IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mary Jane R. Edera
November 19, 2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection: Obtained data from publicly accessible SpaceX API and transformed with web scraping methods.
  - Data Wrangling: The gathered SpaceX data was processed and cleansed.
  - Exploratory Data Analysis (EDA): Performed a preliminary analysis using SQL and visualized the data using Python (Pandas, Matplotlib) for more in-depth insights.
  - Launch Sites examination: Conducted interactive visual analytics of SpaceX launch sites with interactive visual analytics using Folium and Ploty Dash
  - Machine Learning: Created a models using SVM (Support Vector Machine), Decision Tree Classifier, Logistic Regression and KNN (k-Nearest Neighbors) to predict the results of SpaceX landings.
- Summary of all results
  - Successfully collected and cleaned SpaceX data.
  - Visualized launch site data, showing significant patterns and geographical influences on launch outcomes.
  - Developed a predictive machine learning models that all performed equally well, achieving 83.33% test data accuracy, effectively forecasting SpaceX landing success.

Introduction

- Project background and context

The era of commercial spaceflight has arrived, with companies making space exploration more affordable for everyone. Space Y would like to compete with SpaceX, which promotes its Falcon 9 rocket at a cost of $62 million, significantly lower than other providers whose prices start at $165 million each. Much of the savings come from SpaceX's ability to reuse the first stage of the rocket. This study was structured to predict if the Falcon 9 first stage will land successfully.

- Problems you want to find answers
  - How likely is it that a future Falcon 9 first stage landing will be successful, based on data from previous Falcon 9 rocket launches?
  - What attributes are correlated with successful landings?
  - Which model have best accuracy using the training data?

Section 1

Methodology

Methodology

Executive Summary

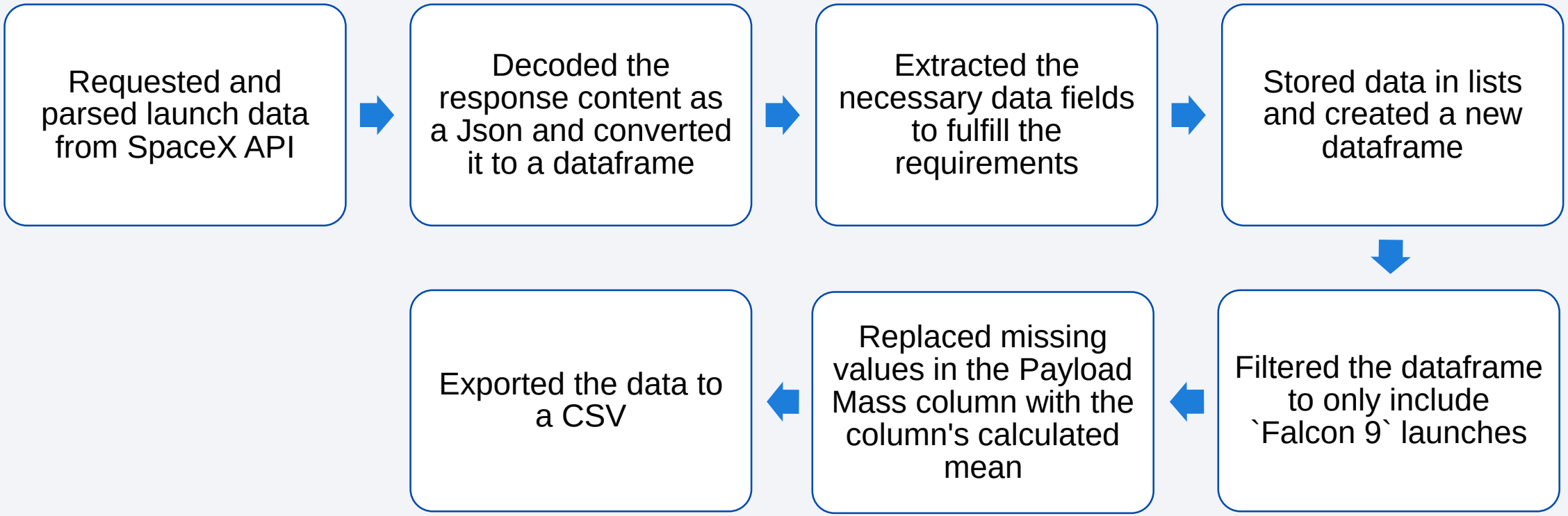
- Data collection methodology:
  - Data was collected from SpaceX REST API and Wikipedia
- Perform data wrangling
  - Training labels for the supervised models were generated by applying one-hot encoding to convert mission outcomes.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

Data was collected from SpaceX REST API and Wikipedia

- SpaceX REST API:
  - https://api.spacexdata.com/v4/rockets/
  - https://api.spacexdata.com/v4/launchpads/
  - https://api.spacexdata.com/v4/payloads/
  - https://api.spacexdata.com/v4/cores/
  - https://api.spacexdata.com/v4/launches/past
- Wikipedia:
  - https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

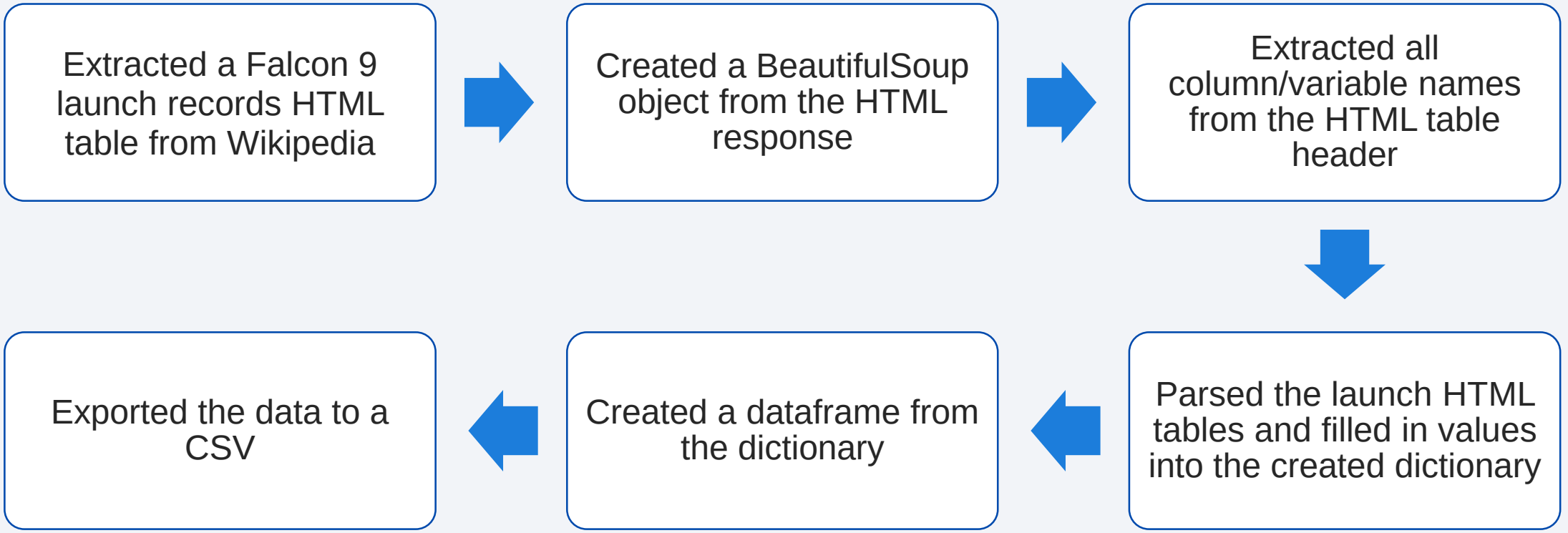
Data Collection – SpaceX API



[Github Link]

https://github.com/Gyned/Applied-Data-Science-Capstone/blob/main/Data%20Collection%20%E2%80%93%20SpaceX%20API.ipynb

Data Collection - Scraping

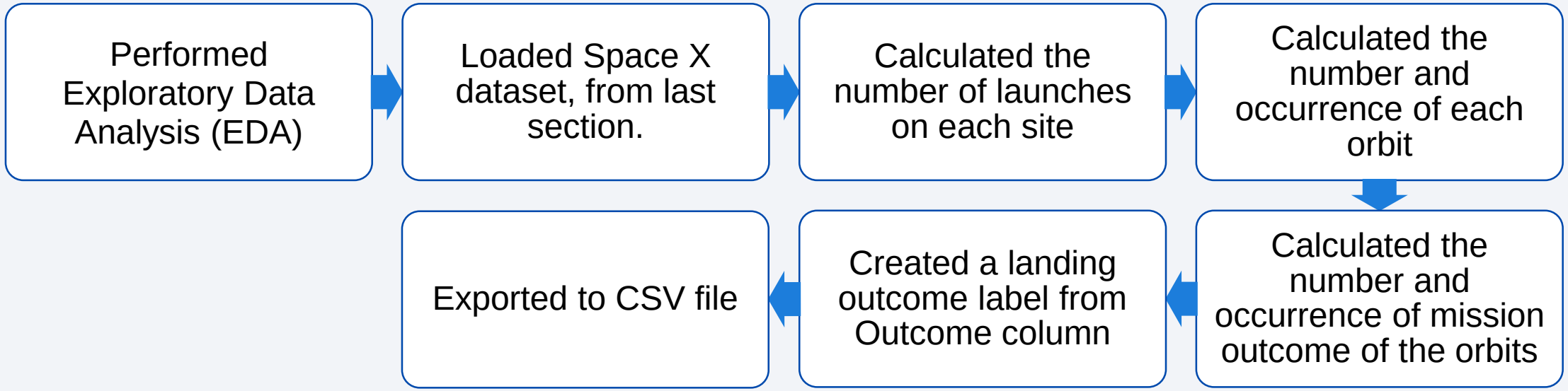


[Github Link]

https://github.com/Gyned/Applied-Data-Science-Capstone/blob/main/Data%20Collection%20-%20Web%20Scraping.ipynb

Data Wrangling

In the data set, there are several different cases where the booster did not land successfully. We will mainly convert those outcomes into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful.



Github Link

https://github.com/Gyned/Applied-Data-Science-Capstone/blob/main/Data%20Wrangling.ipynb

EDA with Data Visualization

- A scatter plot was used to visualize the relationship or correlation between two variables, making it easier to observe patterns and trends.
  - Relationship between Flight Number and Launch Site
  - Relationship between Payload Mass and Launch Site
  - Relationship between Flight Number and Orbit Type
  - Relationship between Payload Mass and Orbit Type
- A bar chart was created to compare variable values at a specific moment, highlighting the highest groups and their relationships, with bar lengths reflecting their values.
  - Relationship between success rate of each orbit type
- Line Chart was used to visualize the yearly trends.
  - Average launch success yearly trend

Github Link

https://github.com/Gyned/Applied-Data-Science-Capstone/blob/main/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

SQL queries performed:

- Displayed the names of the unique launch sites in the space mission
- Displayed 5 records where launch sites begin with the string ‘CCA’
- Displayed the total payload mass carried by boosters launched by NASA (CRS)
- Displayed average payload mass carried by booster version F9 v1.1
- Listed the date when the first successful landing outcome in ground pad was achieved
- Listed the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listed the total number of successful and failure mission outcomes
- Listed the names of the booster versions which have carried the maximum payload mass
- Listed the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
- Ranked 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 Link

https://github.com/Gyned/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Marked all launch sites on a map to give intuitive insights about where are those launch sites
  - Created a folium Map object, with an initial center location to be NASA Johnson Space Center at Houston, Texas.
  - Added a highlighted circle area with a text label on a specific coordinate
  - Added a circle for each launch site in data frame launch_sites
- Marked the success/failed launches for each site on the map to enhance the map and see which sites have high success rates
  - Added colored markers for all launch records to identify successful (green) and failed (red).
- Calculated the distances between a launch site to its proximities to easily find the coordinates of any points of interests
  - Drew a PolyLine between a launch site to the selected coastline point, to its closest city, railway, highway

Github Link

https://github.com/Gyned/Applied-Data-Science-Capstone/blob/main/Build%20an%20Interactive%20Map%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

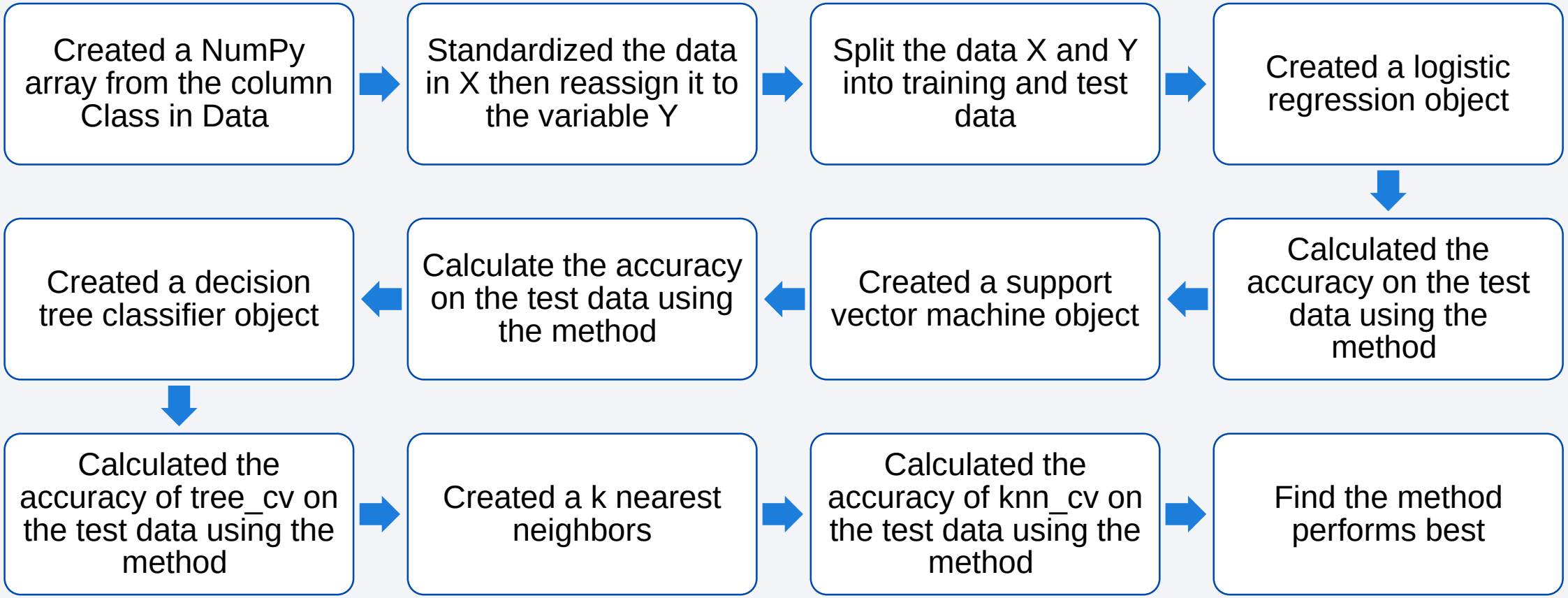
This contains a dropdown list and a range slider to interact with a pie chart and a scatter point chart and to perform interactive visual analytics on SpaceX launch data in real-time.

- Added a Launch Site Drop-down Input Component
  - ➢ to enable selection of different launch sites
- Added a callback function to render success-pie-chart based on selected site dropdown
  - ➢ to show the success count and failed count for the selected site.
- Added a Range Slider to Select Payload
  - ➢ to find if variable payload is correlated to mission outcome and to be able to easily select different payload range and see if we can identify some visual patterns.
- Added a callback function to render the success-payload-scatter-chart scatter plot
  - ➢ to visually observe their correlated with mission for selected site(s).

Github Link

https://github.com/Gyned/Applied-Data-Science-Capstone/blob/main/Build%20a%20Dashboard%20with%20Plotly%20Dash.py

Predictive Analysis (Classification)

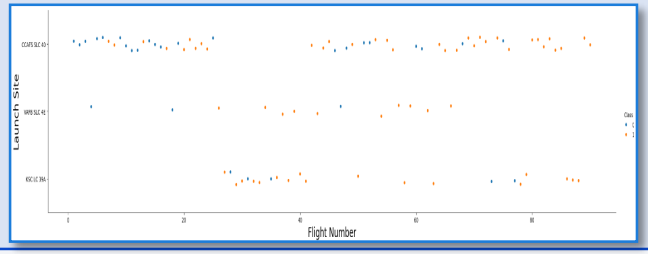


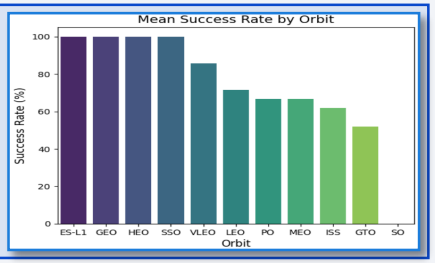
Github Link

https://github.com/Gyned/Applied-Data-Science-Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction.ipynb

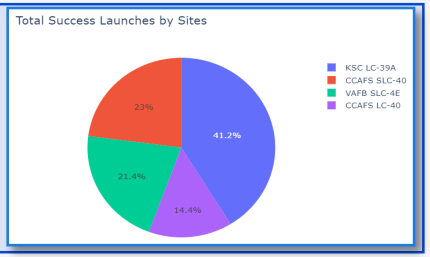
Results

- Exploratory data analysis results



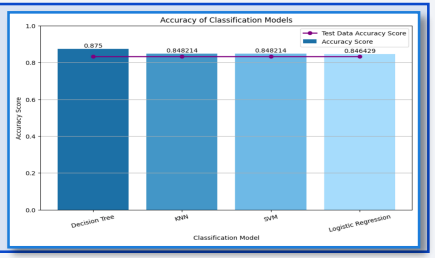


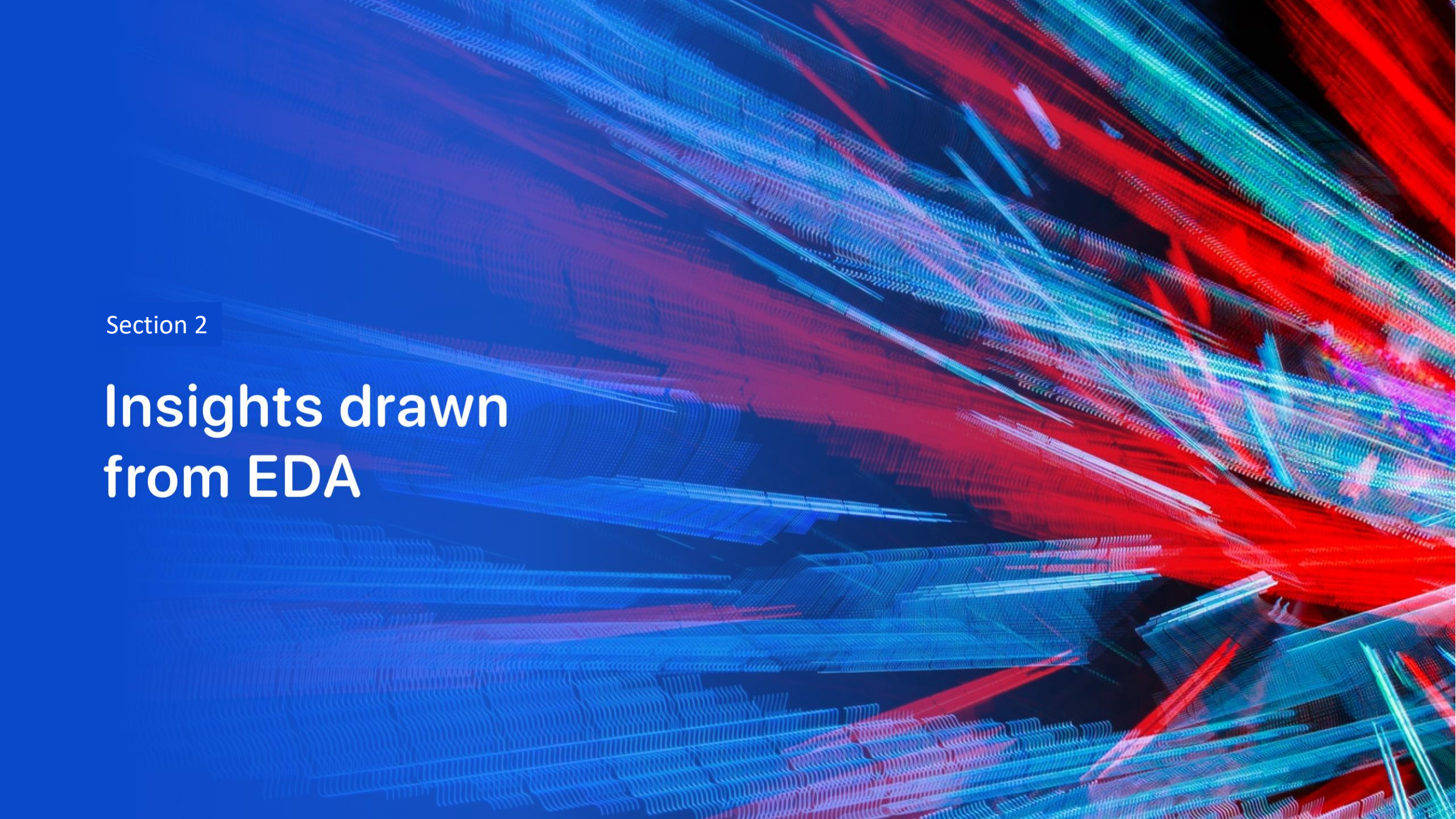
- Interactive analytics demo in screenshots



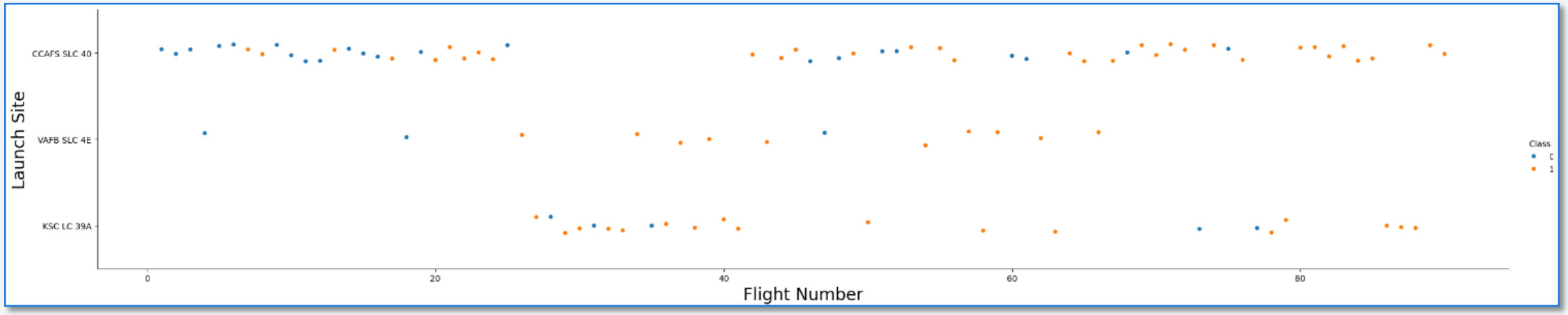
- Predictive analysis results

	Classification Model	Accuracy Score	Test Data Accuracy Score
2	Decision Tree	0.889286	0.833333
3	KNN	0.848214	0.833333
1	SVM	0.848214	0.833333
0	Logistic Regression	0.846429	0.833333



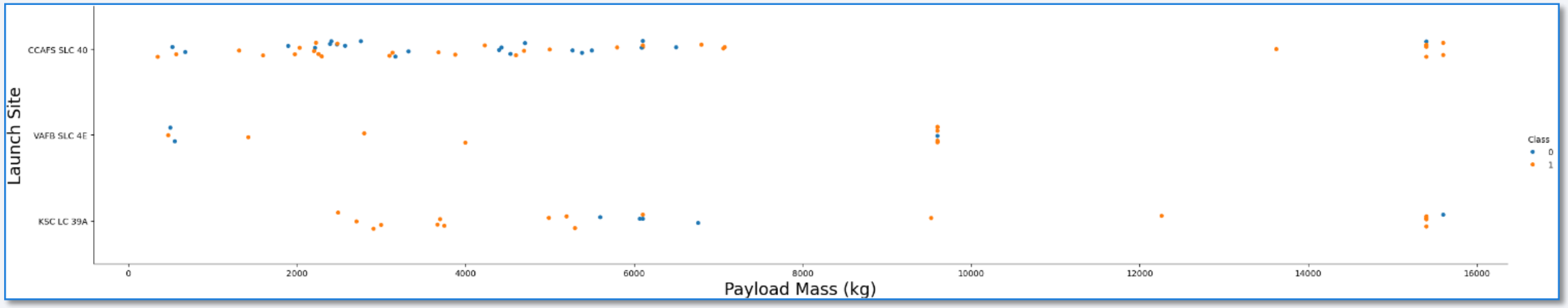

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- Success rates improve as the number of flights increases.
- The earliest flights were unsuccessful, whereas the most recent flights were all successful.
- The CCAFS SLC 40 launch site has for approximately half of all launches.
- Given that VAFB SLC 4E and KSC LC 39A have higher success rates, it can be inferred that each new launch is likely to have a greater chance of success.

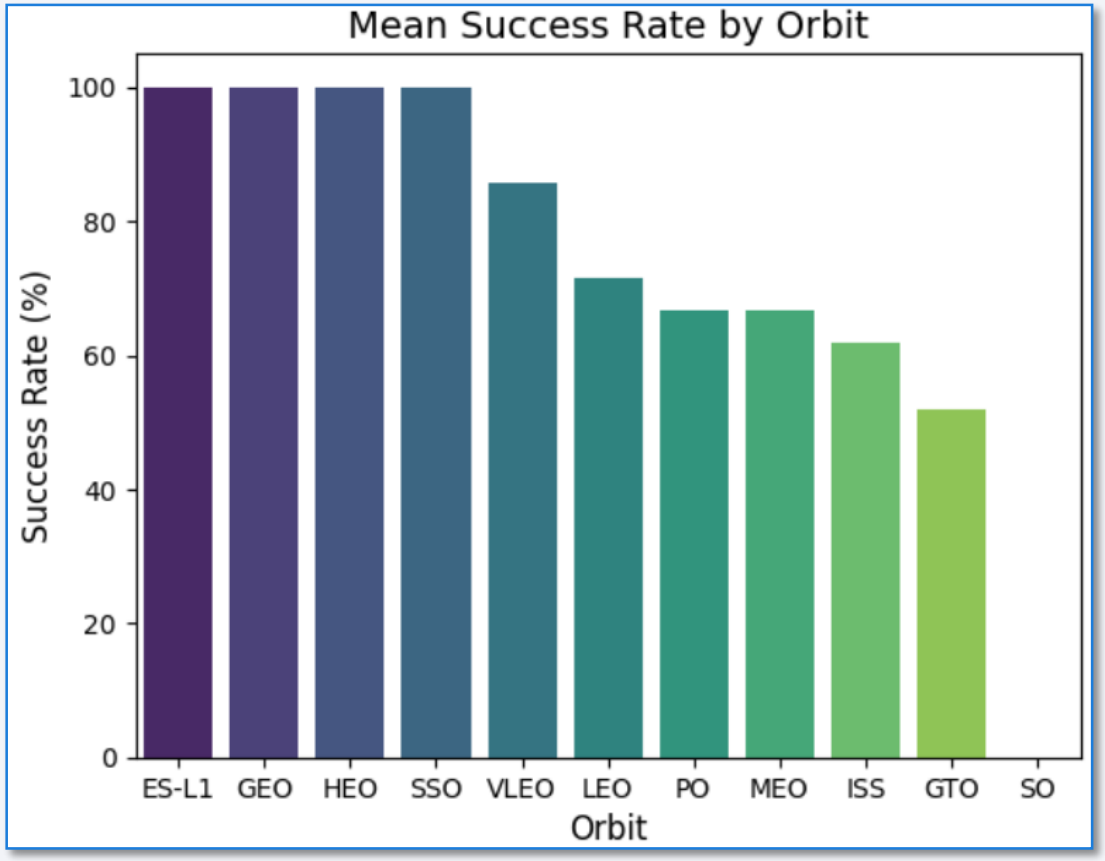
Payload vs. Launch Site



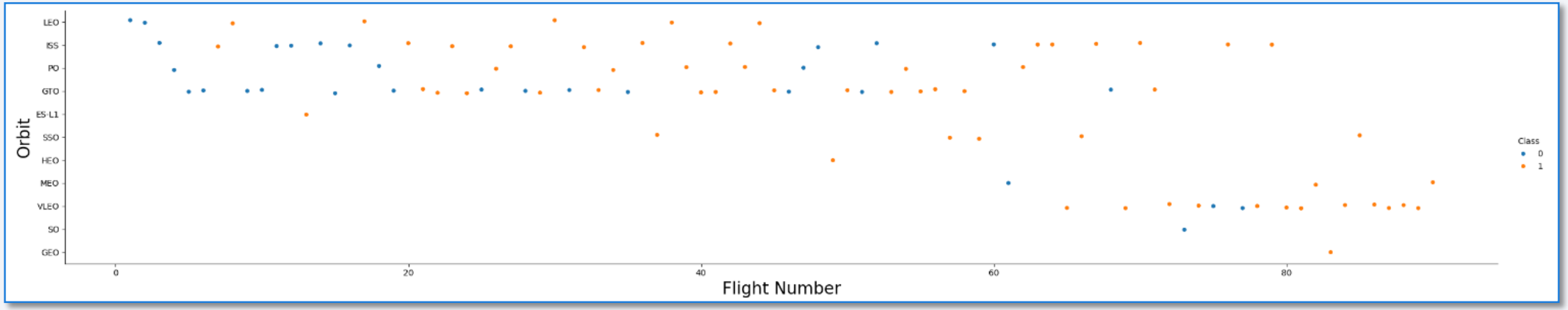
- The VAFB-SLC launch site has no rockets launched for heavy payload mass (greater than 10000).
- At every launch site, a greater payload mass is mostly associated with a higher success rate.
- Majority of launches with a payload mass exceeding 7000 kg were successful.
- The KSC LC 39A with less than 5500 payload mass have 100% success rates.

Success Rate vs. Orbit Type

- The orbit types ES-L1, GEO, HEO, and SSO have a success rate of 100%.
- The orbit type SO has a success rate of 0%.
- The success rate for orbit types such as GTO, ISS, PO, MEO, LEO, and VLEO ranges from 52% to 86%.

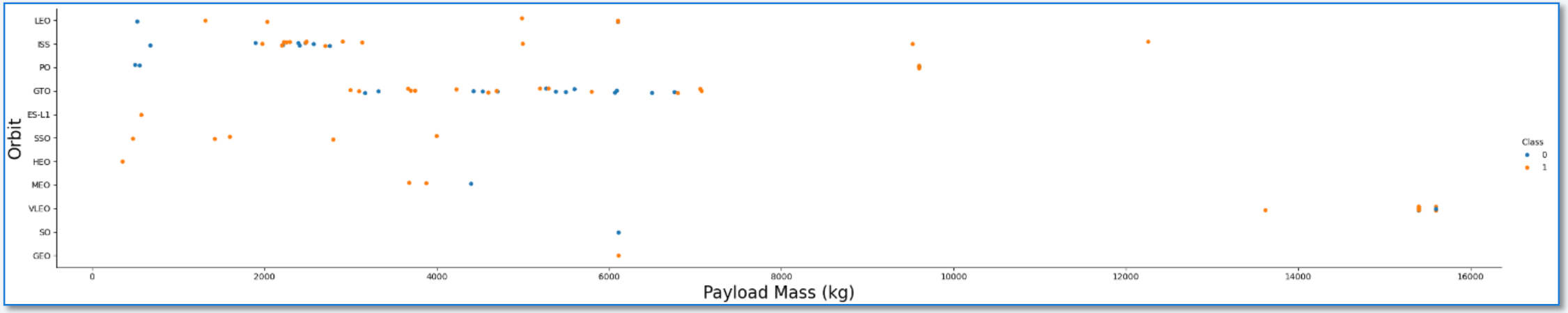


Flight Number vs. Orbit Type



The successful landing rates appear to improve due to the increased number of flights in the LEO orbit. In contrast, it appears to be no correlation between flight numbers and success in the GTO orbit.

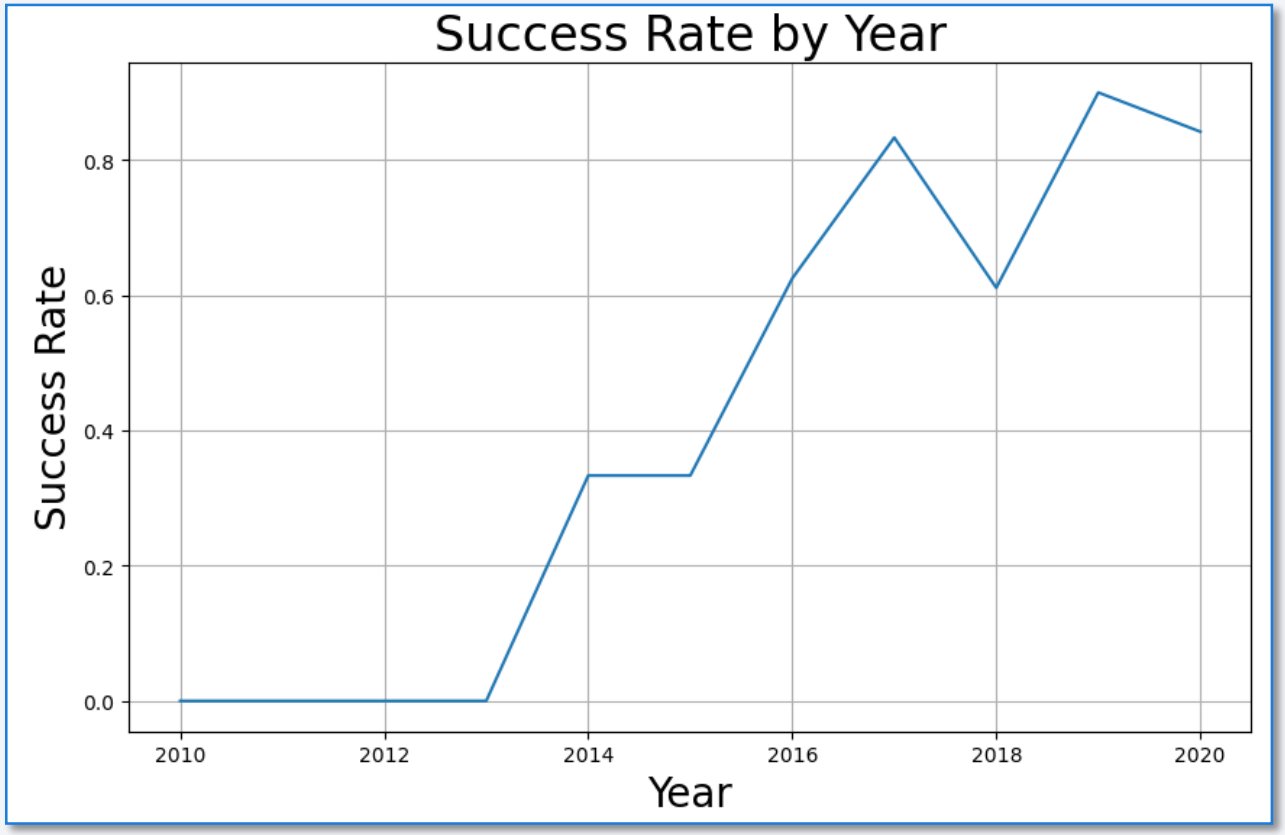
Payload vs. Orbit Type



- Polar, LEO, and ISS have higher successful landing or positive landing rates with heavy payloads.
- However, because both successful and bad landings occur for GTO, it is challenging to differentiate between the two.

Launch Success Yearly Trend

The success rate since 2013 kept increasing till 2020



All Launch Site Names

```
%sql SELECT DISTINCT LAUNCH_SITE as "Launch_Sites" FROM SPACEXTBL;
```

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

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

- Displayed the names of the unique launch sites in the space mission.
- This indicates that there are four launch sites: CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, and CCAFS SLC-40.

Launch Site Names Begin with 'CCA'

```
%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

Displayed 5 records where launch sites begin with the string 'CCA'. This includes columns such as Date, Time, Booster Version, Launch Site, Payload, Payload Mass (kg), Orbit, Customer, Mission Outcome, and Landing Outcomes for launch sites that contain 'CCA.'

Total Payload Mass

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) as "Total Payload Mass (Kg)" FROM 'SPACEXTBL' WHERE Customer = 'NASA (CRS)';
 * sqlite:///my_data1.db
Done.
```

Total Payload Mass (Kg)
45596

Displayed the total payload mass carried by boosters launched by NASA (CRS), totaling 45,596 kg.

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) as "Average Payload Mass (Kg)" FROM 'SPACEXTBL' WHERE Booster_Version LIKE 'F9 v1.1%';
```

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

Average Payload Mass (Kg)
2534.6666666666665

Displayed average payload mass carried by booster version F9 v1.1, which is 2,534.67 kg.

First Successful Ground Landing Date

```
%sql SELECT MIN(DATE) AS "Date of First Successful Landing" FROM 'SPACEXTBL' WHERE Landing_Outcome = "Success (ground pad)";
```

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

Date of First Successful Landing
2015-12-22

Listed the date when the first successful landing outcome in ground pad was achieved, which occurred on December 22, 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT DISTINCT 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

Listed the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT Mission_Outcome, COUNT(*) AS TOTAL_NUMBER FROM 'SPACEXTBL' GROUP BY Mission_Outcome
```

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

Mission_Outcome	TOTAL_NUMBER
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Listed the total number of successful and failure mission outcomes, including 1 failure (in flight), 99 successes, and 1 success with an unclear payload status.

Boosters Carried Maximum Payload

```
%sql SELECT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ IN (SELECT MAX(PAYLOAD_MASS__KG_) AS 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

Listed the names of the booster_versions which have carried the maximum payload mass by utilizing a subquery, resulting in 12 booster versions.

2015 Launch Records

```
%sql SELECT substr(Date, 6,2) as month, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTBL WHERE substr(Date,0,5)='2015' AND Landing_Outcome = "Failure (drone ship)"
```

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

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

- Listed the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
- Since SQLLite does not support month names, case when was used to convert month numbers to month names

```
%sql SELECT CASE substr(Date, 6, 2) WHEN '01' THEN 'January' WHEN '02' THEN 'February' WHEN '03' THEN 'March' WHEN '04' THEN 'April' WHEN '05' THEN 'May' WHEN '06' THEN 'June' WHEN '07' THEN 'July' WHEN '08' THEN 'August' WHEN '09' THEN 'September' WHEN '10' THEN 'October' WHEN '11' THEN 'November' WHEN '12' THEN 'December' END as month_name, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTBL WHERE substr(Date,0,5)='2015' AND Landing_Outcome = "Failure (drone ship)"
```

month_name	Landing_Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

```
%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(*) DESC;
 * 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

Ranked 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.

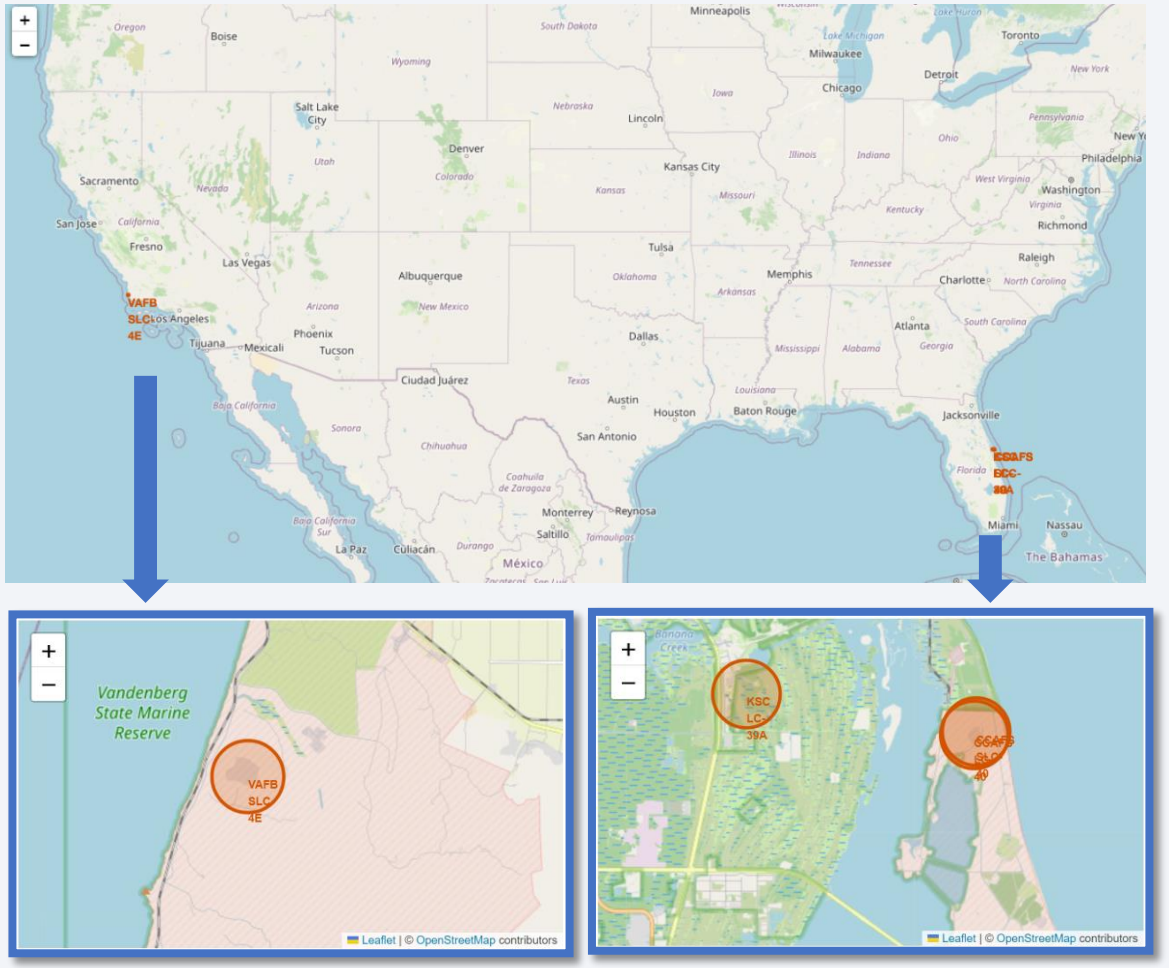
Section 3

Launch Sites Proximities Analysis

Launch Sites Map - SpaceX

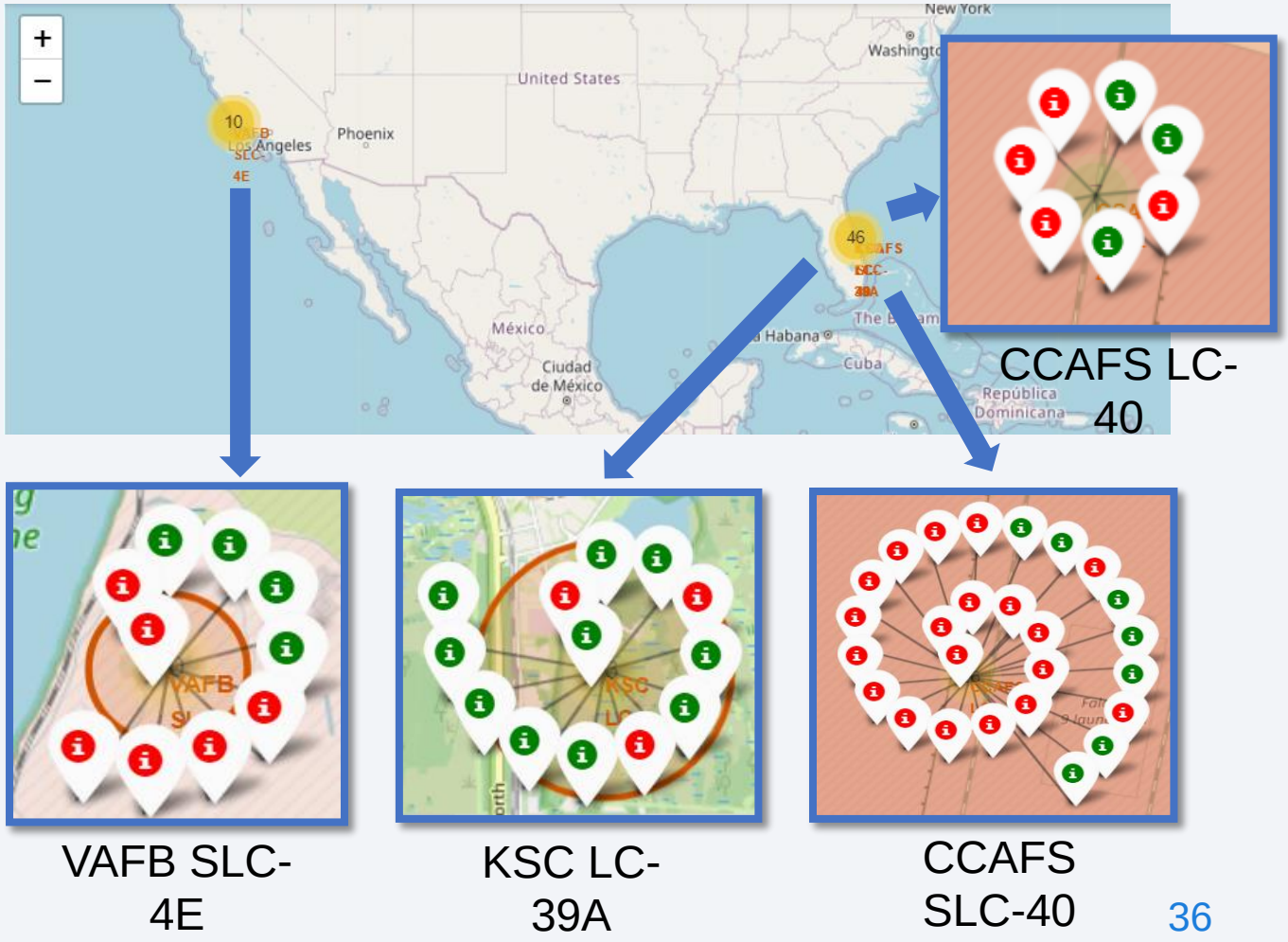
- The highlighted circle area with a text label on a specific coordinate represents Launch sites.

- This indicates that all the launch sites are located near the Equator and are also in very close proximity to the coast.



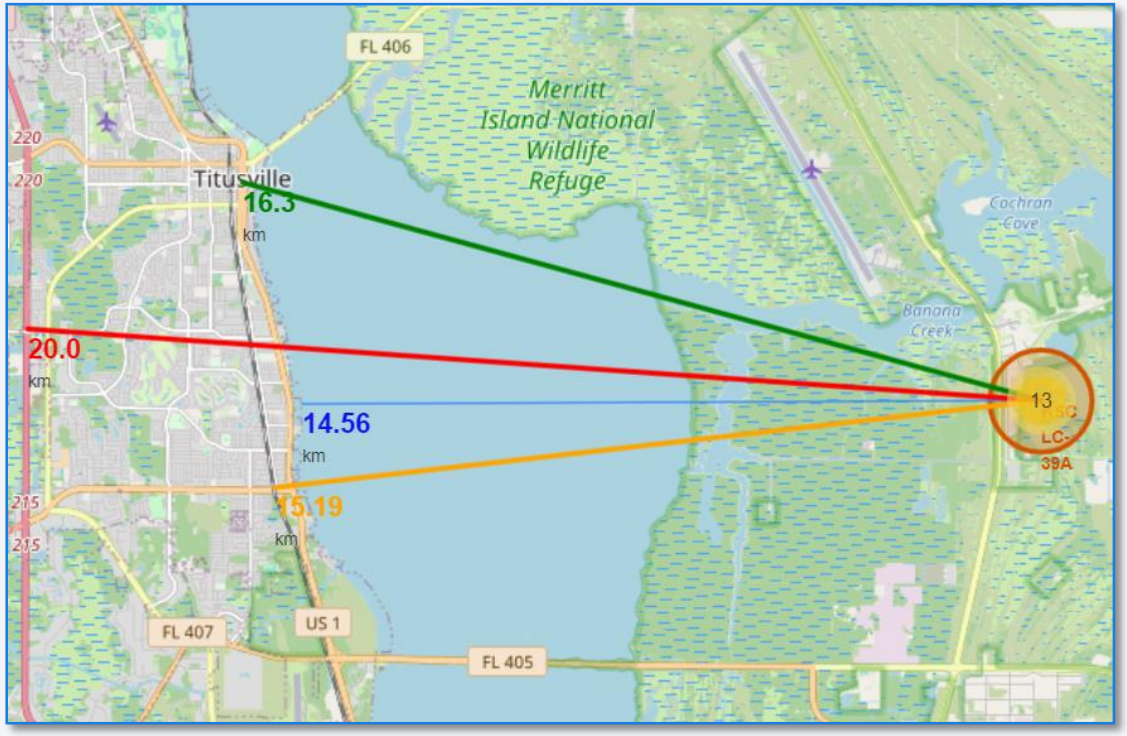
Launch Sites Map for All Success and Failed Launches

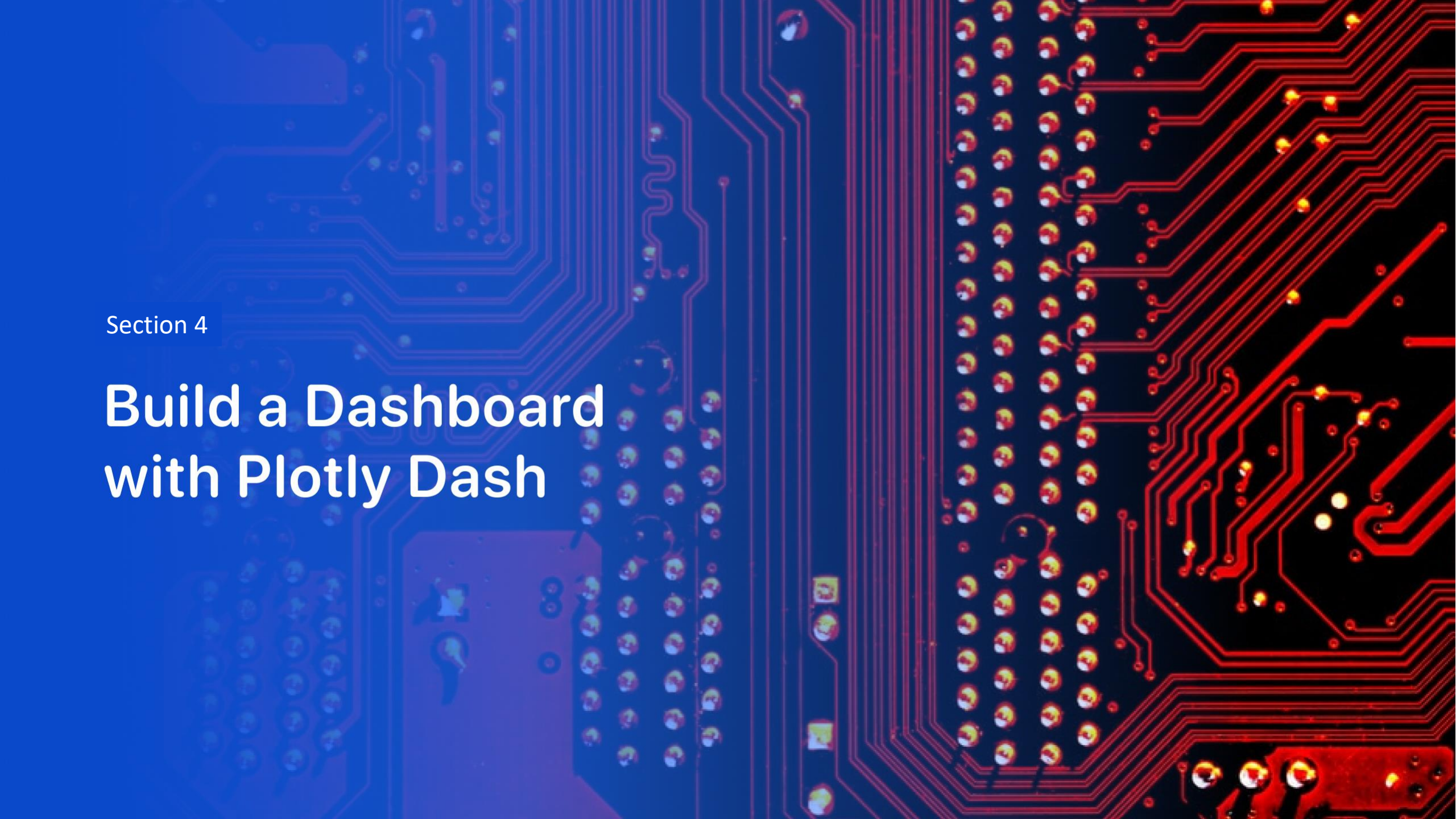
- Colored markers were used to easily identify launch sites with relatively high success rates: Green indicates a successful launch, while Red signifies a failed launch.
- Launch Site KSC LC-39A features a remarkably high success rate.



Launch Site Map Distance to its proximities

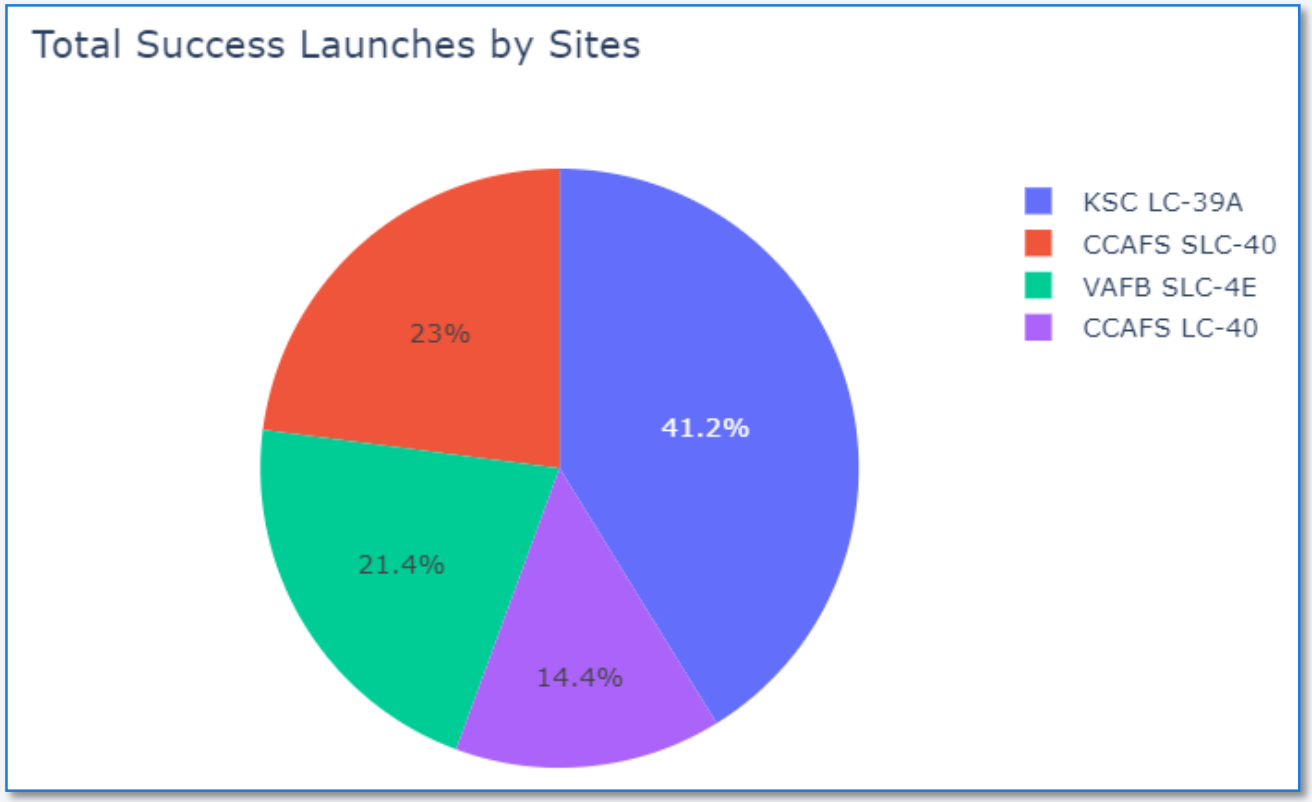
- The screenshot illustrates a sample distance from the KSC LC-39A Launch Site to nearby locations. We can observe that it is:
  - 14.56 km close to the coastline
  - 15.19 km close to the railway
  - 16.3 km close to the nearest city
  - 20 km close to the highway
- It also shows that could cover distances of 14 to 20 km posing a potential danger to populated areas.
- Launch sites are not in close proximity to railways or highways, but they are near coastlines and maintain a certain distance from cities.




Section 4
Build a Dashboard with Plotly Dash

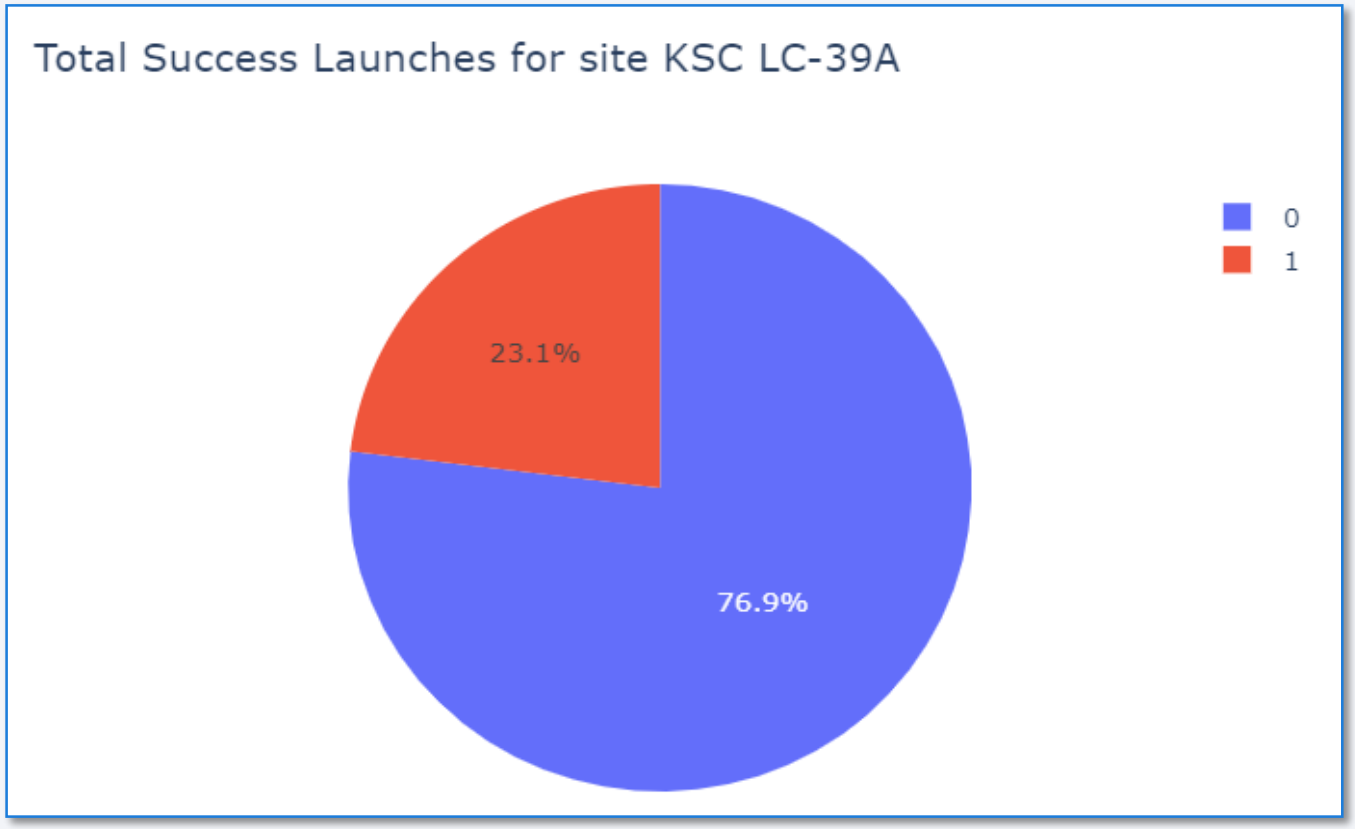
Successful Launches by Site

This pie chart shows that, among the 4 sites, KSC LC-39A has the highest number of successful launches, accounting for 41.2% of total successful launches, while CCAFS LC-40 has the fewest, with 14.4%.

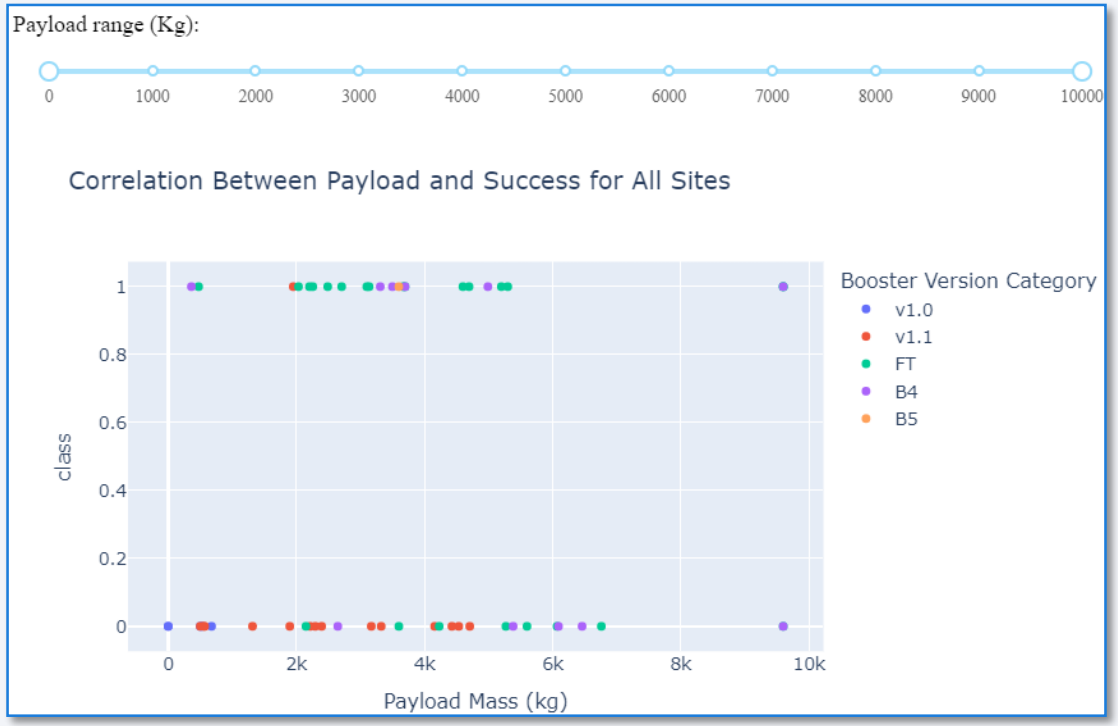


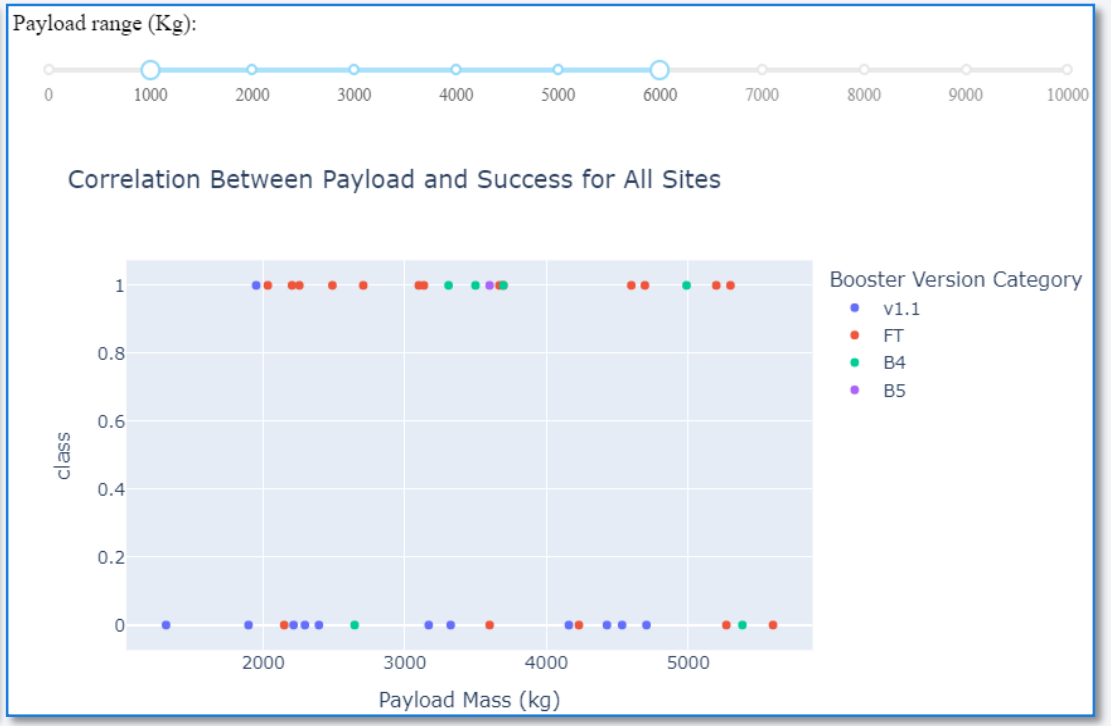
Launch Site with the Highest Success Ratio

The launch site KSC LC-39A boasts the highest success rate at 76.9% (shown in blue, corresponding to a class of 0) and a failure rate of 23.1% (represented in red, corresponding to a class of 1).



Payload vs. Launch Outcome scatter plot for all sites





The charts indicate that payloads ranging from 2000 to 5500 kg achieve the highest success rate.

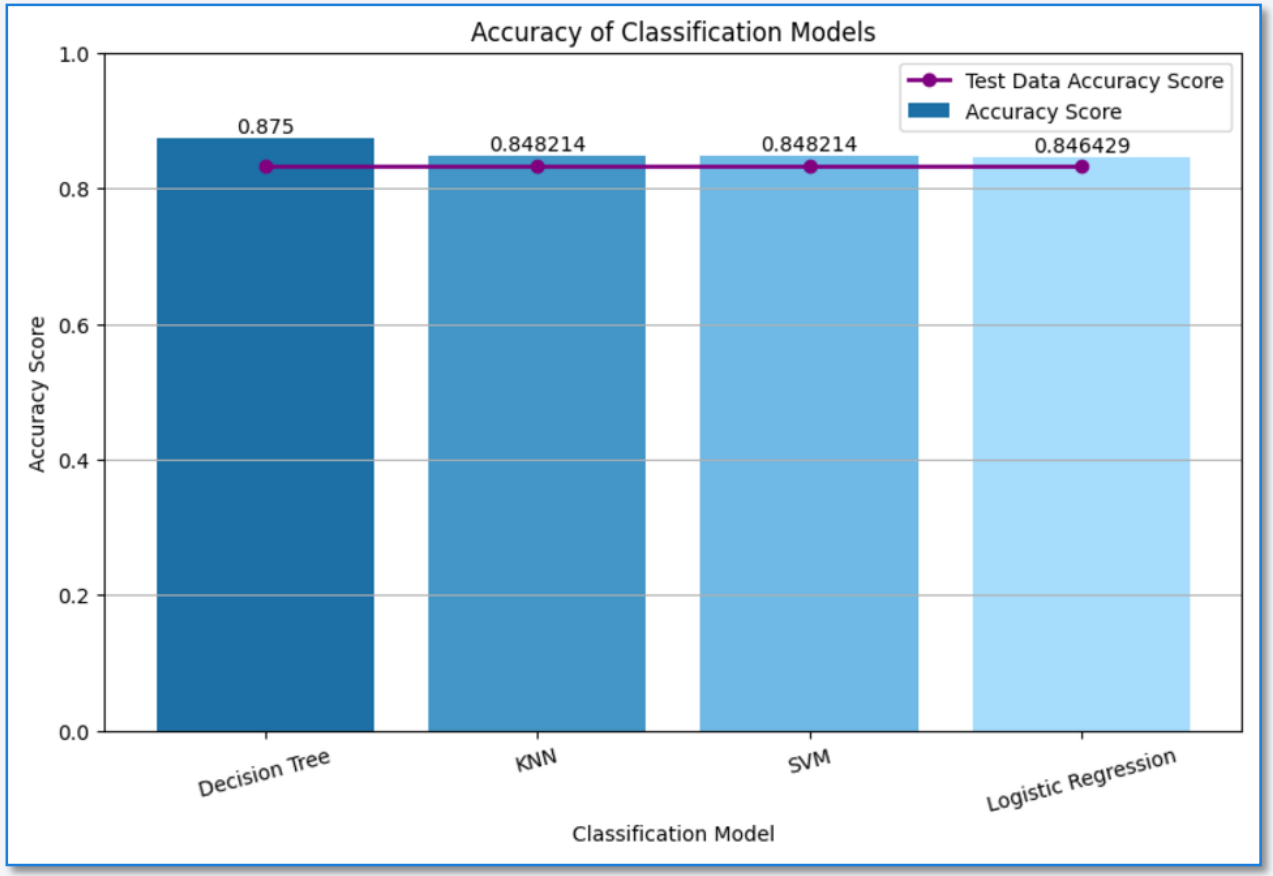
Section 5

Predictive Analysis (Classification)

Classification Accuracy

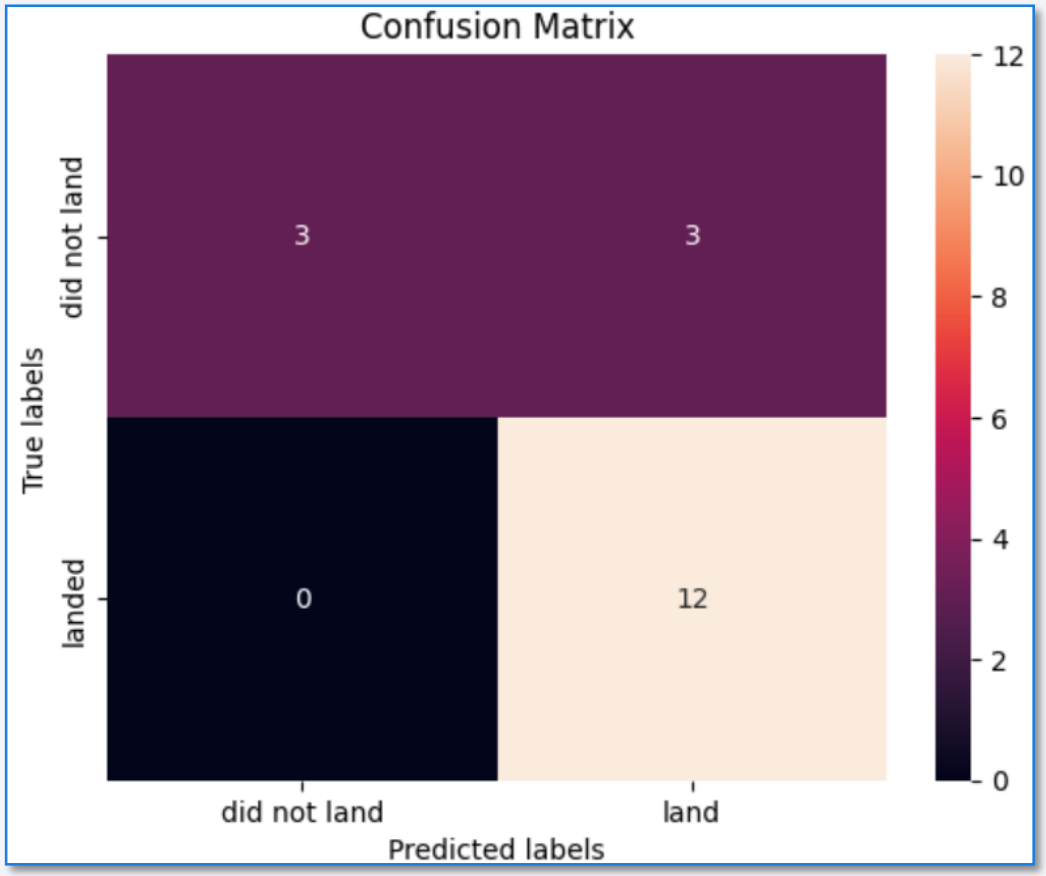
In this analysis, the Decision Tree is the most effective model; however, other models show similar test data accuracy ratings, suggesting that they generalize reasonably well.

Classification Model	Accuracy Score	Test Data Accuracy Score
Decision Tree	0.875000	0.833333
KNN	0.848214	0.833333
SVM	0.848214	0.833333
Logistic Regression	0.846429	0.833333



Confusion Matrix

- The model correctly identified 3 positive outcomes and 12 negative outcomes, with no false positives and 3 false negatives.
- The accuracy score indicates that approximately 83.3% of the predictions are correct.
- All four models have identical confusion metrics with final results of:
  - Precision = 1.0 *(means that all the instances predicted as positive are indeed positive, with no false positives)*
  - Recall = 0.5 *(means that the model correctly identifies 50% of the actual positive instances, with the remaining 50% being missed)*
  - F1 Score = 0.6667 *(mean of precision and recall; moderately performing)*



Conclusions

- Launch site KSC LC 39A has the highest success rates, whereas CCAFS SLC 40 has the lowest.
- The orbit types ES-L1, GEO, HEO, and SSO have a success rate of 100% while the orbit type GTO has the lowest success rate at 52%.
- The success rate since 2013 kept increasing till 2020
- Success rates improve as the number of flights increases.
- Launch sites are not located near railways or highways, but they are situated close to coastlines and maintain a certain distance from cities.
- The optimal algorithm for this dataset is the Decision Tree Model.

Appendix

- The courses taken are part of the IBM Data Science Professional Certificate
- All Python code snippets, SQL queries, charts, notebook outputs, and datasets created during this project were included in the course.

Thank you!