IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Anais Velazquez
October 21, 2024

Outline

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

Executive Summary

We used these data science methodologies:

- Collecting data
- Data Wrangling
- Exploratory Analysis using different tools
- Visual Analytics
- Predictive Analysis

To discovered the price for each launch and see if the first stage will land successfully, using the data science methodologies listed above.

Introduction

Space Y that would like to compete with SpaceX. Space Y was founded by Billionaire industrialist Allon Musk.

In this project, we wanted to find out a couple of things that makes SpaceX successful. Firstly, we needed to gather information about Space X and determine if SpaceX will reuse the first stage. As well as if the first stage lands successfully. To find out this information, we trained a machine learning model and used public information to predict if SpaceX will reuse the first stage.

In additional we also wanted to determine the cost of each launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX REST API (capsules, cores, launches/past)
- Perform data wrangling
  - Normalized the data from JSON to CSV.
  - Cleaned data (NULLs)
  - Sampling data
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using **multiple** classification models.

Data Collection

We collected using SpaceX REST API, specifically on capsules, core, past launches, boosters, launchpad, and payload.

Using these APIs we found information on:

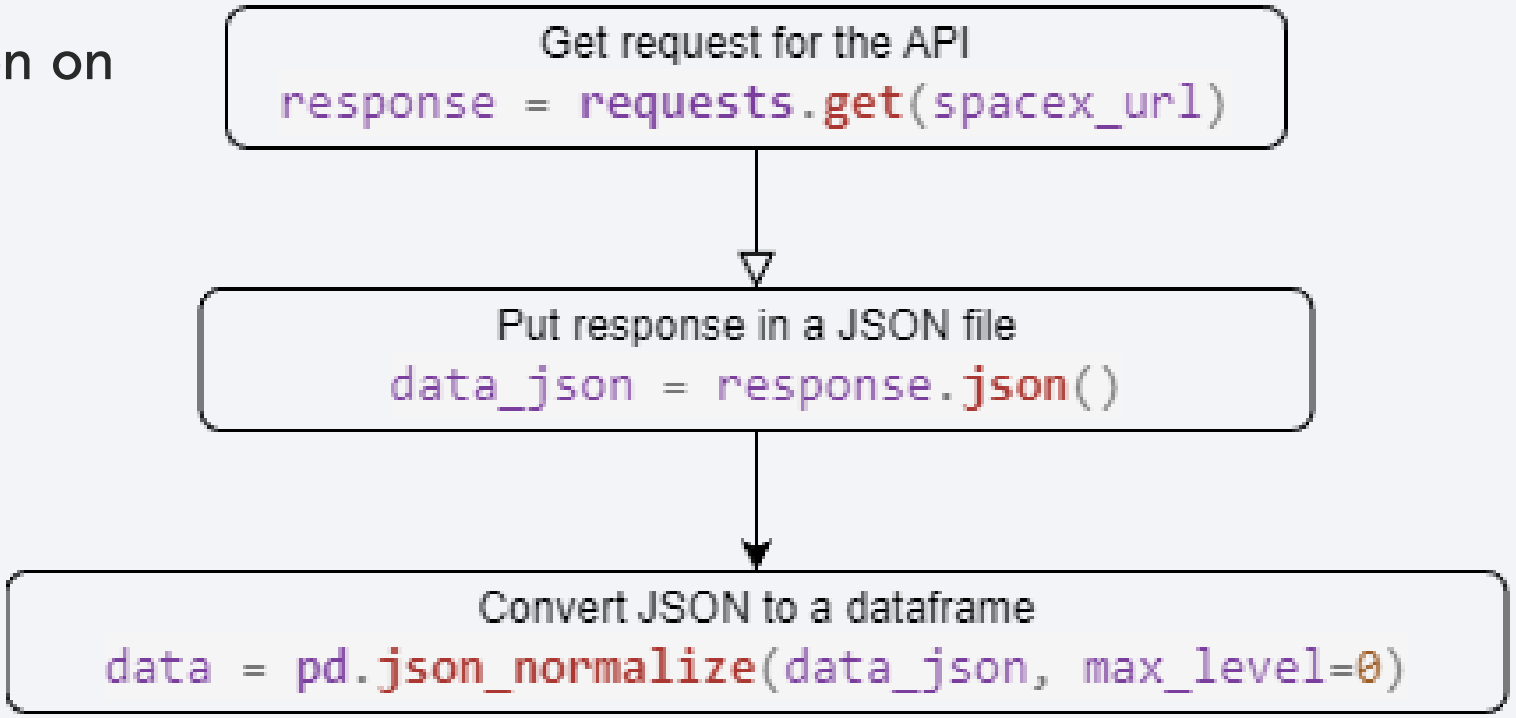
- Booster version
- Launch site (name of site, longitude, and latitude)
- Payload mass and orbit
- etc.

We also conducted web scraping using Wikipedia provide tables on launch records.

Data Collection – SpaceX API

- Here is the API call using past launches to get the information on all past launches' details.

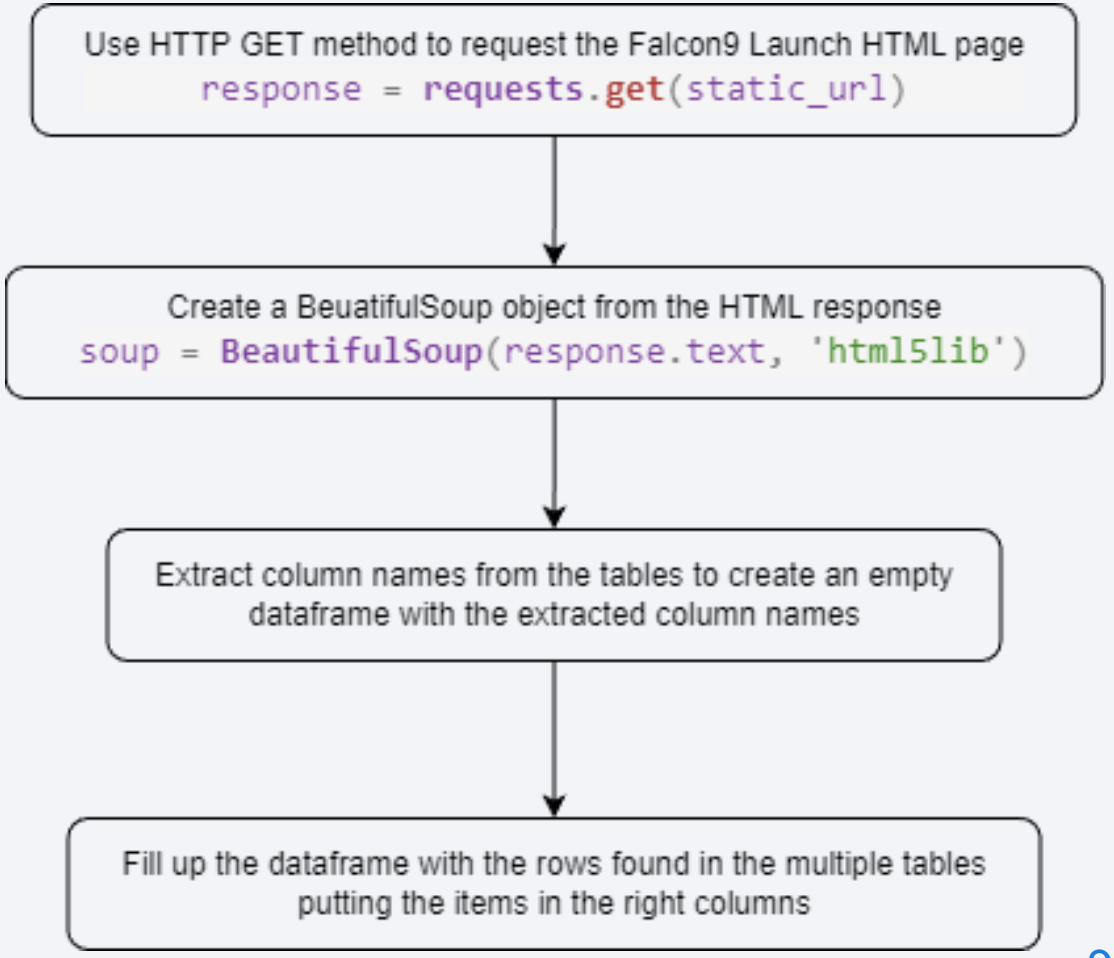
GitHub



Data Collection - Scraping

We using the “List of Falcon 9 and Falcon Heavy launches” Wikipedia page we took launch information from 2010-2021.

GitHub



Data Wrangling

In order to clean up the data I did the following:

- Checked for NULLs within the dataframe, none were found.
- Created a variable that represented the classification variable on the outcome of the launch.
  - The variable was name class and had two classifications, 0 and 1. The meaning of 0 might the first stand did not land successfully, while the 1 meant it did land successfully.

GitHub

EDA with Data Visualization

The plots created through EDA with Data Visualization are:

- Scatterplots:
  - Flight Number vs Payload Mass (kg)
  - Flight Number vs Launch Site
  - Launch Site vs Payload Mass (kg)
  - Flight Number vs Orbit
  - Orbit vs Payload Mass (kg)
- Orbit vs Success Rate bar plot
- Year vs Success Rate line plot

GitHub

EDA with SQL

- Unique launch sites used in the space mission
  - SELECT DISTINCT LAUNCH_SITE FROM SPACEXTABLE
- Launch sites begin with the string 'CCA'
  - `SELECT * FROM SPACEXTABLE WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5`
- Total payload mass carried by boosters launched by NASA (CRS)
  - SELECT SUM(PAYLOAD_MASS__KG_) AS 'TOTAL PAYLOAD MASS BY NASA (CRS)' FROM SPACEXTABLE WHERE CUSTOMER = 'NASA (CRS)'
- Average payload mass carried by booster ver. F9 v1.1
  - SELECT AVG(PAYLOAD_MASS__KG_) AS 'AVERAGE PAYLOAD MASS BY BOOSTER VER. F9 V1.1' FROM SPACEXTABLE WHERE BOOSTER_VERSION = 'F9 V1.1'
- Date of the 1st successful landing on a ground pad
  - SELECT DATE AS 'FIRST SUCCESSFUL LANDING OUTCOME ON GROUND PAD' FROM SPACEXTABLE WHERE LANDING_OUTCOME LIKE '%SUCCESS%GROUND%' LIMIT 1

EDA with SQL

- List of names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
  - SELECT BOOSTER_VERSION AS 'BOOSTERS IN SUCCESSFUL DRONE SHIP LANDINGS W/ 6000>PAYLOAD MASS>4000' FROM SPACEXTABLE WHERE LANDING_OUTCOME LIKE '%SUCCESS%DRONE%' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000
- Total number of successful & failure missions
  - SELECT MISSION_OUTCOME, COUNT(*) AS COUNT FROM SPACEXTABLE GROUP BY MISSION_OUTCOME
- List of names of the booster that have carried the maximum payload mass
  - SELECT DISTINCT BOOSTER_VERSION AS 'BOOSTER VERSIONS W/ THE MAX PAYLOAD MASS' FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)

EDA with SQL

- 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.
  - SELECT SUBSTR(DATE,6,2) AS MONTH, LANDING_OUTCOME, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTABLE WHERE LANDING_OUTCOME LIKE '%FAILURE%DRONE%' AND SUBSTR(DATE,0,5) ='2015'
- Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order.
  - SELECT LANDING_OUTCOME, COUNT(*) AS COUNT FROM SPACEXTABLE WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING_OUTCOME ORDER BY COUNT DESC

GitHub

Build an Interactive Map with Folium

On the interactive map using Folium you will see

- Markers that are used to mark where the launch sites are on the map
- Circle are used to show the launches that were successful and failures at each launch sites
- Lines are used to calculate the proximity of other landmarks from the launch sites

GitHub

Build a Dashboard with Plotly Dash

On the dashboard, you will see the following plots:

- A pie chart with a dropdown menu that shows
  - All sites comparison of successful launches
  - Each individual launch site comparison between successful and failure launches
- A scatterplot with a slider that sets the payload range.
  - Relation between the mission outcome and the payload mass. The color of each data point is also marked for booster version category.

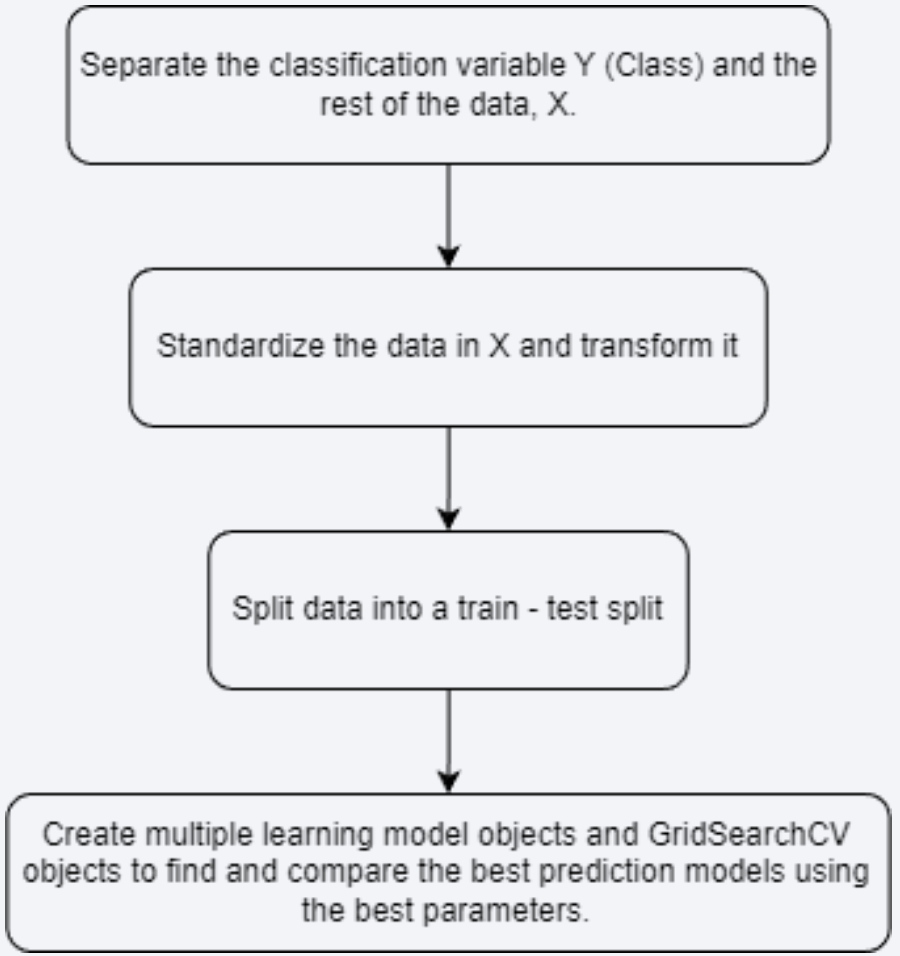
GitHub

Predictive Analysis (Classification)

When creating the prediction models, we used the flowchart to create them. We different predictive models we used are:

- Logistic Regression
- Support Vector Machine
- Decision Tree
- K-nearest Neighbor

GitHub



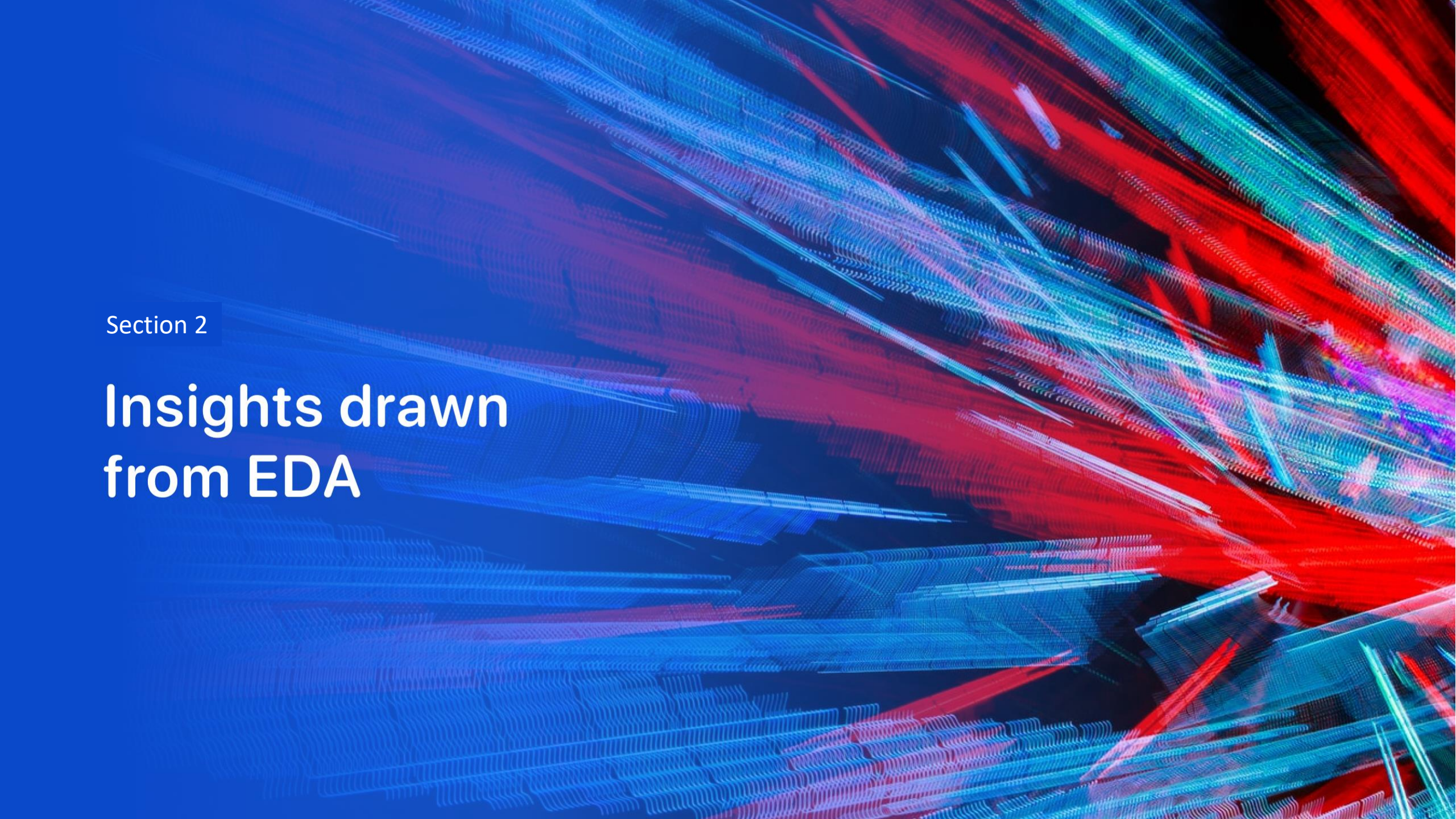
Results

Exploratory Data Analysis (EDA) Results

- Orbits that have been 100% successful were ES-L1, GEO, HEO, and SSO.
- Success rate was rising in 2013 – 2017, except 2014 where it was stable.
- Most mission outcomes were successful except for 1 failure.

Predictive Analysis Results

- The decision tree model had the highest accuracy and best score.
  - Accuracy : 0.889285714285714 5
  - Score: 0.9444444444444444

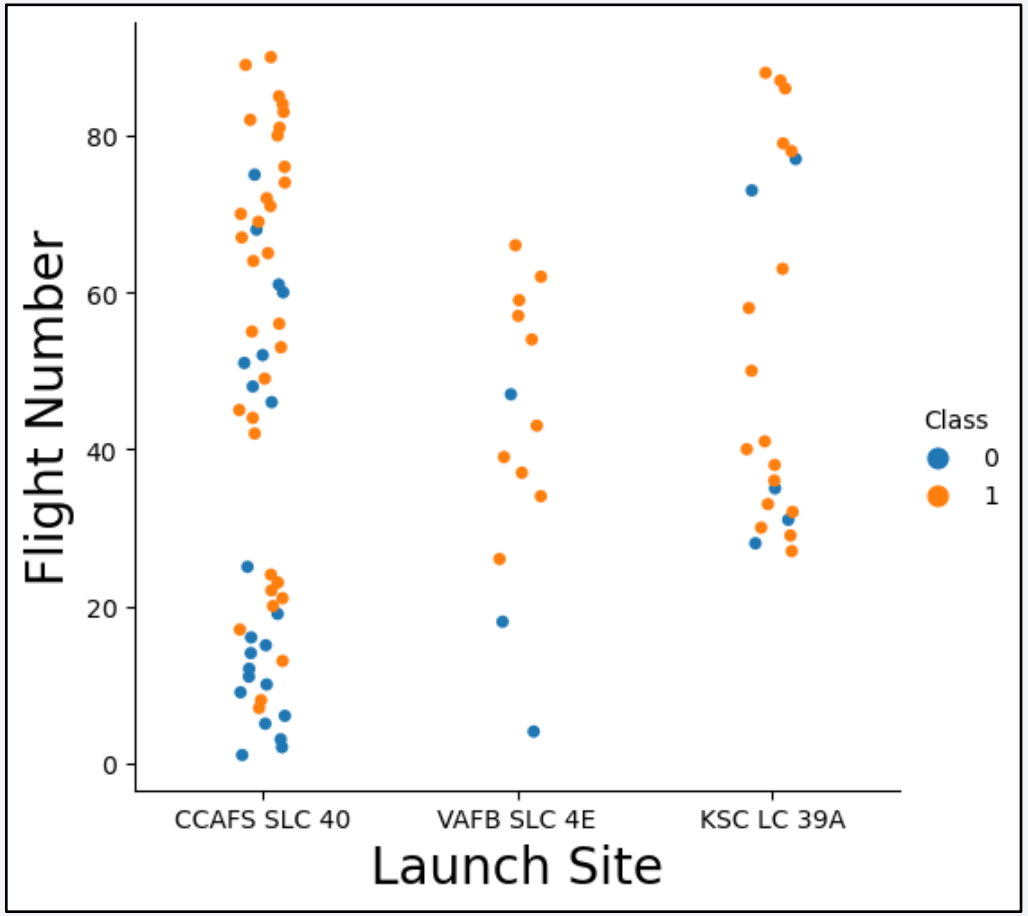

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

This scatter plot shows

- The correlation between the number of flights and their launch sites
- Whether the mission was a success or a failure via the color of the data points.
    - Orange/1 means the mission was a success and Blue/0 means it was a failure.

From the scatter plot we can see that CCAFS SLC 40 launch site had the most launches compared to the rest.

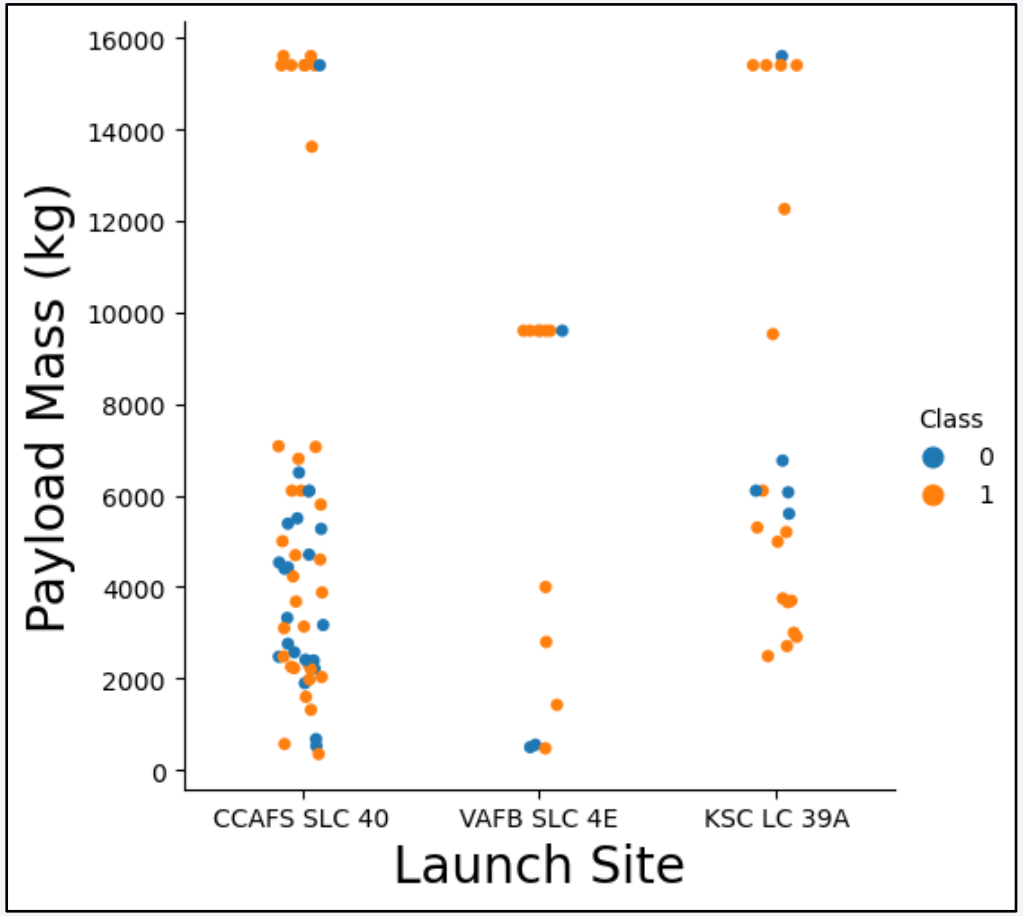


Payload vs. Launch Site

This scatter plot shows

- The correlation between the payload mass in kg and their launch sites
- Whether the mission was a success or a failure via the color of the data points.
    - Orange/1 means the mission was a success and Blue/0 means it was a failure.

From the scatter plot we can see that CCAFS SLC 40 and KSC LC 39A launch sites had the highest payload mass (<16000kg) while VAFB SLC 4E had its highest payload mass at <10000kg.

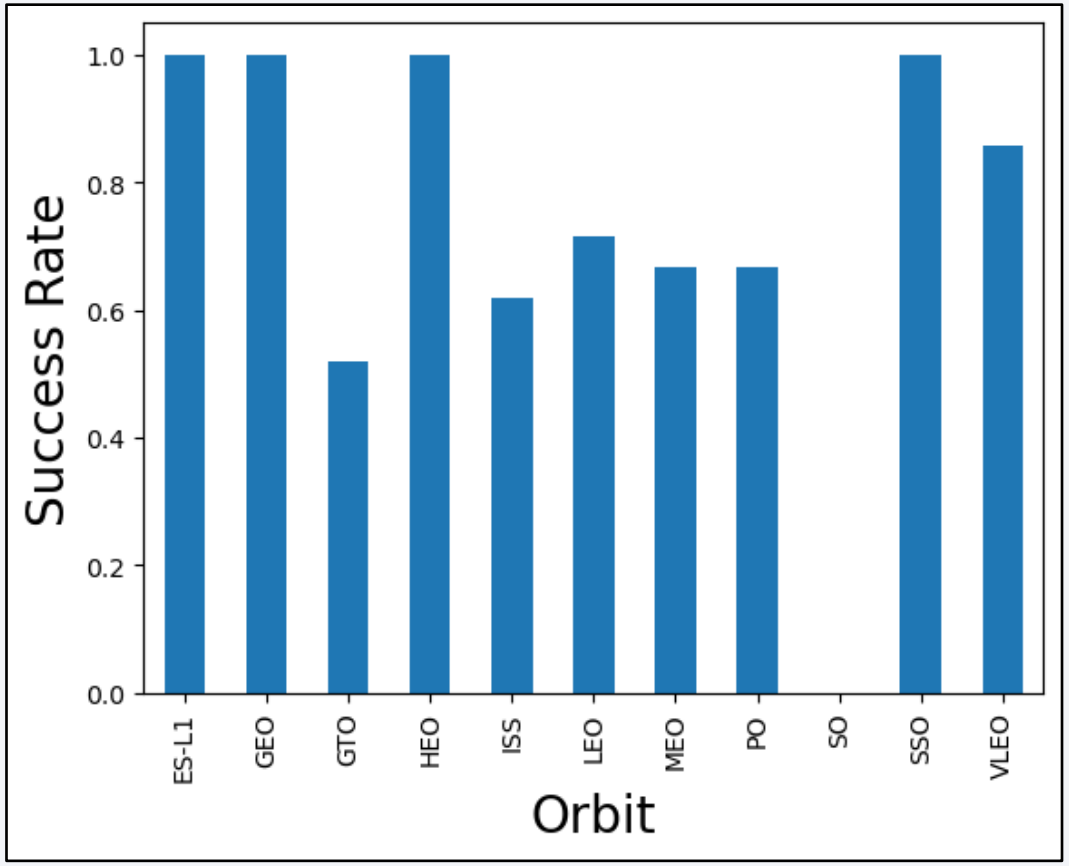


Success Rate vs. Orbit Type

This bar chart shows

- The correlation between the success rate and the orbit the mission was set to.

From this bar chart we can see that the orbits that had a 100% success rate were ES-L1, GEO, HEO, and SSO. While SO had a lowest success rate at 0%.



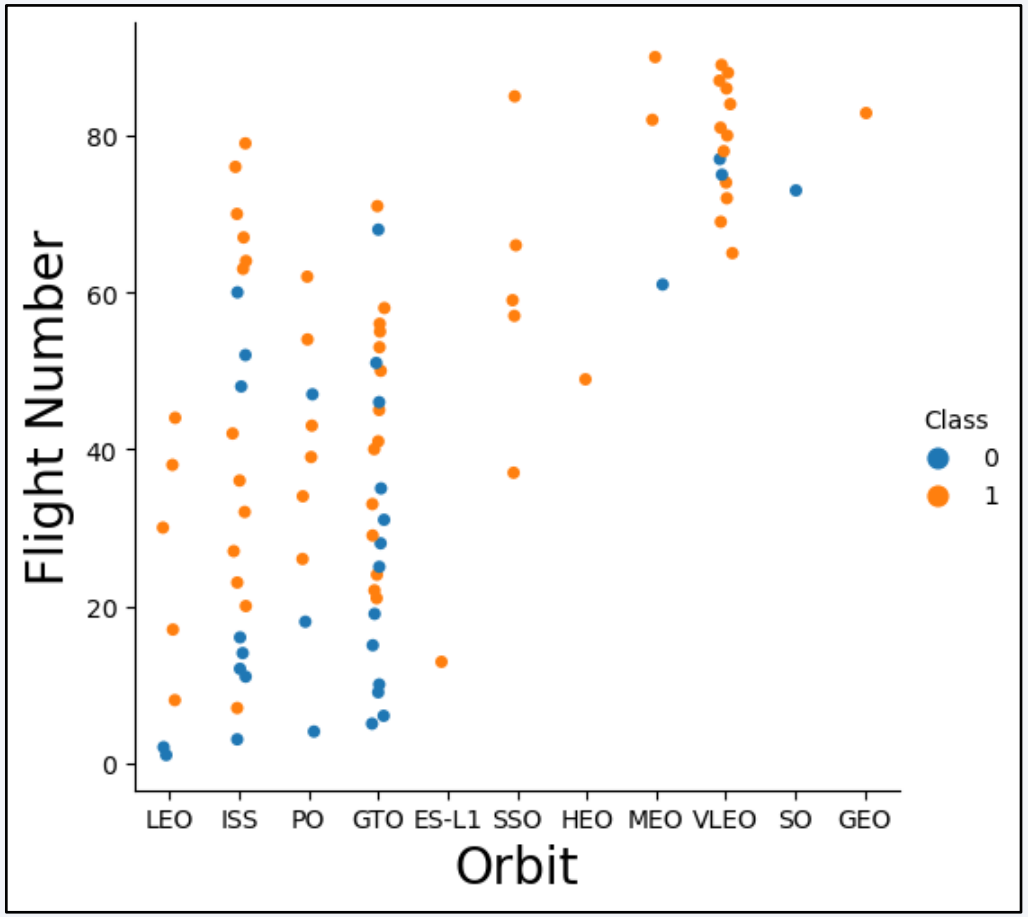
Flight Number vs. Orbit Type

This scatter plot shows

- The correlation between the flight number and the orbit the mission was set to.
- Whether the mission was a success or a failure via the color of the data points.
    - Orange/1 means the mission was a success and Blue/0 means it was a failure.

Taking what we learned from the last slide, we can see why the orbits with the highest success rate were high. Only 1 flight that were successful means the success rate 100%.

The only orbit that has >1 flight and a success rate of 100% is SSO, with 5 flights being successful in that orbit.



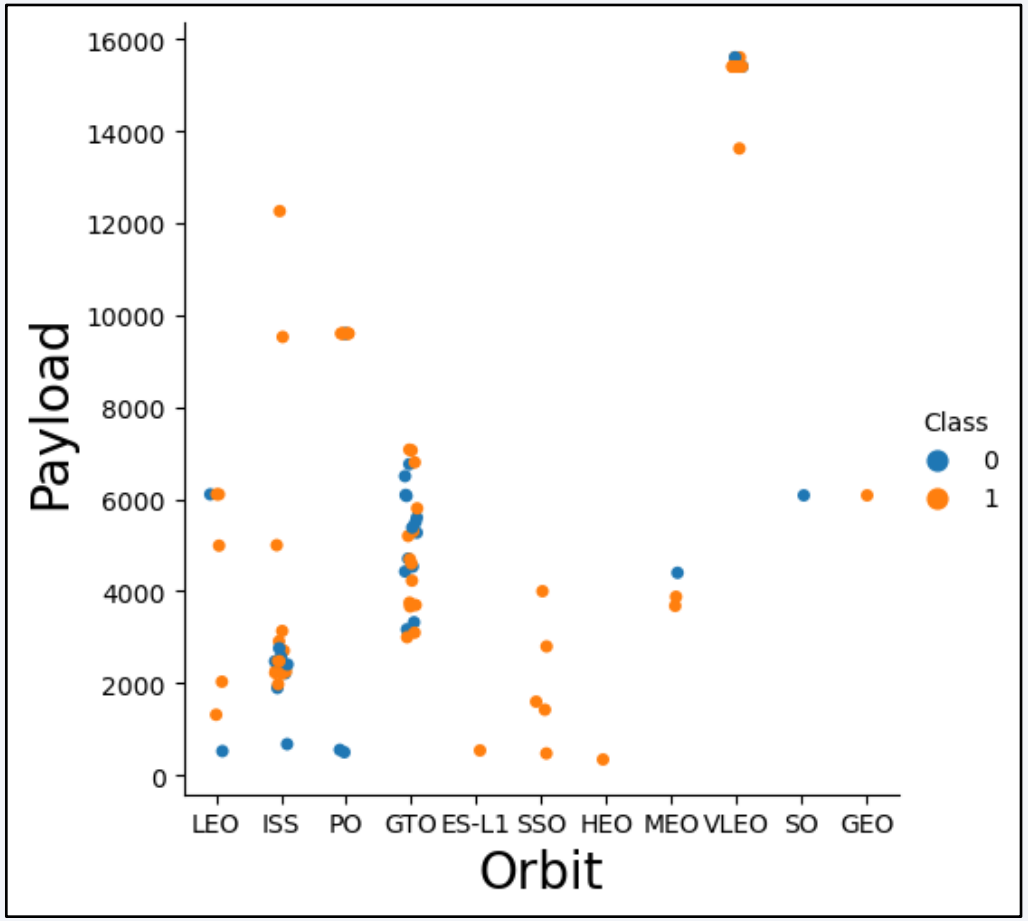
Payload vs. Orbit Type

This scatter plot shows

- The correlation between the payload mass in kg and the orbit the mission was set to.
- Whether the mission was a success or a failure via the color of the data points.
    - Orange/1 means the mission was a success and Blue/0 means it was a failure.

Taking what we learned from the last 2 slide, we can see that SSO having the highest number of flights with 100% success rate had a very low payload mass.

Similarly, the other launches that had 100% success rate are low in payload mass, as well.

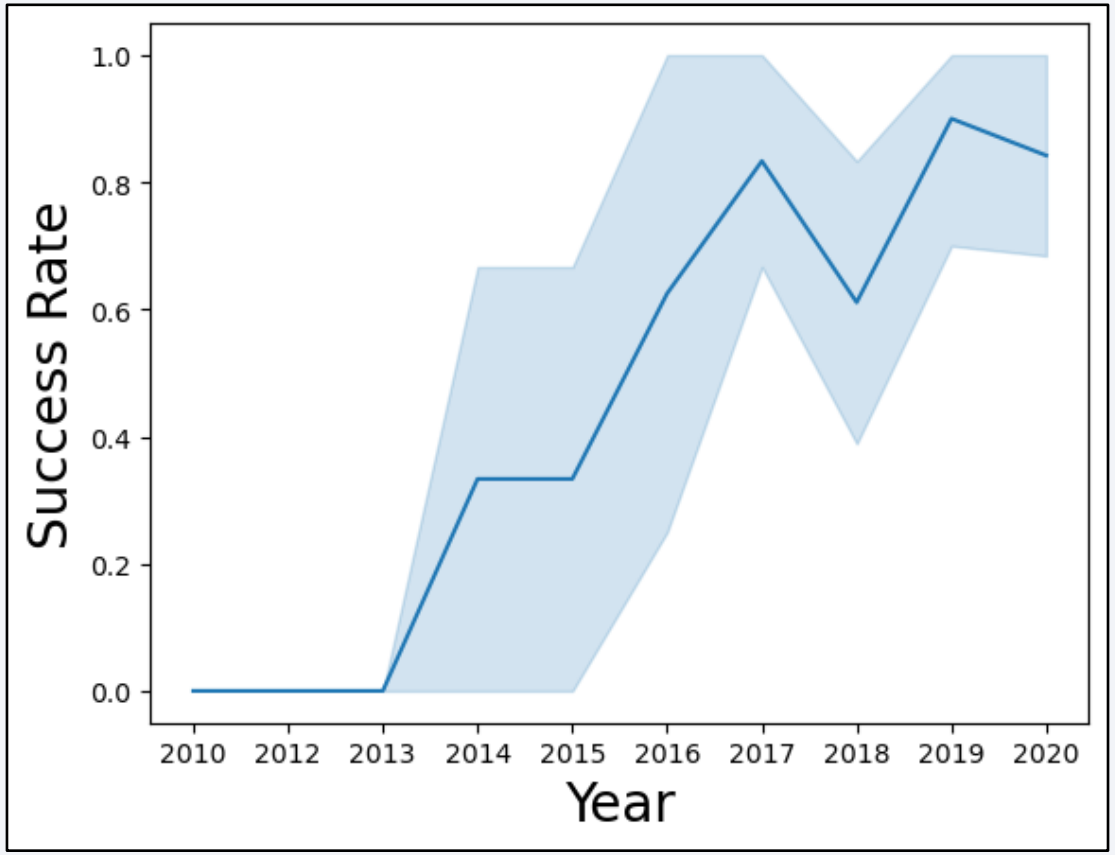


Launch Success Yearly Trend

This line graph shows

- The correlation between the success rate and the year.

As we can see by the line graph the success rate since 2013 kept increasing till 2017 (stable in 2014) and after 2015 it started increasing. In 2018 the success rate took a dip then went back up in 2019.



All Launch Site Names

Query:
Unique launch sites used in the space mission

- SELECT DISTINCT LAUNCH_SITE FROM SPACEXTABLE

This query shows the unique names of the launch sites in all the space missions for SpaceX. As we can see there are only four, CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, and CCAFS SLC-40.

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

Launch Site Names Begin with 'CCA'

Query:

Launch sites begin with the string 'CCA'

- `SELECT * FROM SPACEXTABLE WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5`

This query shows the first 5 launches information that name begins with 'CCA'.

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

Query:
Total payload mass carried by boosters launched by NASA (CRS)

- SELECT SUM(PAYLOAD_MASS__KG_) AS 'TOTAL PAYLOAD MASS BY NASA (CRS)' FROM SPACEXTABLE WHERE CUSTOMER = 'NASA (CRS)'

Total Payload Mass by NASA (CRS)
45596

This query shows the total payload mass by NASA (CRS) which is 45596kg.

Average Payload Mass by F9 v1.1

Query:

Average payload mass carried by booster ver. F9 v1.1

- SELECT AVG(PAYLOAD_MASS__KG_) AS 'AVERAGE PAYLOAD MASS BY BOOSTER VER. F9 V1.1' FROM SPACEXTABLE WHERE BOOSTER_VERSION = 'F9 V1.1'

This query shows that average payload mass carried by booster ver. F9 v1.1 is 2928.4 kg.

Average Payload Mass by Booster ver. F9 v1.1
2928.4

First Successful Ground Landing Date

Query:

Date of the 1st successful landing on a ground pad

- SELECT DATE AS 'FIRST SUCCESSFUL LANDING OUTCOME ON GROUND PAD' FROM SPACEXTABLE WHERE LANDING_OUTCOME LIKE '%SUCCESS%GROUND%' LIMIT 1

This query shows date of the first successful landing outcome on a ground pad was December 22, 2015.

First Successful Landing Outcome on Ground Pad
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Query:

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

- SELECT BOOSTER_VERSION AS 'BOOSTERS IN SUCCESSFUL DRONE SHIP LANDINGS W/ 6000>PAYLOAD MASS>4000' FROM SPACEXTABLE WHERE LANDING_OUTCOME LIKE '%SUCCESS%DRONE%' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000

This query shows that F9 FT B1022, F9 FT B1026, F9 FT B1021.2, and F9 FT B1031.2 were the boosters that have had a successful landing on a drone shop and have a 4000<payload mass<6000.

Boosters in Successful Drone Ship Landings w/ 6000>Payload mass>4000
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

Query:

Total number of successful & failure missions

- SELECT MISSION_OUTCOME, COUNT(*) AS COUNT FROM SPACEXTABLE GROUP BY MISSION_OUTCOME

This query shows that only 1 mission outcome was a failure (in flight).

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

Boosters Carried Maximum Payload

Query:

List of names of the booster that have carried the maximum payload mass

- SELECT DISTINCT BOOSTER_VERSION AS 'BOOSTER VERSIONS W/ THE MAX PAYLOAD MASS' FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)

This query shows that these booster versions have carried the maximum payload mass which is around 16000kg.

Booster Versions w/ the max Payload Mass
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

Query:

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.

- SELECT SUBSTR(DATE,6,2) AS MONTH, LANDING_OUTCOME, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTABLE WHERE LANDING_OUTCOME LIKE '%FAILURE%DRONE%' AND SUBSTR(DATE,0,5) ='2015'

This query shows that in January and in April of 2015 these were the only two launches with failure landing outcomes on drone ships.

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

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

Query:

Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order.

- SELECT LANDING_OUTCOME, COUNT(*) AS COUNT FROM SPACEXTABLE WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING_OUTCOME ORDER BY COUNT DESC

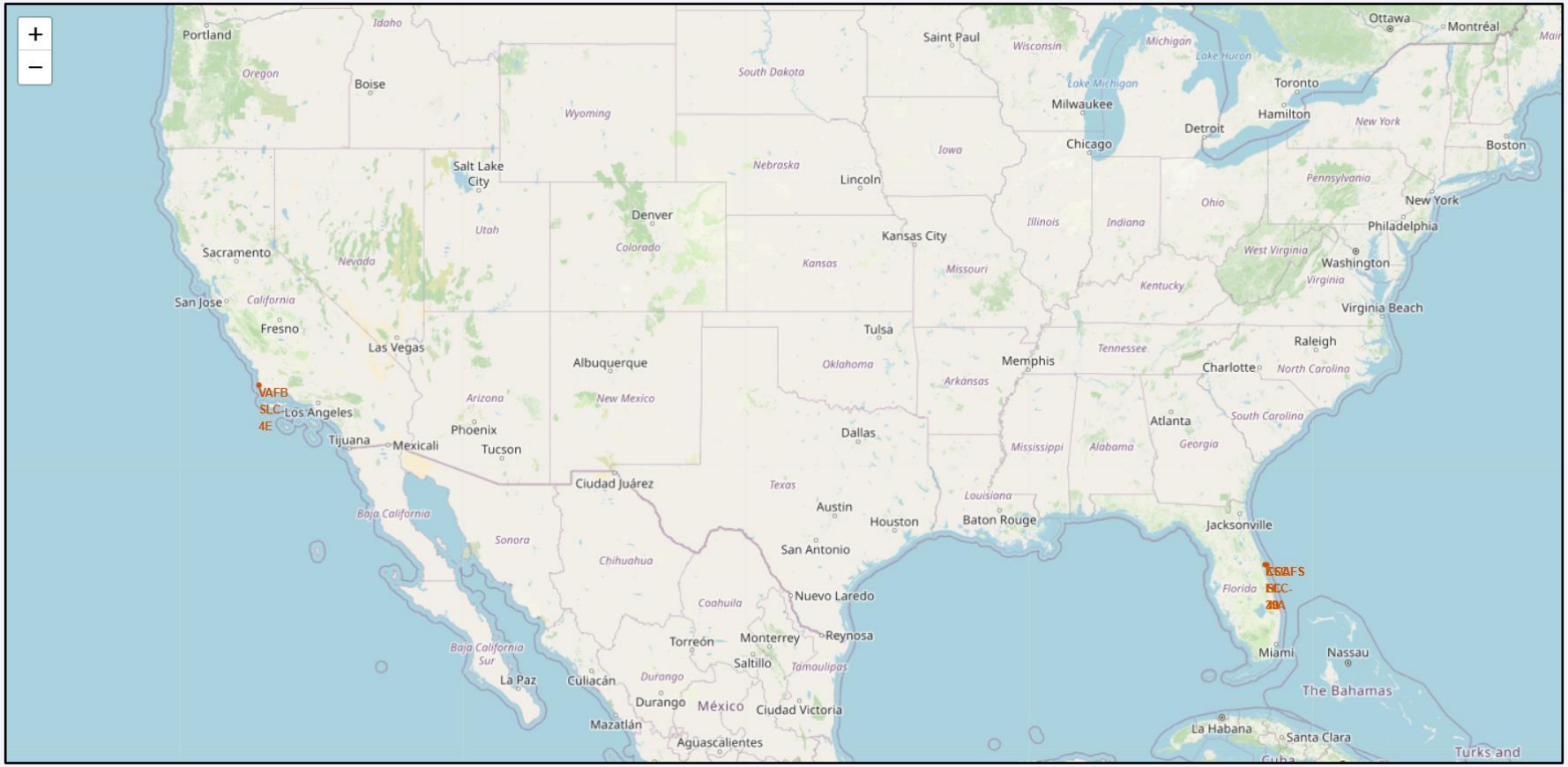
This query shows that no attempt of landing outcome was the highest in count between the dates June 4, 2010 and March 20, 2017.

Landing_Outcome	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

Section 3

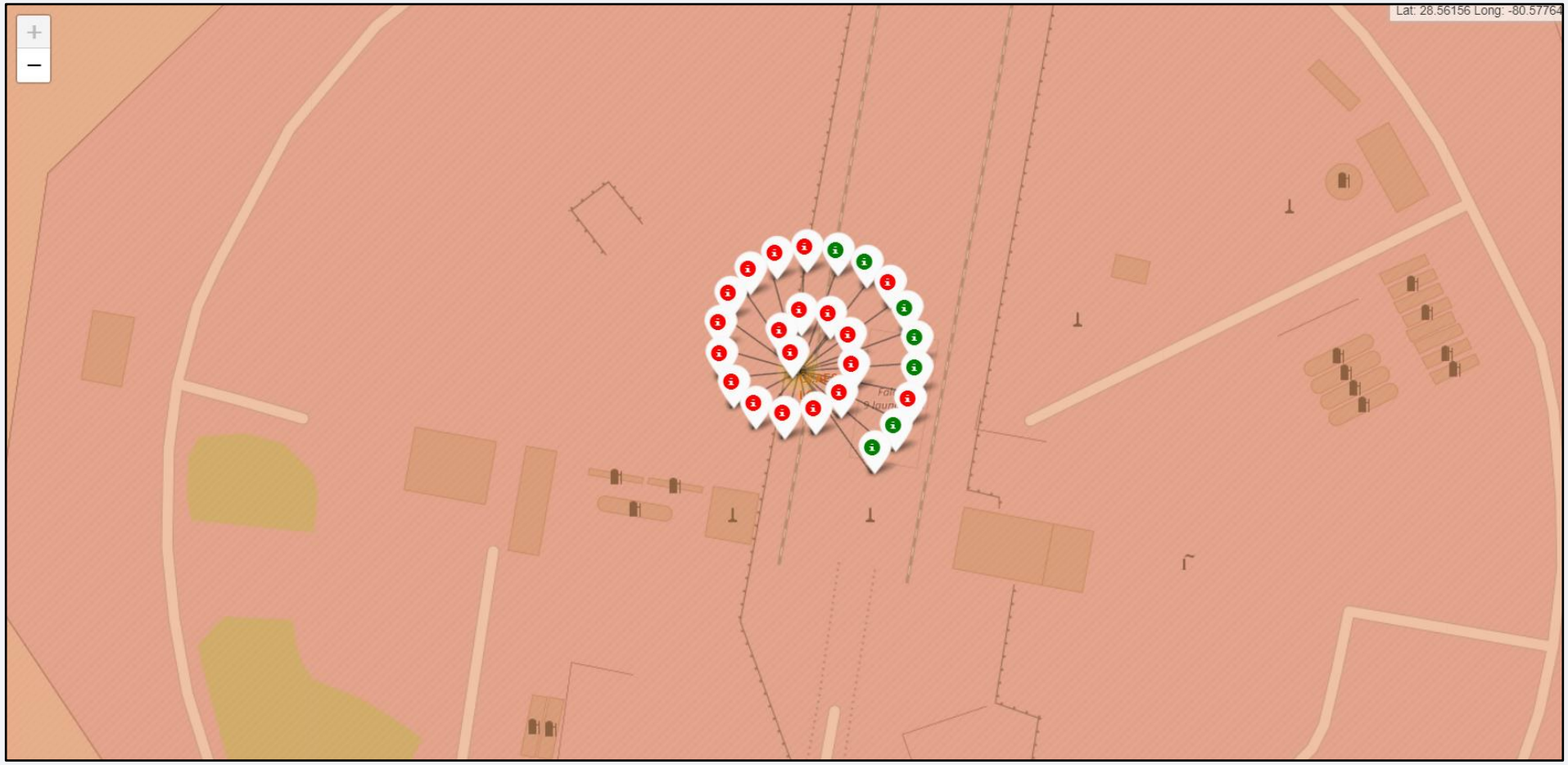
Launch Sites Proximities Analysis

Launch Sites' Markers on a Global Map



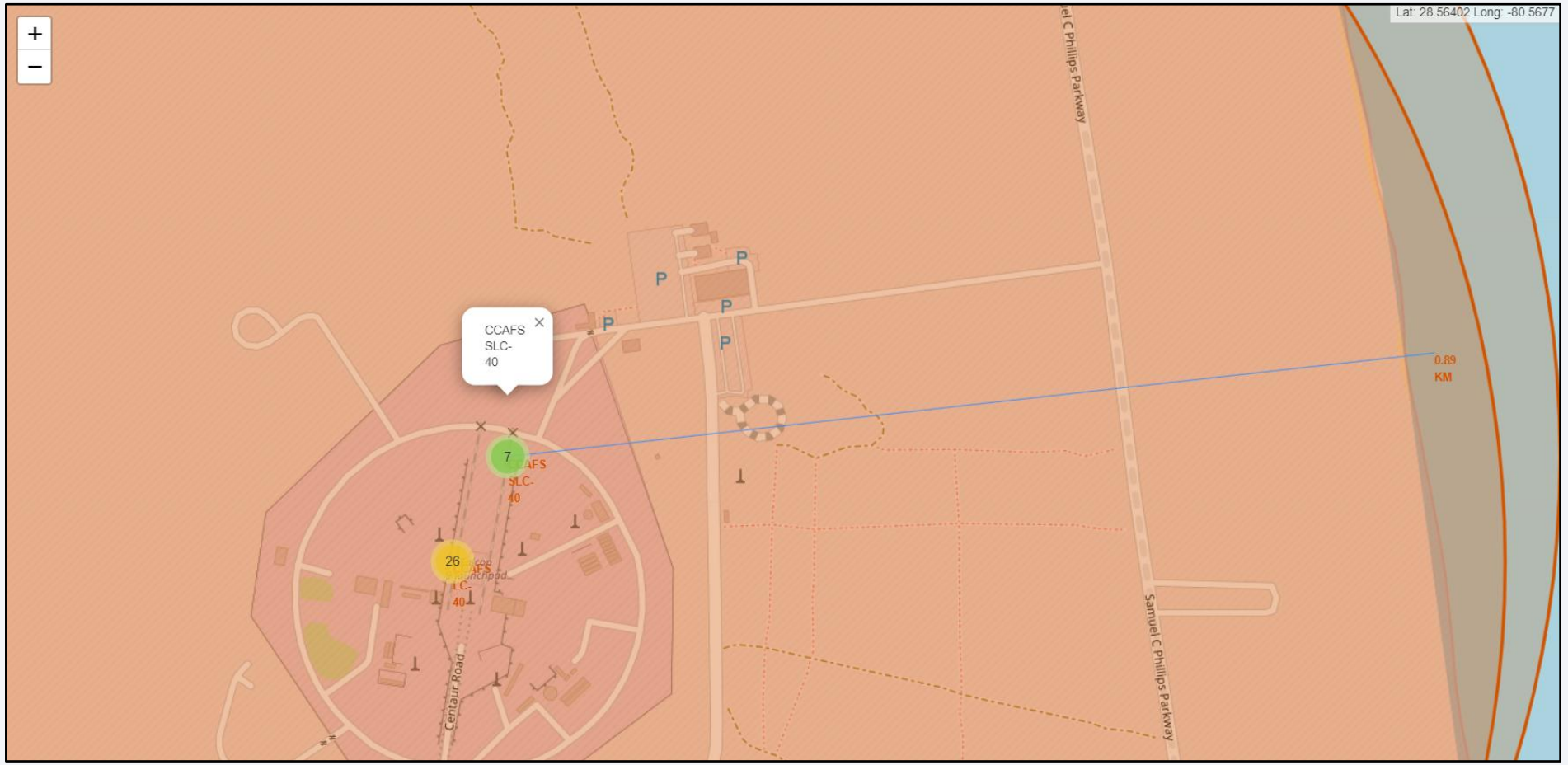
We can see Florida and California are where all the launch sites are located.

Color-labeled Launch Outcomes

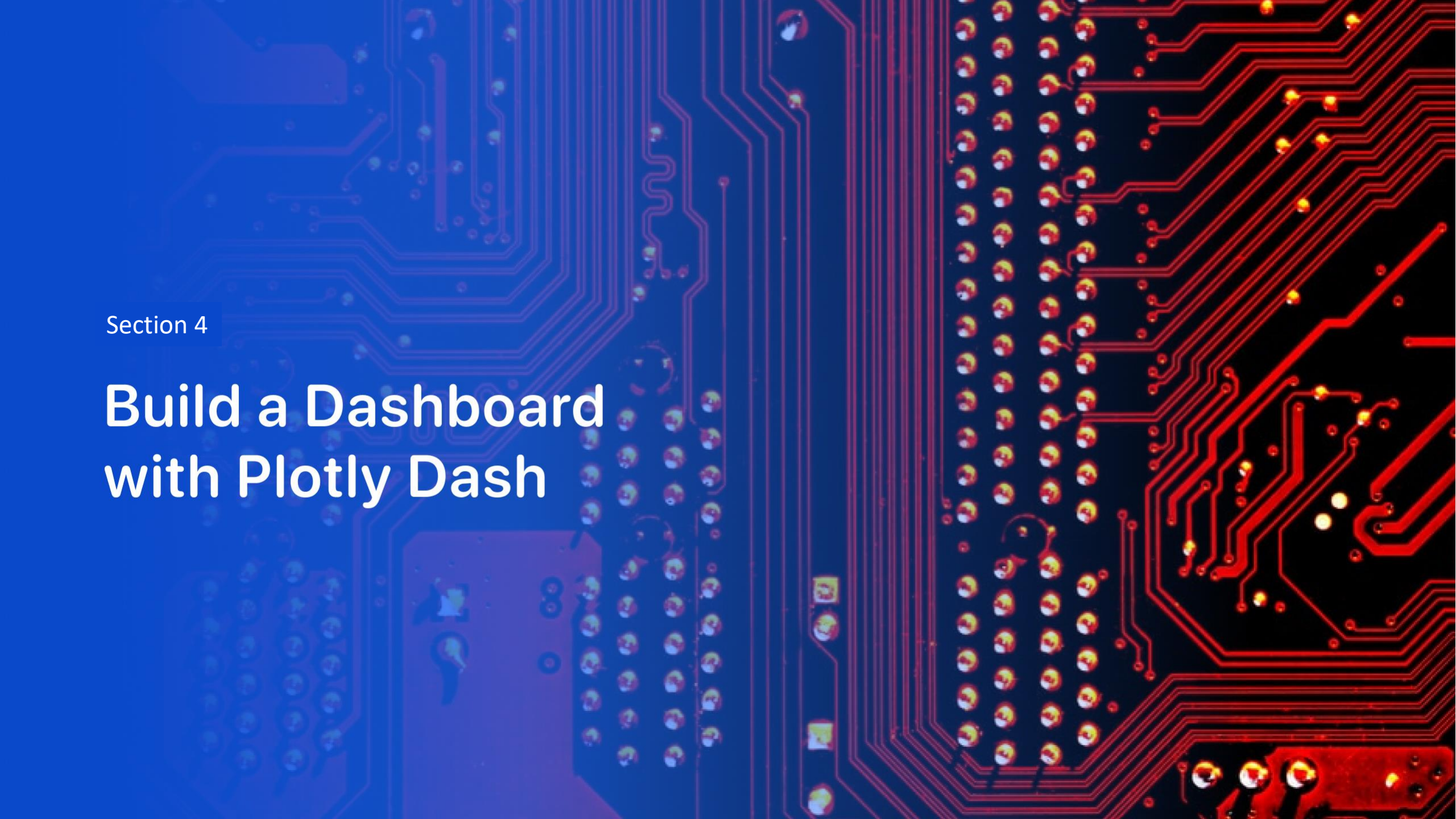


We can see at this launch site, CCAFS LC-40, the launch outcomes tend to be failures.

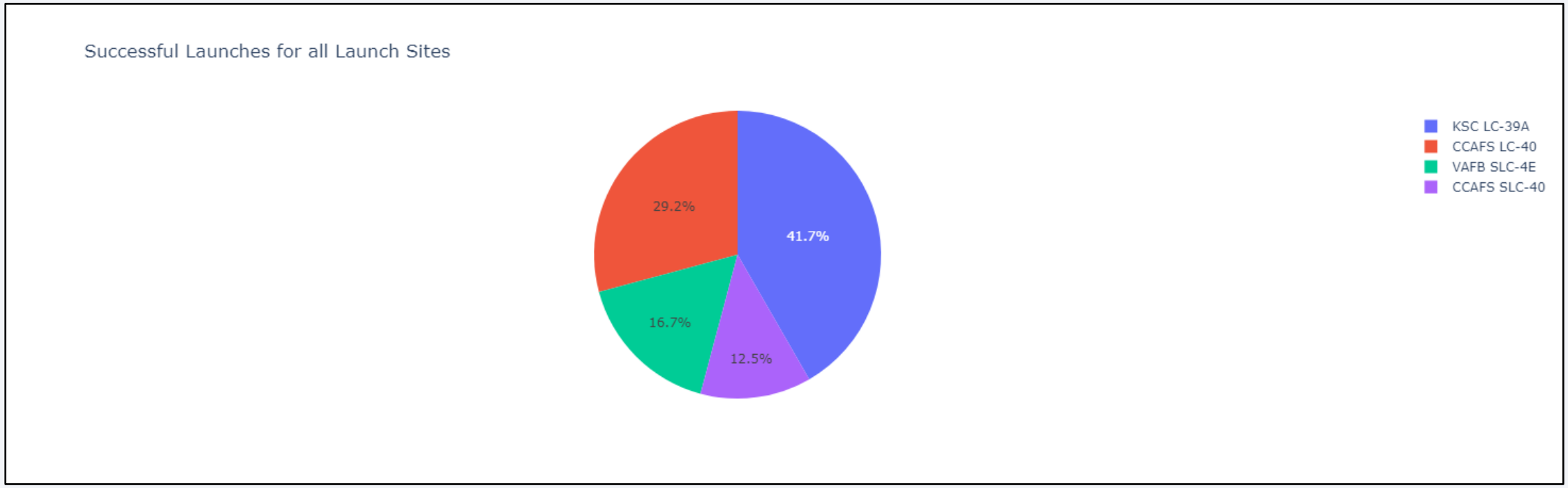
Launch Site and its Proximities to the Coastline



We can see at this launch site, CCAFS SLC-40, is approximately .89km from the coastline.

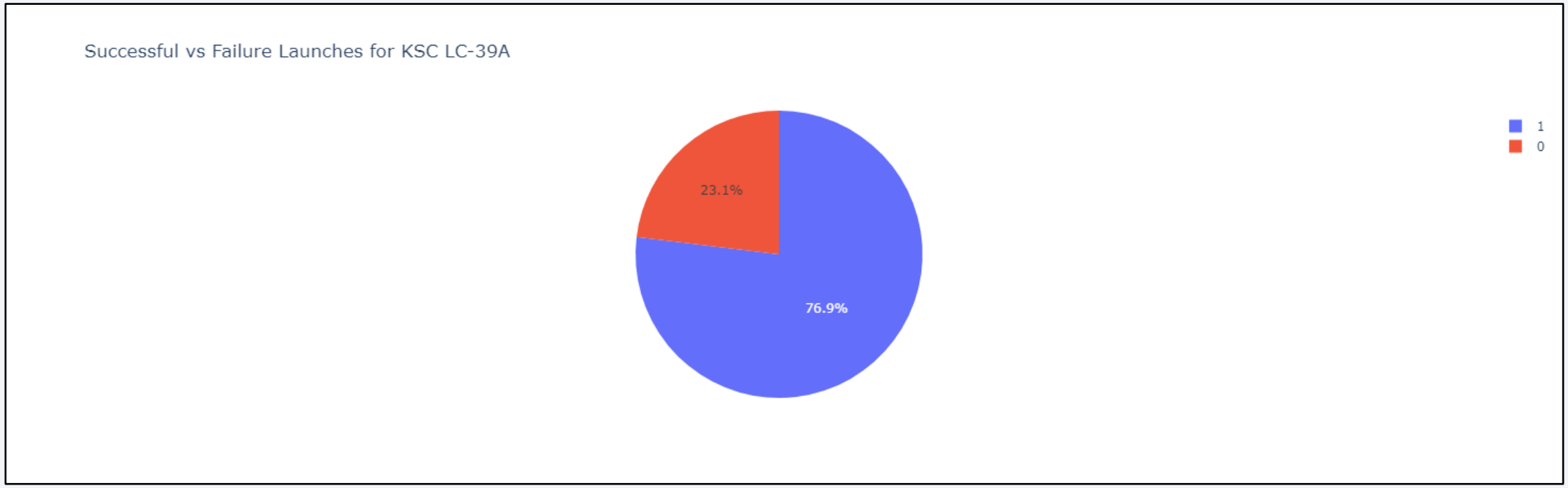

Section 4
Build a Dashboard with Plotly Dash

Launch Success Count for All Sites



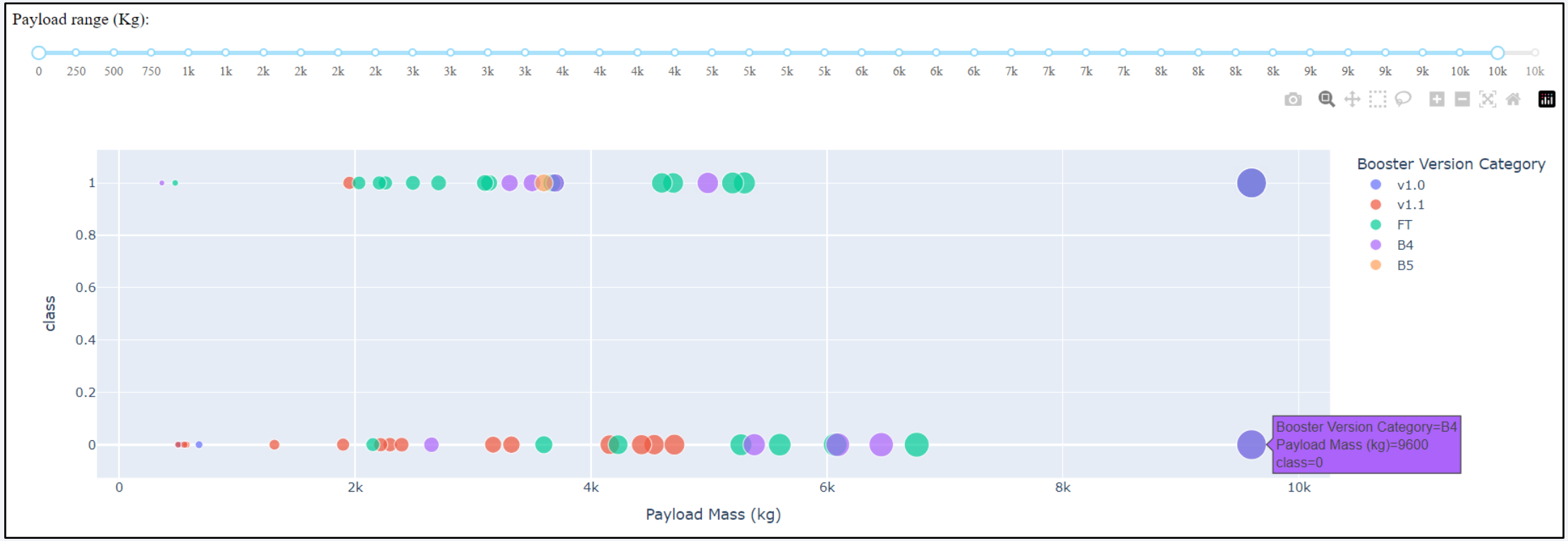
The launch site with the highest launch success ratio is KSC LC-39A.

Launch Site with Highest Launch Success Ratio



KSC LC-39A has a success ratio of 76.9%

Payload vs. Launch Outcome



Shows all the launch sites booster version categories, their outcome (success/failure), and their payload mass.

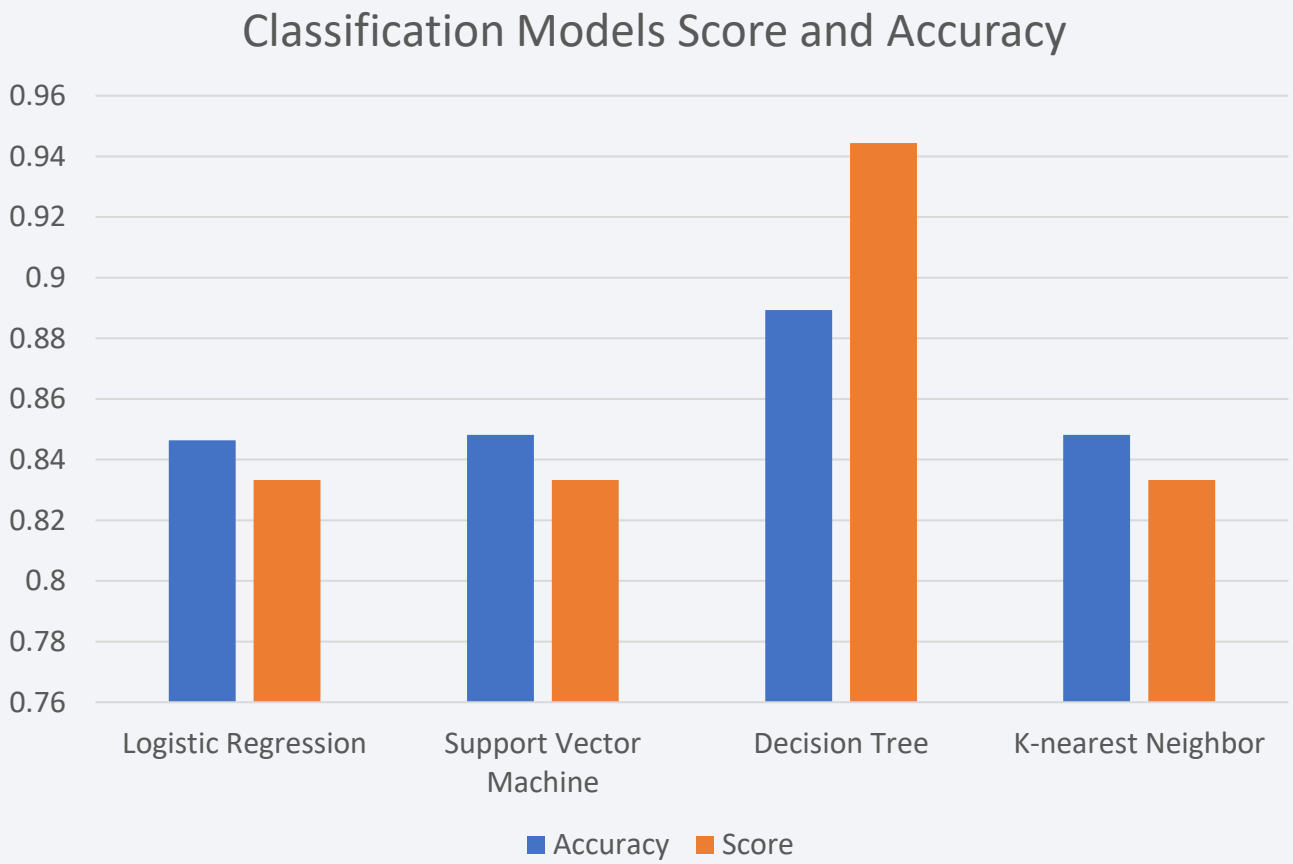
Section 5

Predictive Analysis (Classification)

Classification Accuracy

The models used were:

- Logistic Regression
  - Accuracy : 0.8464285714285713
  - Score: 0.8333333333333334
- Support Vector Machine
  - Accuracy : 0.8482142857142856
  - Score: 0.8333333333333334
- **Decision Tree**
  - **Accuracy : 0.8892857142857145**
  - **Score: 0.9444444444444444**
- K-nearest Neighbor
  - Accuracy : 0.8482142857142858
  - Score: 0.8333333333333334

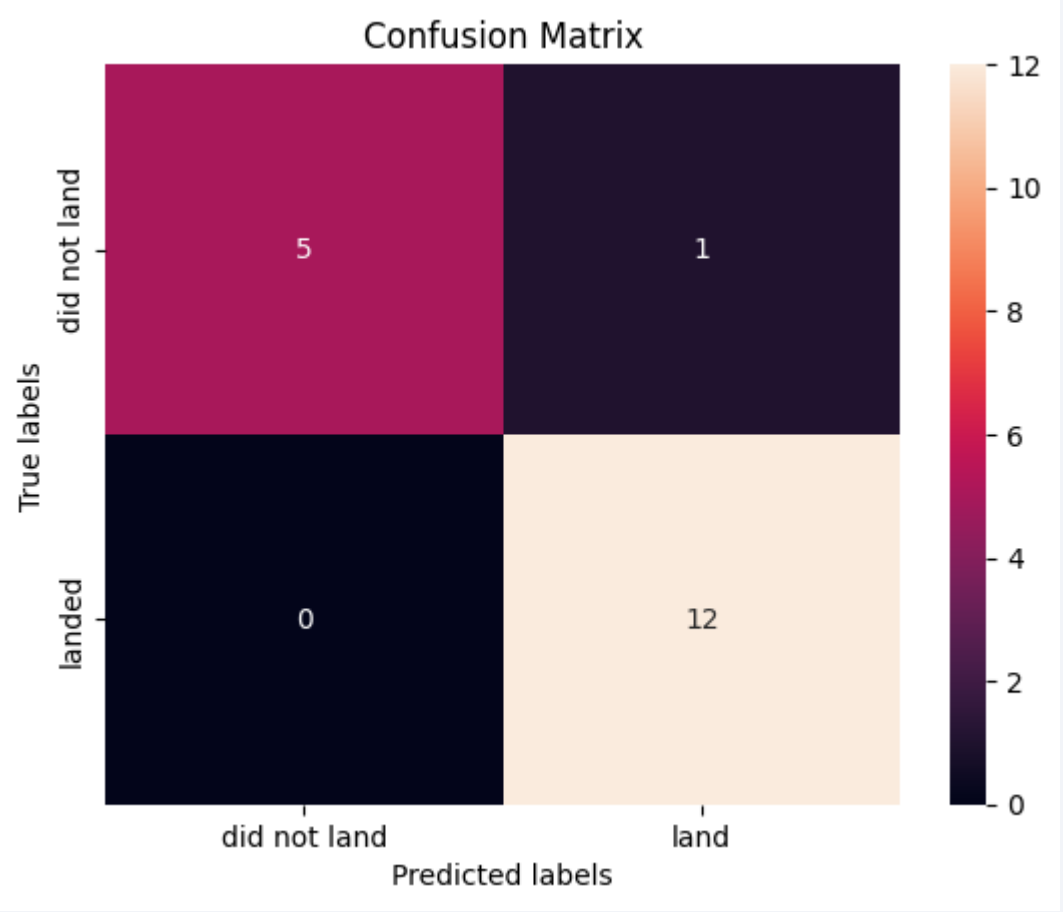


Confusion Matrix

In this confusion matrix, you can see that it incorrectly marked only 1 predicted outcome.

This confusion matrix was created using the **decision tree model** had the highest accuracy and best score.

- Accuracy : 0.889285714285714 5
- Score: 0.9444444444444444



Conclusions

Exploratory Data Analysis (EDA) Results

- Orbits that have been 100% successful were ES-L1, GEO, HEO, and SSO.
  - Learning that SSO has the most launch that were 100% successful.
- Success rate was rising in 2013 – 2017, except 2014 where it was stable.
- Most mission outcomes were successful except for 1 failure.

Predictive Analysis Results

- The decision tree model had the highest accuracy and best score.
  - Accuracy : 0.8892857142857145
  - Score: 0.9444444444444444

Appendix

GitHub Link to complete project files: https://github.com/Avelazquez2/Data-Science-Courses/tree/main/Capstone

Thank you!