Winning Space Race with Data Science

Guosong Li

September 25, 2023

Outline

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

Executive Summary

In this study, we collected Space X Falcon 9 rocket launches data from web. Then, we applied data science tools to process, analyze and visualize the data. Several machine learning algorithms were used to make prediction.

The methodologies used for data processing include data collection, data wrangling and preprocessing, exploratory data analysis, data visualization.

By looking at those plots, some trend and/or characteristics can be found, such as the improvement of successful rate over years.

Four machine learning classifiers were tried. The results showed that among four methods, three methods, K nearest neighbors, Logistic regression and Support vector machine have same accuracy while decision tree classifier has lower accuracy.

Introduction

SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. 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 an alternate company wants to bid against SpaceX for a rocket launch.

In this lab, we will collect and make sure the data is in the correct format from an API. Then, process and visualize the data, and get some insights. In addition, we will apply machine learning technology to fit models and then make predictions.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX launch data that is gathered from an API, specifically the SpaceX REST API
- Perform data wrangling
  - Using data analysis to find some patterns in the data and determine what would be the label for training supervised models, converting label from string to numbers
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Four ML classifiers will be tried to fit model and make predictions

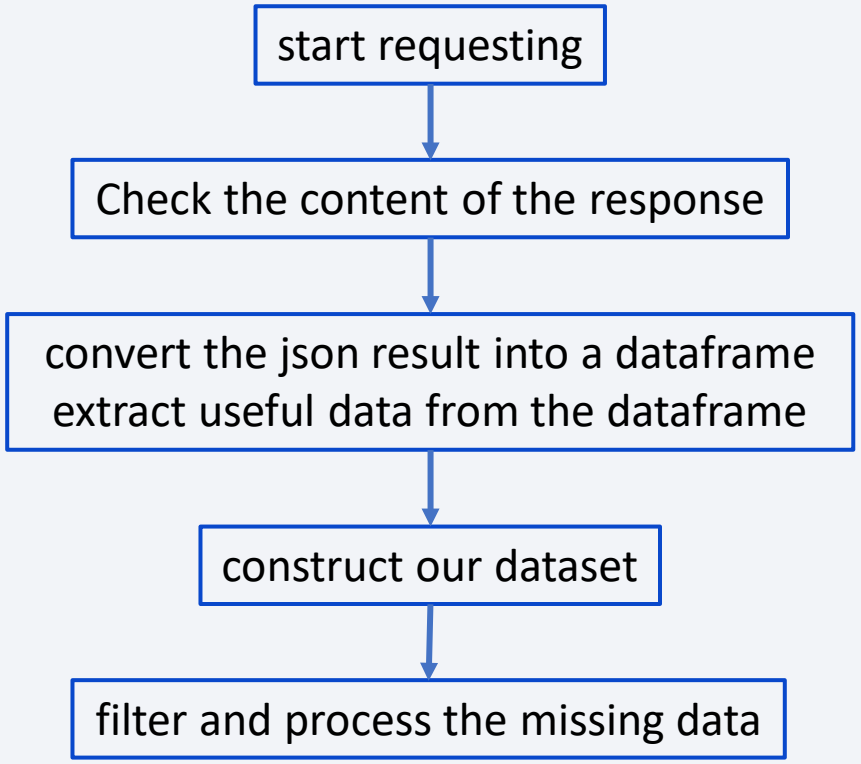
Data Collection

- The data was collected from different sources
  - https://api.spacexdata.com/v4/rockets
  - https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches
  - https://...../IBM-DS0321EN-SkillsNetwork/datasets/dataset_part_1.csv
- Apply different approach to get the data
  - Make a get request to the SpaceX API, then clean the requested data
  - Extract a Falcon 9 launch records HTML table from Wikipedia, and parse the table and convert it into a Pandas data frame

Data Collection – SpaceX API

- Request and parse the SpaceX launch data using the GET request
- Filter the dataframe to only include Falcon 9 launches
- Dealing with Missing Values

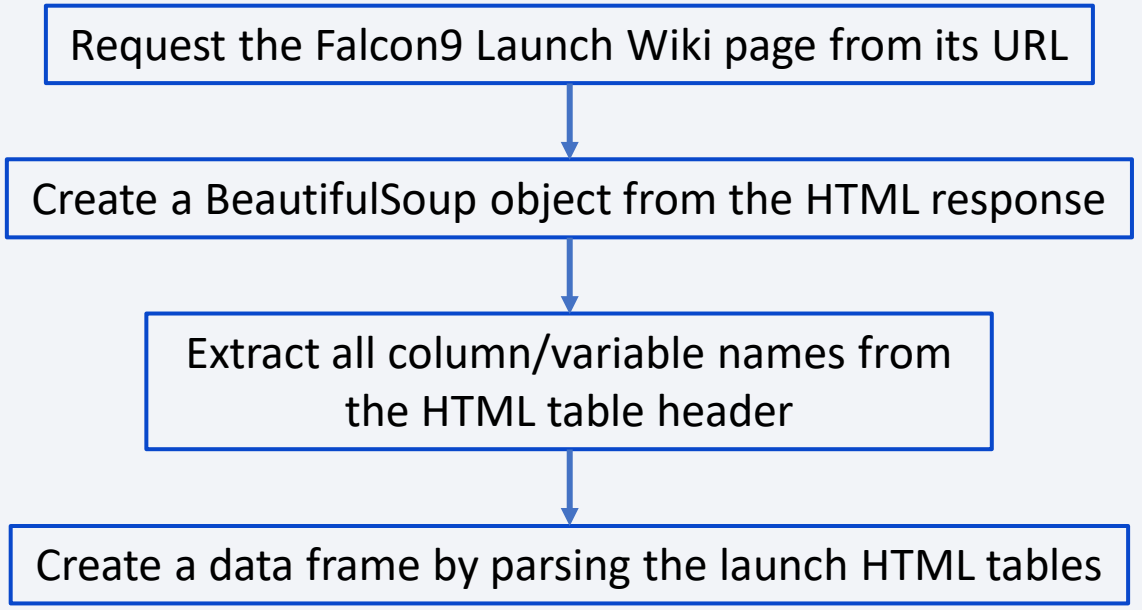
- IBM_Course_Applied-Data-Science-Capstone/1.week_01_jupyter-labs-spacex-data-collection-api.ipynb (github.com)



Data Collection - Scraping

- Extract a Falcon 9 launch records HTML table from Wikipedia
- Parse the table and convert it into a Pandas data frame

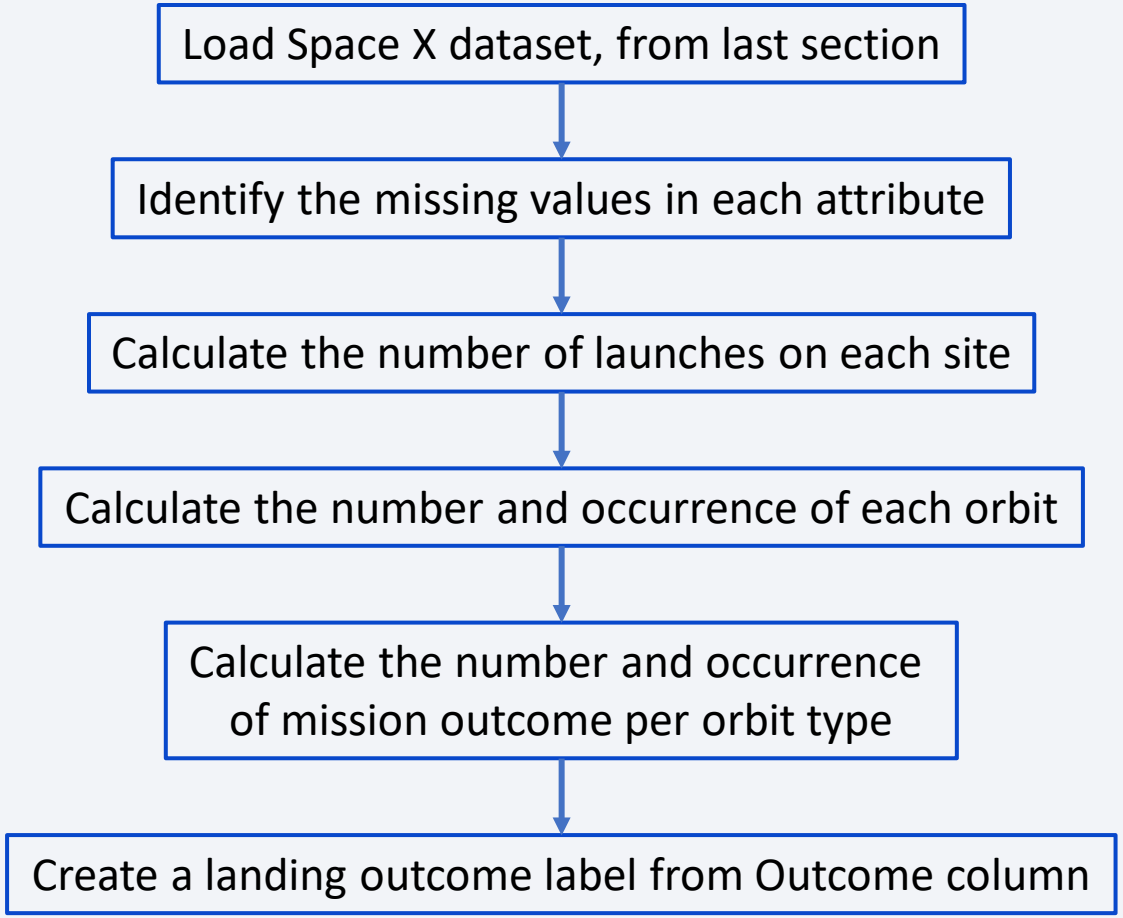
- [IBM_Course_Applied-Data-Science-Capstone/2. week_01_jupyter-labs-webscraping.ipynb (github.com)]



Data Wrangling

- Describe how data were processed
- You need to present your data wrangling process using key phrases and flowcharts

- IBM_Course_Applied-Data-Science-Capstone/3.week_01_labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb (github.com)



EDA with Data Visualization

- Scatter plots: Scatter plots were used to represent the relationship between two variables. Different sets of features were compared such as Flight Number vs. Launch Site, Payload vs. Launch Site, Flight Number vs. Orbit Type and Payload vs. Orbit Type.
- Bar chart: Bar charts were used makes it easy to compare values between multiple groups. The x-axis represents a category and the y-axis represents a discrete value. Bar charts were used to compare the *Success Rate* for different *Orbit Types*
- Line chart: Line charts are used for showing data trends over time. A line chart was used to show Success Rate over a certain number of Years.
- IBM_Course_Applied-Data-Science-Capstone/5. week_02_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb (github.com)

EDA with SQL

- Display different kinds of data, such as
  - the names of the unique launch sites in the space mission, records where launch sites begin with specific string, e.g. 'CCA'
  - total payload mass carried by boosters launched by NASA (CRS), average payload mass carried by booster version F9 v1.1
- List different types of data
  - the date when the first successful landing outcome, the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - the names of the booster_versions which have carried the maximum payload mass
- Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20
- Add the GitHub URL of your completed EDA with SQL notebook, as an external reference and peer-review purpose

Build an Interactive Map with Folium

Finding an optimal location for building a launch site certainly involves many factors and hopefully we could discover some of the factors by analyzing the existing launch site locations. In this lab, objects were created and added to a Folium map. Marker objects were used to show all launch sites on a map as well as the successful/failed launches for each site on the map. Line objects were used to calculate the distances between a launch site to its proximities

- By adding these objects, following geographical patterns about launch sites are found:
  - Are launch sites in close proximity to railways? Yes
  - Are launch sites in close proximity to highways? Yes
  - Are launch sites in close proximity to coastline? Yes
  - Do launch sites keep certain distance away from cities? Yes
- IBM_Course_Applied-Data-Science-Capstone/6.week_03_lab_jupyter_launch_site_location.jupyterlite.ipynb (github.com)

Build a Dashboard with Plotly Dash

In this lab, two type of charts were created using the dashboard application.

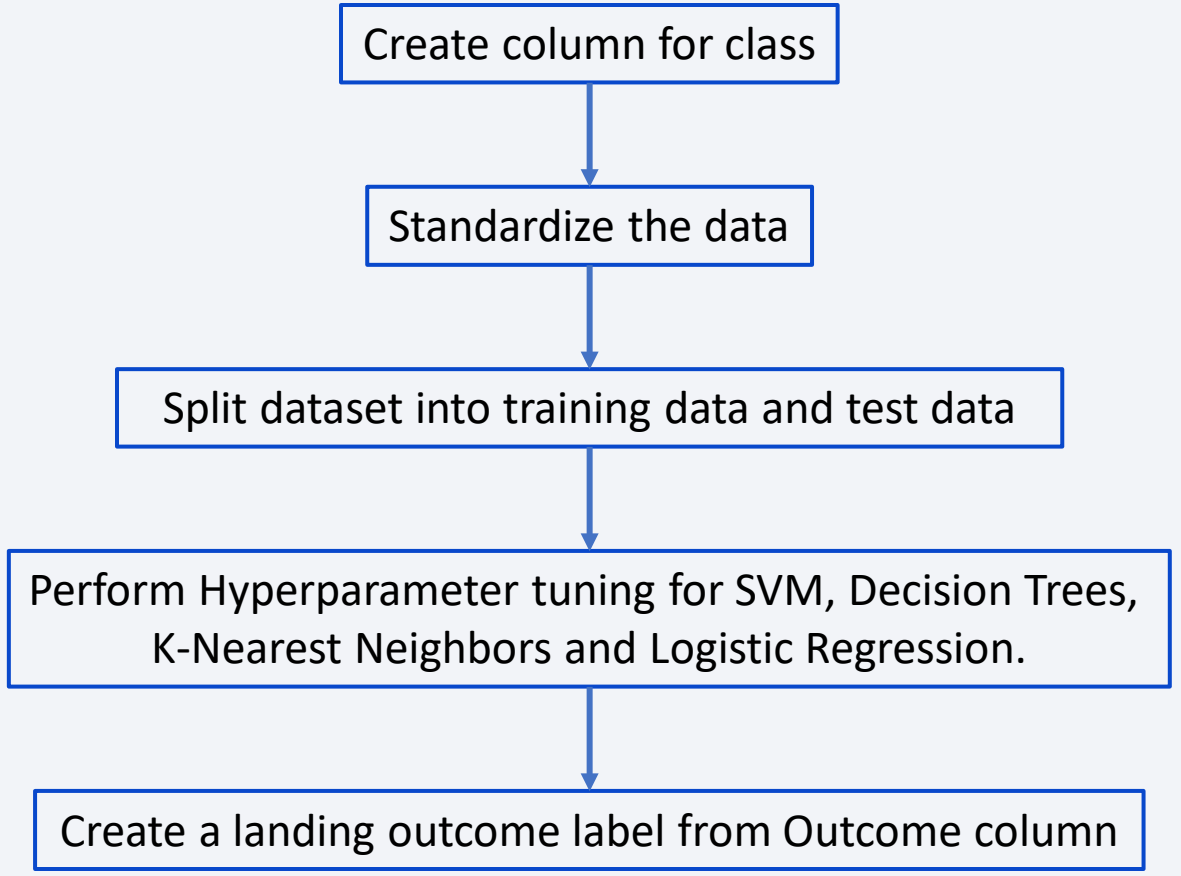
A pie chart that shows the successful launch by each site. This chart is useful as you can visualize the distribution of landing outcomes across all launch sites or show the success rate of launches on individual sites.

A scatter chart that shows the relationship between landing outcomes and the payload mass of different boosters. The dashboard takes two inputs, namely the site(s) and payload mass. This chart is useful as you can visualize how different variables affect the landing outcomes,

Predictive Analysis (Classification)

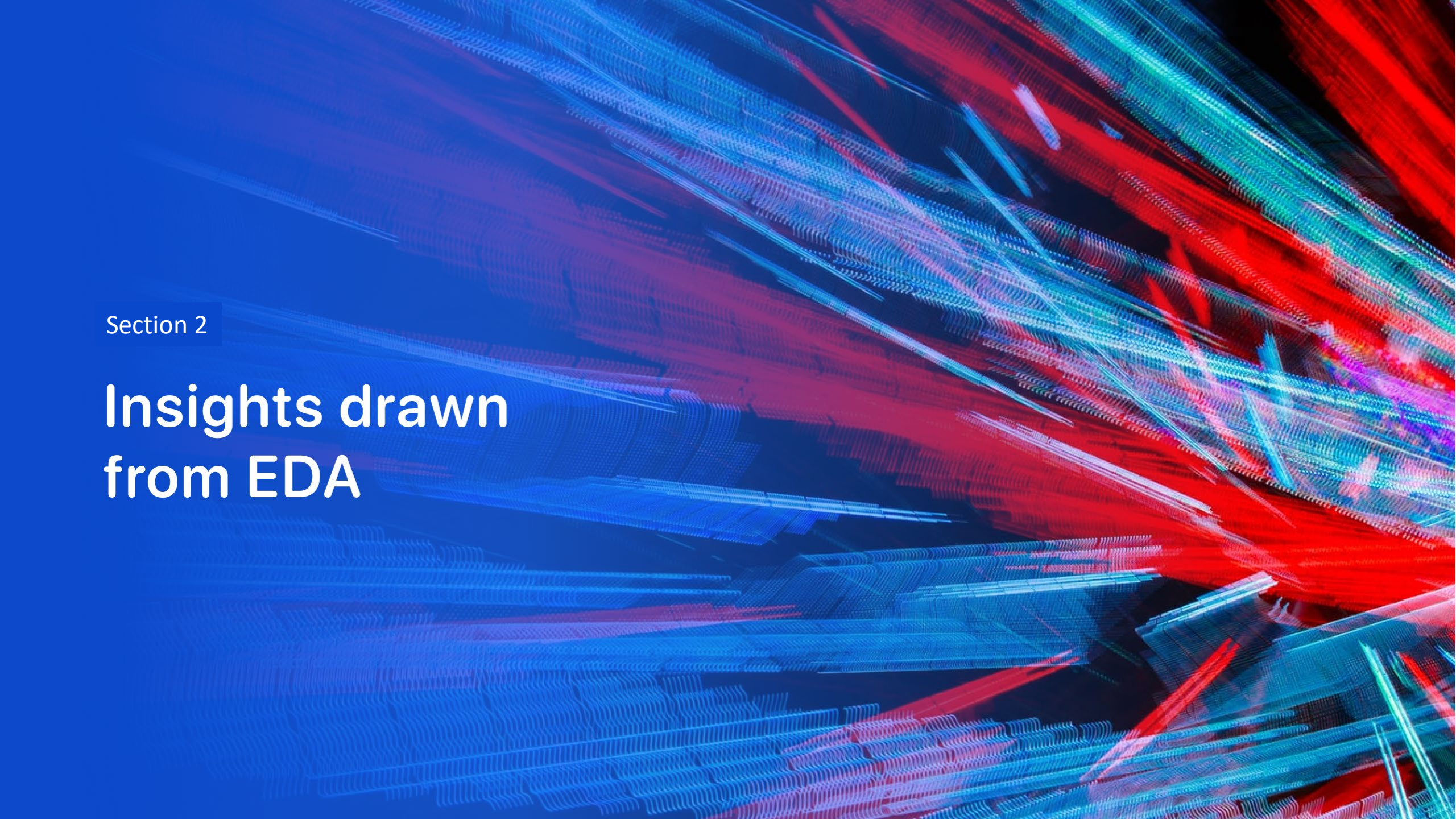
- The flowchart shows the process to perform machine learning study using four different classifiers

- IBM_Course_Applied-Data-Science-Capstone/8. week_04_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb (github.com)



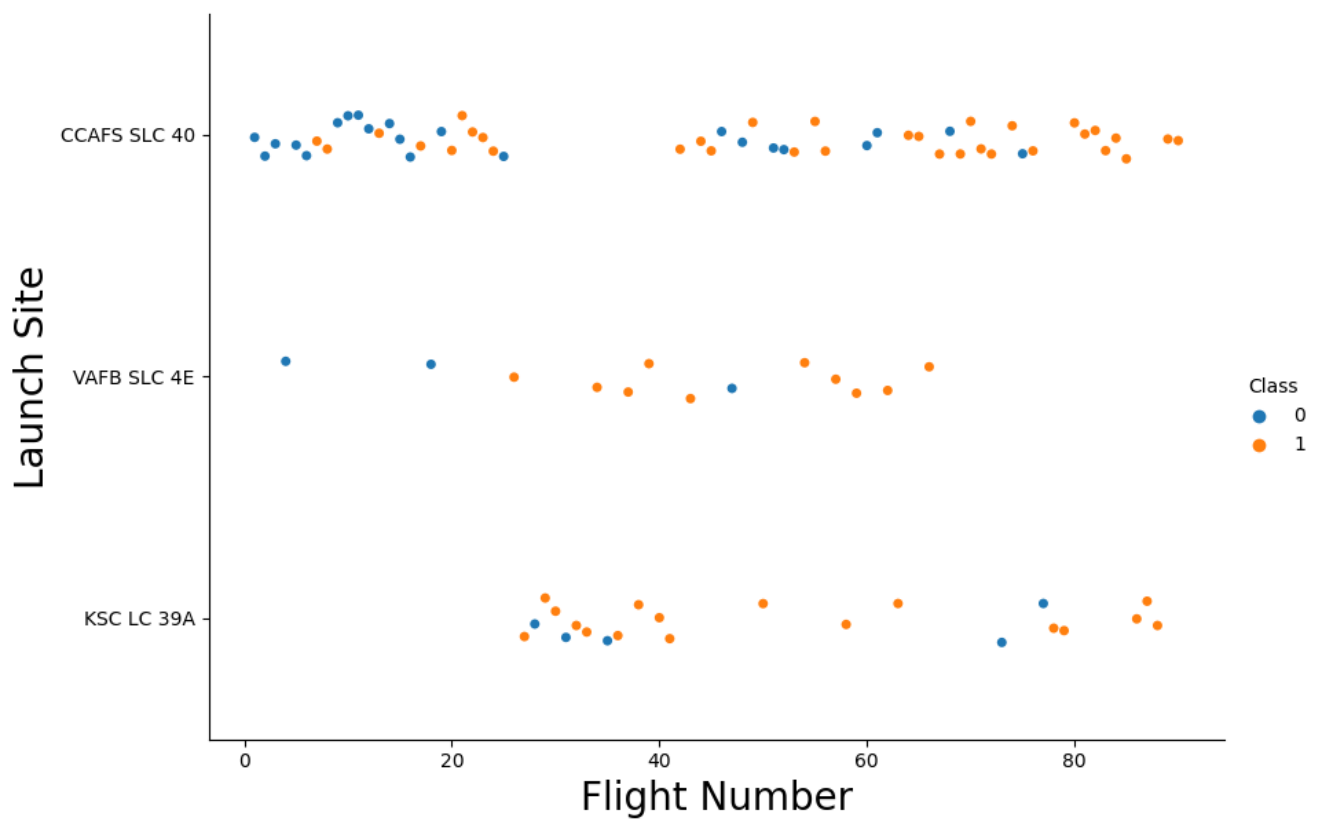
Results

- The results of the exploratory data analysis shows the relationships of Falcon 9 characteristic quantities and trend such as successful ratio.
- The machine learning study results showed that the classifier algorithms can achieve 77.8% - 83.3% accuracy in prediction.


Section 2
Insights drawn from EDA

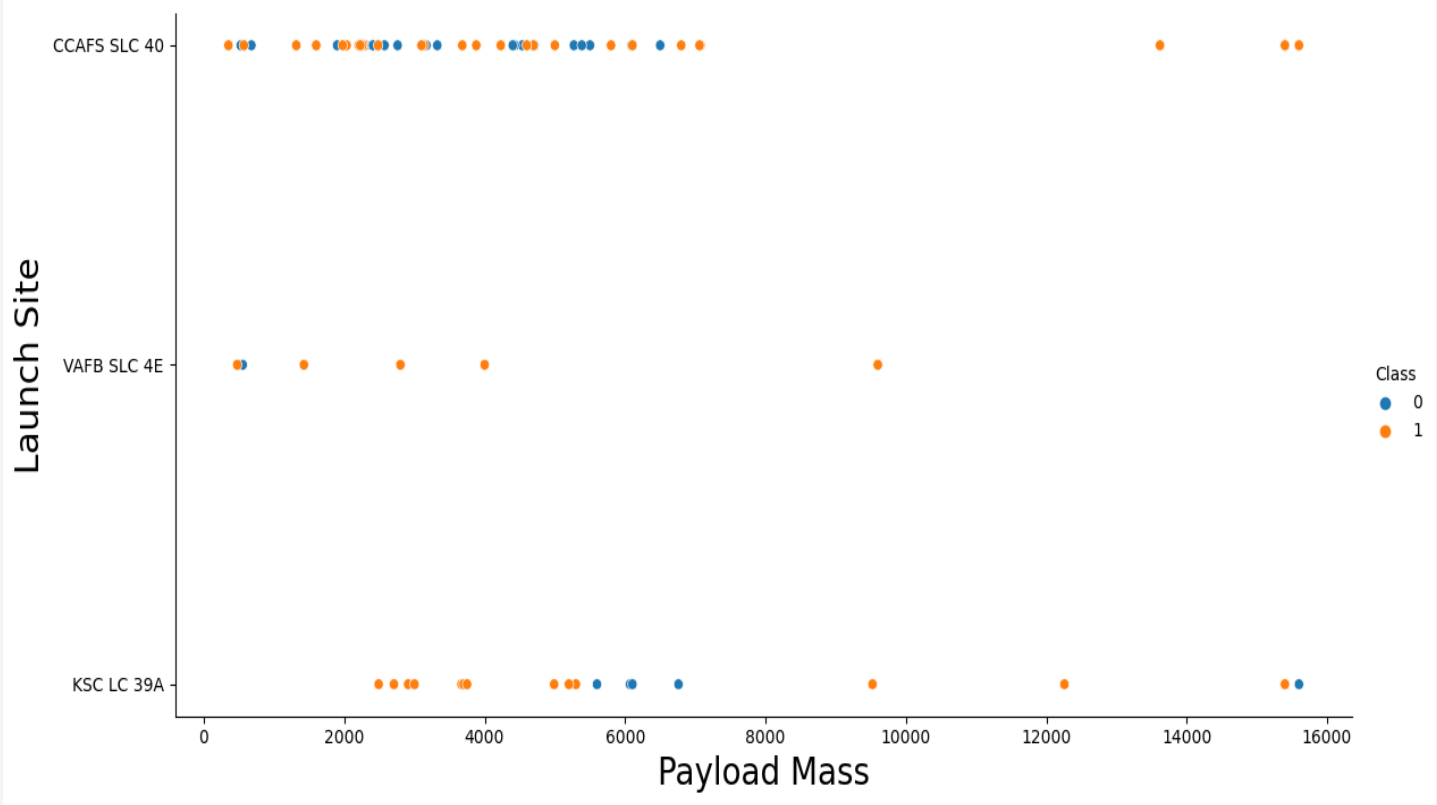
Flight Number vs. Launch Site

- This is the scatter plot of Flight Number vs. Launch Site
- From this plot, you can clear see where those flights were launch, and how are the outcomes.



Payload vs. Launch Site

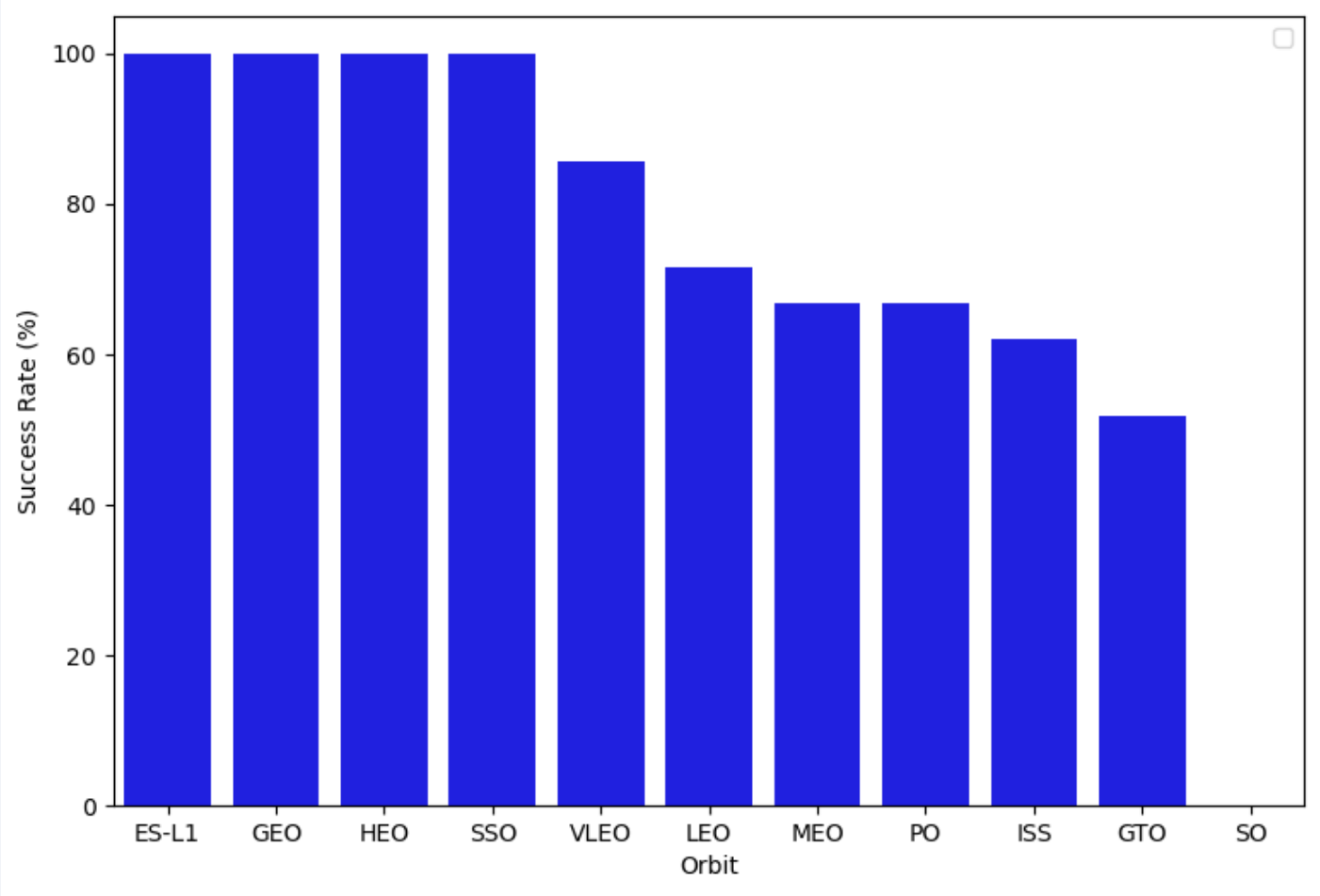
- This is a scatter plot of Payload vs. Launch Site
- This plot clearly shows the pay load data for each launch site



Success Rate vs. Orbit Type

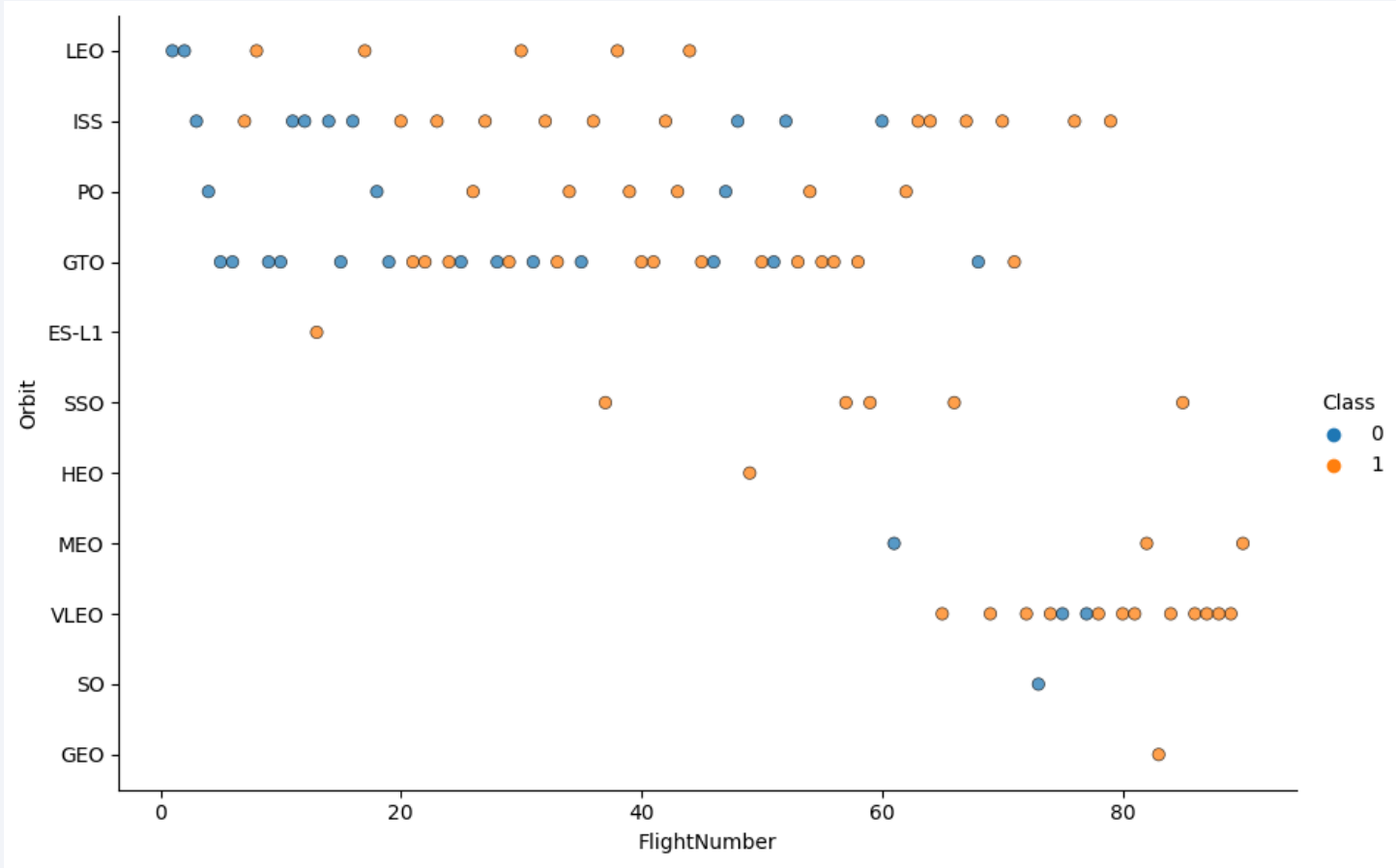
- This is a bar chart for the success rate of each orbit type

- This plot shows that some orbits have higher success rate compared to other orbits



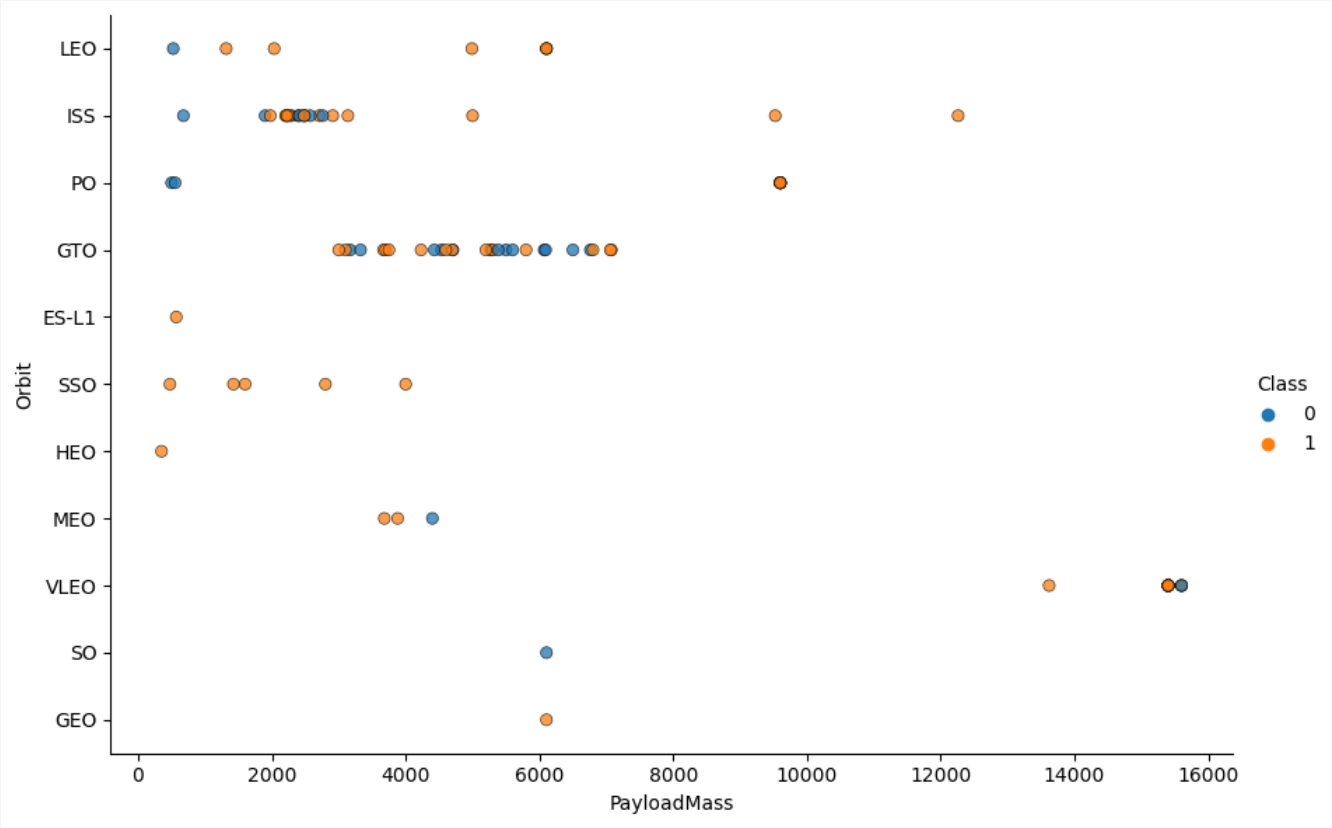
Flight Number vs. Orbit Type

- This is a scatter point of Flight number vs. Orbit type
- This plot shows that in early days, more flights were launched to lower orbits while more flights were launched to higher orbits later



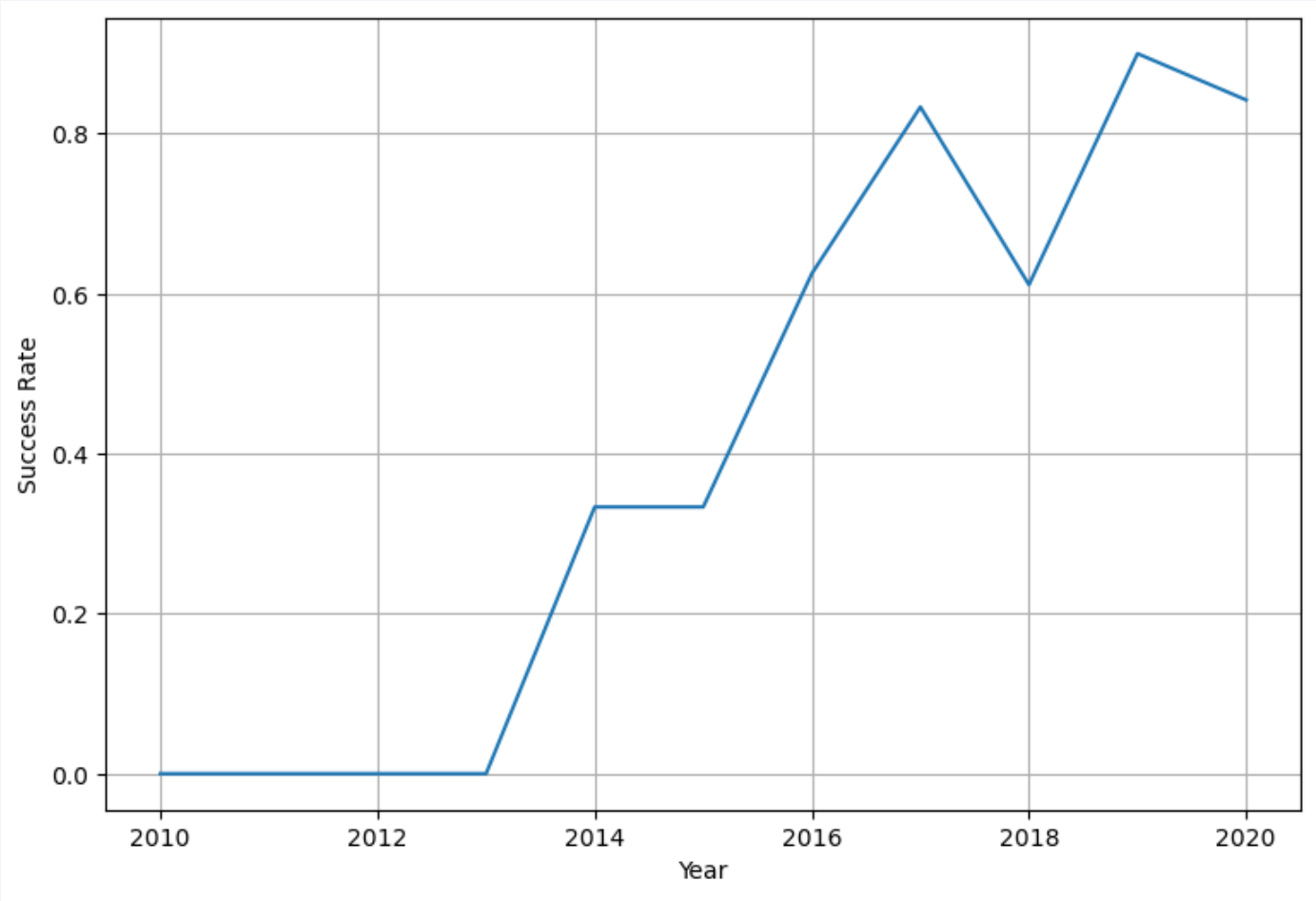
Payload vs. Orbit Type

- This is a scatter point of payload vs. orbit type
- This plot shows that the payloads to lower orbits usually are higher that the payloads to higher orbits. This follows the physics laws.



Launch Success Yearly Trend

- This is a line chart of yearly average success rate
- The plot clearly shows the trend that the success rate improves significantly year by year.



All Launch Site Names

- Find the names of the unique launch sites
- Select those records with distinct launch sites

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

In [7]:
```
%sql select distinct(LAUNCH_SITE) from SPACEXTBL
```

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

Out[7]:

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- Select those record with string ‘CCA’ in Launch_Site

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

In [8]:
```
%sql select * from SPACEXTBL where LAUNCH_SITE like 'CCA%' limit 5
```

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

Out[8]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Missio
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	
2010-08-12	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	
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

- Calculate the sum of payload_mass carried by boosters from NASA

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

```
In [9]: %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where CUSTOMER = 'NASA (CRS)'
 * sqlite:///my_data1.db
Done.
```

Out[9]:

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1

- Select those record with booster version F9 v1.1 and use keyword “avg” to calculate the average payload_mass

```
In [10]: %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where BOOSTER_VERSION = 'F9 v1.1'
 * sqlite:///my_data1.db
Done.
Out[10]: avg(PAYLOAD_MASS__KG_)
         2928.4
```

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- Select record with successful landing outcome on ground pad and use keyword “min” to find out the earliest data.

```
In [11]: %sql select min(DATE) from SPACEXTBL where Landing_Outcome = 'Success (ground pad)'
 * sqlite:///my_data1.db
Done.
Out[11]: min(DATE)
         2015-12-22
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- Select booster record with successful landing and payload mass is greater than 4000 but less than 6000

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

In [13]:
```
%sql select Booster_Version from SPACEXTBL where Landing_Outcome = 'Success (drone ship)' and (PAYLOAD_MASS__KG_ > 4000) and
```

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

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

- Calculate the total number of successful and failure mission outcomes
- Count the total of Mission_Outcome with outcomes are success or failure

List the total number of successful and failure mission outcomes

In [16]:
```
%sql select count(Mission_Outcome) from SPACEXTBL where Mission_Outcome = 'Success' or Mission_Outcome = 'Failure (in flight
```

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

Out[16]:

count(Mission_Outcome)
99

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Select those boosters with payload_mass being equal to the maximum payload.

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

In [19]:
```
%sql select Booster_Version from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

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

Out[19]:

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

- Select record with landing_outcome including 'Failure%drone' in year 2015

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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

```
%sql SELECT SUBSTR(Date,4,2) AS Month, Date, Booster_Version, Launch_site FROM SPACEXTBL WHERE Landing_Outcome LIKE 'Failure%drone%' AND SUBSTR(Date,7,4) =
```

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

Month	Date	Booster_Version	Launch_Site
5-	2015-10-01	F9 v1.1 B1012	CCAFS LC-40
5-	2015-04-14	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

- Select and count landing outcome between 2010-06-04 and 2017-03-20, group them, and rank them in order

Rank the count of landing outcomes (such as Failure (drone ship) or Success (gro

```
%%sql SELECT Landing_Outcome,  COUNT(Landing_Outcome)
FROM SPACEXTABLE
WHERE Date > '2010-06-04' AND Date < '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY COUNT(Landing_Outcome) DESC
```

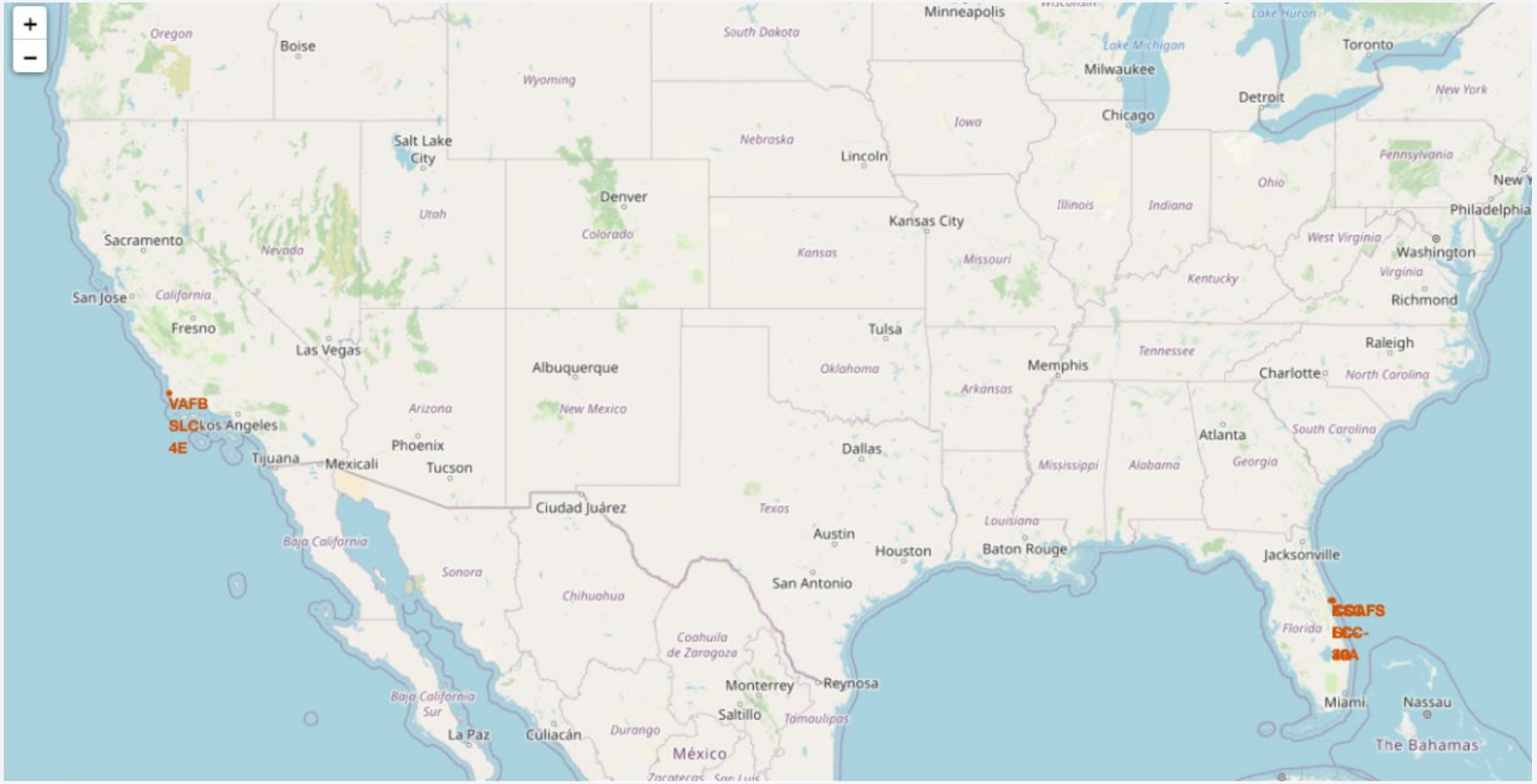
```
 * sqlite:///my_data1.db
Done.
```

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


Section 3
Launch Sites
Proximities Analysis

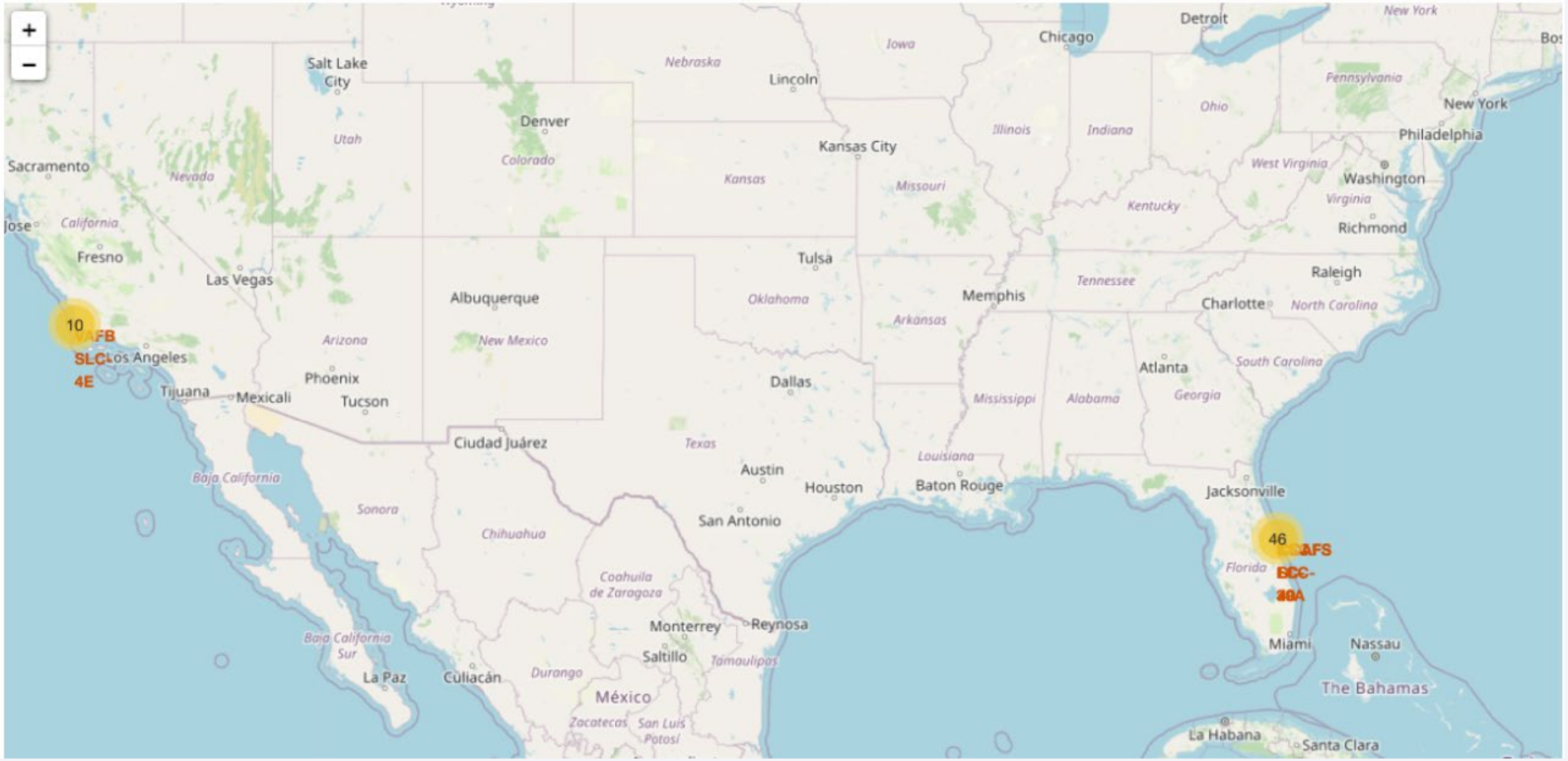
Folium Map of All Launch Sites

- This map shows all launch sites, some in the west, some in the east.



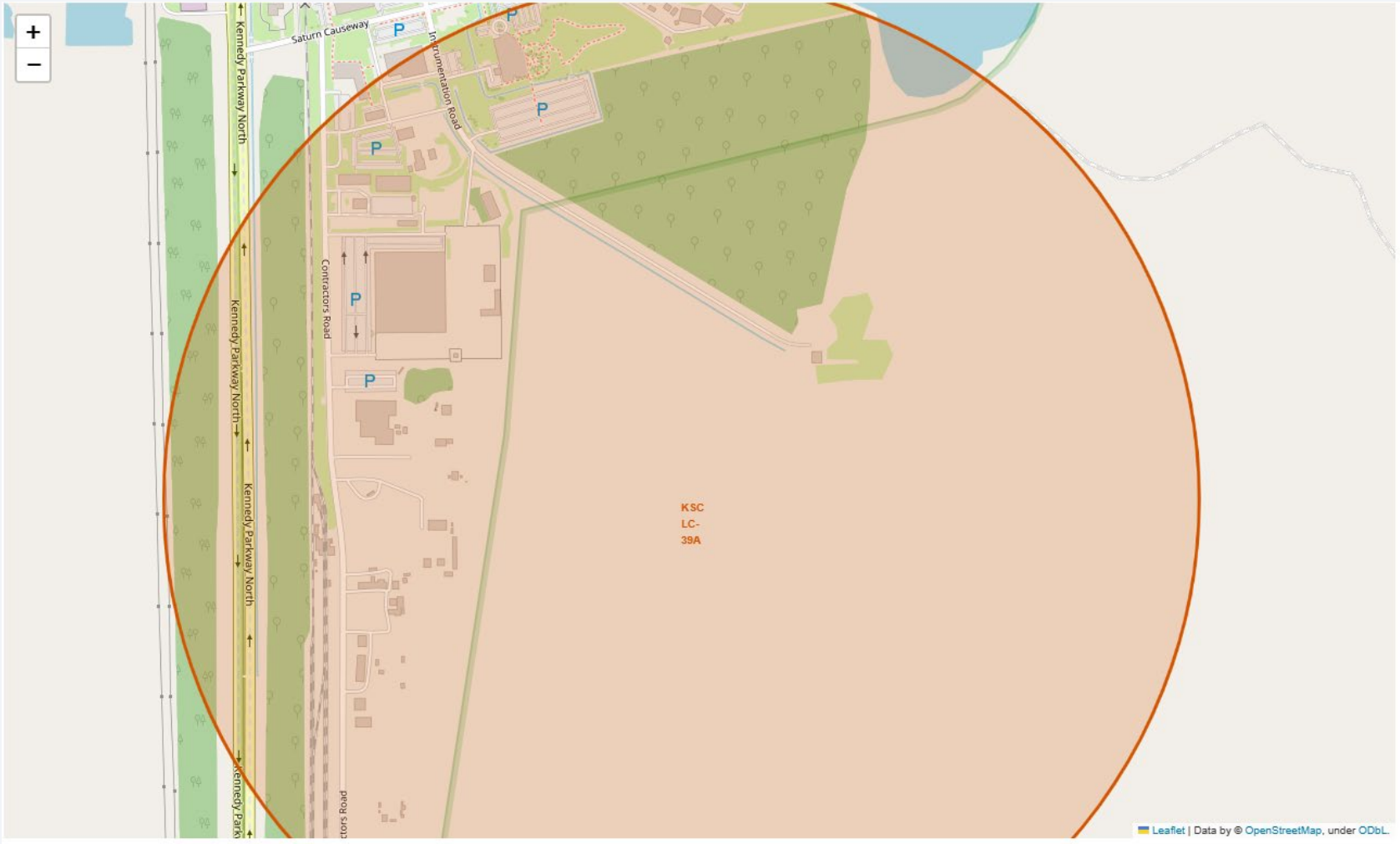
Folium Map of Launch Outcomes

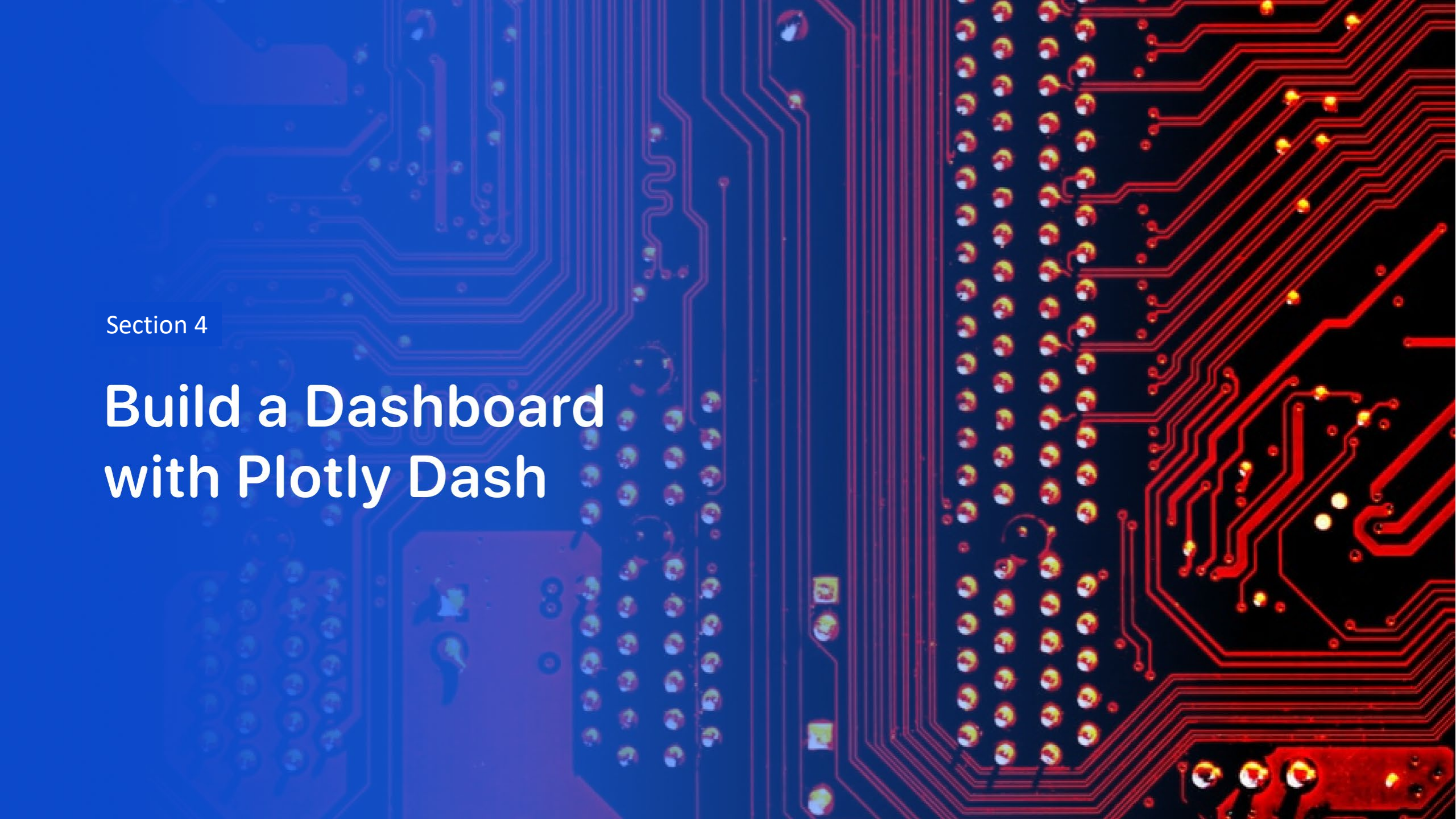
- The maps shows the launch outcomes



Folium Map Launch Site and its Proximities

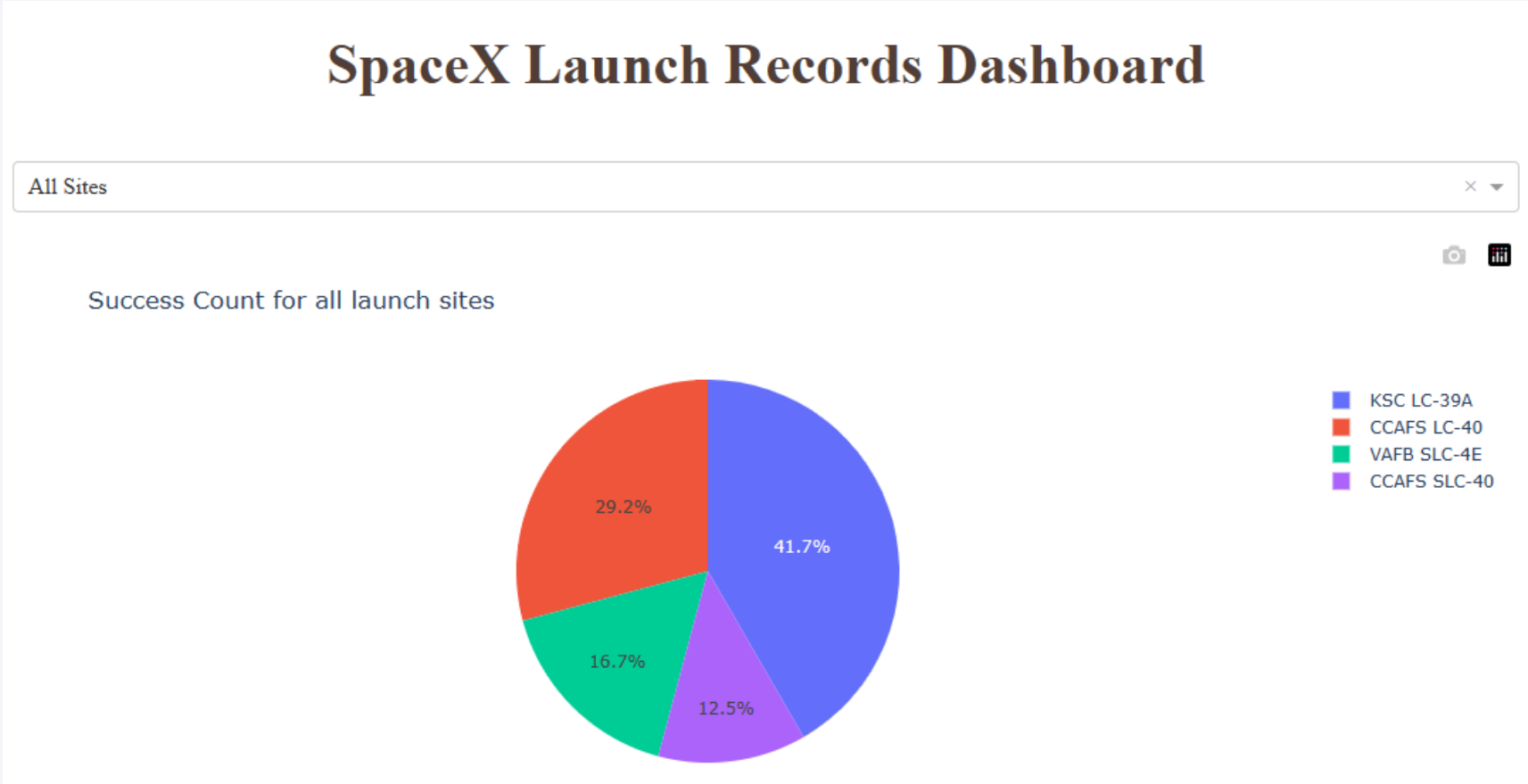
- Launch site and its proximities such as railway, highway can be seen




Section 4
Build a Dashboard with Plotly Dash

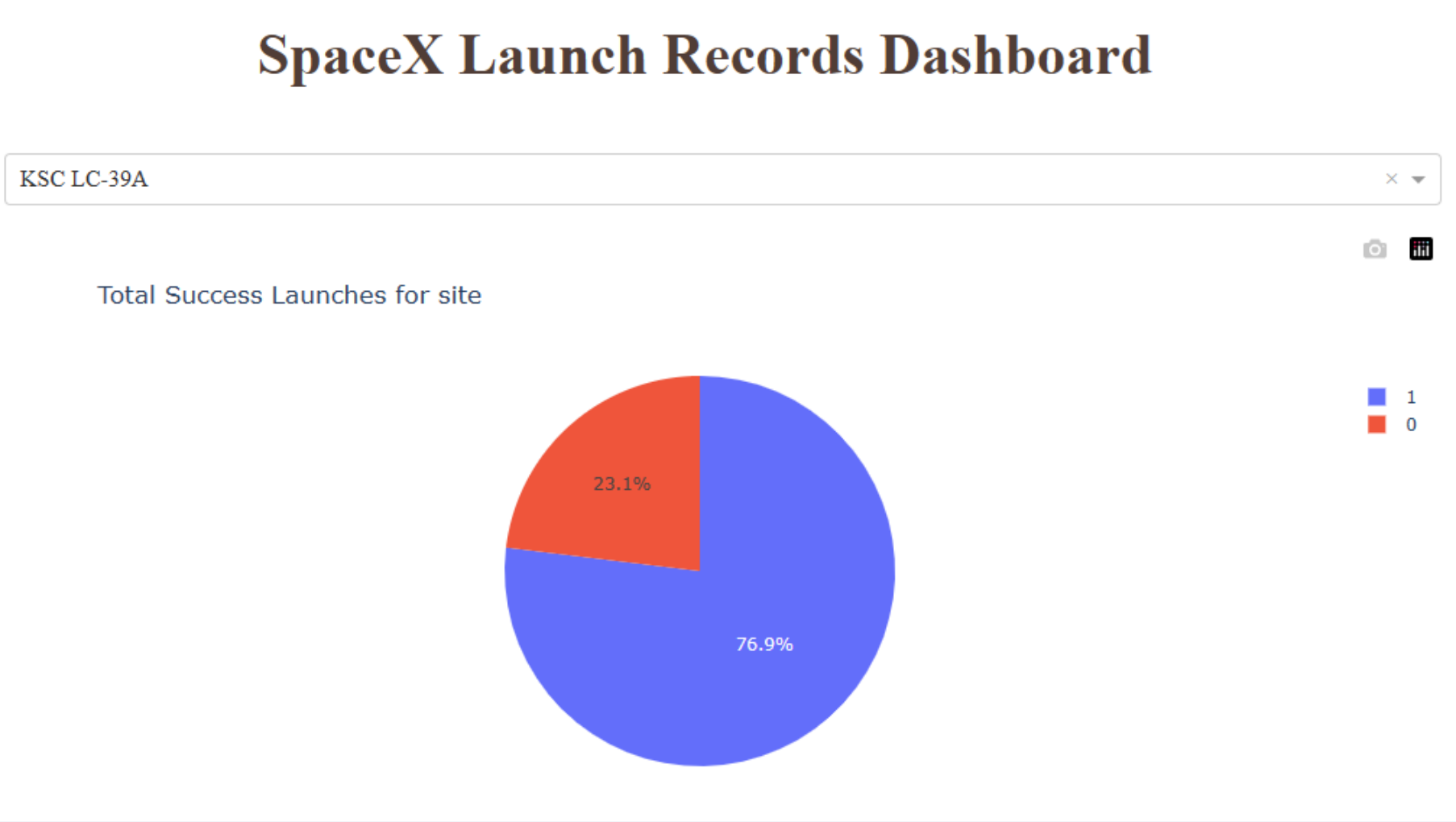
Successful Count for All Launch Sites

- The pie chart shows that launch site KSC LS-39A has highest successful counts



Launch Site with Highest Successful Ratio

- The piechart shows the launch site KSC LS-39A has highest successful ratio



Payload vs. Launch Outcome Plot

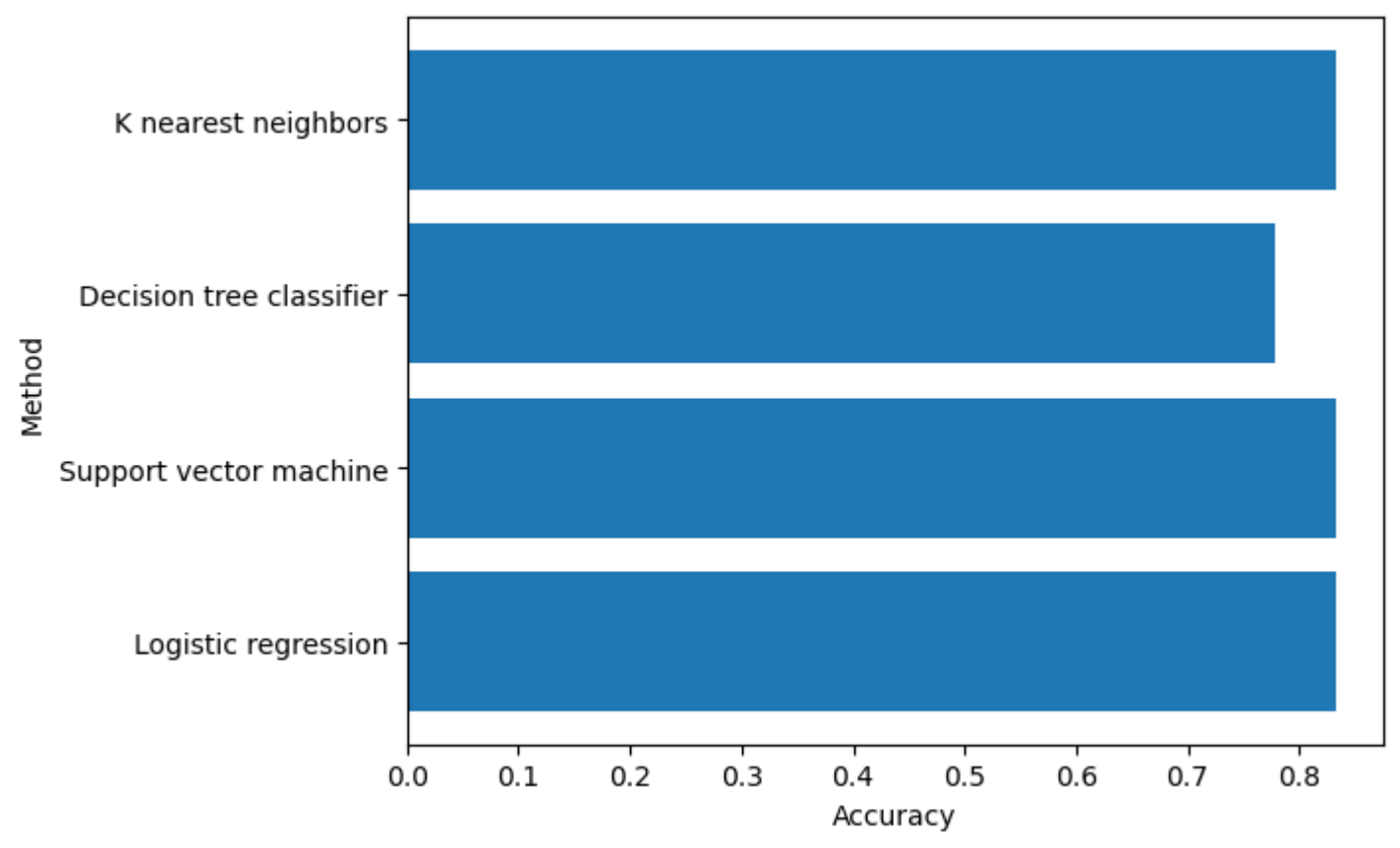
- This plot shows Payload vs. Launch Outcome relationship, and the green color of FT has highest successful ratio

Section 5

Predictive Analysis (Classification)

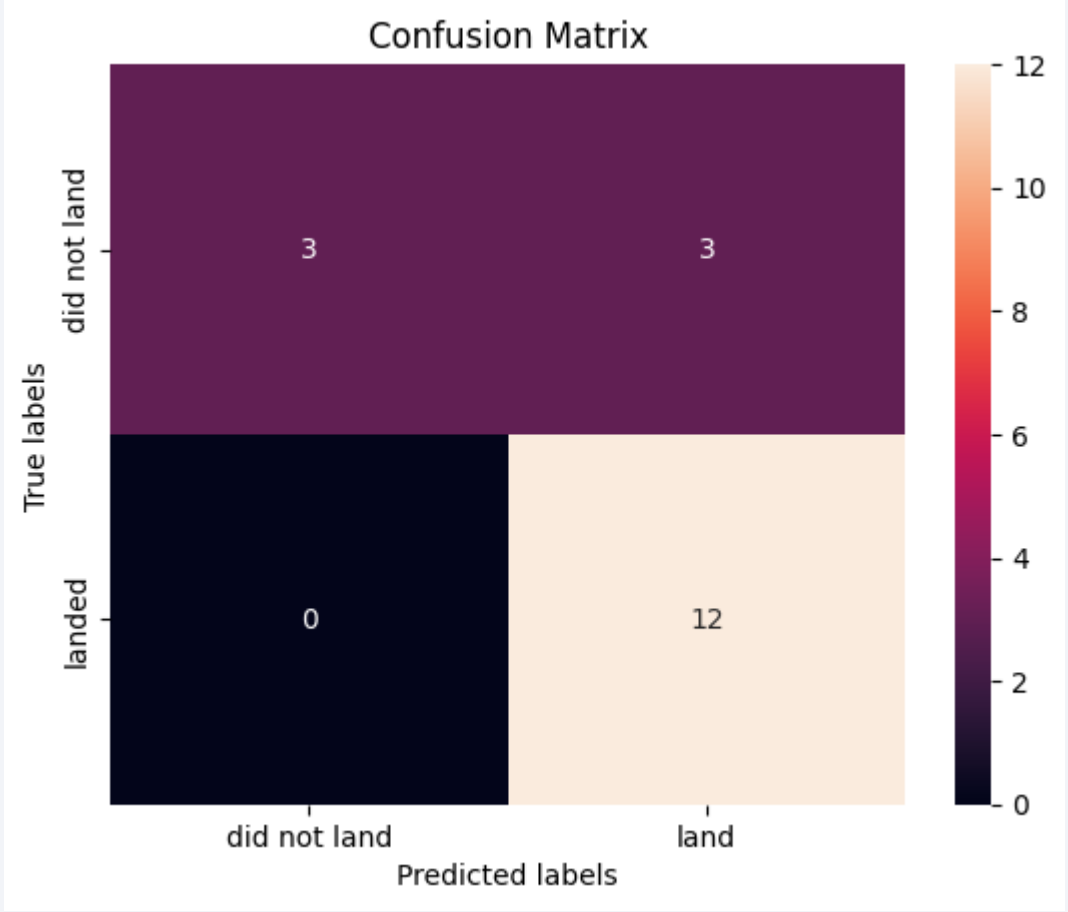
Classification Accuracy

- This is the bar chart for the accuracy of all built classification models
- Among four methods, three methods, K nearest neighbors, Logistic regression and Support vector machine have same accuracy
- Decision tree classifier has lower accuracy



Confusion Matrix

- This plot shows the confusion matrix of the best performing model
- TP = 12, TN = 3, FP = 3, FN = 0



Conclusions

- The data analysis shows that the successful rate improves significantly over years due to technology development and lesson learned
- Some orbits have higher successful rate, such as ES-L1 , GEO, SSO and HEO
- Lower orbits can carry more payload while higher orbits can take less payload
- The launch site selection is important. The site nearby needs to have good transportation system and the site should also have safe distance to cities and towns.
- Four classification algorithms were tested, three of them have same level of accuracy at 0.833 while Decision Tree Classifier has the lowest accuracy at 0.778

Appendix

- Github link: https://github.com/Yg9huatong/IBM_Course_Applied-Data-Science-Capstone

Thank you!