IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Matthias Mueller
01/10/2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection with API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis
  - Interactive Analytics
  - Predictive Analytics

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
  - Predict, if the Falcon 0 first stage will land successfully
  - Which factors determine if the rocket will land successfully

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data collection was done with the SpaceX API and web scraping
- Perform data wrangling
  - One-hot encoding for categorial features and data cleansing was done
- 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

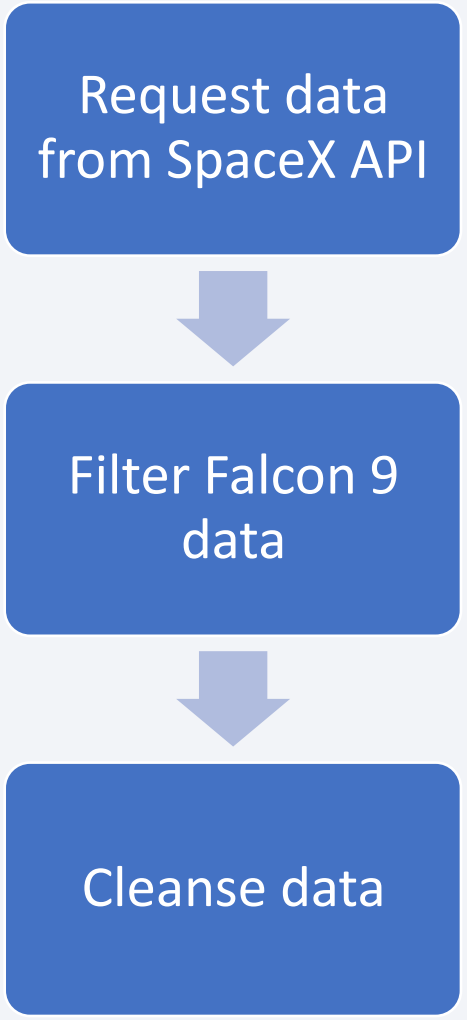
Data Collection

- In the process of data collection several methods were used:
  - In the project we used the get request via the SpaceX API to collect data
  - In the next step, the response content was decoded as json-format and turned it into a pandas dataframe
  - Next the data was cleaned, missing values were removed or filled in with fitting values (like average)
  - The BeautifoulSoup framework was used for web scraping and get the Falcon 9 launch records from Wikipedia as a html table and pass it on to a pandas data frame

Data Collection – SpaceX API

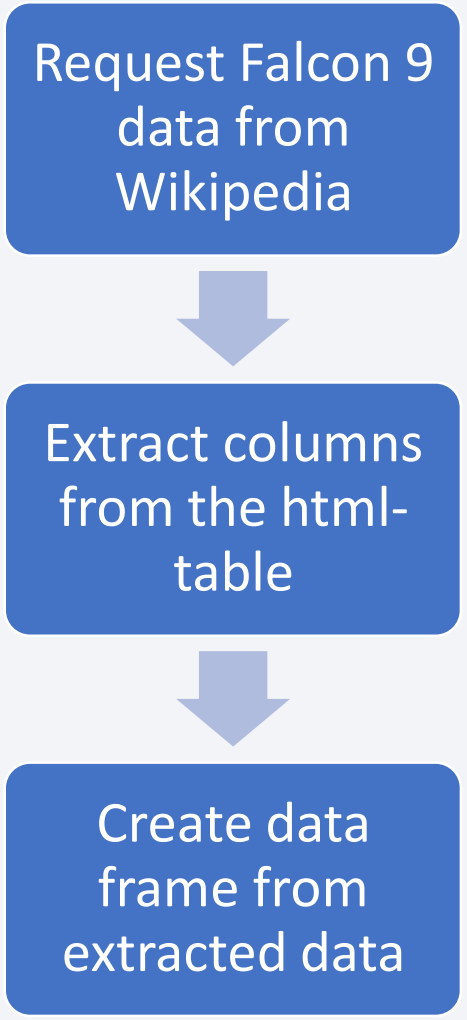
- The data was collected via the SpaceX API and converted to a data frame

- See GitHub for the complete workflow: https://github.com/Code770/IBM-Data-Science-capstone-project/blob/75d213337e8340b7779444ce88b6b27dbcced603/Data%20Collection.ipynb



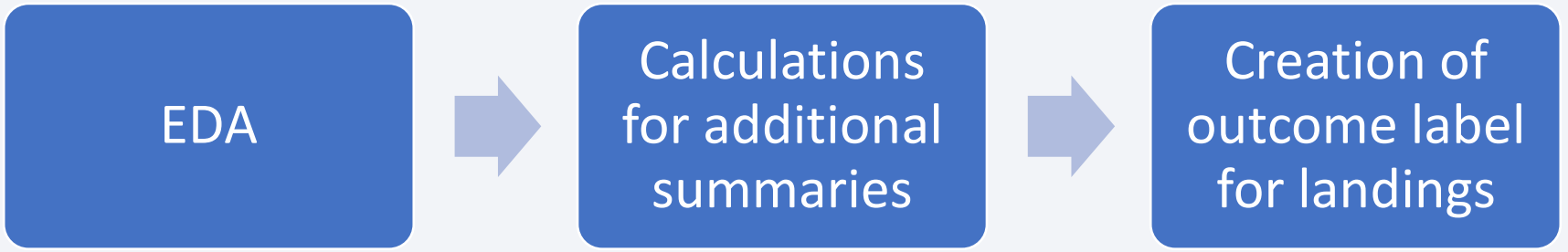
Data Collection - Scraping

- Webscraping from Wikipedia for the Falcon 9 launch records was applied using the BeautifoulSoup framework
- GitHub Url: https://github.com/Code770/IBM-Data-Science-capstone-project/blob/a3e5a70ed6bc5ce218899d5f518de3c65faa4aeb/Data%20Collection%20with%20Web%20Scraping.ipynb



Data Wrangling

- Exploratory Data Analysis (EDA) was used to discover the dataset
- The summaries of the launches per site or the occurrences of mission outcome per orbit were calculated and a landing outcome label was added to the dataframe

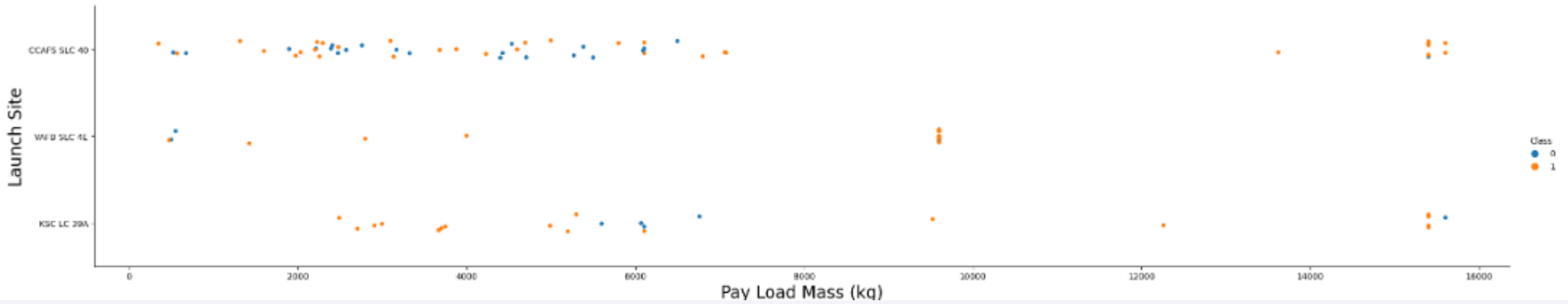


GitHub Url: https://github.com/Code770/IBM-Data-Science-capstone-project/blob/b5af91369a83d4a3ef858a9878b5c51877502820/Exploratory%20Data%20Analysis%20Overview.ipynb

EDA with Data Visualization

- For data exploration and to visualize relationships between features several graphs were used like barplots or scatterplots
  - Payload Mass vs flight number, launch site vs flight number, launch site vs payload mass, orbit vs flight number and payload vs orbit

In [6]:
```
# Plot a scatter point chart with x axis to be Pay Load Mass (kg) and y axis to be the launch site, and hue to be the class value
sns.catplot(y="LaunchSite", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("Pay Load Mass (kg)",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()
```



- GitHub Url: https://github.com/Code770/IBM-Data-Science-capstone-project/blob/9b6e45098fbea339f81602774b29d08bb9453358/EDA%20with%20Data%20Visualization.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 failed landing_outcomes in drone ship, their booster versions, and launch site names for 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/Code770/IBM-Data-Science-capstone-project/blob/043b7b8f163b5e099138082a69c4915da99f95b8/Exploratory%20Data%20Analysis%20Overview.ipynb

Build an Interactive Map with Folium

- Markers, circles, marker cluster and lines were added to the map to highlight areas of interest
- Circles highlight areas around coordinates of the launch sites
- Marker clusters show groups of coordinates, e.g. launches of a launch site
- Lines were used to show the calculated distances between two coordinates

- GitHub URL: https://github.com/Code770/IBM-Data-Science-capstone-project/blob/9955fb1685b8b0d39b4094ec7a7265ce3a49f0d6/Launch_sites_analysis_with_Folium.ipynb

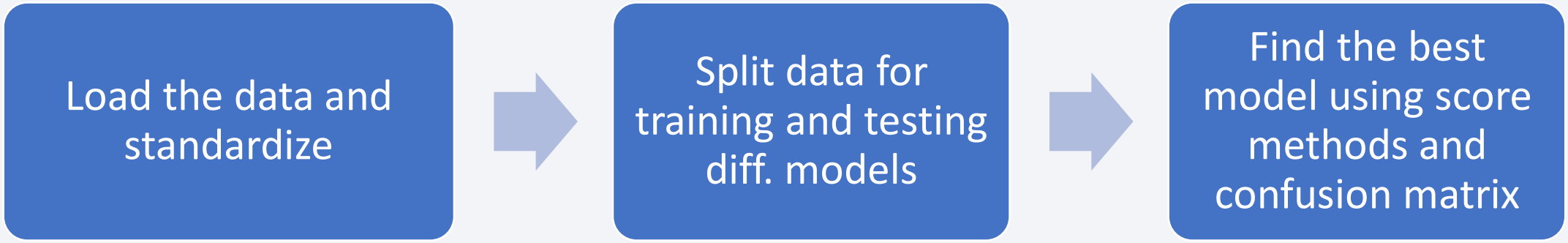
Build a Dashboard with Plotly Dash

- Following graphs haven been created:
  - Pie Chart to visualize percentage of launches per site
  - Scatter plot to visualize payload range
- This plots showed the relation between payloads and launch sites to quickly identify the successful combinations for future launches

- GitHub URL: https://github.com/Code770/IBM-Data-Science-capstone-project/blob/0baaa02eca858dd83c570b089a07b7e5f9d47f3f/spacex_dash_app.py

Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model



- GitHub URL: https://github.com/Code770/IBM-Data-Science-capstone-project/blob/a24923ca6cf9b7078bbc803169176d39e225954a/SpaceX_Machine_Learning_Prediction.ipynb

Results

- Exploratory data analysis results
  - 4 different launch sites
  - Average payload of Falcon 9 v1.1 booster is 2.928 kg
  - First successful landing was in 2015
  - Successful landings occurred more often as years passed
  - Most of the mission outcomes were successful

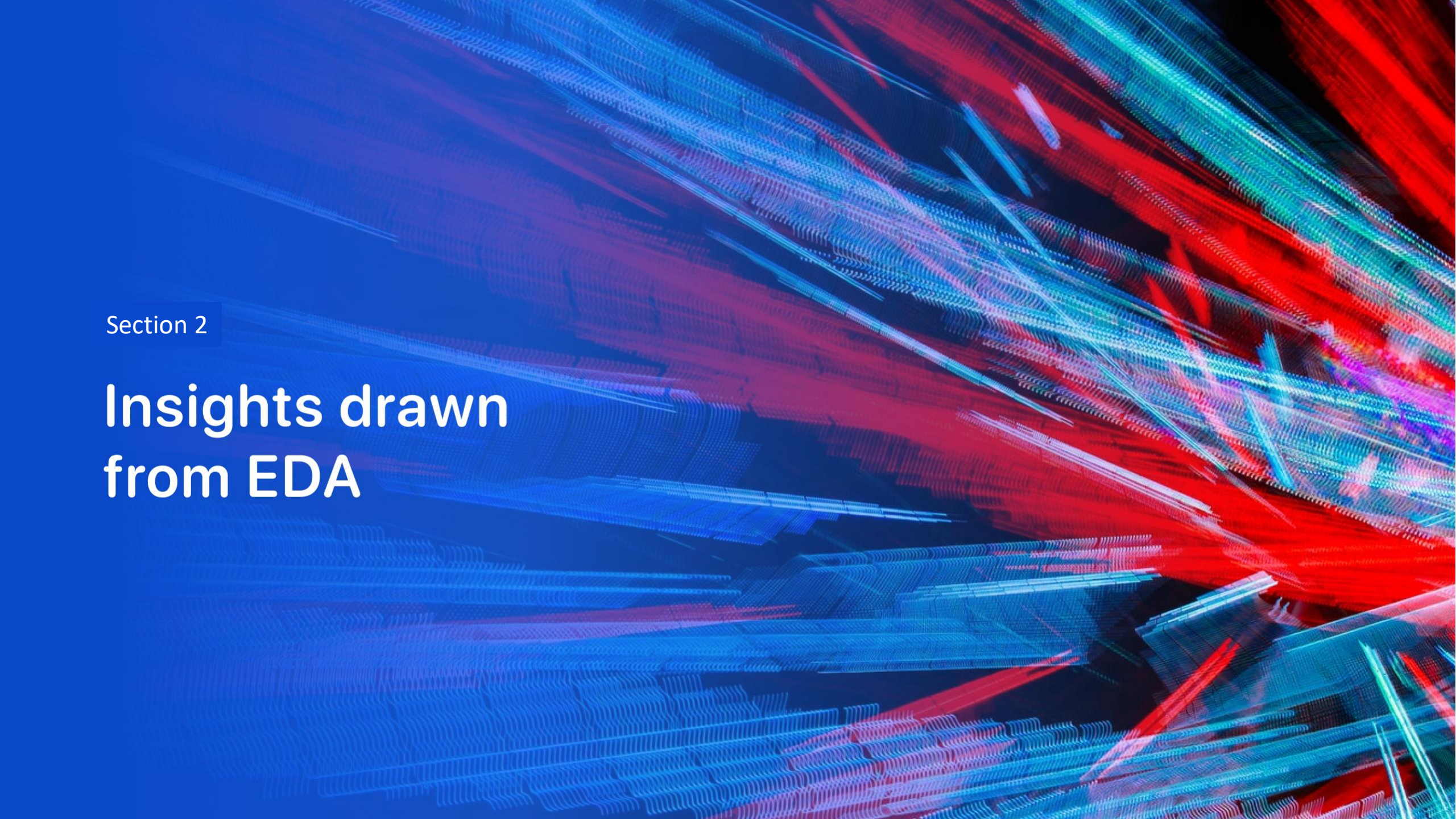
Results

- Interactive analytics were used to study the launch sites and the surroundings
- Most launches were near the equator and near coast lines

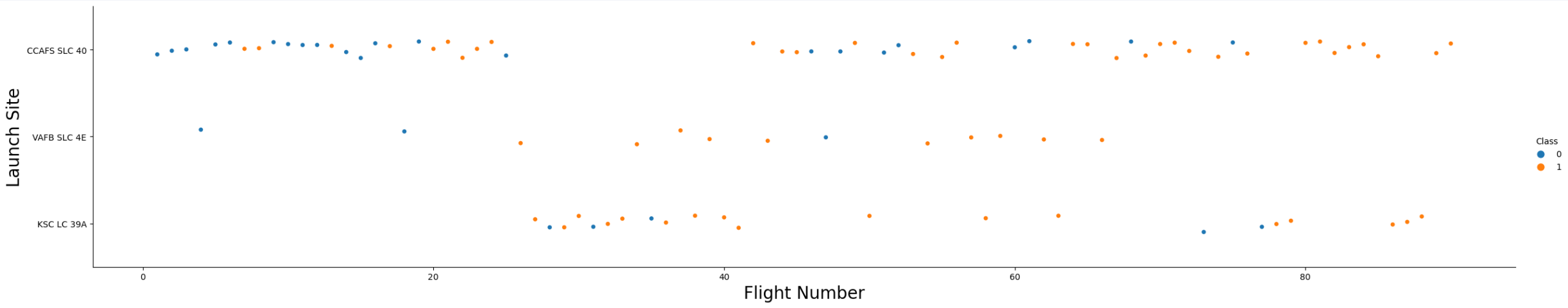
Results

- Predictive analytics showed that the Decision Tree Classifier is the best model to predict successful landings
- With an accuracy over 88%

```
Best model is DecisionTree with a score of 0.8857142857142858
Best params is : {'criterion': 'gini', 'max_depth': 4, 'max_features': 'sqrt', 'min_samples_leaf': 4, 'min_samples_split':
5, 'splitter': 'random'}
```

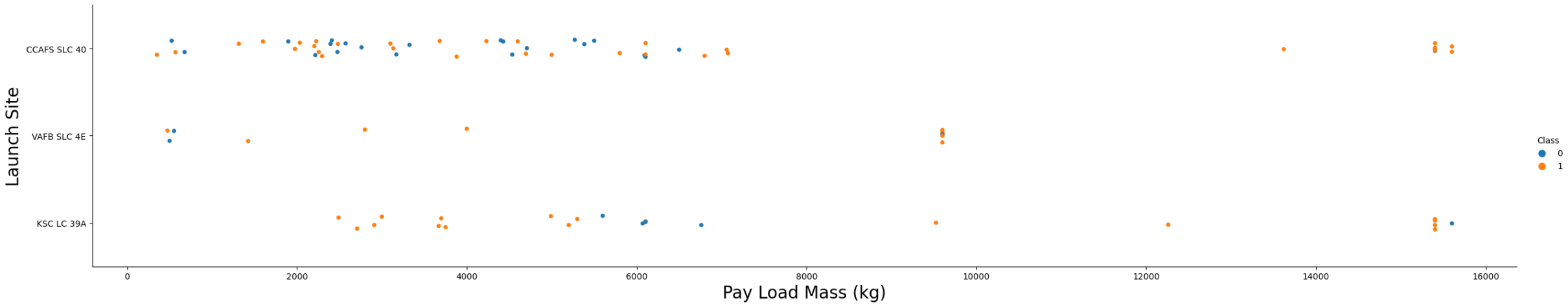

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- The scatter plot shows that the best launch site for recent launches is CCAF5 SLC 40
- The plot also shows that the general success rate improved over time

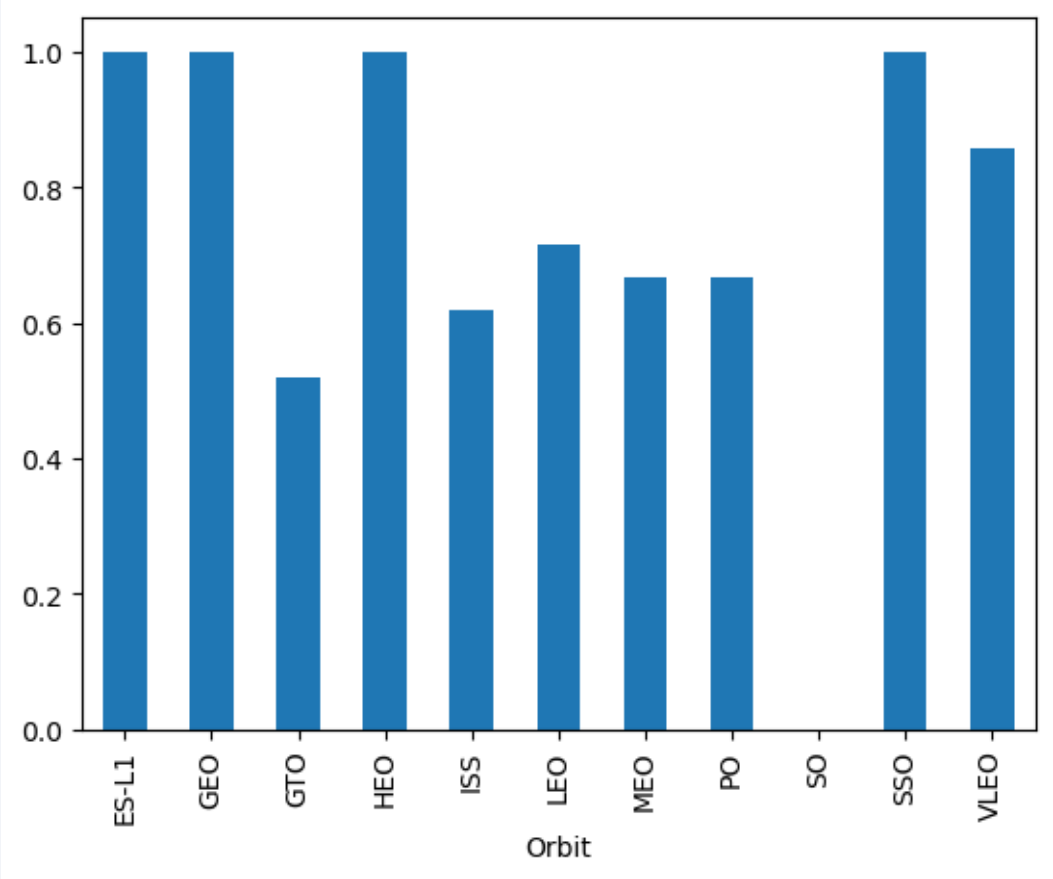
Payload vs. Launch Site



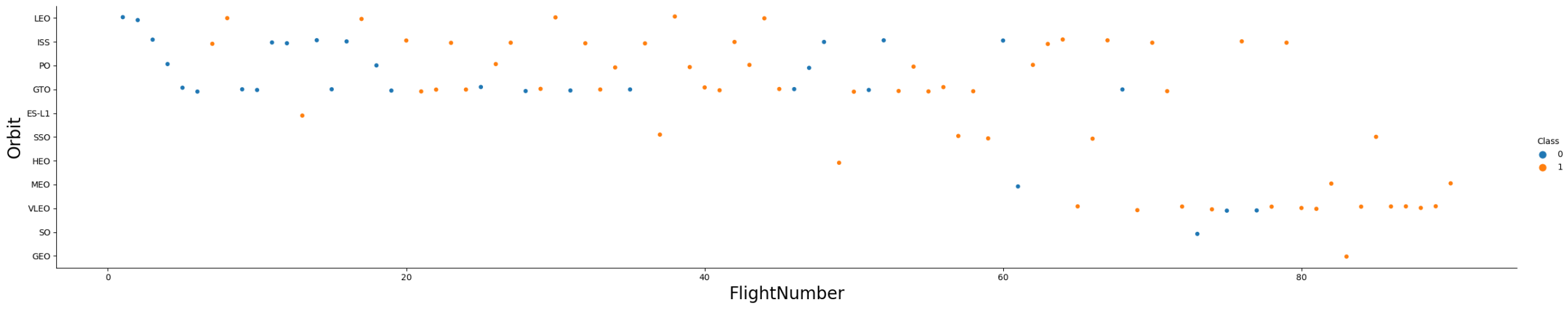
- Payloads over 9000kg are often successful
- No launches on VAFB-SLC over 10000kg

Success Rate vs. Orbit Type

- Highest success for orbits:
  - ES-L1
  - GEO
  - HEO
  - SSO
- Rank below:
  - VLEO
  - LFO

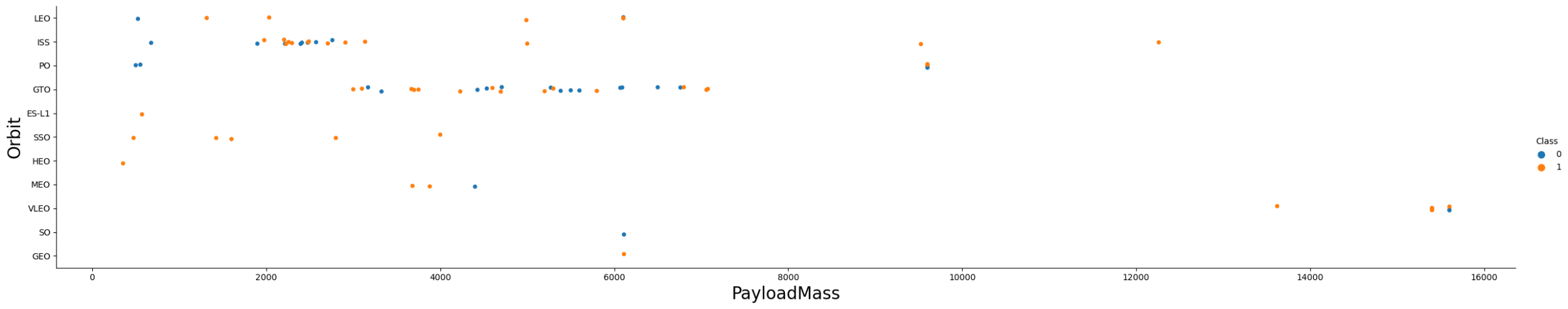


Flight Number vs. Orbit Type



- Success rate improved over time
- More launches to VLEO in recent history

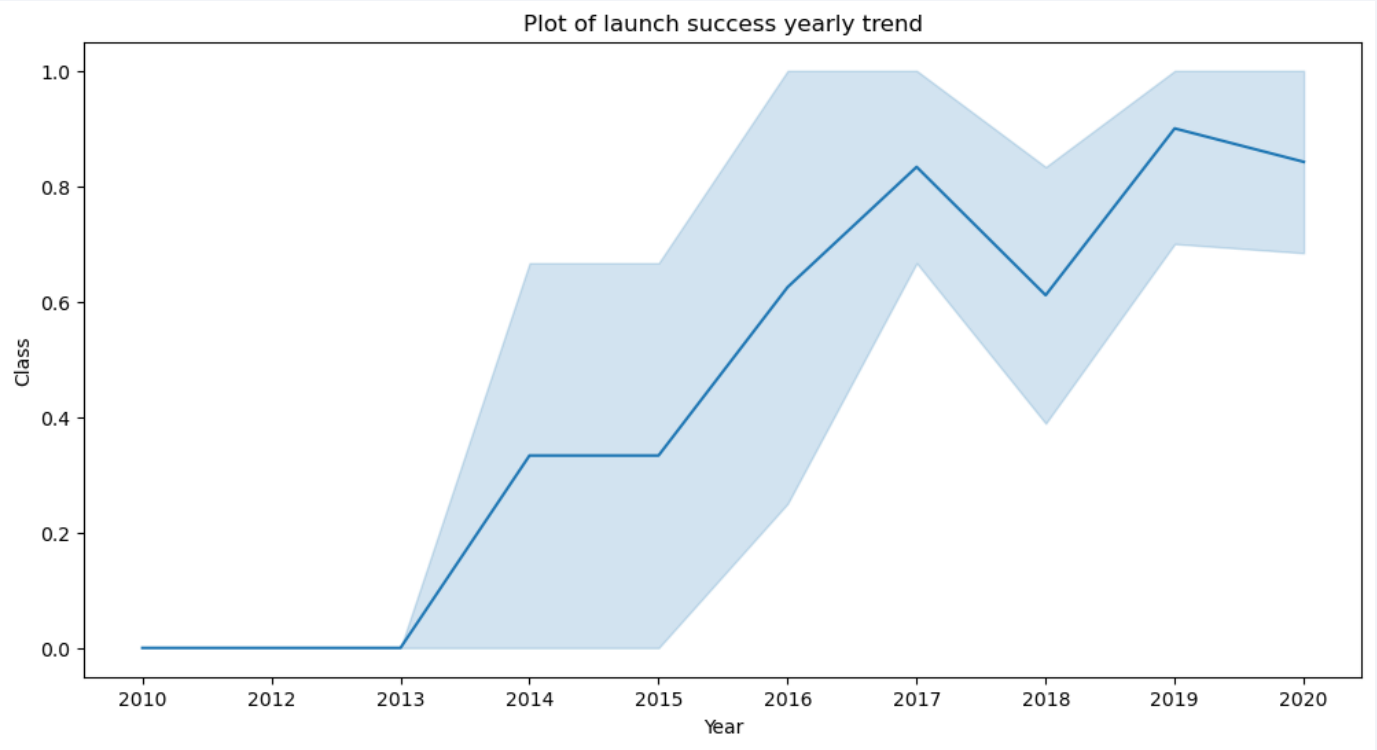
Payload vs. Orbit Type



- No relations between payload and orbit type detected
- Fewer launches to the heigher orbits

Launch Success Yearly Trend

- Success rate increased beginning in year 2013 until 2020
- First 3 years no successful launches



All Launch Site Names

- In the dataset there were 4 different launch sites:
  - CCAFS LC-40
  - CCAFS SLC-40
  - KSC LC-39A
  - VAFB SLC-4E

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- Query used LIMIT 5 to show 5 records

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

In [8]:
```
%sql select * from SPACEXTBL where launch_site like 'CCA%' LIMIT 5
```

```
 * ibm_db_sa://nwp46021:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/BLUD
B
Done.
```

Out[8]:

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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
- Query used the Sum function and the Like ‘string’ statement for calculation

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

```
%sql SELECT SUM(payload_mass__kg_) as Payload_sum from SPACEXTBL WHERE customer LIKE 'NASA (CRS)'
```

```
 * ibm_db_sa://nwp46021:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.ap
B
Done.
```

payload_sum
45596

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1
- Query used the average function and the like ‘string’ statement for calculation

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

```
%sql SELECT AVG(payload_mass__kg_) as Payload_avg_F9_v1_1 from SPACEXTBL WHERE booster_version LIKE 'F9 v1.1%'
```

```
 * ibm_db_sa://nwp46021:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:
B
Done.
```

payload_avg_f9_v1_1
2534

First Successful Ground Landing Date

- Date of the first successful landing outcome on ground pad
- Query used the min function and the like ‘string’ statement for calculation

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

Hint:Use min function

```
%sql SELECT MIN(DATE) FROM SPACEXTBL WHERE LANDING__OUTCOME LIKE 'Success (ground pad)'
```

```
 * ibm_db_sa://nwp46021:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.dat
B
Done.
```

1
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

- SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE LANDING__OUTCOME LIKE 'Success (drone ship)' AND payload_mass__kg_ 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

- Calculate the total number of successful and failure mission outcomes:
- 71
- Query used to count function
- SELECT COUNT(LANDING__OUTCOME) FROM SPACEXTBL WHERE LANDING__OUTCOME LIKE 'Success%' OR LANDING__OUTCOME LIKE 'Failure%'

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Query used subquery

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

```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE payload_mass__kg_ = (SELECT MAX(payload_mass__kg_) FROM SPACEXTBL)
```

```
 * ibm_db_sa://nwp46021:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:3150
B
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

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- SELECT LANDING__OUTCOME, booster_version, launch_site FROM SPACEXTBL WHERE LANDING__OUTCOME LIKE 'Failure (drone_ship)' AND DATE LIKE '2015%'

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
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
- SELECT COUNT(LANDING__OUTCOME) as outc_count, Landing__OUTCOME FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING__OUTCOME ORDER BY outc_count DESC

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

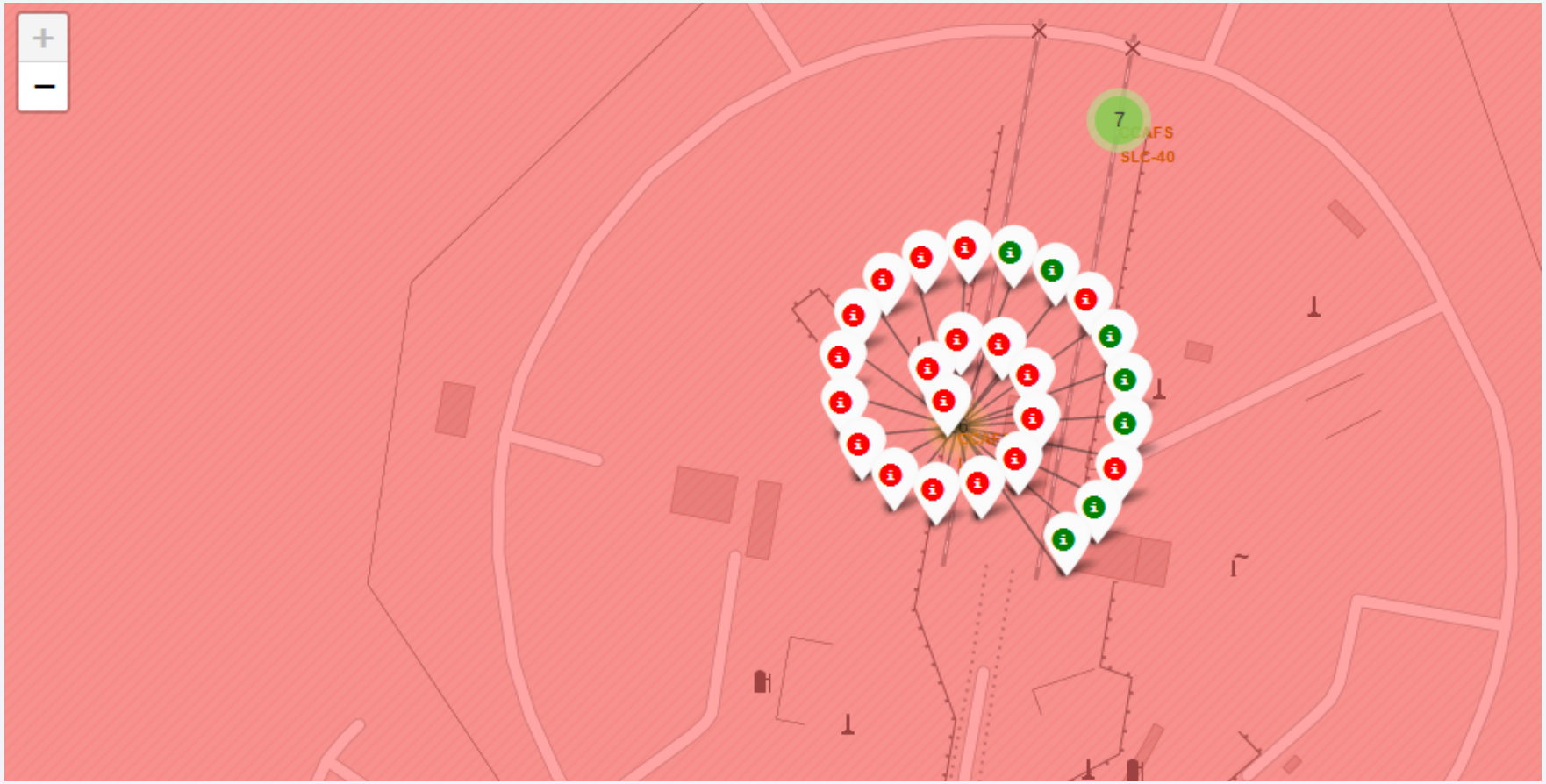

Section 3
Launch Sites
Proximities Analysis

Launch Sites

- Launch sites are near coast lines of the oceans and not too far from roads and railroads

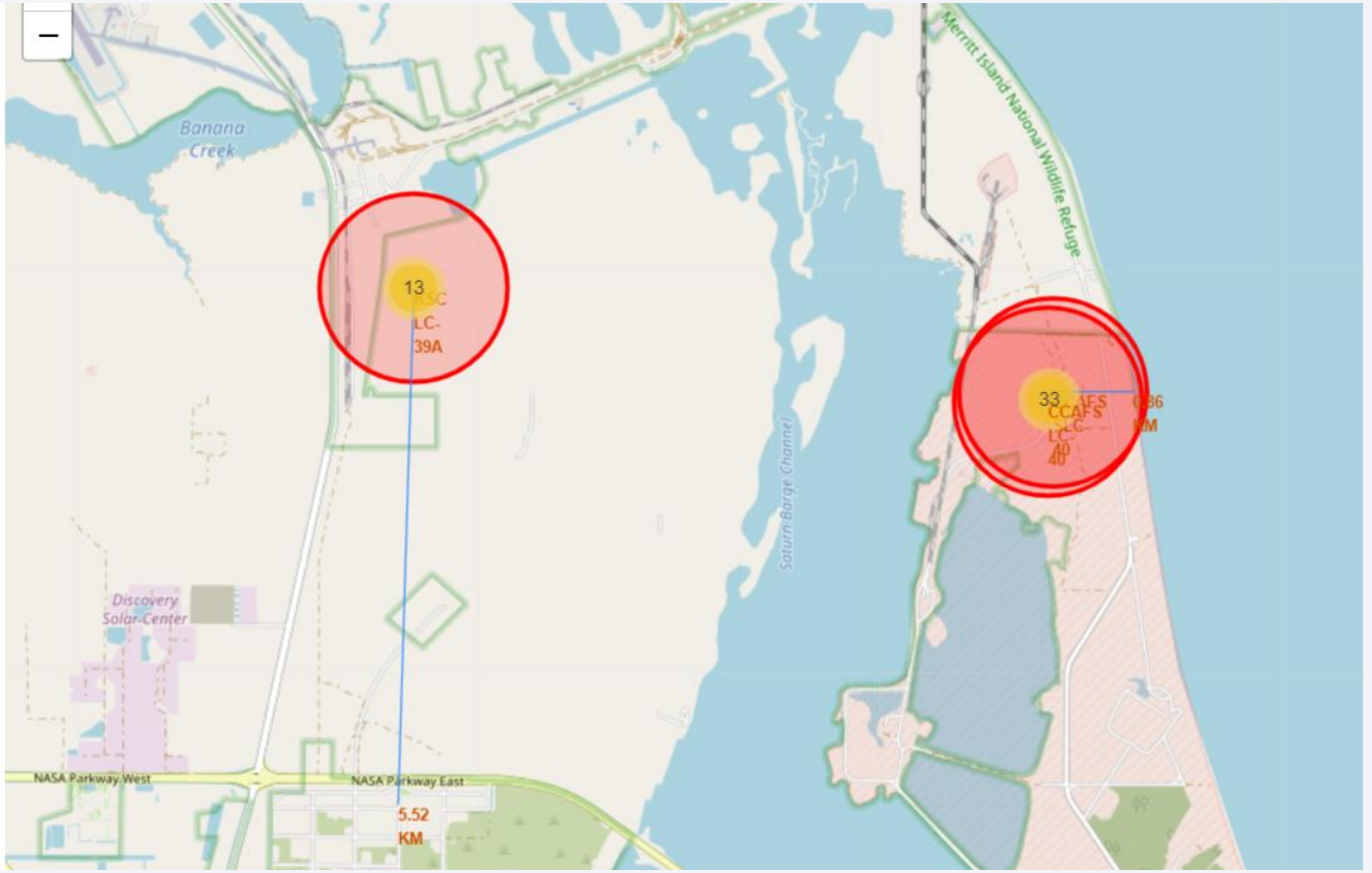
Launch outcomes

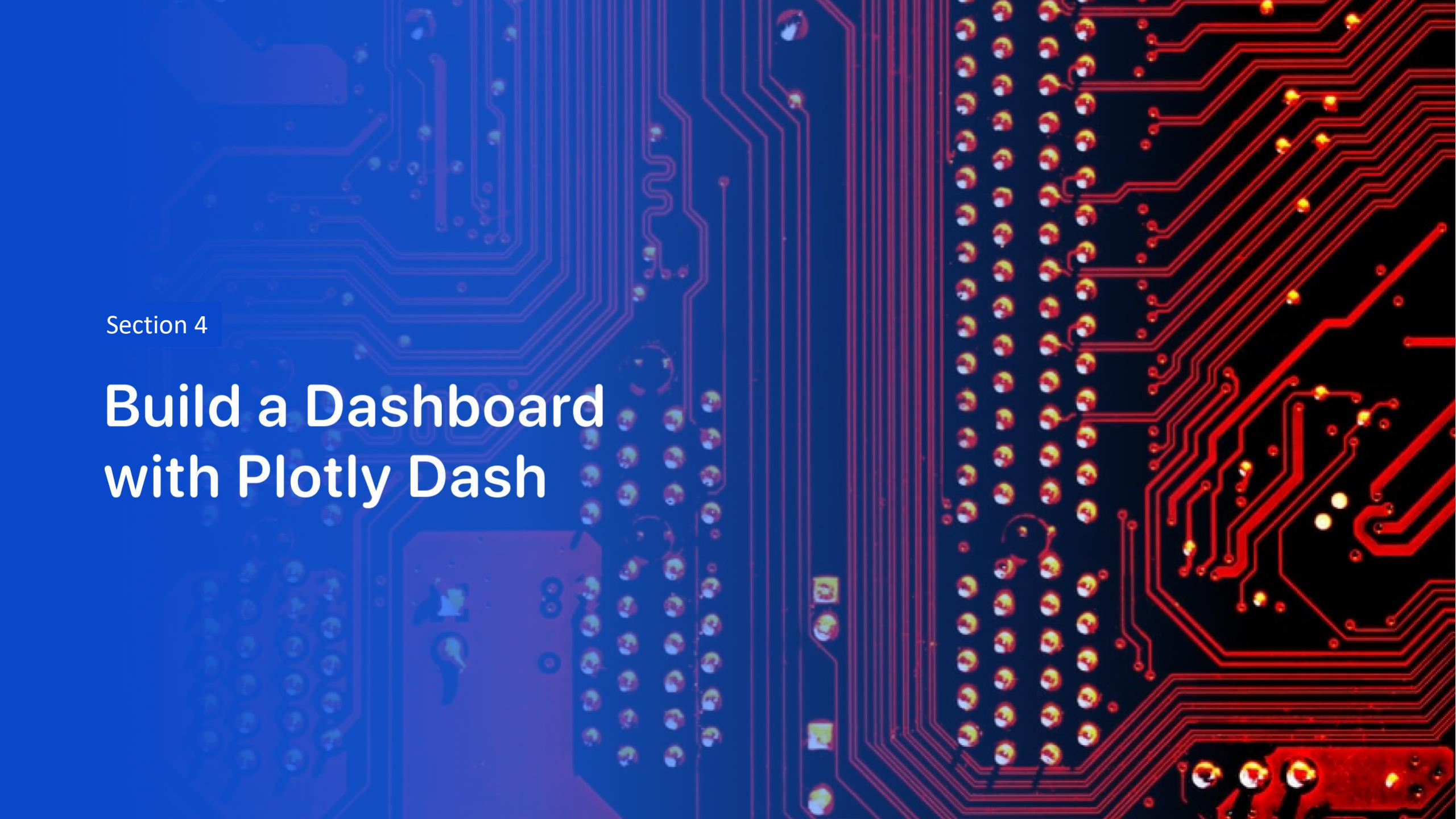
- Colored markers indicate successful and failed launches



Launch sites area

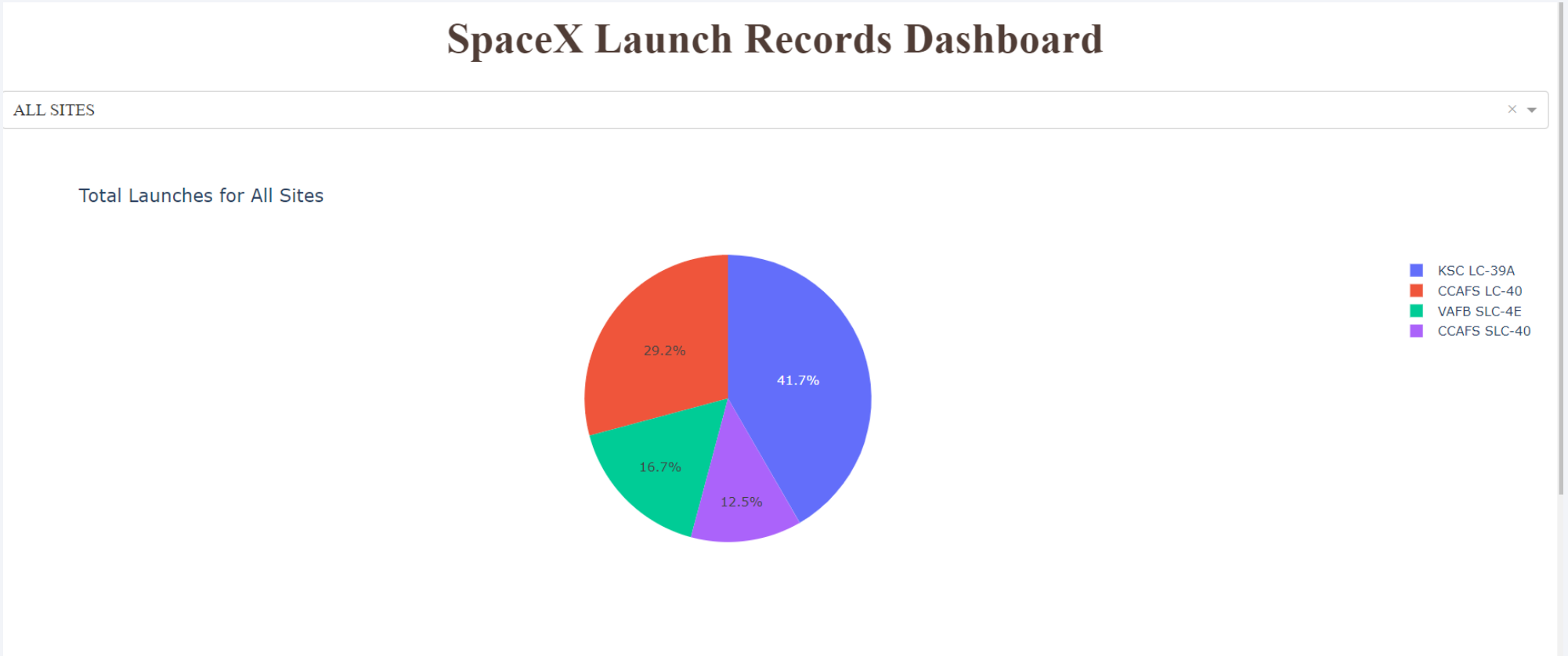
- Launch sites are near important logistic roads and railroads and far from populated areas




Section 4
Build a Dashboard with Plotly Dash

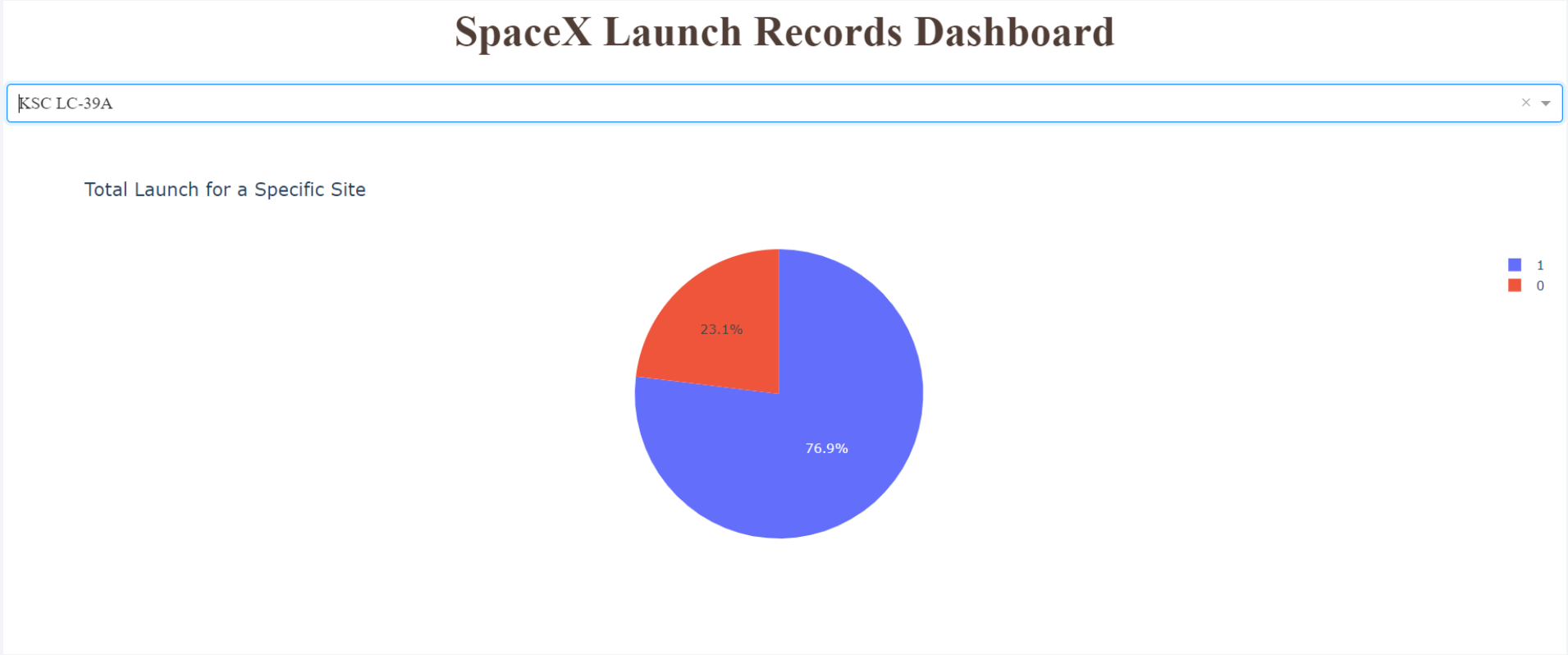
Successful Launches by Site

- Pie chart for total success by launch site



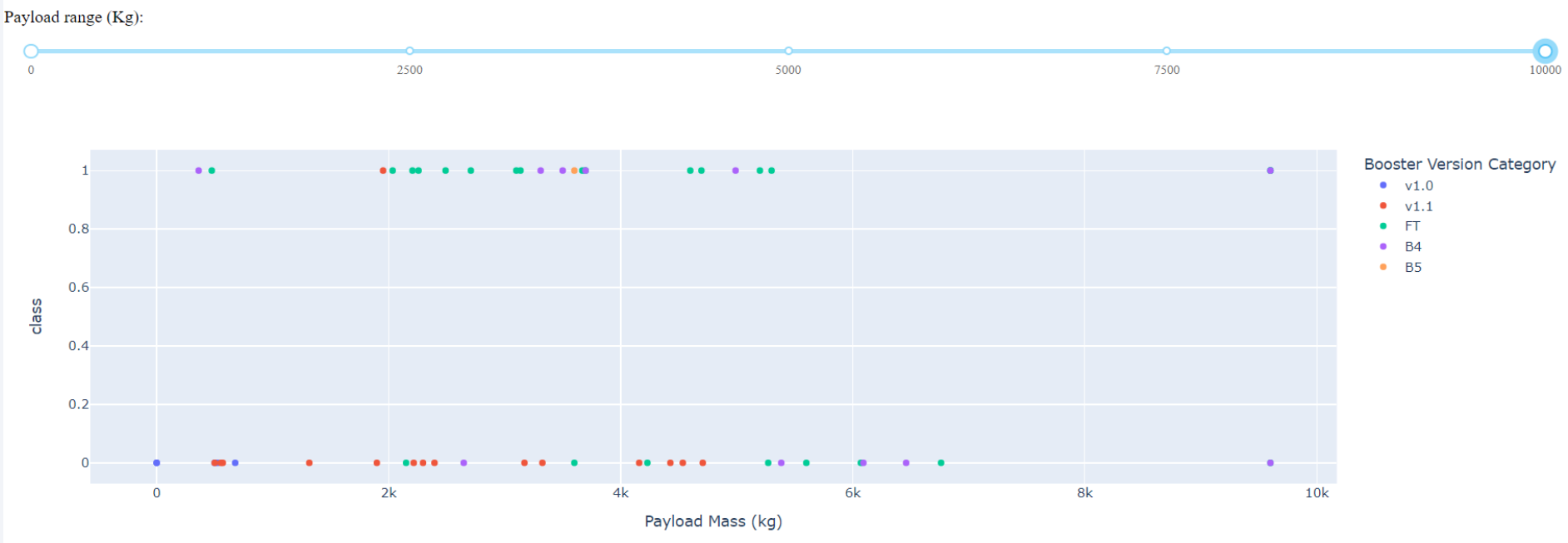
Launch success for KSC LC-39-A

- KSC LC-39-A were the most successful launch site with 76.9% success rate



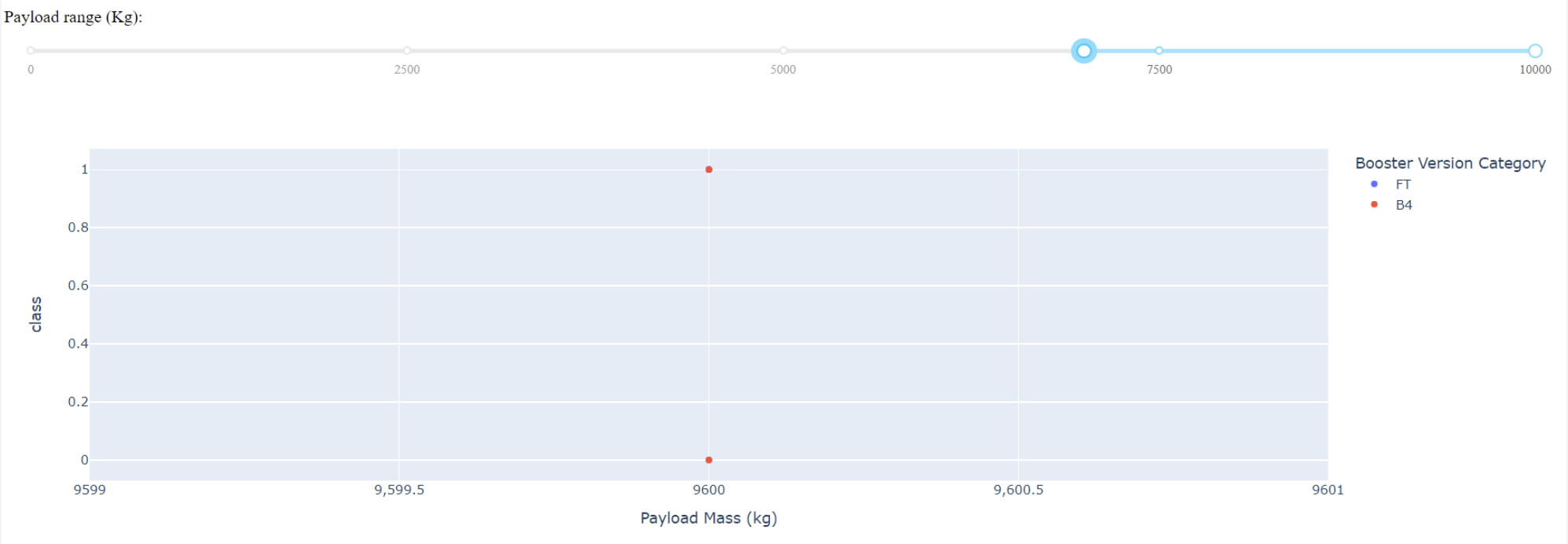
Payload vs. launch outcome

- Payloads below 6000kg and FT boosters are most successful



Payload vs. launch outcome

- There's not enough data for conclusions above 7000 kg

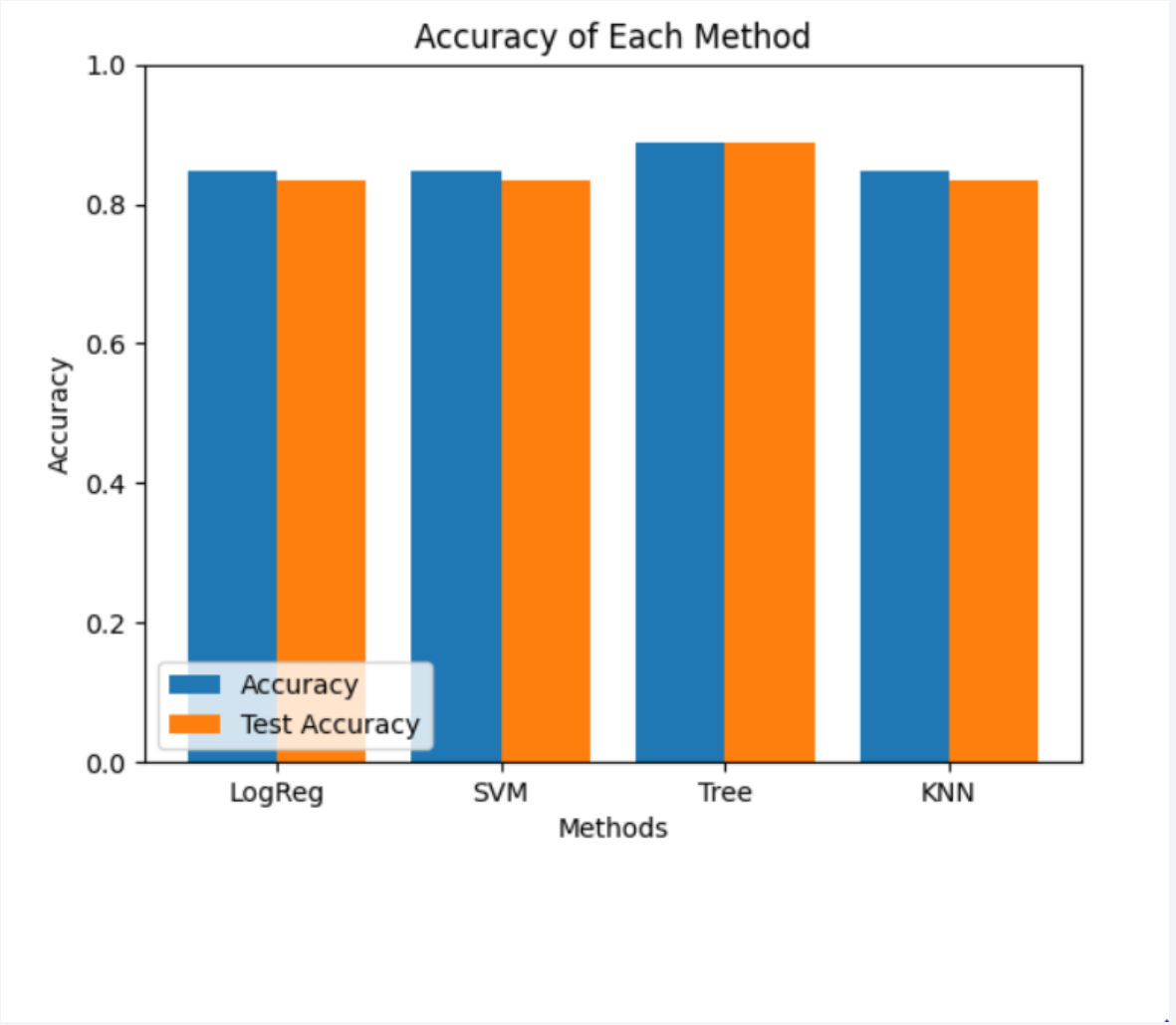


Section 5

Predictive Analysis (Classification)

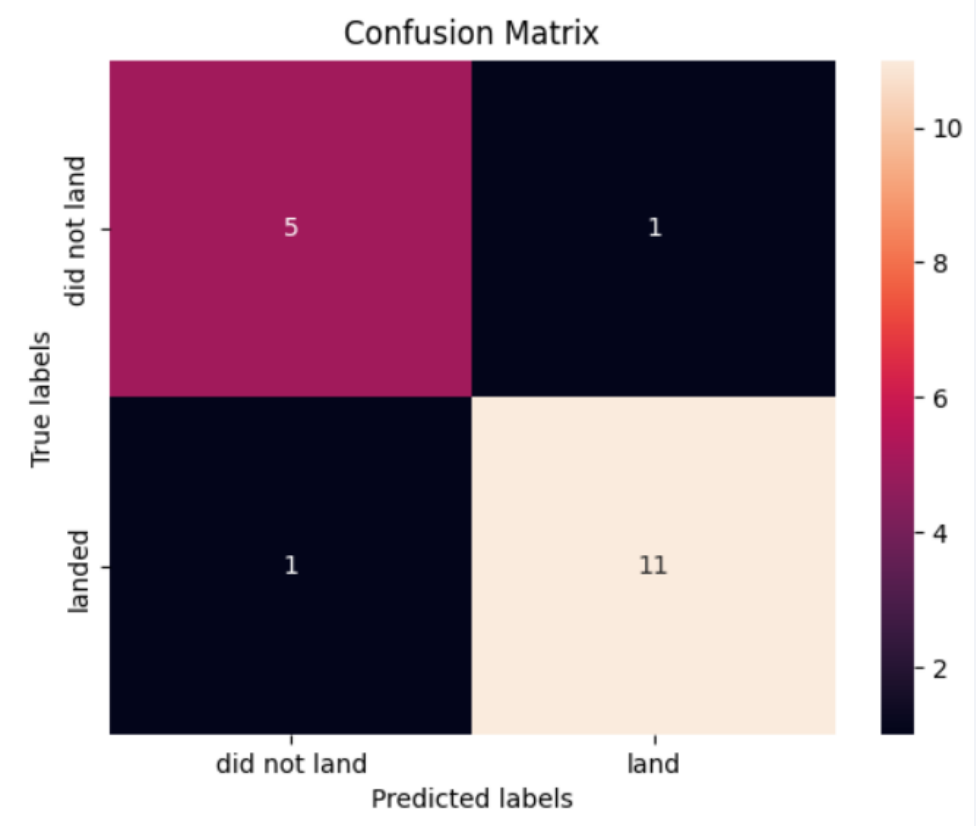
Classification Accuracy

- The best model was the Decision tree classifier with an accuracy of 87%



Confusion Matrix Decision Tree

- Confusion Matrix for the decision tree classifier



Conclusions

- Best launch site is KSC-LC-39A
- Success rate increased over the years
- Launches above 7.000kg are more successful
- Decision tree classifier is best suited to predict launch outcome

Appendix

- All the important data can be found in GitHub Url:
- https://github.com/Code770/IBM-Data-Science-capstone-project

Thank you!