IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Dmytro Pashniev
14.06.2024

Outline

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

Executive Summary

In this research we collected data from SpaceX API and by web scrapping Falcon 9 historical launch records from Wikipedia, preprocessed it by deleting null results and set proper format, wrangled it by simple Exploratory Data Analysis and determined training labels. Then we performed further Exploratory Data Analysis with visualization, SQL, interactive maps and dashboards. At the end, using Machine Learning Prediction we found best hyperparameters for SVM, Classification Trees and Logistic Regression, and revealed the method performs best using test data.

For all explorations we used programing language Python, libraries: pandas, numpy, BeautifulSoup, matplotlib, plotly, seaborn, Folium, Plotly Dash, sklearn etc. We worked in Jupiter notebooks and IDE on the Skills Network Labs platform.

As the result we got an efficient mechanism which allows to predict success landing of the first stage with given characteristics of the mission.

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 the savings is because SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

In this research we will collect data and wrangling it, perform Exploratory Data Analysis to find some patterns in it and determine what would be the label for training supervised models, create predictive models and find the best one.

Thus, we will predict if the Falcon 9 first stage will land successfully.

Section 1

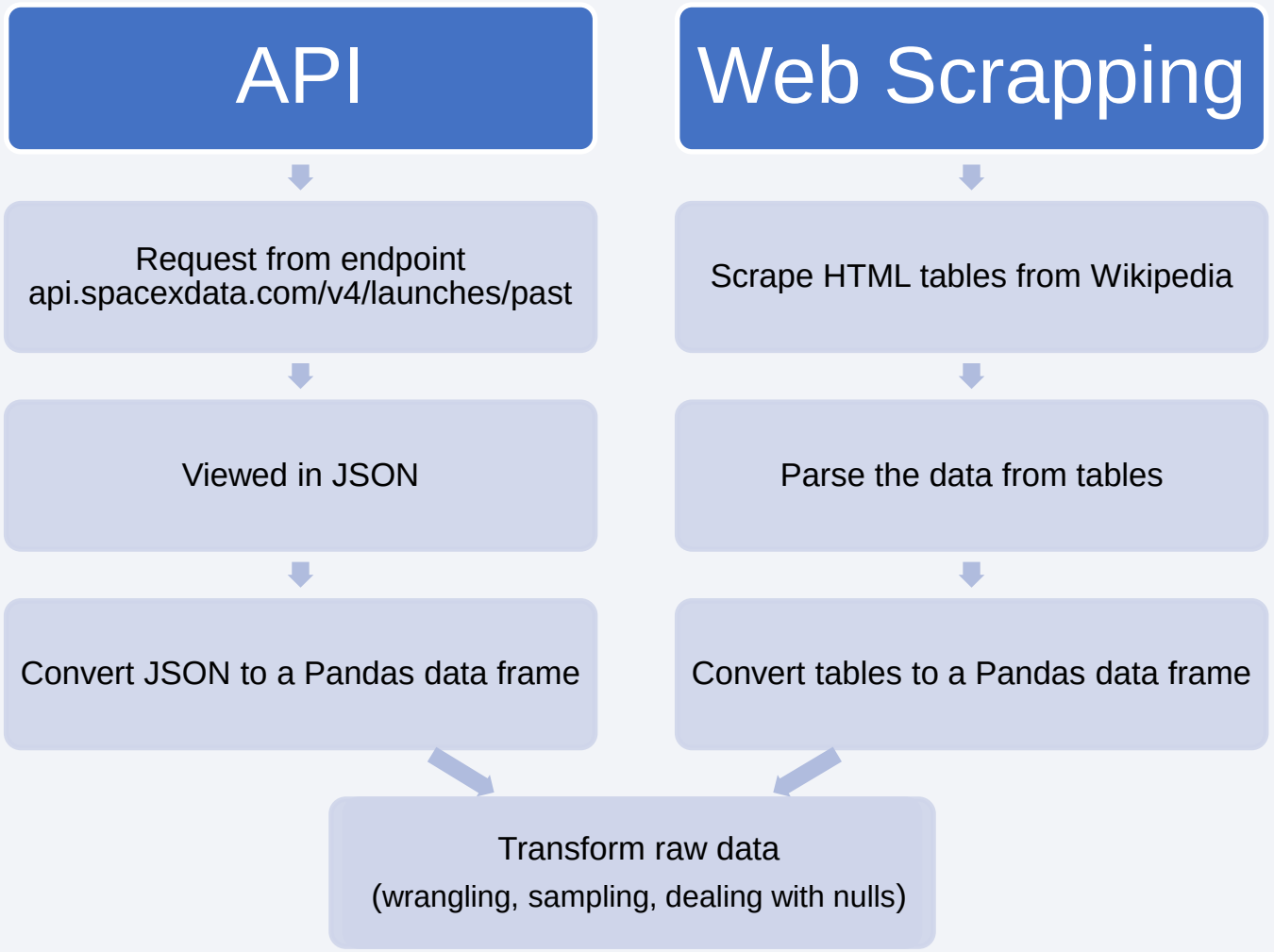
Methodology

Methodology

Executive Summary

- Data collection methodology:
  - request to the SpaceX API and clean the requested data
  - perform web scraping to collect Falcon 9 historical launch records from a Wikipedia page titled `List of Falcon 9 and Falcon Heavy launches`
- Perform data wrangling
  - determine the distinct kinds of outcomes,
  - create a set of outcomes where the second stage did not land successfully
  - using the set, add to the dataframe a classification column that represents the outcome of each launch: if its value is zero, the first stage did not land successfully; one means the first stage landed successfully
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using Machine Learning

Data Collection (explaination on the next page)



Data Collection

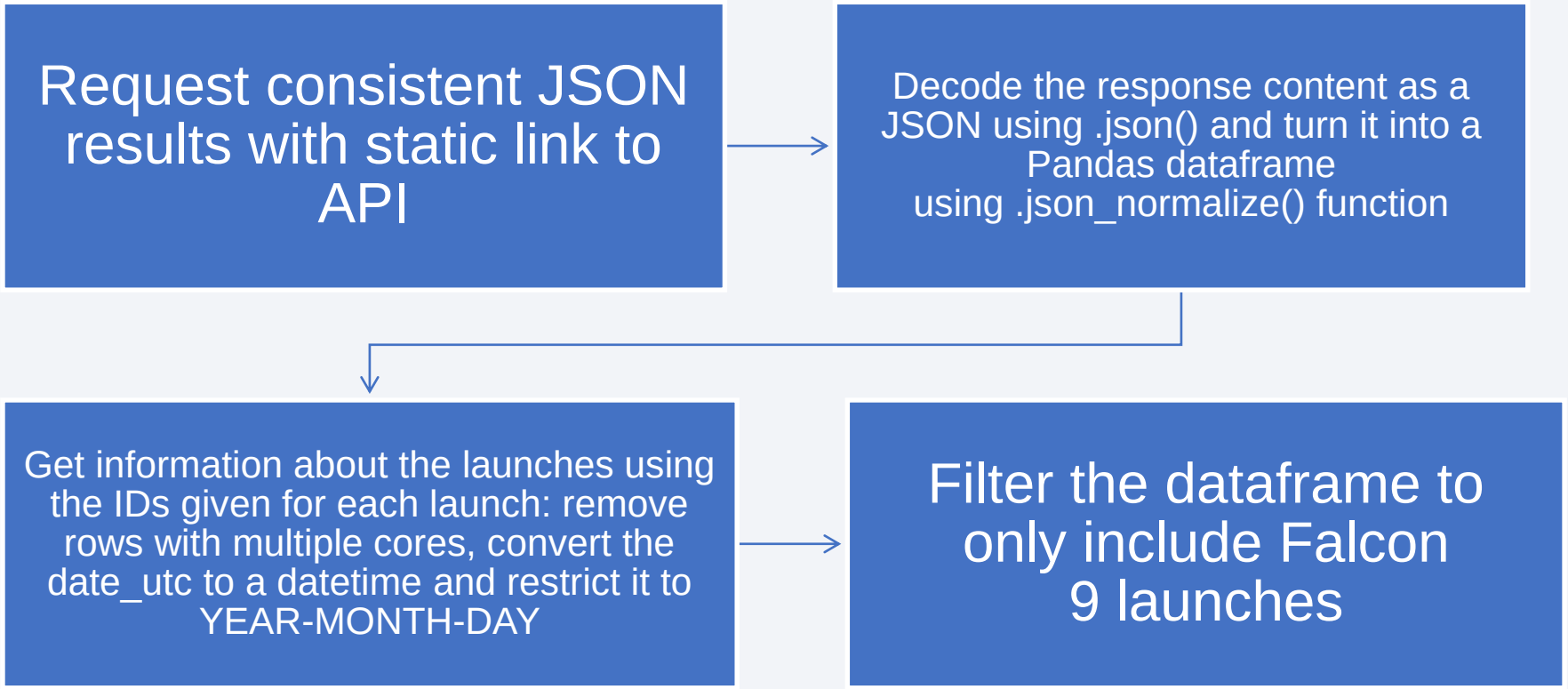
First, the SpaceX launch data was gathered from the **SpaceX REST API** (the rocket used, payload delivered, launch specifications, landing specifications, landing outcome etc.).
We worked with the the SpaceX REST API endpoint api.spacexdata.com/v4/launches/past.
We got past launch data by the requests library and viewed it by calling the .json() method.
Then we converted this JSON to a dataframe with the json_normalize function.
At the end we got a table form of the past launch data.

Second, we use another popular data source for obtaining Falcon 9 Launch data **- web scraping** related Wiki pages.
Using the Python BeautifulSoup package we web scraped some HTML tables that contain valuable Falcon 9 launch records. Then we parsed the data from those tables and convert them into a Pandas data frame for further visualization and analysis.

Finally, we transformed this raw data into a clean dataset which provides meaningful data on the situation we are trying to address: Wrangling Data using an API, Sampling Data, and Dealing with Nulls.

Data Collection – SpaceX API

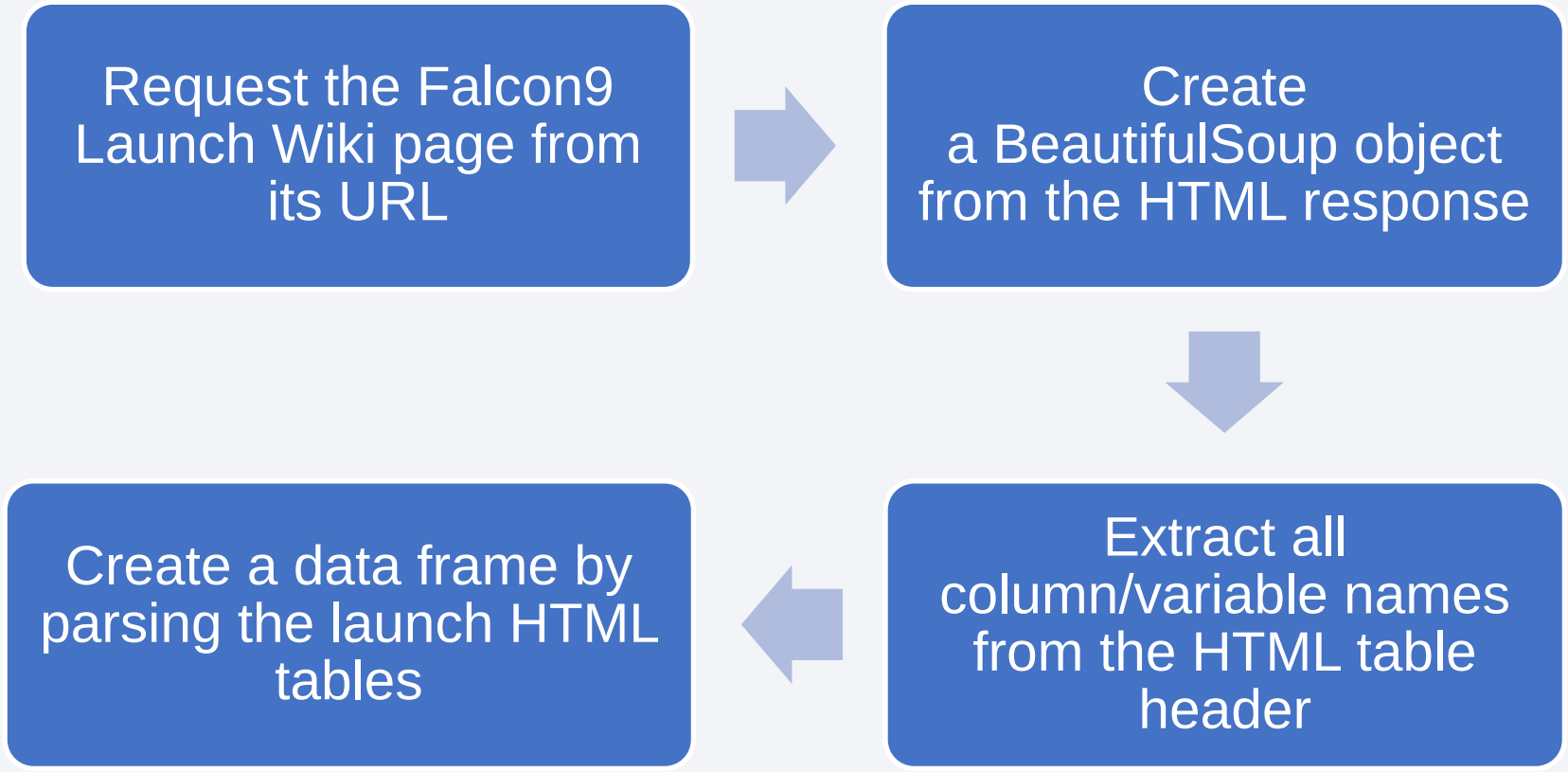
Data collection with SpaceX REST calls



The completed SpaceX API calls notebook - GitHub URL

Data Collection - Scraping

Web scraping process



The completed web scraping notebook - GitHub URL

Data Wrangling

Data wrangling process

Missing Values

- calculate below the mean for the PayloadMass using the .mean()
- use the mean and the .replace() function to replace `np.nan` values in the data with the mean calculated

Launches

- use the method value_counts() on the column LaunchSite to determine the number of launches on each site

Orbits

- Use the method.value_counts() to determine the number and occurrence of each orbit in the column Orbit

Outcomes

- Create a landing outcome label “Class” from Outcome column: if the value is zero, the first stage did not land successfully; one means the first stage landed Successfully

The completed data wrangling related notebook - GitHub URL

EDA with Data Visualization

We plotted the following charts on received data frame:

- the scatter plot FlightNumber vs. PayloadMass – to see their relationship;
- the scatter plot FlightNumber vs. LaunchSite – to see their relationship;
- the scatter plot PayloadMass vs. LaunchSite with hue to be the Class value – to visualize the relationship between Payload and Launch Site;
- bar chart for the success rate of each orbit - to visually check if there are any relationship between success rate and orbit type;
- the scatter plot FlightNumber vs. Orbit with hue to be the Class value – to visualize the relationship between FlightNumber and Orbit type;
- the scatter plot Payload vs. Orbit with hue to be the Class value - to reveal the relationship between Payload and Orbit type;
- the line chart year vs. success rate - to get the average launch success trend.

The completed EDA with data visualization notebook - GitHub URL

EDA with SQL

We performed the following SQL queries to SpaceX table:

- To display the names of the unique launch sites in the space mission - SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE; response - CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, CCAFS SLC-40;
- To display 5 records where launch sites begin with the string 'CCA‘ - SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5;
- To display the total payload mass carried by boosters launched by NASA (CRS) - SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTABLE WHERE Customer LIKE 'NASA (CRS)‘; response – 45596;
- To list the date when the first succesful landing outcome in ground pad was achieved - SELECT MIN("Date") FROM SPACEXTABLE WHERE "Landing_Outcome" LIKE '%Success%'; response – 2015-12-22;
- To list the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000 - SELECT DISTINCT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome" LIKE 'Success (drone ship)' AND "PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MASS__KG_" < 6000;

The completed EDA with SQL notebook - GitHub URL

EDA with SQL

We performed the following SQL queries to SpaceX table (continue):

- The total number of successful and failure mission outcomes - SELECT 'Success outcomes' as 'Kind', COUNT("Date") FROM SPACEXTABLE WHERE "Landing_Outcome" LIKE '%Success%' union SELECT 'Failure outcomes' as 'Kind', COUNT("Date") FROM SPACEXTABLE WHERE "Landing_Outcome" LIKE '%failure%‘;
- The names of the booster_versions which have carried the maximum payload mass. Use a subquery - SELECT DISTINCT "Booster_Version" FROM SPACEXTABLE WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTABLE);
- The records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015 - SELECT DISTINCT substr("Date", 6,2) as 'Month', "Landing_Outcome", "Booster_Version", "Launch_Site" FROM SPACEXTABLE WHERE substr("Date", 0, 5) = '2015' AND "Landing_Outcome" LIKE 'Failure (drone ship)‘;
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order - SELECT "Landing_Outcome", COUNT("Date") FROM SPACEXTABLE WHERE ("Landing_Outcome" LIKE '%success%' OR "Landing_Outcome" LIKE '%failure%') GROUP BY "Landing_Outcome" ORDER BY COUNT("Date") DESC

The completed EDA with SQL notebook - GitHub URL

Build an Interactive Map with Folium

We added the following map objects to a folium map:

- a blue circle at NASA Johnson Space Center's coordinate with a icon showing its name – to mark its location;
- a Circle object based on its coordinate (Lat, Long) values for each launch site with its name as a popup label – to mark their locations;
- markers of the launch outcomes for each site – to see which sites have high success rates;
- a MousePosition – to get coordinate for a mouse over a point on the map;
- markers on closest to CCAFS LC-40 coastline (city, railway, highway) points on the map and a polyLine using the points coordinates and launch site coordinates – to see how close launch sites are proximity to those points/

The completed interactive map with Folium map notebook - GitHub URL

Build a Dashboard with Plotly Dash

We have added to a dashboard:

- A pie chart to show the total successful launches count for all sites. If a specific launch site was selected, show the Success vs. Failed counts for the site
- A scatter chart to show the correlation between payload and launch success with a slider to select payload range

After visual analysis using the dashboard, we should be able to obtain some insights to answer the following five questions:

- Which site has the largest successful launches?
- Which site has the highest launch success rate?
- Which payload range(s) has the highest launch success rate?
- Which payload range(s) has the lowest launch success rate?
- Which F9 Booster version has the highest launch success rate?

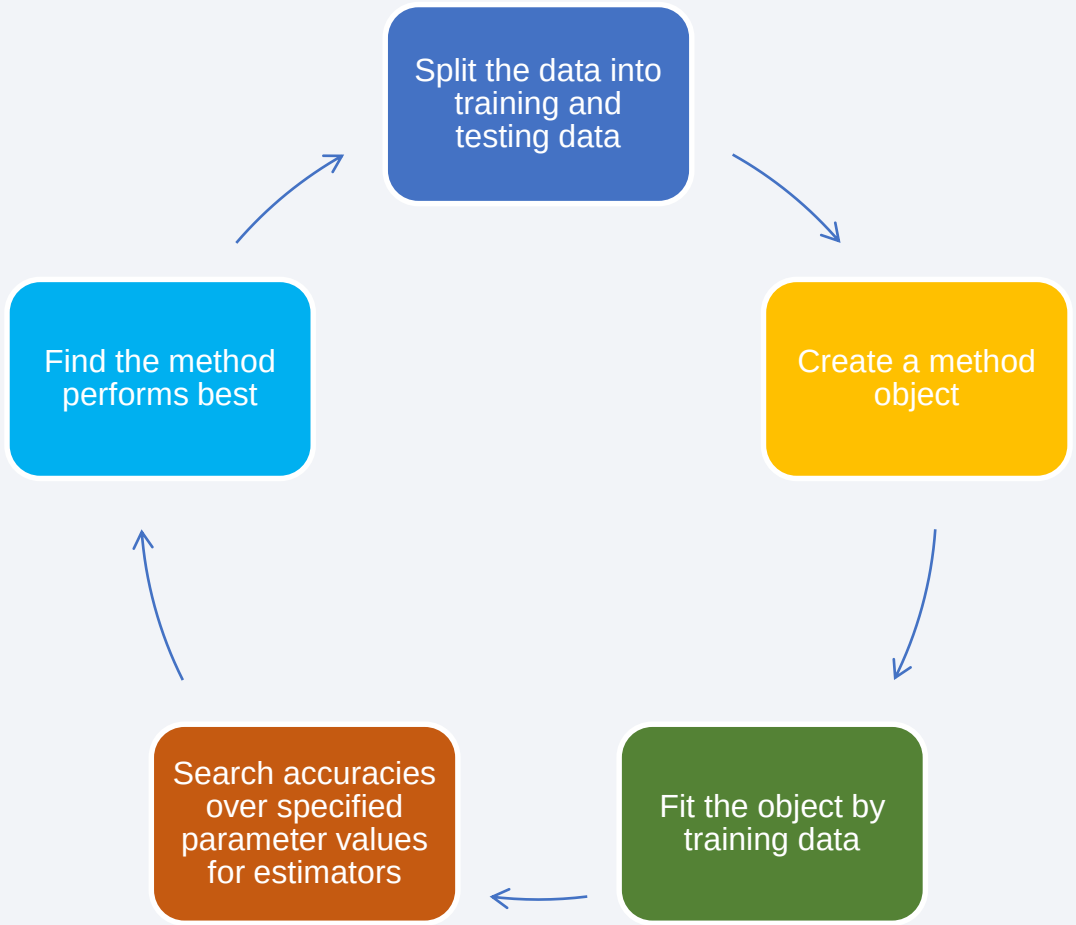
The completed interactive map with Folium map notebook - GitHub URL

Predictive Analysis (Classification)

We built, evaluated, improved, and found the most efficient classification model using the process presented in the flowchart.

Thus, we tested the above methods:

- Logistic Regression
- Support Vector Machines
- Decision Tree
- K-neighbors



The completed interactive map with Folium map notebook - GitHub URL

Results

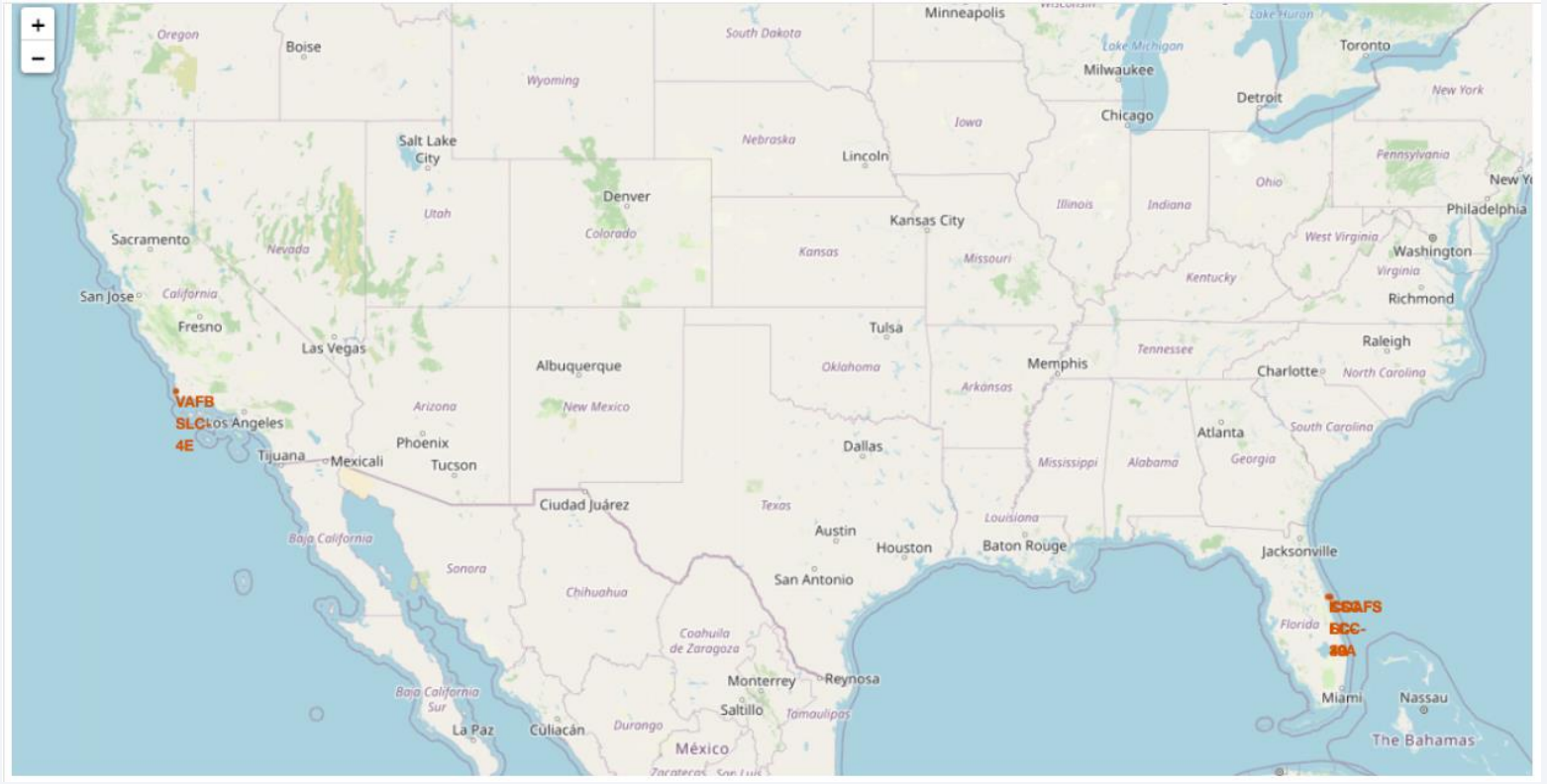
Exploratory data analysis results:

- the common success rate is equal to 0.66;
- as the flight number increases, the first stage is more likely to land successfully. The payload mass is also important; it seems the more massive the payload, the less likely the first stage will return;
- different launch sites have different success rates: CCAFS LC-40 - 60 %, KSC LC-39A and VAFB SLC 4E - 77%;
- for the VAFB-SLC launch site there are no rockets launched for heavy payload mass (greater than 10000)
- the following orbits have high success rate: ES-L1, SSO, HEO, GEO;
- in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit;
- with heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS; however for GTO we cannot distinguish this well as both positive landing rate and negative landing (unsuccessful mission) are both there here;
- the success rate since 2013 kept increasing till 2020.

Results

Interactive analytics results with map:

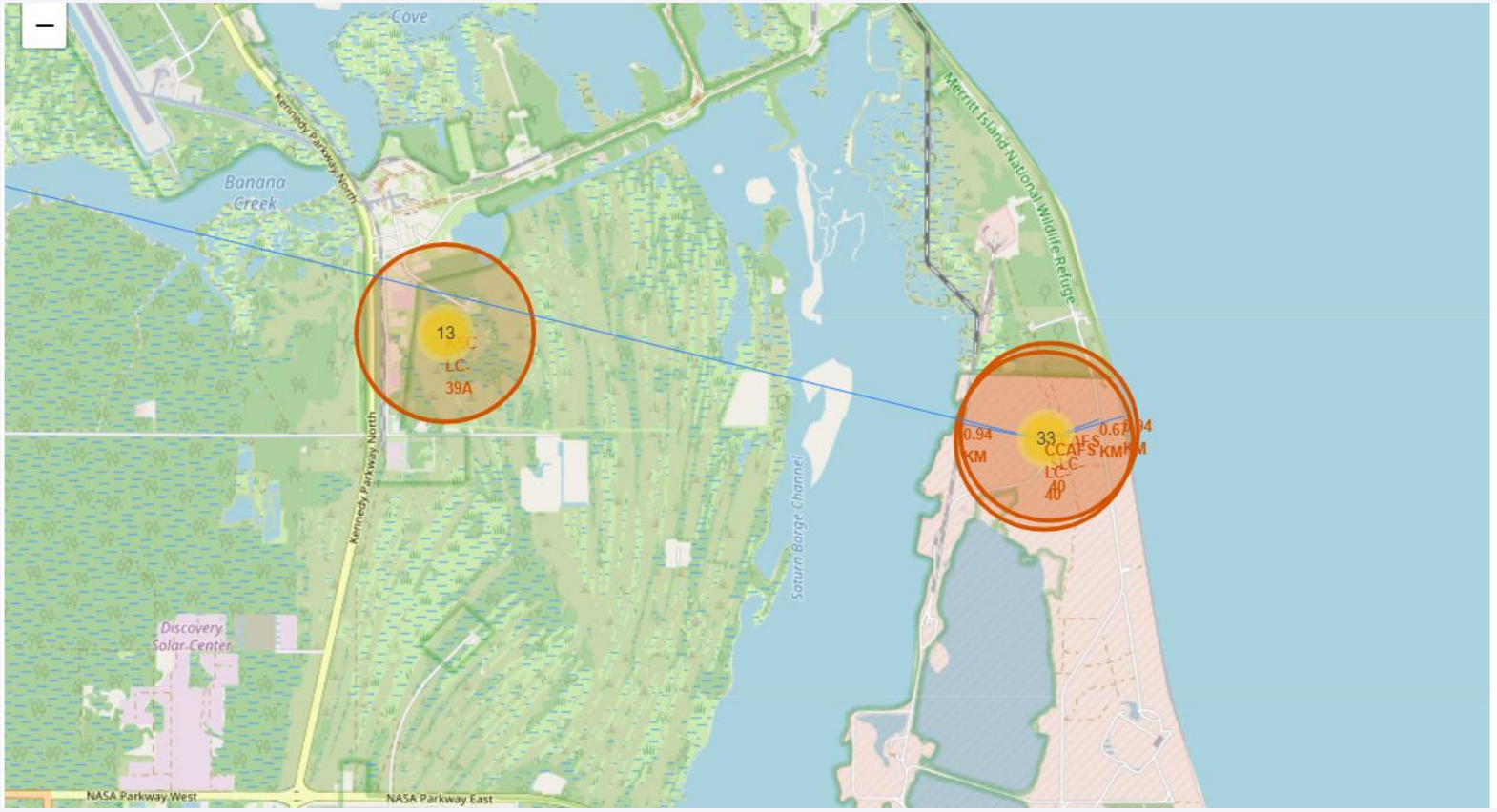
- all launch sites are in proximity to the Equator line because this allows less energy to be used to launch a rocket into orbit
- all launch sites are in very close proximity to the coast because the return sites of the first stage are located on the water to reduce the consequences of a possible accident



Results

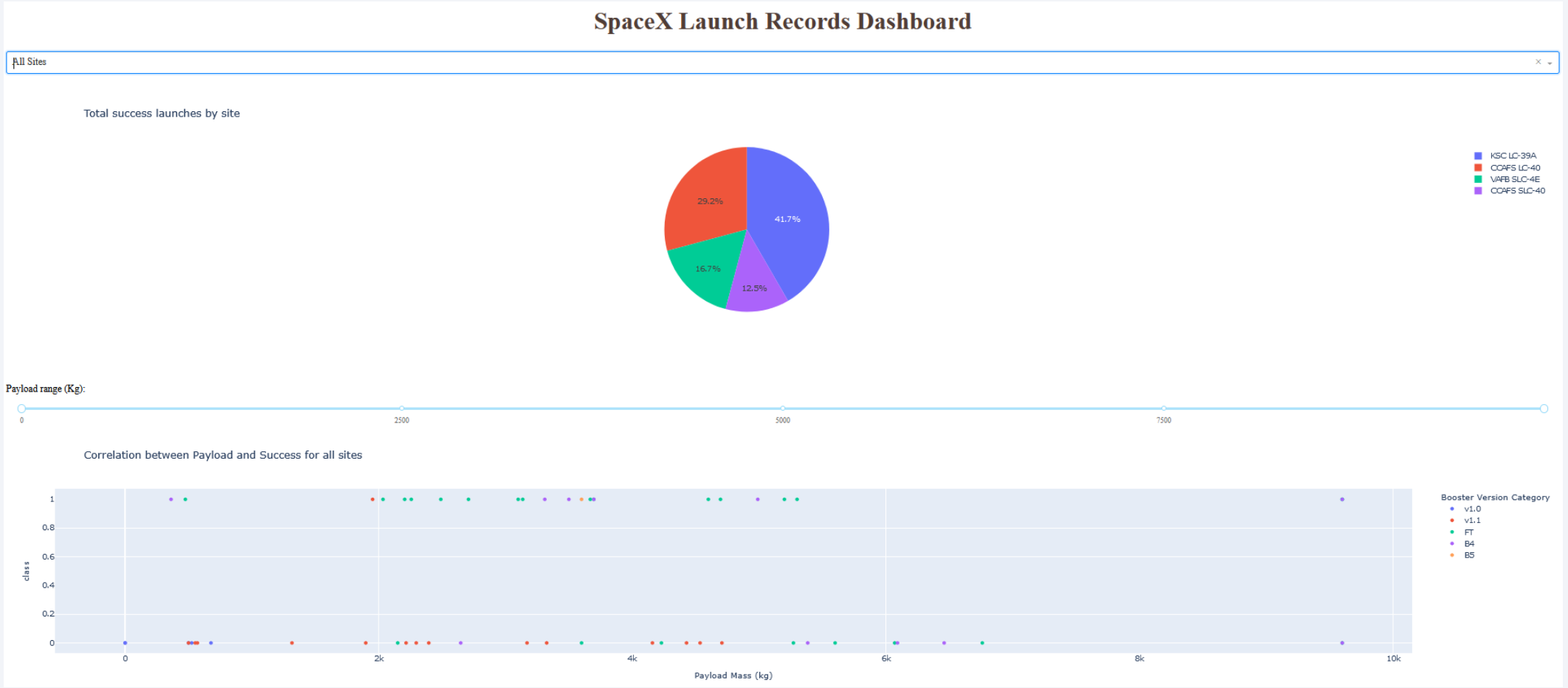
Interactive analytics results with map:

- launch sites are in close proximity to railways and highways to reduce logistics costs
- launch sites are kept at a certain distance from cities to prevent damage to people



Results

Interactive analytics results with dashboard:



Results

Predictive analysis results:

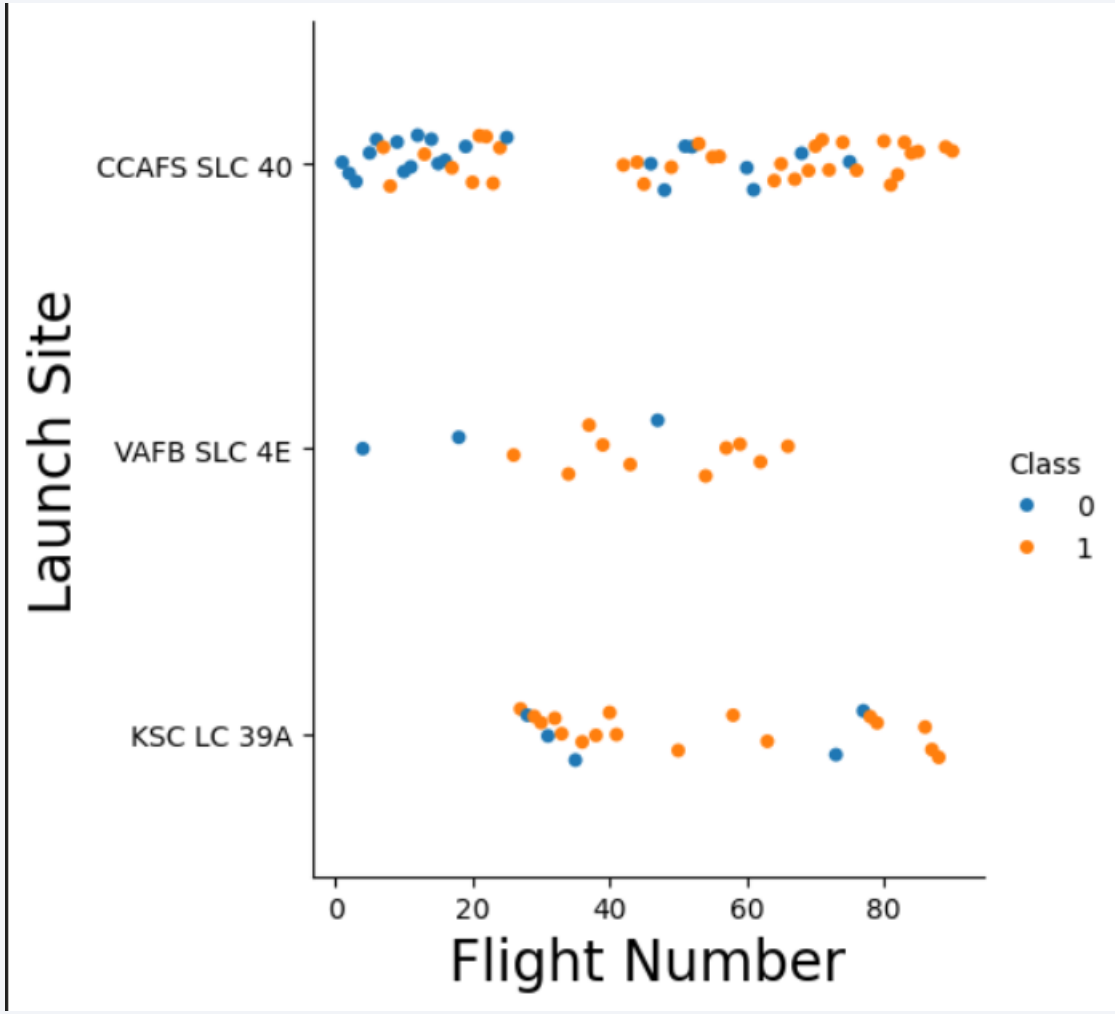
- Logistic Regression method showed the best results with inverse of regularization strength equal 0.01, penalty - 'l2' and solver – 'lbfgs';
- Support Vector Machines got the best results with inverse of regularization strength equal 1.0, gamma - 0.03, kernel = 'sigmoid'
- Decision Tree method showed the best results with 'criterion': 'entropy', 'max_depth': 16, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 10, 'splitter': 'random'
- K-neighbors method got the best results with 'algorithm': 'auto', 'n_neighbors': 10, 'p': 1
- Logistic Regression, Support Vector Machines and K-neighbors estimators had accuracy 0.83, Decision Tree method - 0.94

Section 2

Insights drawn from EDA

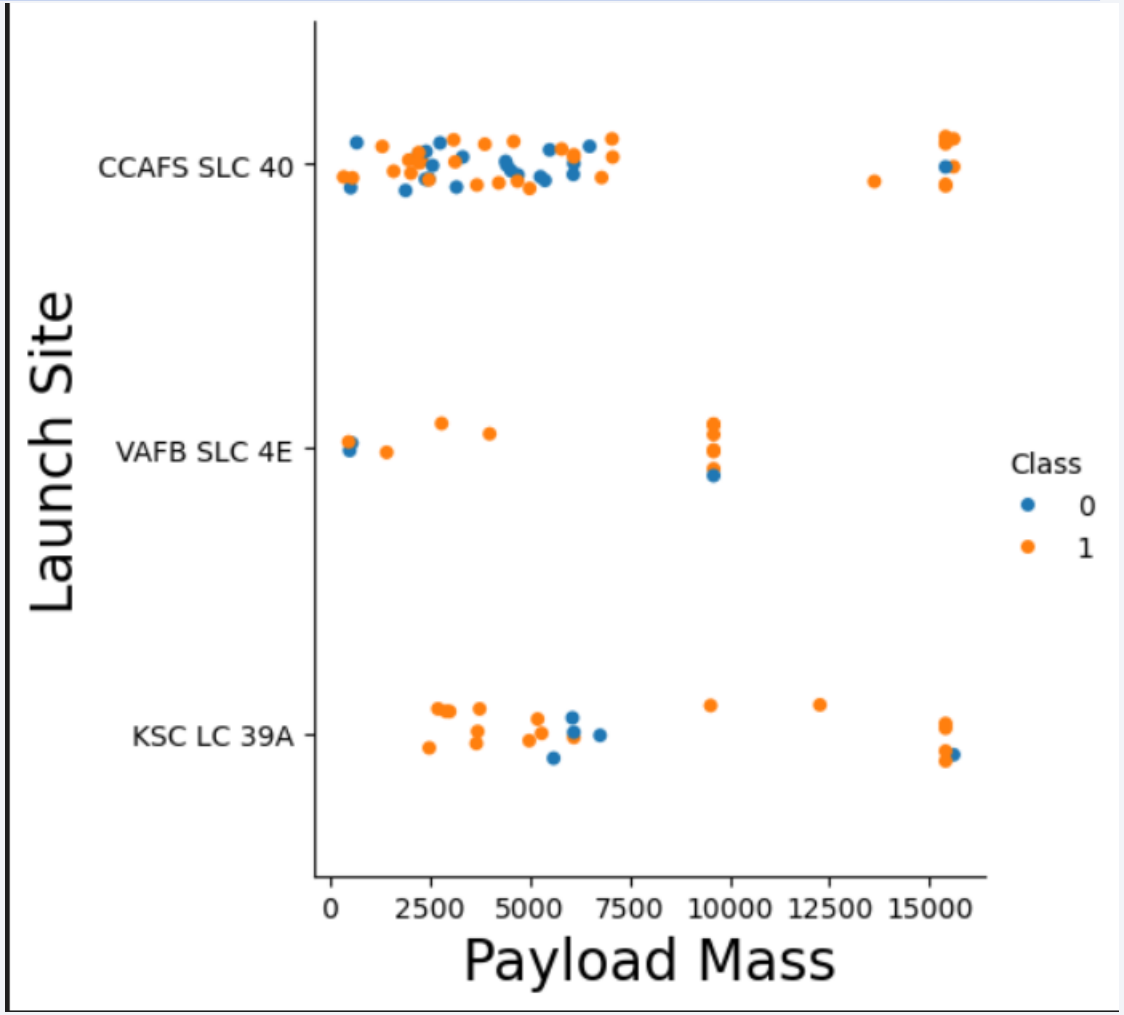
Flight Number vs. Launch Site

In the Flight Number vs. Launch Site scatter point plots we found that as flight number increases, the number of successful first stage returns at all launch sites increases



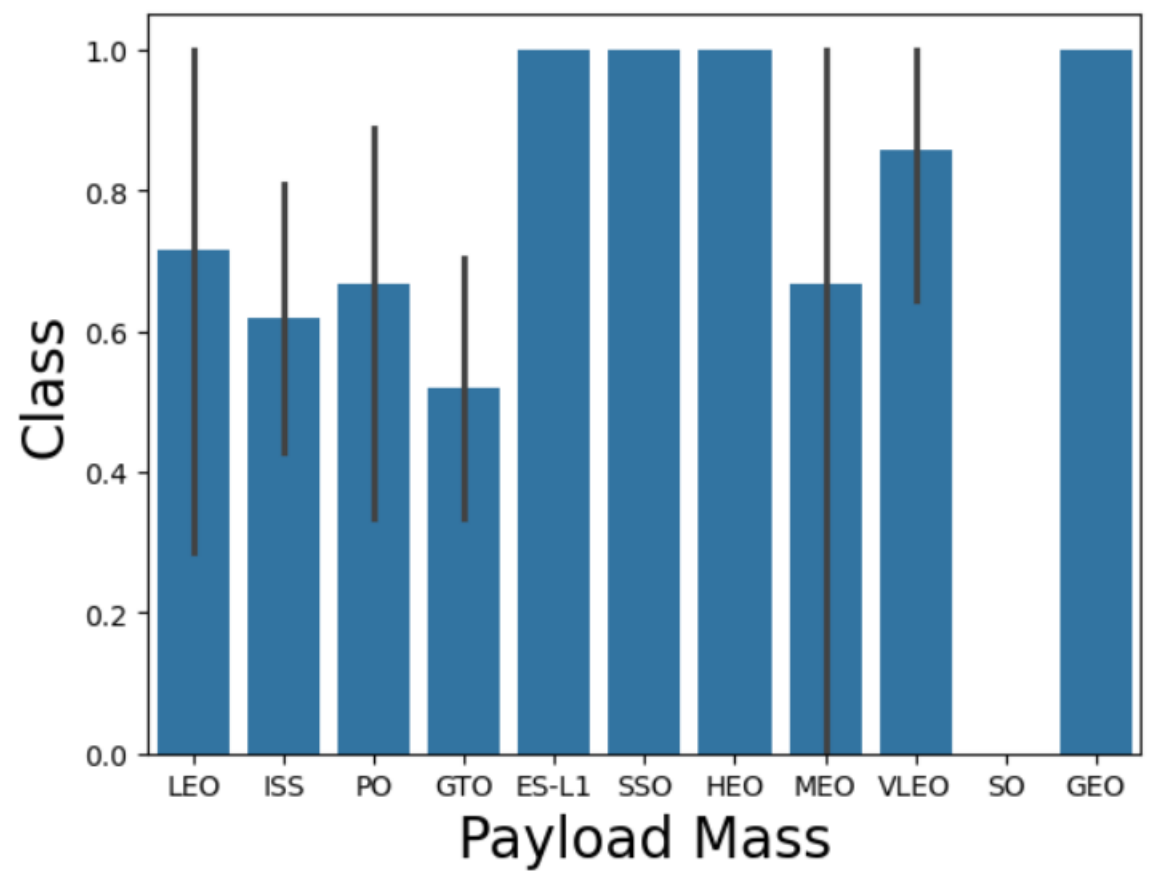
Payload vs. Launch Site

If you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launch site there are no rockets launched for heavy payload mass (greater than 10000)



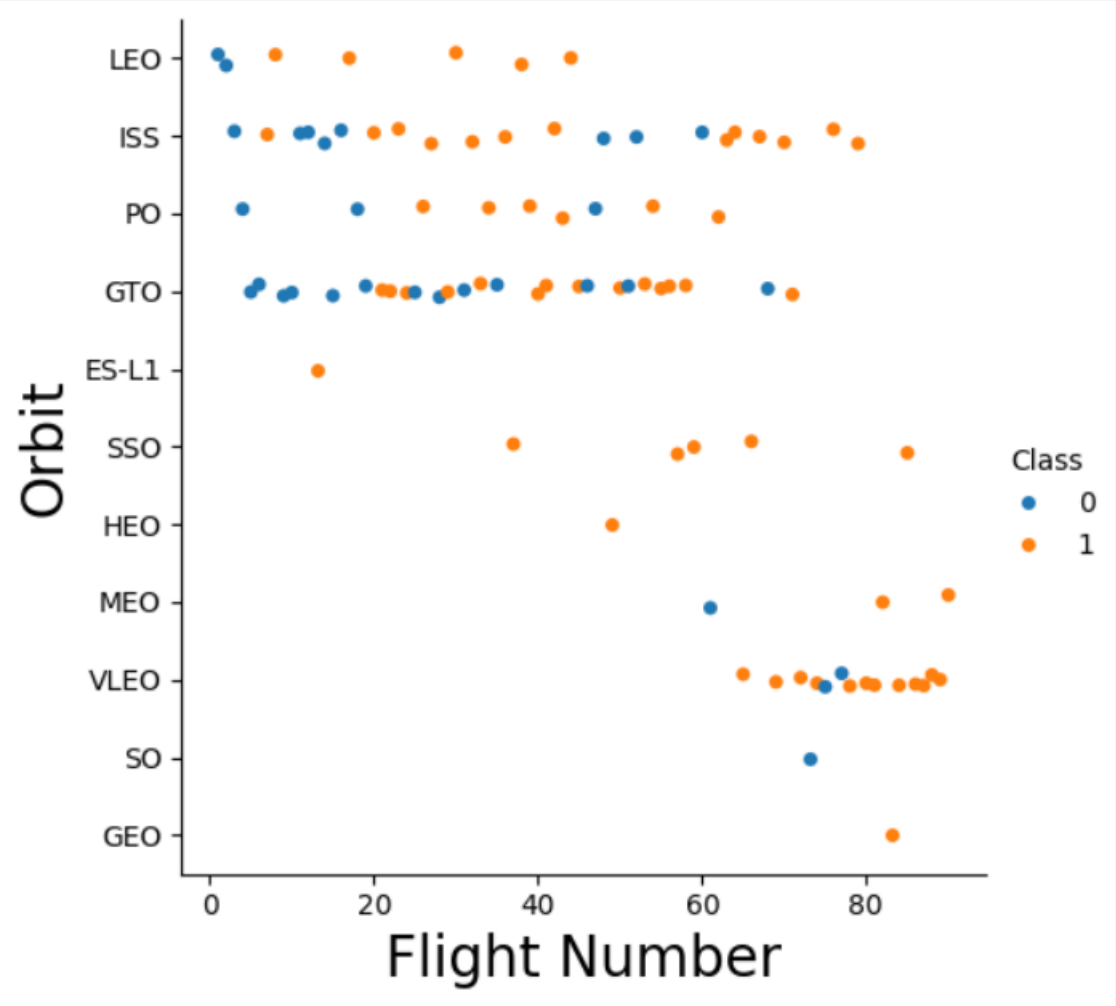
Success Rate vs. Orbit Type

With the bar chart for the success rate of each orbit type we can see which orbits have high success rate: ES-L1, SSO, HEO, GEO



Flight Number vs. Orbit Type

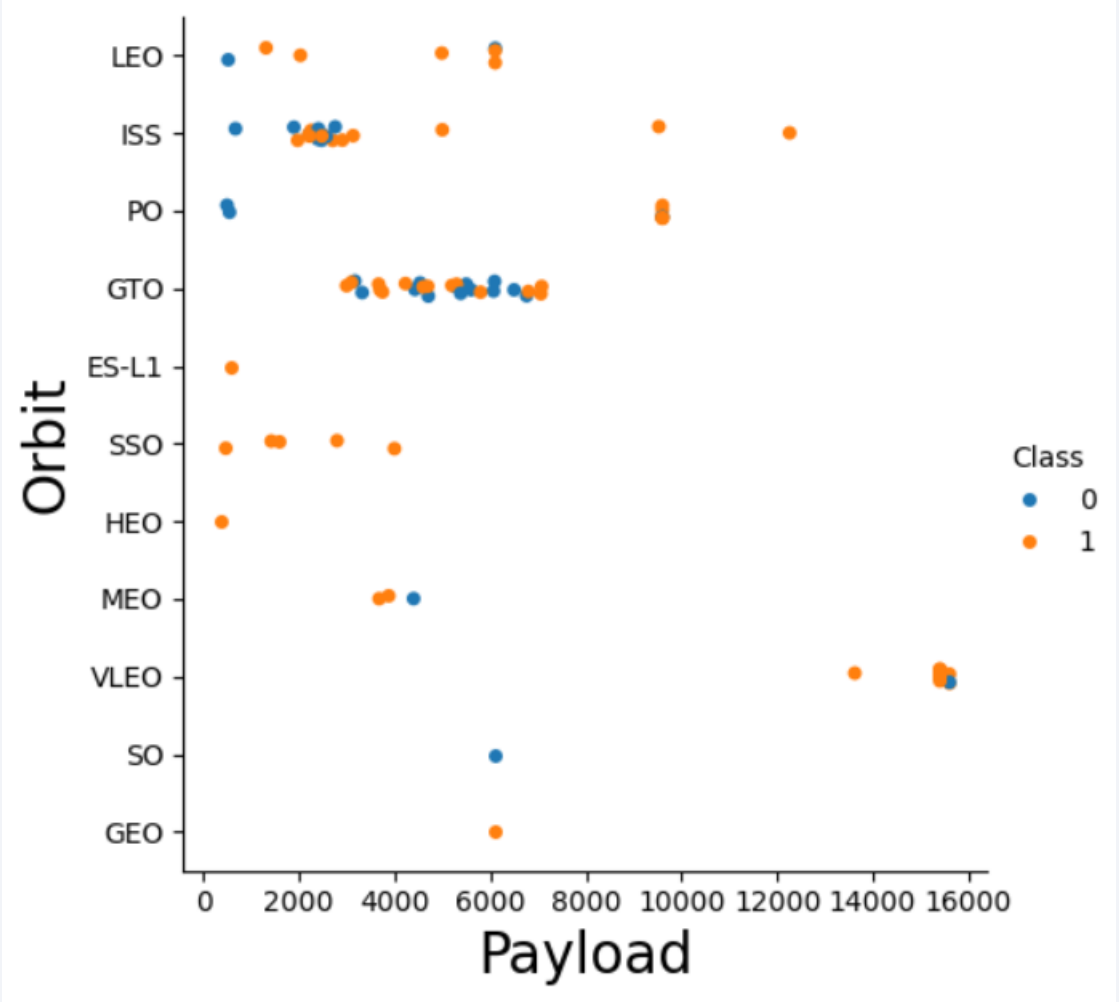
The scatter plot of the number of flights and orbit type shows us that orbits with a high success rate have an insignificant number of launches, suggesting that we cannot say that orbit type affects the success of a first stage return.



Payload vs. Orbit Type

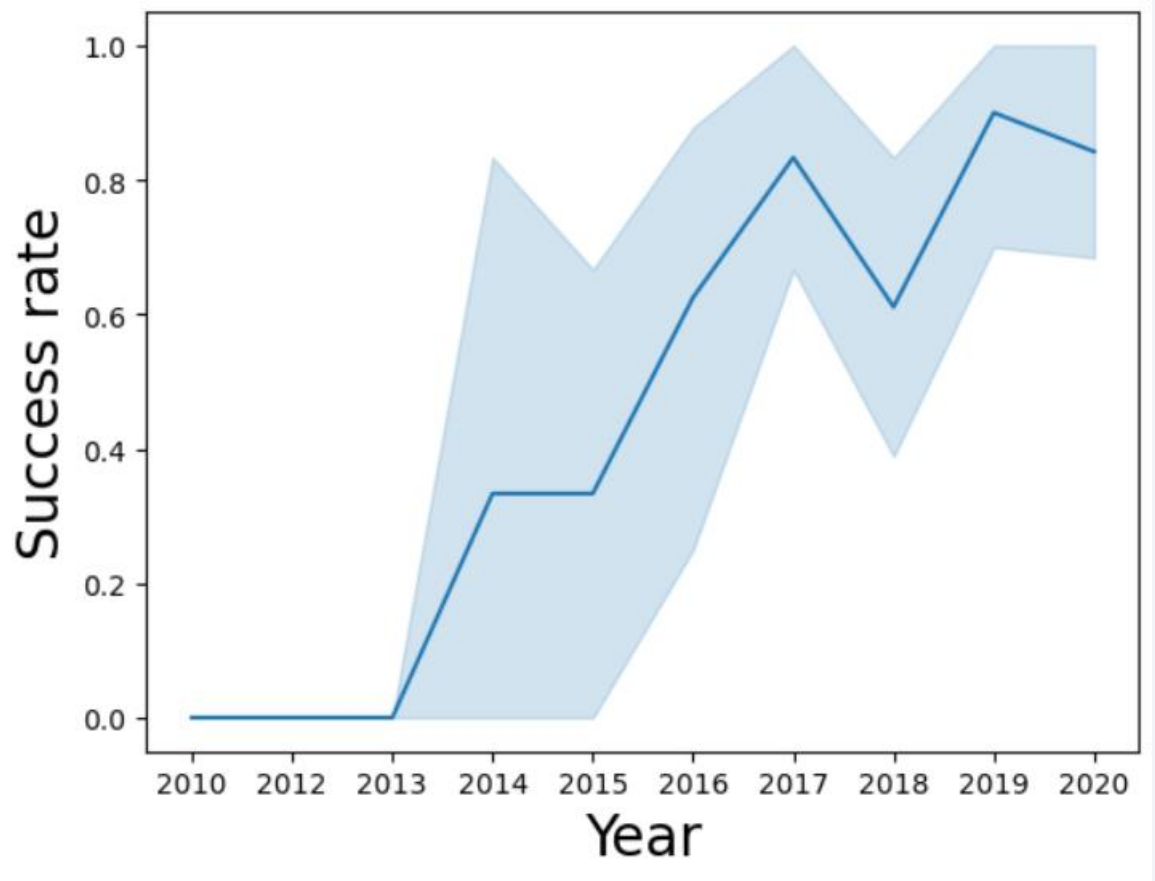
With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.

However for GTO we cannot distinguish this well as both positive landing rate and negative landing (unsuccessful mission) are both there here



Launch Success Yearly Trend

You can observe that the success rate since 2013 kept increasing till 2020



All Launch Site Names

Our query for the names of the unique launch sites returned us four results:

- Cape Canaveral Launch Complex 40
- Cape Canaveral Space Launch Complex 40
- Vandenberg Space Launch Complex 4E
- Kennedy Space Center Launch Complex 39A

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

Launch Site Names Begin with 'CCA'

5 records where launch sites begin with `CCA` means that launch site is a launch pads for rockets located at the north end of Cape Canaveral Space Force Station, Florida

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

Total Payload Mass

The total payload carried by the boosters from NASA is 45596 kg, which is the sum of the payloads on the rockets launched by the Customer as NASA (CRS) for all time

SUM("PAYLOAD_MASS__KG_")
45596

Average Payload Mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1 equals 2928.4 kg

AVG("PAYLOAD_MASS__KG_")
2928.4

First Successful Ground Landing Date

The date of the first successful landing outcome on ground pad is December 22, 2015 means that that result of the first stage return is possible almost 9 years

MIN("Date")
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

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

Total Number of Successful and Failure Mission Outcomes

The total number of successful and failure mission outcomes means that the number of successes is more than six times greater than the number of failures. Total success rate of launches with certain result of the first stage return equals 61 / (10+61) = 0.86

Kind	COUNT("Date")
Failure outcomes	10
Success outcomes	61

Boosters Carried Maximum Payload

List the names of the booster which have carried the maximum 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

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 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

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

Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

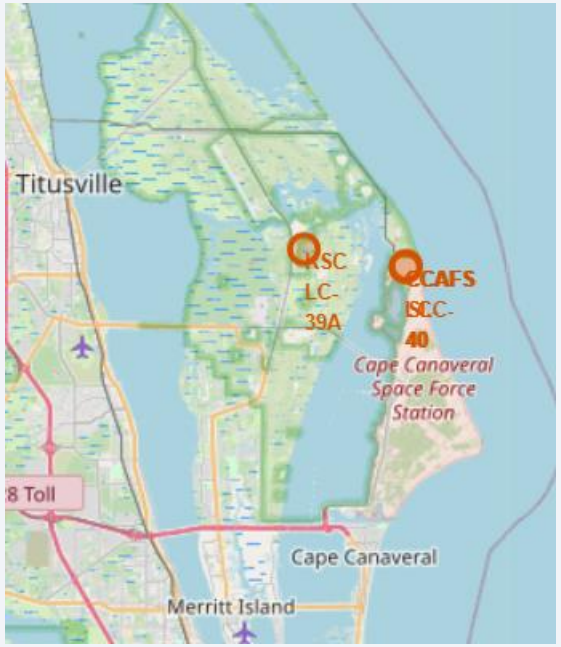
Landing_Outcome	COUNT("Date")
Success	38
Success (drone ship)	14
Success (ground pad)	9
Failure (drone ship)	5
Failure	3
Failure (parachute)	2

Section 3

Launch Sites Proximities Analysis

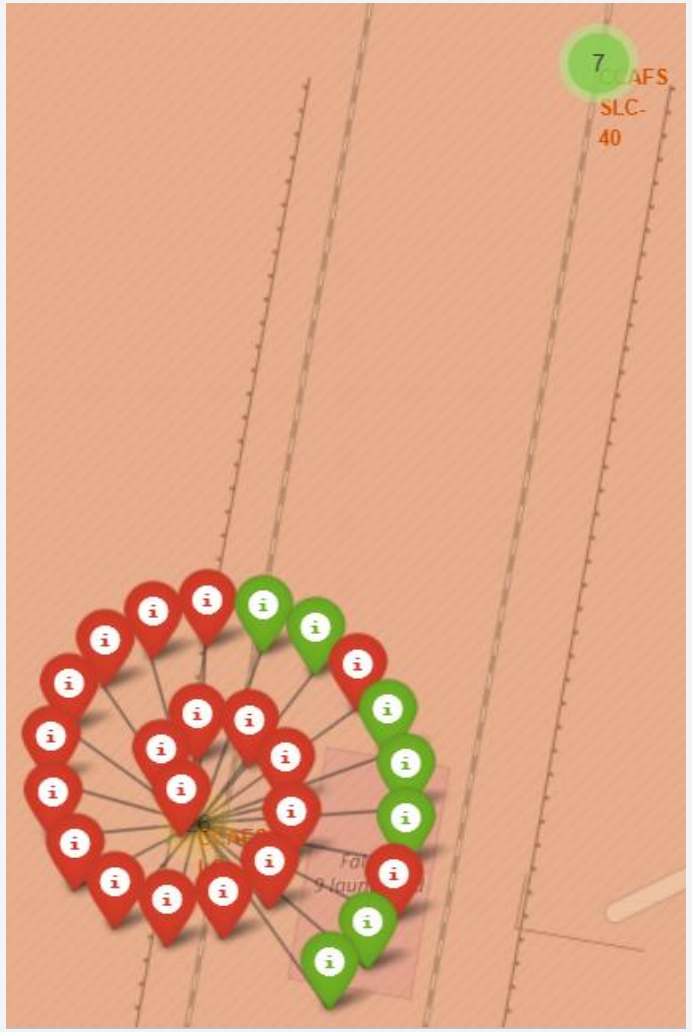
Launch sites’ locations

- Cape Canaveral Launch Complex 40 and Cape Canaveral Space Launch Complex 40 are launch pads for rockets located at the north end of Cape Canaveral Space Force Station, Florida.
- Kennedy Space Center Launch Complex 39A is the first of Launch Complex 39's three launch pads, located at NASA's Kennedy Space Center in Merritt Island, Florida.
- Vandenberg Space Launch Complex 4E is a launch and landing site at Vandenberg Space Force Base, California, U.S.
- all launch sites are in proximity to the Equator line because this allows less energy to be used to launch a rocket into orbit
- launch sites are in close proximity to railways and highways to reduce logistics costs
- launch sites are kept at a certain distance from cities to prevent damage to people
- all launch sites are in very close proximity to the coast because many of the return sites for the first stage are located on the water to reduce the consequences of a possible accident



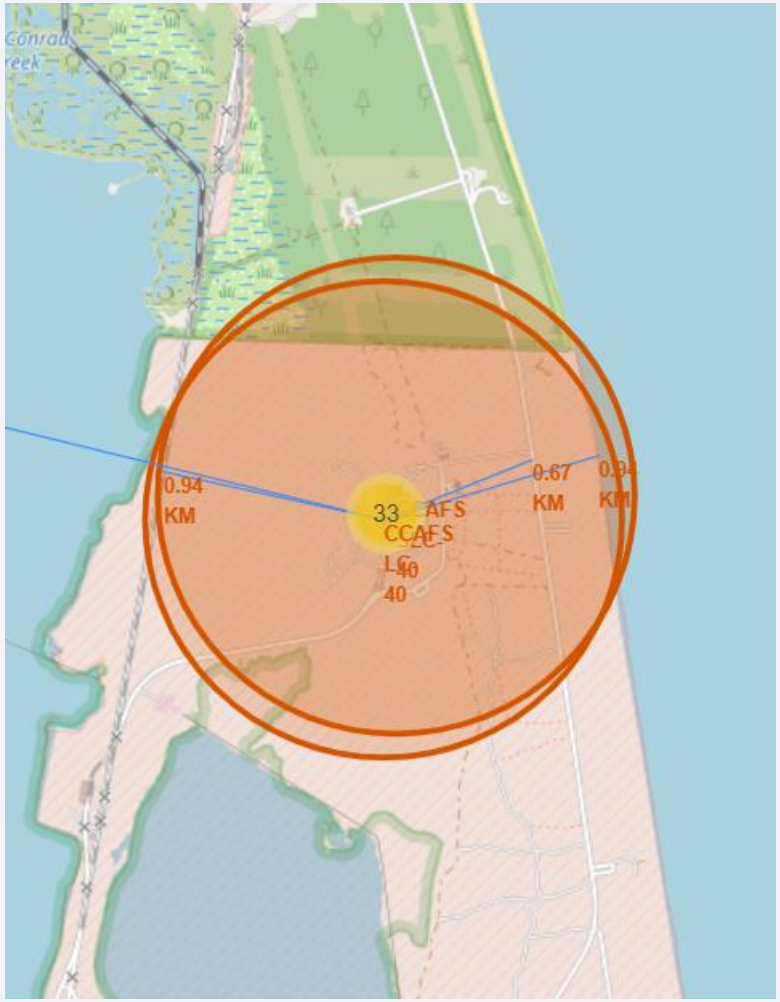
The success/failed launches for each site on the map

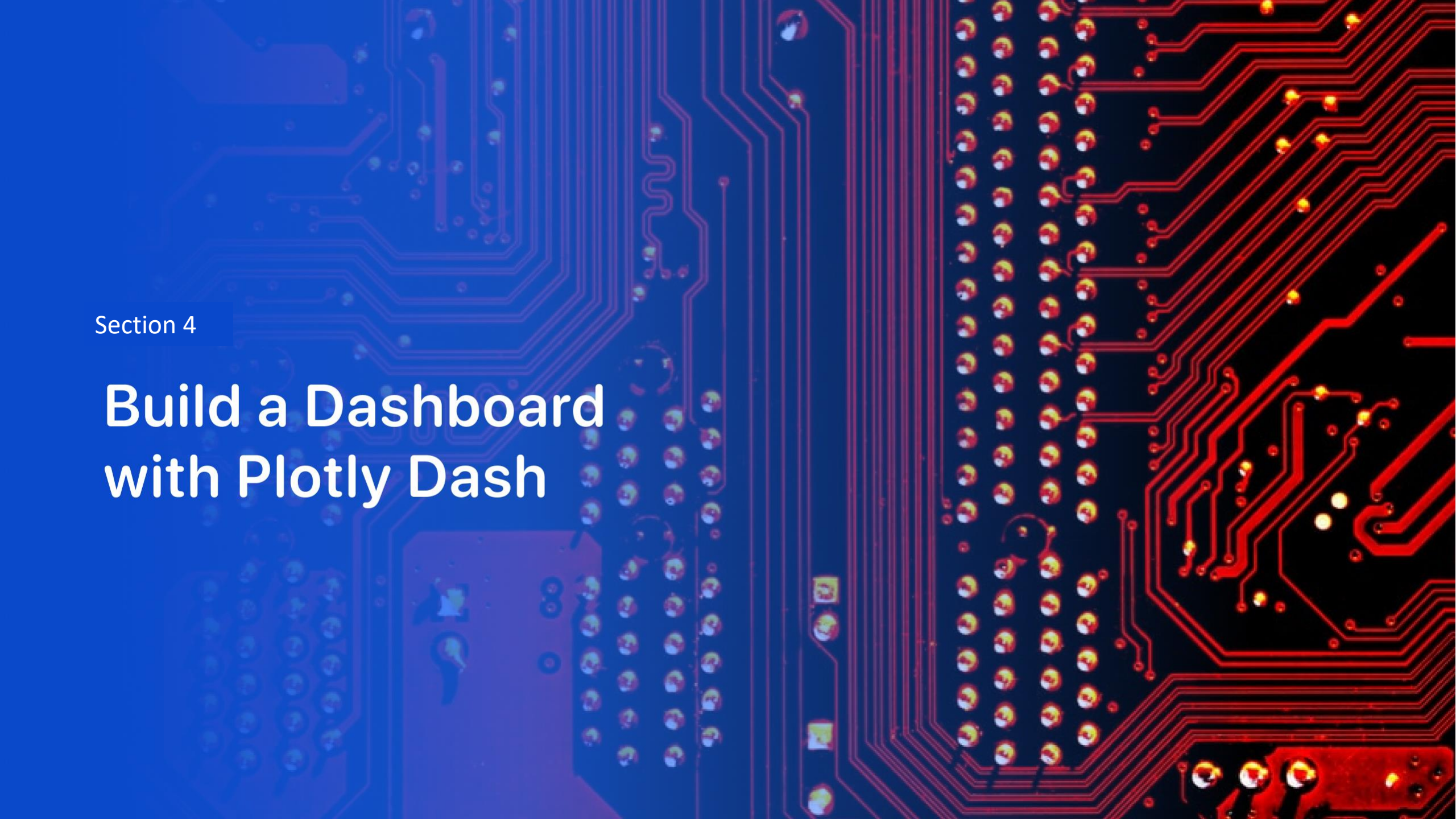
- From marker cluster for launch site we can see the total number of its launches
- From the color-labeled markers in marker clusters, we should be able to easily identify which launch sites have relatively high success rates.



The distances between a launch site to its proximities

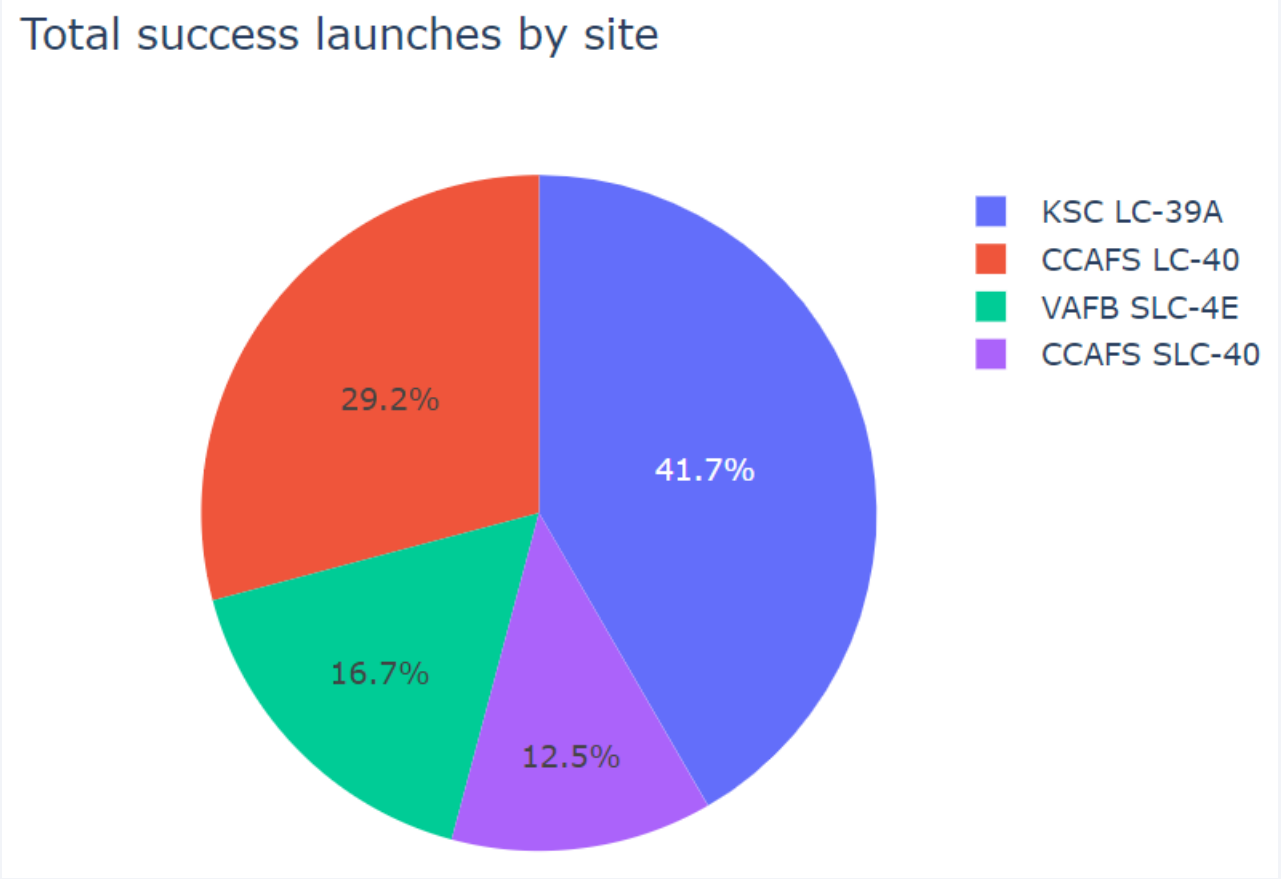
By adding lines between the launch sites and the closest points of the coastline, we show that they are located closer to logistics and away from human settlements




Section 4
Build a Dashboard with Plotly Dash

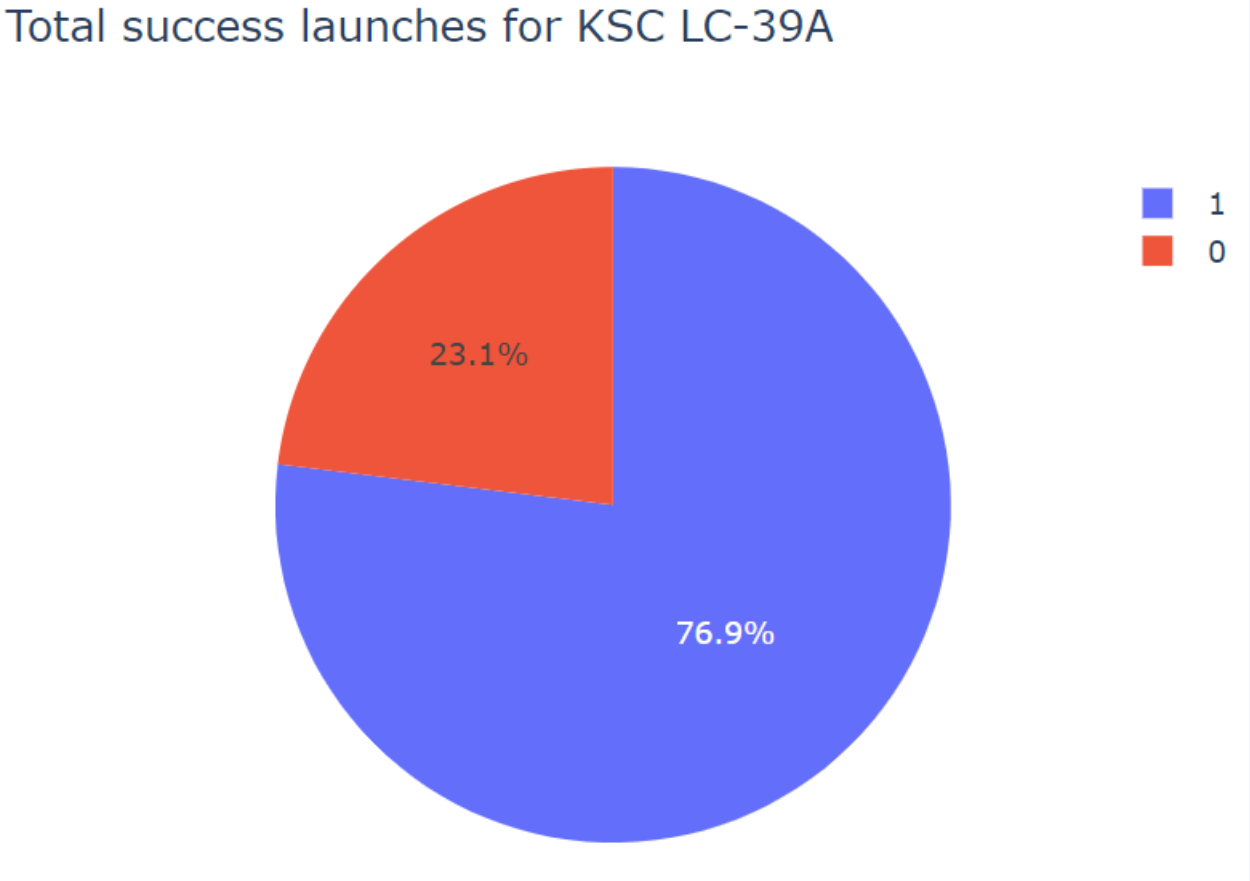
Total success launches for all sites

A piechart of launch success count for all sites shows that **Kennedy Space Center Launch Complex 39A** has the largest successful launches



The launch site with highest launch success ratio

The piecharts of success rate for a specific launch site gave us the one with highest launch success ratio - **Kennedy Space Center Launch Complex 39A**

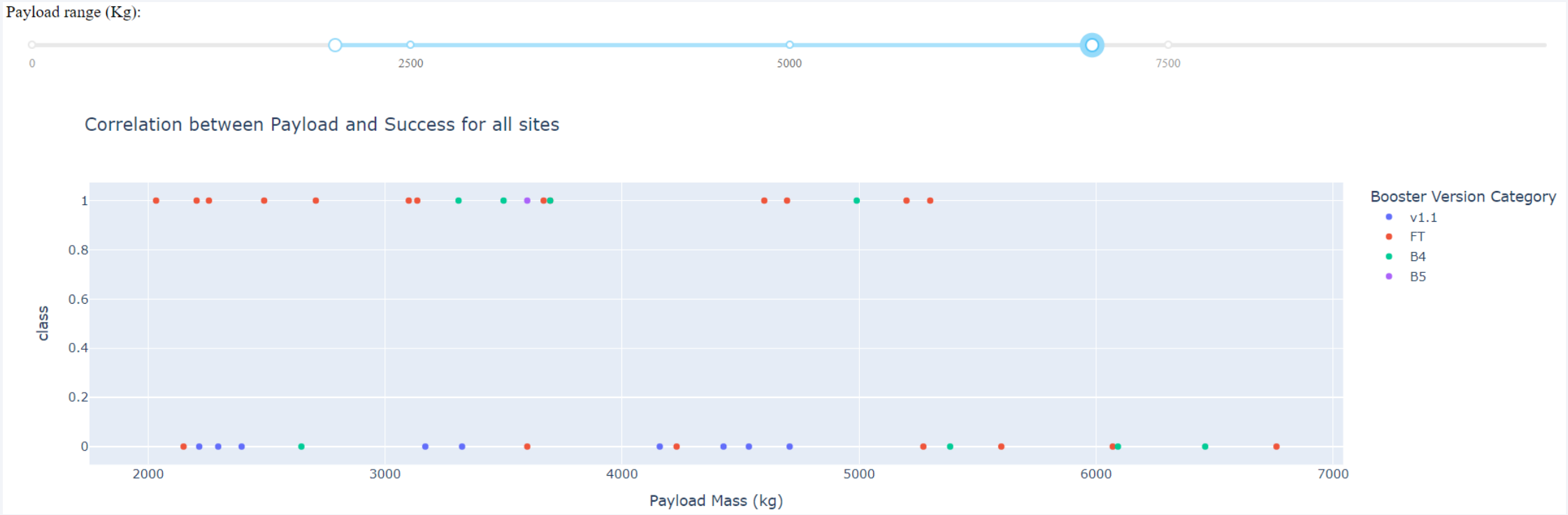


Payload vs. Launch Outcome with different payload ranges

Payload range **from 1952 to 5300 kg** has the highest launch success rate

Payload vs. Launch Outcome with different payload ranges

Payload range **from 2150 to 6761** has the lowest launch success rate



Payload vs. Launch Outcome with different payload ranges

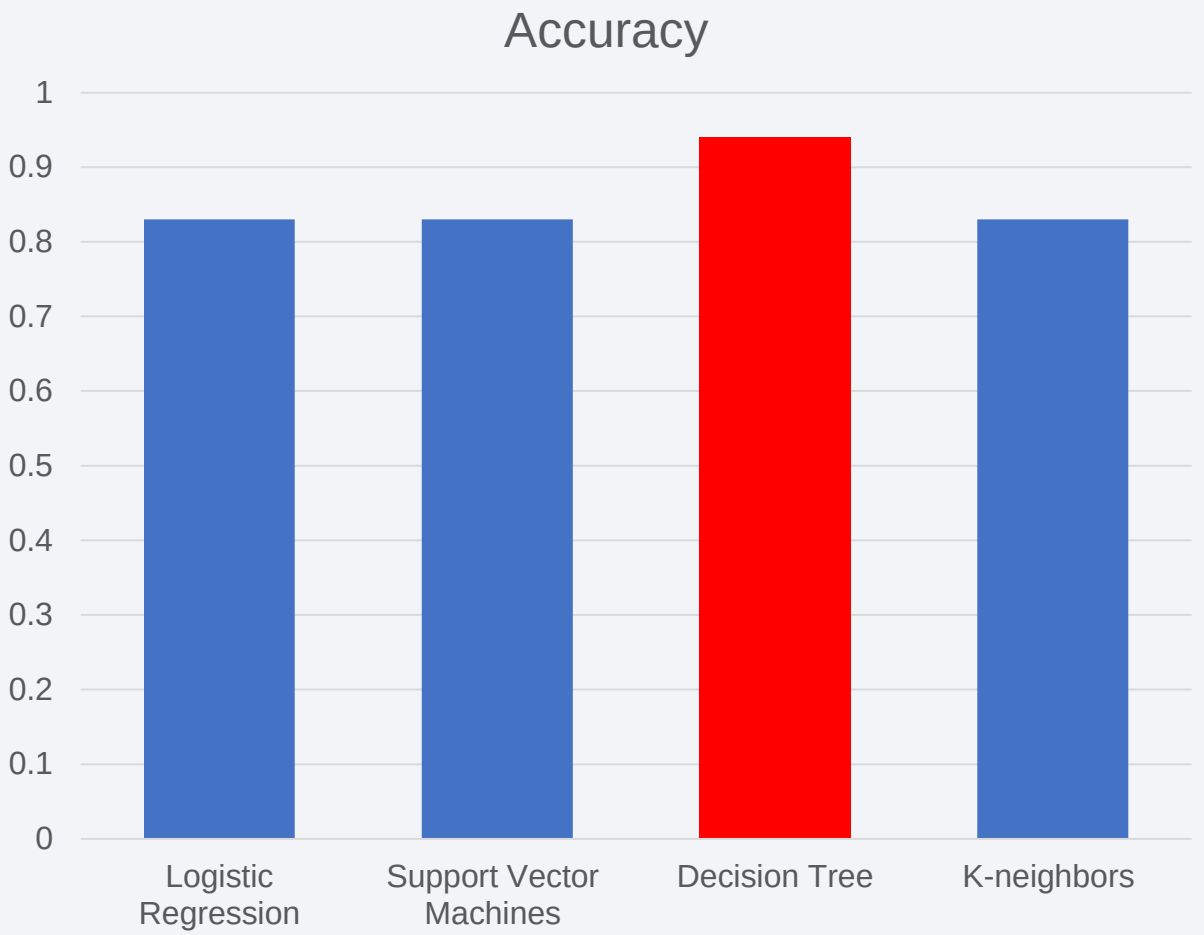
F9 Booster versions **FT and B4** have the highest launch success rate

Section 5

Predictive Analysis (Classification)

Classification Accuracy

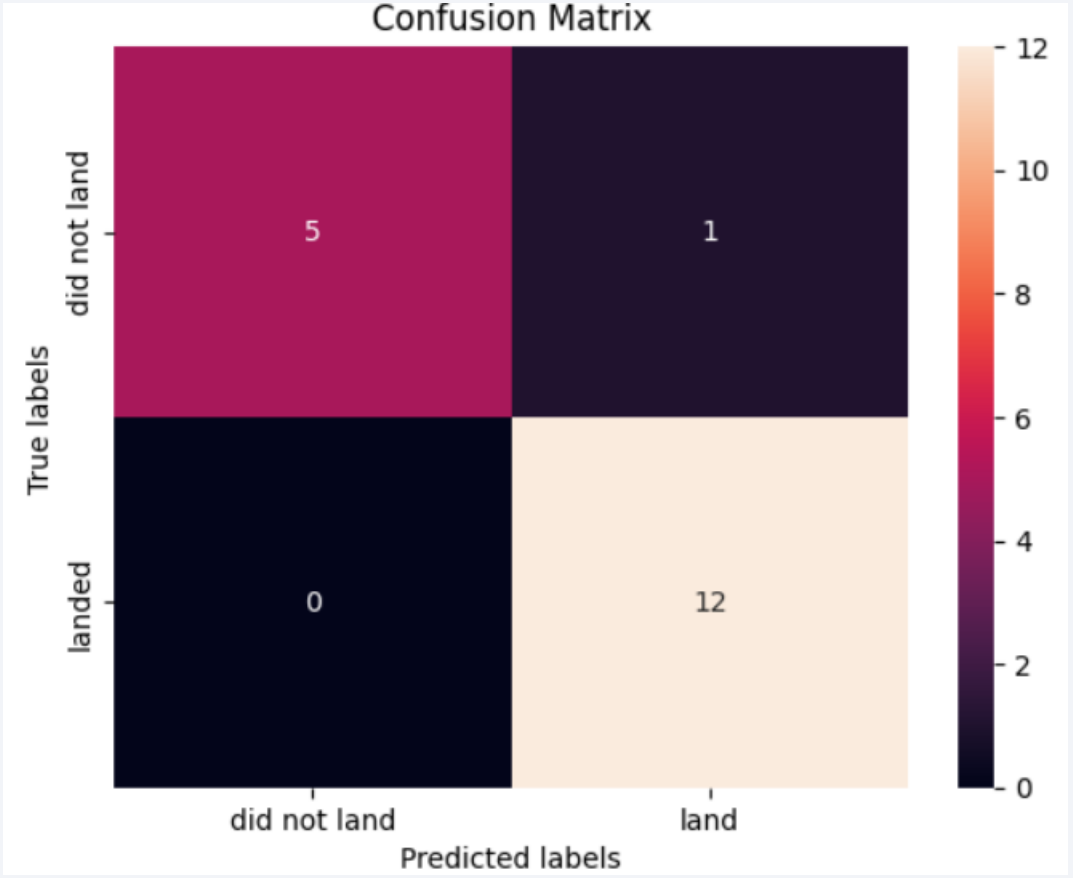
The built model accuracy for all built classification models in a bar chart shows that the Decision Tree method has the highest classification accuracy



Confusion Matrix

The confusion matrix of the best-performing model - Decision Trees - shows that the problem of false positives remains, but only in one case.

This means that the decision tree is the most effective prediction model in this case.



Conclusions

- As flight number increases, the number of successful first stage returns at all launch sites increases. Common success rate of returns kept increasing since 2013 till 2020. Explored data shows that the number of successes is more than six times greater than the number of failures. Total success rate of launches with certain result of the first stage return equals 0.86.
- These facts indicate that nowadays SpaceX has collect enough experience for launch success mission with high probability.
- Spatial analysis of all launch sites gave the following results. In order to spend less energy to launch a rocket into Earth orbit, the launch sites should be located near the equator.
- To reduce logistics costs launch sites must be in close proximity to railways and highways. Success landing on drone ship has high quantity, so launch site has to be in very close proximity to the coast. This is also necessary to reduce the consequences of a possible accident, as well as the fact that the launch pad must be kept at a certain distance from cities to prevent harm to people.
- From all explored launch sites, Kennedy Space Center Launch Complex 39A has the largest successful launches which indicates the need to adopt their experience.

Conclusions

- Data analysis regarding different Earth orbits showed that high success rate relates to: ES-L1, SSO, HEO, GEO, but these orbits have an insignificant number of launches, suggesting that we cannot say that orbit type affects the success of a first stage return.
- The highest launch success rate revealed regarding to payload range from 1952 to 5300 kg and F9 Booster versions FT and B4.
- Decision Tree method has the highest classification accuracy. This means that the decision tree is the most effective prediction model in this case.

All these results allow us to reduce mission costs, avoid unnecessary damage, choose the correct rocket characteristics for successful first stage reentry, and give us a mechanism to predict mission success by adjusting various parameters.

Thank you!