IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Dmitrii Lychev
April 17, 2023

Outline

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

Executive Summary

In this capstone project we will predict whether the SpaceX Falcon9 first stage will land successfully. If we can determine if the first stage will land we can determine the cost of a launch. This will be achieved with the use of different machine learning classification algorithms.

The methodology followed will include Data Collection, Data Wrangling and Preprocessing, Exploratory Data Analysis, Data Visualization and Machine Learning Prediction.

During our investigation the results of our analysis indicate that there are some features of rocket launches that have a correlation with the success or failure launches.

In the end we conclude that the Decision Tree may be the best machine learning algorithm to solve this task.

Introduction

Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against space X for a rocket launch.

This brings us to the main question – will the first stage of the rocket land successfully?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

  Data was collected using following two methods: requesting data from the SpaceX API and web scraping data from the Wikipedia page.

- Perform data wrangling

  Data was filtered to contain only relevant information about Falcon 9 launches. Missing values were replaced with mean values.

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

  Four different machine learning algorithms were used for predictive analysis. These are logistic regression, support vector machines, k-nearest neighbour and decision tree classifier. Each model was trained, tuned and evaluated to find the best one.

Data Collection

Data was collected using two methods:

- SpaceX API
- Web scrapping Wikipedia page

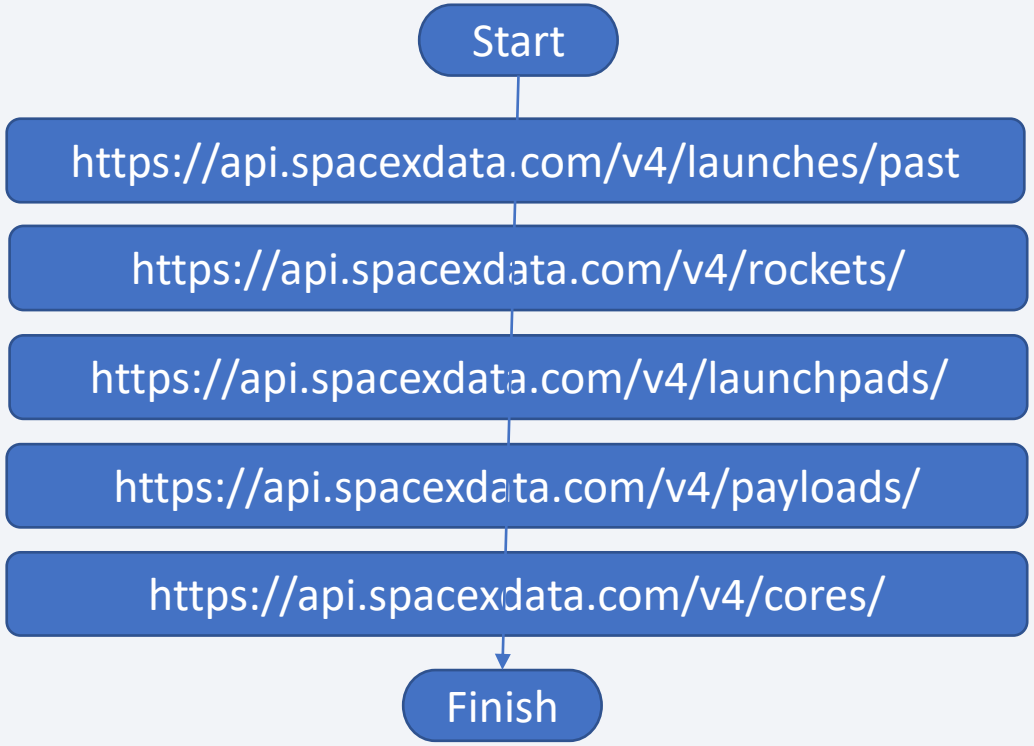
Data Collection – SpaceX API

The sequence of SpaceX API calls:

1. Get rocket launch data
2. Get Booster Version
3. Get Launch Site
4. Get Payload Data
5. Get Core Data

The GitHub URL of the completed SpaceX API calls notebook:
https://github.com/DmLychev/Applied-Data-Science-Capstone/blob/74bf2dee53eef7d5300ef62551b72de868166977/01.%20Data%20Collection%20API.ipynb

The flowchart of SpaceX API calls:



Data Collection - Scraping

The sequence of actions to perform web scraping:

1. Request the Falcon9 Launch HTML page
2. Create a BeautifulSoup object from the HTML response
3. Extract all column/variable names from the HTML table header. Parsing the launch HTML tables
4. Export to CSV.

The GitHub URL of the completed web scraping notebook: https://github.com/DmLychev/Applied-Data-Science-Capstone/blob/d347a20ffbdb0c501d35519e1616868df3de8d19/02.%20Data%20Collection%20with%20Web%20Scraping.ipynb

The flowchart of web scraping:

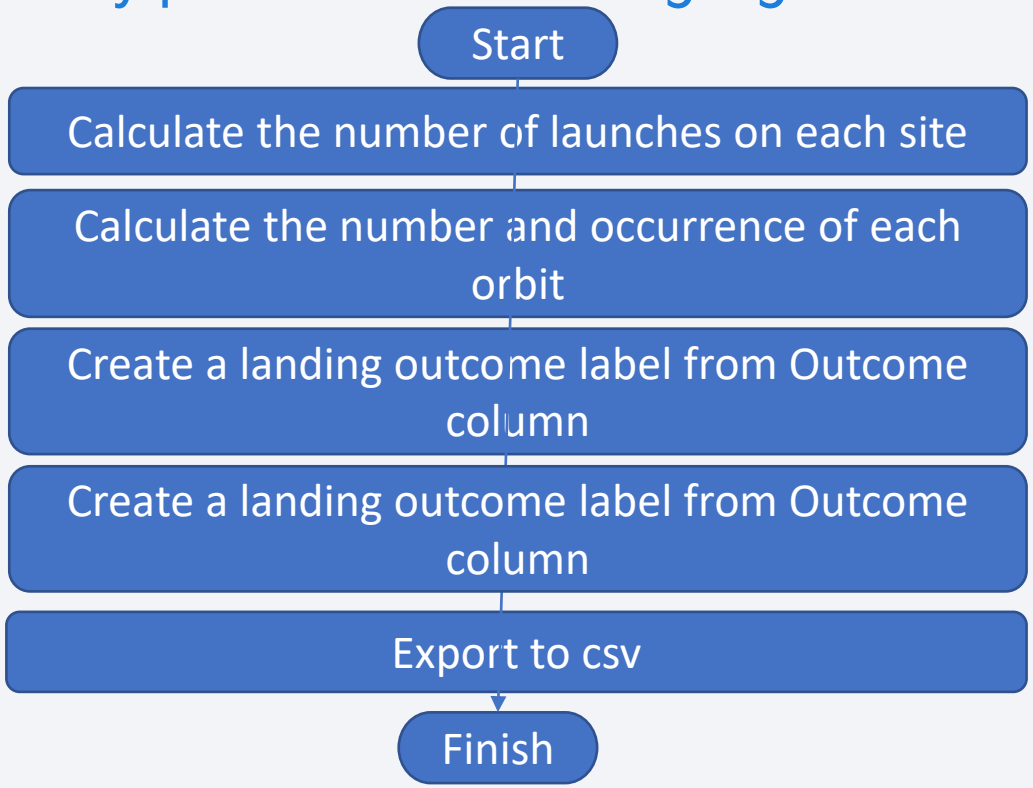
Data Wrangling

The sequence of actions to perform web scraping:

1. Calculate the number of launches on each site
2. Calculate the number and occurrence of each orbit
3. Calculate the number and occurence of mission outcome per orbit type
4. Create a landing outcome label from Outcome column
5. Export to CSV

The GitHub URL of completed data wrangling: https://github.com/DmLychev/Applied-Data-Science-Capstone/blob/1fef970c3bb5994bb6deb79164ab6521a5b964f7/03.%20EDA.ipynb

The key phases of data wrangling:



EDA with Data Visualization

- Scatter plots were used to represent the relationship between two variables. Different sets of features were compared such as Flight Number vs. Launch Site, Payload vs. Launch Site, Flight Number vs. Orbit Type and Payload vs. Orbit Type.
- Bar charts were used to compare values between, multiple groups at a glance. The Success Rate was compared for different Orbit Types using a bar chart.
- Line charts were used for showing data trends over time. The Success Rate over a certain number of Years was shown using a line chart.

The GitHub URL of completed EDA with data visualization notebook:
https://github.com/DmLychev/Applied-Data-Science-Capstone/blob/84c43223842be8bb8fa0f284fd77f7c6726e67b6/05.%20EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA’
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display the average payload mass carried by booster
- List the date when the first succesful landing outcome in ground pad was acheived.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
- 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.
- Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

The GitHub URL of completed EDA with SQL notebook:
https://github.com/DmLychev/Applied-Data-Science-Capstone/blob/84c43223842be8bb8fa0f284fd77f7c6726e67b6/04.%20EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

Following objects were created and added to a Folium Map:

- Markers were used to show all launch sites on a map with Success/Failure mark/
- Line objects were used to calculate the distances between a launch site and coastline, cities, railways and highways,
- The GitHub URL of completed interactive map with Folium map: https://github.com/DmLychev/Applied-Data-Science-Capstone/blob/09915cc32679c236d39db6fdaf66a541119874cf/06.%20Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

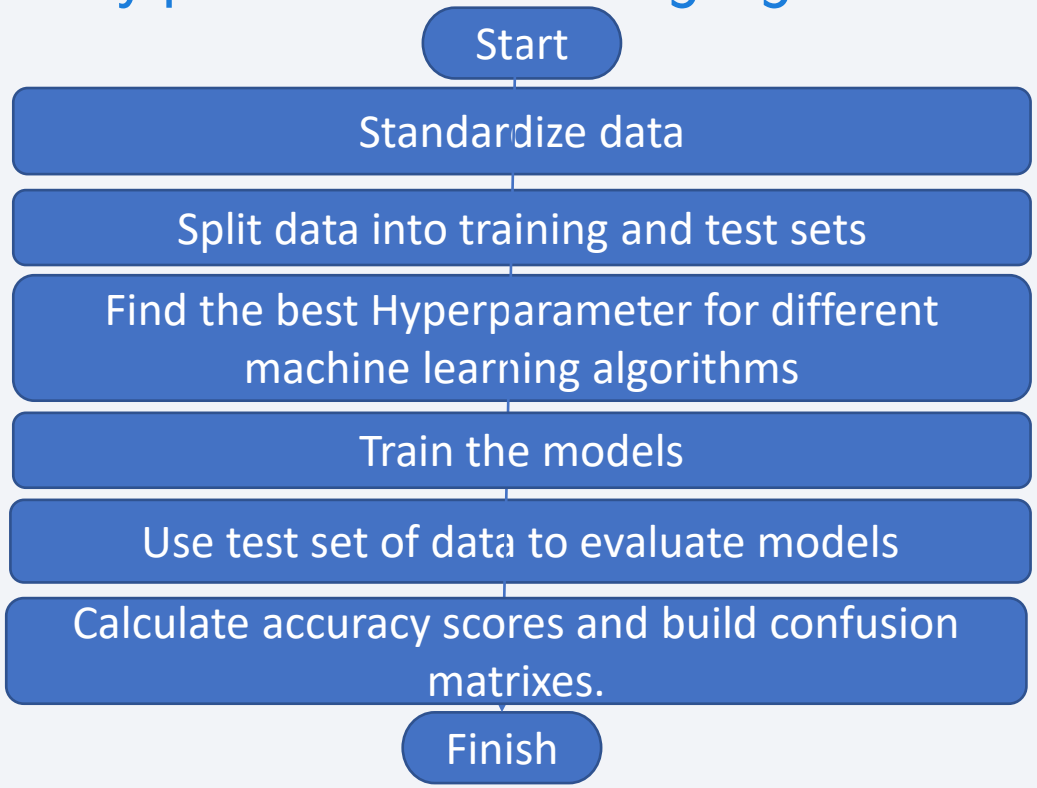
The dashboard application contains:

- A pie chart that shows the successful launch by each site. This chart visualize the distribution of landing outcomes across all launch sites or the success rate of launches on individual sites.
- A scatter plot that shows the relationship between landing outcome and payload mass.
- The GitHub URL of completed Plotly Dash lab: https://github.com/DmLychev/Applied-Data-Science-Capstone/blob/8c97fe242326b49b52f582f55f84c0a48f00cf6c/07.%20spacex_dash_app.py

Predictive Analysis (Classification)

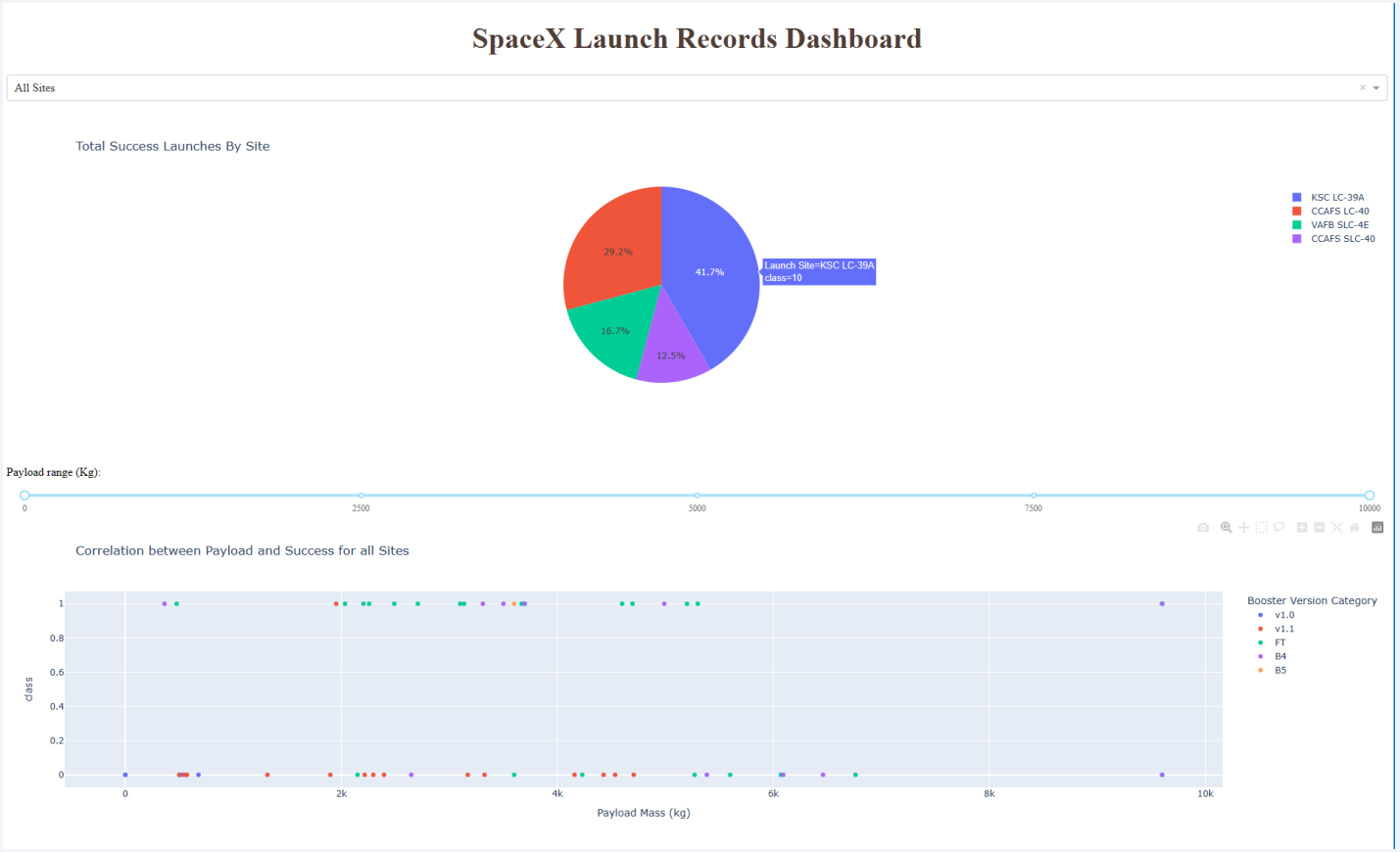
The GitHub URL of completed predictive analysis lab:
https://github.com/DmLychev/Applied-Data-Science-Capstone/blob/0d0de3b533f2d9aba187a2ed4573b49d25f284e4/08.%20Machine%20Learning%20Prediction.ipynb

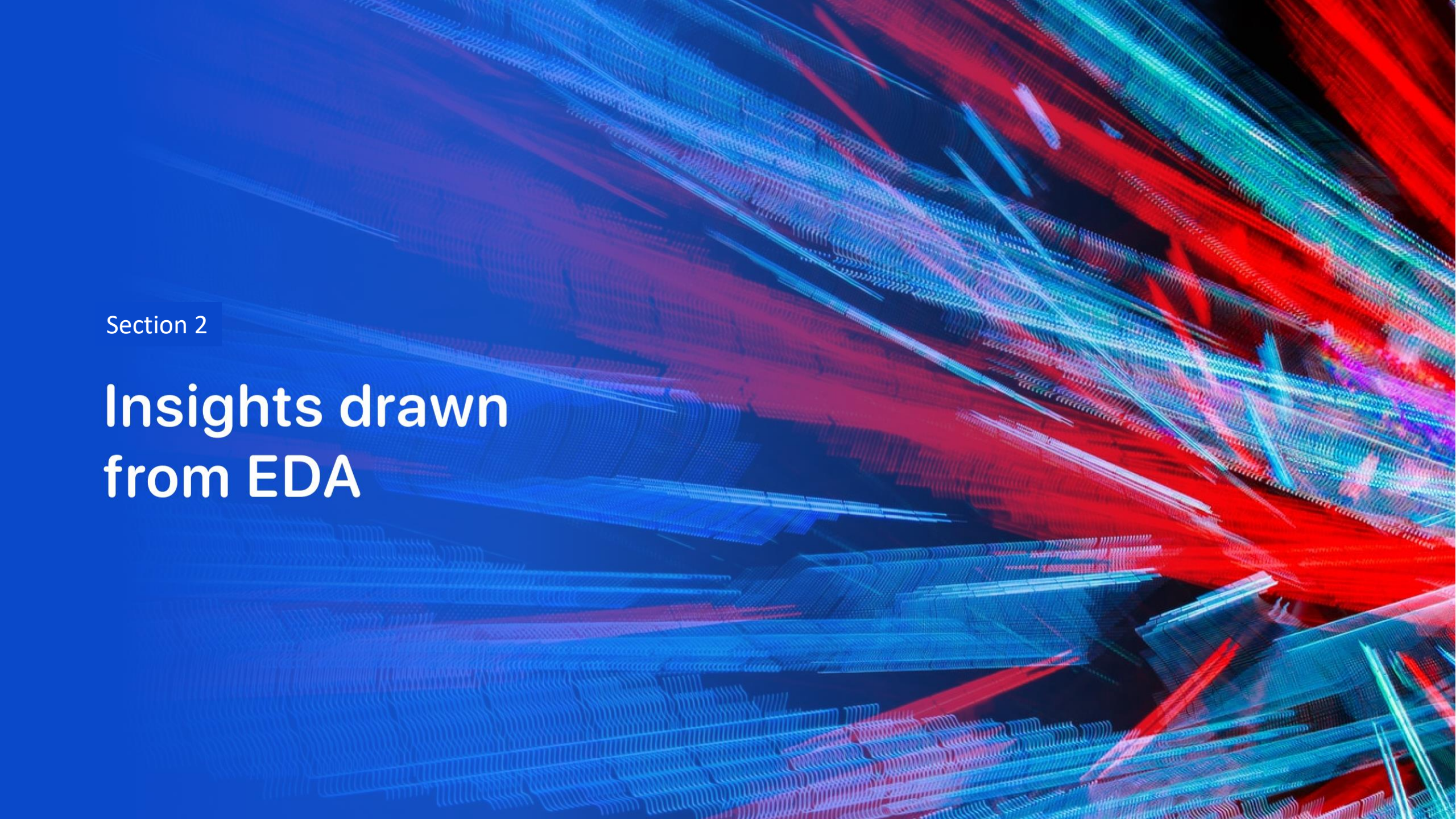
The key phases of data wrangling:



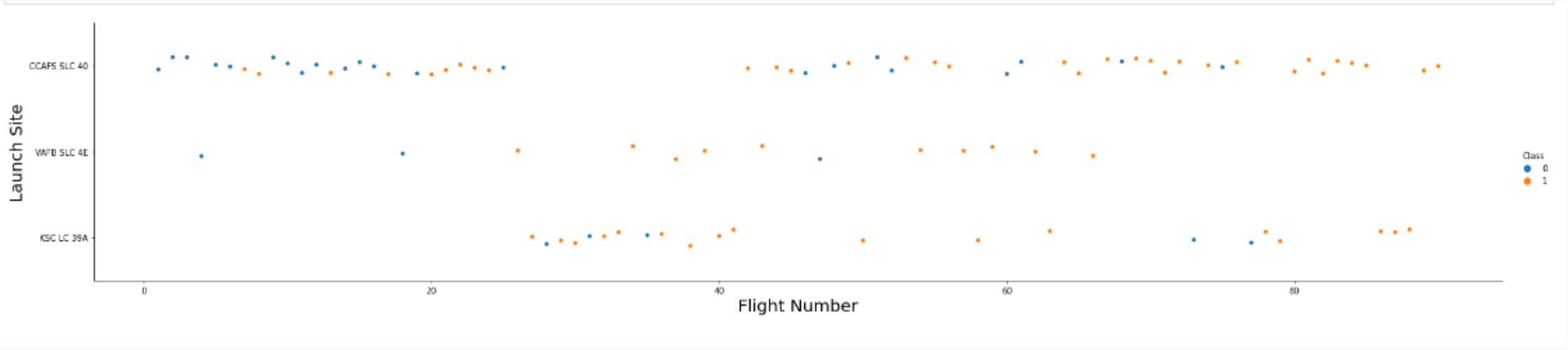
Results

- The results of the EDA revealed that the success rate of the Falcon 99 landing was 66%
- The Predictive Analysis results showed that the decision Tree algorithm was the best classification method with an accuracy of 94%




Section 2
Insights drawn from EDA

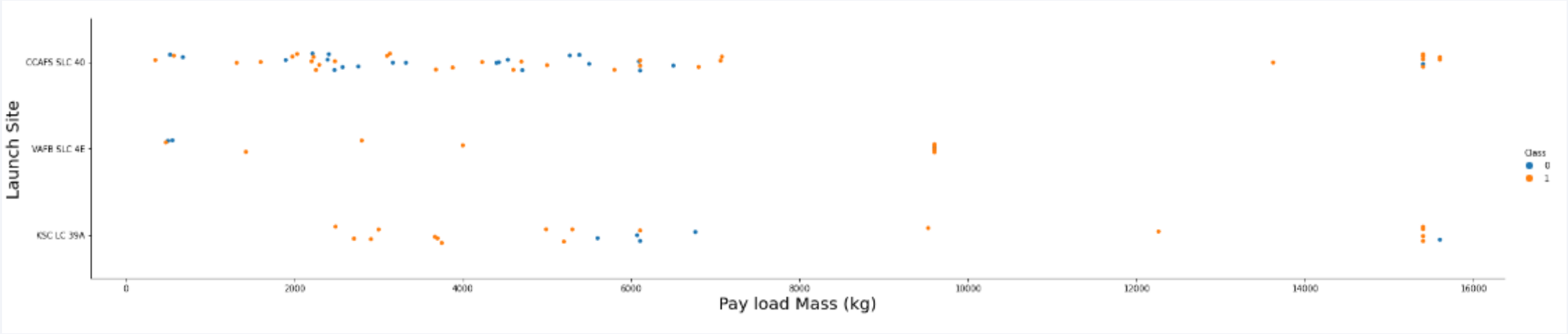
Flight Number vs. Launch Site



The red dots on the figure above represent the successful launches while the blue ones represent unsuccessful launches.

This figure shows that the success rate increased as the number of flights increased especially after th 40th launch.

Payload vs. Launch Site

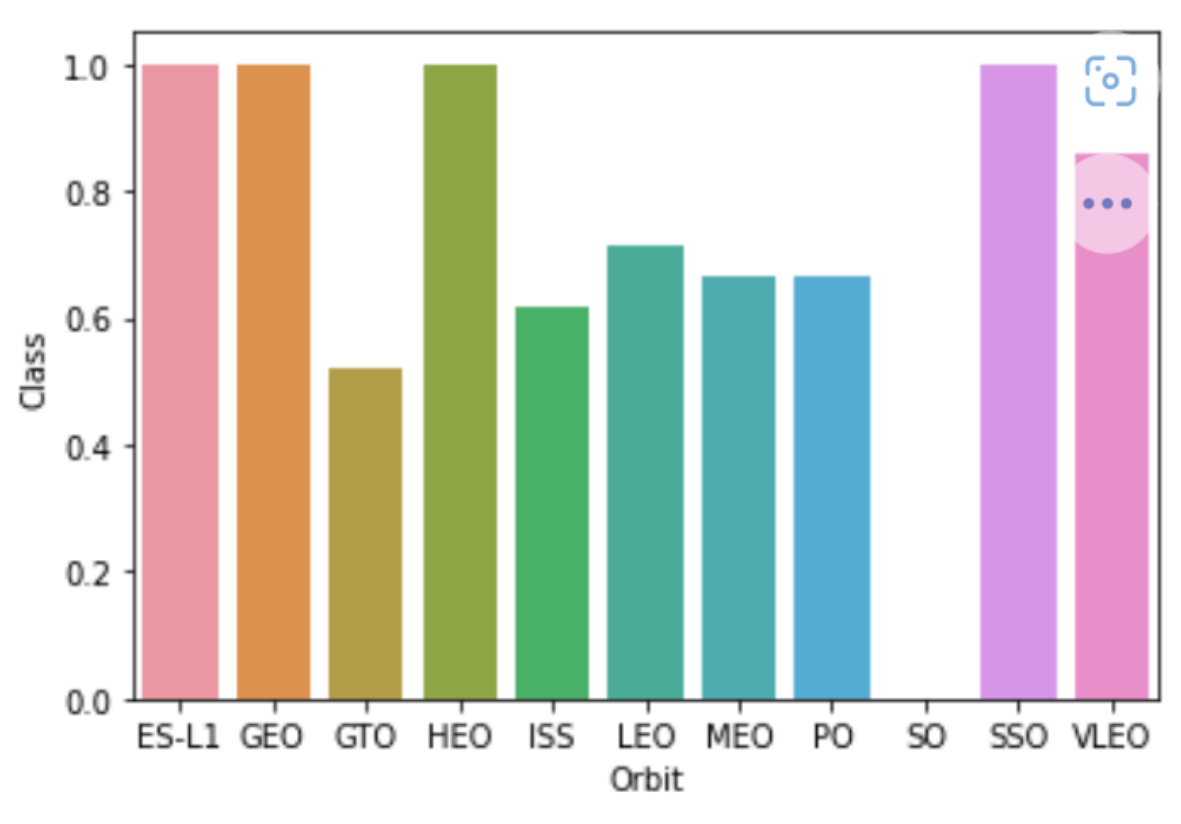


The red dots on the figure above represent the successful launches while the red ones represent unsuccessful launches.

This figure shows that there are no rocket launches with heavy payload mass for the VAFB-SLC. The correlation between Payload and Launch Site is weak therefore decisions can not be made using this metric.

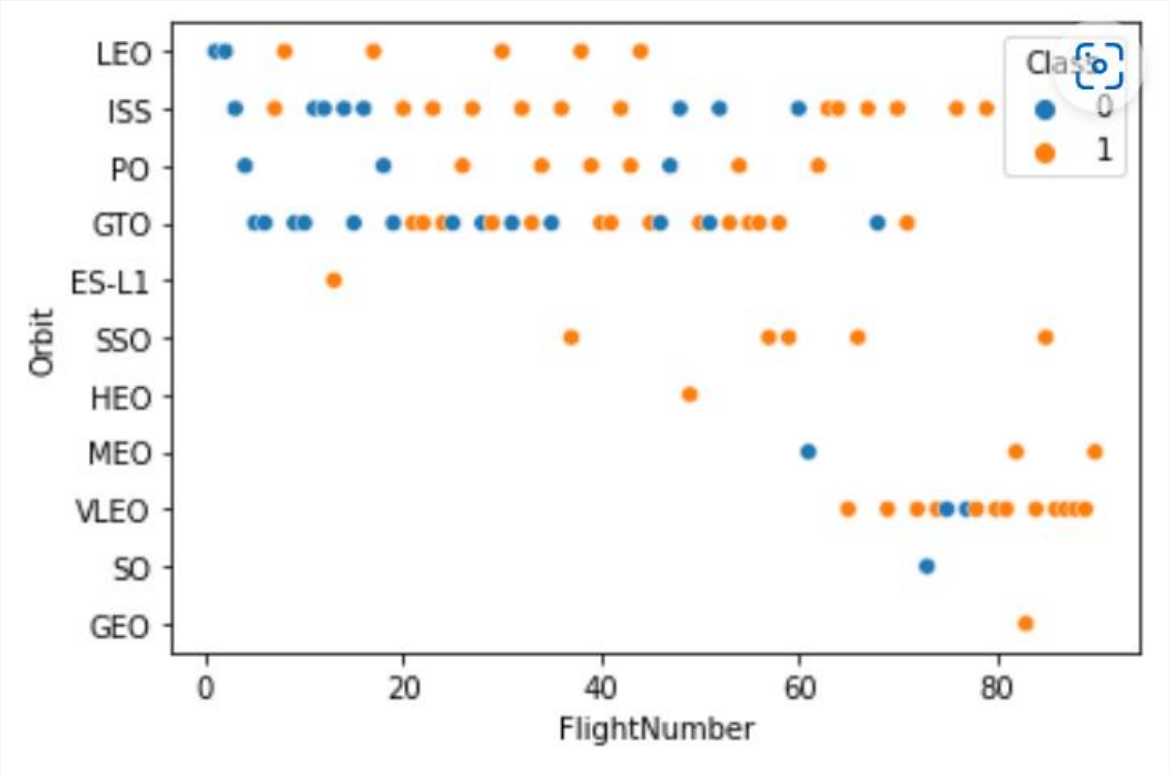
Success Rate vs. Orbit Type

- SO orbit doesn't have any successful launches.
- SSO, HEO, GEO, ES-L1 have 100% success rate.



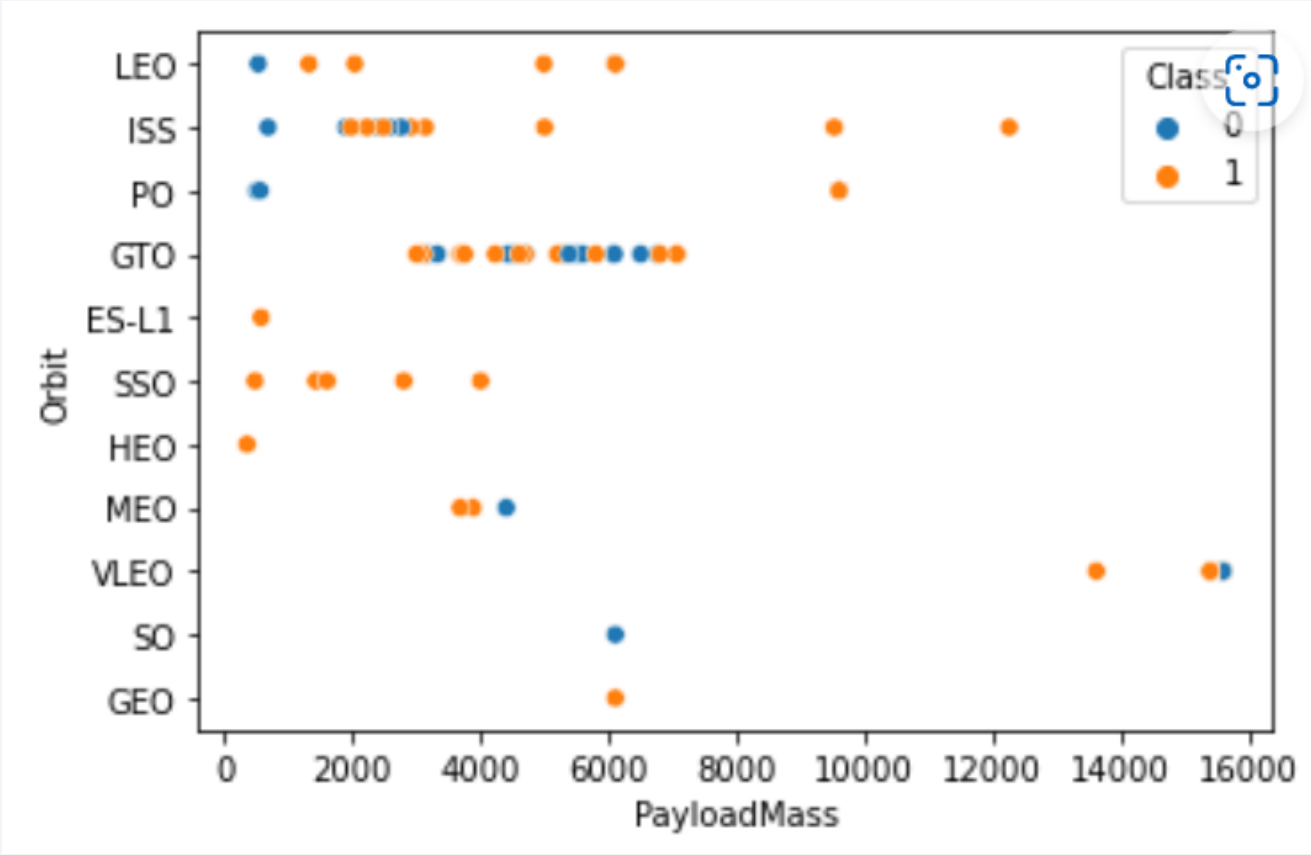
Flight Number vs. Orbit Type

- Flights number greater than 40 have a higher success rate than flights numbers between 0 and 40.
- SSO, HEO, ES-L1, GEO orbits have a 100% success rate
- There is a positive correlation between LEO orbit and the number of flights.
- There seems to be no correlation between GTO orbit and the number of flights.



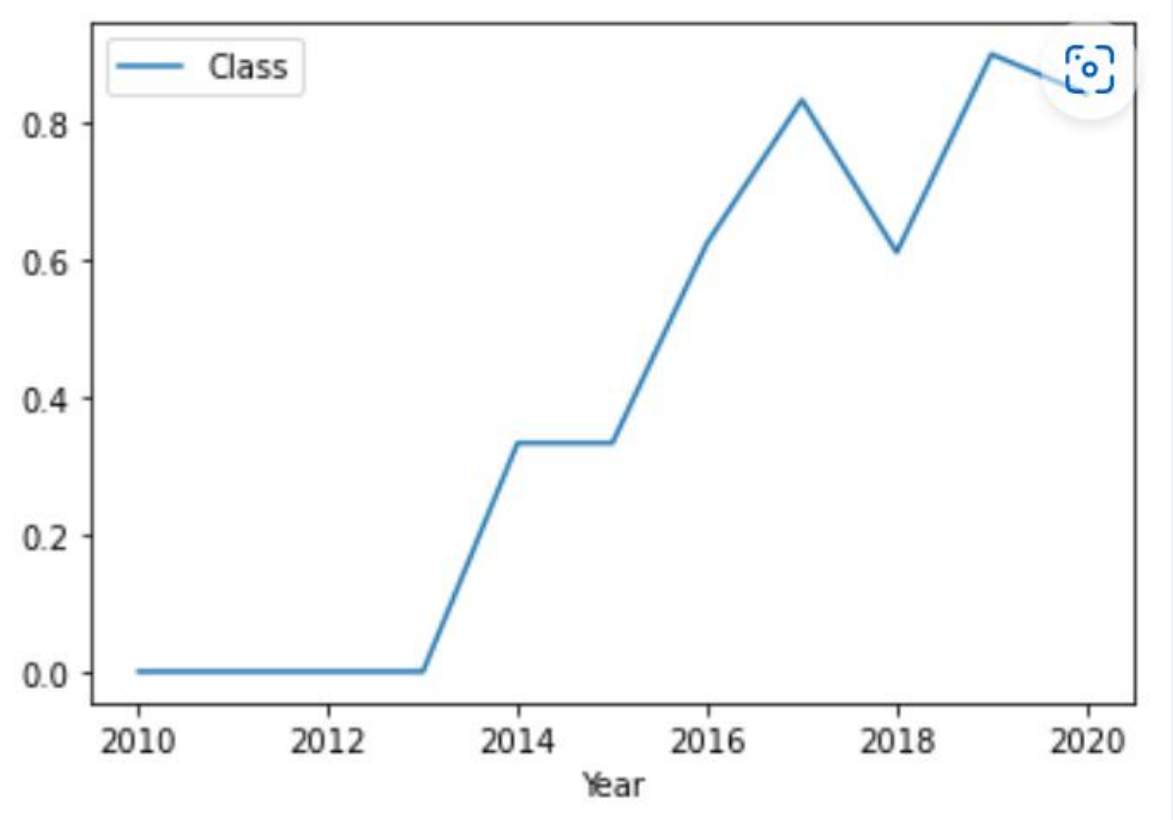
Payload vs. Orbit Type

- For PO, SSO, LEO and ISS orbits the heavier the payload mass, the bigger the success rate.
- There seems to be no correlation between GTO orbit and the payload mass.



Launch Success Yearly Trend

- There is a strong trend to increase in landing success rate through years.



All Launch Site Names

The DISTINCT clause was used to return unique values from the Launch_Site column. There are 4 unique launch sites. They are:

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

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

Launch Site Names Begin with 'CCA'

The LIMIT and LIKE clauses were used to display first five rows where Launch_Site name started with ‘CCA’.

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

The SUM() function was used to calculate the total payload mass carried by boosters launched by NASA (CRS).

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

The AVG() function was used to calculate average payload mass carried by booster version F9 v1.1

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

The MIN() functions and WHERE clause were used to calculate the date when the first succesful landing outcome in ground pad was acheived

min(Date)
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

The BETWEEN and WHERE claused were used to display the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

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

Total Number of Successful and Failure Mission Outcomes

The COUNT() function and WHERE clause were used to calculate the total number of successful and failure mission outcomes

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

Boosters Carried Maximum Payload

The MAX() function in a subquery was used to dicplay the names of the booster_versions which have carried the maximum payload mass.

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

The SELECT statement, WHERE and GROUPBY clauses were used to display the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

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

The COUNT() function, the WHERE, BETWEEN, GROUPBY and ORDERBY with DESC clauses were used to rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

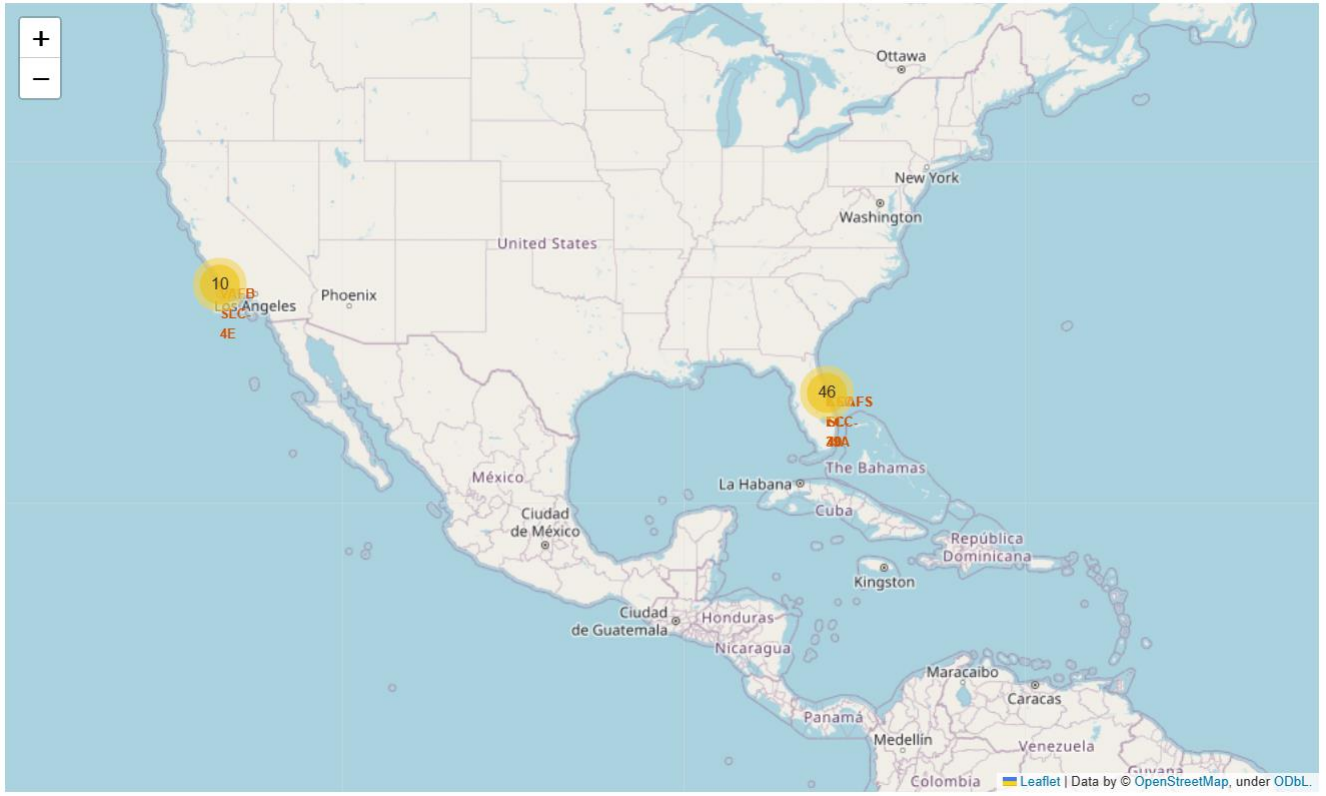
landing__outcome	count_launches
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

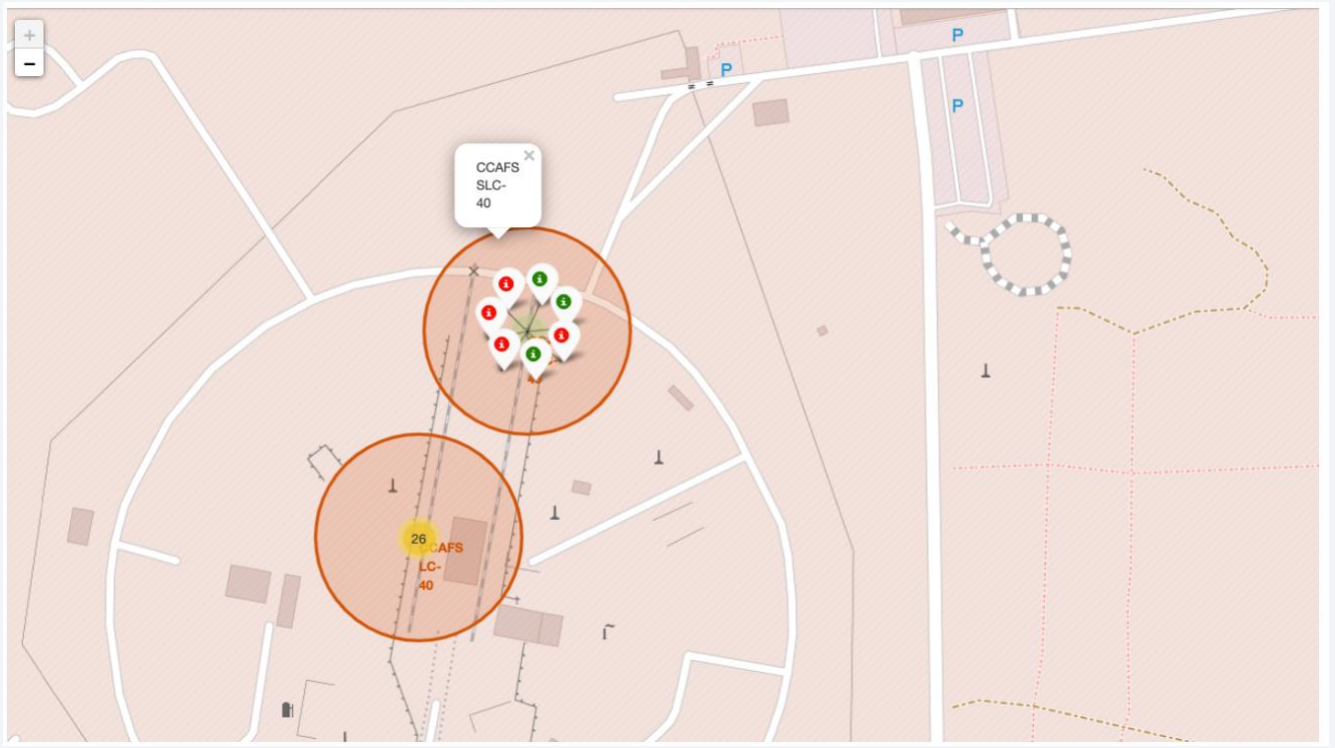
SpaceX Launch Sites Locations

- The yellow markers indicate the locations of all the SpaceX launch sites. All sites are located in the southern part of the US near the coast.



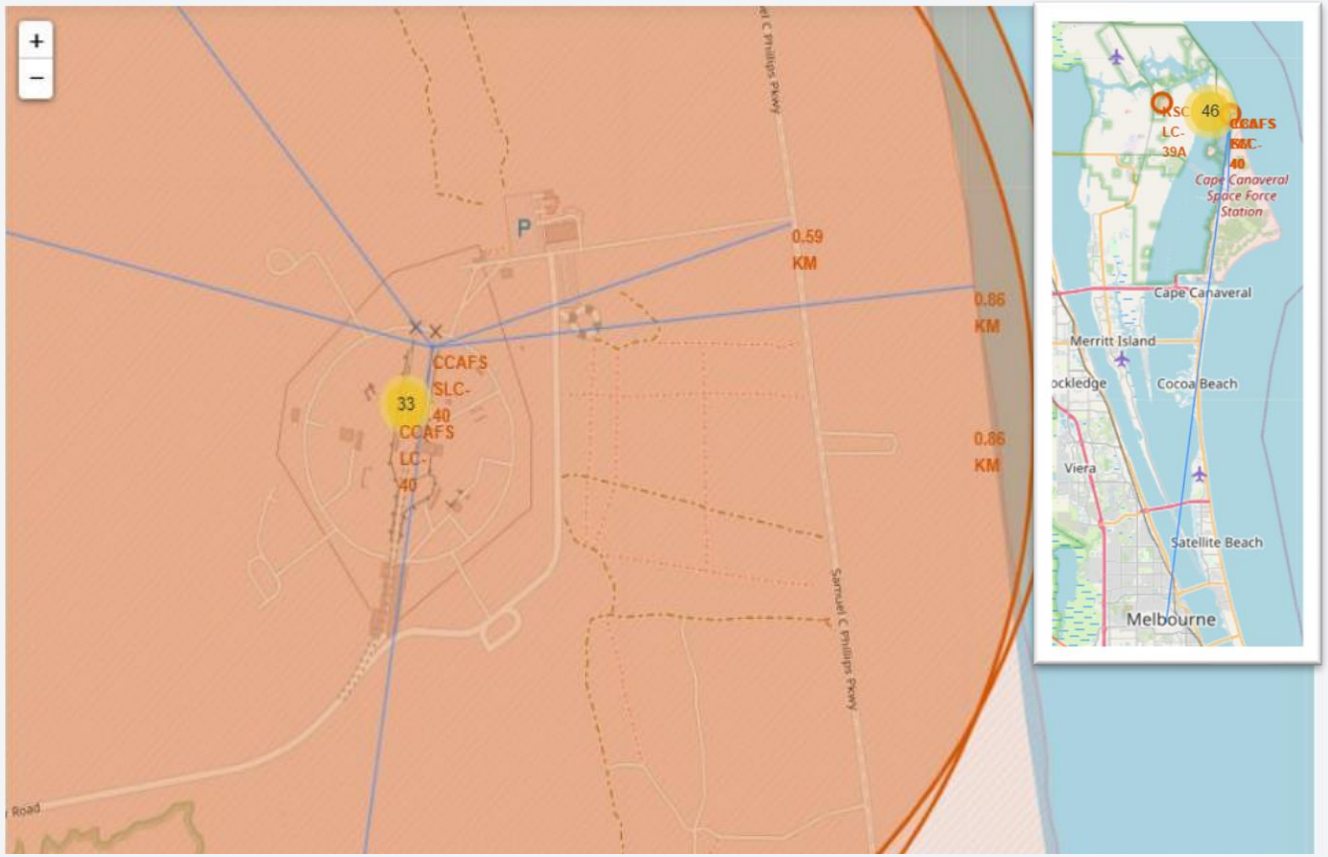
Success/Failure

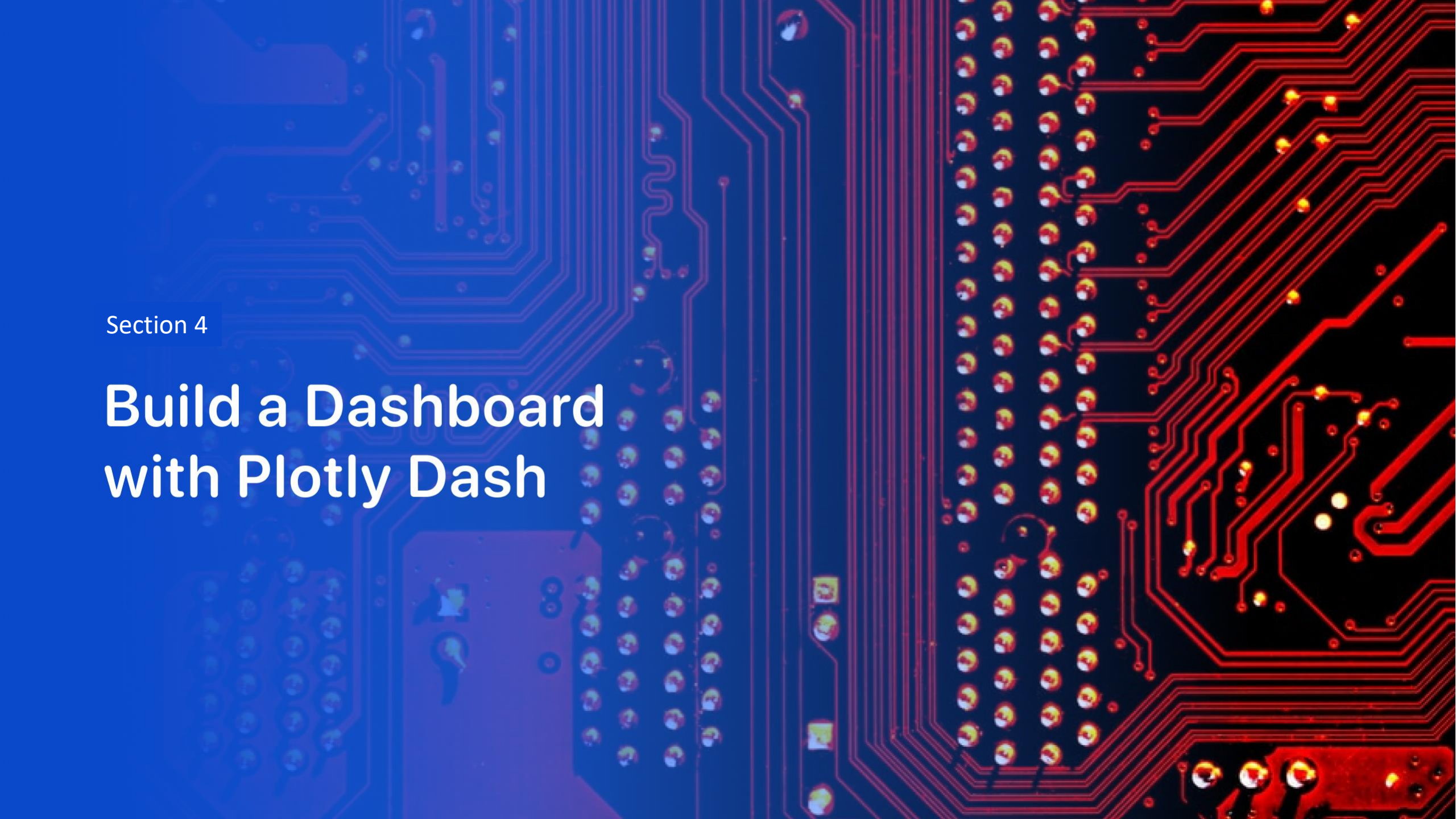
All of the launch outcomes group in clusters where red markers represent failed launches and green markers represent successful launches.



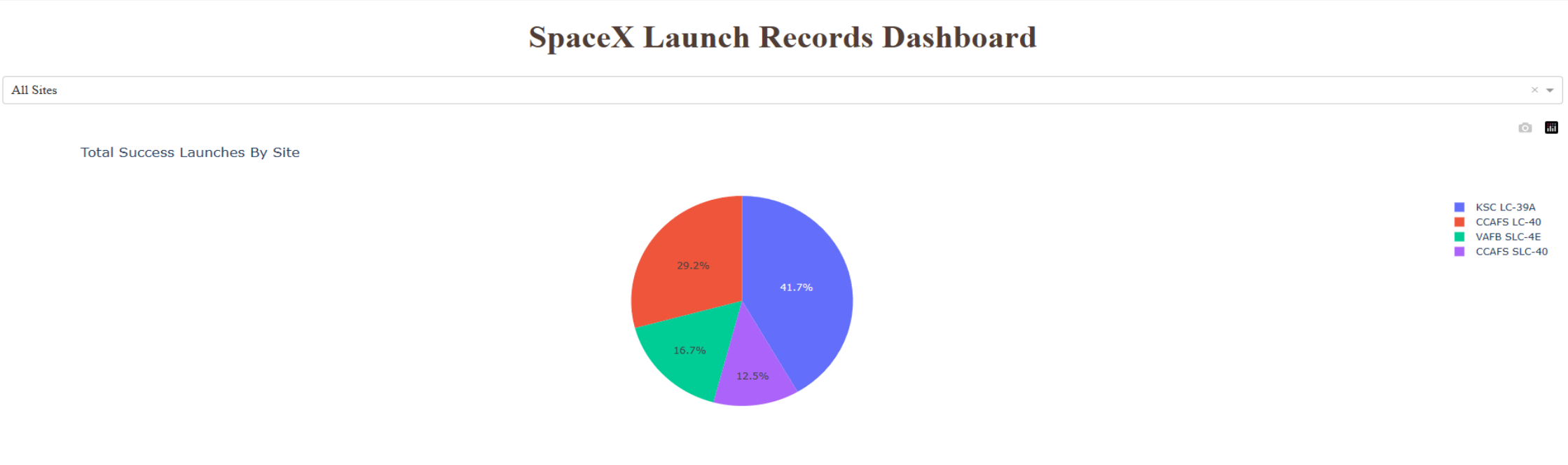
Launch Sites proximities

The map on the right displays distances to the nearest coast, city, highway and railroad. All of them pretty close to the launch site.



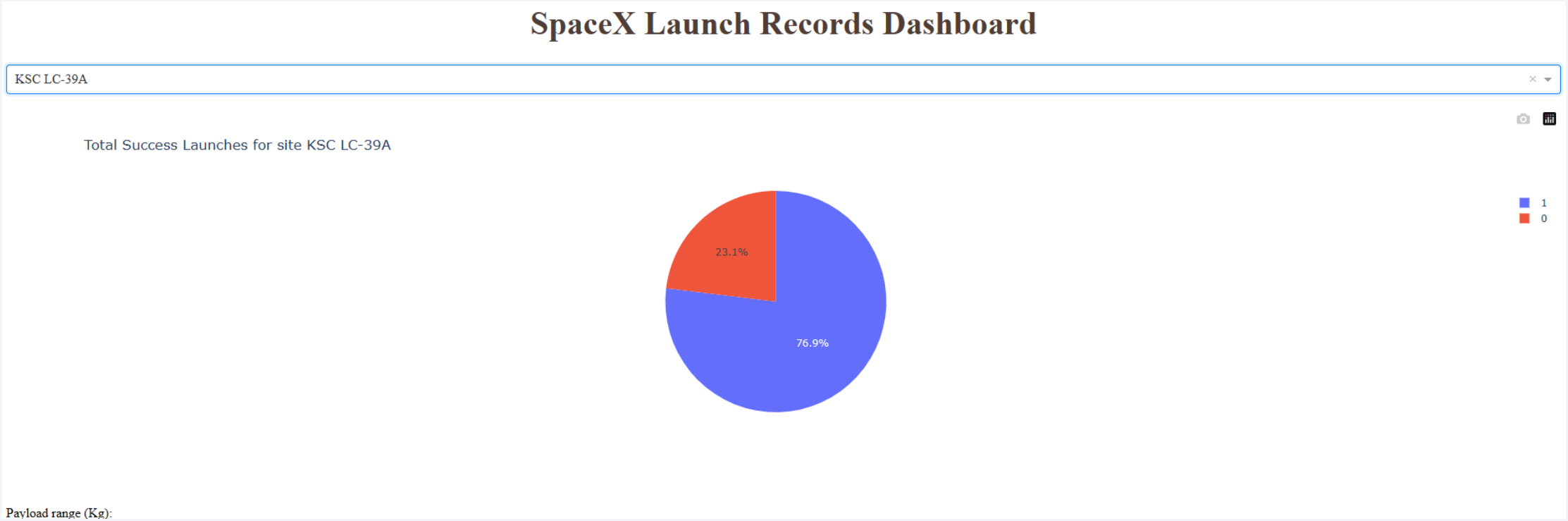

Section 4
Build a Dashboard with Plotly Dash

Total Successful Launches By Site



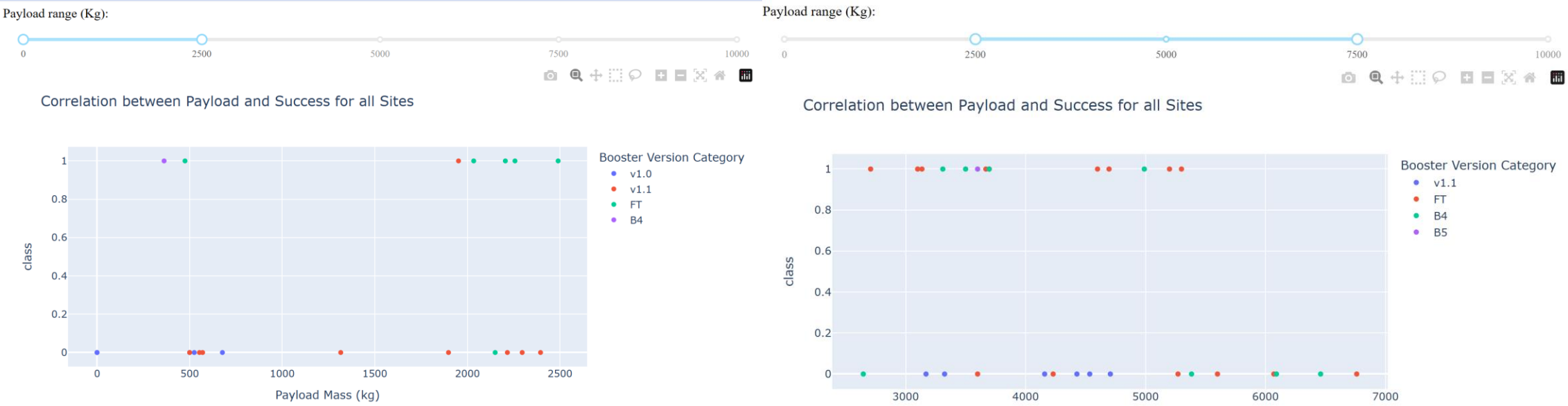
- The KSC LC-39A Launch Site has the most successful launches with 41.7%

Launch Site With Highest Success Ratio



- The KSLC-39A has the highest success rate with 76.9%

Payloads vs Launch Outcome



- The Booster v1.1 has the largest success rate as for payload mass under 2500 kg, as for payload mass in a range from 2500 to 7000 kg.
- Overall the launch success rate for payloads under 2500 kg lower than for 2500 – 7000 kg.

Section 5

Predictive Analysis (Classification)

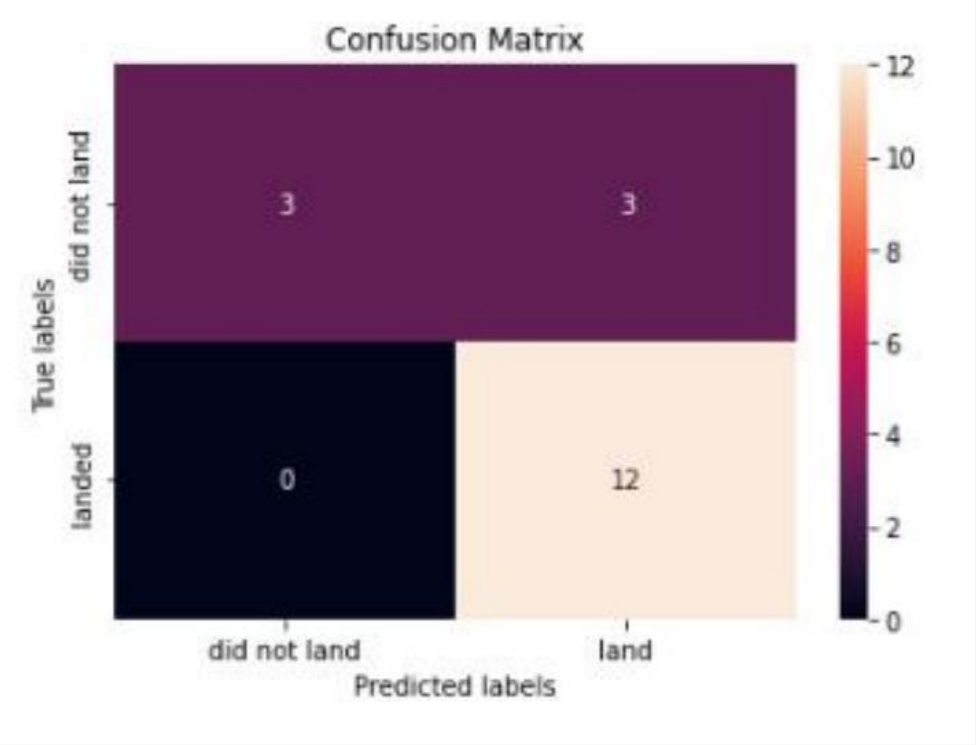
Classification Accuracy

The Decision Tree classifier has the best accuracy at 94%.

	method	accuracy
0	Logistic regression	0.833333
1	Support vector machine	0.833333
2	Decision tree classifier	0.944444
3	K nearest neighbors	0.833333

Confusion Matrix

The Decision Tree model predicts 15 landings correctly and 3 landings incorrectly (false positive).



Conclusions

- The best predictive model to use for this dataset is the Decision Tree Classifier as it had the highest accuracy with 94%.
- There are certain orbits like SSO, HEO, GEO, and ES-L1 where launches were the most successful.
- The analysis showed that there is a positive correlation between number of flights and success rate as the success rate has improved over the years.

Appendix

GitHub Repository: https://github.com/DmLychev/Applied-Data-Science-Capstone

Thank you!