IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

FREDERICKV
ALBORNOZ
ESCOBEDO
19th December, 2023

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- Prediction if the Falcon 9 first stage will land successfully
- Collecting data on the Falcon 9 first-stage landing
- See interactive Plotly Dash
- Applying Machine Learning to determine if the first stage of Falcon 9 will land successfully

Introduction

- 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.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Request to the SpaceX API
  - Clean the requested data
- Perform data wrangling
  - Extract a Falcon 9 launch records HTML table from Wikipedia
  - Parse the table and convert it into a Pandas dataframe
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Create a column for the class, Standardize the data, Split into training data and test data, Find the method perform best using test data

Data Collection

We collected dataset about SpaceX using:

- SpaceX REST API
  - Response JSON and normalizing the data to turn it into a Pandas dataframe
  - Getting information about launches using the IDs given for each launch
  - Applying getBoosterVersion tocreate a Pandas data frame
  - Dealing with Missing Values
- Wikipedia
  - Request the Falcon9 Launch Wiki page from its URL
  - Extract all column/variable names from the HTML table header using BeautifulSoup
  - Create a data frame by parsing the launch HTML tables

Data Collection – SpaceX API

Dataset is collected using the SpaceX API, using ger requests in Python. Then we normalized JSON data. After that we converted dictionary to DataFrame, filtered it and clean missing values.

Results shown here: Data Collection – SpaceX API

Getting data in JSON format and its normalizing

↓

Getting information using the IDs for new columns

↓

Creating new data frame and combining it to dictionary

↓

Filtering the data to get only about Falcon 9

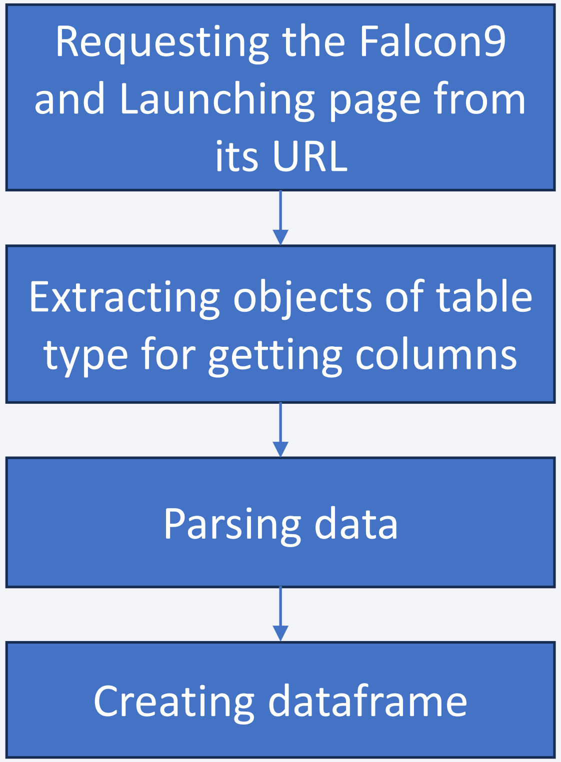
↓

Finding missing values of dataset and dealing with it

Data Collection - Scraping

First, we did the request to get content using BeautifulSoup. After that we extract all column names from HTML table header and created a data frame by parsing the HTML.

Results shown here: Data Collection - Scraping



Data Wrangling

- For data wrangling we performed Exploratory Data Analysis (EDA) and found some patterns in data:
  - Identify and calculate the percentage of the missing values in each attribute, also identify types of columns
  - Calculate the number of launches on each site
  - Calculate the number and occurrence of each orbit
  - Calculate the number and occurrence of mission outcome of the orbit
  - Create a landing outcome label from Outcome column
- Results shown here: Data wrangling

EDA with Data Visualization

- As a result, we got visualization of the relationship between Flight Number and Launch Site, Payload and Launch Site, success rate of each orbit type, FlightNumber and Orbit type, Payload and Orbit type. Visualize the launch success yearly trend.
- Also, we made some features engineering:
  - Create dummy variables to categorical columns
  - Cast all numeric columns to `float64` using one hot encoding
- Results shown here: Data wrangling

EDA with SQL

After connecting to database, we performed EDA with SQL by executing queries to get some insights from data:

- unique launch sites
- launch sites which begin with ‘CCA’
- total payload mass carried by boosters launched by NASA (CRS)
- average payload mass carried by booster version F9 v1.1
- the date when the first successful landing outcome
- the names of the boosters which have success in drone ship
- the total number of successful and failure mission outcomes
- the names of the booster_versions which have carried the maximum payload mass
- the month names, failure landing_outcomes in drone ship, booster versions, launch_site for the months in year 2015
- the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20

- Results shown here: EDA with SQL

Build an Interactive Map with Folium

- During building an interactive map with Folium we used:
  - **Circles:** To mark region where is something placed
  - **Markers:** To highlight the name of circles
  - **Marker Cluster:** To simplify the map as it contains a lot of markers in the same or almost the same coordinates
  - **Mouse Position:** To get a coordinates of position where mouse placed at a map
  - **Polyline:** To draw a line from coordinate1 to coordinate2 and show distance
- Adding this interactive things, we simplify understanding of map and our results
- Results shown here: Launch Sites Locations Analysis with Folium

Build a Dashboard with Plotly Dash

- During building a Dashboards we added:
  - a dropdown list which shows all launch sites
  - a pie chart to show the total successful launches count for all sites and for specific launch site show the Success vs. Failed counts for the site
  - a slider to select payload range
  - a scatter chart to show the correlation between payload and launch success
- This plots and interaction allows us quickly visualize very useful information.
- Results shown here: Application with Plotly Dash

Predictive Analysis (Classification)

- After loading our datasets, we normalize it and split it on test and on train datasets.
- After that we make experiments and trying to fit our data to different methods of ML. We created:
  - Logistic Regression Model
  - Support Vector Machine
  - Decision Tree
  - K-neighbors Classifier
- After that we calculated the score and show confusion matrix of each method
- Results shown here: Machine Learning Prediction

Results

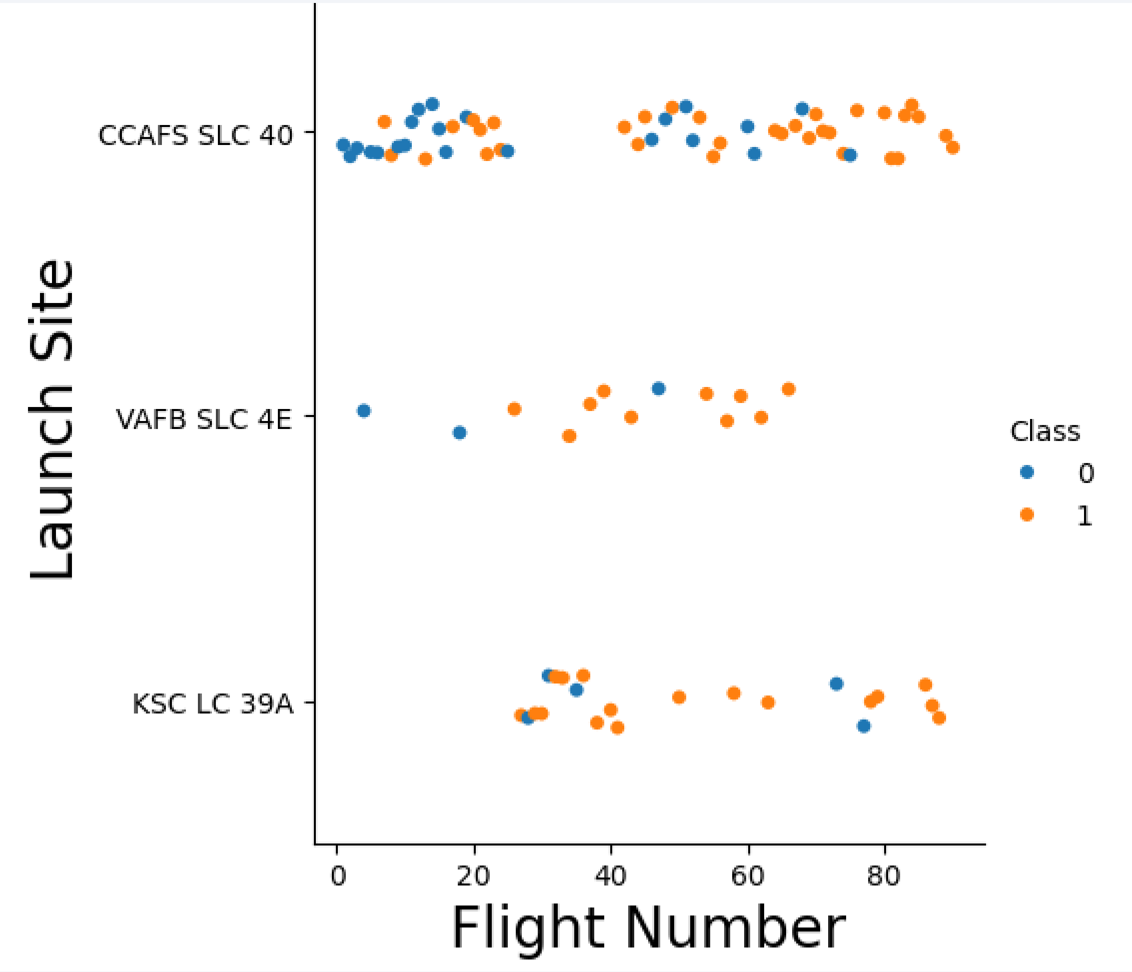
- Exploratory data analysis results
  - The more flights, the more payload mass needed
  - Rate of success depends on orbit type and number of flights
  - Depending on orbit type, we can understand how many payload mass we need
  - Success rate grown with newer technologies
- Interactive analytics
  - Launch sites close to Equator, coastline, railway and highway, but not close to cities
- Predictive analysis results
  - The best results of prediction shows Decision Tree method and the lowest accuracy have KNN

Section 2

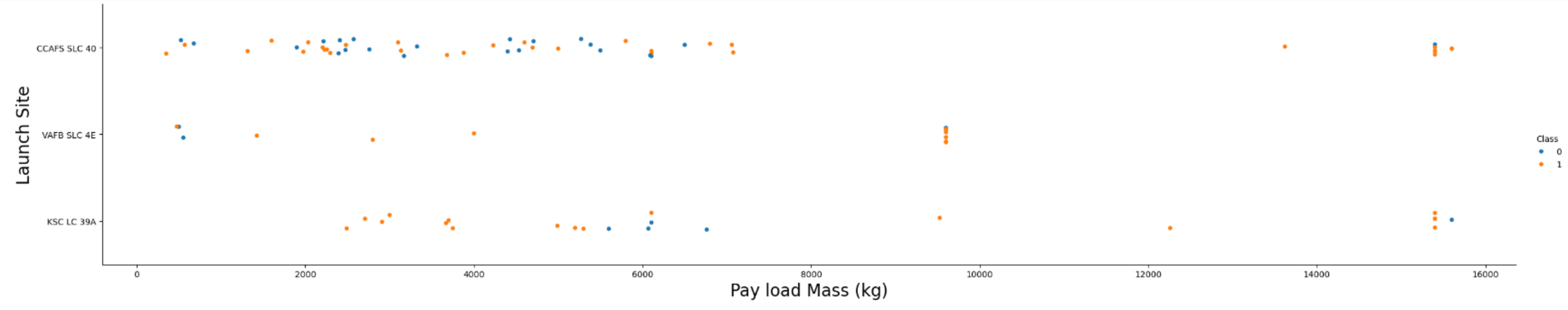
Insights drawn from EDA

Flight Number vs. Launch Site

- The highest number of flights was made from CCAFS SLC 40. At the same time, this launch site have the highest number of flights. A lot of flights was failure.



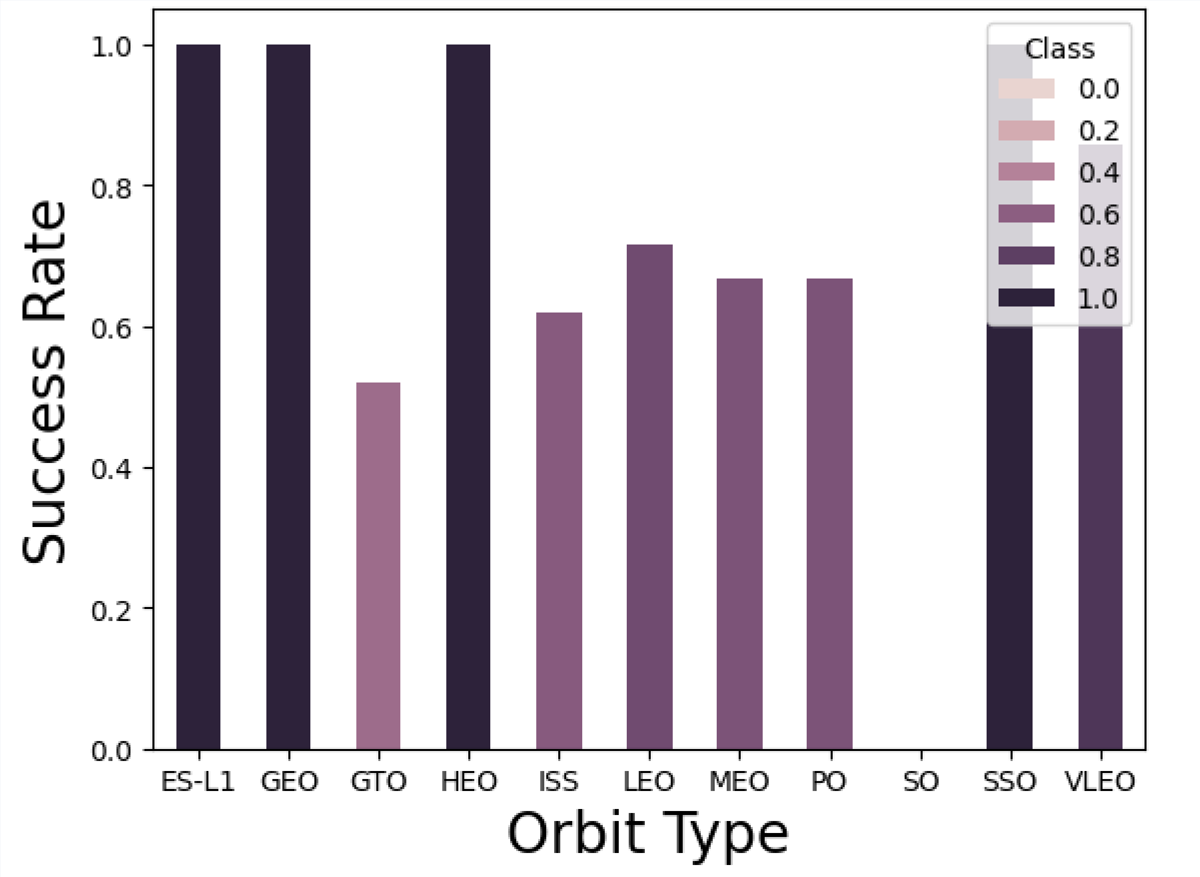
Payload vs. Launch Site



- This scatter shows that most of flights have payload mass less then 8000 kg. Launch site called VAFB SLC 4E did not have payload mass more then 10 000 kg.

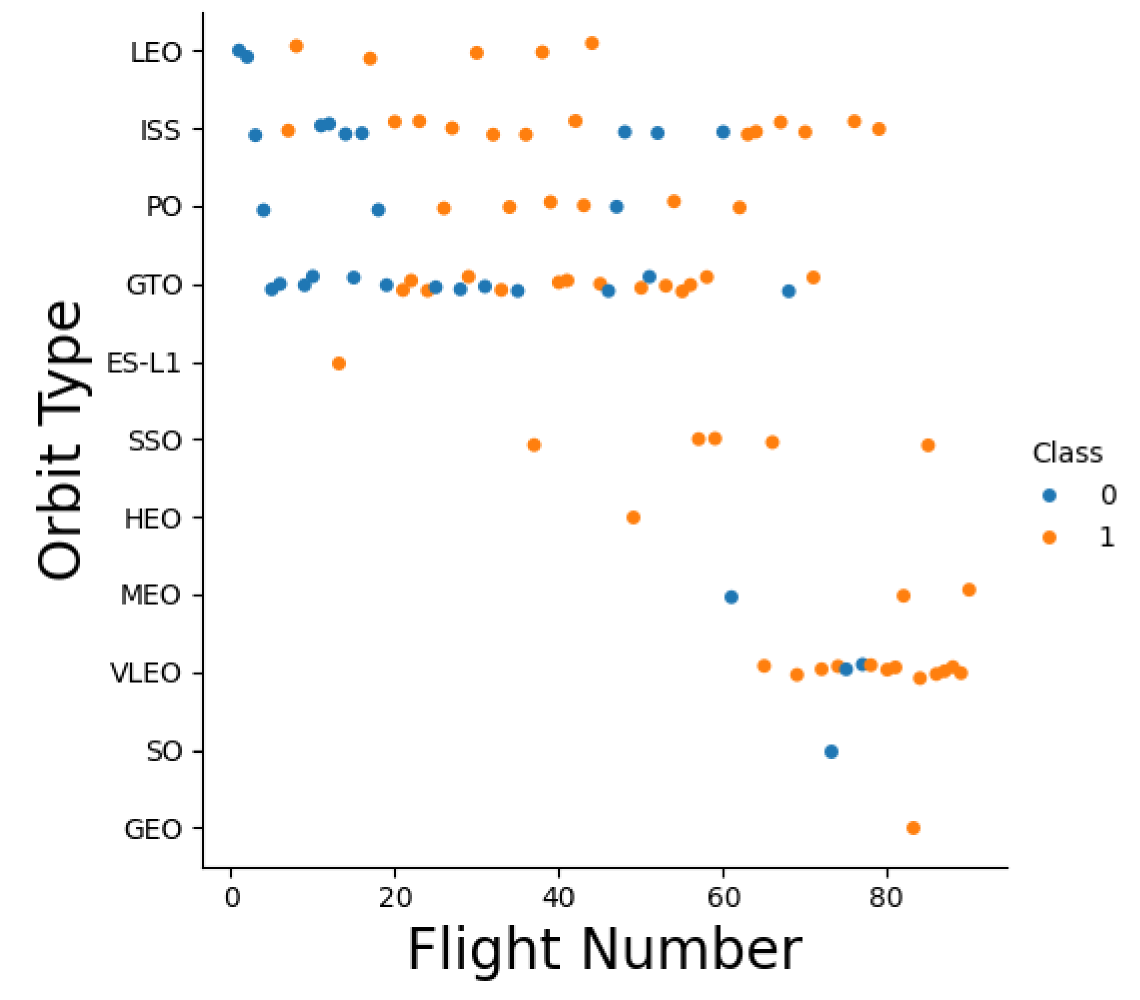
Success Rate vs. Orbit Type

- This bar chart shows that orbit SO does not have success rate. ES-L1, GEO, HEO and SSO orbit types have 100% success rate.



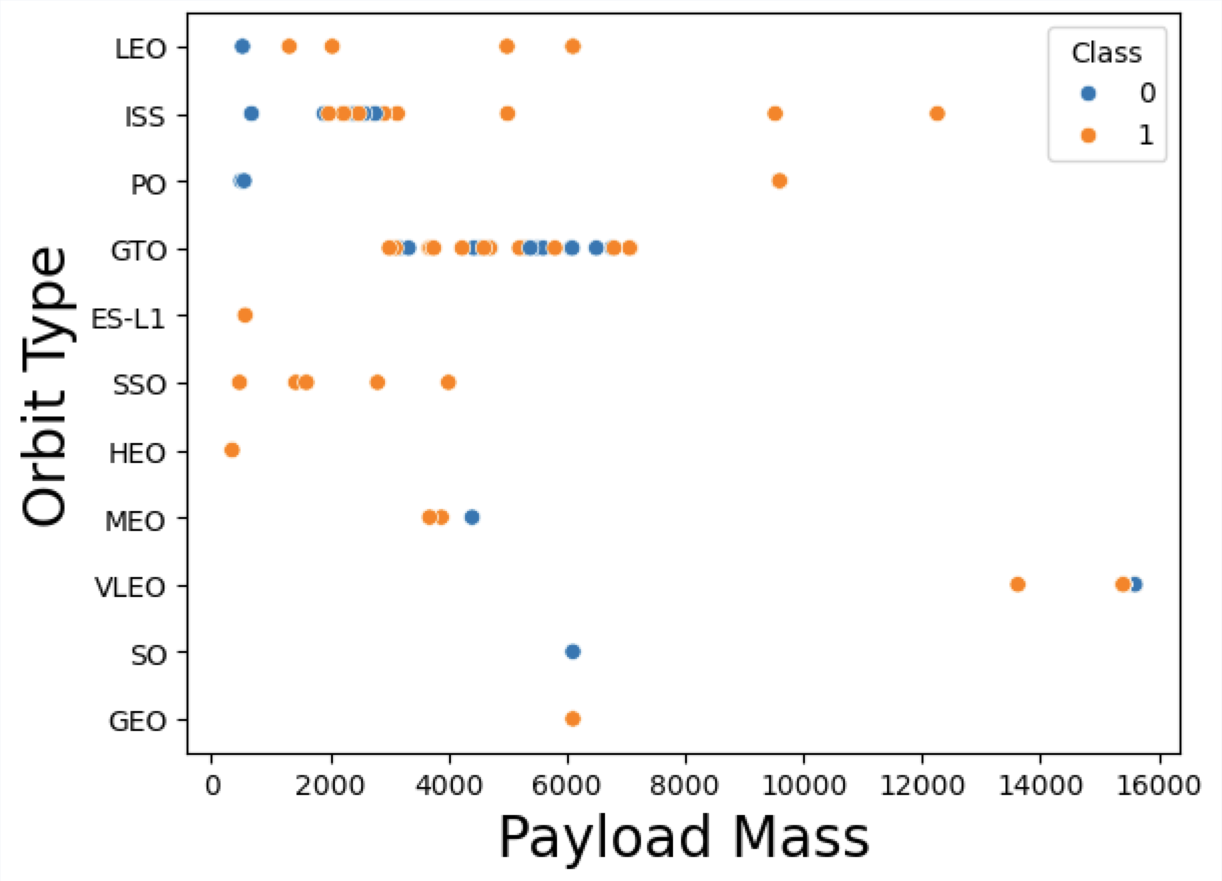
Flight Number vs. Orbit Type

- Flights to LEO, ES-L1, HEO and SSO orbits with increasing number of flights become successful. At the same time, flights to GTO, ISS, PO and VLEO orbits did not depends on flight number.



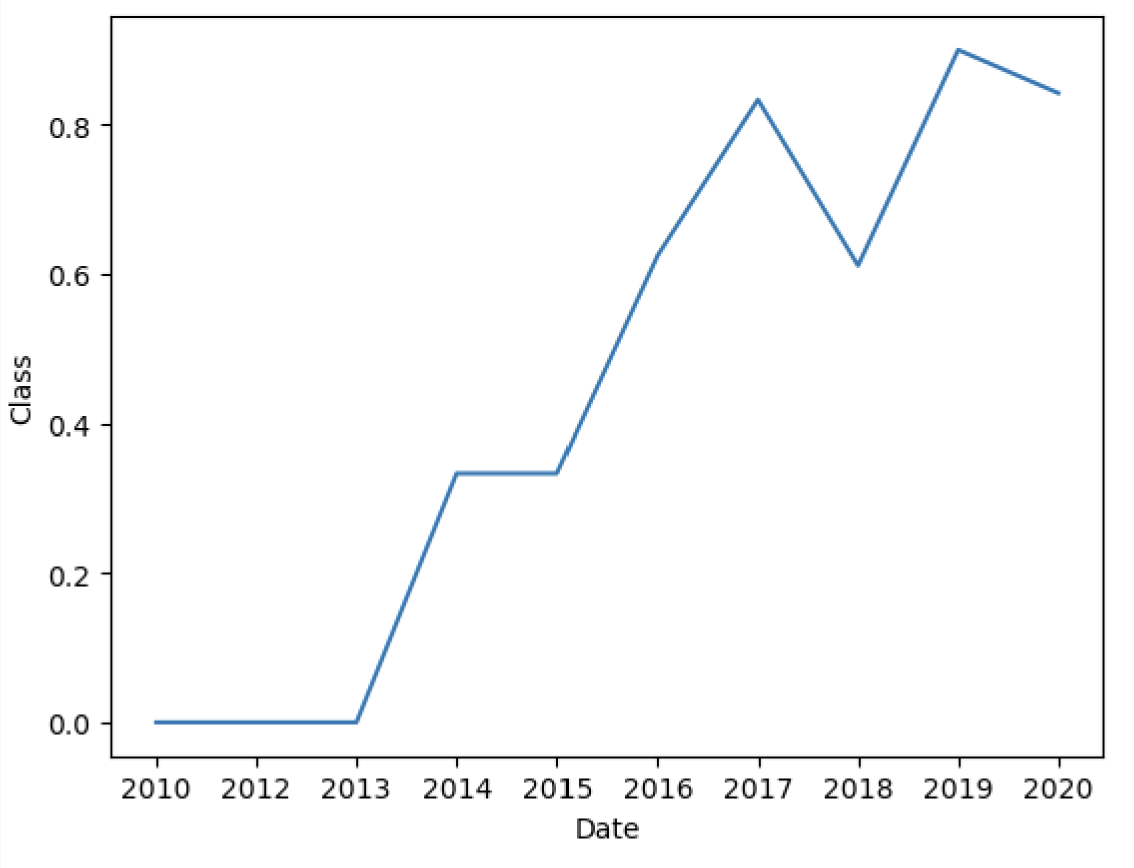
Payload vs. Orbit Type

- Flights to most of orbits need less payload mass then 8 000.
- Flight to VLEO needs more then 12 000 payload mass.



Launch Success Yearly Trend

- From 2010 to 2013 rate was fully failure.
- From 2013 rate becomes bigger and bigger yearly.



All Launch Site Names

- The names of the unique launch sites

```
%%sql
SELECT DISTINCT(Launch_Site) FROM SPACEXTBL
```

```
 * sqlite:///my_data1.db
Done.
```

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`

```
%%sql
SELECT * FROM SPACEXTBL
WHERE Launch_Site LIKE "CCA%"
LIMIT 5
```

```
 * sqlite:///my_data1.db
Done.
```

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

Total Payload Mass

- The total payload carried by boosters from NASA

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) as "Total payload mass", Customer from SPACEXTBL
WHERE Customer = "NASA (CRS)"
```

```
 * sqlite:///my_data1.db
Done.
```

Total payload mass	Customer
45596	NASA (CRS)

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) as "Average payload mass", Booster_Version from SPACEXTBL
WHERE Booster_Version = "F9 v1.1"
```

```
 * sqlite:///my_data1.db
Done.
```

Average payload mass	Booster_Version
2928.4	F9 v1.1

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad

```
%%sql
SELECT MIN(Date), Landing_Outcome from SPACEXTBL
WHERE Landing_Outcome = 'Success'
```

```
 * sqlite:///my_data1.db
Done.
```

MIN(Date)	Landing_Outcome
2018-07-22	Success

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
%%sql
SELECT Booster_Version from SPACEXTBL
WHERE Landing_Outcome = 'Success (drone ship)' AND (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000)
```

```
 * sqlite:///my_data1.db
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes

```
%%sql
SELECT COUNT(Mission_Outcome) as 'Total number', Mission_Outcome from SPACEXTBL
GROUP BY Mission_Outcome
```

```
 * sqlite:///my_data1.db
Done.
```

Total number	Mission_Outcome
1	Failure (in flight)
98	Success
1	Success
1	Success (payload status unclear)

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass

```
%%sql
SELECT Booster_Version from SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

```
 * sqlite:///my_data1.db
Done.
```

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

2015 Launch Records

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

```
%%sql
SELECT SUBSTR(Date, 6, 2) as 'Month', Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTBL
WHERE Landing_Outcome = 'Failure (drone ship)' AND SUBSTR(Date, 0, 5)='2015'
```

```
 * sqlite:///my_data1.db
Done.
```

Month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- 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

```
%%sql
SELECT COUNT(Landing_Outcome) as 'outcome', Landing_Outcome FROM SPACEXTBL
WHERE Landing_Outcome = 'Failure (drone ship)' OR Landing_Outcome = 'Success (ground pad)' AND
        Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY 'outcome' DESC
```

```
 * sqlite:///my_data1.db
Done.
```

outcome	Landing_Outcome
3	Success (ground pad)
5	Failure (drone ship)

Section 3

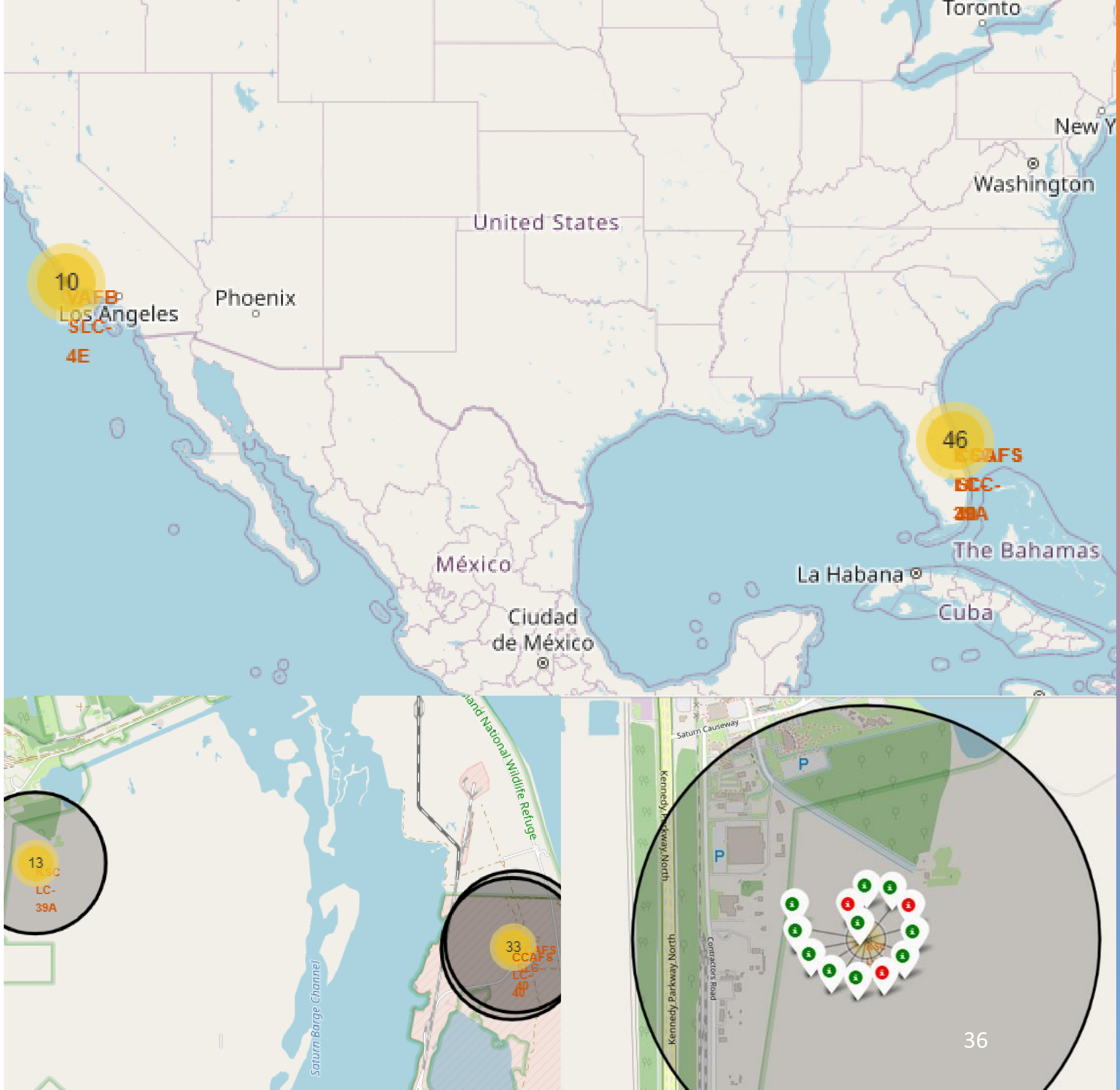
Launch Sites Proximities Analysis

All launch sites

- There is shown all launch sites. As we can see, they placed close to Equator line and to the coasts.

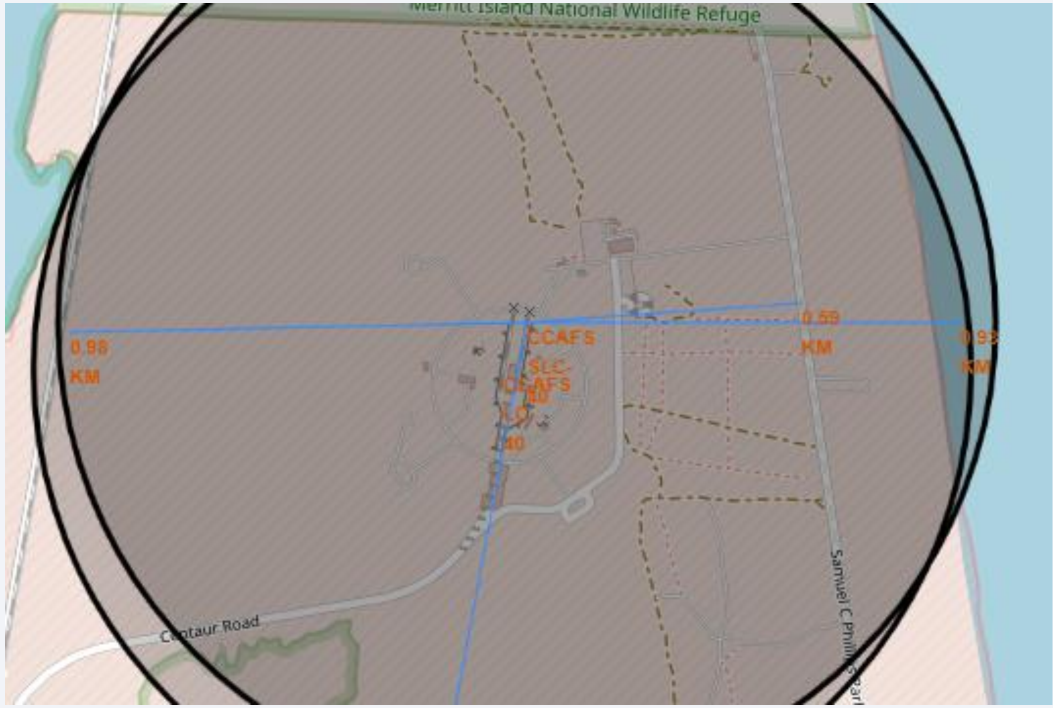
Color-labeled outcomes

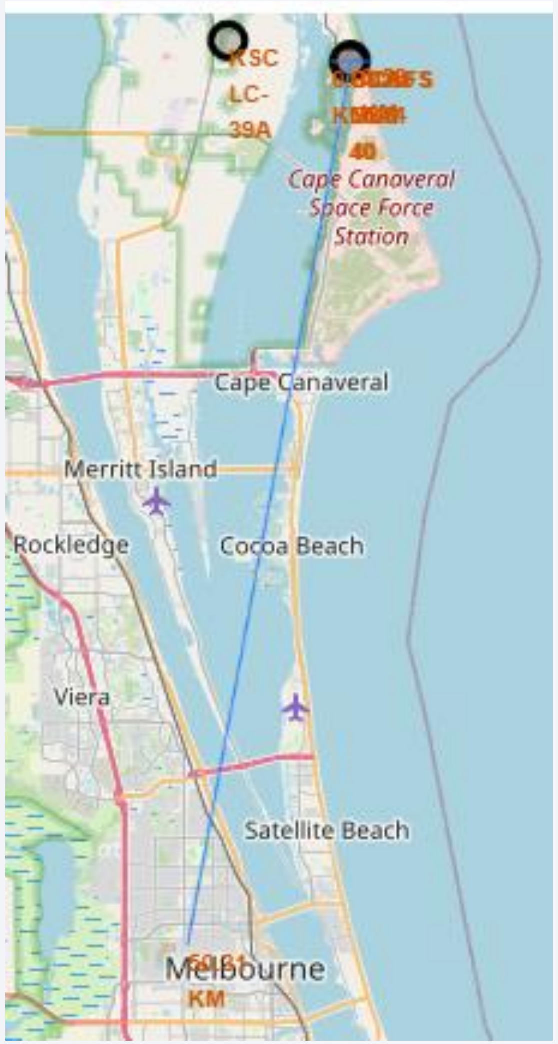
- From the color-labeled markers, we are be able to easily identify which launch sites have relatively high success rates.
- Green markers = success and red markers = failure



Distance from a launch site

- Here is shown distance from one launch site to closest railway, highway and coastline. Also, here is shown distance to Malbourne.



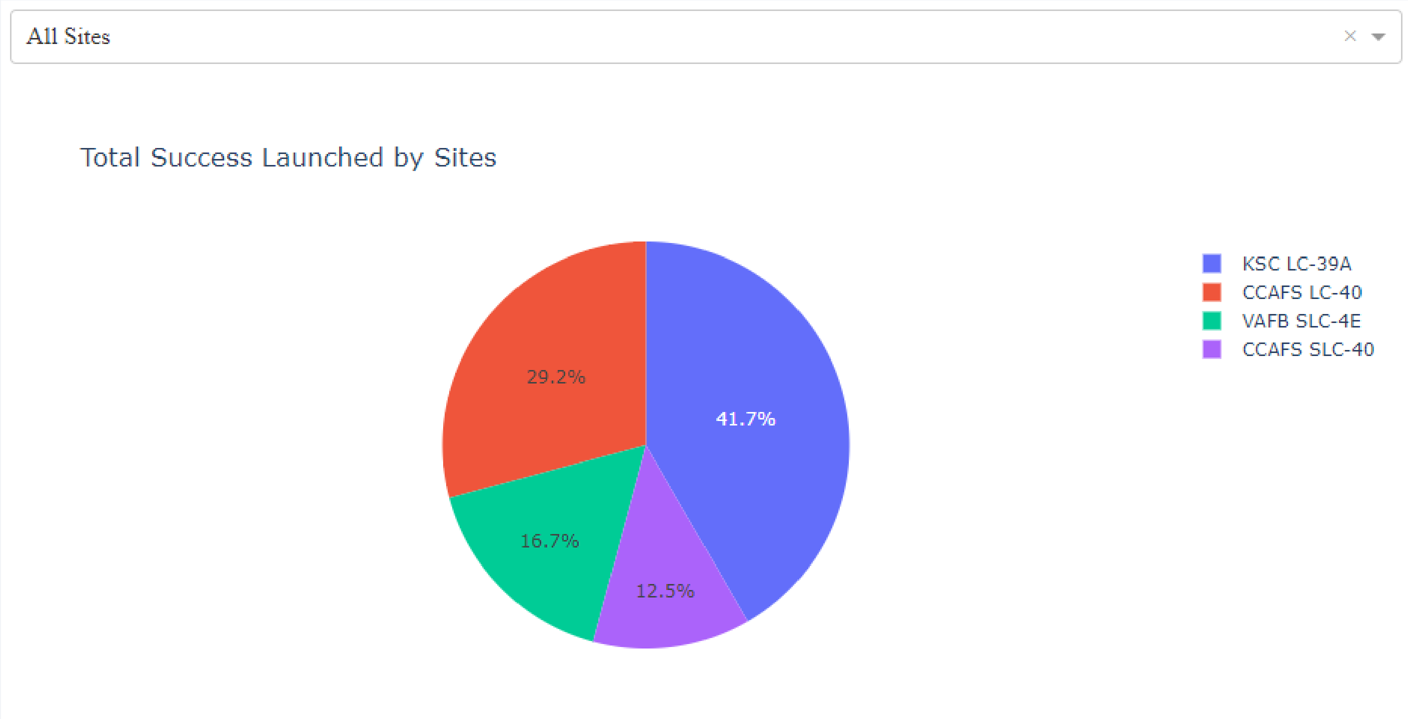


Section 4

Build a Dashboard with Plotly Dash

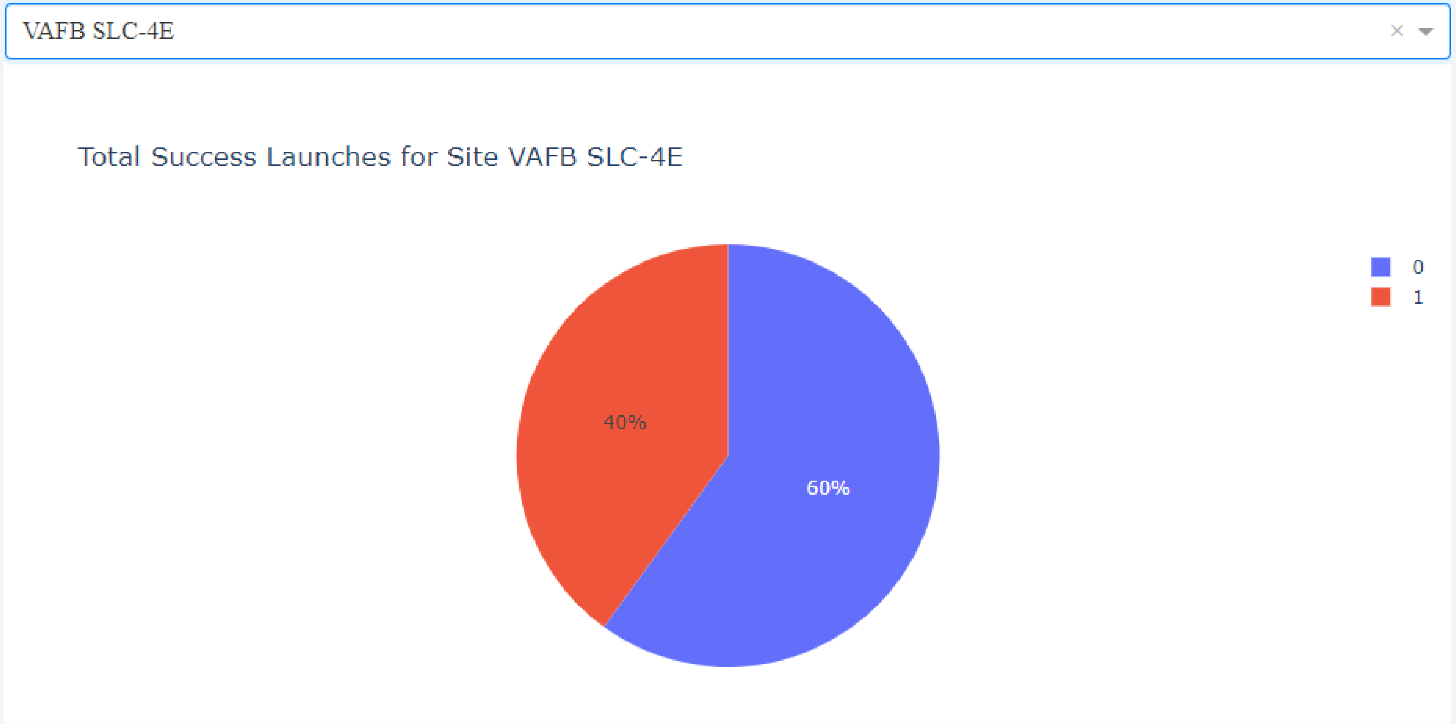
Total success launched by sites

- The pie chart shows percentage of successful launch comparing with other sites



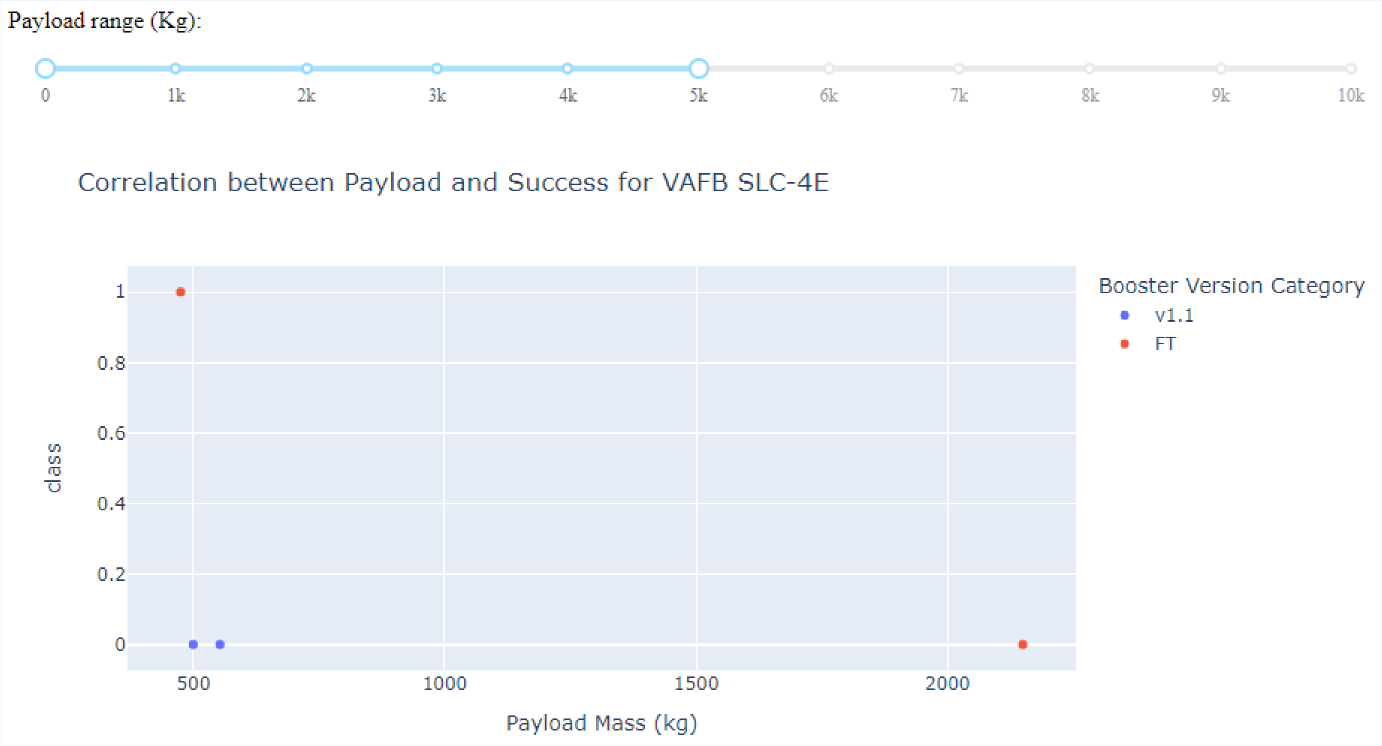
Total Success Launches for Site VAFB SLC-4E

- Here is shown percentage of successful launch comparing to failure for VAFB SLC-4E. Successful rate is 40%



Correlation between Payload and Success for VAFB SLC-4E

- Here is shown outcomes for VAFB SLC-4E for payload from 0 to 5000 kg. Also, shown for what booster version is success or failure.

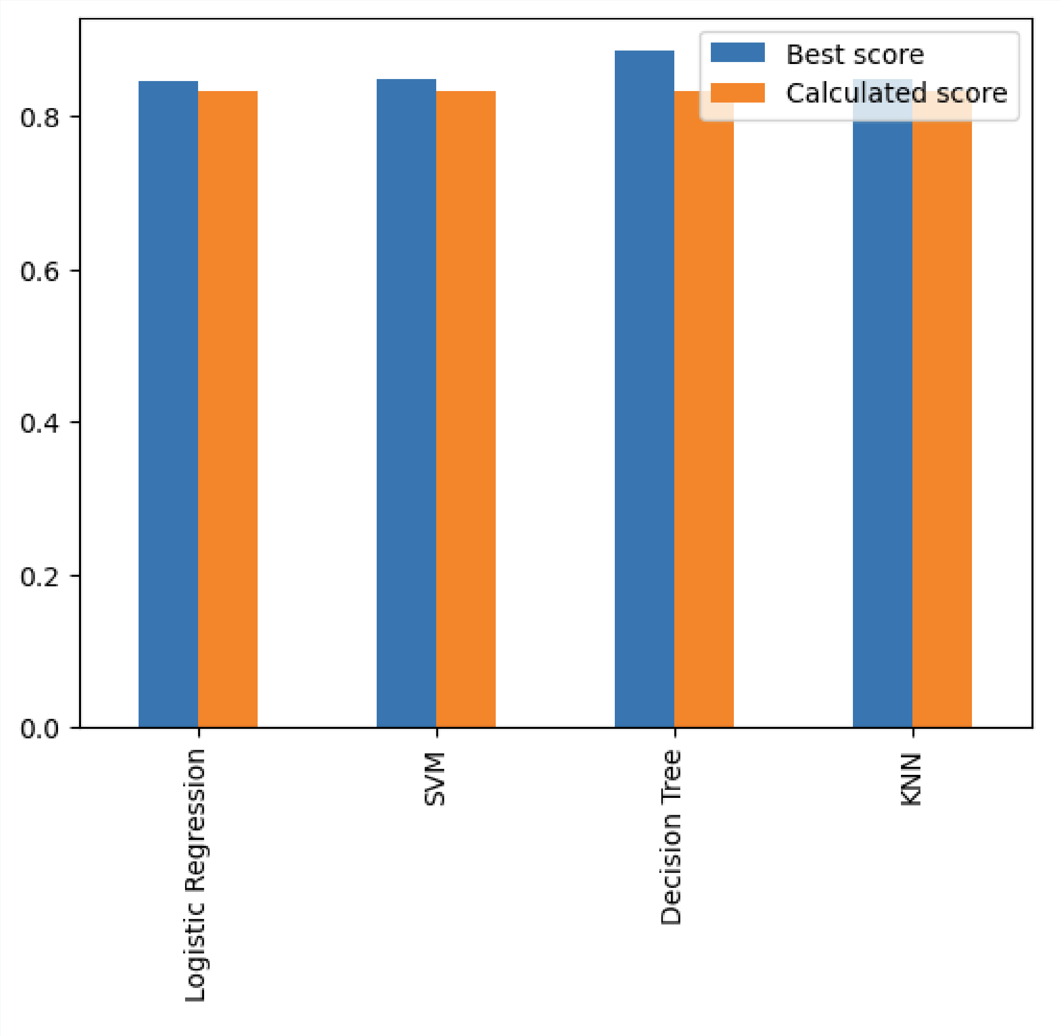


Section 5

Predictive Analysis (Classification)

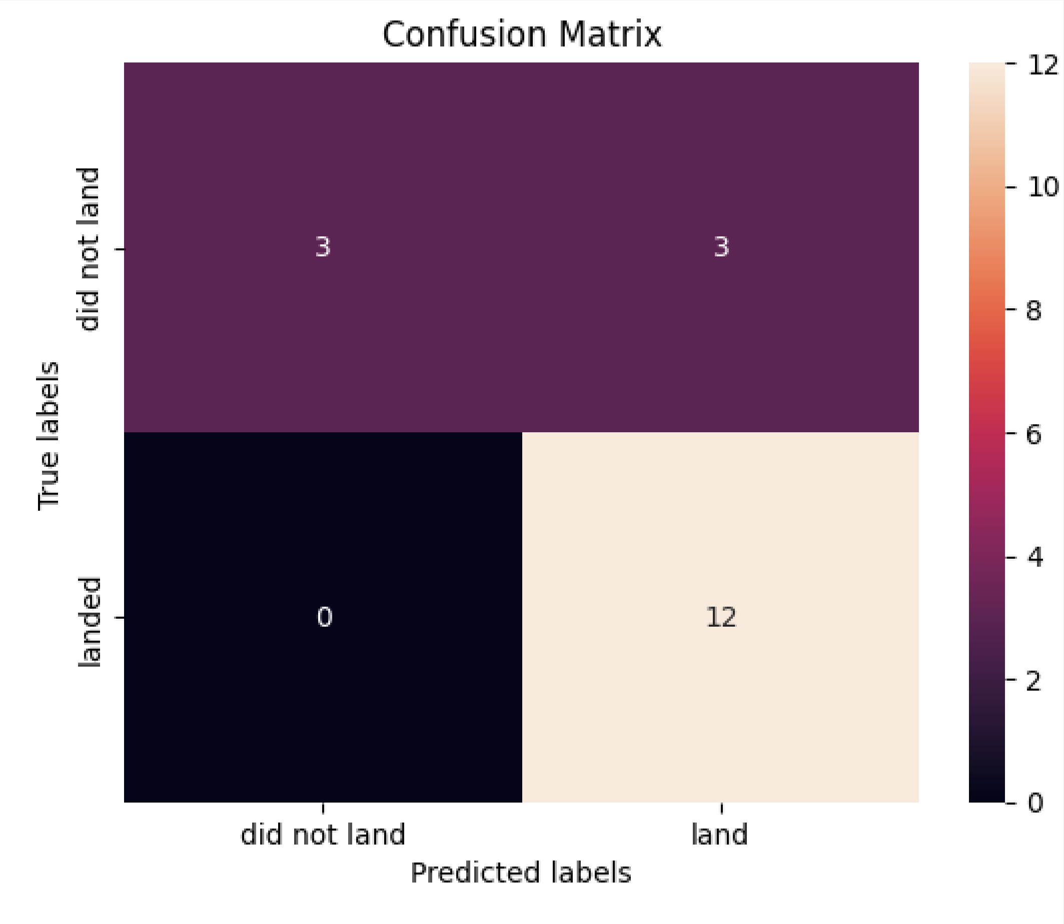
Classification Accuracy

- As we can see, all test score are same. But the best accuracy shows Decision Tree Classifier with best score.



Confusion Matrix

- Prediction about landing absolutely predicted well, however it is hard to predict failure landing.



Conclusions

- KSC LC-39A rate have most of successful launches.
- Decision Tree Classifier works better for prediction.
- All of sites close to Equator and seacoasts.
- Success of mission can depend on orbit type.
- Since 2013 percentage of successful flights increasing.

Thank you!