IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Masoud N.
October 2022

Outline

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

Executive Summary

- **Below methodologies were used to accomplish the project:**
- Web Scraping and SpaceX API for Data Collection
- EDA using SQL, Data Visualization and Interactive Visual Analytics
- Machine Learning for classification

- **Summary of all results**
- Finding the most important features to predict success of launches
- Being able to predict success of launching with accuracy of ~90%

Introduction

- **As a member of data science team at SpaceY, I want to find a solution to compete with SpaceX**

- **Objectives:**
- How to reduce cost of each launch by predicting successful landing of first stage
- What are the essential factors affecting launching
- Best station of launching

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using SpaceX API and Web Scraping the related Wikipedia web page
- Perform data wrangling
  - Cleaning the data, labeling the outcome and finding crucial features using pandas
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Using Grid Search Cross Validation on some classification algorithms, find the best parameters for the most accurate model to classify the objective

Data Collection

- Data sets were collected from:
- SpaceX API (https://api.spacexdata.com/v4/rockets/)
- Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches), using web scraping technics with Request and BeautifulSoup libraries.

Data Collection – SpaceX API

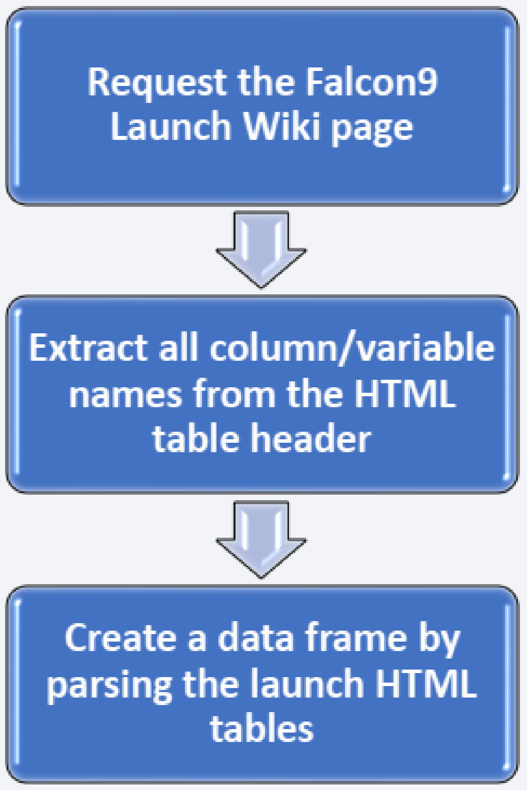
- Using SpaceX REST API, the required data collected according to the flowchart:

- Source Code:

- https://github.com/absurdlyhard/Applied_Data_Science_Capstone/blob/master/Data%20Collection%20API.ipynb

Data Collection - Scraping

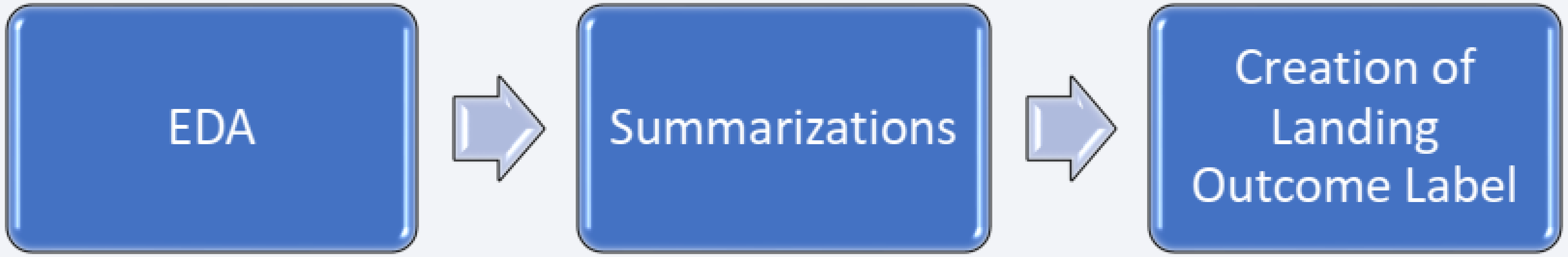
- Wikipedia page on the same subject were scraped according to the following flowchart:



- Source Code:
  - https://github.com/absurdlyhard/Applied_Data_Science_Capstone/blob/master/Data%20Collection%20with%20Web%20Scraping.ipynb

Data Wrangling

- Conducting EDA on the collected dataset
- Summaries of launches for each site, occurrences of each orbit and occurrences of mission per orbit type were claculated.
- At the end, the landing outcome lanel were extracted from outcome column



- Source Code: https://github.com/absurdlyhard/Applied_Data_Science_Capstone/blob/master/EDA.ipynb

EDA with Data Visualization

- Scatter plot, bar plot and line plot were used to find insights from below feature pairs:

- Payload Mass vs Flight No.
- Launch Site vs Flight No.
- Launch Site vs Payload Mass
- Orbit vs Flight No.
- etc.

- Source Code: https://github.com/absurdlyhard/Applied_Data_Science_Capstone/blob/master/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

- **The following SQL queries were performed using SQLite:**

- Names of the unique launch sites in the space mission;
- Top 5 launch sites whose name begin with the string 'CCA';
- Total payload mass carried by boosters launched by NASA (CRS);
- Average payload mass carried by booster version F9 v1.1;
- Date when the first successful landing outcome in ground pad was achieved;
- Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg;
- Total number of successful and failure mission outcomes;
- Names of the 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; and
- 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.

- Source code: https://github.com/absurdlyhard/Applied_Data_Science_Capstone/blob/master/EDA%20with%20SQL%20(Using%20SQLite).ipynb

Build an Interactive Map with Folium

- **Markers, circles, lines and marker clusters were used with Folium Maps**

- Markers indicate points like launch sites;
- Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center;
- Marker clusters indicates groups of events in each coordinate, like launches in a launch site; and
- Lines are used to indicate distances between two coordinates.

- Source code: https://github.com/absurdlyhard/Applied_Data_Science_Capstone/blob/master/Interactive%20Visual%20Analytics%20with%20Folium%20Lab.ipynb

Build a Dashboard with Plotly Dash

- **The following graphs and plots were used to visualize the data:**
- Percentage of Launches by site (Pie Chart with a drop-down menu)
- Payload Range (Scatter Plot with a slider)

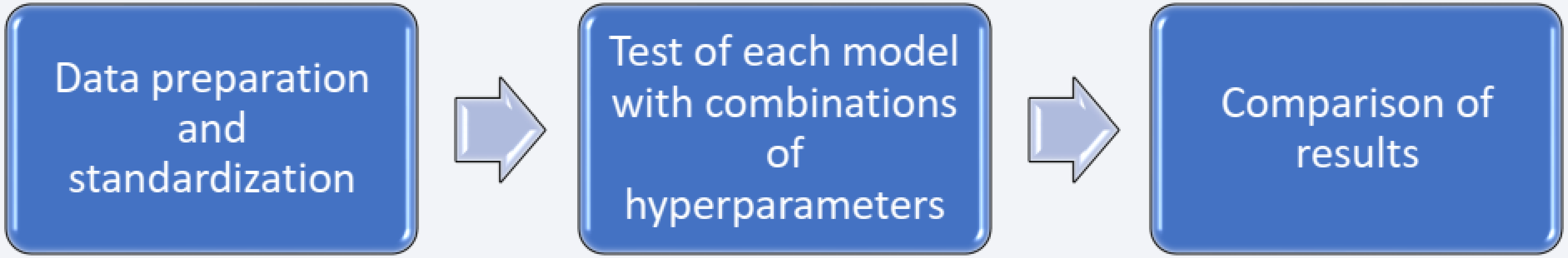
- **Using interactive charts and dashboard enable us to analyze the data in real time by tweaking the features**

- Source code:

https://github.com/absurdlyhard/Applied_Data_Science_Capstone/blob/master/SpaceX%20Dash%20App.py

Predictive Analysis (Classification)

- Four classification models (Logistic Regression, Support Vector Machine, Decision Tree and KNN) were built.
- Using GridSearchCV best hyper parameters of each were found.



- Source code:
- https://github.com/absurdlyhard/Applied_Data_Science_Capstone/blob/master/Machine%20Learning%20Prediction.ipynb

Results

- **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;

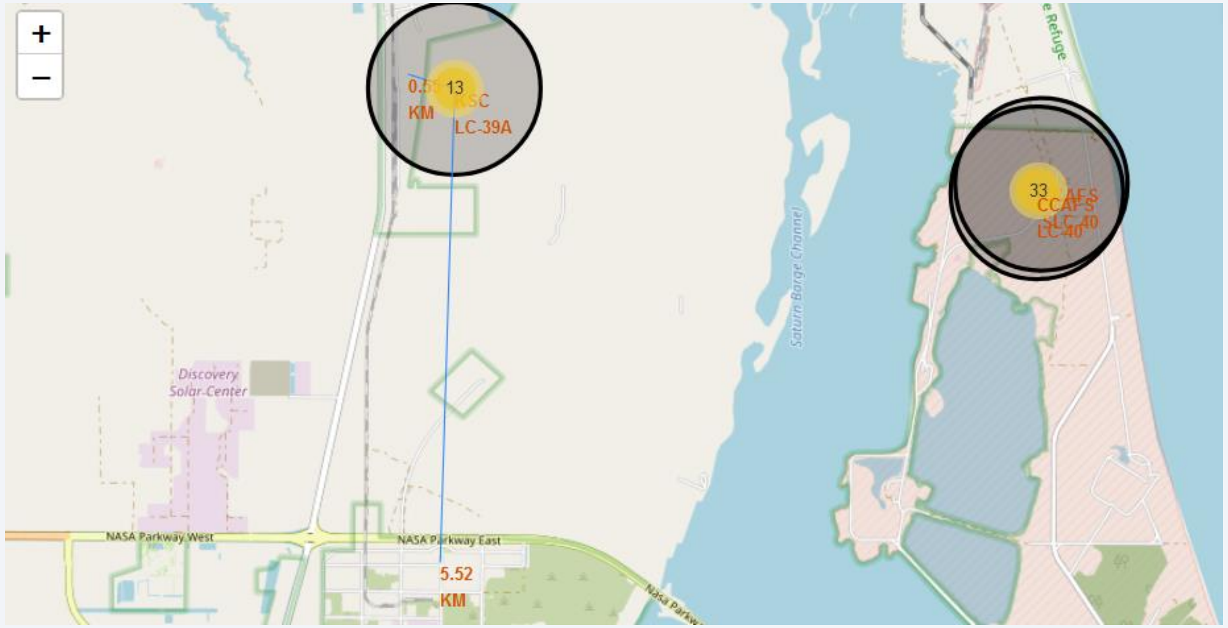
• Almost 100% of mission outcomes were successful;

• 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 Analytics**
- Using interactive analytics, we found distance of strategic places such as railroad, highway, coastline and adjacent city to a launch site



Results

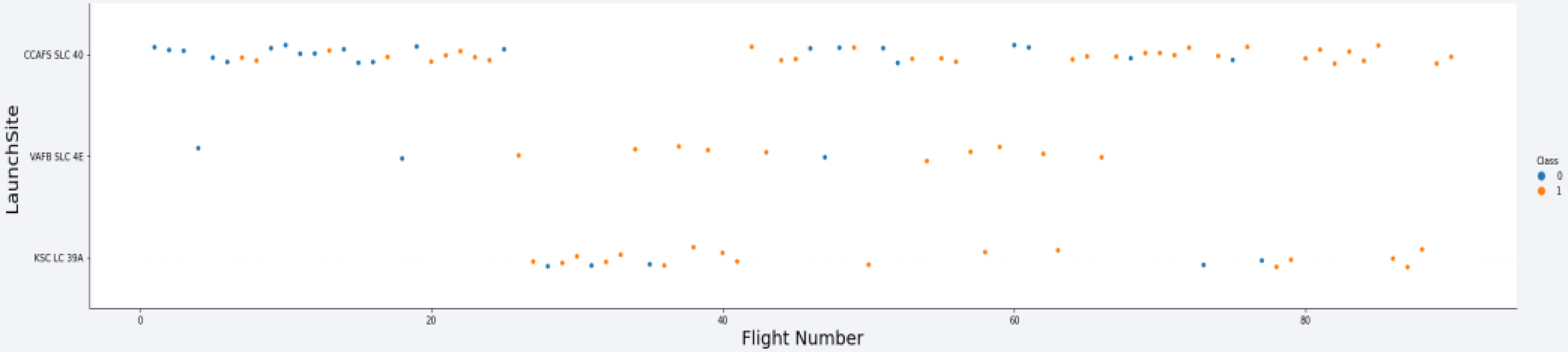
- **Predictive Analysis Results**

- The result showed that the Decision Tree Classifier is the best model to predict success of landing of first stage compared to other models. It has 87% accuracy on train data and around 94% accuracy on the test set.

Section 2

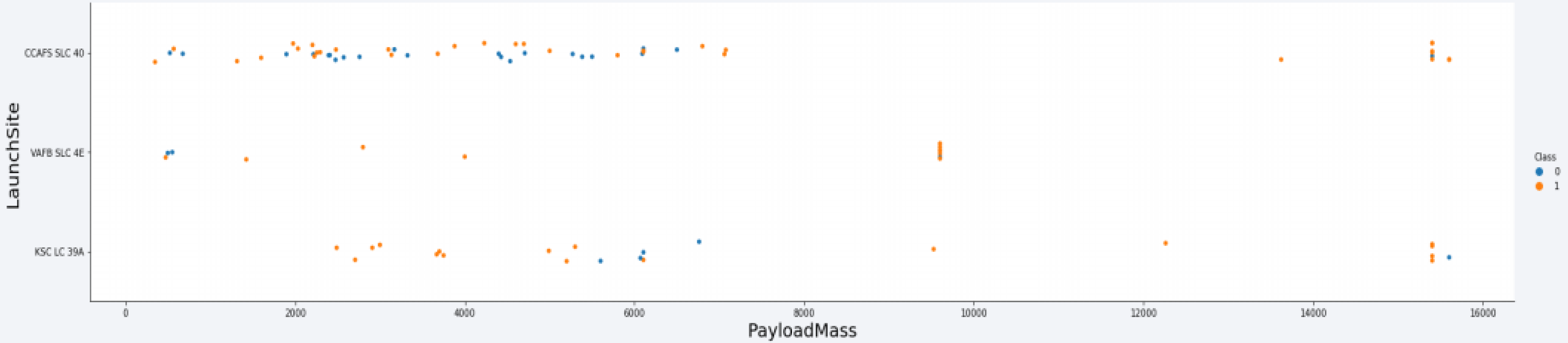
Insights drawn from EDA

Flight Number vs. Launch Site



- The best launch site is CCAF5 SLC 40.
- Rate of success has been increased over time

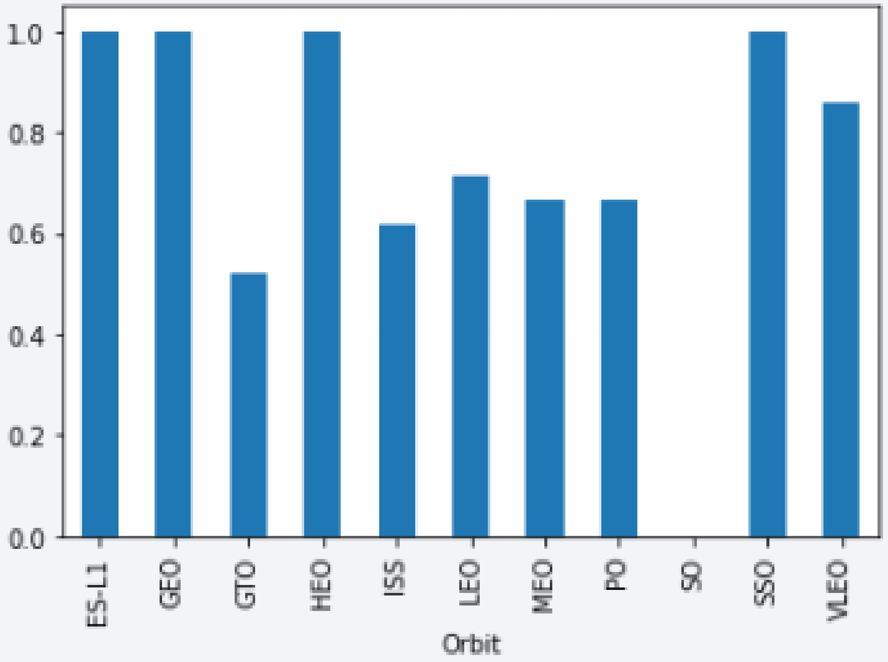
Payload vs. Launch Site



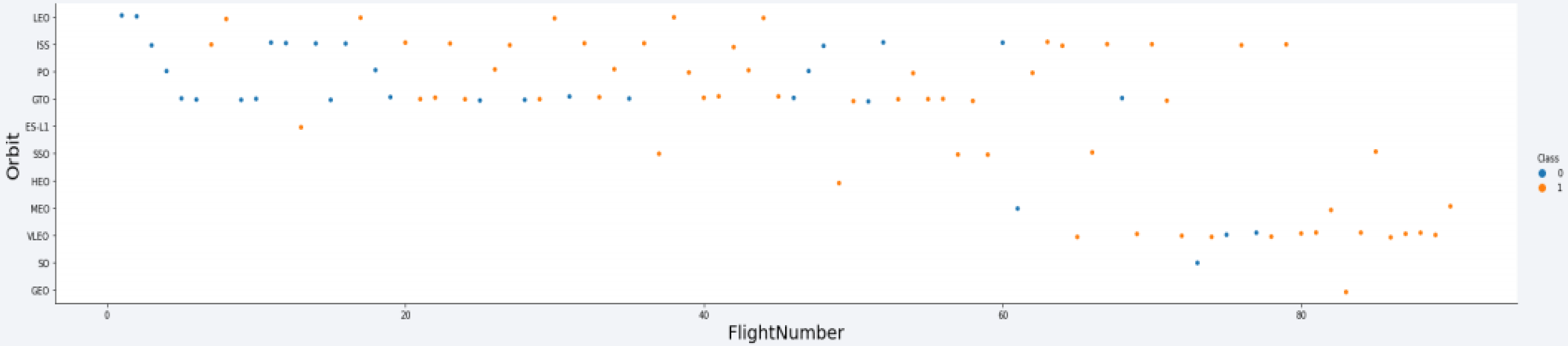
- The more payload mass, the better success rate
- No heavy mass payload (more than 10 ton) were used in VAFB SLC 40 launch site

Success Rate vs. Orbit Type

- The most successful launches were targeted for orbits as follows:
- ES-L1
- GEO
- HEO
- SSO

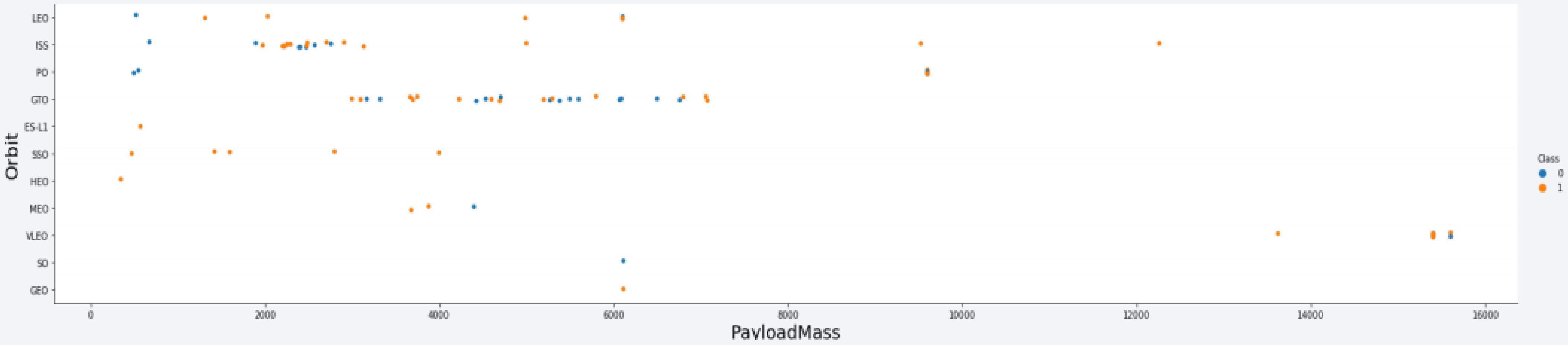


Flight Number vs. Orbit Type



- Success rate increased over time for all the orbits.
- Most recent successful launches has VELO at the orbit target.

Payload vs. Orbit Type

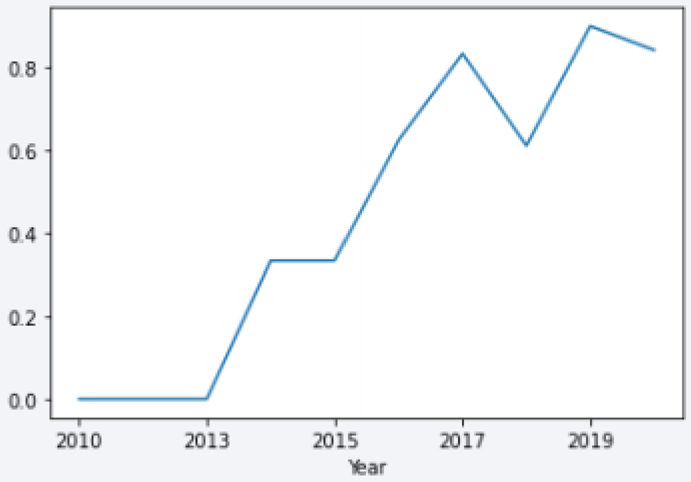


- For GTO, no clear relation can be found between payload mass and success.
- So and Geo has the least launches.
- The heaviest payload masses were launched for VELO.

Launch Success Yearly Trend

- The graph trend shows that success rate has been increased over time.

- In 2018 a drop in success rate can be observed.
- First three years have no tangible achievements.



All Launch Site Names

- Name of all launch sites

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

- Unique name of launch sites provided above table.

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Total payload carried by boosters from NASA

SUM(PAYLOAD_MASS__KG_)
45596

- Summing up all the records containing NASA (CRS) resulted in above figure.

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- Date of the first successful landing outcome on ground pad

Date
22-12-2015

Successful Drone Ship Landing with Payload between 4000 and 6000

- 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

- Total number of successful and failure mission outcomes

Landing _Outcome	QTY
Controlled (ocean)	5
Failure	3
Failure (drone ship)	5
Failure (parachute)	2
No attempt	21
No attempt	1
Precluded (drone ship)	1
Success	38
Success (drone ship)	14
Success (ground pad)	9
Uncontrolled (ocean)	2

Boosters Carried Maximum Payload

- Names of the booster which have carried the maximum payload mass

Booster_Version	
F9 B5 B1048.4	F9 B5 B1049.5
F9 B5 B1049.4	F9 B5 B1060.2
F9 B5 B1051.3	F9 B5 B1058.3
F9 B5 B1056.4	F9 B5 B1051.6
F9 B5 B1048.5	F9 B5 B1060.3
F9 B5 B1051.4	F9 B5 B1049.7

2015 Launch Records

- 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

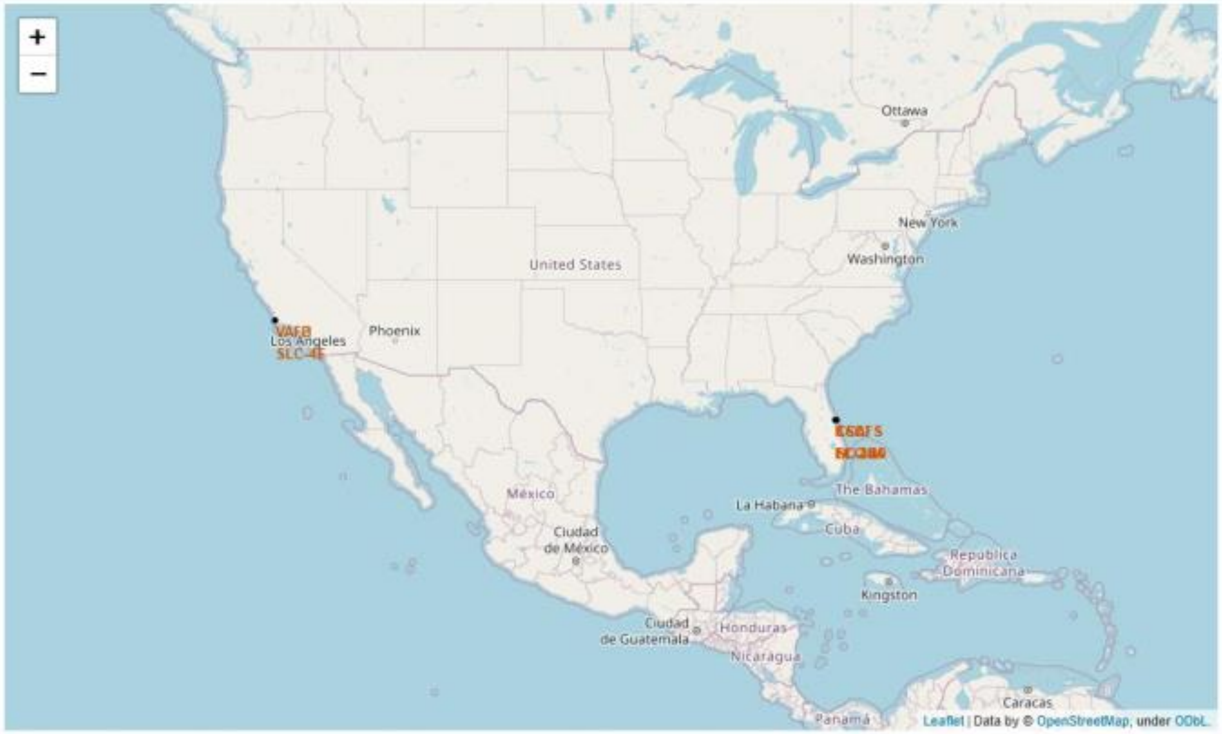
- Rank of quantity of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

Landing _Outcome	QTY
Success	20
Success (drone ship)	8
Success (ground pad)	6

Section 3

Launch Sites Proximities Analysis

Overview of Launch Sites

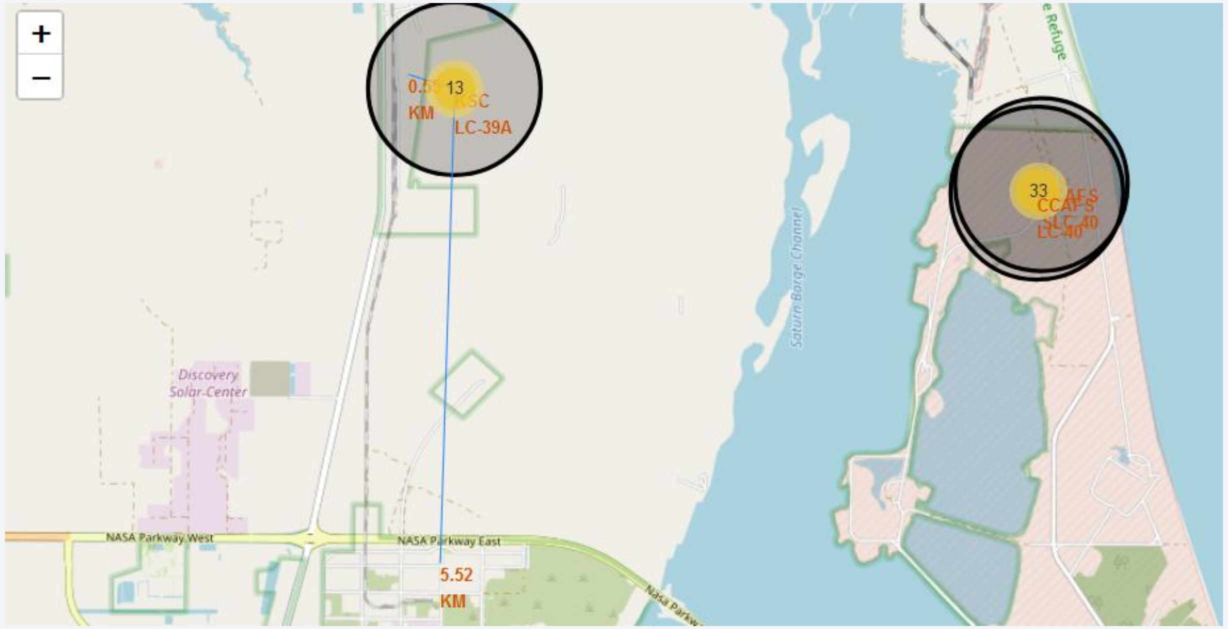


- All sites are located in vicinity of coast lines and near highways and railroads.

Quantity and Launch Results

- Quantity of launches can be seen in yellow circles for each site. In addition, green mark shows successful and red mark shows unsuccessful launch.

Launch Site Distance to Strategic Places in Neighborhood

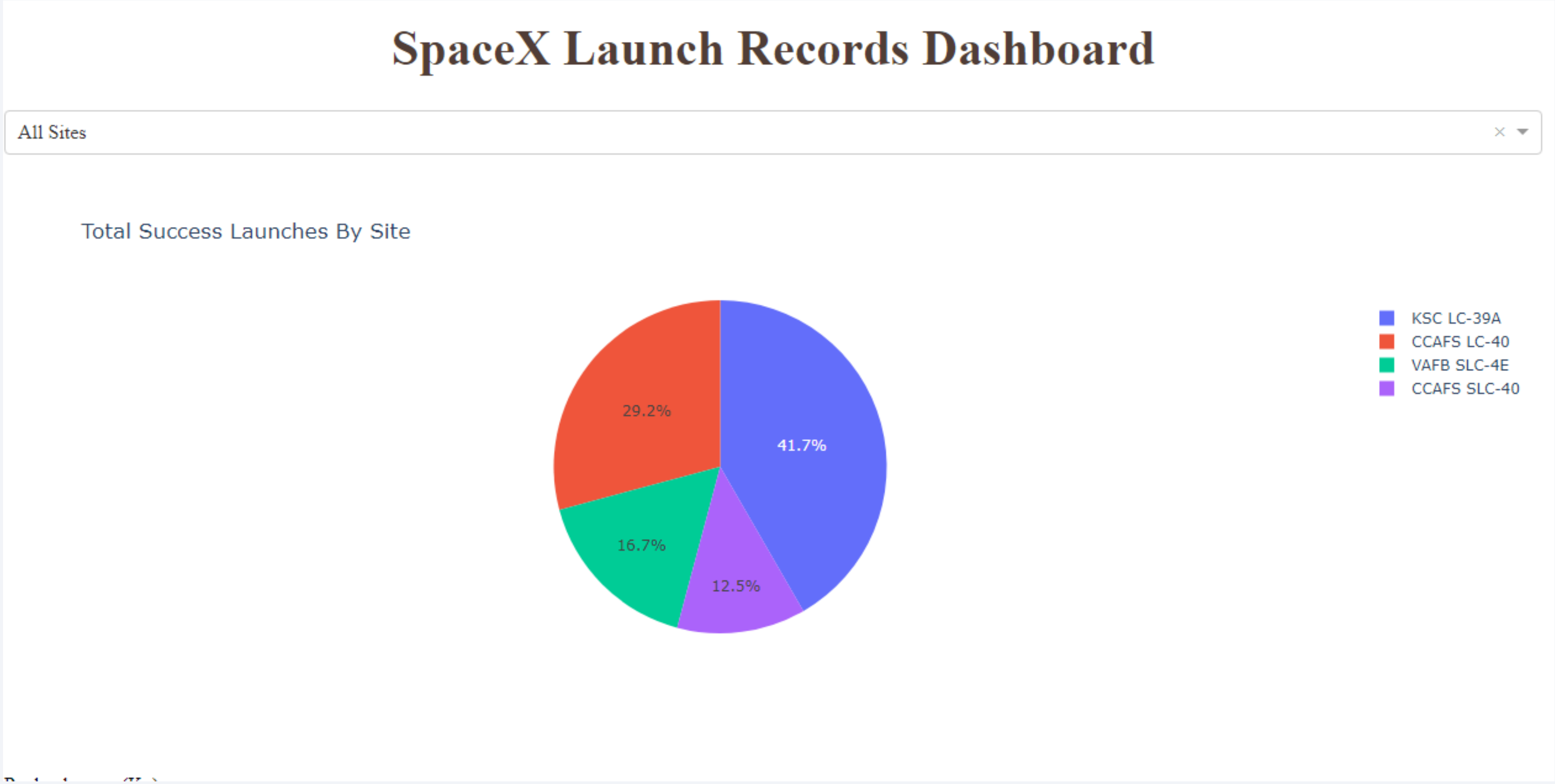


- Distance to strategic points, including railroad, city, etc. in vicinity of one of the launch sites can be observed in above map.

Section 4

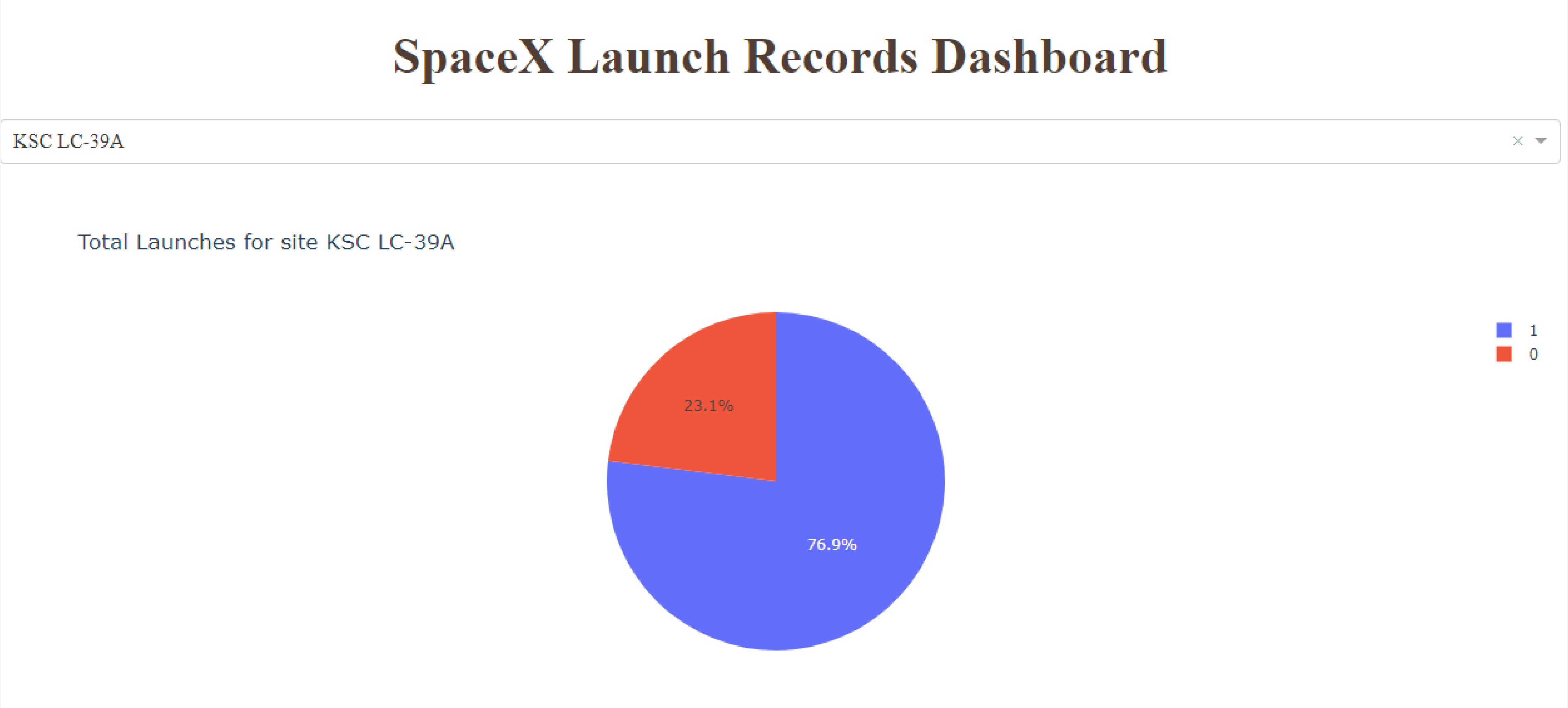
Build a Dashboard with Plotly Dash

Launch Success for each Site



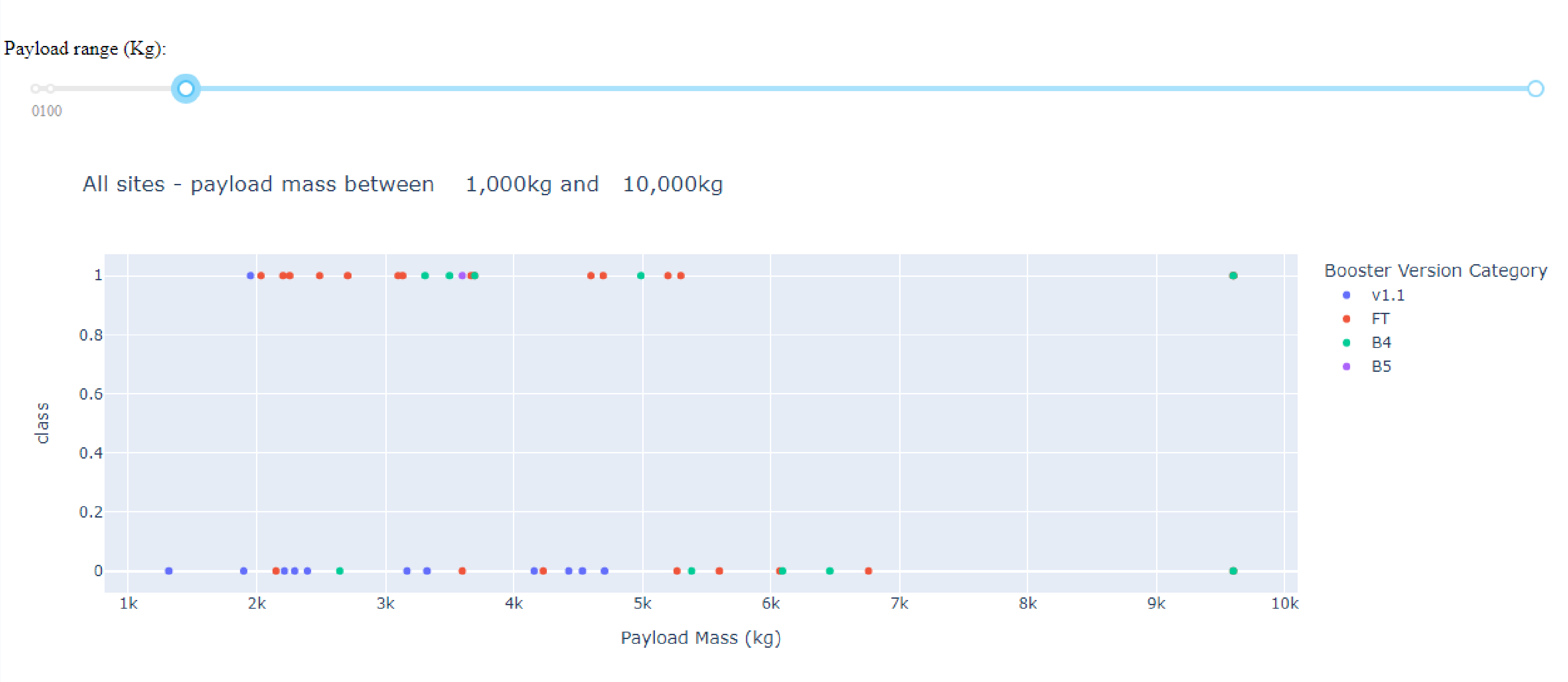
- Quantity of successful launches can be found for each site in the above screenshot.

Total Launches of the Most Successful Site



- ShKSC LC-39A is the launch site with the highest launch success ratio which is 76.9%.

Payload Mass vs. Launch Outcome



- Payload range of 2 to 6 ton and booster version FT have the largest success rate.

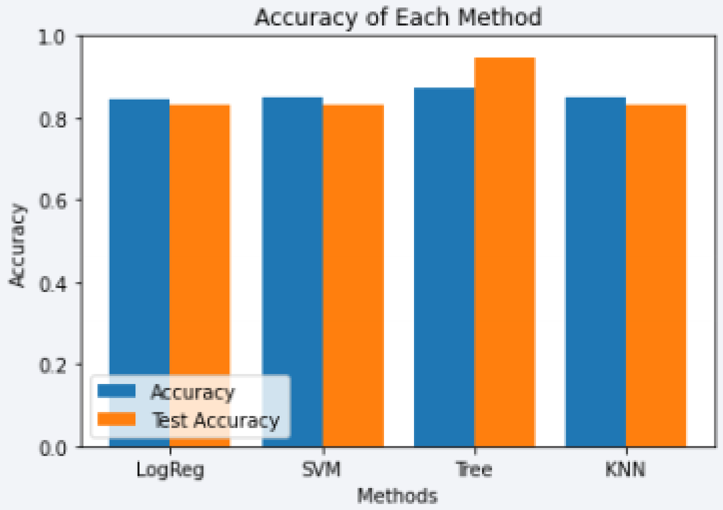
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Accuracy on train and test set for each classification model can be observed in the following bar chart.

- Decision Tree Classifier has the best performance among classification models with 87% accuracy on the train set and 94% accuracy on the test set.



Confusion Matrix

- This is the confusion matrix of the best performing model, Decision Tree Classifier, showing number of TP, TN, FP and FN categories.

Conclusions

- The most successful launch site is KSC LC-39A
- Payload Masses in the range of 2 to 6 ton has the best success rate.
- Heavy payload masses above 7 ton could be risky.
- Due to technological advancements, rate of successful landing of first stage has been improved over time.
- Distance of a launch site from adjacent strategic points, such as highway, railroad, etc. are critical.
- Decision Tree Classifier with best hyper parameters found by GridSearchCV has the best accuracy to predict the success rate of landing.

Appendix

- SQLite worked fine for me, so I decided to use it rather than other options.
- IBM Watson Studio and Google Colab provide fast and flexible environments for coding, testing and debugging data science projects.
- Jupyter Notebook magic commands, such as “%sql”, makes it easier to perform EDA with SQL in python environment.

Thank you!