IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Nijat Alizada

28 December 2021

Outline

- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Introduction

SpaceY is the youngest and most dynamic aerospace manufacturer providing space transportation services. Our mission is to make space travel safe and affordable.

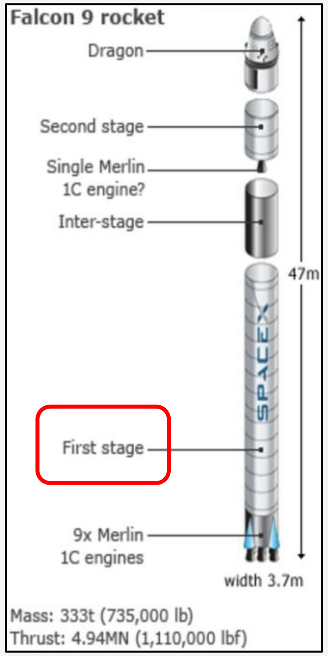
This project intends to answer the following:

- What are the main features to predict whether first stage of rockets will land successfully?

We will be using publicly available data from SpaceX to answer this business question.

Problem Statement:

In current commercial space age, companies are making space travel more affordable. Our biggest competitor is SpaceX, and its main advantage is relatively inexpensive rocket launches. Cost of SpaceX Falcon 9 rocket launch is 62 million dollars, whereas other providers cost upward of 165 million dollars each. **Much of the savings is because SpaceX can reuse the first stage of their rockets after being launched.** For SpaceY to compete, our goal is to be able to recover first stage of rockets after their launch.



Source: SpaceX Falcon 9 rocket enjoys successful maiden flight - BBC News

Section 1

Methodology

Methodology

Executive Summary

- **Data collection**
  - Data was collected from SpaceX API and then Data Scraping and Wrangling took place
- **Exploratory Data Analysis (EDA)**
  - Various scatter charts, bar graphs and pie charts were generated as part of EDA
  - Circa 10 different SQL queries were executed for EDA purposes
- **Interactive visual analytics**
  - Interactive Map was created with Folium, as well as a dashboard using Plotly Dash
- **Predictive analysis using classification models**
  - Four different Machine Learning models were tested and evaluated

Data Collection

Following flowchart describes the process of data collection stages of the analysis:

- **SpaceX API** — Using Python requests library to make SpaceX API calls
- **Data Scraping** — Use of BeautifulSoup library to scrape the response from API call
- **Data Wrangling** — Filtering the data and dealing with missing values

See next pages for more details of each stage

Data Collection – SpaceX API

Import required libraries

- requests, pandas, numpy, datetime

Use GET method of requests library to make an API call to SpaceX URL

- https://api.spacexdata.com/v4/launches/past

Apply pd.json_normalize method on response.json() to convert the response into pandas DataFrame object

```
import requests
import pandas as pd
import numpy as np
import datetime

pd.set_option('display.max_columns', None)
pd.set_option('display.max_colwidth', None)
```

```
spacex_url = 'https://api.spacexdata.com/v4/launches/past'
response = requests.get(spacex_url)
data = pd.jason_normalize(response.json())
```

GitHub URL: [Link]

Data Collection - Scraping

Wikipedia page with the list of Falcon 9 and Falcon Heavy launches was accessed to scrape data.

Following sequence of activities were executed for Web Scraping:

1. Import required libraries (requests, BeautifulSoup, pandas, etc.)
2. GET method to make API call and then create soup object
3. Use find_all method on soup object to extract tables and add them into DataFrame

2021

[hide] Flight No.	Date and time (UTC)	Version, Booster[b]	Launch site	Payload[c]	Payload mass	Orbit	Customer	Launch outcome	Booster landing
104	8 January 2021 02:15[604]	F9 B5 △ B1060.4	CCSFS, SLC-40	Türksat 5A[605]	3,500 kg (7,700 lb)	GTO	Türksat	Success	Success (drone ship)
	A 3,500 kg (7,700 lb) satellite intended to be stationed at 31.0° east.[605] This is the most powerful satellite in Türksat's fleet[606] and will provide Ku-band television broadcast services over Turkey, the Middle East, Europe and Africa. The satellite was injected in to a Super-synchronous transfer orbit of 280 km × 55,000 km (170 mi × 34,180 mi) with 17.6° inclination.[607]								
105	20 January 2021 13:02[608]	F9 B5 △ B1051.8[609]	KSC, LC-39A	Starlink 16 v1.0 (60 satellites)	15,600 kg (34,400 lb)	LEO	SpaceX	Success	Success (drone ship)
	The first booster to successfully launch and land eight times. Achieved a record turnaround time between two launches of the same booster of only 38 days and brought the total of launched Starlink satellites to over 1000.[610] SpaceX stated that the landing would occur during higher winds than usual; this test to expand the landing envelope was successfully passed by the booster.[611]								
106	24 January 2021 15:00[612]	F9 B5 △ B1058.5[613]	CCSFS, SLC-40	*Transporter-1* (143 smallsat rideshare)	~5,000 kg (11,000 lb)	SSO	Various	Success	Success (drone ship)
	First dedicated smallsat rideshare launch, targeting a 525 km (326 mi) altitude orbit.[614] The launch deployed a record 143 satellites, consisting of 120 CubeSats, 11 microsatellites, 10 Starlinks, and 2 transfer stages. In addition, 2 hosted payloads and 1 non-separating dummy satellite[615] were[*failed verification*] launched.[616] These include SpaceBEE (x 36), Lemur-2 (x 8), ICEYE (x 3), UVSQ-SAT,[617] ELaNa 35 (PTD-1),[381] and multiple Kepler nanosats.[618][619] D-Orbit flew their ION SCV LAURENTIUS, 10 Starlink satellites were placed in a polar orbit[620] and 2 of 15 payloads remained attached to SHERPA-FX1. Exolaunch deployed several small satellites and cubesats via their own deployment mechanisms. First flight of a Falcon 9 with a SHERPA-FX transfer stage called SHERPA-FX1.[621][622]								
107	4 February 2021 06:19[623]	F9 B5 △ B1060.5[624]	CCSFS, SLC-40	Starlink 18 v1.0 (60 satellites)	15,600 kg (34,400 lb)	LEO	SpaceX	Success	Success (drone ship)
	This marked the fastest turnaround to date, at 27 days, and the first time a Falcon 9 flies twice within a month.[625]								
108	16 February 2021 03:59:37[626]	F9 B5 △ B1059.6	CCSFS, SLC-40	Starlink 19 v1.0 (60 satellites)[627]	15,600 kg (34,400 lb)	LEO	SpaceX	Success	Failure (drone ship)
	A hole in a heat-shielding engine cover, which likely developed through fatigue, allowed recirculating hot exhaust gases to damage one of the Merlin 1D first-stage engines, causing it to shut down early during ascent. Engine-out capability of the Falcon 9 allowed the mission to continue and successfully deploy the 60 Starlink satellites to orbit.[628] The issue caused the booster to fail its landing attempt and miss the droneship *Of Course I Still Love You* (OCISLY) after its entry burn, breaking the longest streak of 24 landing successes.[629] During this mission, *GO Ms. Tree* and *GO Ms. Chief* were used for the last time to recover the fairings.[630][631] After this mission, both ships were retired because SpaceX no longer plans to catch the fairings.[632]								

Source: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

GitHub URL: Link

Data Wrangling

In the Data Wrangling stage, missing values were identified and handled.

It was identified that "PayloadMass" feature of the dataset had several missing values, which were imputed by the mean value of the feature.

Note: Missing values in the "LandingPad" column meant that no landing pad was used, therefore they were not handled.

Before

FlightNumber	0
Date	0
BoosterVersion	0
PayloadMass	5
Orbit	0
LaunchSite	0
Outcome	0
Flights	0
GridFins	0
Reused	0
Legs	0
LandingPad	26
Block	0
ReusedCount	0
Serial	0
Longitude	0
Latitude	0
dtype: int64	

After

FlightNumber	0
Date	0
BoosterVersion	0
PayloadMass	0
Orbit	0
LaunchSite	0
Outcome	0
Flights	0
GridFins	0
Reused	0
Legs	0
LandingPad	26
Block	0
ReusedCount	0
Serial	0
Longitude	0
Latitude	0
dtype: int64	

GitHub URL: Link

EDA with Data Visualization

Following charts were plotted as part of the Exploratory Data Analysis:

- Scatter plot of Flights vs Payload
    - This chart indicates how continuous launch attempts and Payload variables affect the launch outcome
- Scatter plot of Flights vs Launch Site
    - This plot was used to visualize the relationship of launch outcomes among Launch Sites
- Scatter plot of Payload Mass vs Launch Site
    - This visual was used to understand relationship between Payload and Launch Sites
- Bar chart of Success Rate of each Orbit Type
    - This chart was created to visually check if there are any relationship between success rate and orbit type
- Scatter plot of Flights vs Orbit Type
    - Here we checked if there is any relationship between Flights and Orbit Types
- Line chart for launch success yearly trend
    - To get the average launch success trend over 10-year period

GitHub URL: [Link]

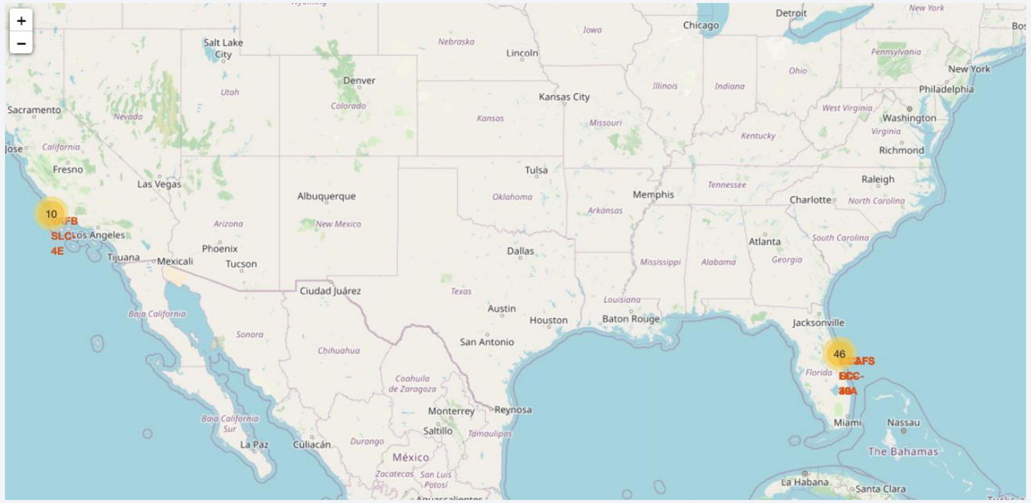
EDA with SQL

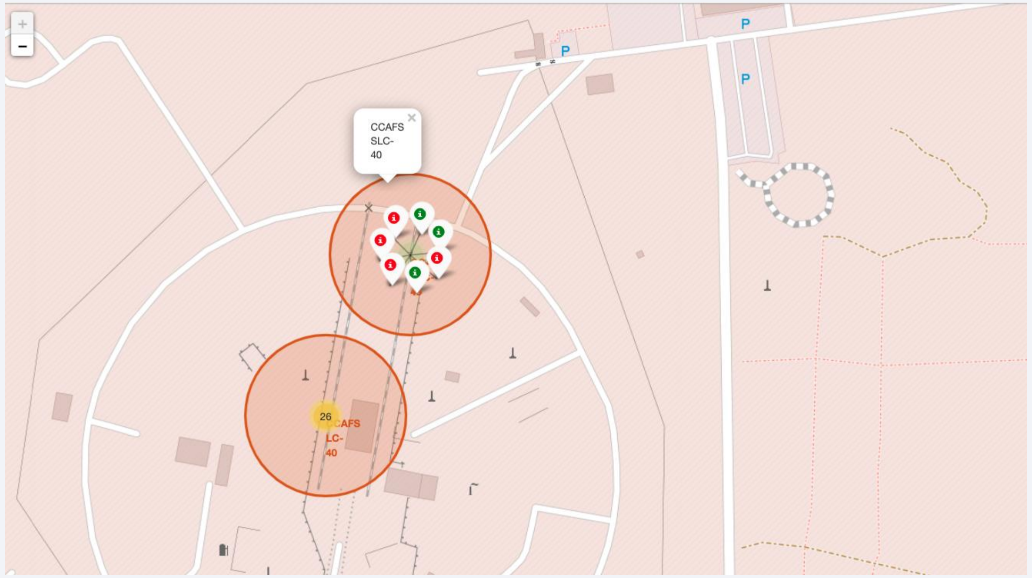
Following information was retrieved through SQL queries during EDA stage:

- The names of unique launch sites in the space mission
- First 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 failed 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 year 2015
- Ranking the count of landing outcomes between 2010-06-04 and 2017-03-20

GitHub URL: [Link]

Build an Interactive Map with Folium





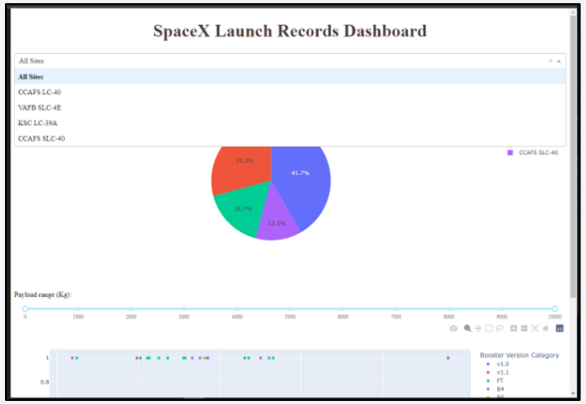
- folium.Circle and folium.Marker objects were added to highlight each launch site
- To indicate the number of failed and success launches at each site, MarkerCluster objects were created
- Finally, distances between a launch site to its proximities were calculated using MousePosition and folium.PolyLine objects

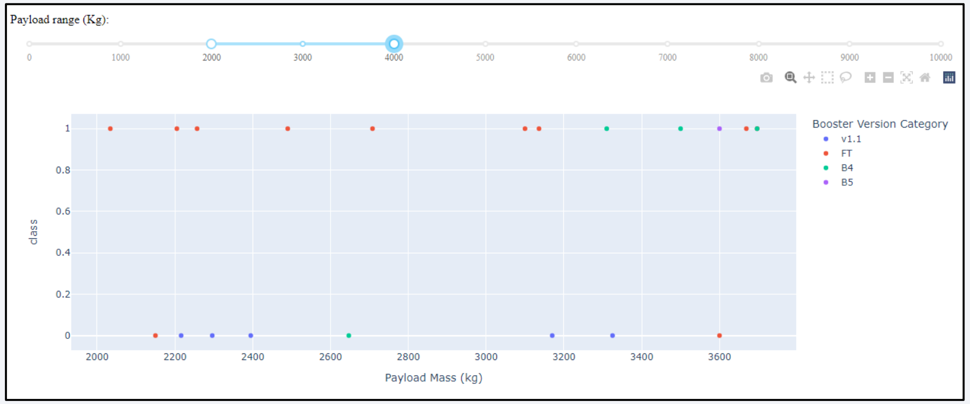
GitHub URL: [Link]

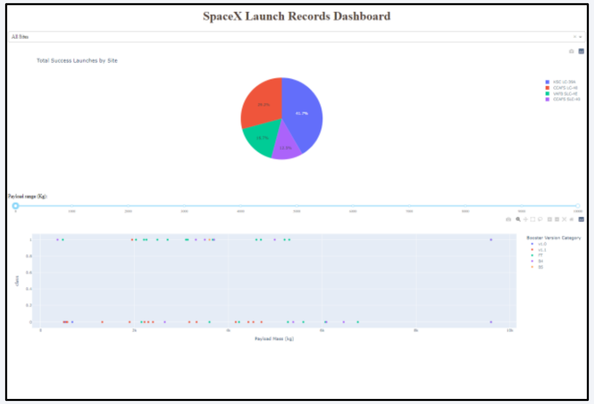
Build a Dashboard with Plotly Dash

Following elements were added into the Dashboard:

- Dropdown of launch sites
- Success launch Pie Chart
- Range Slider for payload mass
- Success vs payload scatter chart







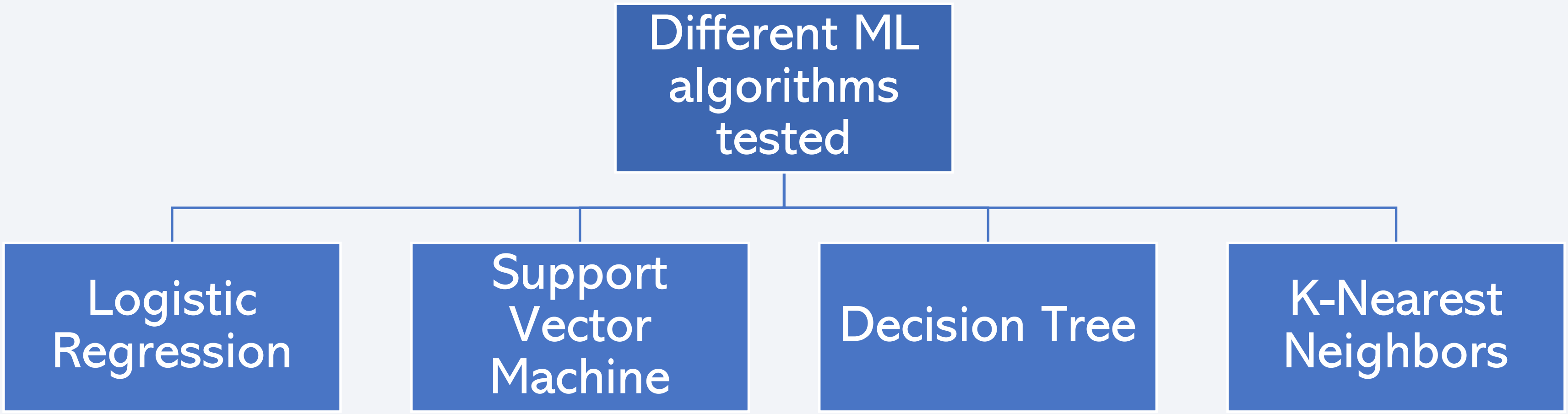
GitHub URL: [Link]()

Predictive Analysis (Classification)

Following steps were used prior to applying Machine Learning algorithms:

1. Standardize the data using preprocessing.StandardScaler()
2. Splitting the data into train and test sets using train_test_split()

Below are different Machine Learning algorithms that were used:

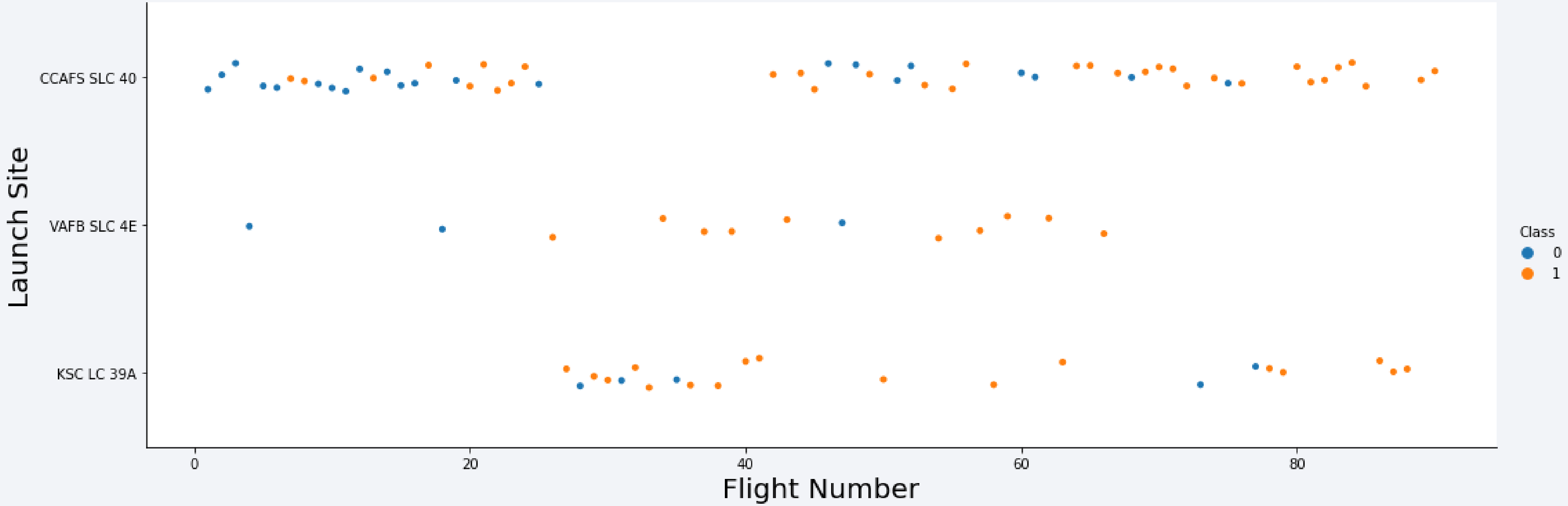


GitHub URL: [Link]

Section 2

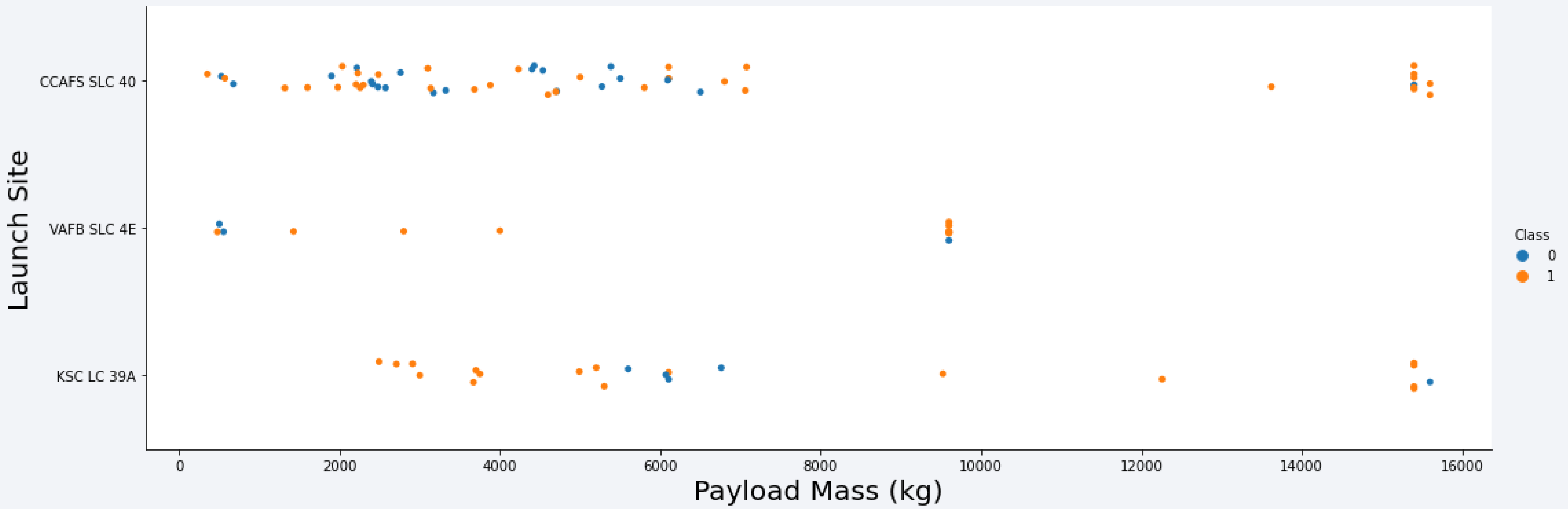
Insights drawn from EDA

Flight Number vs. Launch Site



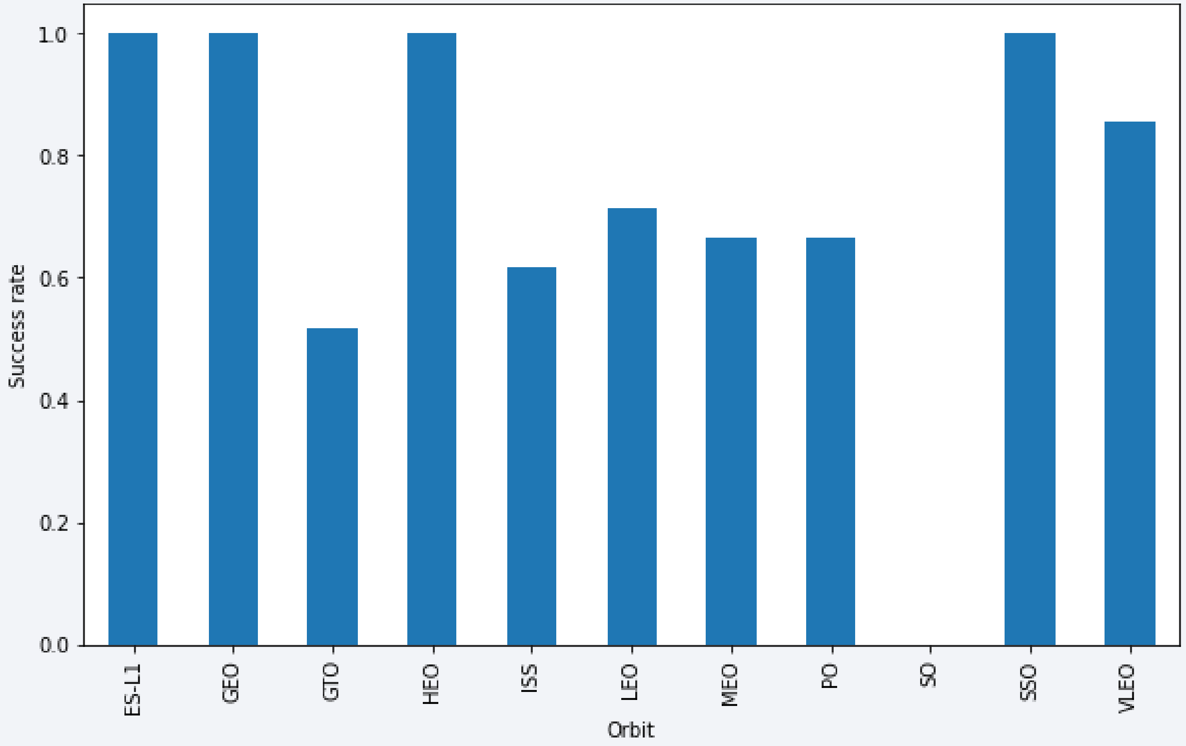
This chart indicates the number of flights per each launch site and whether these launches had successful first stage landing or not (blue dots show unsuccessful landings). We can observe that CCAFS SLC 40 has a greater number of launches compared to other sites, however success rate is 60%, while KSC LC 39A and VAFB SLC 4E have a success rate of 77%.

Payload vs. Launch Site



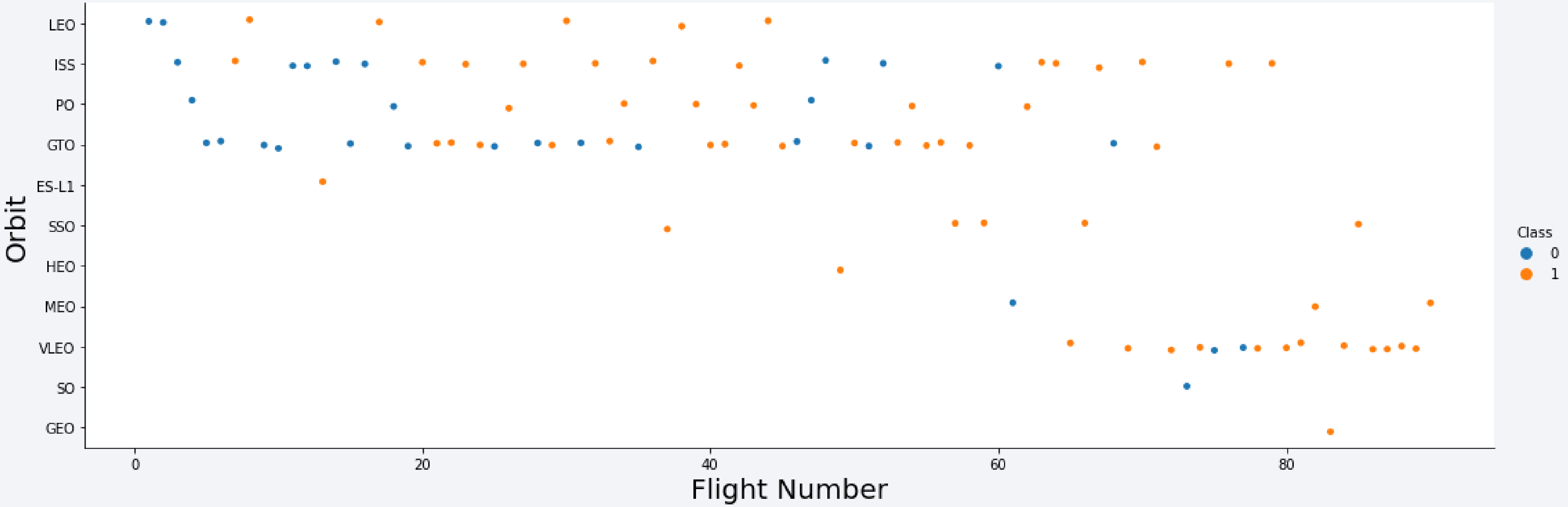
Scatter point chart above demonstrates launches distributed by payload mass. It is observed that significant portion of the launches fall below payload mass of 8000 kg. In fact, there was no launch from VAFB SLC 4E where payload mass was over 10000 kg. Most of such launches were executed from CCAFS SLC 40 with 83% success rate.

Success Rate vs. Orbit Type



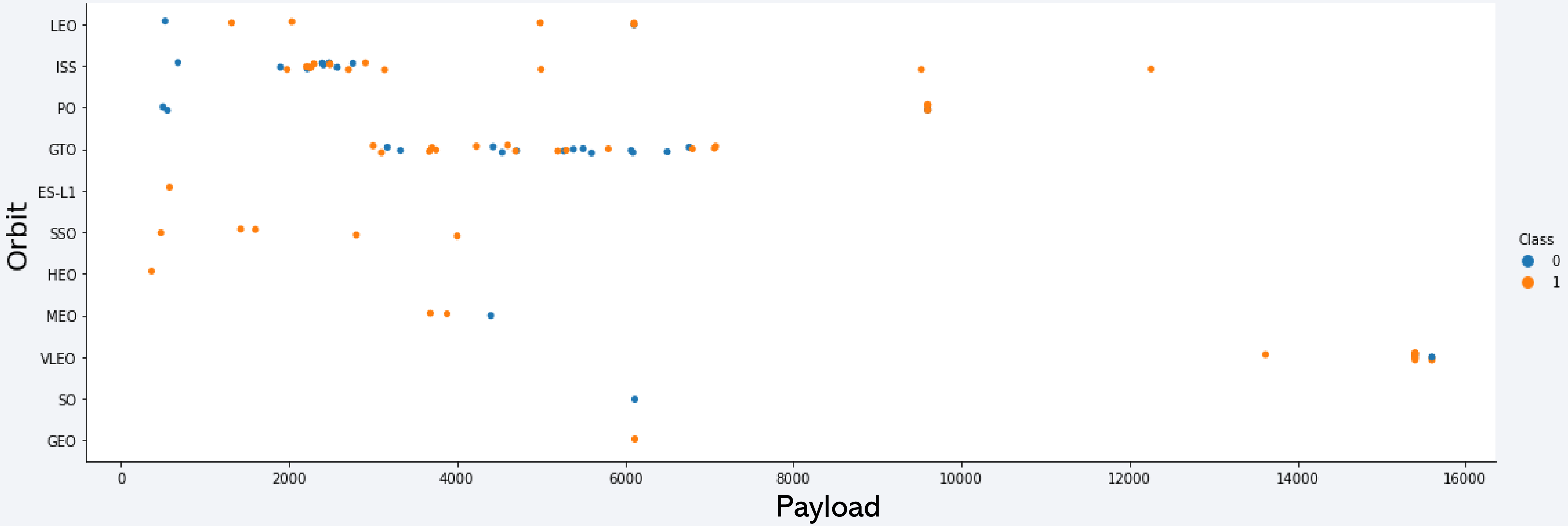
Bar graph above shows orbits and whether launches to these orbits were successful. We can observe that ES-L1, GEO, HEO and SSO have 100% success rate. SO has the lowest success rate – there was no launch with successful first stage landing to this orbit. Similarly, GTO is also one of the orbit types with lowest success rate.

Flight Number vs. Orbit Type



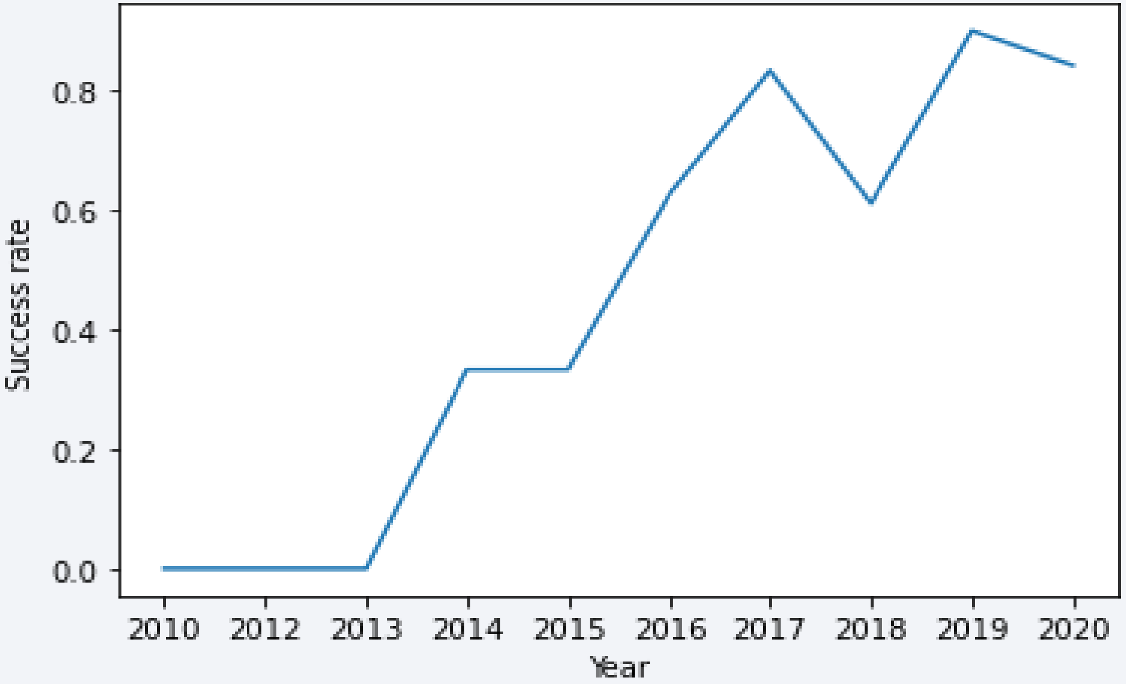
As a continuation of previous topic, we wanted to observe whether success rate of each orbit has any relationship with number of flights. While SO had the lowest success rate in previous visual, now we can see that there was only 1 flight to this orbit, which is not indicative of any pattern.

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 and negative landing are distributed without any pattern.

Launch Success Yearly Trend



Looking into the yearly trend of successful launches, we can observe that the success rate kept increasing since 2013.

All Launch Site Names

```
In [6]: %sql SELECT DISTINCT(LAUNCH_SITE) FROM SPACEXTBL;

 * ibm_db_sa://qsw28198:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

Out[6]:

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

Here, we have extracted list of all launch sites within the dataset:

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

Launch Site Names Begin with 'CCA'

In [12]:
```
%sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE ('CCA%') LIMIT 5
```

```
 * ibm_db_sa://qsw28198:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

Out[12]:

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

Above query displays 5 records where launch site names begin with the string “CCA”.

Total Payload Mass

In [14]:
```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)'
```

```
 * ibm_db_sa://qsw28198:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

Out[14]:

1
45596

Total payload mass carried by boosters launched by NASA (CRS) was 45596 kgs.

Average Payload Mass by F9 v1.1

```
In [15]: %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1'

 * ibm_db_sa://qsw28198:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

Out[15]:

1
2928

Average payload mass carried by booster version F9 v1.1 was 2928 kgs.

First Successful Ground Landing Date

```
In [16]: %sql SELECT MIN(DATE) FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Success (ground pad)';

 * ibm_db_sa://qsw28198:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

Out[16]:

1
2015-12-22

22 December 2015 was the date when the first successful landing outcome in ground pad was achieved.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
In [18]: %sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;

 * ibm_db_sa://qsw28198:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

Out[18]:

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

Above query displays list of the booster names which have successful drone ship landing with payload mass between 4000 and 6000 kgs.

Total Number of Successful and Failure Mission Outcomes

In [22]:
```
%sql SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS COUNT FROM SPACEXTBL GROUP BY MISSION_OUTCOME;
```

```
 * ibm_db_sa://qsw28198:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

Out[22]:

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

We can observe that 99% mission outcomes were successful, with only 1 failure.

Boosters Carried Maximum Payload

```
In [23]: %sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);

 * ibm_db_sa://qsw28198:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

Out[23]:

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

Above query displays list of the booster versions which have carried the maximum payload mass.

2015 Launch Records

In [24]:

```
%sql SELECT LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Failure (drone ship)' AND DATE >= '2015-01-01' AN
```

```
 * ibm_db_sa://qsw28198:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

Out[24]:

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

There were 2 failed landings in drone ship in 2015 and above query displays their booster versions and launch sites – both were launched from CCAFS LC-40.

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

In [27]:
```
%sql SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS COUNT FROM SPACEXTBL WHERE DATE >= '2010-06-04' AND DATE <= '2017-03-20' GROUP BY LANDING__C
```

```
 * ibm_db_sa://qsw28198:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

Out[27]:

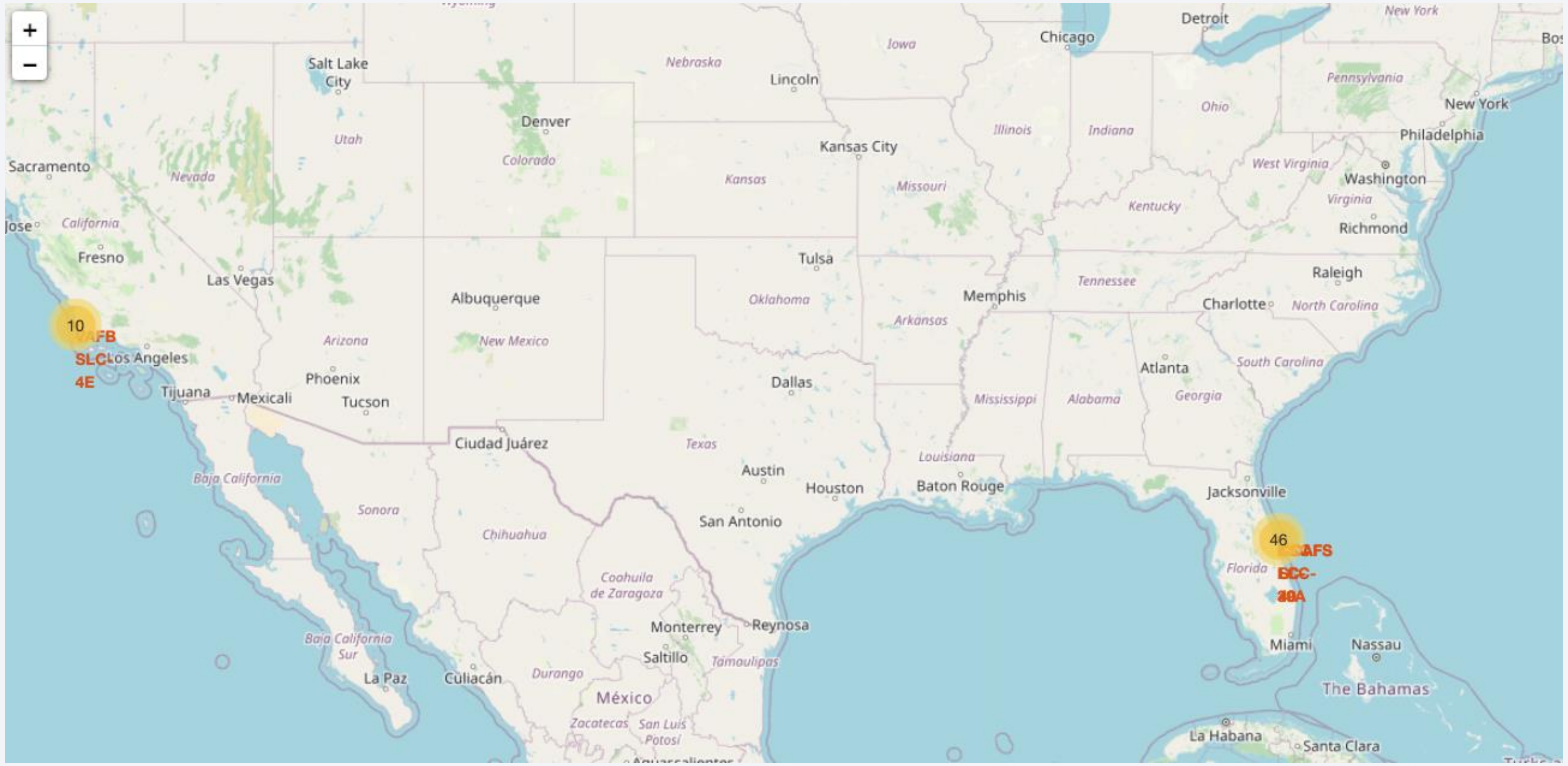
landing__outcome	COUNT
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Ranking of landing outcomes between 6 June 2010 and 20 March 2017 is displayed above. We can observe that the most frequent outcome was "no attempt" with 10 such cases. There were 5 cases of each failed and successful landings from drone ships in the same period.

Section 4

Launch Sites Proximities Analysis

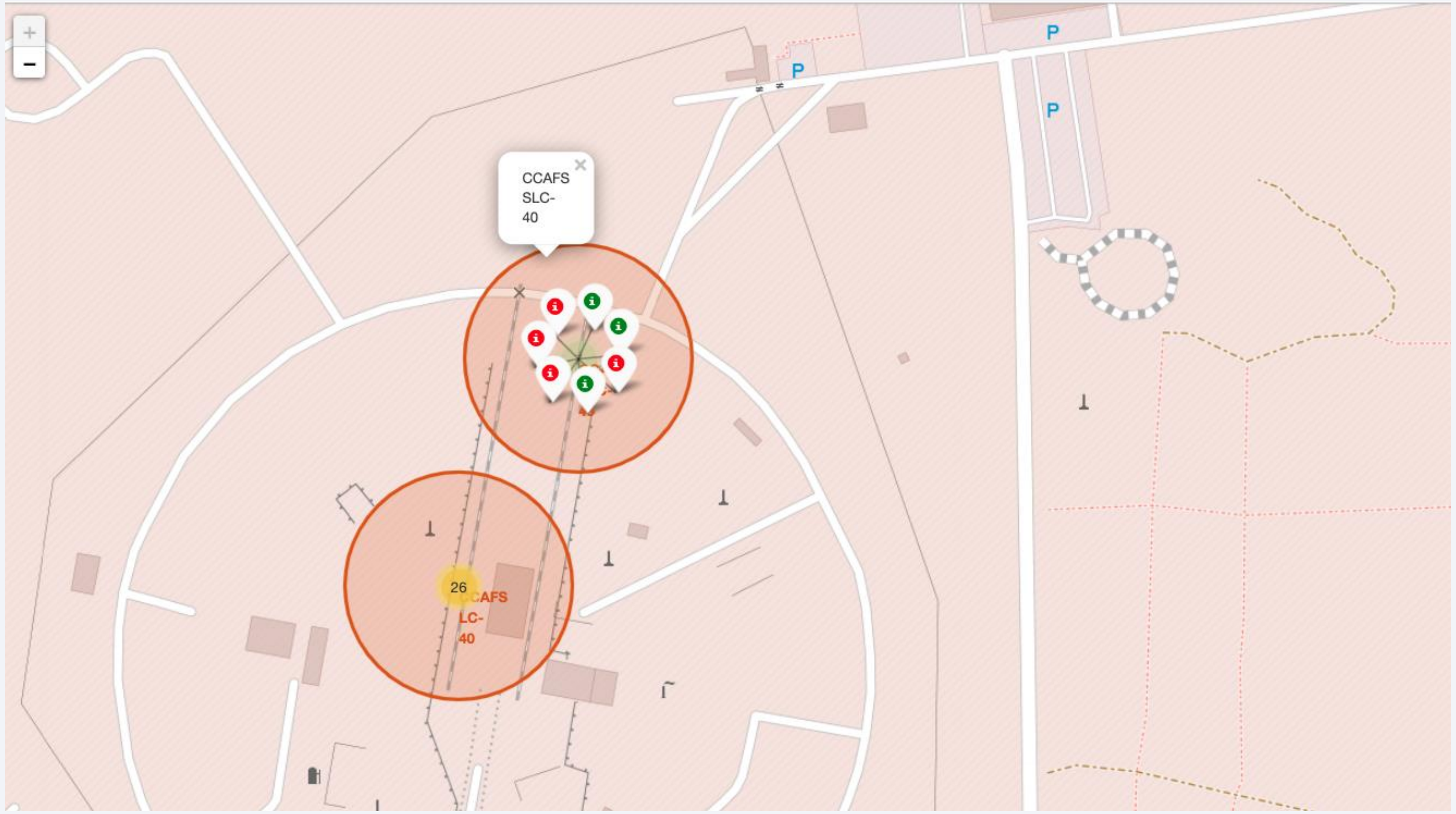
Launch sites



This map displays launch sites:

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

Launch outcomes per sites

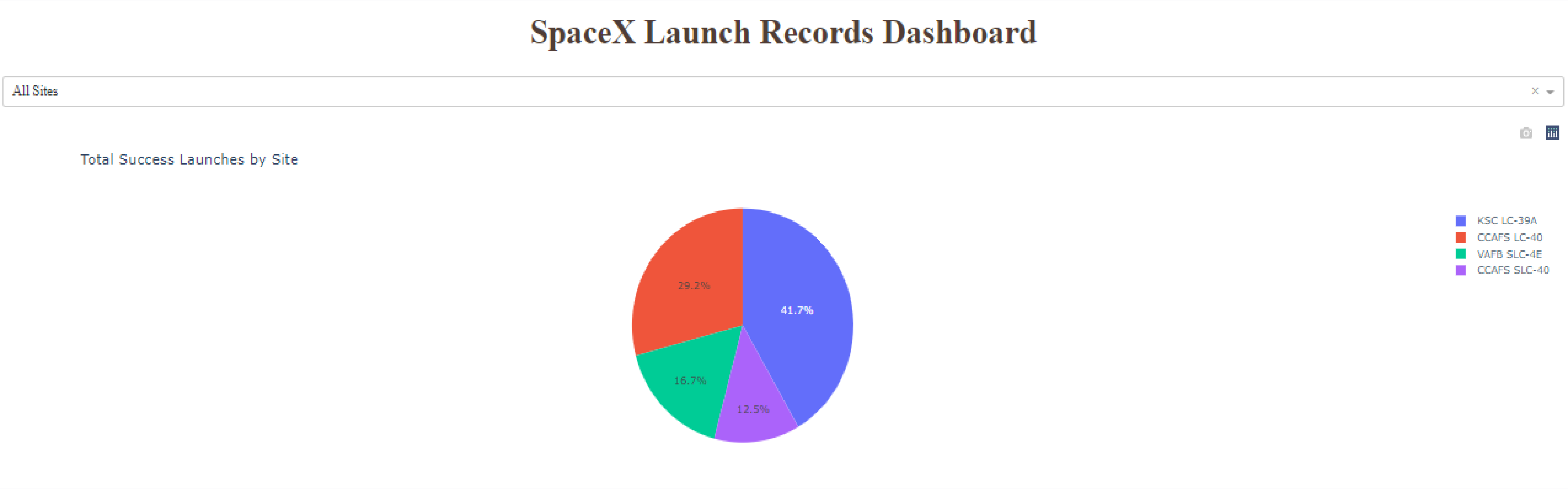


Here, we can easily identify which launch sites have relatively high success rates based on the color-labeled markers.

Section 5

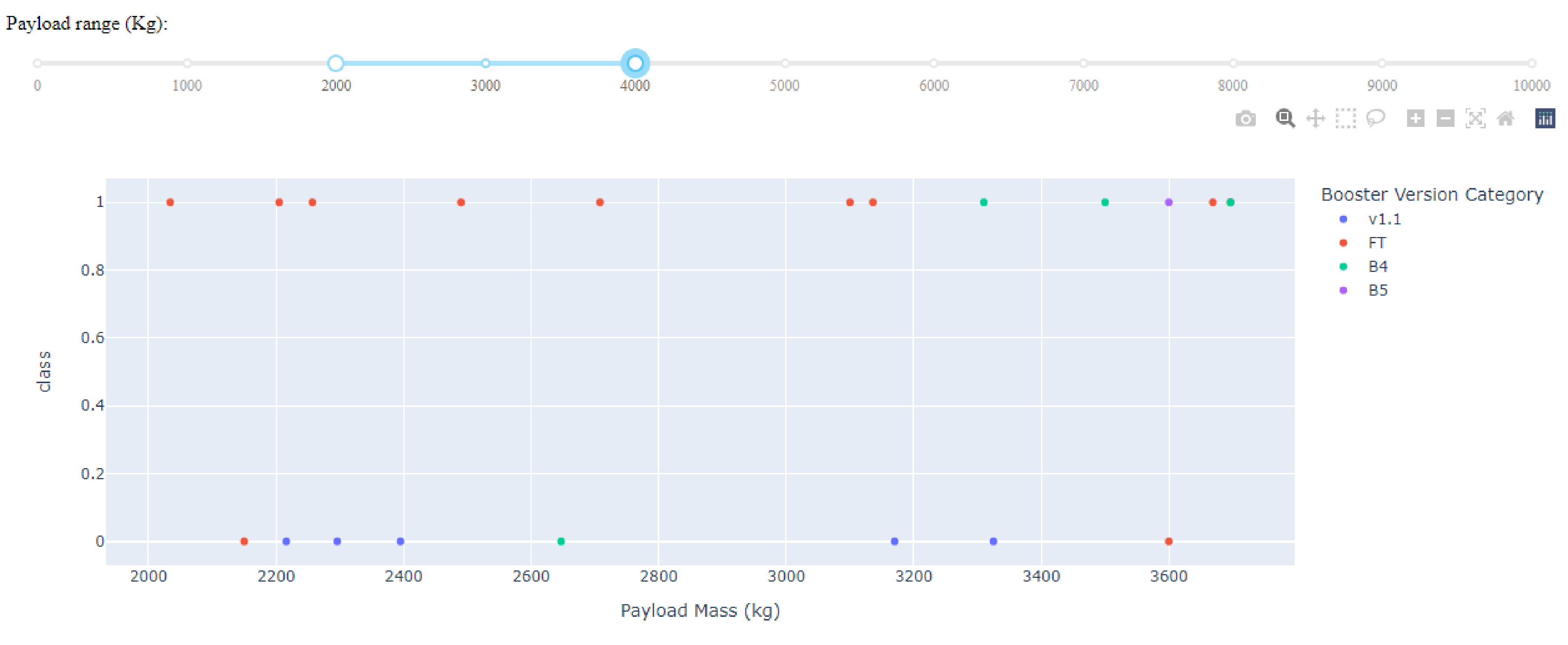
Build a Dashboard with Plotly Dash

Success launch by sites



This pie chart shows the distribution of total success launches by site. We can observe that almost 42% successful launches were from KSC LC-39A. CCAFS LC-40 is the second site with highest number successful launches with 30%.

Payload mass

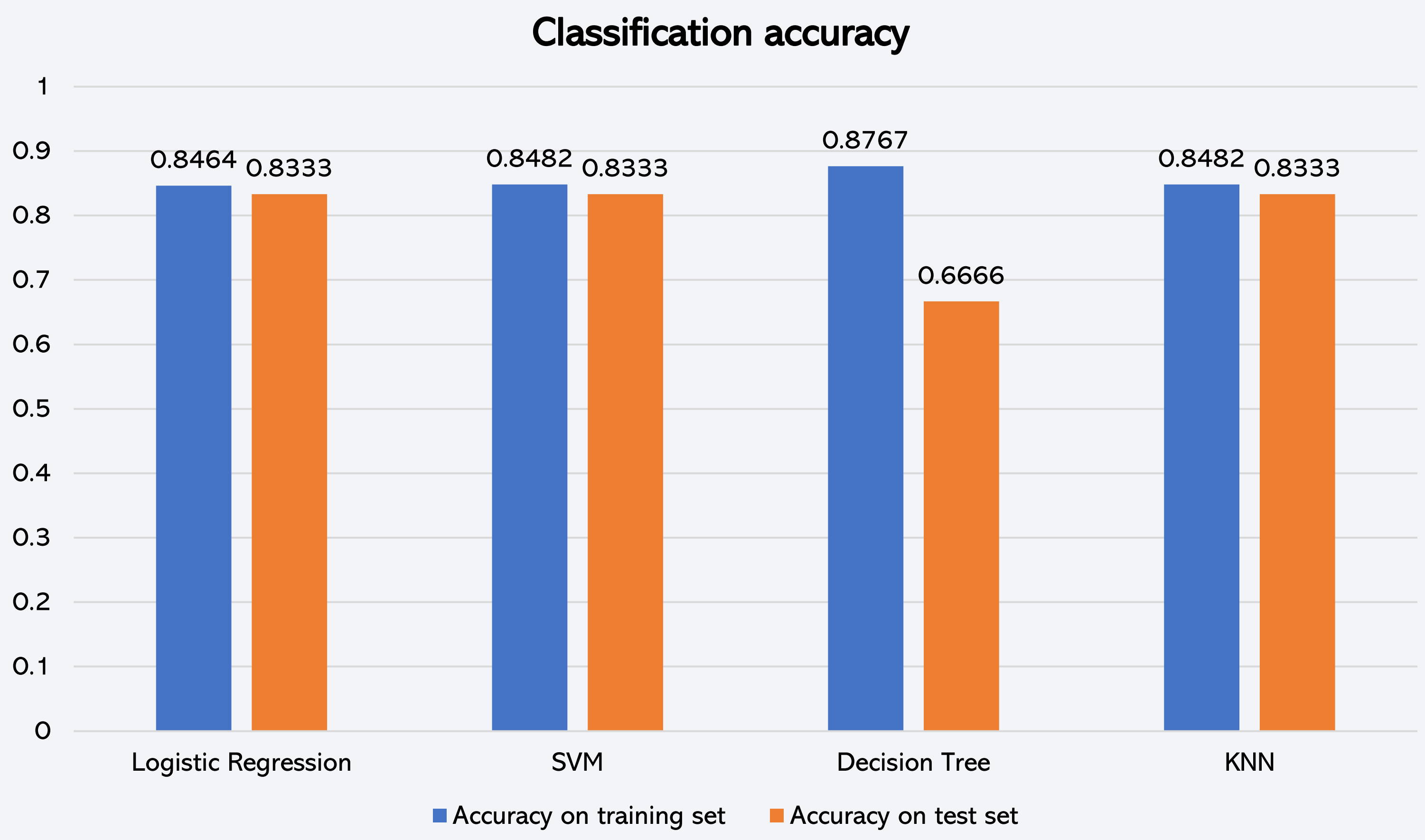


Here, we have selected payload range of 2000 to 4000 kgs. There seems to be no specific pattern of successful launches depending on the payload mass.

Section 6

Predictive Analysis (Classification)

Classification Accuracy



Most of the models indicate similar results on provided training and test data sets. For purposes of further reference, we will look at confusion matrix of KNN model, which has the highest accuracy on both training and test data sets.

Confusion Matrix

Confusion matrix of K-Nearest Neighbors model above shows the results on test data set. Here we can observe that the model predicts 3 out of 6 cases where classification label is “not landed”. For “landed” cases – the model has predicted 15 cases, 3 of which were mis-classified.

Conclusions

- CCAFS SLC 40 has a greater number of launches compared to other sites, however success rate is 60%
- GTO is also one of the orbit types with lowest success rate.
- Looking into the yearly trend of successful launches, we can observe that the success rate kept increasing since 2013
- Almost 42% successful launches were from KSC LC-39A
- Most of the models indicate similar results on provided data, with KNN model having slight edge on both training and test data sets.

Appendix

- Data Collection API notebook: Link
- Data Collection with Web Scraping notebook: Link
- EDA with Data Visualization notebook: Link
- EDA with SQL notebook: Link
- EDA notebook: Link
- Interactive Visual Analytics with Folium notebook: Link
- Build a Dashboard with Plotly Dash notebook: Link
- Machine Learning Prediction notebook: Link
- CSV file:

Spacex CSV file

Thank you!