IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Rakshan Reddy
11/22/2023

Outline

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

Executive Summary

Methodologies Used:
- **Data Collection** → collected data through SpaceX API and Wikipedia web scraping techniques

- **Data Wrangling** → used the collected data to analyze rocket launch successes and failures

- **Data Exploration** → used charts and visuals to observe factors including payload mass, flight type and launch site

- **Data Analysis** → used SQL and the visuals from the data exploration phase to draw conclusions on the relationships between spacecraft features

- **Data Visualization** → created interactive visuals that show success/fail rates of launches and the geographical location of launch sites

- **Predictive Modeling** → predicted landing outcomes through the use of logistic regression, SVM, KNN and decision trees

Results Summary:
- Over the years, launch success has mostly improved
- Orbits GEO, HEO, SSO and ES-L1 have never failed
- The majority of launch sites and near coasts
- The decision tree predictive model had a slight edge as the best predictive model

Introduction

Background + Context:

SpaceX is one of the world's most advanced manufacturers of rockets and spacecrafts in their journey to make life multiplanetary. Part of the reason SpaceX is so successful is because they reuse the first stage of their Falcon 9 spacecraft. Reusing this first stage makes it so that they can launch rockets and spacecrafts at a relatively low price compared to other competing companies. By analyzing the success rate of this first stage launch, we can determine if this method is ideal and if SpaceX should continue to use this model or change their strategy. As a competing company, we can also use some of their same methods if proven successful.

What to look for?

- Success/fail rates of SpaceX launches and how different factors including payload mass, launch site, and flight type affect these rates
- Most successful predictive modeling methods
- What SpaceX has done to increase their success rates

Section
1
Methodology

Methodology

Executive Summary

- **Data Collection** → collected data through SpaceX API and Wikipedia web scraping techniques

- **Data Wrangling** → cleaned and processed the collected data using one hot encoding

- **Data Exploration** → performed exploratory data analysis using charts and SQL

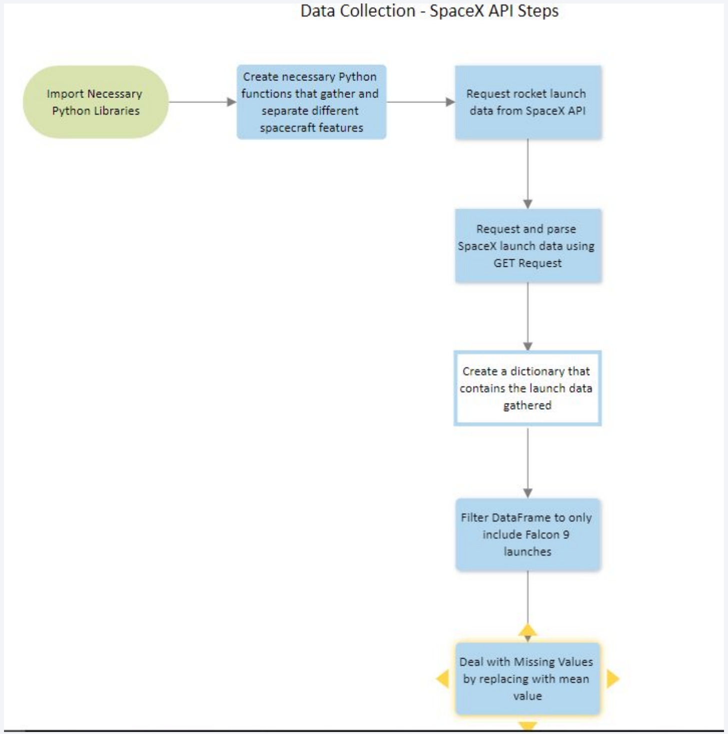
- **Data Visualization** → created interactive visuals that show success/fail rates of launches and the geographical location of launch sites using Folium and Plotly Dash

- **Predictive Modeling** → predicted landing outcomes through the use of logistic regression, SVM, KNN and decision trees while also observing heatmaps on predicted vs actual launch patterns

Data Collection - SpaceX API

- Reference flow chart on the right for process taken to collect data using SpaceX API

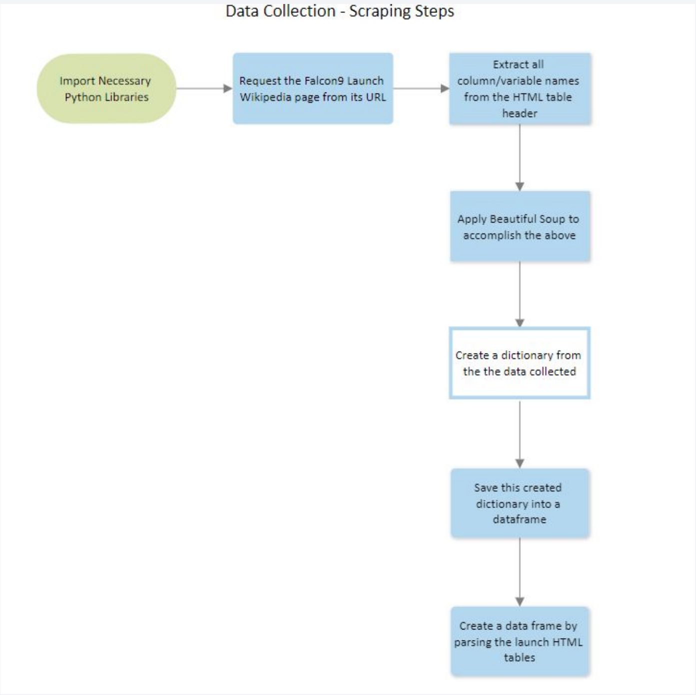
- SpaceX API Jupyter Notebook: https://github.com/Reddy991/SpaceX-Data-Science-Project/blob/main/%231%20Data%20Collection.ipynb



Data Collection - Scraping

- Reference flow chart on the right for process taken to collect data using Web Scraping

- Web Scraping Jupyter Notebook: https://github.com/Reddy991/SpaceX-Data-Science-Project/blob/main/%232%20Webscraping.ipynb



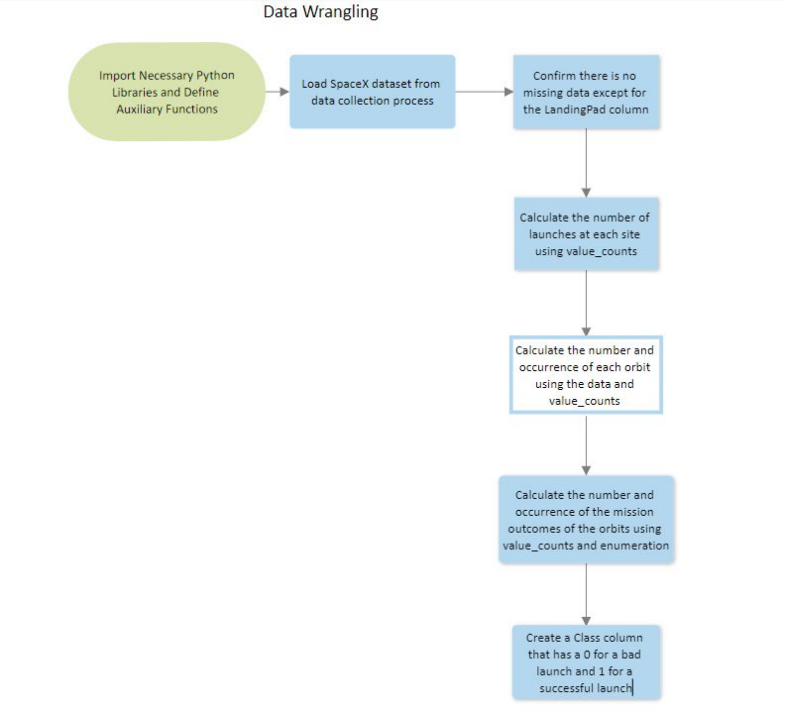
Data Wrangling

- Data was processed by observing the number of spacecrafts launched from different sites. Data was also processed by observing the number of orbits each spacecraft was put into and how this correlated to a successful and unsuccessful launch.

- Reference flow chart on the right for steps taken in the data wrangling process

- Data Wrangling Jupyter Notebook:

https://github.com/Reddy991/SpaceX-Data-Science-Project/blob/main/%233%20Data%20Wrangling.ipynb



EDA with Data Visualization

Charts Plotted:

- Flight Number vs. Payload Mass (kg)
- Flight Number vs. Launch Site
- Orbit Type vs Success Rate
- Flight Number vs. Orbit Type
- Payload Mass (kg) vs. Orbit Type
- Date vs. Success Rate

Use of Charts:

- To view the relationship between these different variables
- Compare what variables and factors are the most important in creating a successful spacecraft

EDA with Data Visualization Jupyter Notebook:

https://github.com/Reddy991/SpaceX-Data-Science-Project/blob/main/%235%20Exploratory%20Data%20Analysis%20with%20Visualizations.ipynb

EDA with SQL

SQL Queries Performed:

- Names of Unique Launch Sites
- 5 Records of Launch Sites Beginning with “CCA”
- Total Payload Mass of Boosters Launched by NASA (CRS)
- Average Payload Mass Carried by Booster F9 v1.1
- Date of First Successful Landing in Ground Pad
- Names of Boosters with Success in Drone Ship and Payload Mass between 4000 and 6000
- Total Number of Successful and Failure Mission Outcomes
- Names of Boosters that have Carried the Maximum Payload Mass
- Records for Failure Landing Outcomes in Drone Ship in 2015
- Count of Landing Outcomes between 2010-06-04 and 2017-03-20, in descending order.

EDA with SQL Jupyter Notebook:

https://github.com/Reddy991/SpaceX-Data-Science-Project/blob/main/%234%20Exploratory%20Data%20Analysis%20with%20SQL.ipynb

Build an Interactive Map with Folium

Launch Site Markers:

- Added red-colored markers at all SpaceX launch sites with appropriate longitude and latitude coordinates by hovering the mouse over the location

Launch Outcome Markers:

- Added green and red markers at all launch sites that indicate a successful launch (green) or a failed launch (red)

Purpose of Adding these Objects:

- To understand the location of all the launch sites in relation to the highways, coastlines and railroads around them and how the location of the launch sites affects a launch to be successful or unsuccessful

Folium Interactive Map Jupyter Notebook:

https://github.com/Reddy991/SpaceX-Data-Science-Project/blob/main/%236%20Launch%20Site%20Map%20Visualization.ipynb

Build a Dashboard with Plotly Dash

Pie Chart of Launches:

- Created a pie chart that showed the percentage of successful and unsuccessful launches compared with the total number of launches

Scatter Plot of Payload Mass (kg) vs. Success Rate:

- Created a scatter plot that shows the relationship between the payload mass of a rocket and its success rate

Payload Mass Adjuster:

- Changes the mass of the rocket so that users can observe the effects that different masses have on success rate

Purpose of Adding these Objects:

- To understand the true relationship between mass and successful launches. This dashboard overall helps us to look at success rates more concisely and the effect of different variables, such as mass, on this success rate.

Plotly Dash Dashboard Jupyter Notebook:

https://github.com/Reddy991/SpaceX-Data-Science-Project/blob/main/%237.1%20Interactice%20Analytics%20Dashboard%20Code.py

Predictive Analysis (Classification)

Summary:

- Found the best classification by using different predictive modeling methods including logistic regression, KNN, SVM, and decision trees and then comparing the methods to see which one had the greatest accuracy

Process:

- Standardardized and fit the data
- Split the data into two training groups and two testing groups
- Created a GridSearch object for each of the methods used (logreg, KNN, SVM, and decision trees) with cv=10
- Calculated the accuracy of each method using the score method
- Compared each of the models by analyzing the accuracy and heatmaps of all the methods

EDA with SQL Jupyter Notebook:

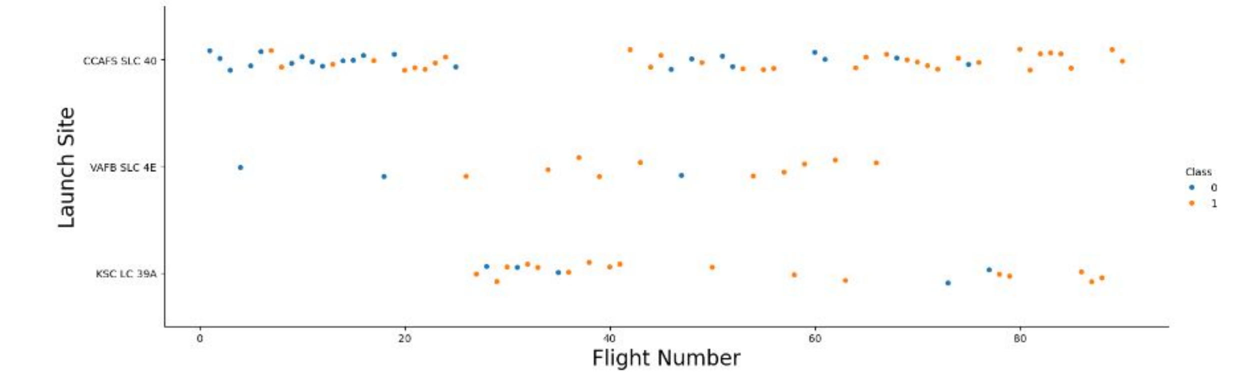
https://github.com/Reddy991/SpaceX-Data-Science-Project/blob/main/%238%20Machine%20Learning%20Predictions.ipynb

Section
2

Insights drawn from EDA

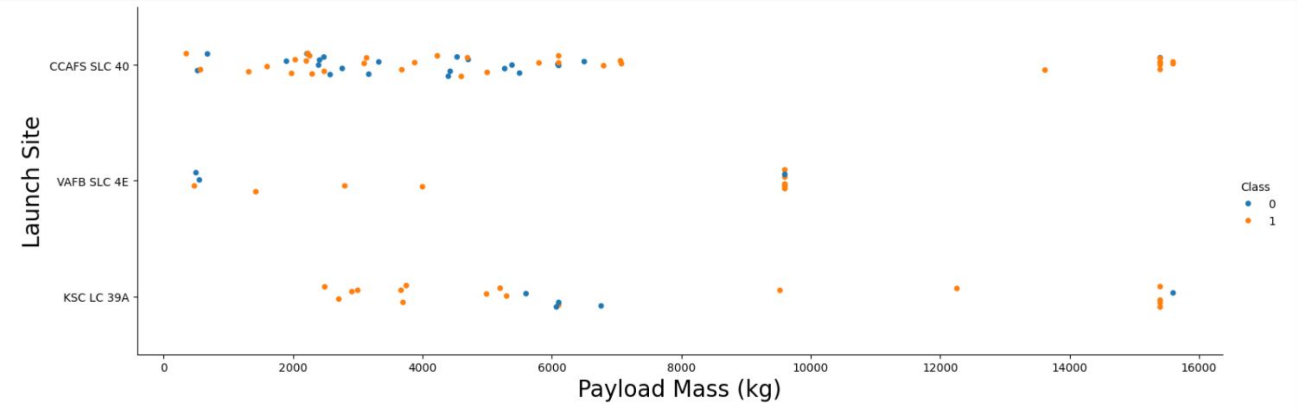
Flight Number vs Launch Site

- Orange = Successful Flight
- Blue = Failed Flight
- As new models of flights were built, the success rate grew compared to its original flights
- The VAFB SLC 4E launch site has not had a many launches after around flight number 68, but their past five launches have been successful
- The majority of launches are from launch site CCAFS SLC 40
- At all launch sites, newer flights have had higher success rates



Payload vs Launch Site

- Orange = Successful Flight
- Blue = Failed Flight
- Heavier launches are performed at the CCAFS SLC 40 and KSC LC 39A launch sites
- Launches with a payload mass more than 7,000 kg have a much higher success rate than launches with less than that payload mass (almost 100%)
- The VAFB SLC 4E launch site has not launched a flight with a payload mass greater than 10,000 kg

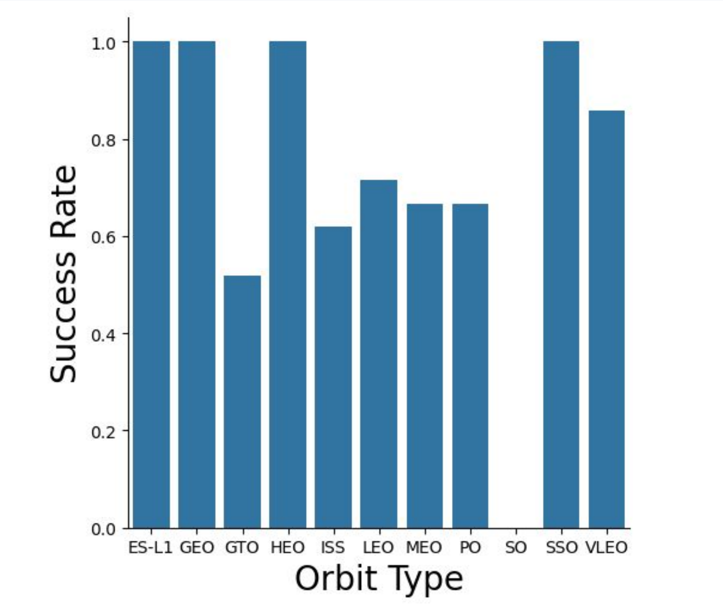


Success Rate vs Orbit Type

- The ES-L1, GEO, HEO, and SSO orbit types have a **100%** success rate

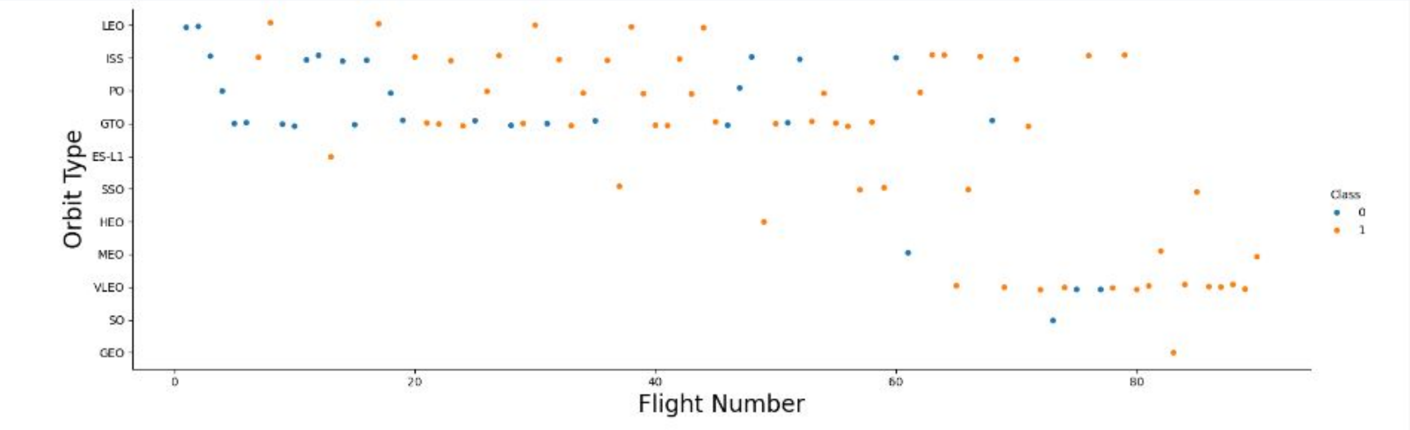
- The SO orbit type has a **0%** success rate

- All other orbit types including GTO, ISS, LEO, MEO, PO, and VLEO have between a **25% and 90%** success rate



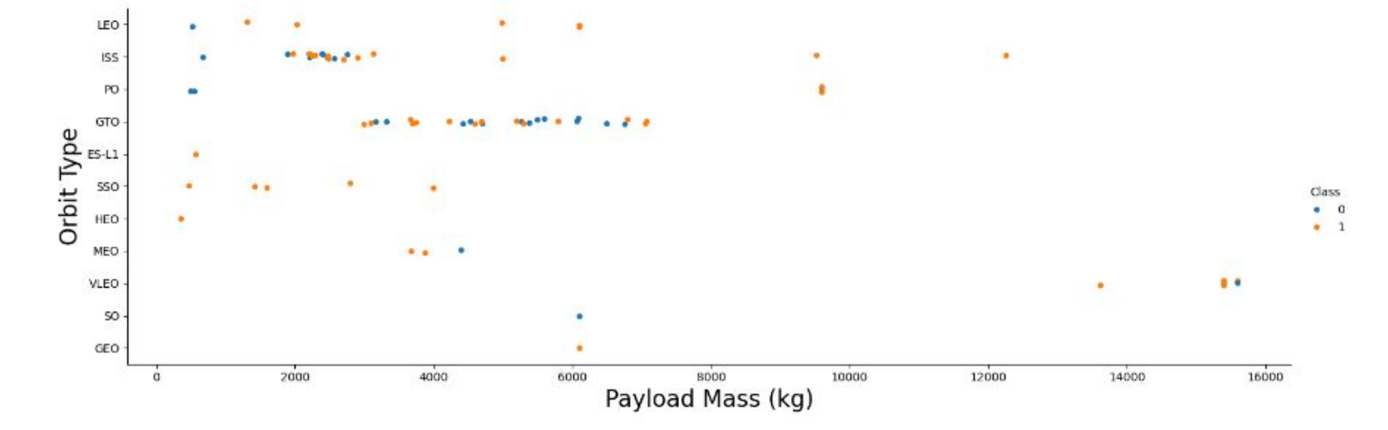
Flight Number vs Orbit Type

- As the flight number increases, the success rate of the orbit types usually increases
- This is seen clearly in the LEO and SO orbits
- However, the GTO orbit still has a failure even after the flight number surpasses 60, and does not follow this pattern



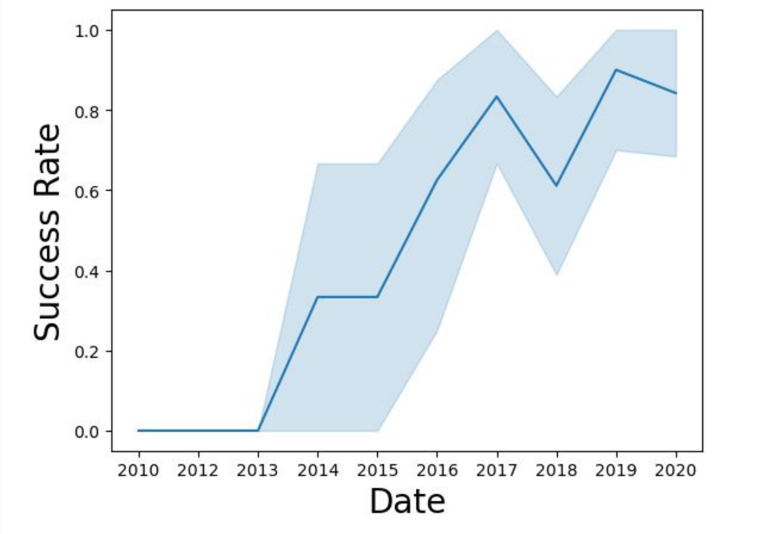
Payload vs Orbit Type

- Heavier payloads in the ISS, and PO orbit types generally have better success rates
- The GTO orbit has a lot of launches, all below 8,000 kg which have a varied success rate
- The VLEO orbit type is not fully successful as payload mass increases



Launch Success Yearly Trend

- From 2013-2017 the launch success rate increased or stayed constant. It never decreased during this period
- From 2017-2018 the launch success rate decreased
- From 2018-2019 the launch success rate increased again
- However, it looks like the launch success rate after 2019 is starting to decrease again



All Launch Site Names

Unique Launch Site Names:

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

SQL Query Explanation:

- Selecting all the unique (distinct) launch site names from the SPACEXTBL database

```
%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 of Launch Site Names beginning with CCA to the right

SQL Query Explanation:

- Selecting all records from the SPACEXTBL database where the launch site has a name that has CCA in it. Also filtering the number of records the dataset gives back to 5

```
%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 mass from NASA (CRS) is 45,596 kg

SQL Query Explanation:

- Selecting the sum of all payload masses from the SPACEXTBL dataframe where the customer/user is NASA (CRS)

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE CUSTOMER = "NASA (CRS)";
```

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

The average payload mass of the F9 v1.1 booster is 2928.4 kg

SQL Query Explanation:

- Selecting the average payload mass from the SPACEXTBL dataframe where the booster version is 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

First Successful Ground Landing Date

The first successful ground landing date was December 22nd, 2015 (2015-12-22)

SQL Query Explanation:

- Selecting the minimum date from the SPACEXTBL dataframe where the landing outcome was a success on ground pad

```
%sql SELECT MIN(DATE) FROM SPACEXTBL WHERE Landing_Outcome = "Success (ground pad)";
```

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

MIN(DATE)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

The booster versions of the successful drone ship landings with payload mass between 4000 kg and 6000 kg are F9 FT B1022, F9 FT B1026, F9 FT B1021.2, and F9 FT B1031.2

SQL Query Explanation:

- Selecting the booster versions from the SPACEXTBL data frame where the landing outcome is a success (drone ship) and the payload mass is between 4000 kg and 6000 kg

```
%sql SELECT Booster_Version FROM SPACEXTBL WHERE Landing_Outcome = "Success (drone ship)" AND PAYLOAD_MASS__KG_ BETWEEN 400(
```

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

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 was 1 in-flight failure, 99 successes, and 1 success where the payload status is unclear

SQL Query Explanation:

- Selecting the mission outcome and count of the all rows from the SPACEXTBL data frame grouped by the mission outcome

```
%sql SELECT Mission_Outcome, COUNT(*) as Total_number_mission_outcomes FROM SPACEXTBL GROUP BY Mission_Outcome
```

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

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

Boosters Carried Maximum Payload

The boosters that carried the maximum payload are shown to the right. All booster versions were F9 B5 versions.

SQL Query Explanation:

- Selecting the booster versions from the SPACEXTBL data frame where the payload mass is equal to the max payload mass from the SPACEXTBL data frame

```
%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 two records of a failed drone ship landing in 2015 were in January and April both from the CCAFS LC-40 launch site and F9 v1.1 booster version.

SQL Query Explanation:

- Selecting the month, landing outcome boost version and launch site from the SPACEXTBL data frame where the landing outcome is failure by drone ship and the year is 2015

```
%sql SELECT substr(Date,6,2) as Month, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTBL WHERE Landing_Outcome =
```

```
 * 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

There were 8 successful landings from ground pad between June 4, 2010 and March 20, 2017.

SQL Query Explanation:

- Selecting the landing outcome and count of all rows from the SPACEXTBL data frame where the date is between 2010-06-04 and 2017-03-20

```
%sql SELECT Landing_Outcome, COUNT(*) as Total FROM SPACEXTBL WHERE Date BETWEEN "2010-06-04" and "2017-03-20" AND ORDER BY
```

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

Landing_Outcome	Total
Success (ground pad)	8

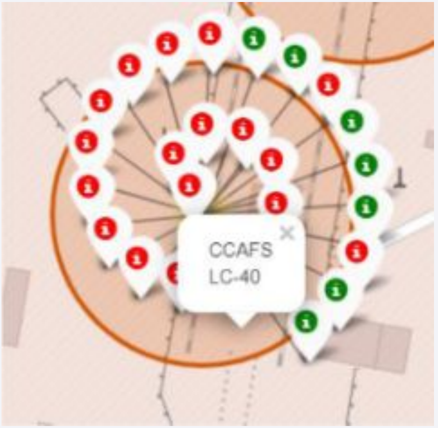
Section
3

Launch Sites Proximities Analysis

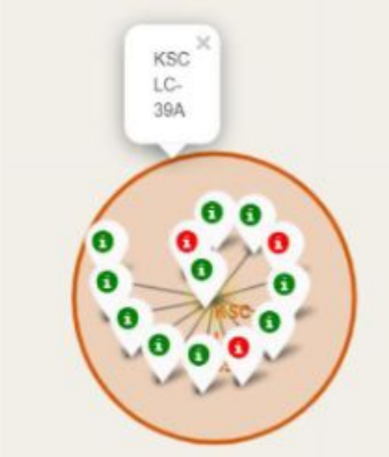
Folium Map Launch Sites

SpaceX launch sites are located in California and Florida exclusively

Colored Labeled Launch Outcomes

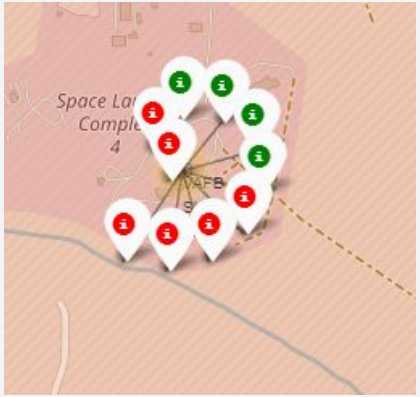


CCAFS LC-40 Launch Site



KSC LC-39A Launch Site

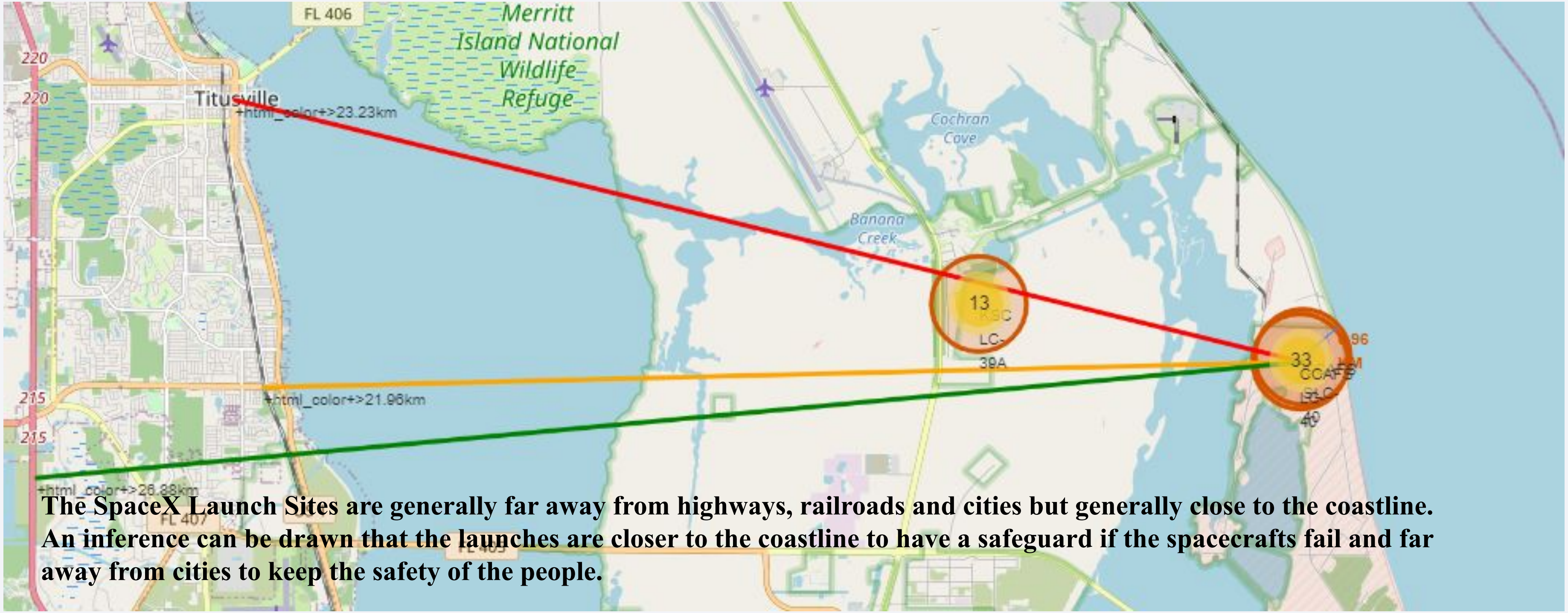
CCAFS SLC-40 Launch Site



CCAFS LC-40 Launch Site

From these maps it can be seen that the KSC LC-39A has the highest launch success rate whereas all the other launch sites have a success rate equal to or less than 50%. It can also be observed that the most number of launches are performed at the CCAFS LC-40 Launch Site which may indicate it as SpaceX's go to launch site.

Launch Site Proximites

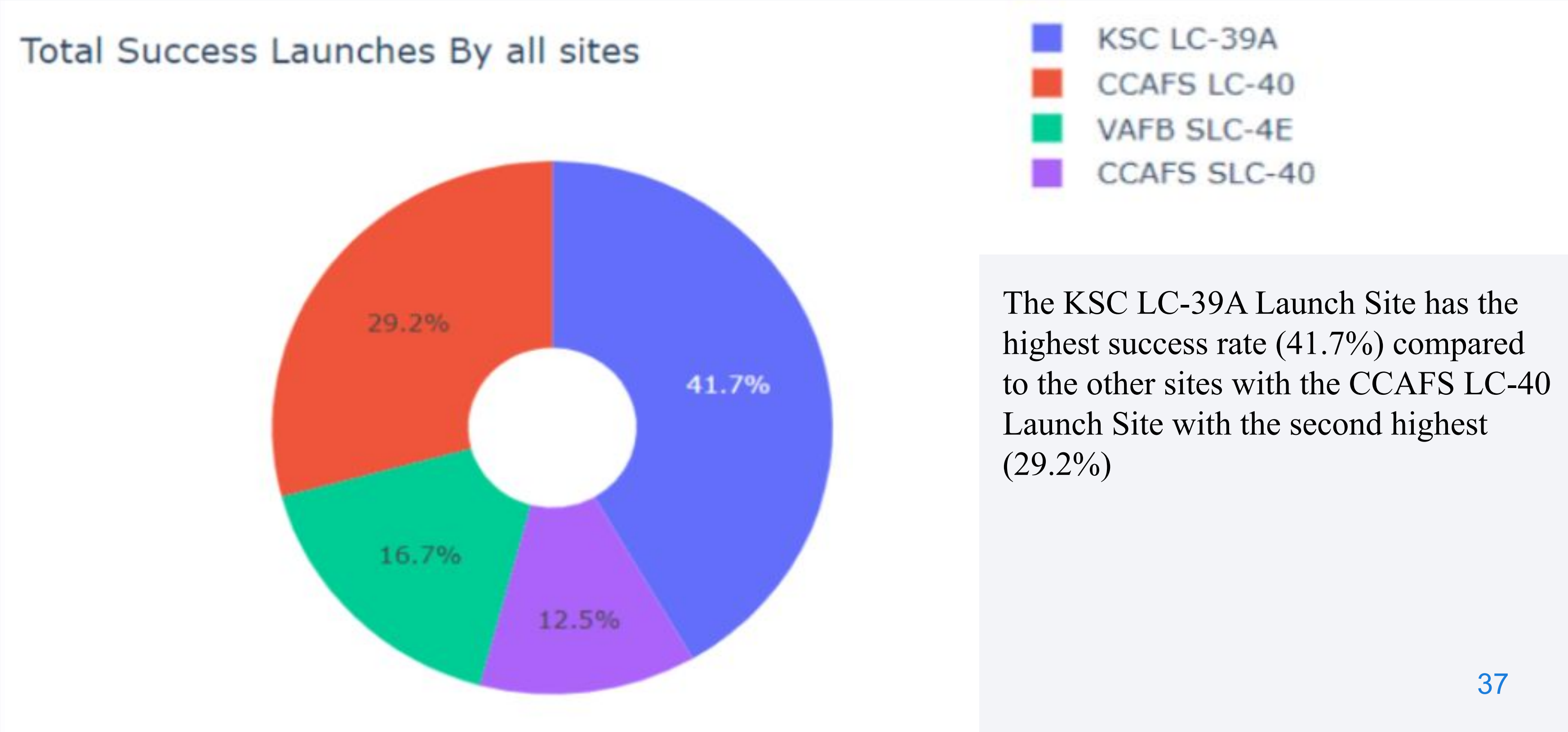


The SpaceX Launch Sites are generally far away from highways, railroads and cities but generally close to the coastline. An inference can be drawn that the launches are closer to the coastline to have a safeguard if the spacecrafts fail and far away from cities to keep the safety of the people.

Section
4

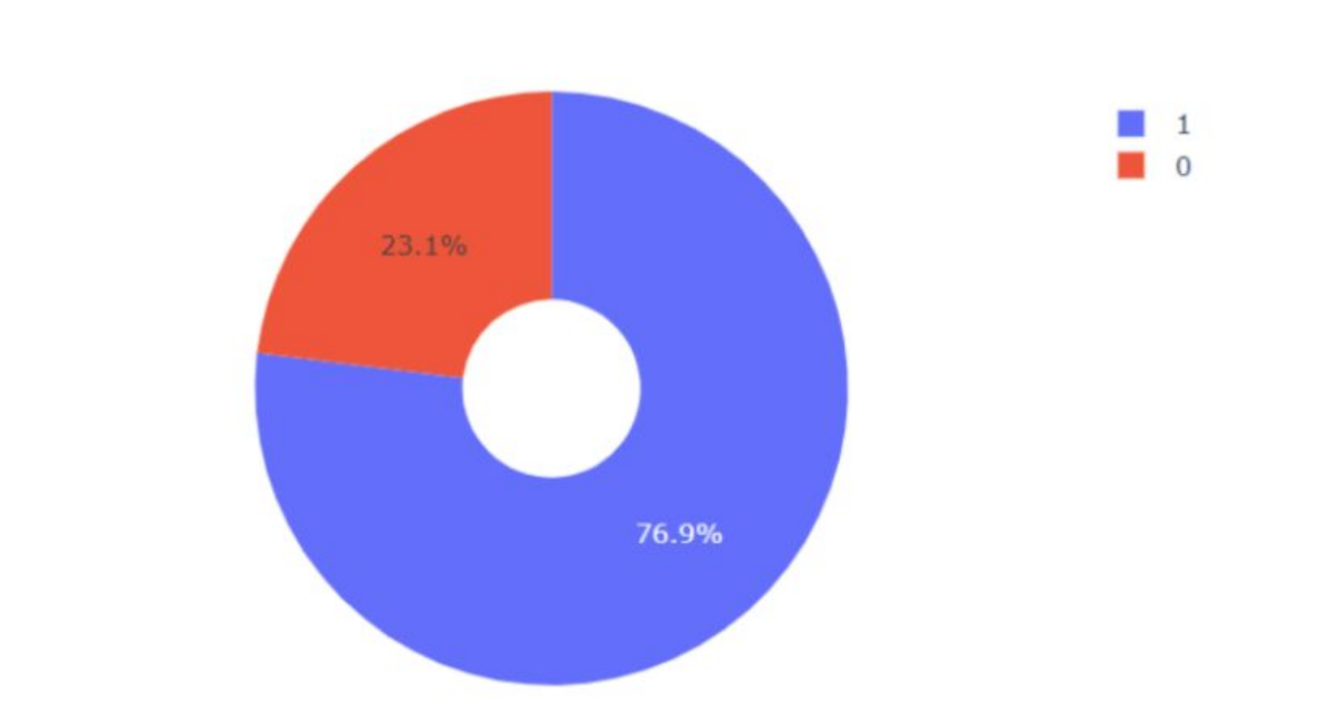
Build a Dashboard with Plotly Dash

Launch Success Pie Chart



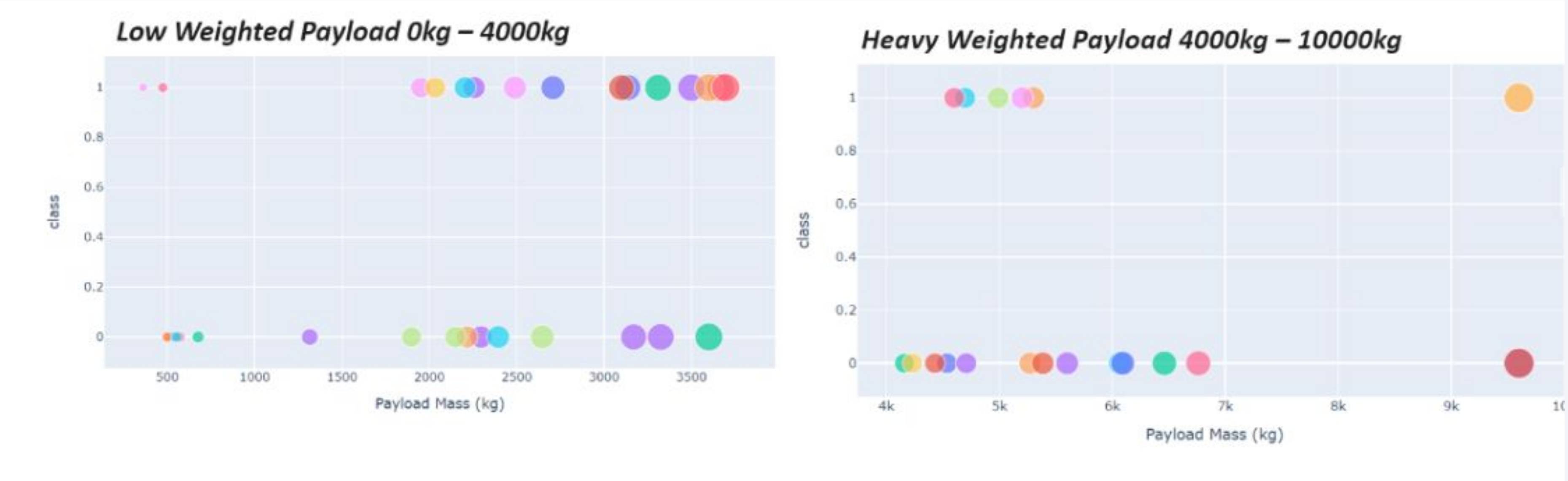
The KSC LC-39A Launch Site has the highest success rate (41.7%) compared to the other sites with the CCAFS LC-40 Launch Site with the second highest (29.2%)

Highest Launch Success Ratio Pie Chart



This pie chart displays the KSC LC-39A Launch Site (which had the highest launch success) success ratio and shows that this launch site had a 76.9% success to failure ratio.

Payload vs. Launch Outcome Scatter Plot



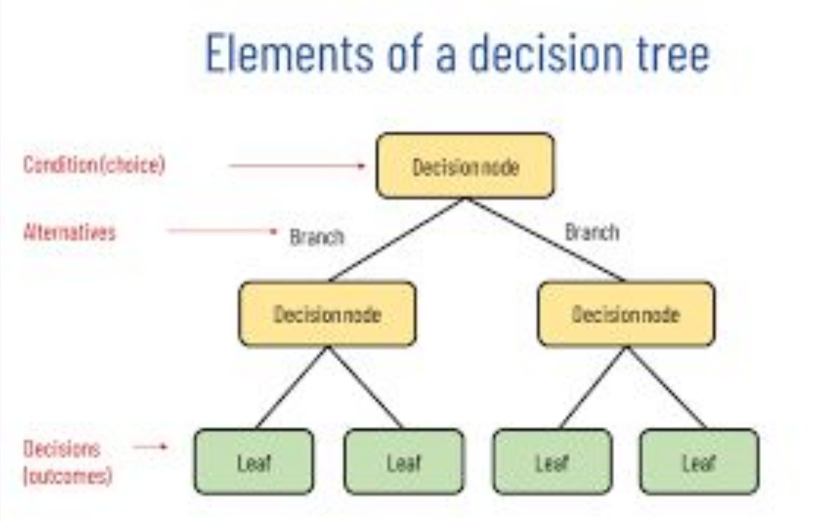
The success rates for the payloads between 2,000 kg and 6,000 kg are generally higher than the lower mass payloads than this and the higher mass payloads than this

Section
5

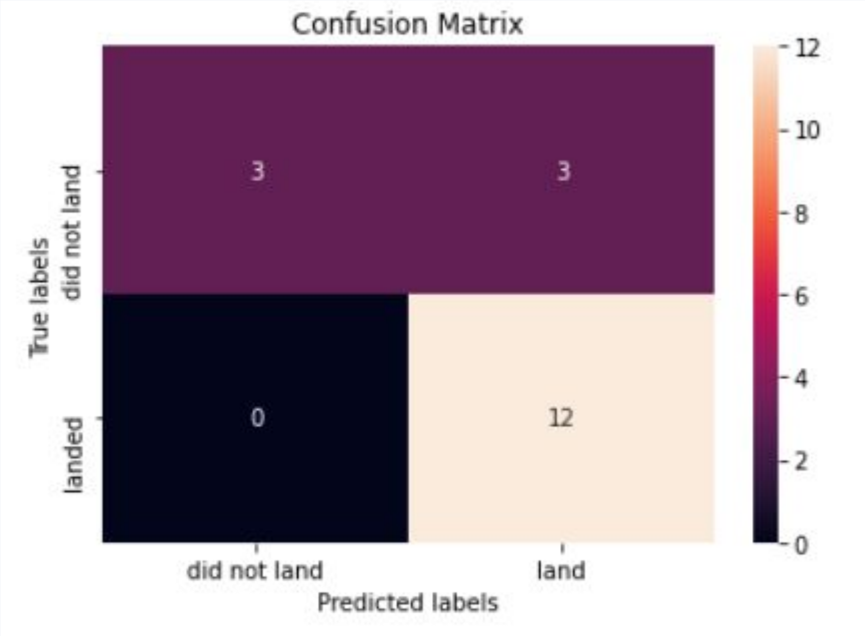
Predictive Analysis (Classification)

Classification Accuracy

- The decision tree model was proven to have the highest classification accuracy with a value of 0.8732 after testing all other possible models including logistic regression, SVM, and KNN.



Confusion Matrix



This is the confusion matrix of the decision tree model that had the highest accuracy out of all the predictive modeling practices. As shown by this matrix, the only problem with this model is the rate of false positives which was collected as 3 total here.

Conclusions

- Launch Success Rates have increased dramatically since 2013
- The KSC LC-39A Launch Site had the highest launch success rate out of all the SpaceX launch sites
- The Decision Tree predictive modeling method is the best to use here because of it having the highest accuracy of all the models
- Orbits ES-L1, GEO, and SSO had some of the highest success rates
- The Location of SpaceX centers is strategically placed close to coastlines and generally a good distance away from civilian areas

Appendix

- **Complete project GitHub Link:**
  https://github.com/Reddy991/SpaceX-Data-Science-Project/tree/main

Thank you!