IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Keith Tutwiler
5/2/2024

Outline

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

Executive Summary

- **Methodology:** To acquire the necessary data, we utilized Python, employing the SpaceX API for primary data collection and the Beautiful Soup library for web scraping and data cleaning. Missing values were addressed to ensure the integrity of our analysis. Subsequently, we conducted an EDA primarily using SQL, supplemented by pandas and Matplotlib for visualizations. Key factors such as flight number, launch site, and booster type were examined to assess their impact on the landing outcome. Furthermore, we employed Folium to create an interactive map of launch sites and developed a Dash application to visualize the proportion of successful and failed launches across different sites. Additionally, we explored the correlation between payload mass and launch success/failure.

Results: Through our analysis, we identified several noteworthy findings:

•Successful landing outcomes were influenced by factors such as launch site and booster type.

•An interactive map highlighted spatial trends in launch outcomes.

•Predictive analysis using KNN, logistic regression, SVM, and decision trees yielded a prediction score of 83%, with minimal false positives.

•Orbit type emerged as a significant determinant of launch success, providing insights into optimal trajectories.

Introduction

- SpaceX is looking for an innovative way to reduce costs and increase efficiency
- One way of doing so is reusing stage 1 rockets.
- I want to find the key metrics that determine what makes a rocket landing success or not.
- This will enable SpaceX to save money by investing in the key factors that make their rockets reusable.

Section 1

Methodology

Methodology

Data collection methodology:

- My primary method of collecting data involve importing data from the SpaceX API
- Also, I scraped the Falcon 9 and Falcon Heavy Launches data from the SpaceX Wiki website

Perform data wrangling

- I use the Requests and Beautiful soup library to obtain the web information and parse the data then convert it to a dataframe
- Then I was able to calculate the number of Launches at each site and determine their success rate

Perform exploratory data analysis (EDA) using visualization and SQL

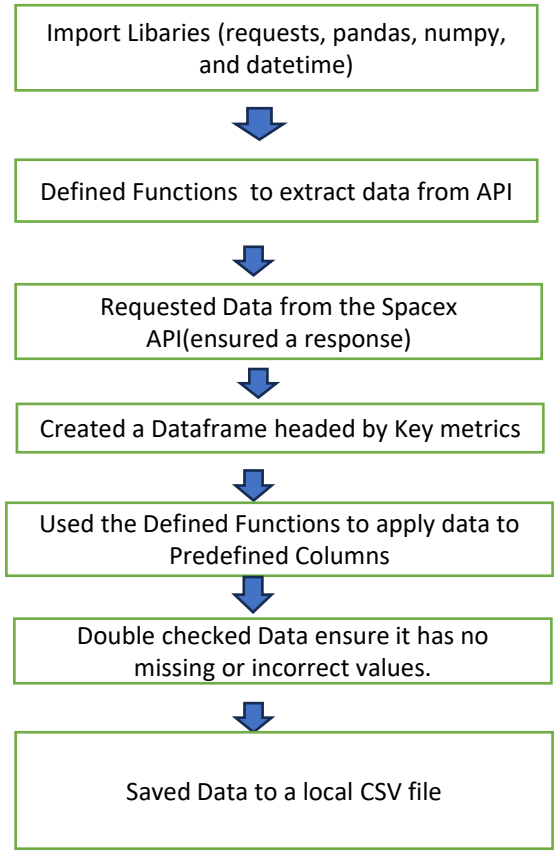
Perform interactive visual analytics using Folium and Plotly Dash

Perform predictive analysis using classification models

- Afterwards I split the data into training and testing data. I then trained the data using, logistic regression, Support Vector Machines, K nearest Neighbor, and a decision tree.
- I used a confusion matrix to help determine the accuracy of the models and used the best_score_ method to determine the accuracy of my models

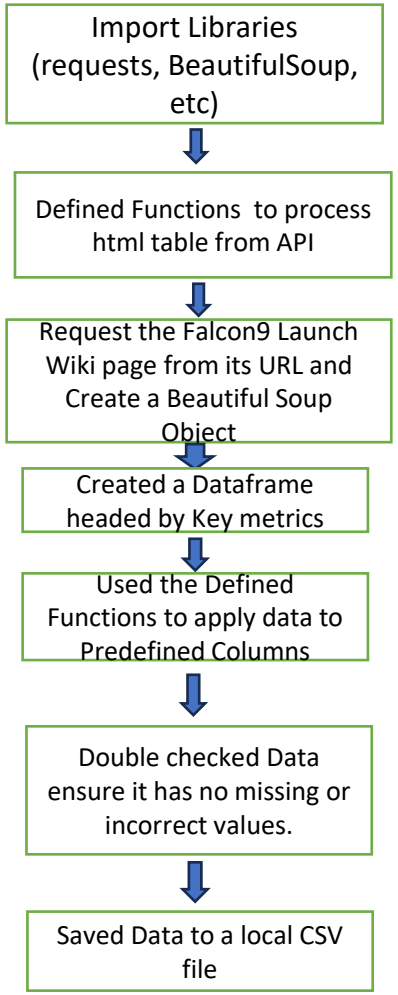
Data Collection – SpaceX API

Here is the notebook that covers the entire process



Data Collection - Scraping

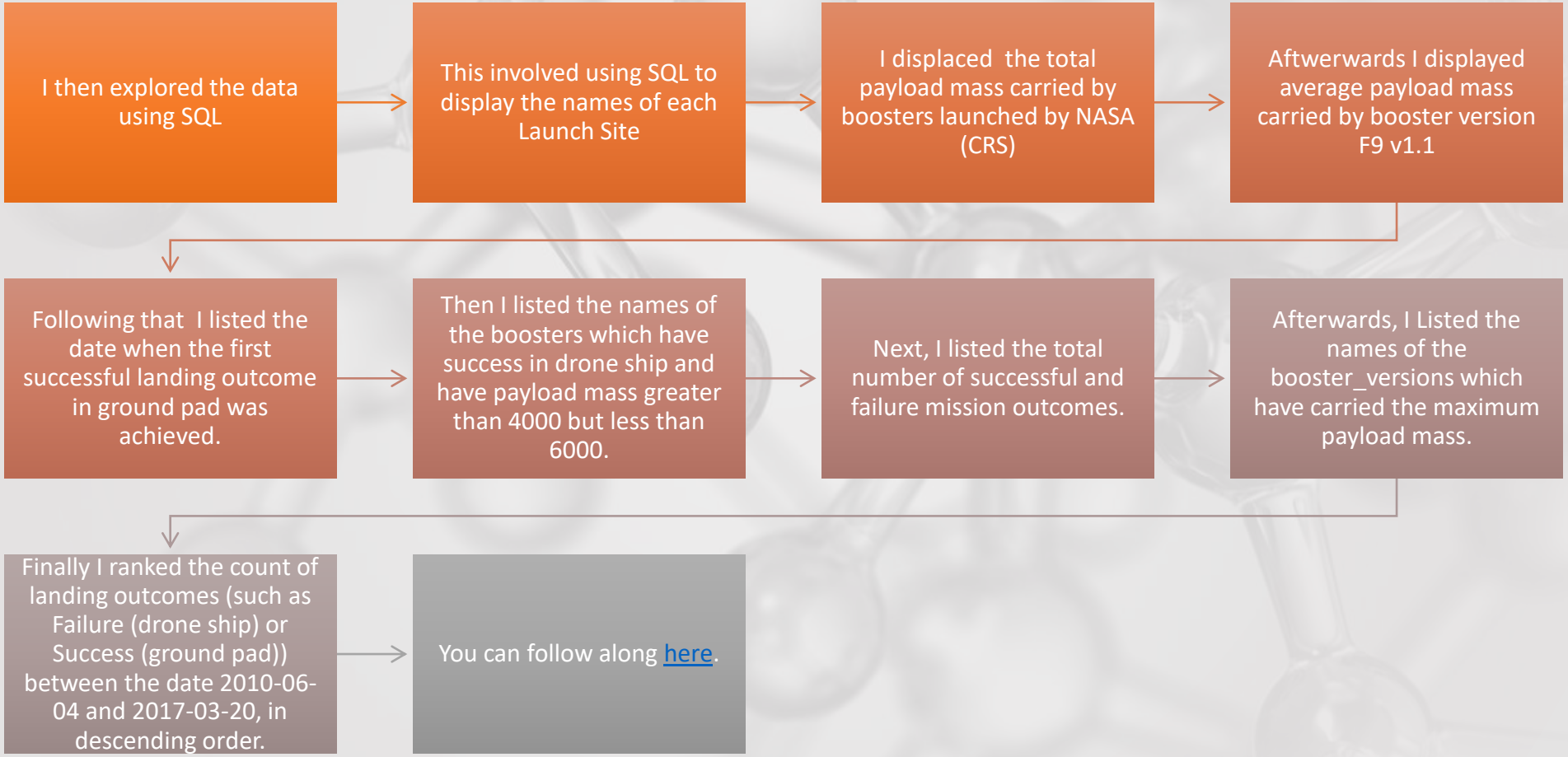
If you interested In seeing the Step by Step Process you can follow along here

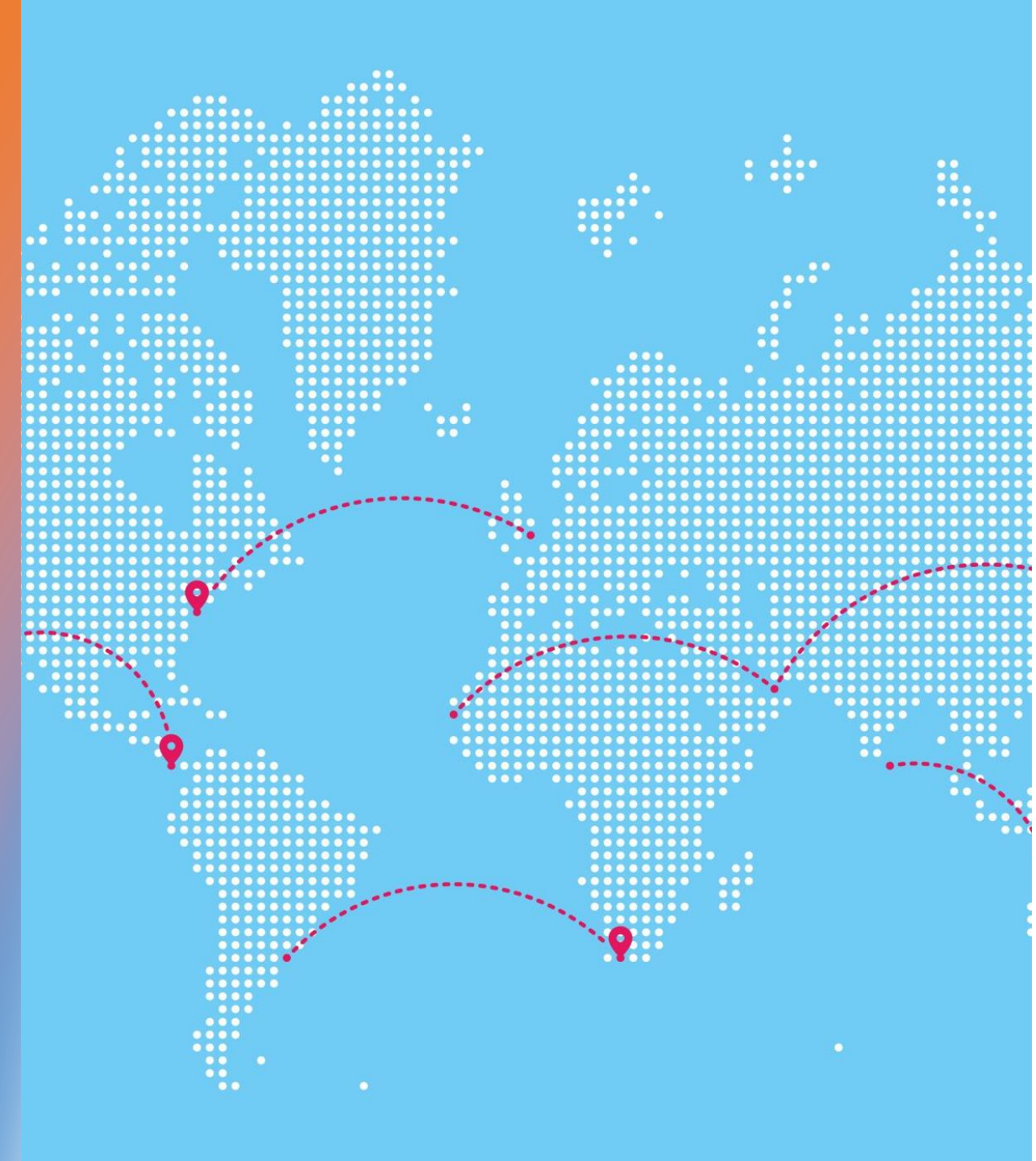


EDA with Data Visualization

- Next, we used a Scatterplot to visualize the correlation between Flight Number and Payload Mass(KG). In the visualization we used a hue to show which launch had successful landings and which one had failures.
- Following that I visualized the correlation between Flight Number and Orbit type while also showing hue for successful and failed Launches
- Then I visualized the correlation between Paload Mass(kg) and Orbit type while also showing hue for successful and failed Launches
- I then showed a comparison of the percentage of Success rate for each orbit type.
- Finally, I showed the correlation of Orbit type and Flight Number, also I showed the correlation of Orbit Type and Payload Mass(Kg)
- You can see each step here.

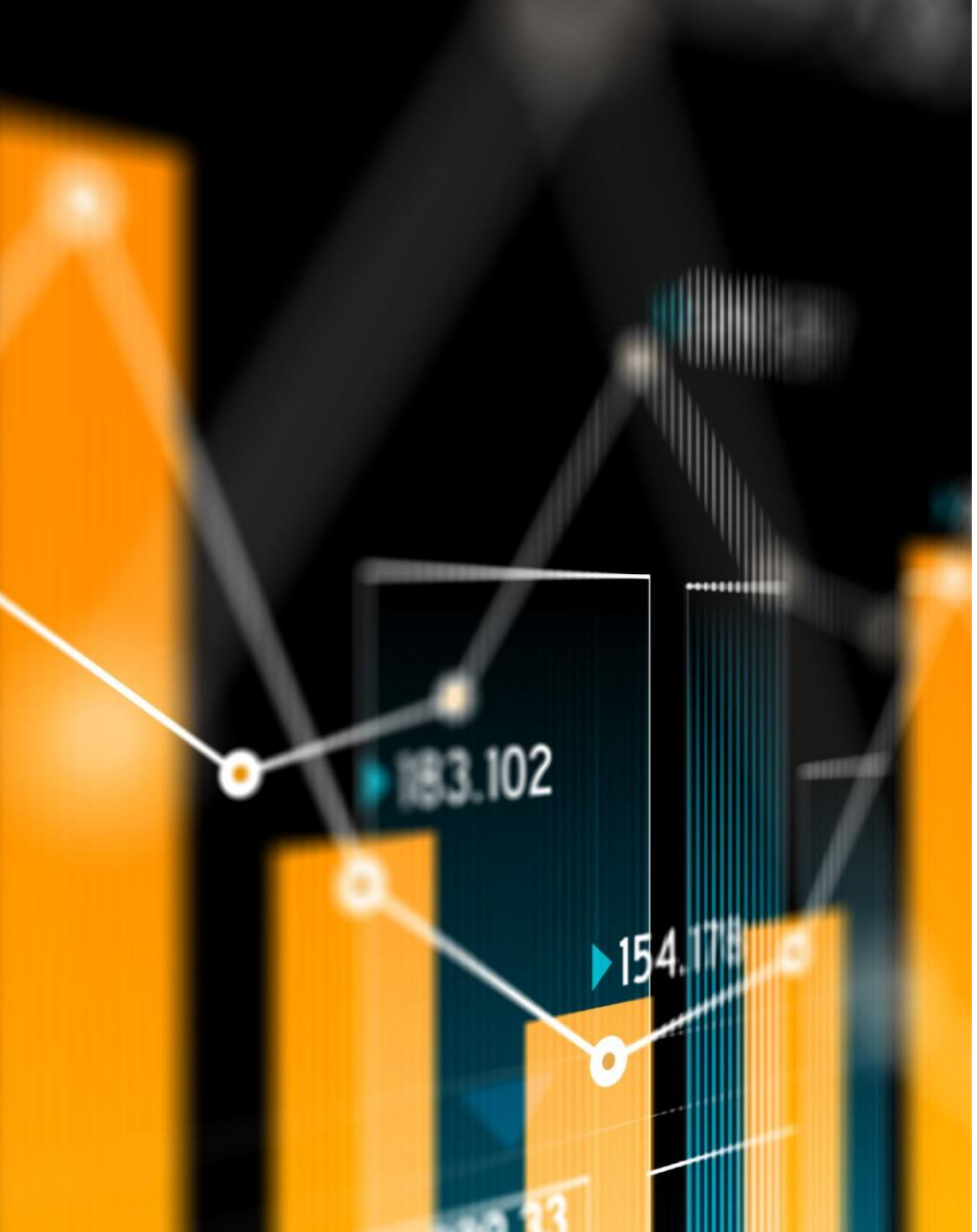
EDA with SQL


I then explored the data using SQL
This involved using SQL to display the names of each Launch Site
I displaced the total payload mass carried by boosters launched by NASA (CRS)
Aftwerwards I displayed average payload mass carried by booster version F9 v1.1
Following that I listed the date when the first successful landing outcome in ground pad was achieved.
Then I listed the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
Next, I listed the total number of successful and failure mission outcomes.
Afterwards, I Listed the names of the booster_versions which have carried the maximum payload mass.
Finally I ranked 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.
You can follow along here.



Build an Interactive Map with Folium

- An interactive map was created to highlight the specific map coordinates the rockets were launched from. The map highlights the successful and failed launches with markers and circles
- We then calculated the distances between launch sites and their proximities
- This helped determine which launch sites had the best chance of success
- To take a look at how this process was performed please click here

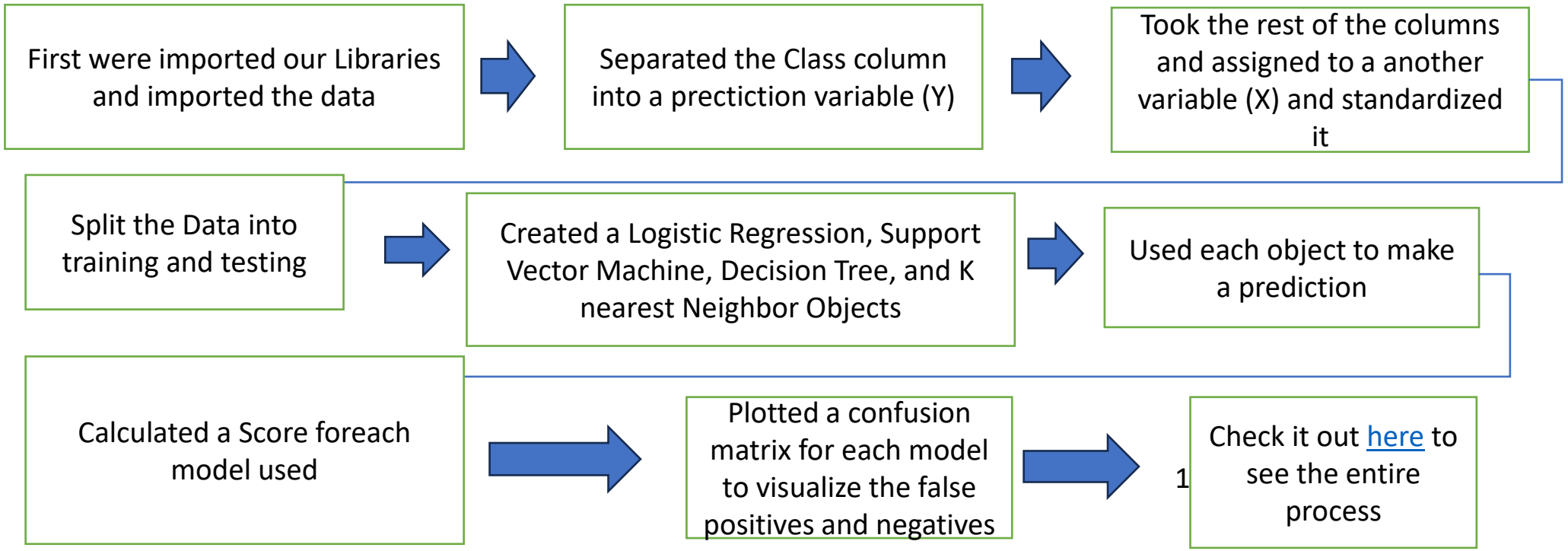


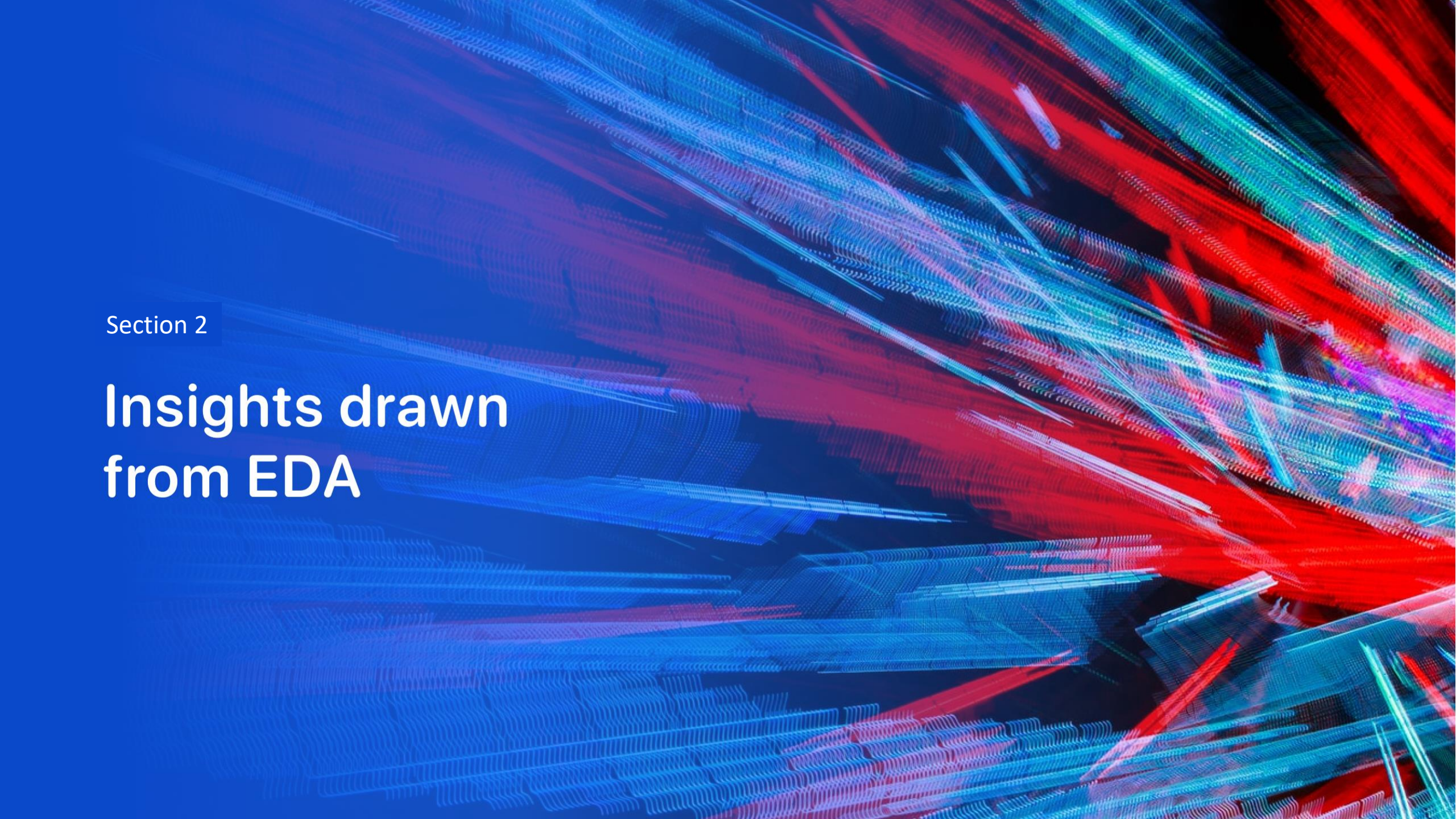
Build a Dashboard with Plotly Dash

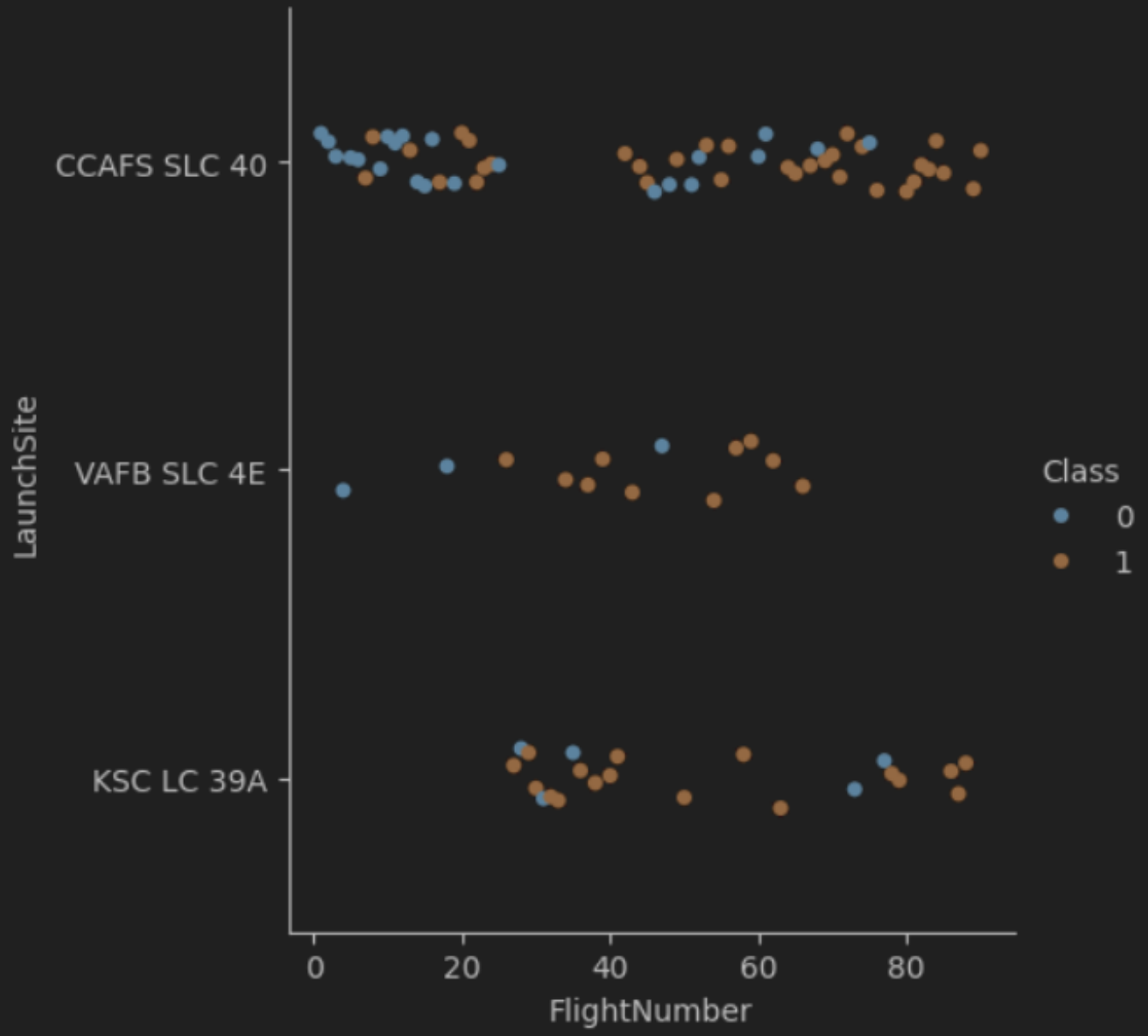
- Here we made two interactive dashboards containing a Pie chart and Scatter Plots using Plotly Dash
- With my Pie chart you can see the success/failure proportion for each Launch Site with a drop-down menu
- With my Scatter plot you can see the correlation of each Launch Site and the Payload Mass(kg) and a hue the shows the successful or failed launch site.
- The Scatter plot is interactive and allows you to select the Payload Mass range preferential analysis
- With these plots you help determine with Launch Site had the highest probability of success and which payload ranges had the highest/lowest success rates
- You can see the code needed to complete the dash app here

Predictive Analysis (Classification)

Finally, python machine learning models (LR, SVM, Decision Trees, and KNN) were used on the data to make predictions for the launches




Section 2
Insights drawn from EDA

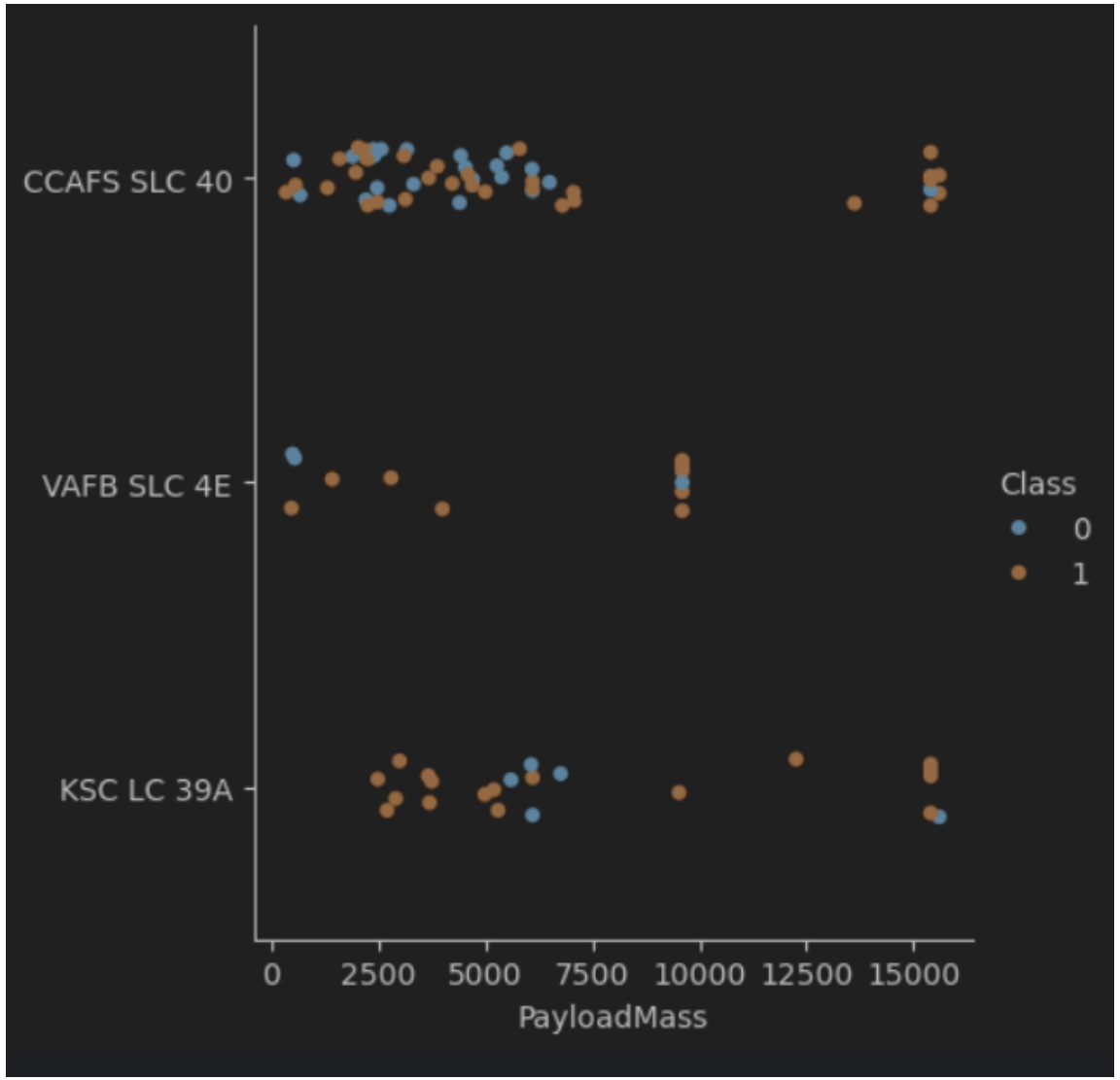


Flight Number vs. Launch Site

- Here you can see that the ‘CCAFS SLC 40’ contained the highest amount of Flight Numbers and the most amount of Successful Launches

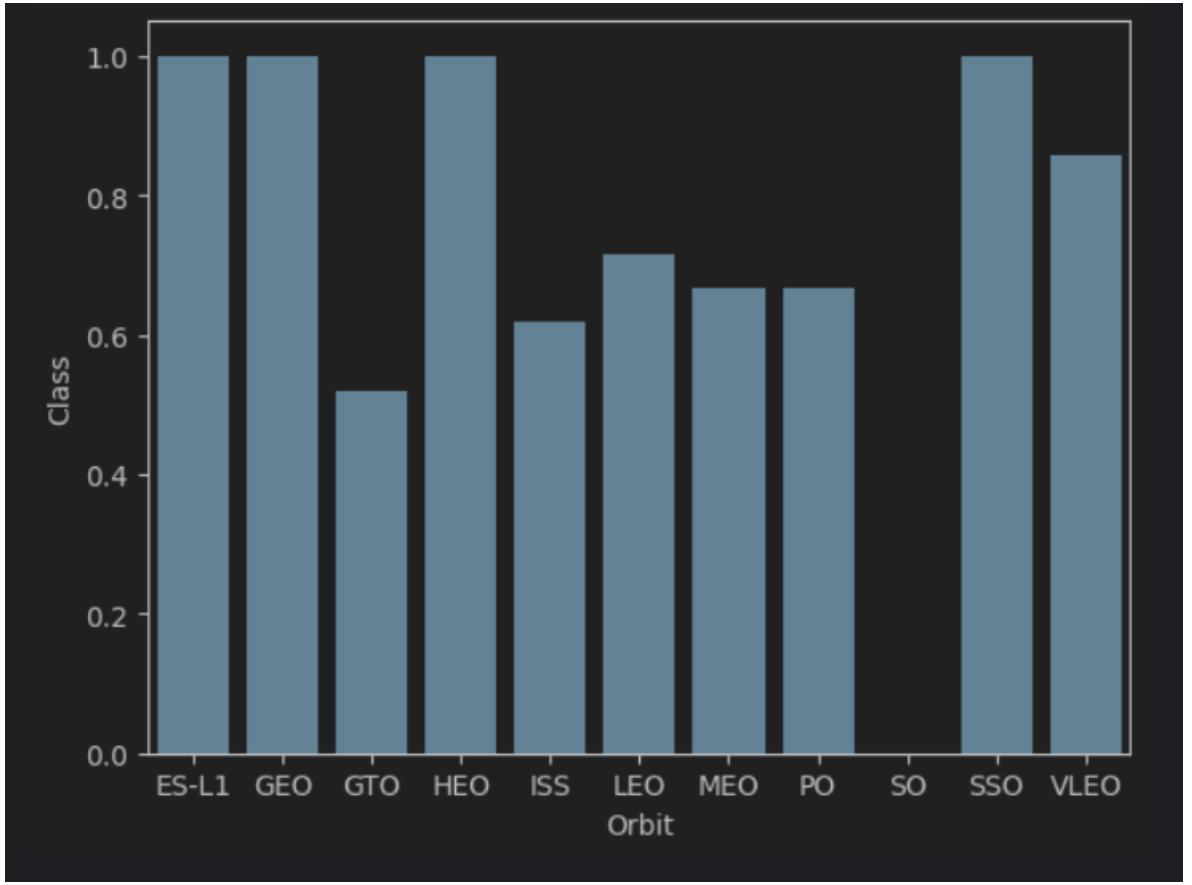
Payload vs. Launch Site

- Here you can see the ‘CCAFS SLC 40 launch site launched with the highest Payload Mass



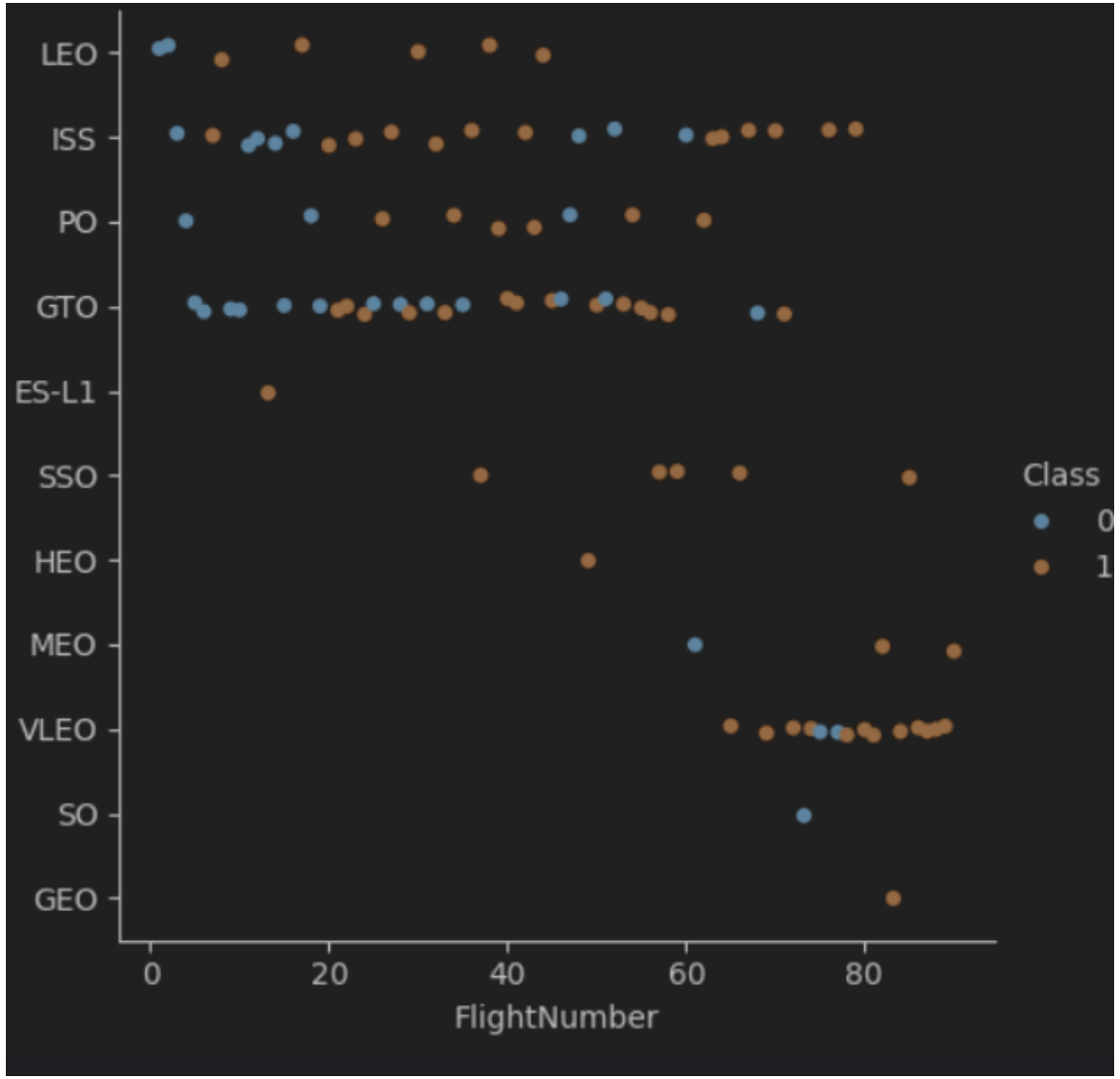
Success Rate vs. Orbit Type

Here you can that the ES-L!, GEO, Heo, and SSO orbit had the highest chance of success



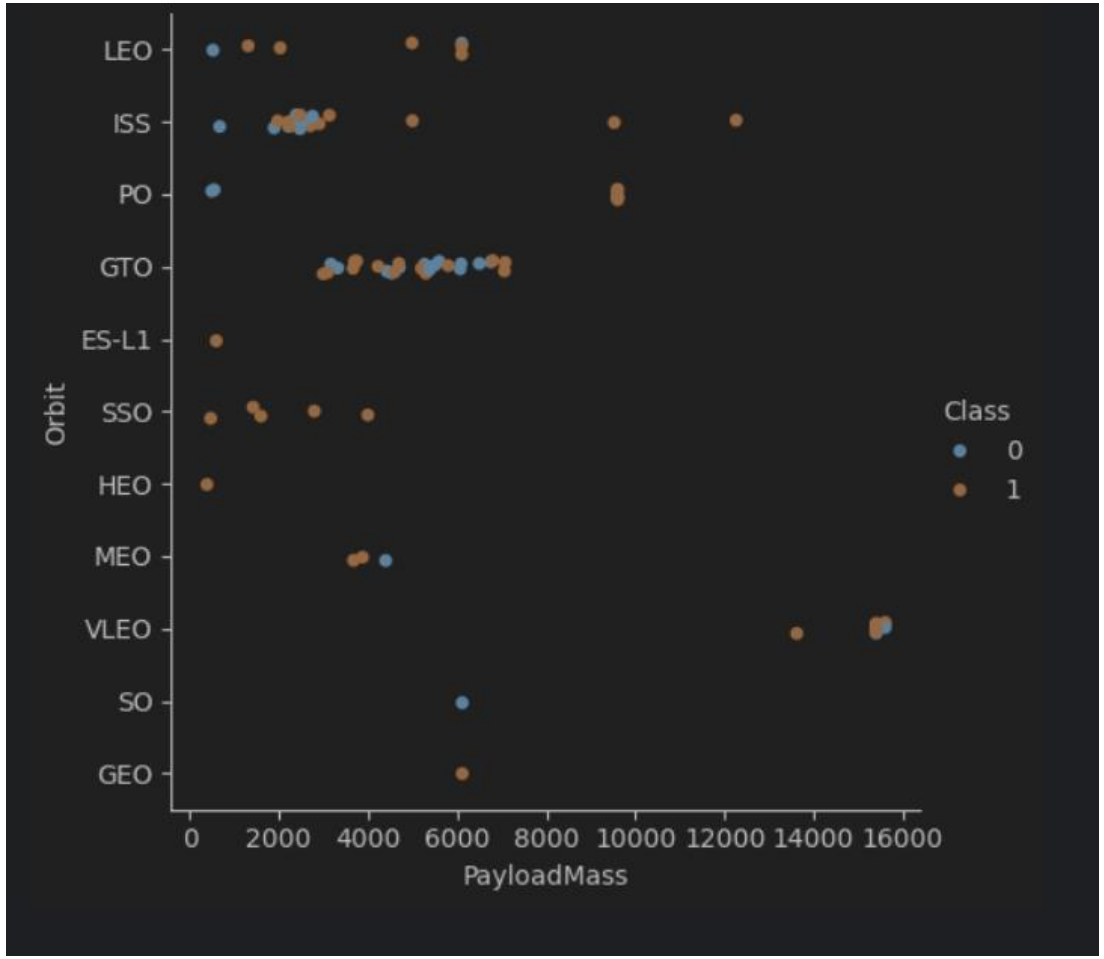
Flight Number vs. Orbit Type

- There seems to be no relationship between Flight number and Orbit Type



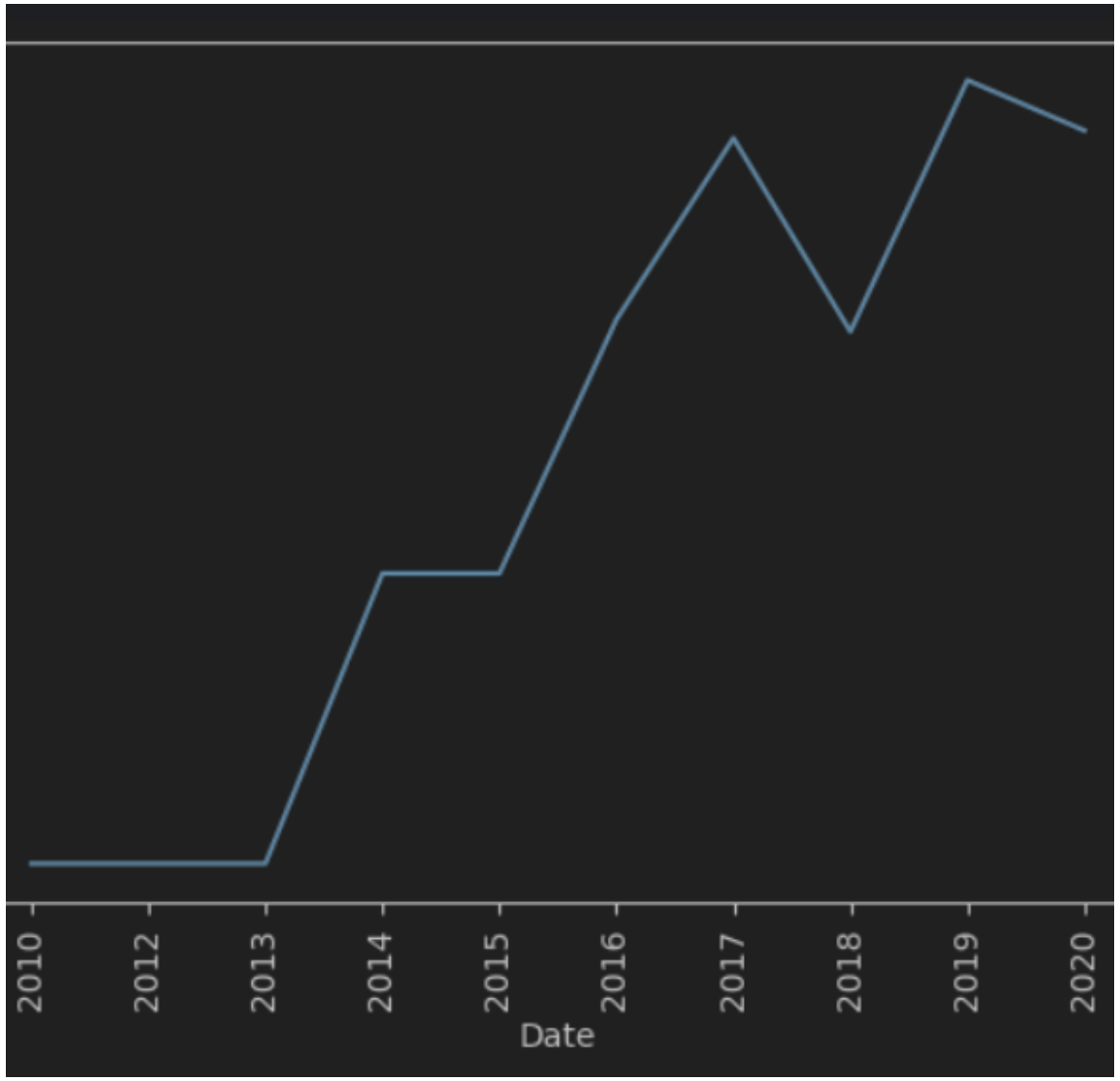
Payload vs. Orbit Type

- You can see here that there doesn't seem to be much relationship between Orbit type and Payload Mass



Launch Success Yearly Trend

- As time progressed you can see that Launches had higher success rates



```
%sql select distinct Launch_Site from SPACEXTABLE

 * sqlite:///my_data1.db
Done.
[('CCAFS LC-40',), ('VAFB SLC-4E',), ('KSC LC-39A',), ('CCAFS SLC-40',)]
```

All Launch Site Names

When Querying the data with SQL, I was able to get a list of all Launch Site Names

Launch Site Names Begin with 'CCA'

```
%sql select * from SPACEXTABLE where Launch_Site like '%CCA%' LIMIT 5

 * sqlite:///my_data1.db
Done.
[('2010-06-04', '18:45:00', 'F9 v1.0  B0003', 'CCAFS LC-40', 'Dragon Spacecraft Qualification Unit', 0, 'LEO', 'SpaceX', 'Success', 'F
 ('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
 ('2012-05-22', '7:44:00', 'F9 v1.0  B0005', 'CCAFS LC-40', 'Dragon demo flight C2', 525, 'LEO (ISS)', 'NASA (COTS)', 'Success', 'No a
 ('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')]
```

I was also able to use SQL to obtain all Launch Site names beginning with CCA

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer like '%NASA (CRS)%'
```

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

```
[(48213,)]
```

Total Payload Mass

Here is a query requesting the Total SUM of Payload Mass in the Data Frame

Average Payload Mass by F9 v1.1

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version like '%F9 v1.1%'

 * sqlite:///my_data1.db
Done.
[(2534.6666666666665,)]
```

This query returned the Average Payload Mass from Booster Version F9 v1.1

```
%sql select min(Date) from SPACEXTABLE where Landing_Outcome like'%Success%'
```

```
 * sqlite:///my_data1.db
Done.
[('2015-12-22',)]
```

First Successful Ground Landing Date

Here how we can use SQL to display the First Successful Ground Landing Date

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select Booster_Version from SPACEXTABLE where Landing_Outcome like'%Success (drone ship)%' and PAYLOAD_MASS__KG_ > 4000 and
 PAYLOAD_MASS__KG_ <6000

  * sqlite:///my_data1.db
 Done.
[('F9 FT B1022',), ('F9 FT B1026',), ('F9 FT  B1021.2',), ('F9 FT  B1031.2',)]
```

This query returns the Successful Drone Ship Landing with Payload between 4000 and 6000

Total Number of Successful and Failure Mission Outcomes

```
%sql select count(*) from SPACEXTABLE where Mission_Outcome like '%Success%' or Mission_Outcome like '%Failure%'

  * sqlite:///my_data1.db
Done.
[(101,)]
```

- Here I used the count(all) function to return the value of Mission Outcomes where there was a Success or Failure

```
%sql select Booster_Version from SPACEXTABLE where (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE) order by PAYLOAD_MASS__KG_ desc
 limit 10
```

```
  * sqlite:///my_data1.db
 Done.
[('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',)]
```

Boosters Carried Maximum Payload

- I used SQL to return the Booster Versions from the SpaceX table where the maximum payload mass existed and listed the values after (up to 10)

2015 Launch Records

This is how you can reurn all failed launch record from 2015

```
%sql select * from SPACEXTABLE where substr(Date,0,5)='2015' and Landing_Outcome like '%Failure (drone ship)%'

 * sqlite:///my_data1.db
Done.
[('2015-01-10', '9:47:00', 'F9 v1.1 B1012', 'CCAFS LC-40', 'SpaceX CRS-5', 2395, 'LEO (ISS)', 'NASA (CRS)', 'Success', 'Failure (drone
 ('2015-04-14', '20:10:00', 'F9 v1.1 B1015', 'CCAFS LC-40', 'SpaceX CRS-6', 1898, 'LEO (ISS)', 'NASA (CRS)', 'Success', 'Failure (dron
```

```
%sql SELECT DISTINCT Landing_Outcome, COUNT(Landing_Outcome) FROM SPACEXTABLE WHERE Landing_Outcome like '%Failure (drone ship)%' or
 Landing_Outcome like '%Success (ground pad)%' and Date >= '2010-06-04' AND Date <= '2017-03-20' GROUP BY Landing_Outcome;
```

```
 * sqlite:///my_data1.db
Done.
[('Failure (drone ship)', 5), ('Success (ground pad)', 3)]
```

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

- This is where I selected the Distinct Landing Outcomes and the count of Landing Outcomes from the spaceX table where the Landing Outcome that had Failure or Success and where the Dates were between 6-4-2010 and 3-20-2017

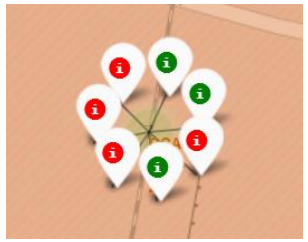
Section 3

Launch Sites Proximities Analysis

Interactive Map (Launch Site)

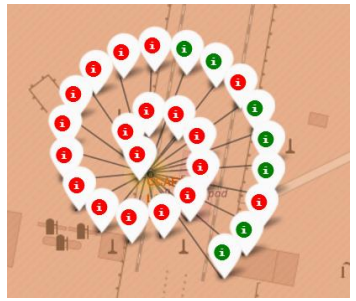
- This map contains the Launch Sites for our rockets (They are clearly shown in Southern California and Central Florida on the East coast.

Which Launch Site Had the most Successful Launch Ratios?

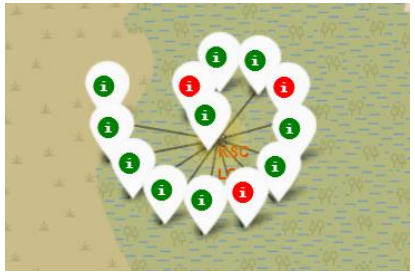


Site CCAFS SFL-40

Site CCAFS LC-40



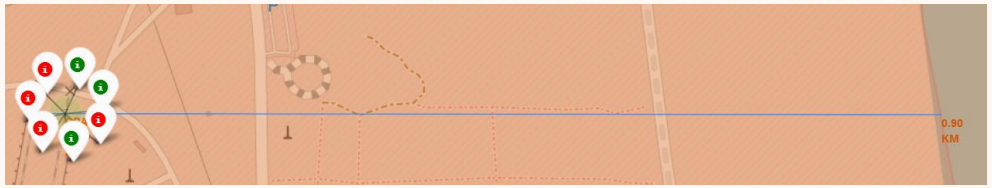
From our Interactive Map you can conclude that Site CCAFF LC-40 had the most Launches, but Site KSC LC-39A had the best Success Rate.



Site KSC LC-39A

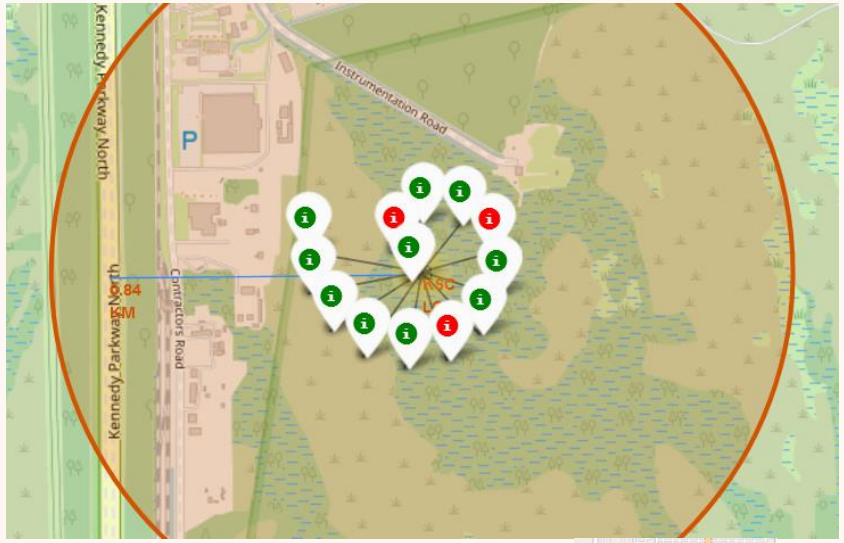
Site VAFB SLC-4e

Are Launch Sites close to Cities, Coastlines, Railroads, or Highways

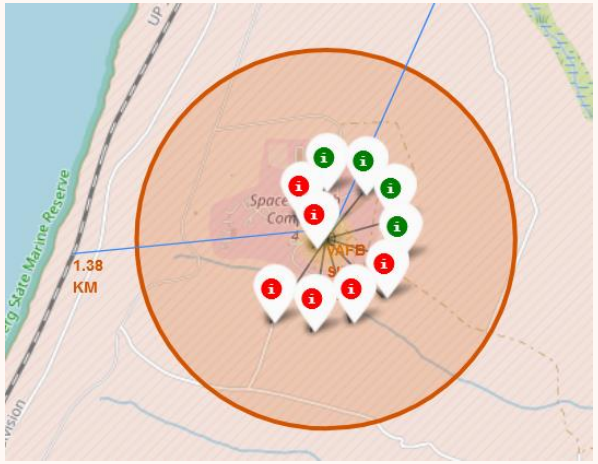


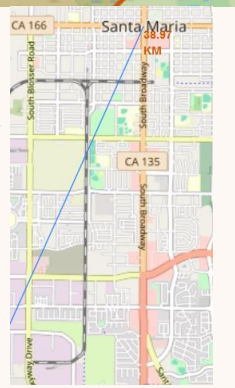
Launch Site CCAFS SLC -40 is .90 km from the Atlantic coastline

Launch Site KSC LC-37A is .84 KM from the nearest Highway

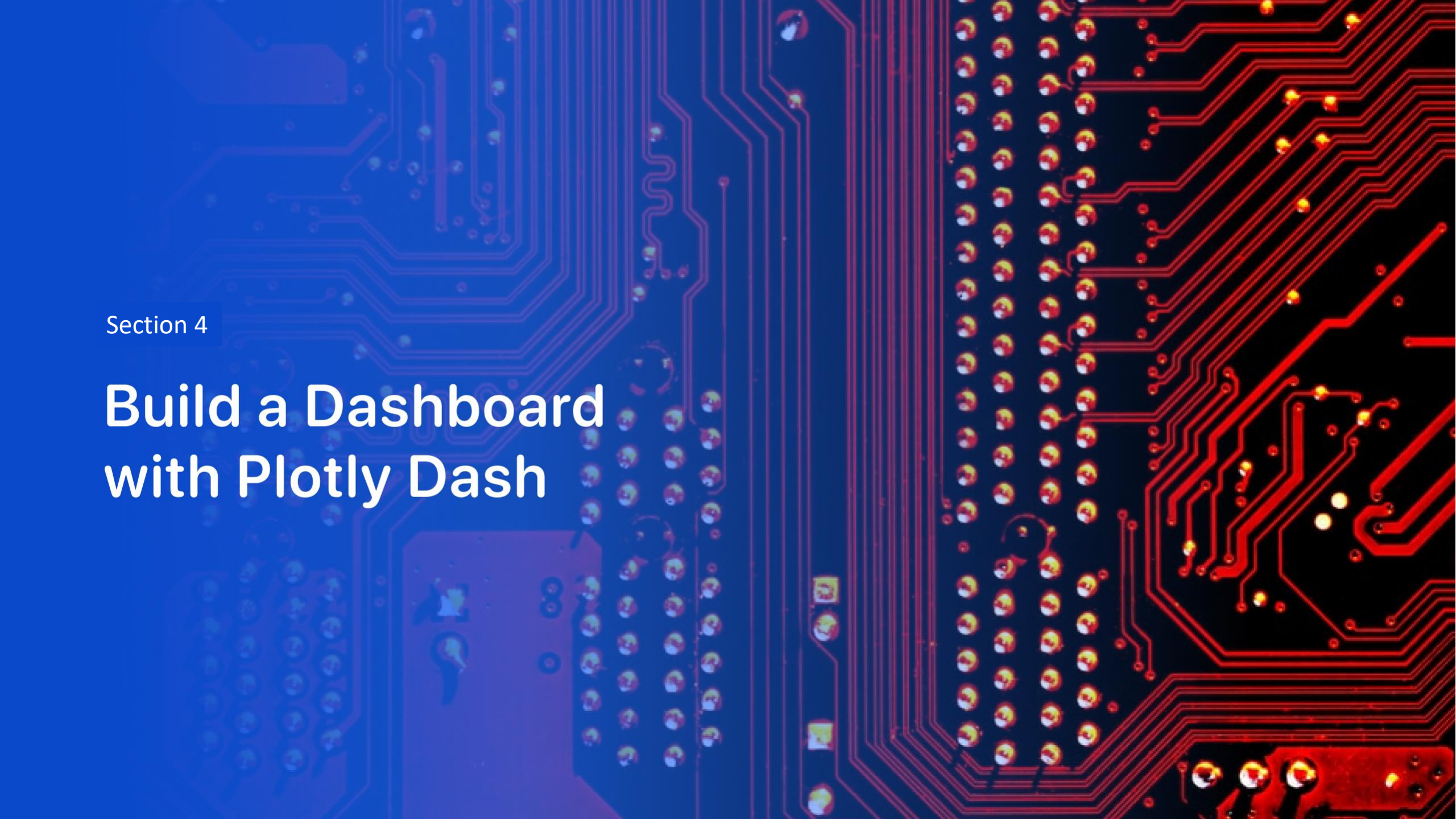


Site VAFB SLC-4e is 1.38 from the nearest railway and 38 Miles from Downtown Santa Maria

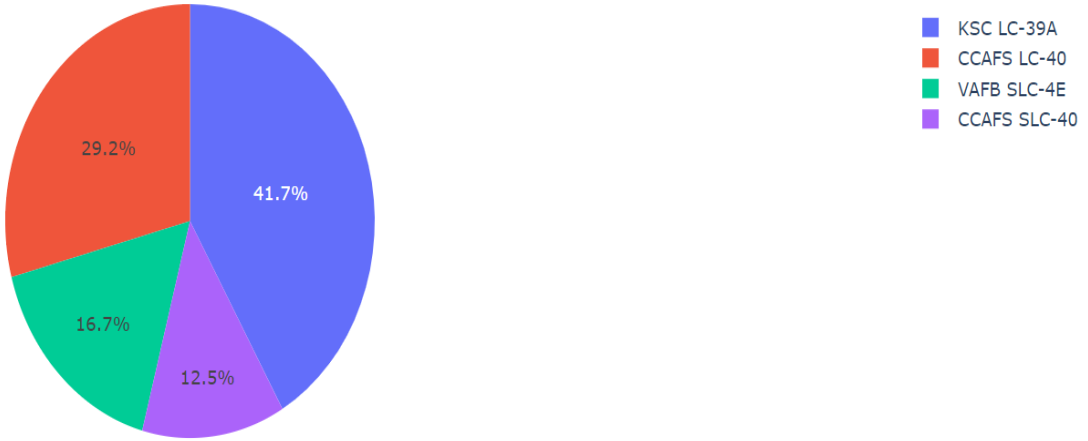




Launch sites are relatively close to Coastlines, Railroads, and Highways, but keep a clear distance from Cities.

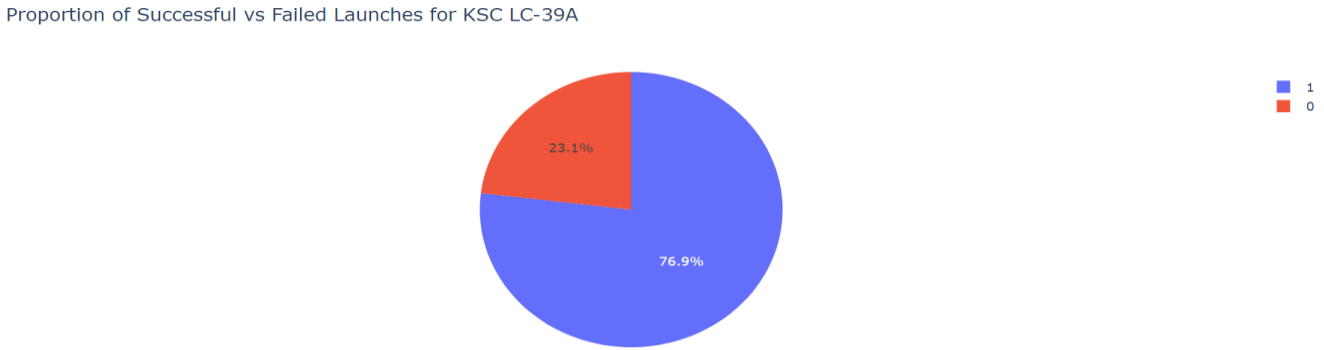

Section 4
Build a Dashboard with Plotly Dash

Which site have had the most Launches



From this Pie chart you could determine that site KSC LC-39A have completed the most amount of launches

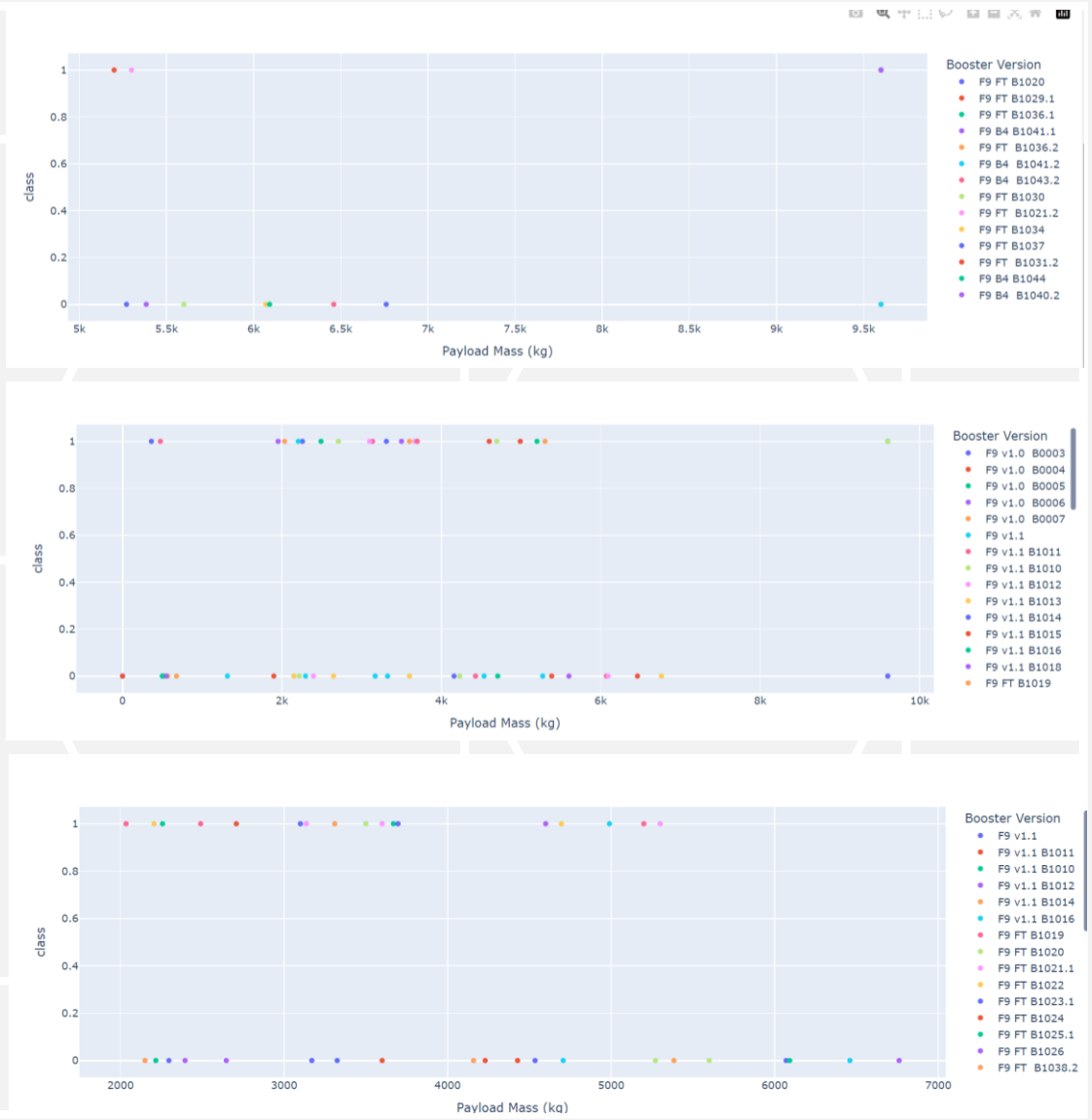
Which Launch Site had the most successful Rocket Launches?



As you can see. Site KSC LC-39A had the greatest proportion of Successful Rocket Launches.

Does booster type matter? Which payload size work best?

- Its looks like most successful missions contained payloads that were less than 6000 kg, but more than 1000 kg. It does not seem that booster variation does not matter.

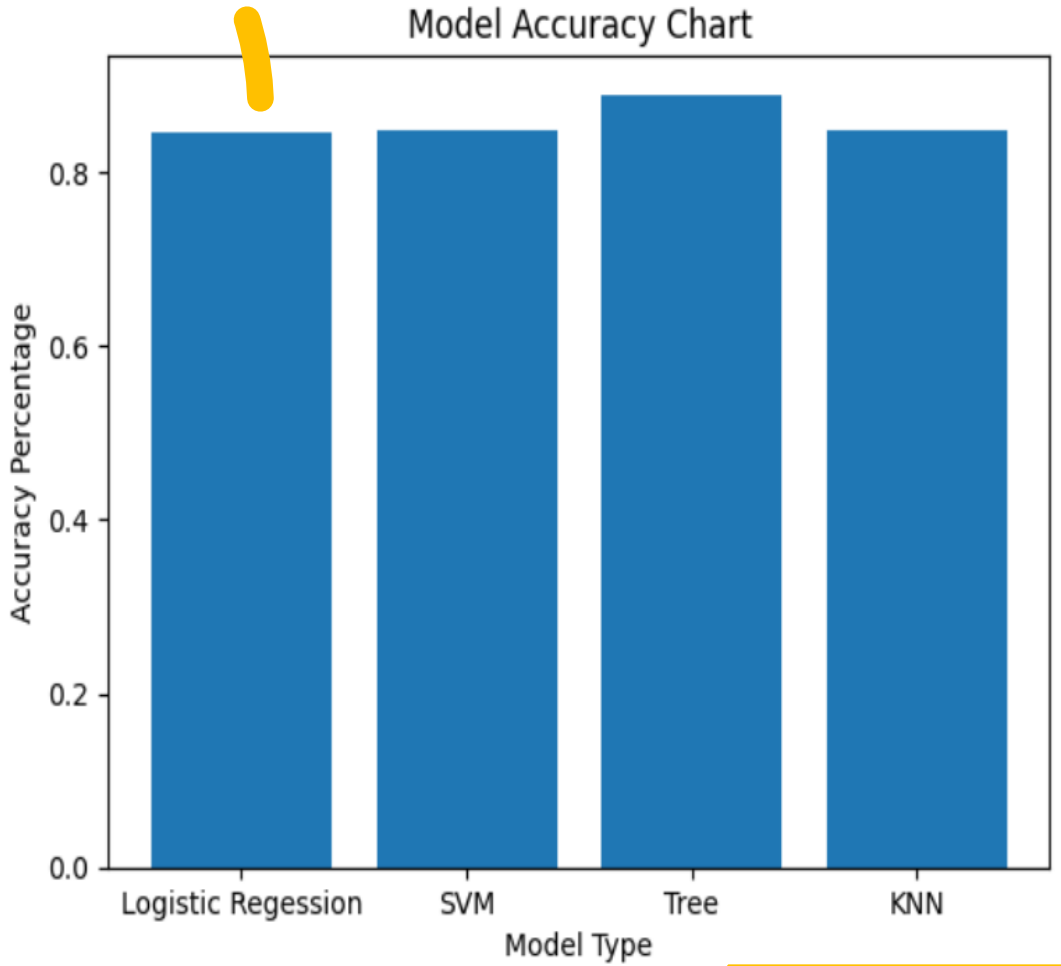


Section 5

Predictive Analysis (Classification)

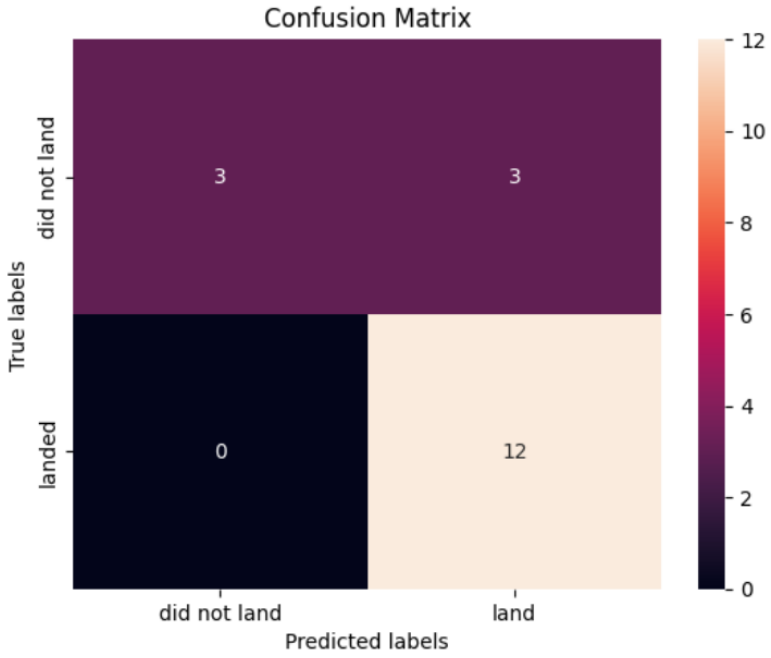
Classification Accuracy

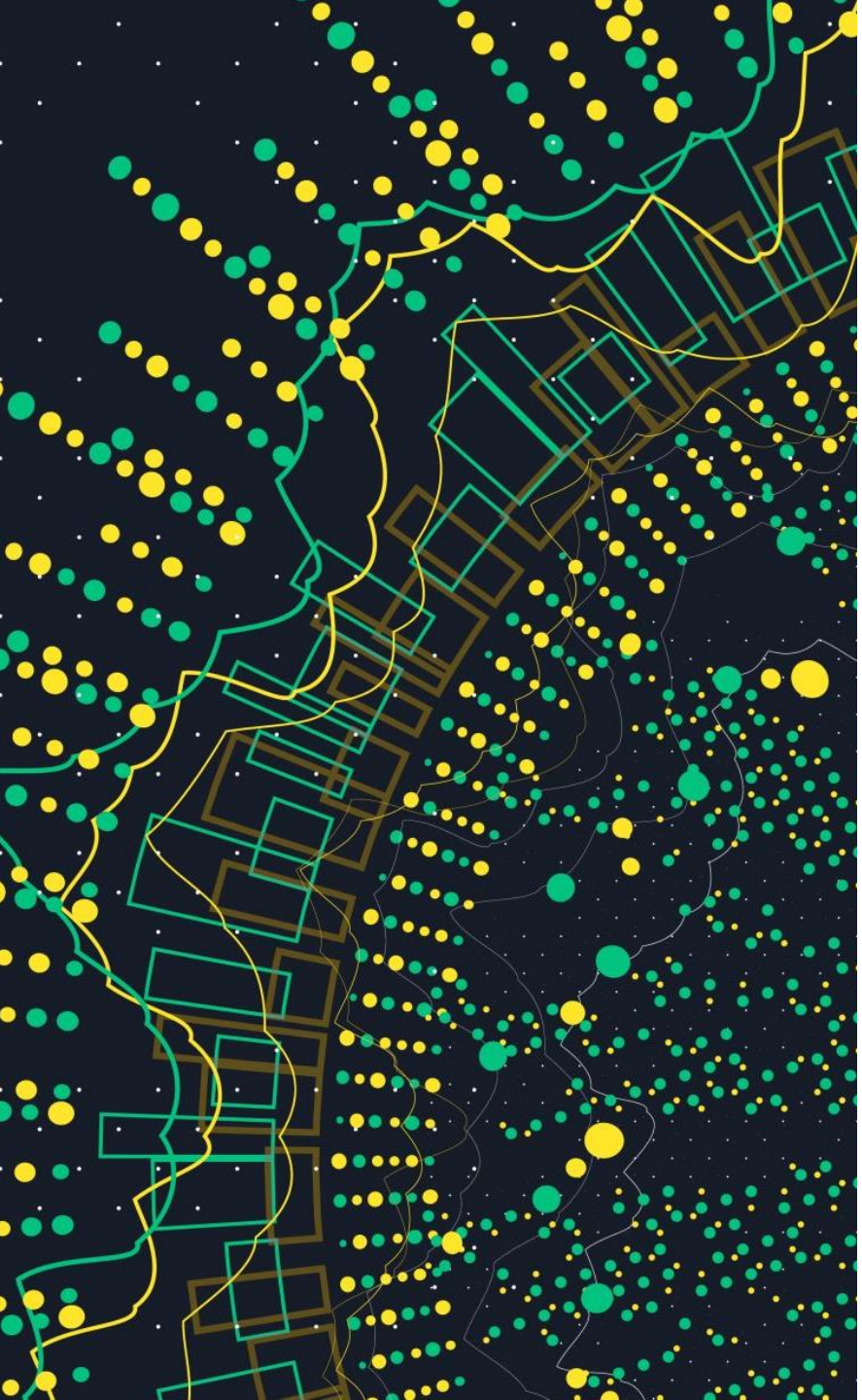
While trying to make a prediction with the SpaceX Data the Decision Tree Model showed the best Accuracy with .89, but the Confusion Matrix showed 6 False positives.



Confusion Matrix

- When using Logistic Regression, K Nearest Neighbor, and Support Vector Machine models I got a high accuracy score, but I perpetually got 3 False positives with each Score.
- I believe any of those models work best with the SpaceX data set.





Conclusions

- ‘CCAFS SLC 40’ contained the highest amount of Flight Numbers and the greatest number of Successful Launches but Launch Site KSC LC-39A had the greatest proportion of Successful Rocket Launches.
- CCAFS SLC 40 launch site had to the greatest success with the highest Payload Mass
- ES-L!, GEO, Heo, and SSO orbit types had the highest success rate
- Besides a small dip in 2018, Success rate for all Launches Steadily increased
- Launch Sites seems to be purposely built away from cities and are install closest to coastlines
- Decision Tree displays the best accuracy with this data set, but poor confusion matrix
- The best Models to use are Logistic Regression, Support Vector Machine, and K Nearest Neighbor.
- This has been a lot of fun, thank you.