IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Martin P.
7.5.2025

Outline

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

Executive Summary

Summary of methodologies

- Data collection through API & web scraping
- Data wrangling
- Exploratory Data Analysis
  - Data Visualization
  - SQL
- Dashboard & Interactive Map analytics
- Classification model pipeline

Summary of all results

- Exploratory Data Analysis result presentation
- Interactive maps and dashboard results
- Classification model results

Introduction

Project background and context

The project is set in the era of commercial spaceflight, where companies like SpaceX have revolutionized the industry by making space launches more affordable, largely due to reusable rocket technology. As a data scientist at a fictional competitor, SpaceY, I'm tasked with analyzing SpaceX's operations to help estimate launch costs and understand the reuse of rocket components.

Problems I want to find answers to

The project aims to answer whether SpaceX will reuse the first stage of a rocket for a given launch, using machine learning and public data. This prediction is crucial because the reuse of the first stage significantly reduces launch costs.

Section 1

Methodology

Methodology

Executive Summary

Data collection methodology: Data was sourced from two endpoints. SpaceX REST API and Wiki pages. These sources provided records about launch dates, sites, payloads, orbit, booster versions, outcomes of the launches and various other attributes of each launch.

Data wrangling: Collected data was enriched by normalizing data, imputing missing data, removing irrelevant columns, feature engineering and by creating a landing outcome label based on the Outcome column

Exploratory data analysis (EDA) using visualization and SQL:

- Visualized various features combinations using Matplotlib and Seaborn
- SQL queries to gain deeper insight on the structure of the data

Interactive visual analytics using Folium and Plotly Dash

- Folium was used for the visualization of geographical launch sites and to check the proximity of various landscape objects from the launch site.
- Developed dashboard for interactive analysis of launch success rates by different combination of launch sites and payload range.
- Normalized data have been divided in training and test data. Evaluated by four different classification models by accuracy metric on test data. Multiple sets of hyperparameters have been considered.

Predictive analysis using classification models

- Analyzed four classification models. The best model was evaluated using the accuracy metric on test data. Multiple sets of hyperparameters have been evaluated utilizing cross validation.

Data Collection

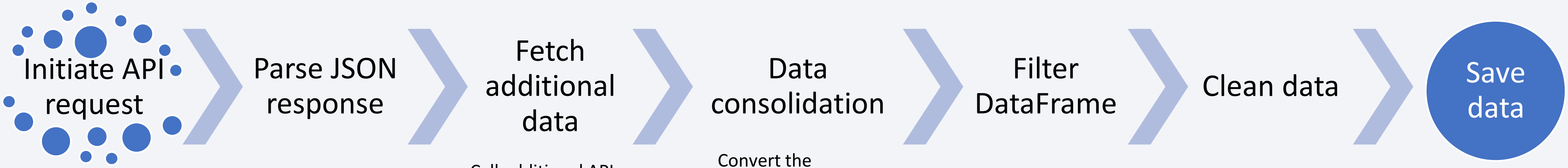
Data on past SpaceX launches have been collected from the SpaceX REST API (api.spacexdata.com/v4/launches/past), using Python's requests library to make GET requests Additional API endpoints (e.g., /rockets, /cores, /payloads) are used to enrich the data by resolving the relevant ID fields into meaningful values.

The BeautifulSoup package is used to scrape Falcon launch data from HTML tables on relevant Wikipedia page ensuring access to the latest lunch information.

Following columns have been fetched:

SpaceX Rest API	Wikipedia
– FlightNumber, – Date, – BoosterVersion, – PayloadMass, – Orbit, – LaunchSite, – Outcome, – Flights, – GridFins, – Reused, – Legs, – LandingPad, – Block, – ReusedCount, – Serial, – Longitude, – Latitude	– Flight No., – Launch site, – Payload, – PayloadMass, – Orbit, – Customer, – Launch outcome, – Version Booster, – Boosterlanding, – Date & Time

Data Collection – SpaceX API



Initiate API request

Sent a GET request to the SpaceX API endpoint

Parse JSON response

Convert JSON response to Python dictionary

Fetch additional data

Call additional API endpoints for detailed lunch data referenced by respective IDs from main API endpoint

Data consolidation

Convert the response data from various endpoint into single comprehensive Python DataFrame object

Filter DataFrame

Filter the DataFrame to include only Falcon 9 lunch entries

Clean data

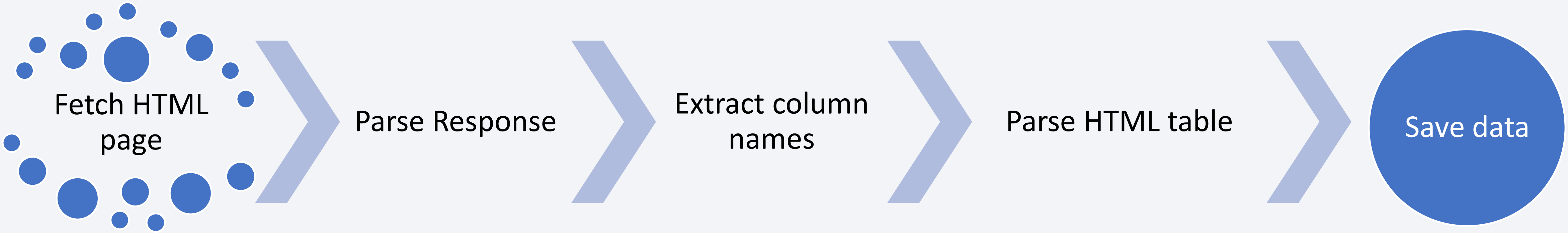
Impute missing values using the mean value

Save data

Save the cleaned DataFrame to csv for further reference

https://github.com/mp-vue/applied-data-science-capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping



Fetch HTML page

Parse Response

Extract column names

Parse HTML table

Save data

Sent a GET request using Python’s requests library

Parse content of the HTML response using the BeautifulSoup library.

Extract column names from the HTML table containing launch data

Locate relevant part of the HTML content and parse the data to DataFrame

Save the cleaned DataFrame to csv for further reference

https://github.com/mp-vue/applied-data-science-capstone/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

Data wrangling involves cleaning, transforming and organizing raw data into a structured format suitable for analytical analysis. Decision on the training labels and preliminary exploratory analysis have been conducted.

Explored the number of launches on each site

Explored the number and occurrence of each orbit

Explored the number and occurrence of mission outcome per orbit type

Conversion of outcomes into training Labels -> 1 means the booster successfully landed 0 means it was unsuccessful.

https://github.com/mp-vue/applied-data-science-capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

1. **Scatter plots – used for pattern / correlation analysis of two variables**
   - Identify any patterns between the launch sequence and the location of launch site.
   - Identify any patterns between the launch sequence and the payload.
   - Identify any patterns between the payload and the launch location
   - Identify any patterns between the launch sequence and the type of orbit the mission aimed to achieve
   - Identify any potential relationships between the payload and the type of orbit targeted by the launch
2. **Bar charts – used for comparison of categorical variables**
   - Visualize the success rate of missions for each specific orbit type (e.g., what percentage of launches achieved their intended orbit for each category)
3. **Line plot – used for identifying time trends**
   - Identify trends in launch success rates over time (yearly). It reveals whether the success rate is improving, declining, or staying consistent across years.

https://github.com/mp-vue/applied-data-science-capstone/blob/main/edadataviz.ipynb

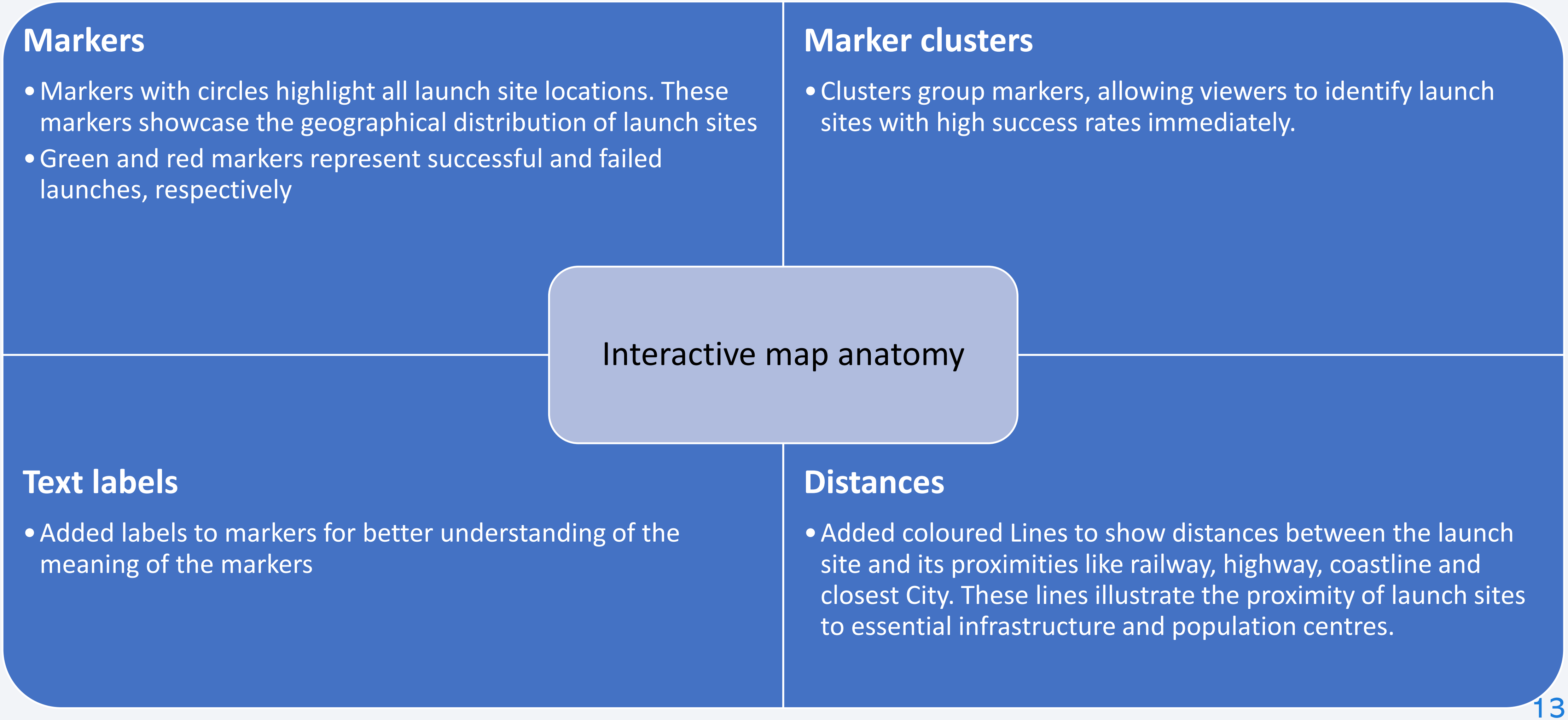
EDA with SQL

Performed SQL queries:

- Displayed the names of the unique launch sites in the space mission
- Displayed 5 records where launch sites begin with the string ‘CCA’
- Displayed the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listed the date when the first successful landing outcome in ground pad was achieved
- Listed the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listed the total number of successful and failure mission outcomes
- Listed the names of the booster versions which have carried the maximum payload mass
- Listed the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
- Ranked the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) in descending order

https://github.com/mp-vue/applied-data-science-capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium



Interactive map anatomy

Markers

- Markers with circles highlight all launch site locations. These markers showcase the geographical distribution of launch sites
- Green and red markers represent successful and failed launches, respectively

Marker clusters

- Clusters group markers, allowing viewers to identify launch sites with high success rates immediately.

Text labels

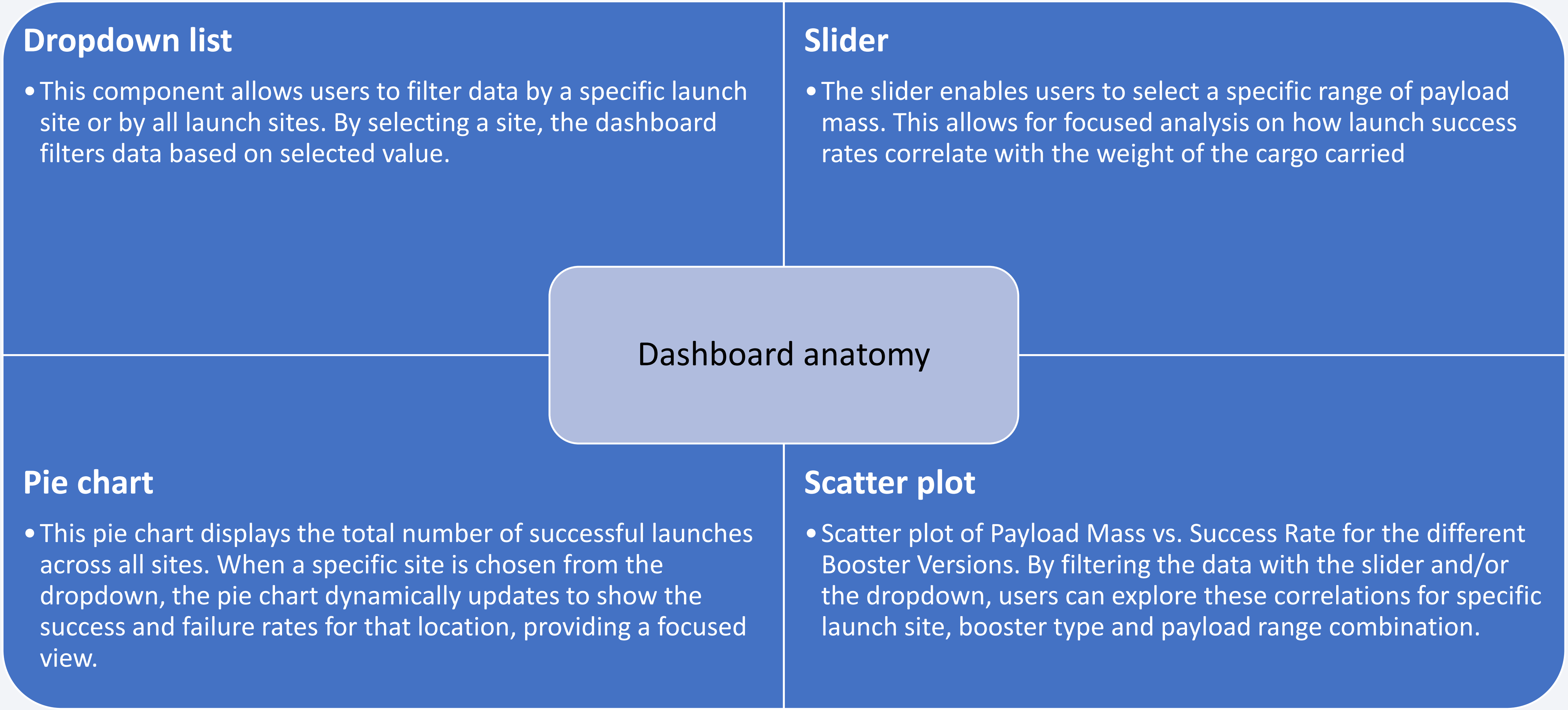
- Added labels to markers for better understanding of the meaning of the markers

Distances

- Added coloured Lines to show distances between the launch site and its proximities like railway, highway, coastline and closest City. These lines illustrate the proximity of launch sites to essential infrastructure and population centres.

Github https://github.com/mp-vue/applied-data-science-capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash



Dropdown list

- This component allows users to filter data by a specific launch site or by all launch sites. By selecting a site, the dashboard filters data based on selected value.

Slider

- The slider enables users to select a specific range of payload mass. This allows for focused analysis on how launch success rates correlate with the weight of the cargo carried

Dashboard anatomy

Pie chart

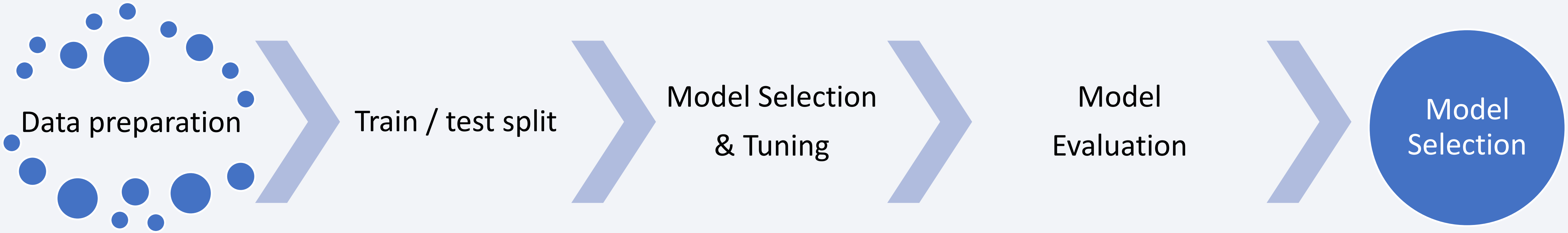
- This pie chart displays the total number of successful launches across all sites. When a specific site is chosen from the dropdown, the pie chart dynamically updates to show the success and failure rates for that location, providing a focused view.

Scatter plot

- Scatter plot of Payload Mass vs. Success Rate for the different Booster Versions. By filtering the data with the slider and/or the dropdown, users can explore these correlations for specific launch site, booster type and payload range combination.

https://github.com/mp-vue/applied-data-science-capstone/blob/main/spacex-dash-app.py

Predictive Analysis (Classification)



Data preparation

Train / test split

Model Selection & Tuning

Model Evaluation

Model Selection

Select relevant features, label column, impute missing values, normalize data

Divide the data into training and testing sets for model training and evaluation.

Apply GridSearchCV with cross-validation (cv=10) to various models (Logistic Regression, SVM, Decision Tree, KNN) to identify the optimal hyperparameter settings for each model

Model has been evaluated using accuracy metric and confusion matrices to assess performance on unseen data.

Select the model with the highest performance based on the evaluation metrics

https://github.com/mp-vue/applied-data-science-capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

The results of the assignment will be categorized to 3 main parts, which is:

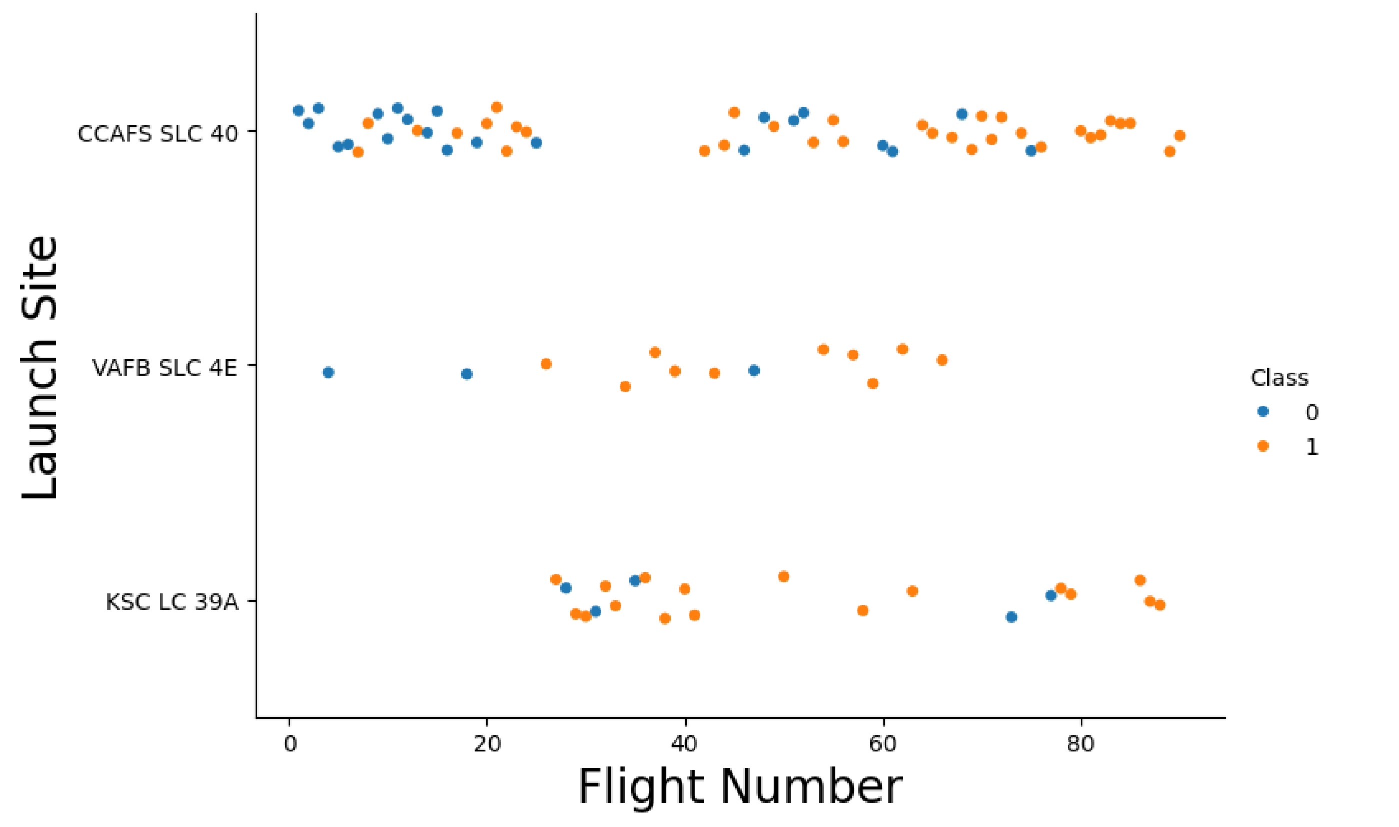
1. Insights drawn from EDA
2. Launch site proximity analysis
3. Predictive analysis results

Section 2

Insights drawn from EDA

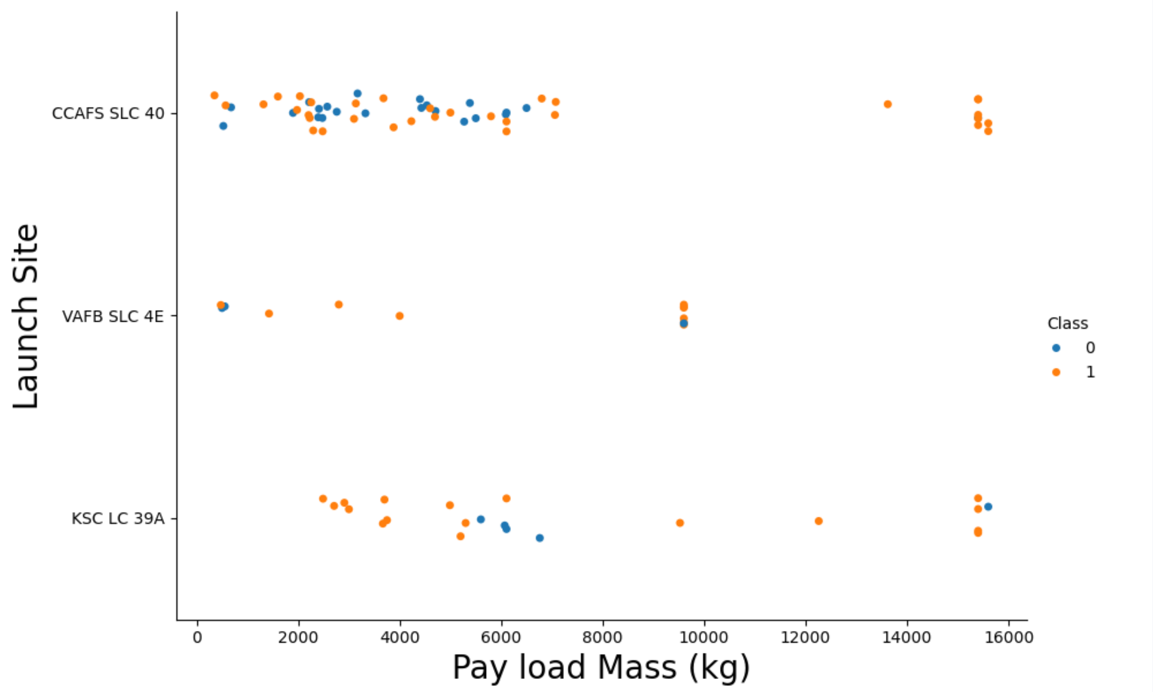
Flight Number vs. Launch Site

- The earliest flights have high failure rate. The CCAFS SLC 40 launch site has about half of all launches. There is significant increase of success rate in the recent launches.



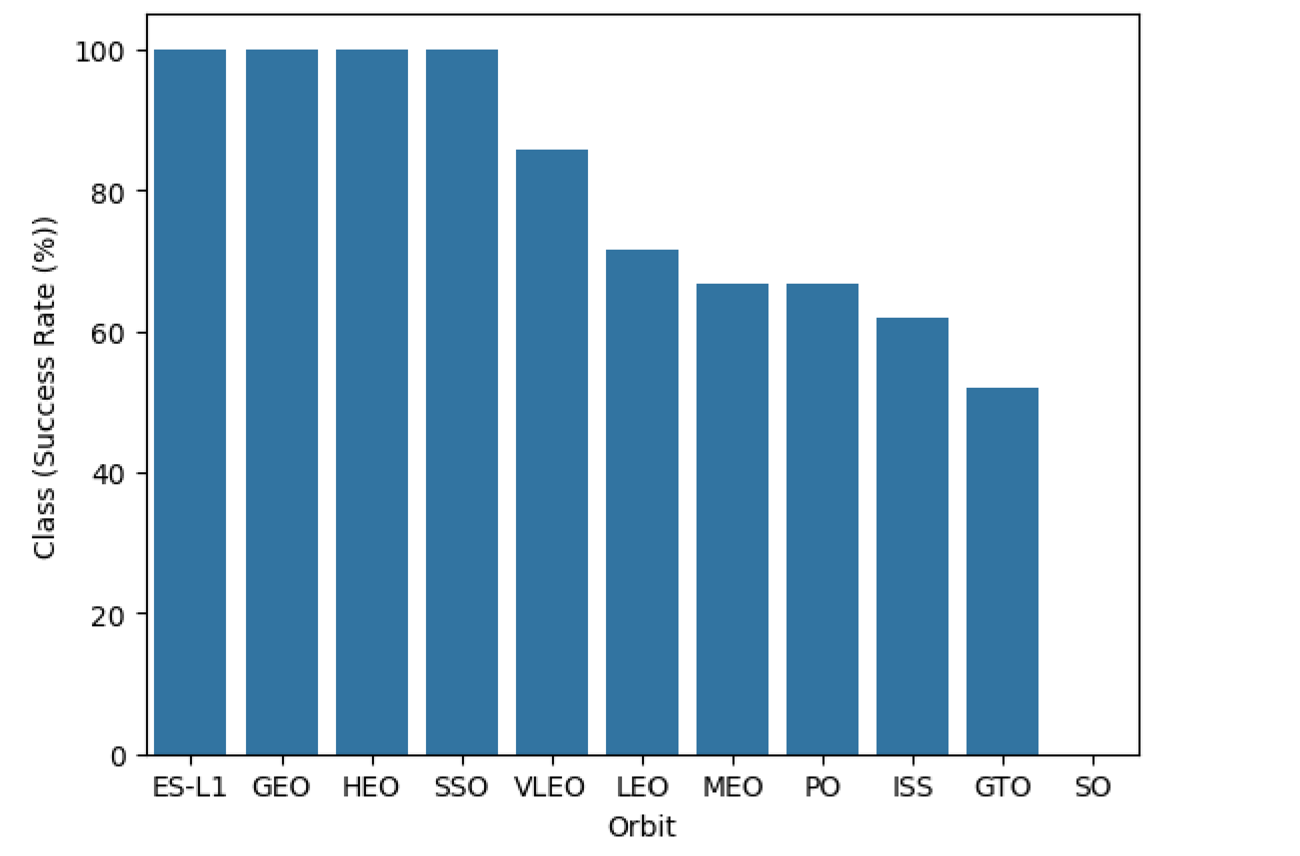
Payload vs. Launch Site

- Depending on the launch site a heavier payload may be a consideration for a successful landing. VAFB SLC-4E launch site was not used with payload greater than 10000. All launch sites were mostly used with payload lighter than 8000kg



Success Rate vs. Orbit Type

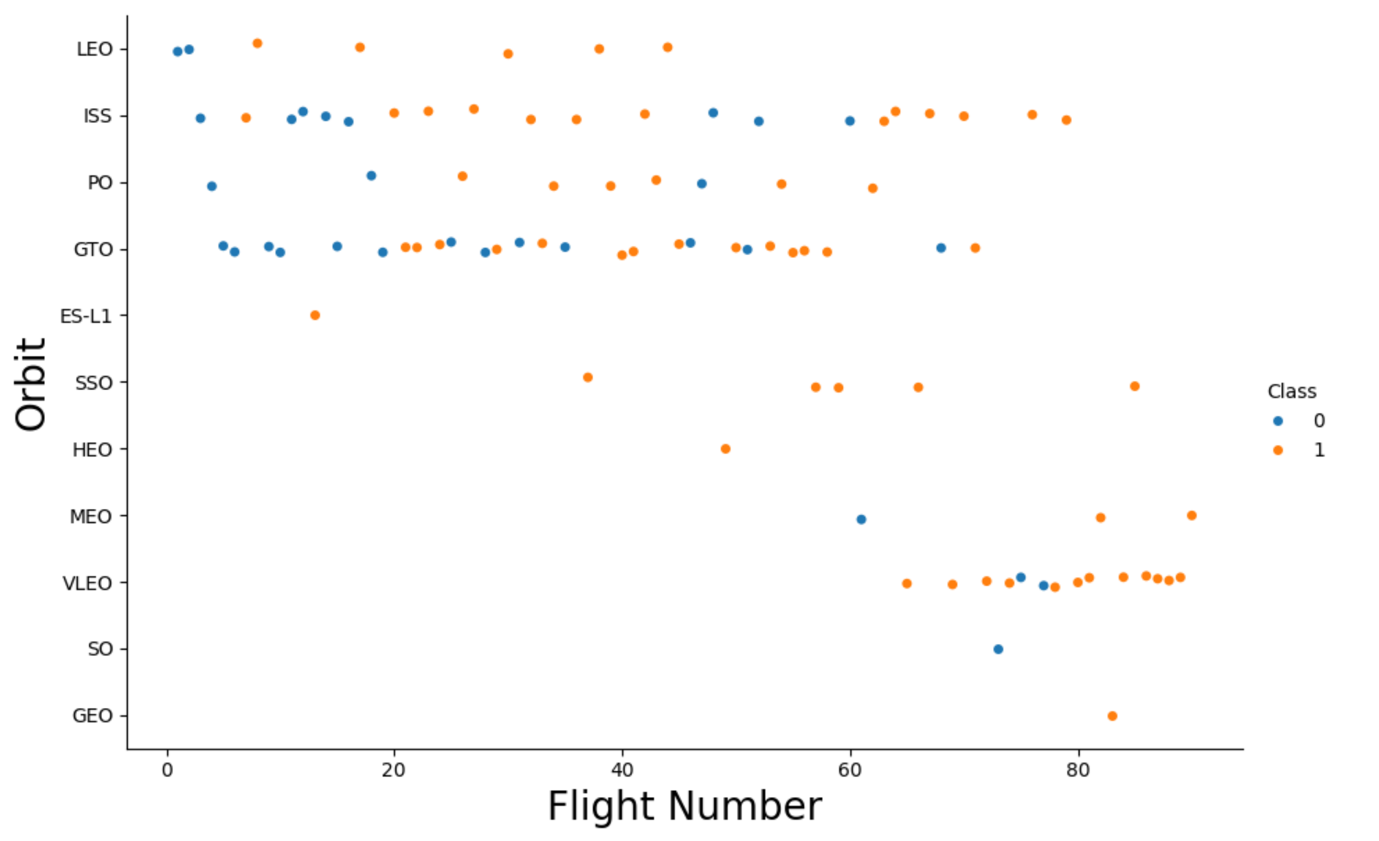
- ES-L1, GEO, HEO, SSO are all orbits with 100% success rate.
- VLEO, LEO, MEO, PO, ISS, GTO are orbits with success rate between 50% and 85%
- SO has no successful launch



Flight Number vs. Orbit Type

- Most of the early flights were done to LEO, ISS, PO and GTO orbits

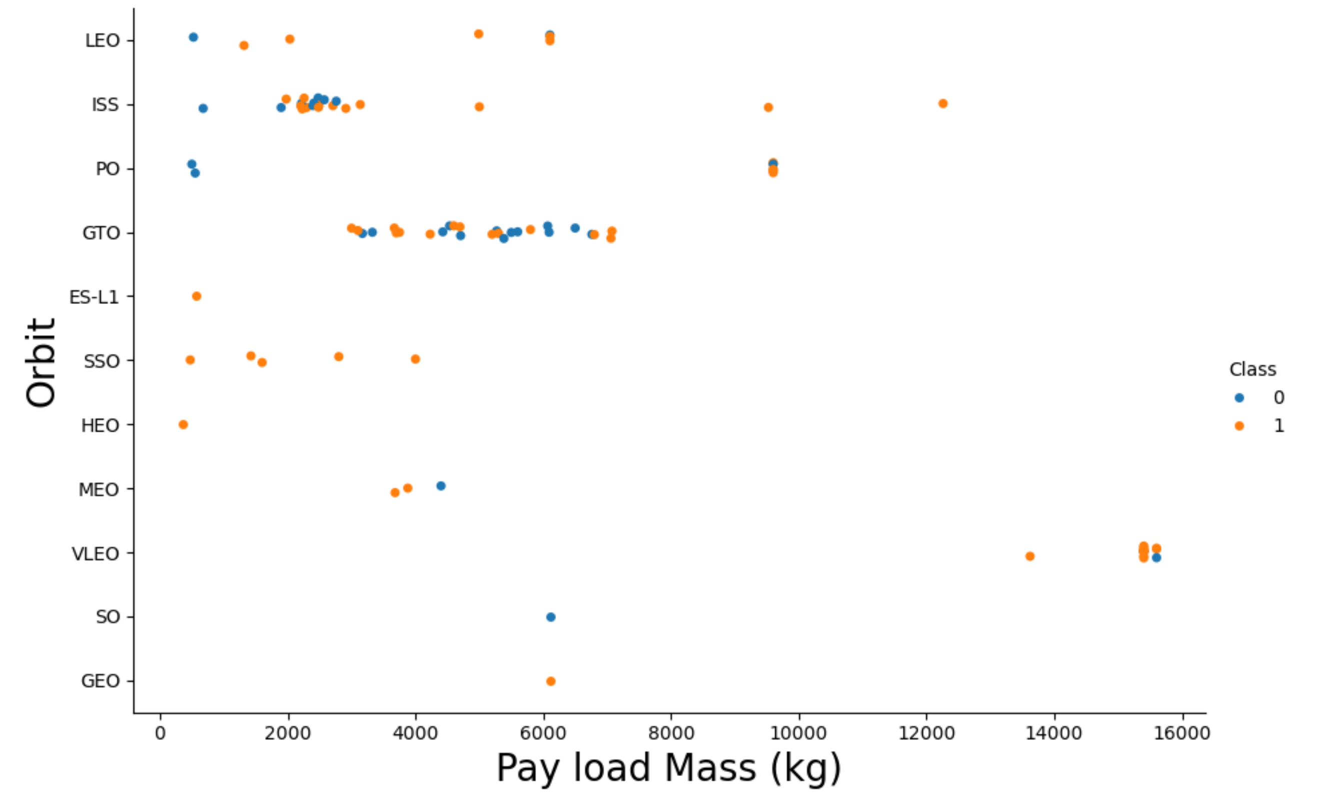
- The success rate increases with the number of flights for the LEO and MEO orbit. For other orbits there seems to be no relation between the success rate and the number of flights



Payload vs. Orbit Type

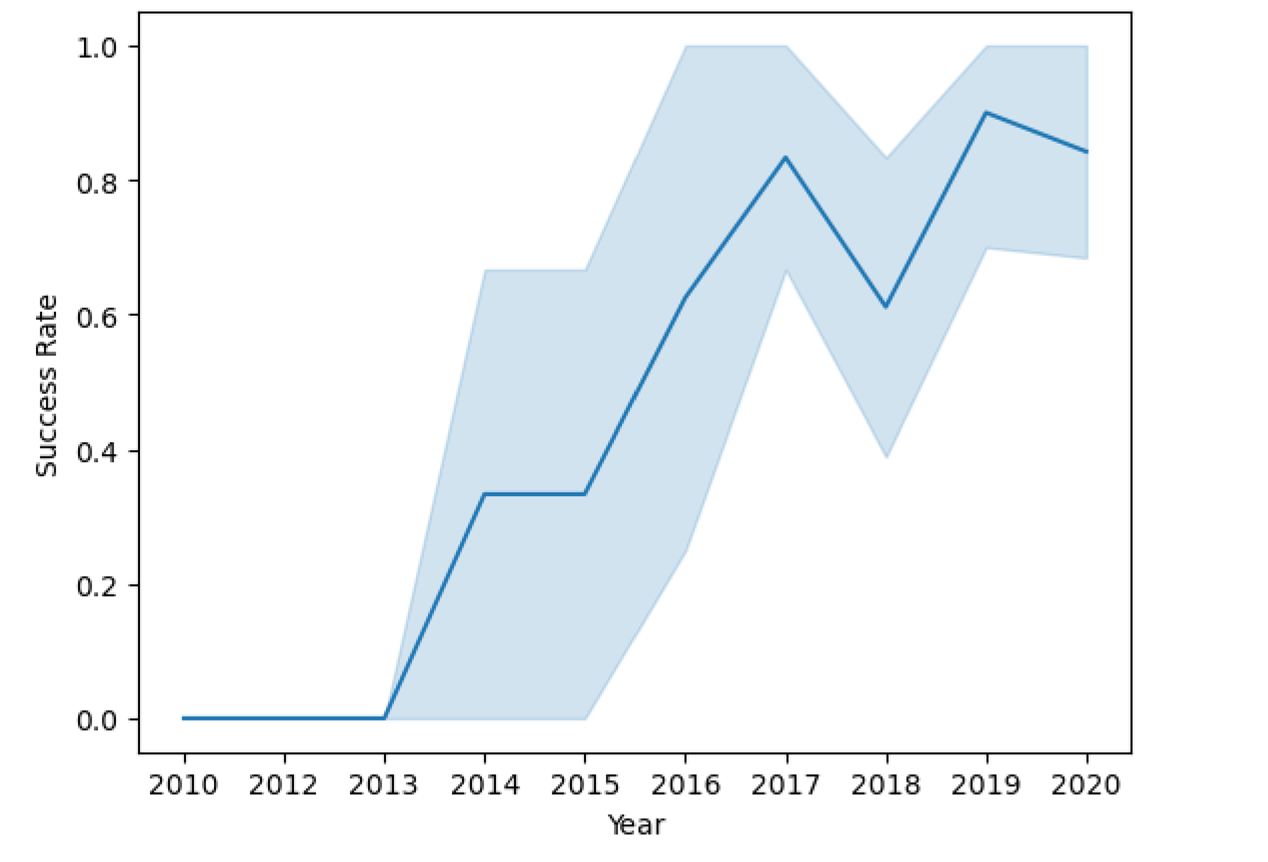
- Most of the orbits were used with a payload less than 10 tons.

- Heavy payloads have a mixed influence on GTO orbits and positive on LEO, POLAR and ISS orbit

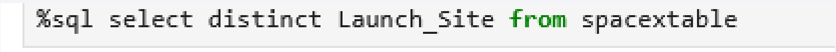


Launch Success Yearly Trend

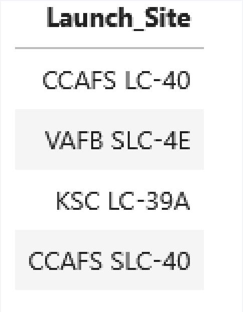
- There is an upward trend in a success rate since 2013 with dip between years 2017-2019.



All Launch Site Names



```
%sql select distinct Launch_Site from spacextable
```

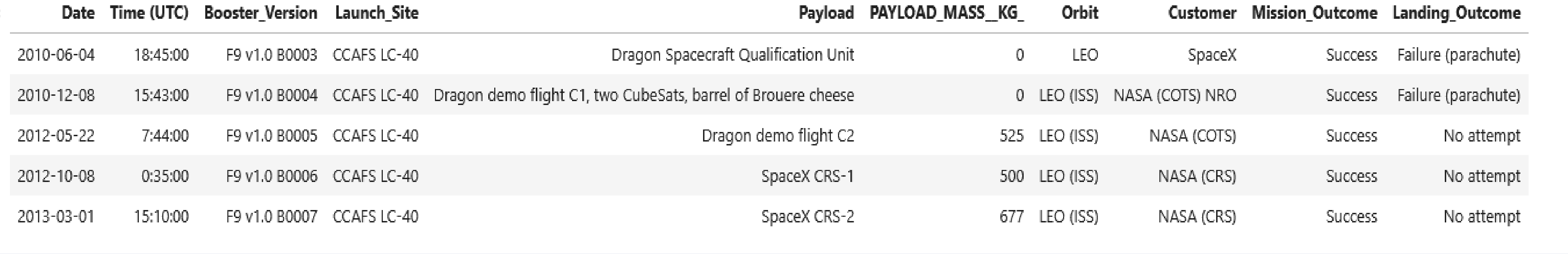


Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Explanation: The use of **DISTINCT** keyword in the query allows to remove duplicate entries for LAUNCH_SITE column

Launch Site Names Begin with 'CCA'

*select * from spacextable where Launch_Site like 'CCA%' limit 5;*

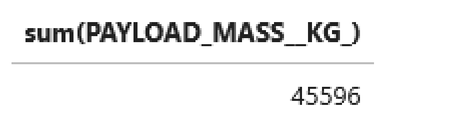


Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight 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

Explanation: SELECT gets data, * means all columns, **FROM** specifies the table, **WHERE** filters rows, **LIKE 'CCA%'** finds values starting with "CCA", and **LIMIT 5** returns only 5 results.

Total Payload Mass

```
select sum(PAYLOAD_MASS__KG_) from spacextable where customer = 'NASA (CRS)’;
```

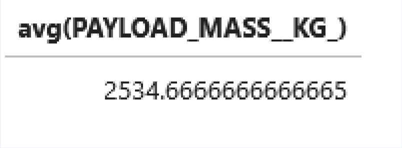


sum(PAYLOAD_MASS__KG_)
45596

Explanation: This SQL query calculates the total payload mass for a specific customer.
SUM(PAYLOAD_MASS__KG_) adds up all payload masses, **FROM spacextable** uses the SpaceX data table, **WHERE customer = 'NASA (CRS)'** filters rows where the customer is NASA (CRS).

Average Payload Mass by F9 v1.1

select avg(PAYLOAD_MASS__KG_) from spacextable where Booster_Version like 'F9 v1.1%'

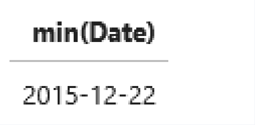


avg(PAYLOAD_MASS__KG_)
2534.6666666666665

Explanation: This query calculates the average payload mass for a specific booster version. **AVG(PAYLOAD_MASS__KG_)** computes the mean payload mass, **FROM spacextable** uses the data table, **WHERE Booster_Version LIKE** *'F9 v1.1%'* filters for booster versions that start with "F9 v1.1"

First Successful Ground Landing Date

select min(Date) from spacextable where Landing_Outcome = 'Success (ground pad)'

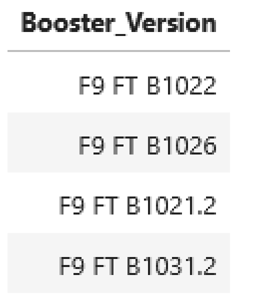


min(Date)
2015-12-22

Explanation: **MIN(Date)** returns the earliest date, **FROM spacextable** uses the data table, **WHERE Landing_Outcome = 'Success (ground pad)'** filters for successful landings on a ground pad.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
select Booster_Version from spacextable where Landing_Outcome = 'Success (drone ship)' and  PAYLOAD_MASS__KG_ between 4000 and 6000
```

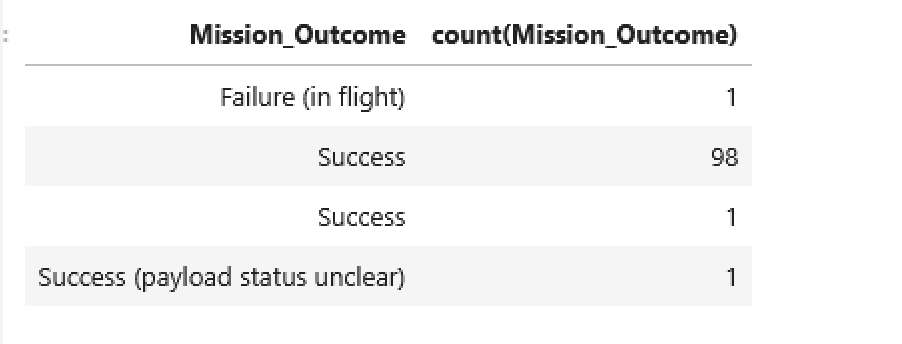


Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Explanation: **SELECT Booster_Version** retrieves the booster version, **FROM spacextable** uses the SpaceX data table, **WHERE Landing_Outcome = 'Success (drone ship)'** filters for successful drone ship landings, and **PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000** limits results to payloads between 4000 and 6000 kg.

Total Number of Successful and Failure Mission Outcomes

select Mission_Outcome, count(Mission_Outcome) from spacextable group by Mission_Outcome

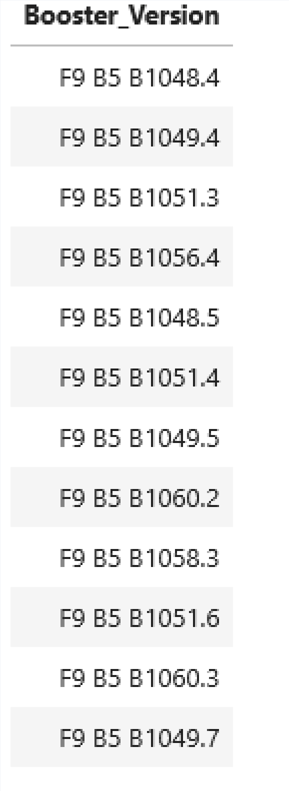


Mission_Outcome	count(Mission_Outcome)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Explanation: SELECT Mission_Outcome, COUNT(Mission_Outcome) gets each unique mission outcome and how many times it happened, **FROM spacextable** uses the data table, and **GROUP BY Mission_Outcome** groups the results by each outcome type.

Boosters Carried Maximum Payload

select distinct Booster_Version from spacextable where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from spacextable)

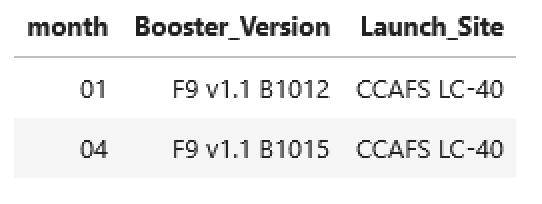


Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

Explanation: SELECT DISTINCT Booster_Version gets unique booster versions, **FROM spacextable** uses the SpaceX data table, and **WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM spacextable)** filters for the row(s) with the maximum payload mass.

2015 Launch Records

```
select substr(Date, 6,2) as month, Booster_Version, Launch_Site from spacextable where
Landing_Outcome = 'Failure (drone ship)' and substr(Date,0,5)='2015'
```

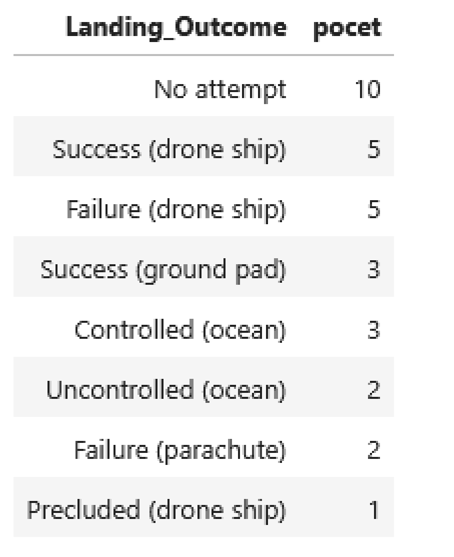


month	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

Explanation: SELECT substr(Date, 6, 2) as month extracts the month from the Date column, **Booster_Version** and **Launch_Site** show the respective booster and launch site details, **FROM spacextable** uses the data table, and **WHERE Landing_Outcome = 'Failure (drone ship)'** filters for failed drone ship landings. **substr(Date, 0, 5) = '2015'** limits the results to launches in the year 2015.

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

select Landing_Outcome, count(Landing_Outcome) as pocet from spacextable where date between '2010-06-04' and '2017-03-20' group by Landing_Outcome order by pocet desc



Landing_Outcome	pocet
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

Explanation: SELECT Landing_Outcome, COUNT(Landing_Outcome) AS pocet retrieves each landing outcome and the count of how often it occurred, **FROM spacextable** uses the data table, and **WHERE date BETWEEN '2010-06-04' AND '2017-03-20'** filters the results to launches that occurred within the specified date range. **GROUP BY Landing_Outcome** groups the results by the landing outcome, and **ORDER BY pocet DESC** sorts the results by the count in descending order.

Section 3

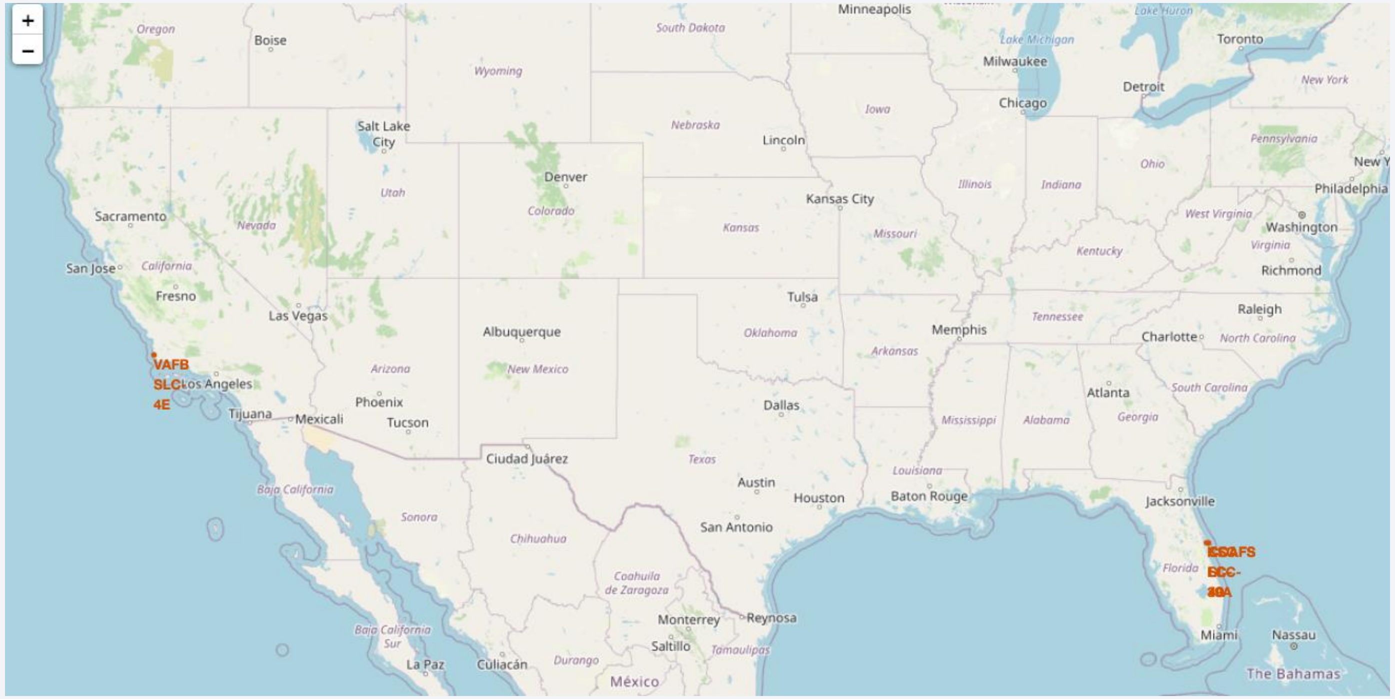
Launch Sites Proximities Analysis

Launch Sites Location on Open Maps

- All the Launch Sites are closes to the Equator and the coastlines

- Rockets launched near the Equator benefit from the inertia due to the rotation of the Earth on its own axis

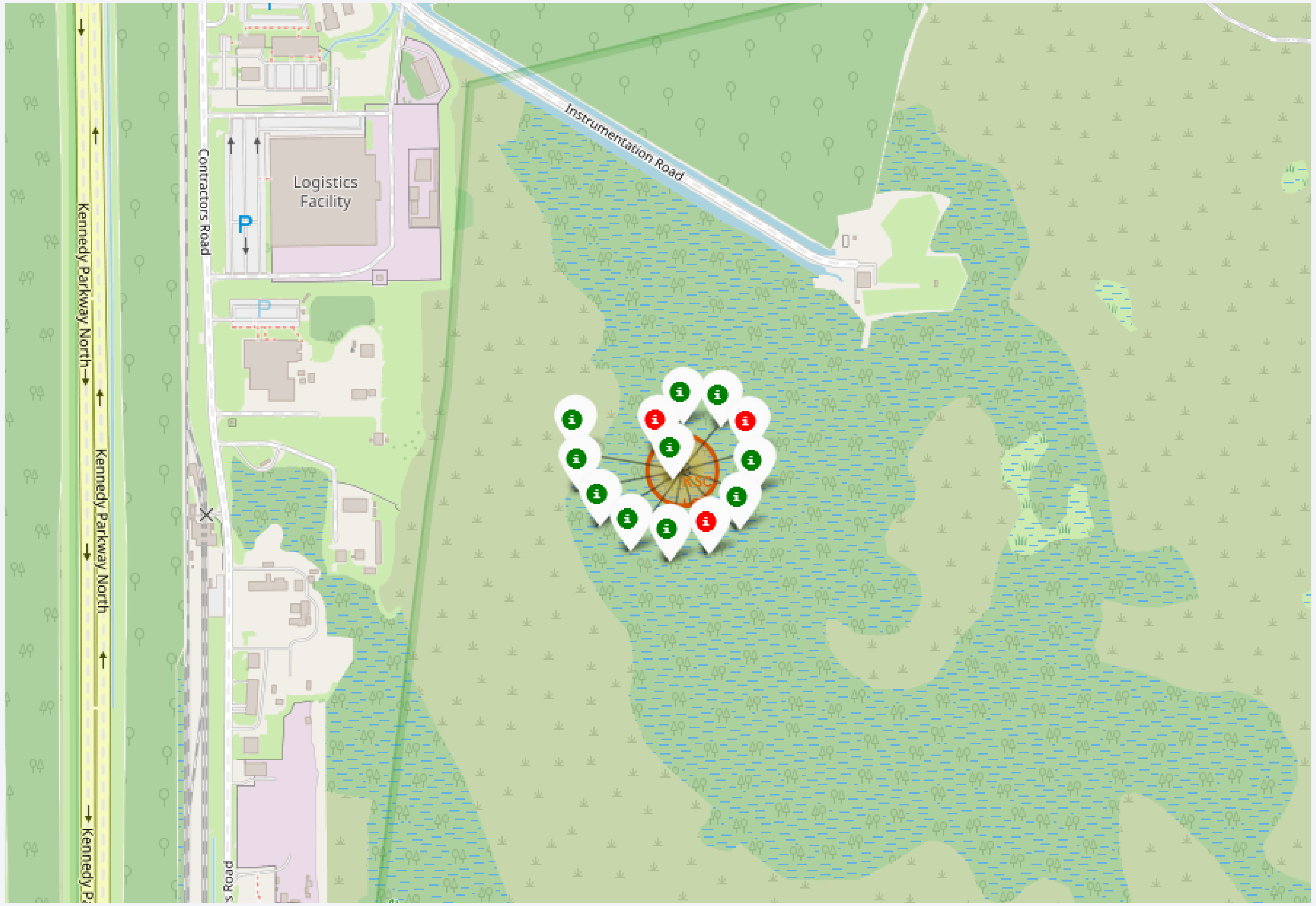
- Launch Sites near coastlines are preferred as the mission can be aborted with minimum damage in the ocean in case of an untoward incident



Color-labeled Launch Outcomes on the Map

- Colour-labeled markers allows us to visually identify which launch sites have relatively high success rates.
  - Green marker -> successful launch
  - Red marker -> successful launch

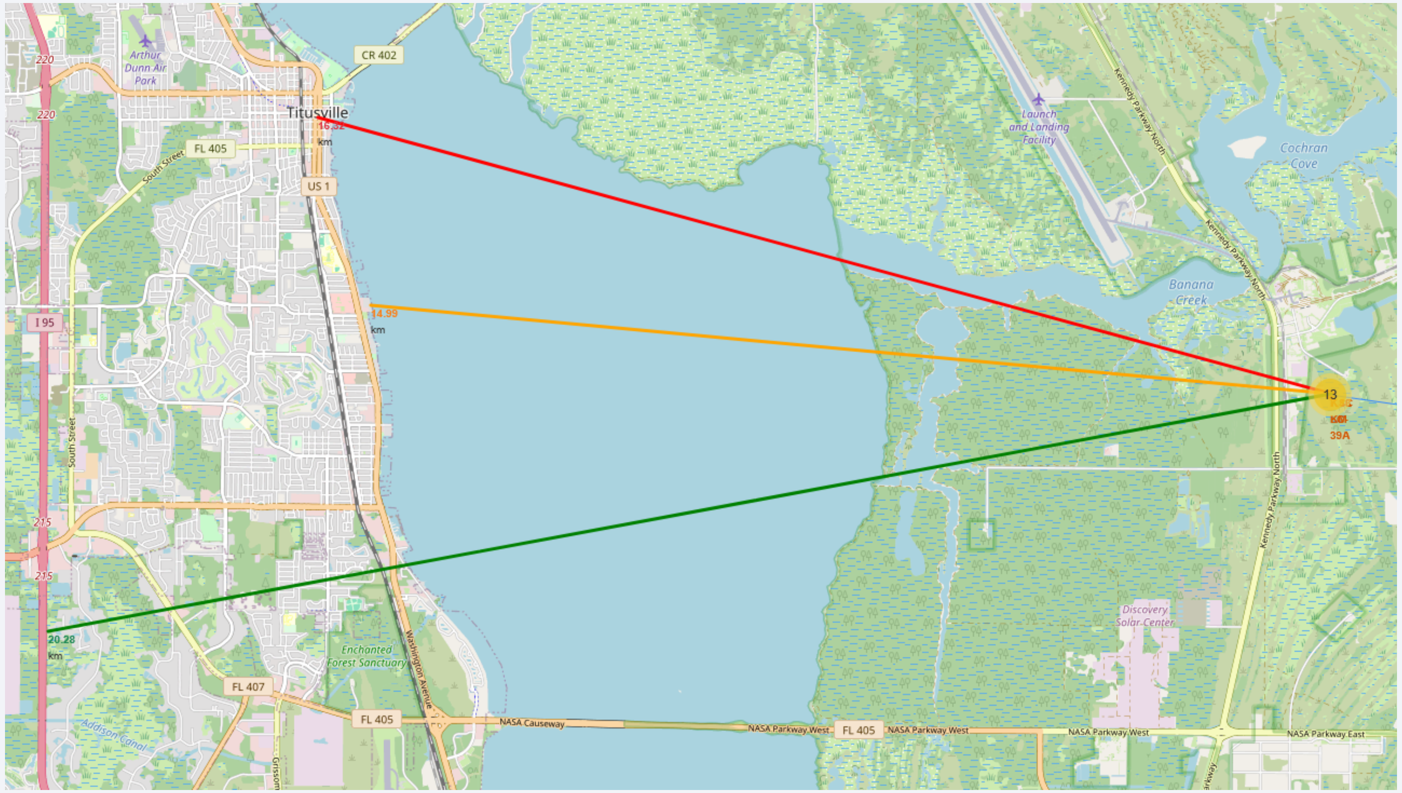
- KSC LC-39A has very high success rate.

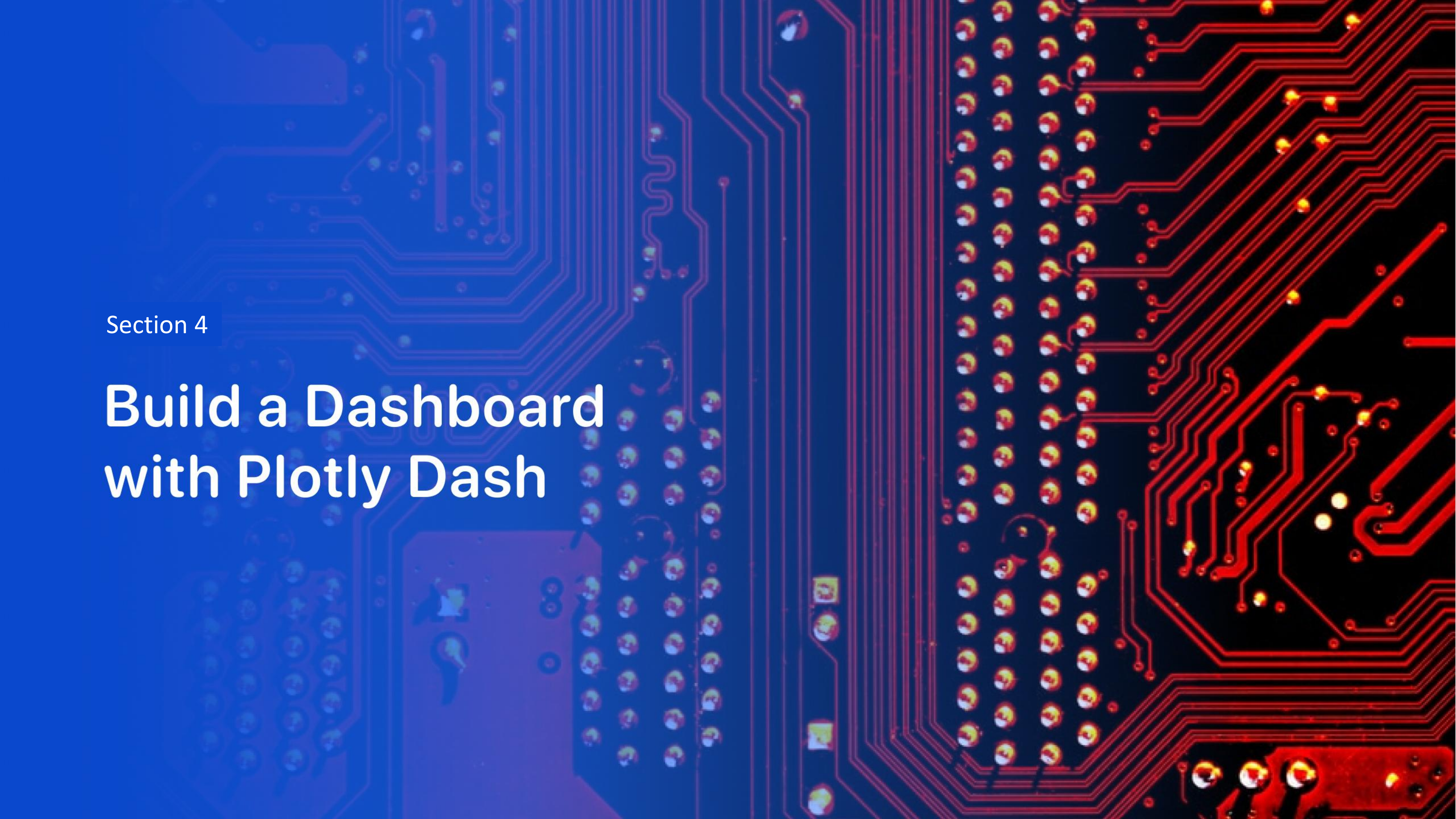


Distance from the launch site KSC LC-39A to selected proximities

- relatively close to railway (15.23 km)
- relatively close to highway (20.28 km)
- relatively close to coastline (14.99 km)

- Based on the findings, the launch site is relatively close to the populous areas





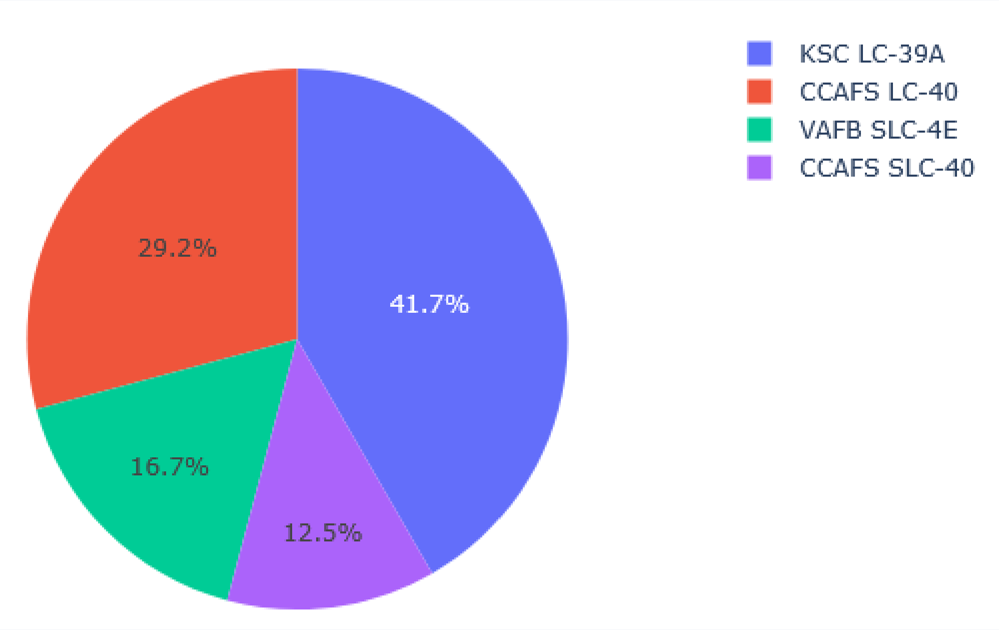
Section 4

Build a Dashboard with Plotly Dash

Total Success Launches for All Sites

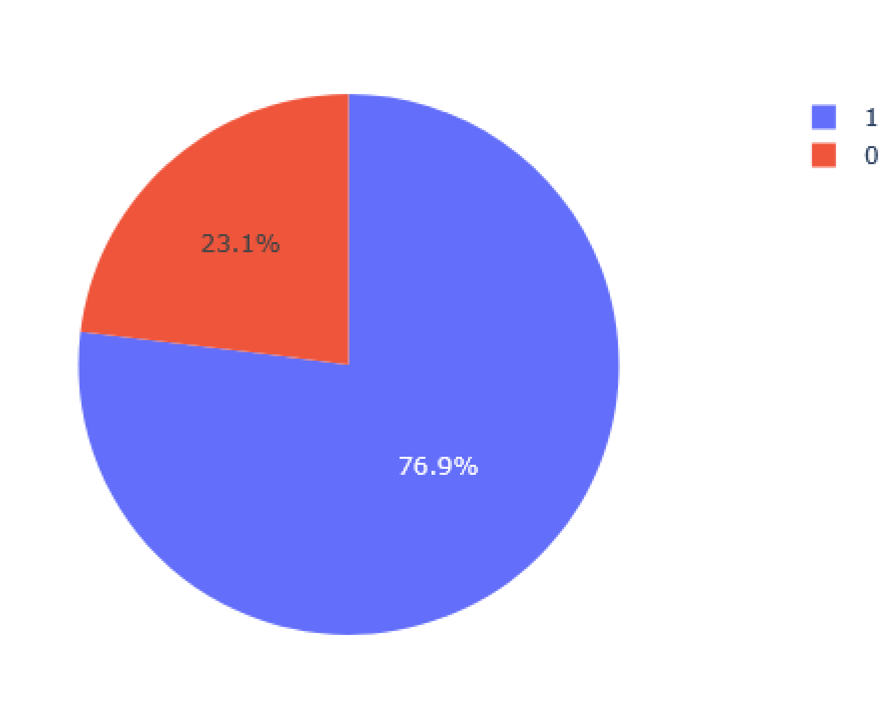
- KSC LC-39A is a launch site with highest portion of successful launches

- VAFB SLC-4E is a launch site with lowest number of successful launches



Launch Site with Highest Launch Success Ratio

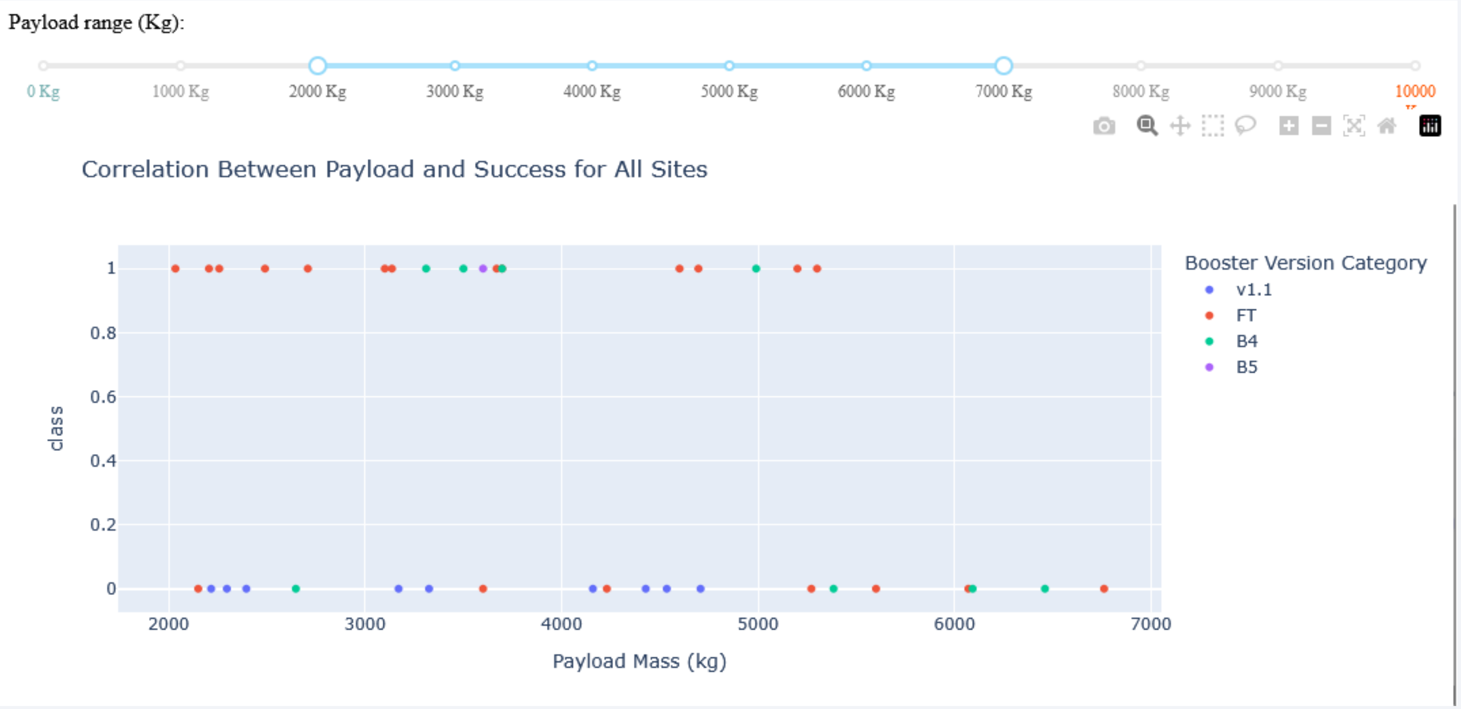
- 76.9 % launches from the site KSC LC-39A were successful



Payload vs. Launch Outcome scatter plot

Payload vs. Launch Outcome scatter plot for all sites with a payload range between 2000 and 7000 kilograms.

- On this payload range the Booster version with highest success rate is FT, whereas Booster version v1.1 scores n o successful landing

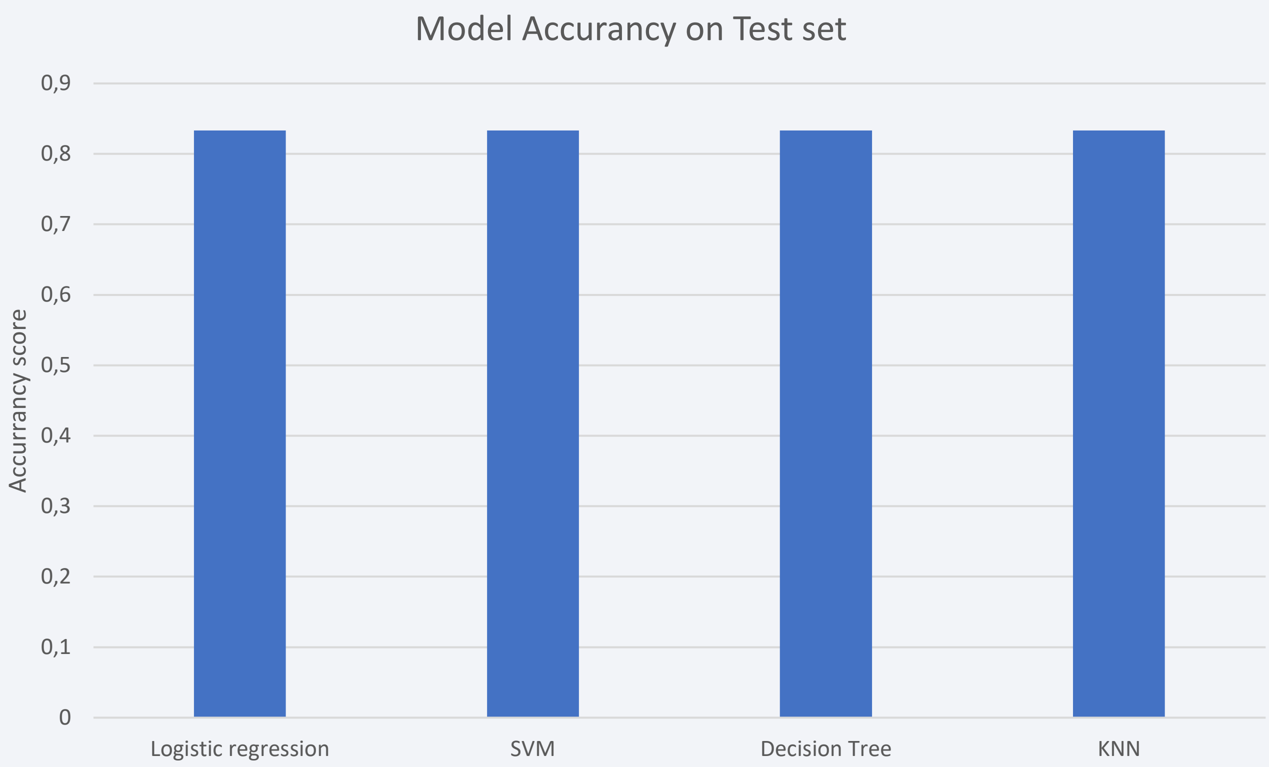


Section 5

Predictive Analysis (Classification)

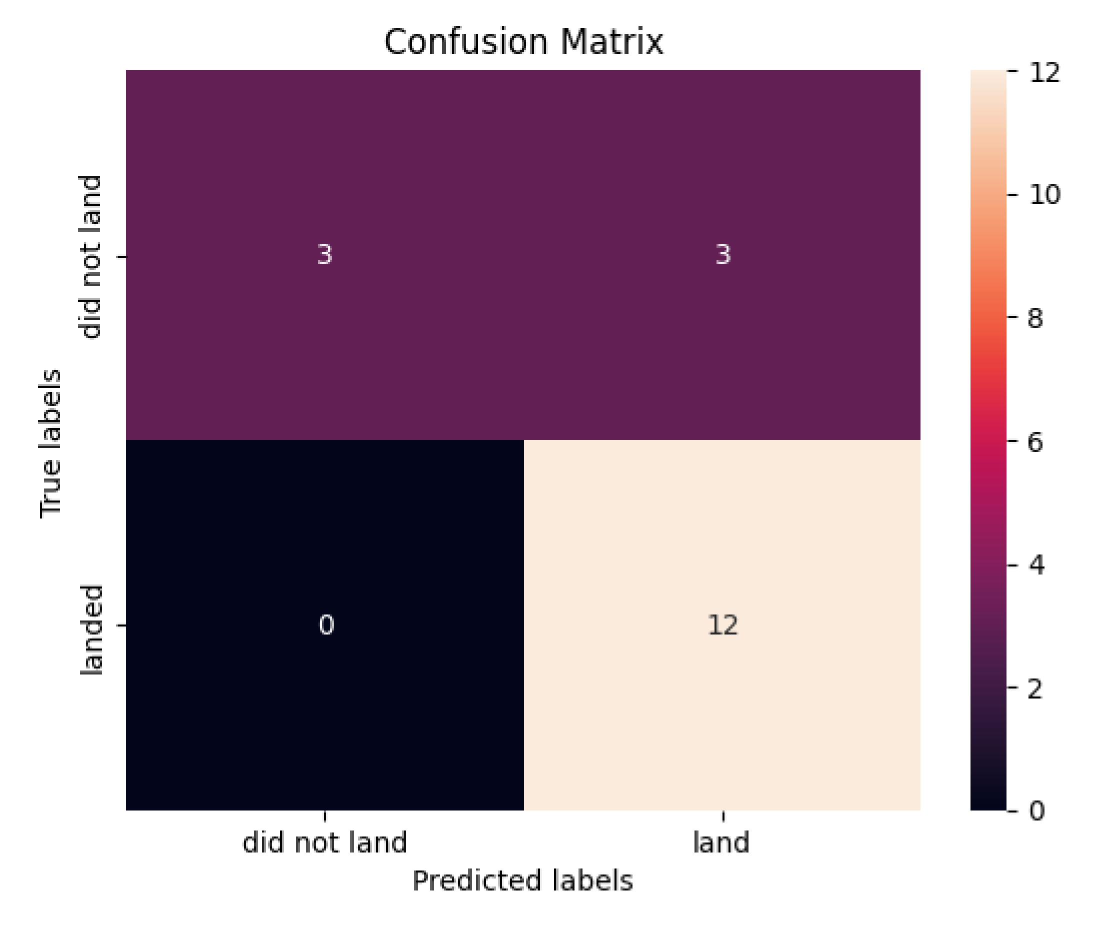
Classification Accuracy

All models have the same accuracy on test data due to small dataset.



Confusion Matrix

- A confusion matrix is a table that is used to define the performance of a classification algorithm. A confusion matrix visualizes and summarizes the performance of a classification algorithm
- The models misclassifies 3 unsuccessful landing attempts as successful.
- All models have the same confusion matrix on test data due to small dataset.



Conclusions

In this project, we try to predict if the first stage of a given Falcon 9 launch will land in order to determine the cost of a launch. Each feature of a Falcon 9 launch, such as its payload mass or orbit type, may affect the mission outcome in a certain way. Major findings are:

- Based on the EDA with visualization success rate improves over time among all launch sites
- Based on the EDA with interactive visual analytics we conclude that launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast
- Based on the EDA with SQL we conclude that KSC LC-39A has the highest success rate of the launches from all the sites
- Based on the test data accuracy score, there is no performance difference between tested models. There is a need for more data for better fine tuning and evaluation of models.

Appendix

Project source code:

https://github.com/mp-vue/applied-data-science-capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Thank you!