Winning Space Race with Data Science

Author: Aleksander Melonashi

Outline

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

Executive Summary

- **Following methodologies were used to analyze data:**
  - Data collection from SpaceX REST API and web scraping
  - Exploratory Data Analysis (Data cleaning, Data exploration & Data Visualization)
  - Machine Learning model (feature engineering, data scaling, hyperparameter tuning)

- **Summary of all results:**
  - Required data was available on the internet
  - Exploratory data analysis helped understand key patterns in data along with recognizing important features for machine learning model development
  - Machine learning models enabled us to predict landing outcome based of existing data

Introduction

- Objective: To evaluate if SpaceY could compete against SpaceX
- Questions we are looking answers for:
  - Which type of rockets have higher landing success rates?
  - What are the ideal parameters of rockets for different requirements?
  - What are the ideal launch sites with higher success rates?
  - What are the optimal landing methods?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from 1 source:

    - SpaceX REST API

- Perform data wrangling
  - Null values were dealt with and insignificant features were removed. After summarization of features, a new categorical target variable for landing outcome was created for further Machine Learning model development.

Methodology

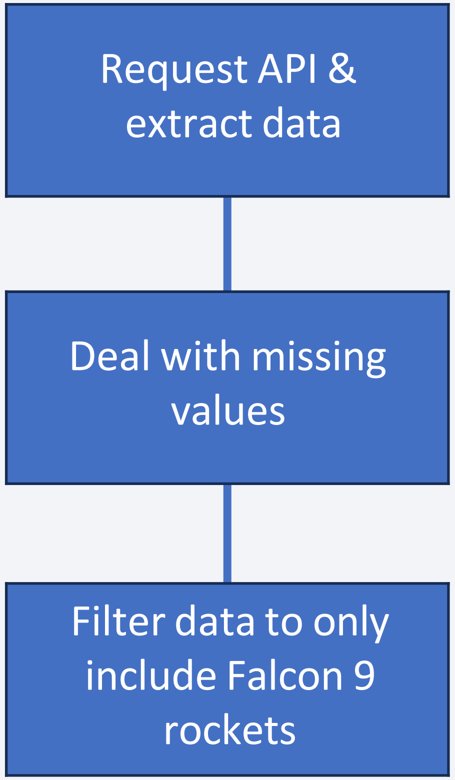
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
  - After data was processed, it was scaled and split into 2 parts (Features & Target) which was fed to 4 major classification models. All models were optimized with Hyperparameter tuning. They were properly tested and evaluated.

Data Collection

- Data sets were collected from SpaceX API and Wikipedia.
  - SpaceX API: https://api.spacexdata.com/v4/rockets/

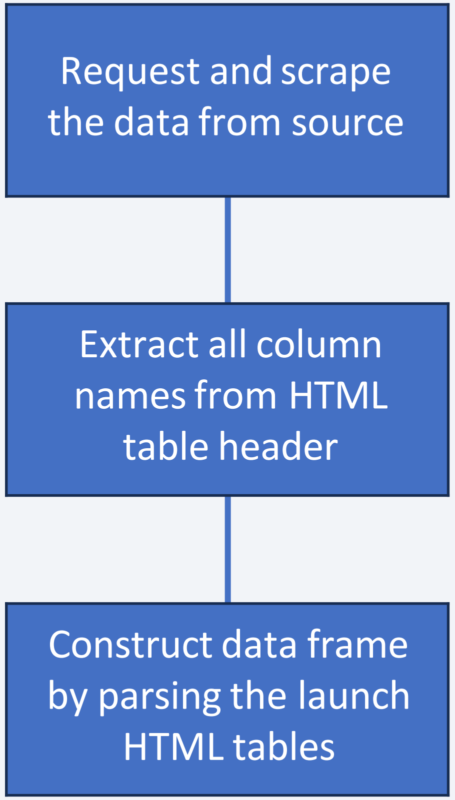
Data Collection – SpaceX API

- We use a public API that enables us to extract data regarding Falcon rocket launches.
- Flowchart illustrates how API was used, continued by further data preprocessing methods.



Data Collection - Scraping

- Data was also collected from Wikipedia through Web Scraping.
- This was enabled by Python libraries like requests and beautiful soup.

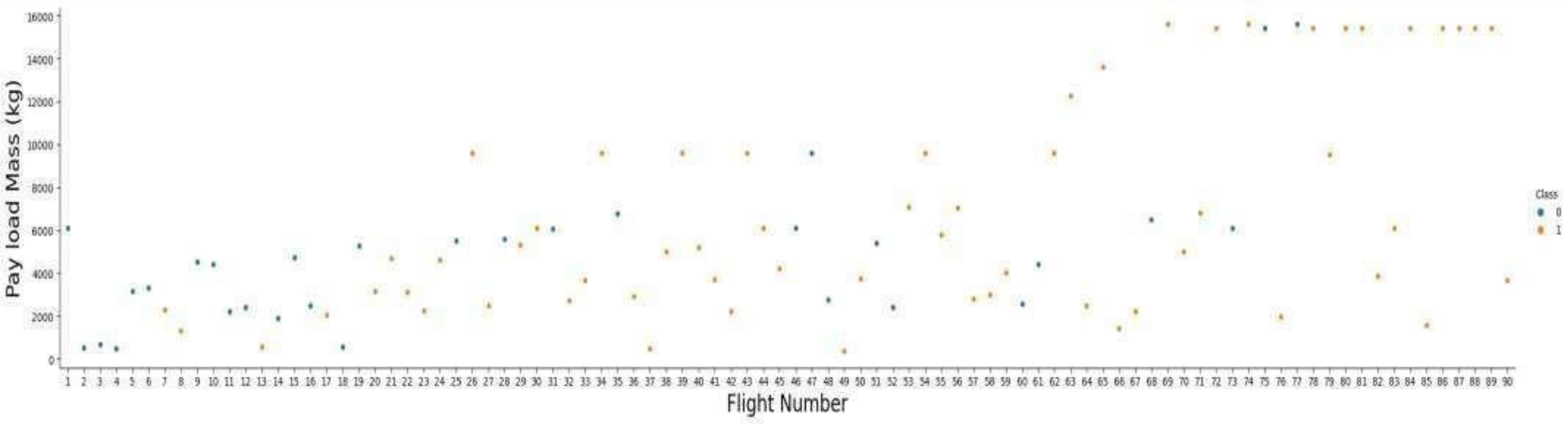


Data Wrangling

- A proper data frame is constructed from the data collected.
- Exploratory data analysis (EDA) is performed on the data frame.
- After summarization, a new ‘landing’ outcome label was created from existing outcome column.

EDA with Data Visualization

- Different scatterplots, bar graphs and line graphs were used to explore data and visualize the relationship between different features.



EDA with SQL

- The following SQL queries were performed:
  - Names of the unique launch sites in the space mission
  - Top 5 launch sites whose name begin with the string 'CCA’
  - Total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1
  - Date when the first successful landing outcome in ground pad was achieved
  - Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg
  - Total number of successful and failure mission outcomes
  - Names of the booster versions which have carried the maximum payload mass;
  - Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.

Build an Interactive Map with Folium

- Markers, circles, lines and marker clusters were used with Maps
  - Markers indicate different launch sites
  - Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Centre
  - Marker clusters indicate groups of events in each coordinate, like launches in a launch site
  - Lines are used to indicate distances between two relevant coordinates.

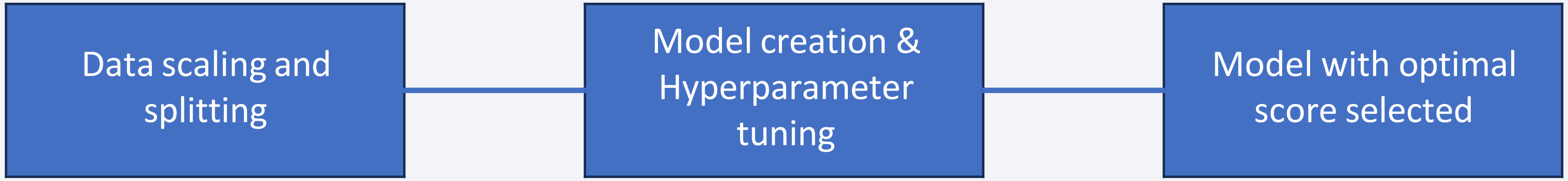
Build a Dashboard with Plotly Dash

- The following graphs and plots were used to visualize data
  - Percentage of launches by site
  - Payload range

- This combination enabled us to analyse the relation between payloads and launch sites, helping to identify the best place to launch according to payloads.

Predictive Analysis (Classification)

- Four models on 4 namely: Logistic regression, Decision tree, Support vector machine and K-nearest neighbors were trained and tested.
- All are trained and tested over different combinations of parameters. Optimal parameters are chosen for each model and the model with the highest accuracy is finally chosen



Results

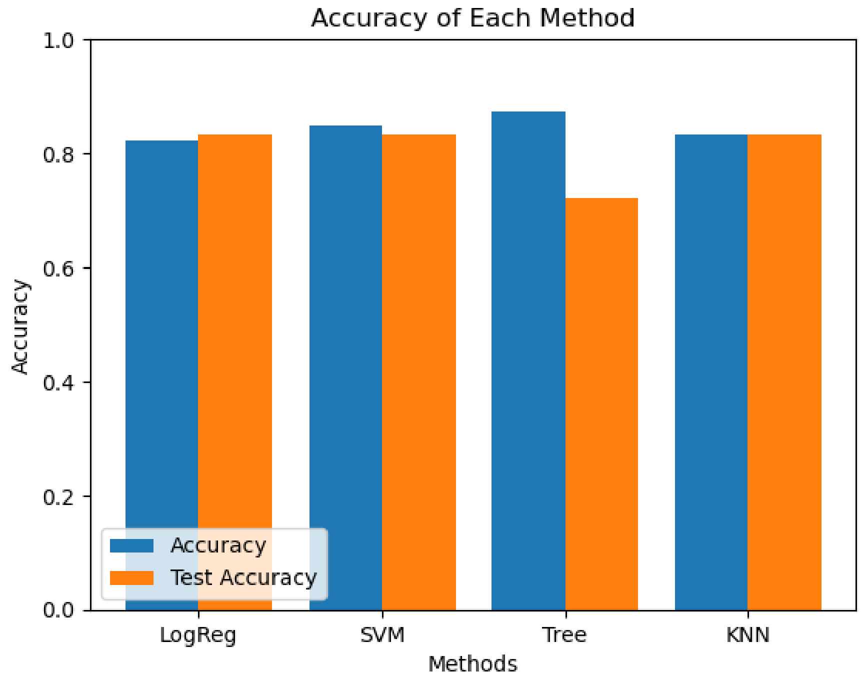
- Space X uses 4 different launch sites
- Most amount of payloads have been sent to Geostationary transfer orbit (GTO)
- The first successful landing outcome occurred in 2015, five years after the first launch
- Drone ships are most extensively used for landing rockets.
- Many Falcon 9 booster versions having payload above the average were successful at landing in drone ships
- All 5 landing attempts of flights that sent payloads to SSO (Sun-synchronous orbit) were successful.
- Flights with top 3 largest payloads have gone to VLEO (Very low earth orbit) and a majority of recent Falcon 9 flights also delivered payloads in the VLEO region
- Cape Canaveral launch site (CCAFS SLC-40) is the most used extensively used launch site by SpaceX.
- Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015;
- Landing success rates have consistently improved over the years.

Results

- All launch sites are in safe places (near sea) and have a good logistic infrastructure around.
- Most launches occur at east coast launch sites.

Results

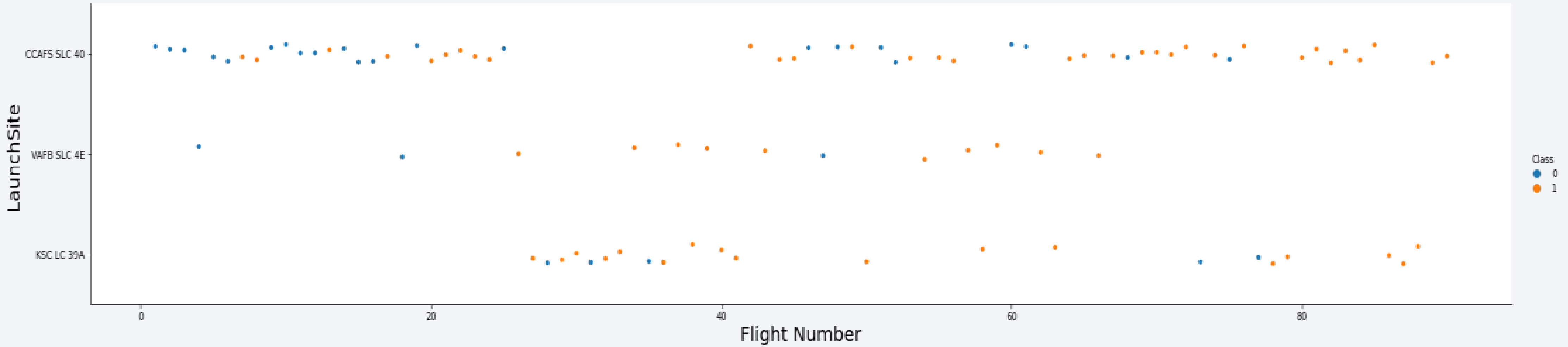
- Support vector machine works best to predict successful landings overall having an accuracy of around 84% and similar test data accuracy.



Section 2

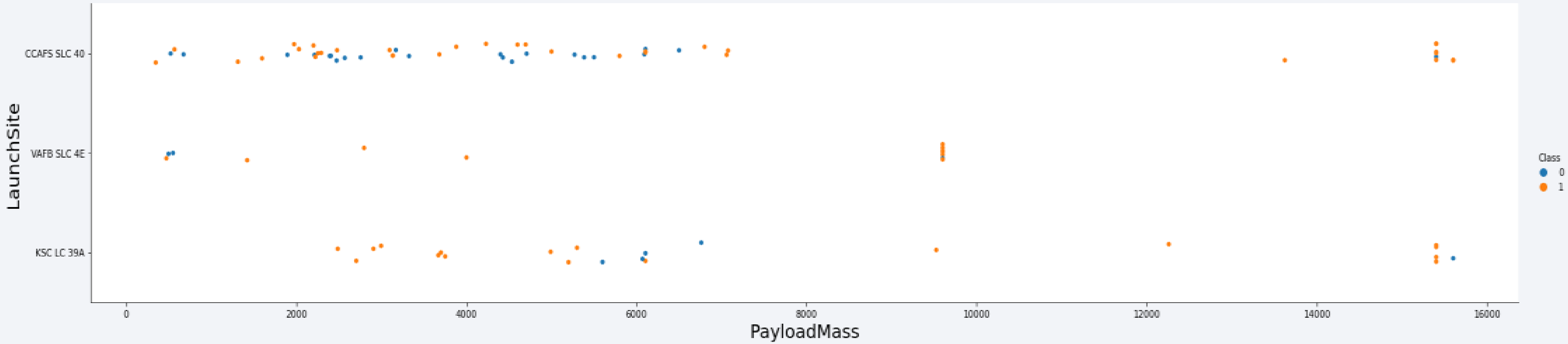
Insights drawn from EDA

Flight Number vs. Launch Site



- From the plot above, we can see that CCAF5SLC-40 is the most used site with also the most success
- In second place VAFB SLC4E and third place KSCLC 39A
- We can further observe that the general successrate improved over time.

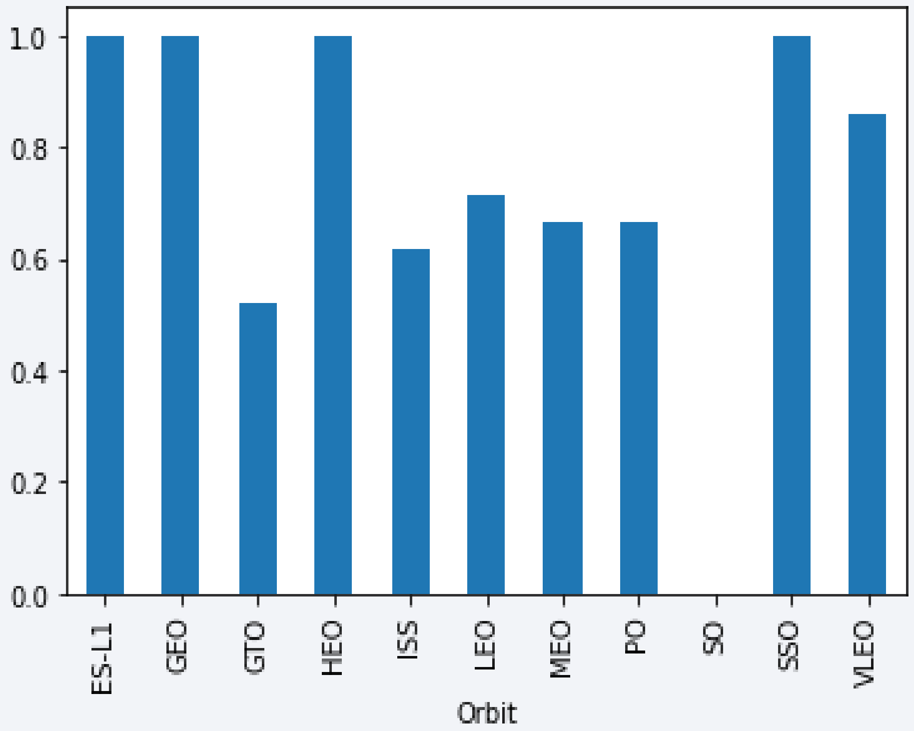
Payload vs. Launch Site



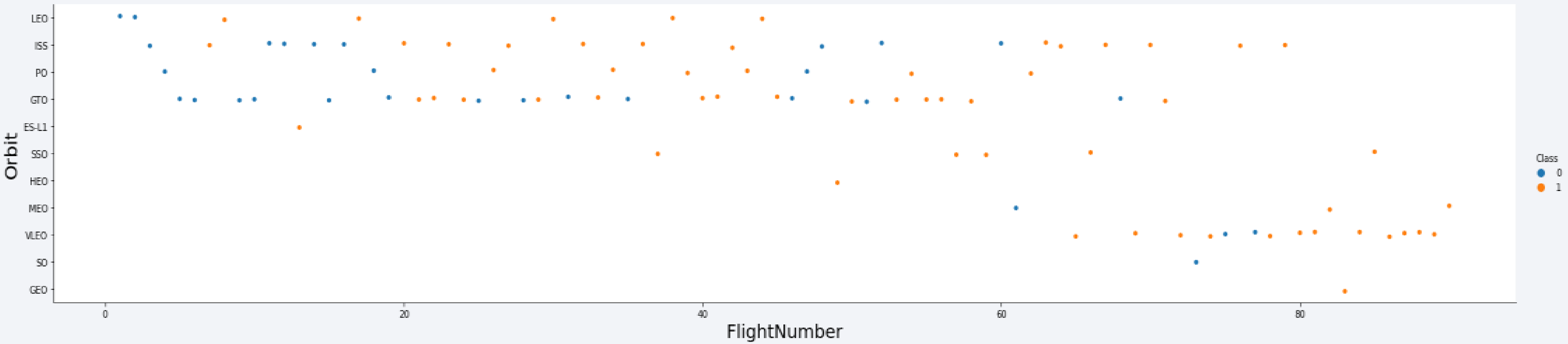
- Payloads over 9,000kg have great success rates
- Payloads over 12,000kg have been sent only from CCAFSSLC40 and KSCLC 39A launch sites.

Success Rate vs. Orbit Type

- Thehighest successrates in orbits:
  - ES-L1
  - GEO
  - HEO
  - …

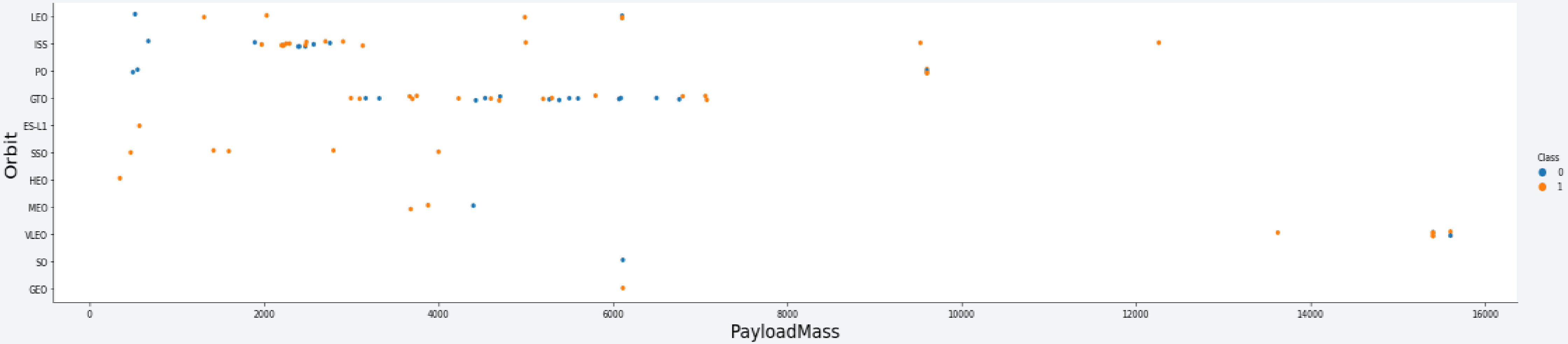


Flight Number vs. Orbit Type



- Successrate improved over time in all orbits
- VLEOorbit seemsto be a new business opportunity, due to recent increase in its frequency.

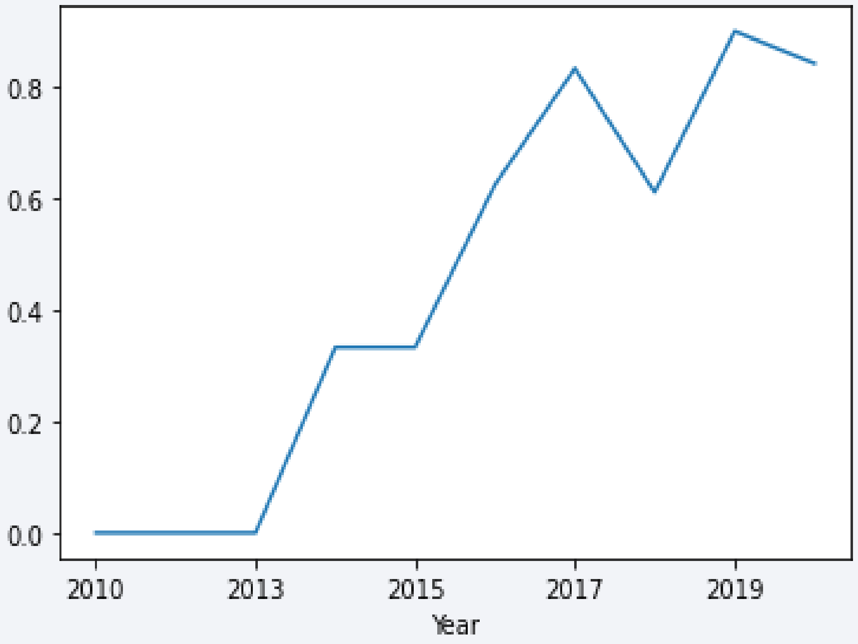
Payload vs. Orbit Type



- There is no relation between payload and successrate to orbit GTO
- The largest payloads have been sent to VLEO
- Thereare very few launches to the orbits SOand GEO.

Launch Success Yearly Trend

- There was no success (or attemps) of landings in the first three years.
- Successrate started increasing in 2013 and kept until recent times.



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 above are obtained by selecting unique occurrences of “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	**CCA**FS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	**CCA**FS 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	**CCA**FS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	**CCA**FS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	**CCA**FS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attemp

Total Payload Mass

- Total payload carried by boosters from NASA:

Total Payload (kg)
111.268

- We arrived at the number above by summing all payloads whose codes contain ‘CRS’,which corresponds to NASA.

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1:

Avg Payload (kg)
2.928

- We came to this number by filtering data by the booster version and calculating average payload mass.

First Successful Ground Landing Date

- First successful landing outcome on ground pad:

Min Date
2015-12-22

- By filtering data based on landing outcome (on ground pad) and getting the minimum value for date we identified the first occurrence.

Successful Drone Ship Landing with Payload between 4000 and 6000

- Boosters which have successfully landed on drone ship and had payload mass greater than 4000 and less than 6000:

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

- Selecting distinct booster versions based on the filters above, we arrived at these results.

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

- Grouping mission outcomes and counting records for each group led us to the result above.

Boosters Carried Maximum Payload

- Boosters which have carried the maximum payload mass

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

- These are the boosters which have carried the maximum payload mass registered in the dataset.

2015 Launch Records

- Failed landing outcomes in drone ships, their booster versions, and launch site names for the 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 dates 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

Section 3

Launch Sites Proximities Analysis

All launch sites

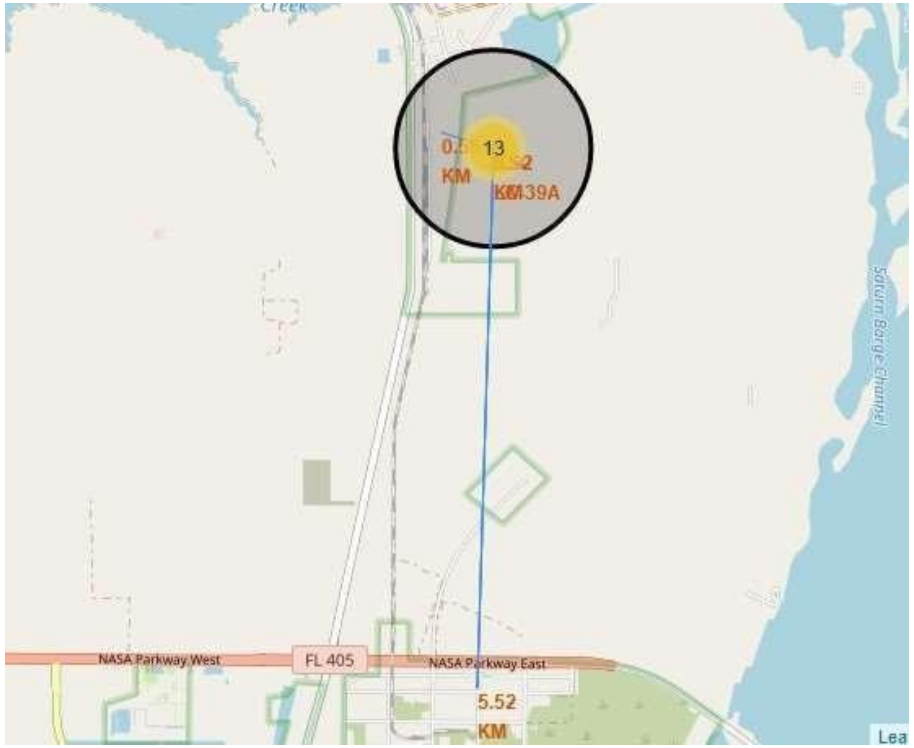
- Launch sites are present on the eastern and western coasts with most launches occurring at the eastern coast.

Launch outcomes by Site

- Example of KSCLC-39A launch site launch outcomes

- Green markers indicate success and red ones indicate failure.

Logistics and safety

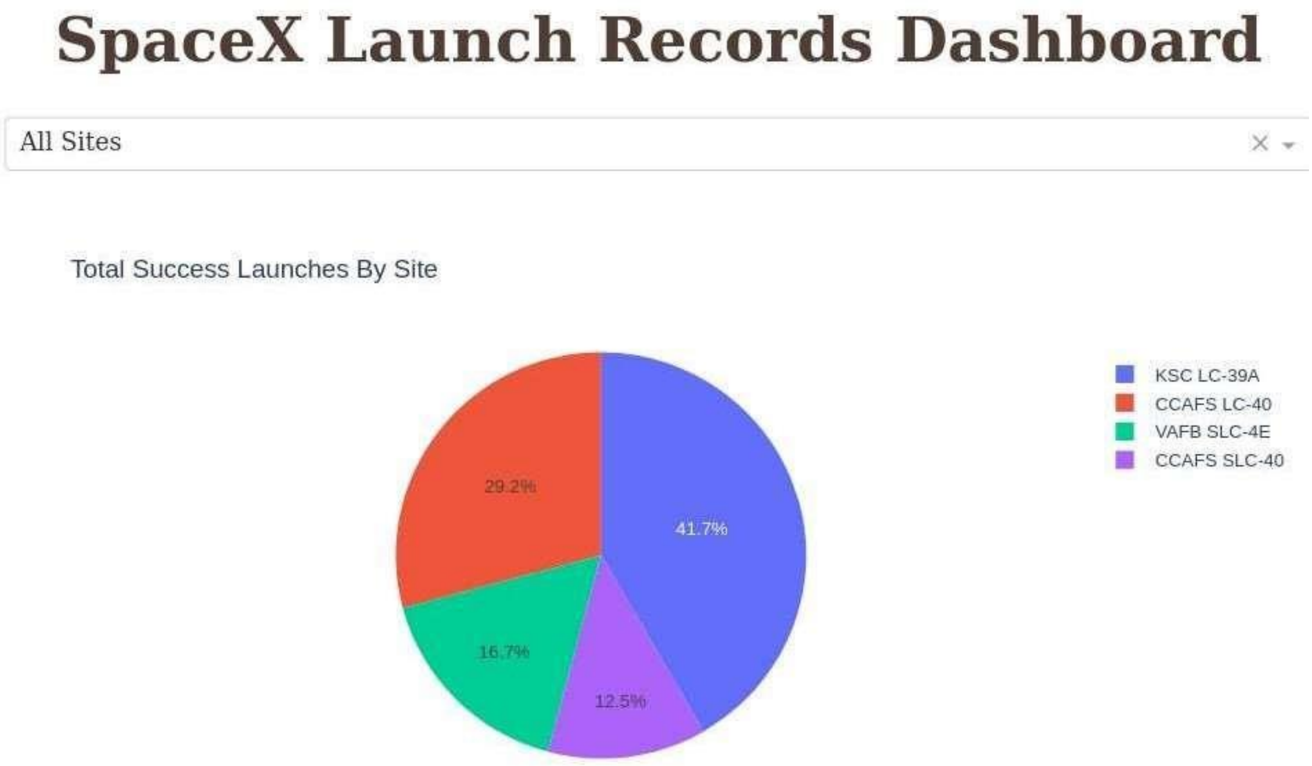


- Launch site KSCLC-39A has good logistical aspects, railroads and roads are present while being relatively far from any population centers.

Section 4

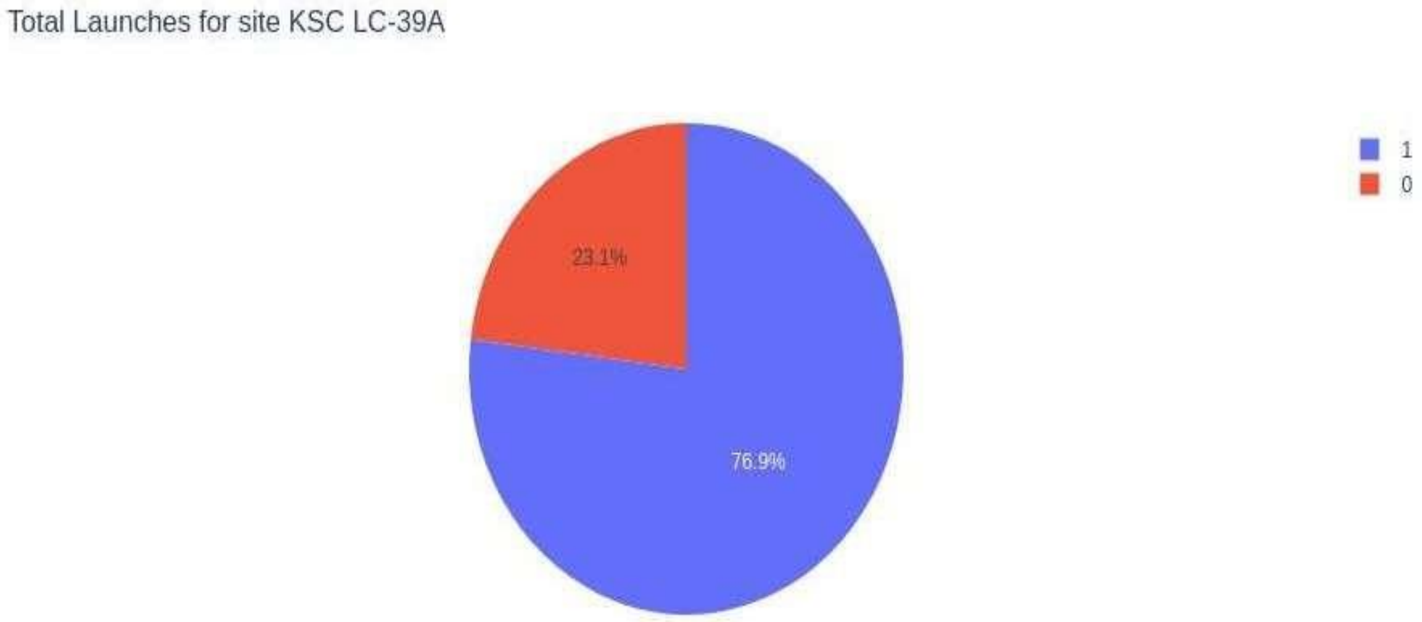
Build a Dashboard with Plotly Dash

Successful launches by Site



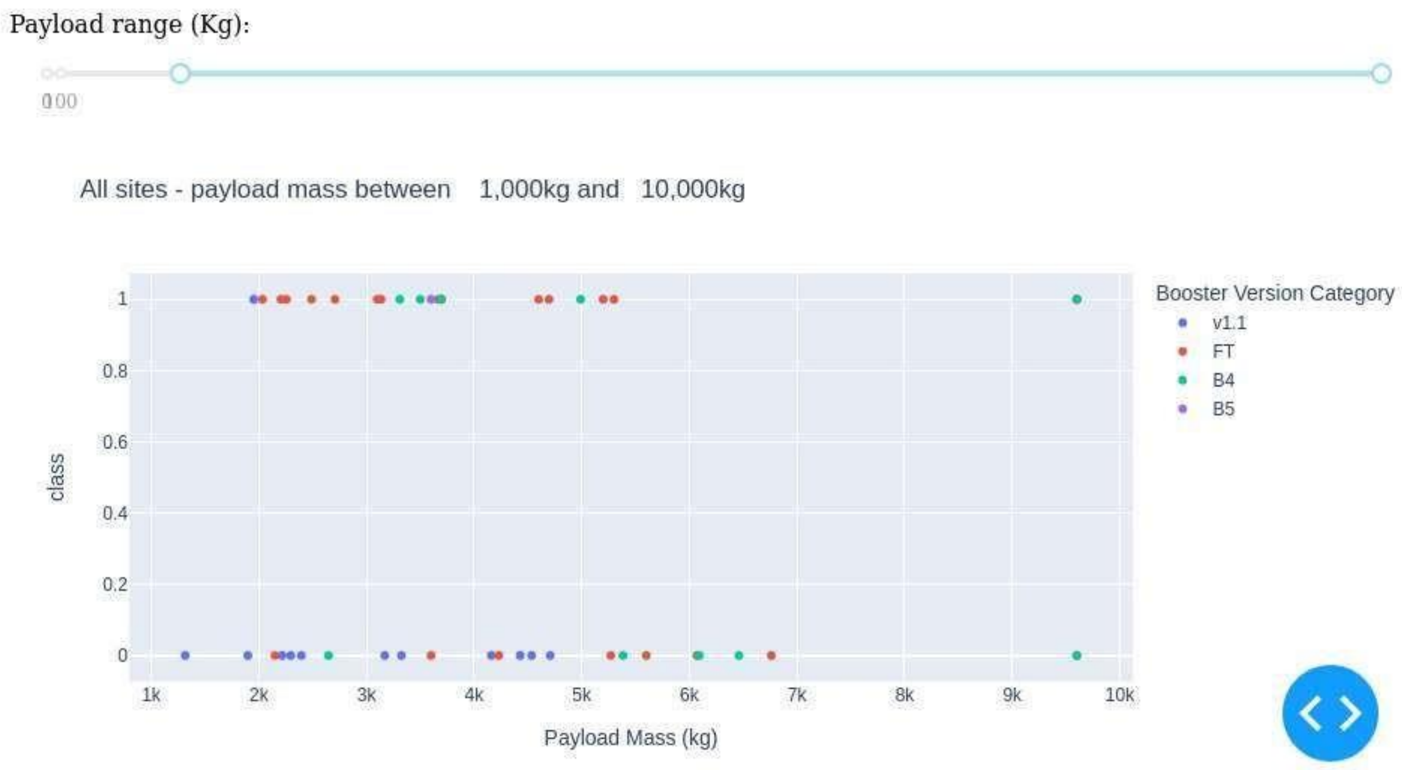
- Thesite from where launches are done is a very important factor for success of different missions.

Launch success ratio for KSC LC-39A



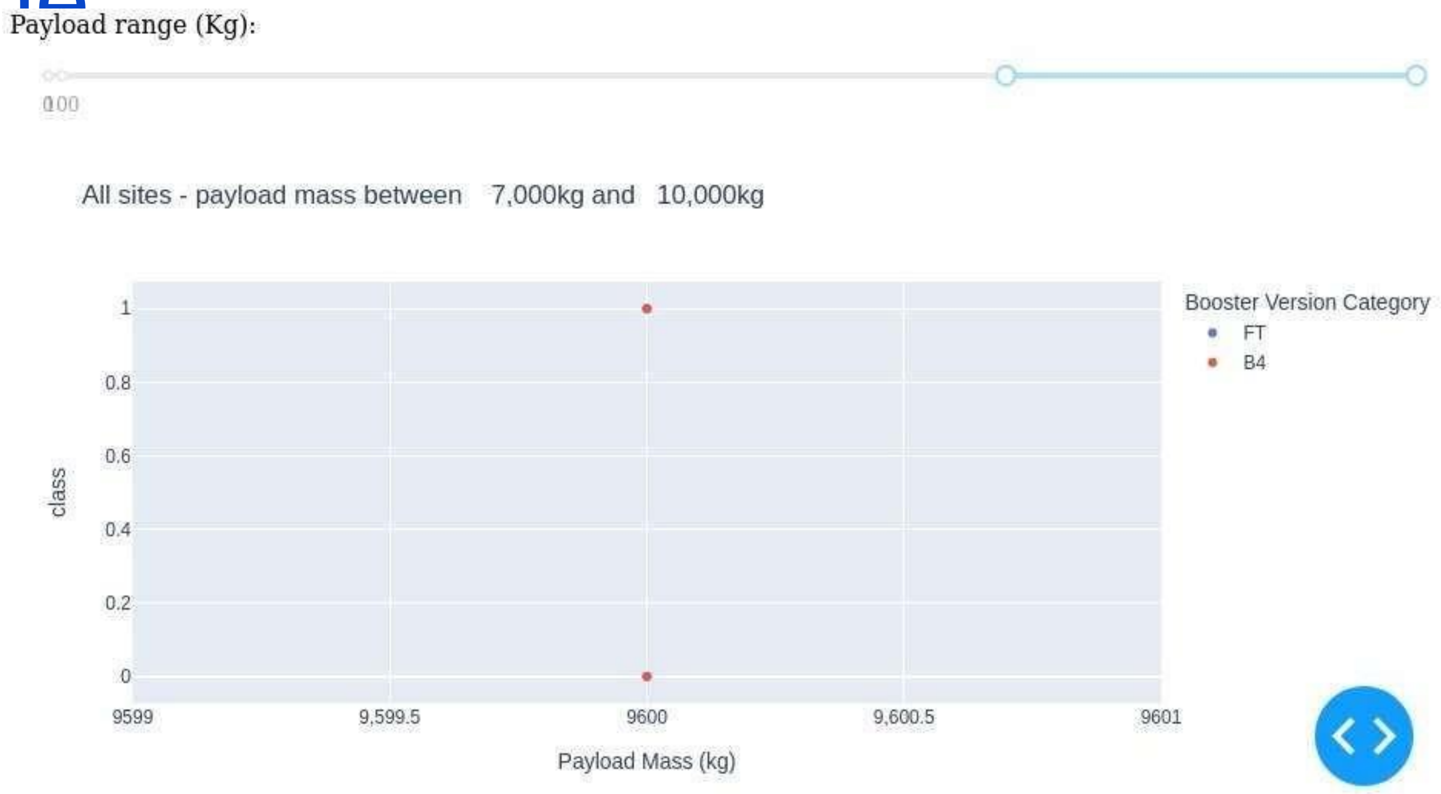
- Almost 77% of launches are successful in this site.

Payload vs Launch outcome



- Payloads under 6,000kg and FTboosters are the most successful combination.

Payload vs Launch outcome



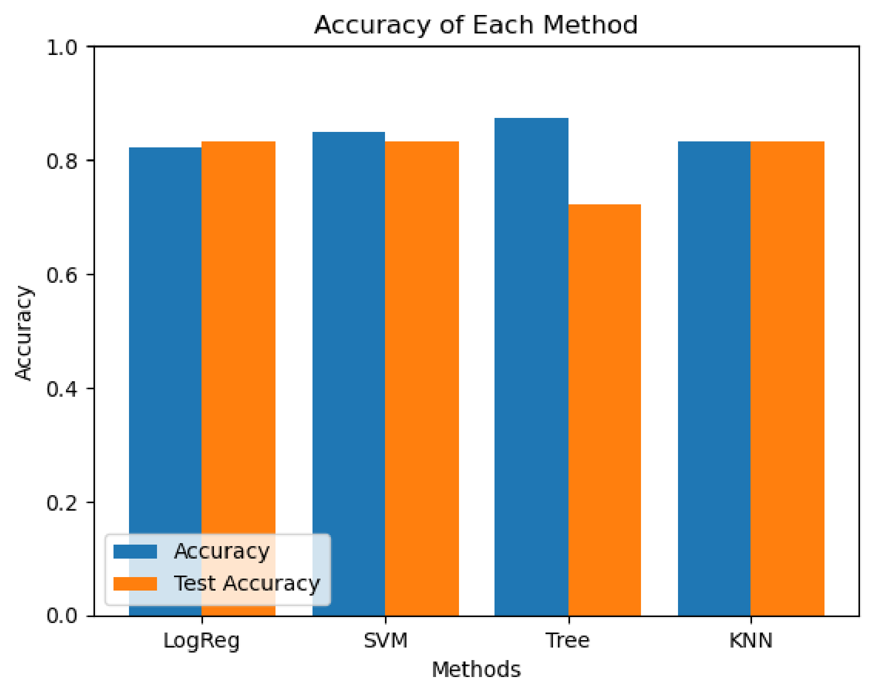
- There's not enough data to estimate risk of launches over 7,000kg

Section 5

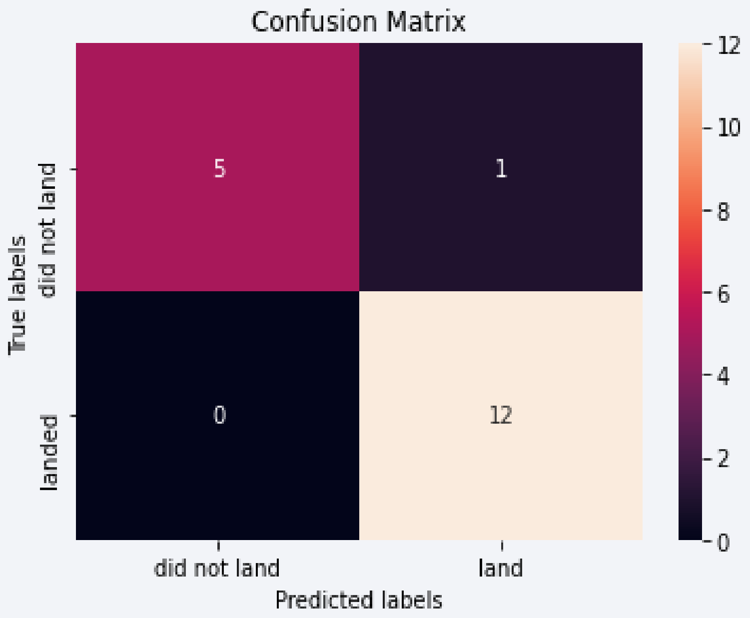
Predictive Analysis (Classification)

Classification Accuracy

- Four classification models were tested.
- Support Vector Machine provides the highest accuracy with test and overall accuracies of around 84%.



Confusion Matrix



- Confusion matrix of Decision Tree Classifier illustrates high accuracy by showing the higher values of true positive and true negative compared to the false ones.

Conclusions

- Data was extracted from multiple sources and properly analyzed
- Launches with payloads above 7,000 kgs have higher success rates.
- Biggest payloads usually go to VLEO.
- Landing success rates consistently improved over time taking in consideration the evolution of rocket system technologies.
- Decision Tree Classifier provides most accurate predictions of the landing outcomes based of the existing data

Appendix

- For consistent accuracy results for models across different runs, set any value to np.random.seed() or use Cross validation.
- Data cutoff is 2020. Improved results & conclusions can be achieved given newer data considering the large amount of major advancements and feats in the recent times.

Thank you!