Winning the Space Race with Data Science

James Porter
15 Aug 2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis Visualization
  - Visual Analytics with Folium
  - Interactive Dashboard with Plotly
  - Predictive Analysis
- Summary of all results
  - Exploratory Data Analysis Results
  - Interactive Analytics Demo
  - Predictive Analysis Results

Introduction

- Project Background and Context:
  - In this capstone, we will predict if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.
- Questions wanting to answer:
  - Price per launch
  - Probability of a successful landing

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX REST API
  - Web Scraping of Wikipedia
- Perform data wrangling
  - Filtering Data
  - Dealing with Missing Values
  - One Hot Encoding to Prepare Data
- Perform exploratory data analysis (EDA) using visualization and SQL

Methodology

Executive Summary

- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Collected data was normalized, divided into training and test data sets, and applied to four different classification models.

Data Collection

- The process of gathering data encompassed a blend of API requests from SpaceX's REST API and the extraction of information through web scraping from a table within SpaceX's Wikipedia page. The utilization of both these methods was essential to acquire a comprehensive dataset for a more intricate analysis.
  - The data columns were populated using SpaceX's REST API, including FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, and Latitude.
  - Furthermore, the data columns sourced from Wikipedia web scraping encompassed Flight Number, Launch Site, Payload, PayloadMass, Orbit, Customer, Launch Outcome, Booster Version, Booster Landing, Date, and Time.

Data Collection – SpaceX API

- SpaceX provides an API to allow access to it's data.
- This data was collected and processed to use as part of our dataset.

- GITHUB – Data Collection API

Request the SpaceX API and collect the launch data

Filter Data to only show Falcon 9 rocket data

Deal with missing values

Data Collection - Scraping

- Wikipedia also contains a wealth of information about SpaceX launches.

Locate the Falcon 9 launch data Wiki Page

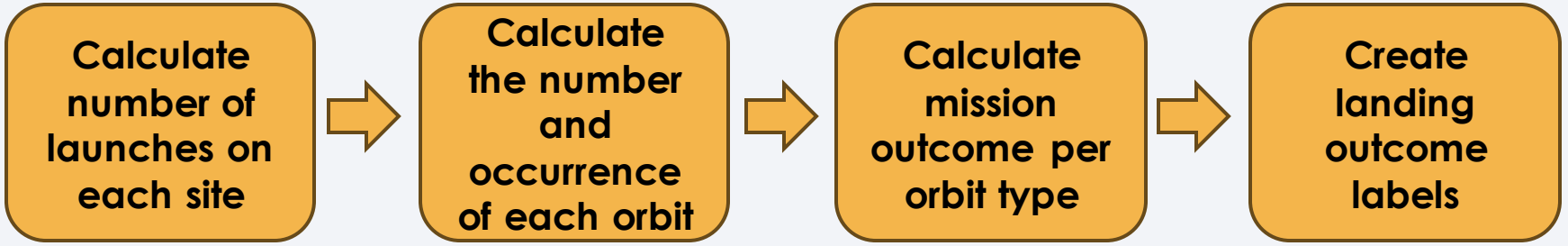
Extract names from the HTML table header

Create a Data Frame from the collected HTML tables

- GITHUB – Data Collection Web Scraping

Data Wrangling

- Here the data was explored to identify the different types of data that we could use to formulate identifying factors that contributed to either successful or failed landing attempts.



- GITHUB – Data Wrangling

EDA with Data Visualization

- The following charts were plotted. These were chosen to see how different variables affected the landing outcome:
  - FlightNumber vs. PayloadMass
  - FlightNumber vs LaunchSite
  - Payload Vs. LaunchSite
  - sucess rate of each orbit
  - FlightNumber and Orbit type.
  - Payload vs. Orbit

- GITHUB – EDA with Data Visualization

EDA with SQL

- SQL queries performed:
  - `SELECT DISTINCT Launch_Site FROM SPACEXTBL`
  - `SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5;`
  - `SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer = 'NASA (CRS)'`
  - `SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version LIKE 'F9 v1.1';`
  - `SELECT MIN(Date) AS MIN_DATE FROM SPACEXTBL WHERE "Landing_Outcome" = 'Success (ground pad)';`
  - `SELECT Booster_Version FROM SPACEXTBL WHERE (PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000)`
  - `AND (Landing_Outcome = 'Success (drone ship)');`
  - `SELECT Mission_Outcome, COUNT(Mission_Outcome) AS counts FROM SPACEXTBL GROUP BY Mission_Outcome;`
  - `SELECT Booster_Version, PAYLOAD_MASS__KG_ FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);`
  - `SELECT Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTBL WHERE Landing_Outcome = 'Failure (drone ship)' AND YEAR(Date) = '2015'`
  - `SELECT Landing_Outcome, COUNT(*) AS LandingCounts FROM SPACEXTBL WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'`
  - `GROUP BY Landing_Outcome`
  - `ORDER BY COUNT(*) DESC;`

- GITHUB – EDA with Data Visualization

Interactive Map with Folium

- Markers Indicating Launch Sites
- • Added blue circle at NASA Johnson Space Center's coordinate with a popup label
- showing its name using its latitude and longitude coordinates
- • Added red circles at all launch sites coordinates with a popup label
- showing its name using its name using its latitude and longitude coordinates

- Colored Markers of Launch Outcomes
- • Added colored markers Of Successful (Green) and Unsuccessful (red) Launches at each launch site to
- show which launch sites have high success rates
- Distances Between a Launch Site to Proximities
- • Added colored lines To show distance between launch site CCAFS SLC- 40 and its proximity to the nearest
- coastline, railway, highway, and city

Dashboard with Plotly Dash

- Dropdown List with Launch Sites
- • Allow user to select all launch sites or a certain launch site

- Pie Chart Showing Successful Launches
- • Allow user to see successful and unsuccessful launches as a percent of the total

- Slider of Payload Mass Range
- • Allow user to select payload mass range

- Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version
- • Allow user to see the correlation between Payload and Launch Success

- GITHUB – Plotly Dashboard App

Predictive Analysis (Classification)

- We loaded the data using numpy and pandas, transformed the data, split
- our data into training and testing.

- We built different machine learning models and tune different
- hyperparameters using GridSearchCV.

- We used accuracy as the metric for our model, improved the model using
- feature engineering and algorithm tuning.

- We found the best performing classification model.

- GITHUB – Predictive Analysis

Results

Exploratory Data Analysis

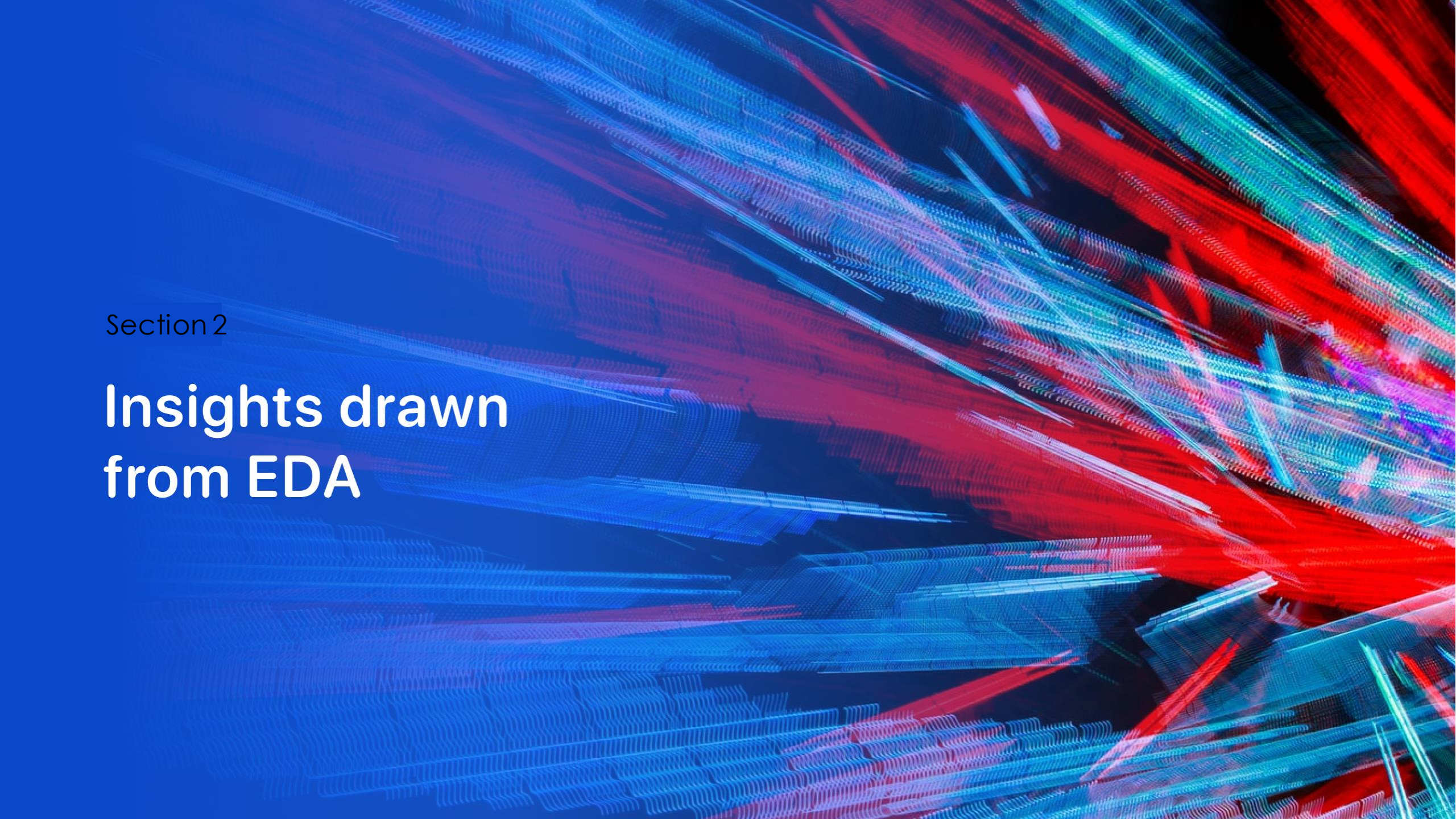
- Launch success has improved over time
- KSC LC-39A has the highest success rate among landing sites
- Orbits ES-L1, GEO, HEO and SSO have a 100% success rate

Visual Analytics

- Most launch sites are near the equator, and all are close to the coast
- Launch sites are far enough away from anything a failed launch can damage
- (city, highway, railway), while still close enough to bring people and material
- to support launch activities

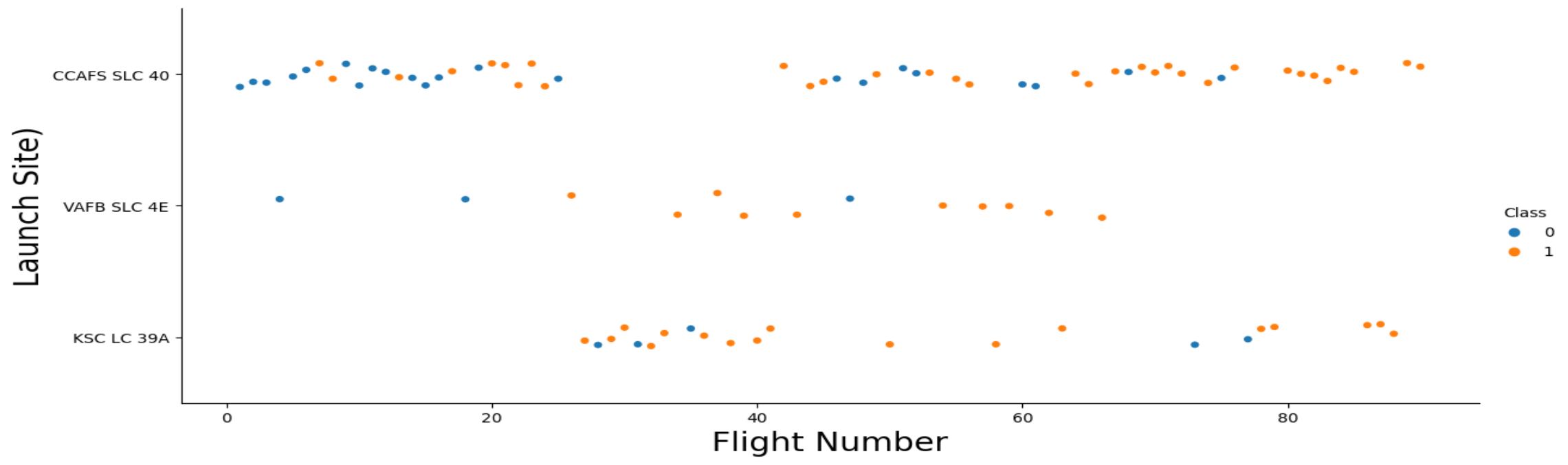
Predictive Analytics

- Decision Tree model is the best predictive model for the dataset


Section 2
Insights drawn from EDA

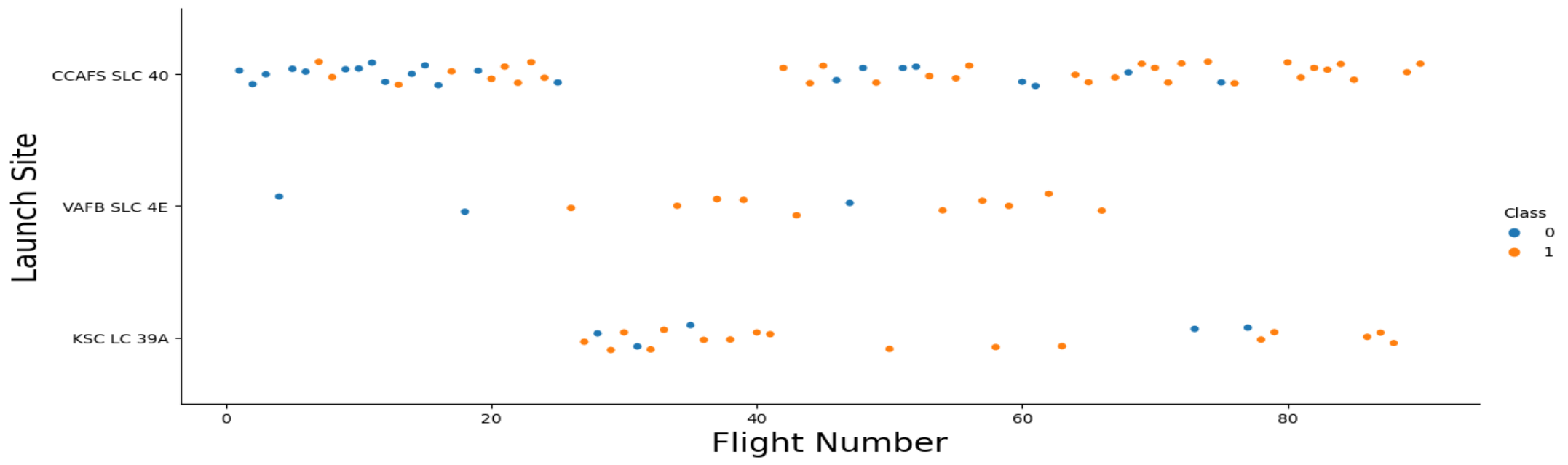
Flight Number vs. Launch Site

- From the plot, we found that the more flights at a launch site, the
- greater the success rate at a launch site.



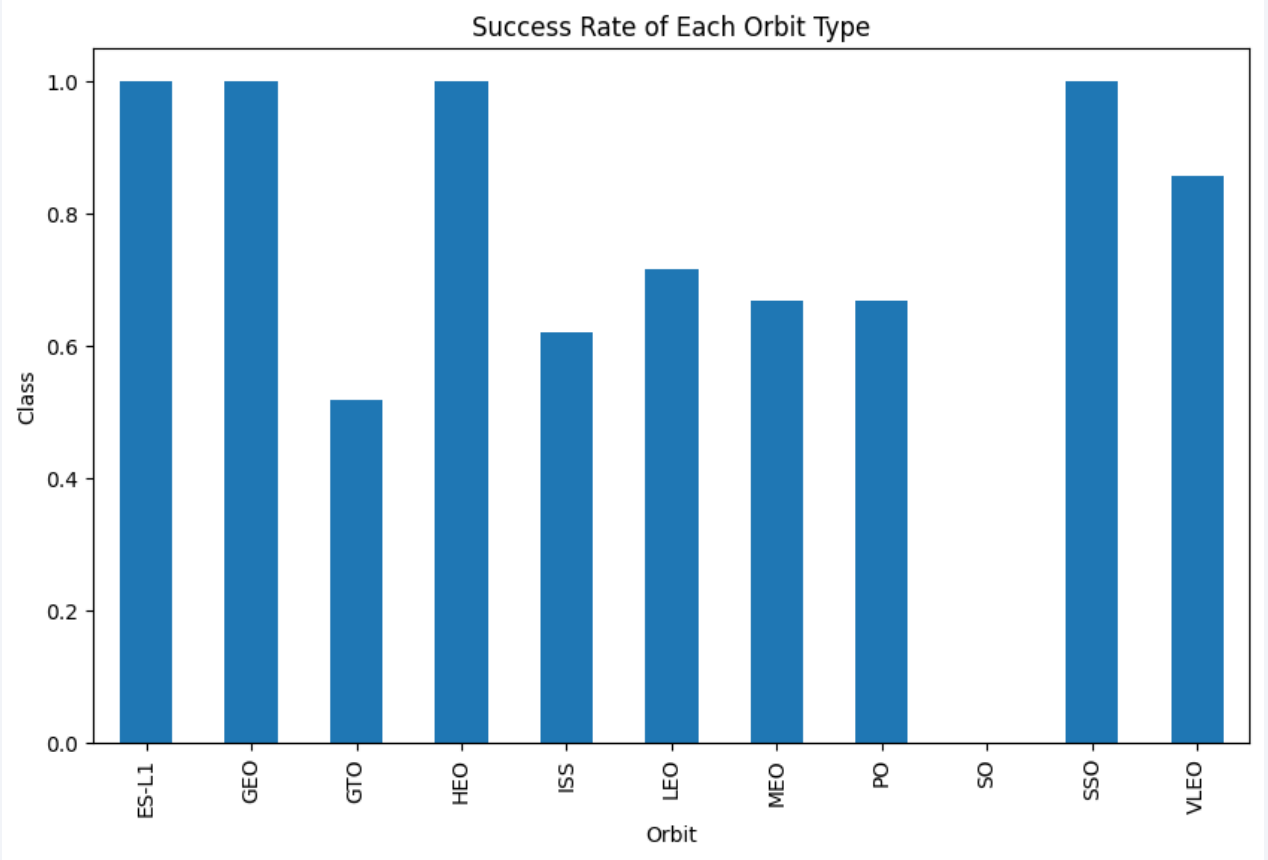
Payload vs. Launch Site

- THE GREATER THE PAYLOAD, THE MORE SUCCESSFUL THE LAUNCH WAS AT SITE CCAFS SLC-40



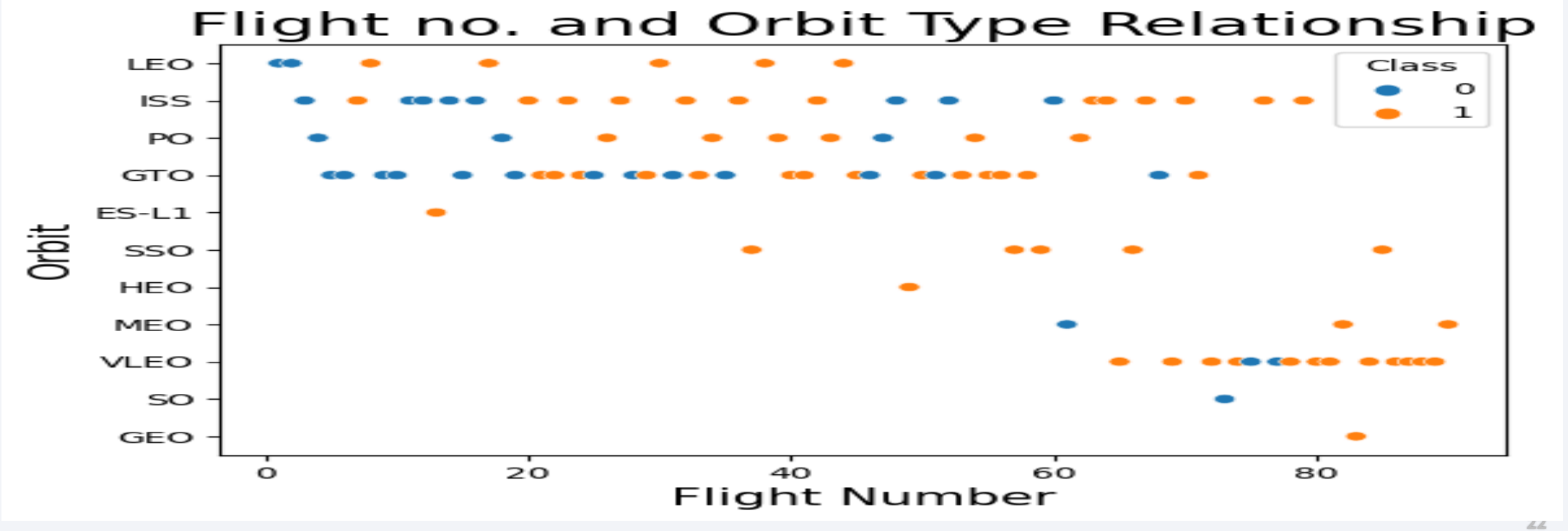
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO,
- SSO, VLEO HAD THE HIGHEST
- SUCCESS RATE.



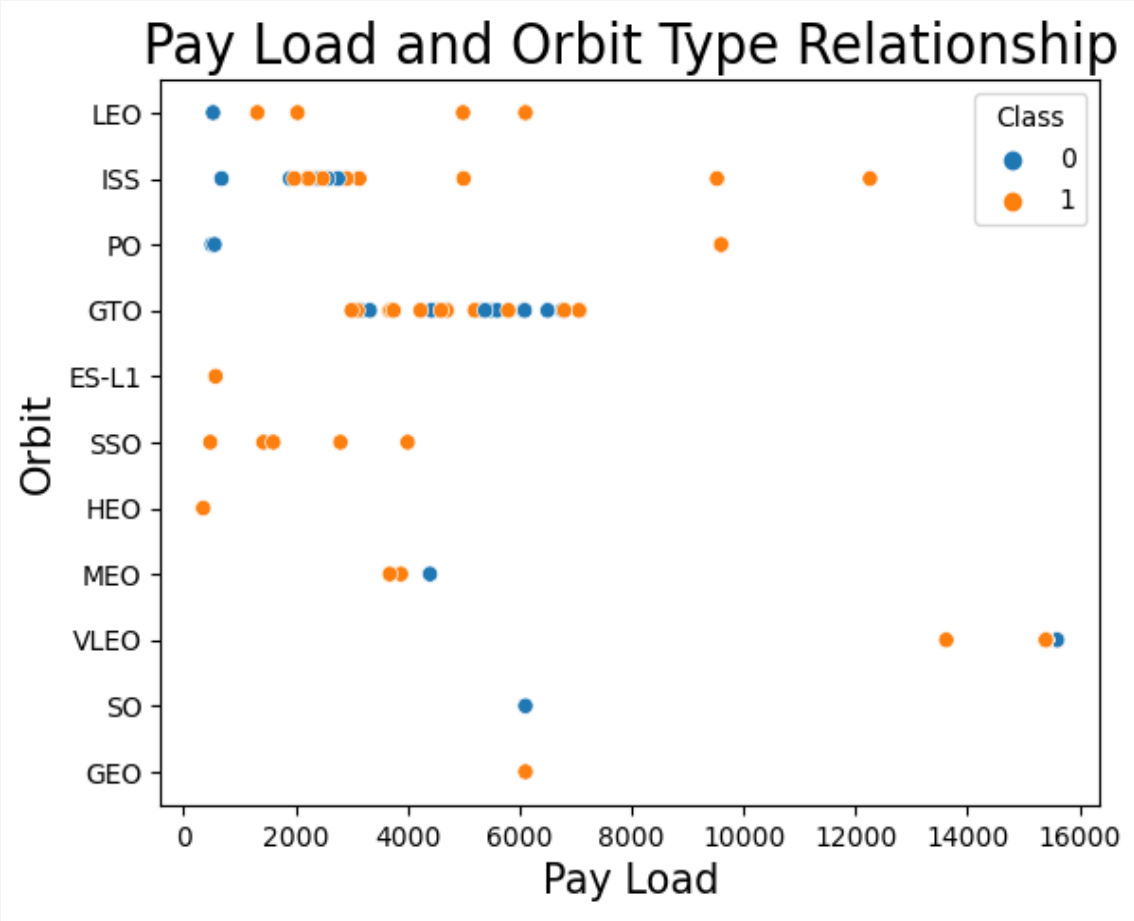
Flight Number vs. Orbit Type

- We see that in the LEO orbit, success is related to the number of flights whereas in the GTO orbit, there is no relationship between flight number and the orbit.



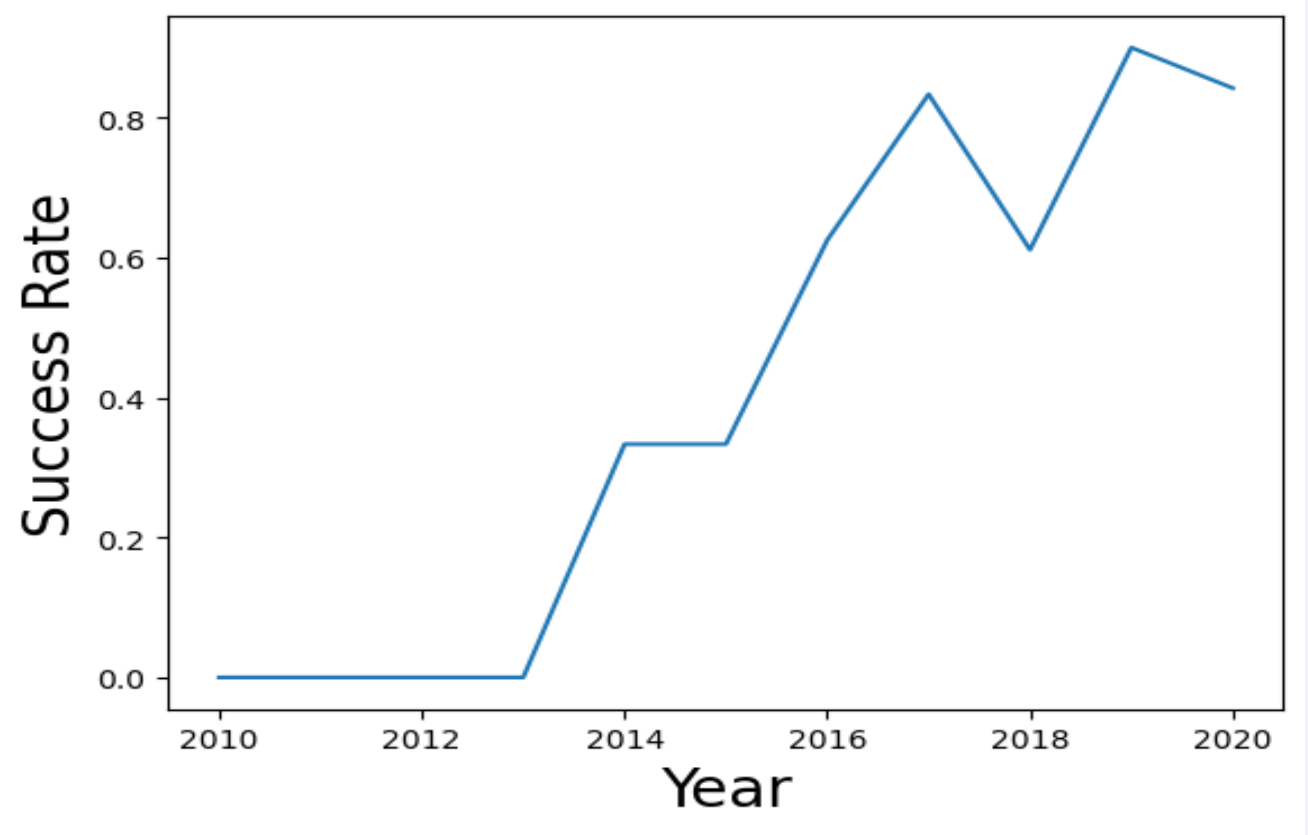
Payload vs. Orbit Type

- PO, LEO AND ISS ORBITS HAD MORE SUCCESSFUL LANDINGS WITH A HEAVIER PAYLOAD.



Launch Success Yearly Trend

- We can see the launch success has increased over time.



All Launch Site Names

- We used the key word **DISTINCT** to show only
- unique launch sites from the SpaceX data.

```
%%sql

select distinct Launch_Site from spacextbl
```

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

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

Launch Site Names Begin with 'CCA'

- We used WHERE to display 5 records where launch sites begin with
- `CCA`

```
%%sql
select * from spacextbl where Launch_Site LIKE 'CCA%' limit 5;
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
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

- We calculated the total payload using SUM and WHERE.

```
%%sql

select sum(PAYLOAD_MASS__KG_) from spacextbl where Customer = 'NASA (CRS)'
```

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

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- We calculated the average
- payload mass with AVG and WHERE

```
%%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

- Dates of the first successful landing date MIN and WHERE

```
%%sql

select min(Date) as min_date from spacextbl where "Landing_Outcome" = 'Success (ground pad)';
```

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

min_date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- We found that the dates of the first successful landing outcome on
- ground pad was 22nd December 2015

```
%%sql

select Booster_Version from spacextbl where (PAYLOAD_MASS__KG_> 4000 and PAYLOAD_MASS__KG_ < 6000)
and (Landing_Outcome = 'Success (drone ship)');
```

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

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 FAILURE MISSION OUTCOMES WERE FOUND COUNT AND FROM

```
%%sql

select Mission_Outcome, count(Mission_Outcome) as counts from spacextbl group by Mission_Outcome;
```

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

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

Boosters Carried Maximum Payload

- THE NAMES OF THE BOOSTER WHICH HAVE CARRIED THE MAXIMUM PAYLOAD MASS WERE FOUND USING MAX

```
%%sql
select Booster_Version, PAYLOAD_MASS__KG_ from spacextbl where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from spac
```

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

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

2015 Launch Records

- The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 using LIKE and BETWEEN.

List the failed landing_outcomes in drone ship, their booster versions,

```
task_9 = '''
        SELECT BoosterVersion, LaunchSite, LandingOutcome
        FROM SpaceX
        WHERE LandingOutcome LIKE 'Failure (drone ship)'
            AND Date BETWEEN '2015-01-01' AND '2015-12-31'
        '''
create_pandas_df(task_9, database=conn)
```

	boosterversion	launchsite	landingoutcome
0	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
1	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- The count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, ranked in descending order using COUNT, GROUP BY, and ORDER BY

```
%%sql

select Landing_Outcome, count(*) as LandingCounts from spacextbl where Date between '2010-06-04' and '2017-03-20'
group by Landing_Outcome
order by count(*) desc;
```

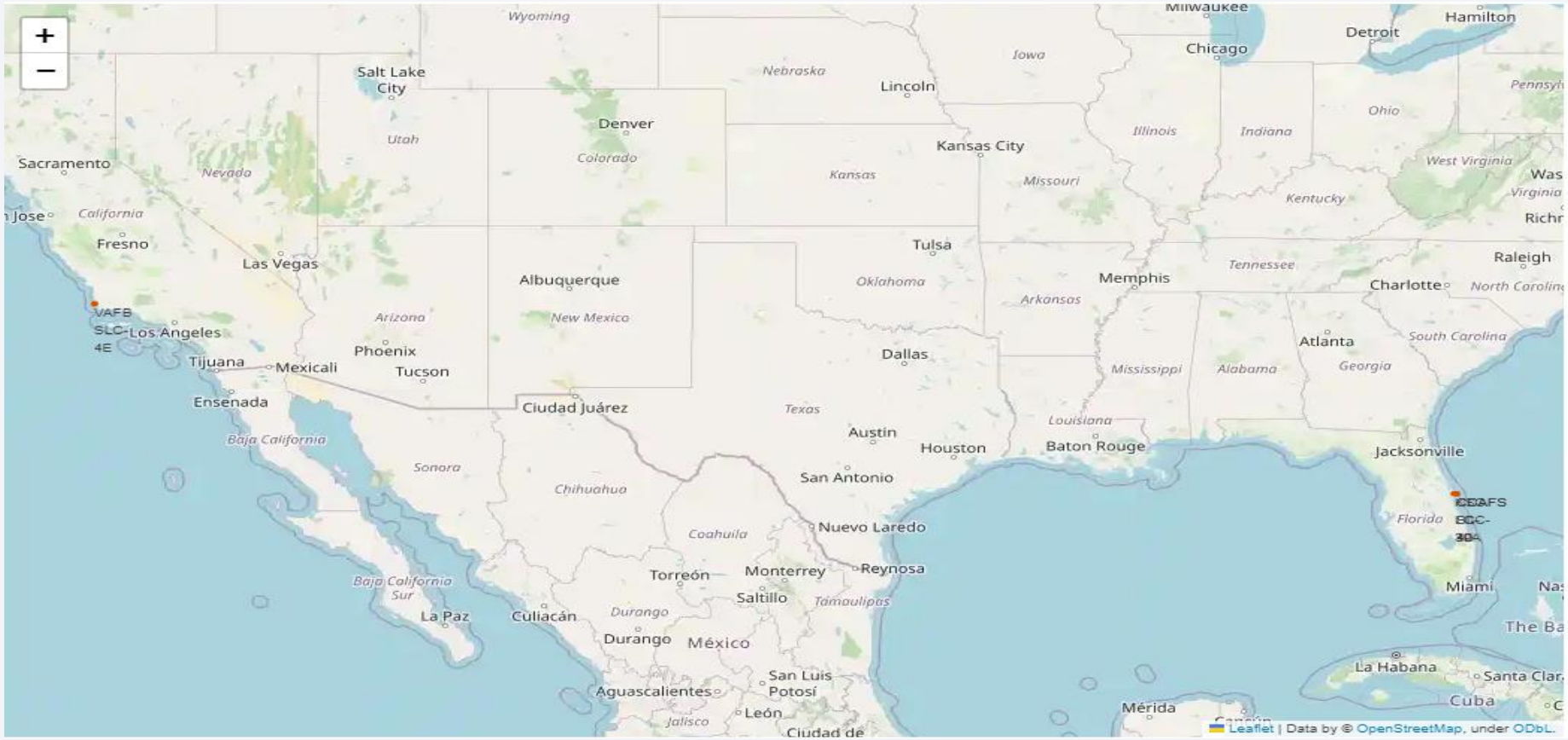
```
 * sqlite:///my_data1.db
Done.
```

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


Section 3
Launch Sites
Proximities Analysis

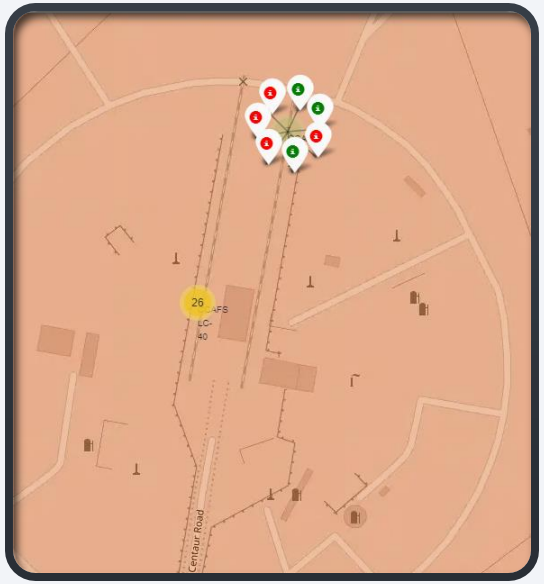
Launch Site Locations

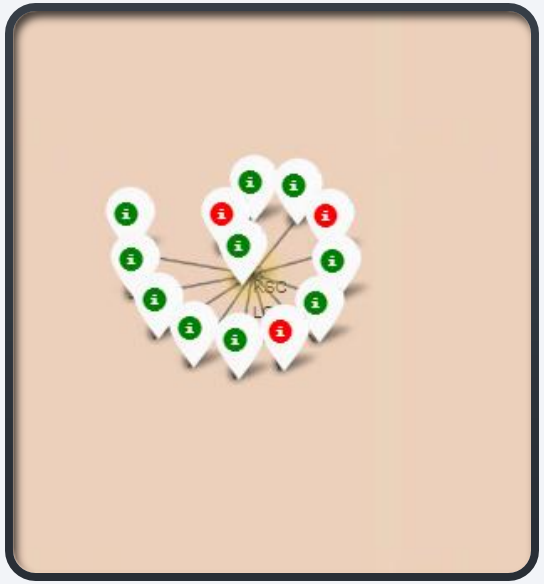
- Launch locations are on the East Coast in Florida, and the West coast in California.

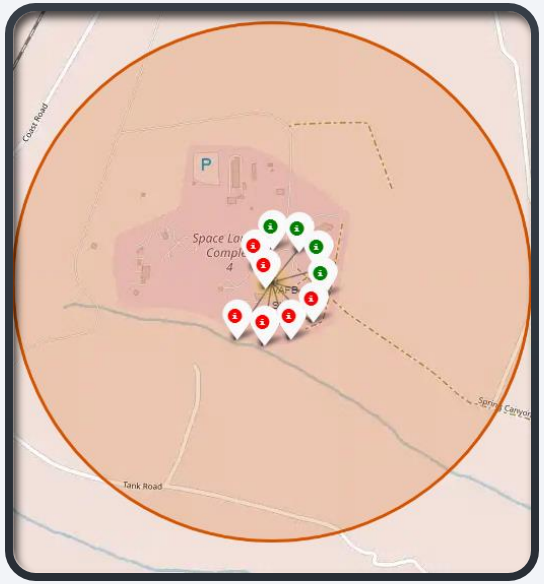


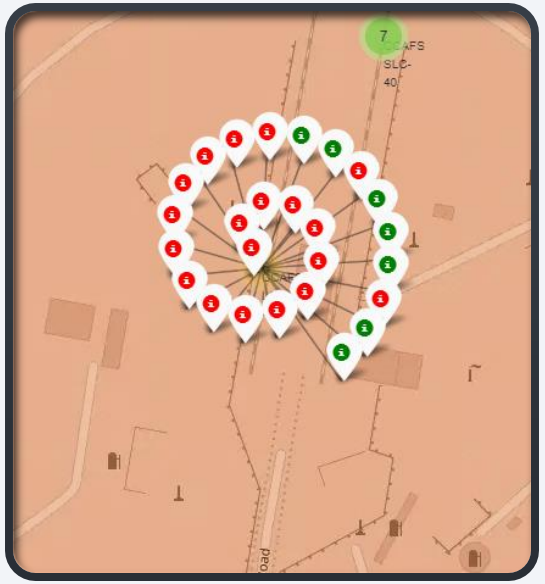
Markers Showing Launch Sites

- Sites with color coded markers indicating successful/failed launches per site.



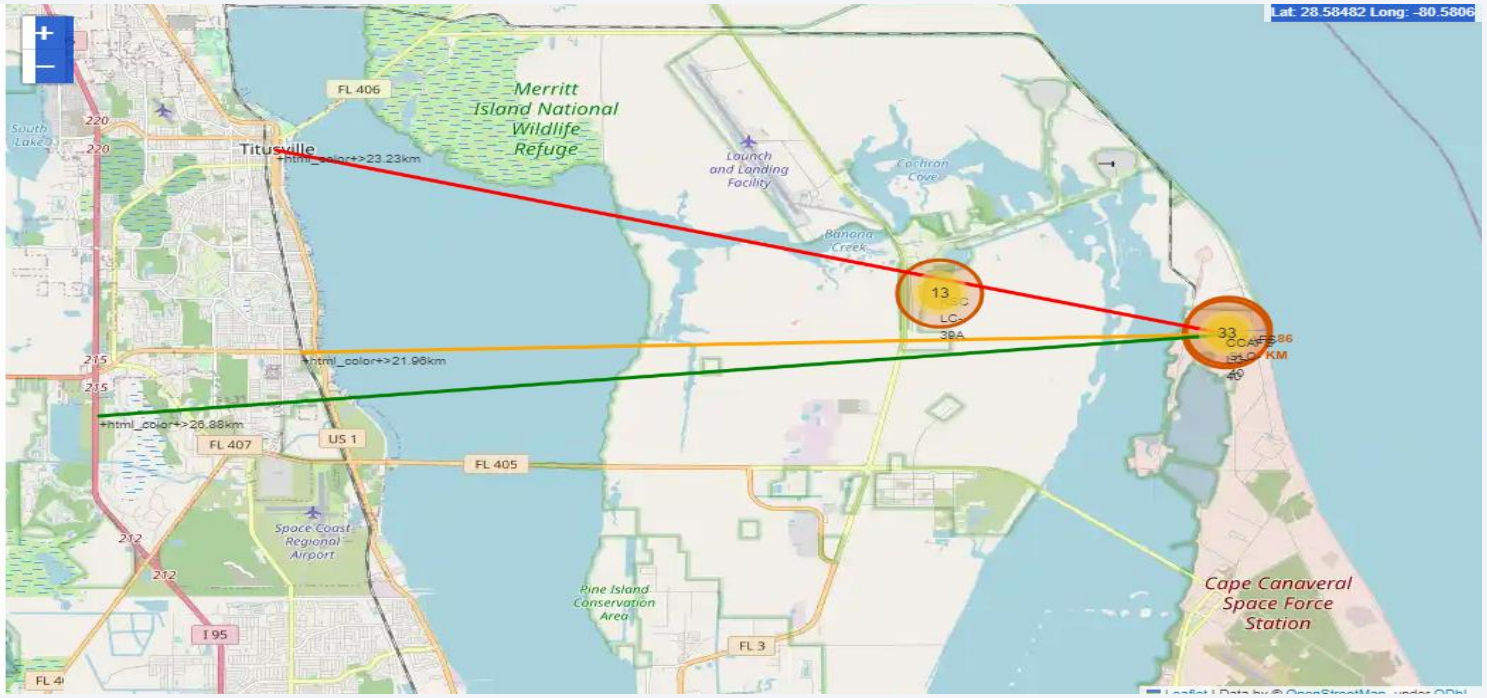


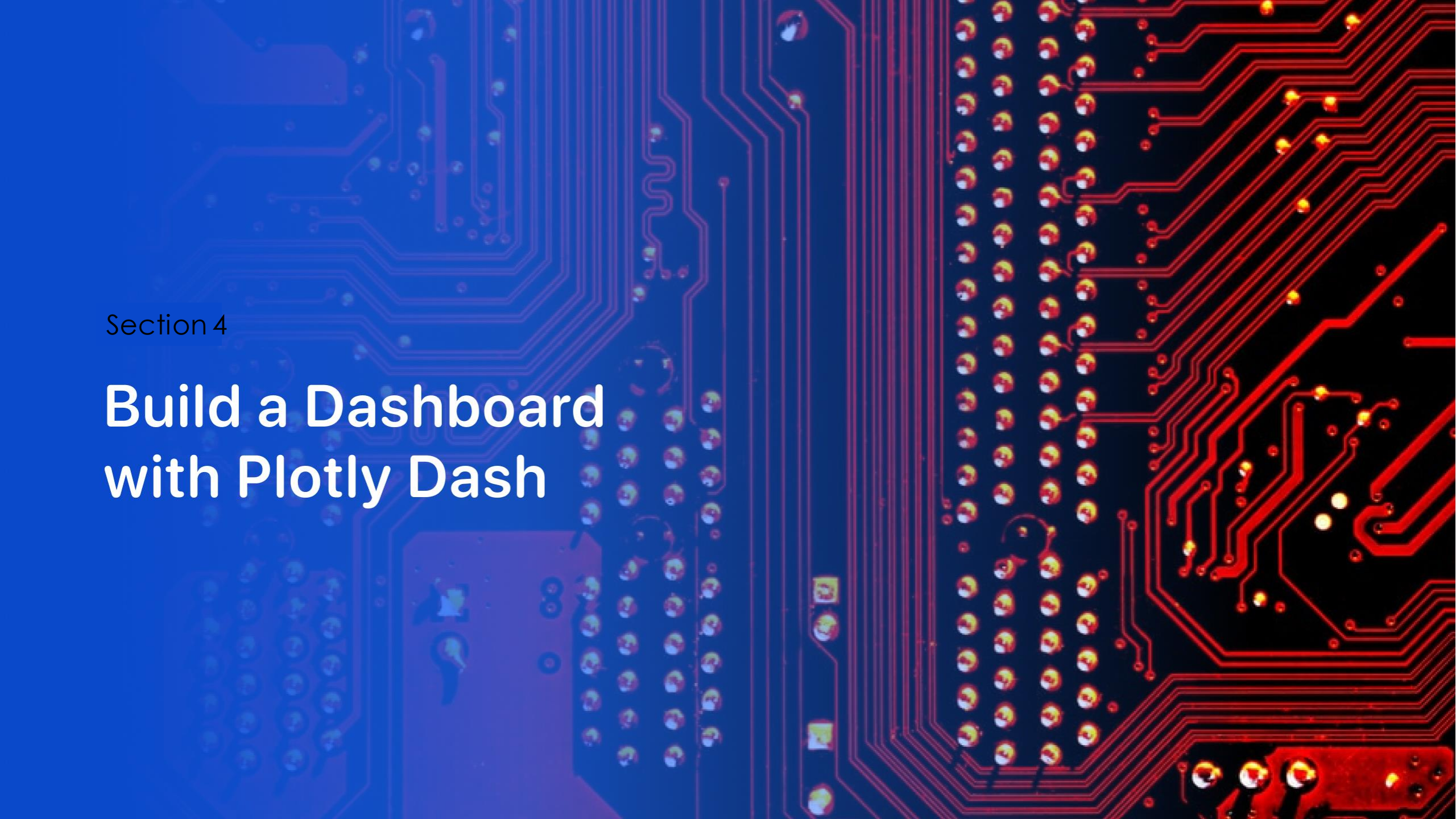




Pertinent Proximities

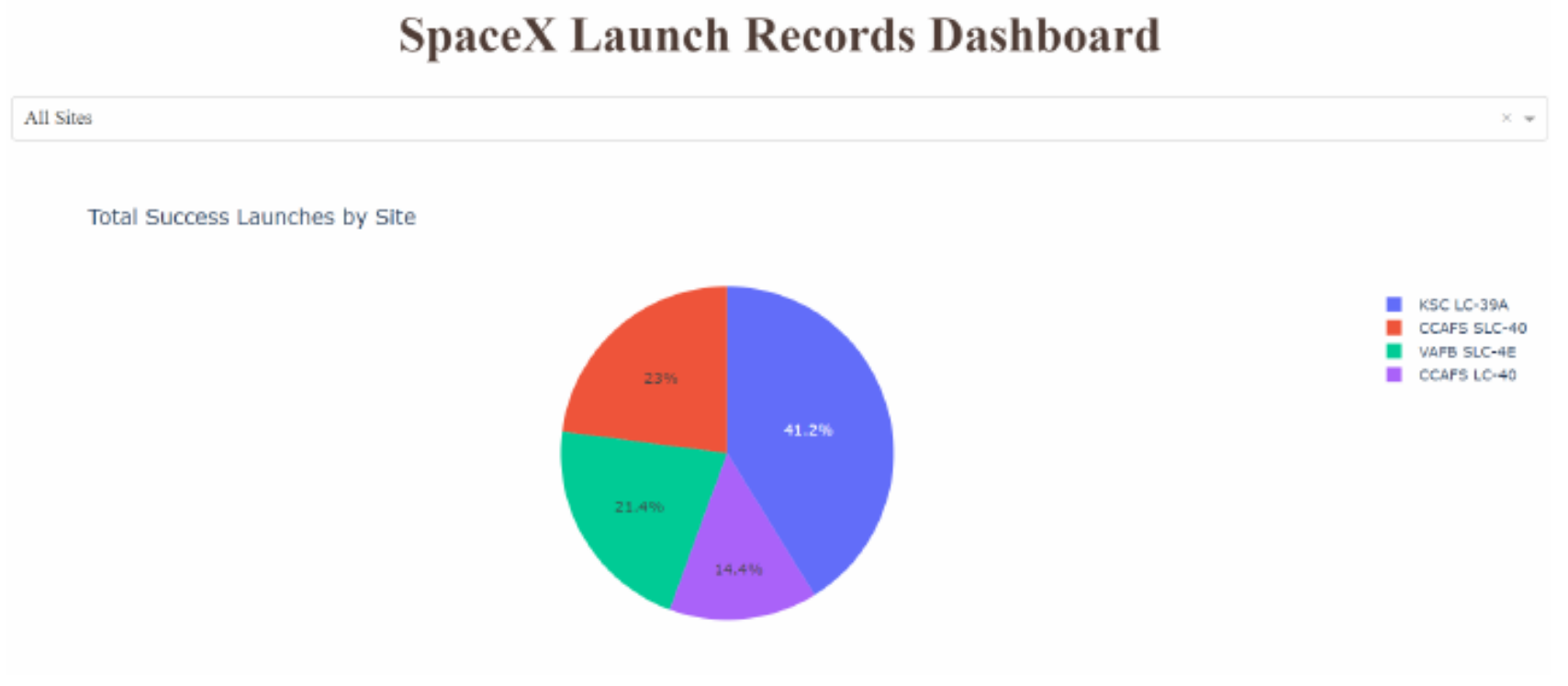
- Map showing proximities to closest railroad, highways, and city. Indicates that care has been taken to pick launch sites with a safe distance from occupied locations




Section 4
Build a Dashboard with Plotly Dash

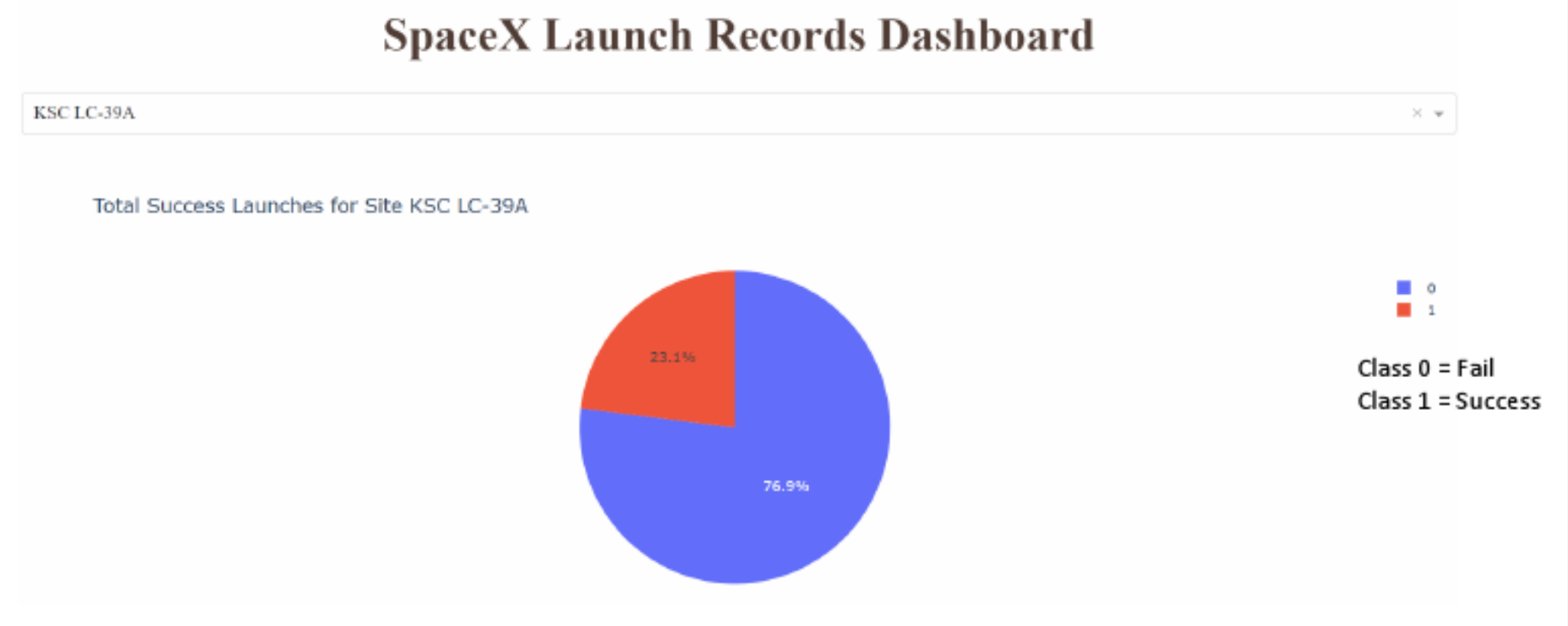
Successful Launches by Site

- KSC LC-39A HAS THE MOST SUCCESSFUL LAUNCHES AMONGST LAUNCH SITES (41.2%)



KSC LC-39A Successful Launches

- KSC LC-39A HAD THE HIGHEST SUCCESS RATE AMONG ALL LAUNCH SITES



Payload Vs Success for all Sites

- Scatter plot of Payload vs Launch Outcome for all sites

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- THE DECISION TREE CLASSIFIER IS THE MODEL WITH THE HIGHEST CLASSIFICATION ACCURACY

```
# Create a DF for algorithm type and respective best scores

Model_Performance_df = pd.DataFrame({'Algo Type': ['Logistic Regression', 'SVM','Decision Tree','KNN'],
'Accuracy Score': [logreg_cv.best_score_, svm_cv.best_score_, tree_cv.best_score_, knn_cv.best_score_],
'Test Data Accuracy Score': [logreg_cv.score(X_test, Y_test), svm_cv.score(X_test, Y_test),
tree_cv.score(X_test, Y_test), knn_cv.score(X_test, Y_test)]})
```

```
Model_Performance_df.sort_values(['Accuracy Score'], ascending = False, inplace=True)
```

```
Model_Performance_df
```

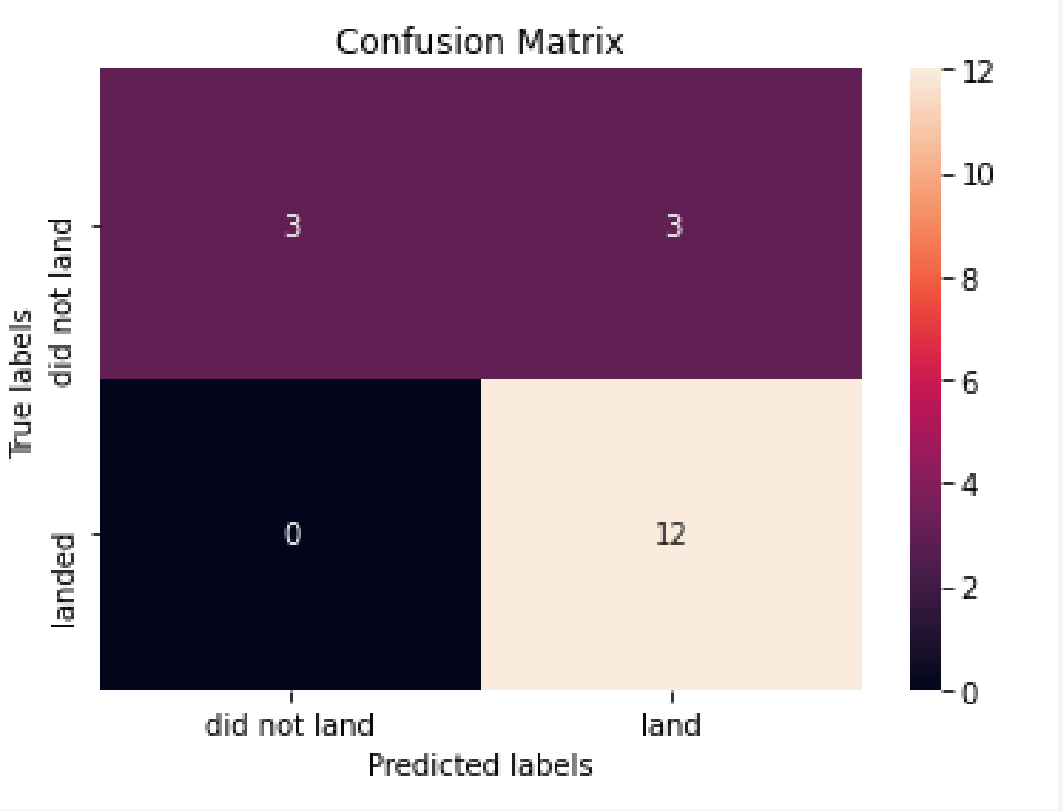
	Algo Type	Accuracy Score	Test Data Accuracy Score
2	Decision Tree	0.875000	0.833333
3	KNN	0.848214	0.833333
1	SVM	0.848214	0.833333
0	Logistic Regression	0.846429	0.833333

```
i = Model_Performance_df['Accuracy Score'].idxmax()
print('The best performing alogrithm is '+ Model_Performance_df['Algo Type'][i]
+ ' with score ' + str(Model_Performance_df['Accuracy Score'][i]))
```

```
The best performing alogrithm is Decision Tree with score 0.875
```

Confusion Matrix

- The confusion matrixes for all the models were the same. False positives were a problem in all models.



Conclusions

- We can conclude that:
- The larger the flight amount at a launch site, the greater the success rate at a
- launch site.

- Launch success rate started to increase in 2013 till 2020.

- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.

- KSC LC-39A had the most successful launches of any sites.

- The Decision tree classifier is the best machine learning algorithm for this task.


Thank You!