IBM Developer
SKILLS NETWORK

SpaceX Falcon 9 first stage Landing Prediction

IBM DATA SCIENCE CAPSTONE PROJECT

Viktoriya Shilina

10.11.2023

Executive Summary

Summary of methodology

To get conclusions were used few steps:

- Data Collection
- Data Wrangling
- Exploratory Data Analysis (includes SQL analysis and Plotting)
- Interactive Visual Analytics using Folium and Dash
- Predictive Analysis based on Machine Learning techniques

Summary of results

The useful information of this project is displayed in these visual outcomes:

- Exploratory Data Analysis (EDA) results
- Geospatial analytics
- Interactive dashboard
- Predictive analysis of classification models

Outline

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

Introduction

- SpaceX advertises launches of Falcon 9 rockets with a cost of around $62mln. This is 2.5 times cheaper than other providers offer. This difference because SpaceX can land, and then re-use the first stage of the rocket.

- Prediction of the Falcon 9 first stage successful land can help to determine the expenses of the launch and this information can be used by companies who wants to bid against SpaceX for a rocket launch.

- In this project to get conclusions was used such a parameters like Launch Site (location), Payload Mass, Orbit which rocket reached, Flight Number to observe the continuous flight attempts.

Section 1

Methodology

Data Collection

Data was taken from different sources :

- https://api.spacexdata.com/v4/launches/past
- https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/dataset_part_1.csv

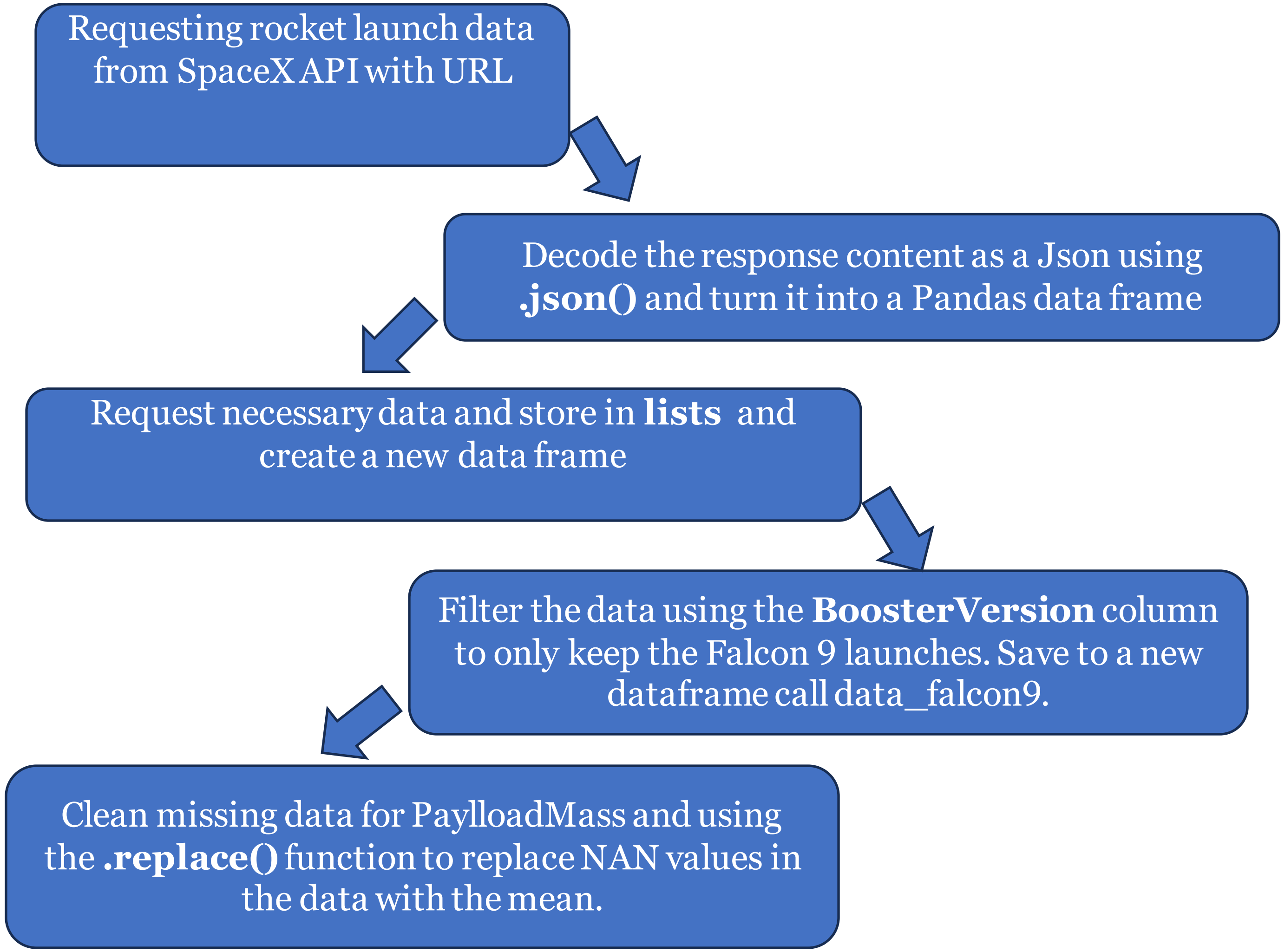
Data Collection-SpaceX API

3 task to complete:

1. Request and parse the SpaceX launch data using the GET request

2. Filter the data frame to only include Falcon 9 launches

3. Dealing with Missing Values



Requesting rocket launch data from SpaceX API with URL

Decode the response content as a Json using **.json()** and turn it into a Pandas data frame

Request necessary data and store in **lists** and create a new data frame

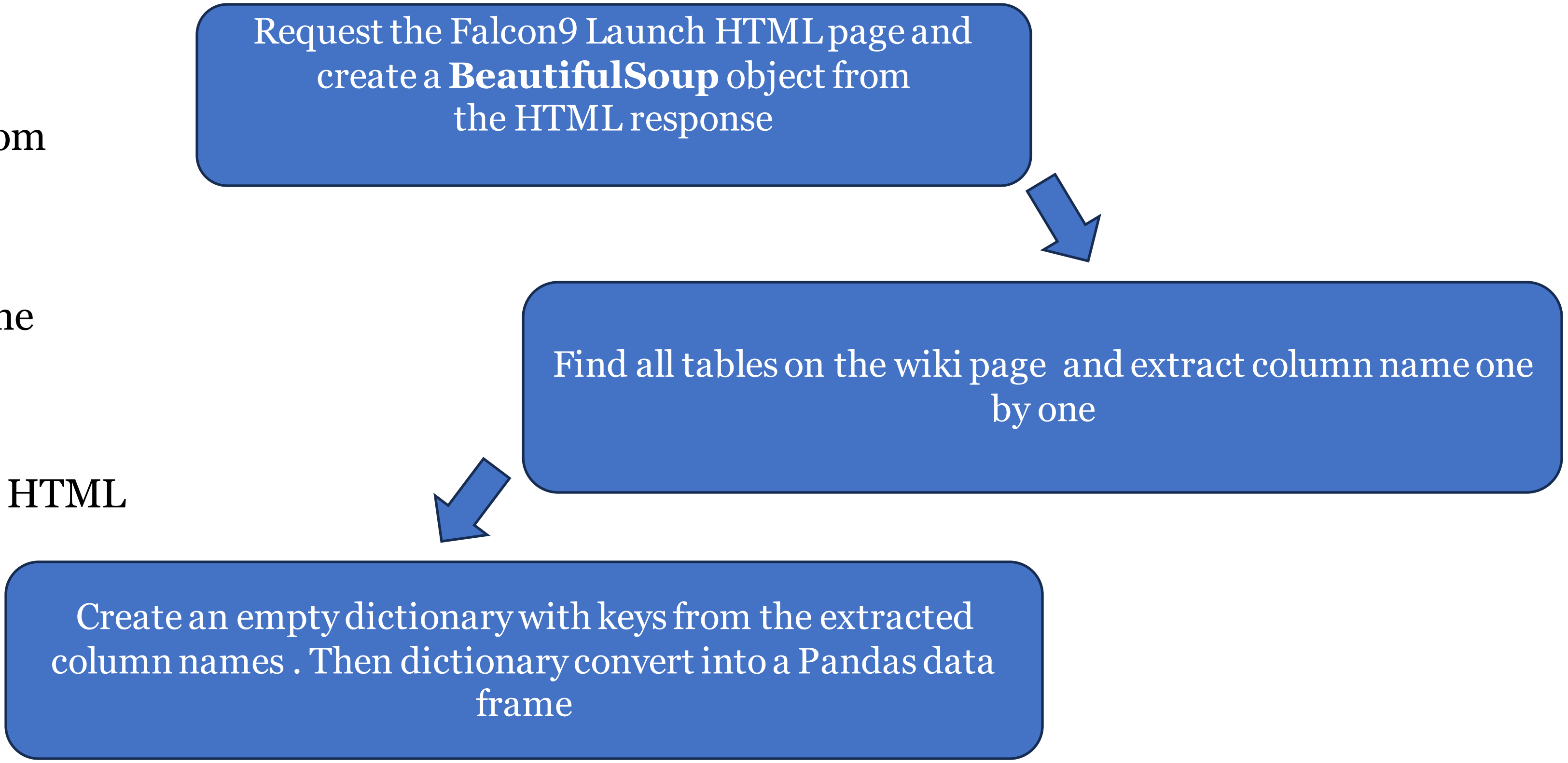
Filter the data using the **BoosterVersion** column to only keep the Falcon 9 launches. Save to a new dataframe call data_falcon9.

Clean missing data for PaylloadMass and using the **.replace()** function to replace NAN values in the data with the mean.

Collection GitHub

Data Collection - Scraping

- Request the Falcon9 Launch Wiki page from its URL
- Extract all column/variable names from the HTML table header
- Create a data frame by parsing the launch HTML tables



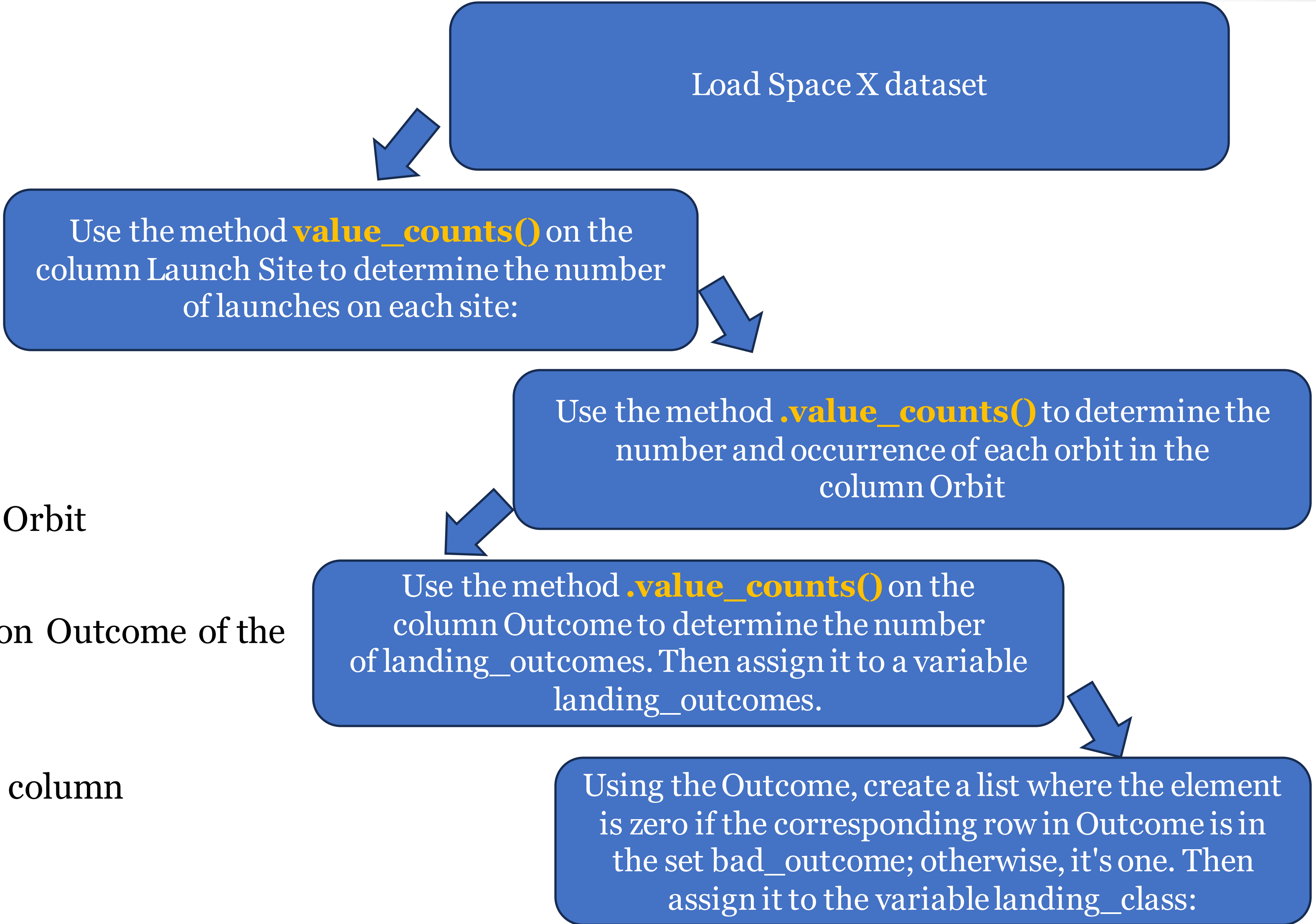
Request the Falcon9 Launch HTML page and create a **BeautifulSoup** object from the HTML response

Find all tables on the wiki page and extract column name one by one

Create an empty dictionary with keys from the extracted column names . Then dictionary convert into a Pandas data frame

Scraping GitHub

Data Wrangling



Load Space X dataset

Use the method **value_counts()** on the column Launch Site to determine the number of launches on each site:

Use the method **.value_counts()** to determine the number and occurrence of each orbit in the column Orbit

Use the method **.value_counts()** on the column Outcome to determine the number of landing_outcomes. Then assign it to a variable landing_outcomes.

Using the Outcome, create a list where the element is zero if the corresponding row in Outcome is in the set bad_outcome; otherwise, it's one. Then assign it to the variable landing_class:

- Calculate the number of launches on each Site
- Calculate the number and occurrence of each Orbit
- Calculate the number and occurrence of Mission Outcome of the Orbits
- Create a landing outcome label from Outcome column

Wrangling GitHub

EDA with Data Visualization

- For visualization were chosen 3 types of charts:

Scatter plot	**Bar plot**	**Line chart**
Suitable to observe relationships between 2 variables	Most useful to compare values between multiple groups	One of the best choice for showing data trends over time
1. FlightNumber vs. PayloadMass 2. FlightNumber vs LaunchSite 3. Payload vs Launch Site 4. FlightNumber vs Orbit type 5. Payload vs Orbit type	Was observed the success rate for each type of orbit	X axis was year and Y axis was average success rate, so we got the average launch success trend

EDA GitHub

EDA with SQL

SQL was used to gather some information from dataset. SQL queries were applied to:

- Display the names of the unique Launch Sites in the space mission.
- Display 5 records where Launch Sites begin with the string 'CCA'.
- Display the total Payload Mass carried by boosters launched by NASA (CRS).
- Display average Payload Mass carried by Booster Version F9 v1.1.
- List the date when the first successful Landing Outcome in ground pad was achieved.
- List the names of the Boosters which have success in drone ship and have Payload Mass greater than 4000 but less than 6000.
- List the total number of successful and failure Mission Outcomes.
- List the names of the Booster Versions which have carried the maximum Payload Mass.
- 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.
- 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.

SQL GitHub

Build an Interactive Map with Folium

First was arranged data set and initialized folium **Map object**, then were added a circle for each Launch Site.

As launch happens in one of the four Launch Sites many launch records will have the same coordinate. So marker clusters is a good way to simplify a map containing.

So were created markers for all launch records. If a launch was successful, then was used a **green** marker and if a launch was failed, was used a **red** marker. To get this point first was defined each class: successful launch (class=1), unsuccessful - (class=0).

Color-labeled markers in marker clusters make easy to identify which launch sites have relatively high success rates.

To explore and analyze the proximities of launch sites was added a **MousePosition** on the map to get coordinate for a mouse over a point on the map.

Knowing coordinates of launch site and proximities we can find distance between them using def **calculate_distance** (lat1, lon1, lat2, lon2), and **folium.PolyLine** help to create distance line between a launch site to the selected point.

Folium GitHub

Predictive Analysis (Classification)

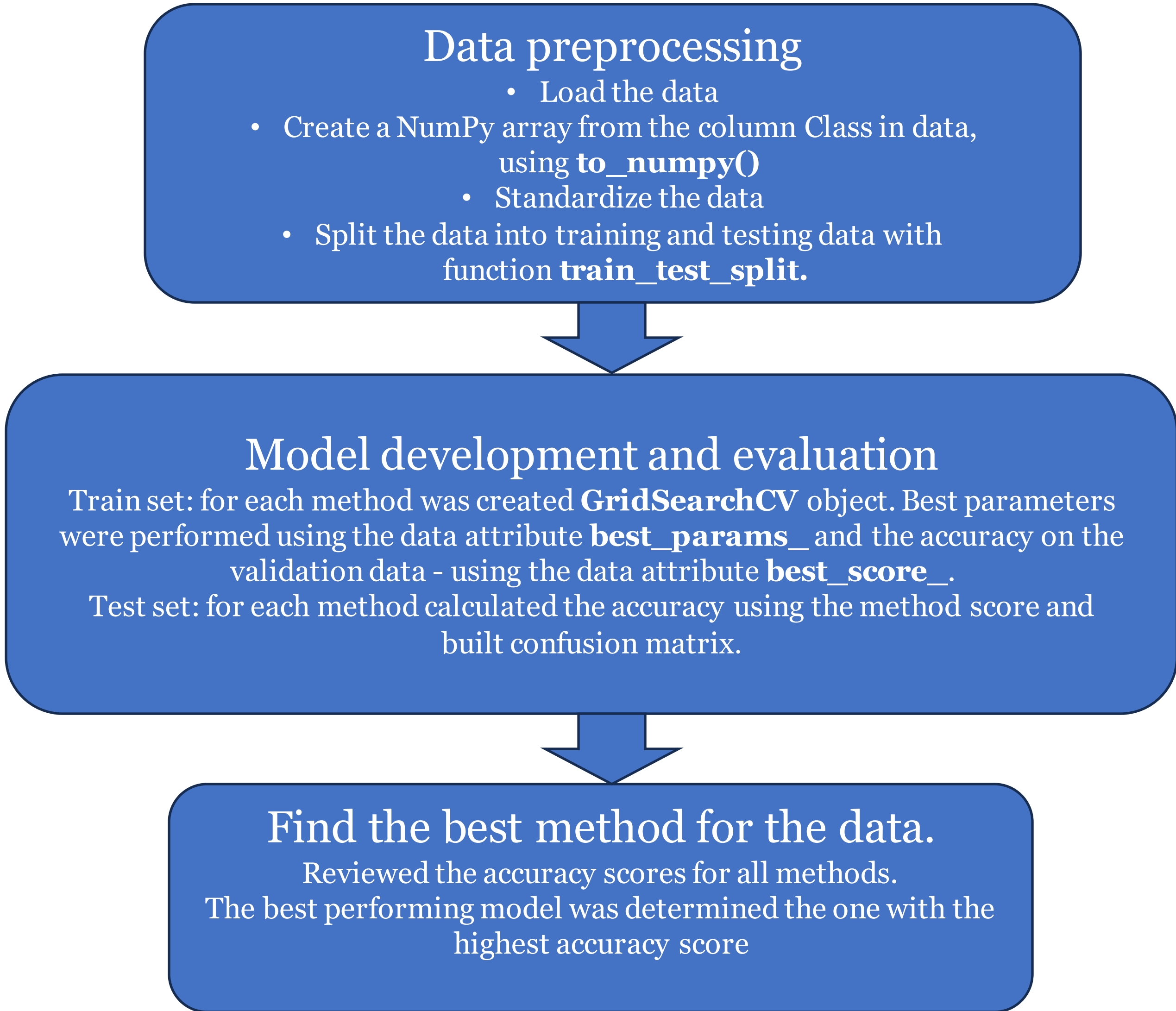
The purpose of this step was to build a machine learning pipeline to predict success land of the Falcon 9.

To get that point our data was **standardized**, and **train_test_split**, used to split the data into training and testing sets.

Four methods were used for prediction: Logistic Regression, Support Vector machines, Decision Tree Classifier, and K-nearest neighbors.

By comparing accuracies of each one we found model that perform best for the data set.

Classification GitHub



Data preprocessing

- Load the data
- Create a NumPy array from the column Class in data, using **to_numpy()**
- Standardize the data
- Split the data into training and testing data with function **train_test_split.**

Model development and evaluation

Train set: for each method was created **GridSearchCV** object. Best parameters were performed using the data attribute **best_params_** and the accuracy on the validation data - using the data attribute **best_score_**.

Test set: for each method calculated the accuracy using the method score and built confusion matrix.

Find the best method for the data.

Reviewed the accuracy scores for all methods.

The best performing model was determined the one with the highest accuracy score

Build a Dashboard with Plotly Dash

In this part were created two types of charts using dashboard application: pie and scatter.

- First was created pie-chart that visualized percentage of successful (and failure in case of separate Launch Site) launches. Dropdown has 5 options:

  1. **summary of successful launches,**
  2. **CCAFS LC-40 site,**
  3. **CCAFS SLC-40 site,**
  4. **KSC LC-39A site,**
  5. **VAFB SLC-4E site.**

- According to each dropdown at the bottom appeared scatter plot showing relationship between lAnding Outcomes and the Payload Mass for different Boosters. This chart visualized the influence of different variables on successful launch.

[Dash GitHub]

Results

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

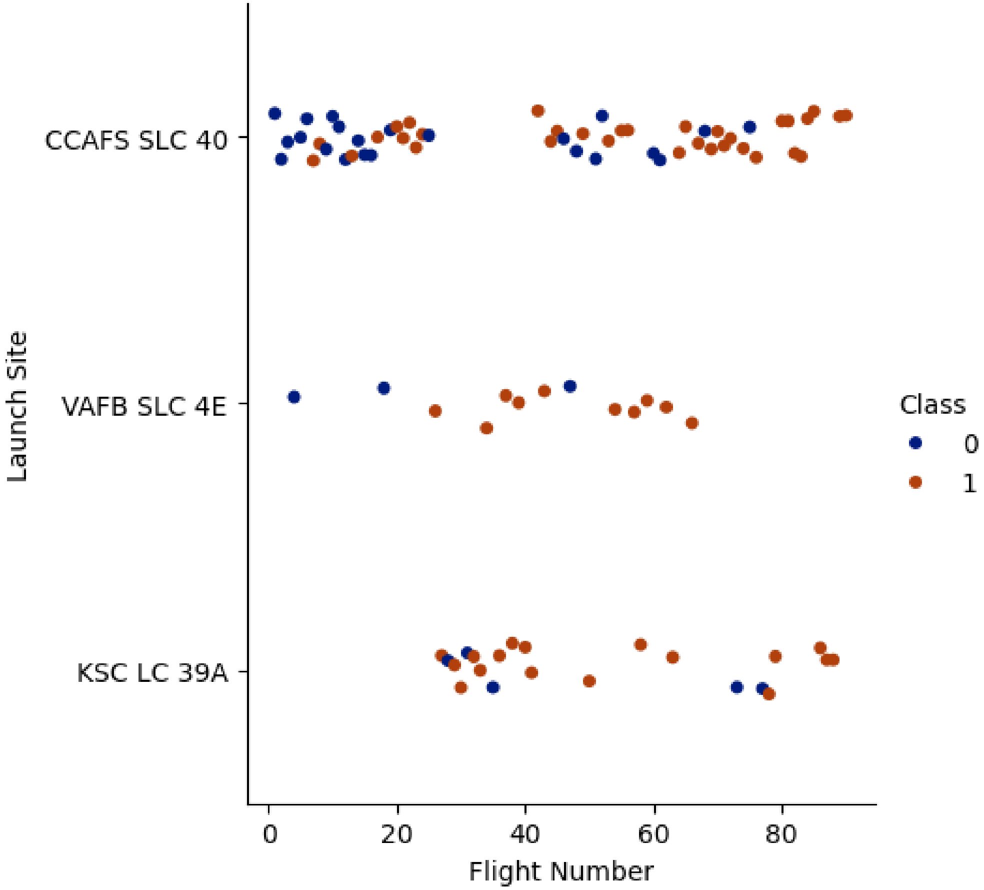
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

What we can notice from this chart:

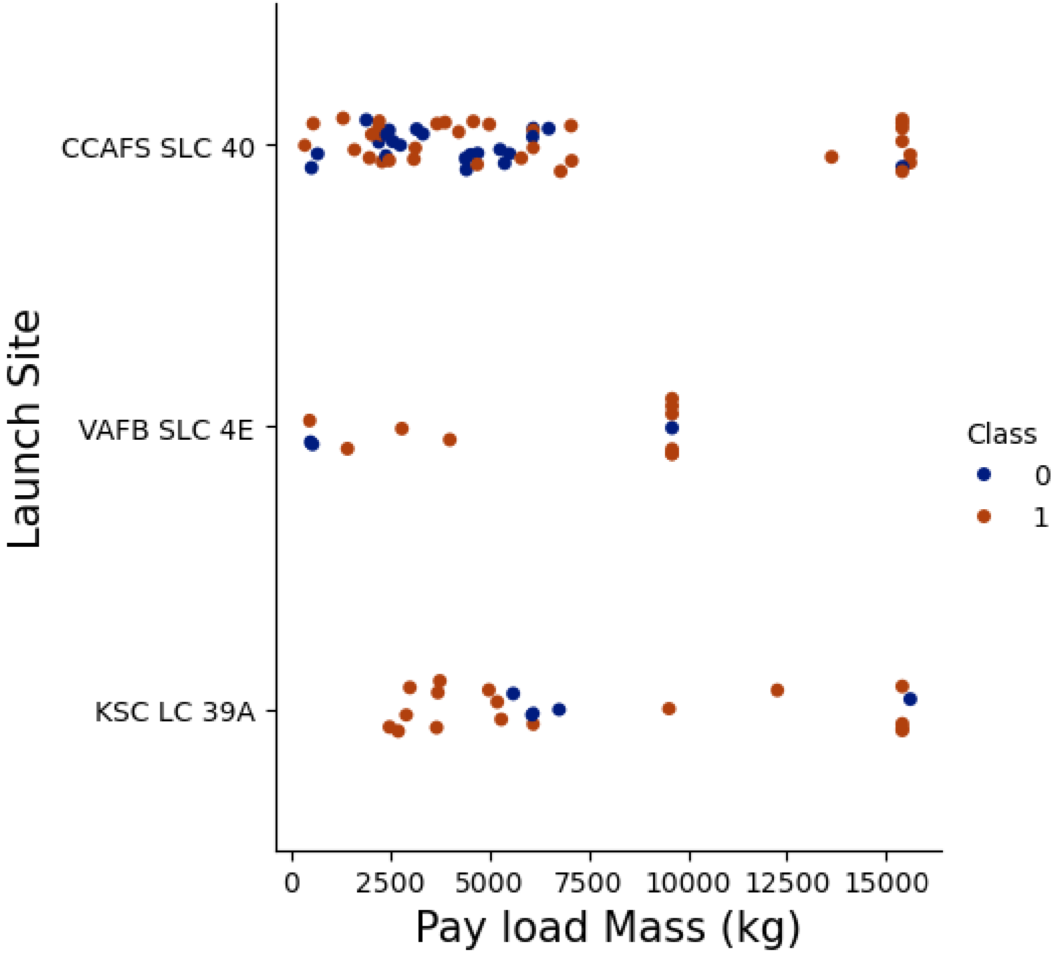
- With the increase of number of flights success rating is growing.
- Earlier launches were done mainly from CCAFS SLC 40-launch site and the success rate was low.
- After number of flight 40 for all launch sites the increase of the success rate is observed.



Payload vs. Launch Site

What we can notice from this chart:

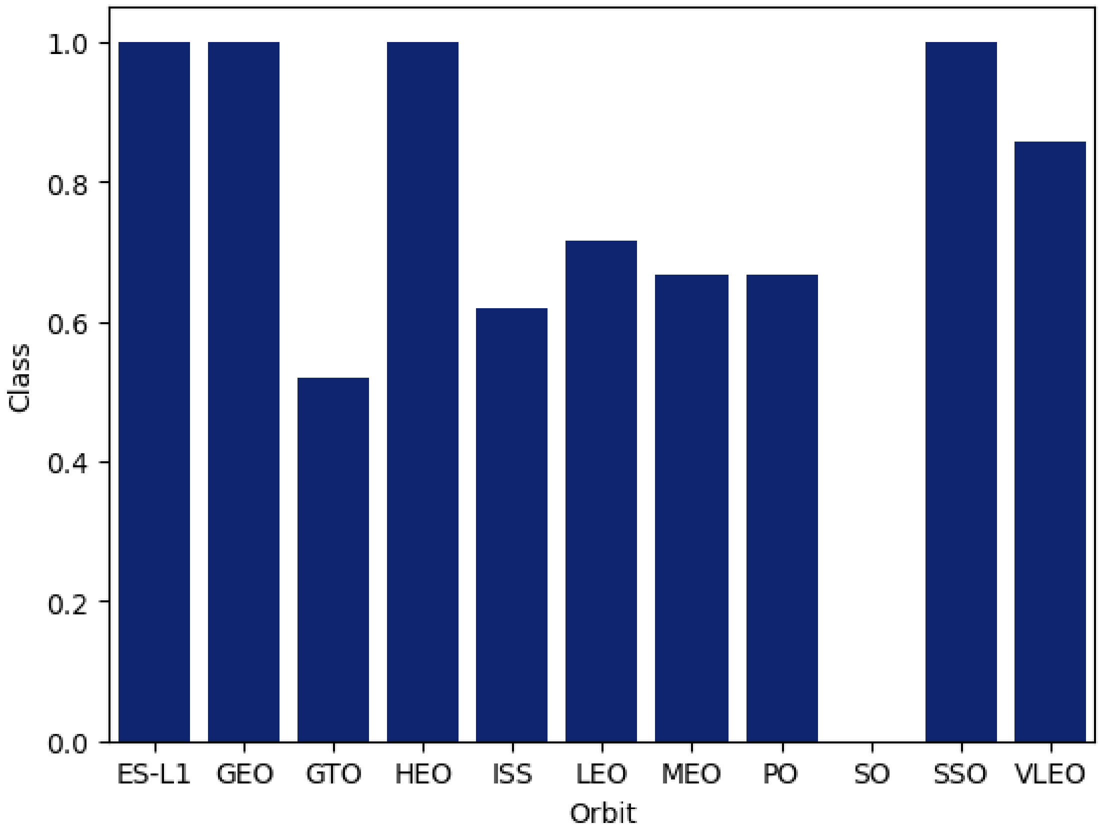
- If Payload Mass is more that 7500 kg the chance of success launch is higher but still there is not enough data for heavy launches.
- Most of the launches were done with Payload Mass, less than 7000 kg.



Success Rate vs. Orbit Type

What we can notice from this chart:

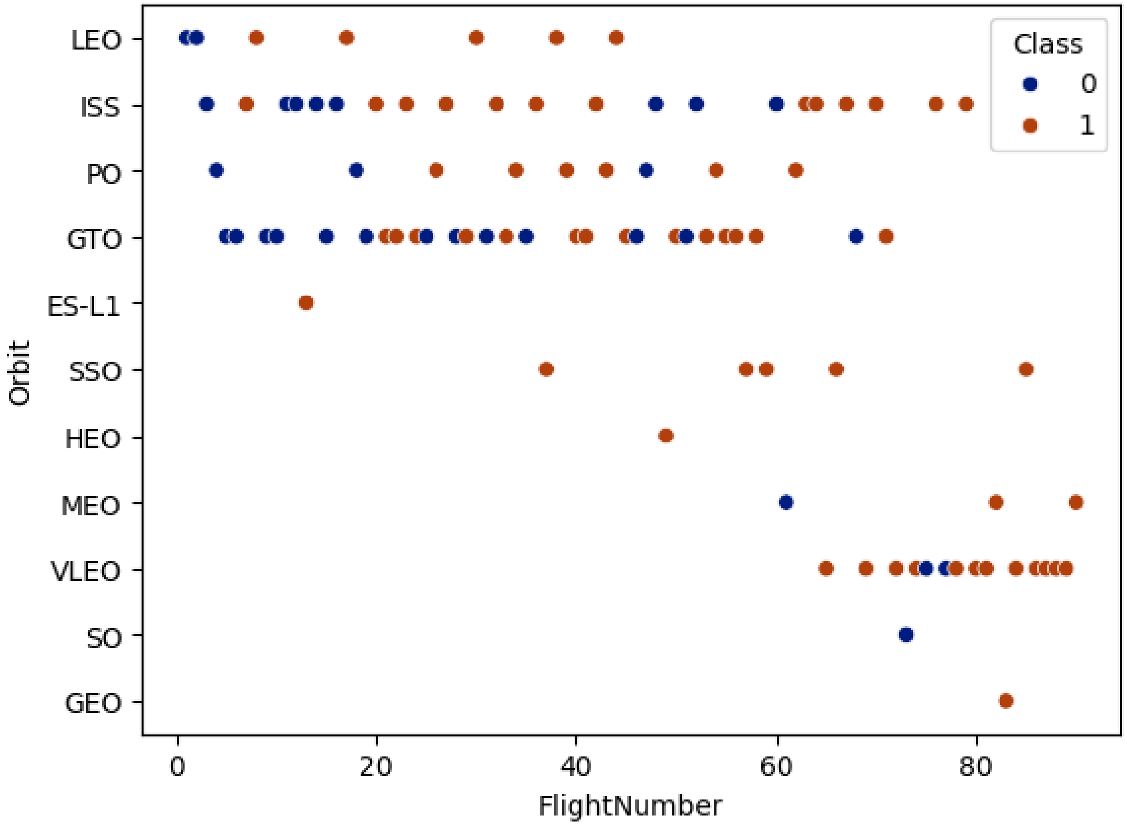
- Orbits with highest success rate are (100%): ES-L1, GEO, HEO, SSO.
- Lowest success rate was observed for SO Orbit.



Flight Number vs. Orbit Type

What we can notice from this chart:

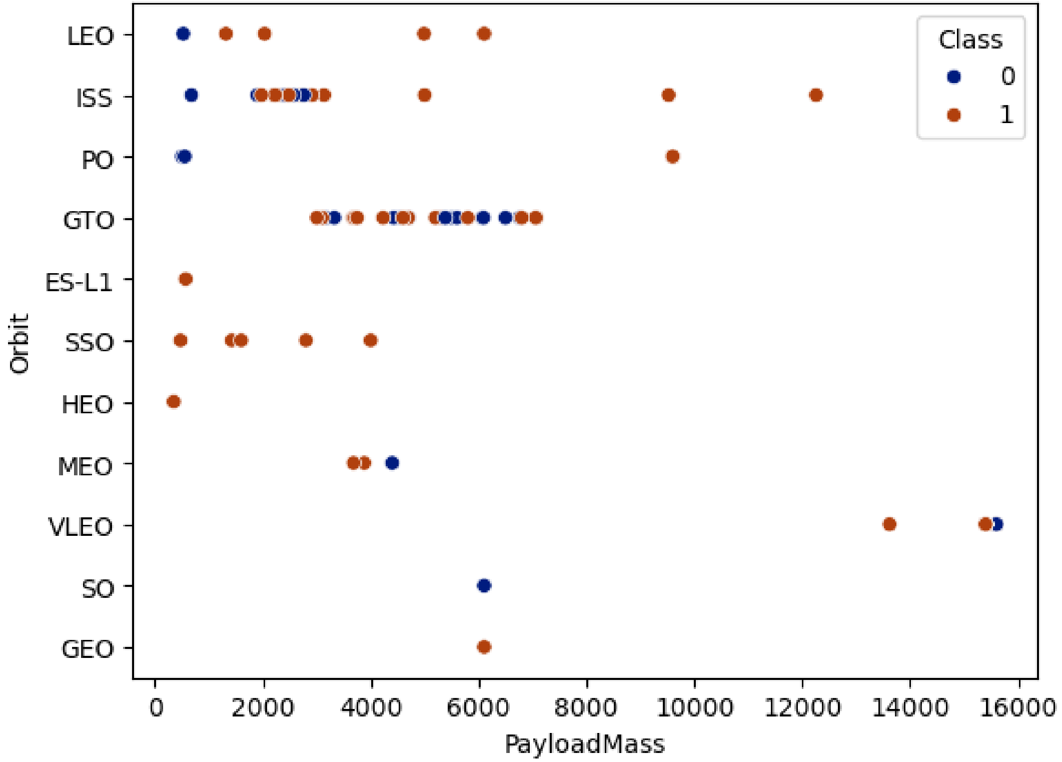
- In previous plot we notice high success rate for ES-L1, GEO, HEO Orbits, but as it was only one launch for each orbit we cannot conclude that these orbits are better to get success launch.
- Amount of a launches for SSO orbit – 5 and all successful. Much better than others with 100% success rate but still not enough data to get some conclusion.
- In general we can see that Flight Number increase and success rate is increase.



Payload vs. Orbit Type

What we can notice from this chart:

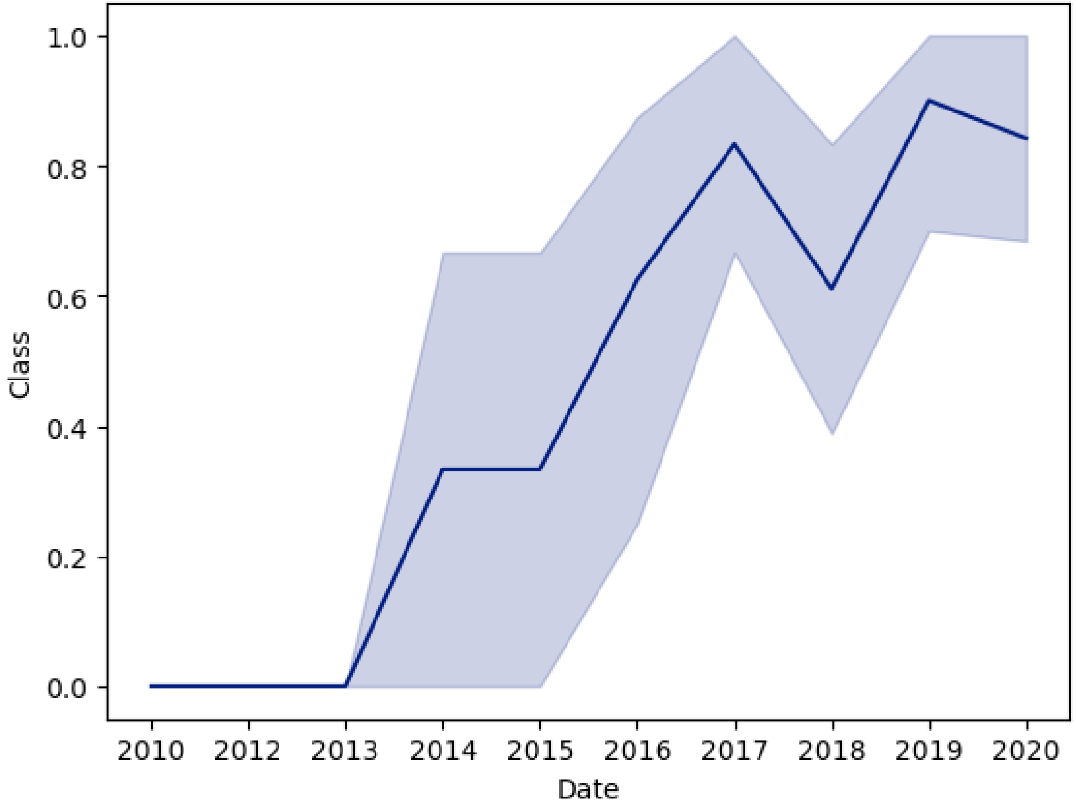
- SSO and LEO Orbit associated with light payload mass.
- For VLEO Orbit, on the contrary, we observe launches only with heavy payload mass.
- All the launches for GTO orbit were done with the mass between 3000 and 8000 kg (medium weight).



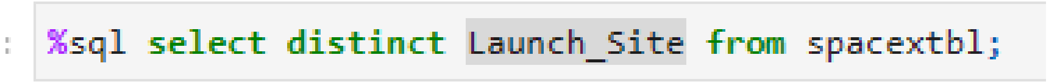
Launch Success Yearly Trend

What we can notice from this chart:

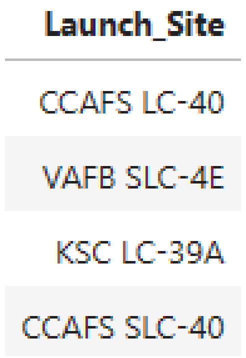
- From 2013 success rate for the launch is increasing.
- After 2015 success rate is reached 50% and kept going higher and even with the drop in 2017-18 never came lower than 60%.



All Launch Site Names



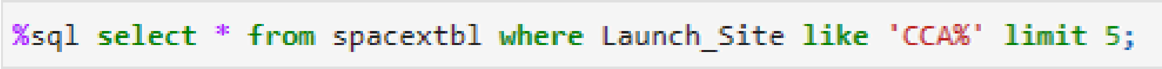
```
%sql select distinct Launch_Site from spacextbl;
```



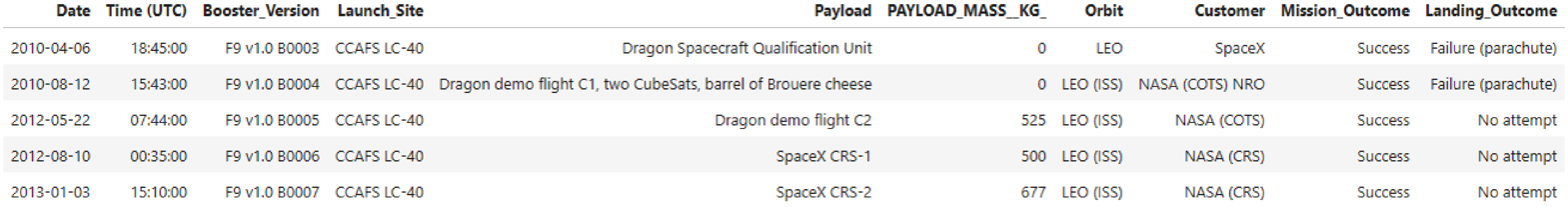
Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

To select unique Launch Sites was used command **DISTINCT**, so we got names without overlapping.

Launch Site Names Begin with 'CCA'



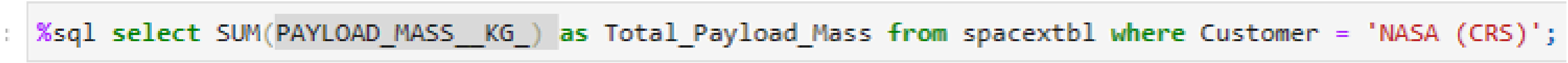
```
%sql select * from spacextbl where Launch_Site like 'CCA%' limit 5;
```



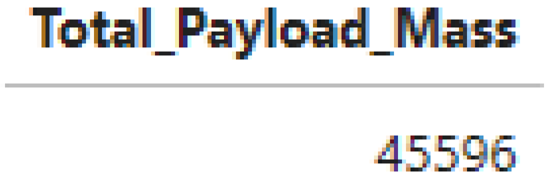
Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

To get records where Launch Site name begin with CCA was used command **LIKE** with wildcard 'CCA%' , as we need only 5 records was used command **LIMIT 5**.

Total Payload Mass launched by NASA



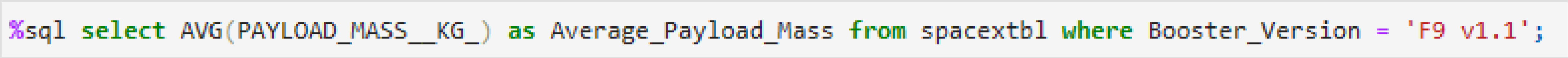
```
%sql select SUM(PAYLOAD_MASS__KG_) as Total_Payload_Mass from spacextbl where Customer = 'NASA (CRS)';
```



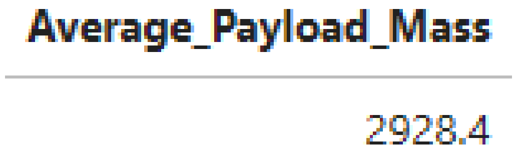
Total_Payload_Mass
45596

To get total sum of Payload Mass was used function **SUM**, to extract only records with Customer NASA(CRS) used command **WHERE.**

Average Payload Mass by F9 v1.1



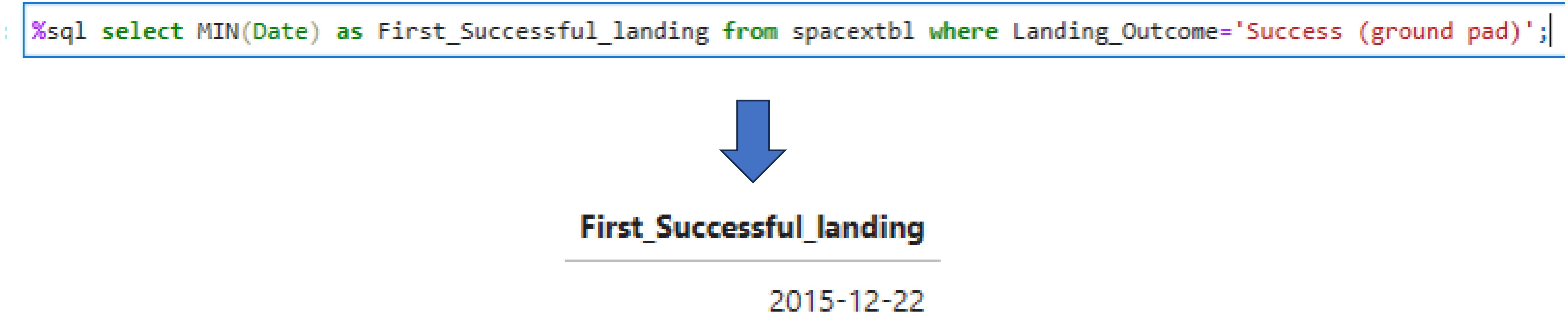
```
%sql select AVG(PAYLOAD_MASS__KG_) as Average_Payload_Mass from spacextbl where Booster_Version = 'F9 v1.1';
```



Average_Payload_Mass
2928.4

To get average Payload Mass was used function **AVG**, to extract only records with Booster Version F9 v1.1 used command **WHERE**.

First Successful Ground Landing Date

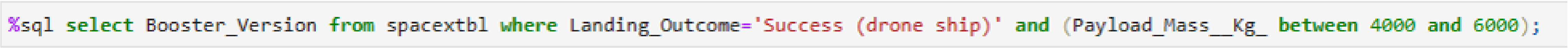


```
%sql select MIN(Date) as First_Successful_landing from spacextbl where Landing_Outcome='Success (ground pad)';
```

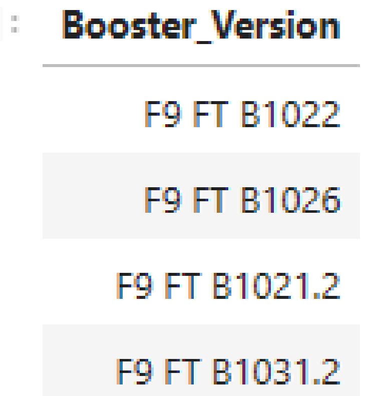
First_Successful_landing
2015-12-22

To get first successful ground landing was used command **MIN** for Date so we got first date. Condition was described by command **WHERE** to specify Landing Outcome that had to be successful on the ground.

Successful Drone Ship Landing with Payload between 4000 and 6000



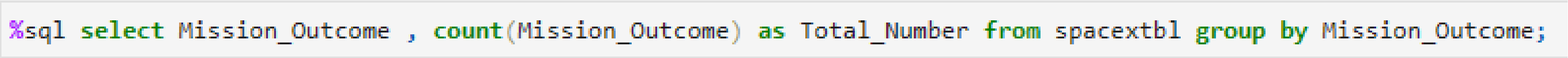
```
%sql select Booster_Version from spacextbl where Landing_Outcome='Success (drone ship)' and (Payload_Mass__Kg_ between 4000 and 6000);
```



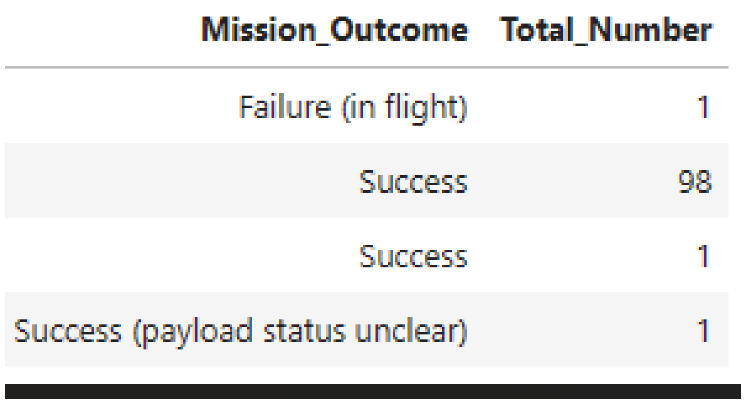
Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

To select Booster Version with special condition was used command **WHERE** (we specified type of landing and range of mass).

Total Number of Successful and Failure Mission Outcomes



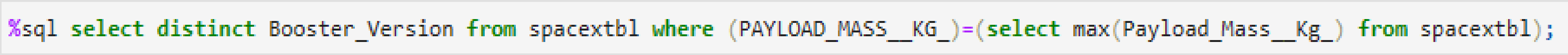
```
%sql select Mission_Outcome , count(Mission_Outcome) as Total_Number from spacextbl group by Mission_Outcome;
```



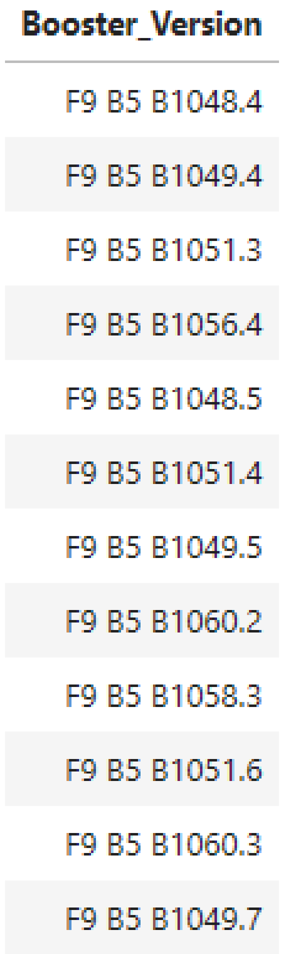
Mission_Outcome	Total_Number
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Function **COUNT** calculate total amount of Mission Outcomes and command **GROUP BY** arrange data into groups according to the types of Mission outcome.

Boosters Carried Maximum Payload



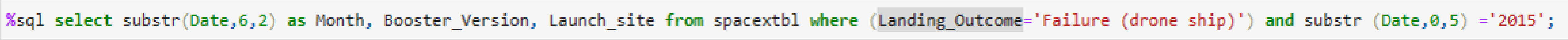
```
%sql select distinct Booster_Version from spacextbl where (PAYLOAD_MASS__KG_)=(select max(Payload_Mass__Kg_) from spacextbl);
```



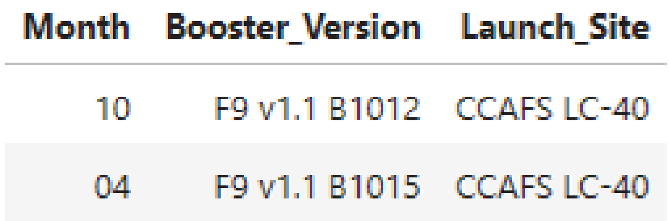
Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

For this solution was used subquery. First command **SELECT** (within the brackets) found the maximum payload – function **MAX**, then it was used in the **WHERE** condition. Also **DISTINCT** command was used to retrieve only unique Booster Versions.

2015 Launch Records



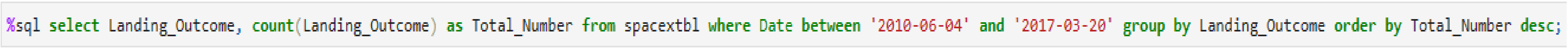
```
%sql select substr(Date,6,2) as Month, Booster_Version, Launch_site from spacextbl where (Landing_Outcome='Failure (drone ship)') and substr (Date,0,5) ='2015';
```



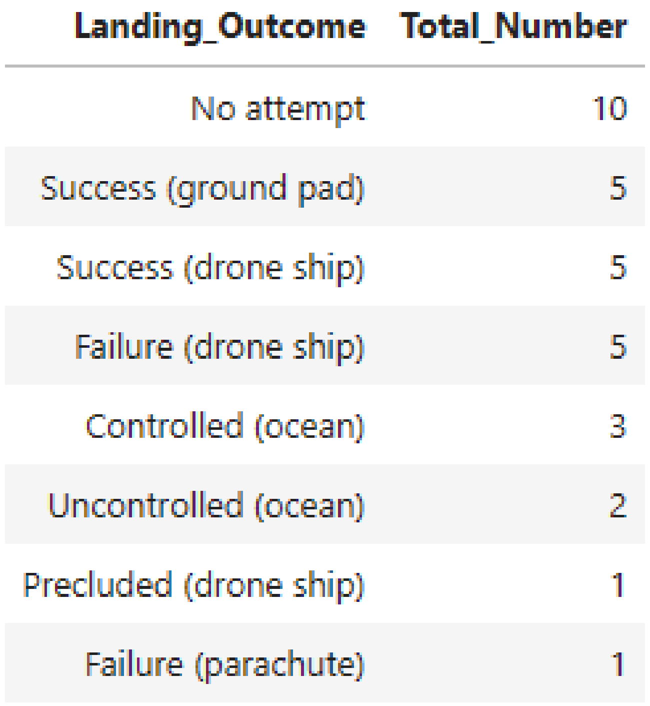
Month	Booster_Version	Launch_Site
10	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

Command **WHERE** filtered the results with 2 conditions: failed landing (drone ship), **AND** the year of 2015. **SUBSTR(DATE,6,2)** was used to get results by months.

Types of Landing Outcomes Between 2010-06-04 And 2017-03-20



```
%sql select Landing_Outcome, count(Landing_Outcome) as Total_Number from spacextbl where Date between '2010-06-04' and '2017-03-20' group by Landing_Outcome order by Total_Number desc;
```



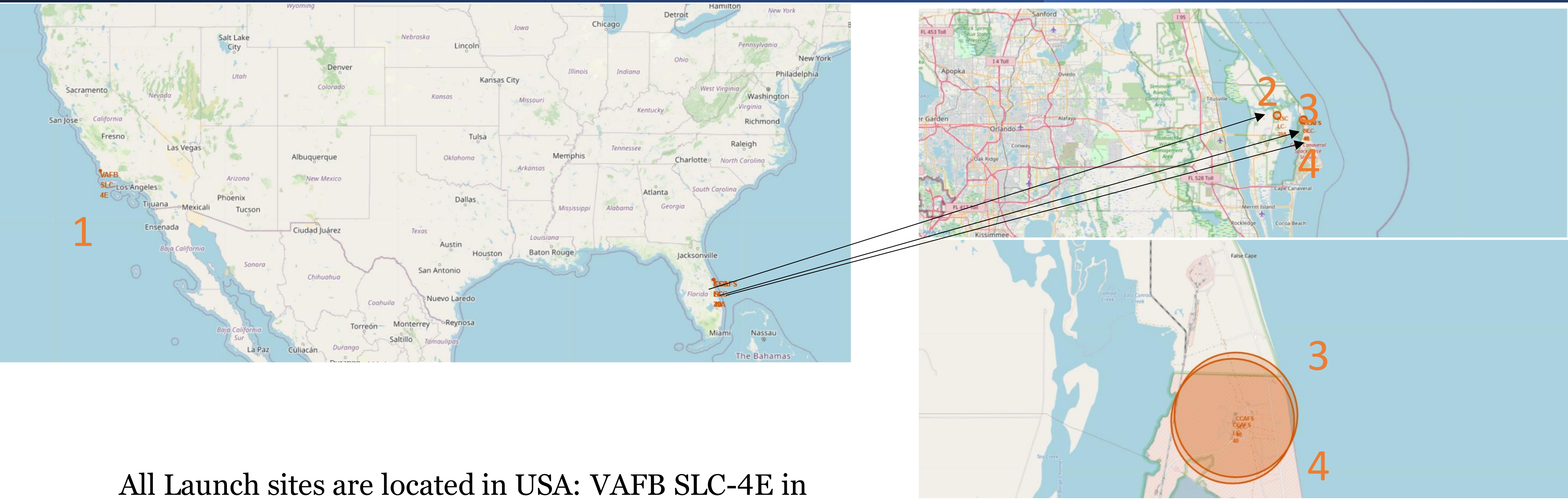
Landing_Outcome	Total_Number
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

Command **WHERE** was supplemented with the **BETWEEN** to filter the results within specified dates. Commands **GROUP BY** and **ORDER BY**, grouped and ordered results. **DESC** was used to specify the descending order.

Section 3

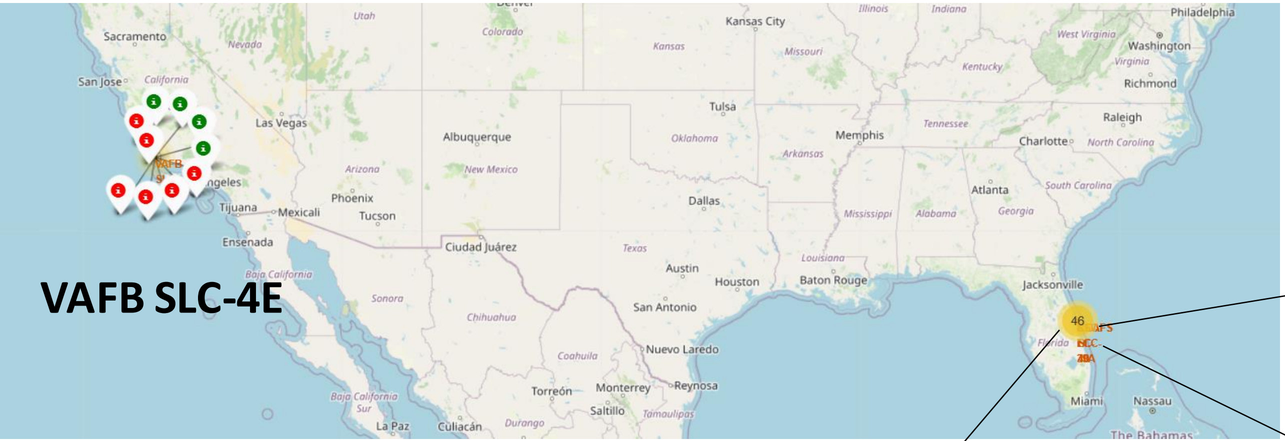
Launch Sites Proximities Analysis

Positions of all launch sites

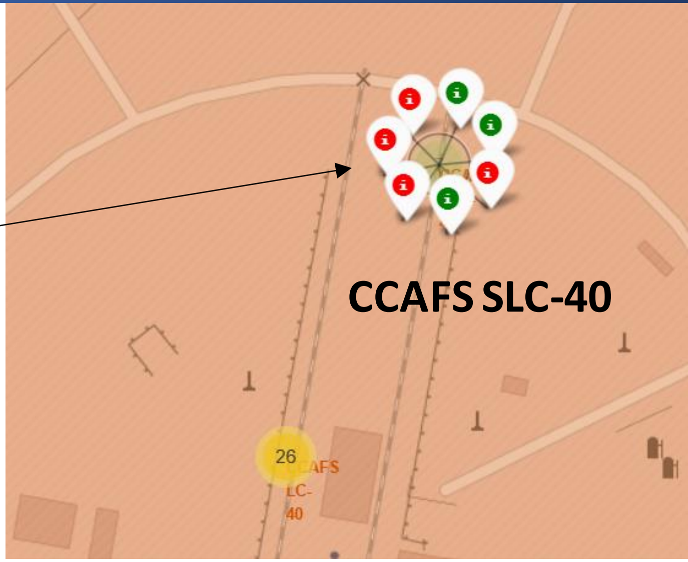


All Launch sites are located in USA: VAFB SLC-4E in California, three others in Florida.

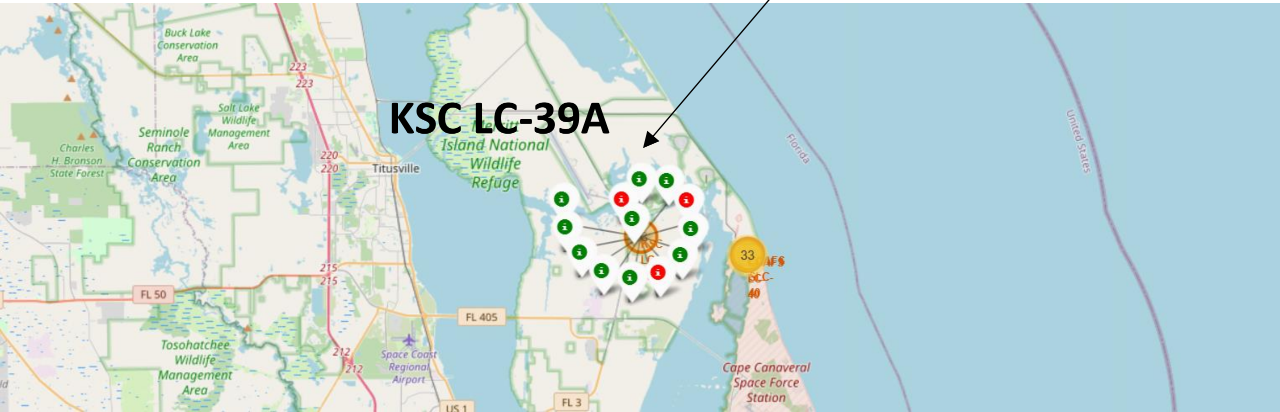
Failure and success marked for each Launch site



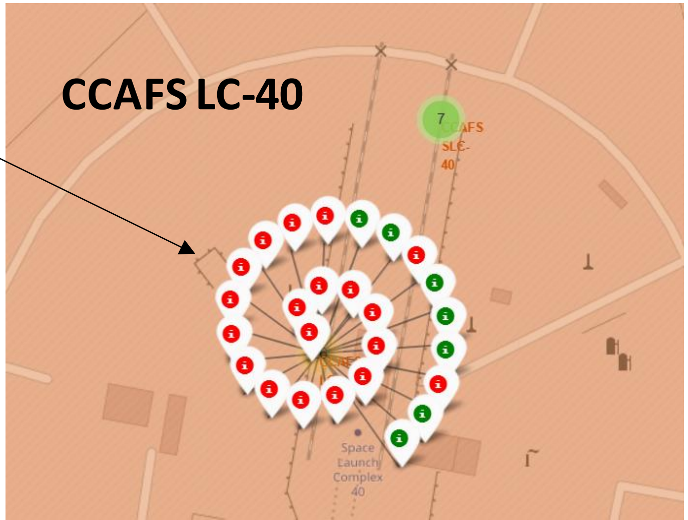
VAFB SLC-4E



CCAFS SLC-40



KSC LC-39A



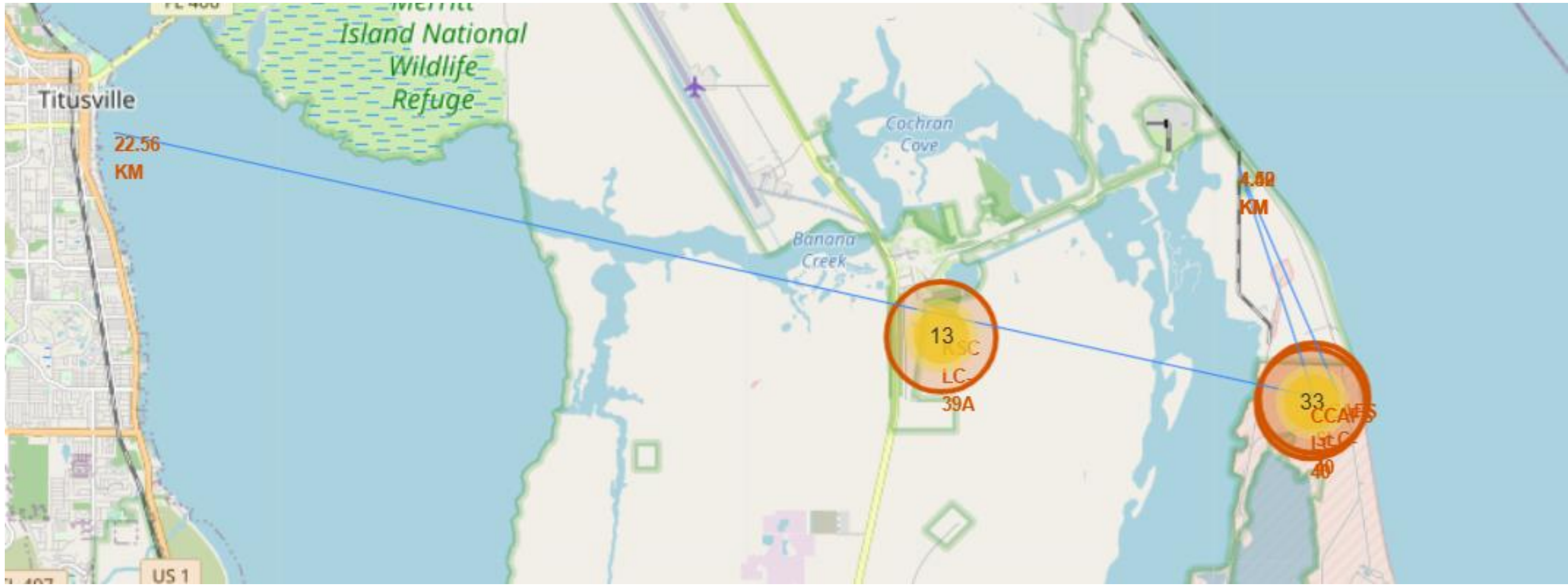
CCAFS LC-40

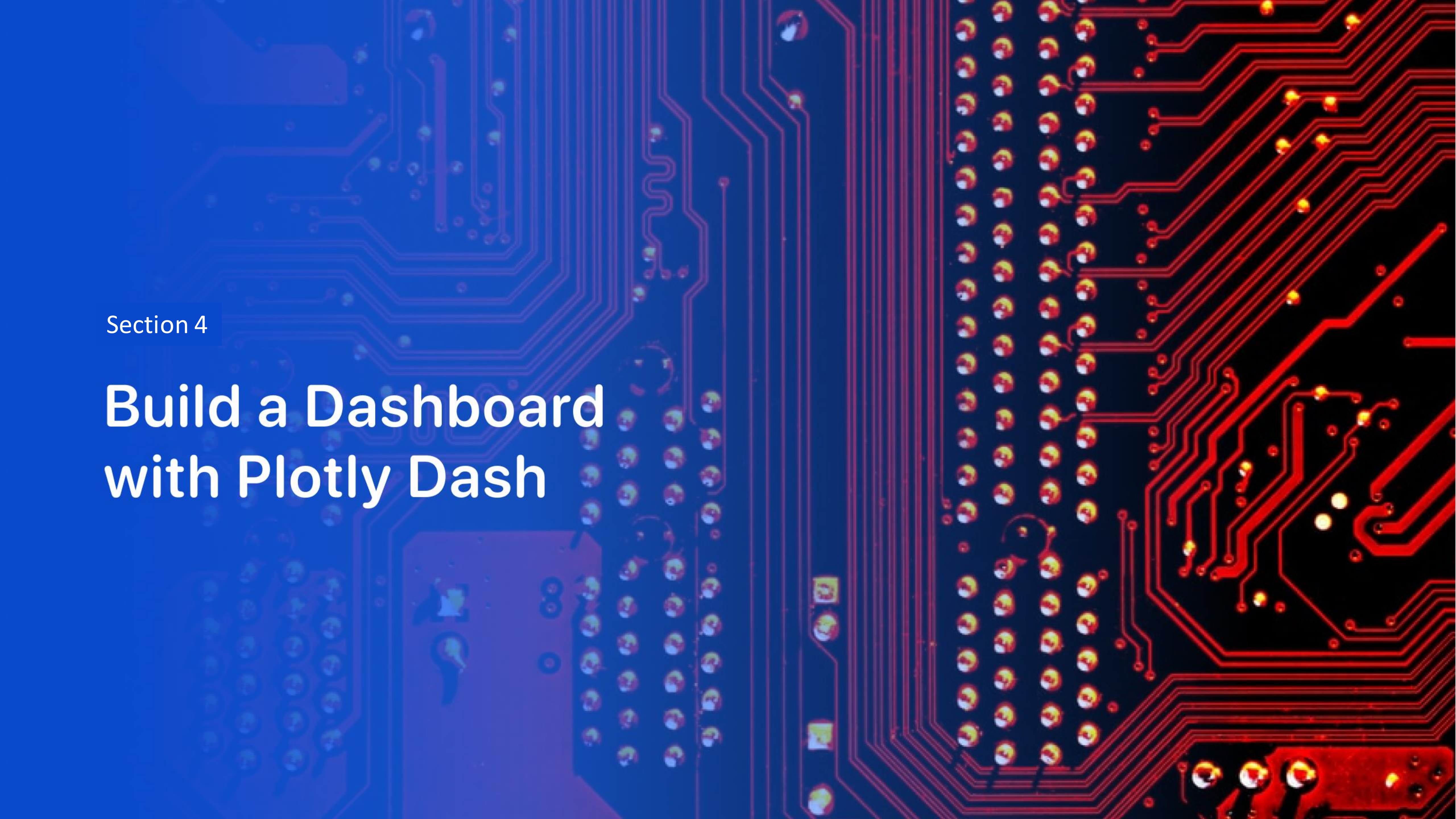
Green labels mean successful landing, red – failure.
Visually is clear that high level of success has KSC LC-39A.

Launch sites and its proximities

Using the CCAFS SLC-40 launch site as an example site, we can understand more about the placement of Launch Sites.

For the site CCAFS SLC-40 we calculated distance from the site to the point on the coast line and to Titusville.

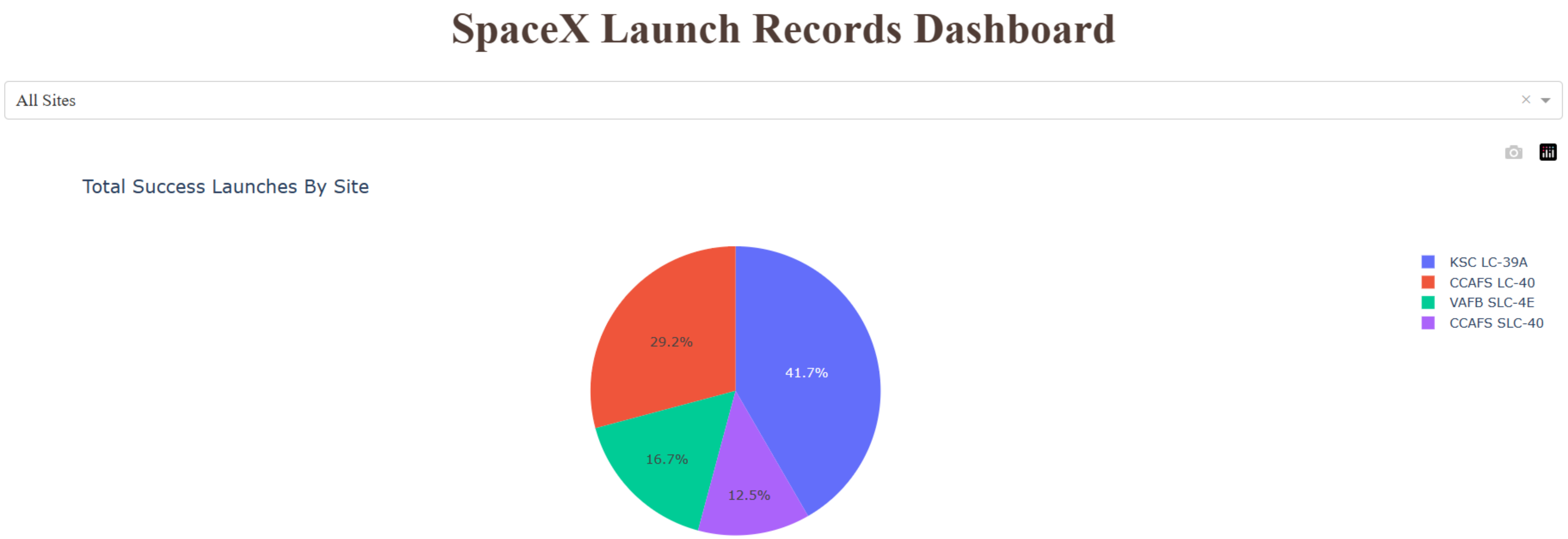




Section 4

Build a Dashboard with Plotly Dash

Successful launches count for all sites

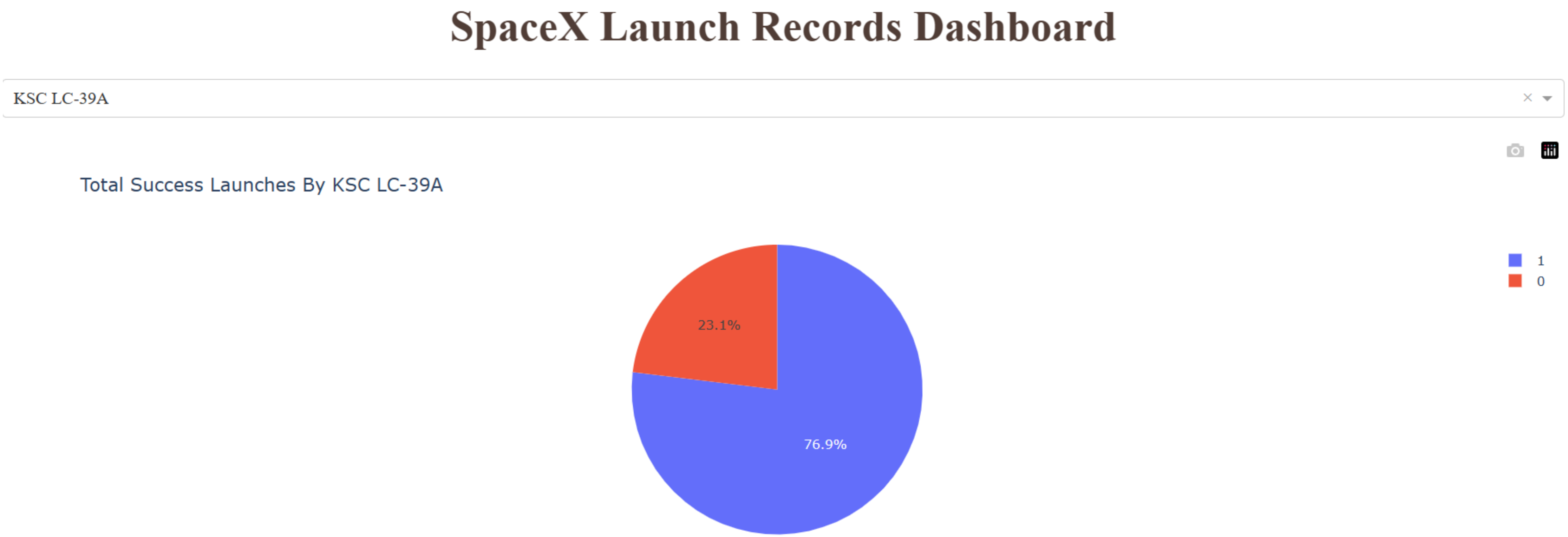


SpaceX Launch Records Dashboard

All Sites

From the pie chart we can see that the Launch Site KSL LC-39A has highest amount of successful launches – 41.7%.

Pie-chart for most successful site - **KSC LC-39A.**



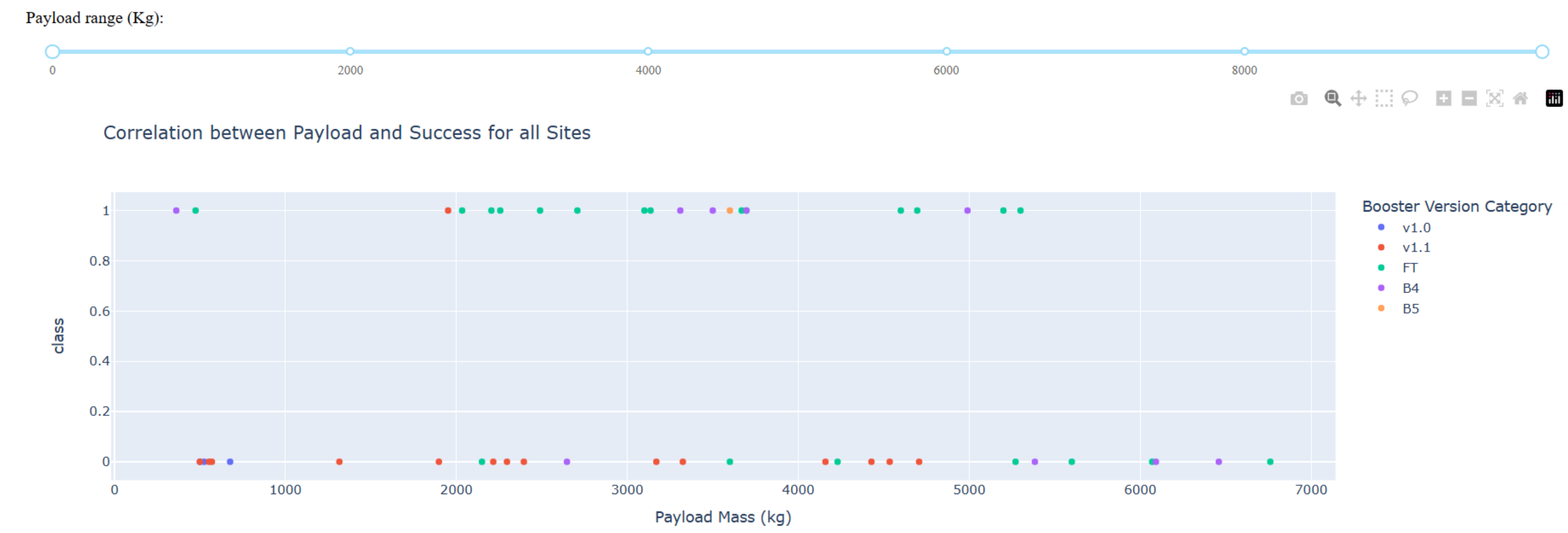
SpaceX Launch Records Dashboard

KSC LC-39A ×

Total Success Launches By KSC LC-39A

From this pie chart we noticed that ¾ from all the launches from this site are successful . It's highest success rate among all sites.

Payload vs. Launch Outcome scatter plot for all sites, with different payload selected

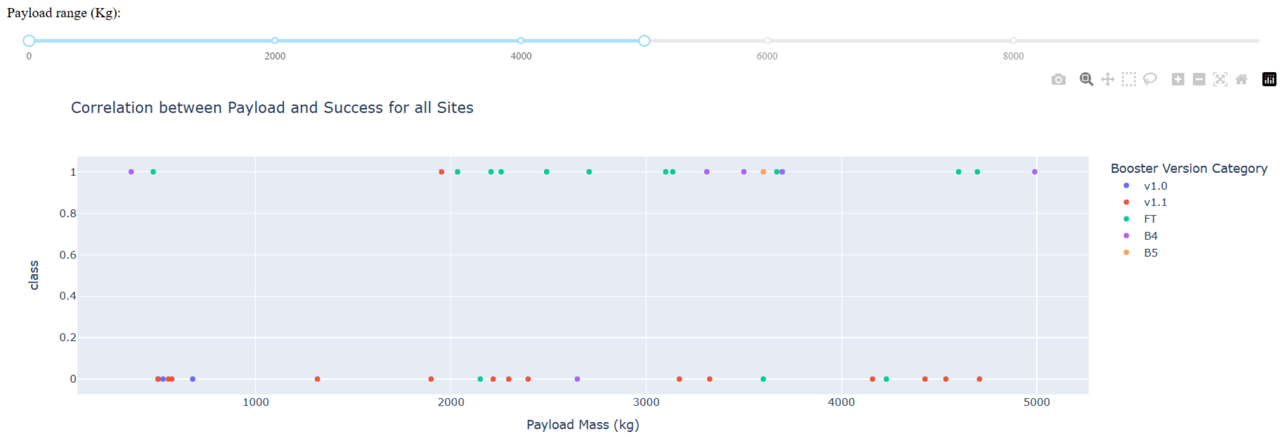


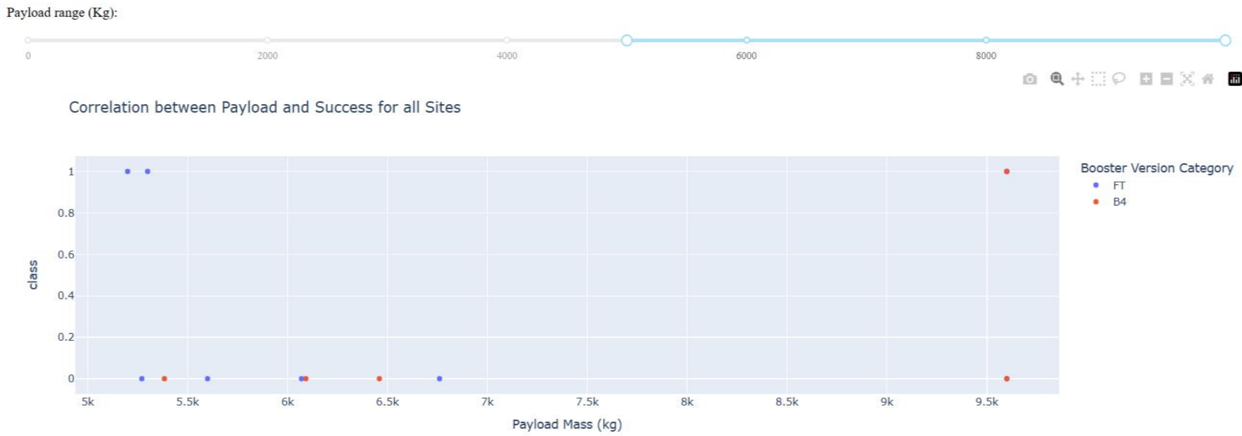
All Payload Mass range was divided into two parts. Between 0 to 5000 and 5000 to 10000.

For Payload Mas more than 5000 kg only two types of Booster Version were used: FT and B4.

Most of the successful launches were made with the Payload Mass between 2000 to 5000 kg.

Success rate for heavy launches is low.

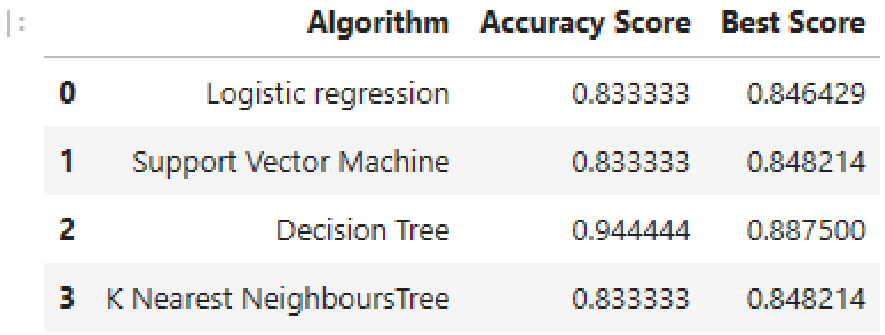




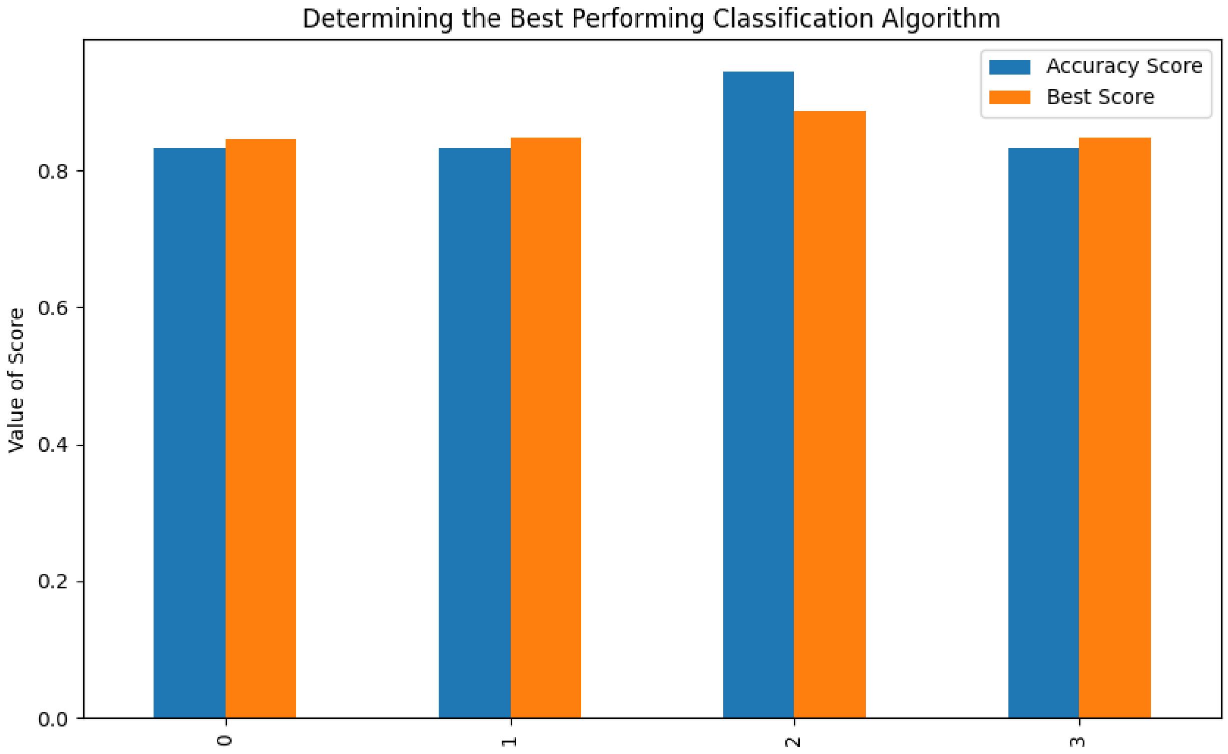
Section 5

Predictive Analysis (Classification)

Classification Accuracy



	Algorithm	Accuracy Score	Best Score
0	Logistic regression	0.833333	0.846429
1	Support Vector Machine	0.833333	0.848214
2	Decision Tree	0.944444	0.887500
3	K Nearest NeighboursTree	0.833333	0.848214

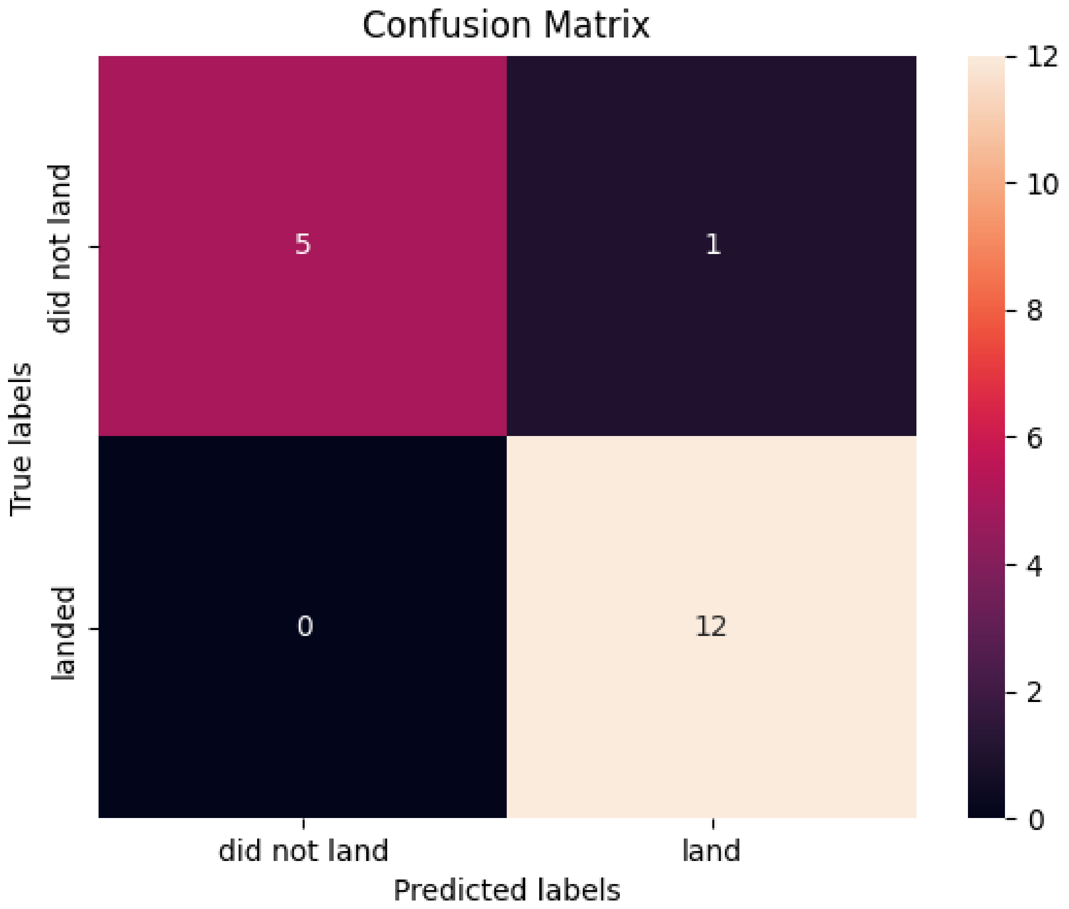


Determining the Best Performing Classification Algorithm

Pivot table and bar chart showed higher Accuracy score and Best score for Decision Tree model. That's mean it's the best model for our data.

Confusion Matrix for Decision Tree Model

- As was performed in the previous slide the best classification model is the Decision Tree model, with an accuracy of 94.44%.
- Confusion matrix showed that from 18 results only one was wrong. So we got 1 False Positive.
- The rest of the result were classified accurately: 5 True Negative (didn't land) and 12 True Positive (landed).



Conclusions

- With the time the success rating of the launches is increasing, that due to technology development and improvements done after failure.
- Orbit SSO has good tendency to be most successful with low payload mass.
- After 2015 success rate is reached 50% and kept going higher even with the drop in 2017-18 never came lower than 60%.
- Folium visualization showed that high level of success has KSC LC-39A Launch Site.
- Same result we got from Dash. KSL LC-39A has highest amount of successful launches – 41.7% among all sites. ¾ from all the launches from this site are successful.
- Most of the successful launches were made with the Payload Mass between 2000 to 5000 kg.
- Pivot table and bar chart showed higher Accuracy score and Best score for Decision Tree model, with an accuracy of 94.44%. Confusion matrix showed that from 18 results only one was wrong.

Appendix

[All Files of Capstone project GitHub]

THANK YOU