IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Aparna Srivastava
12-08-2024

Outline

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

Executive Summary

- Summary of methodologies
  1. Data Collection from SpaceX API and Wikipedia Scrap
  2. Data Wrangling
  3. EDA with Data Visualization and SQL
  4. Interactive Analysis with Folium
  5. Plotly DashBoard
  6. Predictive Analysis Classification
- Summary of all results
  1. Exploratory Data analysis
  2. Model Evaluation
  3. Discussions
  4. Conclusions

Introduction

- This project was developed to train data science skills as a capstone project for the IBM Data Science course at Coursera.
- The Problem consists of collecting, pre-process, cleaning and modeling analyse data from SpaceX company located on 2 sources. The first one, company website, contains technical information about the mission and, the second one, wikipedia contains history of the launches.
- The objective of the analysis is to help a new fictitious company SpaceY to catch up on this new space race.
- Methods like Exploratory data analysis with queries and graphics, interactive map plots, dashboards and machine learning models were used to better understand the relations between the success of a mission and the explanatory variables.

Section 1

Methodology

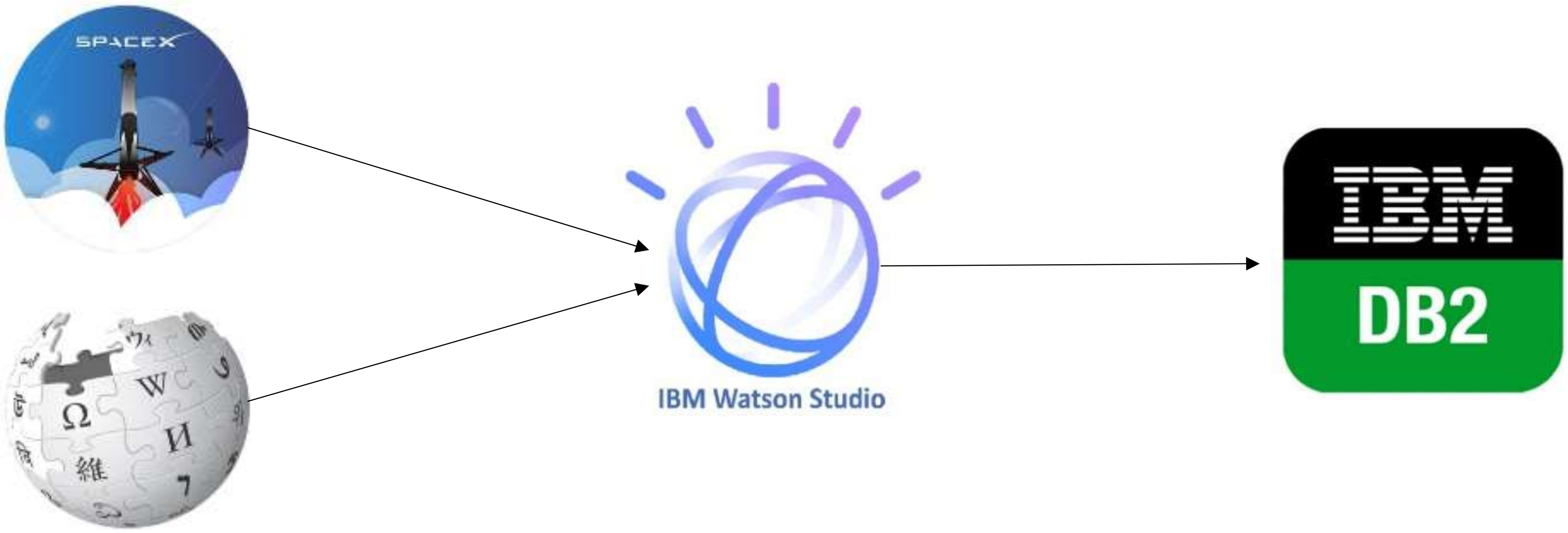
Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- 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

- Data Sources:
  1. [SpaceX API]
  2. [List of Falcon 9 and Falcon Heavy launches (Wikipedia)]
- Data Pipe Line Flow



Data Collection – SpaceX API

python™

- **Request and parse the SpaceX launch data using the GET request.**

```
RangeIndex: 94 entries, 0 to 93
Data columns (total 17 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   FlightNumber    94 non-null     int64  
 1   Date            94 non-null     object 
 2   BoosterVersion  94 non-null     object 
 3   PayloadMass     88 non-null     float64
 4   Orbit           94 non-null     object 
 5   LaunchSite      94 non-null     object 
 6   Outcome         94 non-null     object 
 7   Flights         94 non-null     int64  
 8   GridFins        94 non-null     bool   
 9   Reused          94 non-null     bool   
 10  Legs            94 non-null     bool   
 11  LandingPad      64 non-null     object 
 12  Block           90 non-null     float64
 13  ReusedCount     94 non-null     int64  
 14  Serial          94 non-null     object 
 15  Longitude       94 non-null     float64
 16  Latitude        94 non-null     float64
```

Requests
http for humans

GitHub Notebook

Data Collection - Scraping

python

- Request and parse **List of Falcon 9 and Falcon Heavy launches** using the GET request and retrive data tables using BeautifulSoap.

```
RangeIndex: 121 entries, 0 to 120
Data columns (total 11 columns):
 #   Column           Non-Null Count  Dtype 
---  ------           --------------  ----- 
 0   Flight No.       121 non-null    object
 1   Launch site      121 non-null    object
 2   Payload          121 non-null    object
 3   Payload mass     121 non-null    object
 4   Orbit            121 non-null    object
 5   Customer         120 non-null    object
 6   Launch outcome   121 non-null    object
 7   Version Booster  121 non-null    object
 8   Booster landing  121 non-null    object
 9   Date             121 non-null    object
 10  Time             121 non-null    object
```

Requests
http for humans

+ BeautifulSoup

[GitHub Notebook]

Data Wrangling

- **SpaceX**
  1. Filter the dataframe to only include `Falcon 9` launches.
  2. Dealing with Missing Values.
- **Wikipedia**
  1. Filter the dataframe to only include `Falcon 9` launches.
  2. Dealing with Missing Values.
  3. Inspect value counts from Orbit, Launch Site and Outcome columns.
  4. Define Binary Class column from Outcome categories.
  5. One hot encoding to binarize categorical variables.

[GitHub Notebook]

EDA with Data Visualization

Plots:

- Category plot of Payload Mass by Flight number with Class on color dimension.
- Category plot of Launch Site by Flight number with Class on color dimension.
- Scatter plot of Launch Site by Payload Mass with Class on color dimension.
- Barchart of Succes rate by Orbit.
- Scatter plot of Orbit by Flight Number with Class on color dimension.
- Scatter plot of Orbit by Payload Mass with Class on color dimension.
- Line plot of Success by Year.

GitHub Notebook

EDA with SQL

SQL Queries

- Unique Launch Sites.
- Launch Sites biginning with CAA.
- Total Payload Mass carried by booster lauched from NASA CRS.
- Average Payload Mass caried by F9 v1.1 boosters.
- Date of the first success landing outcome in ground pads.
- *Boosters success in drone ship with payload between 4000 and 6000.*
- *Total number of successful and failure mission outcomes.*
- *Booster_versions which have carried the maximum payload mass.*
- *Failed landing_outcomes in drone ship in 2015.*
- *Descending Landing outcomes between 2010-06-04 and 2017-03-20.*

GitHub Notebook

Interactive Map with Folium

python™

Map elements

- Blue circle on NASA location with name and popup.
- Red circles for launch sites with names and popup.
- Marker cluster for success and faillures on launchs with red and green.
- Distance from nearest Railroad with blue connect line .
- Distance from nearest city Titusville with blu connect line.

[GitHub Notebook]

Dashboard with Plotly Dash

Dashboard

- Interactive Pychart to show parts of a whole to compare number of lauch attemps in all dites or parts of a whole for each lauch site selected on the dropdown menu to compare number success and faillures.
- Interactive Scatterplot to compare the relation of Class by Paylod Mass (KG) allowing to change both launch sites and range of payload mass.

GitHub Python Script

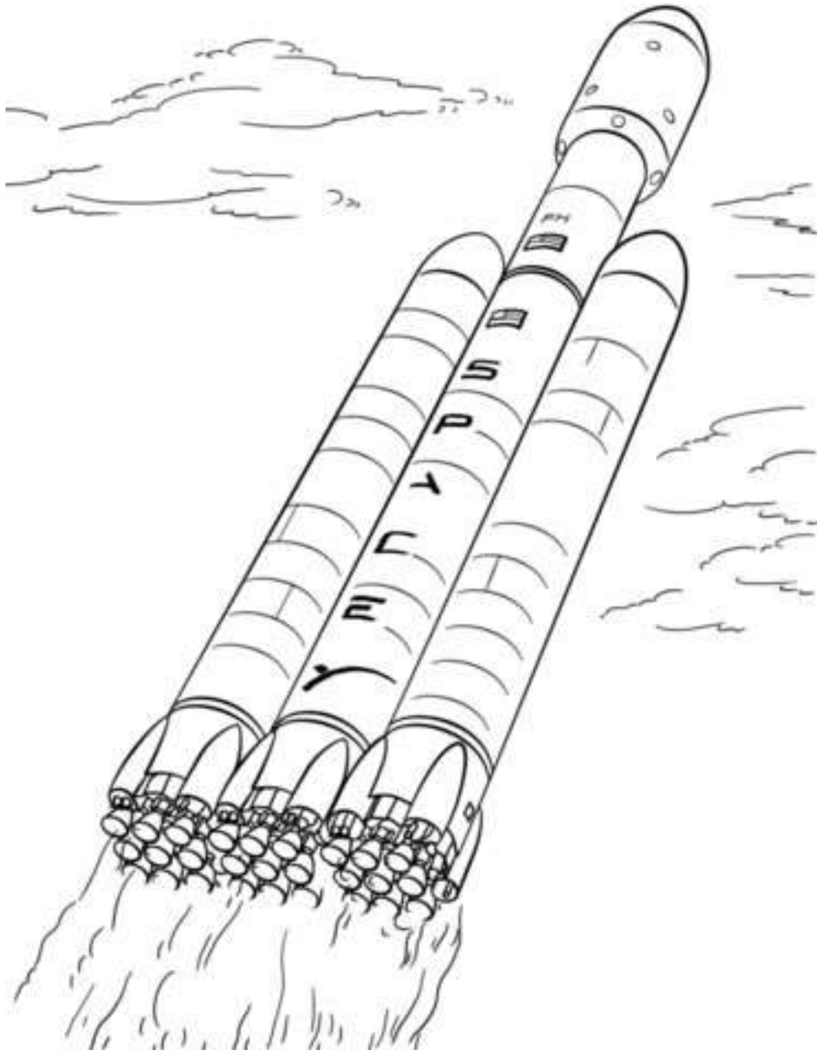
Predictive Analysis (Classification)

- Modeling
  - Load X and Y data from IBM SkillsNetwork in .csv flat files format.
  - Standardize X features with normalization.
  - Split data in train and test samples 80/20
  - Cross vadidation for Logistic Regression score to find best parameters.
  - Cross vadidation for SVM score to find best parameters.
  - Cross vadidation for Decision Tree score to find best parameters.
  - Cross vadidation for KNN score to find best parameters.
  - Plot confusion matrices to visualize positive and negative predictions.

GitHub Python Notebook

Results

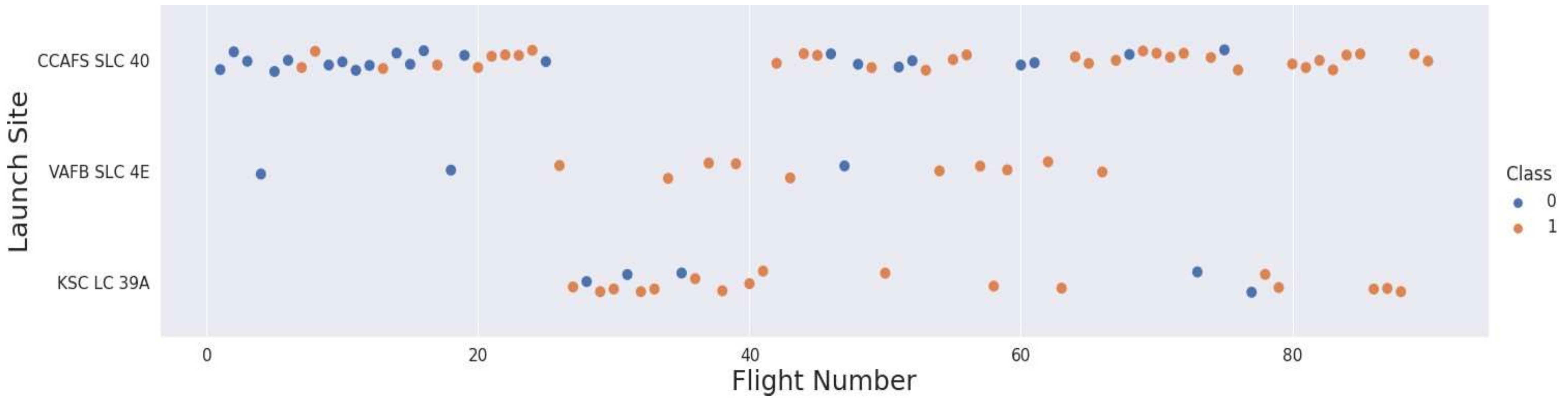
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results



Section 2

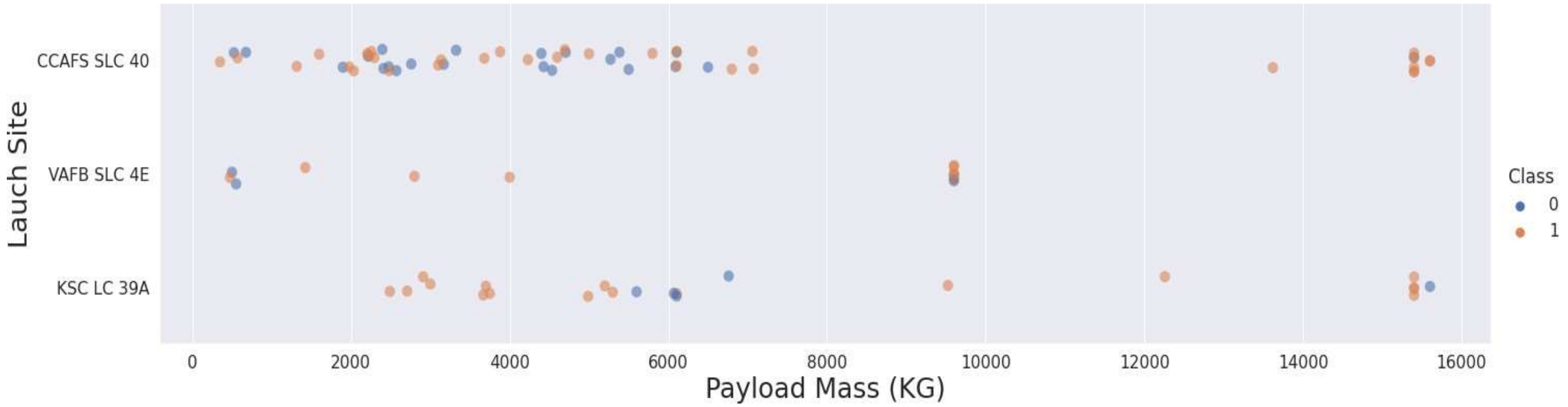
Insights drawn from EDA

Flight Number vs. Launch Site



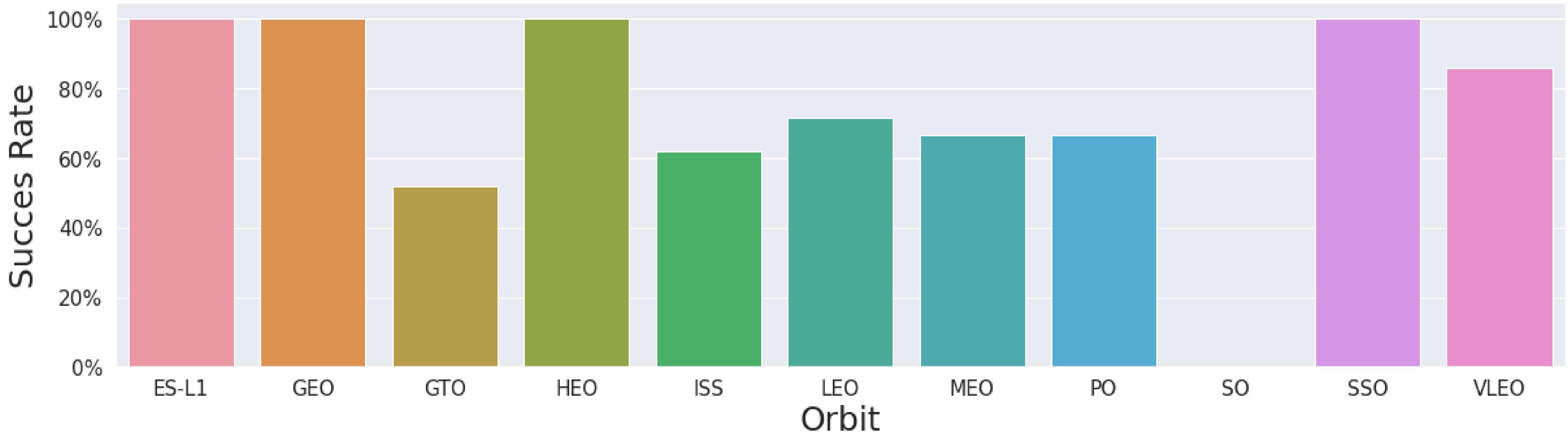
- Most missions occurred at CCAFS SLC 40.
- Most landing faillures occurred before Flight Number 20.
- After Flight Number 80 all landings outcomes were success.

Payload vs. Launch Site



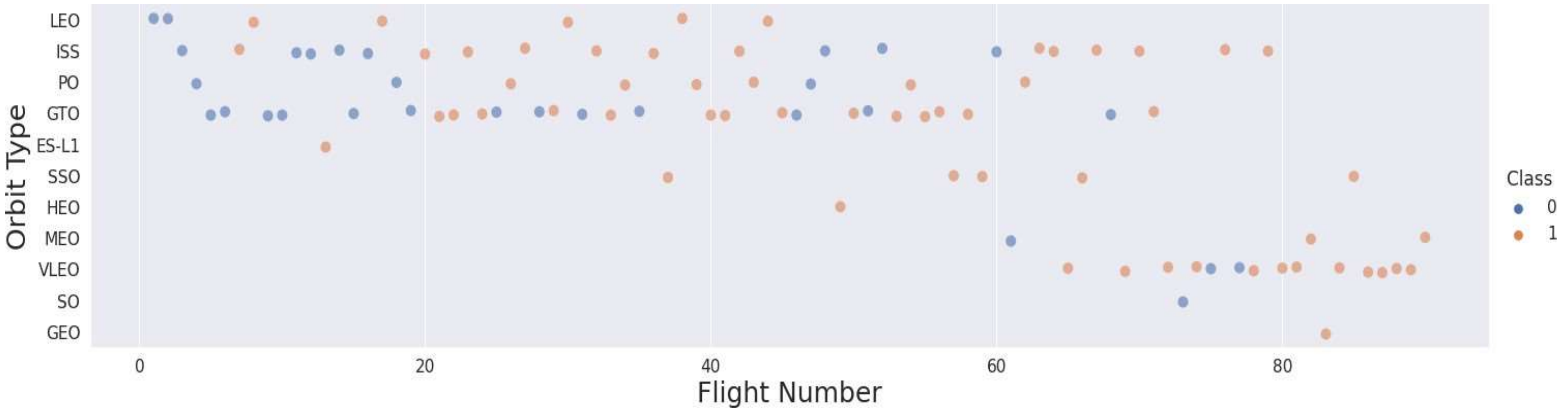
- Most payload were below 8000 kg.
- Most faillures occurred with payload between 2500 and 7500 kg.
- After Flight Number 80 all landings outcomes were success.

Success Rate vs. Orbit Type



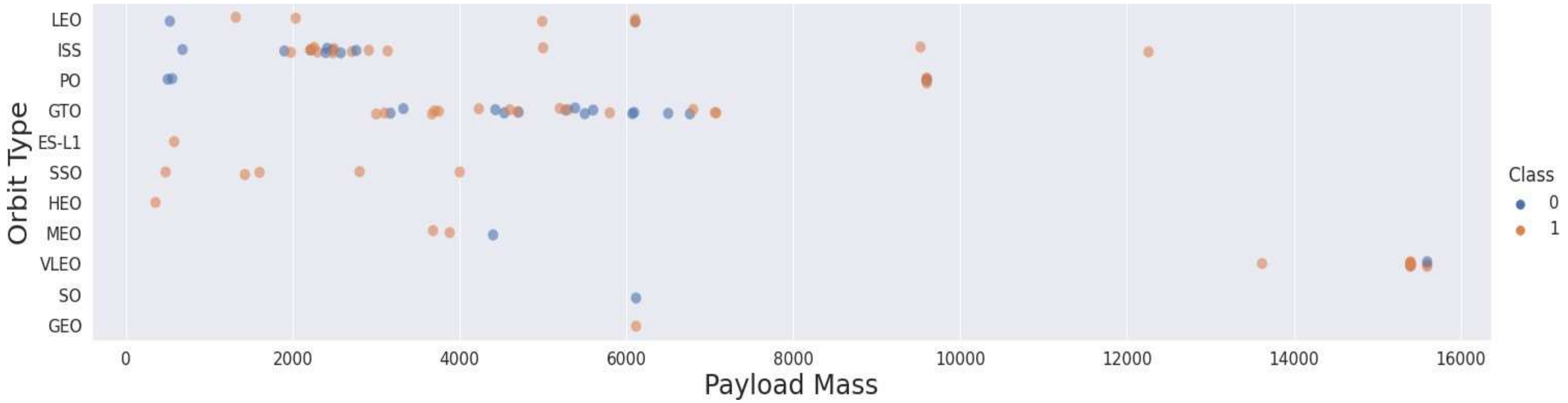
- Success Rate is 100% at ES-L1, GEO, HEO and SSO.
- Succes Rate betwwen 40% and 80% at GTO, ISS, LEO, MEO, PO and VLEO.
- Just at SO orbit the success rate is 0%.

Flight Number vs. Orbit Type



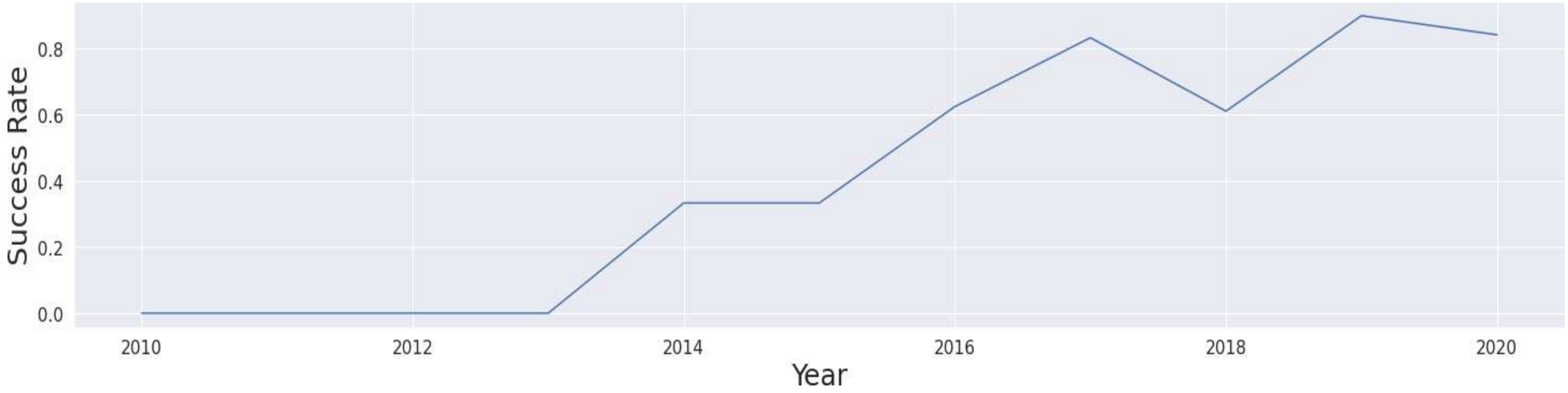
- From flight 0 to 60 just LEO, ISS, PO, GTO, ES-L1 and SSO orbits were contemplated.
- From flight 60 to today HEO, MEO, VLEO, SO and GEO orbits begin to be contemplated.
- It seens the after acomplished the first group of orbits, SpaceX focus on the newest ones to achieve the same level of success.

Payload vs. Orbit Type



- Heavy payload mass over 8000 kg seens to be concentrated on ISS, PO and VLEO orbits.
- Most faillures concentrate on low 1500 to 3000 kg and medium, 2500 and 7500 kg payload mass on ISS and GTO orbits.

Launch Success Yearly Trend



- From years 2010 to 2013 the faillures predominate.
- From years 2013 to 2017 occurred 2 majors improvements on succes rate.
- From 2017 to 2018 there was a fall on succes rate recovered from 2018 to 2019.

All Launch Site Names

- Cape Canaveral Space Launch Complex 40 (CCAFS LC-40).
- Vandenberg Space Launch Complex 4 (VAFB SLC-4E).
- Kennedy Space Center Launch Complex 39 (KSC LC-39A).
- Cape Canaveral Space Launch Complex 40 (CCAFS SLC-40).

Launch Site Names Begin with 'CCA'

	Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit
0	04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO
1	08-12-2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0	LEO (ISS)
2	22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)
3	08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)
4	01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)
6	03-12-2013	22:41:00	F9 v1.1	CCAFS LC-40	SES-8	3170	GTO
7	06-01-2014	22:06:00	F9 v1.1	CCAFS LC-40	Thaicom 6	3325	GTO
8	18-04-2014	19:25:00	F9 v1.1	CCAFS LC-40	SpaceX CRS-3	2296	LEO (ISS)
9	14-07-2014	15:15:00	F9 v1.1	CCAFS LC-40	OG2 Mission 1 6 Orbcomm-OG2 satellites	1316	LEO

Only 10 rows were displayed from 100 Launch Sites that the name starts with CCA.

Total Payload Mass

- Total Payload by booster from NASA is `45596 Kg.`

Average Payload Mass by F9 v1.1

The average Payload Mass of F9 v1.1 is 2928.4 Kg

First Successful Ground Landing Date

- The first Succesful Ground Landing Date is at `01-05-2017`.

Successful Drone Ship Landing with Payload between 4000 and 6000

Boosters:

23	F9 FT B1022
27	F9 FT B1026
31	F9 FT B1021.2
42	F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

Success	98
Success	1
Success (payload status unclear)	1
Failure (in flight)	1

Boosters Carried Maximum Payload

- Booster Version that carried the maximum payload mass is `F9 B5 B1048.4.`

2015 Launch Records DroneShip Failures

Date	Booster_Version	Launch_Site
2015-10-01	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	F9 v1.1 B1015	CCAFS LC-40

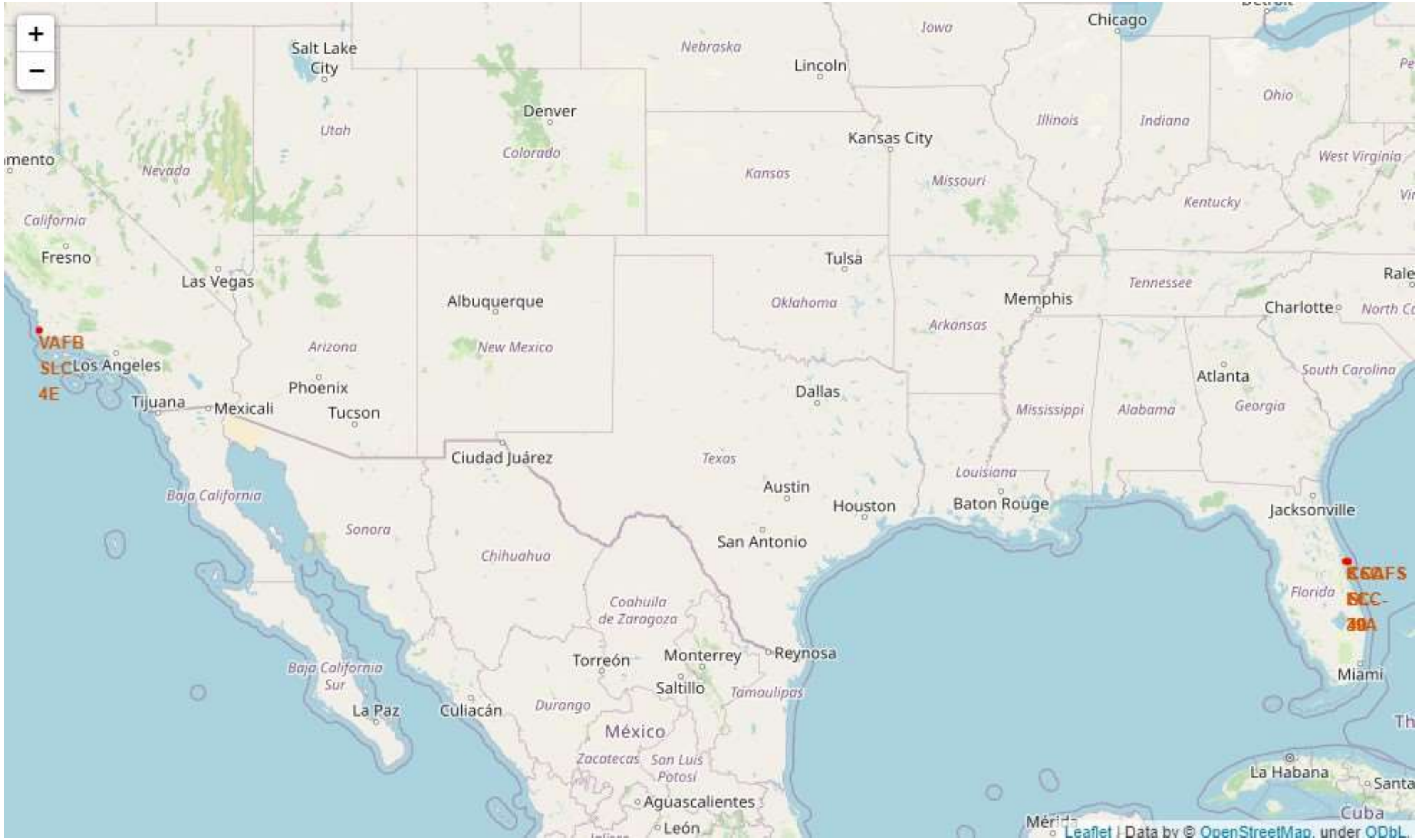
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	1
Precluded (drone ship)	1

Section 4

Launch Sites Proximities Analysis

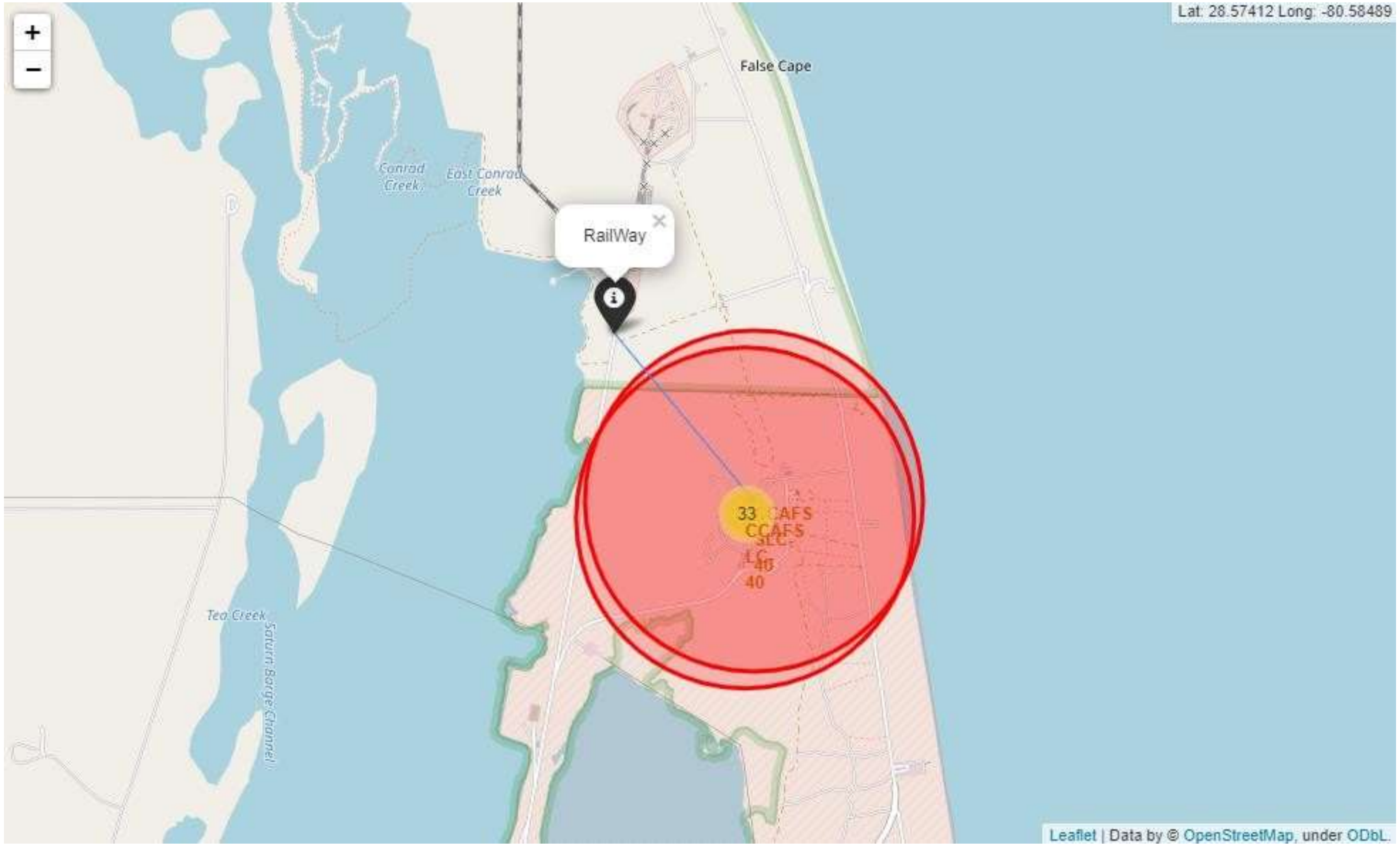
Launch Sites SpaceX Map



- The locations are on both coats os US.

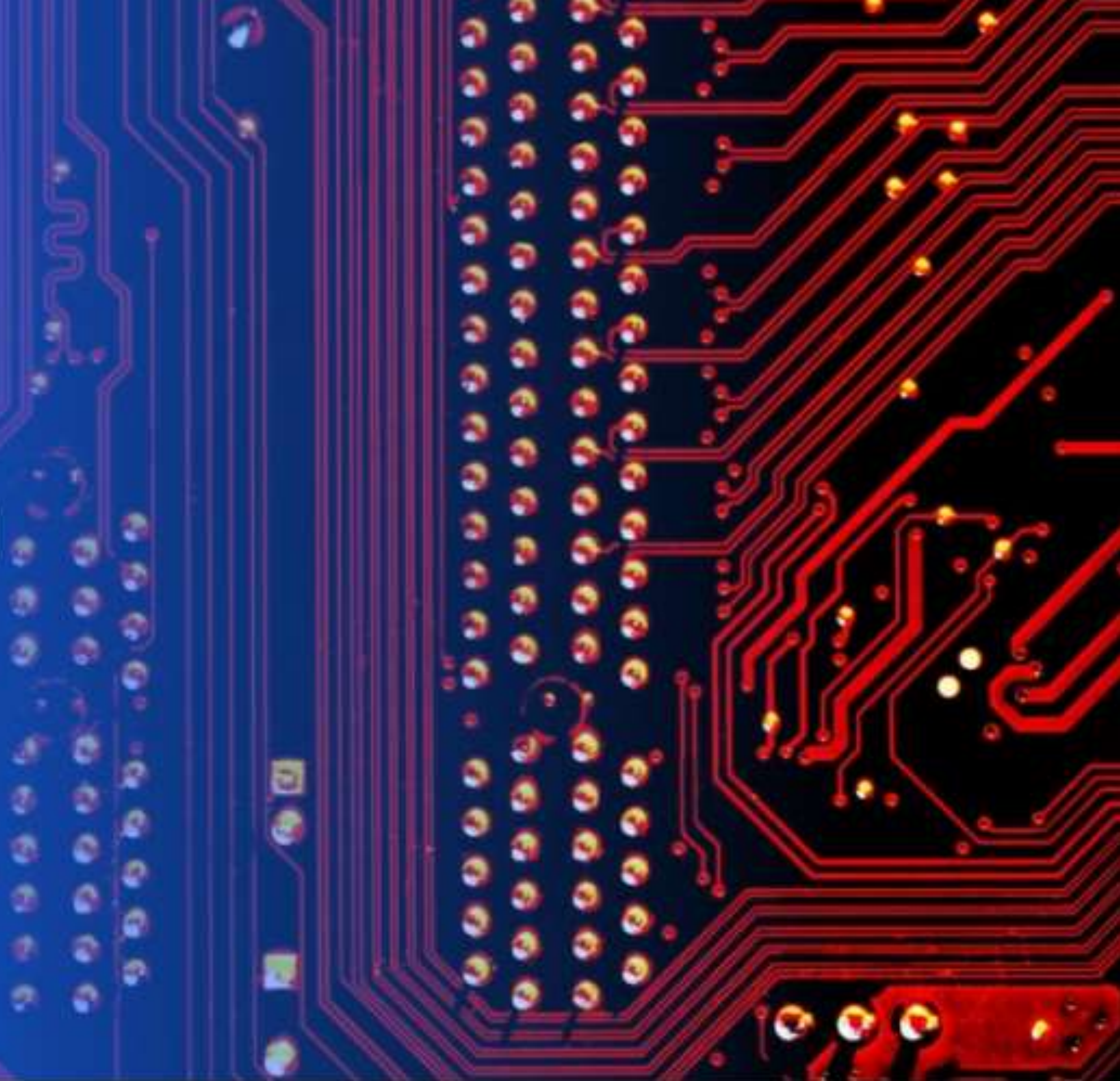
Launch Sites Succes and Failures Map

Launch Site Proximity to Railway

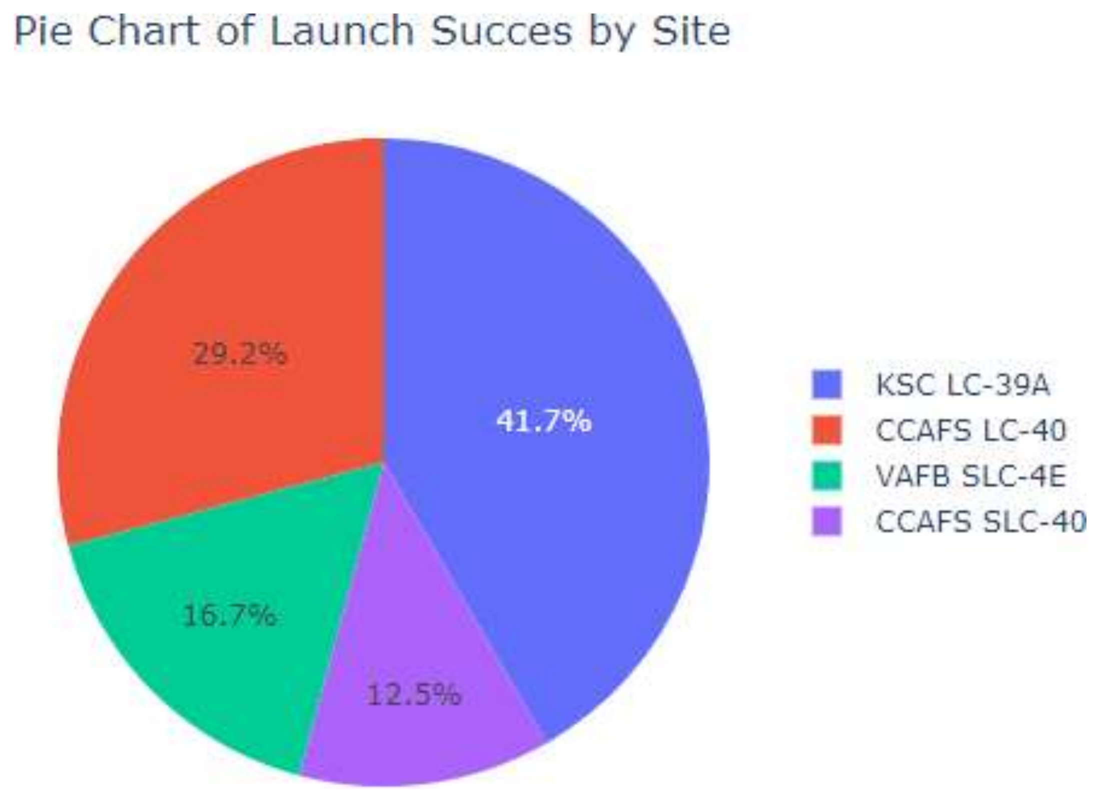


Section 5

Build a Dashboard with Plotly Dash

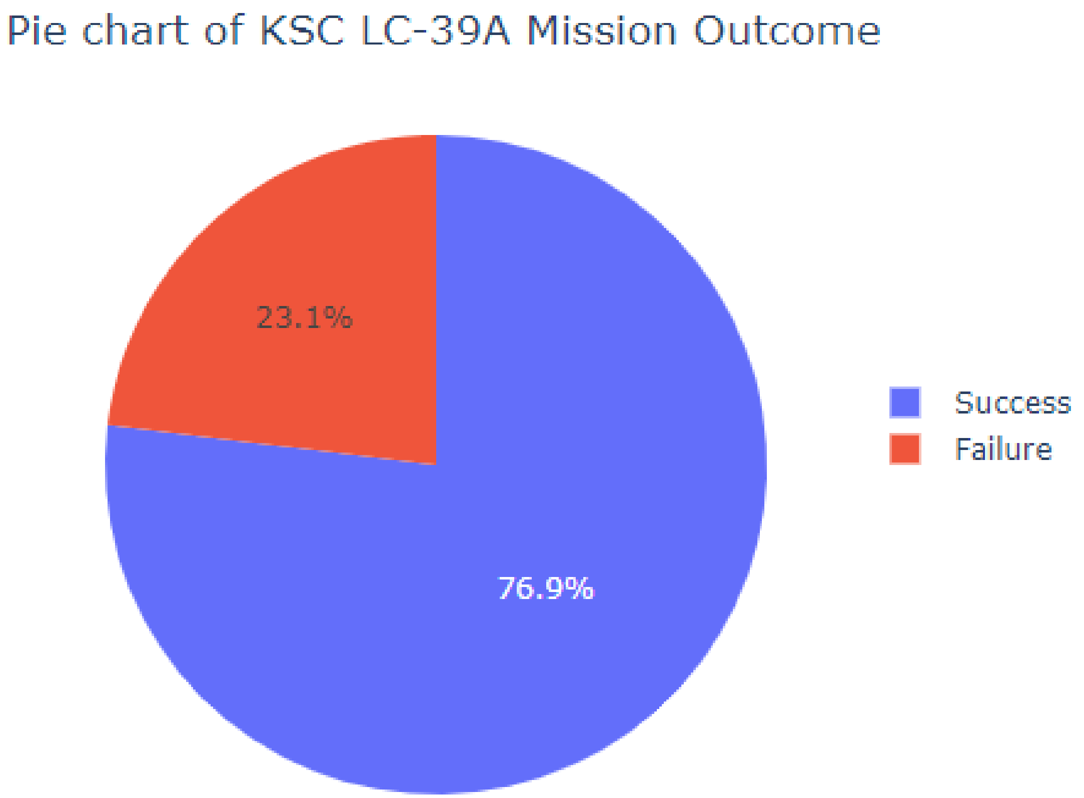


Launch Success by Site



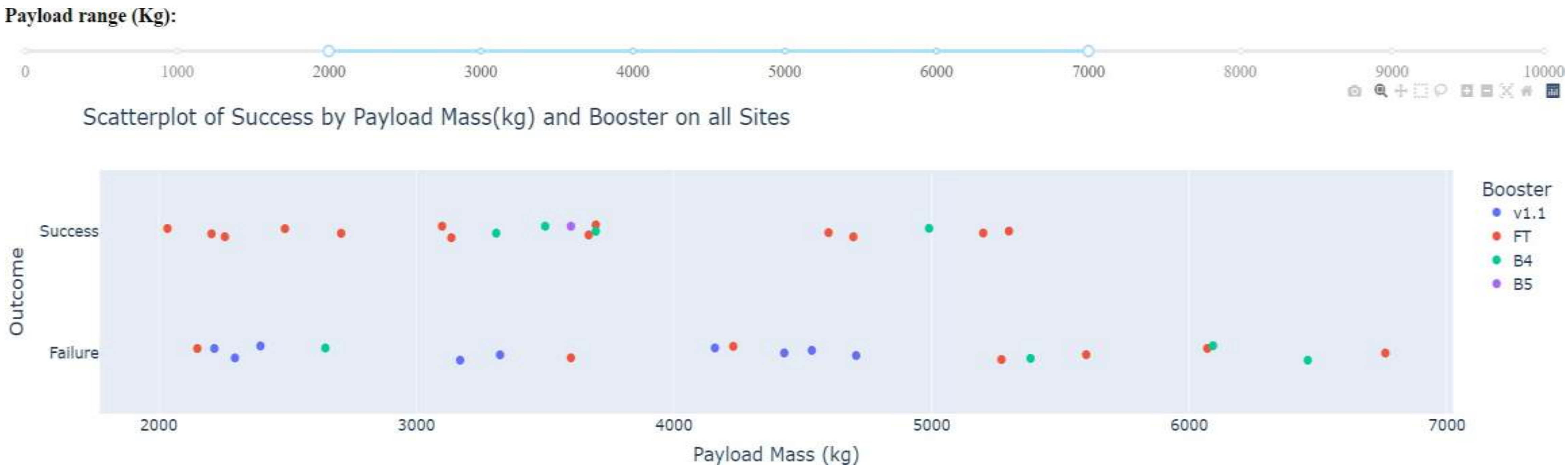
- CCAFS LC-40 and KSC LC-39A sum up 71.9% of the success launchs.

Launch Site with the Highest Success Ratio



- KSC LC-39A performed best with 76.9% of success rate.

Launch Outcomes by Payload Mass and Boosters

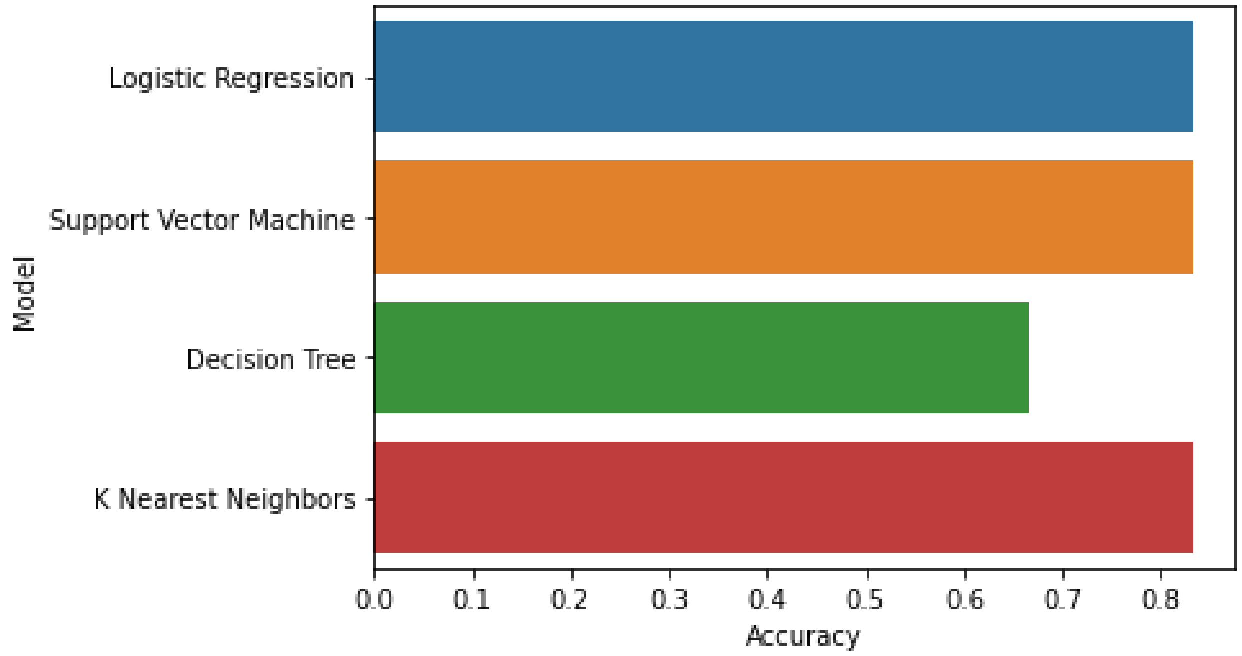


- From 2000 to 7000 kg of payload mass for all launch sites the FT Booster achieved more success then the other boosters.

Section 6

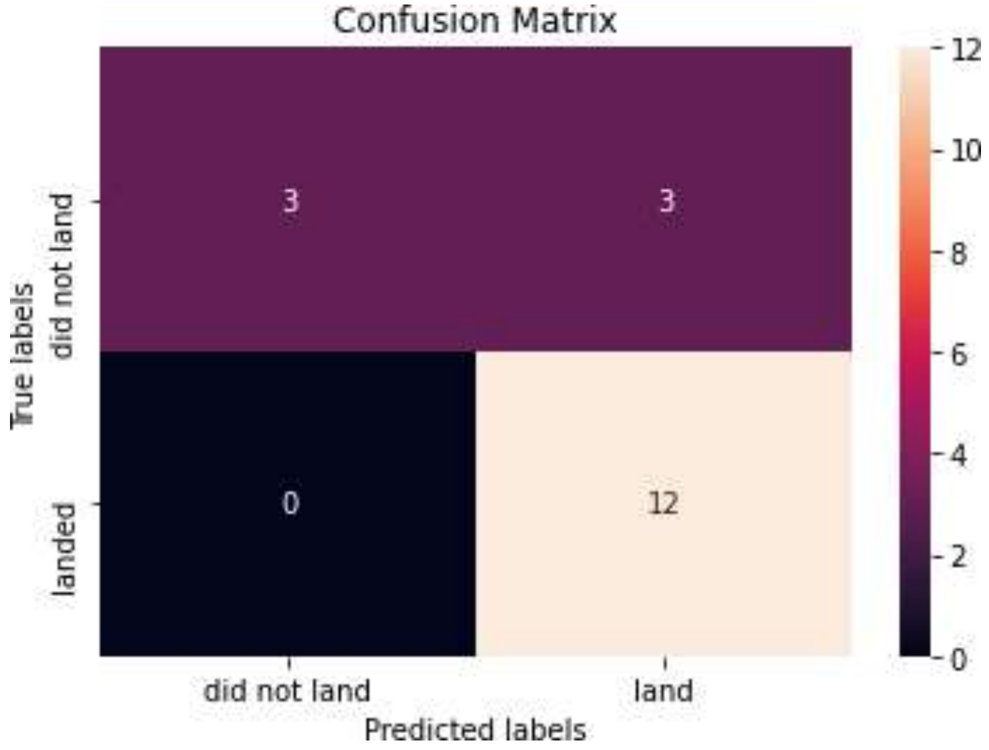
Predictive Analysis (Classification)

Classification Accuracy



- Three methods performed equally with 3 false positives or 3 events were first booster not landed successfully but the model predict in fact it did.Logistic Regression, Support Vector MachineKNN.

Confusion Matrix Logistic Regression



- Logistic Regression performed well on predicting landed and performs randonly on predicting not landed missions with 50% correct predictions.

Conclusions

- The succes rate gets better after 2013 and impproved even more in 2015.
- The best succes rate occurs in orbits ES-L1, GEO, HEO and SSO.
- He best payload mass is between 200 and 7000 kg
- The best succes launch site is CCAFS SLC-40
- The coefficients that impact most positive on the outcomes are, Block, ReusedCount, LandingPad_5e9e3032383ecb267a34e7c7, LandingPad_5e9e3032383ecb6bb234e7ca, GridFins_True, Legs_True.
- The coefficients that impact most positive on the outcomes are Serial_B1017, Serial_B1018, Serial_B1020, Serial_B1028, Serial_B1044, Serial_B1050, GridFins_False, Legs_False.

Appendix

- GitHub Notebook Data Collection Spacex API
- GitHub Notebook Data Collection Web Scraping WikiPedia
- GitHub Notebook Data Wrangling
- GitHub Notebook EDA
- GitHub Notebook Interactive Map
- GitHub Python Script DashBoard
- GitHub Python Notebook Predict Modeling

Thank you!