IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Yaw Bempong
11/18/2025

Outline

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

Executive Summary

This report details the strides made by SpaceX in the commercialization of their landed boosters by leveraging data science techniques . The primary goal of this project was to leverage advanced analytics to understand, predict, and ultimately improve the reliability and commercial viability of reusable launch vehicle technology. Collected data from APIs and scraping, prepared and cleaned it, drew insights, added impactful visualizations and built predictive(classification) models.

Key Findings

- There was a steep yearly progression in success rates from 2013 to 2016 followed by a slight plateau between 2017 and 2020
- The **Full Throttle(FT)** booster version introduced in 2015 has the highest success rate across different payloads.
- The **Kennedy Space Center(KSC LC-39A)** launch site has the highest success rate of all launch sites at **76%**
- **The SVM performed best on the test dataset and as such is the optimal model.** However, due to the nature of the test dataset, there is more to be done to evaluate the model especially on real-world data.
- The dataset is small, this limits the confidence in generalization

Introduction

The advent of the commercial space age has fundamentally shifted the paradigm of space travel, driving down costs and making orbital access more attainable for a broader range of enterprises and governments. **SpaceX** remains a leading innovator and influence in the recent space travel ecosystem. **SpaceY,** a rising competitor, seeks to replicate the successes of SpaceX.

A core competitive advantage enabling SpaceX's success is the relatively inexpensive cost of its launch services. This significant cost reduction is largely attributable to SpaceX's capability to reuse the first stage of its rockets.

Objective:

Will the first stage successfully land?

The objective of this project is to analyze historical launch data for SpaceX to develop a predictive model capable of determining the probability of a successful first-stage landing.

By accurately predicting the landing outcome, we can precisely determine the true cost of a launch and identify key factors that optimize the success rate of booster recovery, thereby enhancing SpaceY's commercial advantage.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from the SpaceX api endpoints such as launches, payload and core
  - Data was collected by scraping the launches Wikipedia page
- Perform data wrangling
  - The data was combined using python and pandas
- Exploratory data analysis (EDA) was done using visualization and SQL
- Interactive visual analytics were done using Folium and Plotly Dash
- Predictive analysis using classification models such as SVM, Decision Trees, and Logistic Regression
  - Models were built, trained on a training dataset and tested on a test set.

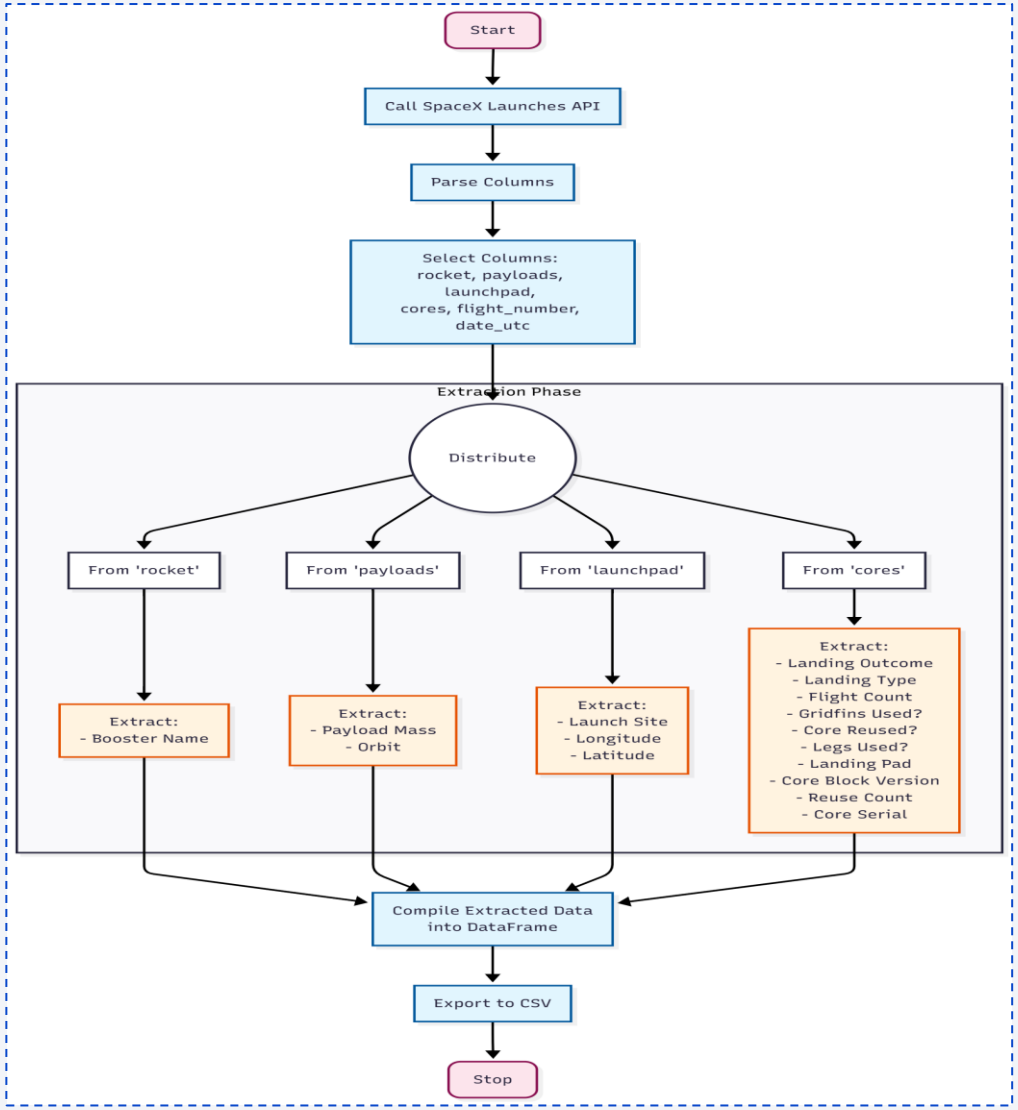
Data Collection

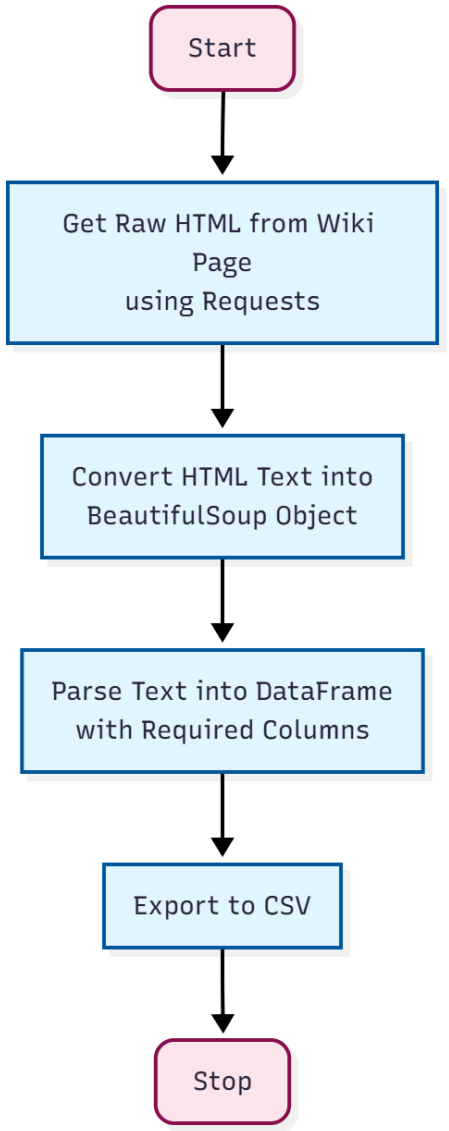
- Data was collected from the SpaceX api endpoints such as launches, payload and core
- Data was collected by scraping the SpaceX launches Wikipedia page

Data Collection – SpaceX API

- Using the launches, launchpads, payloads and cores api endpoint, compile a dataframe of needed columns.

- API datacollection notebook





Data Collection - Scraping

- Using beautifulsoup (bs4), python and pandas. Extract tables from the raw html files
- web scraping data collection notebook

Data Wrangling

- Data Processing Highlights
  - Exploratory Analysis: Loaded dataset and identified data types; detected ~29% missing values in LandingPad.
  - Feature Inspection: Quantified launch distribution across LaunchSite and Orbit categories.
  - Label Encoding: Converted categorical Outcome into a binary Class target for supervised learning.1 = Successful Landing, 0 = Unsuccessful Landing.
  - Validation: Calculated a baseline success rate of 66.7% before exporting.
- Data Wrangling Notebook

EDA with Data Visualization

- Scatter Plots (Relationships)
  - Variables: Analyzed Flight Number and Payload Mass against Launch Site and Orbit.
  - Insights: Revealed correlations between mission experience, payload capacity, and landing success. Identified usage patterns of launch sites and the impact of payload weight on different orbits.
- Bar Chart (Comparison)
  - Variables: Success Rate vs. Orbit Type.
  - Insights: Compared reliability across orbits, highlighting those with 100% success rates (e.g., ES-L1, SSO) to identify key predictive features.
- Line Chart (Trends)
  - Variables: Success Rate vs. Year.
  - Insights: Demonstrated a clear upward trend in mission success from 2013 to 2020, validating the maturation of SpaceX's landing technology.
- EDA and Visualization notebook

EDA with SQL

- **List Unique Launch Sites:** Selected distinct Launch_Site from the SPACEXTABLE.
- **Filter Launch Sites:** Selected 5 records where Launch_Site starts with 'CCA'.
- **Total Payload Mass by NASA (CRS):** Calculated the sum of PAYLOAD_MASS__KG_ for missions where the customer starts with 'NASA (CRS)'.
- **Average Payload Mass:** Calculated the average PAYLOAD_MASS__KG_ for the booster version 'F9 v1.1'.
- **First Successful Landing:** Retrieved the minimum Date for a successful landing on a ground pad.
- GitHub SQL_EDA

EDA with SQL

- **Drone Ship Success:** Selected booster versions with successful drone ship landings and payload mass between 4000 and 6000 kg.
- **Total Outcomes:** Counted the total number of mission outcomes.
- **Maximum Payload Boosters:** Selected booster versions that carried the maximum payload mass using a subquery.
- **Failures in 2015:** Listed month names, landing outcomes, booster versions, and launch sites for failed drone ship landings in 2015.
- **Rank Landing Outcomes:** Ranked the count of landing outcomes between '2010-06-04' and '2017-03-20' in descending order.
- EDA_SQL notebook

Build an Interactive Map with Folium

- **Folium Objects**
  - **Markers**
    - **NASA Johnson Space Center Marker:** A marker was added to the initial map to highlight the starting location.
    - **Launch Site Markers:** Markers were created for each launch site to pinpoint their exact geographical locations.
    - **Success/Failure Markers:** Markers colored green (success) and red (failure) were added to indicate the outcome of individual launches.
    - **Distance Marker:** A marker displaying the calculated distance was placed on a specific point of interest, such as the closest coastline.
  - **Circles**
    - **NASA Johnson Space Center Circle:** A circle was added around the NASA Johnson Space Center to highlight the area.
    - **Launch Site Circles:** Circles were added around each launch site coordinate to visually emphasize the launch site areas on the map.
- **Reasoning for Added Objects**
  - **Markers:** Markers were used to pinpoint specific locations of interest, such as launch sites and the NASA center, providing a clear visual reference for their exact positions
  - **Circles:** Circles were added to highlight the general area surrounding key locations.
- Map Visualization Folium notebook

Build an Interactive Map with Folium

- **Folium Objects**
  - **Marker Clusters**
    - **Marker Cluster Object:** A MarkerCluster object was used to group individual launch outcome markers. This prevents the map from becoming cluttered when multiple markers (launch records) share the same or very close coordinates.
  - **PolyLines**
    - **Equator Line:** A red PolyLine was drawn representing the Equator to visualize the launch sites' proximity to it.
    - **Distance Line:** A PolyLine was drawn connecting a launch site (e.g., CCAFS SLC-40) to a point of interest (e.g., the closest coastline) to visualize the distance between them.
  - **MousePosition**
    - **MousePosition Control:** A MousePosition plugin was added to the map to display the latitude and longitude coordinates of the mouse cursor dynamically. This tool aids in finding coordinates for other points of interest like railways or highways.
- **Reasoning for Added Objects**
  - **Marker Clusters:** A marker cluster aggregates these points, simplifying the map visualization while allowing users to zoom in to see individual records.
  - **PolyLines:** The Equator line helps analyze if launch sites are positioned near the equator for physical advantages. The distance line provides a direct visual representation of the proximity between a launch site and critical infrastructure or natural features like coastlines.
  - **MousePosition:** It allows users to easily determine the coordinates of any point on the map by hovering the mouse pointer.
- Map Visualization Folium notebook

Build a Dashboard with Plotly Dash

- **Dashboard Features**
  - **Pie Chart (Total Success Launches)**
    - **Visualization:** A pie chart that dynamically updates based on the user's selection.
      - **Default View (All Sites):** Displays the proportion of total successful launches contributed by each launch site.
      - **Specific Site View:** Displays the ratio of Successful (1) vs. Failed (0) launches for the selected site.
    - **Interaction:** Driven by a Dropdown menu (site-dropdown) that allows users to select "All Sites" or a specific launch site.
    - **Reasoning:** To provide a high-level overview of mission success. It allows users to instantly compare the performance contribution of different sites or drill down into the reliability of a single site.
  - **Scatter Chart (Correlation between Payload and Success)**
    - **Visualization:** A scatter plot with Payload Mass (kg) on the x-axis and class (Success/Failure) on the y-axis.
      - Points are color-coded by Booster Version Category.
    - **Interaction**: Driven by two inputs:
      - **Dropdown menu (site-dropdown):** Filters data by launch site.
      - **Range Slider (payload-slider):** Filters data to show only missions within a specific payload mass range (0kg - 10,000kg).
    - **Reasoning:** To analyze the relationship between payload weight and mission outcome. The color coding helps identify if specific booster versions perform better with certain payload ranges, and the slider allows users to isolate specific payload scenarios (e.g., heavy vs. light).
- Dash App Source code

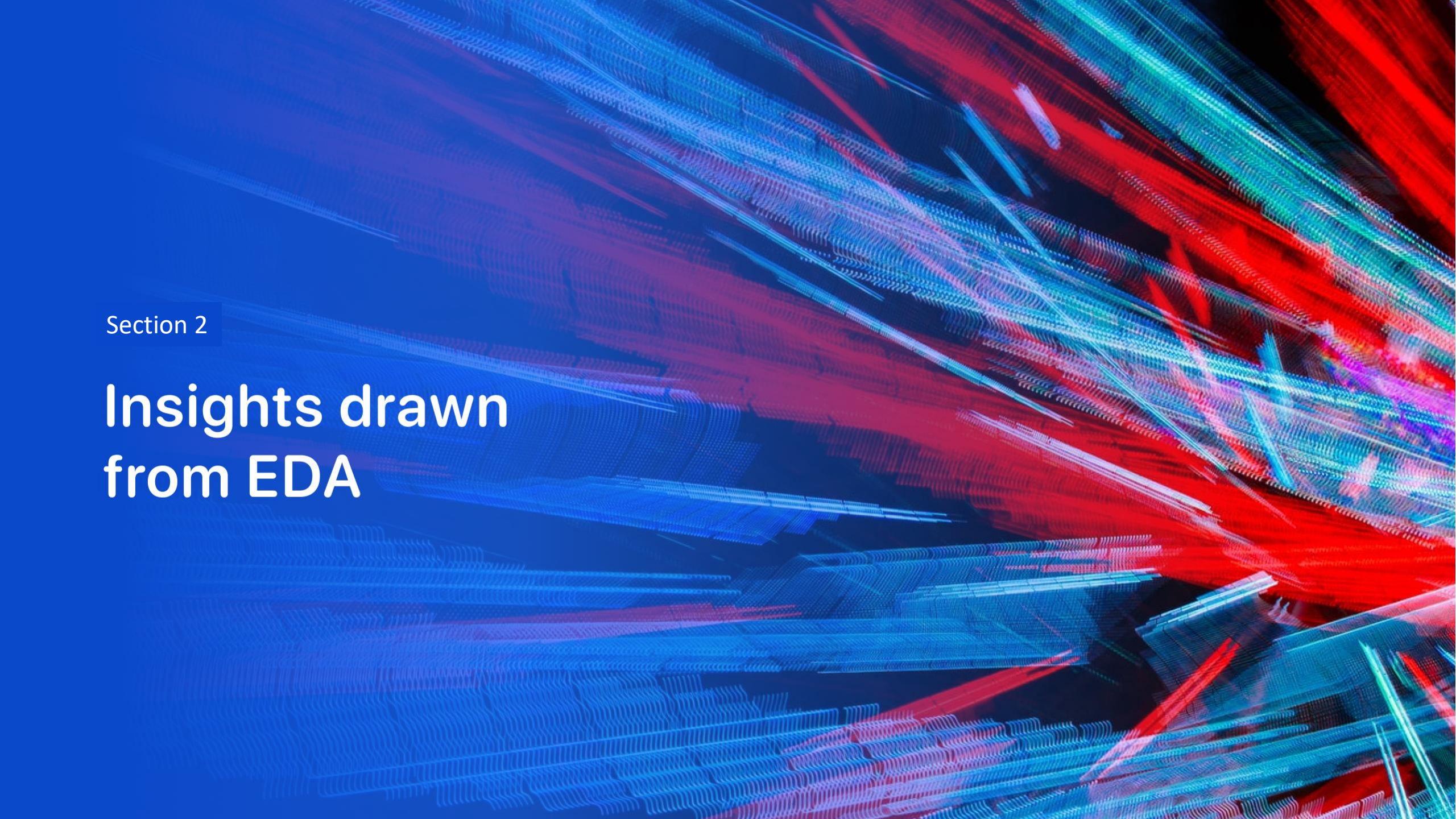
Predictive Analysis (Classification)

To determine the best performing classification model for predicting landing success, the following steps were taken:

- **Data Preparation:** The dataset was loaded, and the target variable (Class) was isolated.
- **Standardization:** The feature data (X) was standardized using StandardScaler to ensure all features contributed equally to the model.
- **Splitting:** The data was split into training (80%) and testing (20%) sets using train_test_split.
- **Model Tuning:** Four different algorithms were trained using **GridSearchCV** with Cross-Validation to find the best hyperparameters:
  - **Logistic Regression:** Tuned C, penalty, and solver.
  - **Support Vector Machine (SVM):** Tuned kernel, C, and gamma.
  - **Decision Tree:** Tuned criterion, splitter, max_depth, etc.
  - **K-Nearest Neighbors (KNN):** Tuned n_neighbors, algorithm, and p.
- **Evaluation:** Each model was evaluated on the Test Set.
  - **Result:** All four models achieved an identical accuracy score of approximately **83.33%** on the test data.
  - **Confusion Matrices:** Visualized to identify false positives and false negatives (all models behaved similarly).
- Predictive Model building notebook

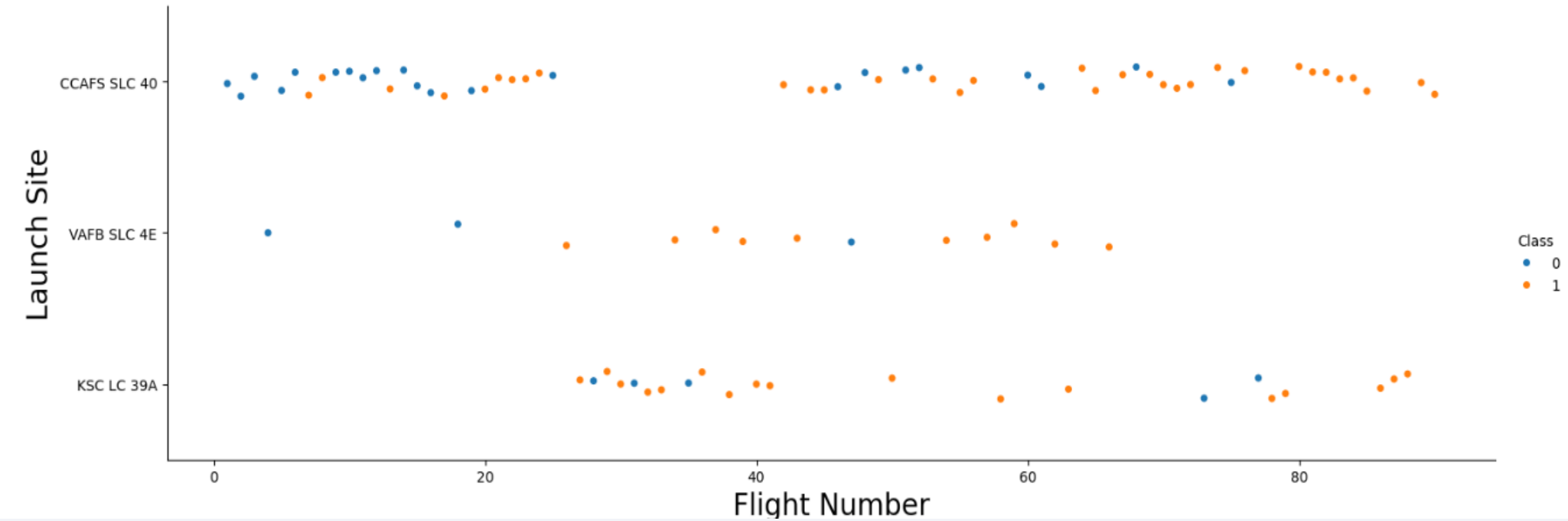
Results

- There was a steep yearly progression in success rates from 2013 to 2016 followed by a slight plateau between 2017 and 2020
- The Full Throttle(FT) booster version introduced in 2015 has the highest success rate across different payloads.
- The KSC LC-39A launch site has the highest success rate of all launch sites at 76%
- The Decision Tree classifier performed best out of all the classification models tested with 87.5% training accuracy. The **SVM performed best on the test dataset and as such is the optimal model*.** However, due to the nature of the test dataset, there is more to be done to evaluate the model especially on real-word data.


Section 2
Insights drawn from EDA

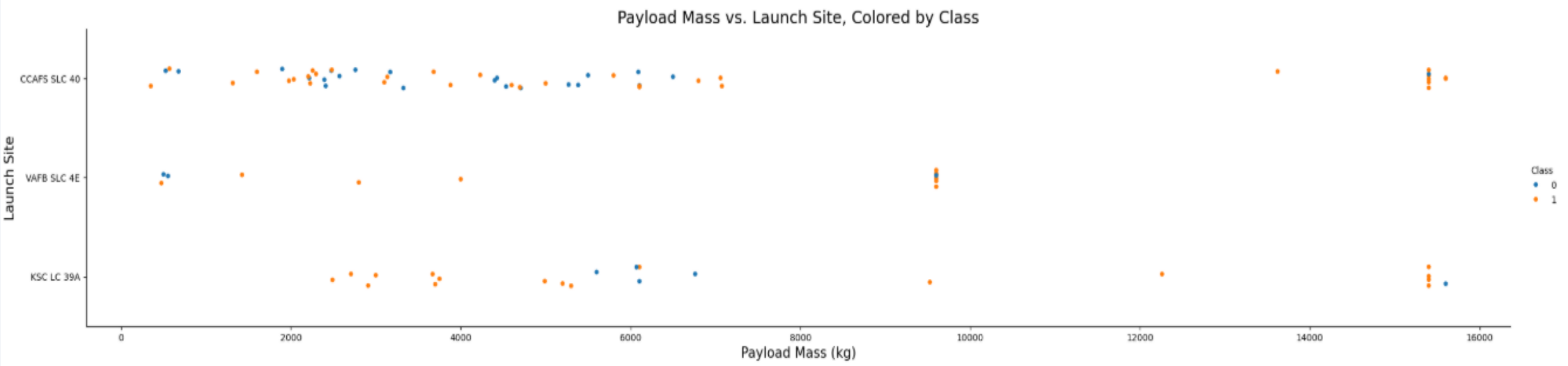
Flight Number vs. Launch Site

- Lower flight numbers (earlier) were concentrated at the Cape Canaveral launch site or “Slick Forty” (CCAFS SLC 40). This indicates it was the initially preferred launch site by SpaceX. It has remained preferred till date with a few launches moved to the Kennedy Space Center’s Launch Complex 39A in recent times.
- The Vandenberg launch site sees the fewest launches and even fewer for the higher (later) flight numbers



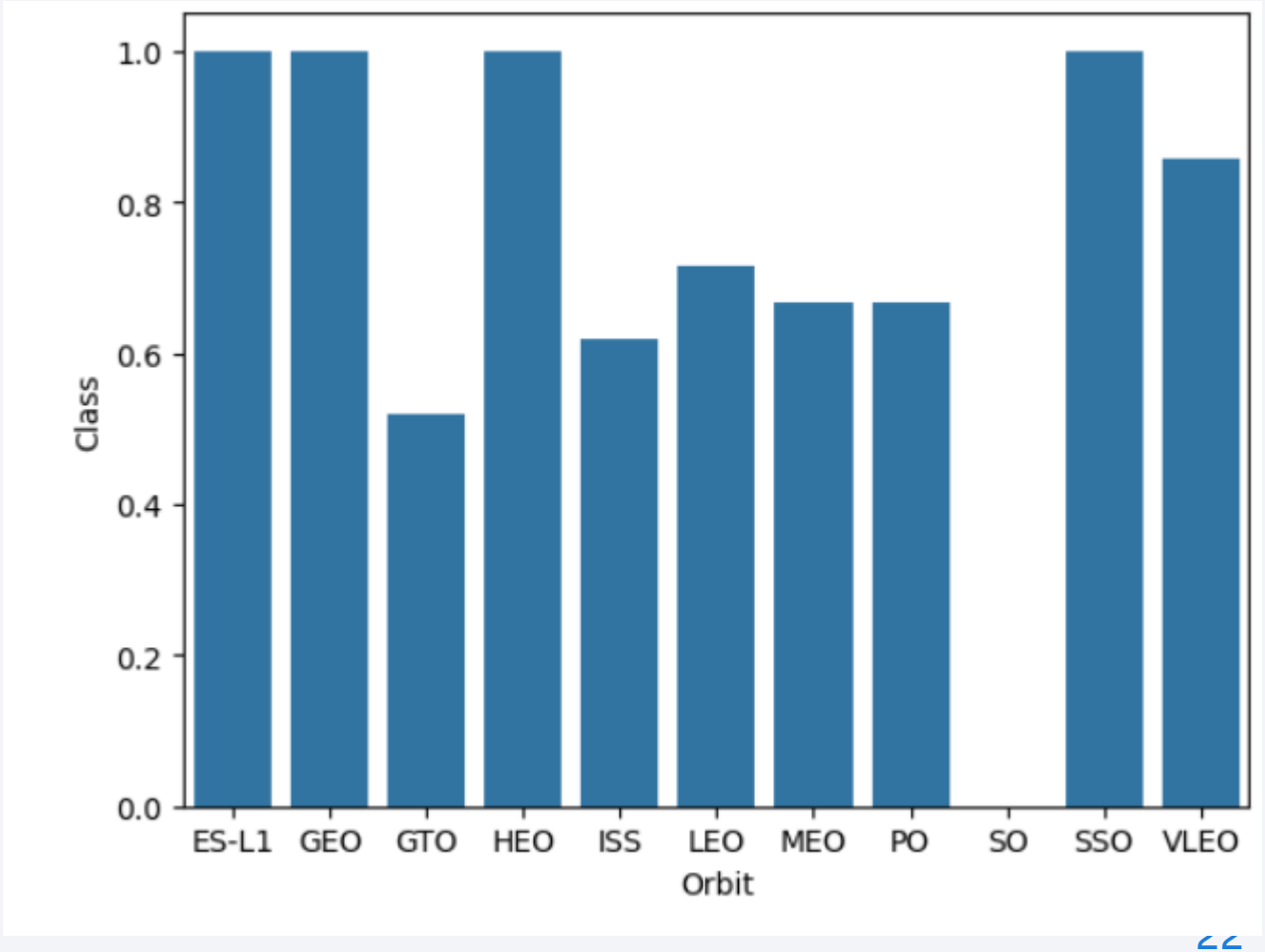
Payload vs. Launch Site

- Below 8000kg payload mass, launches are done from all three sites. The Vandenberg site sees very few flight. “Slick Forty” remains the preferred with the Kennedy Launch center a moderate second.
- Between 8000kg and 14000kg payload mass, the Vandenberg site is the default with almost no launches from the other sites.
- Above 14000kg payload mass, launches are done from only two sites. The Vandenberg site sees no flights from SpaceX. “Slick Forty” remains the preferred with the Kennedy Launch center a moderate second.



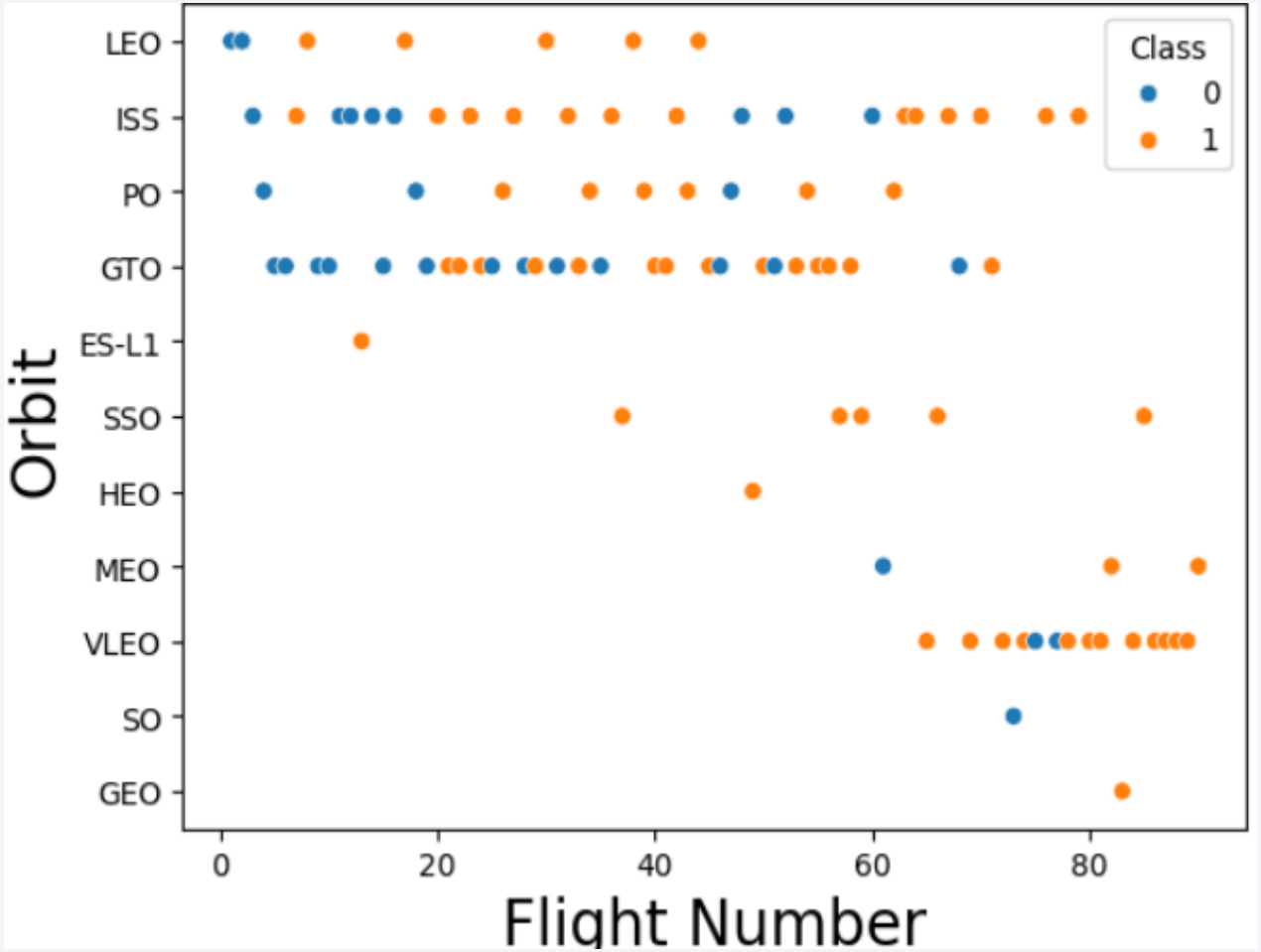
Success Rate vs. Orbit Type

- The Geostationary Transfer Orbit is the least successful and most costly SpaceX launch orbit



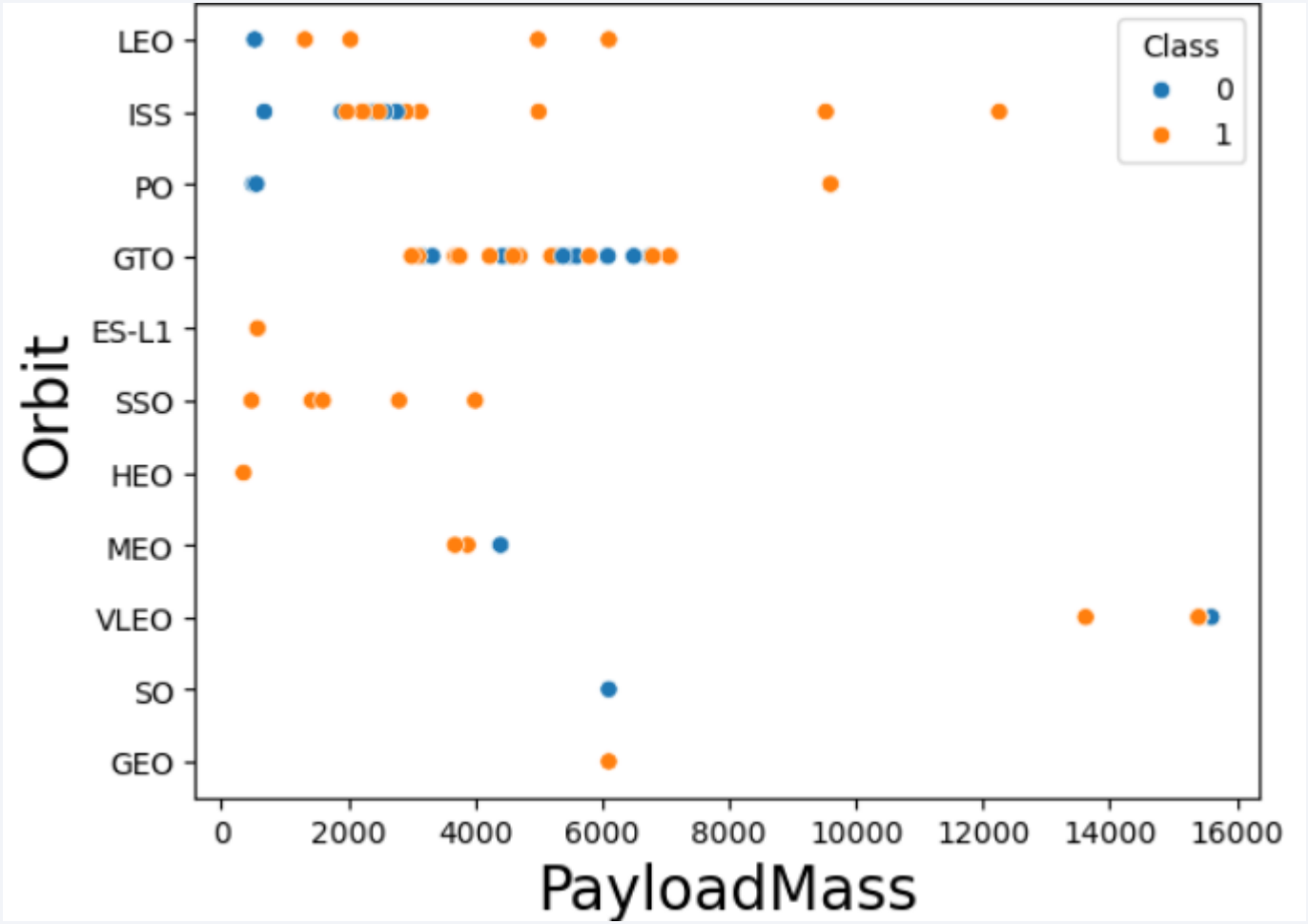
Flight Number vs. Orbit Type

- There are consistent launches to International Space Station(ISS) orbit across all flight numbers (times)
- Later flight numbers have featured the Very Low Earth Orbit which is evident in SpaceX use of that orbit for its communication satellite services such as Starlink which was being built out



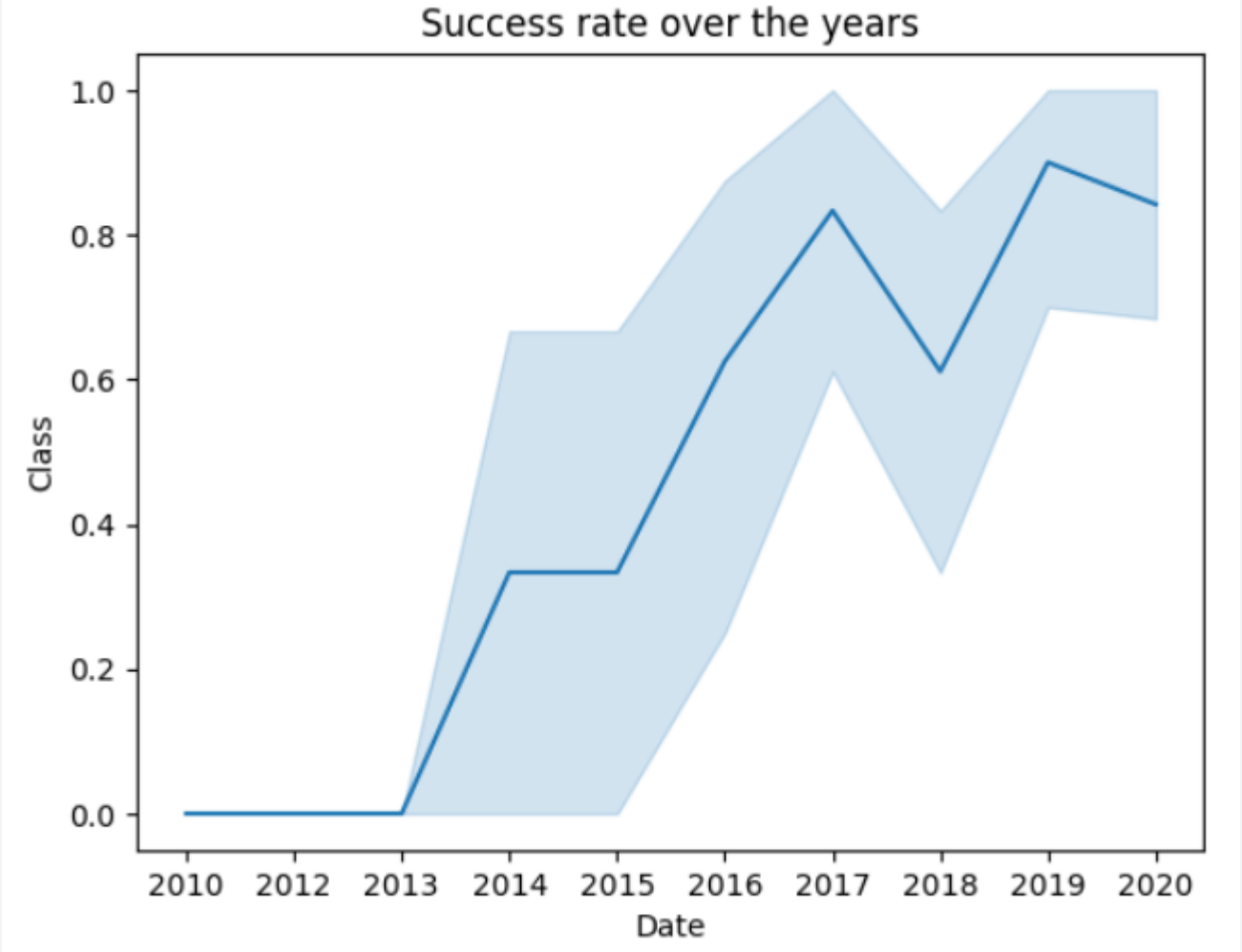
Payload vs. Orbit Type

- Geostationary Transfer Orbit(GTO) has a payload mass below 8000kg.
- The heaviest payloads are on the Very Low Earth Orbit(14000kg and above)



Launch Success Yearly Trend

- There was a steep yearly progression in success rates from 2013 to 2016 followed by a slight plateau between 2017 and 2020



All Launch Site Names

Launch Sites

Using the SQL query ***select distinct("Launch_Site") from SPACEXTABLE*** ***these were the results***

- **Cape Canaveral Space Force Station**, Florida: Specifically, SpaceX uses **Space Launch Complex 40** (**CCAFS SLC-40**) here. This site is a workhorse for Falcon 9 launches, handling a significant portion of their missions, including Starlink deployments and commercial satellite launches. It goes by the nickname “Slick Forty”
- **Kennedy Space Center**, Florida: Located adjacent to Cape Canaveral, SpaceX operates from the historic **Launch Complex 39A** (**KSC LC-39A**). This pad is notable for supporting crewed missions (like Crew Dragon flights to the ISS) and Falcon Heavy launches.
- **Vandenberg Space Force Base**, California: On the West Coast, SpaceX utilizes Space **Launch Complex 4 East** (**VAFB SLC-4E**). This site is crucial for launching payloads into polar and sun-synchronous orbits.

Launch Site Names Begin with 'CCA'

- Using the SQL query ***select * from SPACEXTABLE where "Launch_Site" like 'CCA%' limit 5***
- This filters on the Launch_Site column where “CCA%” indicates characters “CCA” followed by zero or more characters

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

- Using the SQL query *select sum(T."PAYLOAD_MASS__KG_") from (select * from SPACEXTABLE where Customer like "NASA%") as T*
- This creates a view with only data on NASA launches aliased T
- From this view, the sum of the payload is obtained using the sum function.

99980kg

- That is the equivalent of approximately 850,000 bananas.
- All launched by SpaceX rockets

Average Payload Mass by F9 v1.1

- Using the SQL query *select avg(T."PAYLOAD_MASS__KG_") from (select * from SPACEXTABLE where "Booster_Version" = "F9 v1.1") as T*
- This creates a view with only data on F9 v1.1 launches aliased T
- From this view, the average of the payload is obtained using the avg function.

2928.4kg

- That is the equivalent of approximately 24,816 bananas.
- Per launch on average by SpaceX

First Successful Ground Landing Date

- Using the SQL query ***select min(Date) from SPACEXTABLE where "Landing_Outcome" = "Success (ground pad)"***
- With this the earliest date where a ground pad landing was successful is obtained

22nd December, 2015

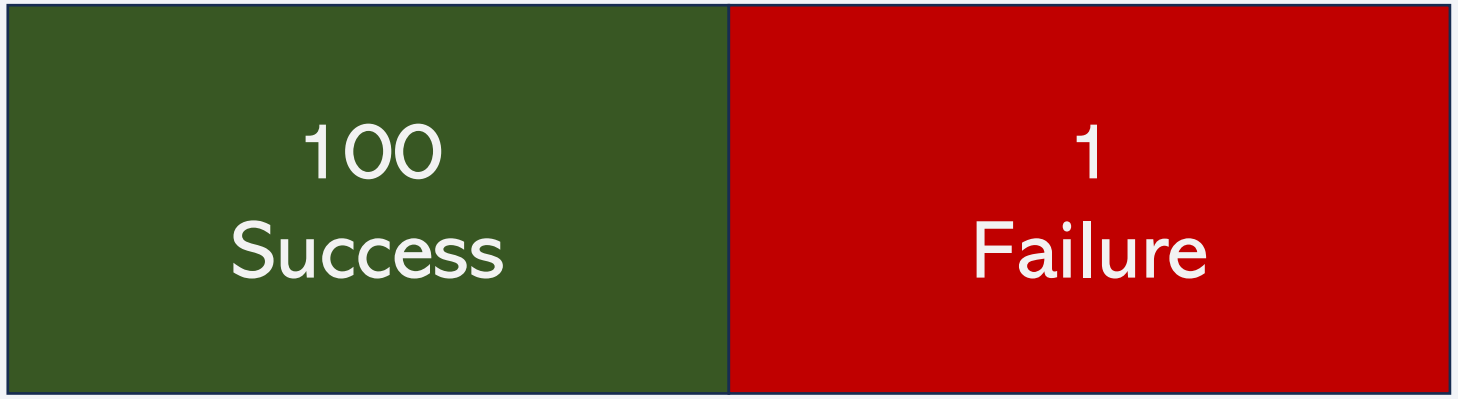
- An impactful date for the scientific community, especially in the aerospace discipline
- It has over 3.1 million views on youtube.

Successful Drone Ship Landing with Payload between 4000 and 6000

- Using the SQL query ***select "Booster_Version" from SPACEXTABLE where ("Landing_Outcome" = "Success (drone ship)") and ("PAYLOAD_MASS__KG_" > 4000 and "PAYLOAD_MASS__KG_" < 6000)***
- The following boosters landed payloads between 4000kg and 6000kg
  - F9 FT B1022
  - F9 FT B1026
  - F9 FT B1021.2
  - F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- There were 101 mission outcomes



- Almost all missions were considered some form of success

Boosters Carried Maximum Payload

- Using the SQL query ***select "Booster_Version" from SPACEXTABLE where "PAYLOAD_MASS__KG_" = (select max("PAYLOAD_MASS__KG_") from SPACEXTABLE)***
- The following boosters carried the 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

2015 Launch Records

- In 2015 there were some failed landings as highlighted below

MonthName	Landing_Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

Below are the landing outcomes in that period.

- Overall, there were some successes when landing were attempted, especially on drone ships

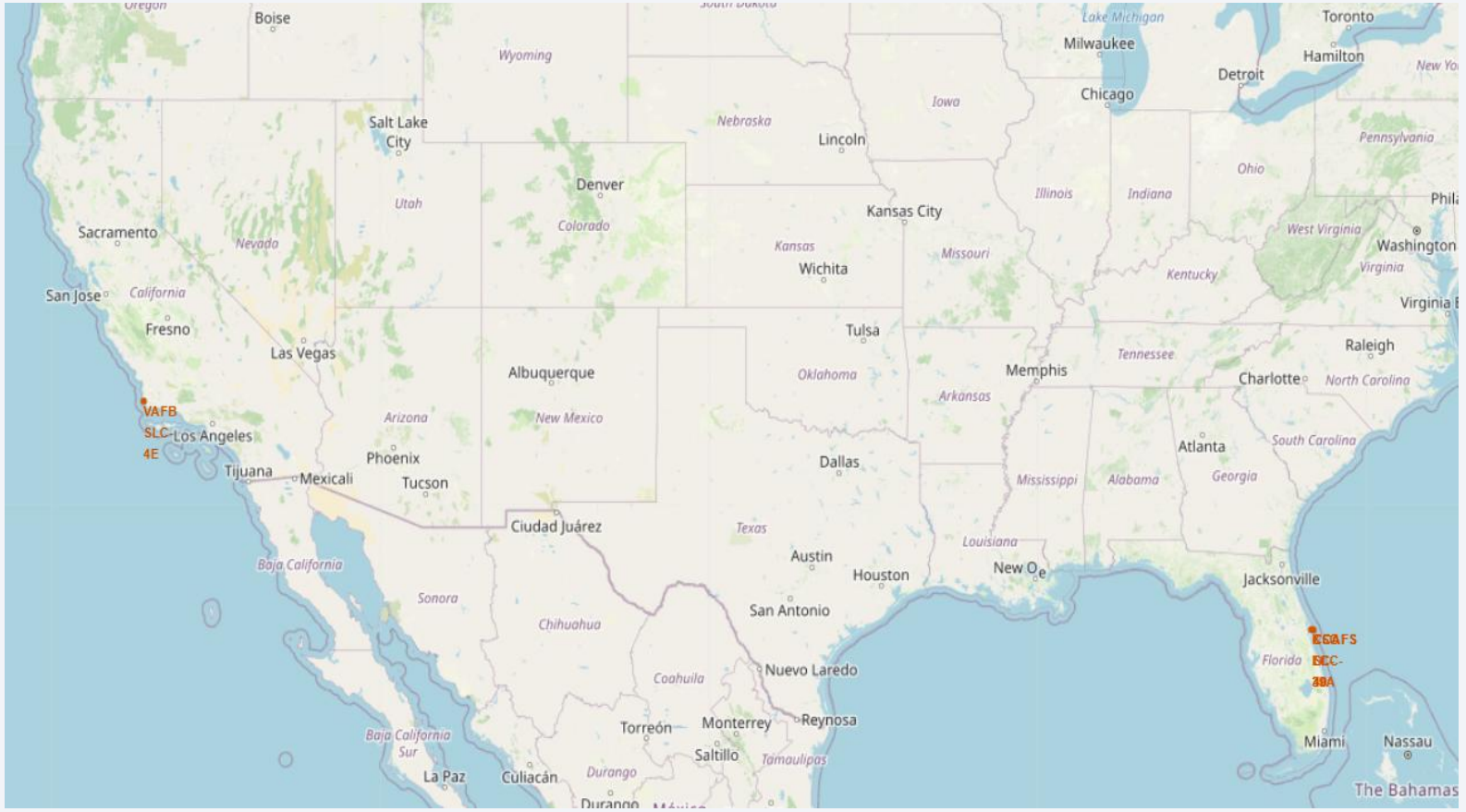
Landing_Outcome	Outcome_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

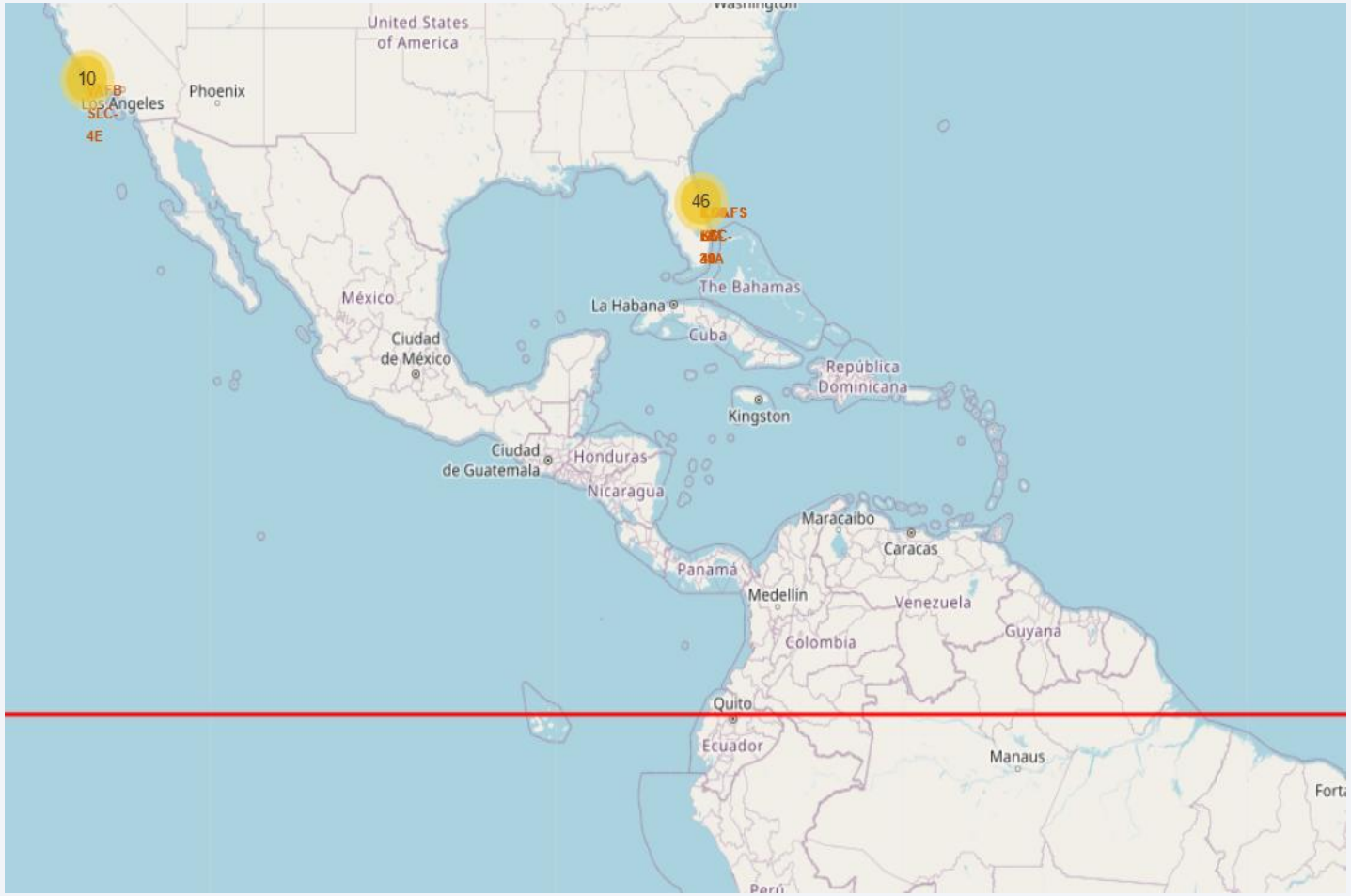
Launch Sites relative to the coastline

- On the West Coast close to Los Angeles sits the Vandenberg
- On the East Coast in Florida is the Kennedy space center and the Cape Canaveral Space Force station within miles of each other



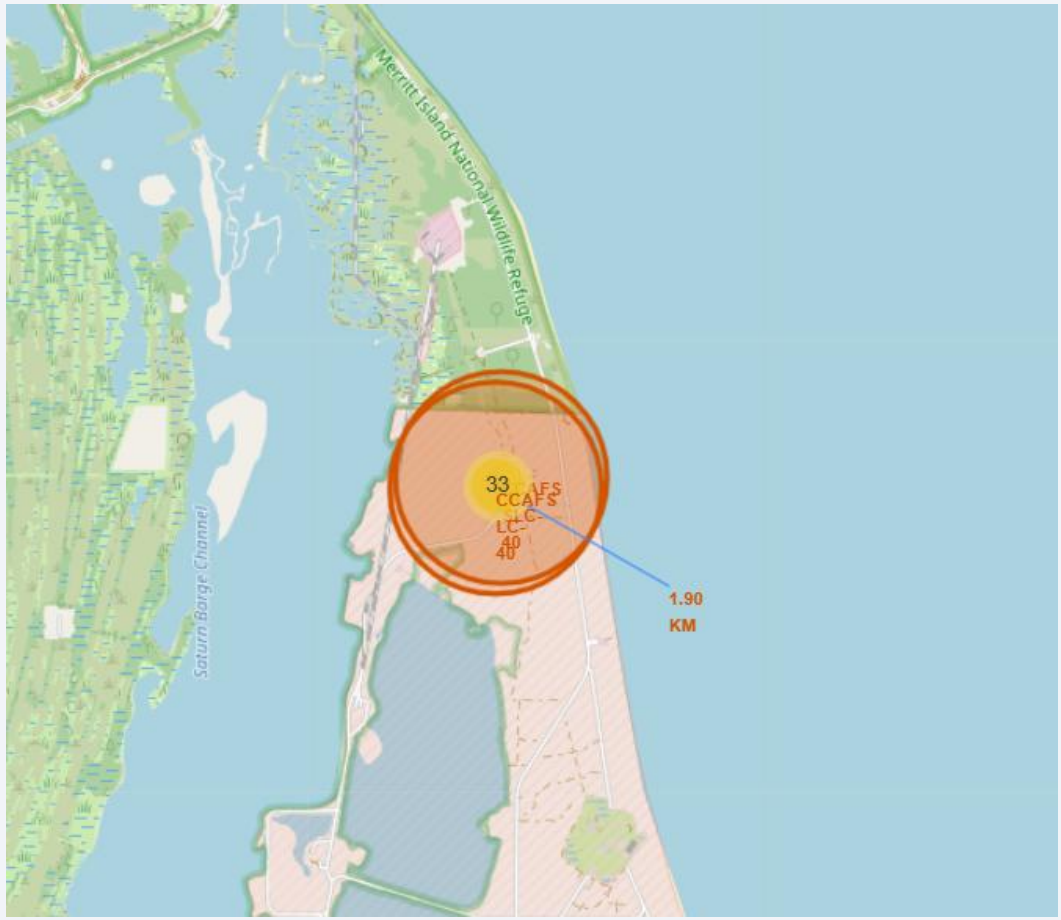
All launch sites are significantly distant from the equator

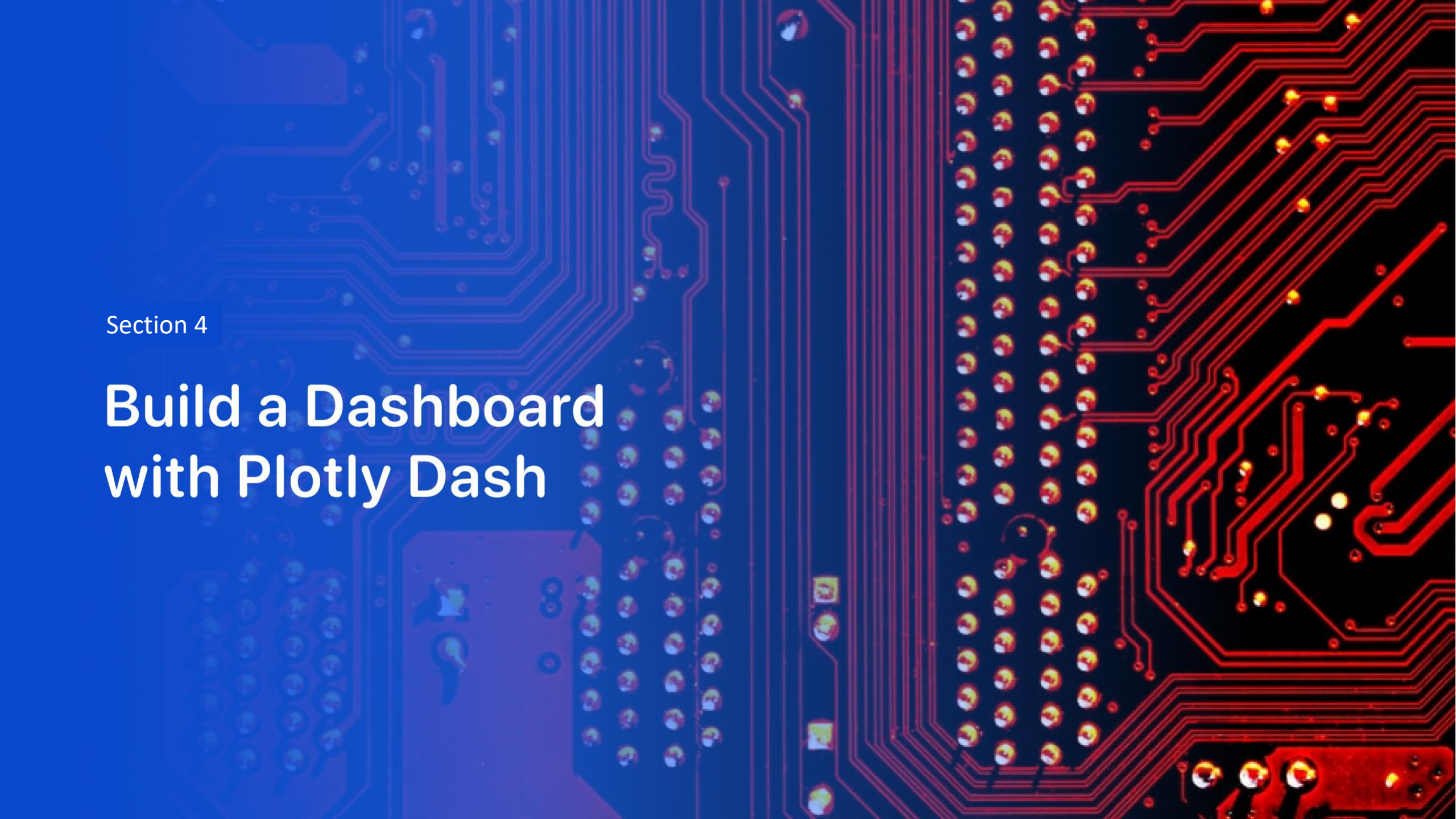
- As illustrated by the red line in the figure, all three launch sites are significantly far from the equator



Cape Canaveral Station is less than 2km to the Sea

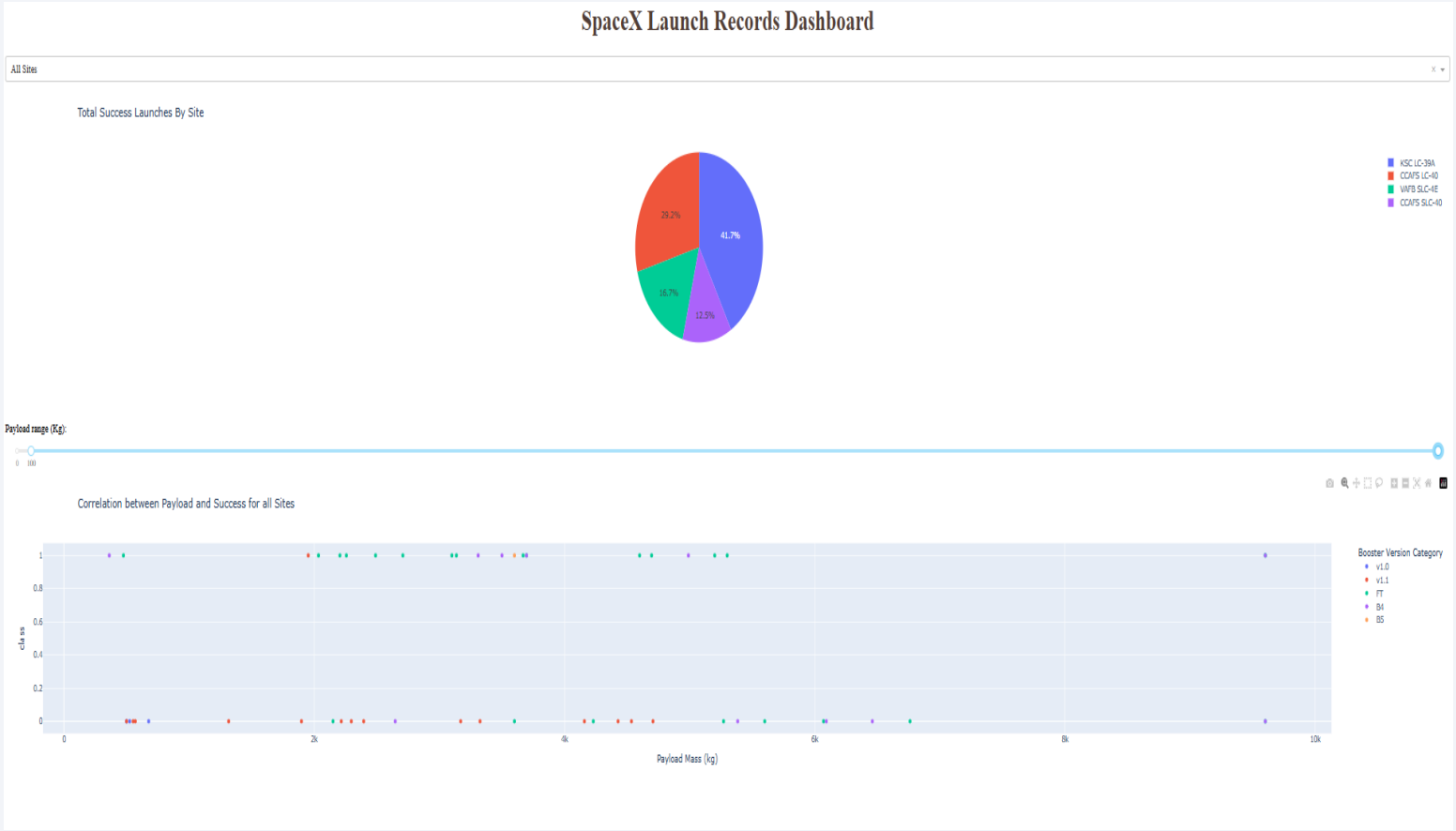
- Cape Canaveral Station lies very close to the sea with less than 2km as indicated by this distance line




Section 4
Build a Dashboard with Plotly Dash

Default Dashboard View

- There are 2 sections to the dashboard
- A success rate visualizer that uses a pie chart and is sorted by sites
- A scatter of the success state and the payload mass. It is colour-coded by booster version.

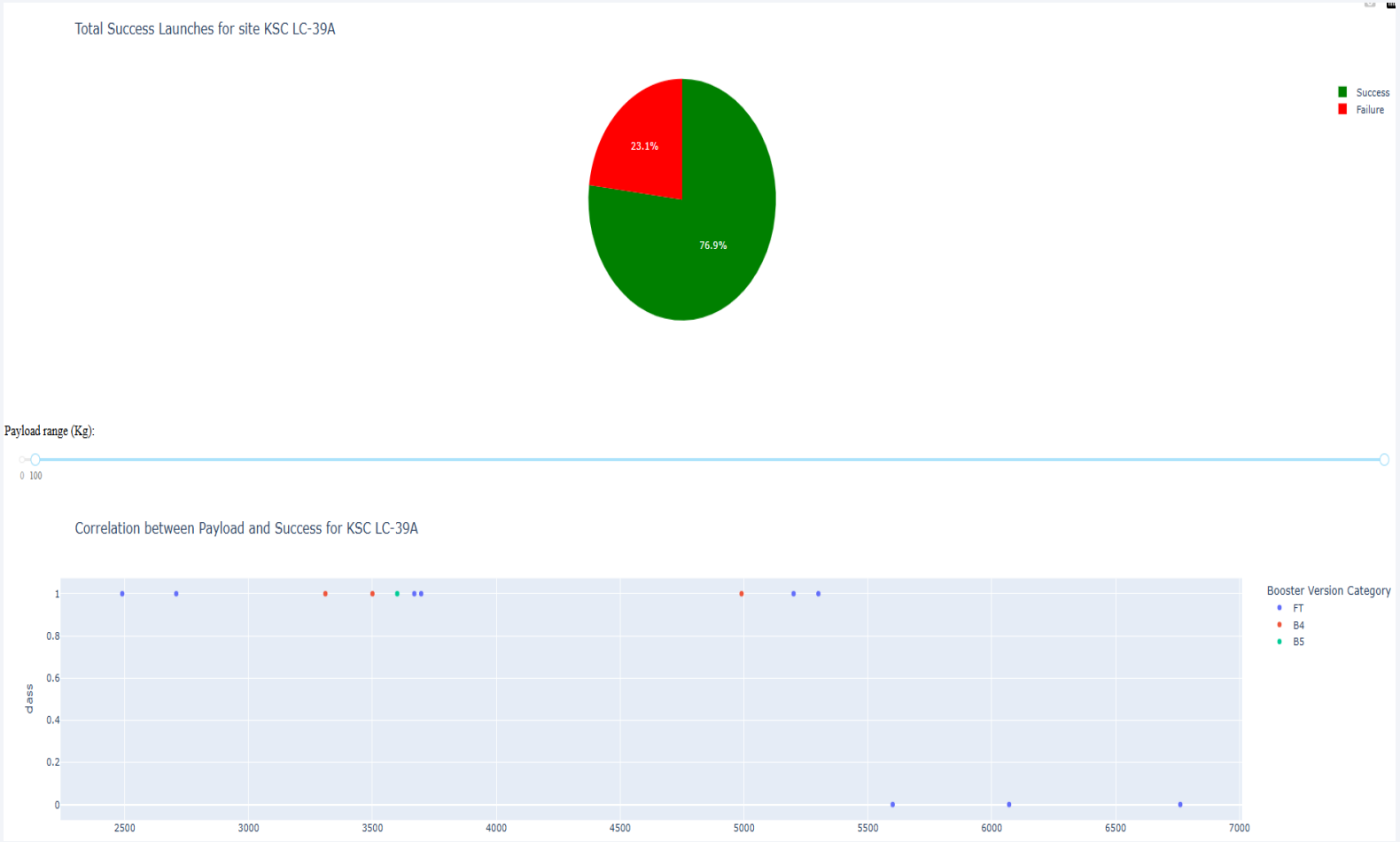


Insight on the Cape Canaveral launch site

- The site boasts a moderate 26% success rate.
- Irrespective of payload mass the full throttle booster(FT) remains a great performer

Insights on the Kennedy Launch Center

- The site boasts a 76% success rate, best of all sites.
- The Full throttle booster(FT) remains a great performer.
- There are more successes with payloads below 5500kgs

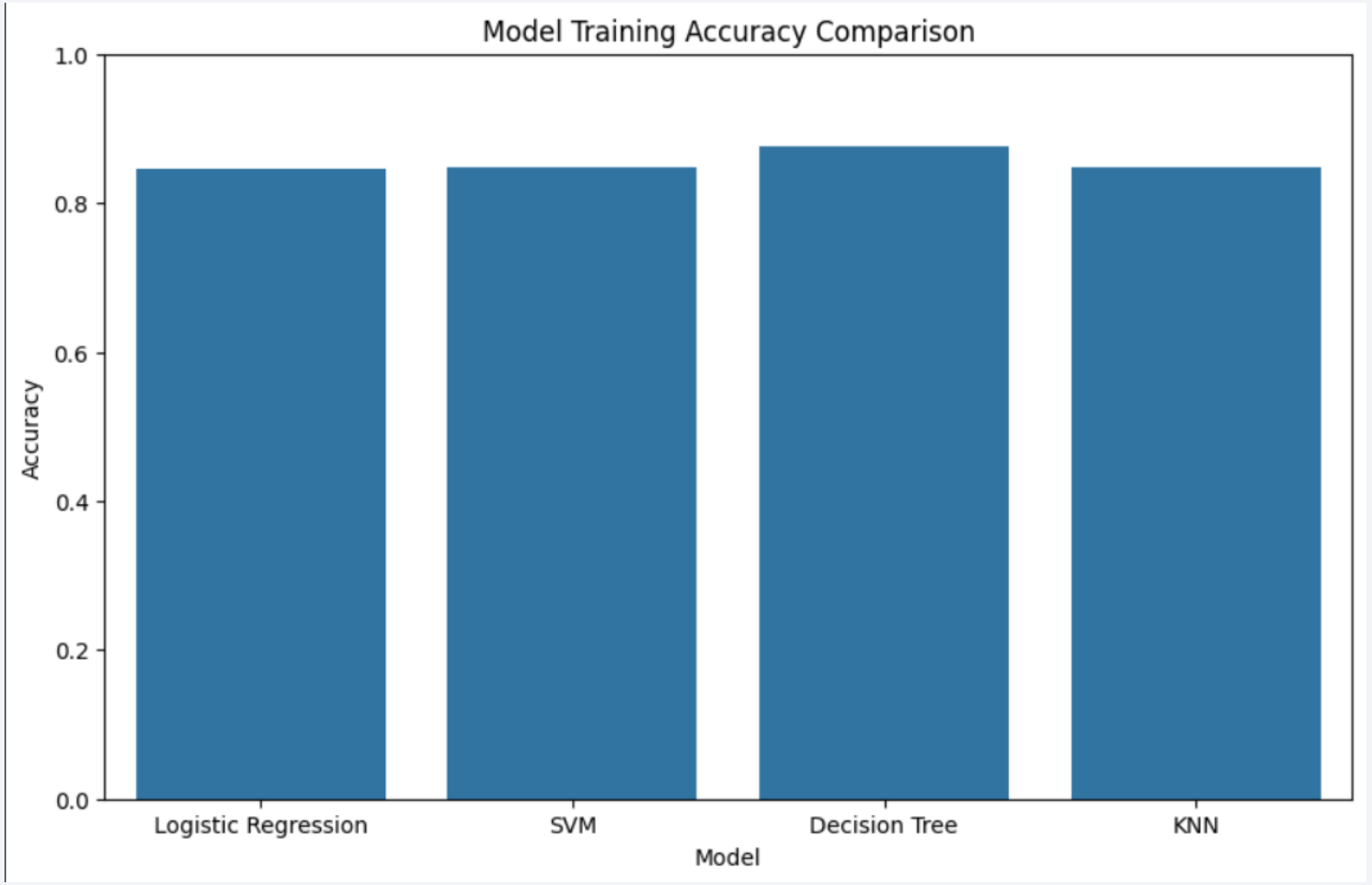


Section 5

Predictive Analysis (Classification)

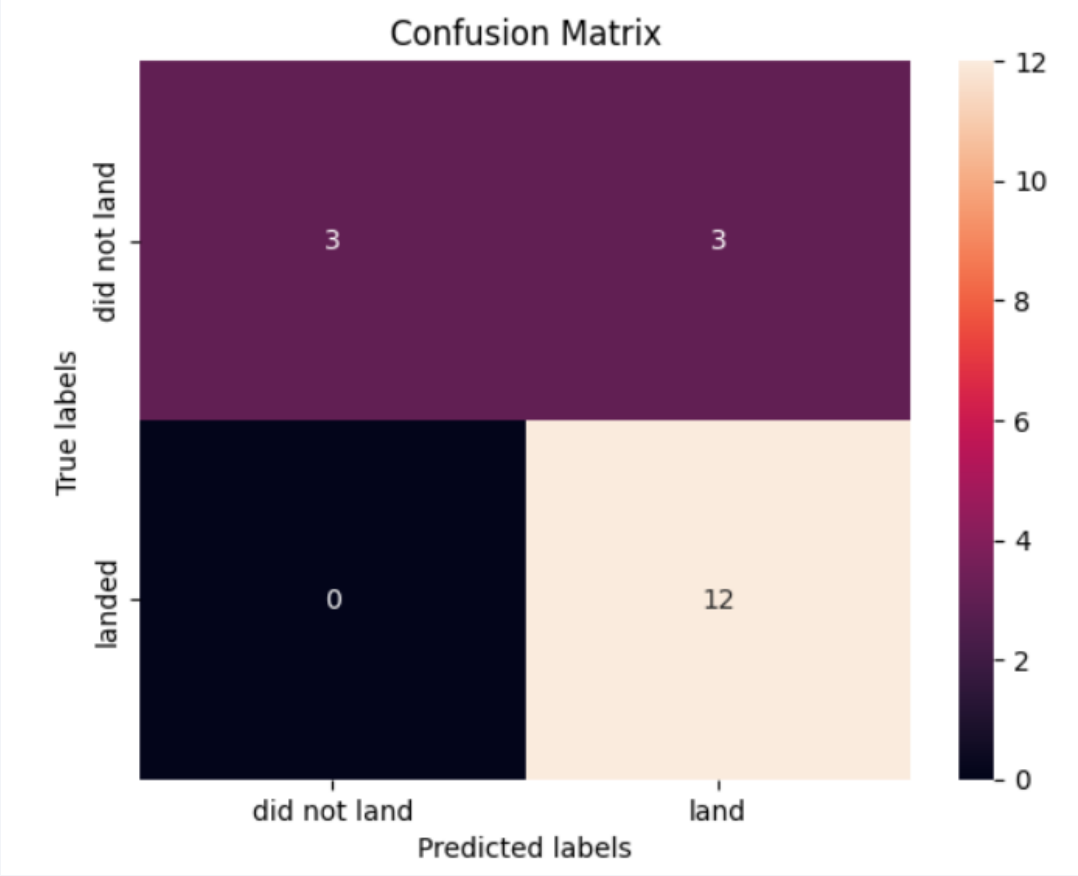
Classification Accuracy

- In training the Decision Tree Classifier provides the highest accuracy at 87%.
- However the SVM and KNN models had the better test accuracy at 83%.
- As such the best model to use would be the **SVM**



Confusion Matrix

- With 18 samples in the test dataset, it is too small a sample to confidently generalize.
- The **SVM** model correctly predicted 3 times where the was not a successful landing and 12 times when there was a successful landing
- However, there were three times that the landing was unsuccessful, but the model predicted a successful launch
- *This is an identical confusion matrix to the **KNN classifier** on the same test dataset. SVM is preferred since it will not need the entire training set for each new inference.*



Conclusions

- There was a steep yearly progression in success rates from 2013 to 2016 followed by a slight plateau between 2017 and 2020
- The **Full Throttle(FT)** booster version introduced in 2015 has the highest success rate across different payloads.
- The **KSC LC-39A** launch site has the highest success rate of all launch sites at **76%.**
- The **SVM classifier** performed best out of all the classification models tested with **83.33% test accuracy**. Due to the nature of the test dataset, there is more to be done to evaluate the model especially on real-world data.

Appendix

- All relevant assets are on github

Thank you!