IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Luis Alberto Sánchez

2024-02-05

Outline

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

Executive Summary

Analysis of data employed the following methodologies:

- Gathering data through web scraping and the SpaceX API.
- Exploratory Data Analysis (EDA), which involved data wrangling, data visualization, and interactive visual analytics.
- Using Machine Learning for prediction.

Summary of the results includes:

- Successful collection of valuable data from public sources.
- Identification, through EDA, of the most predictive features for launch success.
- Machine Learning Prediction revealing the optimal model for predicting key characteristics that drive successful opportunities, leveraging the entirety of the collected data.

Introduction

The goal is to assess the feasibility of the emerging company Space Y in competing with Space X.

Desired outcomes include:

- Developing an optimal method for estimating the overall cost of launches by predicting successful landings of the first stage of rockets.
- Determining the most favorable location for conducting launches.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data from Space X was obtained from:
    - Space X API (https://api.spacexdata.com/v4/rockets/)
    - Web scraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)
- Perform data wrangling
  - Data was labeled through the generation of a landing outcome label, derived from the summarized and analyzed feature data.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash

Methodology

Executive Summary

- Perform predictive analysis using classification models
  - The data collected was normalized and split into training and test datasets, and they were evaluated through four distinct classification models. The accuracy of each model was evaluated using various parameter combinations.

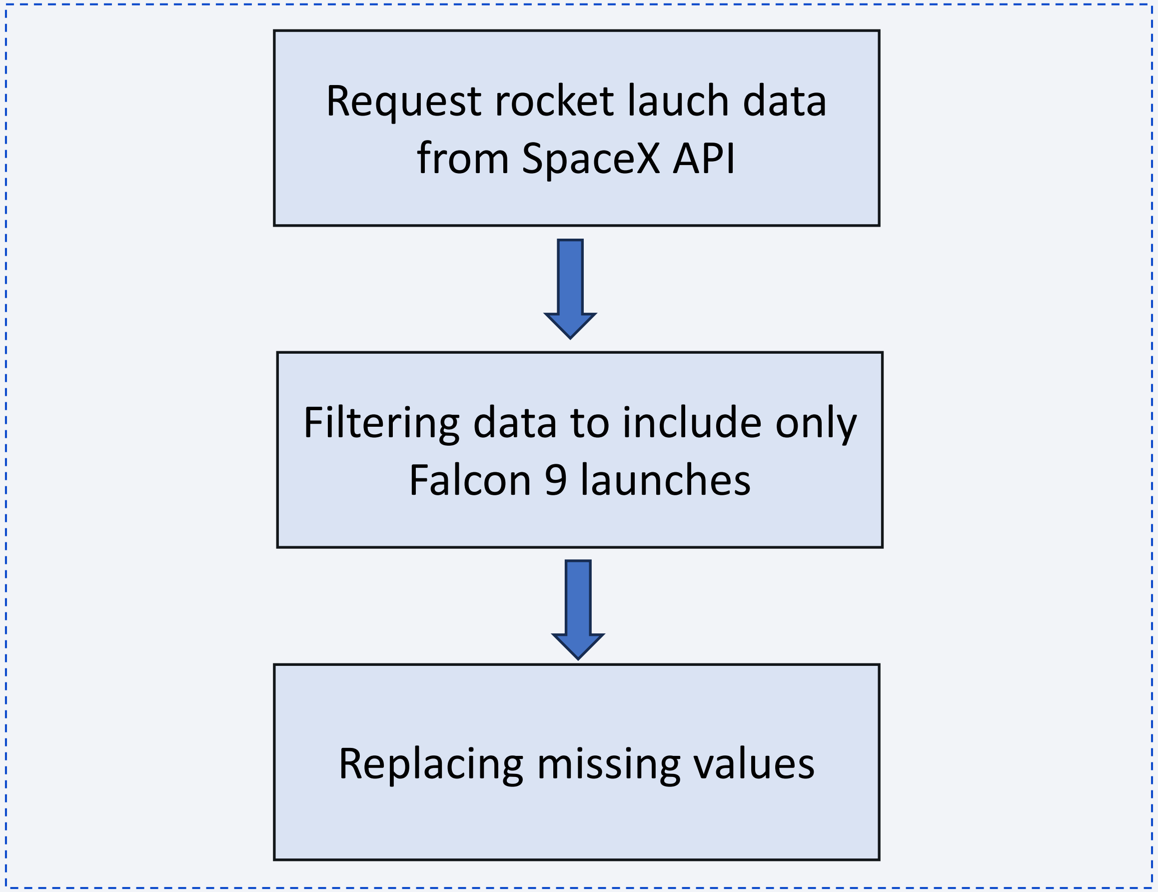
Data Collection

Data from Space X was obtained from:

- Space X API (https://api.spacexdata.com/v4/rockets/)
- Web scraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)

Data Collection – SpaceX API

- SpaceX provides a public API for accessing and retrieving data.
- The data retrieval process follows the flowchart depicted alongside, and the obtained data is subsequently stored for future use.

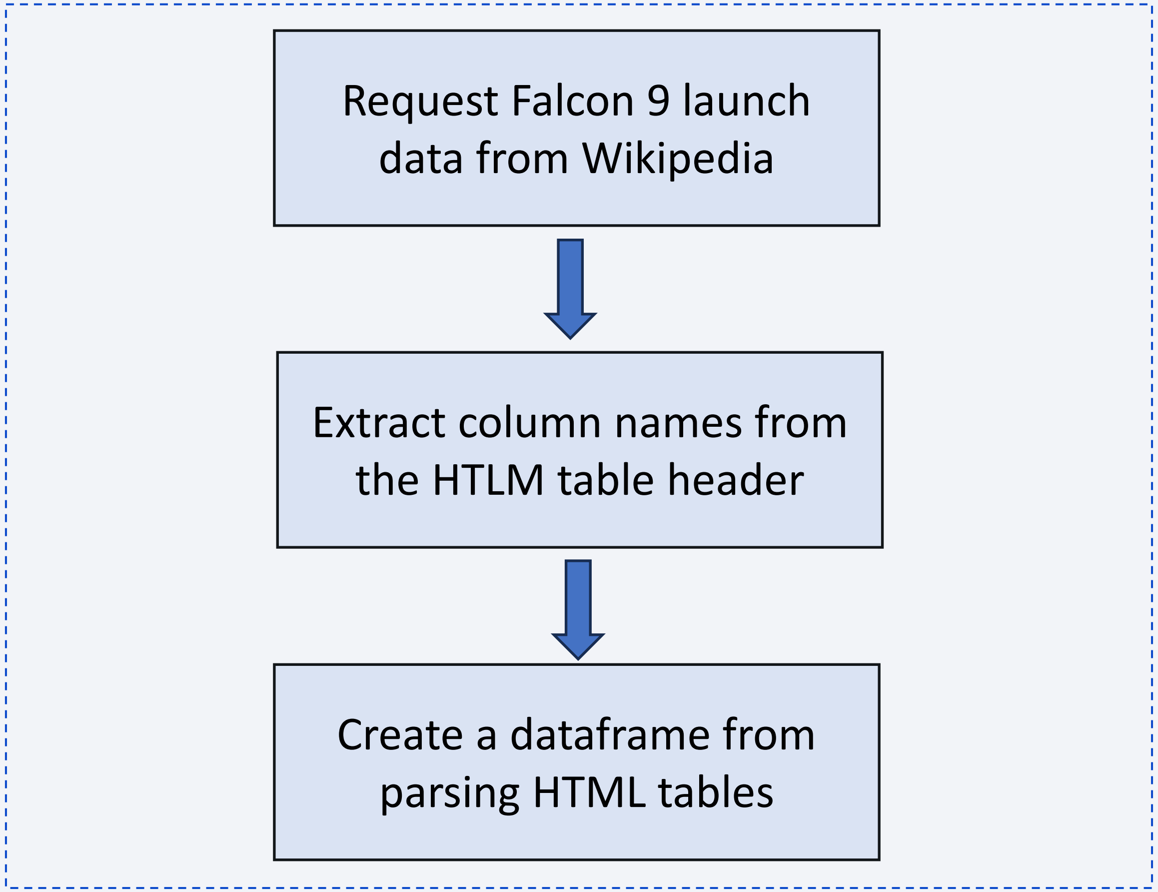


https://github.com/sanchez-luis/ibm-applied-data-science-capstone/blob/main/Lab%201%20Collecting%20the%20data.ipynb

Data Collection - Scraping

- Information regarding SpaceX launches can also be acquired from Wikipedia.
- The data retrieval process adheres to the outlined flowchart, and the downloaded data is subsequently stored for future use.

https://github.com/sanchez-luis/ibm-applied-data-science-capstone/blob/main/Lab%203%20Web%20scraping%20Falcon%209%20and%20Falcon%20Heavy%20Launches%20Records%20from%20Wikipedia.ipynb



Data Wrangling

- Various cases in the dataset involve unsuccessful booster landings, with outcomes determined by factors like accidents.
- Different codes, such as True Ocean, True RTLS, and True ASDS, signify successful landings, while their counterparts (False Ocean, False RTLS, False ASDS) represent unsuccessful attempts.
- The outcomes are predominantly transformed into Training Labels, with "1" denoting a successful landing and "0" indicating an unsuccessful one.

Exploratory Data Analysis → Summaries lauches per site, ocurrencer of each orbit and ocurrences of misión outcome → Create a landing outcome label

https://github.com/sanchez-luis/ibm-applied-data-science-capstone/blob/main/Lab%202%20Data%20wrangling%20.ipynb

EDA with Data Visualization

- Charts plotted:

Flight Number vs. Payload Mass	Flight Number vs. Launch Site
Payload Mass vs. Launch Site	Orbit Type vs. Success Rate
Flight Number vs. Orbit Type	Payload Mass vs Orbit Type and Success Rate Yearly Trend

Scatter plots visually depict the connection between variables, providing potential inputs for machine learning models if a discernible relationship is present.

Bar charts facilitate comparisons across discrete categories, aiming to illustrate the correlation between specific categories and their corresponding measured values.

Line charts, designed for time series data, reveal trends over time, offering insights into the temporal evolution of the dataset.

https://github.com/sanchez-luis/ibm-applied-data-science-capstone/blob/main/Lab%205%20EDA%20with%20Visualization%20Lab.ipynb

EDA with SQL

SQL queries performed:

- Identifying the names of unique launch sites in the space mission dataset.
- Listing the top 5 launch sites with names starting with the string 'CCA.'
- Calculating the total payload mass carried by boosters launched by NASA in the CRS program.
- Determining the average payload mass carried by the booster version F9 v1.1.
- Identifying the date of the first successful landing outcome on a ground pad.
- Listing the names of boosters with successful drone ship landings and payload mass between 4000 and 6000 kg.
- Analyzing the total number of successful and failed mission outcomes.
- Identifying the names of booster versions that carried the maximum payload mass.
- Listing failed landing outcomes on drone ships in 2015, along with their booster versions and launch site names.
- Ranking the count of landing outcomes (e.g., Failure (drone ship) or Success (ground pad)) between June 4, 2010, and March 20, 2017.

https://github.com/sanchez-luis/ibm-applied-data-science-capstone/blob/main/Lab%204%20EDA_SQL.ipynb

Build an Interactive Map with Folium

Using Folium Maps, various elements such as markers, circles, lines, and marker clusters were employed:

- Markers represent specific points, such as launch sites.
- Circles highlight areas around specific coordinates, exemplified by the NASA Johnson Space Center.
- Marker clusters group events within each coordinate, showcasing launches at a launch site.
- Lines are employed to depict distances between two coordinates.

https://github.com/sanchez-luis/ibm-applied-data-science-capstone/blob/main/Lab%206%20Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

Launch Site Selection Dropdown:

- Implemented a dropdown list to facilitate the selection of launch sites.

Pie Chart Displaying Success Launches (All Sites/Specific Site):

- Introduced a pie chart to visualize the overall count of successful launches across all sites, and the distribution of success versus failure counts for a specific launch site if chosen.

Payload Mass Range Slider:

- Integrated a slider feature for selecting the payload mass range.

Scatter Chart Depicting Payload Mass vs. Success Rate for Various Booster Versions:

- Introduced a scatter chart to illustrate the relationship between payload mass and launch success, categorized by different booster versions.

https://github.com/sanchez-luis/ibm-applied-data-science-capstone/blob/main/Lab%207%20spacex_dash_app.py

Predictive Analysis (Classification)

- Four different classification models were compared: logistic regression, support vector machine, decision tree and k-nearest neighbors.

Data preparation and normalization → Hyperparameter optimization of the different models → Comparison of accuracies of the different models

https://github.com/sanchez-luis/ibm-applied-data-science-capstone/blob/main/Lab%208%20Machine%20Learning%20Prediction.ipynb

Results

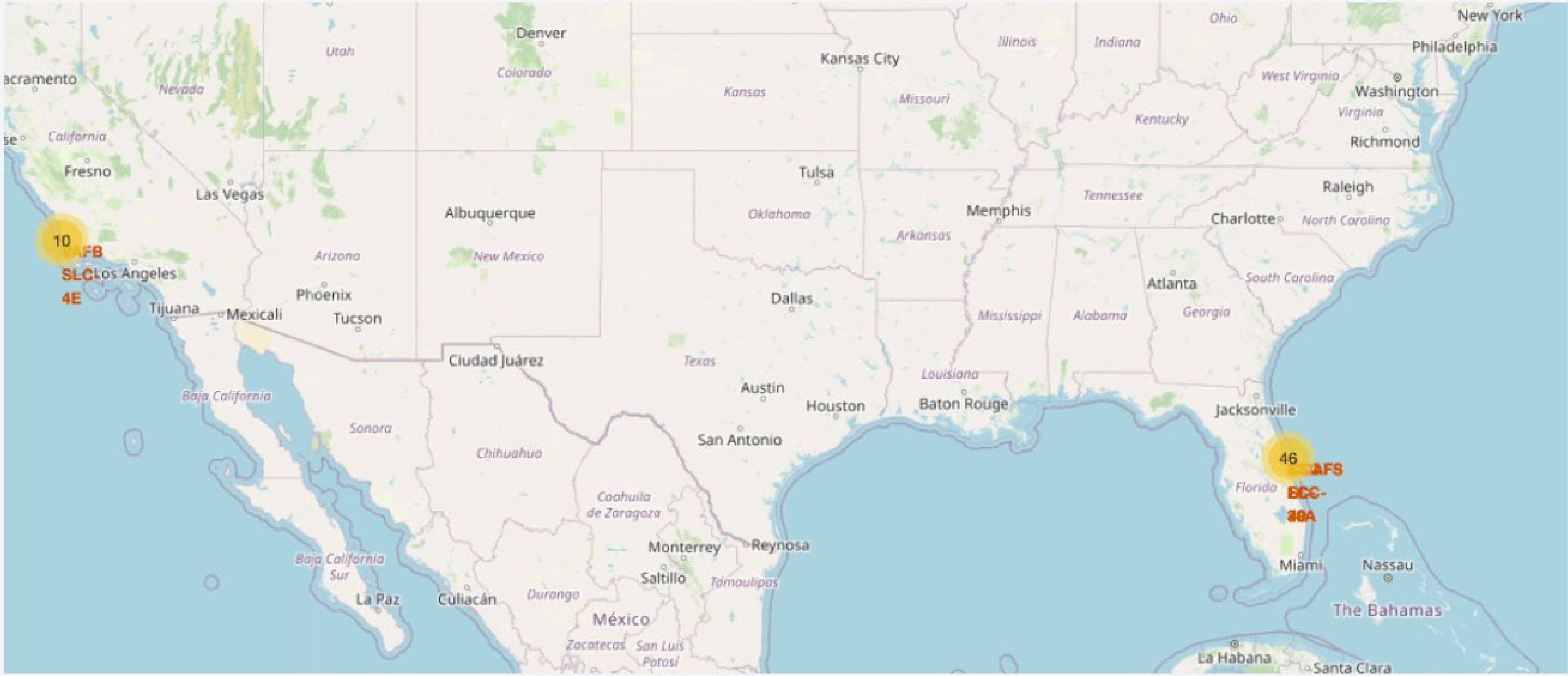
Exploratory data analysis results

- SpaceX operates from four distinct launch sites.
- The initial launches were conducted for SpaceX itself and NASA.
- The average payload of the F9 v1.1 booster is 2,928 kg.
- The first successful landing occurred in 2015, five years after the inaugural launch.
- Several Falcon 9 booster versions achieved successful landings on drone ships with payloads surpassing the average.
- Nearly 100% of mission outcomes were successful.
- In 2015, two booster versions, F9 v1.1 B1012 and F9 v1.1 B1015, experienced failures in landing on drone ships.
- The success rate of landing outcomes improved over the years.

Results

Interactive analytics demo

- Most of the launches happens at east cost



Results

Predictive Analysis results

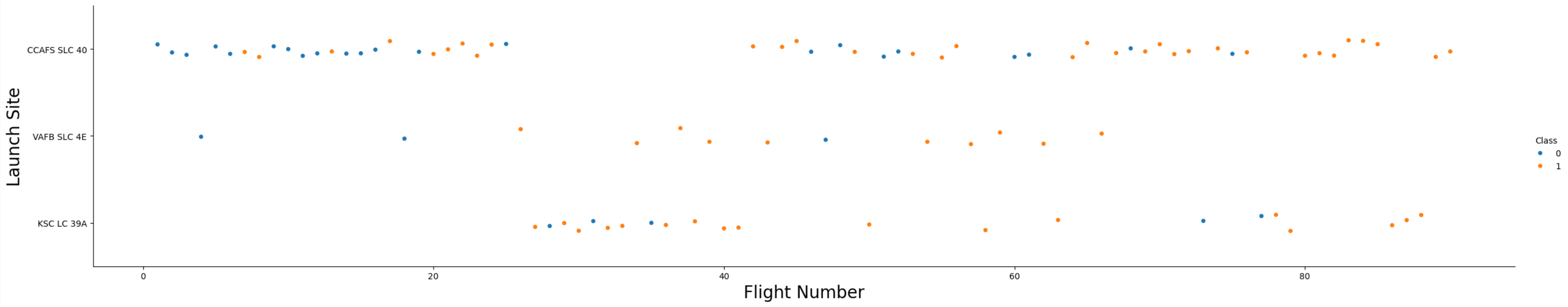
- Predictive Analysis showed that Decision Tree Classifier is the best model to predict successful landings

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.814286	0.819444
F1_Score	0.909091	0.916031	0.897638	0.900763
Accuracy	0.866667	0.877778	0.855556	0.855556

Section 2

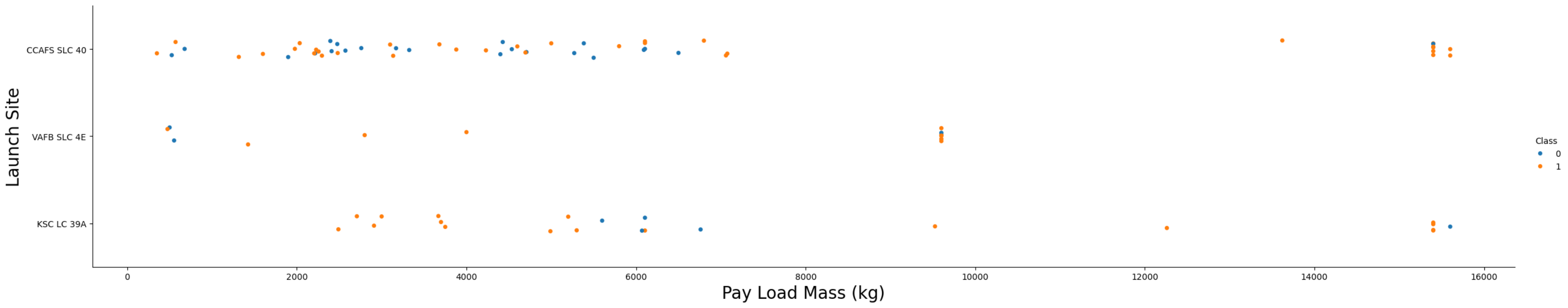
Insights drawn from EDA

Flight Number vs. Launch Site



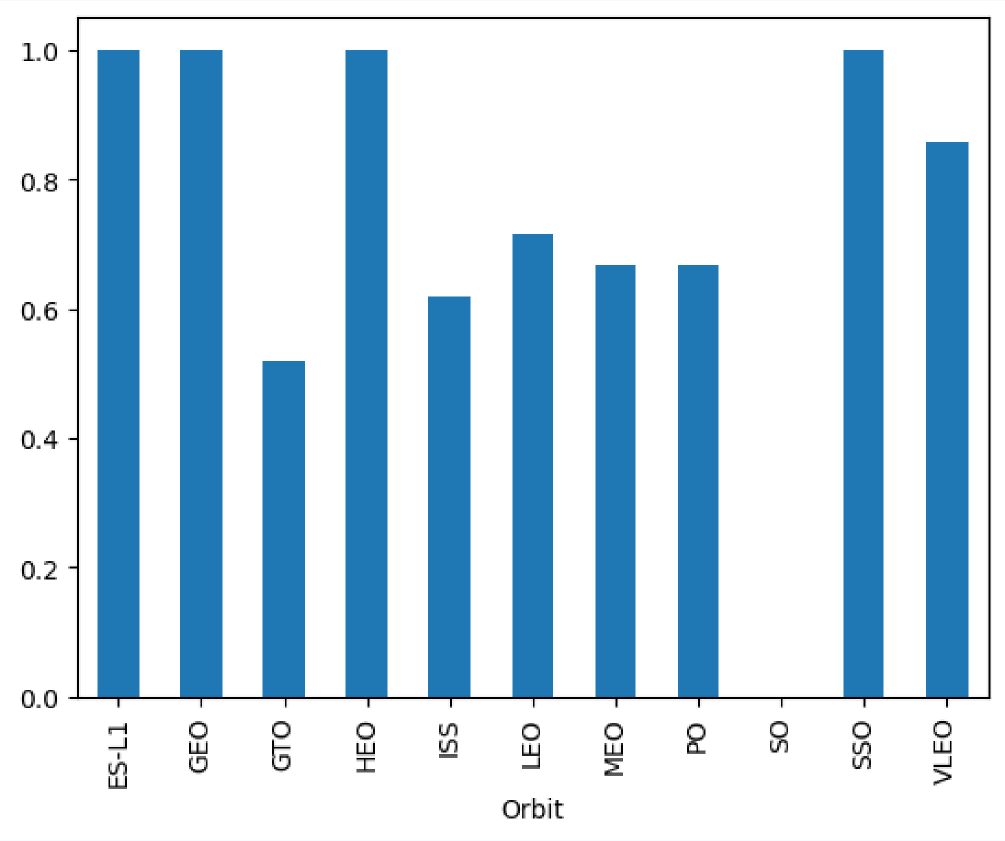
- Initial flights ended in failure, while the most recent flights achieved success.
- Approximately half of all launches occur at the CCAFS SLC 40 launch site.
- VAFB SLC 4E and KSC LC 39A exhibit higher success rates.
- There is an assumption that each successive launch tends to have a higher success rate.

Payload vs. Launch Site



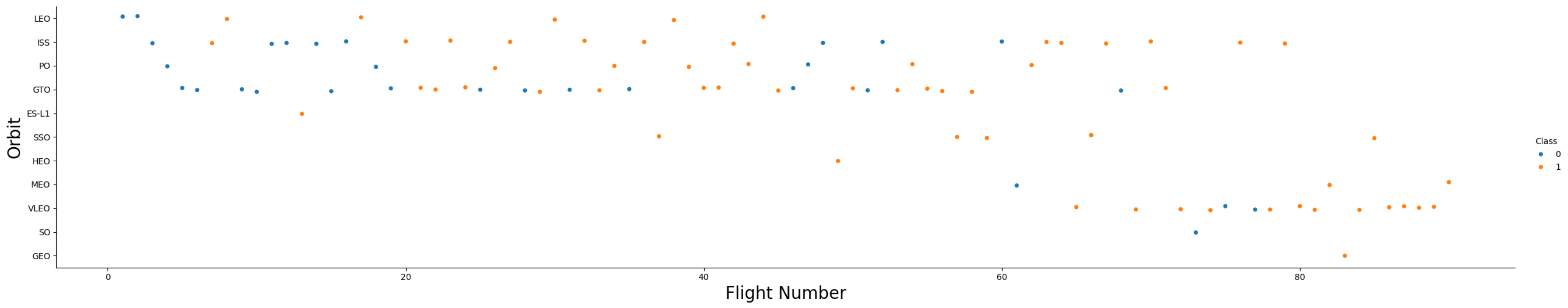
- Across every launch site, there is a positive correlation between payload mass and success rate.
- The majority of launches with a payload mass exceeding 7000 kg resulted in success.
- KSC LC 39A maintains a 100% success rate even for payload masses under 5500 kg.

Success Rate vs. Orbit Type



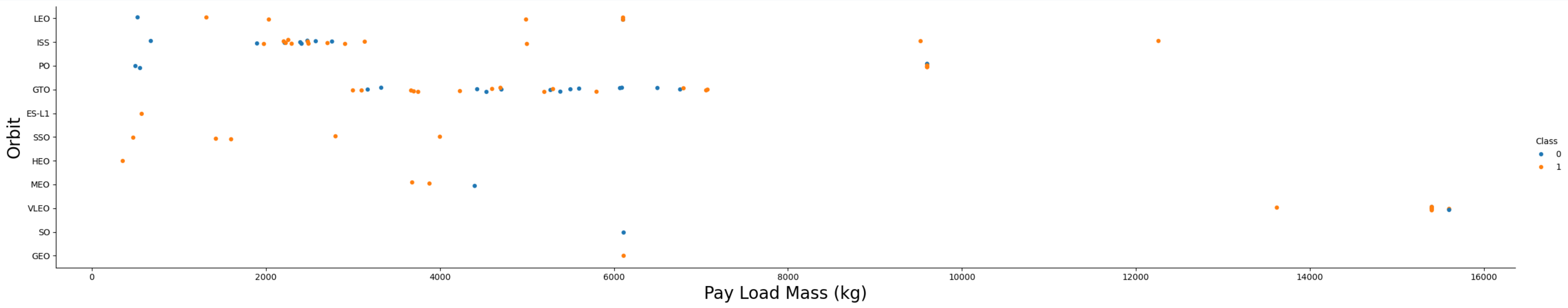
- Orbits achieving a 100% success rate include ES-L1, GEO, HEO, and SSO.
- Orbits with a 0% success rate pertain to SO.
- Orbits demonstrating success rates between 50% and 85% encompass GTO, ISS, LEO, MEO, and PO.

Flight Number vs. Orbit Type



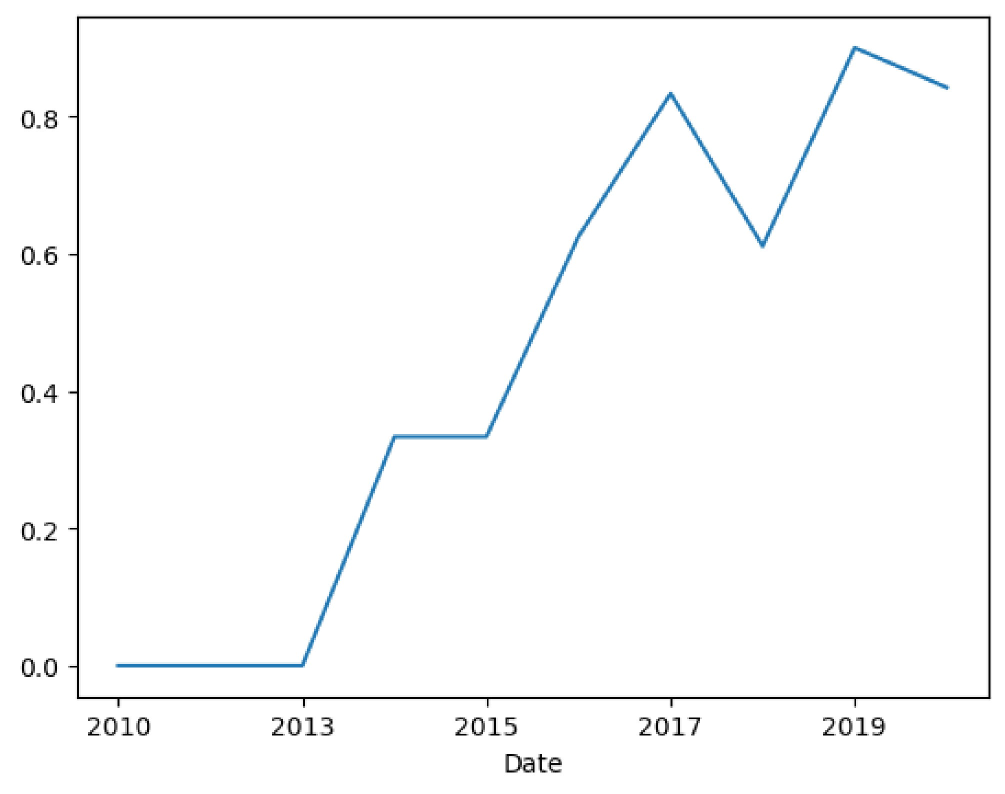
In the Low Earth Orbit (LEO), success appears to be correlated with the number of flights. Conversely, in the Geostationary Transfer Orbit (GTO), there seems to be no discernible relationship between flight number and success.

Payload vs. Orbit Type



Large payloads negatively impact Geostationary Transfer Orbit (GTO) missions, while they have a positive influence on Geostationary Earth Orbit (GEO) and Polar Low Earth Orbit (LEO) missions such as the International Space Station (ISS).

Launch Success Yearly Trend



The success rate has continuously risen from 2013 until 2020.

All Launch Site Names

sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;

Four launch sites:

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

They are obtained by displaying the names of the unique launch sites in the space mission

Launch Site Names Begin with 'CCA'

sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_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

Displaying 5 records where launch sites begin with the string ‘CCA’

Total Payload Mass

```
sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE PAYLOAD LIKE '%CRS%';
```

TOTAL_PAYLOAD
111268

Displaying the total payload mass carried by boosters launched by NASA (CRS).

Average Payload Mass by F9 v1.1

sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';

AVG_PAYLOAD
2928.4

Displaying average payload mass carried by booster version F9 v1.1.

First Successful Ground Landing Date

```
sql SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (ground pad)';
```

FIRST_SUCCESS_GP
2015-12-22

Listing the date when the first successful landing outcome in ground pad was achieved

Successful Drone Ship Landing with Payload between 4000 and 6000

sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND LANDING_OUTCOME = 'Success (drone ship)';

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

Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.

Total Number of Successful and Failure Mission Outcomes

sql SELECT MISSION_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL GROUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME;

Mission_Outcome	**QTY**
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Listing the total number of successful and failure mission outcomes.

Boosters Carried Maximum Payload

```sql
sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ =
(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL) ORDER BY BOOSTER_VERSION;
```

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

Listing the names of the booster versions which have carried the maximum payload mass.

2015 Launch Records

```
sql SELECT substr(Date, 6, 2) as Month, LANDING_OUTCOME, BOOSTER_VERSION, LAUNCH_SITE
FROM SPACEXTBL
WHERE LANDING_OUTCOME = 'Failure (drone ship)' AND substr(Date, 0, 5) = '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

Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015.

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

sql SELECT LANDING_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING_OUTCOME ORDER BY QTY DESC;

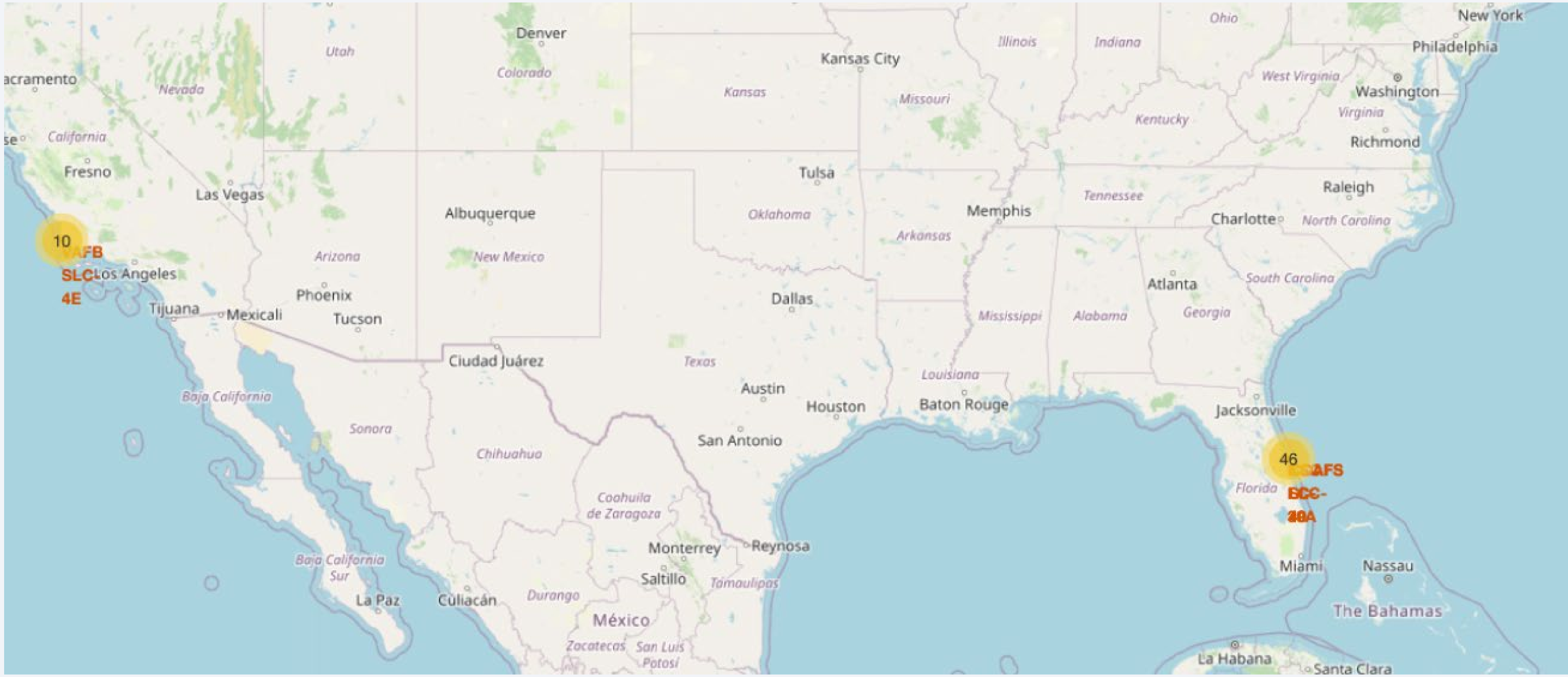
Landing_Outcome	**QTY**
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Ranking the count of landing outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

Section 3

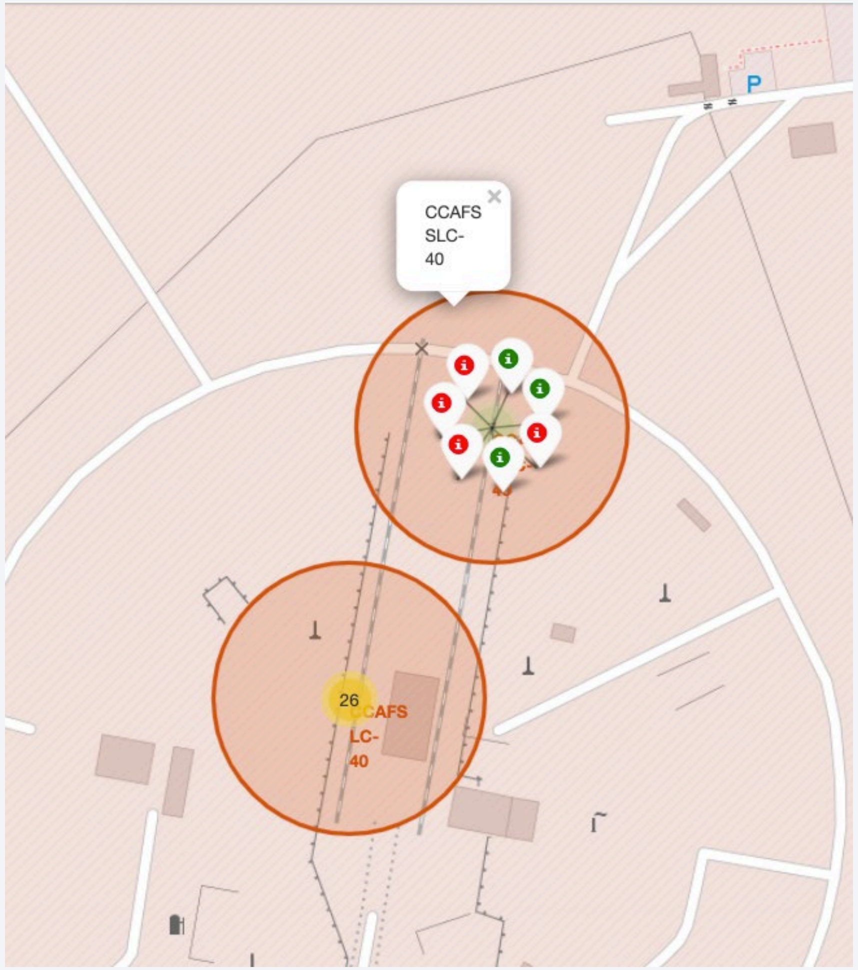
Launch Sites Proximities Analysis

Launch sites



- Most of Launch sites are in proximity to the Equator line.
- All launch sites are close to the coast.

Launch records on the map

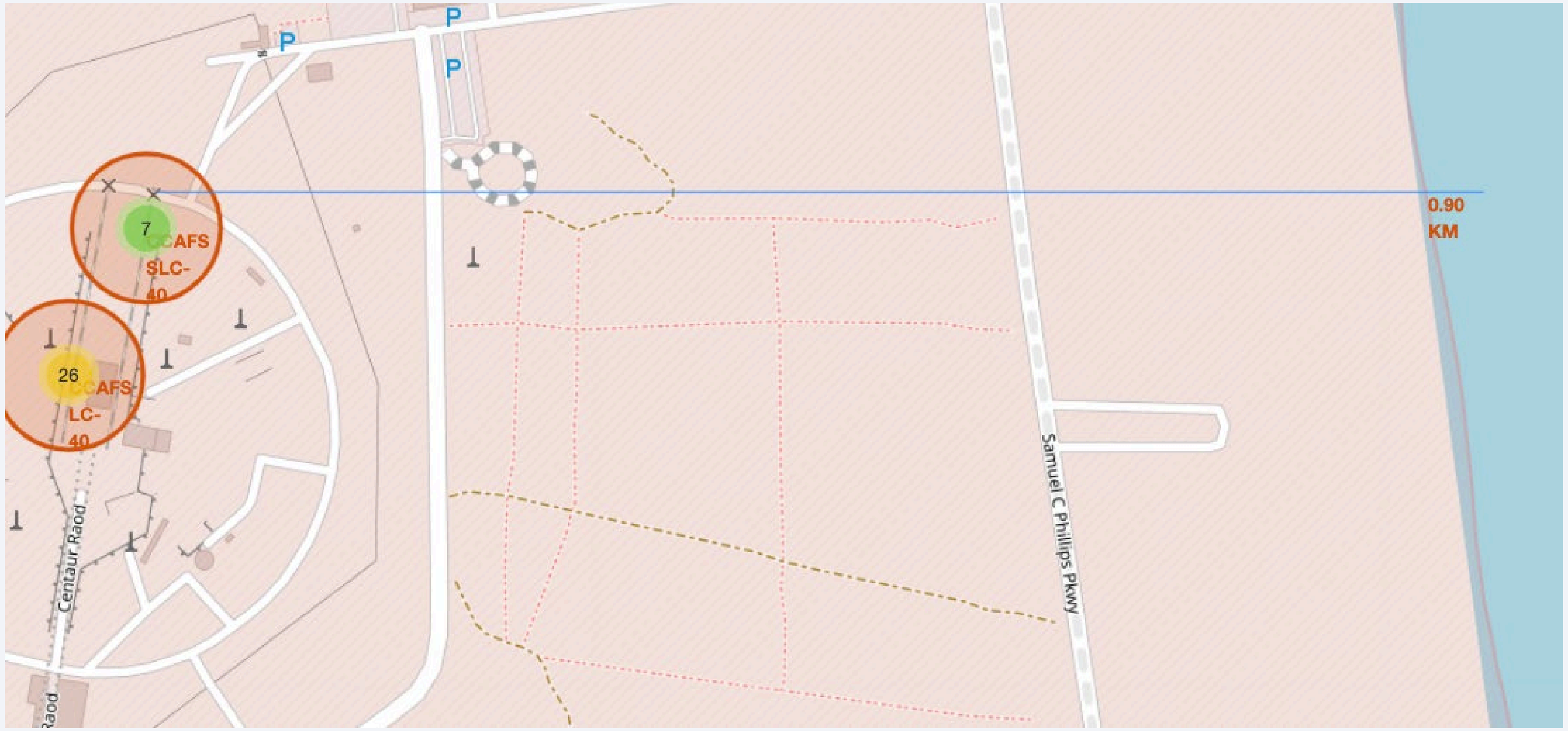


Example of CCAFS SLC-40 launch site

Green Marker = Successful Launch
Red Marker = Failed Launch

Distance from launch site to proximities

Example of CCAFS SLC-40 launch site

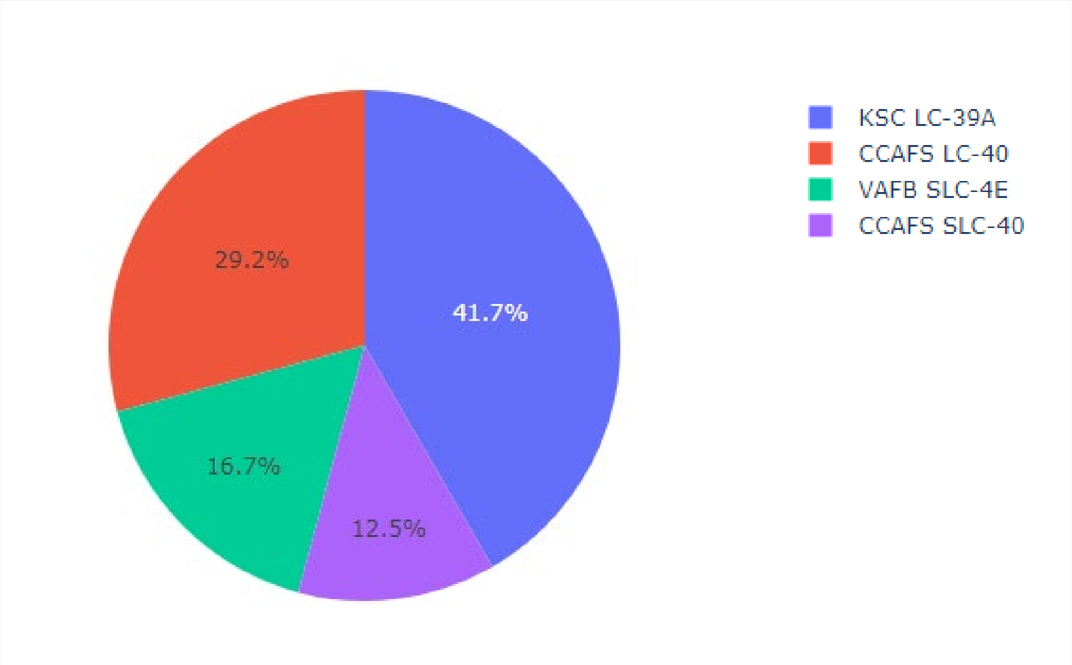


Very close to coastline (0.90 km)

Section 4

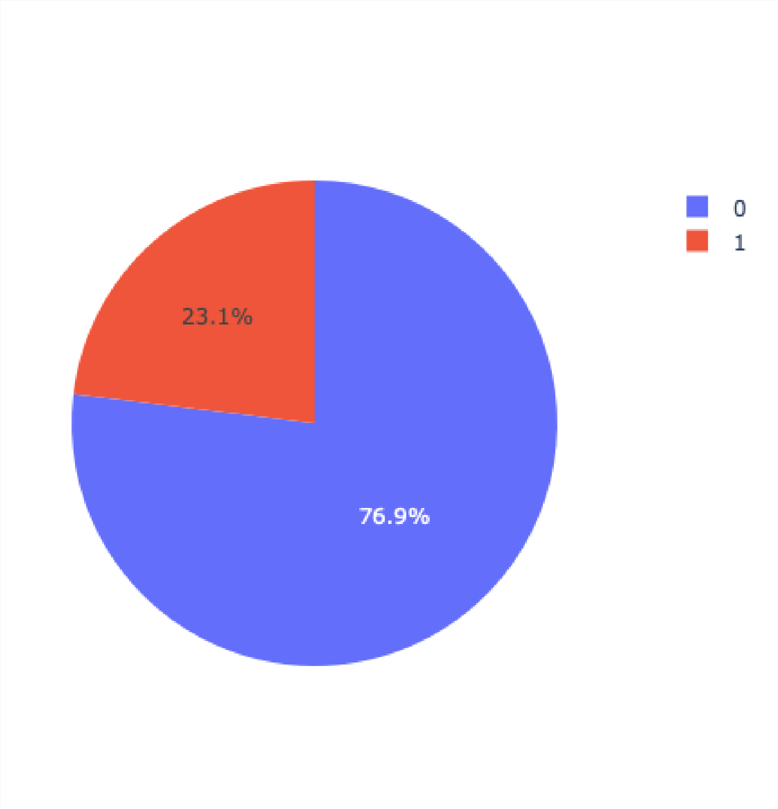
Build a Dashboard with Plotly Dash

Total success launches by site



The chart shows that from all the sites, KSC LC-39A has the most successful launches.

Total success launches for KSC LC-39A



KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings.

Payload vs launch outcome

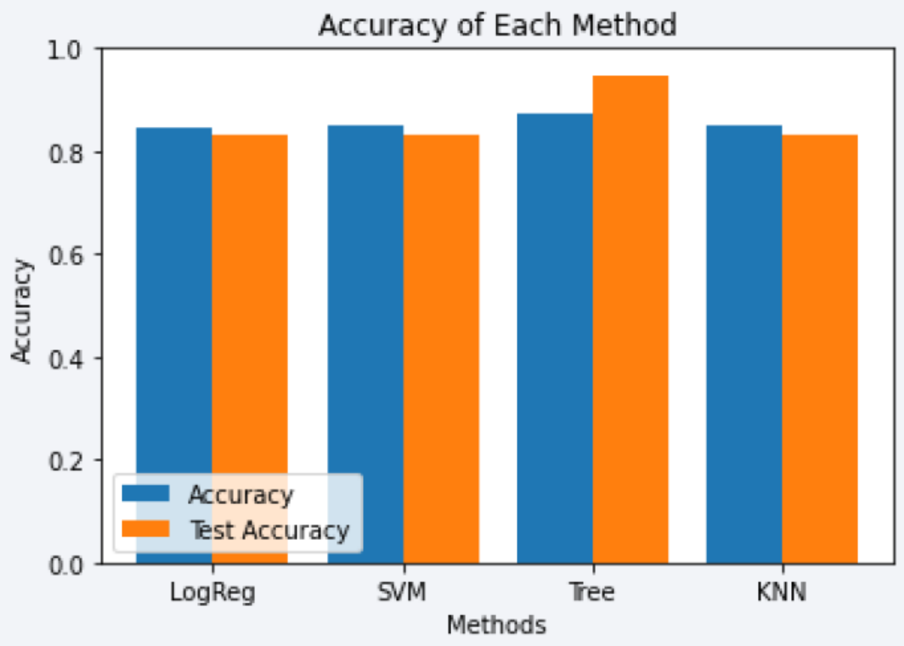
The chart show that payloads between 2000 and 5000 kg have the highest success rate.

Payloads above 5000 kg present a very low success rate

Section 5

Predictive Analysis (Classification)

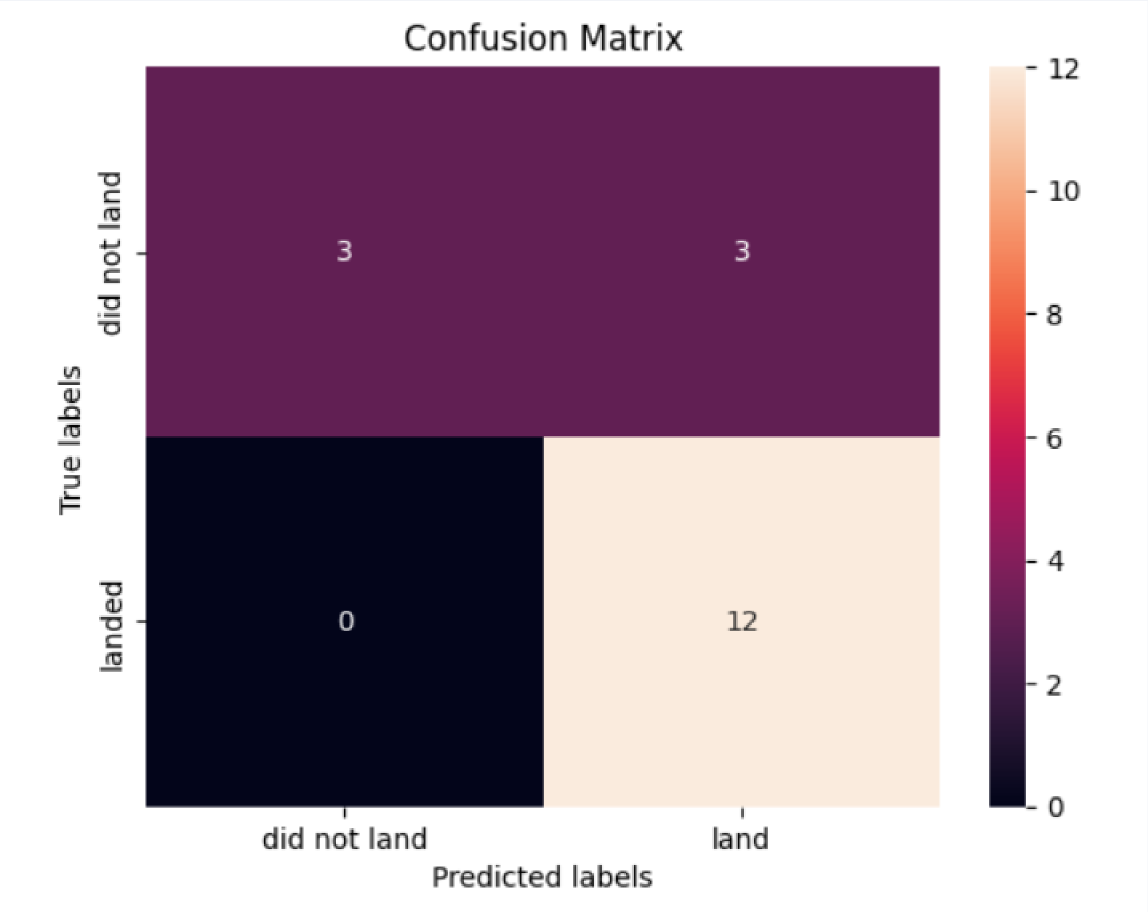
Classification Accuracy



	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.814286	0.819444
F1_Score	0.909091	0.916031	0.897638	0.900763
Accuracy	0.866667	0.877778	0.855556	0.855556

- Four classification models underwent testing, and their respective accuracies are presented in the bar plot.
- The Decision Tree classifier emerged as the model with the highest classification accuracy, surpassing 87%.

Confusion Matrix



- Analyzing the confusion matrix reveals the ability of logistic regression to differentiate between various classes.
- The predominant issue identified is the occurrence of false positives.

Conclusions

- Various data sources were analized, refining conclusions throughout the process.
- The optimal launch site identified is KSC LC 39A.
- Launches with payloads exceeding 7,000 kg appear to carry lower risk.
- While the majority of mission outcomes are successful, the success of landing outcomes seems to improve over time, aligning with the evolution of processes and rockets.
- The Decision Tree Classifier emerges as a viable tool for predicting successful landings and potentially enhancing profits.

Appendix

- Git-hub repository:

https://github.com/sanchez-luis/ibm-applied-data-science-capstone/tree/main

Thank you!