IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Niranjan Vaidya>
<March 2,2024>

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results

Executive Summary

To analyze and visualize the data, following methodologies were applied on the data:

- Data collection performed using web scraping and SpaceX api
- Data Wrangling
- Exploratory Data analysis with SQL
- Predictive Analysis
- Interactive map with folium and plotly dashboard

- Summary of all results
  - EDA analysis helped to understand what features are required for successful launchings
  - It also helped us to predict which factors are important for predicting best model

Introduction

- Project background and context
  - SpaceX, a rocket launch company can launch satellites at low prices because of ability to reuse the first stage. Predicting if first stage will land successfully will help us to determine the cost of launch which can be use if alternate company Space Y wants to big against Space X.
- Problems you want to find answers
  - With the previous data, predicting the probability of success of first stage landing for Falcon 9 rocket and what parameters affect the success of first stage landing
  - Determine the location for successful launches

Section 1

Methodology

Methodology

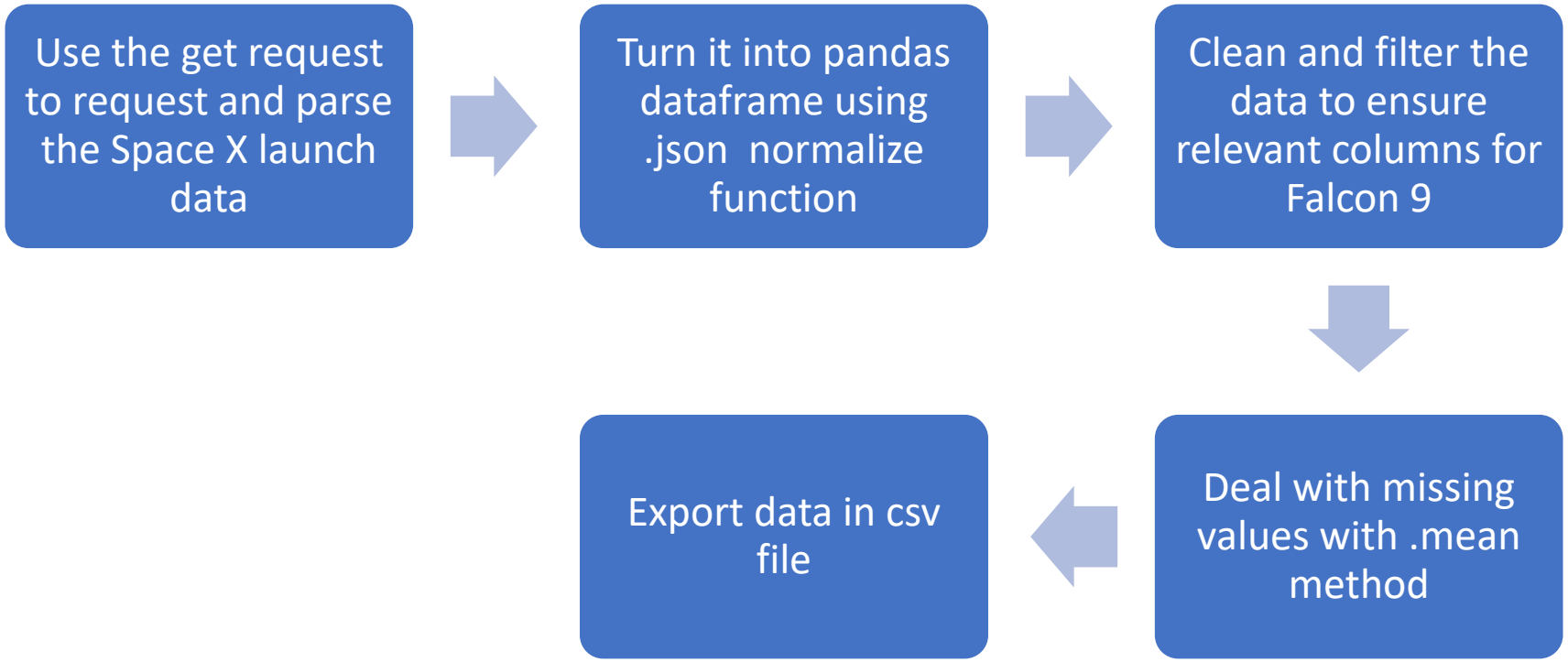
Executive Summary

- Data collection methodology:
  - Performed using SpaceX REST API and json function
  - Webscraping using beautiful soup function
- Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection

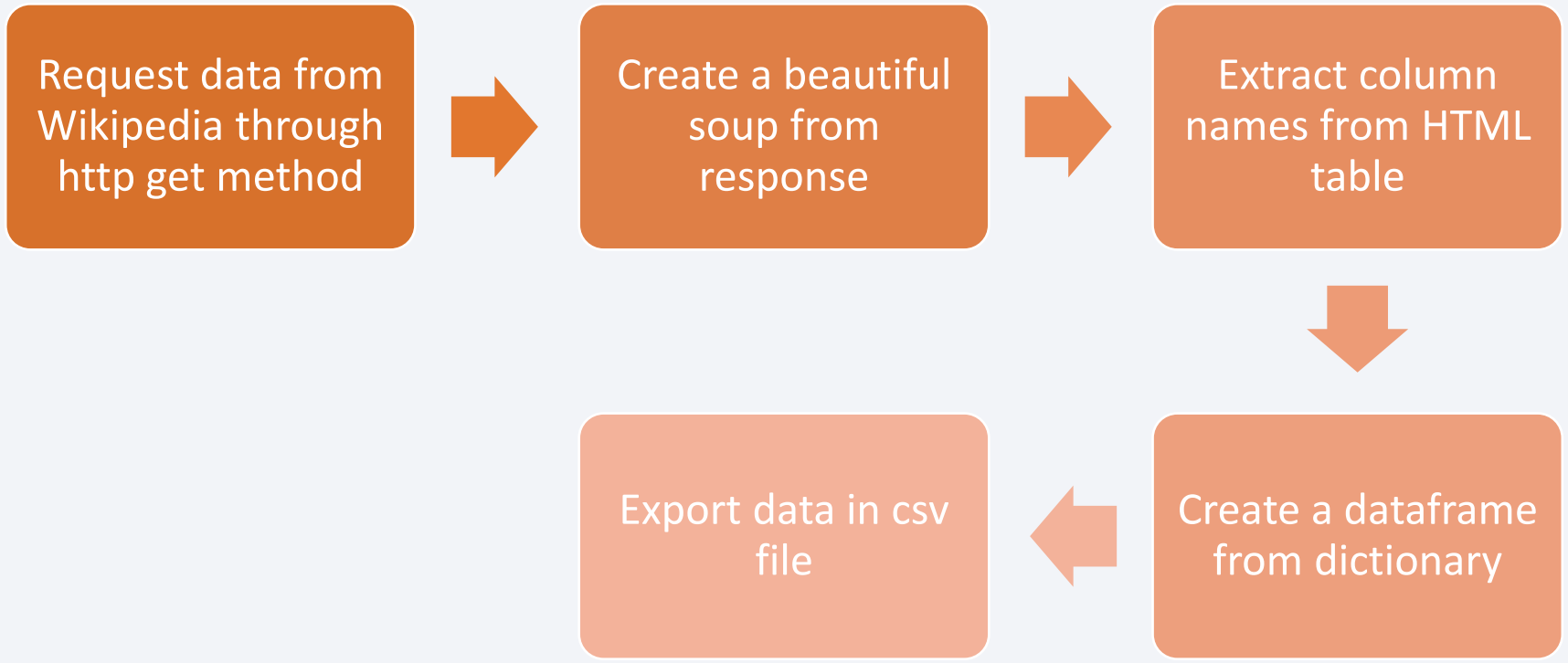
- Data sets were collected using Space X Rest API and from wiki page by using webscraping through beautiful soup function. Using pandas function, it is then transformed into dataframe.

Data Collection – SpaceX API



Link for the file –Data Collection API

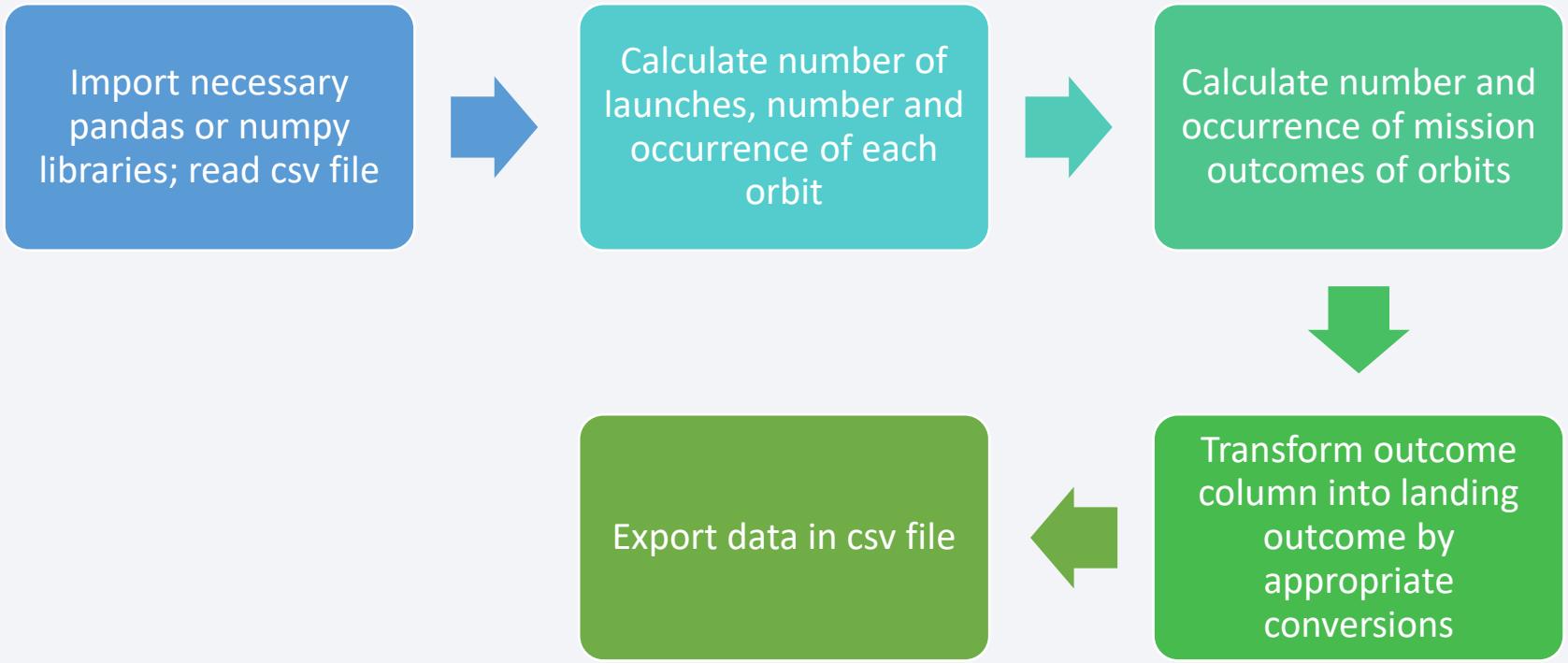
Data Collection - Scraping



Link for the file –Data Collection Webscraping

Data Wrangling

- Data Wrangling is a process of transforming raw data into data suitable for analysis. Some of the attributes are studied and column 'outcome' indicates if first stage was landed successfully. There are 8 outcomes, but our intent is to determine if landing was successful or not. Here landing outcomes is converted into 0 or 1, where 0 is bad outcome and 1 is good outcome based on certain conditions, where booster landed successfully.



Link for the file –Data Wrangling

EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts
- Charts that were plotted are as below:
  - Flight Number vs Payload Mass
  - Flight Number vs Launch Site
  - Payload Mass vs Launch Site
  - Orbit type vs Success Rate
  - Flight Number vs Orbit Type
  - Payload Mass vs Orbit Type
  - Success Rate year wise

Link for the file –EDA with Visualization

EDA with SQL

SQL queries performed are:

- Names of unique launch sites using ‘Select Distinct’ Query
- 5 records where launch sites begin with ‘CCA’ with ‘ using WHERE, LIKE and LIMIT’ operations in SQL
- Total payload mass with ‘ SUM’ query
- Averag payload mass with ‘AVG’ query
- Date when first successful landing outcome in ground pad was achieved with ‘Min(date) query
- Names of distinct boosters greater than 4000 and less than 6000
- Total number of successful and failure mission outcomes with ‘group by’ mission outcomes
- Distinct booster version which have carried max payload mass using subquery operation in SQL
- 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 using ‘group by and Order by Desc’ function in SQL

Link for the file –EDA with SQL

Build an Interactive Map with Folium

- Markers added for all launch sites on US Map with appropriate text
- Colors for success and failed launches
- Distance between one launch site to its nearest coast line

Link for the file –Folium Map

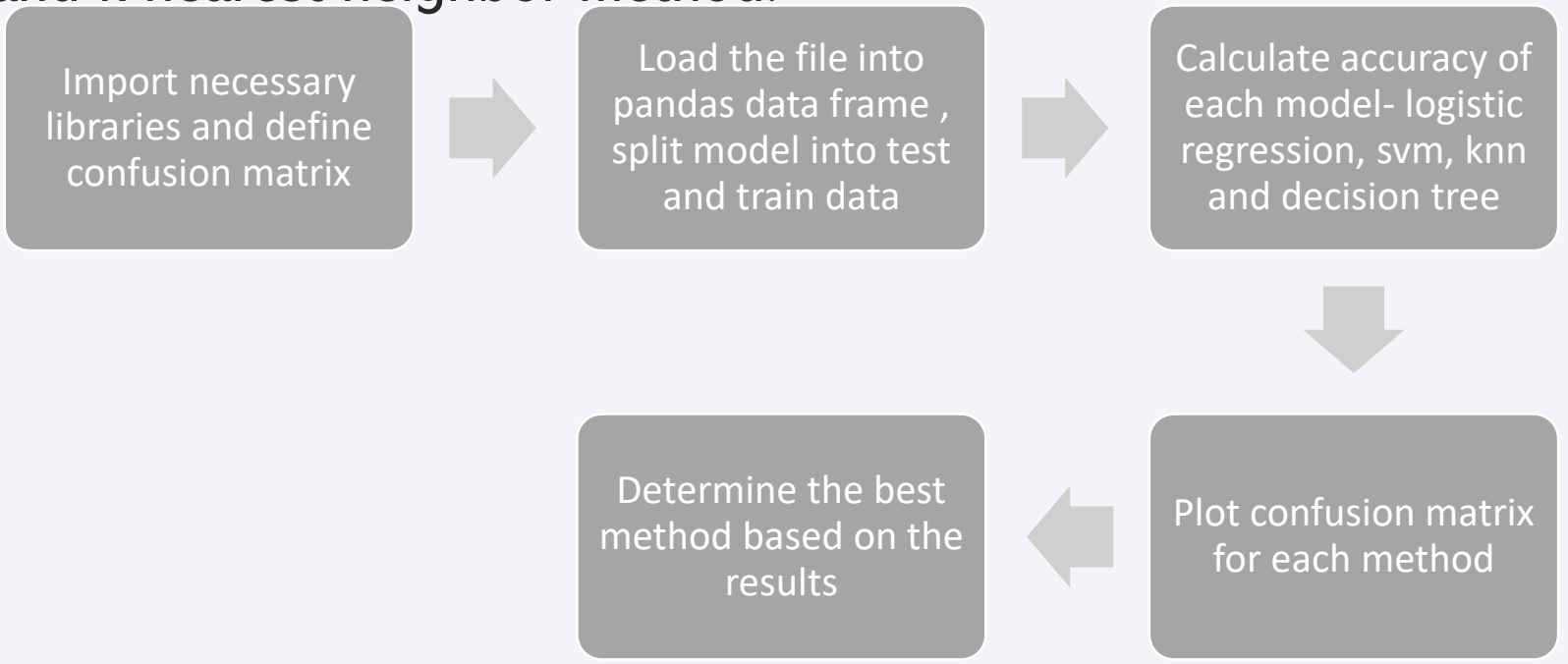
Build a Dashboard with Plotly Dash

- Dropdown to allow user to select a particular launch site to understand its success ratio
- Pie chart for all launch sites showing their proportion in overall successful launches
- Scatter chart for Payload Mass vs Success rate for different booster versions

Link for the file –Plotly Dash

Predictive Analysis (Classification)

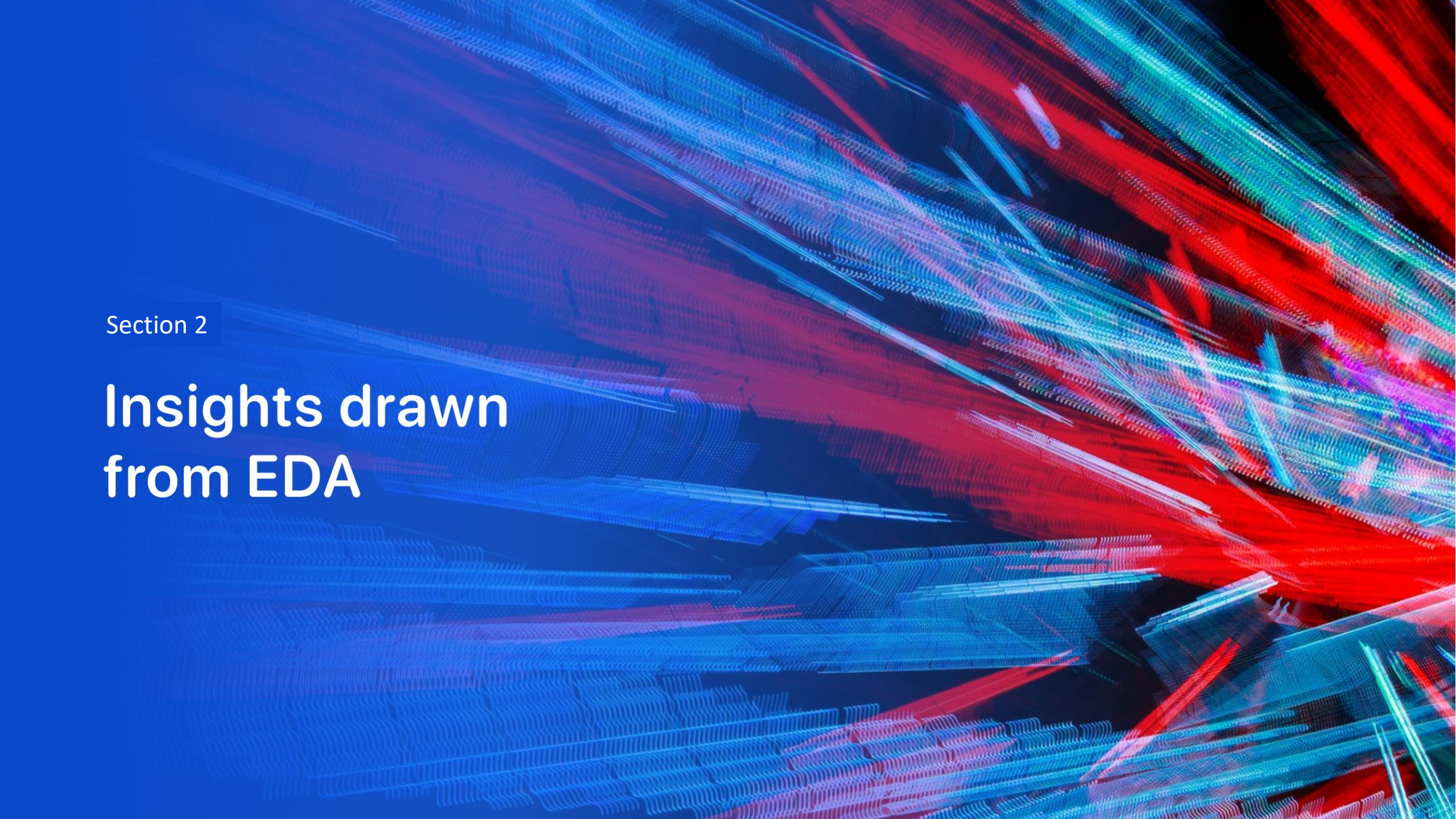
- Summarize how you built, evaluated, improved, and found the best performing classification model
- Four models were used to analyze the model and accuracy for each was determined. These are logistic regression, support vector machine, decision tree and k nearest neighbor method.



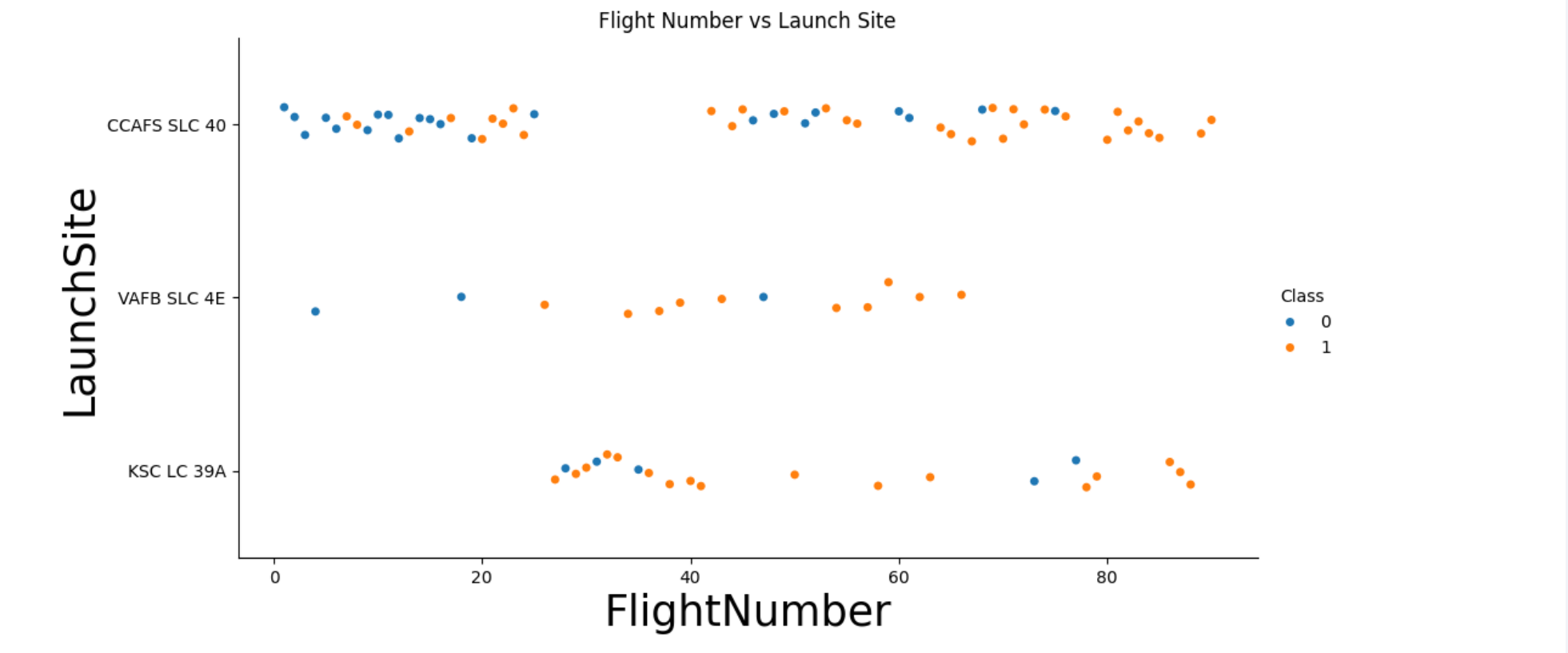
Link for the file –Predictive Analysis

Results

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

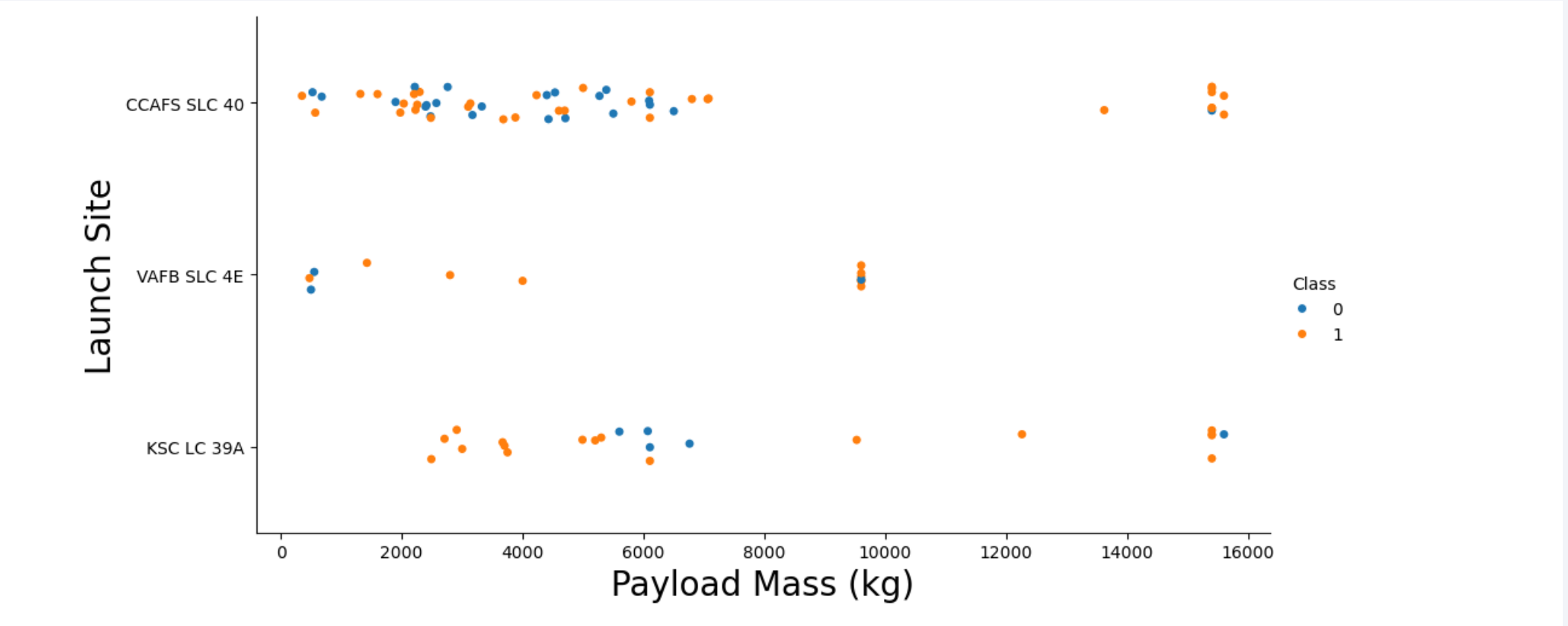

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- This scatter plot explains the relationship between flight number and different launch sites used for launching rockets
- It explains that early flights failed mostly as 0 class means failure.
- Also, site- CCAFS SLC 40 has more launches than any other site
- Recent launches after flight number 80 have all been successful (class 1)

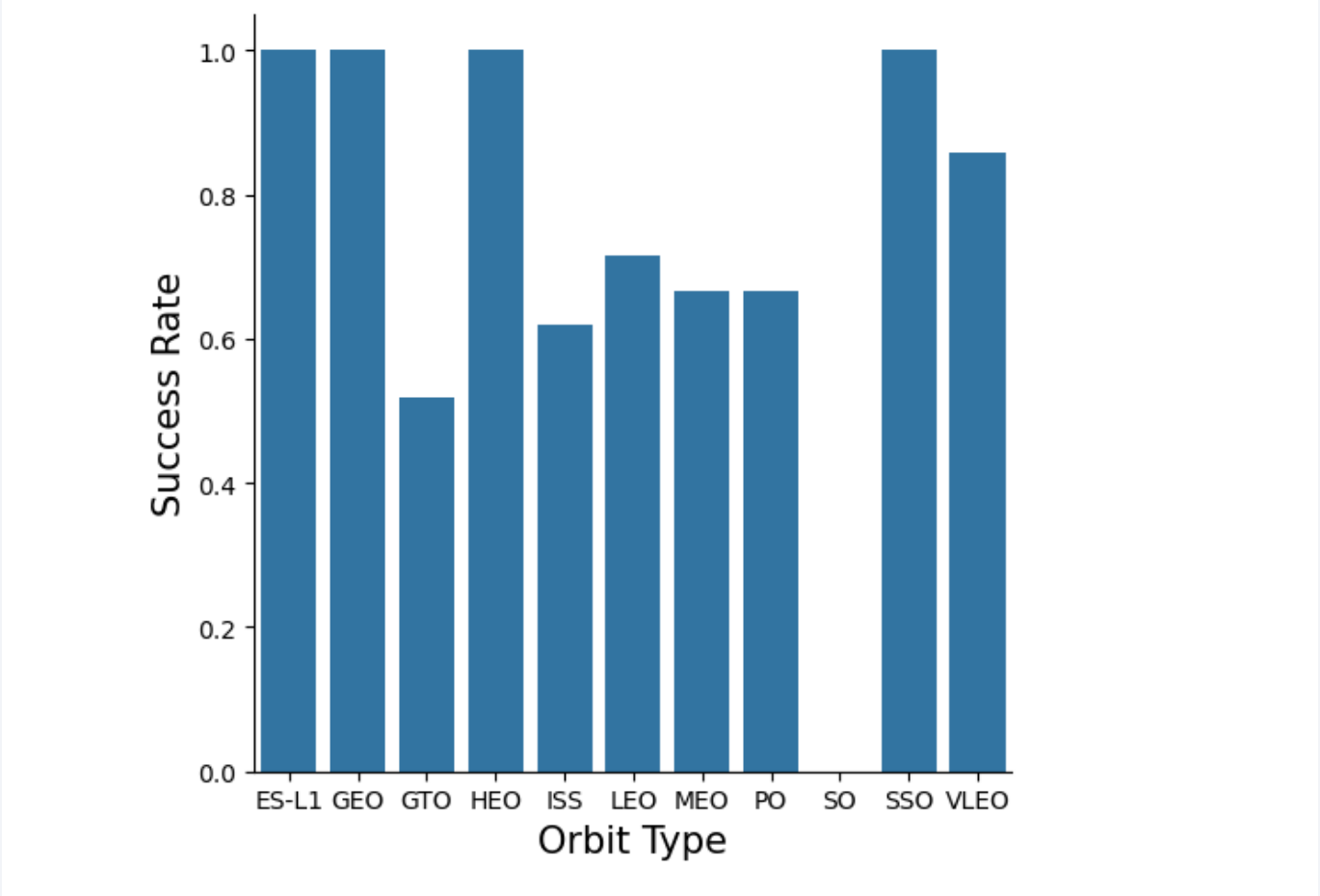
Payload vs. Launch Site



- This scatter plot explains the relationship between Payload Mass and different launch sites used for launching rockets
- It explains that majority of Payload are less than 6000 kg for different launch sites. Further, amount of payload launched from site-CCAFS SLC 40 far exceeds the payload launched from other two sites combined.
- Payload above 8000 have higher success rate than those below.

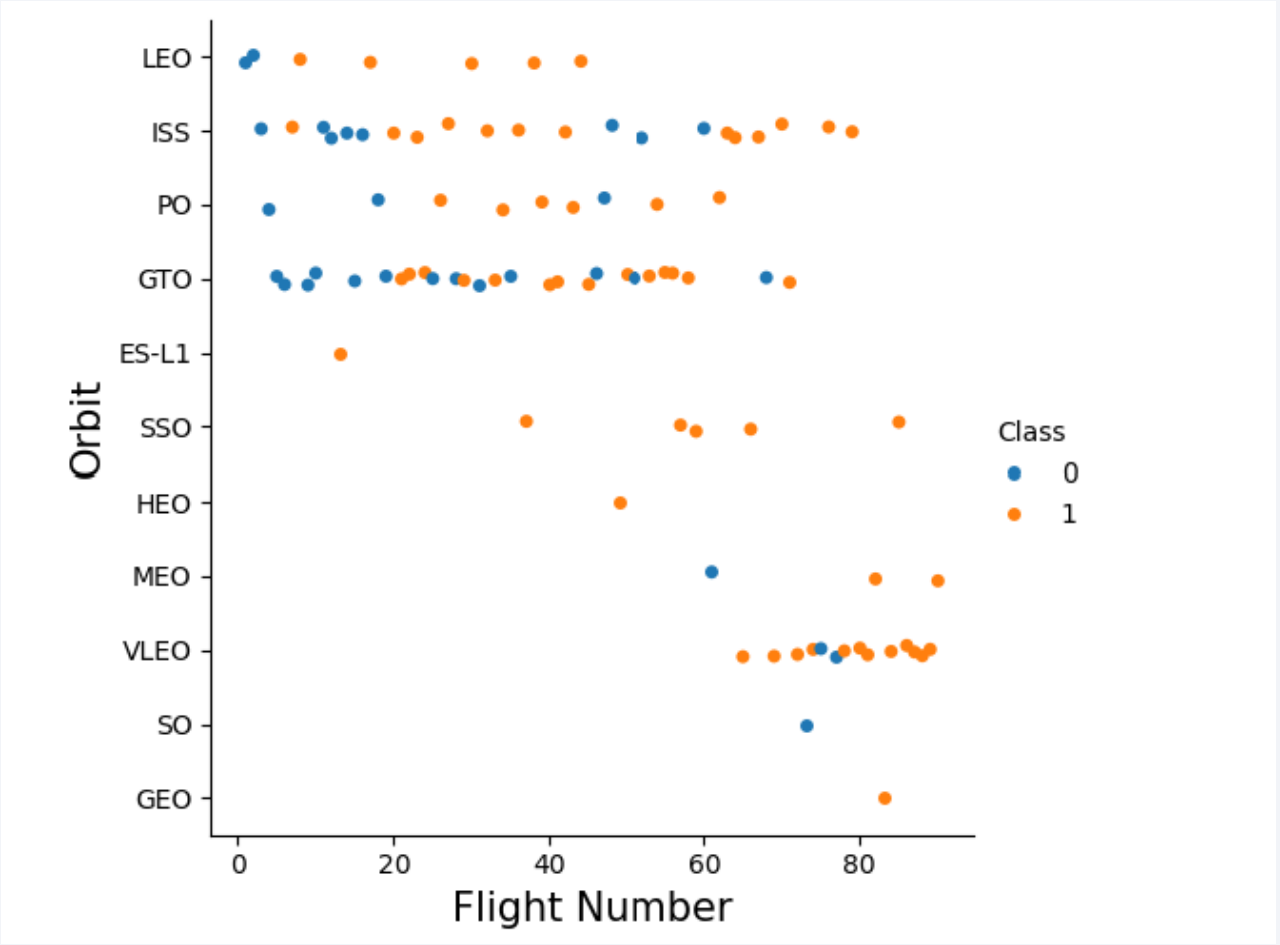
Success Rate vs. Orbit Type

- It explains the relationship between success rate of launch with different orbit type used in form of bar chart.
- Here we can observe that Orbit- ES-L1, GEO , HEO and SSO have 100 % success rate while SO has 0 success rate.

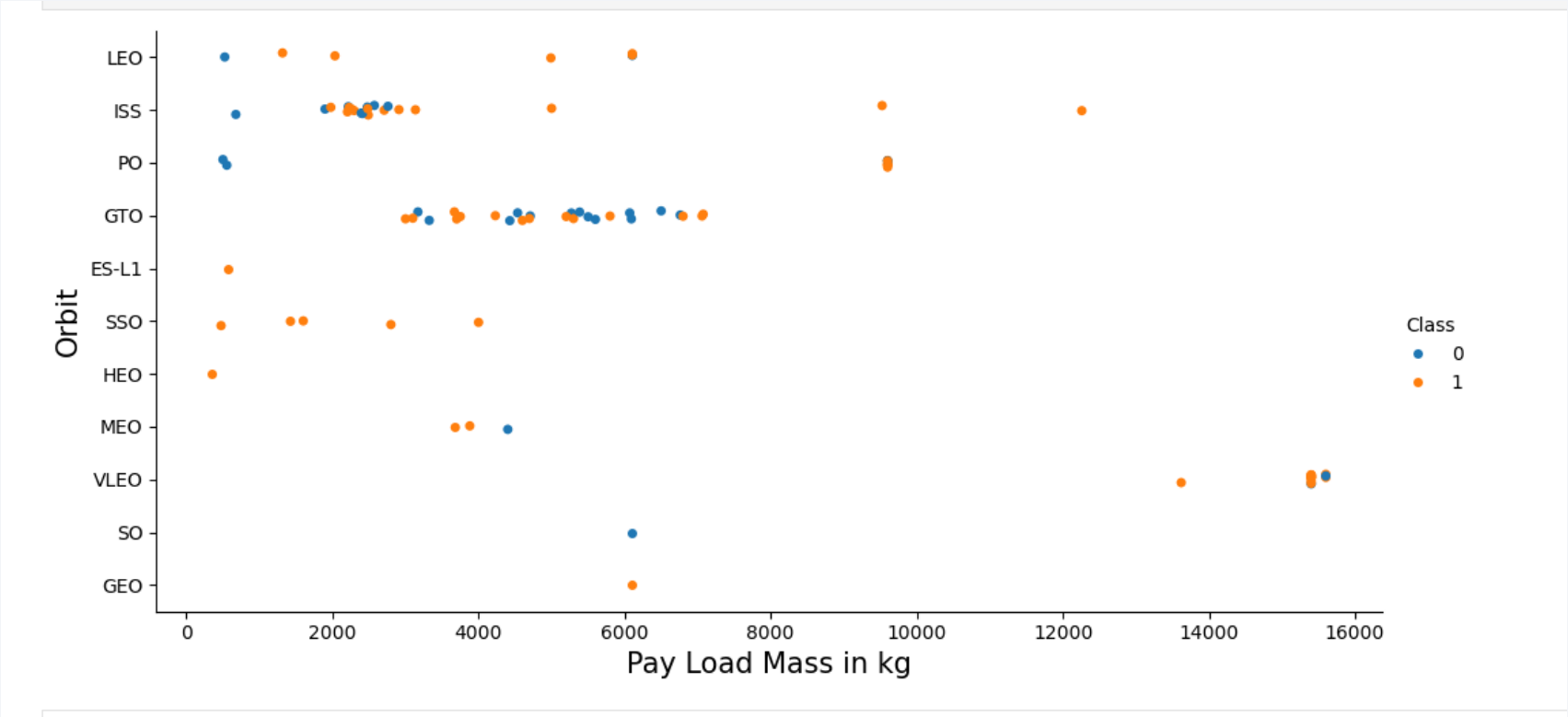


Flight Number vs. Orbit Type

- It can be observed that majority of the launches were from GTO,PO,ISS and LEO Orbit.
- All the initial flight launches before 40 were in LEO,ISS,PO,GTO and ES-L1 orbit while recent launches have been in GEO,SO,VLEO,MEO,HEO and SEO orbit.



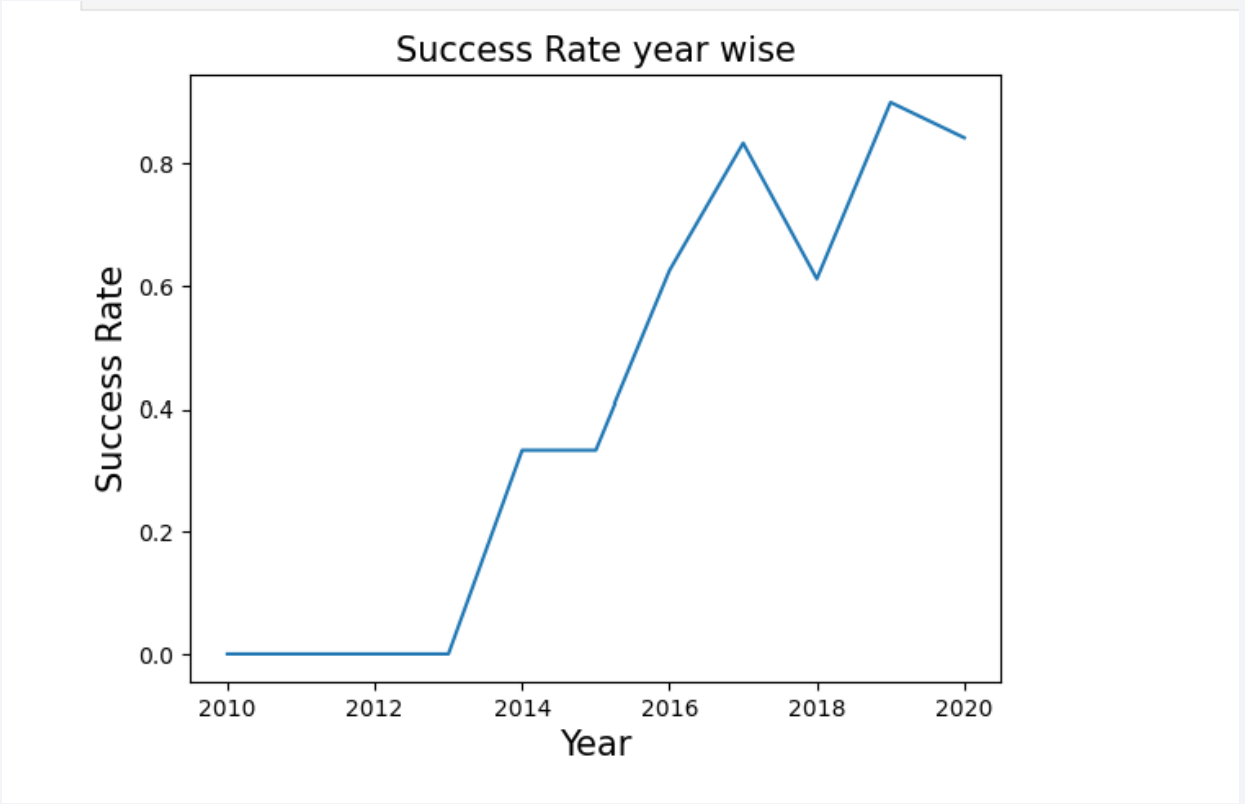
Payload vs. Orbit Type



- Majority of Payload were below 6000 kg.
- SSO which has 100 % success rate had payload below 6000 kg
- VLEO is the only orbit where payload used was always above 12000 kg

Launch Success Yearly Trend

- It can be observed that success rate was 0 till 2013 and after 2013, there has been a continuous increase till 2017.
- Success rate has seen fluctuations from 2017 onwards , with drops in 2018, increase in 2019 and again drop in 2020.
- 2019 remains the best year for launch success



All Launch Site Names

- Names of the unique launch site

Task 1

Display the names of the unique launch sites in the space mission

```
[8]: %sql SELECT DISTINCT Launch_Site FROM SPACEXTBL;
```

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

[8]:

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

Launch Site Names Begin with 'CCA'

5 records where launch sites begin with `CCA

Task 2

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

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

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

[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	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

- Total payload carried by boosters from NASA us 45596 kg

Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

```
[10]: %sql SELECT SUM(PAYLOAD_MASS__KG_) AS TotalPayloadMass FROM SPACEXTBL WHERE Customer = 'NASA (CRS)';
```

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

[10]:

TotalPayloadMass
45596

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1is 2928.4 kg

Task 4

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

```
[11]: %sql SELECT AVG(PAYLOAD_MASS__KG_) AS AveragePayloadMass FROM SPACEXTBL WHERE Booster_Version = 'F9 v1.1';
```

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

[11]:

AveragePayloadMass
2928.4

First Successful Ground Landing Date

- Dates of the first successful landing outcome on ground pad is 22 Dec 2015

Task 5

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

Hint:Use min function

[12]:
```
%sql SELECT MIN(Date) AS FirstSuccessfulLandingDate FROM SPACEXTBL WHERE Landing_Outcome = 'Success (ground pad)';
```

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

[12]:

FirstSuccessfulLandingDate
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are listed below

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

[13]:
```
%sql SELECT DISTINCT Booster_Version FROM SPACEXTBL WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;
```

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

[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

- Total number of successful and failure mission outcome mentioned in below table (image)

Task 7

List the total number of successful and failure mission outcomes

```
[16]: %sql SELECT "Mission_Outcome", COUNT(*) AS TotalOutcomes FROM SPACEXTBL GROUP BY "Mission_Outcome";
```

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

[16]:

Mission_Outcome	TotalOutcomes
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Below table shows all booster versions that have carried max payload mass

Task 8

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

```
[18]: %sql SELECT DISTINCT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

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

[18]:

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 failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

Done.

[13]:

MonthName	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

- 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

Done.

[11]:

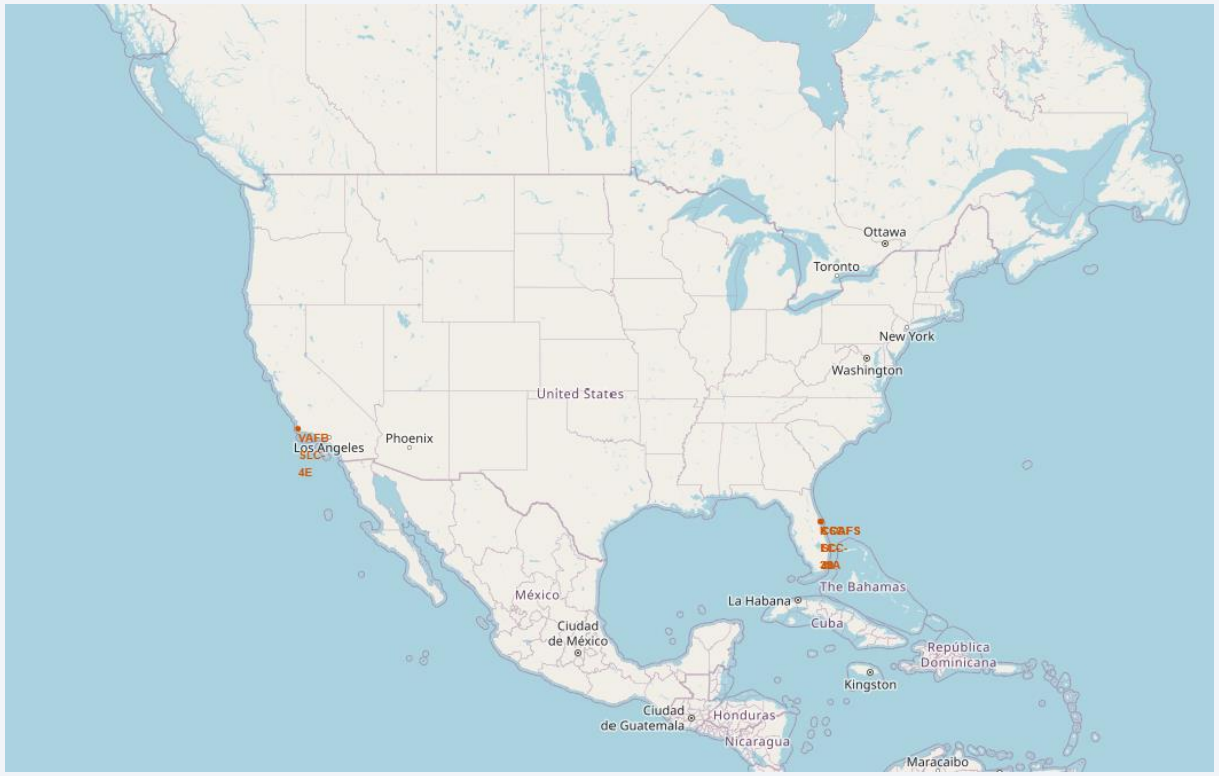
Landing_Outcome	Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

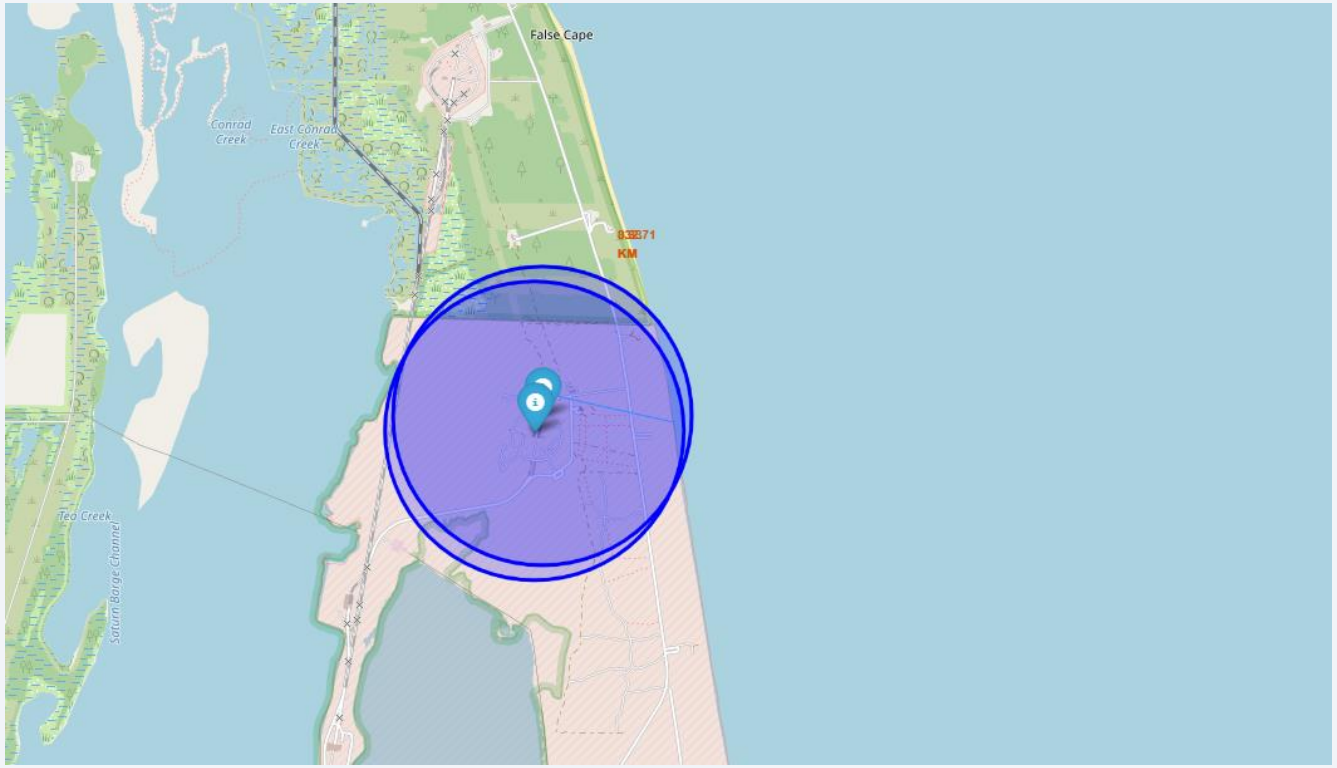
<Folium Map –All Launch Sites markings

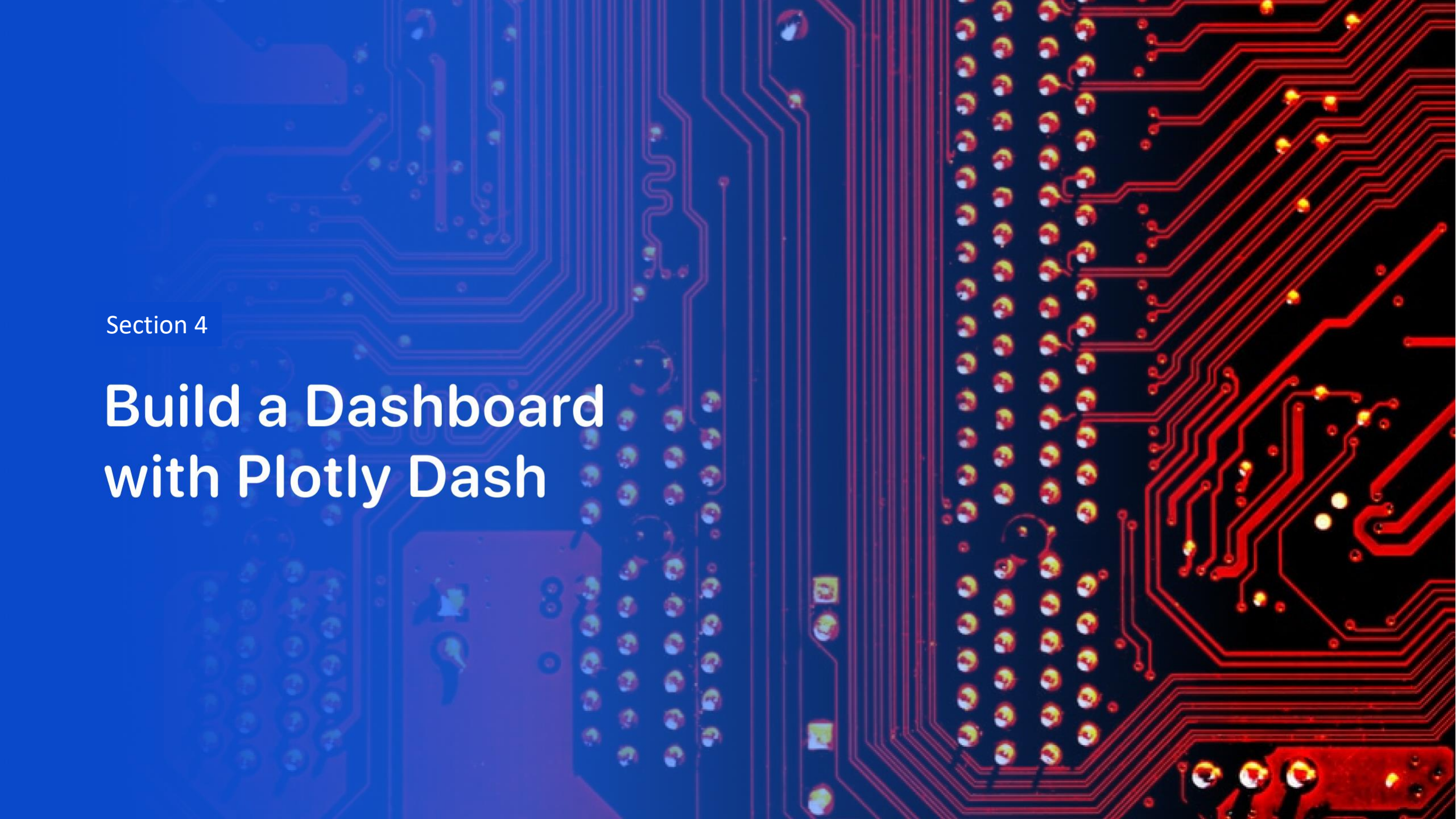
- This shows all the successful launch sites locations across US, being in California and Florida with red markers



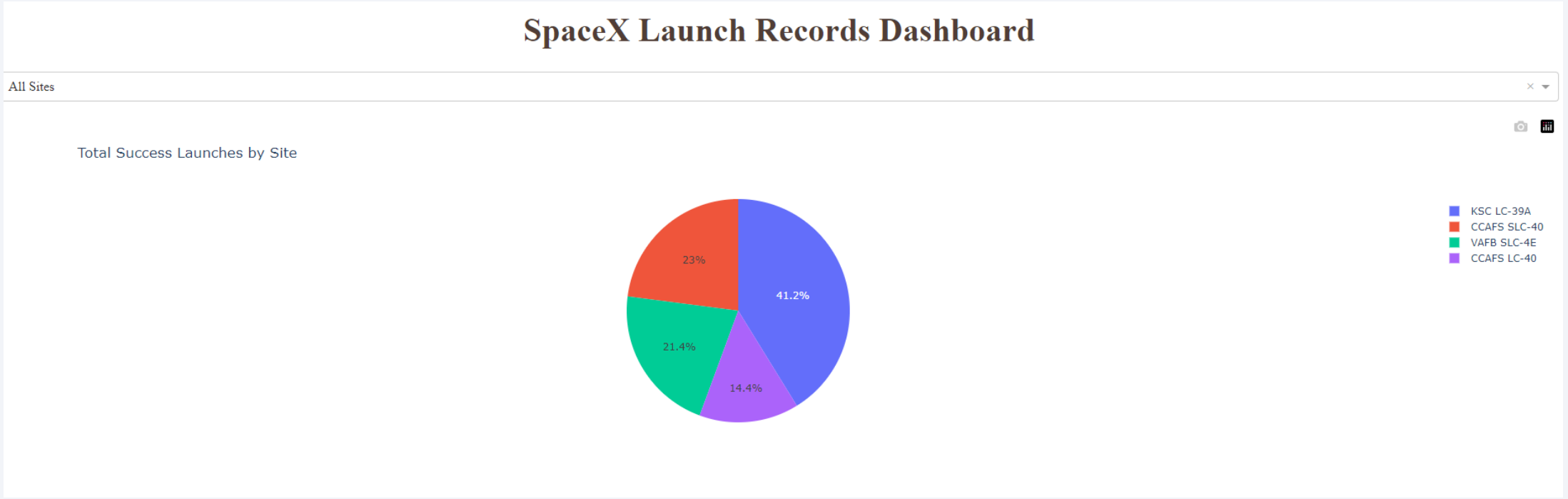
Folium Map – Distance with coastline

- Distance between launch site and coastline is approx 0.9 km



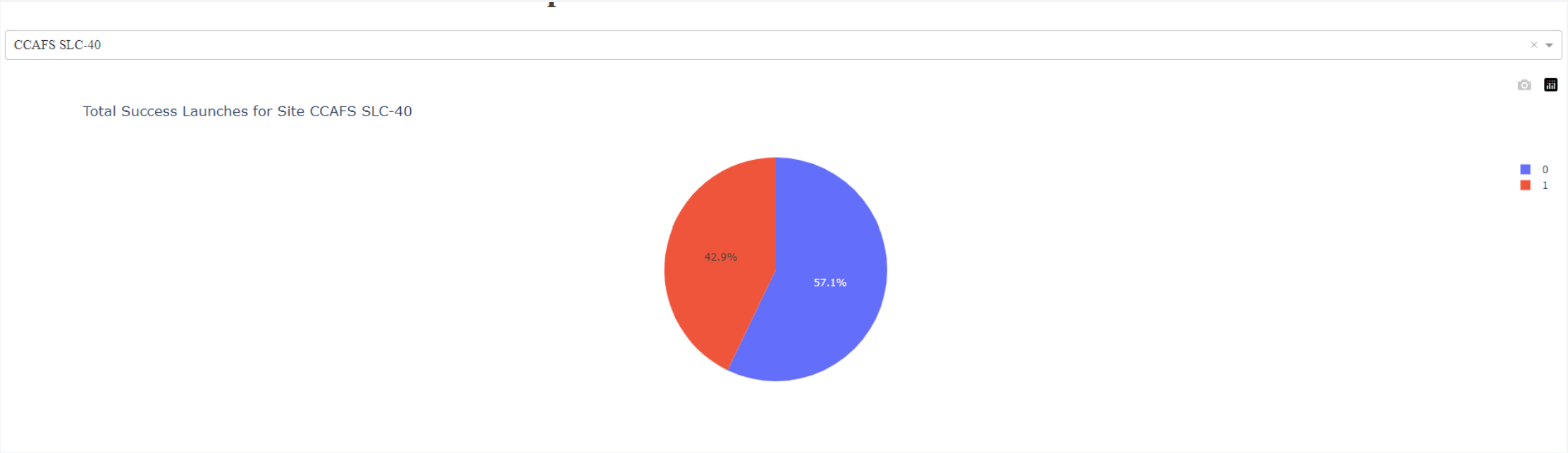

Section 4
Build a Dashboard with Plotly Dash

Dashboard for All sites



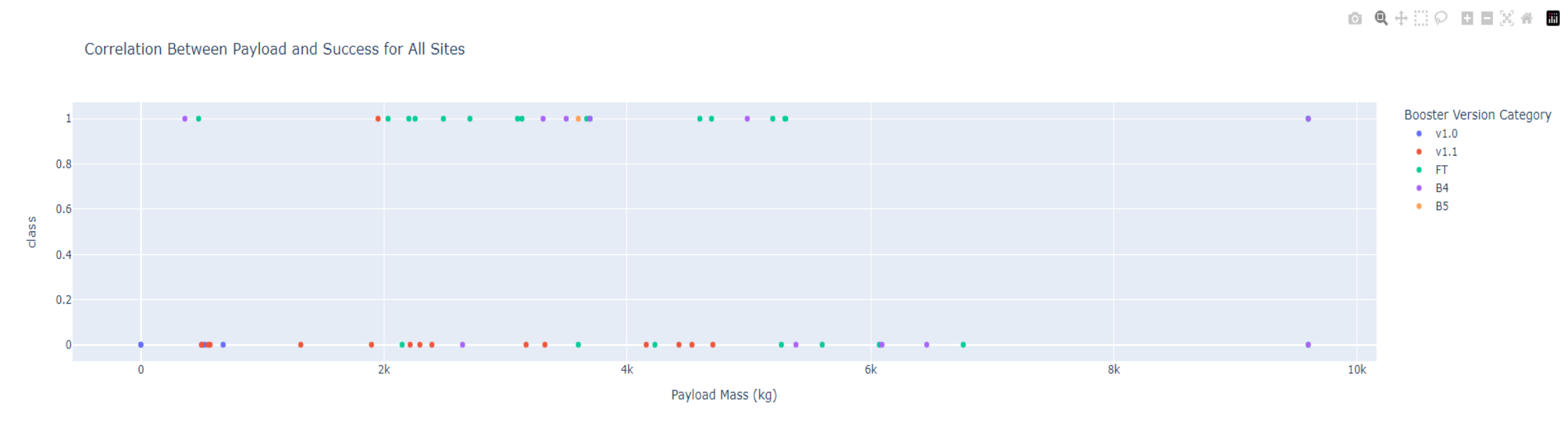
- This chart shows all sites and their proportion in overall launches made
- It can be seen that KSC KC-39 A has higher share compared to other sites

Pie chart for Most successful launch site



- Pie chart for site CCAFS SLC 40 where 1 means success and 0 means failure. This site has more success than other sites as per plotly dashboard analysis

Scatter plot for Payload vs Success ratio



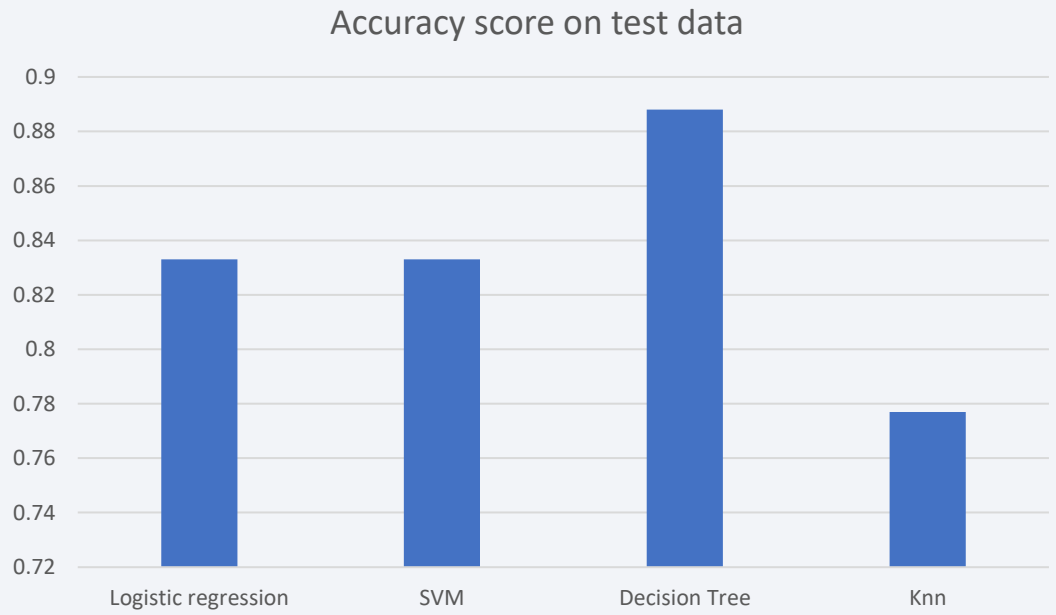
- Replace <Dashboard screenshot 3> title with an appropriate title
- Show screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider
- Explain the important elements and findings on the screenshot, such as which payload range or booster version have the largest success rate, etc.

Section 5

Predictive Analysis (Classification)

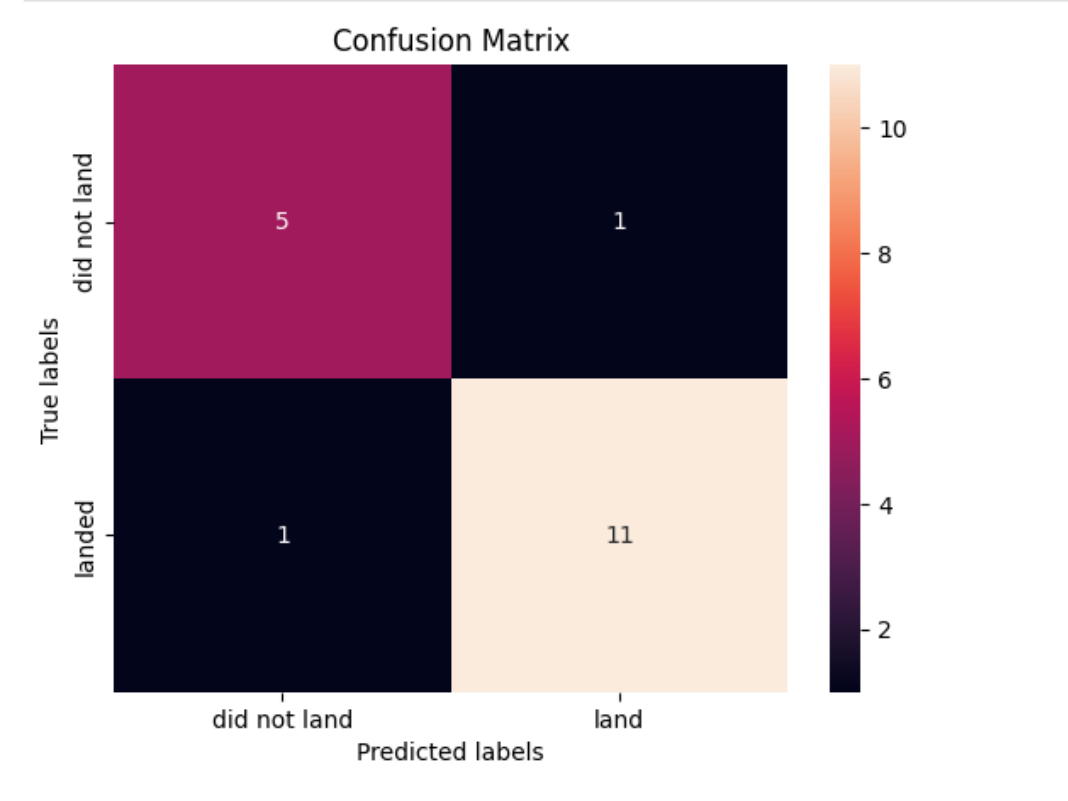
Classification Accuracy

- Bar chart for different models with their accuracy score on test data
- It can be determined from the analysis that Decision tree model has higher score compared to other models.



Confusion Matrix

- Confusion matrix of the Decision tree model.
- Based on the analysis, it can be determined that decision tree model has greater score and also accuracy on test data compared to other 3 models.



Conclusions

- First successful landing occurred on 22 Dec 2015 with overall 99 missions have been successful. Launch success rate has seen increase from 2013 till 2017 with drop afterwards but 2019 remaining most successful year
- Avg payload mass is around 2928 kg. Majority of payload being less than 6000 kg. Launches having less payload is preferred for more success
- Most successful orbits were from GTO, PO, ISS and LEO orbit
- Based on launch sites data, site- CCAFS SLC 40 has more success ratio for landings
- We can ascertain that Decision tree model is best fit for this analysis and for future prediction

Appendix

- Tools used Anaconda environment-Jupyter notebook, plotly dash
- Language used for analysis SQL lite and Python
- IBM Data Science Course

Thank you!