IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sharif Noor Zisad
August 22, 2022

Outline

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

Executive Summary

- Summary of methodologies
- The goal of this assignment is to use a comprehensive data-driven analytic strategy to anticipate positive outcomes in the space race using SpaceX landing data.
- It entails data gathering, data manipulation, EDA with data visualization, EDA with SQL, and so on. Creating an interactive map using Folium, creating a dashboard with Plotly Dash, and classifying data with Predictive Analysis techniques

- Summary of all results
- The findings are reported as follows: exploratory data analysis findings, interactive analytics demo (with screenshots), and predictive analysis findings.

Introduction

- Project background and context
- In the age of commercial space flight, SpaceX says on its website that Falcon 9 rocket launches are 260+% cheaper (62 million dollars against 165 million dollars for other providers), with much of the savings due to SpaceX's ability to reuse the first stage.

- Problems you want to find answers
- Will SpaceX successfully land its first stage?
- What are the factors that impact its ideal landing success?
- How can this information be used to potentially compete with SpaceX for rocket launches?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The information was gathered by issuing a get request to the SpaceX API and web scraping Falcon 9 and Falcon Heavy Launch Records from Wikipedia.
- Perform data wrangling
  - Dealing with missing values, adding additional columns, removing unnecessary columns, and visualizing with Panda's data frames
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Selected the highest performing model after evaluating four different classification models: logistic regression, tree, SVM, and KNN.

Data Collection

- The data is gathered through the SpaceX REST API, and the API will provide us with information regarding launches such as the rocket utilized, payload delivered, launch specs, landing specifications, and landing outcome.
- BeautifulSoup is used to get more data from Wikipedia.

Data Collection – SpaceX API

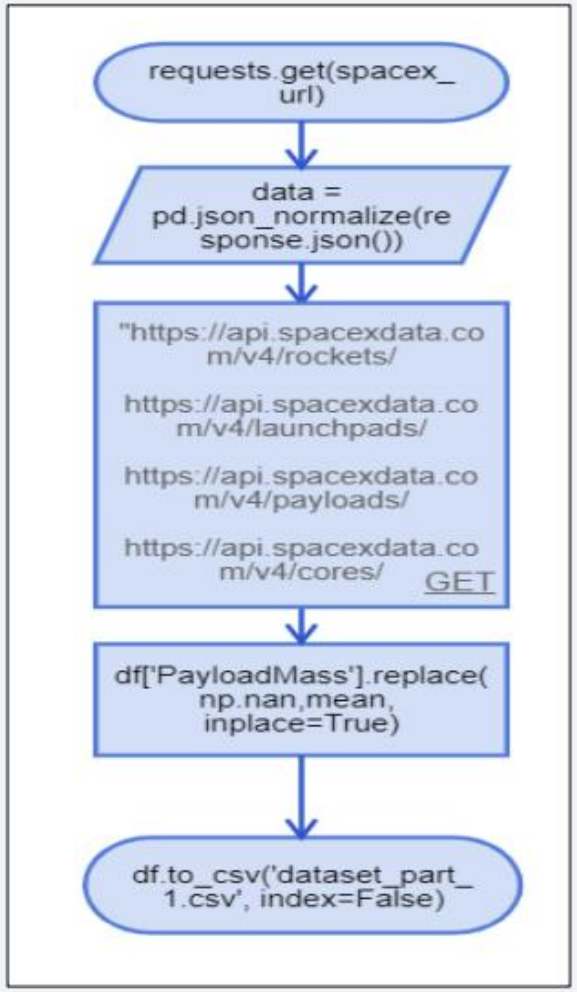
Request and parse the SpaceX launch data using the GET request

Decode the response content as a Json using json () and turn it into a Pandas dataframe using json_normalize

Use the API again to get information about the launches using the IDs given for each launch. Specifically, will be using columns

- Rocket
- Payloads
- Launchpad
- Cores

The mean and the replace () function to replace np.nan ` values in the data with the mean calculated



Github: Data Collections using Spacex API

Data Collection - Scraping

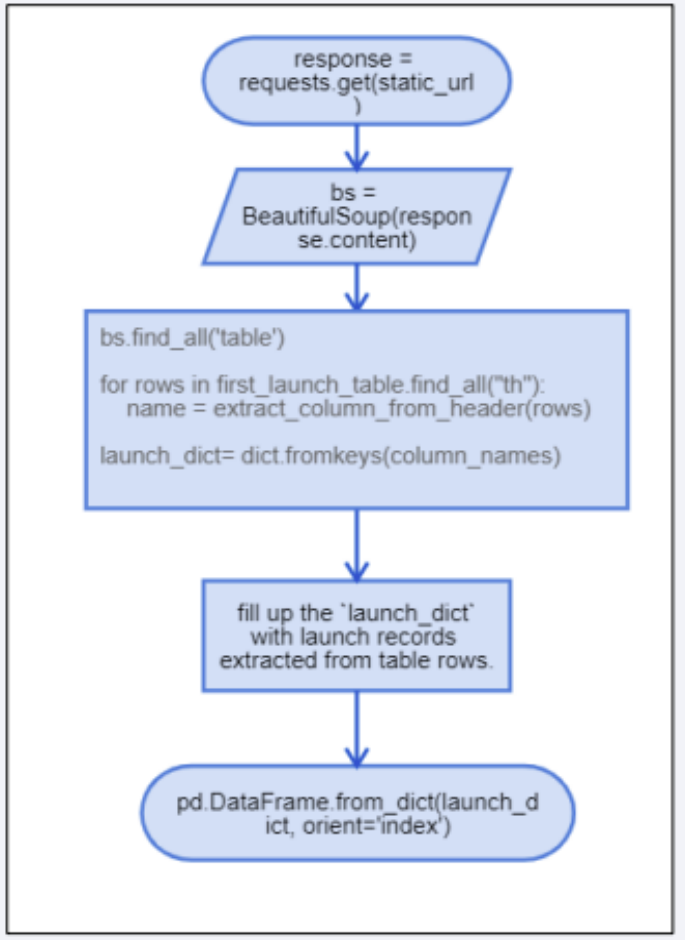
Extract a Falcon 9 launch records HTML table from Wikipedia

Parse the table and convert it into a Pandas data frame

TASK 1: Request the Falcon9 Launch Wiki page from its URL

TASK 2: Extract all column/variable names from the HTML table header

TASK 3: Create a data frame by parsing the launch HTML tables



Github: Data Collections using Webscraping

Data Wrangling

Data was processed using pandas data frame features

Data Wrangling process contains 5 steps for understanding data

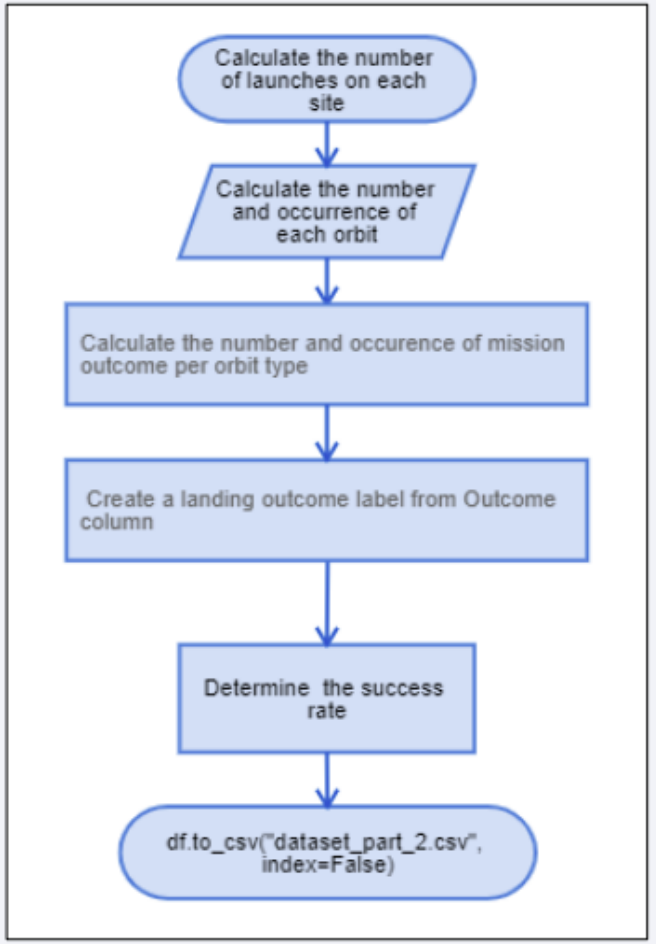
1) Data Analysis

2) Calculate number of launch on each site

3) Calculate the number of occurrence of each orbit

4) Calculate the number and occurrence of mission outcome per orbit type

5) Create a landing outcome label from Outcome column



Github: Data Wrangling

EDA with Data Visualization

- Scatter Graphs aid in data visualization by displaying data patterns and identifying correlations between variables. Scatter plots are made up of a broader body of data, such as Fight Number vs. Payload Mass, Fight Number vs. Launch Site, Payload vs. Launch Site, Orbit vs. Fight Number, Payload vs. Orbit Type, Orbit vs. Payload Mass, and Orbit vs. Payload Mass.
- The Bar Graph allows you to quickly compare data sets from various groupings. On one axis, the graph shows categories, and on the other, a discrete value. The objective is to demonstrate the link between the two axes. Bar charts may also demonstrate significant changes in data over time. Orbit vs. Mean
- Line graphs are valuable because they clearly depict data variables and patterns and can aid in making predictions about the outcomes of data that has not yet been captured.

Github: Data Visualization

EDA with SQL

- Display the names of the space mission's distinct launch locations.
- Show 5 records where the launch locations start with the string 'CCA.'
- Display the total payload mass carried by NASA-launched rockets (CRS).
- Display the average payload mass carried by the F9 v1.1 booster variant.
- List the date of the first successful landing outcome in the ground pad.
- List the names of boosters that have been successful in drone ships with payload masses more than 4000 but less than 6000.
- Count the total number of successful and unsuccessful mission results.
- List the names of the rocket variants that carried the most payload mass. Make use of a subquery.

Github: EDA with SQL

Build an Interactive Map with Folium

- All launch sites were graphically marked on a map with latitude and longitude coordinates circled.
- Marked the successful and unsuccessful launches for each location, then allocated the data frame launch outcomes (failures, successes) to classes 0 and 1 on the map using Green and Red markers in a MarkerCluster ()
- The distance from the launch point to several land markers was calculated. On the map, lines are drawn to indicate the distance between landmarks.

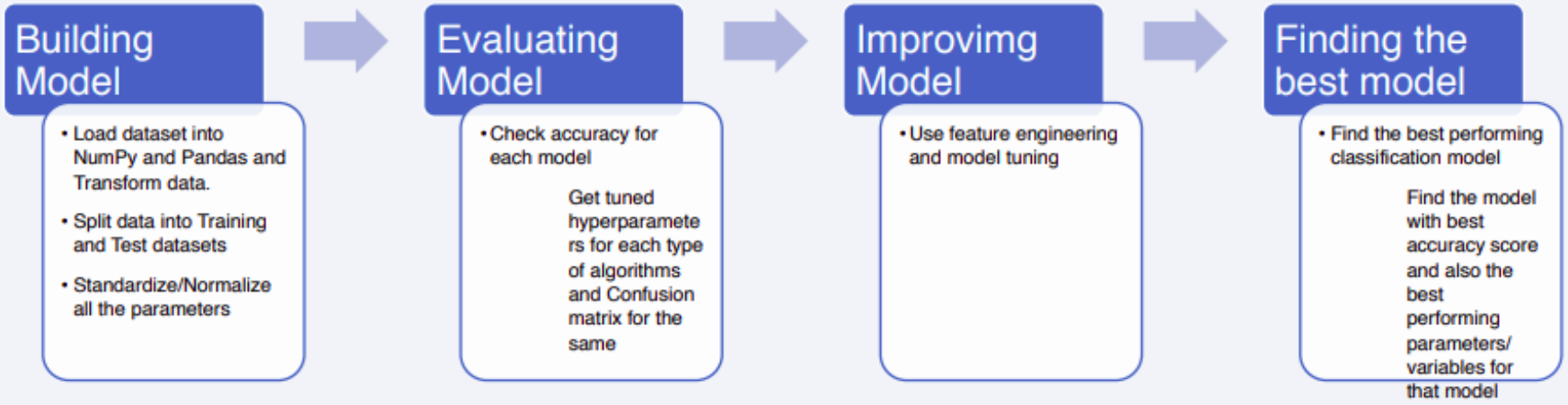
Github: Map with Folium

Build a Dashboard with Plotly Dash

- Plotly is a Python wrapper for the JavaScript library "leaflet." It allows us to interact with our data visualizations and host them on the web:
- The pie chart displays the number of launches from each launch location, as well as the number of successful and unsuccessful launches from those sites.
- Callback function with input'site dropdown' and output'success pie chart'
- Callback function with inputs'site dropdown' and 'payload slider' and output'success payload scatter chart'
- Scatter Graph: Relationship between launch success (Outcome) and payload (in kg) for various booster models.

Github: Plotly Dash

Predictive Analysis (Classification)

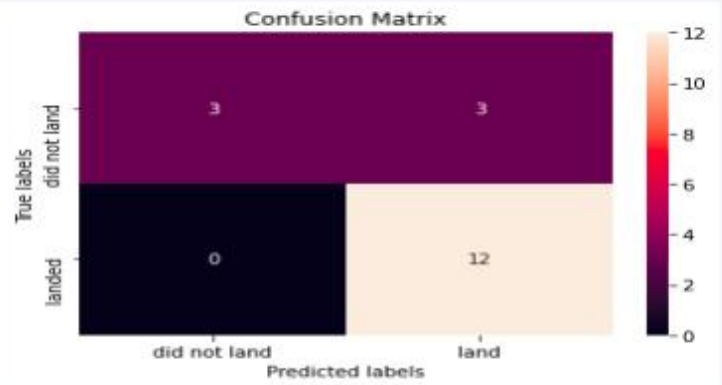


Github: Predictive Analysis

Results

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

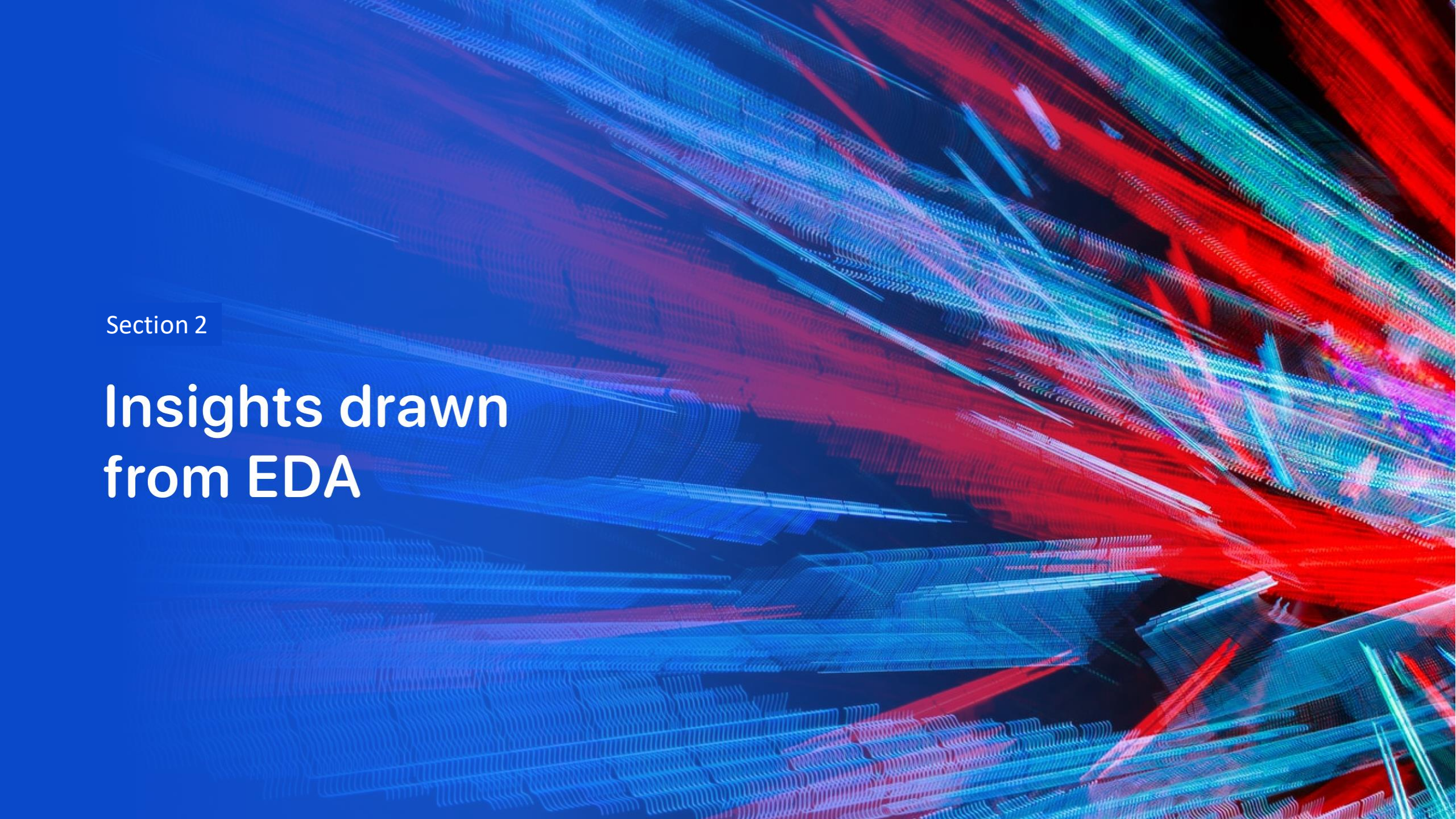
landing__outcome	COUNT
Success (drone ship)	2
Success (ground pad)	2
Failure (drone ship)	1
No attempt	1



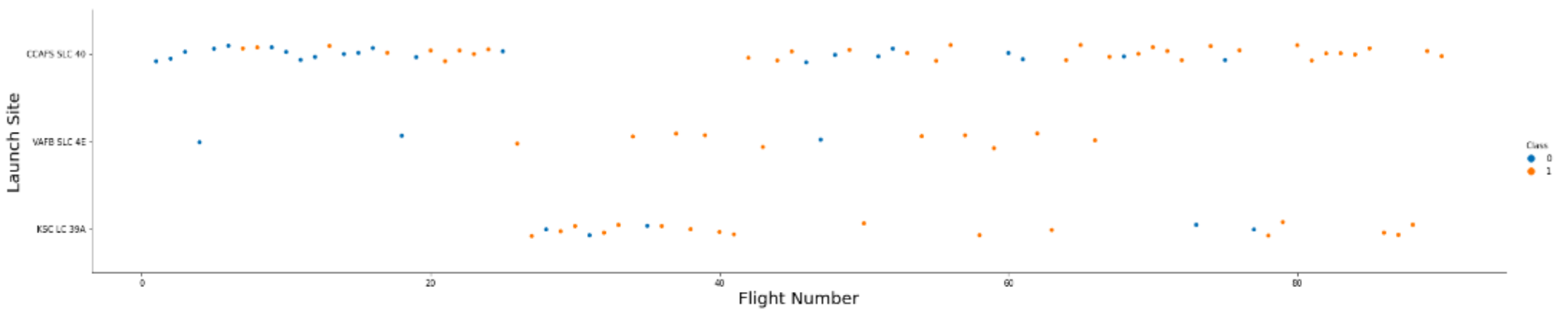

```
Accuracy for Logistics Regression method: 0.8333333333333334
Accuracy for Support Vector Machine method: 0.8333333333333334
Accuracy for Decision tree method: 0.6111111111111112
Accuracy for K nearsdt neighbors method: 0.8333333333333334
```

Section 2

Insights drawn from EDA

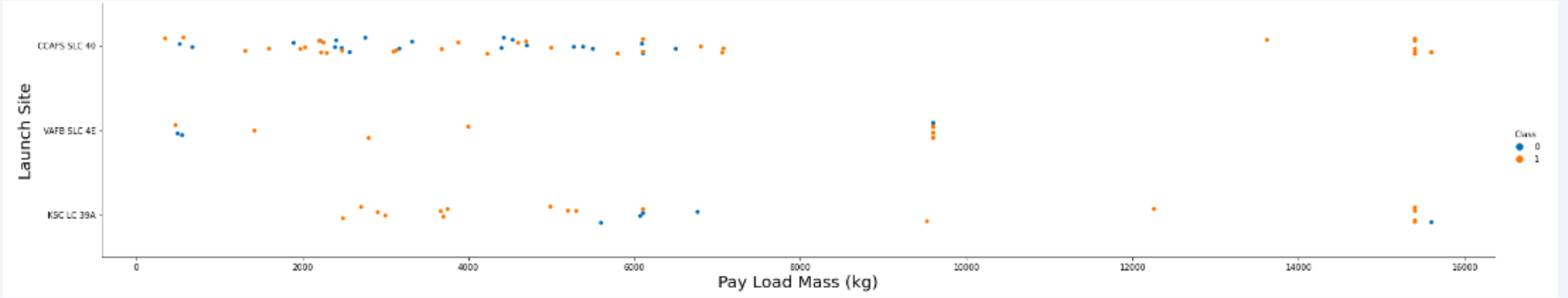


Flight Number vs. Launch Site



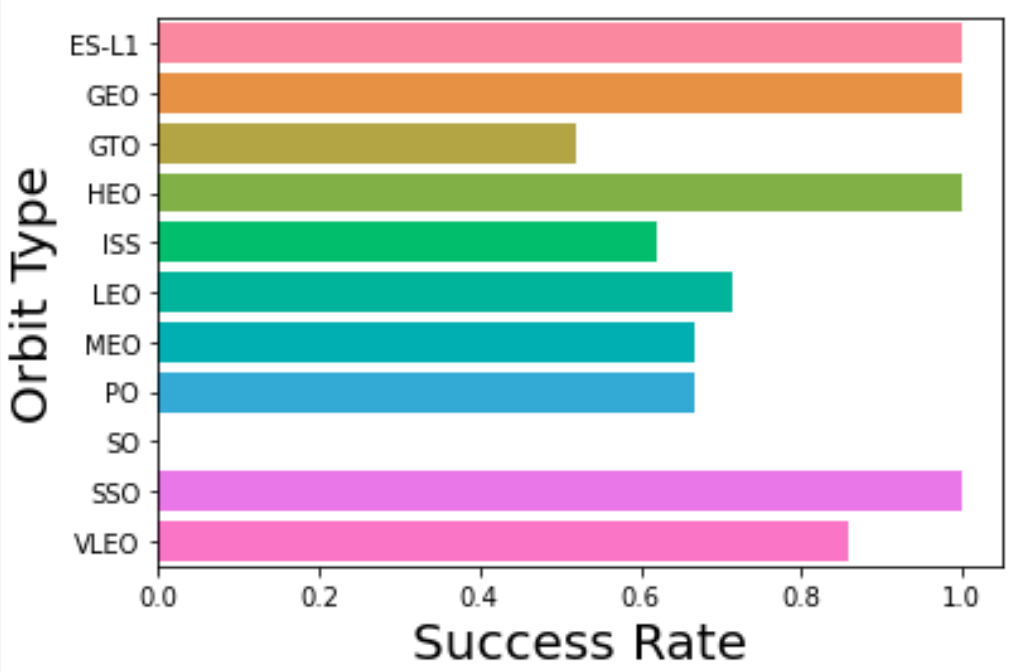
- Depicts the total launches by flight and launch site. The CCAFS SLC 40 has most launches across all flight numbers and have most failed launches in lower flight numbers and reduces as flight number increases

Payload vs. Launch Site



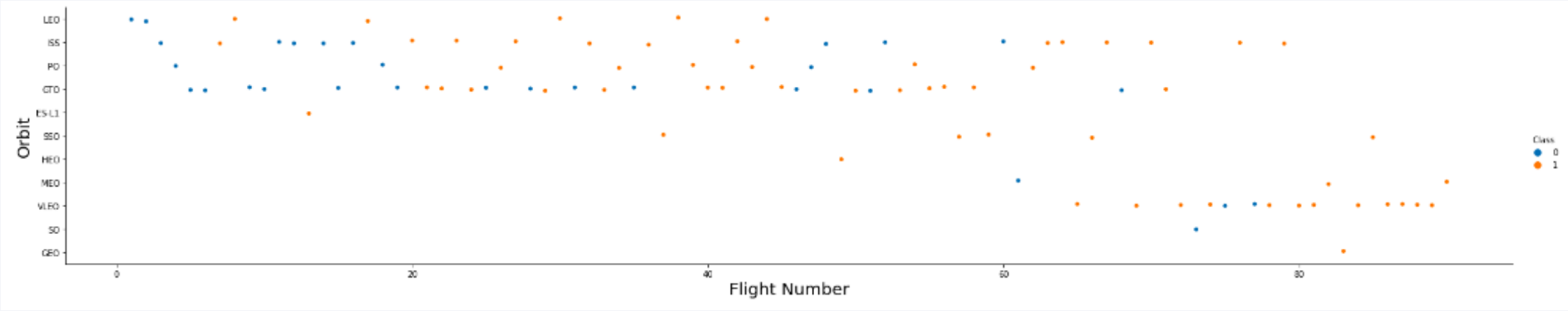
- There will be no VAFB SLC 4E launches that exceed 10000 PayloadMass. There were no failure launches in the 8000-15000 PayloadMass range, and just one failed KSC LC 39A launch in the 8000-16000 range.

Success Rate vs. Orbit Type



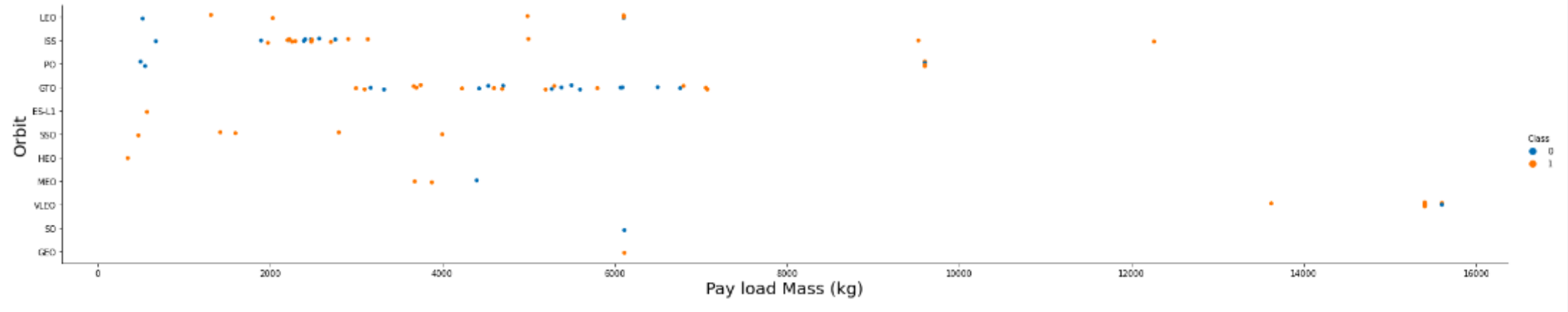
- ES-L1, GEO, HEO and SSO have 100% Success Rate

Flight Number vs. Orbit Type



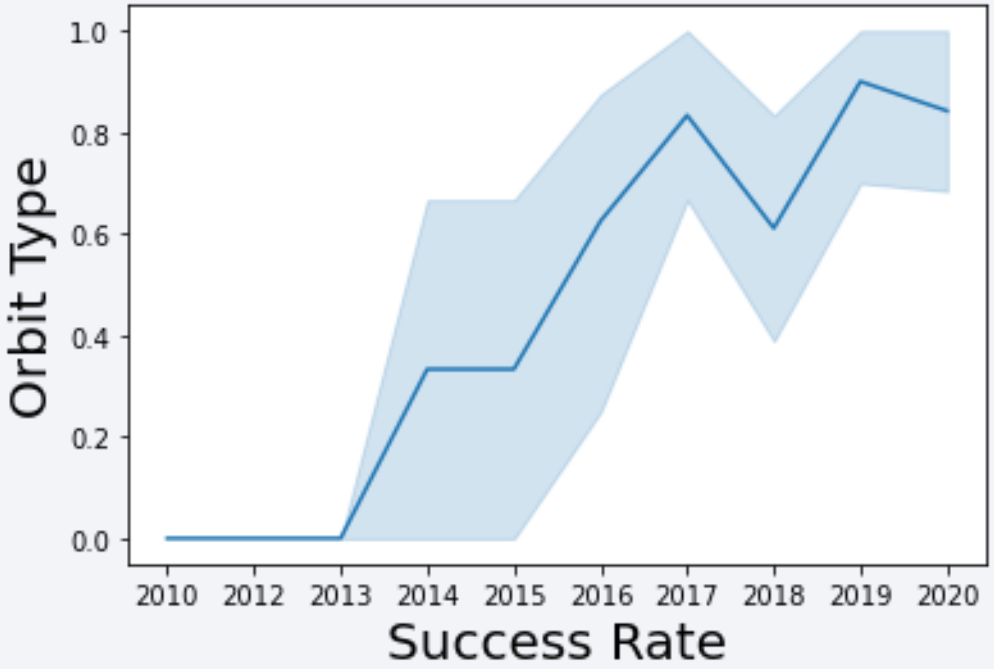
- LEO, SSO, HEO, and GEO orbits appear to have a direct association to success, however GTO orbit appears to have failures at all Flight Numbers.

Payload vs. Orbit Type



- Success by PayloadMass appears to have a clear association with LEO, ISS, PO, and SSO. While SSO has a high success rate in the lower PayloadMass category, LEO, ISS, and PO have a greater success rate in the higher PayloadMass category.

Launch Success Yearly Trend



- Overall, the Success Rate has been steadily increasing between 2013 and 2020. (despite a flat 2014 and a small dip in 2018)

All Launch Site Names

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

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

- Using SQL Query, select unique launch locations from the LAUNCH SITE column of the SpaceX launches dataset.

Launch Site Names Begin with 'CCA'

```
%%sql
SELECT *
FROM SPACEXTBL
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5;
```

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcom
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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attemp
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attemp
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attemp

- Selecting 5 records from SpaceX dataset with Launch Site names starting with CCA

Total Payload Mass

45596

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

- Calculating total of all booster's 'Payload Mass' from SpaceX dataset launches by 'NASA (CRS)' customer

Average Payload Mass by F9 v1.1

2534

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

- Calculating average payload mass of Booster version F9 v1.1 only

First Successful Ground Landing Date

2015-12-22

```
%%sql
SELECT MIN(DATE)
FROM SPACEXTBL
WHERE LANDING__OUTCOME = 'Success (ground pad)';
```

- Calculating first successful ground landing date using MIN function on landing date column

Successful Drone Ship Landing with Payload between 4000 and 6000

booster_version	landing__outcome	payload_mass__kg_
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600

```
%%sql
SELECT DISTINCT(BOOSTER_VERSION), LANDING__OUTCOME, PAYLOAD_MASS__KG_
FROM SPACEXTBL
WHERE LANDING__OUTCOME = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

- Getting list of booster versions which have successful drone ship landing and their payload is in the range of 4000 to 6000 KG

Total Number of Successful and Failure Mission Outcomes

successful_missions
61

failure_missions
10

```
%%sql
SELECT COUNT(LANDING__OUTCOME) AS SUCCESSFUL_MISSIONS
FROM SPACEXTBL
WHERE LANDING__OUTCOME LIKE 'Success%';
```

```
%%sql
SELECT COUNT(LANDING__OUTCOME) AS FAILURE_MISSIONS
FROM SPACEXTBL
WHERE LANDING__OUTCOME LIKE 'Failure%';
```

- Getting count of total successful and failure missions from SpaceX dataset based on landing outcome name starting with 'Success' for successful missions and 'Failure'" for failure missions

Boosters Carried Maximum Payload

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

- Obtained boosters carrying maximum payload by first selecting maximum of payload from SpaceX dataset and then selecting the booster version, payload using a subquery (query inside a query)

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

2015 Launch Records

landing__outcome	booster_version	launch_site	date_year
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	2015
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	2015

```
%%sql
SELECT LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE, YEAR(DATE) AS DATE_YEAR
FROM SPACEXTBL
WHERE LANDING__OUTCOME = 'Failure (drone ship)' AND YEAR(DATE) = '2015'
```

- Get list of failed landing_outcome along with the booster version and launch site for the year 2015 from SpaceX table using WHERE clause

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

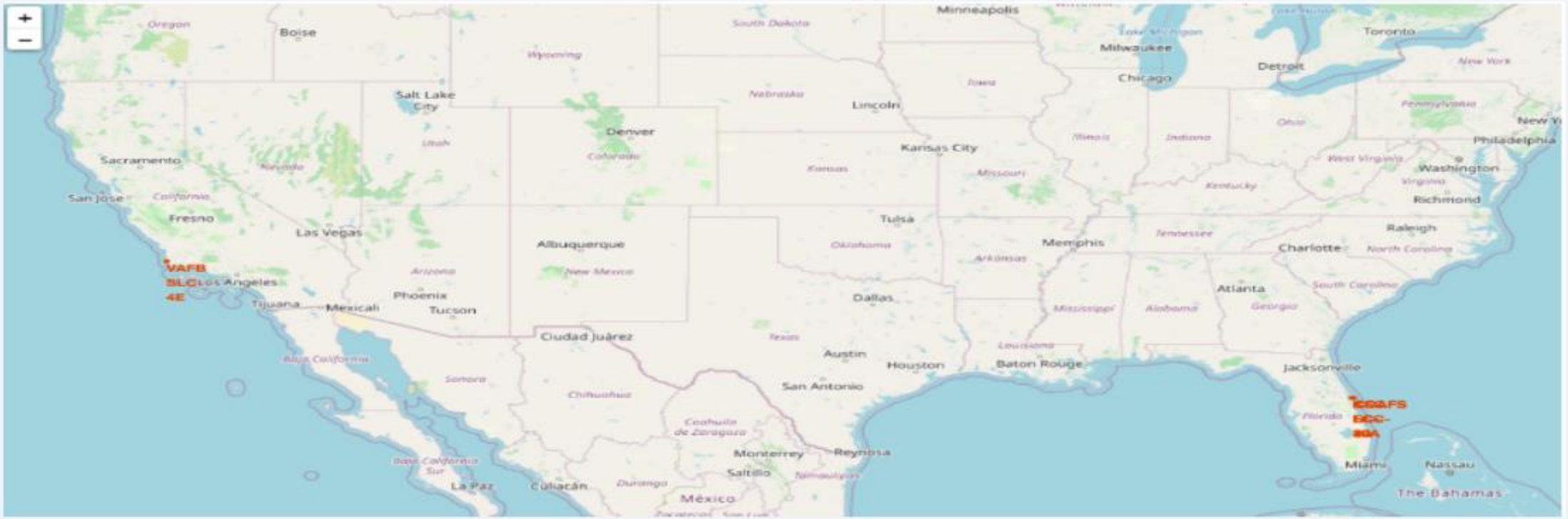
```
%%sql
SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS COUNT
FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04'  AND '2017-03-20'
GROUP BY LANDING__OUTCOME
ORDER BY COUNT DESC
```

- Obtained count of landing_outcome in descending order occurred between above 2 dates by GROUP BY and ORDER BY clauses

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

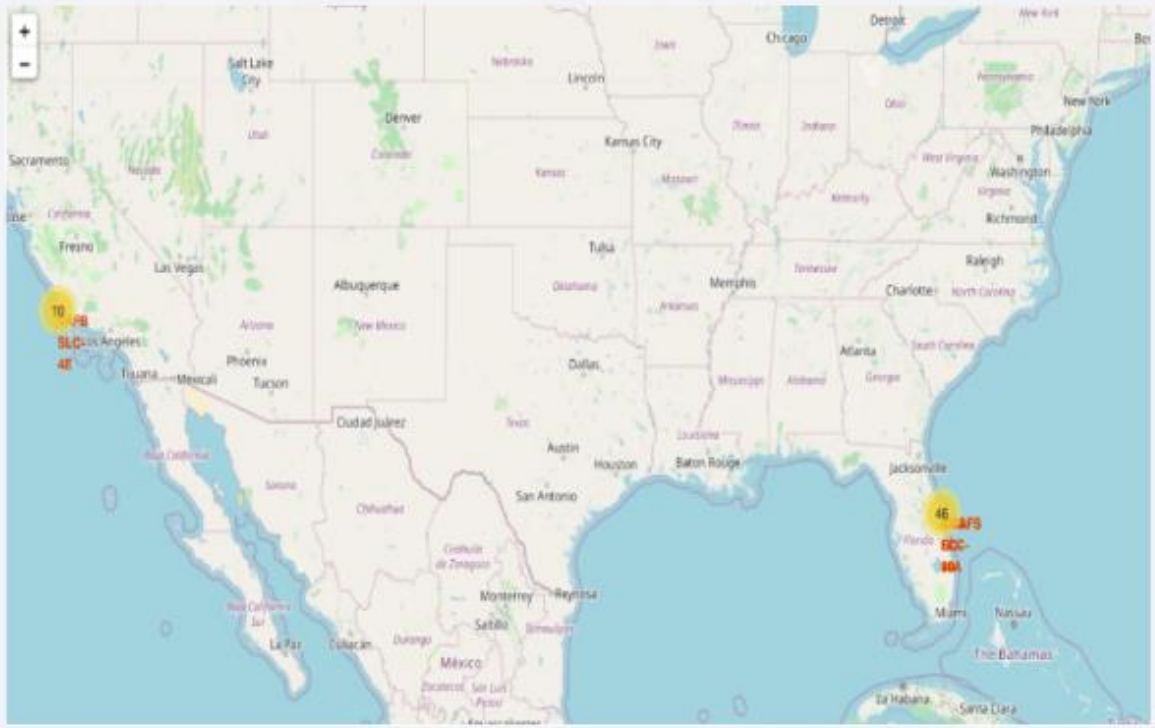

Section 3
Launch Sites
Proximities Analysis

Launch Sites On A Map



- SpaceX launch sites are in– Florida on the USA east coast and California on the USA west coast

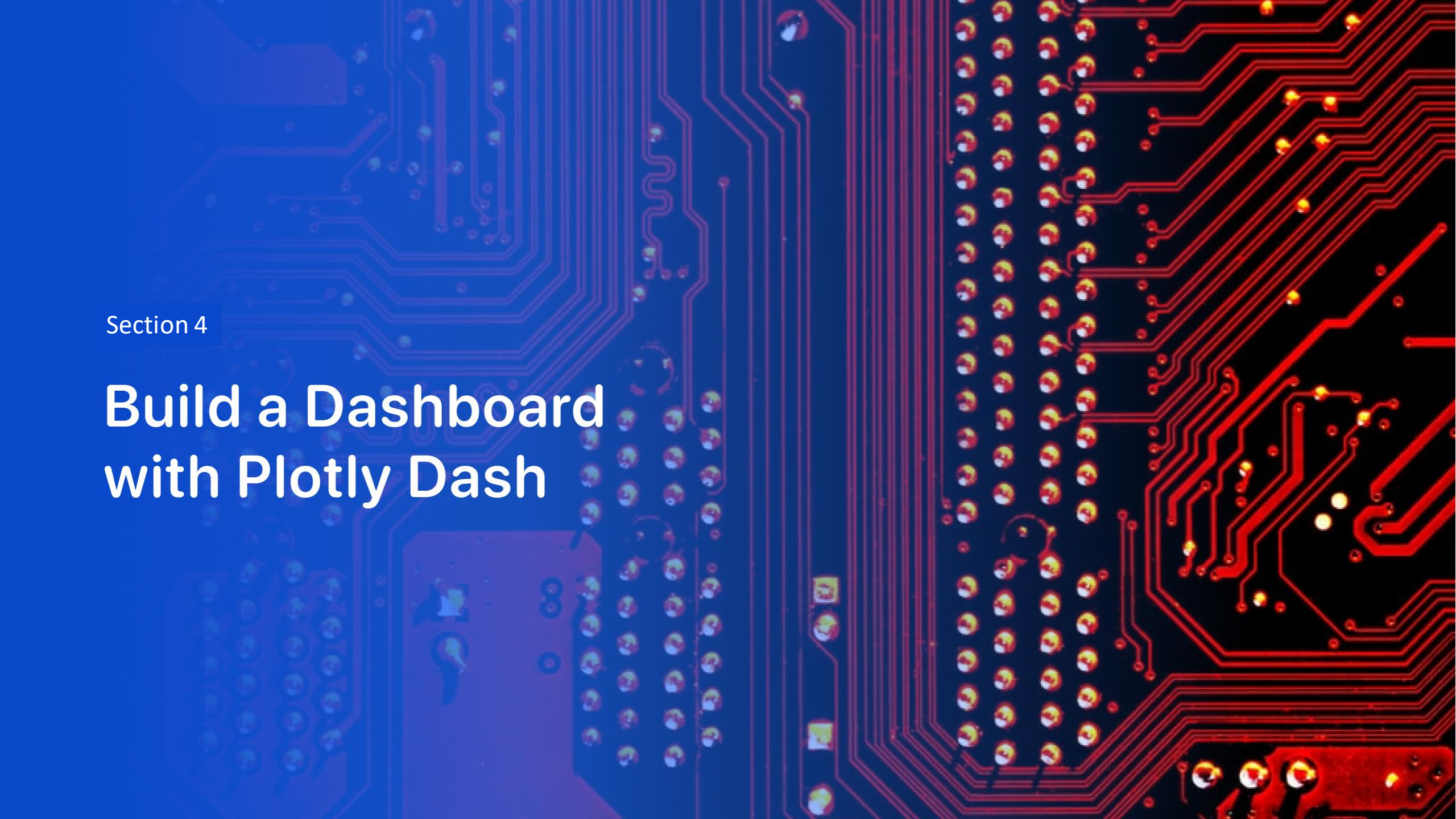
Sites And Launch Outcomes On A Map



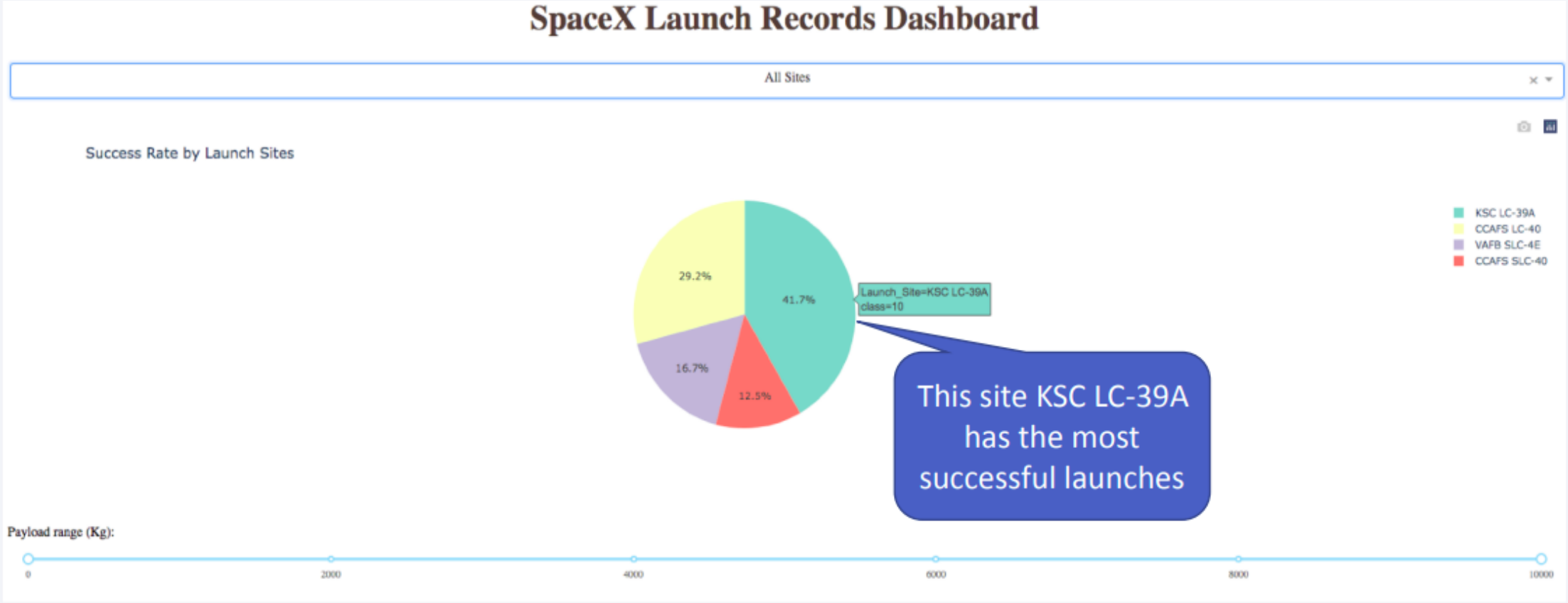
- Green colored are the successful launches and red colored are the failed launches

Launch Site Proximities On A Map

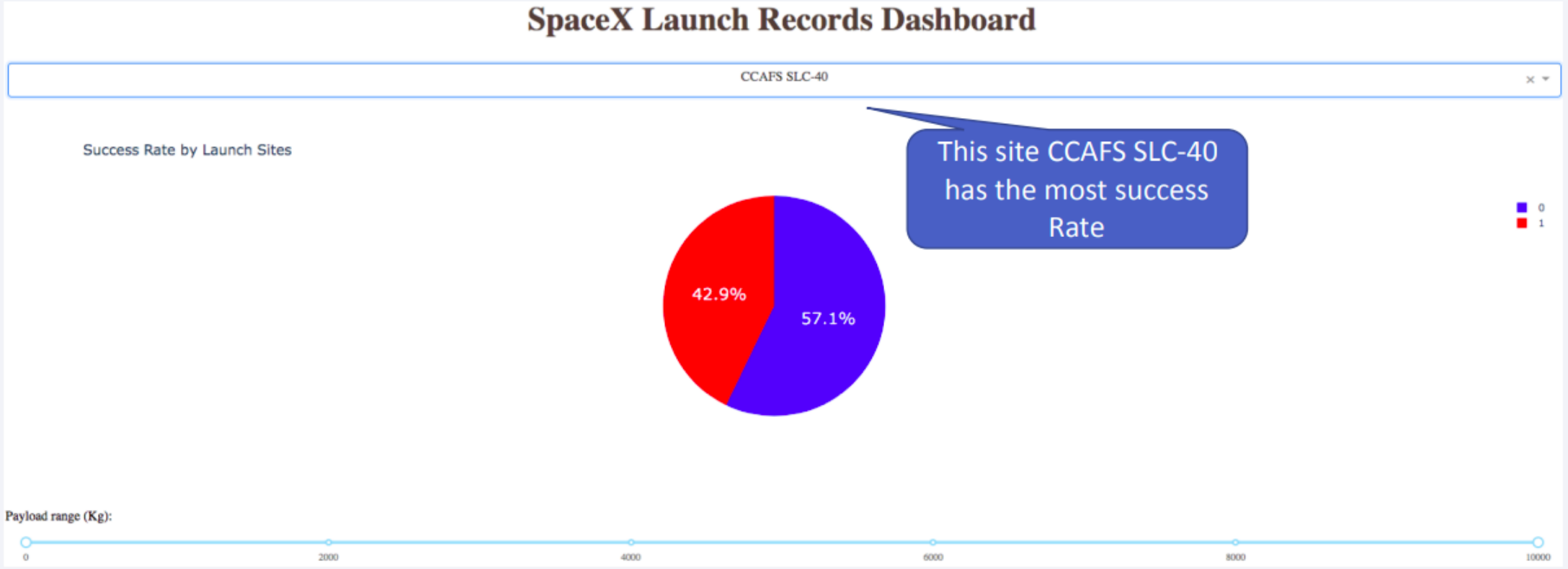
- Here, we can observe distance of launch sites from east coast, highways, key road, railway line visualized


Section 4
Build a Dashboard with Plotly Dash

Success Rate by Launch Sites



Most Successful Launch Site



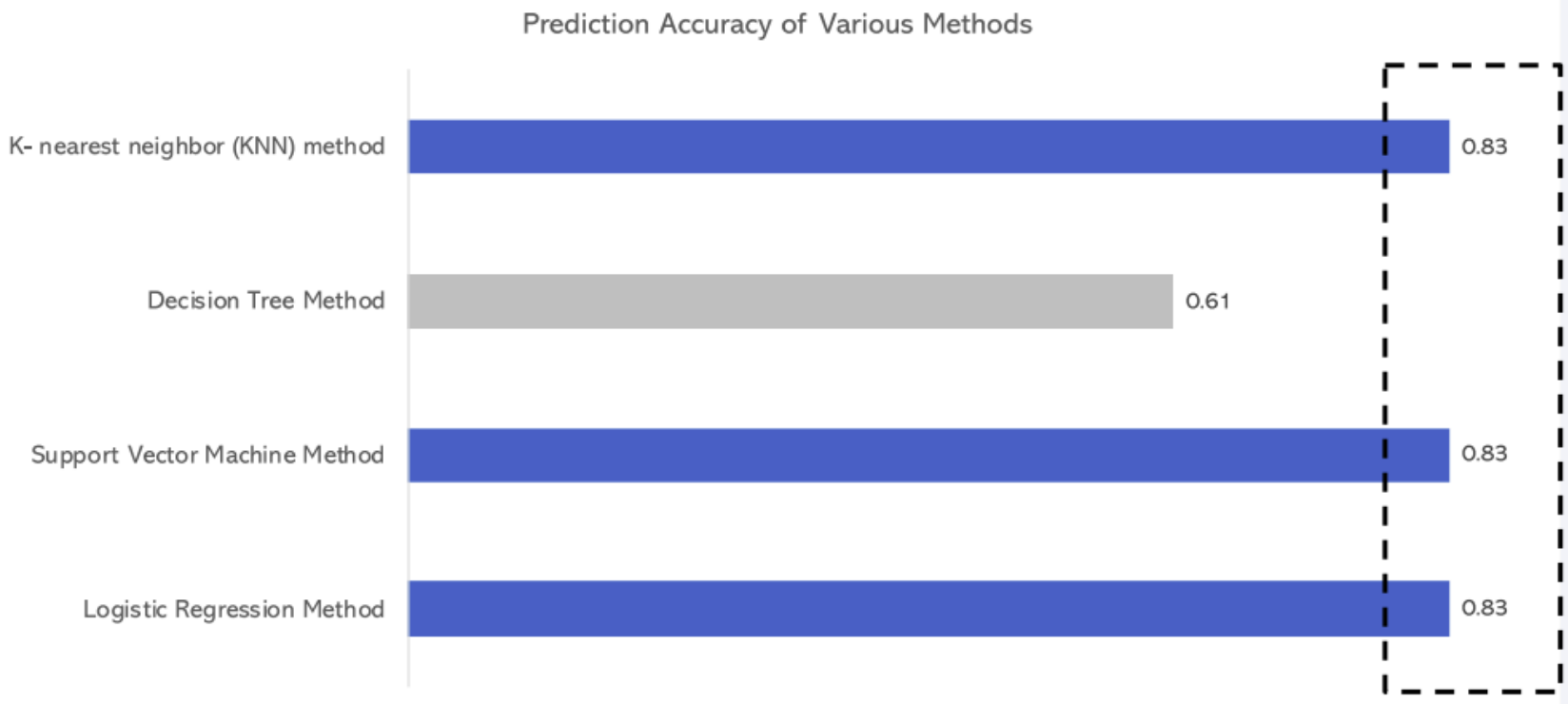
Success by Payload Mass & Booster Version

- Lower Payload launches (up to 6,000 kg) are more successful

Section 5

Predictive Analysis (Classification)

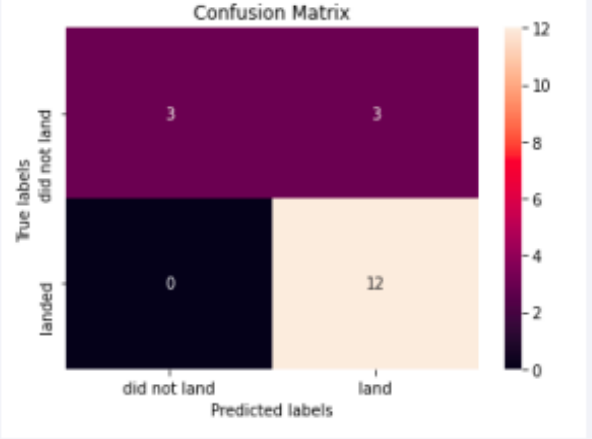
Classification Accuracy



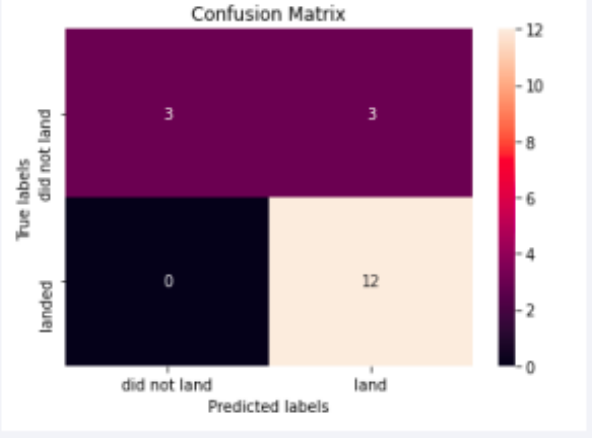
- KNN, Support Vector & Logistic Regression Methods have high accuracy

Confusion Matrix

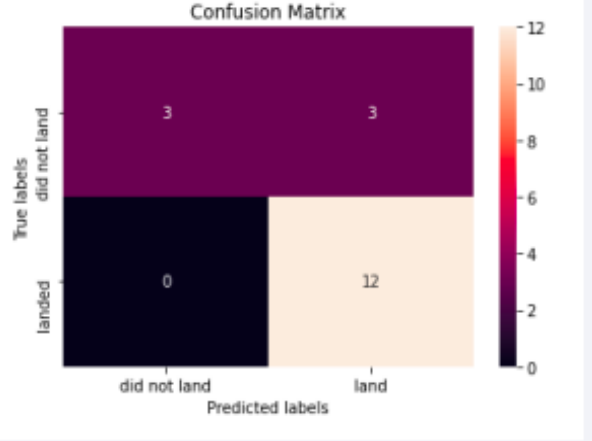
```
yhat = knn_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```




```
yhat=logreg_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```




```
yhat=svm_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```



- The above confusion matrix shows that all 3 models – KNN, Logistic Regression & SVM have highest true positives and least false negatives

Conclusions

- KNN, Logistic Regression and SVM are the best classifier models for this dataset
- The lower payload launches have higher success rate than heavier payloads
- Site KSC LC-39A has the most successful launches from all sites
- F9 Booster versions v1.0, v1.1, FT, B4, B5 have the highest launch success rates
- The SpaceX launches have been continuously getting better from year 2013 to 2020 based on data so they have the best chances for perfecting their launches in the future years

Thank you!