IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

NatalieCheong
October 4, 2022

Table of Content

Executive Summary

SpaceY is a new commercial rocket launch and wants to bit against SpaceX.

SpaceX rocket launches are relatively inexpensive. SpaceX advertise Falcon 9 rocket launches on its website with a cost of 62 million dollars. Other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.

SpaceX public statements indicate Falcon 9 is reusable, two-stage rocket designed. Falcon 9 is allows SpaceX to refly the most expensive parts of the rocket, which in turn drives down the cost of space access.

SpaceY would gathering information about SpaceX and determine if the first stage will land successfully, and will train a machine learning model and use public information to predict of SpaceX will reuse the first stage.

INTRODUCTION: BACKGROUND

- This report has been prepared as part of the Applied Data Science Capstone course.
- I will take a role of a Data Scientist working for a new rocket company named SpaceY. My job is to determine the price of each launch.
- Instead of using rocket science to determine if the first stage will land successfully, I will train a machine learning model and using public infromation to predict if SpaceX will reuse the first stage.

INTRODUCTION: BUSINESS PROBLEMS

- SpaceX advertise Falcon9 rocket launches with a cost of 62 million dollars where the first stage of their rockets can be reused.
- Sometimes SpaceX will sacrifice the first stage due to mission parameters such as payload, orbit, and customers.
- Thus, this report is aim to predict the first stage rocket landing successfully for the cost of a launch.

Section 1

Methodology

METHODOLOGY

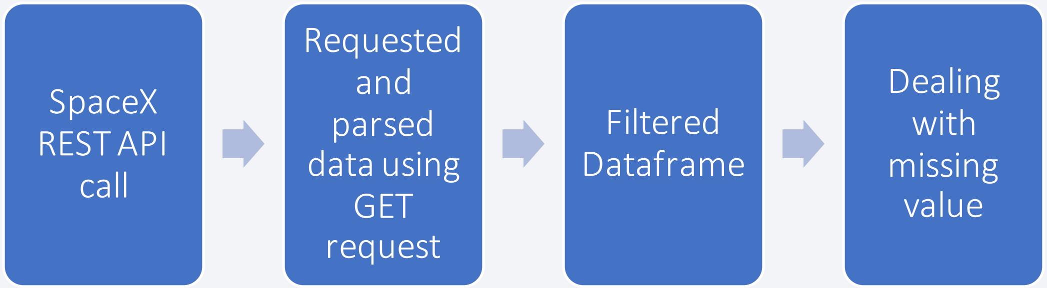
The report of the data science methodology are outlined as such:

- Data collection methodology
- Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
- Report results

Data Collection – SpaceX API

- To determine if the first stage will land, and determine the cost of launch.
- This information can be used if SpaceY wants to bid against SpaceX for a rocket launch.
- Collect and make sure the data is in the correct format from and API.
- API

- Acquired historical launch data from Open Source REST API for SpaceX.
- Requested and parsed the SpaceX launch data using GET request.
- Filtered the dataframe only include Falcon 9 launches.
- Dealing with the missing values.



https://github.com/NatalieCheong/SpaceX-Lab1-Collecting-the-Data/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

- Web scraping:
- Acquired historical launch data from Wikipedia page "List of Falcon9 and Falcon Heavy Launches
- Requested the Falcon9 launch Wiki page from its URL
- Extract all column/variable names from the HTML table header
- Create a data frame by parsing the launch HTML tables

https://github.com/NatalieCheong/SpaceX-Lab2-WebScraping/blob/main/Week1-Lab2-jupyter-labs-webscraping.ipynb

Data Wrangling

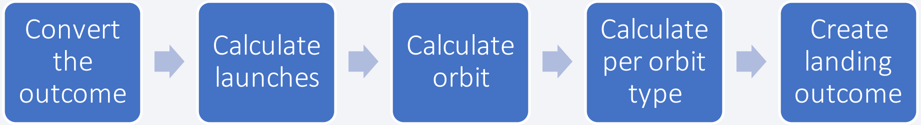
- To perform some Exploratory Data Analysis to find some patterns in the data and determine what would be the label for training supervised model.
- In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident.
- The objectives:
  - Exploratory Data Analysis
  - Determine Training Labels
- Will mainly convert the outcomes into Training Labels with 1 means the booster successfully landed, 0 means it was unsuccessfully.

Task 1 : Calculate the number of launches on each site

Task 2 : Calculate the number and occurance of each orbit

Task 3 : Calculate the number and occurance of missing outcome per orbit type

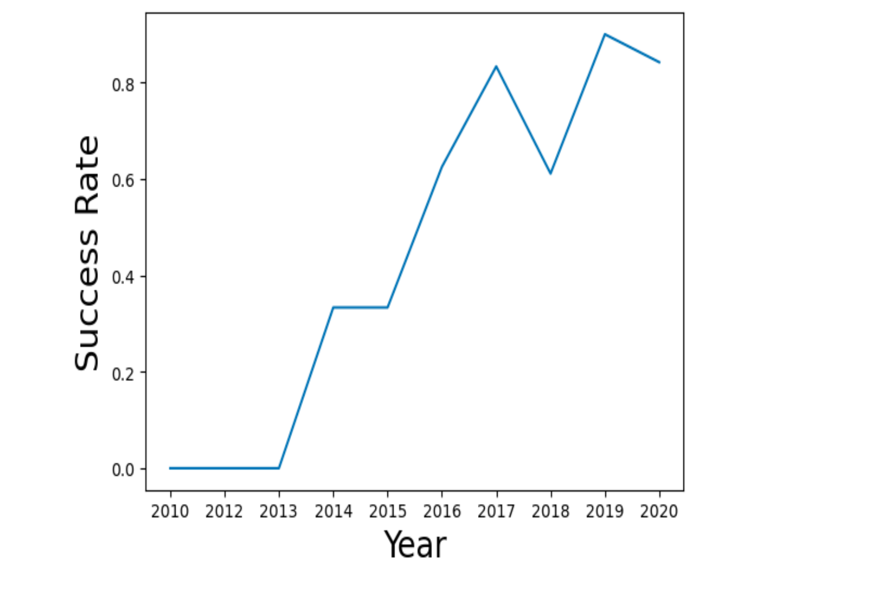
Task 4 : Create a landing outcome label from outcome column



https://github.com/NatalieCheong/SpaceX-Lab2-Data-Wrangling/blob/main/Week1-Lab2-Data%20Wranglinglabs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- We will predict if the Falcon9 first stage will land successfully.
- SpaceX advertises Falcon9 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 due to the fact that SpaceX can reuse the first stage.
- Several examples of an unsuccessful landing are shown here: Most unsuccessful landings are planned. SpceX performs a controlled landing in the oceans.
- We will use Matplotlib and Seaborn visualization libraries to plot the outcome of FlightNumber, PayloadMass, LaunchSite, Orbit type, Success rate, Payload and Year.



The success rate since 2013 kept increasing till 2020

https://github.com/NatalieCheong/SpaceX-Lab-Exploring-and-Preparing-Data/blob/main/Week2-Exploring%20and%20Preparing%20Data%20jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- Run SQL queries to display and list information about:
  - Launch sites
    - Display the names of the launch sites in the Space mission
    - Display 5 records where launch sites begin with 'CCA'
  - Payload Masses
    - Display the total payload mass carried by booster launch by NASA(CRS)
    - Display average payload mass carried by booster version F9

https://github.com/NatalieCheong/SpaceX-Lab-Sql-Notebook/blob/main/Week%202%20SQL.ipynb

- Booster versions
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Mission outcomes
  - List the total number of successful and failure mission outcomes
- Booster landings
  - Rank the count of successful landing outcomes

Build an Interactive Map with Folium

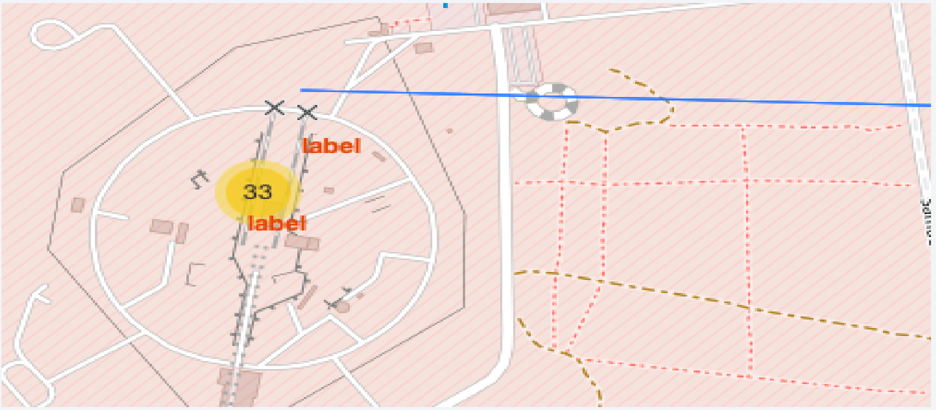
- Launch sites locations analysis with Folium.
- The launch success rate may depend on may factors such as payload mass, orbit type, and so on. It may also depend on the location and proximities of a launch site, i.e, the initial position of rocket trajectories.
- Finding an optimal location for building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

https://github.com/NatalieCheong/Interactive-Visual-Analytics-with-Folium-/blob/main/Week3-Folium-ab_jupyter_launch_site_location.ipynb

Task 1: Mark all launch sites on a map

Task 2: Mark the success/failed launches for each site on the map

Task 3: Calculate the distance between a launch site to its proximities

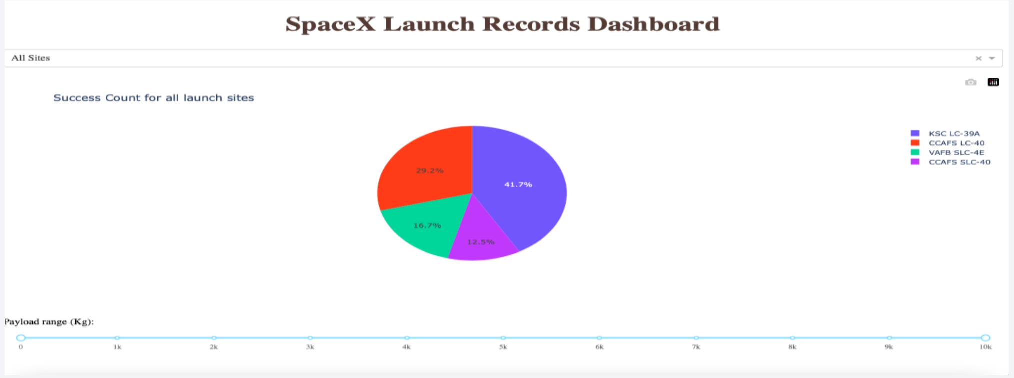


The interactive Folium map showing proximity from CCAFS SLC-40 launch site to nearby railway, highway, and coastline

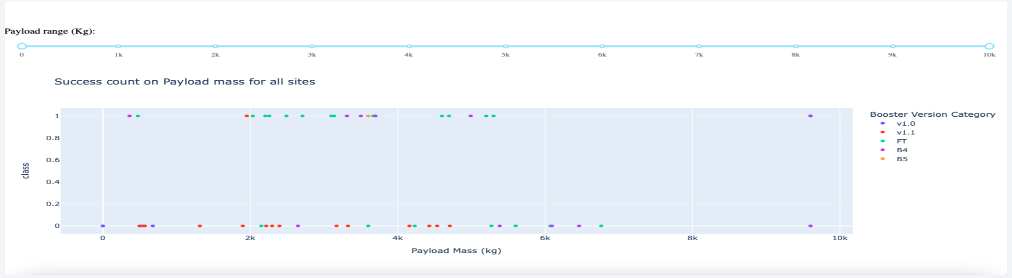
Build a Dashboard with Plotly Dash

- Building a Plotly Dash Application for stakeholders to have better understanding and manipulate data in an interactive visual analytics on SpaceX launch data real-time.
- Pie chart showing success rate
- Scatter chart showing Payload Mass vs Landing Outcome
- Drop-down menu to choose between all sites and individual launch sites.

  https://github.com/NatalieCheong/SapceX-Dashboard-Plotly-Dash/blob/main/Week3-spacex_dash_app.py



Success count for all launch sites



Success count on Payload Mass for all sites

Predictive Analysis (Classification)

- Will create a Machine Learning pipeline to predict if the first stage will land given the data from the preceding outcomes.
- The objectives:
  - Create a column for the class
  - Standardize the data
  - Split into training data and test data
  - Find best Hyperparameter for SVM, Classification Decision Tree And Logistic Regression
- Evaluate the accuracy of each model using test data to select the best model

- Fit the training data to various model types:
  - Logistic Regression
  - Support Vector Machine (SVM)
  - Decision Tree Classifier
  - K Nearest Neighbors Classifier
- Used a cross-validation grid-search over a variety of hyperparameters to select the best ones for each model:
  - Enabled by scikit-learn library function GridSearch CV

https://github.com/NatalieCheong/SpaceX-Lab-Machine-Laerning-Prediction/blob/main/Week4-SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Predictive Analysis (Classification)-Continued

Create Column
- Standardize the data

Split Data
- Train Data
- Test Data

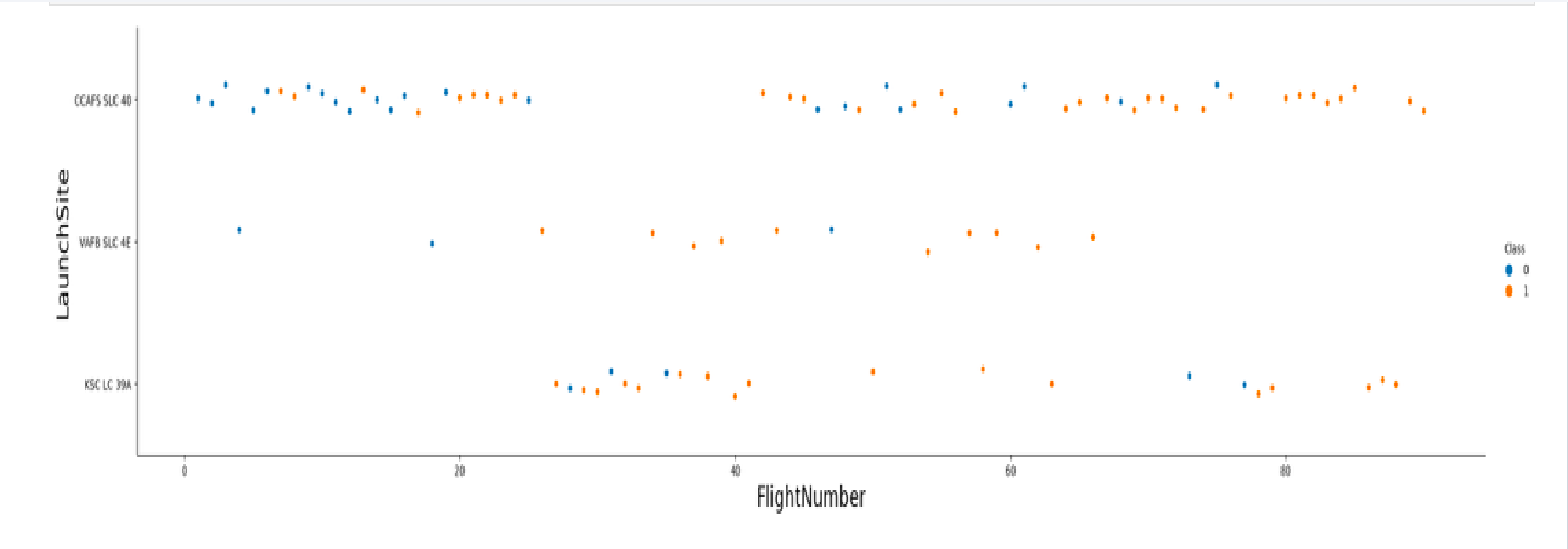
Fit the training data
- Logistic Regression
- Support Vector Machine
- Decision Tree Classifier
- K Nearest Neighbors

The method performs the best
- Bar Chart

Section 2

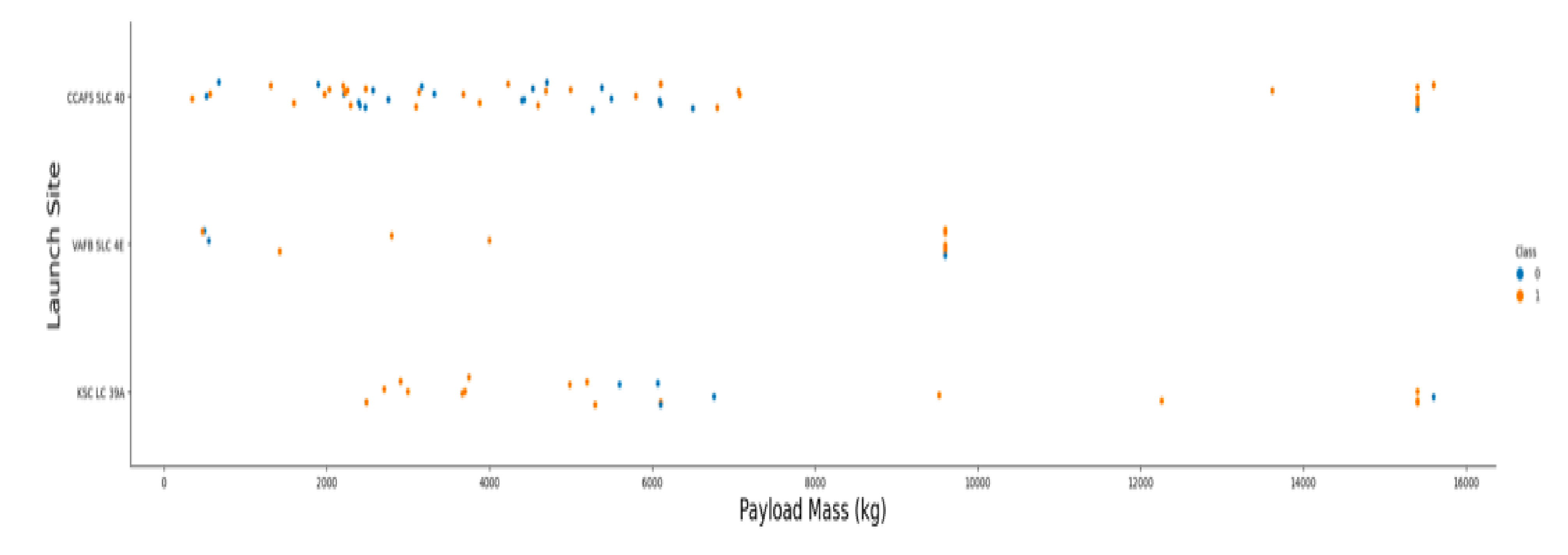
Insights drawn from EDA

Flight Number vs. Launch Site



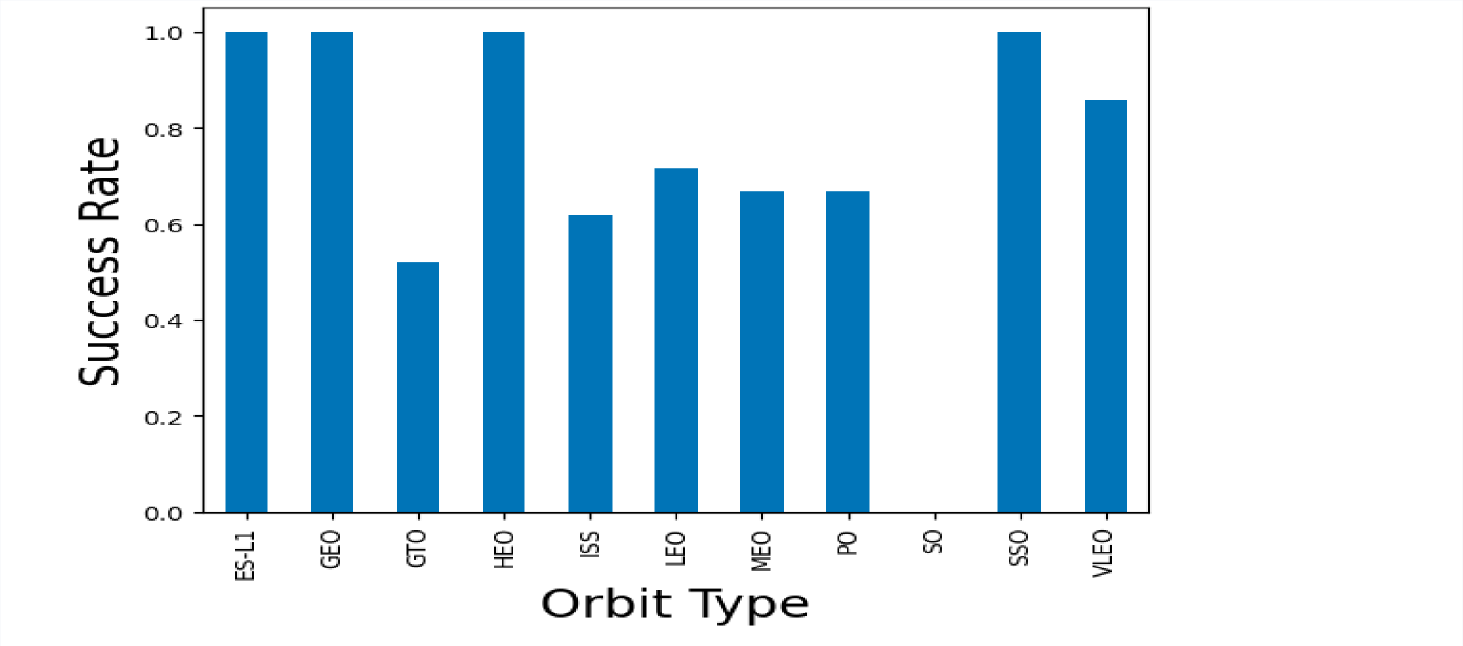
Flight Number vs. Launch Site, CCAFS SLC 40 appears to have been where most of the early 1st stage landing failure took place

Payload vs. Launch Site



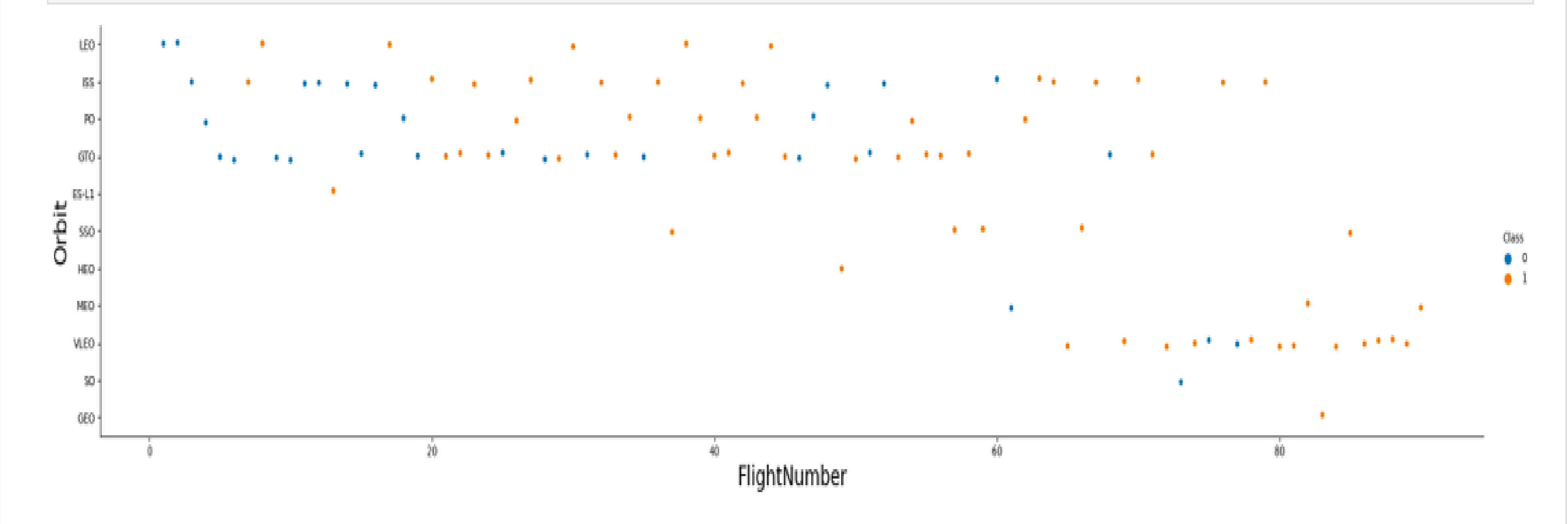
- Payload vs. Launch Site scatter point chart that find out that for the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000). As a result, the above scatter plot shown that CCAFS SLC 40 and KSC LC 39A appear to be favored to heavier payloads.

Success Rate vs. Orbit Type



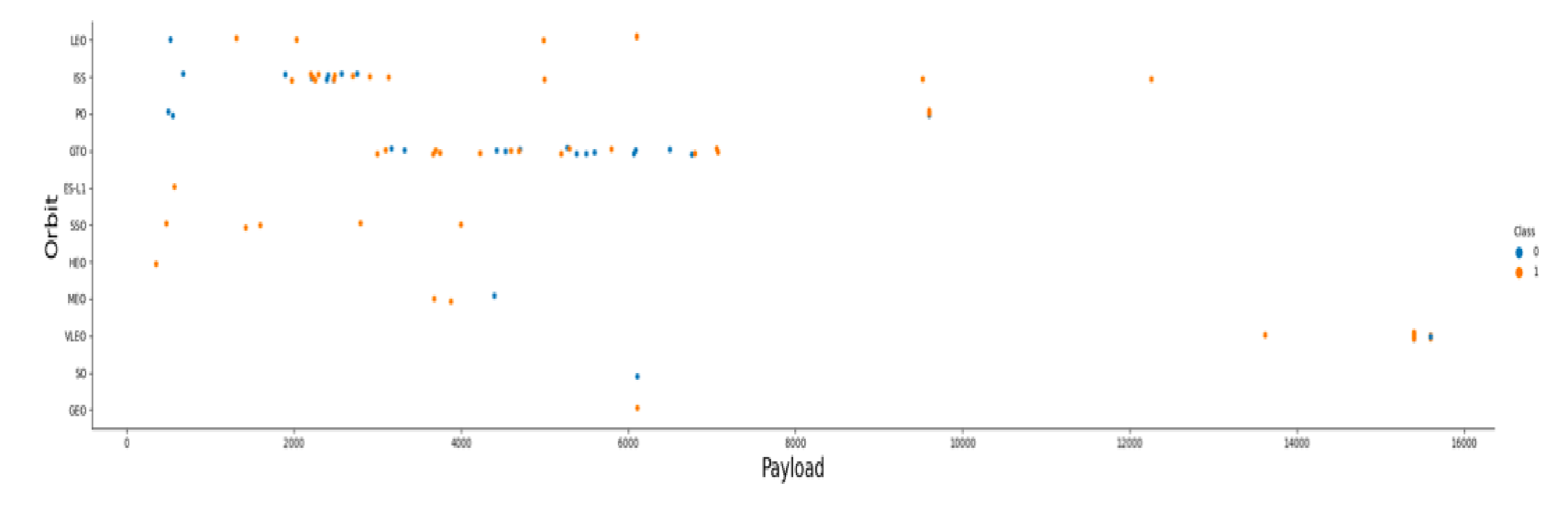
- Success Rate vs. Orbit Type bar chart shown that all orbit types except 'SO' have had successful 1st stage landings

Flight Number vs. Orbit Type



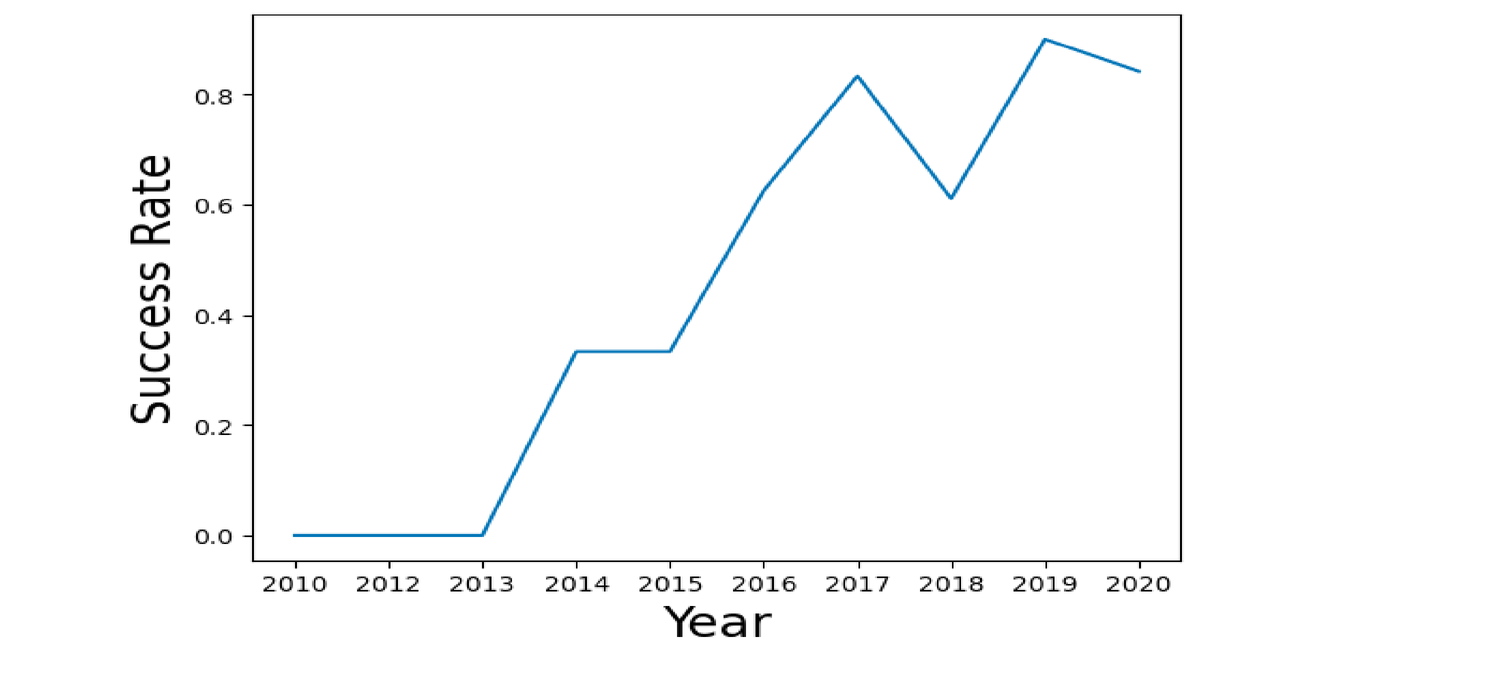
Flight number vs. Orbit Type, flight number positively correlated with 1st stage recovery for all orbit types.

Payload vs. Orbit Type



Payload vs. Orbit Type, heavier payloads have a negative influence on GTO orbits and positive influence on ISS orbits

Launch Success Yearly Trend



The success rate since 2013 kept increasing till 2020

All Launch Site Names

In [29]:
```
%%sql
SELECT DISTINCT LAUNCH_SITE
FROM SPACEXDATASET;
```

```
 * ibm_db_sa://kby38023:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[29]:

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

The Launch Sites has SpaceX used:

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

Launch Site Names Begin with 'CCA'

DATE	TIME__UTC_	BOOSTER_VERSION	LAUNCH_SITE	PAYLOAD
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2

SpaceX Data Set

Display 5 records where launch sites begin with the string 'CCA'

In [32]:

```
%%sql
SELECT LAUNCH_SITE
FROM SPACEXDATASET
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5;
```

```
 * ibm_db_sa://kby38023:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomair
Done.
```

Out[32]:

launch_site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Query Result

Examine Launch Site and date records where Launch Site begin with the string 'CCA':

- Last launch from CCAFS LC-40 was 2016-08-14
- First launch from CCAFS SLC-40 was 2017-12-15

Thus, from the query result above shown that the first 5 rows launch sites begin with string 'CCA' are all CCAFS LC-40

Total Payload Mass

Display the total payload mass carried by boosters launched by NASA (CRS)

In [42]:
```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_)
FROM SPACEXDATASET
WHERE Customer = 'NASA (CRS)';
```

```
 * ibm_db_sa://kby38023:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg
Done.
```

Out[42]:

1
45596

- The total payload mass carried by booster launched by NASA (CRS) are : 45,596 KG

Average Payload Mass by F9 v1.1

Display average payload mass carried by booster version F9 v1.1

In [21]:
```
%%sql
SELECT AVG("PAYLOAD_MASS__KG_")
FROM SPACEXDATASET
WHERE "BOOSTER_VERSION" LIKE '%F9 v1.1%';
```

```
 * ibm_db_sa://kby38023:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqblod8lcg.databases.appdomain.cloud:31198
Done.
```

Out[21]:

1
2534

The average payload mass carried by booster version F9 v1.1 is : 2534 KG

First Successful Ground Landing Date

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

In [22]:

```
%%sql
SELECT MIN(Date)
FROM SPACEXDATASET
WHERE Landing__Outcome = 'Success (ground pad)';
```

```
 * ibm_db_sa://kby38023:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.clou
Done.
```

Out[22]:

1
2015-12-22

The date when the first successful landing outcome in ground pad was achieved on 2015-12-22. It was more than 5 years after the first Falcon 9 launch on 2010-06-04.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
SELECT BOOSTER_VERSION
FROM SPACEXDATASET
WHERE LANDING__OUTCOME = 'Success (drone ship)'
   AND 4000 < PAYLOAD_MASS__KG_ < 6000;
```

```
 * ibm_db_sa://kby38023:***@0c77d6f2-5da9-48a9-81f8-86
Done.
```

booster_version
F9 FT B1021.1
F9 FT B1023.1
F9 FT B1029.2
F9 FT B1038.1
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1

The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000 are:

- F9 FT B1021.1
- F9 FT B1029.1
- F9 B4 B1042.1
- F9 B5 B1046.1
- F9 FT B1021.1
- F9 FT B1038.1
- F9 B4 B1045.1

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
In [24]:
%%sql
SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL_NUMBER
FROM SPACEXDATASET
GROUP BY MISSION_OUTCOME;
```

```
 * ibm_db_sa://kby38023:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90
Done.
```

Out[24]:

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

The total number of successful and failure mission outcomes are:

- 1 – Failure (in flight)
- 99 – Success
- 1 – Success (payload status unclear)

Boosters Carried Maximum Payload

The list of the names of the booster version which have carried the maximum payload mass:

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

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

```
 * ibm_db_sa://kby38023:***@0c77d6f2-5da9-
Done.
```

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

2015 Launch Records

26]:

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

```
 * ibm_db_sa://kby38023:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.b
Done.
```

26]:

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 record displays the failure landing outcome in drone ship, booster versions, and launch site for the year of 2015.

- Failure (drone ship) F9 v1.1 B1012   CCAFS LC-40
- Failure(drone ship) F9 V1.1 B1015   CCAFS LC-40

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

The rank of successful landing outcomes the date 04-06-2010 and 20-03-2017 in descending order:

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

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

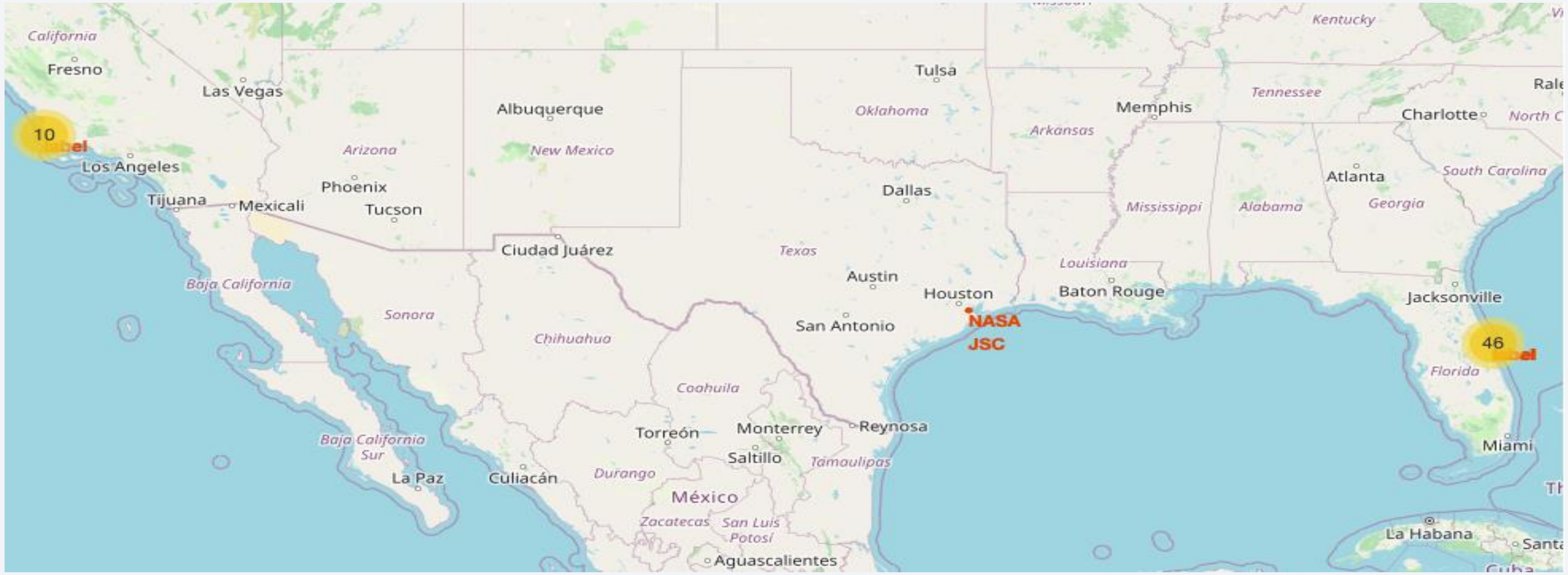
```
 * ibm_db_sa://kby38023:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2ic
Done.
```

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

Section 3

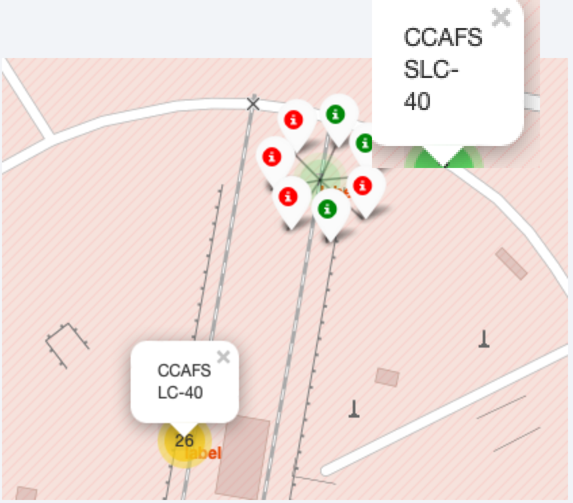
Launch Sites Proximities Analysis

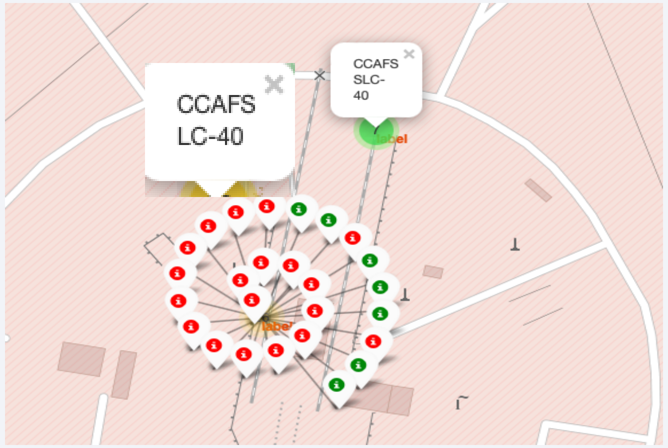
ALL LAUNCH SITE ON MAP

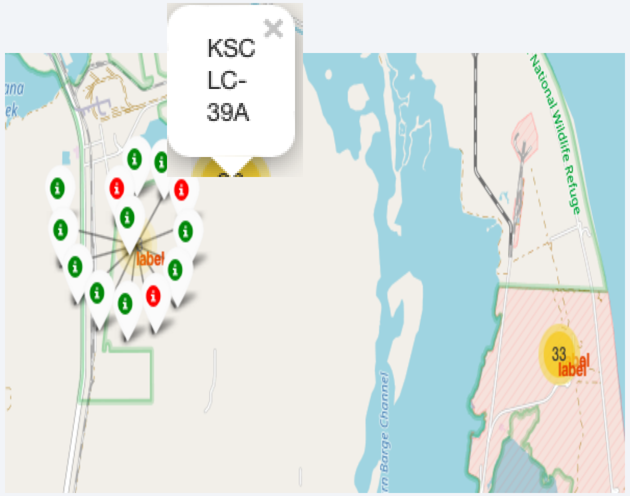


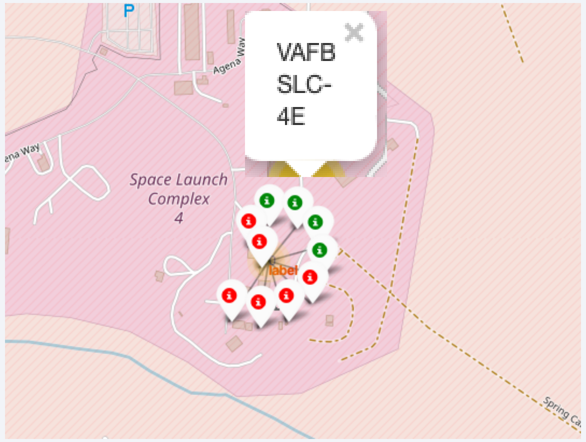
The launch sites on a map highlights the importance of the launch site proximity to the coast and equator.

THE SUCCESS/FAILED LAUNCHES FOR EACH SITE



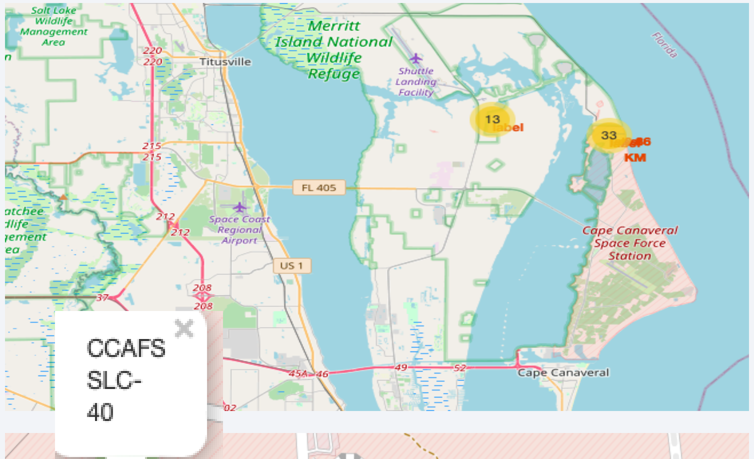


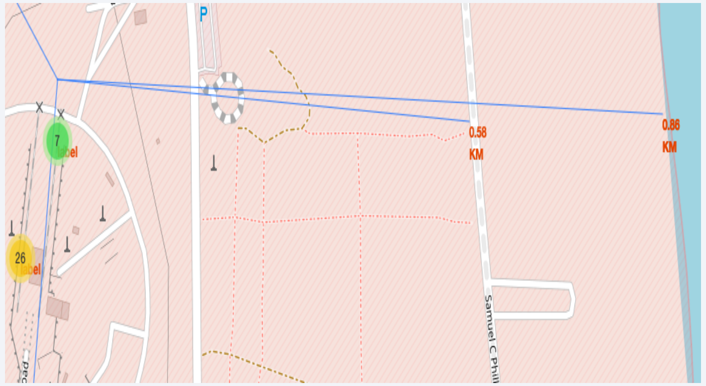


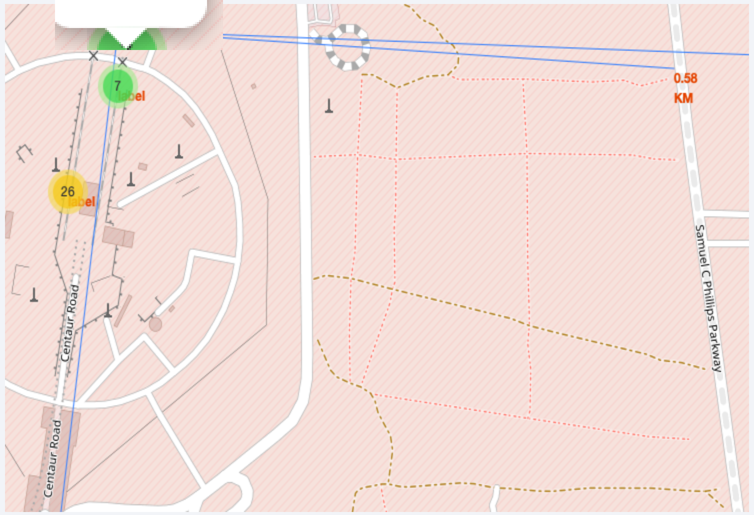


From the color-labeled markers in marker clusters, KSC LC-39A launch site have relatively high success rates.

THE DISTANCES BETWEEN A LAUNCH SITE





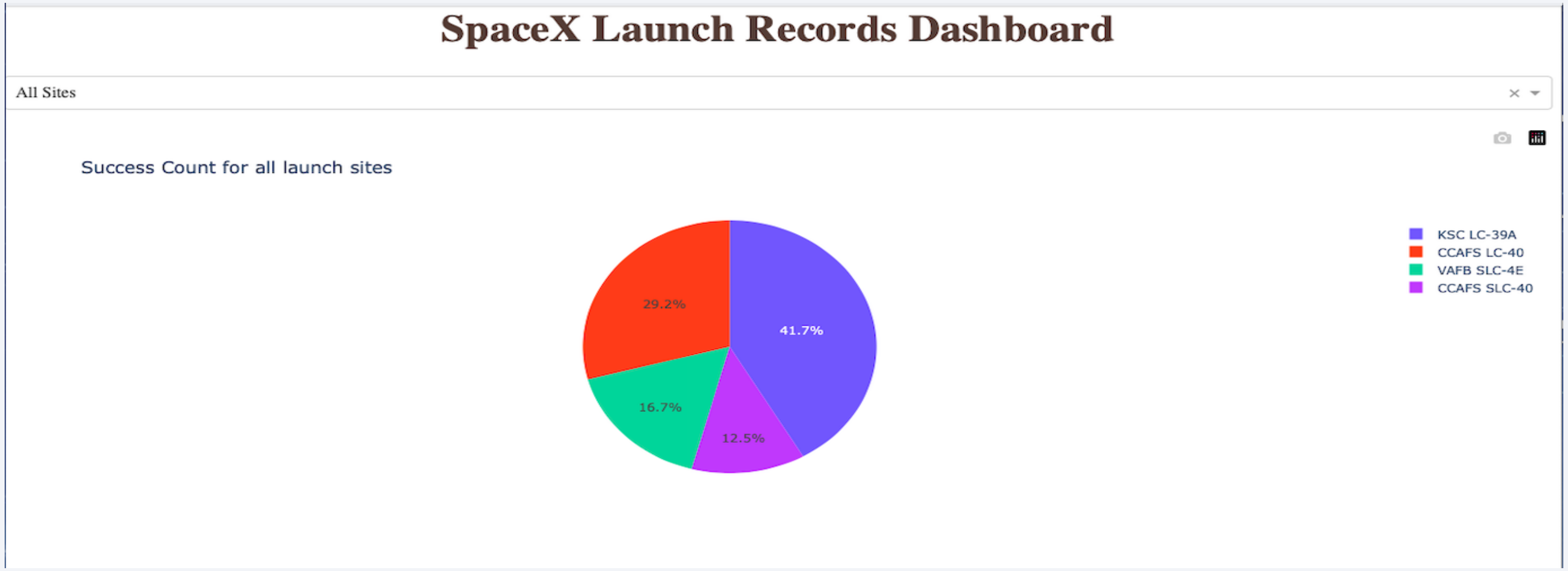


- The launch sites for CCAFS SLC-40 is close to railways at 1.24 km
- The launch sites for CCAFS SLC-40 is close to highways at 0.58 km
- The launch sites for CCAFS SLC-40 is close to coastline at 0.86 km
- The launch sites for CCAFS SLC-40 does keep certain distance away from cities at 51.48 km

Section 4

Build a Dashboard with Plotly Dash

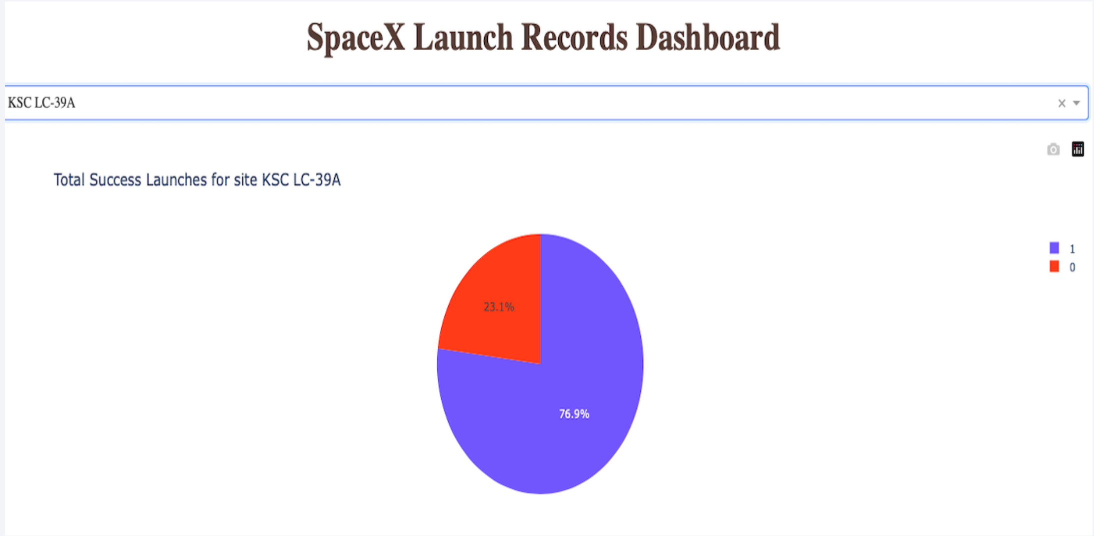
SUCCESS COUNT FOR LAUNCHES SITE

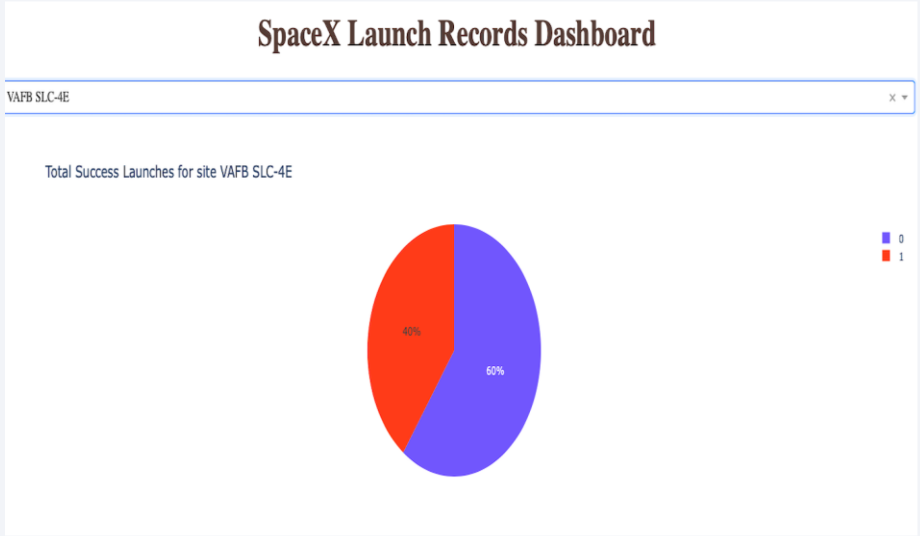


The pie chart showing the booster landing success rate for all launch sites.

- KSC LC-39A has 41.7%
- CCAFS LC-40 has 29.2%
- VAFB SLC-4E has 16.7%
- CCAFS SLC-40 has 12.5%

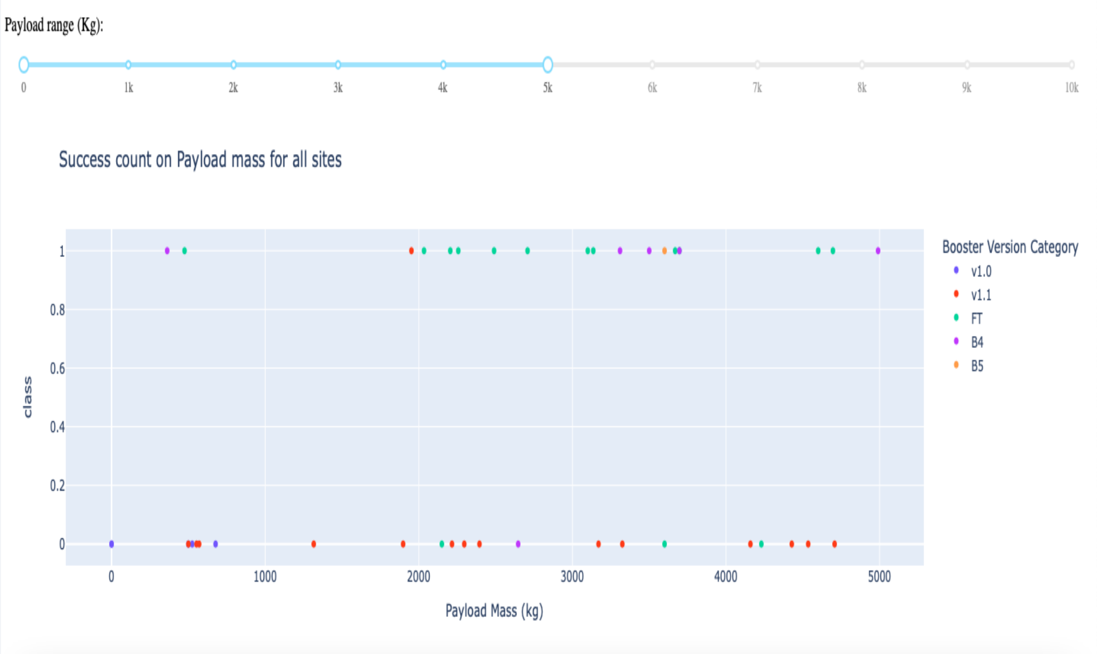
LAUNCH SITE WITH HIGHEST LAUNCH SUCCESS RATIO



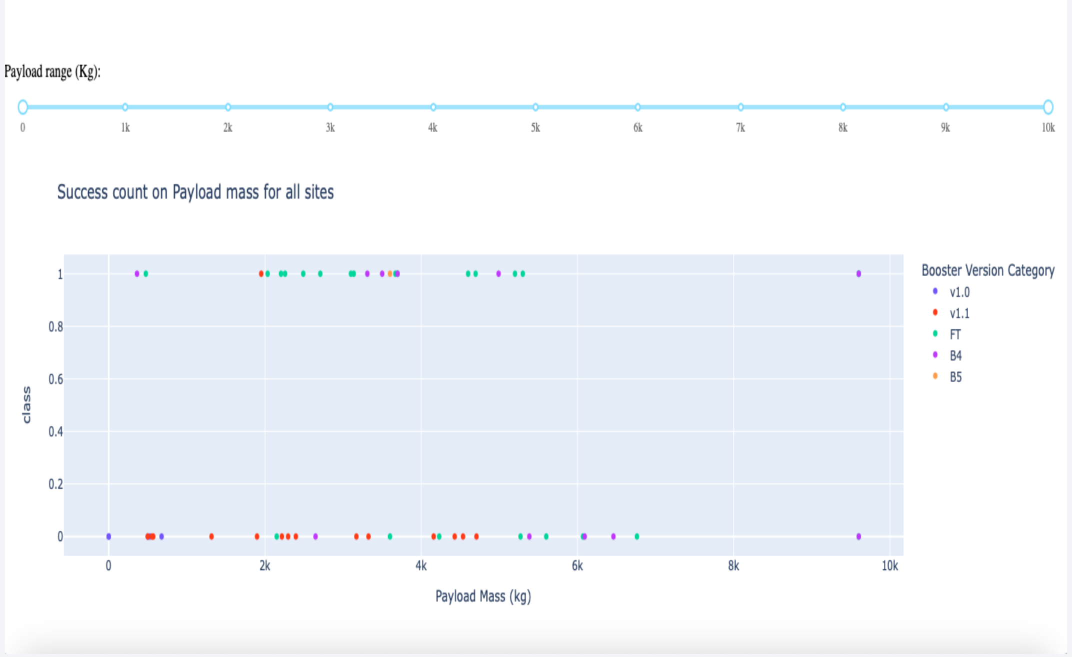


- VAFB SLC-4E has the heaviest successful booster landing at 60% success rate.
- KSC LC-39A has the highest booster landing 76.9% success rate.

PAYLOAD vs. LAUNCH OUTCOME FOR ALL SITES



Low weighted payload (0-5000kg)



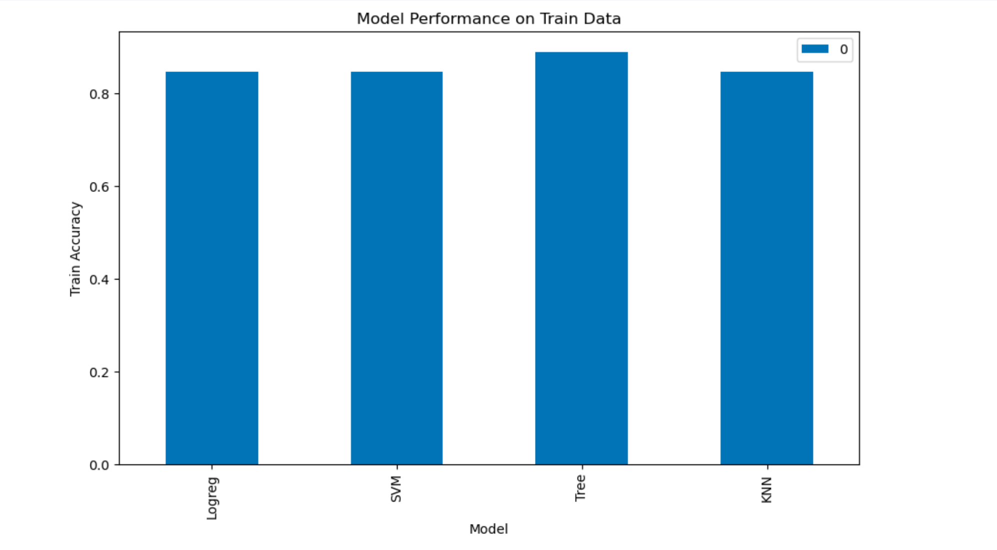
Heavy weighted payload (5000-10000kg)

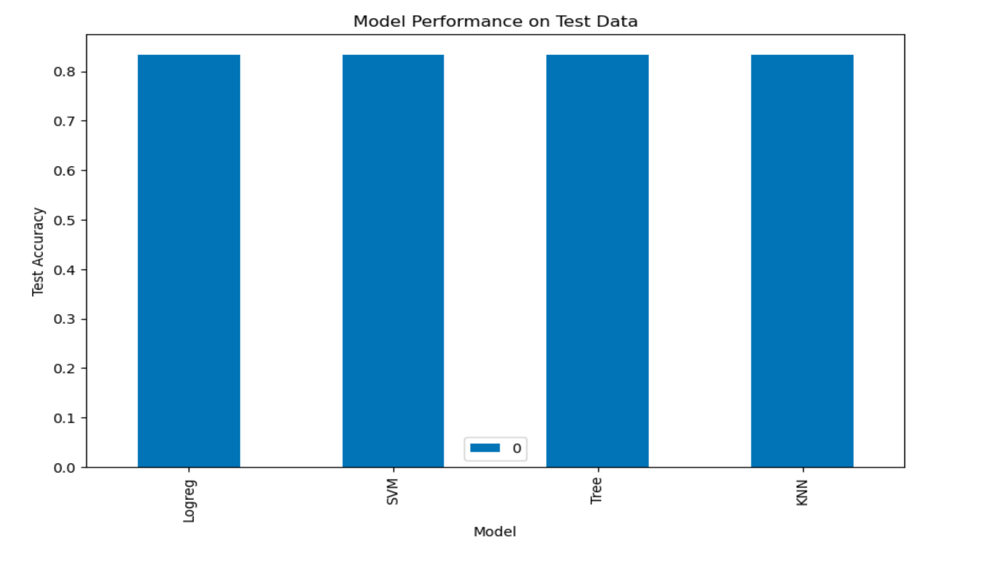
Low weighted payloads have a better success rate than the heavy weighted payloads

Section 5

Predictive Analysis (Classification)

Classification Accuracy

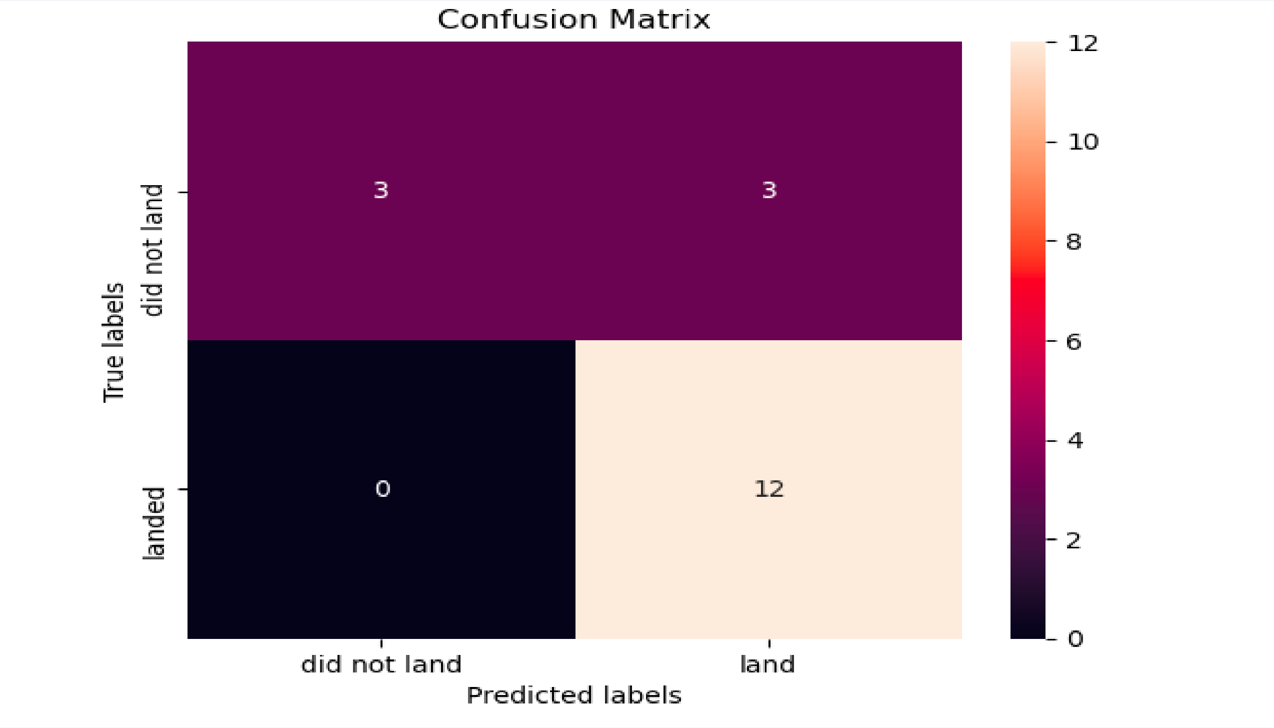




Model	TrainAccuracy	TestAccuracy
Logreg	0.84722	0.83333
SVM	0.84722	0.83333
Tree	0.88889	0.83333
Knn	0.84722	0.83333

For the accuracy test result, all models performed with similar results at 83.33%.

Confusion Matrix



By examining the Confusion Matrix, all the models performed the same results. However, the major problem is False Positives as a result the models incorrectly predicting the1st stage booster landing in the test set.

Conclusions

- Using the models from the report of SpaceY can predict when SpaceX will successfully land the 1st stage with the accuracy of 83.33%.
- This will enable SpaceY to make more information bids against SpaceX, this included the cost and R & D information.
- The success of a mission can be explained by several factors such as the launch site, the orbit, and the previous number of launches. From all these information gather together, we can have a better understanding of the success or failure on each launch.
- Most of the mission outcomes are successfu. However, successful landing outcomes will still need to be improve over time, according to the new rockets strcuture and new launch site.
- From this whole project, the best launch site is KSC LC-39A and the model that predict the higher test accuracy is Decision Tree Classifier.

Appendix

References :

- https://www.spacex.com/vehicles/falcon-9/

Acknowledgement:

- Thank you to IBM for offer this Professional Certificate course
- Thank you to Coursera for providing this Professional Certificate course
- Thank you to GitHub for allowing to create respository for this project

Thank you!