IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Joaquin Koifman

23/03/2023

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

Summary of Methodologies

The goal of this investigation is to identify the factors for the successful landing of the Falcon 9, a Space X rocket. To carry out this investigation, different methodologies were used, such as:

Collect data using several techniques.

Wrangle data to create success/fail outcome variable.

Explore data with data visualization techniques.

Analyze the data with SQL.

Explore launch site success rates and proximity to geographical markers.

Visualize the launch sites with the most success and successful payload ranges.

Build Models to predict landing outcomes.

Search

How payload mass, launch site, number of flights, and orbits affect first-stage landing success

Rate of successful landings over time

Best predictive model for successful landing (binary classification)

SPACEX

Introduction

Context

The aim of this capstone if to predict 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.

Therefore, what we're trying to determine is, if the first stage of Falcon 9 will land successfully or not.

Section 1

Methodology

Methodology

Step by step:

Collecting data was performed by using SpaceX REST API and web scraping techniques

Wrangling data by filtering the data and handling missing values, in order to prepare the data for analysis and modeling

Exploring data analysis using visualization and SQL

Visualizing analytics using Folium and Plotly Dash

Predicting and analyzing data by using classification models as to predict landing outcomes using classification models. Testing each one of them to find the best one.

Data Collection (API)

Steps:

1. Request the rocket launch data from SpaceX API
2. Normalize the response by using JSON to convert it into a dataframe
3. Get information about the launches using the API
4. Create a new dataframe
5. Filter the dataframe to only include Falcon 9 launches
6. Dealing with Missing Values
7. Export data to csv file

Data Collection Web Scraping

Steps:

1. Request the Falcon 9 launch data from Wikipedia
2. Extract column/variable names from HTML table header
3. Create a dataframe by parsing the launch HTML tables
4. Export data to csv file

Data Wrangling

Steps:

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
5. Export to csv

EDA with Data Visualization

Charts:

Scatter plots: are used to represent the relationship between two variables. Some of the variables that were analyzed were:

- Flight Number vs. Launch Site
- Payload vs. Launch Site
- Flight Number vs. Orbit Type
- Payload vs. Orbit Type.

Bar charts: are used to show comparisons among discrete categories. They show the relationships among the categories and a measured value. In this case bar charts were used to analyze the plotted bar chart try to find which orbits have high success rate

Line chart: are useful for showing data trends over time. For example, here it was used to show success rate over a certain number of years.

EDA with SQL

List of the SQL queries and lists that were used on the dataset:

- Names of unique launch sites
- 5 records where launch site begins with ‘CCA’
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1.
- Date of first successful landing on ground pad
- Names of boosters which had success landing on drone ship and have payload mass greater than 4,000 but less than 6,000
- Total number of successful and failed missions
- Names of booster versions which have carried the max payload
- Failed landing outcomes on drone ship, their booster version and launch site for the months in the year 2015
- Count of landing outcomes between 2010-06-04 and 2017-03-20 (desc)

Interactive Map with Folium

Marking launch sites

by using location on the map using site's latitude and longitude coordinates, a highlighted yellow circle area with a text label on a specific coordinate of the launch was added.

Succeeded and failed launches

by adding color-labeled markers in marker clusters for each site, it is easily to identify successful (green) and unsuccessful (red) launches at each launch site to show which launch sites have high success rates

Calculate the distances between a launch site to its proximities

by adding colored lines to show distance between launch sites and its proximity to the nearest coastline, railway, highway, and city using latitudes an longitudes to locate them

Dashboard with Plotly D

The dashboard contains:

Dropdown List with Launch Sites: allows the user to select all launch sites or a certain launch site.

Pie Chart Showing Successful Launches: allows the user to see all successful and unsuccessful launches as a percent of the success rate of each site.

Slider of Payload Mass Range: allows the user to select payload mass range

Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version: allows user to visualize how different variables affect the landing outcomes.

Predictive Analysis (Classification)

- Create a NumPy array from the column Class in data
- Standardize the data. Fit and transform the data.
- Split the data using train_test_split
- Create a logistic regression object then create a GridSearchCV object. Fit the object to find the best parameters.
- Apply GridSearchCV on different algorithms: logistic regression, support vector machine (SVC), decision tree and K-Nearest Neighbor.
- Calculate the accuracy on the test data for each model
- Identify the best model using the test data to evaluate models based on their accuracy scores and confusion matrix

Results

Results

Exploratory data analysis results:

Launch success has improved over time.

KSC LC-39A has the highest success rate among landing sites

Orbits ES-L1, GEO, HEO and SSO have a 100% success rate

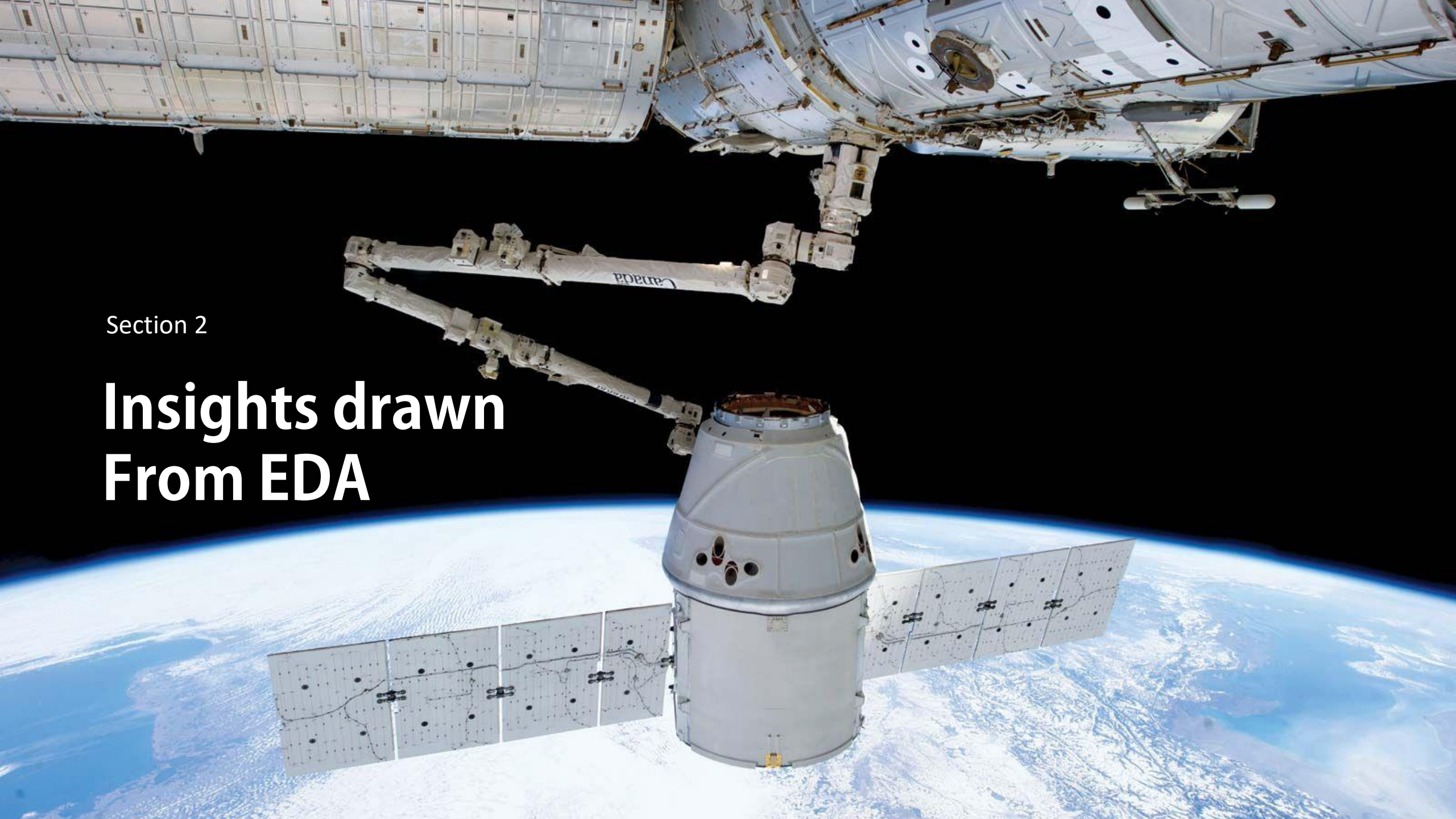
Interactive analytics demo in screenshots:

Most launch sites are between the parallel 35N and 28N (or near the equator), and all of them are close to the coast

Every launch Facility is far enough away from anything a failed launch can damage (city, highway, railway).

Predictive analysis results

The Decision Tree model is the best predictive model for the dataset, with an accuracy of the 83.33% on the data test.



Section 2

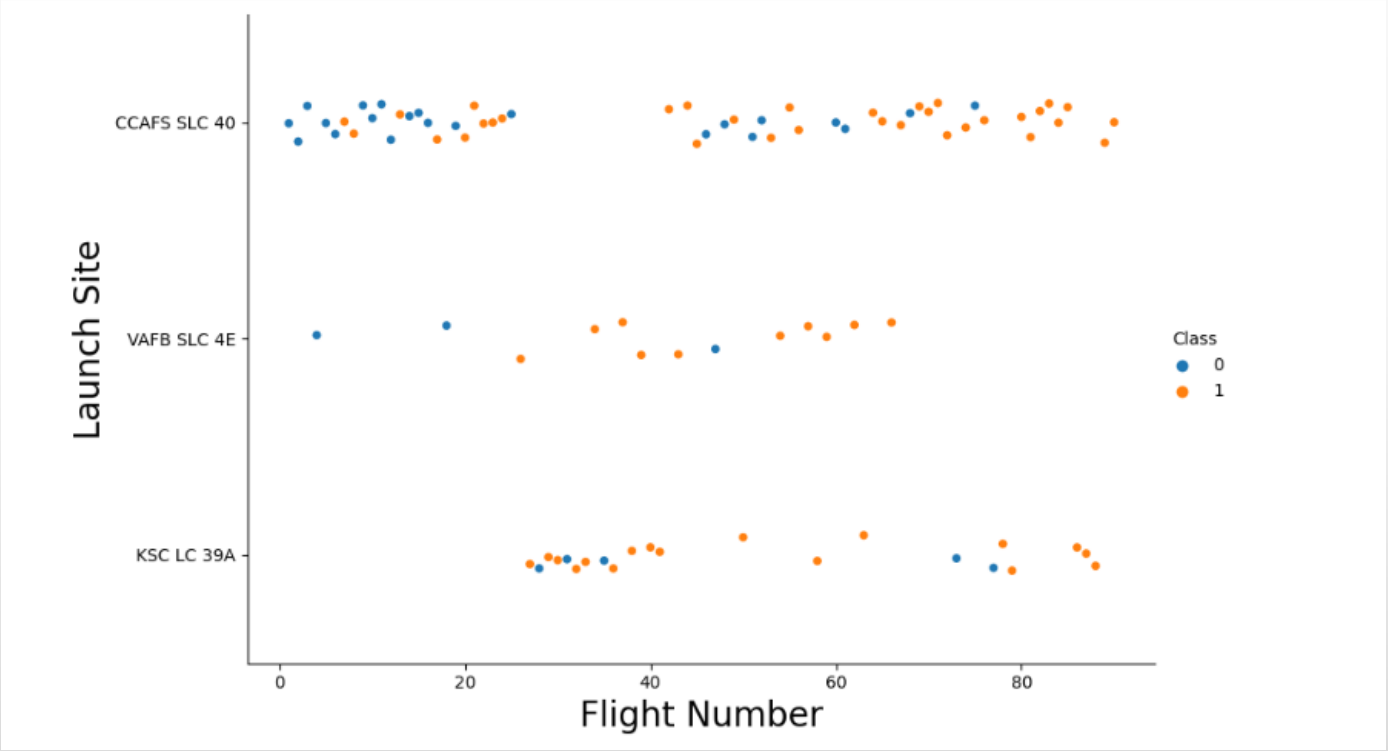
Insights drawn From EDA

Flight Number vs. Launch Site

Exploratory Data Analysis:

Most of the firsts flights tend to had an unsuccessful ending witch are represented by blue dots. After a while, flights were succeeding their landings, these are represented with orange dots

Around half of launches were from CCAFS SLC 40, and VAFB SLC 4E and KSC LC 39A have higher success rates



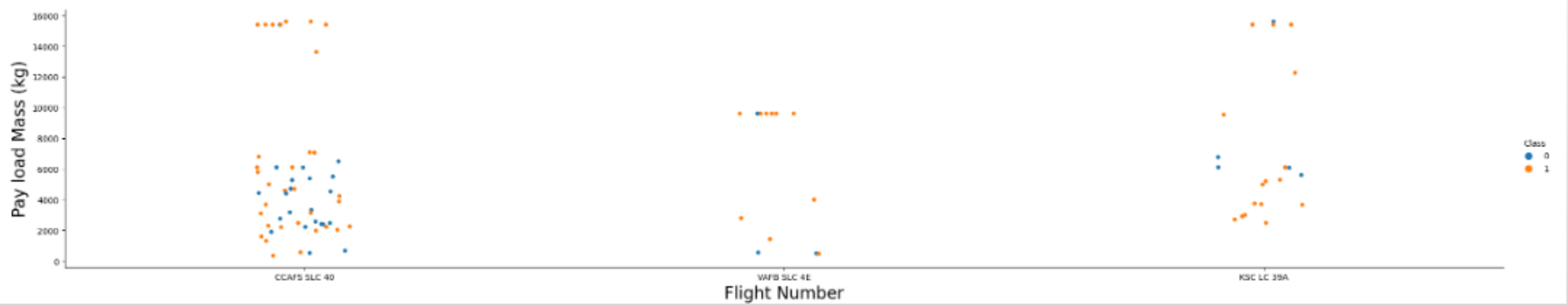
Payload vs. Launch Site

Exploratory Data Analysis:

In general, the success rate is higher, if the payload mass is higher too.

VAFB SKC 4E has only 2 unsuccessful launches and also it has not launches anything greater than 10000 kg

KSC LC 39A has a 100% success rate for launches less than 5,500 kg



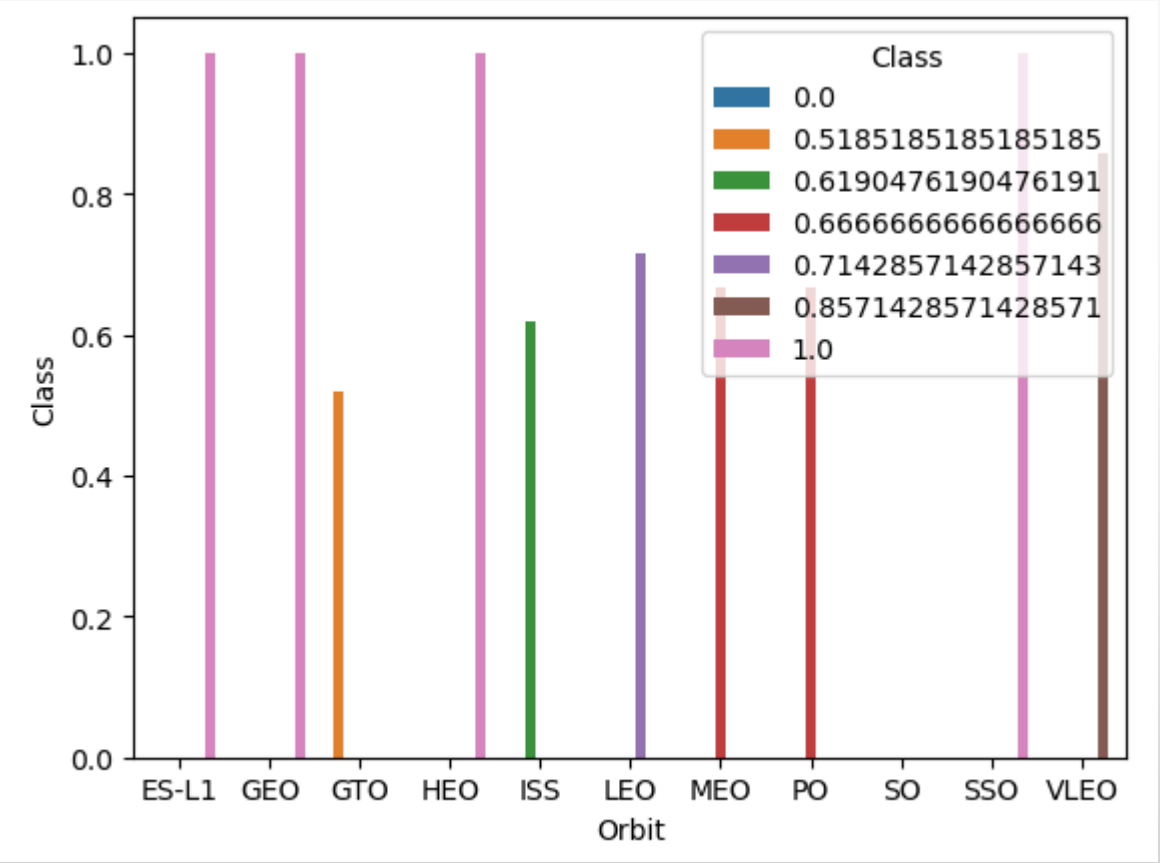
Success Rate vs. Orbit Type

Exploratory Data Analysis:

100% Success Rate:
ES-L1, GEO, HEO and SSO

50%-80% Success Rate:
GTO, ISS, LEO, MEO, PO

0% Success Rate:
SO



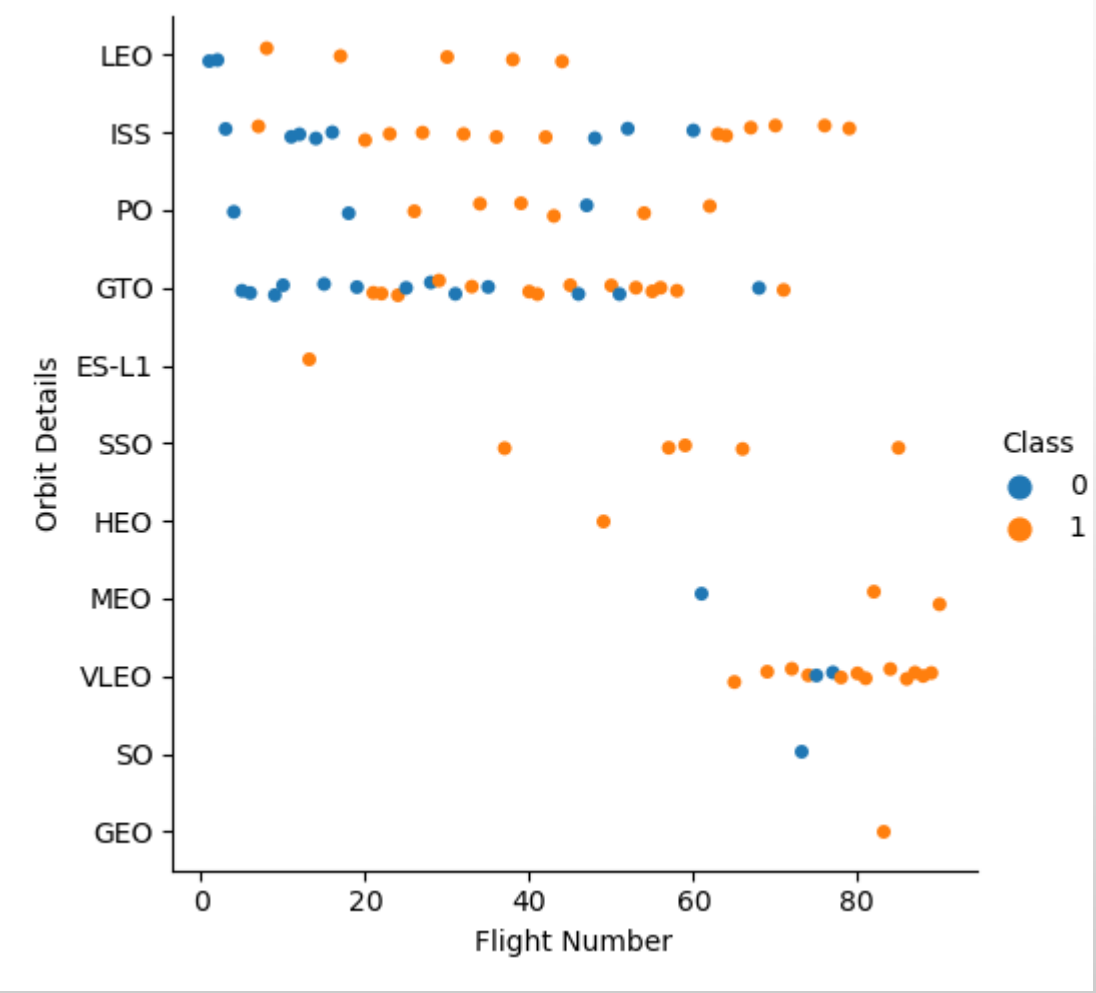
Flight Number vs. Orbit Type

Exploratory Data Analysis:

The success rate typically increases with the number of flights for each orbit (LEO)

ISS and GTO share similarities when it comes to their success rates

The SSO and GEO orbits have a 100% success rate



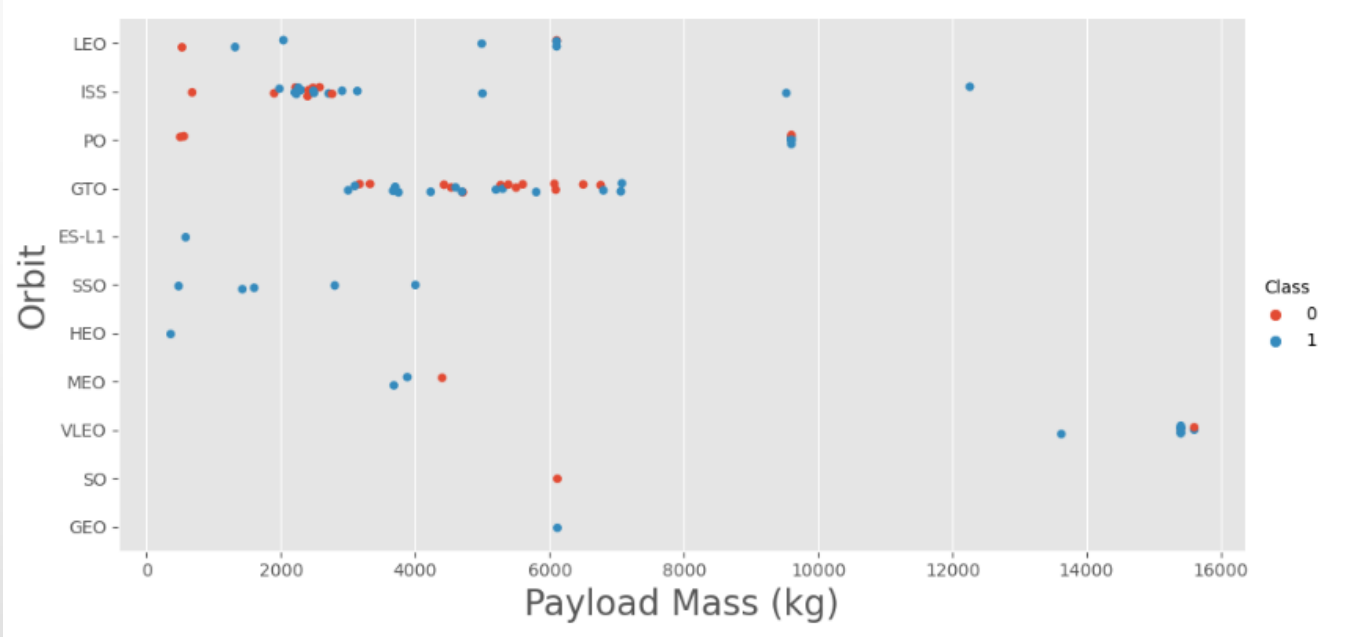
SPACEX

Payload vs. Orbit Type

Exploratory Data Analysis:

As the payloads get heavier, the success rate increases.

GTO orbit, alternates it success as the mass increases

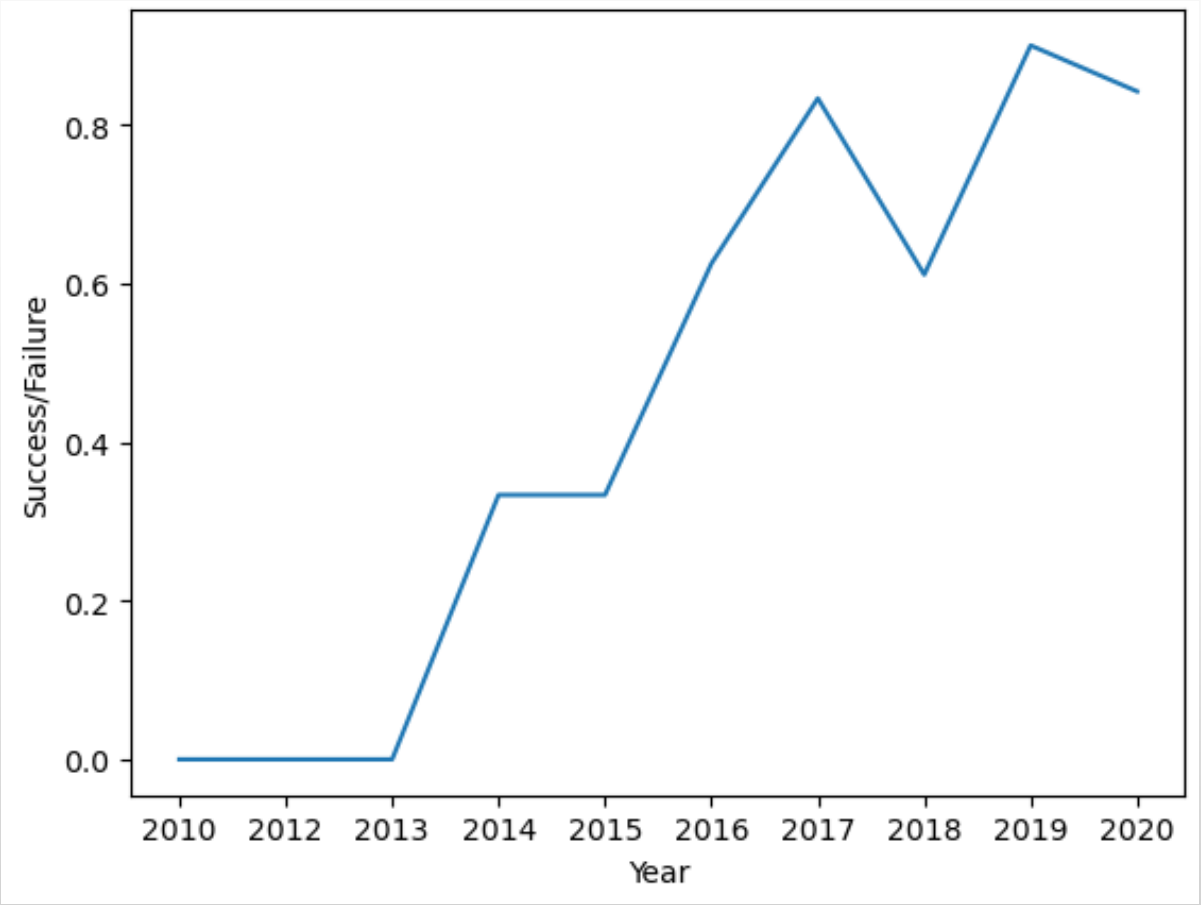


Launch Success Yearly Trend

Exploratory Data Analysis:

In general, the success rate has improved since 2013

However, the success decreased in 2018 and in 2020.



Launch Site Names

Exploratory Data Analysis:

Launch Site Names:

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

```
%sql SELECT DISTINCT LAUNCH_SITE as "Launch_Sites" FROM SPACEXTBL;
```

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

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

'CCA' Launch Site Names

Exploratory Data Analysis:

Records with Launch Site Starting with 'CCA':

Displaying 5 records below

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Payload Mass

Exploratory Data Analysis:

Total Payload Mass: 45,596 kg (total) carried by boosters launched by NASA (CRS)

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) as PM_KG_TOTAL, Customer FROM 'SPACEXTBL' WHERE Customer = 'NASA (CRS)'
```

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

PM_KG_TOTAL	Customer
45596	NASA (CRS)

Average Payload Mass: 2,928 kg (average) carried by booster version F9 v1.1

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

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

AVG(PAYLOAD_MASS__KG_)
2928.4

SPACEX

Landings and Missions

Exploratory Data Analysis:

First successful landing outcome on ground pad

```
%sql SELECT min(DATE) FROM SPACEXTBL WHERE LANDING_OUTCOME ='Success (ground pad)'
```

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

1
2015-12-22

Boosters which have successfully landed on drone ship and had payload mass greater than 4,000 but less than 6,000

```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ between 4000 and 6000 AND LANDING_OUTCOME='Success (drone ship)'
```

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

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

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT Mission_Outcome, COUNT(Mission_Outcome) as Total FROM SPACEXTBL GROUP BY Mission_Outcome;
```

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

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

SPACEX

2015 Failed Launches

Exploratory Data Analysis:

The list displays the month names, failure landings outcomes in drone ship ,booster versions and launch sites for the months in year 2015.

```
%sql SELECT substr(Date,4,2) as month, DATE,BOOSTER_VERSION, LAUNCH_SITE, [Landing _Outcome] \
FROM SPACEXTBL \
where [Landing _Outcome] = 'Failure (drone ship)' and substr(Date,7,4)='2015';
```

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

month	Date	Booster_Version	Launch_Site	Landing _Outcome
01	10-01-2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	14-04-2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Boosters Carried Maximum Payload

Exploratory Data Analysis:

Carrying Max Payload

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

```
%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

SPACEX

Successful Landings

Between 2010-06-04 and 2017-03-20

Exploratory Data Analysis:

The list displays the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

```
%sql SELECT [Landing _Outcome], count(*) as count_outcomes \
FROM SPACEXTBL \
WHERE DATE between '04-06-2010' and '20-03-2017' group by [Landing _Outcome] order by count_outcomes DESC;
```

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

Landing _Outcome	count_outcomes
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

Section 3

Launch Sites Proximities Analysis

Launch Sites

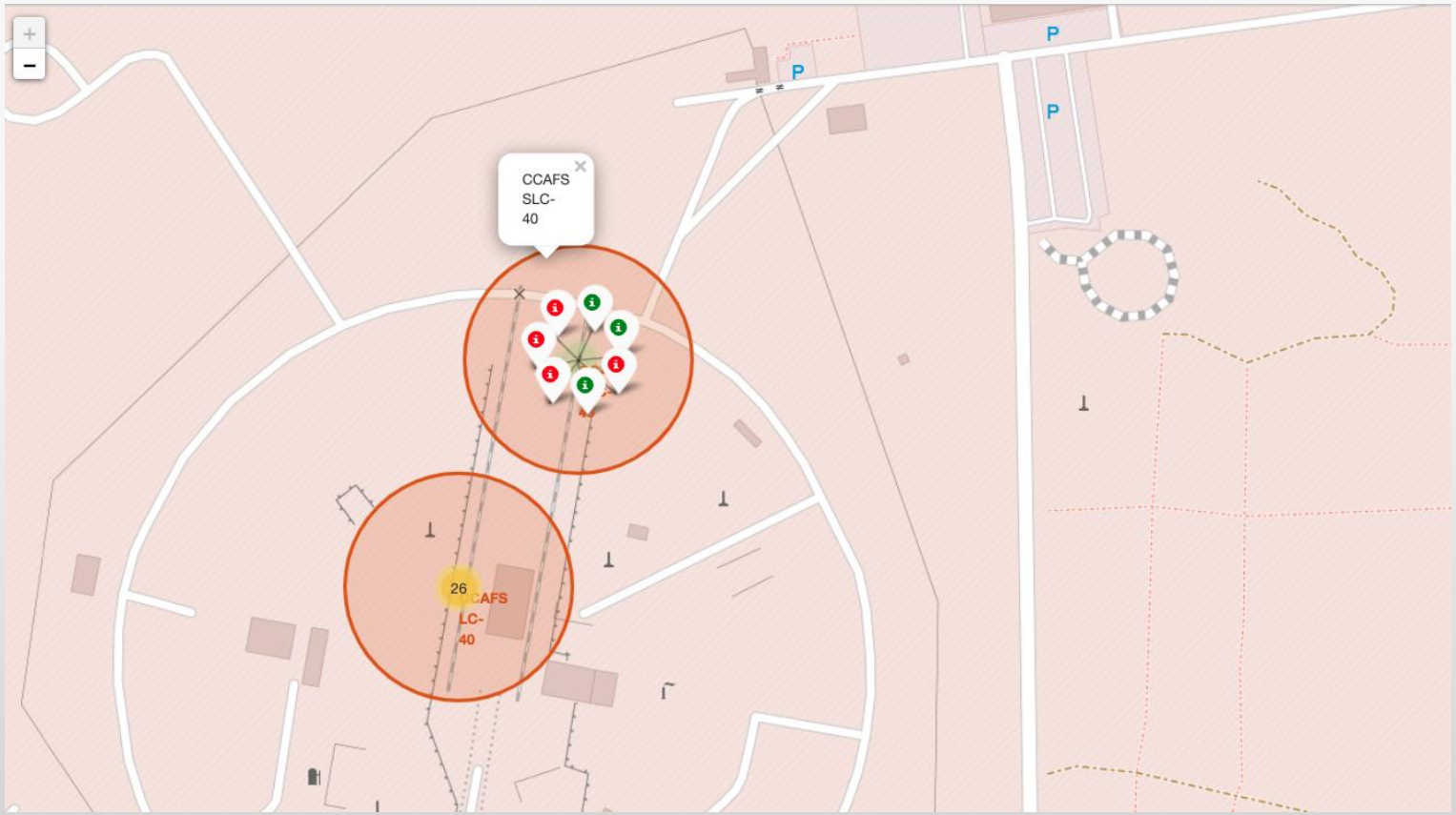
Exploratory Data Analysis:

The black dots represents where the locations of all the Space X launch sites are situated in the US. These locations are strategically located “near” to de Equator because they use the equatorial orbit and the earth’s speed rotation as an additional natural boost to save time and fuel.

Launch Outcomes

Exploratory Data Analysis:

At each launch site there are two possible outcomes: green markers are for successful launches, and red markers are for unsuccessful launches



Launch Site Proximities

Exploratory Data Analysis:

The generated map shows several proximities (railways, highways, closest city and coast line, etc) to the CCAFS SLC-40 launch site with a blue line.

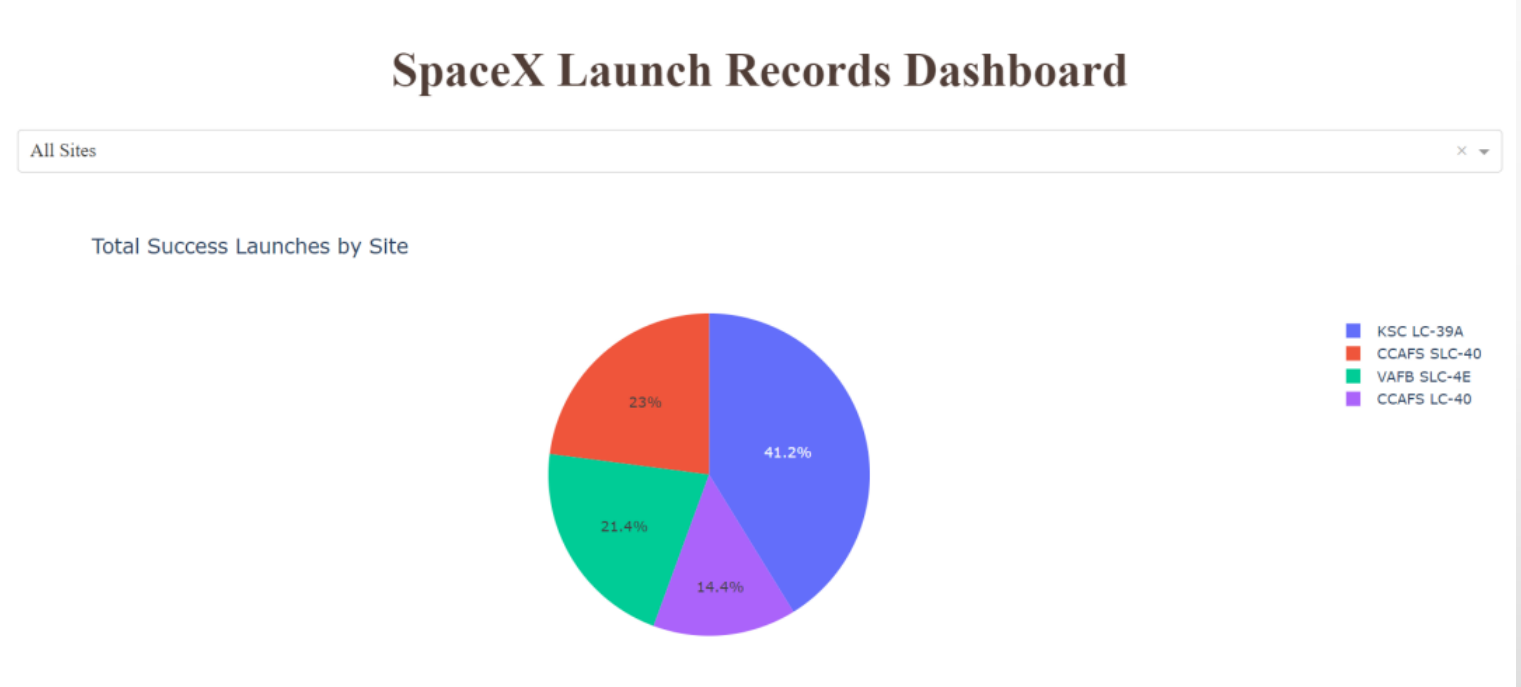
Section 4

Built a Dashboard with Ploty Dash

Total Success Launches by Site

Exploratory Data Analysis:

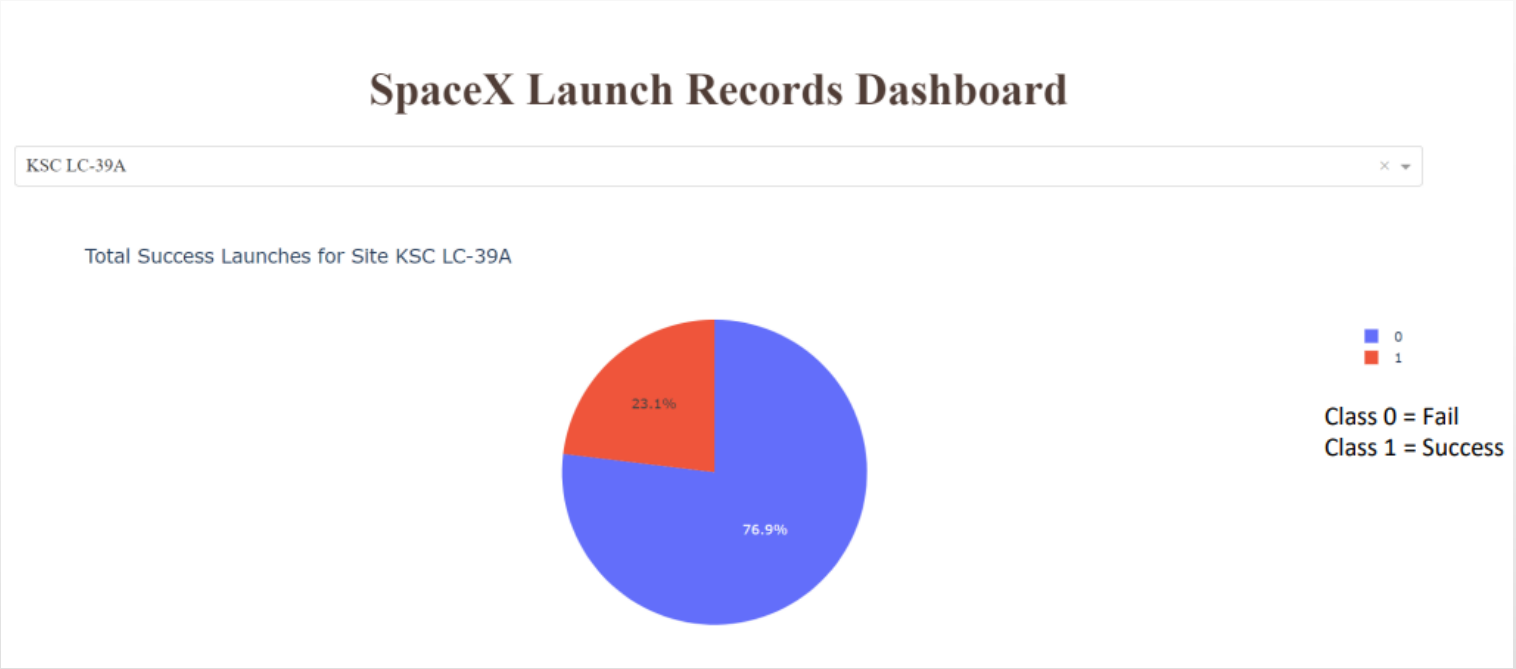
KSC LC-39A has the most successful launches amongst launch sites with a total of 10 (41.2%).



Total Success Launches For Site KSC LC-39A

Exploratory Data Analysis:

KSC LC-39A has the highest success rate amongst launch sites, with 10 successful launches and 3 failed ones (76.9%).



Correlation Between Payload and Success

Exploratory Data Analysis:

Payloads between 2,000 kg and 5,000 kg have the highest success rate (1 indicating successful outcome and 0 indicating an unsuccessful outcome).

Section 5

Predictive analysis (Classification)

Classification Accuracy

Exploratory Data Analysis:

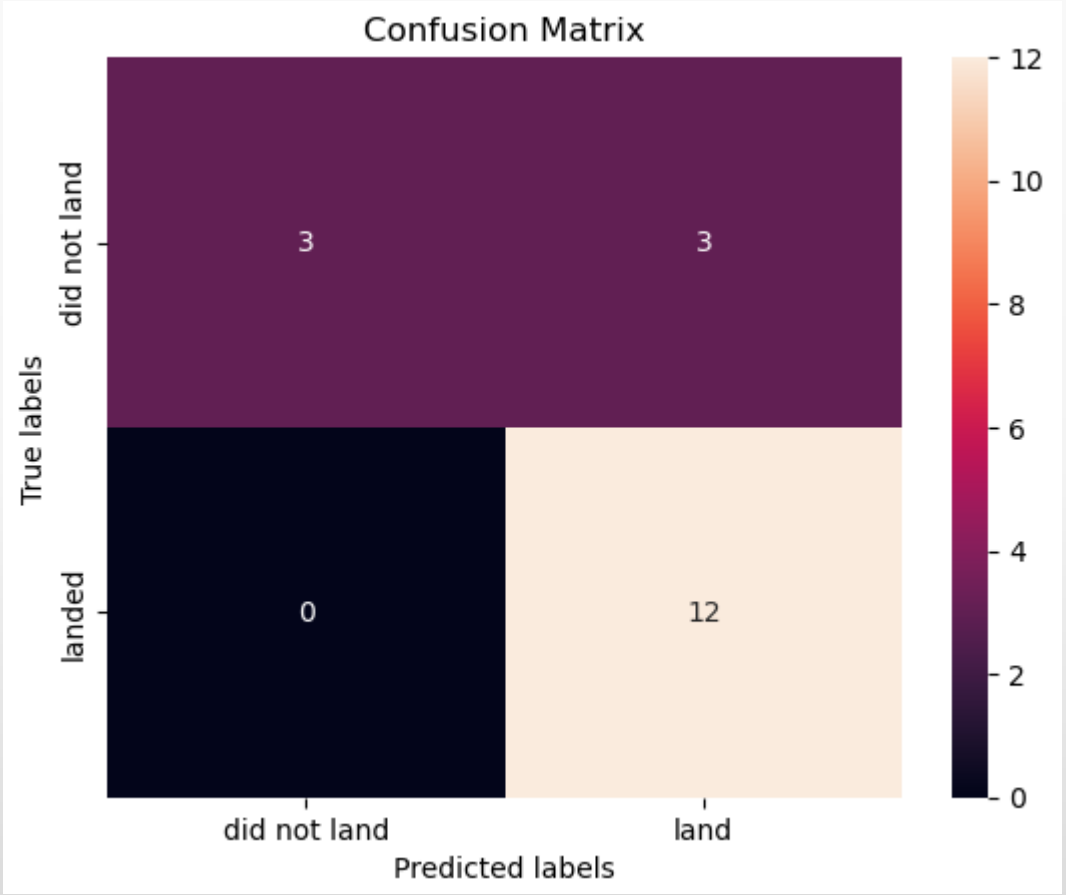
All four models performed at about the same level, having the same scores and accuracy. Although, the Decision Tree model slightly outperformed the rest.

	ML Method	Accuracy Score (%)
0	Support Vector Machine	83.333333
1	Logistic Regression	83.333333
2	K Nearest Neighbour	83.333333
3	Decision Tree	83.333333

Confusion Matrix

Performance Summary

- A confusion matrix summarizes the performance of a classification algorithm
- All the confusion matrices were identical
- The fact that there are false positives (Type 1 error) is not good
- The model predicted 12 successful landings



Section 6

Conclusions

Conclusions

- **Launch sites:** Most launches sites are strategically located “near” to de Equator because they use the equatorial orbit and the earth’s speed rotation as an additional natural boost to save time and fuel. Also, they are strategically located near highways and railways for transportation of personal and cargo, but also far away from cities for safety.
- **Model Performance:** The models performed similarly on the test set with the decision tree model slightly outperforming.
- **Launch Success:** Launches had been increasing over time. Having launch site KSC LC-39A with the highest success rate among launch sites.
- **Payload Mass:** Across all launch sites, the higher the payload mass (kg), the higher the success rate
- **Orbits:** ES-L1, GEO, HEO, and SSO have a 100% success rate.


Thank You!