Commercial Space Race

Data Science Capstone Project

Usman M. Umer

May, 2022

Outline

- Executive Summary
- Introduction
- Data
- Methodology
- Results
- Conclusions

Executive Summary

- Summary of methodologies
  - Data Collection through API and Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis(EDA) using Visualization and SQL
  - Interactive Visual Analytics with Folium and Dashboards
  - Machine Learning Prediction
- Summary of results
  - Exploratory Data Analysis(EDA) using Visualization and SQL result
  - Interactive Visual Analytics with Folium and Dashboards results
  - Machine Learning Predictive Analytics results

Introduction

- Numerous companies are working to make space travel affordable for everyone.
- SpaceX advertises Falcon 9 rocket launches $62 million; while others $165 million each.
- If we can successfully determine the successful landing of first stage Falcon 9, we can determine the cost of a launch.

Spacecraft and Starlink

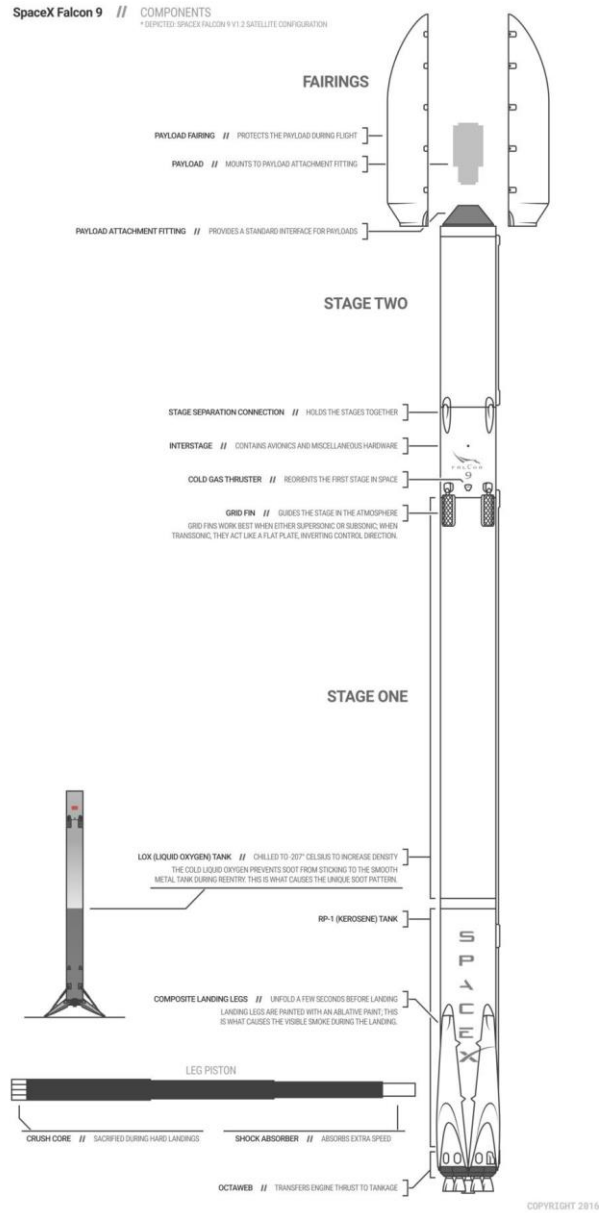
Rockets and launch service

Orbital and sub-orbital reusable rockets

Suborbital space flights

Introduction

- The first stage of Falcon 9 rocket is a large and the most expensive component. Sometimes it does not land, it will crash or sacrifices due to the mission parameters like payload, orbit, and customer.
- Unlike other rocket providers, if SpaceX recover and reuse the first stage of Falcon 9 rocket it drastically reduced manufacturing costs.
- The main objective of this project is to predict if the first stage of SpaceX Falcon 9 rocket will land successfully and to analyze what factors contribute for successful landing rate.
- Machine learning algorithms employed to predict the successful rate of Falcon 9 launches.



Methodology

Data Collection

Rocket launch data from
- SpaceX Reset API
- Web scraping from Wikipedia

Data wrangling

- Transforming data by dropping irrelevant column.
- Make data useful for Machine Learning and training models

Exploratory Data Analysis(EDA) using Visualization and SQL

- Visualize the relationship between parameters

Interactive Visual Analytics with Folium and Dashboards

- Data Visualization with Folium and Plotly Dash

Machine Learning Prediction

- Split dataset into training and test data for Machine Learning prediction.
- Build various models (SVM, KNN, Classification Tree and Logistic Regression) using train data.
- Evaluate the accuracy models and their best parameters using test data

Methodology

DATA COLLECTION
SpaceX API

This API will give us data about FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

Methodology

DATA COLLECTION

SpaceX API

1. Request and getting response from SpaceX API

```
spacex_url="https://api.spacexdata.com/v4/launches/past"
```

```
response = requests.get(spacex_url)
```

2. Convert the json result into a dataframe

```
# Use json_normalize meethod to convert the json result into a dataframe
data = pd.json_normalize(response.json())
```

3. Apply various functions to get more information

```
# Call getLaunchSite
getLaunchSite(data)
```

```
# Call getBoosterVersion
getBoosterVersion(data)
```

```
# Call getPayloadData
getPayloadData(data)
```

```
# Call getCoreData
getCoreData(data)
```

4. construct our dataset using the data

```
launch_dict = {'FlightNumber': list(data['flight_number']),
'Date': list(data['date']),
'BoosterVersion':BoosterVersion,
'PayloadMass':PayloadMass,
'Orbit':Orbit,
'LaunchSite':LaunchSite,
'Outcome':Outcome,
'Flights':Flights,
'GridFins':GridFins,
'Reused':Reused,
'Legs':Legs,
'LandingPad':LandingPad,
'Block':Block,
'ReusedCount':ReusedCount,
'Serial':Serial,
'Longitude': Longitude,
'Latitude': Latitude}
```

5. create data frame from the dictionary launch_dict.

```
# Create a data from launch_dict
launch_df = pd.DataFrame.from_dict(launch_dict)
```

6. Filter the dataframe to only include Falcon 9 launches.

```
# Hint data['BoosterVersion']!='Falcon 1'
data_falcon9 = launch_df[launch_df['BoosterVersion'] == 'Falcon 9']
```

Methodology

DATA COLLECTION

Web Scraping- Wikipedia

1. Request the Falcon9 Launch Wiki page from its URL

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

```
html_data = requests.get(static_url).text
```

2. Create a BeautifulSoup object from the HTML response

```
# Use BeautifulSoup() to create a BeautifulSoup object from a response text content
soup = BeautifulSoup(html_data, 'html5lib')
```

3. Extract all column/variable names from the HTML table header

```
html_tables = soup.find_all('table')
```

4. Extract_column_from_header

```
column_names = []

for row in first_launch_table.find_all('th'):
    name = extract_column_from_header(row)
    if(name != None and len(name) > 0):
        column_names.append(name)
```

5. Create a data frame by parsing the launch HTML tables

```
launch_dict= dict.fromkeys(column_names)

# Remove an irrelvant column
del launch_dict['Date and time ( )']

# Let's initial the launch_dict with each value to be an empty list
launch_dict['Flight No.'] = []
launch_dict['Launch site'] = []
launch_dict['Payload'] = []
launch_dict['Payload mass'] = []
launch_dict['Orbit'] = []
launch_dict['Customer'] = []
launch_dict['Launch outcome'] = []
# Added some new columns
launch_dict['Version Booster']=[]
launch_dict['Booster landing']=[]
launch_dict['Date']=[]
launch_dict['Time']=[]
```

5. create data frame from the dictionary launch_dict.

```
df=pd.DataFrame(launch_dict)
```

Methodology

DATA WRANGLING

1. Check Missing Values

```
data_falcon9.isnull().sum()
```

2. Dealing with Missing Values

```
# Calculate the mean value of PayloadMass column
PayloadMass_mean = data_falcon9.PayloadMass.mean()
# Replace the np.nan values with its mean value
data_falcon9['PayloadMass'] = data_falcon9['PayloadMass'].replace(np.nan, PayloadMass_mean)
```

3. Check missing values of the PayLoadMass change to zero.

```
data_falcon9.isnull().sum()
```

4. Create a new column, 'class', to delineate between successful and unsuccessful recoveries

1 → successful recovery
0 → unsuccessful recovery

- **There are several cases in which the booster failed to successfully land on the dataset.**
- **According to our data frame, there are 8 different outcomes:**
  - **True ASDS → Successful landing to drone ship**
  - **True RTLS → Successful landing on a ground pad**
  - **True Ocean → Successful landing in ocean**
  - **None None → Failed to land**
  - **None ASDS → Failed to land**
  - **False ASDS → Failed to land on drone ship**
  - **False RTLS → Failed to land on ground pad**
  - **False Ocean → Failed to land in ocean**

Methodology

EDA using Visualization

- The relationship between some attributes are visualized using different plots to determine if they helps to determine the first state can be reused.

1. Scatter plot drown between
    - Flight Number and Payload
    - Flight Number and Launch Site
    - Launch Site and Payload
    - Orbit type and Flight Number
    - Orbit type and Payload
- A scatter plot shows how much one variable is affected by another. It also shows the relationship between two variables is called a correlation.

2. Bar Chart drown between
    - Success Rate and Orbit type
- Bar chart is very helpful indicate the relationship between multiple groups at a glance. One axis represents a category and the other axis represents a discrete value.

3. Line chart between
    - Success rate and Time in years
- This chart shows data variables and trends very clearly and helps predict the results of data that has not yet been recorded.

Methodology

EDA using SQL

- SQL allows us to make complicated queries to collect relevant information about the dataset
- Some SQL queries are used to get information from our dataset :
  - Display the names of the unique launch sites in the space mission.
  - Displaying 5 records where launch sites begin with the string 'KSC’
  - Displaying the total payload mass carried by boosters launched by NASA (CRS)
  - Displaying average payload mass carried by booster version F9 v1.1
  - Listing the date where the successful landing outcome in drone ship was achieved.
  - Listing the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000
  - Listing the total number of successful and failure mission outcomes
  - Listing the names of the booster_versions which have carried the maximum payload mass.
  - Listing the records which will display the month names, successful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017
  - Ranking the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

Methodology

Interactive Visual Analytics 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.
- Interactive visual analytics are performed.
  - All launch sites on a map are marked
  - The success/failed launches for each site on the map assigned to class 0 with Green and 1 Red markers on the map; i.e. 0 for failure, and 1 for success.
  - The distances between a launch site to its proximities calculated and answered the following questions
    - Are launch sites in close proximity to railways? Yes
    - Are launch sites in close proximity to highways? Yes
    - Are launch sites in close proximity to coastline? Yes
    - Do launch sites keep certain distance away from cities? Yes

Methodology

Build an Interactive Dashboard with Ploty Dash

- The dashboard application contains a pie chart and a scatter point chart.
- Pie chart
  - Shows the total launches by a certain site or all launch sites
  - Shows total success launches by sites
  - Indicate a successful landing distribution across all launch sites or to indicate the success rate of individual launch sites.
- Scatter chart
  - Shows the relationship between Outcomes and Payload mass(Kg) for the different boosters version.
  - All sites/individual site & Payload mass on a slider between 0 and 10000 kg
  - It helps to visualize how launch success depends on the launch point, payload mass, and booster version categories

Methodology

Machine Learning Prediction

- **Dataset split into training and test data for Machine Learning prediction.**
- **Various models (SVM, KNN, Classification Tree and Logistic Regression) build using train data.**
- **The accuracy of models evaluated using test data.**

Model Building

- 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.

Model Evaluation

- Check accuracy for each model
- Get tuned hyperparameters for each type of algorithms
- Plot Confusion Matrix

Model Improvement

- Improved the model using feature engineering and algorithm tuning.

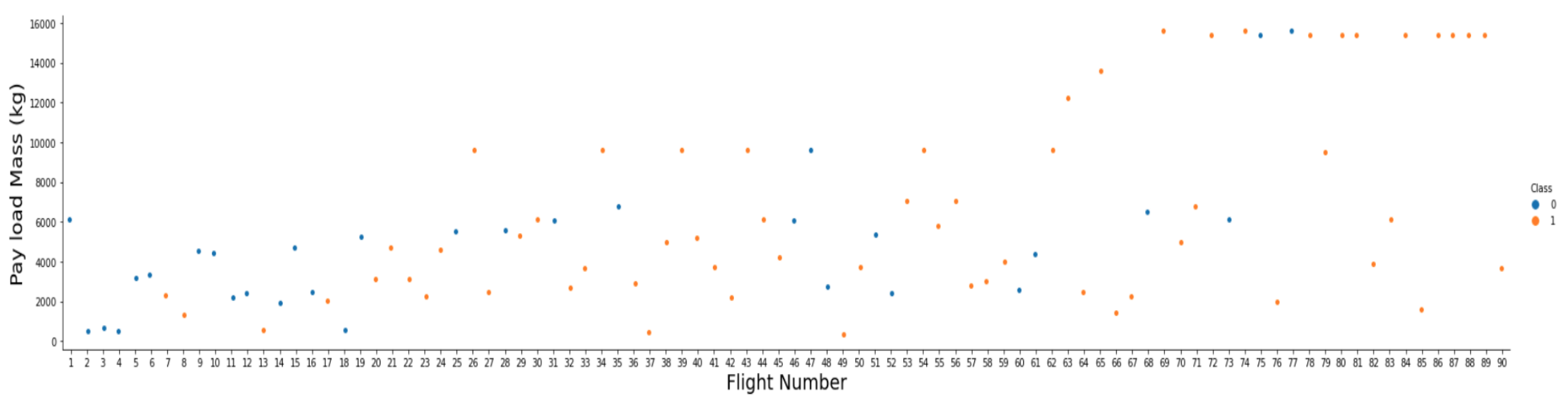
Finding the best performing model

- The model with the best accuracy score wins the best performing model

Results

EDA with visualization

Flight Number vs. Pay Load Mass

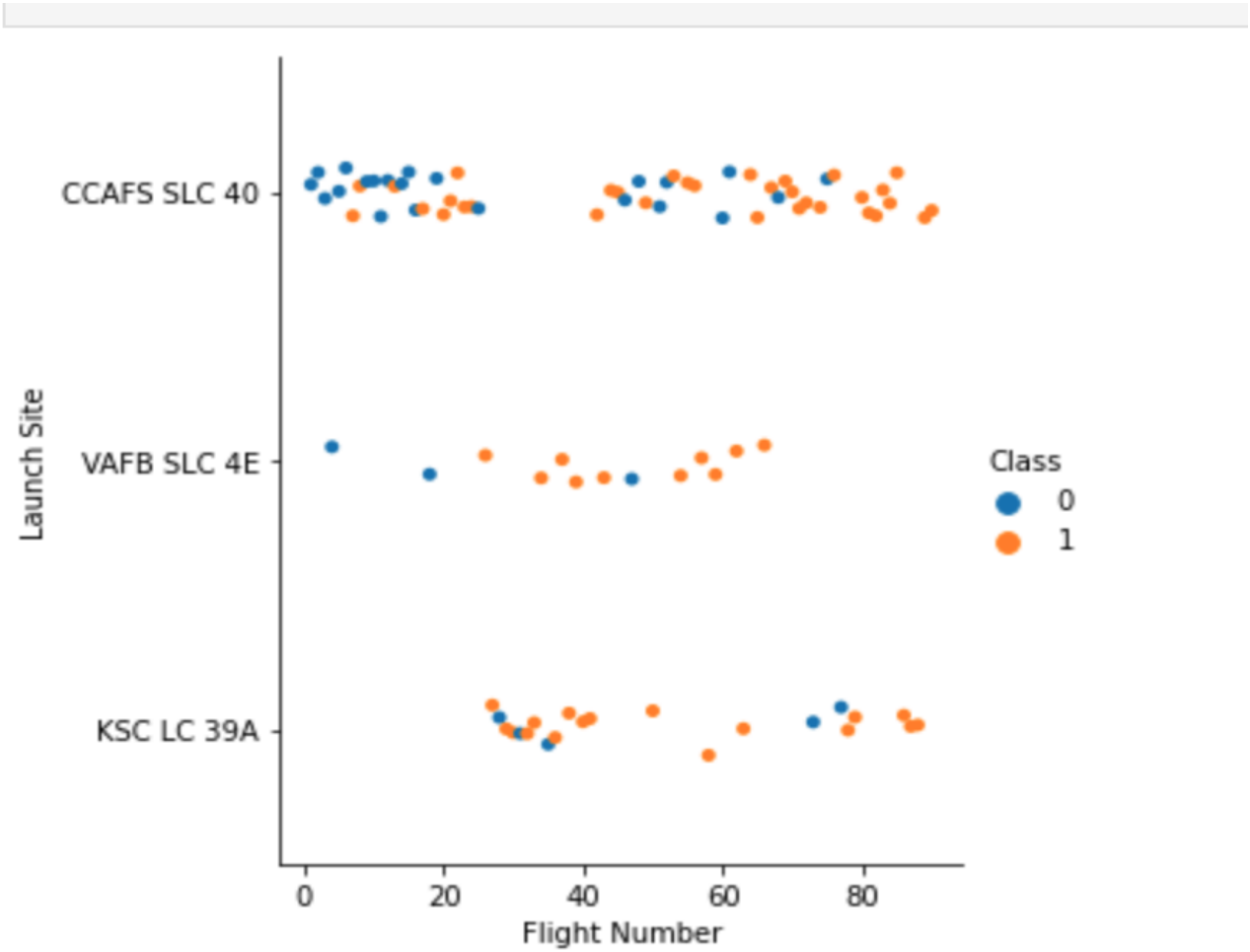


- As the flight number increases, the first stage is more likely to land successfully.
- There is not quite a clear pattern in this visualization to make a decision if the success of first stage landing depends on Pay Load Mass.

Results

EDA with visualization

Flight Number vs Launch Site

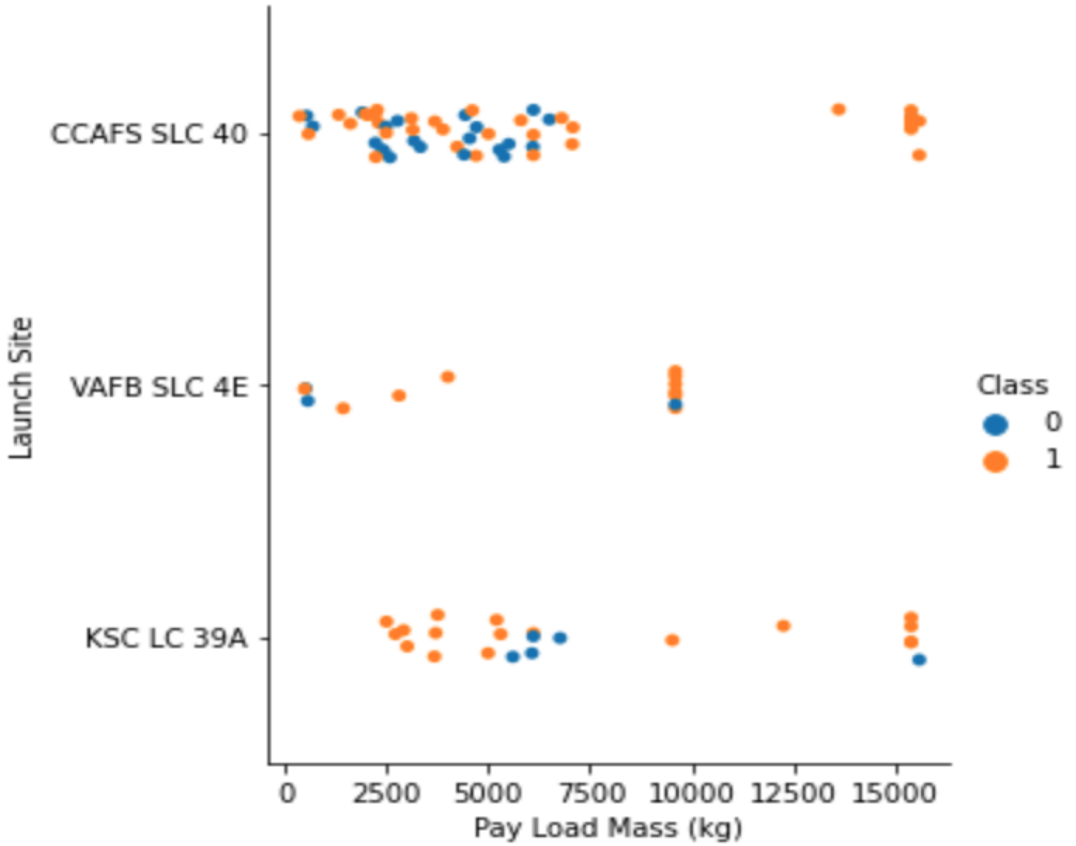


- **Class 0 (blue) represents unsuccessful launch, and Class 1 (orange) successful launch.**
- **The success rate increases as the number of launch increase at each site especially after the 20th launch**

Results

EDA with visualization

Pay Load Mass vs Launch Site

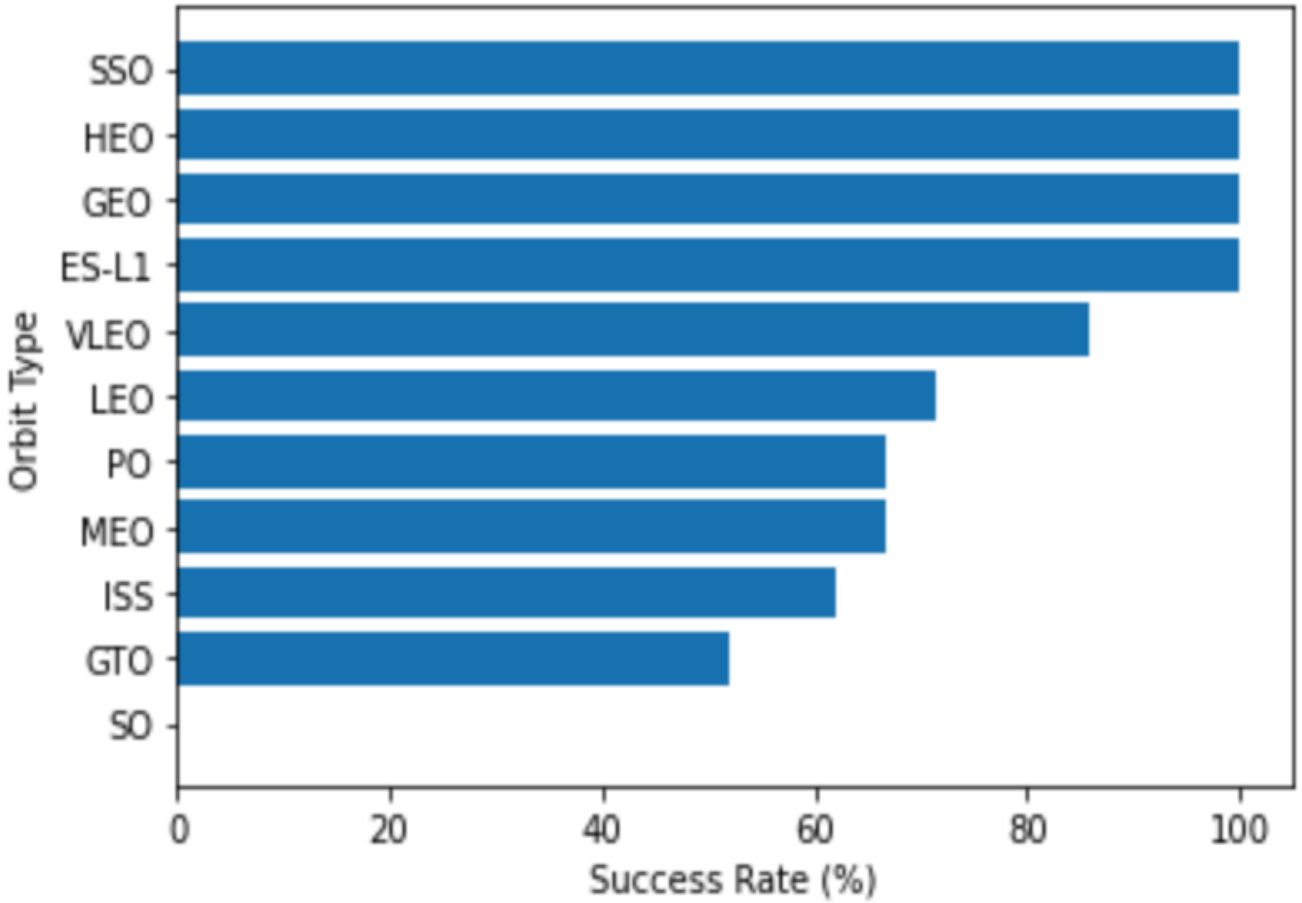


- The greater the payload mass for Launch Site CCAFS SLC 40 the higher the success rate for the Rocket.
- However, KSC LC 39A launch site success rate of low weighted payloads is higher than that of heavy weighted payloads.
- However, no clear pattern found between successful launch and Pay Load Mass.

Results

EDA with visualization

Success rate vs. Orbit type

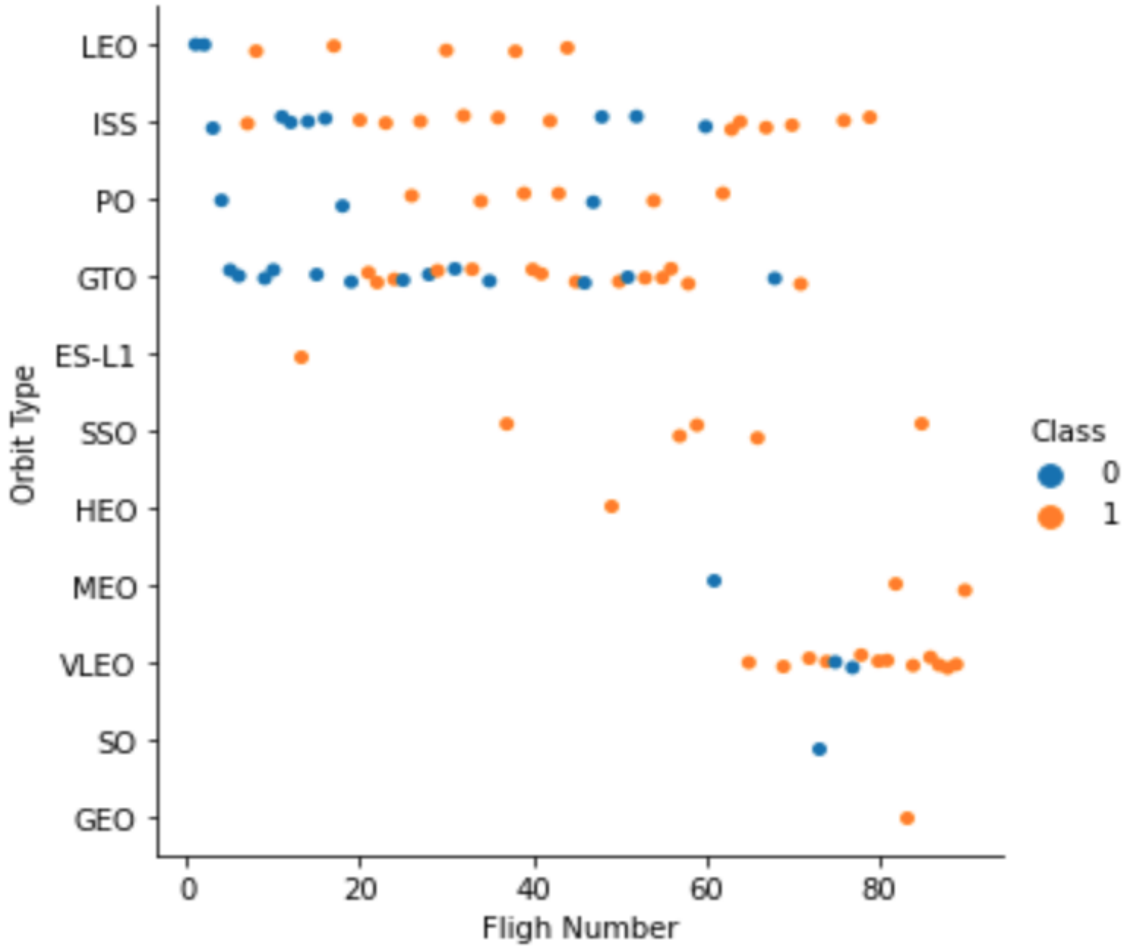


- **Orbit types SSO, HEO, GEO, and ES-L1 have 100% success rates.**
- **Orbit type GTO has a 50% success rates.**
- **While, type SO recorded a failure.**

Results

EDA with visualization

Flight Number vs. Orbit type

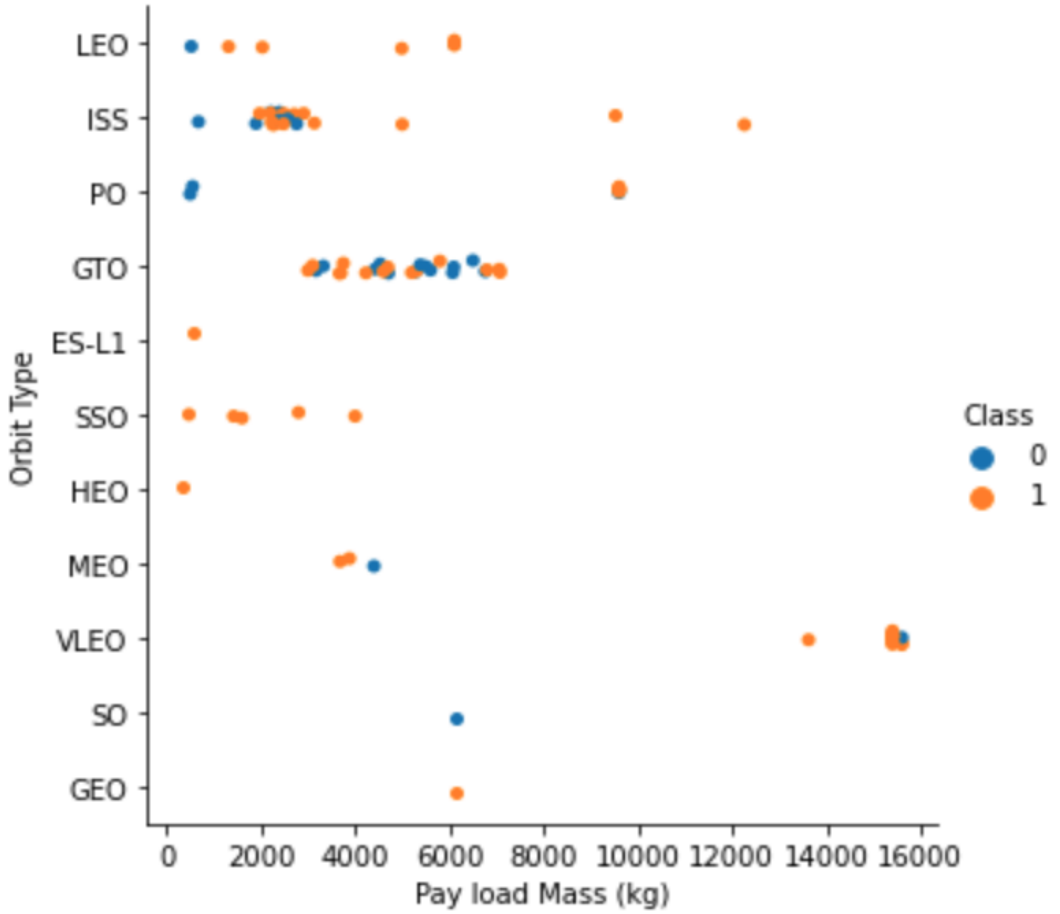


- It is visible that the success rate of orbit LEO , SSO, and VLEO related to the number of flights
- On the other hand, there seems to be no relationship between flight number in GTO orbit.
- The success rate of orbit ES-L1, HEO, and GEO are skewed due having a single launch.

Results

EDA with visualization

Payload Mass vs. Orbit Type

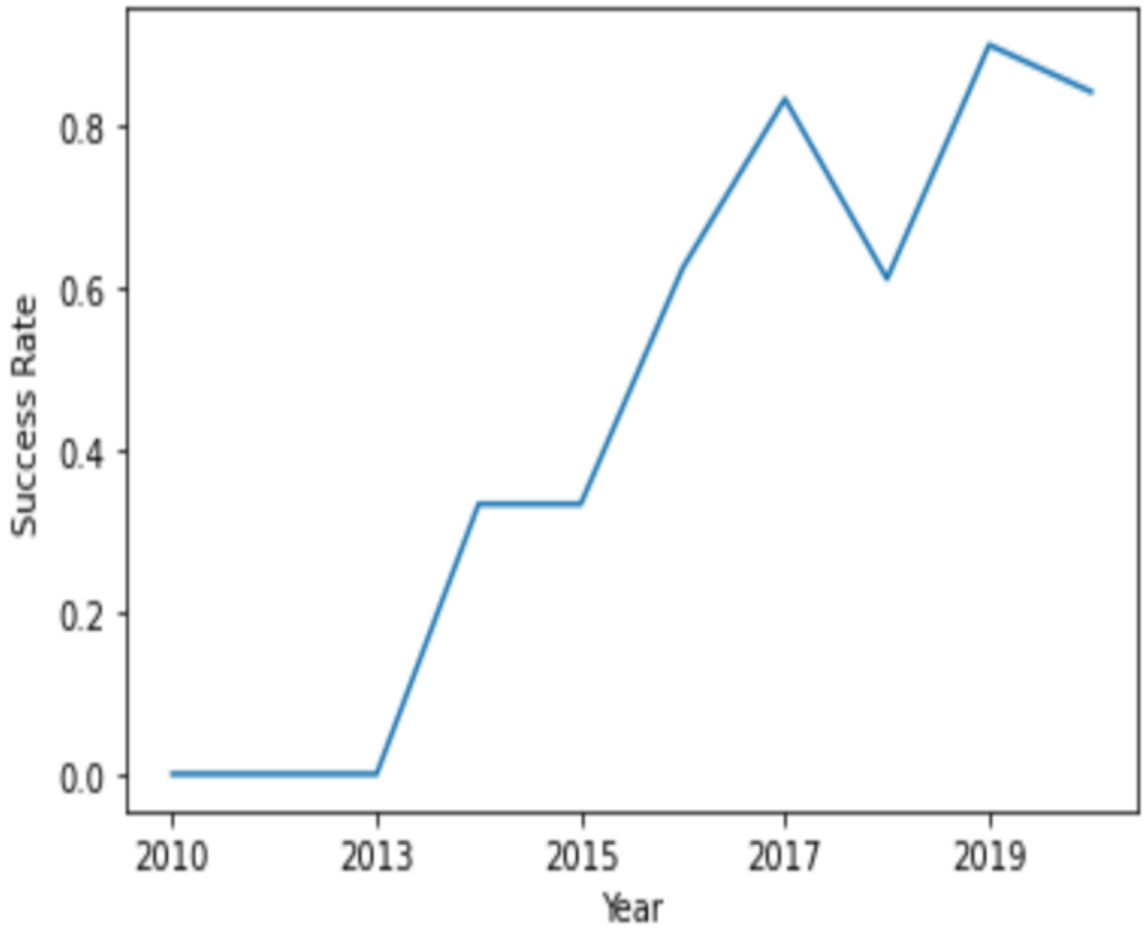


- We can observe that heavy payloads have a positive landing influence for LEO, ISS and PO orbits.
- On the other hand, there is no relationship between payload mass and success rate for GTO orbit.

Results

EDA with visualization

Launch success yearly trend



- **The success rate kept increasing since 2013 till 2020, with a slightly decreased in 2018.**

Results

EDA with SQL

Unique launch sites in the space mission

Query

```
%%sql
SELECT DISTINCT LAUNCH_SITE
FROM SPACEXTBL
```

Results

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

- **When the word DISTINCT is used in the query, it will only shows unique values in the Launch_Site column from the SPACEXTBL table.**

Results

EDA with SQL

Launch sites begin with the string 'CCA'

Query

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

- **When the word LIMIT is used in the query, it will only shows a specified number of records(i.e. 5) from the SPACEXTBL table.**
- **Using the LIKE operator and the percent sign (%) together, the Launch_Site name starting with CAA could be called.**

Results

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

Results

EDA with SQL

The total payload mass carried by boosters launched by NASA (CRS)

Query

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS total_payload_mass_kg
FROM SPACEXTBL
WHERE CUSTOMER = 'NASA (CRS)'
```

Results

total_payload_mass_kg
45596

- **Using the SUM function to summates the total in the column PAYLOAD_MASS_KG_**
- **The WHERE clause filters the dataset to perform calculations only if Customer is NASA (CRS)**

Results

EDA with SQL

Average payload mass carried by booster version F9 v1.1

Query

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS avg_payload_mass_kg
FROM SPACEXTBL
WHERE BOOSTER_VERSION = 'F9 v1.1'
```

Results

avg_payload_mass_kg
2928

- **The AVG() function returns the average value of an expression. Hence, the function AVG works out the average in the column PAYLOAD_MASS_KG_**
- **The WHERE clause filters the dataset to only perform calculations on Booster_version F9 v1.1**
- **On average, rockets with by booster version F9 v1.1 carry a mass of 2928KG.**

Results

EDA with SQL

First successful landing date

Query

```
%%sql
SELECT MIN(DATE) AS first_successful_landing_date
FROM SPACEXTBL
WHERE LANDING__OUTCOME = 'Success (ground pad)'
```

Results

first_successful_landing_date
2015-12-22

- **The MIN() function returns the smallest value of the selected column. The MIN() function used to find out the earliest date in the column DATE.**
- **In the WHERE clause, filter the dataset to perform a search only if Landing__outcome is Success (ground pad)**
- **The first successful Stage One recovery landing occurred on December 12, 2015.**

Results

EDA with SQL

Names of boosters with successful drone landing with payload between 4000 and 6000 KG

Query

```
%%sql
SELECT BOOSTER_VERSION
FROM SPACEXTBL
WHERE LANDING__OUTCOME = 'Success (drone ship)'
    AND (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000)
```

Results

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

- **The WHERE clause used to filter for boosters which have successfully landed on drone ship.**
- **Moreover, the AND condition applied to determine successful landing with payload mass greater than 4000 but less than 6000.**

Results

EDA with SQL

The total number of successful and failure mission outcomes

Query

```
%%sql
SELECT MISSION_OUTCOME, COUNT(*) AS total_number
FROM SPACEXTBL
GROUP BY MISSION_OUTCOME
```

Results

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

- **The COUNT function used to count the number of rows returned in a MISSION_OUTCOME.**
- **The GROUP BY clause is used in collaboration with the SELECT statement to arrange identical values into groups to find the total number in each MISSION_OUTCOME.**
- **There were 100 missions recorded in the database**

Results

EDA with SQL

List of Boosters Carried Maximum Payload

Query

```
%%sql
SELECT DISTINCT BOOSTER_VERSION, PAYLOAD_MASS__KG_
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (
    SELECT MAX(PAYLOAD_MASS__KG_)
    FROM SPACEXTBL);
```

- **List of boosters that have carried the maximum payload selected using a subquery in the WHERE clause and the MAX() function.**

Results

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

Results

EDA with SQL

Failed Landing outcomes, booster versions, and launch site of 2015

Query

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

Results

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

- **The WHERE clause filters the dataset to perform a search of failed landing outcome.**
- **Additional condition is set using the AND operator to display a record of YEAR 2015.**

Results

EDA with SQL

Landing outcomes rank between 2010-06-04 and 2017-03-20

Query

```
%%sql
SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS total_number
FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LANDING__OUTCOME
ORDER BY total_number DESC
```

Results

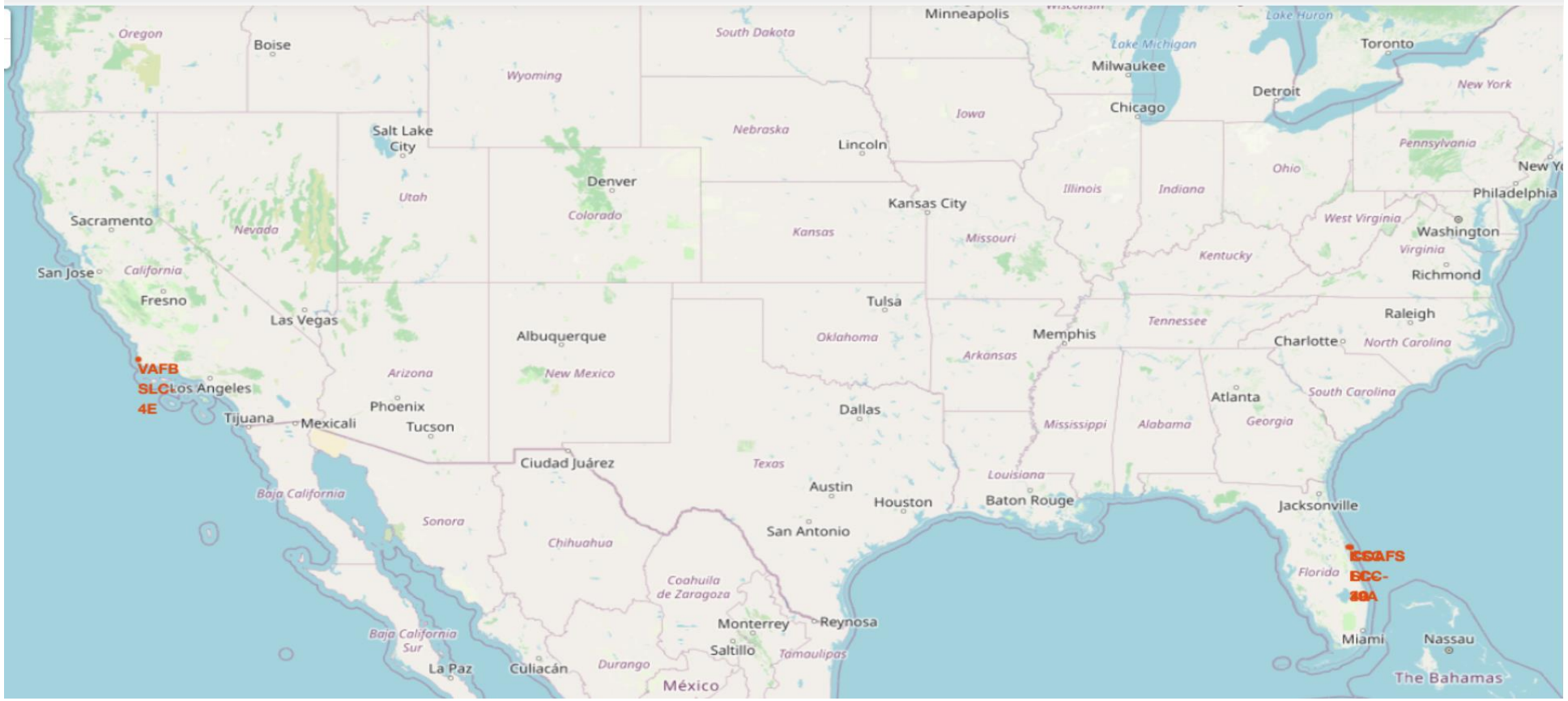
landing__outcome	total_number
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

- **The COUNT function used to count the LANDING_OUTCOME and the WHERE clause to filter date BETWEEN 2010-06-04 to 2010-03-20.**
- **The GROUP BY clause used in collaboration with the SELECT statement to arrange identical values into groups to find the total number in each LANDING_OUTCOME.**
- **Finally, the total number of outcomes ORDER BY descending order.**

Results

Interactive Visual Analytics with Folium

SpaceX launch sites on a map

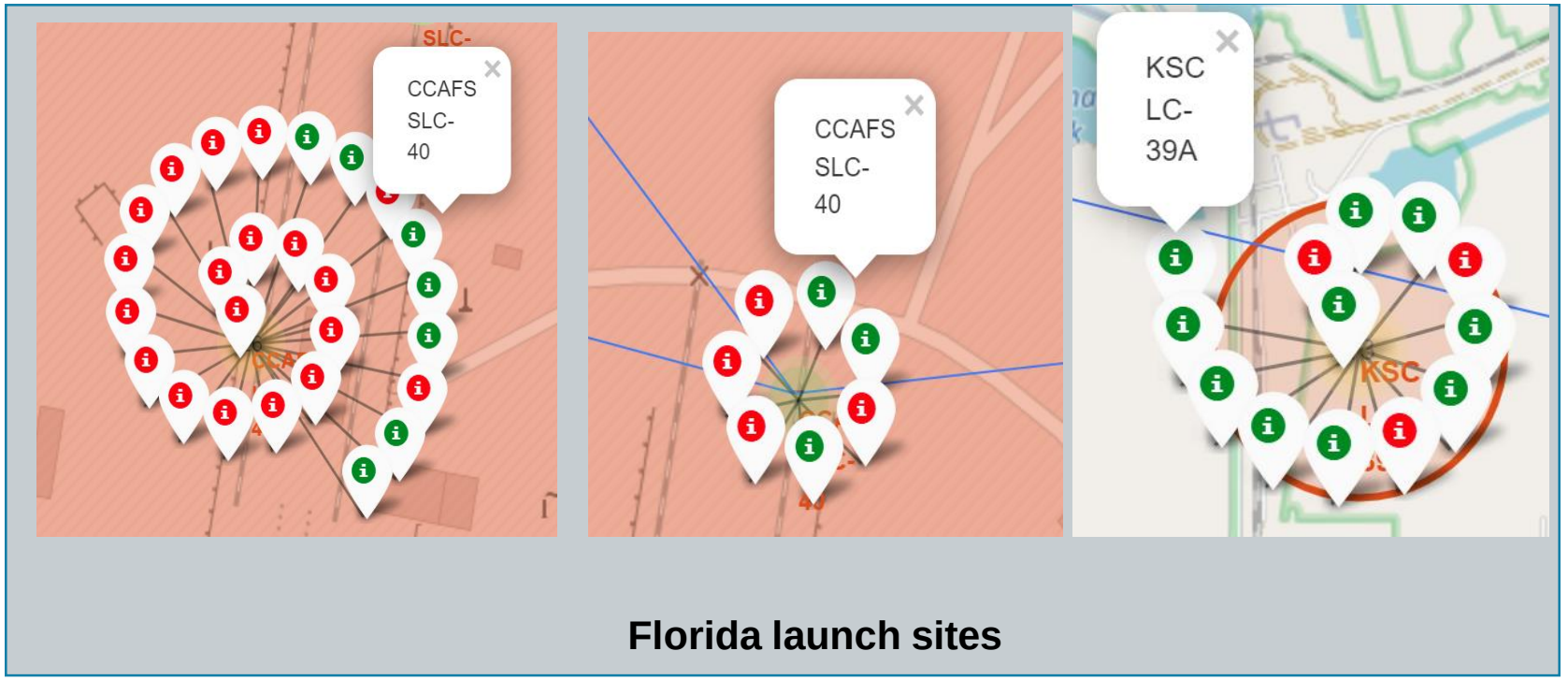


- **The map shows that all launch sites of SpaceX are in the United States of America near to the coasts of Florida and California.**

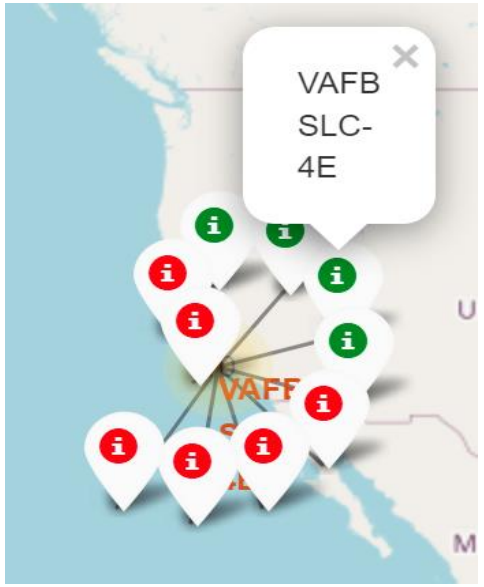
Results

Interactive Visual Analytics with Folium

Color Labelled Markers



Florida launch sites



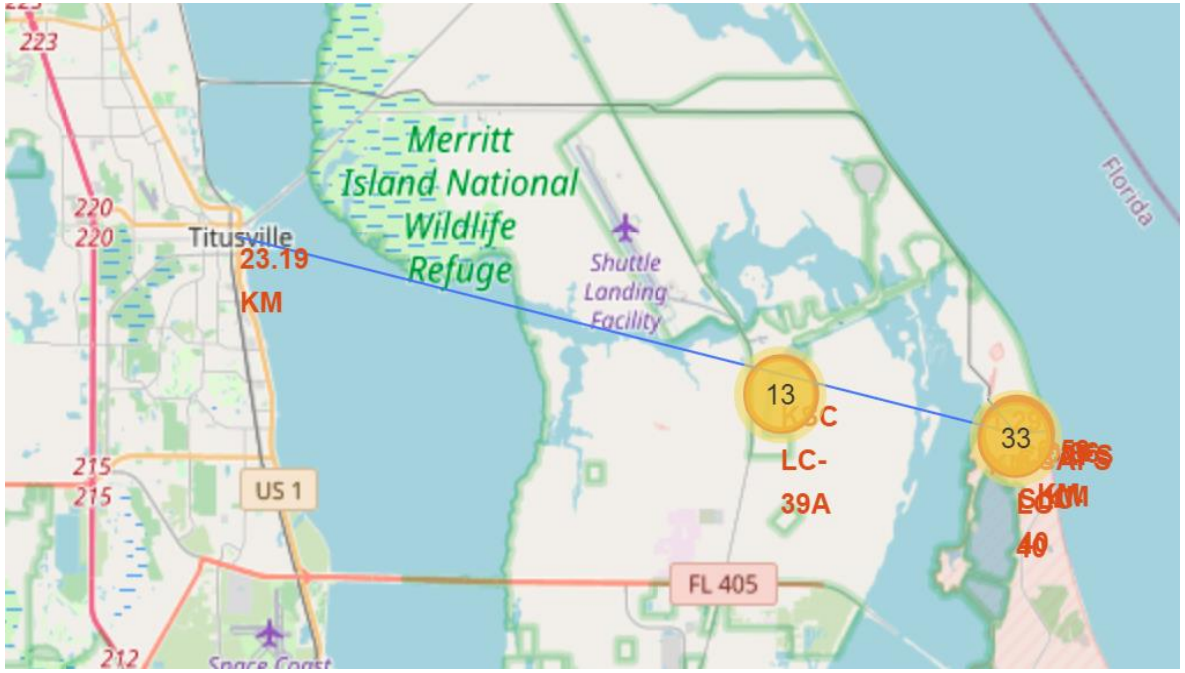
California launch site

- Green Marker shows successful Launches and Red Marker shows Failures
- KSC LC-39A launch site had the highest success rate.

Results

Interactive Visual Analytics with Folium

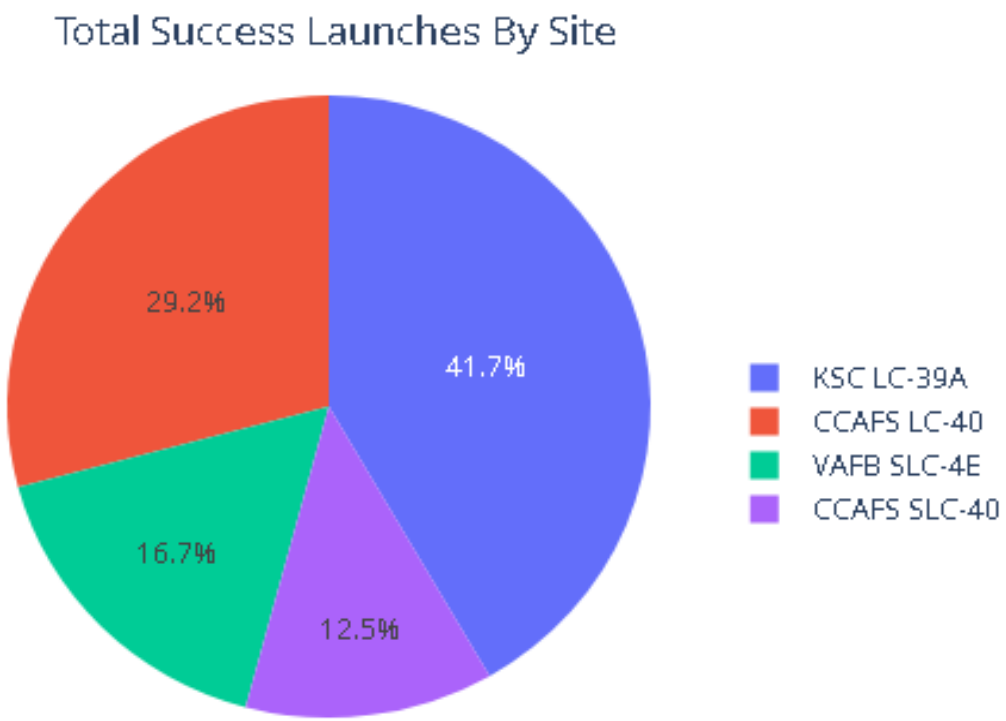
Launch Sites distance to landmarks

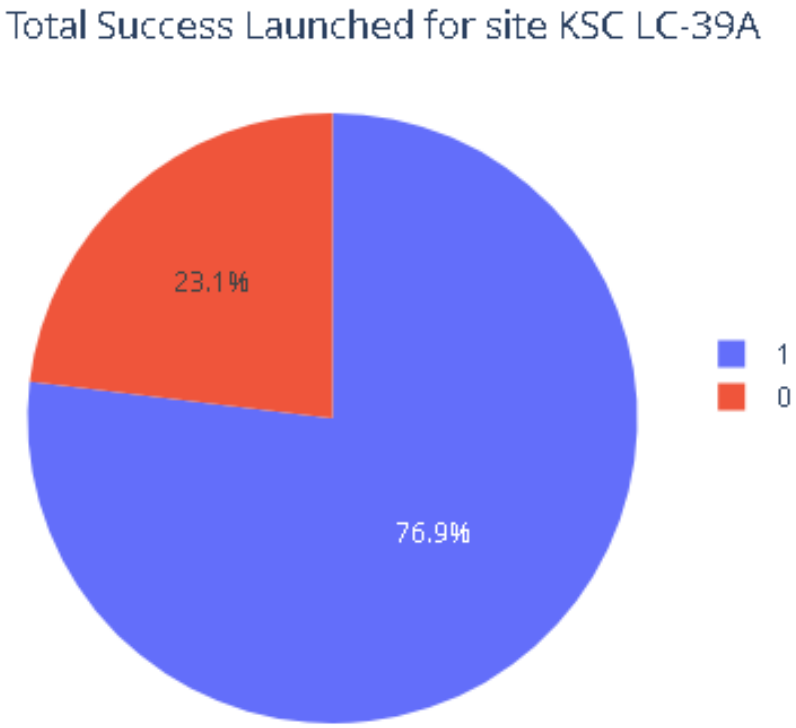


- Are launch sites in close proximity to railways? Yes
- Are launch sites in close proximity to highways? Yes
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes
- Launch site are close to railways and highways for transportation of equipment or personnel.
- They are also close to coastline and relatively far from the cities so that launch failure does not pose a threat.

Results

Dashboard with Plotly Dash





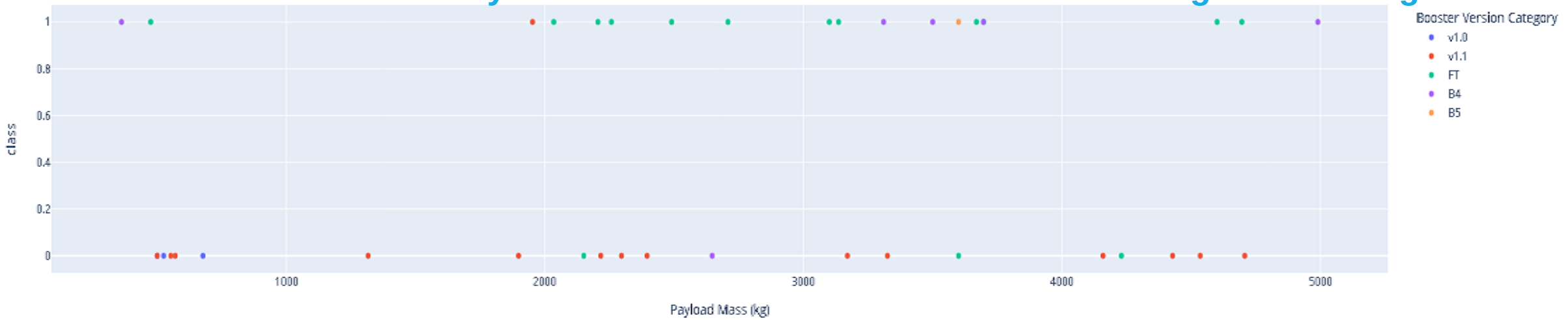
- **KSLC-39A had the highest launch success rate**
- **While VAFB SLC-4E had the lowest launch success rate among all sites.**

- **KSC LC-39A had 76.9% success and 23.1% failure rate**

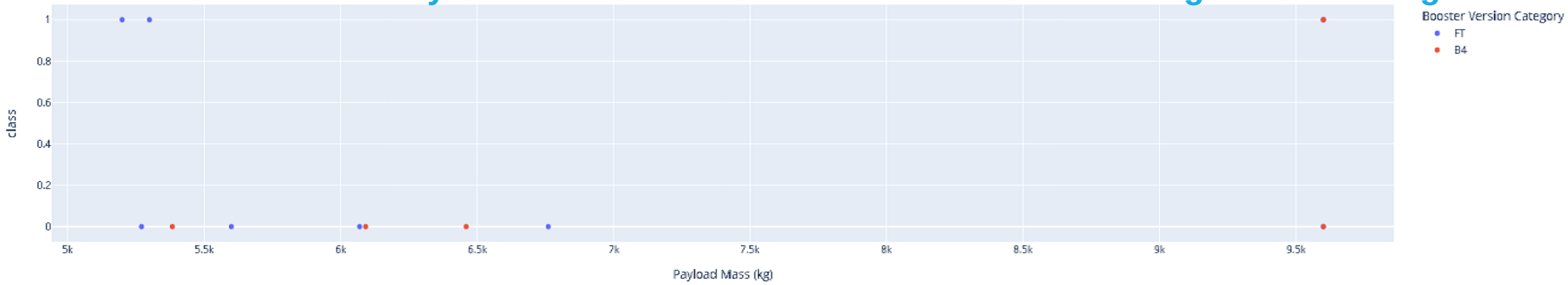
Results

Dashboard with Plotly Dash

Correlation between Payload and Success for all sites between 0kg and 5000 kg



Correlation between Payload and Success for all sites between 5000kg and 10000 kg

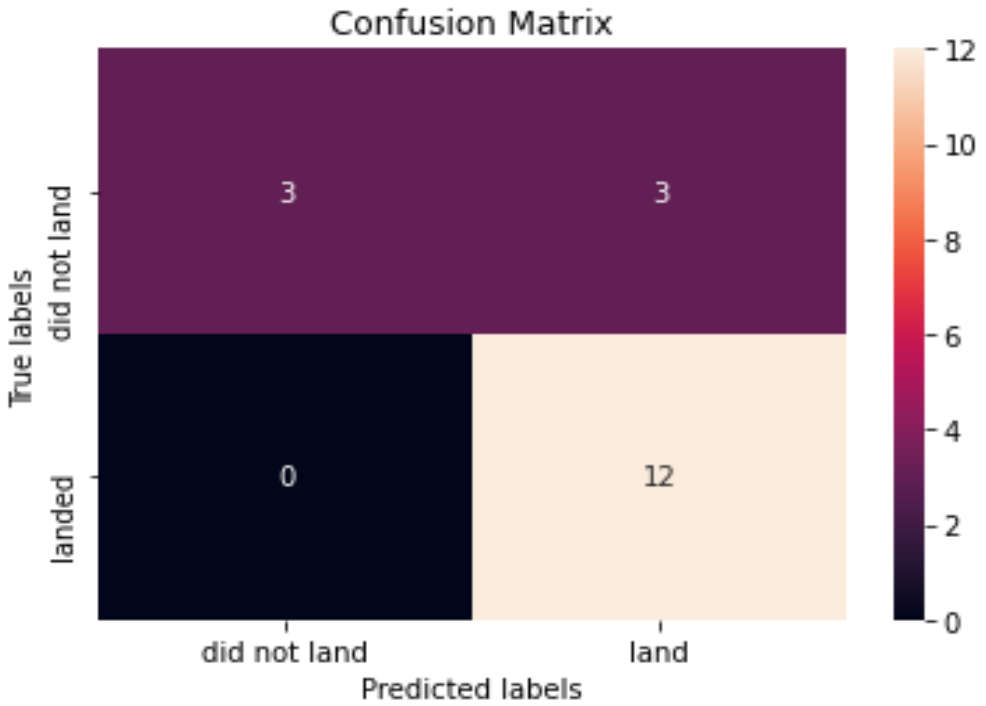


- **The success rates for low weighted payloads is higher than the heavy weighted payloads**

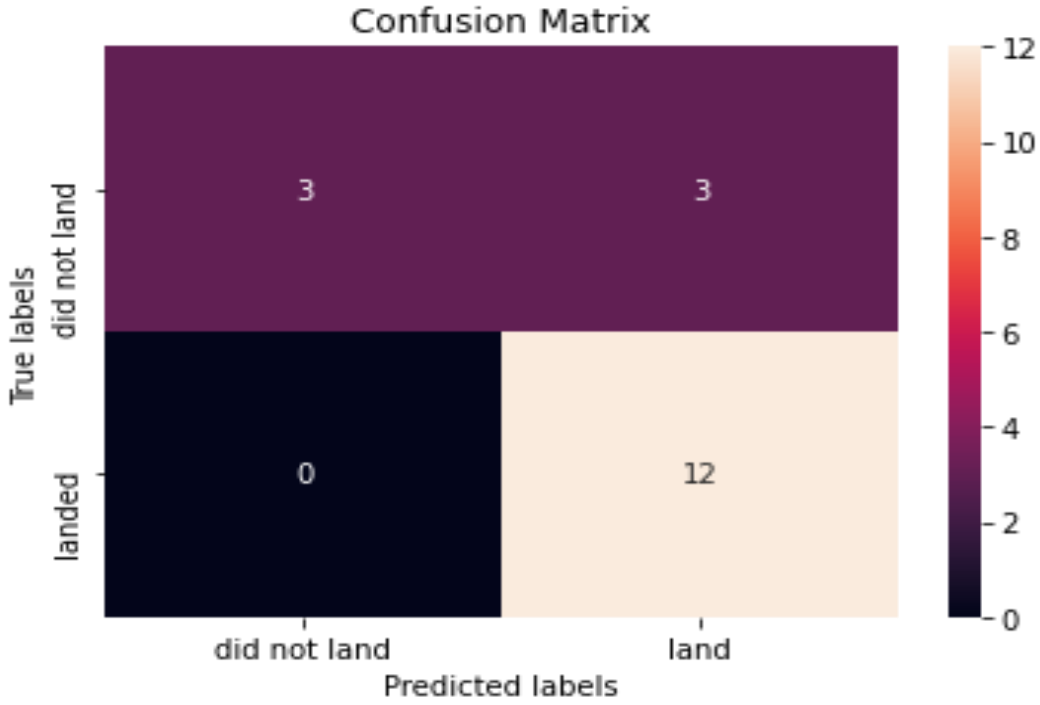
Results

Predictive analysis (Classification)

Logistic regression confusion matrix



Support Vector Machine (SVM) confusion matrix

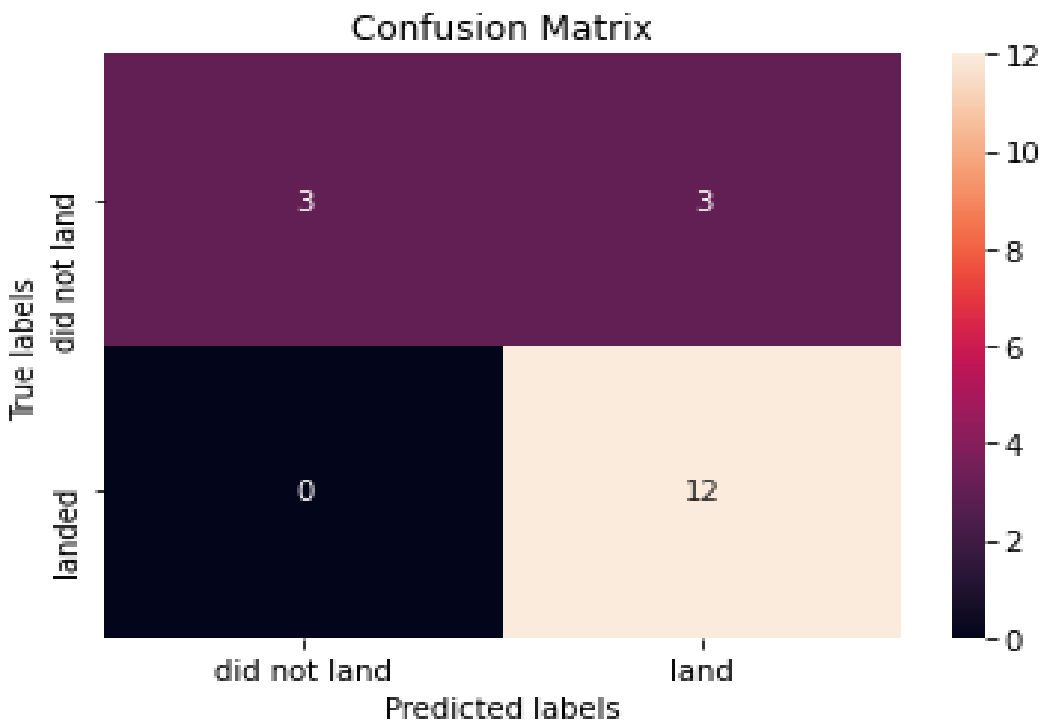


- **Logistic regression and SVM can distinguish between the different classes. However, the major problem we see from confusion matrix is false positives, unsuccessful landing marked as successful landing by the classifier.**

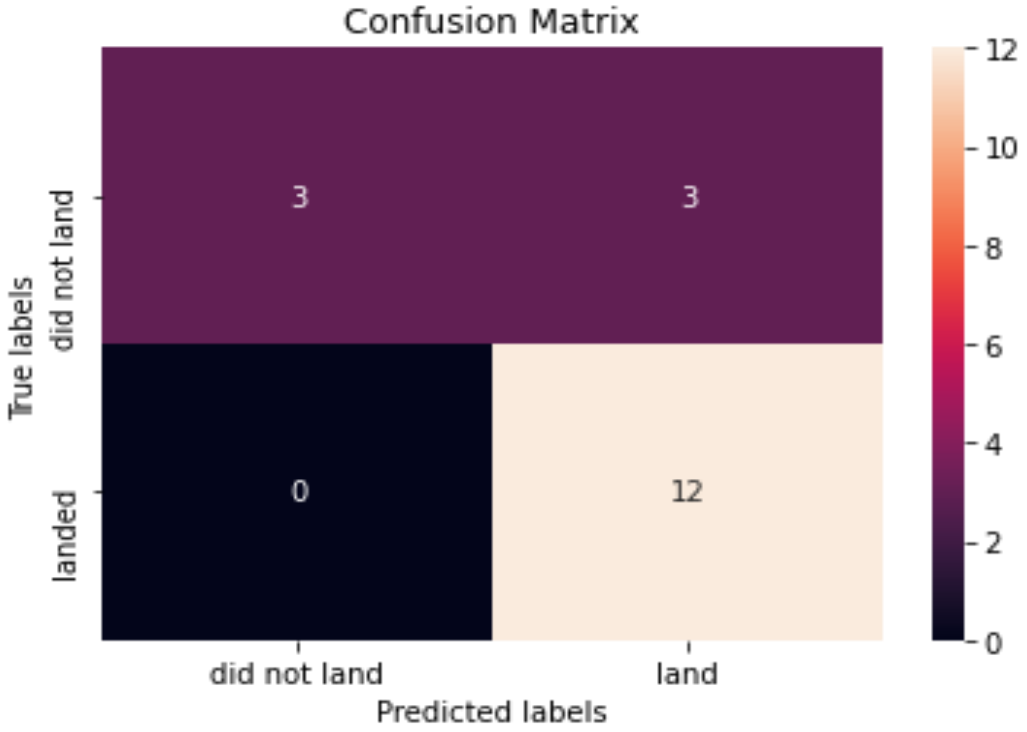
Results

Predictive analysis (Classification)

Decision tree classifier confusion matrix



K-nearest neighbors confusion matrix

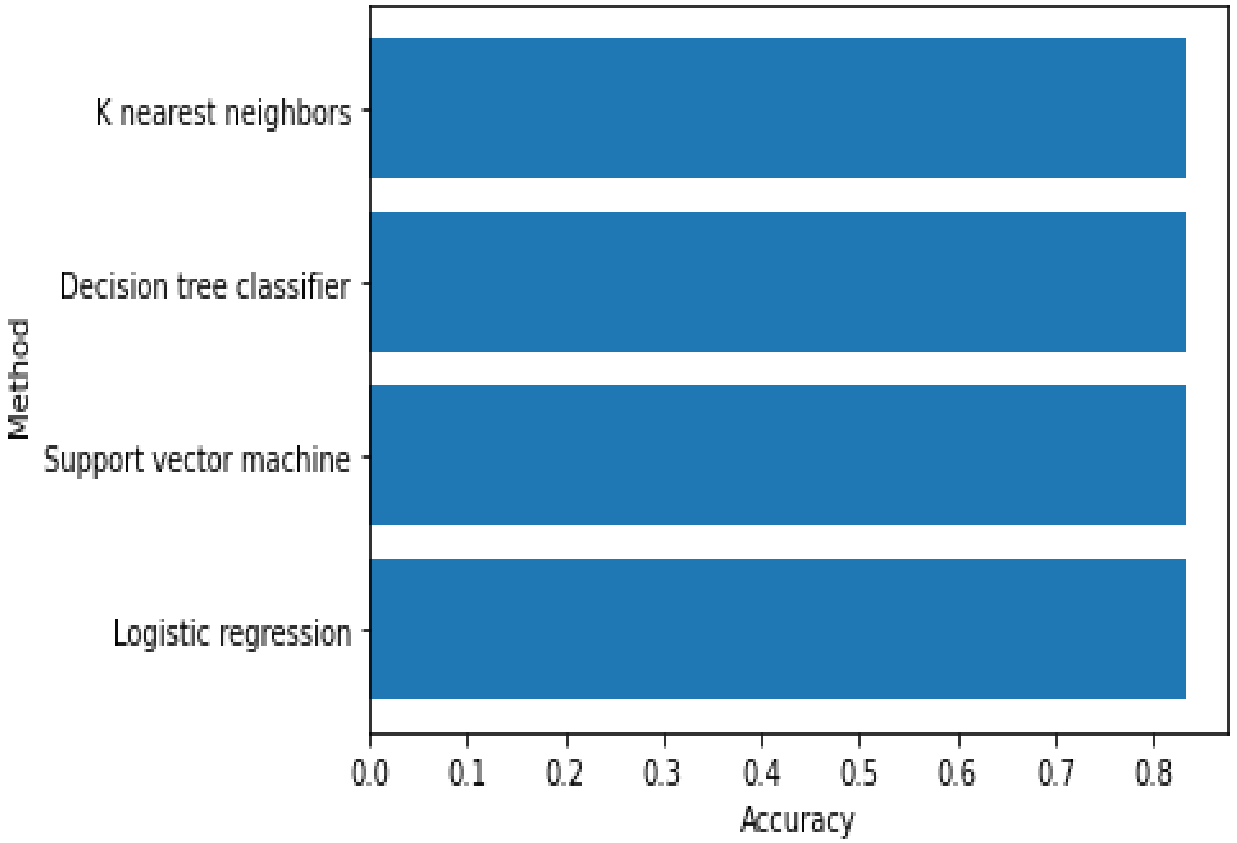


- **Similarly, the confusion matrix of decision tree classifier and K-nearest neighbor has a problem false positives.**

Results

Predictive analysis (Classification)

Models Accuracy Test



- **The accuracy of machine learning algorithms to predict Falcon 9 first stage successful landing is the same, i.e. 83.33% for this dataset.**

Conclusions

- The success rates of SpaceX Falcon 9 launches increase as the number of launches increases.
- Orbit SSO, HEO, GEO, and ES-L1 has the highest success rates.
- KSC LC-39A launch site has the highest success rate.
- Launch site are close to railways and highways for transportation of equipment or personnel. They are also close to coastline and relatively far from the cities so that launch failure does not pose a threat.
- The success rates for low weighted payloads is higher than the heavy weighted payloads


THANK YOU

Appendix

- **Github link**

  https://github.com/usmuh/Applied-Data-Science-Capstone-Project