IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Abdiel Willyar Goni
December 8 ,2021

Outline

Executive Summary

- **Summary of methodologies:**

- Data collection

- Data wrangling

- EDA with data visualization

- EDA with SQL

- Building an interactive map with Folium

- Building a Dashboard with Plotly Dash

- Predictive analysis (Classification)

- **Summary of all results:**

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

Introduction

- **Project background and context**

We predicted if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

- **Problem to be solved**

- What influences if the rocket will land successfully?

- The effect each relationship with certain rocket variables will impact in determining the success rate of a successful landing.

- What conditions does SpaceX have to achieve to get the best results and ensure the best rocket success landing rate.

Section 1

Methodology

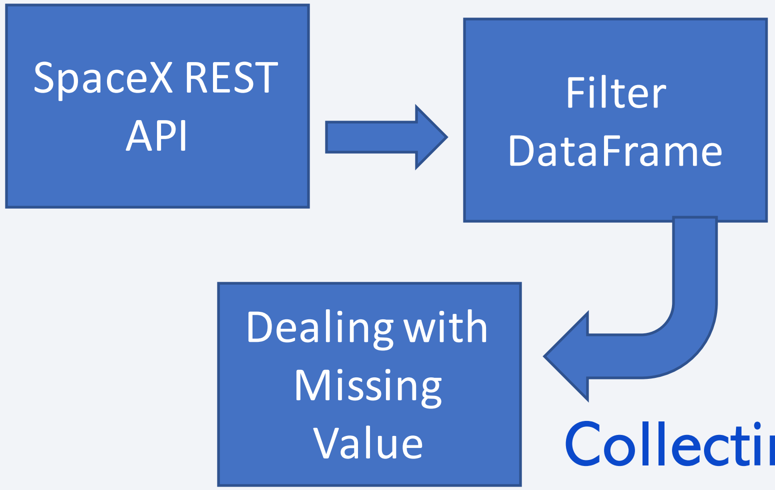
Methodology

Executive Summary

- Data collection methodology:
  - SpaceX Rest API
  - (Web Scrapping) from Wikipedia
- Perform data wrangling
  - One Hot Encoding data fields for Machine Learning and dropping irrelevant columns
- 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

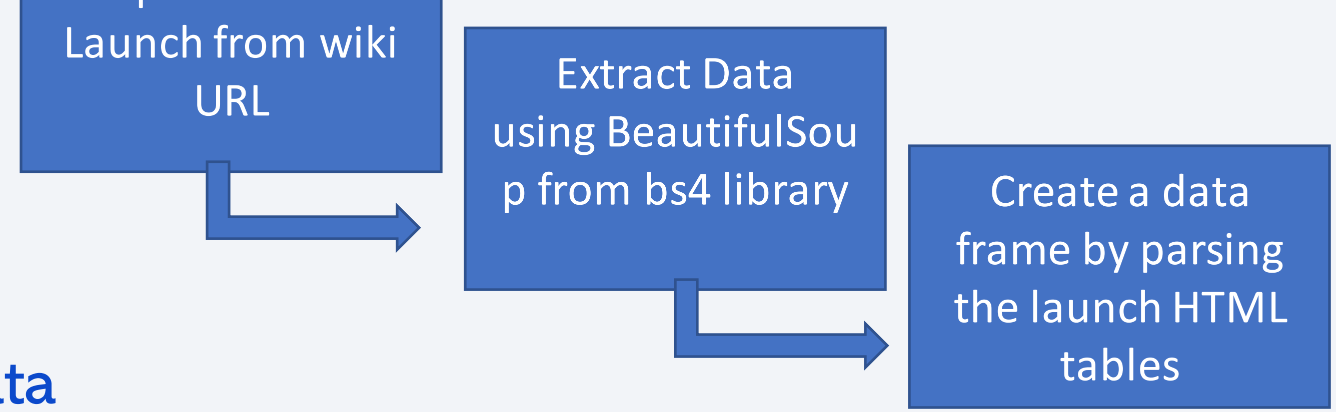
Data Collection

- The following datasets was collected by:
  - Worked with SpaceX launch data that is gathered from the SpaceX REST API.
  - API will give data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.
  - The goal is to use this data to predict whether SpaceX will attempt to land a rocket or not.
  - Web scraping Wikipedia using *BeautifulSoup*.



Collecting Data

Data Wrangling



Data Collection – SpaceX API

Collecting Data

SpaceX REST API

Filter DataFrame

Dealing with Missing Value

API calls

Decode the response content (.json)

Turn it into a Pandas dataframe (.json_normalize)

Filter Data Frame (using BoosterVersion)

Replace the np.nan values with its mean value

Replace the np.nan values with its mean value for the `PayloadMass`

Githublink

Githublink

Data Collection - Scraping

Collecting Data

Request Falcon9 Launch from wiki URL → Extract Data using BeautifulSoup from bs4 library → Create a data frame by parsing the launch HTML tables

Web Scrapping

Use requests.get() method with the provided static_url → Create a BeautifulSoup object from a response text content → find_all function in the BeautifulSoup object → Fill up the `launch_dict` with launch records extracted from table rows → Create a dataframe

Data Wrangling

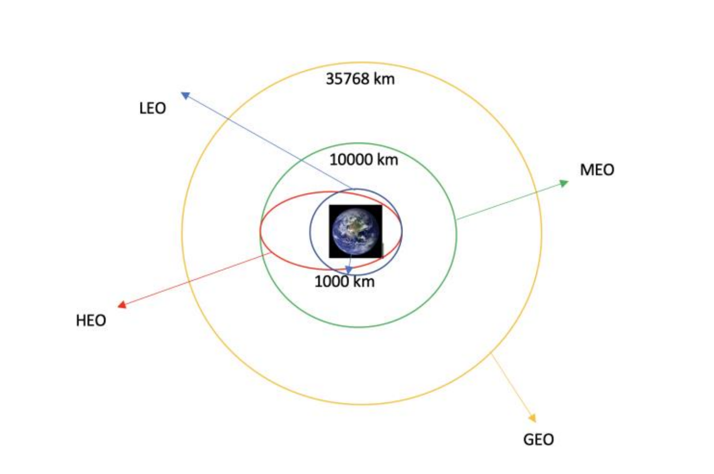
In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground pad.True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship.

Perform exploratory Data Analysis and determine Training Labels

1. **Calculate the number of launches on each site**
2. **Calculate the number and occurrence of each orbit**
3. **Calculate the number and occurence of mission outcome per orbit type**
4. **Create a landing outcome label from Outcome column**

[Githublink]()

Each launch aims to an dedicated orbit, and here are some common orbit types



EDA with Data Visualization

1. Scatter Graphs being drawn:

   - Flight Number VS. Payload Mass
   - Flight Number VS. Launch Site
   - Payload VS. Launch Site
   - Orbit VS. Flight Number
   - Payload VS. Orbit Type
   - Orbit VS. Payload Mass

2. Bar Graph being drawn:

   Mean VS. Orbit

3. Line Graph being drawn:

   Success Rate VS. Year

Githublink

EDA with SQL

1. Understand the Spacex DataSet
2. Load the dataset into the corresponding table in a Db2 database
3. Execute SQL queries to answer assignment questions

SpaceX has gained worldwide attention for a series of historic milestones.
It is the only private company ever to return a spacecraft from low-earth orbit, which it first accomplished in December 2010. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars wheras other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage.
Therefore if we can determine if the first stage will land, we can determine the cost of a launch.
This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.
This dataset includes a record for each payload carried during a SpaceX mission into outer space.

Build an Interactive Map with Folium

The launch success rate may depend on many factors such as payload mass, orbit type, and so on. It may also depend on the location and proximities of a launch site, i.e., the initial position of rocket trajectories. Finding an optimal location for building a launch site certainly involves many factors and hopefully we could discover some of the factors by analyzing the existing launch site locations.

Click to add text

To visualize the Launch Data into an interactive map. We took the Latitude and Longitude Coordinates at each launch site and added a Circle Marker around each launch site with a label of the name of the launch site.

1: Mark all launch sites on a map
2: Mark the success/failed launches for each site on the map
3: Calculate the distances between a launch site to its proximities

Using Haversine's formula we calculated the distance from the Launch Site to various landmarks to find various trends about what is around the Launch Site to measure patterns. L

Githublink

Build a Dashboard with Plotly Dash

Used Python Anywhere to host the website live 24/7, can play around with the data and view the data

Scatter Graph showing the relationship with Outcome and Payload Mass (Kg) for the different Booster Versions

- Shows the relationship between two variables.
- Scatter Graph is the best method to show you a non-linear pattern.

Githublink

Predictive Analysis (Classification)

BUILDING MODEL

- Load our dataset into NumPy and Pandas
- Transform Data
- Split our data into training and test data sets
- Check how many test samples we have
- Decide which type of machine learning algorithms we want to use
- Set our parameters and algorithms to GridSearchCV
- Fit our datasets into the GridSearchCV objects and train our dataset

EVALUATING MODEL

Check accuracy for each model, Tuned hyperparameters for each type of algorithms ,Confusion Matrix

IMPROVING MODEL

Feature Engineering, Algorithm Tuning

FINDING THE BEST PERFORMING CLASSIFICATION MODEL

Githublink

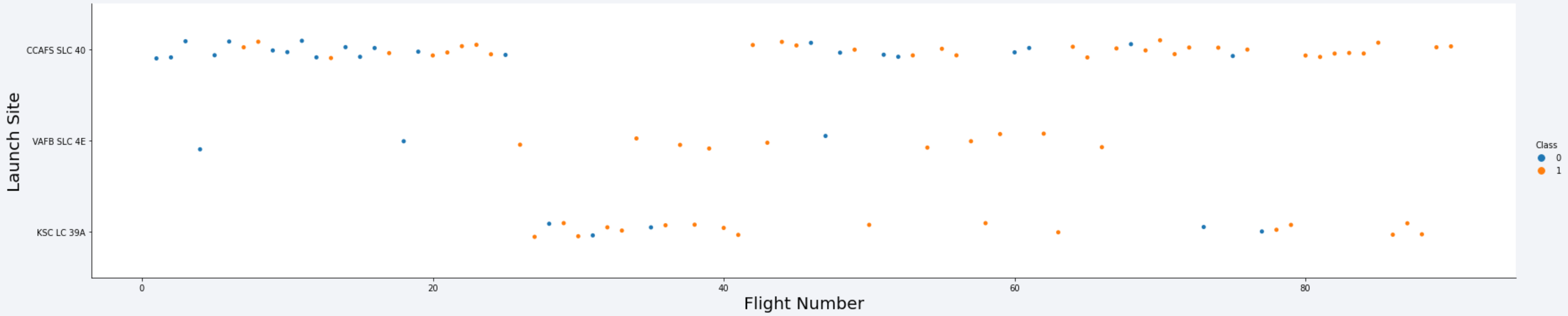
Results

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

Section 2

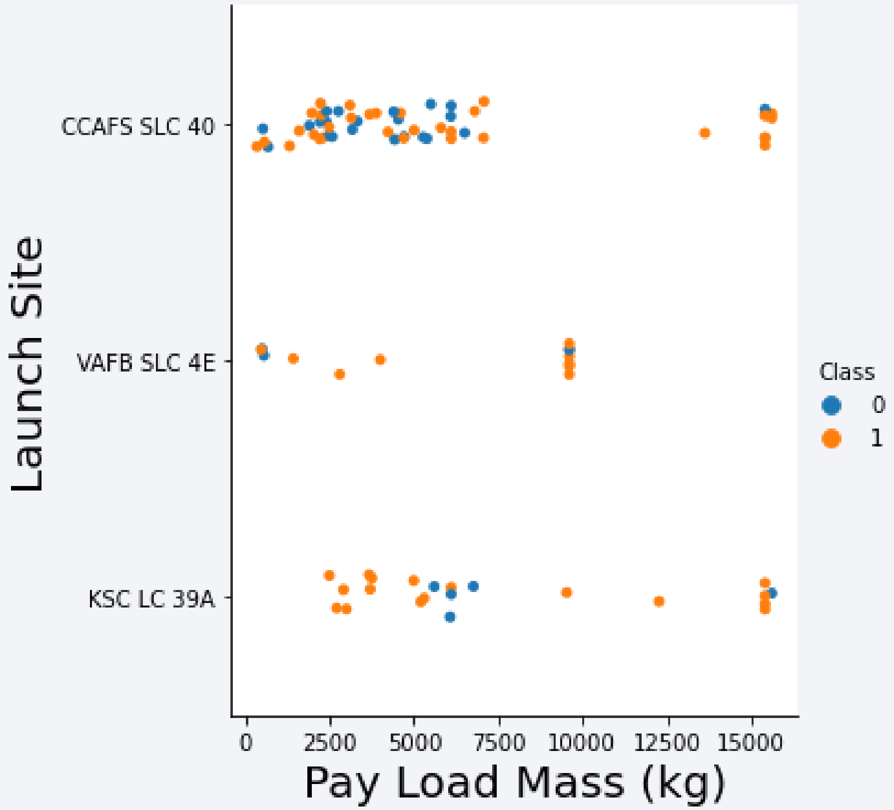
Insights drawn from EDA

Flight Number vs. Launch Site



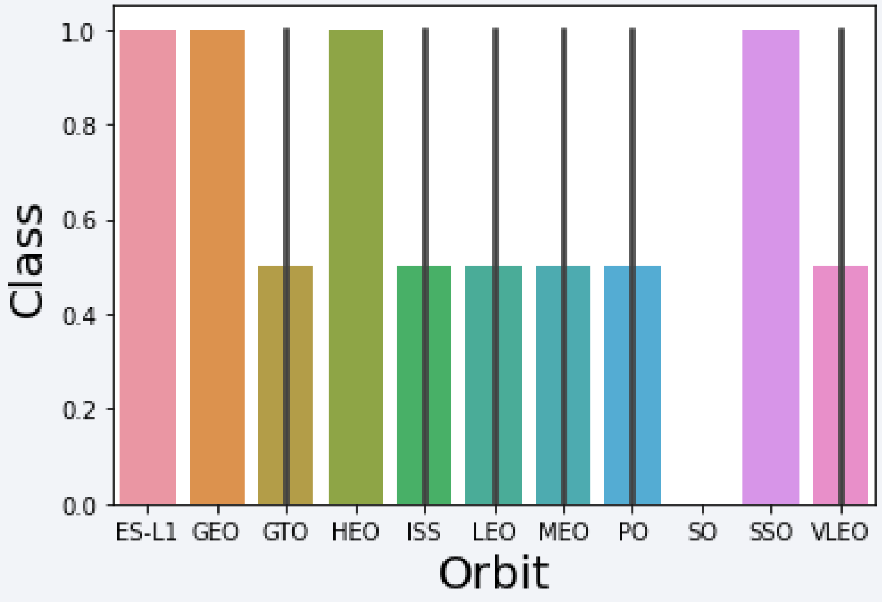
The more the number of flights at the launch site, the higher the success rate at the launch site.

Payload vs. Launch Site



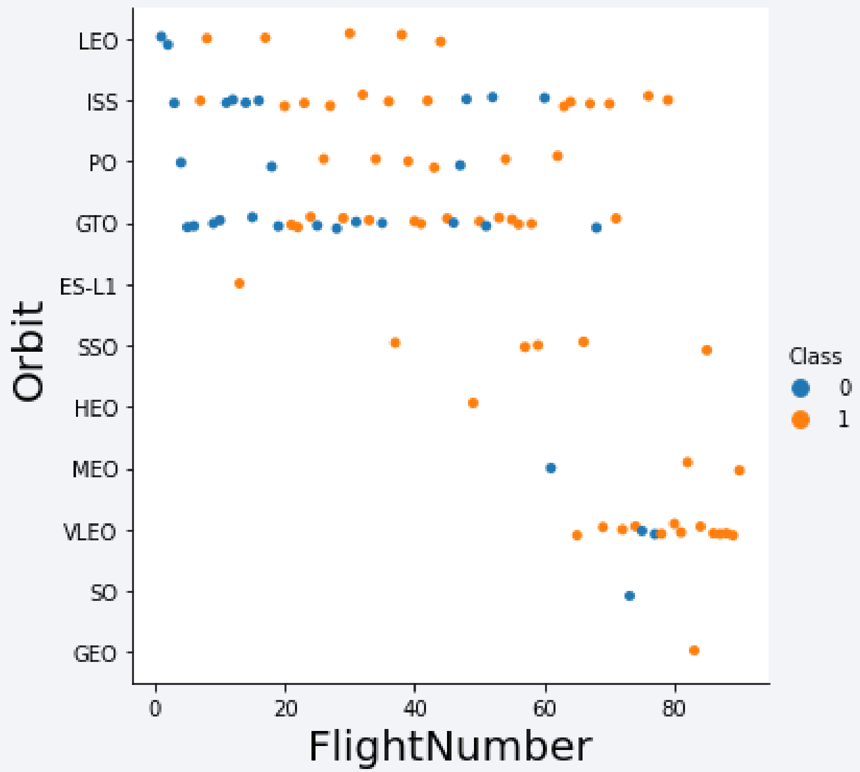
The greater the payload mass for the CCAFS SLC 40 Launch Site, the higher the Rocket success rate.

Success Rate vs. Orbit Type



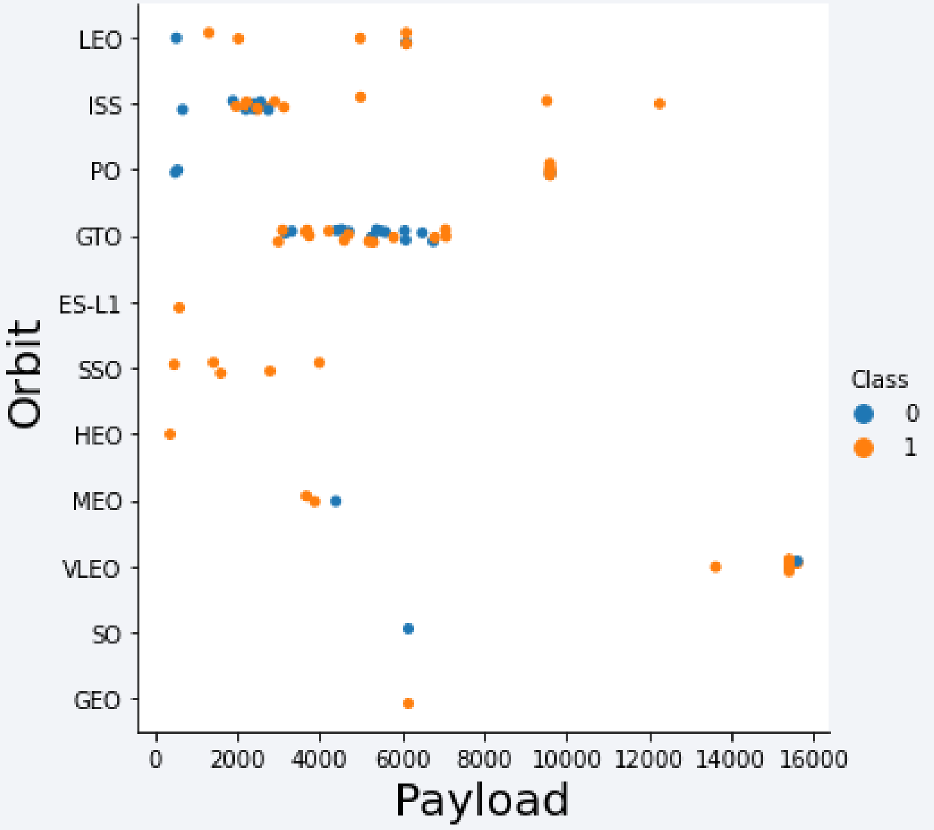
Orbit ES-L1, GEO,HEO,SSO, has the best Success Rate

Flight Number vs. Orbit Type



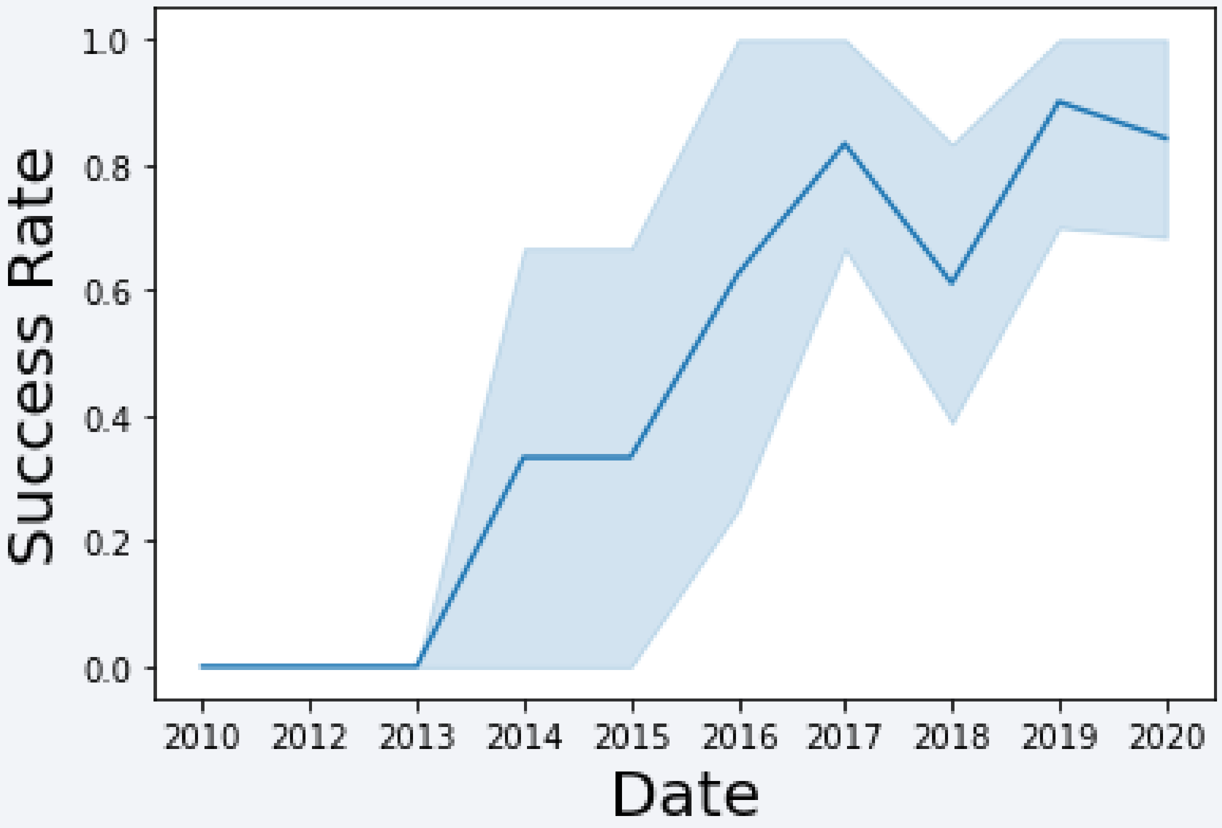
Seems to be no relationship between flight number
when in GTO orbit.

Payload vs. Orbit Type



Heavy
payloads have a negative
influence
on GTO orbits and positive
on GTO
and Polar LEO (ISS) orbits.

Launch Success Yearly Trend



The success rate fluctuates. But it has increased since 2013.

All Launch Site Names

```
%sql select distinct(launch_site) from SPACEX
```

Done.

Out[57]:

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

Using the DISTINCT syntax on the SPACEX database to display all launch_sites

Launch Site Names Begin with 'CCA'

Done.

Out[56]:

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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)
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-12	22:41:00	F9 v1.1	CCAFS LC-40	SES-8	3170	GTO	SES	Success	No attempt

```
%sql select * from SPACEX where launch_site like 'CCA%' limit 5
```

Select all the columns from the SPACEX table which is based on launch_site and takes only 5 records

Total Payload Mass

Done.

Out[55]:

total_payload_mass
22007

%sql select sum(payload_mass__kg_) as total_payload_mass from SPACEX where customer = 'NASA (CRS)'

select the payload mass column then add it up using the SUM syntax and display it with the total payload mass column name based on the customer name 'NASA (CRS'

Average Payload Mass by F9 v1.1

%sql select avg(payload_mass__kg_) as payload_mass__kg_ from SPACEX where booster_version = 'F9 v1.1'

Done.

Out[54]:

payload_mass__kg_
3676

average the payload_mas__kg_ column with AVG syntax based on booster_version = 'F9 v1.1'

First Successful Ground Landing Date

```
%sql select min(DATE) as DATE from SPACEX where Landing__Outcome = 'Success (ground pad)'
```

Done.

Out[50]:

DATE
2017-01-05

use the MIN syntax in the DATE column to see the first successful ground from the landing_outcomes column

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select date, booster_version, payload, payload_mass__kg_ from SPACEX where payload_mass__kg_
between 4000 and 6000  and Landing__Outcome LIKE 'Success (drone ship)'
```

Done.

Out[49]:

DATE	booster_version	payload	payload_mass__kg_
2016-06-05	F9 FT B1022	JCSAT-14	4696
2017-11-10	F9 FT B1031.2	SES-11 / EchoStar 105	5200

select multiple columns from the SPACEX table based on the payload mass column with the given range

Total Number of Successful and Failure Mission Outcomes

```
%sql select count(*), mission_outcome from SPACEX group by mission_outcome
```

Done.

Out[58]:

1	mission_outcome
44	Success
1	Success (payload status unclear)

Count all mission outcomes with COUNT sin syntax

Boosters Carried Maximum Payload

```
%sql select * from SPACEX where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEX)
```

Done.

Out[59]:

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2019-11-11	14:56:00	F9 B5 B1048.4	CCAFS SLC-40	Starlink 1 v1.0, SpaceX CRS-19	15600	LEO	SpaceX	Success	Success
2020-07-01	02:33:00	F9 B5 B1049.4	CCAFS SLC-40	Starlink 2 v1.0, Crew Dragon in-flight abort test	15600	LEO	SpaceX	Success	Success
2020-04-06	01:25:00	F9 B5 B1049.5	CCAFS SLC-40	Starlink 7 v1.0, Starlink 8 v1.0	15600	LEO	SpaceX, Planet Labs	Success	Success
2020-03-09	12:46:14	F9 B5 B1060.2	KSC LC-39A	Starlink 11 v1.0, Starlink 12 v1.0	15600	LEO	SpaceX	Success	Success
2020-06-10	11:29:34	F9 B5 B1058.3	KSC LC-39A	Starlink 12 v1.0, Starlink 13 v1.0	15600	LEO	SpaceX	Success	Success

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

2015 Launch Records

```
%sql select monthname(DATE) as month, landing__Outcome from SPACEX where DATE LIKE '2015%' and Landing__Outcome LIKE 'Failure%'
```

Done.

Out[64]:

MONTH	landing__outcome
October	Failure (drone ship)

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

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

```
%sql select * from SPACEX where Landing__Outcome like 'Success%' and (DATE between '2010-06-04' and '2017-03-20') order by date desc
```

Done.

Out[65]:

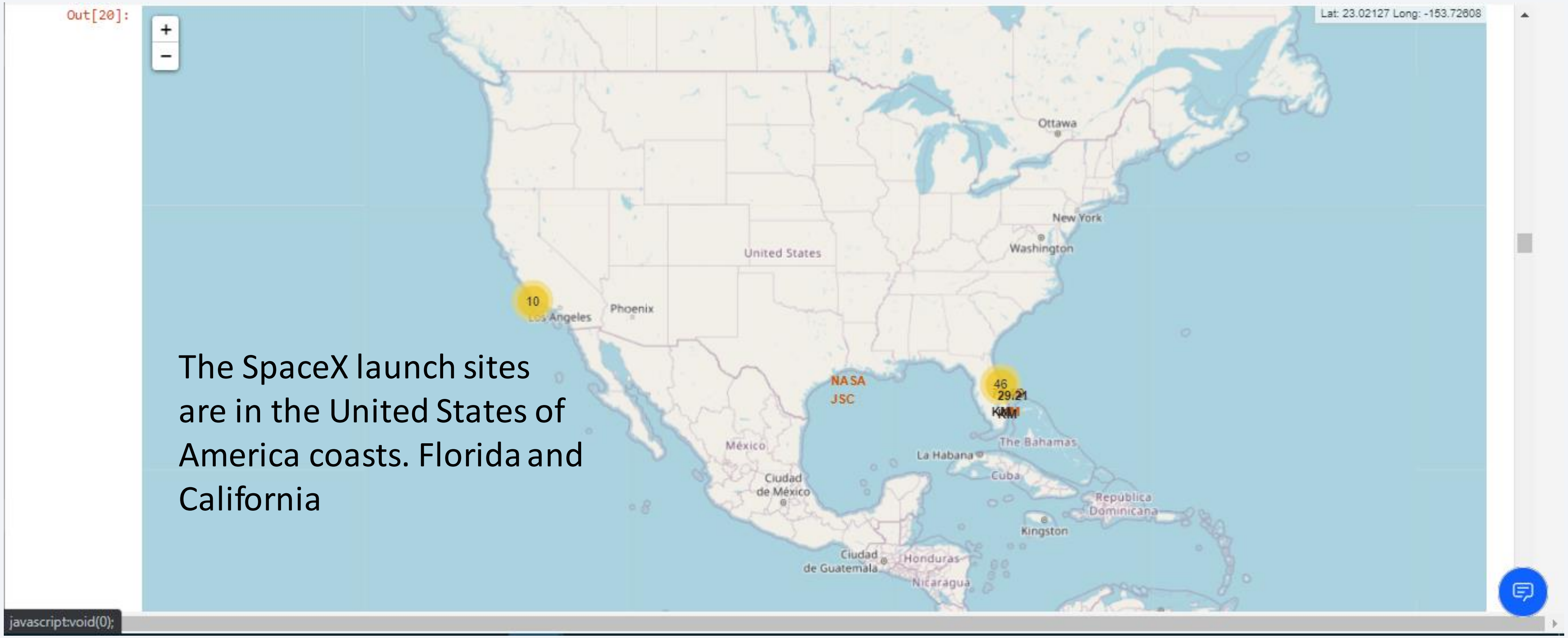
DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2017-03-06	21:07:00	F9 FT B1035.1	KSC LC-39A	SpaceX CRS-11	2708	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-01-05	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
2016-08-04	20:43:00	F9 FT B1021.1	CCAFS LC-40	SpaceX CRS-8	3136	LEO (ISS)	NASA (CRS)	Success	Success (drone ship)
2016-06-05	05:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)

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

Section 4

Launch Sites Proximities Analysis

All Launch Sites Global Map Markers



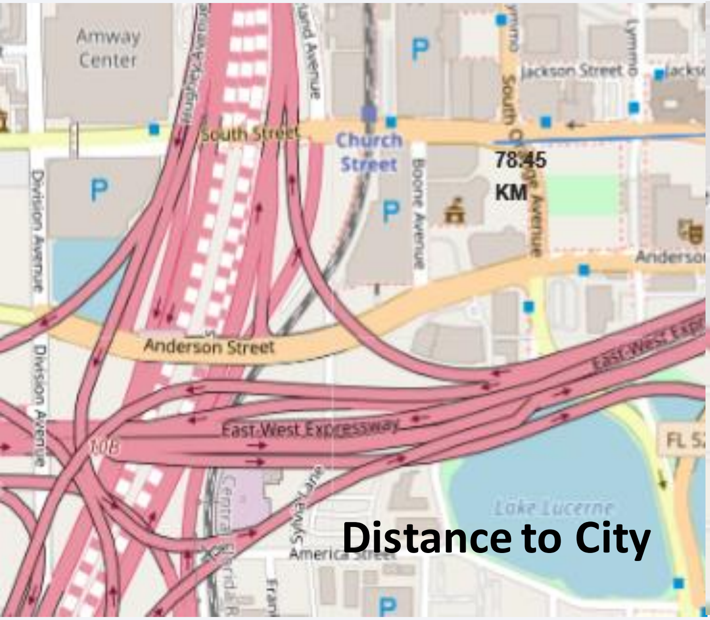
The SpaceX launch sites are in the United States of America coasts. Florida and California

Color Labeled

From the color-labeled markers in marker clusters, you should be able to easily identify which launch sites have relatively high success rates.

Green Marker shows successful Launches and Red Marker shows Failures

Selected Launch Site



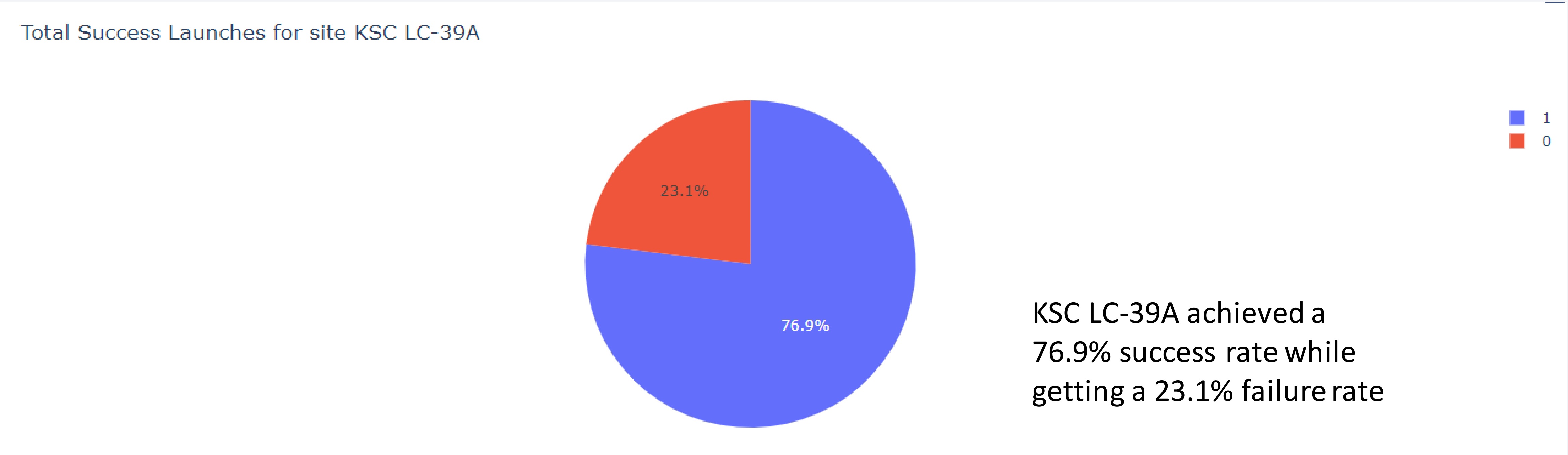
Section 5

Build a Dashboard with Plotly Dash

Launch Success

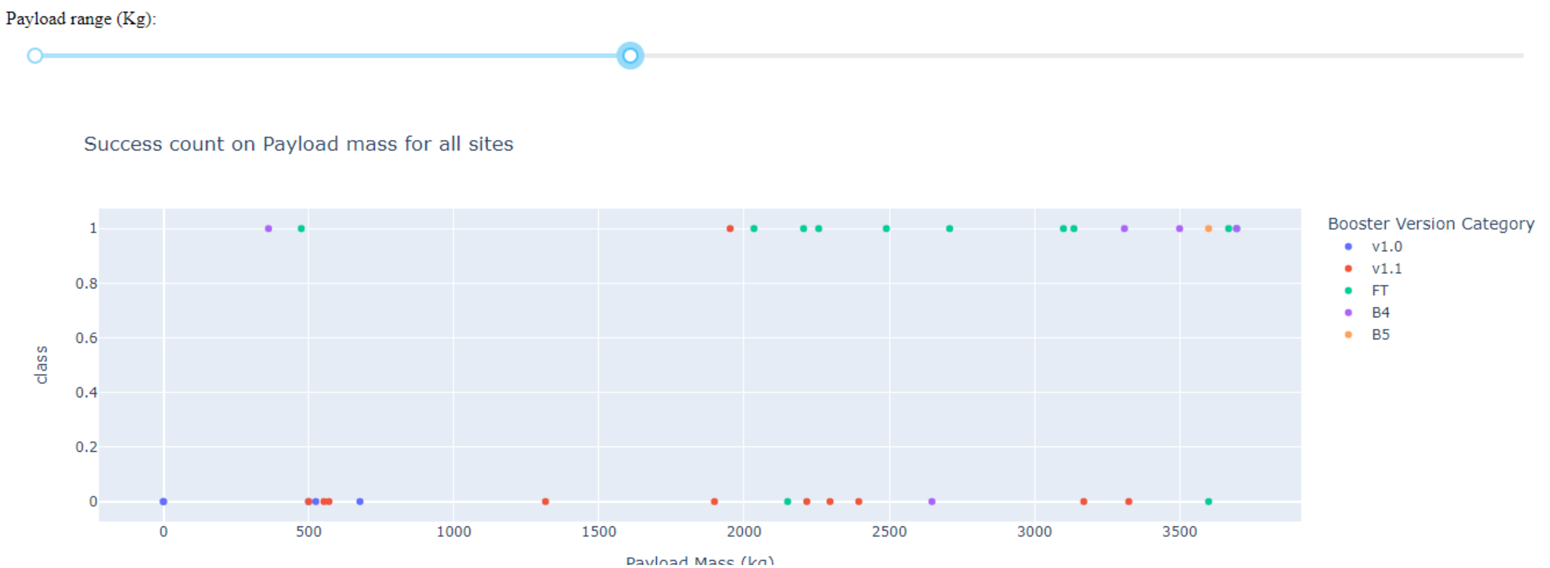
KSC LC-39A had the most successful launches from all the sites

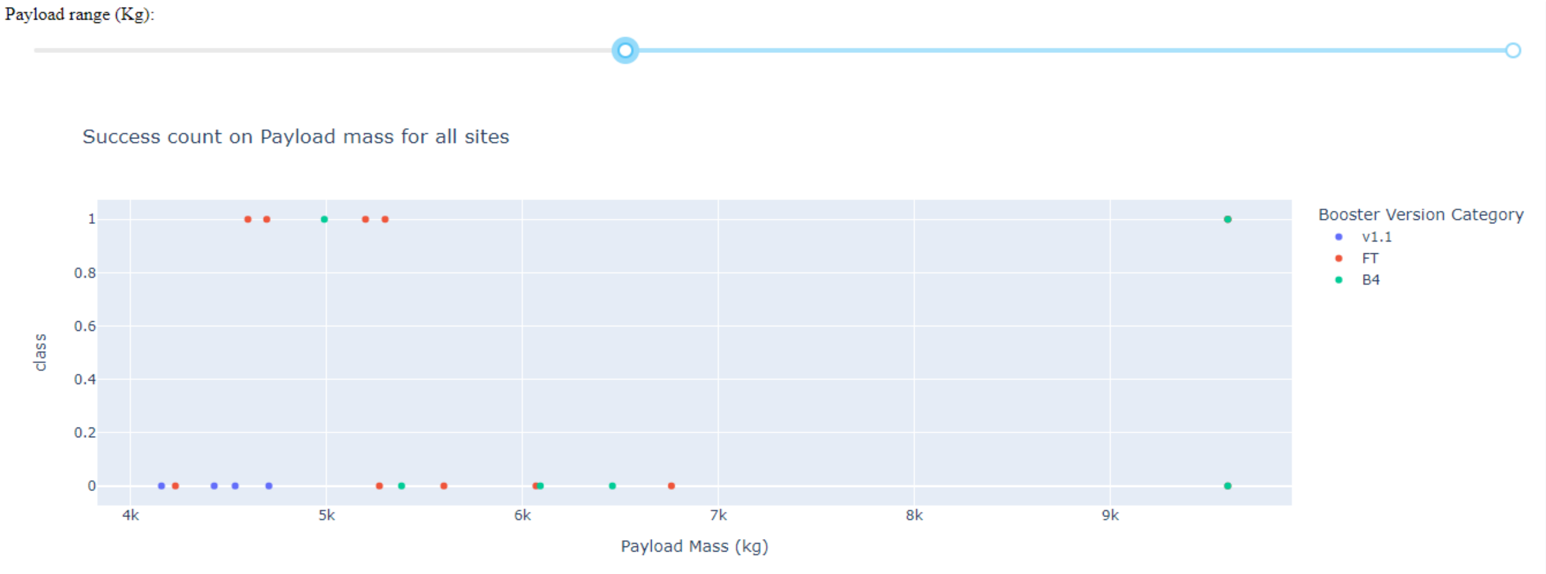
Highest Launch Succes Ratio



KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

Payload vs Launch Outcome





The success rates for low weighted payloads is higher than the heavy weighted payloads

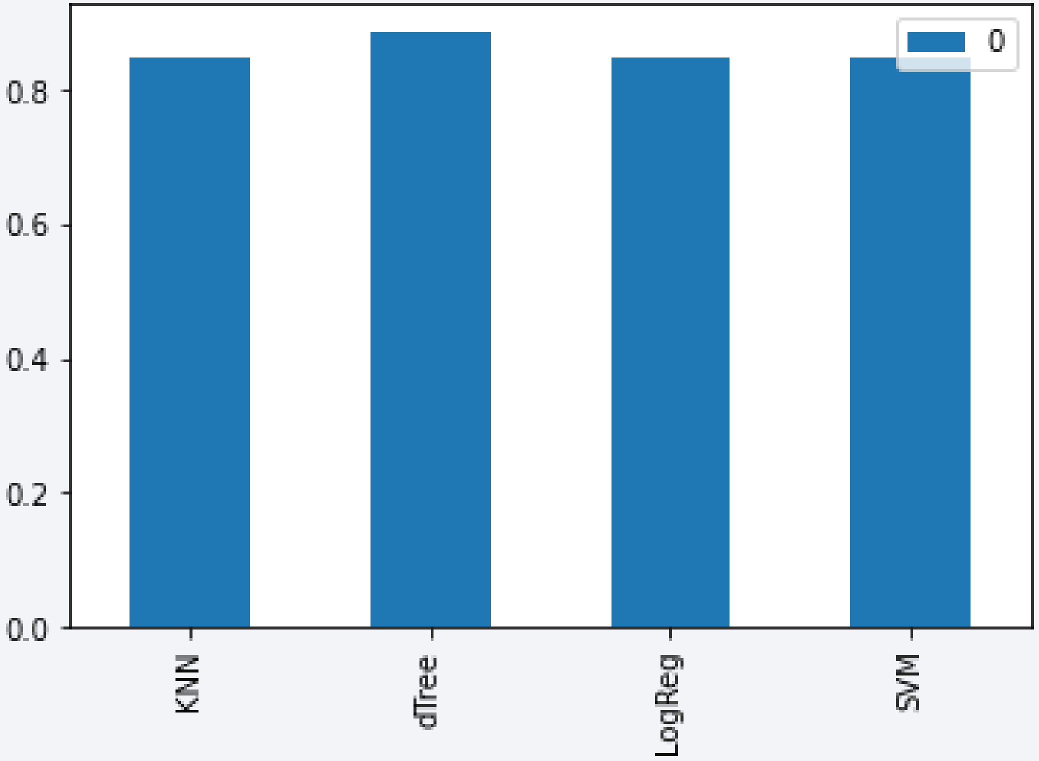
Section 6

Predictive Analysis (Classification)

Classification Accuracy

KNN	0.848214
dTree	0.885714
LogReg	0.846429
SVM	0.848214

Best Algorithm is Tree with a score of 0.8857142857142858



Confusion Matrix

Confusion Matrix for the Decision Tree. Major problem is false positives.

Conclusions

- Orbit GEO,HEO,SSO,ES-L1 has the best Success Rate
- Low weighted payloads perform better than the heavier payloads
- KSC LC-39A had the most successful launches from all the sites
- The Tree Classifier Algorithm is the best for Machine Learning for this dataset
- Success rates for SpaceX launches is directly proportional time in years they will eventually perfect the launches

Thank you!