IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

USHASI DAS

07-10-2024

Outline

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

Executive Summary

Summary of methodologies

- Data collection
- Data wrangling
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly Dash
- Predictive analysis (Classification)

Summary of all results

- Exploratory Data Analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Introduction

Project background and context SpaceX is the most successful company of the commercial space age, making space travel affordable. The company 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 SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage.

Questions to be answered

- How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?
- Does the rate of successful landings increase over the years?
- What is the best algorithm that can be used for binary classification in this case?

Section 1

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from the SpaceX API and web-scraping a Wikipedia and data was filtered to include only Falcon9 launches.
- Data wrangling
  - Missing values were encountered and landing outcomes were summarized to successful and failure landings.
- Exploratory data analysis (EDA) using visualization and SQL
- Interactive visual analytics using Folium and Plotly Dash
- Predictive analysis using classification models
  - Different models were trained and the best one was selected based on their performances on testing data.

Data Collection

SpaceX REST API :

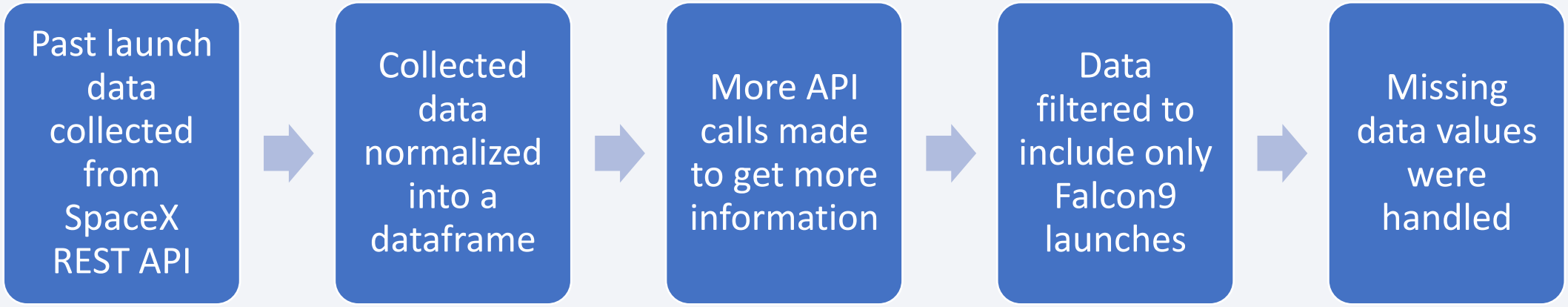
- Past launch data was collected from the API and the result was normalized into a dataframe.
- More API calls were made to get information about rocket booster name and version, payload mass, targeted orbit, details of launch site and core details.
- The data was filtered to include only Falcon9 launches.
- Some basic data wrangling was done to deal with missing values.

Web-scraping Wikipedia page:

- From the Wikipedia page “List of Falcon 9 and Falcon Heavy launches”, launch data was scraped.

Data Collection – SpaceX API

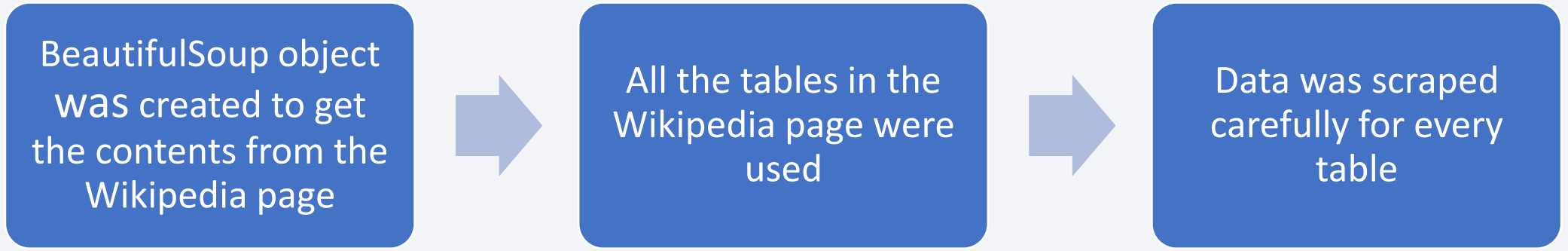
Work process flowchart



GitHub URL : https://github.com/DasUshasi/IBM-data-science-course-Assignments/blob/main/capstone%20project/1-data%20collection%20API.ipynb

Data Collection - Scraping

Work process flowchart

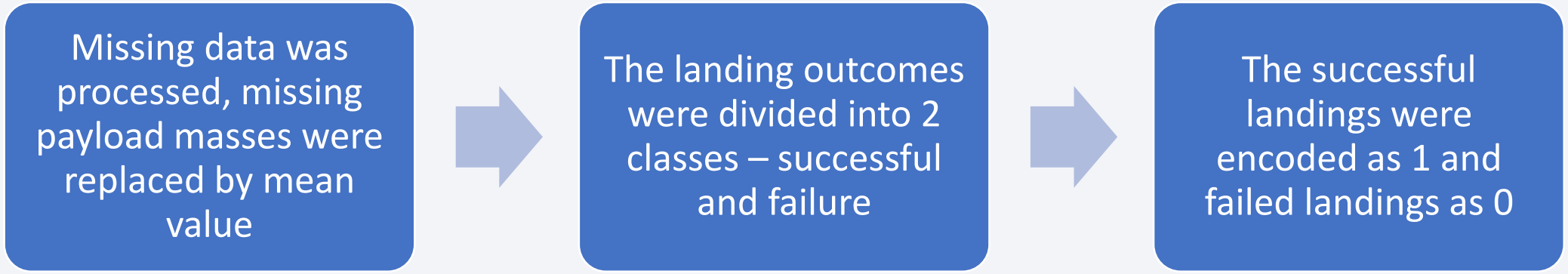


GitHub URL : https://github.com/DasUshasi/IBM-data-science-course-Assignments/blob/main/capstone%20project/2-data%20collection%20webscraping.ipynb

Data Wrangling

2 major jobs were done :

- Missing data were processed
- Landing outcomes were classified into successful and failed landings



GitHub URL : https://github.com/DasUshasi/IBM-data-science-course-Assignments/blob/main/capstone%20project/3-data%20wrangling.ipynb

EDA with Data Visualization

- Scatter plots were used to see how the Payload mass and Launch site would affect the launch outcome and the relationship among them.
- Bar chart was used to visualize the success rate of landing of rockets with respect to their orbits.
- Scatter plots were used to visualize how orbits affected landing outcome and if there is a relationship between orbit type and payload mass.
- Line plot was used to visualize the success trend over the years.

GitHub URL : https://github.com/DasUshasi/IBM-data-science-course-Assignments/blob/main/capstone%20project/5-eda%20python.ipynb

EDA with SQL

- Queries were made to get the unique launch sites and records of few launches made from specific launch sites
- Queries were made to get information about payload mass based on booster versions
- Queries were made to get few information of few specific successful landings and total number of successful and failure mission outcomes
- Queries were made to get information about successful and failed landings in a specified time interval

GitHub URL : https://github.com/DasUshasi/IBM-data-science-course-Assignments/blob/main/capstone%20project/4-eda%20sql.ipynb

Build an Interactive Map with Folium

- Circles and Markers were used to highlight the NASA Johnson Space Center and all the launch sites from the dataset in a Folium Map
- A MarkerCluster was created and for each record, a Marker was added at the record's launch site to the Marker Cluster with colour assigned (green = successful, red = failure) to indicate whether the landing was successful or not
- For launch site CCAFS SLC-40, the nearest coastline, highway, railway and city and their distance from the site were marked using Marker and PolyLine

GitHub URL : https://github.com/DasUshasi/IBM-data-science-course-Assignments/blob/main/capstone%20project/6-visualization%20folium.ipynb

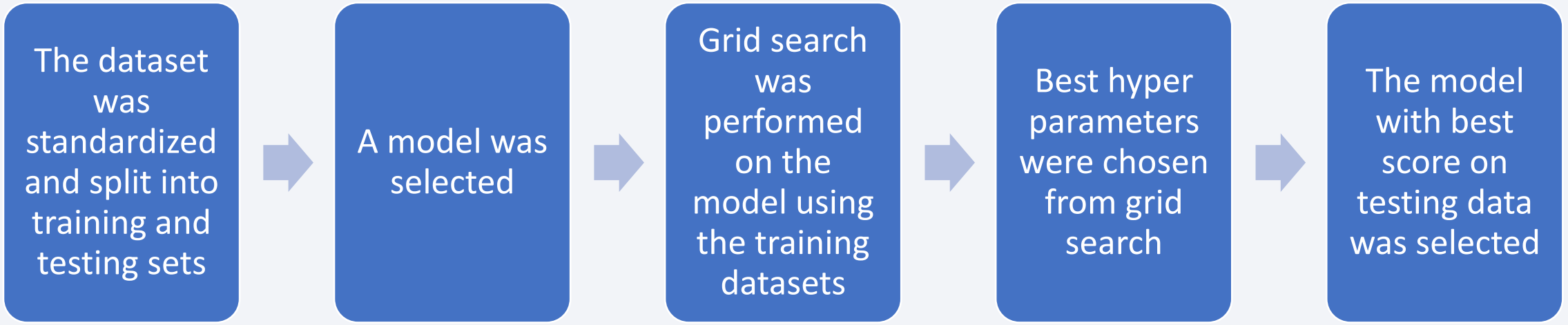
Build a Dashboard with Plotly Dash

- A dropdown was created so that user can select a specific launch site (or all)
- If a specific launch site is selected, the number of successful and failure outcomes from that site were plotted in a pie chart, and if “all sites” is selected then a pie chart is created to show the percentage of successful outcomes from all launch sites
- A slider was created so that the user can select a specific payload mass range
- A scatter plot is created to show the relation between payload mass and successful outcomes for the selected launch site

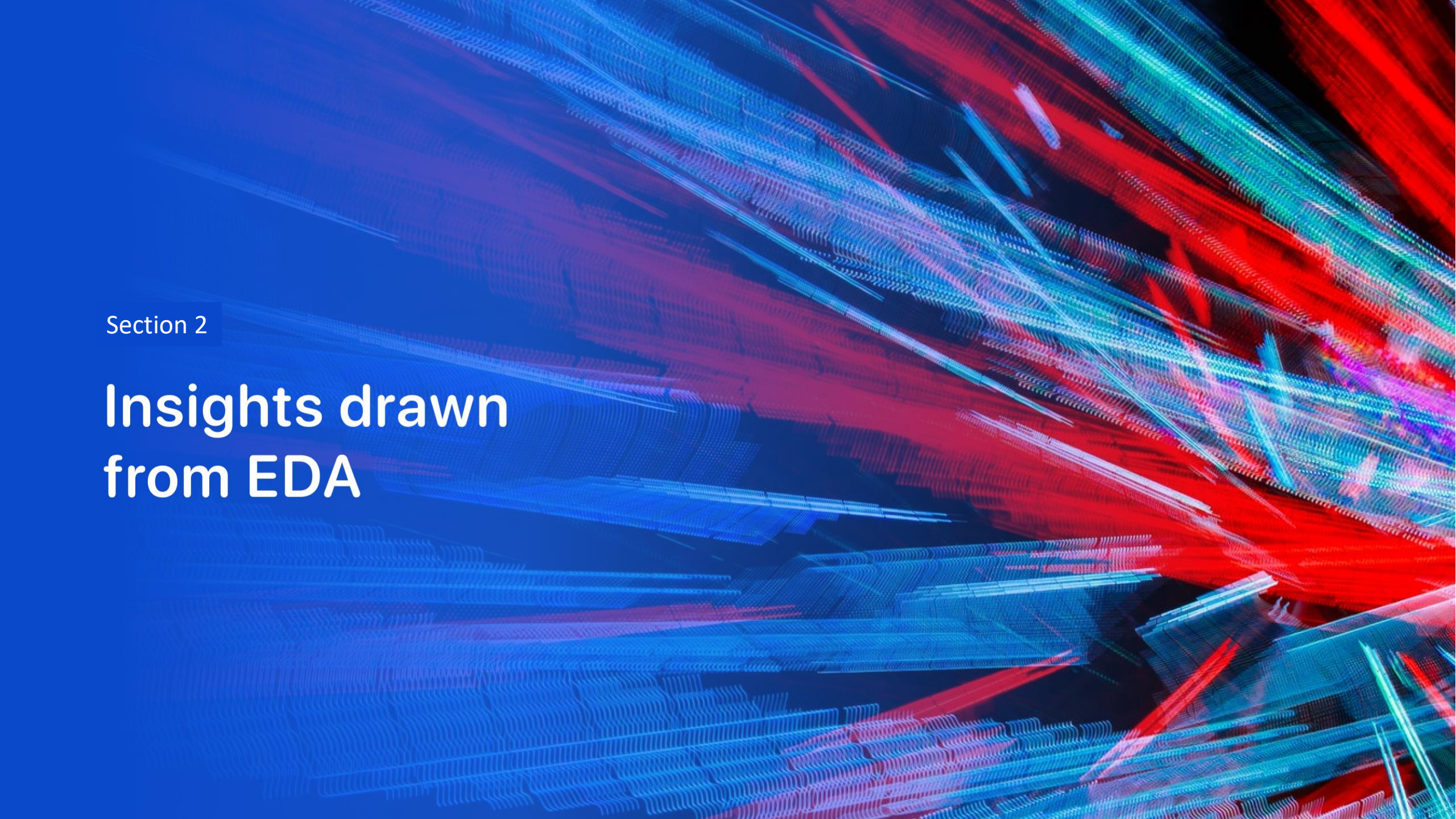
GitHub URL : https://github.com/DasUshasi/IBM-data-science-course-Assignments/blob/main/capstone%20project/7-dashboard.py

Predictive Analysis (Classification)

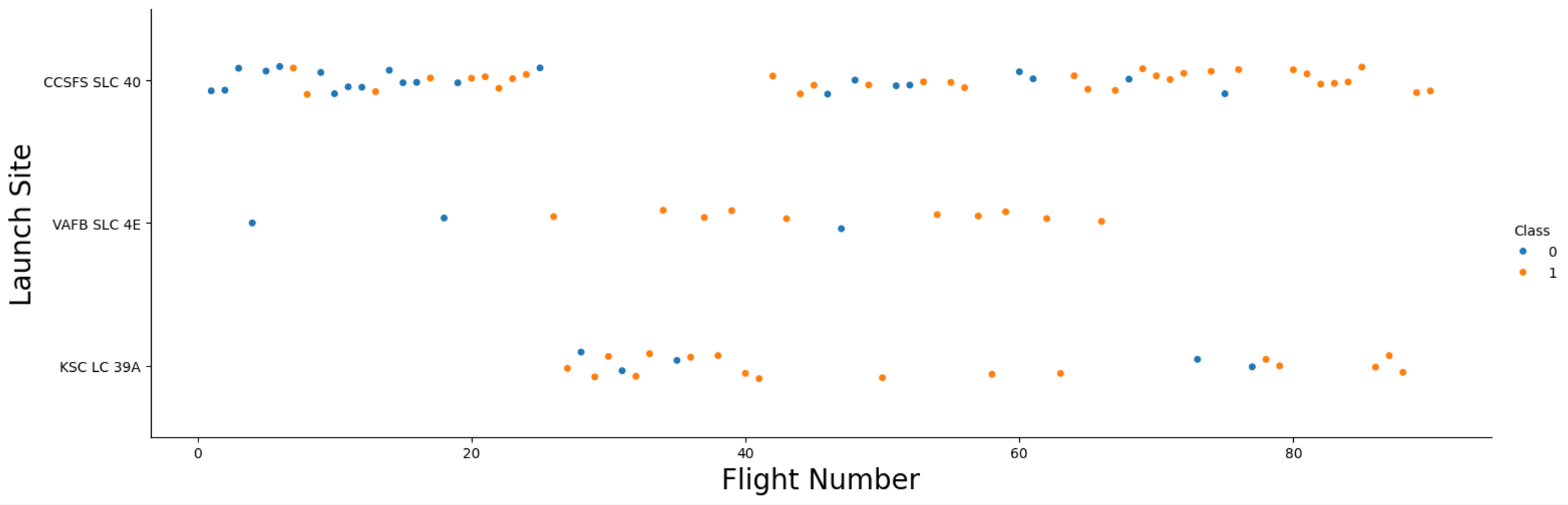
Logistic regression, support vector machine, decision tree classifier and k nearest neighbors were trained and tested on the dataset.



GitHub URL : https://github.com/DasUshasi/IBM-data-science-course-Assignments/blob/main/capstone%20project/8-machine%20learning.ipynb

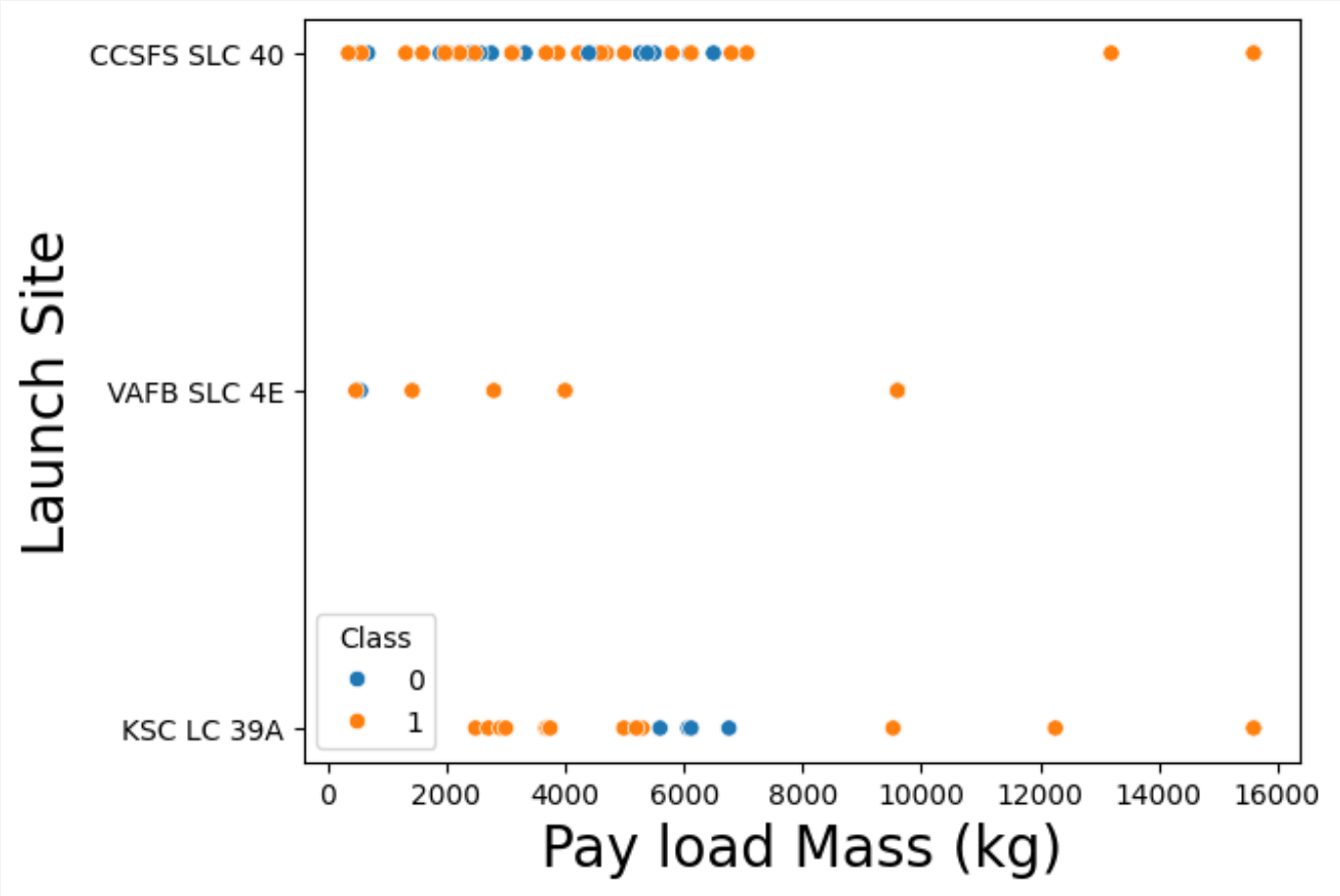

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



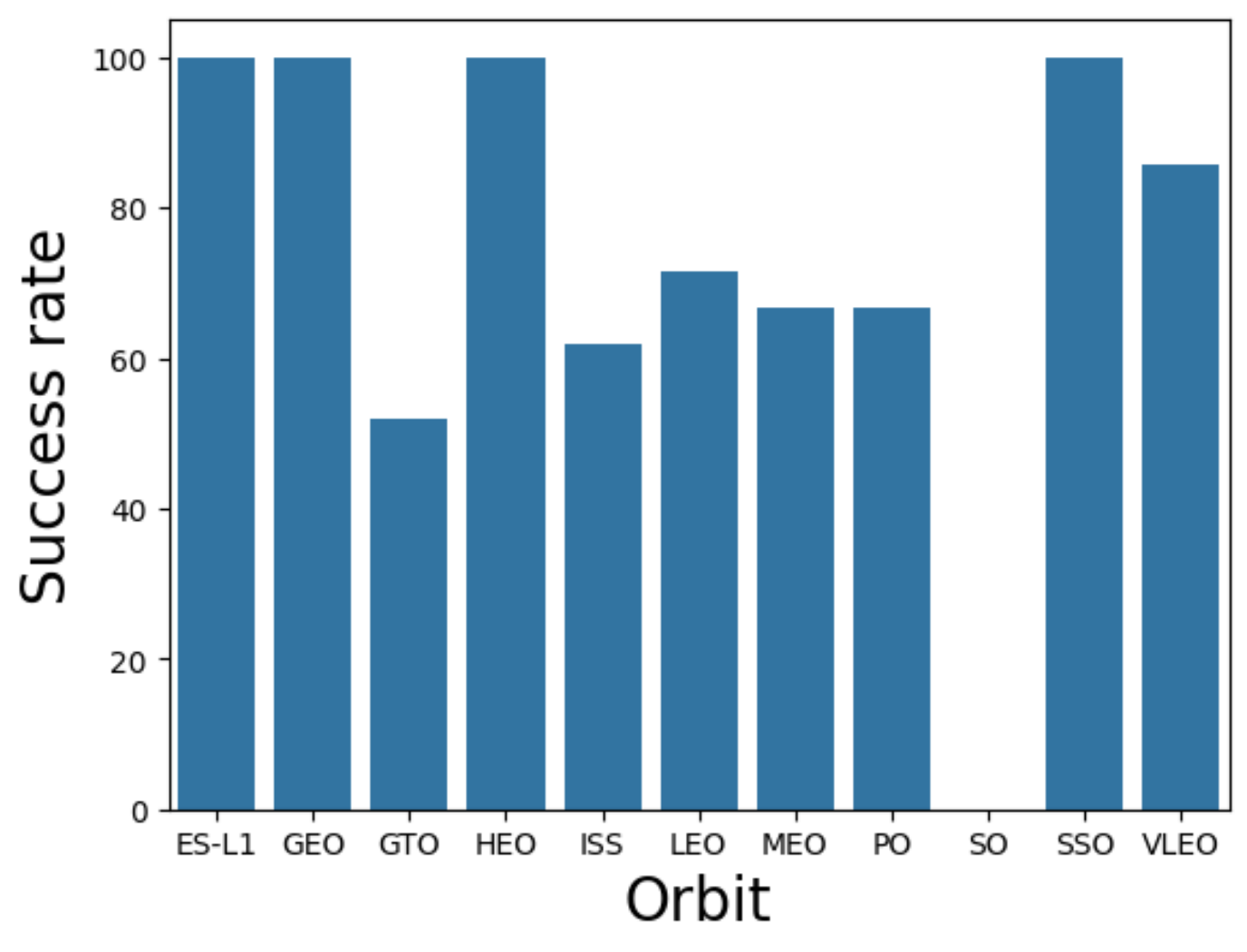
- Higher the flight number, the greater is the chance of successful outcome at any site
- Comparatively less number of launches are made from KSC LC 39A and VAFB SLC 4E but they have a higher success rate than CCSFS SLC 40

Payload vs. Launch Site



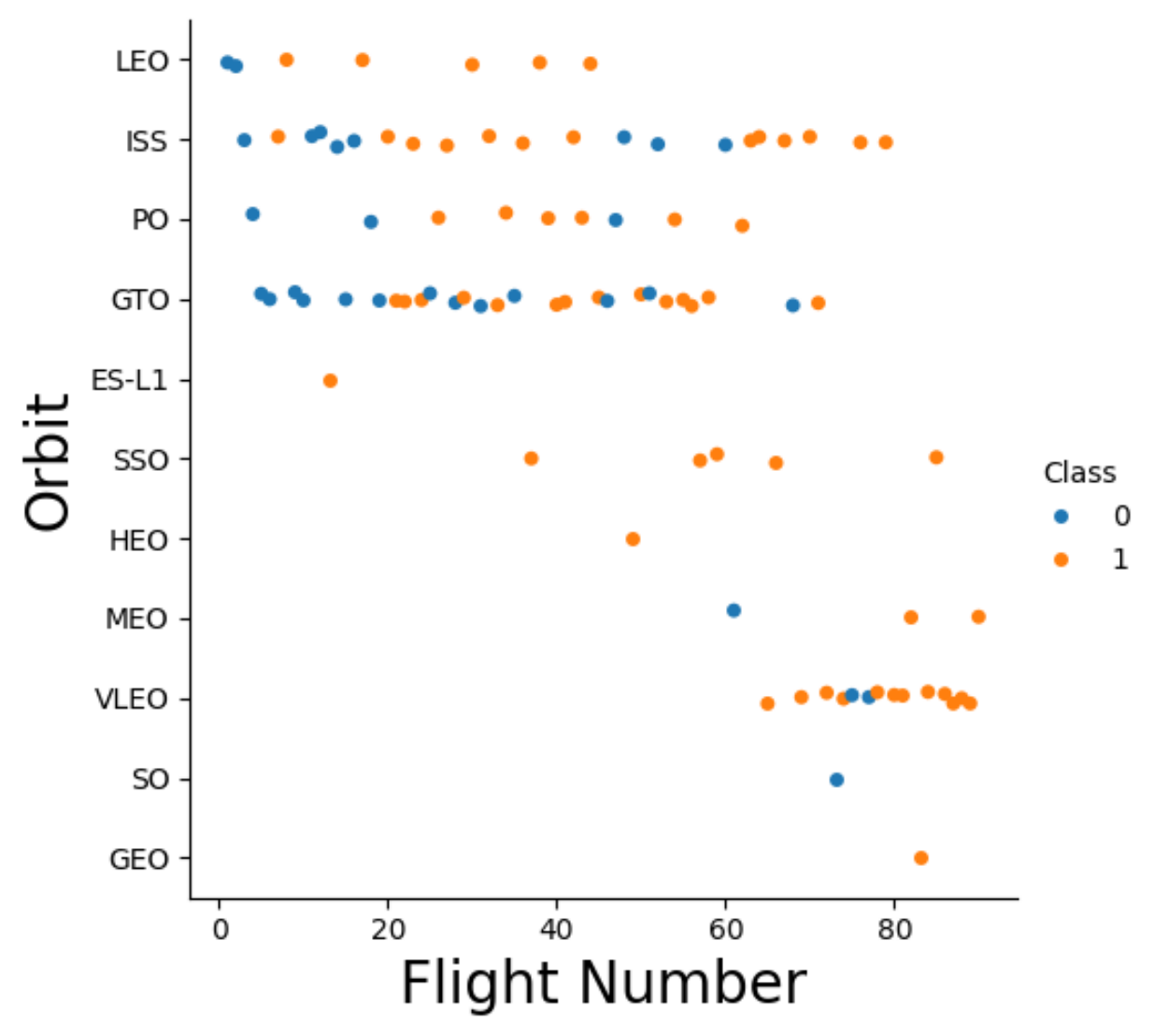
- Higher the payload mass, the greater is the chance of successful outcome at any site
- For the VAFB-SLC 4E launch site there are no rockets launched for heavy payload mass

Success Rate vs. Orbit Type



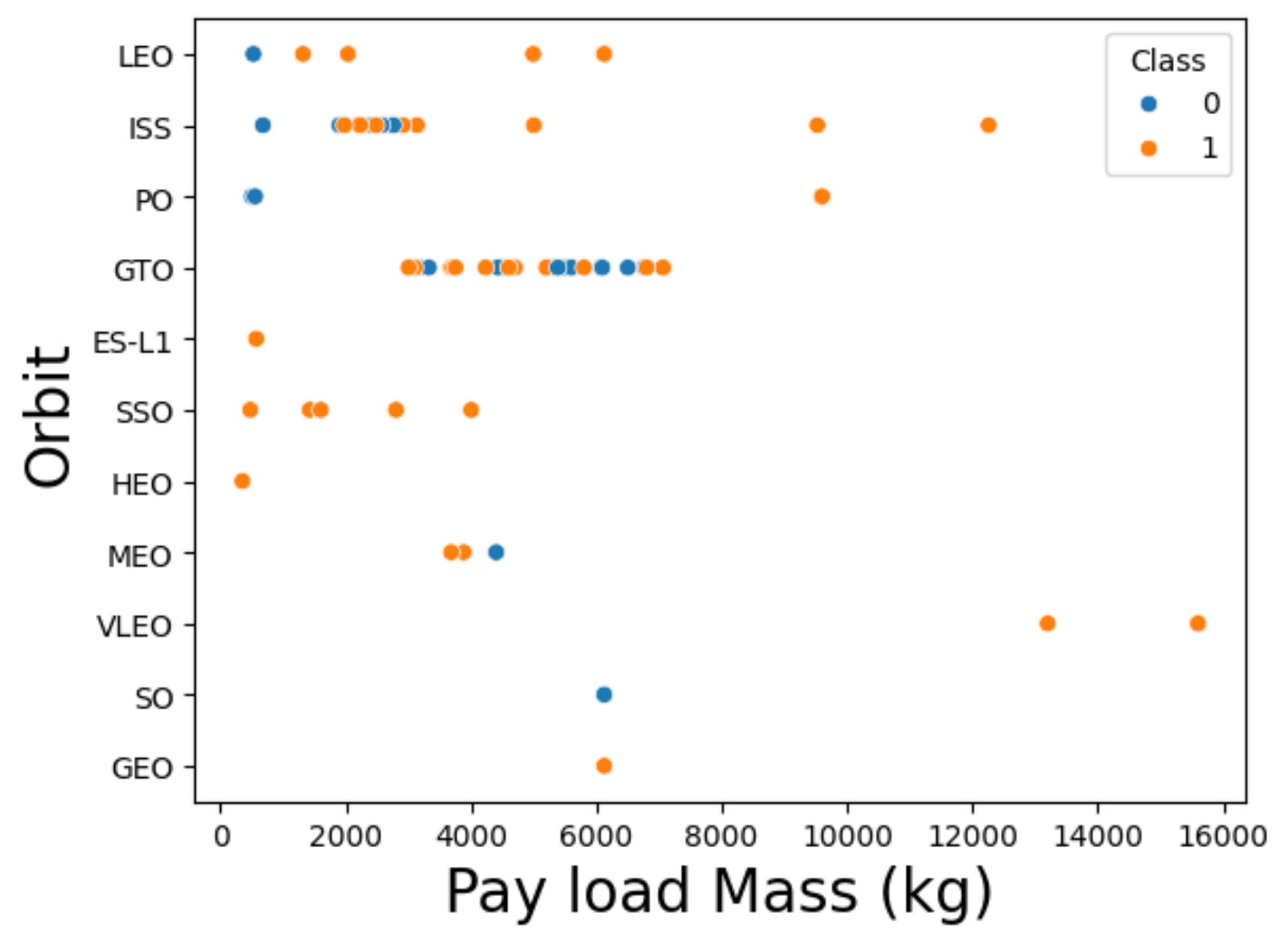
- ES-L1, GEO, HEO and SSO have the highest success rates, SO has the least success rate

Flight Number vs. Orbit Type



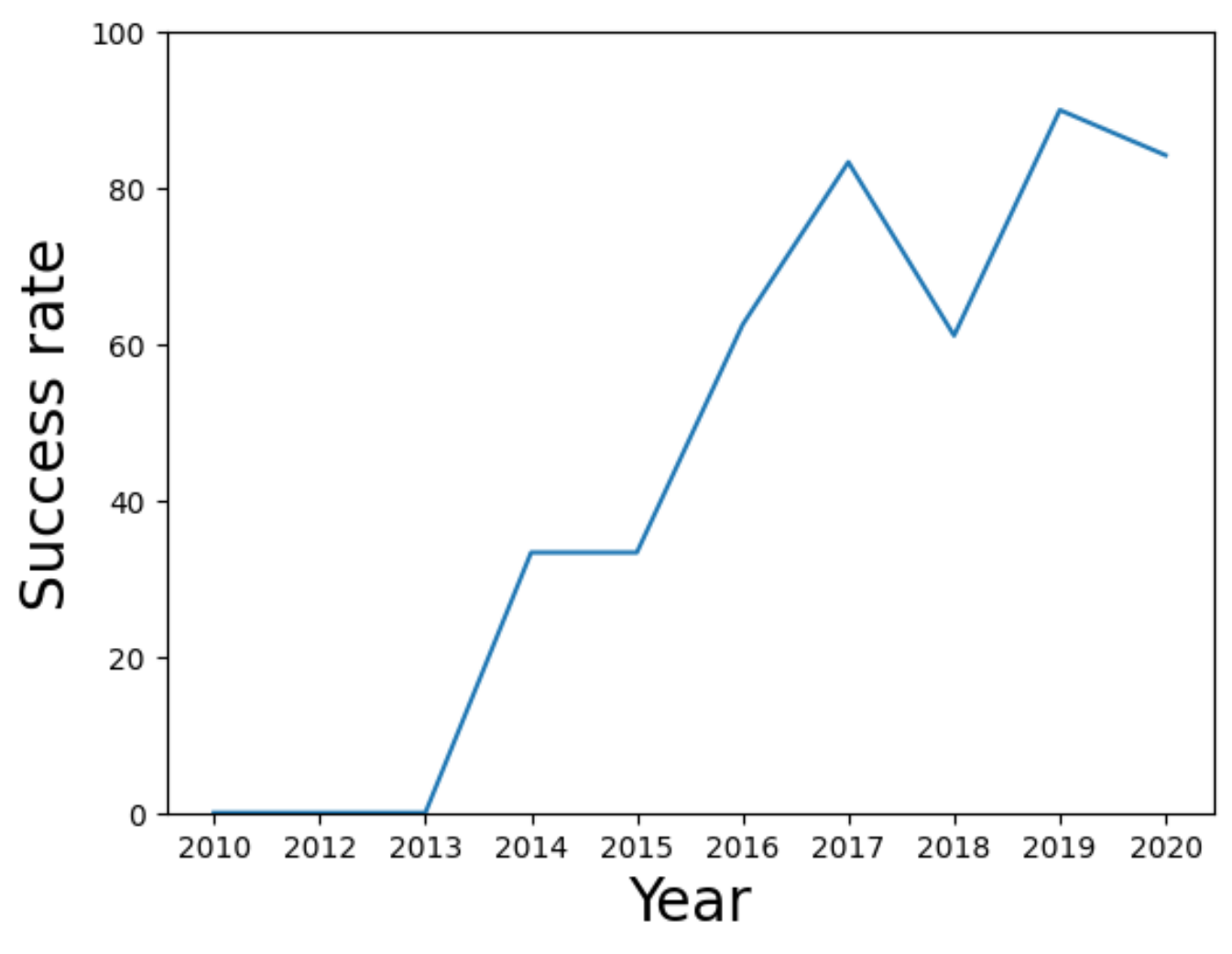
- In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for PO, VLEO and ISS.
- For GTO we cannot distinguish this well as both successful and failed landings are there.

Launch Success Yearly Trend



- The success rate since 2013 kept increasing till 2017 (stable in 2014) and after 2015 it started increasing.
- Success rate dropped at 2018, but started increasing again.

All Launch Site Names

Query :

SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE

The keyword DISTINCT is used to select the distinct launch sites from the dataset

Launch Site Names Begin with 'CCA'

Query :

SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE "CCA%" LIMIT 5;

“%” wildcard was used with LIKE to get the records whose launch site starts 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 :

SELECT SUM("PAYLOAD_MASS__KG_") AS "TOTAL_MASS" FROM SPACEXTABLE WHERE "Customer" LIKE "NASA (CRS)“

SUM function was used to calculate the total payload mass carried by boosters of NASA

Average Payload Mass by F9 v1.1

Query :

SELECT AVG("PAYLOAD_MASS__KG_") AS "AVG MASS" FROM SPACEXTABLE WHERE "Booster_Version" LIKE "F9 v1.1“

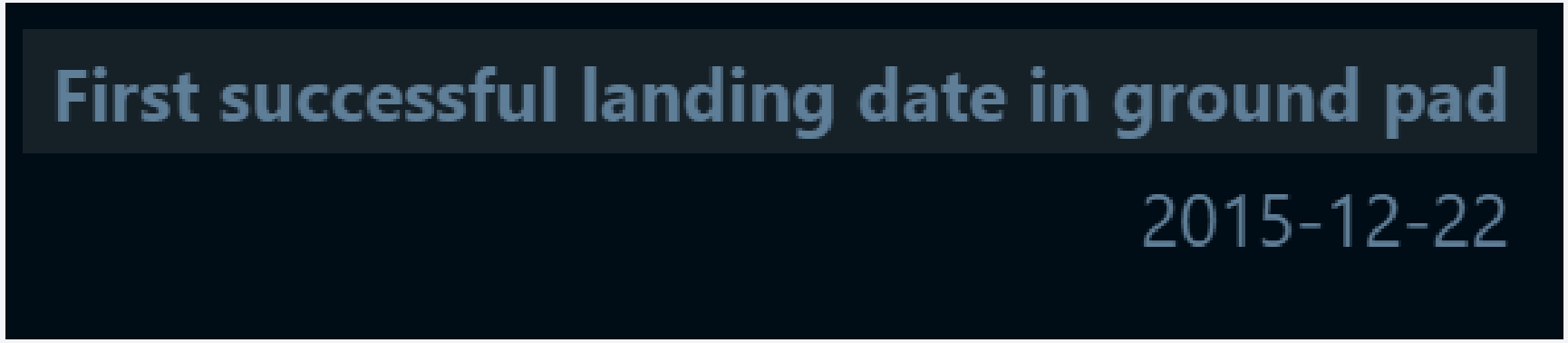
AVG function was used to calculate the average payload mass carried by an F9 v1.1 booster

First Successful Ground Landing Date

Query :

SELECT MIN(DATE) AS "First successful landing date in ground pad" FROM SPACEXTABLE WHERE "Landing_Outcome" LIKE "Success (ground pad)“

LIKE was used to get the dates where the landing outcome was a successful ground landing and MIN function was used to get the minimum date



Successful Drone Ship Landing with Payload between 4000 and 6000

Query :

SELECT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome" = "Success (drone ship)" AND "PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MASS__KG_" < 6000

Data was selected based on landing outcome and payload mass

Total Number of Successful and Failure Mission Outcomes

Query :

SELECT COUNT("Mission_Outcome") AS "SUCCESS" FROM SPACEXTABLE WHERE "Mission_Outcome" LIKE "Success%"

Query :

SELECT COUNT("Mission_Outcome") AS "FAILURE" FROM SPACEXTABLE WHERE "Mission_Outcome" LIKE "Failure%"

“%” wildcard was used with LIKE to get the successful and failed records

Boosters Carried Maximum Payload

Query :

```
SELECT "Booster_Version" FROM
SPACEXTABLE WHERE
"PAYLOAD_MASS__KG_" = (SELECT
MAX("PAYLOAD_MASS__KG_") FROM
SPACEXTABLE);
```

Subquery was used to get the maximum payload mass using MAX function and then booster versions of the records whose payload mass matched the maximum 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

Query :

SELECT substr(Date, 6, 2) AS Month, Booster_Version, Launch_Site, Landing_Outcome FROM SPACEXTABLE WHERE "Landing_Outcome" = "Failure (drone ship)" AND substr(Date, 0, 5) = '2015'

substr function was used on Date to get the year and month for each record, the year was used for selecting the records.

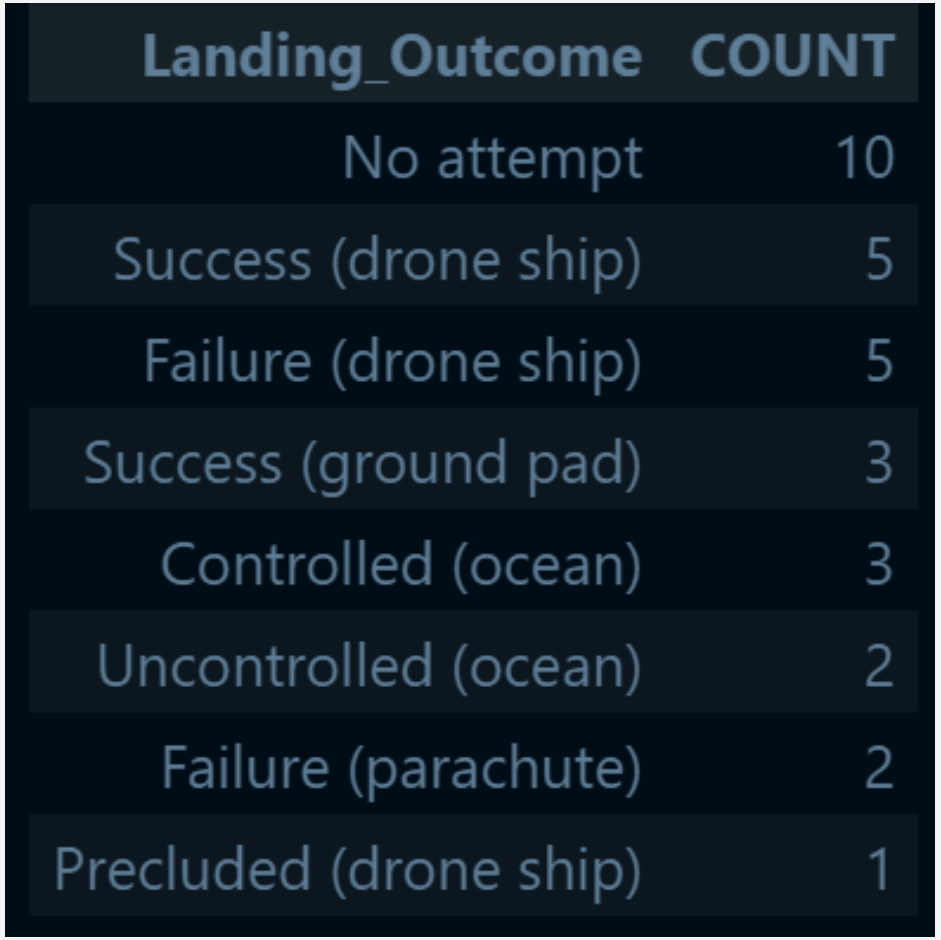
Month	Booster_Version	Launch_Site	Landing_Outcome
01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

Query :

SELECT "Landing_Outcome", COUNT("Landing_Outcome") AS COUNT FROM SPACEXTABLE WHERE "Date" BETWEEN "2010-06-04" AND "2017-03-20“ GROUP BY "Landing_Outcome“ ORDER BY COUNT DESC

BETWEEN is used to get data in the specified date range, GROUP BY is used to group the data based on landing outcomes and the result is sorted in descending order by ORDER BY and DESC



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

All launch sites on map

VAFB SLC-4E was found to be near Los Angeles while CCAFS LC-40, CCAFS SLC-40 and KSC LC-39A were found very close to each other near Florida

Successful and failed launches from sites

The clip of map shows 9 successful launches (marked by green) and 4 failed launches (marked by red) made from the launch site KSC LC-39A

Distances of CCAFS SLC-40 to its proximities

The launch site is at 0.59Km from Samuel C Phillips pkwy and 0.86Km from coastline

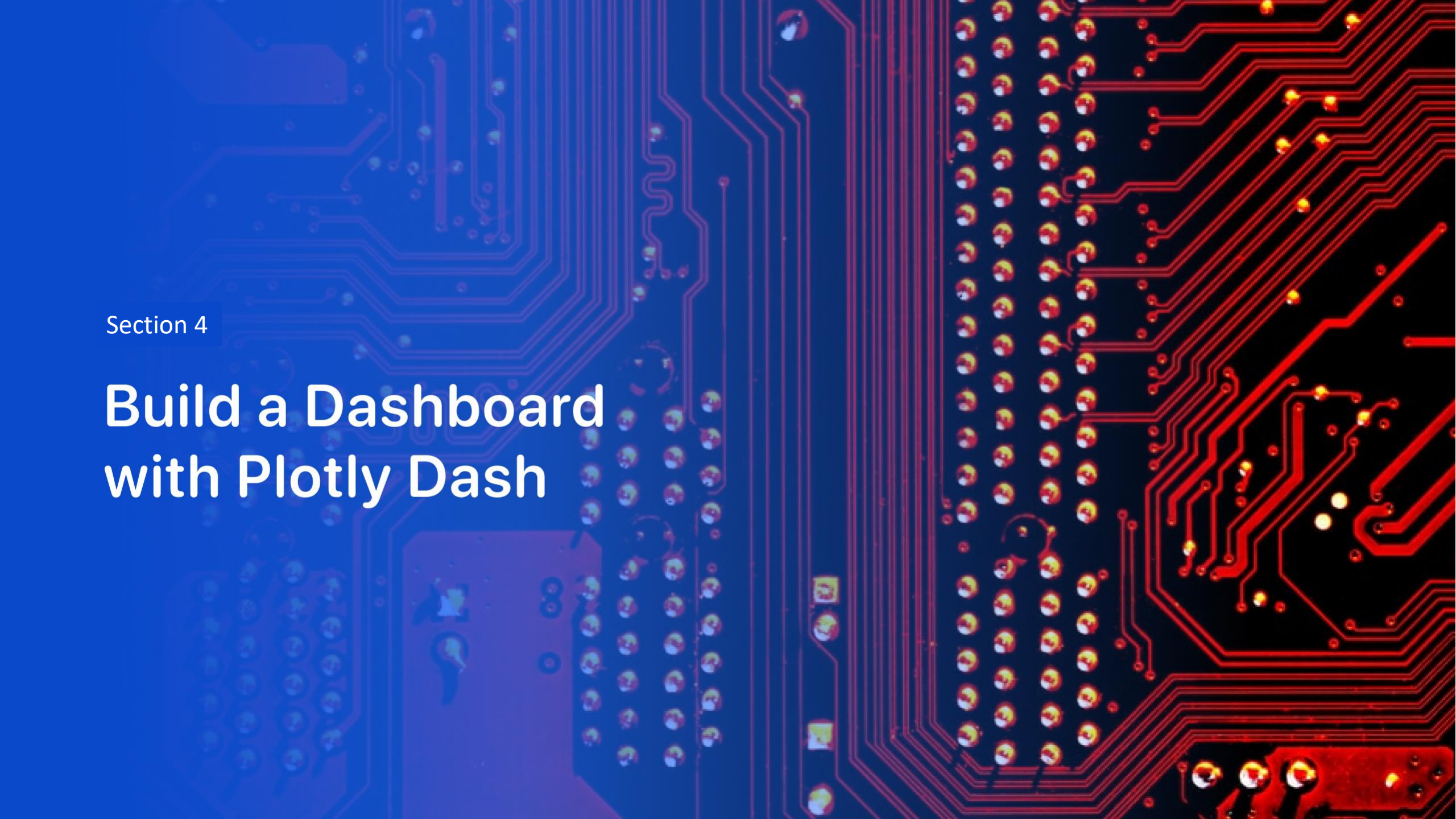
The launch site is at 1.26Km from NASA railway road

The launch site is at 19.26Km from Cape Canaveral city

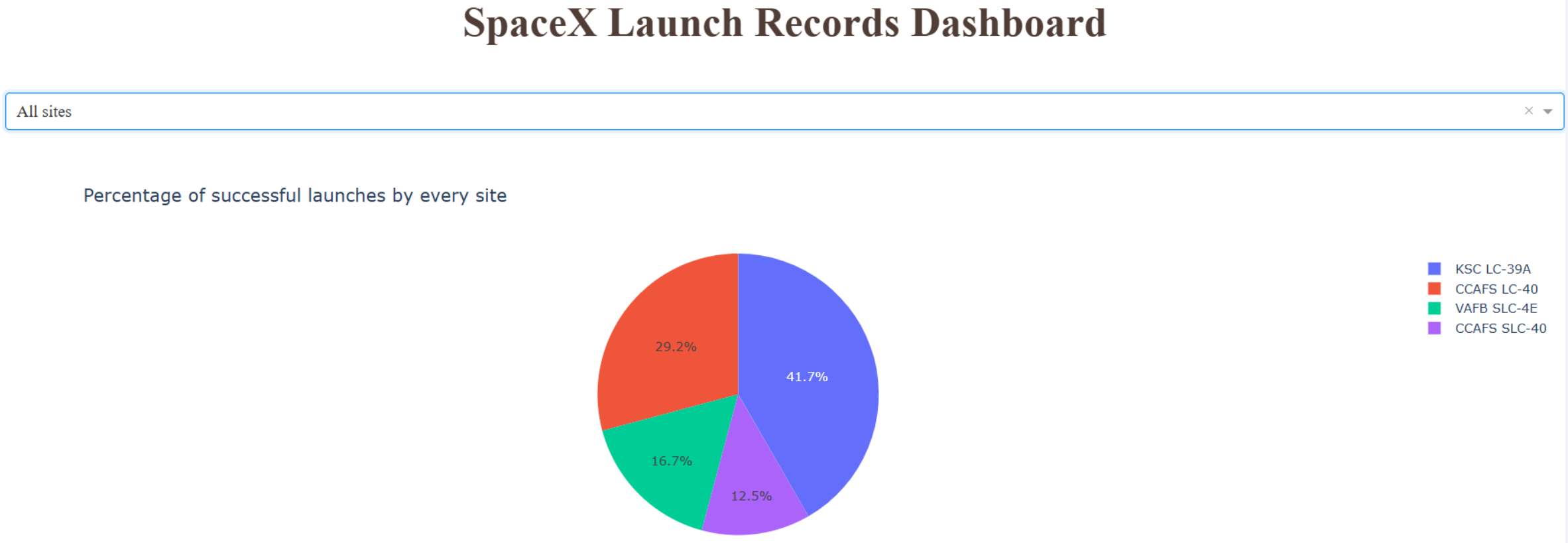
The launch site is close to highway and railway as these are used for transportation of materials.

The launch site is also close to coastline.

The launch site is relatively far from cities, which is obvious as launches can be dangerous.

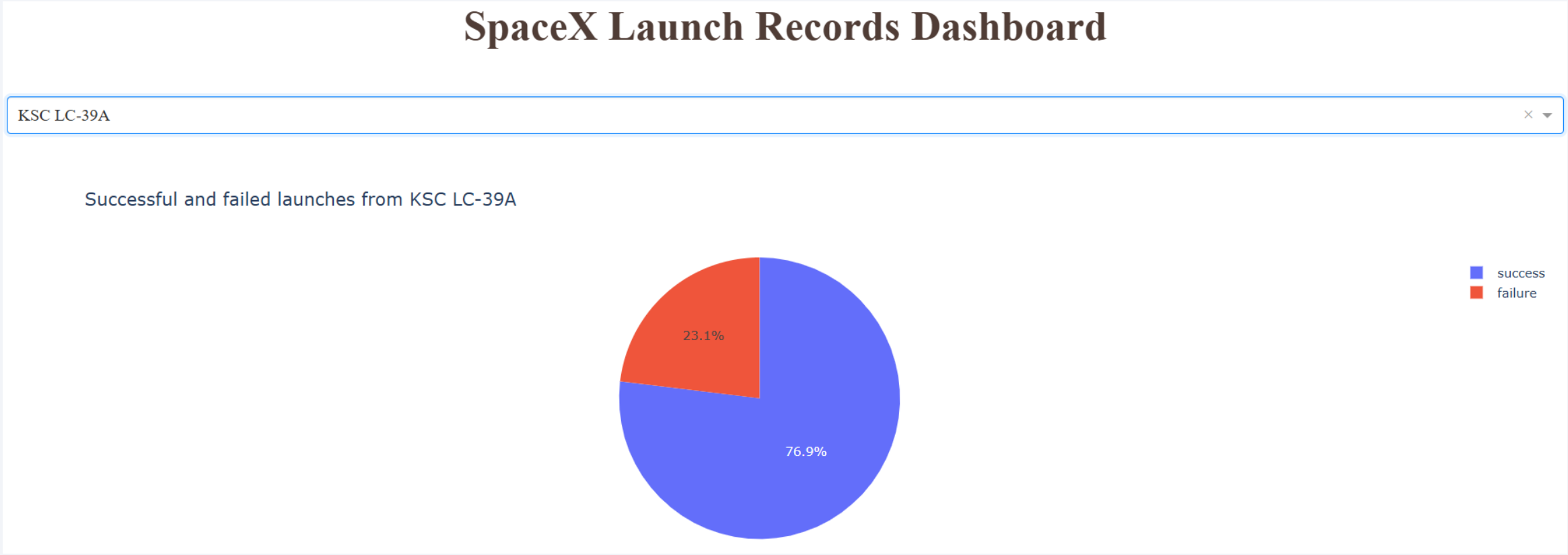

Section 4
Build a Dashboard with Plotly Dash

Launch success count for all sites



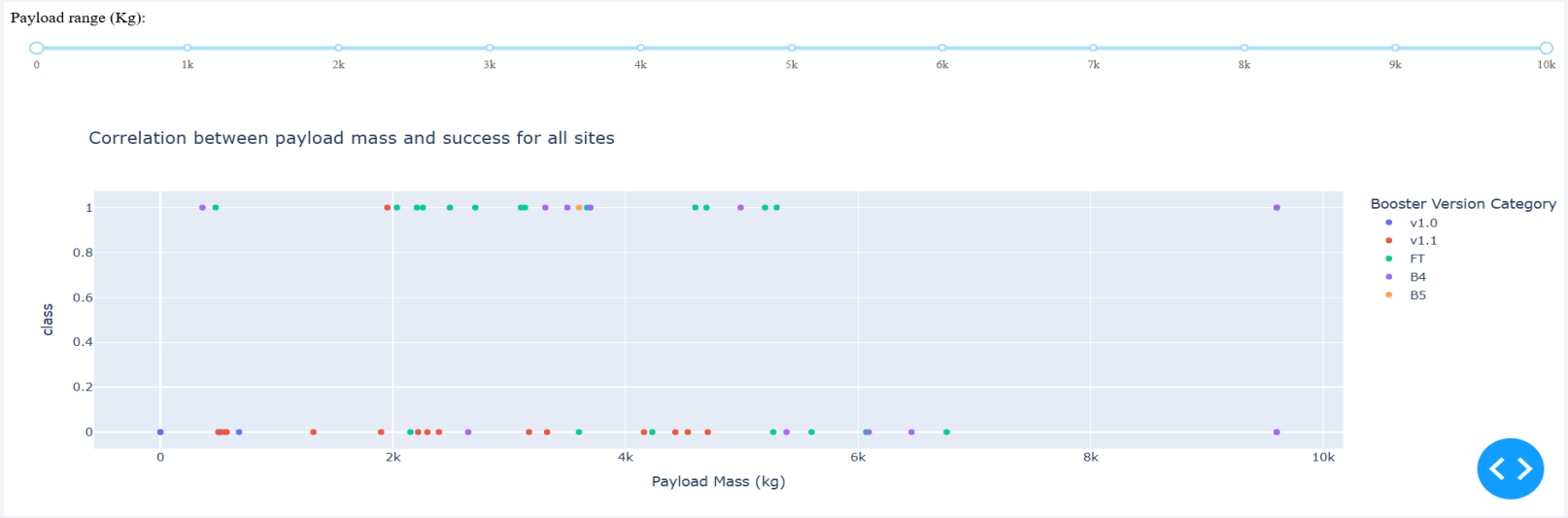
KSC LC-39A contributes most to the successful launches statistics, 41.7% of the successful launches made were from this site

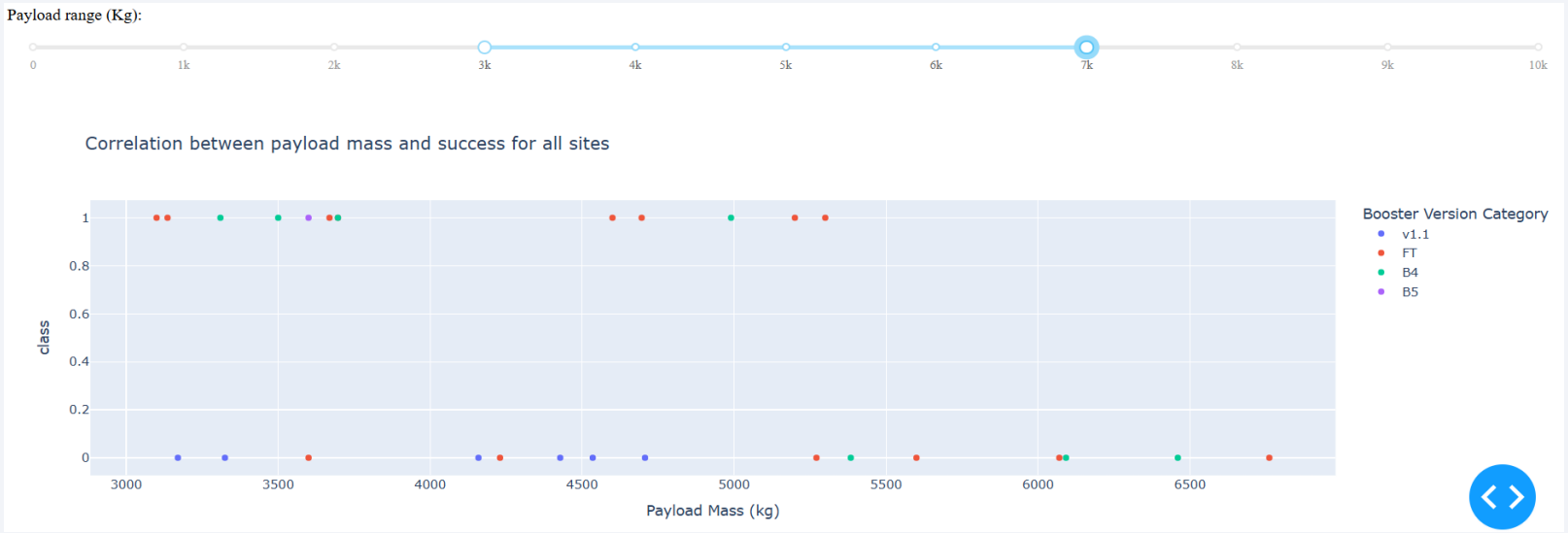
Launch statistics of KSC LC-39A



76.9% of launches made from this site were successful, making KSC LC-39A the sites with most successful launches

Payload mass vs launch outcome for all sites



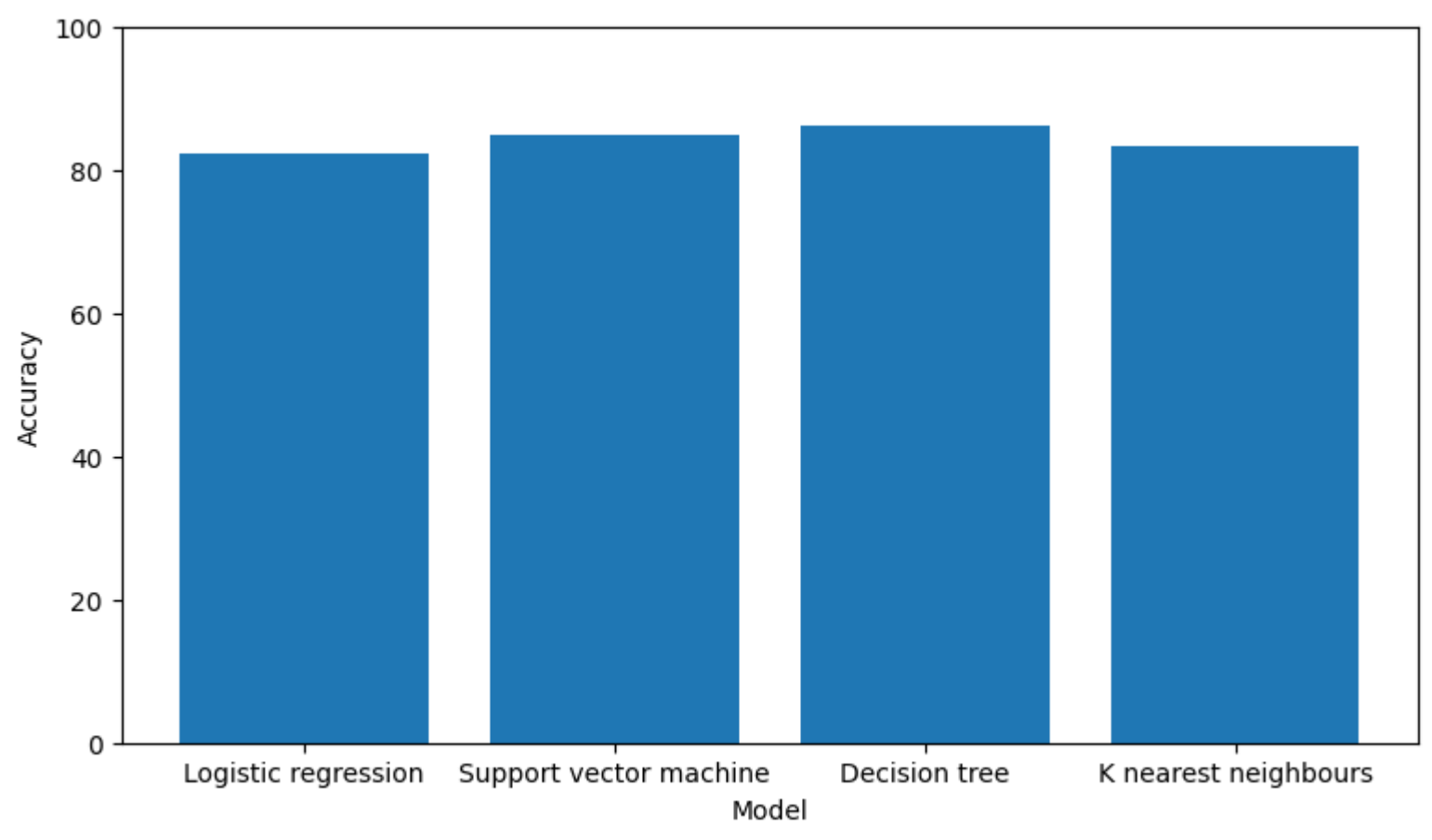


Rockets with payload mass between 2000kg and 6000kg have higher success rate.

Section 5

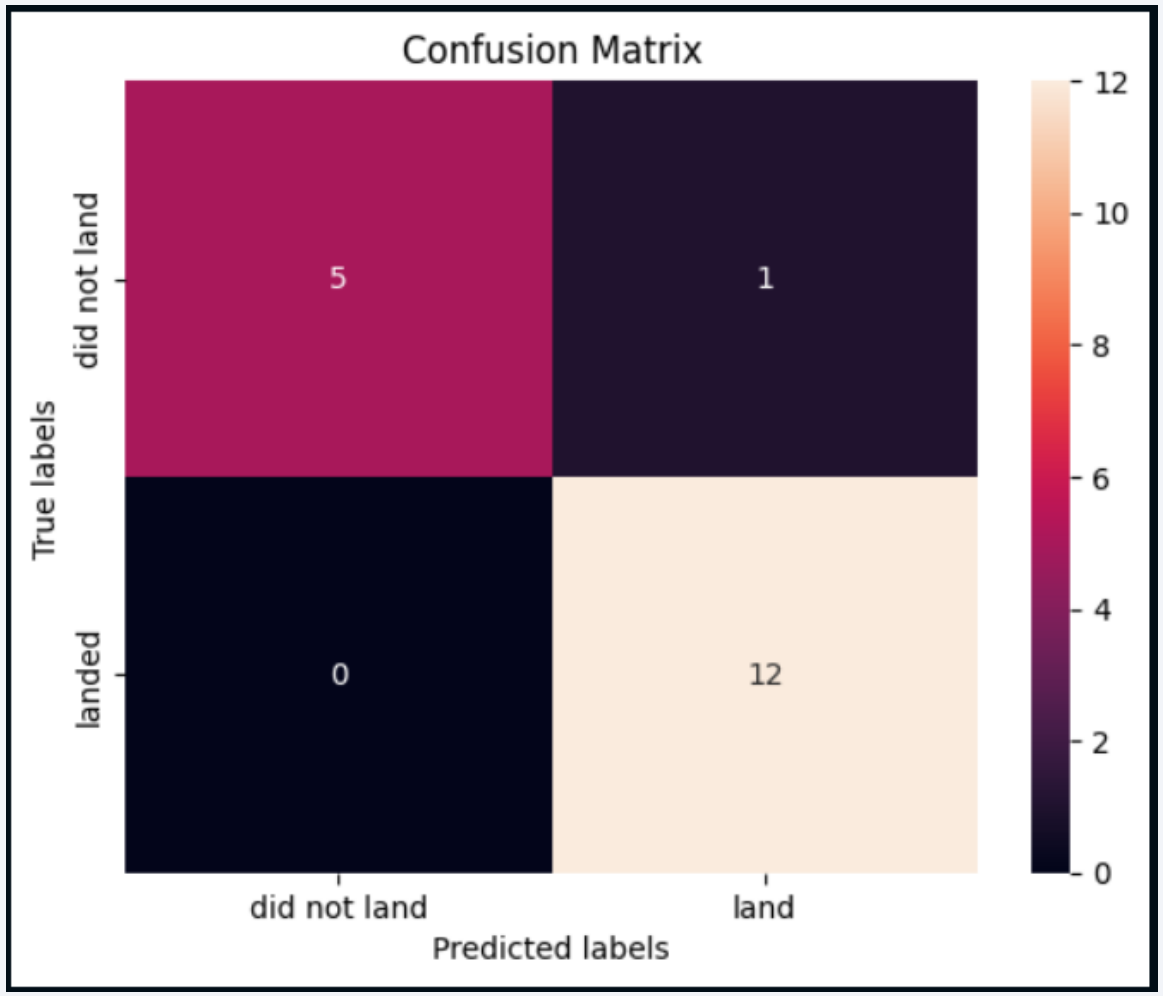
Predictive Analysis (Classification)

Classification Accuracy



Decision tree classifier has the highest accuracy score, which is almost 86%

Confusion Matrix



The confusion matrix of decision tree classifier shows that :

- Out of 6 “did not land” outcomes, the model predicted 5 correctly
- The model predicted all “landed” outcomes correctly

Conclusions

- Decision Tree Model is the best algorithm for this dataset.
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate
- The success rate of launches has increased from 2013 to 2020 but a sudden dip was seen in 2018

Thank you!