Winning Space Race with Data Science

Vincent Tsang
11 DEC 2023

Outline

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

Executive Summary

- Summary of methodologies
  - Launch Data from SpaceX's API and Wikipedia have been collected and wrangled.
  - Exploratory data analysis (EDA) and visualization have been performed on the data.
  - Predictive analysis has been performed on the data to predict success rate.
- Summary of all results
  - Launch site KSC LC-39A has overall 76.9% of success rate.
  - Success rate is the highest when payload mass is between 2000 kg and 6000 kg.
  - Booster version FT has the highest success rate.
  - The predicted success rate (Decision Tree Classification) is 83.33%, which is similar to the trend of success rate in recent years.

Introduction

- Project background and context
  - SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.
  - If we can determine if the first stage will land, we can determine the cost of a launch.
- Problems you want to find answers
  - Predict if the Falcon 9 first stage will land successfully

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Collect launch data by calling SpaceX API and performing web scraping on Wikipedia
- Perform data wrangling
  - Calculate the number of launches and occurrence of mission outcome of the orbits
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Use GridSearchCV to compare the scores of 4 classification models and then choose 1

Data Collection – SpaceX API

- https://github.com/vincent168e/ibm_ds_final_proj/blob/main/1.1_jupyter-labs-spacex-data-collection-api.ipynb

1. Request and parse the SpaceX launch data
2. Normalize the Json file and turn it to Pandas data frame
3. Filter the data frame to only include Falcon 9 launches

Data Collection - Scraping

- https://github.com/vincent168e/ibm_ds_final_proj/blob/main/1.2_jupyter-labs-webscraping.ipynb

1. Request the Falcon9 Launch Wiki page from its URL
2. Create a Beautiful Soap object from the response
3. Extract all column/variable names from the HTML table header
4. Create a data frame by parsing the launch HTML tables

Data Wrangling

- https://github.com/vincent168e/ibm_ds_final_proj/blob/main/1.3_labs-jupyter-spacex-Data%20wrangling.ipynb

1. Calculate the number of launches on each site
2. Calculate the number and occurrence of each orbit
3. Calculate the number and occurence of mission outcome of the orbits
4. Create a landing outcome label from Outcome column

EDA with Data Visualization

- Scatter plot is mainly being used to find relationships between variables.
- Bar plot is being used to show success rate of different orbits.
- Line plot is being used to show success rate of different years.
- https://github.com/vincent168e/ibm_ds_final_proj/blob/main/2.2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- Run SQL queries to find out:
  - All Launch Site Names
  - Launch Site Names Begin with 'CCA'
  - Total Payload Mass
  - Average Payload Mass by F9 v1.1
  - First Successful Ground Landing Date
  - Successful Drone Ship Landing with Payload between 4000 and 6000
  - Total Number of Successful and Failure Mission Outcomes
  - Boosters Carried Maximum Payload
  - 2015 Launch Records
  - First Successful Ground Landing Date
  - Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- https://github.com/vincent168e/ibm_ds_final_proj/blob/main/2.1_jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Location markers, circles and popup labels have been added to a folium map to clearly locate the 4 launch sites.
- Cluster markers of launch details (success or not) have been added to show success rate at first glance.
- Distance lines between launch site and its proximities have been added.
- https://github.com/vincent168e/ibm_ds_final_proj/blob/main/3.1_lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- Pie chart of success rate of launch sites has been added to a dashboard with a dropdown list to select all sites or a specific site, and show the success cases (class) when clicking on a sector.
- Scatter chart of success rate of launch sites has been added to a dashboard with a dropdown list to select all sites or a specific site, and the range of payload mass, and then show the success cases (class) when clicking on a point.
- https://github.com/vincent168e/ibm_ds_final_proj/blob/main/3.2_spacex_dash_app.py

Predictive Analysis (Classification)

- Dataset has been split to train set and test set.
- 4 classification models have been used to fit and train the model, using GridSearchCV to find out the most suitable parameters.
- Decision Tree model has been selected as it has the highest accuracy score in this case. Confusion Matrix has been plotted.
- Further use Decision Tree to predict the success rate.
- https://github.com/vincent168e/ibm_ds_final_proj/blob/main/4.1_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

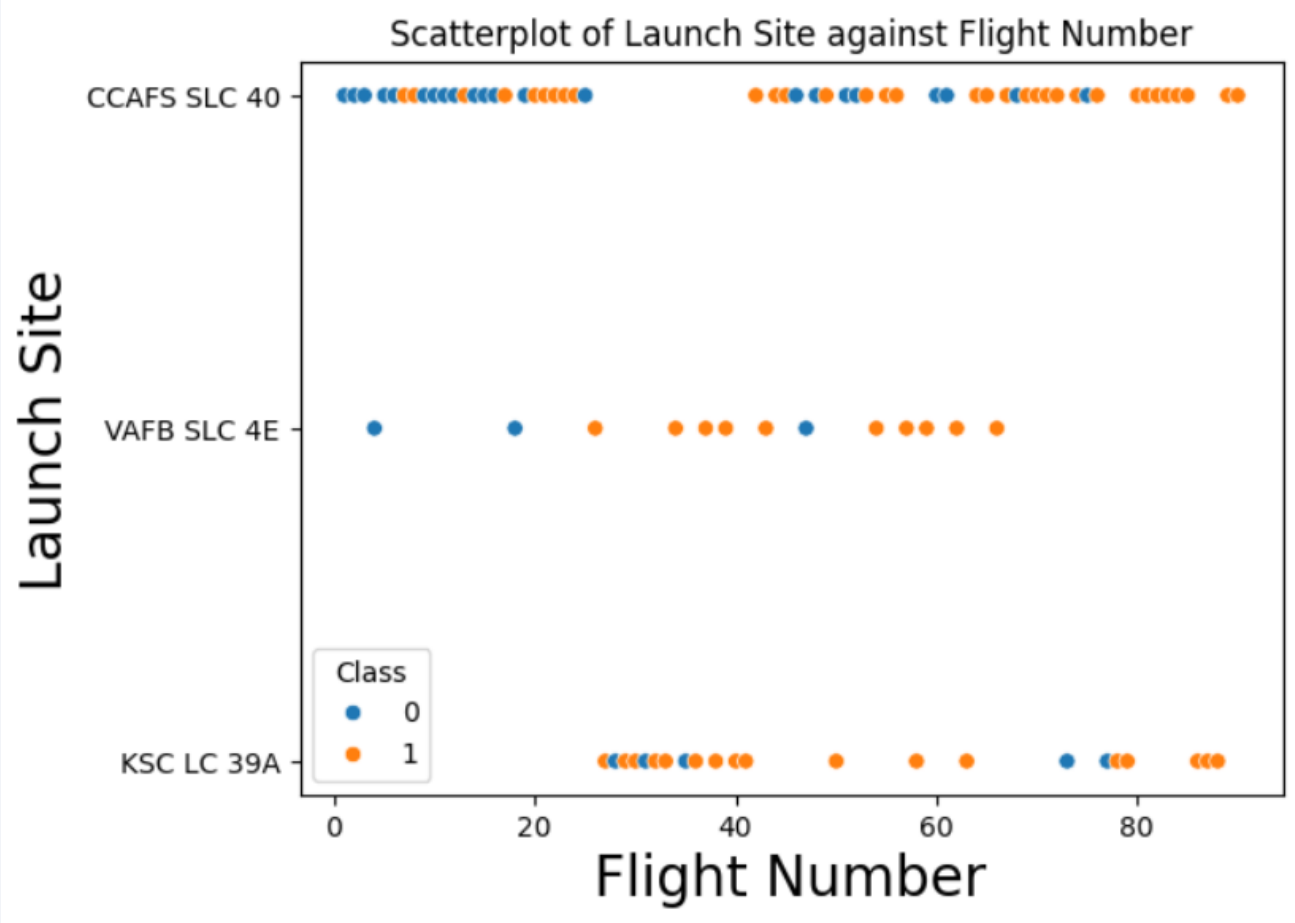
- Exploratory data analysis results (in next section)
- Interactive analytics demo in screenshots (in next section)
- Predictive analysis results (in next section)

Section 2

Insights drawn from EDA

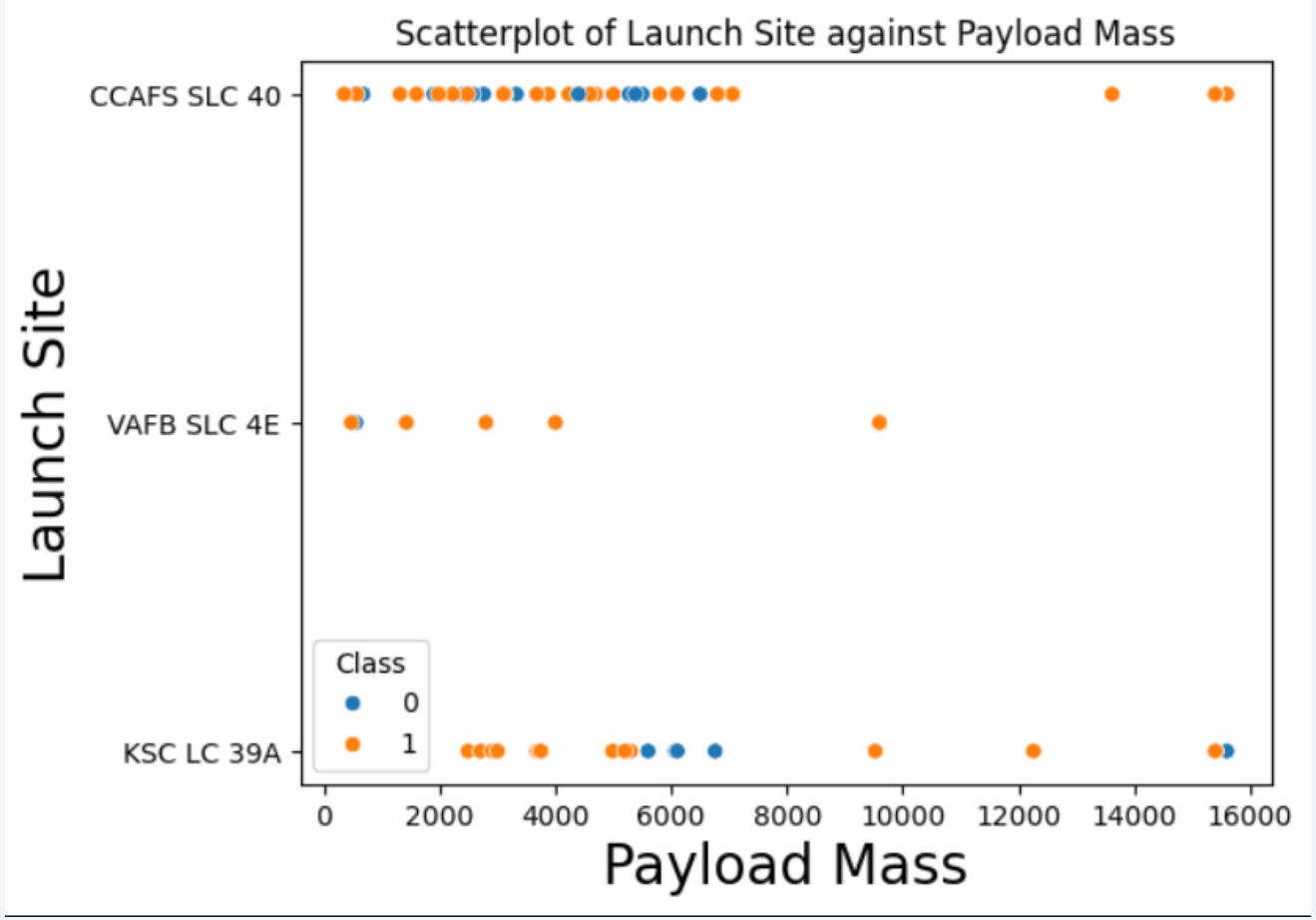
Flight Number vs. Launch Site

- At site CCAFS SLC 40, the success rate is:
  - Low when Flight Number < 30.
  - High when Flight Number > 70.
- At site VAFB SLC 4E, the success rate is:
  - Low when Flight Number < 20.
  - High when Flight Number is between 30 and 70.
- At site KSC LC 39A, the success rate is:
  - High when Flight Number is between 40 and 60 and Flight Number > 80.



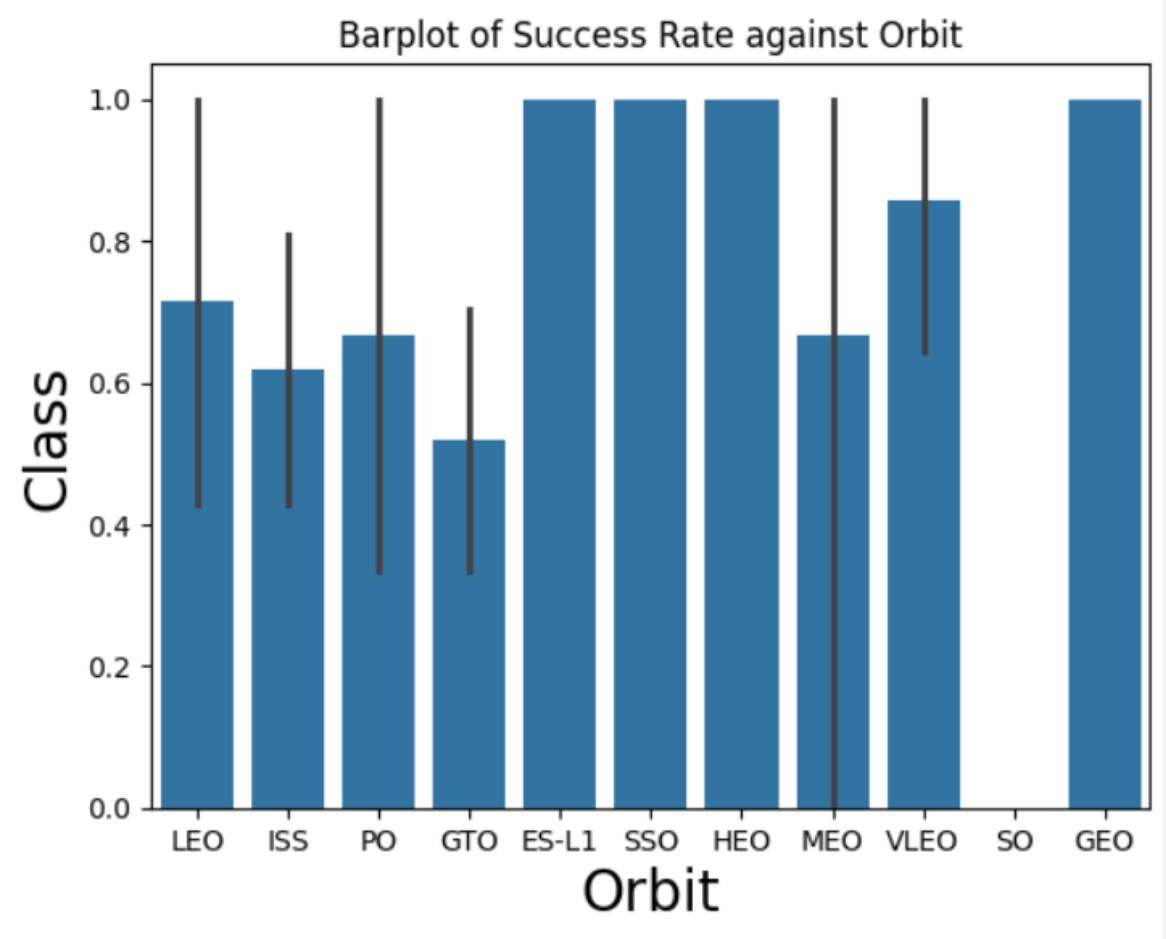
Payload vs. Launch Site

- For the VAFB-SLC launch site there are no rockets launched for heavy payload mass (> 10000).



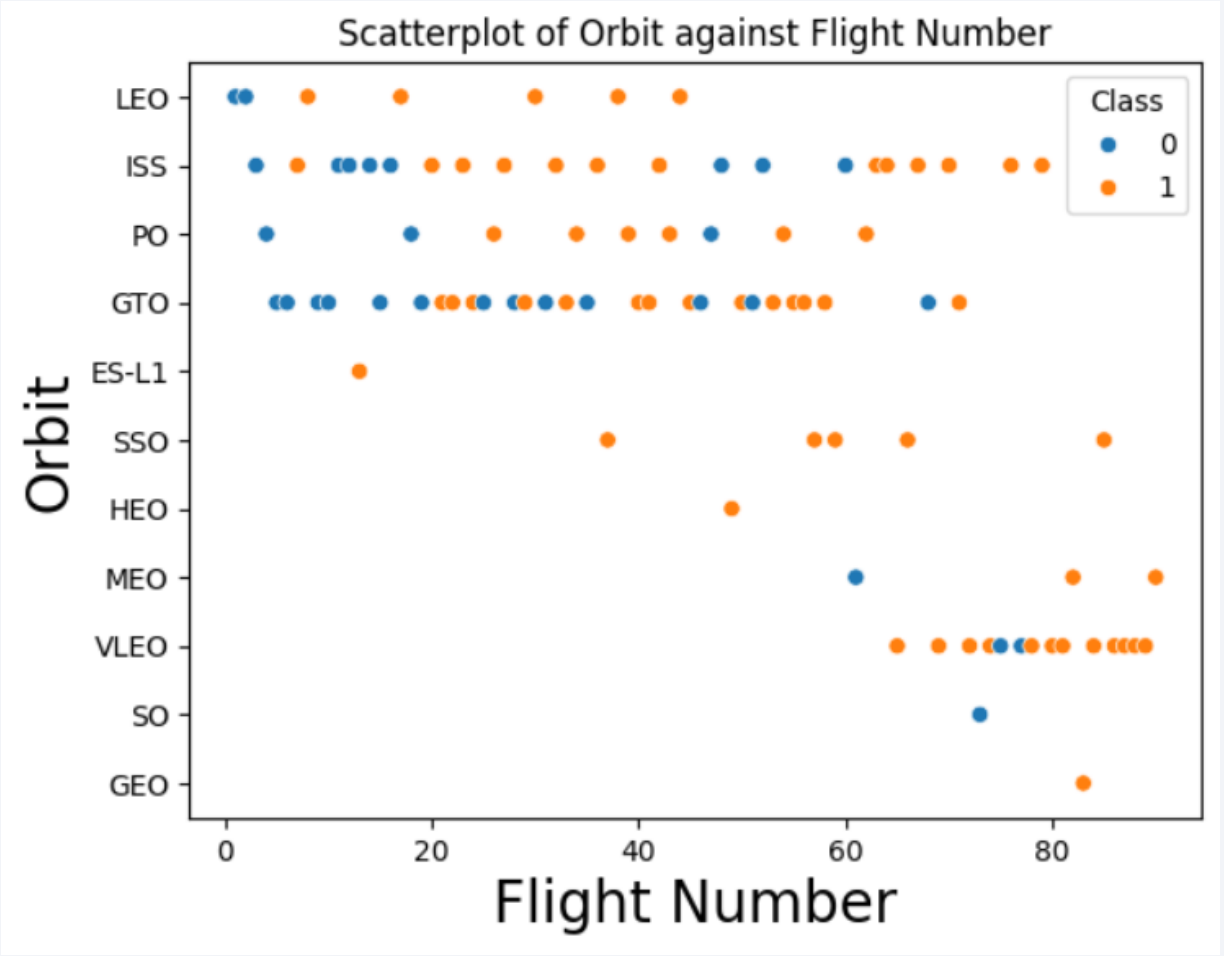
Success Rate vs. Orbit Type

- Orbit ES-L1, SSO, HEO and GEO have 100% success rate. VLEO has around 85% success rate.
- LEO, ISS, PO and MEO have around 60-70% success rate. GTO has around 50% success rate. SO has 0% success rate.



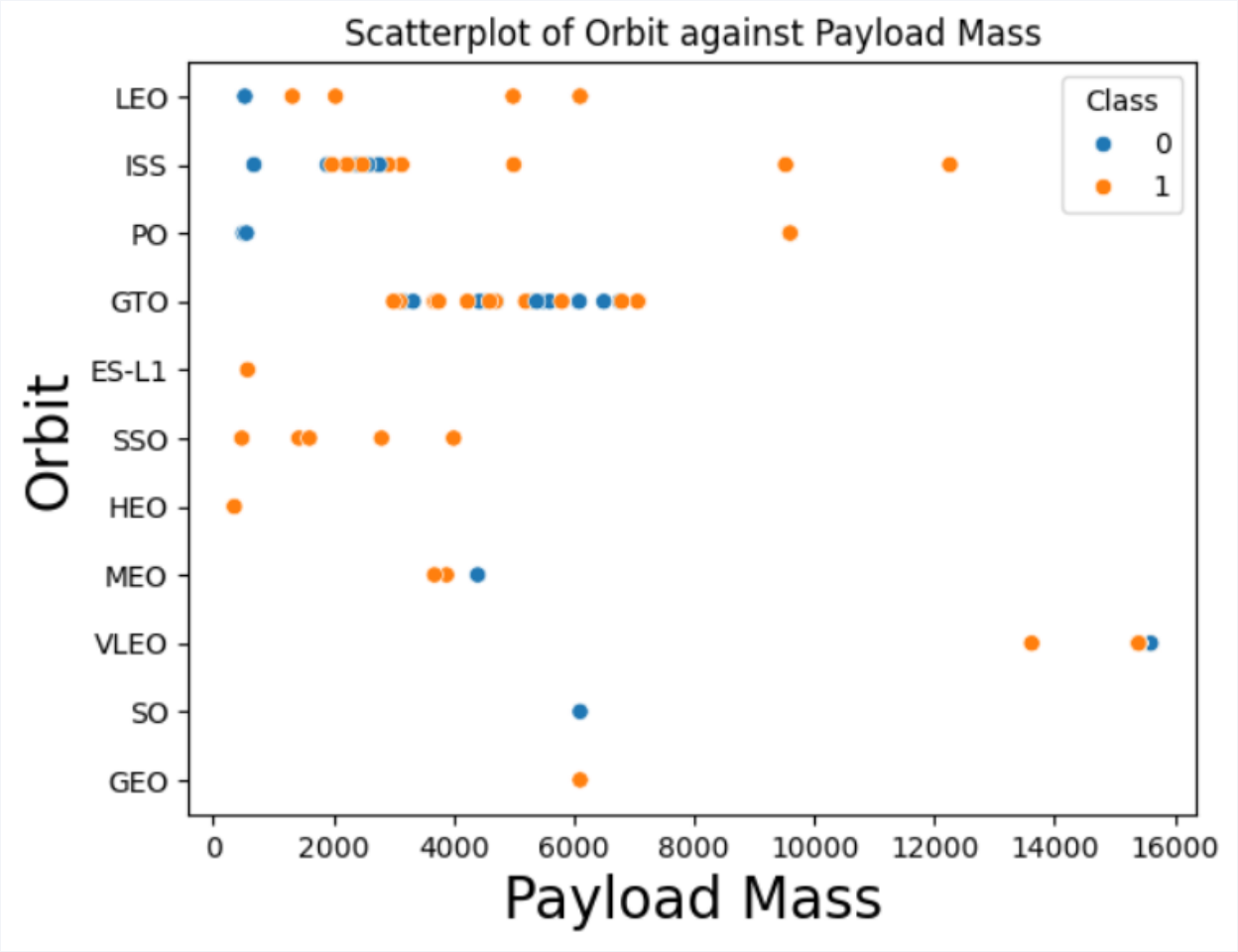
Flight Number vs. Orbit Type

- In LEO orbit, the success rate seems to be related to the flight number.
- In GTO orbit, there seems to be no relationship between flight number and orbit.



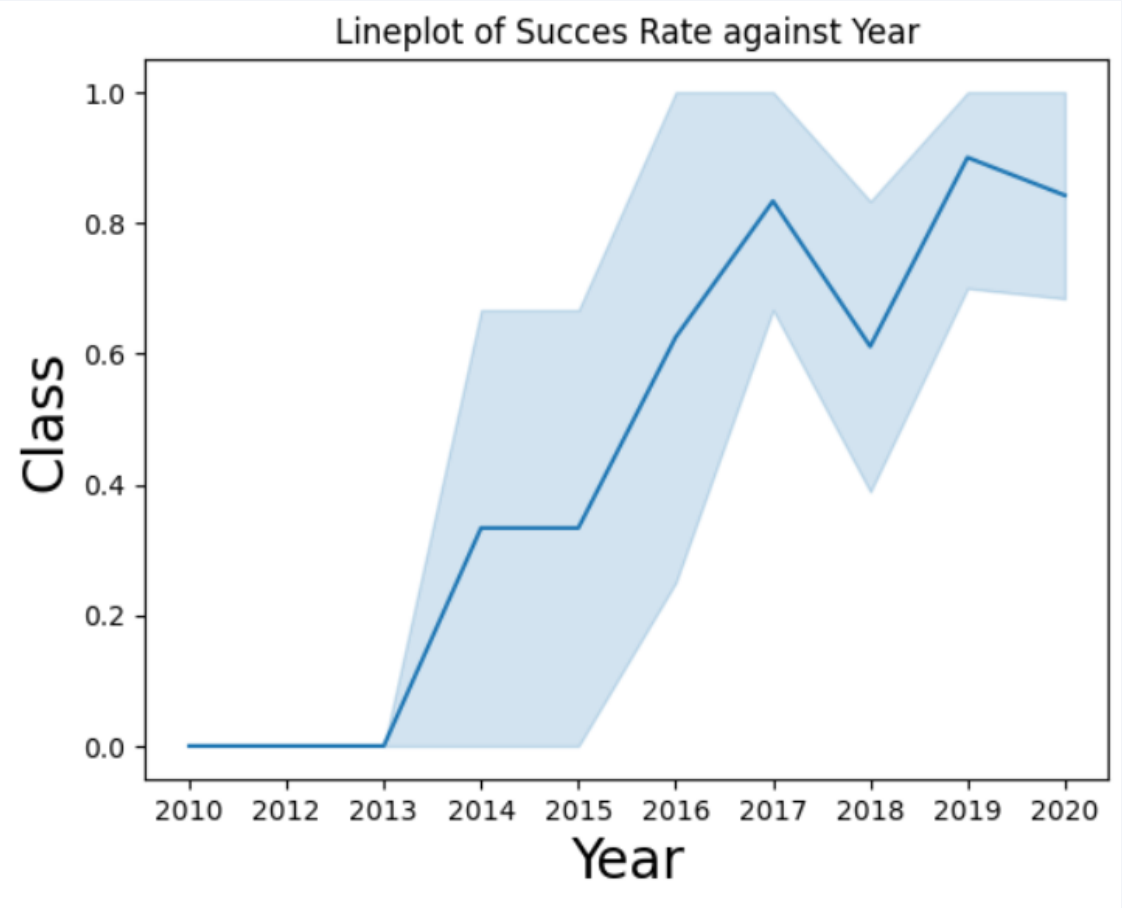
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- For GTO we cannot distinguish this well as both positive landing rate and negative landing (unsuccessful mission) are both here.



Launch Success Yearly Trend

- The success rate since 2013 kept increasing until 2020, exceeding 80%.



All Launch Site Names

- There are 4 unique launch sites:
  - CCAFS LC-40
  - CCAFS SLC-40
  - KSC LC-39A
  - VAFB SLC-4E

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

Launch Site Names Begin with 'CCA'

- Below were 5 records where launch sites begin with `CCA`, from 2010 to 2013.
- 2 were failed to land the 1st stage, while 3 have no attempt.

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- The total payload carried by boosters from NASA is 45596 kg.

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2534.67 kg.

avg(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

- The first successful landing outcome on ground pad was at 2015-12-22.

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000:
  - F9 FT B1021.2
  - F9 FT B1031.2
  - F9 FT B1022
  - F9 FT B1026

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

Total Number of Successful and Failure Mission Outcomes

- The total number of mission outcomes:
  - Successful: 100
  - Failed: 1

Successful Outcomes
100

Failed Outcomes
1

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass (15600 kg):
  - 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

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

- The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015 were at January and April, both at CCAFS LC-40 site.

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

- Rank of the count of landing outcomes between the date 2010-06-04 and 2017-03-20 were:
  - Most: No attempt, with 10
  - 2nd Most: Success (drone ship) and Failure (drone ship), with 5
  - Least: Precluded (drone ship), with 1

Landing_Outcome	count(Landing_Outcome)
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

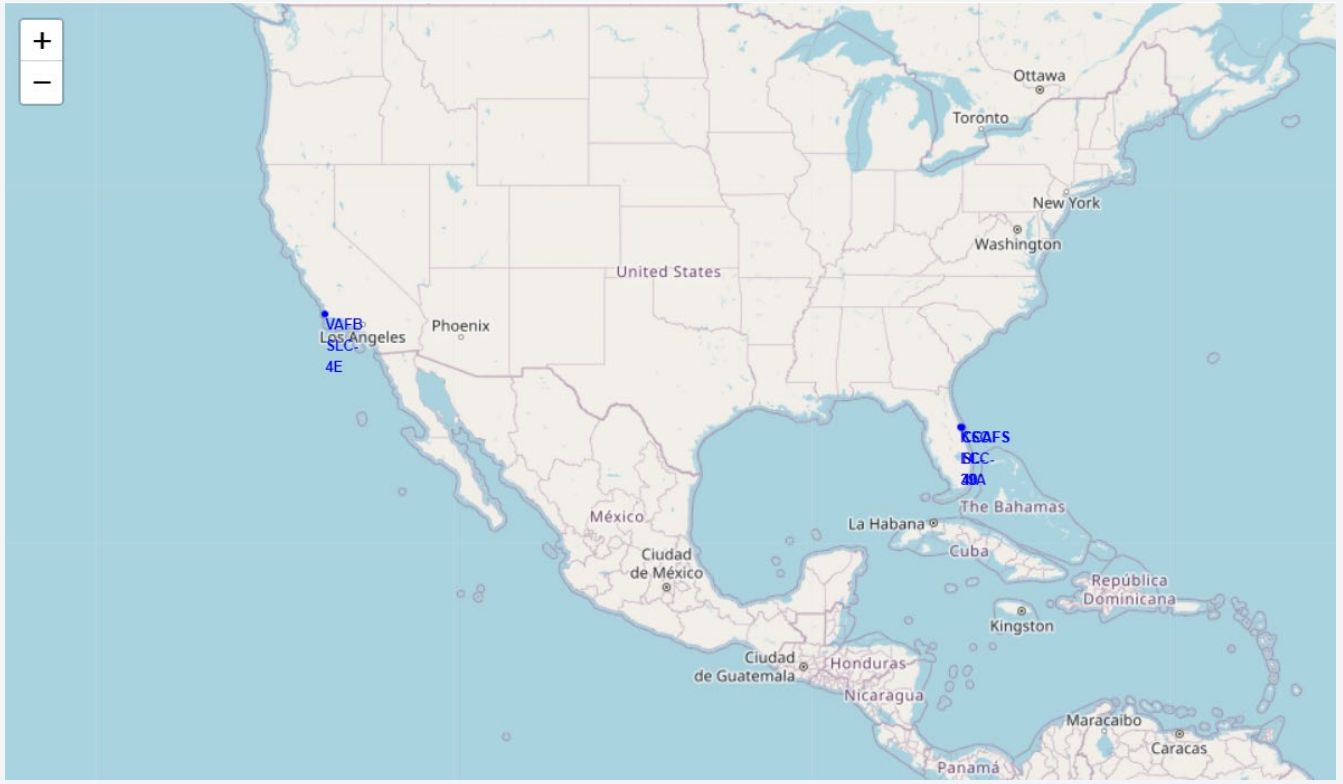
Section 3

Launch Sites Proximities Analysis

Launch Sites' Locations on Map

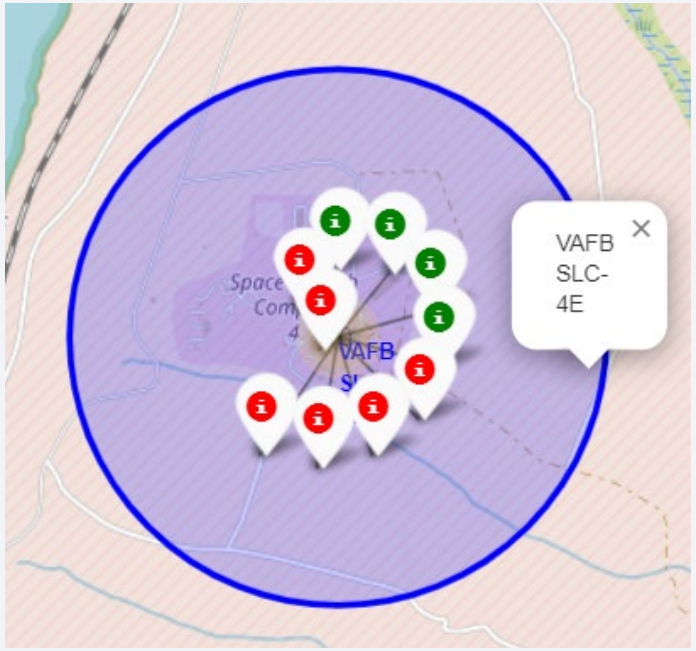
- All launch sites are not close to the Equator line.
- All launch sites are close to the coast.

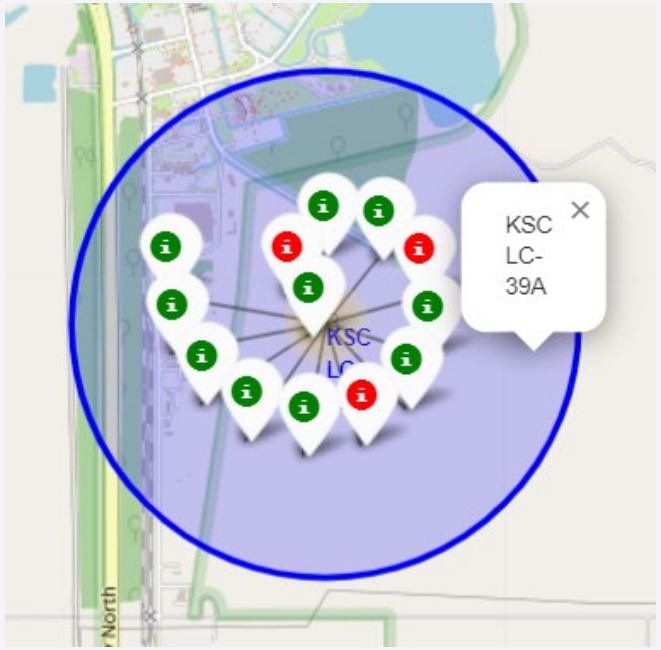
	Launch Site	Lat	Long
0	CCAFS LC-40	28.562302	-80.577356
1	CCAFS SLC-40	28.563197	-80.576820
2	KSC LC-39A	28.573255	-80.646895
3	VAFB SLC-4E	34.632834	-120.610745

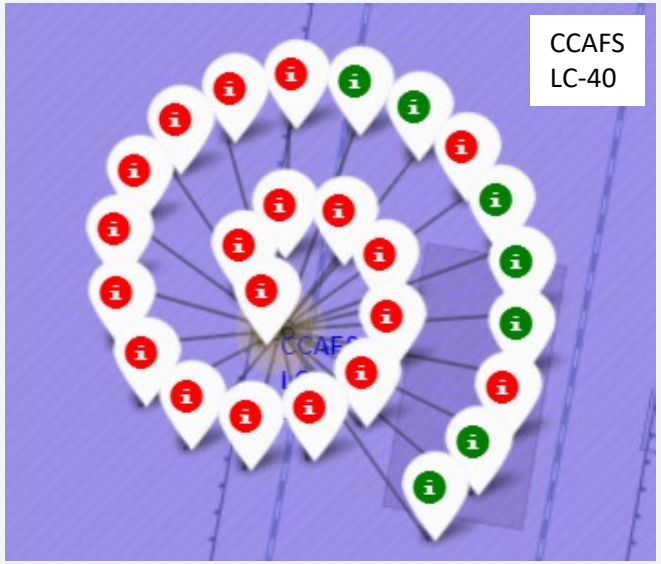


Outcomes Grouped by Launch Sites on Map

- Success rate at KSC LC-39A is the highest.
- Success rate at the other 3 sites are < 50%.

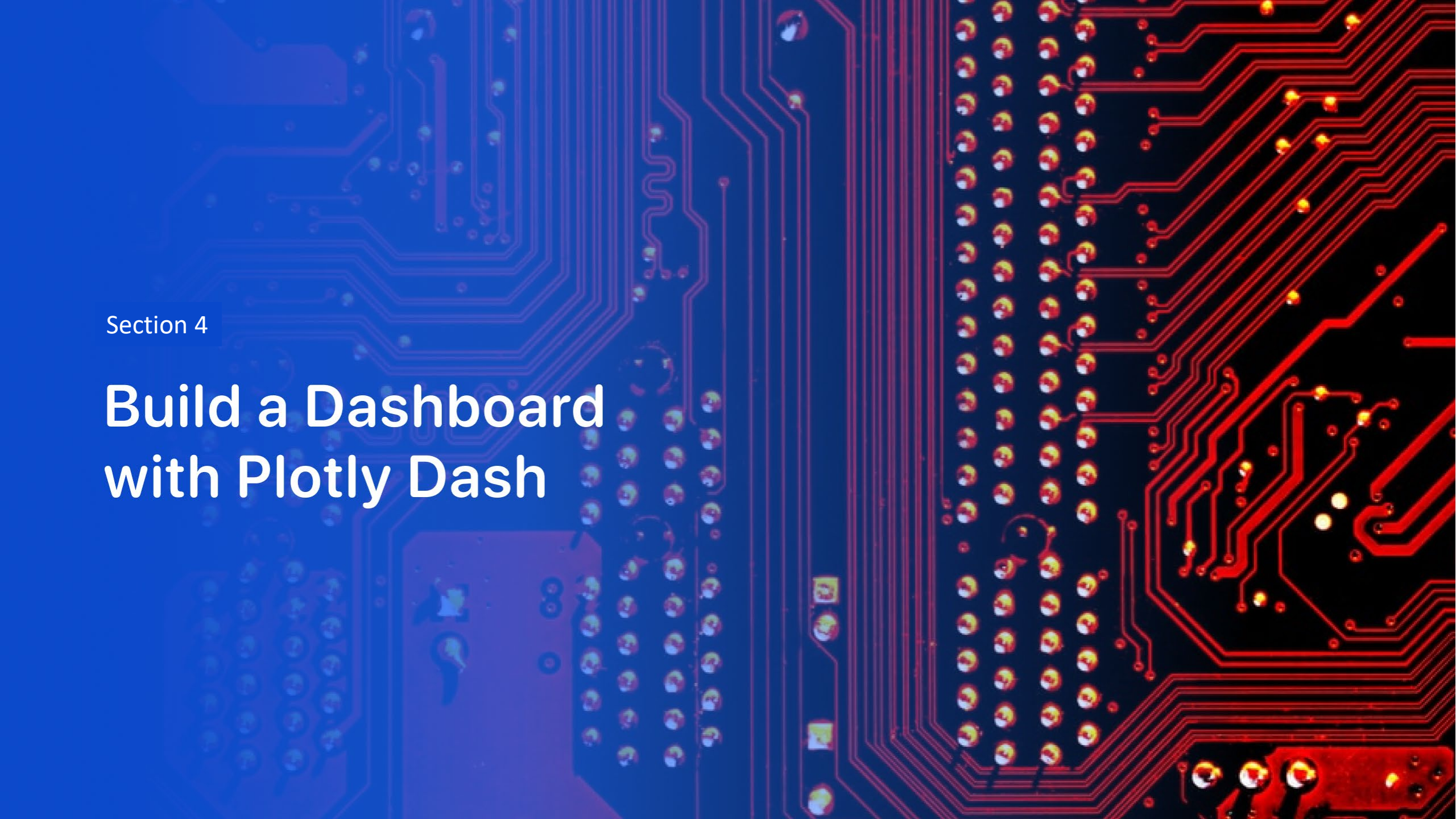




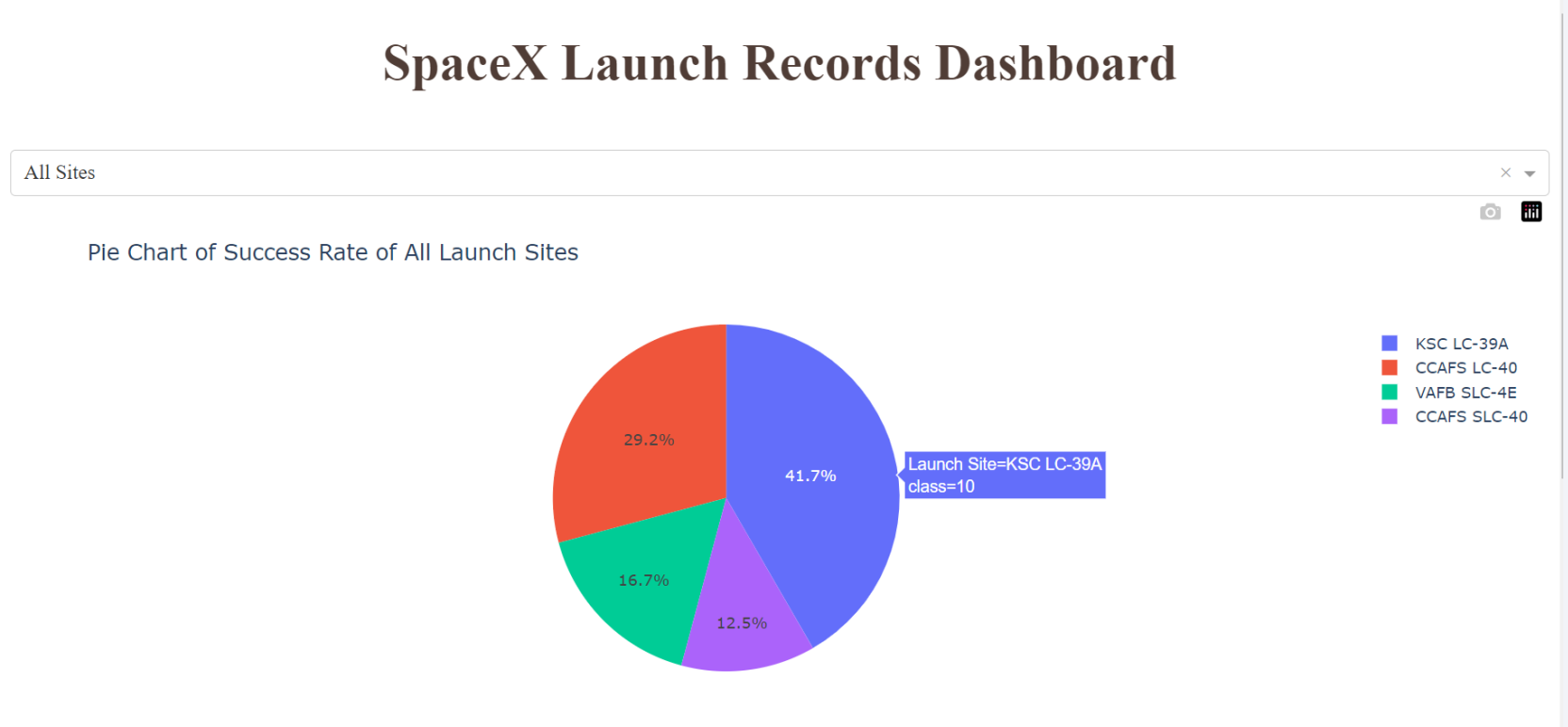


Launch Site (CCAFS LC-40) to its Proximities

- CCAFS LC-40 is
  - 0.99 km far from nearest railway
  - 0.58 km far from nearest highway
  - 0.86 km far from nearest coastline

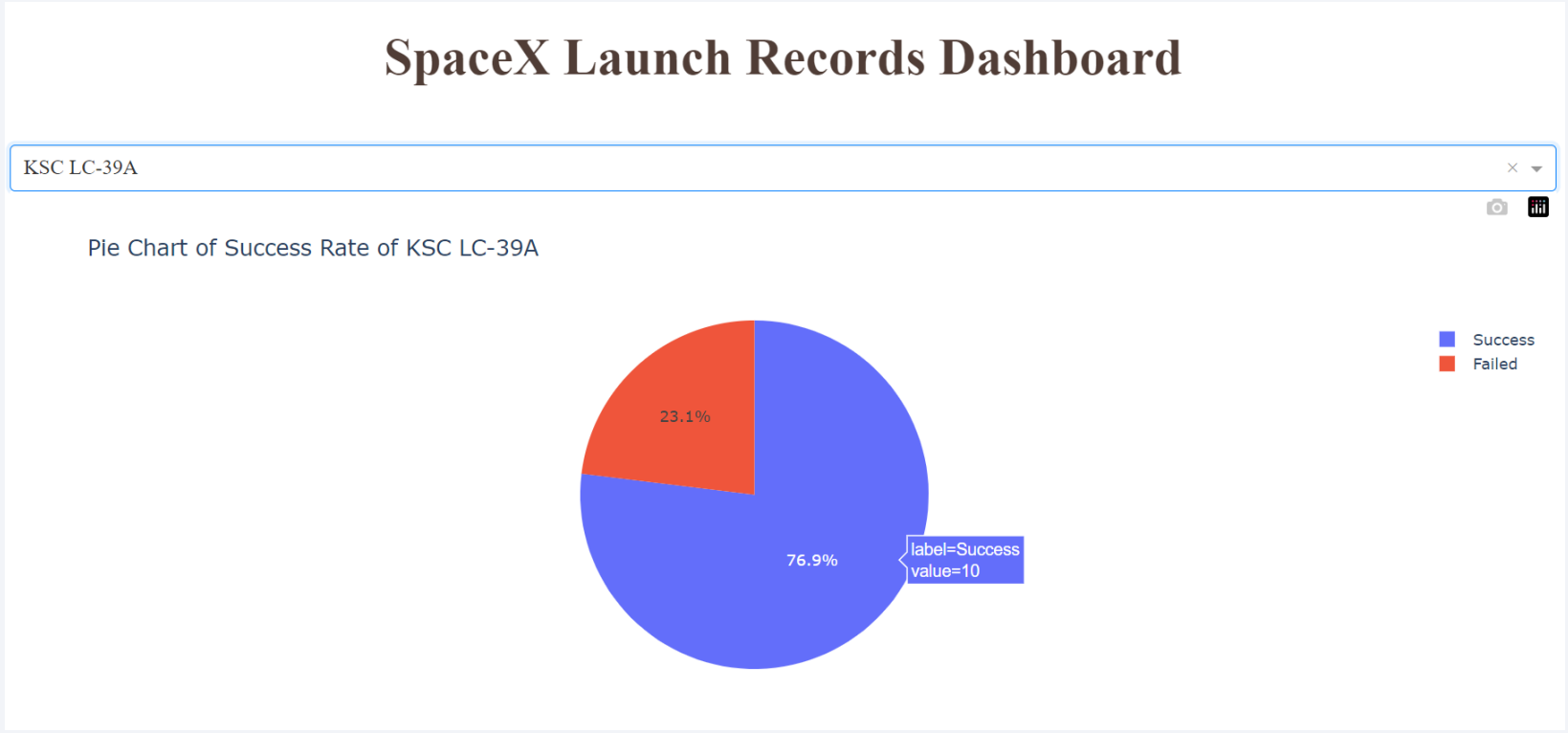

Section 4
Build a Dashboard with Plotly Dash

Pie Chart of Success Rate of All Launch Sites



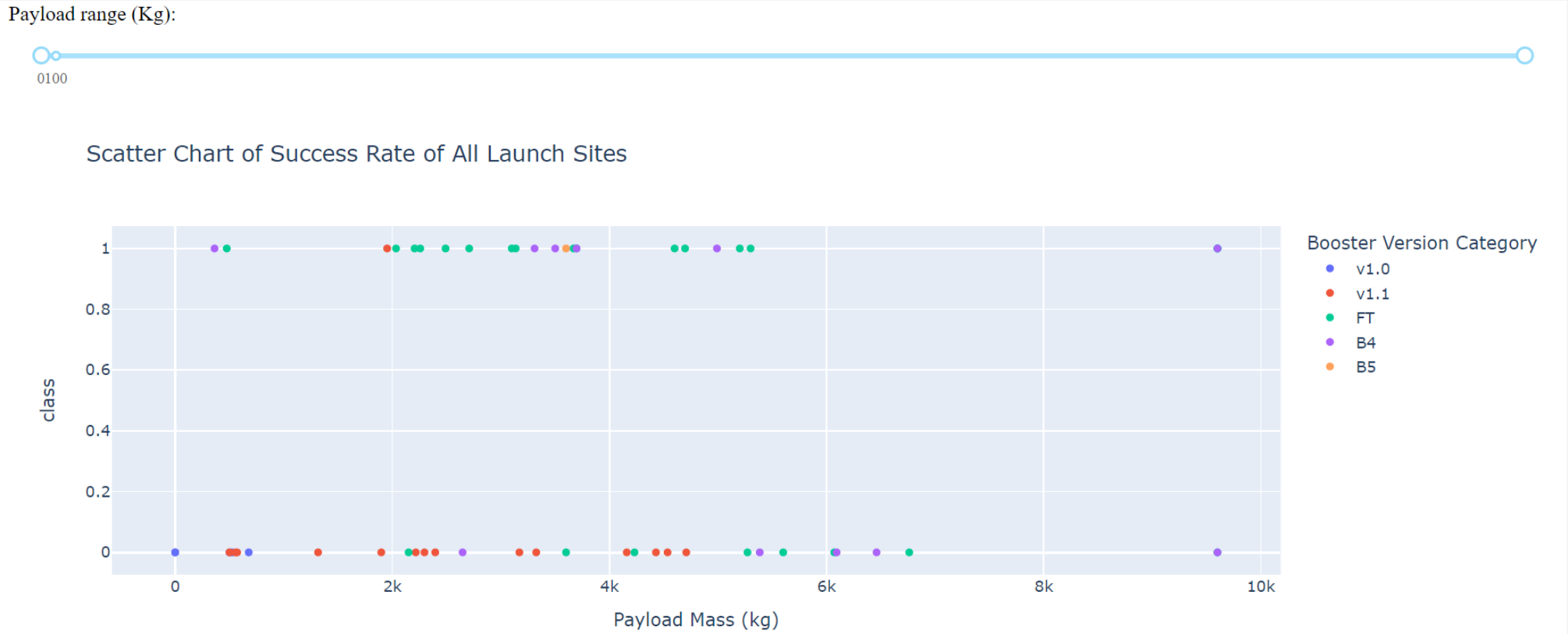
- Site KSC LC-39A has the largest number of successful outcomes (10), holding 41.7%.

Pie Chart of Success Rate of Site KSC LC-39A



- Site KSC LC-39A has overall 76.9% of success rate.

Scatter Chart of Success Rate of All Launch Sites



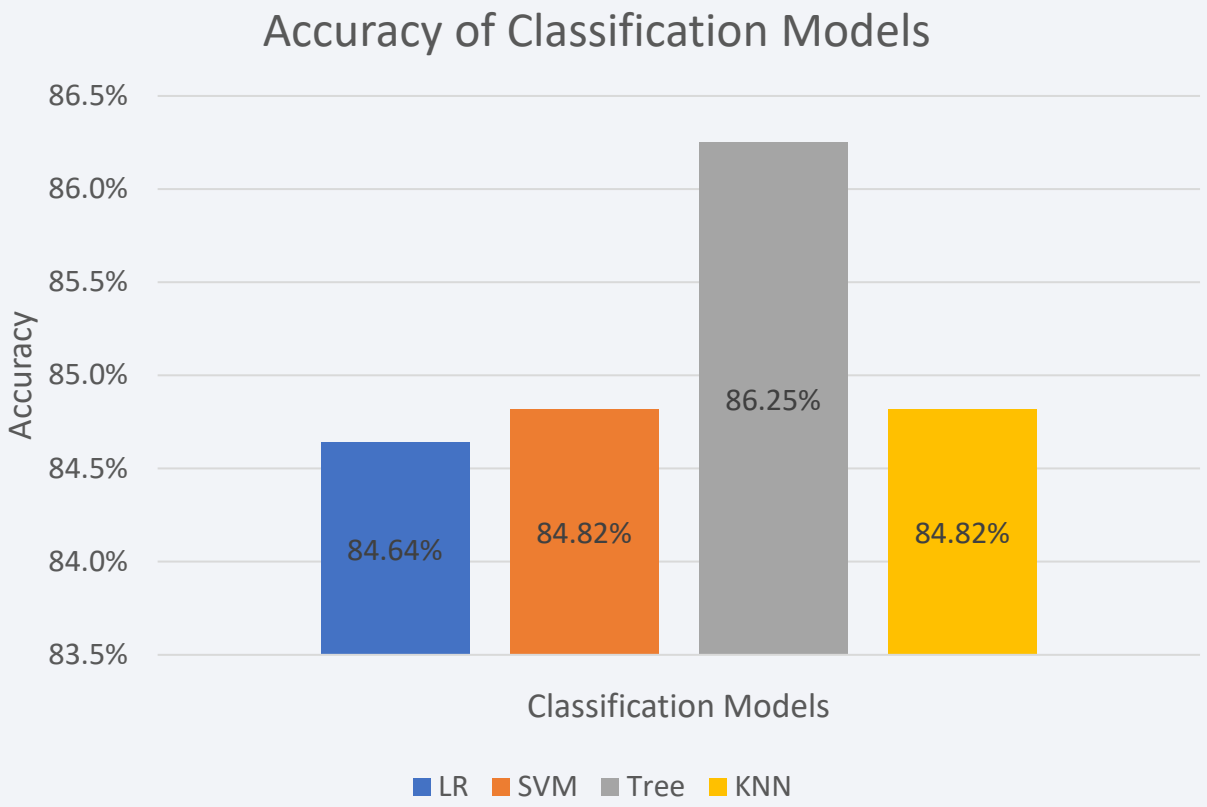
- Success rate is the highest when payload mass is between 2000 kg and 6000 kg.
- Booster version FT has the highest success rate.

Section 5

Predictive Analysis (Classification)

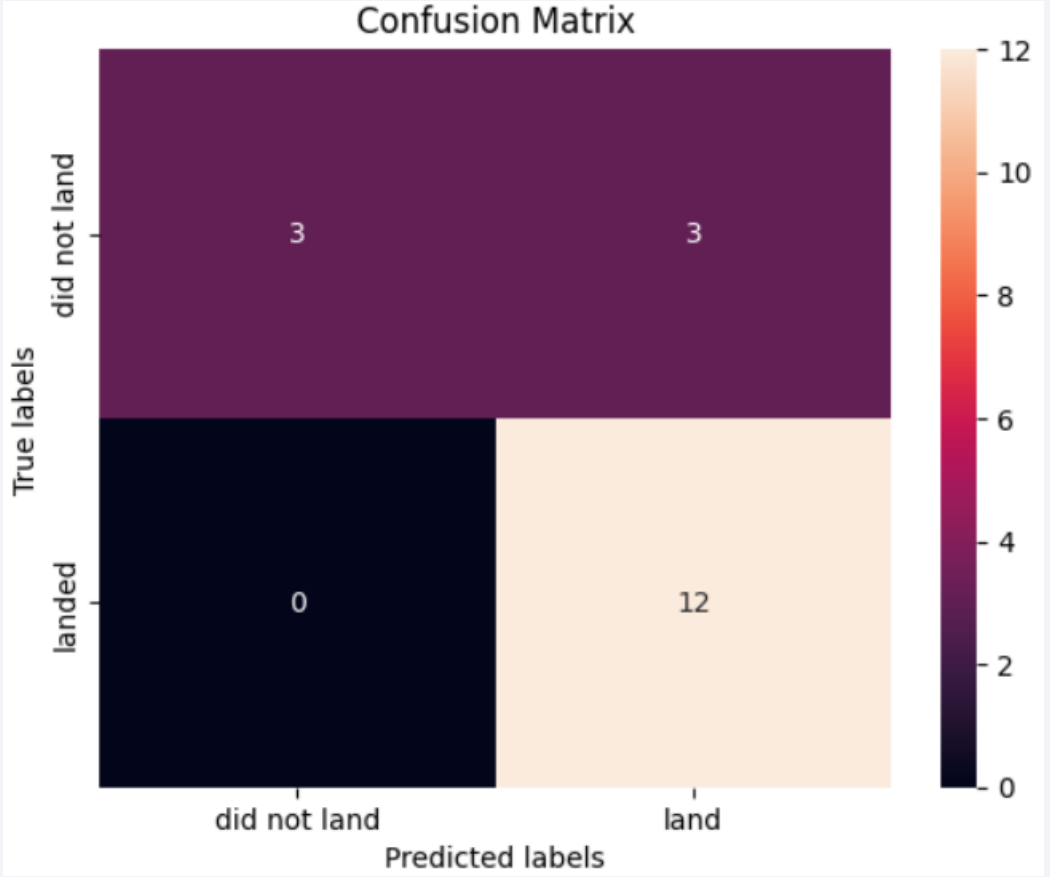
Classification Accuracy

- 4 classification models have been tested:
  - Logistic Regression (LR)
  - Support Vector Machine (SVM)
  - Decision Tree (Tree)
  - K-Nearest Neighbours (KNN)
- Decision Tree has the highest accuracy rate of 87.68%.



Confusion Matrix of Decision Tree Classification

- True Positive (land: 12) is high
- True Negative (not land: 3) is low

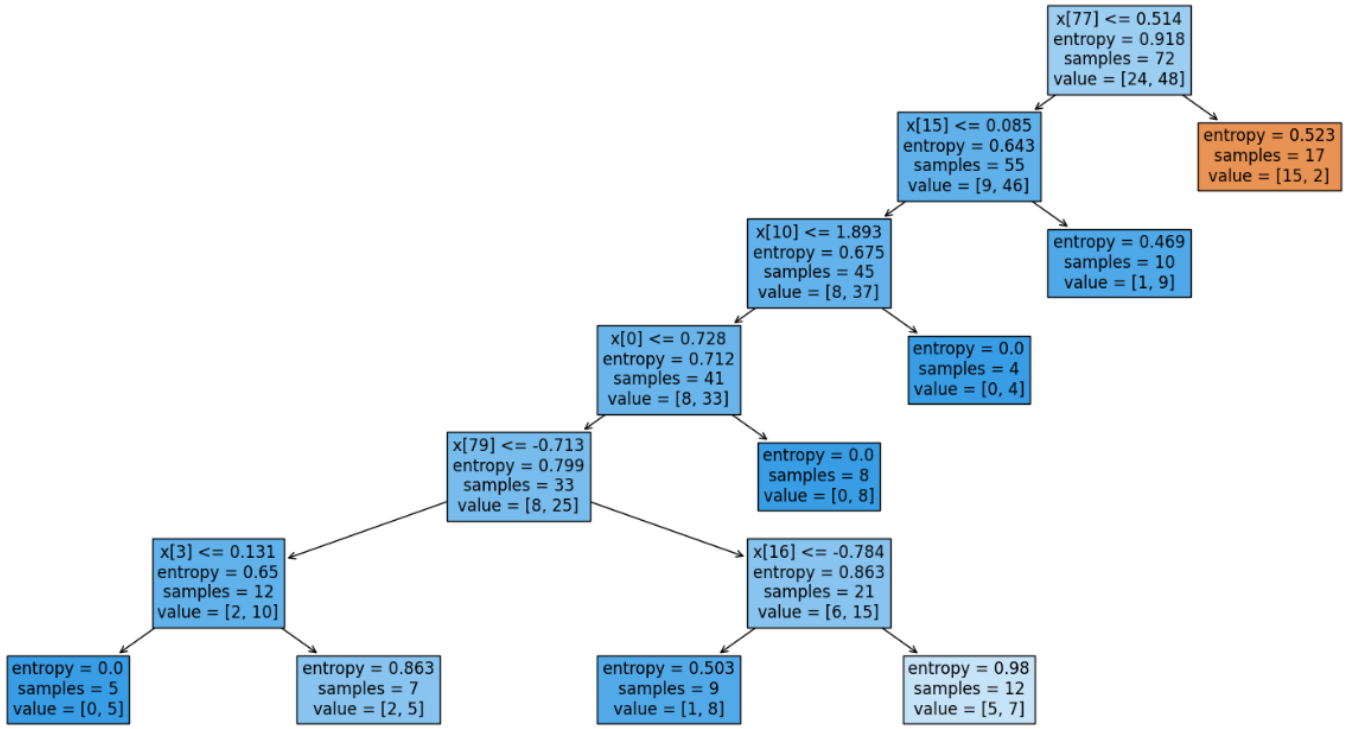


*Predictive Analysis using Decision Tree Classification

```
DecisionTreeClassifier
DecisionTreeClassifier(criterion='entropy', max_depth=8, max_features='sqrt',
                       min_samples_leaf=4, min_samples_split=5,
                       splitter='random')
```

```
Predicted success rate is 83.33333333333334 %
```

- Use GridSearchCV to find the most suitable parameters.
- Fit and predict on the DecisionTreeClassifier object.
- The predicted success rate is **83.33%**, which is similar to the trend of success rate in recent years.



Conclusions

- Launch site KSC LC-39A has overall 76.9% of success rate.
- Success rate is the highest when payload mass is between 2000 kg and 6000 kg.
- Booster version FT has the highest success rate.
- The predicted success rate (Decision Tree Classification) is **83.33%**, which is similar to the trend of success rate in recent years.

Appendix

- Relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets of this project:
- https://github.com/vincent168e/ibm_ds_final_proj/tree/main

Thank you!