IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Lewis Chukwuma

11th October, 2024

Outline

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

Executive Summary

- SpaceX advertises falcon 9 launches on its website having a cost of 62 million dollars whereas other providers cost upward of 165 million dollars each for the same thing. What makes SpaceX stand out? Its due to the fact that they reuse the first stage by making sure the first stage lands successfully saving them millions of dollars. This information from SpaceX REST API can also be used to reduce costs for SpaceY and improve competition.
- Research has been carried out, models have been trained and tested and results show that there is an 83% chance that the first stage will land successfully using Logistic Regression, SVM and KNN, but all these rely heavily on criterias such as booster version, payload mass, orbit, launch site etc.

Introduction

Commercial space age is here and companies are making space travel affordable for everyone. The front runner of it all is SpaceX and some of their accomplishments include sending spacecraft to the international space station, Starlink which is a satellite internet constellation providing satellite internet access and sending manned missions to space. The reason why SpaceX is able to do this is because their rocket launches are relatively inexpensive. They advertise falcon 9 launches at a cost of 62million dollars while other providers cost an upward of 165 million dollars and it is due to the fact that unlike other providers, SpaceX's falcon 9 can recover its first stage, thereby saving a lot of money. Here we would like to understand the factors that can guarantee a successful first stage landing.

Some questions we would like to find answers to include, how the launch site affects the landing, how the payload mass affects the landing, how the orbit type affects the landing, how the booster affects the landing, how these factors affect each other and also predicting the outcome of a launch by modelling all these factors from historic data.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

Like every other leading company in the world, SpaceX store their data and this can be gotten by requesting rocket launch data from SpaceX API using a GET request with the following URL; "SpaceX data link"

- Perform data wrangling

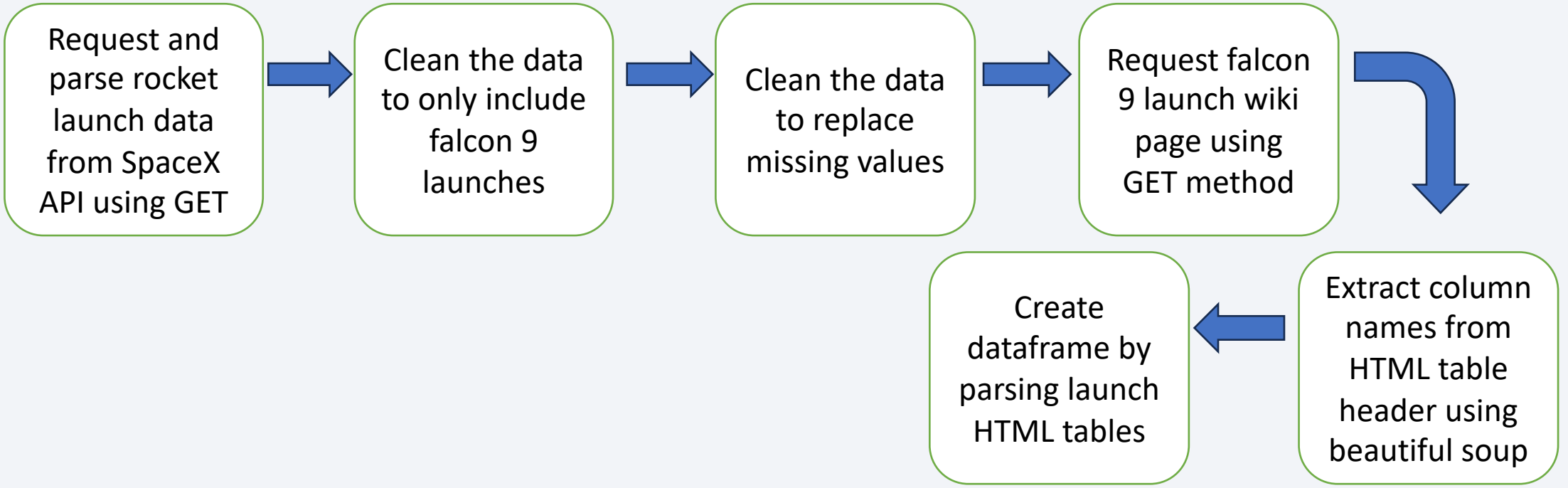
Next we have to clean the data such as removing unnecessary data and removing missing values in the data set and then extract falcon 9 launch records with Beautiful Soup

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Using models such as Logistic Regression, SVM, Decision Tree Classifier and KNN to train and test the data to see if it will be able to predict data outside the training set.

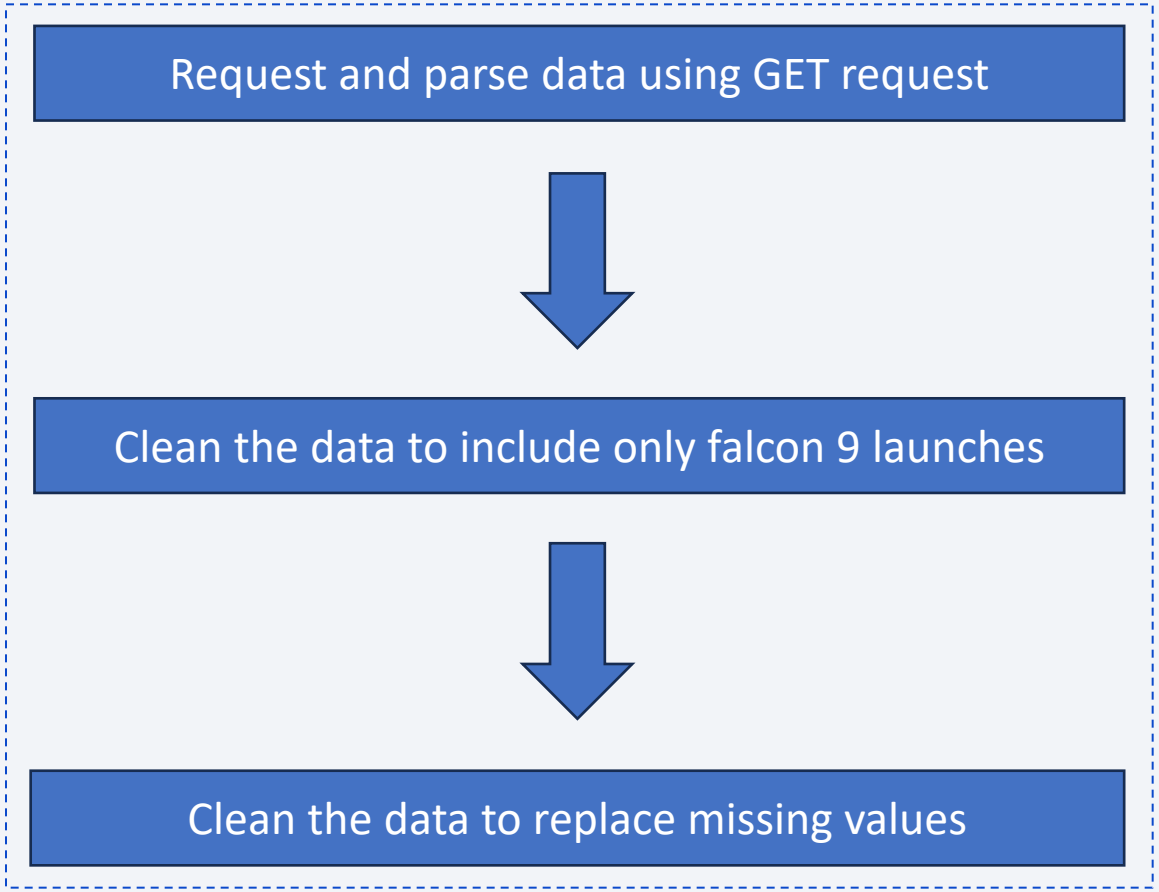
Data Collection

Like every other leading company in the world, SpaceX store their data and this can be gotten by requesting rocket launch data from SpaceX API using a GET request with the following URL; "SpaceX data Link"



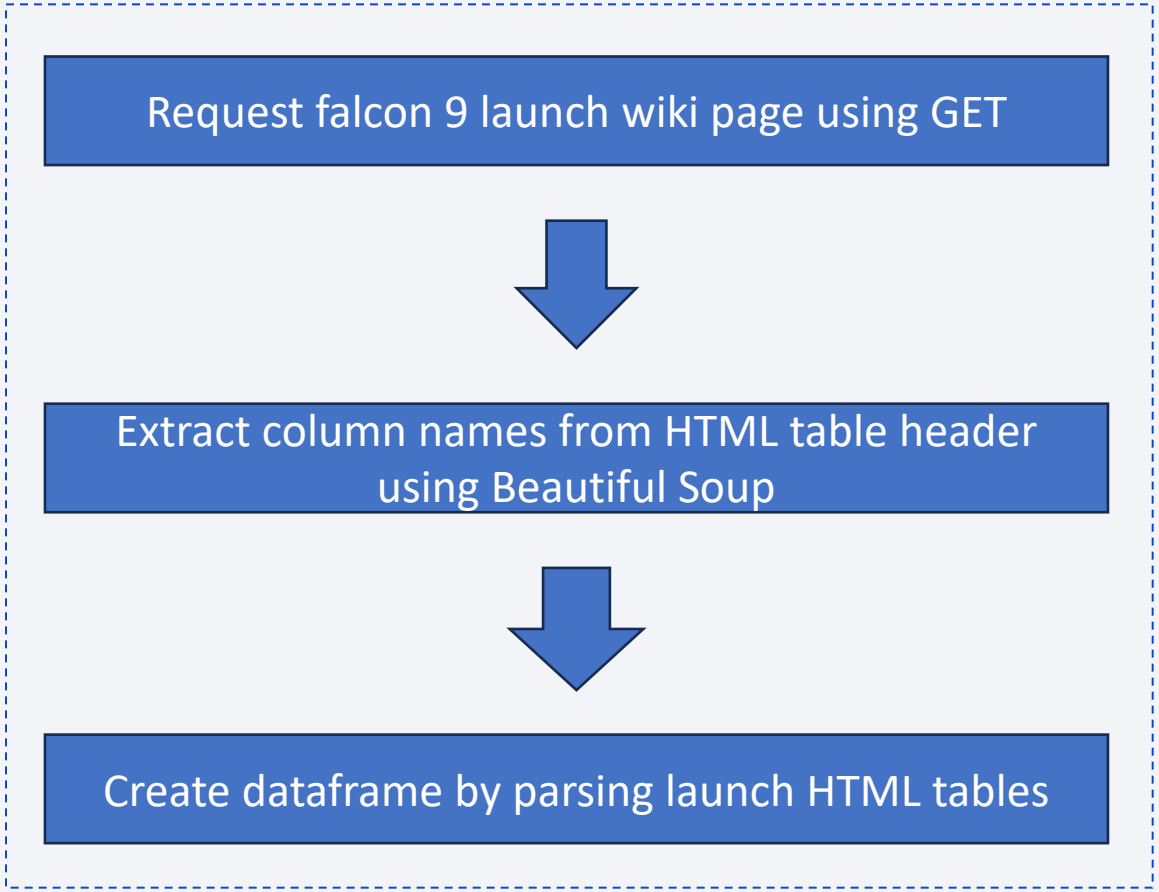
Data Collection – SpaceX API

- Lewis Chukwuma Data Collection-API link



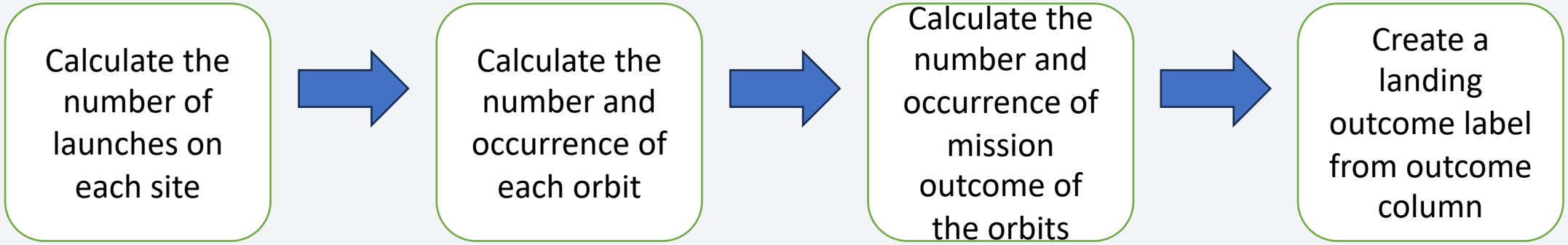
Data Collection - Scraping

- Lewis Chukwuma Data Collection - Scraping link



Data Wrangling

Performing Exploratory data analysis, it is important to find some patterns in the data to determine the label for training supervised models. There are several different cases where the booster did not land successfully due to an accident, so we transformed those training labels with 1 which means the booster successfully landed and 0 which means it was unsuccessful



Lewis Chukwuma Data Wrangling link

EDA with Data Visualization

- Plotting the relationship between parameters such as FlightNumber and LaunchSite, Payload Mass and Launch Site, FlightNumber and Orbit type, Payload Mass and Orbit type because we want to observe if there is any relationship between any of them and then the success rate of each orbit type and also visualize the launch success yearly trend to visually check if there are any relationships between success rate and orbit type and to also observe the yearly success trend.

Lewis Chukwuma EDA with Visualization link

EDA with SQL

- Displayed the names of unique launch sites in the space mission
- Displayed 5 records where launch sites begin with the string “CCA”
- Displayed the total payload mass carried by boosters launched by NASA (CRS)
- Displayed average payload mass carried by booster version F9 v1.1
- Listed the date when the first successful landing outcome in ground pad was achieved
- Listed the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listed the total number of successful and failure mission outcomes
- Listed the names of booster versions which have carried the maximum payload mass
- Listed the records that will display the month names, failure landing outcomes in drone ship, booster versions, launch site for the month in year 2015
- Ranked the count of landing outcomes between the dates 2010-06-04 and 2017-03-20 in desc order
- Lewis Chukwuma EDA with SQL link

Build an Interactive Map with Folium

Firstly we marked all launch sites on the map using the sites longitude and latitude coordinates then we visualized it on the map by pinning them but to make it easier to pin point on the map, we added circles and markers in places such as CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E launch site coordinates. Next we enhanced the map by adding the launch outcomes for each site and created markers for class = 1 to represent green and class = 0 to represent red. Lastly we analyzed and explored the proximities of launch sites using polyline and also used mouseposition to find coordinates on the map

Lewis Chukwuma Interactive Visual Analytics with Folium link

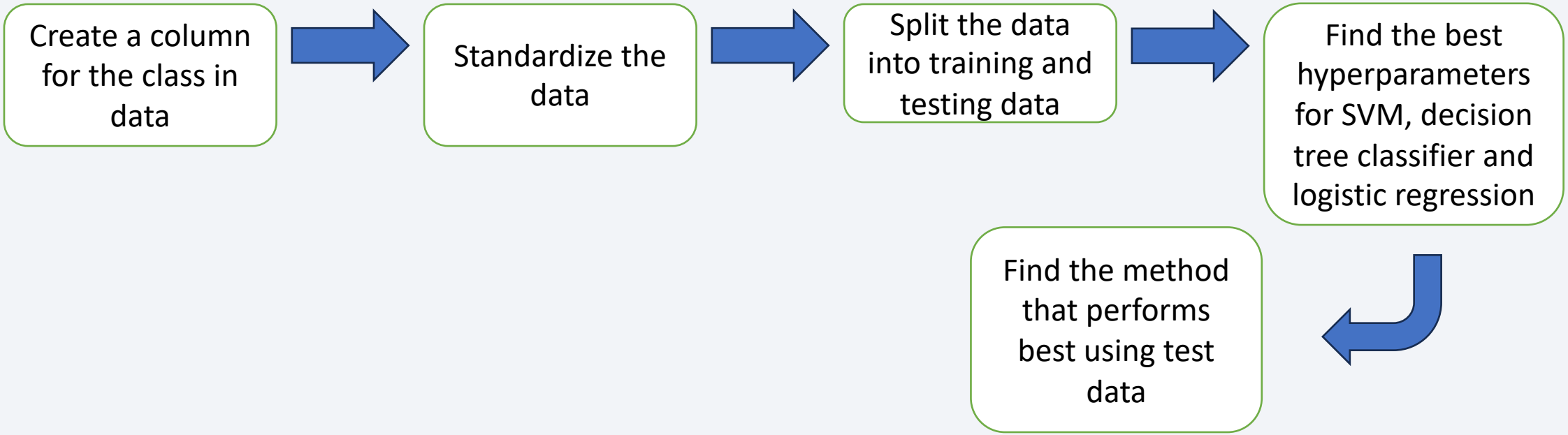
Build a Dashboard with Plotly Dash

I created a pie chart and scatter plot to interact with the data, whereas the piechart showed the success count for all launch sites and also individual launch sites. The dashboard also contains a drop-down list and a range slider so as to be able to select and view each launch sites and also select a payload range respectively. All these was done so as to answer questions such as which site has the largest successful launches, which site has the highest launch success rate, which payload range has the highest launch success rate, which payload range has the lowest launch success rate and which F9 booster version has the highest launch success rate.

Lewis Chukwuma Interactive Dashboard with Plotly Dash link

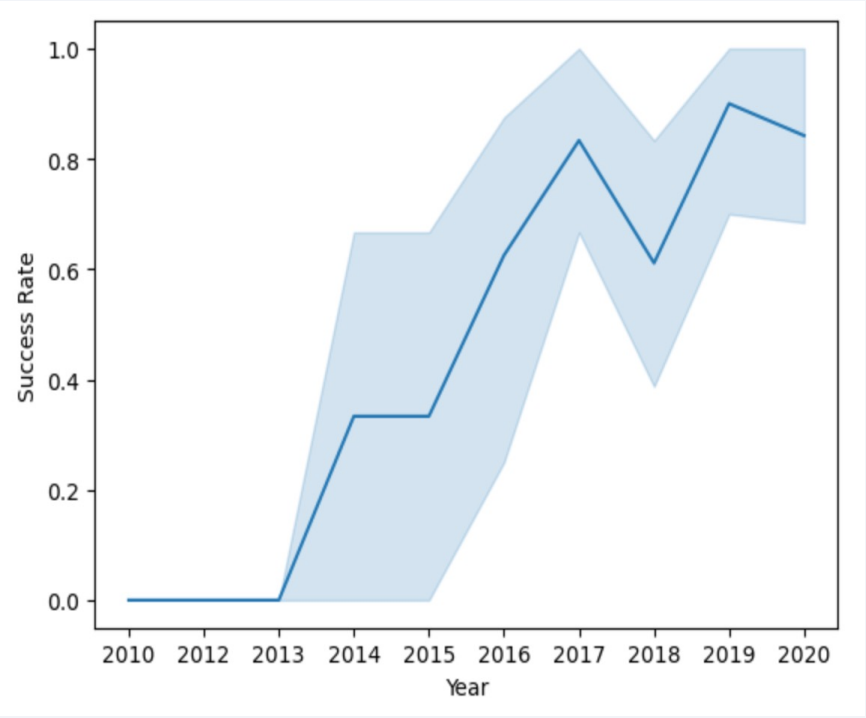
Predictive Analysis (Classification)

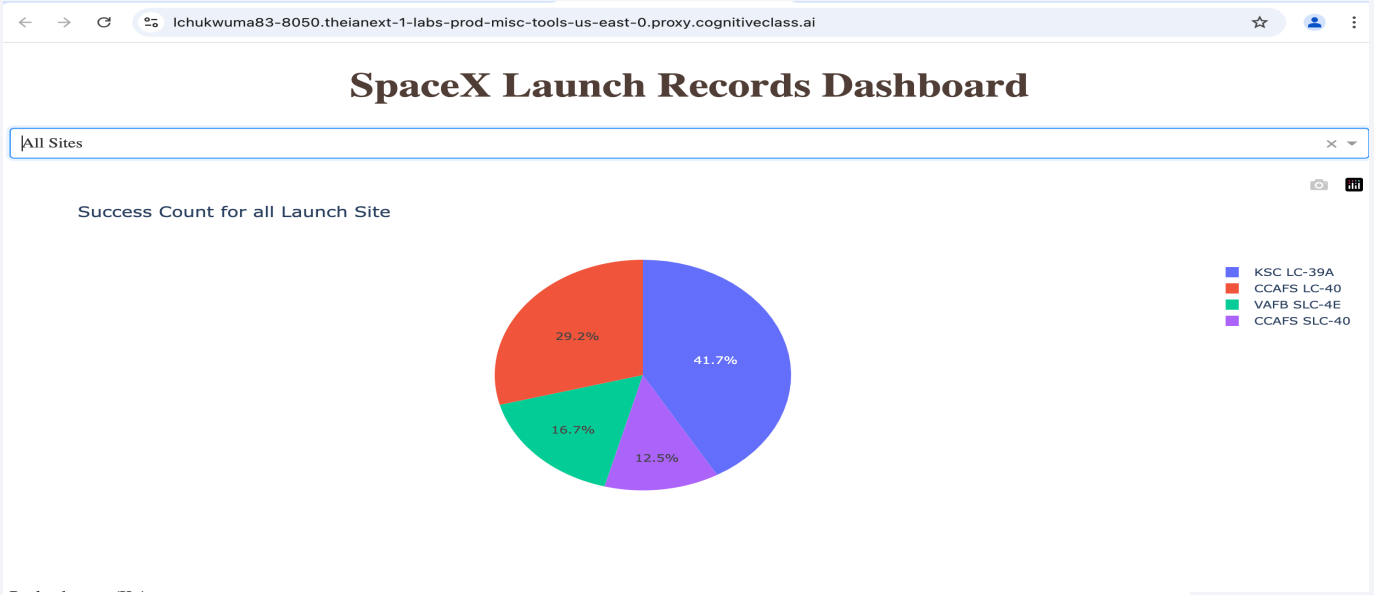
First you have to create a Numpy array from the column class in data by applying the method to_numpy() and assign it to a variable, then standardize the data and split it into training and testing data, then create a logistic regression object, support vector machine object, decision tree classifier object and K-nearest neighbor object which will be used to train the data and after each model, testing is done by calculating the accuracy on the test data. Finally the best model is selected by selecting the model closest to 100%.

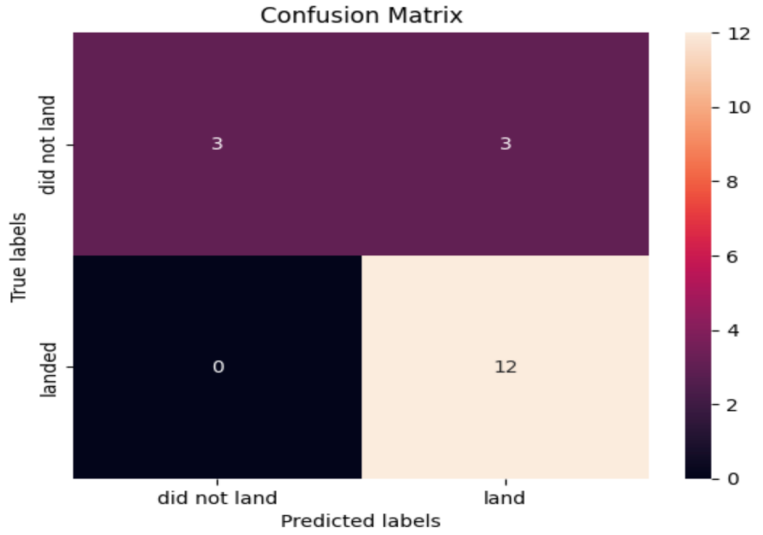


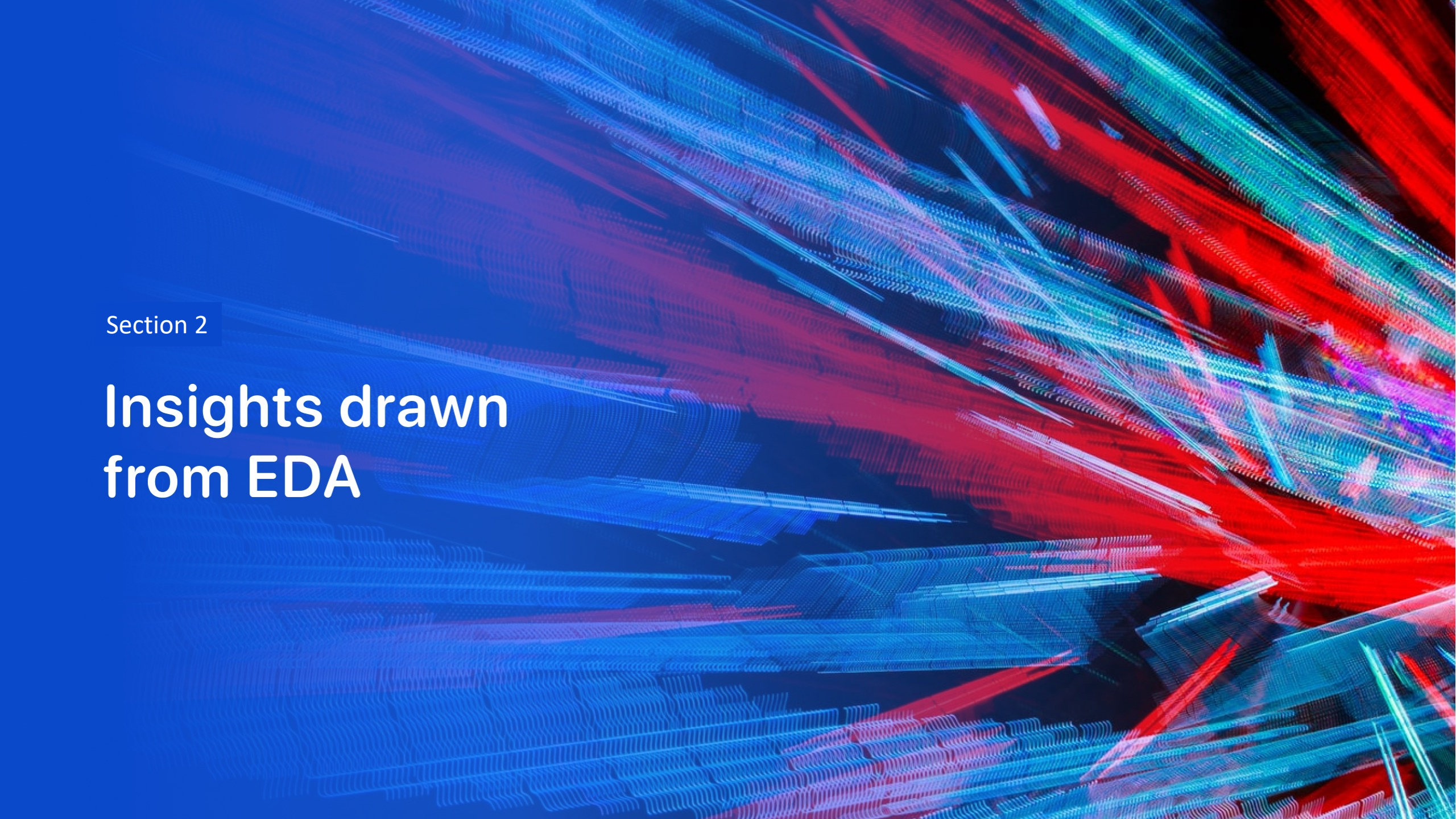
Lewis Chukwuma Machine Learning Prediction link

Results


1.0
0.8
0.6
0.4
0.2
0.0
Success Rate
2010
2012
2013
2014
2015
2016
2017
2018
2019
2020
Year

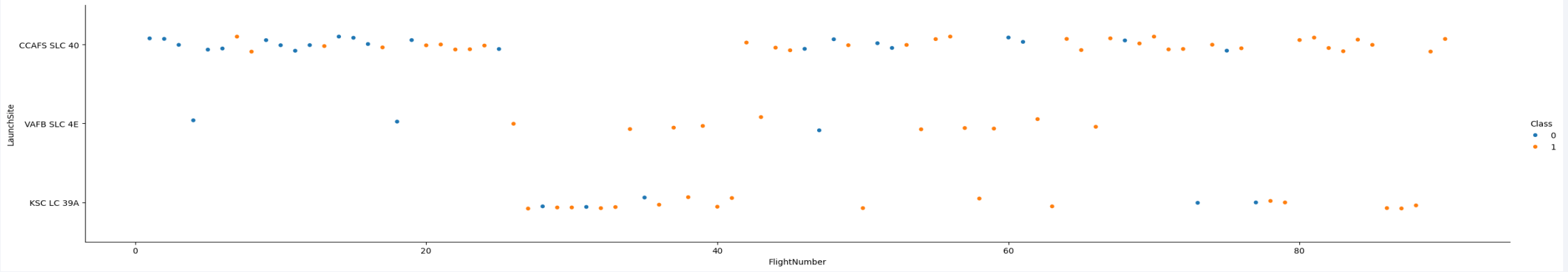

lchukwuma83-8050.theianext-1-labs-prod-misc-tools-us-east-0.proxy.cognitiveclass.ai
SpaceX Launch Records Dashboard
All Sites
Success Count for all Launch Site
29.2%
41.7%
16.7%
12.5%
KSC LC-39A
CCAFS LC-40
VAFB SLC-4E
CCAFS SLC-40


Confusion Matrix
True labels
did not land
landed
3
3
0
12
did not land
land
Predicted labels
12
10
8
6
4
2
0

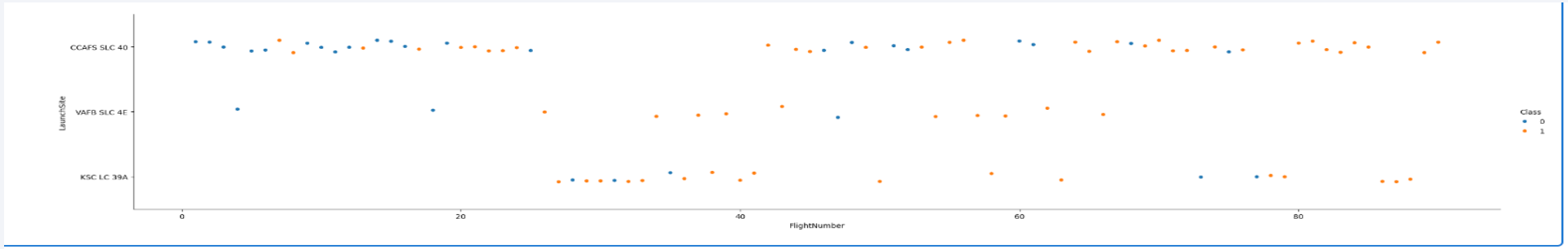

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

Flight Number vs. Launch Site



Scatter plot with explanations

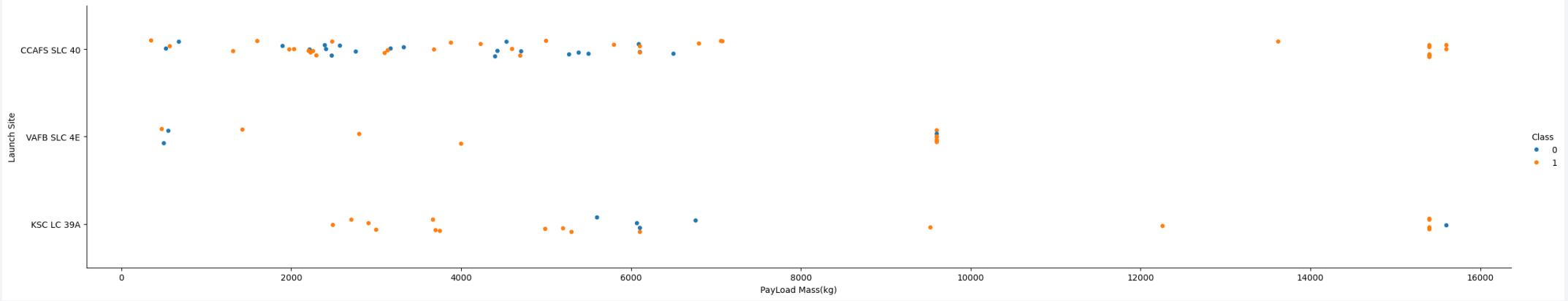


Now try to explain the patterns you found in the Flight Number vs. Launch Site scatter point plots.

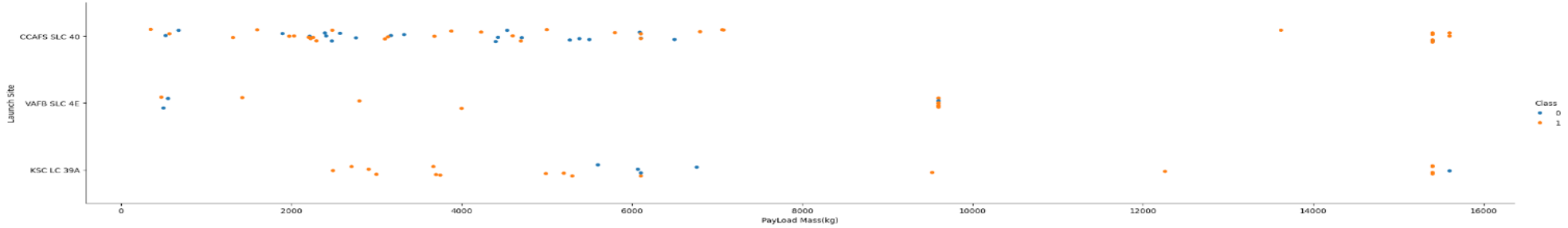
Here, there are no flights below 20 for KSC LC 39A Launch Site, also the same for VAFB SLC 4E above 65 flights and between 24 to 40 in CCAFS SLC 40

Payload vs. Launch Site

Payload vs. Launch Site



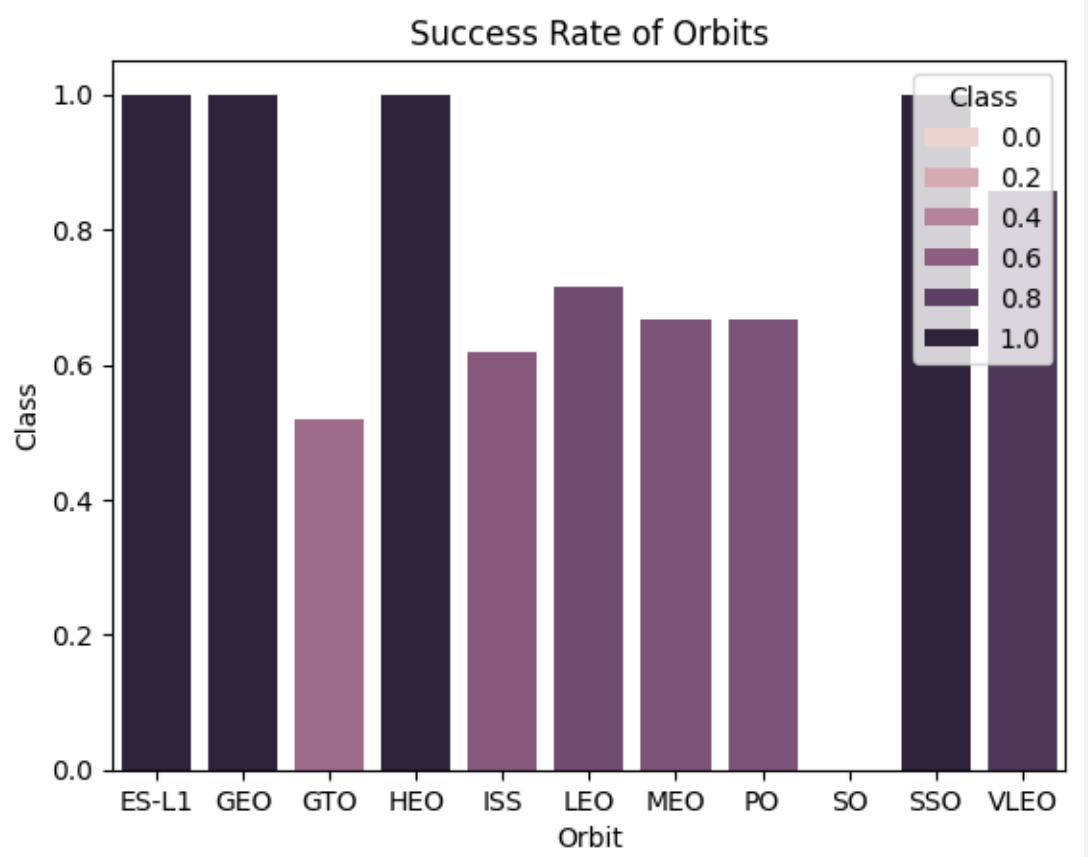
Scatter plot with explanations



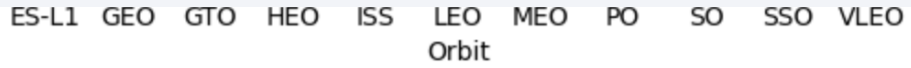
Now if you observe Payload Mass Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type

Success rate of each orbit type



Scatter plot with explanations

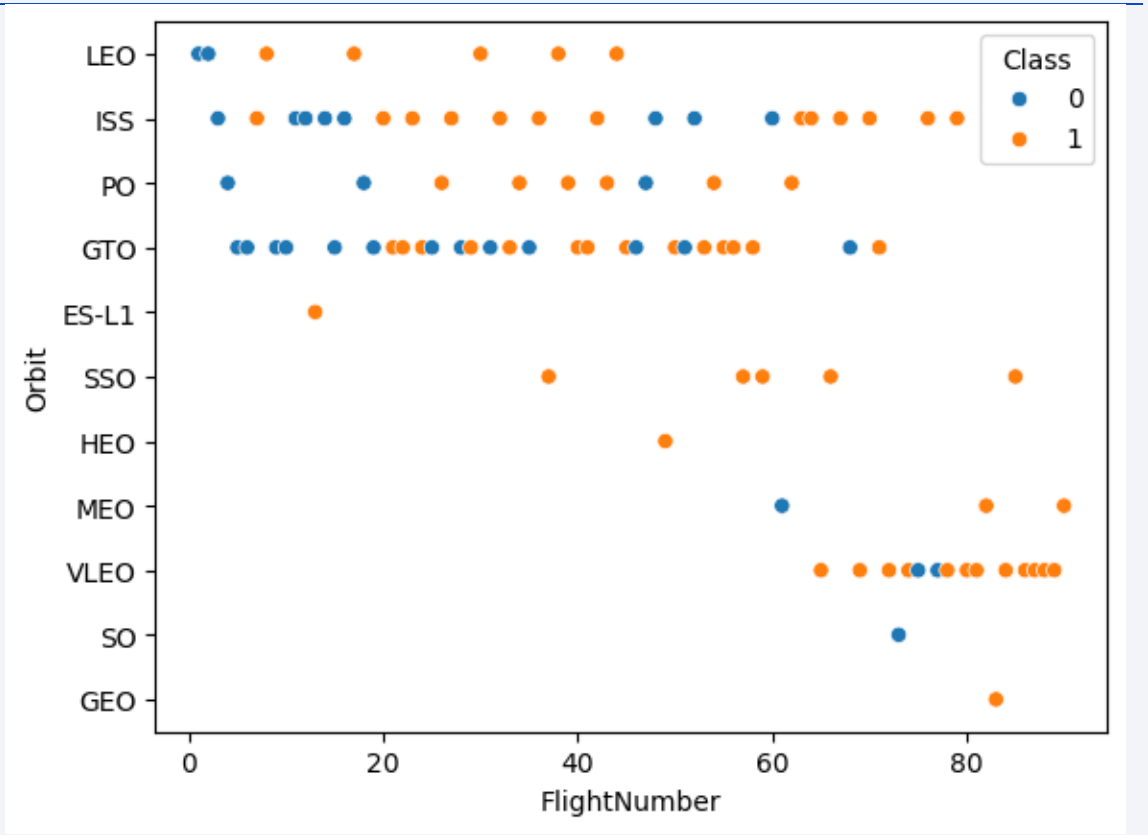


Analyze the plotted bar chart to identify which orbits have the highest success rates.

From the chart above, orbits with the highest success rates are ES-L1, GEO, HEO and SSO, while orbits with the lowest success rate is GTO

Flight Number vs. Orbit Type

Flight number vs. Orbit type



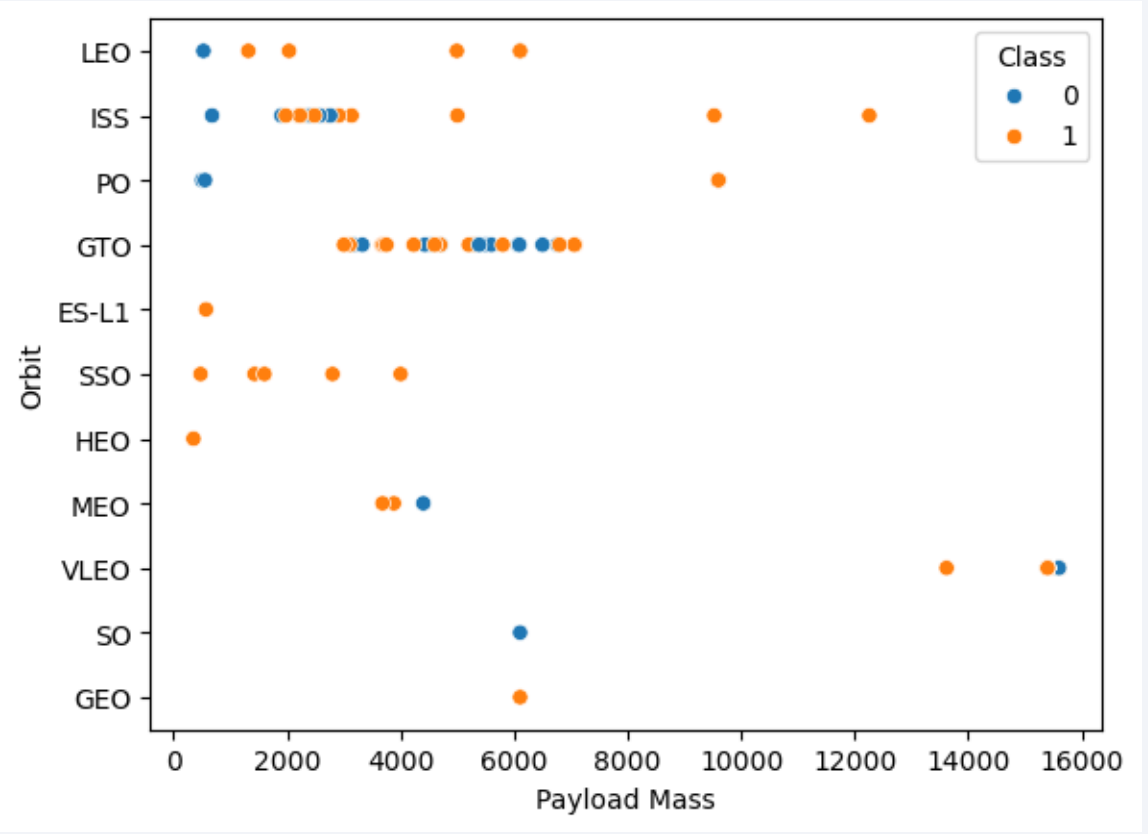
Scatter plot with explanations

FlightNumber

You can observe that in the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.

Payload vs. Orbit Type

Scatter point of payload vs. orbit type



Scatter plot with explanations

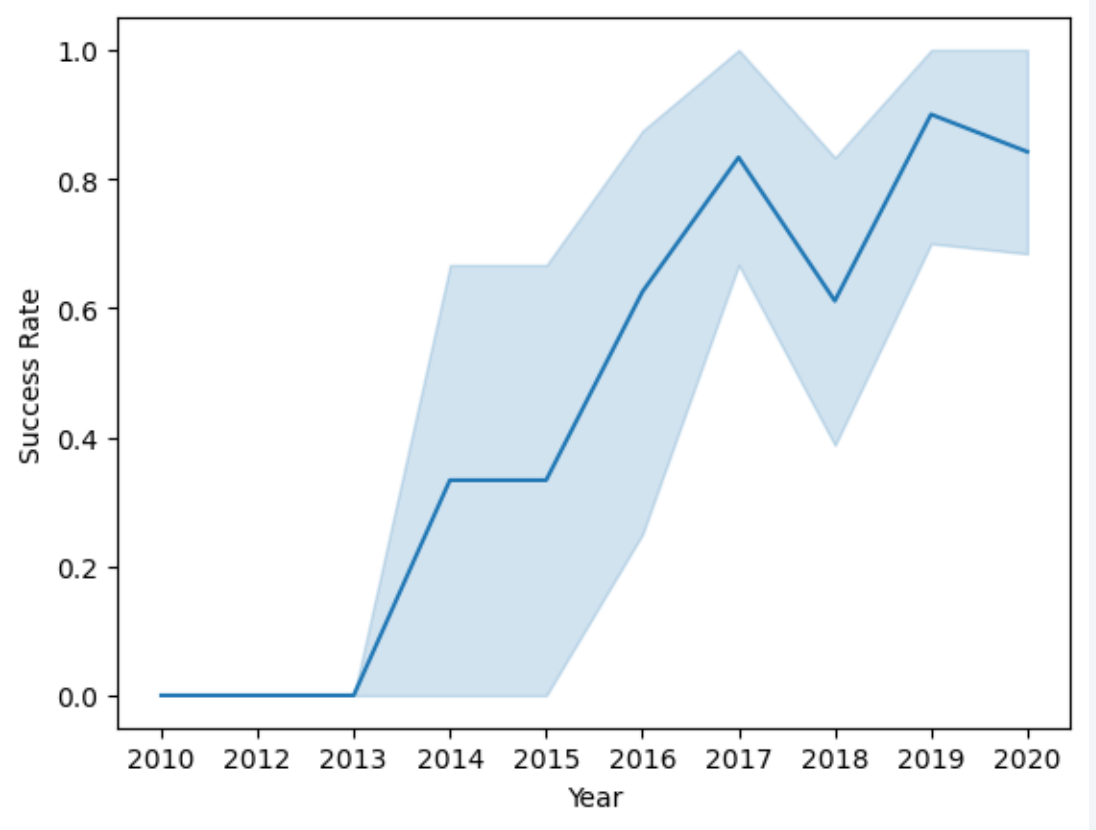
Payload Mass

With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

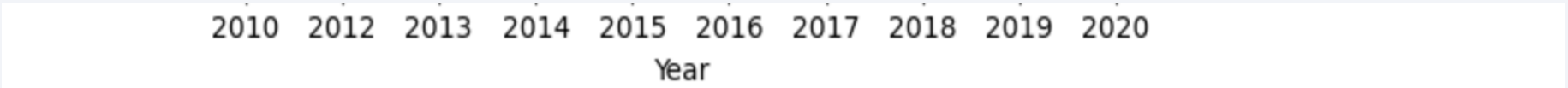
However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend

Line chart of yearly average success rate



Line chart with explanations



you can observe that the sucess rate since 2013 kept increasing till 2020

All Launch Site Names

- Names of the unique launch sites

```
[47]: %sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE;
 * sqlite:///my_data1.db
Done.
```

[47]:

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

The result shows all the unique launch sites, instead of repeating them because of the keyword DISTINCT, as it is used to display words once instead of repeating it according to how many times it was shown on the table.

Launch Site Names Begin with 'CCA'

From the table, you have to select all launch sites that begin with CCA by using * and the keyword WHERE as it is used to meet a condition and that condition is to find sites beginning with CCA and you use the % symbol to fill up unknown words when searching for a keyword and lastly you use LIMIT because you are only looking for specific number of records and not all the records.

```
[11]: %sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE "CCA%" LIMIT 5;
 * sqlite:///my_data1.db
Done.
```

[11]:

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attemp
2012-10-08	0:35:00	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

Total Payload Mass

- The total payload carried by boosters from NASA

```
[58]: %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE;
 * sqlite:///my_data1.db
Done.
[58]: SUM(PAYLOAD_MASS__KG_)
619967
```

To get the total of a column, you have to make use of the SUM() keyword, with the name of the column inside the parenthesis and ofcourse direct it to the table name using the keyword FROM.

Average Payload Mass by F9 v1.1

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

```
[60]: %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version = "F9 v1.1";
 * sqlite:///my_data1.db
Done.
[60]: AVG(PAYLOAD_MASS__KG_)
2928.4
```

To calculate the average of a column, you have to make use of the AVG() keyword with the name of the column inside the parenthesis but since we are looking for the average payload mass of a specific booster version which is F9 v1.1, we have to use the conditional keyword WHERE to assign the average to booster version F9 v1.1 only.

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad

```
[70]: %sql SELECT MIN(Date) FROM SPACEXTABLE WHERE Landing_Outcome = "Success (ground pad)";

 * sqlite:///my_data1.db
Done.
[70]: MIN(Date)
      2015-12-22
```

To find the first successful landing outcome, you have to make use of the keyword MIN() as this will take you to the lowest point but in this case the first date, also you have to make use of the conditional keyword WHERE as we are asked to find a specific location which is ground pad, that was success.

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

```
[71]: %sql SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = "Success (drone ship)"\n
          AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
 * sqlite:///my_data1.db
Done.
```

[71]:

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Here we have to make use of the keyword DISTINCT so as to get unique booster versions and then a conditional keyword because we are looking for a specific location which is drone ship and lastly BETWEEN because we are trying to find values between two values which is 4000 and 6000.

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes

```
[73]: %sql SELECT Mission_Outcome, COUNT(Mission_Outcome) FROM SPACEXTABLE GROUP BY Mission_Outcome;
 * sqlite:///my_data1.db
Done.
```

[73]:

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

Here we are trying to get the total number of mission outcomes, so we have to use the COUNT() keyword with the column name inside the parenthesis and we also need it to be grouped according to the success and failure outcomes using the keyword GROUP BY.

Boosters Carried Maximum Payload

To, find the boosters that carried max payload, we have to make use of a conditional keyword such as WHERE on the column we are looking into, the max() keyword to find the maximum and the column name placed inside the parenthesis and all this would be gotten from the table SPACEXTABLE.

```
[75]: %sql SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) \n
          FROM SPACEXTABLE);
 * sqlite:///my_data1.db
Done.
```

[75]:

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

```
[83]: %sql Select substr(Date, 6,2) as month, Date, Booster_Version, Launch_Site, Landing_Outcome FROM SPACEXT
          where Landing_Outcome = 'Failure (drone ship)' and substr(Date, 0,5)='2015';
       * sqlite:///my_data1.db
      Done.
```

[83]:

month	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Here, we have to select several columns such as date, booster versions, launch site, landing outcome of the record from 2015 making use of the conditional keyword WHERE to query the result to only provide records from the failed drone ship landing outcome.

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

```
[86]: %sql SELECT Landing_Outcome, COUNT(Landing_Outcome)FROM SPACEXTABLE WHERE Date BETWEEN "2010-06-04" \n
          AND "2017-03-20"  GROUP BY Landing_Outcome ORDER BY Date DESC;
 * sqlite:///my_data1.db
Done.
```

[86]:

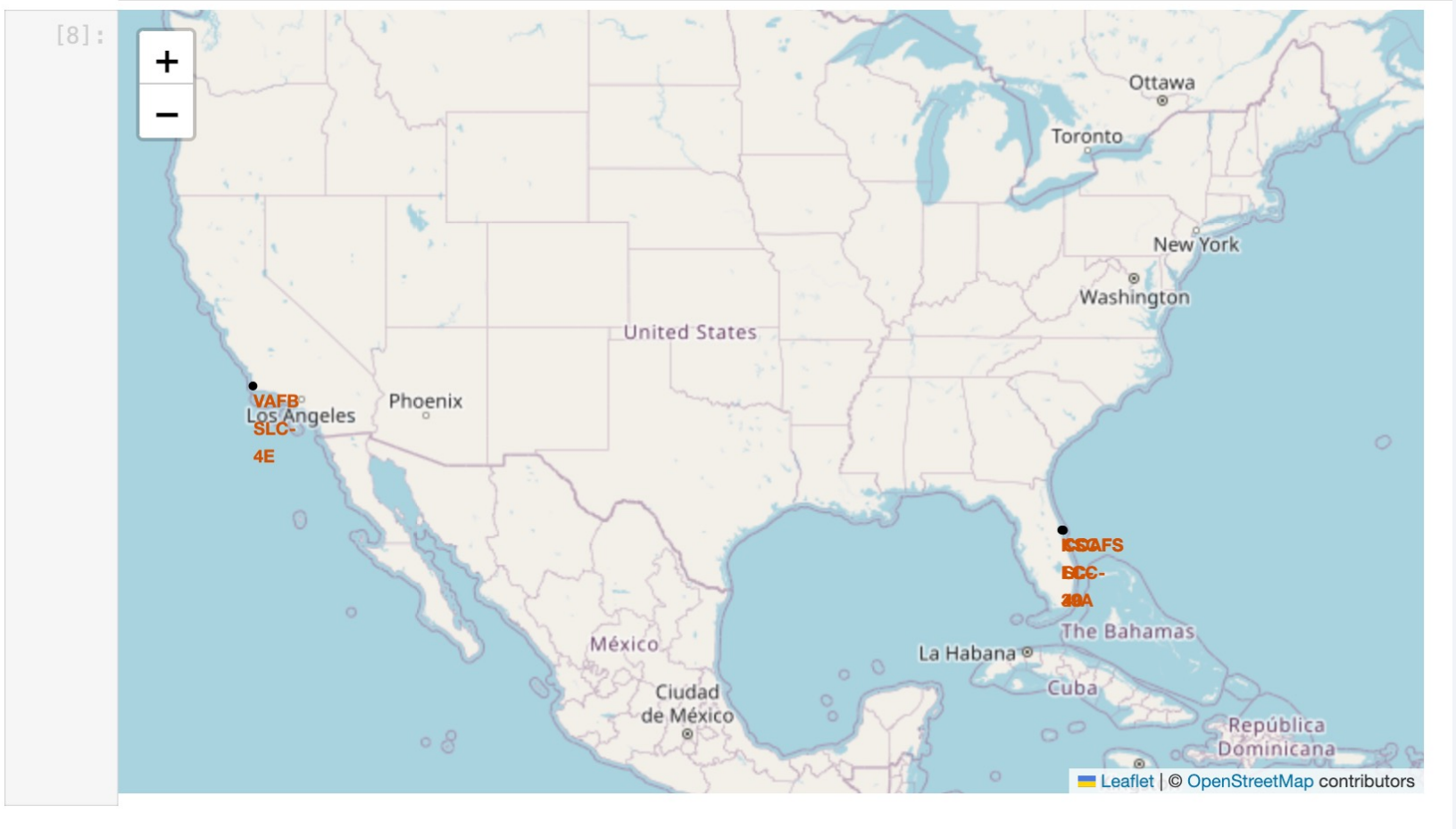
Landing_Outcome	COUNT(Landing_Outcome)
Success (drone ship)	5
Success (ground pad)	3
Precluded (drone ship)	1
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
No attempt	10
Failure (parachute)	2

Here, we have to make use of keywords such as COUNT(), WHERE, AND, GROUP BY, ORDER BY and DESC where count is used to count the numbers of specific entity in a column, where is a conditional keyword, and is used as a logical keyword, group by is used to group a specific entity, order by is used to arrange a specific group of entities and dese is also descending order, how it would be arranged.


Section 3
Launch Sites
Proximities Analysis

Generated Map of All Launch Sites

From the generated map, we can deduce that all launch sites are located at coastal lines. Launch site VAFB SLC-4E is located in California while Launch sites CCAFS LC-40, CCAFS SLC-40 and KSC LC-39A are all located in the same area of Florida.

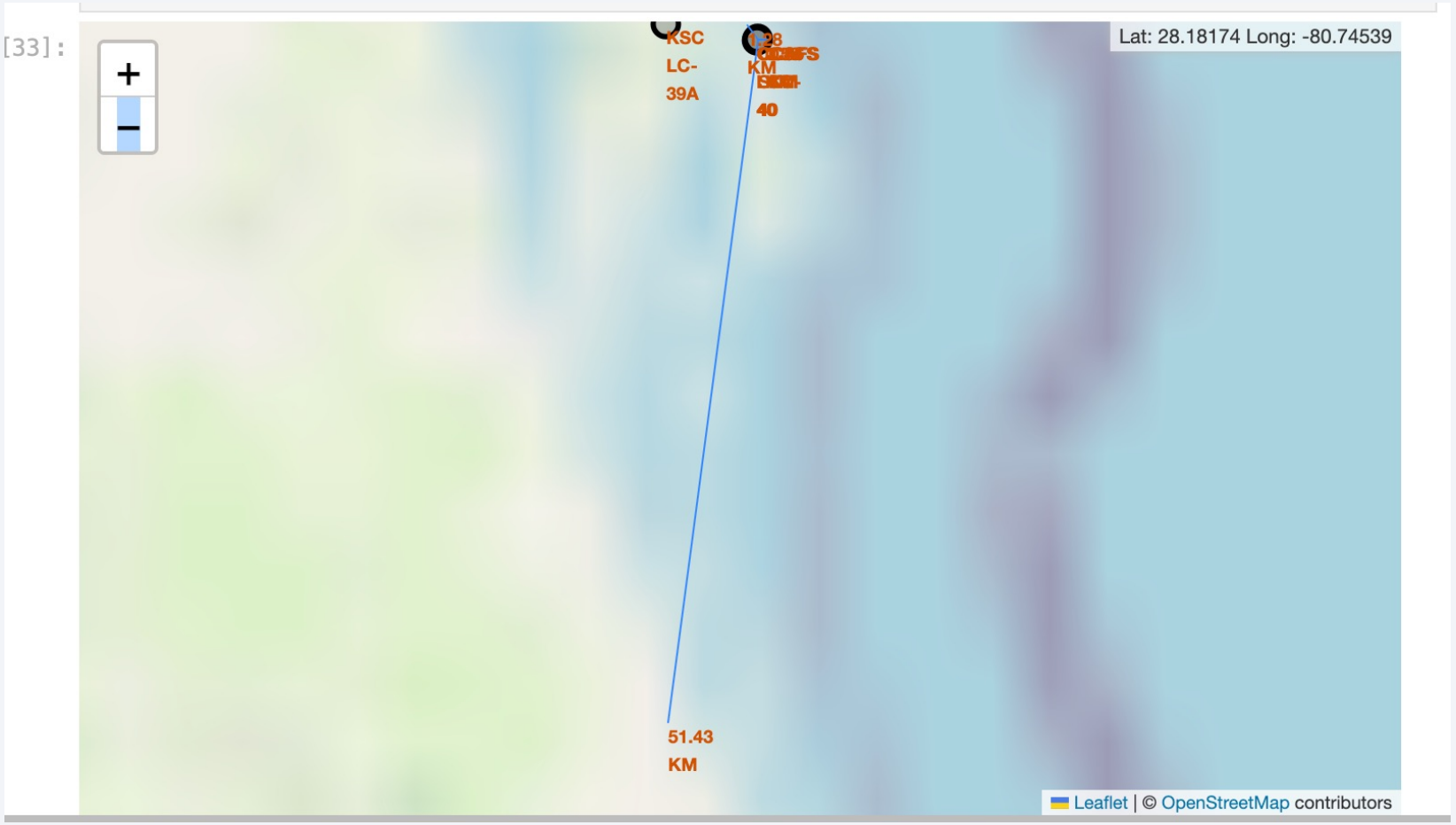


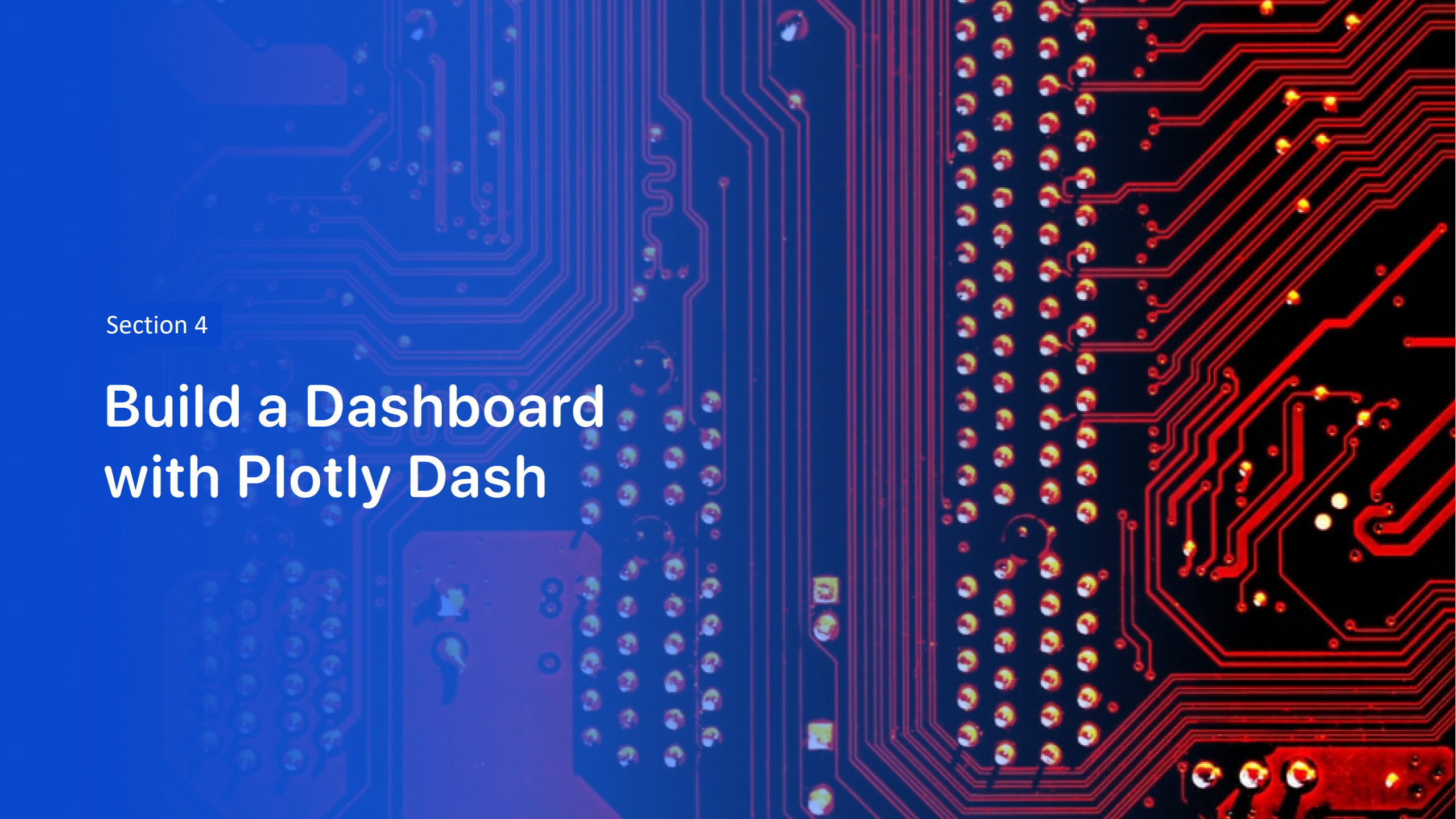
Color Labeled Map of Launch Outcomes

From the generated map, launch site CCAFS SLC-40 has a 43% success rate and a 57% failure rate and its in a close proximity with launch site CCAFS LC-40

Generated Map of CCAFS SLC-40 and its Proximites

From the generated map, launch site CCAFS SLC-40 is 1.28km to the railways and therefore in close proximity, 0.58km to the highway which is also in close proximity, coastline is 0.86km to the launch site which is also in close proximity but 51.43km to the city which is a hugh distance compared to the rest.

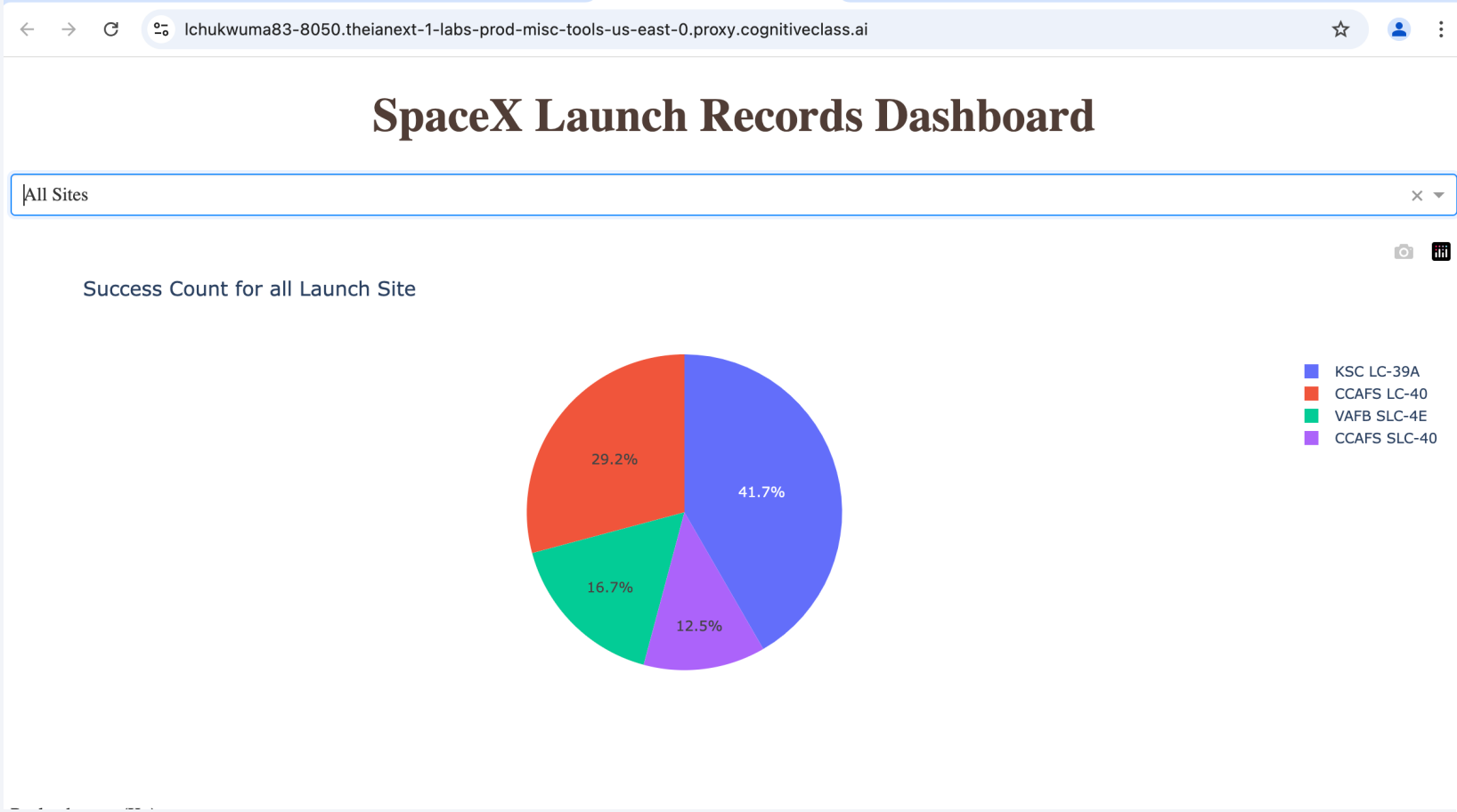



Section 4
Build a Dashboard with Plotly Dash

Success Count for All Launch Sites

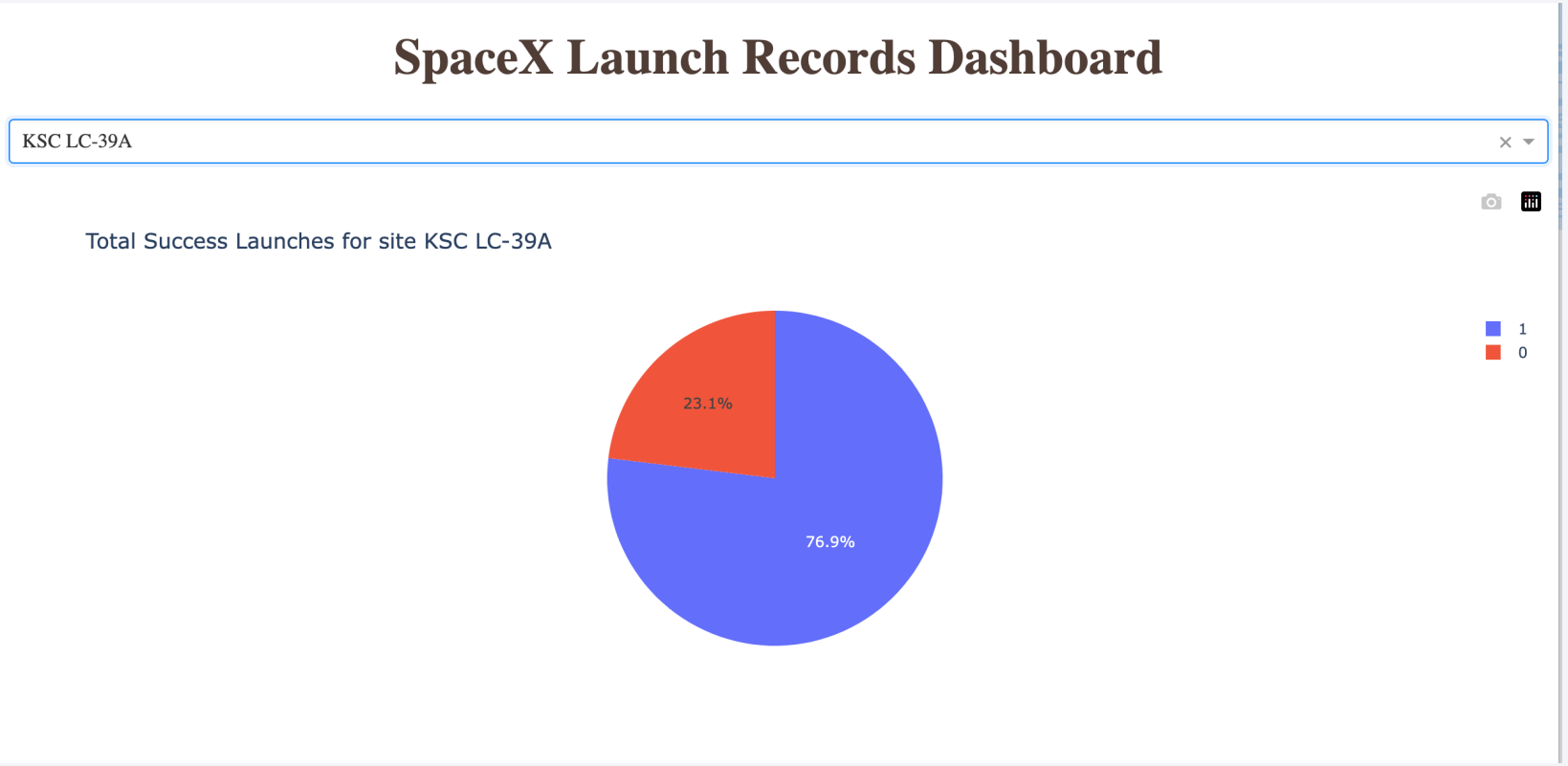
Explain the important elements and findings on the screenshot

From the piechart, launch site KSC LC-39A has the highest launch success count, while launch site CCAFS SLC-40 has the lowest launch success count



Total Success Count for site KSC LC-39A

From the piechart, there is a 76.9% success rate in launch site KSC LC-39A to 23.1% failure rate. The purple color in the legend which says 1 signifies success while the orange color which says 0 signifies fail.



Success Count on Booster Version and Payload Mass

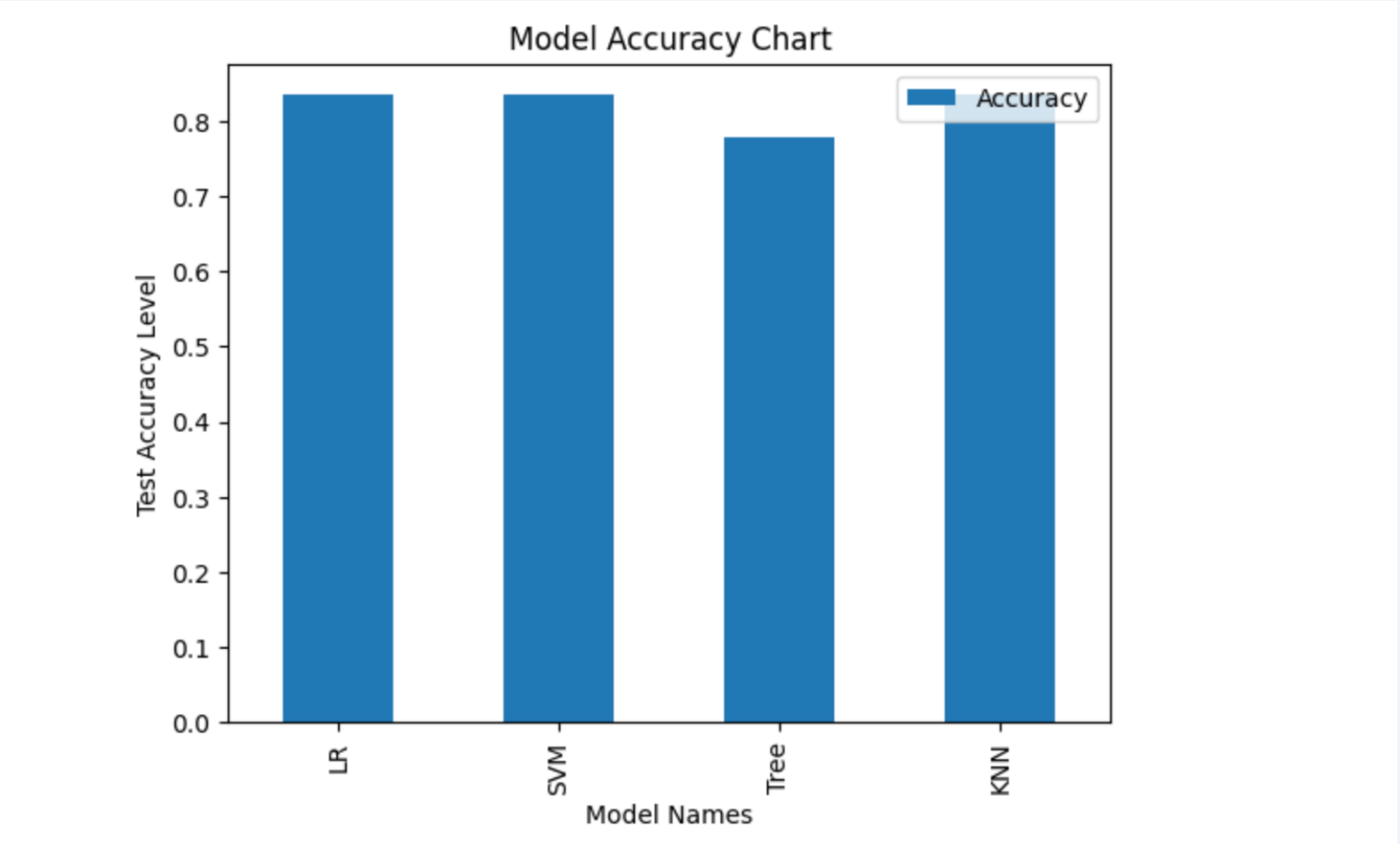
From the scatter plot, we can see that for site KSC LC-39A, which is said to be the most successful launch site, booster version B4 has 100% success rate but FT has a success rate of 100% if the payload mass is less than 5500 and 0% if it is above. Meanwhile for site CCAFS SLC-40, B4 booster version has a success rate of 40% while booster version FT has a success rate of 50%.

Section 5

Predictive Analysis (Classification)

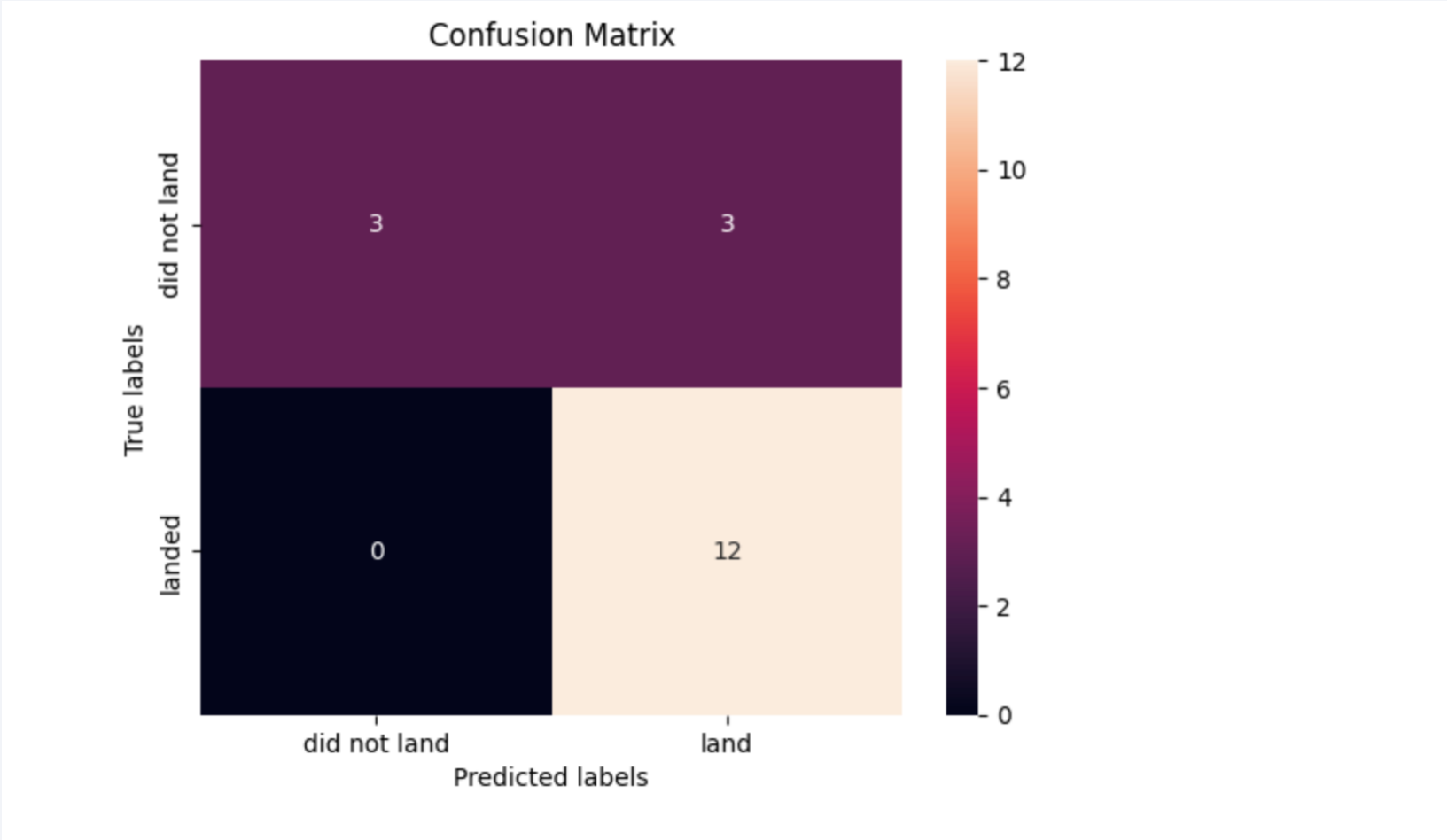
Classification Accuracy

From the model accuracy chart, logistic regression, support vector machine and k nearest neighbor performed best with an accuracy of 83.33% unlike the decision tree classifier with an accuracy of 77.8%.



Confusion Matrix

From the confusion matrix, we can see that there's a true label and a predicted label. A true label is what actually happened while a predicted label is what was predicted to happen. We can see that 12 was predicted to land and they landed, so that is a TP, while 0 was predicted to not land, that is FN. 3 was also predicted not to land and 3 didn't land, rather 12 so that is TN. Lastly 3 was predicted to land but 3 didn't land so that is FP. TP (true positive), FN(false negative), TN(true negative), FP(false positive).



Conclusions

- Landing outcome depends on flight parameters like launch sites, payload mass, the orbit which it is launched into and the boosters.
- Launch sites are closer to places like railway lines and the coastlines but farther from the cities.
- Orbits such as ES-L1, GEO, HEO and SSO had the highest success rate.
- Launch success rate kept increasing overtime.
- KSC LC-39A had the highest launch site success rate.
- Models such as Logistic Regression, Support Vector Machine and K-Nearest Neighbor had the highest test accuracy of 83.3%

Appendix

- Code snippet for the model accuracy chart

[49]:
```
Log = logreg_cv.score(X_test, Y_test)
svm = svm_cv.score(X_test, Y_test)
tree = tree_cv.score(X_test, Y_test)
knn = knn_cv.score(X_test, Y_test)

data = {"Name" : ["LR", "SVM", "Tree", "KNN"], "Accuracy": [Log, svm, tree, knn] }
df = pd.DataFrame(data)
df.set_index("Name", inplace=True)
df.plot(kind= "bar")
plt.title("Model Accuracy Chart")
plt.xlabel("Model Names")
plt.ylabel("Test Accuracy Level")
plt.show()
```

Thank you!