IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Conor McDonnell
29 November 2024

Outline

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

Executive Summary

- Summary of methodologies:
  - Data Collection of Rocket Flight Information and Metrics
  - Wrangling of Rocket Flight Data
  - EDA with Data Visualisation
  - EDA with SQL
  - Interactive Launch Map with Folium
  - Active Data Dashboard with Dash
  - Machine Learning Predictive Analysis
- Summary of all results:
  - Descriptive Statistics and Graphs
  - Predictions of Future Flight Outcomes

Introduction

- Project background and context
  - SpaceY aims to compete with SpaceX, an industry leading rocket flight company that is able to launch rockets much more cheaply than competitors by reusing the first stage of their rockets. With a flight costing competitors around $165m, SpaceX flights cost only $62m when the first stage can be reused after a successful landing.
- Problems you want to find answers
  - What data is available to predict successful first stage landings?
  - How can we use this data to predict when a successful landing will occur?
  - If we are able to predict successful landings, can we compete with SpaceX in terms of cost?

Section 1

Methodology

Methodology

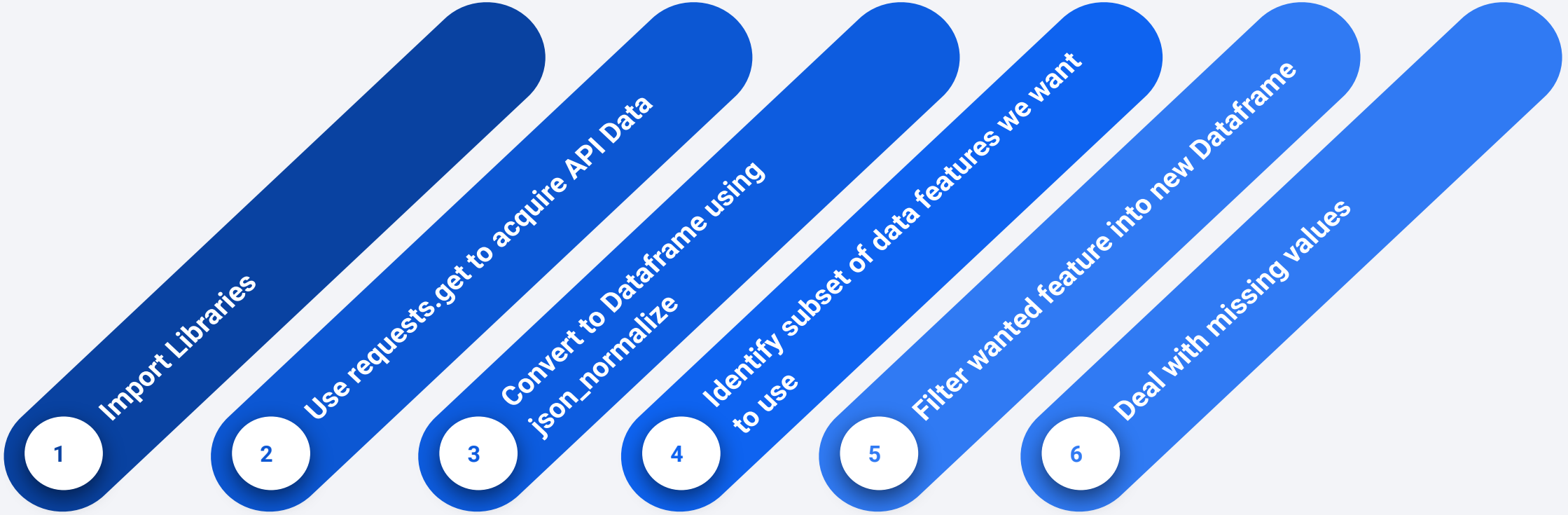
Executive Summary

- Data collection methodology:
  - API Requests and Webscraping
- Perform data wrangling
  - Data was processed using standard Data Science techniques in Python and it's open source libraries, including Numpy and Pandas
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Scikit-Learn

Data Collection

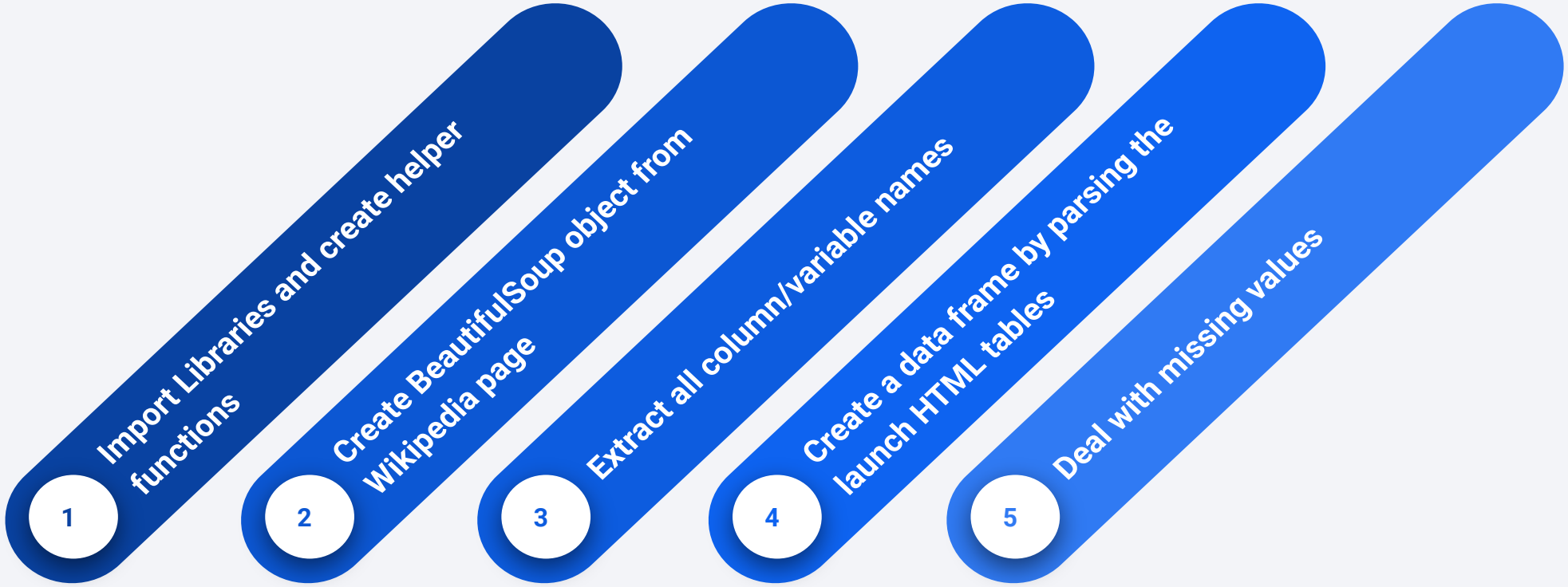
- SpaceX flight data was acquired by requesting data through their own Rest API and by webscraping their wikipedia page
- API:
  - Get response from API
  - Convert response to a JSON file
  - Clean the data with Numpy and Pandas
  - Assign list to dictionary then create a Dataframe
  - Filter the Dataframe
- Wiki:
  - Get response from HTML
  - Create beautifulsoup object
  - Find tables in soup object
  - Get columns names and create dictionary
  - Append data to keys
  - Convert dictionary to a Dataframe

Data Collection – SpaceX API



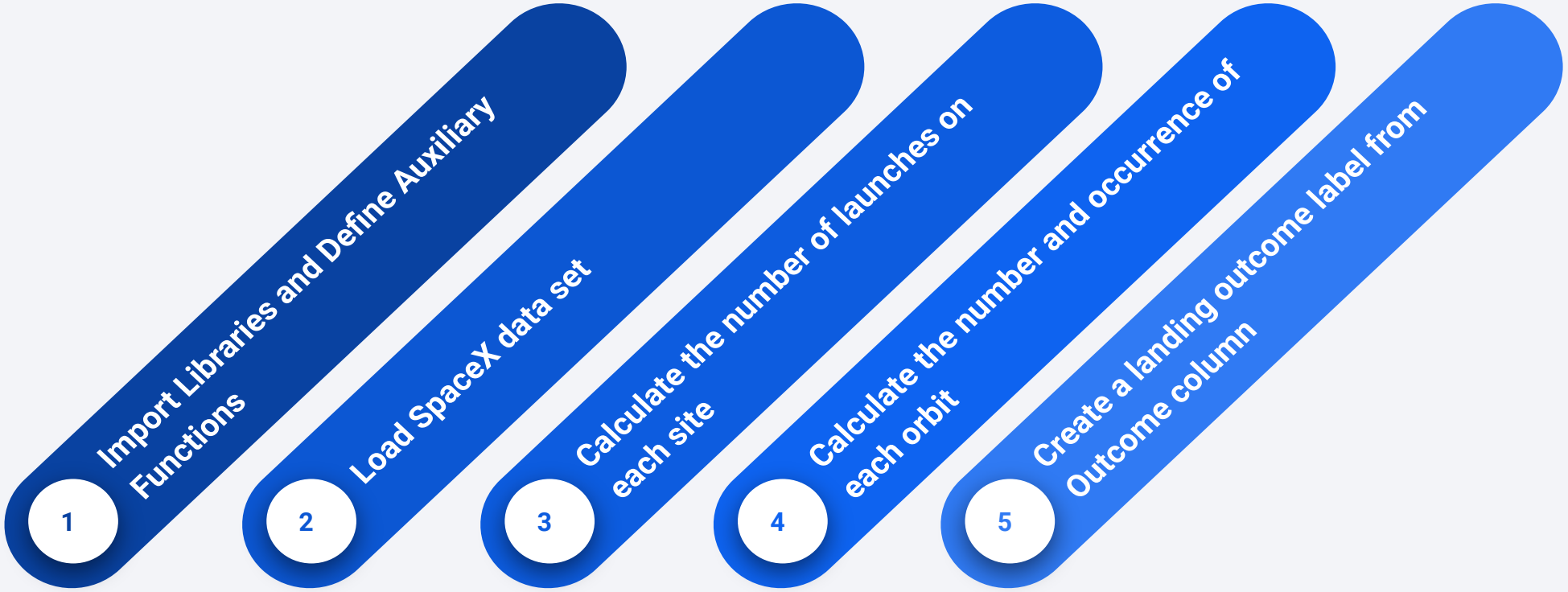
- https://github.com/cmcd17/CourseraDataScience/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping



- https://github.com/cmcd17/CourseraDataScience/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling



- Exploratory data analysis and training labels were created by using variable counts and creating categorical variables for successful or failed landing outcomes
- https://github.com/cmcd17/CourseraDataScience/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Visualize the relationship between Flight Number and Launch Site - Scatter
- Visualize the relationship between Payload Mass and Launch Site - Scatter
- Visualize the relationship between success rate of each orbit type - Bar
- Visualize the relationship between FlightNumber and Orbit type - Scatter
- Visualize the relationship between Payload Mass and Orbit type - Scatter
- Visualize the launch success yearly trend - Line

https://github.com/cmcd17/CourseraDataScience/blob/main/edadataviz%20(1).ipynb

EDA with SQL

- 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. 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

- https://github.com/cmcd17/CourseraDataScience/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Location of launch sites were labelled on the map with circle markers and their name
- Zooming in to each launch site allows you to see success (green) and failure (red) icons, this was done with MarkerCluster
- Using other location data we calculated the distance from a particular launch site to nearby landmarks, such as the coast line, highway, and airport, these lines are shown in blue when zoomed in to the launch site

https://github.com/cmcd17/CourseraDataScience/blob/main/lab_jupyter_launch_site_location%20(1).ipynb

Build a Dashboard with Plotly Dash

- Add a Launch Site Drop-down Input Component
- Add a callback function to render success-pie-chart based on selected site dropdown
- Add a Range Slider to Select Payload
- visually observe how payload may be correlated with mission outcomes for selected site(s)

Which site has the largest successful launches?

Which site has the highest launch success rate?

Which payload range(s) has the highest launch success rate?

Which payload range(s) has the lowest launch success rate?

Which F9 Booster version (v1.0, v1.1, FT, B4, B5, etc.) has the highest launch success rate?

Predictive Analysis (Classification)

- Build Model
  - Model is built from data that has been organized and cleaned with Numpy and Pandas
  - Data is split into training and test sets
  - We will try multiple models, including: decision tree, SVM, and KNN
  - Datasets will be trained with GridSearchCV
- Evaluate Model
  - CHeck accuracy for each model
  - Create confusion matrices for each model
  - Calculate hyperparameters for each model
- Improve Model
  - Select the best features
  - Algorithm tuning
- Picking the Best Model
  - We will pick the model with the highest accuracy and best performance

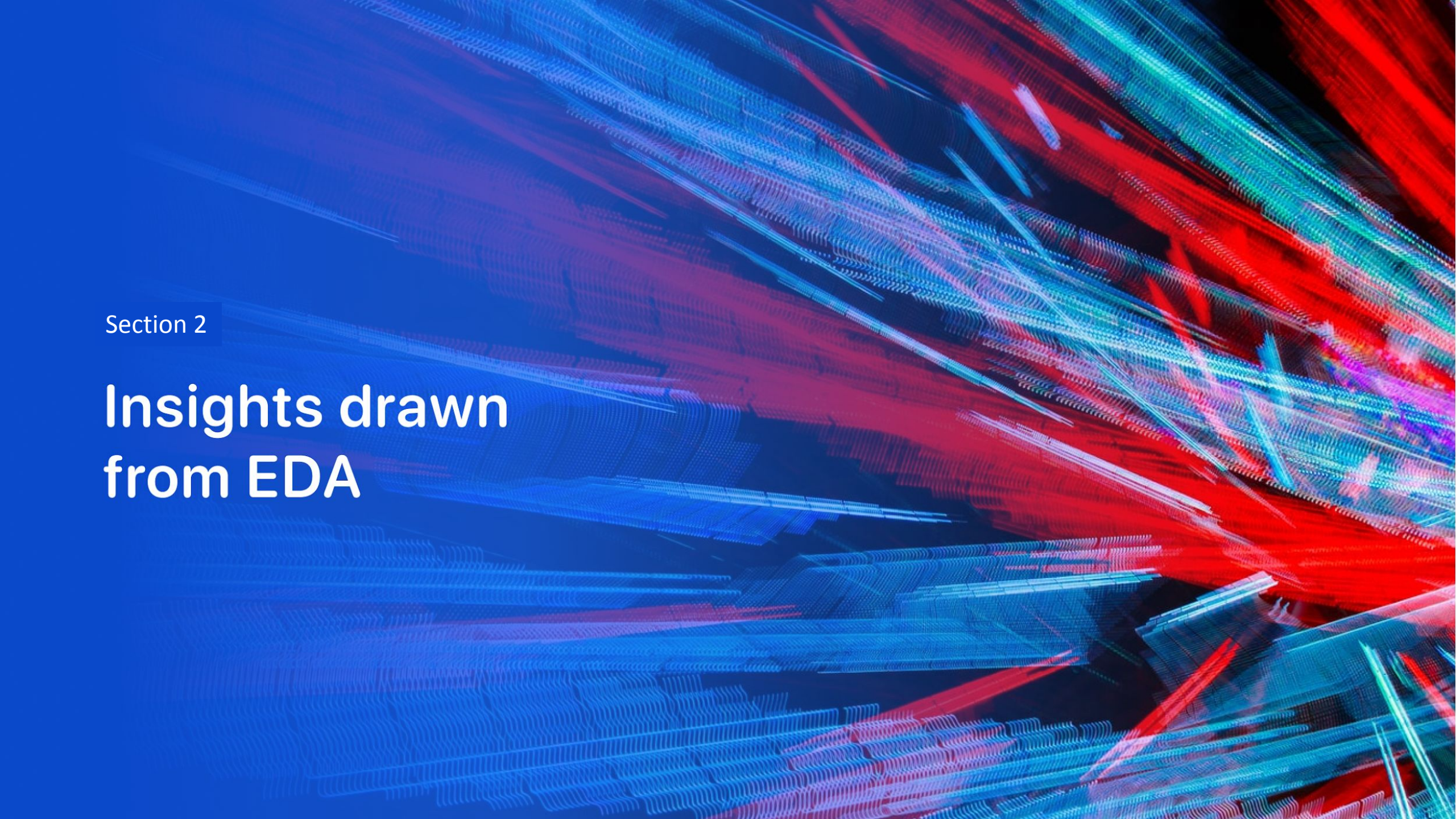
https://github.com/cmcd17/CourseraDataScience/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5%20(1).ipynb

Results

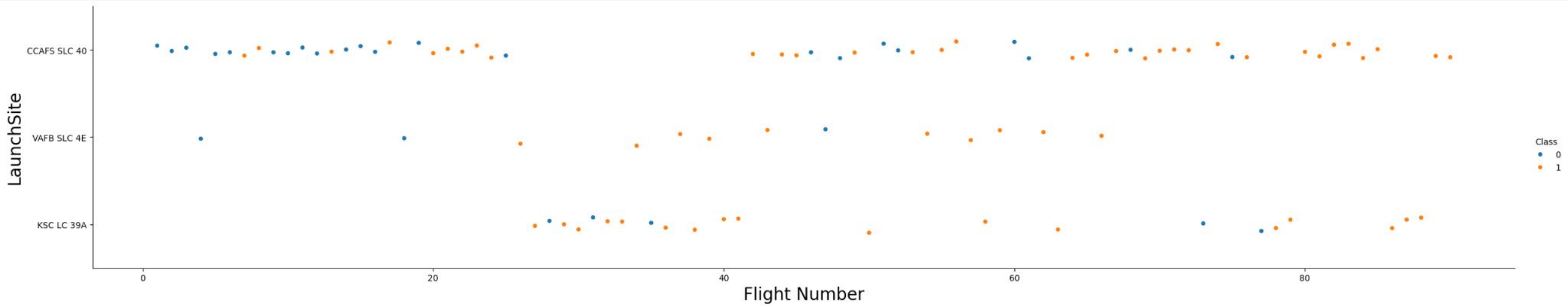
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

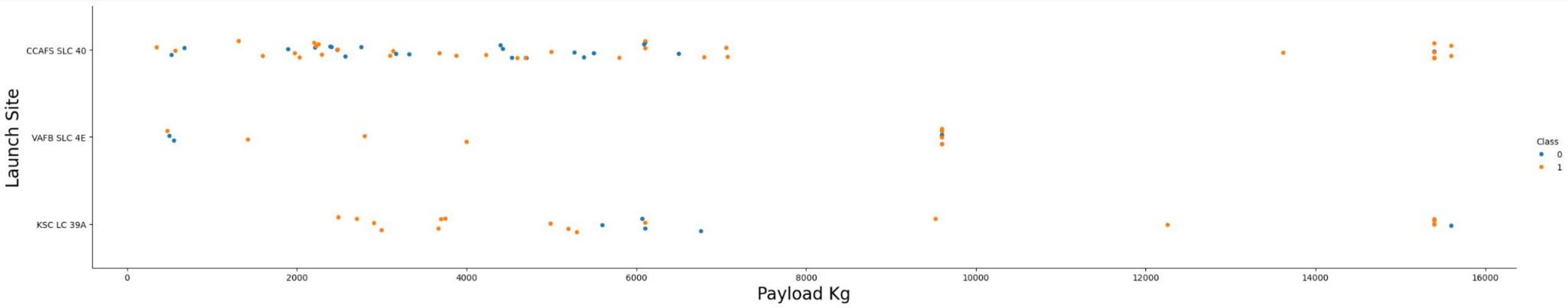


Flight Number vs. Launch Site



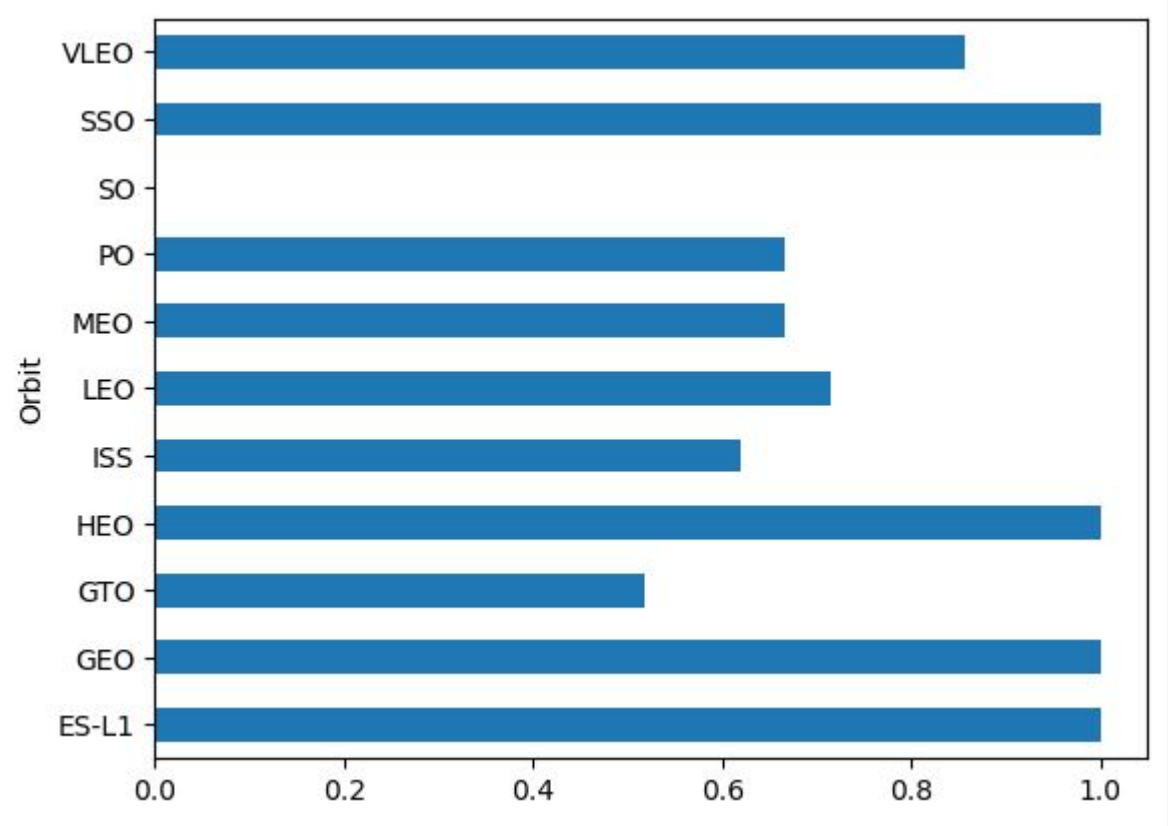
- The majority of flights use the CCAFS SLC 40 launch site
- There is no clear pattern of successful flights based on launchsite using visual inspection

Payload vs. Launch Site



- The CCAFS SLC 40 launch site is used for the majority of low payload flights
- The VAFB SLC 4E launch site is not used for payloads greater than 10000 Kg

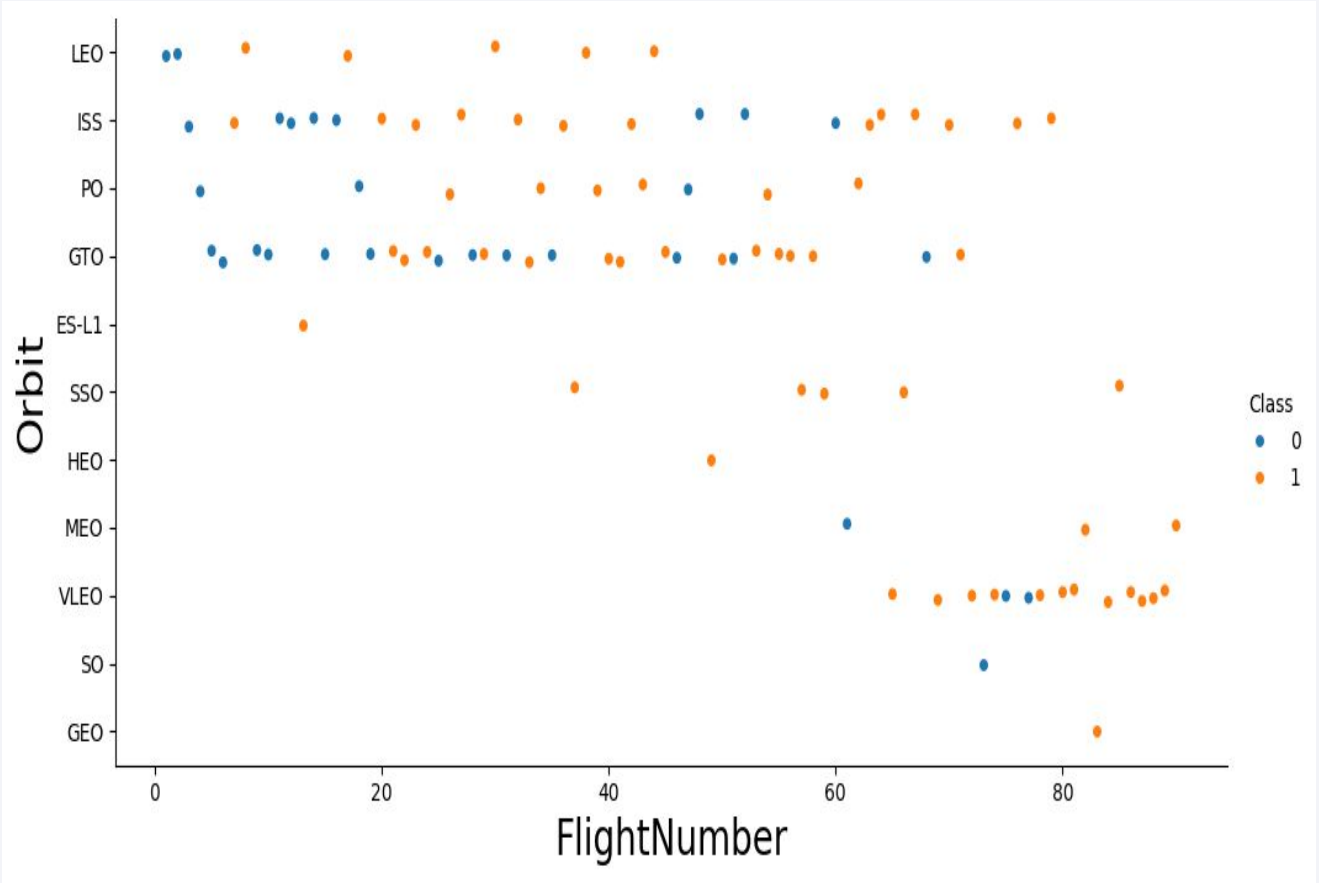
Success Rate vs. Orbit Type



- SO orbit has no successful flights
- SSO, HEO, GEO, and ES-L1 orbits have a 100% success rate
- Most launch sites have a success rate between 50 and 80%

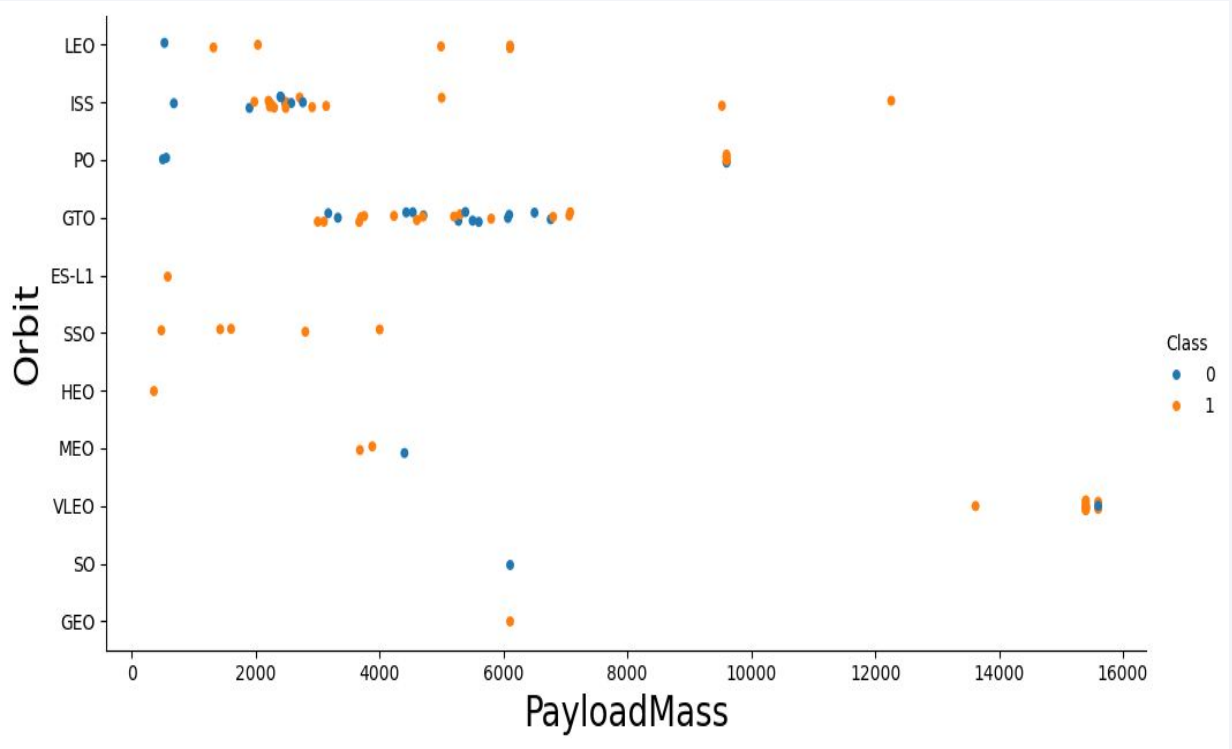
Flight Number vs. Orbit Type

- SO orbit has no successful flights but has only been attempted once
- GEO orbit has had one successful flight from one attempt
- Early flights numbers were all failures



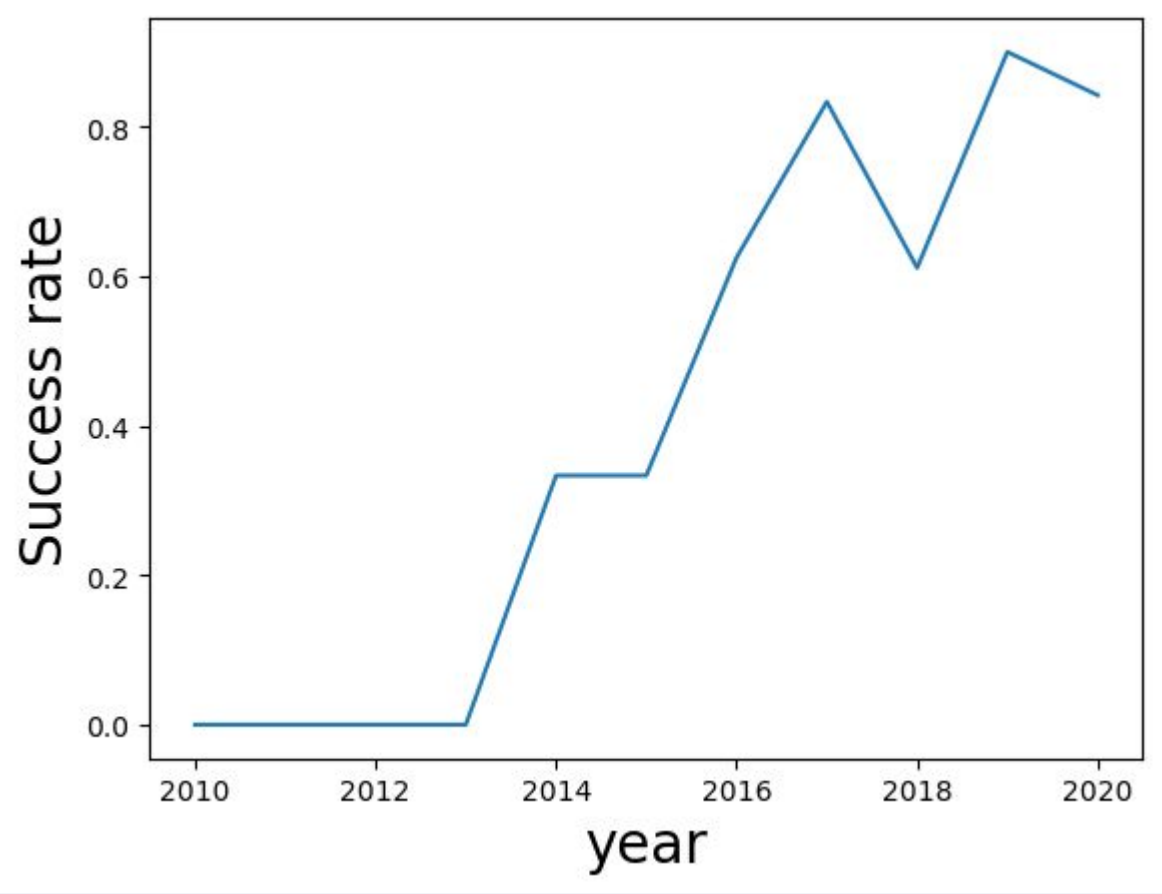
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS
- However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present



Launch Success Yearly Trend

- Success rate was 0% from 2010 to 2013
- Success rate rose sharply starting from 2013 to 2017 and from 2018 to 2019
- Success rate saw dips from 2017 to 2018 and from 2019 to 2020



All Launch Site Names

- There are 4 launch sites

In [30]:
```
%sql select Distinct(Launch_Site) from SPACEXTABLE;
```

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

Out[30]:

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

Launch Site Names Begin with 'CCA'

- The 5 records contain information for launch site CCAFS LC-40
- Data includes Date and Time, Booster version,Payload, Payload Mass, Orbit, Mission Outcome, etc.

In [13]:
```
%sql SELECT * \
    FROM SPACEXTABLE \
    WHERE LAUNCH_SITE LIKE'CCA%' LIMIT 5;
```

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

Out[13]:

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

- The sum total payload mass carried by boosters launched by NASA (CRS) is: 45596 Kg

```
In [16]: %sql SELECT SUM(PAYLOAD_MASS__KG_) \
             FROM SPACEXTBL \
             WHERE CUSTOMER = 'NASA (CRS)';
 * sqlite:///my_data1.db
Done.
Out[16]: SUM(PAYLOAD_MASS__KG_)
         45596
```

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is: 2928.4 Kg

```
In [17]: %sql SELECT AVG(PAYLOAD_MASS__KG_) \
             FROM SPACEXTBL \
             WHERE BOOSTER_VERSION = 'F9 v1.1';
 * sqlite:///my_data1.db
Done.
Out[17]: AVG(PAYLOAD_MASS__KG_)
         2928.4
```

First Successful Ground Landing Date

- The date of the first successful ground landing was December 22 2015

In [31]:
```
%sql SELECT MIN(DATE) \
FROM SPACEXTBL \
WHERE LANDING_OUTCOME = 'Success (ground pad)'
```

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

Out[31]:

MIN(DATE)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- the boosters which have success in drone ship and have 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

In [33]:
```
%sql select BOOSTER_VERSION from SPACEXTBL where LANDING_OUTCOME="Success (drone ship)" and PAYLOAD_MASS__KG_ BETWEEN 4000 a
```

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

Out[33]:

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- There have been 100 total successful F9 missions
- There has been 1 failure in flight for F9 missions

In [21]:
```
%sql SELECT MISSION_OUTCOME, COUNT(*) as total_number \
FROM SPACEXTBL \
GROUP BY MISSION_OUTCOME;
```

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

Out[21]:

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

Boosters Carried Maximum Payload

- The list in the image contains all the booster versions that have carried the maximum payload mass

In [22]:
```
%sql SELECT BOOSTER_VERSION \
FROM SPACEXTBL \
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

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

Out[22]:

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

- Two failures are recorded for failure in landing outcome drone ship
- Both missions used the same launch site but had different booster versions

In [36]:
```
%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.
```

Out[36]:

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

- Counts for success and failure types are shown in this image in descending order with no attempt have the most outcomes and precluded (drone ship) having the least outcomes for the specified date range

```
[21]: %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_outcomes DESC;
```

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

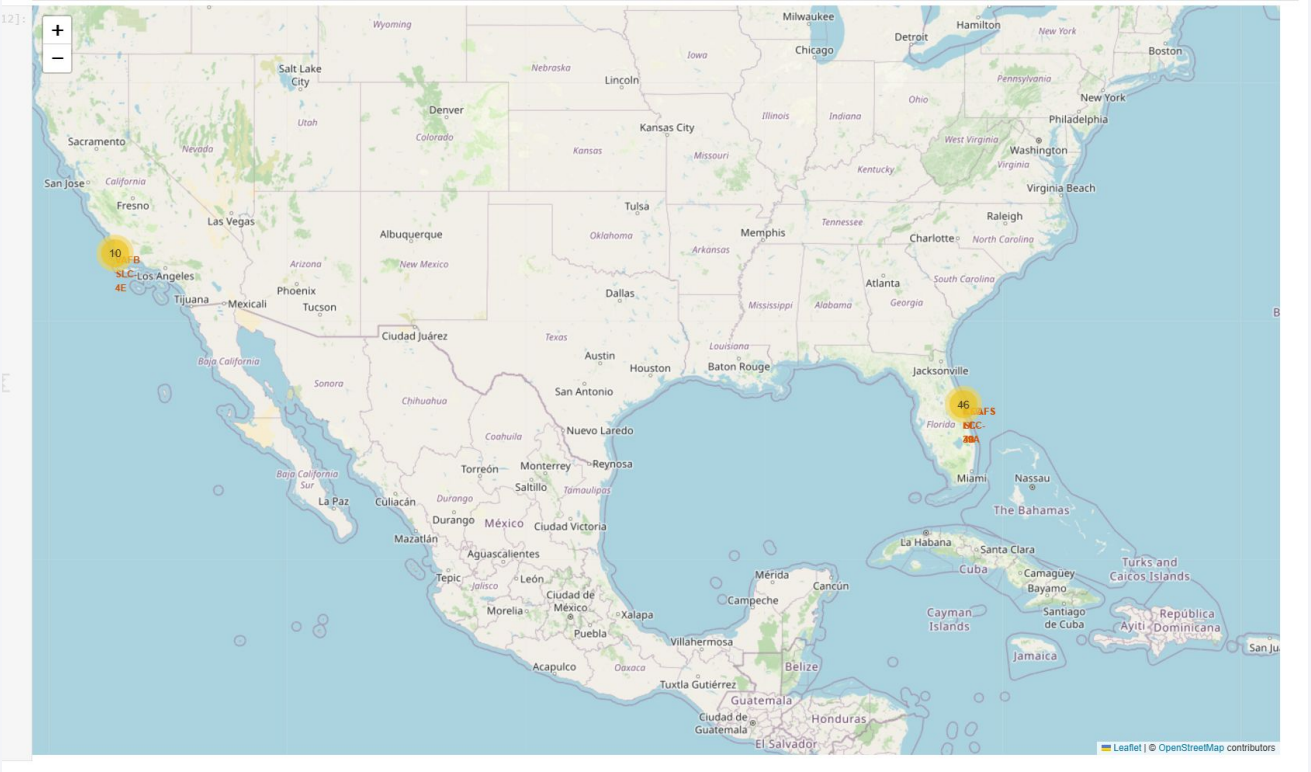
[21]:

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

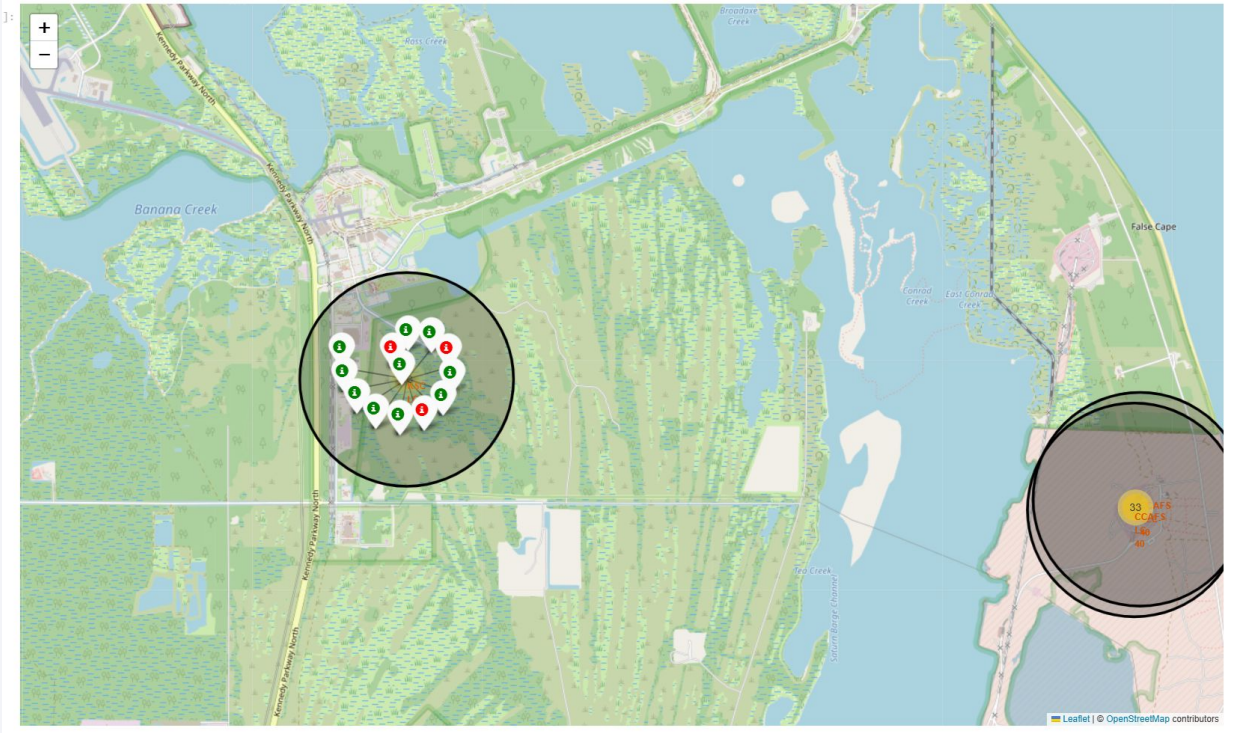
<Folium Map Screenshot 1>

- Map shows launch sites on the east and west coasts of the United States
- Zooming in allows you to see better detail between sites and the number of success and fails of each site



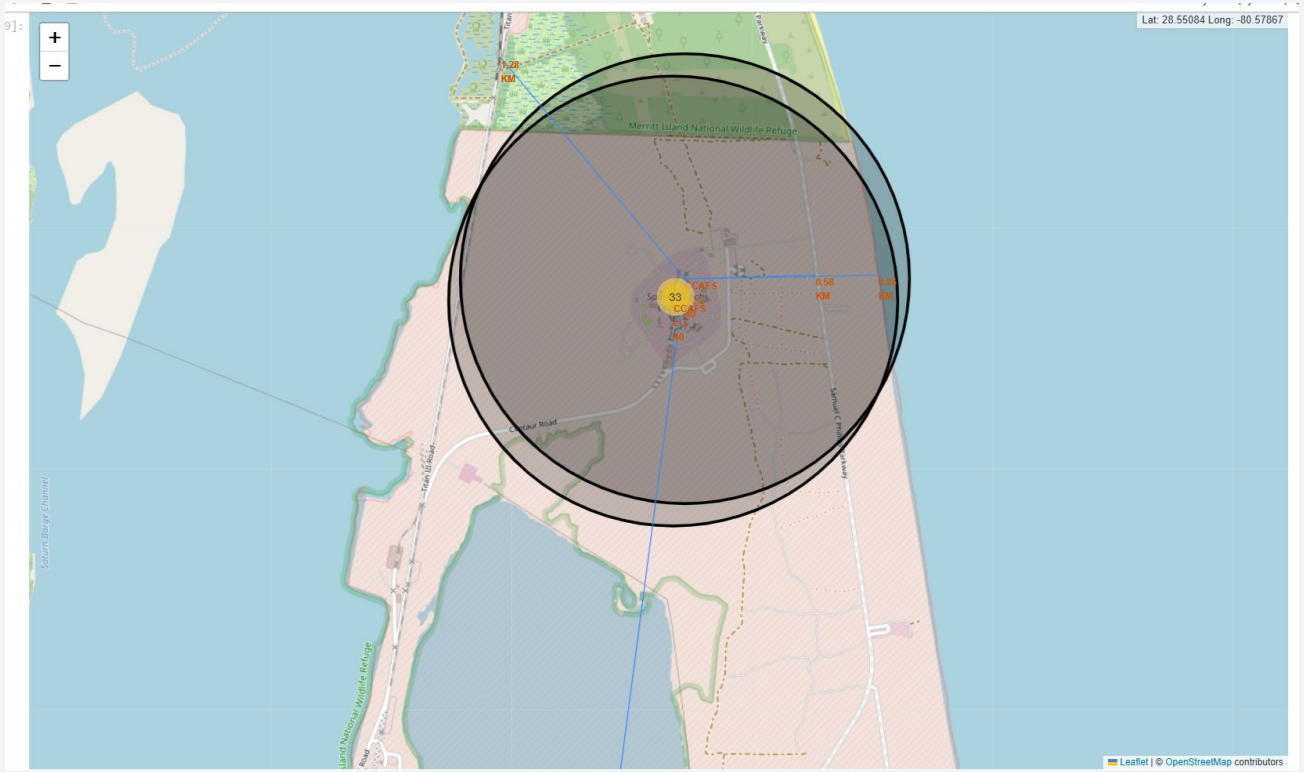
<Folium Map Screenshot 2>

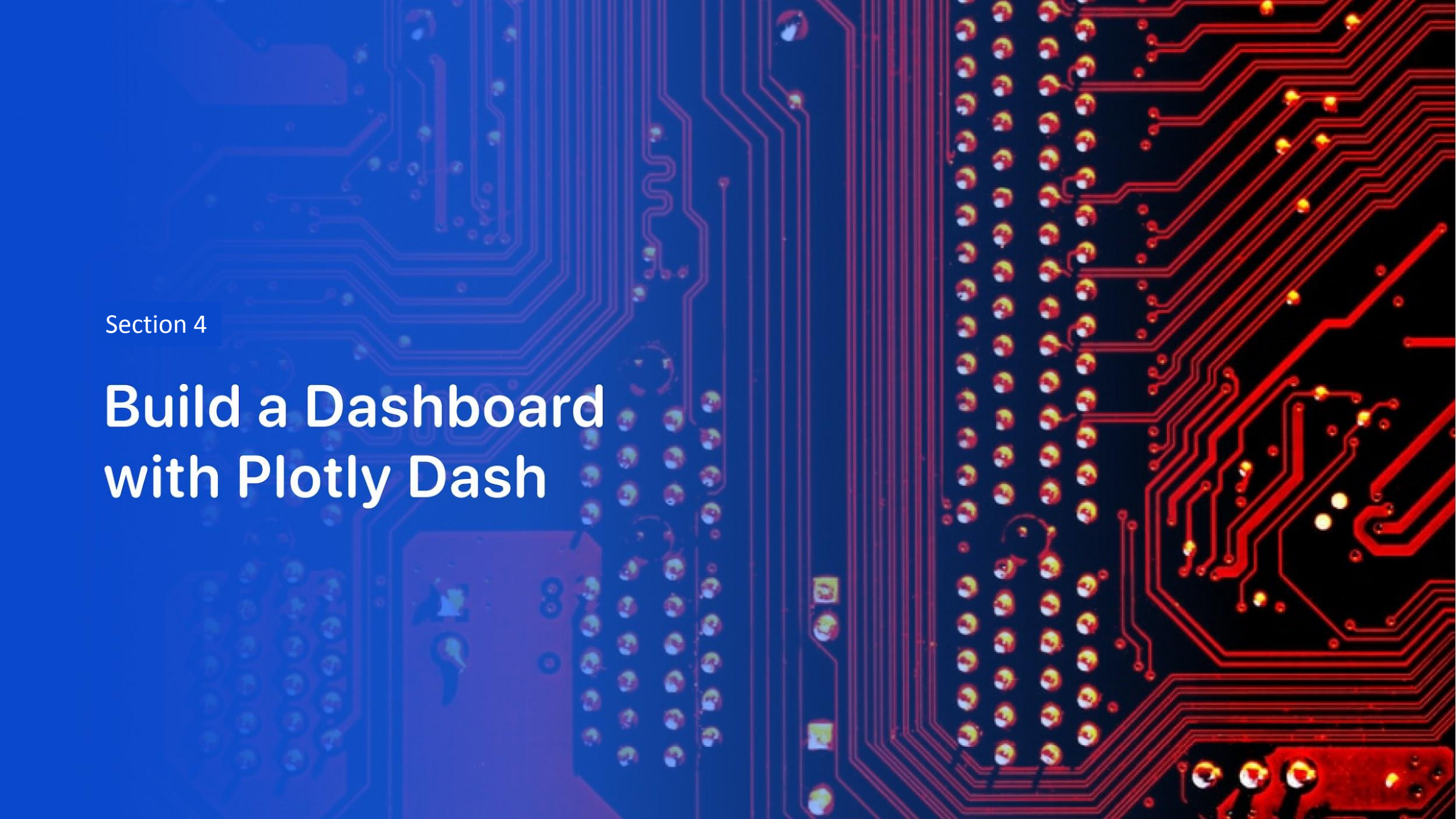
- Image shows the number of successful and failed outcomes for launch site KSC LC-39A



<Folium Map Screenshot 3>

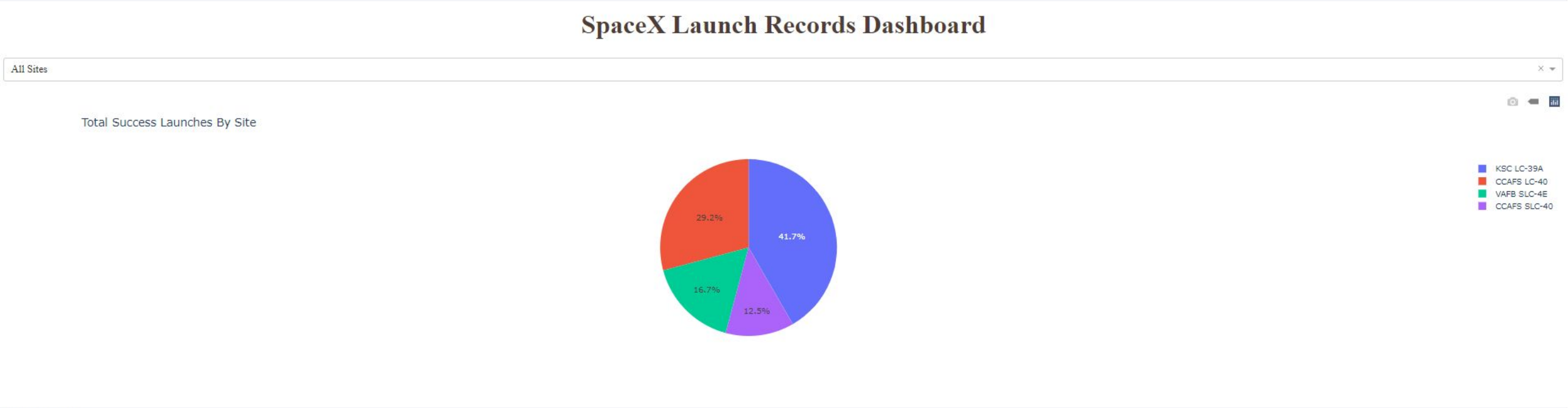
- Blue lines show distance of highway, coastline, and airport (out of view) to launch site CCAFS SLC-40




Section 4
Build a Dashboard with Plotly Dash

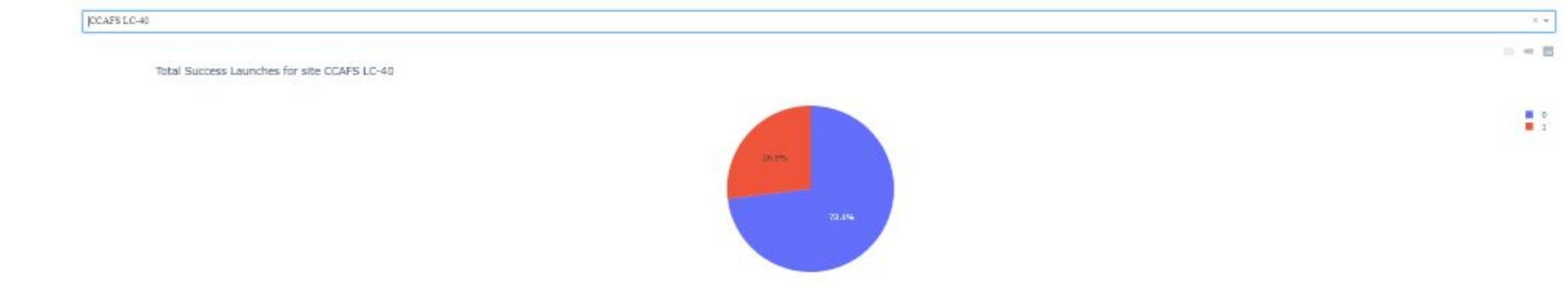
<Dashboard Screenshot 1>

- Launch site KSC LC-39A is the most successful with 41.7%
- Launch site CCAFS SLC-40 is the least successful with 12.5%

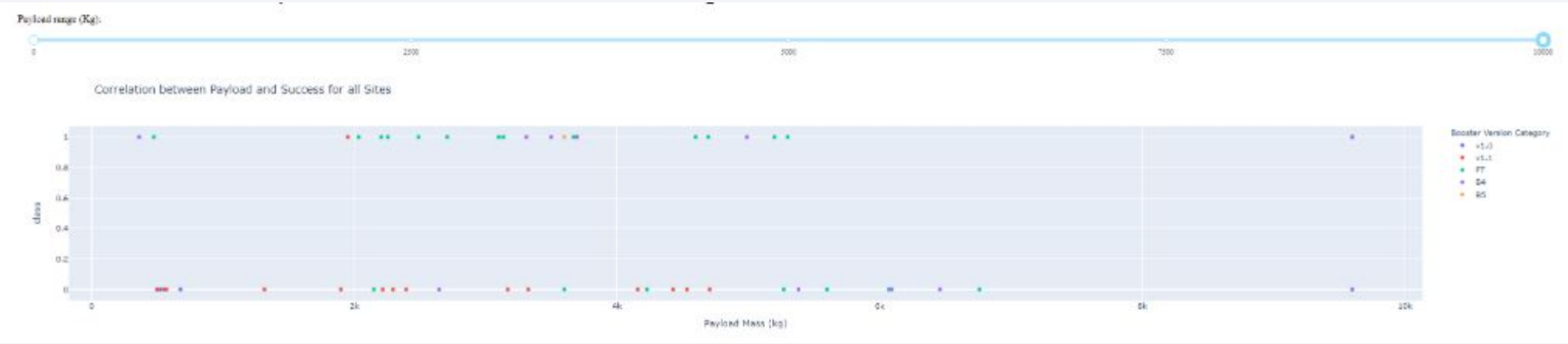


<Dashboard Screenshot 2>

- Launch site CCAFS LC-40 has a 72.1% success rate and a 26.9% failure rate



<Dashboard Screenshot 3>



Section 5

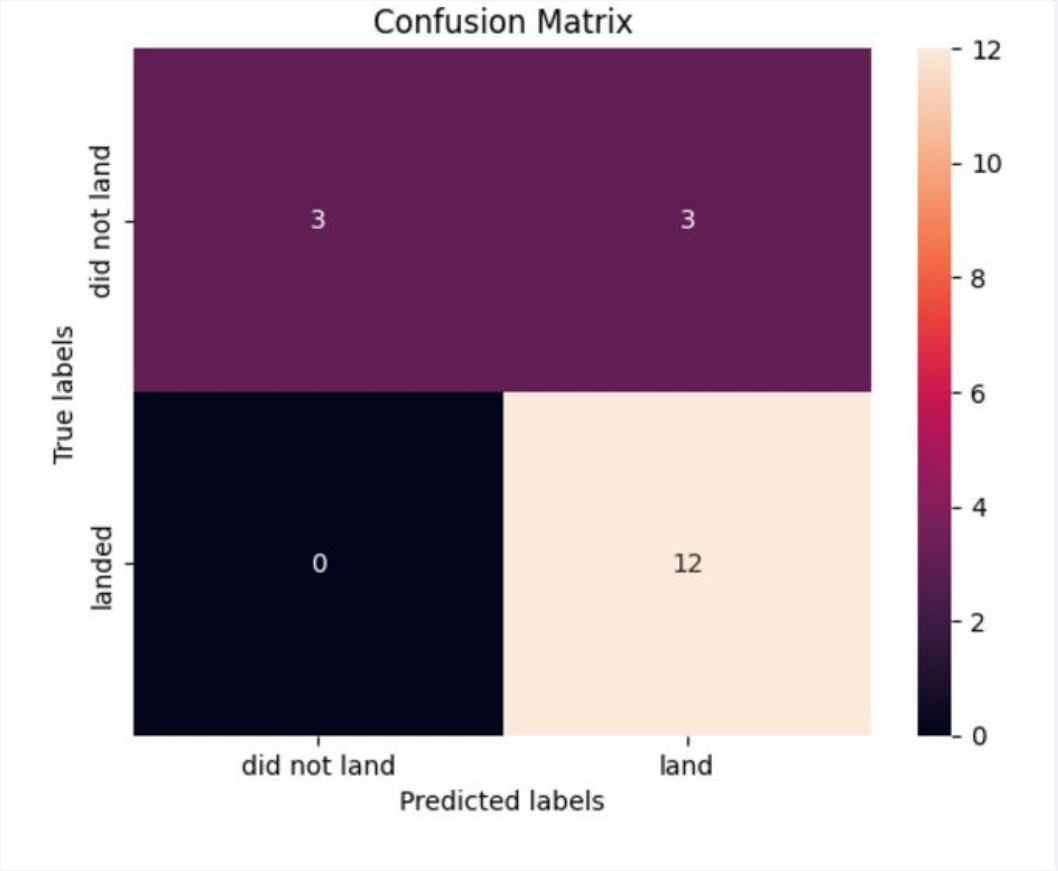
Predictive Analysis (Classification)

Classification Accuracy

- The model with the highest accuracy is the decision tree classifier with an accuracy score of 0.8767

Confusion Matrix

- Decision Tree Classifier:
- True Positive - 12 (True label is landed, Predicted label is also landed)
- False Positive - 3 (True label is not landed, Predicted label is landed)



Conclusions

- SpaceX had more successful launches as time progressed
- Light payloads perform better than heavy payloads
- KSC LC-39A is the most successful launch site
- GEO, HEO, SSO, and ES-L1 orbits have the highest success rate
- The decision tree classifier is the most robust machine learning technique for this data set

Appendix

- We did not use R in this project but if we wanted to complete additional statistical analyses RStudio might be useful

Thank you!