IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Shlok Divyam
18th June 2025

Outline

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

Executive Summary

Goal: In this Capstone Project, we will try to predict if the SpaceX Falcon 9 first stage will land successfully, using different machine learning classification algorithms.

Methods:

Data collection via SpaceX API + web scraping

Data Wrangling and Formatting

Exploratory Data Analysis (EDA)

Visualization of Collected Data

Machine learning classification models

Outcome:

The outcome of landing is directly correlated with some of the features of the rocket launch, such as its payload and site of launch.

The Decision Tree model performs better than other models, and hence, can be used to predict the outcomes in this case.

Introduction

- The question that we will try to answer is, will the launched SpaceX Falcon 9 first stage successfully land?
- SpaceX advertises that Falcon 9 rocket launches costs 62 million dollars on its website. Other launching vehicles cost more than 165 million dollars. This savings is due to the fact that SpaceX is able to reuse the first stage of rocket launch. Therefore, if we determine whether the first stage will successfully land, we will be able to determine the cost of launch.
- Most of the times, the unsuccessful landings are planned. Sometimes, SpaceX will perform a controlled landing in the ocean.
- This information can be used to if an alternate company wants to bid against SpaceX for a rocket launch, boosting their chances of winning.

Section 1

Methodology

Methodology

Data collection methodology	SpaceX API Web scraping
Perform data wrangling	Pandas Numpy
Perform exploratory data analysis (EDA)	SQL Visualization with Bar Charts, Scatter Plot, and Line Charts
Perform interactive visual analytics	Folium Plotly Dash
Perform predictive analysis using classification models	Logistic Regression Support Vector Machines Decision Trees K Nearest Neighbours

Data Collection

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite
4	1	2010-06-04	Falcon 9	6123.547647	LEO	CCSFS SLC 40
5	2	2012-05-22	Falcon 9	525.000000	LEO	CCSFS SLC 40
6	3	2013-03-01	Falcon 9	677.000000	ISS	CCSFS SLC 40
7	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E
8	5	2013-12-03	Falcon 9	3170.000000	GTO	CCSFS SLC 40
9	6	2014-01-06	Falcon 9	3325.000000	GTO	CCSFS SLC 40
10	7	2014-04-18	Falcon 9	2296.000000	ISS	CCSFS SLC 40
11	8	2014-07-14	Falcon 9	1316.000000	LEO	CCSFS SLC 40
12	9	2014-08-05	Falcon 9	4535.000000	GTO	CCSFS SLC 40
13	10	2014-09-07	Falcon 9	4428.000000	GTO	CCSFS SLC 40
14	11	2014-09-21	Falcon 9	2216.000000	ISS	CCSFS SLC 40
15	12	2015-01-10	Falcon 9	2395.000000	ISS	CCSFS SLC 40
16	13	2015-02-11	Falcon 9	570.000000	ES-L1	CCSFS SLC 40
17	14	2015-04-14	Falcon 9	1898.000000	ISS	CCSFS SLC 40
18	15	2015-04-27	Falcon 9	4707.000000	GTO	CCSFS SLC 40

- SpaceX API
- The API used is https://api.spacexdata.com/v4/rockets/, and is available on the SpaceX website.
- The API provides data about many types of rocket launches done by SpaceX. Therefore, we filtered it to include only Falcon 9 launches.
- Every missing value in the Payload Mass column is replaced with its mean.
- There were 90 rows and 17 columns. The picture on the left shows a hint of the first few rows of the data.

Data Collection - Scraping

- Web scraping
  - The data is scraped from the Wikipedia page of all the Falcon 9 launches: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
  - The website contains many tables, out of which, only one table was of our use.
  - There were 121 rows and 11 columns. The picture given shows the first few rows of the data.

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	NaN	Dragon Spacecraft Qualification Unit	LEO	SpaceX	Success	F9 v1.07B0003.18	Failure	4 June 2010	18:45
1	2	CCAFS	NaN	Dragon	LEO	NASA	Success	F9 v1.07B0004.18	Failure	8 December 2010	15:43
2	3	CCAFS	NaN	Dragon	LEO	NASA	Success	F9 v1.07B0005.18	No attempt\n	22 May 2012	07:44
3	4	CCAFS	NaN	SpaceX CRS-1	LEO	NASA	Success	F9 v1.07B0006.18	No attempt	8 October 2012	00:35
4	5	CCAFS	NaN	SpaceX CRS-2	LEO	NASA	Success	F9 v1.07B0007.18	No attempt\n	1 March 2013	15:10
5	6	VAFB	NaN	CASSIOPE	Polar orbit	MDA	Success	F9 v1.17B10038	Uncontrolled	29 September 2013	16:00
6	7	CCAFS	NaN	SES-8	GTO	SES	Success	F9 v1.1	No attempt	3 December 2013	22:41
7	8	CCAFS	NaN	Thaicom 6	GTO	Thaicom	Success	F9 v1.1	No attempt	6 January 2014	22:06
8	9	Cape Canaveral	NaN	SpaceX CRS-3	LEO	NASA	Success	F9 v1.1	Controlled	18 April 2014	19:25
9	10	Cape Canaveral	NaN	Orbcomm-OG2	LEO	Orbcomm	Success	F9 v1.1	Controlled	14 July 2014	15:15
10	11	Cape Canaveral	NaN	AsiaSat 8	GTO	AsiaSat	Success	F9 v1.1	No attempt	5 August 2014	08:00

Data Wrangling

- The data is later processed so that there are no missing entries. Categorical features are encoded using one-hot encoding.
- An extra column called ‘Class’ is also added to the data frame. This column contains numerical values, 0 if a given launch was unsuccessful and 1 if it was successful.
- In the end, the dataframe had 90 rows and 83 columns, which was then saved in a csv format.
- A glimpse of the saved dataframe is showed in the picture below.

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude	Class
0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003	-80.577366	28.561857	0
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005	-80.577366	28.561857	0
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007	-80.577366	28.561857	0
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.0	0	B1003	-120.610829	34.632093	0
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1004	-80.577366	28.561857	0

EDA with Data Visualization

- To visualize the data, the python libraries that were used are:
  - Pandas
  - Numpy
  - Matplotlib
  - Seaborn
- Some graphs were plotted which showed the relationships between:
  - Flight Number vs. Launch Site
  - Payload vs. Launch Site
  - Success Rate vs. Orbit Type
  - Flight Number vs. Orbit Type
  - Payload vs. Orbit Type
  - Yearly Trend of Success Rate
- Link to Notebook: https://github.com/ShlokDivyam1109/Data-Science-Capstone-Project/blob/main/edadataviz.ipynb

EDA with SQL

- By querying SQL, we found answer to some questions, which were:
  - All the names of Launch Sites
  - Launch Sites whose name begins with 'CCA'
  - Total Payload of all the launches till date
  - Average Payload weight carried by booster version F9 v1.1
  - First Successful Ground Landing date
  - Names of the boosters which have success in Drone Ship Landing and have Payload Mass between 4000 and 6000
  - Total Number of Successful and Unsuccessful Launches
  - Names of boosters which carried the maximum Payload
  - Records from the year 2015
  - Count of all the outcomes with Launch Date between the date 2010-06-04 and 2017-03-20, in descending order

- Link to Notebook: https://github.com/ShlokDivyam1109/Data-Science-Capstone-Project/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Folium provides functions to create Interactive Maps, with Markers, Labels, Popups, and many more objects. To visualise data about the Launch sites, we used Folium maps, and created:
  - Markers: To locate Launch Sites, and its nearest Costline, Highway, Railway and City, along with Launch Status.
  - Marker Cluster: To make smart group of Launch Stutus Markers.
  - Labels: To define distances from Launch Site to its nearest landmarks as mentioned.
  - Lines: To show the distances between Launch Site to its nearest landmarks.
- Link to Notebook: https://github.com/ShlokDivyam1109/Data-Science-Capstone-Project/blob/main/lab_jupyter_launch_site_location.ipynb

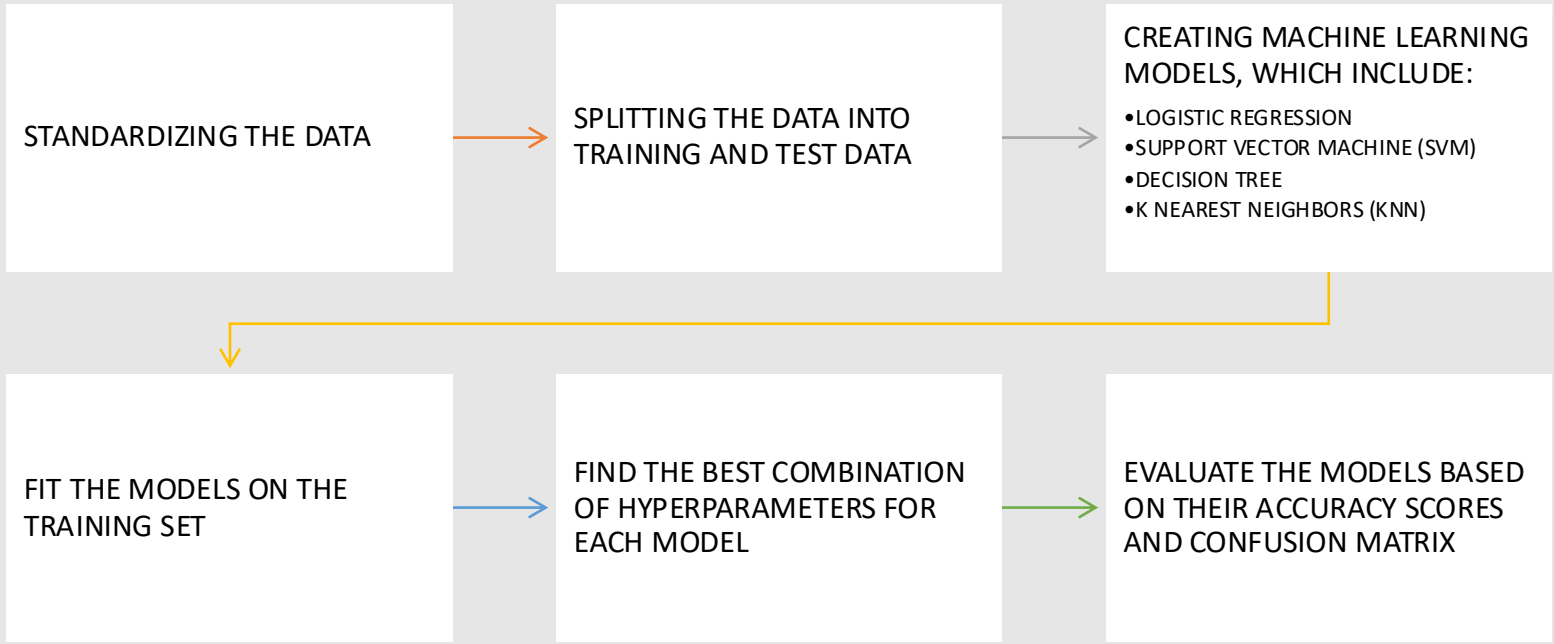
Build a Dashboard with Plotly Dash

- Using Plotly Dash, we made a Dashboard with following plots and input options:
    - Input Options:
        - Dropdown -> To select whether the data should belong to a particular site, or all the sites.
        - Range Slider -> To select the Payload Range
    - Output Plots:
        - Pie -> If all sites selected, will show number of successful launches in every sites, else, will show number of successful and unsuccessful launches of the particular site selected from the dropdown.
        - Scatter Plot -> To show the success of a launch based on Payload in x axis, whose range is selected from the slider, out of the data of selected site.
- Link to Notebook: https://github.com/ShlokDivyam1109/Data-Science-Capstone-Project/blob/main/Dashboard.py

Predictive Analysis (Classification)

Scikit-Learn provided functions to create machine learning models on python

The machine learning prediction phase include the following steps:



Link to the notebook: https://github.com/ShlokDivyam1109/Data-Science-Capstone-Project/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

EDA Results

Landing Success by Launch Site
- CCAFS SLC 40 and KSC LC 39A: more frequent launches, but lower landing success rates.
- VAFB SLC 4E: fewer launches, but higher success rate.

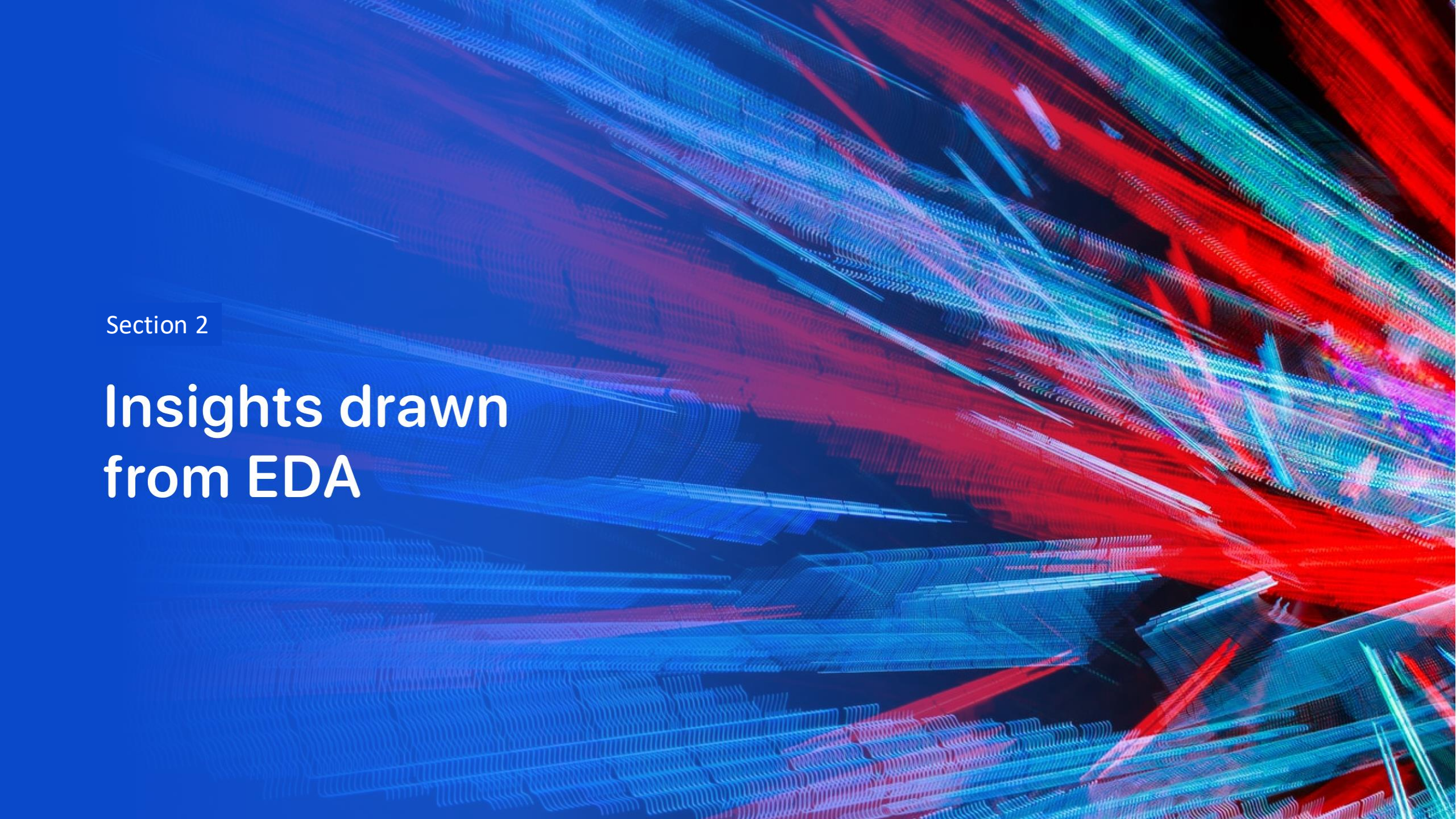
Orbit vs Landing Success
- GTO (Geostationary Transfer Orbit) had more failed landings.
- LEO and SSO showed higher success rates.
- Complex orbits generally correlated with higher success, likely due to higher-profile missions.

Payload Mass vs Success
- Lower payloads (< 4000 kg) had mixed outcomes.
- Payloads between 4000–6000 kg had better success rates.
- Very high payloads (>7000 kg) had few data points, but leaned toward higher success.

Flight Number vs Landing Success
- Strong positive correlation: As the flight number increased, so did the landing success indicating SpaceX improved over time.

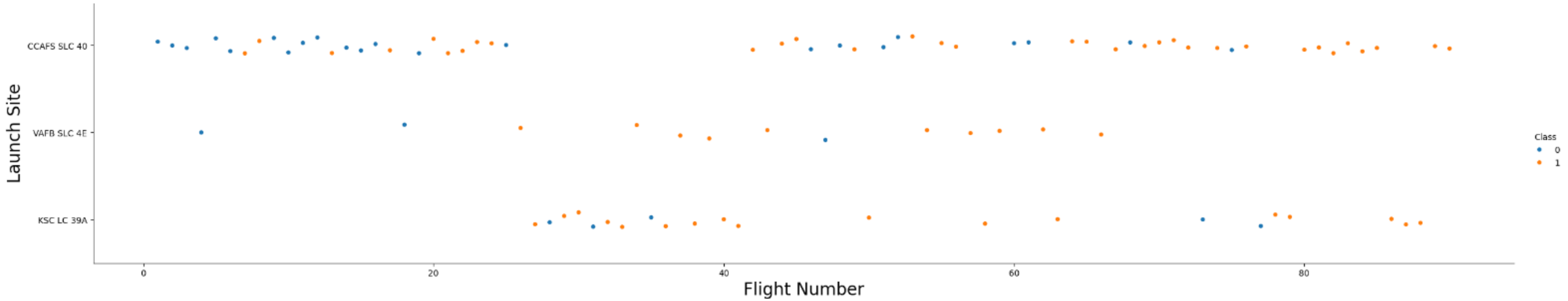
Booster Version Category
- Newer booster versions had much better success rates than older versions.
- Shows the impact of iterative engineering upgrades on performance.

Geospatial Observations
- Landings closer to Florida/California coastlines had higher success, consistent with barge/ground landings.
- Barge landings were more risky than land-based landings but improved over time.

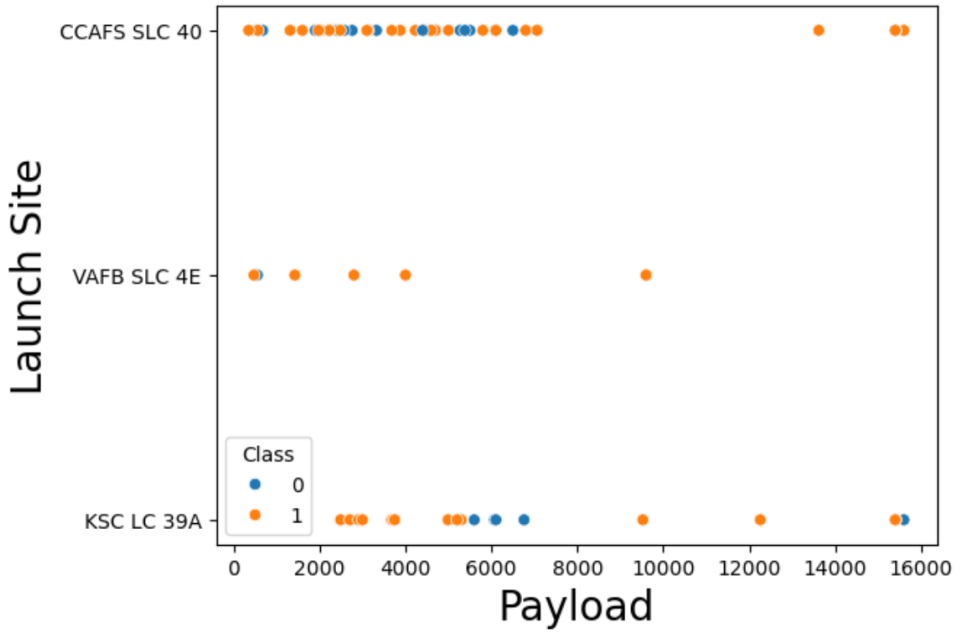

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

- Most of the launches took place in CCAFS SLC-40 Launch Site.
- However, there was a change in the preference from flight number 27 to 41, and KSC LC-39A emerged as the most prefered site.
- A few launches took place from VAFB SLC 4E



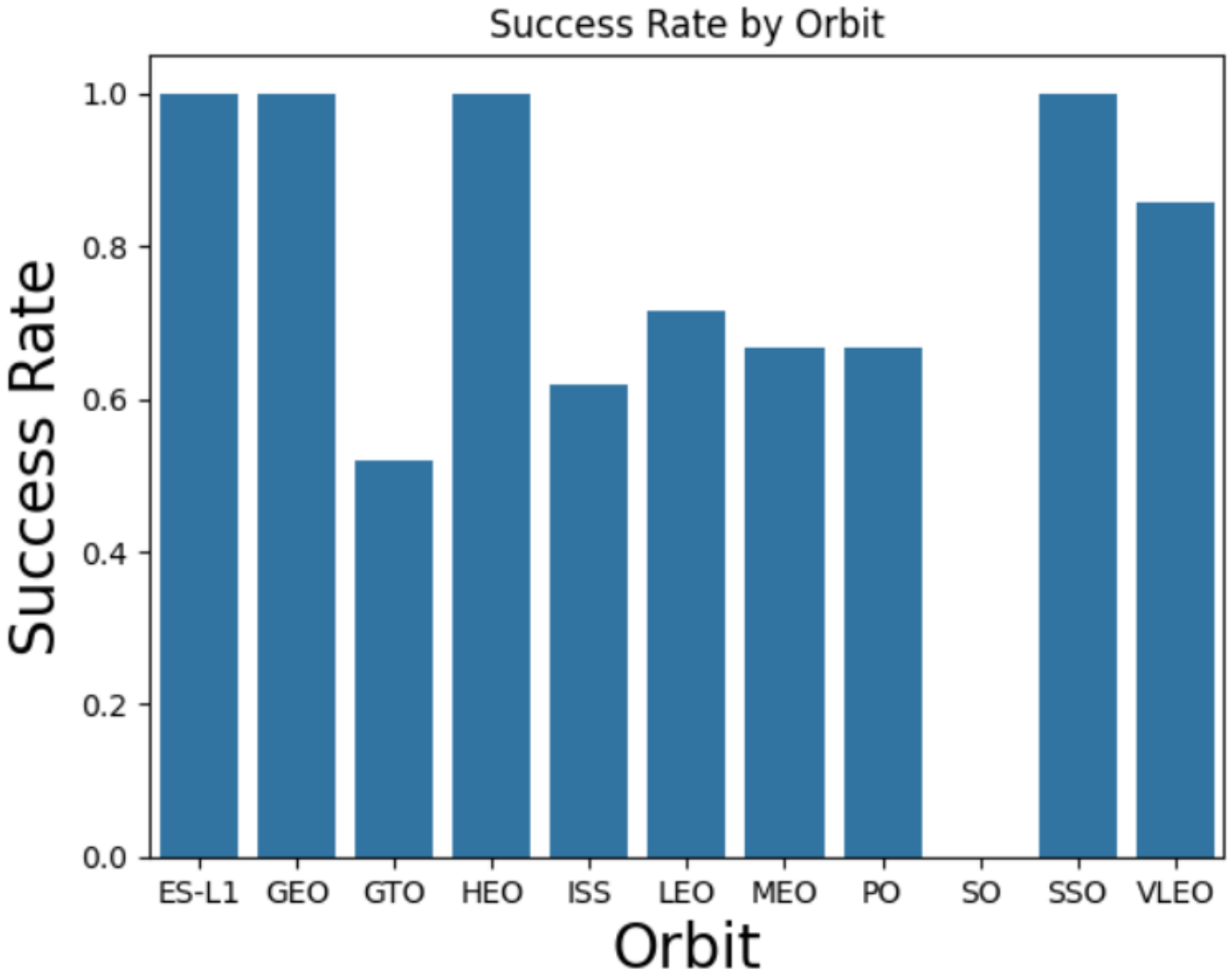
Payload vs. Launch Site



- Lower payloads (< 4000 kg) had mixed outcomes.
- Payloads between 4000–6000 kg had better success rates.
- Very high payloads (>7000 kg) had few data points, but leaned toward higher success—possibly due to newer, upgraded boosters.
- VAFB SLC 4E had no launches for heavy payload mass (>10000 kg)

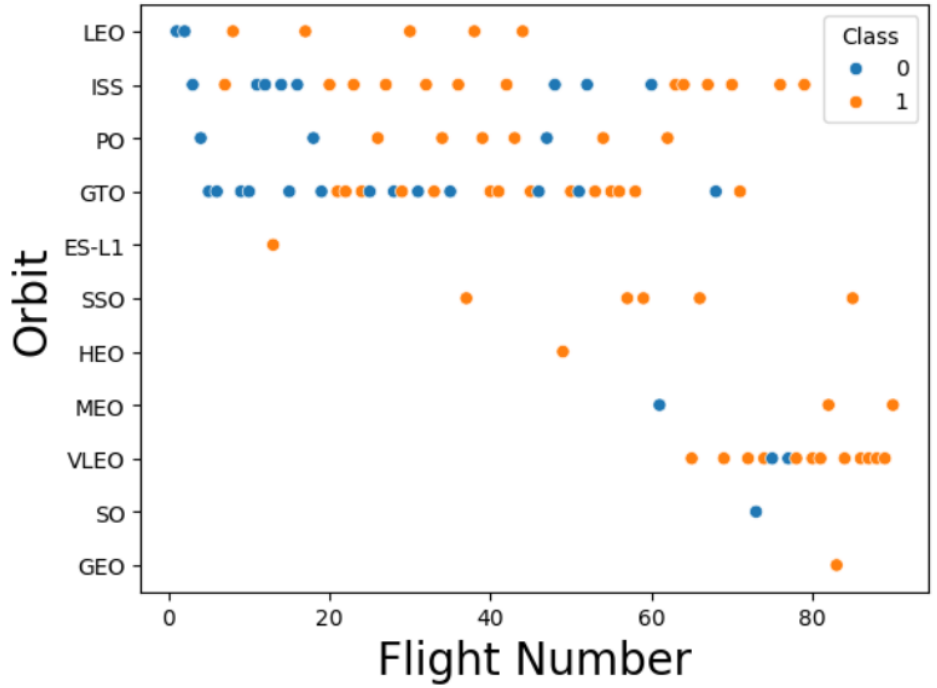
Success Rate vs. Orbit Type

- GTO (Geostationary Transfer Orbit) had more failed landings.
- LEO (Low Earth Orbit) and SSO (Sun Synchronous Orbit) showed higher success rates.
- Complex orbits (like ES-L1, HEO) generally correlated with higher success, likely due to higher-profile missions.

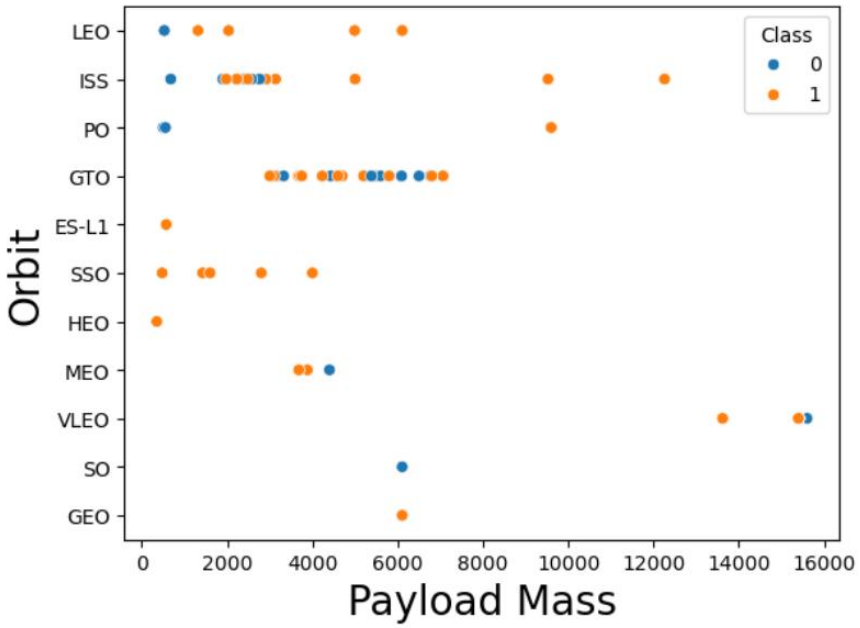


Flight Number vs. Orbit Type

- Orbit diversity increases with Flight Number – SpaceX took on more complex and diverse orbital missions over time.
- Landing success probability improves for newer flights, regardless of orbit.



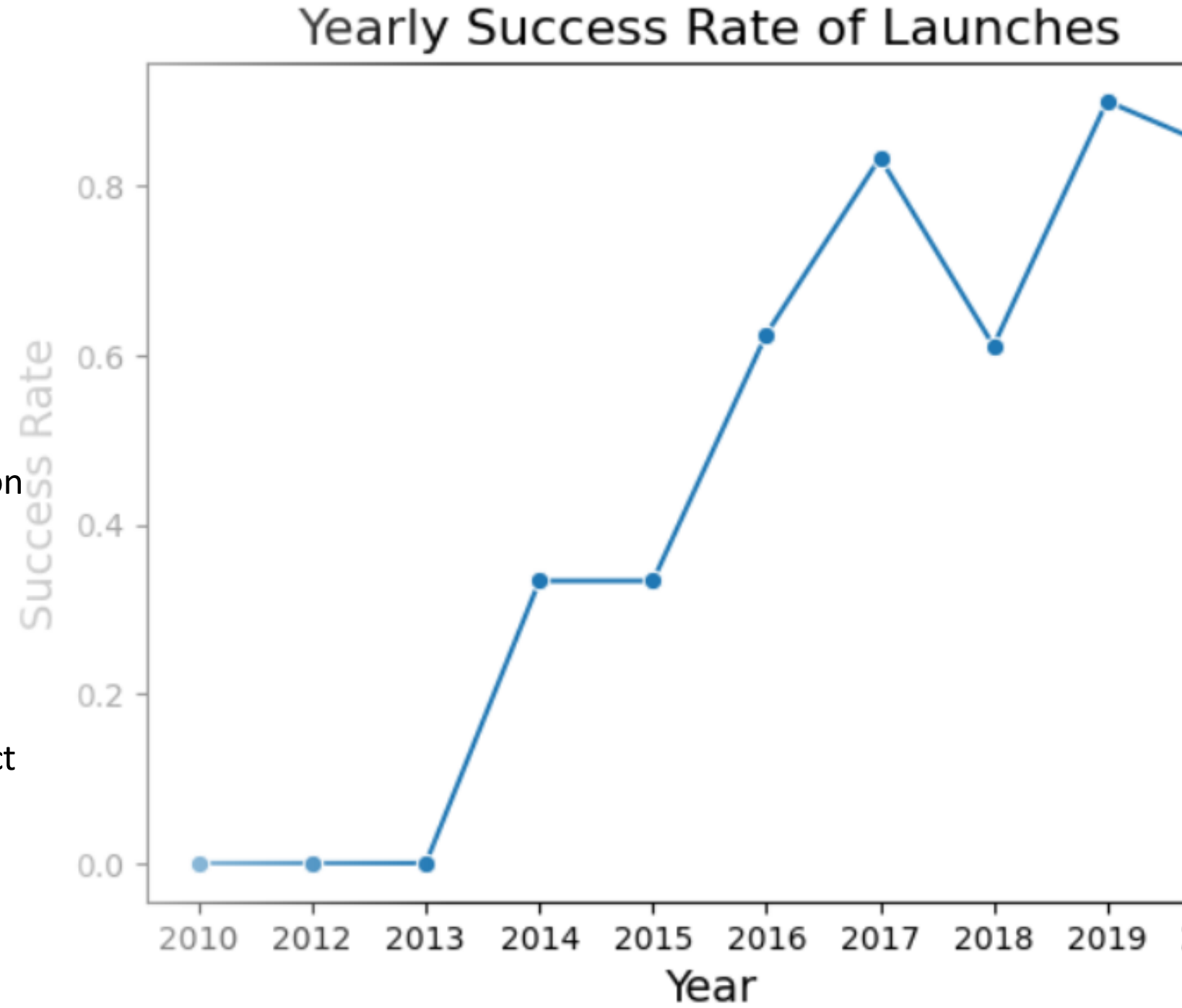
Payload vs. Orbit Type



- Heavier payloads are typically sent to higher orbits (GTO, HEO, MEO).
- LEO and ISS missions have lower payloads and higher landing success rates.
- Failures were more common in early GTO missions

Launch Success Yearly Trend

- The overall trend is strongly upward, demonstrating continuous improvement in SpaceX's technology and mission execution.
- A key turning point occurred in 2016–2017, where the success rate exceeded 60%, and continued to grow from there.
- The brief dip in 2018 may reflect a spike in launch volume or mission complexity but was quickly recovered by 2019.



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

All Launch Site Names

- There were only four sites from which the launch missions took place.
- Out which, two launch sites are of CCAFS
- Quesry: SELECT DISTINCT "Launch_Site" FROM SPACEXTBL;

Launch Site Names Begin with 'CCA'

- The record show that there are many different rows of records, with Launch Site Names beginning with 'CCA'.
- Query: SELECT * FROM SPACEXTBL where Launch_Site like 'CCA%' LIMIT 5;

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

Total Payload Mass

- The large number of Total Payload Mass indicates that SpaceX has launched many rockets, some of which have large Payload mass.
- Query: Select sum(PAYLOAD_MASS__KG_) from SPACEXTBL
-

sum(PAYLOAD_MASS__KG_)
619967

Average Payload Mass by F9 v1.1

- The low average Payload mass of F9 v1.1 indicates that it was used during initial launches, and that, it might have been used in the latest heavy payload launches.
- Query: `Select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version like 'F9 v1.1%'`

avg(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

- The First Successful Ground Landing Date found from the query supports our observation that, after the year 2015, successful launches has significantly kept on increasing.
- Query: Select min(Date) from SPACEXTBL where Landing_Outcome = 'Success (ground pad)'

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The large number of booster versions with successful Drone Ship Landing and Payload between 4000 and 6000 supports our observation that launches with medium Payload have higher success frequency.
- Query: SELECT Booster_Version FROM SPACEXTBL WHERE Landing_Outcome = 'Success (drone ship)' AND Payload_Mass__kg_ > 4000 AND Payload_Mass__kg_ < 6000;
-

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

Total Number of Successful and Failure Mission Outcomes

- Large number of Successful Mission Outcomes indicates that, regardless of success of first stage landing, the mission outcome has been mostly successful.
- Query: Select Mission_Outcome,count(*) from SPACEXTBL group by Mission_Outcome;

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

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

Boosters Carried Maximum Payload

- The booster names which caried maximum Payload mass are all from the latest booster versions, showcasing SpaceX's technological evolution.
- Query: SELECT Booster_Version FROM SPACEXTBL WHERE Payload_Mass__kg_ = (Select max(Payload_Mass__kg_) from SPACEXTBL);

2015 Launch Records

- The Launch Records of the year 2015 between the given dates shows two of the unsuccessful launches.
- Query: SELECT substr(Date, 6, 2) AS month, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTBL WHERE substr(Date, 0, 5) = '2015' AND Landing_Outcome LIKE '%Failure (drone ship)%';

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

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

- Between the given dates, in most of the missions, there was no attempt made to land the first stage. But second most recurring outcome was successful landing on drone ship, which indicates that SpaceX has made advancements in their technology.
- Query: SELECT Landing_Outcome, COUNT(*) FROM SPACEXTBL WHERE date > '2010-06-04' AND date < '2017-03-20' GROUP BY Landing_Outcome ORDER BY COUNT(*) DESC

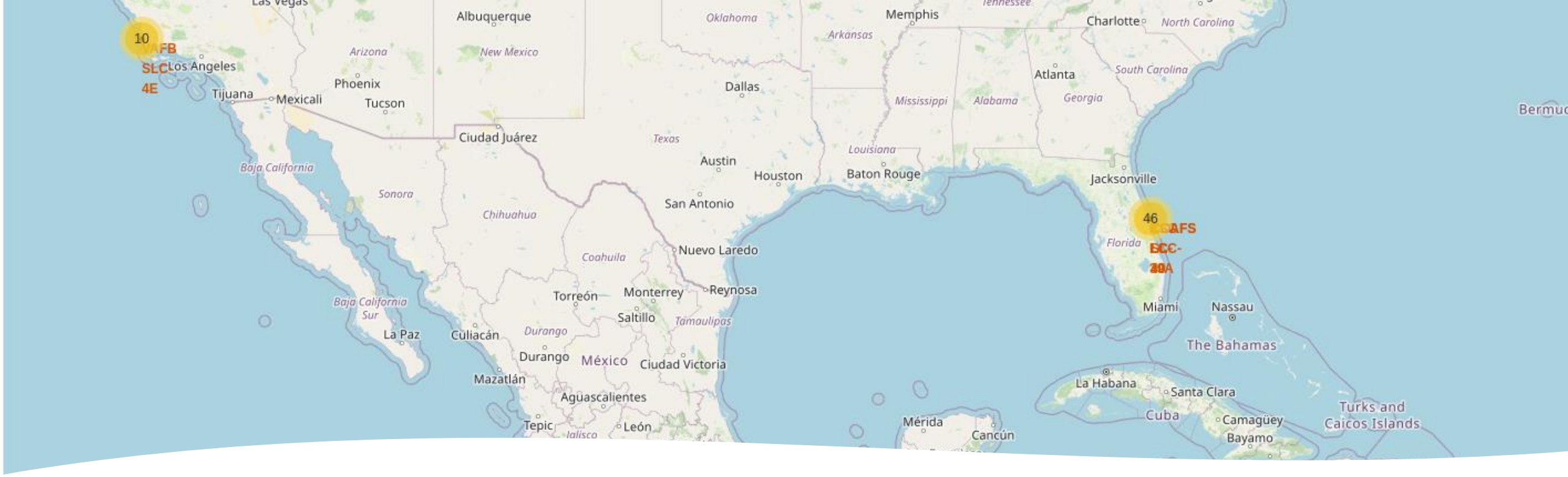
Landing_Outcome	COUNT(*)
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

Section 3

Launch Sites Proximities Analysis

Mark All Launch Sites

- The snippet Interactive Folium Map shown here has Markers at the location of different Launch Sites of Falcon 9 rockets.

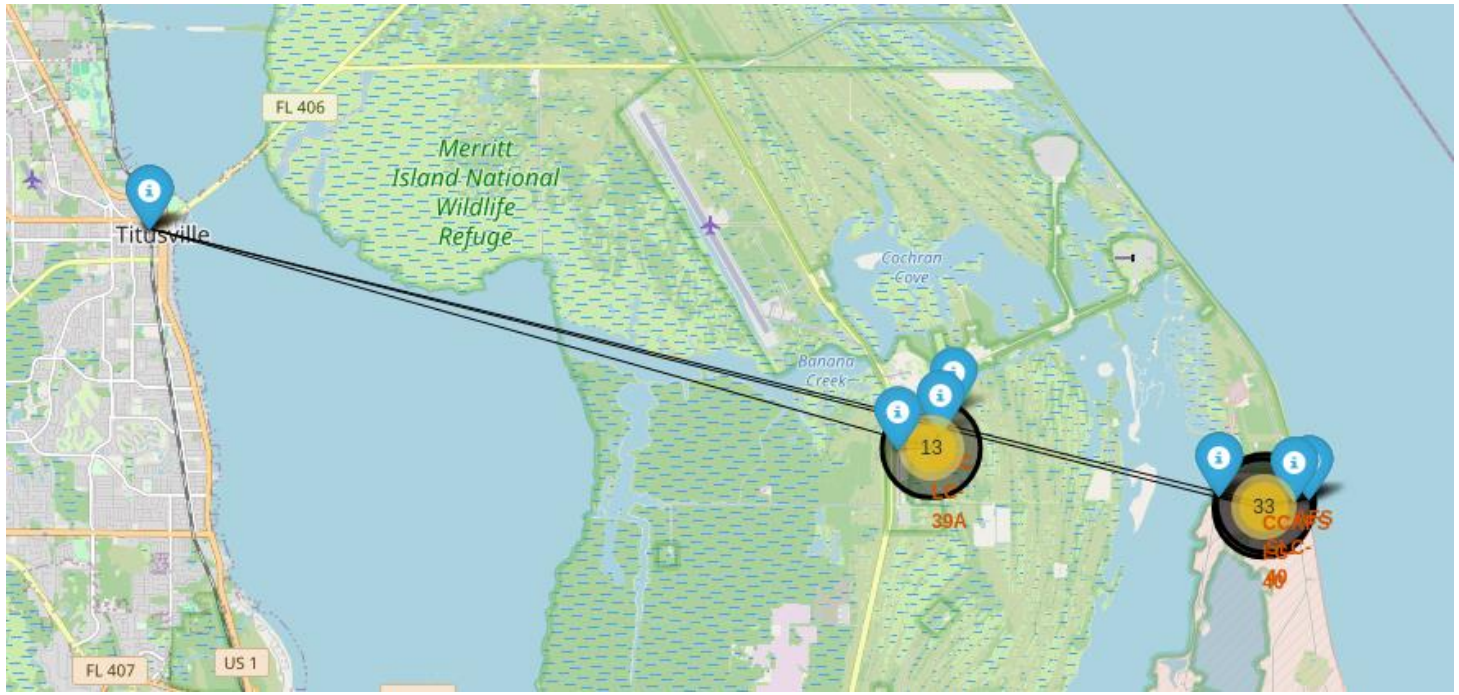


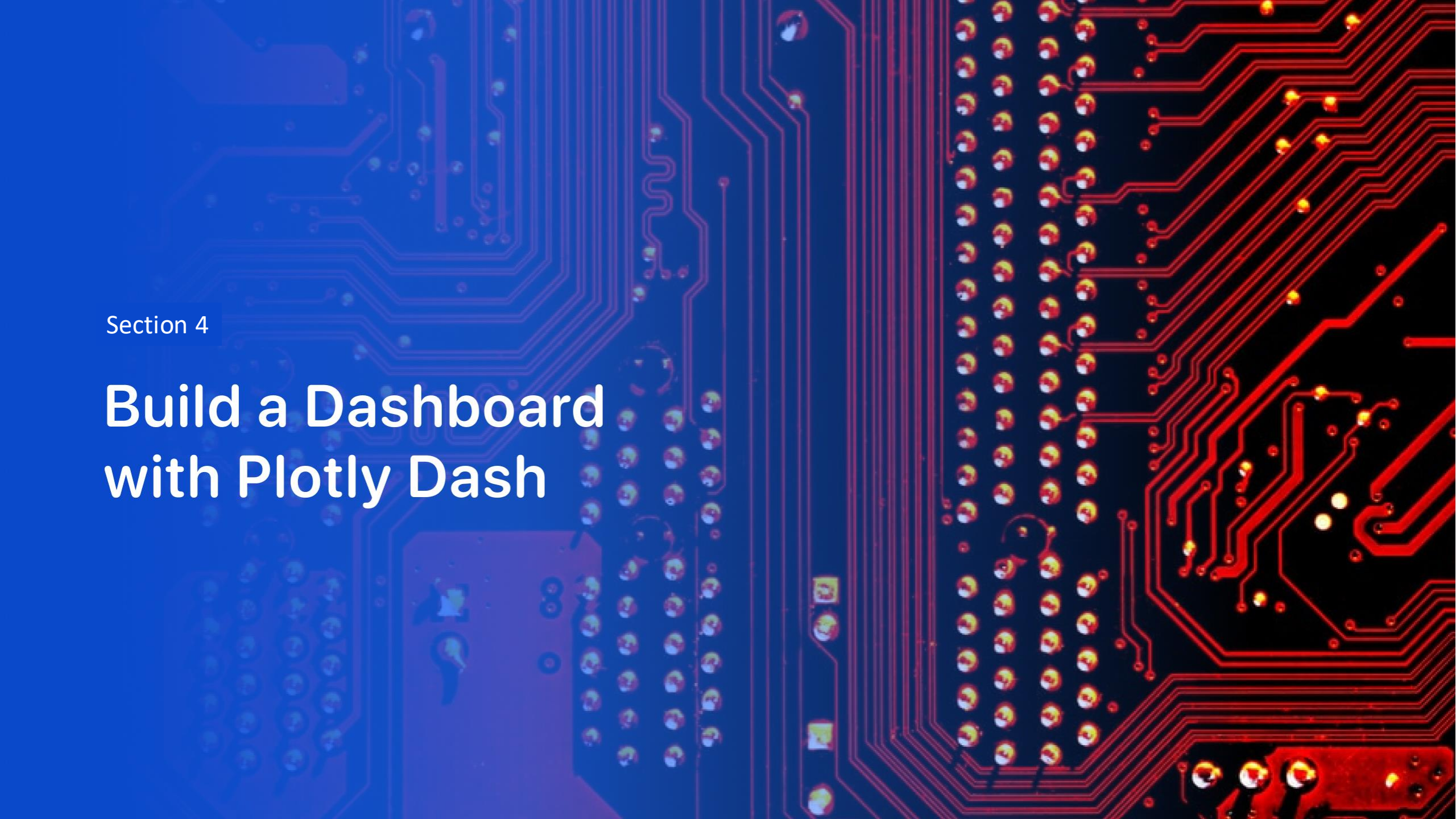
Mark the success/failed launches for each site on the map

- The MarkerCluster Object smartly groups the Markers based on their proximity. This helps greatly to visualize the number of Launches at a given location. This makes the markers, which indicates the success/failure of all the launches, to pop up when clicked.

Launch Site and its proximities

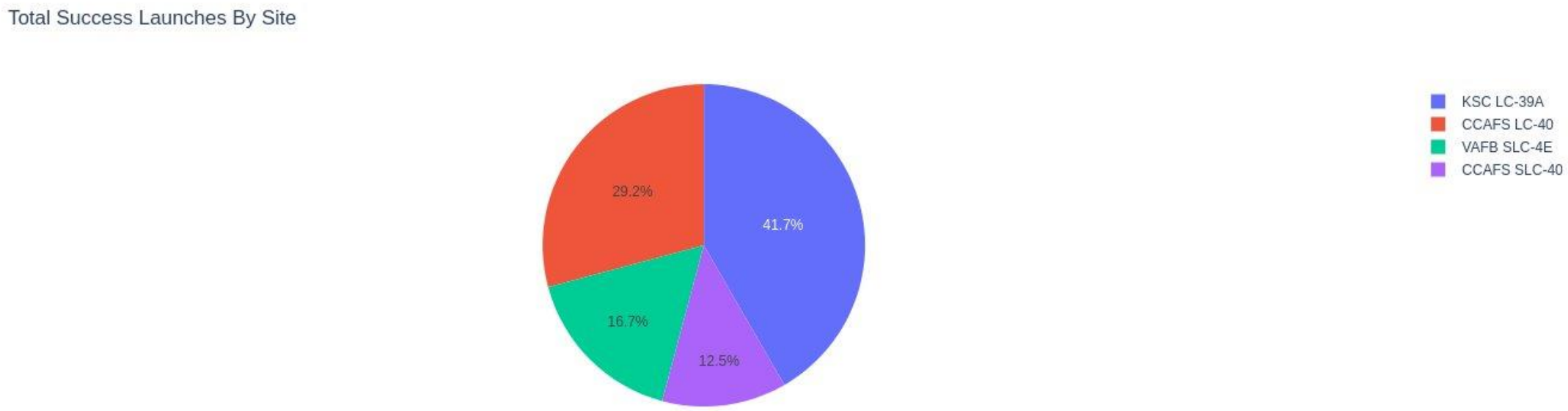
- The given snippet of the Folium map shows that Launch Site location is chosen such that:
  - It is nearest to any of the transportation facility: Highway, Railway or Coastline
  - It is far from cities to mitigate the impact of unsuccessful launches.




Section 4
Build a Dashboard with Plotly Dash

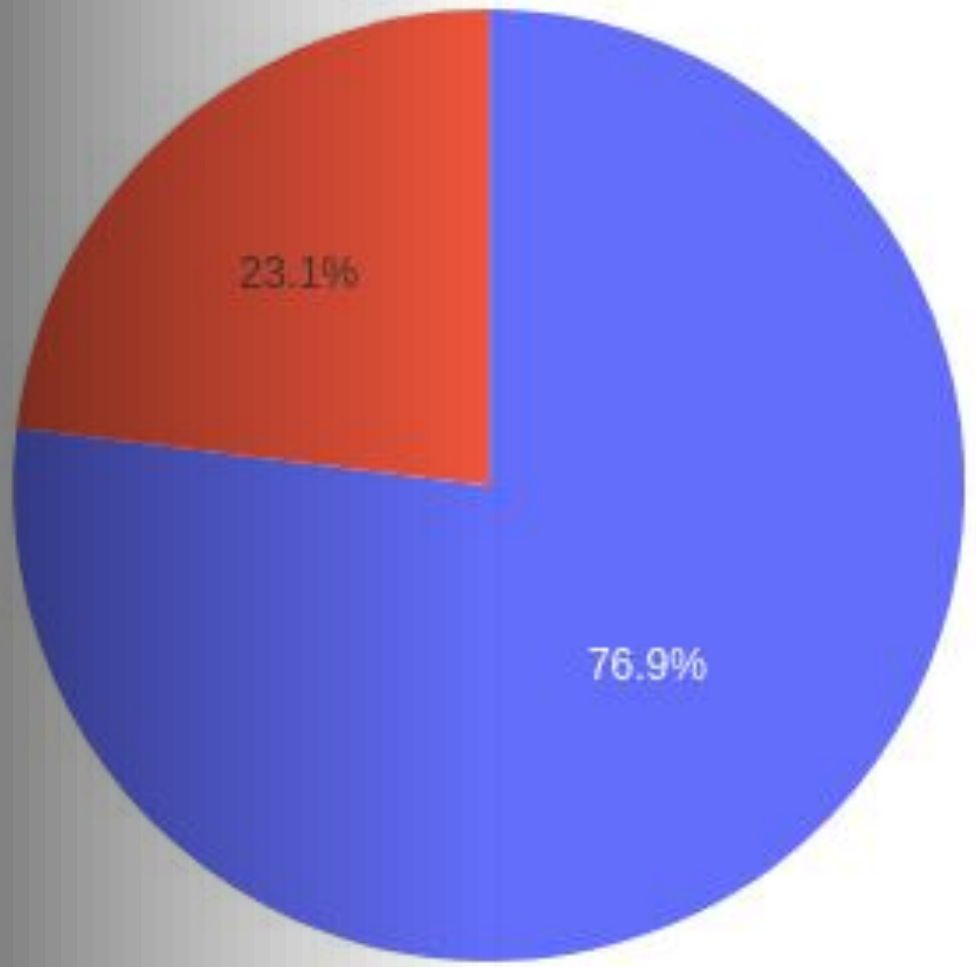
Pie Chart to show Success count for all sites

The given picture shows the pie chart indicating the success count of all sites. It shows that, amongst all other launch sites, the KSC LC-39A has most successful launches, accounting for about 41.2% of all successful launches.



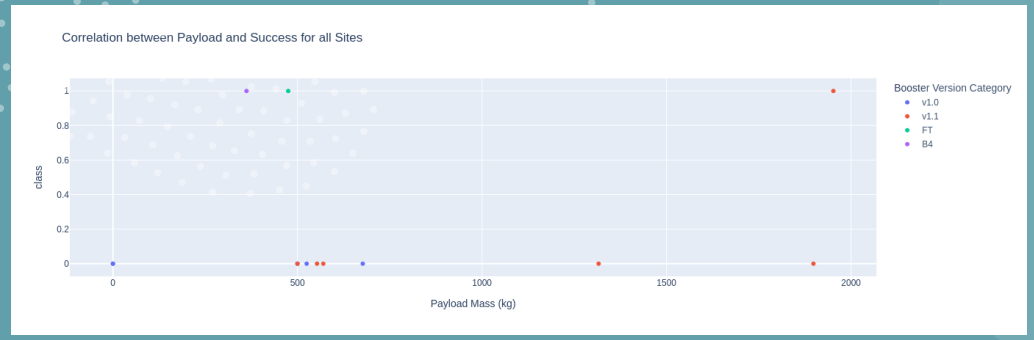
Pie Chart with highest success rate

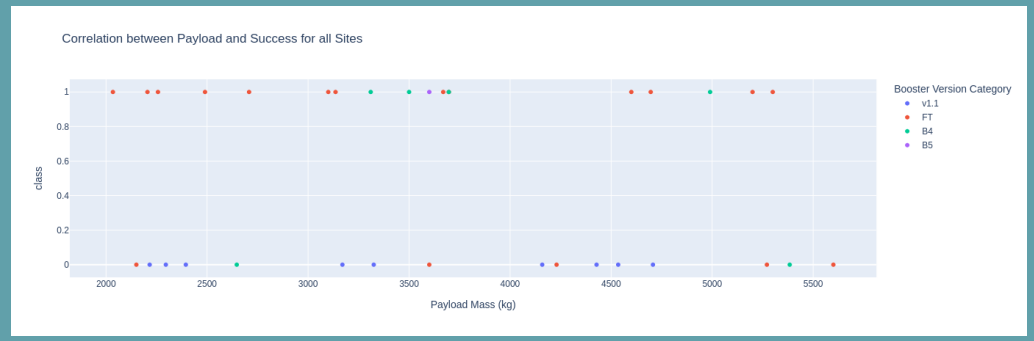
- The pie chart shown in the picture is of date pertaining to Launch Site KSC LC-39A. It shows that about 76.9% of all the launches from the site were successful. This is the highest percentage of success with respect to other launch sites.

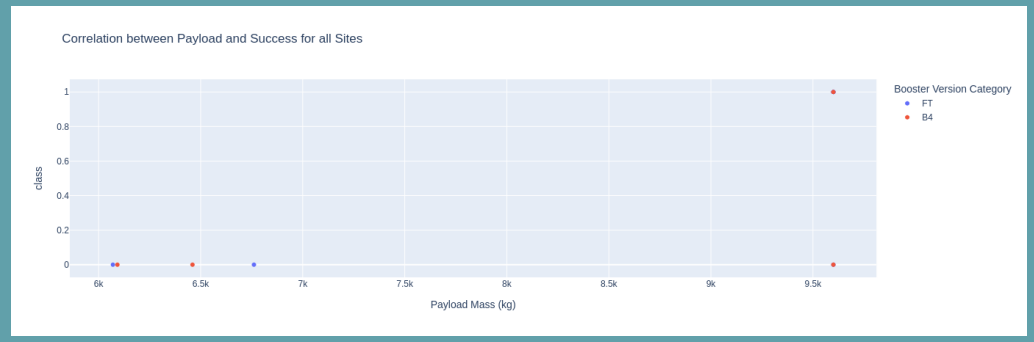


Scatter Plot of Payload vs. Launch Outcome

- The first Scatter Plot shows that for small Payload, i.e., less than 2000kg, the success rate was very low. This may be due to the fact that lesser Payload was used during the initial launches, when landing of first stage were not even attempted.
- The second Scatter Plot shows that for medium Payload, i.e., between 2000kg to 6000kg, the success rate is almost at about 50%, showing increase in success as Payload increased over time.
- The third Scatter Plot shows that for larger Payload, i.e., more than 6000kg, success rate is very low, indicating insufficient technological capabilities to handle larger payloads.







Section 5

Predictive Analysis (Classification)

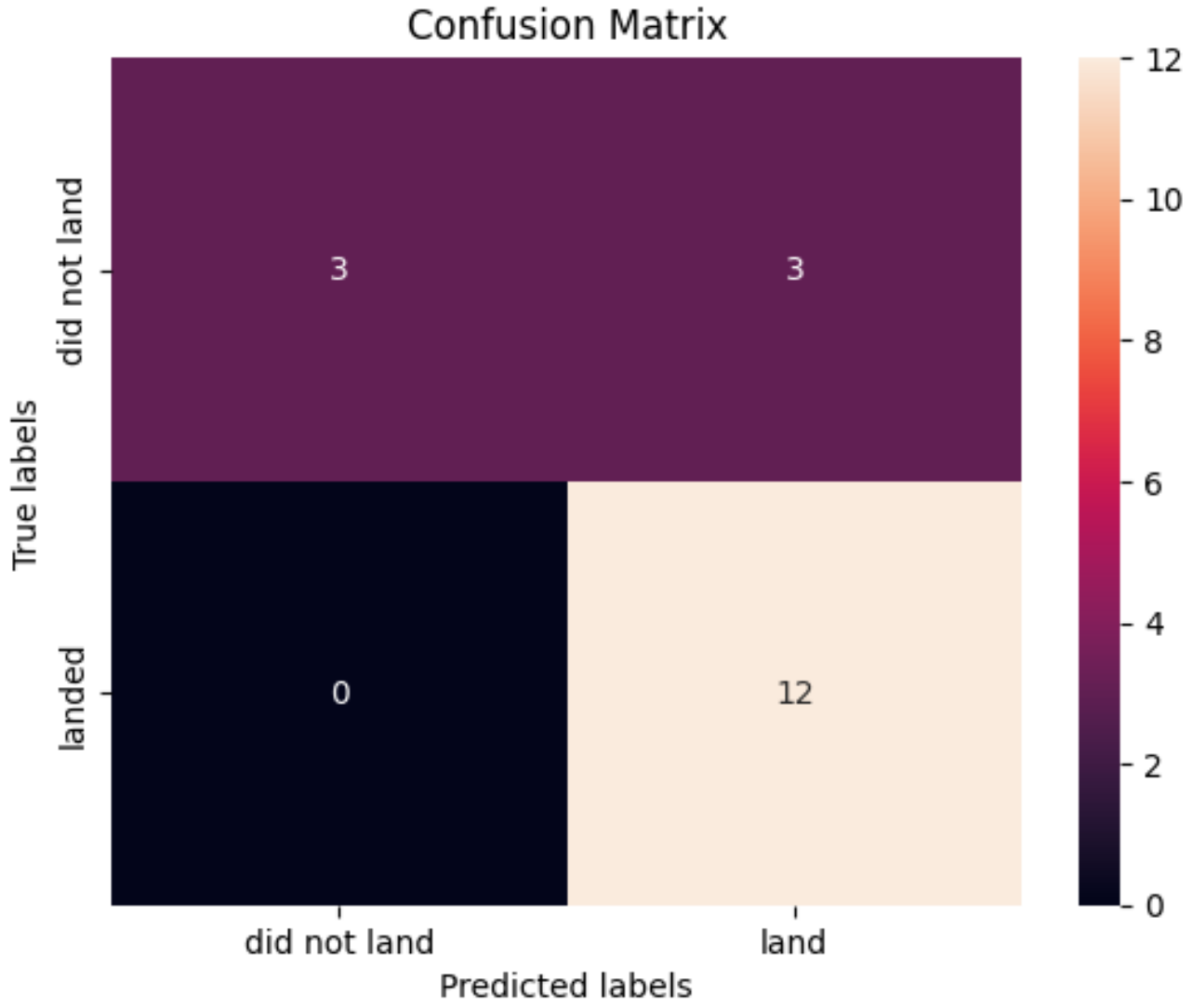
Classification Accuracy

Model	Accuracy
Decision Tree Model	0.91111
Support Vector Machine	0.87778
Logistic Regression	0.86667
K Nearest Neighbours	0.85556

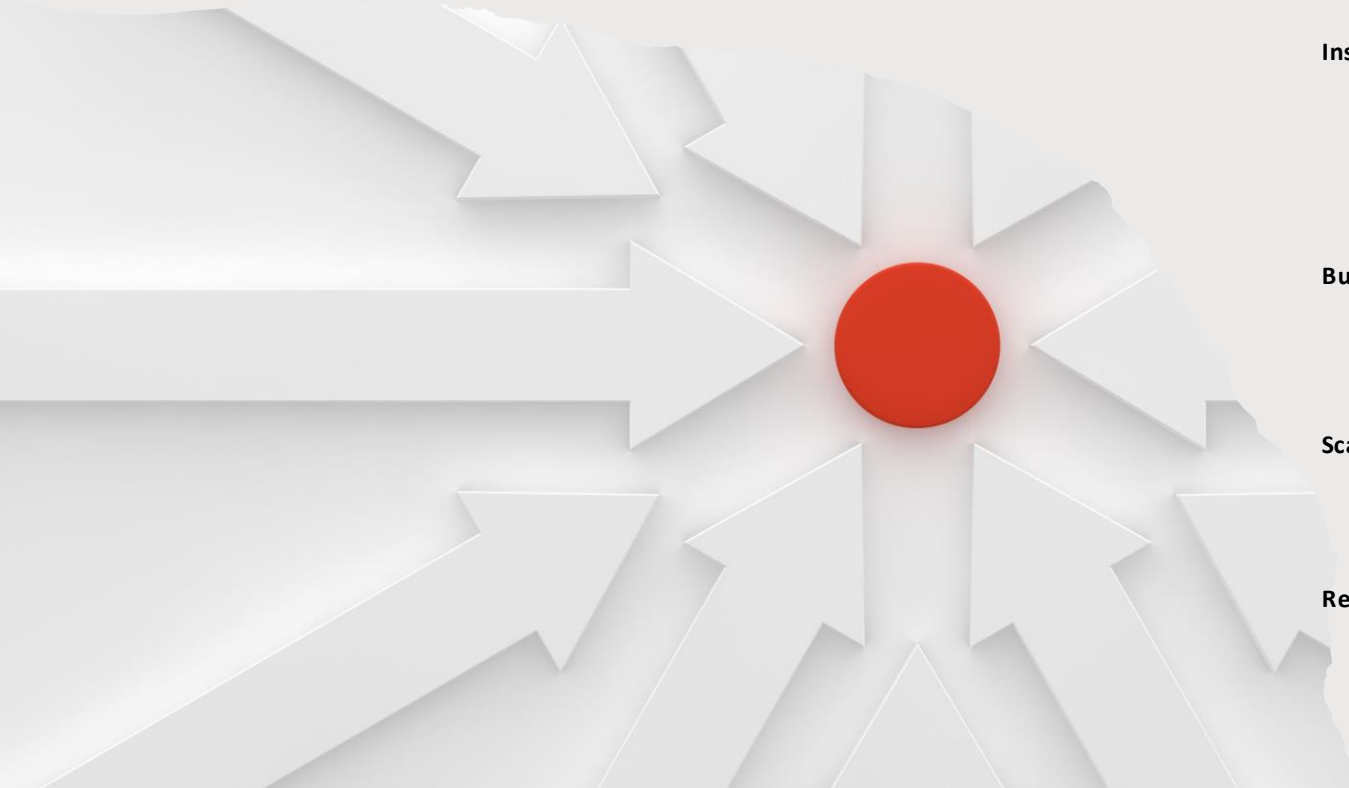
- Based on the scores of Test set, we cannot confirm which method performs the best.
- Same Test set scores may be due to the small size of test sample data. Hence, we tested all methods on the entire dataset.
- The scores of the whole Dataset confirm that the best model is Decision Tree Model. It not only has the highest accuracy, but also has the highest F1 score.

Confusion Matrix

Examining the confusion matrix of Decision Tree Model, we can see that the model can distinguish between different classes, but struggles with False positive cases



Conclusions



Comprehensive Data Pipeline

- Accurate prediction hinges on robust data workflows: **collected via SpaceX REST API and Wikipedia web scraping**, followed by thorough **data cleaning**, including handling missing values and one-hot encoding.

Insightful Exploratory Analysis

- EDA revealed major factors influencing landing outcomes:
  - **Flight number**, **orbit type**, **payload mass**, **booster version**, and **launch site** significantly impact success.

Interactive Visualizations

- Tools like **Folium maps** and **Plotly Dash dashboards** proved invaluable for visualizing spatial and categorical relationships, aiding stakeholder decision-making.

Predictive Modeling & Tuning

- Multiple classifiers were evaluated: **Logistic Regression**, **KNN**, **Decision Tree**, **SVM**.
- After hyperparameter optimization, the **Decision Tree** (or KNN in some implementations) demonstrated the **highest accuracy** in predicting successful landings — often exceeding **90%**.

Insights Backed by Stakeholders

- Models identified critical business insights:
  - Launch experience (higher flight numbers) boosts success.
  - High payloads and complex orbits correlate with more challenging landings.
  - Some launch sites and booster versions are associated with better outcomes.

Business Impact & Cost Savings

- Predicting landing success helps **estimate launch costs** (each unsuccessful landing may increase cost by ~$62M).
- This insight empowers competitive bidding and strategic planning.

Scalable Analysis Foundation

- The project established a **reproducible data science workflow** covering data ingestion, wrangling, visualization, modeling, and reporting — applicable to broader aerospace and industrial problems.

Recommendations & Future Work

- **Extend dataset** with latest launches and booster configurations.
- Experiment with **advanced ML models**.
- **Deploy solutions in production**, integrating dashboards and model pipelines.
- Continuously monitor model performance and retrain periodically.

Appendix

Github Link to all the Notebooks and Pyton codes used in this Capstone project:

https://github.com/ShlokDivyam1109/Data-Science-Capstone-Project.git

Thank you!