IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Emma Sung
5/31/2024

Outline

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

Executive Summary

- Summary of methodologies:
  - Data collection
  - 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

Introduction

- Objective:
  - Create a new rocket company based off of information from Space X rocket launches.
  - Predict whether Falcon 9 will land successfully, which will save money for the company
- Methodology:
  - Gather information about Space X and creating dashboards
  - Determine if SpaceX will reuse the first stage by training a machine learning model and use public information to predict if SpaceX will reuse the first stage.

Section 1

Methodology

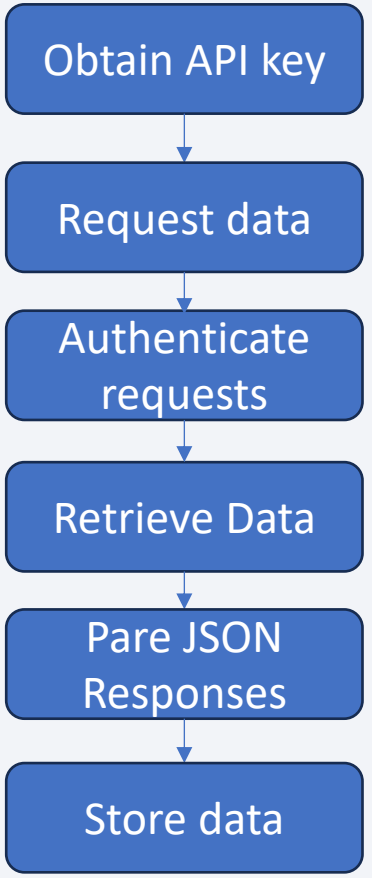
Methodology

Executive Summary

- Data collection methodology
- 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

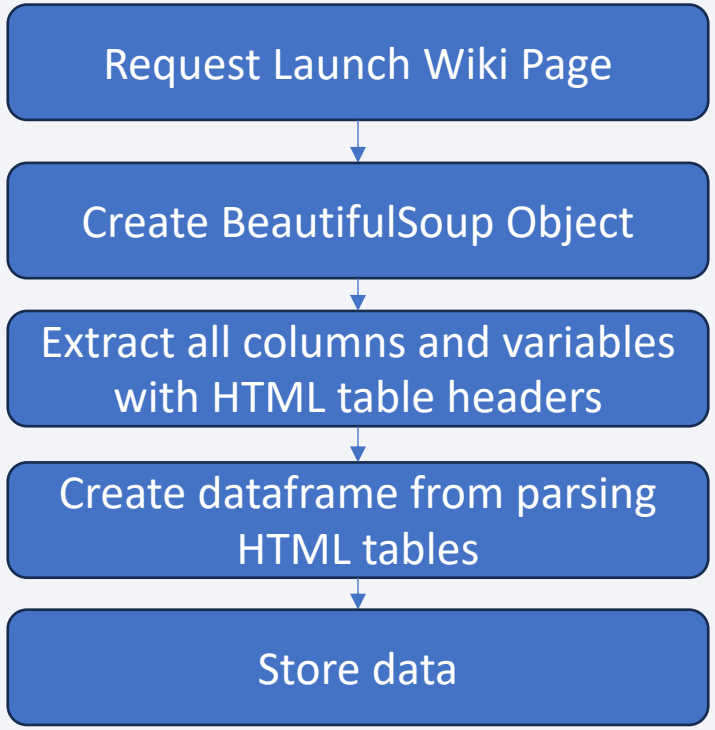
Data Collection – SpaceX API

- Data set was collected by requesting rocket launch data from SpaceX API
- Decoded the response content using Json and Json.normalize()
- Parsed and converted to dataframe
- Filtered the dataframe to only include Falcon 9
- https://github.com/emmasung1/Applied-Data-Science-Capstone/blob/main/Capstone/Week1/jupyter-labs-spacex-data-collection-api.ipynb



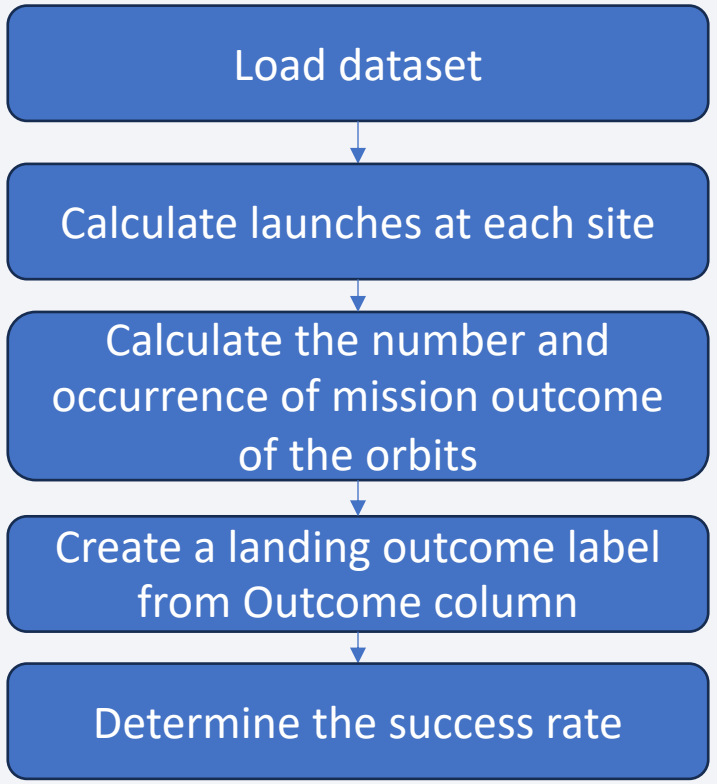
Data Collection - Scraping

- Perform web scraping to collect Falcon 9 historical launch records from a Wikipedia page
- Use the request to create BeautifulSoup object that allows us to parse through the HTML document of the Wiki page
- https://github.com/emmasung1/Applied-Data-Science-Capstone/blob/main/Capstone/Week1/jupyter-labs-webscraping.ipynb



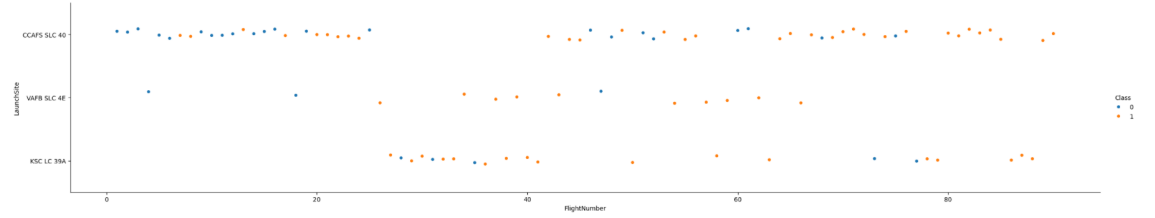
Data Wrangling

- Perform some Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models.
- Load dataset and parse through the data to create a list of possible outcomes
- Use the list of outcomes to label each data as success (value = 1) or fail (value = 0)
- Calculate the success rate by the average of outcome value
- https://github.com/emmasung1/Applied-Data-Science-Capstone/blob/main/Capstone/Week1/labs-jupyter-spacex-Data%20wrangling.ipynb

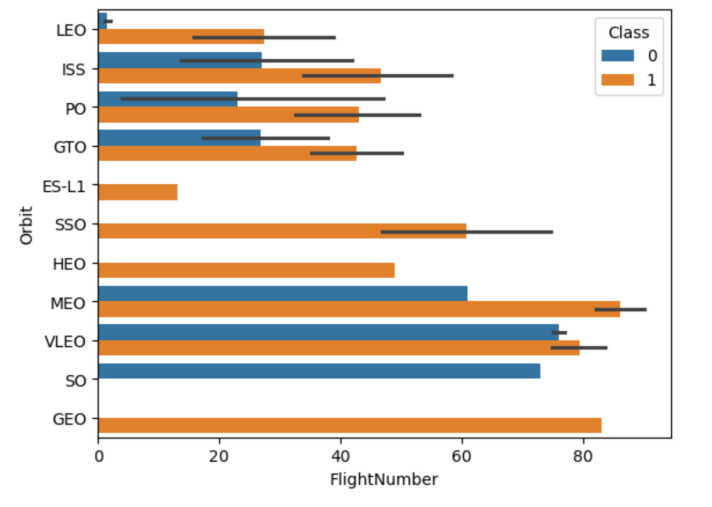


EDA with Data Visualization

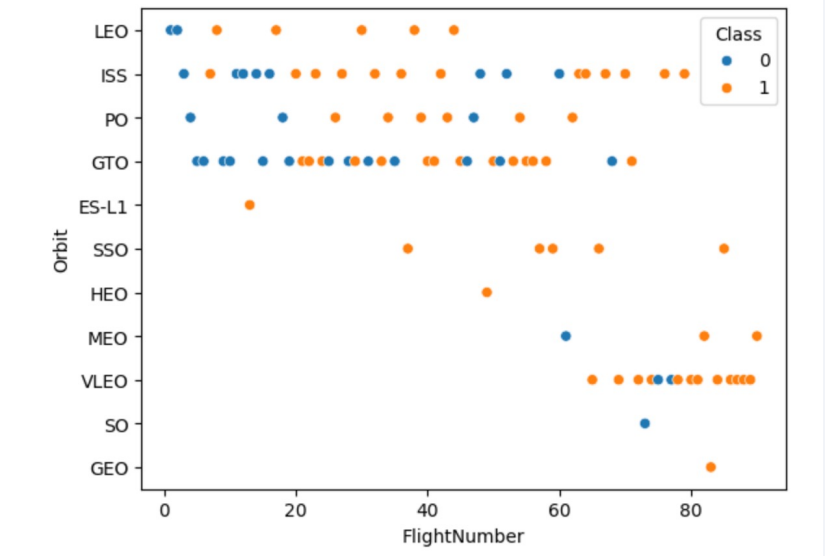
- Scatterplot of flight number vs. launch site
  - Visualized that there are consistent successful launches from flight number ~80+



- Bar plot of flight number vs. orbit
  - Visualize the average flight number for success/failure for each orbit.

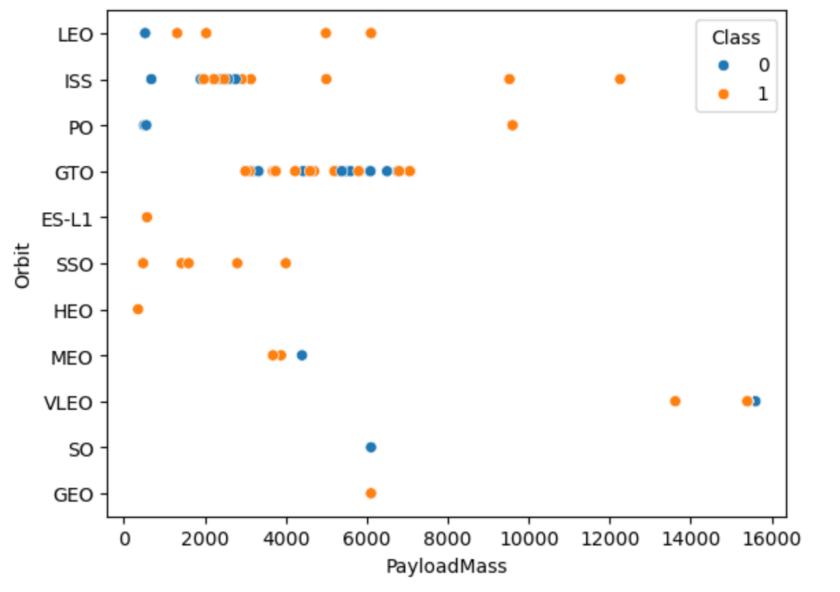


- Scatterplot of Flight Number and Orbit type
  - LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

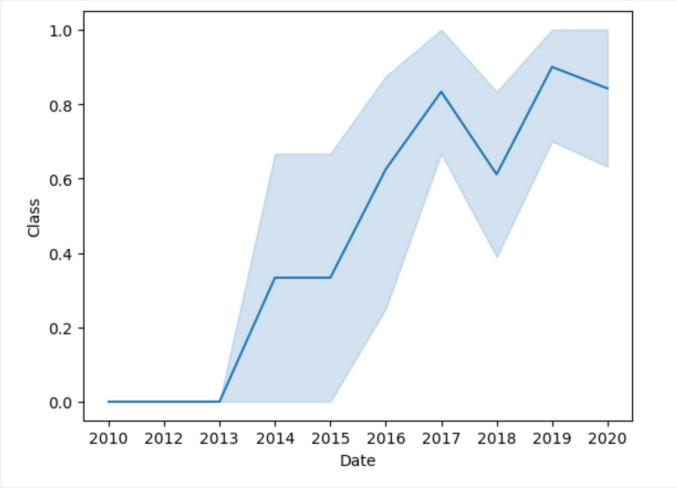


EDA with Data Visualization

- Scatterplot of Payload and 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.



- Line plot of yearly success rate
  - observe that the success rate since 2013 kept increasing till 2020



- https://github.com/emmasung1/Applied-Data-Science-Capstone/blob/main/Capstone/Week2/edadataviz.ipynb

EDA with SQL

- SQL queries performed:
  - Select
  - Distinct
  - Like
  - Average (AVG)
  - Sum
  - Min
  - Between
  - Count
  - Subquery

- https://github.com/emmasung1/Applied-Data-Science-Capstone/blob/main/Capstone/Week2/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

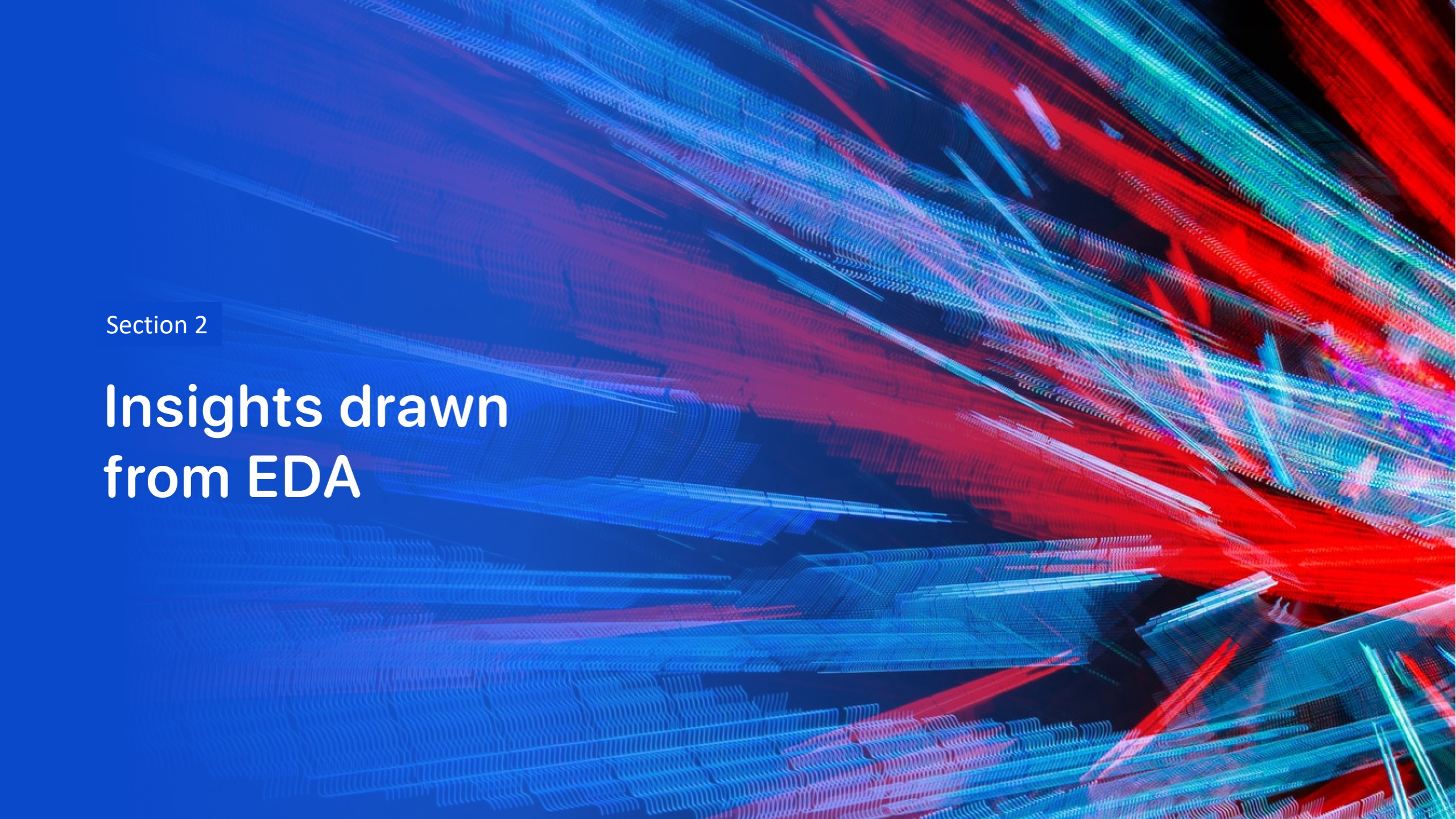
- Added markers and circles for each launch site to display visually where the launch sites are located
- Lines and distance markers to display distances between a launch site to its proximities (highway, beach, city)
- While launch sites are close in proximity to highways and beaches, they are further away from cities.
- https://github.com/emmasung1/Applied-Data-Science-Capstone/blob/main/Capstone/Week3/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Add a Launch Site Drop-down Input Component
  - There are four different launch sites
  - Visualize which one has the largest success count
  - Be able to select one specific site and check its detailed success rate (class=0 vs. class=1).
- Add a callback function to render pie chart based on selected site dropdown
  - get the selected launch site from site-dropdown and render a pie chart visualizing launch success counts.
- Add a Range Slider to Select Payload
  - Helps visualize if variable payload is correlated to mission outcome
- Add a callback function to render the payload scatter plot
  - Visually observe how payload may be correlated with mission outcomes for selected site(s)
- https://github.com/emmasung1/Applied-Data-Science-Capstone/blob/main/Capstone/Week3/spacex_dash_app.py

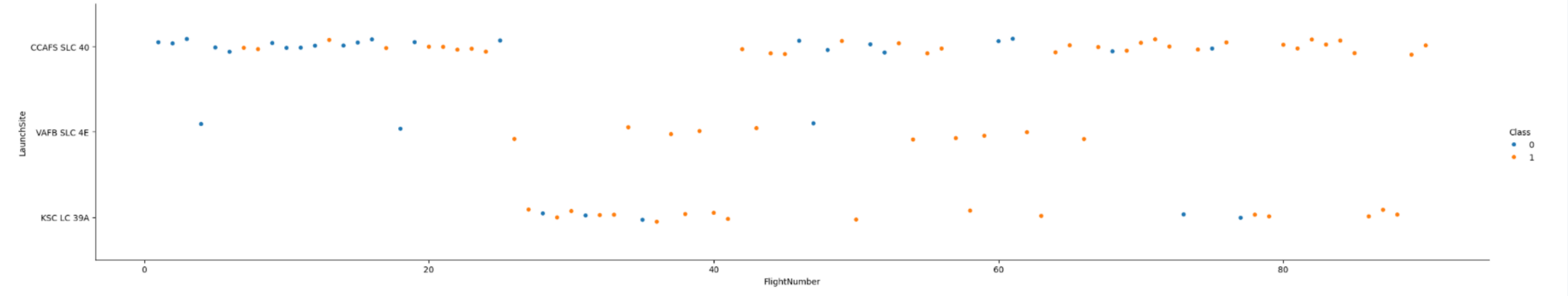
Predictive Analysis (Classification)

- Separate the dataset into training and testing groups
- Use the following models:
  - Logistic Regression
  - SVM
  - Decision Tree
  - KNN
- Train each model using the training data and output training accuracy
- Test the model using the testing data and generate testing data accuracy using score
- The best performing model was – based on the testing data accuracy.
- Gitlab: https://github.com/emmasung1/Applied-Data-Science-Capstone/blob/main/Capstone/Week4/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb


Section 2
Insights drawn from EDA

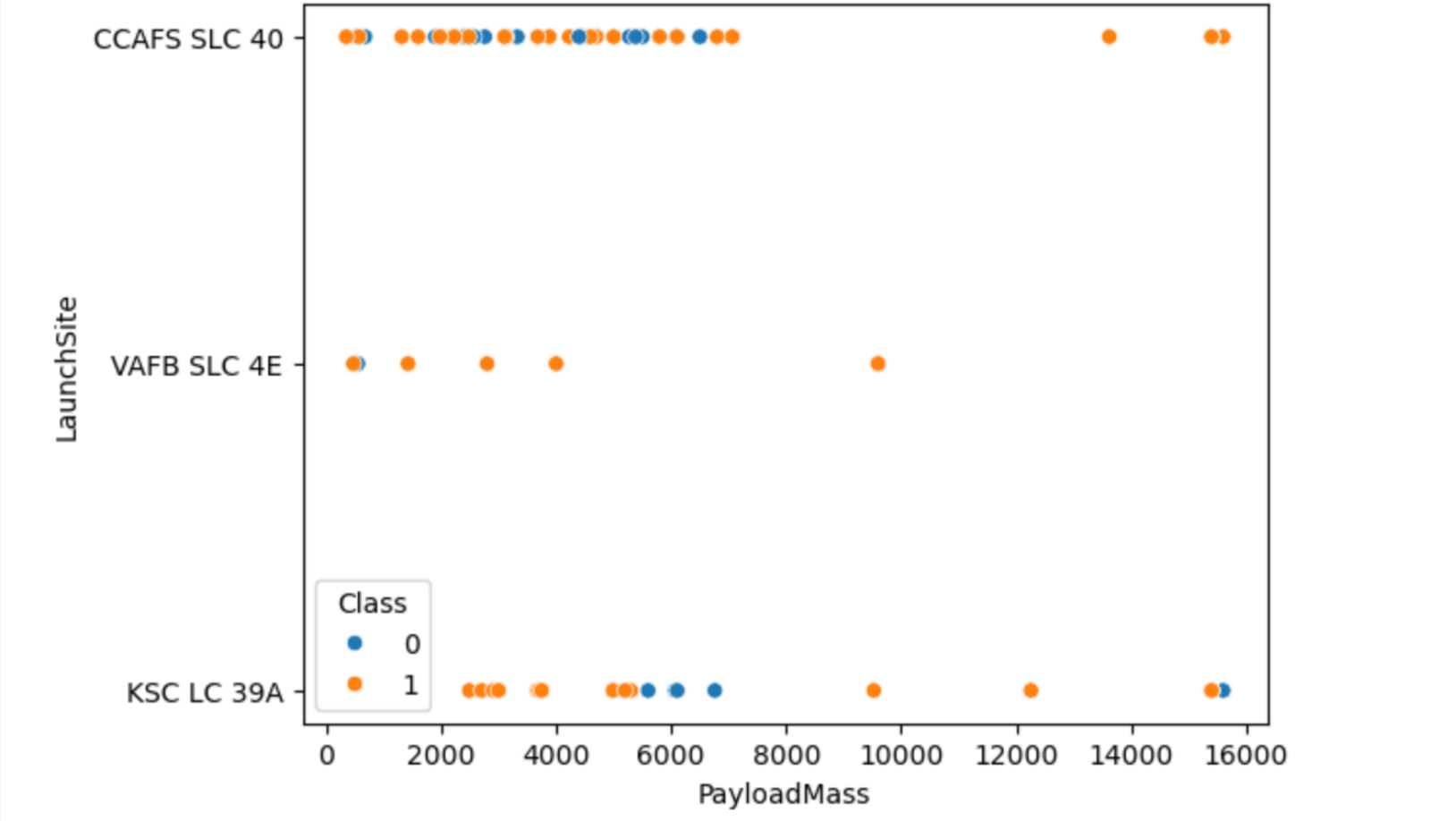
Flight Number vs. Launch Site

- Scatterplot of flight number vs. launch site
  - There are consistent successful launches above flight number ~78



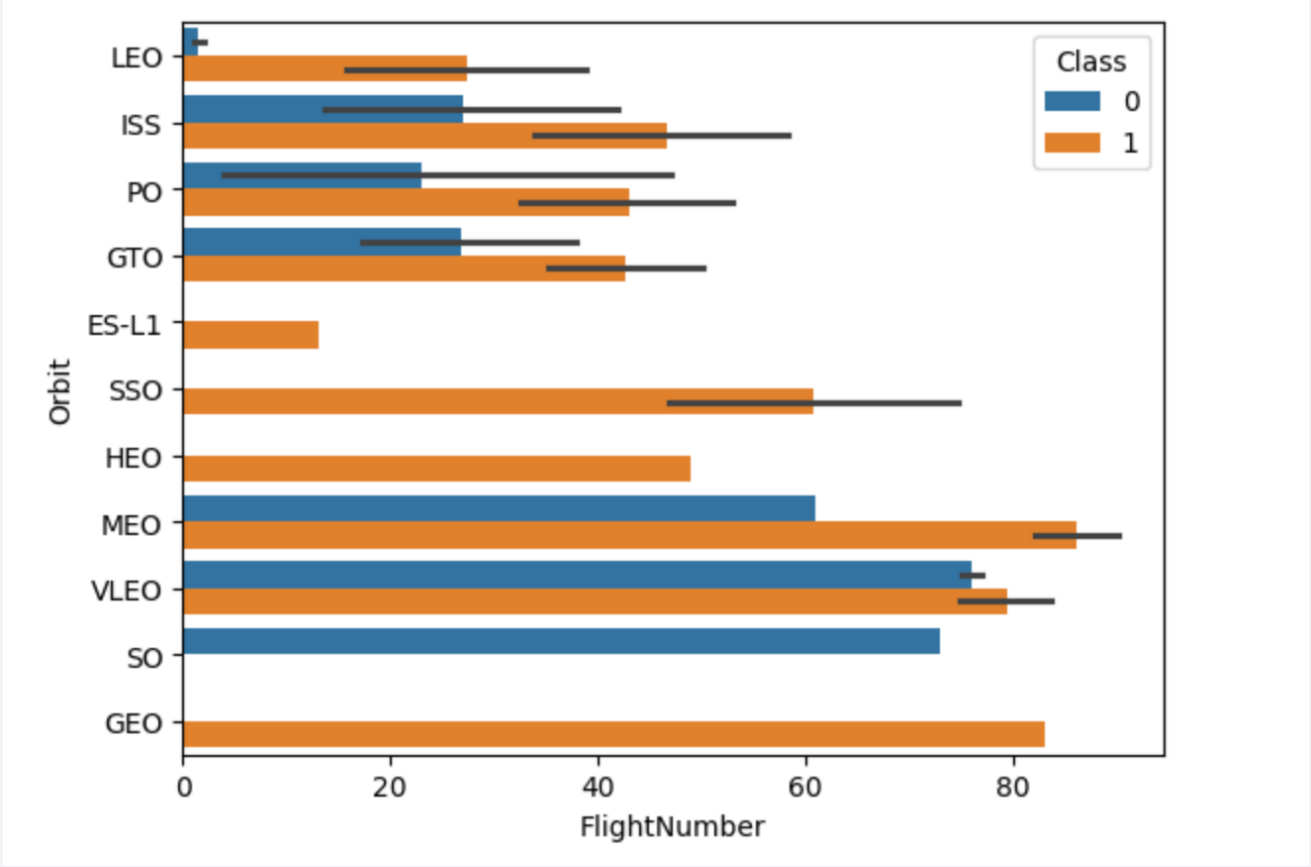
Payload vs. Launch Site

- For VAFB-SLC launch site, there are no rockets launched for heavy payload mass (greater than 10000).



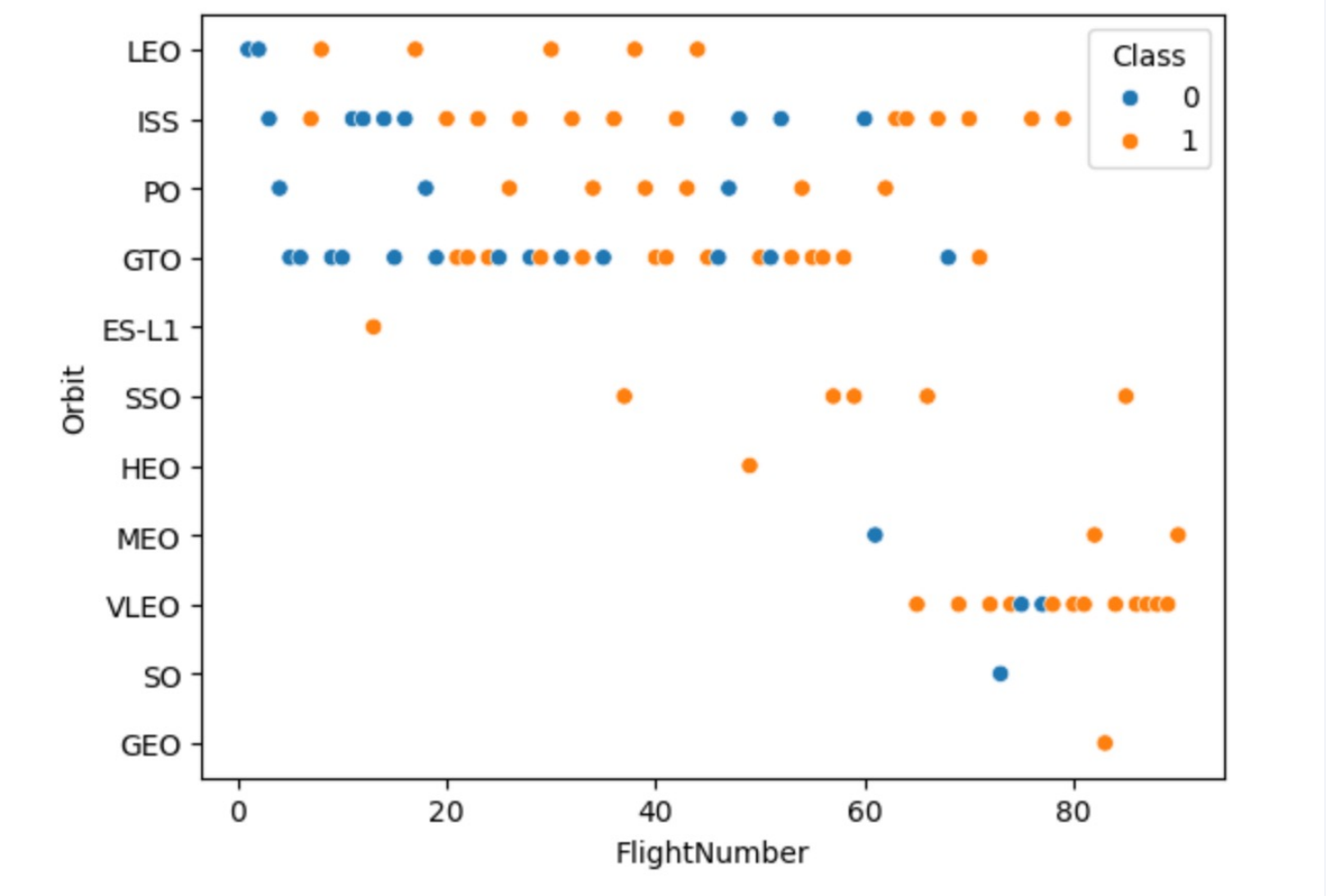
Success Rate vs. Orbit Type

- MEO and GEO has the highest success rate out of all of the orbits.



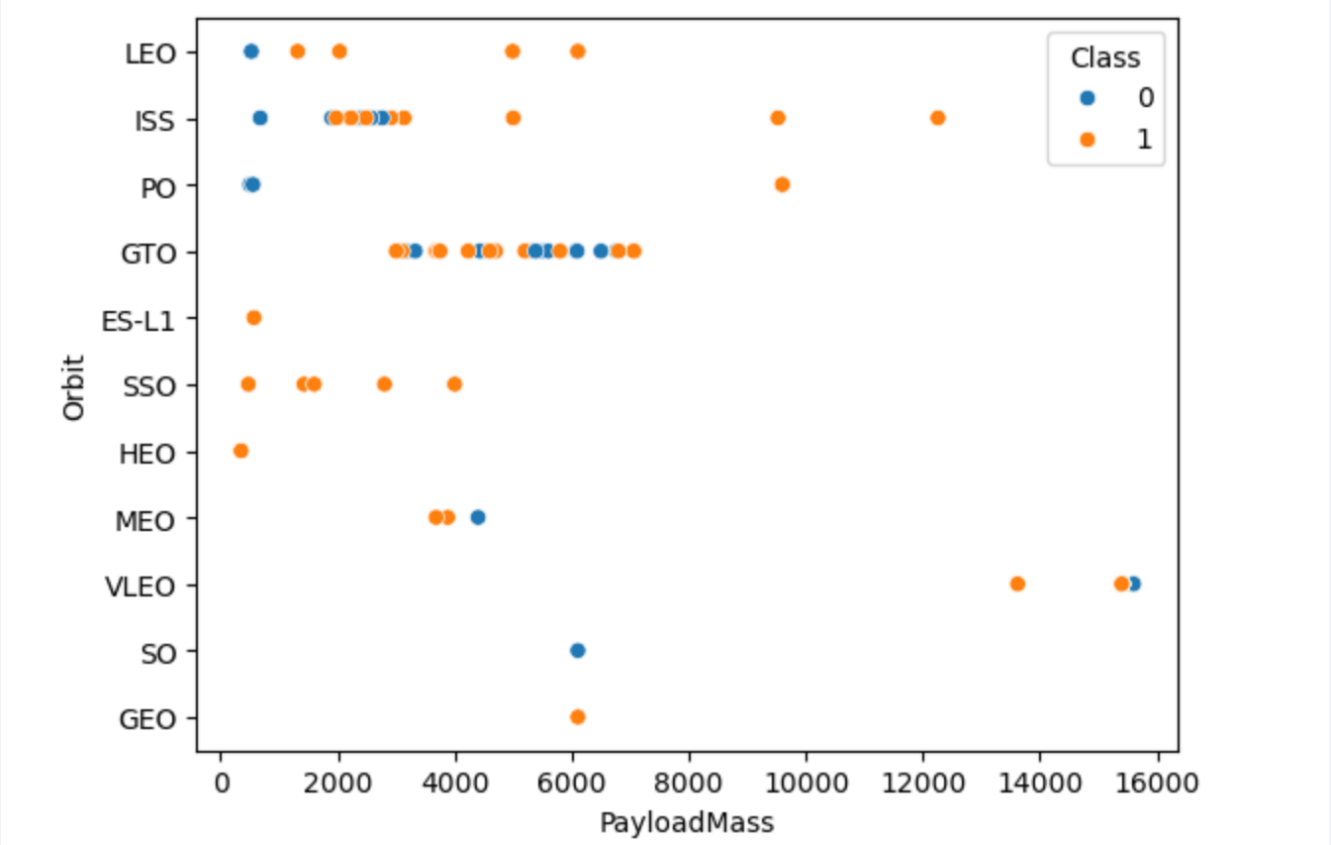
Flight Number vs. Orbit Type

- 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



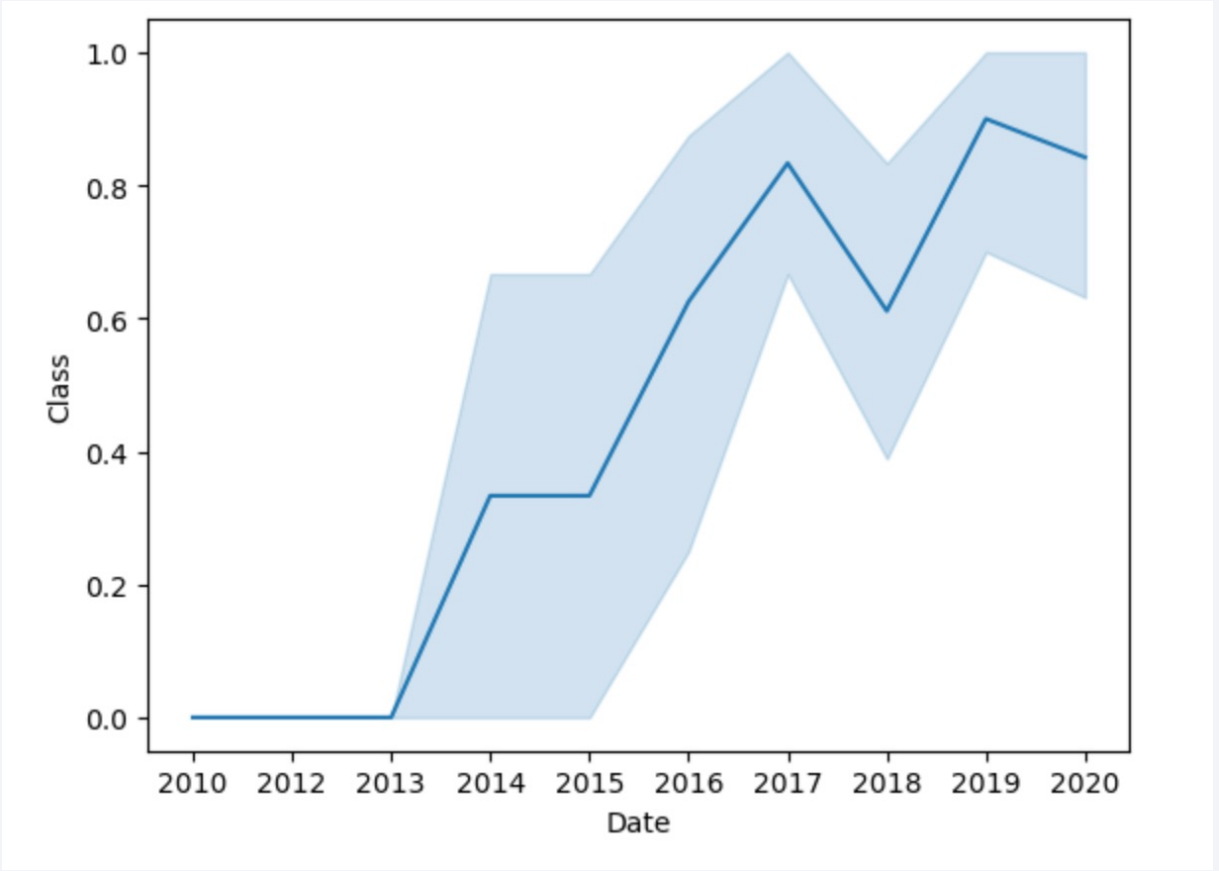
Payload vs. Orbit Type

- 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

- Observe that the success rate since 2013 kept increasing till 2020



All Launch Site Names

- Found names of unique launch sites by using Distinct query:
  - CCAFS LC-40
  - VAFB SLC-4E
  - KSC LC-39A
  - CCAFS SLC-40

Launch Site Names Begin with 'CCA'

- By using Like query and % in quotations, I can find the launch sites that begin with CAA

```
%sql select * from SPACEXTBL where Launch_Site like 'CCA%' limit 5
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (p
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 (p
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	N
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	N
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	N

Total Payload Mass

- Total payload carried by boosters from NASA = 45,596 kg
- This can be found by using SUM( ) and specifying WHERE the customer is NASA

```
%sql select SUM(PAYLOAD_MASS__KG_)as total from SPACEXTBL where Customer = "NASA (CRS)"
```

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

total
45596

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1 = 2,928.4 kg
- The average can be calculated by AVG( ) query and by specifying WHERE the version is F9 v1.1.

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

```
%sql select AVG(PAYLOAD_MASS__KG_) as average from SPACEXTBL where Booster_Version = "F9 v1.1"
```

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

average
2928.4

First Successful Ground Landing Date

- The first successful landing outcome on ground pad = 2018-07-22
- SQL will output the earliest date when the MIN query is used to find the first occurrence of success.

```
%sql select MIN(Date) from SPACEXTBL where Landing_Outcome = "Success"
```

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

MIN(Date)
2018-07-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The image shows the list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000.
- This was achieved by using the BETWEEN query.

```
%sql select booster_version from SPACEXTBL where PAYLOAD_MASS__KG_ between 4000 and 6000
```

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

Booster_Version
F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1014
F9 v1.1 B1016
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1030
F9 FT B1021.2
F9 FT B1032.1
F9 B4 B1040.1
F9 FT B1031.2
F9 B4 B1043.1
F9 FT B1032.2
F9 B4 B1040.2
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1046.3
F9 B5B1054
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5B1060.1
F9 B5 B1058.2
F9 B5B1062.1

Total Number of Successful and Failure Mission Outcomes

- Total number of successful outcomes = 100
- Total number of failure mission outcomes = 1

```
%sql SELECT Mission_Outcome, COUNT(*) AS Total_Outcomes FROM SPACEXTBL GROUP BY Mission_Outcome;
```

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

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

Boosters Carried Maximum Payload

- Below image shows the names of the booster which have carried the maximum payload mass

```
[22]: %sql SELECT booster_version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
 * sqlite:///my_data1.db
Done.
```

[22]:

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

- The image below shows the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015.
- This occurred only on the launch site CCAFS LS-40.

```
%sql SELECT substr(Date,6,2) as Month, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTBL where landing
```

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

Month	Landing_Outcome	Booster_Version	Launch_Site
01	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

- The data below shows 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.
- All successful outcomes were accomplished on 2016.

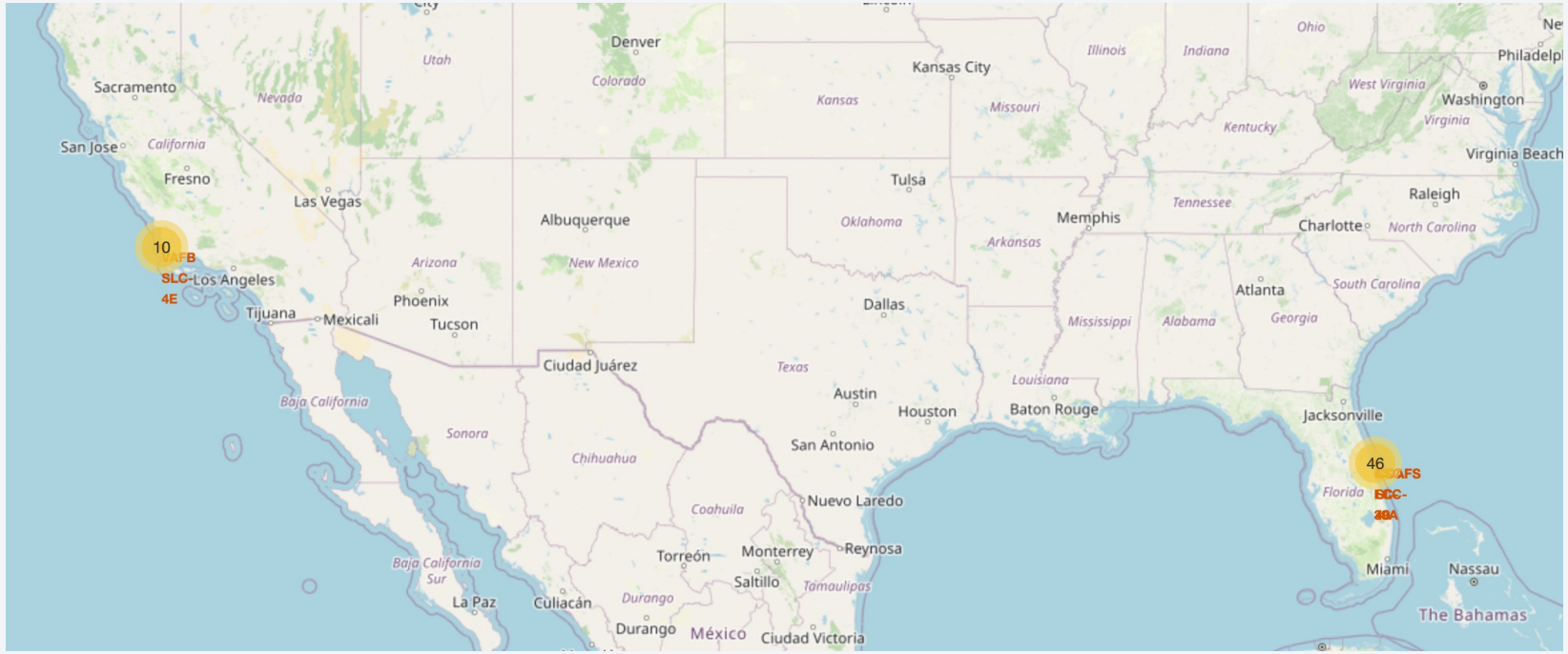
```
%sql SELECT date, Landing_Outcome, count(*) as outcome_count FROM SPACEXTBL where date between '2016-06-04' and '
```

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

Date	Landing_Outcome	outcome_count
2016-07-18	Success (ground pad)	2
2016-08-14	Success (drone ship)	2
2017-03-16	No attempt	1
2016-06-15	Failure (drone ship)	1


Section 3
Launch Sites
Proximities Analysis

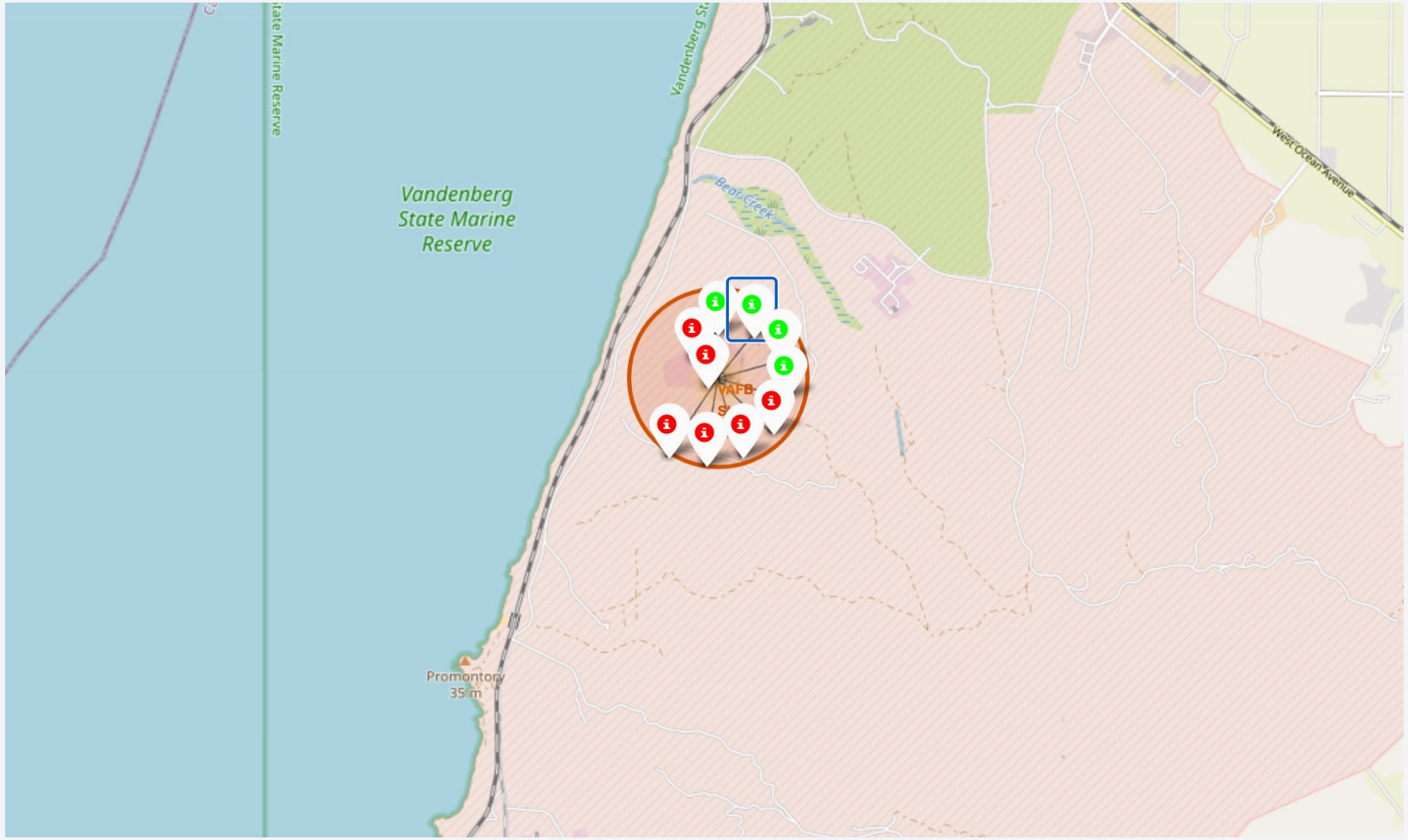
All Launch Sites' Location



- There are 10 launch sites located in California and 46 launch sites located in Florida.
- From the map shown above, launch sites are located near shores and locations with mild weather.

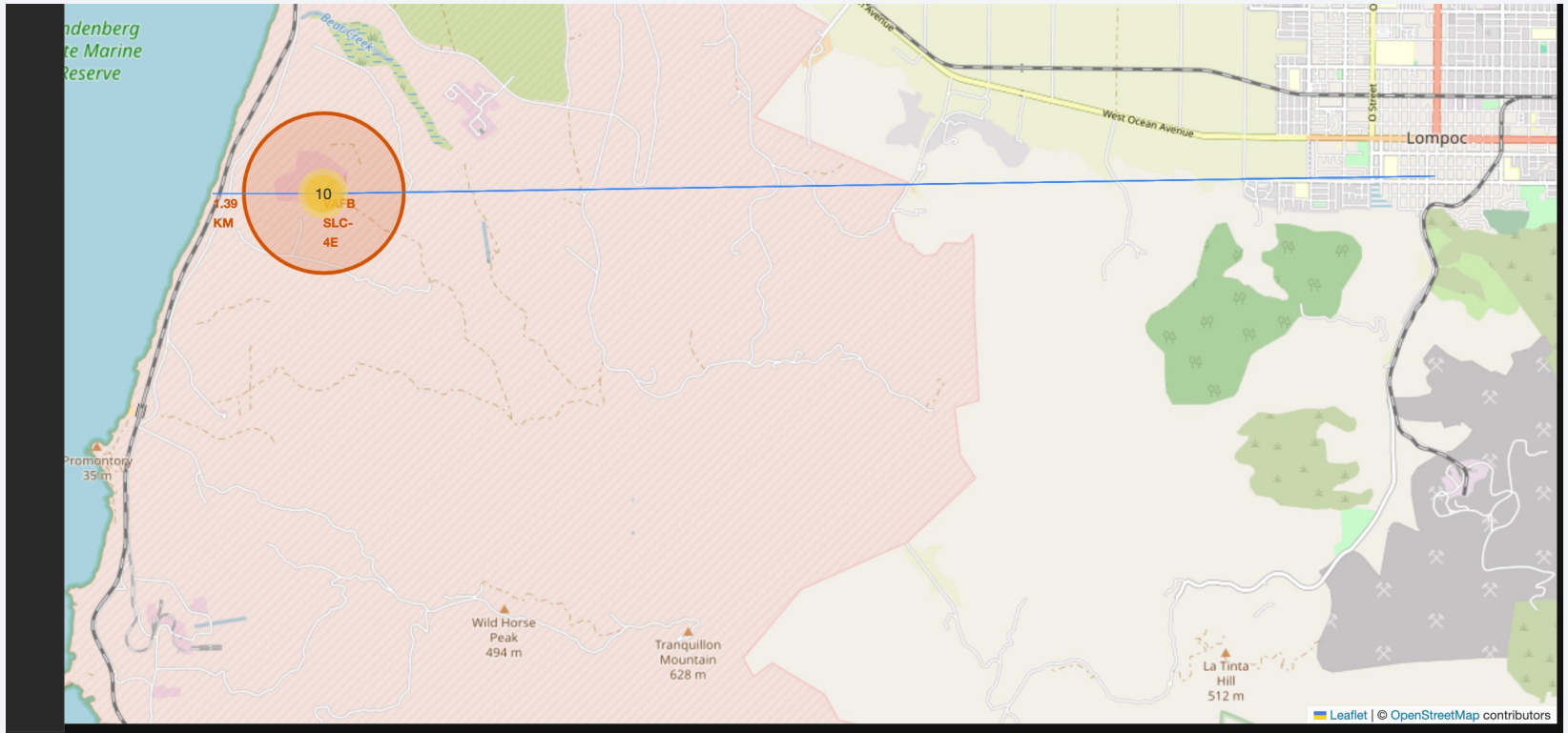
Mark the success/failed launches for each site on the map

- Folium map shows the color-labeled launch outcomes on the map
- Green indicates successful launches and red indicates failed launches.

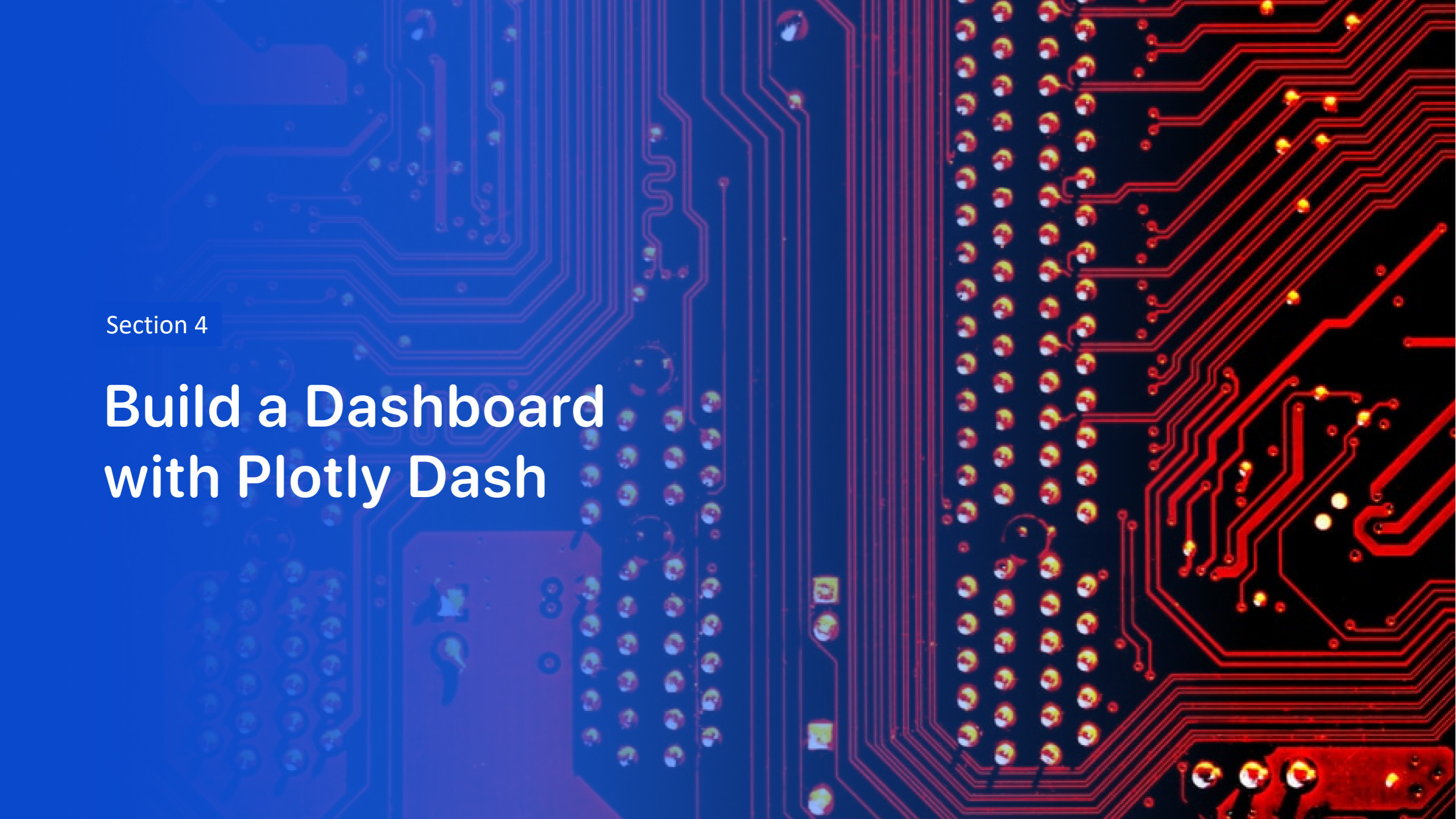


Calculate the distances between a launch site to its proximities

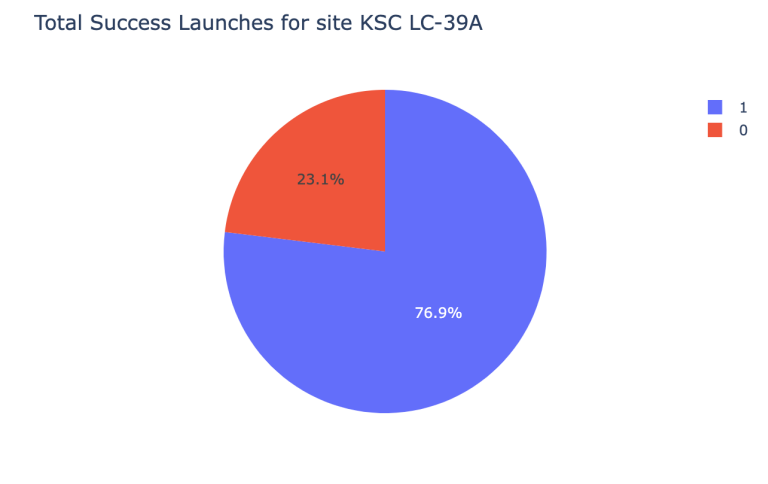
- The generated folium map shows proximities such as railway, highway, coastline to the site.
- Highway and railways are nearby a launch site. However, a city is far away.

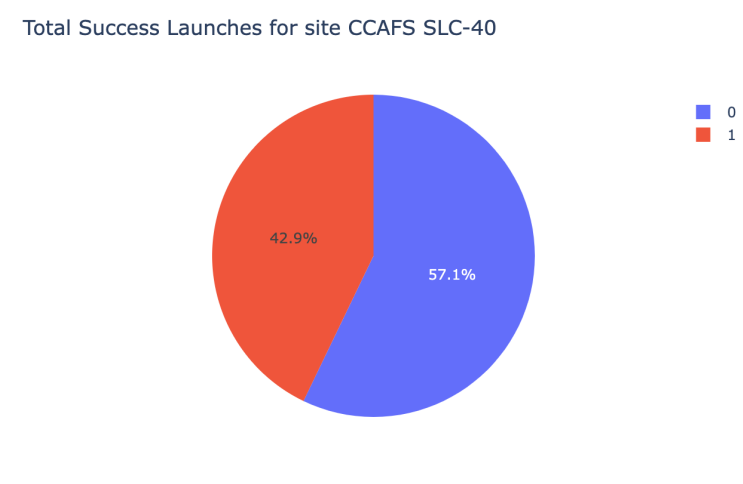


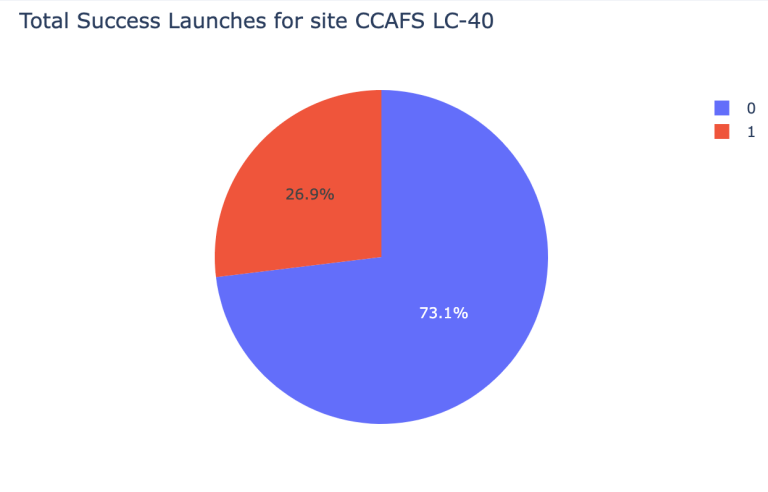

```
[38]: distance_city = calculate_distance(launch_site_lat, launch_site_lon, city_coordinate[0], city_coordinate[1])
      distance_railway = calculate_distance(launch_site_lat, launch_site_lon, railway_coordinate[0], railway_coordinate[1])
      distance_highway = calculate_distance(launch_site_lat, launch_site_lon, highway_coordinate[0], highway_coordinate[1])
      print(distance_city)
      print(distance_railway)
      print(distance_highway)

      13.96202164293696
      1.3005150452431087
      1.184283229262027
```

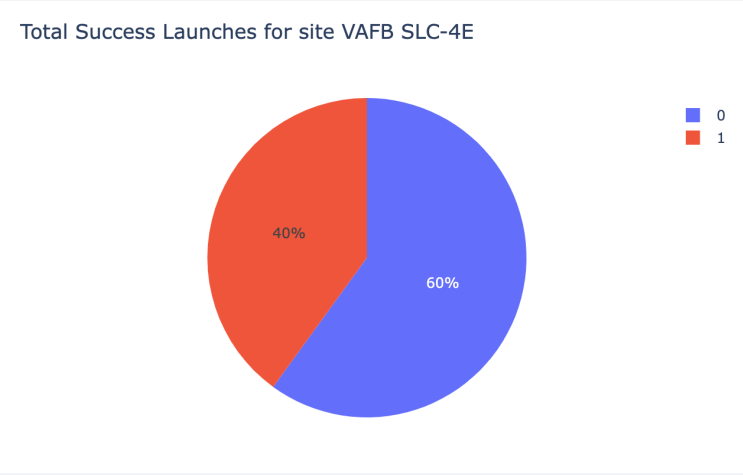

Section 4
Build a Dashboard with Plotly Dash

Launch Success of All Sites



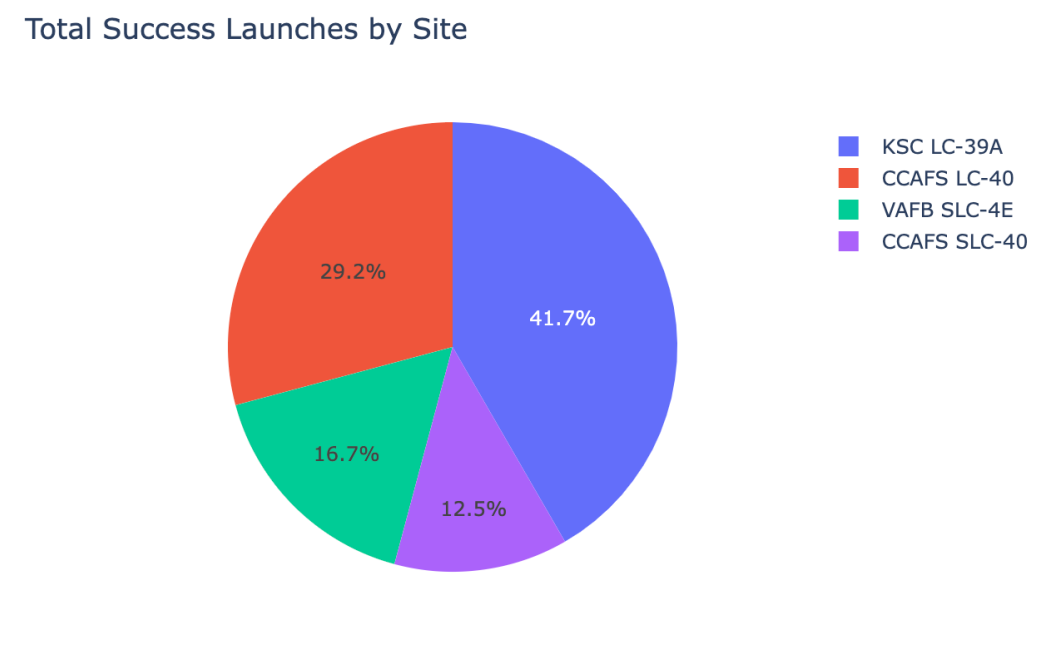


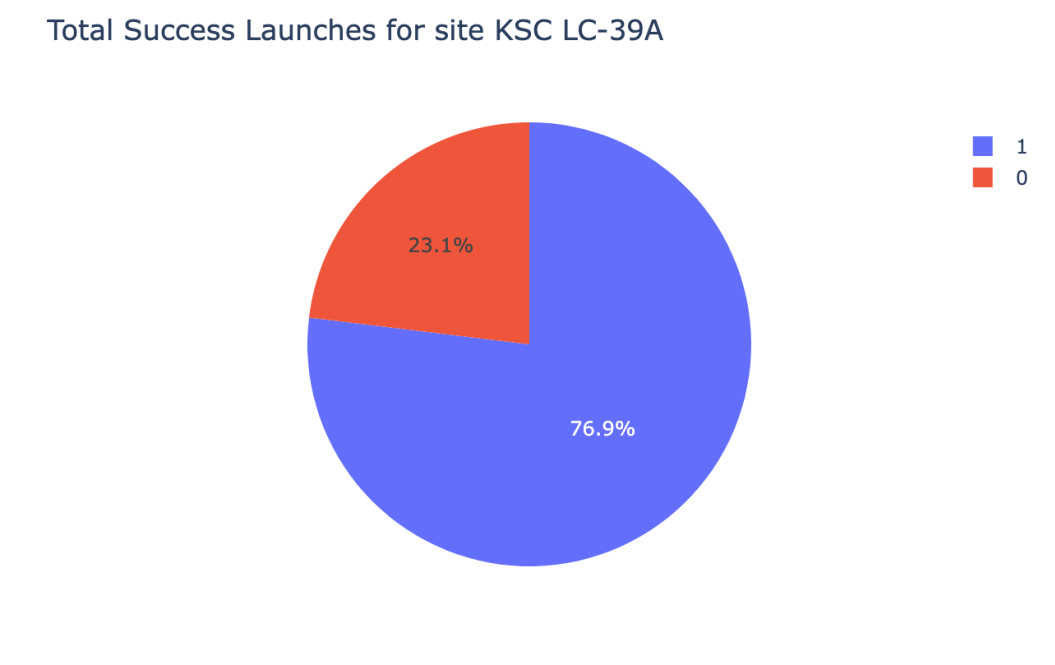




- Site KSC LC-39A has the highest successful launches. Followed by CCAFS LC-40.

High Launch Success Ratio





- KSC LC-39A has the highest launch success ratio compared to other launch sites.

Payload vs Launch Outcome Scatter Plots

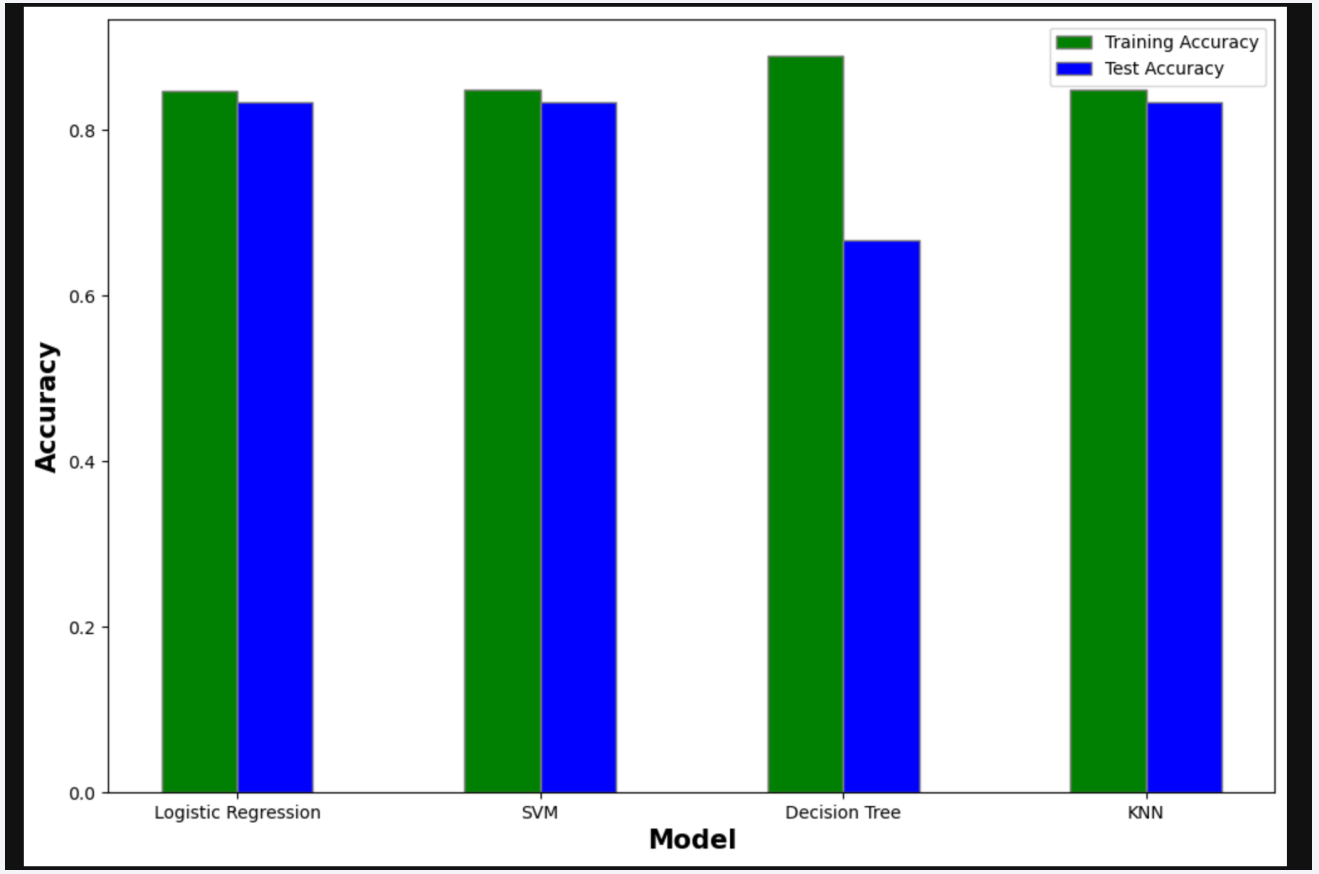
- FT has the largest success rate while v1.1 has the lowest success rate.
- Payload range of 2k-4k has the highest number of successful launches
- Only B4 has attempted launches above 7k and has a 50/50 chance of success.
- FT has a higher success rate between payload ranges of 2k-3k.

Section 5

Predictive Analysis (Classification)

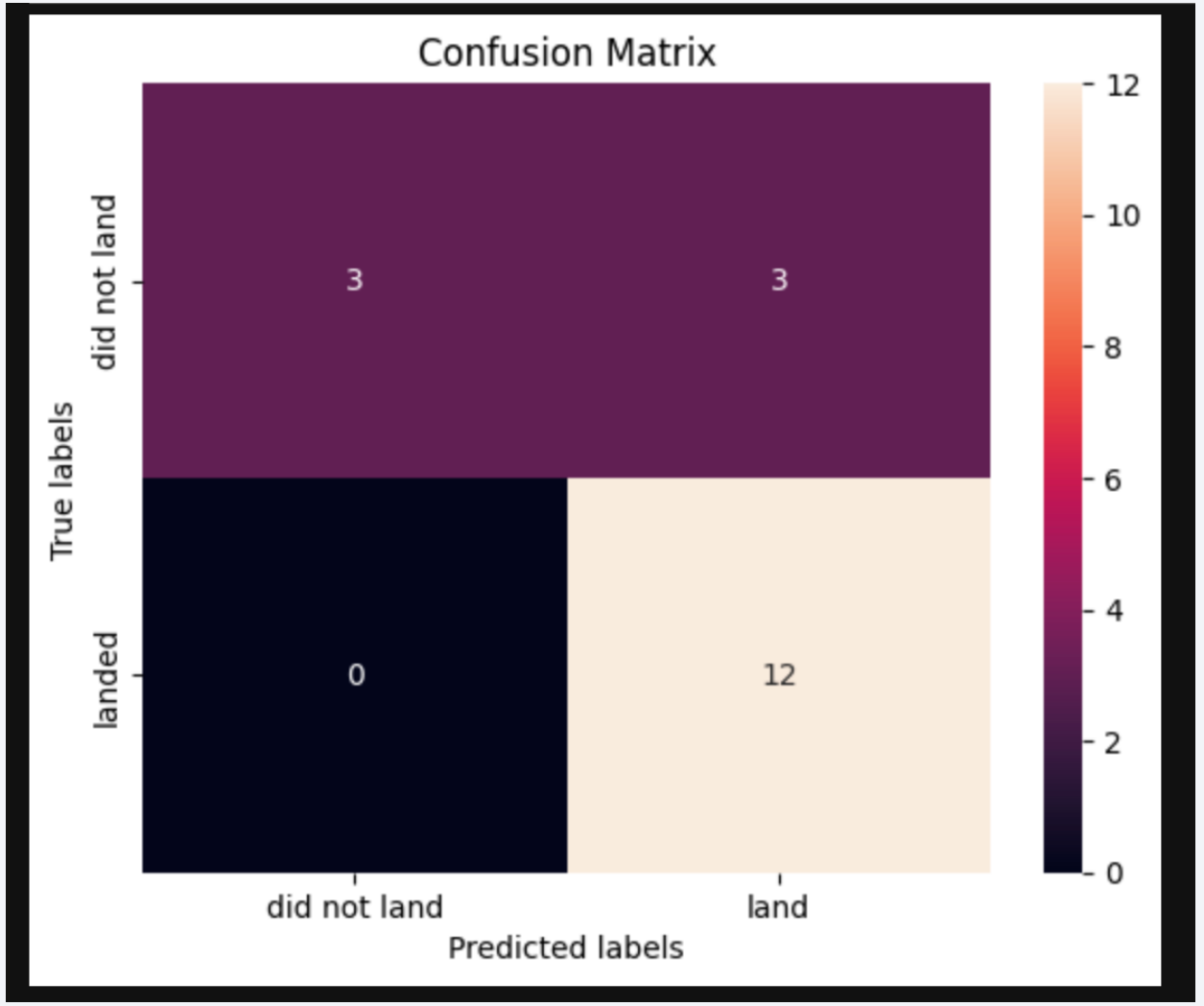
Classification Accuracy

- Decision tree has the highest classification accuracy of 88.9%. However, it has the lowest test accuracy.
- Logistic regression, SVM and KNN all have very similar test accuracies.



Confusion Matrix

- Logistic regression, SVM and KNN all have very similar test accuracies.
- They also have the same confusion matrixes as shown on the side.



Conclusions

- Success rates generally increasing with time.
- Sites CCAFS LC-40 or CCAFS SLC-40 have highest chances of success.
- Payloads below 6000kg have higher changes of success.
- Logistic regression, SVM and KNN all have very similar test accuracies, approx. 83%.

Thank you!