IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Eric Monsalve Cuevas
December 13, 2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive Analytics result

Introduction

- Project background and context

  SpaceX offers Falcon 9 rocket launches at a listed cost of $62 million, significantly lower than other providers, whose prices exceed $165 million per launch. This cost advantage largely stems from SpaceX's ability to reuse the rocket's first stage. By predicting whether the first stage will land successfully, the cost of a launch can be estimated, providing valuable insight for companies aiming to compete with SpaceX in the rocket launch market. The objective of this project is to develop a machine learning pipeline capable of predicting the success of first-stage landings.

- Key questions to address

  - What factors influence the success of a rocket's landing?
  - How do various features interact to impact the probability of a successful landing?
  - What operational conditions are necessary to achieve consistent landing success?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data from SpaceX was obtained from 2 sources:
    - SpaceX API (https://api.spacexdata.com/v4/rockets/)
    - WebScraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)
- Perform data wrangling:
  - Collected data was enriched by creating a landing outcome label based on outcome data after summarizing and analyzing features

Methodology

Executive Summary

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - The data collected up to this point was normalized, split into training and test sets, and analyzed using four distinct classification models. Each model's accuracy was assessed based on various parameter combinations.

Data Collection

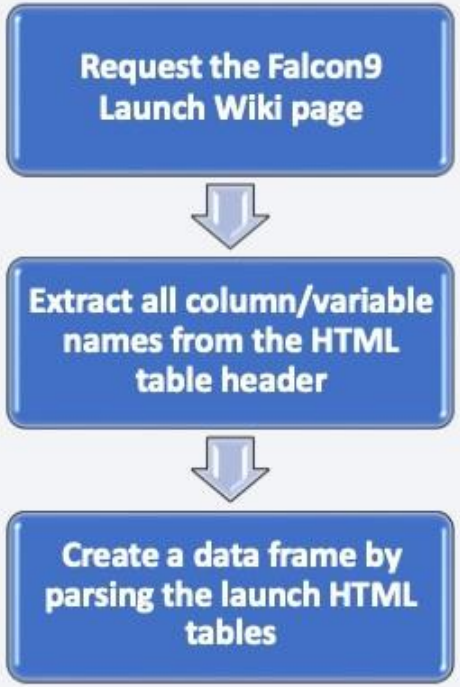
- Describe how data sets were collected.
- Datasets were collected from SpaceX API **(https://api.spacexdata.com/v4/rockets/)** and from Wikipedia **(https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)**, using web scraping technics.

Data Collection – SpaceX API

- SpaceX provides a public API that allows data retrieval for further use.
- The API was utilized following the outlined flowchart, and the data was subsequently stored.
- Source code: https://github.com/SouRitra01/IBM-Data-Science-Project/blob/main/Data%20Collection%20API.ipynb

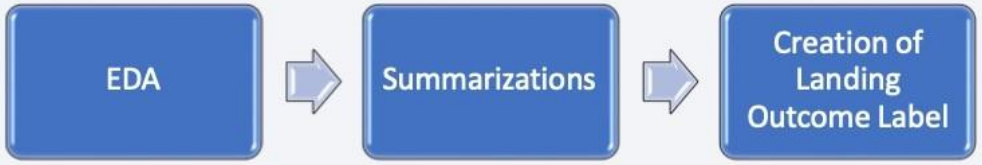
Data Collection - Scraping

- Information on SpaceX launches is also available on Wikipedia.
- Data was retrieved from Wikipedia following the flowchart and subsequently stored.
- Source code: https://github.com/SouRitra01/IBM-Data- Science-Project/blob/main/Data%20Collection%20wit h%20Web%20Scraping.ipynb



Data Wrangling

- An initial Exploratory Data Analysis (EDA) was conducted on the dataset.
- Summaries were generated for launches by site, orbit occurrences, and mission outcomes by orbit type.
- Lastly, the landing outcome label was derived from the Outcome column.



- Source code: https://github.com/SouRitra01/IBM-Data-Science-Project/blob/main/Data%20Wrangling.ipynb

EDA with Data Visualization

- The following SQL queries were executed:
    - Retrieve the unique names of launch sites used in space missions.
    - Identify the top 5 launch sites starting with 'CCA'.
    - Calculate the total payload mass carried by NASA (CRS) boosters.
    - Determine the average payload mass for booster version F9 v1.1.
    - Find the date of the first successful landing outcome on a ground pad.
    - List boosters with successful drone ship landings and payloads between 4000 and 6000 kg.
    - Count the total number of successful and failed mission outcomes.
    - Identify booster versions that carried the highest payload mass.
    - Retrieve failed drone ship landings in 2015, including their booster versions and launch site names.
    - Rank the count of landing outcomes (e.g., Failure on droneship or Success on ground pad) between 2010-06-04 and 2017-03-20.
- Source code: https://github.com/SouRitra01/IBM-Data-Science-Project/blob/main/EDA.ipynb

EDA with SQL

- Scatterplots and bar plots were utilized to explore the data and visualize relationships between feature pairs.
- Payload Mass vs. Flight Number, Launch Site vs. Flight Number, Launch Site vs. Payload Mass, Orbit vs. Flight Number, and Payload vs. Orbit.

- Source code: https://github.com/SouRitra01/IBM-Data-Science-Project/blob/main/EDA%20with%20Data%20Visualization.ipynb

Build an Interactive Map with Folium

- Folium Maps were enhanced with markers, circles, lines, and marker clusters.
- Markers represent points of interest, such as launch sites.
- Circles highlight areas around specific locations, like the NASA Johnson Space Center.
- Marker clusters group events at the same coordinates, such as multiple launches at a single site.
- Lines depict distances between two coordinates.
- Source code: https://github.com/SouRitra01/IBM-Data-Science-Project/blob/main/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

- An interactive dashboard was created using Plotly Dash.
- Pie charts were included to display total launches by specific sites.
- A scatter plot visualized the relationship between Outcome and Payload Mass (kg) across different booster versions.
- The notebook can be accessed at: https://github.com/chuksoo/IBM-Data-Science- Capstone-SpaceX/blob/main/app.py

Predictive Analysis (Classification)

- Data was loaded using NumPy and Pandas, transformed, and split into training and testing sets.
- Various machine learning models were developed, with hyperparameters optimized using GridSearchCV.
- Accuracy was used as the evaluation metric, and the model was enhanced through feature engineering and algorithm tuning.
- The best-performing classification model was identified.
- The notebook is available at: https://github.com/chuksoo/IBM-Data-Science- Capstone-SpaceX/blob/main/Machine%20Learning%20Prediction.ipynb

Results

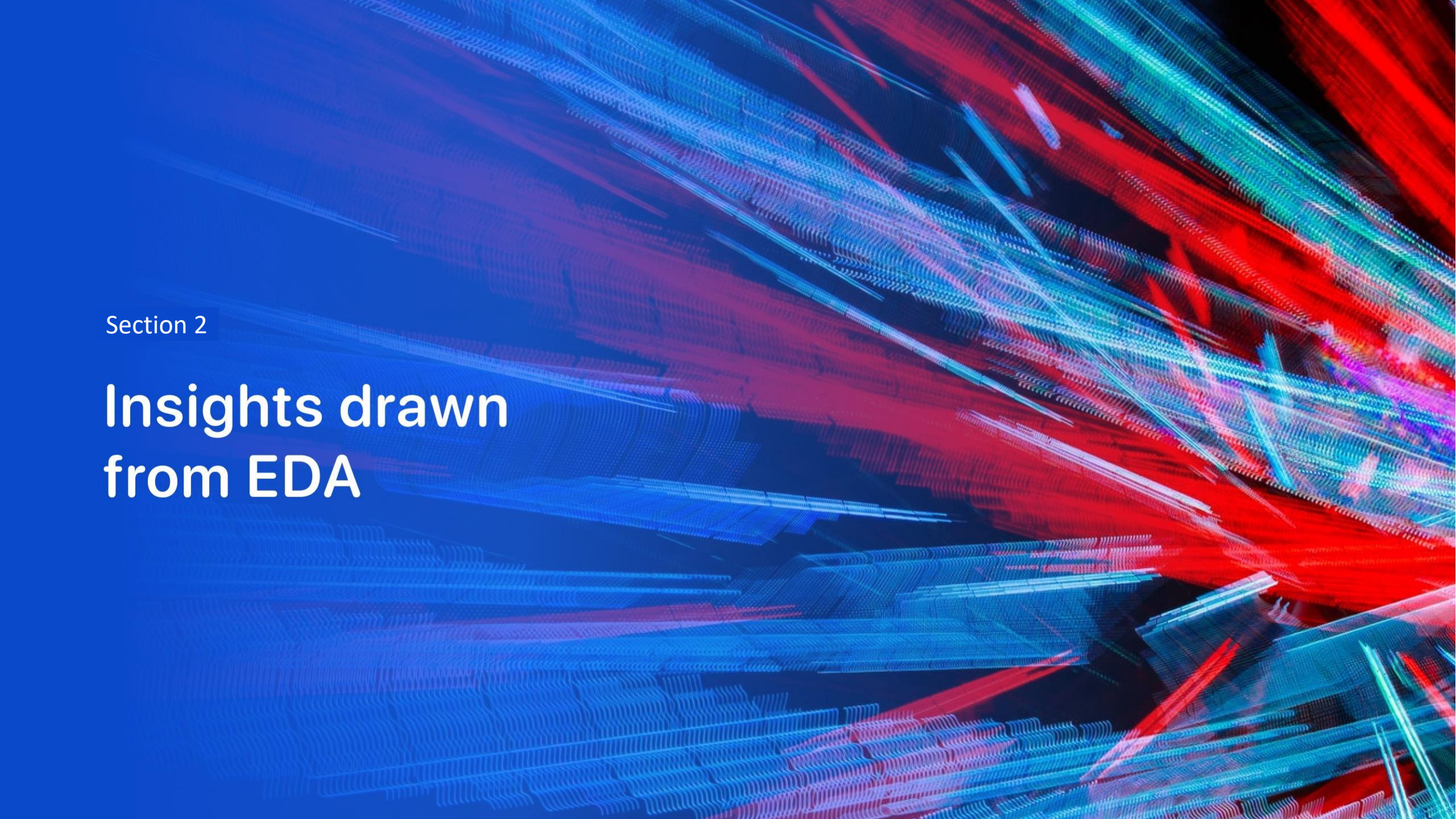
- Exploratory Data Analysis Findings:
  - SpaceX operates four distinct launch sites.
  - Initial launches were conducted for SpaceX itself and NASA.
  - The F9 v1.1 booster has an average payload of 2,928 kg.
  - The first successful landing occurred in 2015, five years after the inaugural launch.
  - Several Falcon 9 booster versions successfully landed on drone ships with payloads above average.
  - Nearly all mission outcomes were successful.
  - Two booster versions, F9 v1.1 B1012 and F9 v1.1 B1015, failed drone ship landings in 2015.
  - The success rate of landing outcomes has improved over the years.

Results

- Interactive analytics revealed that launch sites are typically located in safe areas, such as near the sea, and are supported by strong logistical infrastructure.
- The majority of launches take place at East Coast locations.

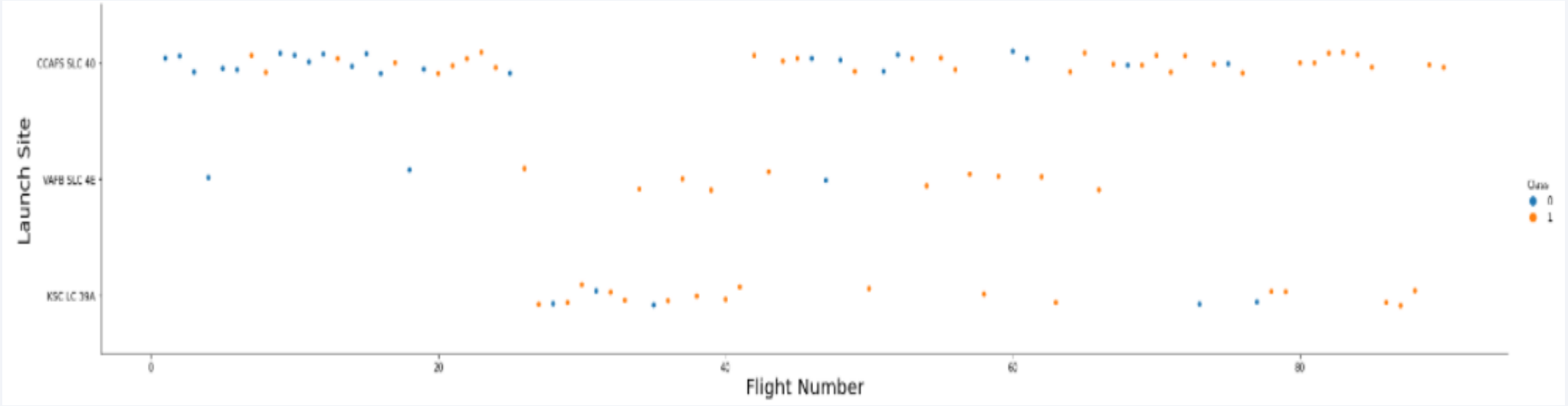
Section 2

Insights drawn from EDA



Flight Number vs. Launch Site

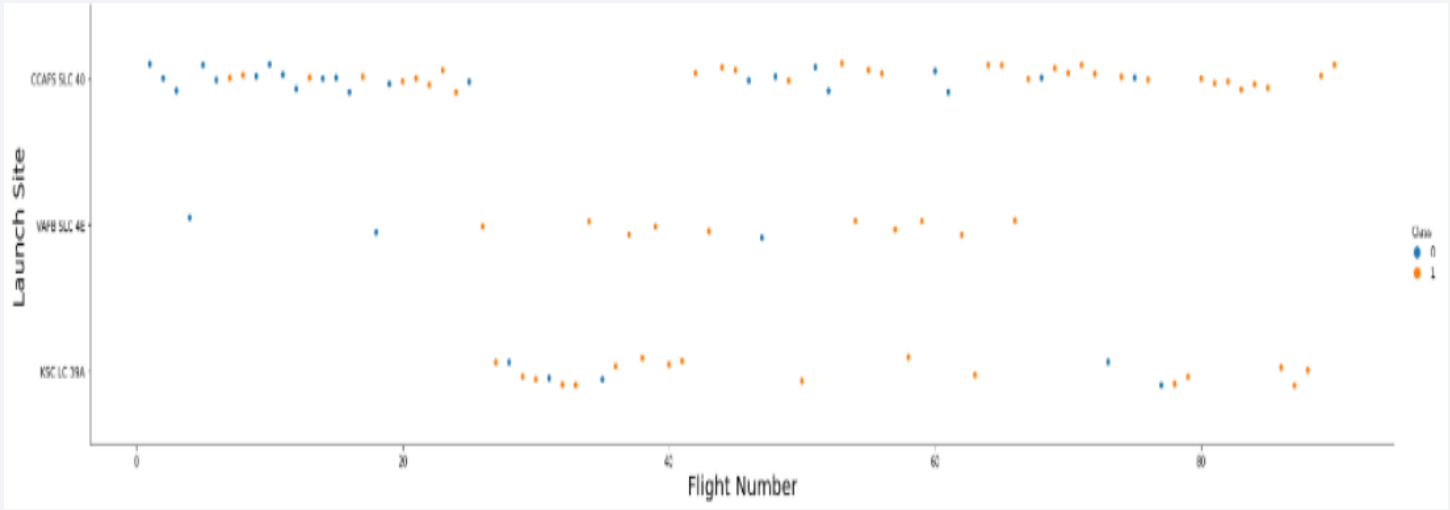
- The plot indicates that higher flight volumes at a launch site correlate with increased success rates.



Payload vs. Launch Site

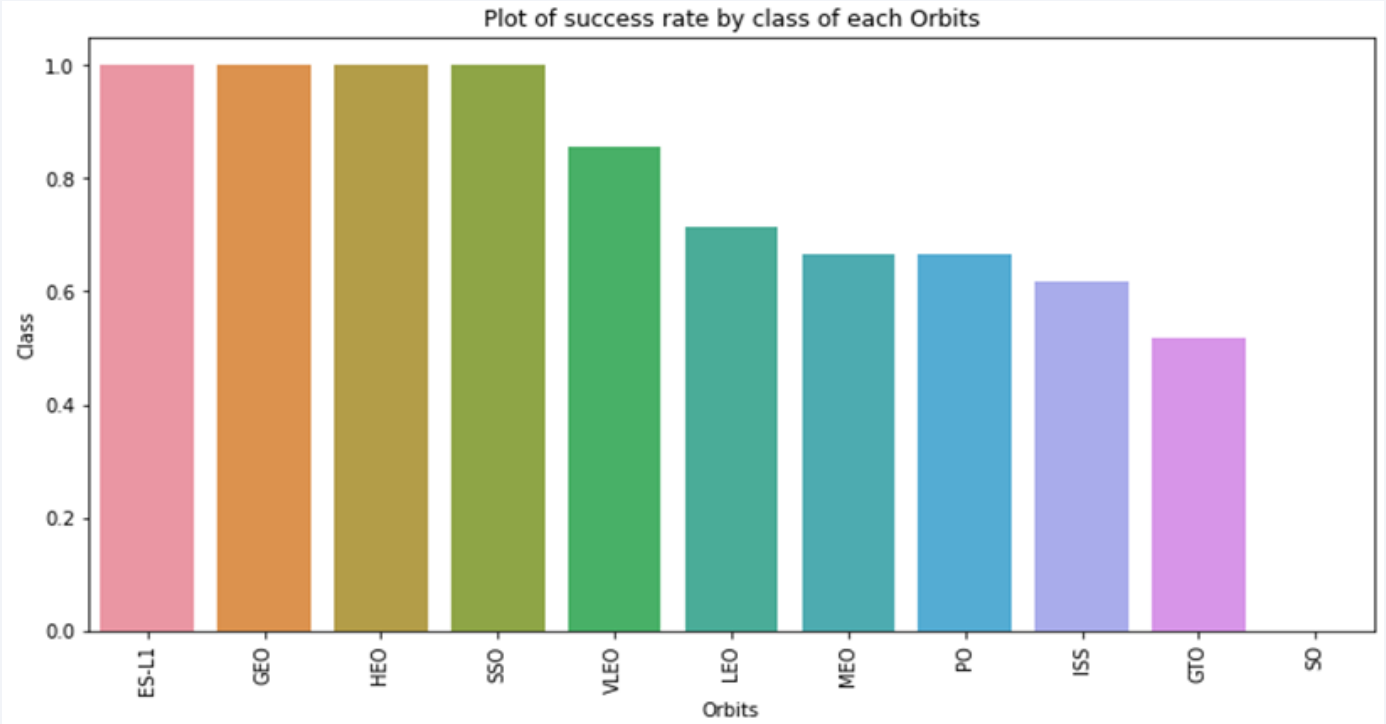
Payload vs. Launch Site

The greater the payload mass for launch site CCAFS SLC 40 the higher the success rate for the rocket.



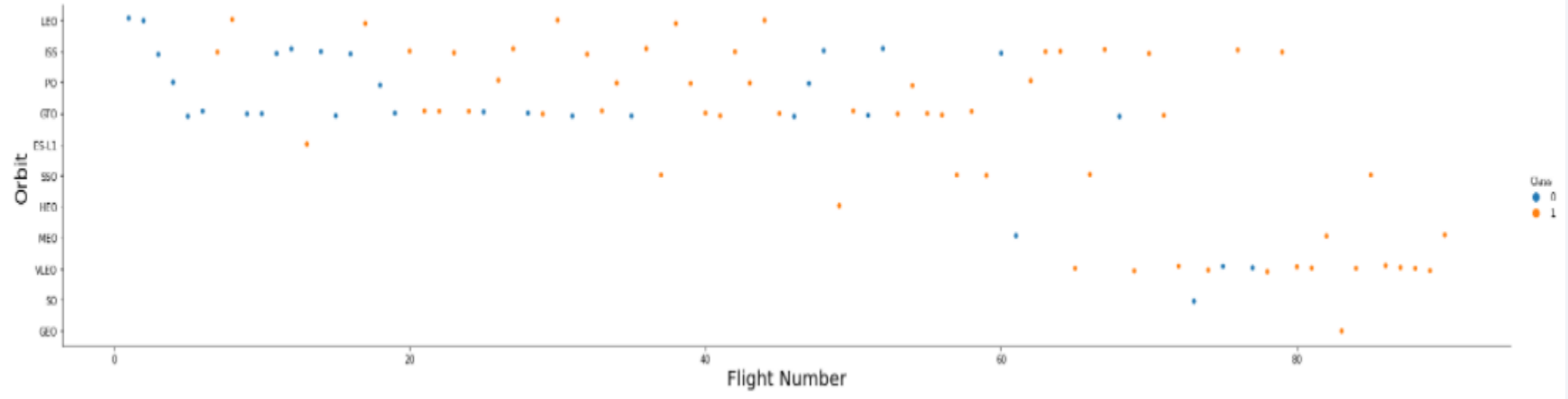
Success Rate vs. Orbit Type

- The plot shows that ES-L1, GEO, HEO, SSO, and VLEO achieved the highest success rates.



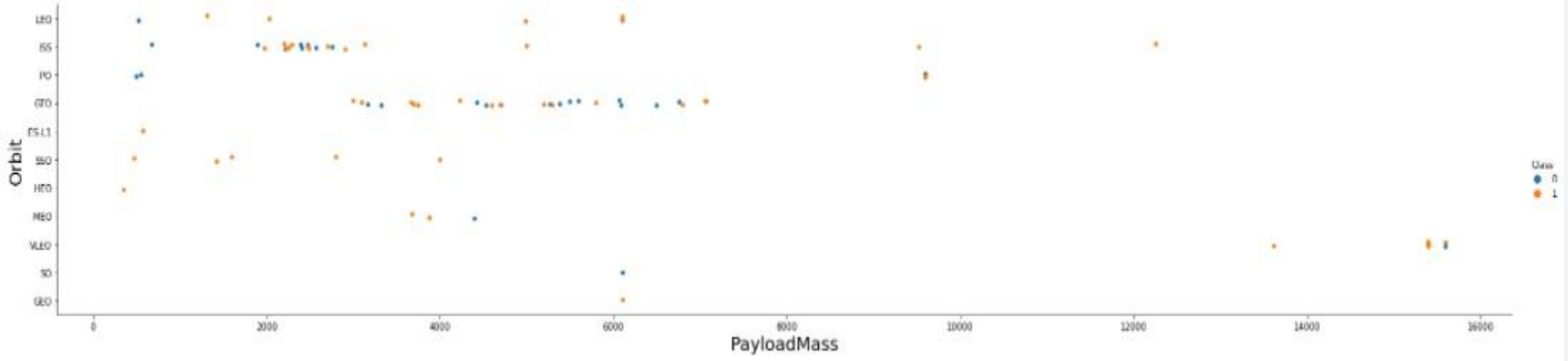
Flight Number vs. Orbit Type

- The plot illustrates the relationship between Flight Number and Orbit Type. In LEO orbits, success correlates with the number of flights, while in GTO orbits, no such relationship is observed.



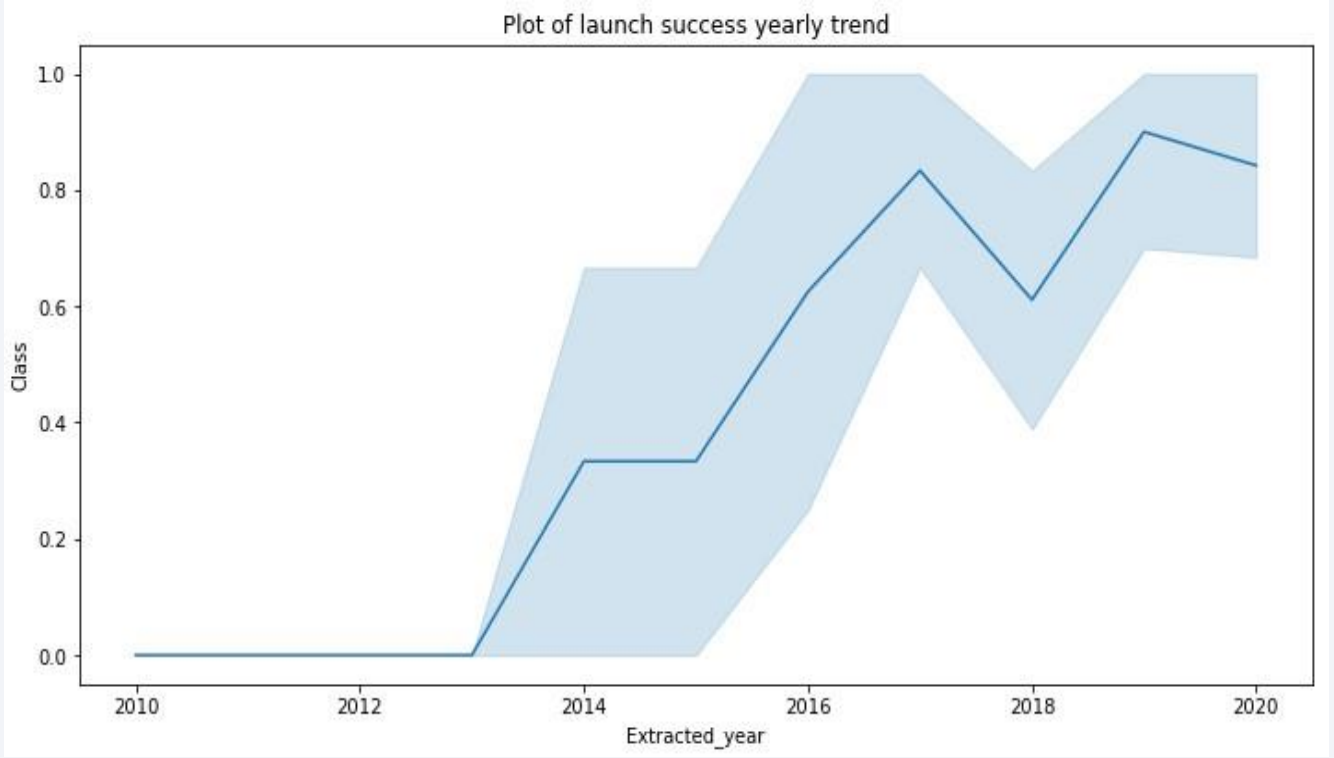
Payload vs. Orbit Type

- Heavy payloads show higher successful landing rates in PO, LEO, and ISS orbits.



Launch Success Yearly Trend

- The plot shows a steady increase in success rates from 2013 to 2020.



All Launch Site Names

- The **DISTINCT** keyword was used to display unique launch sites from the SpaceX data.

Display the names of the unique launch sites in the space mission

In [10]:
```
task_1 = '''
        SELECT DISTINCT LaunchSite
        FROM SpaceX
'''
create_pandas_df(task_1, database=conn)
```

Out[10]:

	launchsite
0	KSC LC-39A
1	CCAFS LC-40
2	CCAFS SLC-40
3	VAFB SLC-4E

Launch Site Names Begin with 'CCA'

Display 5 records where launch sites begin with the string 'CCA'

In [11]:
```
task_2 = '''
        SELECT *
        FROM SpaceX
        WHERE LaunchSite LIKE 'CCA%'
        LIMIT 5
        '''
create_pandas_df(task_2, database=conn)
```

Out[11]:

	date	time	boosterversion	launchsite	payload	payloadmasskg	orbit	customer	missionoutcome	landingoutcome
0	2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	2010-08-12	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2	2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
3	2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- The query was used to retrieve five records of launch sites starting with ‘CCA’.

Total Payload Mass

- The query below calculated the total payload carried by NASA boosters as 45596.

Display the total payload mass carried by boosters launched by NASA (CRS)

In [12]:
```
task_3 = '''
        SELECT SUM(PayloadMassKG) AS Total_PayloadMass
        FROM SpaceX
        WHERE Customer LIKE 'NASA (CRS)'
        '''
create_pandas_df(task_3, database=conn)
```

Out[12]:

	total_payloadmass
0	45596

Average Payload Mass by F9 v1.1

- The average payload mass for booster version F9 v1.1 was calculated as 2928.4.

Display average payload mass carried by booster version F9 v1.1

In [13]:
```
task_4 = '''
        SELECT AVG(PayloadMassKG) AS Avg_PayloadMass
        FROM SpaceX
        WHERE BoosterVersion = 'F9 v1.1'
        '''
create_pandas_df(task_4, database=conn)
```

Out[13]:

	avg_payloadmass
0	2928.4

First Successful Ground Landing Date

- The first successful ground pad landing occurred on December 22, 2015

In [14]:

```
task_5 = '''
        SELECT MIN(Date) AS FirstSuccessfull_landing_date
        FROM SpaceX
        WHERE LandingOutcome LIKE 'Success (ground pad)'
        '''
create_pandas_df(task_5, database=conn)
```

Out[14]:

	firstsuccessfull_landing_date
0	2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

In [15]:

```
task_6 = '''
        SELECT BoosterVersion
        FROM SpaceX
        WHERE LandingOutcome = 'Success (drone ship)'
            AND PayloadMassKG > 4000
            AND PayloadMassKG < 6000
        '''
create_pandas_df(task_6, database=conn)
```

Out[15]:

	boosterversion
0	F9 FT B1022
1	F9 FT B1026
2	F9 FT B1021.2
3	F9 FT B1031.2

- The **WHERE** clause was used to filter boosters that successfully landed on a drone ship, combined with the **AND** condition to select those with payload masses between 4000 and 6000

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

In [16]:
```
task_7a = '''
        SELECT COUNT(MissionOutcome) AS SuccessOutcome
        FROM SpaceX
        WHERE MissionOutcome LIKE 'Success%'
        '''

task_7b = '''
        SELECT COUNT(MissionOutcome) AS FailureOutcome
        FROM SpaceX
        WHERE MissionOutcome LIKE 'Failure%'
        '''
print('The total number of successful mission outcome is:')
display(create_pandas_df(task_7a, database=conn))
print()
print('The total number of failed mission outcome is:')
create_pandas_df(task_7b, database=conn)
```

```
The total number of successful mission outcome is:
```

	successoutcome
0	100

```
The total number of failed mission outcome is:
```

Out[16]:

	failureoutcome
0	1

- A wildcard '%' was used to filter **WHERE** MissionOutcome was either a success or a failure.

Boosters Carried Maximum Payload

- The booster carrying the maximum payload was identified using a subquery in the **WHERE** clause with the **MAX()** function.

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

In [17]:
```
task_8 = '''
        SELECT BoosterVersion, PayloadMassKG
        FROM SpaceX
        WHERE PayloadMassKG = (
                                SELECT MAX(PayloadMassKG)
                                FROM SpaceX
                                )
        ORDER BY BoosterVersion
        '''
create_pandas_df(task_8, database=conn)
```

Out[17]:

	boosterversion	payloadmasskg
0	F9 B5 B1048.4	15600
1	F9 B5 B1048.5	15600
2	F9 B5 B1049.4	15600
3	F9 B5 B1049.5	15600
4	F9 B5 B1049.7	15600
5	F9 B5 B1051.3	15600
6	F9 B5 B1051.4	15600
7	F9 B5 B1051.6	15600
8	F9 B5 B1056.4	15600
9	F9 B5 B1058.3	15600
10	F9 B5 B1060.2	15600
11	F9 B5 B1060.3	15600

2015 Launch Records

- A combination of **WHERE**, **LIKE**, **AND**, and **BETWEEN** conditions was used to filter failed drone ship landings, their booster versions, and launch site names for 2015.

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

In [18]:
```
task_9 = '''
        SELECT BoosterVersion, LaunchSite, LandingOutcome
        FROM SpaceX
        WHERE LandingOutcome LIKE 'Failure (drone ship)'
            AND Date BETWEEN '2015-01-01' AND '2015-12-31'
        '''
create_pandas_df(task_9, database=conn)
```

Out[18]:

	boosterversion	launchsite	landingoutcome
0	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
1	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad))

In [19]:
```
task_10 = '''
        SELECT LandingOutcome, COUNT(LandingOutcome)
        FROM SpaceX
        WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
        GROUP BY LandingOutcome
        ORDER BY COUNT(LandingOutcome) DESC
        '''
create_pandas_df(task_10, database=conn)
```

Out[19]:

	landingoutcome	count
0	No attempt	10
1	Success (drone ship)	6
2	Failure (drone ship)	5
3	Success (ground pad)	5
4	Controlled (ocean)	3
5	Uncontrolled (ocean)	2
6	Precluded (drone ship)	1
7	Failure (parachute)	1

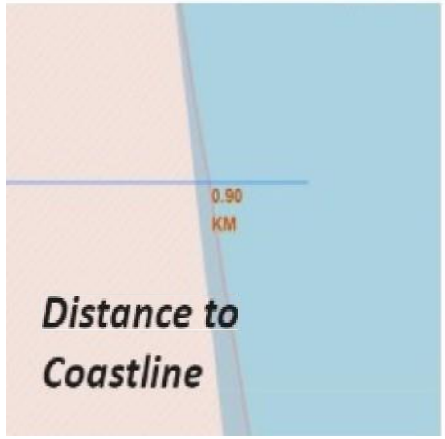
- LandingOutcome and its **COUNT** were selected, using the **WHERE** clause to filter outcomes between 2010-06-04 and 2010-03-20.
- The **GROUP BY** clause was applied to group LandingOutcome, and the **ORDER BY** clause sorted the groups in descending order.

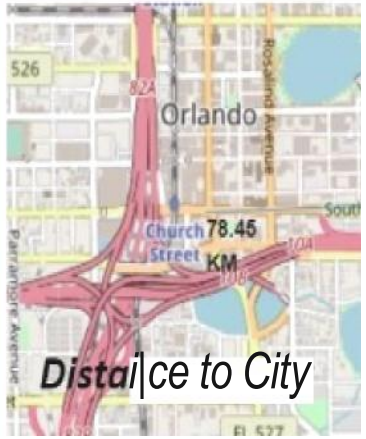

Section 3
Launch Sites
Proximities Analysis

All launch sites global map markers

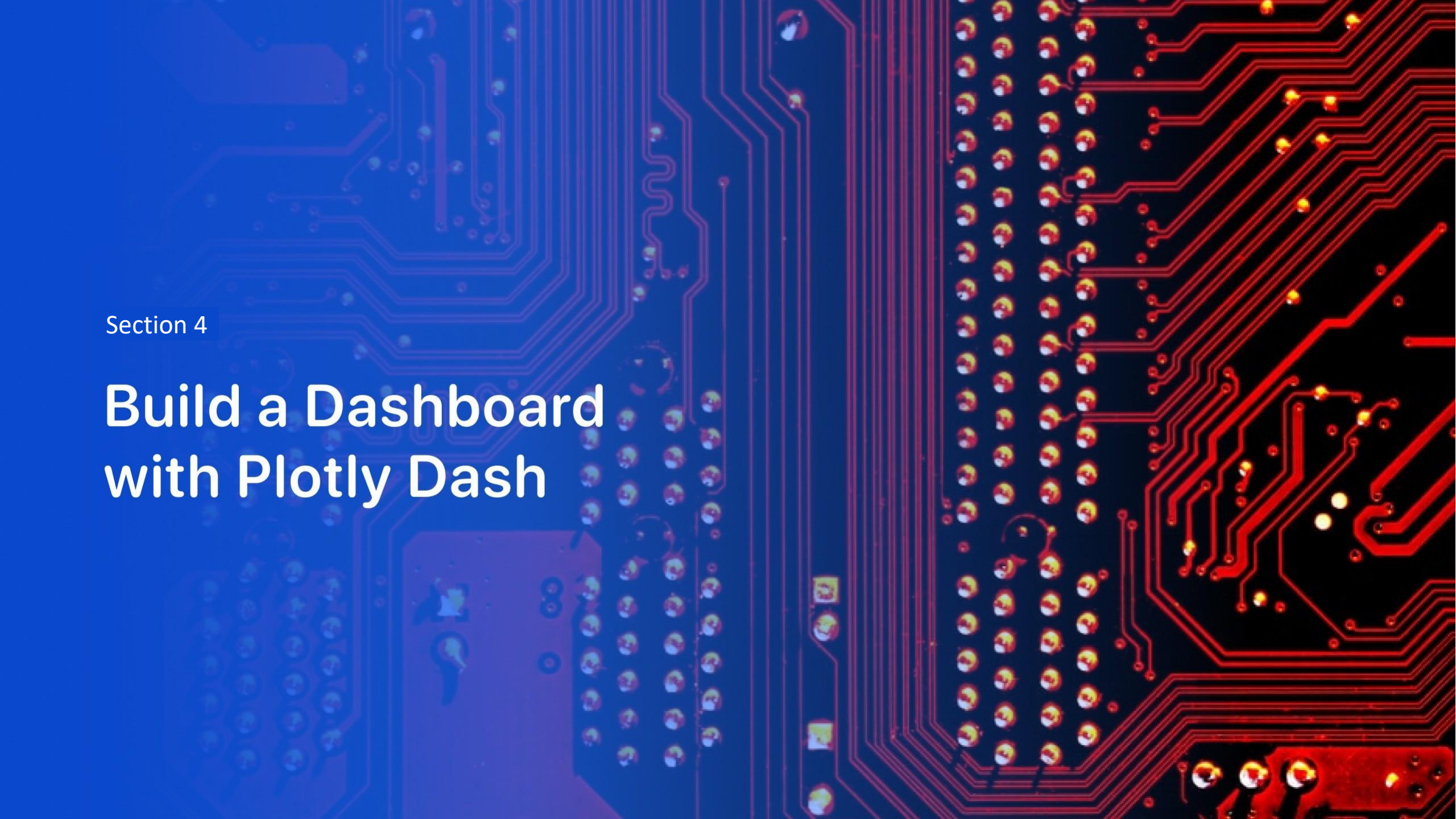
Markers showing launch sites with color labels

Launch Site distance to landmarks

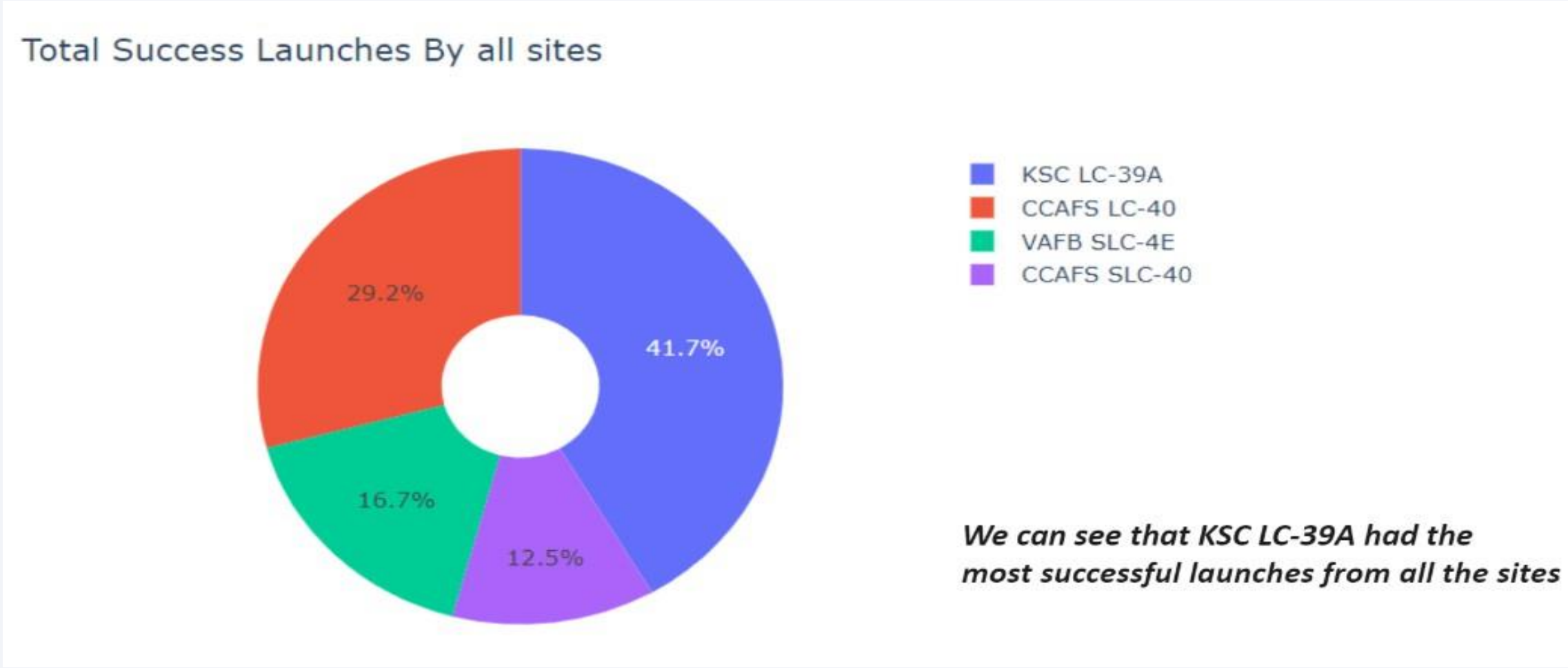




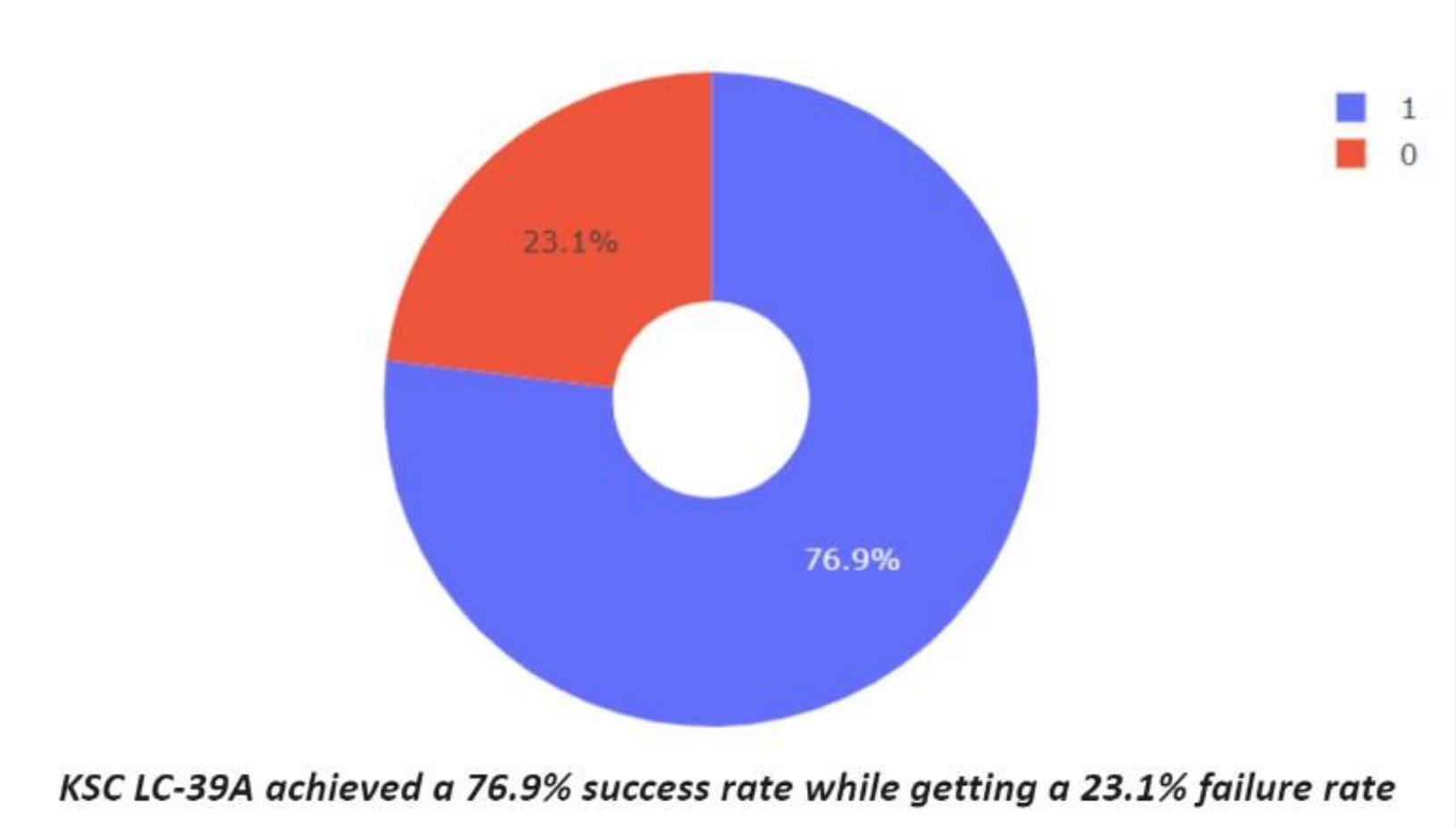
- Are launch sites near railways? No
- Are launch sites near highways? No
- Are launch sites near the coastline? Yes
- Do launch sites maintain a certain distance from cities? Yes


Section 4
Build a Dashboard with Plotly Dash

A pie chart illustrating the success percentage for each launch site.

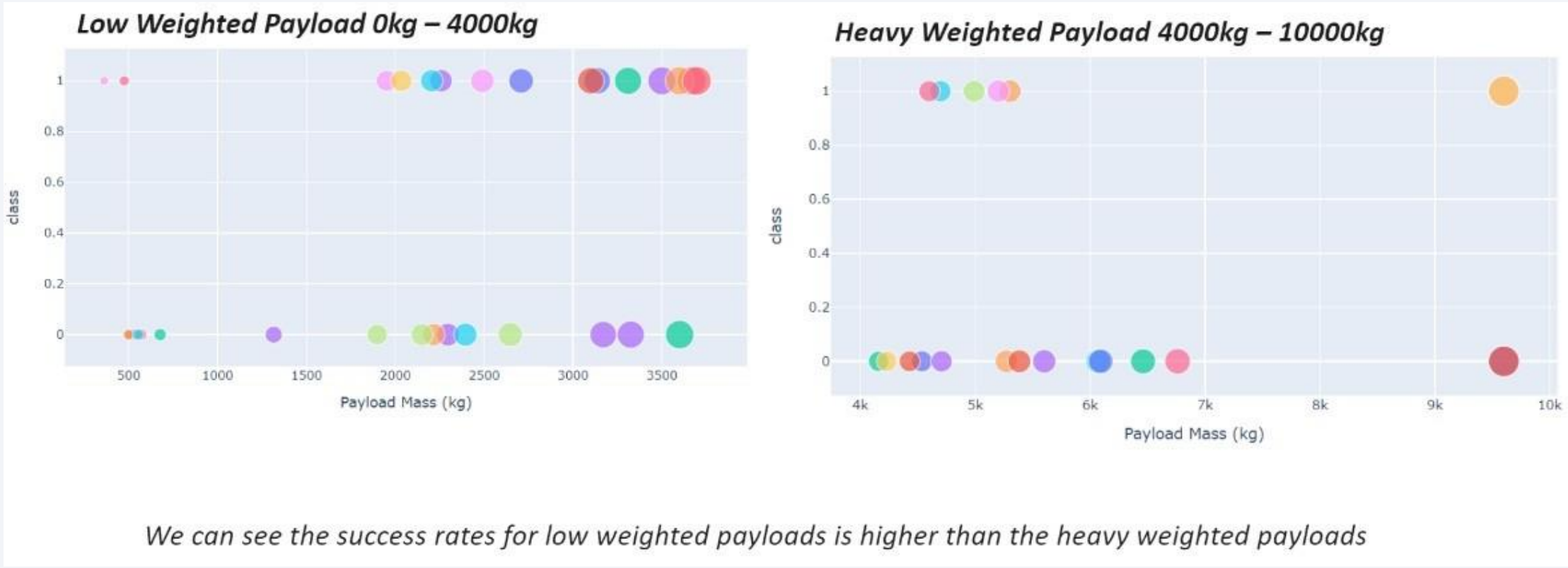


A pie chart displaying the launch site with the highest success ratio.



KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

A scatter plot of Payload vs. Launch Outcome for all sites, with payload ranges adjustable via a slider.



We can see the success rates for low weighted payloads is higher than the heavy weighted payloads

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- The decision tree classifier achieved the highest classification accuracy.

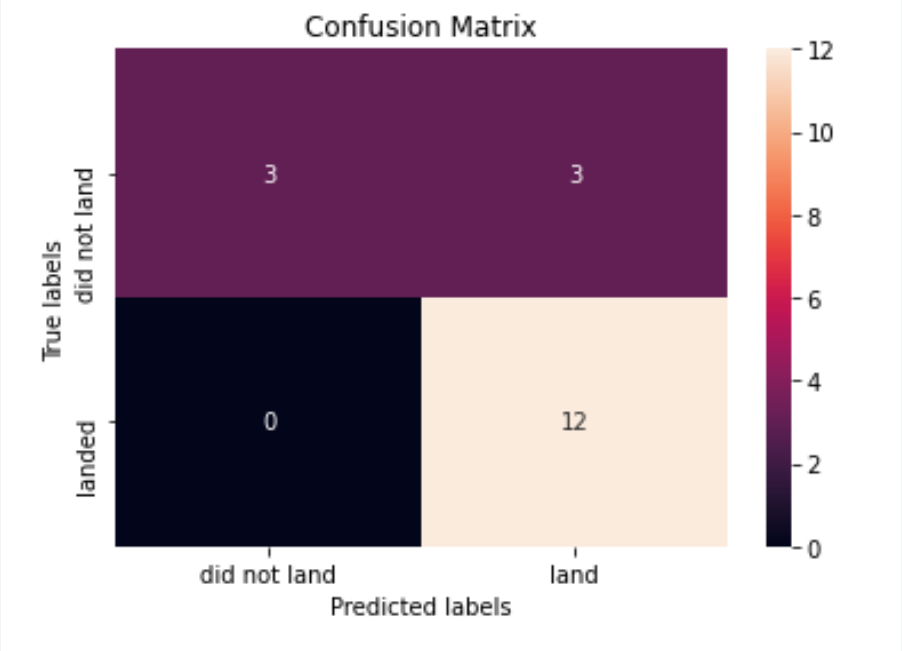
```
models = {'KNeighbors':knn_cv.best_score_,
              'DecisionTree':tree_cv.best_score_,
              'LogisticRegression':logreg_cv.best_score_,
              'SupportVector': svm_cv.best_score_}

bestalgorithm = max(models, key=models.get)
print('Best model is', bestalgorithm,'with a score of', models[bestalgorithm])
if bestalgorithm == 'DecisionTree':
    print('Best params is :', tree_cv.best_params_)
if bestalgorithm == 'KNeighbors':
    print('Best params is :', knn_cv.best_params_)
if bestalgorithm == 'LogisticRegression':
    print('Best params is :', logreg_cv.best_params_)
if bestalgorithm == 'SupportVector':
    print('Best params is :', svm_cv.best_params_)
```

```
Best model is DecisionTree with a score of 0.8732142857142856
Best params is : {'criterion': 'gini', 'max_depth': 6, 'max_features': 'auto', 'min_samples_leaf': 2, 'min_samples_split': 5, 'splitter': 'random'}
```

Confusion Matrix

- The confusion matrix for the decision tree classifier indicates it can differentiate between classes, but it struggles with false positives, marking unsuccessful landings as successful.



Conclusions

In conclusion:

- Higher flight volumes at a launch site correspond to greater success rates.
- Launch success rates steadily increased from 2013 to 2020.
- The orbits ES-L1, GEO, HEO, SSO, and VLEO achieved the highest success rates.
- KSC LC-39A recorded the most successful launches among all sites.
- The decision tree classifier proved to be the most effective machine learning algorithm for this task.

Thank you!