



IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Siddharth Patondikar
31-01-2023



Outline

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

Executive Summary

- Following methodologies are used to analyze the data:
 - Data collection using SpaceX API and Web Scrapping.
 - Data Wrangling, Exploratory Data Analysis (EDA) using visualization tools and SQL. Interactive visualization using Folium and Plotly Dash.
 - Predictive analysis using machine learning.
- Summary of all results:
 - Data collection was possible due to availability on public sources.
 - EDA allowed to identify best features for predicting successful launches.
 - Machine learning helped finding best classification models for predicting launches from total acquired data.

Introduction

- The objective of this report is to evaluate the price of launch for a company Space Y and the viability of the company to compete with Space X.
- Possible best ways to evaluate price of launch is:
 - Predicting successful landing of the first stage
 - Location of launch

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
 - SpaceX data was collected using two ways:
 - SpaceX API (<https://api.spacexdata.com/v4/launchpads/>)
 - Web Scrapping
(https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)
- Perform data wrangling
 - Analyzed mission outcome per orbit
 - A new landing label was created to enrich data based on outcome data.

Methodology

Executive Summary

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
 - After this stage, data was standardized, divided into training and test data and then tested on four machine learning classification models.
 - The accuracy of all models were calculated using combinations of hyperparameters and the model with best accuracy was selected.

Data Collection

Data collection was done using public sources.

SpaceX data was collected using two ways:

- SpaceX API (<https://api.spacexdata.com/v4/launchpads/>)
- Web Scrapping
(https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)


Data Collection – SpaceX API

- SpaceX provides its data to public through its API.
- The API was used as shown in the flowchart to obtain the data.

- Source Code:

[https://github.com/sidpatondikar/Applied-Data-Science-Capstone/blob/main/Data%20Collection%20\(SpaceX%20API\).ipynb](https://github.com/sidpatondikar/Applied-Data-Science-Capstone/blob/main/Data%20Collection%20(SpaceX%20API).ipynb)

Request and parse the SpaceX launch data using the GET request



Filter the data frame to only include Falcon 9 launches



Dealing with Missing Values

Data Collection - Scraping

- Data is obtained from Wikipedia page containing Falcon 9 launches.
- Data is obtained through web scrapping as shown in flowchart

- Source Code:

[https://github.com/sidpatondikar/Applied-Data-Science-Capstone/blob/main/Data%20Collection%20\(Web%20Scrapping\).ipynb](https://github.com/sidpatondikar/Applied-Data-Science-Capstone/blob/main/Data%20Collection%20(Web%20Scrapping).ipynb)

Request the Falcon9 Launch Wiki page from its URL



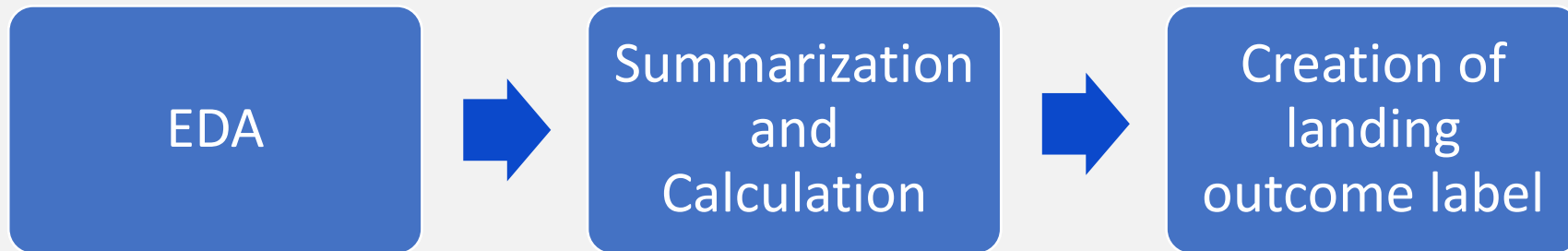
```
graph TD; A[Request the Falcon9 Launch Wiki page from its URL] --> B[Extract all column/variable names from the HTML table header]; B --> C[Create a data frame by parsing the launch HTML tables];
```

Extract all column/variable names from the HTML table header

Create a data frame by parsing the launch HTML tables

Data Wrangling

- Initially some exploratory data analysis (EDA) was done on the data.
- Then summarizations were made for launches per site, occurrence of each orbit.
- Then calculations were made for mission outcome per orbit.
- Finally a landing outcome label was created to enrich the data.



- Source Code: <https://github.com/sidpatondikar/Applied-Data-Science-Capstone/blob/main/Data%20wrangling.ipynb>

EDA with Data Visualization

- To explore data various bar plots, scatter plots and line charts were visualized for various features in data.
- Following charts were visualized:
 - Launch Site v Flight Number
 - Payload Mass v Launch Site
 - Success rate of each orbit
 - Flight Number v Orbit type
 - Payload v Orbit type
 - Launch Success yearly trend

Source Code: <https://github.com/sidpatondikar/Applied-Data-Science-Capstone/blob/main/EDA%20using%20Data%20Visualization.ipynb>

EDA with SQL

- The following SQL queries were performed:
 - Names of the unique launch sites in the space mission;
 - Top 5 launch sites whose name begin with the string 'CCA';
 - Total payload mass carried by boosters launched by NASA (CRS);
 - Average payload mass carried by booster version F9 v1.1;
 - Date when the first successful landing outcome in ground pad was achieved;
 - Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg;
 - Total number of successful and failure mission outcomes;
 - Names of the booster versions which have carried the maximum payload mass;
 - Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015; and
 - Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.

Source Code: <https://github.com/sidpatondikar/Applied-Data-Science-Capstone/blob/main/EDA%20using%20SQL.ipynb>

Build an Interactive Map with Folium

- Markers, circles, lines and marker clusters were used with Folium Maps
 - Markers indicate points like launch sites;
 - Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center;
 - Marker clusters indicates groups of events in each coordinate, like launches in a launch site; and
 - Lines are used to indicate distances between two coordinates.

Source Code: <https://github.com/sidpatondikar/Applied-Data-Science-Capstone/blob/main/Interactive%20Visualization%20using%20Folium.ipynb>

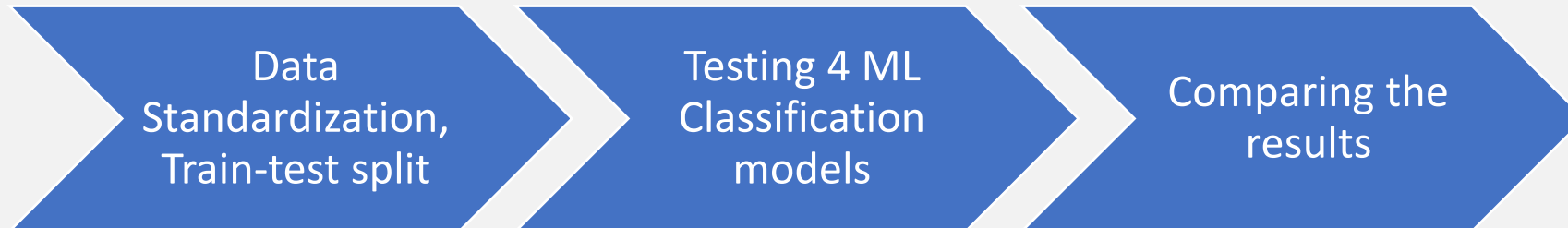
Build a Dashboard with Plotly Dash

- The following graphs and plots were used to visualize data
 - Percentage of launches by site
 - Payload range
 - This combination allowed to quickly analyze the relation between payloads and launch sites, helping to identify where is best place to launch according to payloads.

Source Code : https://github.com/sidpatondikar/Applied-Data-Science-Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Data Standardization, splitting into train and test group were done. Then 4 ML classification models were tested on data and the results were compared.
- The 4 models tested are: Logistic Regression, KNN, SVM and Decision Tree Classifier



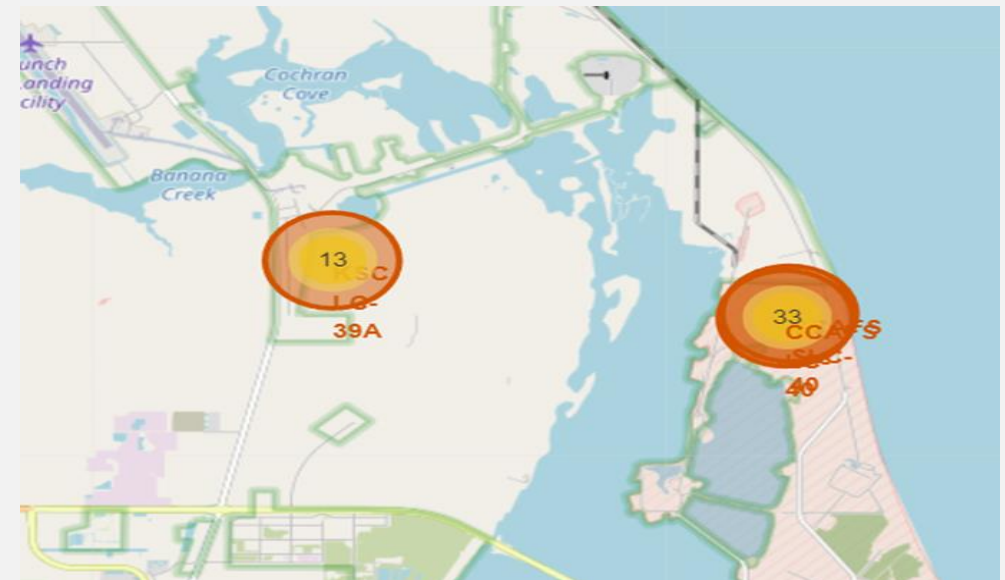
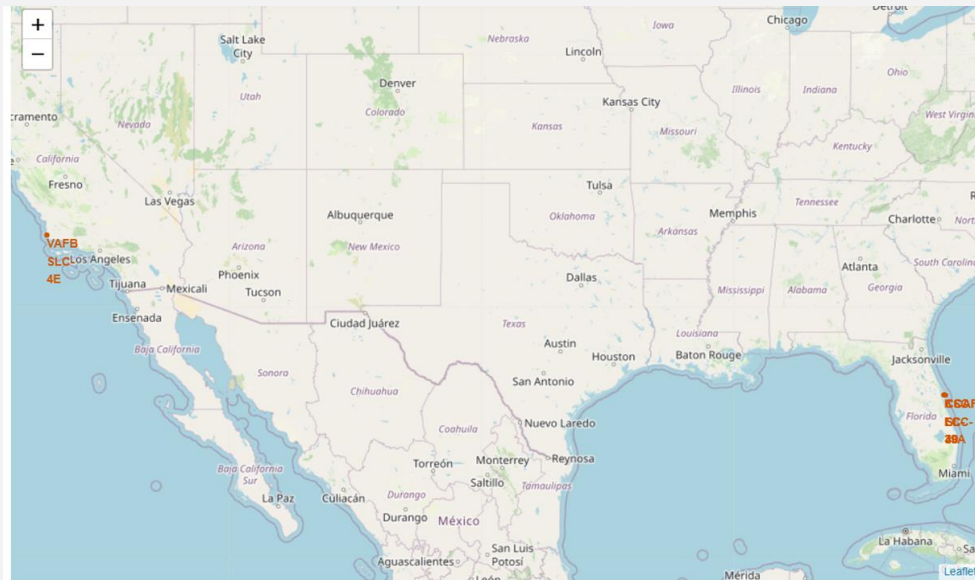
- Source code : https://github.com/sidpatondikar/Applied-Data-Science-Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction.ipynb

Results

- Exploratory data analysis (EDA) Results:
 - There are 4 unique launch sites.
 - The earliest launches were for NASA and SpaceX itself.
 - Average payload mass carried by F9 v1.1 was 2928 KG.
 - The first successful landing was achieved in 2015.
 - 99% of total mission outcomes were success.
 - Two booster version had failed landing in 2015: F9 v1.1 B1012 and F9 v1.1 B1015.
 - Successful launches per year grew from 2010 till 2020.

Results

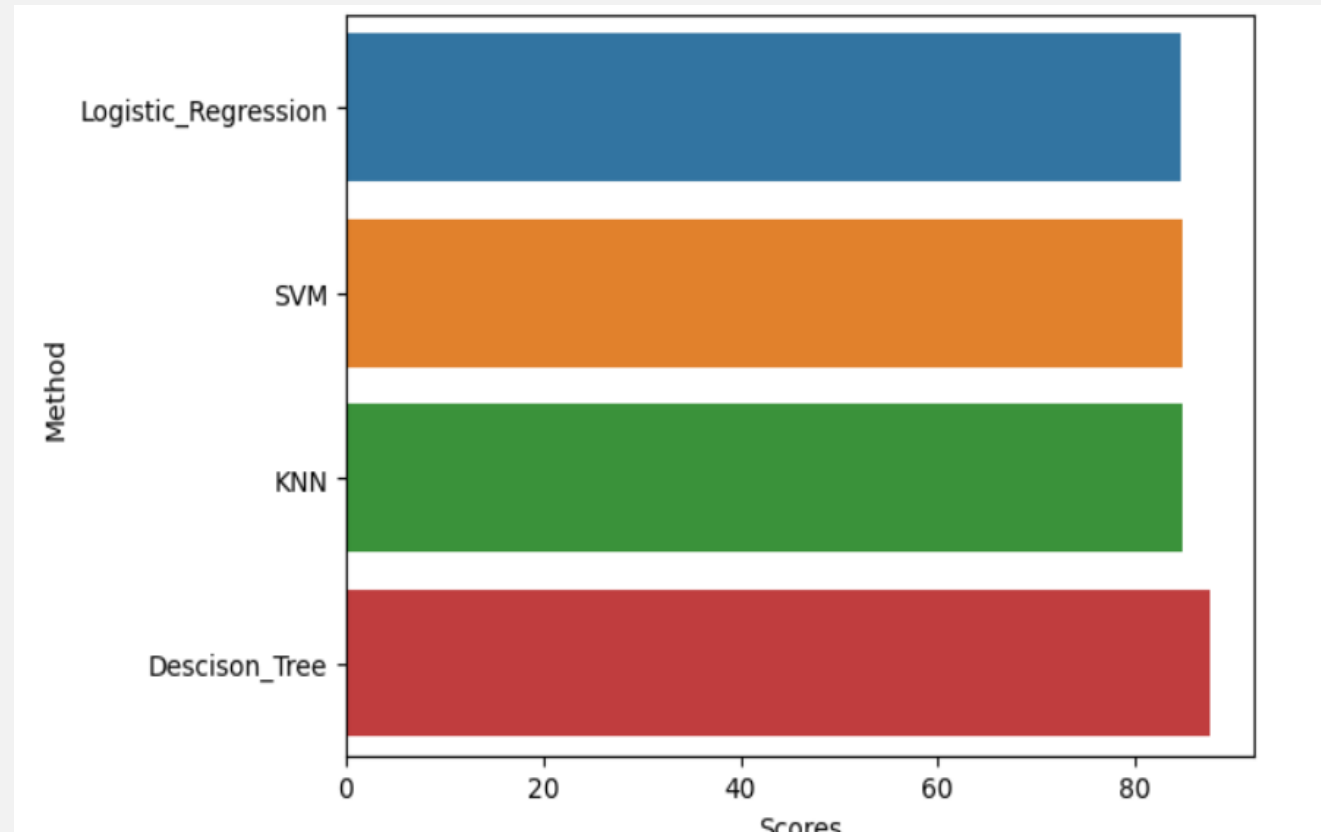
- Interactive analytics demo in screenshots
 - All of the launch sites are located on coasts for safety.
 - However, they are close to road and rail for connectivity.



Results

Predictive analysis results:

- 4 ML classification models are deployed to get the best model and outcome.
- Decision Tree Classifier gives the best accuracy of 87.7% and thus can be deployed to gain maximum profits.



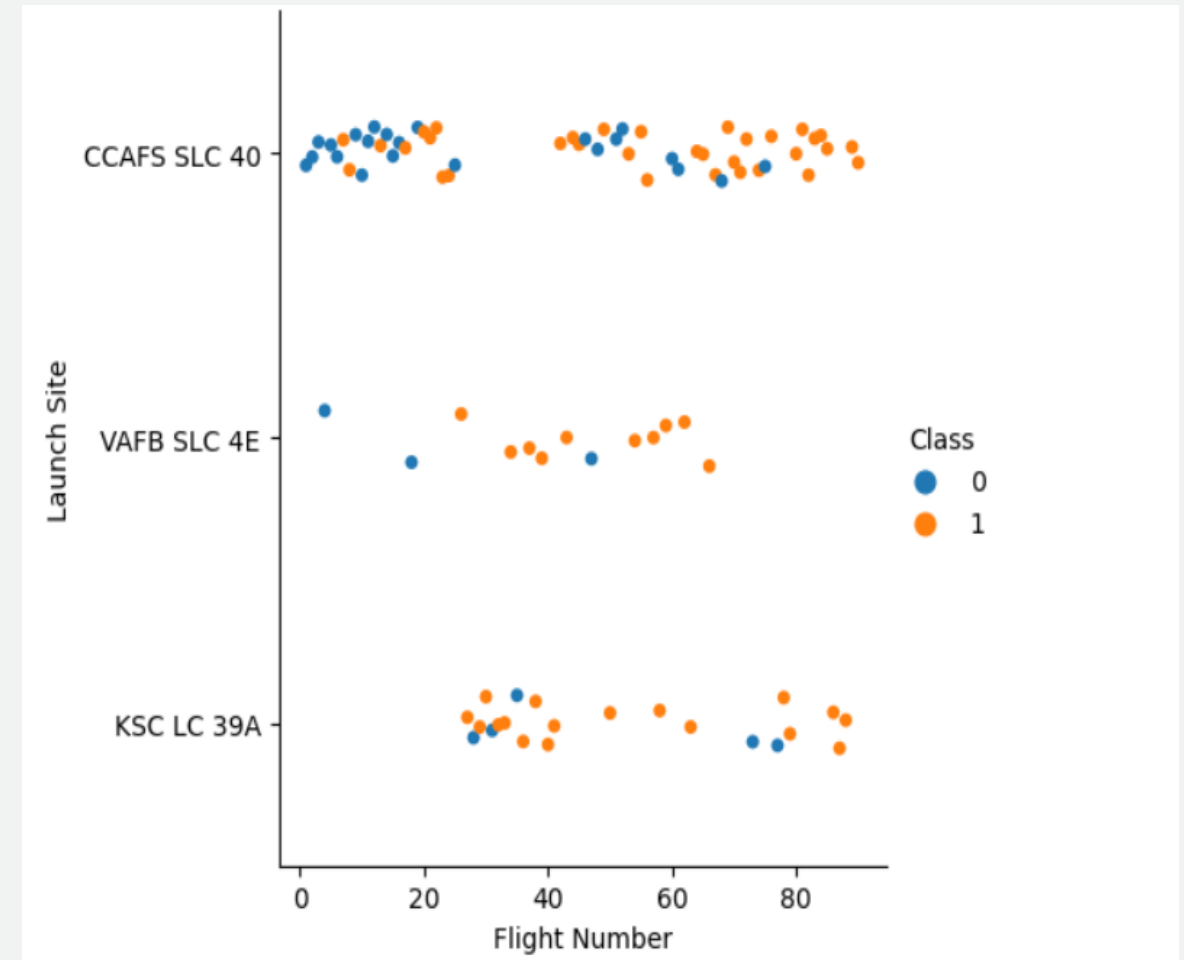
The background of the slide is an abstract composition. It features a dark blue base color. Overlaid on this are numerous diagonal streaks in shades of red and cyan. A faint, light blue grid pattern is also visible, particularly in the lower-left quadrant. The overall effect is dynamic and technological.

Section 2

Insights drawn from EDA

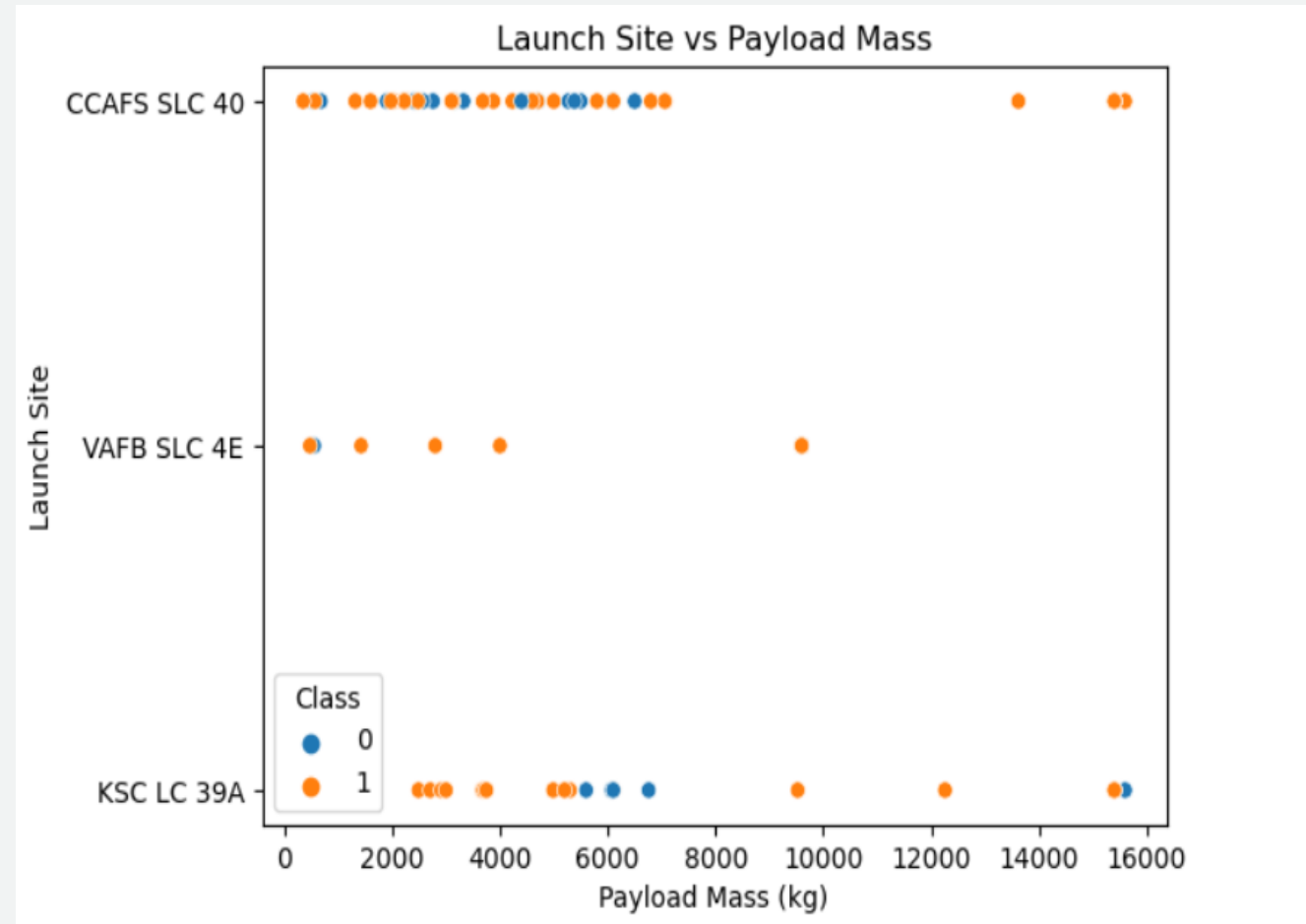
Flight Number vs. Launch Site

- We can observe for CCAFS SLC 40, success of launches increased with time.
- KSC LC 39A was not used for the first 20 flights. Its success rate didn't change much over time.
- From this we can conclude that CCAFS SLC 40 is the best launch site followed by VAFB SLC 4E and KSC LC 39A.



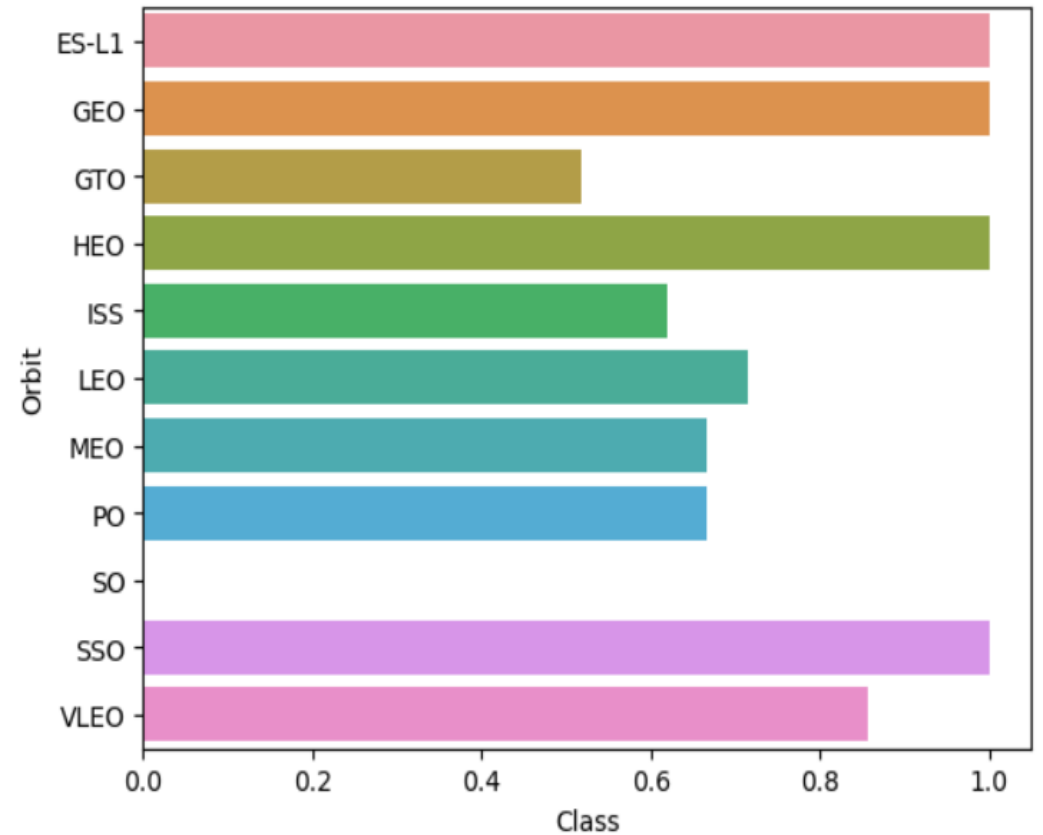
Payload vs. Launch Site

- Payload over 9000 KG has excellent success rate as observed from the plot.
- For payload over 12000 KG, only CCAFS SLC 40 and KSC LC 39A are used.



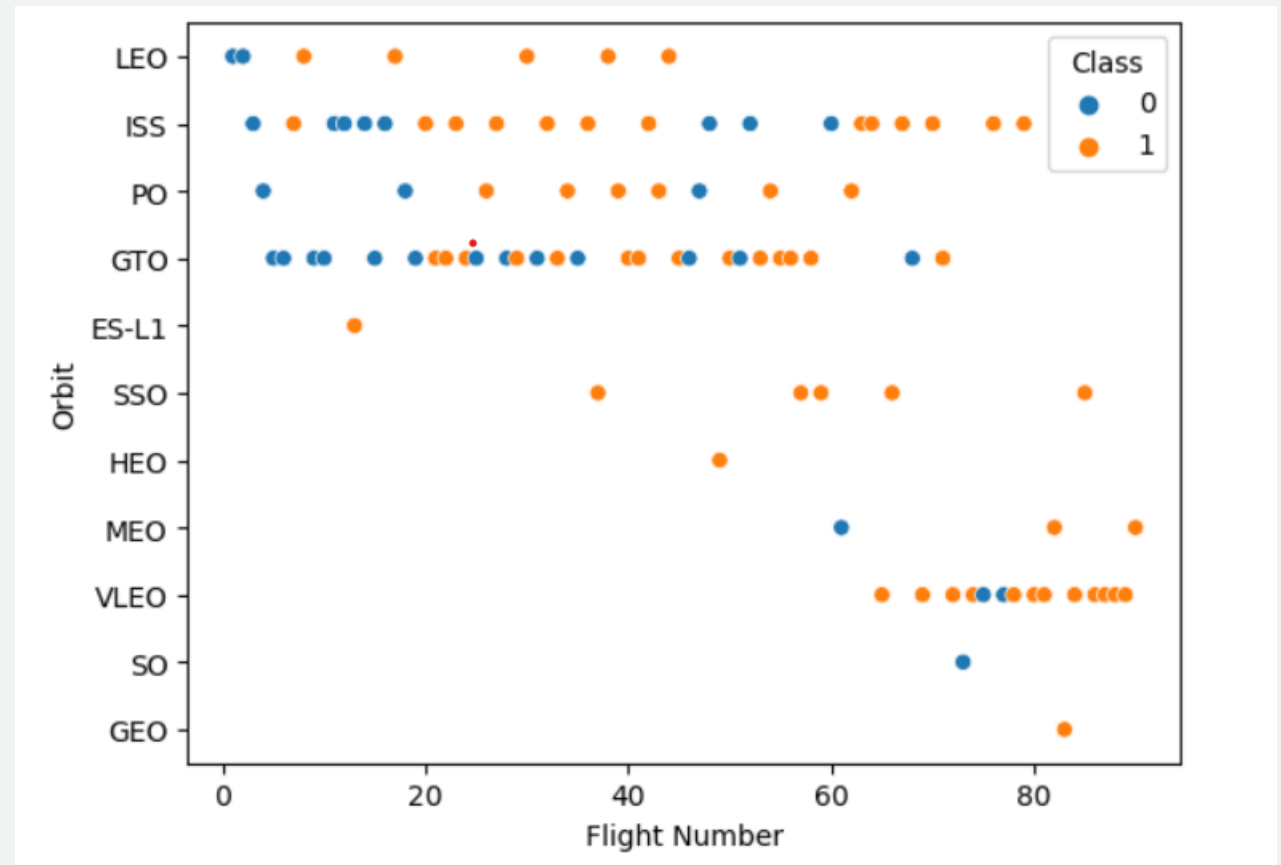
Success Rate vs. Orbit Type

- The most successful orbits for launch are:
 - ES-L1
 - GEO
 - HEO
 - SSO
- The least successful orbit is:
 - GTO



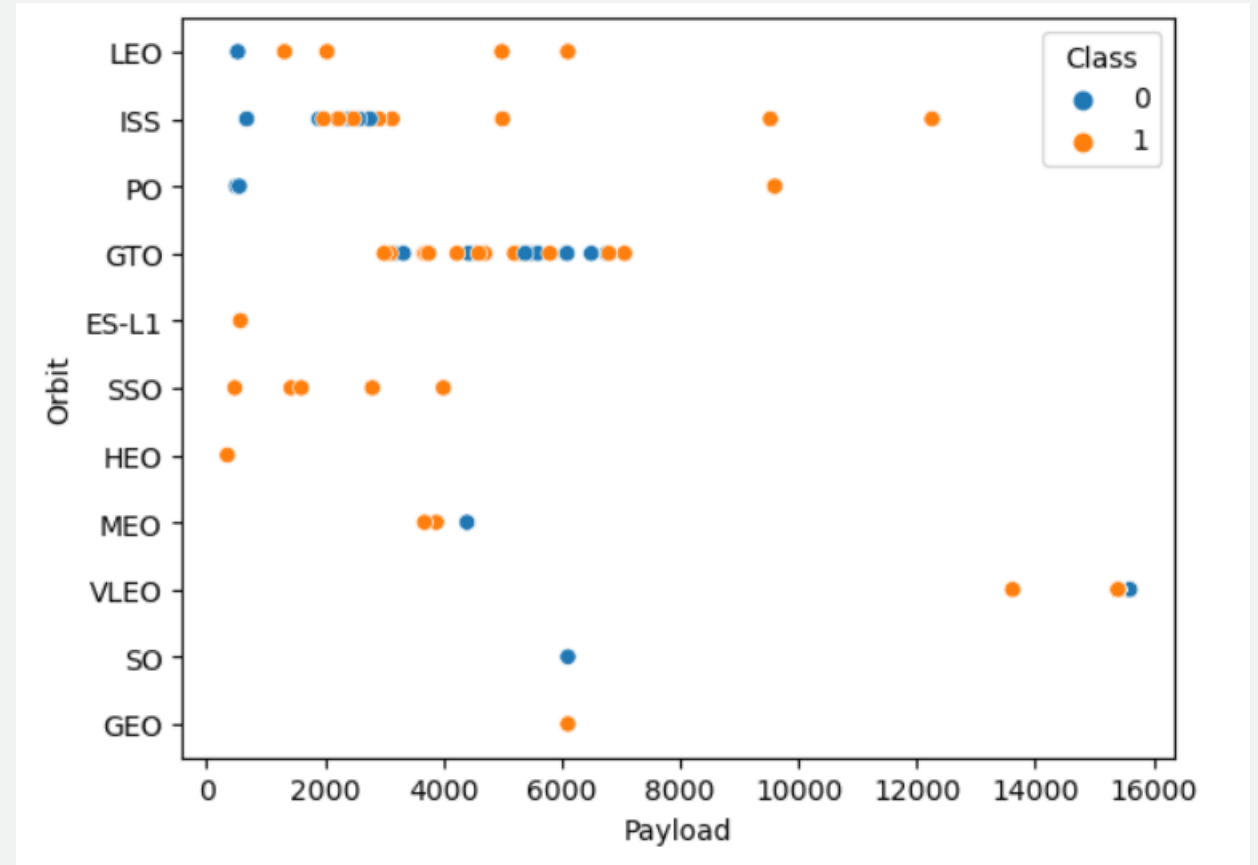
Flight Number vs. Orbit Type

- In LEO orbit the Success appears related to the number of flights
- There seems to be no relationship between flight number and success for GTO



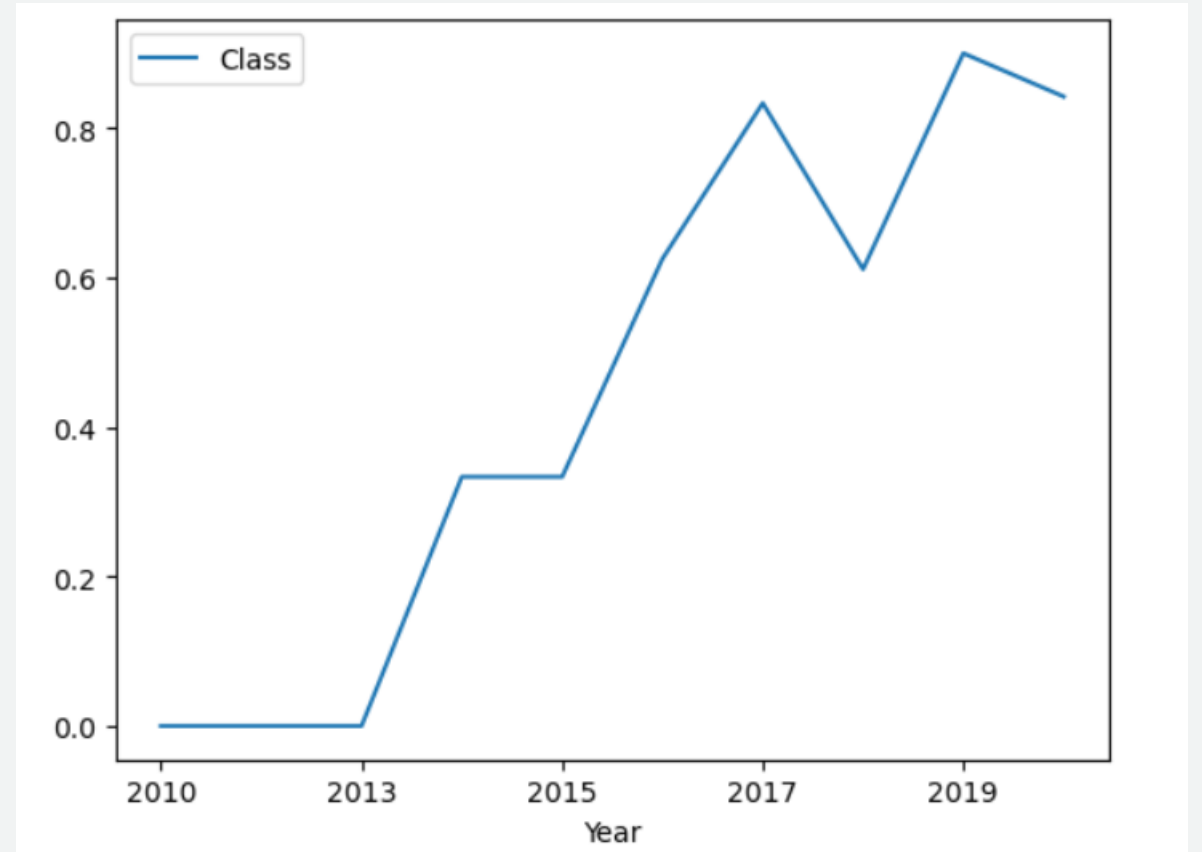
Payload vs. Orbit Type

- For heavy payload successful landing are more for LEO, Polar and ISS
- Payload doesn't seem to have much effect on landing outcome of GTO orbit



Launch Success Yearly Trend

- Success rate has increased yearly.



All Launch Site Names

- There are 4 unique launch sites

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

- These are obtained by using distinct () function on column launch site.

Launch Site Names Begin with 'CCA'

- These 5 records where launch sites 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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

- We can observe most of the initial launches were done for NASA.

Total Payload Mass

- The total payload in KG mass is:

```
total_payload_mass
```

```
45596
```

- The sum function is used for column payload mass in the query to obtained the above result.

Average Payload Mass by F9 v1.1

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

`avg_payload_mass`

2928

- The avg function is used on payload mass column followed by where clause for filtering out F9 v1.1 booster version.

First Successful Ground Landing Date

- The first successful landing outcome on ground pad

`min_date`

`2015-12-22`

- In this query the min function was used for date column followed by a where clause for filtering out successful outcome in landing outcome column.

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

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

- Distinct function was used for booster version column, where clause and between were used for filtering out payload and like was used for filtering out landing outcome for drone ship.

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes

mission_outcome	total
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- Count function was used for counting total rows belonging to given mission outcomes. Where and like were used to filter out mission outcome

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

- Subquery was used to get max payload.

2015 Launch Records

- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

booster_version	launch_site	landing__outcome
F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

- Year function was used to filter out 2015 and like was used for landing outcome to get failure drone ship

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

- 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

landing__outcome	total_count	rnk
No attempt	10	1
Failure (drone ship)	5	2
Success (drone ship)	5	3
Controlled (ocean)	3	4
Success (ground pad)	3	5
Failure (parachute)	2	6
Uncontrolled (ocean)	2	7
Precluded (drone ship)	1	8

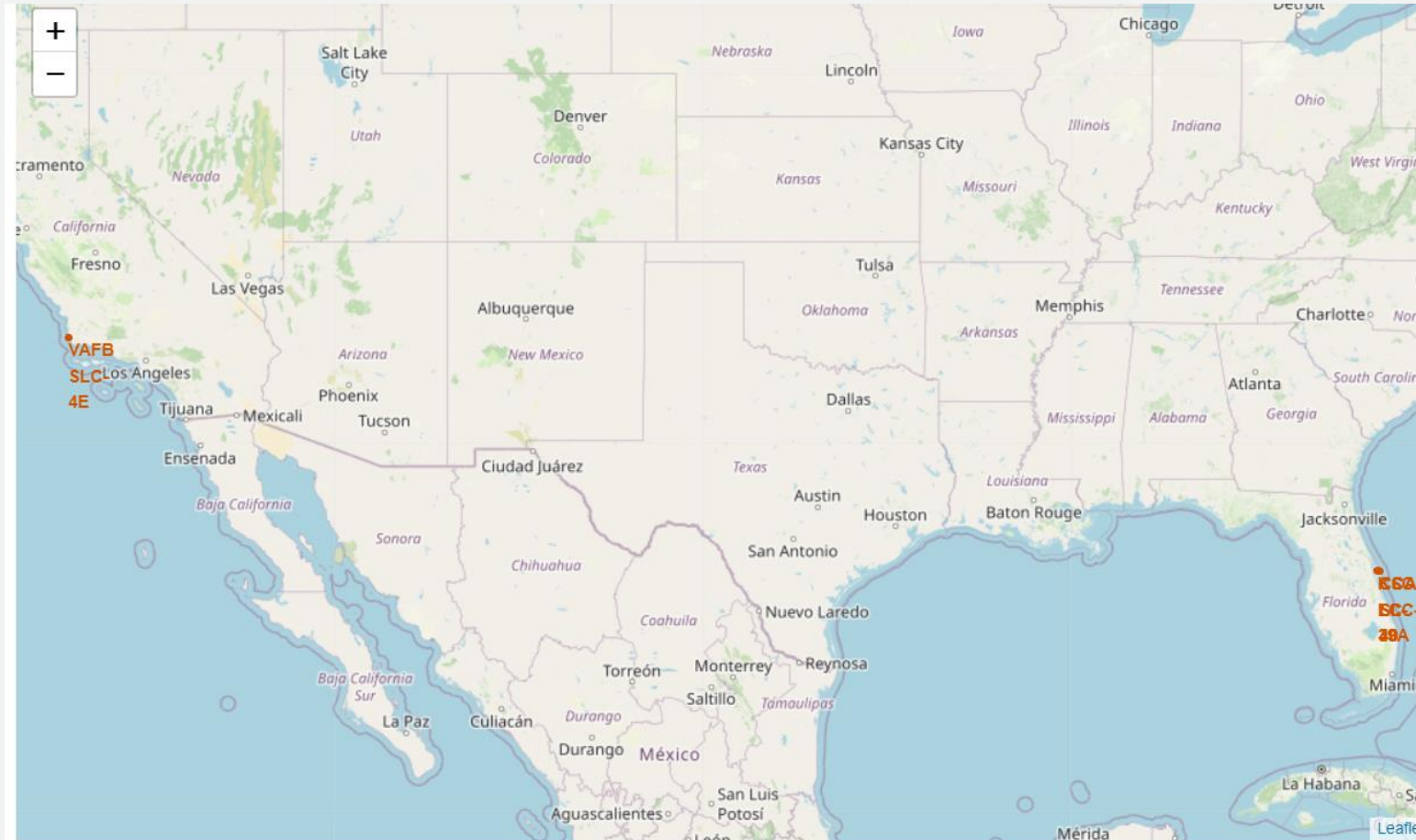
- Row Number function was used to rank the landing outcomes.

A satellite view of Earth from space, showing the curvature of the planet and city lights at night. The background is a deep blue gradient.

Section 3

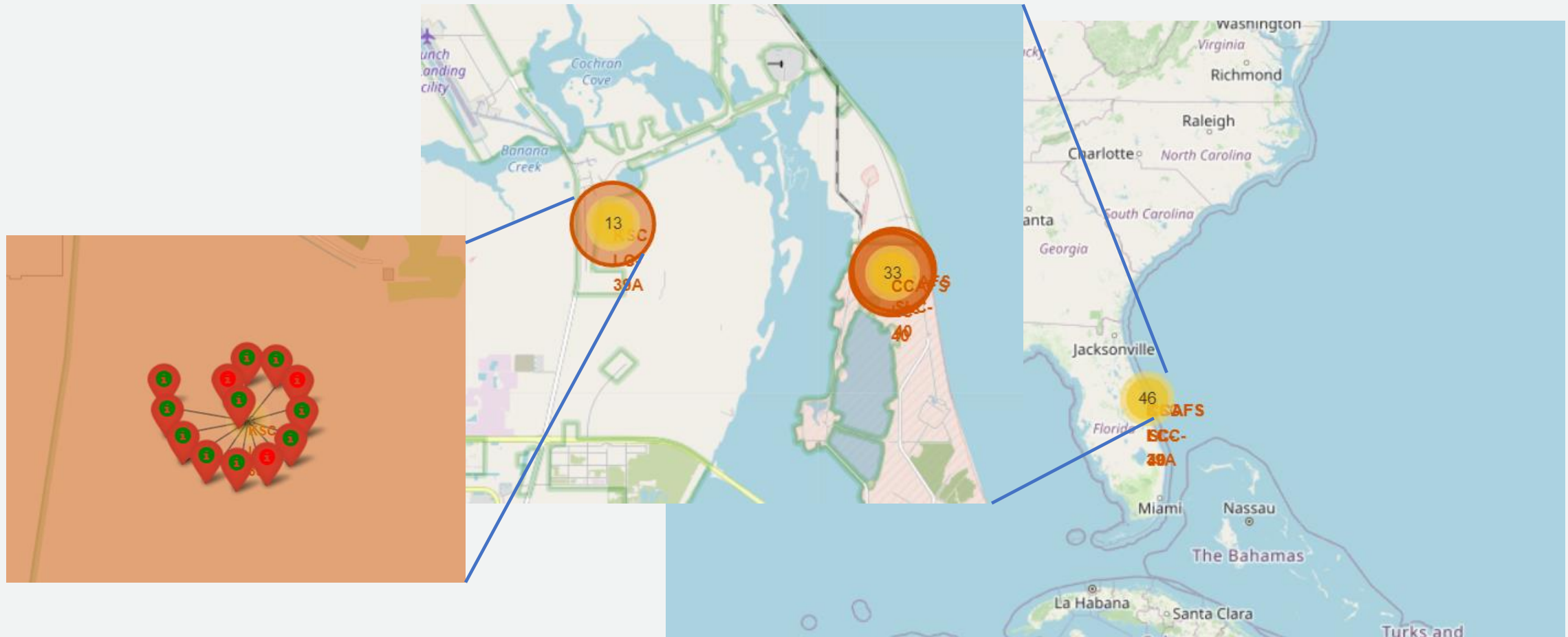
Launch Sites Proximities Analysis

Launch Sites for SpaceX



- All the launch sites are near sea for safety, but not too far from rail and road.

Launch Outcome by Site



Green markers are successful outcomes and red are failed outcomes

Proximity with coastline



- Most of the launch sites are close to coasts.



Section 4

Build a Dashboard with Plotly Dash

Total Success Launch by Sites

Total Success Launch By Sites



- KSC LC-39A has had the most successful launches.

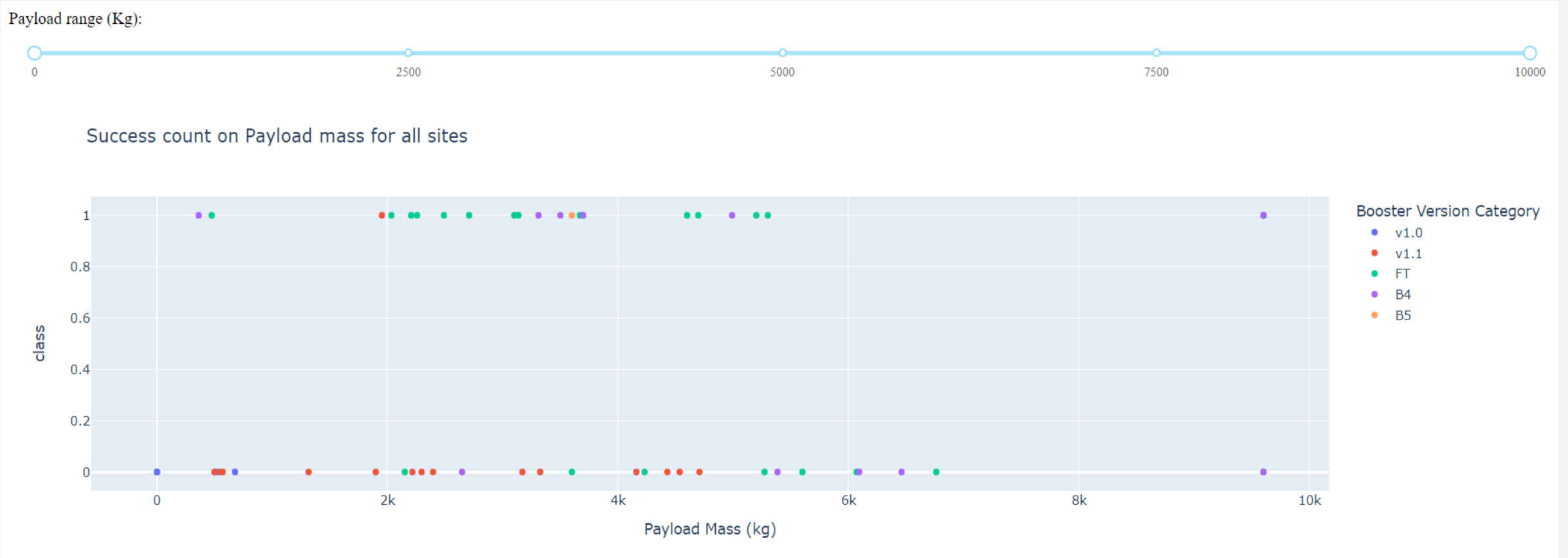
Highest Launch Success Ratio

Total Success Launch by Site KSC LC-39A



- 76.9% of launches are successful for the above site which is the highest amongst all.

Payload Mass for all sites



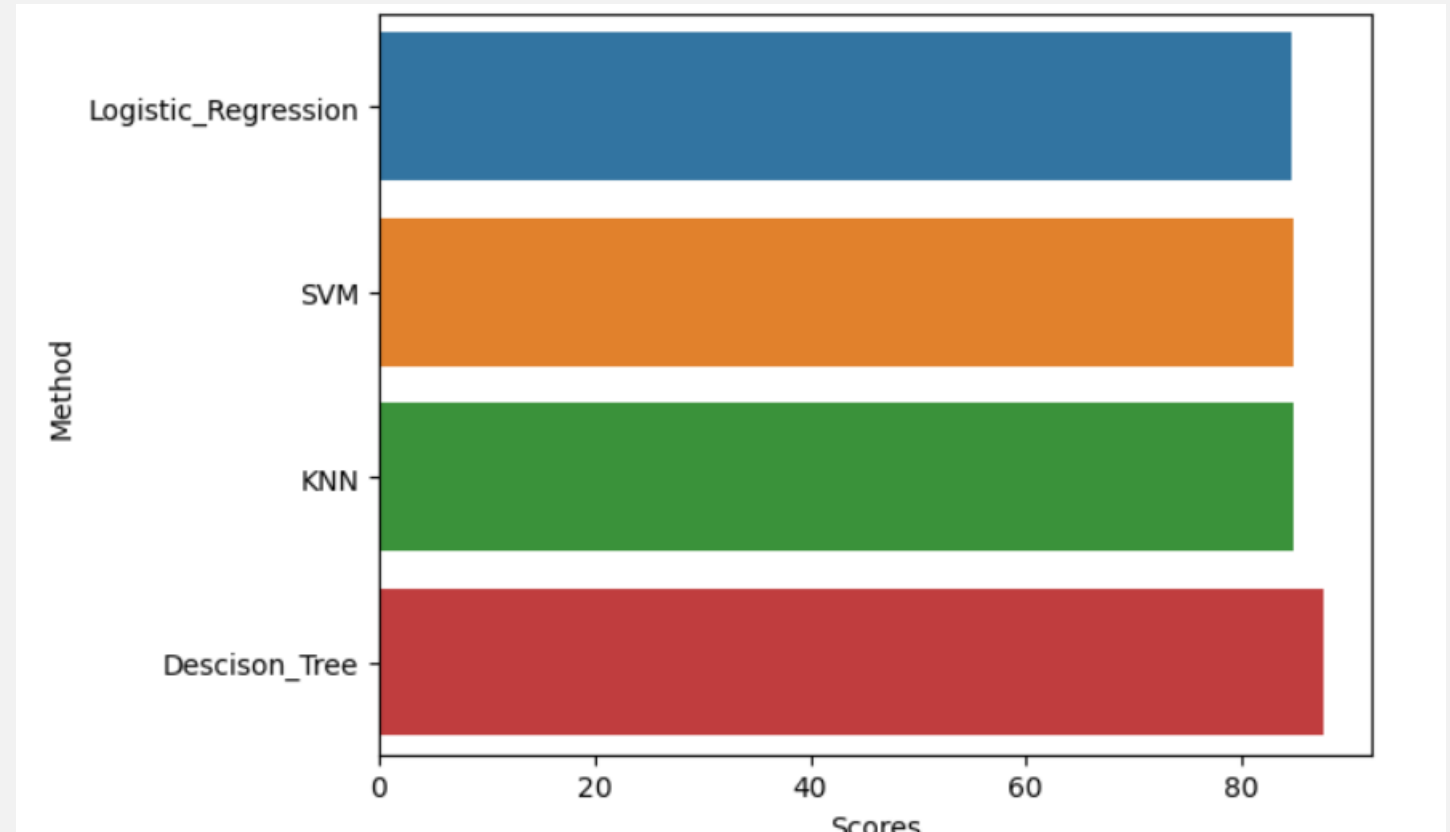
- Payload Mass under 6000 KG and booster version being FT is the combination seeing most successful launches.

Section 5

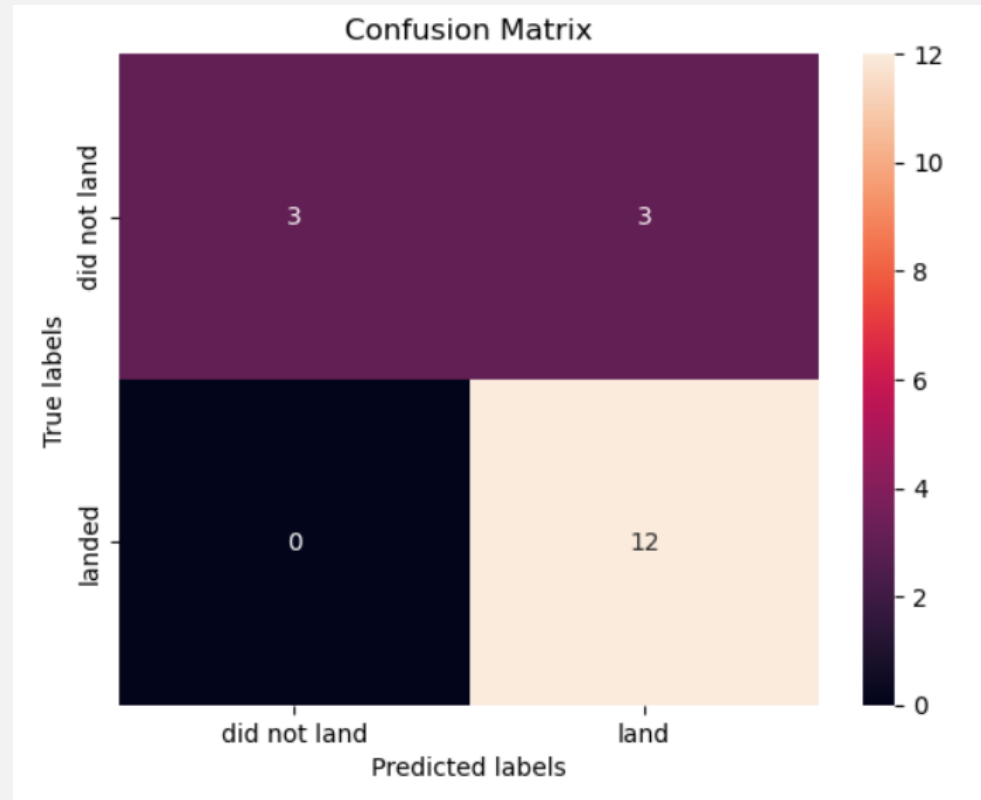
Predictive Analysis (Classification)

Classification Accuracy

- From the bar plot we can see that Decision Tree Classifier performs the best with accuracy of 87.7%



Confusion Matrix



- Decision Tree Classifier performed the best with mostly true positive and true negative and less false ones.

Conclusions

- Different data was analyzed and visualized to gain multiple insights.
- Most successful orbits for launch are ES L1, GEO, HEO and SSO.
- KSC LC-39A is the best site for launch.
- Best combination for successful launch is booster version FT with payload mass under 6000 KG.
- 99% of mission outcomes are success, however landing outcomes are more and more successful over the years due to increasing technology.
- Decision Tree Classifier is the best model for predicting successful launch and maximizing profits for Space Y.

Appendix

- Folium interactive map not visible on github. Therefore, a folder containing all the relevant screenshot are uploaded.
- Additional bar charts are added in predictive analysis to display best performing model

Thank you!

