IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Name>
<Date>

Outline

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

Executive Summary

In this capstone, we will determine the success of landing for Spaces X's Falcon 9 first stage in order to determine the cost of a rocket launching. This will be achieved using machine learning algorithms and following Data Analysis methodologies:

- Data Collection from API and Web Scraping
- Data Wrangling and Preprocessing
- Exploratory Data Analysis
- Data Visualization
- Machine Learning Prediction Model

After the analysis, we found that there are some features and requirements have a correlation with the success or failure and first stage landing. In Summary, Decision Tree may be the best machine learning model to solve this problem.

Introduction

Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against space X for a rocket launch.

Based on the objective, we will find out what are the factors have the higher rate of successful landing for first stage.

Section 1

Methodology

Methodology

To solve the problem stated, data science methodology has been applied with following step:

- Business understanding & analytic – understand to background of the business, identify the problem and objective of the analysis. In this case, we decided to study on competitor SpaceX Falcon 9 to analyze the success rate of landing for first stage in order to determine the cost of launching.
- Data collection – data was collected from SpaceX API and web scraping from Wikipedia.
- Data Wrangling – to transform and clean the data using Python’s pandas.
- Perform exploratory data analysis (EDA) – using seaborn and matplotlib visualization and SQL to analysis the data.
- Perform interactive visual analytics – using Folium and Plotly dashboard for visualization
- Build machine learning model for predictive analysis – using four different classification method to find the best accuracy model for future prediction.

Data Collection

- First data was collected using SpaceX RESTful API by using get request method. This set of data included all the rocket boost version. We will only filter Falcon 9 boostversion in this case study. Relative information to analyze from the dataset are mass of payload, orbit, launch site and its location and others cores related factors.
- Secondly, we will extract data from Wikipedia using web scraping from HTML. We will get the launch record Wikipedia table. From the table, we will extract the related information such as payload, orbit, launch site, landing status and so on.

Data Collection – SpaceX API

- Data collected from SpaceX API RESTful using GET request.
- The response request data is JSON, turn JSON into dataframe using json_normalize
- Get the relevant information for analysis and filtering only Falcon 9 boostversion as our case study
- Perform data wrangling by analysis missing value data on PayloadMass column and replace with mean of Payload Mass
- Refer to below GitHub link :

Place your flowchart of SpaceX API calls here

Data Collection - Scraping

- Data was collected from Wikipedia using web scraping in HTML format. Apply BeautifulSoup libraries to extract launch record from table
- Extract all the column and turn to into dictionary and dataframe
- Refer to below GitHub link::

Place your flowchart of web scraping here

Data Wrangling

- After getting data, we will perform some EDA to find the pattern of data including:
  - Identify the number of launch site.
  - Calculate the number of occurrence of orbit
  - Convert the landing outcome into training label 1 and 0
- Result: The success rate of Falcon 9 landing is 67%
- Refer to below GitHub link:

EDA with Data Visualization

- Using Pandas and Matplotlib to perform several data visualization
  - Scatter plots to find the relationship between factors such as flight number vs launch site, payload vs launch site, flight vs orbit type and payload vs orbit type.
  - Bar chart to visualize the relationship between success rate of each orbit type.
  - Line plot to visualize the trend of success rate over the years.
- Refer to URL:

EDA with SQL

- The following SQL queries were performed for EDA:
  - Find out the unique launch sites name
  - Find out 5 records of launch sites begin with “CCA”
  - Calculate to total payload mass carried launched by customer NASA (CRS)
  - Calculation the average payload mass carried by Falcon 9 v1.1
  - Find the date of first successful landing outcome
  - Find the total number of successful and failure mission outcome.
- Refer to below GitHib URL:

Build an Interactive Map with Folium

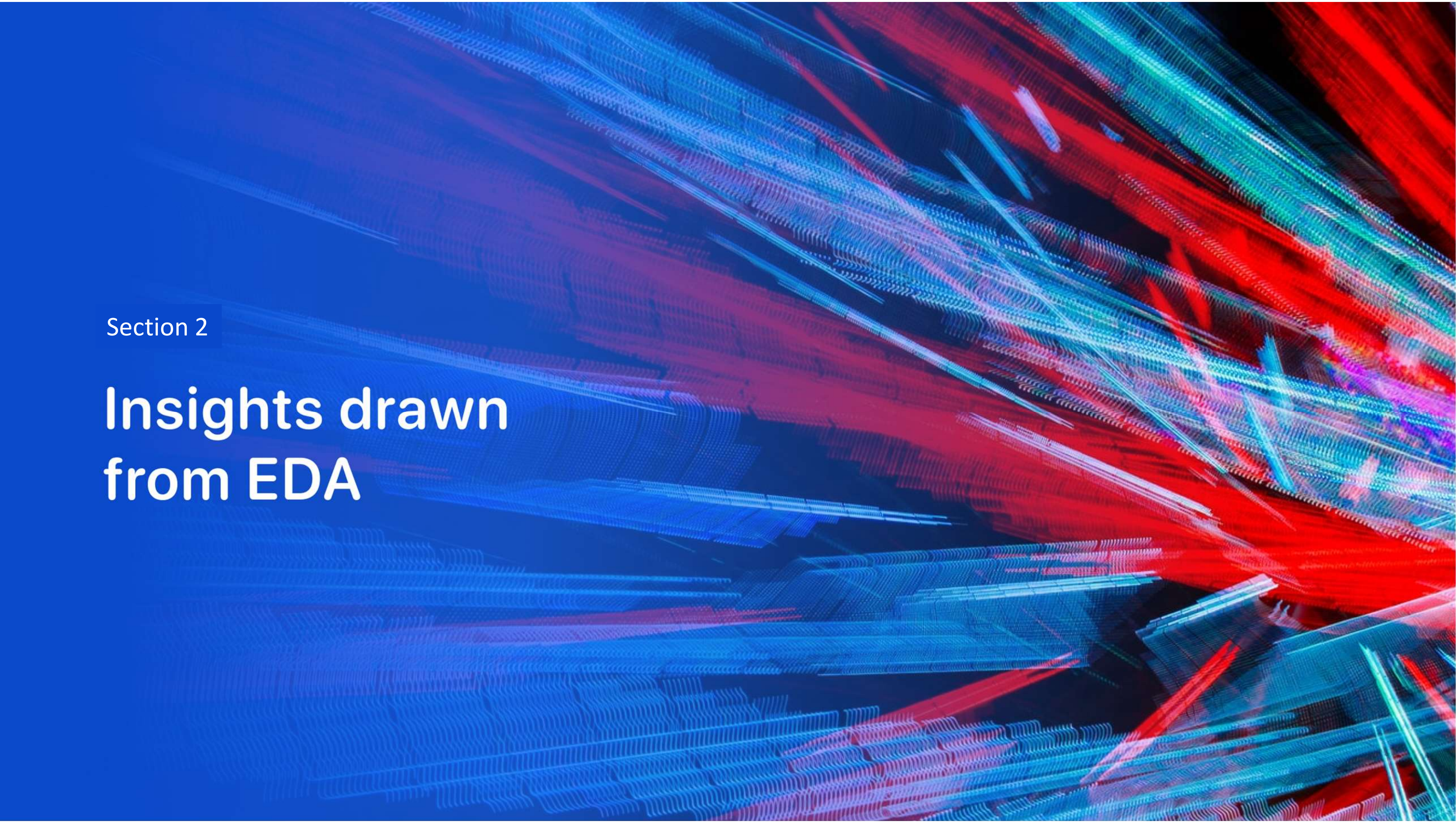
- Marker objects are the show location of all launch sites and the success and failure of the launching of each sites.
- Besides, line objects are used to identify the distance between site and its proximities
    - Calculate distance between launch site and nearest railway.
    - Calculate distance between launch site and nearest highway.
    - Calculate distance between launch site and coastline.
- Refer to Github link:

Build a Dashboard with Plotly Dash

- Following Interactive dashboard has been built using Ploty Dash:
  - Pie chart – to show the distribution of success launching for each sites.
  - Scatter chart – to show the relationship between landing outcome and payload mass of different boosters.
- Refer to GitHub link

Predictive Analysis (Classification)

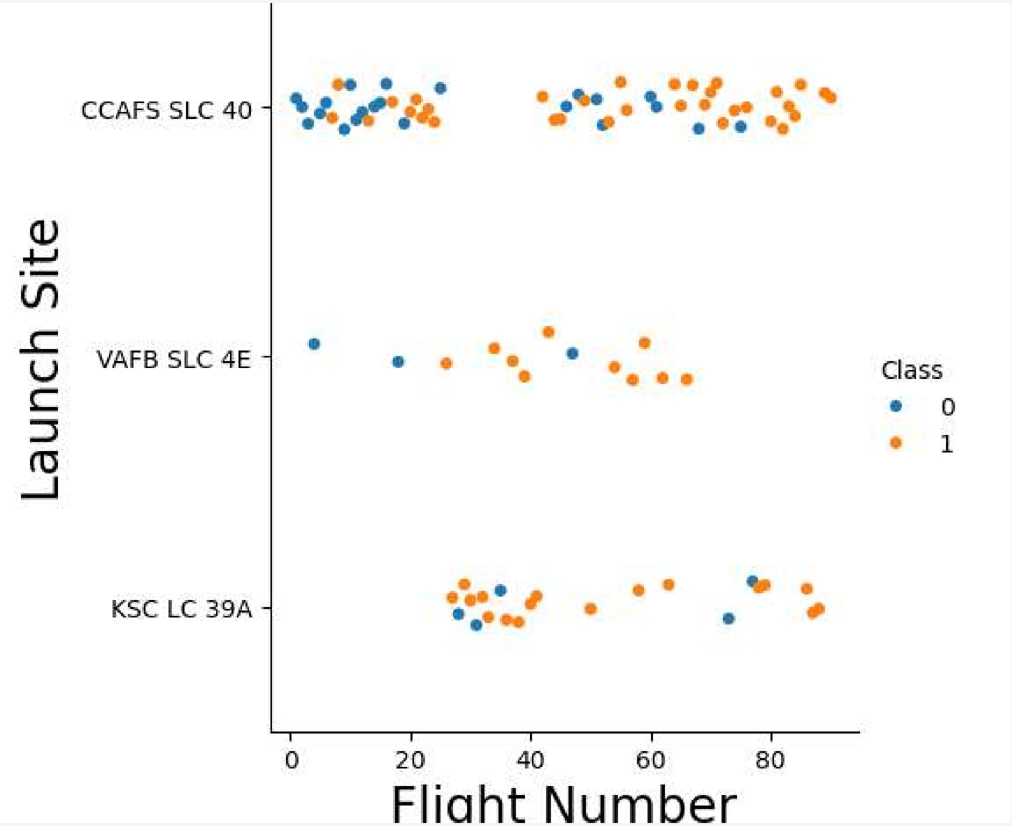
- To determine the best performing classification model between Logistic Regression, SVM, Classification Tree and K Nearest Neighbors:
  - Split the existing data into training and testing set to test on each models.
  - Fit the training model and predict the result of testing.
  - Using confusion matrix to evaluate the accuracy of predicted outcome and actual for each model.
- Refer to GitHub link:



Section 2

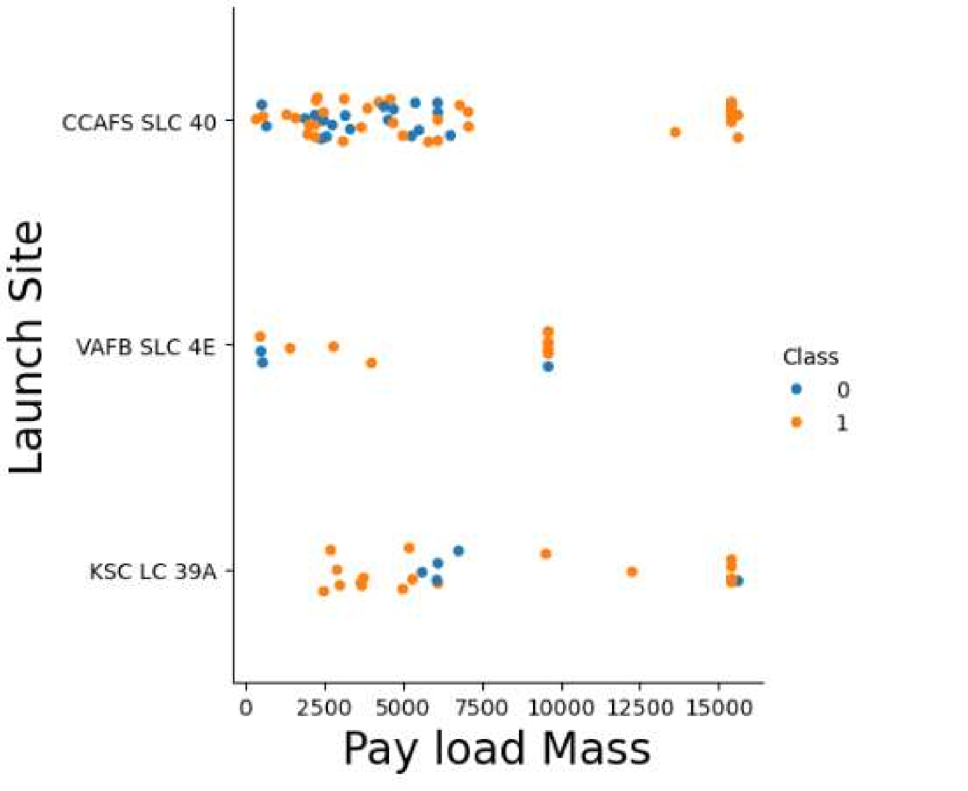
Insights drawn from EDA

Flight Number vs. Launch Site



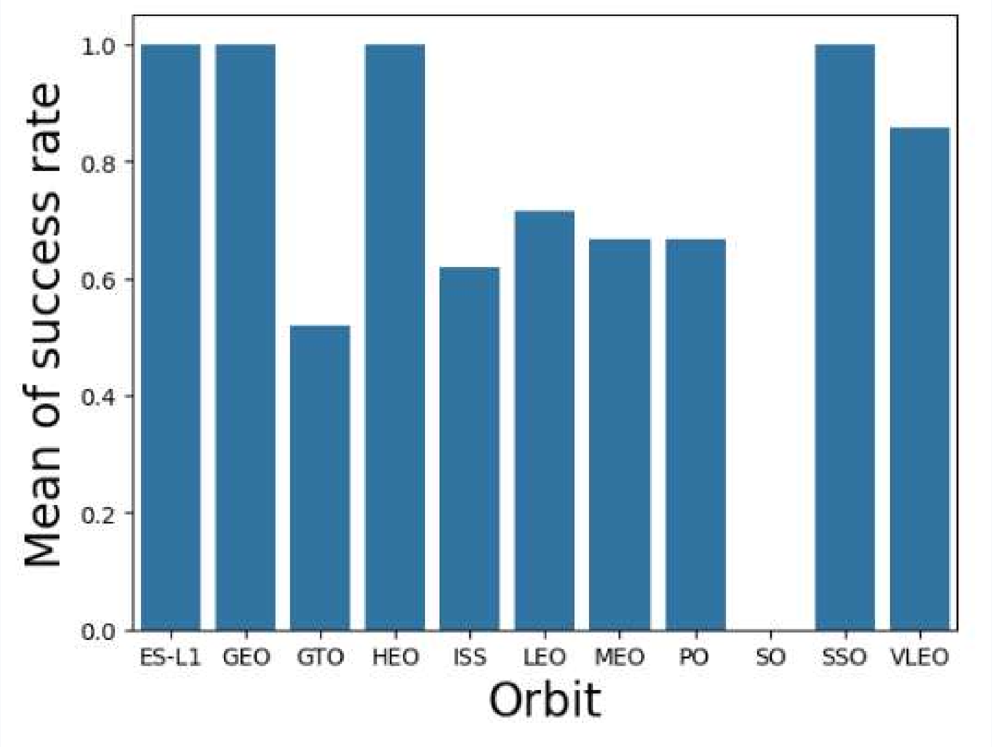
- CCAFS SLC 40 launch site have the most flight number.
- Relationship has been determined for launch site CCAFS SLC 40. The more the flight number the higher the success rate of landing

Payload vs. Launch Site



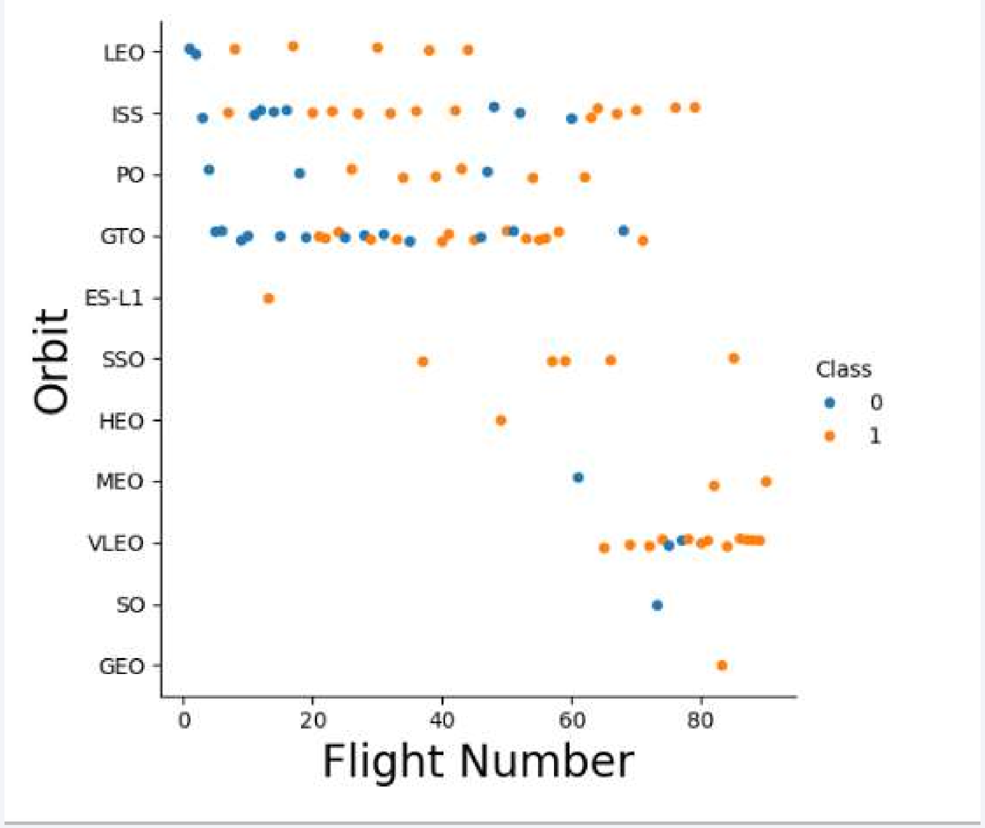
- Launch Site VAFB SLC 4E doesn’t have heavyload launching (more than 10000)
- There is no pattern, meaning no relationship between payload and launch site
- A weak insight can be found on site CCAFS SLC 40 where it has 100% success landing rate for mass 15000.

Success Rate vs. Orbit Type



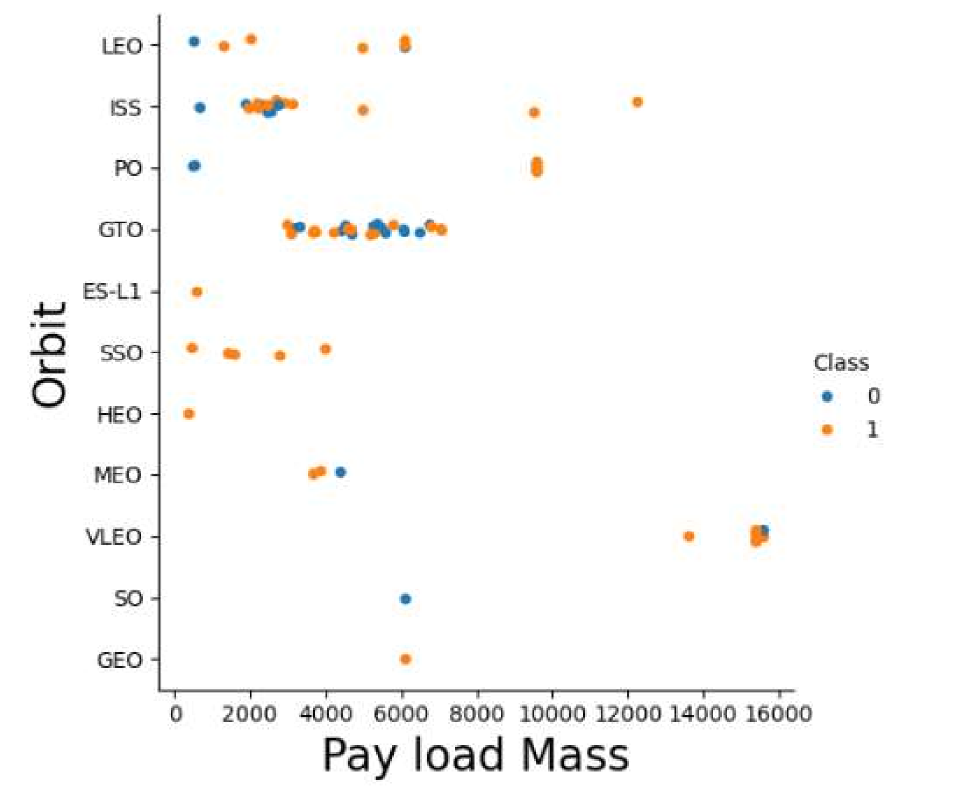
- Orbits including ES-L1, GEO and SSO have 100% of success rate.
- The lowest success rate is orbit GTO.

Flight Number vs. Orbit Type



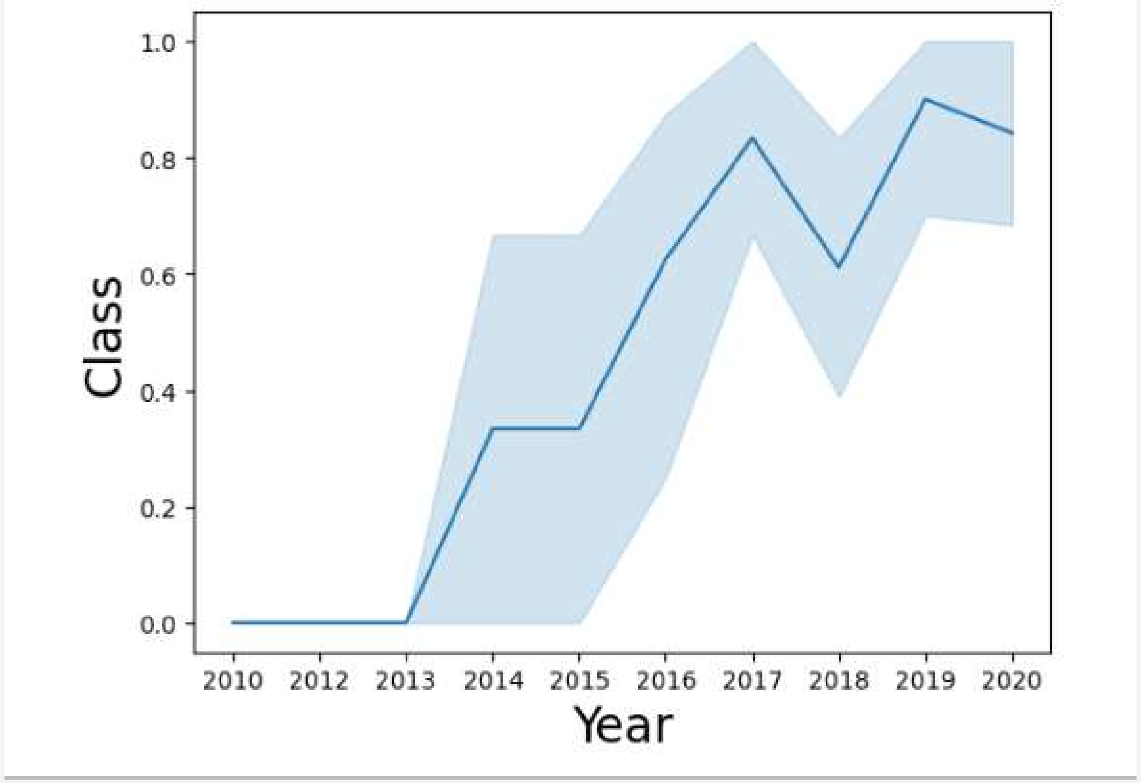
- There seem no relationship on GTO site.
- The highest success rate of landing is SSO orbit which is 100% for all 5 launching. Although ES-L1 and GEO also have 100% success but both only launch once.

Payload vs. Orbit Type



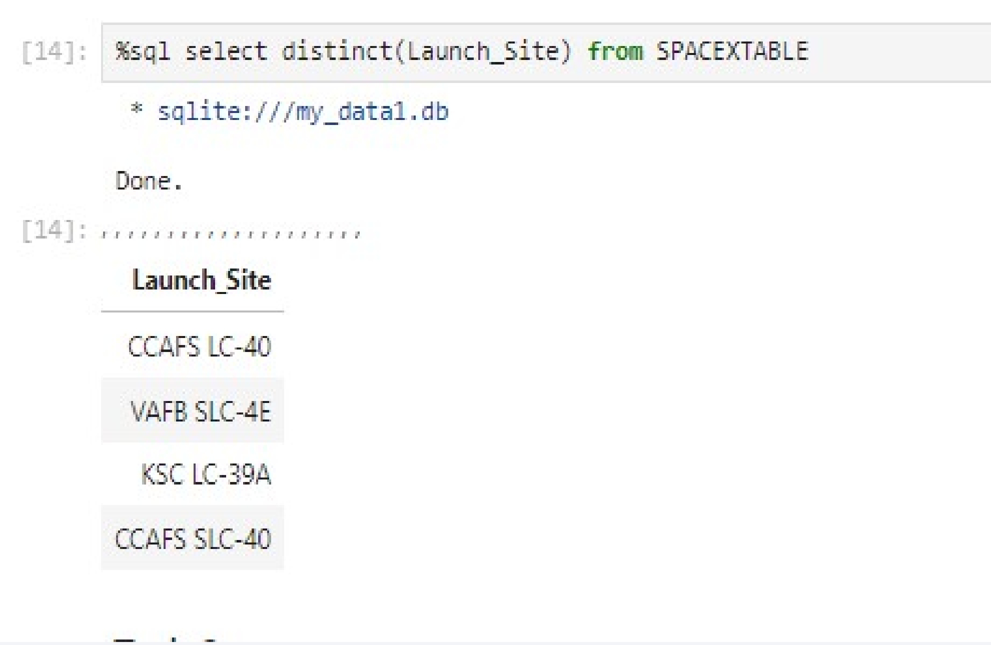
- Pattern can be found which the higher the load mass the higher the success rate on orbit LEO, ISS, VLEO
- However, above hypothesis doesn’t apply to GTO which there is no relationship detected.

Launch Success Yearly Trend



- Average success rate was increase since 2013 until 2020.
- This may because Space X learn the factors of success landing throughout the number of year and amount of number launching.

All Launch Site Names

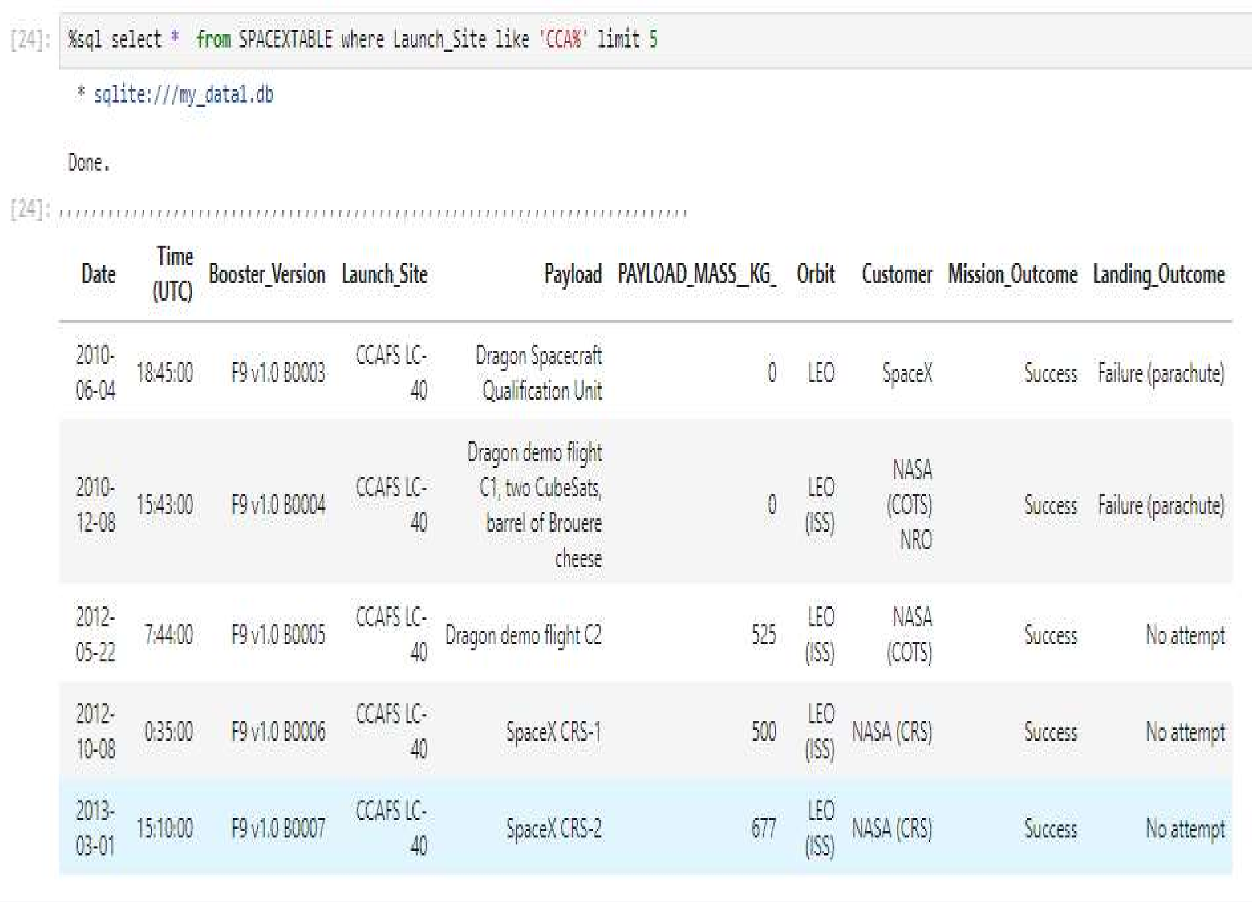


```
[14]: %sql select distinct(Launch_Site) from SPACEXTABLE

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

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

- There are 4 launch site which are
  - CCAFS LC-40
  - VAFB SLC-4E
  - KSC LC-39A
  - CCAFS SLC-40
- Using SQL distinct function to filter unique value for launch site

Launch Site Names Begin with 'CCA'



```
[24]: %sql select * from SPACEXTABLE where Launch_Site like 'CCA%' limit 5
```

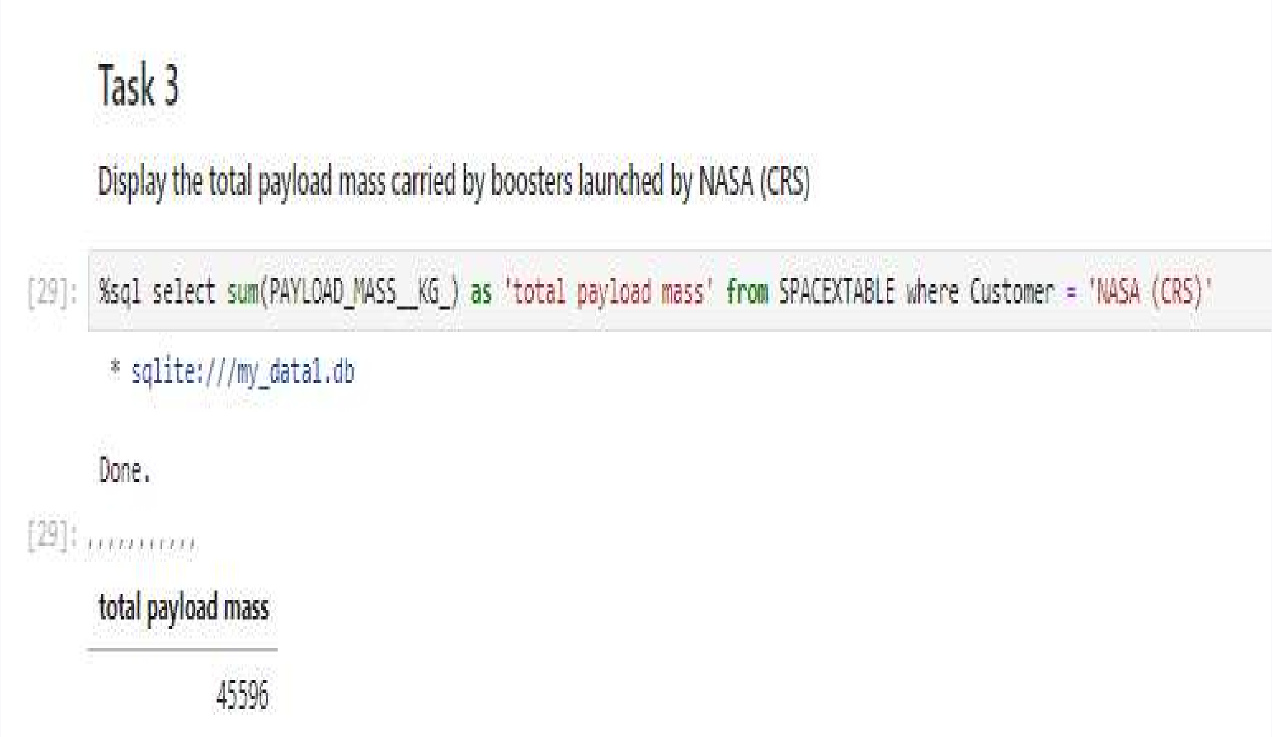
```
 * sqlite:///my_data1.db
Done.
```

[24]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
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 attempt
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

- Using SQL LIKE function to filter launch site name begin with CCA and limited to 5 rows.

Total Payload Mass



Task 3

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

```
[29]: %sql select sum(PAYLOAD_MASS__KG_) as 'total payload mass' from SPACEXTABLE where Customer = 'NASA (CRS)'
```

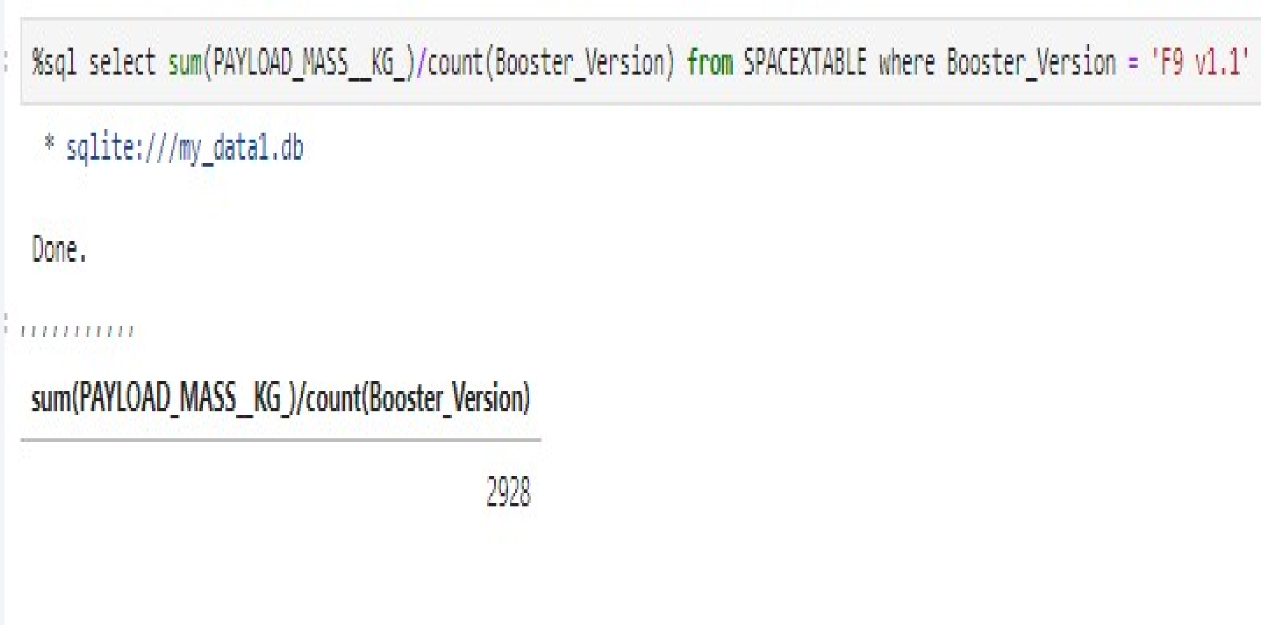
```
 * sqlite:///my_data1.db
Done.
```

[29]:

total payload mass
45596

- Using SQL to sum up payload mass in kg and filter customer which is NASA.
- Total payload mass is 45596 kg.

Average Payload Mass by F9 v1.1



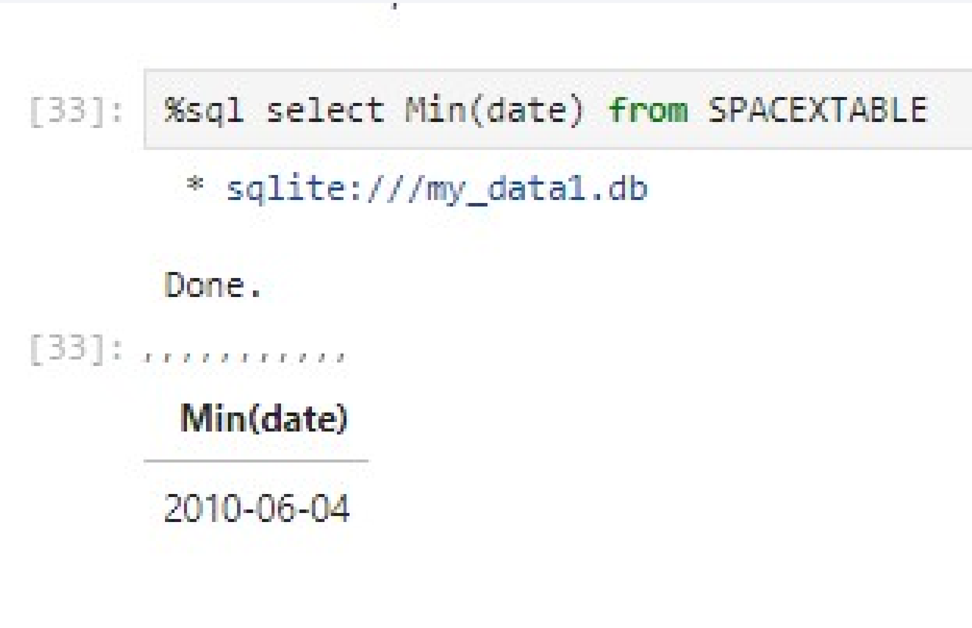
```
%sql select sum(PAYLOAD_MASS__KG_)/count(Booster_Version) from SPACEXTABLE where Booster_Version = 'F9 v1.1'
```

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

sum(PAYLOAD_MASS__KG_)/count(Booster_Version)
2928

- Using SQL to calculate the average payload mass of F9 v1.1 booster (payload mass / count of booster).
- The average payload mass by F9 v1.1 is 2928 kg.

First Successful Ground Landing Date

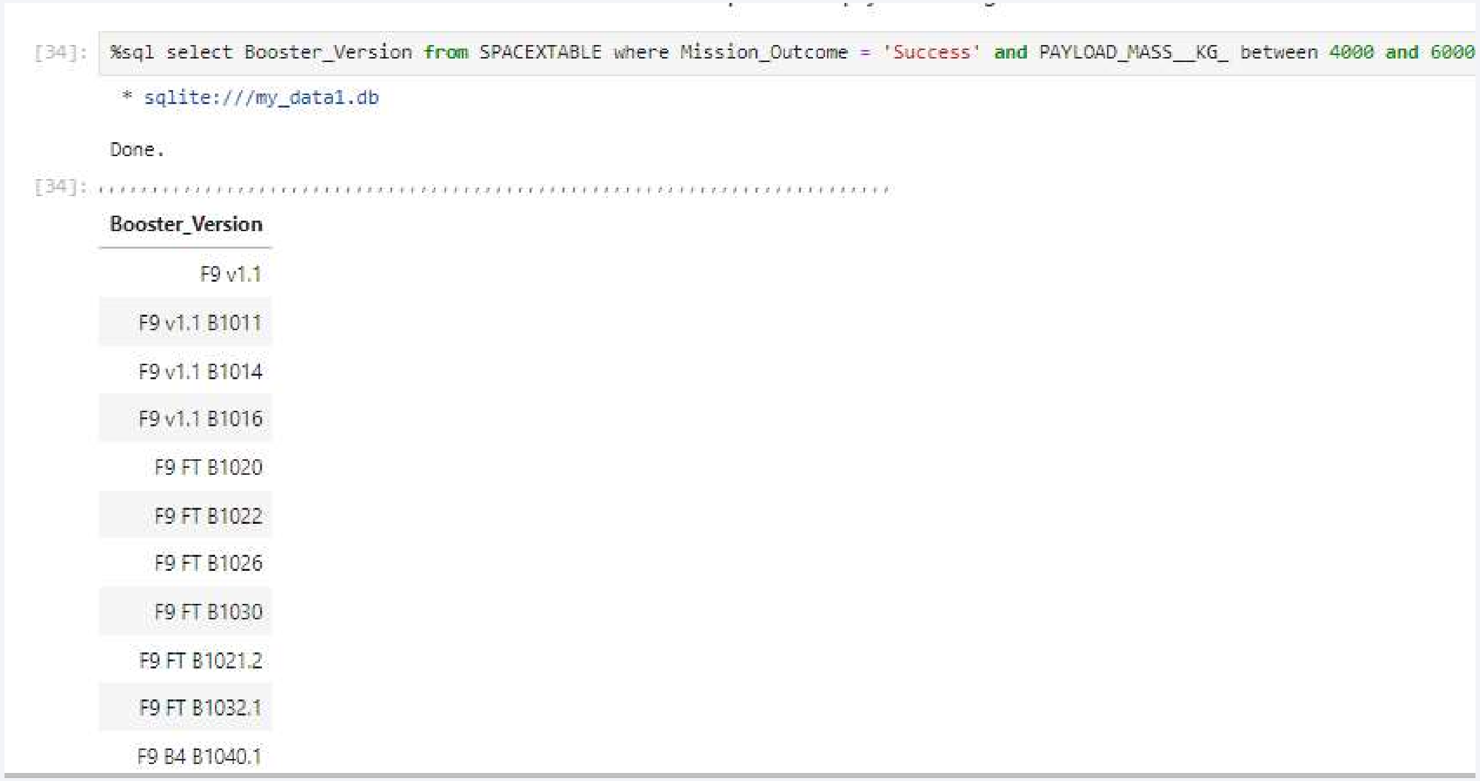


```
[33]: %sql select Min(date) from SPACEXTABLE

 * sqlite:///my_data1.db

Done.
[33]:
```

Min(date)
2010-06-04

- Using SQL to filter to earlier date using Min function
- First successful ground landing date is 04/06/2010.

Successful Drone Ship Landing with Payload between 4000 and 6000



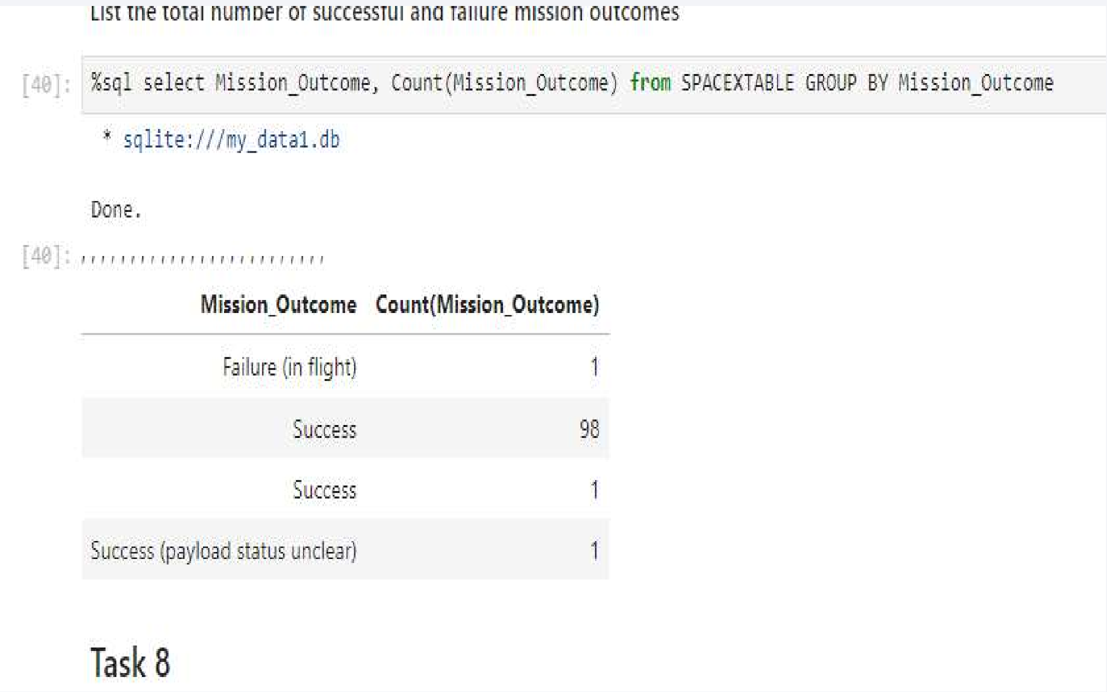
```
[34]: %sql select Booster_Version from SPACEXTABLE where Mission_Outcome = 'Success' and PAYLOAD_MASS__KG_ between 4000 and 6000

       * sqlite:///my_data1.db

      Done.
```

[34]:

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

- Using SQL to filter success outcome and payload mass is between 4000 and 6000.

Total Number of Successful and Failure Mission Outcomes



List the total number of successful and failure mission outcomes

```
[40]: %sql select Mission_Outcome, Count(Mission_Outcome) from SPACEXTABLE GROUP BY Mission_Outcome

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

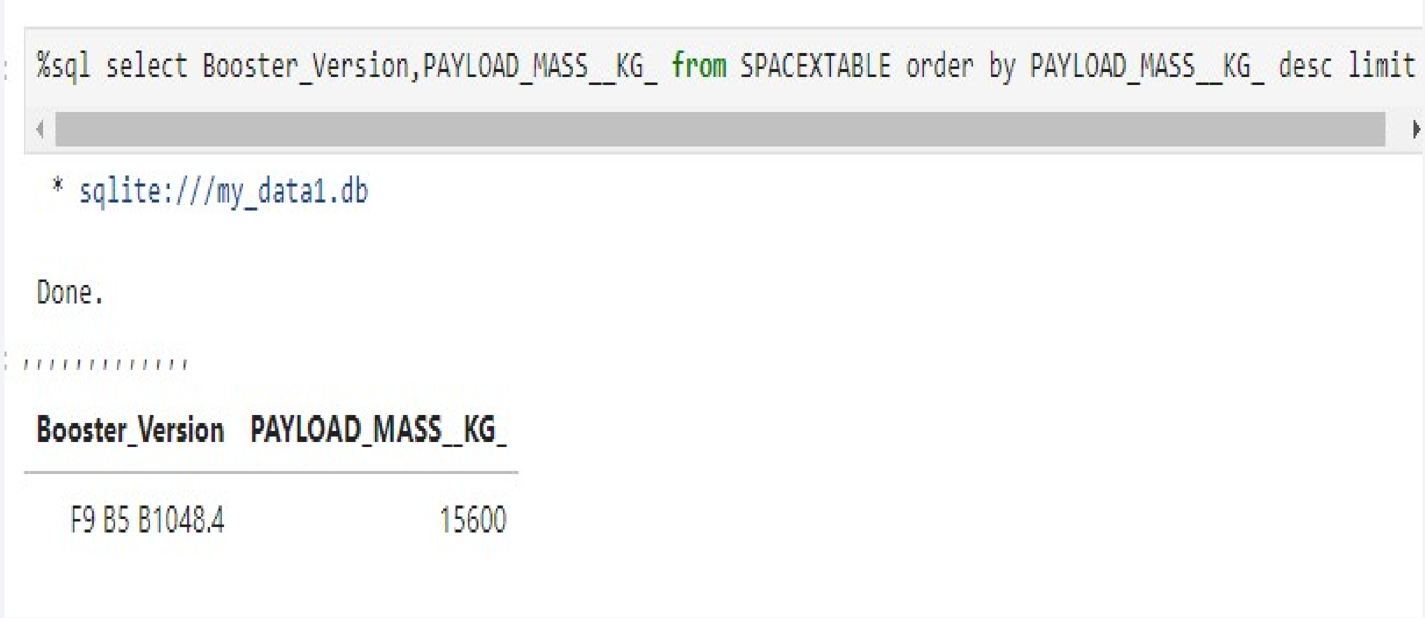
[40]:

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

Task 8

- Using SQL to group mission outcome and count the record of each unique value.
- From the result, 100 out of 101 consider success. Only One failure for mission from mission outcome itself.

Boosters Carried Maximum Payload

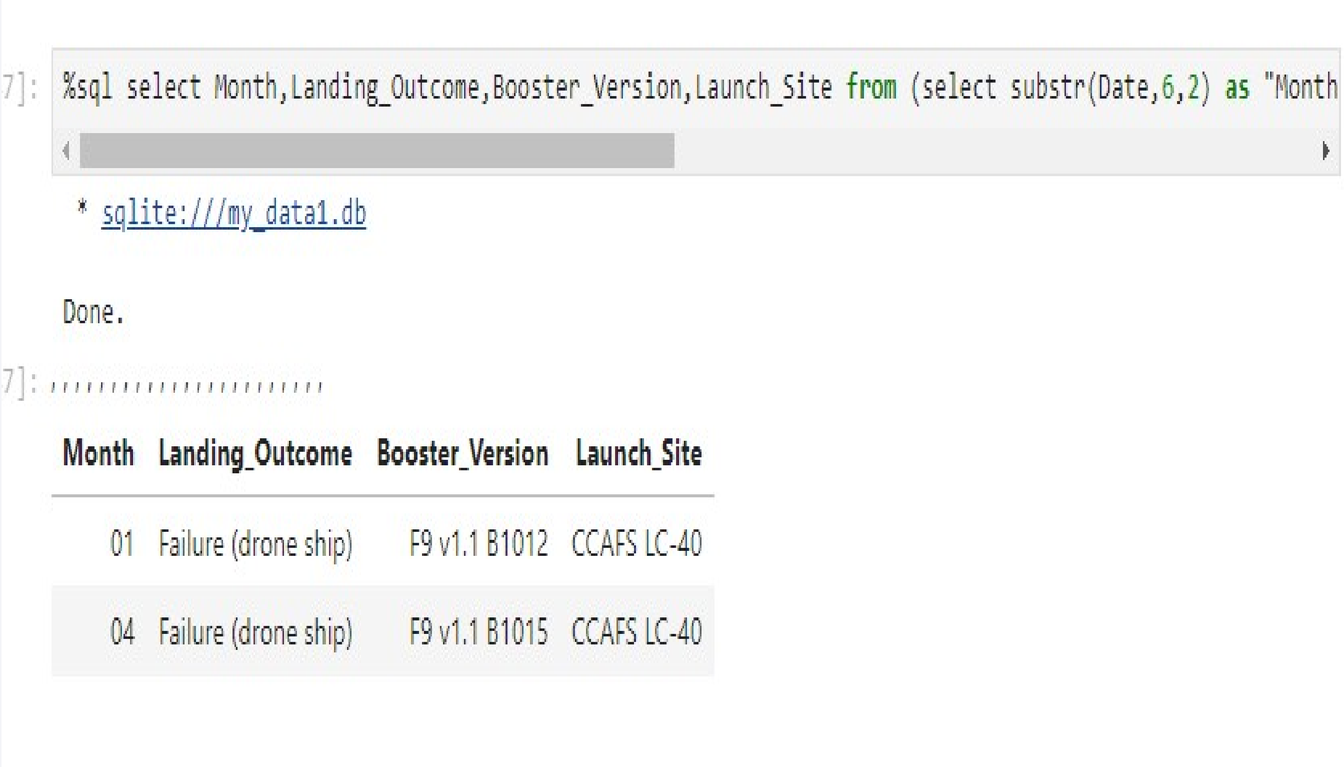


```
%sql select Booster_Version,PAYLOAD_MASS__KG_ from SPACEXTABLE order by PAYLOAD_MASS__KG_ desc limit
```

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

Done.

Booster_Version   PAYLOAD_MASS__KG_
   F9 B5 B1048.4              15600
```

- Using SQL to find booster version carried max payload mass by descending order and limit 1.
- Booster carried the maximum payload is F9 B5 B1048.4 and the max payload is 15600 kg.

2015 Launch Records



```
7]: %sql select Month,Landing_Outcome,Booster_Version,Launch_Site from (select substr(Date,6,2) as "Month
```

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

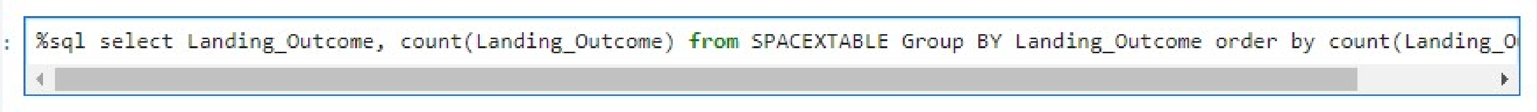
```
7]: ,,,,,,,,,,,,,,,,,,,,,,,
```

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

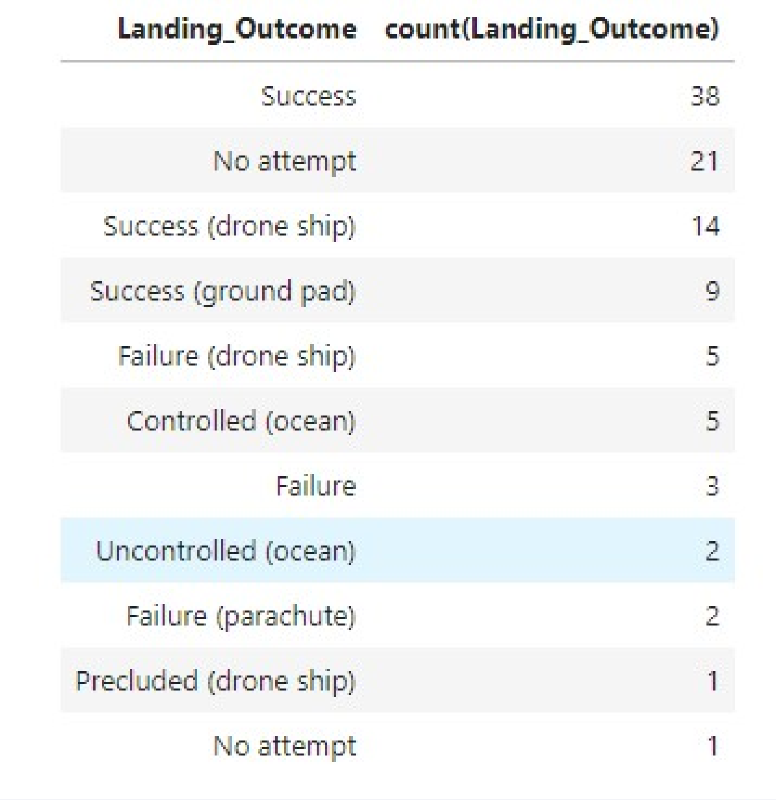
- Using SQL retrieve booster version and launch site by filtering date 2015 and landing outcome is failure.

- As result. There are 2 failure landing in 2015 both come from CCAFS LC-40 and booster versions are F9v1.1. B1012 and .F9v1.1. B1015

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



```
%sql select Landing_Outcome, count(Landing_Outcome) from SPACEXTABLE Group BY Landing_Outcome order by count(Landing_O
```



Landing_Outcome	count(Landing_Outcome)
Success	38
No attempt	21
Success (drone ship)	14
Success (ground pad)	9
Failure (drone ship)	5
Controlled (ocean)	5
Failure	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1
No attempt	1

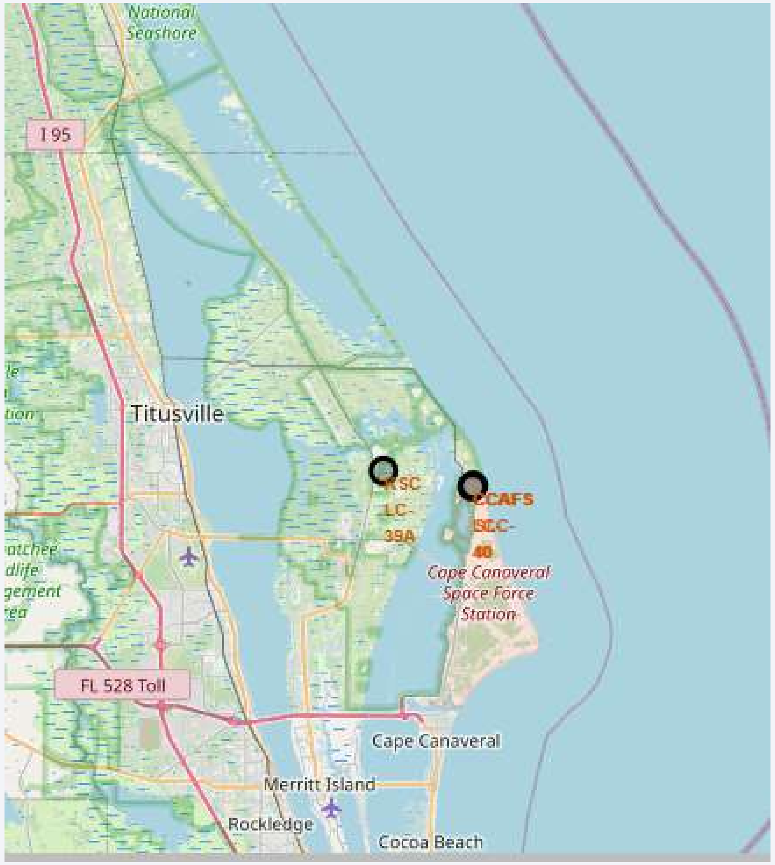
- Using SQL to group landing outcomes and count between the date 2010-06-04 and 2017-03-20, in descending order ranking

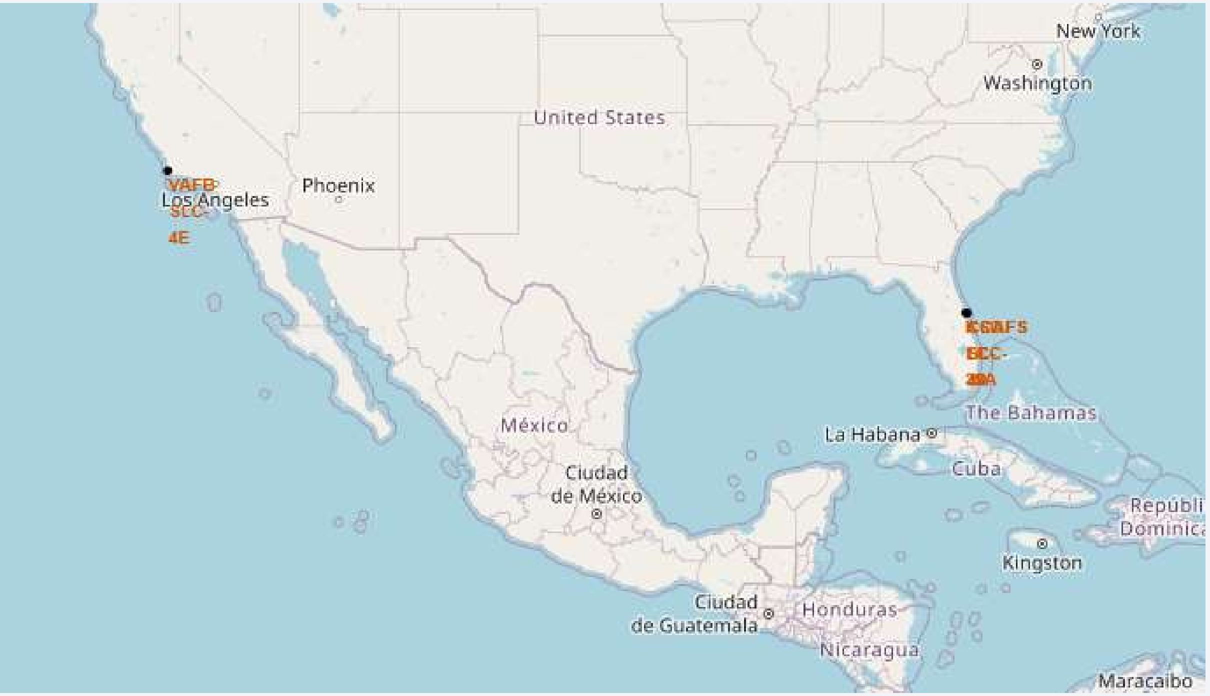
- Top 1 Landing outcome in between the period is “Success” follow by “No attempt”.

Section 3

Launch Sites Proximities Analysis

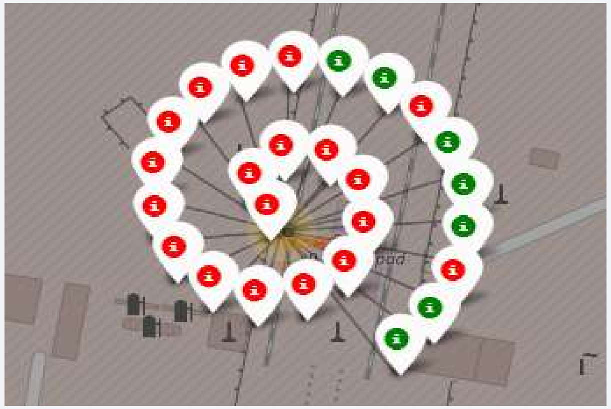
Folium Map Launch Site



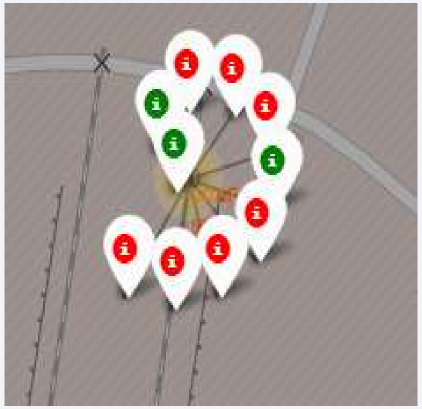


- There are 4 launch sites, 1 in Los Angeles and other 3 sites are located nearly together in Titusville.

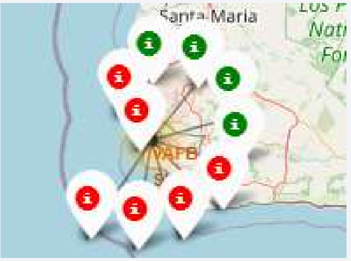
Folium Map with Label of Outcomes



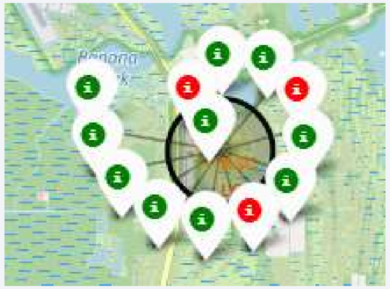
CCAFS LC-40



CCAFS SLC-40



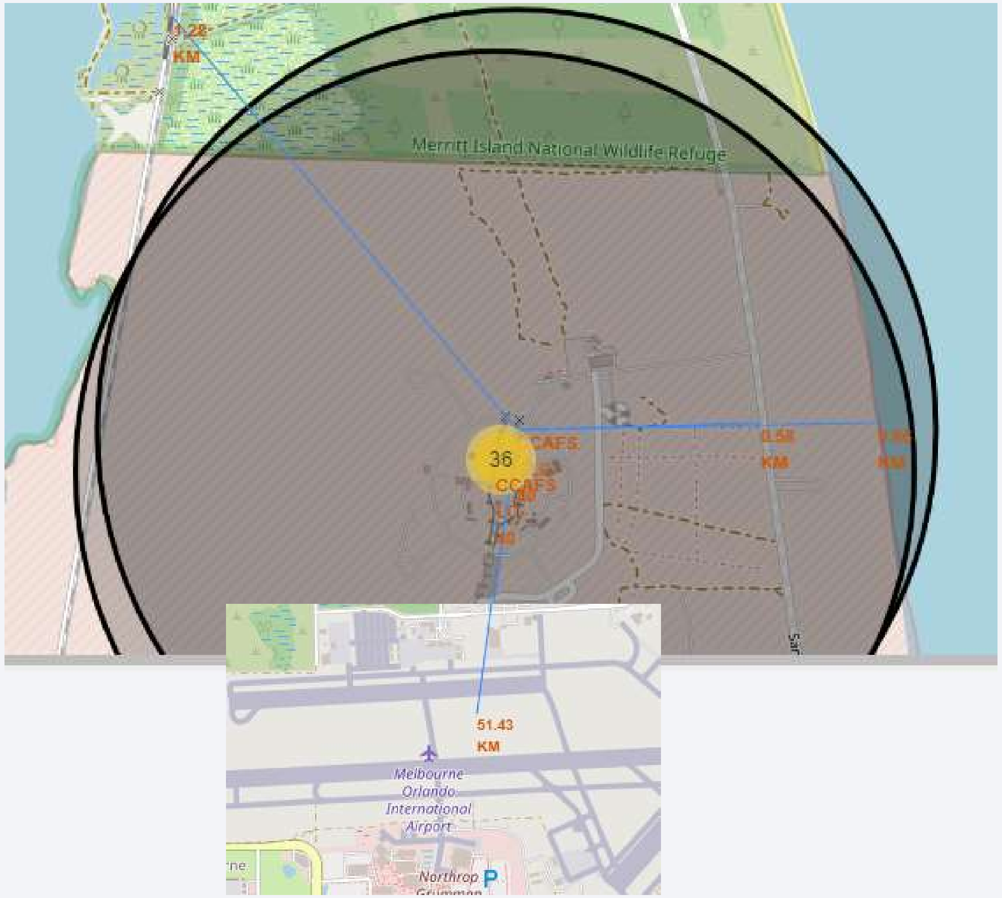
VAFB SLC-4E



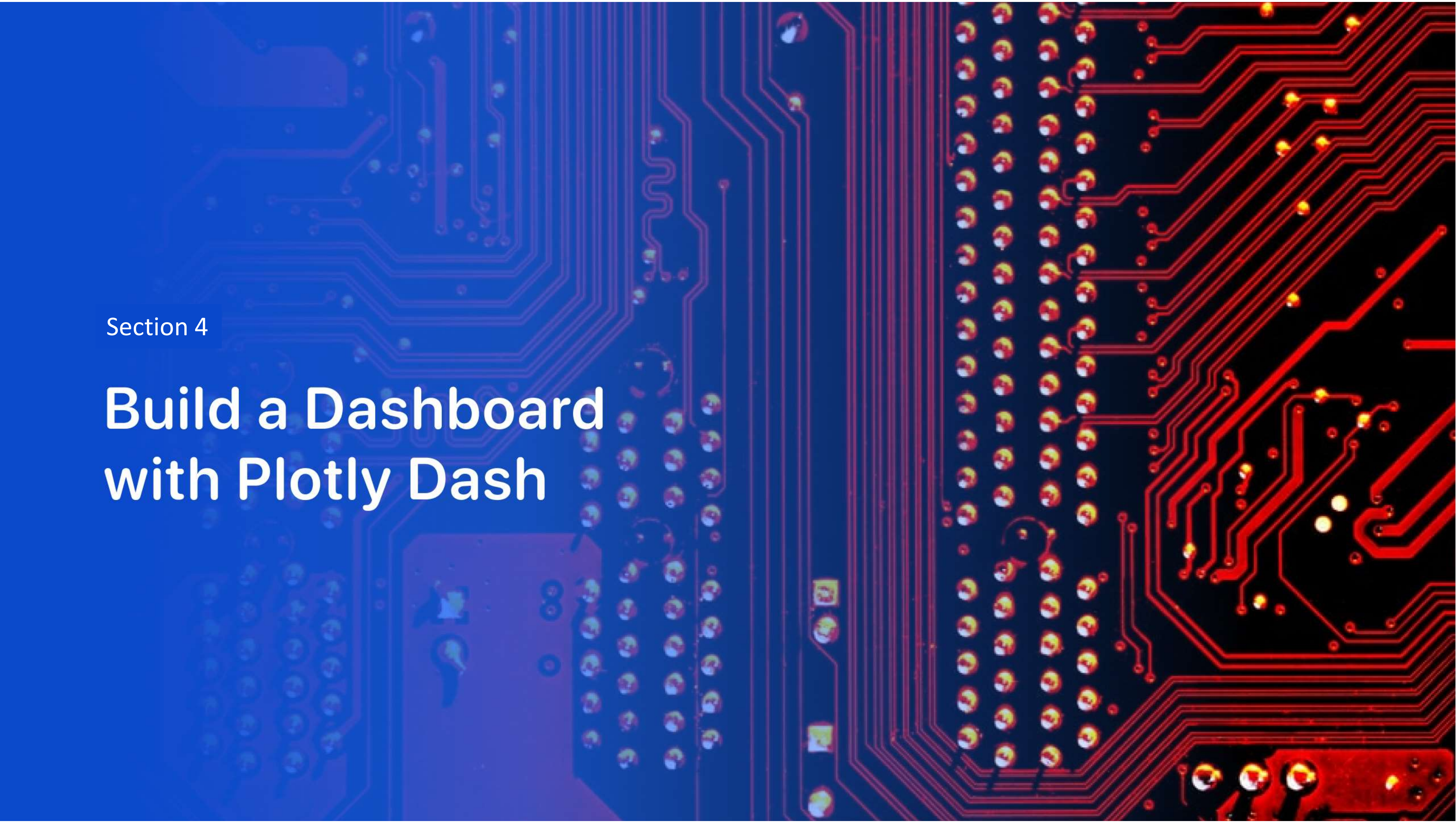
KSC LC 39A

- CCAFS LC-40 is the most launching site.
- KSC LC39A has the highest success rate.
- Overall, SpaceX has most of the launching in Titusville areas.

Folium Map with Launch Site to Its Proximities



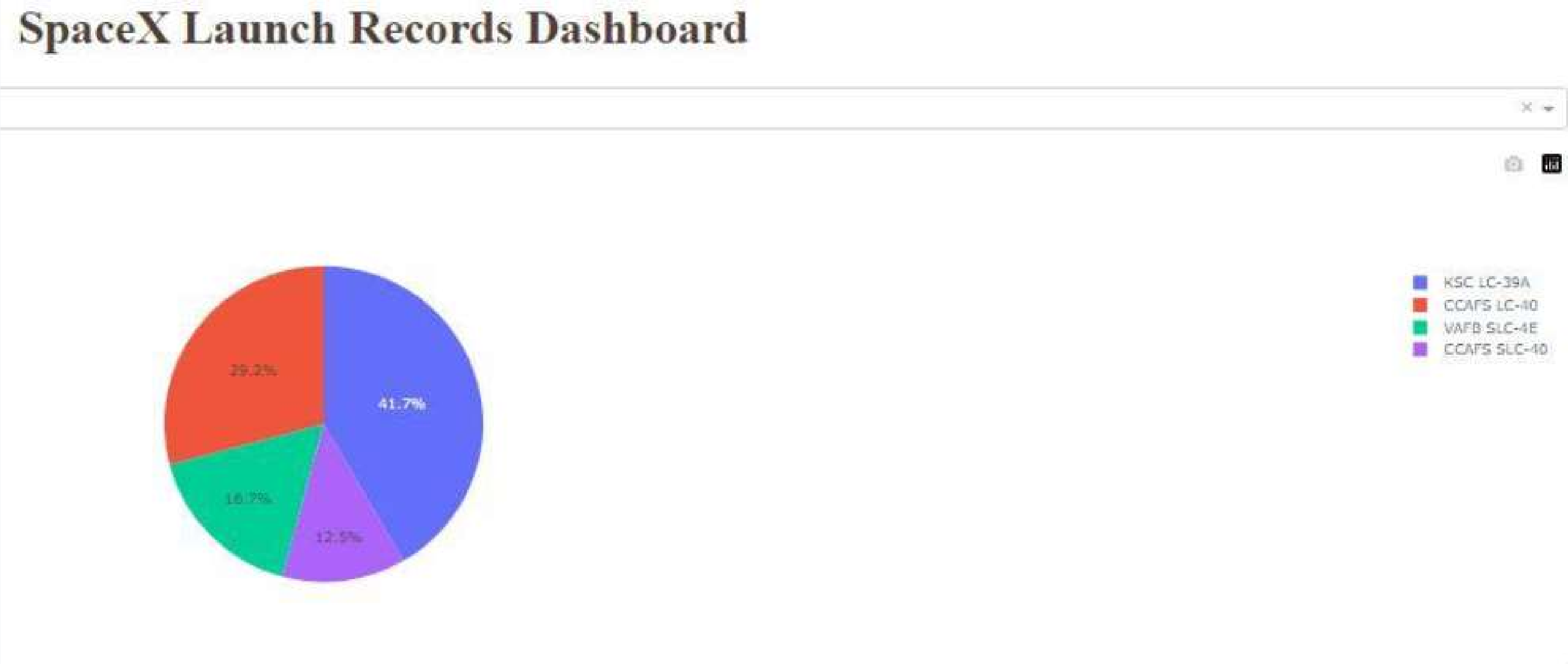
- Folium map show the distance of CCAFS SLC-40 launch site to nearest railway, highway, airport and coastline.
- The site is near to railway and coastline which is 0.58km and 0.86km.
- 1.28km to the nearest highway.
- However, there is quite a distance to the nearest airport which is 51.43km away.



Section 4

Build a Dashboard with Plotly Dash

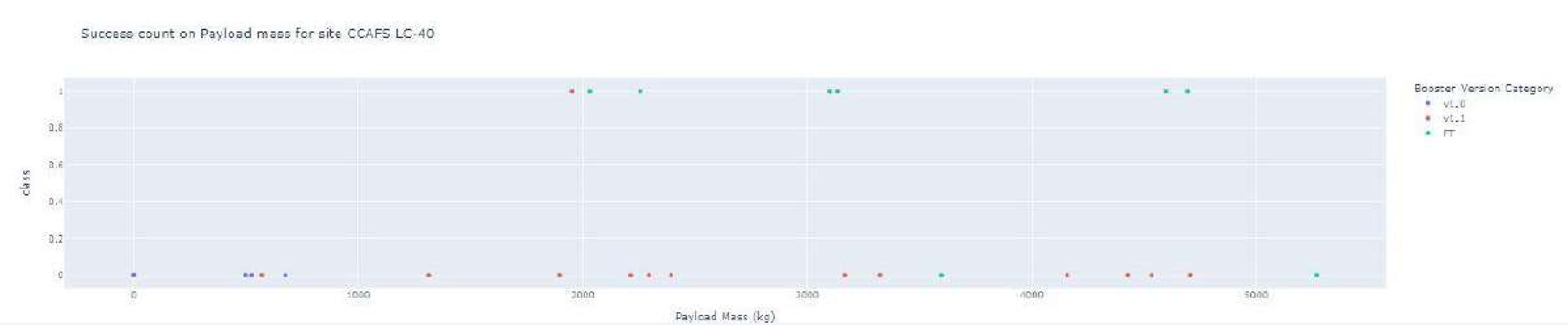
<Dashboard Screenshot 1>



SpaceX Launch Records Dashboard

Based on pie chart, KSC LC-39A has the most launch record which is 41.7% and CCAFS SLC-40 has the lowest number of launch which contribute 12.5% of total.

<Dashboard Screenshot 3>

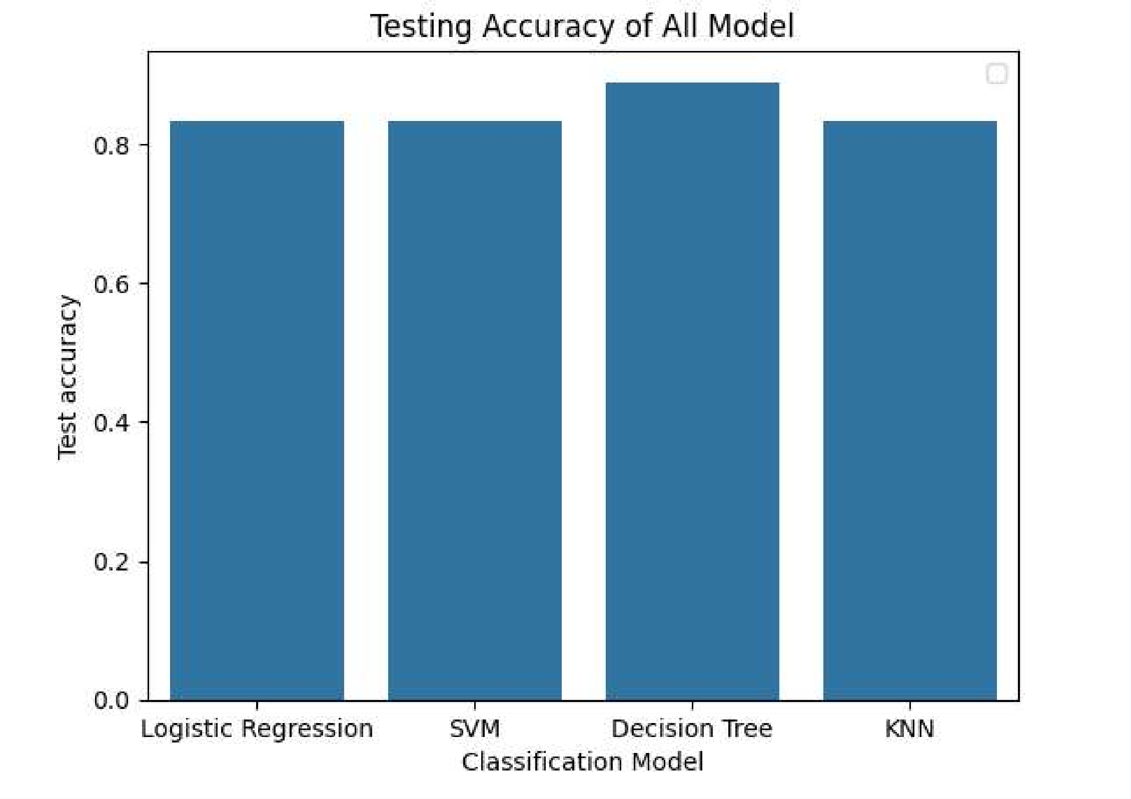


- For site CCAFS-LC-40, booster version FT has higher success rate for heavy load rocket.

Section 5

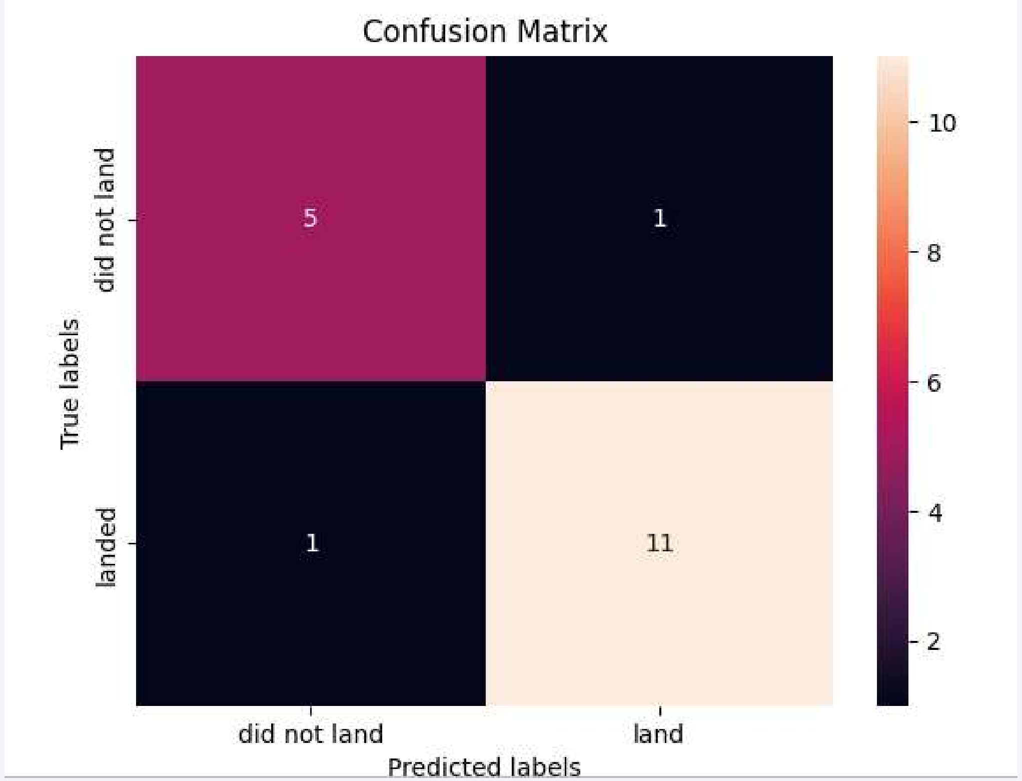
Predictive Analysis (Classification)

Classification Accuracy



Decision Tree is the best classification model to predict the landing outcome. The accuracy of the model is 88.88% which is higher compare to other models.

Confusion Matrix



Here is the confusion matrix of Decision Tree:

- Total test data is 18.
- One variance of predicted <did not land> out of 6.
- One variance of predicted <land> out of 12.

Conclusions

- The success rate of landing showing the trend of increase from year 2013 to 2020.
- There is positive relationship between flight number and success outcome especially site CCAFS SLC-40.
- For heavy load mass launching, CCAFS SLC-40 site and orbit including ISS, LEO and VLEO have higher success rate of landing.
- The best model to predict the success of landing is decision tree which the accuracy of predicted outcome achieve 88.88%.

Thank you!