IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Dmitry Emelianenko
2024-07-02

Outline

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

Executive Summary

- Summary of methodologies
  - Type of research: Building predictive models based on historical data of SpaceX space launches.
  - Data collection Process: Data was collected using SpaceX API's and web scrapping from https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
  - Data Analysis process included data pre-processing and transforming data. Visual summary was created, and an interactive board presented. This allowed to draw some insights on the data. At last, 4 different predictive models were created that would determine the success rate of launches based on Launch Location, Booster Version, Payload and Orbit.
- Summary of all results
  - As a result, we can determine how successful the potential launch would be based on the Booster Version, Orbit, Launch Location and Payload of the spacecraft.

Introduction

- Project background and context
    - The goal of the research is to provide a model that would be useful in determining the success rate of future launches for potential competitive company. The training data used is from previous launches by provided by SpaceX.
- Problems you want to find answers
    - If a new company would try enter the market to compete with SpaceX, what is the cusses rate of a potential launch. The goal is to build a model that would help to determine if the potential new competitor would be successful enough to compete with SpaceX.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from publicly available sources such as SpaceX API's and Wikipedia page with SpaceX Launch history.

Methodology

Executive Summary

- Performed data wrangling:
  - Calculated the number of launches on each site, the number and occurrence of each orbit, the number and occurrence of mission outcome per orbit type.
  - Loaded SpaceX dataset into SQLite Table and used SQL queries to gather the aforementioned information.
  - Created an interactive map using Folium map and created marks for all the launching sites along with useful information such as distances to the objects of importance (such as railroads, highways, cities, shorelines, etc.). Additionally, created markers for successful and failed launches for each launch site.
  - Created a dashboard using Plotly Dash. The dashboard included drop-down to filter information by launch suite and payload range.
  - Performed predictive analysis using classification models such as logistic regression, support vector machine, decision tree classifier and k-nearest neighbors. Accuracy score calculated for each model and the best performing models were logistic regression and decision tree classifier models.
  - Data was split into testing and training set, for each models the best hyperparameters were used.

Data Collection

- Data was collected by accessing SpaceX data via API’s and web scrapping from wiki page
    - https://api.spacexdata.com/v4/rockets/
    - https://api.spacexdata.com/v4/payloads/
    - https://api.spacexdata.com/v4/cores/
    - https://api.spacexdata.com/v4/launches/past
    - https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- Data collection process:
    - Access data using API’s, load data into pandas data frame, deal with categorial variables by encoding it using dummy variables and finally, deal with Null data by replacing it with averages from the same column.

Data Collection – SpaceX API

- Github link with the completed notebook – data collection using api:
    - https://github.com/dmitryemelianenko/Winning-Space-Race-with-Data-Science/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Get booster names:

- requests.get("https://api.spacexdata.com/v4/rockets/"+str(x)).json()

Get launch site names:

- requests.get("https://api.spacexdata.com/v4/launchpads/"+str(x)).json()

Get mass of the payload and orbit data:

- requests.get("https://api.spacexdata.com/v4/payloads/"+load).json()

Get outcome of the landing, the type of the landing, number of flights:

- requests.get("https://api.spacexdata.com/v4/cores/"+core['core']).json()

Get rocket launch data:

- spacex_url="https://api.spacexdata.com/v4/launches/past"
- response = requests.get(spacex_url)

def booster_version(table_cells):

Data Collection - Scraping

GitHub URL of the completed web scraping notebook:

- https://github.com/dmitryemelianenko/Winning-Space-Race-with-Data-Science/blob/main/jupyter-labs-webscraping.ipynb

Define helper functions to process HTML table:

- def date_time(table_cells):
- def booster_version(table_cells):
- def landing_status(table_cells):
- def get_mass(table_cells):
- def extract_column_from_header(row):

Use HTTP GET method to request HTML page and then Create BeautifulSoup object to convert data into a string for further processing:

- static_url = https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- soup = BeautifulSoup(response.text, 'html.parser')

Extract column names:

- for th in first_launch_table.find_all('th'):

  column_names.append(th.text.strip())

Create data frame by parsing the launch HTML talbes using the helper functions and for loop:

- for table_number,table in enumerate(soup.find_all('table',"wikitable plainrowheaders collapsible")):

Data Wrangling

- Firstly, number of launches on each site was calculated:
  - df['LaunchSite'].value_counts()
- Then number and occurrence of each orbit was calculated:
  - df['Orbit'].value_counts()
- The number and occurence of mission outcome of the orbits was calculated:
  - landing_outcomes = df['Outcome'].value_counts()
- A landing outcome label from Outcome column was created and data was classified as 0 = failure and 1 = succcess:

```
landing_class = []
for outcome in df['Outcome']:
    if outcome in bad_outcomes:
        landing_class.append(0)
else:
        landing_class.append(1)
df['Class']=landing_class
```

- Finally, we can calculate the success rate of a landing:
  - df["Class"].mean()
- Github page with the completed notebook:
  - https://github.com/dmitryemelianenko/Winning-Space-Race-with-Data-Science/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Scatter plot of FlightnNumber vs PayloadMass classified by the outcome. From this plot, we can see that as flight number increases, the success rate of first stage rocket returning is also increasing. Additionally, the higher the mass payload, the less likely first stage is to return.

- Scatter plot of Flight Number vs Launch site classified by the outcome was created. From this plot, we can see that different launch sites have different success rates. For example, CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.

- Then, a scatter plot of Paylaod vs Launch Site, lets us determine which site are more successful with different payload mass launches. For example, VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

- Bar chart of success rate by orbit type provides the insight on which orbit type is more likely to be successful.

- Scatter plot of Flight Number vs Orbit type reveals that the success rate of LEO orbit is related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

- Payload vs Orbit type plot suggests that With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS. However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

- Finally, Launch success yearly trend line plot provides an overall trend of success rate. SpaceX seems to be steadily improving yearly.

- Github page with completed notebook: https://github.com/dmitryemelianenko/Winning-Space-Race-with-Data-Science/blob/main/edadataviz.ipynb

EDA with SQL

- Display the names of the unique launch sites in the space mission

  %sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE;

- Display 5 records where launch sites begin with the string 'CCA'

  %sql SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5;

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

  %sql SELECT SUM("PAYLOAD_MASS__KG_") as total_payload FROM SPACEXTABLE WHERE "Customer" = "NASA (CRS)";

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

  %sql SELECT AVG("PAYLOAD_MASS__KG_") AS average_payload FROM SPACEXTABLE WHERE "Booster_Version" = "F9 v1.1";

- List the date when the first succesful landing outcome in ground pad was acheived.

  %sql SELECT MIN("Date") as First_Successful_Landing_Date FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (ground pad)';

- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

  %sql SELECT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (drone ship)' AND "PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MASS__KG_" < 6000;

- List the total number of successful and failure mission outcomes

  %sql SELECT "Mission_Outcome", COUNT(*) as Total_Count FROM SPACEXTABLE GROUP BY "Mission_Outcome";

- List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

  %sql SELECT "Booster_Version" FROM SPACEXTABLE WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTABLE);

- 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.
  - %sql SELECT SUBSTR("Date", 6, 2) AS Month, "Landing_Outcome" AS Failure_Landing_Outcome, "Version_Booster", "Launch_Site" FROM SPACEXTABLE WHERE SUBSTR("Date", 0, 5) = '2015' AND "Landing_Outcome" LIKE 'Failure (drone ship)';
- 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 SELECT "Landing_Outcome", COUNT(*) AS Count FROM SPACEXTABLE WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY "Landing_Outcome" ORDER BY Count DESC;
- GitHub page with completed noteboot: https://github.com/dmitryemelianenko/Winning-Space-Race-with-Data-Science/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Circles on the map locate the launching sites. Within the circles, we have markers that correspond to launch attempts, color coded to distinguish successful launches from failed launches. Lines on the map illustrate proximity (distance) to the objects of interest, such as railroads, shorelines, cities, highways and etc.
- These objects help to visualize the data and provide insights on how successful the launching site is, its proximity to the closest city and major infrastructures.
- GitHub link with the completed notebook: https://nbviewer.org/github/dmitryemelianenko/Winning-Space-Race-with-Data-Science/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

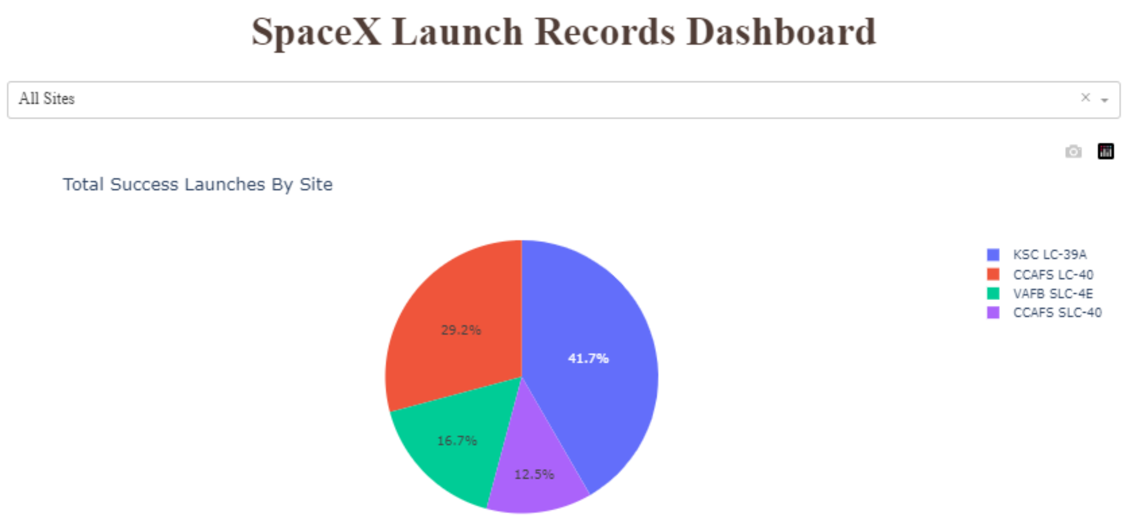
- I have created a PieChart to showcase which launching site had most total successful launches and which launch site had highest success rate of launches.
- Scatter plot provided in the dashboard shows the correlation between Payload and success rate for all sites.
- I've added these plots for visual representation of some of the data. They allow us to draw some preliminary conclusions about the Success rate. For example, which Launch Site has the highest success rate, which payload has the highest success rate, etc.
- GitHub link to the completed dashboard file - https://github.com/dmitryemelianenko/Winning-Space-Race-with-Data-Science/blob/main/spacex_dash_app.py
- GitHub link to the screenshot of the dashboard - https://github.com/dmitryemelianenko/Winning-Space-Race-with-Data-Science/blob/main/Dashboard.png

Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model
- First step in building the model was to split the data into training and validation sets. Then, each model was trained on the training set and accuracy scores for each model were calculated. The accuracy scores when applied to the validation set were compared and the highest accuracy score corresponded to the best performing model.
- For each model, the best performing hyperparameters were selected. In order to determine the best hyperparameters, the accuracy of the model was calculated on the training set and highest score yielded the best performing hyperparameters.
- Additionally, I have used the Confusion matrix for each model to visualize the accuracy, here we want to have all results split between top-left (True Positive) and bottom-right (True Negative) squares.
- 4 models were created:
    - logistic regression
    - support vector machine
    - decision tree classifier
    - k nearest neighbors
- All models performed equally well with accuracy of 83%
- GitHub link to the completed notebook - https://github.com/dmitryemelianenko/Winning-Space-Race-with-Data-Science/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

- Exploratory data analysis results:
  - CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.
  - For the VAFB-SLC launchsite there are no rockets launched for heavy payload mass(greater than 10000).
  - ES-L1, GEO, HEO and SSO orbits have highest success rate.
  - LEO orbit has the Success Rate related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.
  - With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
  - However, for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.
  - Success rate of SpaceX Launches kept increasing since 2013 till 2020.
- Predictive analysis results
  - Most accurate models are logistic regression and decision tree classifier, with accuracy of 83%

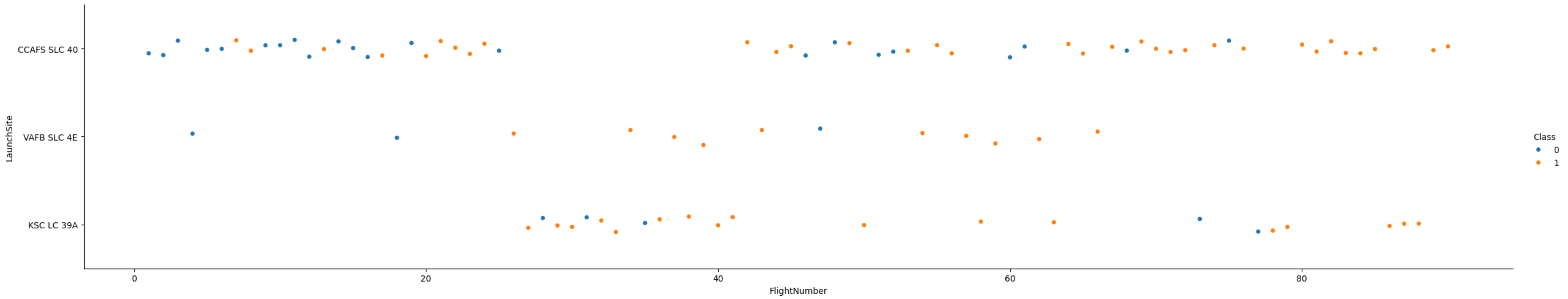


Section 2

Insights drawn from EDA

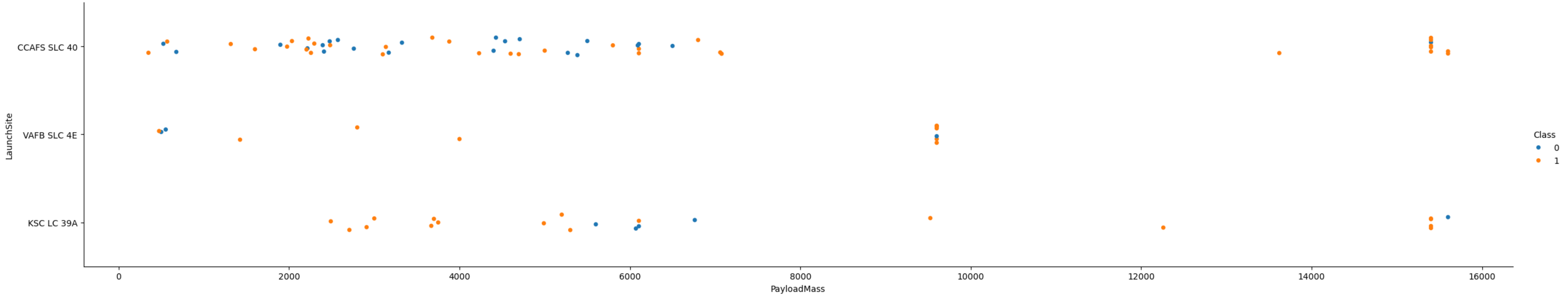
Flight Number vs. Launch Site

We see that different launch sites have different success rates. CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.



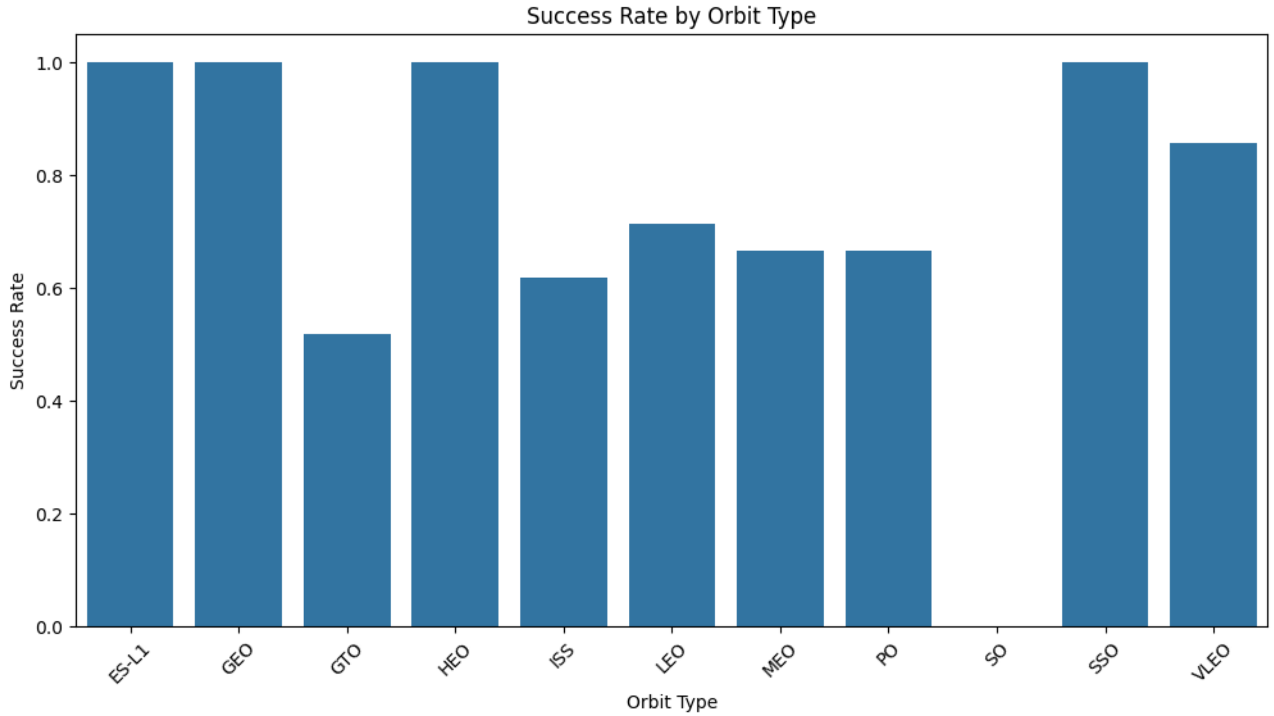
Payload vs. Launch Site

VAFB-SLC Launch Site there are no rockets launched for heavypayload mass(greater than 10000).



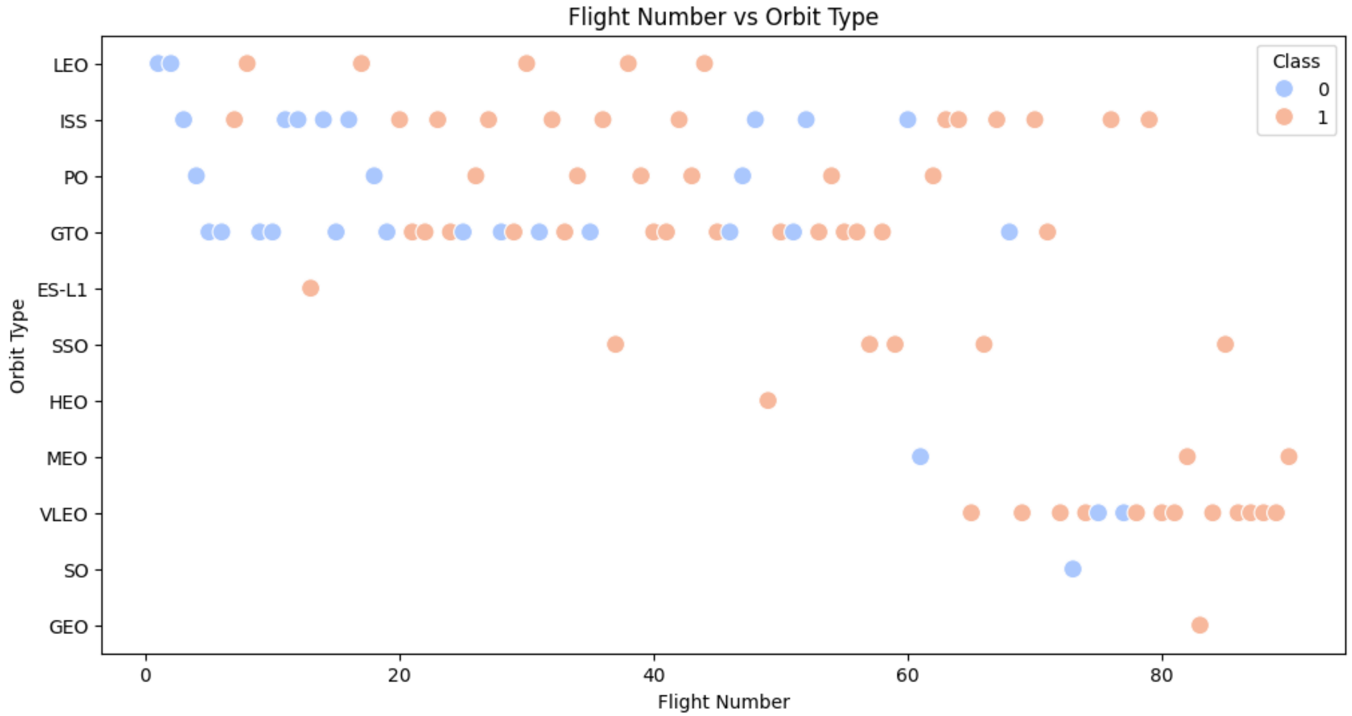
Success Rate vs. Orbit Type

ES-L1, GEO, HEO and SSO Orbit types have the highests success rates. At the same time GTO and SO Orbit types have the lowest success rate



Flight Number vs. Orbit Type

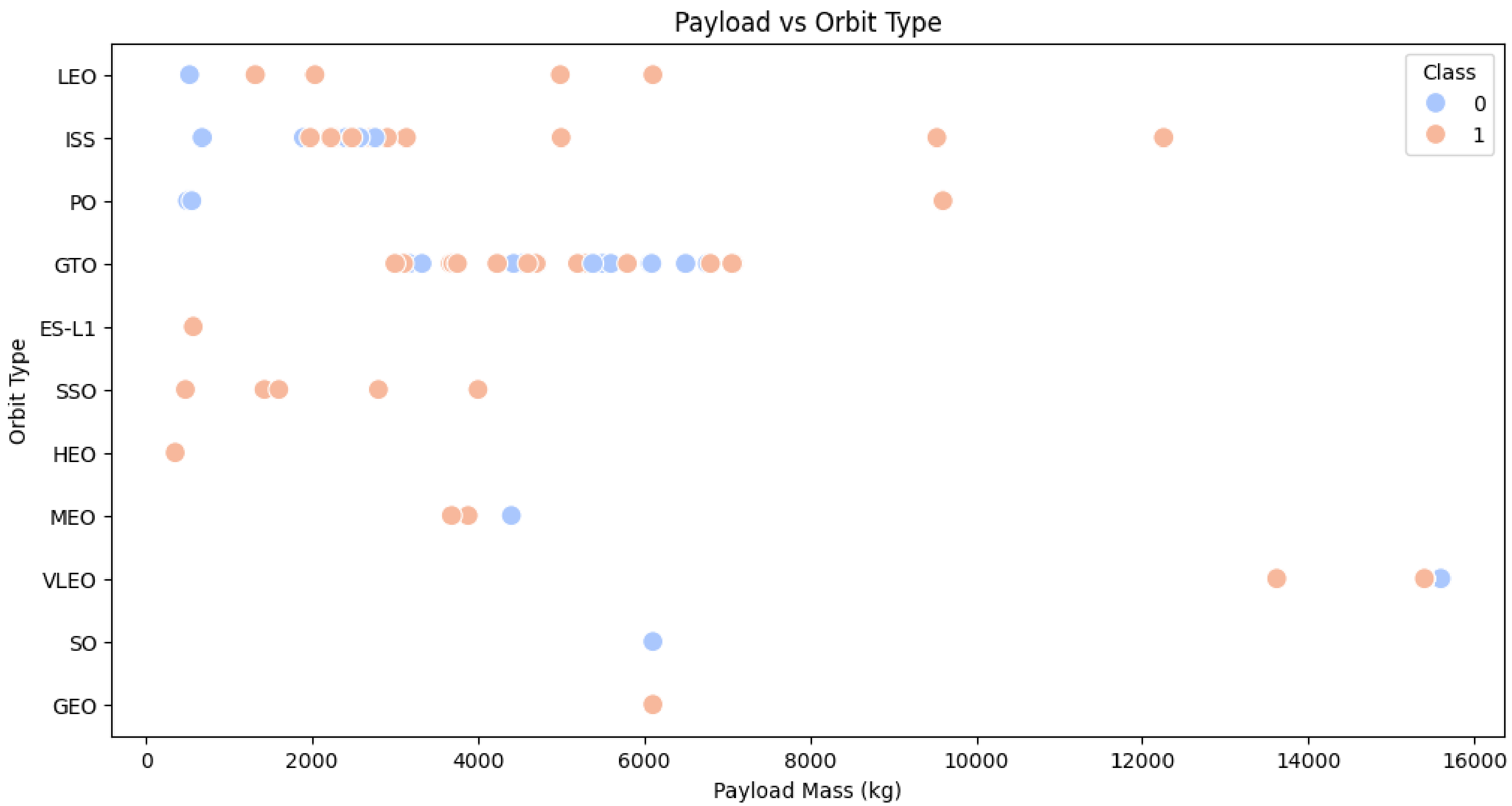
LEO orbit's Success rate appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.



Payload vs. Orbit Type

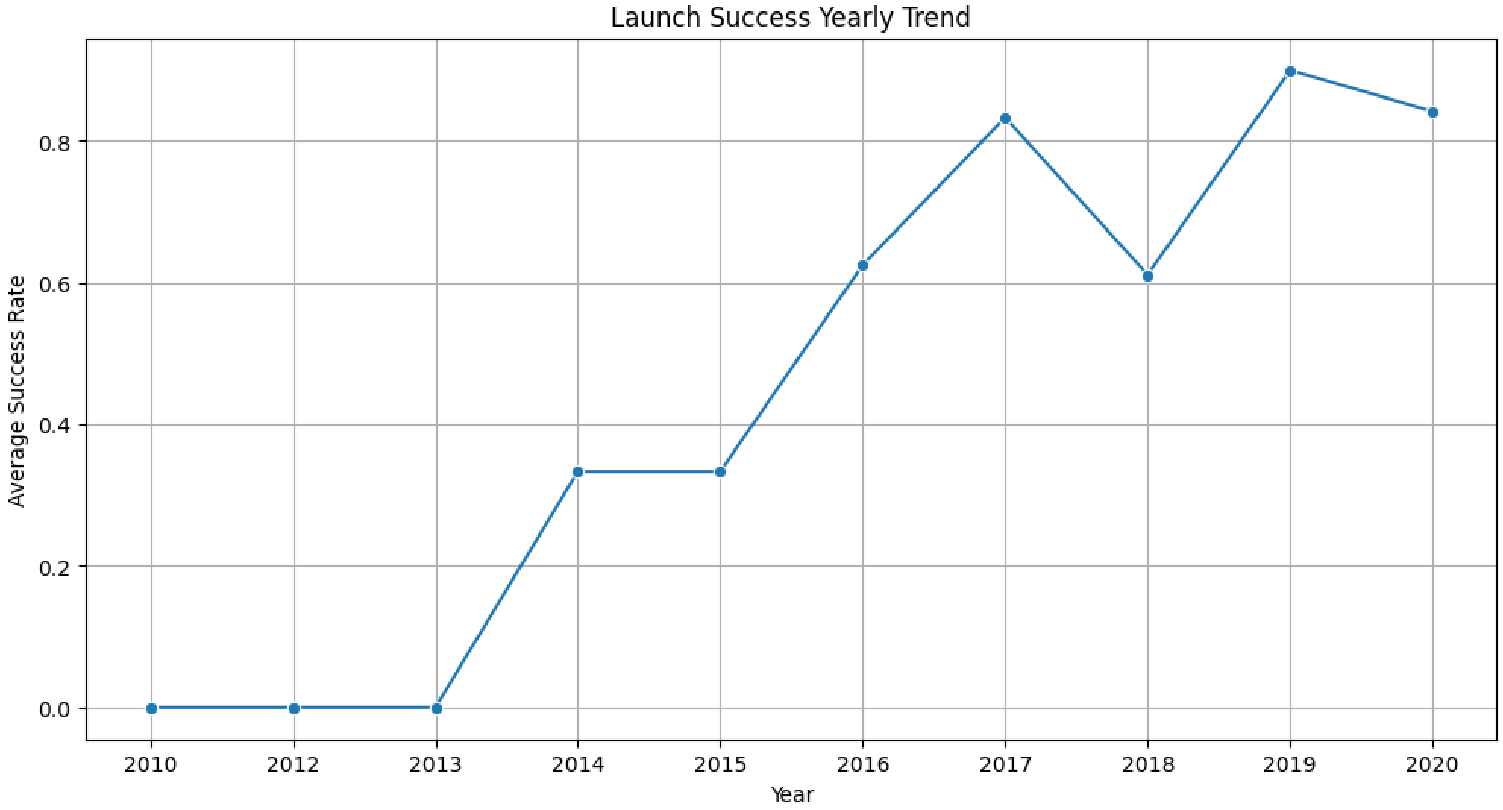
With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

However, for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.



Launch Success Yearly Trend

Success rate of Launches has been steadily increasing since 2010 all the way to 2020



All Launch Site Names

- There are 4 unique Launch sites in the data:
  - %sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE;

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

Launch Site Names Begin with 'CCA'

- %sql SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE 'CCA%' LIMIT

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

Total Payload Mass Launched by NASA (CRS)

- %sql SELECT SUM("PAYLOAD_MASS__KG_") as total_payload FROM SPACEXTABLE WHERE "Customer" = "NASA (CRS)";

total_payload
45596

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG("PAYLOAD_MASS__KG_") AS average_payload FROM SPACEXTABLE WHERE "Booster_Version" = "F9 v1.1";
```

average_payload
2928.4

First Successful Ground Landing Date

```
%sql SELECT MIN("Date") as First_Successful_Landing_Date FROM SPACEXTABLE WHERE "Landing_Outcome" =
'Success (ground pad)';
```

First_Successful_Landing_Date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- %sql SELECT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (drone ship)' AND "PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MASS__KG_" < 6000;

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

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT "Mission_Outcome", COUNT(*) as Total_Count FROM SPACEXTABLE GROUP BY "Mission_Outcome";
```

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

Boosters Carried Maximum Payload

- %sql SELECT "Booster_Version" FROM SPACEXTABLE WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTABLE);

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

```
%sql SELECT SUBSTR("Date", 6, 2) AS Month, "Landing_Outcome" AS Failure_Landing_Outcome, "Version_Booster", "Launch_Site" FROM SPACEXTABLE WHERE SUBSTR("Date", 0, 5) = '2015' AND "Landing_Outcome" LIKE 'Failure (drone ship)';
```

Month	Failure_Landing_Outcome	"Version_Booster"	Launch_Site
01	Failure (drone ship)	Version_Booster	CCAFS LC-40
04	Failure (drone ship)	Version_Booster	CCAFS LC-40

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

```
%sql SELECT "Landing_Outcome", COUNT(*) AS Count FROM SPACEXTABLE WHERE "Date" BETWEEN '2010-06-04'
AND '2017-03-20' GROUP BY "Landing_Outcome" ORDER BY Count DESC;
```

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

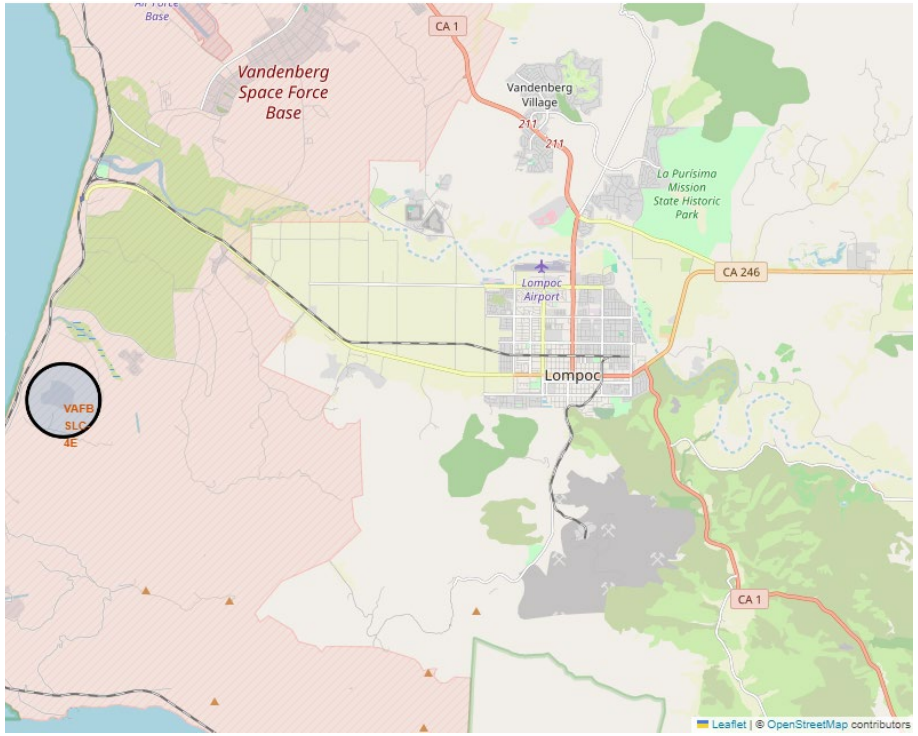
Section 3

Launch Sites Proximities Analysis

All Launch Sites Locations

All Launch Sites appear to be close to the shoreline.

Three of the Launch sites are located on the East Coast and the last on is on the West Coast

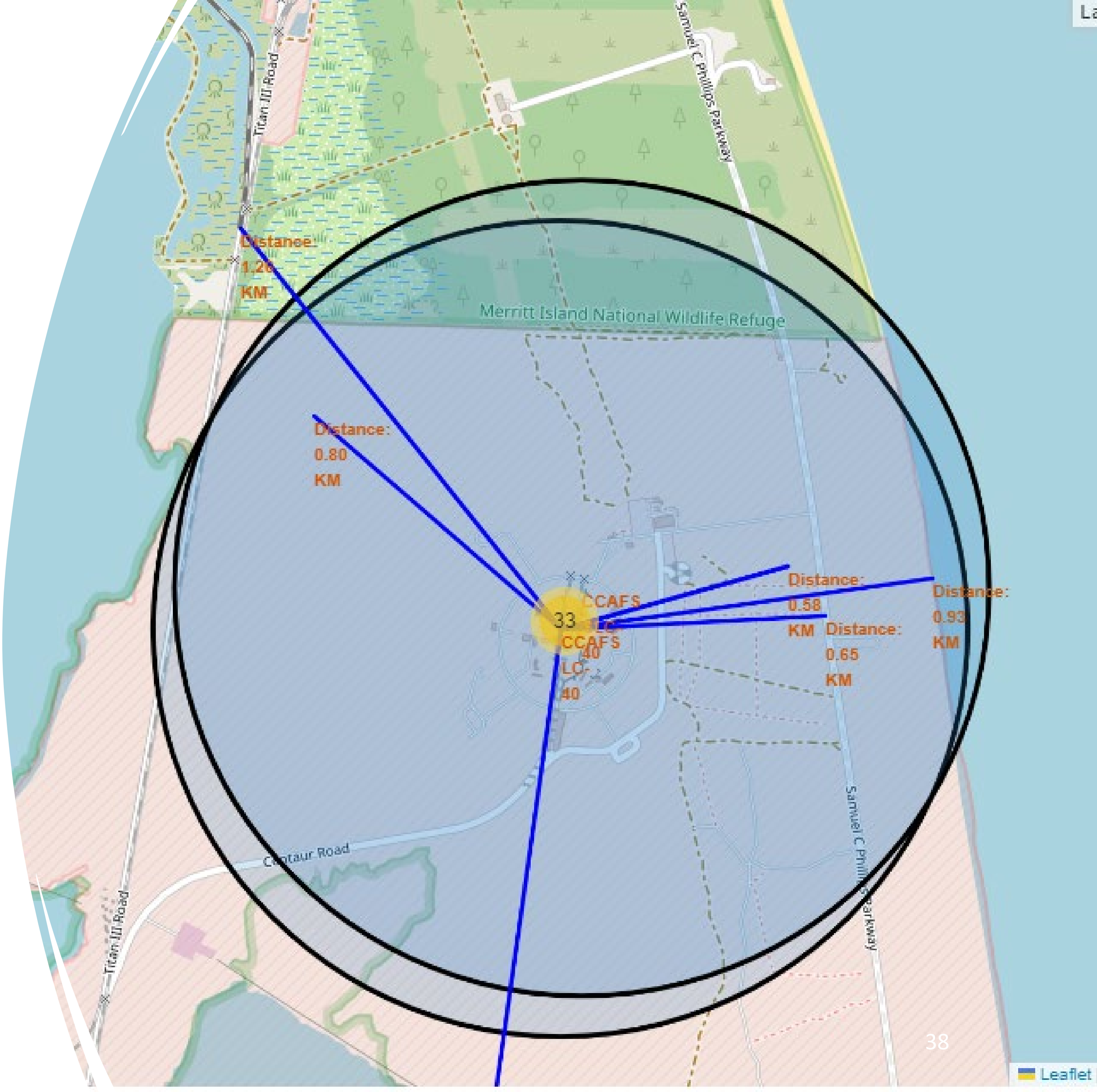


Mark Success/Failed Launches for each site on the map

- KSC-LC-39A has the most successful launches
- CCAFS LC-40 has the most Launches

Distance from Launch Site to Proximities

- We can see that the nearest railroad is 1.2km away from CCAFS LC-40 Launch Site
- The shoreline is 0.93km away
- The closest highway is 0.65km away
- And the closest city is 19.58 km away.

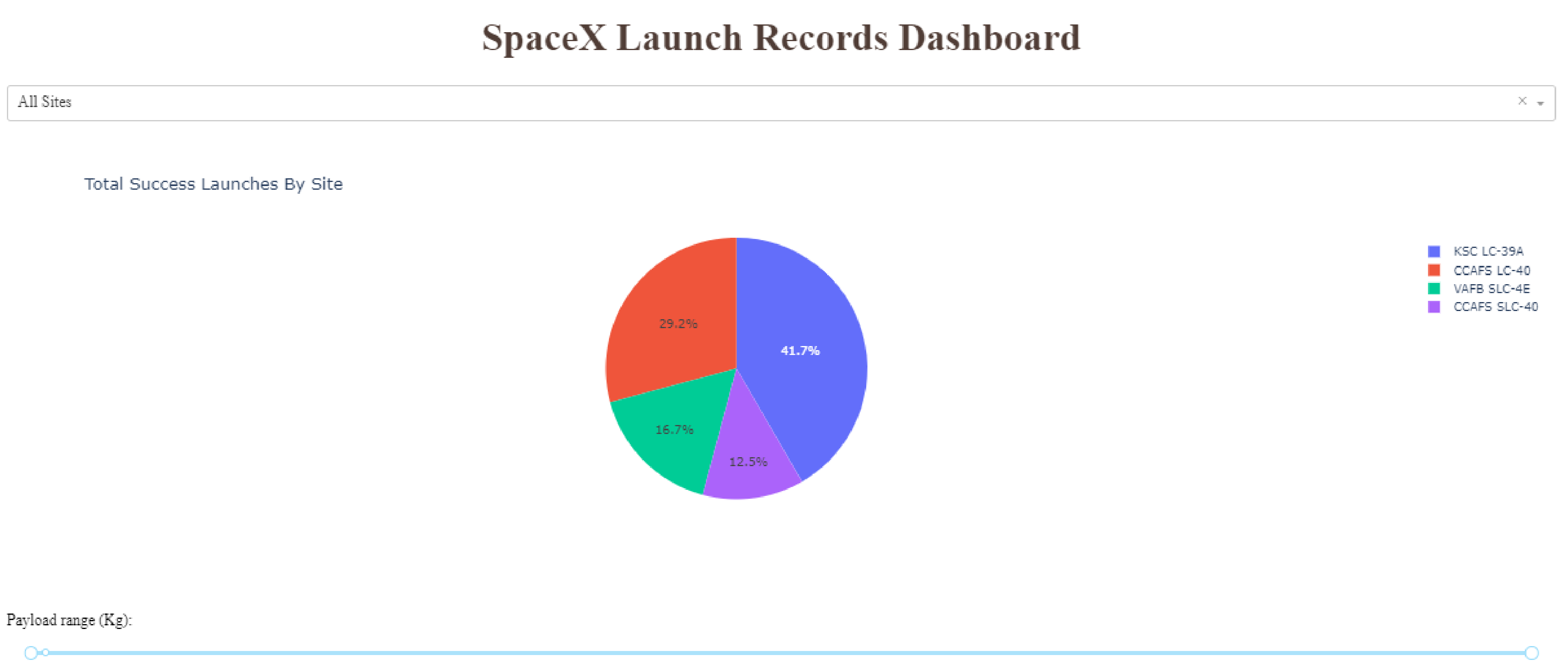


Section 4

Build a Dashboard with Plotly Dash

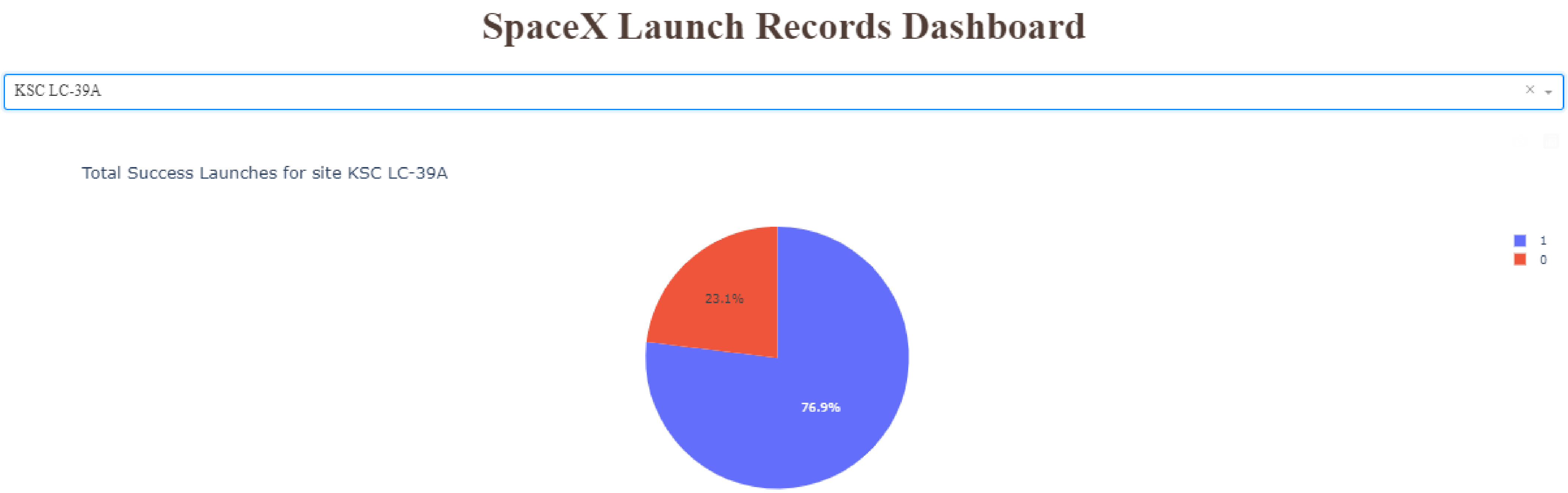
Total Success Launches by Site

From this Pie Chart we can see that the most total successful launches were made from KSC LC-39A site.



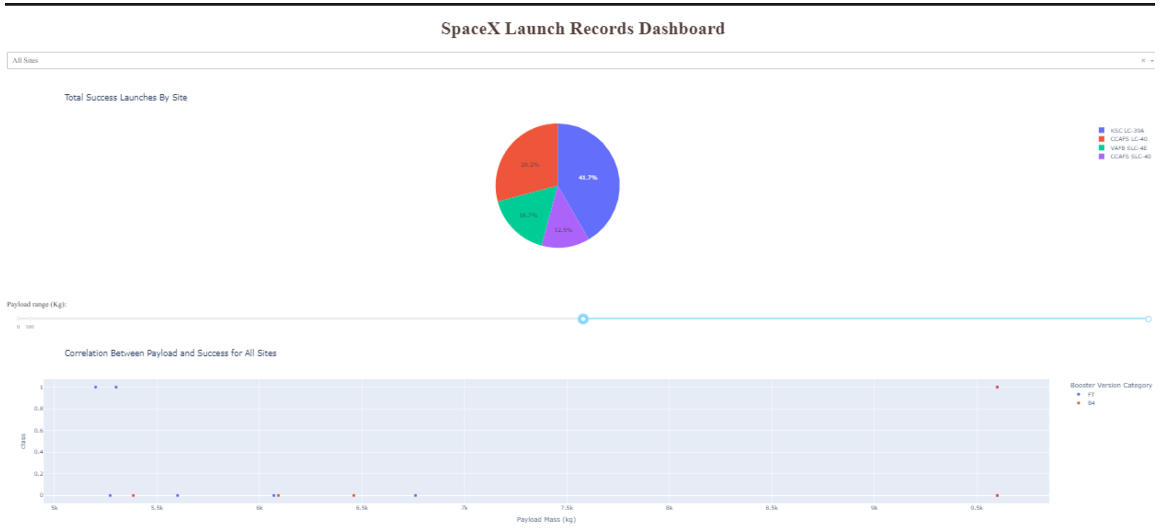
Total Success Launches for site KSC LC-39A

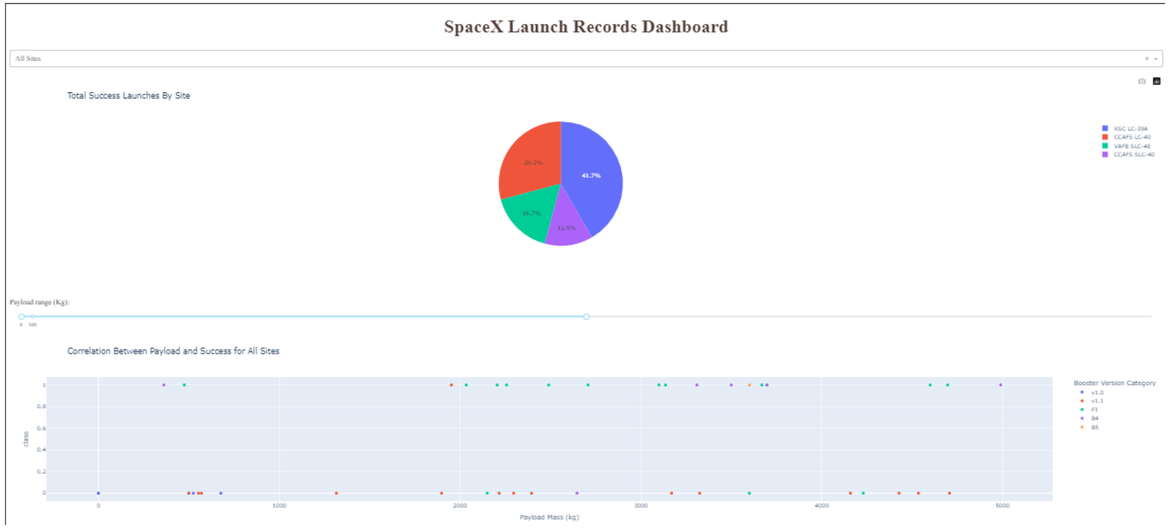
The Highest success rate among the Launch sites is at KSC LC-39A at 76.9%



<Dashboard Screenshot 3>

- For the low payload mass (Below 5000kg) FT Booster (green) version was the most successful
- For the heavy payload mass (Above 5000kg) the overall success rate is smaller for all boosters.
- Only two category of boosters - FT and B4 – launch with the heavy payload mass. FT appears to have a higher success rate, but no launches with payload mass above 7000kg were made with that booster.



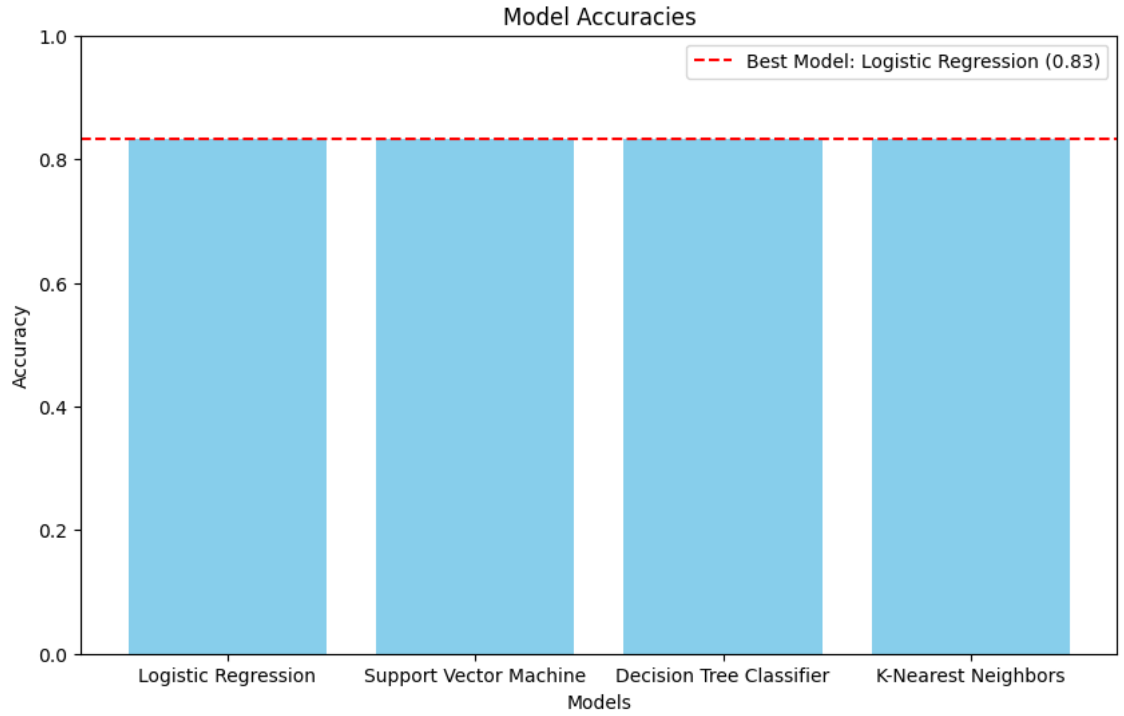


Section 5

Predictive Analysis (Classification)

Classification Accuracy

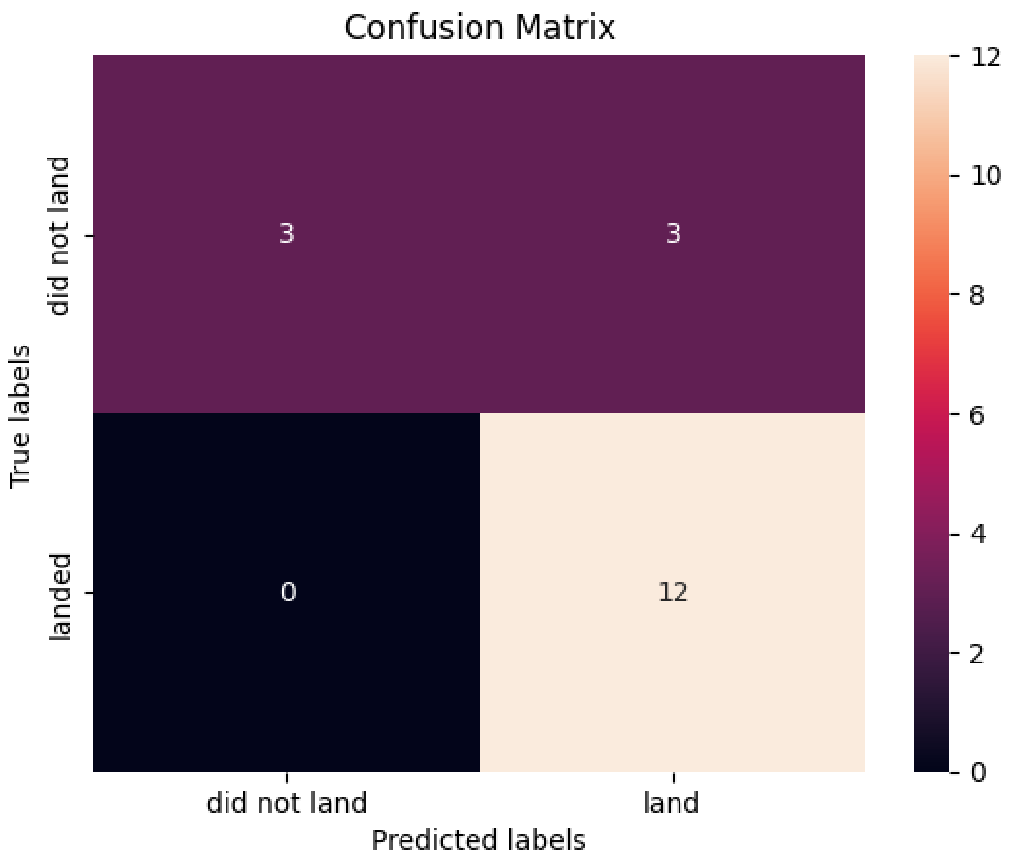
- All models exhibit the same accuracy of 83%



Confusion Matrix

Logistic regression model has the highest accuracy – 83%.

3/18 are True negative, 12/18 are True Positive and only 3/18 predictions did not match the validation data (model predicted 3 landigns that did not occur)



Conclusions

- The most successful Booster Category is FT Booster.
- The most successful Launch site is KSC LC-39A
- Payload seems to be related to the success rate of the launch. Heavier payload mass decrees chances of success.
- SpaceX Success rate has been steadily increasing since 2010

Appendix

- Datasets created during the project:
  - https://github.com/dmitryemelianenko/Winning-Space-Race-with-Data-Science/blob/main/dataset_part_1.csv
  - https://github.com/dmitryemelianenko/Winning-Space-Race-with-Data-Science/blob/main/dataset_part_2.csv
  - https://github.com/dmitryemelianenko/Winning-Space-Race-with-Data-Science/blob/main/dataset_part_3.csv
- Plotly Dash .py file - https://github.com/dmitryemelianenko/Winning-Space-Race-with-Data-Science/blob/main/spacex_dash_app.py

Thank you!