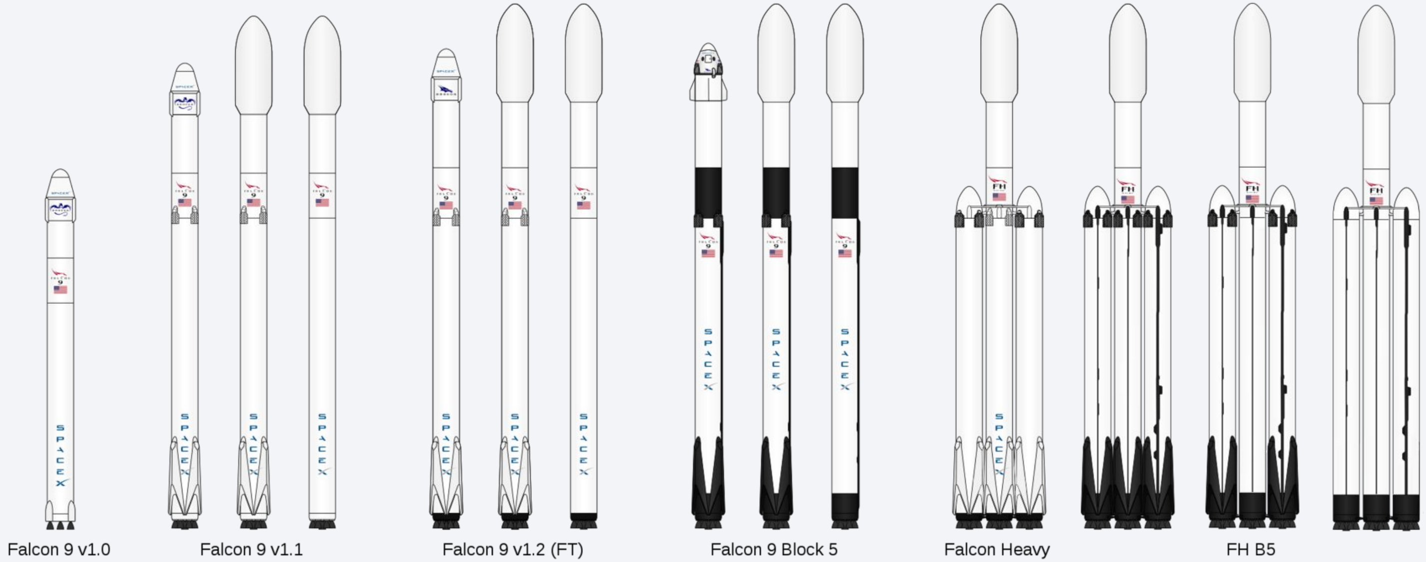
IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

David Nicewonder
June 24, 2024

Outline

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



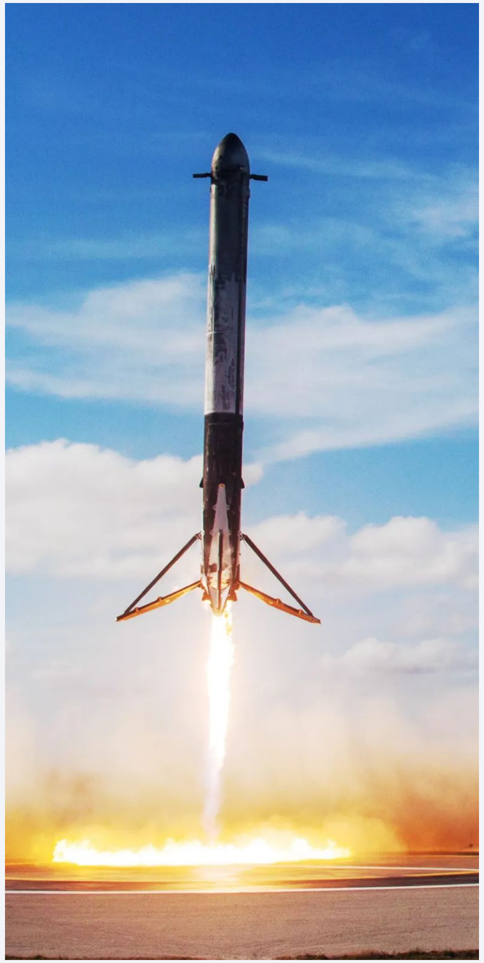
Executive Summary

In this capstone project, data was collected and modeled to predict if the Falcon 9 first stage will land successfully. The dataset includes a record for each payload carried during a SpaceX mission. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

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 successfully, it can assist with determining the cost of a launch.

Introduction



The project goal is to use data to predict whether SpaceX will attempt to land a rocket or not.

In this capstone project, SpaceX launch data was gathered from the SpaceX REST API.

The data included information about launches, the rocket used, payload delivered, launch specifications, landing specifications, and landing outcomes.

Exploratory Data Analysis (EDA) and data visualization to understand the data and to answer project questions.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX launch data was collected using the SpaceX REST API.
- Perform data wrangling:
  - The API data was converted from structured JSON and normalized to a dataframe flat table saved to a CSV file.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection

- SpaceX launch data was collected using the SpaceX REST API.
- The API data was converted from structured JSON and normalized to a Pandas dataframe, which was saved to a CSV file.
- Additional data was gathered by web scraping HTML tables from Wikipedia for SpaceX launch records information using BeautifulSoup. This data was combined with the dataframes.

Data Collection – SpaceX API

- Request rocket launch data from SpaceX API: https://api.spacexdata.com/v4/launches/past

- GitHub URL of the completed SpaceX API calls notebook: https://github.com/davidnicewonder/IBM-Applied-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Request and parse the SpaceX launch data using the GET request

Decode the response content as a Json using `.json()` and turn it into a Pandas dataframe using `.json_normalize()`

Filter the dataframe to only include Falcon 9 launches

Data Collection - Scraping

- Wikipedia: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

- GitHub URL of the completed web scraping notebook: https://github.com/davidnicewonder/IBM-Applied-Data-Science-Capstone/blob/main/jupyter-labs-webscraping.ipynb

Request the Falcon9 Launch Wiki page from its URL

Create a BeautifulSoup object from the HTML response

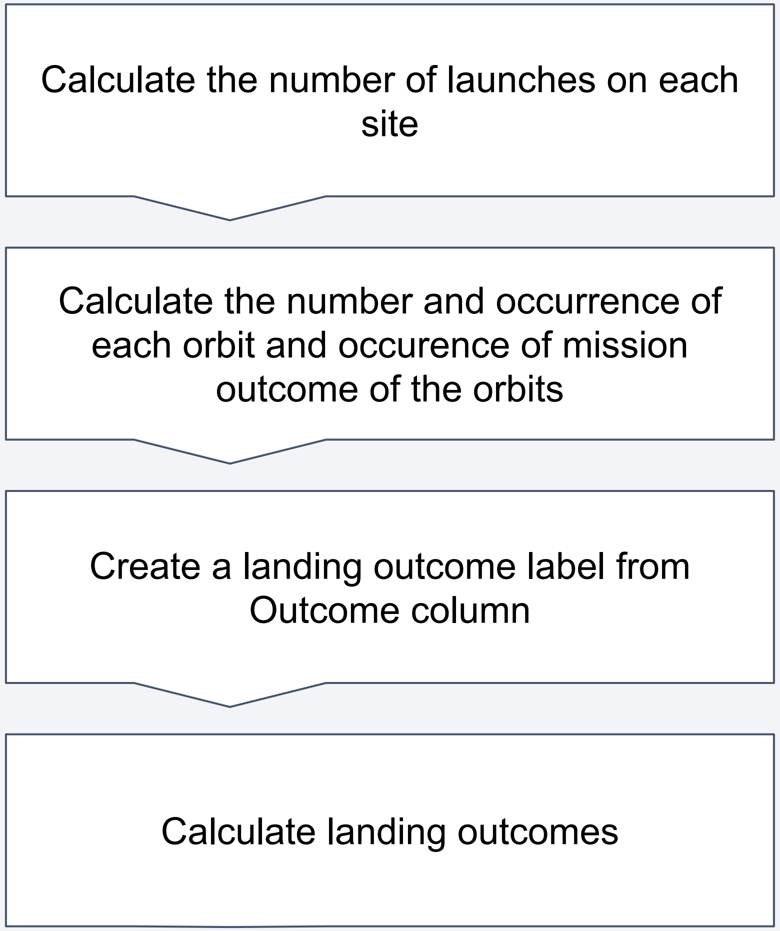
Extract all column/variable names from the HTML table header

Create a data frame by parsing the launch HTML tables

Data Wrangling

- Performed Exploratory Data Analysis (EDA) to find patterns in the data and determine what would be the label for training supervised models.
- Calculated the number of landing outcomes using value_counts() on the column Outcome to determine the number of landing outcomes.

- GitHub URL of your completed data wrangling related notebook: https://github.com/davidnicewonder/IBM-Applied-Data-Science-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- Summary of charts (launch outcomes):
  - FlightNumber vs. PayloadMass
  - FlightNumber vs. LaunchSite
  - Payload vs. Launch Site
  - Orbit Success Rate
  - FlightNumber and Orbit Type
  - Payload vs. Orbit
  - Maps
- GitHub URL of EDA with data visualization notebook: https://github.com/davidnicewonder/IBM-Applied-Data-Science-Capstone/blob/main/edadataviz.ipynb

EDA with SQL

- Summary of SQL queries:
    - Display the names of the unique launch sites in the space mission.
    - 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 names of the boosters 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_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
    - 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.

- GitHub URL of EDA with SQL notebook: https://github.com/davidnicewonder/IBM-Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Summary of maps and objects:
  - Mark all launch sites on a map.
  - Mark the success/failed launches for each site on the map.
  - Calculate the distances between a launch site to its proximities
- Map objects such as markers, circles, and lines were added to the Folium maps to better visualize and orient for real-world locations of launch sites and landing success rates.
- GitHub URL of interactive map with Folium map:
  https://github.com/davidnicewonder/IBM-Applied-Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Summary of Dashboard plots/graphs and interactions:
  - Pie chart
  - Scatter plot
  - Launch site dropdown selector
  - Payload mass slider selector
- The plots and interactions allow decision makers to interact directly with the data to update the plots and charts. This information can help answer their questions in real-time.
- GitHub URL of Plotly Dash lab: https://github.com/davidnicewonder/IBM-Applied-Data-Science-Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model:
  - Four different models were used: Decision Tree, Logistic Regression, Support Vector Machine, and K Nearest Neighbors.
  - Each model used a GridSearchCV cross validation to find the best fit from a set of hyperparameters.
  - Accuracy score and confusion matrix were used to evaluate and compare results.

- GitHub URL of predictive analysis lab: https://github.com/davidnicewonder/IBM-Applied-Data-Science-Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

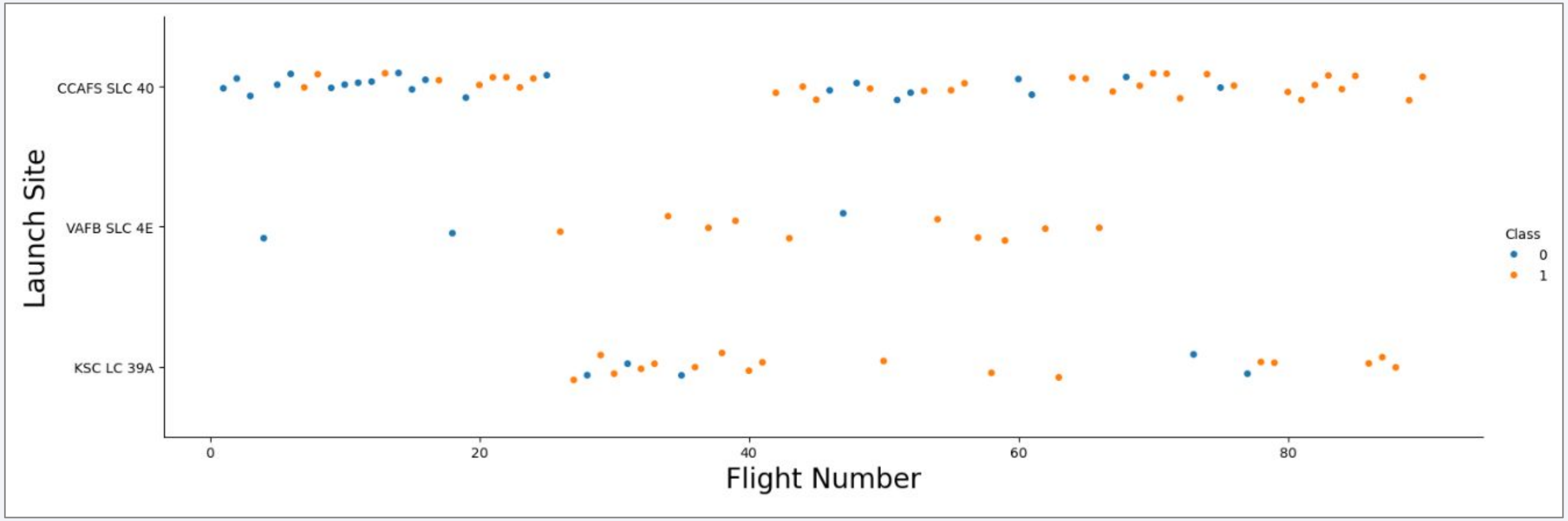
- Exploratory data analysis results:
  - Different launch sites have different success rates. CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.
- Interactive analytics demo in screenshots
  - See following slides.
- Predictive analysis results:
  - Although all models had an accuracy near 85%, the Decision Tree performed slightly better.

Section 2

Insights drawn from EDA

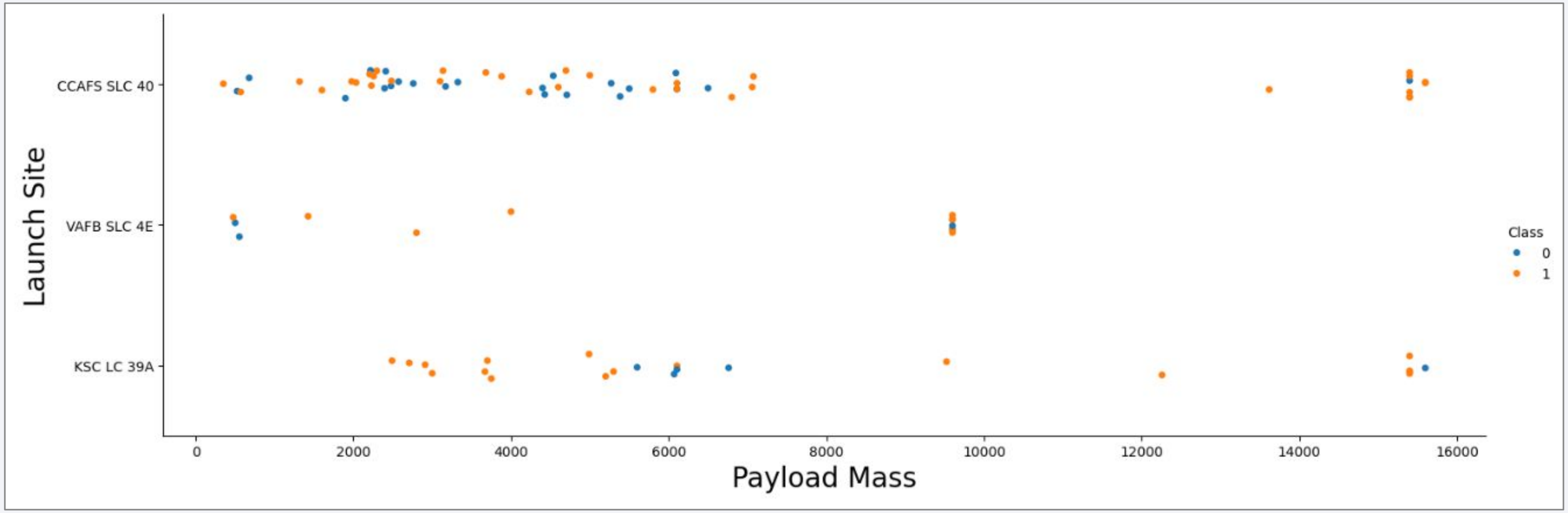
Flight Number vs. Launch Site

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%.



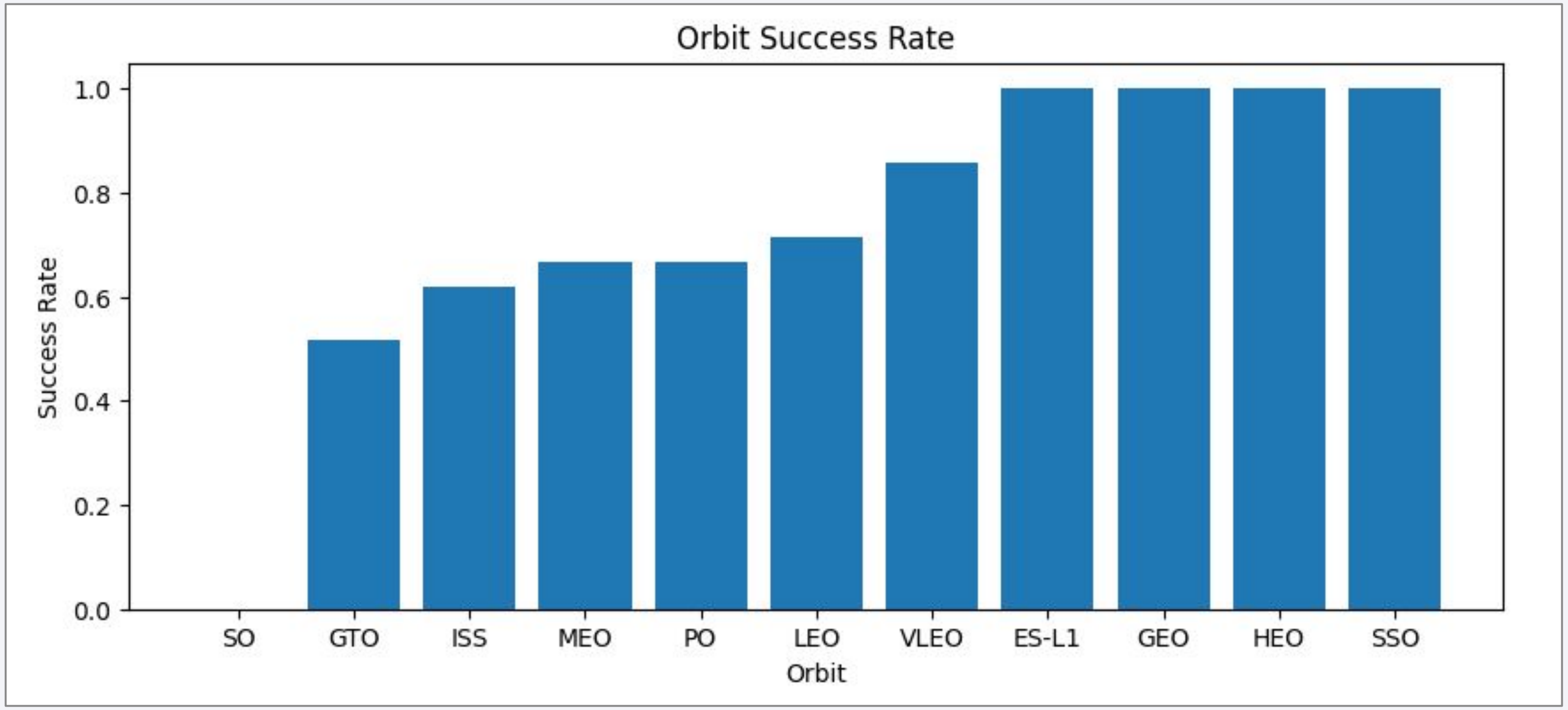
Payload vs. Launch Site

The more massive the payload, the less likely the first stage will return. For the VAFB-SLC launchsite there are no rockets launched for payload mass greater than 10000.



Success Rate vs. Orbit Type

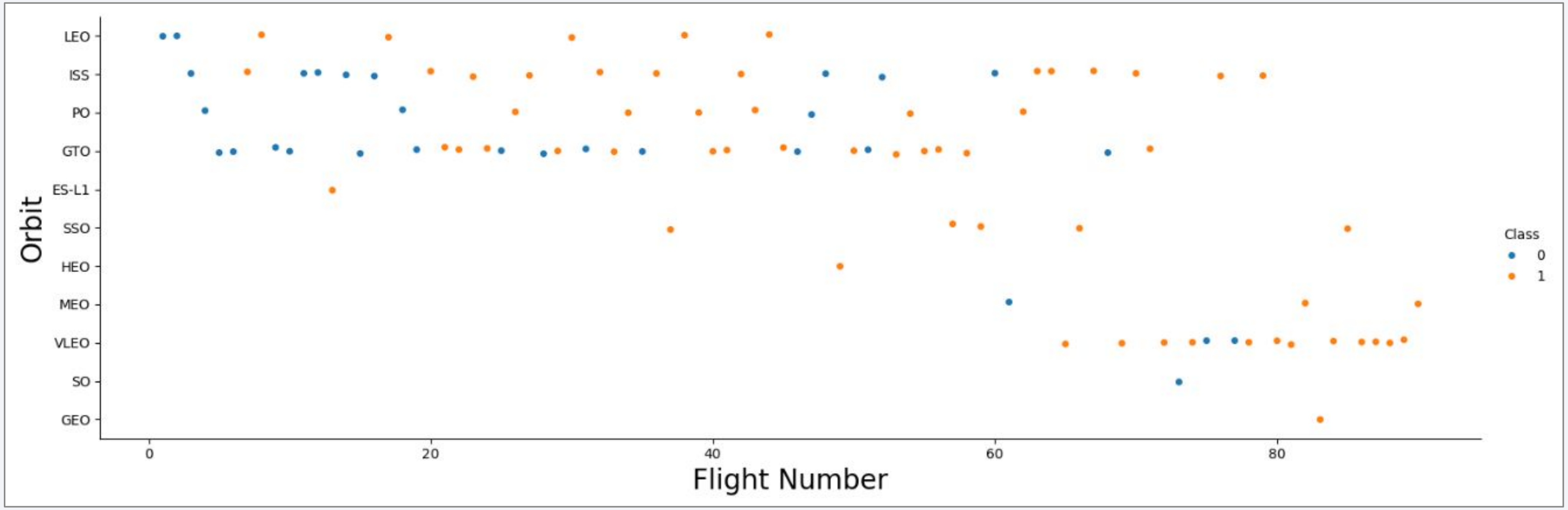
The orbits ES-L1, GEO, HEO, and SSO has the highest success rate.



Orbit	Success Rate
SO	0.000000
GTO	0.518519
ISS	0.619048
MEO	0.666667
PO	0.666667
LEO	0.714286
VLEO	0.857143
ES-L1	1.000000
GEO	1.000000
HEO	1.000000
SSO	1.000000

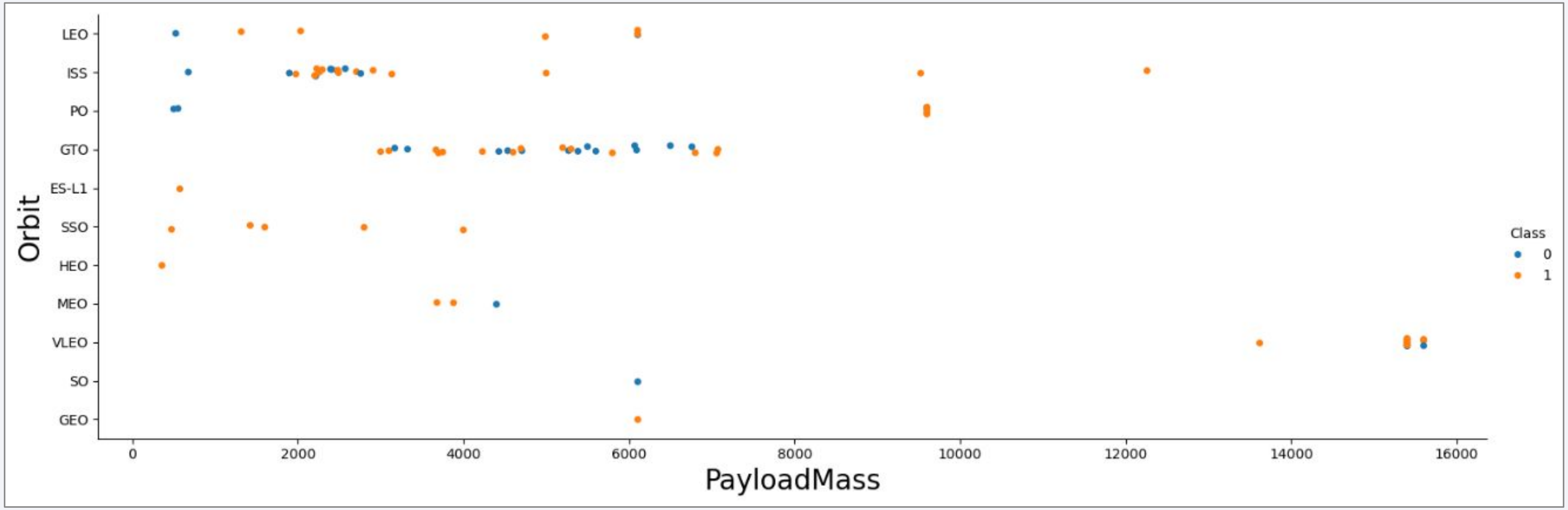
Flight Number vs. Orbit Type

The LEO orbit the Success appears related to the number of flights. There seems to be no relationship between flight number when in GTO orbit.



Payload vs. Orbit Type

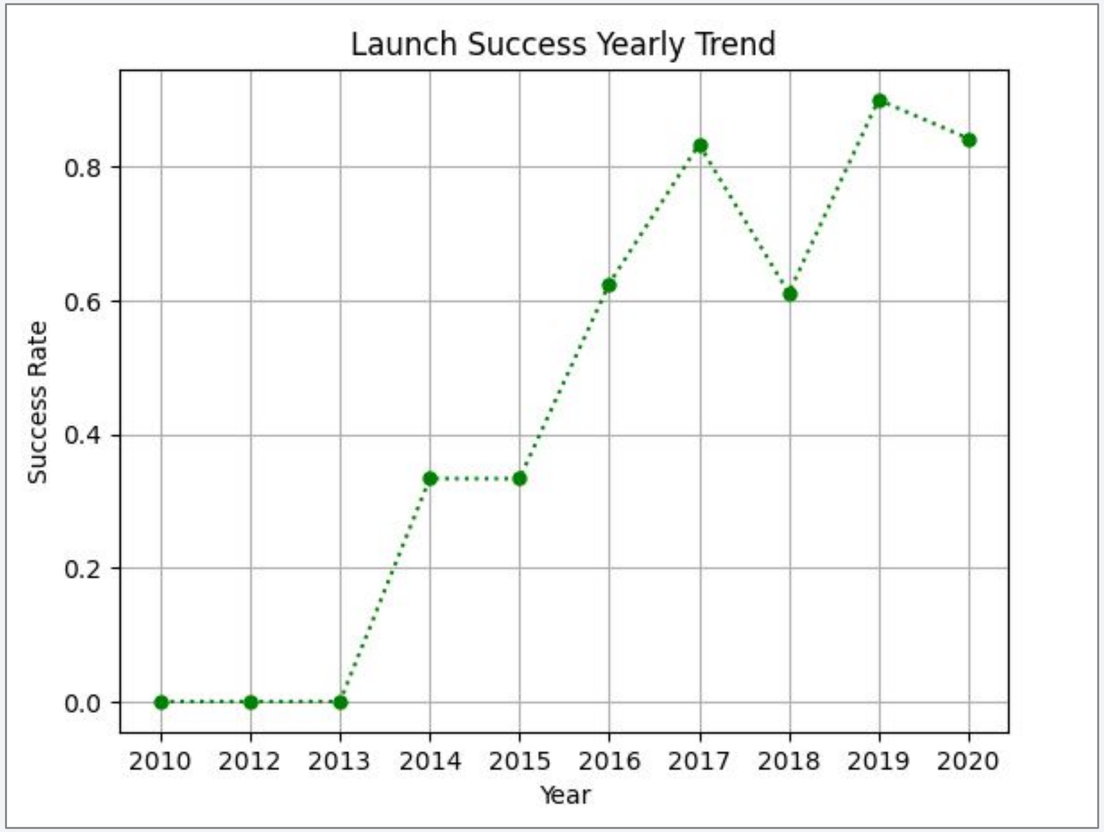
With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.



Launch Success Yearly Trend

The success rate since 2013 kept increasing till 2020.

The largest increase in trend was the four years between 2013 and 2017.



All Launch Site Names

SpaceX launch facilities:

- Cape Canaveral Space Launch Complex 40 (CCAFS SLC-40, CCAFS LC-40)
- Vandenberg Space Force Base Space Launch Complex 4E (VAFB SLC-4E)
- Kennedy Space Center Launch Complex 39A (KSC LC-39A)

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

```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE
```

```
 * 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'

Find 5 records where launch sites begin with `CCA`

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

```
%sql SELECT * FROM SPACEXTABLE 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
2010-06-04	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.

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

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)'
```

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

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

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

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1'
```

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

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

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

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

Hint:Use min function

```
%sql SELECT MIN(Date) FROM SPACEXTABLE 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

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

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, PAYLOAD_MASS__KG_ FROM SPACEXTABLE
WHERE Landing_Outcome = 'Success (drone ship)'
AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

Total Number of Successful and Failure Mission Outcomes

Calculate the total number of successful and failure mission outcomes.

List the total number of successful and failure mission outcomes

```
%%sql
SELECT * FROM
(SELECT COUNT(Mission_Outcome) AS Mission_Success
FROM SPACEXTABLE
WHERE Mission_Outcome LIKE '%Success%'),
(SELECT COUNT(Mission_Outcome) AS Mission_Failure
FROM SPACEXTABLE
WHERE Mission_Outcome NOT LIKE '%Success%')
```

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

Mission_Success	Mission_Failure
100	1

Boosters Carried Maximum Payload

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

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

```
%%sql
SELECT * FROM
(SELECT Booster_Version, MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
```

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

Booster_Version	MAX(PAYLOAD_MASS__KG_)
F9 B5 B1048.4	15600

2015 Launch Records

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

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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
%%sql
SELECT substr(Date, 6,2) AS Month, Landing_Outcome, Booster_Version, Launch_Site
FROM SPACEXTABLE
WHERE substr(Date,0,5)='2015'
AND Landing_Outcome = 'Failure (drone ship)'
```

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

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.

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(Landing_Outcome) AS Count
FROM SPACEXTABLE
GROUP BY Landing_Outcome
HAVING Date >= '2010-06-04' AND Date <= '2017-03-20'
ORDER BY Count DESC
```

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

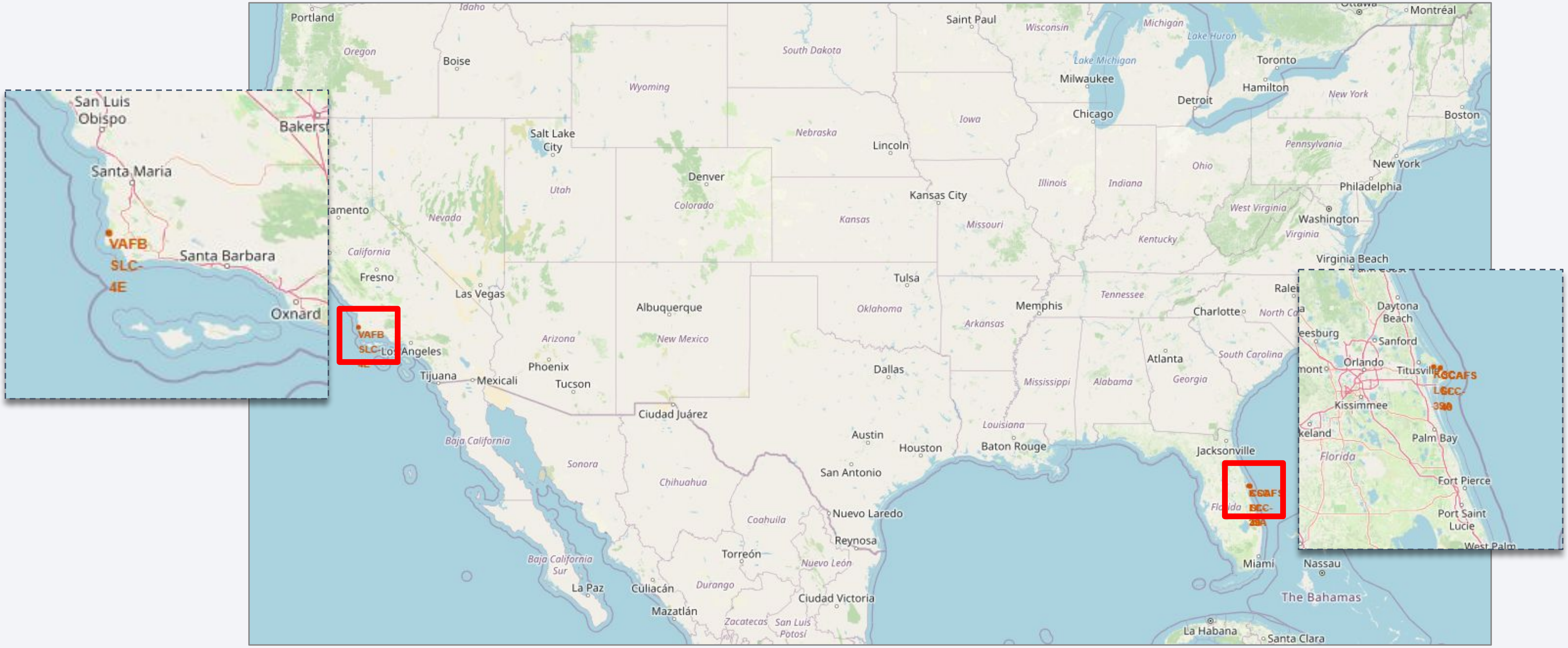
Landing_Outcome	Count
No attempt	21
Success (drone ship)	14
Success (ground pad)	9
Failure (drone ship)	5
Controlled (ocean)	5
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

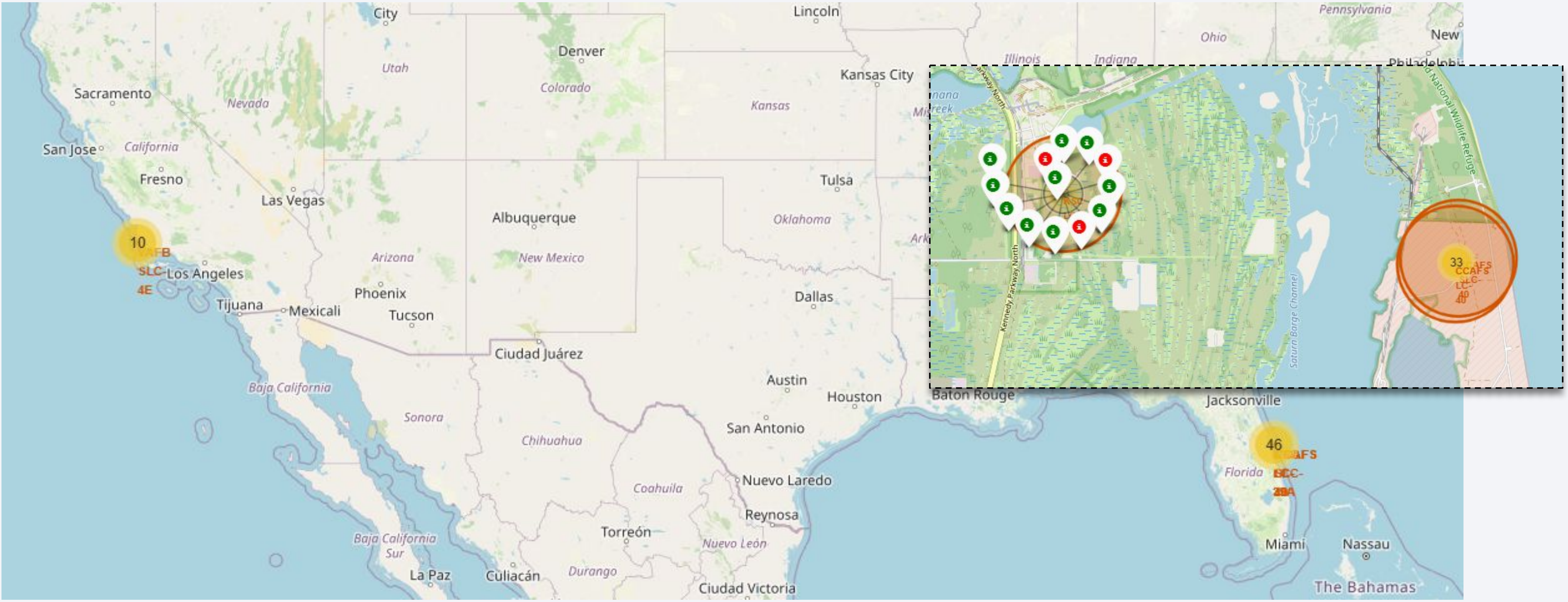
SpaceX Launch Sites

SpaceX launch sites are located on the East and West coasts in the southern United States, Florida and California (nearer to the equator).



Launch Outcomes at Launch Sites

Launch outcome markers at launch sites (green=success, red=failure).



Launch Site Proximities

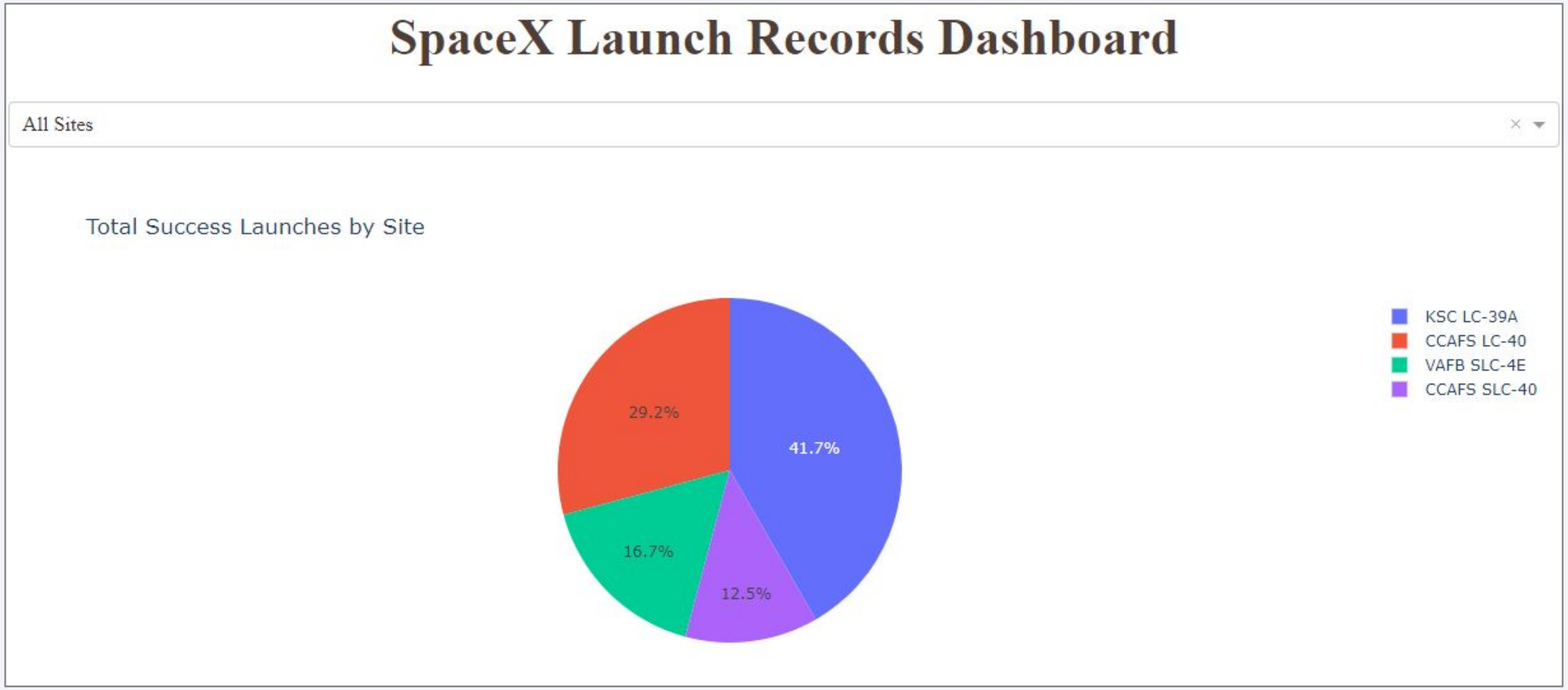
Exploring the generated map shows the CCAFS LC 40 launch site is almost one kilometer (about a half mile) from the Atlantic coast.

Section 4

Build a Dashboard with Plotly Dash

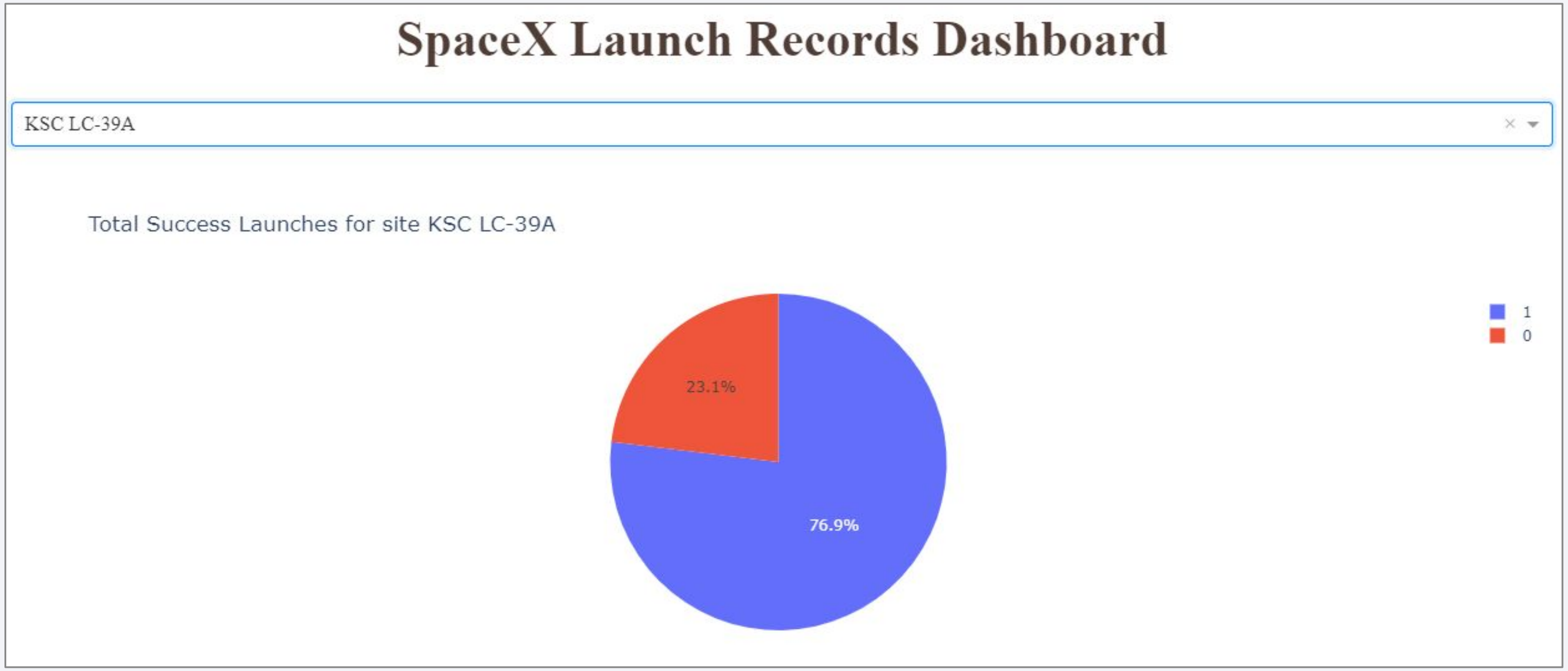
Total Success Launches for All Sites

KSC LC-39A has the highest success rate and CCAFS SLC-40 the lowest.



Launch Site KSC LC-39A

Site KSC LC-39A has the highest success rate of 76.9%.



Payload vs. Launch Outcome for All Sites

Payload vs. Launch Outcome scatter plot for all sites, with payload selected for 4000 and higher.

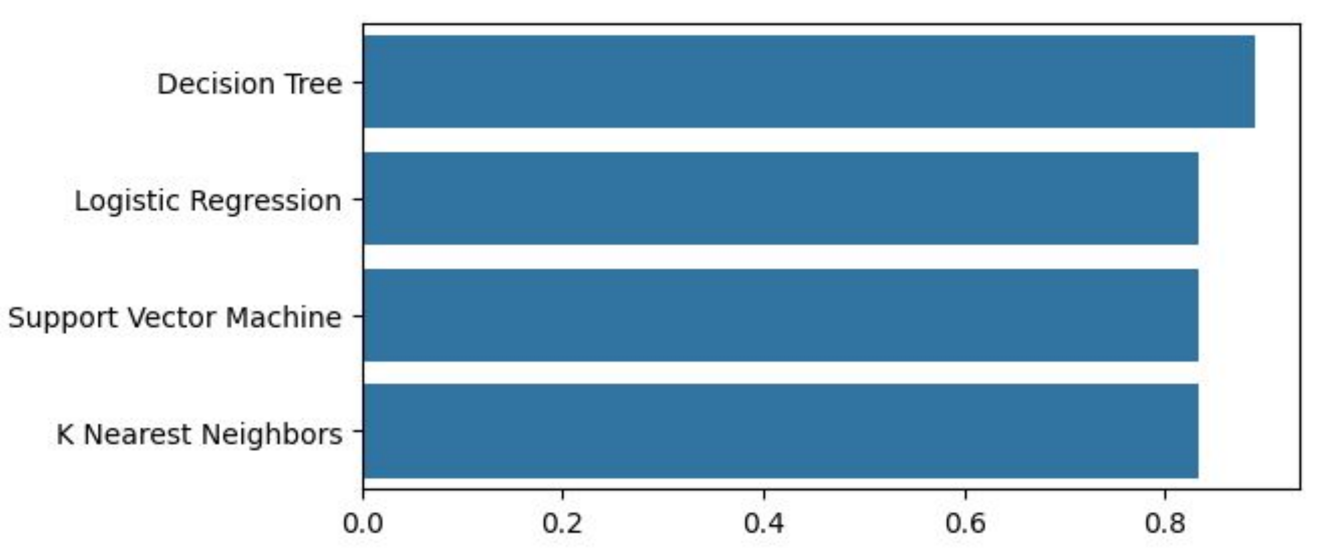
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Overall the models had nearly the same accuracy, with the Decision Tree model performing slightly better.

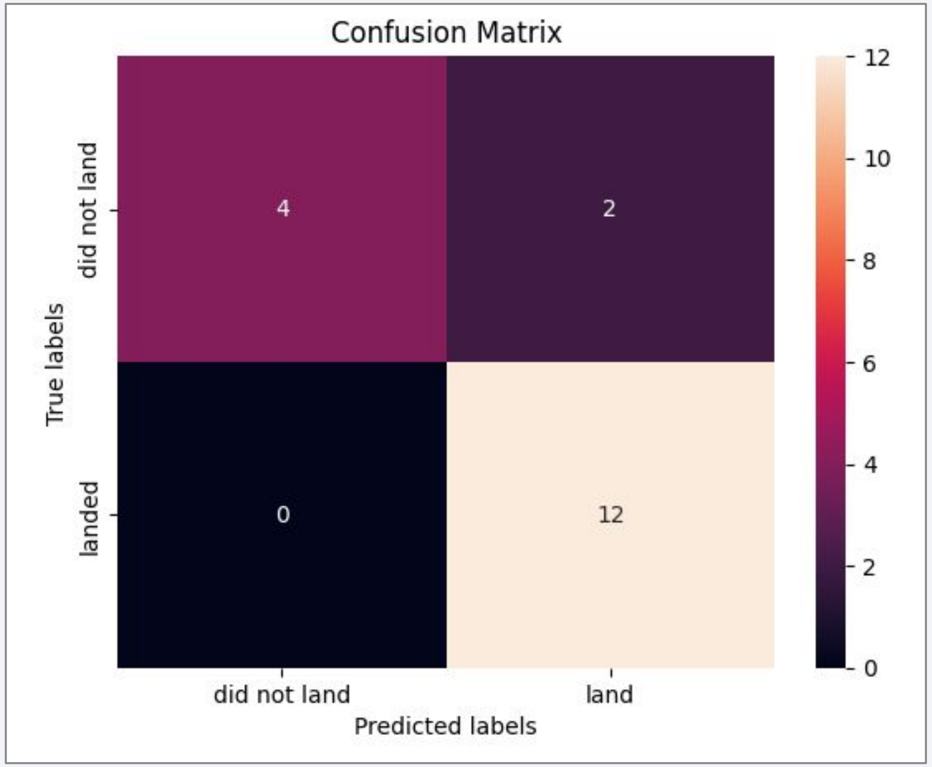
Model	Score	Confusion Matrix
Decision Tree	**0.889**	[4 , 2] [0 , 12]
Logistic Regression	0.833	[3 , 3] [0 , 12]
Support Vector Machine	0.833	[3 , 3] [0 , 12]
K Nearest Neighbors	0.833	[3 , 3] [0 , 12]



Confusion Matrix for Decision Tree

Overall the models had nearly the same accuracy, with the Decision Tree model performing slightly better.

The Decision Tree models benefits from being highly interpretable and its features easy to understand by decision makers.



Conclusions

- SpaceX has gained worldwide attention for a series of historic milestones. It is the only private company ever to return a spacecraft from low-earth orbit.
- SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars wheras other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage.
- If it can be determine if the first stage will land successfully, we can determine the cost of a launch.
- The dataset includes a record for each payload carried during a SpaceX mission into outer space. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

Appendix

Github project repository:
https://github.com/davidnicewonder/IBM-Applied-Data-Science-Capstone

Thank you!