IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Leopoldo Gomez Caudillo
5 January 2028

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection:
    - SpaceX launch data from wikipedia
  - Data wrangling:
    - Clean and store data
  - EDA
    - SQL queries and visualizations
  - Interactive visualizations
    - Map of launch sites and success rate with folium
    - Interactive dashboard with Plotly and Dash
  - Predictive analysis (Classification)
    - Model building and evaluation (LR, SVM, Decision tree, KNN)
    - Tuning hyperparameters
    - Model evaluation
- Summary of all results
  - Data insights
    - It was identified factors correlated with Falcon 9 first stage landings.
    - It was obtained geographical patterns and success rates correlated with Launches
  - Predictive analysis
    - LR, SVM and KNN had 83% of accuracy
    - Decision tree has 94% of accuracy
  - Findings
    - Payload mass affect landing success

Introduction

- Project background and context
  - SpaceX makes space travel affordable due to launches at a significantly lower cost compared to other providers. This savings is because SpaceX can reuse the first stage of the launch. Therefore, by accurately prediction of landing success, it is possible estimate launch costs and obtain valuable feedback.
- Questions
  - Which factors affect the successful landing of the Falcon 9 first stage?
  - Which ML model is the best to predict the landing success of the first stage?

Section 1

Methodology

Methodology

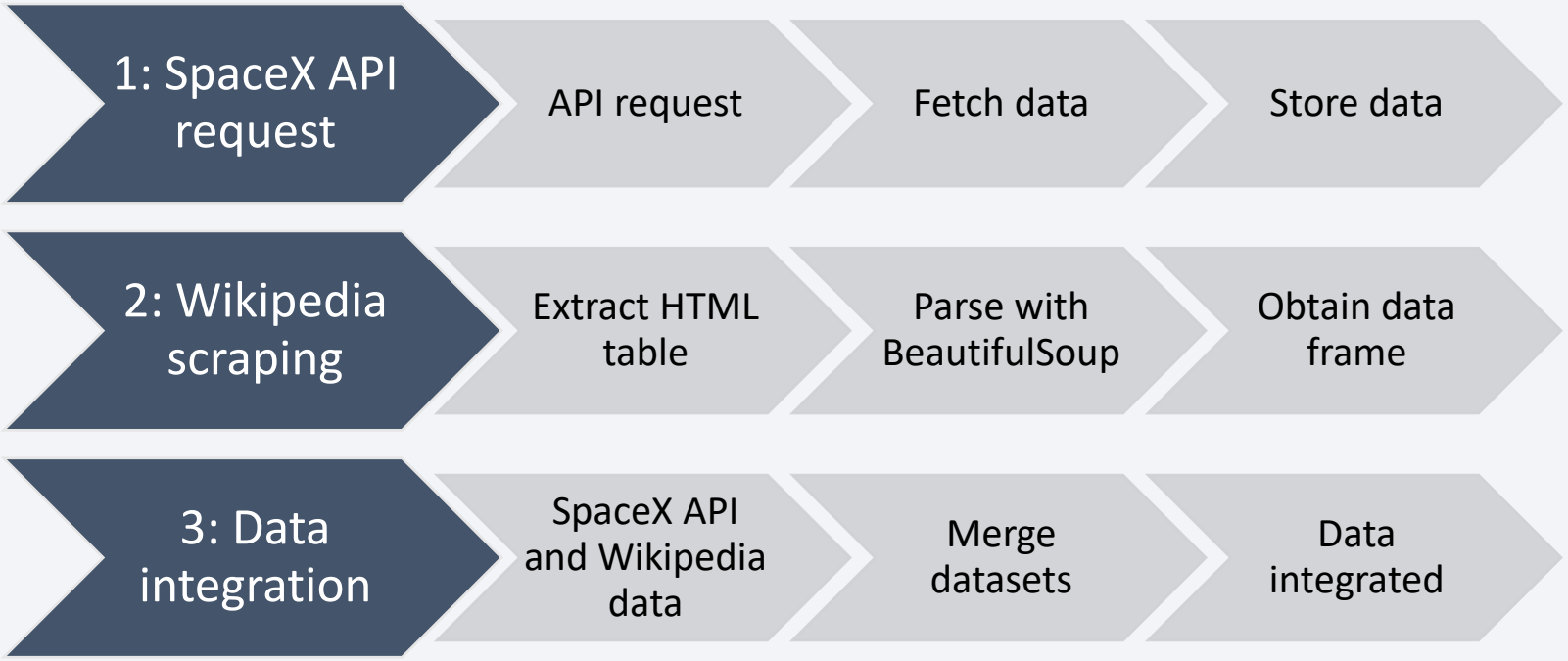
Executive Summary

- Data collection methodology:
  - Data was collected from the SpaceX API using web scraping from Wikipedia.
- Perform data wrangling
  - The data was filtered, cleaned (handling missing values and standardizing formats) and engineered to enrich the dataset.
- Perform exploratory data analysis (EDA) using visualization and SQL
  - SQL queries to obtain insights and answers for specific questions regarding the dataset
  - Visualizations of correlations between launch success rates with launch sites and payloads using Matplotlib and Seaborn

Methodology

- Perform interactive visual analytics using Folium and Plotly Dash
  - Interactive maps for launch sites and autcomes with Folium
  - Interactive dashboard with dropdowns and sliders to analyze the relationship between launch success rates and payload mass, using Plotly and Dash
- Perform predictive analysis using classification models
  - To predict the landing success LR, SVM, KNN and decision tree models were built.
  - Hyperparameter tuning was made with GridSearchCV
  - And evaluation models was made based their performance and accuracy

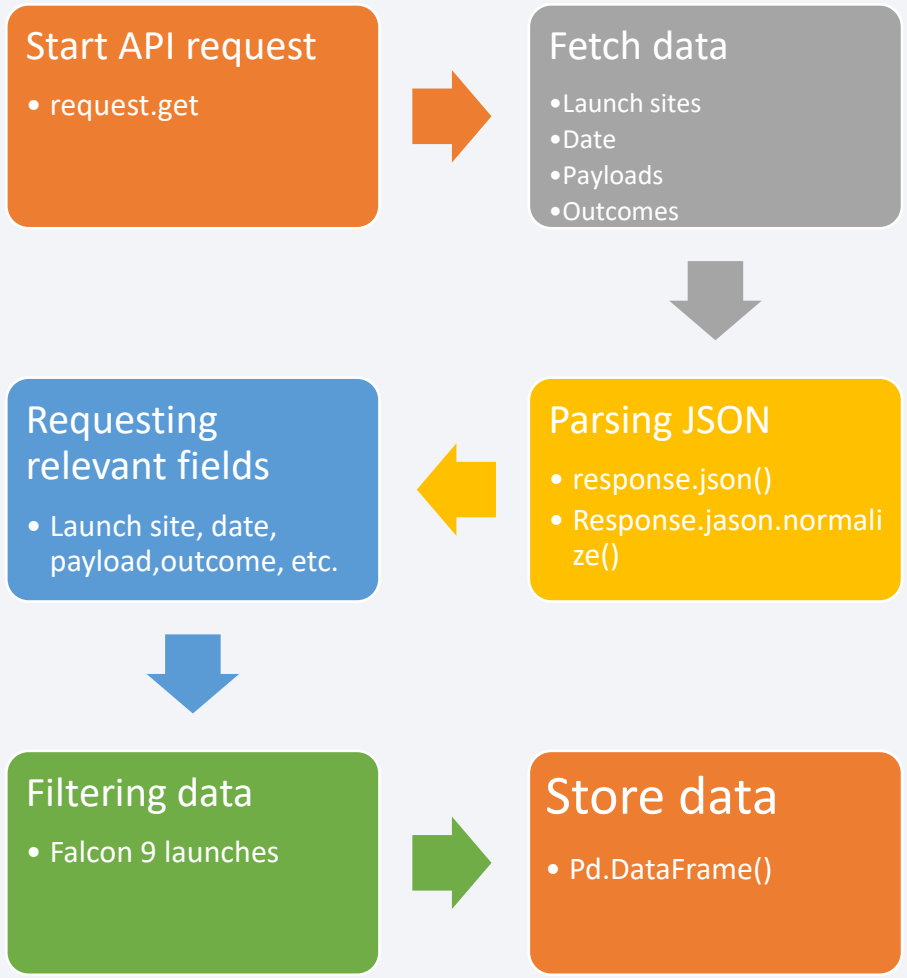
Data Collection



Data Collection – SpaceX API

- 1: API request
  - Connection to the SpaceX API ("https://api.spacexdata.com/v4/launches") through Python's "requests" library
- 2: Decoding the response
  - Conversion from JASON to Python dictionary
  - Extract relevant fields (launch site, date, payload mass, rocket type, outcome, etc.)
- 3: Data filtering
  - Store the data into a Pandas DataFrame
- 4: Data store
  - pd.DataFrame()

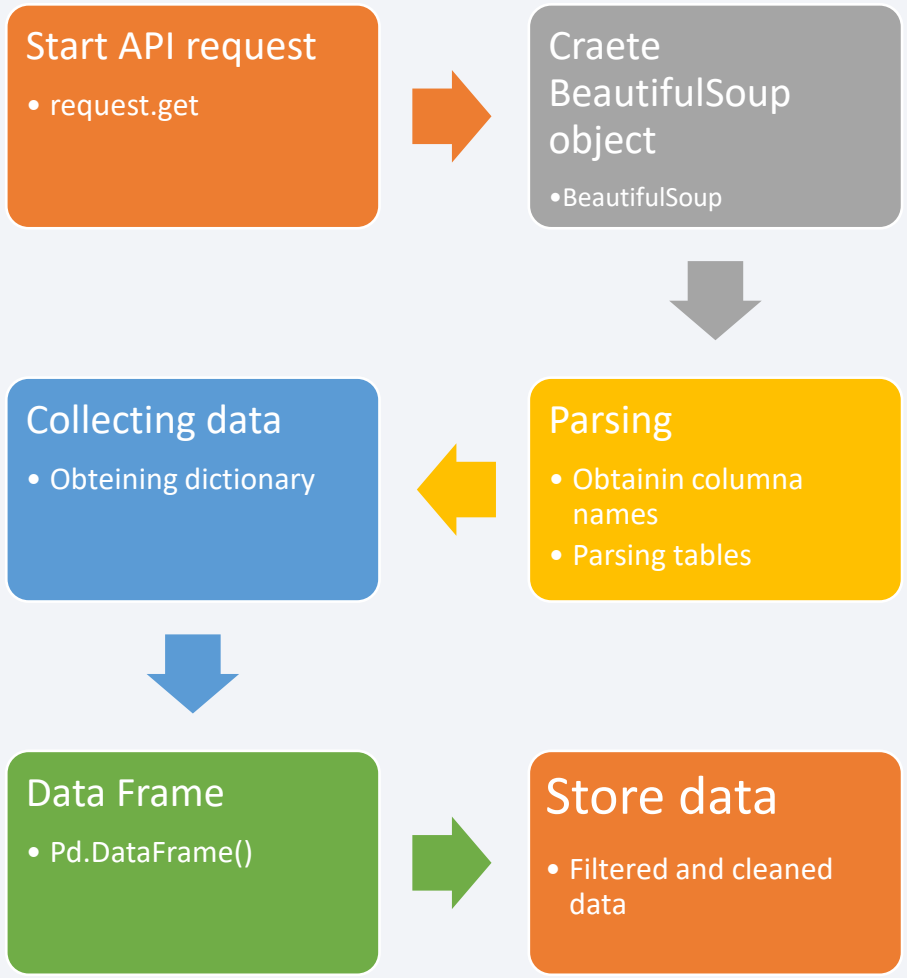
GitHub: https://github.com/pologoca/IBM-Applied-Data-Science-Capstone/blob/main/1-spacex-data-collection-api.ipynb



Data Collection - Scraping

- 1: Web Scraping
  - Request from Wikipedia page https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- 2: Parsing HTML Content
  - Creating aBeautifulSoup object to parse the HTML content and extracts the HTML table containing Falcon 9 data launch records
- 3: Collectig data
  - Obteining dictionaries from the parsed table
  - Creating DataFrame from the dictionary
- 4: Data store
  - Filtering and cleaning the data

Github: https://github.com/pologoca/IBM-Applied-Data-Science-Capstone/blob/main/2-webscraping-data-collection.ipynb



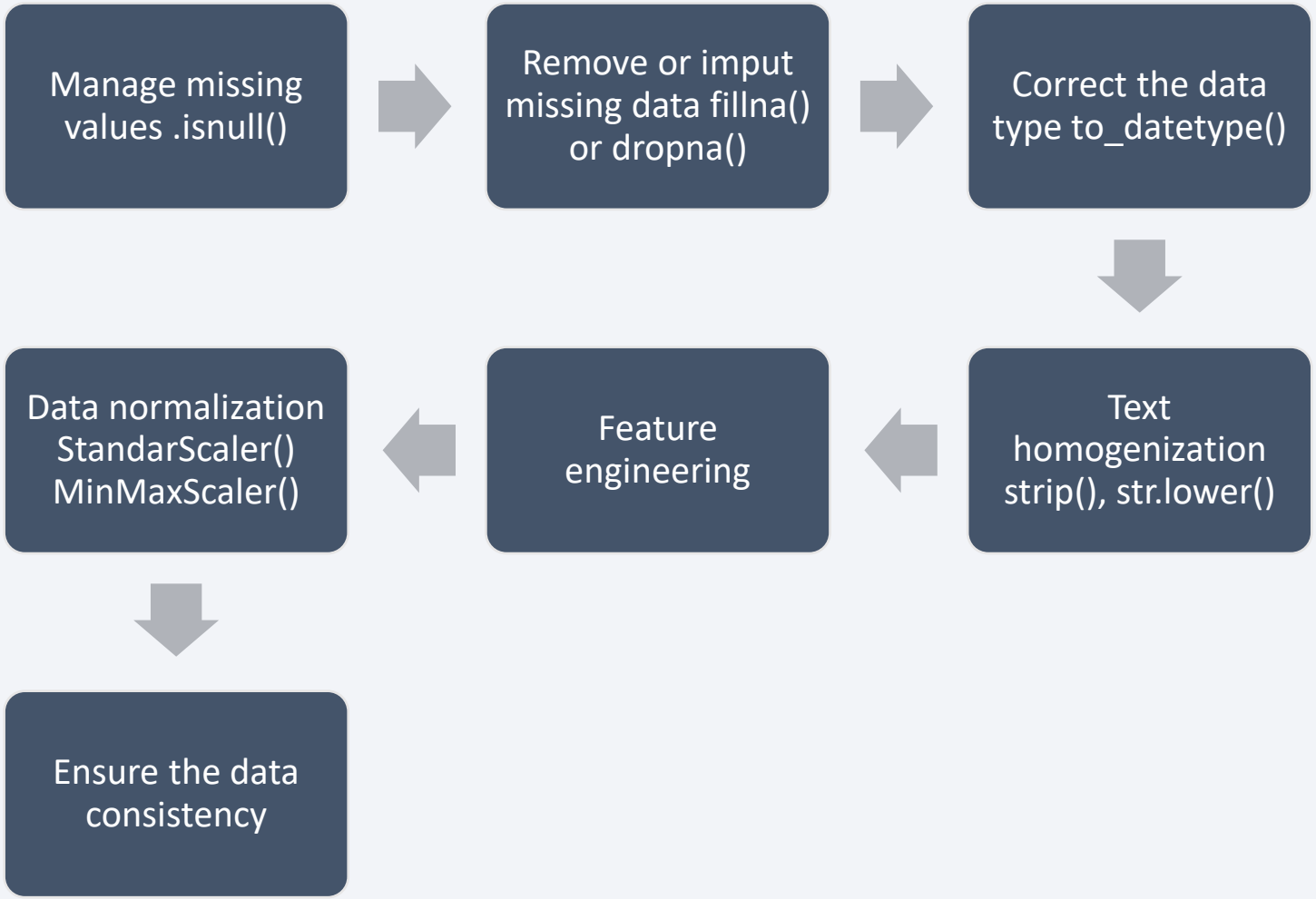
Data Wrangling

Data wrangling refers cleaning, transforming and organizing messy data into structured data for analysis

1. Data cleaning involves handling missing and inconsistent data
2. Data transformation involves
   - Homogenize the text
   - Put the data in correct formats
   - Create new features (variables)
   - Normalize values to ensure consistency
3. Data organizing refers to integrating and validating data into tables for analysis

GitHub: https://github.com/pologoca/IBM-Applied-Data-Science-Capstone/blob/main/3-spacex-data%20wrangling.ipynb

Data Wrangling



EDA with Data Visualization

Exploratory Data Analysis (EDA) includes tables and visualizations for summarize and describes patterns to understand the data's distribution and relationships between features.

Performed visualizations

- Scatter plot: to identifies relationships between two numerical variables
  - Flight number vs. Pay load mass
  - Flight number vs. Launch site
  - Flight number vs. Orbit
  - Launch site vs. Pay load mass
  - Pay load mass vs. Orbit
- Bar plot: to compares frequencies between categorical features
  - Orbit vs. Success rate
- Line plot: to identifies temporal patterns between numerical features
  - Pay load mass vs. Orbit

GitHub: https://github.com/pologoca/IBM-Applied-Data-Science-Capstone/blob/main/4-EDA-data-visualization.ipynb

EDA with SQL

Performed SQL queries:

- Select the unique launch sites in the space mission
- Select 5 records where launch sites begin with the string ‘CCA’
- Select the total payload mass carried by boosters launched by NASA (CRS)
- Select average payload mass carried by booster version F9 v1.1
- Select the date of the first successful landing outcome in ground
- Select the boosters which have success in drone ship and have payload mass greater than 4000 and less than 6000
- Select the total number of successful and failure mission outcomes
- Select the booster versions which have carried the maximum payload mass
- Select the failed landing outcomes in drone ship, their booster versions and launch site for the months in year 2015
- Order the count of landing outcomes (failure or success) between the date 04/06/2010 and 20/03/2017 in descending order

GitHub: https://github.com/pologoca/IBM-Applied-Data-Science-Capstone/blob/main/5-EDA-SQL.ipynb

Build an Interactive Map with Folium

- Markers
  - Markers was used to pinpoint the location of NASA Johnson Space Center for center the map into the window and to locate all launch sites on the map generated
- Circles
  - Circles was used to visualize the potential influence area around the launch sites
- Lines
  - Lines was used to connect launch sites with other locations, like coastline

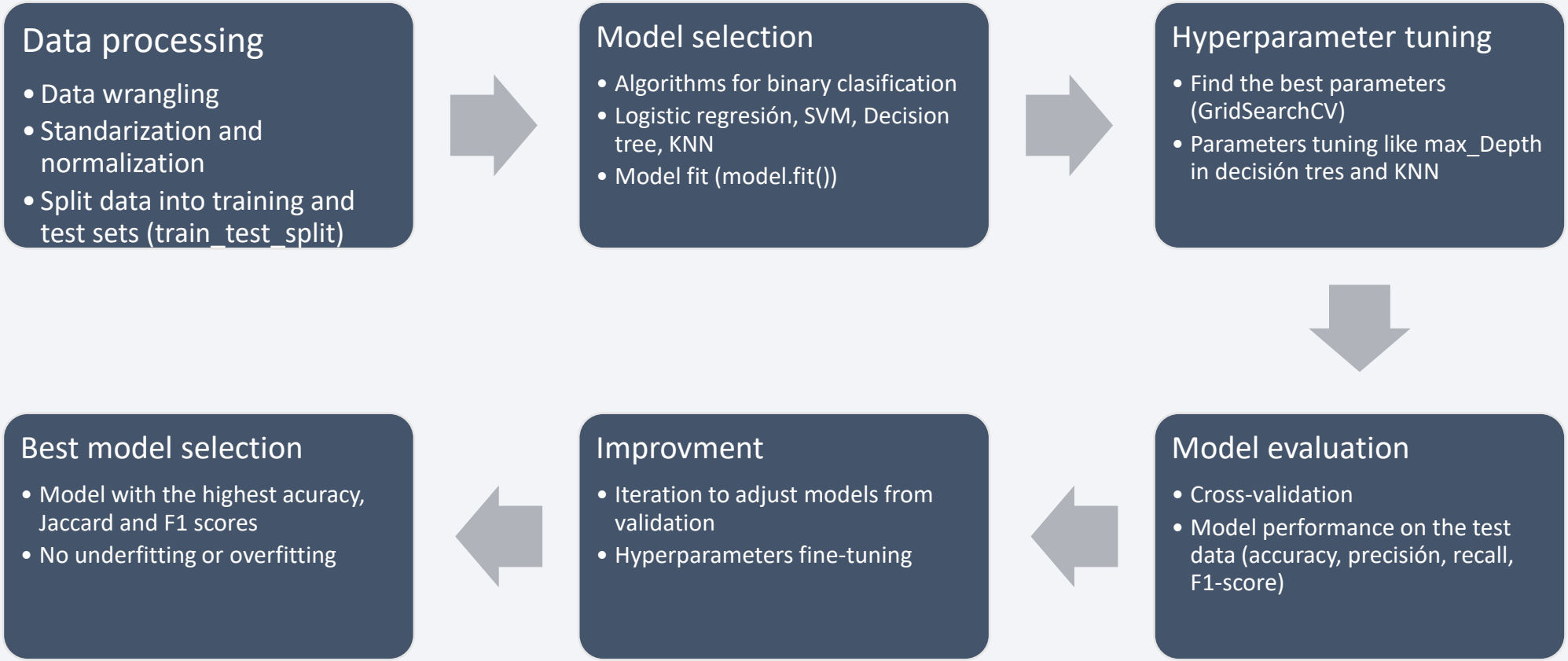
GitHub: https://github.com/pologoca/IBM-Applied-Data-Science-Capstone/blob/main/6-launch-site-locations.ipynb

Build a Dashboard with Plotly Dash

- Dropdown object:
  - Implemented to list launch sites for select one
- Pie chart for success launches:
  - Obtained to show the total successful launches for all sites and the success vs. failed events for in a specific launch site
- Slider object:
  - Implemented to select a pay load mass range
- Scatter plot of pay load mass vs. success rate:
  - Obtained to show the correlation between pay load mass and launch success

GitHub: GitHub: https://github.com/pologoca/IBM-Applied-Data-Science-Capstone/blob/main/7-spacex_dash_app.py

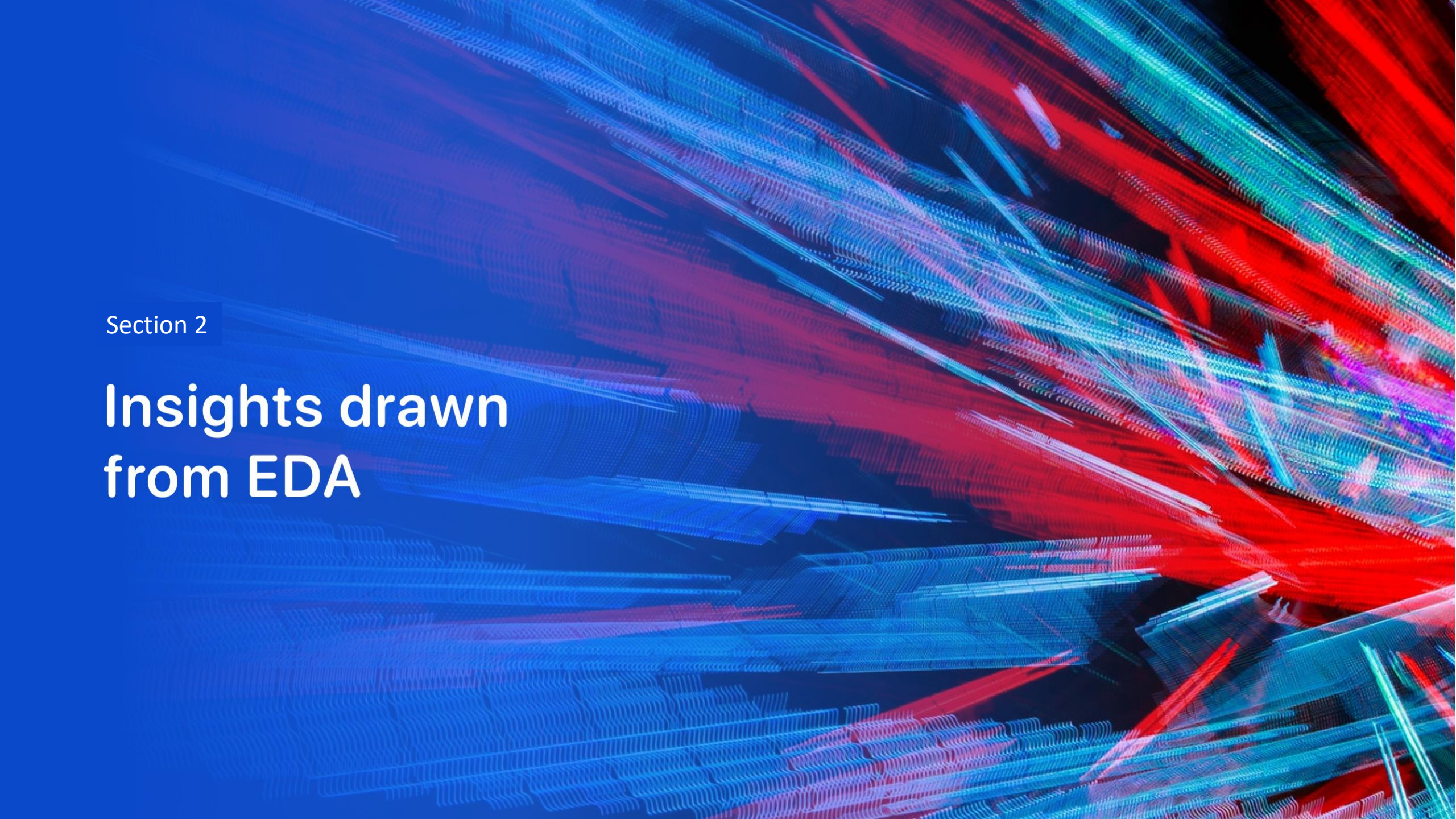
Predictive Analysis (Classification)



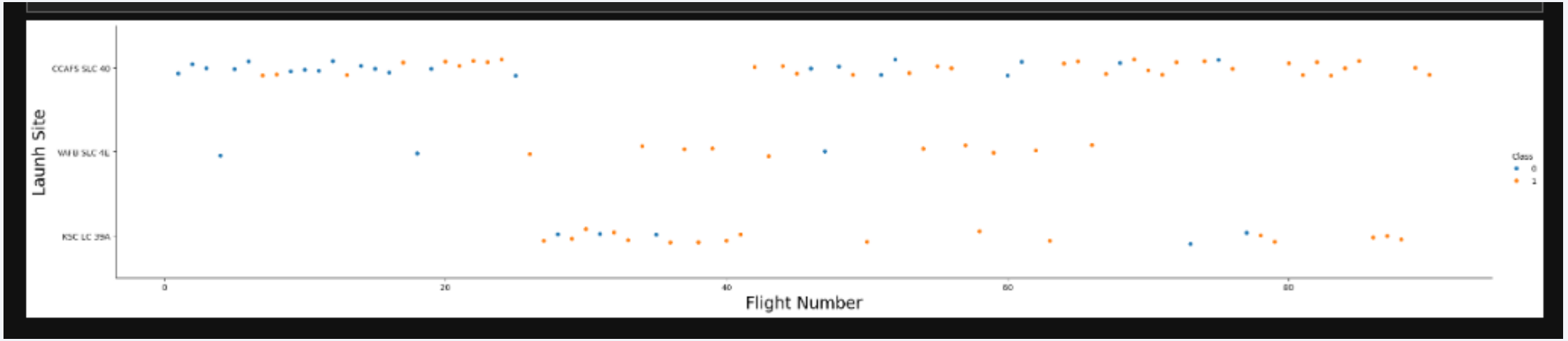
GitHub: https://github.com/pologoca/IBM-Applied-Data-Science-Capstone/blob/main/8-spacex-machine-learning-prediction.ipynb

Results

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


Section 2
Insights drawn from EDA

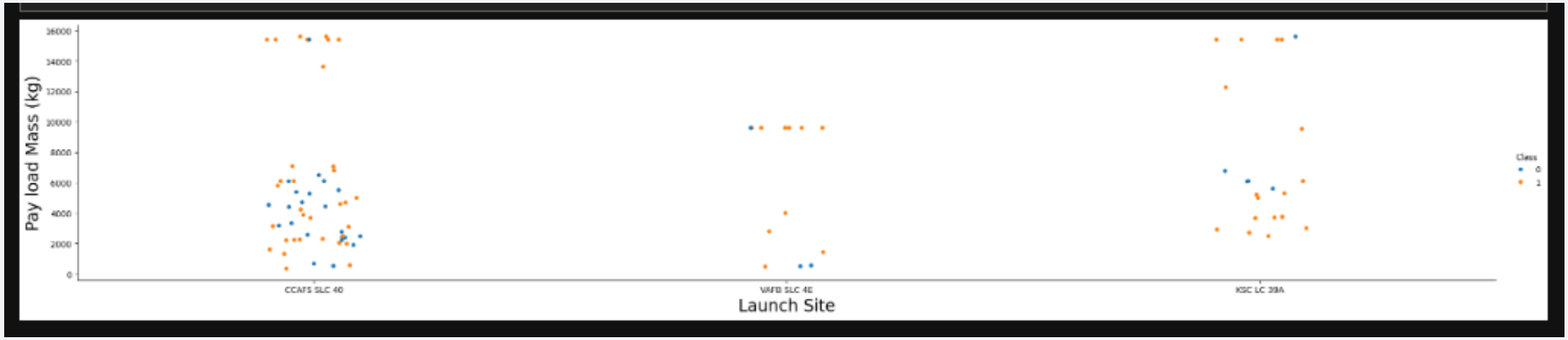
Flight Number vs. Launch Site



Explanation:

- Most of the first launches failed, while most of the later ones were successful
- VAFB SLC 4E and KSC LC 39A have higher success rates than CCAFS SLC 40
- It can be concluded that new launches have higher rates of success.

Payload vs. Launch Site

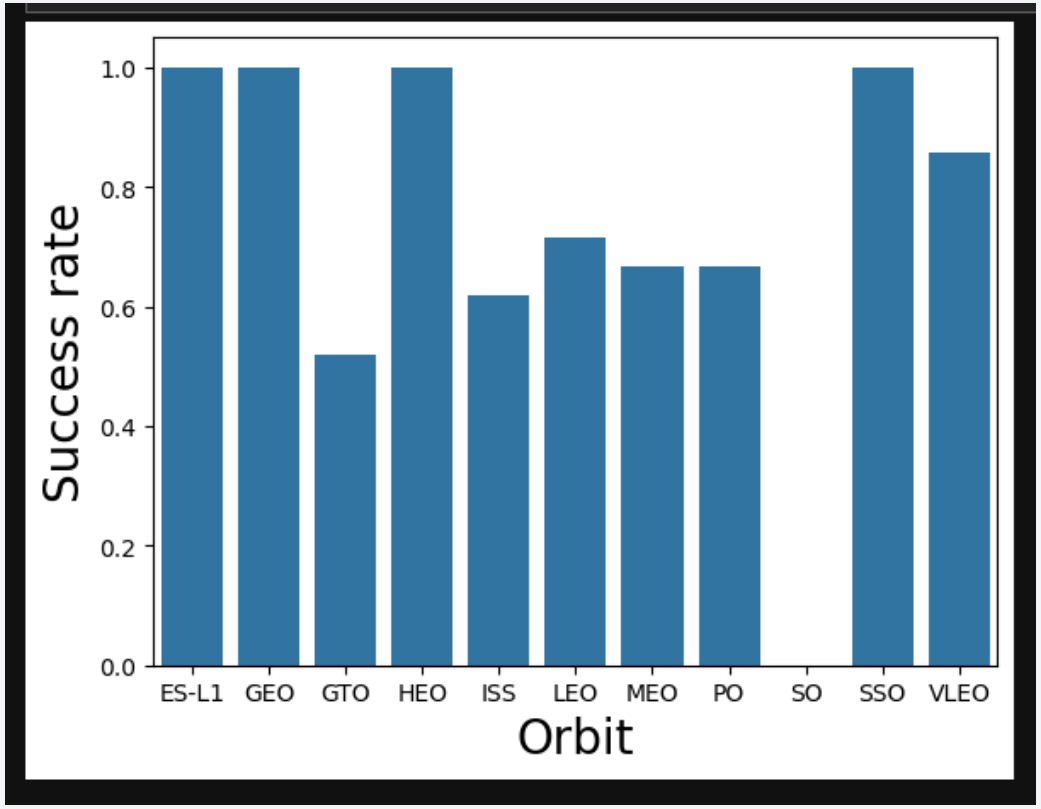


Explanation

- All launches from the VAFB SCL 4E site handle payloads below 10,000 kg

- For all launch sites the higher the payload mass, the higher the success rate

Success Rate vs. Orbit Type



Explanation

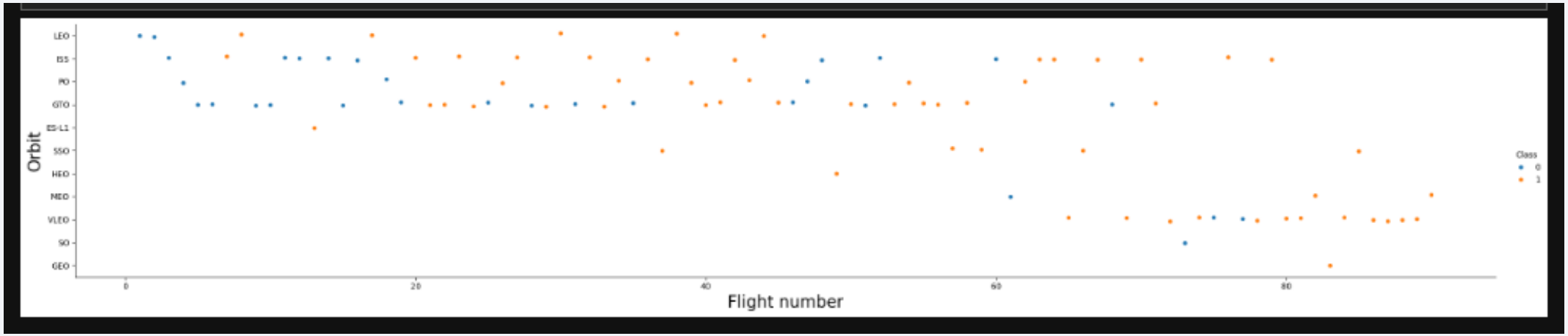
- ES-L1, GEO, HEO, SSO and VLEO show high success rates

- SO, GTO, ISS, MEO, PO and LEO show lower success rates

- All launches from ES-L1, GEO, HEO, SSO are successful

- SO has no successful launches

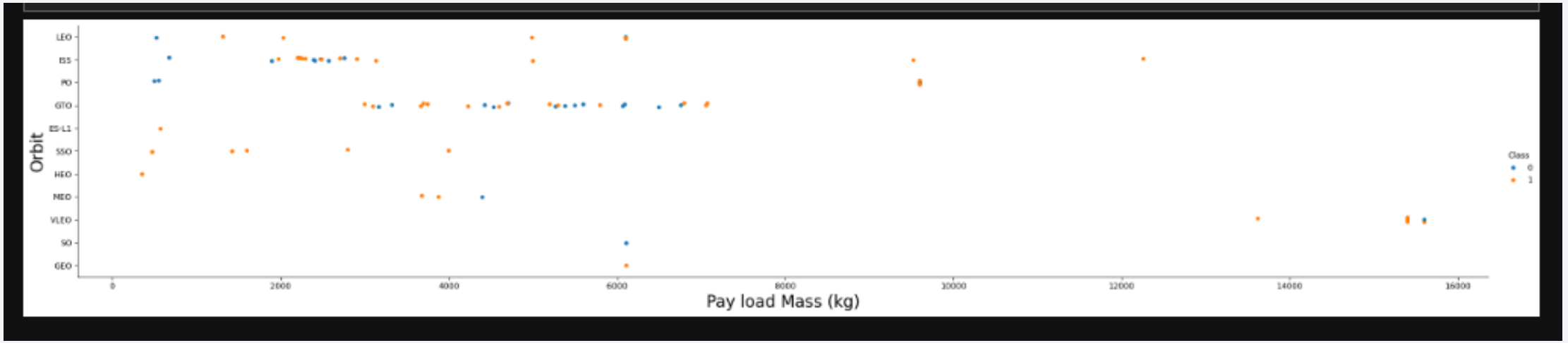
Flight Number vs. Orbit Type



Explanation

- In LEO orbit success appears related to the number of flights
- In ISS, PO and GTO orbits, it seems to be no relationship between flight and orbit
- The other orbits show late flights, most of them are successful

Payload vs. Orbit Type

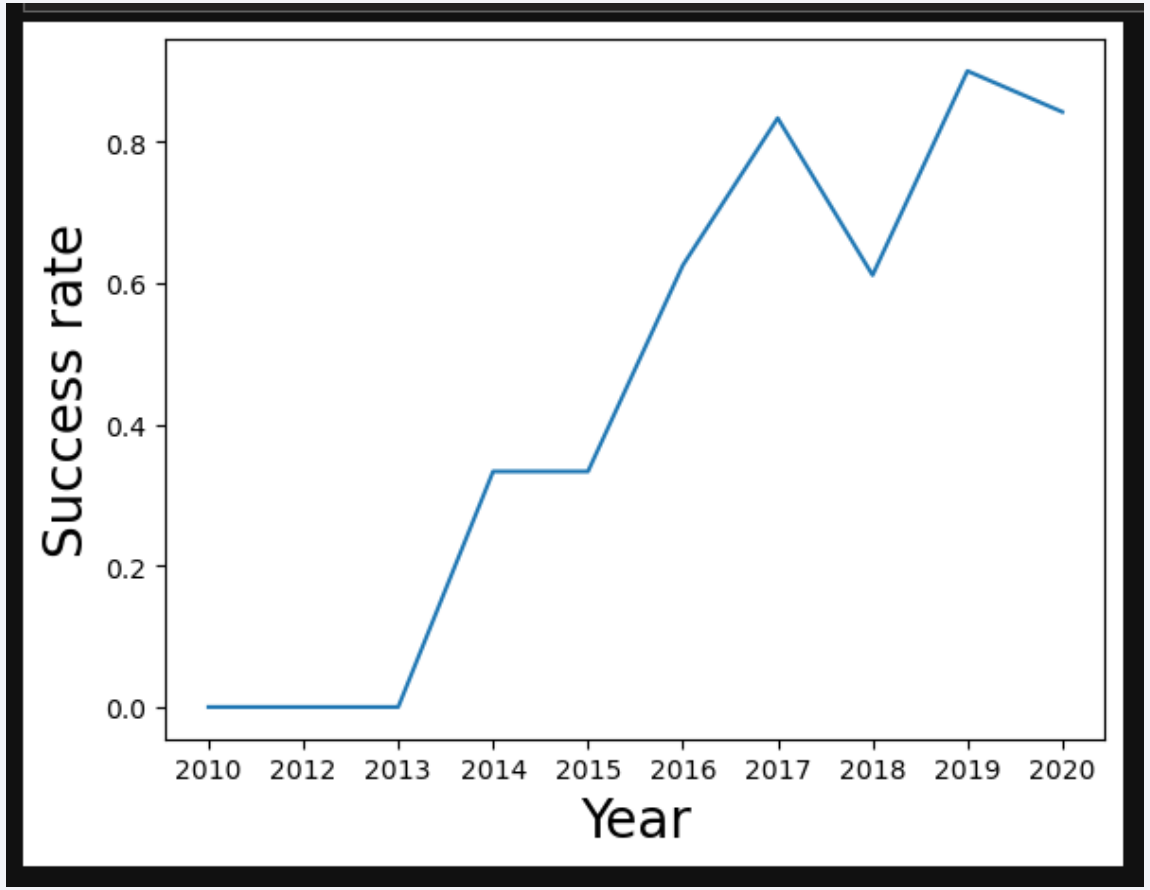


Explanation

- Heavy pay load mass is positive related with successful landing for PO, LEO and ISS orbits

- And negative correlation with GTO orbit

Launch Success Yearly Trend



Explanation

- Successes increased from 2013 to 2017.

- Since 2017, they seem to have stabilized.

All Launch Site Names

Task 1

Display the names of the unique launch sites in the space mission

```
[10]: %sql select distinct(LAUNCH_SITE) from SPACEXTBL
```

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

[10]:

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

The query list the unique name of the launch sites

Launch Site Names Begin with 'CCA'

Task 2

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

[11]:
```
%sql select * from SPACEXTBL where LAUNCH_SITE like 'CCA%' limit 5
```

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

[11]:

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

The query list five records where launch sites begin with “CCA”

Total Payload Mass

Task 3

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

```
[12]: %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where CUSTOMER = 'NASA (CRS)'
```

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

[12]:

sum(PAYLOAD_MASS__KG_)
45596

The query display the total pay load mass carried by boosters launched by NASA (CRS)

Average Payload Mass by F9 v1.1

Task 4

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

```
[13]: %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where BOOSTER_VERSION = 'F9 v1.1'
```

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

[13]:

avg(PAYLOAD_MASS__KG_)
2928.4

The query display the average pay load mass carried by boosters F9 v1.1

First Successful Ground Landing Date

Task 5

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

Hint:Use min function

```
[35]: %sql select min(date) as Date from SPACEXTBL where "Landing_Outcome" = 'Success (ground pad)'
```

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

[35]:

Date
2015-12-22

The query display the date for the first successful landing in ground pad

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
[36]: %sql select BOOSTER_VERSION from SPACEXTBL where "Landing_Outcome" = 'Success (drone ship)' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__
```

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

[36]:

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

The query lists the boosters successful in drone ship and pay load mass greater than 4000 and less than 6000 kg.

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

```
[37]: %sql select count(MISSION_OUTCOME) from SPACEXTBL where MISSION_OUTCOME = 'Success' or MISSION_OUTCOME = 'Failure (in flight)'
 * sqlite:///my_data1.db
Done.
[37]: count(MISSION_OUTCOME)
                          99
```

The query displays the number of successful and failure missions.

Boosters Carried Maximum Payload

Task 8

List all the booster_versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function.

```
[19]: %sql select BOOSTER_VERSION from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
 * sqlite:///my_data1.db
Done.
```

[19]:

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

The query lists boosters that have carried the maximum pay load mass.

2015 Launch Records

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.

```
[40]: %sql select substr(date, 6,2) as Month BOOSTER_VERSION, LAUNCH_SITE from SPACEXTBL where substr(date,0,5)='2015' and "Landing_Outcome" = 'F
```

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

[40]:

month	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

The query lists the month numbers, booster version and launch site for the failed landing outcome in 2015.

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

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.

[42]:
```
%sql select "Landing_Outcome", count("Landing_Outcome") as Landing_Count from SPACEXTBL \
    where date between '2010-06-04' and '2017-03-20' group by "Landing_Outcome" order by count("Landing_Outcome") desc
```

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

[42]:

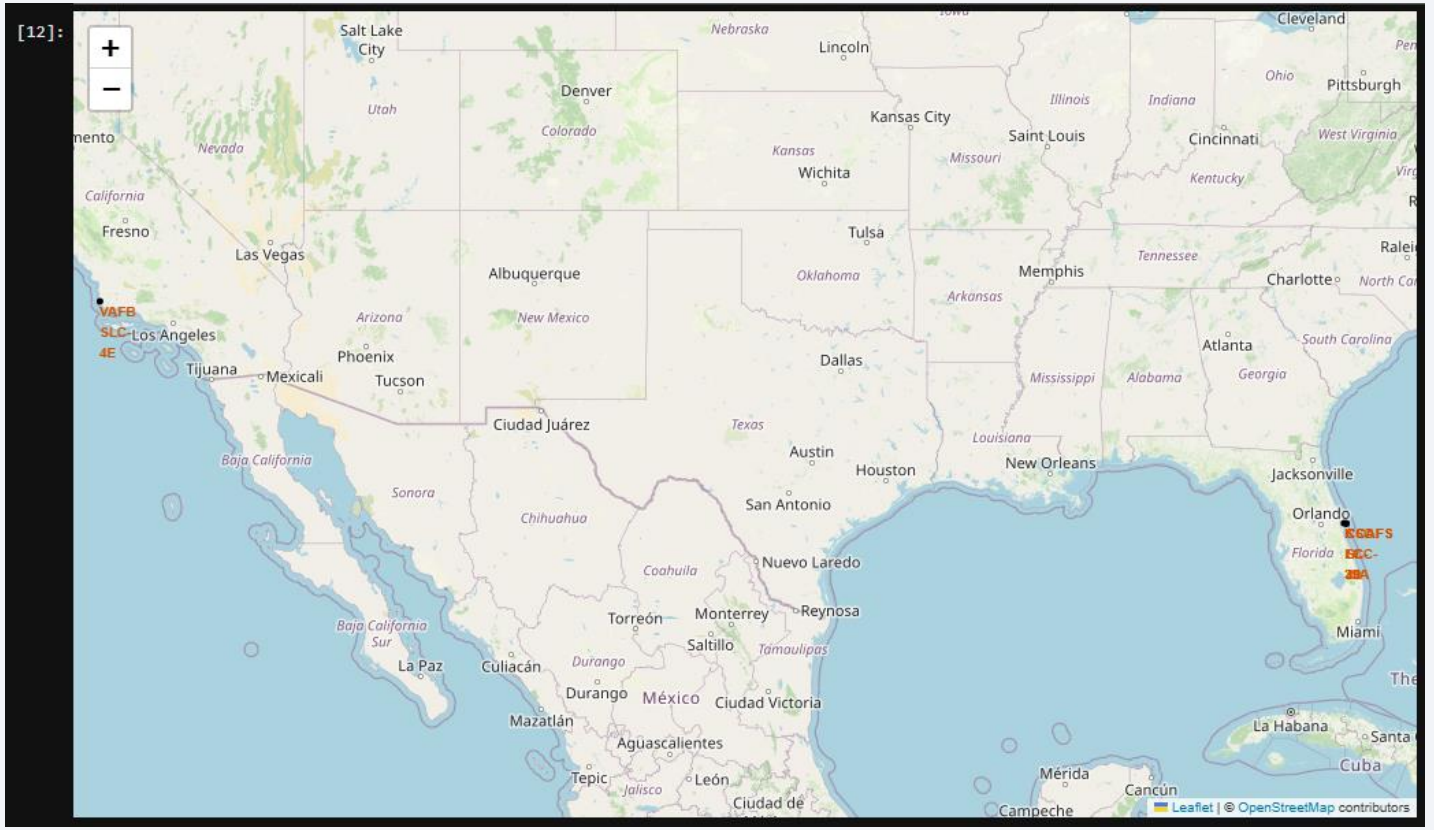
Landing_Outcome	Landing_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

The query displays the ranking of landings that have drone ship failure or group pad success between 04/06/2010 and 20/03/2017.

Section 3

Launch Sites Proximities Analysis

Launch sites



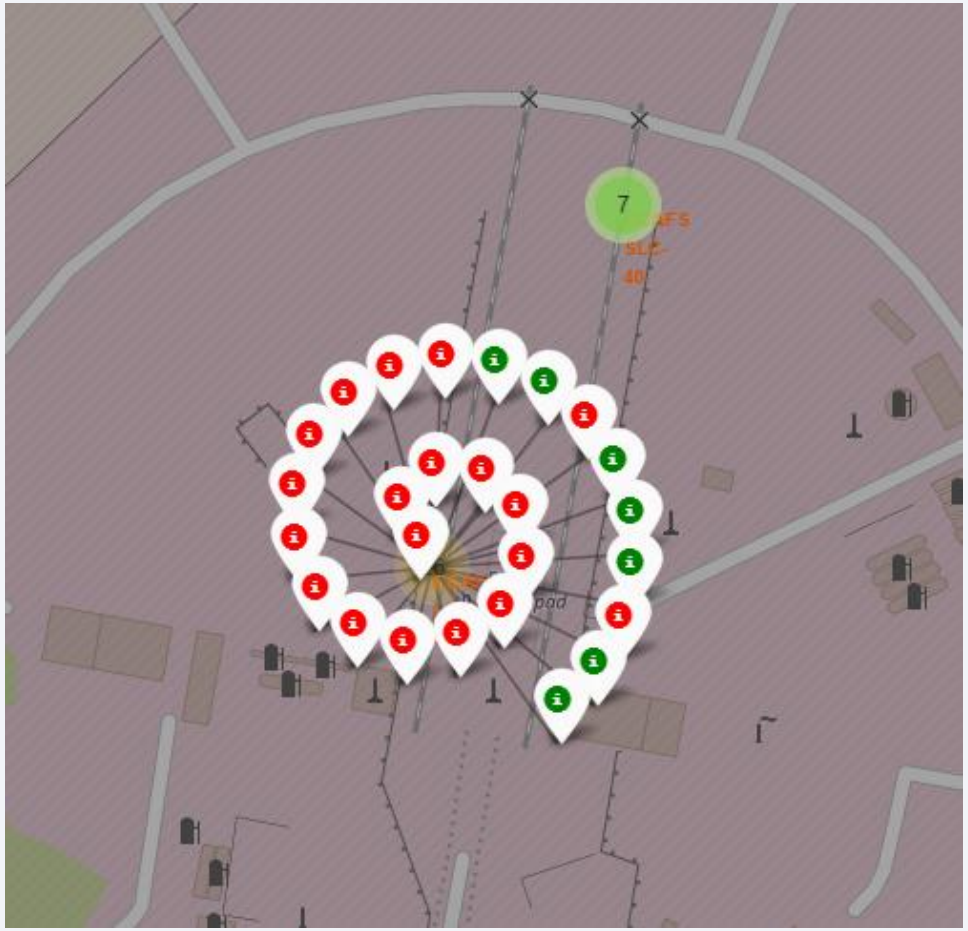
Explanation

The markers shows launch sites on the map

- The launch sites are near the geographic equator.
Some are in the state of Florida on the east coast and others in the state of California on the west coast

- All the sites in California are very close together, as are those in the state of Florida.

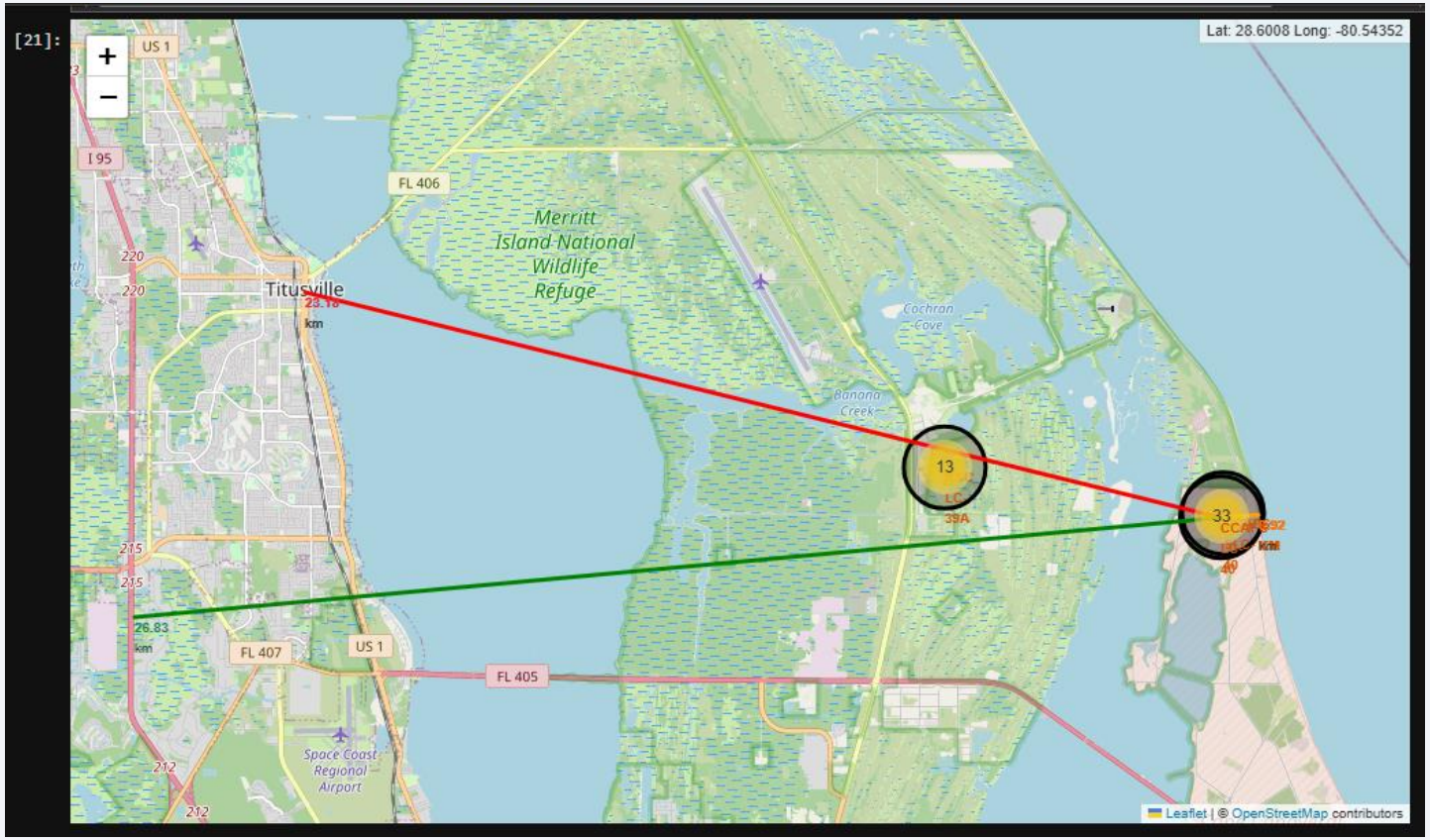
Successful and failure launches



Explanation

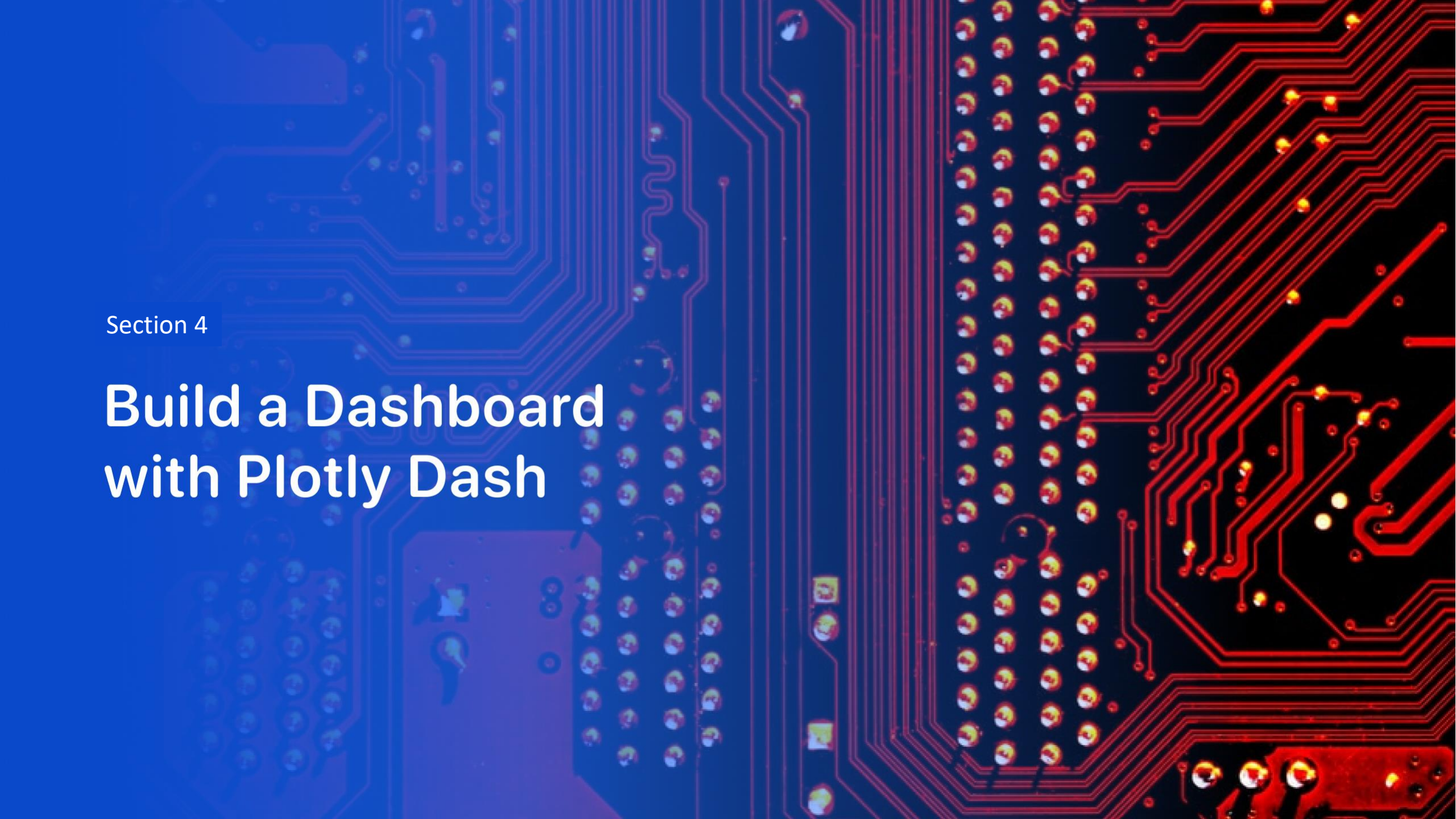
- Green marker shows successful launches for each sites
- Red markers shows failure launches for all sites

Distance from KSC LC-39A launch site to its proximities

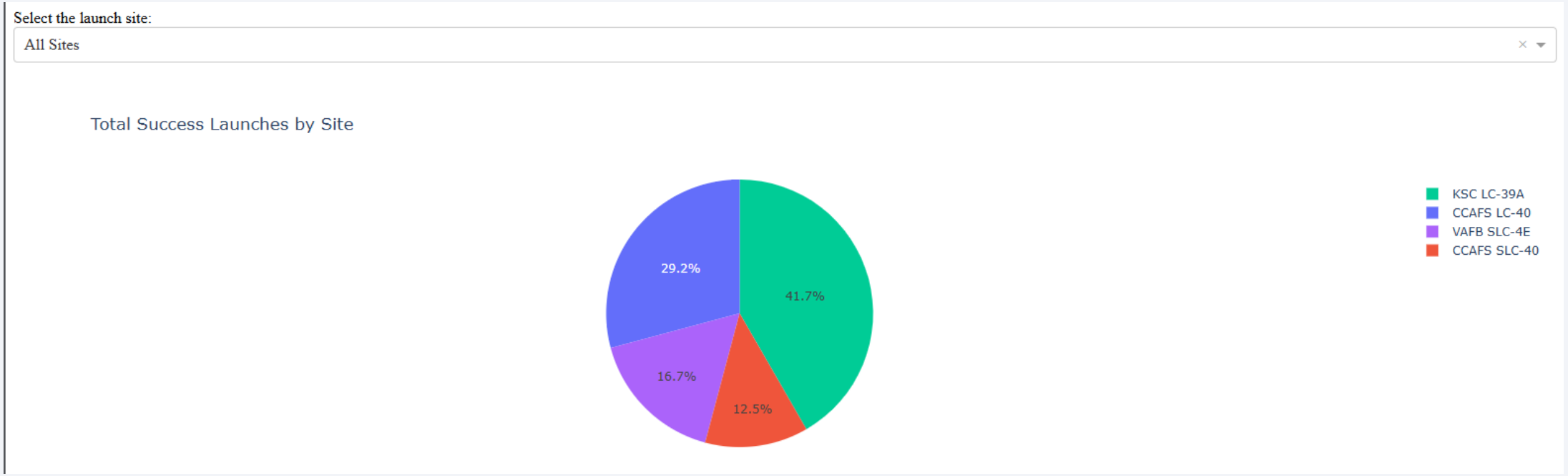


Explanation

- Proximity to Titusville (red line)
- Proximity to railway (green line)
- Proximity to coastline (yellow line)

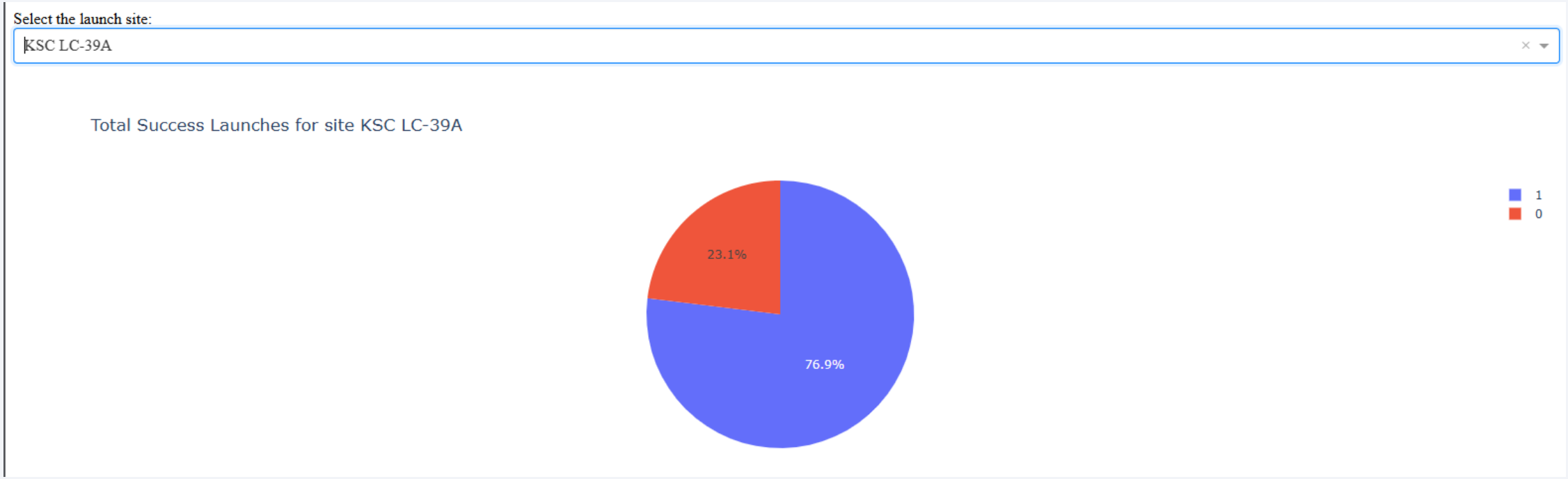

Section 4
Build a Dashboard with Plotly Dash

Distribution of launch success between sites



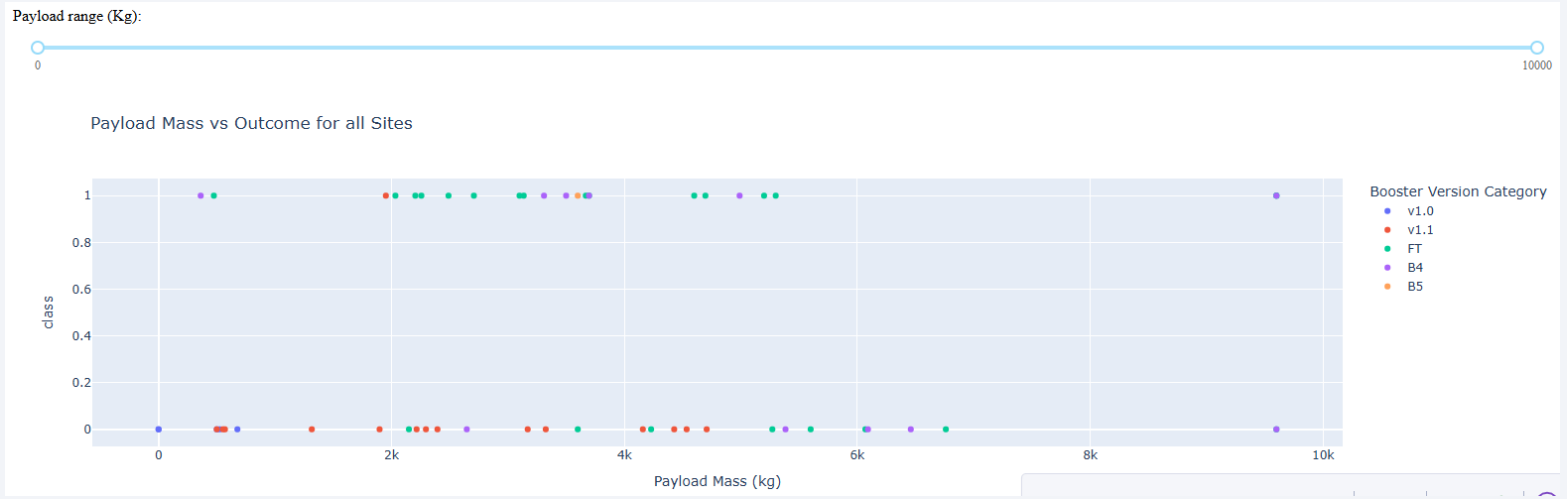
The pie chart shows that KSC LC-39A has the most successful launches. Although, CCAFS LC-40 has the worst rate of success

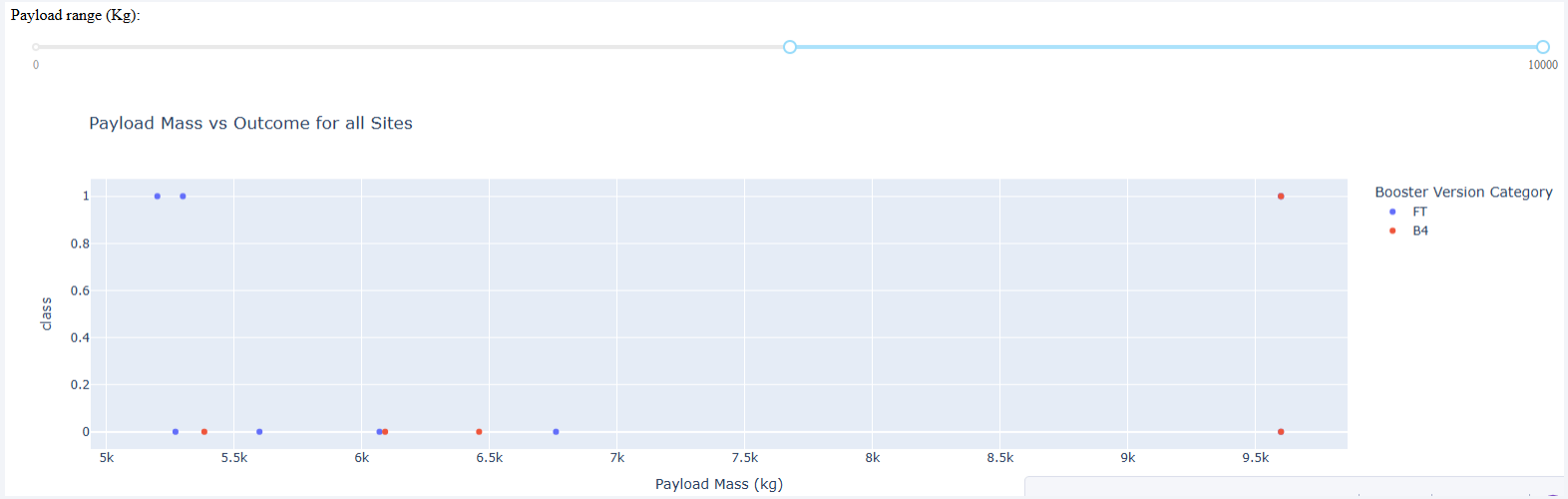
Comparison between successful and failure launches in KSC LC-39A



The pie chart shows that KSC LC-39A has 3 times more successful launches than failures

Pay load mass vs. launch outcome for all sites



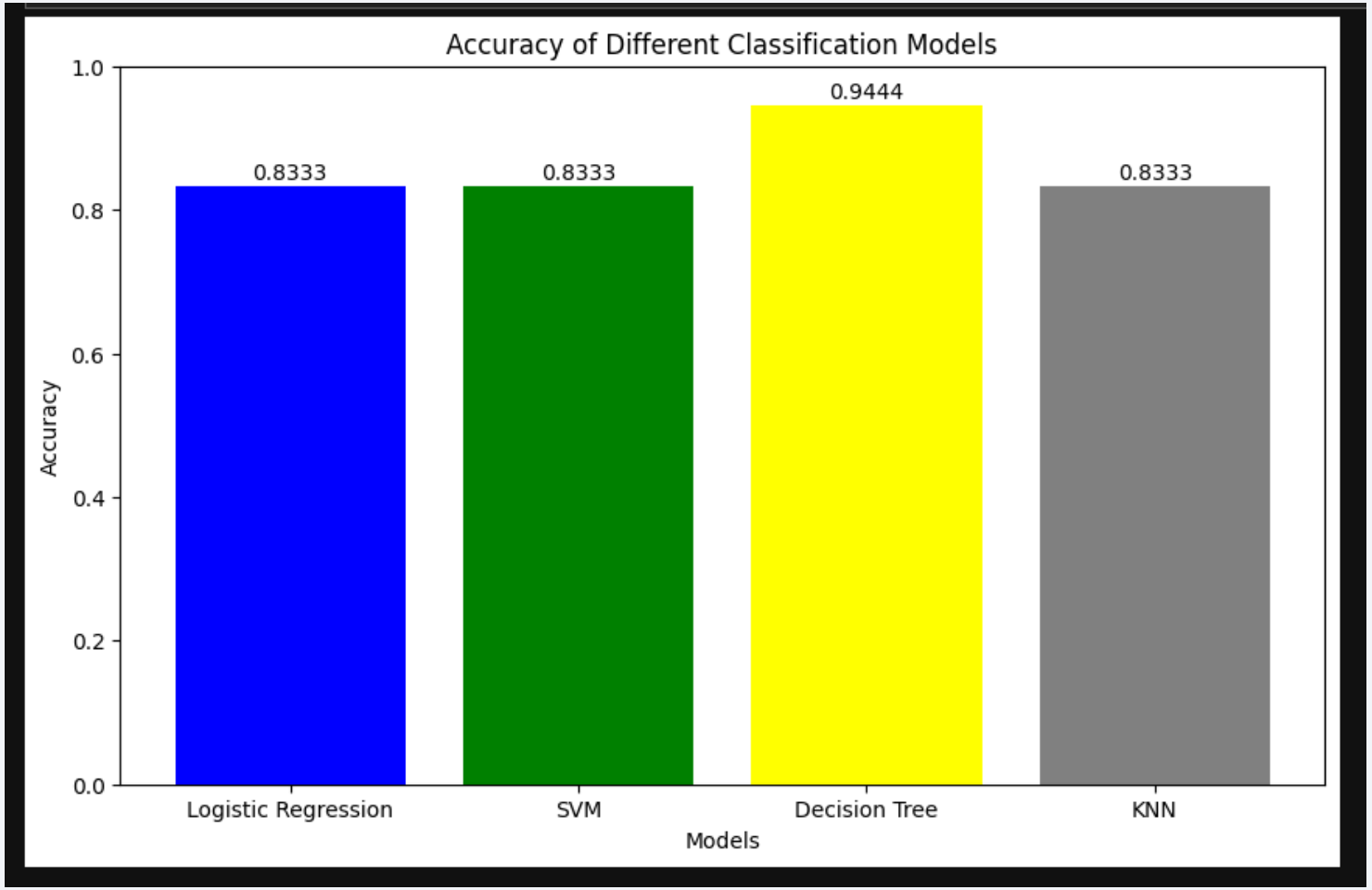


The scatter plots show that pay loads mass between 5,000 and 10,000 kg have the lowest success rate

Section 5

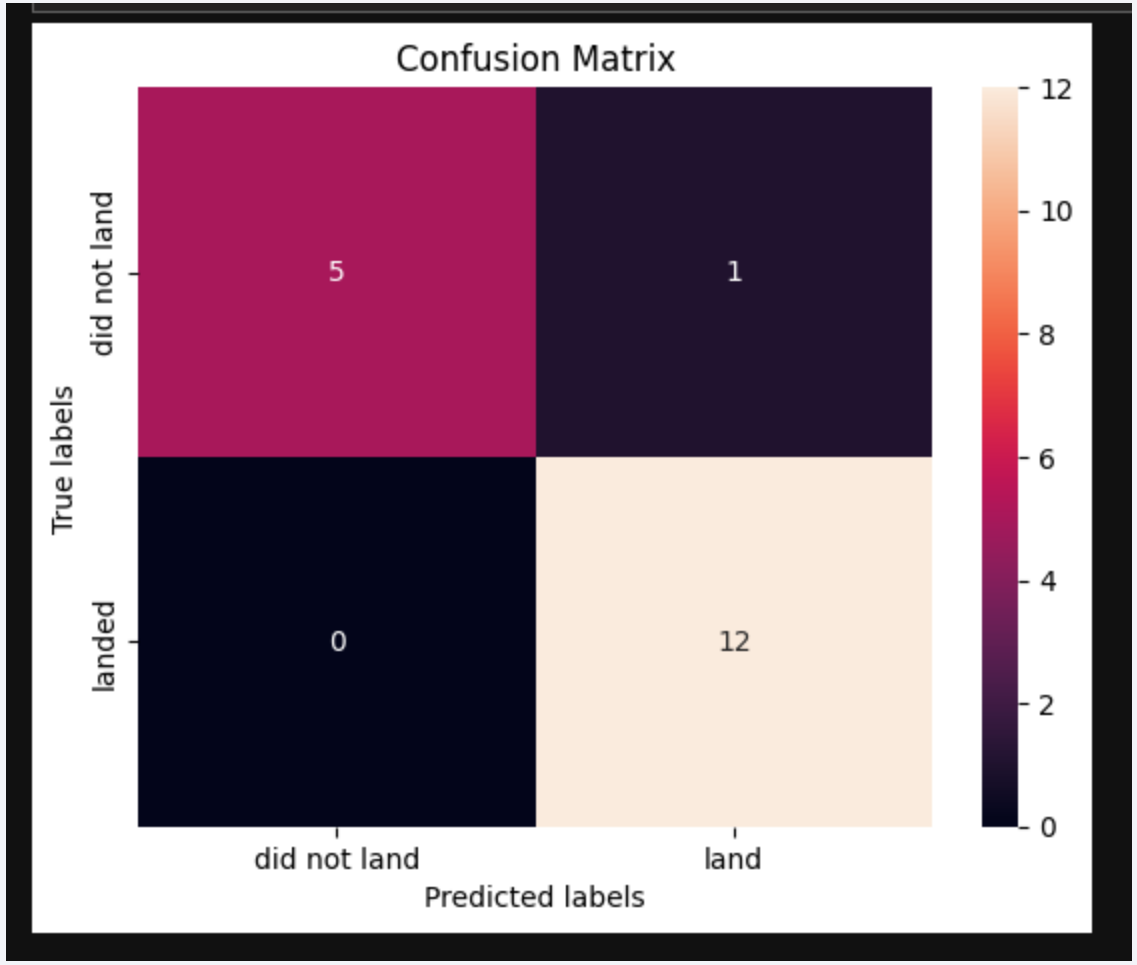
Predictive Analysis (Classification)

Classification Accuracy



The bar chart shows that, the Decision Tree model has the highest classification accuracy on the test data (0.9444).
This suggests that the Decision Tree model is better suited for this dataset compared to Logistic Regression, Support Vector Machine, and K Nearest Neighbors, all of which present an accuracy of 0.8333

Confusion Matrix



Explanation

- Decision tree model shows the bigest accuracy score of 94.44%, with a significant number of true positives and true negatives.

- The absence of false negatives indicates that the model reliably predicts successful landings.

- There is 1 false positive, this is less critical than false negatives in aerospace operations.

- The model shows a balanced performance with a slight bias towards predicting successful landings.

Conclusions

1. KSC LC-39A has the highest success rate (43.7%).

2. Launches with a low payload mass show better results
than launches with a larger payload mass.

3. The success rate of launches increases with the number of launches.

4. Orbits ES-L1, GEO, HEO and SSO have 100% success rate.

5. Interactive data visualizations using Folium, Plotly and Dash provide relevant information related to spatial features of SpaceX launches

6. Decision Tree Model is the best classification model for this dataset.

Acknowledgments

Special thanks to:

Instructors, Coursera and IBM

Thank you!