IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Shrutika Mishra
July 15th , 2024

Outline

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

Executive Summary

- Summary of methodologies
  - **Collect** Data using SpaceX REST API and web scraping techniques
  - **Wrangle** data to create categorical success/fail variable
  - **Explore** data with data visualization techniques considering factors such as payload, launch site, flight number, and yearly trend
  - **Analyze** data with SQL, calculating statistics such as: total payload, payload range for successful launches, and total # of successful and failed outcomes
  - **Explore** launch site success rates and proximity to geographical markers
  - **Visualize** launch sites with most success and successful payload ranges
  - **Build Models** to predict landing outcomes using logistic regression, support vector machine (SVM), decision tree, and K-nearest neighbor (KNN)

Executive Summary Cont.

- Summary of all results
  - Exploratory Data Analysis Result - Launch Success has increased over time, orbits ES-L1, GEO, HEO, and SSO have 100% success rate, The higher the payload mass the higher the success rate
  - Interactive analytics in Screenshots – KSCLC-39A has the highest success rate
  - Predictive Analytics Result – Decision Tree is the best Predictive model

Introduction

- Background
  - SpaceX, advertises Falcon 9 rocket launches on its website as costing $62 million – other providers cost upward of $165 million each. Much of SpaceX's savings come from reusing the first stage of propulsion. Therefore, if we can reliably determine if the first stage will land, we can attempt to determine the price of a launch. This information can be used if an alternate company wants to bid against SpaceX for a launch.
- Problems you want to find answers
  - How payload mass, launch site, number of flights, and orbits affect if the first-stage will land successfully
  - Rate of successful landings over time and
  - Best predictive model for successful landing (binary classification)

Section 1

Methodology

Methodology

Executive Summary

- **Collect** Data using SpaceX REST API and web scraping techniques
- **Wrangle** data by filtering, handling missing values, and applying one hot encoding to prepare data for analysis and modelint
- **Explore** data via EDA with SQL and data visualization
- **Visualize** the data using Folium and Plotly Dash
- **Build Models** to predict landing outcomes using classification models – tune and evaluate models to find best model and best parameters

Data Collection – API

Steps

- **Request data** from SpaceX API (rocket launch data) using GET request
- **Decode response** using .json() and convert to Pandas datagrame using .json_normalize()
- **Request information** about the launches from SpaceX API using custom functions
- **Create dictionary** from the data
- **Create dataframe** from the dictionary
- **Filter dataframe** to only contain Falcon 9 launches
- **Replace missing Values** of Payload Mass with calculated .mean()
- **Export data** to csv file

https://github.com/shrutikamishra/IBM-Data-Science-Capstone_SpaceX/blob/14fe9433c6c1873ba63dc02bf9647c3c65471f7a/01_SpaceX_Data_Collection_API.ipynb

Data Collection – Web Scraping

Steps

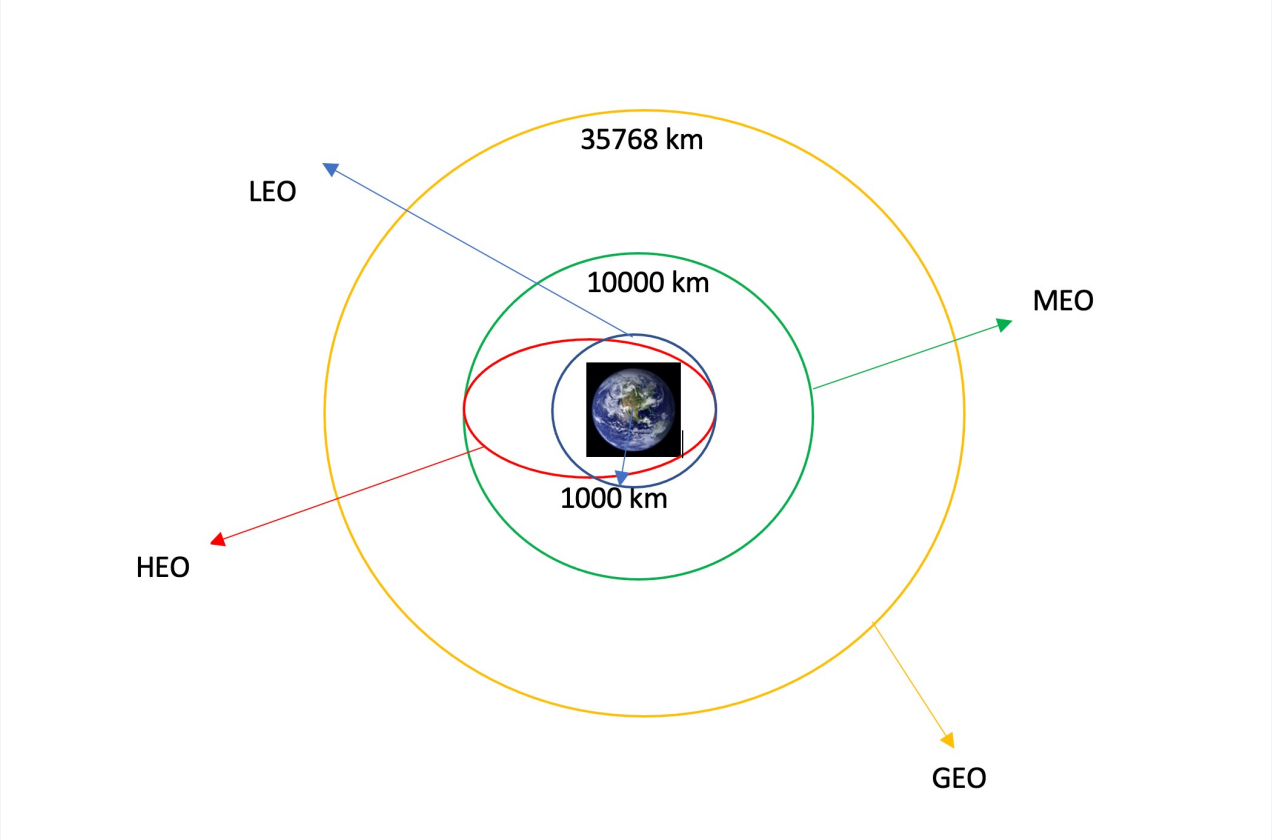
- **Request data** (Falcon 9 launch data) from Wikipedia URL
- **Create BeautifulSoup. object** from HTML response
- **Extract column names** from HTML table header
- **Create dictionary** from the data
- **Create dataframe** from the dictionary
- **Export data** to csv file

https://github.com/shrutikamishra/IBM-Data-Science-Capstone_SpaceX/blob/14fe9433c6c1873ba63dc02bf9647c3c65471f7a/02_SpaceX_Web_Scraping.ipynb

Data Collection - Scraping

Steps

- **Load Data** from previous sections
- **Identify missing values**
- **Identify numerical and categorical columns**
- **Calculate** number of launches on each site
  - Each launch aims for a different orbit
- **Calculate** number and occurence of each orbit
- **Calculate** number and occurrence of mission outcome of orbits
- **Create** landing outcome label from Outcome column for classification
- **Export to CSV**



https://github.com/shrutikamishra/IBM-Data-Science-Capstone_SpaceX/blob/14fe9433c6c1873ba63dc02bf9647c3c65471f7a/03_SpaceX_Data_Wrangling.ipynb

EDA with SQL

SQL Queries

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

https://github.com/shrutikamishra/IBM-Data-Science-Capstone_SpaceX/blob/14fe9433c6c1873ba63dc02bf9647c3c65471f7a/04_SpaceX_EDA_SQL.ipynb

EDA with Data Visualization

Charts

- Flight Number vs. Launch Site (scatter)
  - Earlier flights were generally launched from Cape Canaveral and generally failed. The majority of flights have been launched from CCAFS SLC 40 and the last 13 flights regardless of launch site have all been successful
- Payload Mass (kg) vs. Launch site (scatter)
  - The VAFB-SLC launch site has no rockets launched for heavy payload mass. As shown in the first scatter plot, the majority of the lower payload mass rockets were launched from CCAFS SLC 40.
- Success rate vs. Orbit Type (bar)
  - The most successful Orbit types are ES-L!, GEO, HEO, and SSO. The least successful are GTO, ISS, MEO, and PO.
- Flight Number vs. Orbit Type (scatter)
  - In Leo orbit, the Success appears related to the number of flights. In GTO, there seems to be no relationship between flight number and success.
- Payload Mass (kg) vs. Orbit Type (scatter)
  - For heavy payloads, Polar, LEO, and ISS orbits are generally more successful. GTO is undistinguishable.

https://github.com/shrutikamishra/IBM-Data-Science-Capstone_SpaceX/blob/14fe9433c6c1873ba63dc02bf9647c3c65471f7a/05_SpaceX_Data_Vizualization.ipynb

Build an Interactive Map with Folium

Steps

- Mark all launch sites on a map
  - Marker with popup on NASA Johnson Space Center
  - Circle for each launch site
- Mark the success/failed launches for each site on the map
  - Marker Cluster marking success and failed launches for each site
- Calculate the distances between a launch site to its proximities
  - Calculate distance to coastline and Cape Canaveral
  - Draw line from launch site to points

https://github.com/shrutikamishra/IBM-Data-Science-Capstone_SpaceX/blob/14fe9433c6c1873ba63dc02bf9647c3c65471f7a/06_SpaceX_Interactive_Visual_Analytics_Folium.ipynb

Build a Dashboard with Plotly Dash

- Dropdown List with Launch Sites
  - Select all launch sites or certain launch site
- Pie Chart Showing Successful Launches
  - Displays successful and unsuccessful launches as percent of total launches per site
- Slider of Payload Mass Range
  - Select payload mass range
- Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version
  - Shows correlation between Payload and Launch Success

https://github.com/shrutikamishra/IBM-Data-Science-Capstone_SpaceX/blob/14fe9433c6c1873ba63dc02bf9647c3c65471f7a/07_SpaceX_Interactive_Visual_Analytics_Plotly.py

Predictive Analysis (Classification)

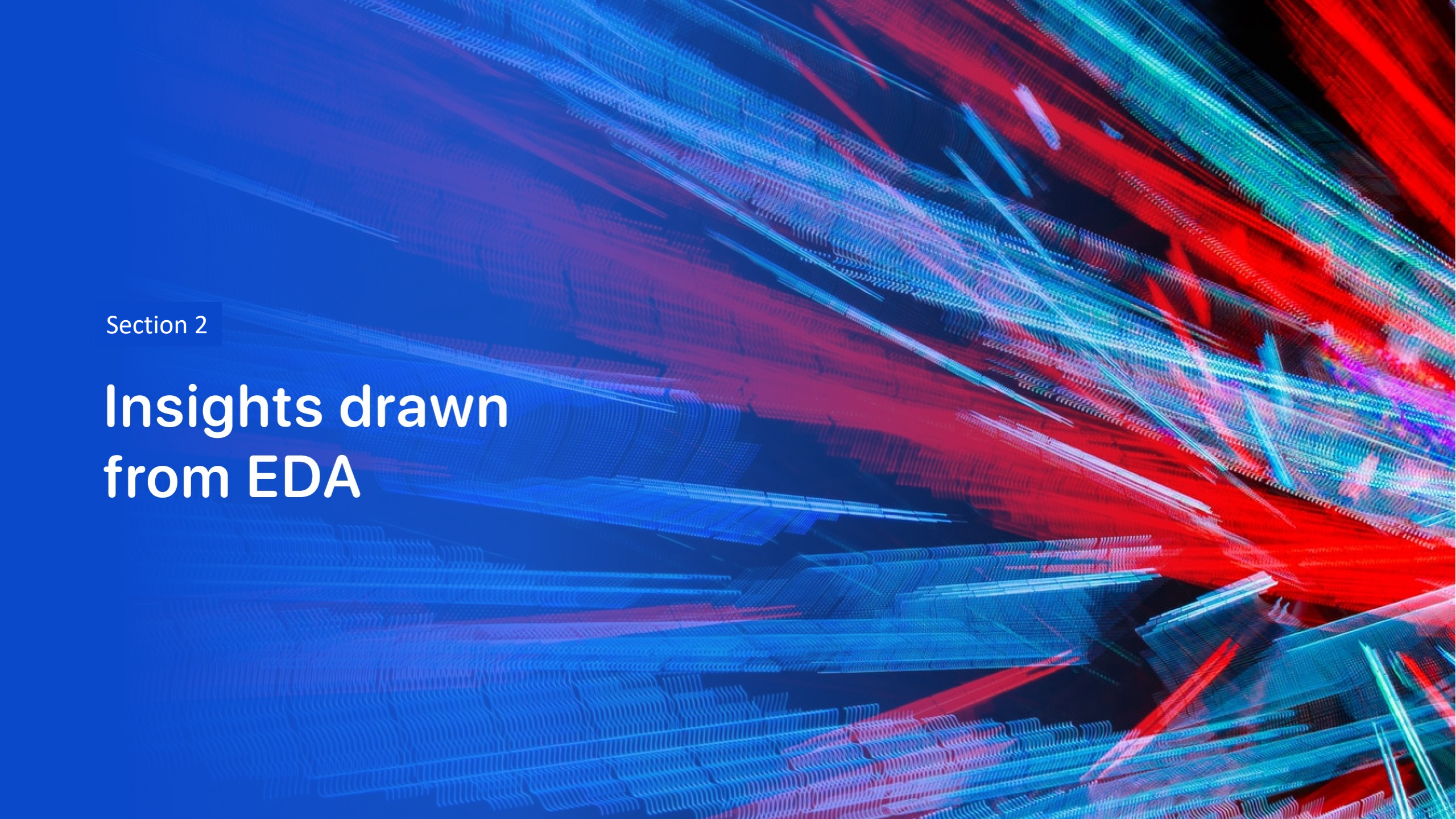
Steps

- **Create** NumPy array from the Class column
- **Standardize** the data with Standard Scaler. Fit and transform the data
- **Split** the data using train_test_split
- **Create** a GridSearchCV object with cv=10 for parameter optimization
- **Apply** GridSearchCV on different algorithms: logistic regression, support vector machine, decision tree, and K-nearest Neighbor
- **Calculate** accuracy on the test data using .score() for all models
- **Assess** the confusion matrix for all models
- **Identify** the best model using

https://github.com/shrutikamishra/IBM-Data-Science-Capstone_SpaceX/blob/14fe9433c6c1873ba63dc02bf9647c3c65471f7a/08_SpaceX_Predictive_Analytics.ipynb

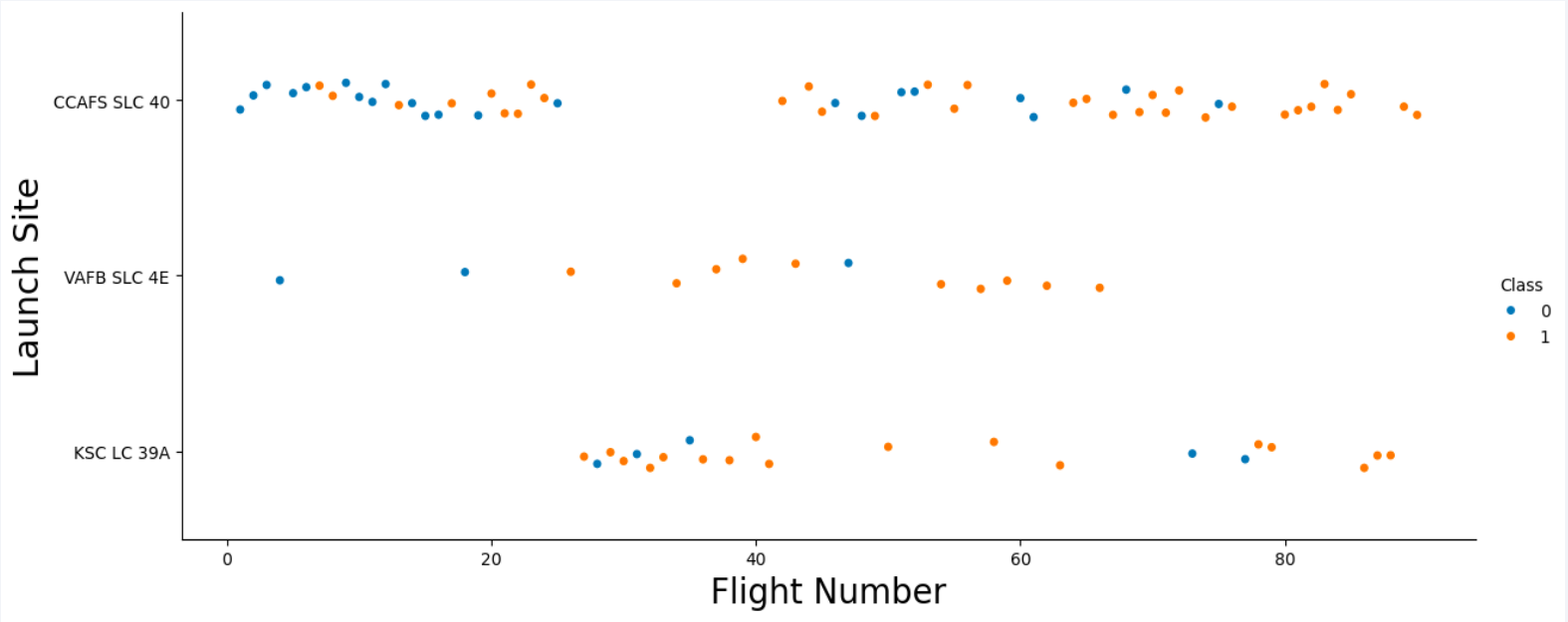
Results

- Exploratory data analysis
  - Launch Success has improved over time, especially from 2013-2020.
  - Orbits ES-L1, GEO, GEO, and SSO have a 100% success rate
  - All launches to the SSO orbit of all payload masses have been successful
- Interactive analytics demo in screenshots
- Predictive analysis results
  - Decision Tree is the best predictive model for the dataset


Section 2
Insights drawn from EDA

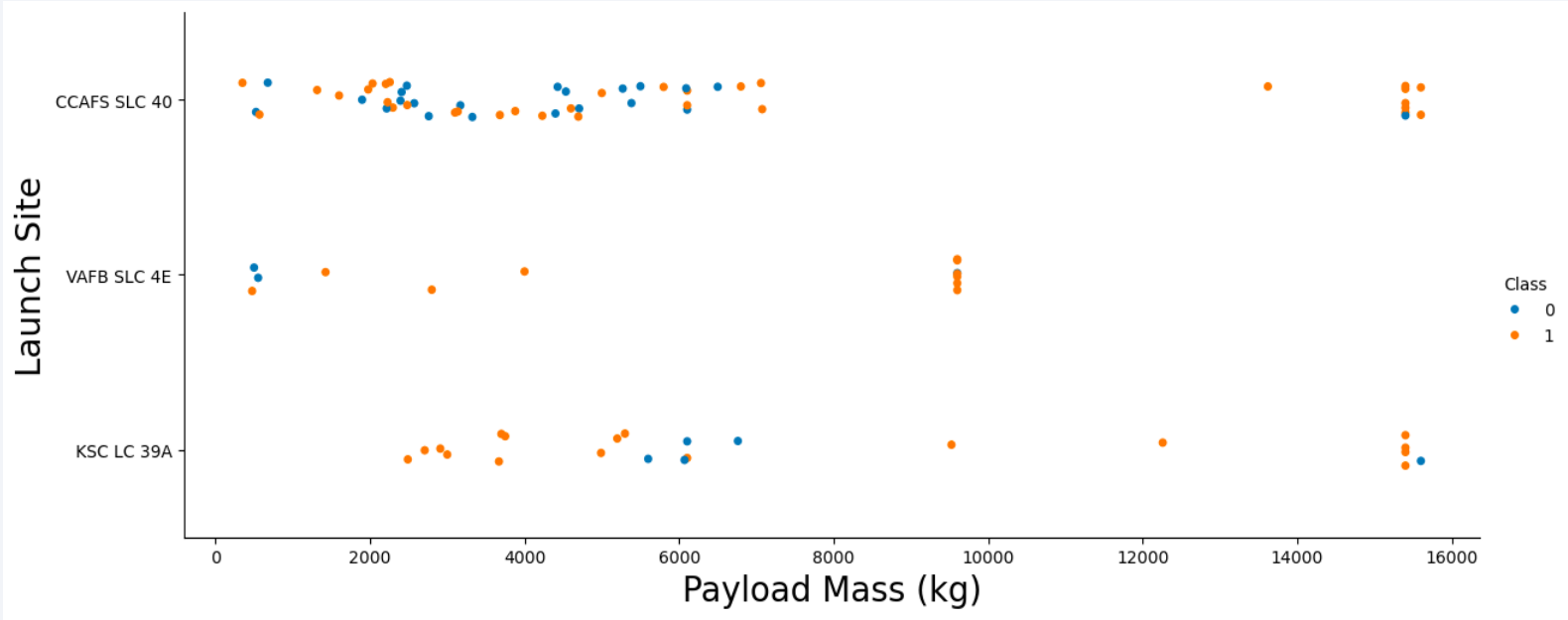
Flight Number vs. Launch Site

- Blue/0 is failure
- Orange/1 is success
- Earlier launches had lower success
- The majority of launches have been from CCAFS SLC 40.
- Recent launches have higher success rates.



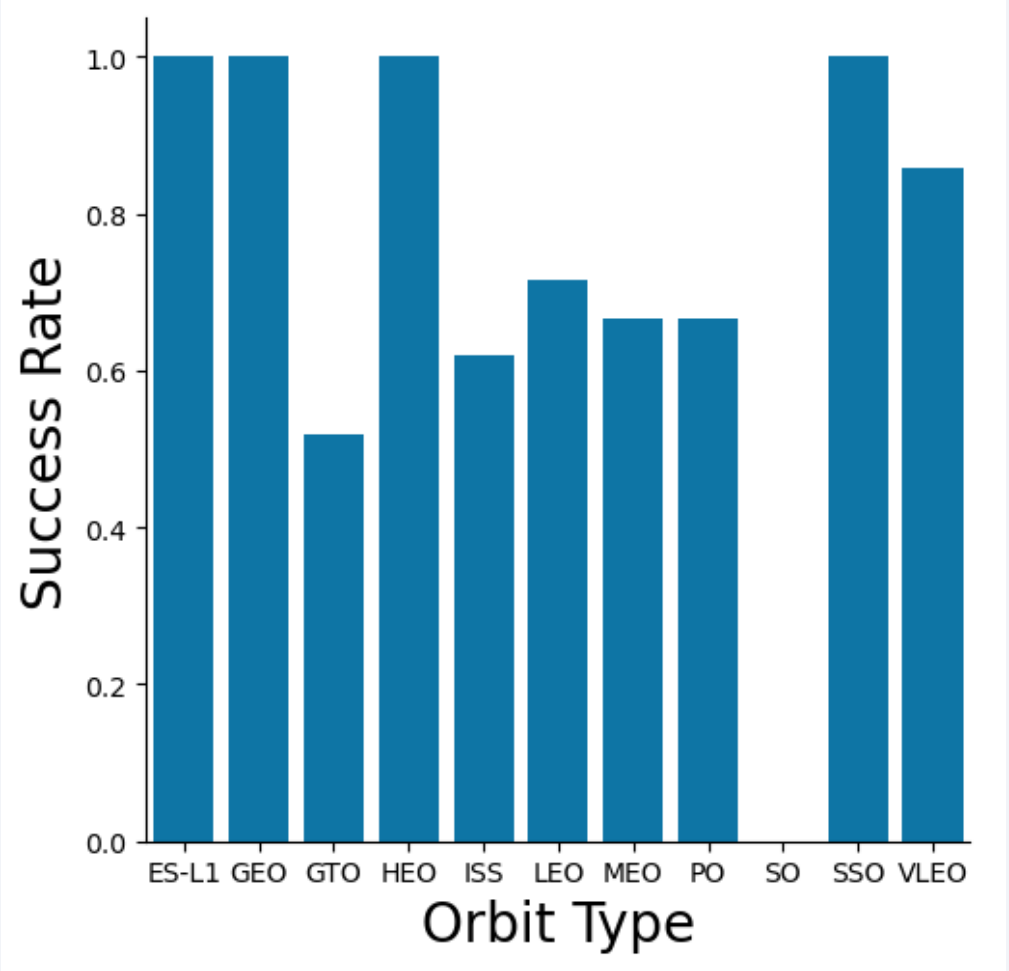
Payload vs. Launch Site

- KSC LC 39A has a 100% successrate for launches less than 5,500 kg
- VAFB has not launched anything greater than 10,000 kg
- The majority of lower payload launches were launched from CCAFS SLC 40



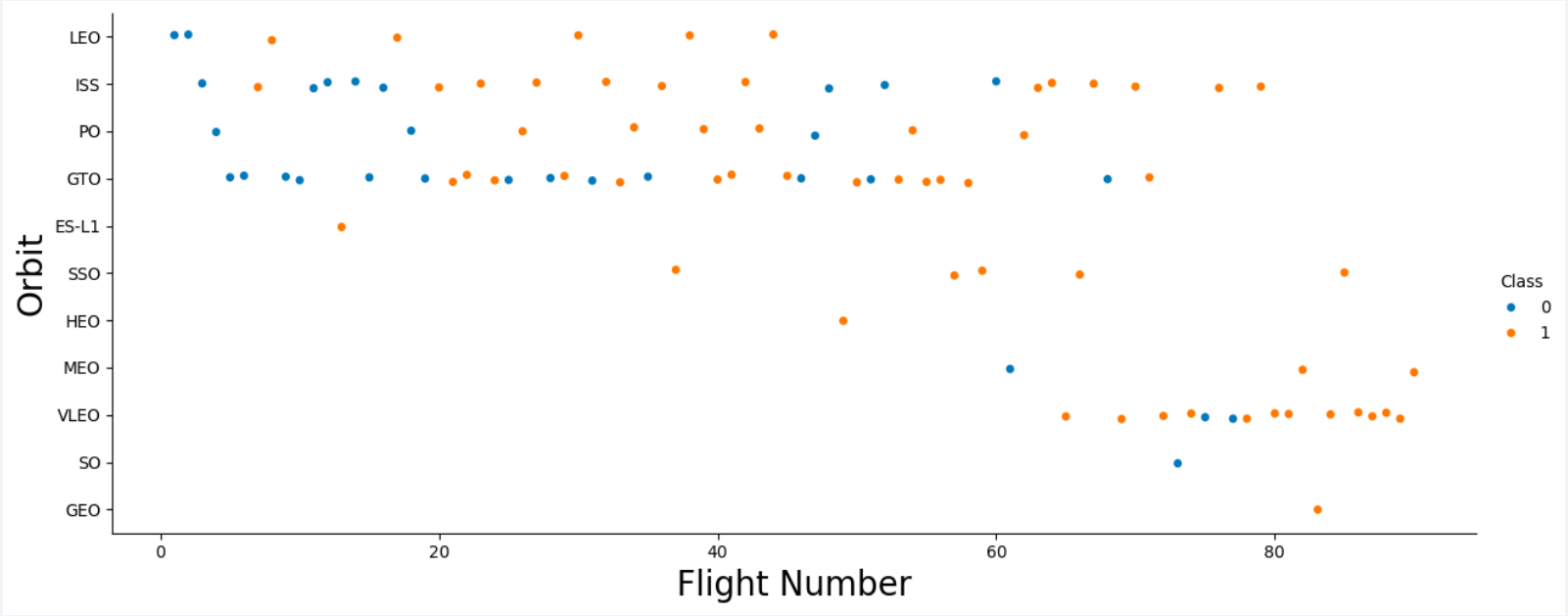
Success Rate vs. Orbit Type

- EA-L1, GEO, HEO, and SSO have 100% success rates
- GTO, ISS, LEO, MEO, and PO have 50-80% success rates.
- SO has a 0% success rate



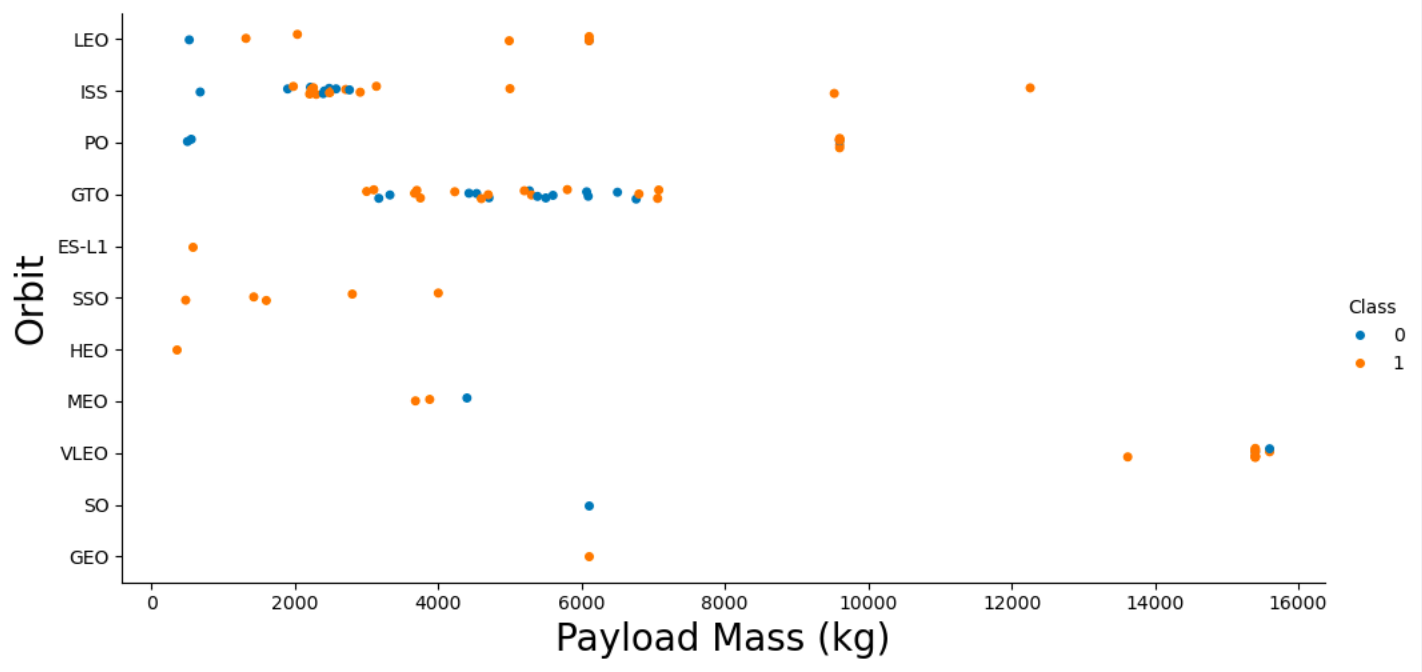
Flight Number vs. Orbit Type

- The success rate generally increases with the number of flights per orbit
- This is especially shown for the LEO and VLEO orbits
- The GTO orbit does not follow this trend though.



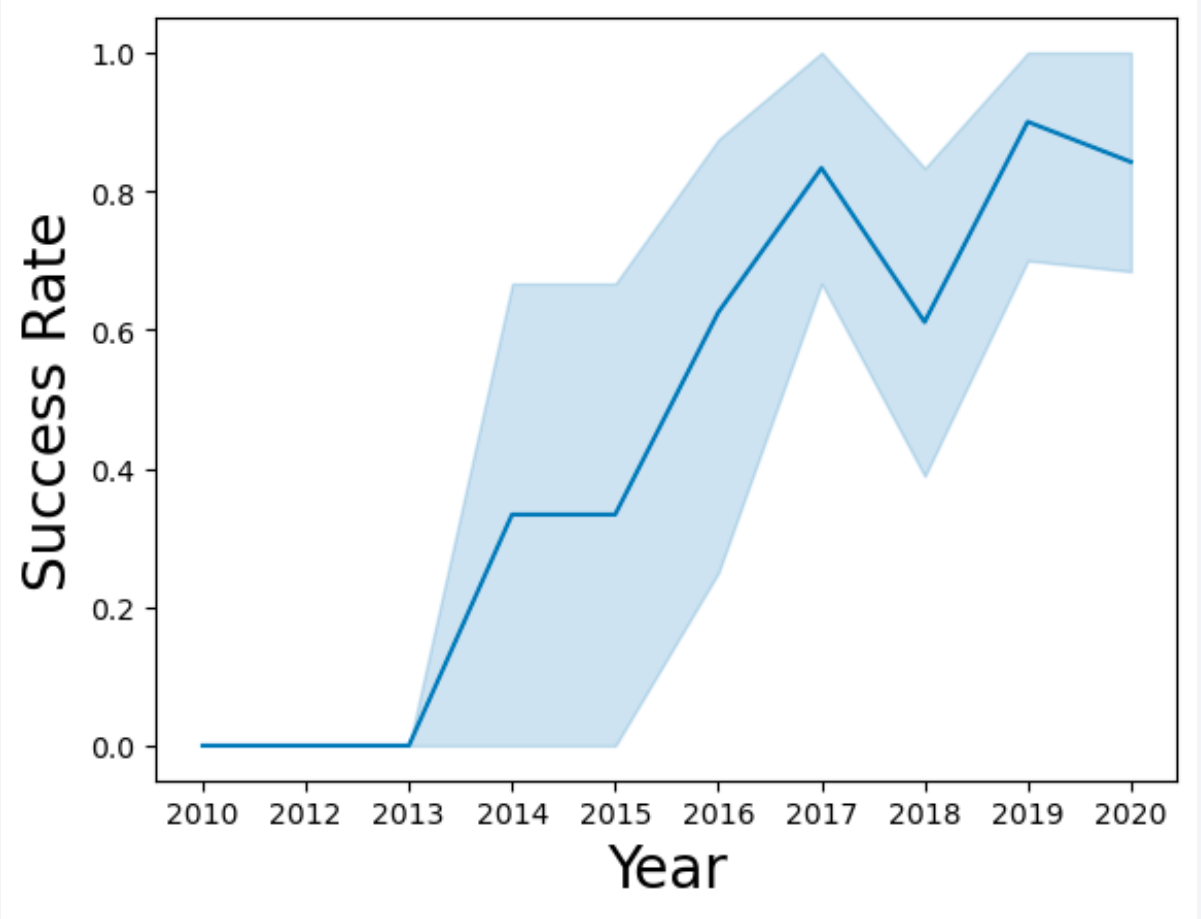
Payload vs. Orbit Type

- LEO, ISS, and PO are better with heavy payloads
- The highest successful payloads have been in VLEO orbits, but there has been a failure with the heaviest payload
- The GTO has mixed success with all payloads



Launch Success Yearly Trend

- The success rate has had an average increase over the past years
- It rose dramatically from 2013 to 2017
- Average success rate in 2020 hovers around 80%



All Launch Site Names

```
%%sql
SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL;
```

[15] ✓ 0.0s Python

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

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

- Used keyword DISTINCT

Launch Site Names Begin with 'CCA'

- Used the ”like” query to display records that begin with ‘CCA’

```
%%sql
select * from SPACEXTBL where LAUNCH_SITE like "CCA%" limit 5
```

[16] ✓ 0.0s Python

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677

Total Payload Mass

```
%%sql
select sum(PAYLOAD_MASS__kg_) from SPACEXTBL where customer = 'NASA (CRS)'
```

[21] Python

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

sum(PAYLOAD_MASS__kg_)
45596

- 45596 KG total

Average Payload Mass by F9 v1.1

```
%%sql
select avg(PAYLOAD_MASS__kg_) from SPACEXTBL where BOOSTER_VERSION = 'F9 v1.1'
```

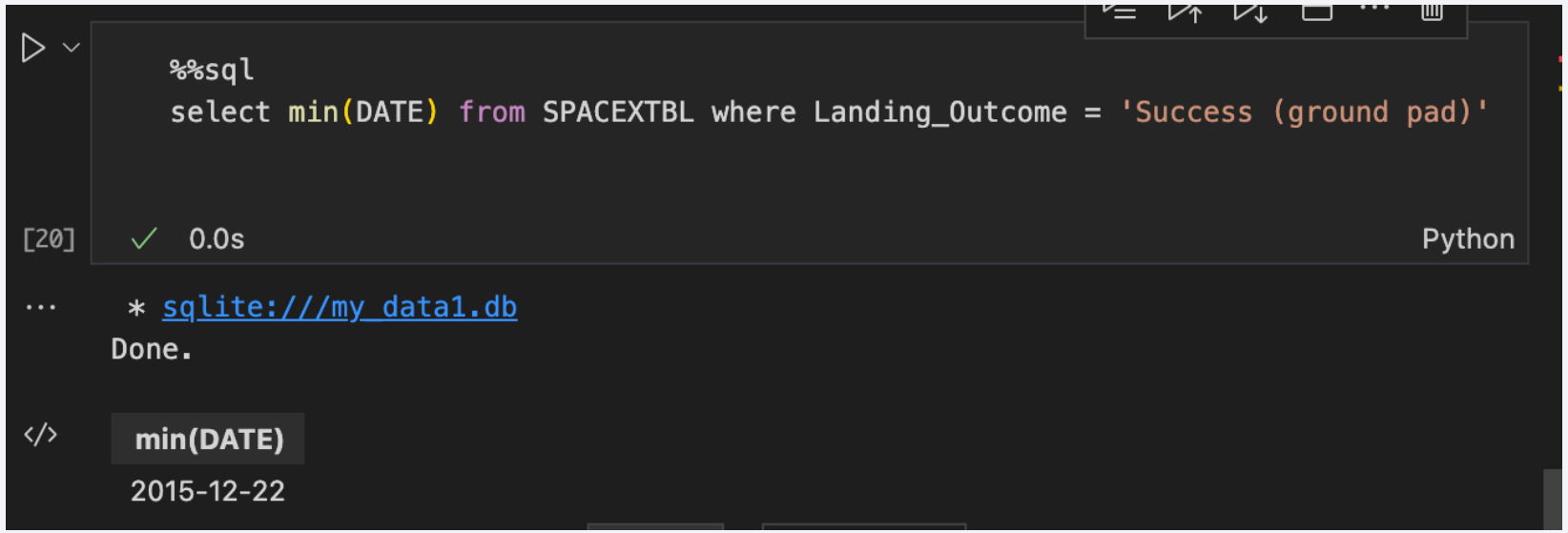
[19] ✓ 0.0s Python

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

avg(PAYLOAD_MASS__kg_)
2928.4

- 2,928 kg average

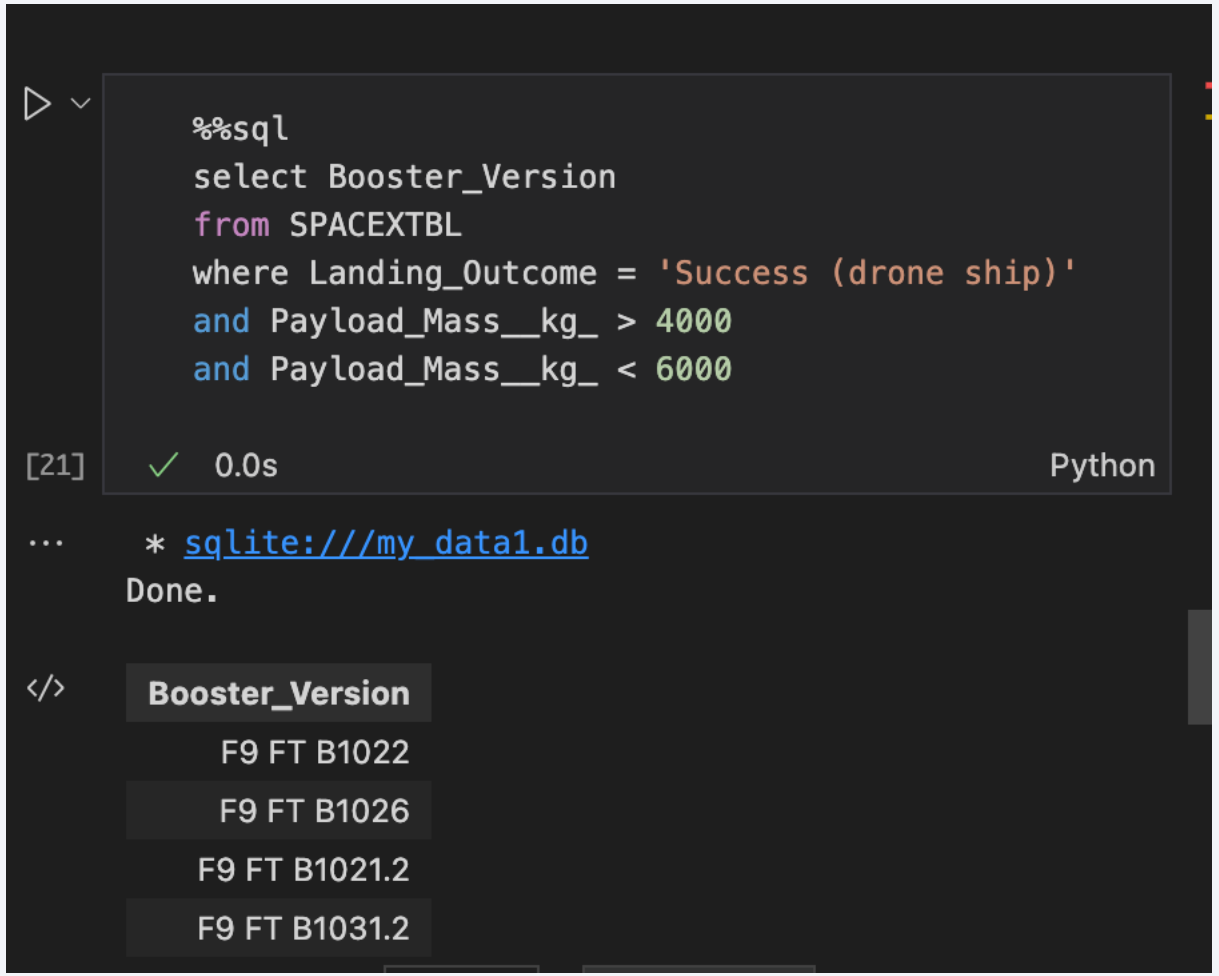
First Successful Ground Landing Date



- 12/22/15

Successful Drone Ship Landing with Payload between 4000 and 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

- 1 failure in flight
- 99 successes
- 1 success (payload status unclear)

```
%%sql

select MISSION_OUTCOME, count(*) as total_number
from SPACEXTBL
group by MISSION_OUTCOME;
```

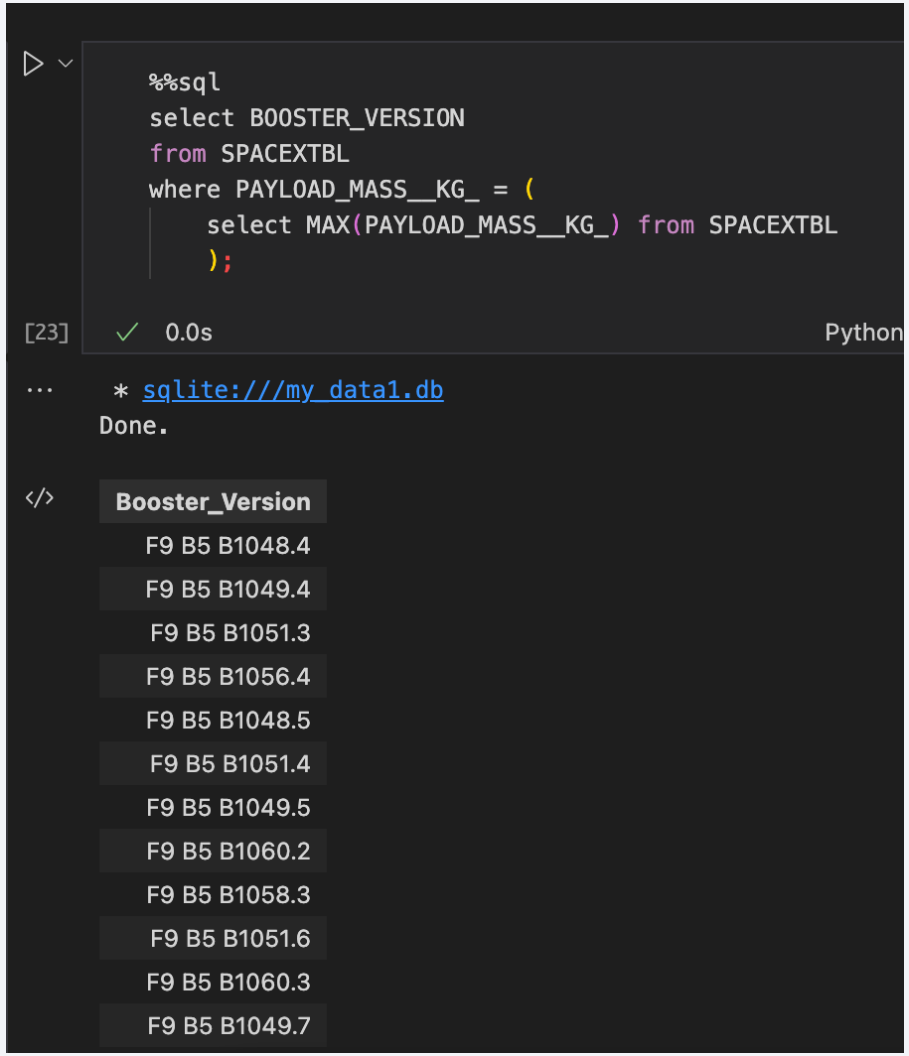
[22] ✓ 0.0s Python

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

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

Boosters Carried Maximum Payload

- F9 B5 B1048.4
- F9 B5 1049.4
- F9 B5 1051.3
- F9 B5 1056.4
- F9 B5 1048.5
- F9 B5 1051.4
- F9 B5 1049.5
- F9 B5 1060.2
- F9 B5 1058.3
- F9 B5 1051.6
- F9 B5 1060.3
- F9 B5 1049.7



2015 Launch Records

- Shows Month, date, booster version, launch site, and landing outcome

```
%%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';
```

[24] ✓ 0.0s Python

```
 * 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

- Ranked in descending order

```
%%sql

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

[25] ✓ 0.0s Python

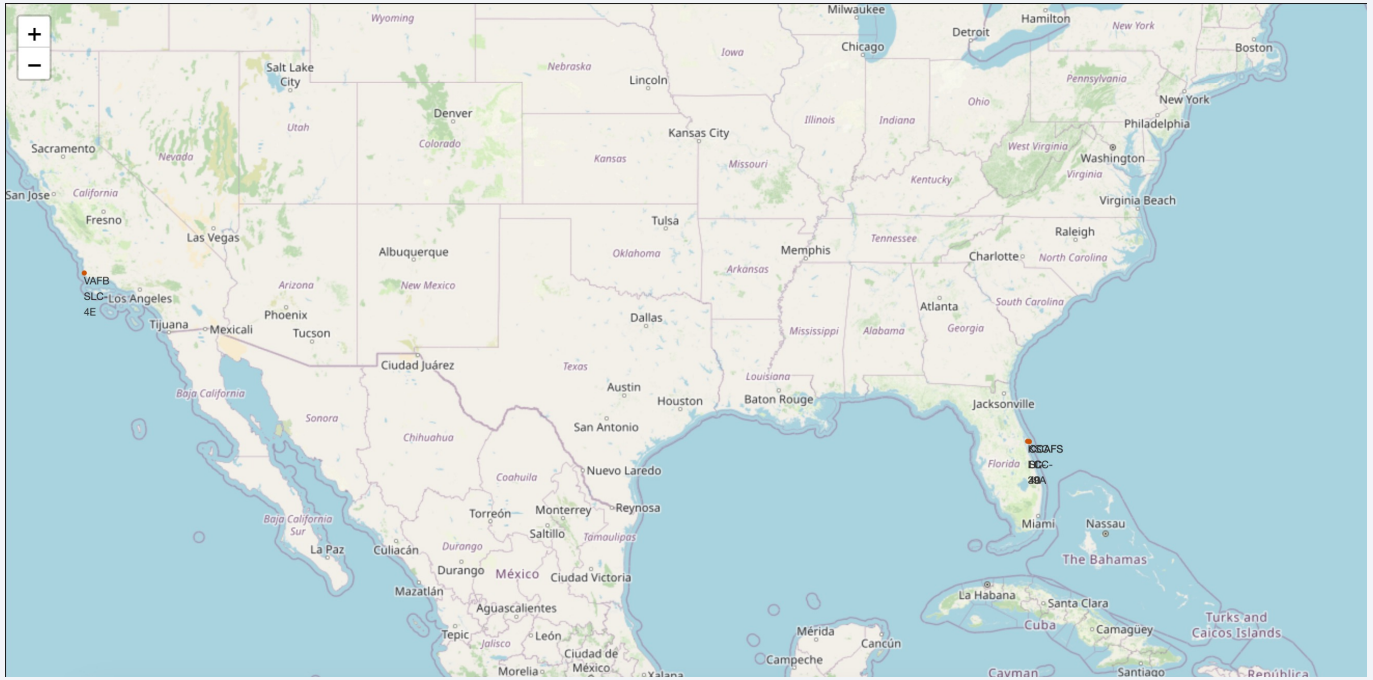
```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	count_outcomes
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

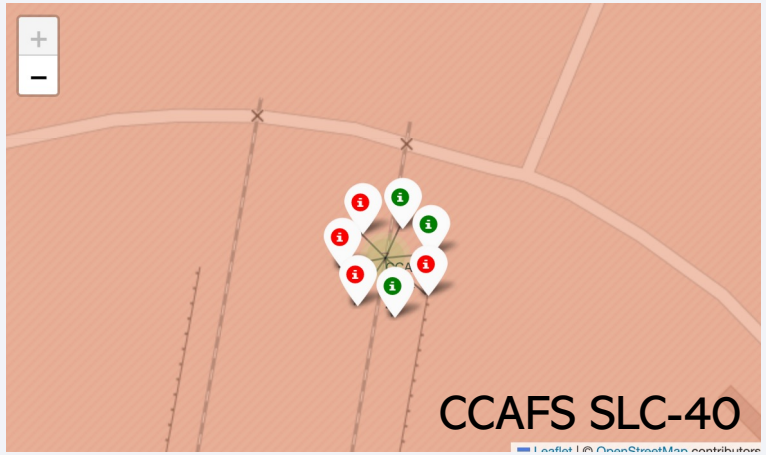
Folium – Launch Sites

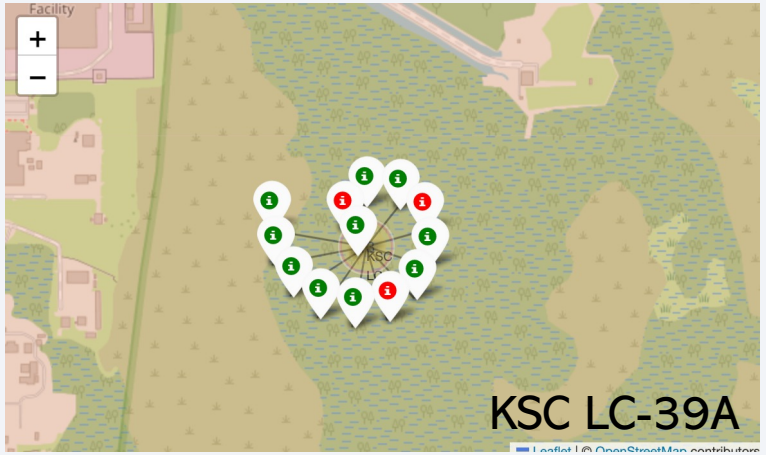
- The launch sites are generally closer to the southern United States and therefore closer to the equator
- The sites are near the coast

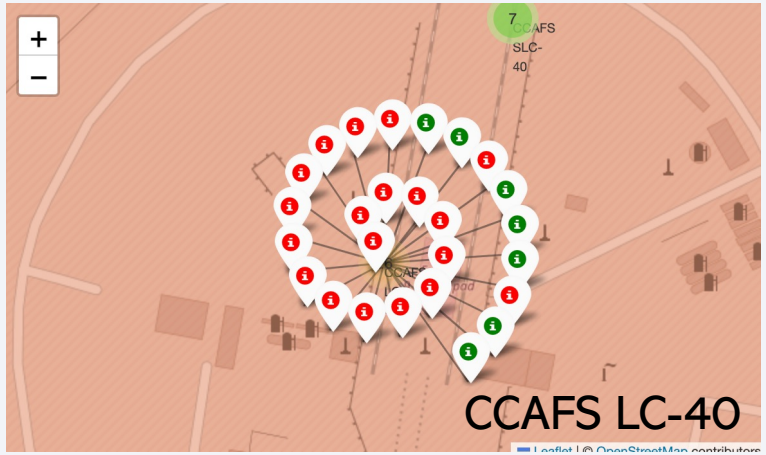


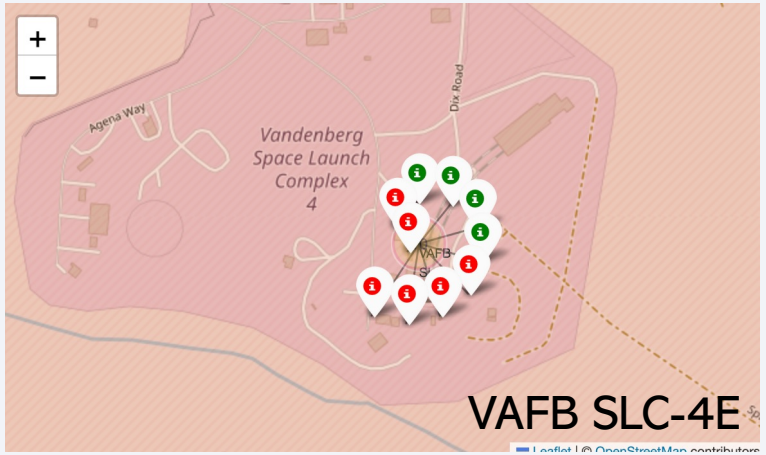
Folium – Success/Failed Launches

- The green represents successful launches and red represents failed launches
- KSC LC-39A has been the most consistently successful
- CCAFS LC-40 started out with the most failures but has started to be more consistently successful



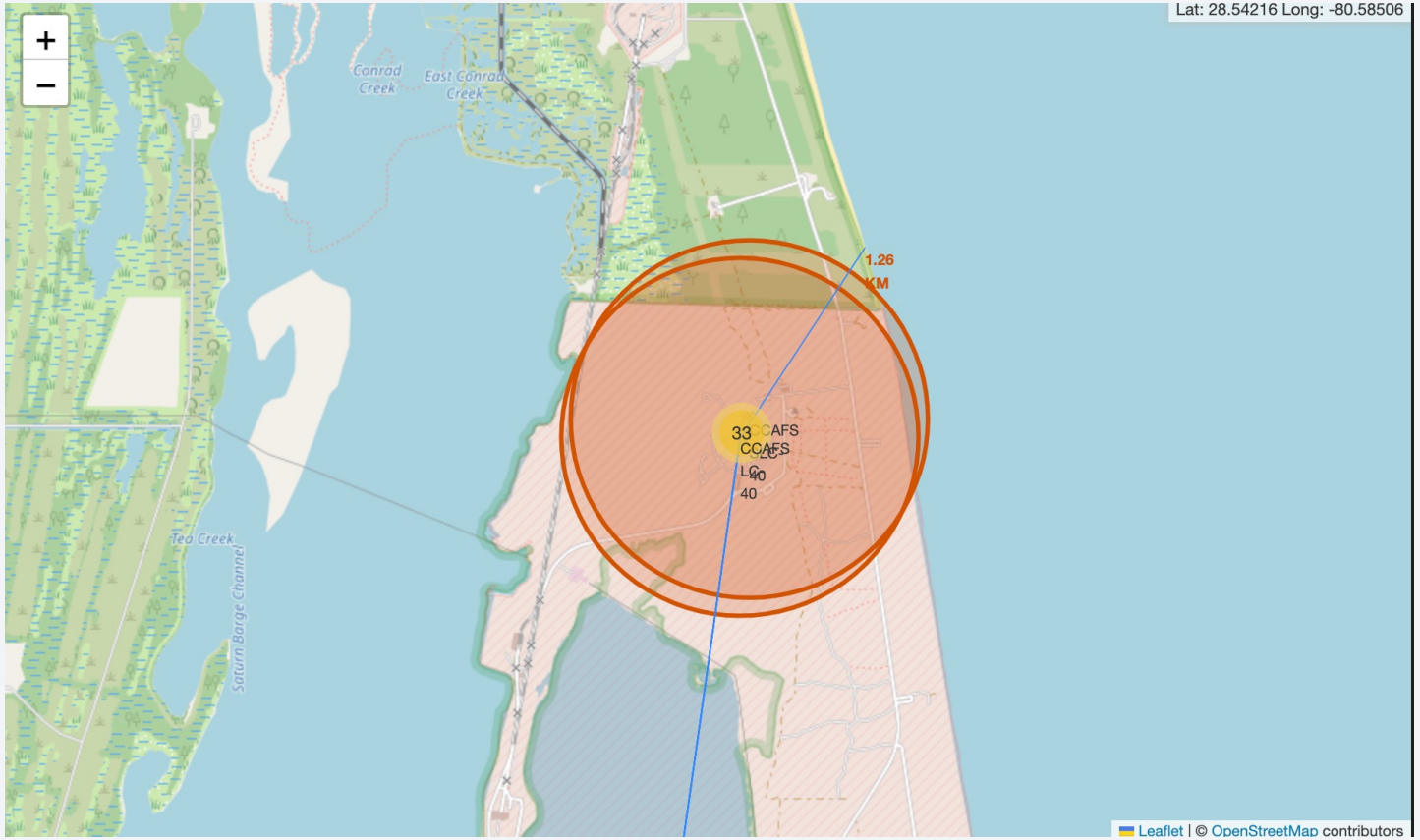


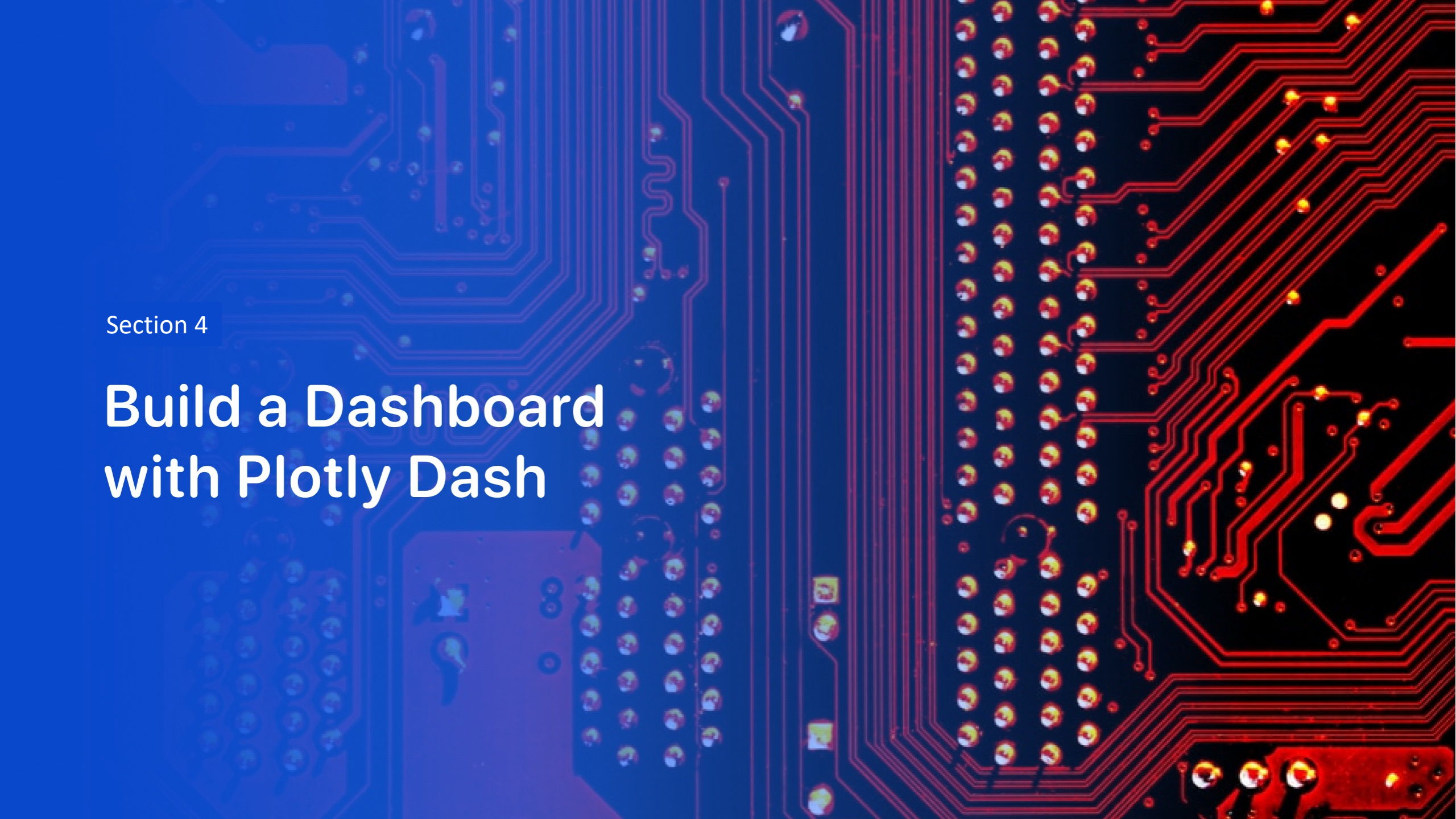




Folium – Launch Site distance to Proximities

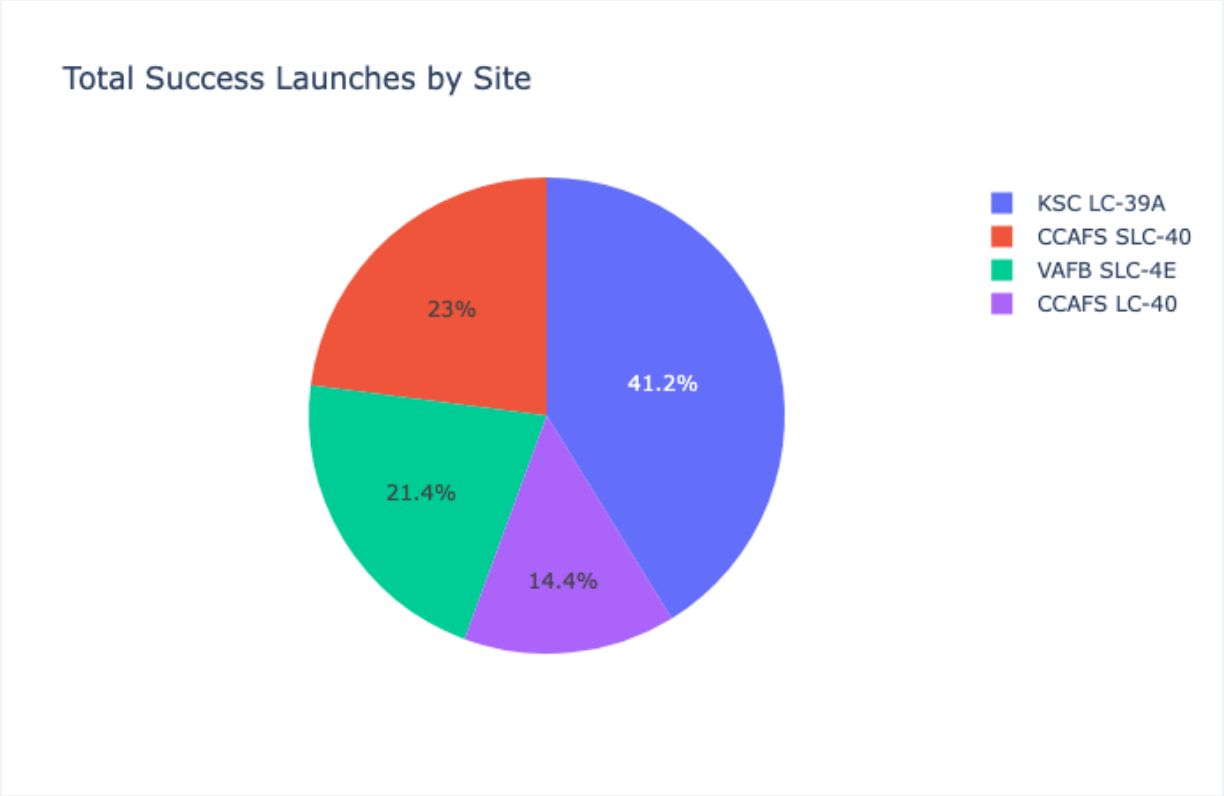
- The launch sites are not in close proximities to railways and highways.
- They are in close proximity to the coastline and do keep distance from cities.




Section 4
Build a Dashboard with Plotly Dash

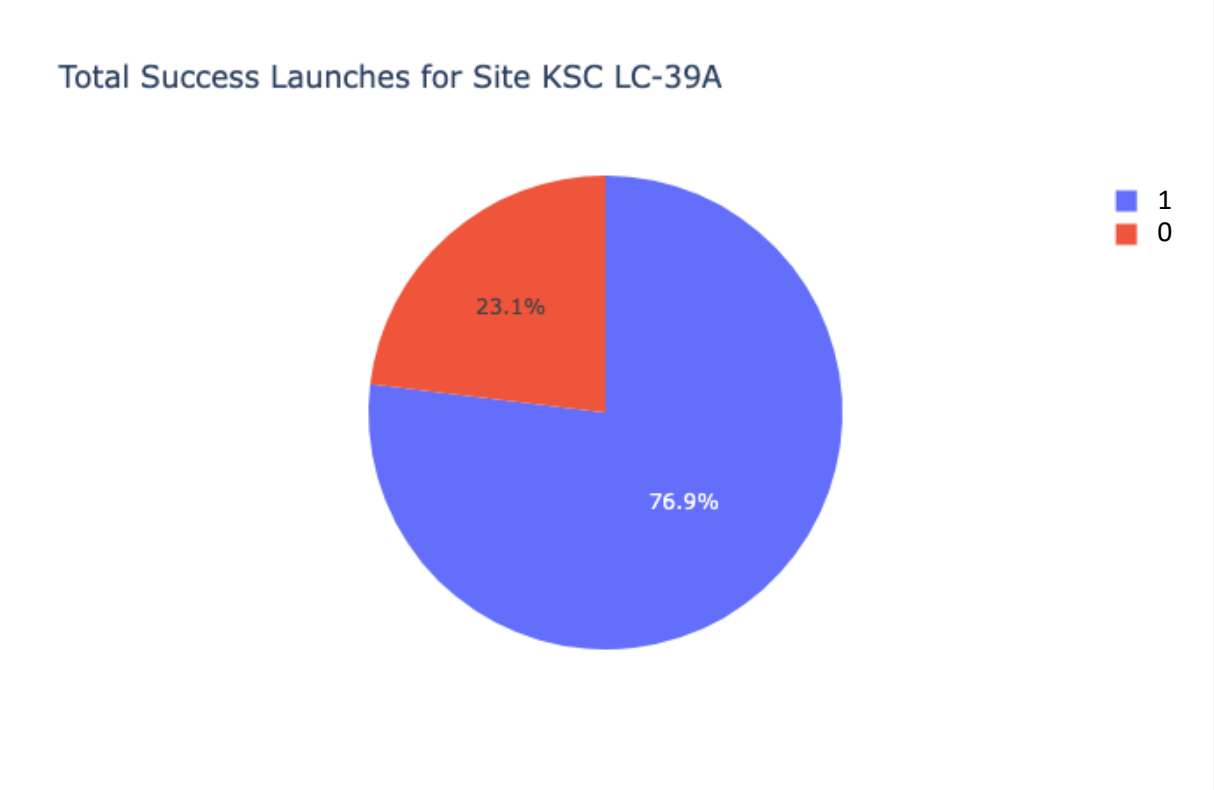
Dashboard – Total Launch Success Count by Site

- 41.2% of all total successful launches were launched from KSC LC-39A
- 23% was from CCAFS SLC-40
- 21.4% was from VAFB SLC-4E
- 14.4% was from CCAFS LC-40



Dashboard – Total Successful launches for KSC LC-39A

- 76.9% of launches from site KSC LC-39A were successful
- Only 23.1% of launches were unsuccessful



<Dashboard Screenshot 3>

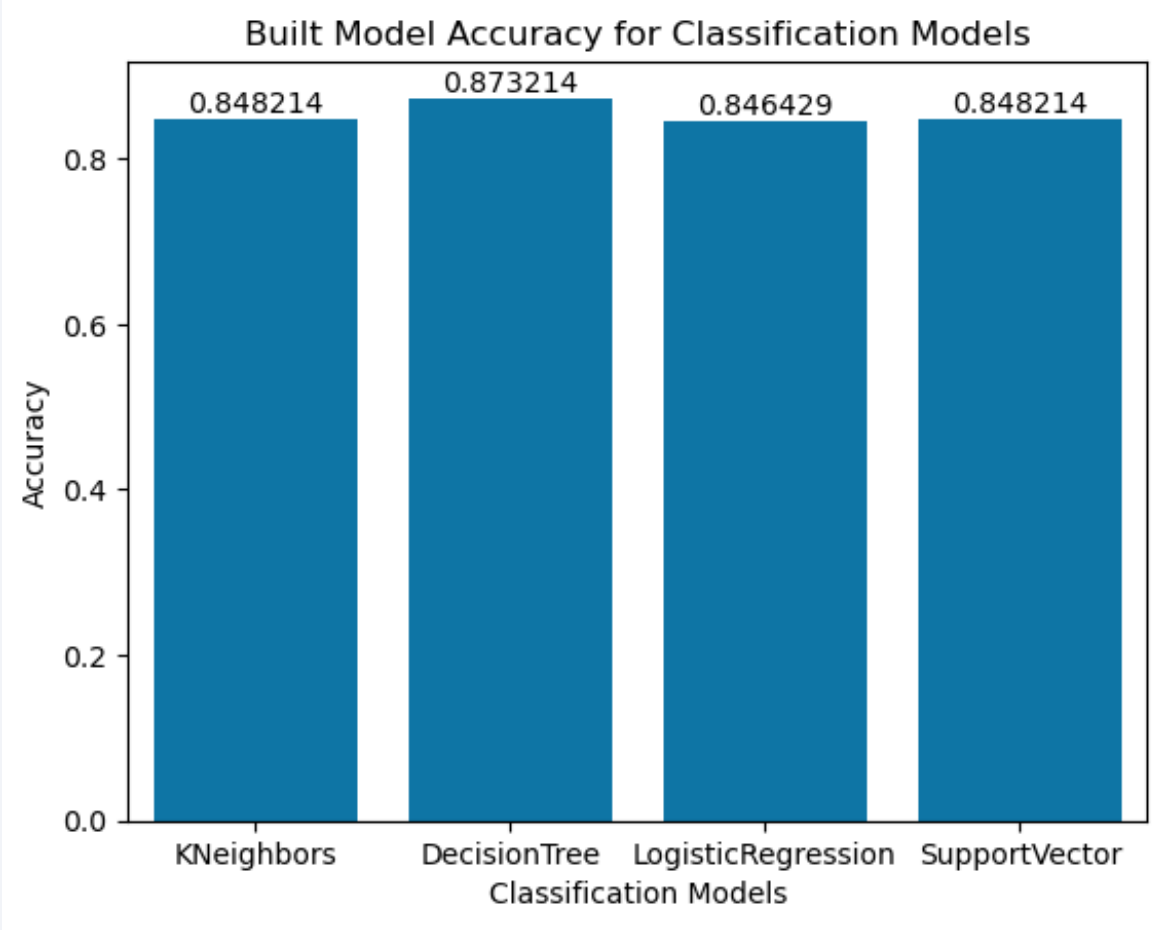
- Most launches with Booster Version 1.1 failed at all payloads
- The launch with the highest payload that was successful was launched from VAFB SLC-4E with a B4 booster, and weighed 9600 kg

Section 5

Predictive Analysis (Classification)

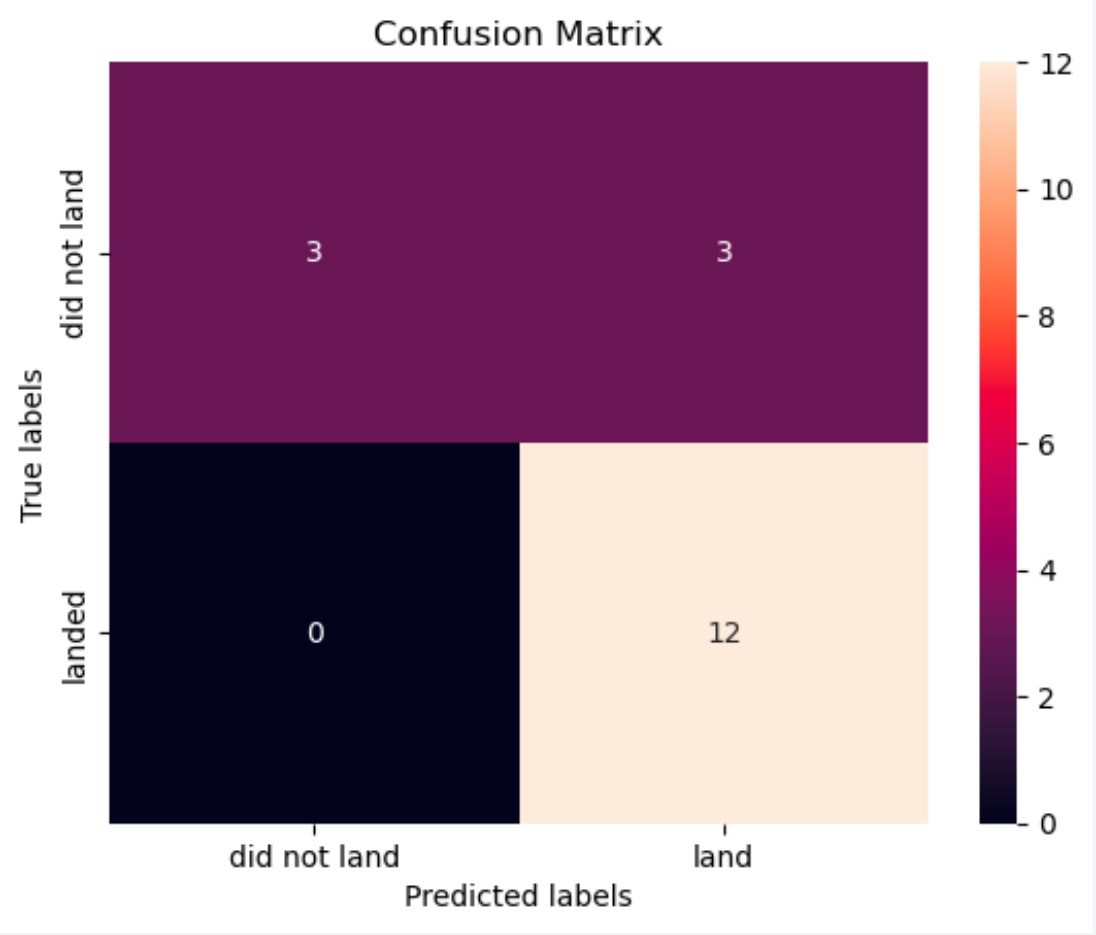
Classification Accuracy

- The Decision Tree model has the best model accuracy with 87.3214%



Confusion Matrix

- The classifier can distinguish between the different classes
- The biggest issue is with false positives i.e. unsuccessful landings being marked as successful by the classifier.



Conclusions

- How payload mass, launch site, number of flights, and orbits affect if the first-stage will land successfully
  - The higher the payload mass, the higher the success rate
  - ES-L1, GEO, HEO, and SSO orbits have a 100% success rate
- Rate of successful landings over time
  - Launch success increases over time
- Best predictive model for successful landing (binary classification)
  - The Decision tree model slightly outperforms other models

Thank you!