IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Thomas Peeden
02/22/2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection – SpaceX API
  - Data Collection – Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis (EDA) using Visualization and SQL
  - Interactive Visual Analytics using Folium and Plotly Dash
  - Predictive Analysis using Classification Models
- Summary of all results
  - Exploratory Data Analysis Results
  - Interactive Analytics screenshots
  - Predictive Analysis results

Introduction

- Project background and context
  - 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.
- Problems you want to find answers
  - What factors influence landing outcome?
  - Where are the location of launch sites?
  - What factors when combined ensure the best outcome?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected by using SpaceX REST API and web scraping HTML tables from Wikipedia
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- Collected data using the following datasets:
  - SpaceX API (https://api.spacexdata.com/v4/)
    - Columns: rocket, launchpad, payloads, and cores
  - Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)

Data Collection – SpaceX API

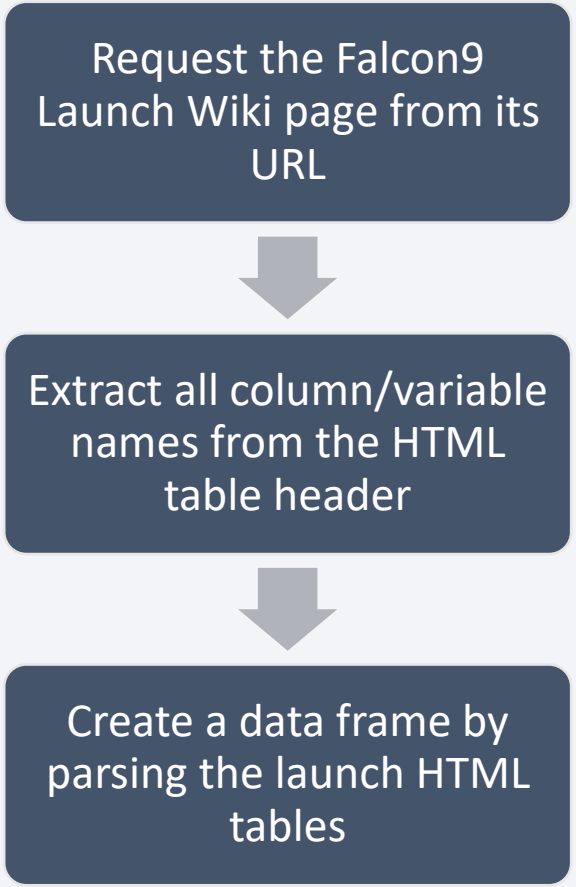
- GET request to collect SpaceX API data
- Decoded the response content as a Json using .json() and turned it into a Pandas dataframe using .json_normalize().
- Filled missing data of Payload Mass using .mean() and the .replace() function to replace np.nan values.

- Source Code: https://github.com/TBPeeden/IBM-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api(1).ipynb

Data Collection - Scraping

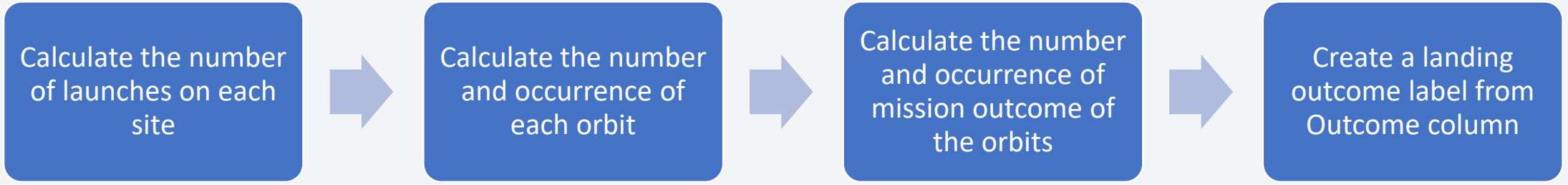
- Used Python BeautifulSoup package to web scrape HTML tables in Wikipedia that contain Falcon 9 launch records.
- Parsed the data from those tables and converted them into a Pandas data frame for further visualization and analysis.

- Source code: https://github.com/TBPeeden/IBM-Data-Science-Capstone/blob/main/jupyter-labs-webscraping(1).ipynb



Data Wrangling

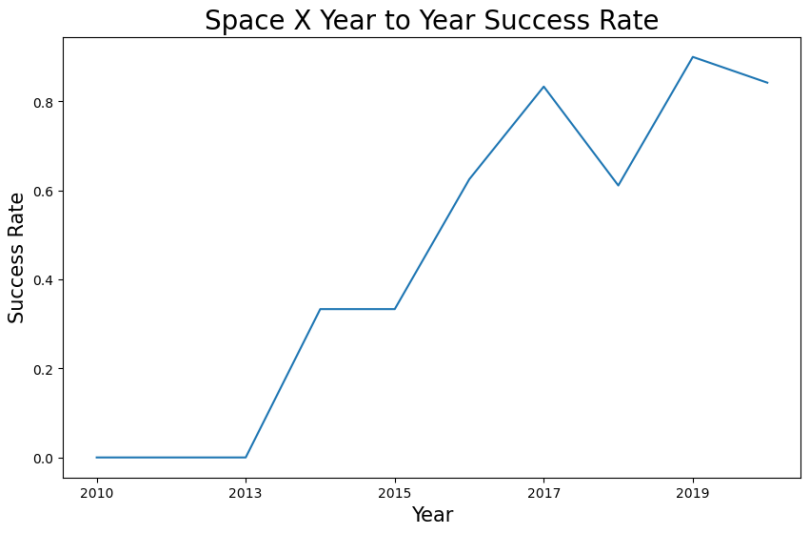
- Use the method value_counts() to determine the number of launches on each site, the number and occurrence of each orbit, and to determine the number of landing_outcomes.
- Using the Outcome, create a list where the element is zero if the corresponding row in Outcome is in the set bad_outcome; otherwise, it's one. Then assign it to the variable landing_class

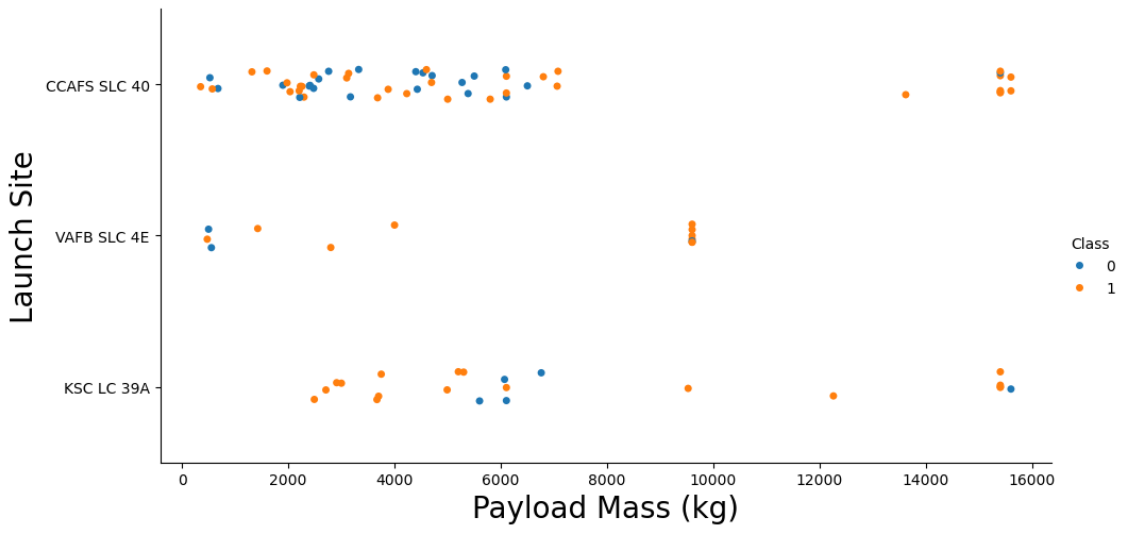


Source code:
https://github.com/TBPeeden/IBM-Data-Science-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling(1).ipynb

EDA with Data Visualization

- Space X overall success rate has trended upward from 2013 to 2020.
- A large payload mass has coincided with more success and most of the launches have been at the CCAFS SLC 40.
- Source code: https://github.com/TBPeeden/IBM-Data-Science-Capstone/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb





EDA with SQL

- SQL queries performed
  - 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.

- Source Code: https://github.com/TBPeeden/IBM-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite(1).ipynb

Build an Interactive Map with Folium

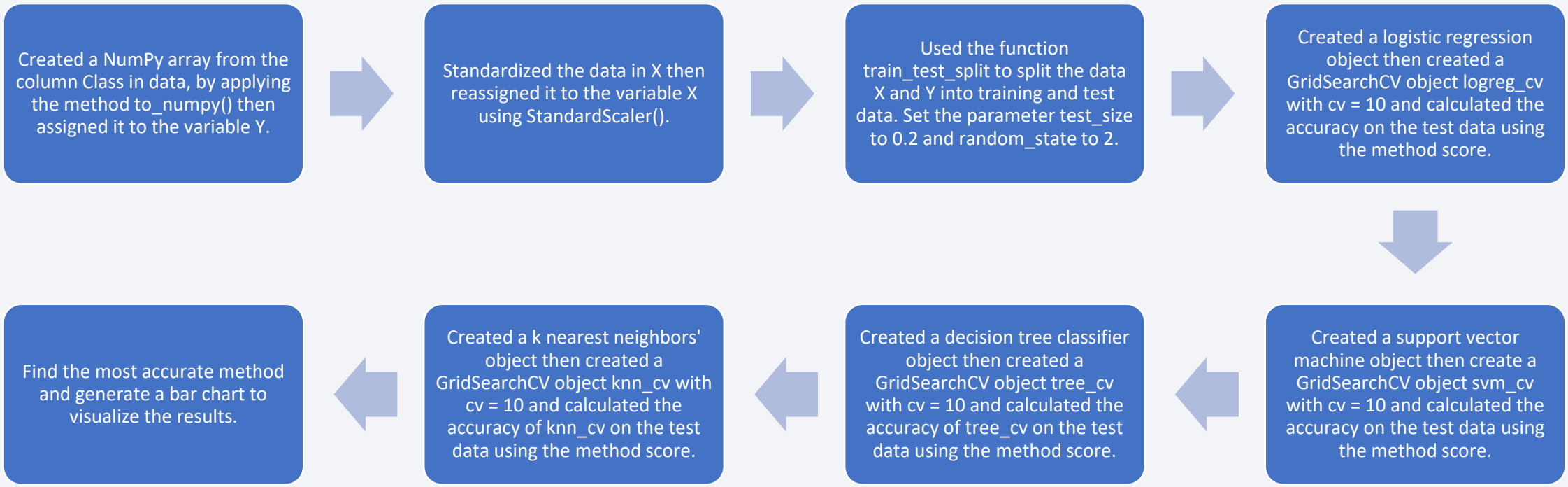
- Marked all launch sites to visualize where the launch sites were located.
- Marked the success/failed launches for each site on the map using markercluster() to simplify the map that contains many markers with the same coordinates.
- Calculated the distances between sites to nearest coastline, city, railway, parkway and airport. All sites are in proximity of these. The following are the distances for CLAFS LC-40
  - Coastline: 0.86 km
  - City: 18.05 km
  - Railway: 1.01 km
  - Parkway: 0.60 km
  - Airport: 10.77 km

- Source code: https://github.com/TBPeeden/IBM-Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location.jupyterlite%20-%20folium.ipynb

Build a Dashboard with Plotly Dash

- Built a dashboard using Plotly Dash by adding a launch site drop-down input component, a callback function to render a success by launch site pie chart. Also added a range slider to select payload and a callback function to render a success by payload scatter plot chart.
- Used these charts to visualize which site has the largest successful launches, which site has the highest launch success rate, which payload range has the highest or lowest launch success rate, and which F9 booster version had the highest launch success rate.
- Source Code: https://github.com/TBPeeden/IBM-Data-Science-Capstone/blob/main/spacex_dash_app(1).py

Predictive Analysis (Classification)



- Source Code: https://github.com/TBPeeden/IBM-Data-Science-Capstone/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite(1).ipynb

Results

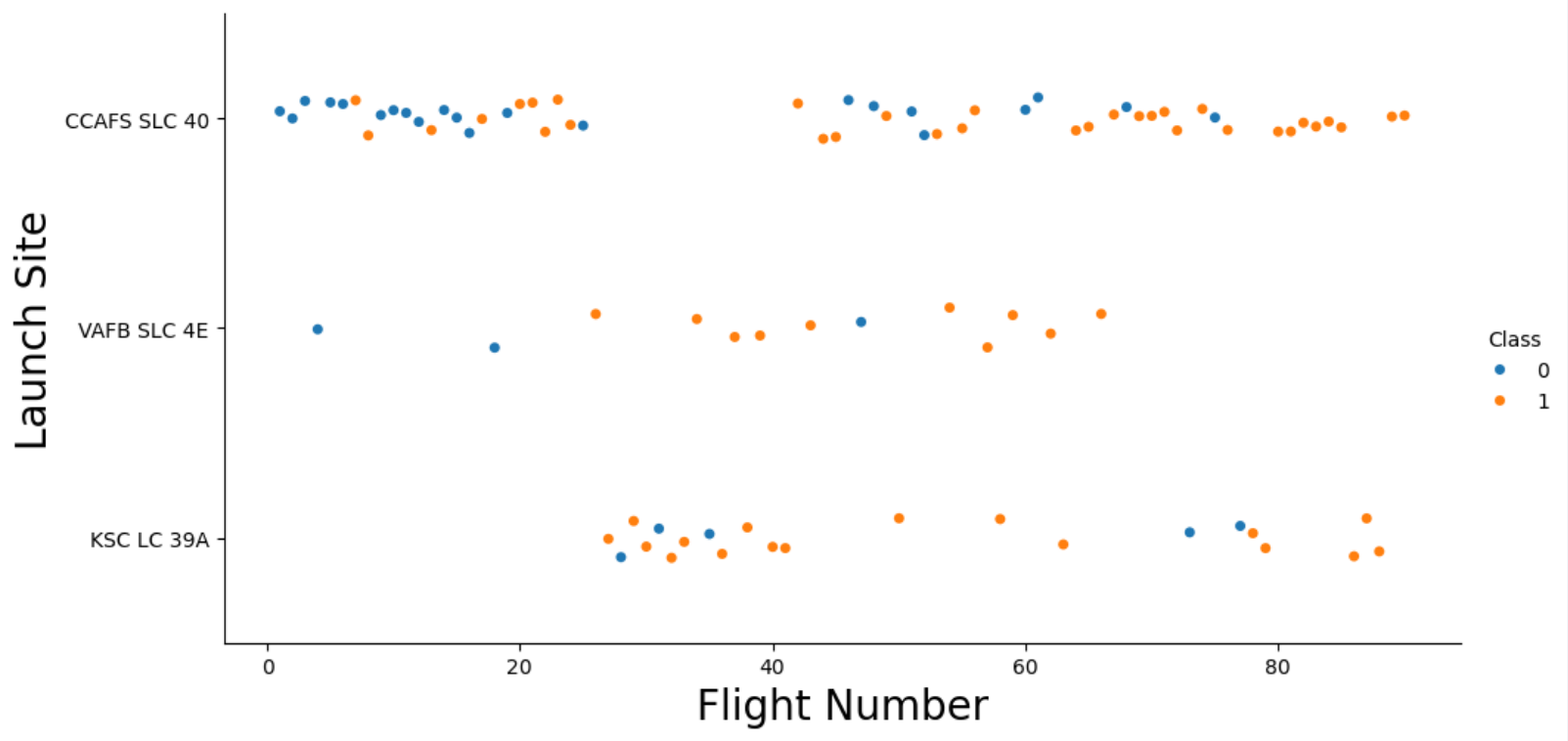
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights Drawn from EDA

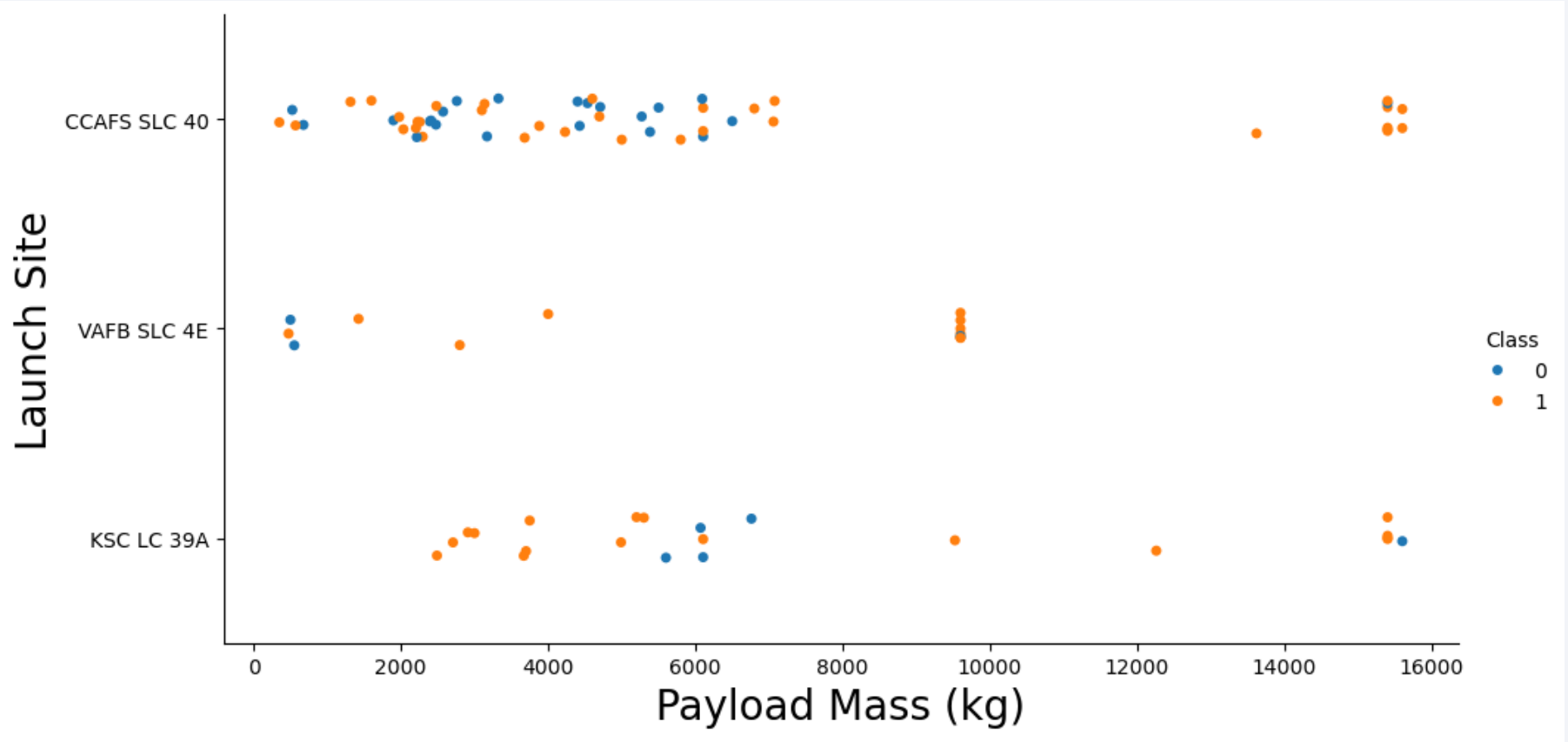
Flight Number vs. Launch Site

- Most flights are out of launch site CCAFS SLC 40 and shows a greater success rate with each new flight.



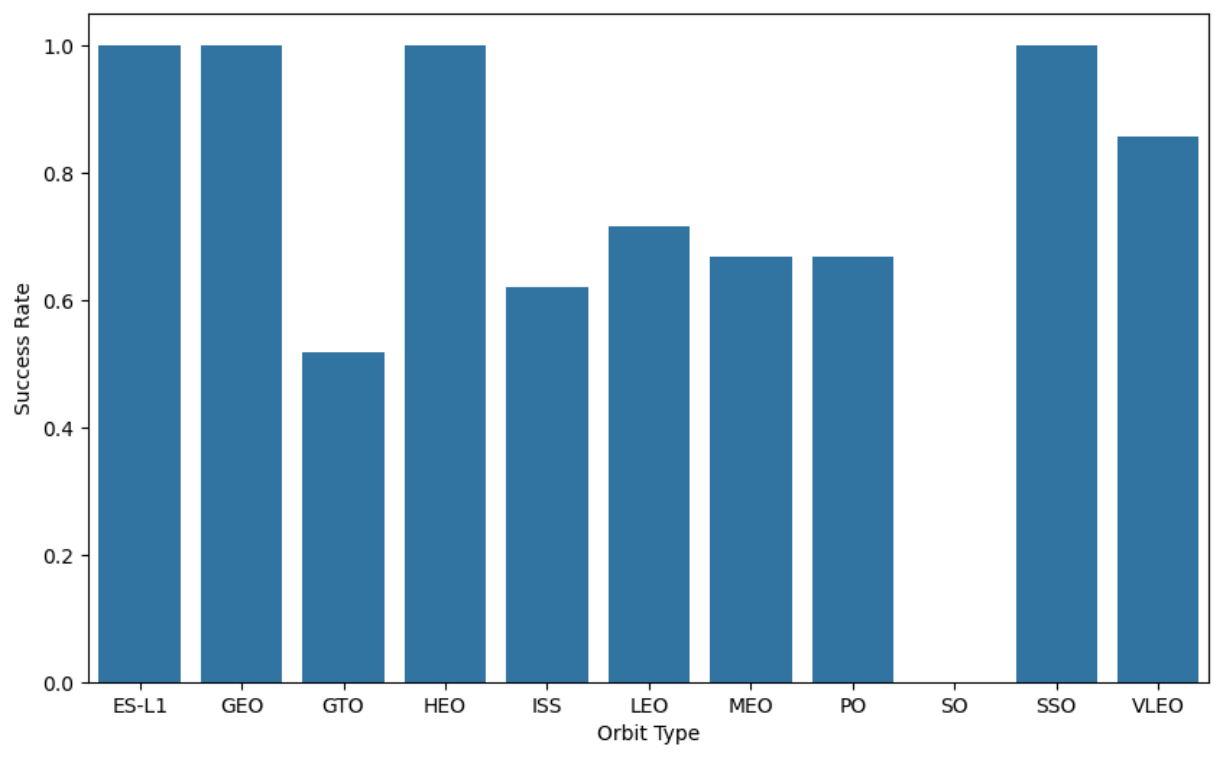
Payload vs. Launch Site

- Success rate increases with a payload mass of more than 7000 kg.



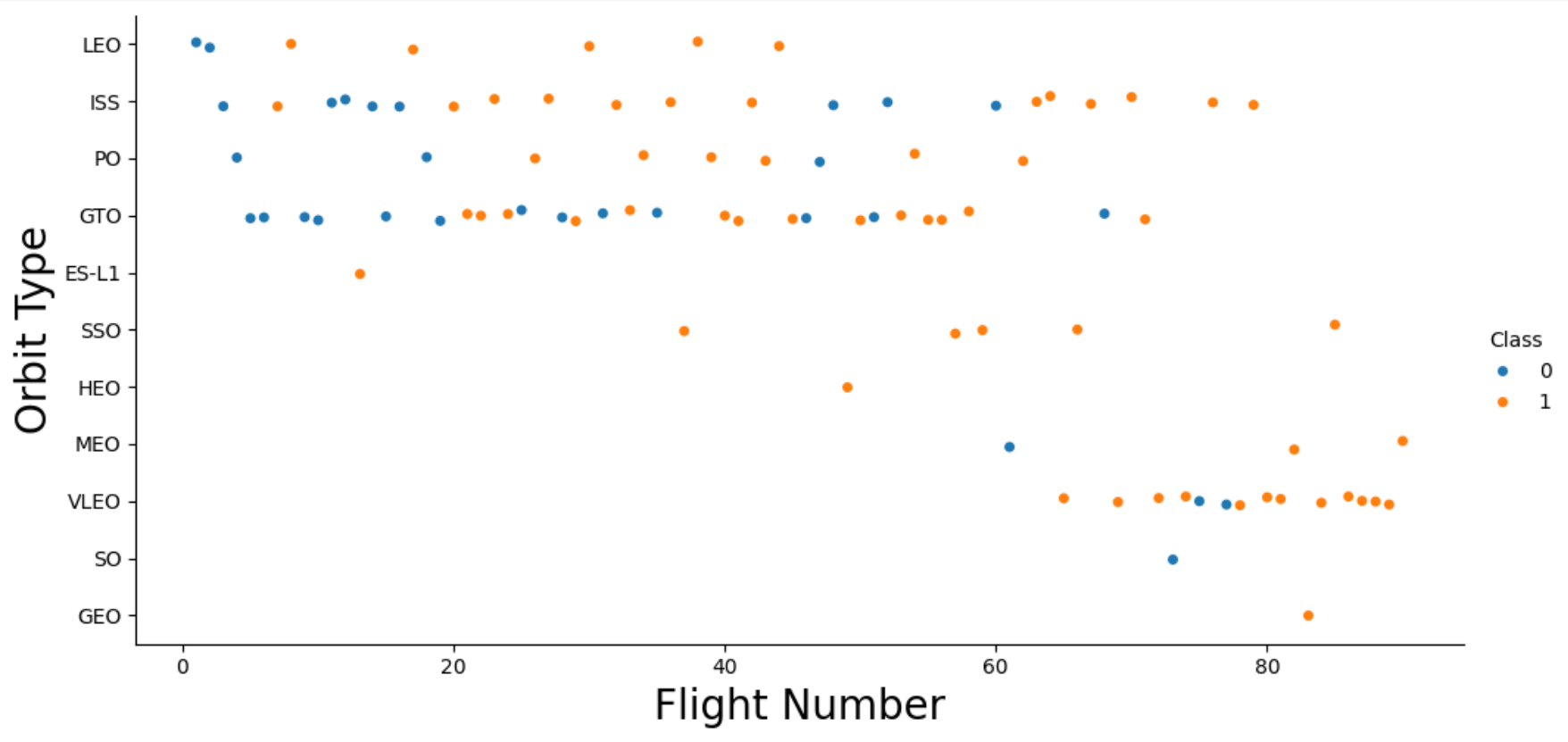
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, and SSO all had 100% success rate.
- ES-L1, GEO, and HEO only had one flight so would need more data to conclude they would consistently be more successful than the others.



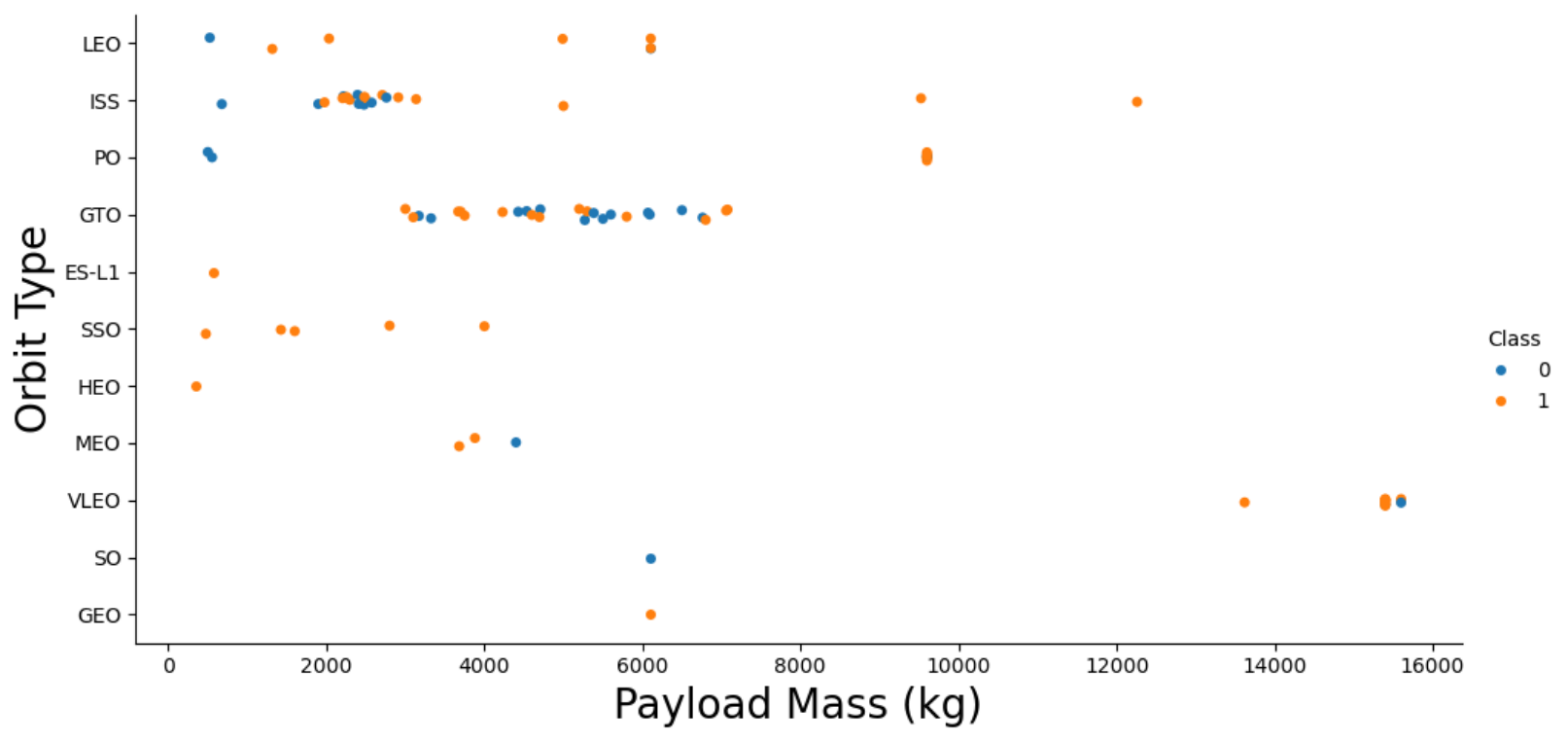
Flight Number vs. Orbit Type

- Success rate increased with each additional flight in each orbit except for GTO orbit type



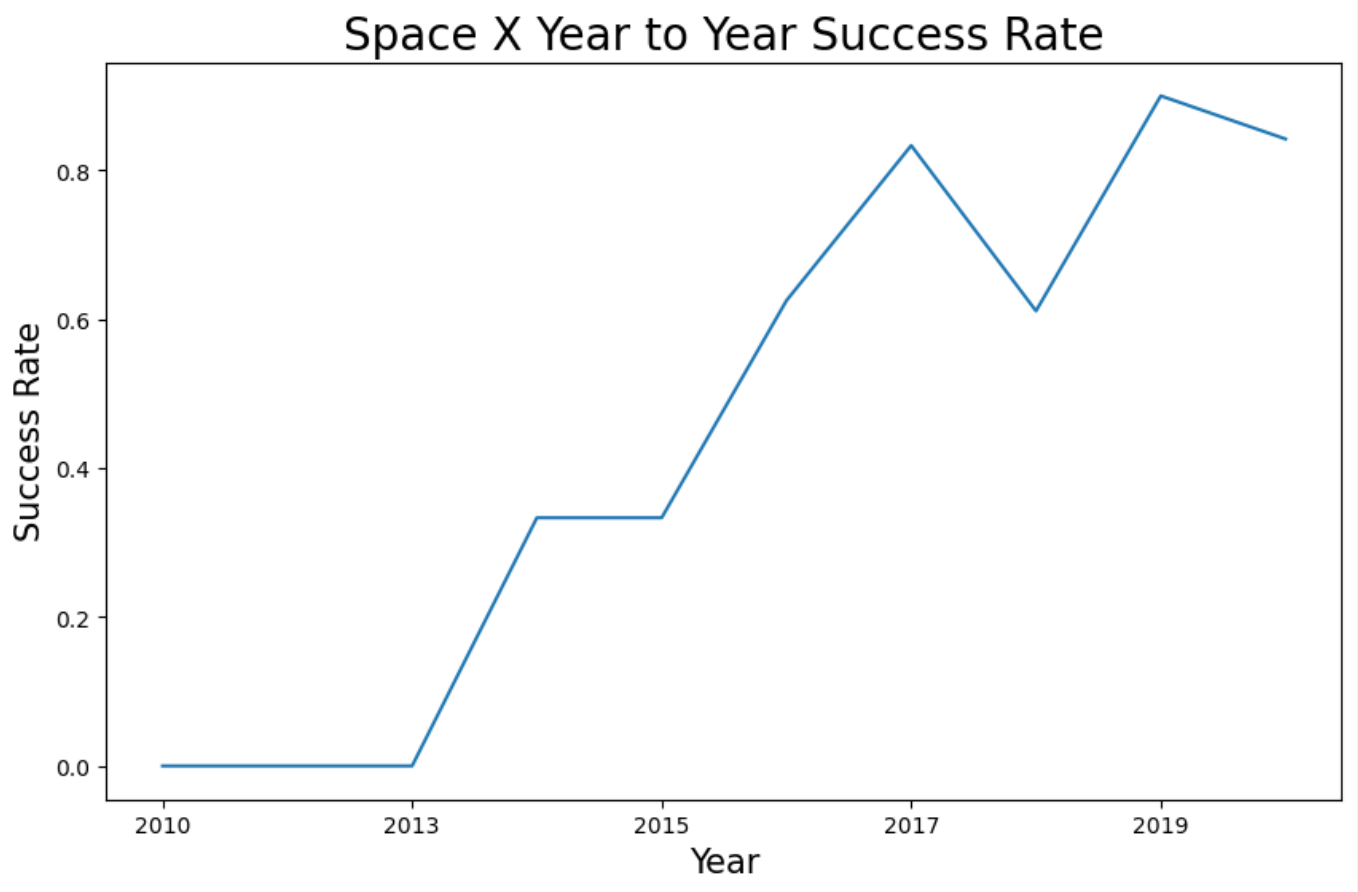
Payload vs. Orbit Type

- Positive relation between increase payload mass and success for LEO, ISS, and PO orbit
- No relation between payload mass and success for GTO.



Launch Success Yearly Trend

- Space X overall success rate has trended upward from 2013 to 2020.



All Launch Site Names

- Used SELECT DISTINCT to query the 5 Space X launch sites from the database.

```
%sql SELECT DISTINCT LAUNCH_SITE as "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'

- Used LIKE to query launch sites beginning with CCA and LIMIT to pull only 5 results from the database.

```
%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
2010-06-04	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

- Total payload carried by boosters from NASA was 45,596 kg.
- Used the SUM function in SELECT to calculate the total payload carried and WHERE clause to only include NASA (CRS)

```
%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

- The average payload mass carried by booster version F9 v1.1 was 2,928.4 kg.
- Used the AVG function in SELECT to calculate the average payload carried and WHERE clause to only include F9 v1.1.

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

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

AVG (PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad was December 22, 2015.
- Used the MIN function in SELECT to find the oldest date and WHERE clause to only include Landing_Outcome = ‘Success (ground pad)’.

```
%sql SELECT 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

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are F9 FT B1022, F9 FT B1026, F9 FT B1021.2, and F9 FT B1031.2.
- Used the WHERE clause to only include successful drone ship landings.
- Used the AND clause to only include payloads between 4,000 and 6,000.

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

```
 * 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

- Used COUNT to calculate the total number of successful and failure mission outcomes.
- Used Group_By to summarize the totals by their outcomes.

```
%sql SELECT Mission_Outcome, COUNT(*) as total FROM SPACEXTBL GROUP BY Mission_Outcome;
```

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

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

Boosters Carried Maximum Payload

- The image to the right are the names of the booster which have carried the maximum payload mass
- Used MAX function in SELECT to find the largest payload carried.

```
%sql SELECT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

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

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

- Below are the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015.
- Used substr to specify the months and years and WHERE to specify it was a failed drone ship landing.

```
%sql SELECT substr(Date,6,2) as month, Date, Booster_Version, Launch_Site, [Landing_Outcome] FROM SPACEXTBL where [Landing_Outcome] = 'Failure (drone ship)' and substr(Date,0,5)='2015';
```

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

month	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- Rank below are 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.
- Used Group_By to summarize the totals by their outcomes and ORDER BY DESC to have the outcomes in descending order.

```
%sql SELECT [Landing_Outcome], count(*) as countoutcomes FROM SPACEXTBL WHERE DATE between '2010-06-04 ' and '2017-03-20' group by [Landing_Outcome] order by countoutcomes DESC;
```

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

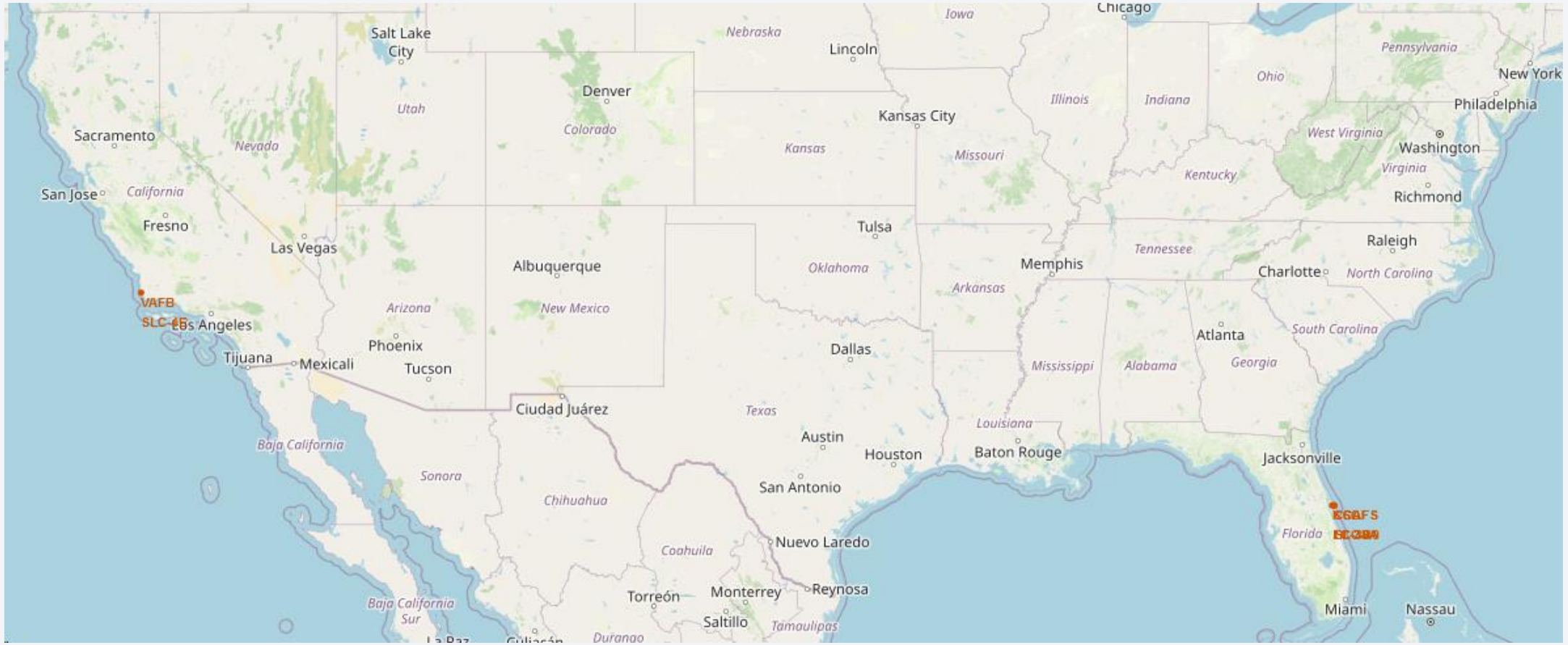
Landing_Outcome	countoutcomes
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

Section 3

Launch Sites Proximities Analysis

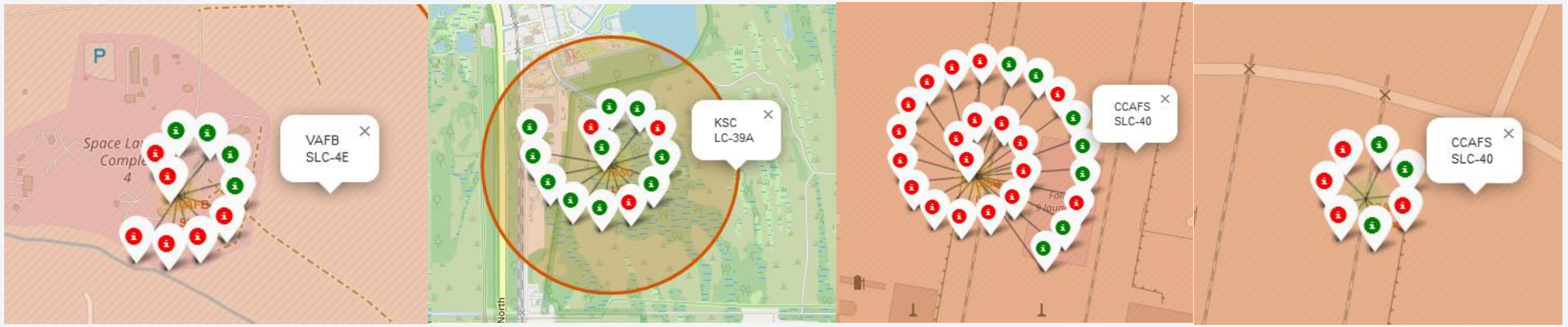
Map of SpaceX Launch Sites

- Launch sites are on the coasts of California and Florida.



Successful Launches by Launch Site – Folium Map

- California Launch Sites – VAFB SLC-4E
- Florida Launch Sites – KSC LC-39A, CCAFS LC-40, and CCAFS SLC-40
- KSC LC-39A has the best success rate.
- Green markers = Successful and Red Markers = Failed



Launch Site Proximity to Points of Interest

- All launch sites have close proximities to an airport, railway, highway, and coastline.
- CCAFS LC-40 is 10.77 km from the nearest airport, 0.86 km from the nearest coastline, 1.01 km from the nearest railway, and 0.60 from the nearest parkway.


```
print("Coastline Distance", distance_coastline)
print("City Distance", city_dist)
print("Railway Distance", railway_dist)
print("Parkway Distance", parkway_dist)
print("Airport Distance", airport_dist)
```

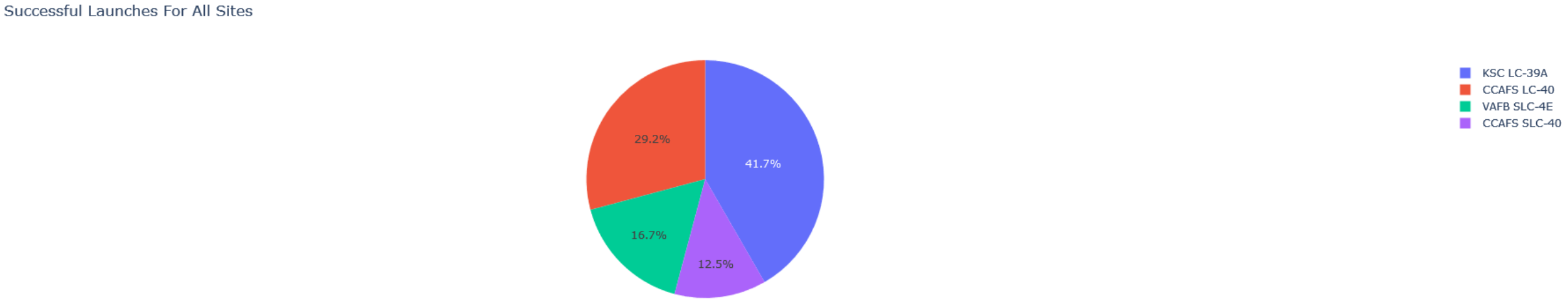
```
Coastline Distance 0.8631600594521457
City Distance 18.04623109911245
Railway Distance 1.0102606589550305
Parkway Distance 0.6006884415304674
Airport Distance 10.772986280435758
```

Section 4

Build a Dashboard with Plotly Dash

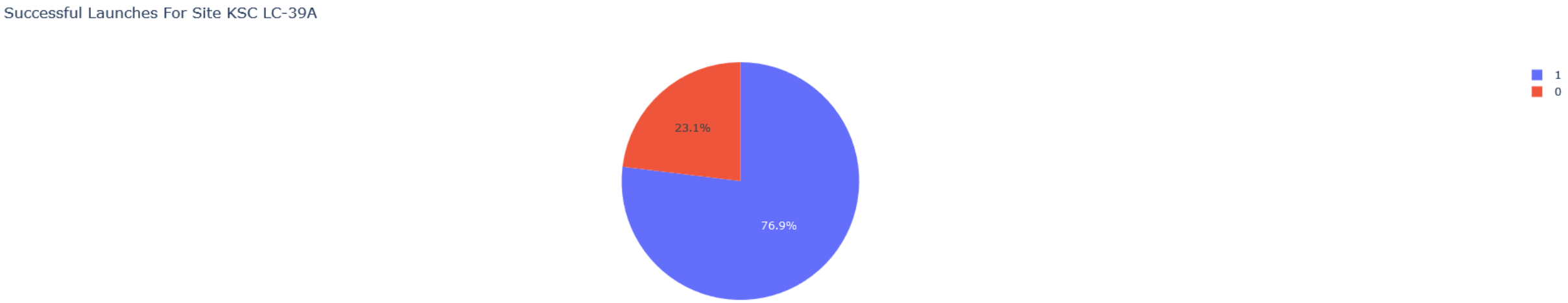
Successful Launches for All Sites – Pie Chart

- Site KSC LC-39A has the most successful launches by site.



KSC LC-39A Launch Success Pie Chart

- KSC LC-39A had a launch success of 76.9%



Payload vs. Launch Outcome Scatter Plot

Payload weight less than 5000 kg

More data and success in this range

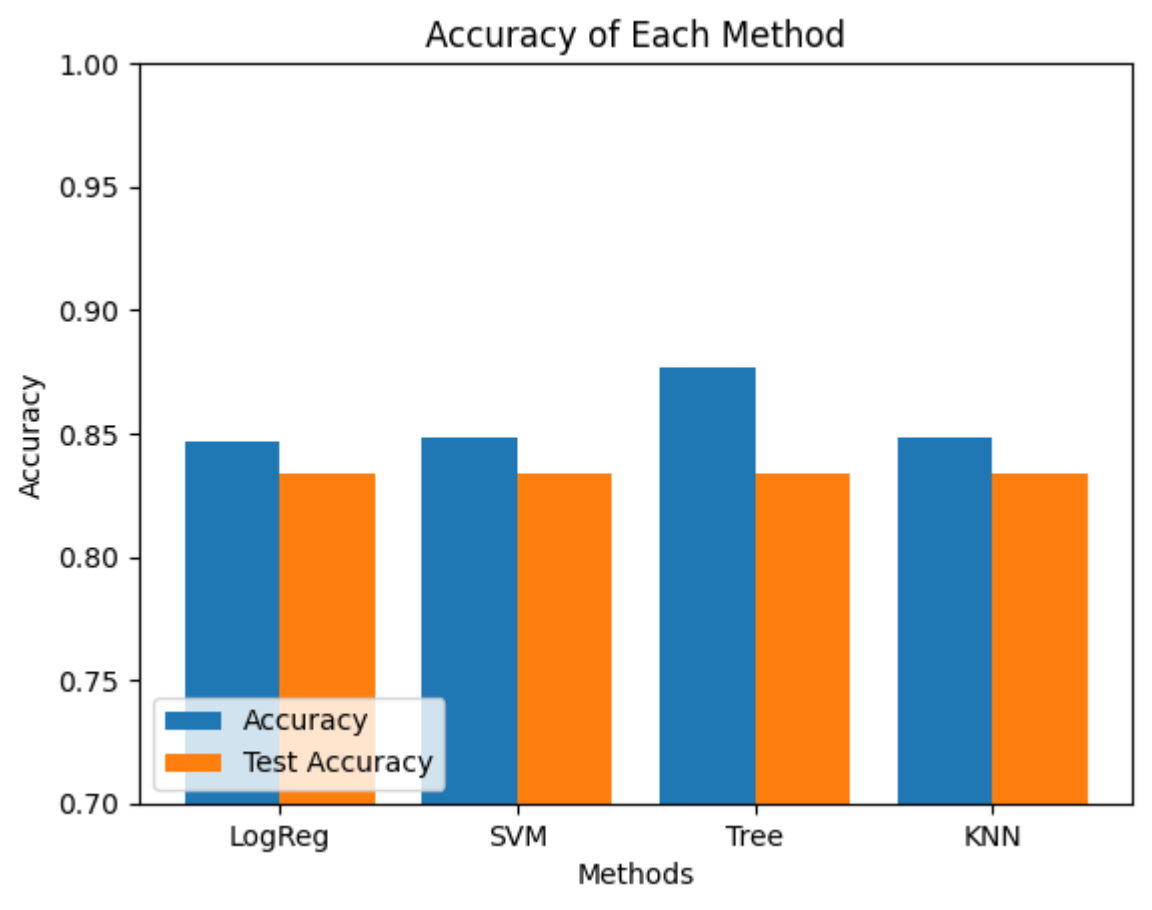
Payload weight more than 5000 kg

Section 5

Predictive Analysis (Classification)

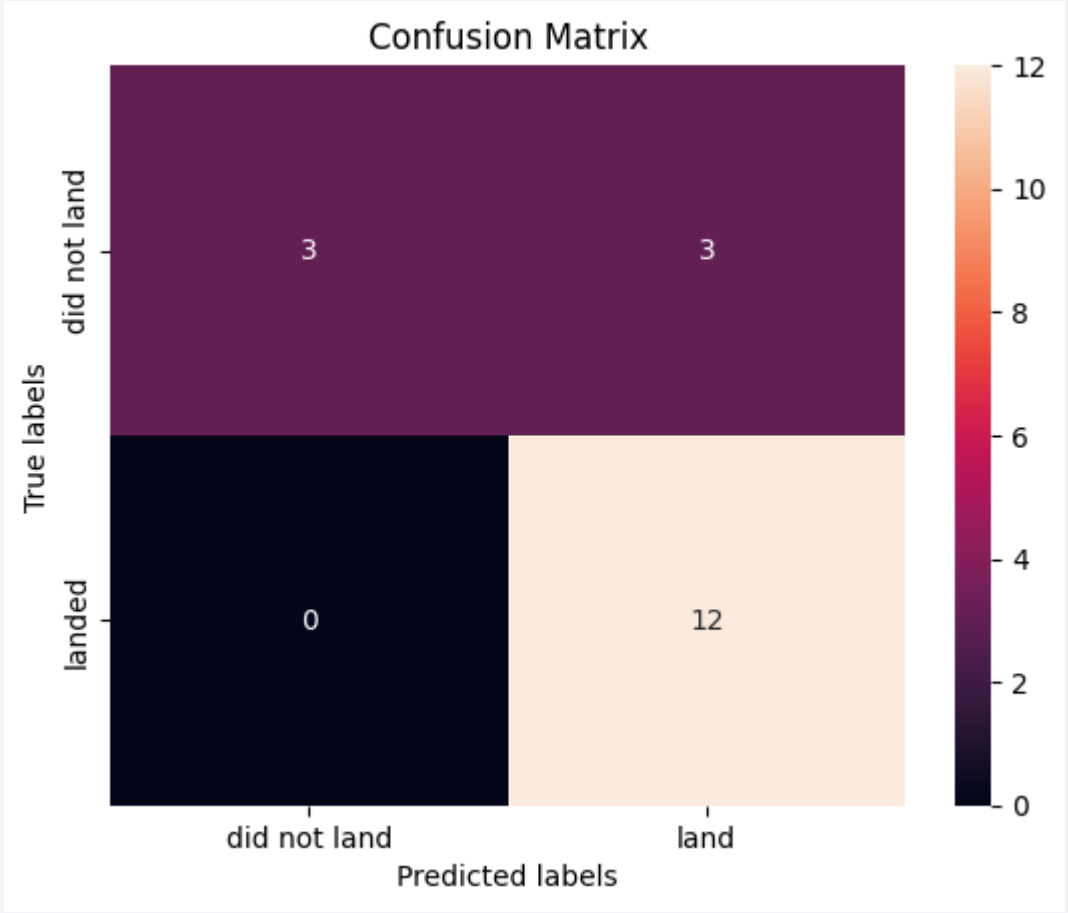
Classification Accuracy

- Visualized the built model accuracy for all built classification models, in a bar chart
- Decision Tree model has the highest classification accuracy at 87%



Confusion Matrix for Decision Tree

- Decision Tree confusion matrix shows 12 predicted landings that correspond with 12 actual landings and also 3 predicted crashes with 3 actual crashes. Only 3 out of the 18 were not predicted correctly.



Conclusions

- Launches have become more success over time and should increase with each launch and more data.
- KSC LC-39A is the most successful launch site.
- ES-L1, GEO, HEO, and SSO orbits all had 100% success rate. ES-L1, GEO, and HEO only had one flight so would need more data to conclude they would consistently be more successful than the others.
- Decision Tree Classifier is the best classification model to predict if landings will be successful.
- Payloads weighing less than 5,000 kg have more successes than payloads of over 5,000 kg but there is also less data for payloads over 5,000 kg.

Thank you!