IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Andrew Lia

7th September 2025

Outline

EXECUTIVE SUMMARY

INTRODUCTION

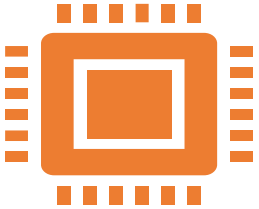
METHODOLOGY

RESULTS

CONCLUSION

APPENDIX

Executive Summary



Summary of methodologies

SpaceX data was collected through API requests and web scraping.

The data was preprocessed, including feature selection, creating, and null handling.

Exploratory Data Analysis was performed using standard charts and SQL.

Further analysis was done using interactive maps and dashboards.

Machine Learning models were trained on the data to predict the landing outcome.

Summary of all results

SpaceX made massive improvements to their rate of successful landings between 2010-2017.

The decline in progress after 2017 indicates variables out of the control of SpaceX, or potentially unknown to SpaceX preventing further improvement.

It is also possible that a certain limit of progress has been met, preventing more substantial improvement until further discoveries, research, testing, etc. are concluded.

Introduction

SpaceX is one of the key players in the modern-day space race to make special travel safe, predictable, and generate revenue

To maximize revenue, the more parts that can be reused from previous launches, the better

One of the vital, and expensive, parts of a rocket launch is the first stage

If we can determine the causes for the highest chance of the first stage to land successfully and be recovered, we can cut a dramatic amount of costs from future launches

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API
  - Web scraping Wikipedia
- Perform data wrangling:
  - Label existing data with a binary (true/false) feature to use as target
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models:
  - KNN, SVM, Logistic Regression, and Decision Tree
  - Grid Search used to find best hyperparameters for each model

Data Collection

Data was collected in two primary ways:

SpaceX API requests.

Web scraping a Wikipedia article with extensive Space X flight data.

The API was used first, and a Pandas Dataframe was made by joining the results of the different endpoints based on relevant shared data (e.g. flight number).

After that, web scraping was performed by selecting the desired table and extracting the text from the otherwise unneeded symbols, links, and nested HTML tags.

Data Collection – SpaceX API

- GitHub URL: https://github.com/Alias012/IBM-DataScience-Capstone/blob/master/data-collection-api.ipynb

1. Call API for IDs
2. Filter and Edit Data
3. Use IDs To Obtain Detailed Info
   - Booster Version
   - Launch Site
   - Payload Data
   - Core Data
4. Filter New Data
5. Handle Missing Values

Data Collection - Scraping

- GitHub URL: https://github.com/Alias012/IBM-DataScience-Capstone/blob/master/data-collection-webscraping.ipynb

1. Request Page Data
2. Initialize BeautifulSoup Object With Page Data
3. Find All Tables and Select The Desired One
4. Extract Text From Header Row
5. Remove Unneeded Columns
6. Extract Data From Table and Fill Column Lists

Data Wrangling

- GitHub URL: https://github.com/Alias012/IBM-DataScience-Capstone/blob/master/data-wrangling.ipynb

01

Calculate Percentage of Missing Data

02

View Value Counts of Relevant Data

- Launch Site
- Orbit
- Outcome

03

Assign Binary (Success/Failure) Label Based On Outcome

04

Calculate Success Rate

EDA with Data Visualization

- GitHub URL: https://github.com/Alias012/IBM-DataScience-Capstone/blob/master/eda-dataviz.ipynb
- Used Charts:
  - Scatter plot of Flight Number vs Payload Mass; useful to see the change of the chosen payload mass over time and how successful the flights of certain masses were.
  - Category plot of Flight Number vs Launch Site; useful to see the scale to which each launch site has been used and how successful it is.
  - Category plot of Payload Mass vs Launch Site; useful to see the distribution of choice of launch site depending on Payload Mass as well as how successful they were.
  - Bar plot of Orbit vs Success Rate; useful to see the rate of success per each orbit type.
  - Category plot of Flight Number vs Orbit; useful to see the change of the chosen orbit type over time and how successful they the flights of certain orbits were.
  - Category plot of Payload Mass vs Orbit; useful to see the choice of orbit type depending on the payload mass, and how successful those choices were.
  - Line plot of Date vs Success Rate; useful to see the change of the rate of success over time.

EDA with SQL

- GitHub URL: https://github.com/Alias012/IBM-DataScience-Capstone/blob/master/eda-sql.ipynb
- SQL queries performed:
  - List of launch sites.
  - First 5 flights that were launched from a site starting with “CCA”.
  - Total payload mass sent by NASA (CRS).
  - Average payload mass sent by flights that used booster version F9 v1.1.
  - First date that a successful launch occurred.
  - Booster versions used on flights that successfully landed on a drone ship and whose payload mass was between 4000 to 6000.
  - Count total number of successes and failures.
  - Booster versions used on flights that carried the highest payload mass.
  - The months, booster versions used and launch sites used where the flight failed to land on a drone ship in 2015.
  - The ordered count of landing outcomes.

Build an Interactive Map with Folium

- GitHub URL: https://github.com/Alias012/IBM-DataScience-Capstone/blob/master/analytics-folium-sns-plt.ipynb
- Map objects:
  - Circle around each launch site; useful to emphasize the locations of interest.
  - Markers at the location of each launch; useful to see the distribution of launches between the sites.
  - Marker cluster to organize the launch markers; useful to organize the vast amount of markers.
  - Latitude and Longitude of current mouse location; useful for finding the specific location of points of interest on the map.
  - Marker of the closest coast-line to a launch site, with a line between them; useful to emphasize the (short) distance between the two.
  - Marker of the closest road to a launch site, with a line between them; useful to emphasize the distance between the two.

Build a Dashboard with Plotly Dash

- GitHub URL: https://github.com/Alias012/IBM-DataScience-Capstone/blob/master/analytics-dash-express.py
- Dashboard elements:
  - Dropdown of all Launch Sites; useful to let user choose what site they are interested in.
  - Pie chart (Success vs Failure if launch site chosen, distribution of Success across all launch sites if no site chosen); useful to give the user the visualization of the specific success data they are interested in.
  - Range slider of Payload Mass; useful to let the user choose payload mass they are interested in.
  - Scatter plot of Payload Mass vs Success; useful to give the user the visualization of the specific payload mass vs success data they are interested in.

Predictive Analysis (Classification)

- GitHub URL: https://github.com/Alias012/IBM-DataScience-Capstone/blob/master/machine-learning-classification.ipynb

01

Prepare and Split Data

02

Initialize ML Models

- Logistic Regression
- Support Vector Machine
- Decision Tree
- K Nearest Neighbors

03

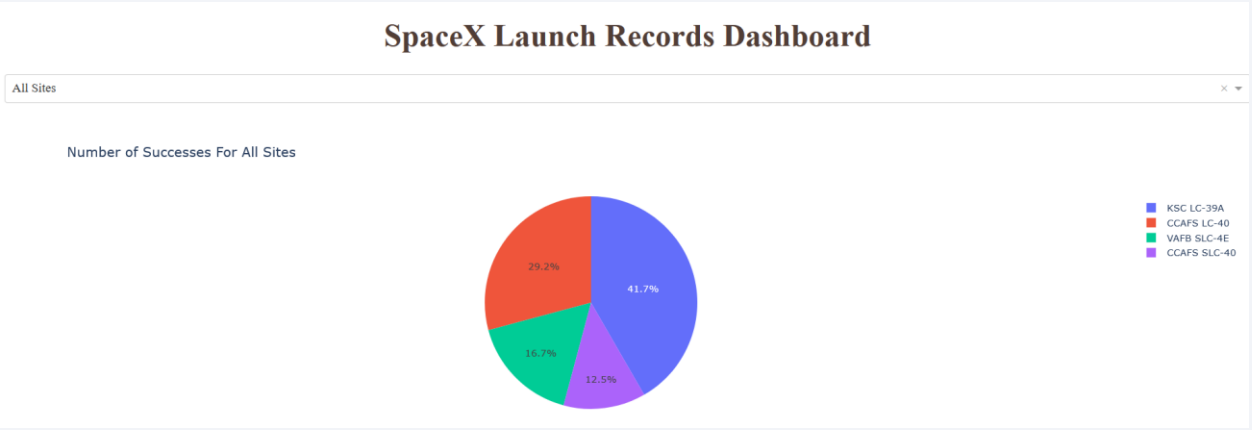
Grid search with many hyperparameters to find the best

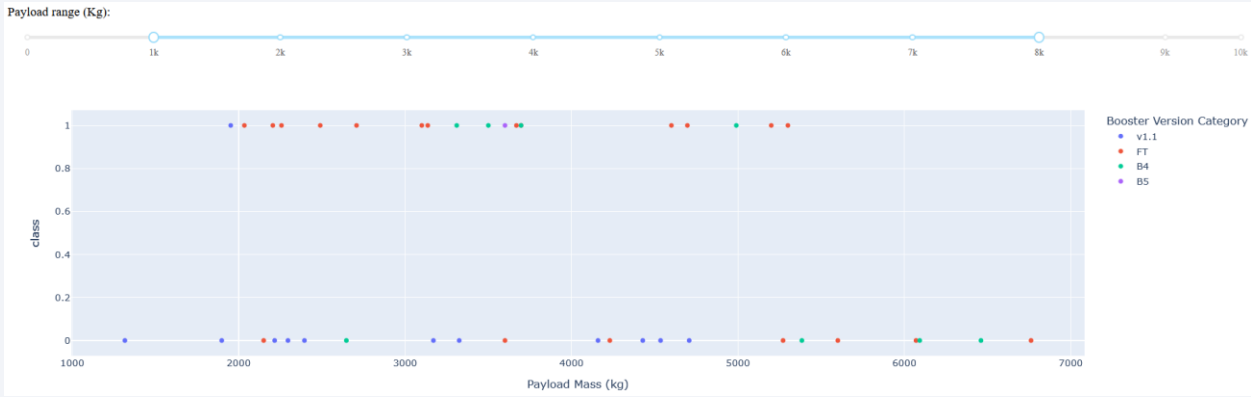
04

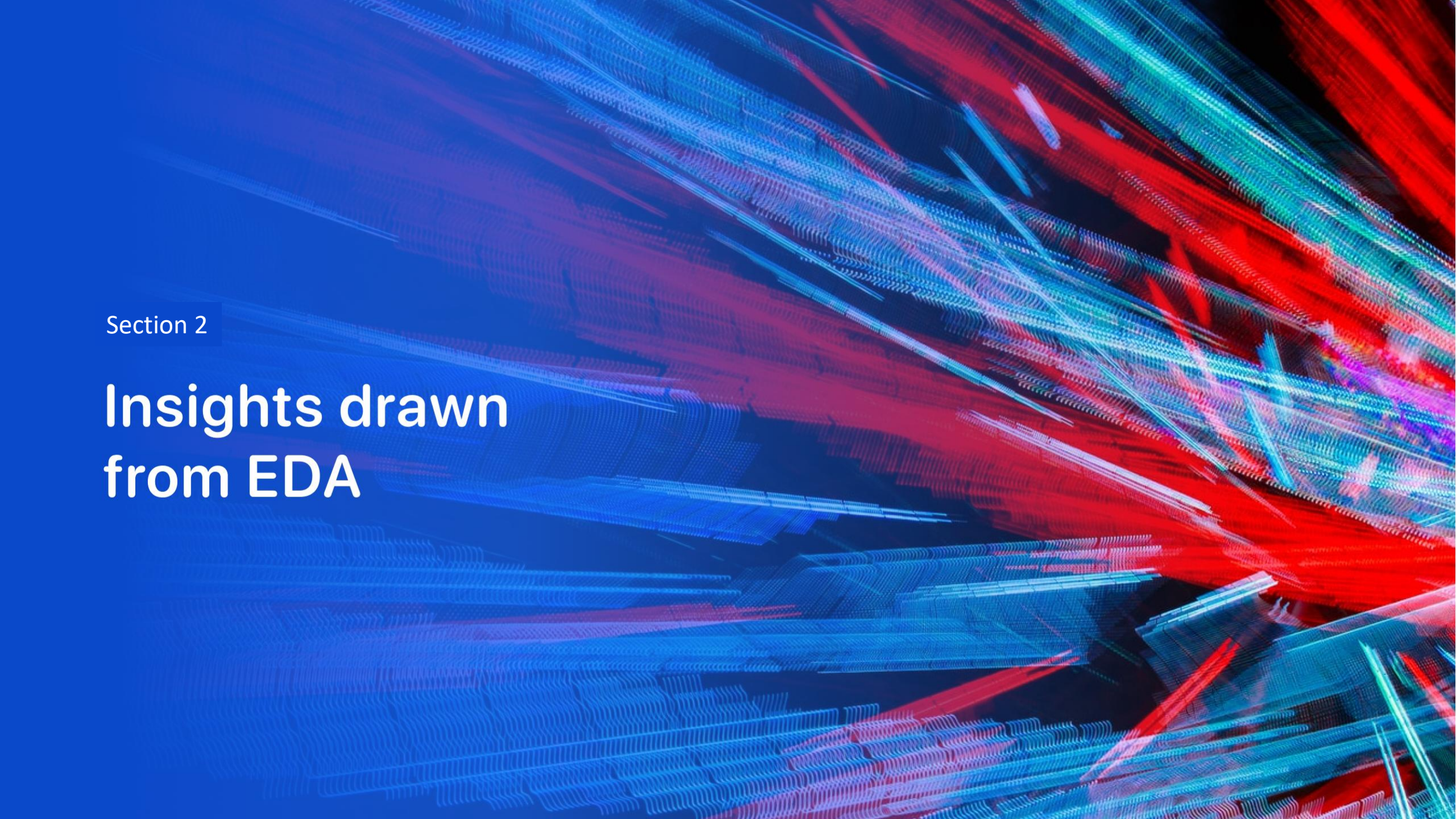
Evaluate accuracies of all models and find the best

Results

- EDA Results: clear correlations between certain features, such as an increase of successful landings over time.
- Predictive Analysis Results: best predictive models achieved an 83% accuracy.
- Plotly Dashboard:



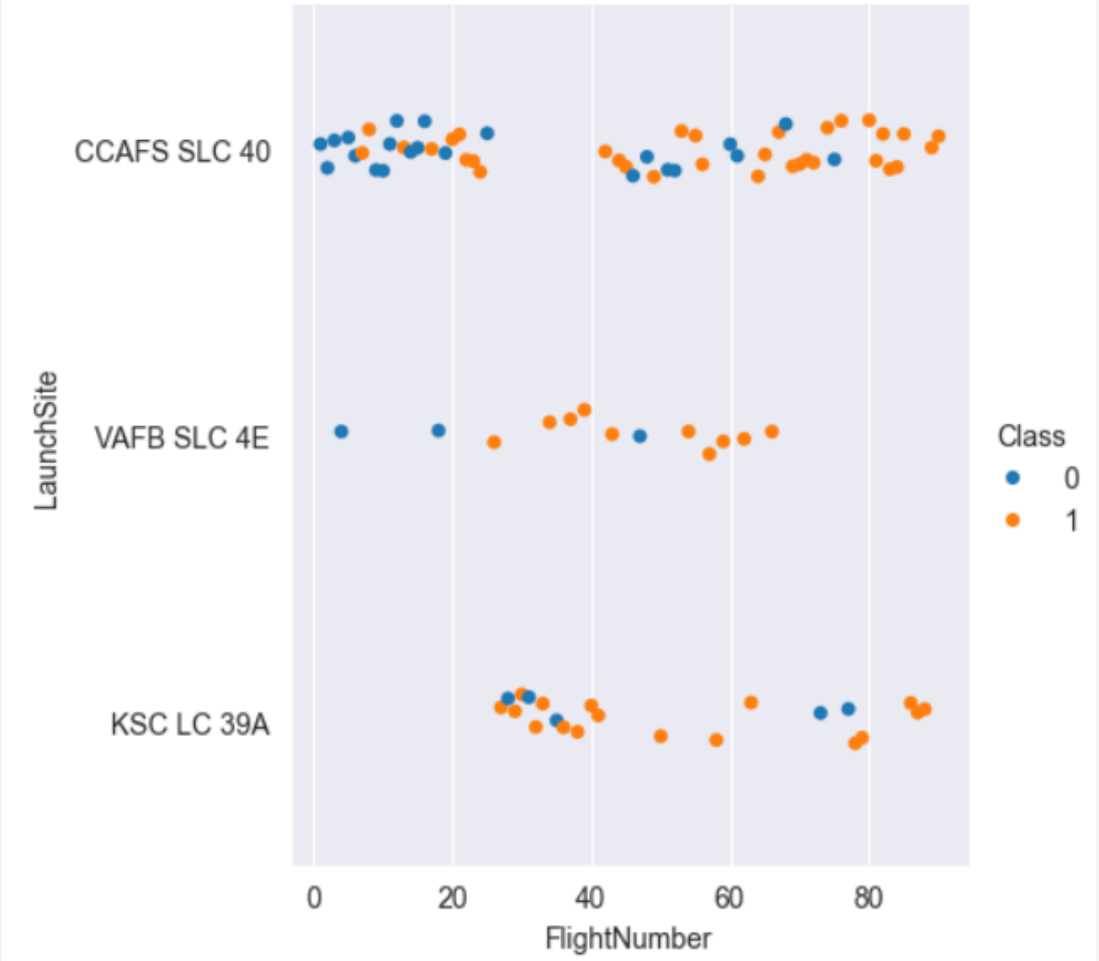



Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

Key Takeaways:

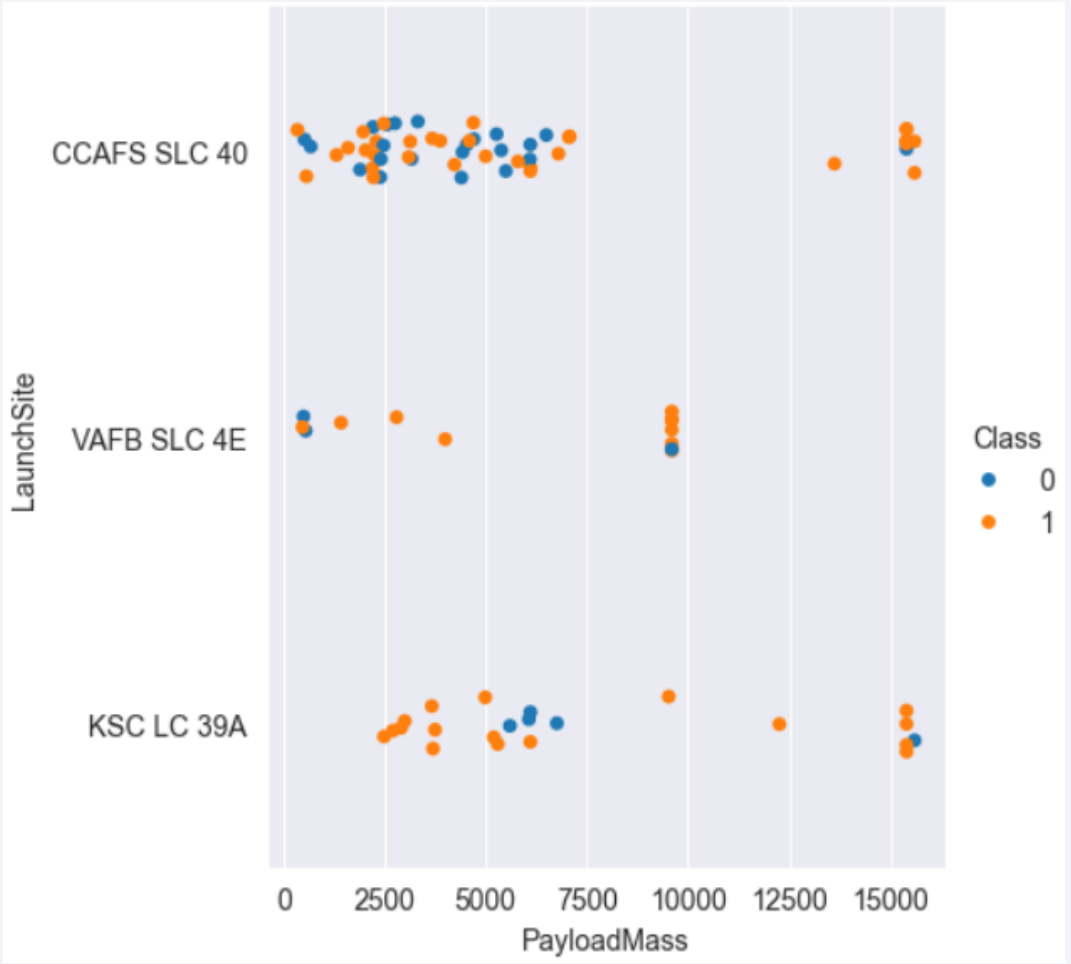
- The earlier flights were more likely to fail.
- CCAFS SLC 40 has been chosen to use as a launch site far more than KSC LC 39A, which as been chosen more than VAFB SLC 4E.
- The most recent flights from all launch sites were successful.



Payload vs. Launch Site

Key Takeaways:

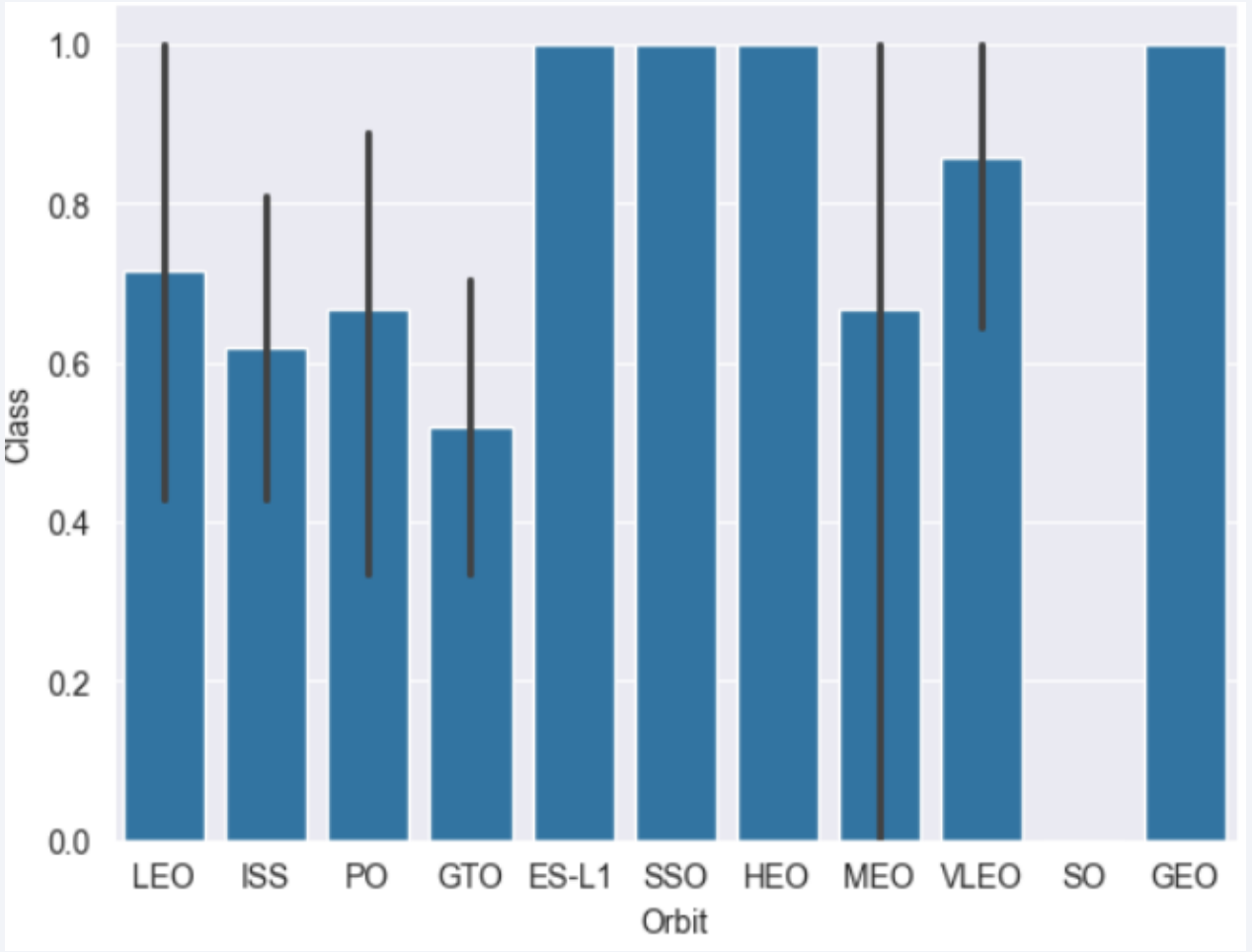
- VAFB SLC 4E is not chosen for flights with a very high payload mass.
- Flights launched from CCAFS SLC 40 with a low payload mass have a relatively low success rate.
- Flights tend to have relatively low or relatively high payload masses, with little middle ground.



Success Rate vs. Orbit Type

Key Takeaways:

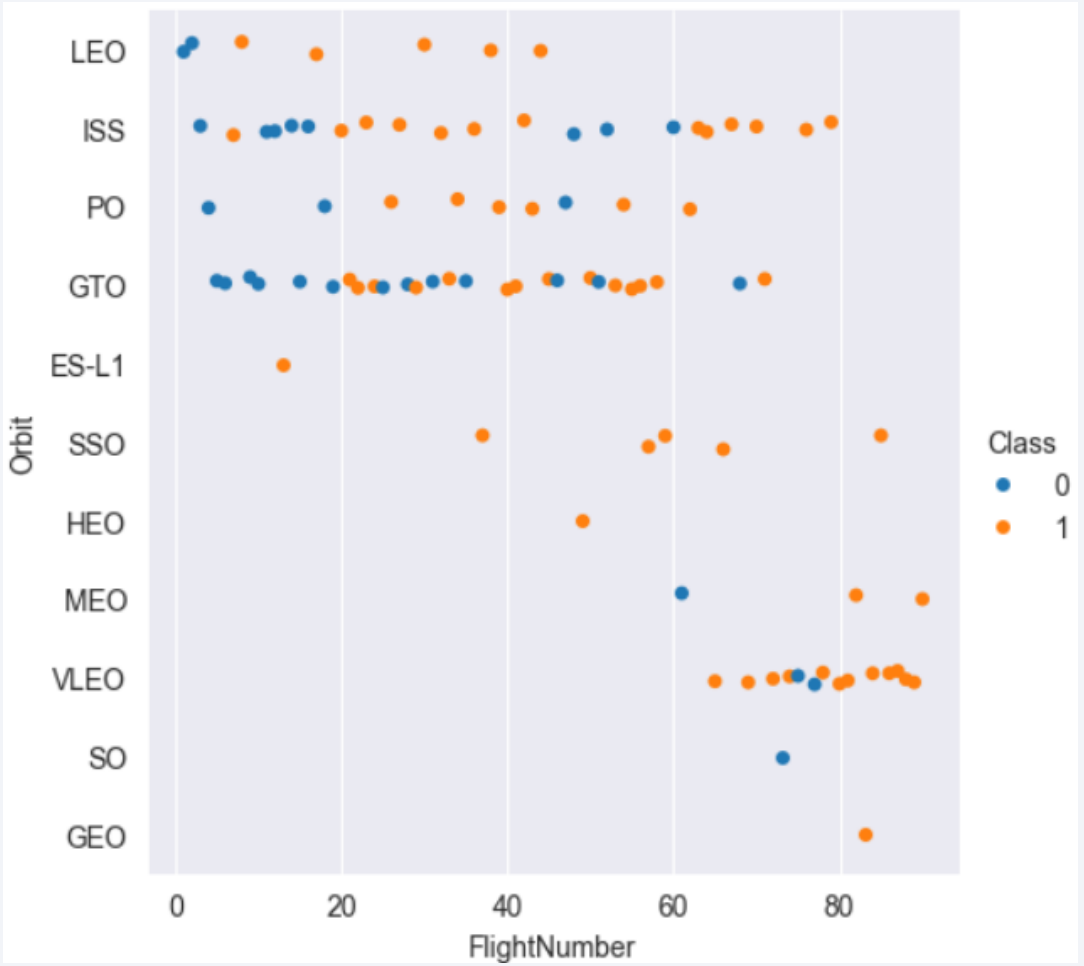
- Four orbit types have a 100% success rate.
- One orbit type, SO, has a 0% success rate.
- The remaining orbit types are in a relative middle ground of success rate, with VLEO being higher than the rest.



Flight Number vs. Orbit Type

Key Takeaways:

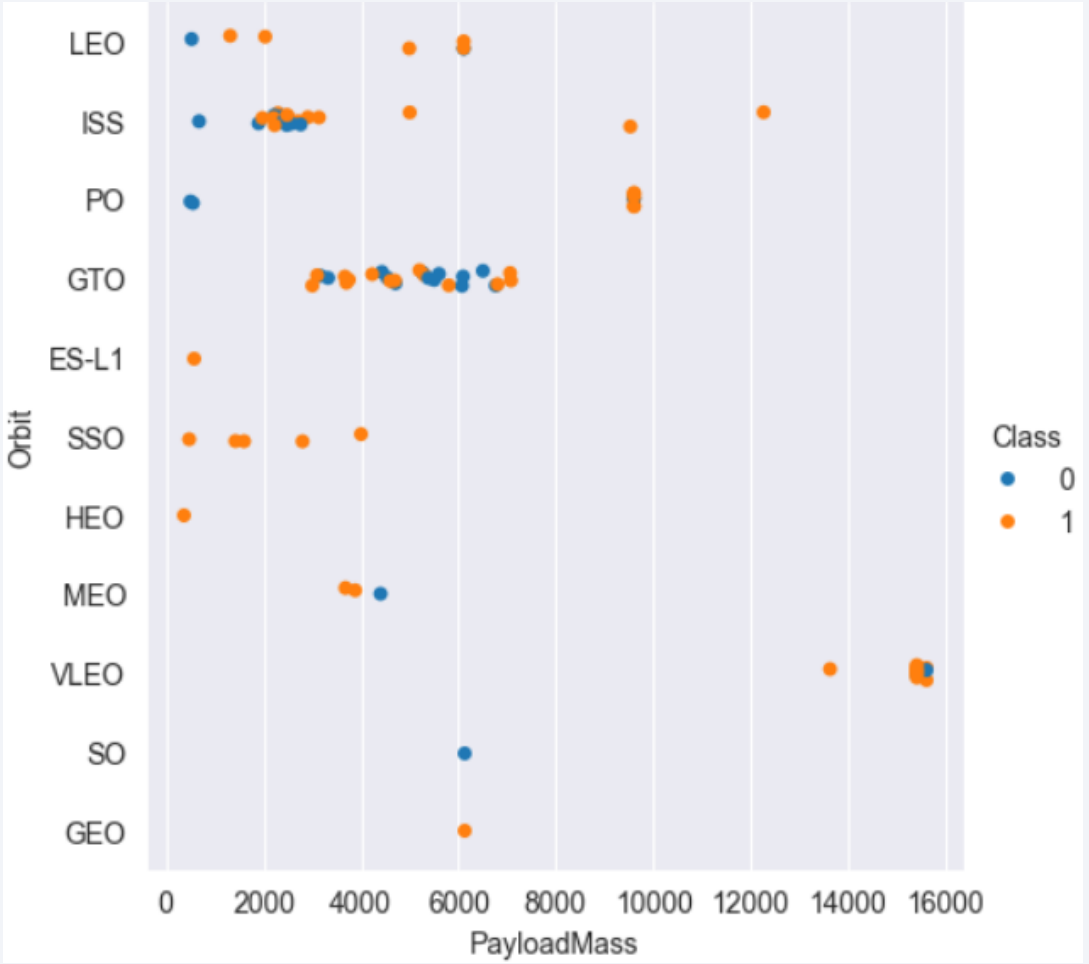
- Earlier flights had a lower success rate (already discovered).
- As time progressed, more orbit types were attempted, and these orbits generally have a higher success rate than the originally attempted ones.



Payload vs. Orbit Type

Key Takeaways:

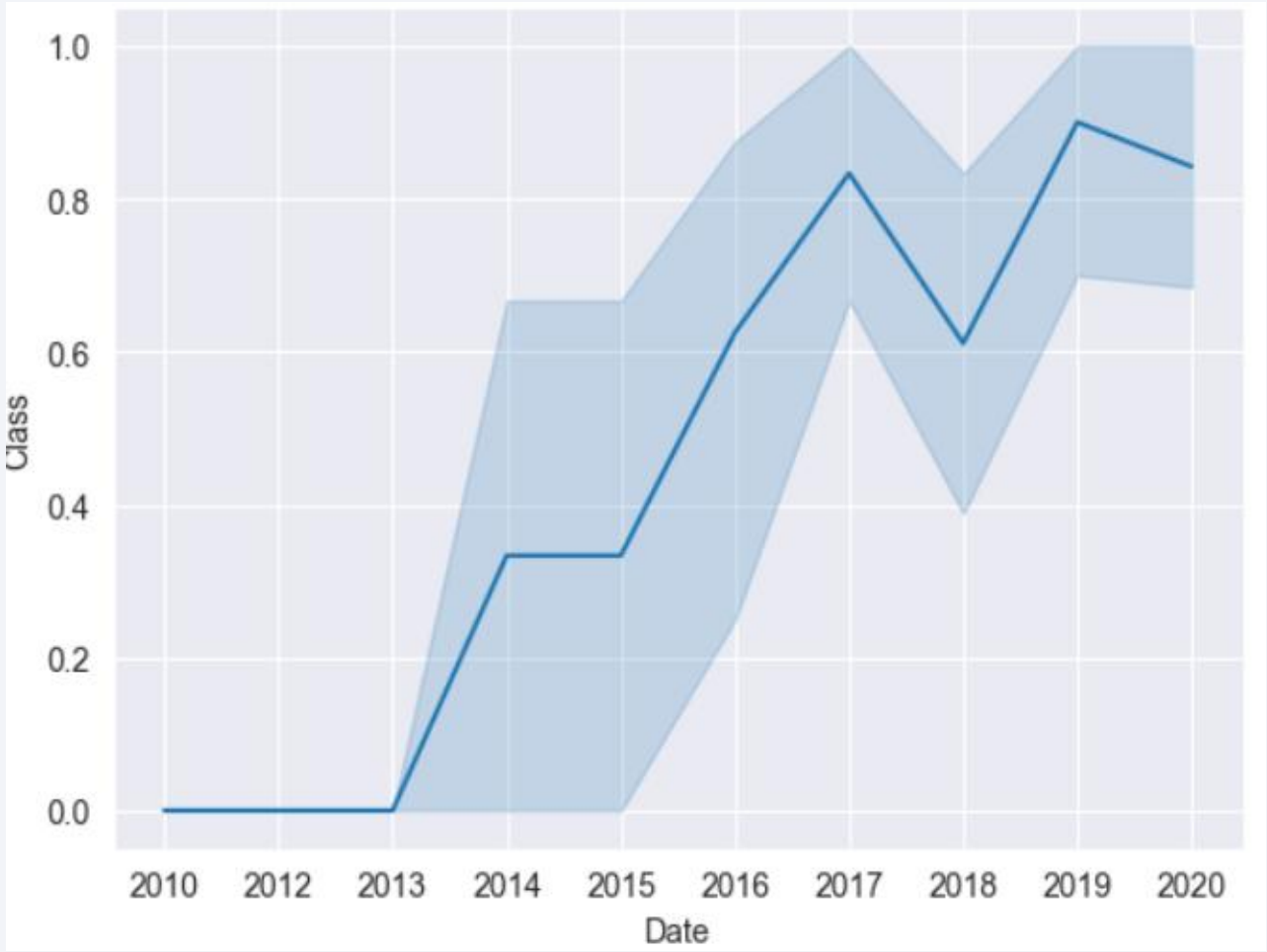
- Flights with a low to mid-payload mass have multiple successful orbit types.
- Flights with a very high payload mass have only been attempted with a VLEO orbit and have been relatively successful.



Launch Success Yearly Trend

Key Takeaways:

- For the first three years of flights, SpaceX had zero successes.
- Afterwards, SpaceX's success rate continued to climb until 2017.
- In 2017, a steep decline in success was noticed, which was rectified in 2018, and a new peak was reached in 2019.
- After 2019, a steady decline in success was noticed.



All Launch Site Names

Explanation of Query:

- All launch sites with a restriction of uniqueness.

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

Launch Site Names Begin with 'CCA'

Explanation of Query:

- All records, with a restriction that the launch site must start with “CCA”. A limit of 5 was applied.

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outc
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (paracl
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 (paracl
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No atte
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No atte
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No atte

Total Payload Mass

Explanation of Query:

- The sum of all payload masses where the name of the customer was “Nasa (CRS)”.

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

Explanation of Query:

- The average of all payload masses where the name of the booster version was “F9 v1.1”.

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

Explanation of Query:

- The oldest date where the type of landing outcome started with “Success”.

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Explanation of Query:

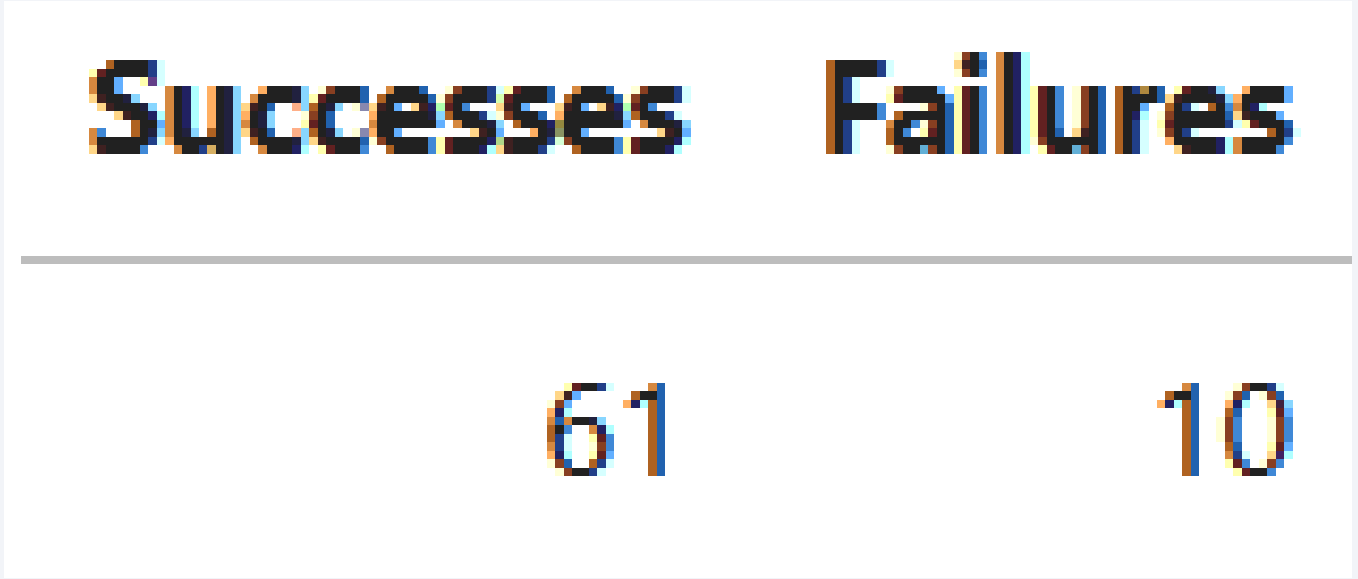
- The booster versions where the landing outcome was “Success (drone ship)”, and a payload mass between 4000 and 6000. A restriction of uniqueness was applied.

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

Total Number of Successful and Failure Mission Outcomes

Explanation of Query:

- Successes: the sum of all successes when landing outcomes that start with “Success” were mapped to 1 and otherwise were mapped to 0.
- Failures: the sum of all failures when landing outcomes that start with “Failure” were mapped to 1 and otherwise were mapped to 0.



Successes	Failures
61	10

Boosters Carried Maximum Payload

Explanation of Query:

- The booster versions that carried the maximum payload mass in the dataset (found through subquery). A restriction of uniquness was applied.

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

Explanation of Query:

- The month (in numerical format), landing outcome, booster version, launch site from flights in 2015 that failed on a drone ship.

substr(Date, 6,2)	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

Explanation of Query:

- The landing outcomes with total occurrences of that type that happened between 2010-06-04 and 2017-03-20 in order from most to least.

Landing_Outcome	Rank
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

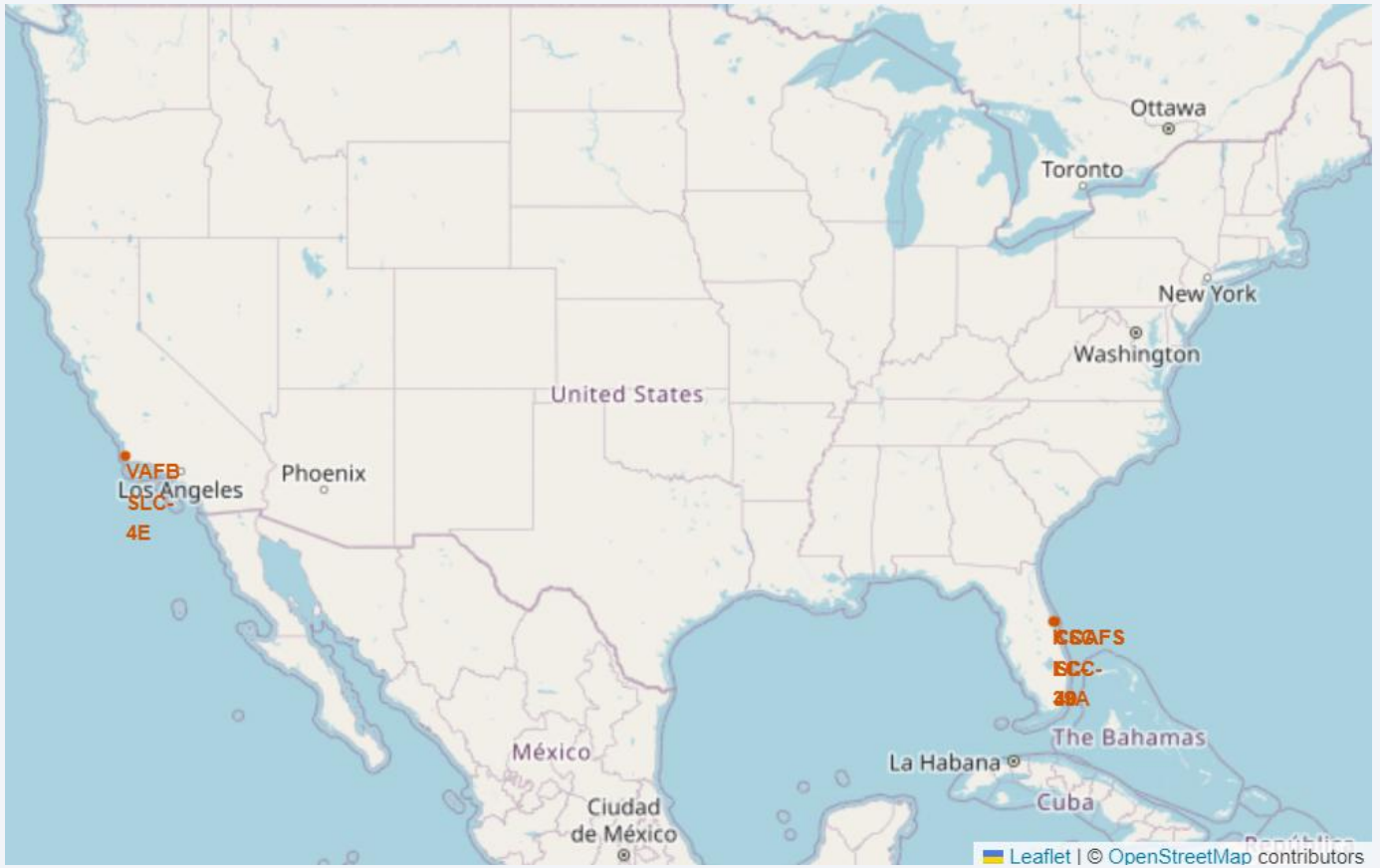
Section 3

Launch Sites Proximities Analysis

Pinpoints of Launch Locations

Explanation of Map:

- The SpaceX launch locations are pin-pointed on the map. One is in Southwest California, three are in East Florida.



Color-Coded Markers of Launches

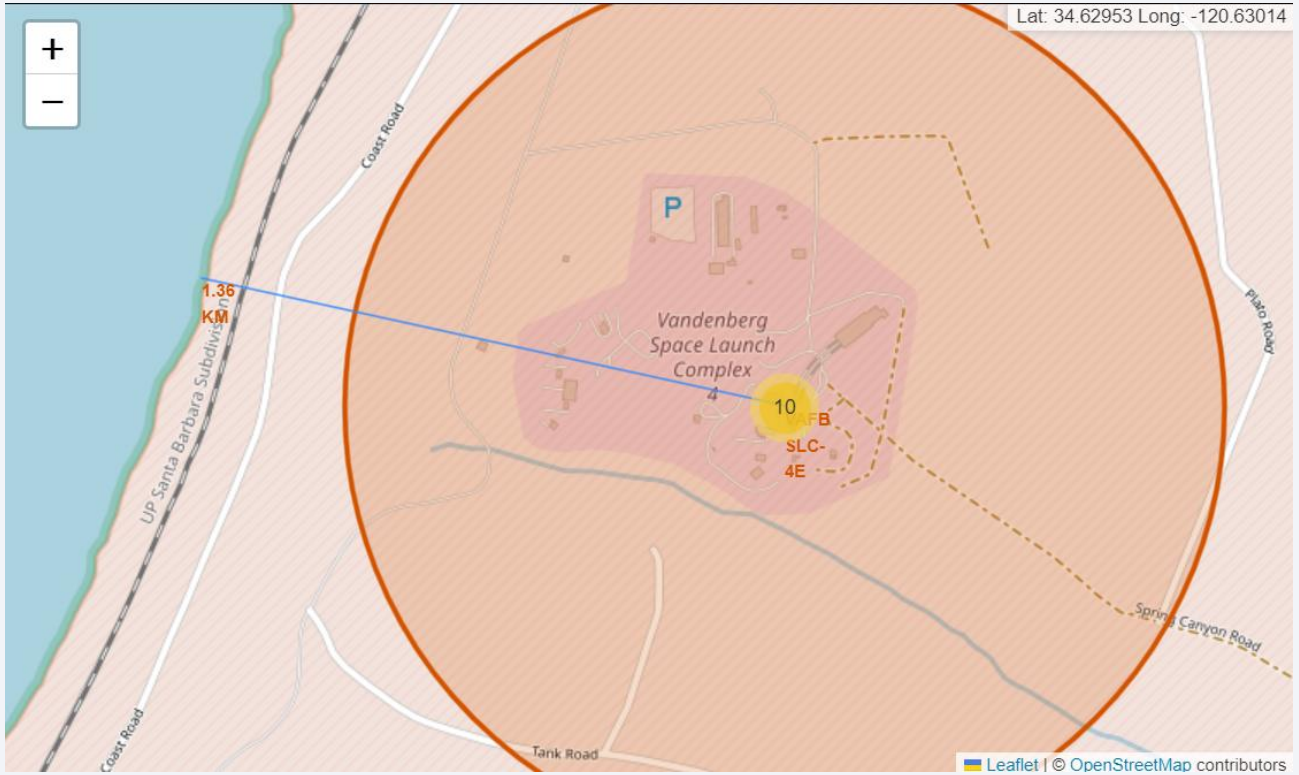
Explanation of Map:

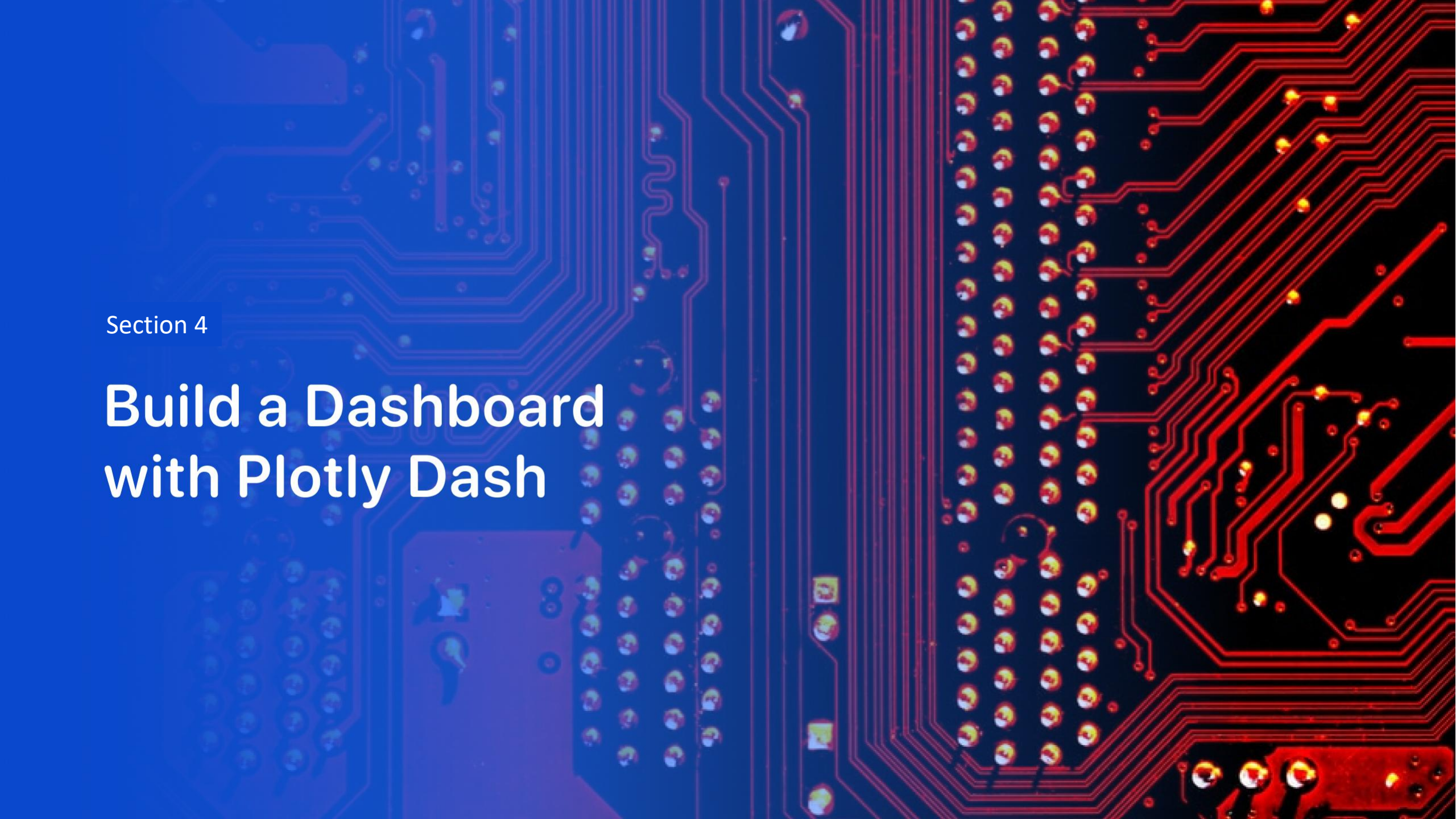
- The markers each indicate a launch at a given launch site. A green marker indicates a launch with a successful landing, whereas a red marker indicates a failed landing.

Distance from Launch Site to Nearest Coastline

Explanation of Map:

- The red circle indicates the radius around a launch site in Southern California. The blue line indicates the direction towards the nearest coastline; at the end of this line there is written “1.36 KM”, which is the distance between the launch site and the end of the line.

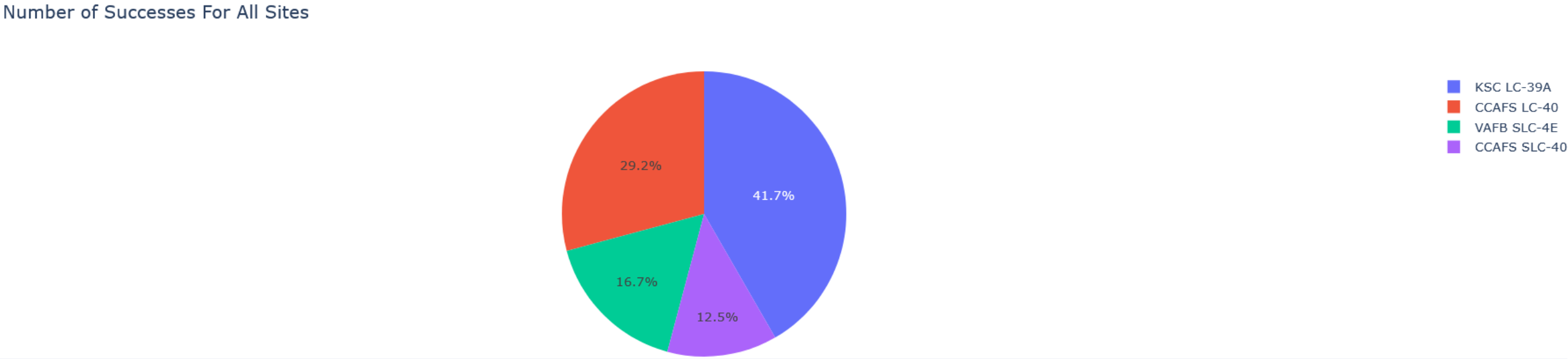



Section 4
Build a Dashboard with Plotly Dash

Pie Chart of Successes Across Launch Sites

Explanation of Chart:

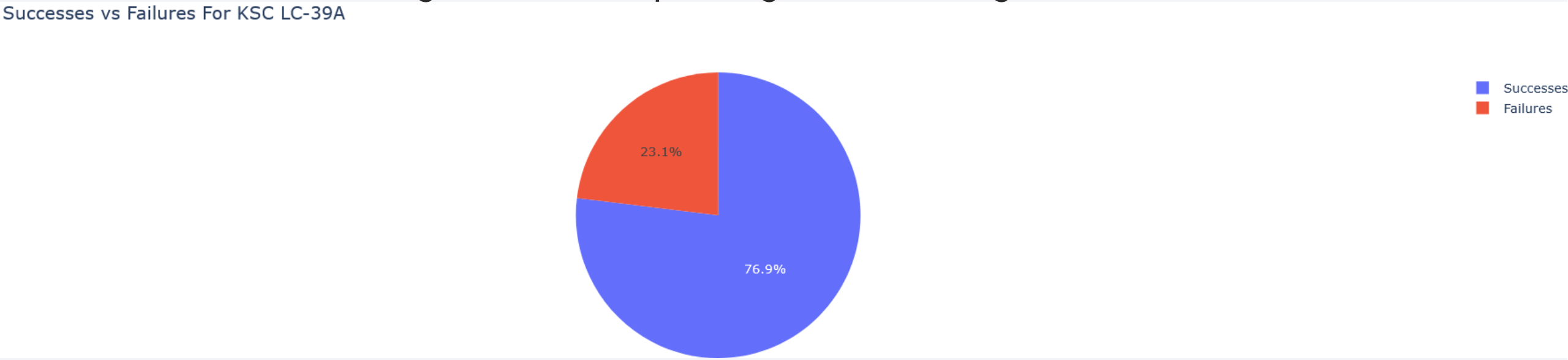
- The pie chart indicates the percentage of the total successes that each launch site claims. It is evident in the chart that KSC LC-39A claims the most at 41.7%.



Pie Chart of Success Rate at Specific Site

Explanation of Chart:

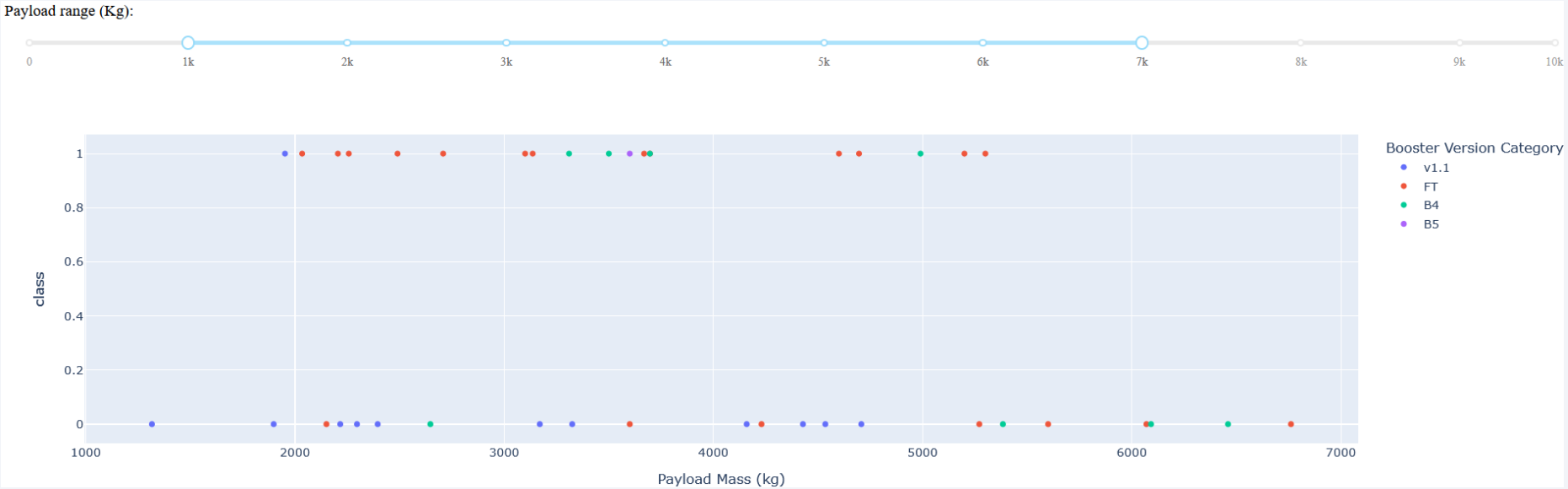
- The pie chart indicates the percentage of the success vs failures at site KSC LC-39A. It is evident in the chart that 76.9% of launches from this side ended in a successful landing, with a corresponding 23.1% ending in failure.



Scatter Plot of Payload Mass vs Outcome

Explanation of Chart:

- The scatter plot the payload mass vs landing outcome. The range slider on top can be used to filter which payload masses to include/exclude. In this example, only flights with more than 1000 kg and less than 7000 kg are shown.



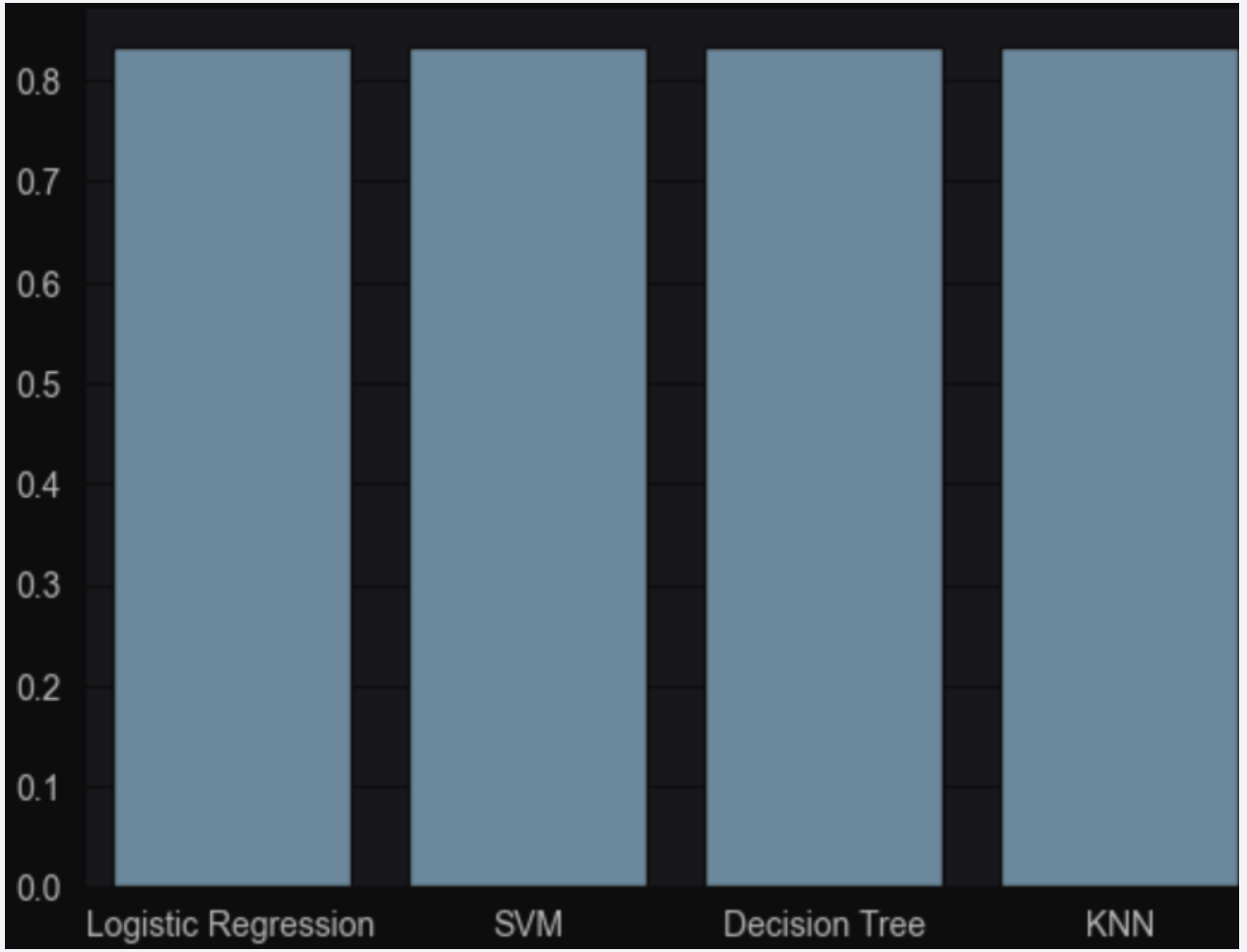
Section 5

Predictive Analysis (Classification)

Classification Accuracy

Explanation of Chart:

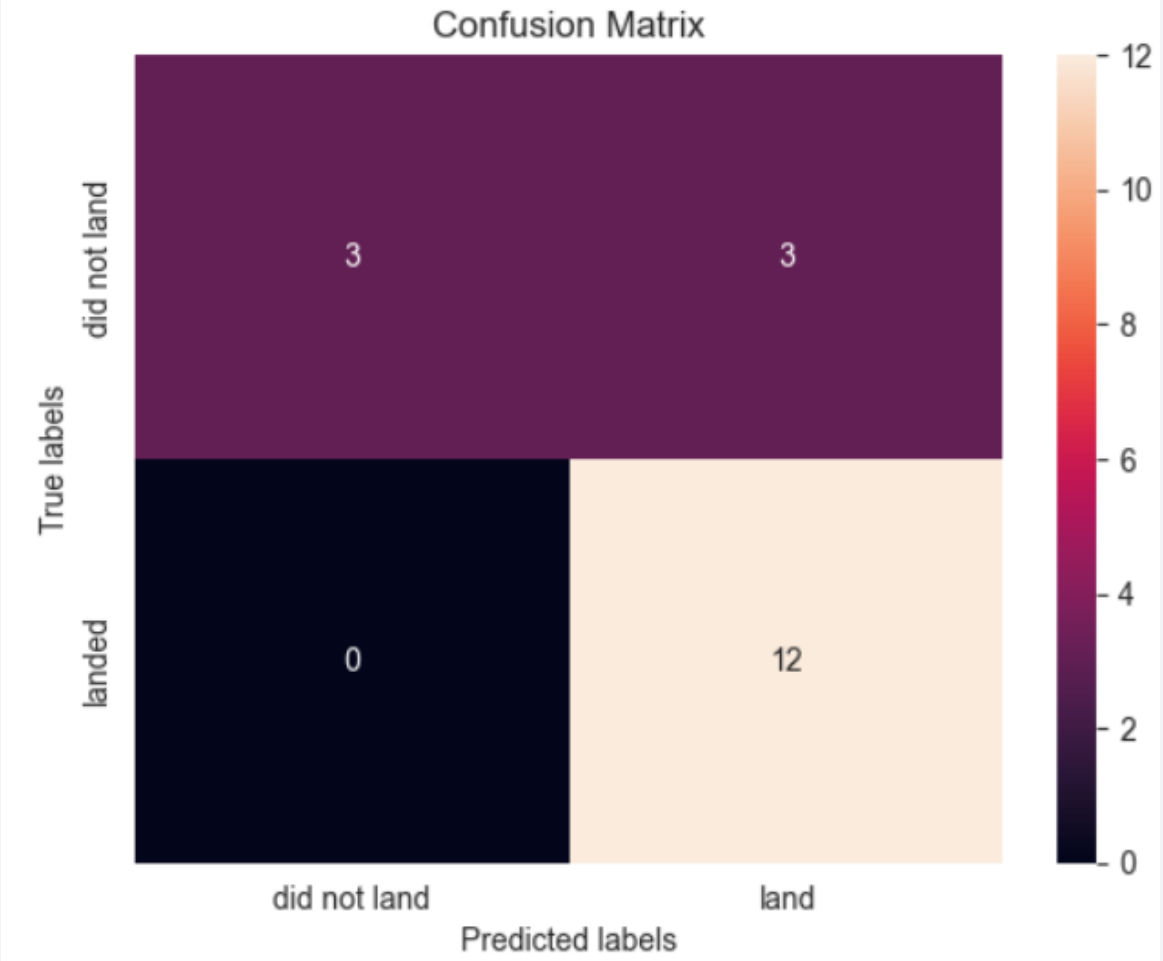
- All tested and optimized models achieved an accuracy of 83%. They also share the exact same confusion matrix. Given this, there is not a clear best model.



Confusion Matrix

Explanation of Chart:

- Accuracy: 0.83
- Precision: 0.8
- Recall: 1.0
- F1-Score: 0.89



Conclusions

- We have determined that SpaceX started as relatively unsuccessful at landing the stage 1 rocket. Over time alterations were made that dramatically increased their rate of success.
- In approximately 2017, the steady increase in success rate stopped. This may indicate that major progress had plateaued, and future progress would come with more work for lesser benefit, after many years of continual improvement.
- The rather high, yet still not incredibly high, prediction capabilities of the machine learning models indicates that most of the information needed to figure out if a landing would be successful was contained within the dataset, although there may be missing data that could improve the model more as an accuracy of 0.83 still has much room for improvement.
- It is possible, given the second and third points, that there is a randomness to each flight that cannot be prepared for or predicted. This randomness could be a form of information that was not (properly) collected or something that is not yet understood. This would support the idea of the plateaued improvement and decent but far from great models, as there may be other factors determining the outcome of the landing whose data is not being reported and/or properly managed.

Appendix

- All code, charts, data, and any other items mentioned in this report cand be found in the following GitHub repository: https://github.com/Alias012/IBM-DataScience-Capstone.

Thank you!