IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ezechias Mirindi

April, 2025

Outline

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

Executive Summary

Aerospatiale companies struggle with building sophisticated spacecrafts as they bear huge costs. To minimize the construction costs, SpaceX wants to reuse the first stage of Falcon 9 for its next space mission. This project aims at predicting whether that first stage will be launch and land successfully based on the past launch data of Falcon 9 rockets.

For this project, data was collected using SpaceX API and Webscraping. After data collection, Data went through the wrangling process for cleaning it. To understand the data, we performed Exploratory Data analysis by writing SQL requests, visualizing data and analyzing data on map using Folium library. To make predictions we used four Machine learning algorithms – Logistic Regression, Support Vector Machine, Decision Tree and K Nearest Neighbors.

The results showed that launches were successful as time passed and with bigger payloads. Most of the launch sites were near coastlines to reduce the risk of impacting population and infrastructures during launching operations. Finally, after training the algorithms, The Decision Tree resulted to be the best algorithm for SpaceX data with a prediction accuracy of 94% on the test data.

Introduction

Nowadays, space exploration has become a true passion for some big firms like Galactic, Rocket-Lab, Blue Origin, etc. Some of these firms build and send spacecrafts to space stations and one of the most successful companies is SpaceX. As competition grows among companies, the latter wants to explore space by optimizing their costs.

This project aims at providing cost information to companies who want to bid against SpaceX. SpaceX will use the first stage of falcon 9 for its next mission. If we can predict whether the first stage will land successfully then we can determine the cost of the launch. And this information can be important to competitors.

Section 1

Methodology

Methodology

- Data collection methodology
- Data wrangling
- Exploratory data analysis (EDA) using visualization and SQL
- Interactive visual analytics using Folium and Plotly Dash
- Predictive analysis using classification models

Data Collection

Data from SpaceX was obtained from two sources:

- **SpaceX REST API**

  [SpaceX API link]

- **WebScraping**

  [Wikipedia link]

Data set view

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial
0	1	2006-03-24	Falcon 1	20.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin1A
1	2	2007-03-21	Falcon 1	NaN	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2A
2	4	2008-09-28	Falcon 1	165.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2C
3	5	2009-07-13	Falcon 1	200.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin3C
4	6	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003

Data collection steps

Step 1: Collect Data from SpaceX API and Convert data to .json file

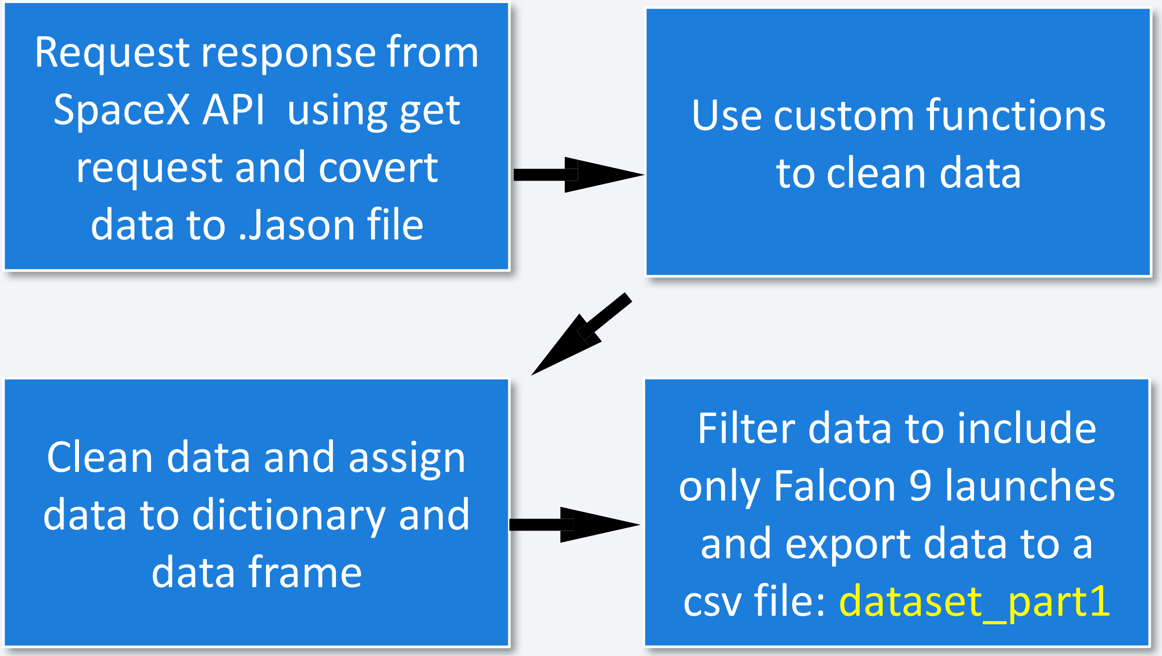
Step 2: Scrap and filter data to include Falcon 9 data, assign data to dataframe and dictionary, and export data to a csv file

Step 3: Plot and visualize the data

Data Collection – SpaceX API

- We target the API endpoint, perform a get request, call the json function to view data, Convert json object into a dataframe and finally save the dataset

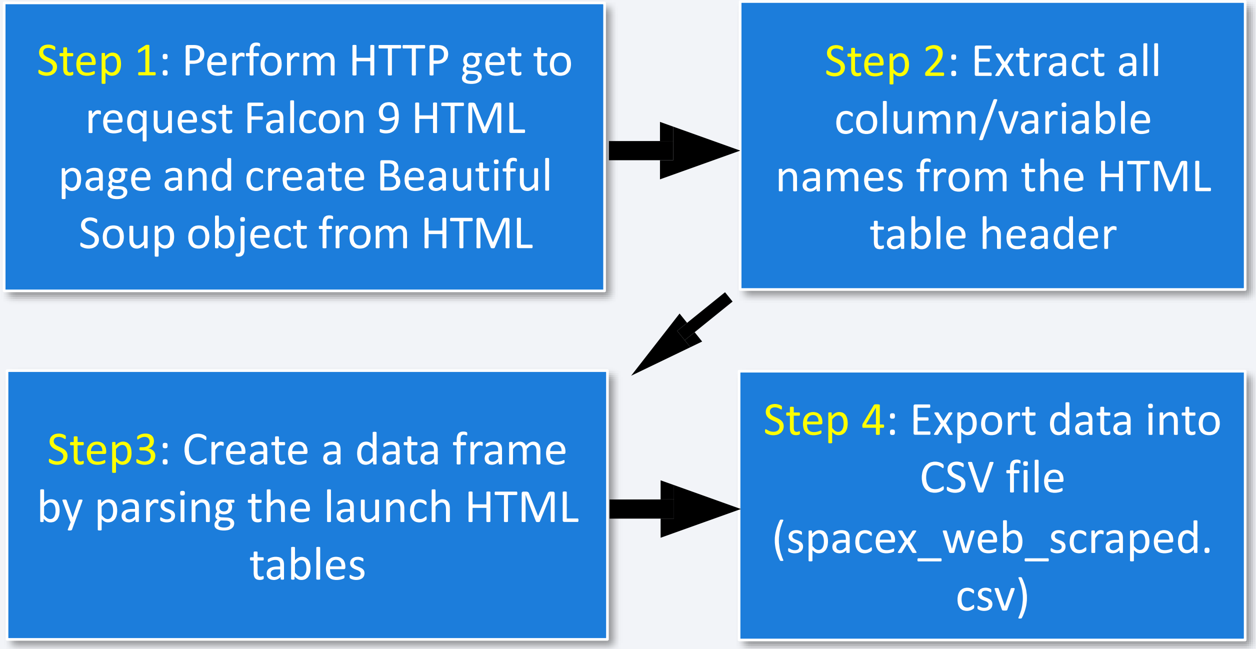
- [Data Collection – SpaceX API]()



Data Collection - Scraping

- We perform a HTTP get request, create a Beautiful Soup object from the HTML response, Extract the variables, create a dataframe from the HTML table and save the dataset as a csv file.

- Data Collection – Scraping



Data Wrangling

Steps

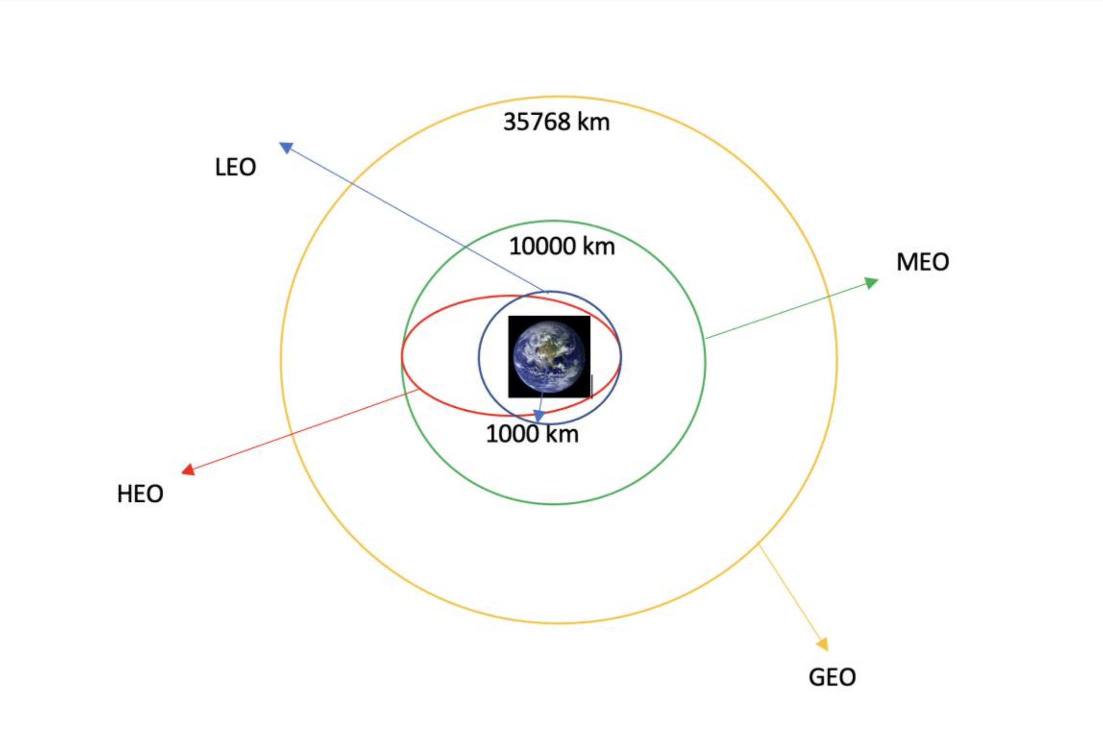
Step 1: Load data from dataset_part1.csv file and calculate the number of launches on each site

Step 2: Calculate the number and the occurrence of each orbit

Step3: Calculate the number and occurrence of mission outcome of the orbits

Step 4: Create a landing outcome label from outcome column and export data into dataset_part2.csv file

Below are orbits found in the dataset



[Data wrangling]

EDA with Data Visualization

Step 1: Visualize the relationship between flight number and launch site

Step 2: Visualize the relationship between payload and launch site

Step3: Visualize the relationship between success rate of each orbit type

Step 4: Visualize the relationship between flight number and orbit type

Step 5: Visualize the relationship between payload and orbit type

Step 6: Visualize the launch success yearly trend

Step 7: Create dummy variable to categorical columns

Step 8: Cast all numeric columns to float64

[EDA with visualization steps]()

EDA with SQL

Step 1: Display the names of the unique launch sites in the space mission

Step 2: Display 5 records where launch sites begin with the string ‘CCA’

Step 3: Display average payload mass carried by booster launched by NASA (CRS)

Step 4: Display average payload mass carried by booster version F9 v1.1

Step 5: List the date when the first successful landing outcome in ground pad was achieved

Step 6: List the names of the boosters which have success in drone ship and mass > 4000 & <6000

Step 7: List the total number of successful and failure mission outcomes

Step 8: List the names of the booster_ve which have carried the max payload mass

Step 9: List the records which display the month, failure landing, booster version ..etc.

Step 10: Rank the count of landing outcomes or success

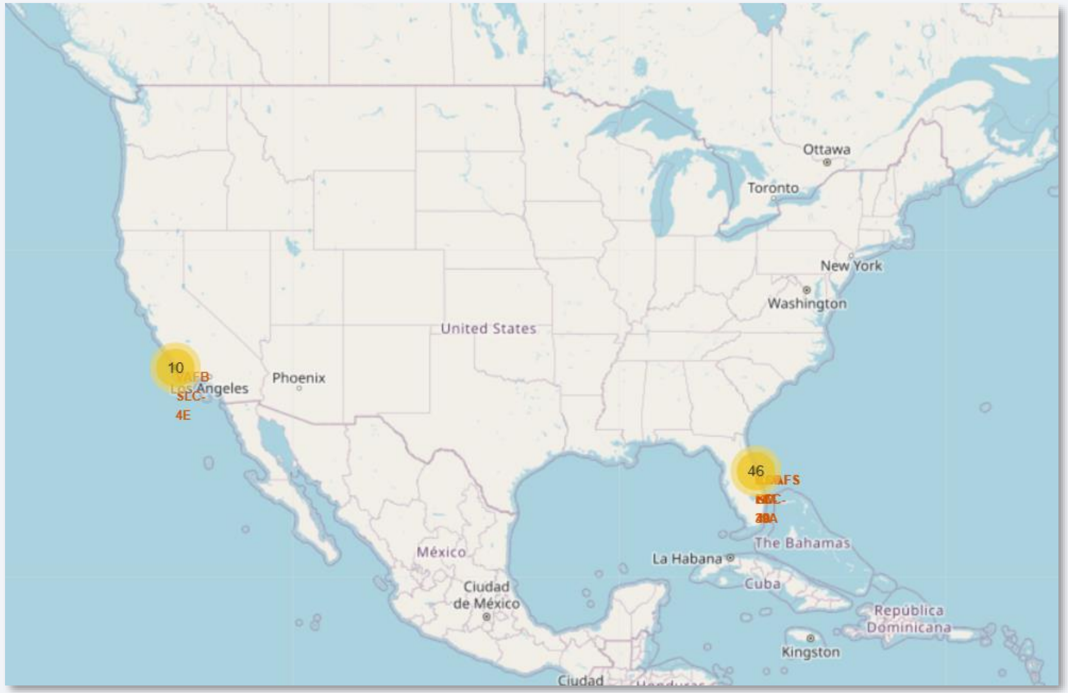
EDA with SQL

Build an Interactive Map with Folium

Step 1: Mark all launch sites on a map created using Folium by adding markers with circle , popup label and text label to each site using its longitude and latitude coordinates to show the geographical location approximately to the equator.

Step 1: Mark the success/failed launches for each site on the map using colored markers.

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



The launch sites are identified with yellow circles

[Visual Analytics with Folium]

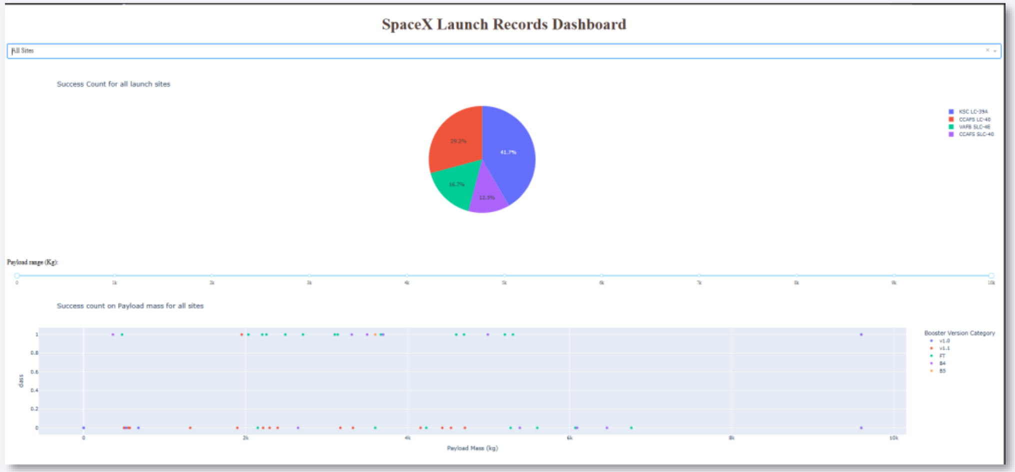
Build a Dashboard with Plotly Dash

Step 1: Add dropdown list to enable launch site selection

Step 2: Add pie chart to show the total successful launches count for all sites and the success vs. failed counts

Step 3: Add a range slider to select payload

Step 4: Add a scatter chart of payload mass vs. success rate of different booster versions



The dashboard shows the number of successful launches for each site

[SpaceX Dash]

Predictive Analysis (Classification)

- Step 1: Load the dataframe
- Step 2: Create Numpy array from the column class in the data
- Step 3: Standardize the data in X and assign it variable Y
- Step 4: Use train_test_split to split the X & Y data into test and training data
- Step 5: Create a logistic object then create a GridSearchCV object
- Step 6: Calculate the accuracy on the test data using score method
- Step 7: Create a support vector machine object then create a GridSearchCV
- Step 8: Calculate the accuracy on the test data using the method score
- Step 9: Create a decision tree classifier object then create a GridSearchCV object
- Step 10: Calculate the accuracy of tree- cv on the test data using the method score
- Step 11: Create a K nearest neighbors object then GridSearchCV
- Step 12: Calculate the accuracy of Knn-cv & find the method performs best

Machine Learning Prediction

Results

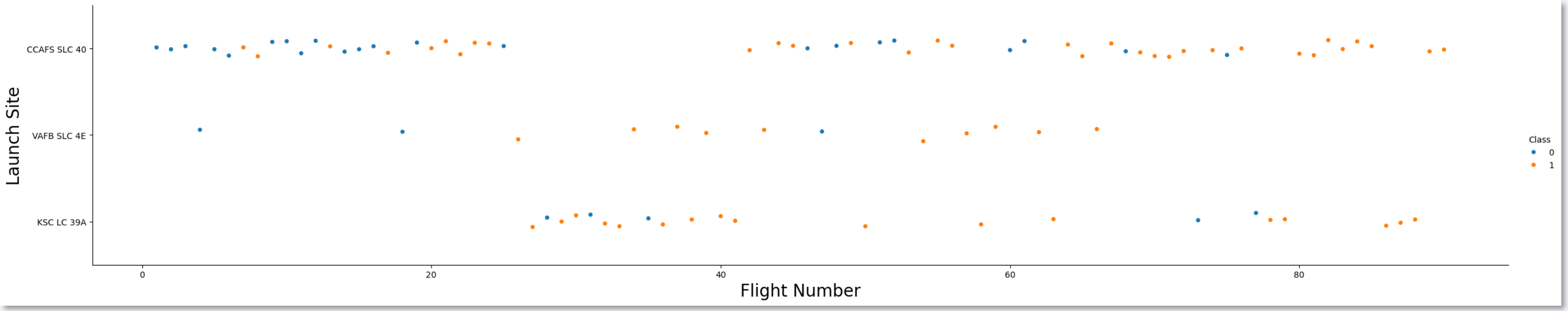
- Space X dataset contains 4 different launch sites;
- The first launches were done to Space X itself and NASA;
- The average payload of F9 v1.1 booster is 2,534 kg;
- The first successful landing outcome happened in 2015 fiver year after the first launch;
- Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average;
- The majority of mission outcomes were successful;
- Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015;
- The number of landing outcomes became as better as years passed;
- All the launch sites were located near coastlines;
- For predictive analysis, the decision tree algorithm showed the highest performance.

Section 2

Insights drawn from EDA

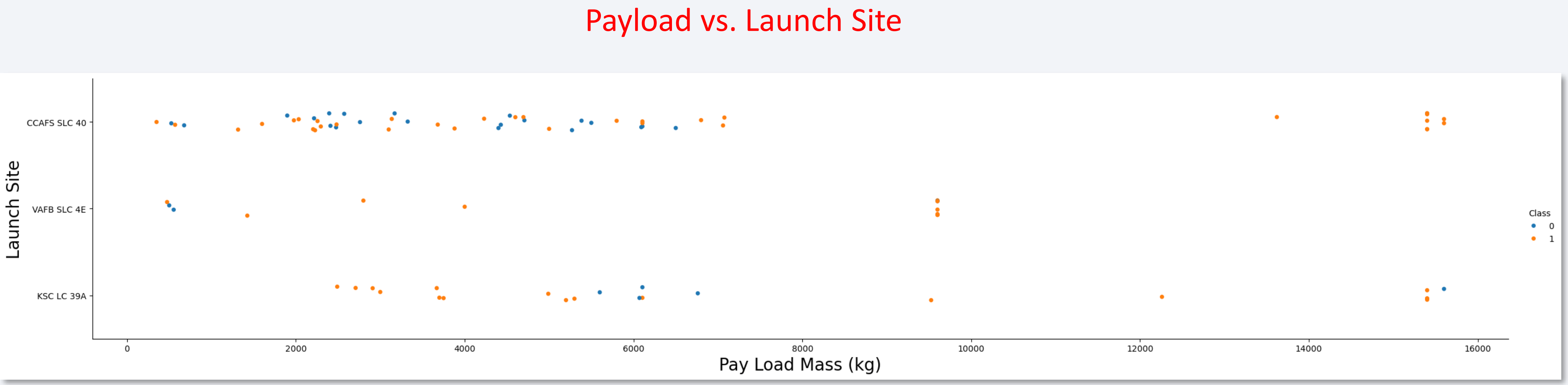
Flight Number vs. Launch Site

Flight Number vs. Launch Site



- The majority of the flights were launched from the CCAFS SLC 40 sites.
- The VAFB SLC 4E and KSC LC39A sites have higher success rates than other sites.
- Newer flights have higher success rates than older flights.

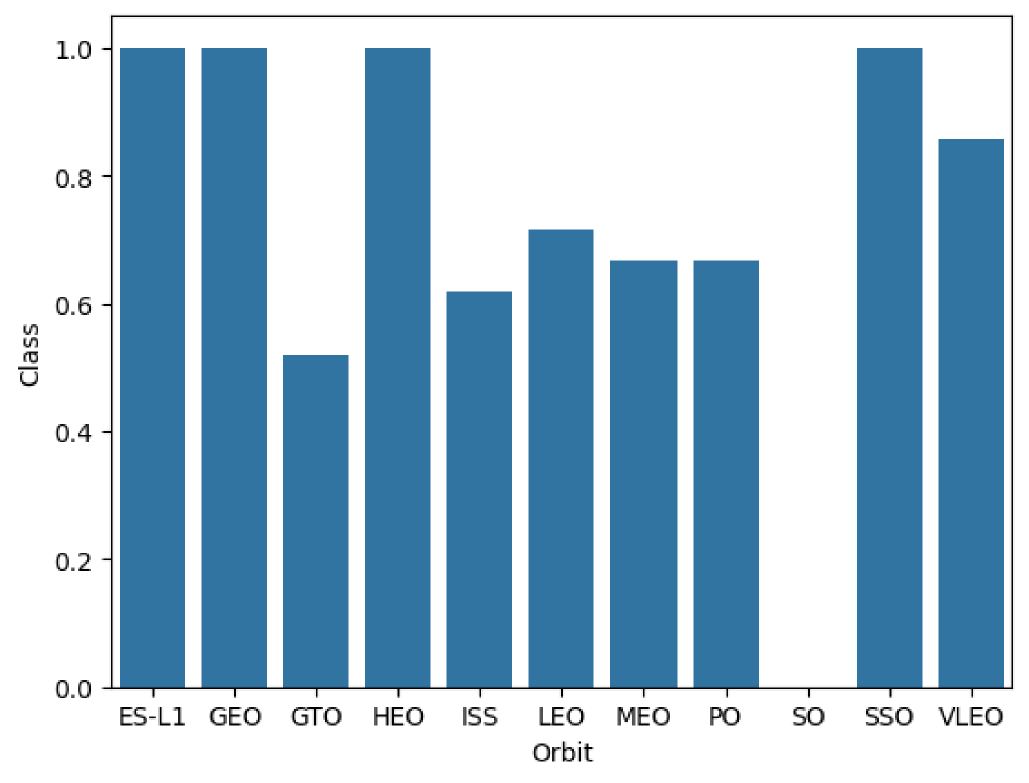
Payload vs. Launch Site



- The majority of the flights with payload mass above 7000 Kg were successful.
- KSC LC 39A success rate for payload mass under 5500 kg is 100%.
- For all launch sites the success rate is proportional to the payload mass.

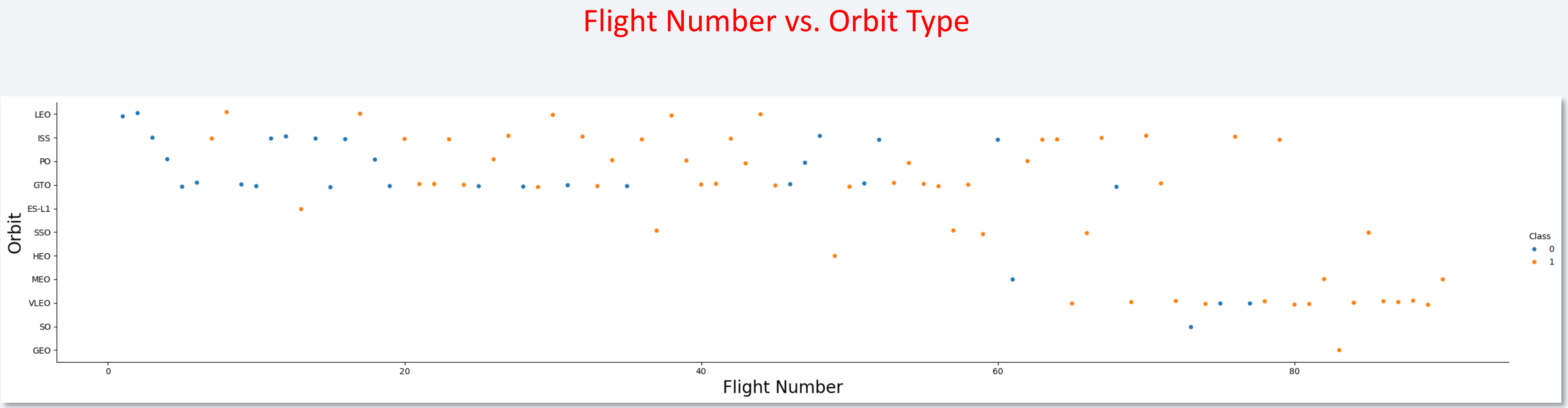
Success Rate vs. Orbit Type

Success Rate vs. Orbit Type



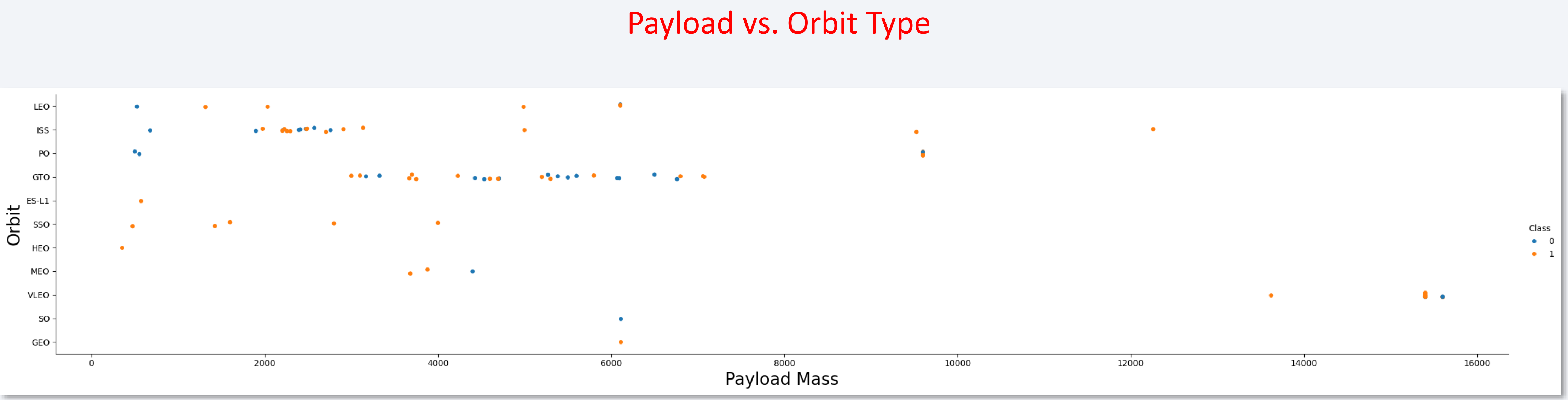
- The SO orbit has 0% success rate.
- The ELS-1, GEO, HEO and SSO orbits have 100% success rate.
- Orbits GTO, ISS, LEO, MEO and PO success rate is higher than 50% and less than 75%.

Flight Number vs. Orbit Type



- The majority of the flights were launches to the ISS and GTO orbits.
- The data suggests that there is no relationship between the flight number and the orbit type.

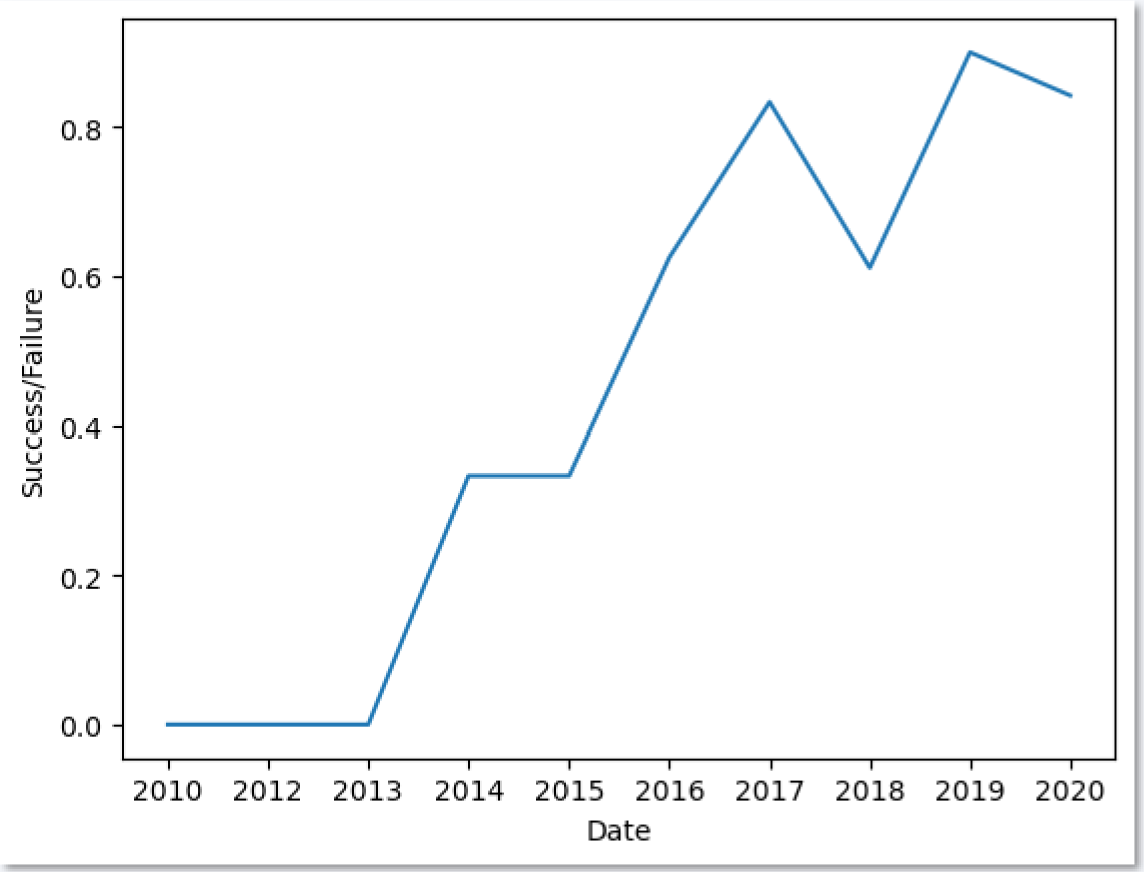
Payload vs. Orbit Type



- Payload masses above 10000Kg were placed in PO, ISS and LEO orbits.
- Payload masses above 4000Kg and less than 8000Kg were placed in the GTO orbit.

Launch Success Yearly Trend

Launch Success Yearly Trend



- The launches success rate increased steadily since 2013.
- The increase in the success rate between 2013 and 2017 was remarkable.
- During 2018 there was a drop in the launches success rate.

All Launch Site Names

SQL query

```
%sql select distinct Launch_Site from SPACEXTABLE
```

Launch sites

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

Launch Site Names Begin with 'CCA'

Launch Site Names Begin with 'CCA'

```
%sql select * from SPACEXTABLE 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

Payload Mass

```
%sql select sum(PAYLOAD_MASS__KG_) as Tot_Payload_Mass from SPACEXTABLE where Customer LIKE "%NASA (CRS)%"
```

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

Tot_Payload_Mass
48213

The total payload mass for customer NASA (CRC) est de 48213Kg

Average Payload Mass by F9 v1.1

Average payload mass

```
%sql select avg(PAYLOAD_MASS__KG_) as AVg_Payload_Mass from SPACEXTABLE where Booster_version like "F9 v1.1%"
```

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

AVg_Payload_Mass
2534.6666666666665

The average payload mass for Booster version F9 v1.1 is 2534 Kg

First Successful Ground Landing Date

First successful Date

```
%sql select min(Date) as First_Successful_Landing_Date from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)'
```

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

First_Successful_Landing_Date
2015-12-22

The first time there was a successful ground landing was in December 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

Below the query and results

```
%sql select Booster_Version from SPACEXTABLE where Landing_Outcome = 'Success (drone ship)' and 4000 <PAYLOAD_MASS__KG_<6000
 * sqlite:///my_data1.db
Done.
```

Booster_Version
F9 FT B1021.1
F9 FT B1022
F9 FT B1023.1
F9 FT B1026
F9 FT B1029.1
F9 FT B1021.2
F9 FT B1029.2

The list in the picture above shows the booster versions that landed on Drone ship with payload between 4000 and 6000

Total Number of Successful and Failure Mission Outcomes

The query used :

```
%sql select Mission_Outcome, count(*) as Total from SPACEXTABLE where Mission_Outcome like "%Success%" or Mission_Outcome like "%Failure%" group by Mi
```

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

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

We can see that most of the outcomes were successful

Boosters Carried Maximum Payload

The query used :

```
%sql select Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE )
```

```
 * 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

The above picture shows the boosters with maximum payload

2015 Launch Records

The query used :

```
%sql select substr(Date, 6,2) as Month, Landing_Outcome, Booster_Version, Launch_Site from SPACEXTABLE where Landing_Outcome = 'Failure (drone ship)'
```

```
 * 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

The results list the failed landing outcomes in drone ship, their booster versions, and launch site names in year 2015

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

```
%sql select Landing_Outcome, count(Landing_Outcome) as count from SPACEXTABLE where Date between '2010-06-04' and '2017-03-20' group by Landing_Outcom
```

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

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

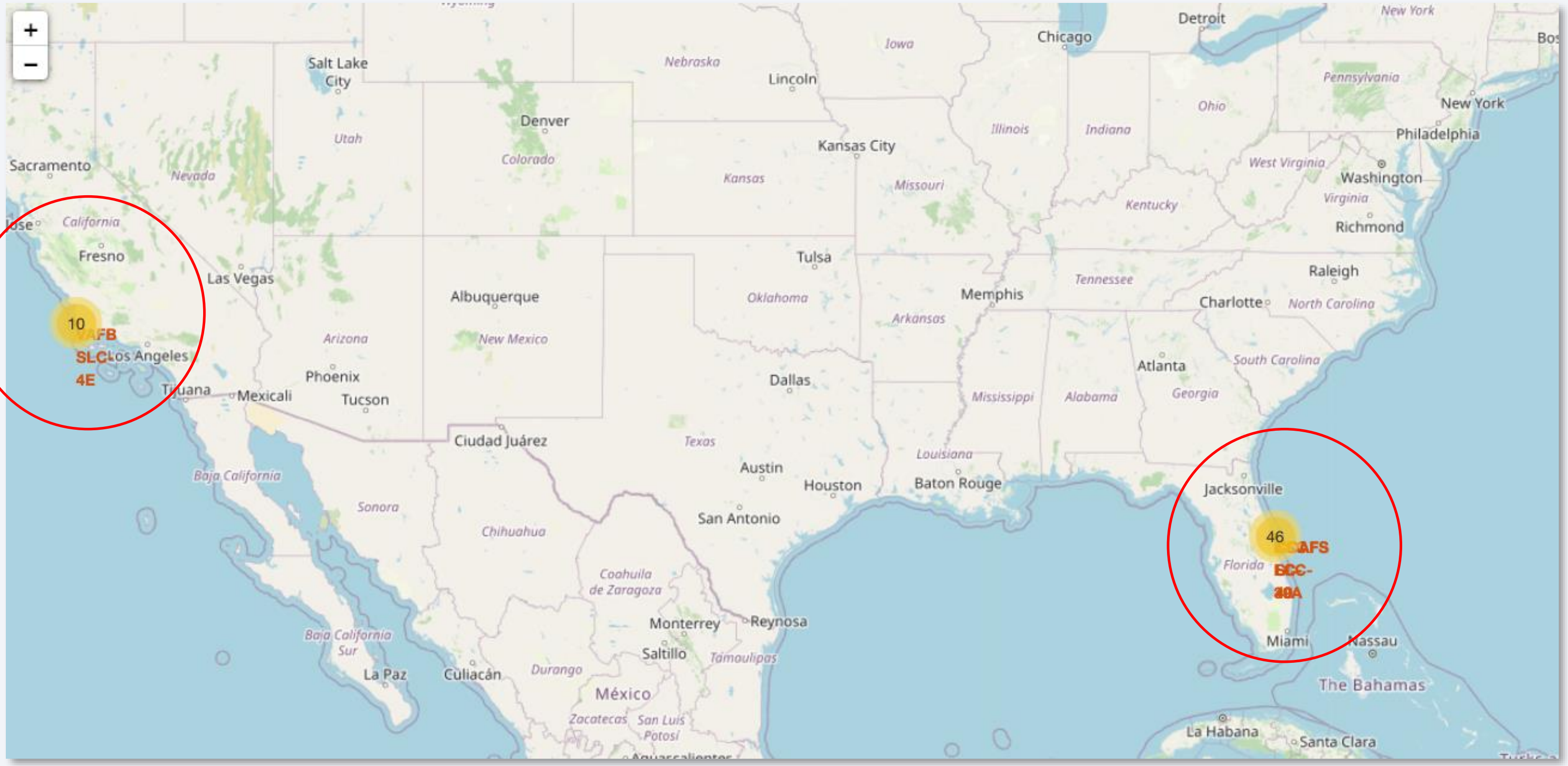
The results rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad) between the date 2010-06-04 and 2017-03-20, in descending order

Section 3

Launch Sites Proximities Analysis

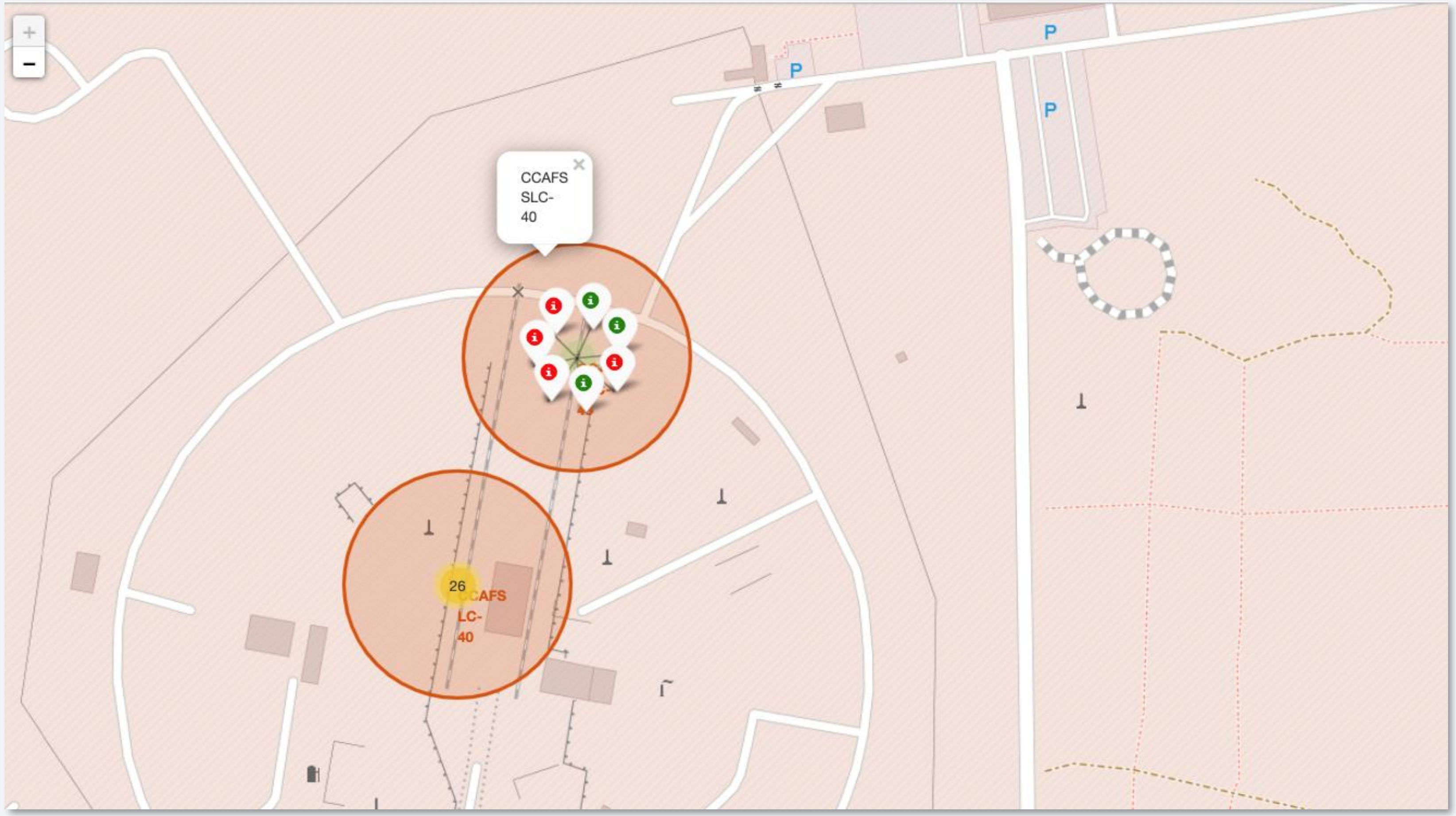
USA Launch Sites in California and Florida

Map generated with Folium



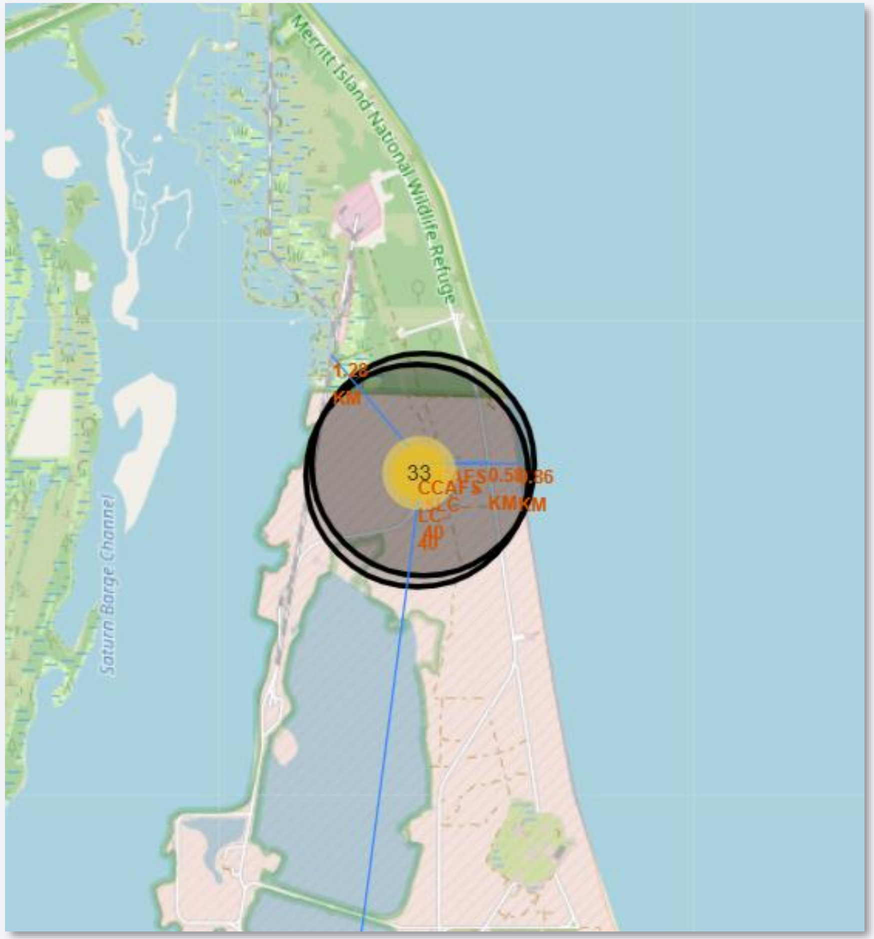
Launch sites are indicated in yellow circles on the Map

Color Labels Showing the Launch Sites on a Map



Launch sites are circles in red and in each site there are red markers for failed outcomes and green markers for successful outcomes

Safe Distance to Launch Site

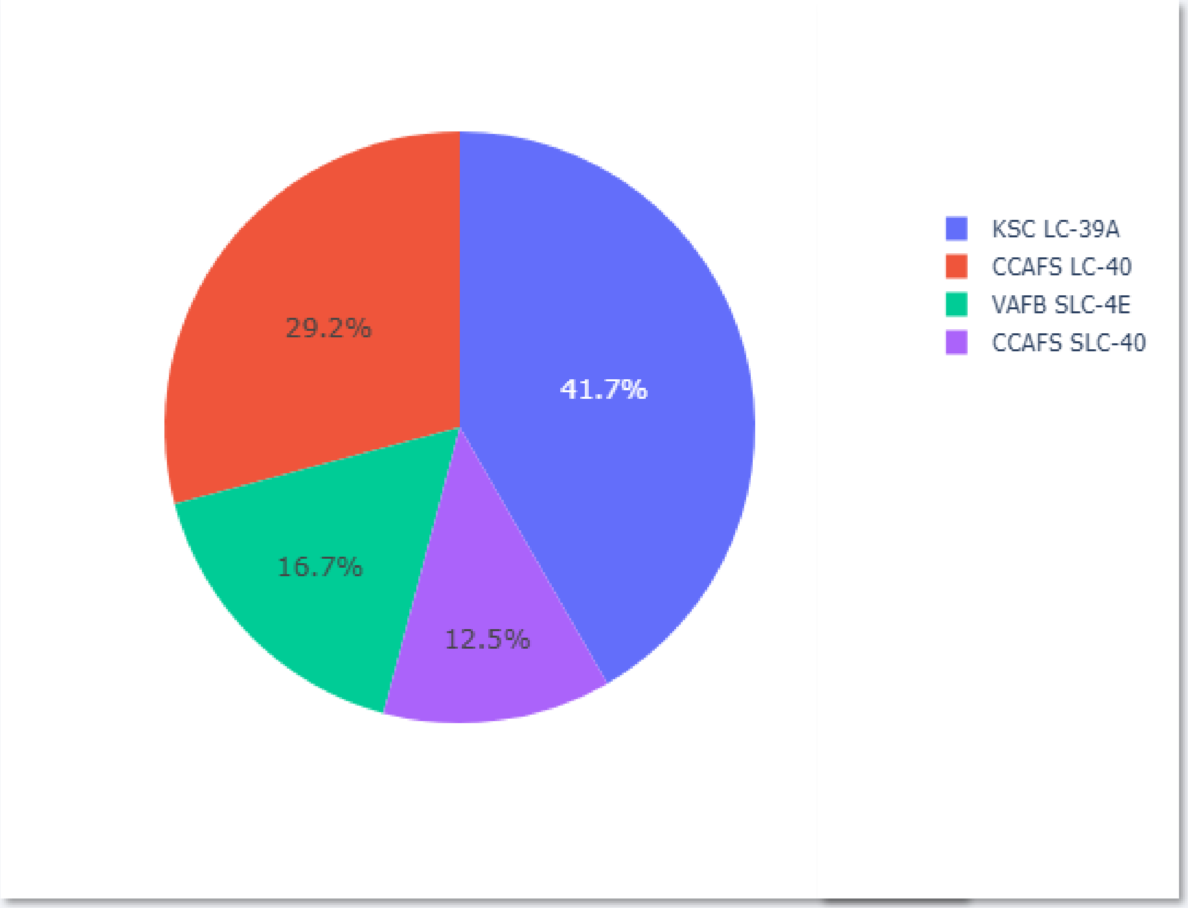


The obtained results indicate that all launch sites are at safe distance from railway lines, airports and cities.

Section 4

Build a Dashboard with Plotly Dash

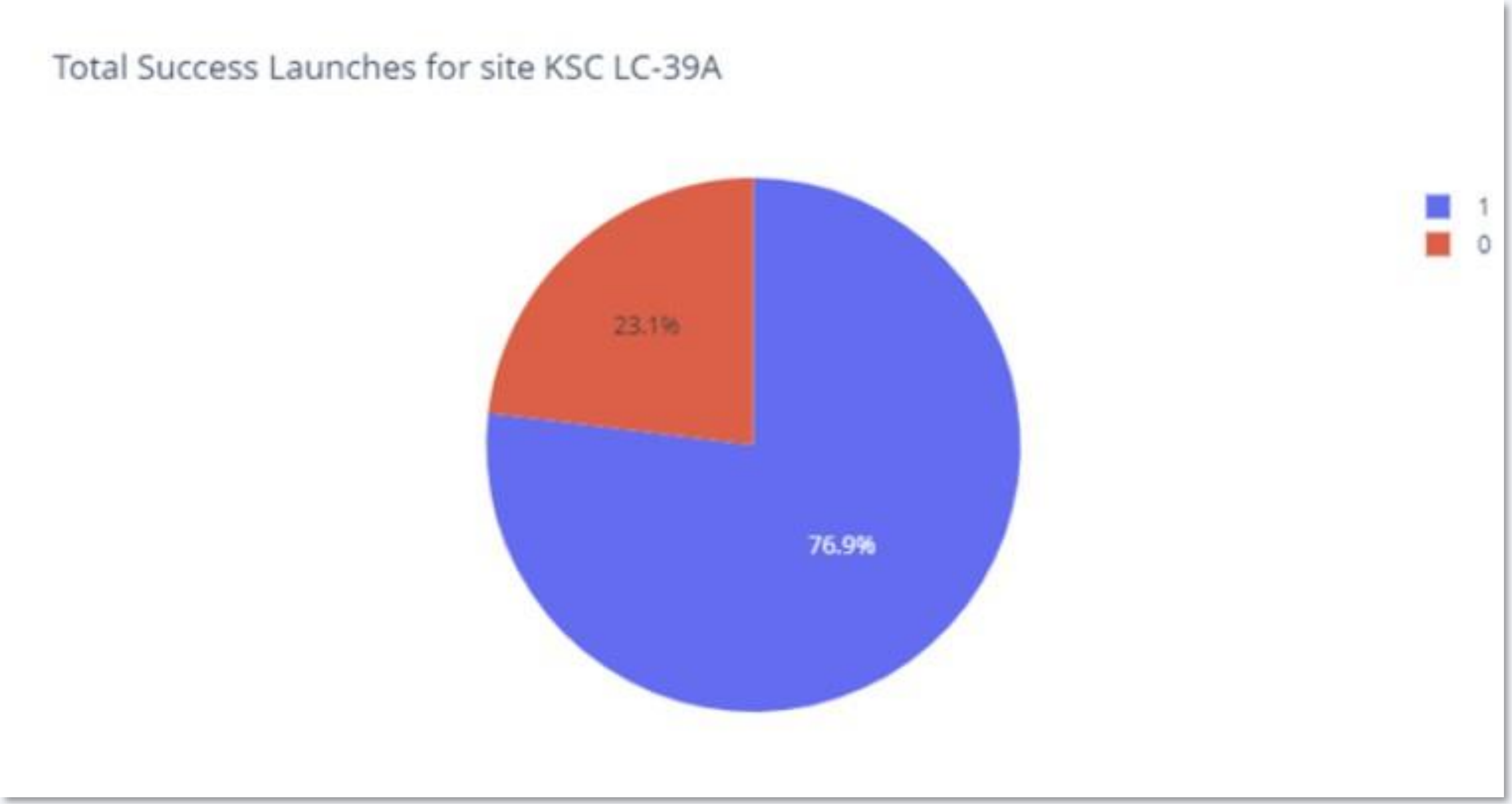
Total Launch Success for All Sites



The highest success launch rates were recorded at these sites :

1. K S C LC-39A (41.7%)
2. C CAFS LC-40 (29.2%)

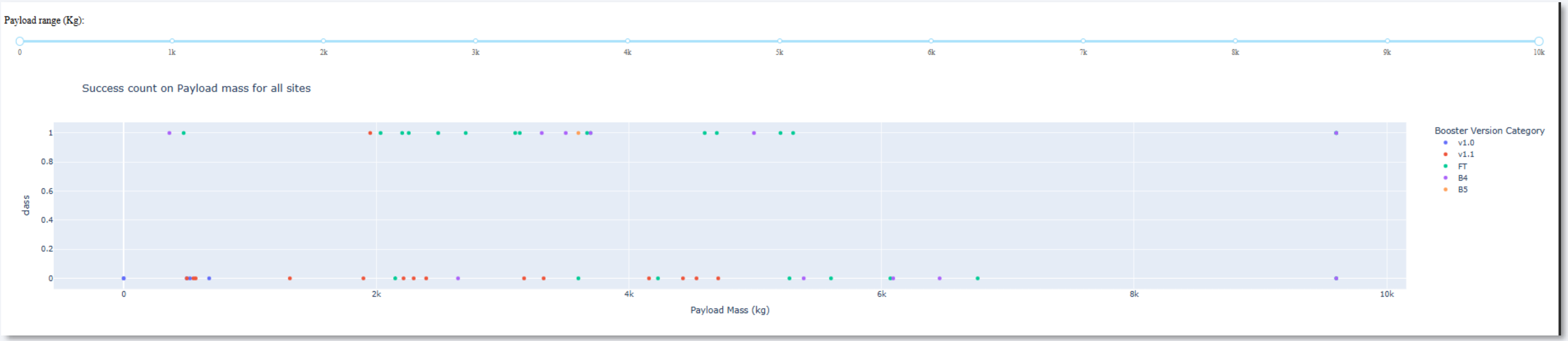
KSC LC-39 Launch Site Success Rate



Site KSC LC-39 success rate is 76.9%

Payload vs. Launch Outcome for All Sites

Payload vs. Launch



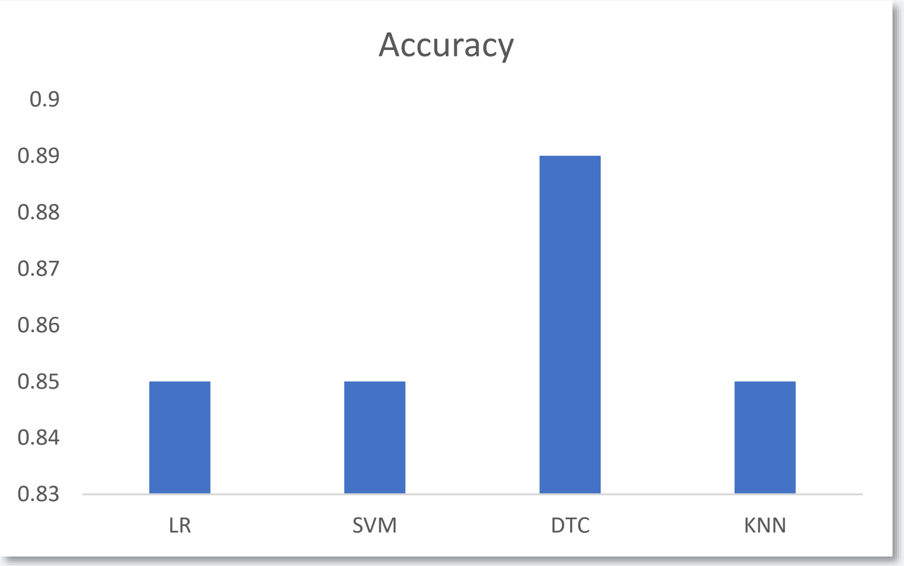
Highest success rate for payloads is between 2000 and 5500 Kgs

Section 5

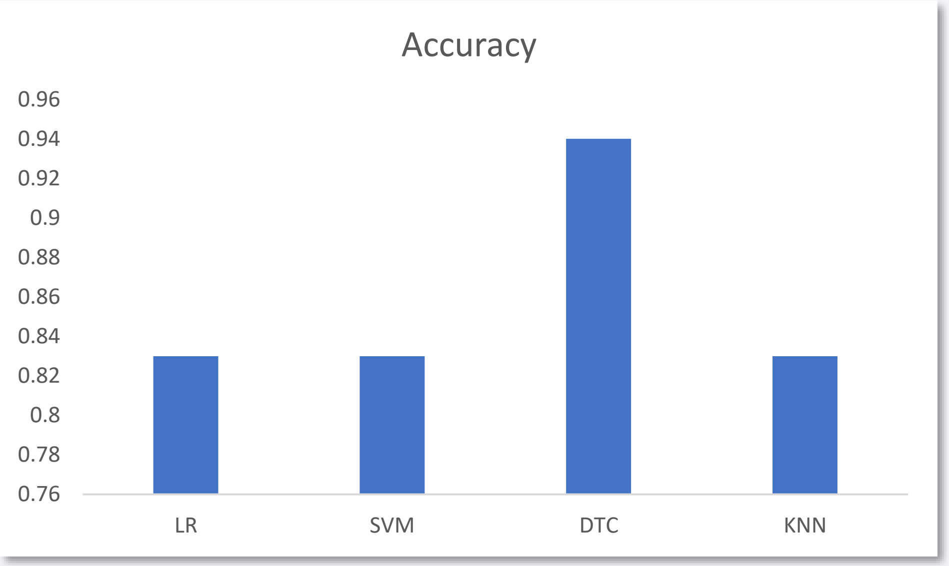
Predictive Analysis (Classification)

Classification Accuracy

Cross validation



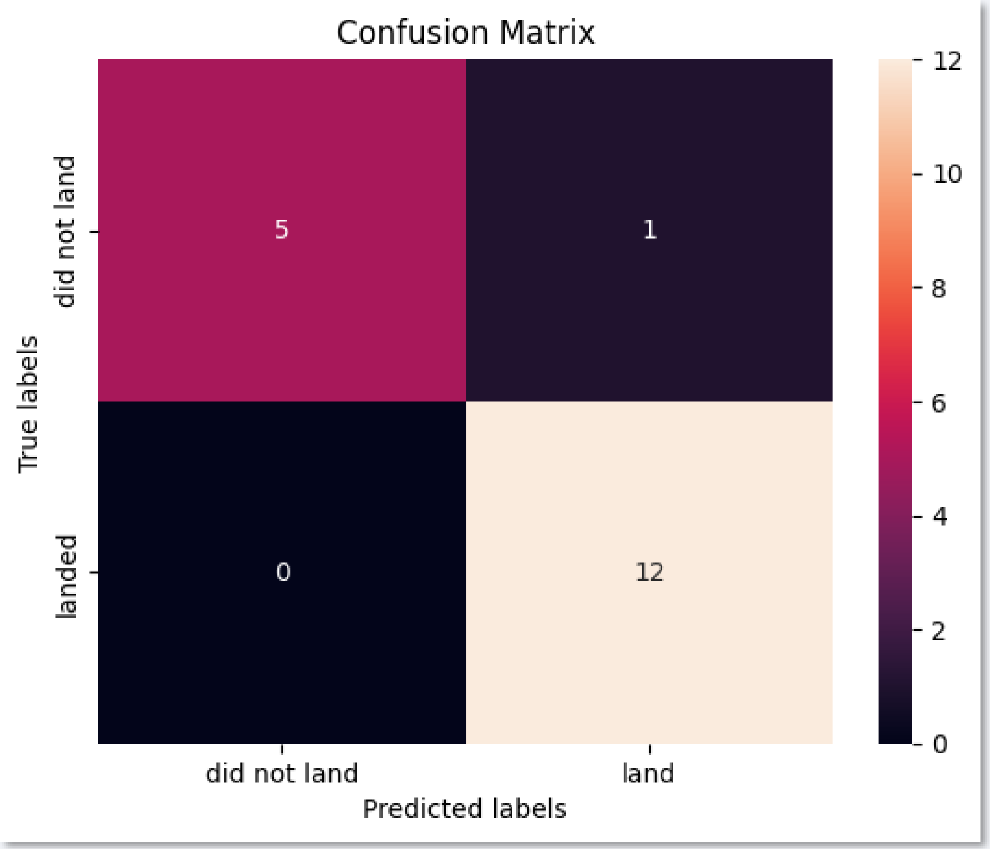
Test data



The Tree Model provided the best accuracy results for the cross-validation stage and the test set

Confusion Matrix

Decision Tree



- The confusion matrix analysis suggests that the best performing model is the decision tree classifier.
- The confusion matrix predicts 12 true positives, 1 false positive, 5 true negatives, and 0 false negative.

Conclusions

- SpaceX data was collected using SpaceX API and web Scraping
- Date wrangling and EDA were performed
- The success rate for the rocket launches increased after 2013.
- Orbits GEO, HEO, ES-L1 and SSO have 100% launch success rate.
- Launch site KSC LC-39A has the highest success rate.
- Predictive analysis was performed using Logistic Regression, Support Vector Machine, Decision Tree and K-Nearest Neighbors.
- The Decision Tree model is the best ML algorithm for analyzing the SpaceX data set and provided the best accuracy results.

Appendix

- Applied Data Science Capstone Project – Coursera Link
- DS-Professional-Certificate-Project - Git Hub Link

Thank you!