IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Srikrishnaa Jayaraman
05/01/2022

Outline

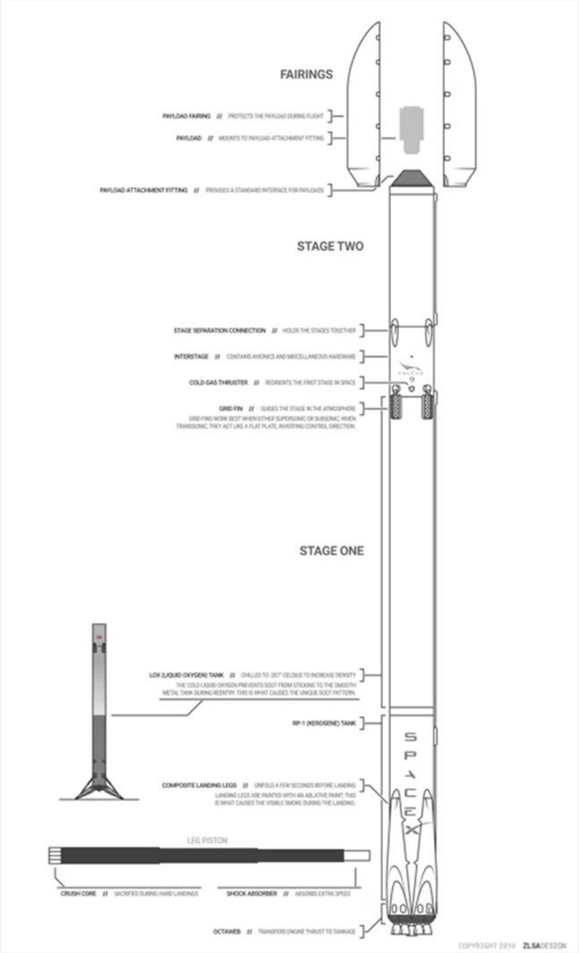
- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

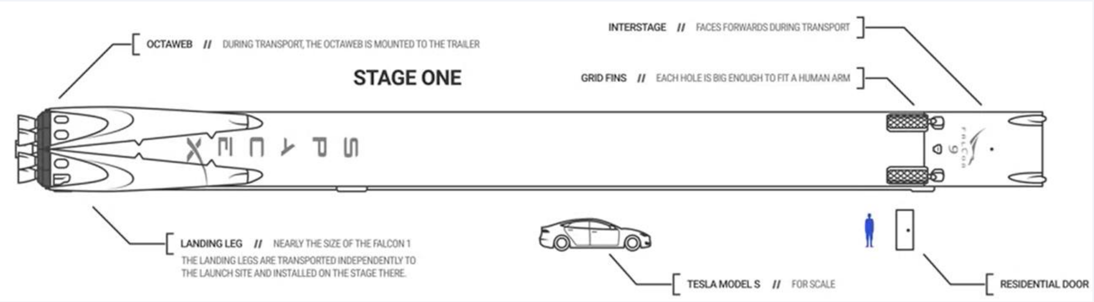
Executive Summary

- SpaceX's Falcon 9 rocket boasts a much cheaper price (at about 1/3rd the cost of other providers) for launch compared to other rockets due to the reusability of its first stage after every mission
- In this project, the likelihood of recovering the first stage is predicted (using data science and machine learning techniques) before launch – which helps determine SpaceX's Falcon 9 launch price.
- From the data available on SpaceX's API and on the internet, we were able to successfully predict the first stage recovery with an accuracy score of 83.33%

Introduction

- SpaceX's Falcon 9 rocket boasts a much cheaper price (at about 1/3rd the cost of other providers) for launch compared to other rockets due to the reusability of its first stage after every mission
- For a competing startup like SpaceY that is new to the space race, it is useful to know the price bid of SpaceX's launch so that SpaceY can undercut them and grab SpaceX's potential customers
- The cost of launch can be determined once we know the likelihood of first stage recovery





Introduction

- Question statement: Is it possible to predict the likelihood of Falcon 9's first stage recovery from previous attempts over the years?
- Technique:
  - Extract Falcon 9 launch data from various sources on the internet
  - Analyze the data for patterns
  - Train a classification machine learning model to predict the fate of the first stage
  - Test the model with existing data
  - Deploy the model for use at SpaceY

Section 1

Methodology

Methodology

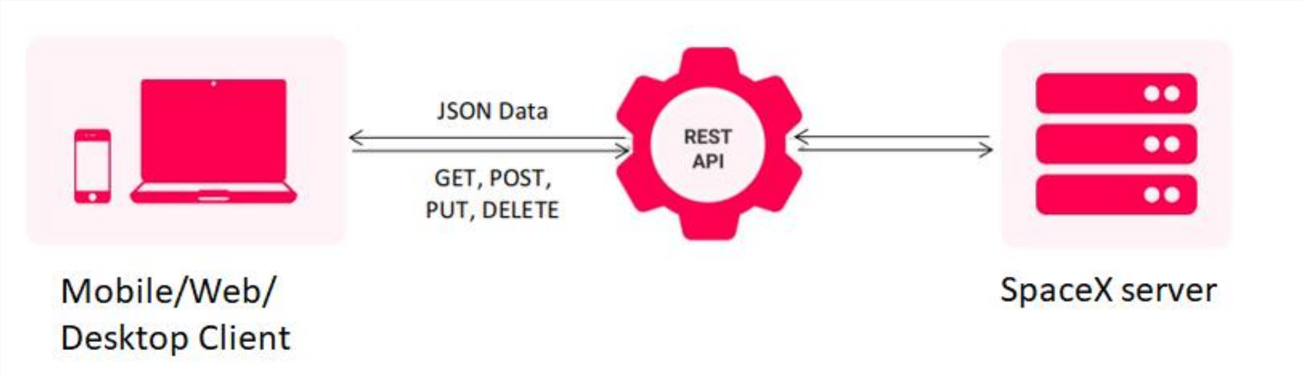
Executive Summary

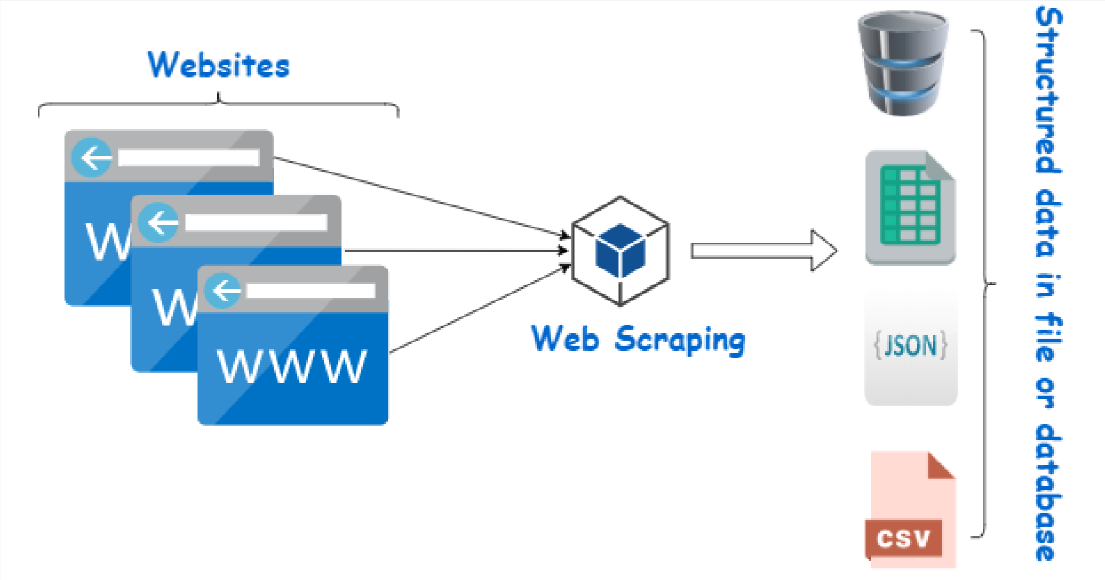
- Data collection methodology:
  - SpaceX REST API and web scraping List of SpaceX Falcon 9 launches wiki article
- Perform data wrangling
  - Assigning appropriate data types and dealing with NaN values appropriately
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Built several classification models and evaluated the accuracy of each to determine the best model

Data Collection

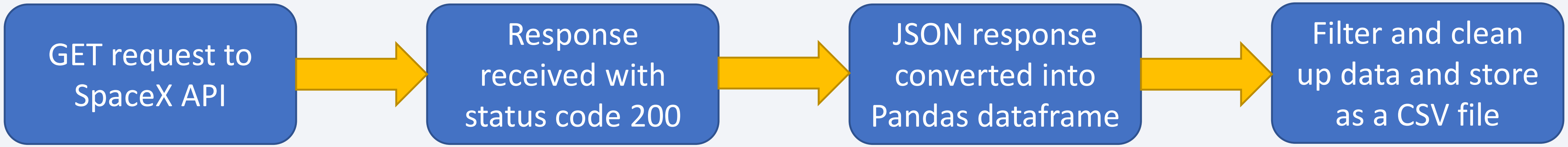
REST API: "Client" requests data from a "Server" through HTTP requests and the "Server" responds with the data in the JSON format via a HTTP response

Web scraping: method to automatically obtain large amounts of (mostly) unstructured data from websites and convert them into structured data for further processing



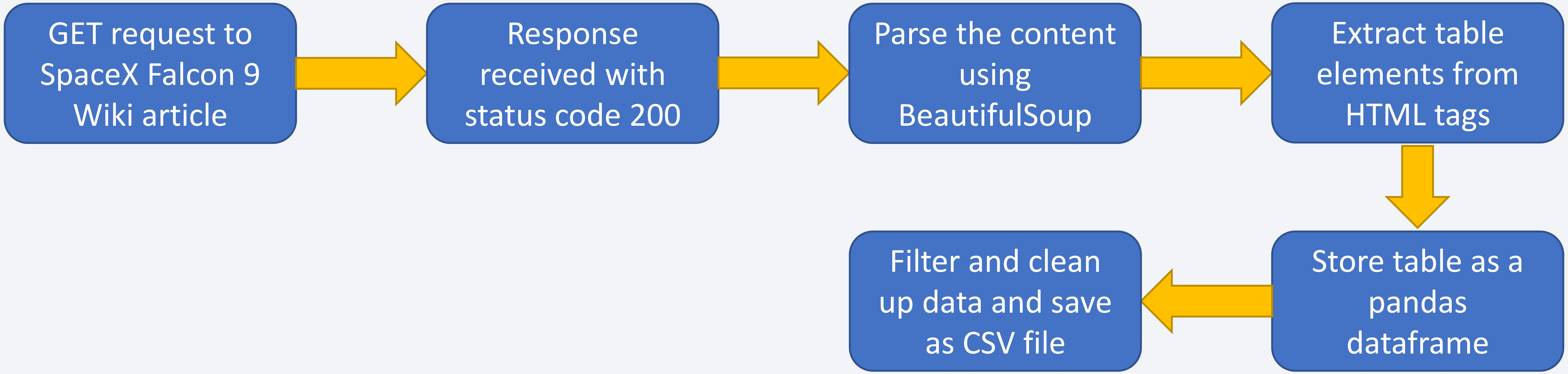


Data Collection – SpaceX API



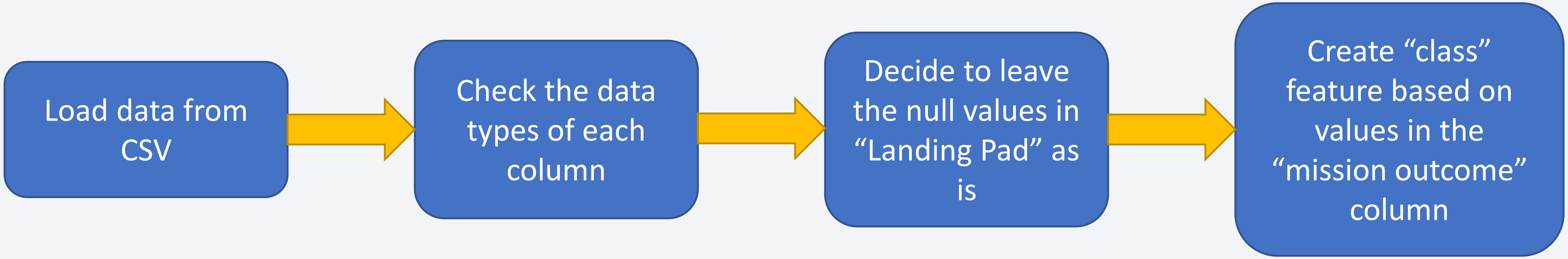
[API notebook]

Data Collection - Scraping



[Webscraping notebook]

Data Wrangling



Wrangling notebook

EDA with Data Visualization

Some of the charts that were plotted:

- Landing success rate versus launch site – to see if any particular site was significantly more successful or less successful
- Landing success rate versus Orbit type – to see if orbit type influenced the landing success rate
- Yearly trend of launch success rate – to see how the average success rate has varied over the years
- And more…

EDA Data Viz notebook

EDA with SQL

Some of the queries:

- Date of first successful landing
- Booster versions with successful drone ship landing having payload mass between 4 and 6 tonnes
- Total number of successful and unsuccessful mission outcomes
- Ranking the count of landing outcomes in a given date range
- And more…

EDA - Data Querying

Build an Interactive Map with Folium

Some of the markers on the map:

- Circular markers were added to indicate launch sites
- Red and green marker clusters were added to indicate successful and failed landings at each launch site
- Line markers were added to indicate distance of launch site from coastline, nearest city, highway, and railway
- Text markers indicating the distance of site from said entities were added as well

[Interactive Maps with Folium]

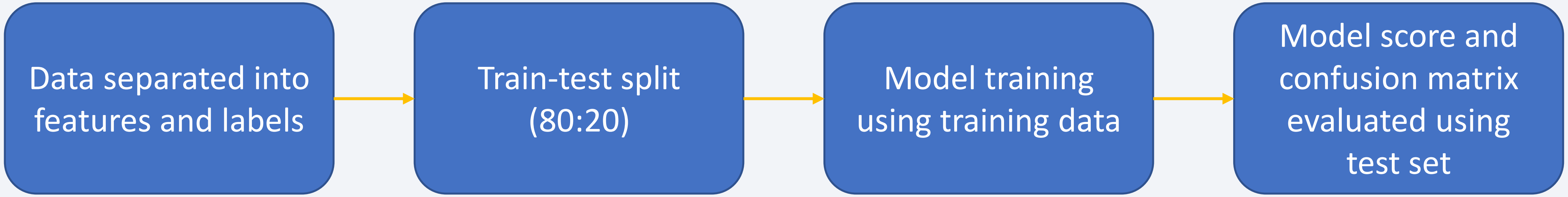
Build a Dashboard with Plotly Dash

Some of the charts and interactions on the dashboard:

- User input dropdown menu for launch site – user can select the launch site for which they would like to see the data
- Pie chart showing success rate at the user chosen site. If all sites were picked, then the percentage of total successes at each site is shown as a pie chart – to quickly visualize success rates at each site
- User input payload mass range slider – user can change the range of payload mass for which they can see the plot of success class versus payload mass graph (for the user chosen launch site), color coded by booster version – to see if class can be separated at a critical payload mass

Plotly Interactive Dashboard

Predictive Analysis (Classification)

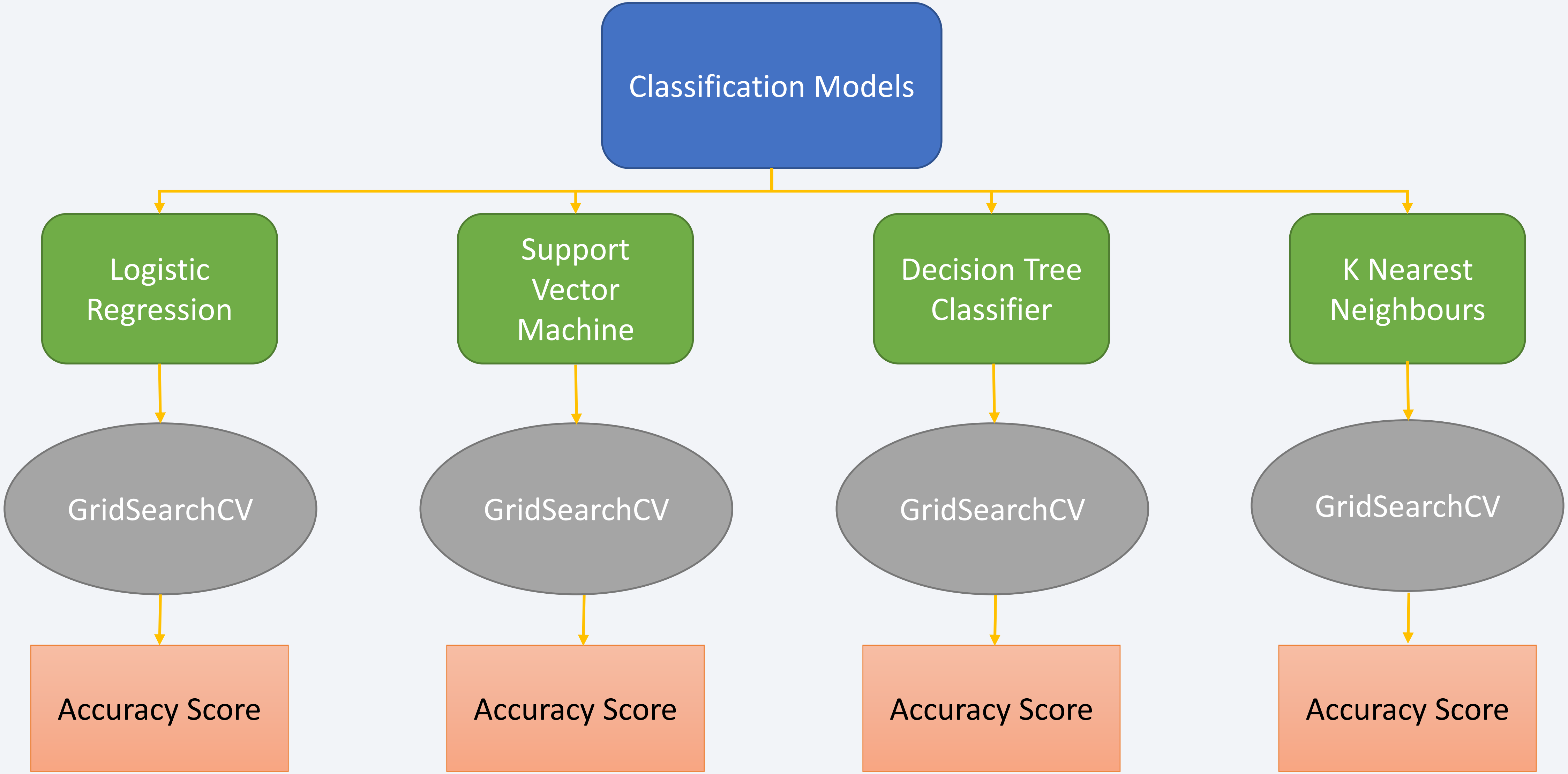


This process is repeated for all the models

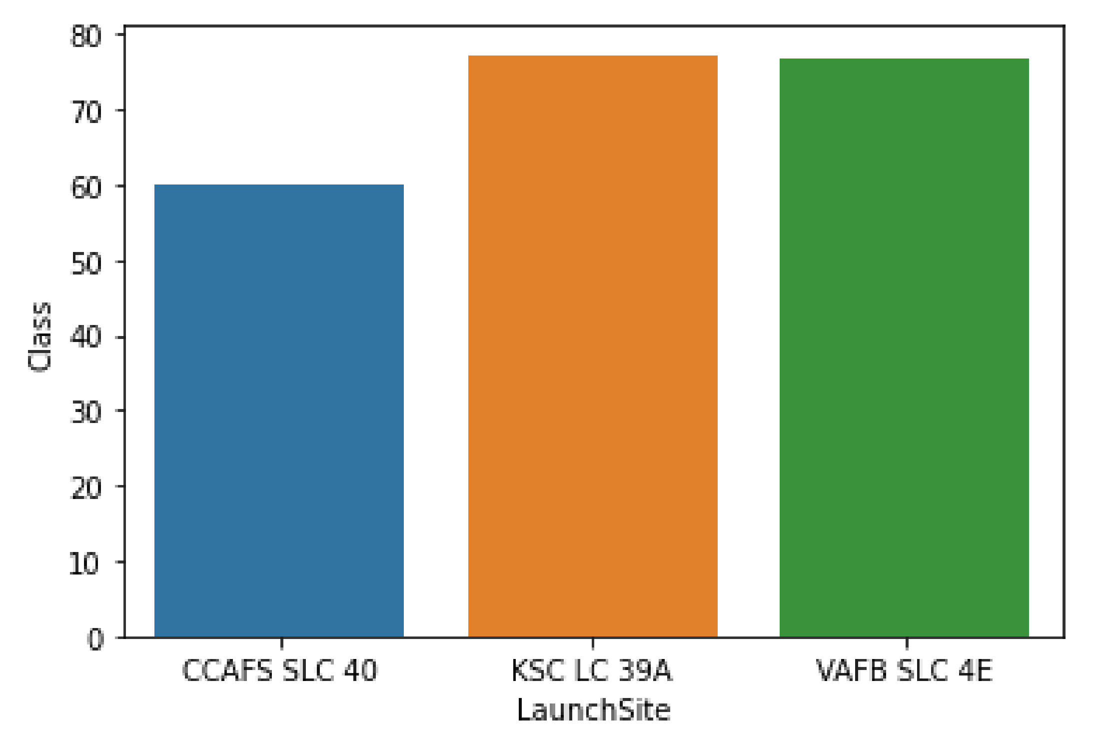
The best model is the one that has the best accuracy score with the test set

[Machine Learning - Classification]

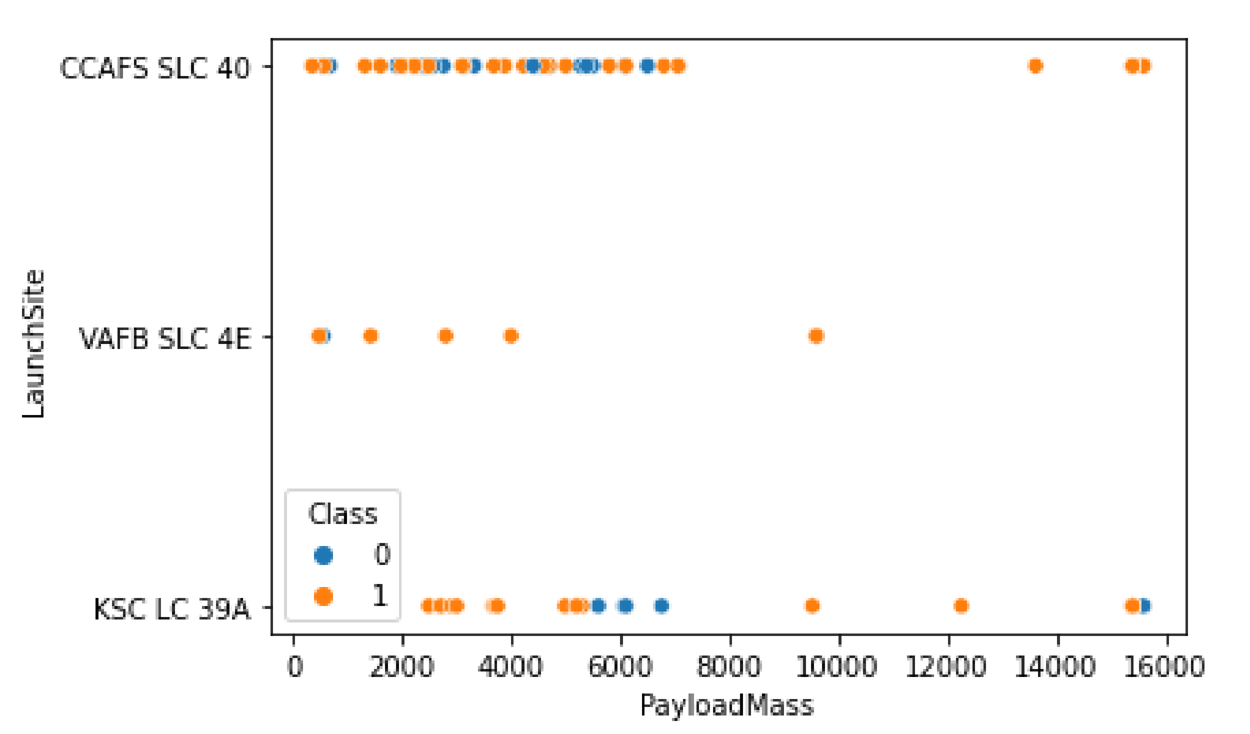
Model Training Process



Results - EDA

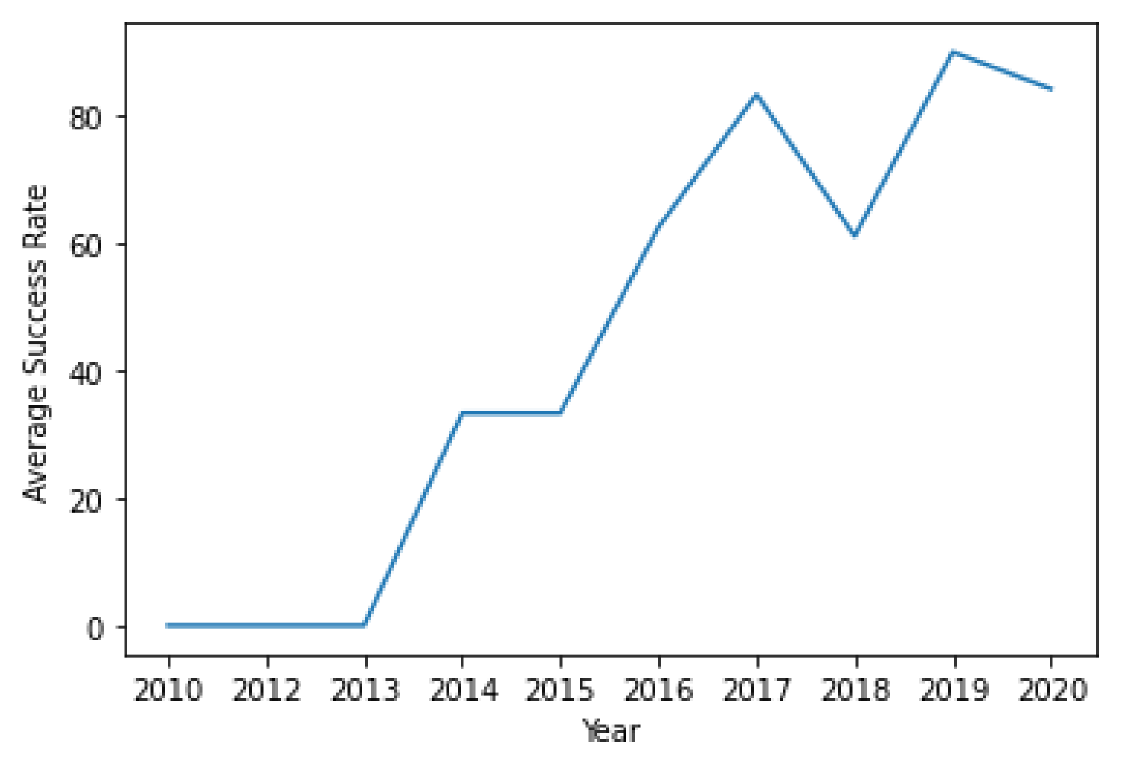


80
70
60
50
40
30
20
10
0
Class
CCAFS SLC 40
KSC LC 39A
VAFB SLC 4E
LaunchSite

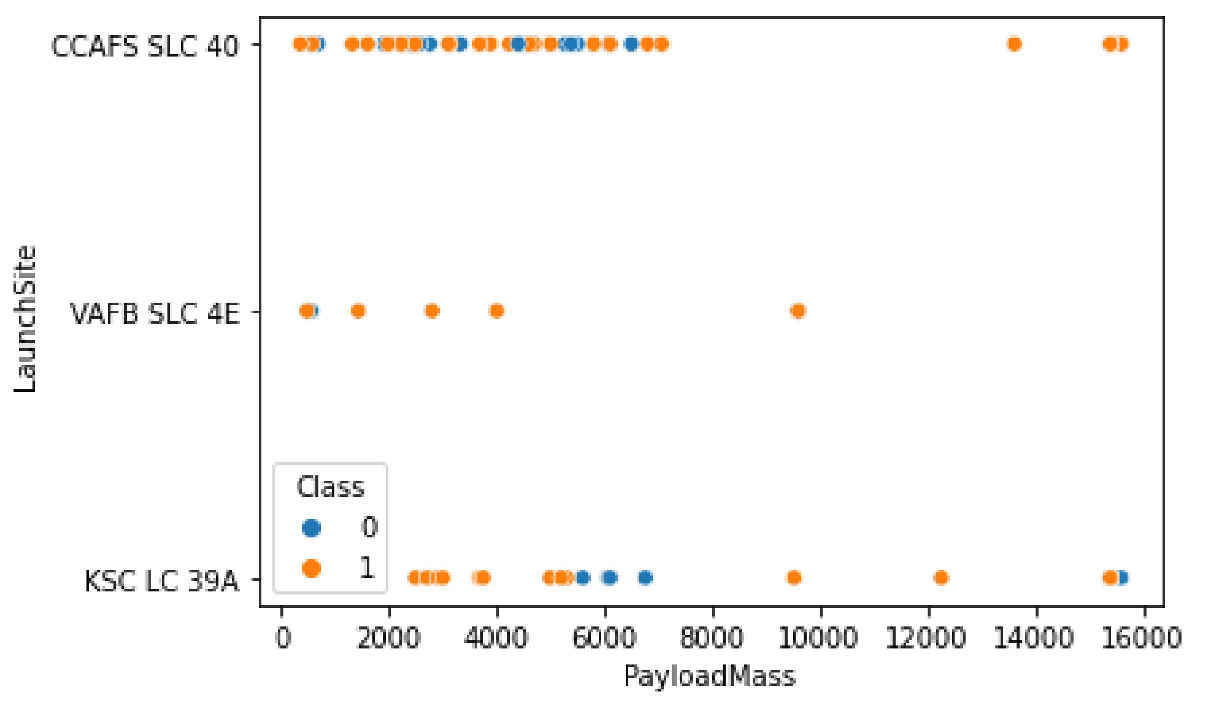


CCAFS SLC 40
VAFB SLC 4E
KSC LC 39A
LaunchSite
Class
0
1
0
2000
4000
6000
8000
10000
12000
14000
16000
PayloadMass

Results - EDA

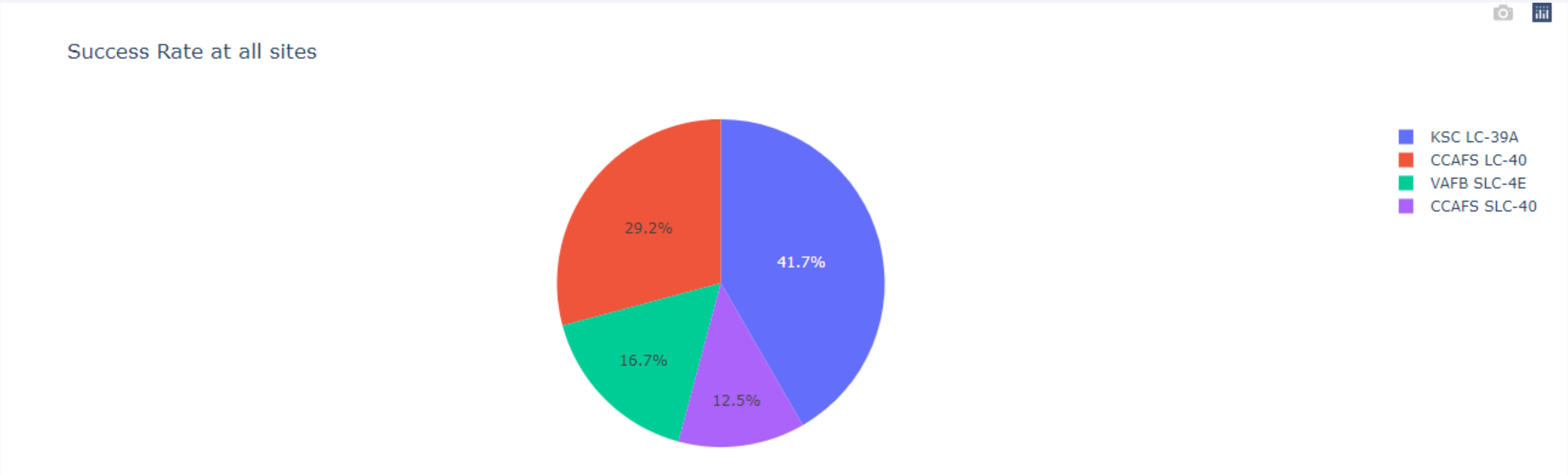


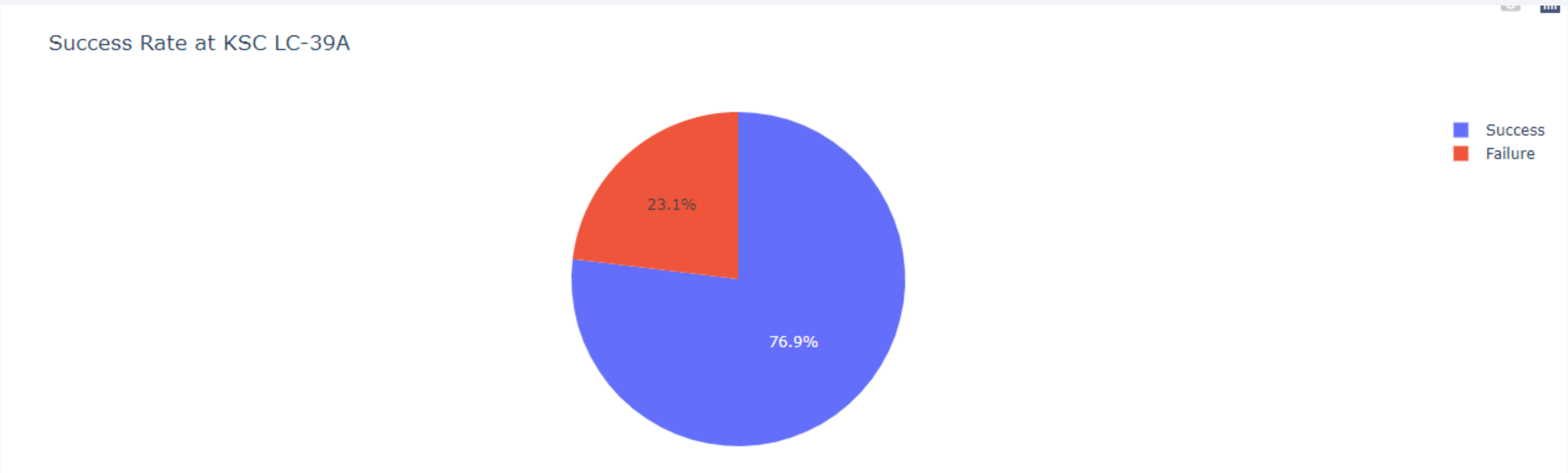
Average Success Rate
80
60
40
20
0
2010
2012
2013
2014
2015
2016
2017
2018
2019
2020
Year



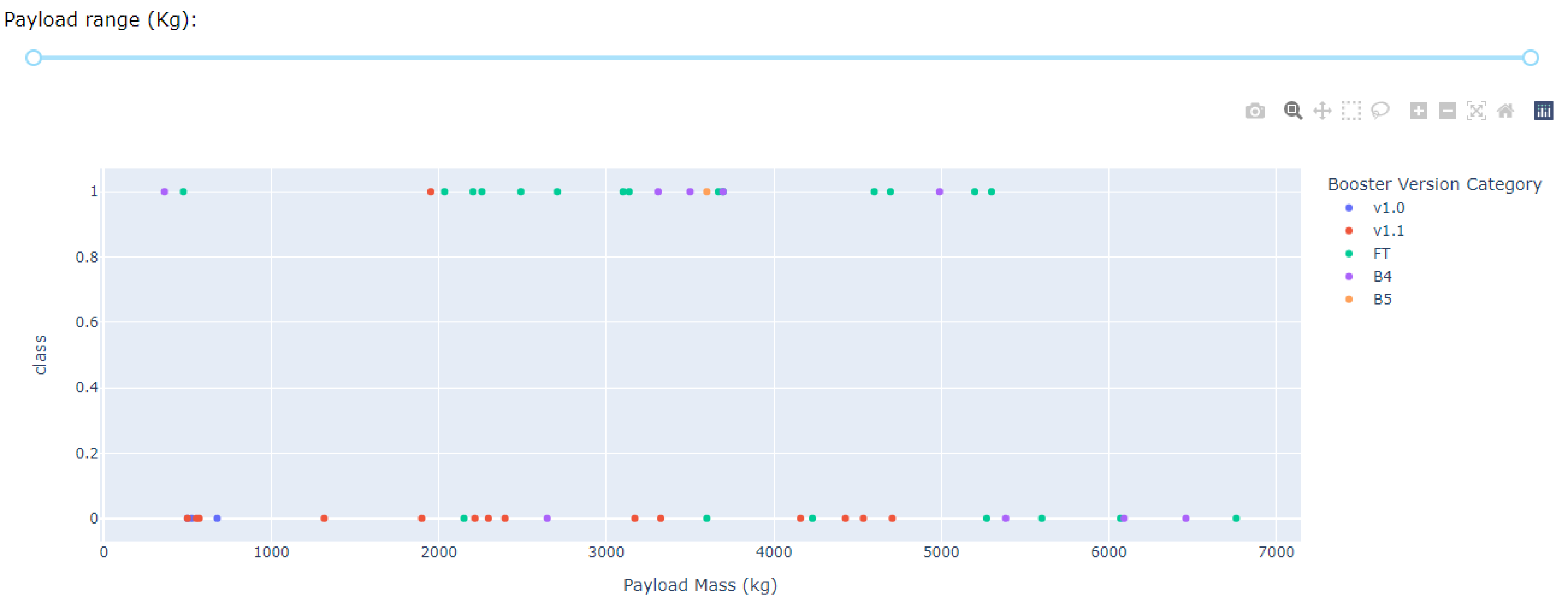
LaunchSite
CCAFS SLC 40
VAFB SLC 4E
KSC LC 39A
Class
0
1
0
2000
4000
6000
8000
10000
12000
14000
16000
PayloadMass

Results - Dashboard

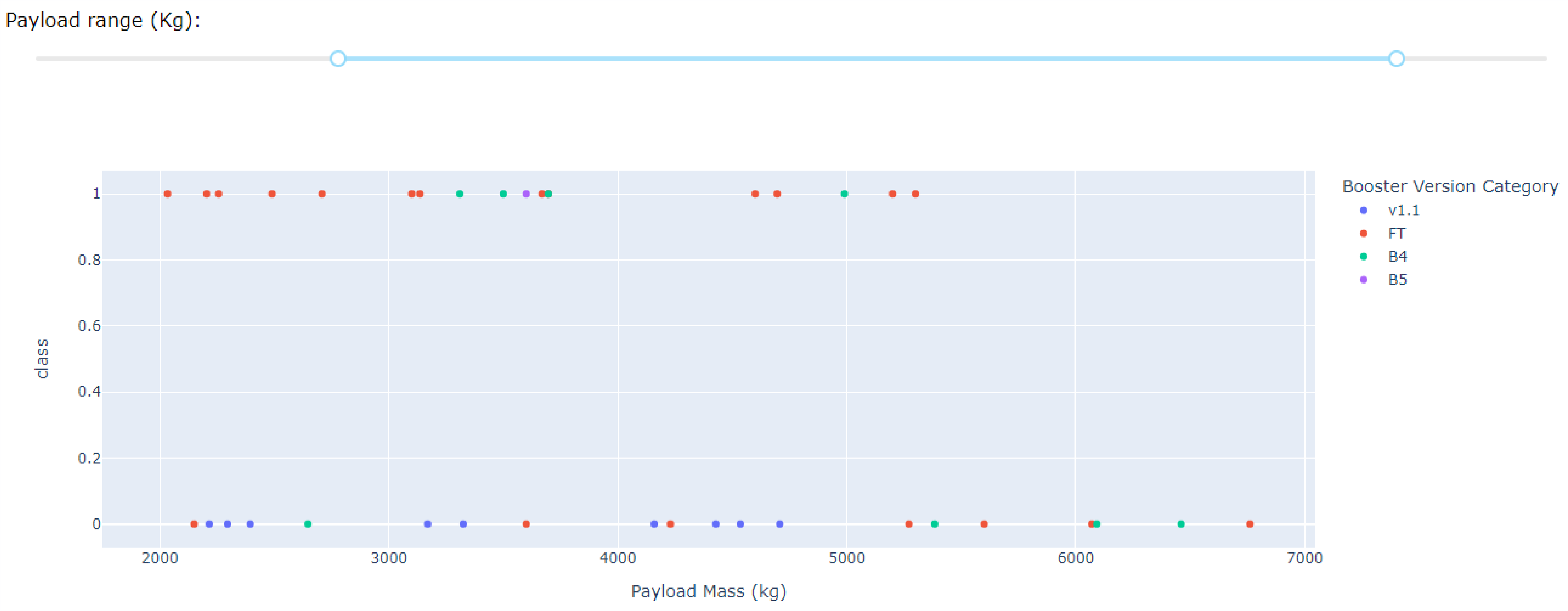




Results - Dashboard



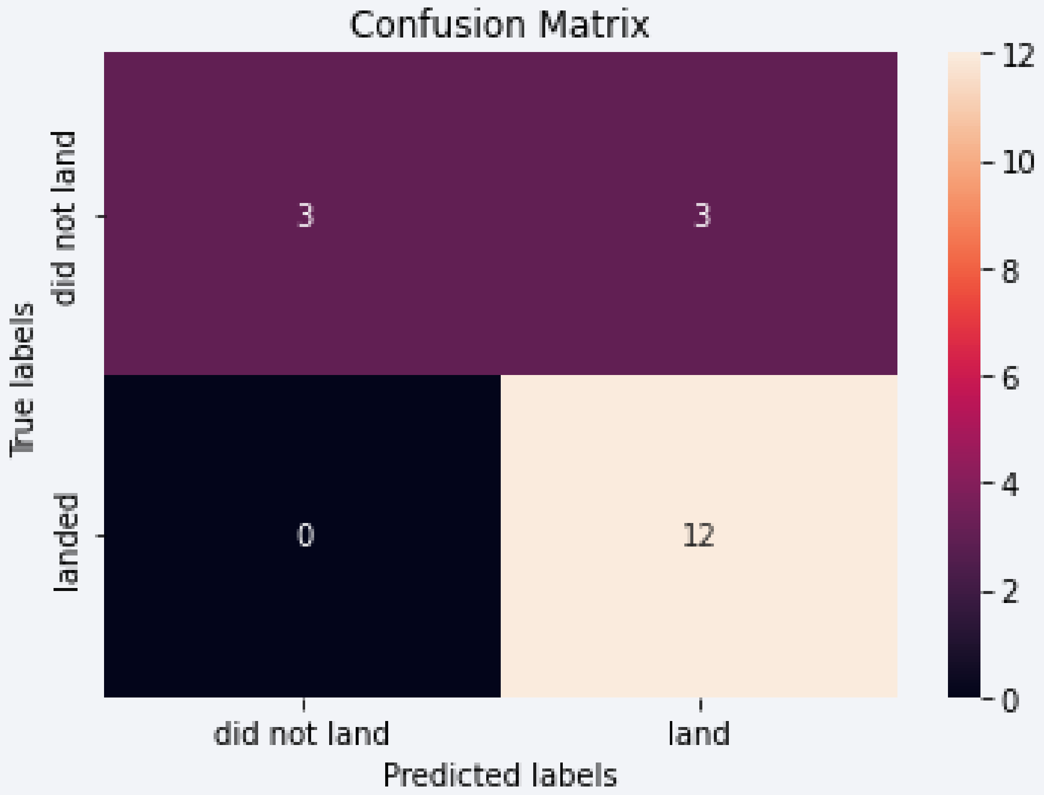
Results - Dashboard



Results – Predictive Analysis

Logistic Regression

Accuracy: 83%



Support Vector Machine

Accuracy: 83%

Results – Predictive Analysis

Decision Tree

Accuracy: 72%

KNN Classifier

Accuracy: 83%

Section 2

Insights drawn from EDA

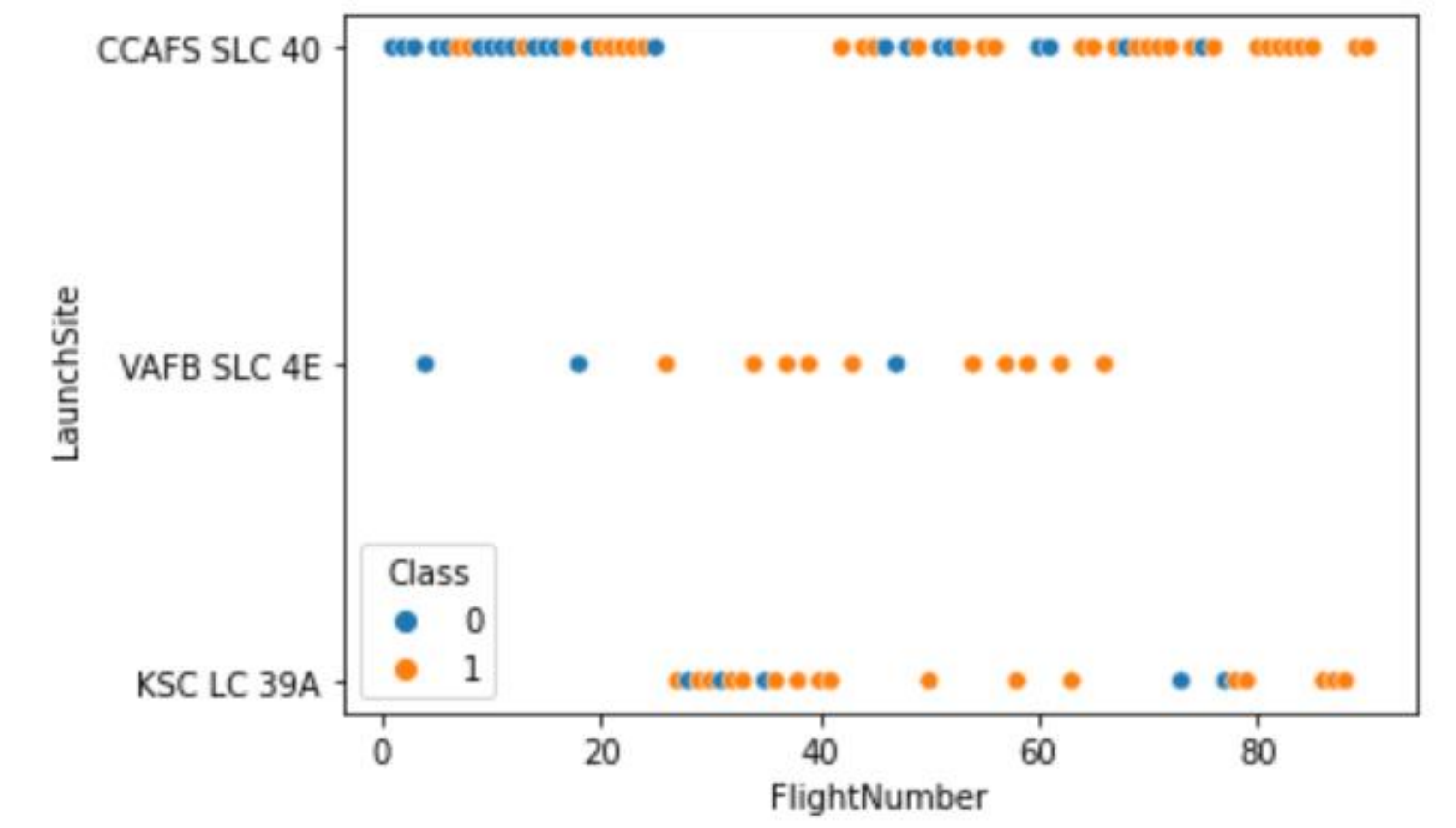
Flight Number vs. Launch Site

CCAFS SLC-40 site was not used for flight numbers between 25 and 40 – KSC LC 39A was used for those numbers instead

VAFB SLC 4E is not used after light number 70.

VAFB SLC 4E is the least used launch site whereas CCAFS SLC 40 is the most used launch site

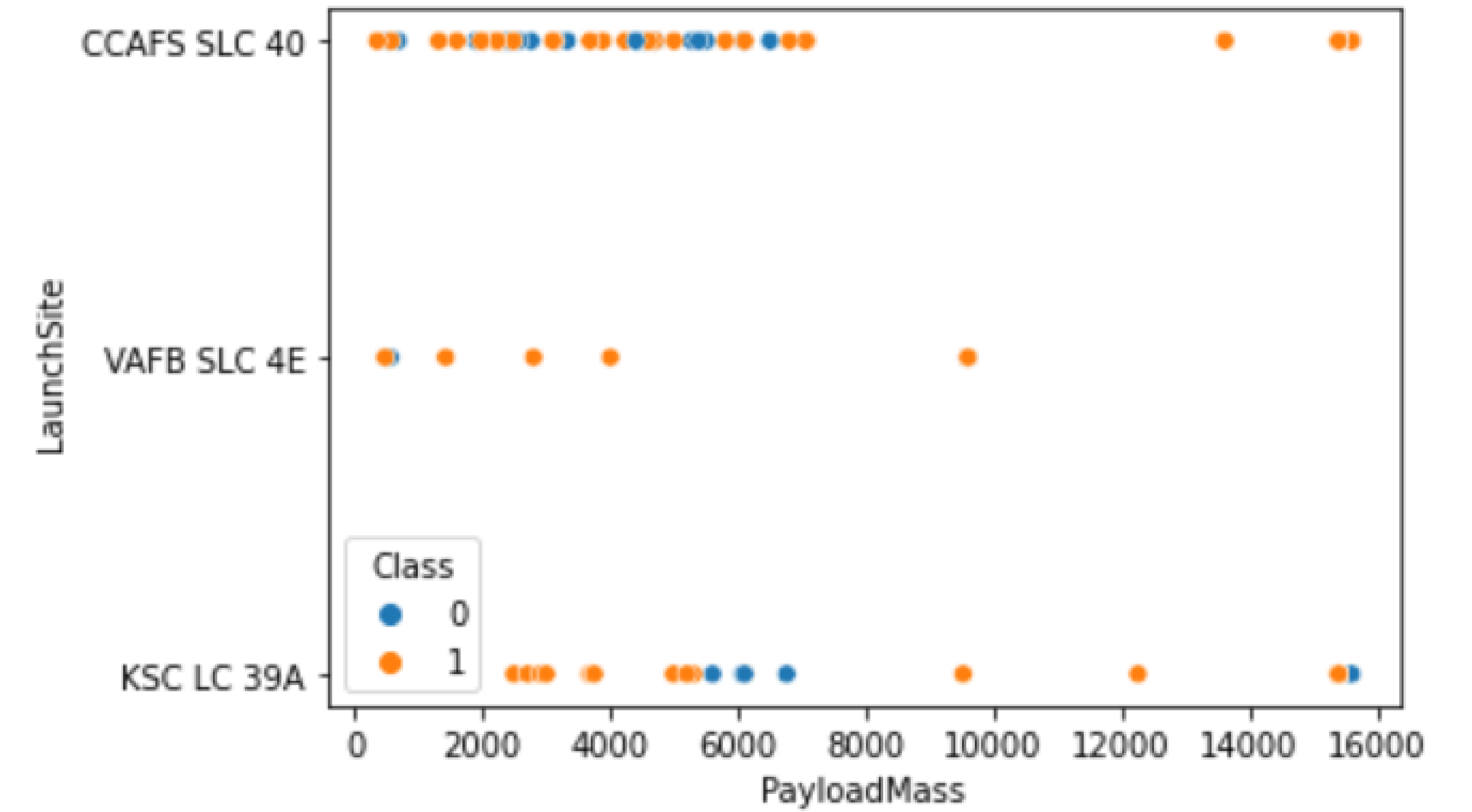
At higher flight numbers, the frequency of successful flights is higher



Payload vs. Launch Site

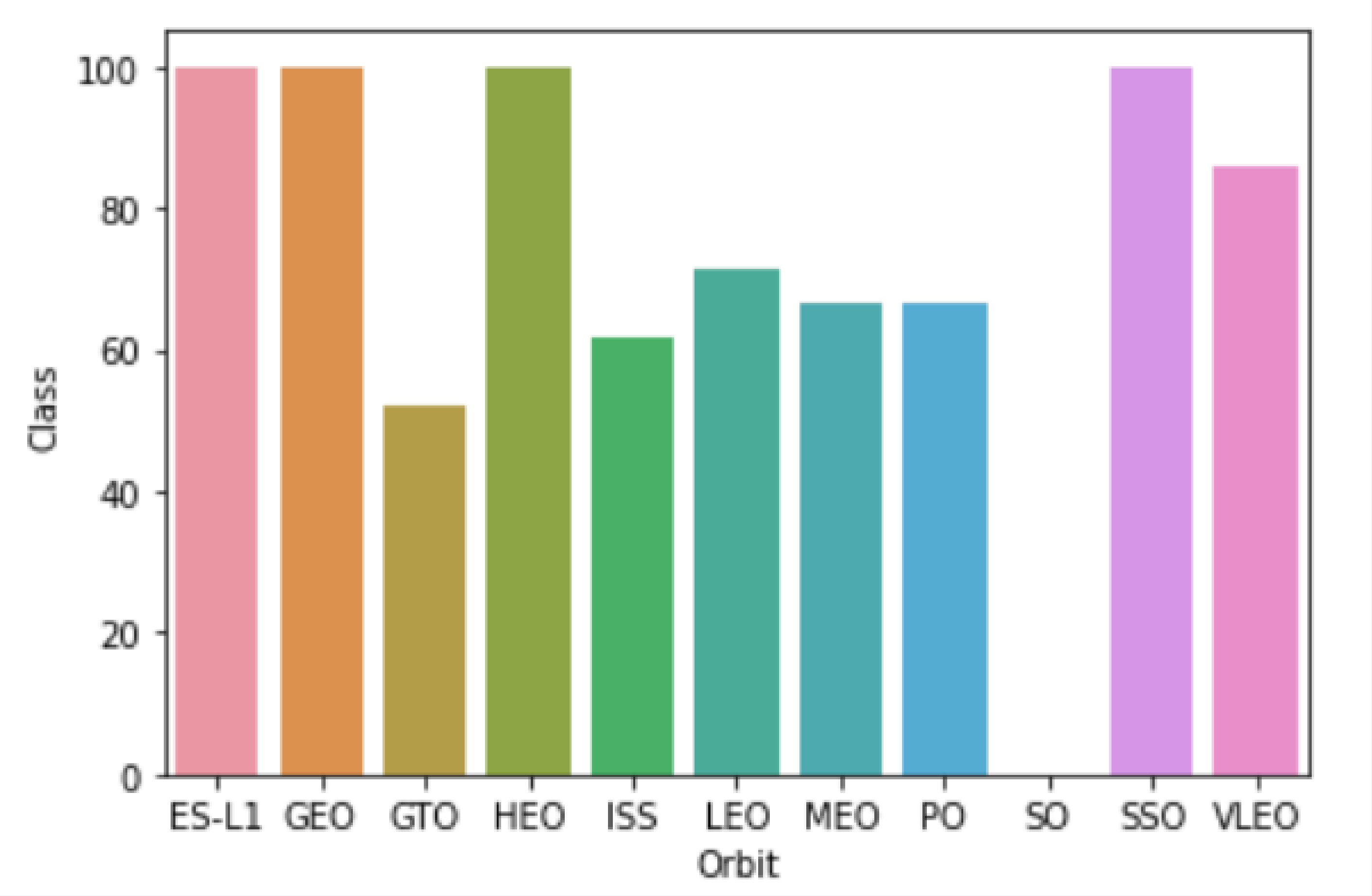
At VAFB SLC 4E, higher payload mass guaranteed successful landing

At VAFB SLC 4E the payload mass is always less than 10k



Success Rate vs. Orbit Type

Success rate for SO orbit is the lowest whereas for ES L1, GEO, HEO, SSO are the highest

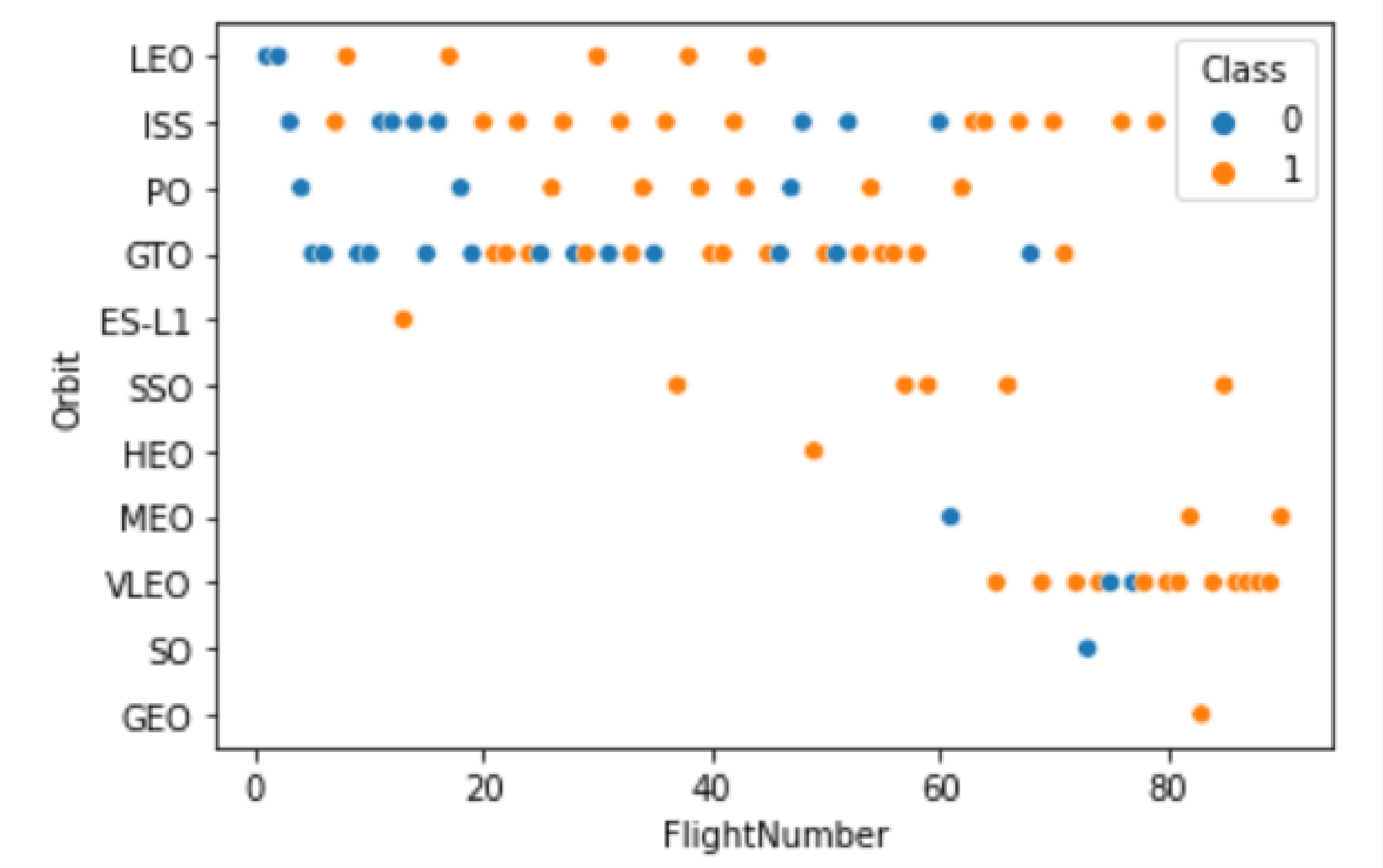


Flight Number vs. Orbit Type

ISS and GTO are more popular than the rest

VLEO was attempted only after flight 60

LEO was stopped after flight 50

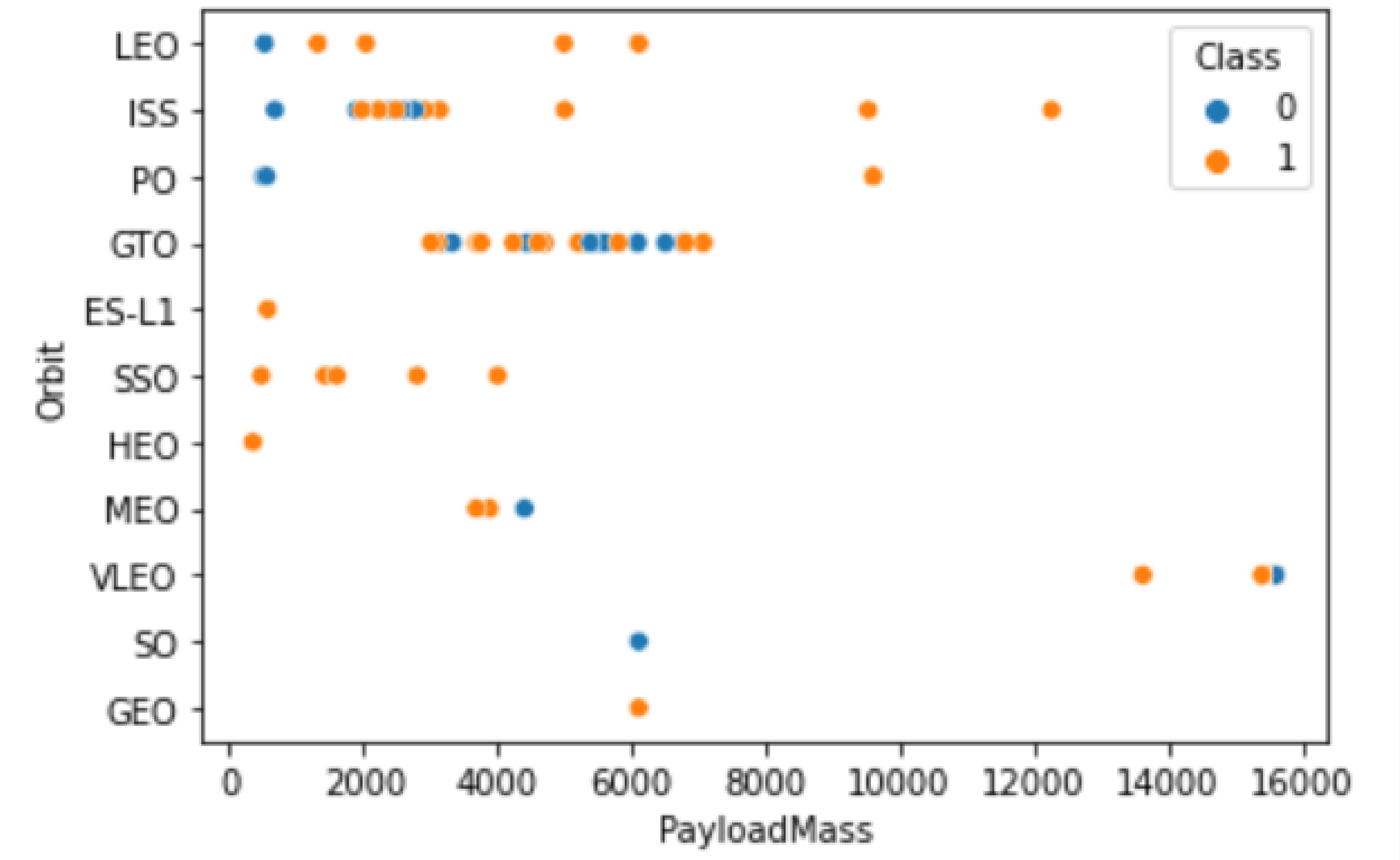


Payload vs. Orbit Type

For LEO, SSO, HEO, ES-L1 the payload is light

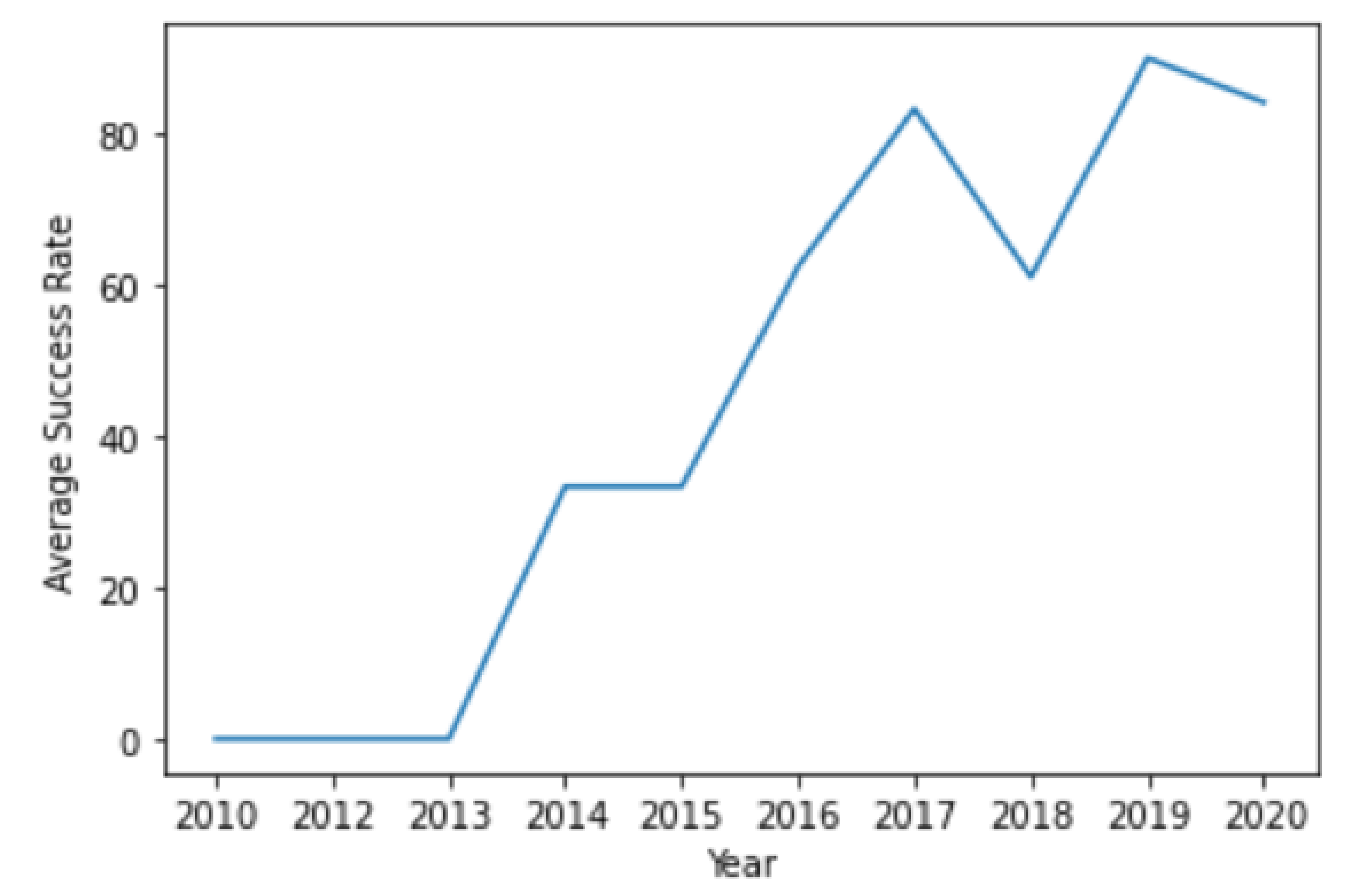
For VLEO the payload is the heaviest

For PO and LEO, landing fails at low payload mass while it is successful at higher payload mass



Launch Success Yearly Trend

The average success rate increases over the years from 2013



All Launch Site Names

```
SELECT UNIQUE(LAUNCH_SITE) from SPACEXTBL

Run time (seconds): 0.070
Status            : succeeded

LAUNCH_SITE
------------
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E
```

Launch Site Names Begin with 'CCA'

```
SELECT * from SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5

Run time (seconds): 0.011
Status            : succeeded

DATE       Time (UTC) BOOSTER_VERSION LAUNCH_SITE PAYLOAD                                                       PAYLOAD_MASS__KG_ ORBIT
---------- ---------- --------------- ----------- ------------------------------------------------------------- ----------------- ---------
2010-06-04 18:45:00   F9 v1.0  B0003  CCAFS LC-40 Dragon Spacecraft Qualification Unit                                          0 LEO
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)
2012-05-22 07:44:00   F9 v1.0  B0005  CCAFS LC-40 Dragon demo flight C2                                                       525 LEO (ISS)
2012-10-08 00:35:00   F9 v1.0  B0006  CCAFS LC-40 SpaceX CRS-1                                                                500 LEO (ISS)
2013-03-01 15:10:00   F9 v1.0  B0007  CCAFS LC-40 SpaceX CRS-2                                                                677 LEO (ISS)
```

Total Payload Mass – NASA (CRS)

```
SELECT SUM(PAYLOAD_MASS__KG_) from SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)'

Run time (seconds): 0.008
Status            : succeeded


1
-----
45596
```

Total payload mass for NASA (CRS) = 45596 kg

Average Payload Mass by F9 v1.1

```
SELECT AVG(PAYLOAD_MASS__KG_) from SPACEXTBL WHERE BOOSTER_VERSION LIKE 'F9 v1.1%'

Run time (seconds): 0.006
Status            : succeeded


1
----
2534
```

Average payload mass by F9 v1.1 = 2534kg

First Successful Ground Landing Date

```
SELECT MIN(DATE) from SPACEXTBL WHERE "Landing _Outcome" = 'Success (ground pad)'

Run time (seconds): 0.007
Status            : succeeded


1
----------
2015-12-22
```

First successful ground landing date: 2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
SELECT BOOSTER_VERSION,PAYLOAD_MASS__KG_ FROM SPACEXTBL WHERE "Landing _Outcome" = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000

Run time (seconds): 0.007
Status            : succeeded

BOOSTER_VERSION PAYLOAD_MASS__KG_
--------------- -----------------
F9 FT B1022                  4696
F9 FT B1026                  4600
F9 FT  B1021.2               5300
F9 FT  B1031.2               5200
```

Total Number of Successful and Failure Mission Outcomes

```
SELECT UNIQUE(MISSION_OUTCOME),COUNT(*) from SPACEXTBL GROUP BY MISSION_OUTCOME

Run time (seconds): 0.008
Status            : succeeded

MISSION_OUTCOME                  2
-------------------------------- --
Failure (in flight)               1
Success                          99
Success (payload status unclear)  1
```

Mission successful 100 times out of 101

Boosters Carried Maximum Payload

```
SELECT BOOSTER_VERSION,PAYLOAD_MASS__KG_ from SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) from SPACEXTBL)

Run time (seconds): 0.009
Status            : succeeded

BOOSTER_VERSION PAYLOAD_MASS__KG_
--------------- -----------------
F9 B5 B1048.4               15600
F9 B5 B1049.4               15600
F9 B5 B1051.3               15600
F9 B5 B1056.4               15600
F9 B5 B1048.5               15600
F9 B5 B1051.4               15600
F9 B5 B1049.5               15600
F9 B5 B1060.2               15600
F9 B5 B1058.3               15600
F9 B5 B1051.6               15600
F9 B5 B1060.3               15600
F9 B5 B1049.7               15600
```

2015 Launch Records

```
SELECT DATE, "Landing _Outcome", BOOSTER_VERSION, LAUNCH_SITE from SPACEXTBL WHERE "Landing _Outcome" = 'Failure (drone ship)' AND DATE LIKE '2015-%'

Run time (seconds): 0.008
Status            : succeeded

DATE       Landing _Outcome     BOOSTER_VERSION LAUNCH_SITE
---------- -------------------- --------------- -----------
2015-01-10 Failure (drone ship) F9 v1.1 B1012   CCAFS LC-40
2015-04-14 Failure (drone ship) F9 v1.1 B1015   CCAFS LC-40
```

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

```
SELECT UNIQUE("Landing _Outcome"), COUNT(*) as ct from SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY "Landing _Outcome" ORDER BY ct DESC

Run time (seconds): 0.009
Status            : succeeded

Landing _Outcome       CT
---------------------- --
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
```

Section 4

Launch Sites Proximities Analysis

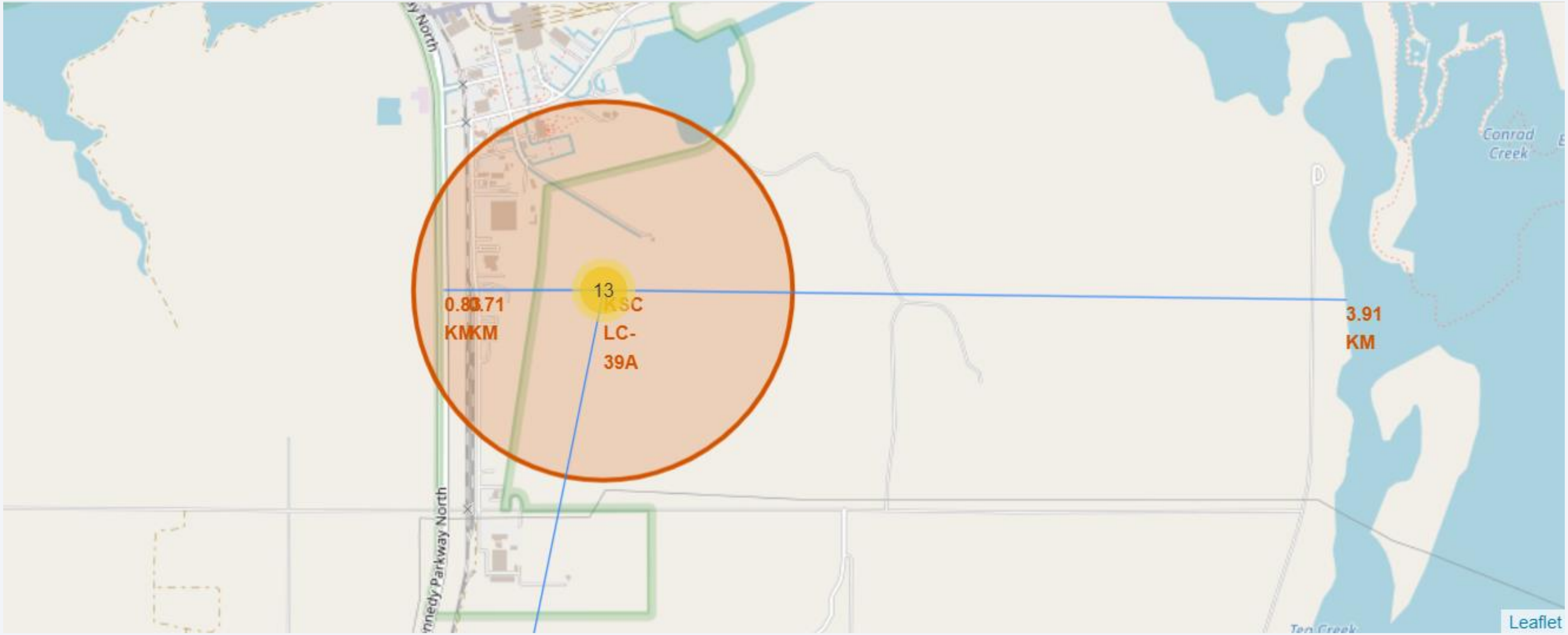
Falcon 9 Launch Sites

The launch sites are marked with a red circle and labelled in red font

Launch Outcomes at VAFB SLC 4E

Red markers indicate failed outcomes whereas green ones indicate successful outcomes

Distance of KSC LC 39A from Amenities

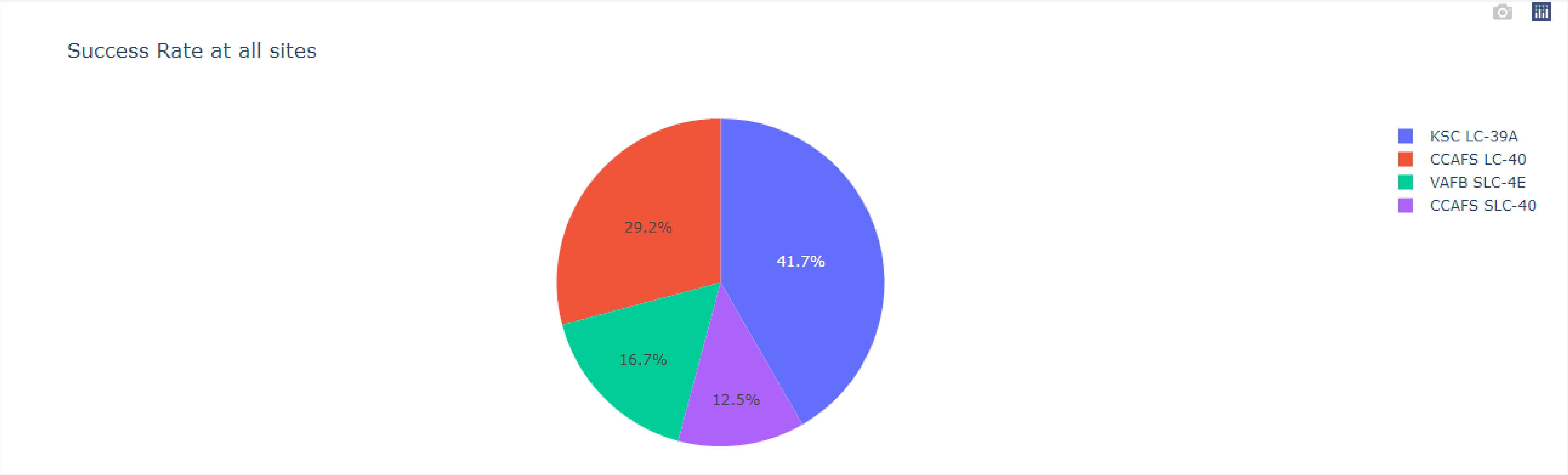


Distance to coastline: 3.91km, railway: 0.71km, highway: 0.80km

Section 5

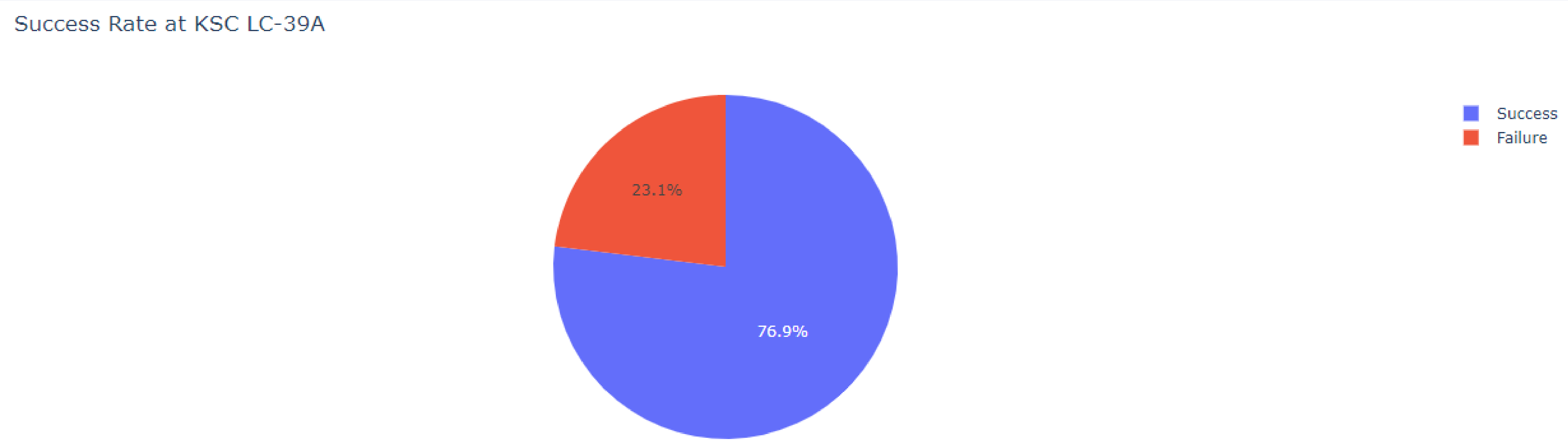
Build a Dashboard with Plotly Dash

Proportion of Total Successes at each Launch Site

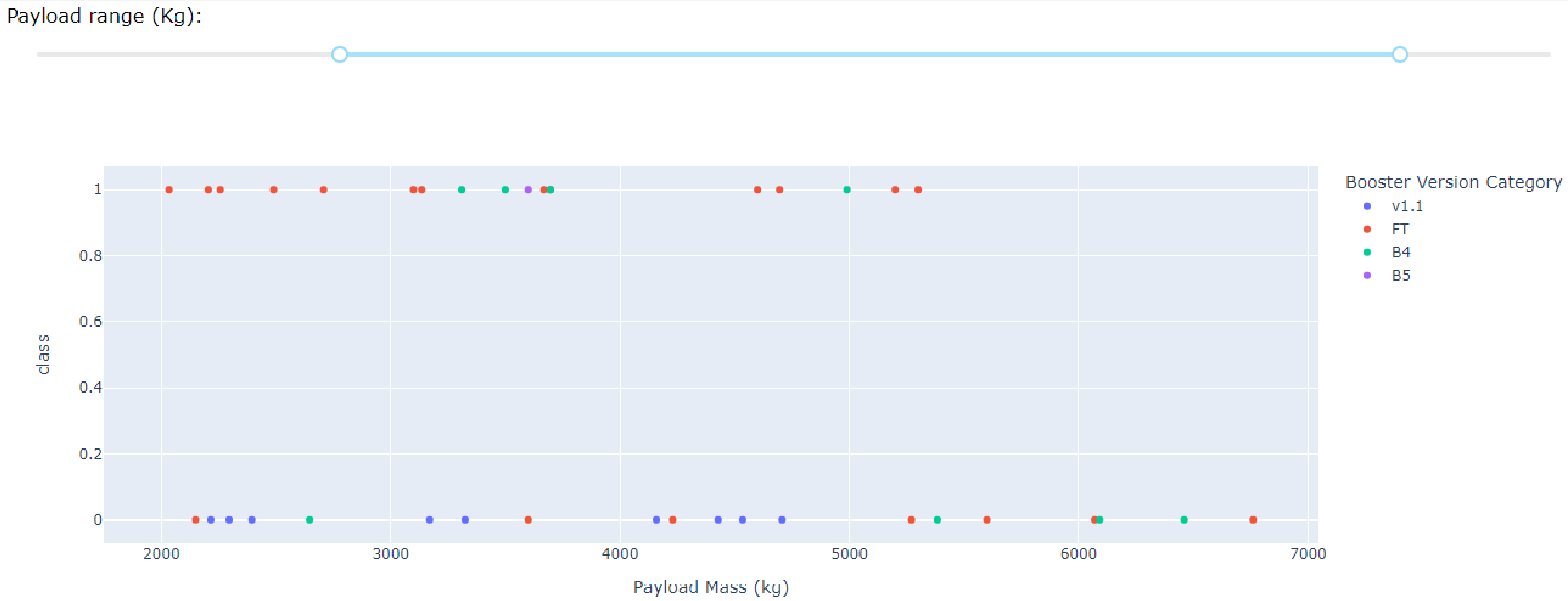


Most of the successes are from KSC LC-39A launch site

Launch Site with the Highest Proportion of Success



Launch Outcome versus Payload Mass for All Sites



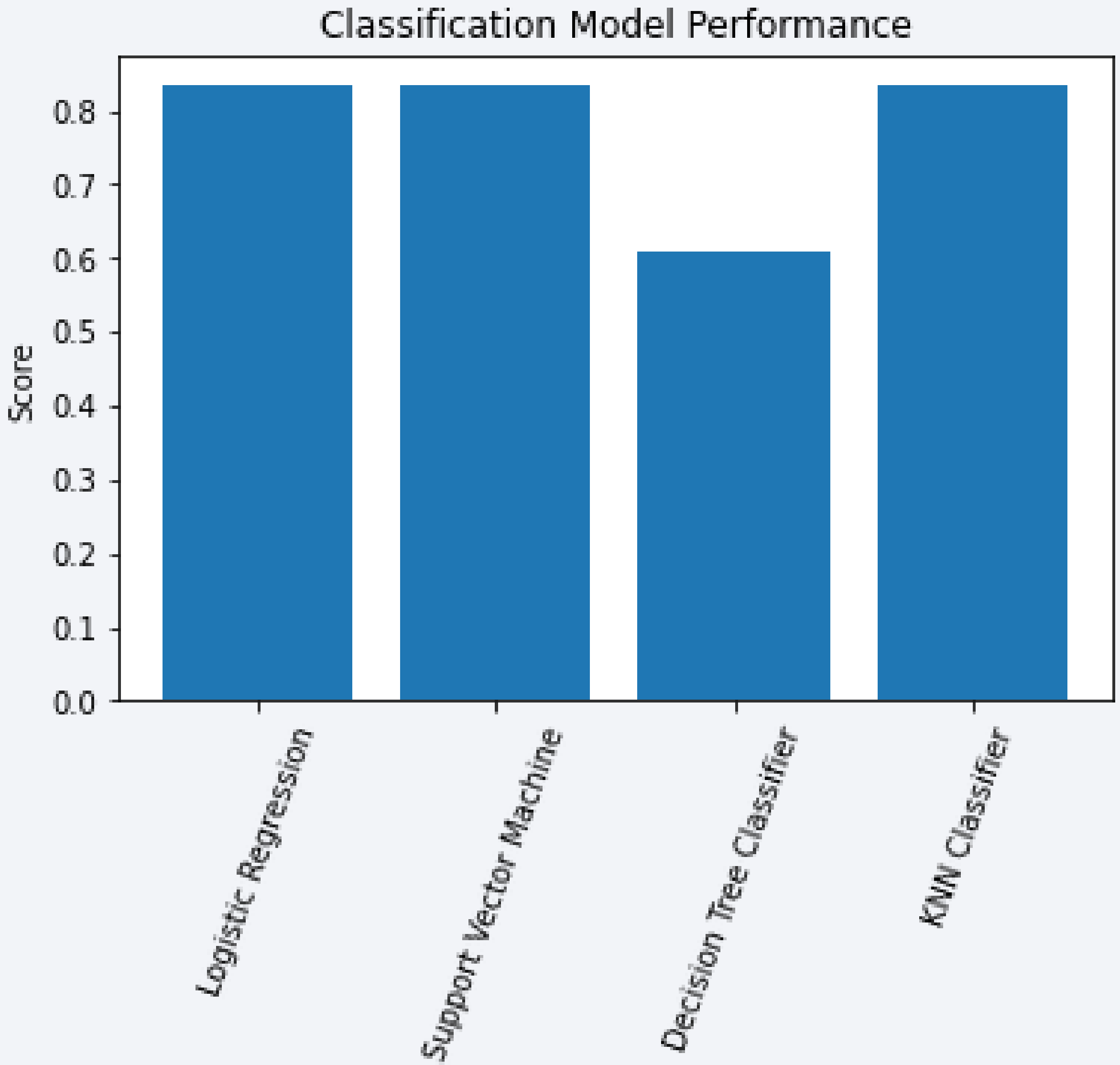
The payload mass range is between 2000kg and 9000kg

FT has the highest success rate while v1.1 has the lowest

Section 6

Predictive Analysis (Classification)

Classification Accuracy



Any of Logistic Regression, SVM, and KNN classifier can be used since they all have the same accuracy

Confusion Matrix

Logistic Regression

Accuracy: 83%

The Logistic Regression model has no problem identifying cases where the first stage truly landed

When it comes to cases where the first stage did not land, the model fails 50% of the time

Conclusions

- Over the years, the average success rate of the Falcon 9 first stage landing has increased
- Rocket launch sites generally tend to be close to the coastline, railway and highway and far away from cities
- KSC LC 39A is the launch site where the likelihood of first stage recovery is the highest
- It is possible to predict the fate of the first stage of Falcon 9 with a reasonably high degree of accuracy, given feature information such as launch site, payload mass, booster version, orbit type etc, using a logistic regressor
- As more launches take place, more data points could be added to the master data set and the regressor can be trained to be better at predicting recovery

Thank you!