IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Hakam Yassine
16/03/2025

Outline

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

Executive Summary

Methodologies

- Data Collection
- Data Wrangling
- Exploratory Data Analysis
- Model Building
- Interactive visualization

Executive Summary

Results

- Model Performance
- Key Insights
- Interactive Visualizations:

Introduction

Project Background and Context:

- **Commercial Space Age**: The industry is evolving, with companies like **Virgin Galactic**, **Rocket Lab**, and **Blue Origin** making space travel more accessible.
- **SpaceX's Leadership**: SpaceX has set a benchmark with its achievements, including cost-effective launches and reusable rocket technology, specifically the **Falcon 9**.

Problems to Find Answers:

- **Launch Cost Determination**: How can we accurately determine the price of each launch for the competing startup, **Space Y**?
- **First Stage Landing Prediction**: What factors influence whether the Falcon 9's first stage will land successfully, and how can we predict this using machine learning?
- **Data Utilization**: How can we effectively gather and analyze public data related to SpaceX to inform our predictions and strategies?

Section 1

Methodology

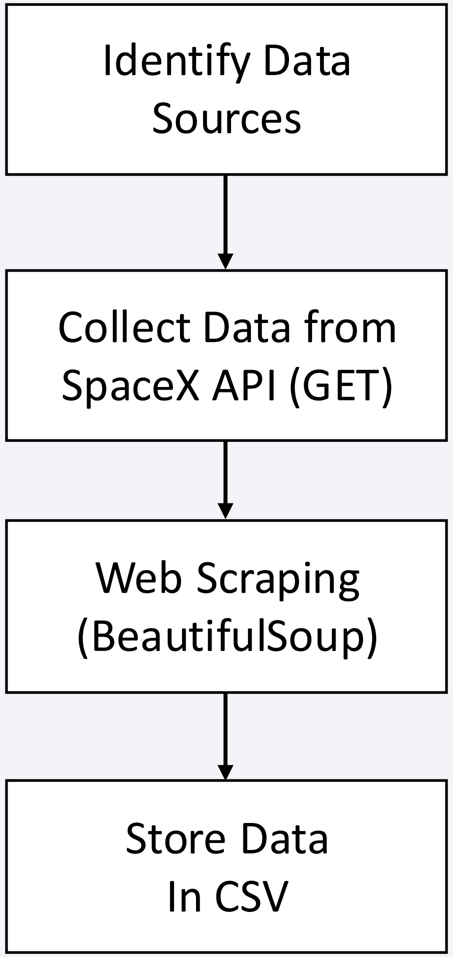
Methodology

Executive Summary

- Data Collection :
    - SpaceX API (GET Requests) and Web Scraping (BeautifulSoup)
- Data Wrangling :
    - Data cleaning, missing values handling, feature engineering (launch sites, payload types)
- Exploratory Data Analysis:
    - SQL Analysis and Visualization
- Model Building:
    - Classification Model (Decision Trees, Random Forest)
    - Evaluation (Accuracy, precision, recall, F1-score)
- Interactive visualization
    - Folium and Plotly Dash

Data Collection

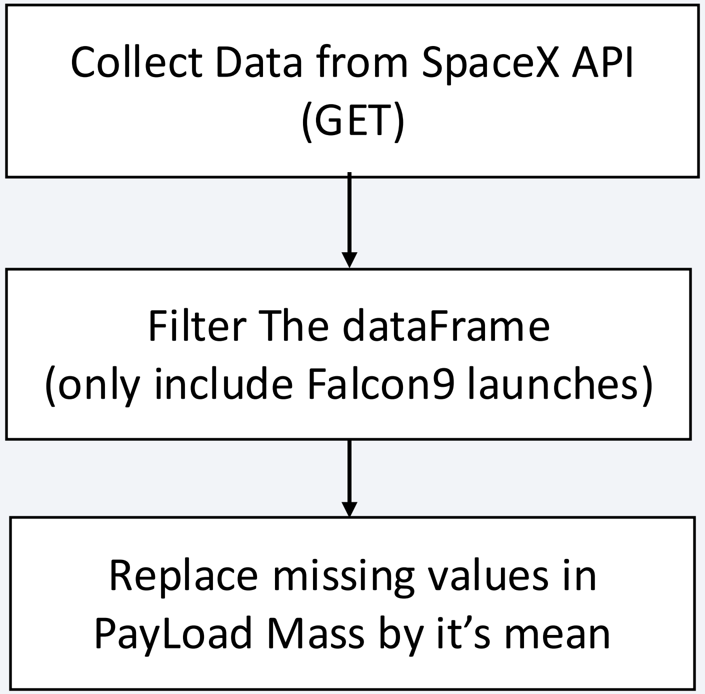
- Data was collected by a get request to the SpaceX API
  - I Retrieved data for SpaceX launches, including launch success or failure.
- Web Scraping with BeautifulSoup
  - To extract additional information from SpaceX's official website or other relevant sources
- Then Store Data in CSV Format for processing
  - Simple and easy to work with for tabular data.
  - Compatible with most data analysis libraries like pandas in Python.
  - Easily exportable for sharing and further analysis.



Data Collection – SpaceX API

- **Objective**: Gather data related to SpaceX launches, including launch success, failure, dates, and other important metrics.

- https://github.com/yhakam/IBM-Final-Project/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

Objectives:

- Extract specific details about SpaceX launches
- Websites such as SpaceX's official site or third-party sources (e.g., mission updates, launches, landing sites).
- Sending HTTP Requests:

Use requests library to send a GET request to a website.

- Fetch the HTML content of the page to be parsed.
- Parse HTML with BeautifulSoup:

Use BeautifulSoup to parse the HTML content.

- Navigate through the parsed HTML tree to find specific data points (e.g., launch success, mission details, etc.).

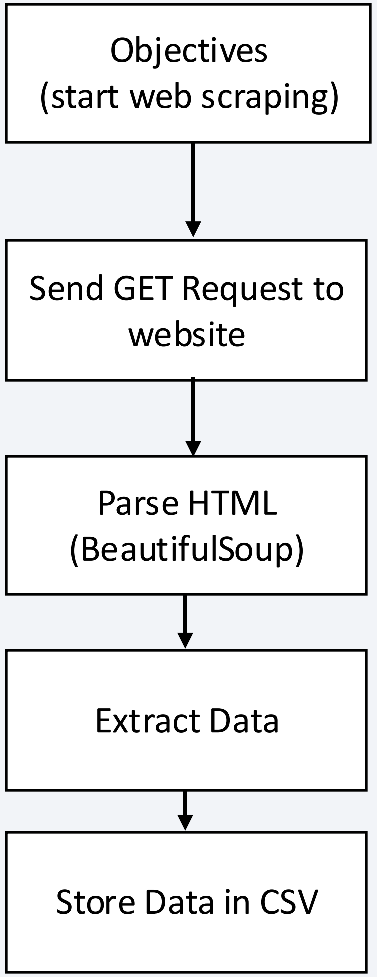
Data Extraction:

- Extract specific elements such as tables, launch dates, success/failure, etc.
- Filter the required data using BeautifulSoup's methods like .find(), .find_all(), and .text.

Store Data:

- Store the scraped data in a structured format like CSV or Excel for further analysis.

https://github.com/yhakam/IBM-Final-Project/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

Data wrangling refers to cleaning and transforming raw data into a usable format for analysis. This process typically includes handling missing values, converting data types, and feature engineering.

- Import Data: Loaded CSV file using pandas.read_csv().
- Missing Data: Handled using dropna() or fillna().
- Feature Engineering: Created Season and extracted date components.
- Encoding: Converted categorical variables to numerical (e.g., success = 1, failure = 0).
- Outliers: Removed extreme values in numerical columns.
- Normalization: Scaled numerical features using Min-Max.
- Final Prep: Selected relevant features for modeling.

https://github.com/yhakam/IBM-Final-Project/blob/main/jupyter-labs-webscraping.ipynb

EDA with Data Visualization

Bar Chart (Launch Success vs Orbit Type):

https://github.com/yhakam/IBM-Final-Project/blob/main/edadataviz.ipynb

- Purpose: To visualize the success rate based on the orbit type.
- Why: Helps understand the Orbit types their success rate.

Catplot

- Purpose: To explore how the payload mass affects the flight number and the success (class) of the landing.
- Why: This plot helps identify if there's a relationship between the payload mass and the success rate of the mission, as well as how the success varies across different flight numbers. This can offer insights into whether certain payload masses or certain flight numbers are linked to higher chances of success or failure.

Scatter Plot (Flight Number/Payload Mass vs Launch Site):

- Purpose: To visualize the relationship between flight number/payload mass and Launching Sites
- Why: Helps understand how payload mass and flight number depends on the Launching site, which can influence predictions on landing outcomes.

- Line Plot (Success Rate Throughout the Year):
  - Purpose: To track how the success rate of SpaceX landings has changed over time throughout the year.
  - Why: This plot helps visualize trends and patterns in the success rate, showing whether there has been an improvement or fluctuation in the success of landings as time progresses. It can help identify seasonal trends or operational changes that might have impacted the mission success rates

EDA with SQL

- Create Table: Defined SPACEXTBL with mission details.
- Select Launch Sites: Queried unique Launch_Site. (DISTINCT)
- Total Payload for NASA (CRS): Summed Payload_Mass__KG_ for NASA. (SUM)
- First Successful Landing: Retrieved the earliest successful landing date.
- Conditional Queries (Where, Like)

https://github.com/yhakam/IBM-Final-Project/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Markers

- What: Placed at each launch site's latitude/longitude.
- Why: To pinpoint the exact location of each launch site and provide popups with site-specific details.

Circles

- What: Drawn around launch site markers with a defined radius.
- Why: To visually emphasize the site's area of influence or highlight its significance.

Lines (Polylines)

- What: (If included) Connect multiple launch sites.
- Why: To illustrate spatial relationships or potential flight paths between sites.

https://github.com/yhakam/IBM-Final-Project/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

Pie Chart (Launch Success Distribution)

- What: Displays overall success count for all sites or success vs. failure for a specific site.
- Why: Quickly shows the proportion of successful vs. failed launches.

Range Slider (Payload Mass Filter)

- What: Selects a range for payload mass (in kg).
- Why: Filters data to analyze how payload mass influences launch success.

Scatter Plot (Payload vs. Launch Success)

- What: Plots payload mass against launch outcome, with points colored by booster version.
- Why: Visualizes correlation between payload and success, highlighting performance differences across booster versions.

Interactive Callbacks

- What: Update the pie and scatter charts based on dropdown and slider inputs.
- Why: Enhances user engagement by providing real-time, interactive data exploration.

https://github.com/yhakam/IBM-Final-Project/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

Data Preparation:

- Split dataset into training and test sets.
- Selected relevant features for prediction.

Model Building:

- Built multiple classification models (e.g., Logistic Regression, Decision Tree, SVM, KNN).
- Model Evaluation:

Employed cross-validation and calculated metrics (accuracy, precision, recall, F1-score).

- Used confusion matrices for detailed error analysis.
- Model Improvement:

Applied GridSearchCV for hyperparameter tuning.

- Compared cross-validated scores to determine the best model.

Best Model Selection:

- Chose the model with the highest mean cross-validation score and balanced performance metrics. (Linear Regression)

https://github.com/yhakam/IBM-Final-Project/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

Exploratory data analysis results :

- Launch Outcomes & Trends:
  - Visualized success/failure distributions across launch sites : KSC LC-39A represents 41,7% of the total Success launches, with a 76,9% success rate
  - Identified correlations between payload mass, Booster Version, and landing success. FT is the version with the highest success rate
- Temporal Analysis:
  - Detected seasonal trends in landing success via line plots. Success rate is growing since 2013
- Feature Relationships:
  - Catplot (Flight Number, Payload Mass, and Class):
    - Pattern: Successes (class = 1) tend to cluster at certain flight numbers and lower payload mass ranges.
    - Insight: Lower payload masses (typically under 6000 kg) and early flight numbers show higher success rates, suggesting an operational “sweet spot” where performance is optimal.
  - Scatter Plot (Payload Mass vs. Landing Success):
    - Pattern: A negative correlation is evident—launches with lower payload mass are more likely to be successful.
    - Insight: There’s a distinct clustering where many successful landings occur at lower payload masses, while higher payloads show more variability in outcomes. Additionally, grouping by booster version highlights that some versions consistently perform better across payload ranges. FT version especially
  - Bar Plot (Orbit types, Flight number and Success Rates):
    - Pattern: Certain Orbit type have way higher success Rates : GEO, HEO, SSO Because of their high flight numbers

Results

Folium map

Markers for Launch Sites:

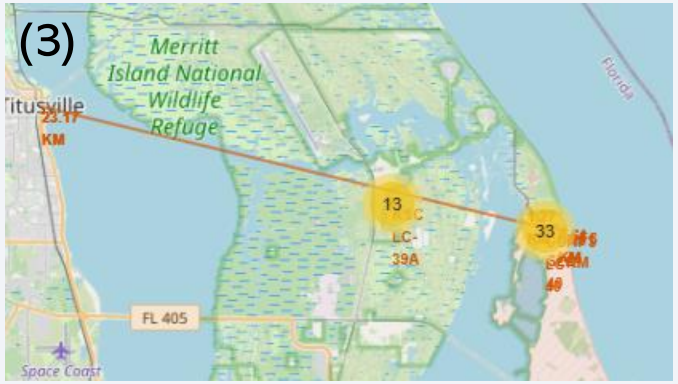
- Placed markers at each launch site's latitude and longitude. (1)

Spatial Distribution:

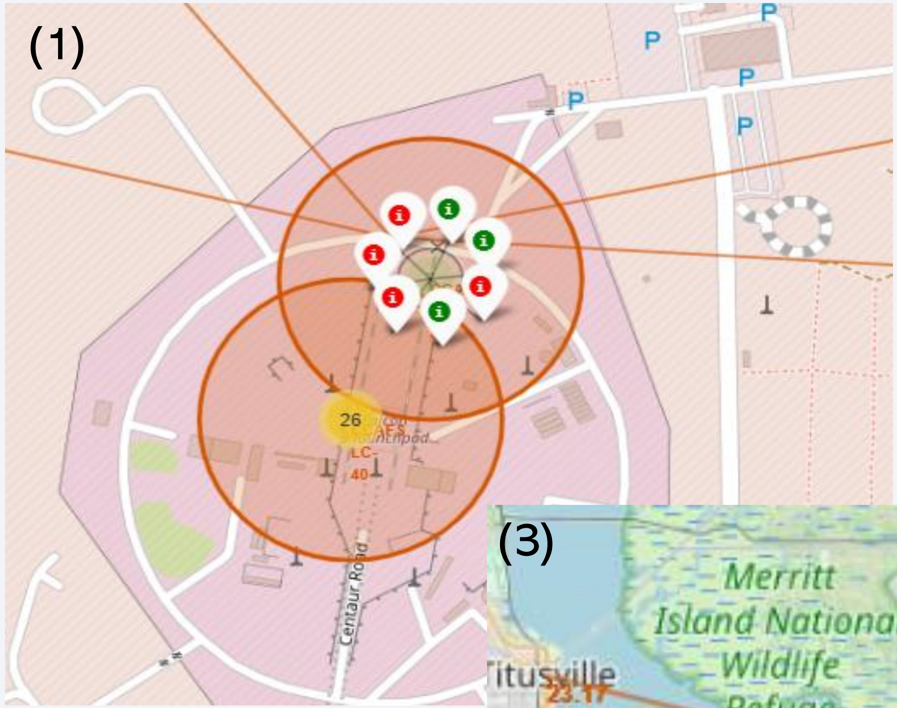
- Reveals clustering of launch sites (1).
- Highlights geographical patterns that may influence launch outcomes: Launching sites are near Highway and railway, but far from cities. (2) (3)

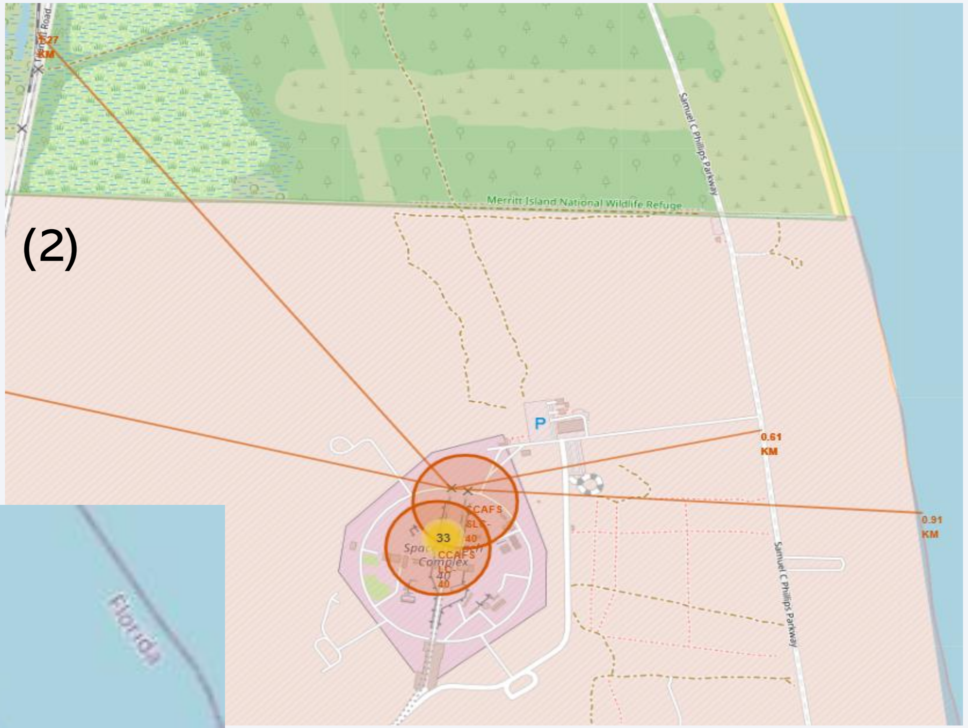
Interactive Elements:

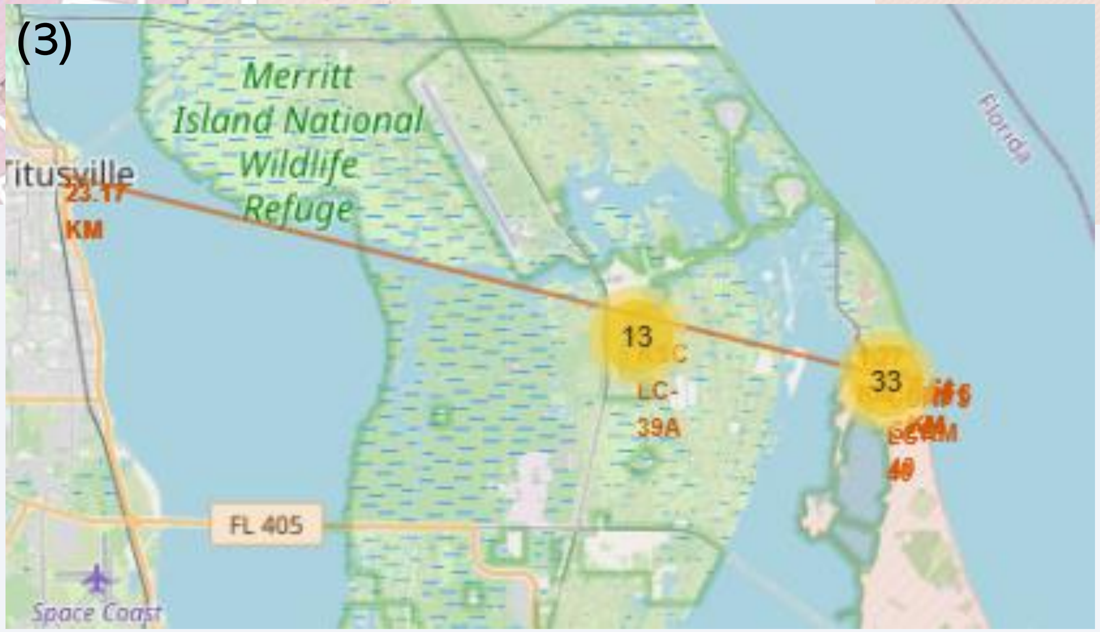
- Enables users to zoom and pan for detailed local insights.
- Provides a visual context for operational regions.



Results


(1)
26
LC-
40
Centaur Road

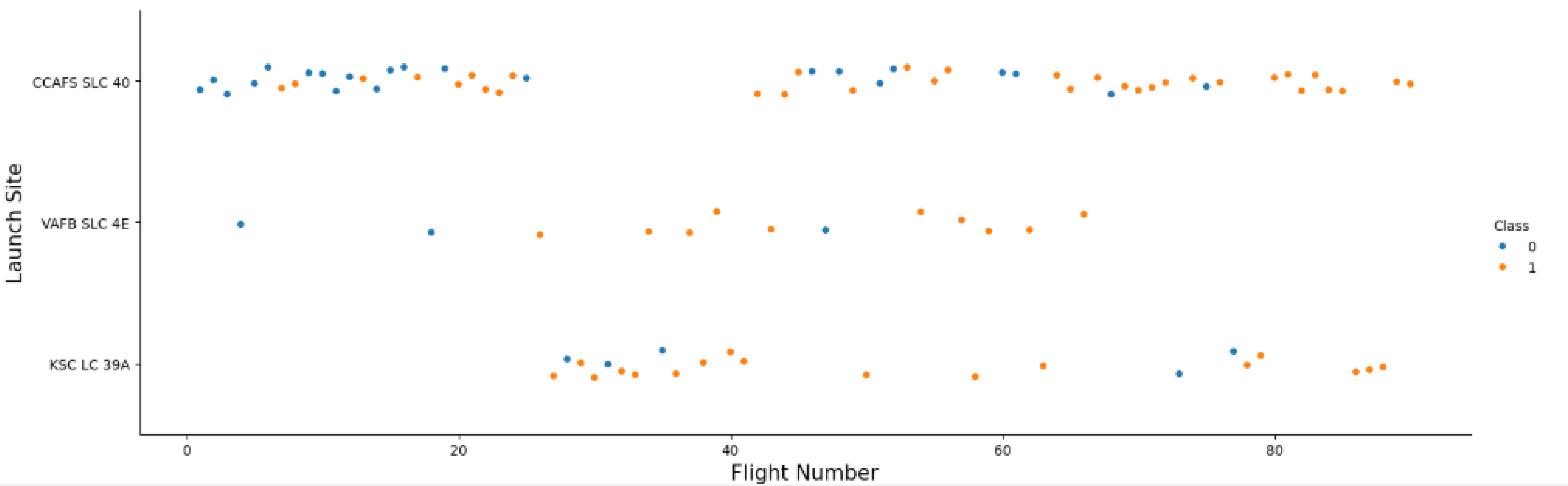

(2)
Merritt Island National Wildlife Refuge
0.61
KM
0.91
KM
33


(3)
Merritt
Island National
Wildlife
Refuge
Titusville
23.17
KM
13
LC-
39A
33
Florida
FL 405
Space Coast

Section 2

Insights drawn from EDA

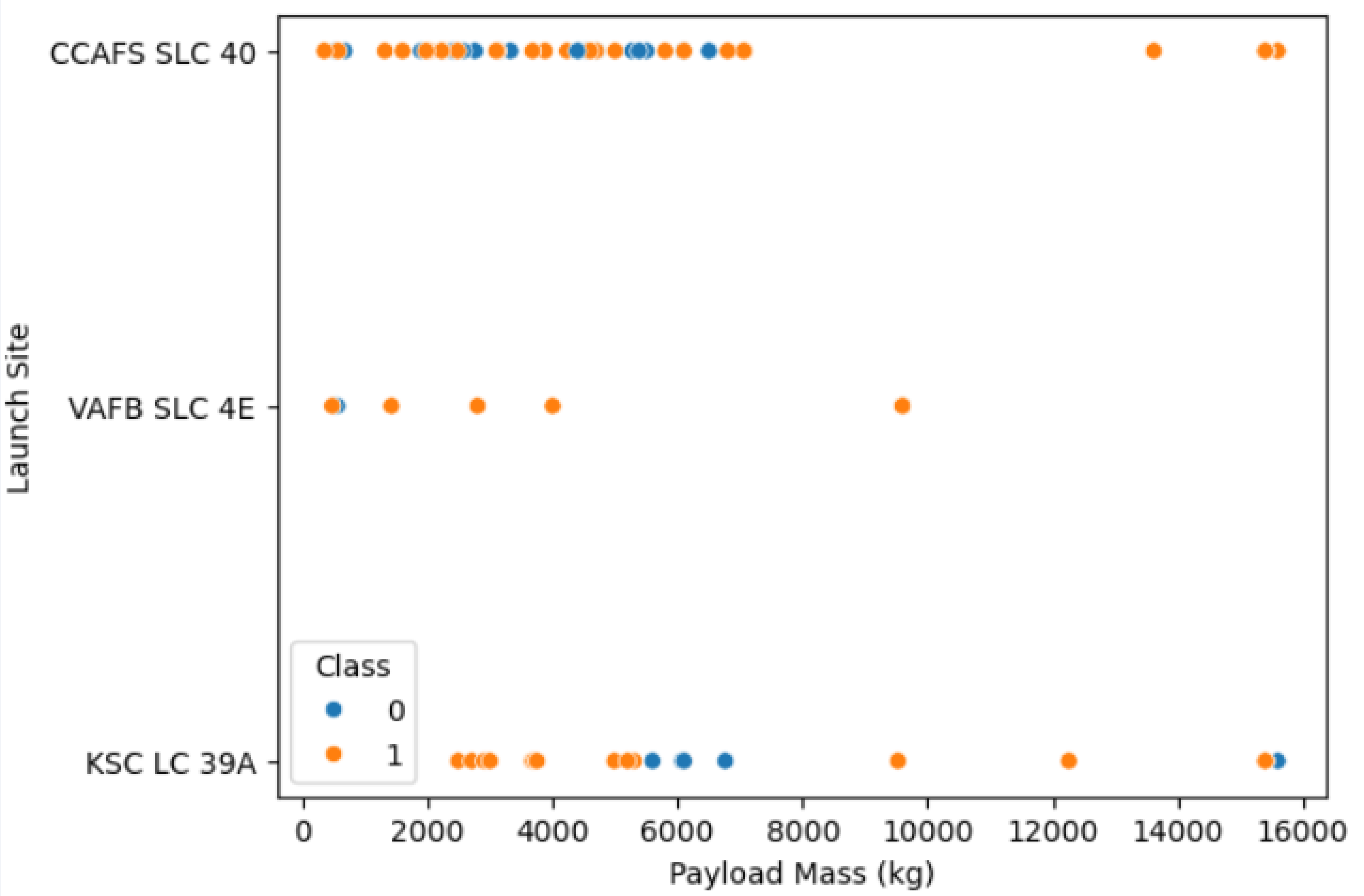
Flight Number vs. Launch Site



- We can see in this catplot that CCAFSSLC40 has the highest number of Flights and his success is pretty much balanced, VAFB SLC 4E has fewer Flights but he still has more success than failure, we can explain that because maybe they invest more in each flights, so they have less in terms of numbers but they succeed way more, and Finaly KSC LC 39A has a moderate number of Flights, but he has the most success out of them all

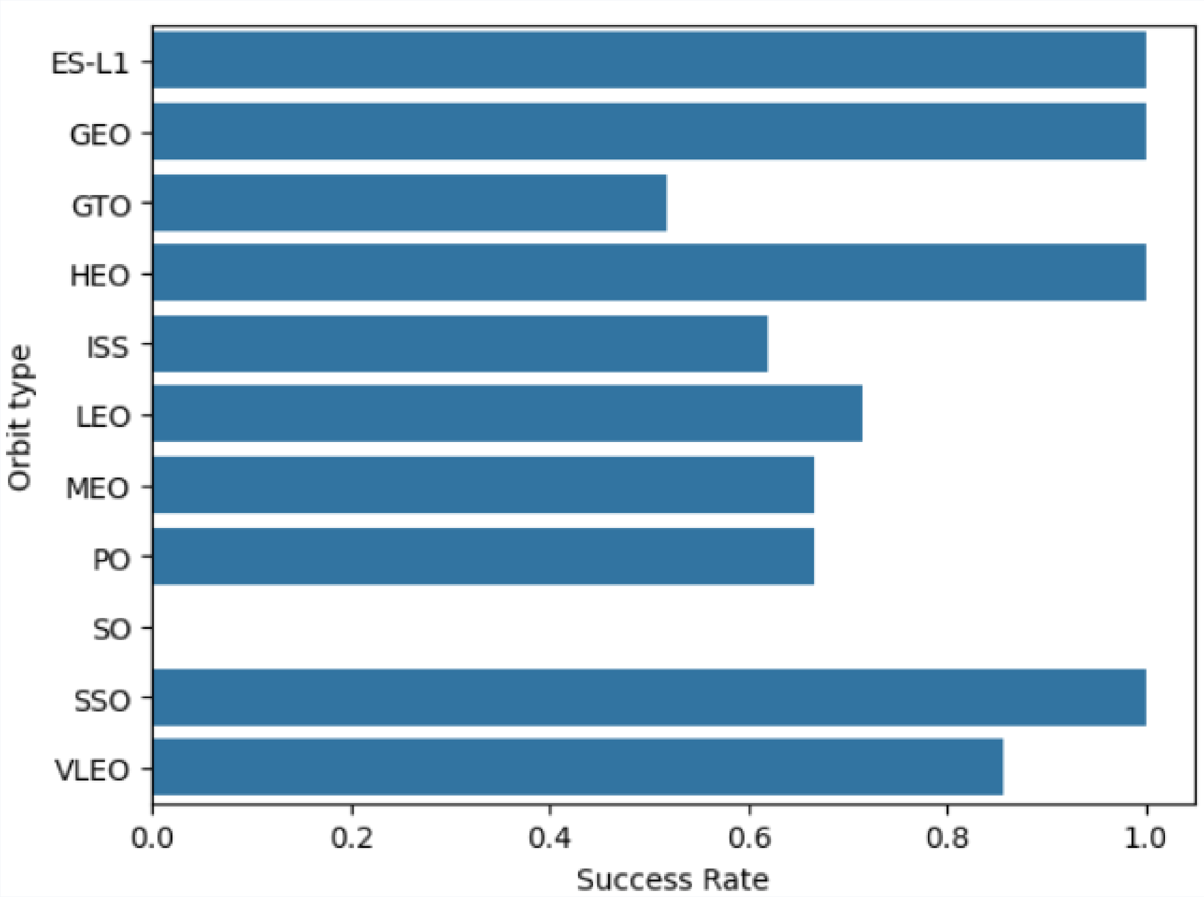
Payload vs. Launch Site

- Here we see that The Majority of success comes from Payload Mass between 0 and 5000KG



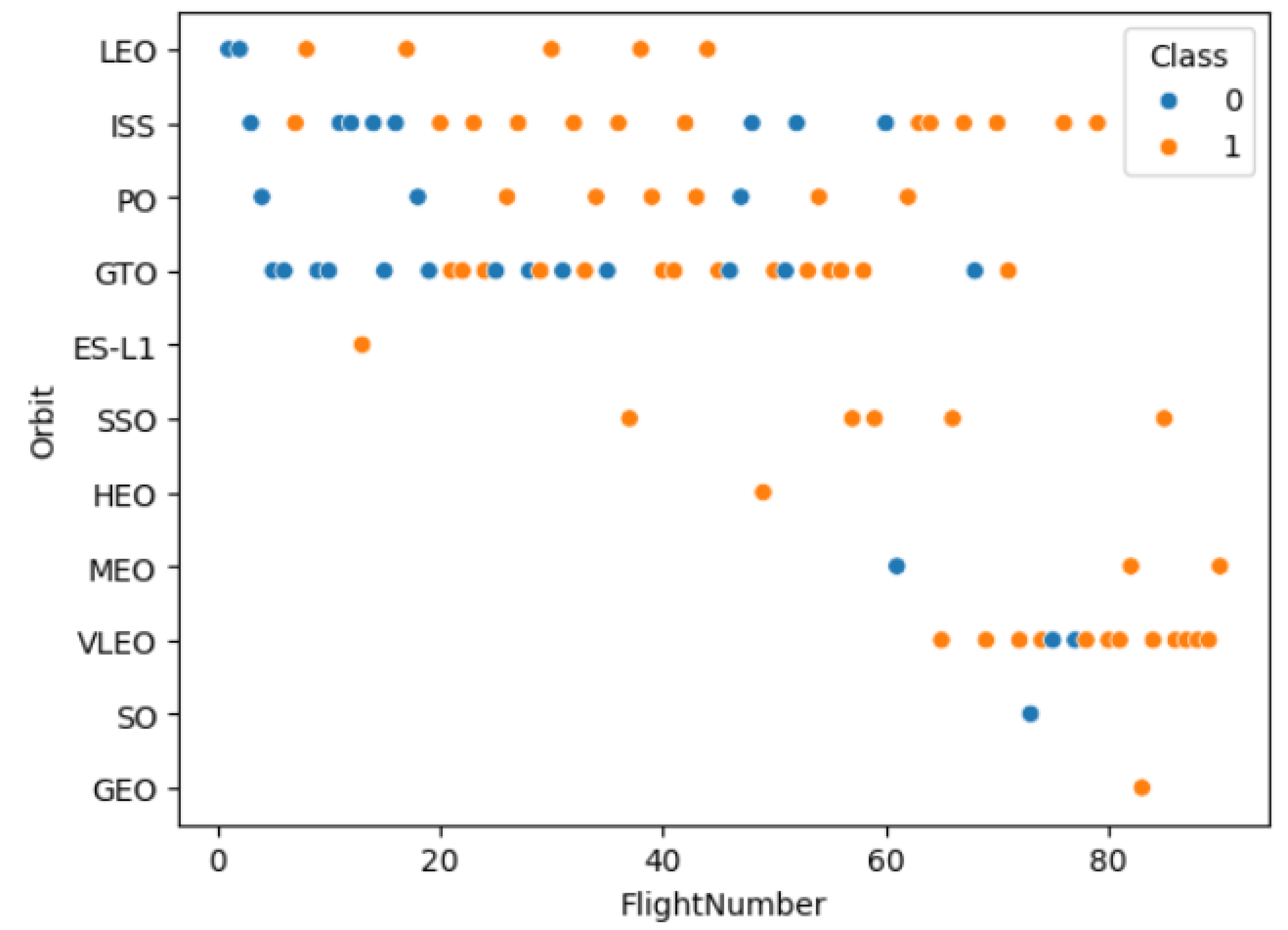
Success Rate vs. Orbit Type

- This Bar Chart Shows us That the orbit Types with the highest success rate are GEO, HEO and SSO, ES-L1 has a very low number of Flight so this is not very representative



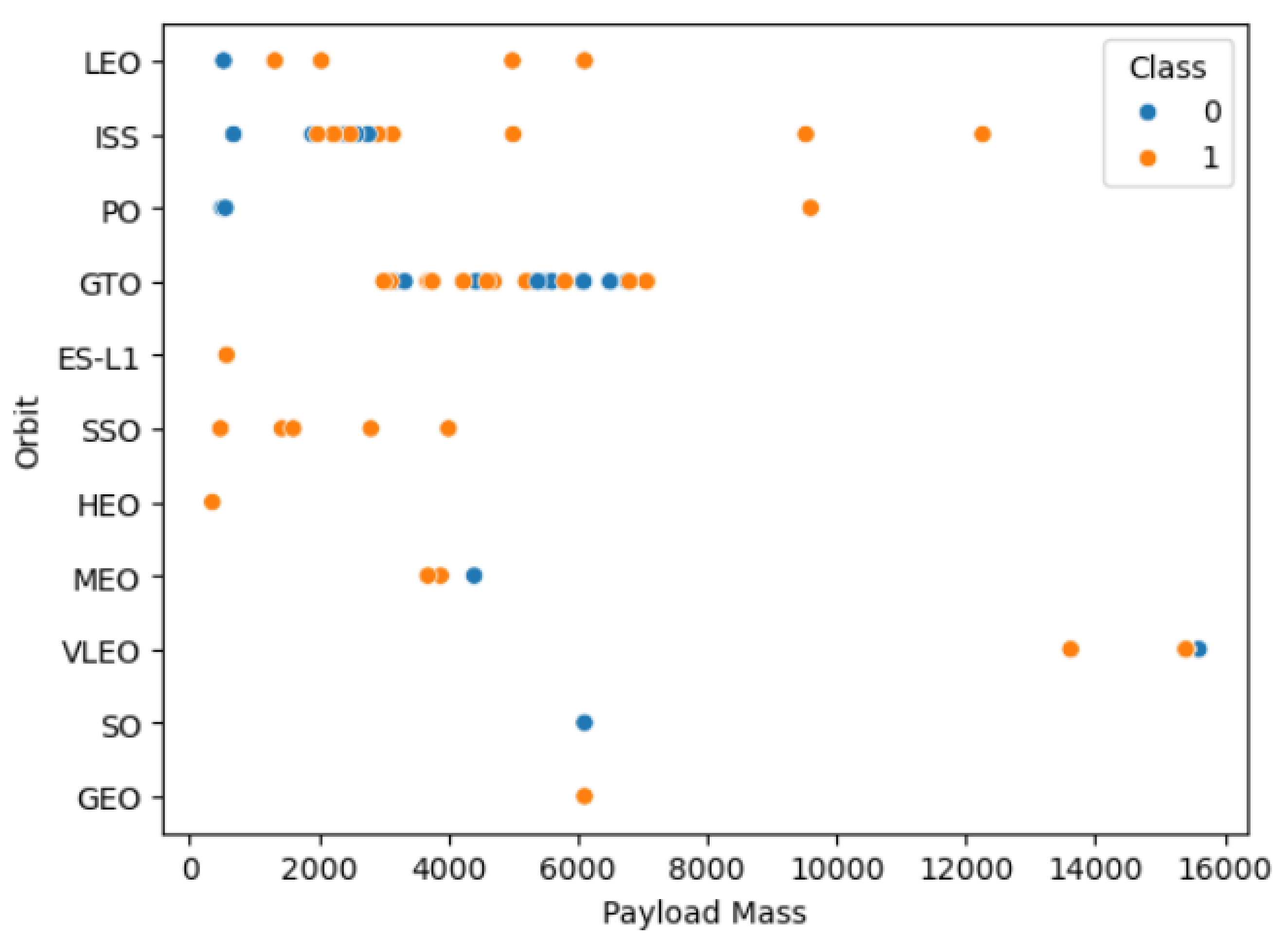
Flight Number vs. Orbit Type

- This Scatterplot shows us that for some orbit Types, the more SpaceX make flights the more the success rate increase, this is especially the case with SSO, VLEO, ISS and LEO we recognize a pattern, than the majority of Flight that exceed 40 are successes



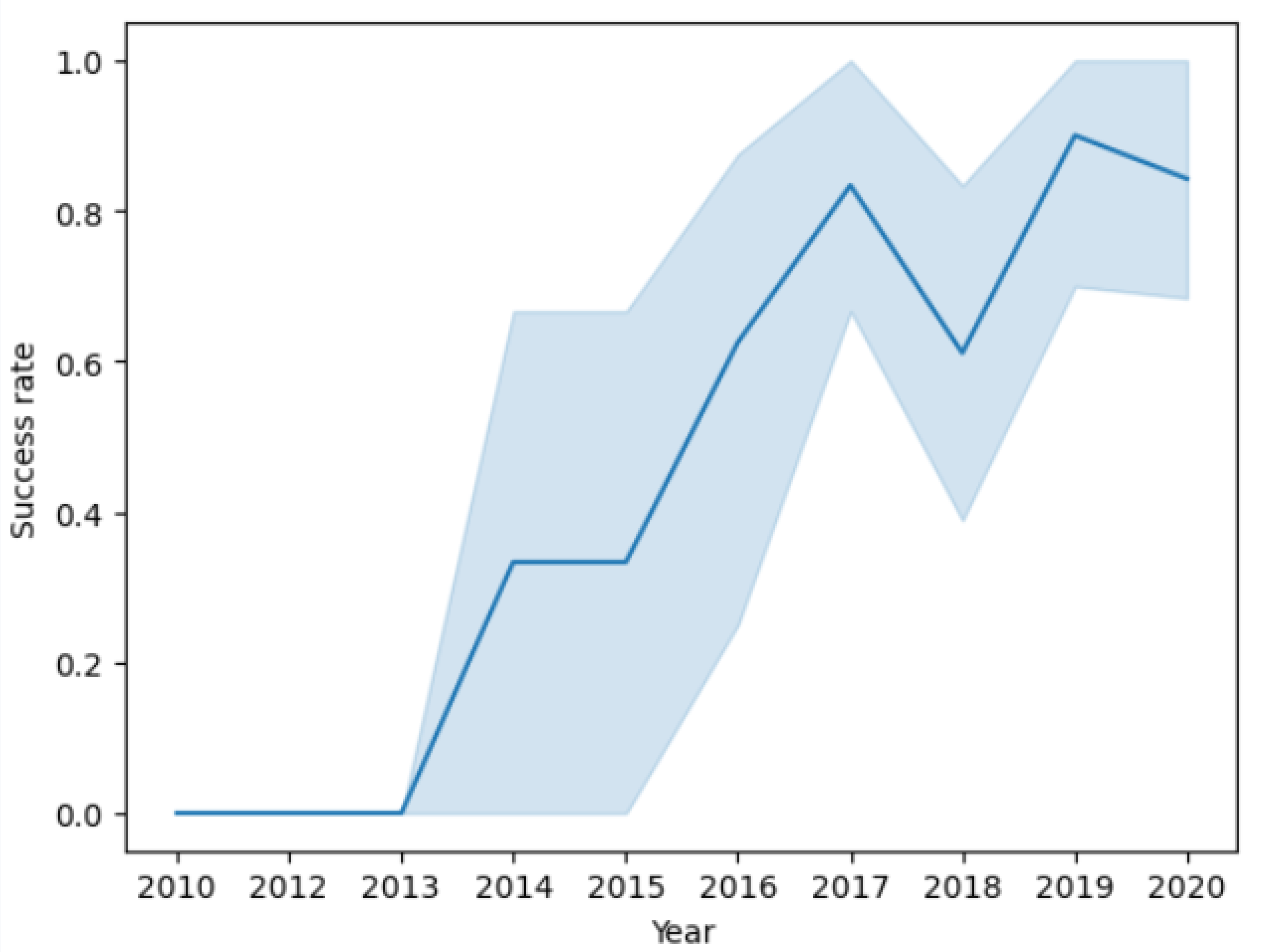
Payload vs. Orbit Type

- We can see in this Scatter Plot that the Majority of Orbit Type doesn’t have record after a Payload Mass of 6000KG, we can note that the trend is the contrary for VLEO, we don’t see any record before 12 000KG



Launch Success Yearly Trend

- This Show us The Success rate throughout the year, we can clearly see that since 2013 we have a big increase in success rate until 2020



All Launch Site Names

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

- This shows us all the Launch Sites

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

- This Shows us the Five First record from Launch Site that begins by “CCA”

Total Payload Mass

Payload	Total payload mass carried by boosters launched by NASA (CRS)
SpaceX CRS-1	111268

- This Shows us The total Payload mass Carried by boosters launched by Nasa (CRS)

Average Payload Mass by F9 v1.1

Booster_Version	AVG
F9 v1.1 B1003	2534.6666666666665

- This shows us the Average Payload mass in KG of the Booster Version F9 v1,1

First Successful Ground Landing Date

Date	Landing_Outcome
2015-12-22	Success (ground pad)

- This shows us the First successful landing in ground pad

Successful Drone Ship Landing with Payload between 4000 and 6000

Booster_Version	Landing_Outcome	PAYLOAD_MASS__KG_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

- This query shows us the Booster version wich have success in drone ship and have a payload mass between 4000 and 6000 KG

Total Number of Successful and Failure Mission Outcomes

Mission_Outcome	total number of successful and failure mission outcomes
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- We had 100 mission successs and only one Failure, That shows that the majority of Landing Failure are still considered a mission Success because some of the landing failure are predetermined

Boosters Carried Maximum Payload

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

- This is the list of All Booster Version that reached the Max Payload Mass (15600)

2015 Launch Records

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

- In 2015 we only Had two drone ship Failure in January and April

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

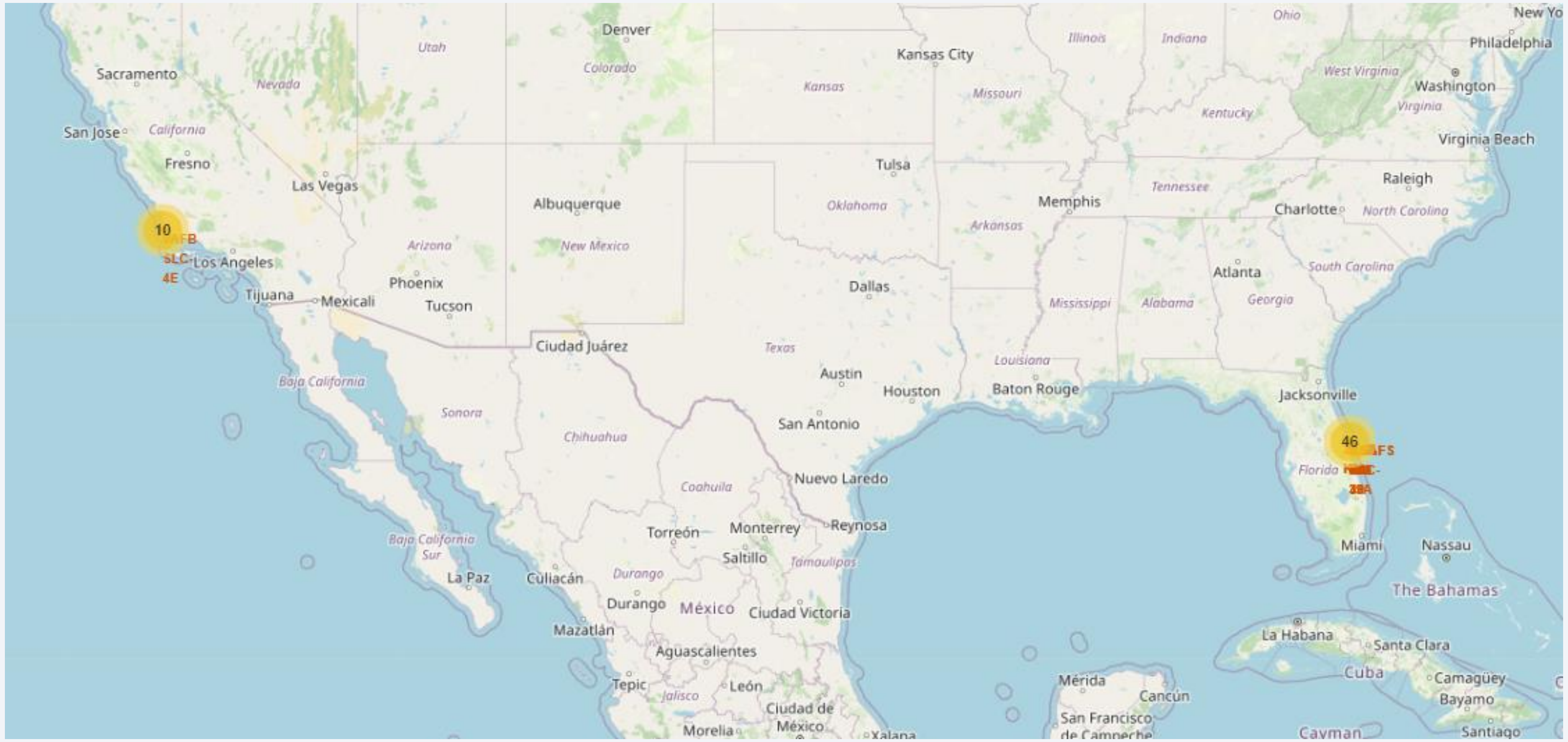
Date	Landing_Outcome	COUNT(Landing_outcome)
2015-12-22	Success (ground pad)	3
2015-01-10	Failure (drone ship)	5

- Here we can see that in 2015 SpaceX had 5 Failure on Drone Ship and 3 Success on ground pad

Section 3

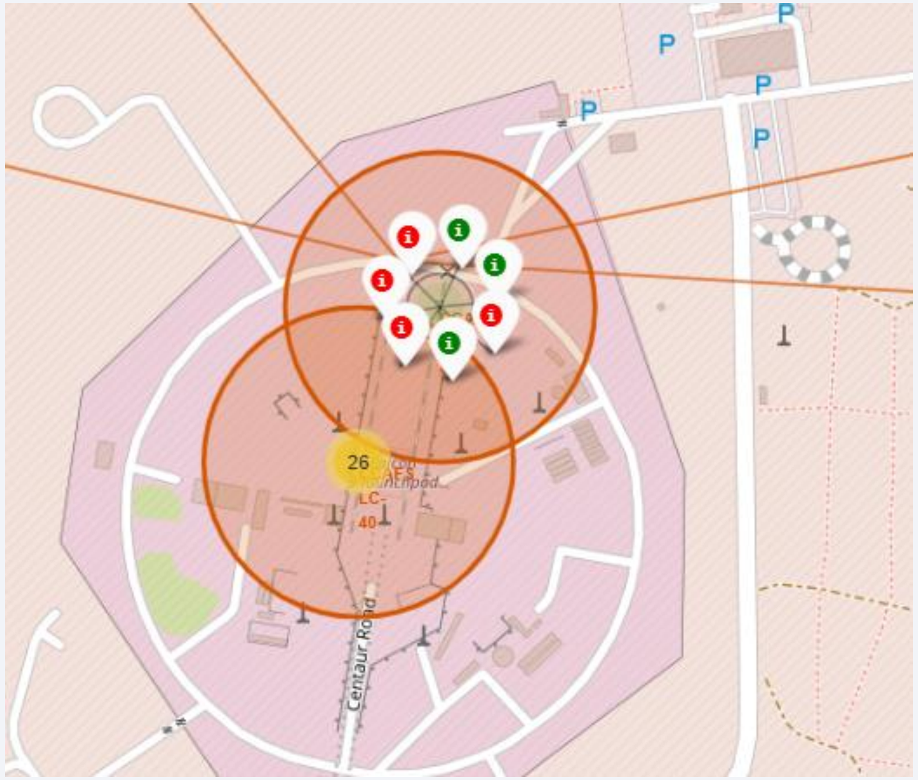
Launch Sites Proximities Analysis

Launching sites Clusters in the Global Map



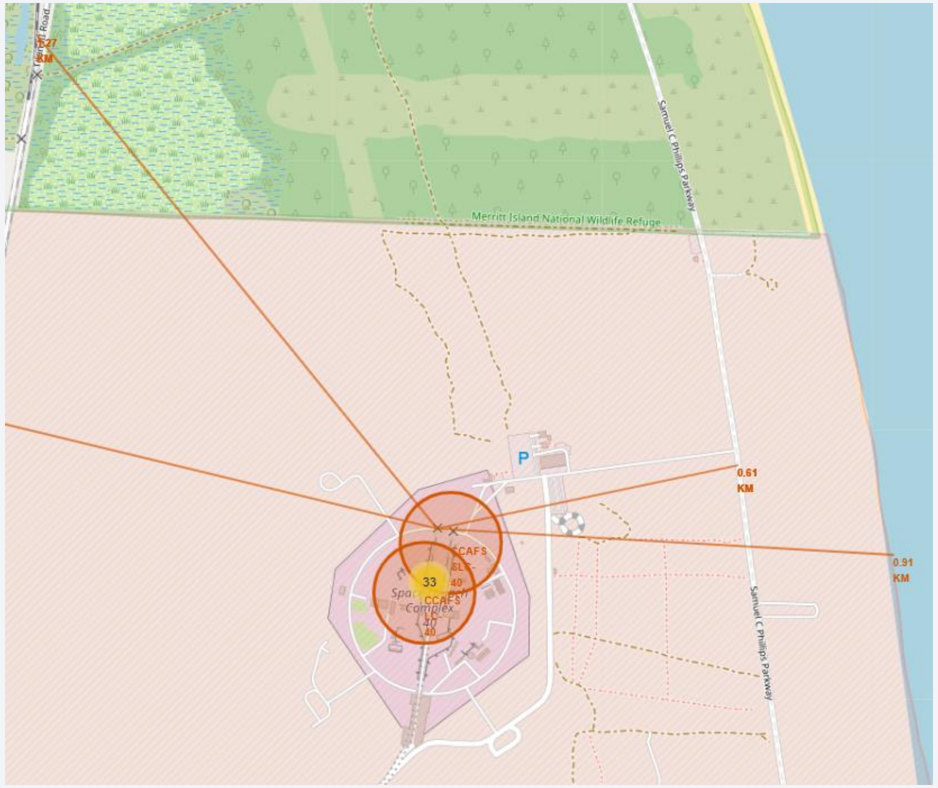
- We can see that launching sites are located in the east and west coast of America, because they are usually plain areas ideal to launch spaceships

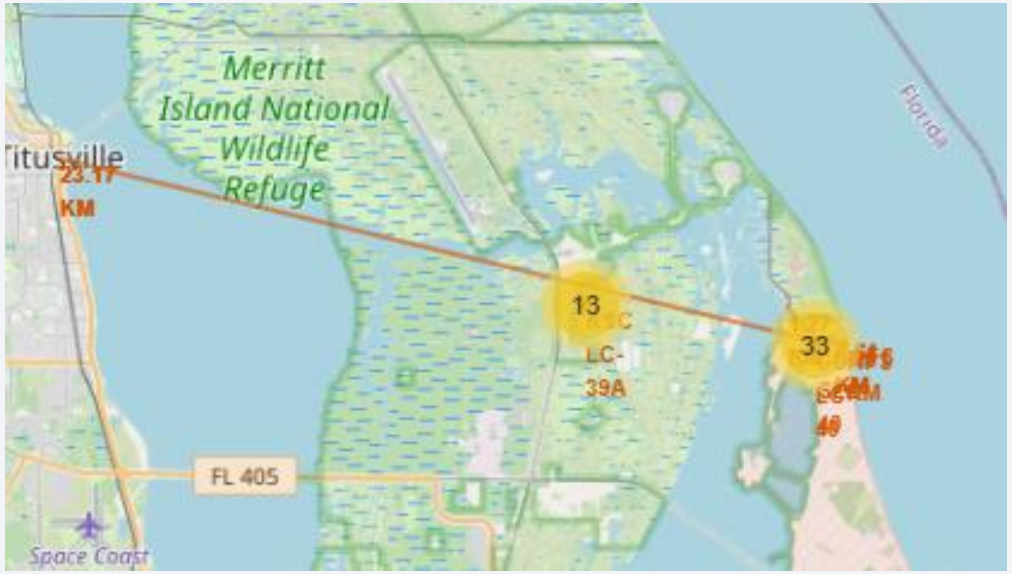
Markers of Succesful landings and Failed ones



- In green we see the successful landings, and in red we see the failed ones, we can see that they are all concentrated at around the same place, that’s why their success rates are pretty much the same

Distance Between Launching sites and highway,railway and cities



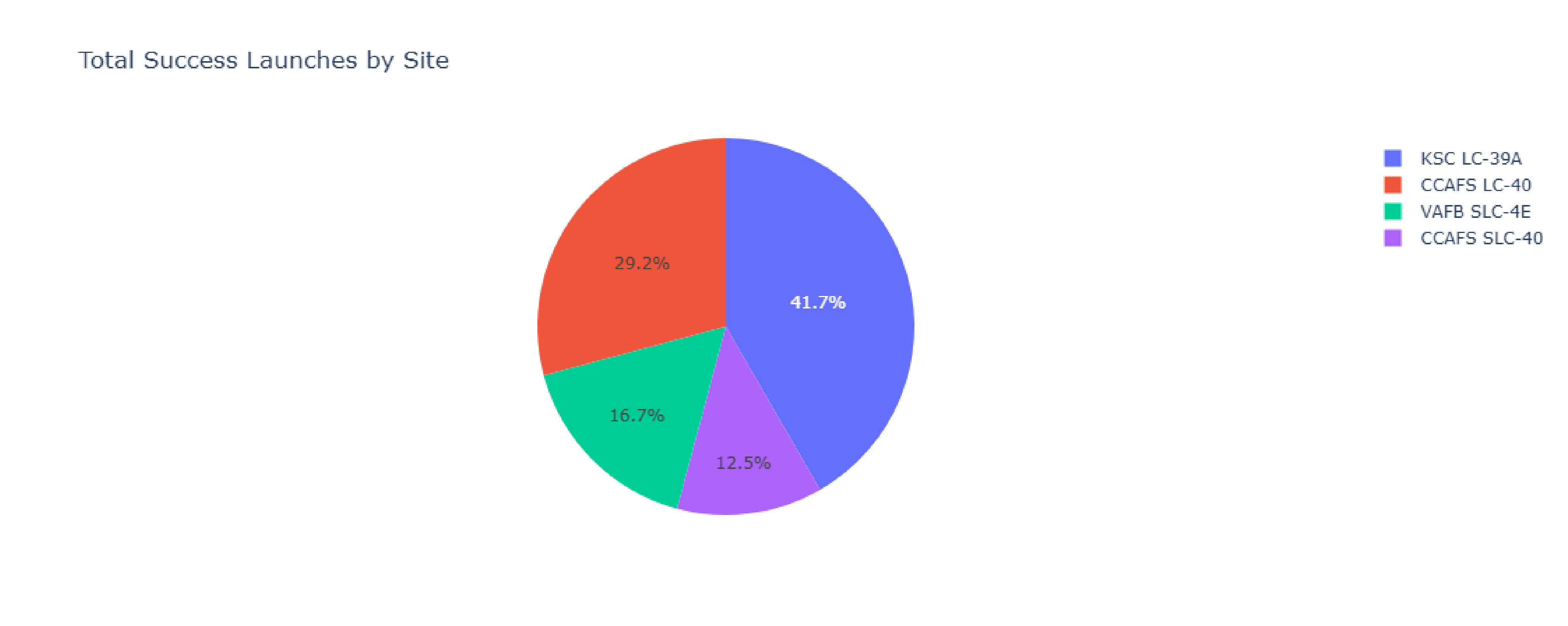


- We can see that the launching sites are near highway and railways but far from cities

Section 4

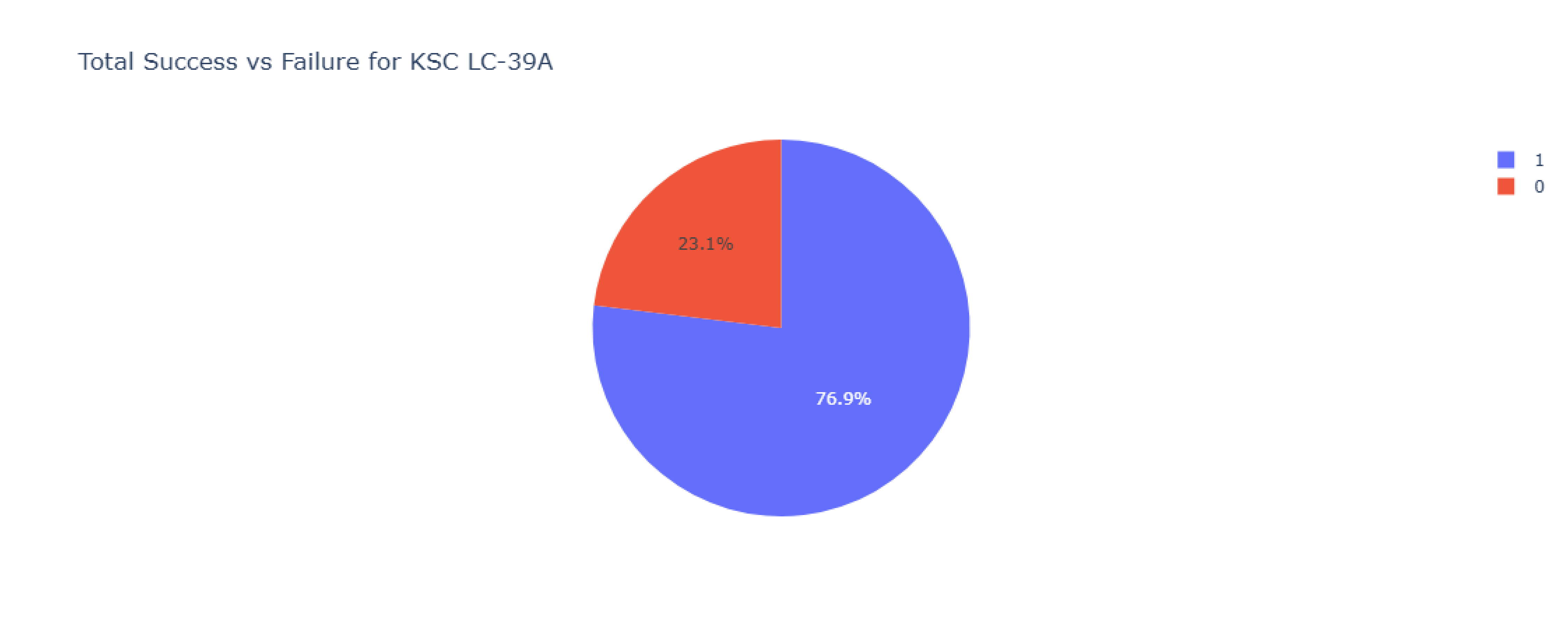
Build a Dashboard with Plotly Dash

Total Success Launches by Site



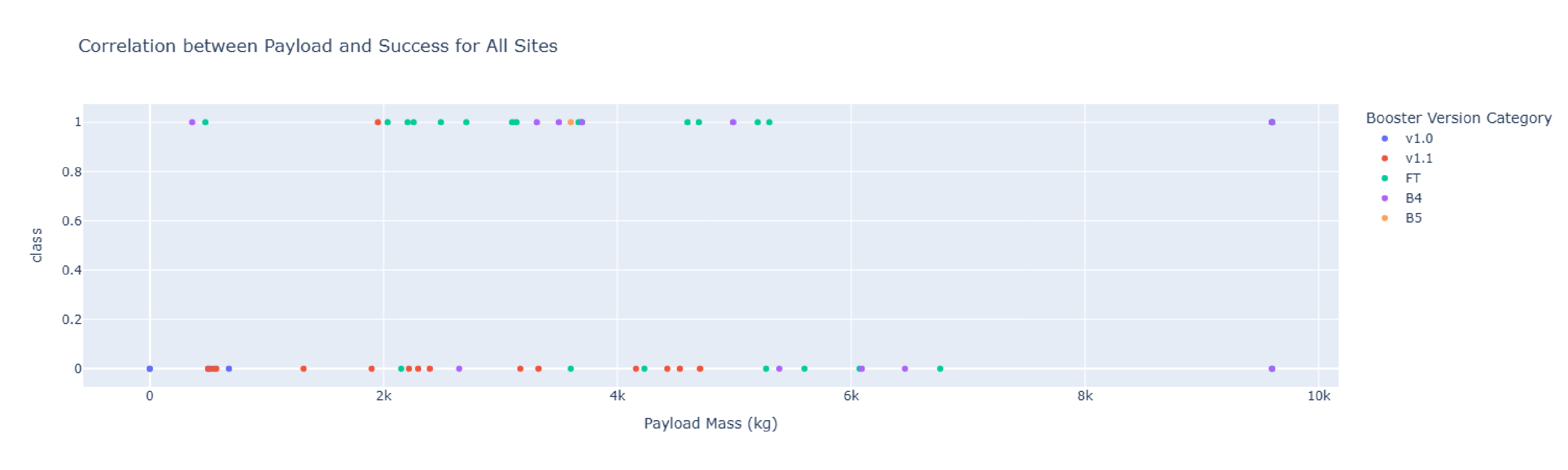
- We can see that KSC LC-39A has the highest number of success landings

Total Success vs Failure for KSC LC-39A



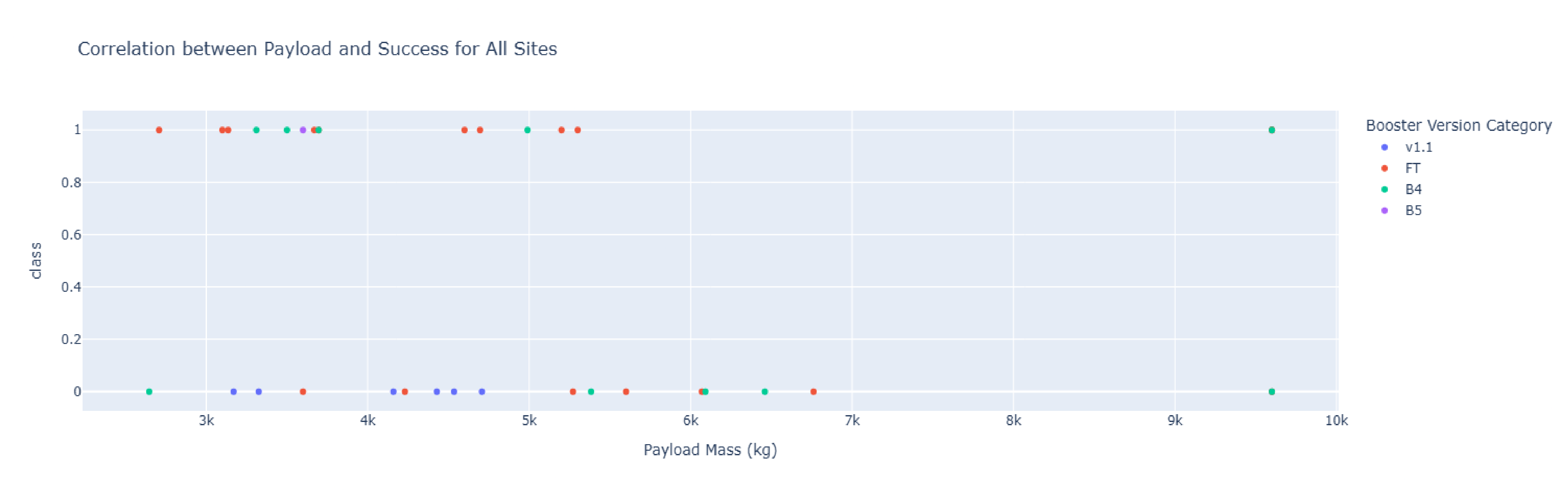
- KSC LC-39A also has the highest success rate

Correlation Between Payload Mass and Success rate for All Sites



- We can see that the Booster version with the highest success landings is FT
- And the highest number of success landings are found when Payload Mass is between 2000 and 6000 KG

Correlation Between Payload Mass and Success rate for All Sites

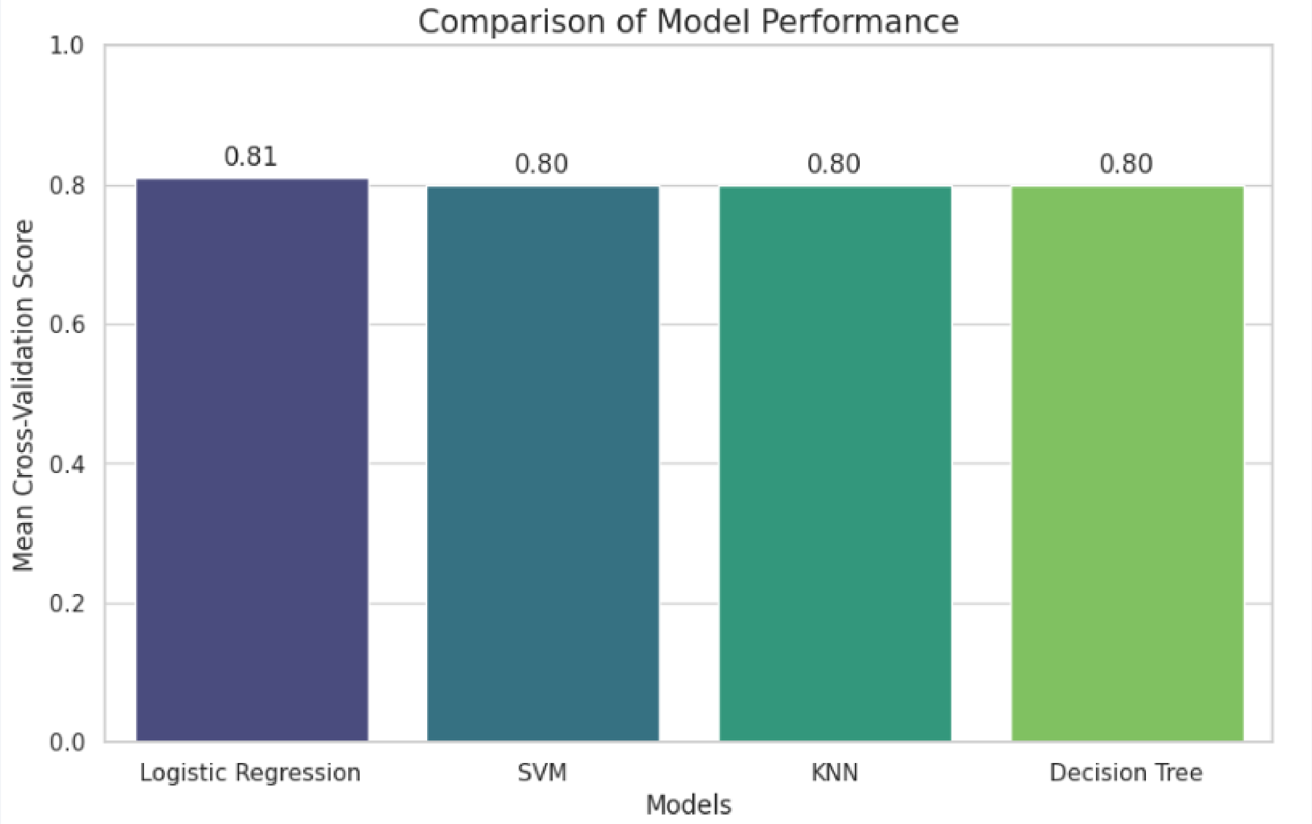


- Even if we change the payload Mass FT is still the best version in terms of success rate

Section 5

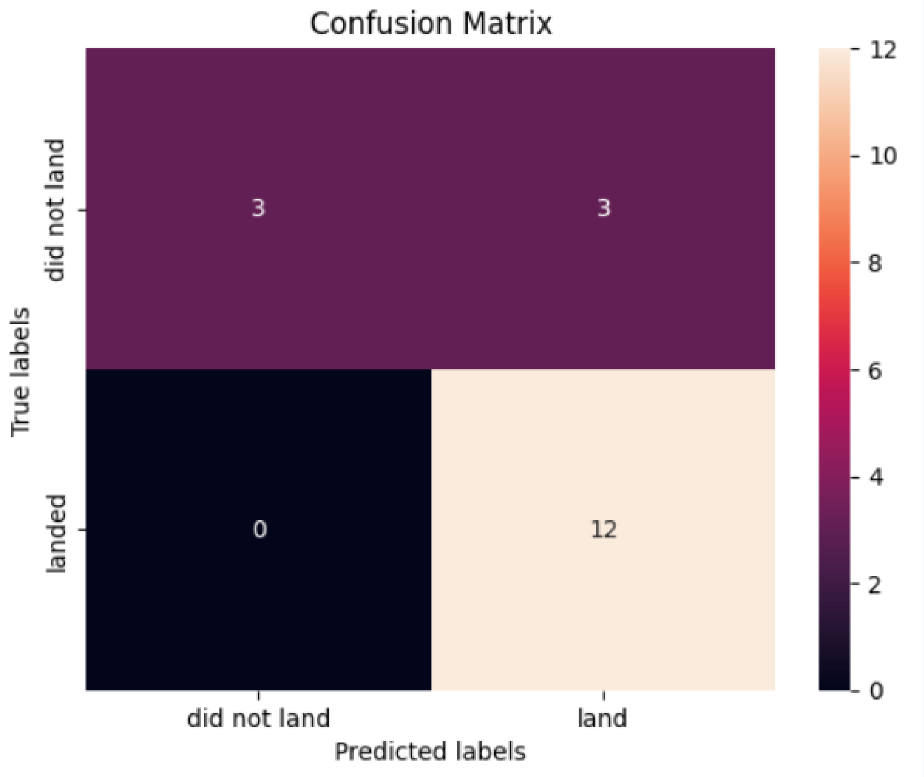
Predictive Analysis (Classification)

Classification Accuracy



- I did a Cross validation Score test, because all the confusion matrix was similar, so base on the mean Cross-Validation Score, Linear Regression has a better accuracy than the other models

Confusion Matrix



- Here we have 12True Positive, 3 False Positive, and 3 True Negative, wich gives us an accuracy of 83,33%, the model predict that the landing will be successful with an accuracy of 83,33%

Conclusions

- To conclude, All our researches showed that SpaceX despite having a lot of Landing Failure has 99% of mission success, that shows that thoses Landing Failures was wanted, and by predicting the outcome with the data at our disposition, we saw that the model predicted 12 times out of 18 that the landing will be successfully, With all those Data we can now assume that indeed the Falcon 9 will cost 62 Million dollars, with an accuracy of 83,33%

Thank you!