IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Calixthe MATTEI
27/07/2023

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies
  - Data Collection with SpaceX API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Maps with Folium
  - Interactive Dashboard with Plotly Dash
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive Maps screenshots
  - Interactive Dashboard screenshots
  - Predictive Analytics result

Introduction

- Rocket launches are extremely expensive. In the recent years, SpaceX came with tremendous savings thanks to the reuse of the first stage. Being able to predict if the first stage of the rocket will land allow competitors to predict the cost of their launch.

- Purpose of this project is to analyze previous SpaceX launches to build a model that can predict if the first stage will land.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
    - SpaceX API and Wikipedia Webscraping.
- Perform data wrangling
    - Null values analysis, value counts study and simple outcome analysis.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Standardization, Gridsearch parameter study and scoring.

Data Collection

- Datasets were collected in two manners: using SpaceX API and by webscraping Wikipedia.
- The process for each activity will be further explained in the following slides

Data Collection – SpaceX API

- Link :

SpaceX API Notebook

Workflow

- The JSON file was normalised to become a dataframe
- Data was cleaned to get only relevant Falcon 9 inputs
- Missing values were handled

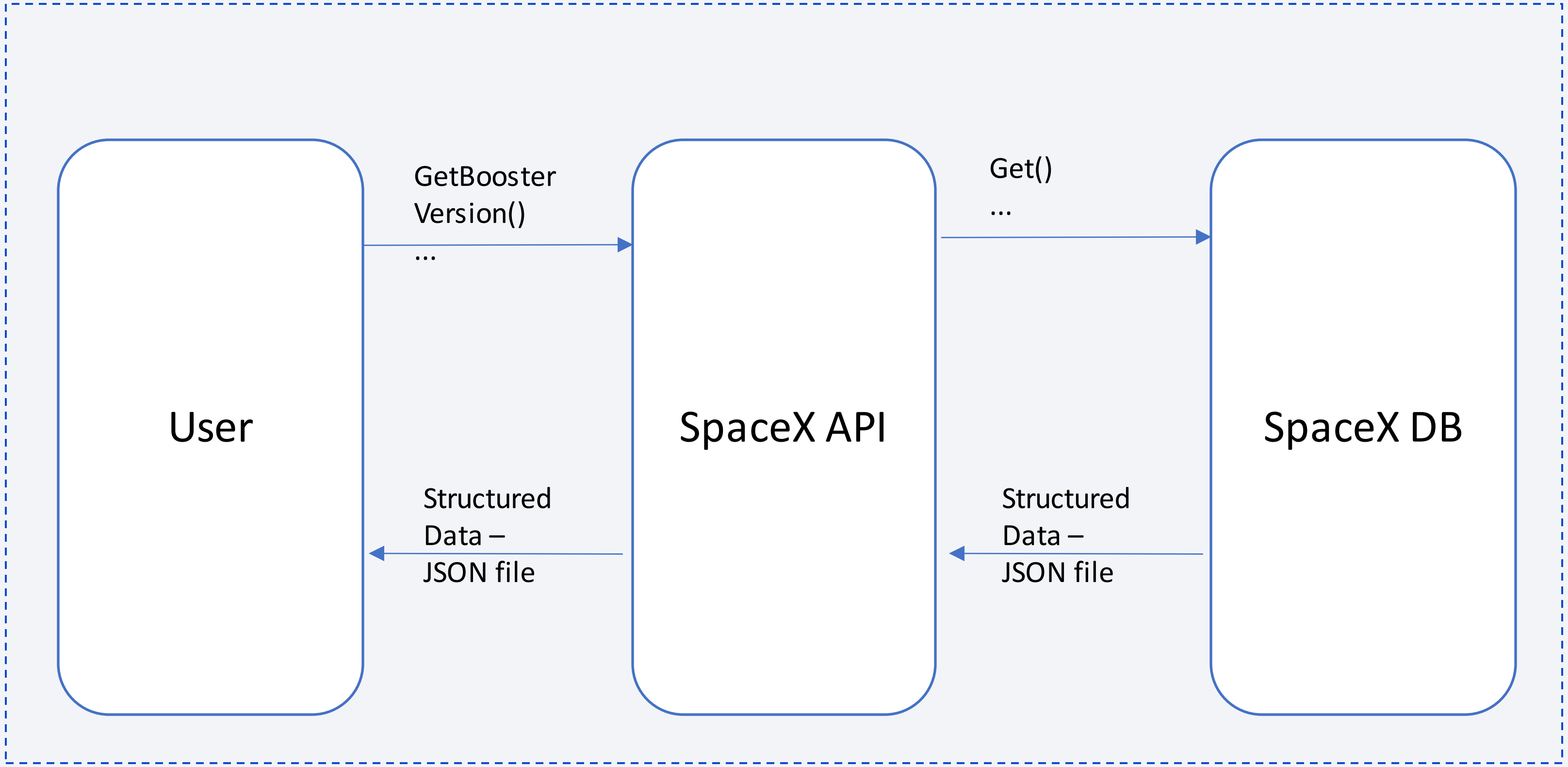


Figure 1: Flowchart of the different activities

Data Collection – Scraping

- Link :

[Webscraping Notebook]

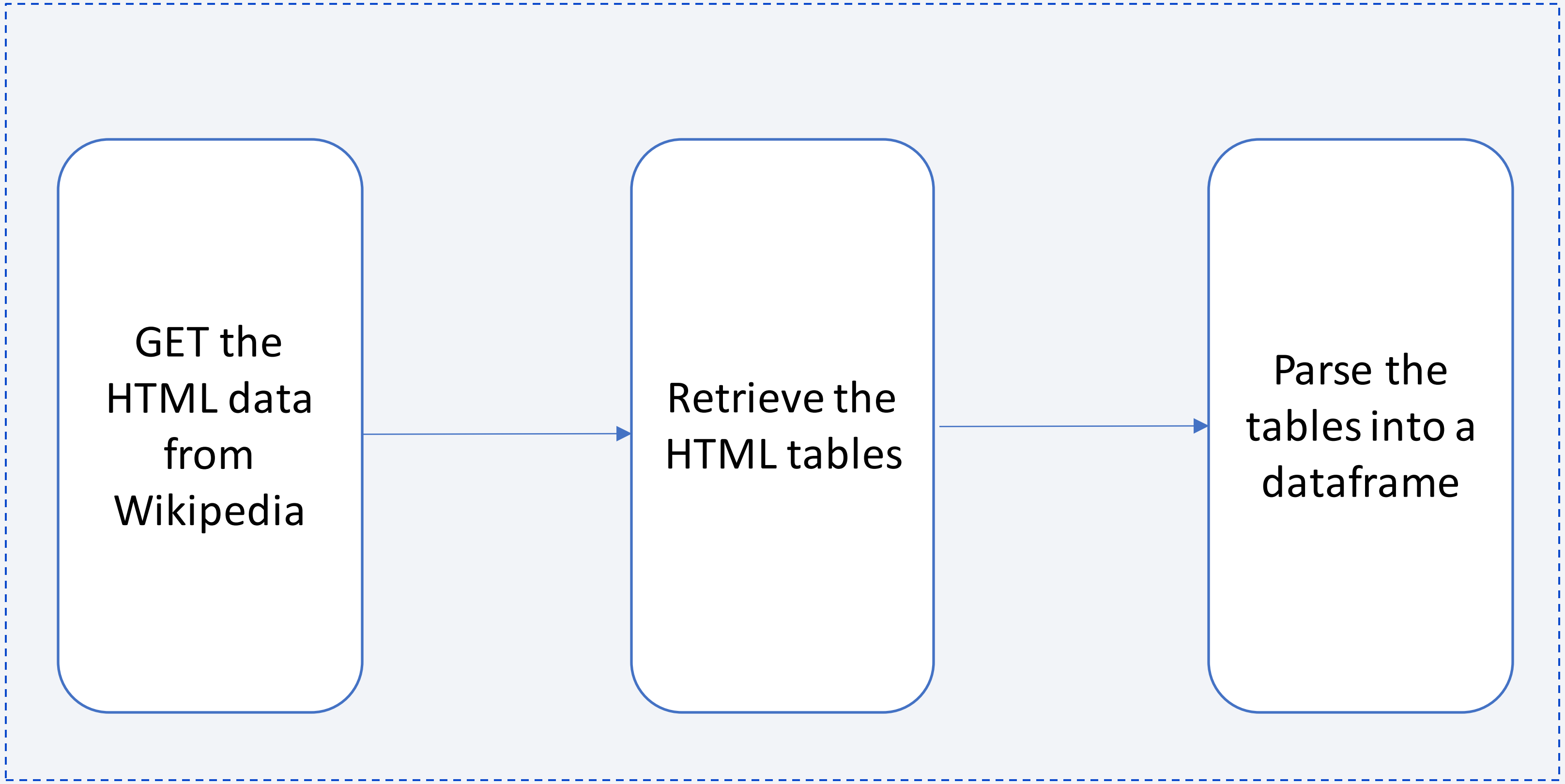


Figure 2: Workflow of the different activities for Webscraping

Data Wrangling

Workflow:

- Types of each column were investigated
- Values were looked at to get a glimpse of the data
- A landing outcome feature was computed

Link:

Data Wrangling Notebook

EDA with Data Visualization

- The purpose of this EDA visualization is to grasp some insights from the data quickly in a visual manner.
  The different plots were:
  - Flight Number vs Launch Site scatter plot
  - Payload vs Launch Site scatter plot
  - Success Rate vs Orbit Type bar chart
  - Flight Number vs Orbit Type scatter plot
  - Payload vs Orbit Type scatter plot
  - Launch Success Yearly Trend line plot

Link:

EDA with Data Visualization Notebook

EDA with SQL

The purpose of this EDA with SQL is to put some numbers in perspective to grasp some insights from the data quickly.

The following requests were performed:

- Names of the unique launch sites in the space mission
- 5 records where launch sites 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 greater than 4000 but less than 6000
- Total number of successful and failure mission outcomes
- Names of the booster versions which have carried the maximum payload mass.
- Records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months 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

Link:

EDA with SQL Notebook

Build an Interactive Map with Folium

- The purpose of this activity is to get some geospatial insights from the data.
- Circles on the launch sites locations, markers regarding their launch success rate and lines for distance calculation analysis were performed.

Link:

[Interactive Maps with Folium]

Build a Dashboard with Plotly Dash

- Pie chart displaying the Success rate for all and each sites, and scatter plot showing the Payload Mass vs Success Rate of all sites were displayed.
  - This allows to quickly understand which site and payload are the most successful.

Link:

[Dashboard with Plotly Dash]()

Predictive Analysis (Classification)

Logistic regression, SVM, Decision Tree and KNN classifier were trained on the data to find the most accurate classification model.

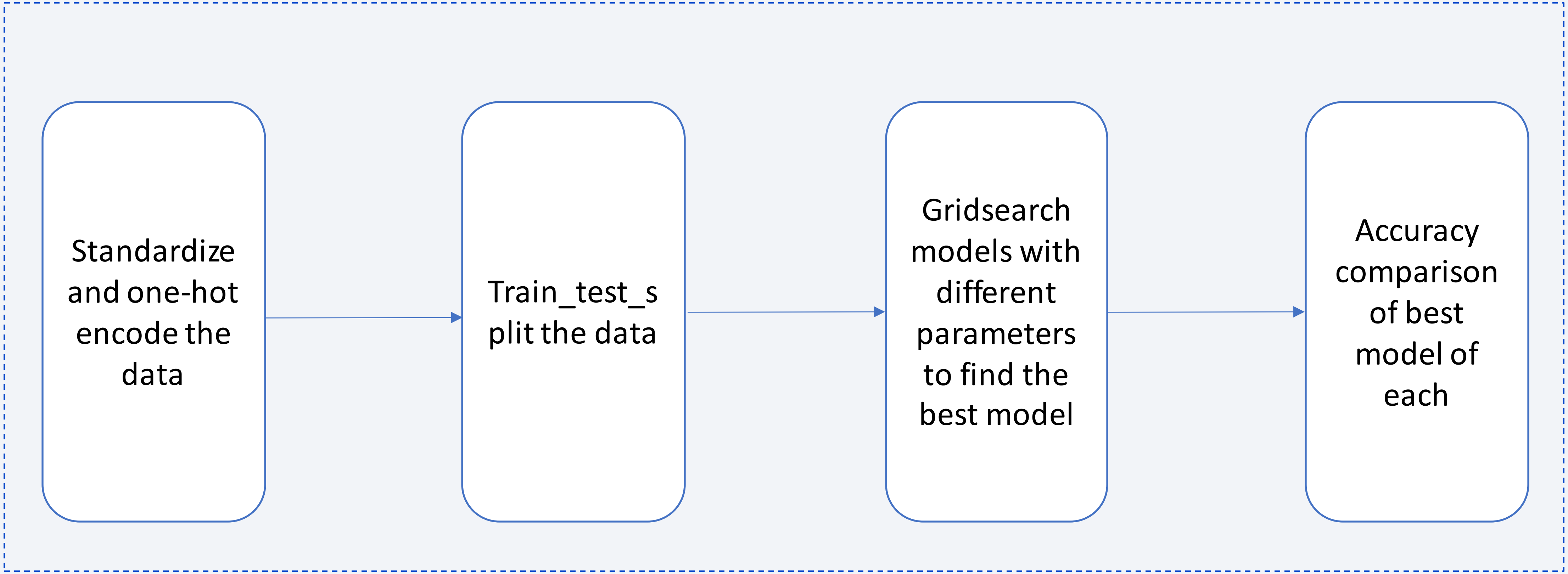


Figure 3: Workflow of the different activities for Predictive Analysis

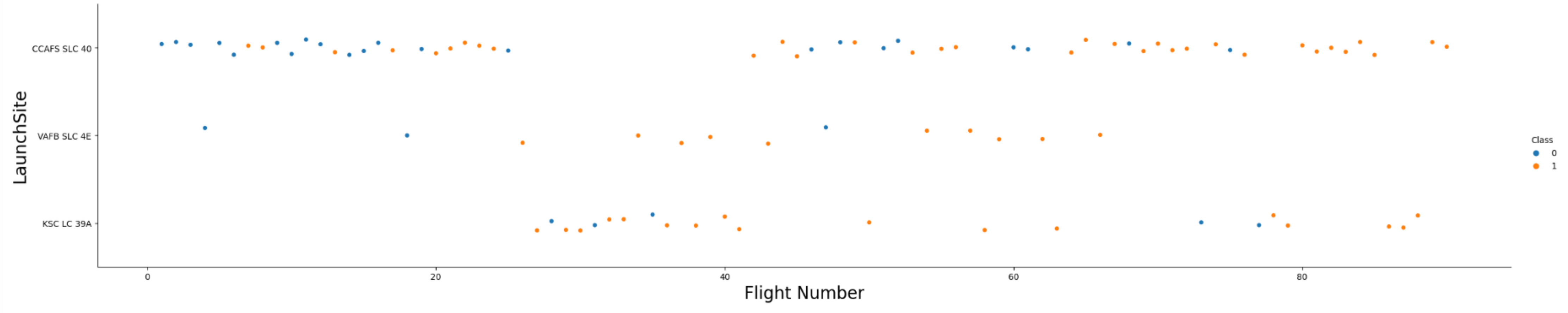
Results

- Insightful data have been retrieved from the different analysis:
  - All launch site success rate, orbit and payload are not equal. More launches and lower payload tends to increase the success rate.
  - The TLR increases, giving an increasing overall success rate.
  - Launch sites share several characteristics in their proximity and location.

Section 2

Insights drawn from EDA

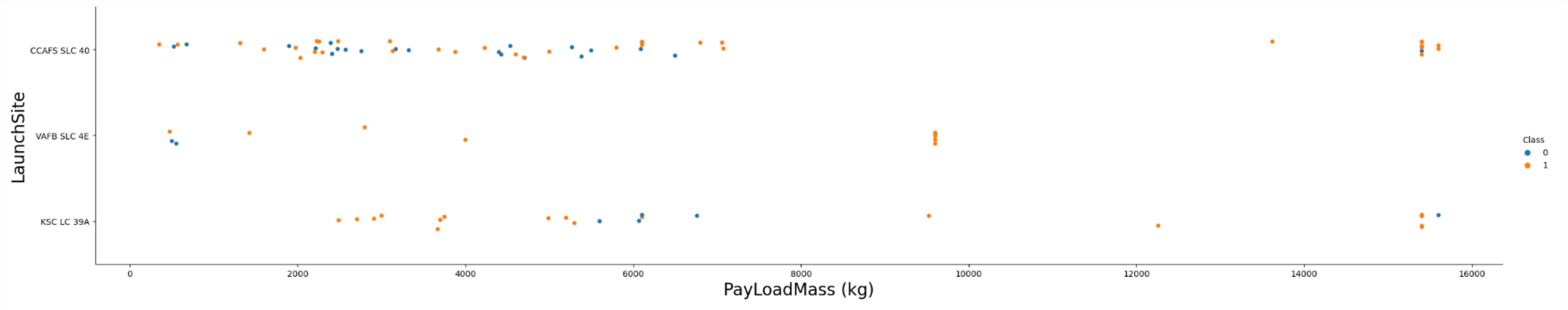
Flight Number vs. Launch Site



2 main insights:

- CCAFS SLC 40 is the main launch site, and as become more mature for launching around the 50th flight
- The two other sites have a high success rate

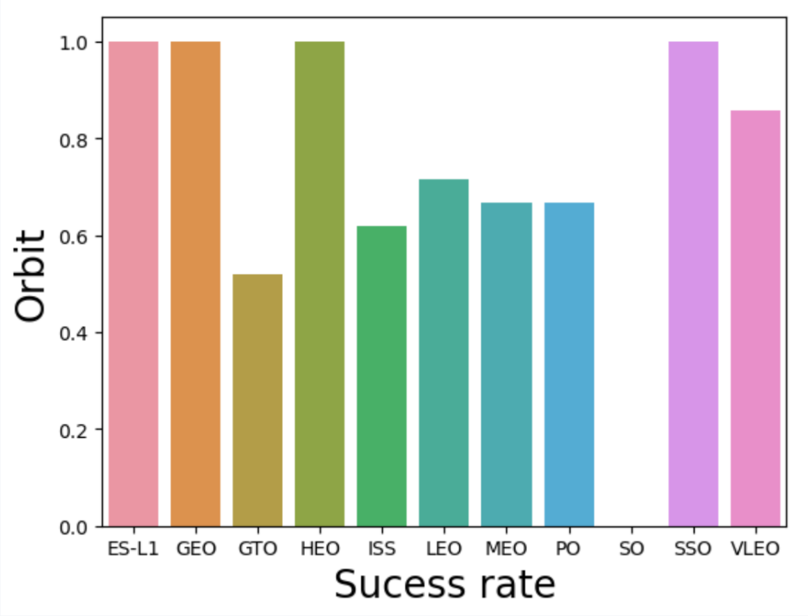
Payload vs. Launch Site



1 main insight:

- The higher the payload, the better the success rate

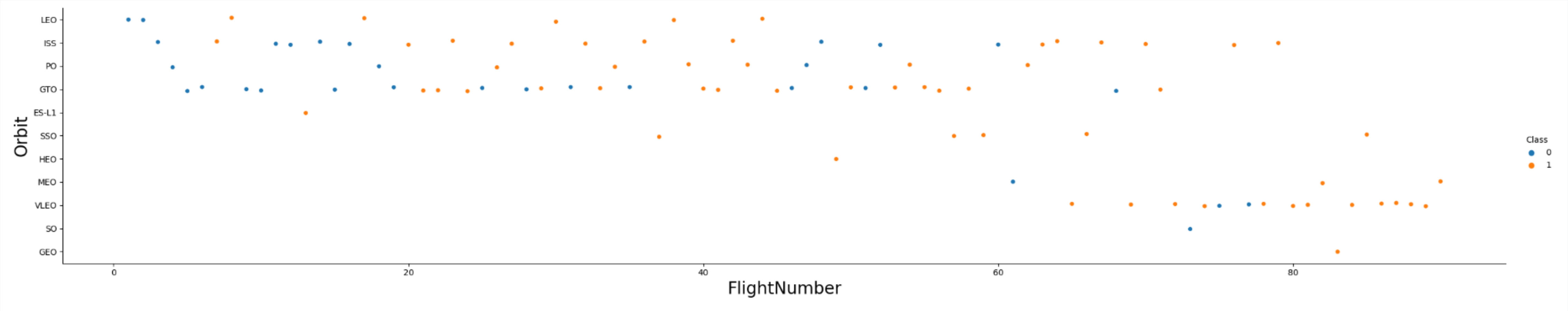
Success Rate vs. Orbit Type



1 main insight:

- All orbits seems to have a 50% + success rate except the GTO – the geosynchronous orbit. ES-L1, GEO, HEO and SSO stand out with their 100% success rate.

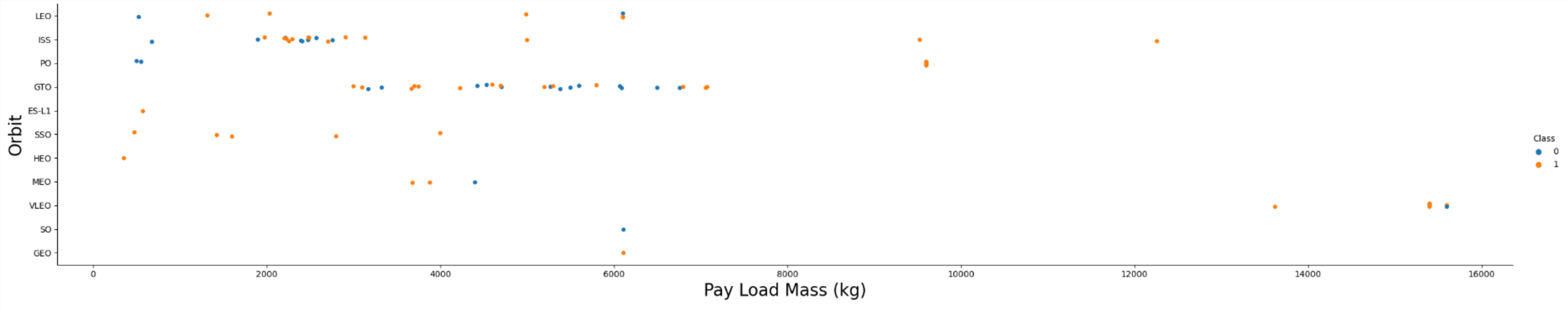
Flight Number vs. Orbit Type



1 main insight:

- The higher the Flight number, the better the success rate. The Technology Level Readiness reached a mature level. Since Flight 80, no failure.

Payload vs. Orbit Type

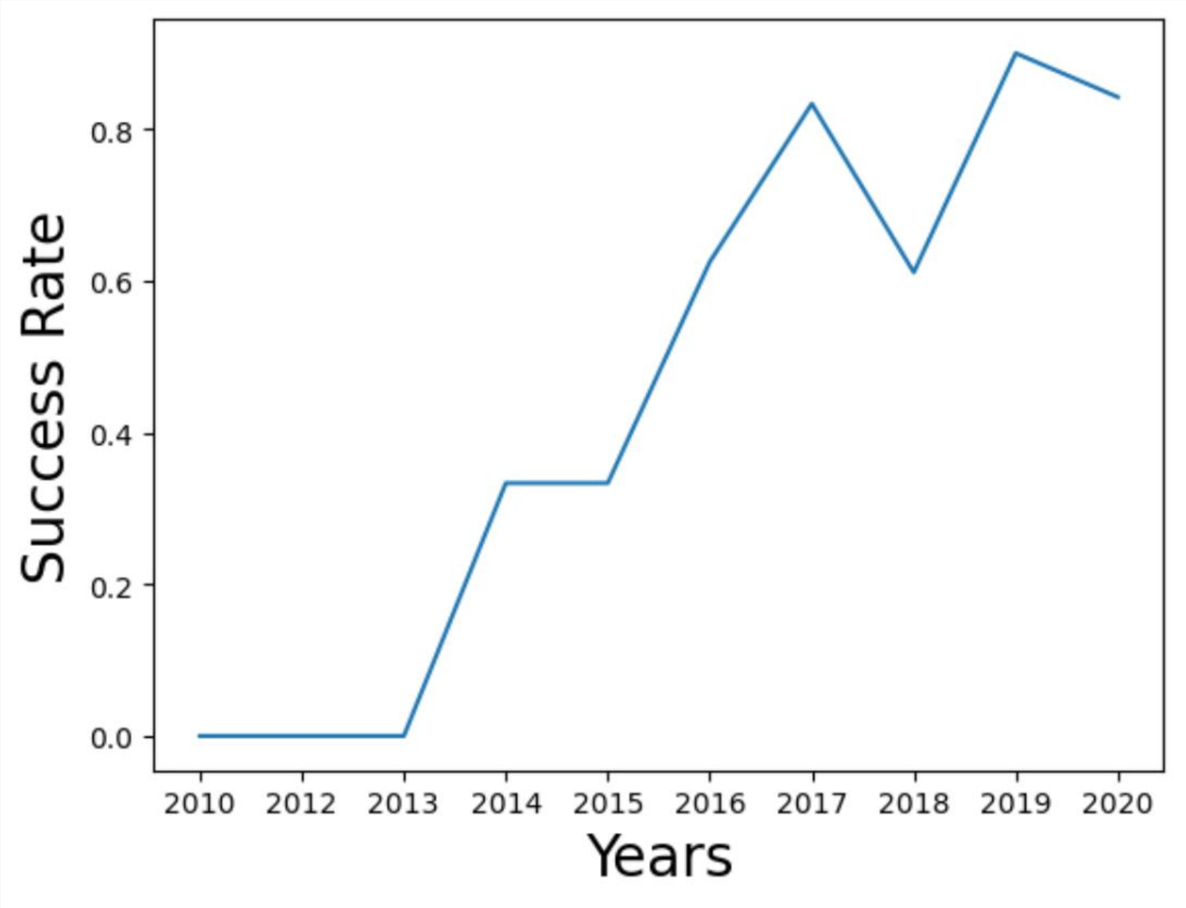


1 main insight:

- The higher the payload, the better the success rate

Launch Success Yearly Trend

- The success rate keeps increasing since 2013.



All Launch Site Names

- SpaceX is using 4 different launch sites.

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`.

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

Total Payload Mass

- The total payload carried by boosters from NASA is 45 596kg.

SUM(PAYLOAD_MASS__KG_)
45596.0

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2928.4.

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- The first successful landing outcome on ground pad occurred the 01/08/2018.

MIN(Date)
01/08/2018

Successful Drone Ship Landing with Payload between 4000 and 6000

- 4 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 is respectively 61 and 10.

num_successes	num_failures
61	10

Boosters Carried Maximum Payload

- The boosters which have carried the maximum payload mass are the following ones.

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

2015 Launch Records

- The 2 failed landing in drone ship, their booster versions, and launch site names for year 2015 can be found below.

Month	Landing_Outcome	Booster_Version	Launch_Site
10	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

- The count of each landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order, can be found below.

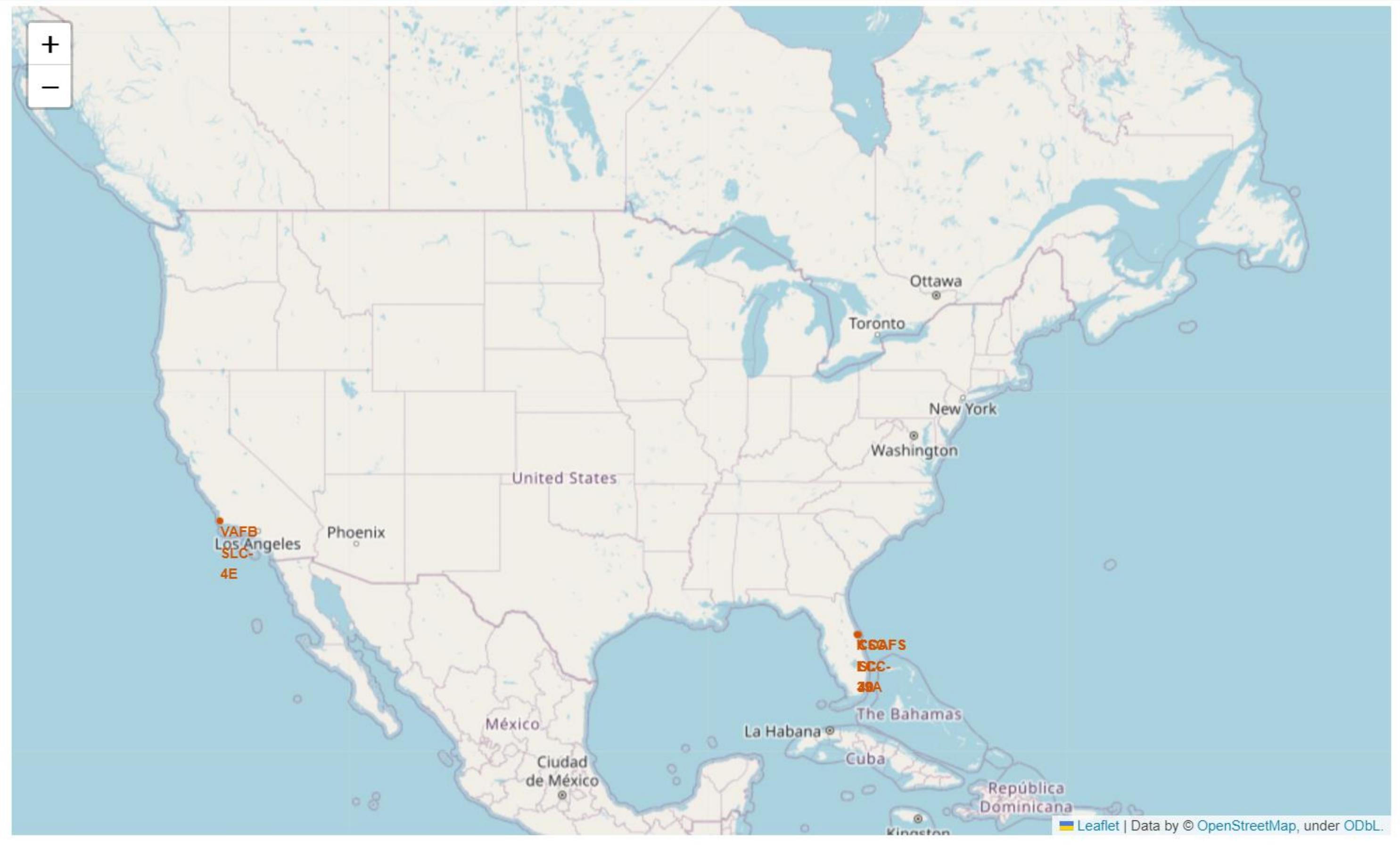
Landing_Outcome	Rank
Success	20
No attempt	9
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1

Section 3

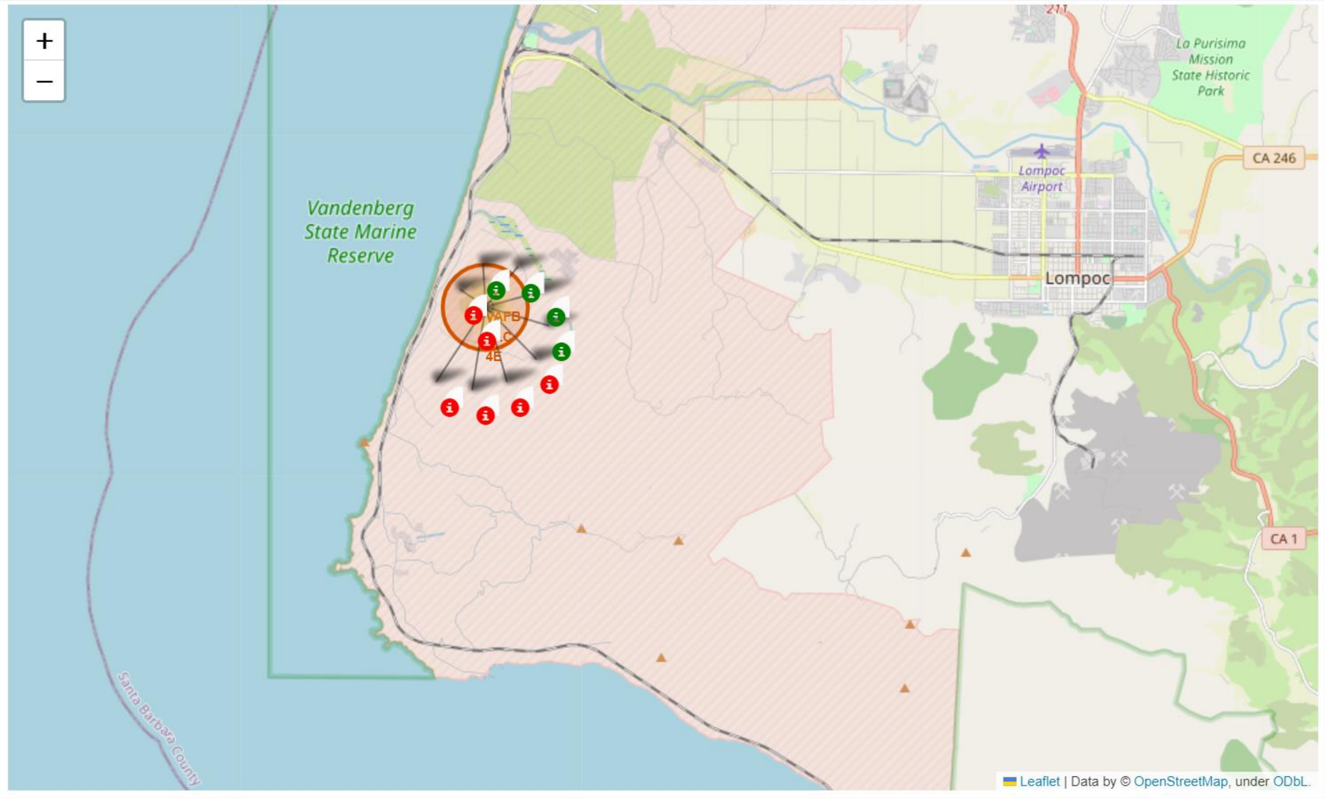
Launch Sites Proximities Analysis

Launch Site locations

- Launch sites of SpaceX can be found on both coasts of the USA, mainly in the southern area.



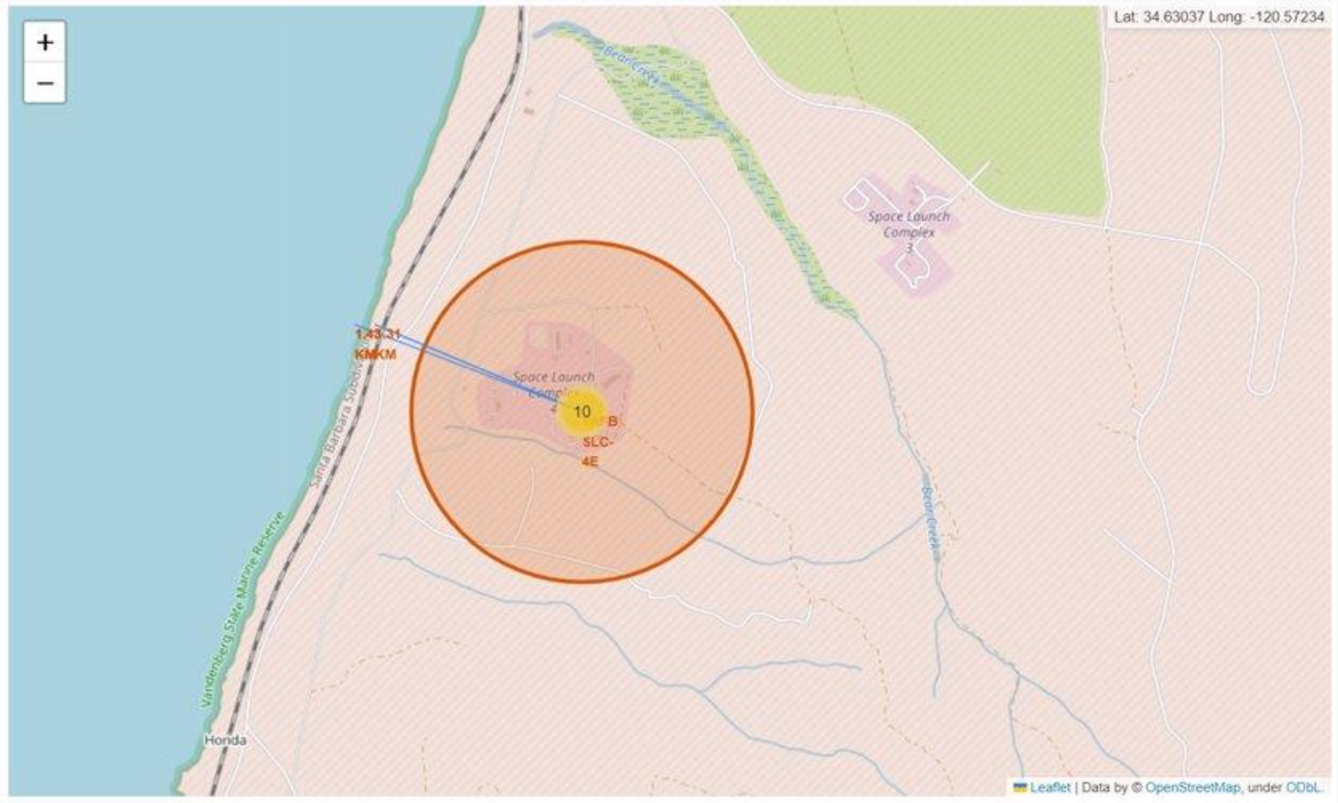
Launch outcomes for VAFB SLC-4E



- Launch outcomes can be easily displayed on the map. This launch site have had quite a few failures.

Launch Site proximities - **VAFB SLC-4E**

- Water proximity (< 2km) – probably to avoid debris to hurt neighboring accomodations
- Railway proximity (< 2km) – probably to be able to receive/deliver cumbersome parts.
- (not displayed) City proximity: cities are further away but close enough to allow employees to live nearby.
- (not displayed: Highway proximity: no highways are nearby.

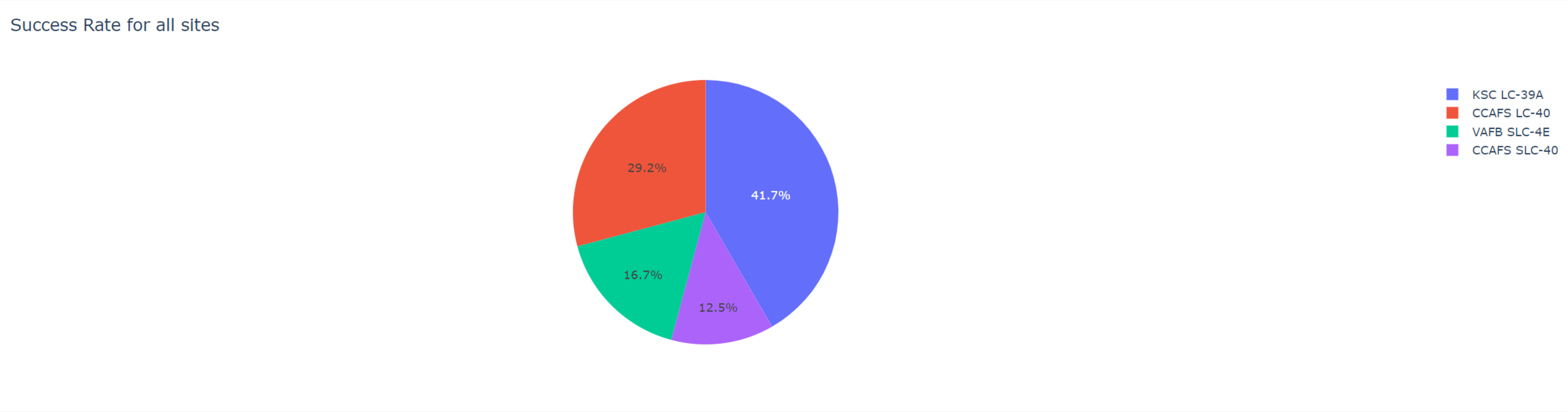


Section 4

Build a Dashboard with Plotly Dash

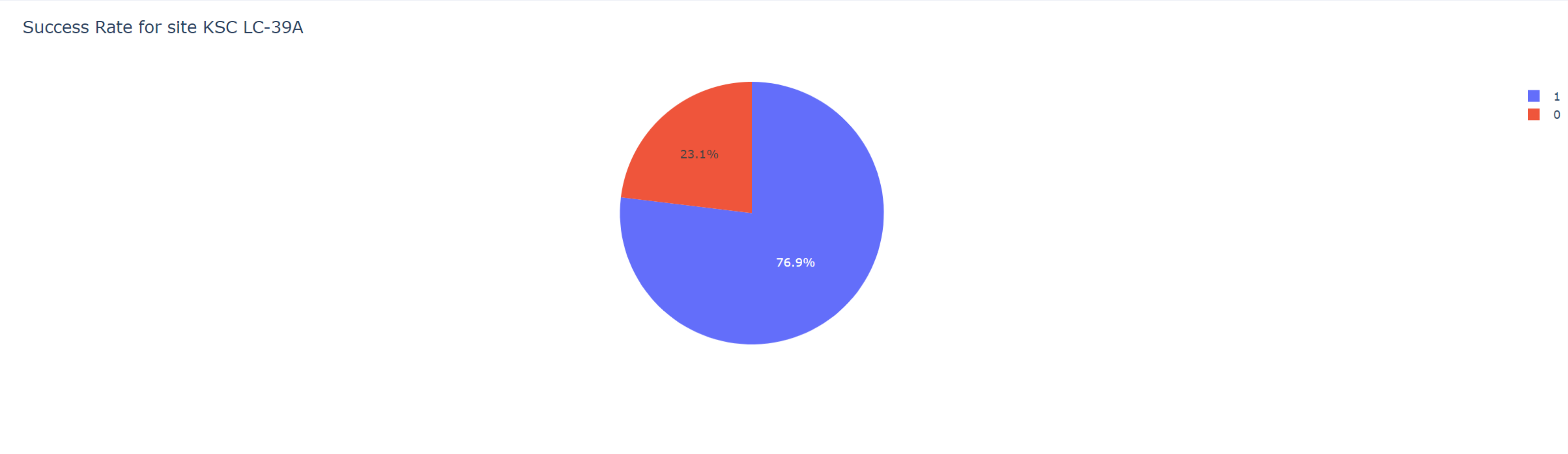
Success Rate – Launch Sites

- The most successful site is the KSC LC-39A.



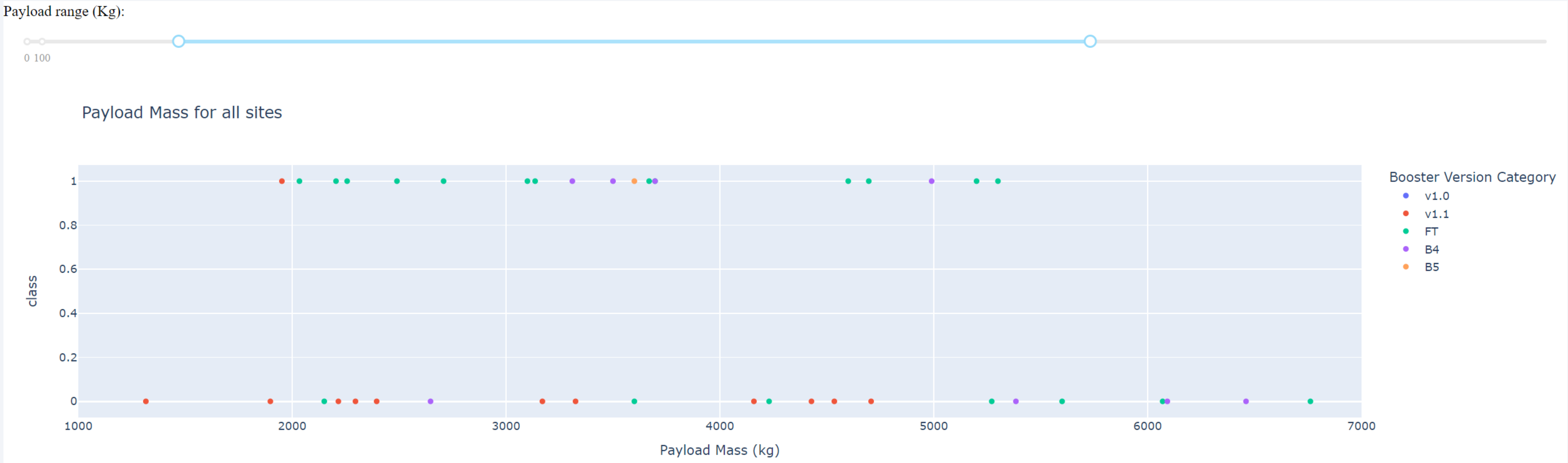
Most successful Launch Site

- The most successful site has a success rate of 76.9%.



Payload and Success Rate analysis

- Higher launch rate on payload below 4000kg.
- More success on payload below 4000kg.



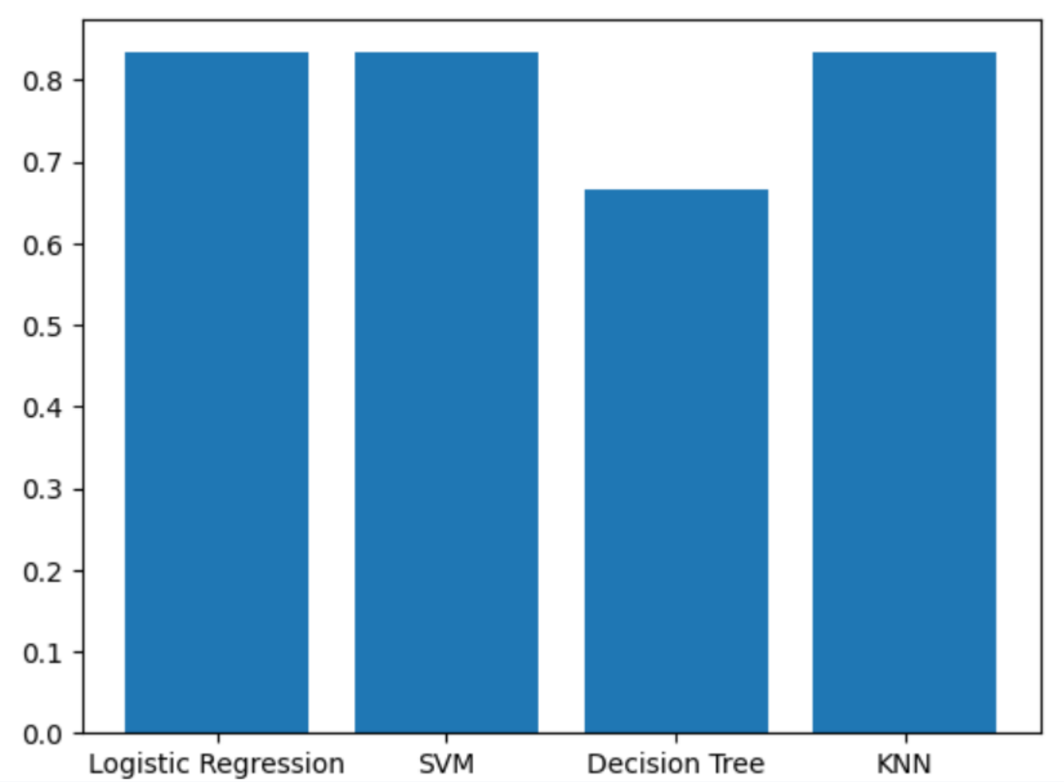
Section 5

Predictive Analysis (Classification)

Classification Accuracy

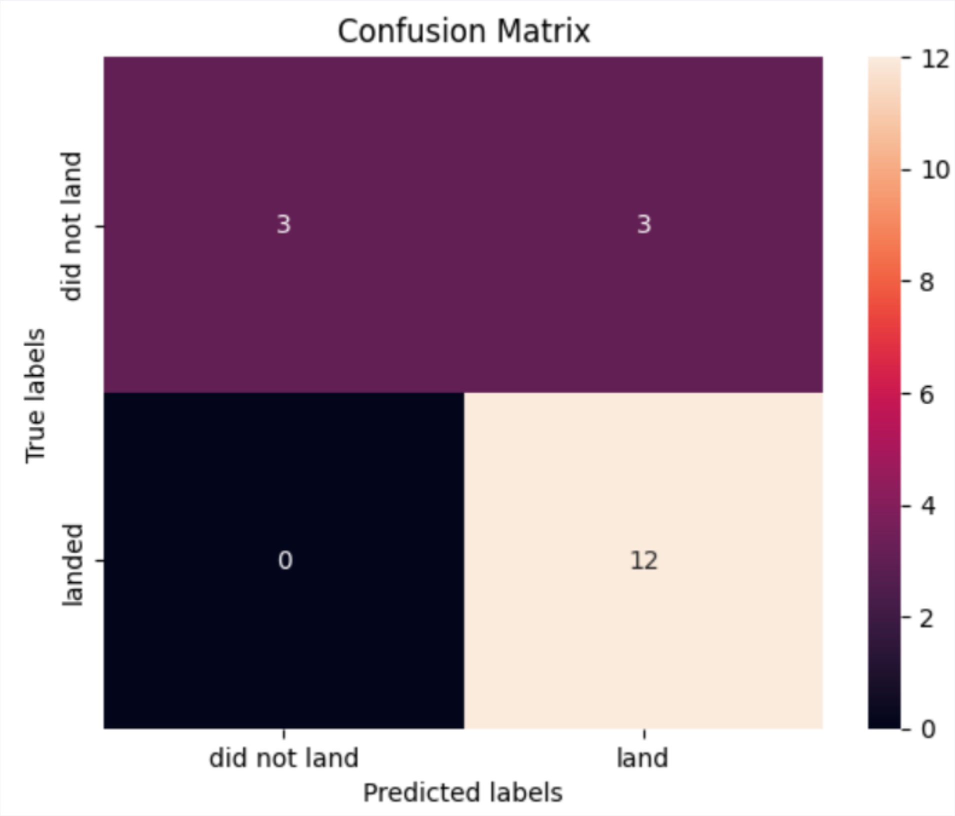
The logistic regression, SVM and KNN models have the same best classification accuracy on the test data of 83.3%.

The Logistic regression has been chosen for the rest of the analysis.



Confusion Matrix

- Row 2 shows us that all real landed outcomes were correctly classified. No false negative.
- The first row shows that out of 6 not-landed scenarios, 3 were classified as landed. These false positives are quite problematic.



Conclusions

- SpaceX data has been extensively processed.
- Insightful data have been retrieved from the different analysis:
  - All launch site success rate, orbit and payload are not equal. More launches and lower payload tends to increase the success rate.
  - The TLR increases, giving an increasing overall success rate.
  - Launch sites share several characteristics in their proximity and location.
- Models are trained to predict if the first stage will land depending on different independent variables

Thank you!