IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Pedro Paulo Macedo
05-29-2022

Outline

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

Executive Summary

- The following methodologies were used to analyze data:

    Data Collection using web scraping and SpaceX API;

    Exploratory Data Analysis (EDA), including data wrangling, data visualization and interactive visual analytics;

    Machine Learning Prediction.

- Results

    It was possible to collect valuable data from public sources;

    EDA allowed to identify which features are the best to predict success of launches;

    Machine Learning Prediction showed the best model to predict which characteristics are important to drive this opportunity by the best way.

Introduction

The objective is to evaluate the viability of the new company Space Y to compete with Space X.

- Determinate the price of each launch by predicting successful landings of the first stage of rockets that comeback to landing;
- Determinate if Space Y will reuse the first stage;
- Determinate where are the best places to make launches.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Space X API (https://api.spacexdata.com/v4/rockets/)
  - WebScraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

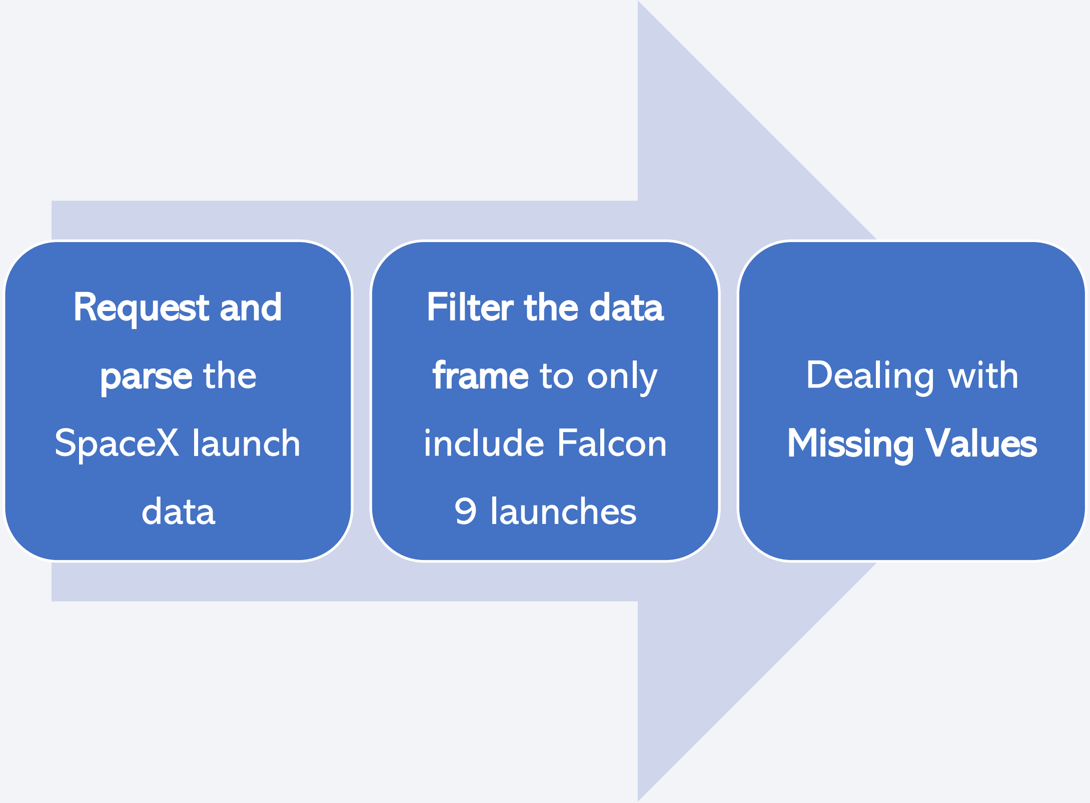
We’ve started collecting data

- Request to the SpaceX API
- Clean the requested Data

Data Collection – SpaceX API

- Imported Libraries and Define Auxiliary Functions;
- Used GET request to request data from a SpaceX public API.

[Data Collection - API . GitHub]



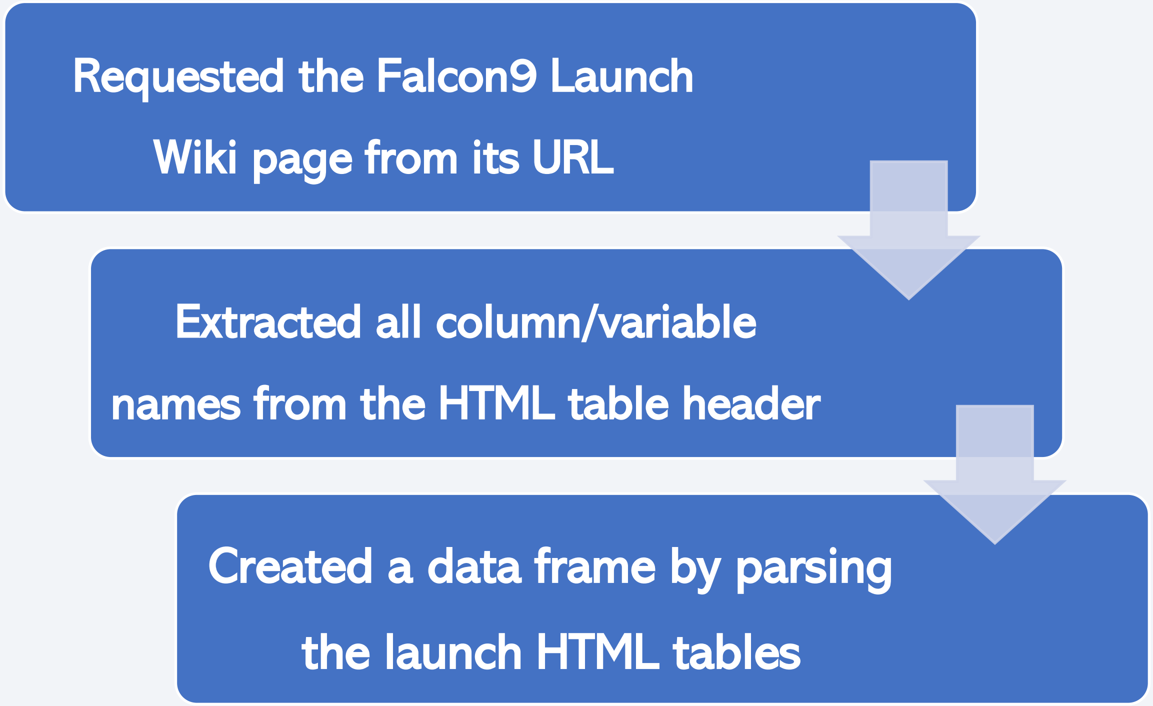
Data Collection - Scraping

- Records with BeautifulSoup
  - Extracted a Falcon 9 launch records HTML table from Wikipedia;

[Falcon 9 – Wikipedia]

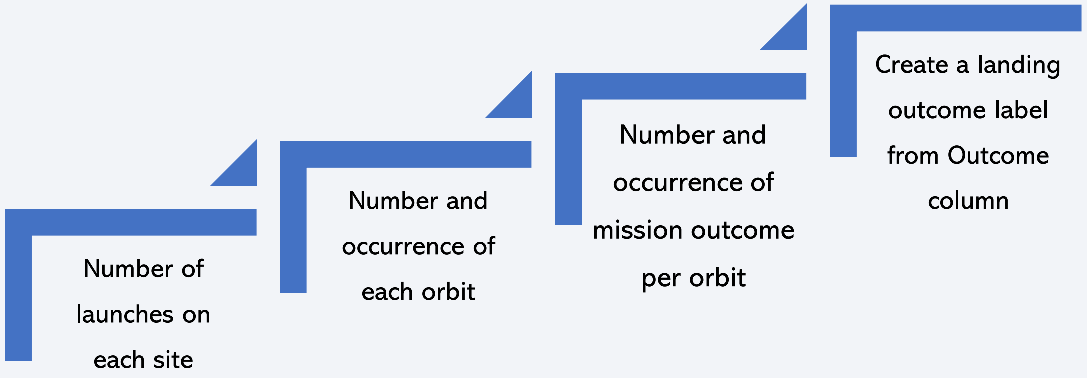
  - Parsed the table and converted it into a Pandas data frame.

[Data Collection - Web Scraping .GitHub]



Data Wrangling

- EDA was performed in the dataset
    - Loaded Space X dataset, from last section;
    - Summarized numbers;
    - Created training labels.



[Data Wrangling .GitHub]

EDA with Data Visualization

EDA used scatterplots and barplots to visualize relationships

- Flight Number vs. Payload Mass
  - See how the number of flights and the payload mass would affect the launch outcome
- Flight Number vs. Launch Site
  - See the relationship between the flight number and the launch site, trying to understand if it can correlate time with outcome
- Payload vs. Launch Site
  - Observe if there is any relationship between launch site and their payload

EDA with Data Visualization

- Success Rate vs. Orbit Type
  - Understand which type of orbit have high success rate
- Flight Number vs. Orbit Type
  - We can see the relationship between the flight number and orbit type, trying to understand if there’s any orbit type increasing success outcome over time
- Payload vs. Orbit Type
  - Observe if there is any relationship between payload and orbit type
- Success Yearly Trend
  - Understand if the success rate average is increasing over time

EDA Data Visualization .GitHub

EDA with SQL

The following SQL queries were performed:

- Unique launch sites in the space mission;
- 5 records where launch sites begin with the string 'CCA';
- The total payload mass carried by boosters launched by NASA (CRS);
- Average payload mass carried by booster version F9 v1.1;
- The date when the first successful landing outcome in ground pad was achieved;
- Boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000;
- Total number of successful and failure mission outcomes;
- 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

[Exploratory Data Analysis .GitHub]

Build an Interactive Map with Folium

- Marked all launch sites on a map
  - Indicate the coordinates where launch sites are on the map
  - Circle indicate highlighted area around this coordinates
- Marked the success/failed launches for each site on the map
  - Indicate by color the result of the launches
  - Green – successful
  - Red - unsuccessful
- Calculated the distances between a launch site to its proximities
  - Lines that indicate the distance

Interactive Visual Analytics .GitHub

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 allowed to quickly analyze the relation between payloads and launch sites, helping to identify where is best place to launch according to payloads.

EDA - Plotly Dash .GitHub

Predictive Analysis (Classification)

Performed exploratory Data Analysis and determinate Training Labels

- Standardize the data
- Find best Hyperparameter for SVM, Classification Trees and Logistic Regression
- Find the method performs best using test data

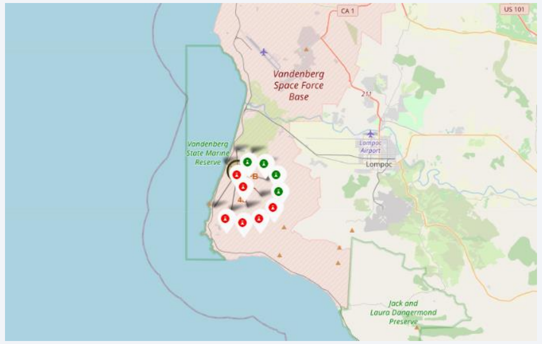
[Machine Learn Prediction .GitHub]

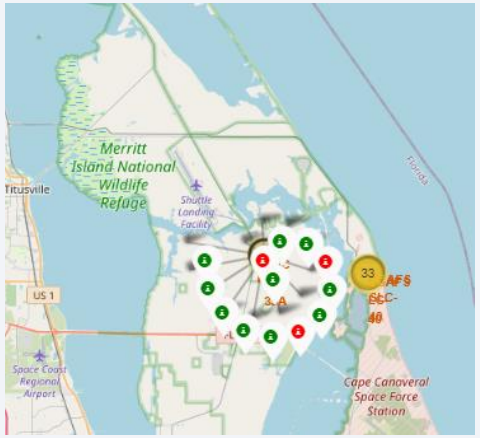
Results - EDA

- Exploratory data analysis results
  - Space X uses 4 different launch sites;
  - The first launches were done to Space X itself and NASA;
  - The average payload of F9 v1.1 booster is 2,928 kg;
  - The first success landing outcome happened in 2015 fiver year after the first launch;
  - Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average;
  - Almost 100% of mission outcomes were successful;
  - Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015;
  - The number of landing outcomes became as better as years passed.

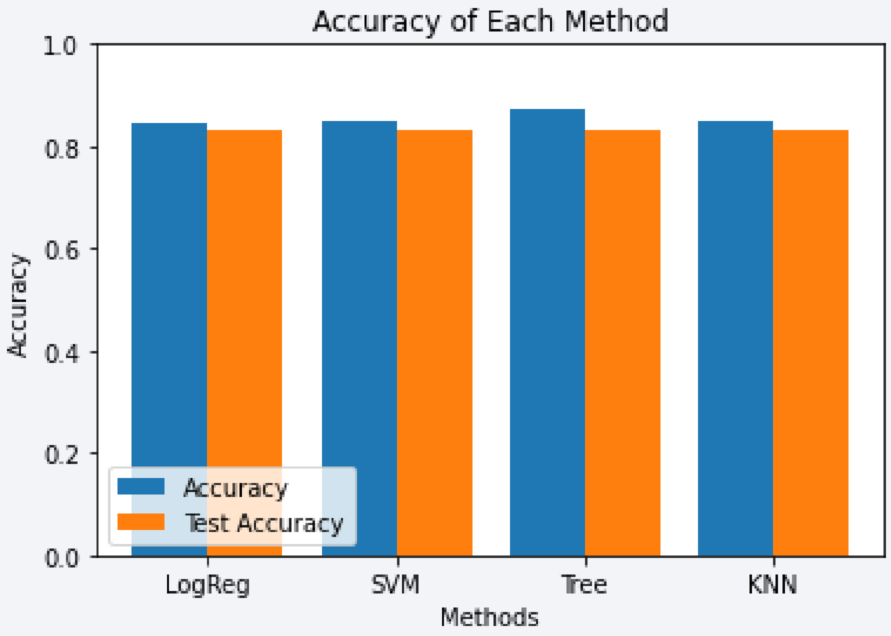
Results – Interactive Analysis

The launch sites have strategic locations which can favor the transport being near railroads, roads and can keep the distance from high population;





Results – Predictive Analysis



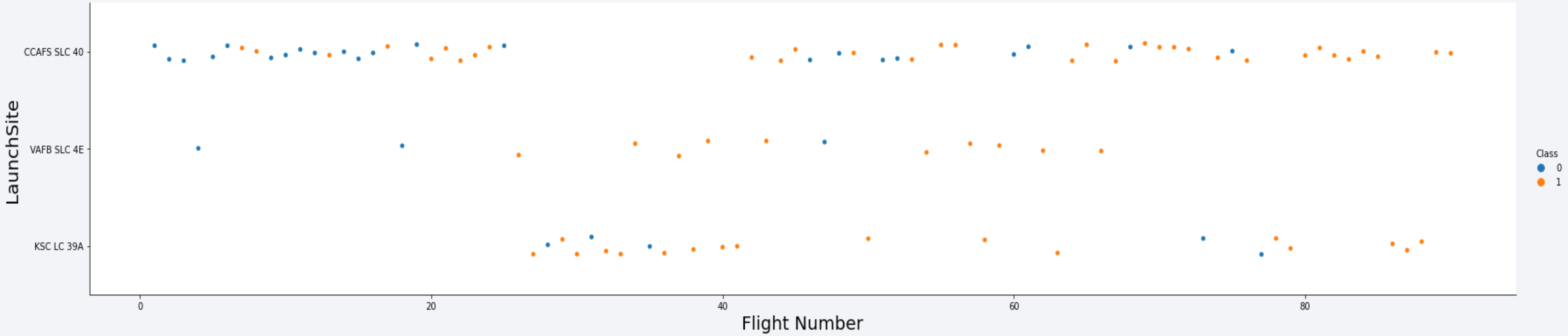
- The model with the highest classification accuracy is Decision Tree Classifier, which has accuracies over than 87,5%.

Model	Accuracy	TestAccuracy
LogReg	0.84643	0.83333
SVM	0.84821	0.83333
Tree	0.875	0.83333
KNN	0.84821	0.83333

Section 2

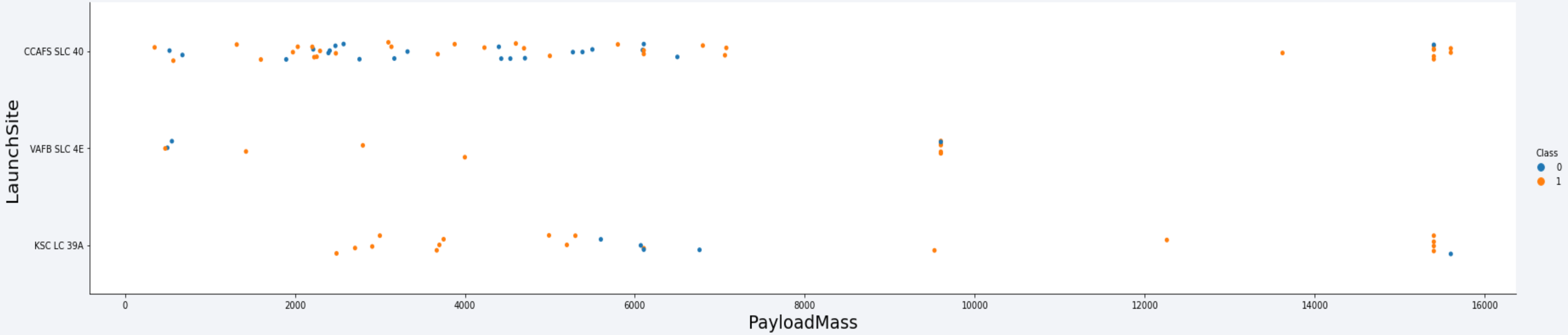
Insights drawn from EDA

Flight Number vs. Launch Site



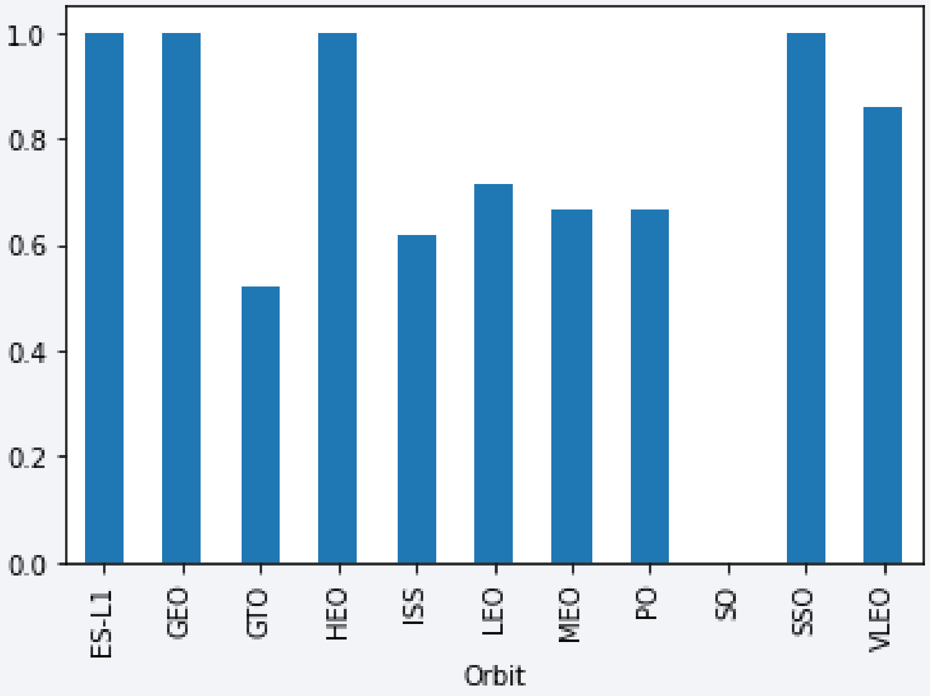
- The Launch Sites with the best success rate over time are VAFB SLC 4E and KSC LC 39A
- It's possible to verify that the most used nowadays is CCAF5 SLC 40 and its success rate improved over time

Payload vs. Launch Site



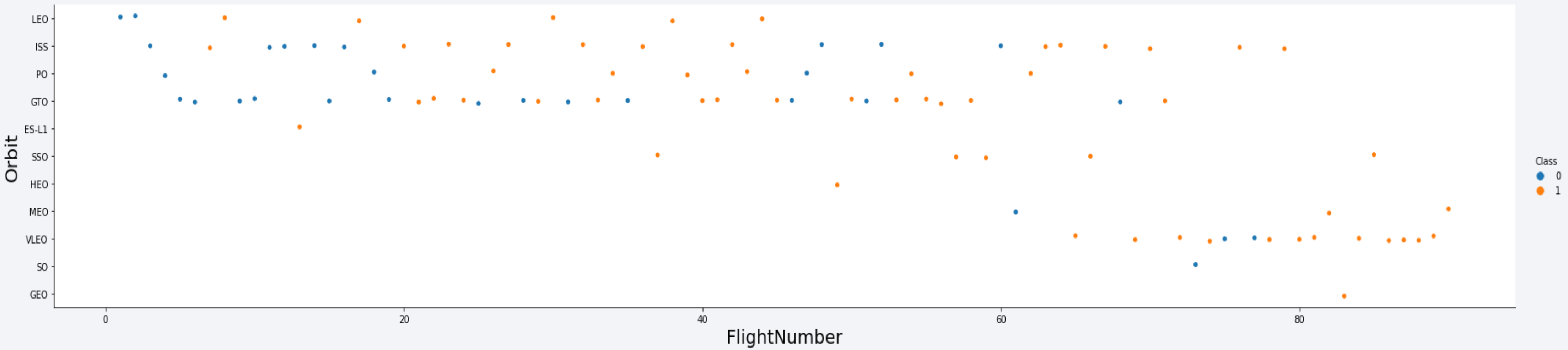
- The success rate increase in heavy payload mass (greater than 10000)
- There are no rockets launched for heavy payload mass from VAFB SLC 4E

Success Rate vs. Orbit Type



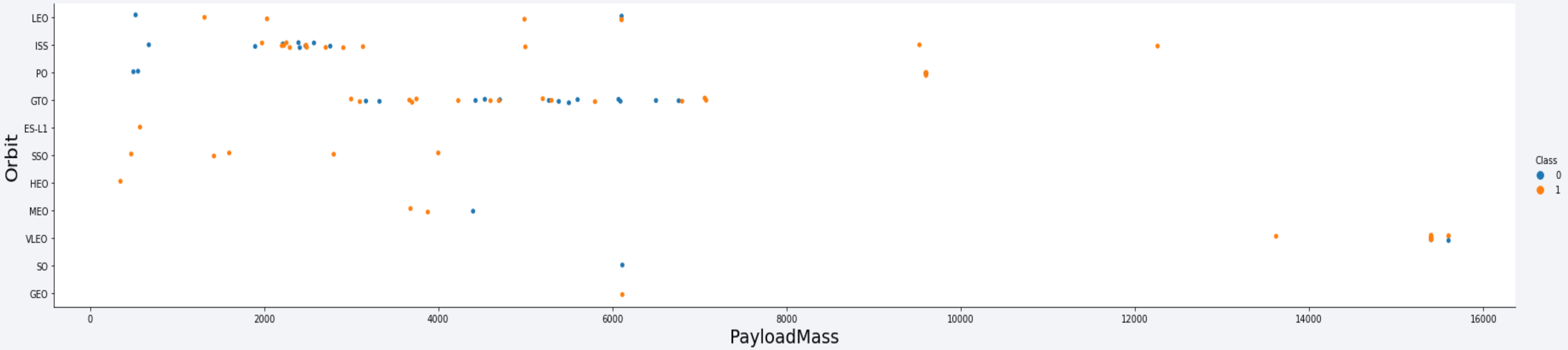
- There are four Orbit Type that have high success rate
  - ES-L1
  - GEO
  - HEO
  - SSO

Flight Number vs. Orbit Type



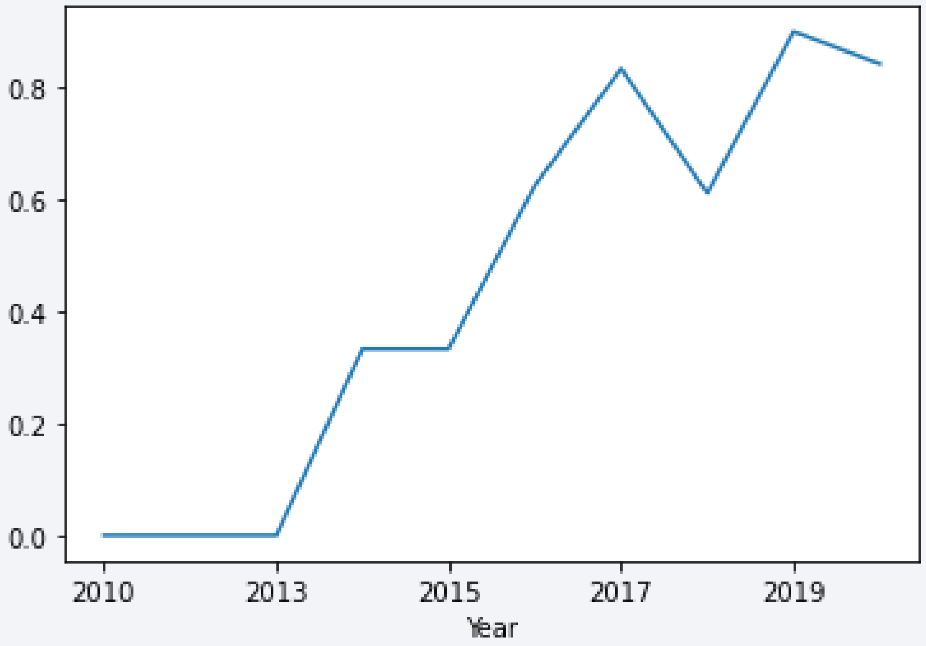
- LEO orbit the Success appears related to the number of flights;
- On the other hand, there seems to be no relationship between flight number when in GTO orbit;
- VLEO orbit and LEO have higher success rate

Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for PO, 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



- The success rate since 2013 kept increasing till 2020
- 2018 appear to be a period of adjust and improvement to keep increasing the success rate

All Launch Site Names

- There are four launch sites

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

- Obtained by "SELECT DISTINCT LAUNCH_SITE"

Launch Site Names Begin with 'CCA'

Cape Canaveral Launches

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	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

- Obtained by “SELECT WHERE LAUNCH_SITE LIKE ‘CCA%’

Total Payload Mass

The Total Payload Mass carried by boosters launched by NASA

total_payload
111268

- Obtained “SUM(PAYLOAD_MASS_KG_)” which corresponds to the sum of all payloads code contains ‘CRS%’

Average Payload Mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1 corresponds to:

avg_payload
2928

- Obtained “AVG(PAYLOAD_MASS_KG_)” which corresponds to the average from booster version F9 v1.1

First Successful Ground Landing Date

First successful landing outcome on ground pad

first_success_gp
01-05-2017

- Obtained by “SELECT MIN(DATE) WHERE LANDING_OUTCOME = SUCCESS”

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 but less than 6000

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

- There are 4 different results obtained by “SELECT DISTINCT BOOSTER_VERSION WHERE PAYLOAD_MASS_KG_ BETWEEN 4000 AND 6000”

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes

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

- As we can see SpaceX do not have a problem with the mission, just with the landing outcome. From 101 missions, 99 had success and just 1 failed.

Boosters Carried Maximum Payload

Names of the booster 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
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

- Failed landing outcomes

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40

- The only landing that failed was launched from CCAFS LC-40 and with F9 v1.1 booster version, obtained by "SELECT WHERE LANDIND__OUTCOME = 'FAILURE (DRONE SHIP)' AND DATE_PART ('YEAR', DATE) = 2015"

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

- Rank of landing outcomes between the date 2010-06-04 and 2017-03-20

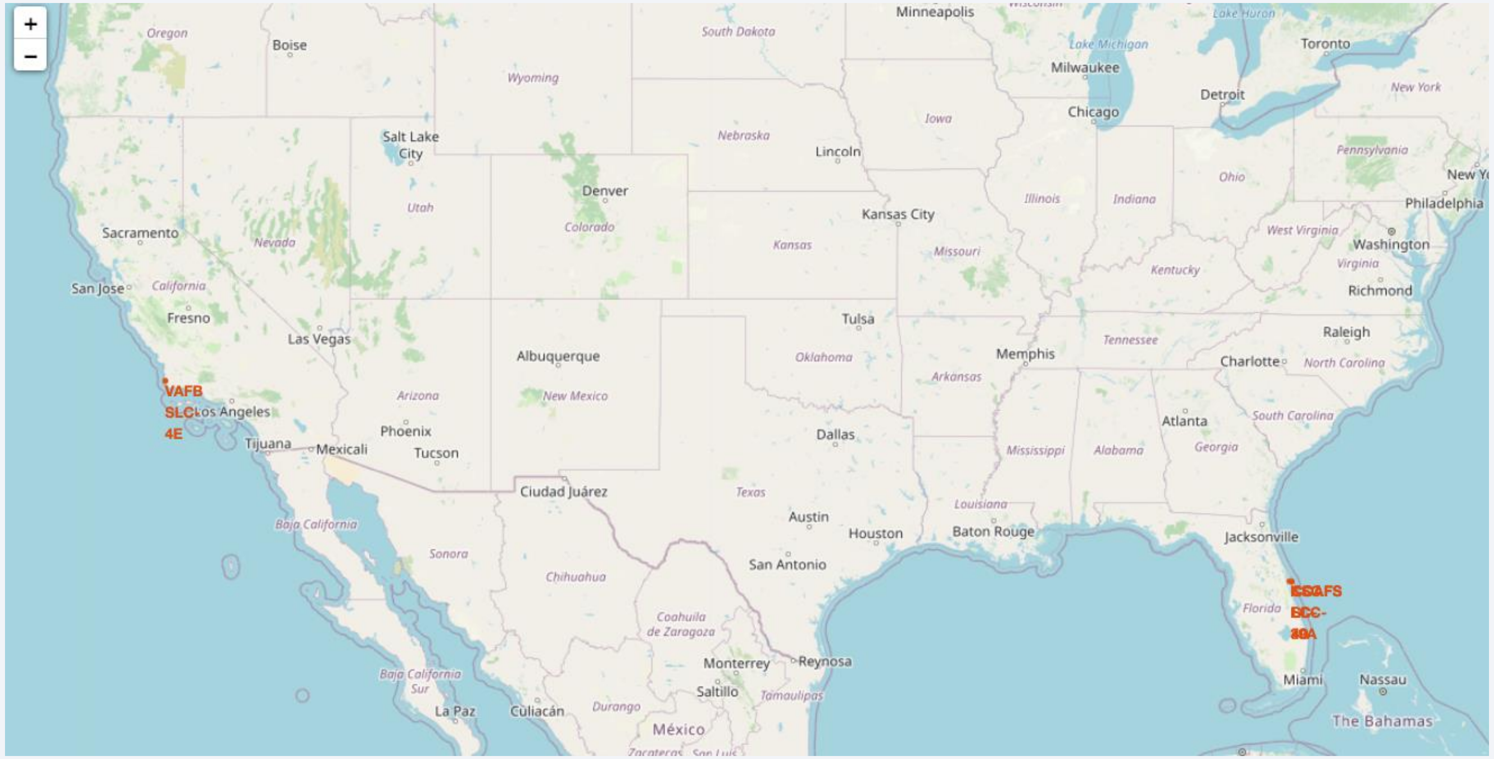
landing_outcome	qty
Success	20
No attempt	11
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Controlled (ocean)	3
Failure	3
Failure (parachute)	2

- Rank obtained by “SELECT LANDING OUTCOME AS QTY WHERE DATE BETWEEN”

Section 3

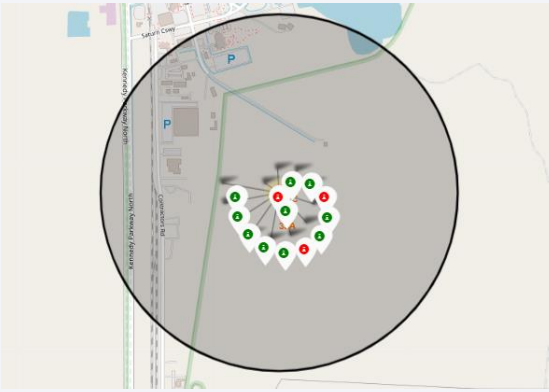
Launch Sites Proximities Analysis

Launch Sites Over The Globe



- Both launch sites are near coast and It's probably because of :
  - Safety
  - Logistic

Launch Outcomes



- We can see both Launch Outcomes from KSC LC-39A and CCAFS LC-40
- Green markers indicate successful launches and red indicates unsuccessful launches

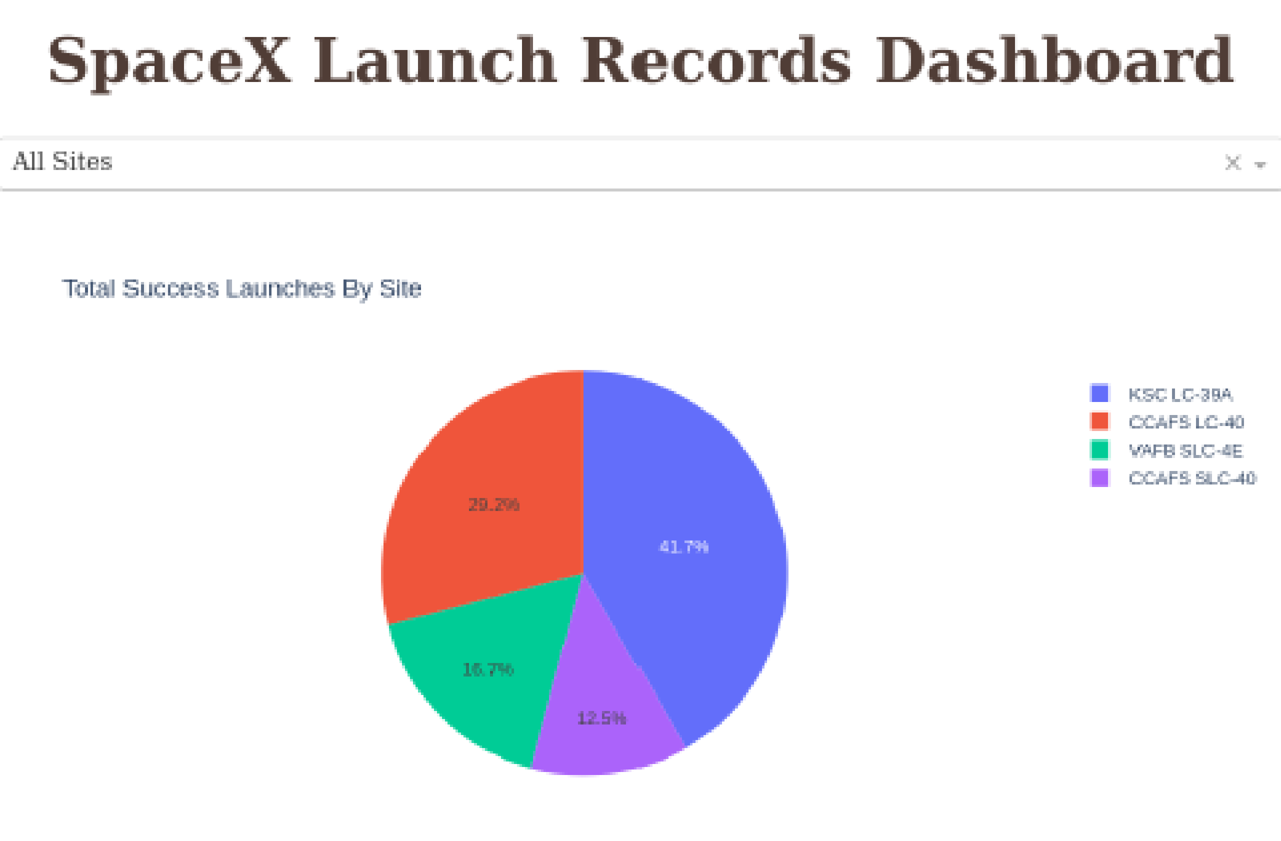
Strategic Location

- Launch sites have strategic locations which can favor the transport being near railroads and roads
- In other hand they keep the distance from cities to make sure that won't risk anyone's life

Section 4

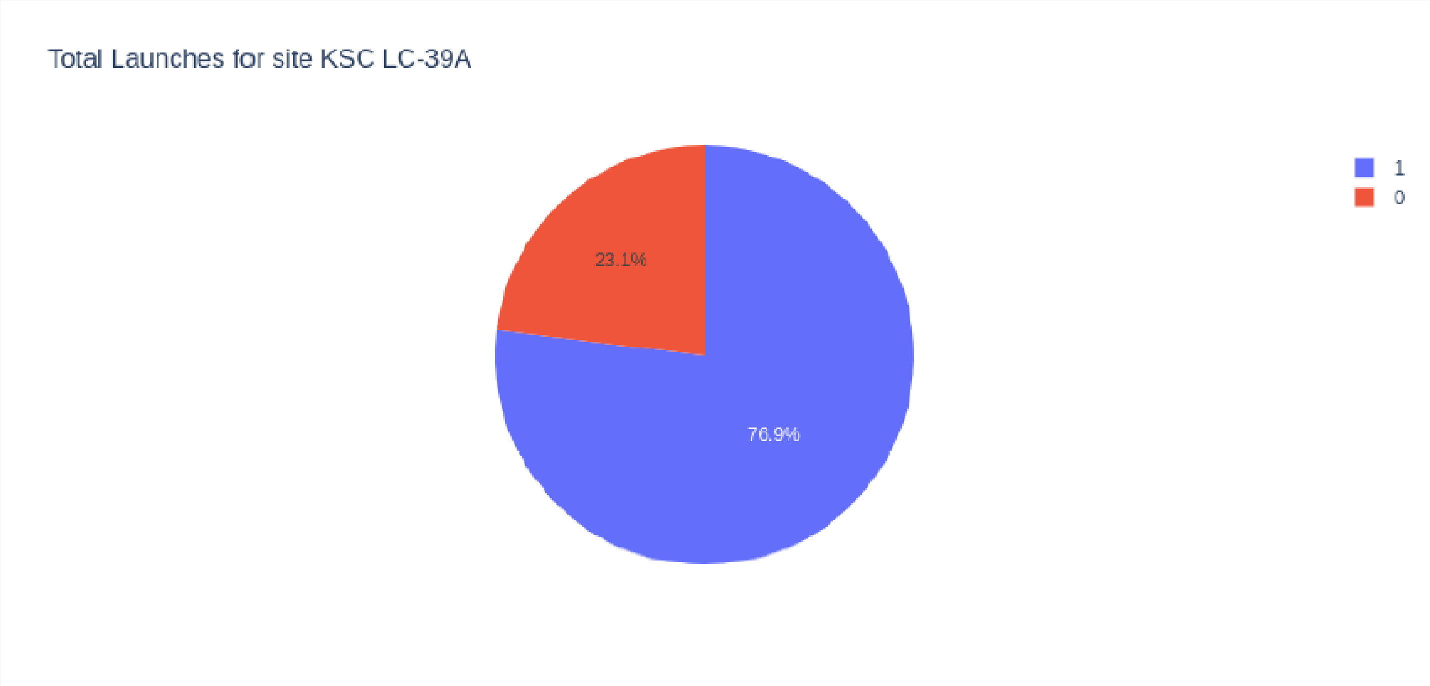
Build a Dashboard with Plotly Dash

Successful Launches – By Site



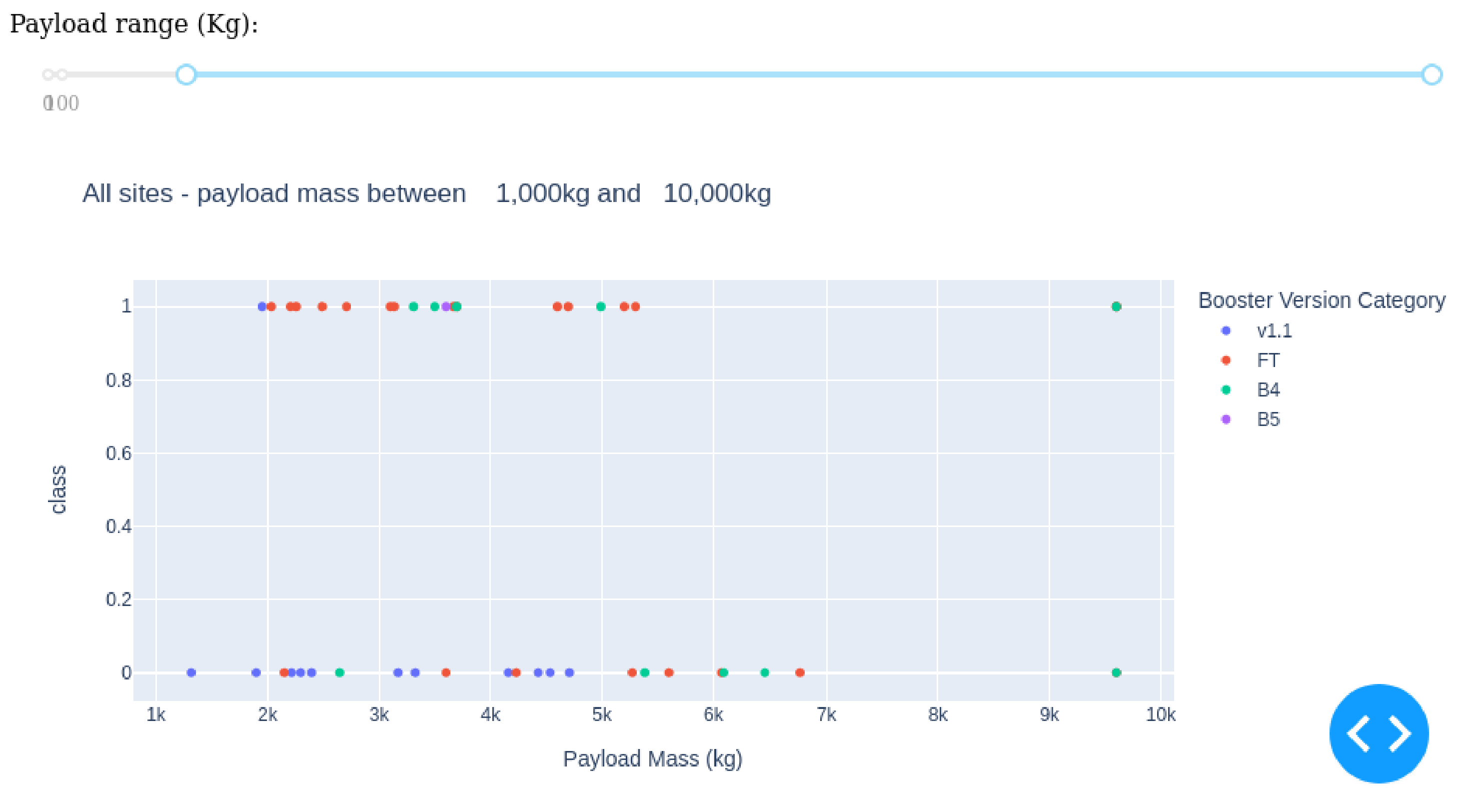
- As we can see in the dashboard, KSC LC-39A is the most successful site with 41,7% of all succeeded launches

KSC LC-39A Launches



- And then, when we zoom in in total launches from KSC we can conclude that this site have both number of succeeded launches and ratio greater than the other sites

Payload Range vs. Launch Outcome



- If we compare the payload (between 1.000 and 10.000kg) with the launch outcome, we can conclude that FT Category have more accuracy than the other booster versions
- And its accuracy increases when the payload keeps between 2.000 and ~3.000kg

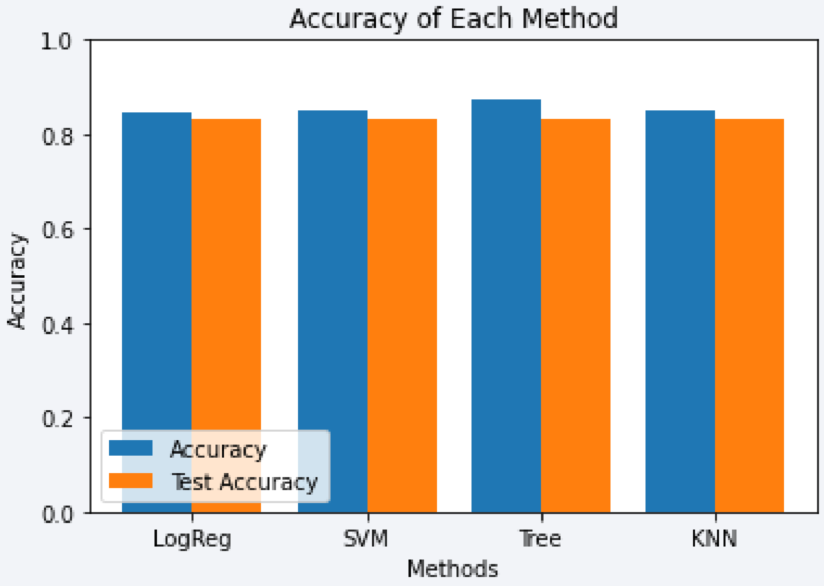
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- The model with the highest classification accuracy is Decision Tree Classifier, which has accuracies over than 87,5%.

Model	Accuracy	TestAccuracy
LogReg	0.84643	0.83333
SVM	0.84821	0.83333
Tree	0.875	0.83333
KNN	0.84821	0.83333



Confusion Matrix

- Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the major problem is false positives.

Conclusions

- Different data sources were analyzed, refining conclusions along the process let us to understand that:
  - The best launch site is KSC LC-39A;
  - Launches between 1.000kg and 10.000kg are less successful than above;
  - If we need to launch with less than 10.000kg, is better to launch with payload between 2.000 and 3.00kg;
- The launch sites have strategic locations which can favor the transport being near railroads, roads and can keep the distance from high population;
- Most of successful mission and landing outcomes seems to improve over time according the evolution of process and rockets;
- Decision Tree Classifier can be used to predict successful landings and increase profits.

Appendix

[Analyses Capstone .GitHub](https://github.com/pedropmacedo/applied-capstone/blob/master/Spacex.csv)

- https://github.com/pedropmacedo/applied-capstone/blob/master/Spacex.csv

Thank you!