IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Qing Yan
1/21/2026

Outline

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

Executive Summary

- This project analyzes SpaceX launch records collected using SpaceX's API. After data cleaning and exploratory analysis, the project predicts first-stage landing success using multiple classification models with relevant factors selected from the exploratory analysis. Among the models tested, the decision tree model performs best with a 94.4% prediction accuracy.

Introduction

- Project background and context
  - Space launches are costly and operationally complex, making mission failures expensive. Understanding the factors that influence first-stage landing success is essential for improving reliability and reducing risk. This project analyzes historical SpaceX launch data to predict the first-stage landing outcome.
- Problems you want to find answers
  - What factors are related to landing success?
  - How do success rates vary across launch sites, payload ranges, and booster versions?
  - Can machine learning models trained on historical data predict future launch outcomes?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The launch data were collected using SpaceX's API and supplemented with scraped launch records data from Wikipedia.
- Perform data wrangling
  - Data were prepared by standardizing, handling missing values, and labeling outcome variable for machine learning methods.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection

1. Make a request to the SpaceX API

    1.1. Make a HTTP request from SpaceX API

    1.2. Convert the requested data into a dataframe

    1.3. Save relevant variable such as features and dates in a subset dataframe and only include Falcon 9 launches records

2. Scrap Falcon 9 launch records from Wikipedia

    2.1. Request the Falcon9 Launch Wiki page from its URL

    2.2. Parse HTML tables using BeautifulSoup

    2.3. Save relevant variables in a dataframe

3. Merge data collect from SpaceX API and Wikipedia

Data Collection – SpaceX API

- 1. Request launch data from SpaceX API using GET request and the following URL:
  - https://api.spacexdata.com/v4/launches/past
- 2. Convert response content as a JSON and convert into a dataframe using .json( ) and .json_normalize( )
- 3. Keep the relevant columns for the prediction.
  - Save columns such as rocket, payloads, launchpad, cores.
- GitHub URL of SpaceX API calls notebook

Data Collection - Scraping

- 1. Request the Falcon9 Launch Wikipedia page using requests.get( )
  - static_url = https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- 2. Create a BeautifulSoup object from the requested content
- 3. Extract relevant variables from the content and save them into a dataframe.
  - Variables such as Flight No., Date, Time, Version Booster, Launch site, payload, and payload mass.
- GitHub URL of web scraping notebook

Data Wrangling

- The goal of data wrangling is to prepare the dataset for further exploratory analysis and modeling.
- Steps taken in the project include:
  - Identify missing values and replace missing values with mean value
  - Clean the format of dates.
  - Extract key features such as payload mass, orbit, booster version, launch site.
  - Create dummy variables for categorical variables
  - Label outcome variable 1 or 0 or a successful or failed landing
  - Combine all the features and outcome variable for further exploratory analysis and machine learning models for prediction.
- GitHub URL of data wrangling notebook

EDA with Data Visualization

- To visualize the relation between different variables, following chats are plotted:
  - A scatter chat of flight number vs. payload mass with color coded landing success.
    - To see whether launch experience and payload mass are related to landing success.
  - A scatter chat of flight number vs. launch site with color coded landing success.
    - To compare the patterns of fight number and landing success across different launch sites.
  - A scatter chat of payload mass vs. launch site with color coded landing success.
    - To see how payload mass varies across different launch sites and the patten with landing success.

EDA with Data Visualization (continued)

- A bar chat to show the landing success rates across different orbits
- A scatter chat of flight number vs orbit type with color coded landing success
  - To examine whether the landing success improves by experience (flight number) for each orbit
- A scatter chat of payload mass and orbit type with color coded landing success
  - To examine the relation between payload and orbit type and how payload mass related to landing success for each orbit.
- A line chat of yearly average success rate
  - To see how success rate changes overtime

- GitHub URL of EDA with data visualization notebook

EDA with SQL

- Identify unique launch sites using SELECT DISTINCT Launch_Site to understand site distribution.
- Filter launches by pattern
- Compute total payload mass for specific customers such as NASA using SUM( ).
- Calculate average payload mass for a specific booster version using AVG( ).
- Find earliest successful ground-pad landing using MIN( ).
- Count mission outcomes by grouping landing success/failure categories using GROUP BY.
- Identify boosters carrying the maximum payload using a MAX( ) subquery.
- Rank landing outcomes within a date range using COUNT( ) + ORDER BY
- GitHub URL of EDA with SQL notebook

Build an Interactive Map with Folium

- Launch site markers are added with each site's latitude and longitude to identify the location on the map
- Circles around each launch site are added to show the boundaries of launch sites on the map
- Launch outcomes are colored coded in the map with green markers as success landings and red markers as failed landings
- Mouse position tool is added as an interactive tool allowing users to extract the latitude and longitude of a location the mouse is hovered over on the map
- Distance lines are added on the map to show the distance between a launch site and nearest coastline, railway, highway or a city.
- GitHub URL of interactive map with Folium map

Build a Dashboard with Plotly Dash

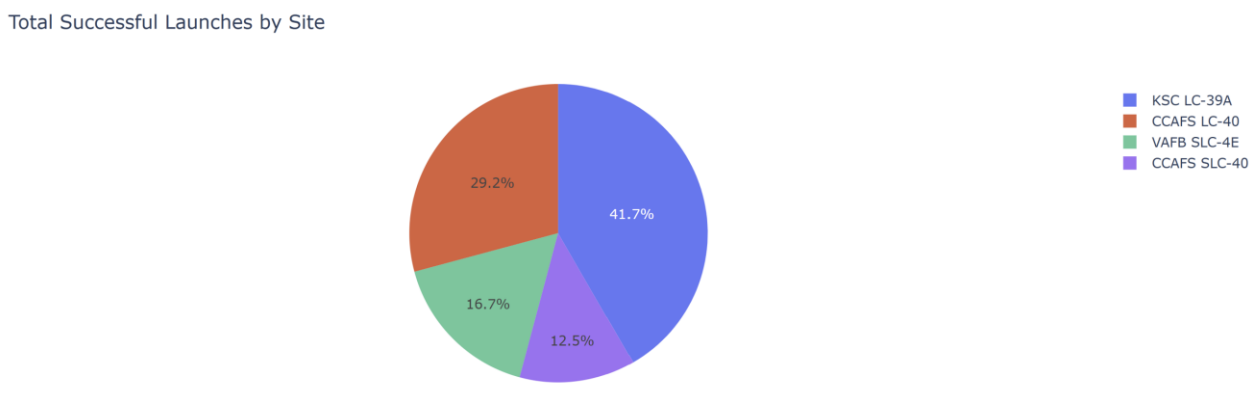
- In the interactive dashboard a launch site dropdown is added to allow users select “All Sites” or a specific launch site to examine their success rates.
- A pie chart is created to show the distribution of success rates by different launch sites on the “All Sites” page. For any specific launch site, a pie chart is created to show the success rate and failure rate for the selected site.
- A payload range slider is added to allow users examine how payload mass is related to the landing outcome.
- A scatter chat is plotted for payload mass vs. launch outcome with color coded booster versions to examine how payload mass and booster type are related to the launch outcome for a selected launch site.
- GitHub URL of Plotly Dash lab

Predictive Analysis (Classification)

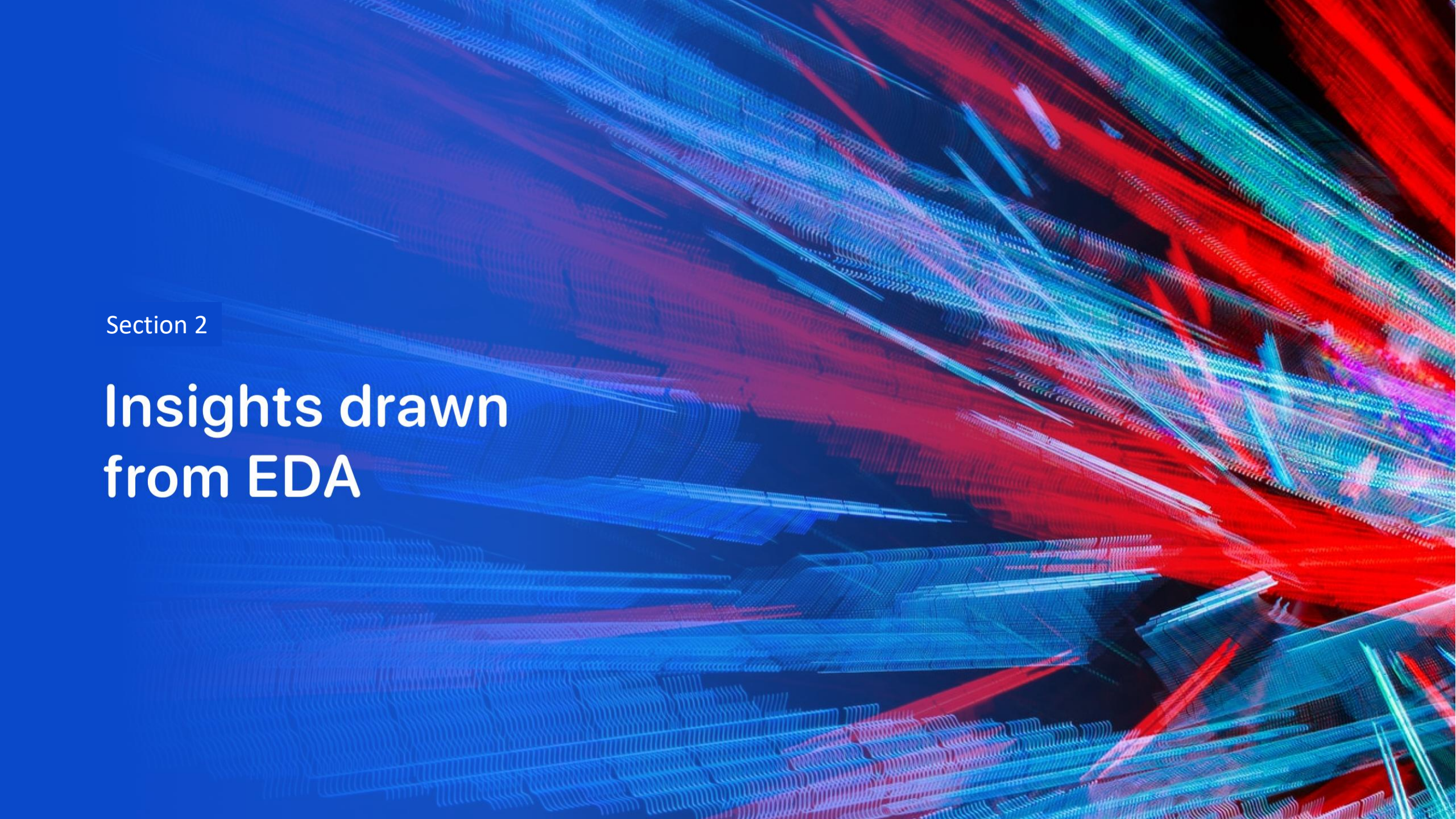
- Prepare the datasets
  - Split data into 80% training data and 20% test data
- Train the following classification models
  - Logistic regress, support vector machine, decision tree, and K-nearest neighbors
- Evaluate prediction using test data with confusion matrix
- Improve the models with hyperparameter tuning
  - Use GridSearchCV (10-fold cross-validation) to search for the best combination of parameters
- Identify the best model by comparing the prediction accuracy
- GitHub URL of predictive analysis lab

Results

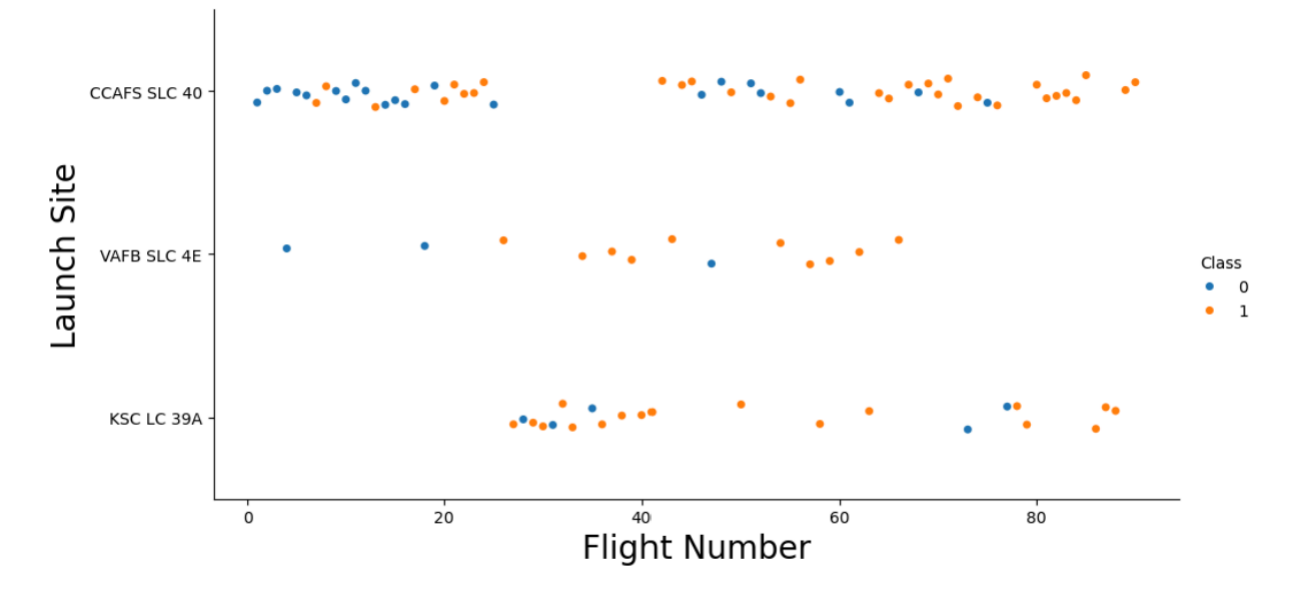
- Exploratory data analysis shows that lunch success improves with experience (flight number). Site KSC LC-39A shows relatedly higher success rate than others. Payload mass is positively correlated with the landing success rate. Certain obit types show higher success rates.
- Interactive analytics demo shows that site KSC LC-39A has the highest success rate than other sites.



- Predictive analysis results show that decision tress model has the higher prediction accuracy (94.4%) than other models trained and tested.

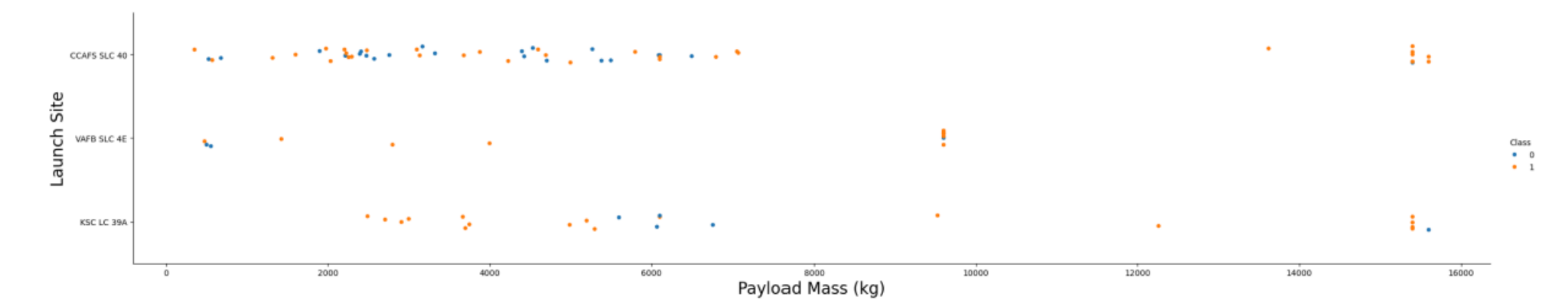

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



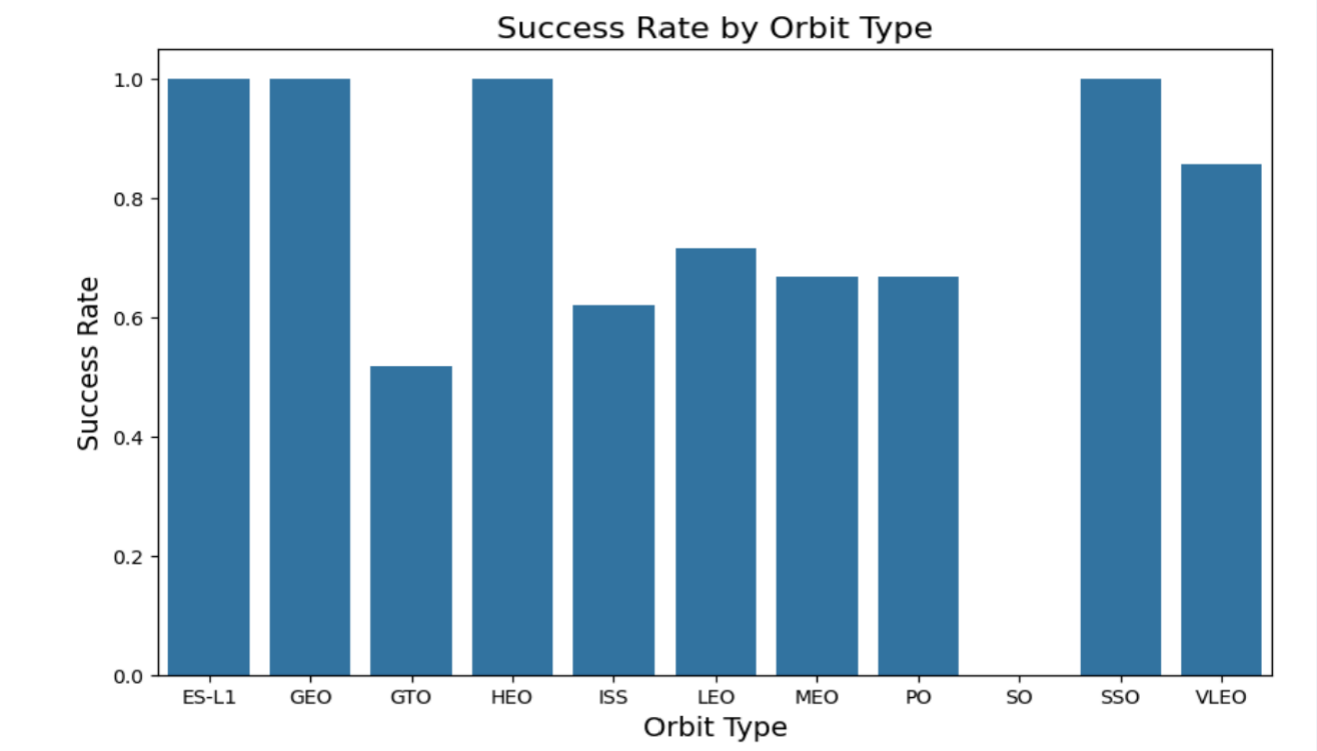
- The scatter chat shows that landing success improves as flight numbers increase across all launch sites.
- CCAFS SLC-40 handled many of the earliest missions, while KSC LC-39A had mostly successful landings during later flights.
- Across all the sites, flight number seems positively correlated with the landing success.

Payload vs. Launch Site



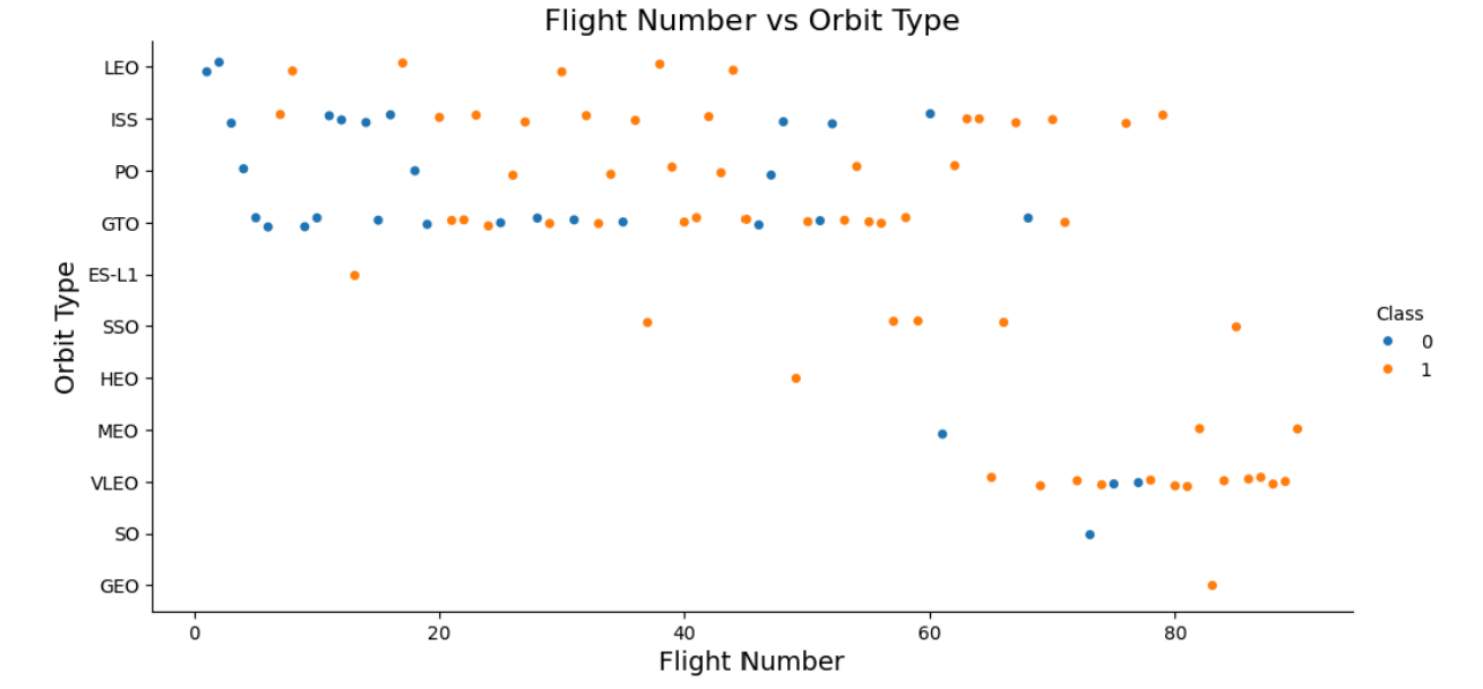
- CCAFS SLC-40 handles a wide range of payload masses.
- VAFB SLC-4E did not launch rockets with heavy payloads (>10,000 kg).
- KSC LC-39A also launches rockets with very heavy payloads, including the heaviest payloads in the dataset (up to ~16,000 kg)
- Landing success seems more common at both CCAFS SLC-40 and KSC LC-39A across all payload ranges.

Success Rate vs. Orbit Type



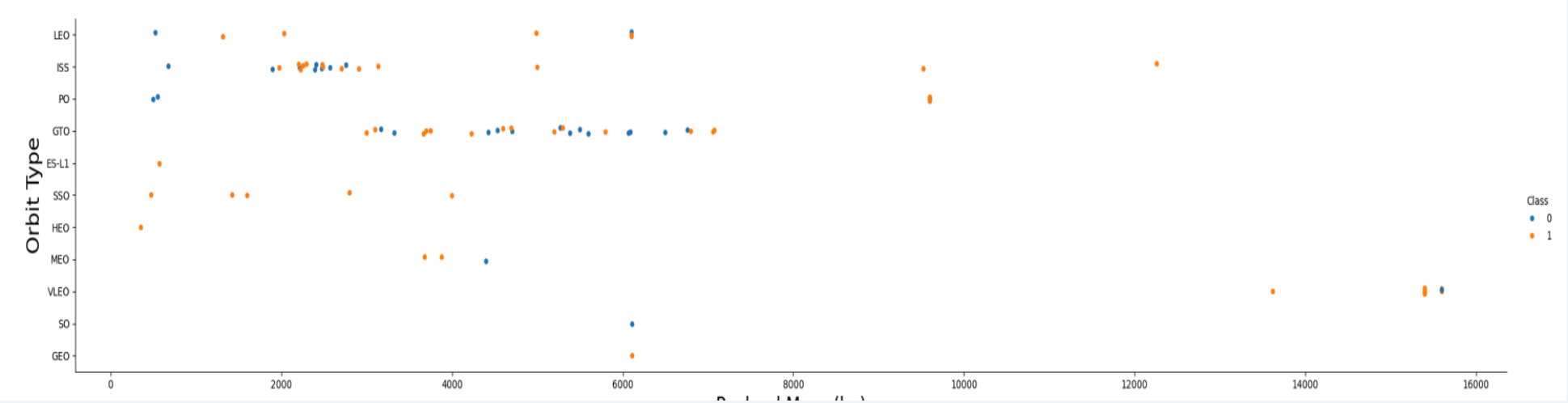
- Lower orbits (LEO, VLEO, MEO, PO) show relatively higher success rates, suggesting easier conditions for booster recovery.
- Higher orbits (GTO, ISS missions) have lower success rates, reflecting more challenging missions.

Flight Number vs. Orbit Type



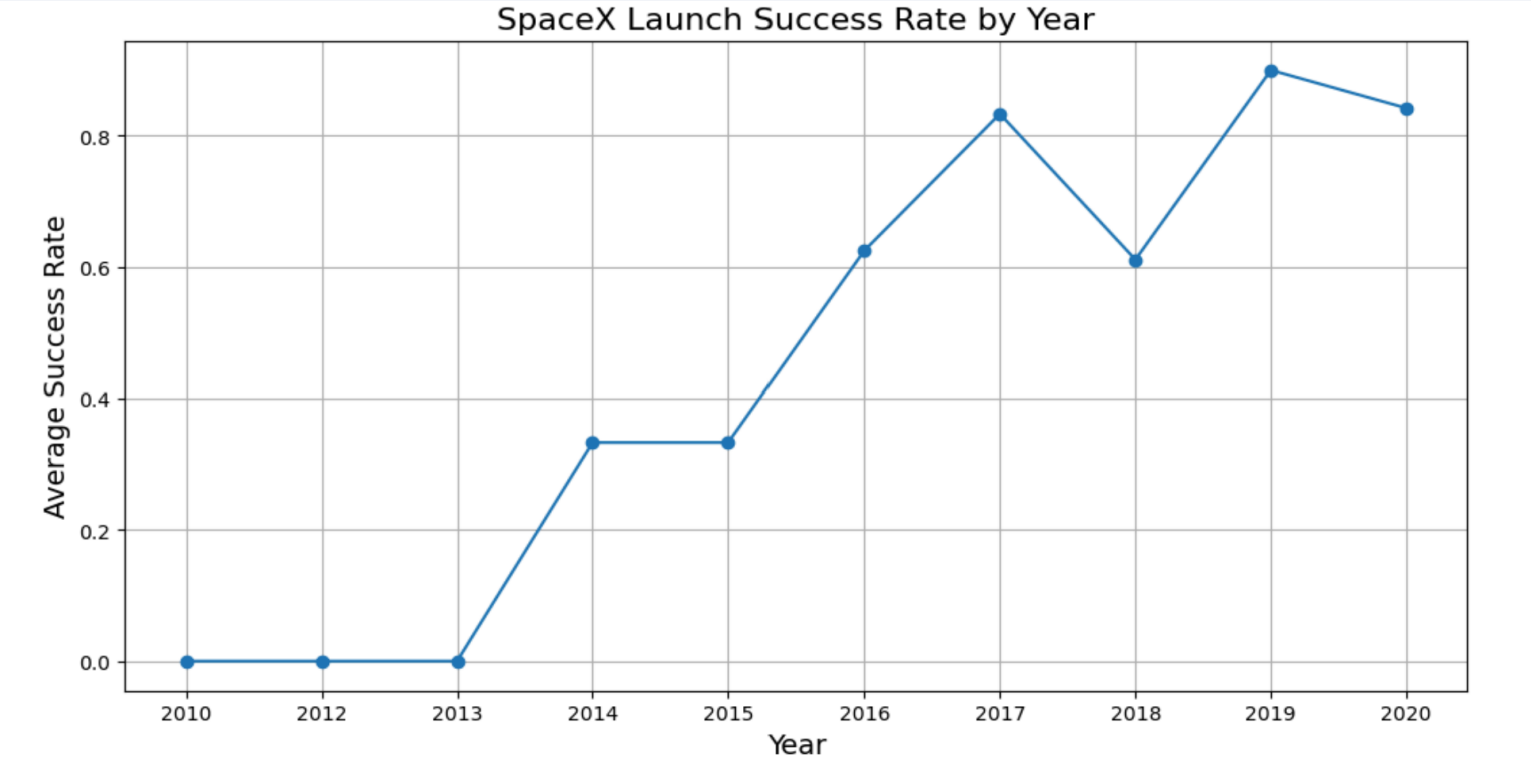
- Flight number is positively correlated with landing success, suggesting the launch experience is positively related with landing success.
- LEO, VLEO, and SSO missions have more successful landings.

Payload vs. Orbit Type



- LEO and ISS missions have heavy payloads and high success rates.
- SSO missions have lighter payload but high success rate.

Launch Success Yearly Trend



- Launch success rate improved overtime
- 2014 had a breakthrough from the low success between 2010 to 2013
- 2019 had the higher success rate between the period of 2010 to 2020

All Launch Site Names

```
[22]: %%sql
SELECT DISTINCT "Launch_Site"
FROM SPACEXTABLE;
```

```
 * sqlite:///my_data1.db
Done.
```

[22]:

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

- There are four unique launch sites and results show their names.

Launch Site Names Begin with 'CCA'

- %%sqlSELECT *FROM SPACEXTABLEWHERE "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

- The results return five entries of lunch site name begein with ‘CCA’

Total Payload Mass

[27]:
```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS TotalPayload
FROM SPACEXTABLE
WHERE Customer = 'NASA (CRS)';
```

```
 * sqlite:///my_data1.db
Done.
```

[27]:

TotalPayload
45596

- The total payload mass carried by boosters launched by NASA (CRS) is 45596KG.

Average Payload Mass by F9 v1.1

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS AvgPayload
FROM SPACEXTABLE
WHERE Booster_Version = 'F9 v1.1';
```

```
 * sqlite:///my_data1.db
Done.
```

AvgPayload
2928.4

- The average payload mass carried by booster version F9 v1.1 is 2928.4 KG.

First Successful Ground Landing Date

```
%%sql
SELECT MIN(Date) AS First_GroundPad_Success
FROM SPACEXTABLE
WHERE Landing_Outcome = 'Success (ground pad)';
```

```
 * sqlite:///my_data1.db
Done.
```

First_GroundPad_Success
2015-12-22

- The date when the first successful landing outcome in ground pad was achieved was December 22, 2025.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
SELECT DISTINCT Booster_Version
FROM SPACEXTABLE
WHERE Landing_Outcome = 'Success (drone ship)'
  AND PAYLOAD_MASS__KG_ > 4000
  AND PAYLOAD_MASS__KG_ < 6000;
```

```
 * sqlite:///my_data1.db
Done.
```

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

- The query result shows the list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

```
%%sql
SELECT Mission_Outcome, COUNT(*) AS Total
FROM SPACEXTABLE
GROUP BY Mission_Outcome;
```

```
 * sqlite:///my_data1.db
Done.
```

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

- The query results counts the total number of successful and failure mission outcomes.

Boosters Carried Maximum Payload

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

- The query result show a list the unique names of the booster which have carried the maximum payload mass in the dataset

2015 Launch Records

```
%%sql
SELECT
    substr(Date, 6, 2) AS Month,
    Landing_Outcome,
    Booster_Version,
    Launch_Site
FROM SPACEXTABLE
WHERE Landing_Outcome = 'Failure (drone ship)'
  AND substr(Date, 1, 4) = '2015';
```

```
 * sqlite:///my_data1.db
Done.
```

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

- The query results provide a list of the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

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

Landing_Outcome	OutcomeCount
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

- The query results provide the ranking of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

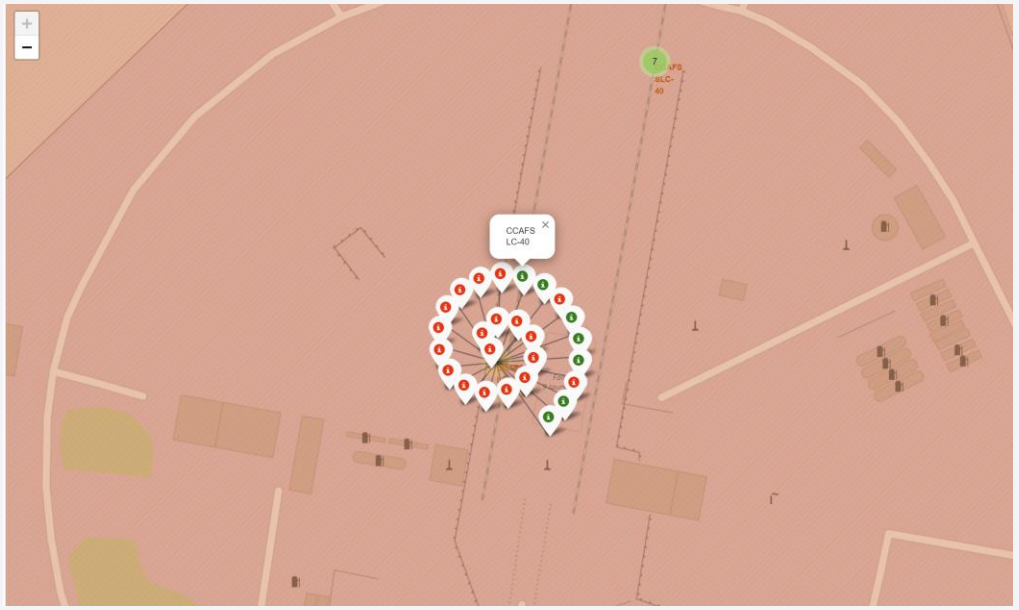
Section 3

Launch Sites Proximities Analysis

Launch sites locations in the US

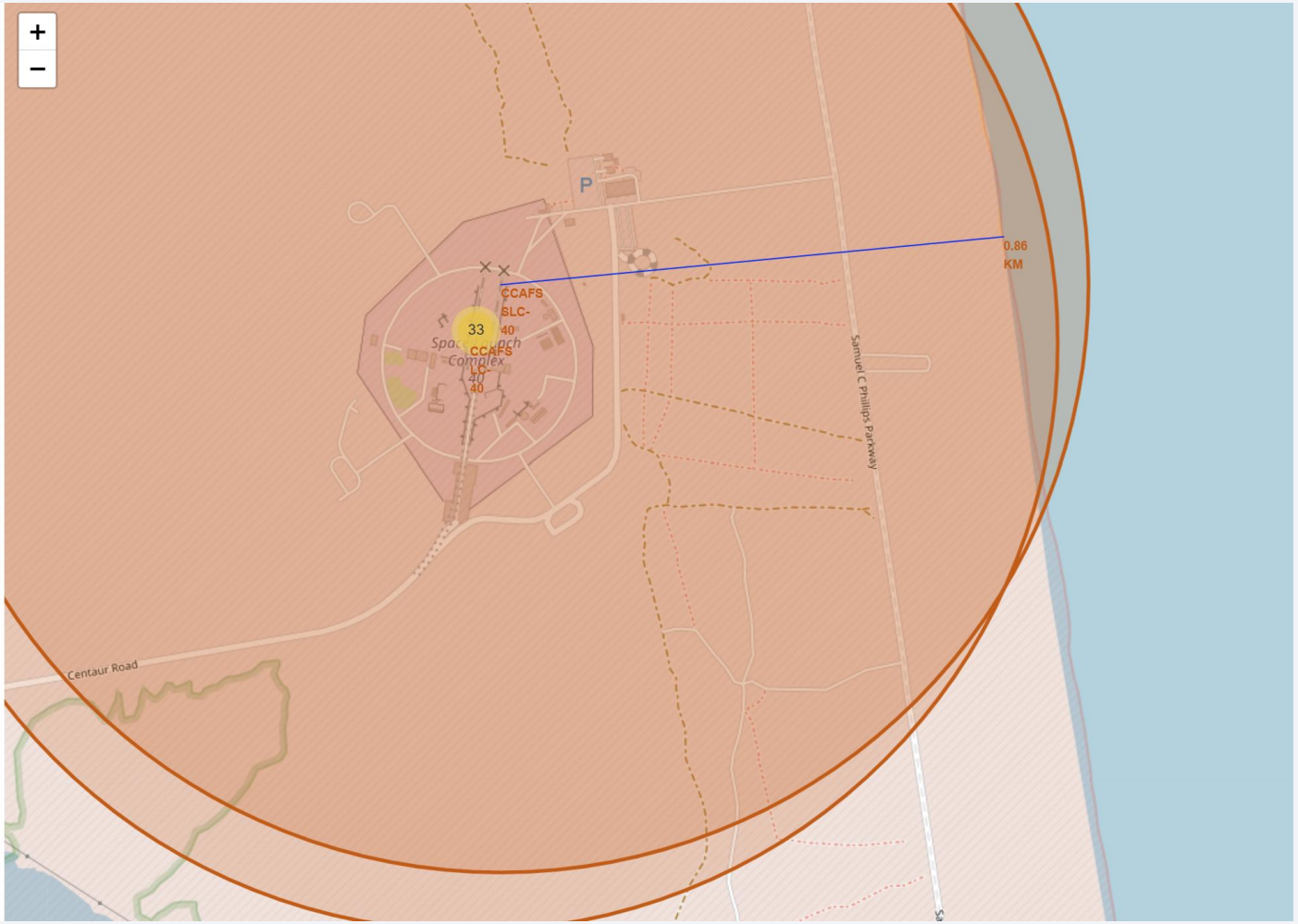
- Launch site VAFB SLC-4E is located in California near Vandenberg Space Force Base
- Launch sites KSC LC-39A, CCAFS LC-40 and CCAFS SLC-40 are located in Florida near Cape Canaveral Space Force Station

Launch results at site CCAFS LC-40

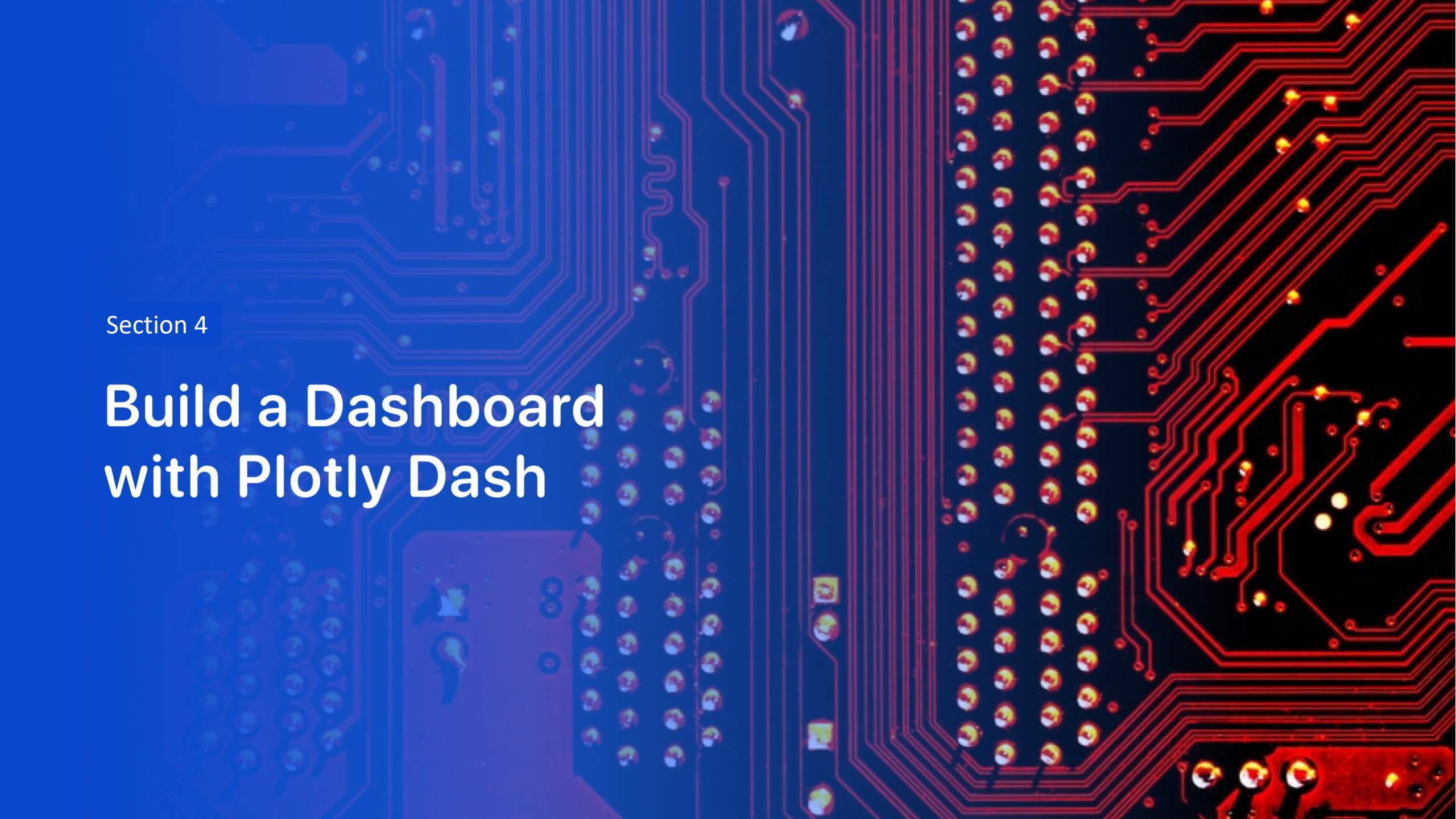


- Among 26 launches at site CCAFS LC-40, 7 succeed and 19 failured, showing a success rate of 26.9%

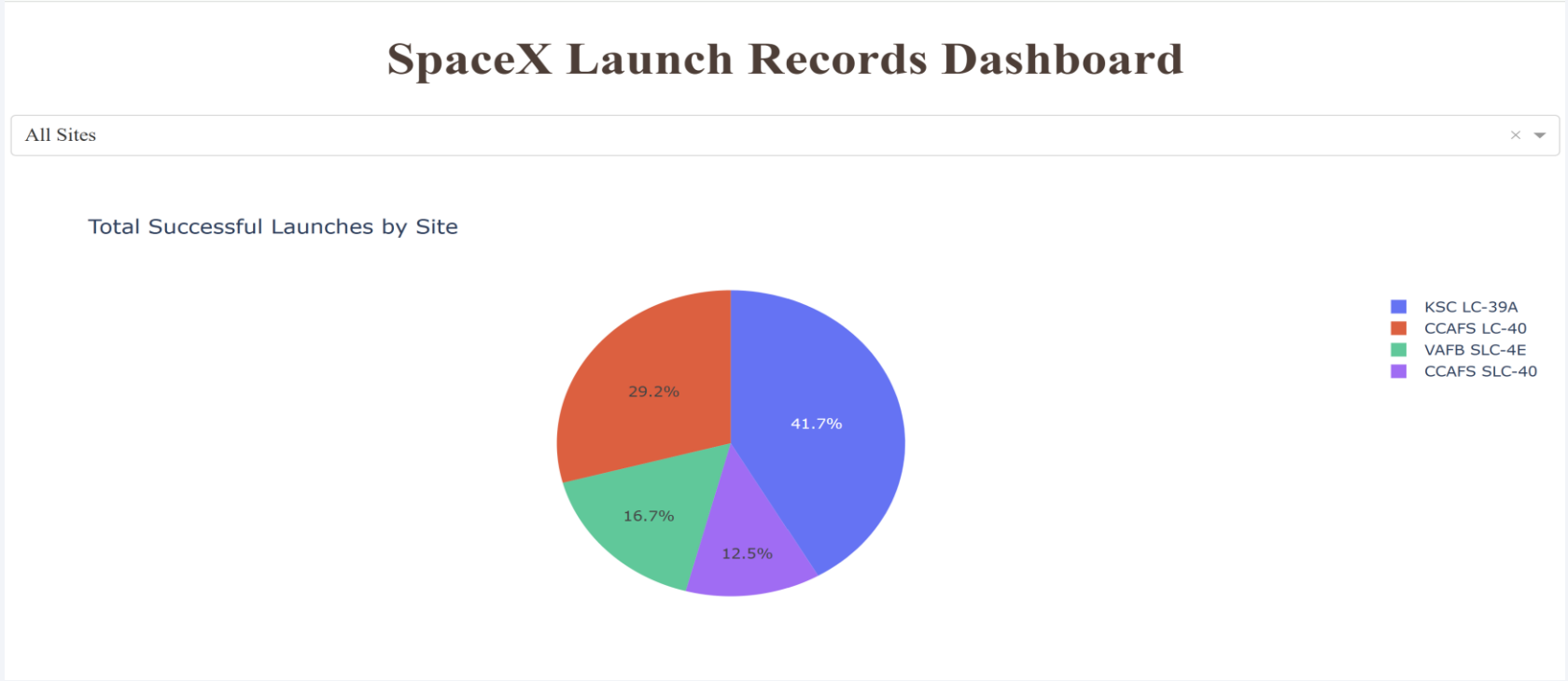
Distance between site CCAFS SLC-40 and coastline



- The distance between site CCAFS SLC-40 and coastline is about 0.86 KM.
- The site is located very close to the ocean so that any failed launches, debris, or boosters could fall into water instead of populated areas.

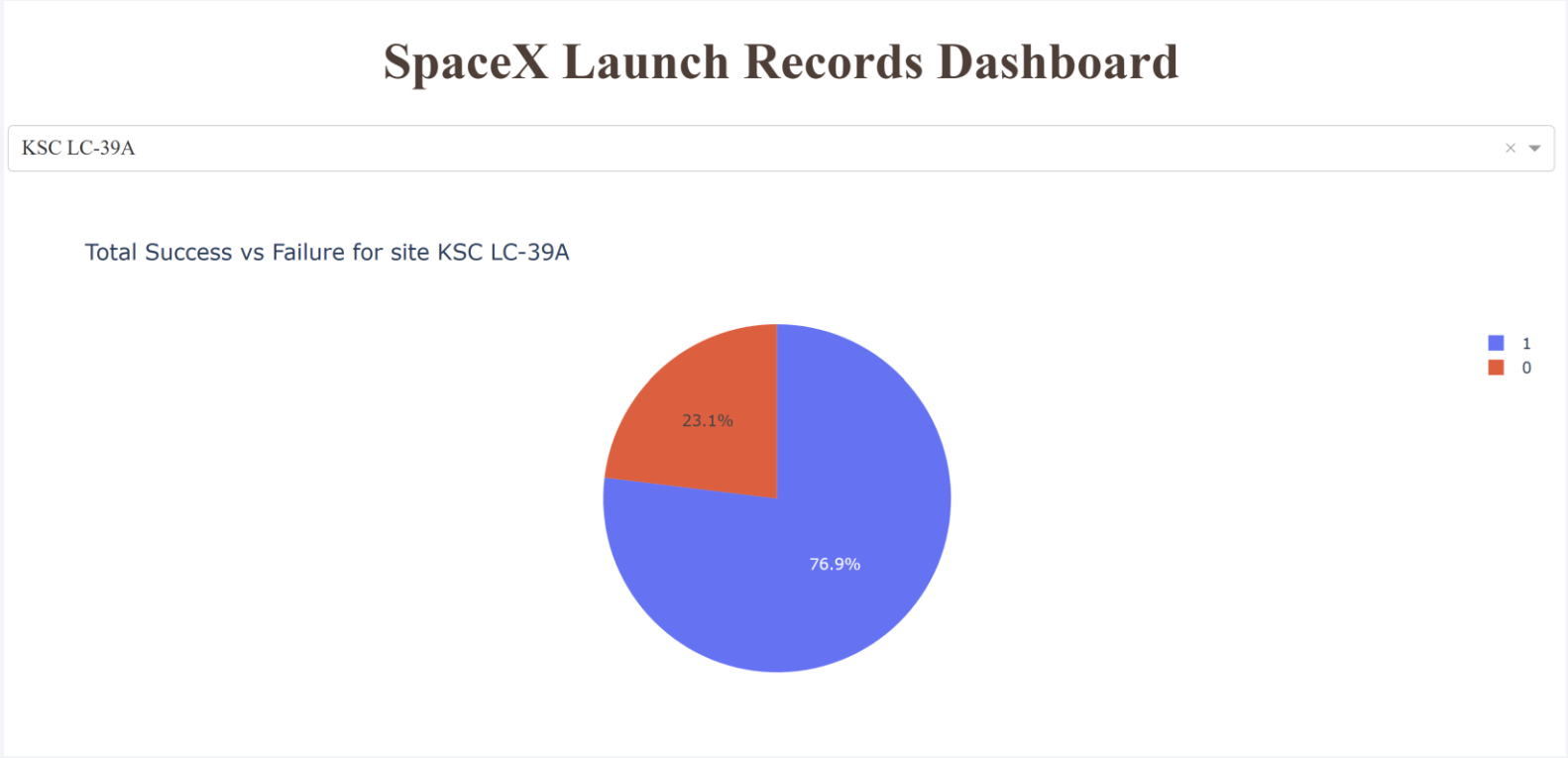

Section 4
Build a Dashboard with Plotly Dash

Launch Records by all sites



- The pie chart shows the launch success by site. Among the four sites, KSC LC-39A contributed 41.7 %, the highest percent of the successful launches. Site CCAFS SLC-40 contributed 12.5% the lowest percent of the successful launches.

Launch records of site KSC LC-39A



- If we take a closer look at site KSC LC-39A, among all the launches on this site, 76.9% succeed and 23.1% failed.

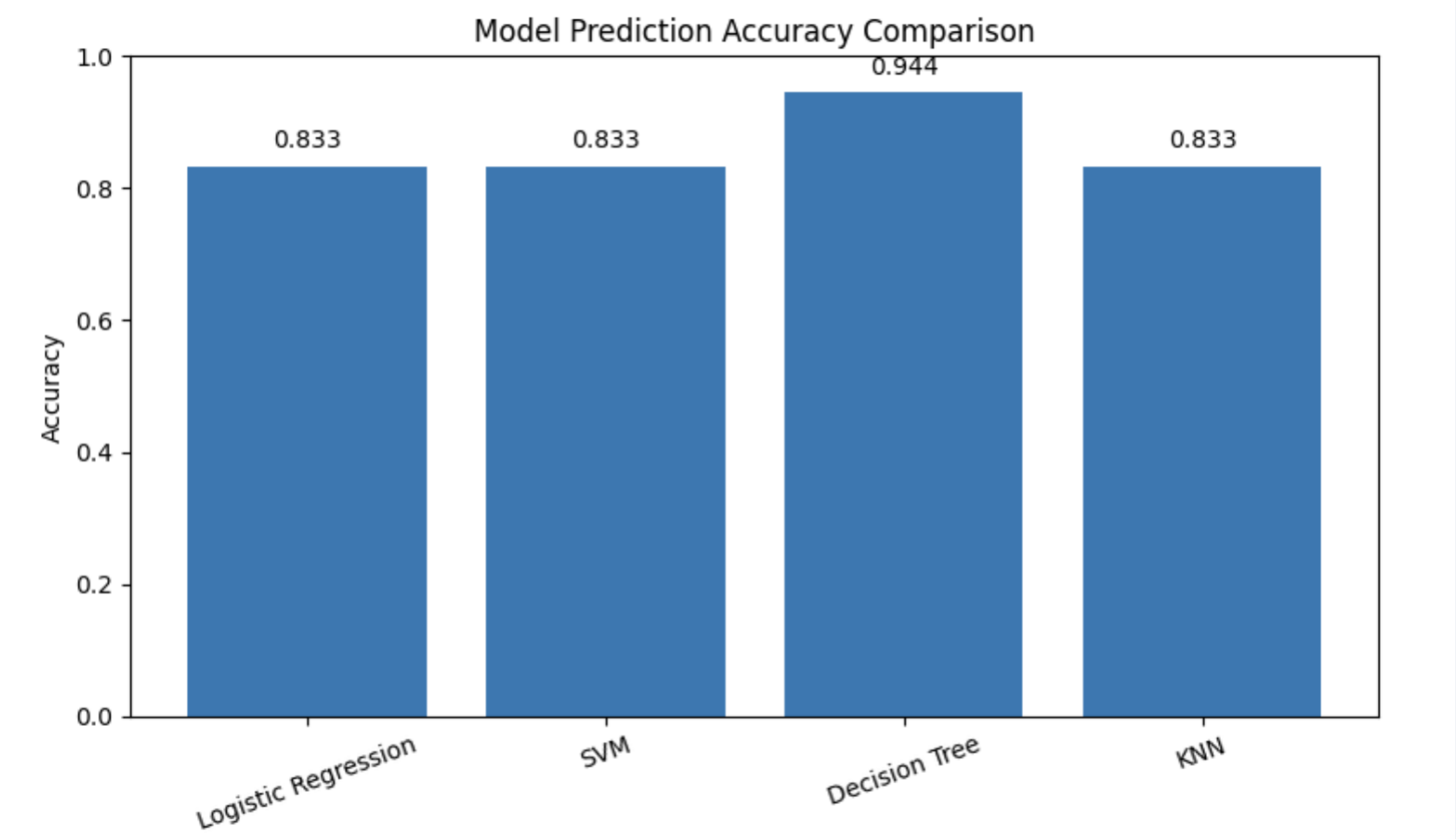
Payload vs. Launch Outcome scatter plot for all sites

- Launches in the low-payload range show a mix of successes and failures.
- v1.0 and v1.1 boosters appear more frequently in low-payload range and show more failures compared with newer boosters.
- High-payload range launches have fewer records, but all high-payload launches succeed.
- Boosters FT and B4 boosters are mainly used in heavier missions

Section 5

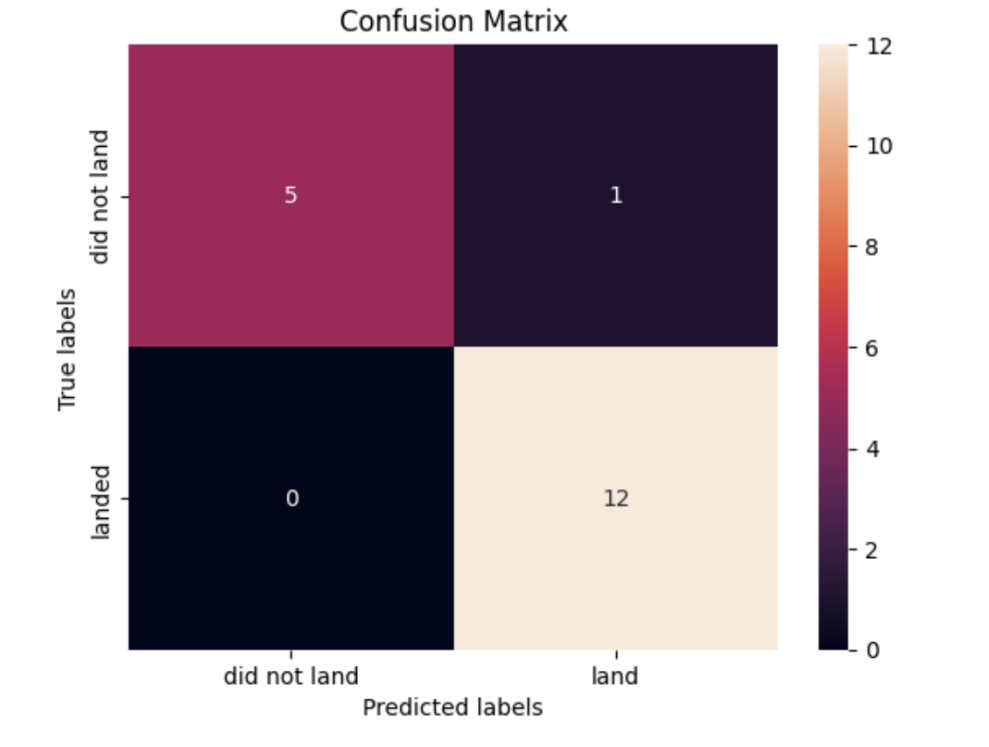
Predictive Analysis (Classification)

Classification Accuracy



- Decision tree model has highest classification accuracy with a 99.4% accuracy.

Confusion Matrix



- Comparing 18 predicted outcomes and 18 true outcomes in the test dataset, the confusion matrix shows that the decision tree model predicted 5 true negative and 12 true positive outcomes. Therefore, the total accurate predictions are 17 out of 18 (94.4%).

Conclusions

- 1. This project shows that factors such as launch site, payload mass, mission obit type play import roles in the launch outcome
- 2. After a breakthrough in 2014, the success rate has been increasing overtime. It shows SpaceX has been learning through the iterations of launches.
- 3. Using the SpaceX Falcon 9 launch data between 2010 and 2020, this project uses four classification models to predict the launch outcome with selected predictors. After training and turning the model parameters, testing dataset suggests decision tree model performs best among the four tested models with a 94.4% prediction accuracy.

Appendix

- GitHub URL of this project

Thank you!