IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Santiago Paniagua
July 30th 2025

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies

for EDA pandas NumPy

for visualization seaborn matplotlib plotly Dash by plotly

for data collection Requests

for predictions scikit learn

for maps Folium

BeautifulSoup for webscraping

Data Base IBM DB2

- Summary of all results

-What makes the rocket land successfully?

-What does SpaceX have to take in count for future successful landings?

Before, we have the libraries used to understand the data and be able to have predictions on the data.

Introduction

- Project background and context

SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.

- Problems you want to find answers

We want to determine if the first stage will land so that we are able to determine the cost of a launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX's API
  - Web Scrapping performed on Wikipedia
- Perform data wrangling
  - The data was filtered using Pandas. Just Falcon 9 data was used.
  - One-hot encoding was performed whether successful landing.
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Using different types of charts and diagrams, and SQL queries to show relationship between variable, to reveal patterns of the data, and to understand the data
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Scikit-learn was used to find the best model among multiple.

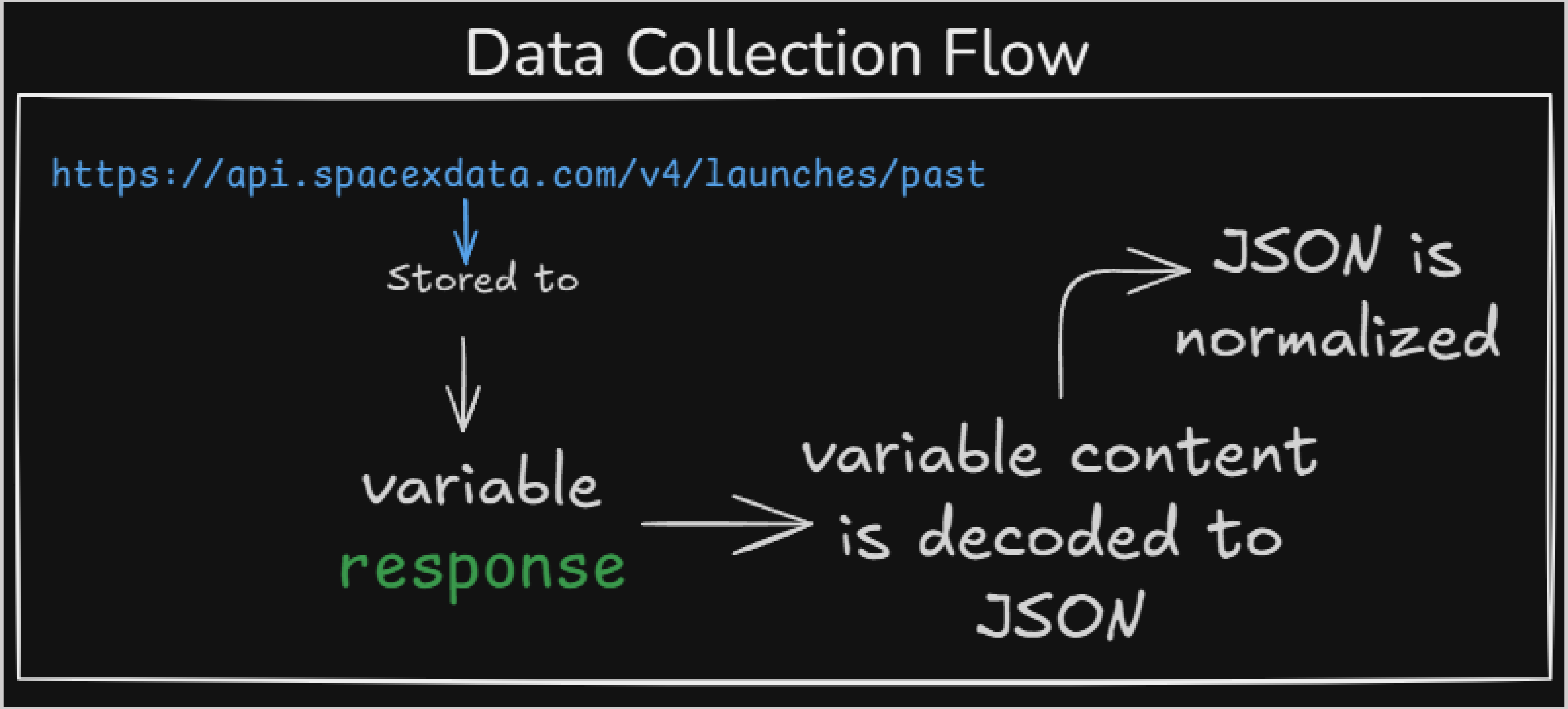
Data Collection

- How data sets were collected:

The data is from Spacex's database firstly functions where executed to call specific columns (rocket, launchpad, payloads, cores...). It was done with the library Requests and auxiliary functions to have the columns as list.

Then, the data had a process: from an URL was stored and decoded to JSON. To be able to work with Pandas, that JSON was transformed into Data Frame. The Data Frame was cleaned just to have Falcon 9 information as well as missing values.

Data Collection – SpaceX API



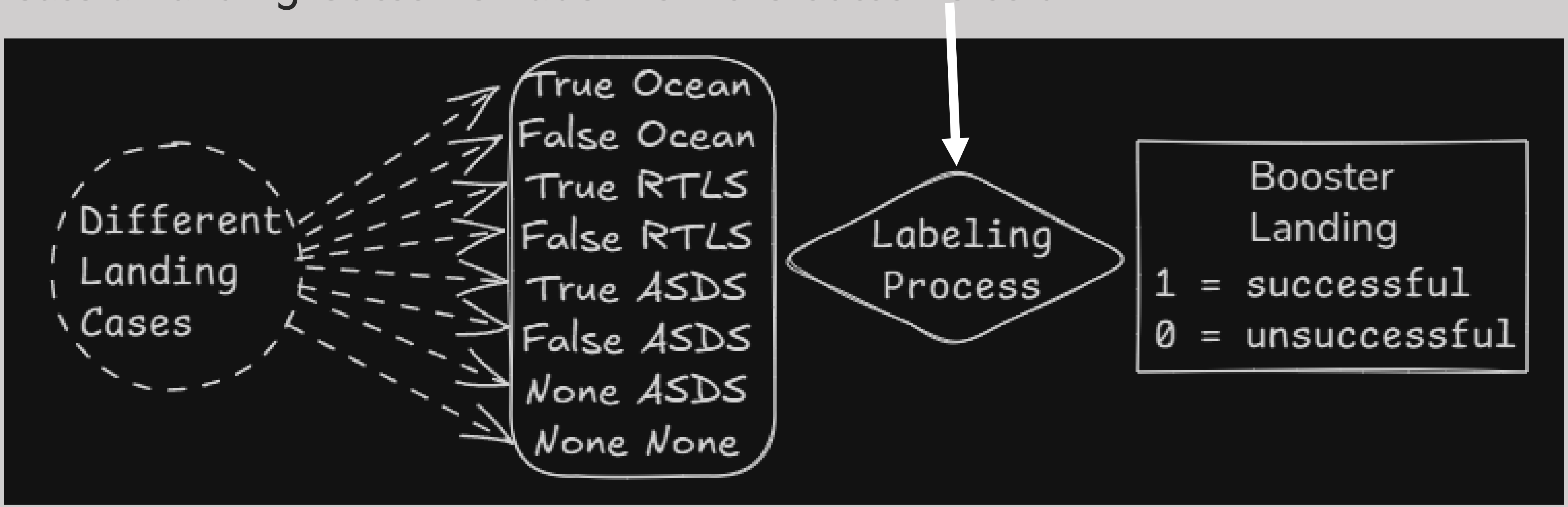
[Jupyter notebook where the data collection happens]

Data Collection - Scraping

Jupyter notebook where Scraping was performedh

Data Wrangling

- Missing values were found to handle them, the count of launches on each site was done.
- The number and occurence of mission outcome of the orbits was put together to create a Landing Outcome Label from the outcome column.



[See the Jupyter Notebook]

EDA with Data Visualization

Three types of visualizations where done: Scatter Plots, Barchar, Stepchar

The scatter plots were done to understand the relationship of the next pairs of vaiables:

- Launch Site vs Flight Number (p21)
- Launch Site vs Payload Mass (p22)
- Flight Number vs Orbit (p24)
- Payload Mass vs Orbit (p25)

Barchar

Used to see the Success Rate by Orbit. (graph on p23)

Stepchar

Used to see the Success Rate Over TIme (p26)

Jupyter Notebook where these plots were obtained

EDA with SQL

- A table is created with no null date records (cell #9)
- Table names are found (SPACEXTBL, SPACEXTABLE)(cell #11)
- Distinct "Launch_Sites" are found (CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, CCAFS SLC-40)(cell #13)
- Launch sites beginning with CCA are found (cell #15)
- Total payload mass carried by boosters launched by NASA (CRS)=45596 (cell #18)
- Average payload mass carried by booster version F9 v1.1=2928.4 (cell #19)
- Date when the first succesful landing outcome in ground pad was achieved=2018-07-22 (cell #21)
- Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000 = F9 FT B1022, F9 FT B1026, F9 FT B1021.2, F9 FT B1031.2
- Total number of successful and failure mission outcomes (cell #24)(first table on the right)
- All the booster_versions that have carried the maximum payload mass (cell #25)
- Records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015 (2nd table on the right)

Mission_Outcome	Total_PayloadMass_KG
Failure (in flight)	1952
Success	608615
Success	4400
Success (payload status unclear)	5000

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

[Jupyter Notebook with SQL queries]

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. (cell #27)

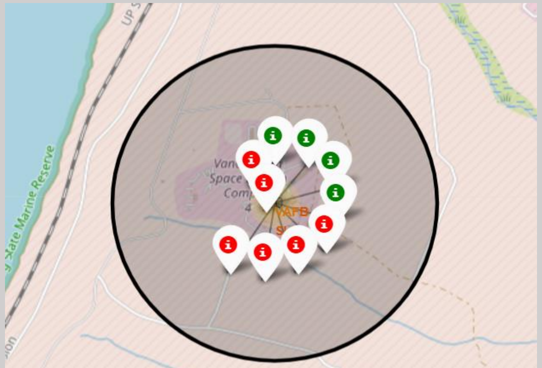
Landing_Outcome	Count	Rank
No attempt	10	1
Success (drone ship)	5	2
Failure (drone ship)	5	2
Success (ground pad)	3	4
Controlled (ocean)	3	4
Uncontrolled (ocean)	2	6
Failure (parachute)	2	6
Precluded (drone ship)	1	8

Jupyter Notebook with SQL queries

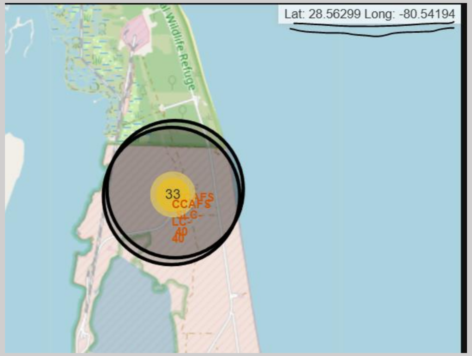
Build an Interactive Map with Folium

1)Launch Sites were tagged.

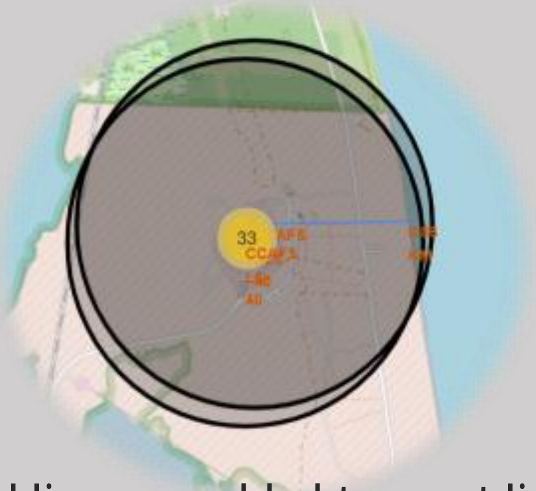
2)Number of Launches was added.



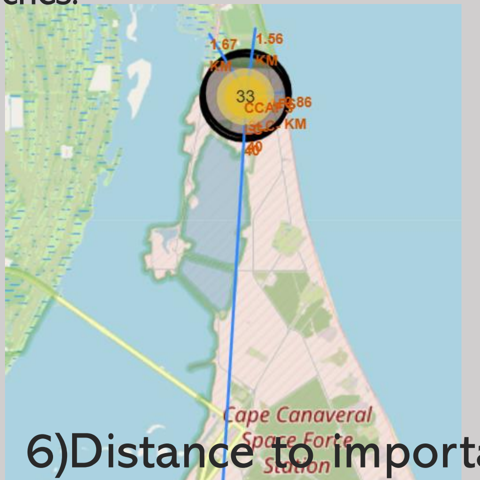
3)Tags added for succesfull and unsuccessful lauches.



4)On the top-right-corner, mouse position was added to be able to know distance from launch sites to important infrastructure



5)Trial line was added to coast line.



6)Distance to important infrastructure was added.

[Jupyter Notebook in GitHub]

Build a Dashboard with Plotly Dash

All sites, CCAF LC-40, VAFB SLC-4E, KSC LC-39A, CCAFS SLC-40 are the options available from the drop down menu.

For all sites the piechart plots percentage of launches whithin the total for each launch site.

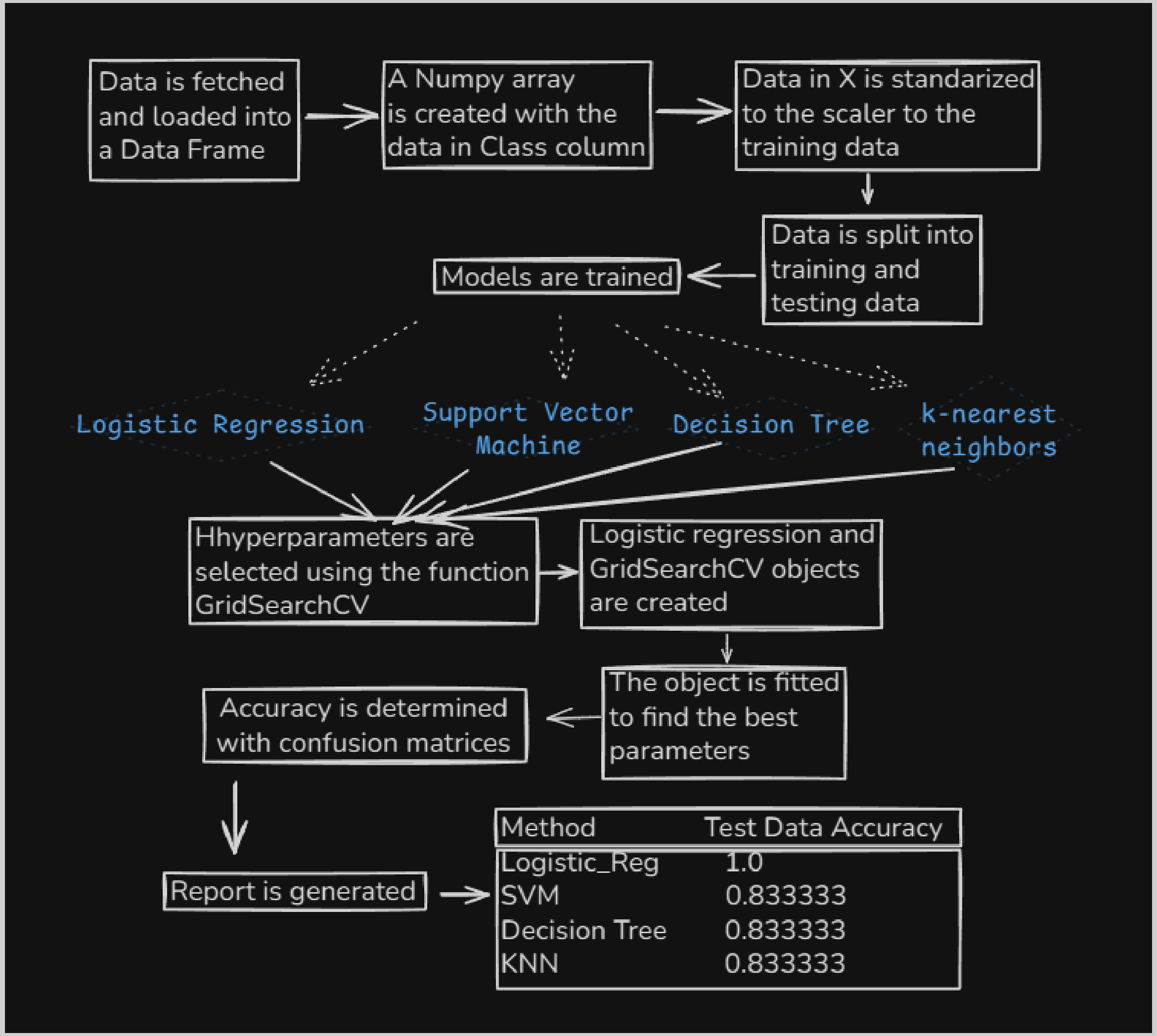
For the rest, the piechart plots successful launches vs unsuccesful launches percentage.

The scatter plot with a range slider allows to filter what is plotted according to the payload.

GitHub URL

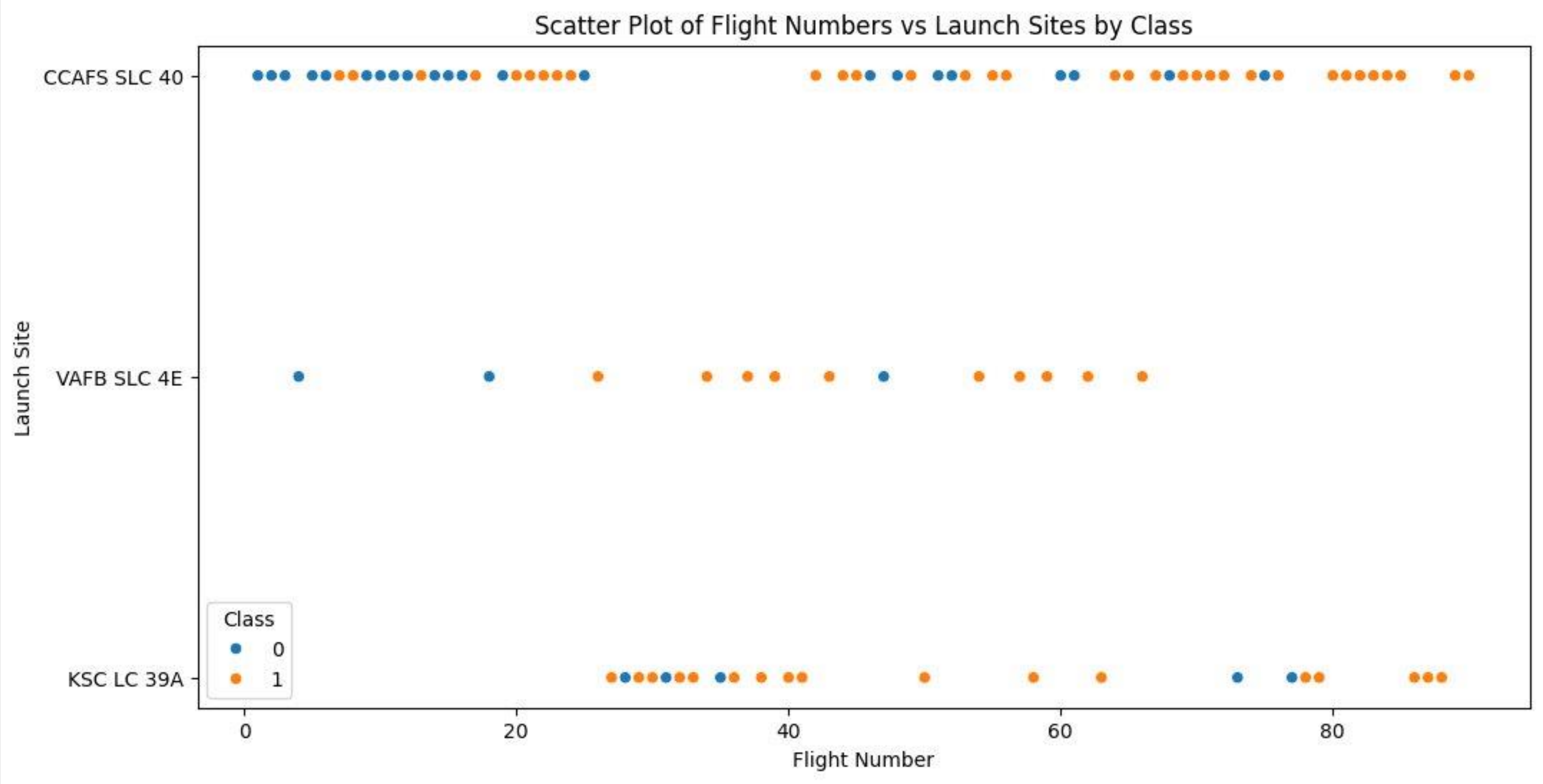
Predictive Analysis (Classification)

The evaluated models are: Logistic Regression, Support Vector Machine, Decision Tree and K-nearest Neighbors.

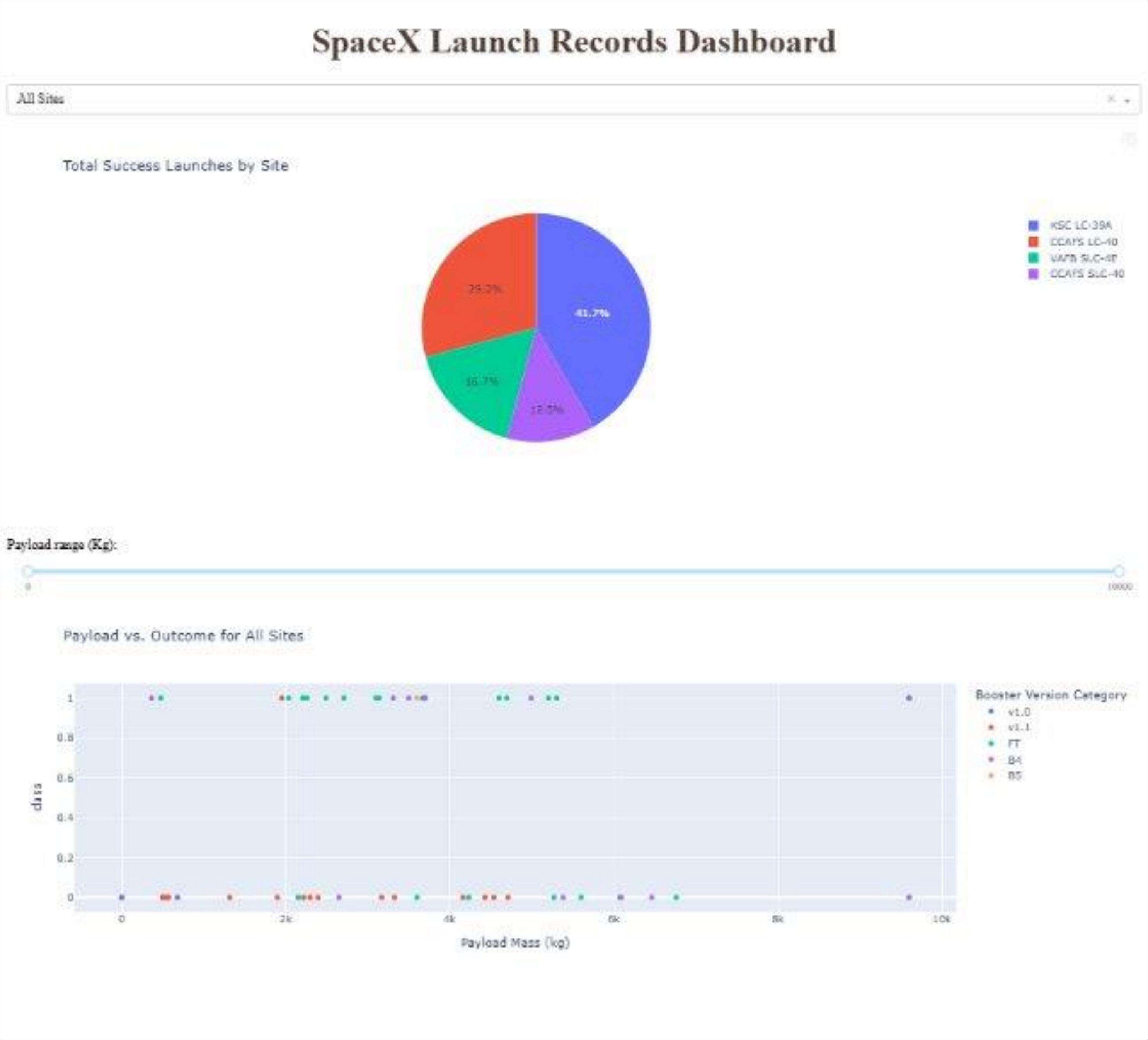


[Jupyter Notebook in Github]

Results



Scatter Plot of Flight Numbers vs Launch Sites by Class
Launch Site
CCAFS SLC 40
VAFB SLC 4E
KSC LC 39A
Flight Number
0
20
40
60
80
Class
0
1


SpaceX Launch Records Dashboard
All Sites
Total Success Launches by Site
41.7%
29.2%
16.7%
12.5%
KSC LC-39A
CCAFS LC-40
VAFB SLC-4E
CCAFS SLC-40
Payload range (Kg):
0
10000
Payload vs. Outcome for All Sites
class
Payload Mass (kg)
Booster Version Category
v1.0
v1.1
FT
B4
B5

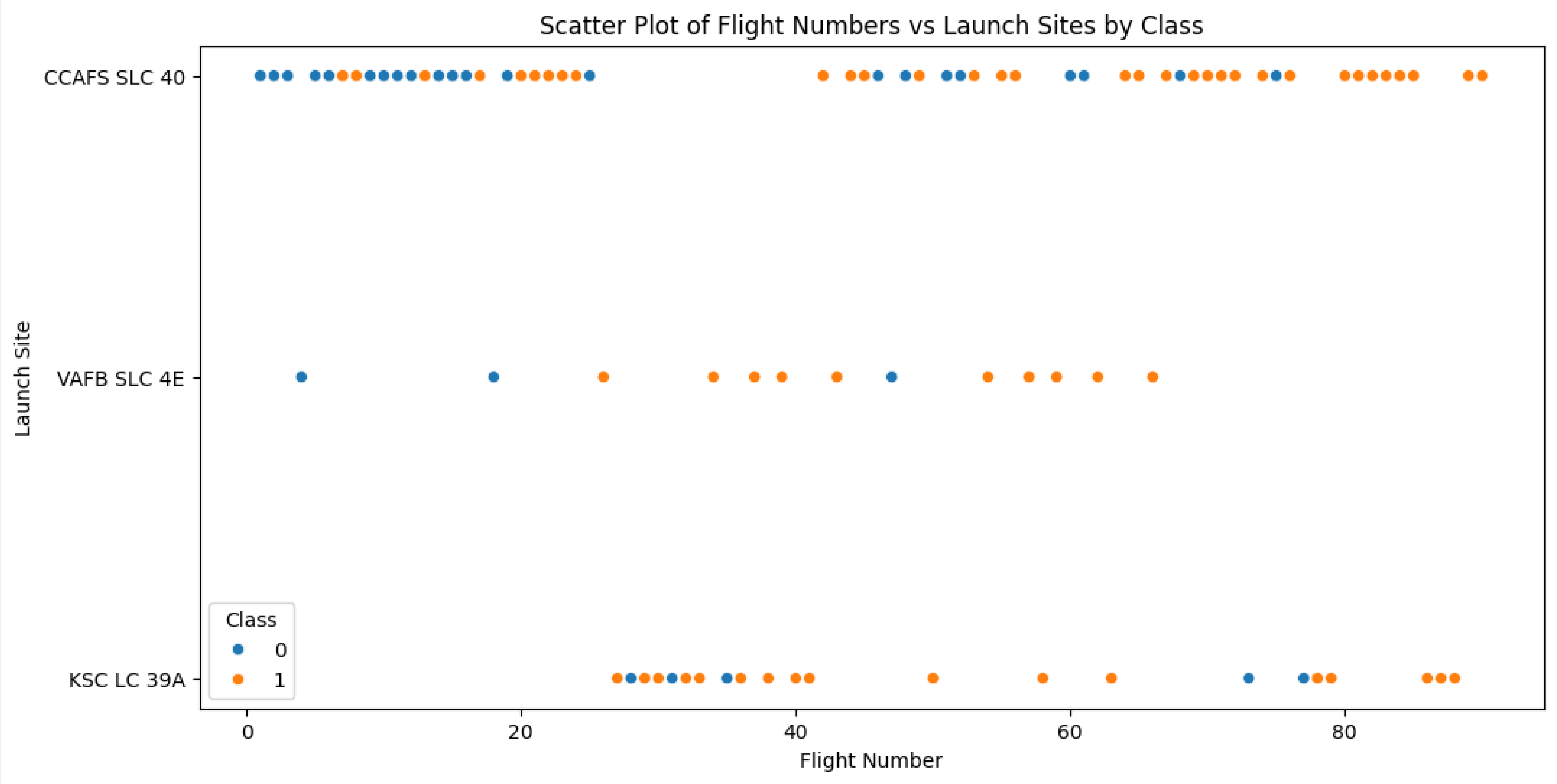
Method	Test Data Accuracy
Logistic_Reg	1.0
SVM	0.833333
Decision Tree	0.833333
KNN	0.833333

Section 2

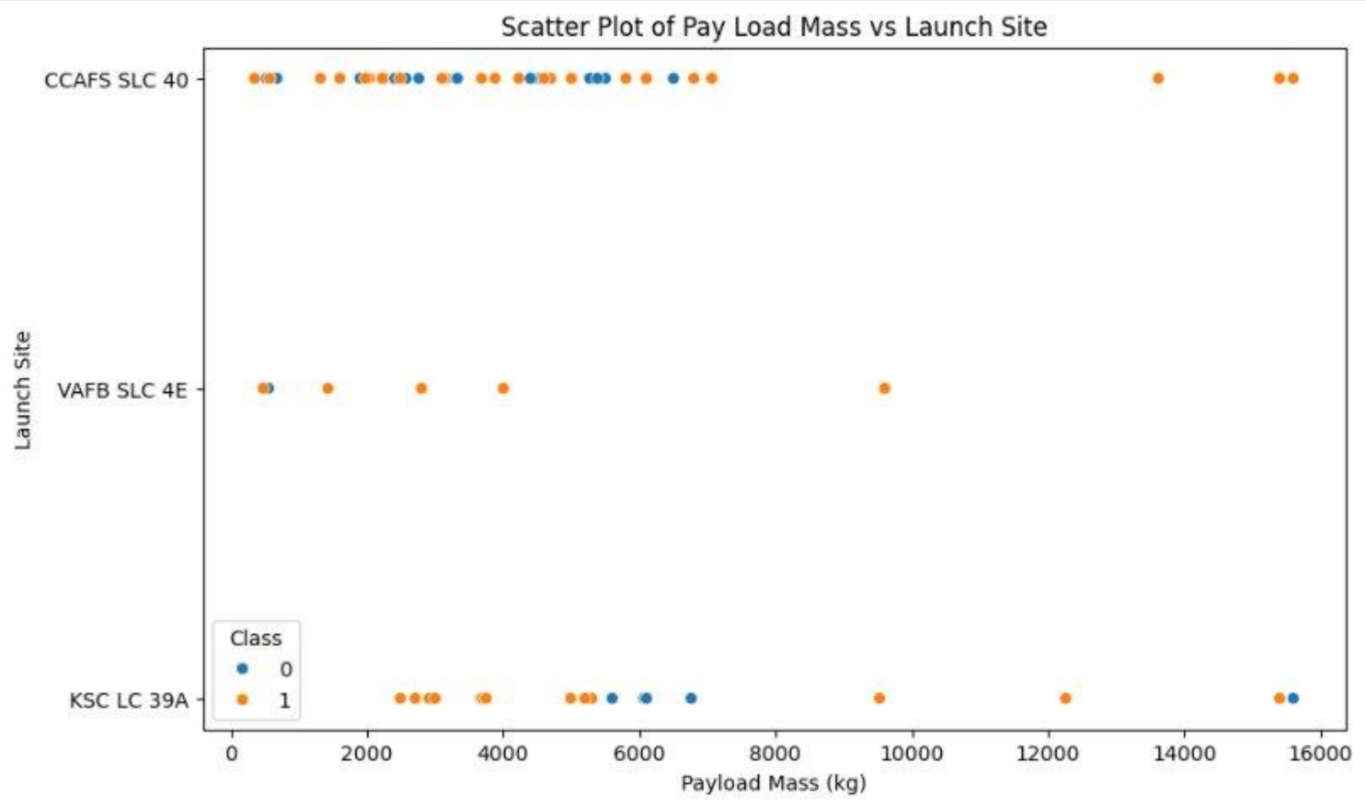
Insights drawn from EDA

Flight Number vs. Launch Site

Approximately on Flight Number 53, the missions began to be successful.



Payload vs. Launch Site

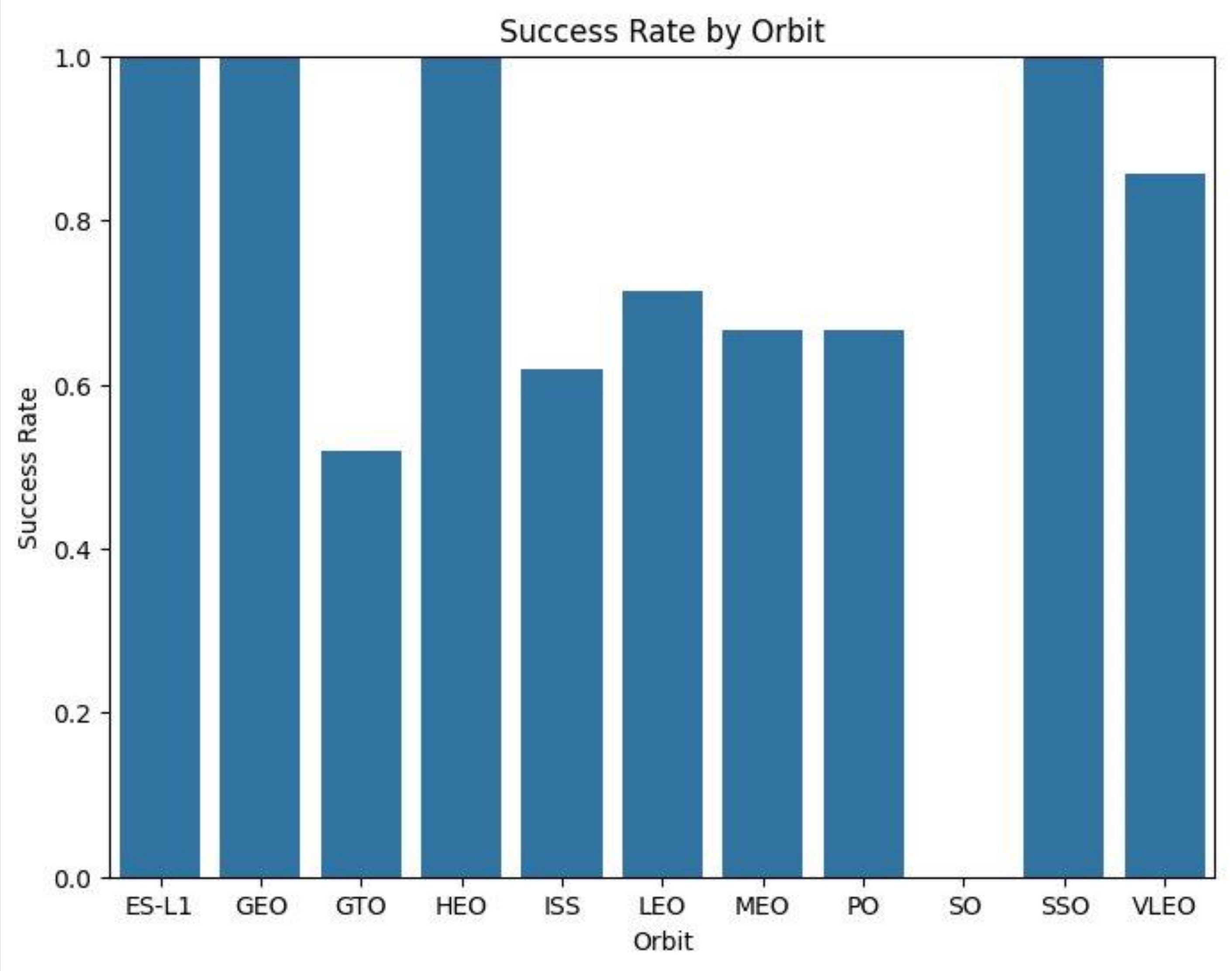


The best Launch Site being Payload the priority is Cape Canaveral Space Launch Complex 40.

Success Rate vs. Orbit Type

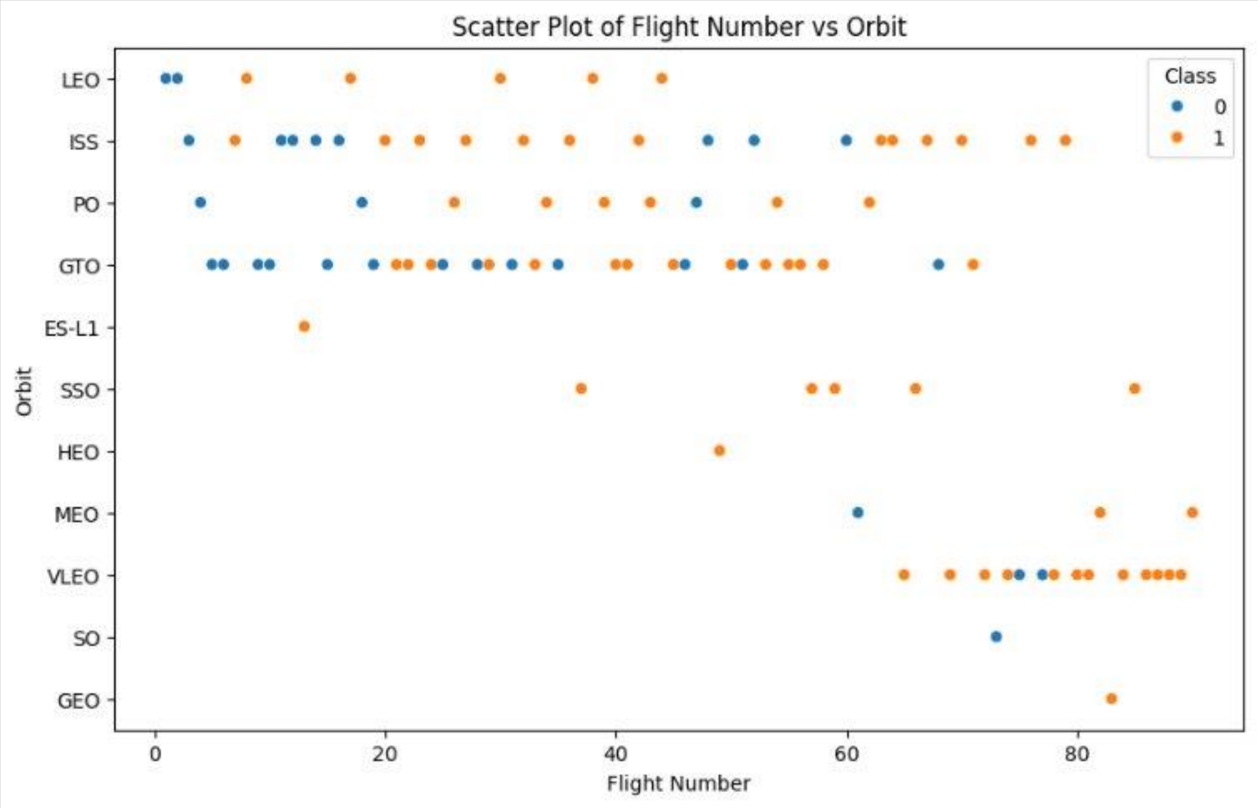
The best orbits are ES-L1, GEO,HEO and SSO in regards of success rate.

The worst performing orbit is GTO



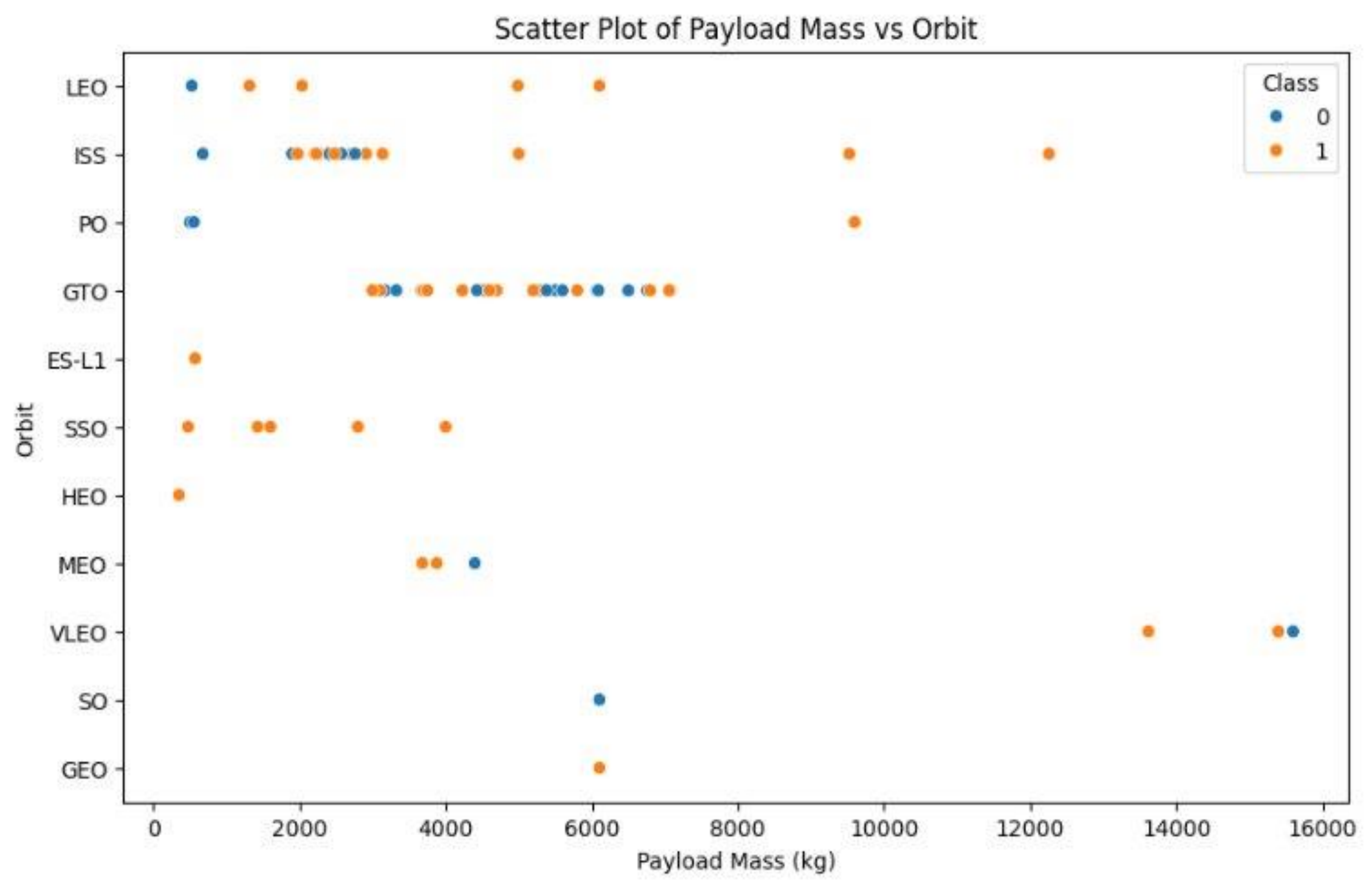
Flight Number vs. Orbit Type

The best orbits are ES-L1, GEO,HEO and SSO in regards of success rate.



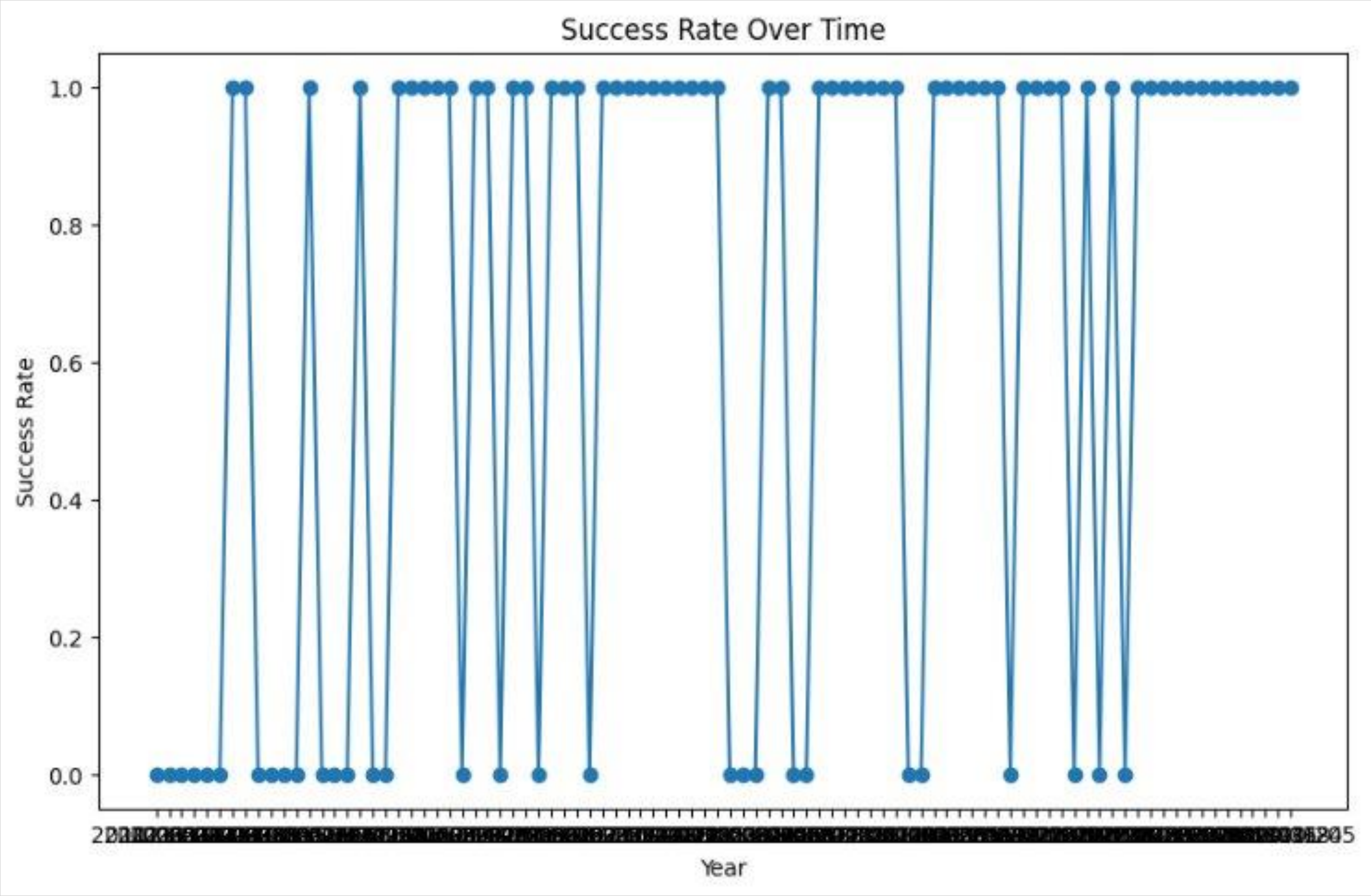
Payload vs. Orbit Type

The best orbits taking in count the payload are ISS and VLEO



Launch Success Yearly Trend

It was very difficult to properly print the years, but the success rate stays successful from 2022.



All Launch Site Names

The next sql query was performed to obtain launch site names:

```
[13]: %%sql
      SELECT DISTINCT Launch_Site
      FROM SPACEXTBL;

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

[13]:

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

Launch Site Names Begin with 'CCA'

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

This launch site is the one with the best performance, here a quere was performed to obtain the table above.

Total Payload Mass

Below, we have a screenshot where the sql query was performed with its result.

```
[18]: %%sql
      SELECT SUM(PAYLOAD_MASS__KG_)
      FROM SPACEXTBL
      WHERE Customer = 'NASA (CRS)';

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

[18]:

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

```
[19]: %%sql
      SELECT AVG(PAYLOAD_MASS__KG_)
      FROM SPACEXTBL
      WHERE Booster_Version LIKE 'F9 v1.1';

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

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

```
[21]: %%sql
      SELECT MIN(Date) AS First_Success_Date
      FROM SPACEXTBL
      WHERE Landing_Outcome = 'Success';

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

[21]:

First_Success_Date
2018-07-22

The very first try was on 2015, but the very first successful landing was on 2018.

Successful Drone Ship Landing with Payload between 4000 and 6000

[23]:
```
%%sql
SELECT Booster_Version
FROM SPACEXTBL
WHERE Landing_Outcome = 'Success (drone ship)'
  AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

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

[23]:

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

Total Number of Successful and Failure Mission Outcomes

[24]:
```
%%sql
SELECT Mission_Outcome,
       SUM(PAYLOAD_MASS__KG_) AS Total_PayloadMass_KG
FROM SPACEXTBL
GROUP BY Mission_Outcome;
```

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

[24]:

Mission_Outcome	Total_PayloadMass_KG
Failure (in flight)	1952
Success	608615
Success	4400
Success (payload status unclear)	5000

Other factors determine more than the payload whether the outcome will be successful or not.

Boosters Carried Maximum Payload

```
[25]: %%sql
SELECT Booster_Version
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = ( SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL );
```

[25]:

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

```
[26]: %%sql
SELECT
    STRFTIME('%Y-%m', Date) AS Month,
    Landing_Outcome,
    Booster_Version,
    Launch_Site
FROM
    SPACEXTBL
WHERE
    Landing_Outcome = 'Failure (drone ship)'
AND
   SUBSTR(Date, 1, 4) = '2015';

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

[26]:

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

In 2015 there was not any successful landing

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

```
[27]: %%sql
SELECT
    Landing_Outcome,
    COUNT(*) AS Count,
    RANK() OVER (ORDER BY COUNT(*) DESC) AS Rank
FROM
    SPACEXTBL
WHERE
    Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY
    Landing_Outcome;

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

[27]:

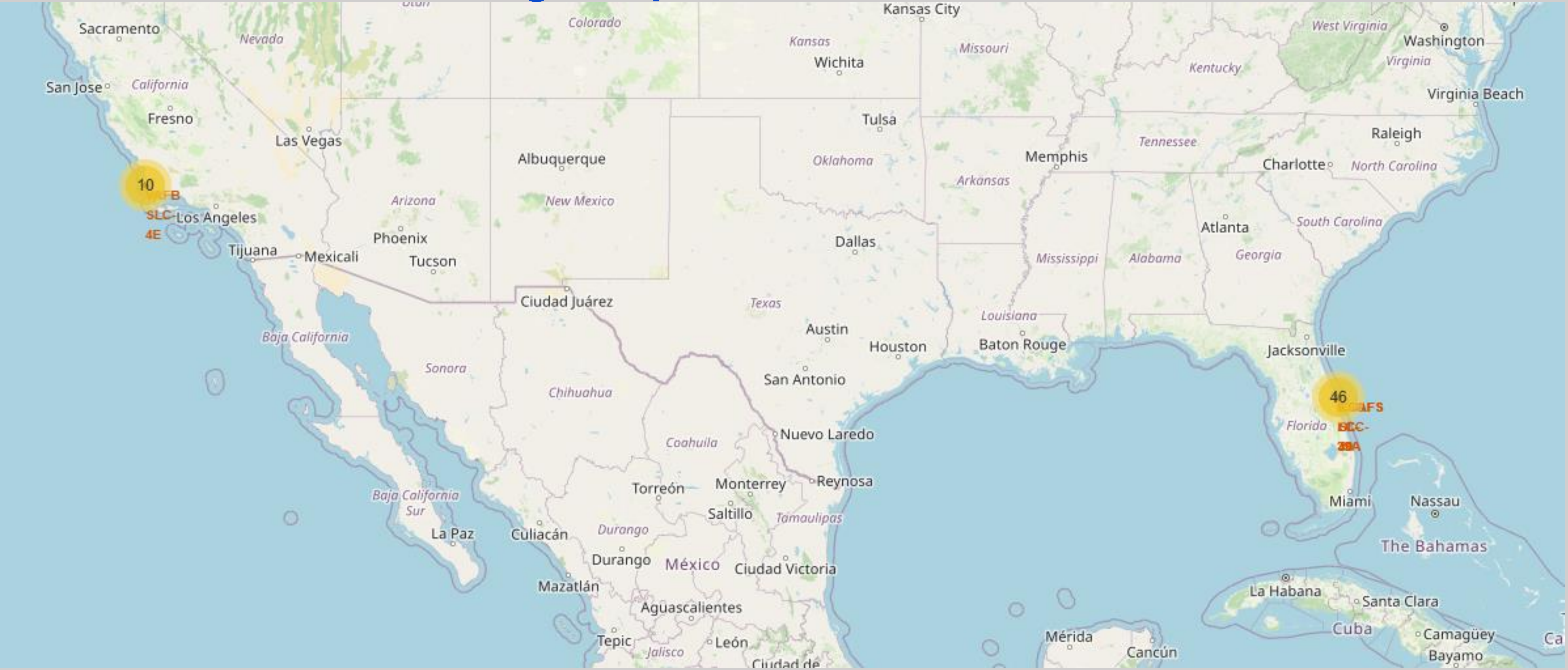
Landing_Outcome	Count	Rank
No attempt	10	1
Success (drone ship)	5	2
Failure (drone ship)	5	2
Success (ground pad)	3	4
Controlled (ocean)	3	4
Uncontrolled (ocean)	2	6
Failure (parachute)	2	6
Precluded (drone ship)	1	8

Drone ship is the safest option according to the data to perform the booster landing.

Section 3

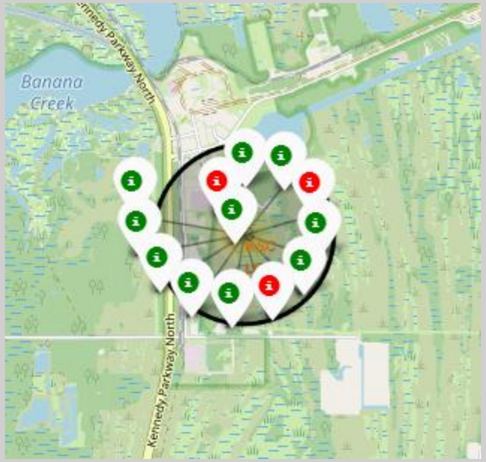
Launch Sites Proximities Analysis

All locaton sites groupded

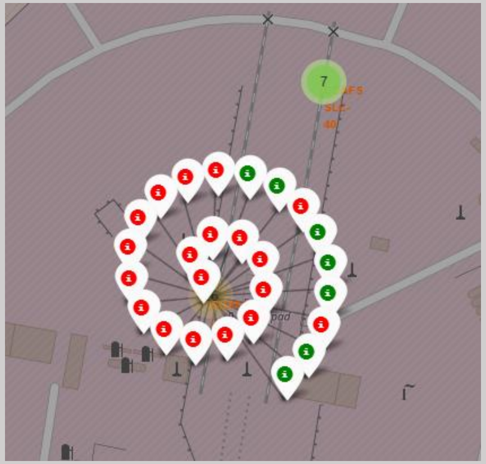


Both launchsites are located in Space Force facilities, **Vandenberg Space Launch Complex 4** on the left and **Cape Canaveral Space Launch Complex 40** on the right where SpaceX's launches are done.

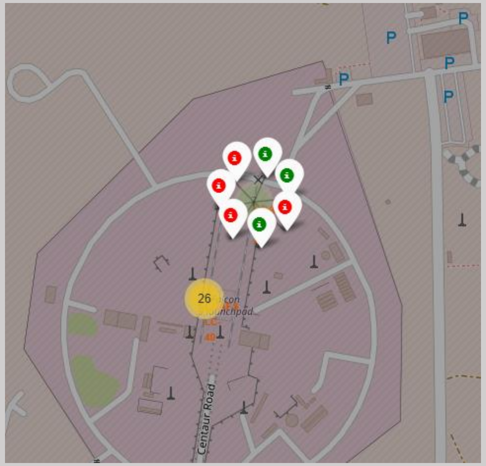
Landing Sites with Tags



KSC LC-39A

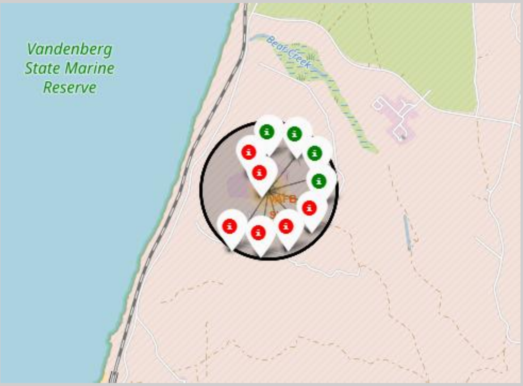


CCAFS SLC-40



CCAFS SLC-40

East Coast

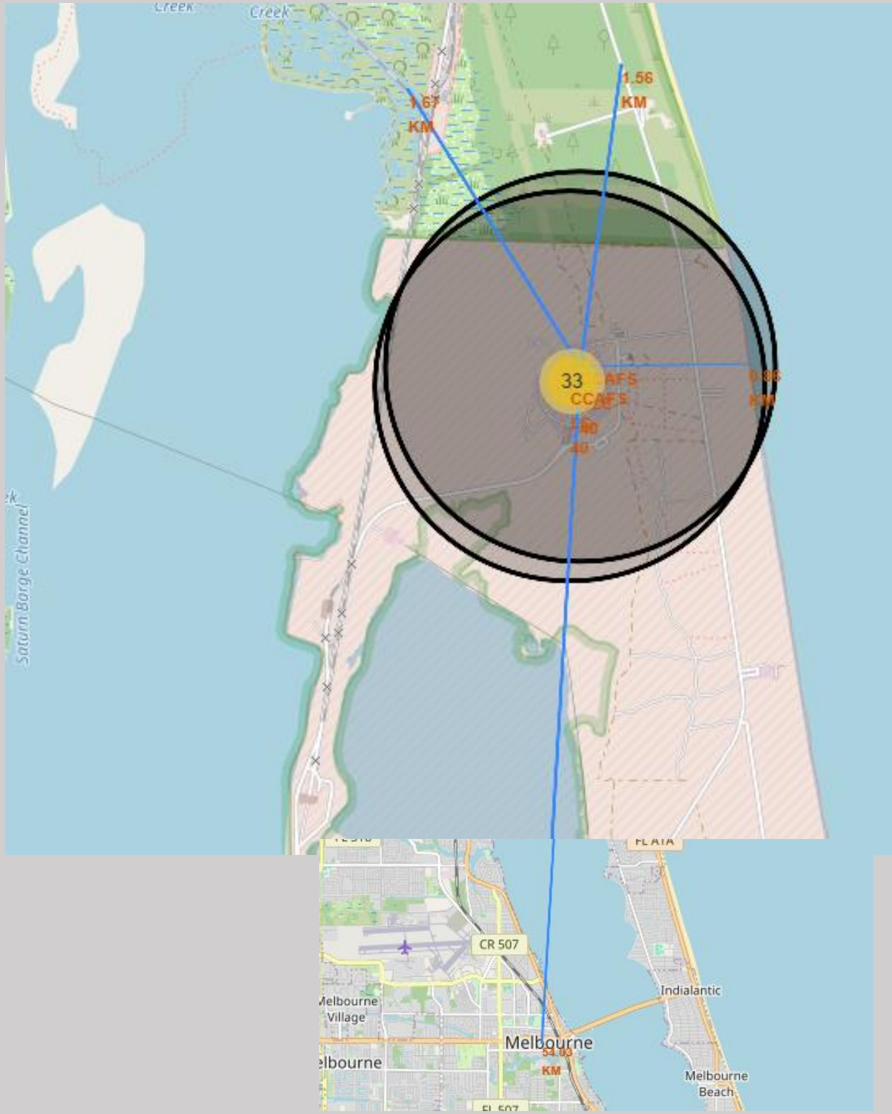


West Coast

VAFB SLC-4E

- KSC LC-39A is the location with the highest number of successful landings.

Important points near the location



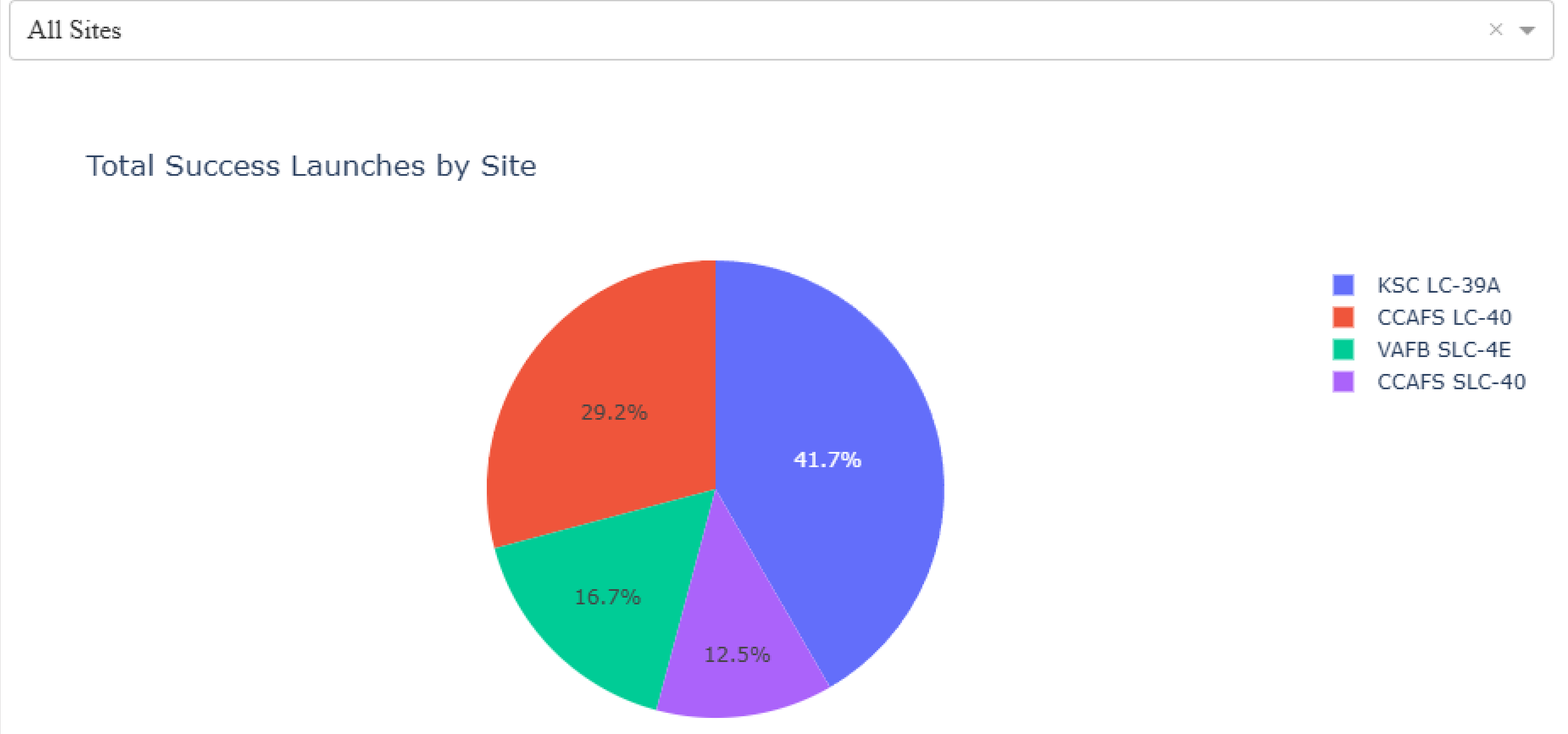
Cape Canaveral Space Force Station’s SLC-40 is about 1Km away from the Atlantic Ocea and about 1.5Km away from a railway and road.

The closest city is Melbourn-Florida

Section 4

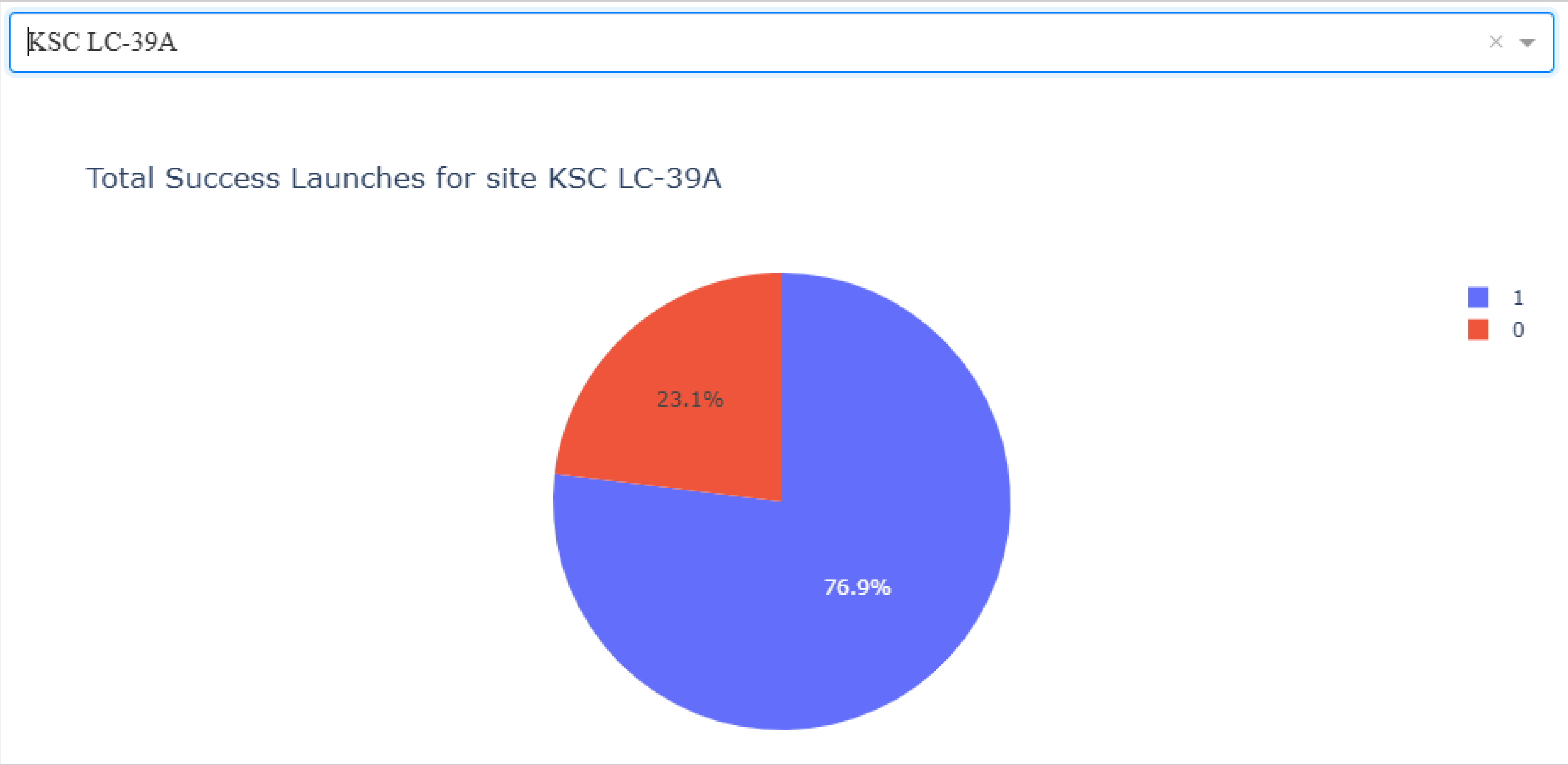
Build a Dashboard with Plotly Dash

All Total Success Launches by Site



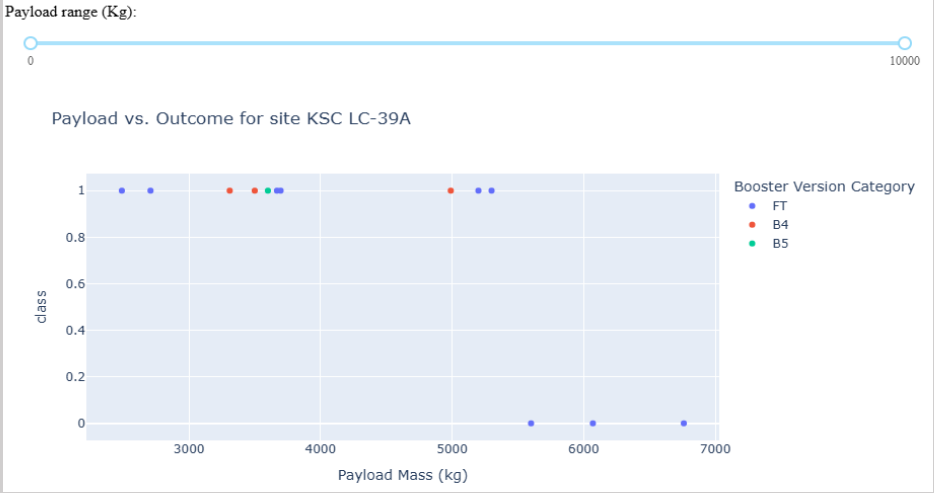
KSC LC-39A is the Launch Site with the most succesful launches and CCAFS SLC-40 is the one with the lowest ammount.

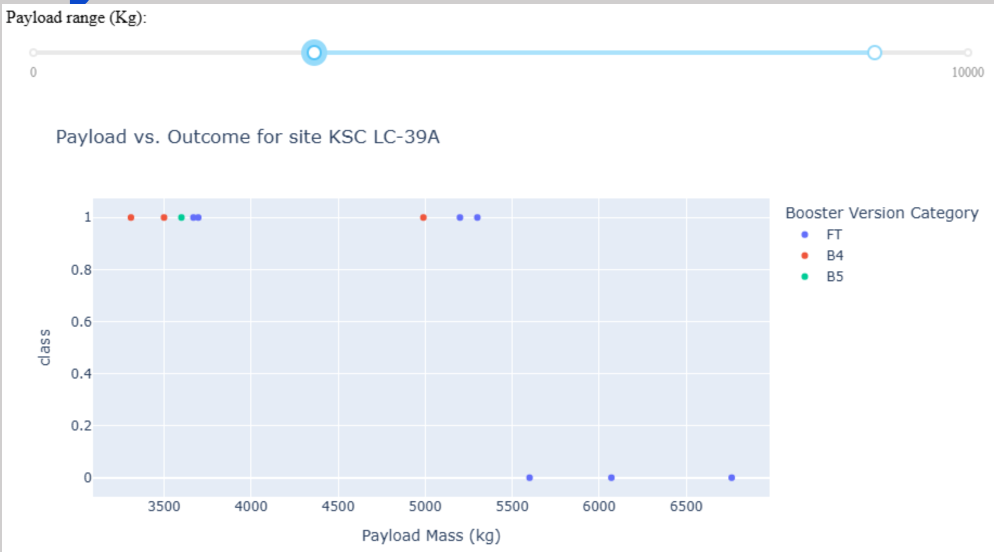
Successful vs Unsuccessful Launches for KSC LC-39A

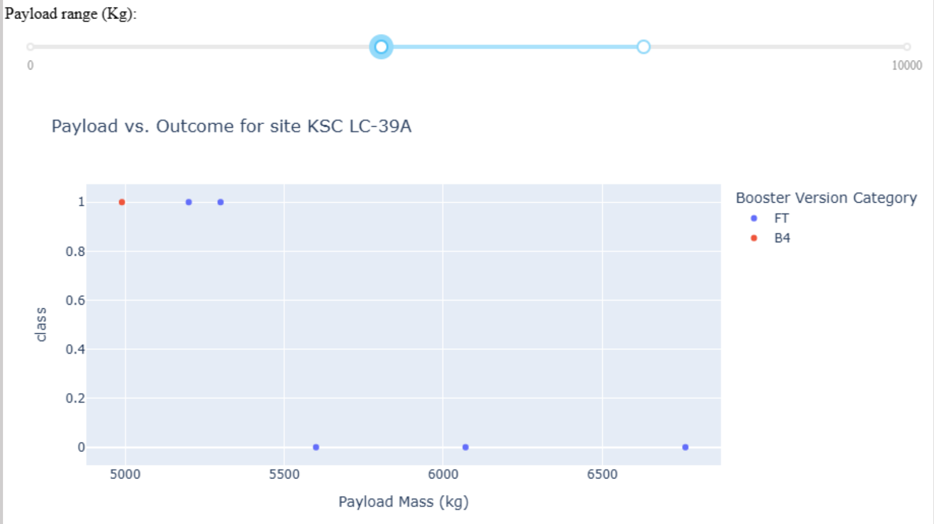


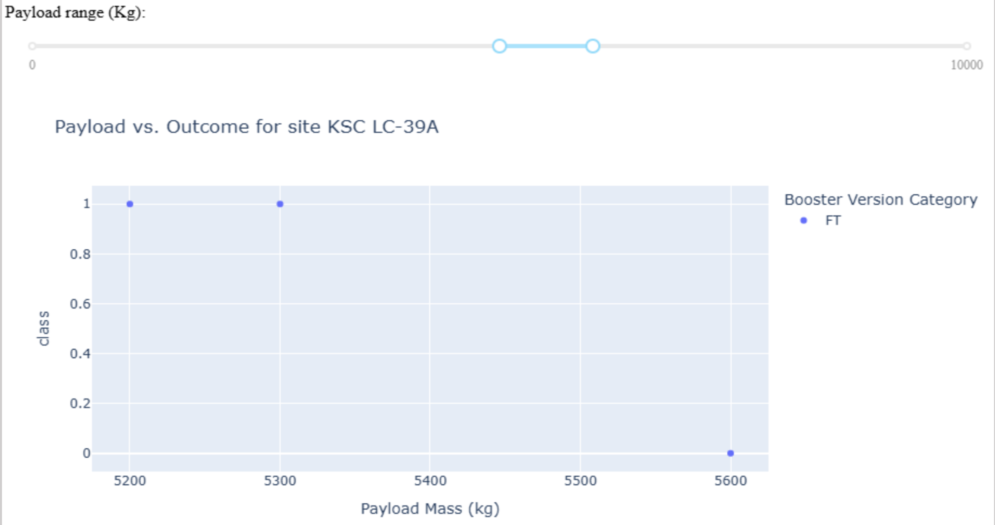
Successful Launches expressed in ratio are 3.34 : 1. It means that are 3.34 times bigger than the unsuccessful ones

Plots depending on the Payload







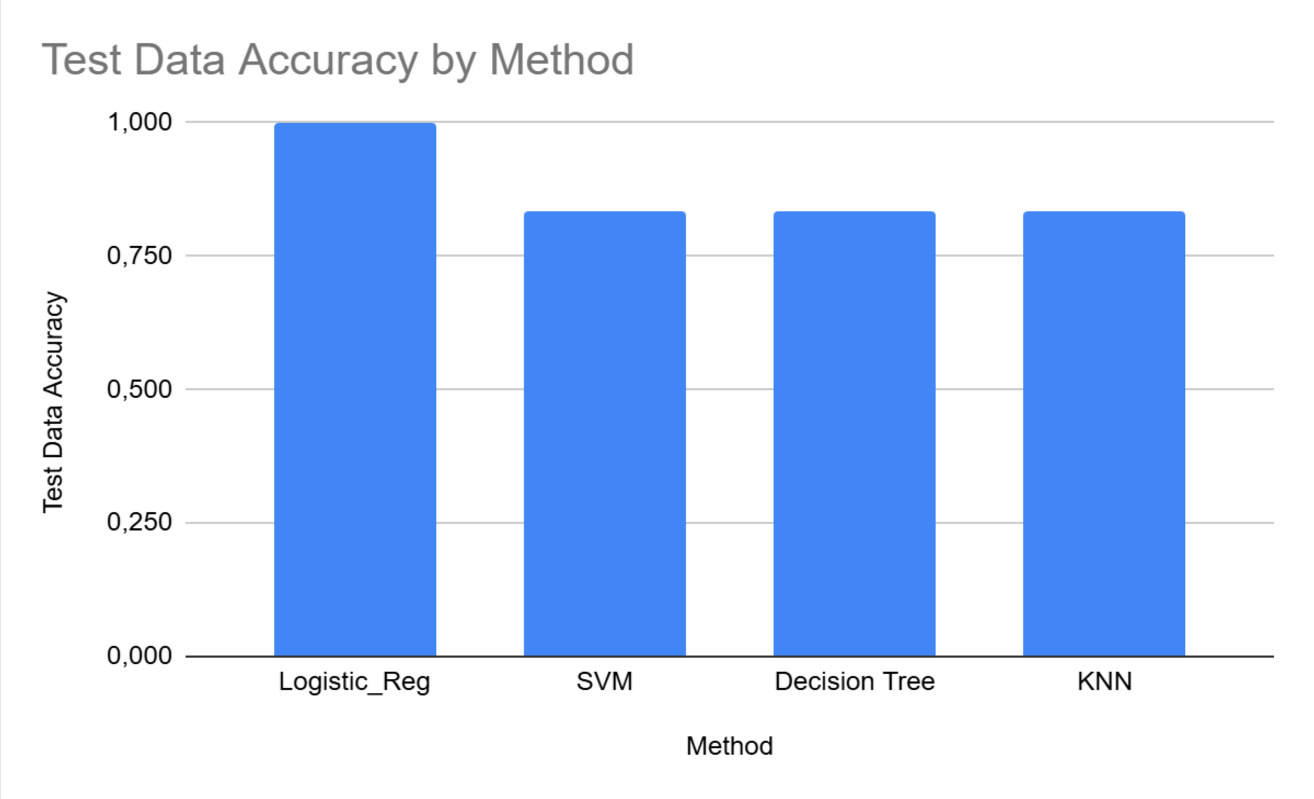


The settings on the top left corner shows the payload interval with highest success rate.

Section 5

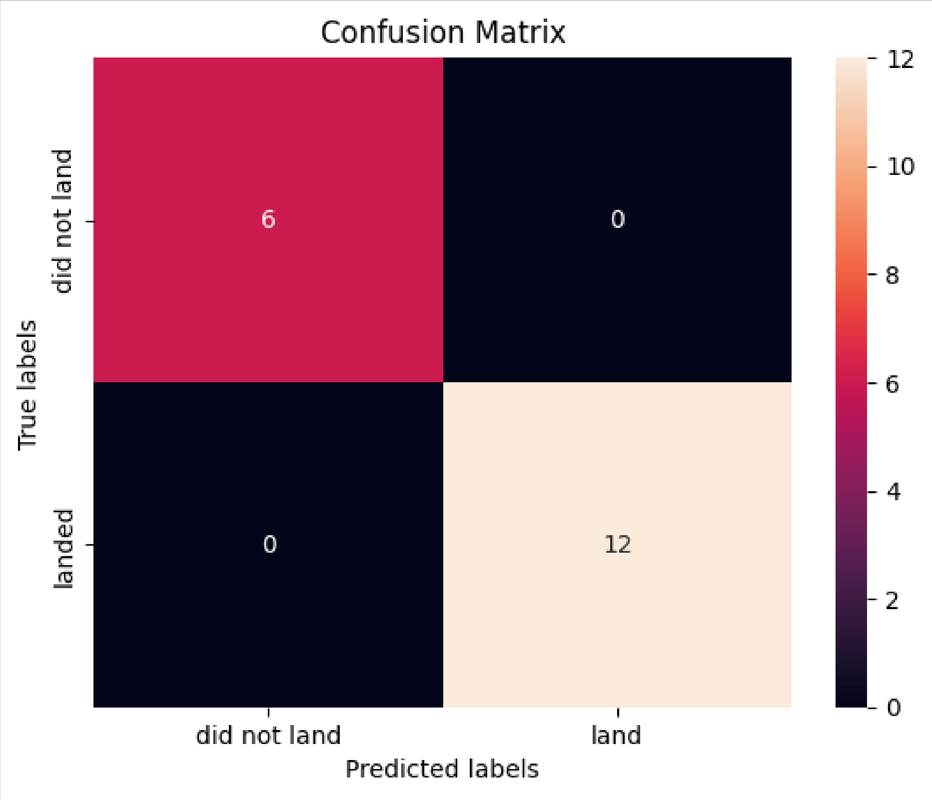
Predictive Analysis (Classification)

Classification Accuracy



The model with the highest accuracy was Logistic Rregression.

Confusion Matrix



This is a confusion matrix for the Logistic Regresion. There were no instances where the model made an error, we can say that the model has achieved an excellent level of accuracy as it correctly predicted all instances with 100% precision and recall. here were no instances where the model made an error

Conclusions

- Cape Canaveral Space Launch Complex 40 has the highest number of successful landings.
- The model that best predicts the lading of Falcon 9 is Logistic Regression with a score of 1.
- ES-L1, GEO, HEO and SSO have a 100% success rate
- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

Thank you!