Winning Space Race with Data Science

<Name>
<Date>

Outline

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

Executive Summary

- Summary of methodologies

1. Data collection
2. Data wrangling
3. EDA with SQL
4. EDA with Visualization
5. Interactive Visual Analytics and Dashboards
6. predictive analysis

- Summary of all results

1. Performed EDA and extract insights
2. Built interactive dashboards to view data
3. Built machine learning models

Introduction

- Project background and context

SpaceX can recover the first stage of its Falcon 9 rocket launches with a cost of 62 million dollars while other providers mention costs upward of 165 million dollars each.

- Problems you want to find answers

1. Whether the first stage of the Falcon 9 rocket launch will land successfully or not?
2. How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?

Section 1

Methodology

Methodology

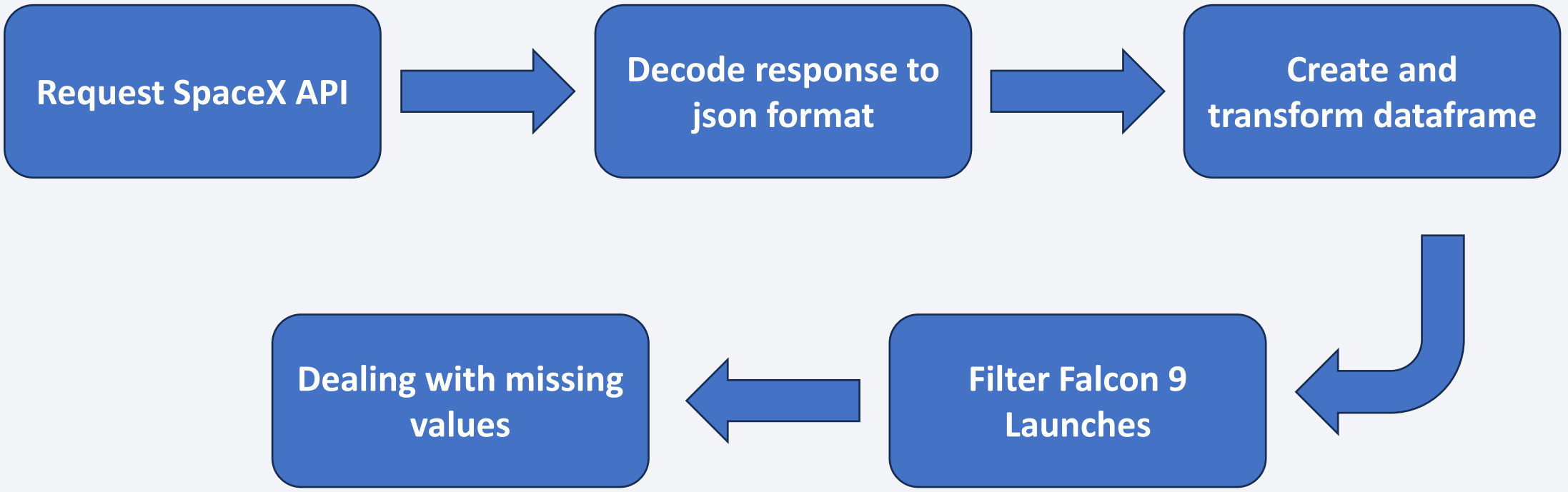
Executive Summary

- Data collection methodology:
  - Data were collected from the SpaceX API and performing Web Scraping from Wikipedia
- Perform data wrangling
  - Dealt with missing values, encoded categorical features and filtered data
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Used Scikit-Learn to standardize the data, trained and evaluated classfication models

Data Collection

- Data about rocket specification and launch information were collected by SpaceX API
- Historical Falcon 9 launch information were collected performing Web Scraping

Data Collection – SpaceX API

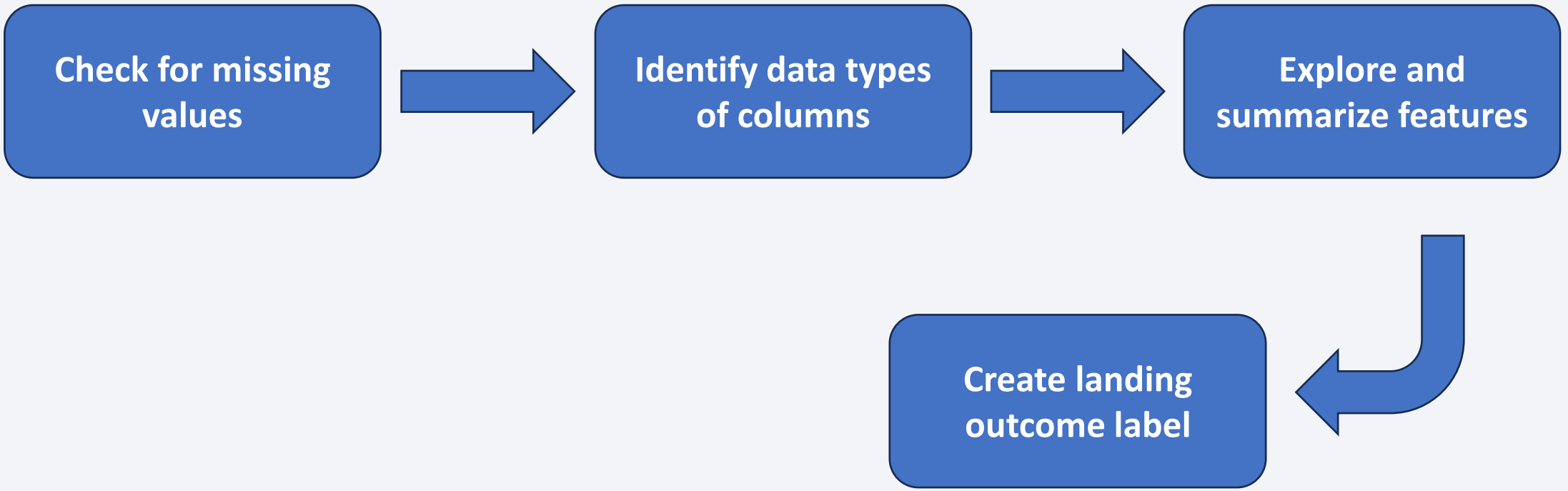


GitHub url: https://github.com/Jackch56/IBM_Data_Science_Professional_Certification/blob/main/Data-Science-Capstone-Project/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

GitHub url: https://github.com/Jackch56/IBM_Data_Science_Professional_Certification/blob/main/Data-Science-Capstone-Project/jupyter-labs-webscraping.ipynb

Data Wrangling



GitHub url: https://github.com/Jackch56/IBM_Data_Science_Professional_Certification/blob/main/Data-Science-Capstone-Project/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- The following chart types were used:

1. Scatter Plot: 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

2. Bar Plot: Orbit Success Rate

3. Line Plot: Success Rate Over Time

Scatter Plots were used to identify relationships between features, Bar Charts helped to find the success rate distribution of orbits, and Line Chart was useful to identify an increase in success rate over time.

GitHub url: https://github.com/Jackch56/IBM_Data_Science_Professional_Certification/blob/main/Data-Science-Capstone-Project/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

The below SQL queries were used to better understand the data:

- 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 succesful landing outcome in ground pad was acheived
- 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. Use a subquery
- 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: https://github.com/Jackch56/IBM_Data_Science_Professional_Certification/blob/main/Data-Science-Capstone-Project/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- **Added Markers of all Launch Sites:**

1. Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates
2. Added the same for all Launch Sites using their respective latitude and longitude coordinates
3. Added coloured markers for landing outcomes for each Launch Site using MarkerCluster to identify Sites with relatively higher success rate. Success is green and Failure is red

- **Distance between a Launch Site and its proximities:**

Added colored lines between the Launch Site and proximities around it like the railway, highway, coastline and closest city

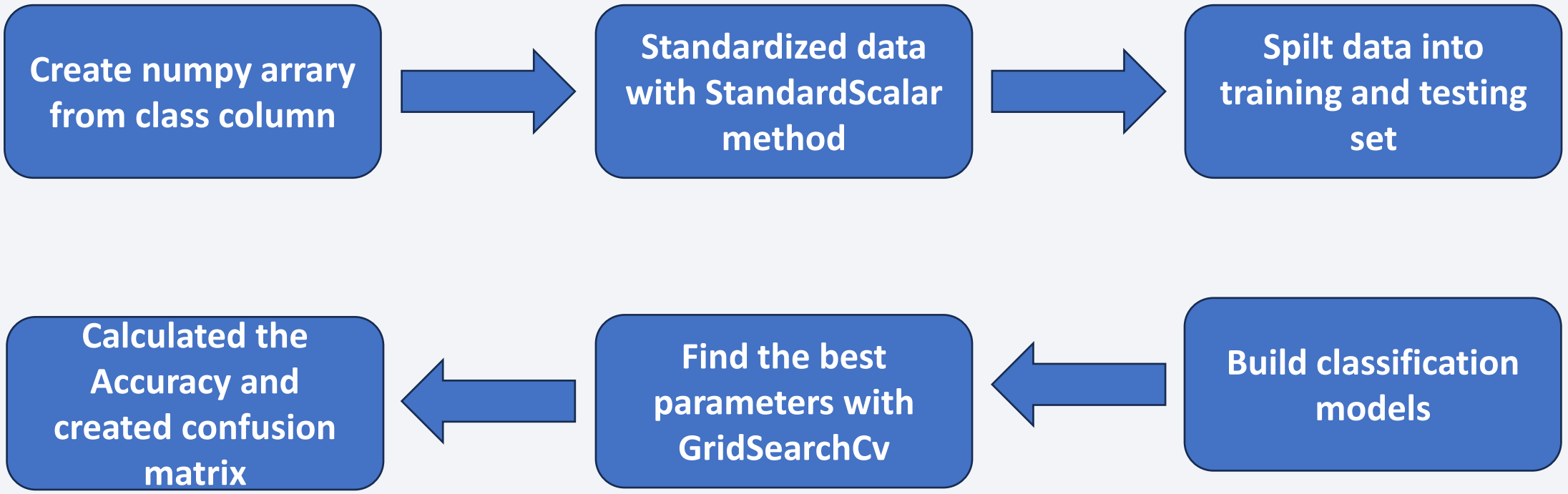
GitHub url: https://github.com/Jackch56/IBM_Data_Science_Professional_Certification/blob/main/Data-Science-Capstone-Project/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- Added a Launch Site drop-down menu for launch site selection
- Added a pie chart to visualize total launches for all sites or successful vs failed launches for a specific launch site
- A scatter plot is displayed to visualize the relationship between payload mass and successful launches
- A slider is included to select payload mass range

GitHub url: https://github.com/Jackch56/IBM_Data_Science_Professional_Certification/blob/main/Data-Science-Capstone-Project/spacex_dash_app.py

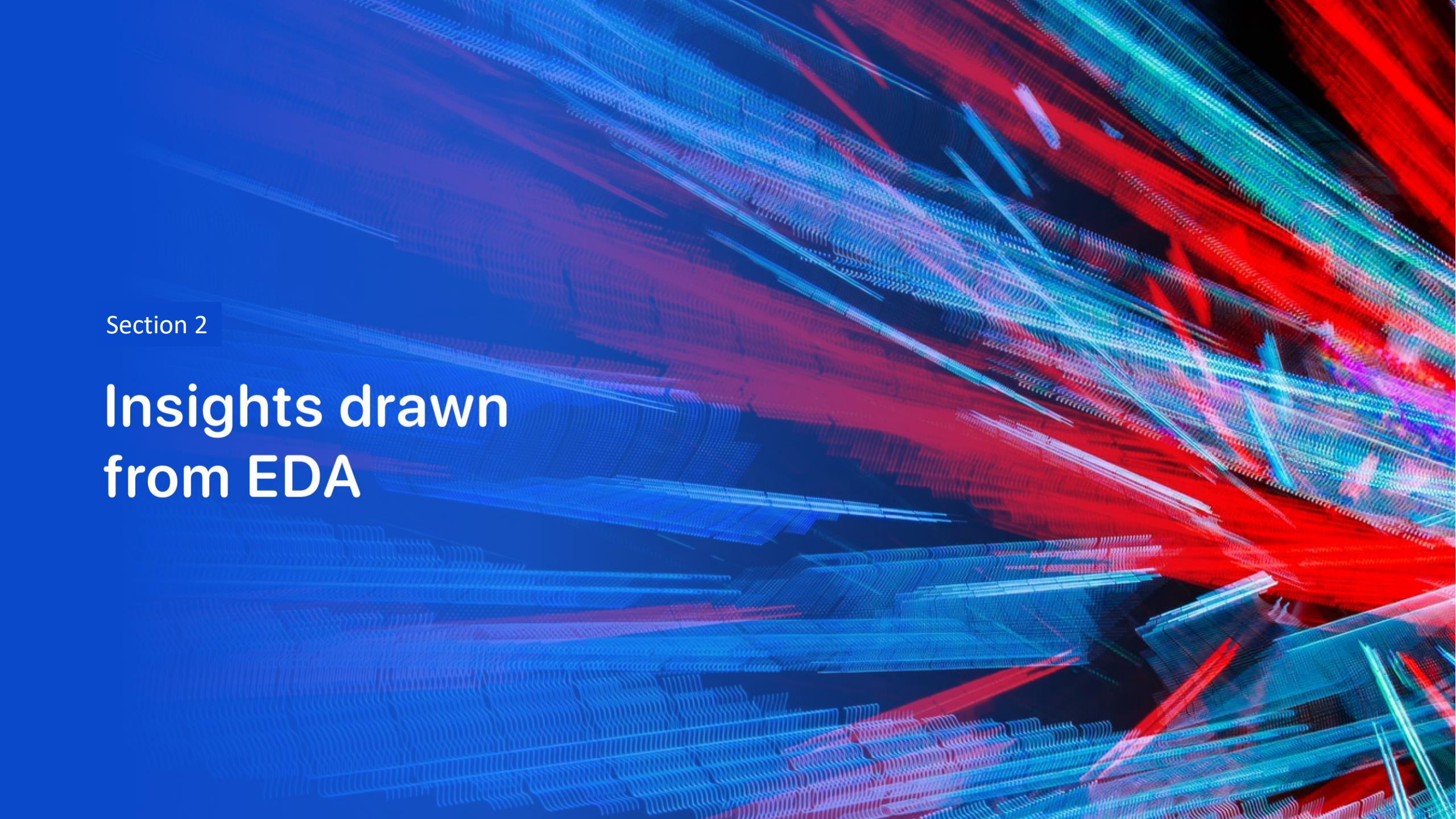
Predictive Analysis (Classification)



GitHub url: https://github.com/Jackch56/IBM_Data_Science_Professional_Certification/blob/main/Data-Science-Capstone-Project/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

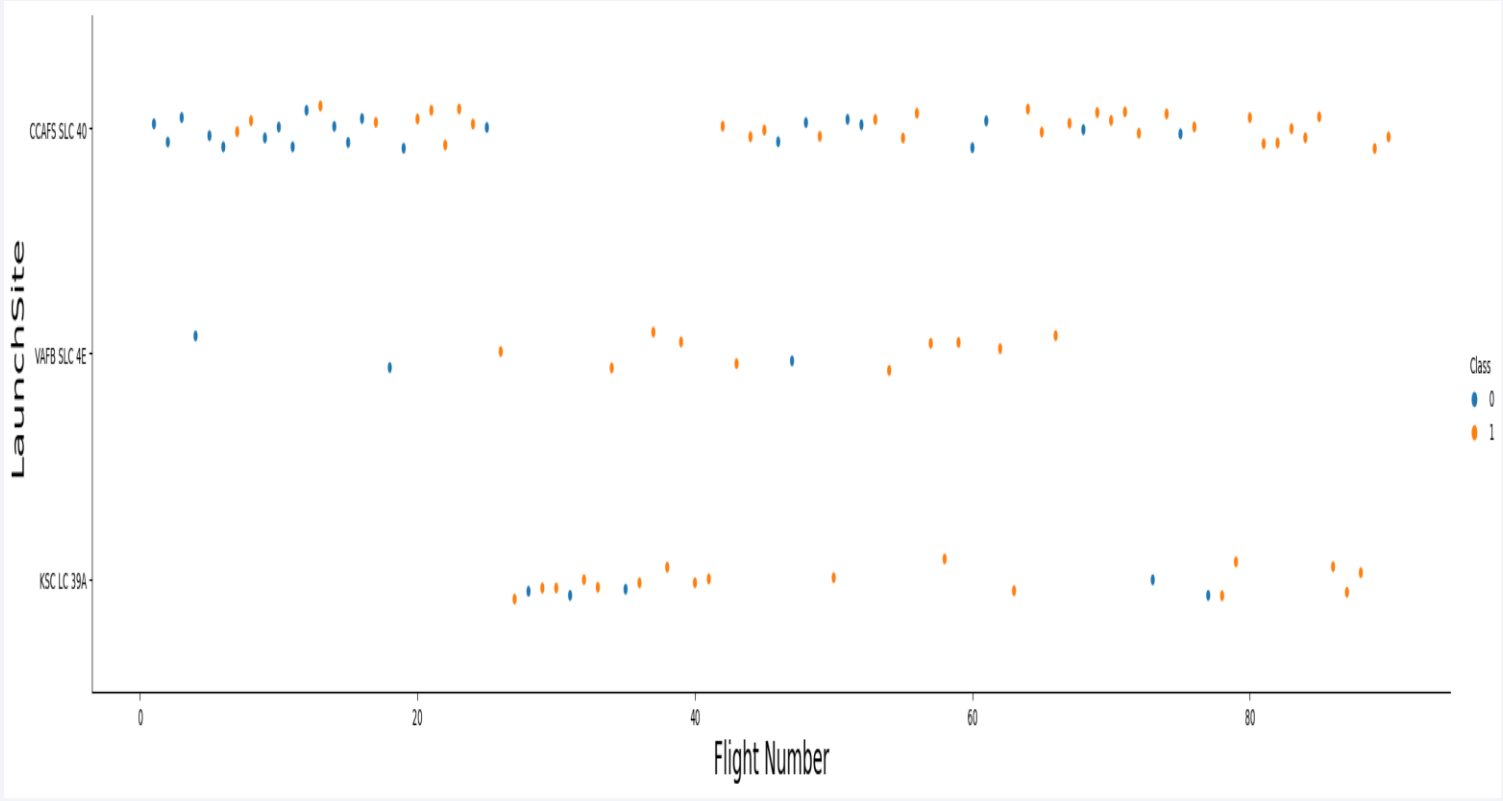
Results

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


Section 2
Insights drawn from EDA

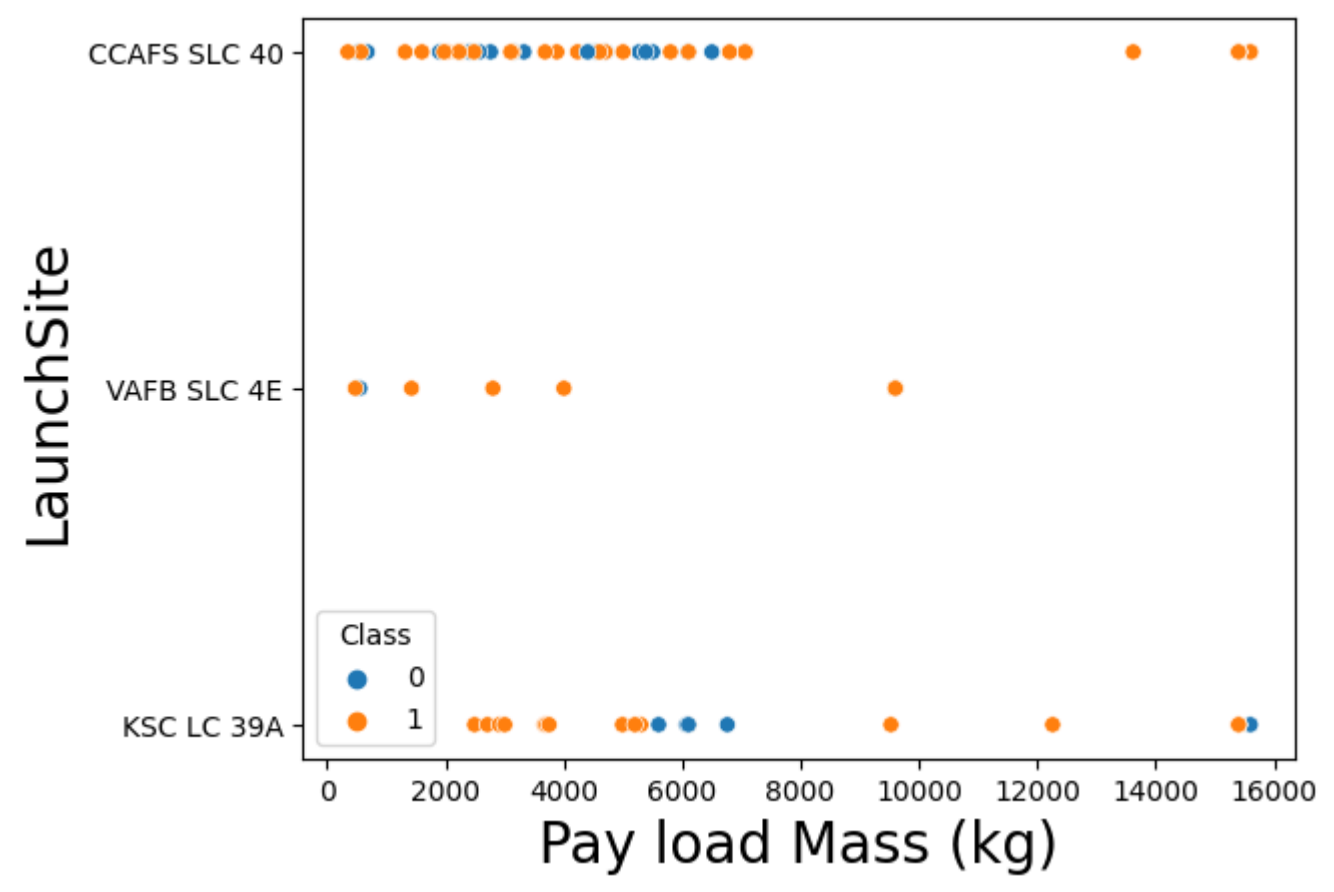
Flight Number vs. Launch Site

- The CCAFS SLC 40 launch site was the most common
- the VAFB SLV 4E and KSC LC 39A boast of a higher success rate as compared to CCAFS SLC 40



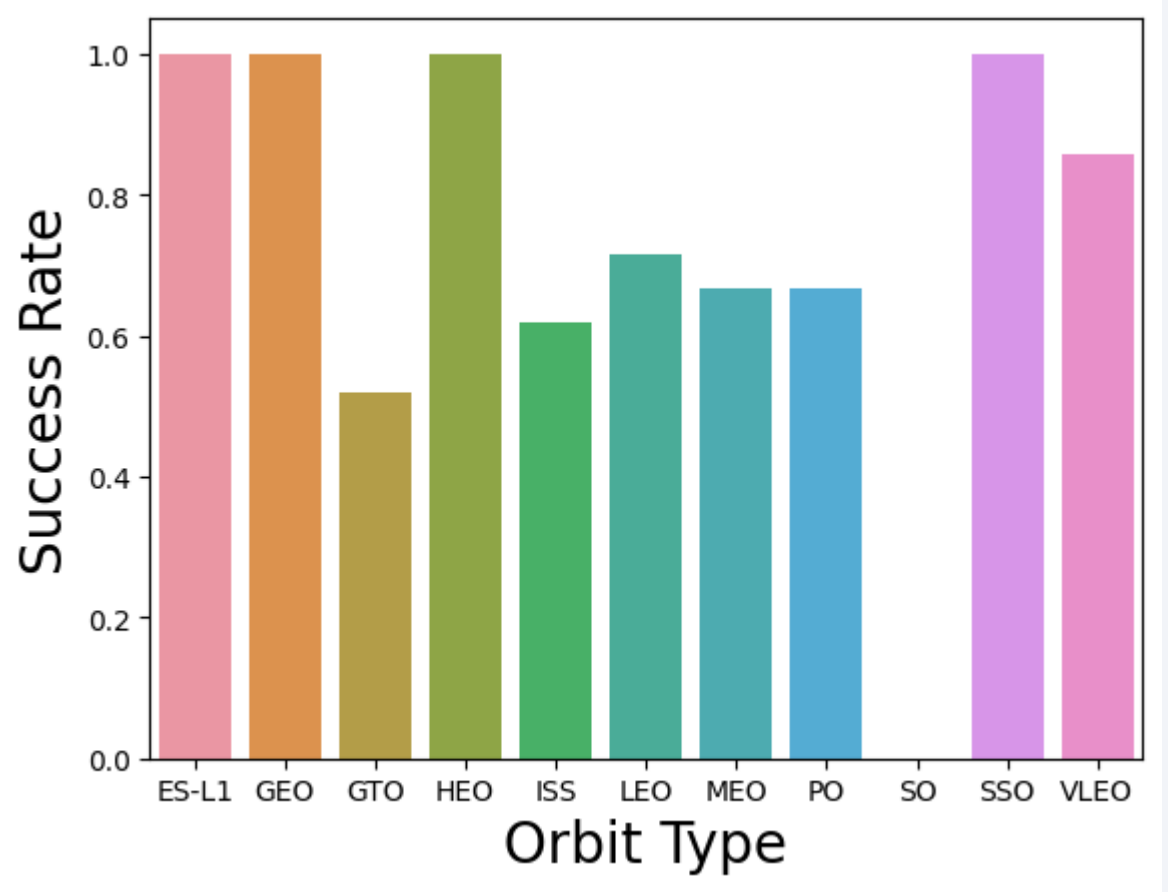
Payload vs. Launch Site

- Higher payload mass have higher success rates
- Launches with a payload mass under 7000kg are more common
- VAFB SLC 43 launch site there are no rocket launches with payload mass greater than 10000 kg.



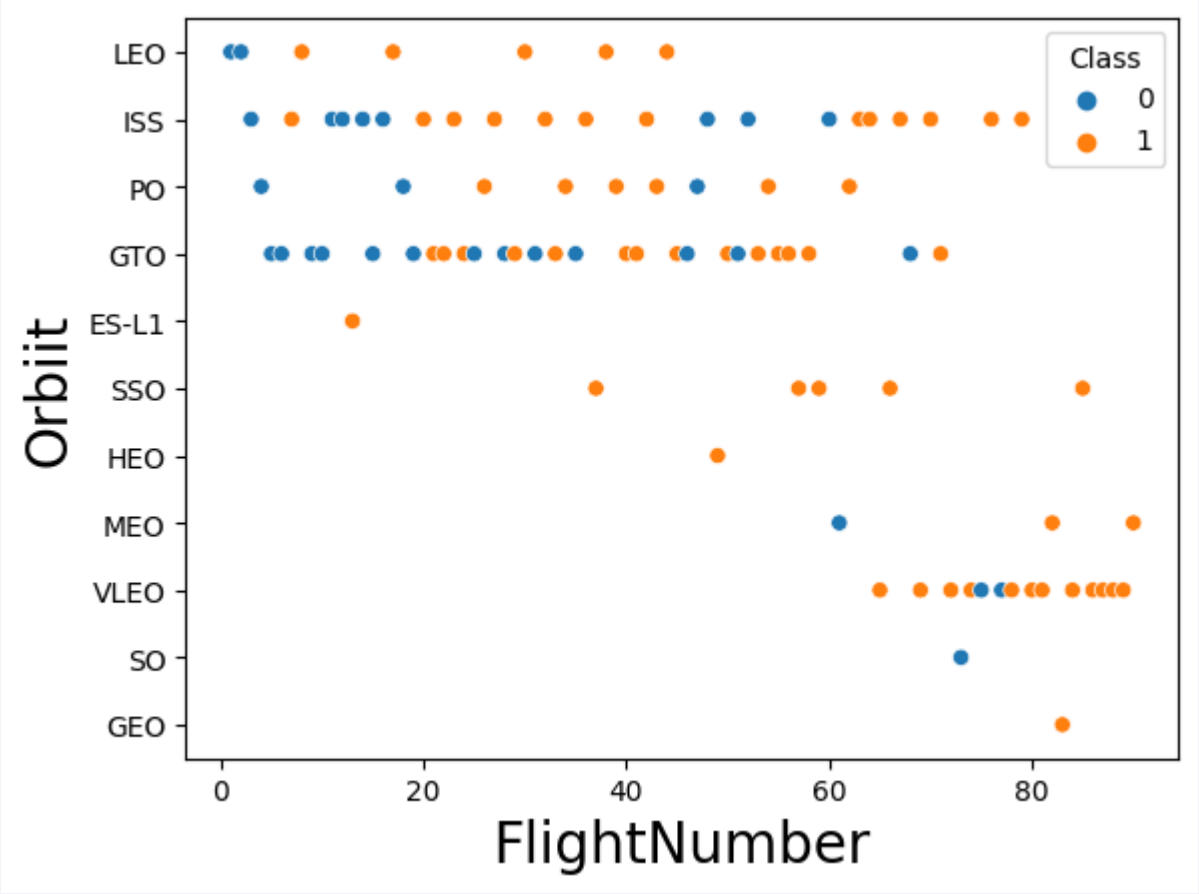
Success Rate vs. Orbit Type

- The ES-L1, GEO, HEO and SSO orbits all have 100% success rate
- The SO orbit has a 0% success rate



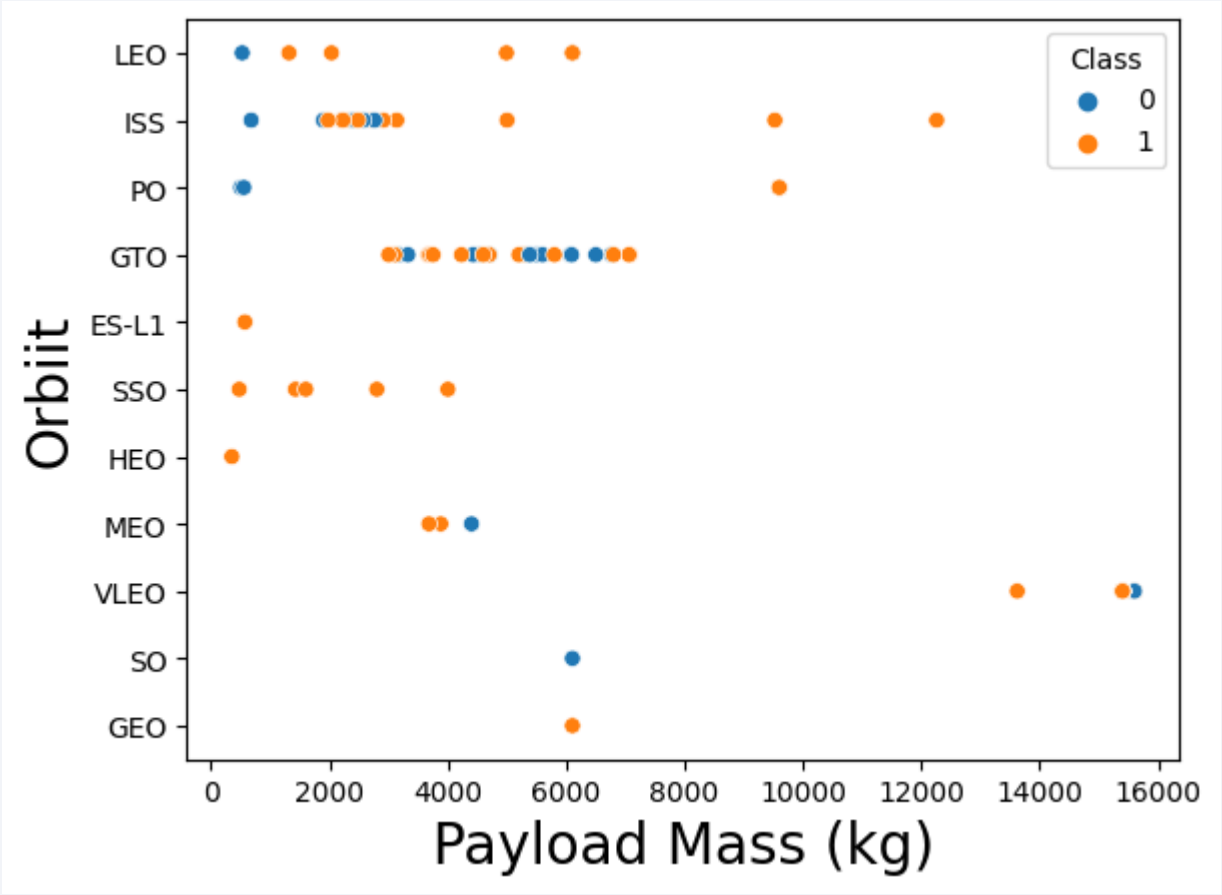
Flight Number vs. Orbit Type

- missions with orbit VLEO have a high success rate
- All mission with ES -L1, SSO, GEO were successful



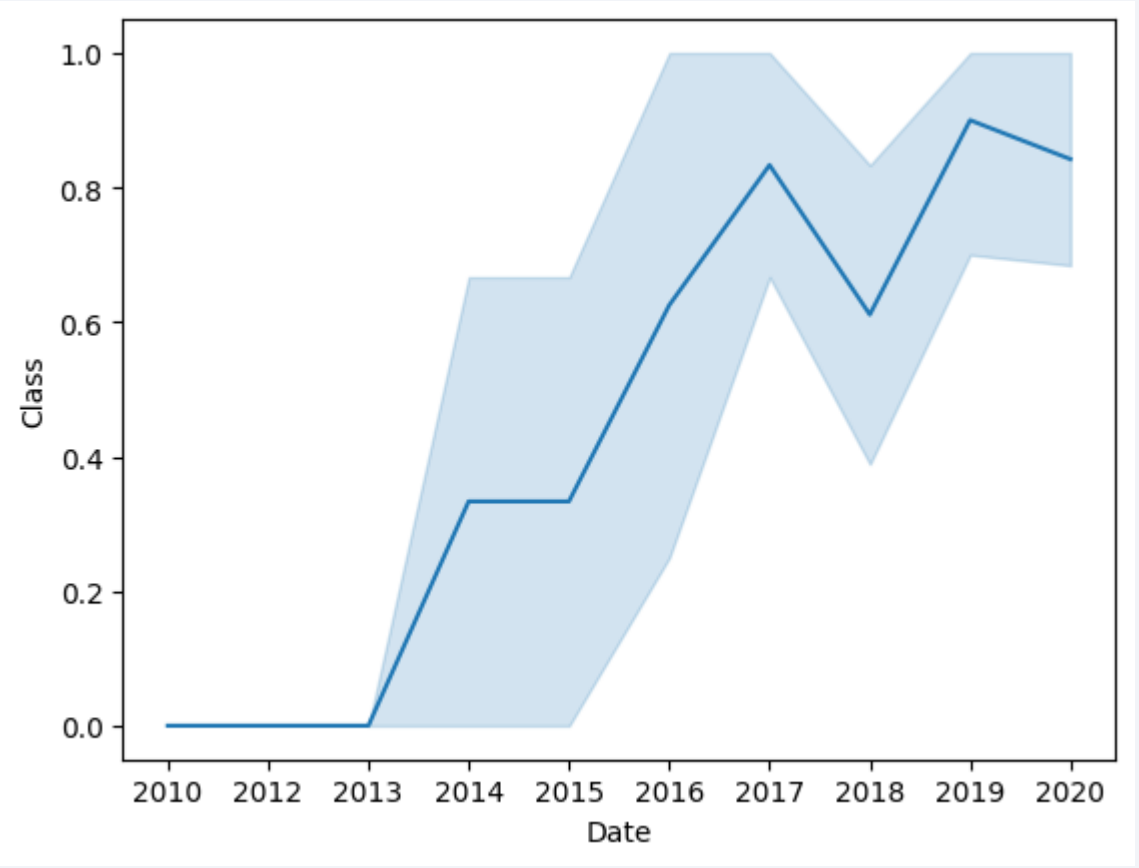
Payload vs. Orbit Type

- Only ISS, PO and VLEO have payloads over 7000kg
- High Payloads are used for missions with orbits VLEO, PO, ISS
- Higher Payload missions have relatively higher success rates



Launch Success Yearly Trend

- The Landing success has steadily increased over the years from 2013
- 2018 saw a dip in success rate
- The success rate has been above 50% since 2016



All Launch Site Names

- Names of Launch Sites has been displayed

Task 1

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'

- 5 records with launch sites begin with ‘CCA’ displayed

Task 2

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
2010-04-06	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

- The total payload mass carried by boosters from NASA has been calculated and presented

Task 3

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

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS total_mass FROM SPACEXTABLE
WHERE customer = 'NASA (CRS)';
```

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

total_mass
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 has been calculated and presented

Task 4

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

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS avg_mass FROM SPACEXTABLE
WHERE Booster_Version LIKE '%F9 v1.1%';
```

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

avg_mass
2534.6666666666665

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad has been extracted

Task 5

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

Hint:Use min function

```
%%sql
SELECT MIN(Date) AS 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

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 has been extracted and presented

Task 6

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 DISTINCT Booster_Version FROM SPACEXTABLE
WHERE (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000) & (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 and their count have been summarized and presented

Task 7

List the total number of successful and failure mission outcomes

```
%%sql
SELECT Mission_Outcome, COUNT(*) FROM SPACEXTABLE
GROUP BY Mission_Outcome;
```

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

Mission_Outcome	COUNT(*)
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 have been extracted

Task 8

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

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

```
 * 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

- The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 have been extracted and presented

Task 9

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, Booster_Version, Launch_Site FROM SPACEXTABLE
WHERE (substr(Date,0,5) = '2015') & (Landing_Outcome = 'Failure (drone ship)');
```

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

month	Booster_Version	Launch_Site
10	F9 v1.1 B1012	CCAFS LC-40
04	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 have been ranked, in descending order and presented

Task 10

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 counts FROM SPACEXTABLE
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY counts DESC;
```

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

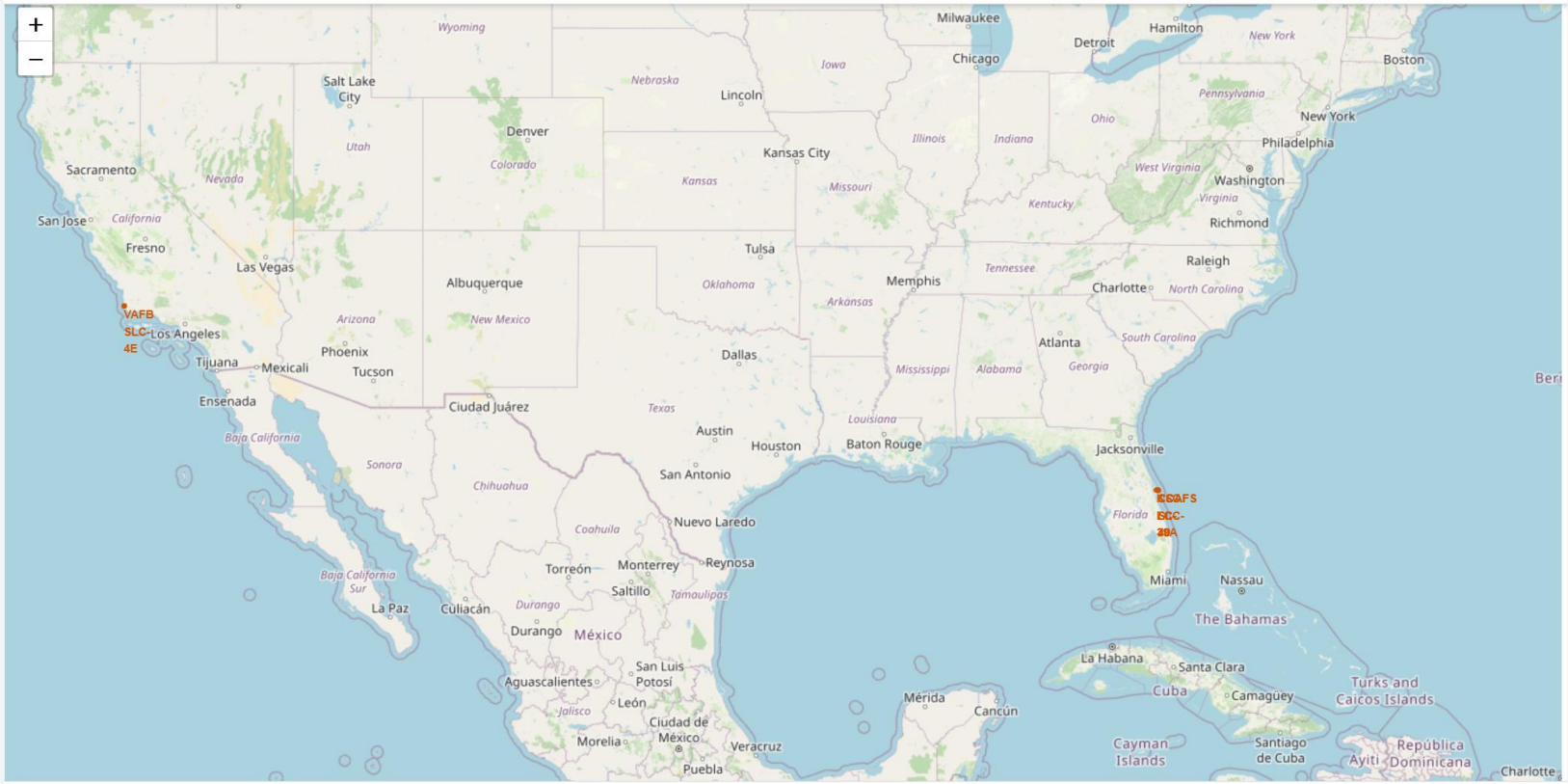
Landing_Outcome	counts
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

Section 3

Launch Sites Proximities Analysis

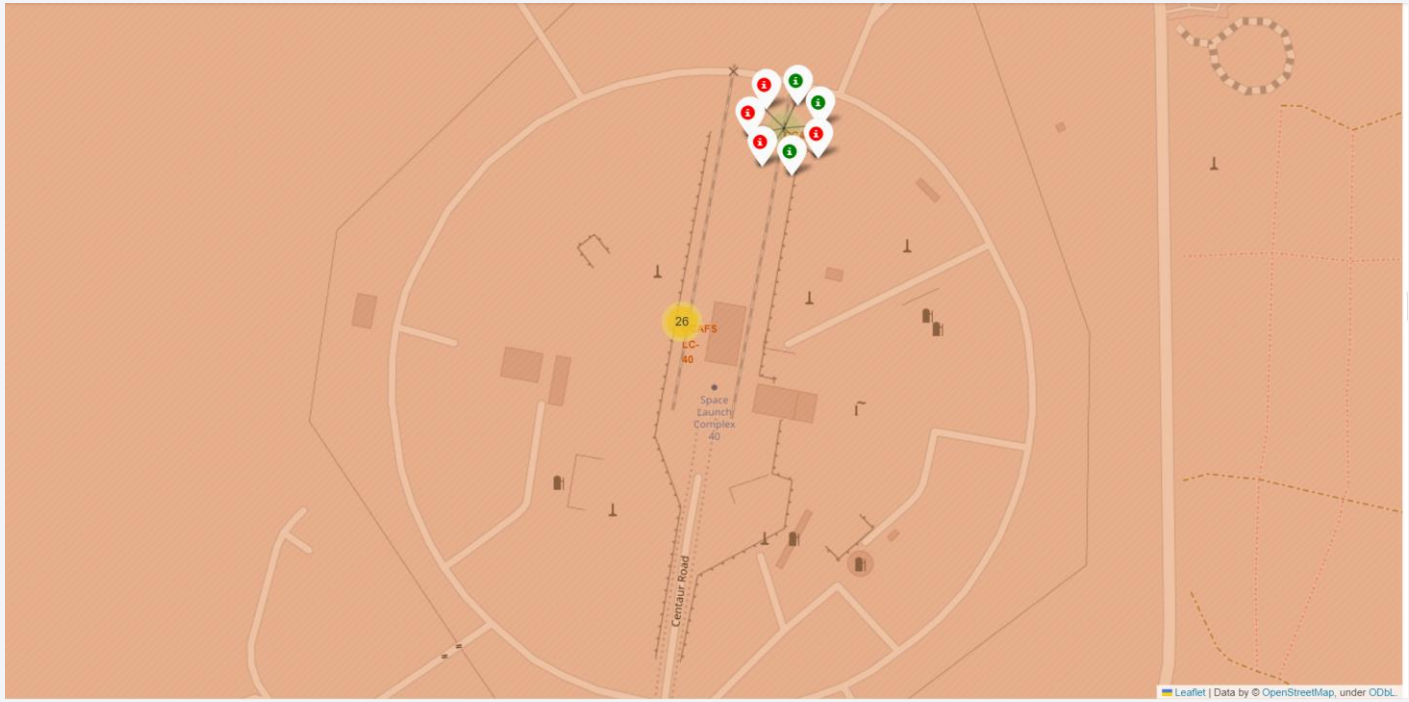
Launch Site Locations

- The launch sites are all on the coast of the USA, specifically in the South East (Florida) and South West (California)



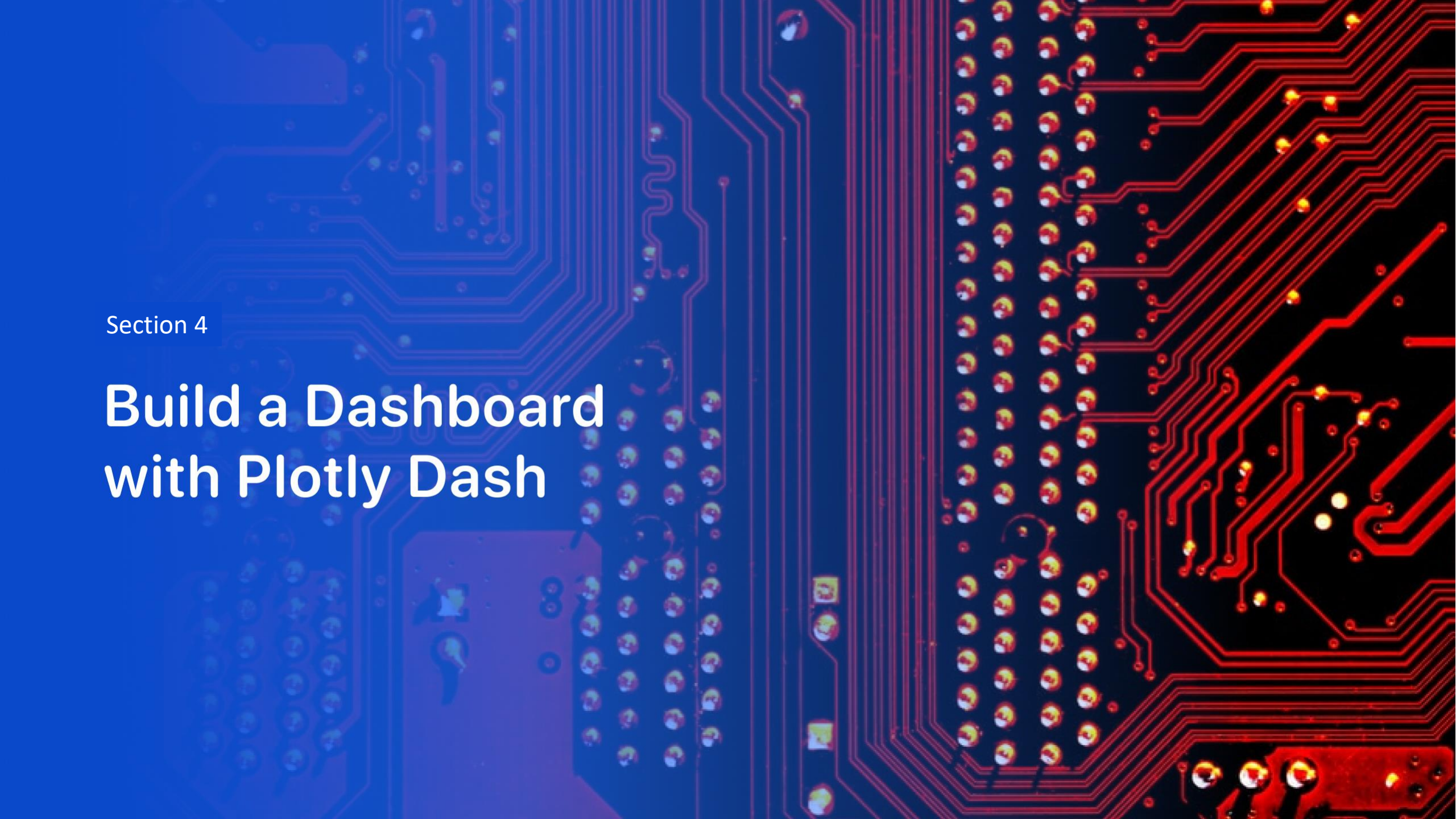
Successful and Failed Launches

- Successful launches are colored green and failed launches are colored red



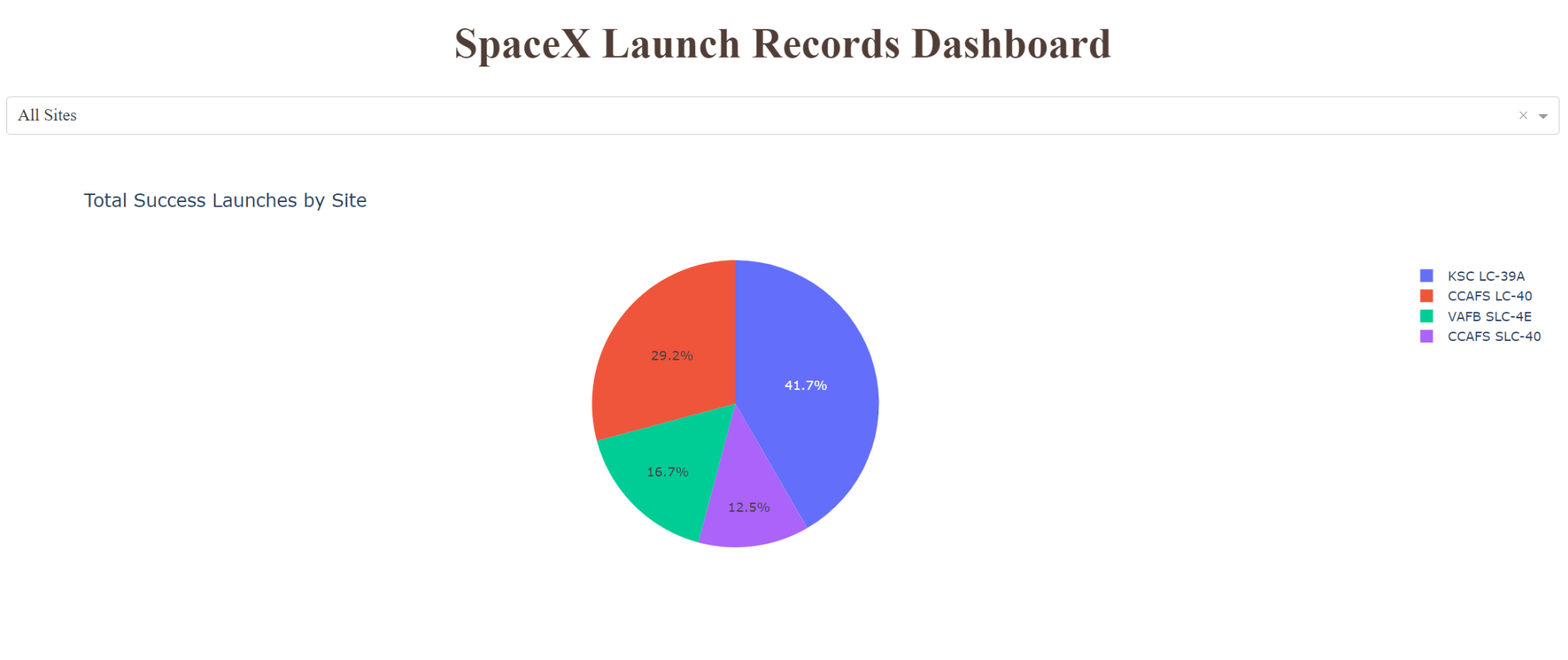
Launch Site Proximity to Points of Interest

- The coast is 0.86km to the East
- A railway is nearby at 1km away
- A road is nearby at 0.59km away
- The nearest major city of Orlando is further at almost 80km


Section 4
Build a Dashboard with Plotly Dash

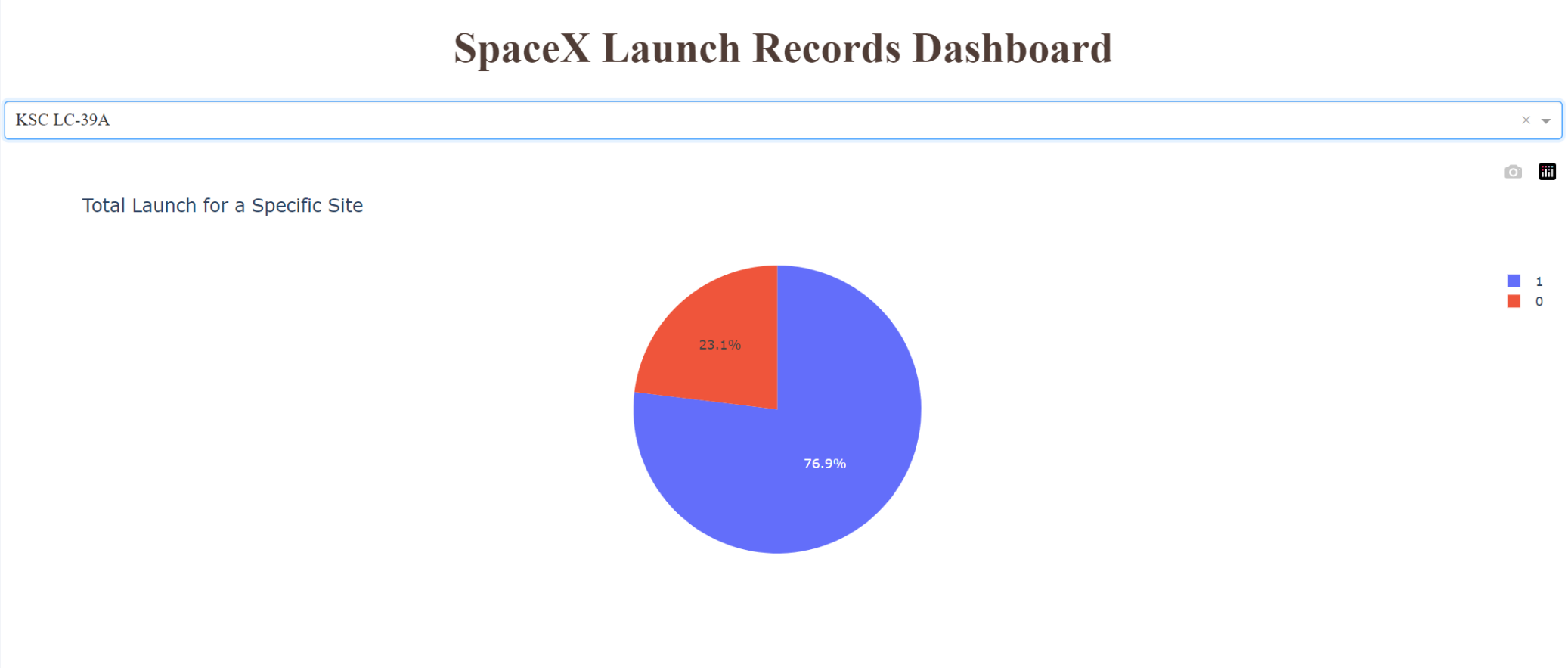
<Dashboard Screenshot 1>

- The launch site with more successful launches was KSC LC 39A



<Dashboard Screenshot 2>

- The KSC LC 39A launch site had 76.9% of successful launches



<Dashboard Screenshot 3>

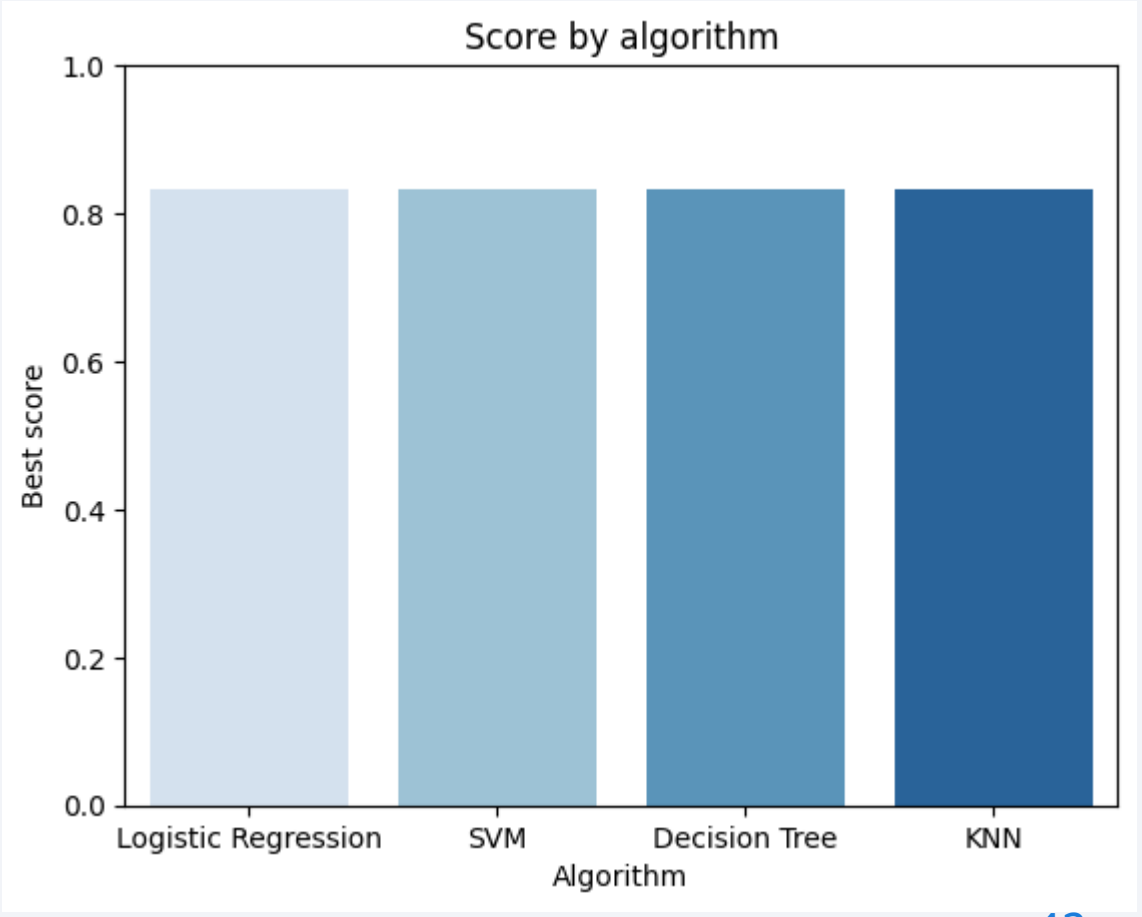
- Booster Version FT showed the highest number of successful launches, on the other hand, version v1.1 showed the highest number of failed launches

Section 5

Predictive Analysis (Classification)

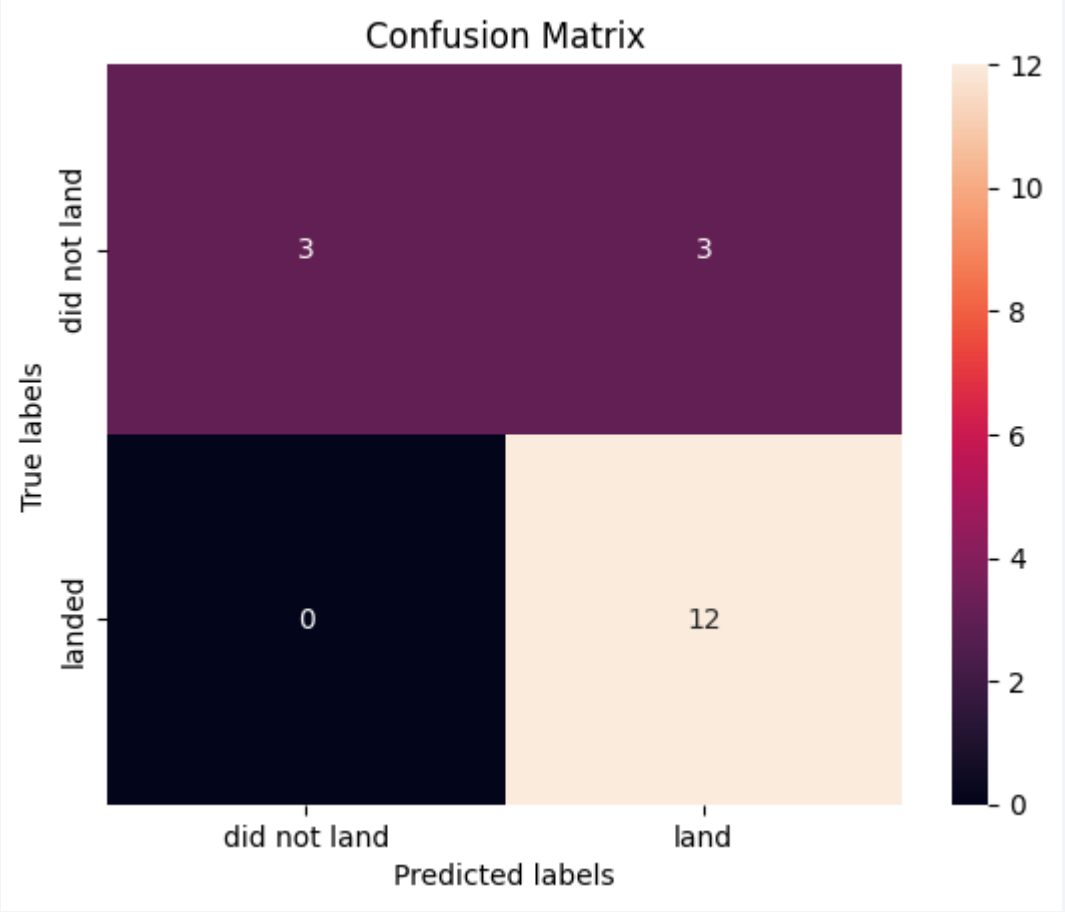
Classification Accuracy

- All models have the same accuracy



Confusion Matrix

- It correctly predicted 15 of the 18 values
- It incorrectly predicted that three results would land when the correct label was a failure



Conclusions

- The Landing success has steadily increased over the years from 2013
- Heavier payloads have fewer data points but a higher success rate
- Launch sites are located on the coast, near to transportation infrastructure (roads and railways) but further from major cities
- The classification models had the best performance with an accuracy of 83.3% on the test set

Appendix

https://github.com/Jackch56/IBM_Data_Science_Professional_Certification/tree/main/Data-Science-Capstone-Project

This link shows the repository with the completed notebooks for this project

Thank you!