IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Rizyana Putri
May 28, 2024

Outline

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

Executive Summary

- Summary of methodologies:
  - Data collection
  - Data wrangling
  - EDA with SQL
  - EDA with visualization
  - Building interactive map with Folium
  - Building dashboard with Plotly Dash
  - Predictive analysis (Classification)
- Summary of results:
  - EDA result
  - Interactive Visual Analytics result
  - Predictive analysis

Introduction

- SpaceX has achieved significant milestones in space exploration, including sending spacecraft to the International Space Station, deploying the Starlink satellite internet constellation, and launching manned missions to space.
- One of the key factors behind SpaceX's success is the relatively low cost of its rocket launches. While other providers charge upwards of $165 million per launch, SpaceX offers Falcon 9 rocket launches for $62 million, due to the reusability of the rocket's first stage.
- The ability to reuse the first stage of the Falcon 9 rocket is crucial for reducing launch costs. However, not every first stage lands successfully—some crash, while others are intentionally sacrificed due to specific mission requirements like payload, orbit, and customer.
- As a data scientist in Space Y, the objective is to predict if Falcon 9 first stage will land successfully; to determine the cost of each launch.

Section 1

Methodology

Methodology

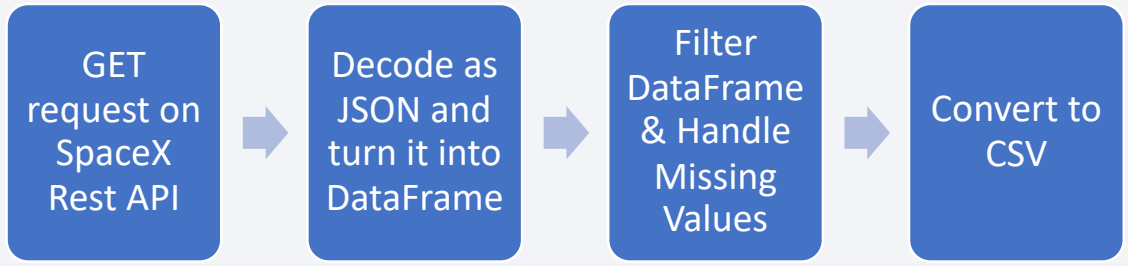
Executive Summary

- Data collection methodology:
  - Data was collected by using SpaceX REST API with the requests library and scraped from Wikipedia using BeautifulSoup.
- Perform data wrangling
  - The `value_counts` method was used to analyze the data. Training labels were created from `Outcomes`.
- Perform exploratory data analysis (EDA) using visualization and SQL
  - EDA performed by querying SQL database. Pandas and Matplotlib were used to visualize the relationships between attributes.
- Perform predictive analysis using classification models
  - Data preprocessing included standardizing, fitting, and transforming features, and splitting the data into training and testing sets. GridSearchCV was used for hyperparameter tuning. Models were evaluated using a confusion matrix and accuracy score.

Data Collection – SpaceX API

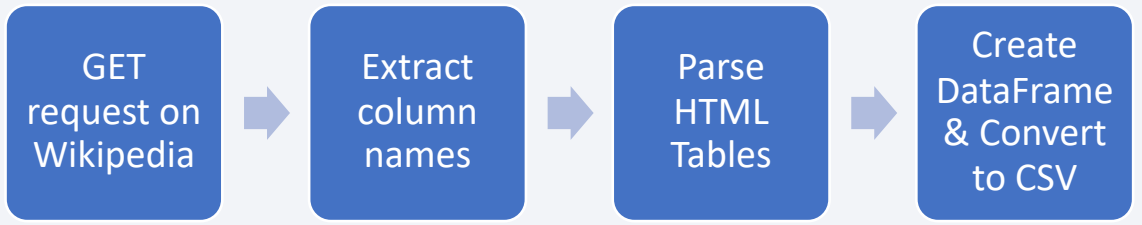
- SpaceX launch data was gathered using the SpaceX REST API (api.spacexdata.com/v4).
- The endpoint v4/launches/past was utilized to retrieve past launch data. Rocket, payloads, launchpad, and cores was obtained from other endpoints.
- The process involved performing a GET request, converting JSON data to a Pandas DataFrame, filter DataFrame to only include Falcon 9 records, and replace missing values with its mean.

External reference: GitHub URL



Data Collection - Scraping

- Web scraping was conducted using data from a Wikipedia page with the BeautifulSoup package.
- HTML tables containing Falcon 9 launch records were scraped, and the data was parsed and converted into a Pandas DataFrame.



External reference: GitHub URL

Data Wrangling

- Initial EDA was performed using `value_counts` to count the number of launches at each site, the number and occurrence of each orbit, and the number and occurrence of mission outcomes for each orbit.
- Training labels were created by converting `Outcomes` into binary classes, 0 indicating failed launch and 1 indicating successful launch.

External reference: GitHub URL

EDA with Data Visualization

- Scatter plot was used to visualize relationship between:
  - Flight Number and Launch Site
  - Payload and Launch Site
  - FlightNumber and Orbit type
  - Payload and Orbit type
- Bar chart was used to visualize success rate of each orbit type
- Line chart was used to visualize launch success yearly trend

External reference: GitHub URL

EDA with SQL

EDA involved querying the SpaceX dataset to gather information such as:

- Unique launch site names,
- Specific records with launch sites starting with 'CCA',
- Total payload mass of boosters launched by NASA (CRS),
- Average payload mass of booster version F9 v1.1,
- The date of the first successful landing outcome on a ground pad,
- Boosters successfully landing on a drone ship with payload mass between 4000 and 6000,
- Total successful and failed mission outcomes,
- Booster versions with maximum payload mass,
- Records displaying month names, failure landing outcomes on drone ships, booster versions, and launch sites for 2015, and
- Ranking landing outcomes between 2010-06-04 and 2017-03-20 in descending order.

External reference: GitHub URL

Build an Interactive Map with Folium

- Launch site locations and their close proximities were marked on an interactive map using several map objects:
  - Marker to mark launch sites based on success/failed launches, and to create icon of text label ‘NASA JSC’ and names of launch sites.
  - Circles to highlight NASA Johnson Space Center’s and launch site’s coordinate.
  - Lines to show the distance between launch sites and nearest railways, highways, cities, and coastlines.
  - MousePosition to get the coordinate.

- External reference: GitHub URL

Build a Dashboard with Plotly Dash

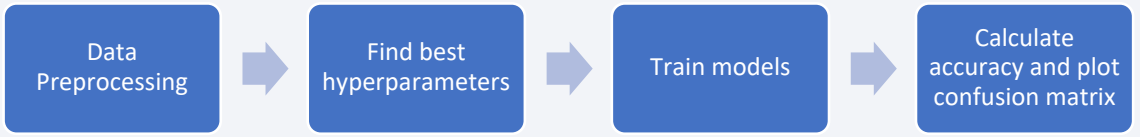
- This dashboard incorporated several plots and graphs:
  - Pie charts to visualize total successful launches for all sites and each launch sites.
  - Scatter plot to visualize the correlation between success and payloads for all sites and each launch sites.
- Along with input components such as:
  - Dropdown list to select the launch sites.
  - Range slider to interact with scatter point chart.

External reference: GitHub URL

Predictive Analysis (Classification)

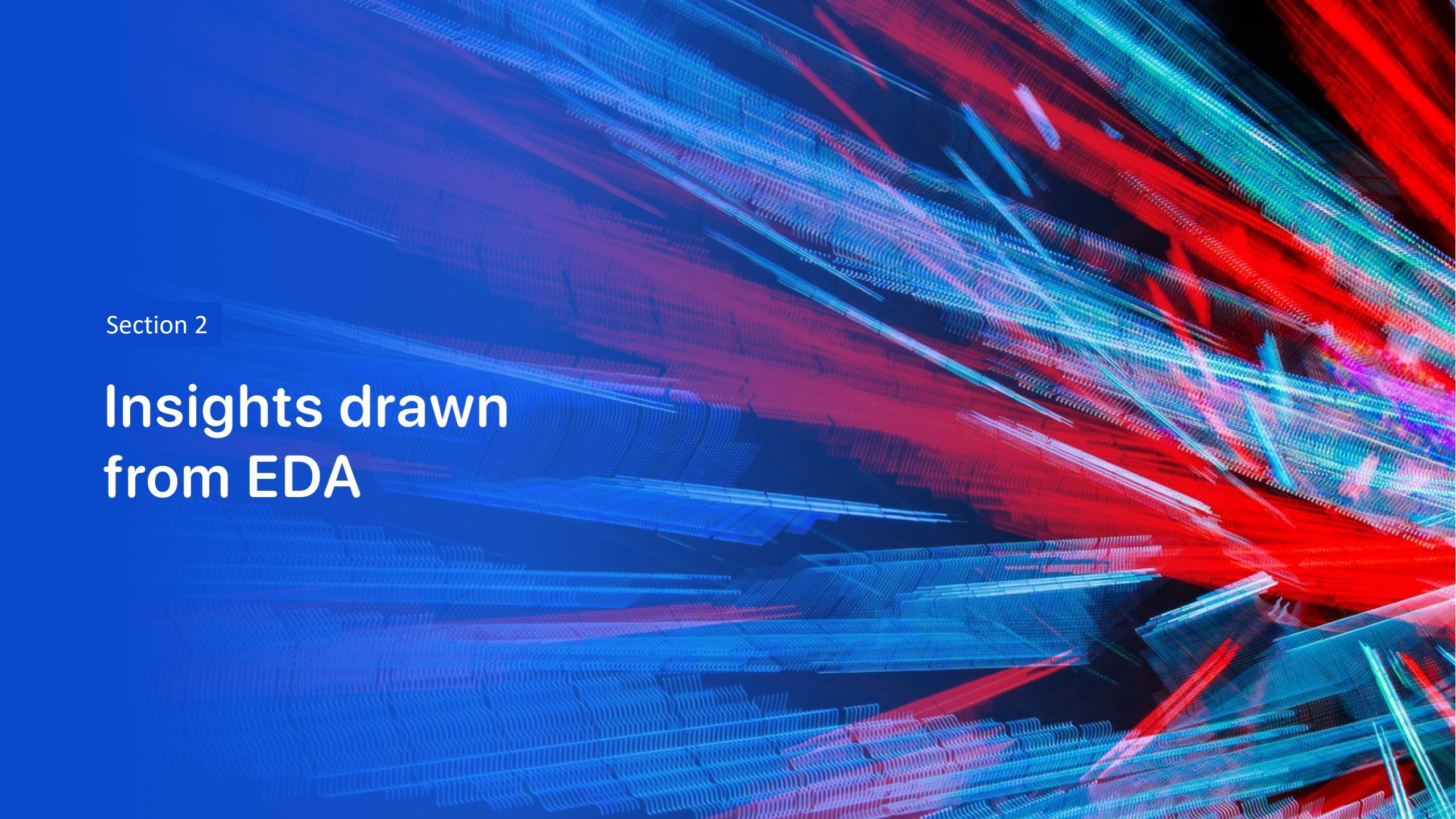
- Data preprocessing, which involved standardizing, fitting, transforming data, and splitting the data into training and testing sets.
- Grid Search was performed to find the best hyperparameters.
- Four models were trained using training set evaluated using accuracy score on the testing set and confusion matrix.

External reference: GitHub URL



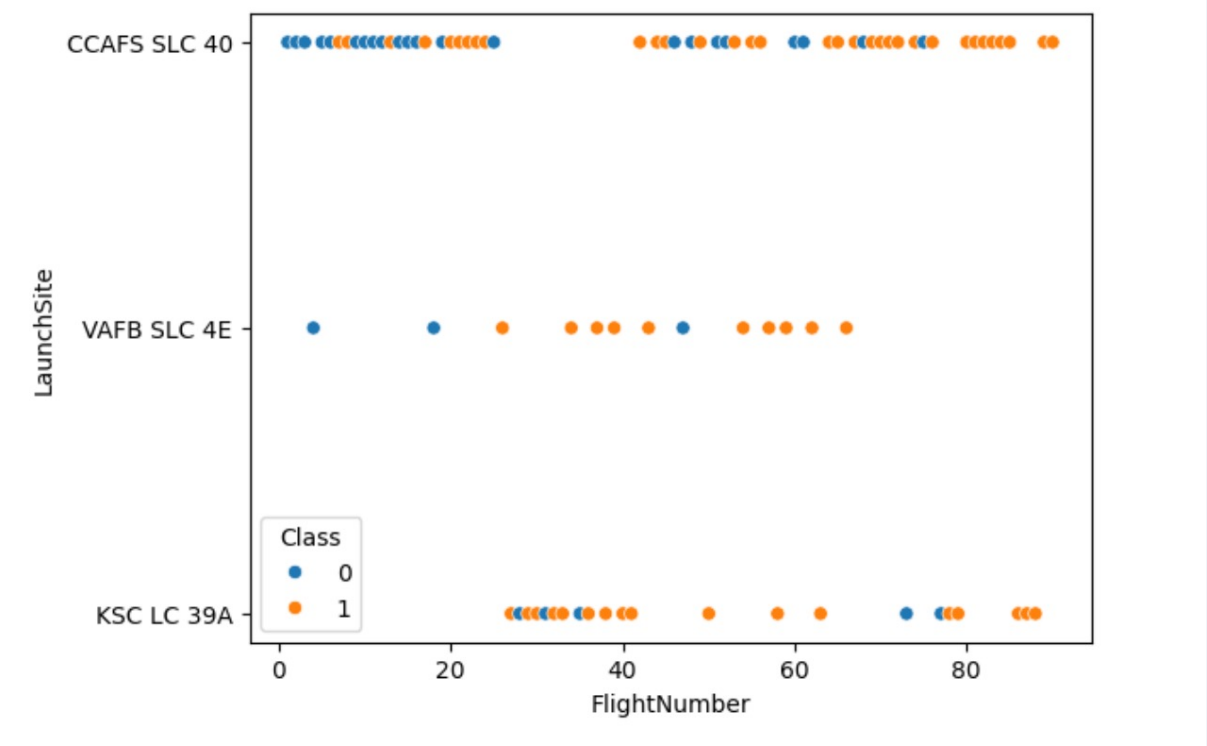
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results


Section 2
Insights drawn from EDA

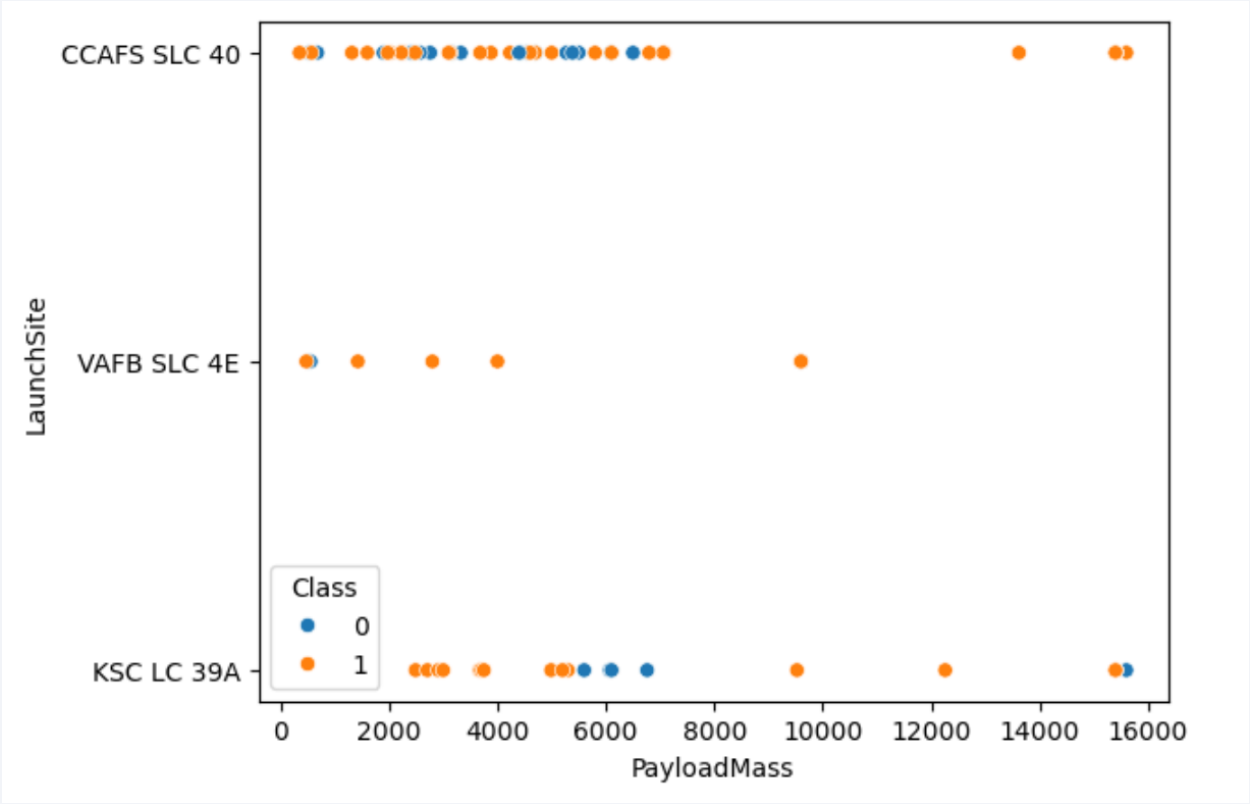
Flight Number vs. Launch Site

- Launches from site CCAFS SLC 40 are significantly higher than other sites.
- VAFB SLC 4E has fewer launches compared to other sites.
- Flight number 30 to 40 are mostly launched from KSC LC 39A.



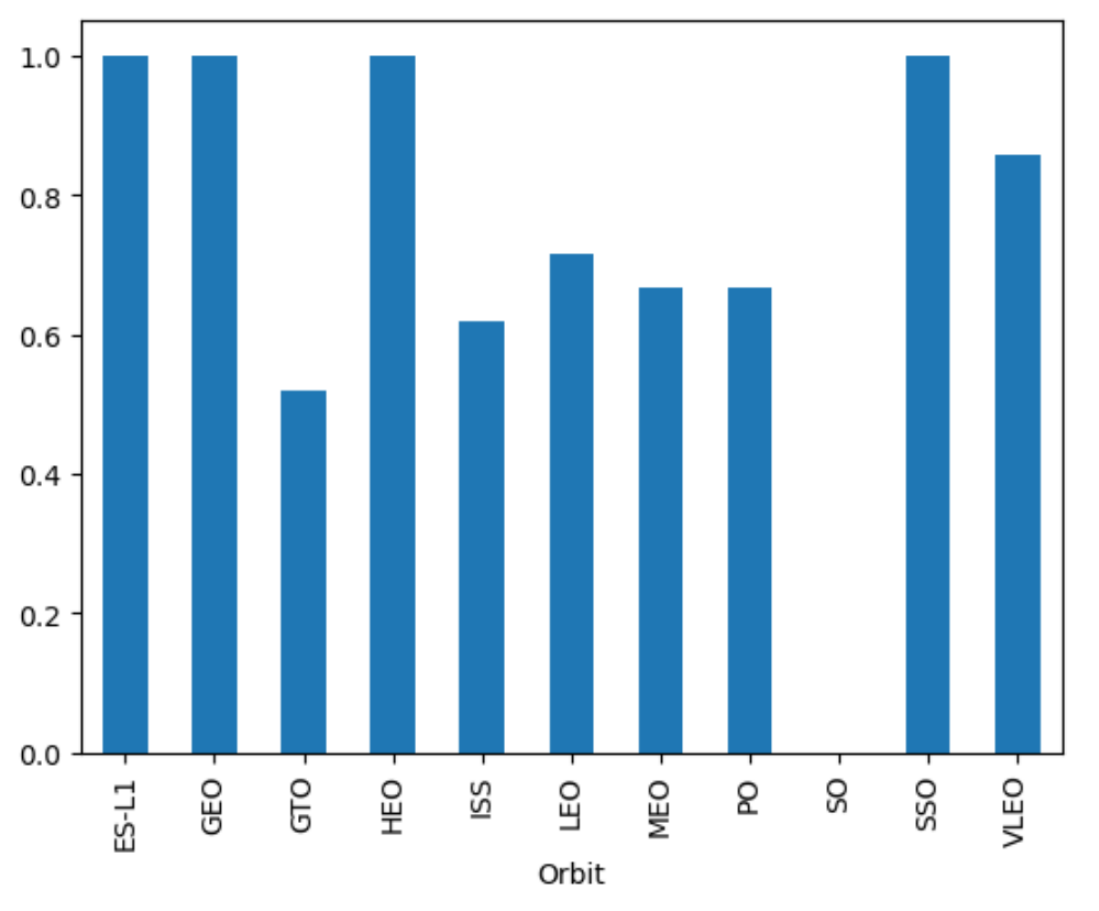
Payload vs. Launch Site

- CCAFS SLC-40 has carried the majority of payload masses, predominantly ranging between 0 and 8,000 kg, and shows a weak correlation with successful launches carrying heavier payload masses.
- VAFB SLC-4E has fewer launches and fewer failed launches.



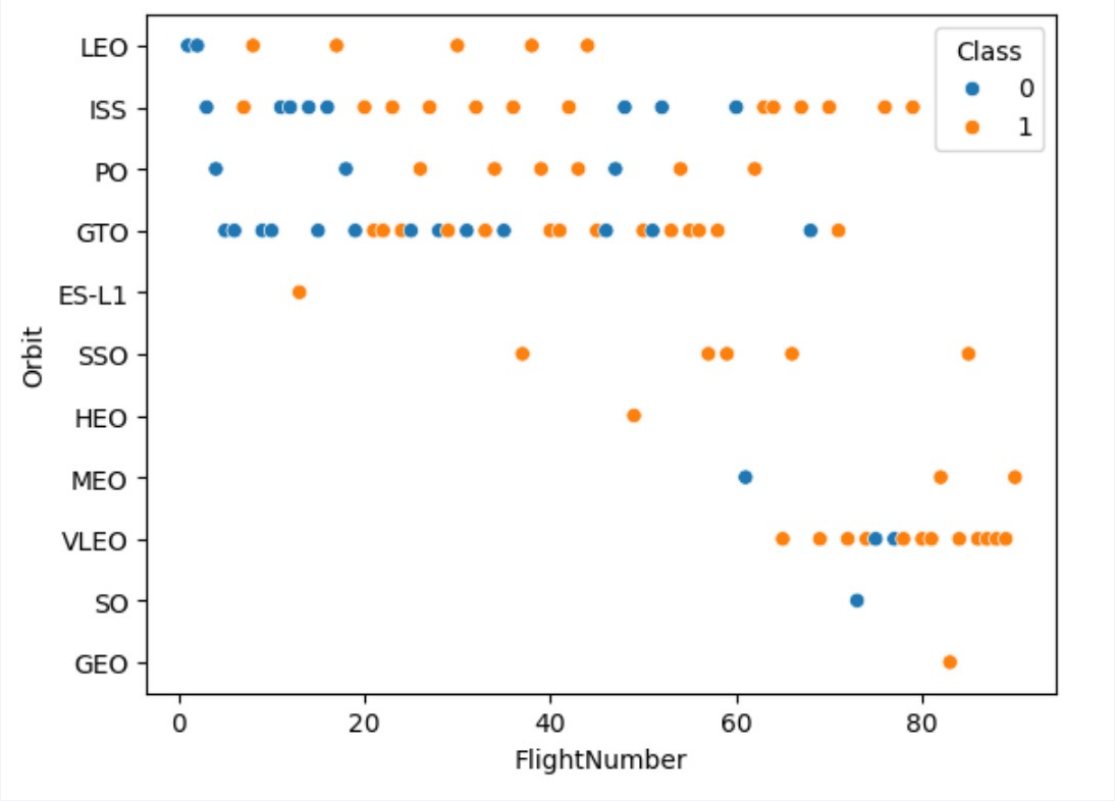
Success Rate vs. Orbit Type

- Orbit types ES-L1, GEO, HEO, and SSO have achieved a 100% success rate for launches, indicating that all missions targeting these orbits have been successful.
- In contrast, the GTO (Geostationary Transfer Orbit) has the lowest success rate at 50%, suggesting greater challenges with missions to this orbit.



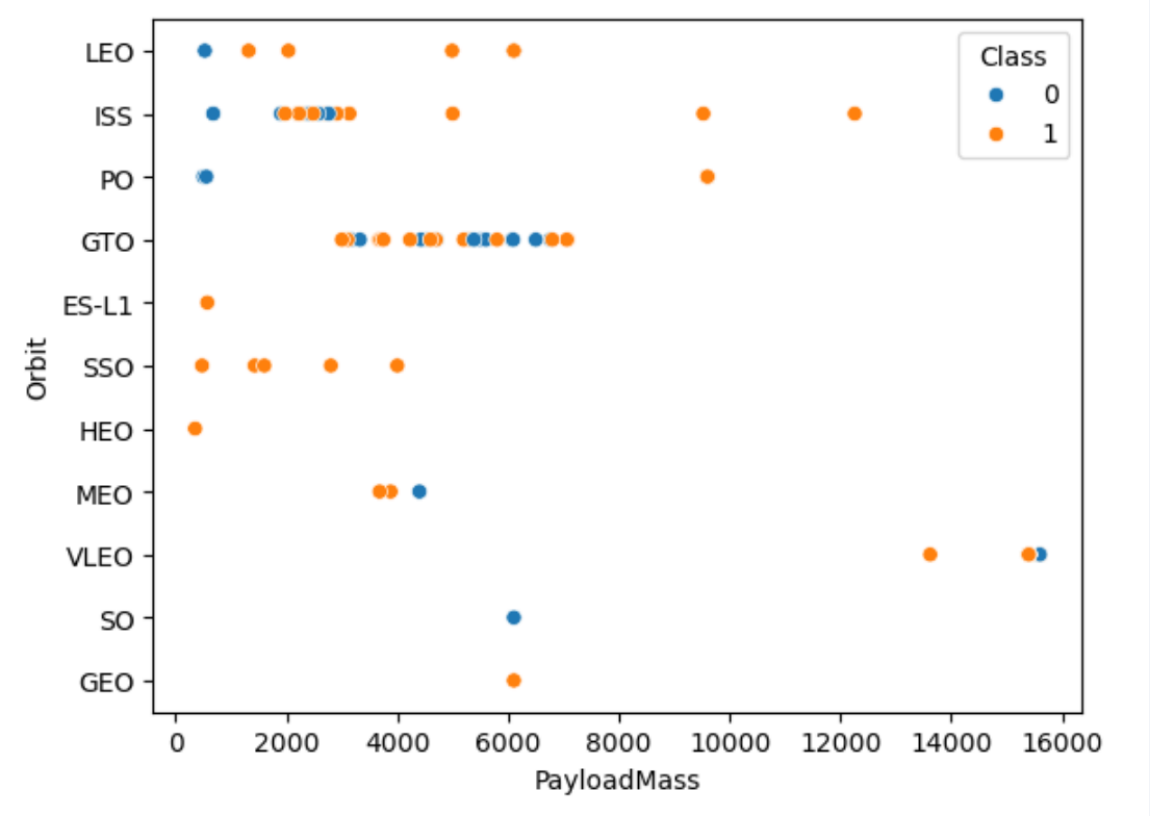
Flight Number vs. Orbit Type

- The plot shows that as the flight number increases, the likelihood of successful launches also tends to increase.
- Recent flight (FN > 60), seem to have a strong correlation with VLEO orbit type.



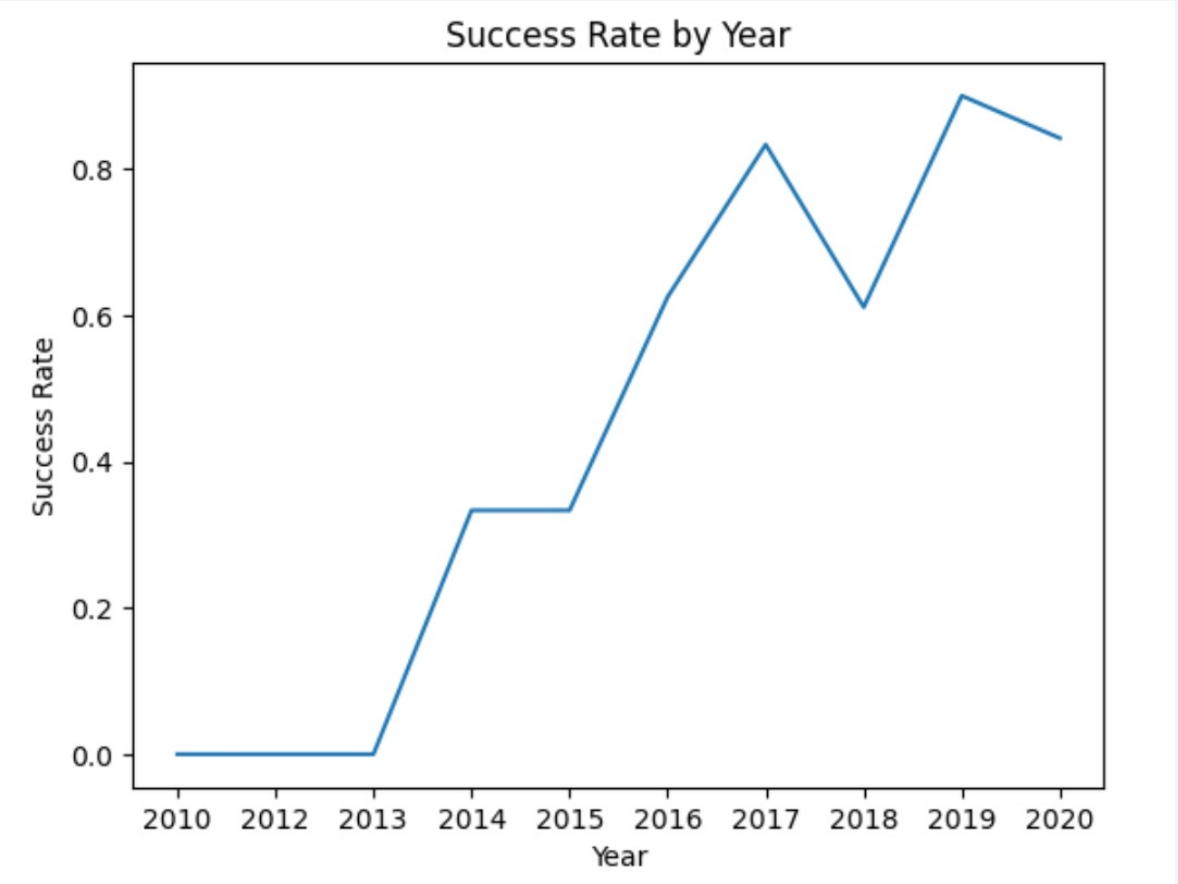
Payload vs. Orbit Type

- There are strong correlation between orbit type ISS and payload mass within range 2,000 to 4,000 kg.
- There are strong correlation between orbit type GTO and payload mass within range 4,000 to 8,000 kg.



Launch Success Yearly Trend

- The graph shows a clear upward trend in launch success rates from 2010 to 2020, with some fluctuations.
- The initial years show low success rates, but there is significant improvement starting from 2014, with the highest success rates achieved in the late 2010s.



All Launch Site Names

- There are four sites where SpaceX has conducted launches:
  - CCAFS LC-40 (Cape Canaveral Air Force Station Launch Complex 40),
  - VAFB SLC-4E (Vandenberg Air Force Base Space Launch Complex 4E), and
  - KSC LC-39A (Kennedy Space Center Launch Complex 39A).
  - CCAFS SLC-40 (Cape Canaveral Air Force Station Space Launch Complex 40)
- Each launch site represents a specific facility where SpaceX has conducted its rocket launches.

Launch Site Names Begin with 'CCA'

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 table displays data for five SpaceX missions launched from Cape Canaveral Air Force Station (CCAFS) Launch Complex 40 between 2010 and 2013, all using the Falcon 9 v1.0 booster.
- The payloads, ranging from 0 to 677 kg, included Dragon spacecraft, demo flights, with all missions targeting Low Earth Orbit (LEO), often for the ISS. Customers included SpaceX, NASA (under COTS and CRS programs), and the NRO. All missions were successful, with landing outcomes varying: two failures involving parachutes and three with no landing attempts.

Total Payload Mass

- The total payload mass carried by SpaceX boosters for NASA CRS missions is 45,596 kilograms. This indicates the substantial role SpaceX has played in transporting cargo to support NASA's missions, demonstrating their capability in handling significant payloads for space resupply operations.

Total_Payload
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by the booster version F9 v1.1 used by SpaceX is 2,928.4 kilograms. This figure reflects the typical capacity of the F9 v1.1 booster in terms of payload weight, indicating its efficiency and reliability in handling moderate-sized payloads for various missions.

Average_Payload
2928.4

First Successful Ground Landing Date

- The first successful landing on a ground pad occurred on December 22, 2015 at 1:29 UTC, with the booster version FB FT B1019 launched from CCAFS LC-40.

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2015-12-22	1:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

- The boosters that successfully landed on a drone ship and carried a payload mass within the specified range are:
  - F9 FT B1022
  - F9 FT B1026
  - F9 FT B1021.2
  - F9 FT B1031.2
- These boosters are part of the Falcon 9 Full Thrust (F9 FT) series, and their successful landings demonstrate SpaceX's capability to recover first stages even with relatively heavy payloads.

Total Number of Successful and Failure Mission Outcomes

- Out of all recorded outcomes, SpaceX has achieved a high success rate with 100 successful missions, with one of them being successful but the payload status remained unclear.
- There was one mission failure in flight.

Outcome	Count
Failure	1
Success	100

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

Boosters Carried Maximum Payload

- Out of the total booster versions listed in the dataset, there are 92 unique booster versions. The 12 booster versions mentioned have been identified as carrying the maximum payload mass in SpaceX's missions.

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

- Below are records for the failed landing outcomes in drone ship in 2015.

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

- Landing outcomes failure on a drone ship occurred twice in 2015; one failure happened in January and the other in April. The booster versions used were F9 v1.1 B1012 and F9 v1.1 B1015, both launched from CCAFS LC-40.

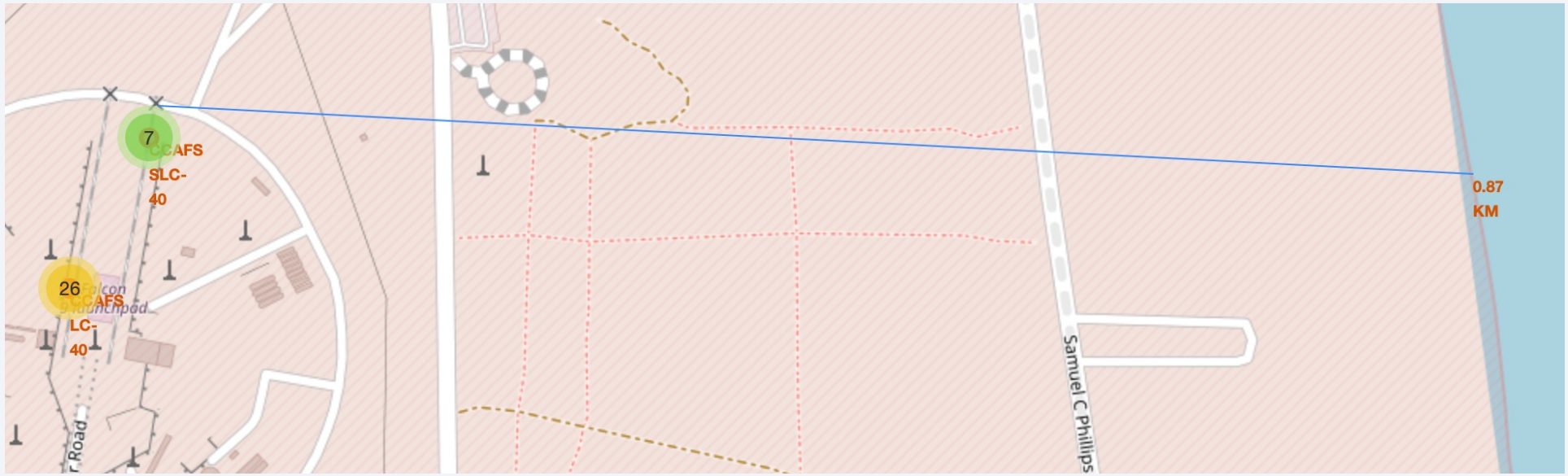
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- No attempt is the most occurring landing outcomes. Followed by an equal number of successes and failures on the drone ship, with each outcome occurring 5 times.

Landing_Outcome	Outcome_Count
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

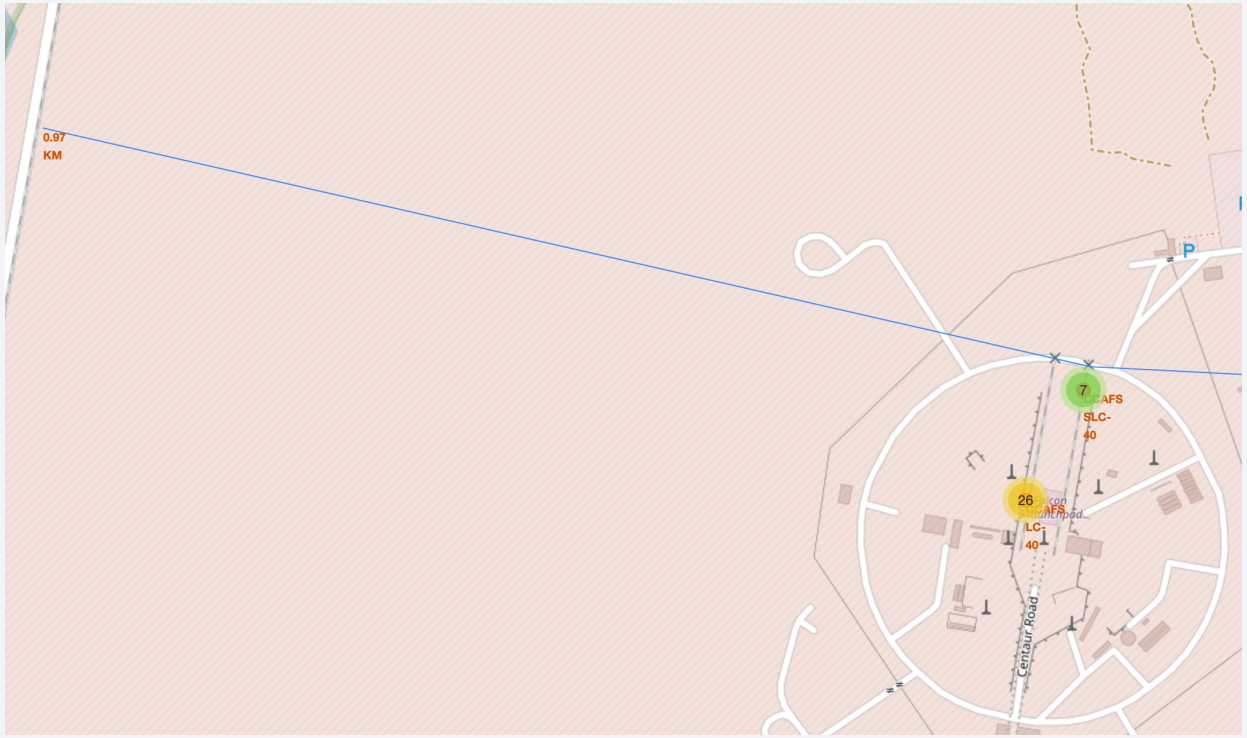
Closest Coastline to Launch Site



- The distance between CCAFS SLC-40 and nearest coastline is 0.87 KM.

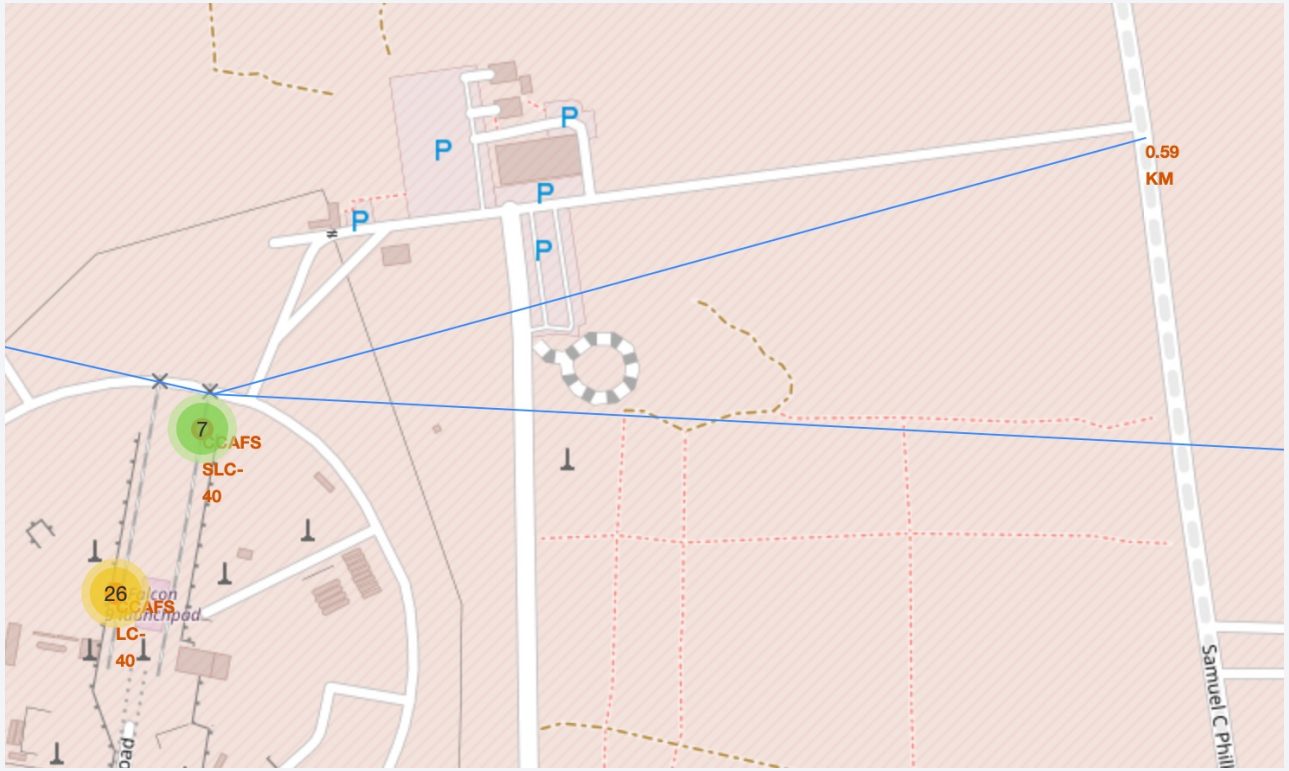
Closest Railway to Launch Site

- The distance between CCAFS SLC-40 and nearest railway is 0.97 KM.



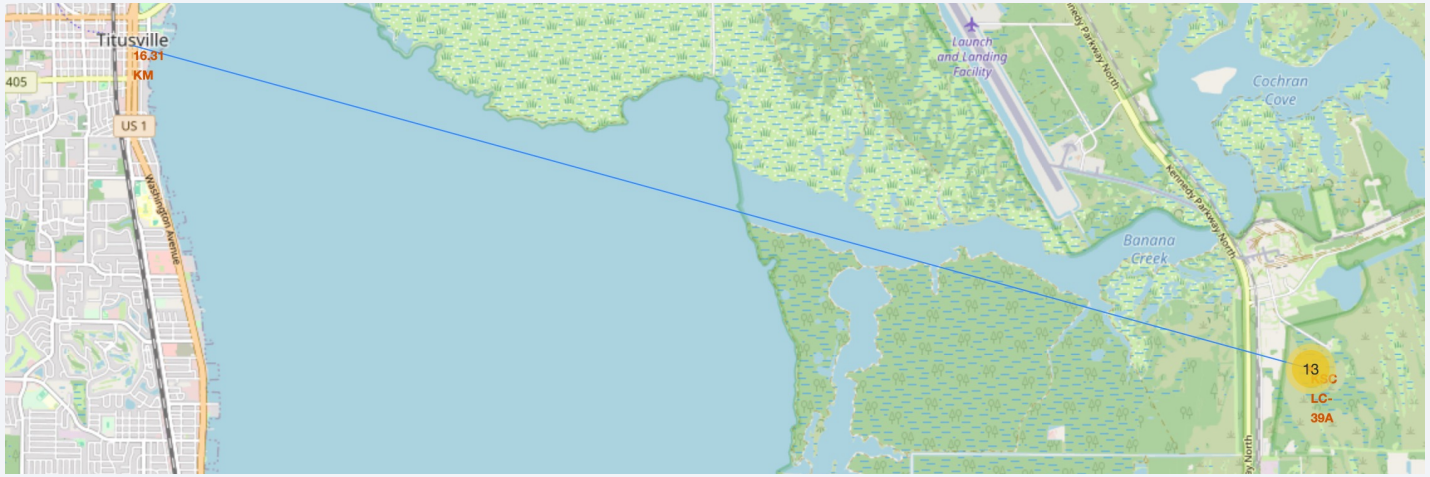
Closest Highway to Launch Site

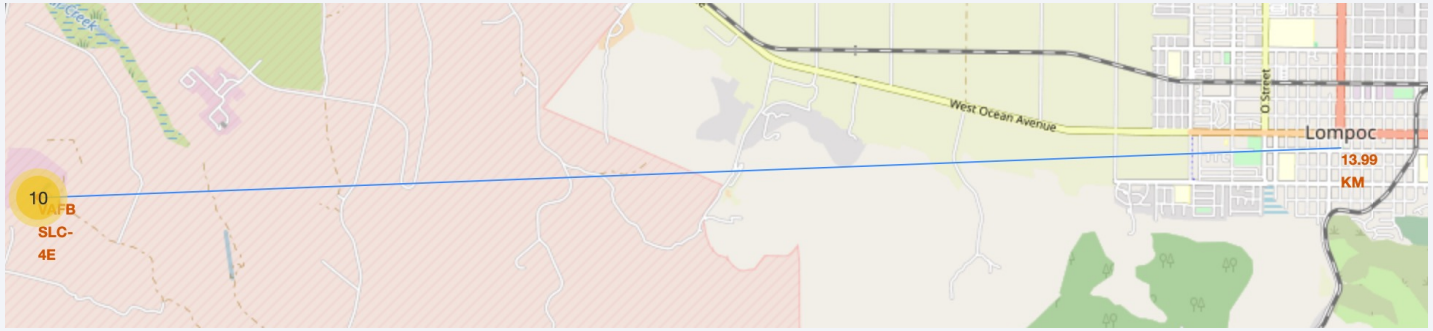
- The distance between CCAFS SLC-40 and nearest highway is 0.57 KM.



Closest City to Launch Site

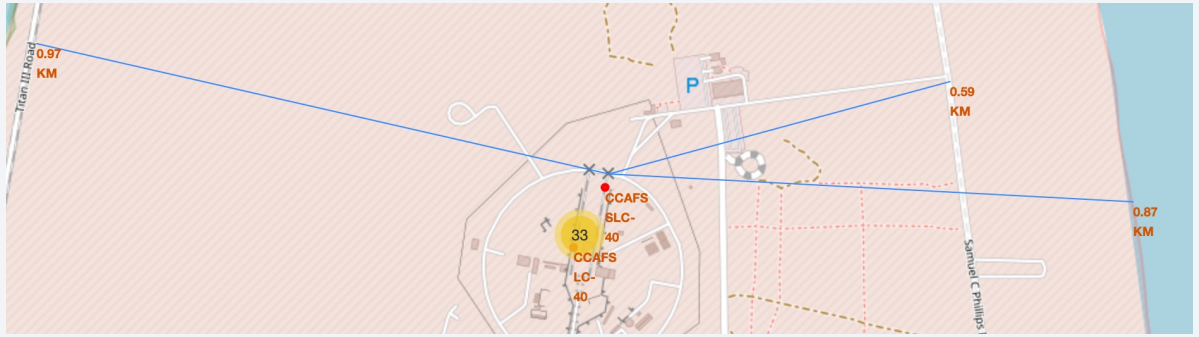
- Launch sites seem to keep certain distances away from cities. The distance between KSC LC-39A and nearest city, Titusville, is 16.31 KM, while VAFB SLC-4E to nearest city, Lompoc, is 14 KM.

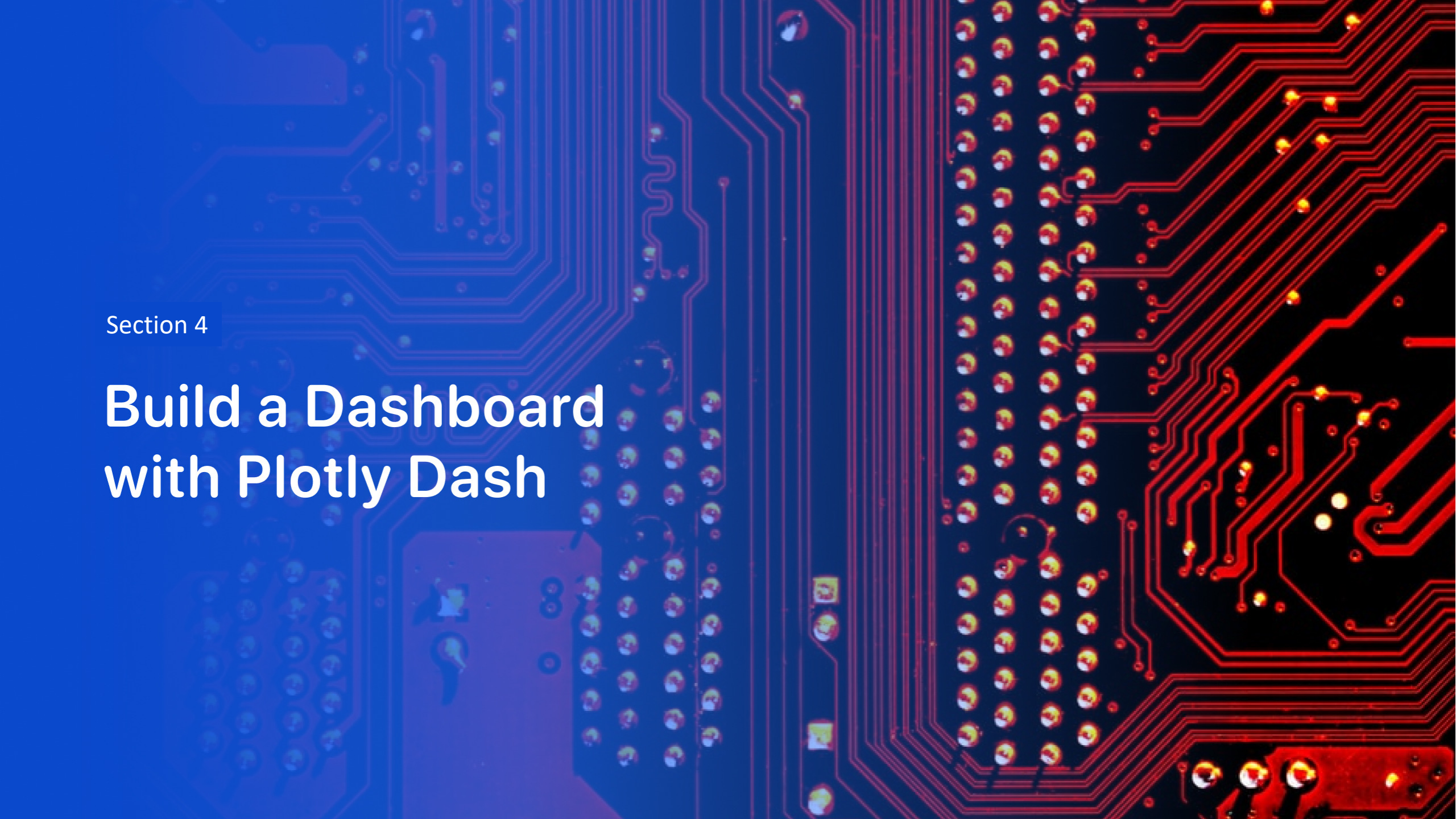




Additional Insights

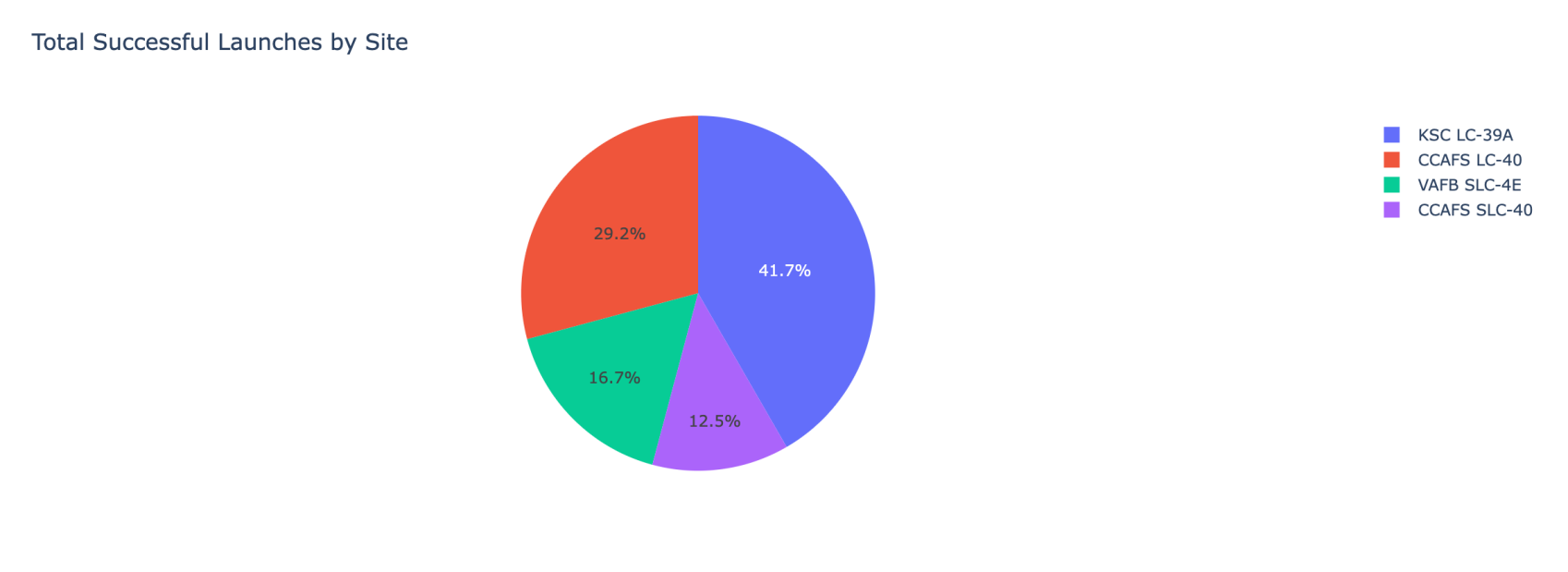
- The launch sites are in close proximity to railways, highways, and coastlines.




Section 4
Build a Dashboard with Plotly Dash

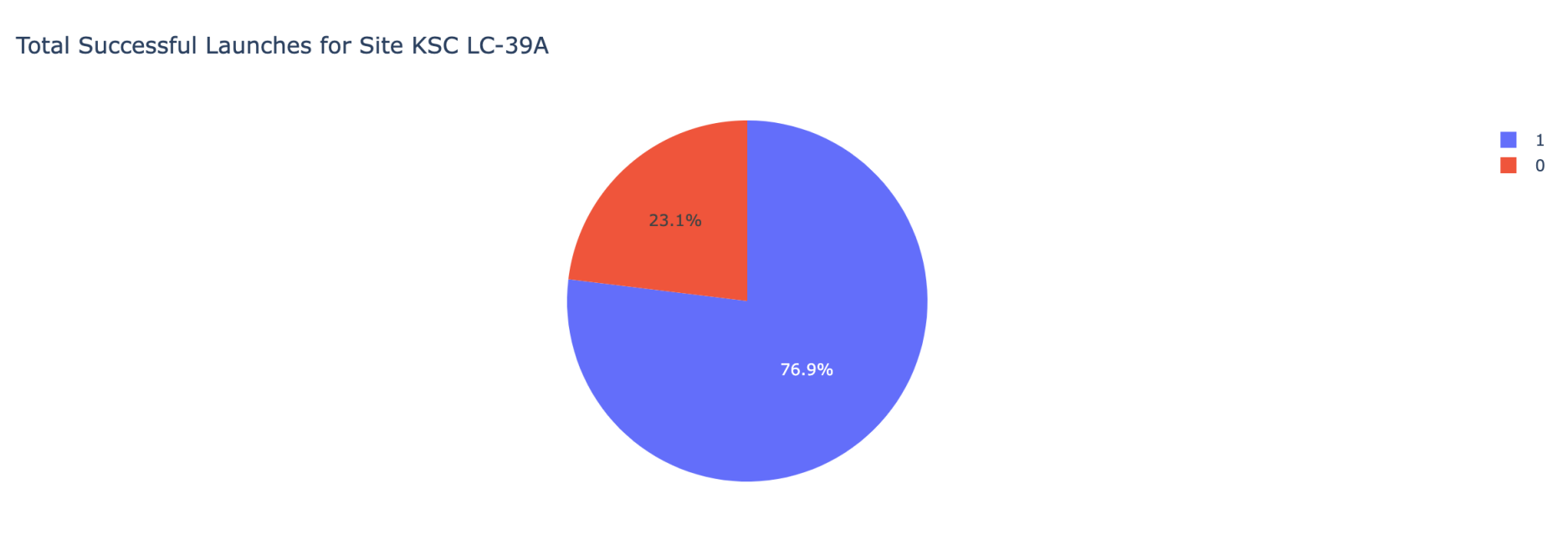
Total Successful Launches by Site

- Most successful launches, 41.7%, occurred at Kennedy Space Center Launch Complex 39A (KSC LC-39A). CCAFS LC-40 follows with 29.2%, VAFB SLC-4E, accounts for 16.7%, while CCAFS SLC-40 contributes 12.5% of the successes.



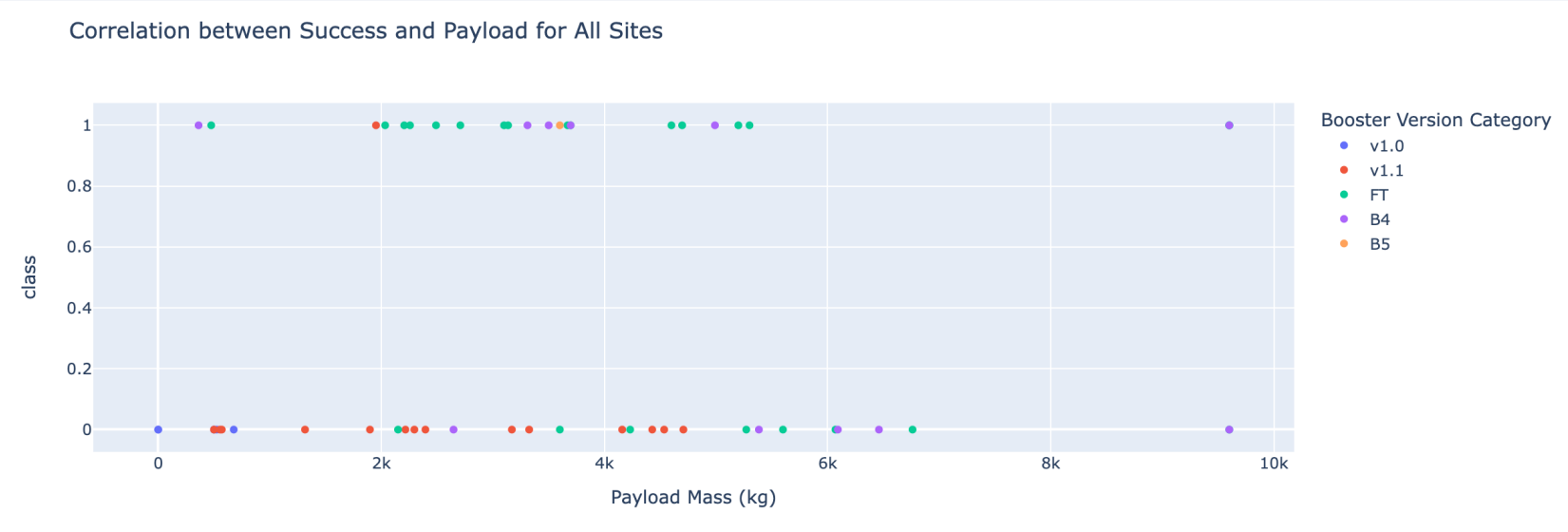
Site with Highest Launch Success Ratio

- Highest launch success ratio occurred at Kennedy Space Center Launch Complex 39A (KSC LC-39A) where 77% of launches landed successfully, and only 23% failed.



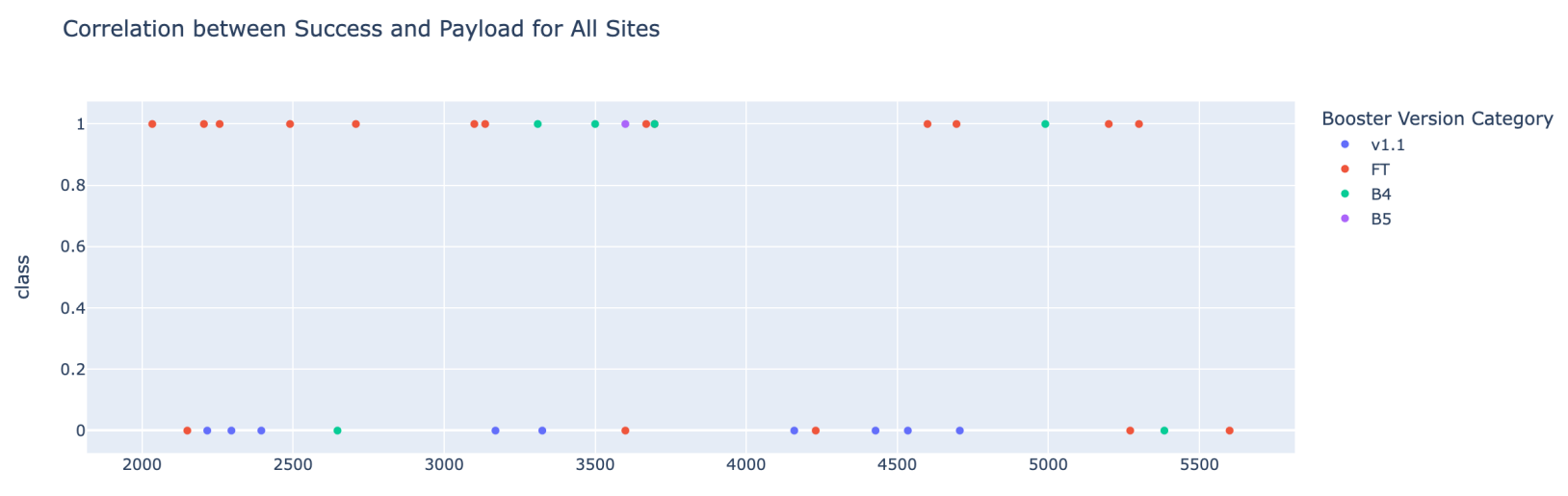
Payload vs Launch Outcome

- There seem to be a moderate and negative relationship between Payload and Launch Outcomes. Most successful launches carried payload masses within the range of 2k to 6k kilograms. Failed launches are spread across the entire range from 0 to 10k kilograms. The booster version v1.1 is the category associated with these failures.



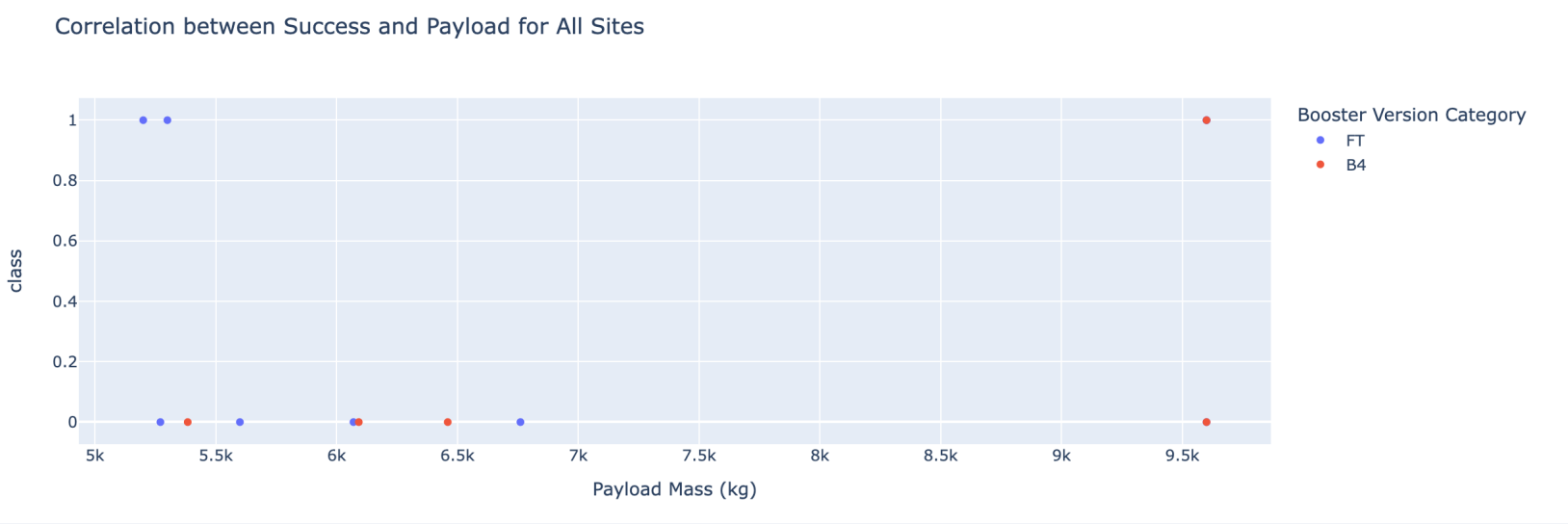
Medium Payload and FT Booster have the highest success rate

- Payload masses within the range of 2,000 to 6,000 kilograms have the highest success rate for launches, with the FT (Full Thrust) booster version dominating these successful missions. This combination highlights the FT booster version's effectiveness in handling medium payload masses, significantly contributing to SpaceX's overall mission success rate.



Heavier payload masses associated with higher failure rates

- Heavier payload masses, ranging from 5,000 to 10,000 kilograms, are associated with higher failure rates in launches. The booster versions that carried these heavy payloads were primarily the FT and B4 categories.

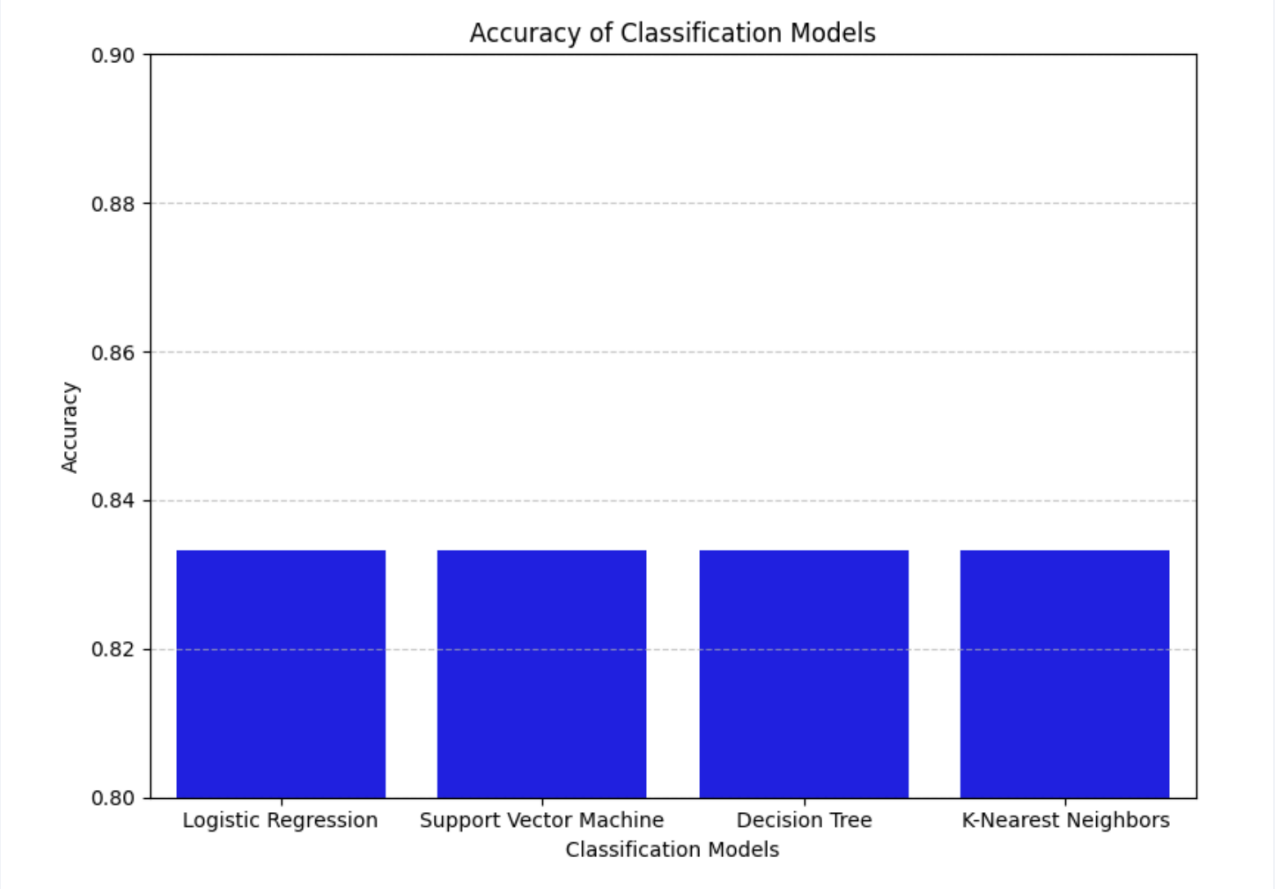


Section 5

Predictive Analysis (Classification)

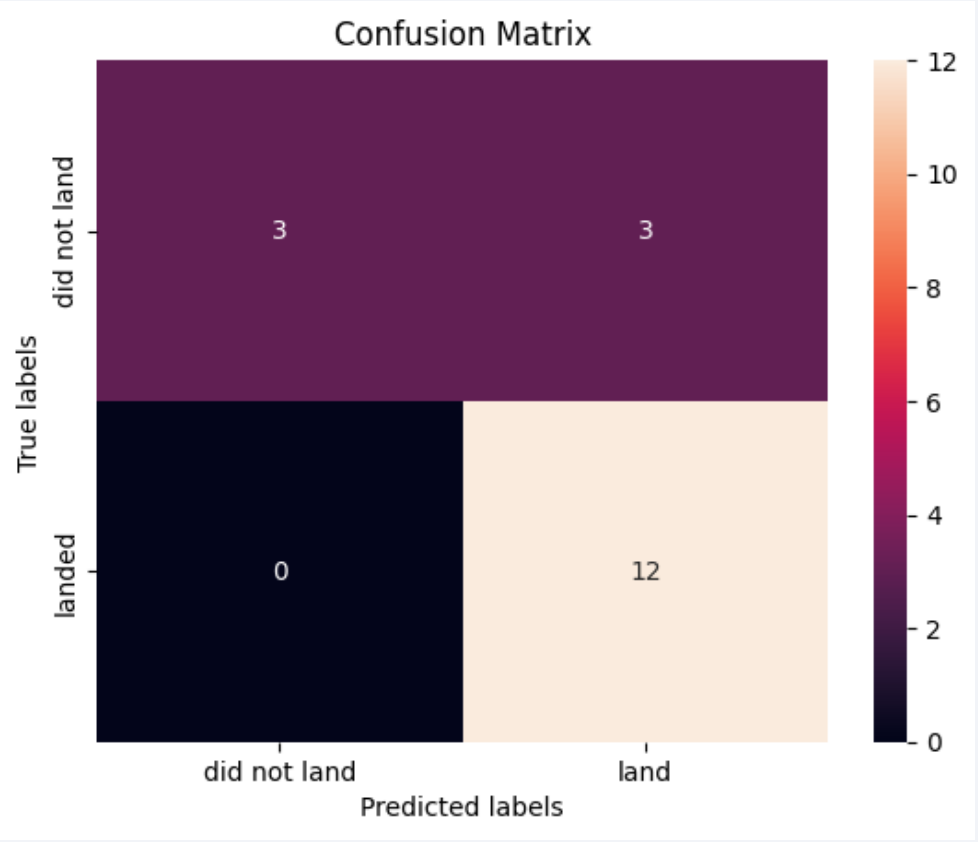
Classification Accuracy

- All four models have the same accuracy score, 0.83.
- There is no significant difference in performance among these models in terms of accuracy.



Confusion Matrix

- Four classification models have the same confusion matrix.
  - TP: Accurately predicted the majority of “landed”
  - FP: Misclassified "landed" as "did not land" for 3 instances.
  - TN: Accurately predicted "did not land" for 3 instances.
  - FN: There were no instance where the model predicted "did not land" when it "landed"



Conclusions

- To predict if first stage Falcon 9 successfully land, EDA, Interactive Visual Analytics, and Predictive Analysis were performed.
- Medium payloads performed better than higher payloads.
- KSC LC 39A had the most successful launches.
- Orbit types ES-L1, GEO, HEO, and SSO had the highest success rate.
- FT Booster had the highest success rate.
- The launch sites are in close proximity to railways, highways, and coastlines, each within approximately 1 km from the launch site. Additionally, the launch sites are situated away from cities, with distances of 16 km to Titusville and 13 km to Lompoc.
- There is no significant difference in performance among Logistic Regression, Decision Tree, K-Nearest Neighbor, and Support Machine Vectors.

Appendix

- https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- https://github.com/rizyanaptr/spacex

Thank you!