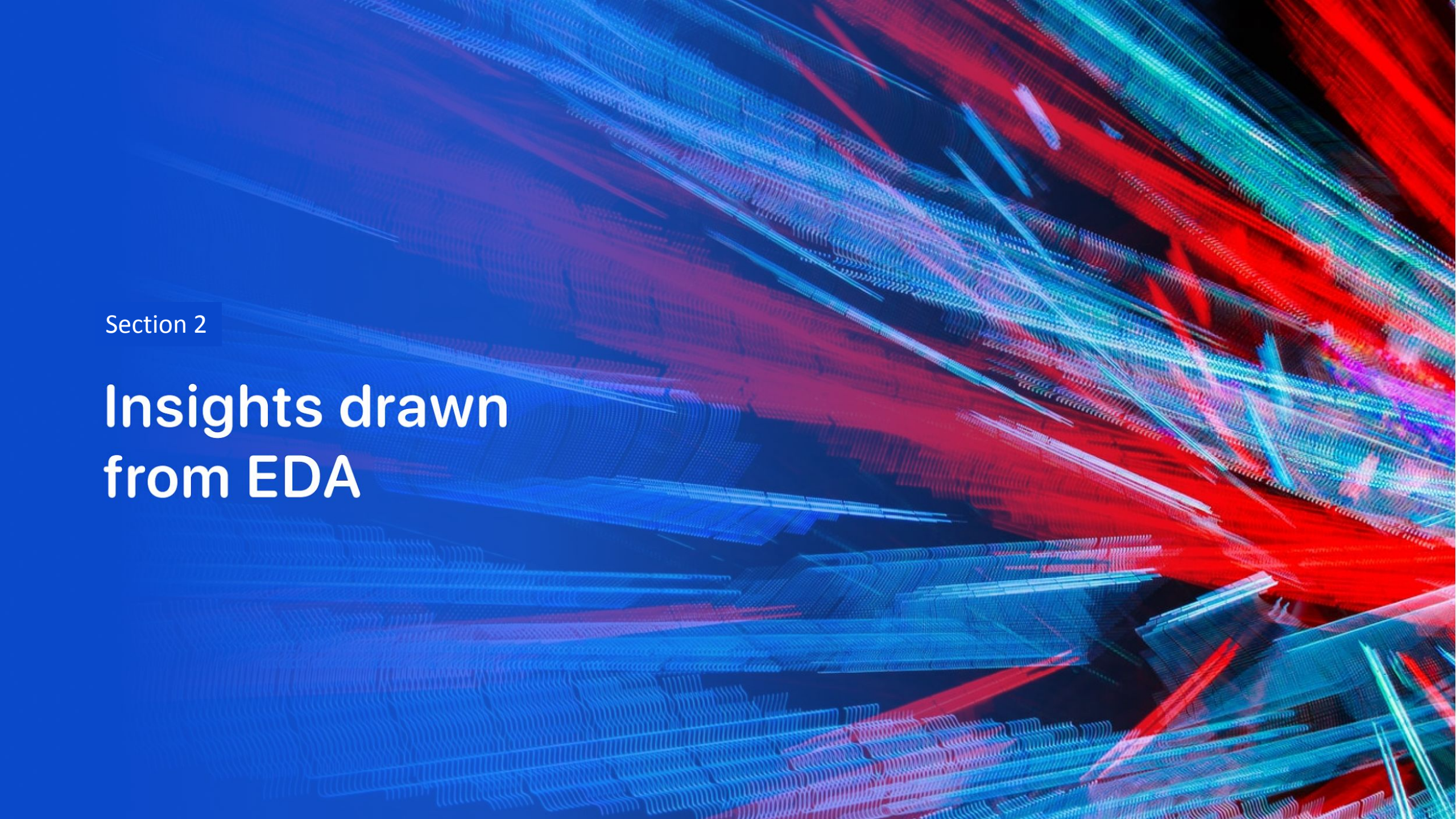
For more details follow the link below
https://github.com/marwanbahgat/Applied-Data-Science-Capstone/blob/9d889c6102ef791ab451e0a1923bf8d9597e75ec/07%20Hands-on%20Lab%20Complete%20the%20Machine%20Learning%20Prediction%20lab.ipynb

Results

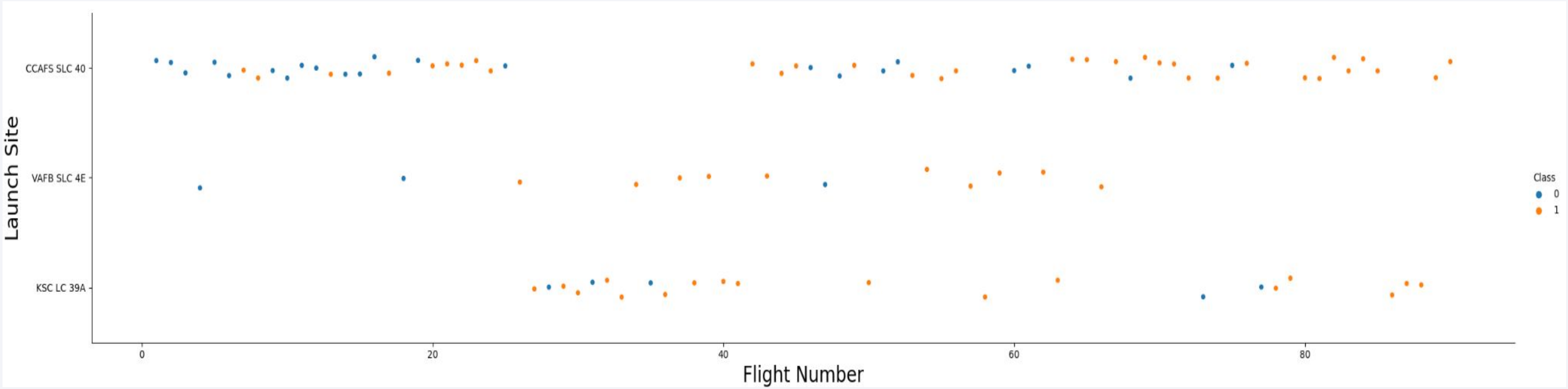
- The SVM, KNN, and Logistic Regression models are the best in terms of prediction accuracy for this dataset.
- Lightweight payloads perform better than heavier payloads.
- SpaceX's launch success rates directly proportional to the number of years in they will finish perfecting the launches.
- KSC LC 39A had the most successful launches of all sites.
- Orbit GEO,HEO,SSO,ES L1 has the best Success Rate

Section 2

Insights drawn from EDA

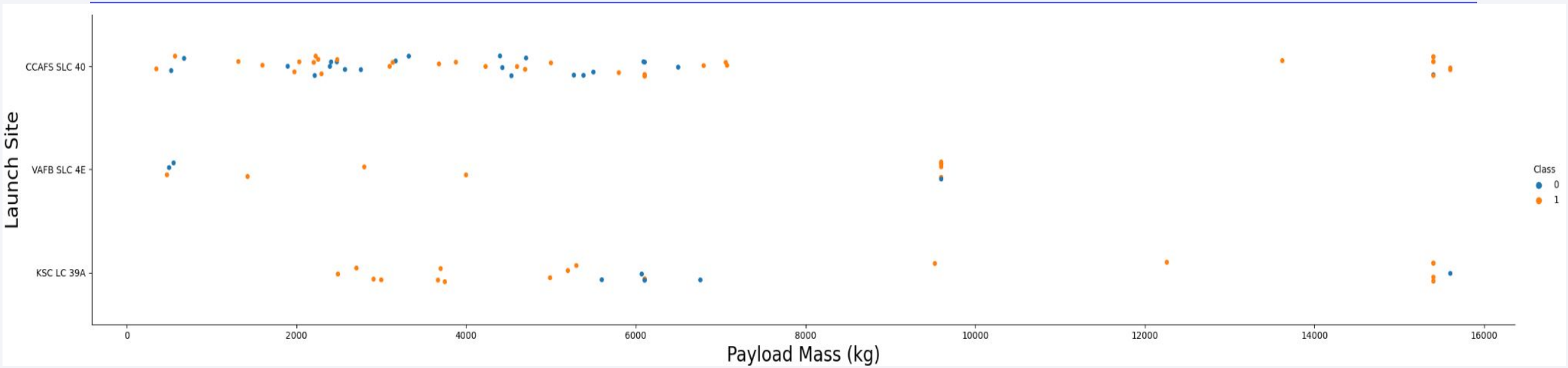


Flight Number vs. Launch Site



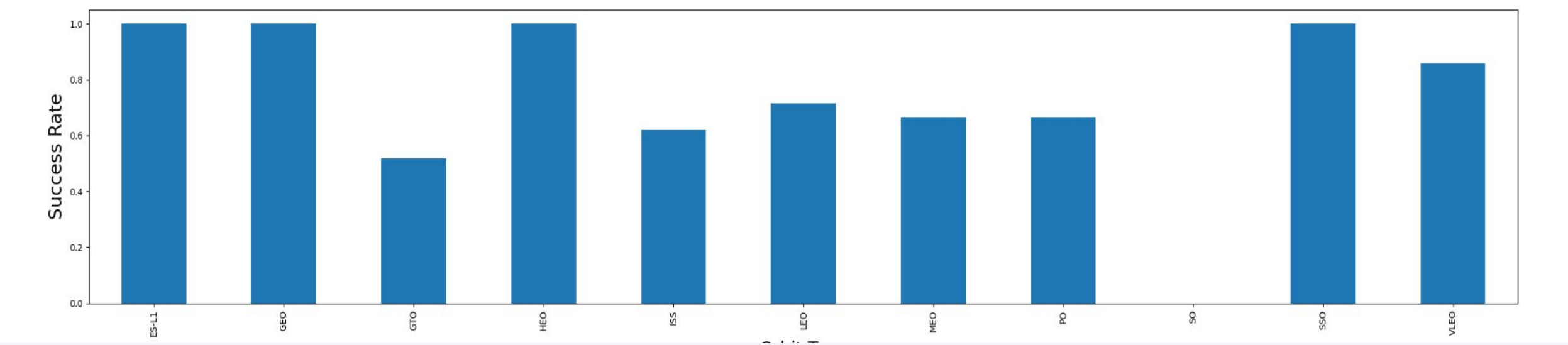
- CCAFS SLC 40 has highest success rate
- With increase of the flight number the success rate increases

Payload vs. Launch Site



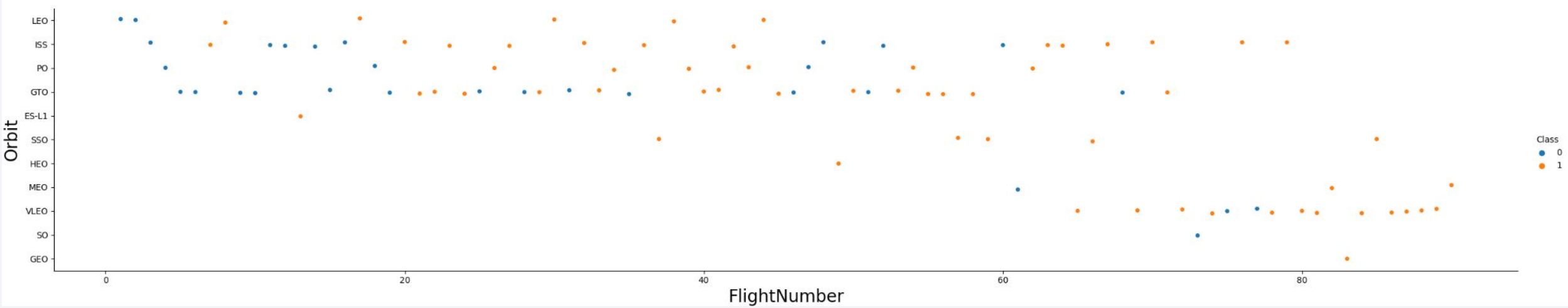
- for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).Hence, the other location is favored for heavypayload

Success Rate vs. Orbit Type



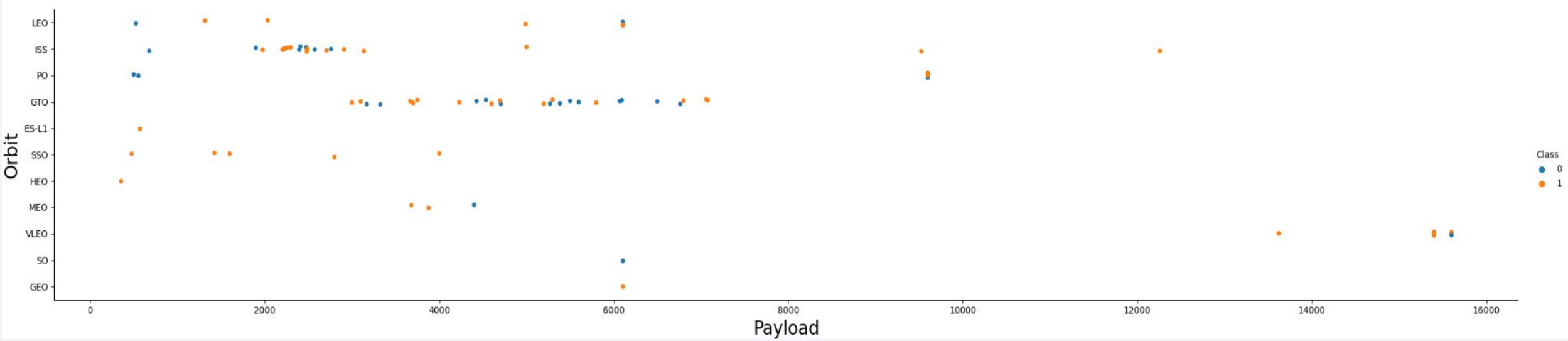
- ES-11, GEO, HEO, SSO has the highest successful rate

Flight Number vs. Orbit Type



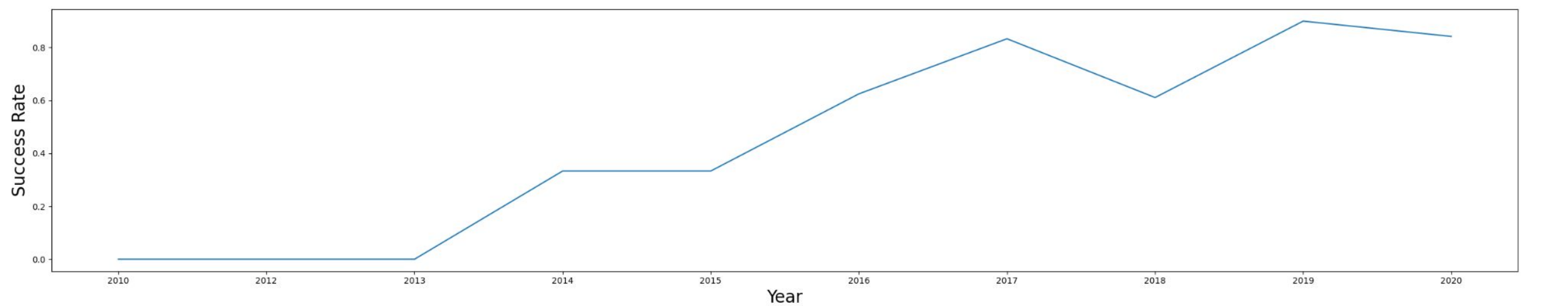
- LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend



- Launch successful rate start increases since 2013
- The success rate since 2013 kept increasing till 2020

All Launch Site Names

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

- We get location site by the following code

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

Launch Site Names Begin with 'CCA'

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

- 5 records where launch sites begin with the string 'CCA'

Total Payload Mass

- the total payload mass carried by boosters launched by NASA (CRS)
  45596
- We get this by the following code

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_)
FROM SPACEXTBL
WHERE Customer = 'NASA (CRS)';
```

Average Payload Mass by F9 v1.1

- the average payload mass carried by booster version F9 v1.1 is 2534.6666666666665
- We get this by the following code

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_)
FROM SPACEXTBL
WHERE Booster_Version LIKE 'F9 v1.1%';
```

First Successful Ground Landing Date

- The first successful landing outcome on ground pad is 22/12/2015
- We get this by the following code

```
%%sql
SELECT min(DATE)
FROM SPACEXTBL
WHERE [Landing_Outcome] = 'Success (ground pad)';
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- We get this by the following code

```
%%sql
SELECT BOOSTER_VERSION
FROM SPACEXTBL
WHERE [Landing_Outcome] = 'Success (drone ship)' AND
4000 < PAYLOAD_MASS__KG_ < 6000;
```

Booster_Version
F9 FT B1021.1
F9 FT B1022
F9 FT B1023.1
F9 FT B1026
F9 FT B1029.1
F9 FT B1021.2
F9 FT B1029.2
F9 FT B1036.1
F9 FT B1038.1
F9 B4 B1041.1
F9 FT B1031.2
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1

Total Number of Successful and Failure Mission Outcomes

- The total number of successful Mission is 100 and failure mission outcomes 1
- we got the result by count the mission outcomes group by the mission outcome as the following code

```
%%sql

SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL_NUMBER

FROM SPACEXTBL

GROUP BY MISSION_OUTCOME;
```

Boosters Carried Maximum Payload

- We get this by the following code

```
%%sql
SELECT DISTINCT BOOSTER_VERSION
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (
    SELECT MAX(PAYLOAD_MASS__KG_)
    FROM SPACEXTBL);
```

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

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

- We get this by the following code

```
%%sql
SELECT [Landing_Outcome], BOOSTER_VERSION, LAUNCH_SITE
FROM SPACEXTBL
WHERE [Landing_Outcome] = 'Failure (drone ship)'
    AND Date like '%2015';
```

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

- We get this by the following code

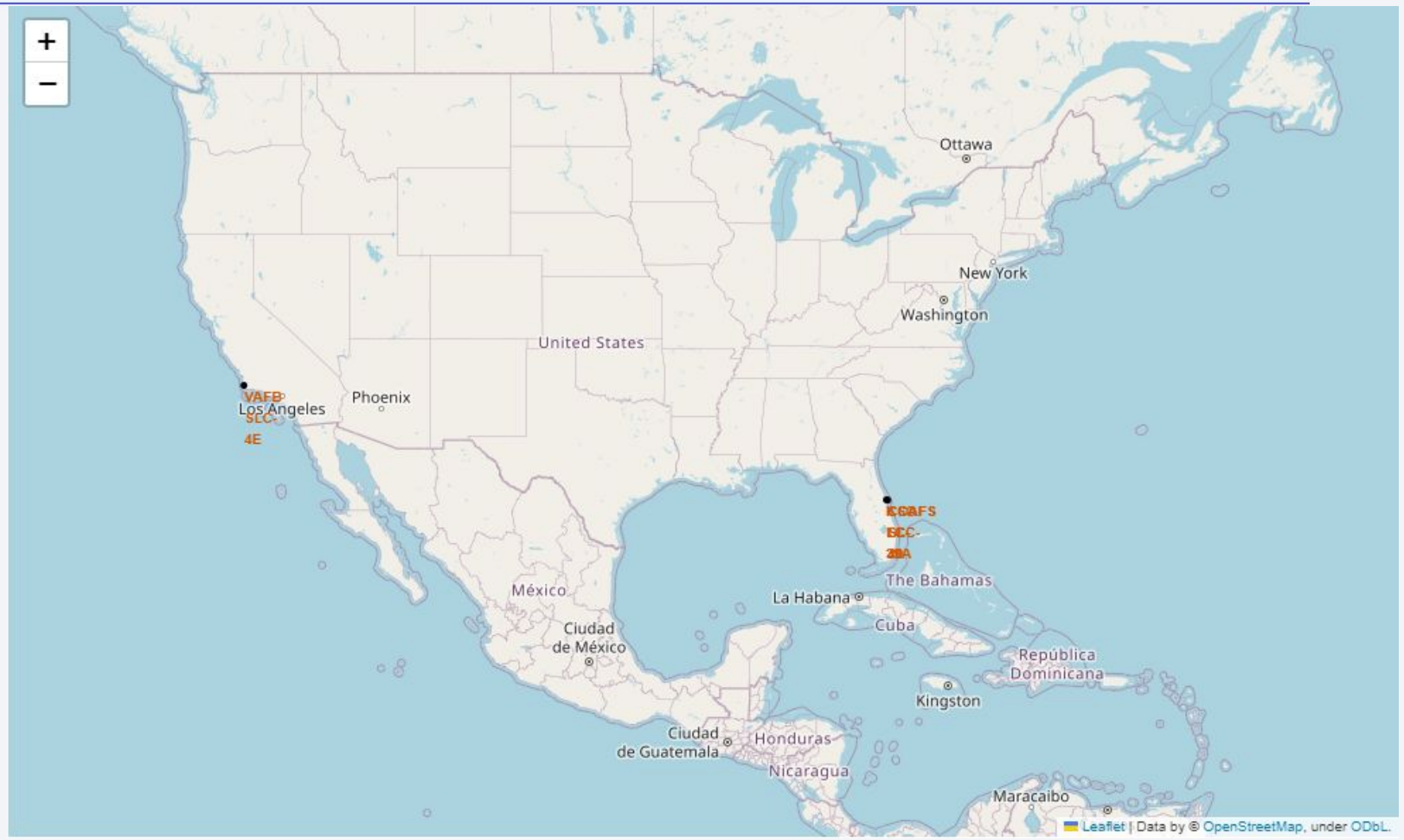
```
%%sql
SELECT [Landing_Outcome], COUNT([Landing_Outcome]) AS
TOTAL_NUMBER
FROM SPACEXTBL
WHERE Date BETWEEN '04-06-2010' AND '20-03-2017'
GROUP BY [Landing_Outcome]
ORDER BY TOTAL_NUMBER DESC
```

Landing_Outcome	TOTAL_NUMBER
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1

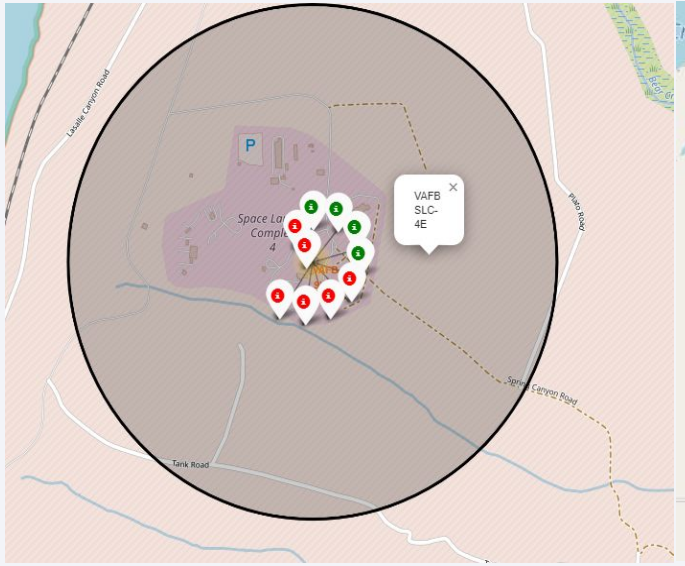

Section 3
Launch Sites
Proximities Analysis

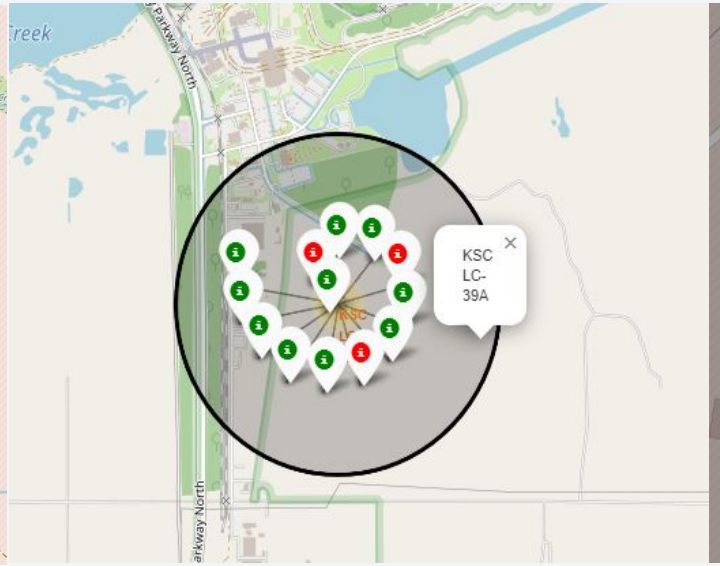
launch sites

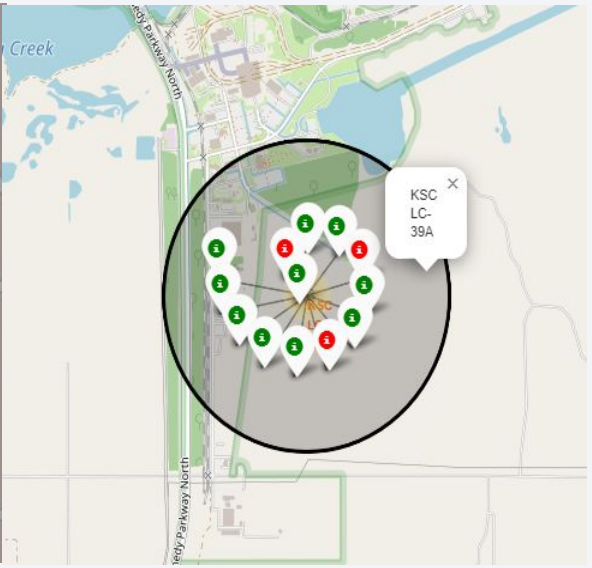
- All Launch site are marked on the map



Success and failed Location

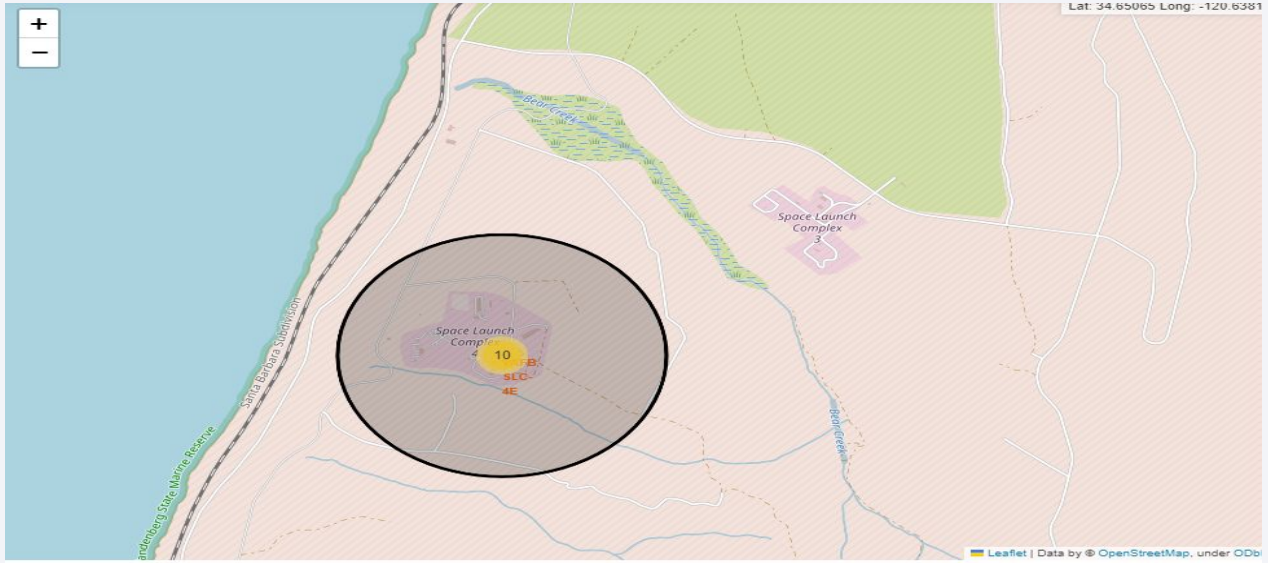




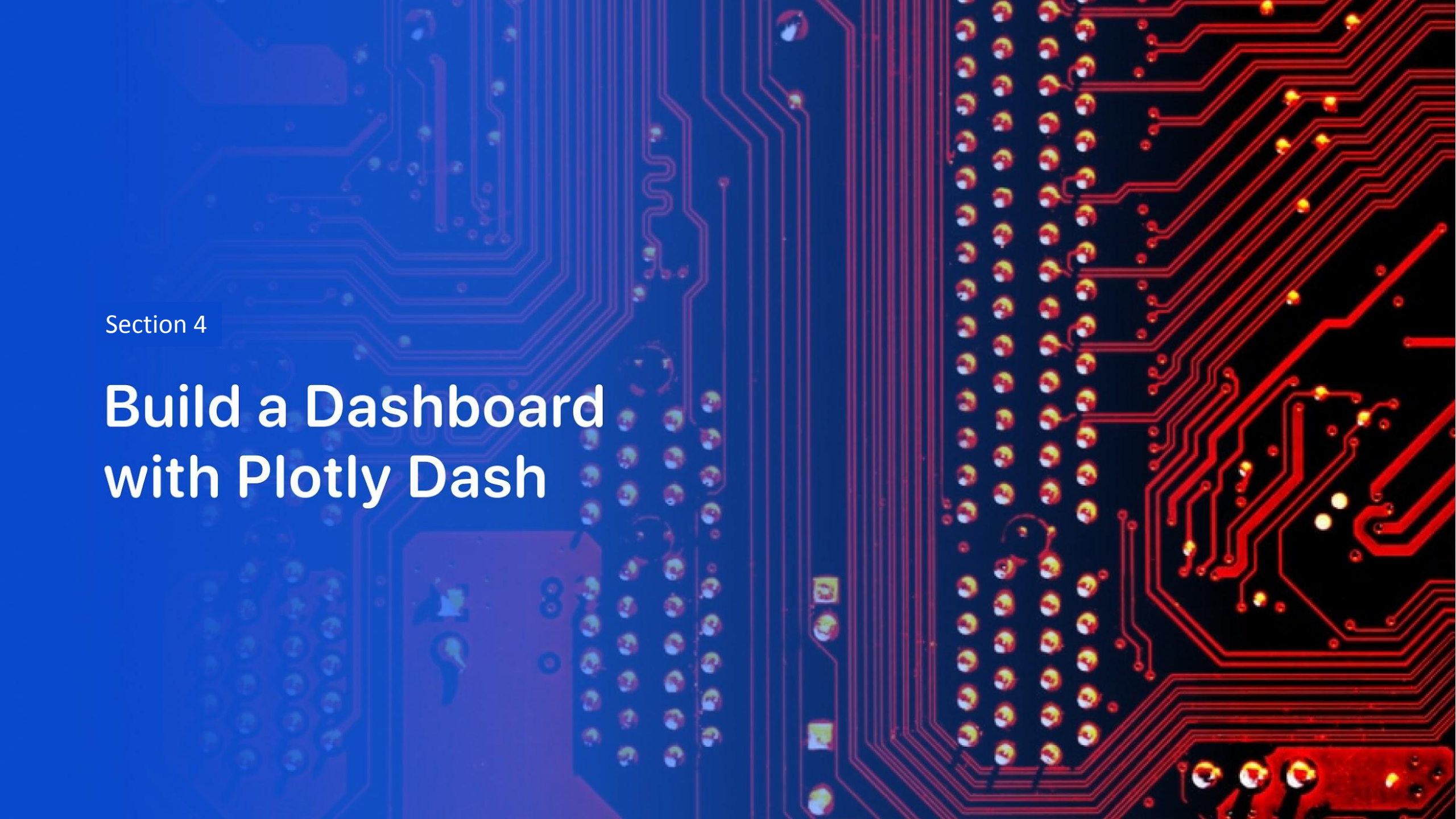


Success and failed launch location shown on the map as green represent success and red represent failed launches

Launch location and its proximity

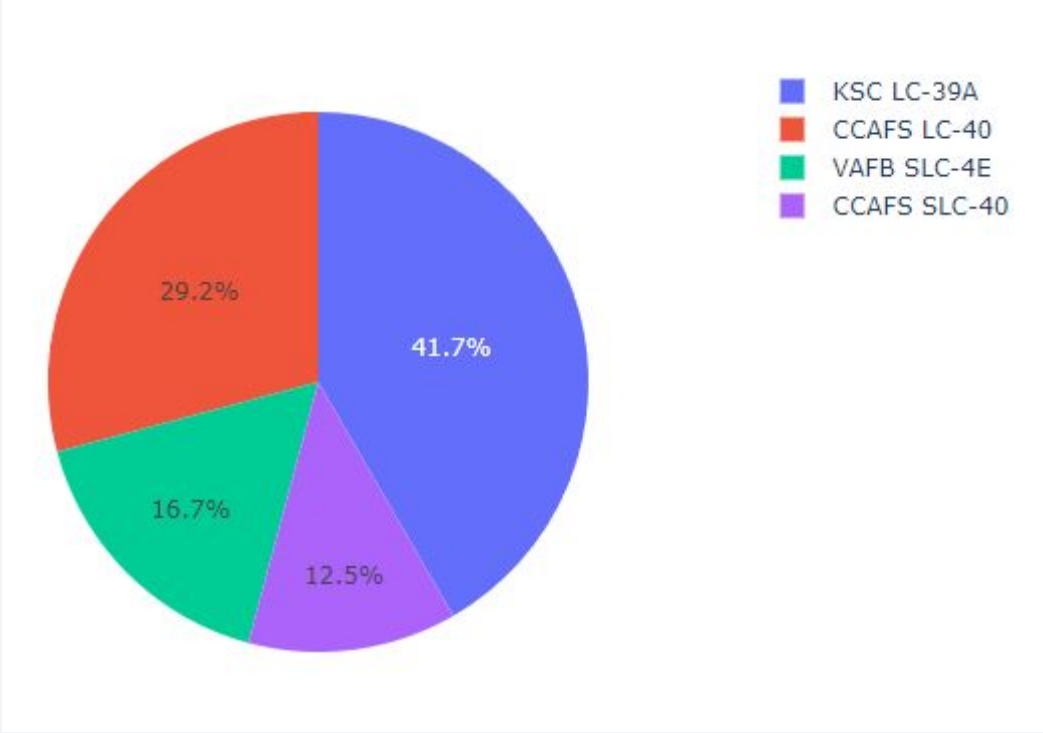


All Location is near the cost and various transportation means


Section 4
Build a Dashboard with Plotly Dash

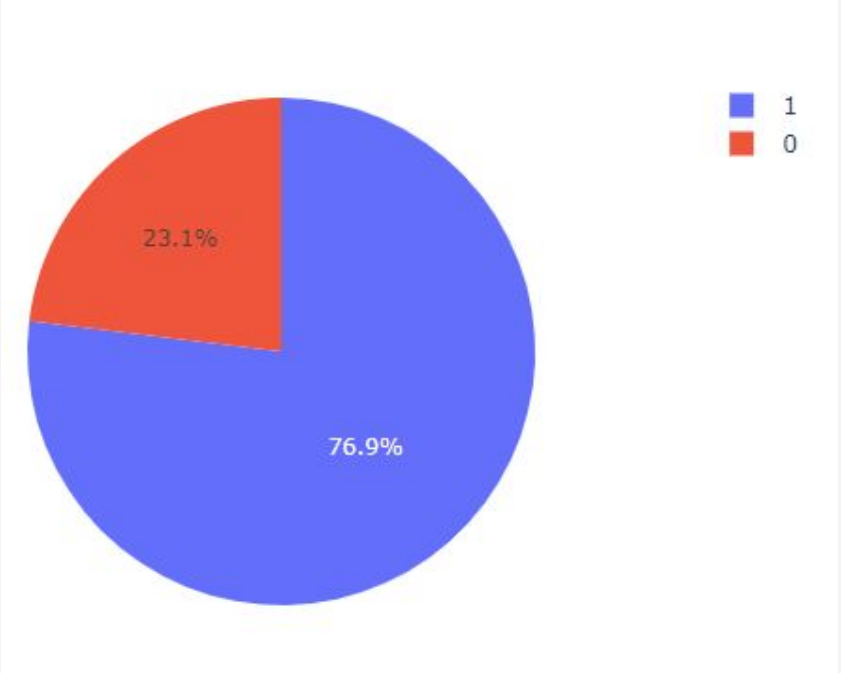
Success Launches by Location

- We Can recognize that KSC LC-39A has the most successful launches by 41.70%

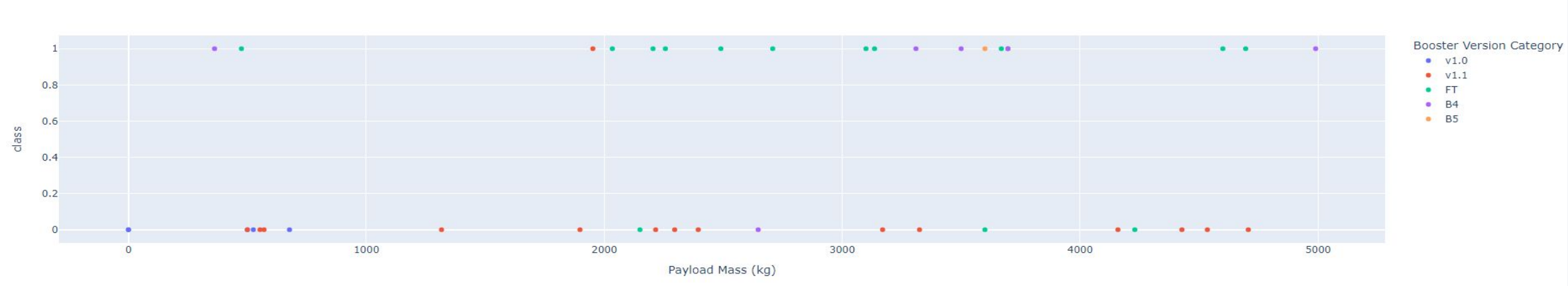


Success rate by Location

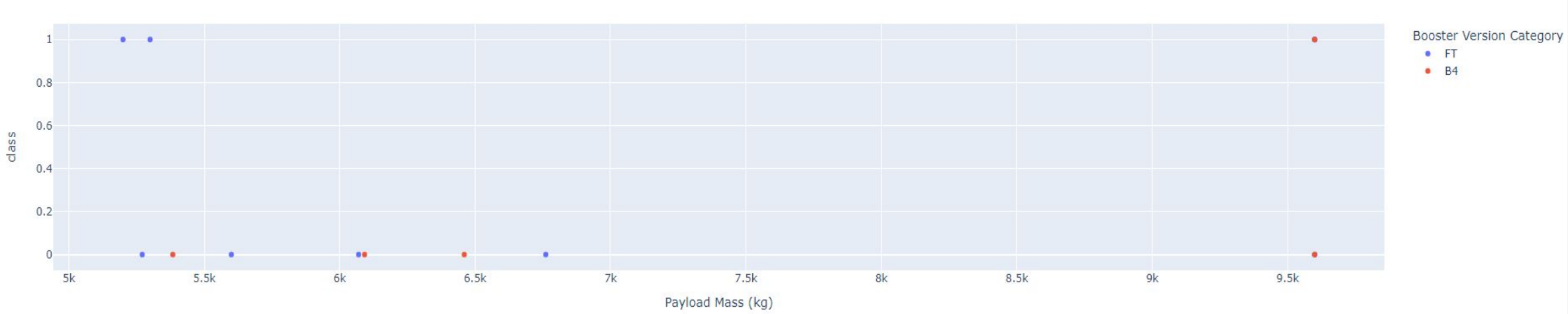
KSC LC-39A has the most successful launches by 41.70% with 23.10 failure rate



Payload vs Launches Outcome



Payload from 0 kg to 5000kg



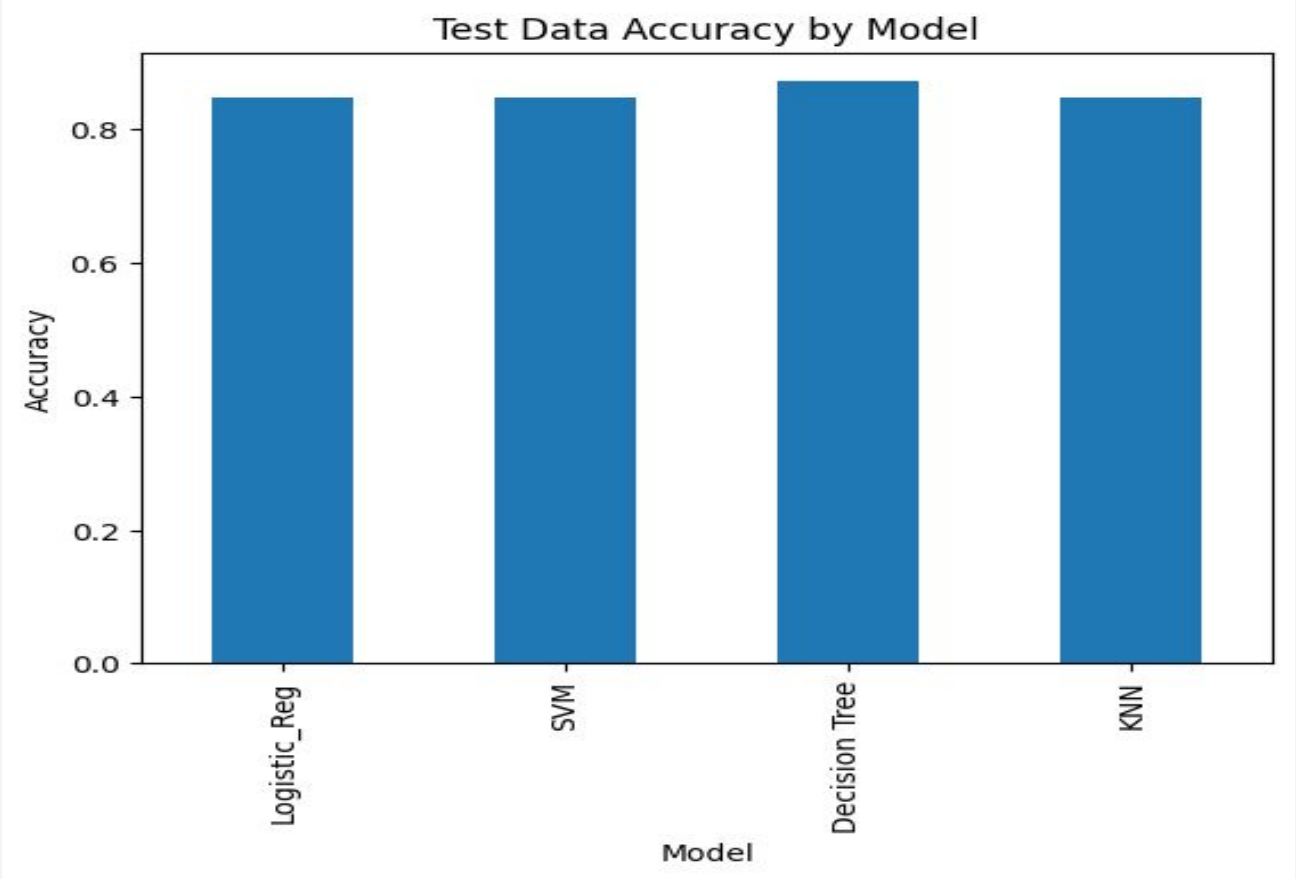
Payload from 5000 kg to 10,000kg

Section 5

Predictive Analysis (Classification)

Classification Accuracy

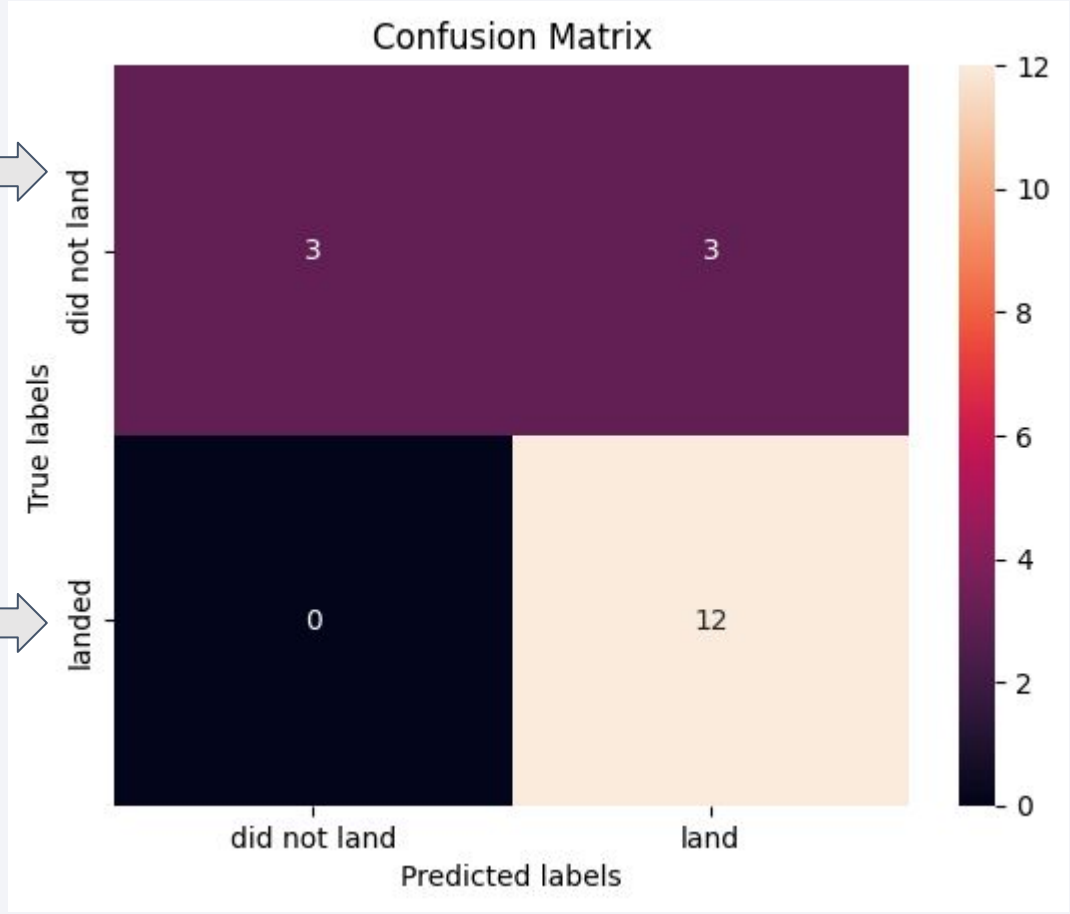
Decision Tree has the highest score



Confusion Matrix

True Negative
TN

False Negative
FN



False Positive
FP

True Positive
TP

Conclusions

- Decision Tree is the best model for prediction accuracy
- Low payload perform better than high payloads
- Success rate improve over the time
- KSC LC 39A is Most successful launch location

Appendix

All relevant assets Python code snippets, SQL queries, charts, Notebook outputs, or data sets you can find it in the following github repo link.

https://github.com/marwanbahgat/Applied-Data-Science-Capstone

Thank you!