ROCKET SCIENCE WITH DATA SCIENCE

Arturo León
2024-03-06

Outline

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

Executive Summary

- The data analysis was carried out through the following processes:
  - Gathering data through the means of web scraping and the SpaceX API;
  - Engaging in Exploratory Data Analysis (EDA), which encompassed the processes of data wrangling, creating data visualizations, and employing interactive visual analytics;
  - Utilizing Machine Learning for predictive analysis.
- In summarizing the outcomes:
  - We were able to acquire substantial data from publicly available sources;
  - EDA techniques helped in pinpointing the key features to forecast the success of rocket launches;
  - Predictive Machine Learning identified the optimal model to ascertain the critical attributes that can capitalize on this opportunity most effectively, based on the entirety of the data gathered.

Introduction

The aim is to assess whether the emerging company Space Y can be a contender against Space X in the aerospace industry.

Sought-after responses include:

- A superior method to gauge the aggregate expenditure for space missions, which involves forecasting the first stage landings of rockets successfully;
- Identifying the prime location for conducting space launches.

Section 1

Methodology

Methodology

In the executive summary:

The data collection process was two-pronged, utilizing:

- The Space X API available at their official API endpoint;
- Web scraping from the Wikipedia page listing Falcon 9 and Falcon Heavy launches.

Subsequent data manipulation involved:

- Augmenting the dataset by generating a label for the landing outcomes based on the summarized and analyzed outcome data.

For the exploratory data analysis (EDA):

- Techniques including visualization and SQL queries were employed to examine the data thoroughly.

Methodology

The executive briefing includes the following points:

- Interactive visual analytics were conducted using Folium for mapping and Plotly Dash for creating interactive dashboards.

- Predictive analytics were carried out with the use of classification models.

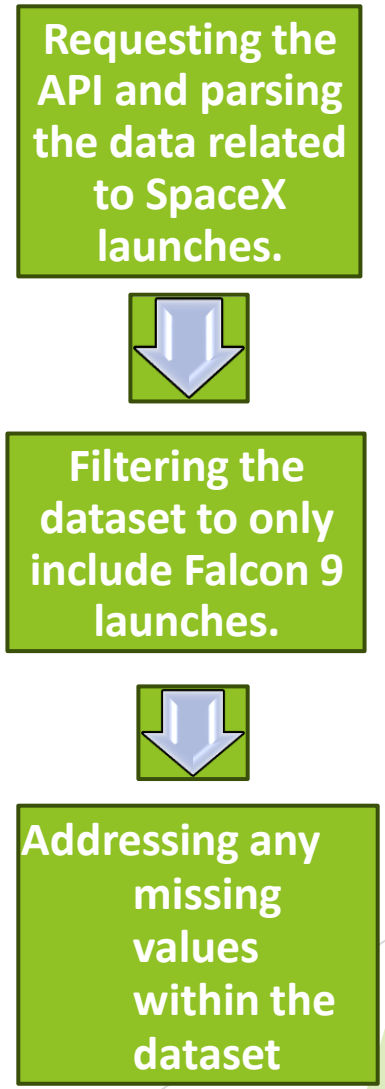
- The collected data was standardized and segmented into training and testing sets. Four distinct classification models were then applied to the data. The precision of each model was assessed using various parameter configurations.

Data Collection

- Data sets were collected from Space X API (https://api.spacexdata.com/v4/rockets/) and from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches), using web scraping technics.

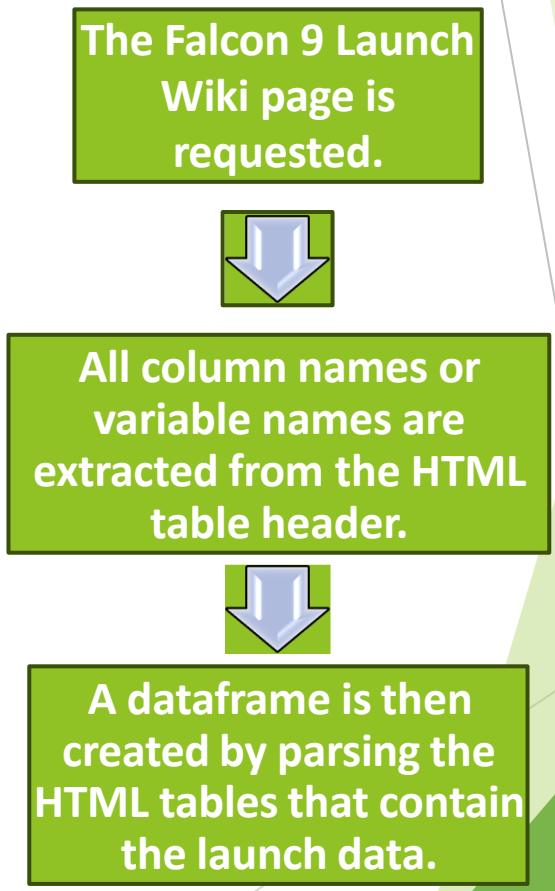
Data Collection – SpaceX API

- SpaceX provides access to a publicly available API for the retrieval and subsequent use of data.
- This API has been utilized as per an adjacent flowchart, following which the data has been stored persistently.



Data Collection - Scraping

- Launch data for SpaceX is also available via Wikipedia, which can be downloaded, processed according to a specified flowchart, and then stored.



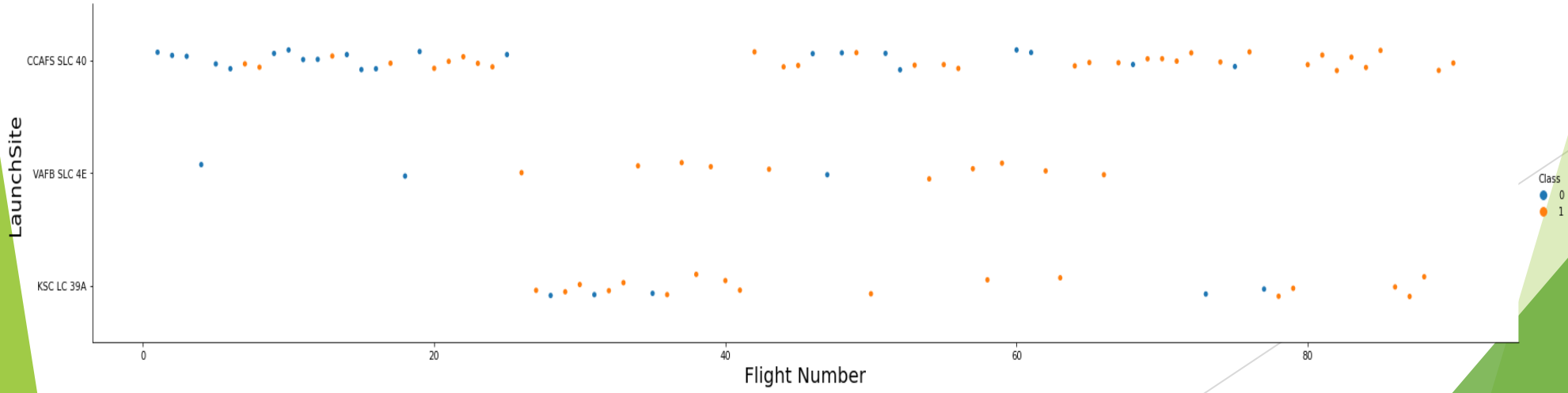
Data Wrangling

- The data analysis commenced with an Exploratory Data Analysis (EDA) on the collected information.
- Subsequently, comprehensive summaries were compiled detailing the number of launches at each site, as well as the frequency of various orbits and mission outcomes for each orbit classification.
- To conclude, a label describing the outcome of the landings was derived from the 'Outcome' column in the dataset.

EDA with Data Visualization

- Scatterplots and barplots were employed as visual tools to investigate the correlations between various feature pairs in the dataset. These comparisons included:

- The relationship between Payload Mass and Flight Number,
- The association of Launch Site with Flight Number,
- The correlation between Launch Site and Payload Mass,
- The interaction between Orbit type and Flight Number,
- And the relationship between the Payload characteristics and Orbit type.



EDA with SQL

- The executed SQL queries included:
  - Retrieval of the distinct names of launch sites used in space missions.
  - Identification of the top five launch sites with names starting with 'CCA'.
  - Calculation of the total payload mass lifted by NASA (CRS) mission boosters.
  - Computation of the average payload mass hoisted by the Falcon 9 version 1.1 booster.
  - Determination of the date on which the inaugural successful landing on a ground pad occurred.
  - Compilation of the list of boosters that successfully landed on a drone ship carrying a payload mass between 4000 and 6000 kg.
  - Tallying the total mission outcomes, categorized into successes and failures.
  - Identification of the booster versions that have transported the heaviest payloads.
  - Analysis of failed landing outcomes on drone ships, including specific booster versions and launch site names, within the year 2015.
  - Ranking of landing outcomes, such as failures on drone ships or successes on ground pads, within the timeframe from June 4, 2010, to March 20, 2017.

Build an Interactive Map with Folium

- Within Folium Maps, the following visual elements were utilized:
  - Markers were placed to denote specific points of interest, such as launch sites.
  - Circles were drawn to highlight significant areas situated at particular coordinates, for instance, the NASA Johnson Space Center.
  - Marker clusters were created to represent collections of events situated at the same coordinates, which could be launches occurring at a single launch site.
  - Lines were incorporated to graphically represent the distances between pairs of coordinates.

Build a Dashboard with Plotly Dash

For the purpose of data visualization, the following graphical representations were utilized:

- Graphs depicting the percentage of launches that occurred at each site.
- Plots showing the range of payloads.

Together, these graphical elements facilitated a swift examination of the correlation between the payloads and the launch sites. This, in turn, aided in determining the most optimal launch locations based on the varying payload requirements.

Predictive Analysis (Classification)

A comparative analysis was conducted on four different classification models:

- Logistic Regression
- Support Vector Machine (SVM)
- Decision Tree
- K Nearest Neighbors (KNN)

Data preparation and standardization

Test of each model with combinations of hyperparameters

Comparison of results

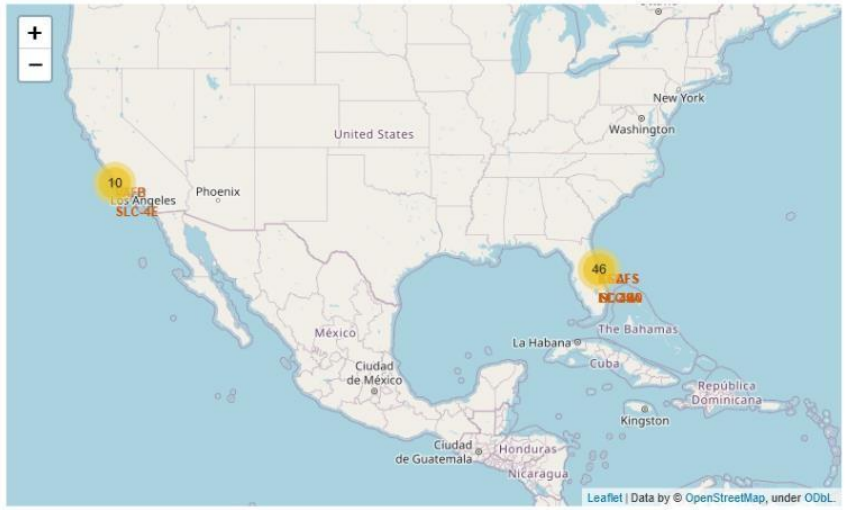
Results

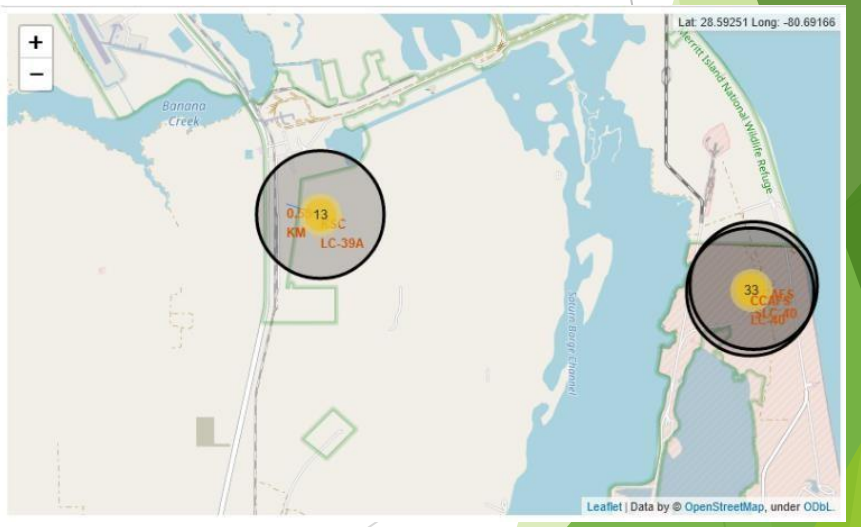
The exploratory data analysis yielded the following insights:

- SpaceX operates four distinct launch sites.
- Initial launches were executed by SpaceX and for NASA.
- The Falcon 9 v1.1 booster's average payload is 2,928 kg.
- The maiden successful landing occurred in 2015, five years after the first launch.
- Numerous Falcon 9 booster variants successfully landed on drone ships carrying payloads exceeding the average.
- Nearly all mission outcomes were successful.
- In 2015, two versions of the boosters, F9 v1.1 B1012 and F9 v1.1 B1015, failed to land successfully on drone ships.
- The success rate of landing outcomes improved over the years.

Results

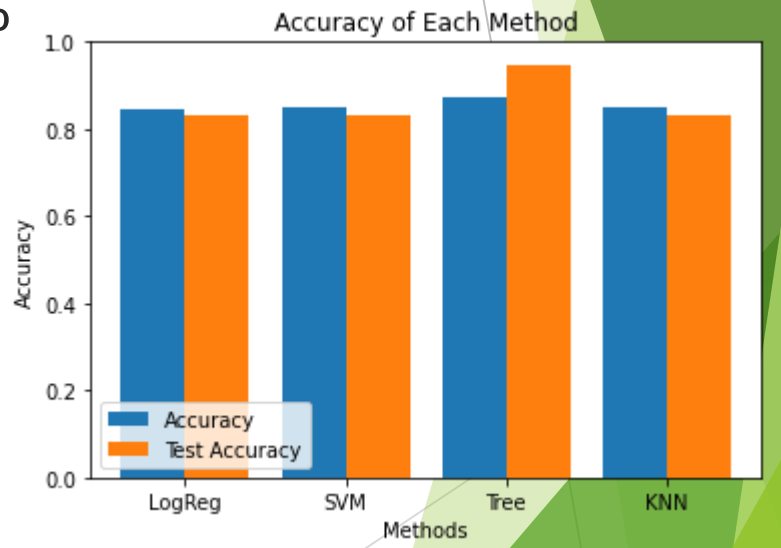
- Interactive analytics enabled the identification of launch sites typically located in safe areas, such as near the sea, and surrounded by robust logistical infrastructure. The majority of launches took place from sites on the east coast.





Results

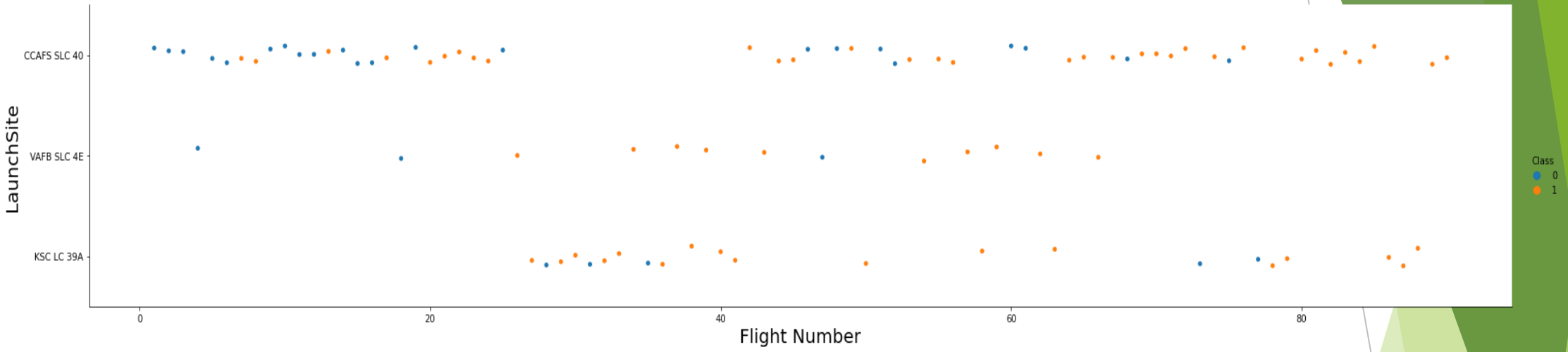
- From the predictive analysis, the Decision Tree Classifier emerged as the most accurate model for predicting successful landings, with an accuracy exceeding 87% for the training set and surpassing 94% for the test data.



Section 2

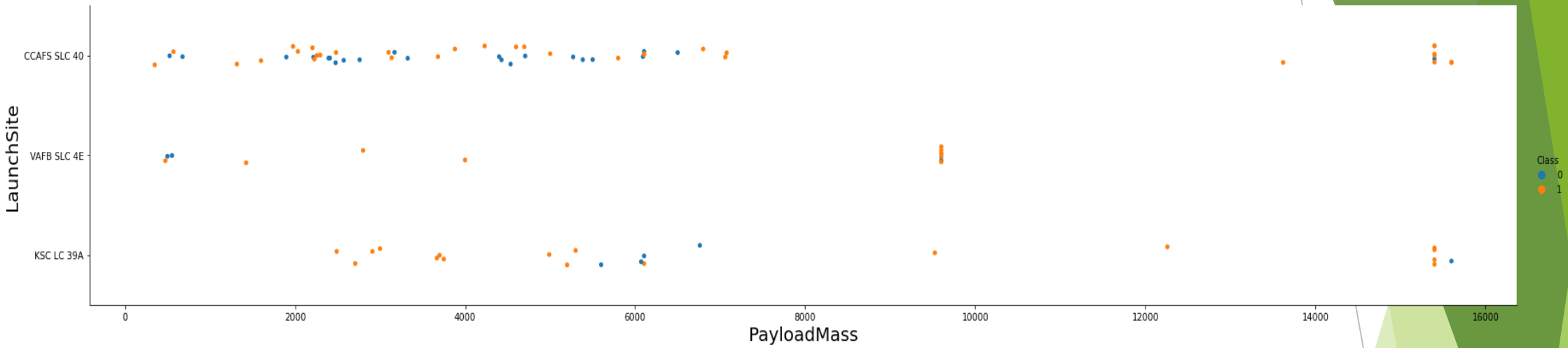
Insights drawn from EDA

Flight Number vs. Launch Site



- The graphical analysis indicates that currently, the most successful launch site is CCAFS SLC 40, demonstrating a high rate of successful recent launches. Following in effectiveness are VAFB SLC 4E and KSC LC 39A. Additionally, there is a visible trend showing that the overall success rate of launches has been on an upswing over time.

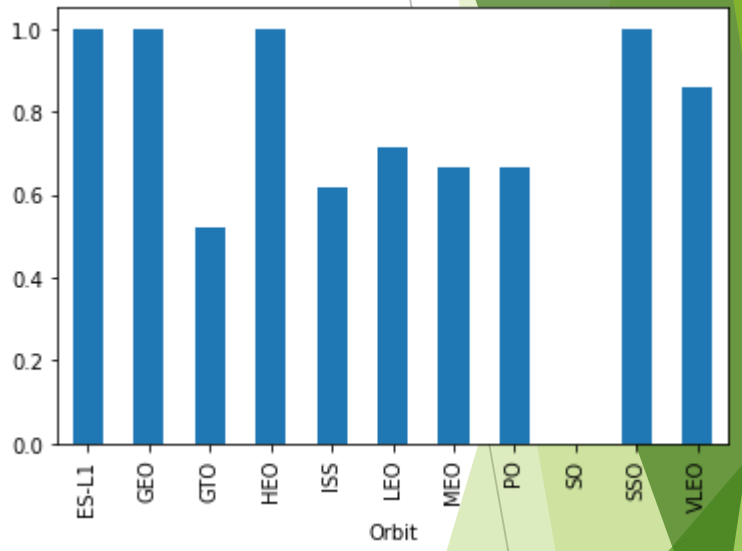
Payload vs. Launch Site



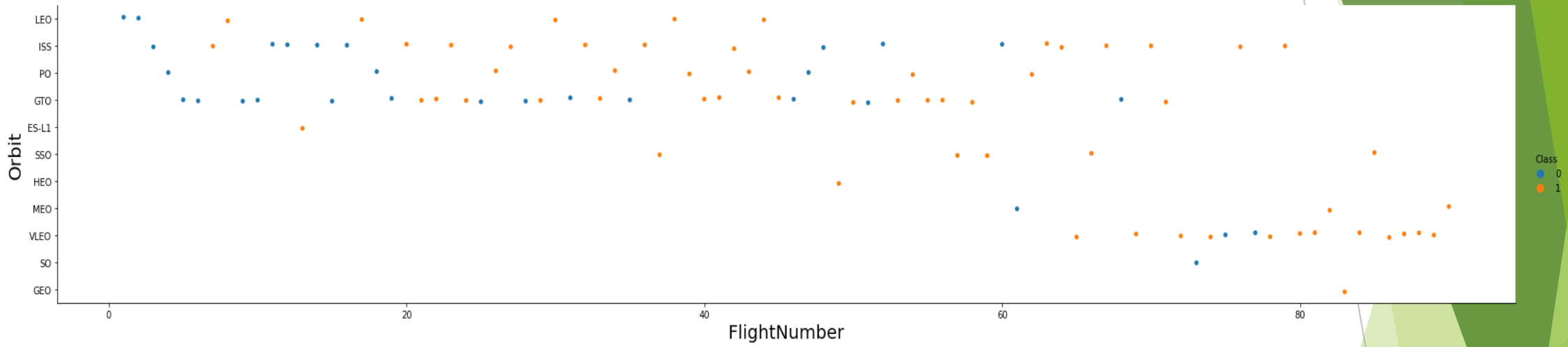
- In terms of payload capacity, rockets carrying payloads over 9,000 kg, which is roughly equivalent to the weight of a school bus, have an excellent success rate. However, the capability to send payloads exceeding 12,000 kg is apparently exclusive to launches from CCAFS SLC 40 and KSC LC 39A.

Success Rate vs. Orbit Type

- Orbital success rates are highest for missions targeting the ES-L1, GEO, HEO, and SSO orbits. Other orbits with notable success rates include VLEO, with over 80% success, and LFO, with success rates above 70%.



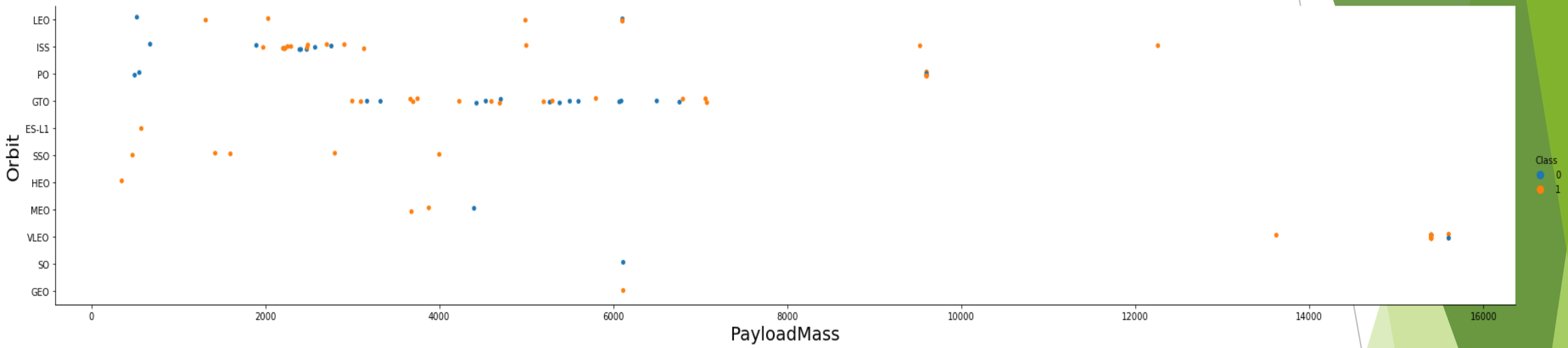
Flight Number vs. Orbit Type



The data analysis indicates the following:

- There has been an overall improvement in success rates across all orbits over time, suggesting advancements in launch reliability and technology.
- The VLEO (Very Low Earth Orbit) has emerged as a new area of business opportunity, likely due to its increasing launch frequency.

Payload vs. Orbit Type



- There seems to be no direct correlation between the payload mass and the success rate for GTO (Geostationary Transfer Orbit) missions.
- The ISS (International Space Station) orbit demonstrates the ability to accommodate the widest range of payload masses while maintaining a good success rate.
- There have been a limited number of launches to SO (Suborbital) and GEO (Geostationary Orbit) orbits, which could suggest either a lesser demand or a more specialized use case for these orbits.

Launch Success Yearly Trend

- The increase in success rates started around 2013 and continued consistently until 2020.
- It appears that the initial years of this period were marked by significant adjustments and improvements in spaceflight technology, contributing to the later success.

All Launch Site Names

- According to data, there are four launch sites:

Launch Site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

The extraction of specific data points was conducted through the following processes:

- Unique launch site identifiers were determined by selecting distinct "launch_site" values from the dataset..

Launch Site Names Begin with 'CCA'

- 5 records where launch sites 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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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 attemp

Total Payload Mass

- Regarding the total payload carried by NASA boosters:

Total Payload (kg)
111.268

- The total payload was calculated by summing all payloads associated with codes containing 'CRS', indicative of NASA's Commercial Resupply Services.

Average Payload Mass by F9 v1.1

- In relation to the average payload mass for the Falcon 9 v1.1 booster:

Avg Payload (kg)
2.928

- The dataset was filtered for the specific booster version mentioned. Subsequently, the average payload mass was computed, resulting in a value of 2,928 kg.

First Successful Ground Landing Date

- First successful landing outcome on ground pad:

Min Date
2015-12-22

- By applying a filter to the dataset for successful landings on a ground pad and selecting the earliest date from the resulting data, the initial successful ground landing was pinpointed to December 22, 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

- When identifying boosters that had achieved successful landings on a drone ship with a payload mass greater than 4,000 kg but less than 6,000 kg, the filter criteria produced four distinct booster versions that met these conditions.

Booster Version
F9 FT B1021.2
F9 FT B1031.2
F9 FT B1022
F9 FT B1026

Total Number of Successful and Failure Mission Outcomes

- Number of successful and failure mission outcomes:

Mission Outcome	Occurrences
Success	99
Success (payload status unclear)	1
Failure (in flight)	1

- The compilation and quantification of mission outcomes, categorized into their respective groups, produced the summarized data previously noted.

Boosters Carried Maximum Payload

- Regarding the boosters with the maximum payload mass, a specific selection was made from the dataset to identify those that have carried the heaviest payloads.

Booster Version (...)
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3

Booster Version
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 Launch Records

- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

Booster Version	Launch Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

- The list above has the only two occurrences.

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

- Ranking of all landing outcomes between the date 2010-06-04 and 2017-03-20:

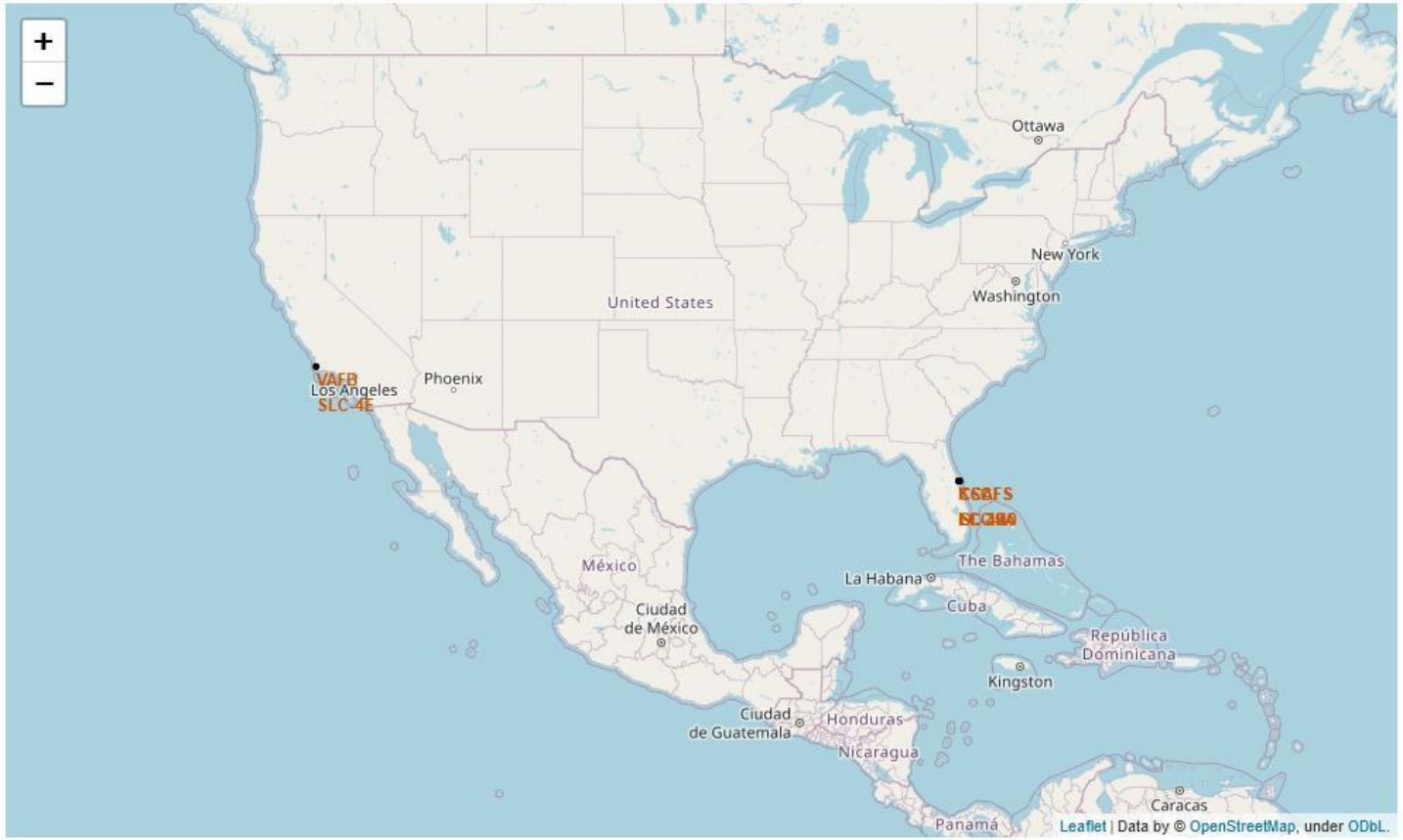
Landing Outcome	Occurrences
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

- This view of data alerts us that “No attempt” must be taken in account.

Section 4

Launch Sites Proximities Analysis

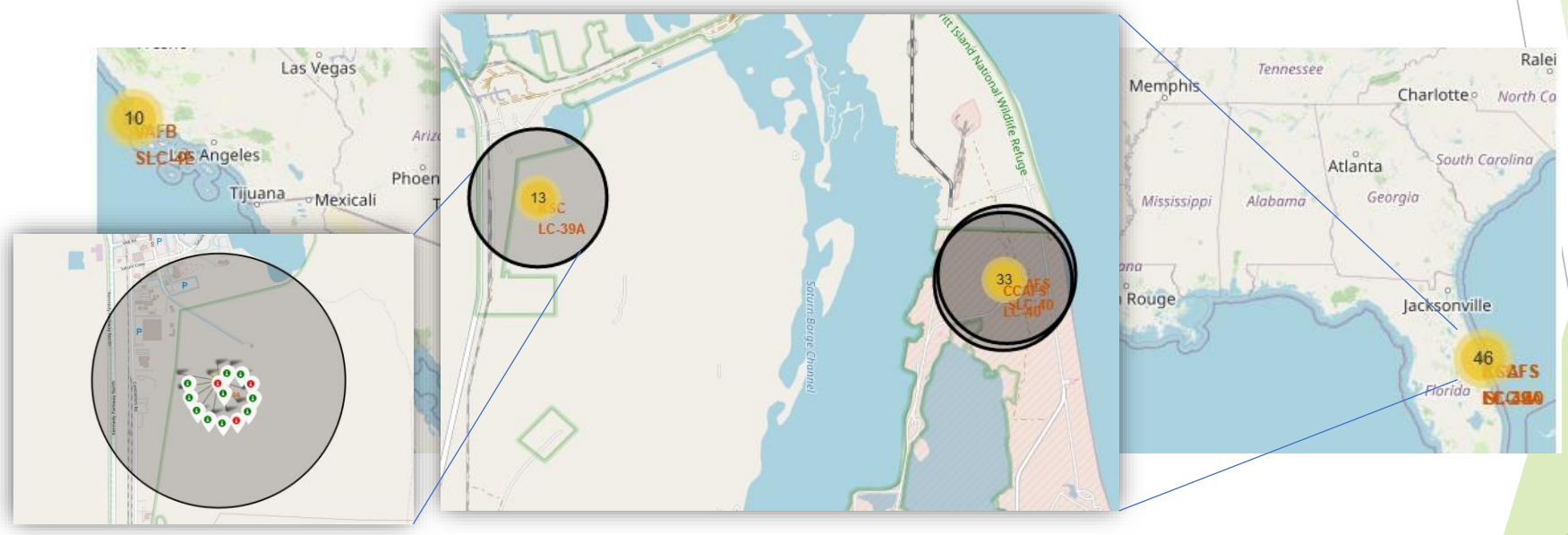
All launch sites



- Launch sites are near sea, probably by safety, but not too far from roads and railroads.

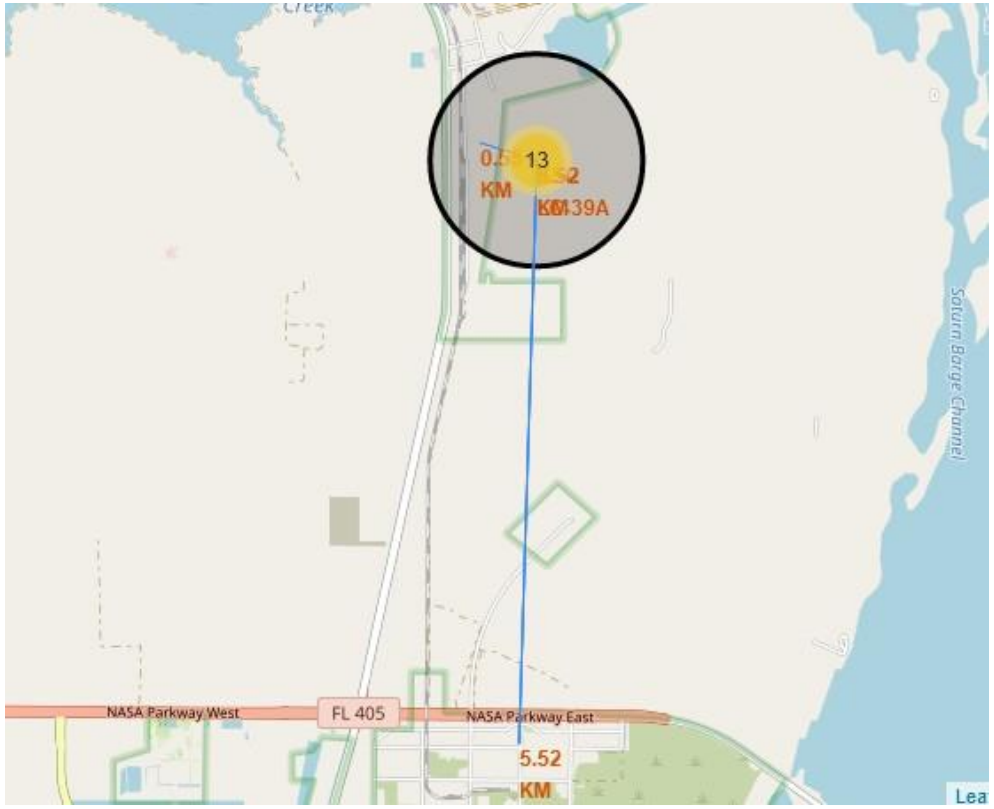
Launch Outcomes by Site

- Example of KSC LC-39A launch site launch outcomes

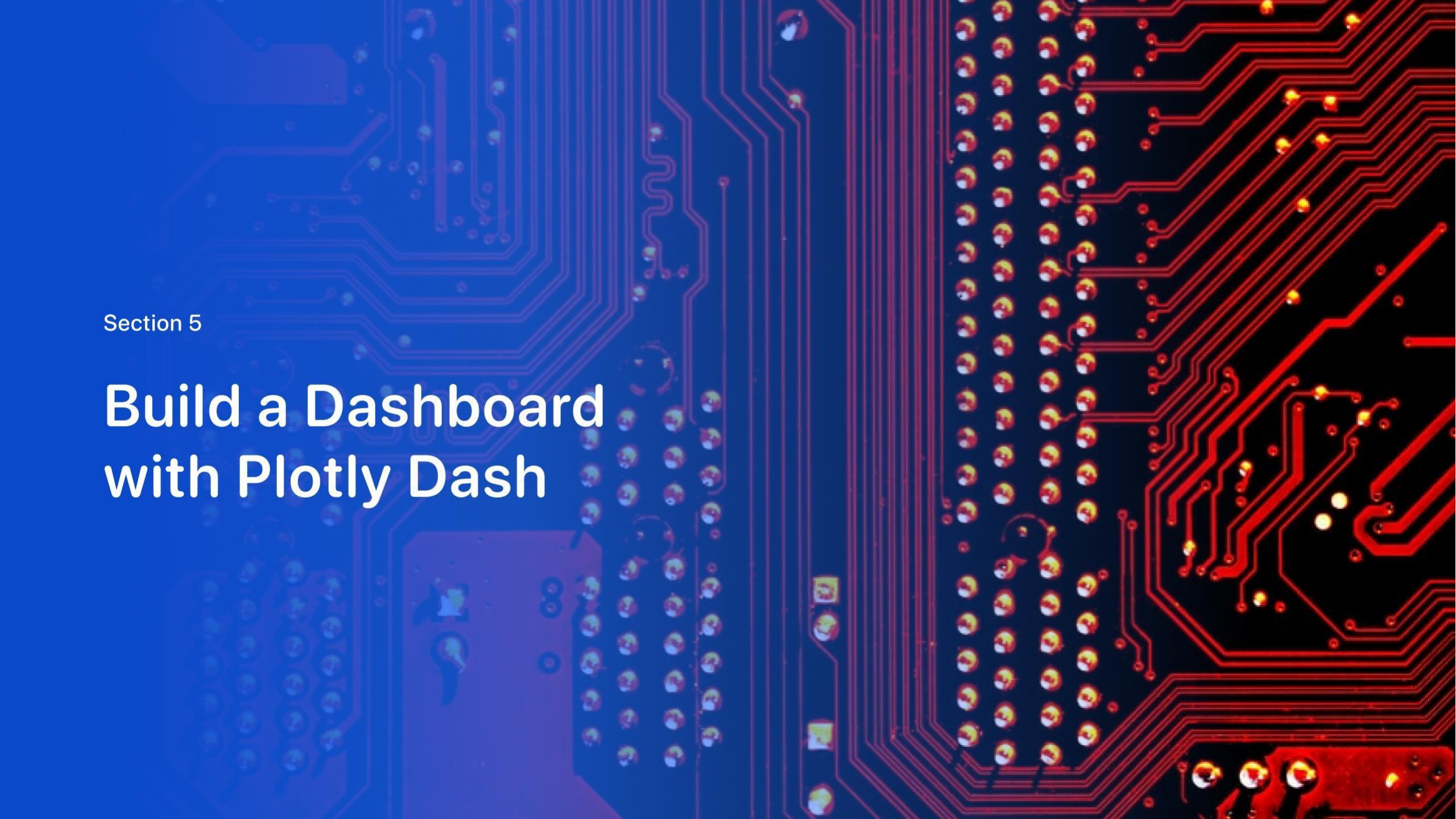


- Green markers indicate successful and red ones indicate failure.

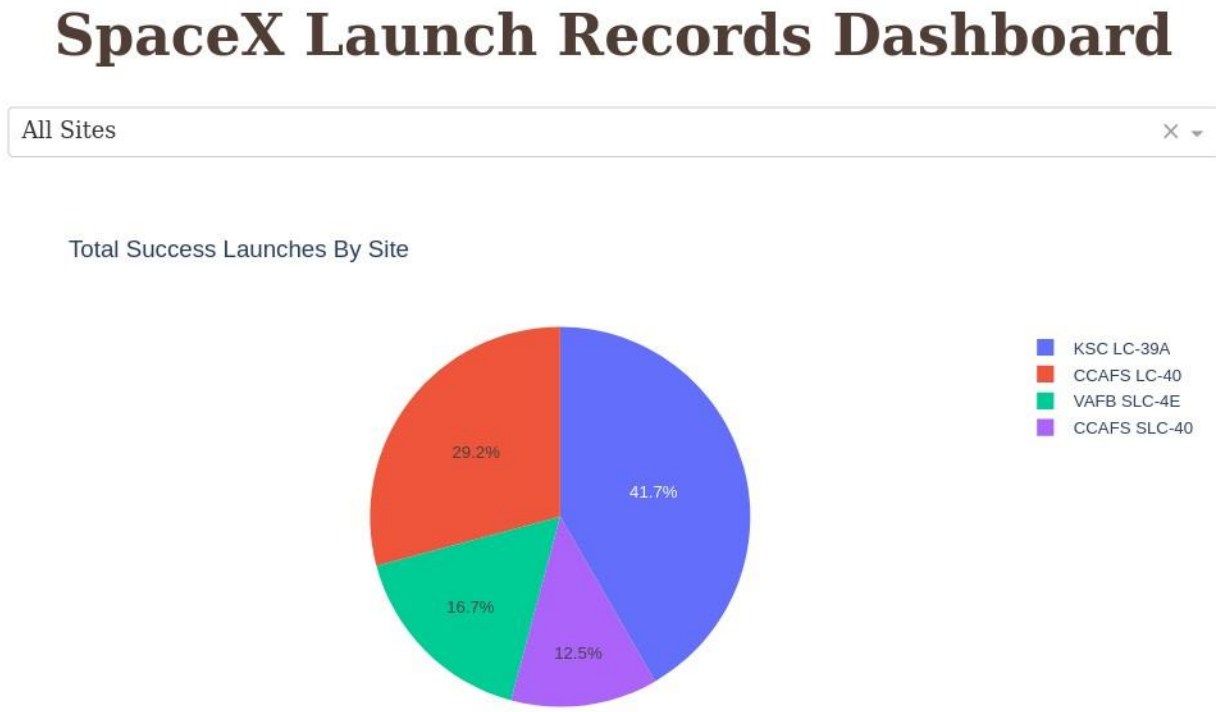
Logistics and Safety



- The selection of launch sites appears to be a strategic balance between safety requirements and logistical convenience. Proximity to the sea is likely for safety considerations, offering a clear area where rocket stages can fall without risk to populated areas.

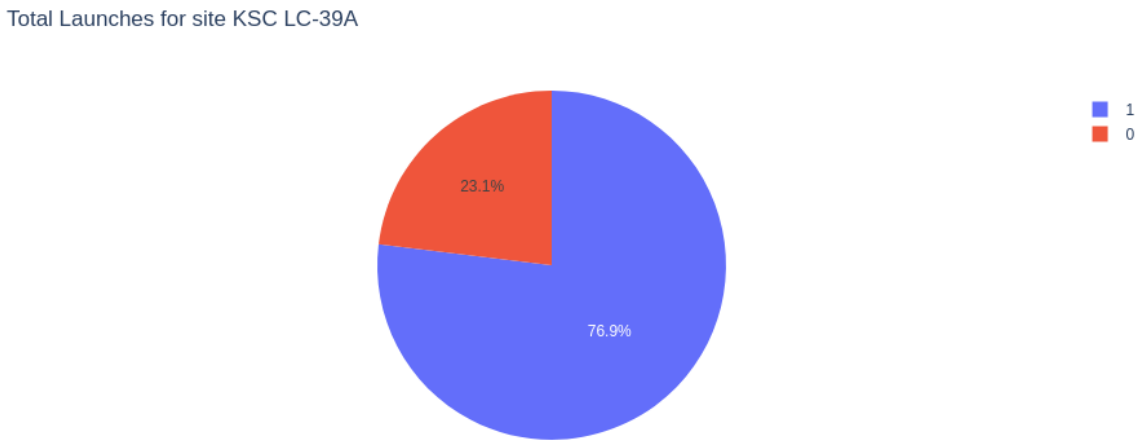

Section 5
Build a Dashboard with Plotly Dash

Successful Launches by Site



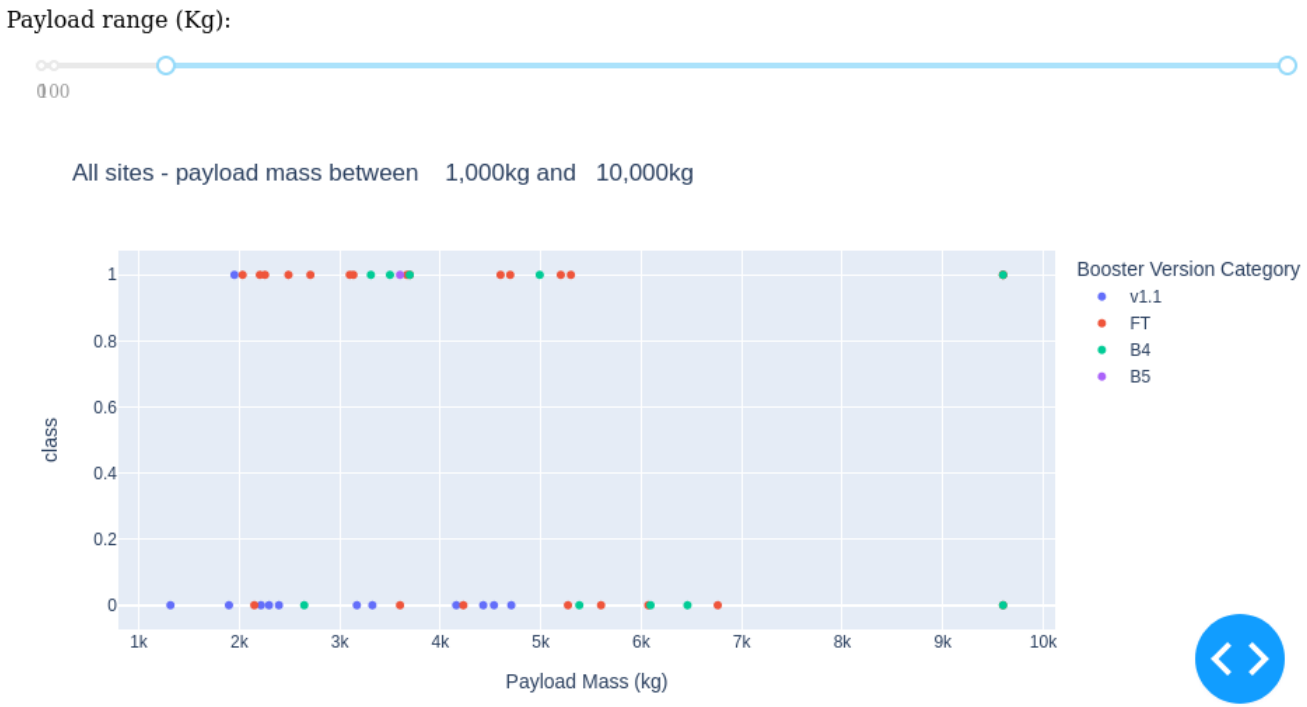
- The place from where launches are done seems to be a very important factor of success of missions.

Launch Success Ratio for KSC LC-39A



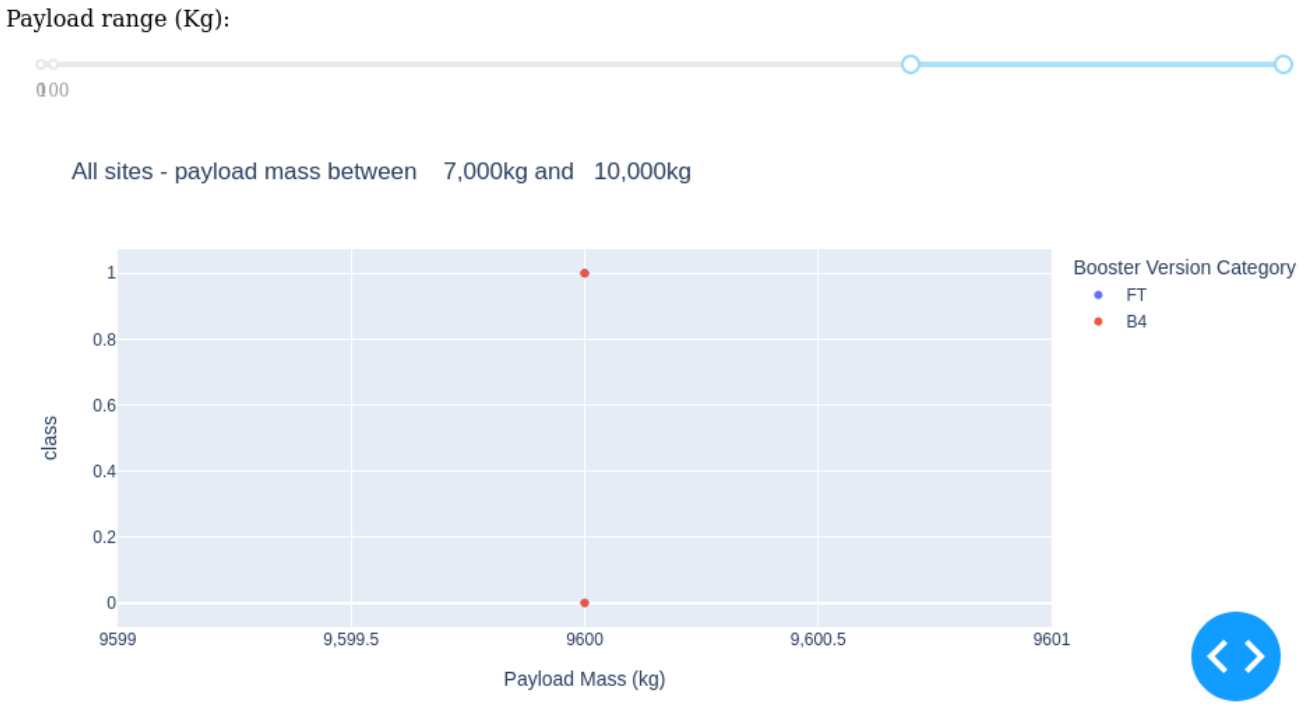
- 76.9% of launches are successful in this site.

Payload vs. Launch Outcome



- . Lastly, payloads weighing less than 6,000 kilograms combined with FT (Falcon 9 Full Thrust) boosters have emerged as the most successful pairing, suggesting this specific booster is well-suited for payloads of this mass range.

Payload vs. Launch Outcome



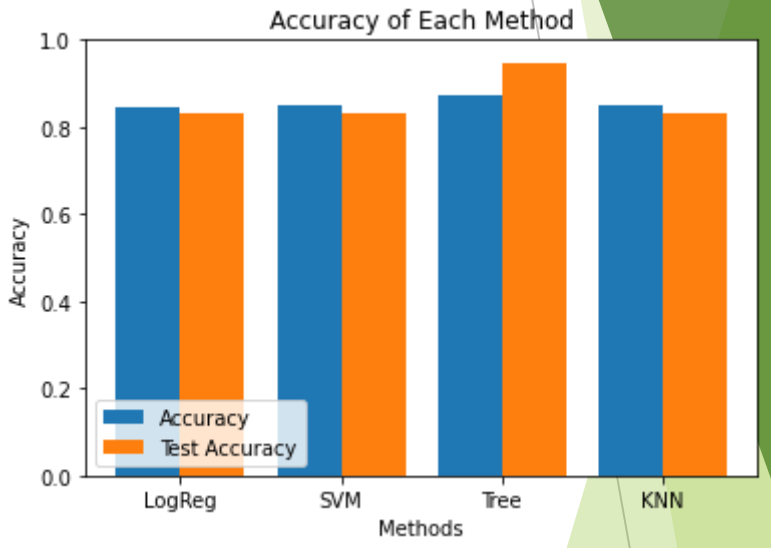
- There's not data to estimate the risk of launches over 7,000kg

Section 6

Predictive Analysis (Classification)

Classification Accuracy

- A comparative analysis of four distinct classification models was undertaken, with their accuracy rates depicted graphically alongside each model.
- The Decision Tree Classifier emerged as the most accurate model, demonstrating an impressive accuracy rate of over 87%.



Confusion Matrix of Decision Tree Classifier

- The effectiveness of the Decision Tree Classifier is further confirmed by the confusion matrix, which shows a predominance of correct predictions, both true positives and true negatives, over incorrect ones.

Conclusions

- The investigation incorporated multiple data sources, leading to refined conclusions. The KSC LC-39A site stood out as the optimal location for launch activities. It was noted that payloads heavier than 7,000kg tend to be less risky. Consistently successful mission outcomes are seen, likely due to advancements in processes and rocket technology. Employing the Decision Tree Classifier could aid in forecasting successful landings, potentially boosting profitability.


Thank you!
IPOSPACE