Data Science Capstone Project

Lucas Felipe Silva
2023/07/01

Outline

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

Executive Summary

- Collected data from SpaceX api and SpaceX wikipedia page.
- Feature engineering on column ‘class’ which classifies successful landings, categorical values to one hot encoding format, and standardized data.
- Explore data using SQL
- Exploratory Data Analysis with SQL, Folium maps and dashboards.
- Use of GridSearchCV to find the best parameters for machine learning models.
- Visualization and comparison of all models (4 in total).

Introduction

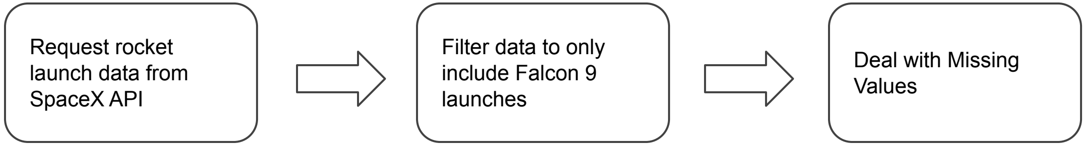
- SpaceX is one of the most successful companies of the space age.
- SpaceX’s accomplishments include:
  - Sending spacecraft to the International Space Station.
  - Starlink, a satellite internet constellation providing satellite internet access;
  - Sending manned missions to space.
- One of the main reasons of these accomplishments is that SpaceX rockets launches are relatively inexpensive.
  - The company advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards 165 million dollars each.
- This difference is because spaceX can reuse the first stage.
- Therefore if we can determine if first stage will land, we can determine the cost of a lauch.

Methodology

- Data Collection
  - Using SpaceX rest API
  - Using web scraping from wikipedia
- Data Wrangling
  - Filtering data
  - Feature Engineering
  - Dealing with Missing values
- Exploratory Data Analysis
  - Visualizations and SQL
- Interactive Data visualization
  - Using Folium Maps and Plotly Dash
- Predictive Analysis
  - Building, Tuning and evaluation

Data Collection

- Data Obtained by using SpaceX rest API

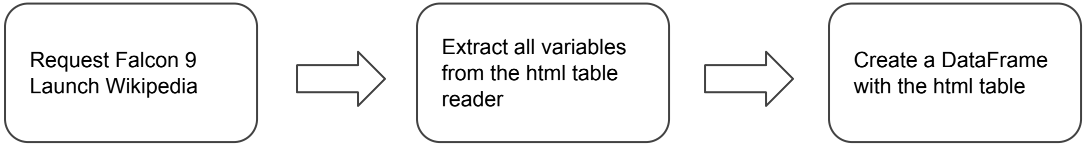


Request rocket launch data from SpaceX API

Filter data to only include Falcon 9 launches

Deal with Missing Values

- Data obtained by using wikipedia web scraping



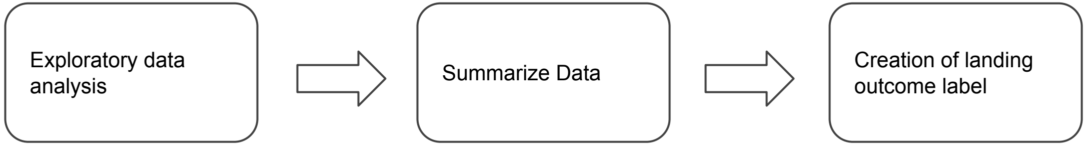
Request Falcon 9 Launch Wikipedia

Extract all variables from the html table reader

Create a DataFrame with the html table

Data Wrangling

- The location of each launch is stored on LaunchSite Column
    - Cape Canaveral Space (**VAFB SLC 4E**), Vandenberg Air force Base Space Launch (CCAFS SLC 40), Kennedy Space Center Launch (KSC LC 39A)
- Each launch aims to an dedicated orbit, and here are some common orbit types. That information is stored on Orbit Column.
- In the outcome variable True Ocean means the mission was successfully landed on a specific region of the ocean; True RTLS means the mission was successfully landed to a ground pad; True ASDS means the mission outcome was successfully landed to a drone ship; None ADS and None None represents failure to land.
- We converted those labels to one hot encoding (1 for True and 0 For False)



Exploratory data analysis

Summarize Data

Creation of landing outcome label

EDA With Data Visualization

- Charts vizualization:
  - Flight Number x PayloadMass, Flight Number x LauchSite, PayloadMass x LauchSite, Orbit Type x Succes Rate, Flight Number x Orbit type, Payload Mass x Orbit Type, and Success rate Yearly trend.
- Scatter Plots show the relationship between variables. If the relationship exists, they could be used in machine learning models.
- Bar Charts shows comparisons among discrete values.
- Line Charts shows trends over time.

EDA With SQL

- Performed SQL Queries
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA'
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first successful landing outcome in ground pad was acheived
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
  - List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - 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

Interactive Map with Folium

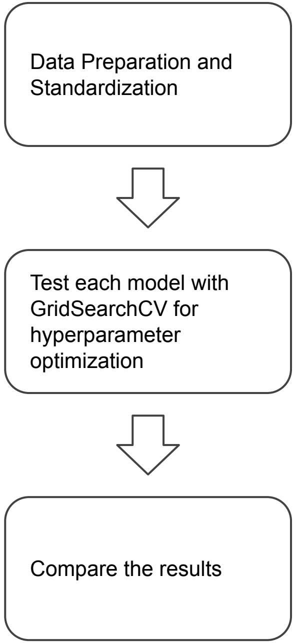
- Markers of all Launch Sites:
- Coloured Markers of the Launch outcomes for each site
- Distances Between a Launch Site to its proximities.

Dashboard with Plotly Dash

- Lauch Sites with drop down list
  - Added a dropdown list to enable Lauch Site selection
- Pie Charts showing Success launches (All Sites/Certain Site)
  - Added a Pie Chart to show the total successful launches count for all sites and the success x Failed counts for the site, if a specific site was selected
- Slider of PayloadMass range
  - Added a slider to select Payload Range
- Scatter Chart of Payload Mass x Success Rate for the different Booster Versions
  - Added a Scatter chart to show the correlation between Payload and Launch Success

Predictive Analysis

- Creating a Numpy array from the column Class
- Standardized data with StandardScaler
- Splitting data using train_test_split
- Create a GridSearchCV with cv=10
- Applying the gridSearch with Different models
  - Logistic Regression
  - SVM
  - Decision tree
  - KNN
- Calculate the accuracy
- Plot the Confusion Matrix
- Finding the best performance by comparing with Jaccard_score and f1_score



Data Preparation and Standardization

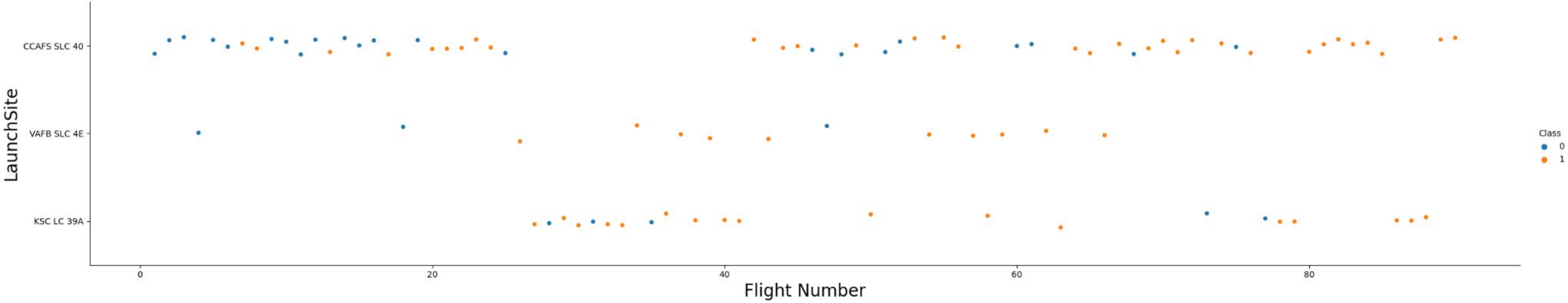
Test each model with GridSearchCV for hyperparameter optimization

Compare the results

Insights Drawn from EDA

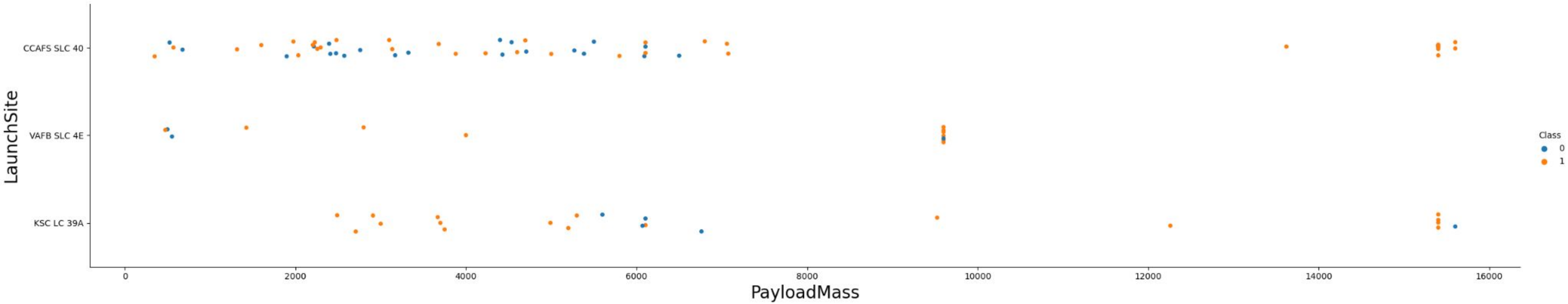
Flight Number vs. Launch Site

- According to the plot above, the most of launches and the successful ones were done on CCAF5 SLC 40.
- The less used launch site is VAFB SLC 4E.
- It’s also possible to see that the successful improved the launches.



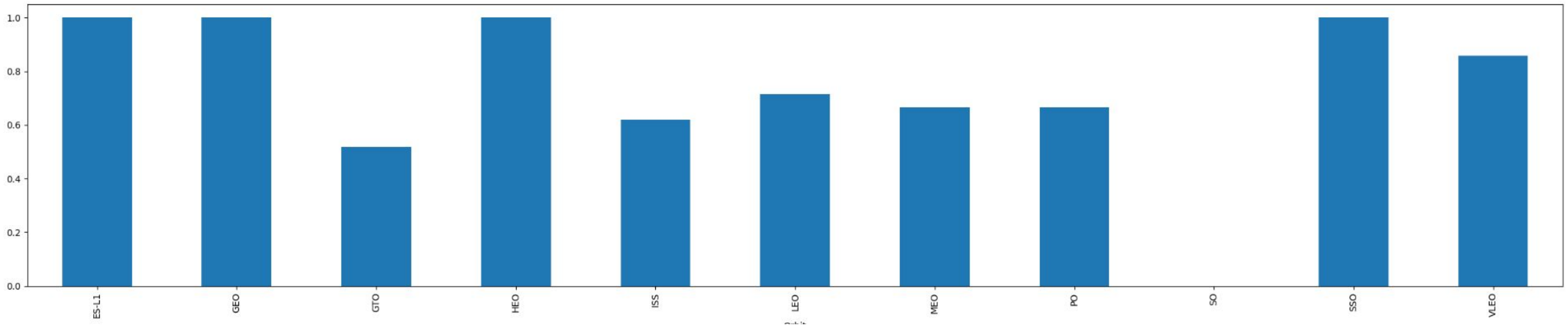
Payload vs. Launch Site

- We can see that VAFB-SLC launchsite there are no rockets launched for heavy payload mass(greater than 10000).



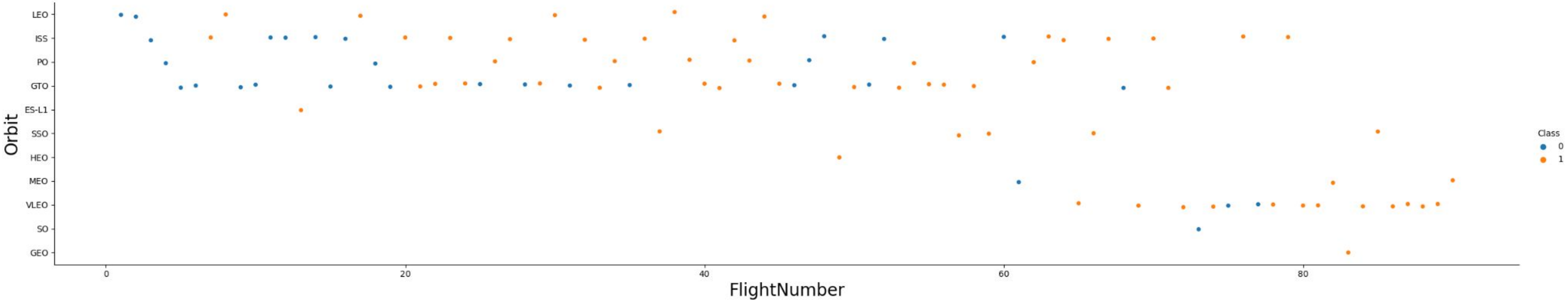
Success Rate vs. Orbit Type

- The big success rate happened on
    - ES-L1, GEO, HEO, SSO



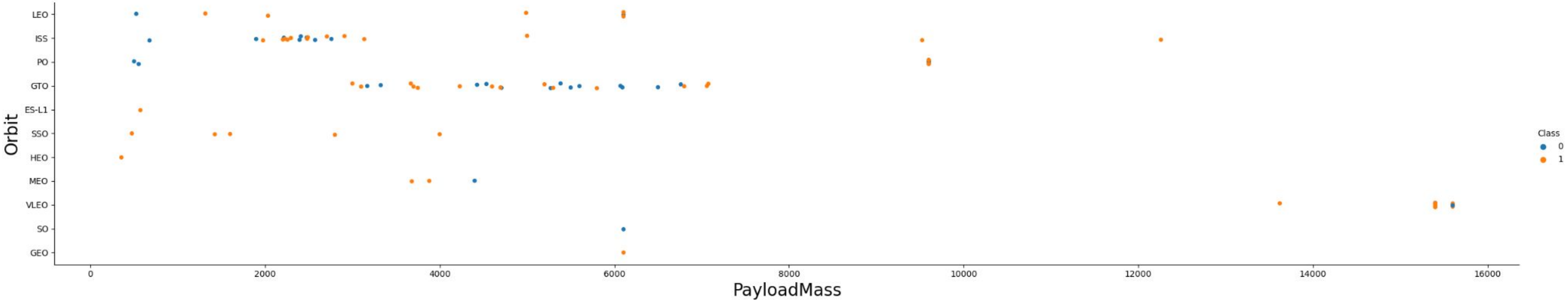
Flight number vs. Orbit Type

- VLEO appears most in the latest launches with a high successful rate.
- The less used is MEO, HEO, SSO.
- LEO Success rate increased over the lanches;
- GTO in the other hand did not increase that much.



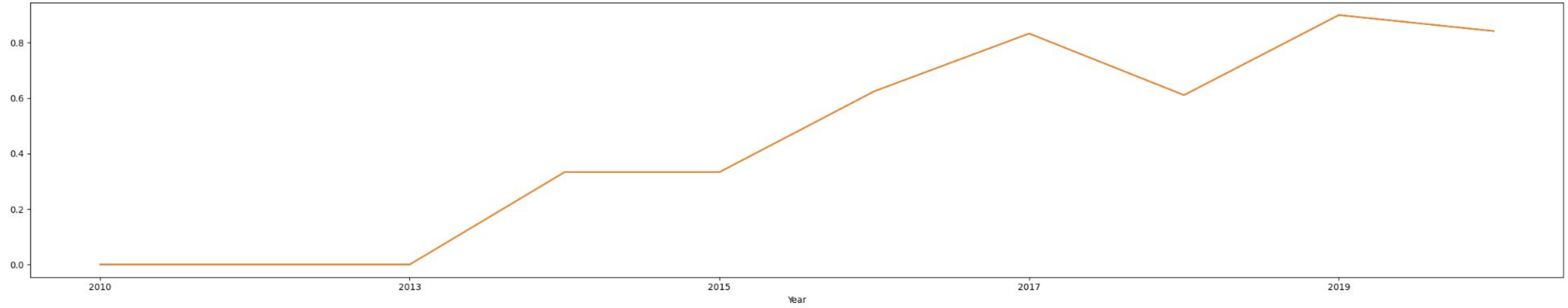
Payload vs. Orbit Type

- 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.



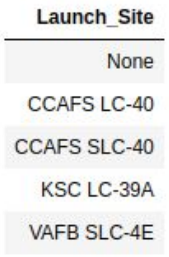
Launch Success Yearly Trend

- It’s possible to see that success rate improved since 2013.



All launch Sites names

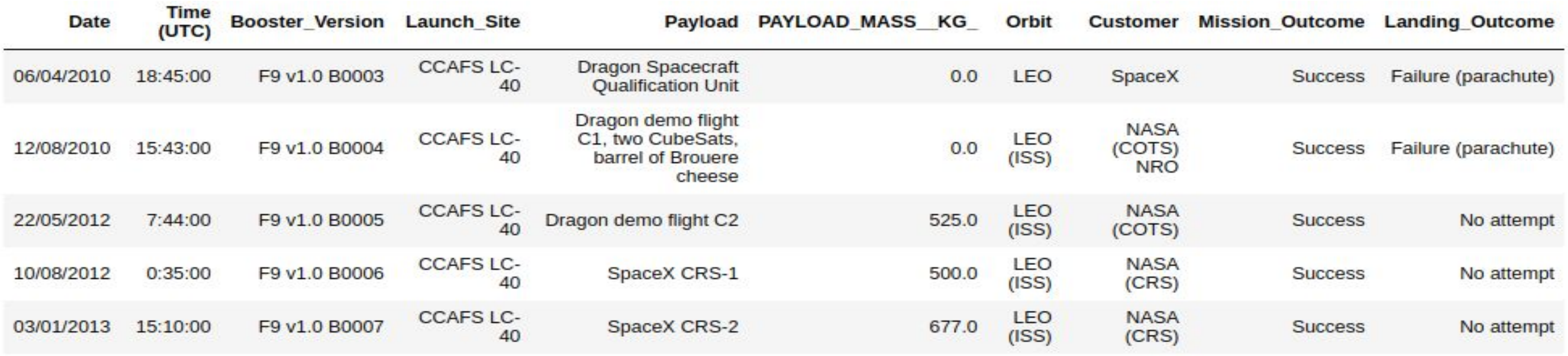
- According to the data there a four Launch sites
- This value was obtained by selecting distinct values from launch site column.



Launch_Site
None
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

Launch Site names begin with ‘CAA’

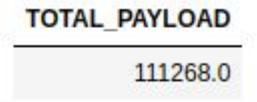
- There are in total 5 rows that launch site begins with ‘CAA’
- This value was obtained by using a sql command ‘LIKE “CCA%” that matches anything that starts with CAA and % represents anything that follows.



Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

Total payload mass

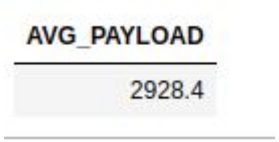
- Total payload mass is equal to 111,268,00
- This data was obtained by using sum function in the payload mass and filtering “PAYLOAD LIKE '%CRS%'”



TOTAL_PAYLOAD
111268.0

Average payload mass by F9 v1.1

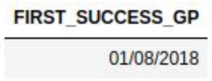
- The average is 2928.4
- This value was obtained by using the AVG sql function and filtering BOOSTER_VERSION = 'F9 v1.1'.



AVG_PAYLOAD
2928.4

First Successful Ground Landing Date

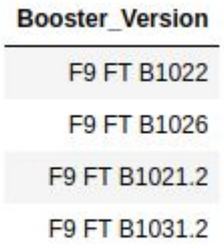
- The value was obtained by filtering Landing_Outcome = 'Success (ground pad)' and using the MIN sql function on date column,



FIRST_SUCCESS_GP
01/08/2018

Successful drone ship landing with payload between 4000 and 6000

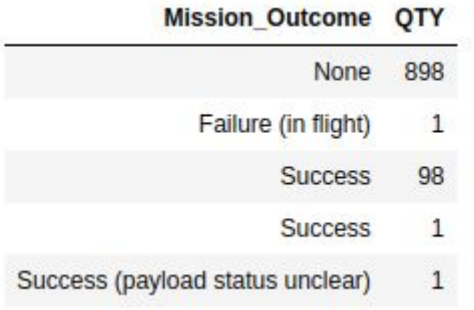
- The value was obtained by
  - Filtering “PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND Landing_Outcome = 'Success (drone ship)'”
  - And selecting distinct values of booster name



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

Total number of successful and failure missions outcomes

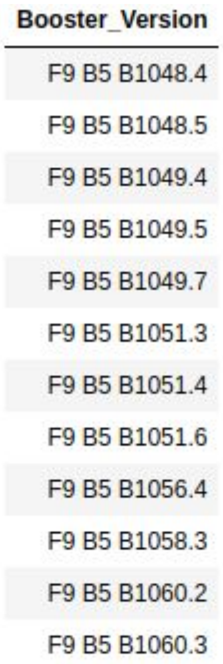
- This value was obtained by grouping the data by mission outcome and count the records.



Mission_Outcome	QTY
None	898
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Booster Carried maximum payload

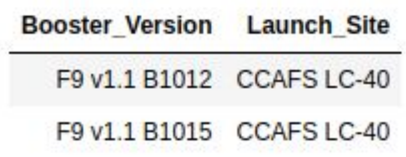
- This value was obtained by using a subquery to filter the maximum values of payload.



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

2015 Launch records

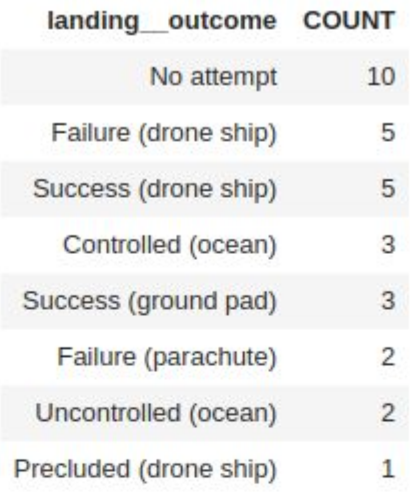
- This value was obtained by filtering the year (from date column) equals to 2015 and filtering Landing_Outcome = 'Failure (drone ship)'



Booster_Version	Launch_Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

Ranking lading outcomes between 2010-06-04 and 2017-03-20

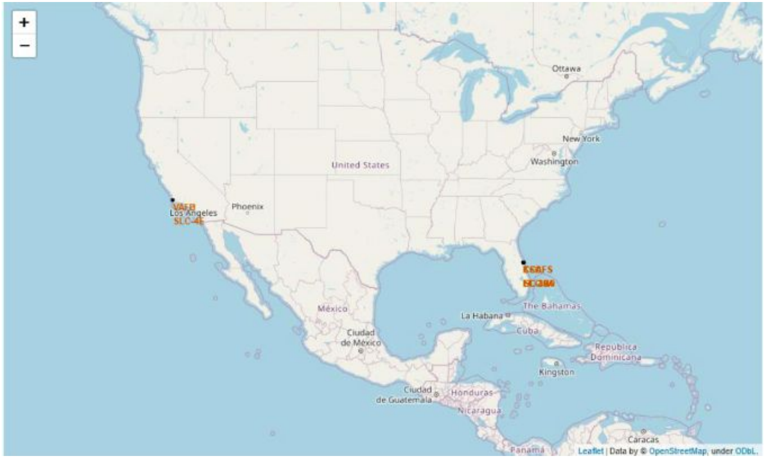
- This value was obtained by grouping the data by landing outcome and counting the records inside the date range specified.



landing__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

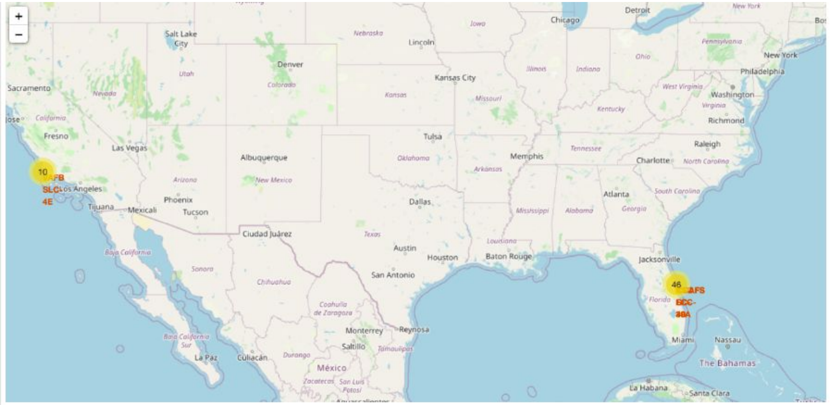
Launch sites proximities analysis

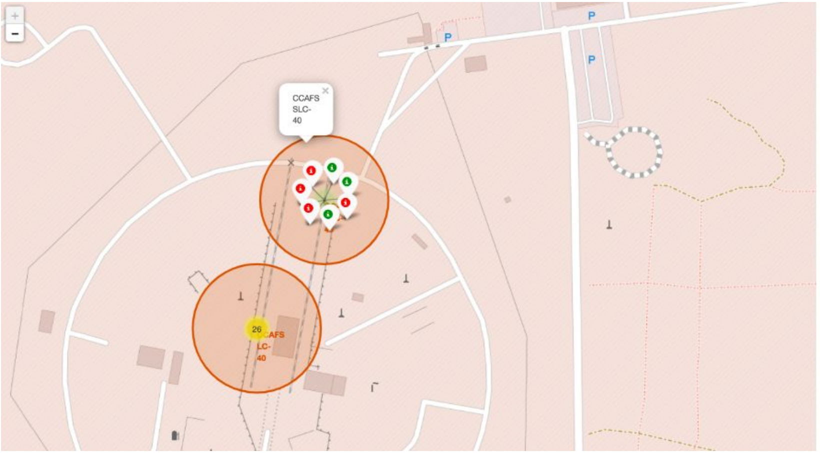
Launch sites



- Launch sites are near sea.

Success/Failure launches from each site

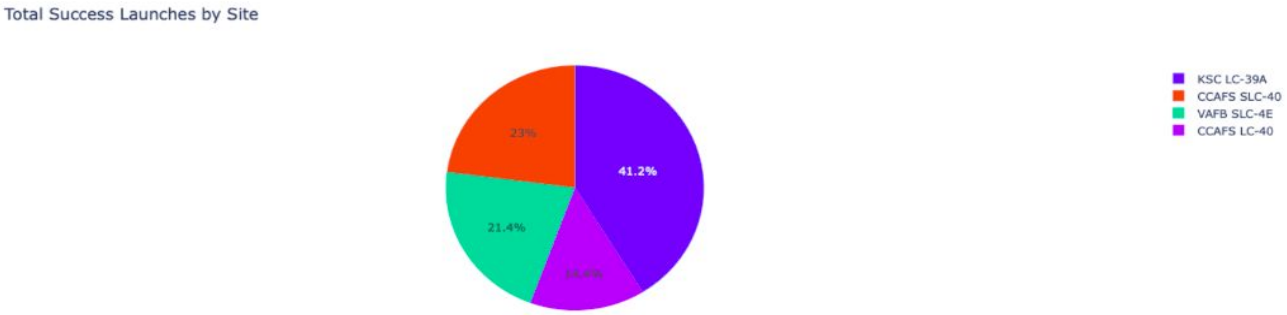




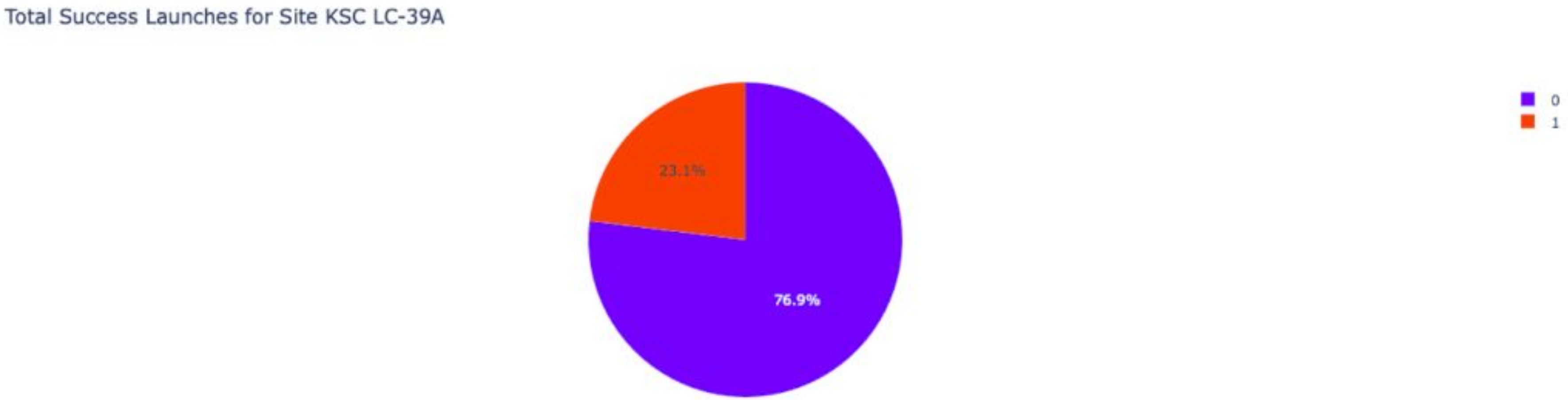
Build a dashboard with Plotly Dash

Launch success from all sites

- This chart show to us that KSC LC-39A has the most successful rate



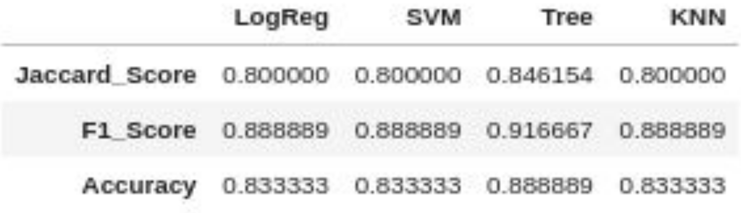
Launch site with highest successful rate



Predictive analysis results

Results from Classification

- The predictive analysis showed that the Decision Tree Classifier is the best model with all metrics that we used.



	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.846154	0.800000
F1_Score	0.888889	0.888889	0.916667	0.888889
Accuracy	0.833333	0.833333	0.888889	0.833333