IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Miguel P. Bento
January 23, 2023

Outline

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

Executive Summary

- Data collection, analysis and predictions:
  - Data Collection using SpaceX API and web scraping;
  - Exploratory Data Analysis (EDA) and Preprocessing using data wrangling, SQL, data visualization and interactive data visualization;
  - Predictions with Machine Learning (ML).
- Summary of all results:
  - Data Collection enabled quality data for EDA and ML;
  - EDA and Preprocessing provided insight into the data allowing to identify and clean relevant features;
  - ML predictions were able to show what makes a successful launch.

Introduction

- We analyze the possibility of a company SpaceY to compete with SpaceX.

- To that end, our goal is to predict the price of each launch:
  - We collect public data from SpaceX to determine if the first stage has a successful landing;
  - We study launch sites to determine its features and viability;
  - We determine the most important features in a successful landing.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - From SpaceX API using requests;
  - From a Wikipedia page using BeautifulSoup.
- Perform data wrangling
  - The data was preprocessed, by adding a landing outcome called "Class".
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Several models were trained on training data and tested on the split testing data.

Data Collection

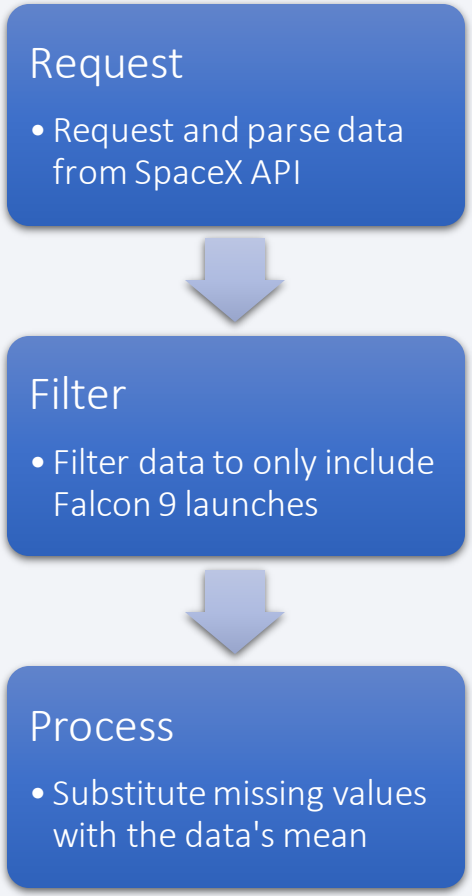
- The data was collected from SpaceX API and an Wikipedia page.

- The methods used to collect and process such data:
  - Requests module in Python;
  - BeautifulSoup module in Python.

Data Collection – SpaceX API

- SpaceX public API was used to collect data regarding past launches.

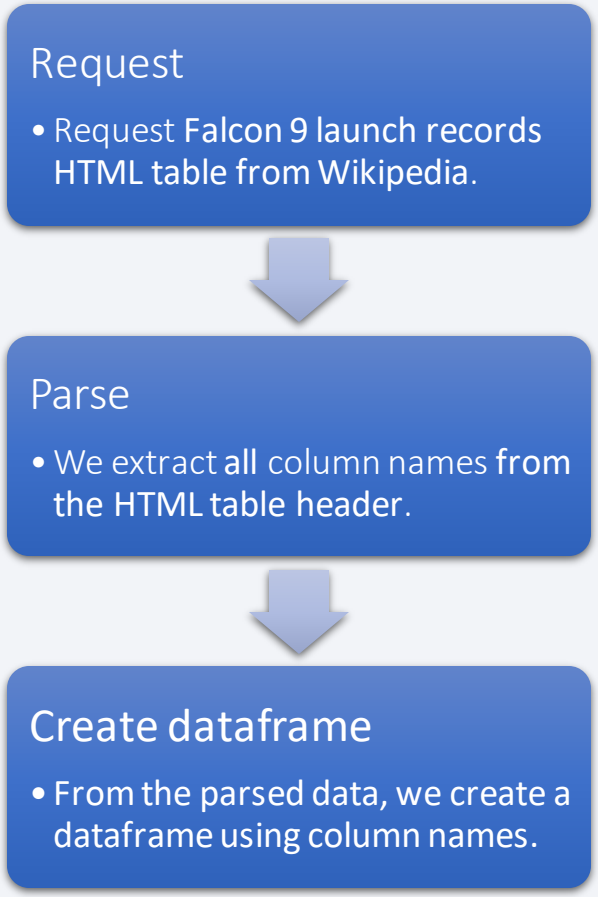
- Source code of the collection process.



Data Collection - Scraping

- A Wikipedia page on past SpaceX launches was used to collect the data through web scraping.

- Source code of the collection process.



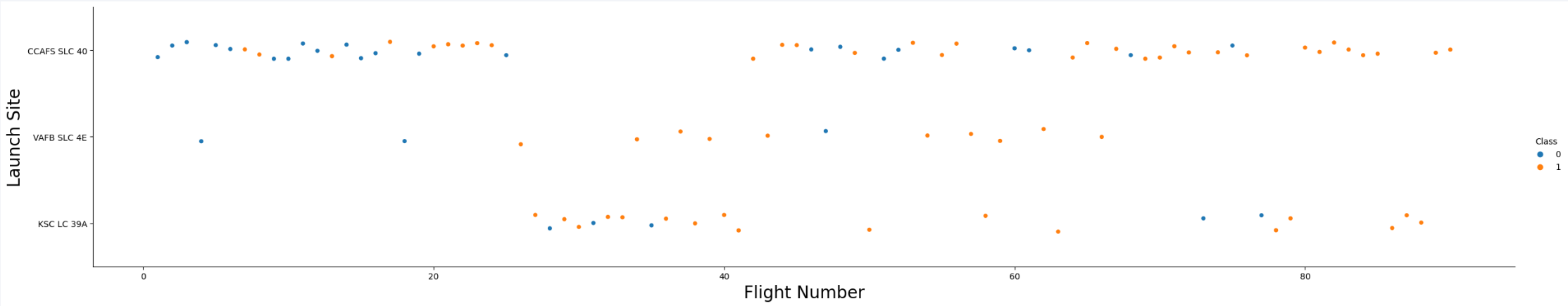
Data Wrangling

- The first step consisted on Exploratory Data Analysis (EDA).
  - The number of launches on each site;
  - The number and occurrence of each orbit;
  - The number and occurrence of mission outcome per orbit type.
- Then, we created the outcome label "Class": 1 for success, 0 for failure.

- Source code of the wrangling process.

EDA with Data Visualization

- Scatter plots, bar plots and line charts were used to explore correlation of features with success:
  - Flight Number, Payload Mass, Launch Sites, Orbit Type, Success Rate were analyzed.



- Source code of the EDA with data visualization.

EDA with SQL

- With SQL we performed EDA by displaying:
  - The names of launch sites;
  - Five records where launch sited began with the string "CCA";
  - The total payload mass carried by boosters from NASA (CRS);
  - The average payload mass carried by booster version F9 v1.1;
  - The date of the first successful landing in ground pads;
  - The names of the boosters successfully landing in drone ships and with payload mass between 4000 kg and 6000 kg;
  - The number of total successful and failed mission outcomes;

EDA with SQL (continued)

- With SQL we performed EDA by displaying:
  - The booster versions that carried the maximum payload mass;
  - The landing outcomes in drone ship, their booster versions and launch site names for the year 2015;
  - Ranked count of the landing outcomes between 2010-06-04 and 2017-03-20.

- Source code of EDA with SQL.

Build an Interactive Map with Folium

- Markers, circles, marker clusters and lines were added to the interactive map:
  - Circles were added to highlight the location of launch sites;
  - Markers label various locations by name;
  - Marker clusters group various markers into one single object as not to overwhelm the map when zoomed out;
  - Lines were added to display the distance between points in the map.

- Source code to Interactive Map with Folium.

Build a Dashboard with Plotly Dash

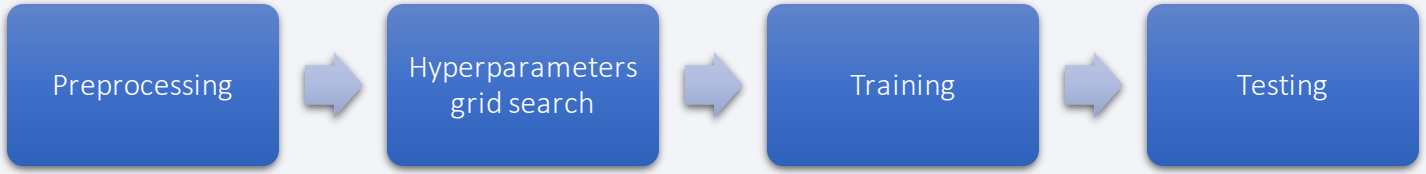
- We added the charts:
  - A pie chart with information of the success/failure on launch sites;
  - A scatter chart with a payload range describing Payload Mass vs Success.

- These charts were added as to explore the relationship between launch sites and landing success at a given payload range.

- Source code to Plotly Dash.

Predictive Analysis (Classification)

- A logistic regression, a support vector machine (SVM), a decision tree and a k-nearest neighbors were used as models to predict landing success:
  - These models were trained on collected data;
  - The hyperparameters were chosen through grid search.



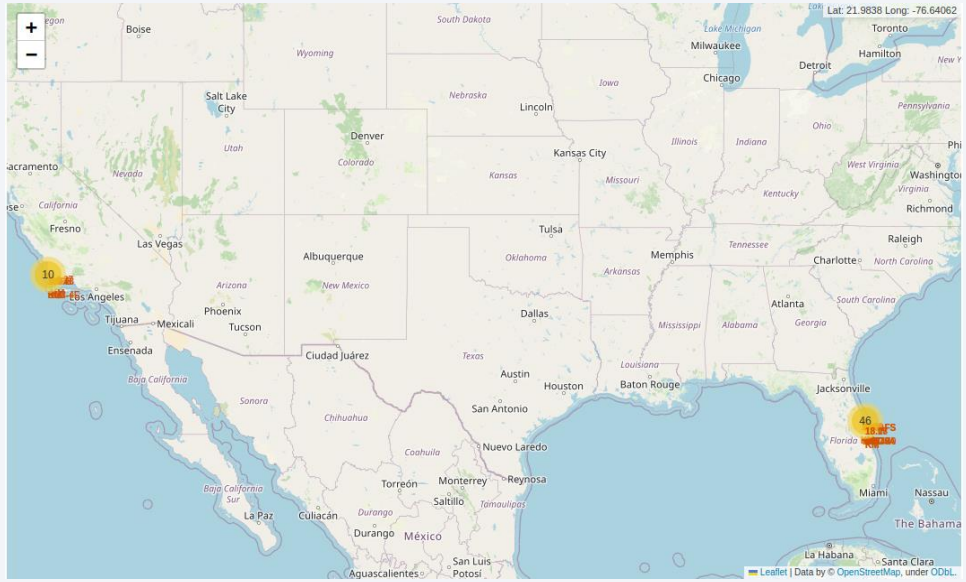
- Source code to Predictive Analysis.

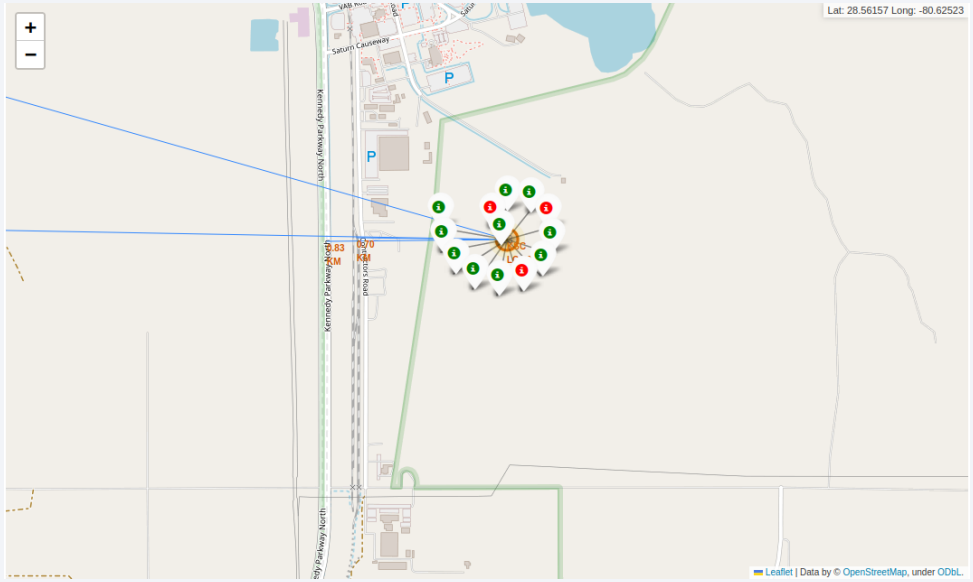
Results

- Our EDA results:
  - Both payload mass and orbit type correlate with success;
  - Over the years the success rate has been improving;
  - There are four launch sites and those are correlated with success;
  - Most landing have been attempted in drone ships;
  - Heavier payloads tend to be more successful;
  - KSC LC-39A launch site produces better outcomes;
  - Launch sites are typically close to railroads, highways and coastlines, but far from cities.

Results

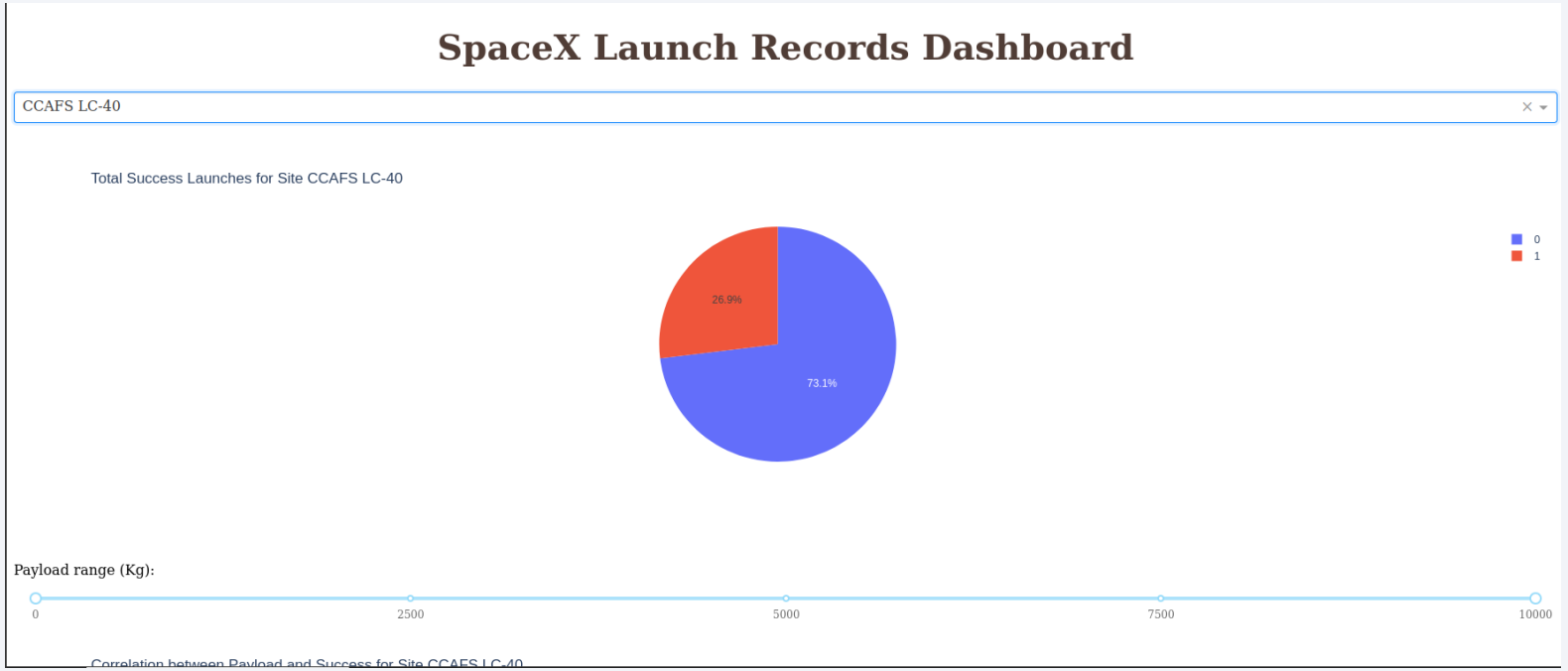
- Interactive maps allowed to understand the relationship between launch sites, logistics and safety.
- Close to logistical infrastructure and far from cities is the typical pattern.





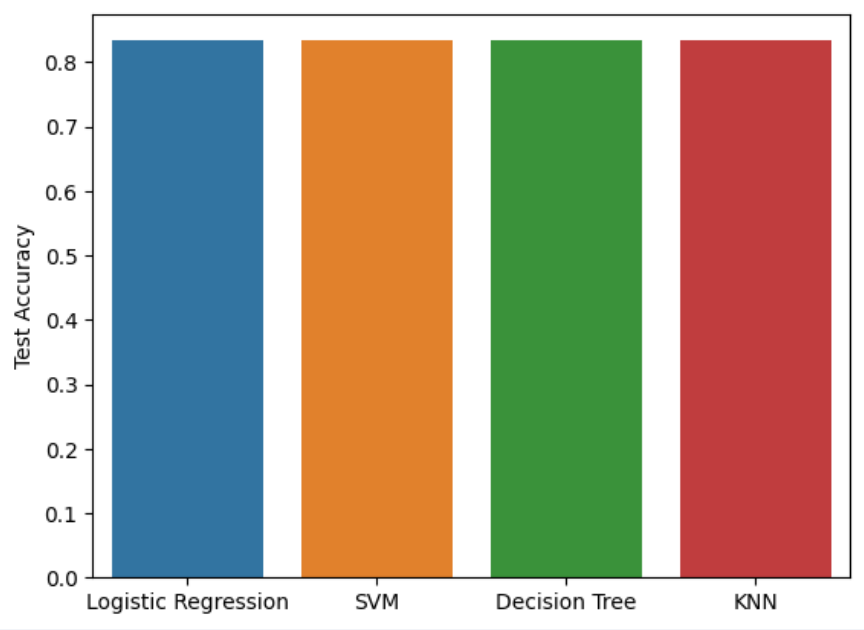
Results

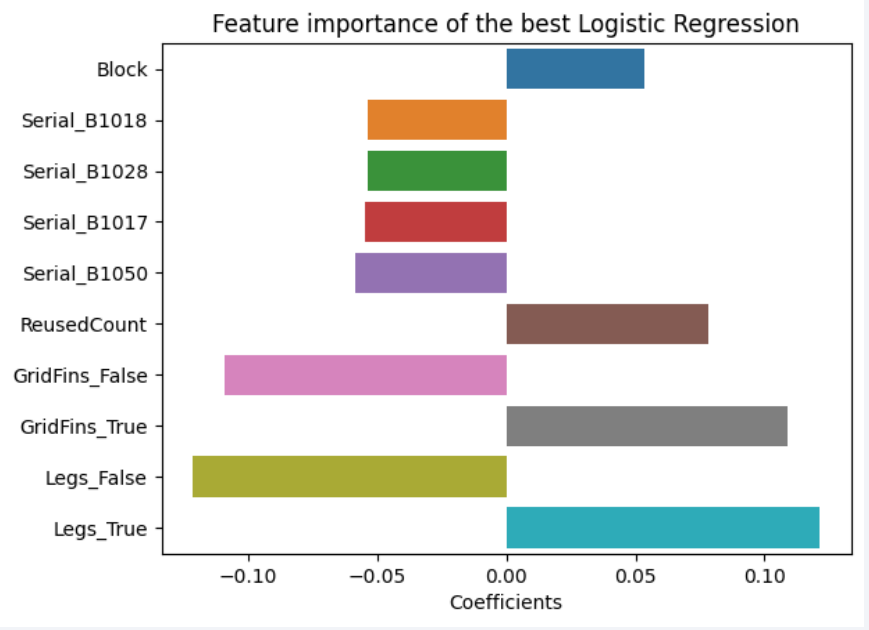
- Plotly Dashboard allowed quick access to launch sites information and payload range selection.



Results

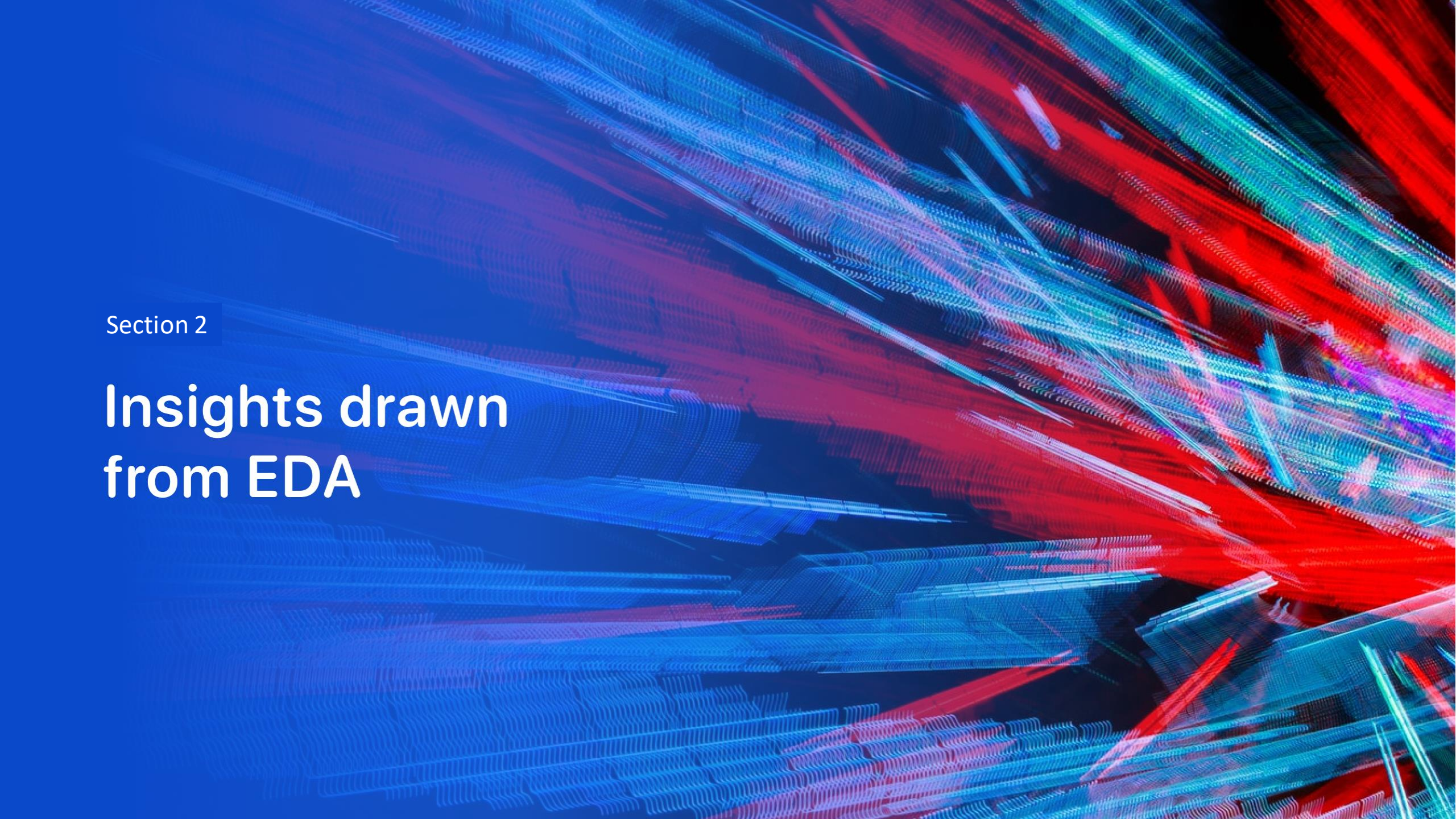
- Machine learning models showed to be equivalent in testing accuracy and confusion matrices, with accuracy of 87%.
- With logistic regression we were able to study the impact of given features.



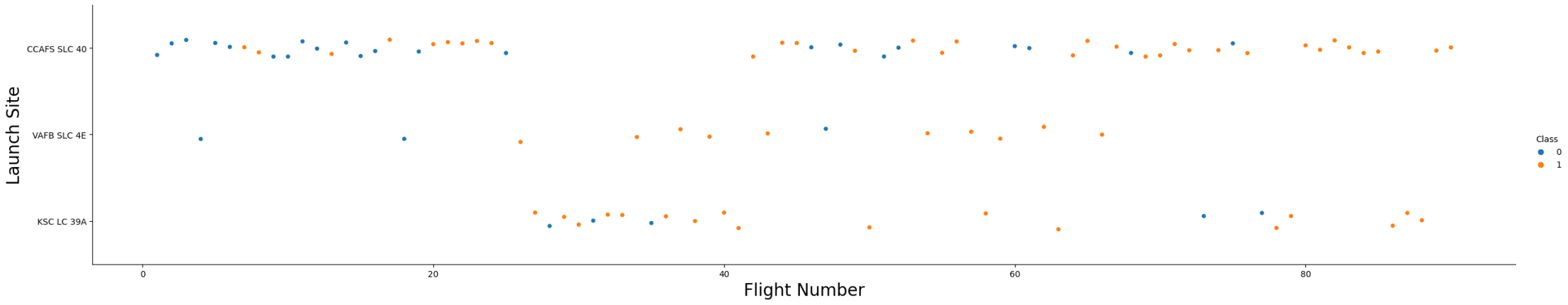


Section 2

Insights drawn from EDA

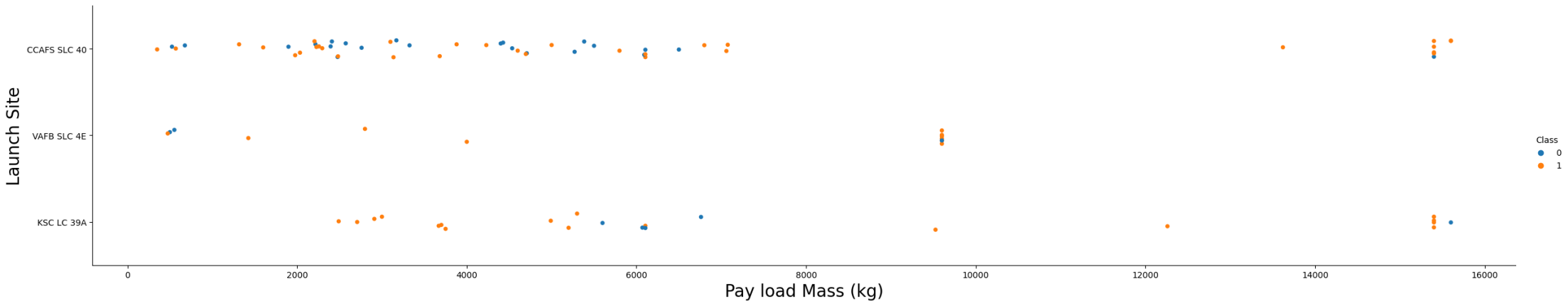


Flight Number vs. Launch Site



- Outcomes improved with flight number regardless of the launch site;
- Improvements are more apparent in CCAFS SLC 40;
- VAFB SLC 4E has been less used recently.

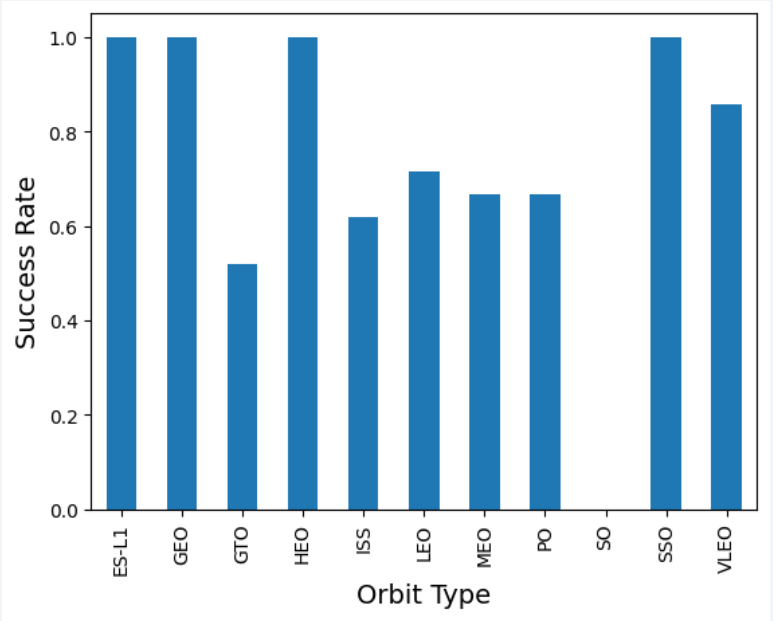
Payload vs. Launch Site



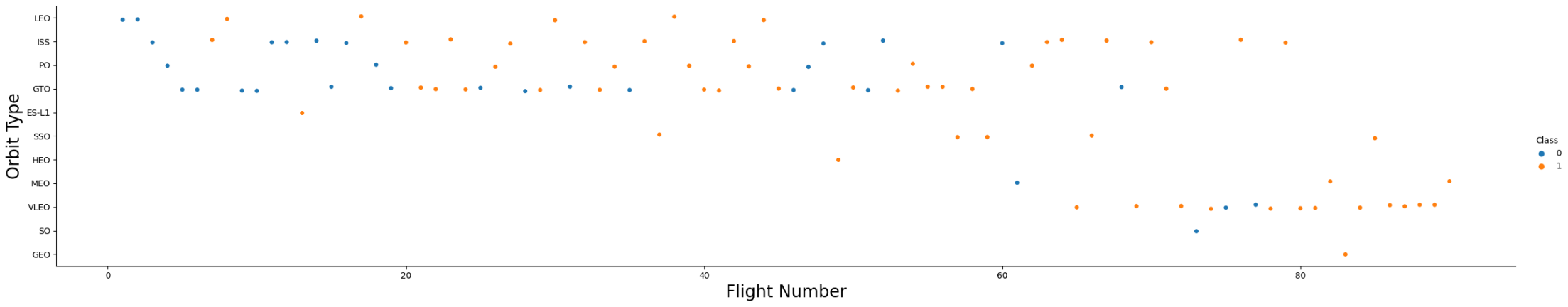
- With increasing payload mass, the success rate improves;
- VAFB SLC 4E site does not seem to launch heavy payloads;
- Above 8000 kg the outcomes were mostly successful.

Success Rate vs. Orbit Type

- Success rate was perfect in ES-L1, GEO, ISS and SSO orbit types;
- VLEO has a good success rate followed by a number of mixed results in specific orbits;
- SO orbit does not have a successful landing.

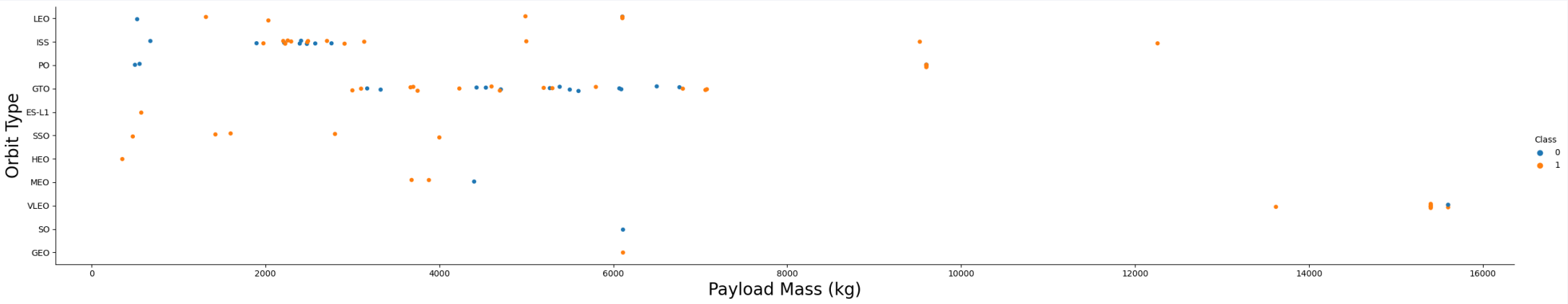


Flight Number vs. Orbit Type



- We can see that LEO orbits improved with flight number;
- In the remaining orbits, flight number does not seem to have an impact;
- SSO and VLEO orbits seem to be successful regardless of flight number.

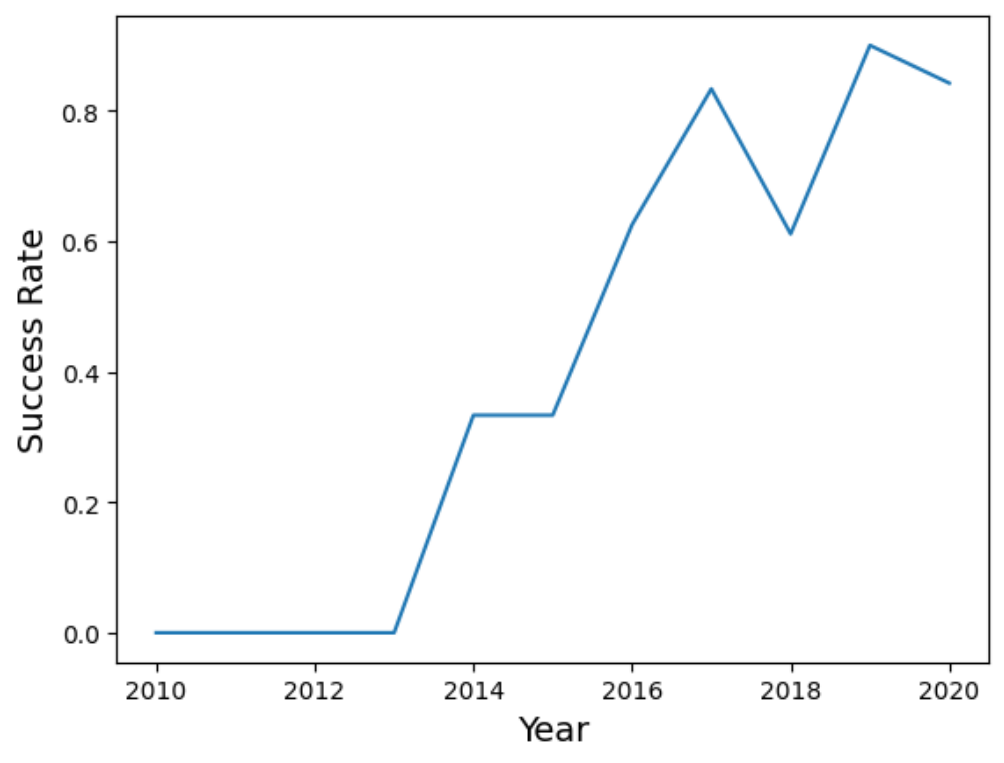
Payload vs. Orbit Type



- PO, LEO and ISS perform better landings under heavy payload masses;
- The remaining outcomes of orbit types do not seem to correlate well with payload masses.

Launch Success Yearly Trend

- Over the years, the success rate has been steadily improving.
- From 2010 and 2013, success rate was stalled at zero, with a quick improvement beginning in 2014.



All Launch Site Names

- The names of the launch sites are:

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

- The query was given by asking for unique "launch_site" entries in the dataset.

Launch Site Names Begin with 'CCA'

- Five records with launch site names beginning with 'CCA':

Date	Time (UTC)	Booster version	Launch Site	Payload	Payload Mass	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 attempt

- Five records were given naturally ordered by date.

Total Payload Mass

- The total payload mass launched by NASA (CRS) is:

Total Payload Mass (kg)
48213

- The total mass was obtained by summing the payload masses of launches of NASA (CRS).

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 was:

Average Payload Mass (kg)
2928

- We refined our search to booster version F9 v1.1 and proceeded to average the payload mass of these same launches.

First Successful Ground Landing Date

- The date of the first successful landing on a ground pad was:

Date
2015-12-22

- We selected successful launches on ground pads, and queried the minimum of the given dates.

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 was:

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

- We selected distinct booster versions from launches where the payload mass was given by a particular range.

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes is given by:

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

- Where by grouping the failures and successes we see there were 101 mission outcomes either successful or failed.

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass is given in the table.
- We used a subquery to select the maximum payload mass in the dataset.
- We then selected the booster versions which have carried this same payload mass.

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

- The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015 are:

Landing Outcome	Booster Version	Launch Site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- With the above conditions, only two records were given.

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

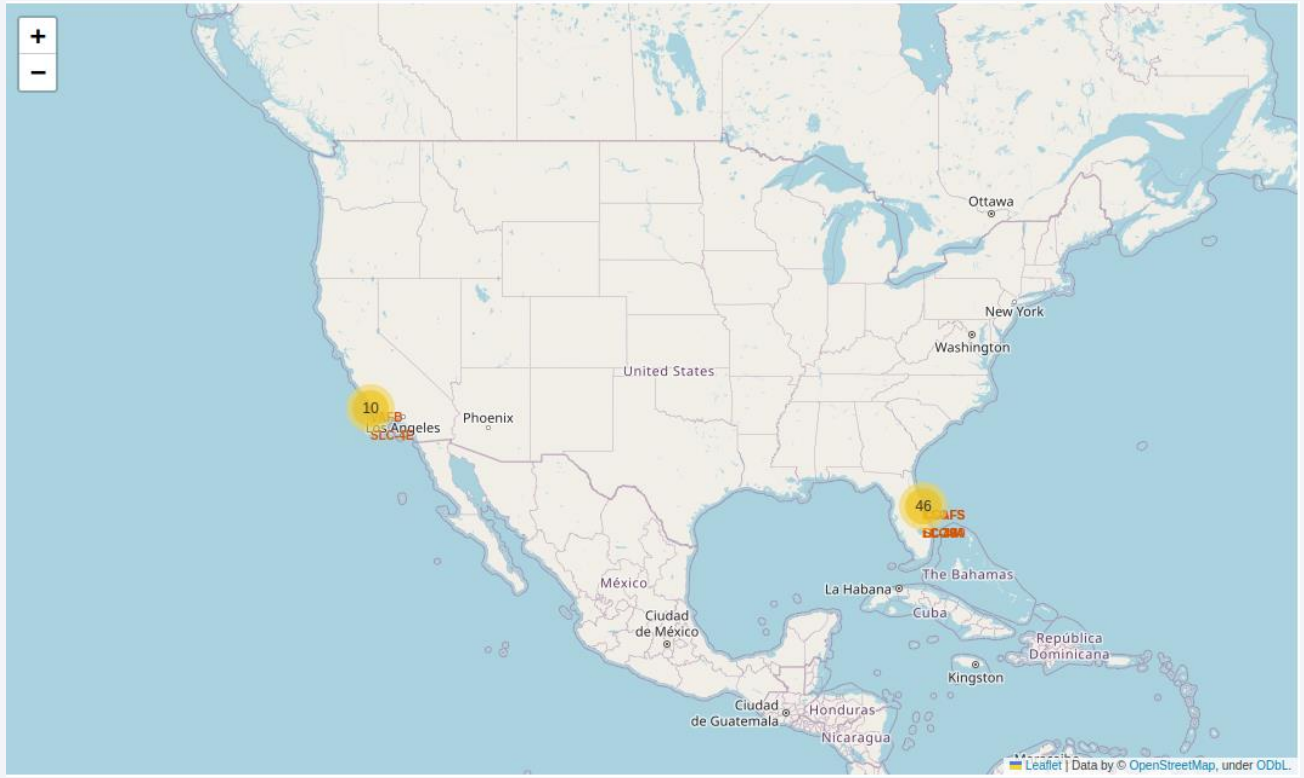
- We ranked the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order:

Landing Outcome	Outcome Count
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

- No attempts in landing were found in the dataset.

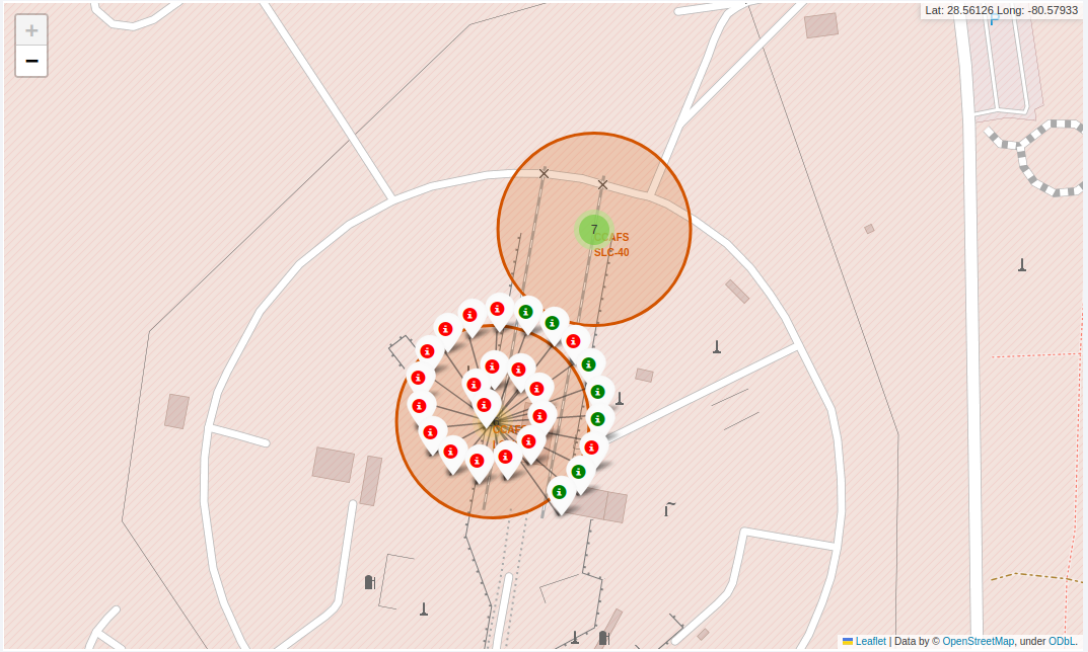

Section 3
Launch Sites
Proximities Analysis

All Launch Sites



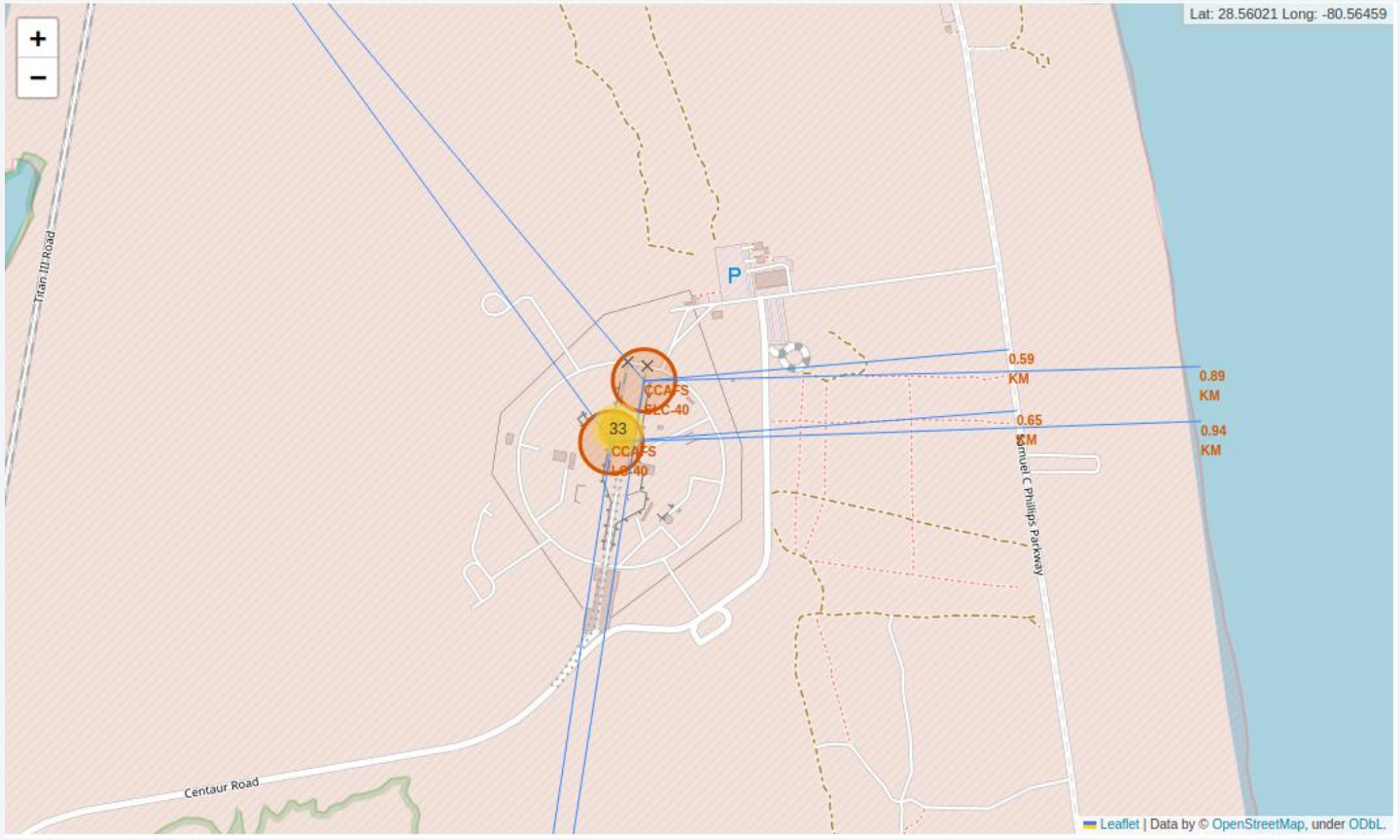
- Launch sites are near coastlines and in locations with good meteorological conditions.

Launch Outcomes by Site



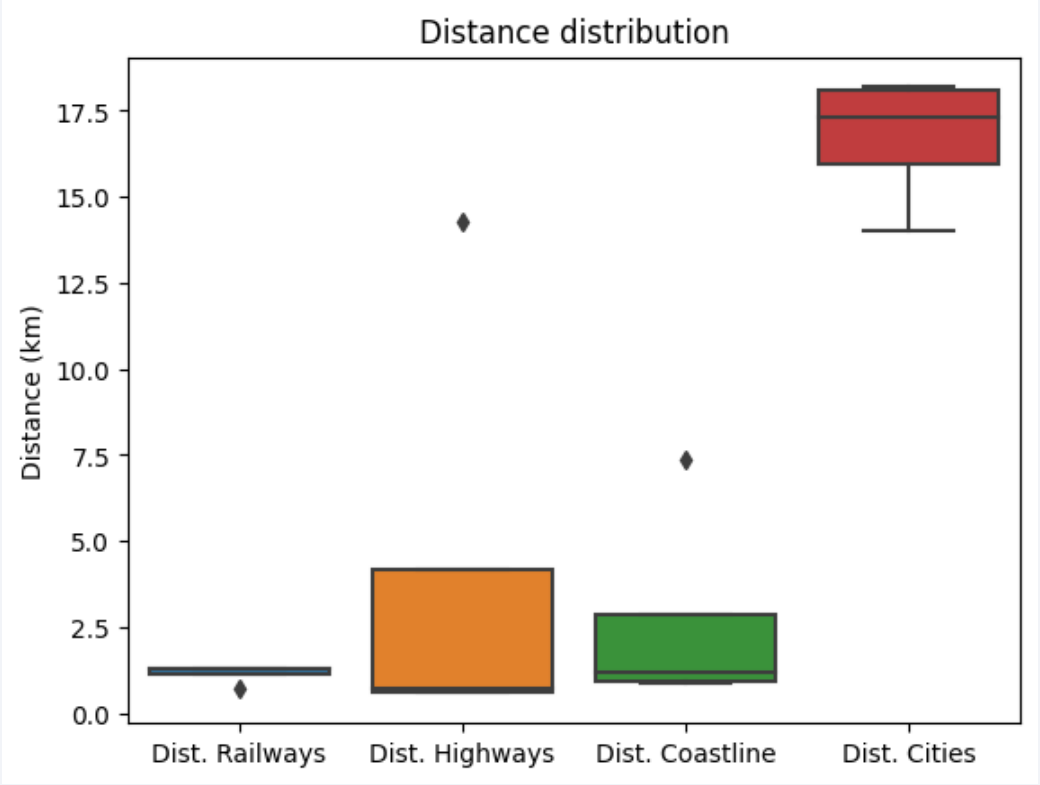
- By looking at CCAFS LC-40 the color of the clusters yields the success rate, while markers are either red (failure) or green (success).

Safety and Logistics in Launch Sites

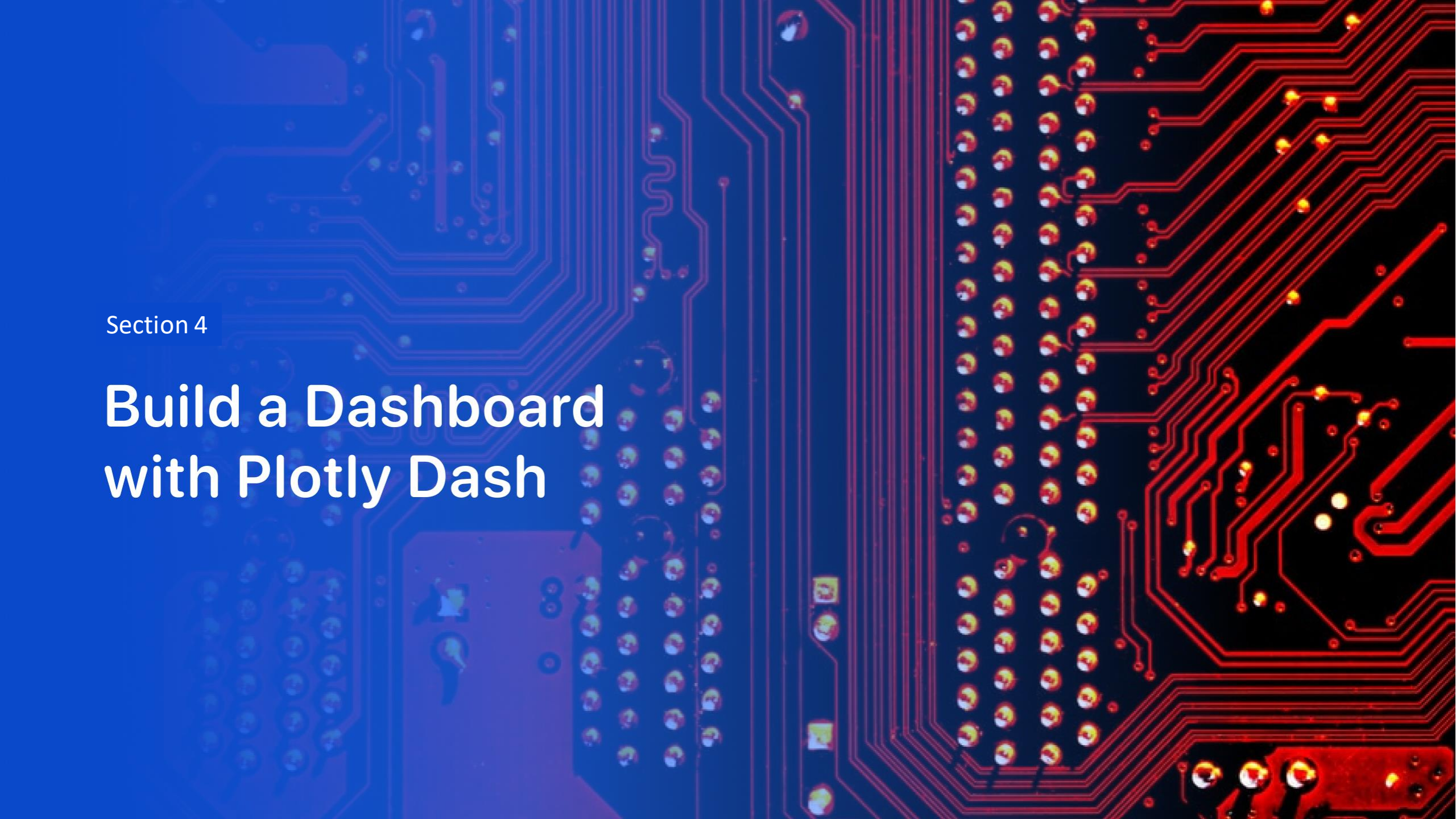


- Launch sites are built near railroads, highways and coastlines, but far from cities.

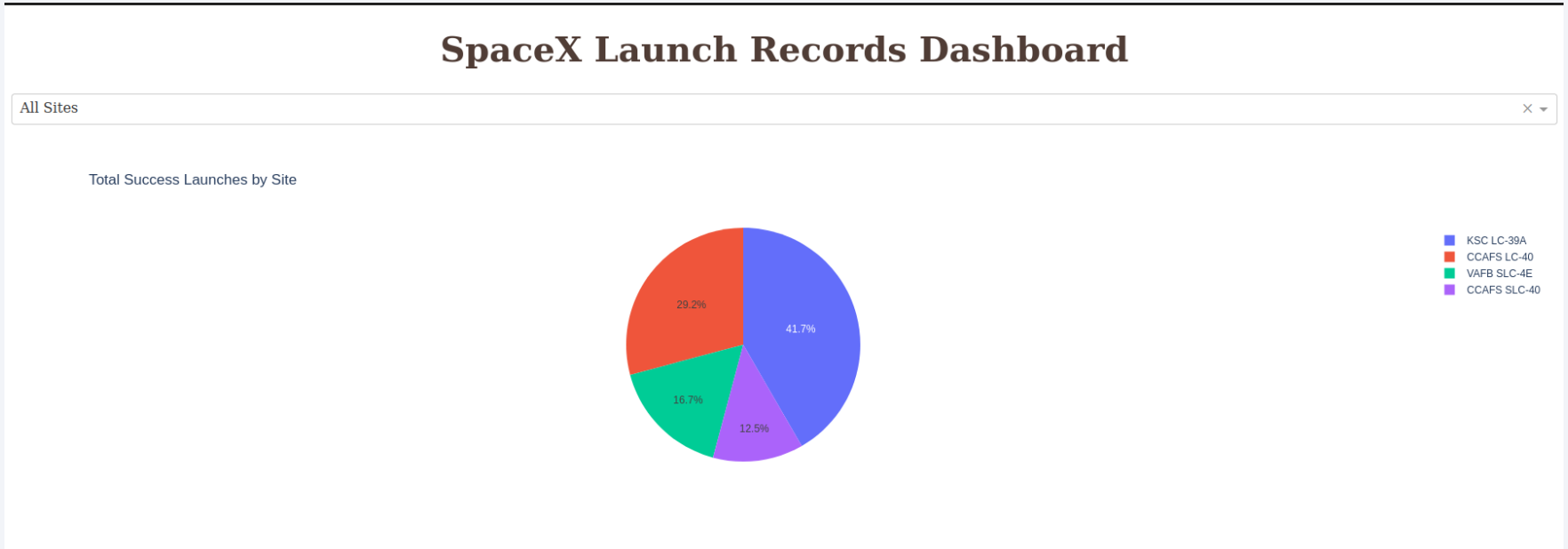
Safety and Logistics in Launch Sites



- A box plot is given showing the trend for all launch sites;
- Proximity to railways is a priority as is distance from cities.

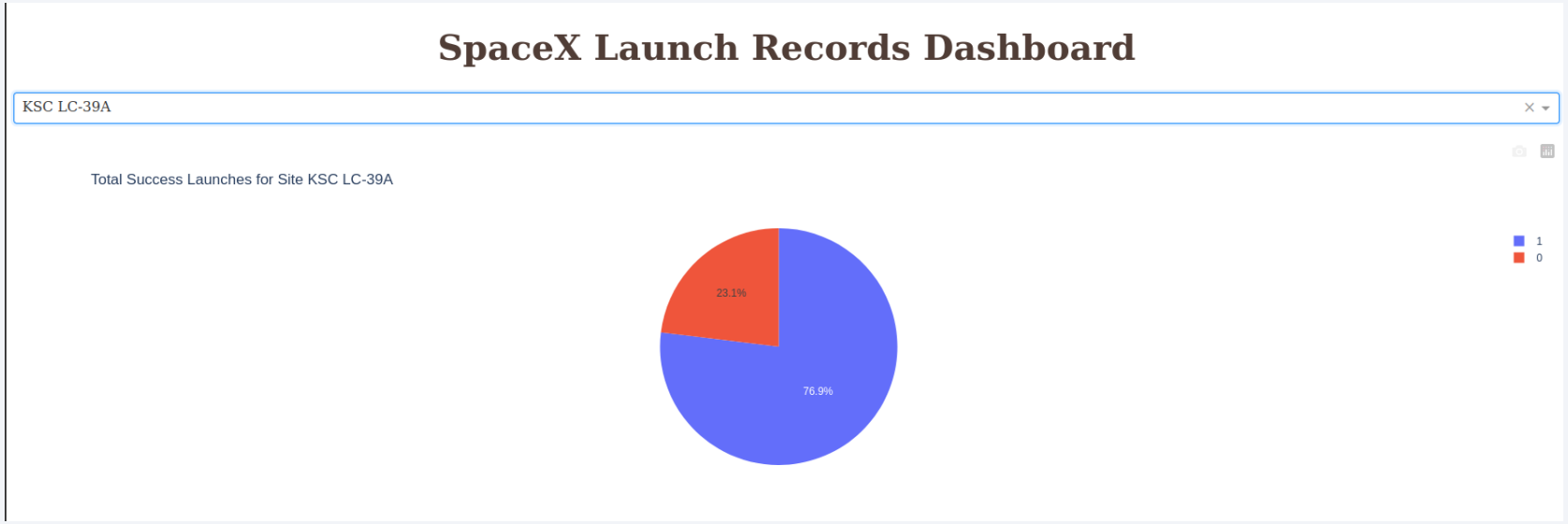

Section 4
Build a Dashboard with Plotly Dash

Successful Launches by Site



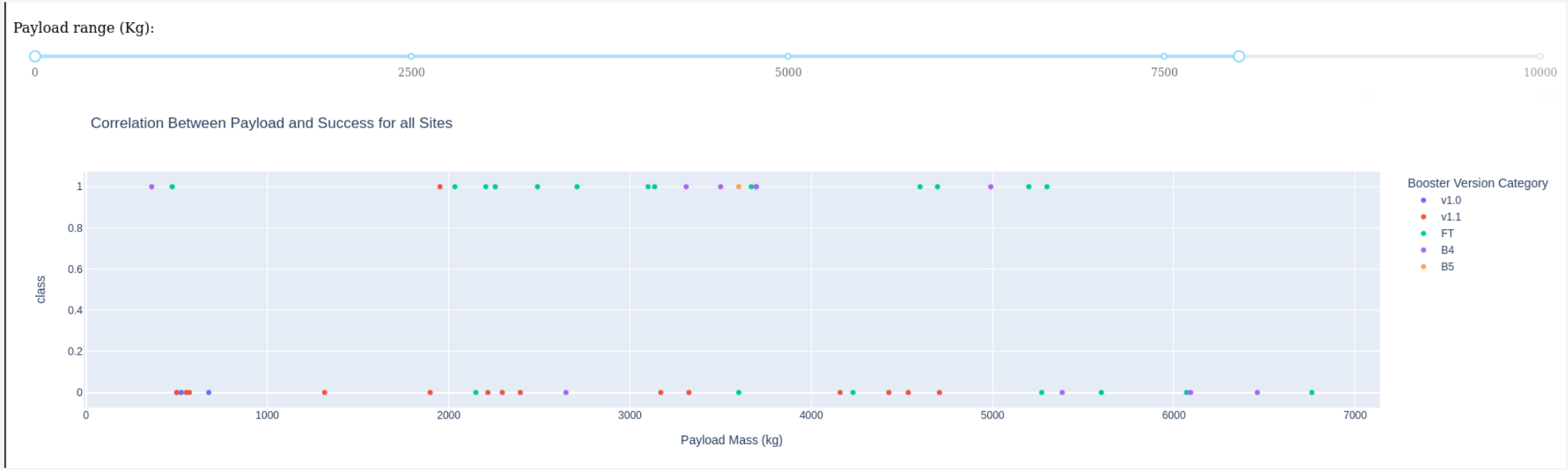
- Launch sites are correlated with success, with KSC LC-39A with the highest number of successes.

Success Rate of KSC LC-39A



- KSC LC-39A has the highest rate of success;
- 76.9% of the launches were successful and 23.1% were not.

Booster Version and Payload Mass



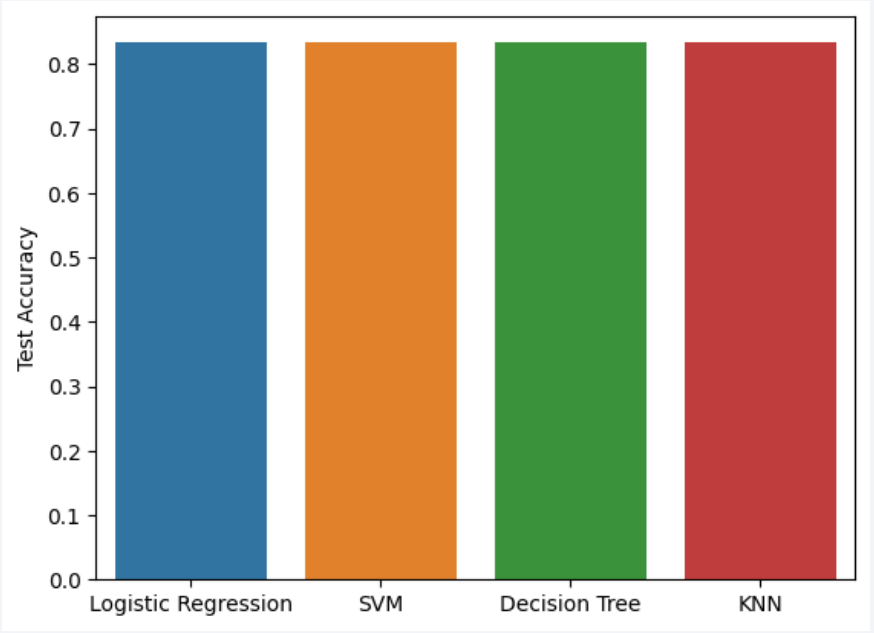
- Payloads under 5500 kg were more successful than heavier ones in general;
- FT boosters have a very good success rate, while v1.0 and v1.1 have a poor performance.

Section 5

Predictive Analysis (Classification)

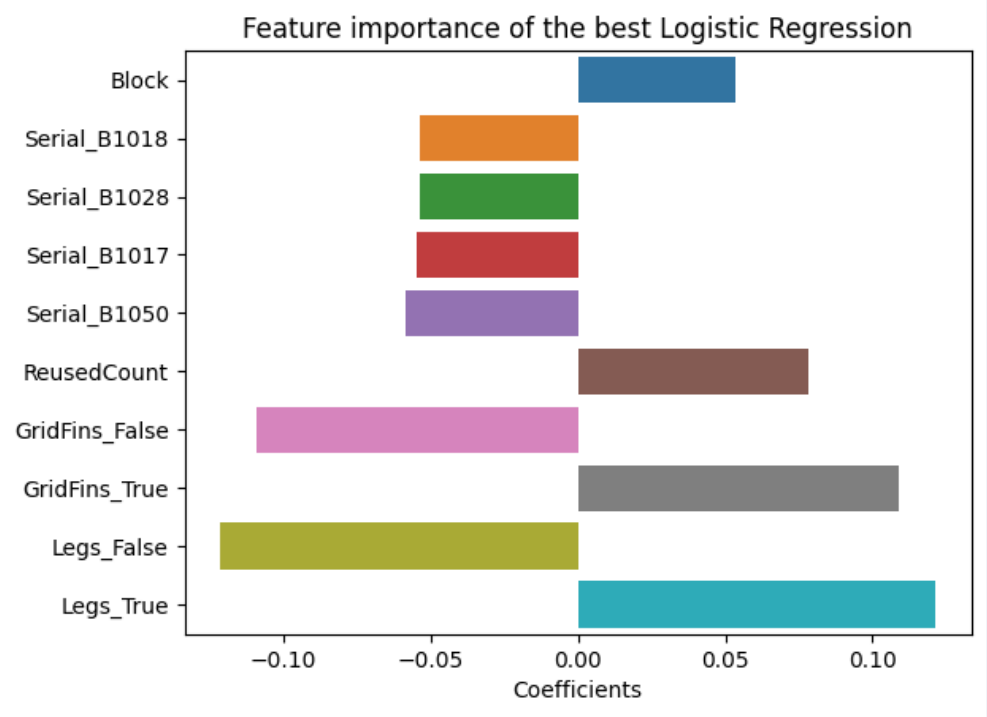
Classification Accuracy

- A bar chart was built from the testing data accuracy;
- All models have the same testing accuracy of 83%;
- Decision trees are not deterministic by nature with some instances performing better than others.



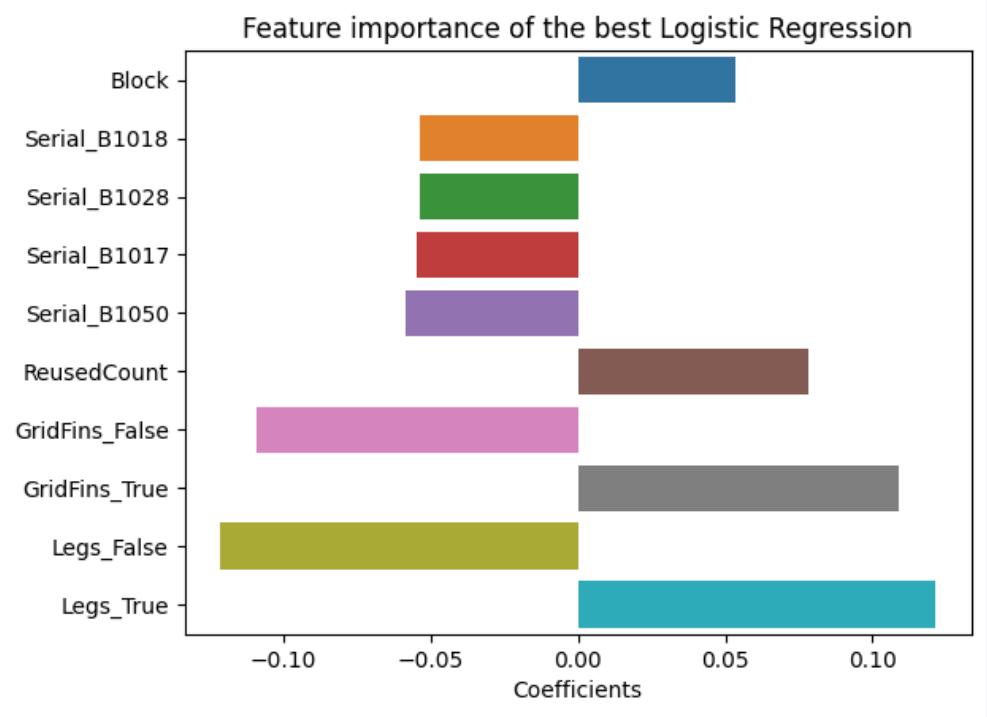
Best Model

- Because logistic regression is deterministic, we extracted the importance of features;
- The same can be done in decision trees but the most impactful features are not always the same;
- From interpretability, logistic regression is considered the best model.

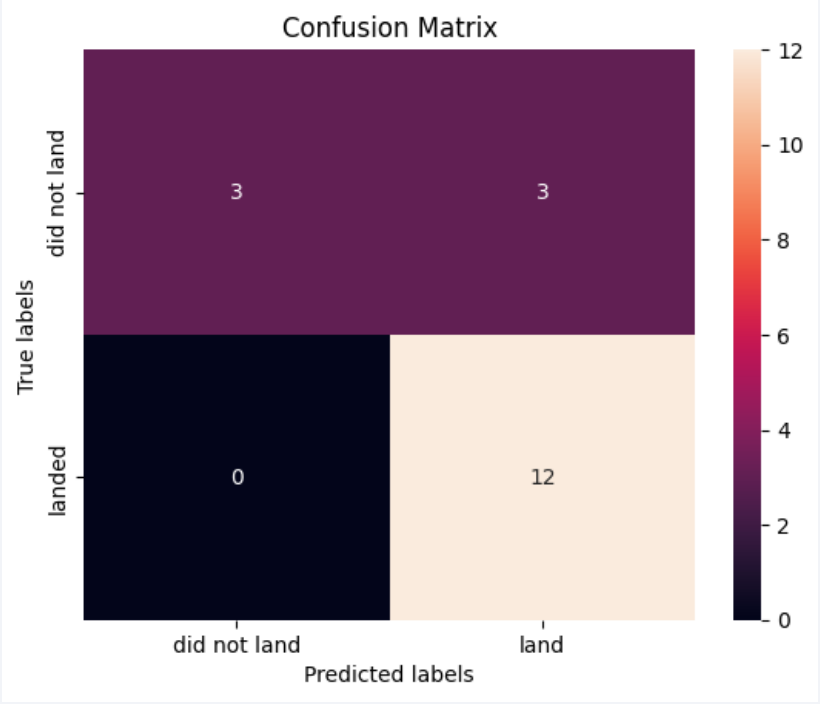


Feature Importance

- From the perspective of competing with SpaceX, we learn which factors make a successful landing;
- The existence of **legs** and **grid fins** were considered the most important factors;
- To have legs and grid fins is significantly correlated with success.



Confusion Matrix of the Best Model



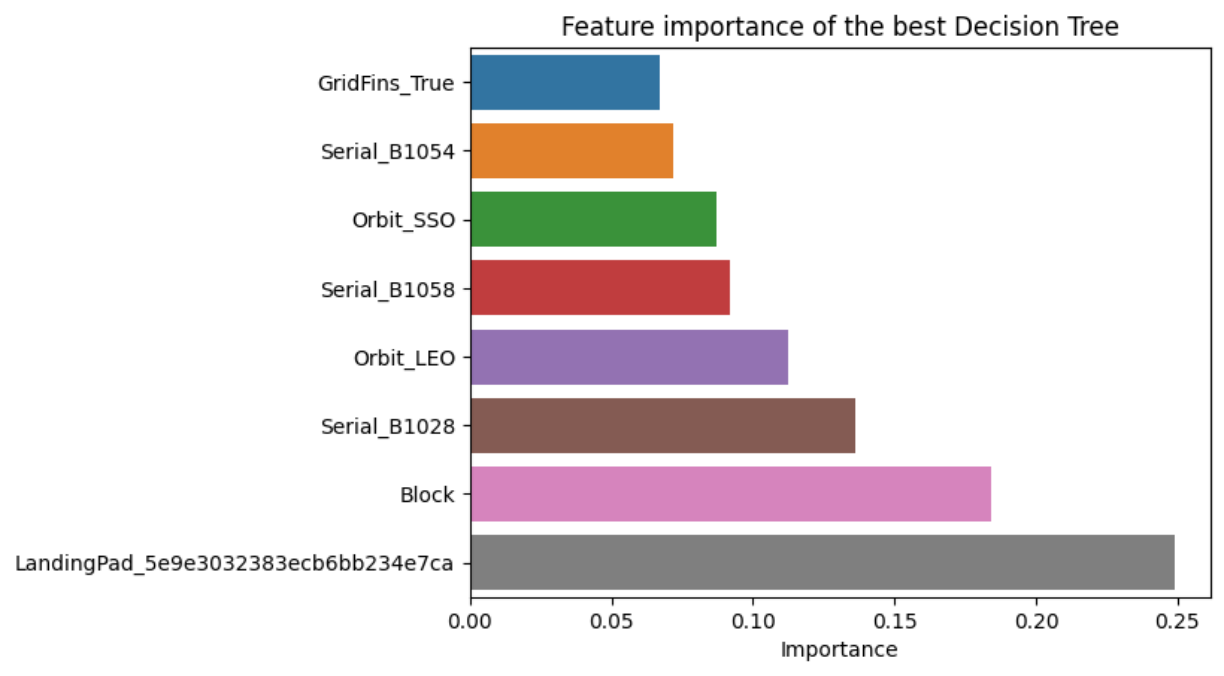
- We were able to predict successful landings in the testing data;
- Nevertheless, there were three predicted successes which were not successful;
- More data may be needed to explain these data points.

Conclusions

- Data was successfully collected from SpaceX;
- With time, SpaceX drastically improved the landing outcomes;
- A good launch site is close to coastlines, railroads and highways but far from cities;
- KSC LS-39A site has the best success rate;
- Payload masses are clearly correlated with success but success depends on the orbit type;
- Feature importance allowed us to find the most impactful technological decisions of SpaceX;
- These are grid fins and legs;
- This finding allows us to consider this advances from the start in SpaceY.

Appendix

- Feature importance of the decision tree:



- We check that it does not agree with logistic regression.

Thank you!