Winning Space Race with Data Science

Virginia Redondo Antón

Sept-2024

Outline

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

Executive Summary

The **objective** was identify the factors for a successful rocket landing.

- Summary of methodologies
  - Data Collection through API and Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis using SQL and Data Visualization
  - Interactive Visual Analytics and Dashboard
  - Machine Learning Prediction
- Summary of all results
  - Launch success has improved over time.
  - KSC LC-39 has the highest success rate.
  - Orbits ES-L1, GEO, HEO, and SSO have a 100% success rate.
  - All launch sites are near the coast.
  - All classification models performed similarly.

Introduction

- Project background and context
  - SpaceX is a space company whose costs per rocket launch are much lower than those of the competition.
  - Much of the savings is because SpaceX can reuse the first stage of the rocket. Therefore, the cost of the launch can be calculated if it can be determined whether the first stage will land.
  - The objective of the new fictitious rocket company SpaceY is to assess if SpaceX will reuse the first stage using public information to train a machine learning model.
- Problems you want to find answers
  - Identify the key factors that determine if the rocket will land successfully.
  - Analyze relationships between variables and their contribution to the success rate of the landing.
  - Determine the best predictive model for successful landing.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX API and web scraping from Wikipedia.
- Data wrangling
  - Data was processed filtering desired data, handling missing values, and applying one-hot encoding to categorical features.
- Exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Accomplish predictive analysis using classification models
  - Data were normalized, divided into training and test datasets, and tuning hyperparameters of different classification models.

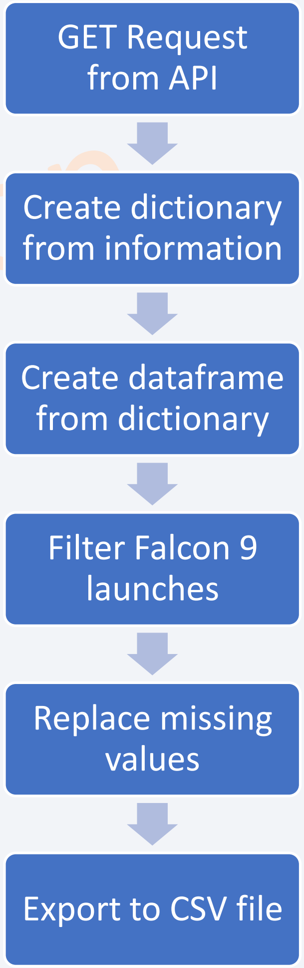
Data Collection

The data for this project was obtained from public information through the SpaceX API and web scraping from Wikipedia.

- SpaceX API (https://api.spacexdata.com/v4/rockets/past)
  - Send an GET request to the SpaceX API to obtain a JSON response.
- Web Scraping from Wikipedia page for List of Falcon 9 and Falcon Heavy launches updated on 9th June 2021.
  - Get the text from the Wikipedia page and find the information from the table with the Falcon 9 launches records using the BeautifulSoup library.

Data Collection – SpaceX API

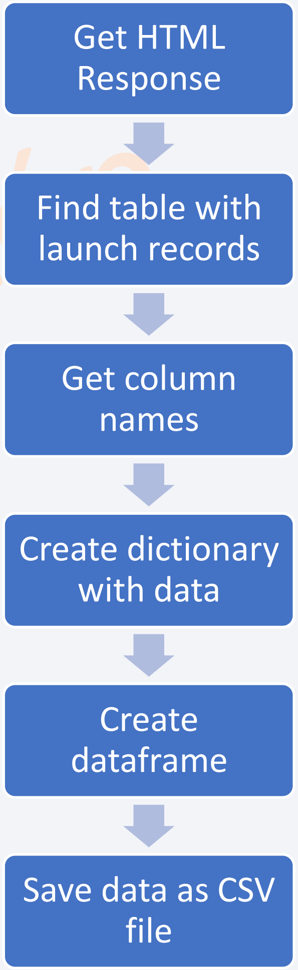
- Process:
  - Request and parse the SpaceX launch data using the GET request.
  - Filter the dataframe to only include Falcon 9 launches.
  - Dealing with Missing Values in PayloadMass column.
  - Export to a CSV file.
- Data obtained:
  - Dataframe with 90 rows and 17 columns



[GitHub URL – Lab1]

Data Collection - Scraping

- Process:
  - Request the Falcon 9 Launch Wiki page from its URL.
  - Extract column names from the HTML table header using BeautifulSoup library.
  - Create a data frame by parsing the launch HTML tables.
  - Export to a CSV file.
- Data obtained:
  - Dataframe with 121 rows and 11 columns



GitHub URL – Lab2

Data Wrangling

- Number of launches by site:

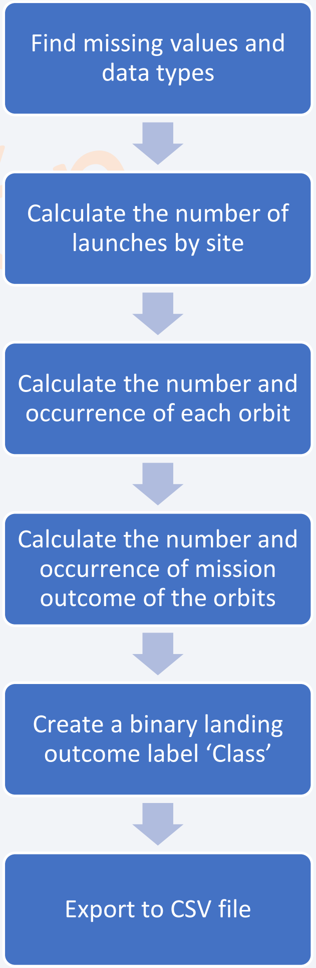
CCAFS SLC 40	55
KSC LC 39A	22
VAFB SLC 4E	13

- Number of each orbit:

GTO	27
ISS	21
VLEO	14
PO	9
LEO	7
SSO	5
MEO	3
ES-L1	1
HEO	1
SO	1
GEO	1

- Number of mission outcome:
  - The success rate is 66%.

True ASDS	41
None None	19
True RTLS	14
False ASDS	6
True Ocean	5
False Ocean	2
None ASDS	2
False RTLS	1



[GitHub URL – Lab3]

EDA with Data Visualization

- **Scatter plots**: Visualize correlation between variables
  - Flight Number vs Payload Mass
  - Flight Number vs Launch Site
  - Payload Mass vs Launch Site
  - Flight Number vs Orbit Type
  - Payload Mass vs Orbit Type
- **Bar char**: Visualize the value of items
  - Success rate of each orbit type
- **Line char**: Visualize trend over time
  - Launch success yearly trend

EDA with SQL

- Display the names of the unique launch sites.
- Display 5 records where launch sites begin with the string ‘CCA’.
- Display the total payload mass carried by boosters launched by NASA (CRS).
- Display average payload mass carried by booster version F9 v1.1.
- List the date when the first successful landing outcome in ground pad was achieved.
- List the name of the booters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
- List the total number of successful and failure mission outcomes.
- List the names of the booster versions which have carried the maximum payload mass.
- List the records which will display the month names, failure landing outcome in drone ship, booster versions, launch site for the months in year 2015.
- Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order.

[GitHub URL – Lab4]

Build an Interactive Map with Folium

- Mark all launch sites on a map
  - Add a blue circle in the NASA Johnson Space Center coordinates with a popup label showing its name using CircleMarker and Marker objects.
  - Add red circles at the coordinates of all launch sites coordinates with a popup label showing their names.
- Mark the successful/unsuccessful launches for each site on the map
  - Add colored markers for successful (green) and unsuccessful (red) launches at each launch site using a MarkerCluster object to group all the markers.
- Calculate the distances between a launch site to its proximities
  - Add PolyLines objects to show distances between a launch site and point of interest in its proximity, such as coastal areas, railways, highways and cities.

[GitHub URL – Lab6]

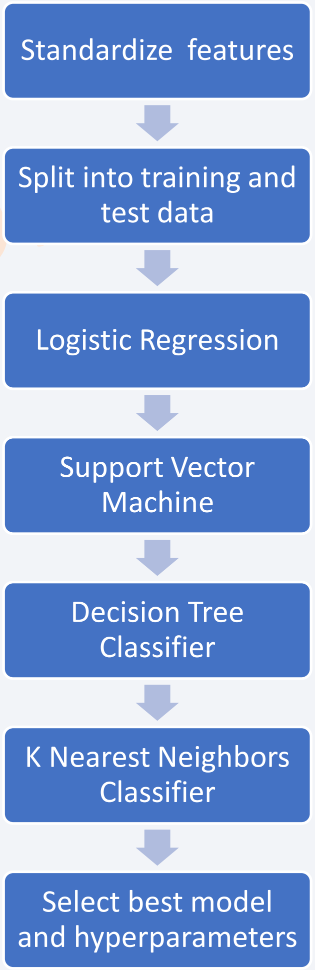
Build a Dashboard with Plotly Dash

- Add a Launch Site drop-down input component
  - Select All launches sites for overall information or a give launch site for selected information.
- Add a pie chart showing Success Rate based upon selected Launch Site
  - Shows Total Success Launches by Site or Total Success Launches for selected site.
- Add a range slider to select Payload Mass
  - Select payload mass range.
- Add a scatter plot chart showing Payload Mass vs Success Rate by Booster Version
  - Shows correlation between Payload Mass and Success Rate distinguishing between different Booster Versions.

[GitHub URL – Lab7]

Predictive Analysis (Classification)

- Standardize features.
- Split data into training and test data with test size of 20%.
- Fit the training data with four models.
- Use GridSerarchCV to find the best parameters of each model.
- Calculate the accuracy on the test data.
- Plot confusion matrix.
- Identify the best model and hyperparameters using Jaccard Score, F1 Score and Accuracy.



[GitHub URL – Lab8]

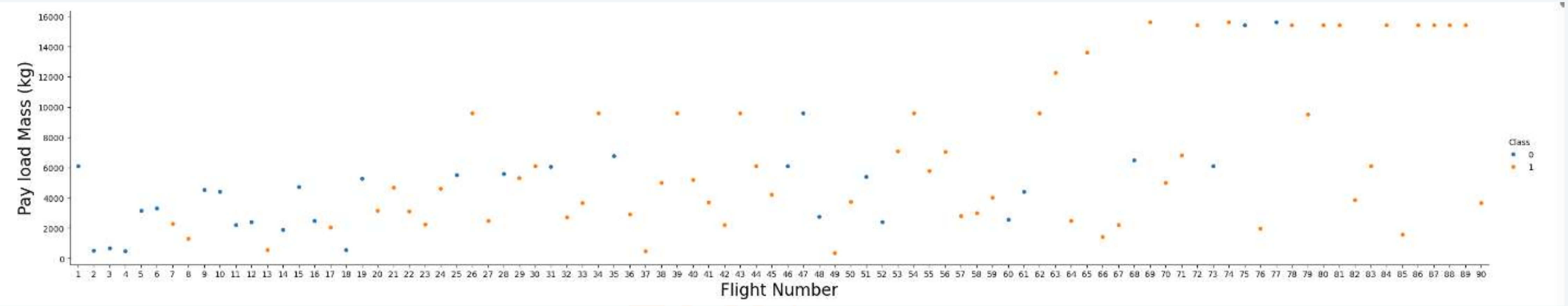
Results

- Exploratory Data Analysis results
  - Launch success has improved over time.
  - KSC LC-39A has the highest success rate among landing sites.
  - Orbits ES-L1, GEO, HEO and SSO have a 100% success rate.
- Interactive Analytics demo in screenshots
  - Most launch sites are near the equator, and all are close to the coast.
  - Launch sites are close enough to towns to facilitate the transport of people and goods, but far enough way from infrastructure to avoid accidents.
- Predictive analysis results
  - Decision Tree model is the best predictive model for the dataset.

Section 2

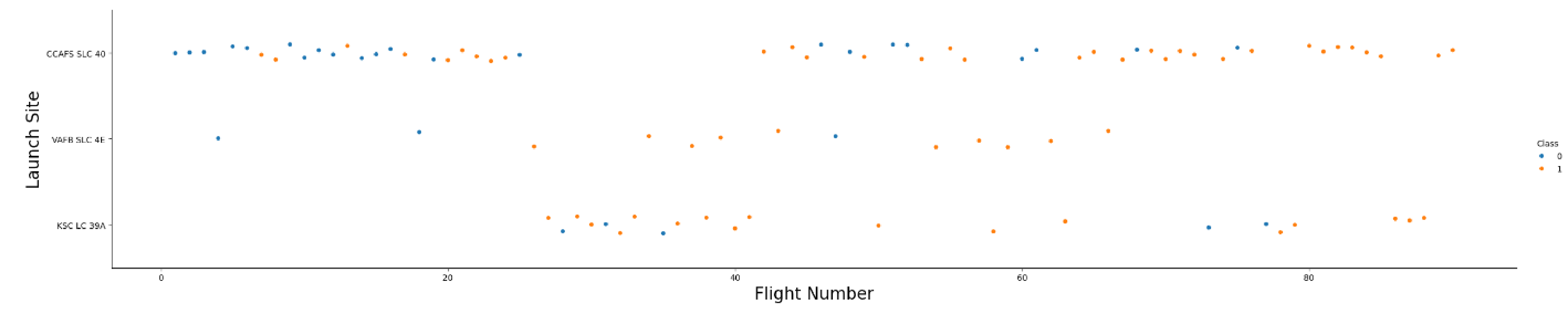
Insights drawn from EDA

Flight Number vs. Payload



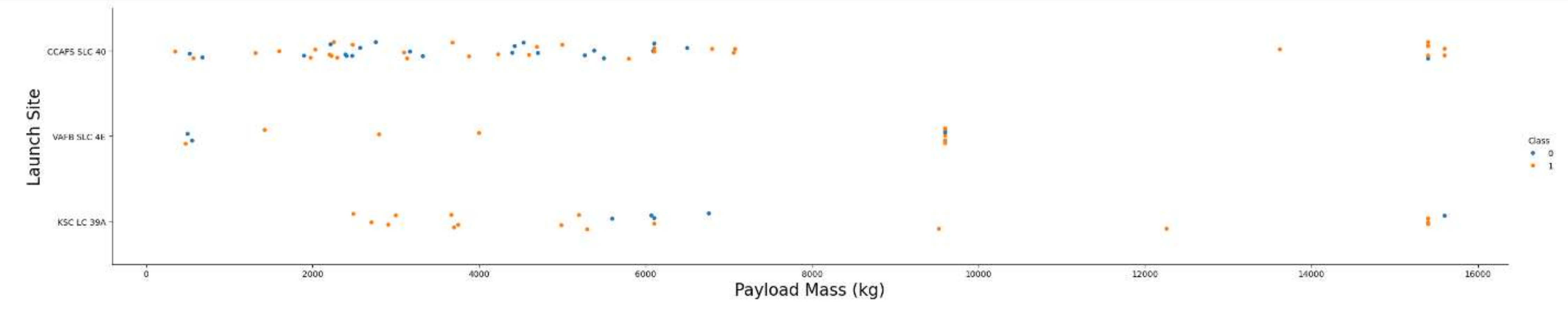
- Payload mass has increased as the flight number also grown.
- The first flights were lighter and often failed to land.
- More recent flights were heavier and succeeded in landing.

Flight Number vs. Launch Site



- Most of the flights were launched from CCAFS SLC 40.
- The first flights often failed to land, while more recent flights succeeded in landing.
- Show the screenshot of the scatter plot with explanations

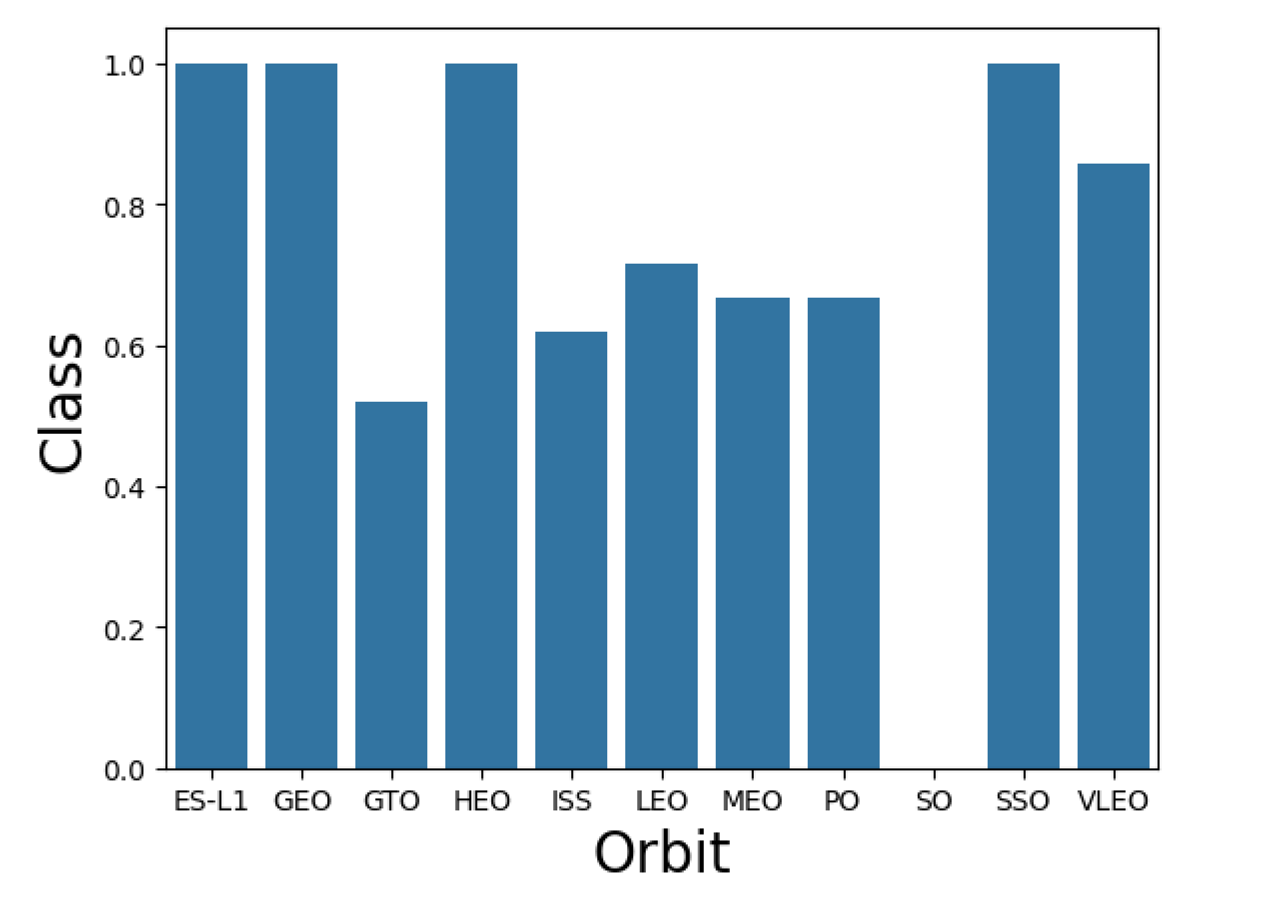
Payload vs. Launch Site



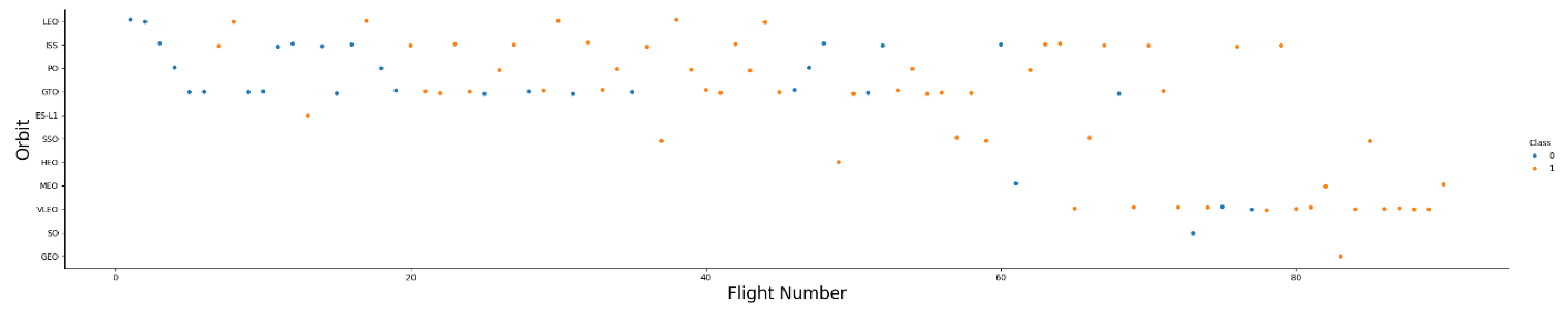
- A higher mass seems to be more successful in landing.
- CCAFS SLC 40 is a good site to launch heavy rockets successfully on landing.
- KSC LC 39A is the best site to launch lighter rockets.

Success Rate vs. Orbit Type

- ES-L1, GEO, HEO and SSO orbits had a 100% success rate.
- SO orbit had a 0% success rate.

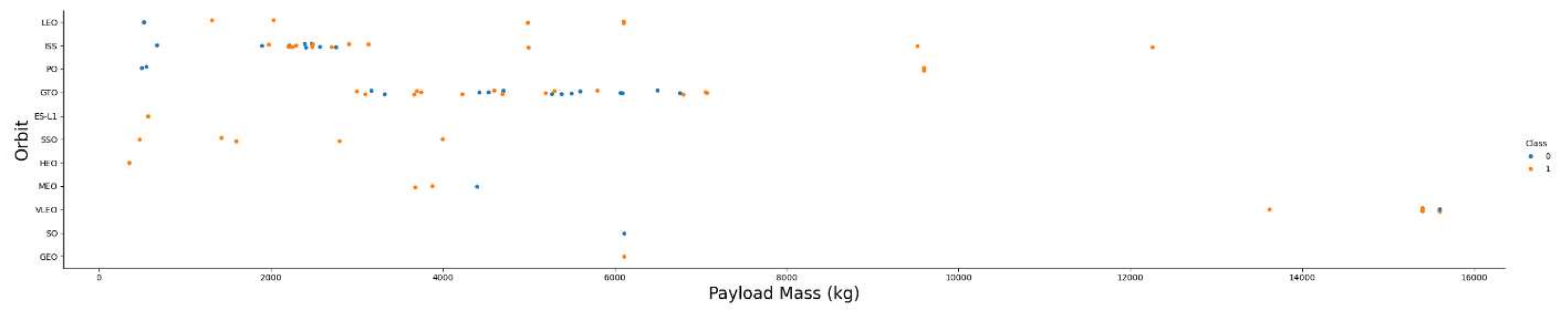


Flight Number vs. Orbit Type



- The success rate had increased with respect the flight number in all orbits.
- VLEO orbit had a great success rate in later flights.

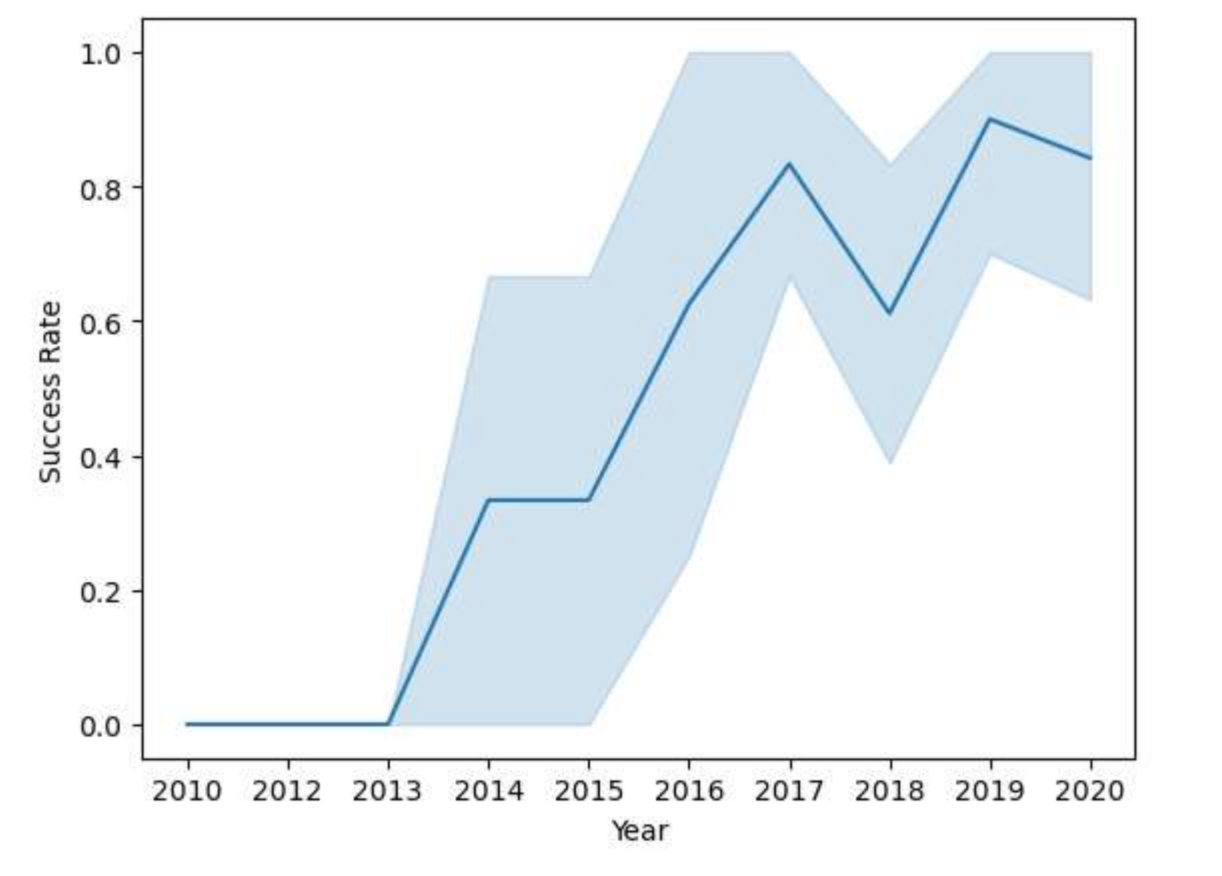
Payload vs. Orbit Type



- Heavy payloads had success at ISS, PO and VLEO orbits.
- SSO had a 100% success rate with payloads less than 4000 kg.

Launch Success Yearly Trend

- Since 2013, success rate increased.
- Success rate declined between 2017 and 2018, but recovered in 2019.



All Launch Site Names

- Find the names of the unique launch sites

```
%sql select distinct "Launch_Site" from SPACEXTABLE
```

- There are 4 distinct launch sites:

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`

```
%sql select * from SPACEXTABLE where "Launch_Site" like "CCA%" limit 5
```

- This query returned 5 records of CCAFS LC-40 site:

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

- Calculate the total payload carried by boosters from NASA

```
%sql select sum("PAYLOAD_MASS__KG_") as "Total_payload_Mass (Kg)" from SPACEXTABLE where "Customer" = "NASA (CRS)"
```

- This query sum all the payload mass of rockets who Customer was NASA.

Total_payload_Mass (Kg)
45596

Average Payload Mass by F9 v1.1

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

```
%sql select avg("PAYLOAD_MASS__KG_") as "Average Payload F9 v1.1" from SPACEXTABLE where "Booster_Version" = "F9 v1.1"
```

- The average payload mass of rocket with booster version F9 v1.1 was 2928 kg.

Average Payload F9 v1.1
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

```
%sql select min(date) as "First Successful Date" from SPACEXTABLE where "Landing_Outcome" = "Success (ground pad)"
```

- The first successful launch on ground pad was in 2015.

First Successful Date
2015-12-22

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

```
%sql select "Booster_Version" from SPACEXTABLE where "Landing_Outcome" = "Success (drone ship)" and "PAYLOAD_MASS__KG_" between 4000 and 6000
```

- This 4 booster versions have successfully landed on drone ship with a payload between 4000 and 6000 kg.

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

```
%sql select "Mission_Outcome", count(*) as "Total Number" from SPACEXTABLE group by "Mission_Outcome"
```

- There are 100 success missions and 1 failure.

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

Boosters Carried Maximum Payload

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

```
%sql select "Booster_Version" from SPACEXTABLE where "PAYLOAD_MASS__KG_" = (select max("PAYLOAD_MASS__KG_") from SPACEXTABLE)
```

- There are 12 booster versions rockets that have carried the maximum payload mass.

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

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%%sql select substr(Date, 6,2) as Month, "Landing_Outcome", "Booster_Version", "Launch_Site"
from SPACEXTABLE where "Landing_Outcome" = "Failure (drone ship)" and substr(Date, 0,5) = "2015"
```

- In 2015, only 2 flights failed in drone ship, in January and April.

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

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

```
%%sql select "Landing_Outcome", count(*) as Rank
from SPACEXTABLE
where Date between "2010-06-04" and "2017-03-20"
group by "Landing_Outcome"
order by Rank desc
```

Landing_Outcome	Rank
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

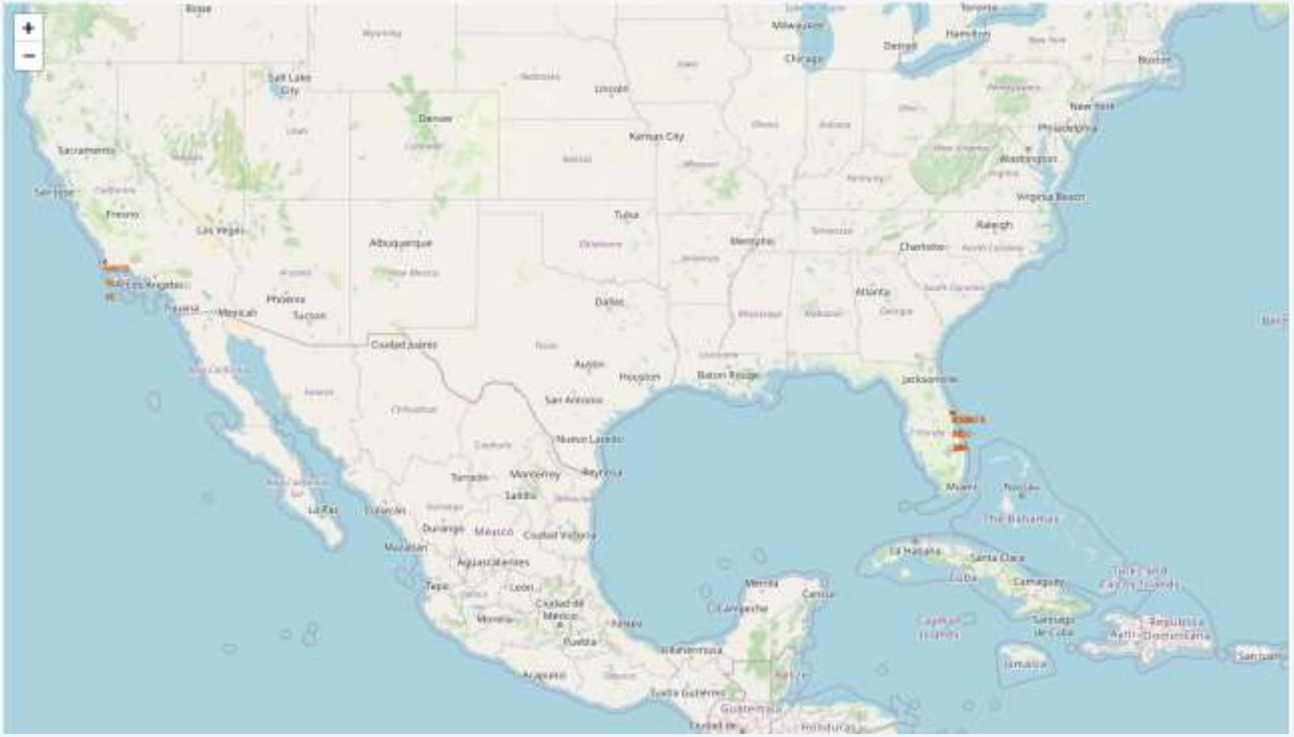
- The most common landing outcome was no attempt.

Section 3

Launch Sites Proximities Analysis

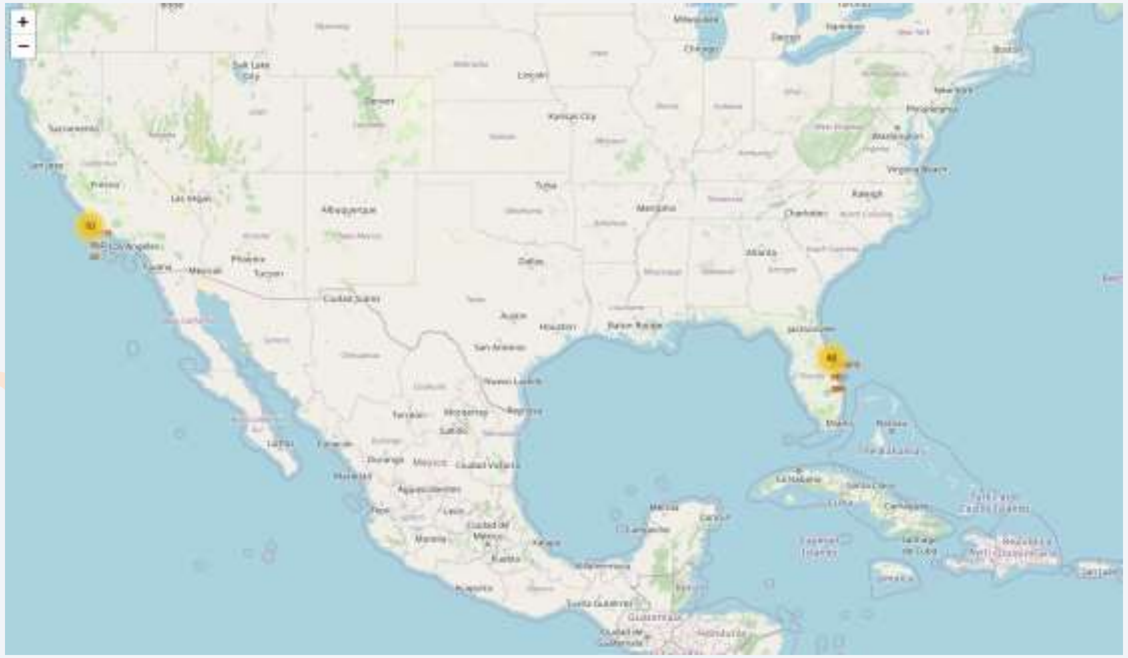
Location of Launch Sites

- All the launch sites (red marks) are near the coast and as far south in the United States to be closer to the equator.



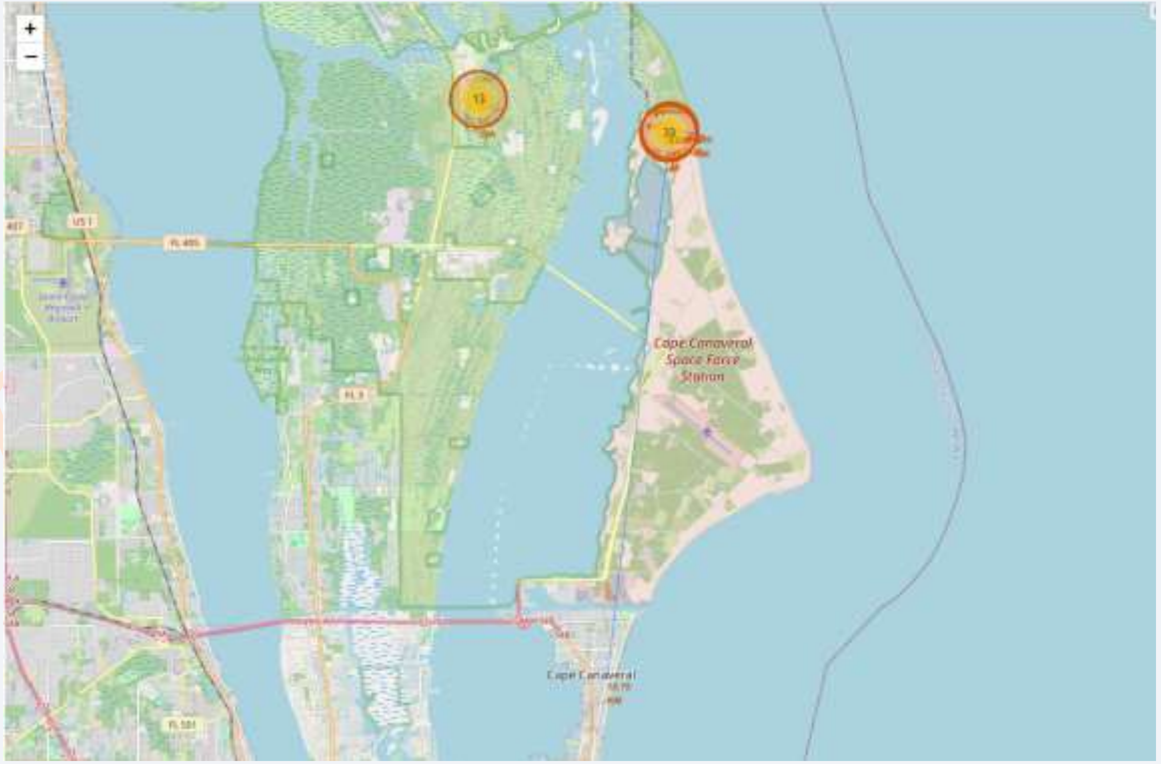
Launch Outcomes by Launch Site

- Yellow circles shows number of launch in each launch site.
- A green mark shows a successful launches and a red mark shows an unsuccessful launches.



Launch Site Proximities

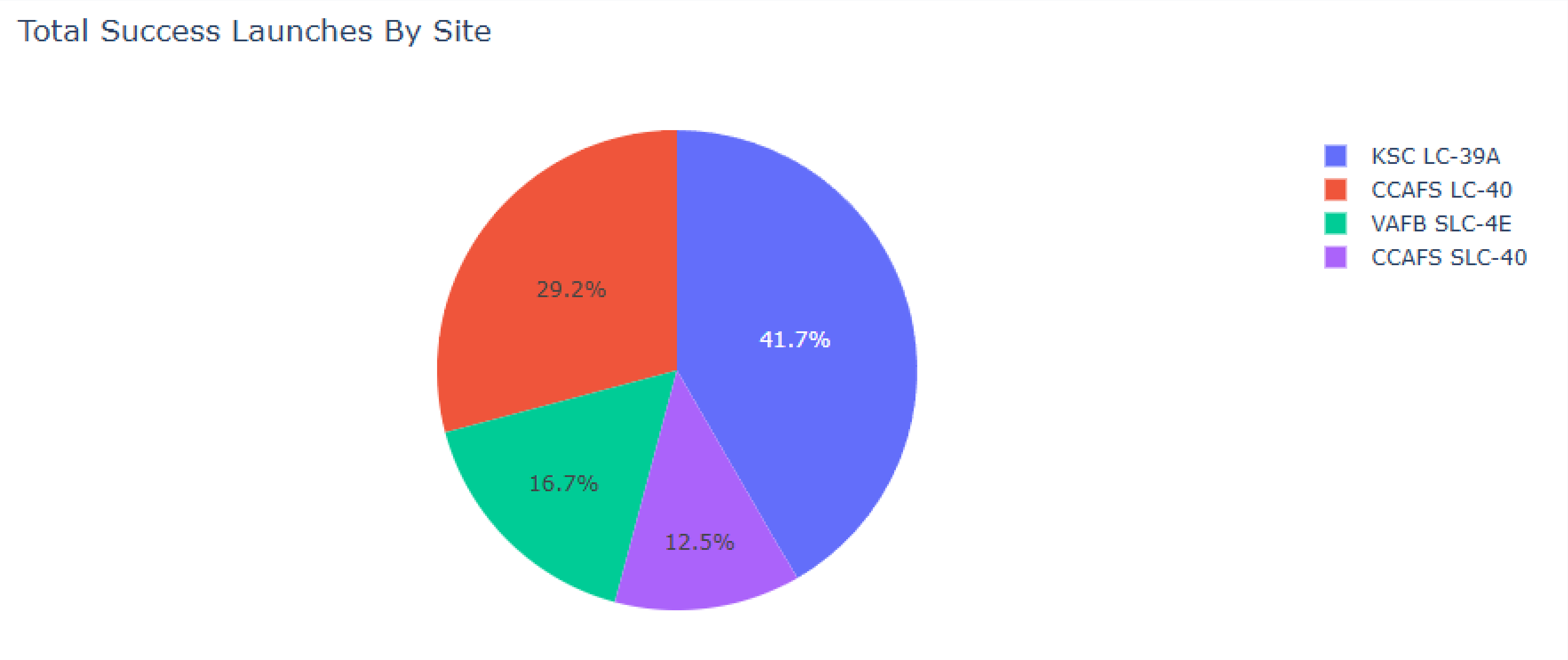
- Launch sites are close to coasts.
- Launch sites are close enough to towns to facilitate the transport of people and goods, but far enough way from infrastructure to avoid accidents.



Section 4

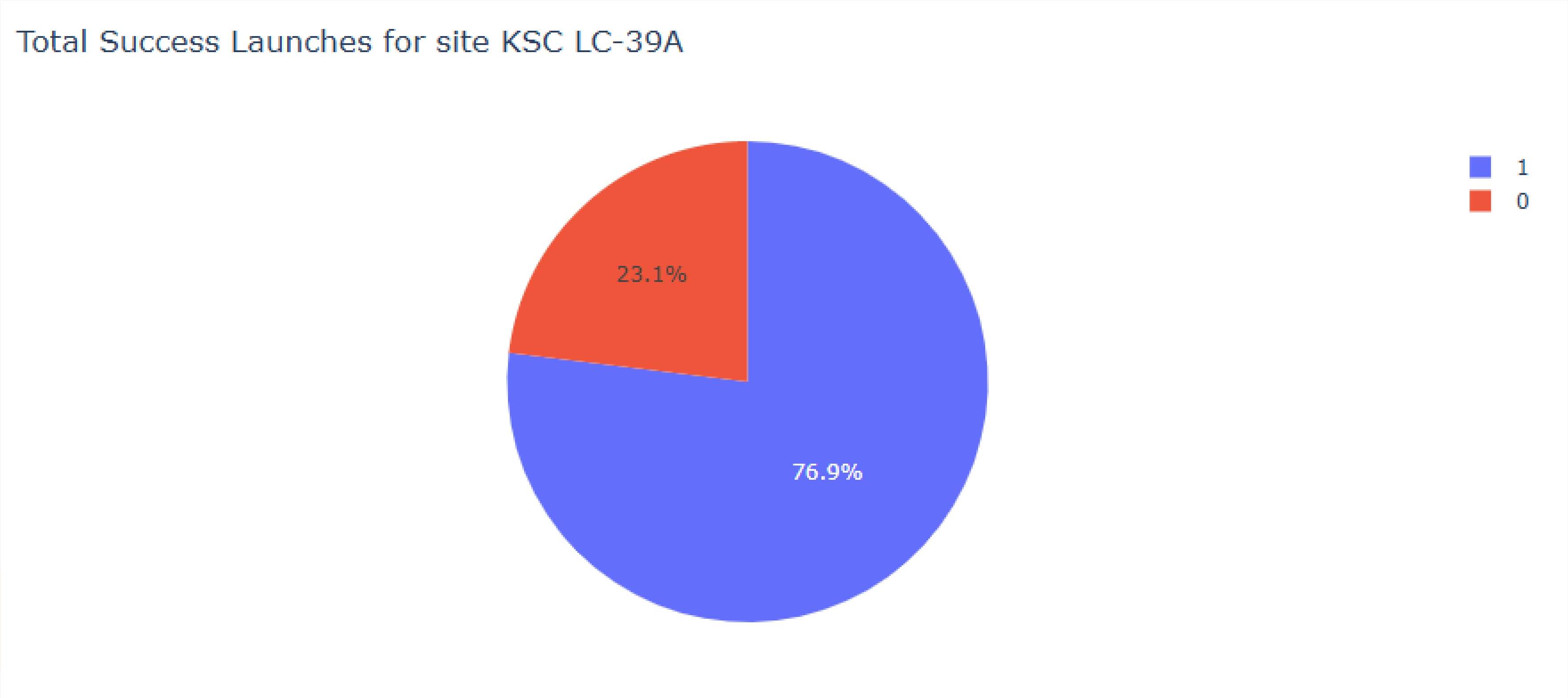
Build a Dashboard with Plotly Dash

Total Success Launches by Site



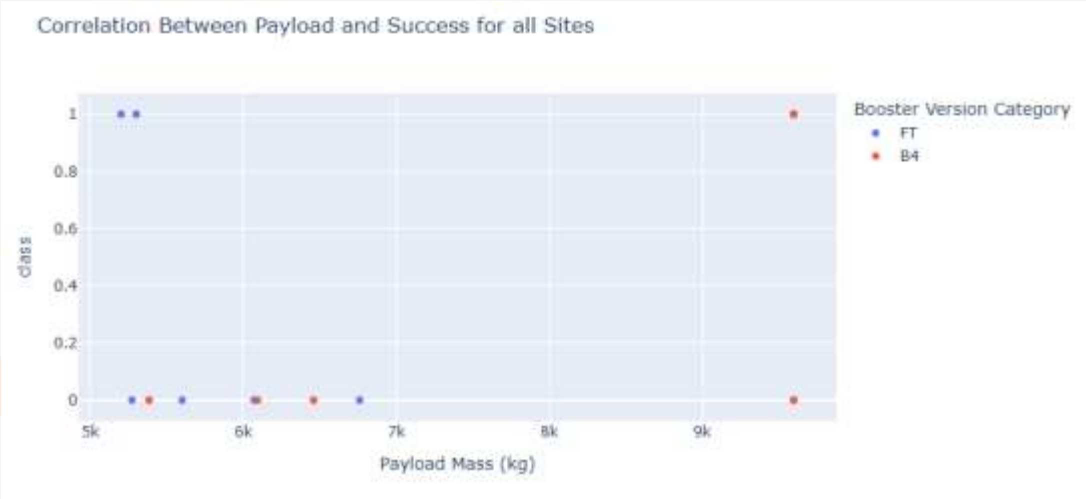
- The launch site with the most successful landing was KSC LC-39A.

Launch Success Ratio for KSC LC-39A



- KSC LC-39A is the launch site with the highest success ratio (76.9%).

Payload Mass vs. Launch Outcome by Range



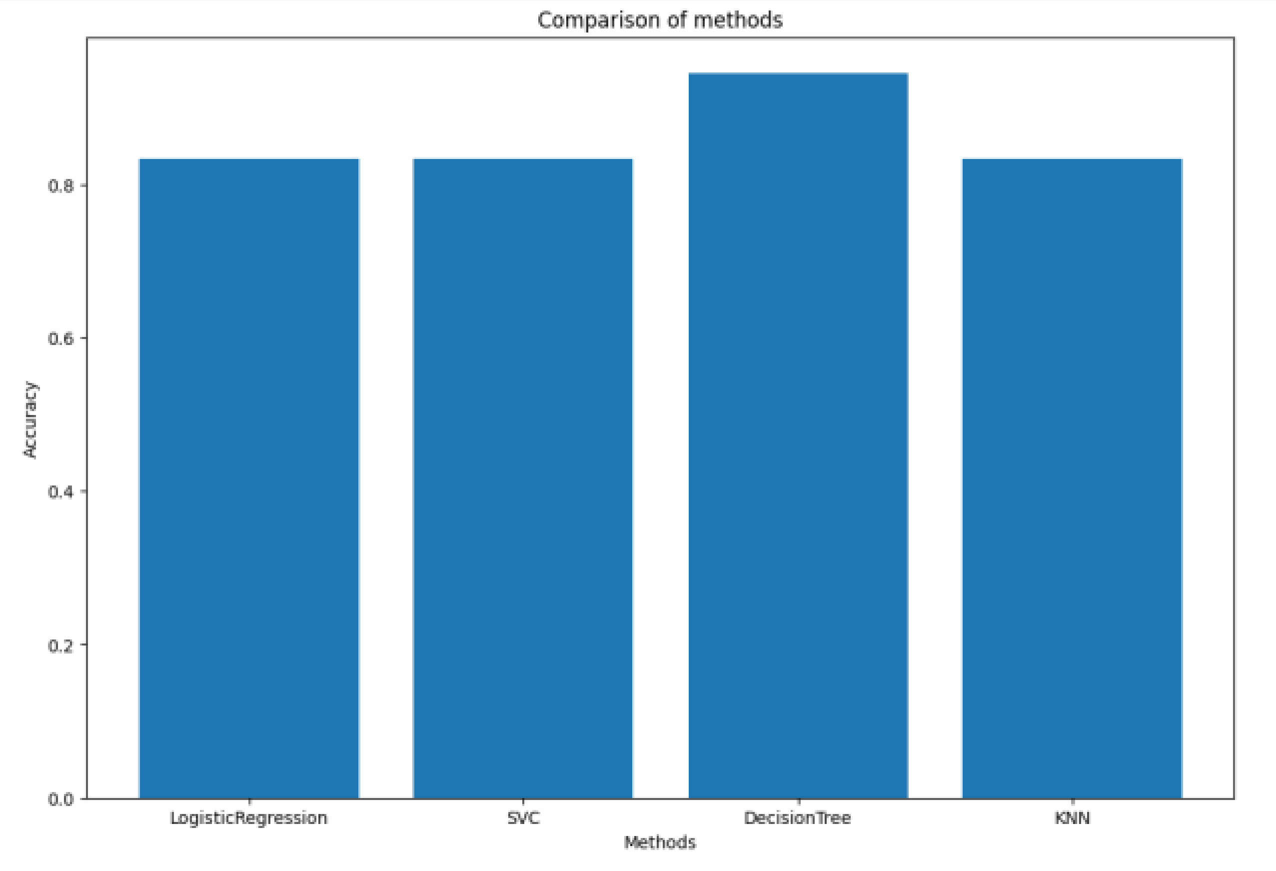
- Payloads less than 5000 kg has a high success ratio than heaviest payloads.

Section 5

Predictive Analysis (Classification)

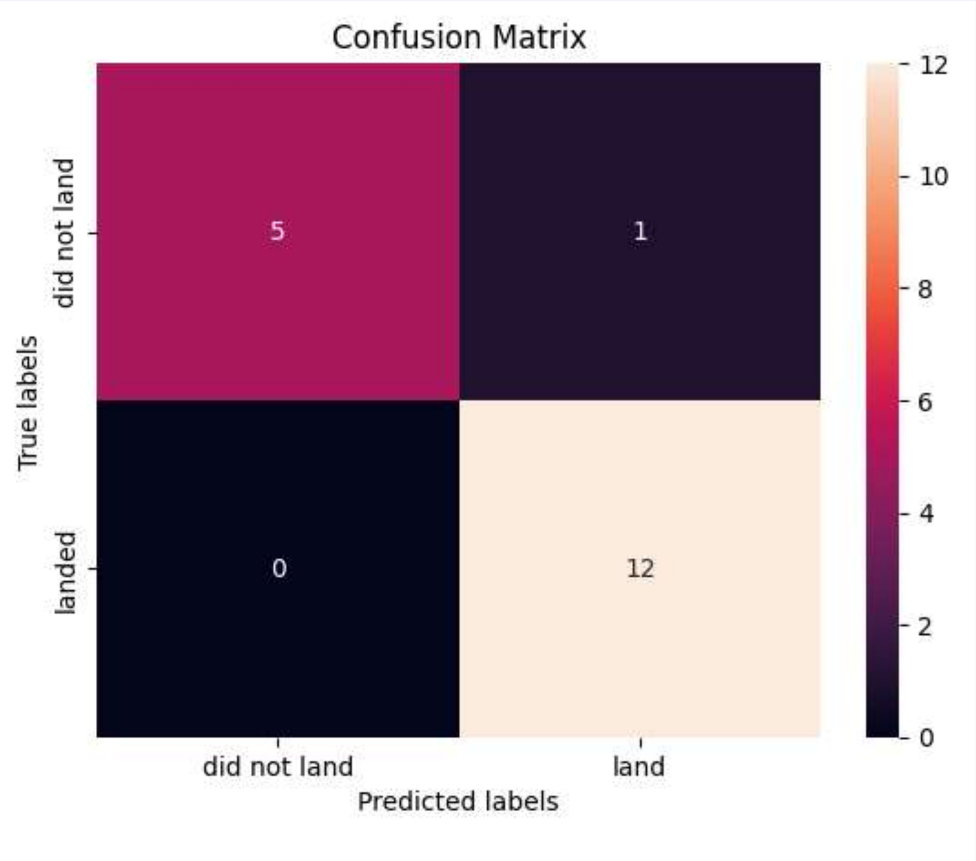
Classification Accuracy

- Decision Tree was the model with the best accuracy (94%).



Confusion Matrix

- The Decision Tree model correctly predicted 12 true positives and 5 true negatives. Only 1 false positive was predicted.



Conclusions

- The yearly trend was a high success rate since 2013.
- ES-L1, GEO, HEO and SSO orbits had a 100% success rate.
- All the launch sites are close to the coast and near the equator.
- KSC LC-39A had the highest success ratio (76.9%).
- Payloads less than 5000 kg has a high success ratio than heaviest payloads.
- Decision Tree was the higher accuracy (94%).

Thank you!