Winning Space Race with Data Science

Marco Vesco
21/02/2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Data on Falcon 9 rocket launches was gathered through the SpaceX API, prepared and explored using visualization techniques such as plots, charts, interactive maps and dashboards. Various classification models were trained to predict the first stage landing outcome of Falcon 9 rockets based on a number of relevant features.
- Overall, Success Rate has been improving significantly in time. Most launches did not succeed in landing the first stage before the 20th attempt.
- Logistic Regression, Support Vector Machine, Decision Tree Classifier and K-Nearest Neighbour classification models perform the same on the available data – they can accurately predict about 83% of first stage landing outcomes on the test data. All models tend to return false positives.

Introduction

SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each. Much of SpaceX's savings can be credited to the reuse of the first stage. Therefore **if we can determine whether the first stage will land, we can determine the cost of a launch**. This information can be used by a competitor wanting to bid against SpaceX for a rocket launch.

What determines if the Falcon 9 first stage will land successfully? Can we predict the outcome of a launch? That is what we aim to find out in this study.

Section 1

Methodology

Methodology

- Extracted data from SpaceX API and processed to create data frame
- Performed data wrangling
  - Dealt with null values
  - Counted number of launches across different variables
  - Created outcome label for successful/unsuccessful launches
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
  - Logistic Regression, Support Vector Machine, Decision Tree, K-Nearest Neighbours
  - Tuned hyperparameters using Grid Search
  - Evaluated models based on accuracy score and confusion matrix

Data Collection

Data sets on Falcon 9 launch records can be collected in two ways:

Space X API

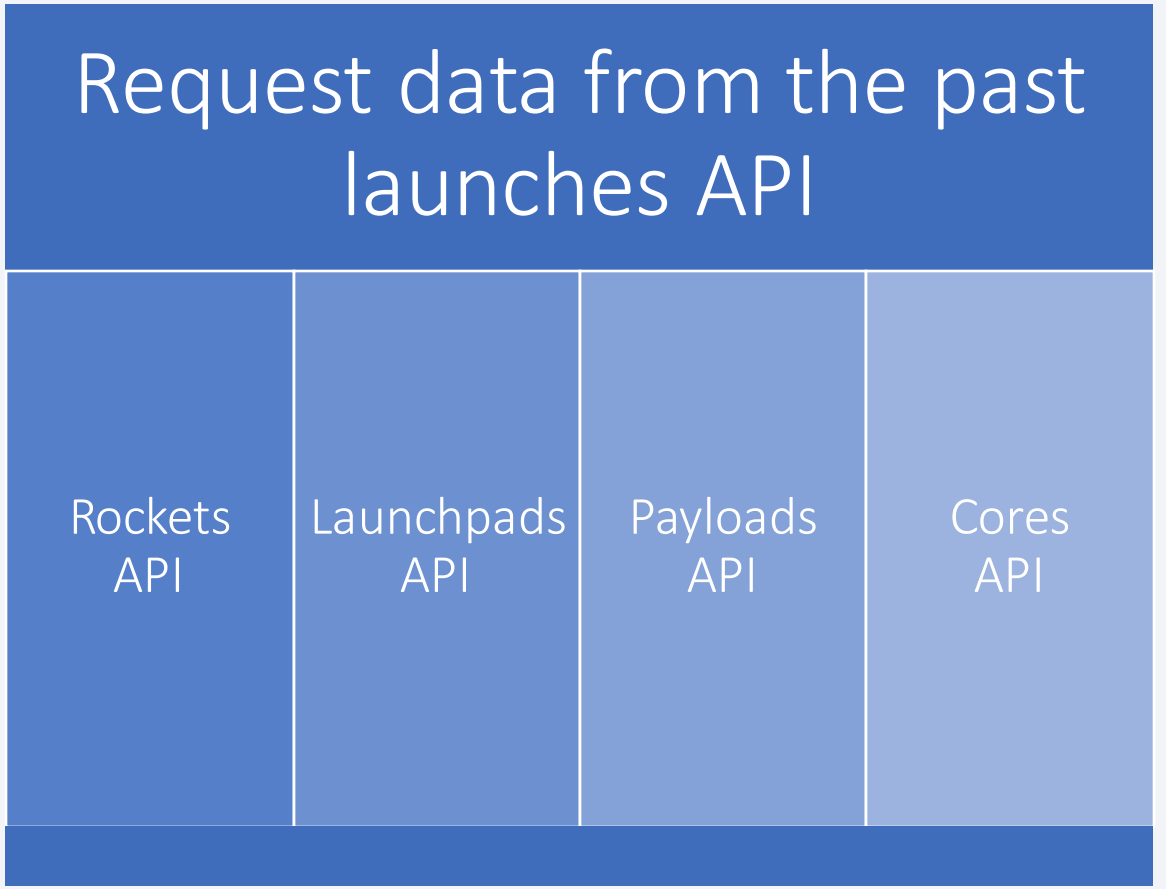
- Get request

Wikipedia page

- Scraping HTML table

Data Collection – SpaceX API

- Extracted launch data using GET request to SpaceX API
- Processed and appended select data to list objects
- Created pandas data frame

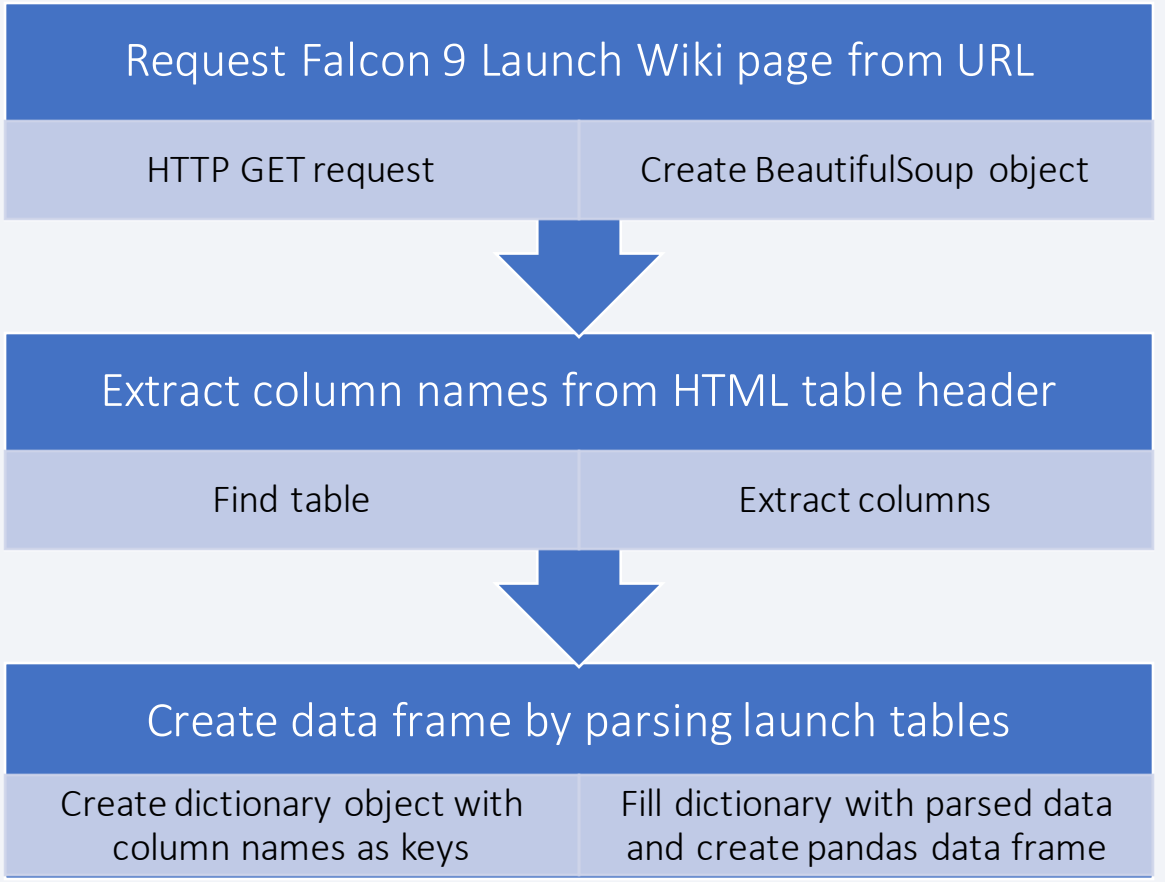


GitHub link to notebook

Data Collection - Scraping

- Scraped the **List of Falcon 9 and Falcon Heavy launches HTML table** from Wikipedia using GET request
- Processed data using BeautifulSoup and appended to dictionary object
- Created pandas data frame

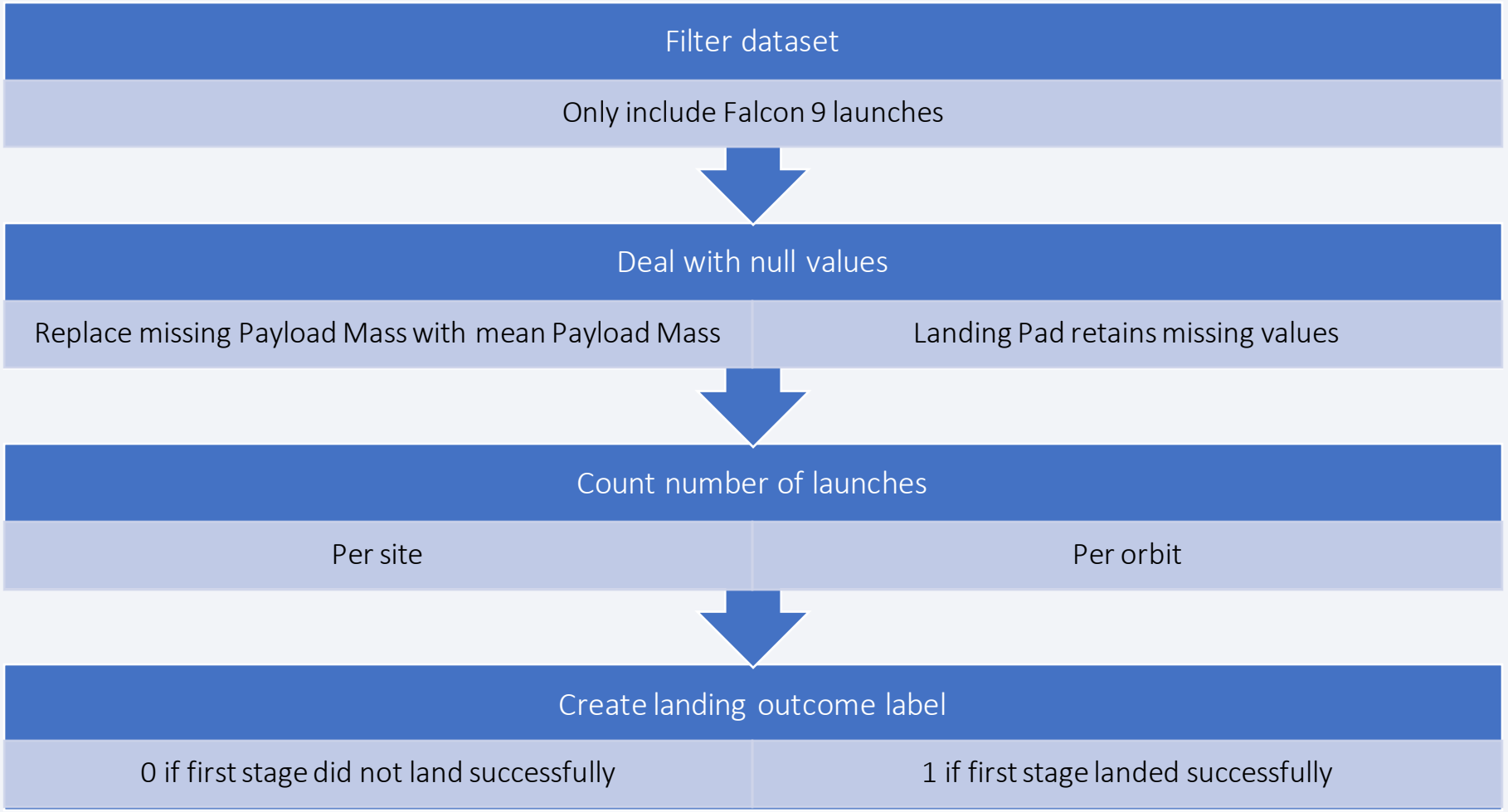
GitHub link to notebook



Data Wrangling

GitHub link to first data wrangling notebook

GitHub link to second data wrangling notebook



EDA with Data Visualization

- Categorical scatter plot of **Flight Number** vs. **Payload Mass** (hue overlay = **Success**)
  - Show if the continuous launch attempts and payload size correlate with first stage landing outcome
- Categorical scatter plot of **Flight Number** and **Launch Site** (hue overlay = **Success**)
  - Compare the continuous launch attempts and different launch sites based on first stage landing outcome
- Bar plot of **Success Rate** per **Orbit** type
  - Show the first stage landing success rate for each orbit
- Categorical scatter plot of **Flight Number** and **Orbit** type (hue overlay = **Success**)
  - Compare the continuous launch attempts and orbit type based on first stage landing outcome
- Categorical scatter plot of **Payload Mass** and **Orbit** type (hue overlay = **Success**)
  - Compare the payload size and orbit type based on first stage landing outcome
- Line chart of **Year** vs. **Success Rate**
  - Visualize the yearly trend of first stage landing outcome

GitHub link to notebook

EDA with SQL

- Displayed names of unique launch sites in the space mission
- Displayed 5 records of launch sites beginning with the string 'CCA'
- Displayed total payload mass carried by boosters launched by NASA (CRS)
- Displayed average payload mass carried by booster version F9 v1.1
- Listed date of first successful landing outcome in ground pad
- Listed names of boosters with success in drone ship and payload mass greater than 4000 but less than 6000
- Listed total number of successful and unsuccessful mission outcomes
- Listed names of booster versions which have carried the maximum payload mass
- Listed failed landing outcomes in drone ship along with booster versions and launch site names in year 2015
- Counted and ranked landing outcomes between 2010-06-04 and 2017-03-20, in descending order

GitHub link to notebook

Build an Interactive Map with Folium

- Marked each launch site on the map with a **circle** and **marker** object
  - Visualize each site's actual location
- Marked all successful and failed launches for each site on the map with a **marker cluster** object (populated with specific markers)
  - Visualize the distribution of launches across sites, and which sites have high first stage landing success rate
- Displayed the distance between a launch site and the nearest coastline, railway, highway and city using **line** objects and corresponding **marker** objects
  - Visualize the features of an optimal location for a launch site

GitHub link to notebook

Build a Dashboard with Plotly Dash

- **Drop-down list** of launch sites
  - Filter graphs and compare data on different launch sites
- **Pie chart** of total successful launches by site
  - Easily glance at the amount of successful launches for the selected site(s)
- Payload **range slider**
  - Filter scatter plot to focus on a specific range of payload mass
- **Categorical scatter plot** of Payload Mass and Success (hue overlay = Booster Version Category)
  - Visualize how payload mass may correlate with first stage landing outcomes for the selected site(s) and booster version(s)

GitHub link to code

Predictive Analysis (Classification)

Data preparation
- Standardize dataset
- Split into training and testing data

Create classifier model objects
- Logistic Regression
- Support Vector Machine
- Decision Tree
- K-Nearest Neighbours

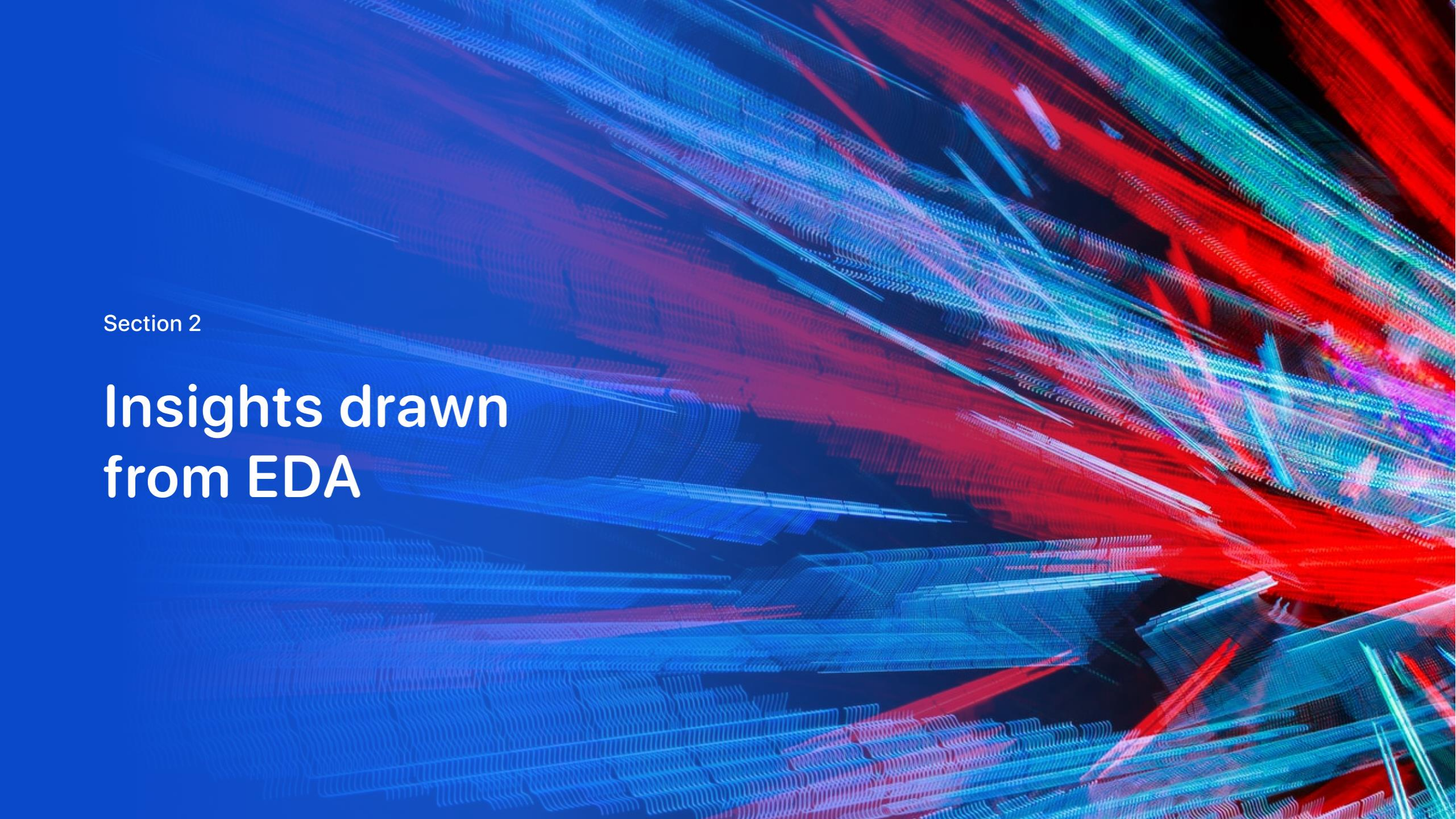
Model tuning
- Find best hyperparameters with Grid Search
- Fit model objects to training data

Model evaluation
- Calculate accuracy score on test data
- Plot confusion matrix
- Compare accuracy scores to find best performing model

[GitHub link to notebook]

Results

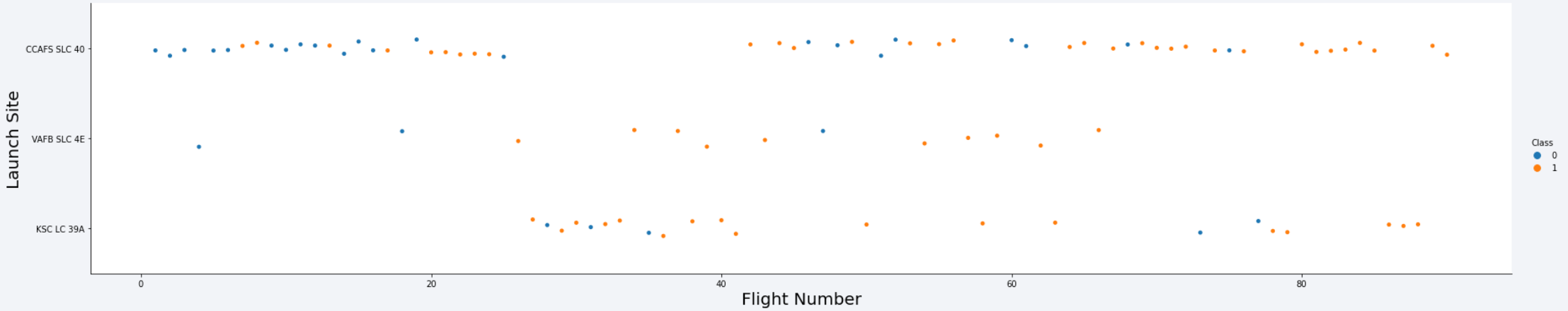
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results



Section 2

Insights drawn from EDA

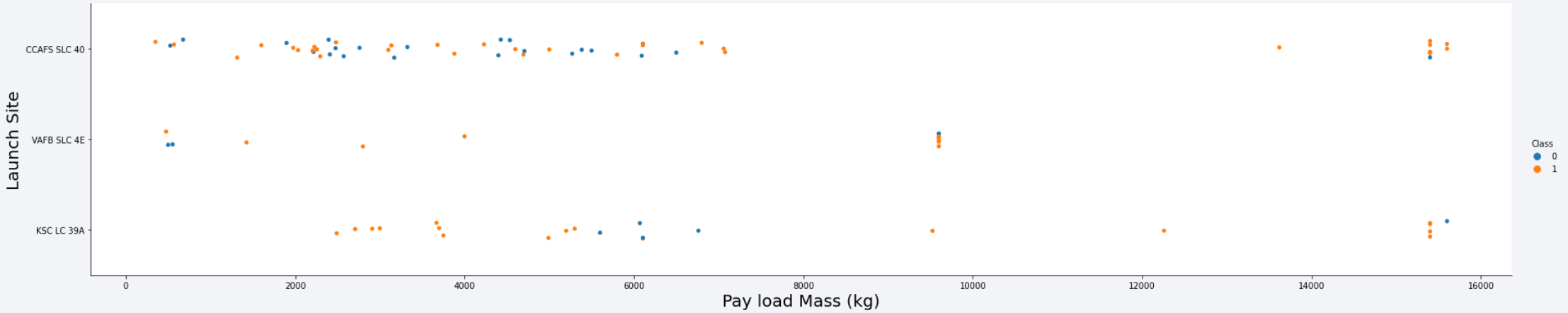
Flight Number vs. Launch Site



Most first stage landing failures seem to be concentrated in the early launches. Almost exclusively from launch site CCAFS SLC 40.

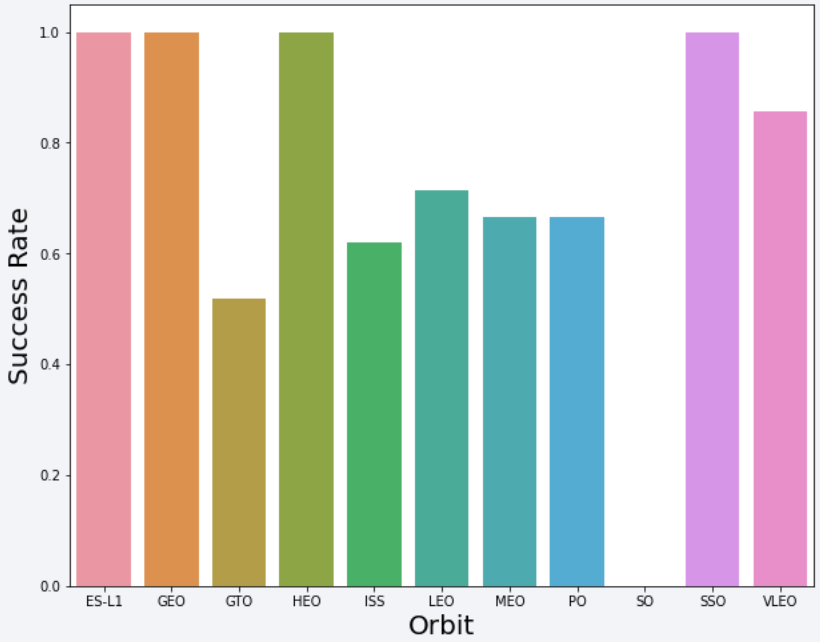
We can notice a stark improvement in success rate across all launch sites after the 20th launch mark.

Payload vs. Launch Site



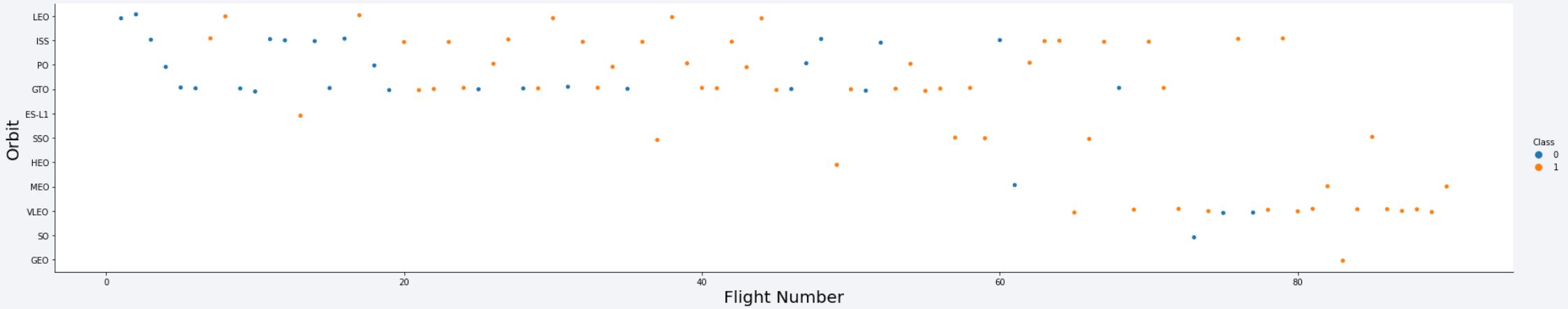
Looking at the VAFB SLC 4E launch site, we can see that no rockets were launched from this site with payload mass greater than 10000 Kg.

Success Rate vs. Orbit Type



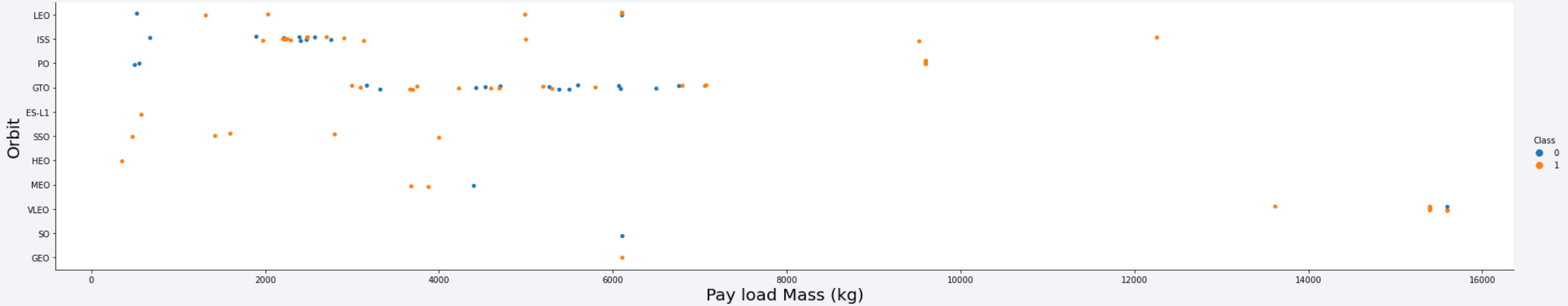
The chart shows a difference in success rate between orbit types. However, we need to consider the number of flights per orbit to see if the success rate is backed by a significant number of attempts.

Flight Number vs. Orbit Type



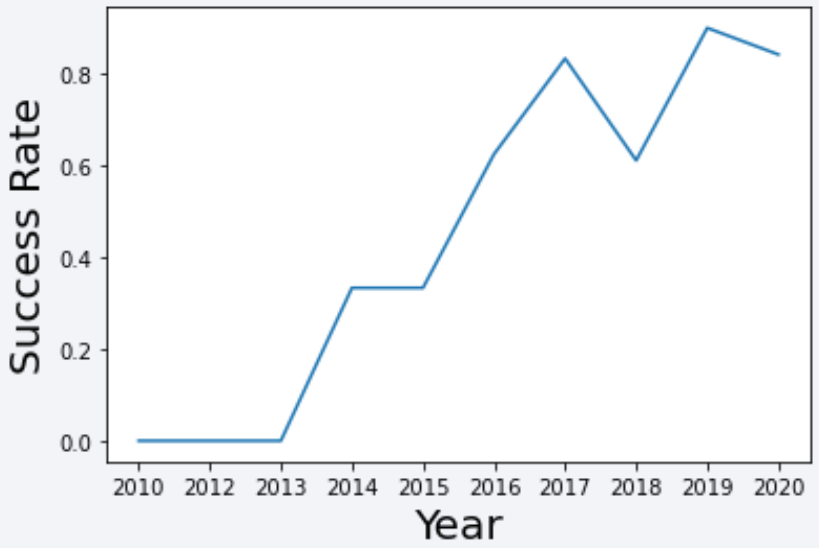
The first stage landing success for most orbits appears to correlate with the number of flights. The SSO orbit is the only one with a Success Rate of 100% over multiple launches.

Payload vs. Orbit Type



Most first stage landing successes with heavy payloads were registered for the PO, LEO, VLEO and ISS orbits. On the contrary, the GTO orbit shows many unsuccessful attempts in the heavy payload range.

Launch Success Yearly Trend



Starting from 2013, the trend of the success rate is positive, with the highest peak in 2019 (about 90% first stage landing successes).

All Launch Site Names

Query result:

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

Explanation:

The name of each unique launch site. Found using a SQL query with

group by LAUNCH_SITE

Launch Site Names Beginning with 'CCA'

Query result:

launch_site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Explanation:

5 records where launch site begins with 'CCA'. Found using a SQL query with

where LAUNCH_SITE like 'CCA%' limit 5

Total Payload Mass

Query result:

Explanation:

Total payload mass carried by boosters launched by NASA. Found using a SQL query with

sum (PAYLOAD_MASS__KG_)

Average Payload Mass by F9 v1.1

Query result:

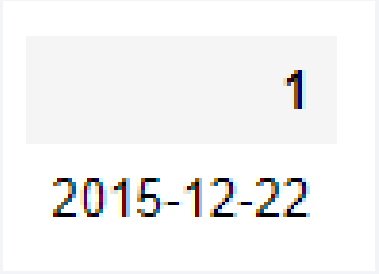
Explanation:

Average payload mass carried by booster version F9 v1.1. Found using a SQL query with

```
select avg(PAYLOAD_MASS__KG_)
```

First Successful Ground Landing Date

Query result:



Explanation:

Date in which the first successful landing in ground pad was achieved. Found using a SQL query with

select min(DATE)

Successful Drone Ship Landing with Payload between 4000 and 6000

Query result:

booster_version	avg_mass
F9 B4 B1040.1	4990
F9 B4 B1043.1	5000
F9 FT B1032.1	5300

Explanation:

Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000. Found using a SQL query with

PAYLOAD_MASS__KG_ between 4001 and 5999

Total Number of Successful and Failure Mission Outcomes

Query result:

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Explanation:

Total number of successful and unsuccessful mission outcomes. Found using a SQL query with

count(MISSION_OUTCOME)

Boosters Carried Maximum Payload

Query result:

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

Explanation:

Names of the booster versions which have carried the maximum payload mass. Found using a SQL query with

where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_)

2015 Launch Records

Query result:

landing__outcome	booster_version	launch_site	DATE
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	2015-01-10
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	2015-04-14

Explanation:

Failed landing outcomes in drone ship, their booster versions, and launch site names in year 2015. Found using a SQL query with

where LANDING__OUTCOME like '%Failure%(drone%ship)%' and DATE like '%2015%'

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

Query result:

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

Explanation:

Count of landing outcomes between the date 2010-06-04 and 2017-03-20, ranked in descending order. Found using a SQL query with

where DATE > '2010-06-04' and DATE < '2017-03-20'

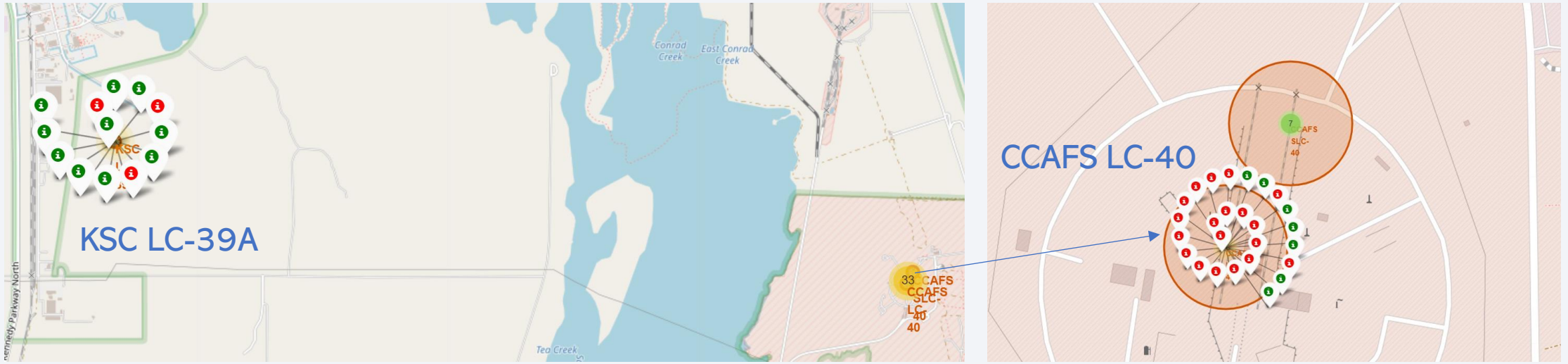
Section 4

Launch Sites Proximities Analysis

Launch Sites Map Overview

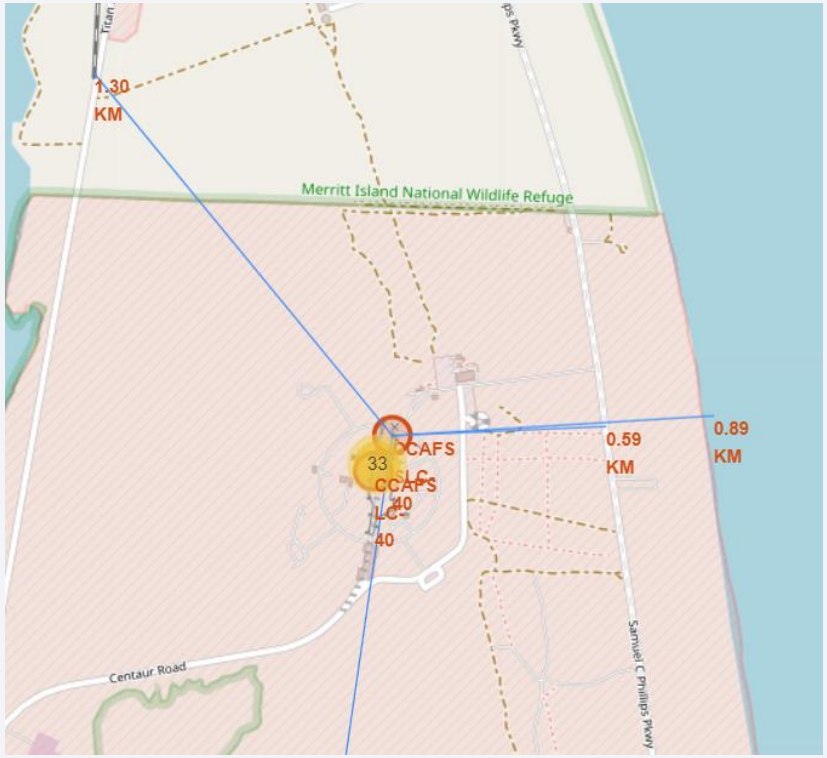
Three launch sites are located on the east coast of the US and one on the west coast. They are North of the equator and at approximately the same latitude, in very close proximity to the coast.

Launch Outcomes Map

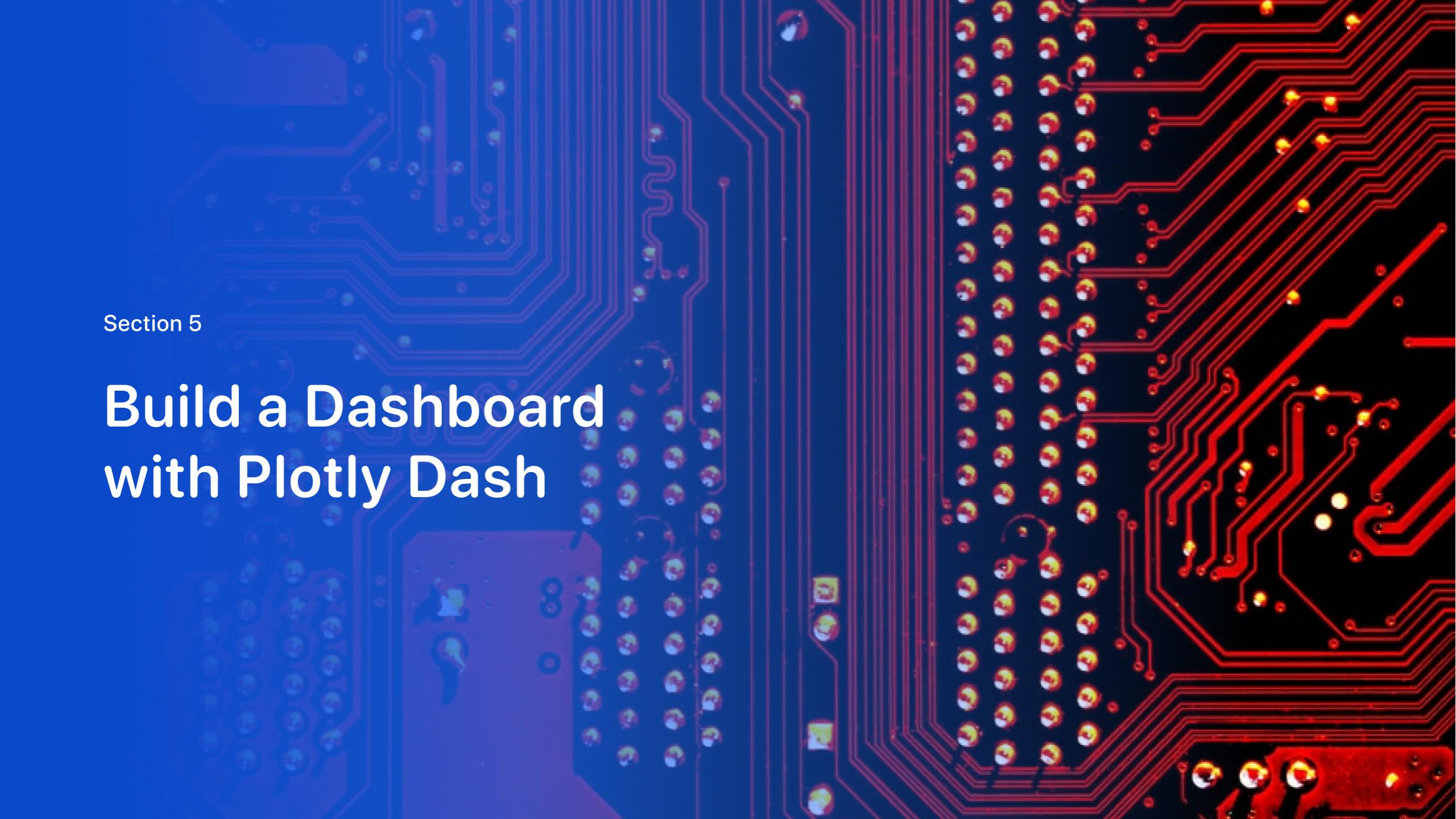


KSC LC-39A has the highest success rate out of all the launch sites, while CCAFS LC-40 has the lowest. They are both located on the east coast.

Launch Site Proximity Map



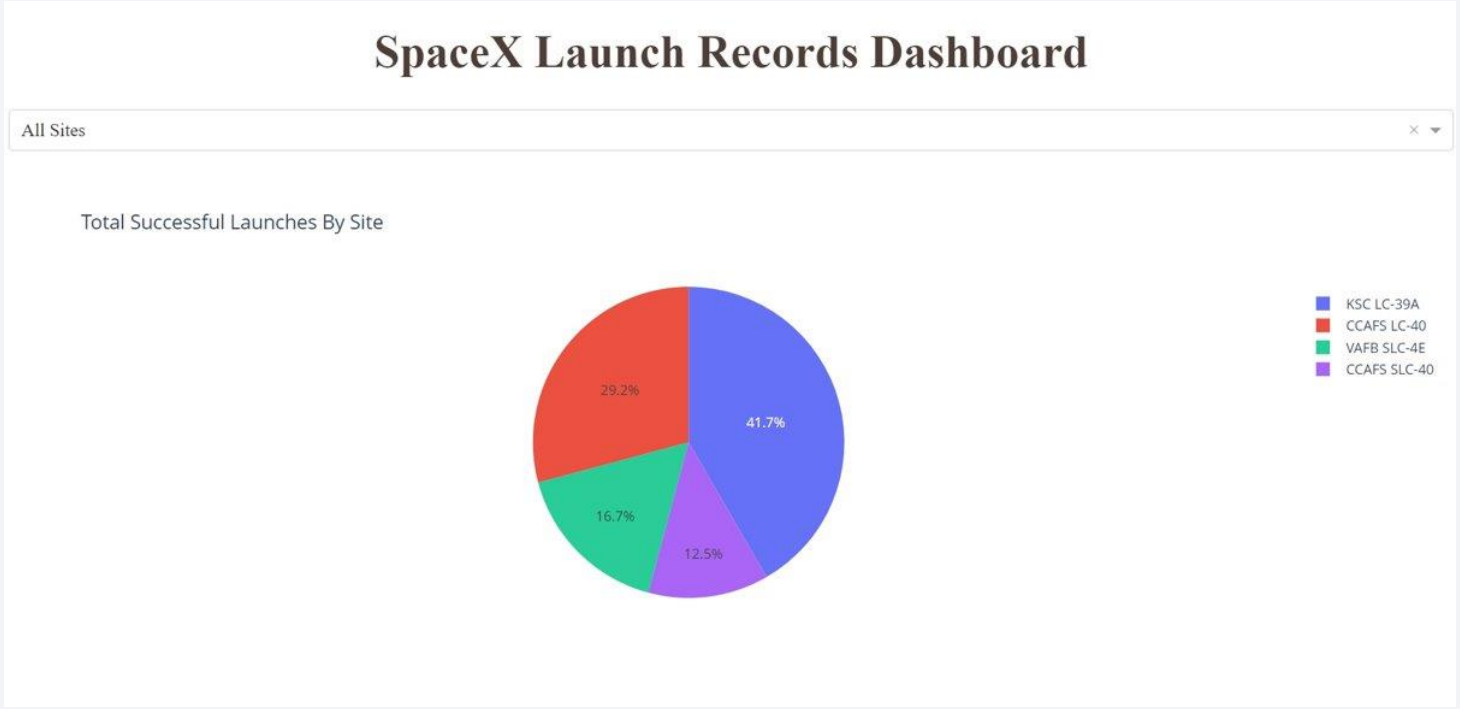
In this example, the launch site CCAFS SLC-40 is in close proximity to the highway, coastline and railway (the average distance is about 1 km). At the same time, the closest city is much further away – approximately 18 km.



Section 5

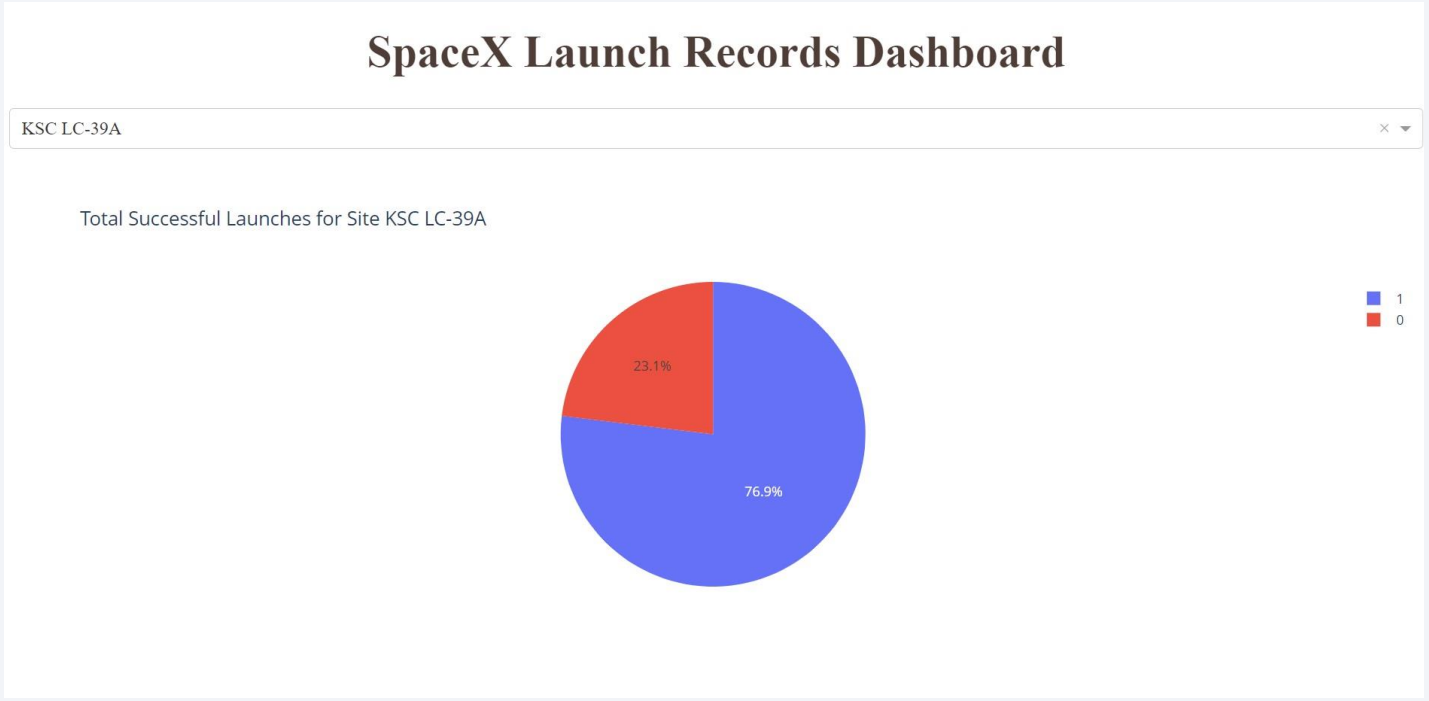
Build a Dashboard with Plotly Dash

Amount of Successful Launches by Site



Most successful first stage landings (about 42%) were launched from KSC LC-39A, while CCAFS LC-40 contributed the least with 12.5% of total successful launches.

Site with Highest Success Rate



On top of recording the highest count of successful launches across all sites, KSC LC-39A also has the highest success rate: the first stage landed successfully in close to 77% of launches from KSC LC-39A.

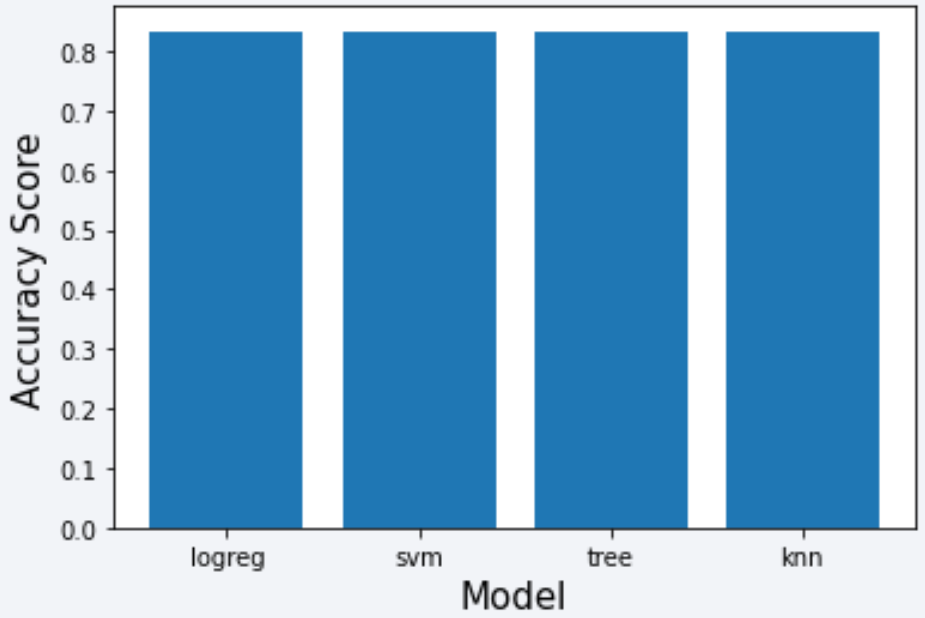
Payload Mass vs. First Stage Landing Outcome

The first stage did not land successfully in most launches with payload mass above 4000 Kg or below 2000 Kg.

Section 6

Predictive Analysis (Classification)

Classification Accuracy



All models practically perform the same. Their accuracy score is roughly 0.83 - the models accurately predict the first stage landing outcome 83% of the time.

Confusion Matrix

All models show the same confusion matrix. They return 3 false positives (Type II Error) out of 6 actual negative values, and no false negative.

Conclusions

- Launches directed to the ES-L1, GEO, HEO and SSO orbits have the highest Success Rate.
- Overall, Success Rate has been improving significantly in time. Most launches did not succeed in landing the first stage before the 20th attempt.
- KSC LC-39A is the launch site with the highest count of successful first stage landings as well as the highest Success Rate.
- Payloads with mass between 2000 and 4000 Kg tend to perform best.
- Logistic Regression, Support Vector Machine, Decision Tree Classifier and K-Nearest Neighbour models perform the same with this dataset. This is likely due to the limited size of the data. They can accurately predict about 83% of first stage landing outcomes on the test data. All models tend to return false positives.

Thank you!