IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

MI Umar
21 October 2023

Outline

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

Executive Summary

The methodologies that are used in this project to analyze the data include:

- ✓ Collection of data by SpaceX API and web scraping using BeautifulSoup
- ✓ The exploratory analysis of data with SQL; data wrangling and data visualization using python's libraries such as Pandas, Numpy and Matplotlib
- ✓ Interactive visual data analysis using Folium and Plotly Dash
- ✓ Predictive analysis using Logistic Regression, Support Vector Machines, Decision Tree and K Nearest Neighbors.

- Summary of all results
  - ✓ The factors that were found correlated with success rate include launch sites, orbits, payload mass and flight numbers.
  - ✓ Decision tree was the best performing classifier with a validation score of 0.87.

Introduction

- **Project background and context**

- SpaceX is the most famous space-travel company that launch Falcon 9 rockets for space travel.

- The company has put tremendous efforts to make space travel affordable for the public, since it costs around 62 million dollars for one rocket launch. At such nominal price, the company has outpaced its competitors that charge around 165 million dollars for a similar launch.

- One of the major reasons behind such a low cost for a launch is the ability of the company to successfully land the rocket and re-use the first stage after that the payload is released in the orbit. However, not all attempts of landing the rocket have been successful. Hence, the cost of a future rocket launch depends on whether the rocket has successfully landed or not.

Image adopted from URL: https://en.wikipedia.org/wiki/File:CRS-8_(26239020092).jpg

Introduction

Problems you want to find answers

What is the impact of different factors such as orbit type, payload mass, flight numbers and launch site on the success of the landing of first stage?

Does the rate of successful landings covary with time scale?

Which machine learning algorithm predicts the success of a future rocket launch with greater accuracy?

Image adopted from URL: https://spacenews.com/falcon-9-launches-dragon-with-heavy-science-payload/

Section 1

Methodology

Methodology

- Executive Summary
- Data collection methodology: Data was collected via
    - SpaceX Rest API
    - Wikipedia using web scrapping
- Perform data wrangling: Data wrangling was performed by
    - Selecting and filtering the required columns from the data
    - Replacing the missing values with mean values
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - The predictive analysis was performed by using different classification algorithms such as logistic regression, SVM, K nearest neighbors and decision clouds.

Data Collection

- Since the complete information required for this project was not available at one single source, we obtained the information about Falcon 9 rocket launches from two different sources:

  a) SpaceX Rest API by using get request function

  b) Wikipedia Page by using web scraping (get request and BeautifulSoup)

Columns Obtained from SpaceX Rest Api
BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude.

Columns Obtained from Wikipedia
Flight No., Date and time, Launch site, Payload, Payload mass, Orbit, Customer, Launch outcome.

Data Collection – SpaceX API

- The data was collected in the form of a Json file from SpaceX Rest API by the Get Request.
- The Json file was normalized and later, it was converted into the data frame.
- The data related to the launch of Falcon 1 was removed and only Falcon 9 launch data was retained.
- The missing values were replaced by the mean value and the data was converted into cvs file for further processing.

GitHub URL: https://github.com/mihtishamu/DS-Capstone/blob/main/Lab%201%20Data%20Collection%20SpaceX%20API.ipynb

Data Collection – SpaceX API

Data With Missing Values

```
FlightNumber       0
Date               0
BoosterVersion     0
PayloadMass        5
Orbit              0
LaunchSite         0
Outcome            0
Flights            0
GridFins           0
Reused             0
Legs               0
LandingPad        26
Block              0
ReusedCount        0
Serial             0
Longitude          0
Latitude           0
dtype: int64
```

Missing Values Replaced by Means

```
FlightNumber       0
Date               0
BoosterVersion     0
PayloadMass        0
Orbit              0
LaunchSite         0
Outcome            0
Flights            0
GridFins           0
Reused             0
Legs               0
LandingPad        26
Block              0
ReusedCount        0
Serial             0
Longitude          0
Latitude           0
dtype: int64
```

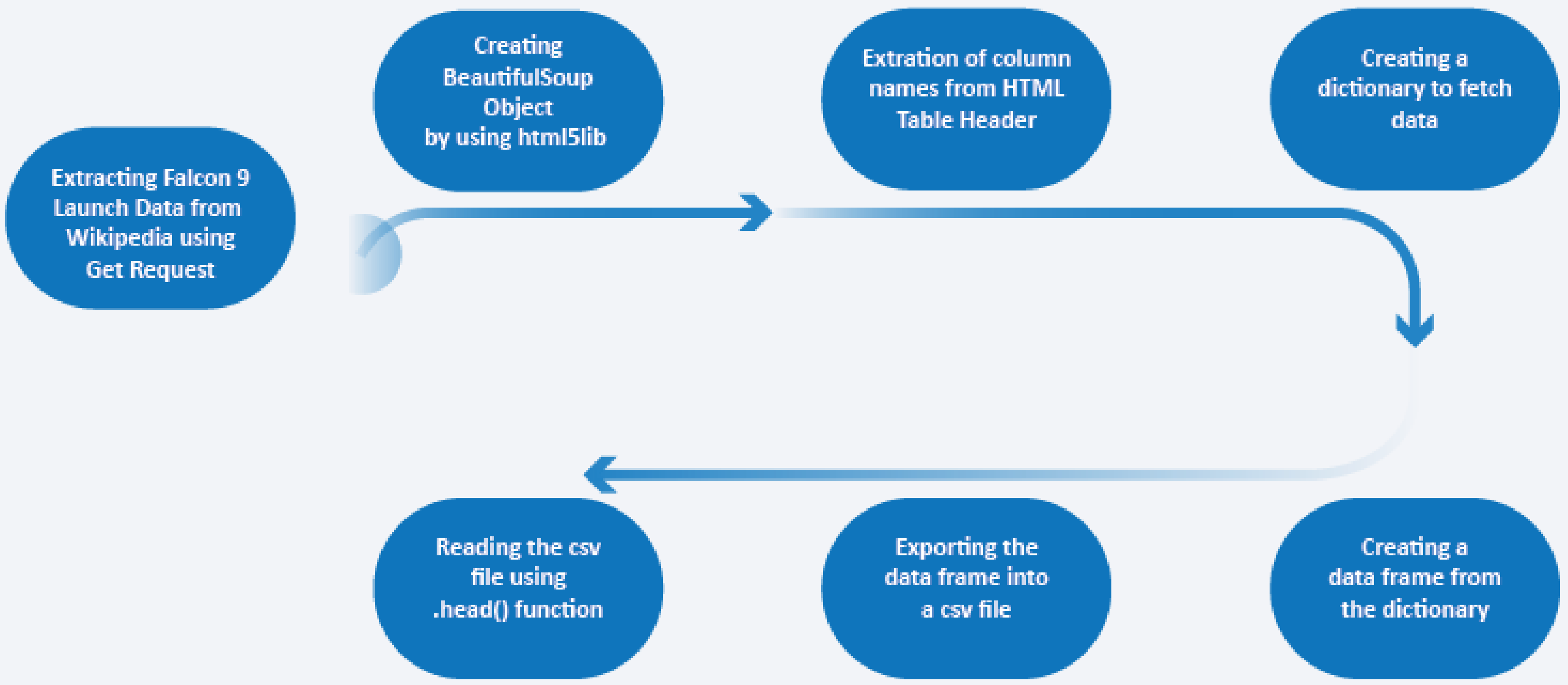
Data Frame Retrieved from SpaceX API

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.561857
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.561857
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.632093
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577366	28.561857
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
89	86	2020-09-03	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	2	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	12	B1060	-80.603956	28.608058
90	87	2020-10-06	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	3	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	13	B1058	-80.603956	28.608058
91	88	2020-10-18	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	6	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	12	B1051	-80.603956	28.608058
92	89	2020-10-24	Falcon 9	15600.0	VLEO	CCSFS SLC 40	True ASDS	3	True	True	True	5e9e3033383ecbb9e534e7cc	5.0	12	B1060	-80.577366	28.561857
93	90	2020-11-05	Falcon 9	3681.0	MEO	CCSFS SLC 40	True ASDS	1	True	False	True	5e9e3032383ecb6bb234e7ca	5.0	8	B1062	-80.577366	28.561857

90 rows × 17 columns

GitHub URL: https://github.com/mihtishamu/DS-Capstone/blob/main/Lab%201%20Data%20Collection%20SpaceX%20API.ipynb

Data Collection - Scraping



- The data was obtained from Wikipedia SpaceX page by using Get Request.
- First a BeautifulSoup Object was created and later a dictionary was created to be filled in with data.
- The dictionary was converted into a data frame.
- The data frame was converted into cvs file and the first five rows of were displayed in a table form by using .head() method.

GitHub URL: https://github.com/mihtishamu/DS-Capstone/blob/main/Lab%202%20Web%20Scraping.ipynb

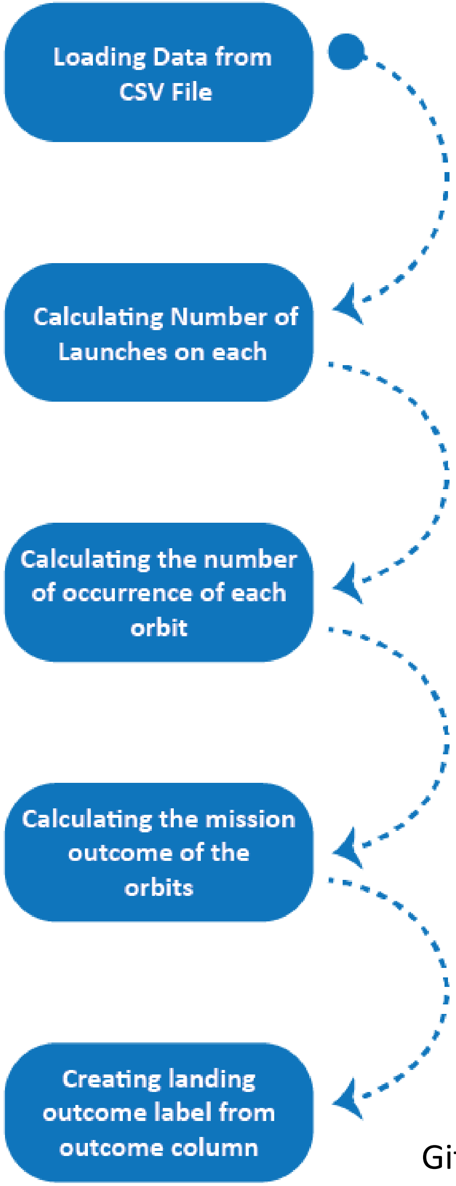
Data Collection – Scraping

Data Frame Retrieved from Wikipedia Using Web Scraping

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	1	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.0B0003.1	Failure	4 June 2010	18:45
2	2	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.0B0004.1	No attempt\n	8 December 2010	15:43
3	3	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.0B0005.1	No attempt	22 May 2012	07:44
4	4	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.0B0006.1	No attempt\n	8 October 2012	00:35

GitHub URL: https://github.com/mihtishamu/DS-Capstone/blob/main/Lab%202%20Web%20Scraping.ipynb

Data Wrangling



- The data was recovered from the previously saved csv file using pandas read csv function.
- The number of missing values in each column were calculated by using isnull().sum() function.
- The type of data of each column was identified using dtype function.
- The data clearly indicated three different launch sites (CCAFS SLC 40, KSC LC 39A, VAFB SLC 4E) and eleven different orbits (GTO, ISS, VLEO, PO, LEO, SSO, MEO, ES-L1, HEO, SO GEO) for Falcon 9 rockets. The number of launches for each site and the number and occurrence of each orbit were estimated by using value_counts() function.
- The number and occurrence of mission outcome of the orbits was estimated and a landing outcome label from Outcome column was created.

GitHub URL: https://github.com/mihtishamu/DS-Capstone/blob/main/Lab%203%20Data%20Wrangling.ipynb

Data Wrangling

Number of Launches on Each Site	
Cape Canaveral Space Launch Complex 40 **(CCAFS SLC 40)**	**55**
Kennedy Space Center Launch Complex 39 A **(KSC LC 39A)**	**22**
Vandenberg Air Force Base Space Launch Complex **(VAFB SLC 4E)**	**13**

Number of Landing Outcomes	
Missions successfully landed to a drone ship (True ASDS)	**41**
Missions unsuccessfully landed to a drone ship (False ASDS)	**06**
Missions successfully landed to a ground pad RTLS (True RTLS)	**14**
Missions unsuccessfully landed to a ground pad RTLS (False RTLS)	**01**
Missions successfully landed to a ocean (True Ocean)	**05**
Missions that were unsuccessful to land in ocean (False Ocean)	**02**
Missions that failed to land to a drone ship (None ASDS)	**02**
Missions that did not land at all (None None)	**19**

GitHub URL: https://github.com/mihtishamu/DS-Capstone/blob/main/Lab%203%20Data%20Wrangling.ipynb

EDA with Data Visualization

- The following charts were plotted:

Comparison	Visualization
Flight number vs Launch Site	Scatterplot
Payload vs Launch Site	Scatterplot
Success rate of each orbit type	Bar chart
Flight Number vs Orbit Type	Scatterplot
Payload vs Orbit Type	Scatterplot
Launch success yearly trend	Line Graph

GitHub URL: https://github.com/mihtishamu/DS-Capstone/blob/main/Lab%205%20Exploratory%20Data%20Analysis%20and%20Visualization.ipynb

EDA with SQL

- The following SQL queries were performed:

1. The names of unique launch sites were displayed
2. Five records of launch sites with names starting with ‘CCA’ were displayed
3. The total payload mass that was carried by boosters launched by NASA was displayed
4. The average payload mass carried by the booster version F9 v1.1 was displayed
5. The date of the first successful landing outcome in ground pad was displayed
6. The boosters that successfully landed on drone ship and contained a payload mass between 4000 and 6000 were displayed
7. The total number of successful and failed mission outcomes were counted.
8. The names of all such booster versions which carried maximum payload mass were listed
9. The names of months in year 2015 with failure landing outcomes in drome ship, booster version and launch site were displayed.
10. The counts of landing outcomes were ranked between 2010-06-04 and 2017-03-20 in descending order.

GitHub URL: https://github.com/mihtishamu/DS-Capstone/blob/main/Lab%204%20SQL.ipynb

Build an Interactive Map with Folium

- All the four launch sites (i.e., CCAFS LC-40, CCAFS SLC-40, KSC LC-39A and VAFB SLC-4E) were marked on the map.
- The success and failure of launches for each site were marked on the map. The purpose of mapping the success and failures of each site was to identify the site with the highest success rate.
- The distance between the launch sites and their proximities was calculated and a line was drawn on the map to display the said distance. The purpose of such visualization was to explore and analyze the proximities of the launch sites so that the success or failures of a launch could be related with the proximities around the launch sites.

GitHub URL: https://github.com/mihtishamu/DS-Capstone/blob/main/Lab%206%20Launch%20SIte%20Location%20Folium.ipynb

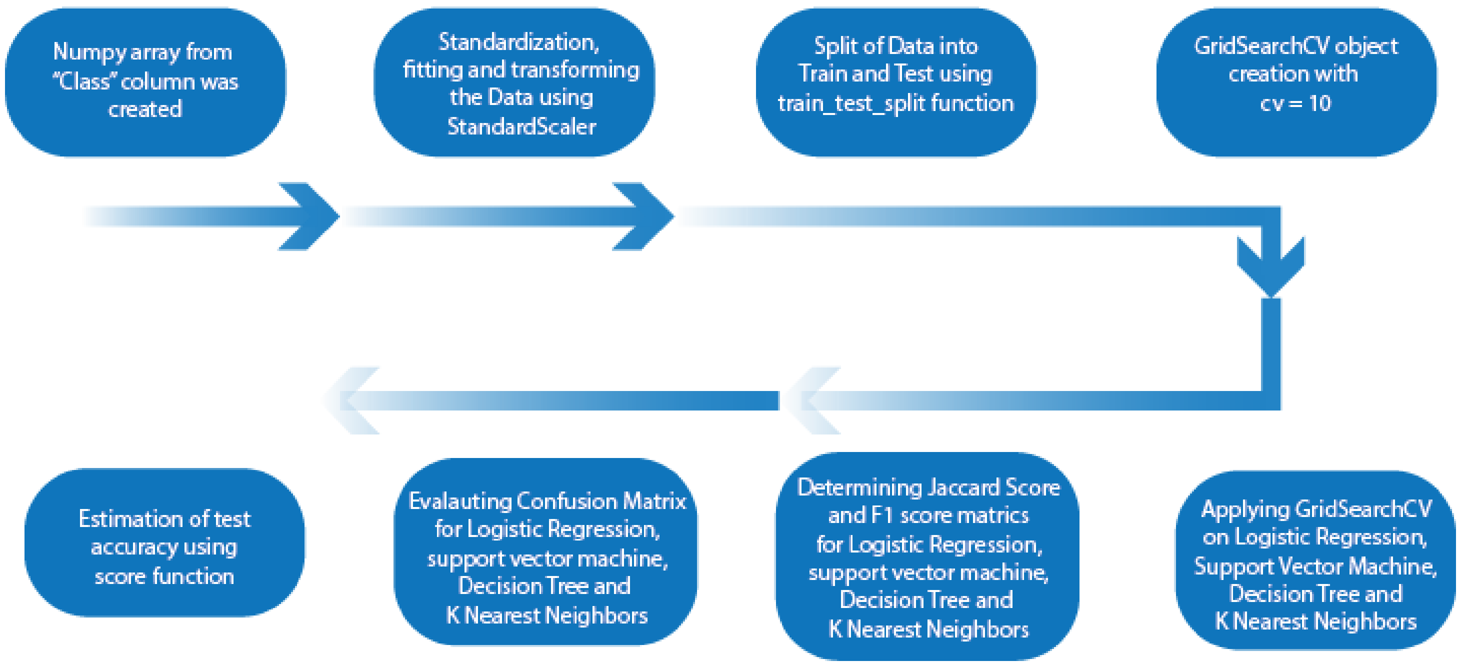
Build a Dashboard with Plotly Dash

- A dropdown list was added to enable the selection of launch sites.
- A pie chart was displayed to depict the count of successful lanches from all different sites.
- A slider was added to select payload range.
- Finally, a scatterplot was added to show correlation between the launch success and payload.

GitHub URL: https://github.com/mihtishamu/DS-Capstone/blob/main/Lab%207%20Plotly%20Dash%20Lab

Predictive Analysis (Classification)

- Since the target variable in this project involved categorical data, we used classification algorithms to predict the target variable.
- For that reason, logistic regression, support vector machine, decision tree and k nearest neighbor models were used.
- The evaluation matrices included Jaccard index, F1 score. Confusion matrix were plotted for all mentioned classification models.
- Finally, the accuracy scores were calculated and the best predictive model was evaluated.



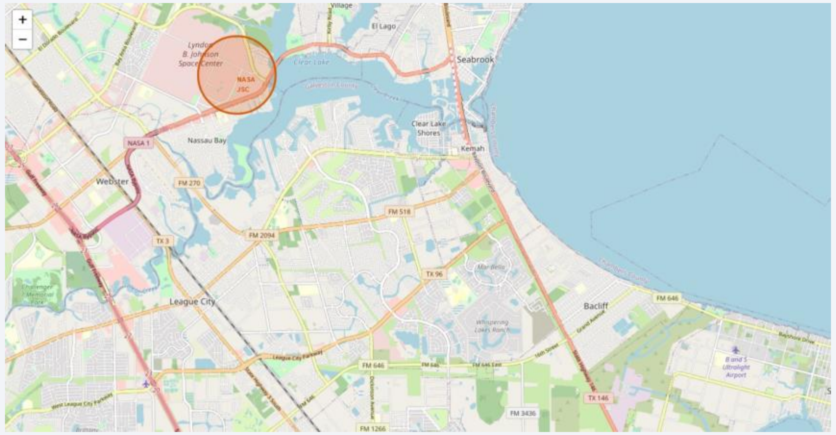
GitHub URL: https://github.com/mihtishamu/DS-Capstone/blob/main/Lab%208%20Machine%20Learning.ipynb

Results

- Exploratory data analysis results

  - The lauch site VAFB SLC 4E showed maximum success rate of 77%, followed by CCAFS LC40 with 60% success rate. Interestingly, VAFB-SLC launch site launched rockets only with a payload mass less than 10000.
  - The success rate was correlated with flight number in case of LEO orbits. However, no such relation was found in case of GTO orbits.
  - The successful landing of rockets with heavy payload mass was found correlated with LEO and ISS orbits.

Results



<u>Interactive analytics demo in screenshots</u>

- Most of the rockets were launched on the site that were located at the eastern cost.

The launch sites were found to be located near sea for more safety and for appropriate infrastructure around

Results

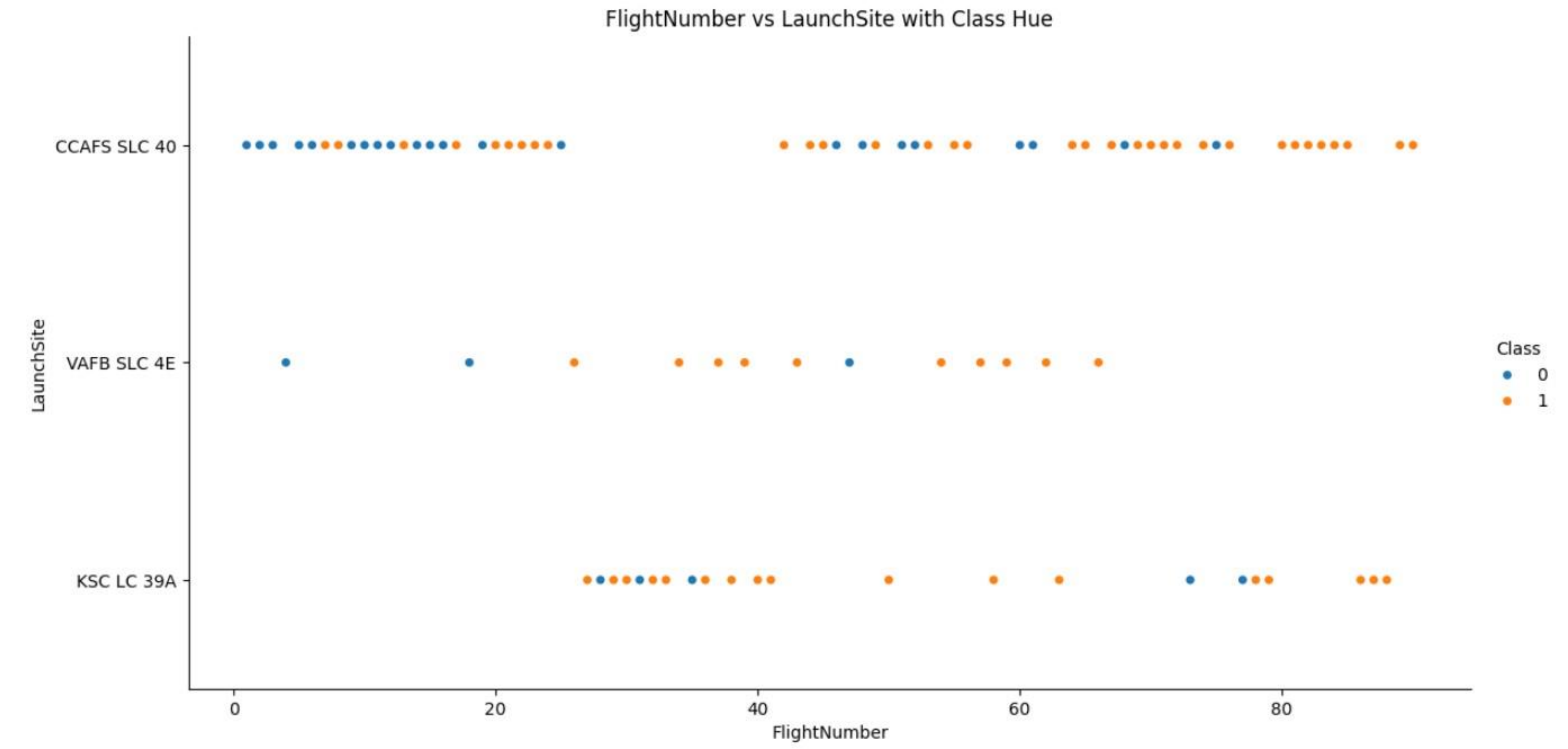
Accuracy Score of Classifiers	
Logistic Regression	0.83
Support vector machine	0.83
Decision tree	0.87
K Nearest Neighbors	0.75

Predictive analysis results

- Decision tree classifier was found to be the best predictive algorithm with a validation score of 0.87.

Section 2

Insights drawn from EDA

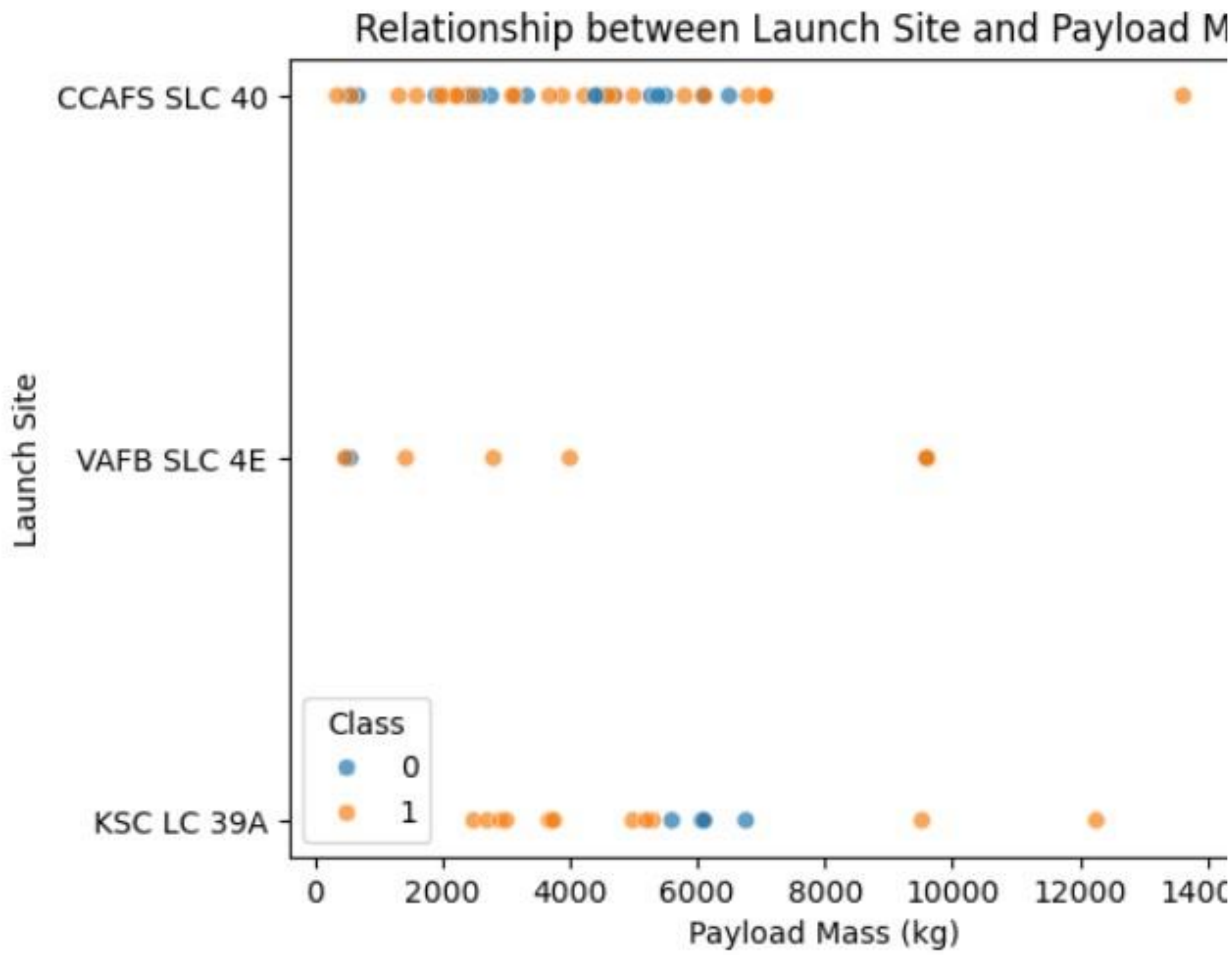


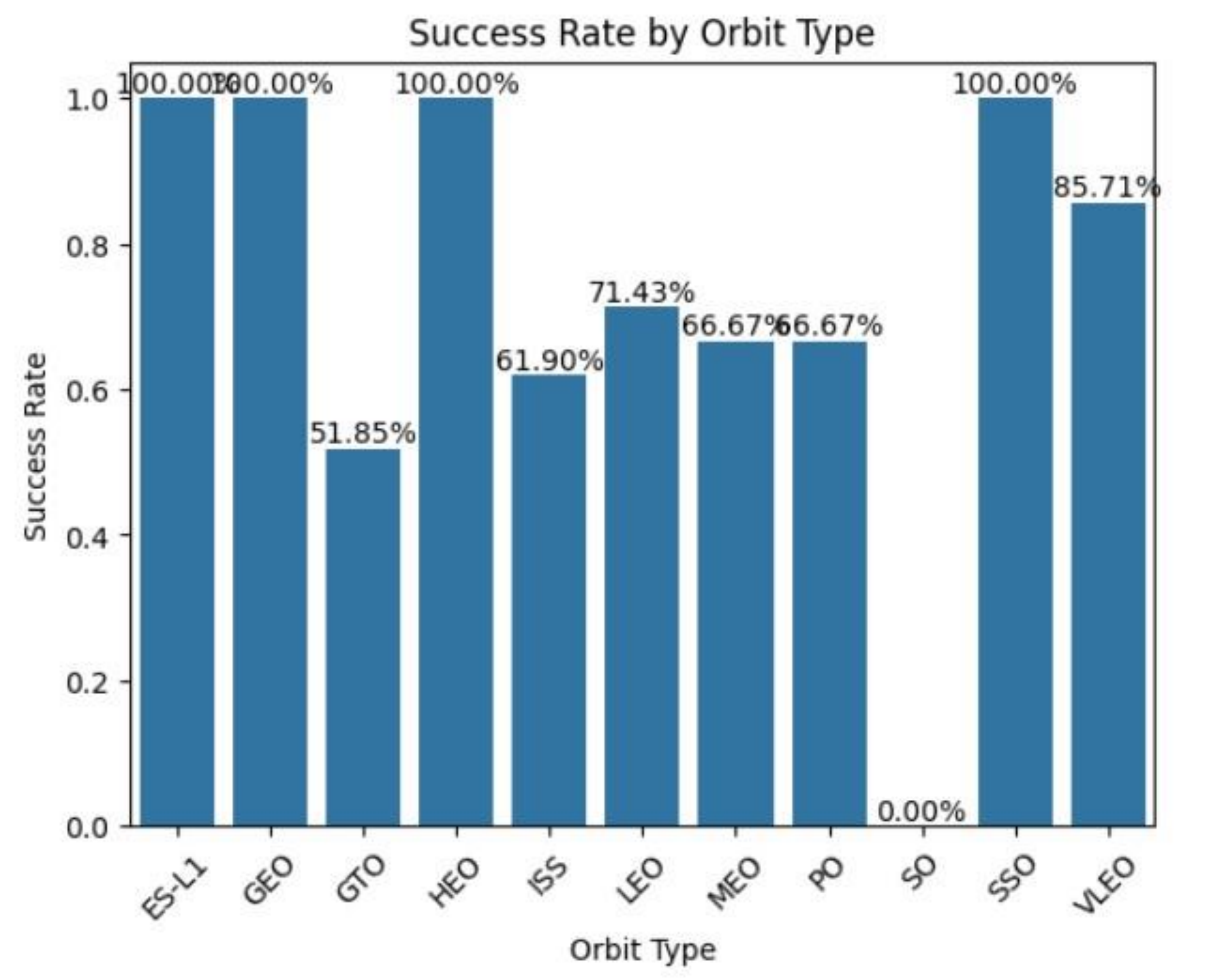
- As per the successful launches, CCAF5 SLC 40 launch site was found to be at the top of the list, followed by VAFB SLC 4E and KSC LC 39 A respectively.

Flight Number vs. Launch Site

Payload vs. Launch Site

- The most successful launches were found to be those with a payload over 9000 kg.
- Only CCAFS SLC 40 and KSC LC 39A launch sites had the ability to carry payloads above 12000 kg.

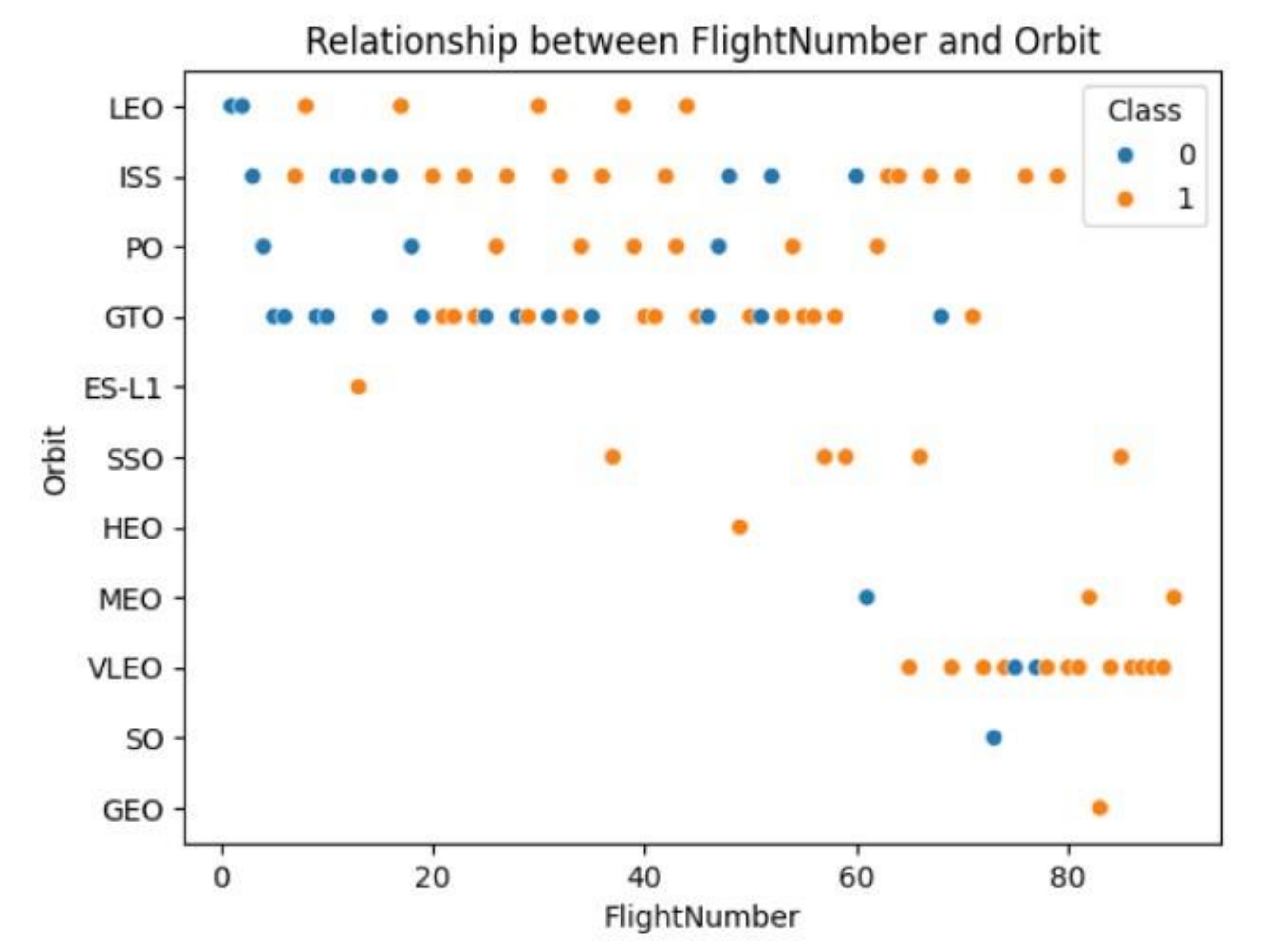




Success Rate vs. Orbit Type

- It is evident from the graph that ES-L1 orbit showed maximum success rate, followed by GEO and HEO respectively.

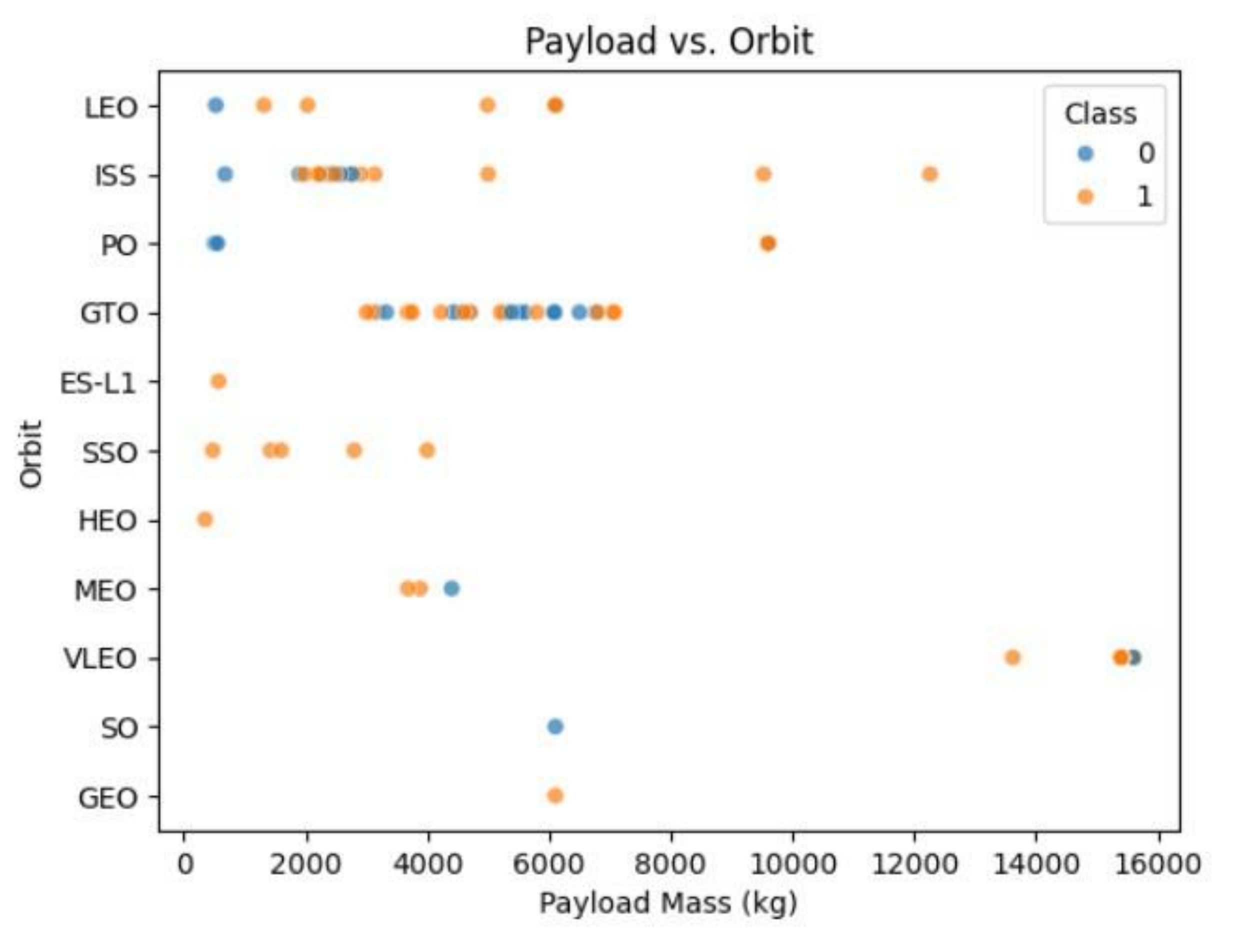
Flight Number vs. Orbit Type



- The success rate of almost all orbits was found to be improved over time.
- A recent increase in VLEO orbit expeditions was observed.

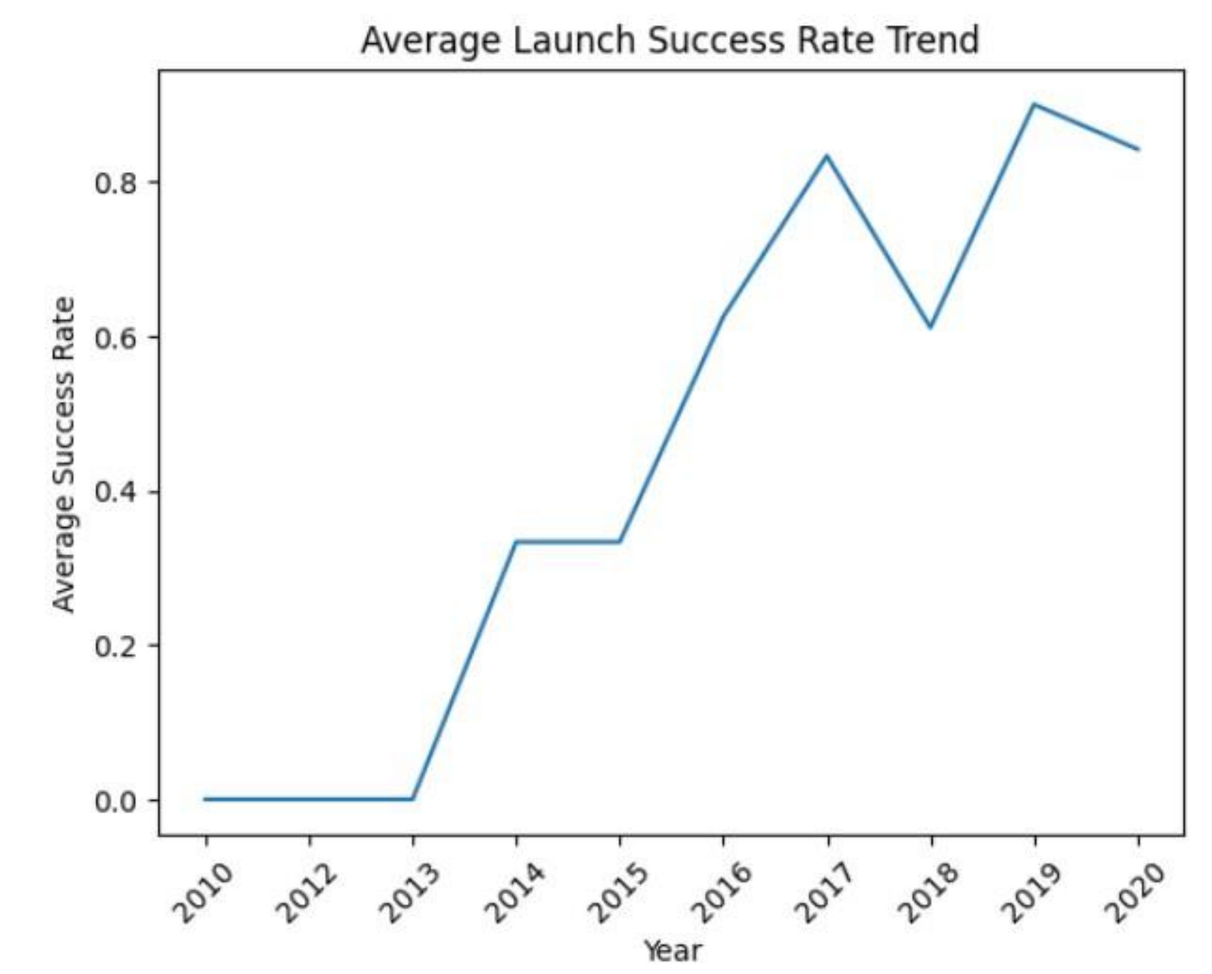
Payload vs. Orbit Type

- The rockets with heavier payloads had successful landings for Polar, LEO and ISS orbits.
- Contrarily, the GTO orbit did not show such correlation between payload and successful landings.



Launch Success Yearly Trend

- There was a significant increase in success rate since 2013.
- The increasing trend was observed until the year 2020.



All Launch Site Names

- The names of the unique launch sites was queried using SQL Distinct function.
- The query fetch four unique launch sites from 'Launch_site' column that are presented below:

Names of Unique Launch Sites	
Cape Canaveral Launch Complex 40	**CCAFS LC 40**
Cape Canaveral Space Launch Complex 40	**CCAFS SLC 40**
Kennedy Space Center Launch Complex 39 A	**KSC LC 39A**
Vandenberg Air Force Base Space Launch Complex	**VAFB SLC 4E**

Launch Site Names Begin with 'CCA'

- Using SQL, five data entries were queried with names that begin with CCA using “CCA% LIMIT 5” code.

Launch Site Nems that begin with CCA

	Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
0	2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	2010-08-12	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2	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
3	2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- The total payload mass carried by NASA (CRS) Boosters was queried using SQL SUM() function. The result was found to be 45596 kg.
- The results are presented below:

```
payload_query = "SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer = 'NASA (CRS)'"
total_payload_mass = cur.execute(payload_query).fetchone()[0]

print("Total Payload Mass Carried by NASA (CRS) Boosters:", total_payload_mass)
```

```
Total Payload Mass Carried by NASA (CRS) Boosters: 45596
```

Average Payload Mass by F9 v1.1

- The average Payload Mass that was carried by the Booster Version v1.1 was queried by SQL AVG() function. The result was found to be 2928.4 kg.

```
# Execute SQL query to display the average payload mass carried by booster version F9 v1.1
average_payload_query = "SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version = 'F9 v1.1'"
average_payload_mass = cur.execute(average_payload_query).fetchone()[0]

print("Average Payload Mass Carried by Booster Version F9 v1.1=", average_payload_mass)
```

```
Average Payload Mass Carried by Booster Version F9 v1.1= 2928.4
```

First Successful Ground Landing Date

- The first successful landing outcome date was queried by SQL using Order By Date with ascending order. The first date was selected using limit 1 function.
- The date was found to be 22nd December 2015, as shown below:

```
# Execute SQL query to retrieve the date of the first successful landing outcome on a ground pad
first_successful_landing_query = "SELECT Date FROM SPACEXTBL WHERE Landing_Outcome = 'Success (ground pad)' ORDER BY Date ASC LIMIT 1"
first_successful_landing_date = cur.execute(first_successful_landing_query).fetchone()[0]

print("Date of First Successful Landing Outcome on Ground Pad:", first_successful_landing_date)
```

```
Date of First Successful Landing Outcome on Ground Pad: 2015-12-22
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters that had successful landings on drone ship and carried payload mass between 4000 and 6000 kg was queried using SQL query.
- The query fetch four boosters' names that are presented below:

Booster Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failed outcomes are presented below:

Outcome	Counts
Success Mission	98
Failure Mission	0

Boosters Carried Maximum Payload

- The names of the booster versions with maximum payload mass were queried using SQL command. The results are presented below:

Booster Version with Maximum Payload Mass
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

- We found two records with failed landing outcomes in drone ship in year 2015. The details of the launch site and boosters is given below:

Booster Version	Launch Site
F9 V1.1 B1012	CCAFS LC 40
F9 V1.1 B1015	CCAFS LC 40

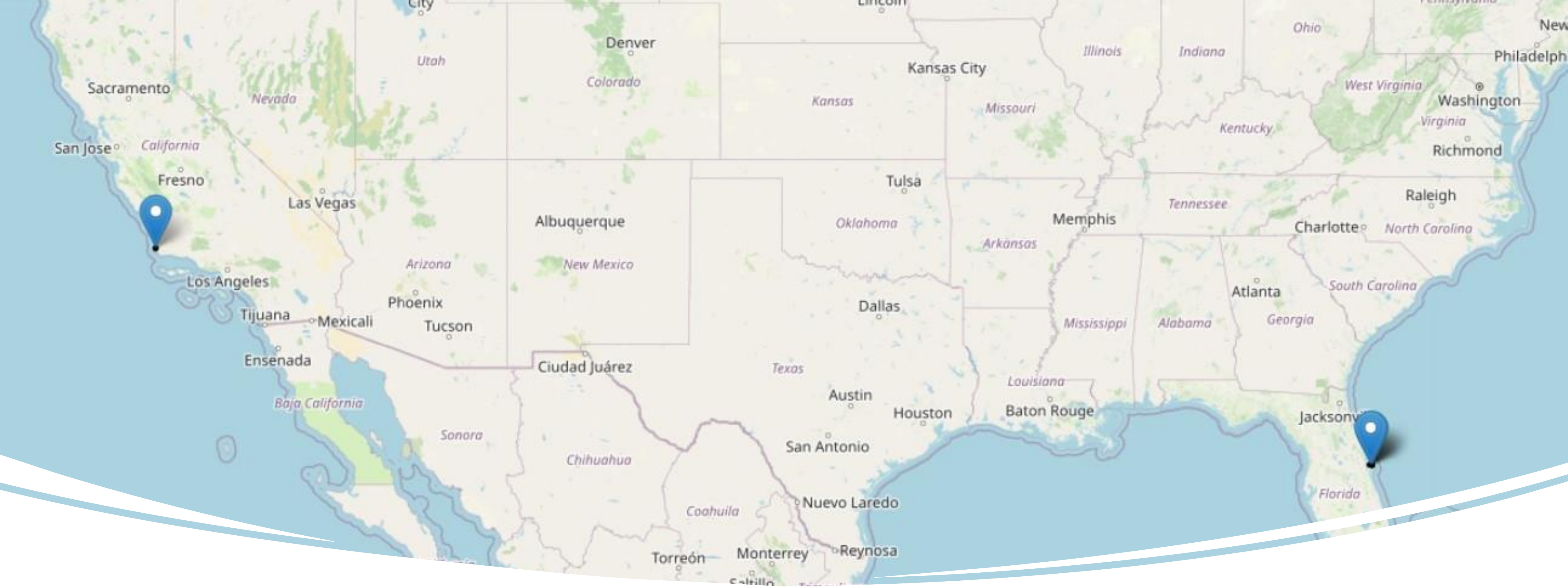
- The rankings of all landing outcomes between 4th July 2010 and 20th March 2017 are presented below.

```
Ranking of Landing Outcomes Count between 2010-06-04 and 2017-03-20:
Rank 1: No attempt - Count: 10
Rank 2: Success (ground pad) - Count: 5
Rank 3: Success (drone ship) - Count: 5
Rank 4: Failure (drone ship) - Count: 5
Rank 5: Controlled (ocean) - Count: 3
Rank 6: Uncontrolled (ocean) - Count: 2
Rank 7: Precluded (drone ship) - Count: 1
Rank 8: Failure (parachute) - Count: 1
```

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

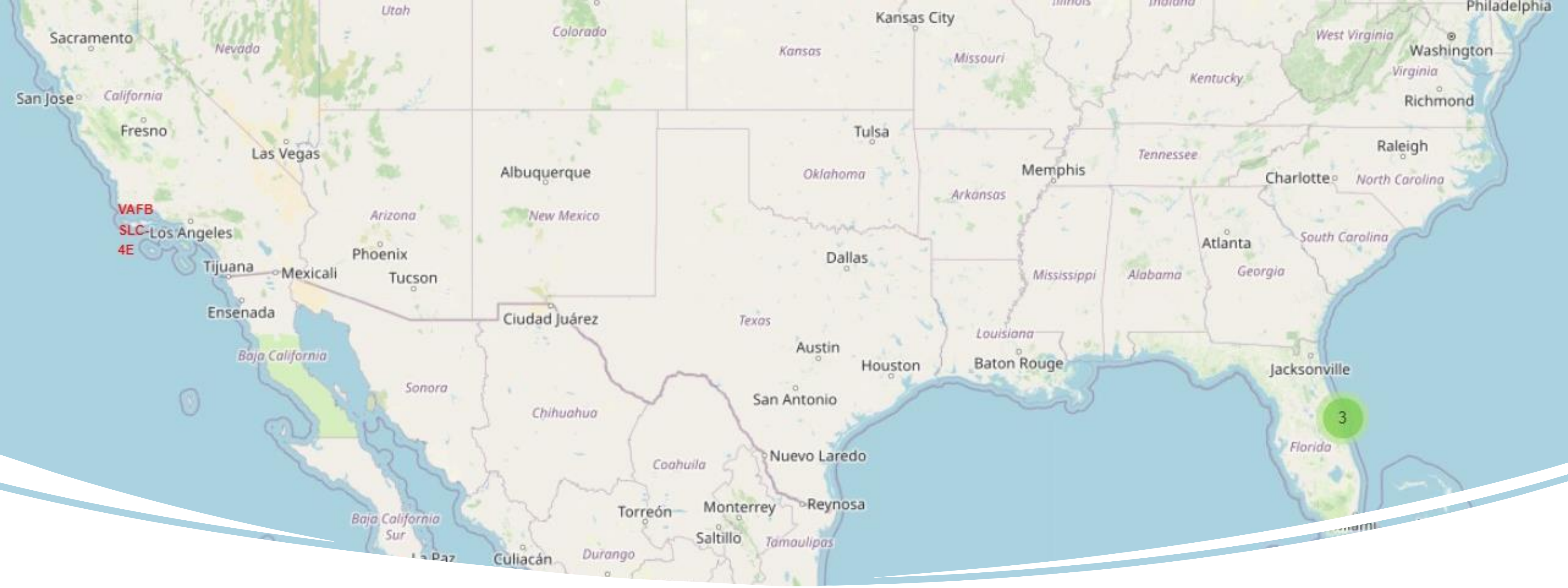
Section 3

Launch Sites Proximities Analysis



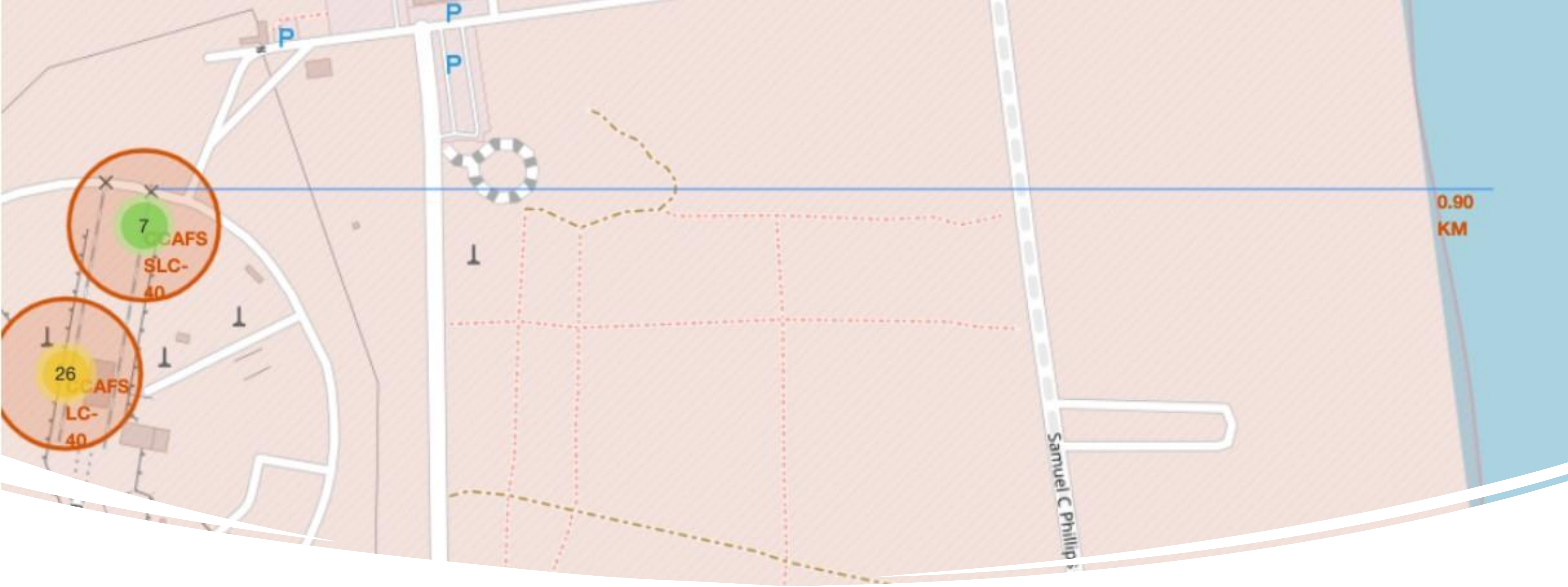
Location of Launch Sites

- For safety measures, the launch sites were found to be near the sea with appropriate logistics and infrastructure.



Launch Outcomes

- The successful launches are presented as green marker whereas the unsuccessful launches are presented as red.



Launch Site Proximities

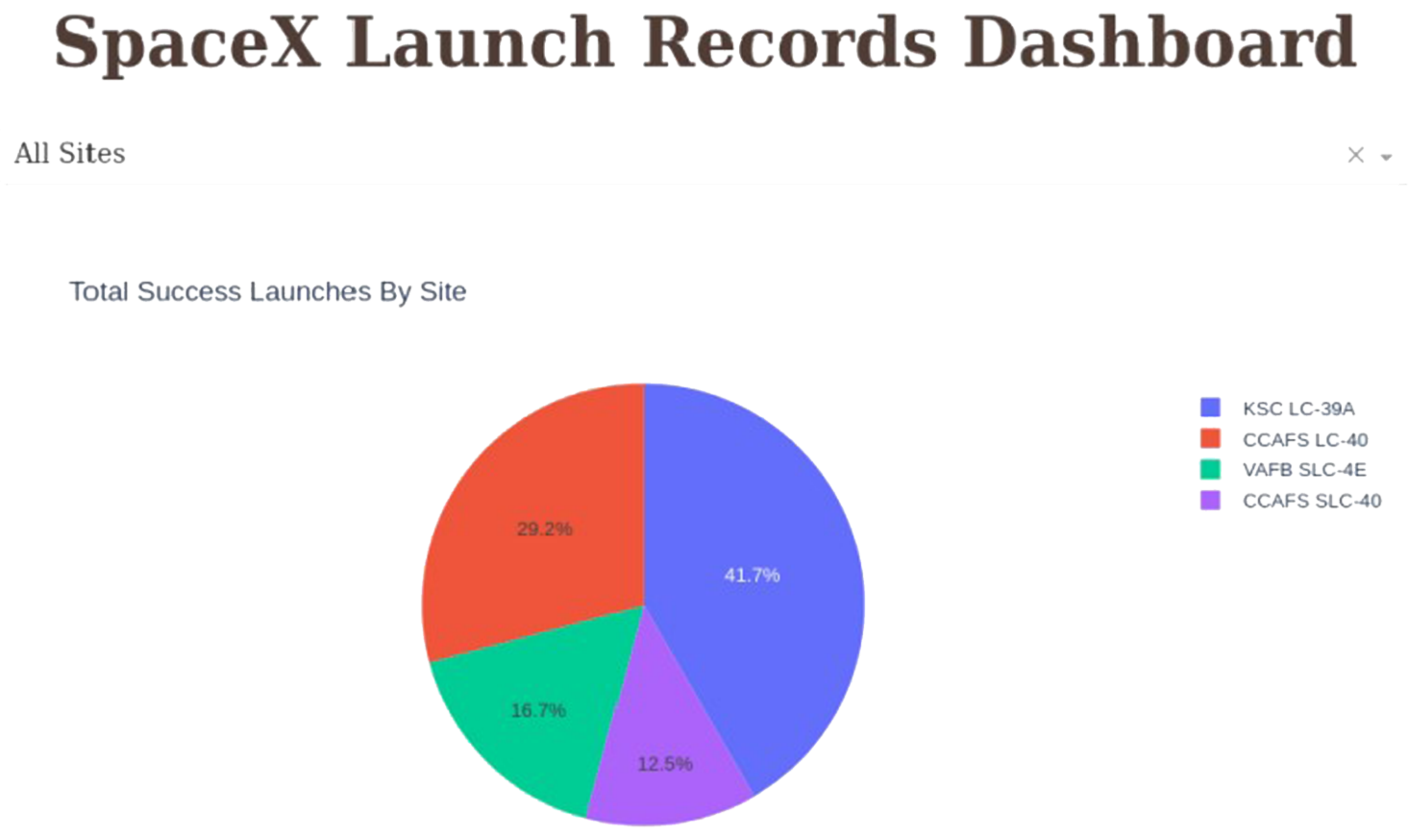
- The infrastructure near the launch site CCAFS SLC 40 is presented in the map. The distance of the launch site from sea is also calculated and presented here.

Section 4

Build a Dashboard with Plotly Dash

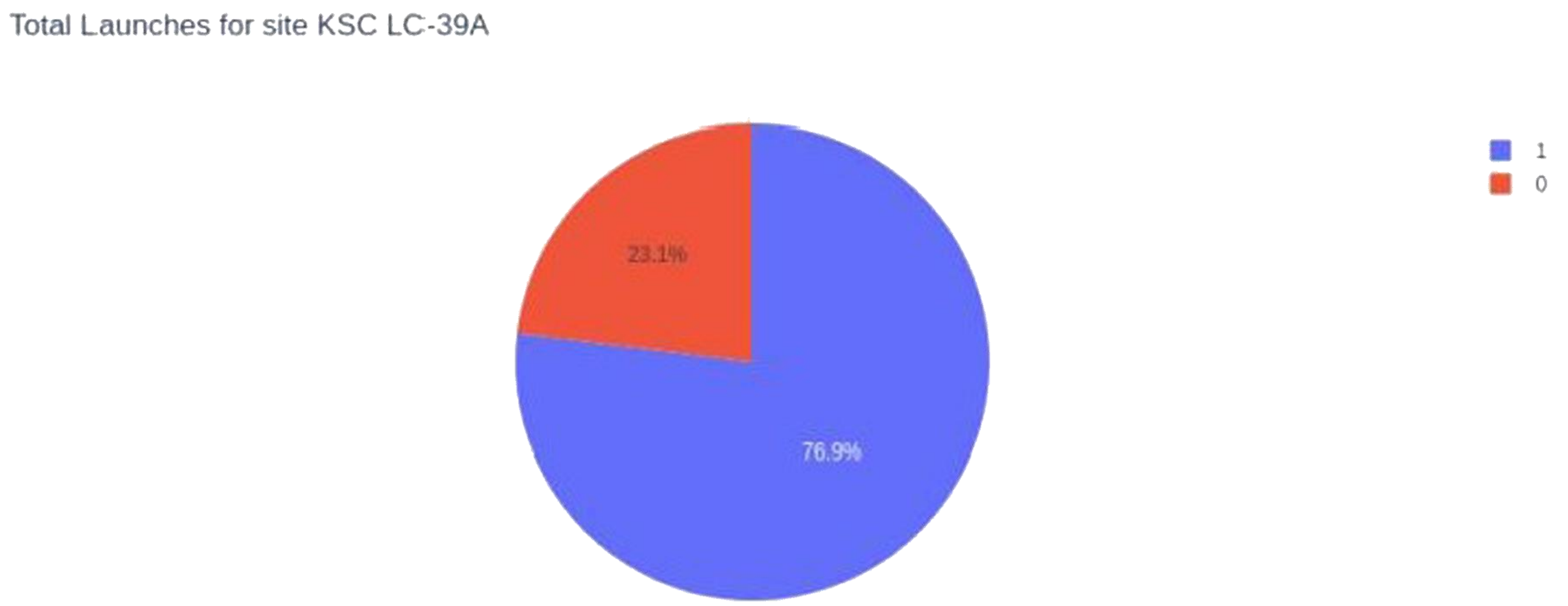
Successful Launch Counts Per Launch Site

- The launch site is one of the most important factors in predicting the successful landing of a launch.

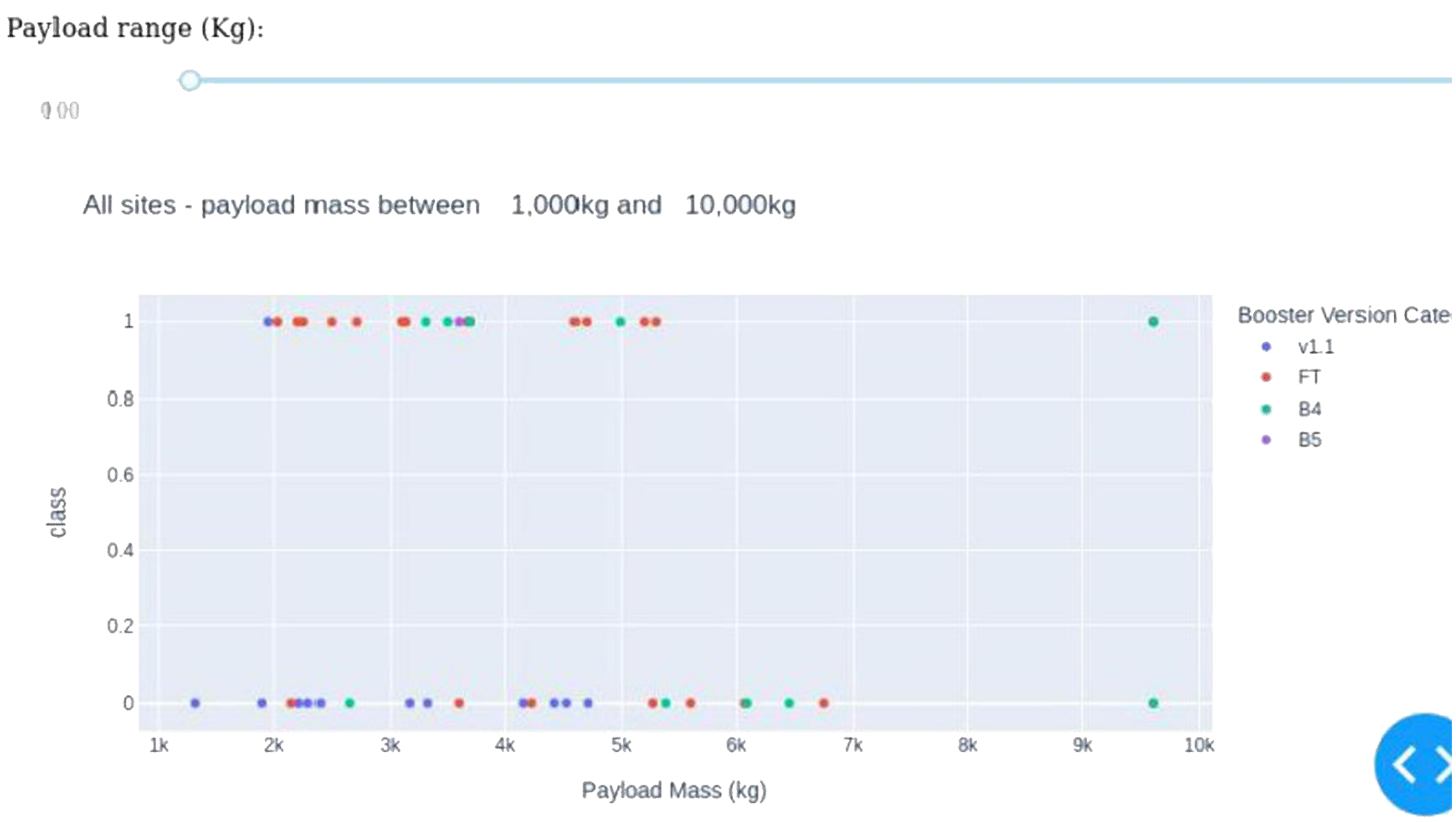


Launch Site With Highest Launch Success Ratio

- The launch site KSC LC-39 A showed maximum successful landings i.e., 76.9%.



Payload Vs Launch Outcome Plot



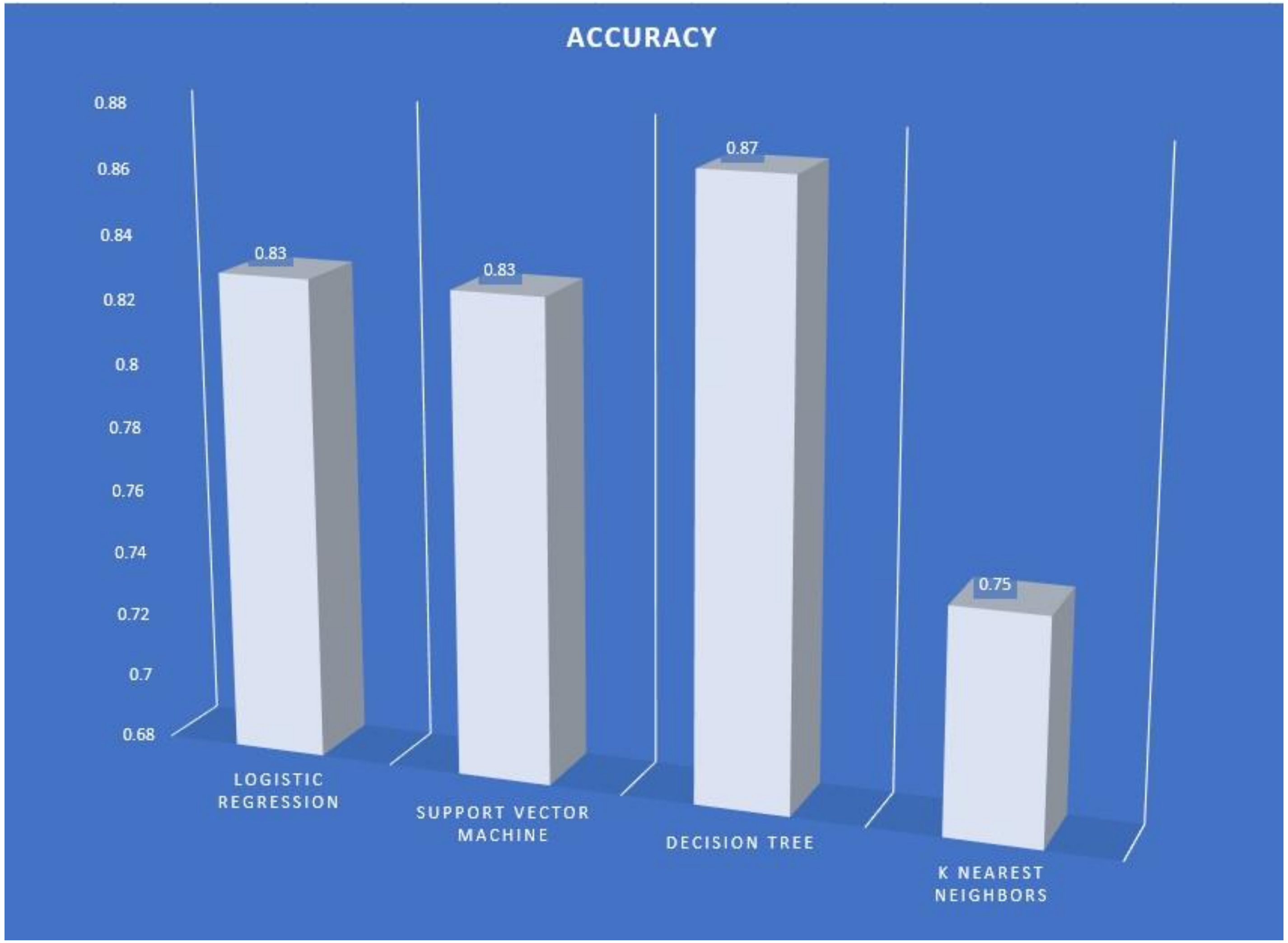
- The FT Boosters with payloads less than 6000 kg were found to be the most successful launches among all.

Section 5

Predictive Analysis (Classification)

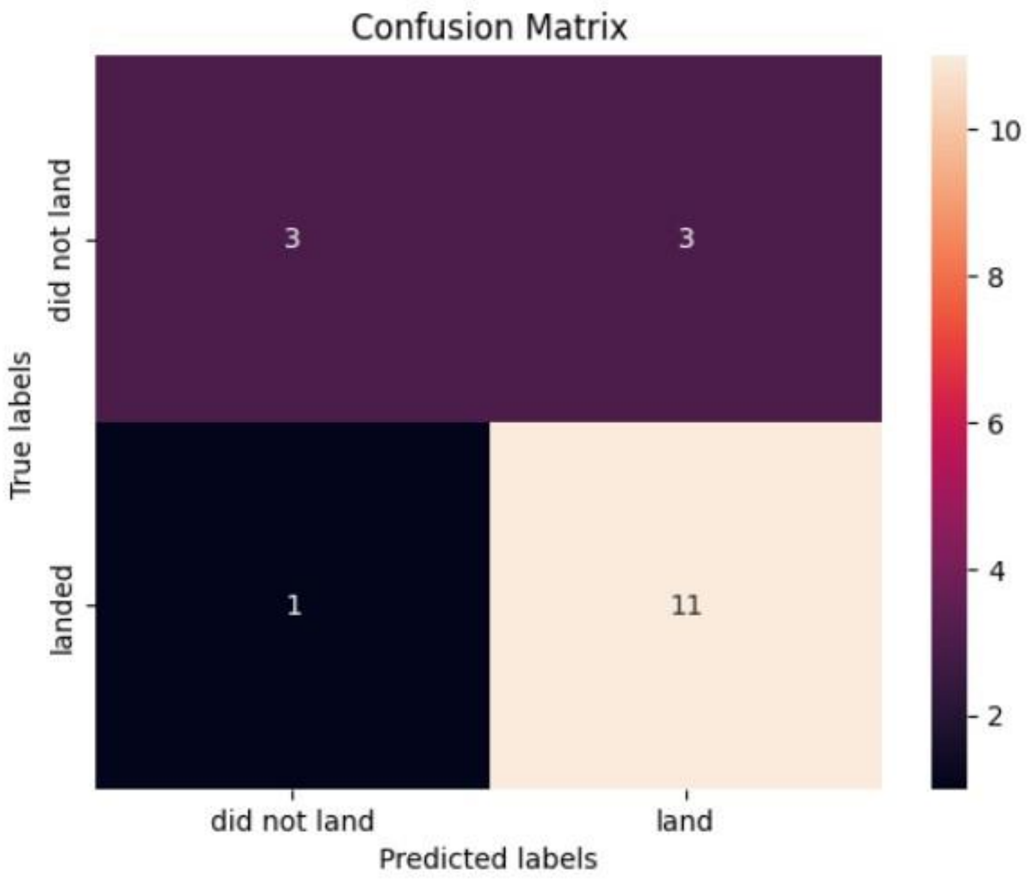
Classification Accuracy

- Decision tree was found to be the best predictive classification model with 87% accuracy.



Confusion Matrix

- The confusion matrix of the best performing model (decision cloud) is presented below.
- The model predicted the test data with 87% accuracy.



Conclusions

The lauch site with most successful missions was found to be KSC LC 39A.

The success rate was found higher for lauches with a payload mass above 7,000 kg.

The rate of successful landings increased over time.

Decision tree was found to be the best predictive classification model.

Thank you!