IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

ALAN VLADIMIR
PALAFOX GARDUÑO
24/11/2024

Outline

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

Executive Summary

- Summary of methodologies

For this project, I'm taking on the role of Data Scientist to analyze data from SpaceX, Elon Musk's company, with the goal of creating a competitor named Space Y. The study was carried out using data extraction through APIs and web scraping, data analytics, and machine learning modeling algorithms.

- Summary of all results

The analysis revealed that many launches failed for various reasons, including rocket weight, rocket system model, launch location, and the orbit in which the rocket would travel through space.

Introduction

- Project background and context

Predict whether the first stage of the Falcon 9 rocket will successfully land. SpaceX advertises Falcon 9 rocket launches on their website at a cost of $62 million, while other providers charge over $165 million per launch. A significant portion of this cost reduction is due to SpaceX's ability to reuse the first stage. Therefore, if we can predict whether the first stage will land, we can estimate the cost of a launch.

- Problems you want to find answers

By analyzing historical launch data from SpaceX, we can predict whether the first stage of the Falcon 9 launch will successfully land, allowing us to estimate the launch cost. This information could give the alternative company, SpaceY, a competitive edge against SpaceX in the rocket launch market.

Section 1

Methodology

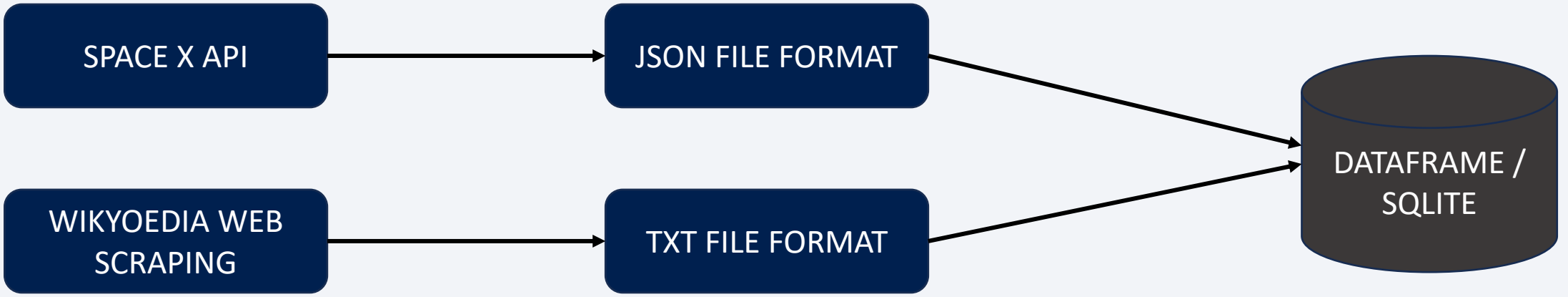
Methodology

Executive Summary

- Data collection methodology:
  - The data was collected using SpaceX's REST API and extracting information from Wikipedia through web scraping
- Perform data wrangling
  - The data was processed using the Pandas and NumPy libraries. One-Hot Encoding was used to remove unnecessary columns. Additionally, the data was normalized and standardized.
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Seaborn and Matplotlib were used for data visualization, and SQL was utilized for data extraction and manipulation.
- Perform interactive visual analytics using Folium and Plotly Dash
  - The Folium and Dash libraries were used.
- Perform predictive analysis using classification models
  - The main library used was Scikit-Learn. First, the data was split into predictor variables and the target variable., The data was normalized.
  - The dataset was divided into training and testing sets.
  - GridSearchCV was used to find the best hyperparameters for the models.
  - The classification methods were compared, and the one with the best performance and results was selected.

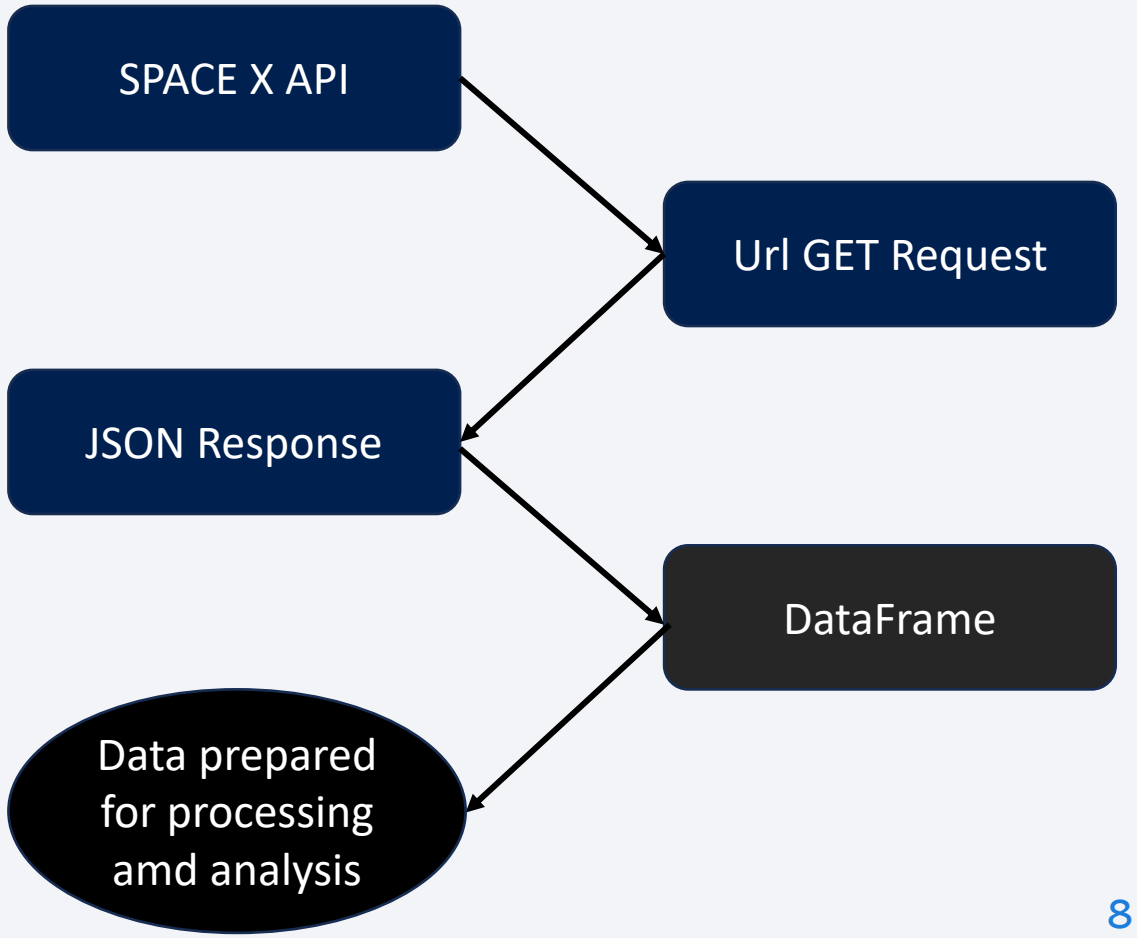
Data Collection

- The data was primarily obtained from two sources:
  - SpaceX API: An open-source REST API that contains historical information on past launches.
  - Web Scraping: Information publicly available on the Wikipedia website was extracted.



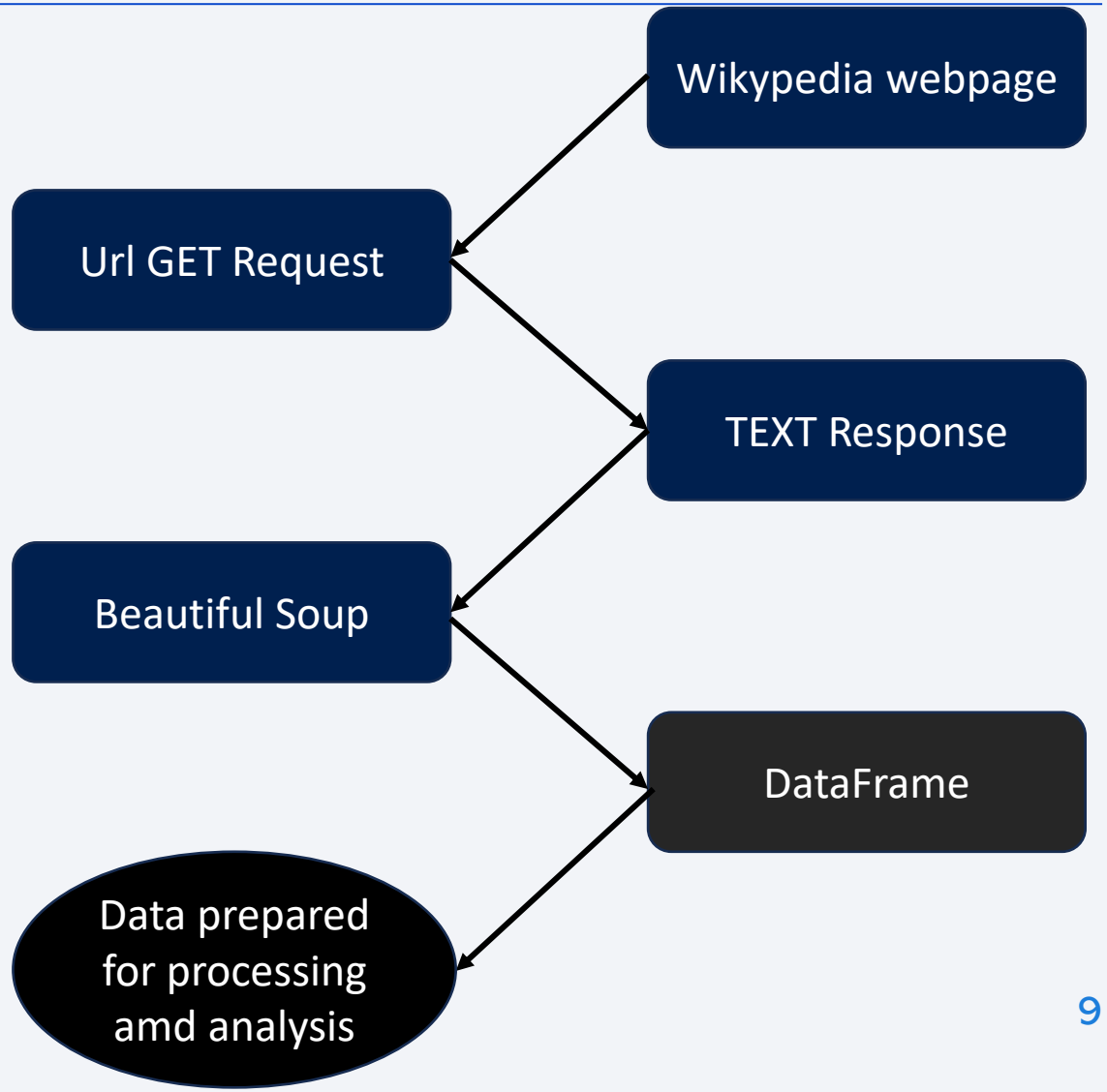
Data Collection – SpaceX API

- The data was extracted from the API using the Pandas, NumPy, and Requests libraries. A GET URL request was made with the Requests library, and the information was then extracted in JSON format. Finally, the data was converted into a DataFrame and stored in a .csv file or an SQL database.
- Add the GitHub URL of the completed SpaceX API calls notebook (Link)



Data Collection - Scraping

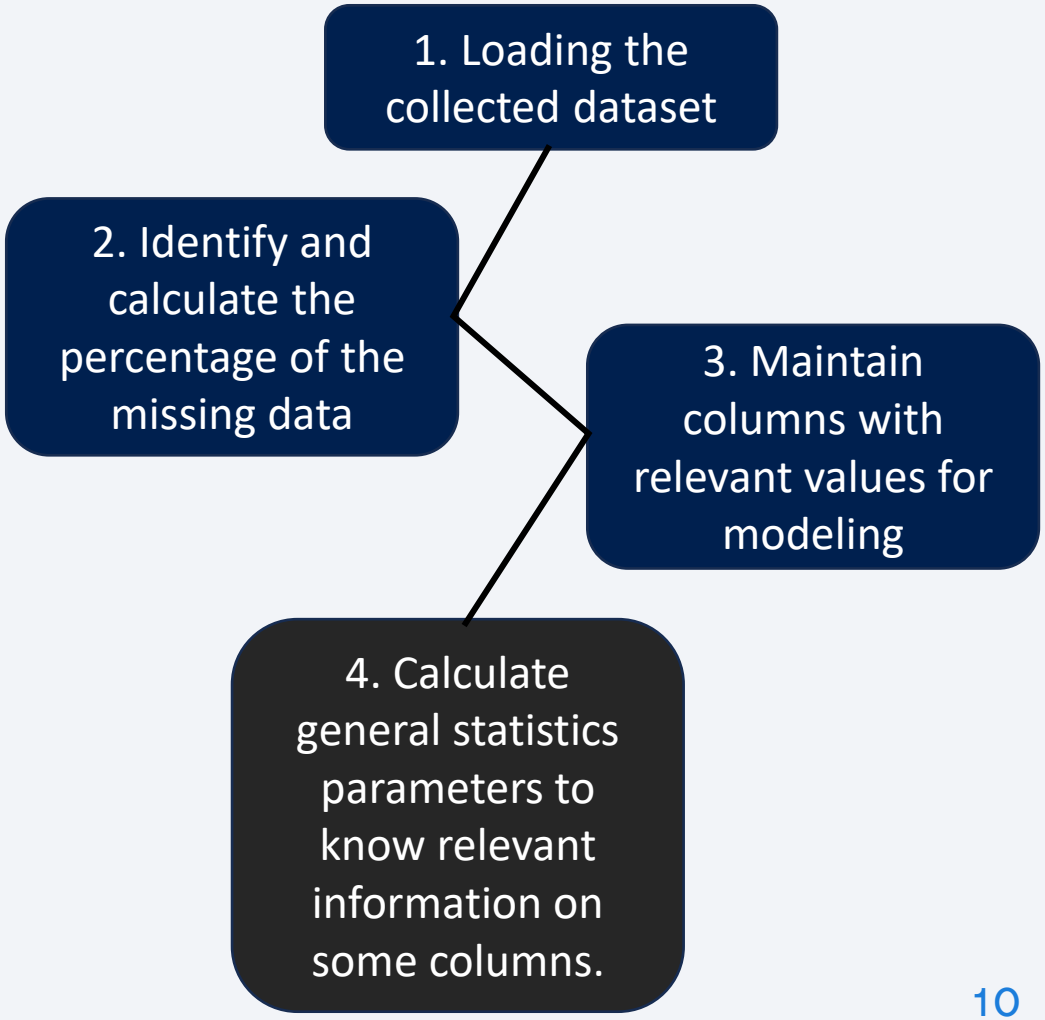
- The data was extracted through web scraping using the Pandas, NumPy, Requests, and Beautiful Soup libraries. A GET URL request was made using the Requests library, and the information was extracted in TEXT format. Then, Beautiful Soup was used to parse the data into a table based on HTML tags. Finally, the data was converted into a DataFrame and stored in a .csv file or an SQL database.
- Add the GitHub URL of the completed web scraping notebook (Link)



Data Wrangling

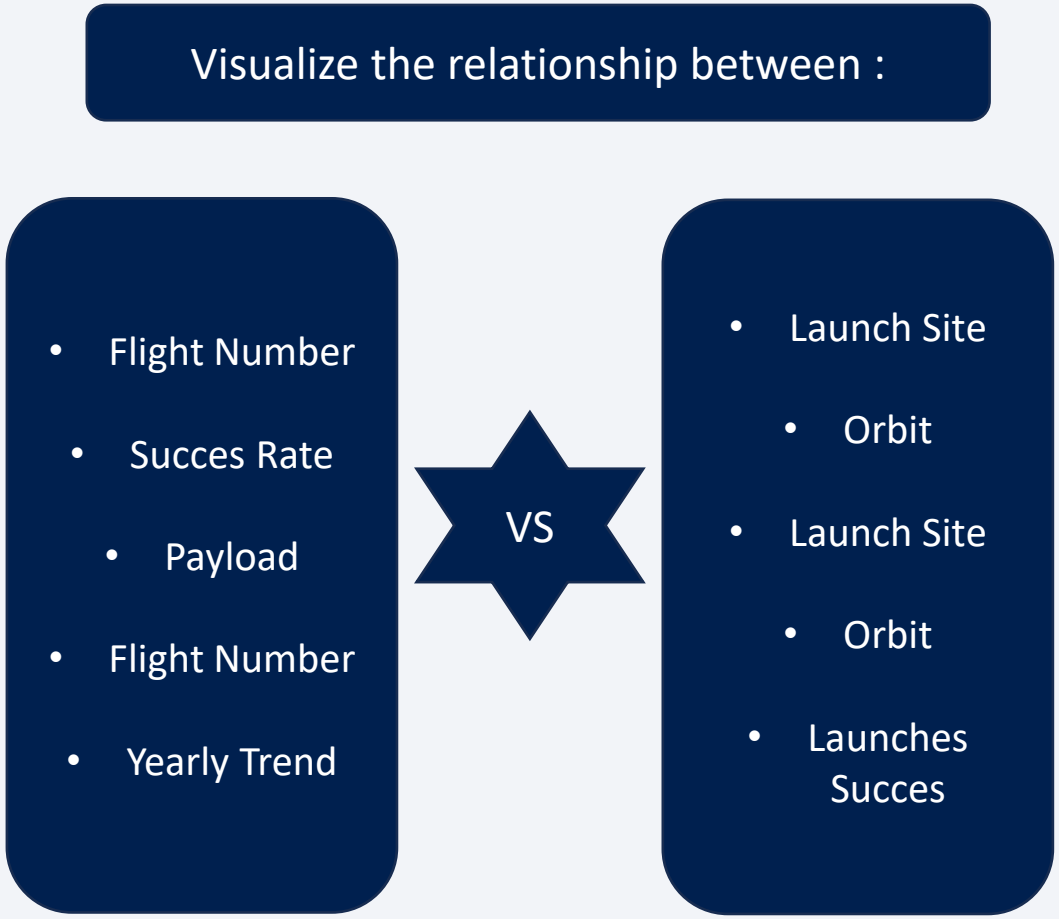
- The Pandas and NumPy libraries were loaded to handle the data obtained in the previous stage. The data was cleaned, information was validated, missing values were filled in, and columns with irrelevant values for machine learning modeling were removed.

- Add the GitHub URL of your completed data wrangling related notebooks (Link)



EDA with Data Visualization

- The EDA process was carried out using various visualizations with the Seaborn and Matplotlib libraries to identify correlations between attributes and the target variable. Additionally, categorical variables were transformed into dummy variables, and feature reduction was performed.
- Add the GitHub URL of your completed EDA with data visualization notebook (Link)



EDA with SQL

- The SQL queries used for the EDA stage are listed on the right:
- Add the GitHub URL of your completed EDA with SQL notebook (Link)

SQL Queries

1. Display the names of the unique launch sites in the space mission
2. Display 5 records where launch sites begin with the string 'CCA'
3. Display the total payload mass carried by boosters launched by NASA (CRS)
4. Display average payload mass carried by booster version F9 v1.1
5. List the date when the first succesful landing outcome in ground pad was acheived.
6. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
7. List the total number of successful and failure mission outcomes
8. List the names of the booster_versions which have carried the maximum payload mass. Using a subquery
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
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

Build an Interactive Map with Folium

- In this stage, the Folium library was used to visualize geospatial data. By adding color-coded markers to the launch sites, it became easier to pinpoint their locations and better understand the data.
- Add the GitHub URL of your completed interactive map with Folium map (Link)

Creation of an interactive map with Folium

- I started by creating a global map and then plotted the four launch sites with their labels.
- Colored markers were added to represent the failed and successful landings of rocket launches at these sites.
- Finally, the distance was calculated between the coastline, the road, and the nearest city to the CCAF LC-40 launch site.
- The launch sites are:

Launch Sites	Lat	Long
CCAFS LC-40	28.562302	-80.577356
CCAFS SLC-40	28.563197	-80.576820
KSC LC-39A	28.573255	-80.646895
VAFB SLC-4E	34.632834	-120.610745

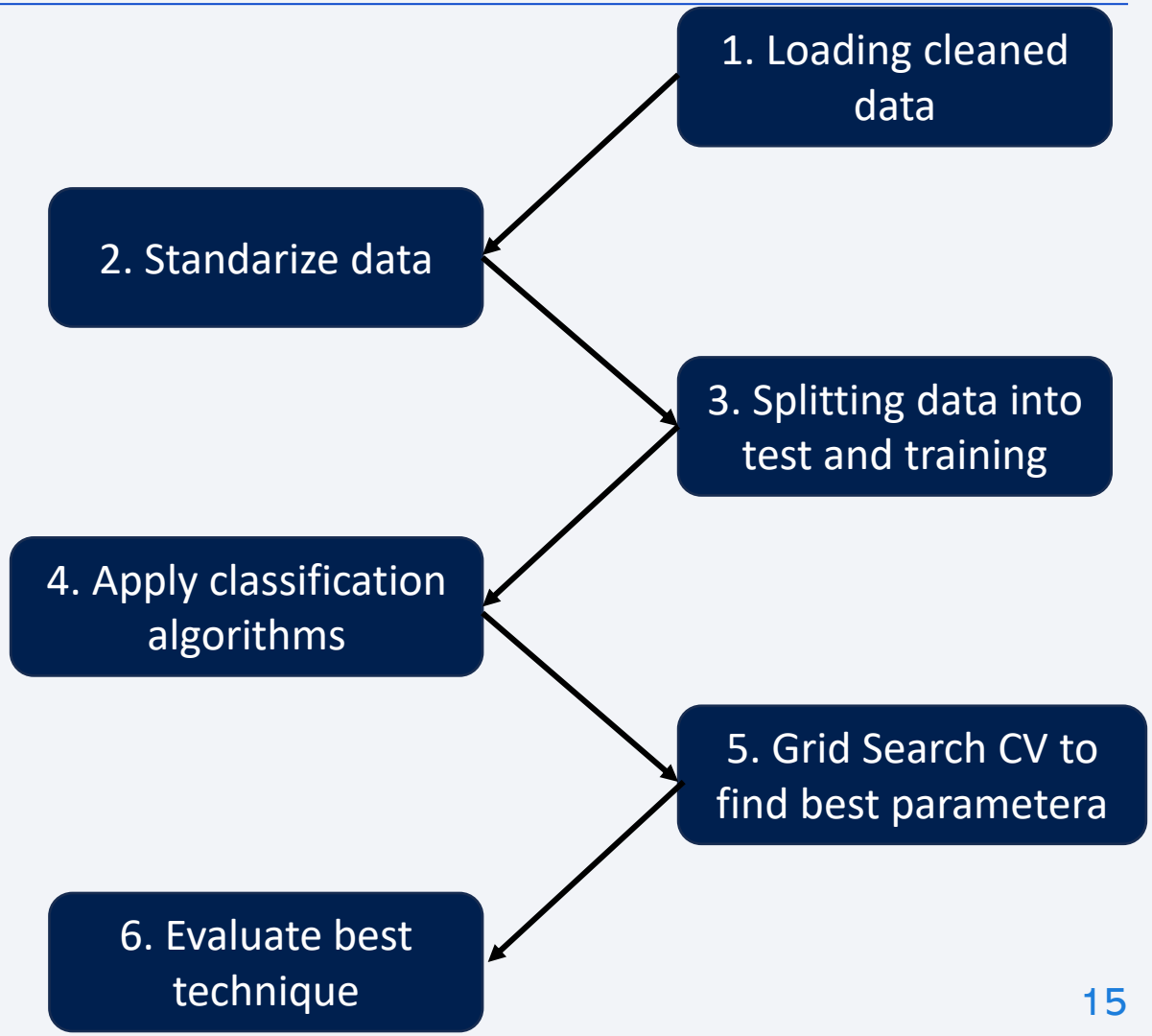
Build a Dashboard with Plotly Dash

- 1. A dropdown menu was added to enable the selection of a Launch Site for the Dashboard.
- 2. A pie chart was included to display the total successful launches for each site.
- 3. A slider was added to select payload within a specified range.
- 4. Finally, a scatter plot was created to show the correlation between payload and launch success.

- Add the GitHub URL of your completed Plotly Dash lab. (Link)

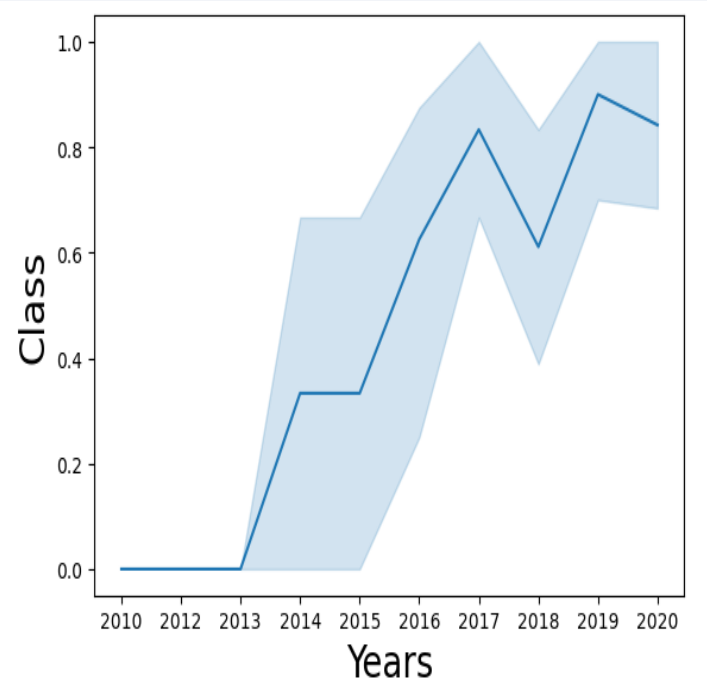
Predictive Analysis (Classification)

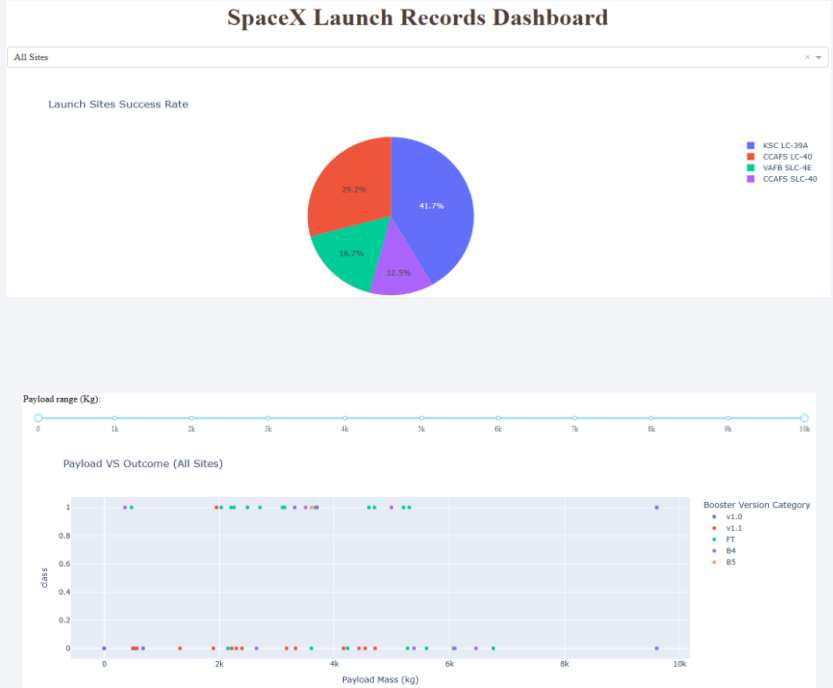
1. Loading cleaned data
2. Standardize data to prevent bias
3. Splitting the data into 20% and 80% for testing and training, respectively
4. Apply 4 different classification algorithms:
    1. Logistic Regression (LR)
    2. Suport Vector Machine (SVM)
    3. Decision Tree (DT)
    4. K Nereast Neighbors (KNN)
5. Using Grid Search to find best hyper parameters
6. Evaluating techniques applying Confusion matrix, F1 score, Jaccard Score and use the best algorithm (Link)

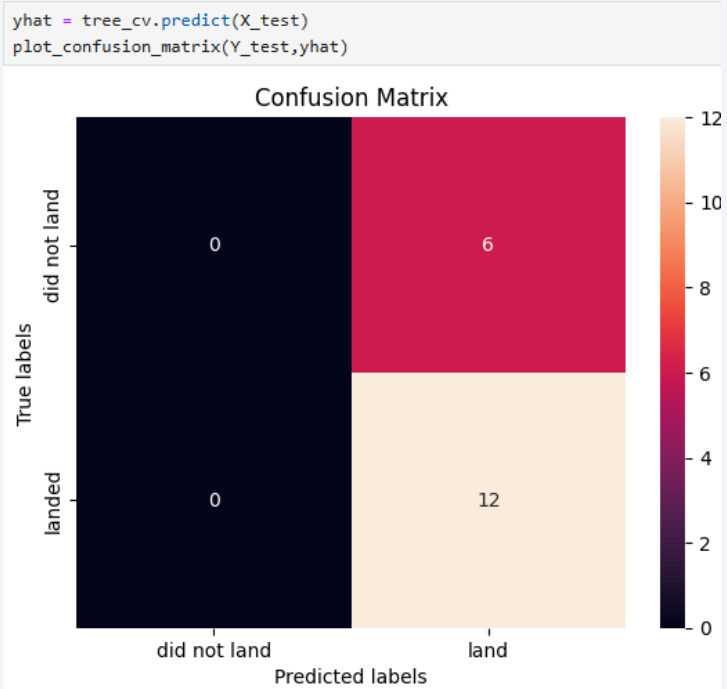


Results

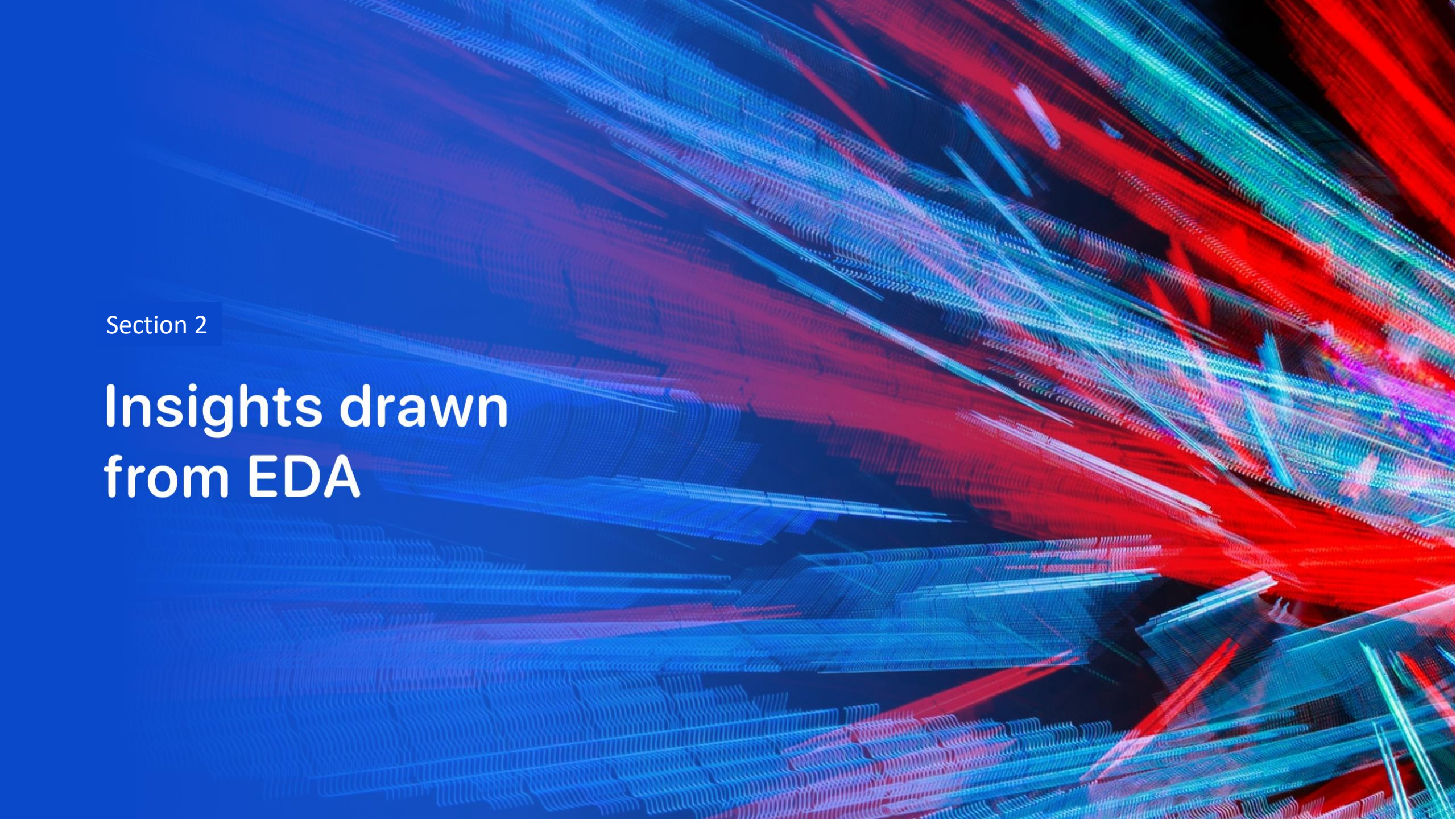
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results





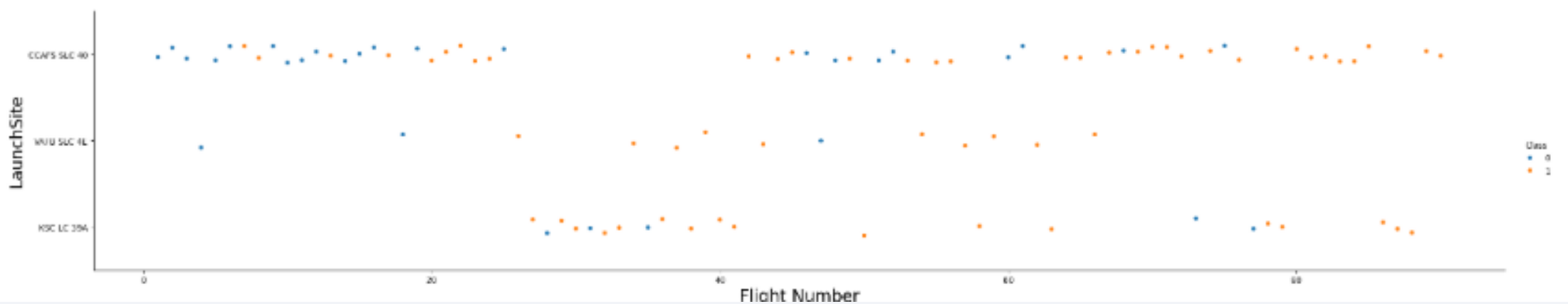


The exploratory data analysis revealed that successful landings have been steadily increasing since 2013. The best classification algorithm for making accurate predictions is the Decision Tree.


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

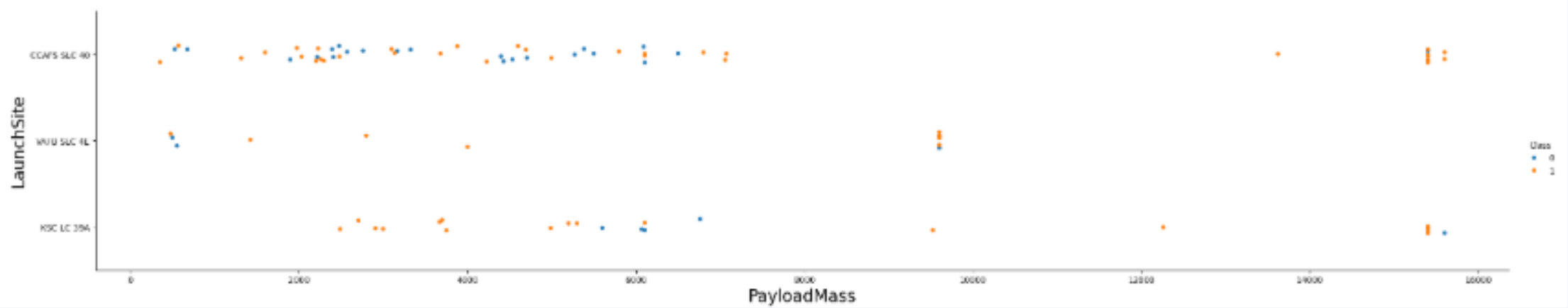
- Plot of Flight Number vs. Launch Site



- CCAFS SLS 40: This is the site with the highest number of launches by SpaceX, making it the site with the most successful landings.
- VAFB SLC 4E: This is the site with the fewest launches overall, with only three successful landings and ten failures.
- KSC LC 39A: This is the most recently used launch site, and it has fewer successes compared to failures at the same site.

Payload vs. Launch Site

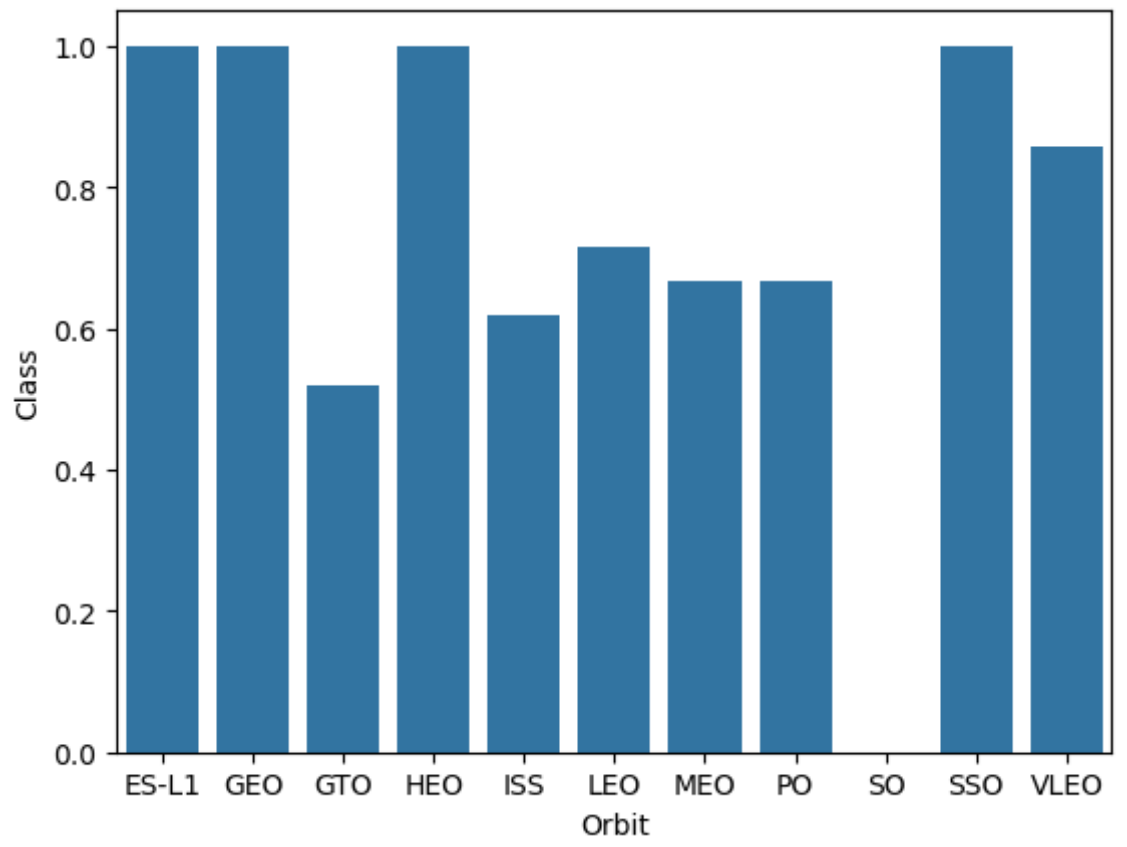
- Show a scatter plot of Payload vs. Launch Site



- This chart clearly shows the relationship between rocket mass and landing success. The lighter the rocket, the higher the probability of success; the heavier the rocket, the greater the likelihood of failure

Success Rate vs. Orbit Type

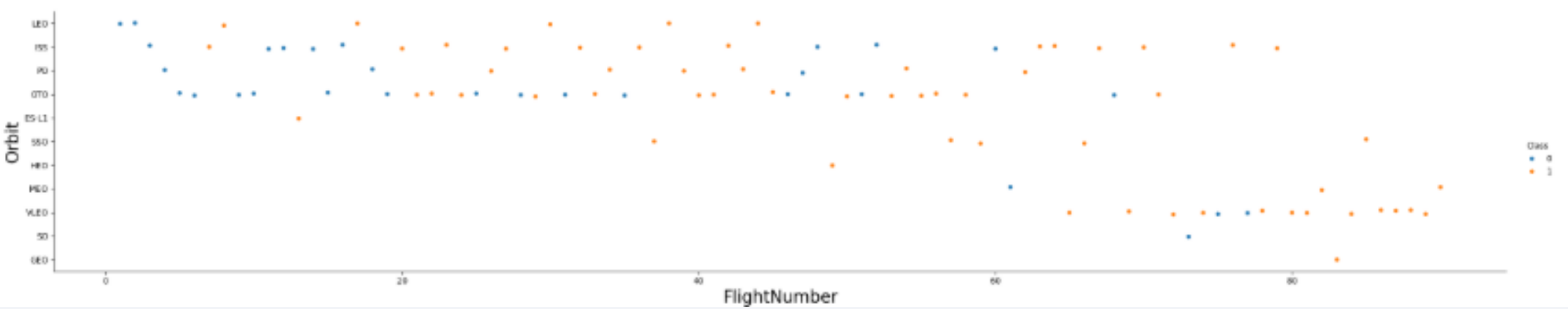
- Barplot – Orbit vs Class



- The orbit followed by rockets during launch significantly impacts their landing success. Satellites in ES-L1, GEO, HEO, and SSO orbits have contributed to higher launch success rates. On the other hand, GTO is the orbit with the highest failure rate.

Flight Number vs. Orbit Type

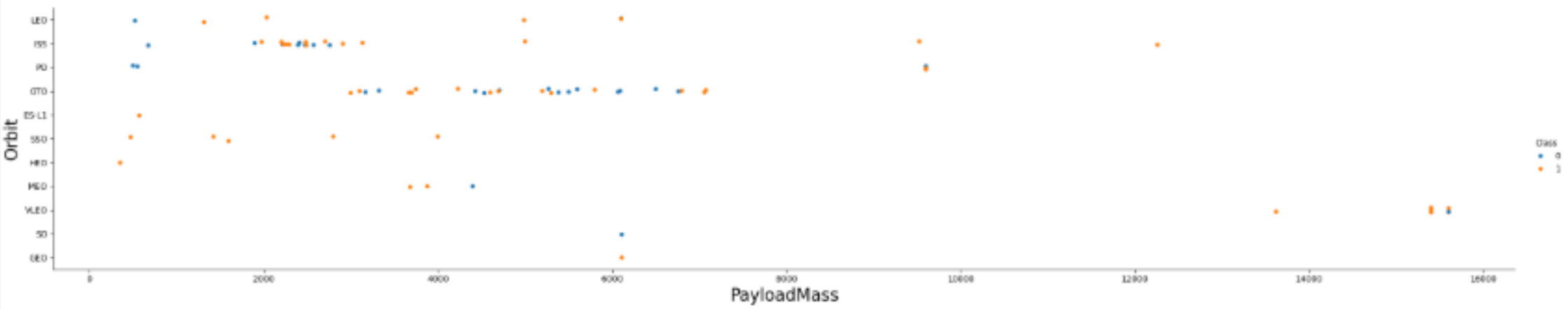
- Scatterplot - Flight number vs. Orbit type



- This chart shows the number of flights and the orbits followed by the rockets. It's clear that the orbits with the highest number of launches are the most successful, while the newer orbits tested in recent flights have only achieved four successful landings.

Payload vs. Orbit Type

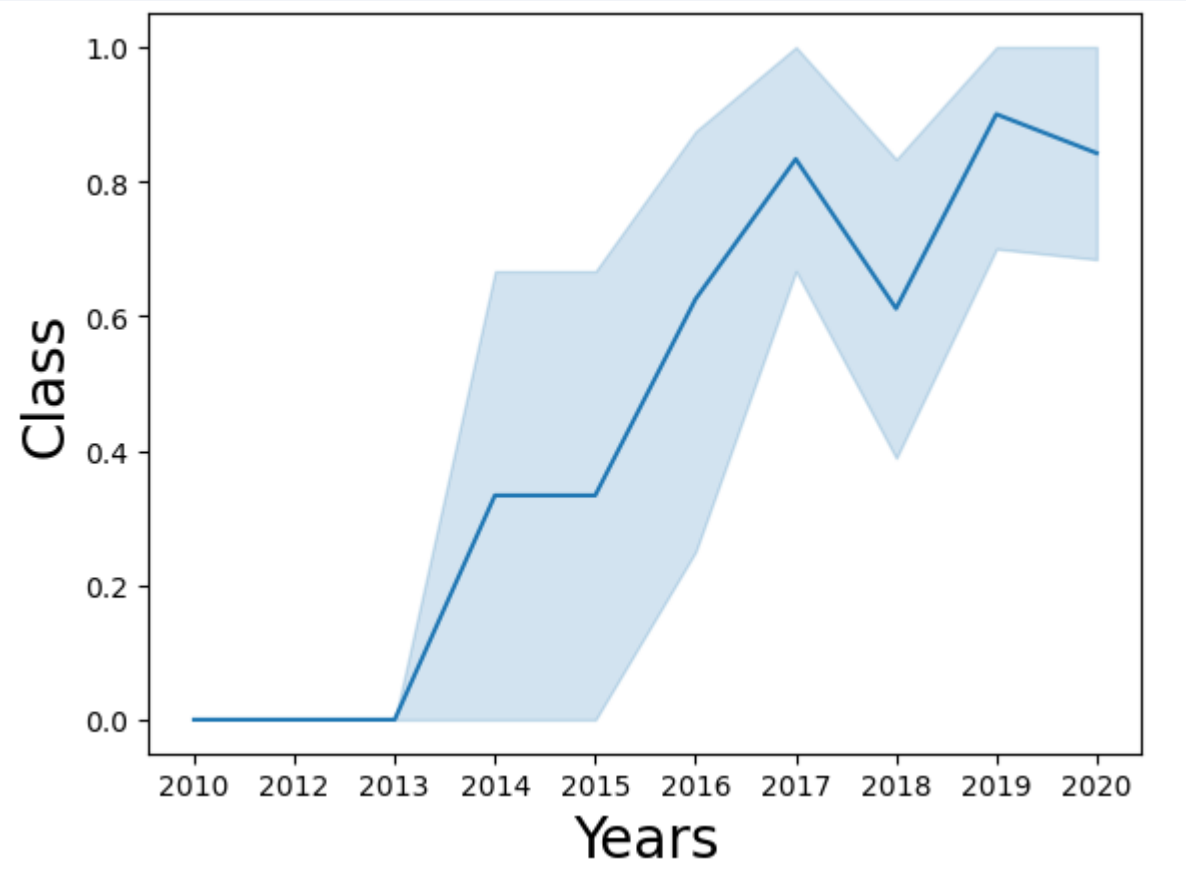
- Scatterplot – Payload vs. orbit type



- This chart displays the mass of the rockets launched and the orbits they followed. It highlights that weight has a greater impact on launch success than the orbits, reaffirming the trend: the lighter the mass, the higher the success rate, while heavier rockets have a greater likelihood of failure

Launch Success Yearly Trend

- Lineplot - Yearly average success rate



- The success rate of rocket launches has been steadily increasing over time, with the upward trend starting in 2013. Since then, the probability of success has continued to grow.

All Launch Site Names

- Find the names of the unique launch sites

```
%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'

- Find 5 records where launch sites begin with `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_Ou
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (par
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 (par
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No a
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No a
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No a

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE CUSTOMER = 'NASA (CRS)'
```

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE BOOSTER_VERSION LIKE 'F9 v1.1'
```

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

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

```
%sql SELECT MIN(DATE), LANDING_OUTCOME, MISSION_OUTCOME FROM SPACEXTABLE \
    WHERE MISSION_OUTCOME LIKE 'SUCCESS'
```

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

MIN(DATE)	Landing_Outcome	Mission_Outcome
2010-06-04	Failure (parachute)	Success

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

```
%sql SELECT Booster_Version, PAYLOAD, LANDING_OUTCOME, MISSION_OUTCOME \
    FROM SPACEXTABLE \
    WHERE LANDING_OUTCOME LIKE '%DRONE SHIP%' AND \
        MISSION_OUTCOME LIKE 'SUCCESS' AND \
        PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
```

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

Booster_Version	Payload	Landing_Outcome	Mission_Outcome
F9 FT B1020	SES-9	Failure (drone ship)	Success
F9 FT B1022	JCSAT-14	Success (drone ship)	Success
F9 FT B1026	JCSAT-16	Success (drone ship)	Success
F9 FT B1021.2	SES-10	Success (drone ship)	Success
F9 FT B1031.2	SES-11 / EchoStar 105	Success (drone ship)	Success

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

```
%sql SELECT COUNT(S.Mission_Outcome), COUNT(F.Mission_Outcome) \
    FROM (SELECT * FROM SPACEXTABLE WHERE MISSION_OUTCOME LIKE '%SUCCESS%') S,\
         (SELECT * FROM SPACEXTABLE WHERE MISSION_OUTCOME LIKE '%FAILURE%') F
```

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

COUNT(S.Mission_Outcome)	COUNT(F.Mission_Outcome)
100	100

Boosters Carried Maximum Payload

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

```
%sql SELECT BOOSTER_VERSION, PAYLOAD_MASS__KG_, PAYLOAD, LANDING_OUTCOME \
    FROM SPACEXTABLE \
    WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
```

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

Booster_Version	PAYLOAD_MASS__KG_	Payload	Landing_Outcome
F9 B5 B1048.4	15600	Starlink 1 v1.0, SpaceX CRS-19	Success
F9 B5 B1049.4	15600	Starlink 2 v1.0, Crew Dragon in-flight abort test	Success
F9 B5 B1051.3	15600	Starlink 3 v1.0, Starlink 4 v1.0	Success
F9 B5 B1056.4	15600	Starlink 4 v1.0, SpaceX CRS-20	Failure
F9 B5 B1048.5	15600	Starlink 5 v1.0, Starlink 6 v1.0	Failure
F9 B5 B1051.4	15600	Starlink 6 v1.0, Crew Dragon Demo-2	Success
F9 B5 B1049.5	15600	Starlink 7 v1.0, Starlink 8 v1.0	Success
F9 B5 B1060.2	15600	Starlink 11 v1.0, Starlink 12 v1.0	Success
F9 B5 B1058.3	15600	Starlink 12 v1.0, Starlink 13 v1.0	Success
F9 B5 B1051.6	15600	Starlink 13 v1.0, Starlink 14 v1.0	Success
F9 B5 B1060.3	15600	Starlink 14 v1.0, GPS III-04	Success
F9 B5 B1049.7	15600	Starlink 15 v1.0, SpaceX CRS-21	Success

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%sql SELECT STRFTIME('%Y', DATE) AS YEAR, STRFTIME('%m', DATE) AS MONTH, MISSION_OUTCOME, LAUNCH_SITE, LANDING_OUTCOME
    FROM SPACEXTABLE \
    WHERE LANDING_OUTCOME LIKE '%DRONE SHIP%' AND STRFTIME('%Y', DATE) LIKE '2015'\
    AND LANDING_OUTCOME LIKE '%FAILURE%';
```

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

YEAR	MONTH	Mission_Outcome	Launch_Site	Landing_Outcome	Booster_Version
2015	01	Success	CCAFS LC-40	Failure (drone ship)	F9 v1.1 B1012
2015	04	Success	CCAFS LC-40	Failure (drone ship)	F9 v1.1 B1015

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

```
%sql SELECT COUNT(F.Landing_Outcome) AS FAILURE, COUNT(S.Landing_Outcome) AS SUCCESS \
    FROM (SELECT * FROM SPACEXTABLE WHERE Landing_Outcome LIKE '%drone ship%' AND DATE BETWEEN '2010-06-04' AND '2017-
         (SELECT * FROM SPACEXTABLE WHERE Landing_Outcome LIKE '%ground pad%' AND DATE BETWEEN '2010-06-04' AND '2017-
```

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

FAILURE	SUCCESS
33	33

```
%sql ALTER TABLE SPACEXTABLE ADD COLUMN CLASSIF
```

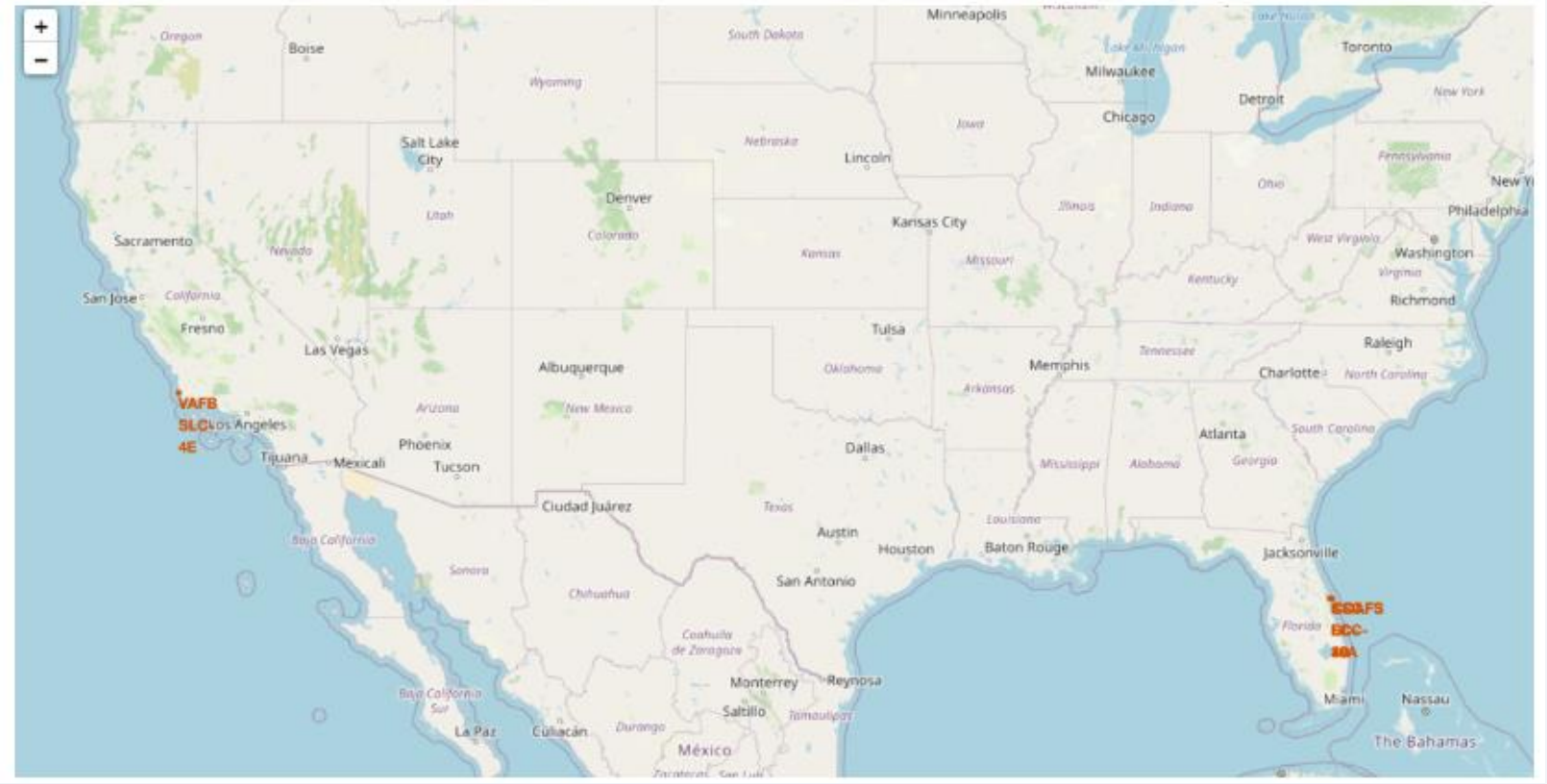
```
 * sqlite:///my_data1.db
(sqlite3.OperationalError) duplicate column name: CLASSIF
[SQL: ALTER TABLE SPACEXTABLE ADD COLUMN CLASSIF]
(Background on this error at: https://sqlalche.me/e/20/e3q8)
```

```
%sql UPDATE SPACEXTABLE \
    SET CLASSIF = CASE \
        WHEN Landing_Outcome LIKE '%drone ship%' THEN 'FAILURE' \
        WHEN Landing_Outcome LIKE '%ground pad%' THEN 'SUCCESS' \
        ELSE '0' \
    END
```

Section 3

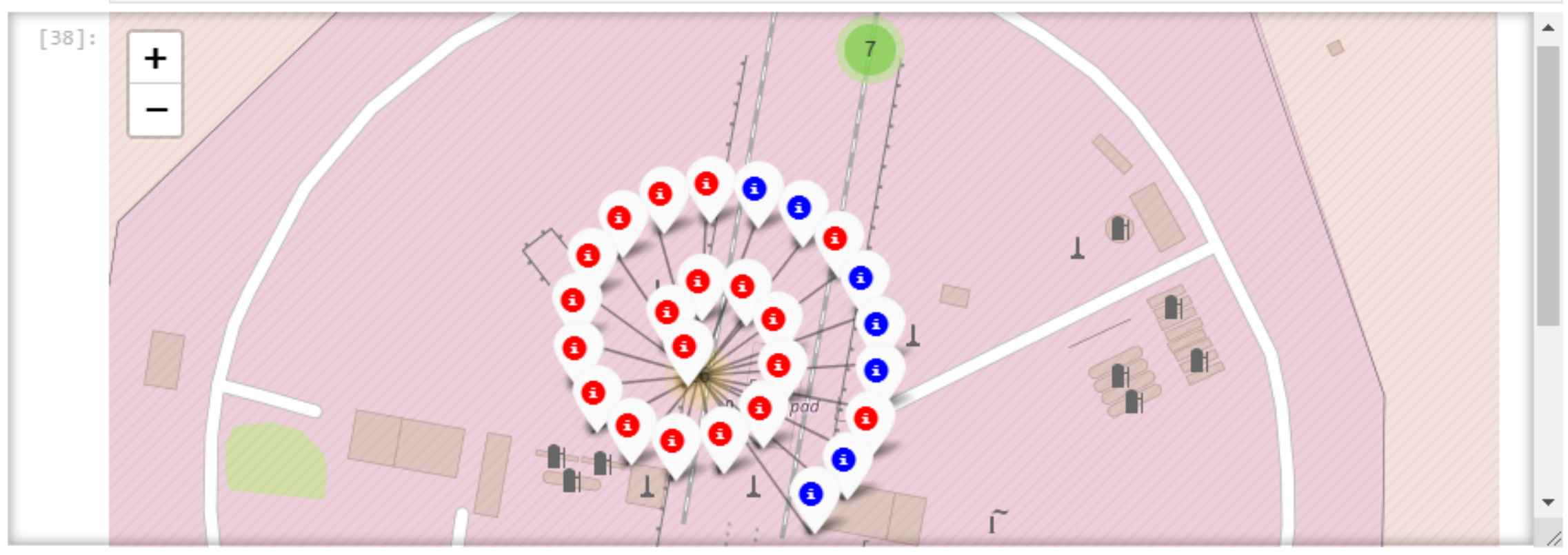
Launch Sites Proximities Analysis

Folium Map: Launch Sites.



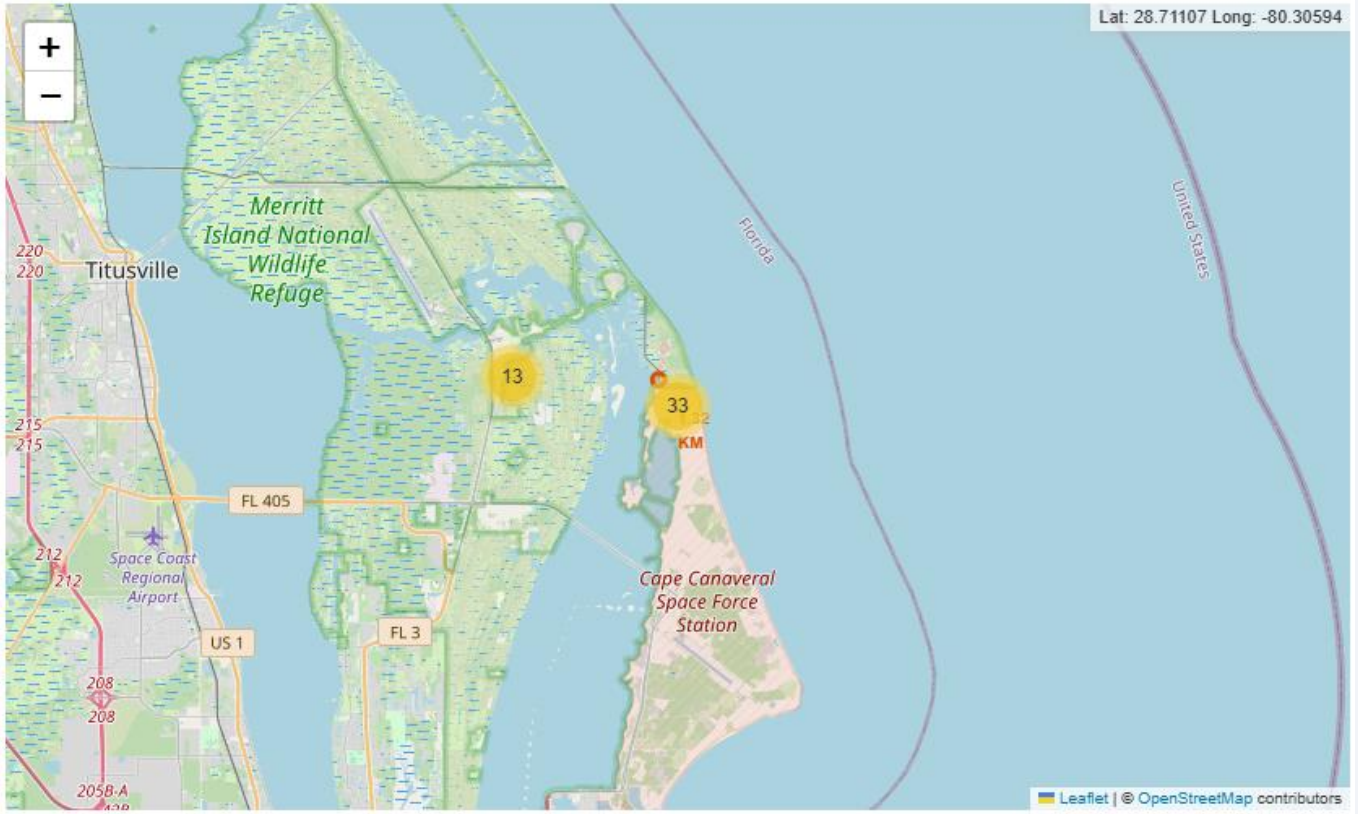
- All SpaceX launches have taken place in the southeastern and southwestern regions of the U.S., specifically in California and Florida.

Folium Map: Success rate for each launch location

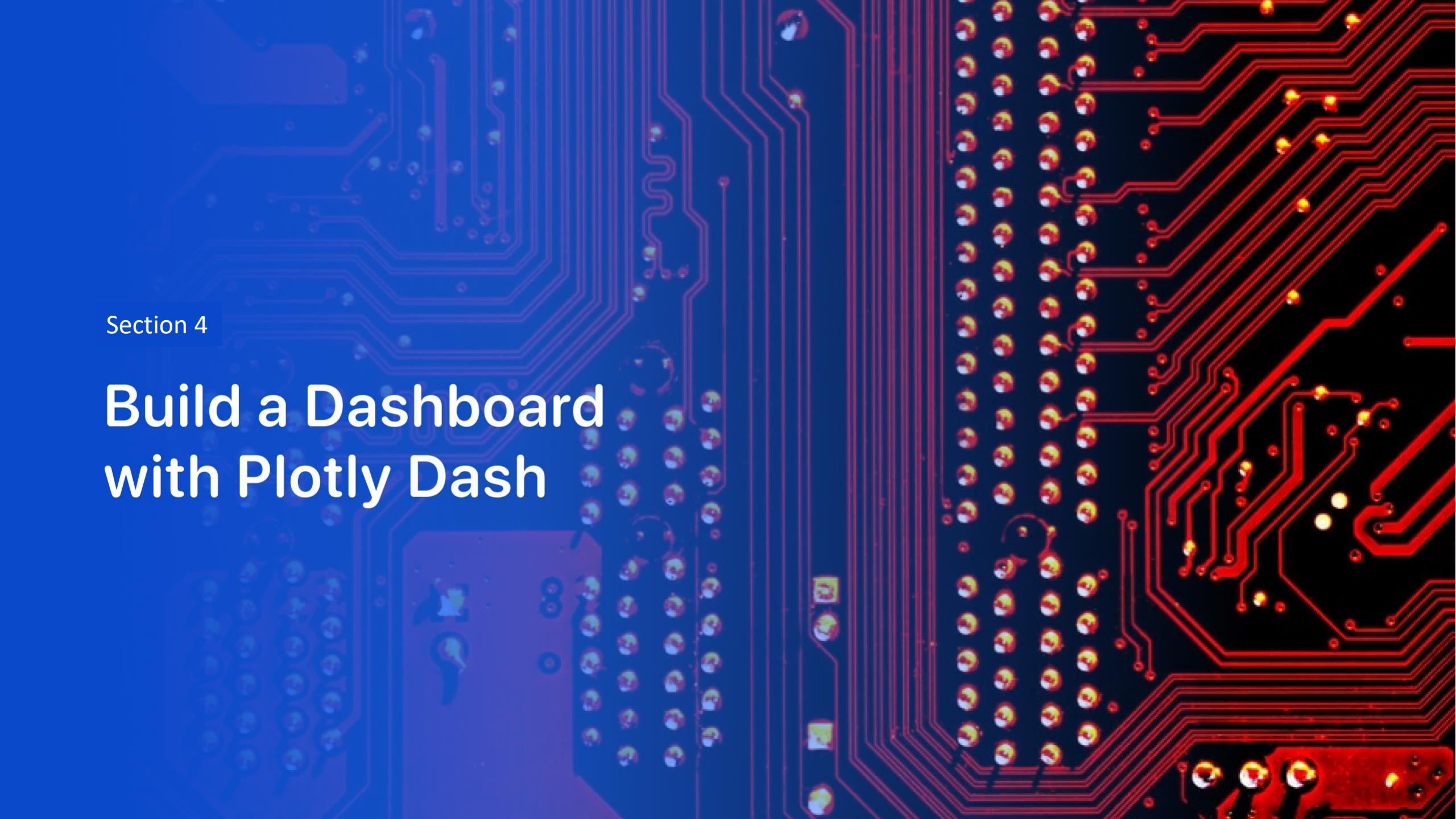


- The flights were classified into two groups: Blue for successful landings and Red for failed landings.

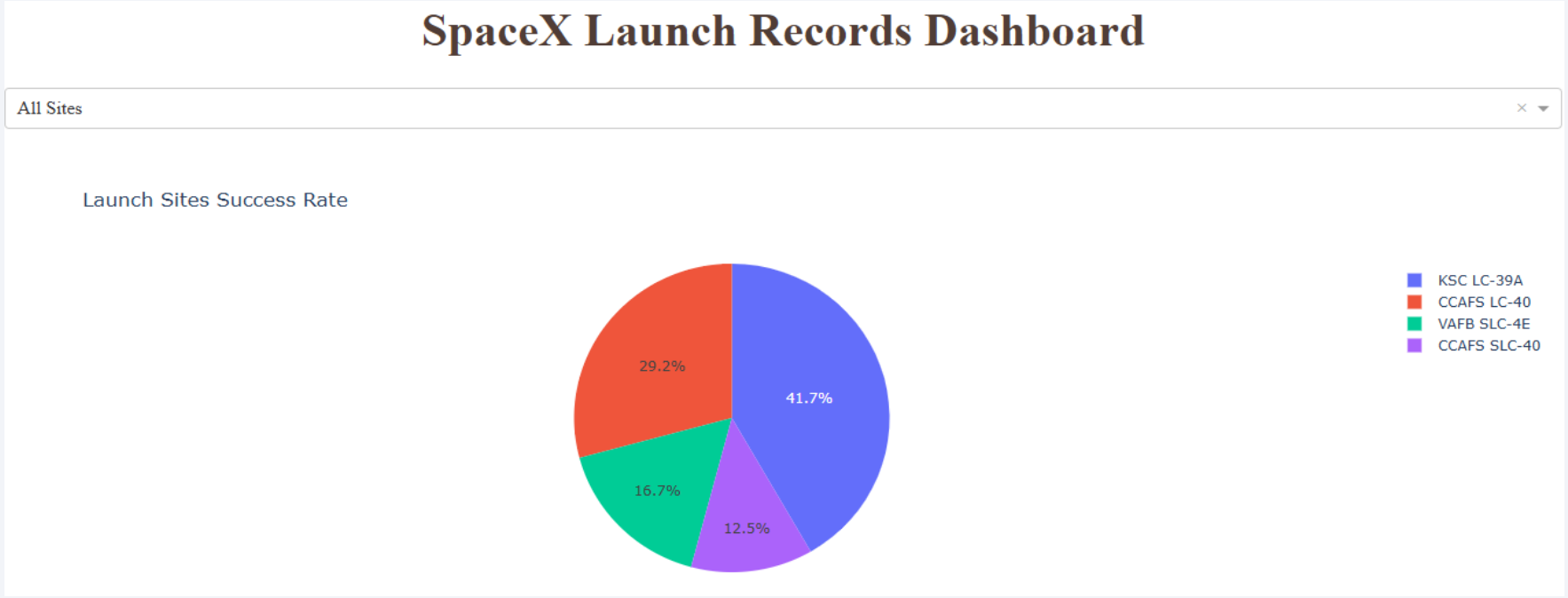
Folium Map: Closest Proximities to CCAFS LC-40



- The distance between a launch site and the nearest road and coastline was calculated.

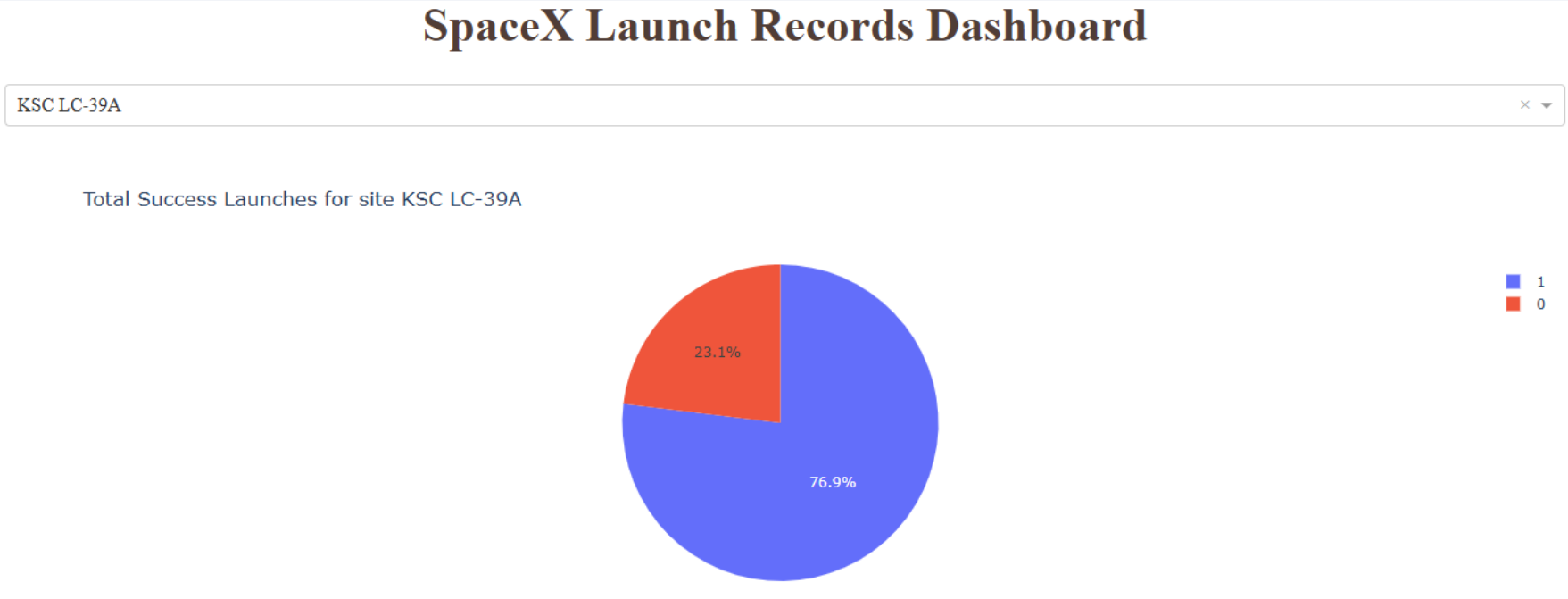

Section 4
Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>



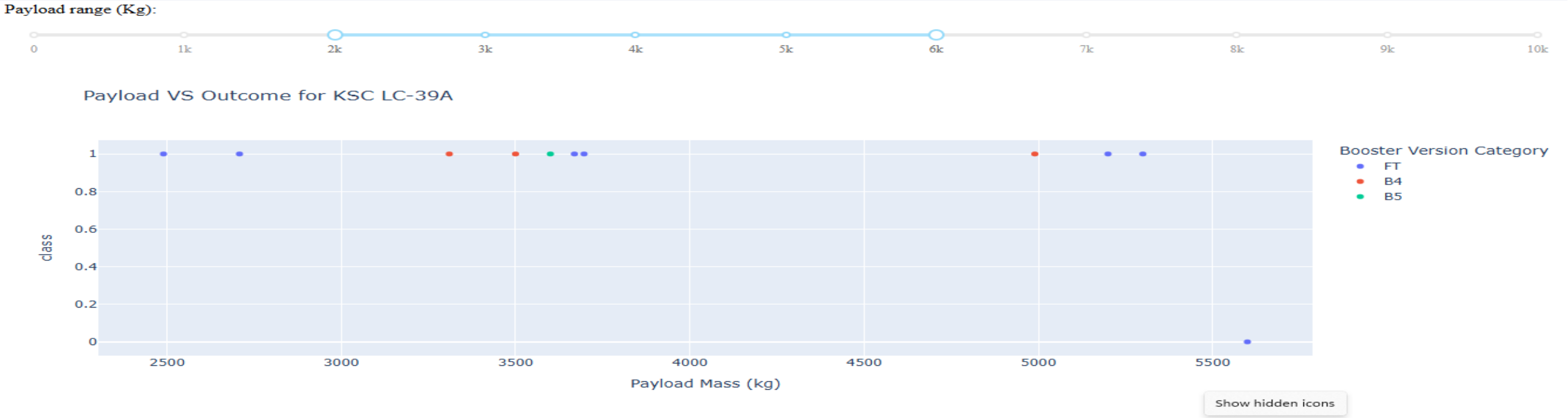
- A pie chart showing the percentage of launches conducted at each site.

<Dashboard Screenshot 2>



- A pie chart showing the percentage of successful and failed landings at the site with the highest number of launches by SpaceX.

Dashboard: Launch success count for all sites

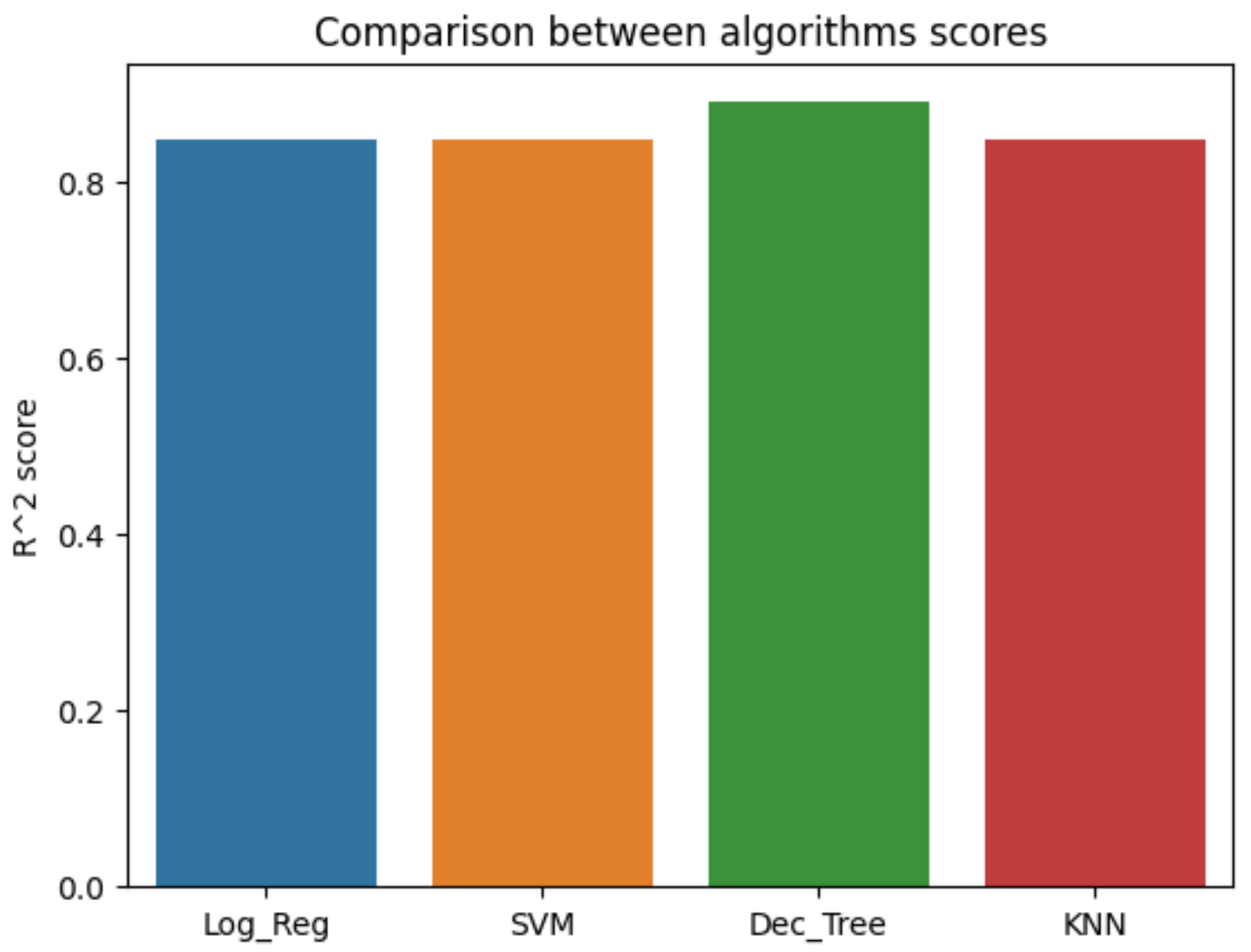


- A chart showing the relationship between landing success and the rocket's mass in kilograms

Section 5

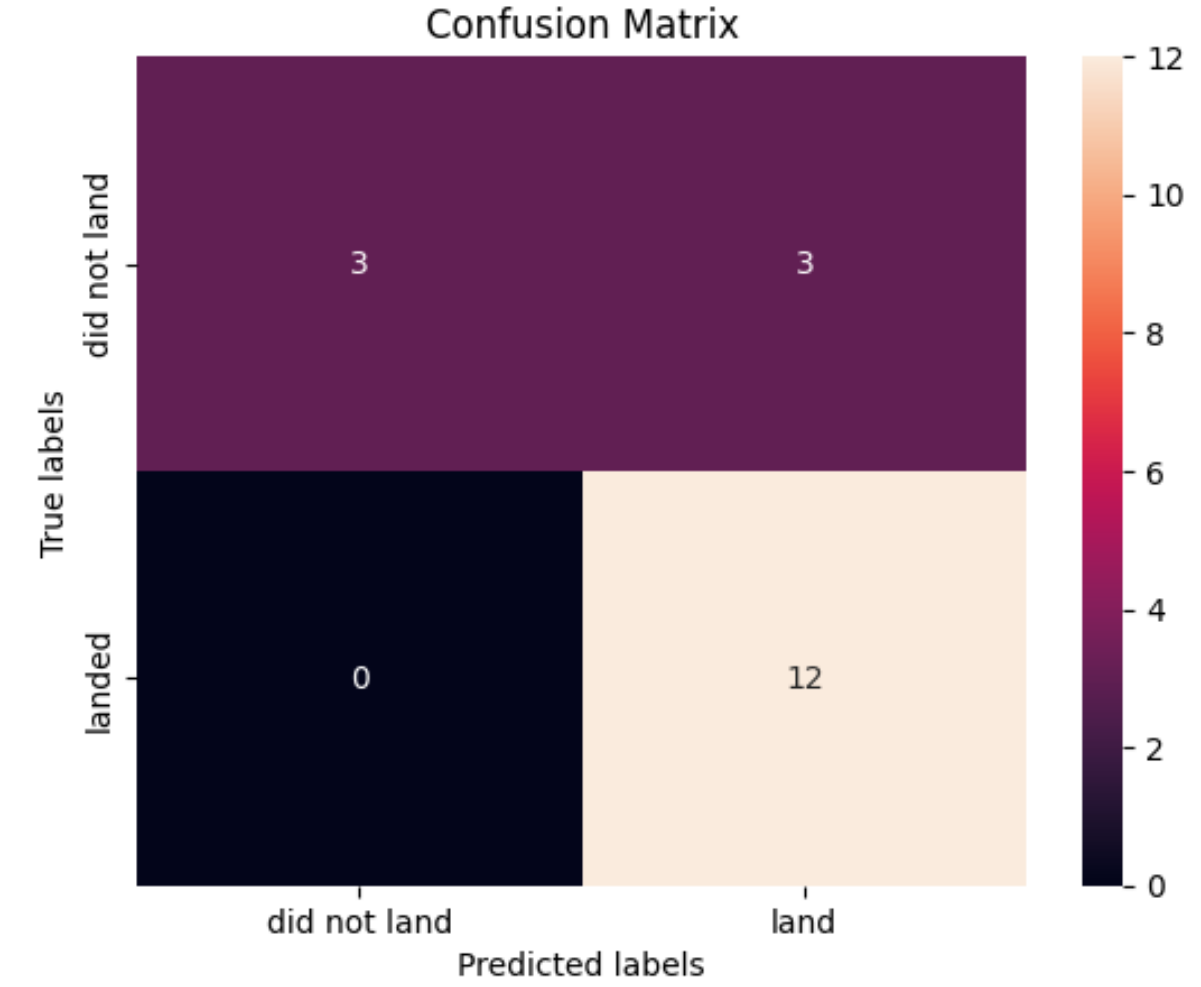
Predictive Analysis (Classification)

Classification Accuracy



- The bar chart represents the accuracy percentage of each model trained on the data, aiming for greater precision in classifying the data and improving predictions on whether future landings will be successful or failed.

Confusion Matrix



- The confusion matrix for the test data, using decision tree classification, as it was the model with the highest accuracy.

Conclusions

- Point 1: The success rate of first stage landing leads to a huge probability of savings for future rocket launches.
- Point 2: The attributes that can affect the success of a successful first stage landing are varied; in this case more than eighty attributes will be considered for the estimation of the calculations.
- Point 3: SpaceX's Falcon 9 launch sites were all close to a highway, a railroad, and a coastline, making them easy to access and reduce transportation costs.
- Point 4: SpaceX’s success rate has increased over the years, with the KSC LC-39A site having the highest success rate.
- Point 5: Orbit and booster version greatly affect success rate

Thank you!