IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Yanting LEO

September 20, 2023

Outline

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

Executive Summary

In this capstone project, we will predict if the SpaceX Falcon 9 first stage will land successfully using several machine learning classification algorithms. The main steps in this project include: Data collection,

wrangling, and formatting Exploratory data analysis Interactive data visualization Machine learning prediction Our graphs show that some features of the rocket launches have a correlation with the outcome of the launches, i.e., success or failure. It is also concluded that decision tree may be the best machine learning algorithm to predict if the Falcon 9 first stage will land successfully.

Introduction

- SpaceX rocket launches cost significantly less than other company launches due to SpaceX's ability to reuse Stage One of their Falcon 9 rocket launcher. It would be helpful to gain valuable insights on what conditions can be met to increase the chance of the Stage One recovery and reuse after a launch by answering the following:
- How often are SpaceX able to recover and reuse the Stage One components?
- What launch site properties impact successful landings?
- What other factors contribute to a successful lancing?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using the SpaceX API, I retrieved launch data involving the rockets used, launch site information, mass payload and orbit details, along with the landing outcomes and the respective core information.
- Perform data wrangling
  - To clean and organize the data for analysis, I built our dataset only on the launch data for the Falcon 9 rockets launches on or before November 11, 2020. To clean the data, we addressed missing data values to maintain dataset integrity.
- Perform exploratory data analysis (EDA) using visualization and SQL
  - I analyzed the rocket launch data by retrieving and comparing the launch sites, booster type, orbits and flight numbers to determine if there were any patterns that could affect successful or failed missions.

Methodology

Executive Summary 2/2

- Perform interactive visual analytics using Folium and Plotly Dash
    - I implemented a dashboard to allows users to view success rates and the correlation between Payload mass and Success rates by Launch Site.
- Perform predictive analysis using classification models
    - I used different classification algorithms to train on the a partition of the data from the SpaceX dataset.
- How to build, tune, evaluate classification models
    - Using Grid Search, I identified parameters to use in the respective Classification models to identify the best model to use to achieve high accuracy during predictions.

Data Collection – SpaceX API

- Launch History:

GET https://api.spacexdata.com/v4/launches/past

- Rocket Information:

GET https://api.spacexdata.com/v4/rockets/ID

- Launchpad Information:

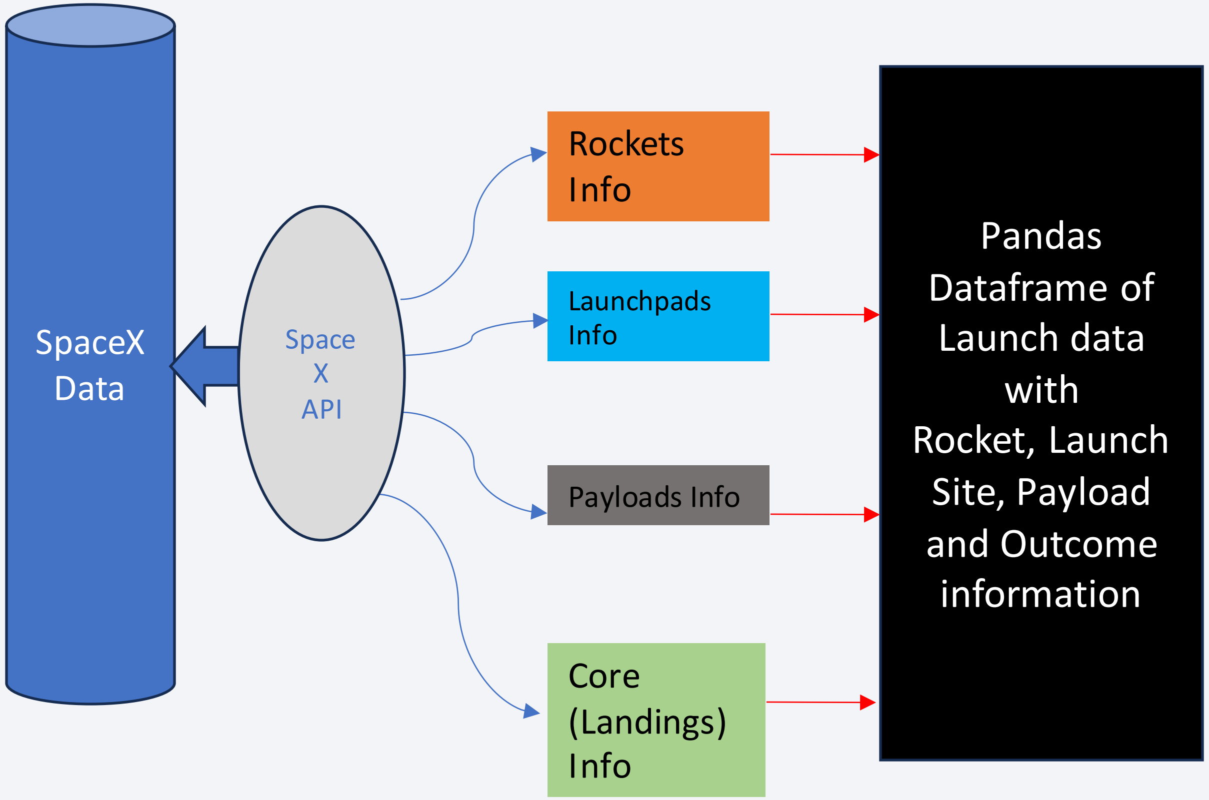
GET https://api.spacexdata.com/v4/launchpads/ID

- Payload Information

GET https://api.spacexdata.com/v4/payloads/ID

- Cores (Landing) Information

GET https://api.spacexdata.com/v4/cores/ID



Jupyter notebook:

https://github.com/Tigrisriver/IBM-Date-Science/blob/main/spacex-data-collection-api.ipynb

Data Collection - Scraping

Extract Falcon 9 launch records from HTML table Wikipedia

↓

Parse the HTML table to extract the data into a Python Dictionary

↓

Convert the Python dictionary into a Pandas DataFrame

https://github.com/Tigrisriver/IBM-Date-Science/blob/main/spacex-data-wrangling.ipynb

Data Wrangling

- The SpaceX data was pre-processed in from the launch history data to extract
  - The launch sites where the launches took place.
  - The orbits rockets took during their respective journeys.
  - The landing outcomes for each launch
- Feature engineering was used to add a binary column, Class, to represent whether a launch landing was considered a success or failure
- Converting categorical data numerically for prediction models.

Jupyter notebook:

https://github.com/Tigrisriver/IBM-Date-Science/blob/main/spacex-data-wrangling.ipynb

EDA with Data Visualization

To get a clearer picture of the possible affects the different rockets, orbits, and payload details may have had on the mission outcomes, various data visualizations were generated to compare the results between successful landings and unsuccessful landings.

Jupyter notebook:

https://github.com/Tigrisriver/IBM-Date-Science/blob/main/spacex-eda-dataviz.ipynb

EDA with SQL

- Date of first successful landing outcome
- Boosters
  - With largest payloads
  - With smaller payloads but successful landings
- Boosters of Successful Landings with small payloads
- Types of landings:
  - More Successes than Failures
  - Controlled Ocean landings were the most frequent

https://github.com/EveThan/IBM-Applied-Data-Science-Capstone-Project/blob/main/4_EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

My Folium Map has circle markers to represent the three different launch sites used for the SpaceX launches. I also included cluster markers of the landing results successes and failures for the viewer to observe the clustering of like results per landing site. I also added distance line segments between launch sites and the closest railway, highways, coastline or city viewers were interested in the proximity of these locations to the lunch sites.

Build a Dashboard with Plotly Dash

- The interactive dashboard shows a pie chart of the proportions of landing success and failures as well as a scatter plot of the relationship between payload mass and landing successes and failures. In order for stakeholders to be able to gain insight on any patterns behind successful and failed SpaceX launches, it is important that they are able to interactively see the relationships between launch sites, rocket boosters, and payload mass and the success of the launch.

https://github.com/Tigrisriver/IBM-Date-Science/blob/main/spacex_dash_appv1.py

Predictive Analysis (Classification)

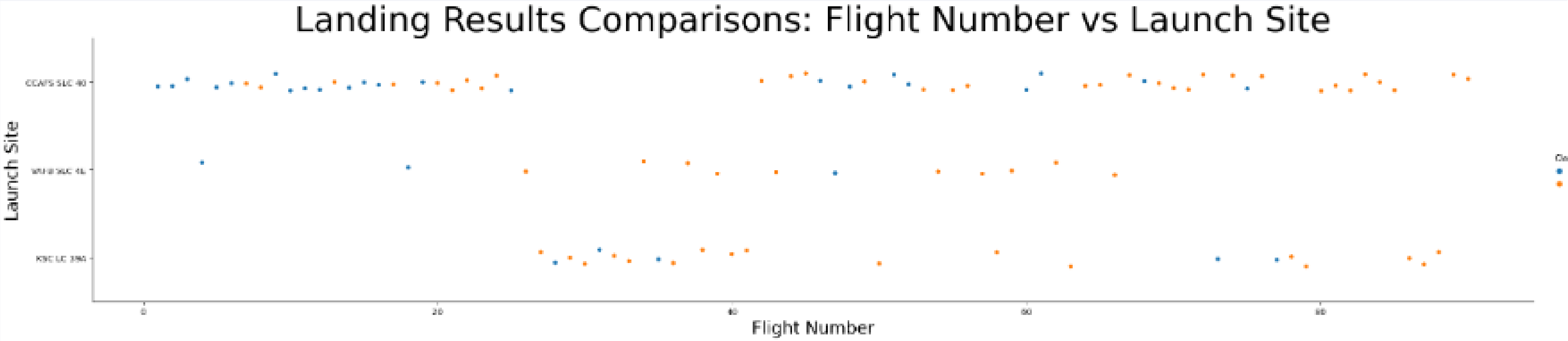
- In order to prepare the data for the Machine Learning prediction process, I standardized the data, and split it into a training set and a test set.
- I then trained the data on the following Machine Learning Algorithms and used GridSearch to tune the hyperparameters
  - Logistic Regression
  - Support Vector Machine (SVM)
  - Decision Tree Classification
  - *k*-Nearest Neighbors (KNN)
- We then compared the training accuracy of each model and determined model produced the best accuracy.

https://github.com/Tigrisriver/IBM-Date-Science/blob/main/spacex-machine-learning-predictions.ipynb

Section 2

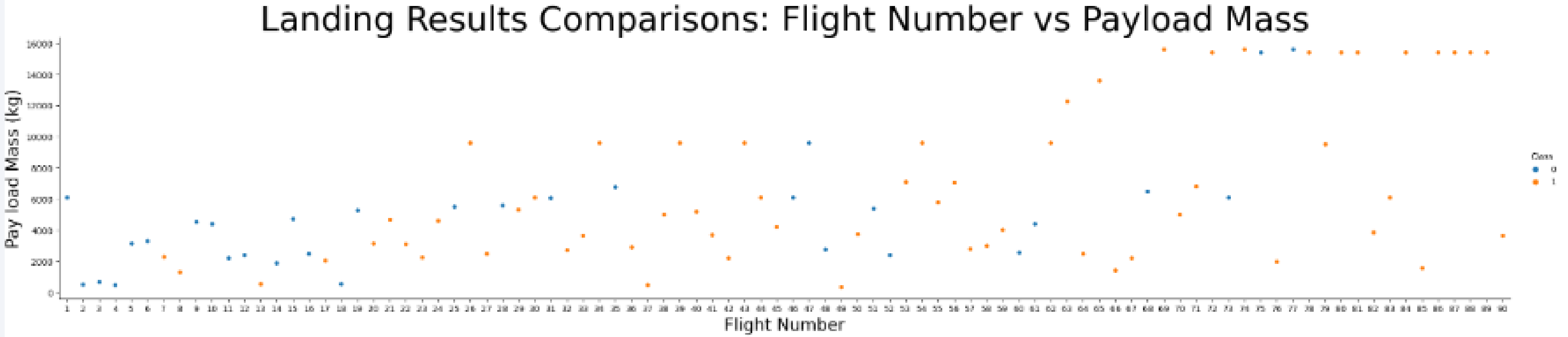
Insights drawn from EDA

Flight Number vs. Launch Site



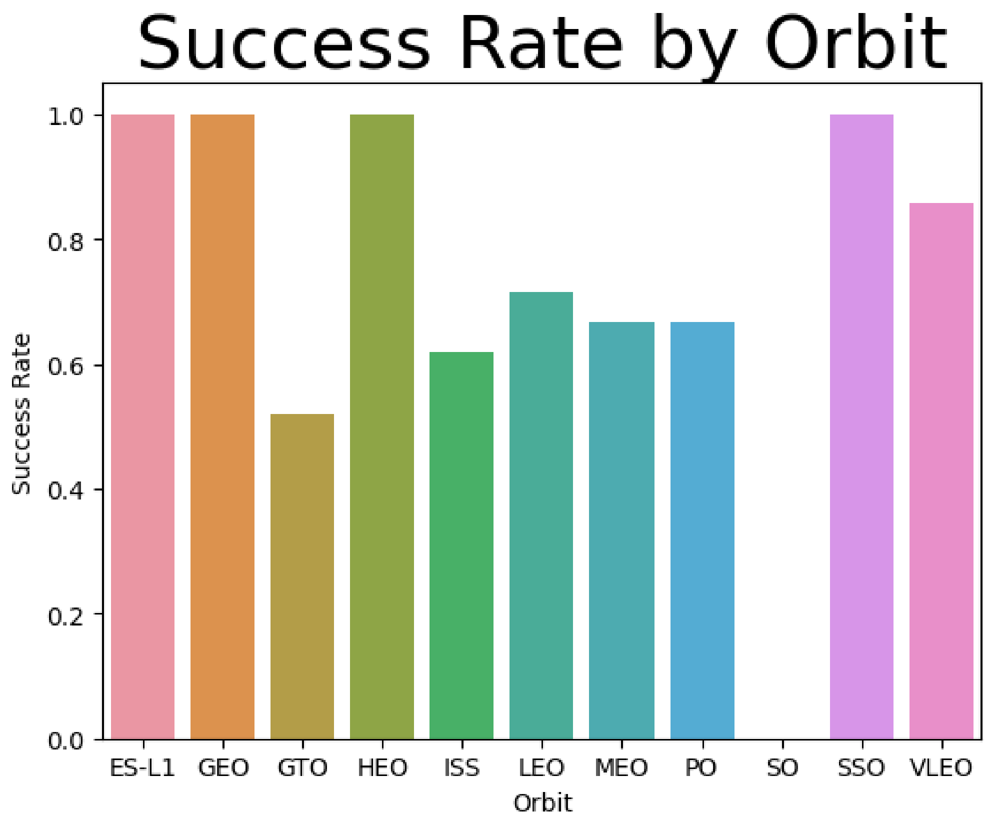
The scatterplot show that as the flight number increases, the flights are more likely to land successfully regardless of the launch site. I can also see that the `CCAFS LC-40` and the `VAFB SLC 4E` sites, initially had several failed landings before the flights started landing successfully. The scatterplot shows that although the `VAFB SLC 4E` launch site had the overall higher success rate in comparison to other sites, it also had less flight launches than the other launch sites. The most flight launches initiated from `CCAFS LC-40` site.

Flight Number vs Payload



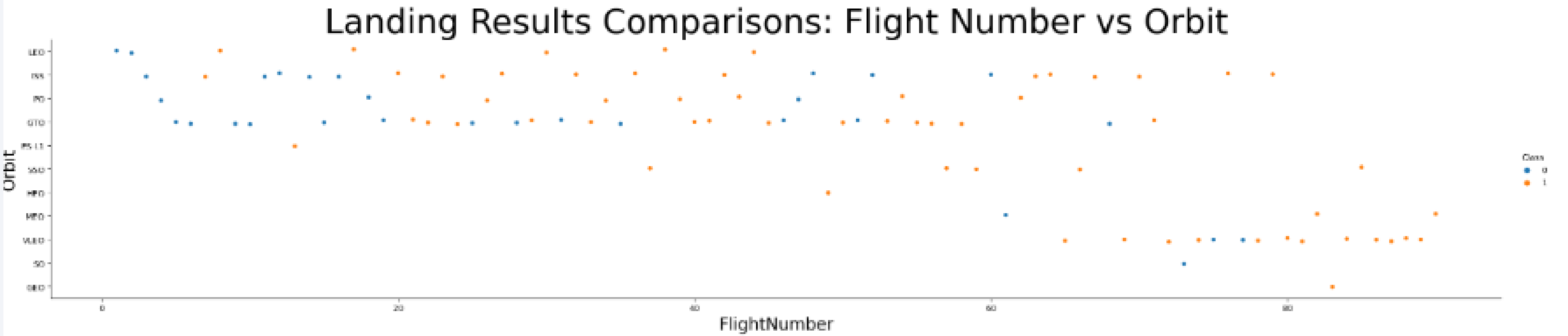
We see that different launch sites have different success rates. `CCAFS LC-40`, has a success rate of 60 %, while `KSC LC-39A` and `VAFB SLC 4E` has a success rate of 77%.

Success Rate vs. Orbit Type



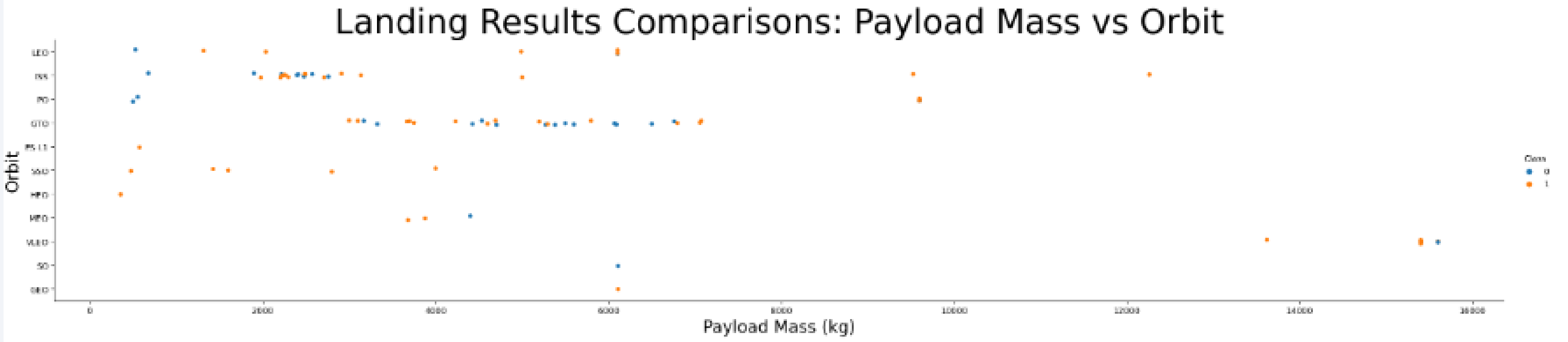
The flights that travelled the `ES-L1`, `GEO`, `HEO`, `SSO` orbits all landed successfully. Although the `VLEO` orbit flights had some landing failures, over all, the vast majority of the flights traveling the `VLEO` orbit were successful. The flight that used the `GTO` orbits had the lowest successful landing rate.

Flight Number vs. Orbit Type



In the `LEO` orbit the sucessful landing appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in `GTO` orbit.

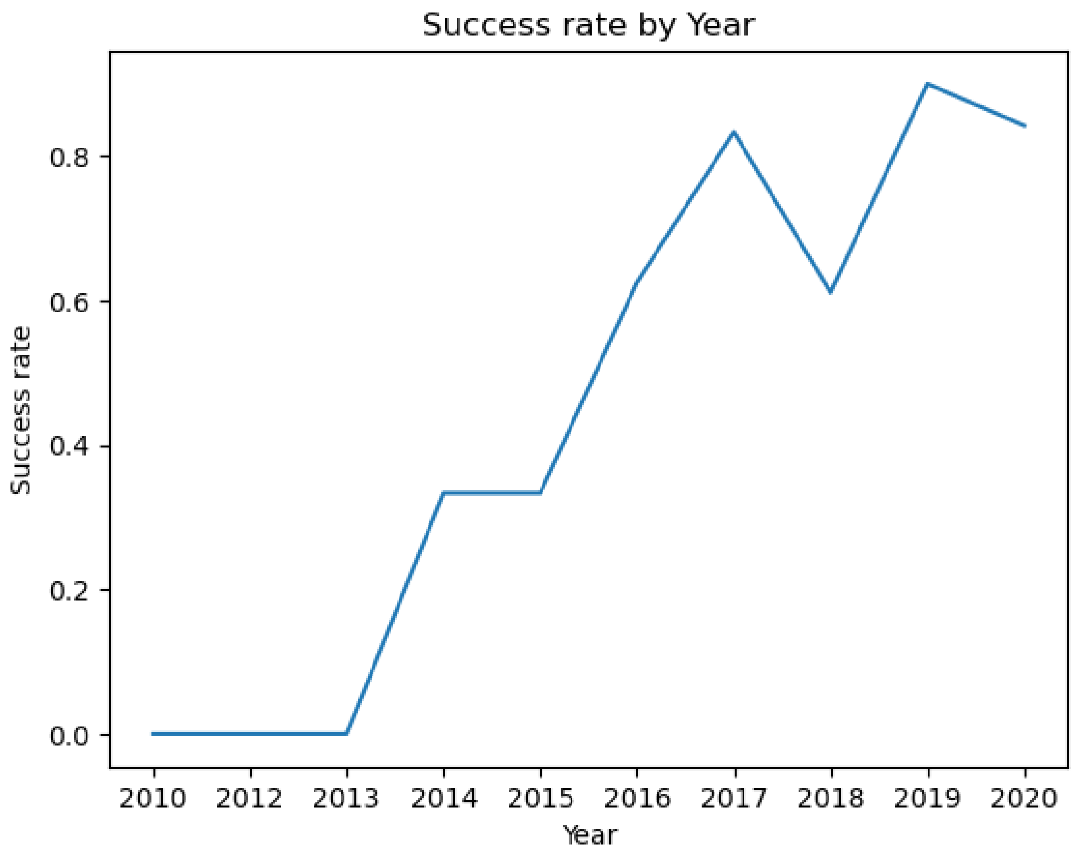
Payload vs. Orbit Type



With heavy payloads there are more for succesful landings for `Polar`, `LEO and ISS orbits than there are failed landings.` `For the lighter payloads (less than 4,000 kg), there are only successful landings for the SSO , HEO and MEO orbits.`

However for the GTO orbits we cannot distinguish this well as both positive landing rate and negative landing (unsuccessful mission) are both there here.

Launch Success Yearly Trend



In the beginning of the program the success rate was minimal until 2013. I can see that that the success rate increased consistently and significantly between 2013 and 2017, with a significant dip between 2017 and 2018. Since 2020, the success rate has been declining but not significantly.

All Launch Site Names

To identify the names of the various launch sites contained in the historical data retrieved of the SpaceX launches, we evaluated the following query to retrieve the complete list of the four launch sites.

```
%sql SELECT DISTINCT Launch_site FROM SPACEXTABLE
```

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

Launch Site Names Begin with 'CCA'

To review launch data for the Florida launch sites (CCAFLS SLC-40 and CCAFLS LC-40), we ran the following query:

```
%sql SELECT * FROM SPACEXTABLE WHERE Launch_site LIKE 'CCA%' LIMIT 5;
```

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

Total NASA Payload Mass

- To determine the amount of payload NASA utilized SpaceX launches for, we queried the database of all the launches with NASA as a customer and summed the payload mass.

Customer	Total_Payload_Mass
NASA (CRS)	48213

```
%sql SElECT Customer, SUM(PAYLOAD_MASS__KG_) AS Total_Payload_Mass FROM SPACEXTABLE WHERE Customer LIKE '%NASA (CRS)%';
```

Average Payload Mass by F9 v1.1

As different Rocket Boosters were used in the launch data, I calculated the average payload mass that launches involving the F9 v1.1 B1003 carried.

Booster_Version	Average_Payload
F9 v1.1 B1003	2534.6666666666665

```
%sql SELECT Booster_Version, AVG(Payload_MASS__KG_) AS Average_Payload FROM SPACEXTABLE
    WHERE Booster_Version LIKE 'F9 v1.1%';
```

First Successful Ground Landing Date

Curious about when the first launch with a successful ground pad landing? I queried the database and identified it to have occurred on December 22, 2015.

First_Landing
2015-12-22

```
%sql SELECT MIN(Date) AS First_Landing FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)';
```

Successful Drone Ship Landing with Payload between 4000 and 6000

Below are the names of boosters involved with launches carrying between 4,000 and 6,000 kg payloads that successfully landed on drone ships

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

```
%sql SELECT DISTINCT Booster_Version FROM SPACEXTABLE
    WHERE Landing_Outcome = 'Success (drone ship)' AND
        Payload_mass__KG_ BETWEEN 4000 and 6000;
```

Total Number of Successful and Failure Mission Outcomes

Missions fail when their landings are not successful. To understand the different types of successful and failed landings thus missions, we queried launch data to compile a list and associated frequency of their occurrences in the historical SpaceX launch data.

Successful	Failed
61	40

```
%%sql SELECT DISTINCT
(SELECT COUNT(*) FROM SPACEXTABLE WHERE Landing_Outcome LIKE 'Success%') AS Successful,
(SELECT COUNT(*) FROM SPACEXTABLE WHERE Landing_Outcome NOT LIKE 'Success%') AS Failed
FROM SPACEXTABLE;
```

Landing_Outcome	Frequency
Controlled (ocean)	5
Failure	3
Failure (drone ship)	5
Failure (parachute)	2
No attempt	21
No attempt	1
Precluded (drone ship)	1
Success	38
Success (drone ship)	14
Success (ground pad)	9
Uncontrolled (ocean)	2


```
%sql SELECT DISTINCT Landing_Outcome, COUNT(*) as Frequency FROM SPACEXTABLE GROUP BY Landing_Outcome;
```

Boosters Carried Maximum Payload

- To identify which Boosters are capable of carrying the largest payloads during the launch, I queried the launch data to compile the list.

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

```
%%sql SELECT DISTINCT Booster_Version, Payload_mass__KG_ FROM SPACEXTABLE
      WHERE Payload_mass__KG_ =
            (SELECT Max(Payload_mass__KG_) as Max_Payload
            FROM SPACEXTABLE)
      ORDER BY Booster_Version;
```

2015 Launch Records

The details regarding the 2015 launches that resulted in failed drone ship landings are below.

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

```
%%sql SELECT Date, Landing_outcome, Booster_version, Launch_site
      FROM SPACEXTABLE
      WHERE Landing_outcome = 'Failure (drone ship)'
          AND Date LIKE '2015%';
```

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

Below are the landing outcomes (in order of decreasing frequency) of SpaceX launches between June 4, 2010 and March 3, 2017

Landing_Outcome	Frequency
Controlled (ocean)	30
Failure (drone ship)	25
Failure (parachute)	24
No attempt	14
Precluded (drone ship)	13
Success (drone ship)	8
Success (ground pad)	3
Uncontrolled (ocean)	1

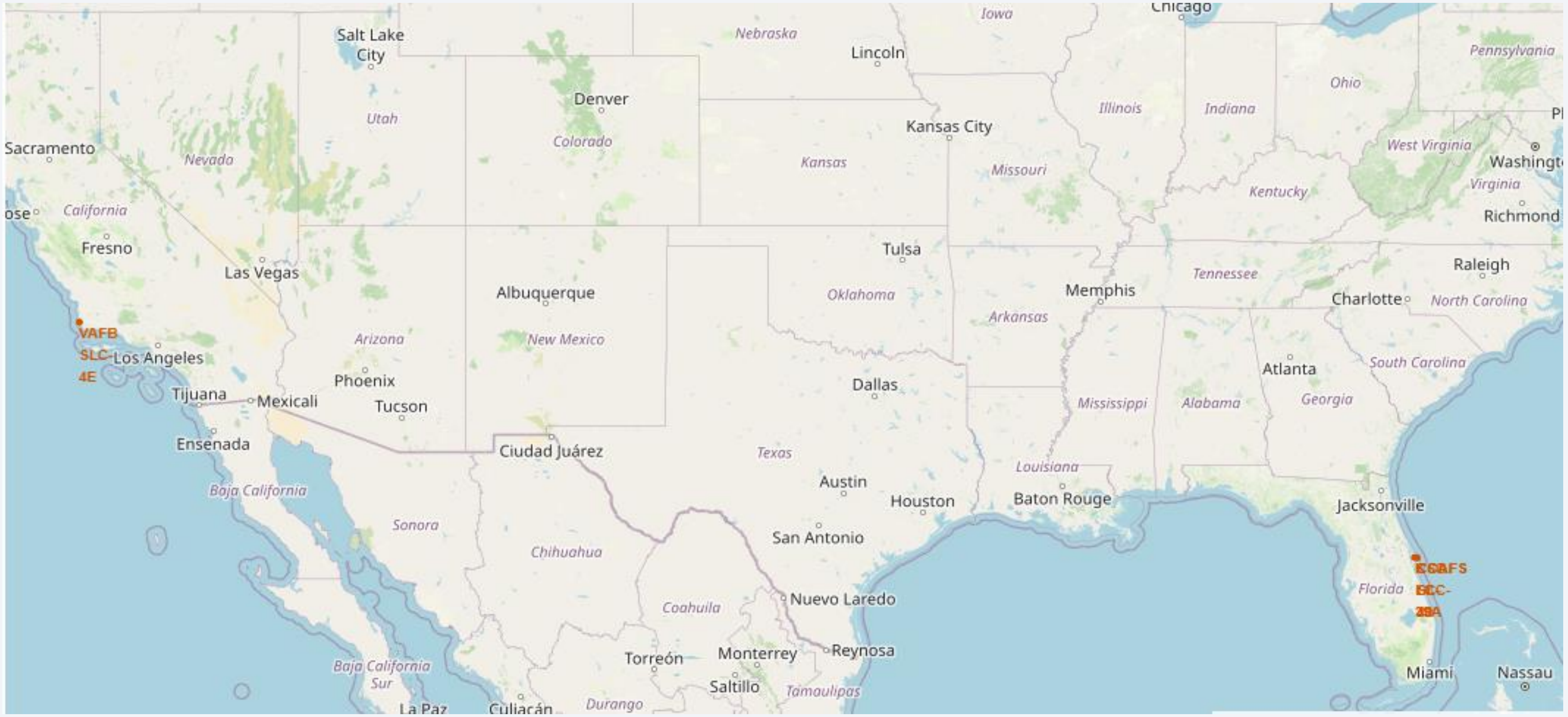
```
%%sql SELECT DISTINCT Landing_Outcome,
                      RANK() OVER (ORDER BY Landing_Outcome DESC) as Frequency
      FROM SPACEXTABLE
      WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
      ORDER BY Frequency DESC;
```

Section 3

Launch Sites Proximities Analysis

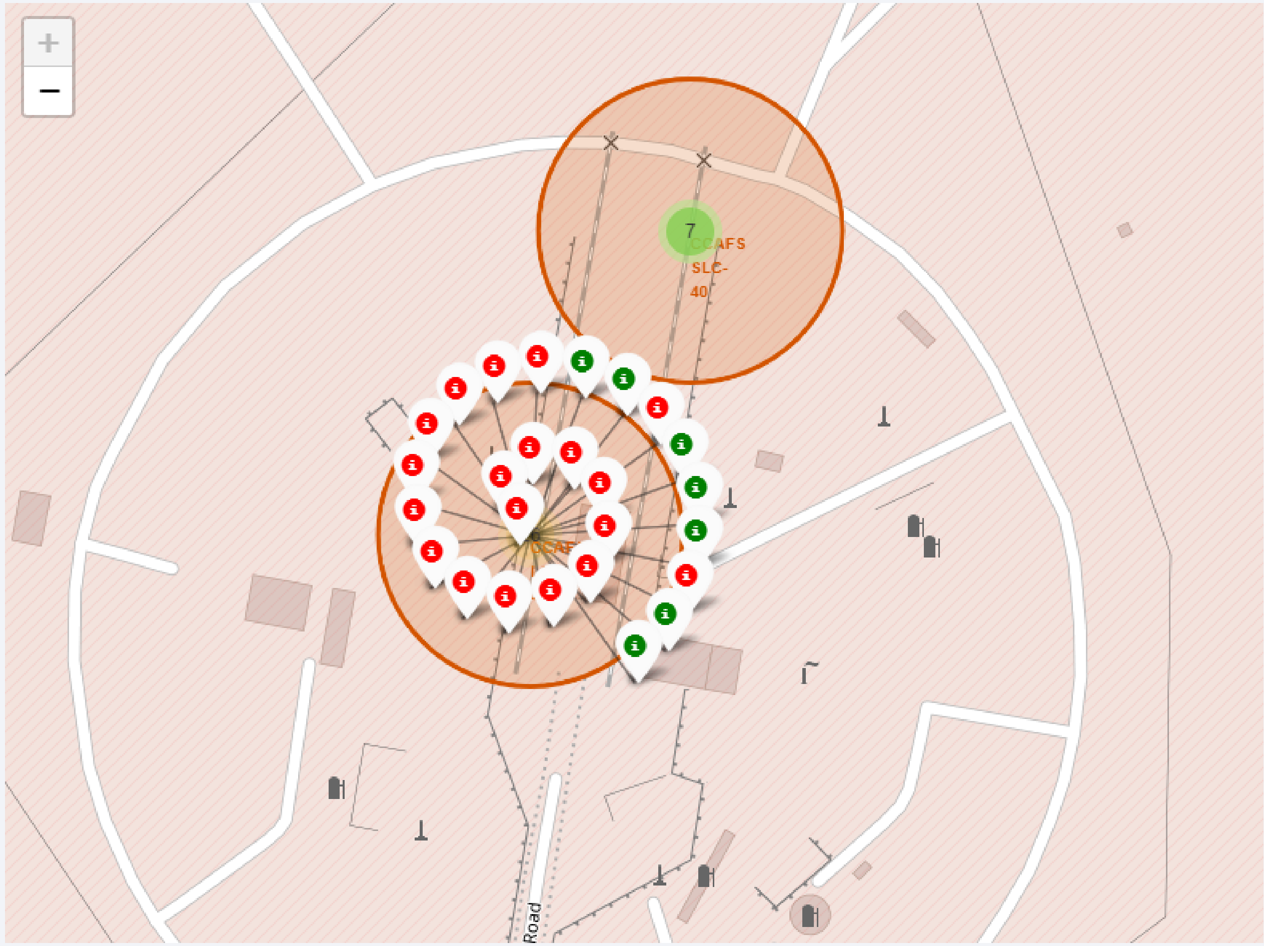
Map of SpaceX Launch Sites

In the SpaceX data, there are four different launch sites identified by the orange point and text on the map below. The sites were located in the United States of America, with three of them being in relative close distance to each other in Florida, while the other one was located in California.



SpaceX Landing Results for CCAFS LS-40 Launches

For the 26 launches from the CCAFS LS-40, the successful landings and missions are dark green markers, and the failed landings are red.



CCAFS SLC-40 Launch Site Proximity

The CCAFS SLC-40 Launch site is

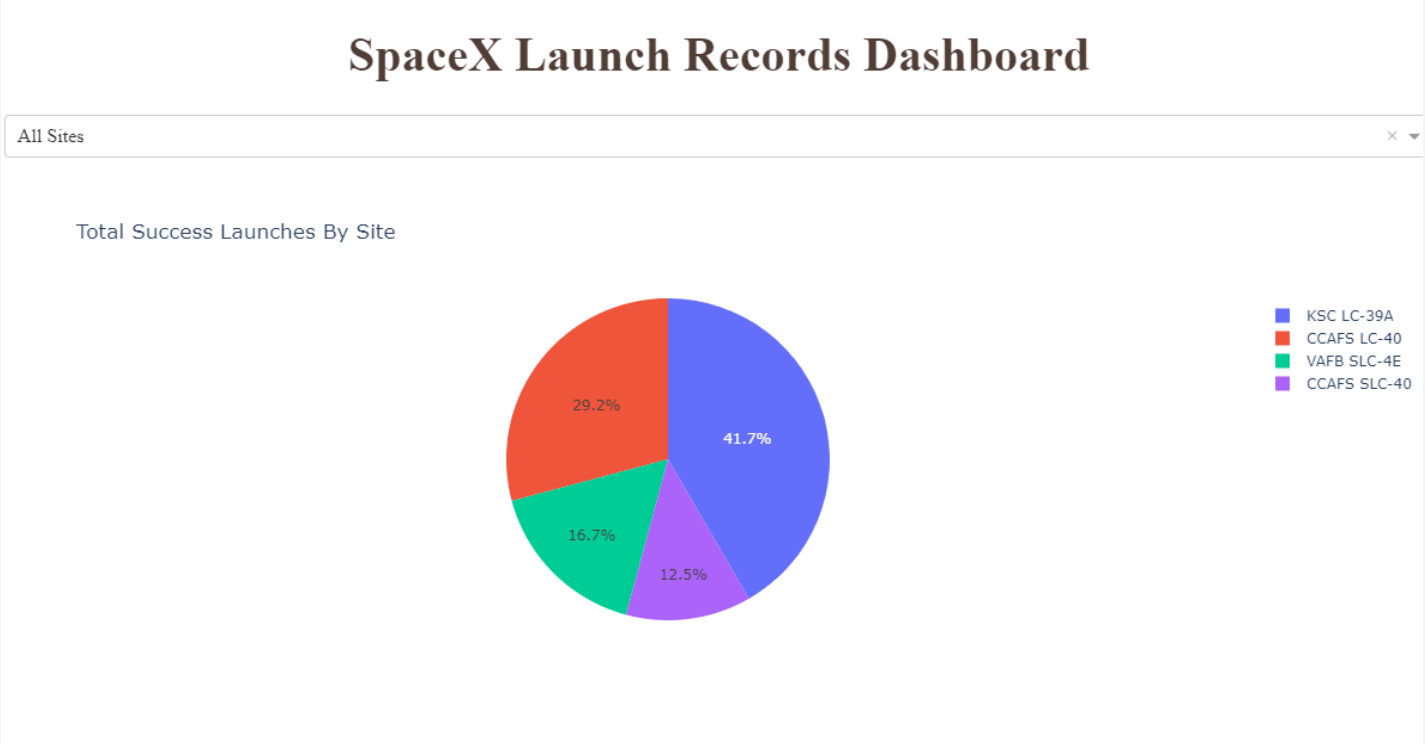
- Relatively close to a coastline as indicated by the shorter, horizontal blue line segment,
- In close proximity (medium, slanted blue line segment) to a railway
- Considerably farther away (longer blue line segment) from the closest city of Melbourne / Cape Canaveral

Section 4

Build a Dashboard with Plotly Dash

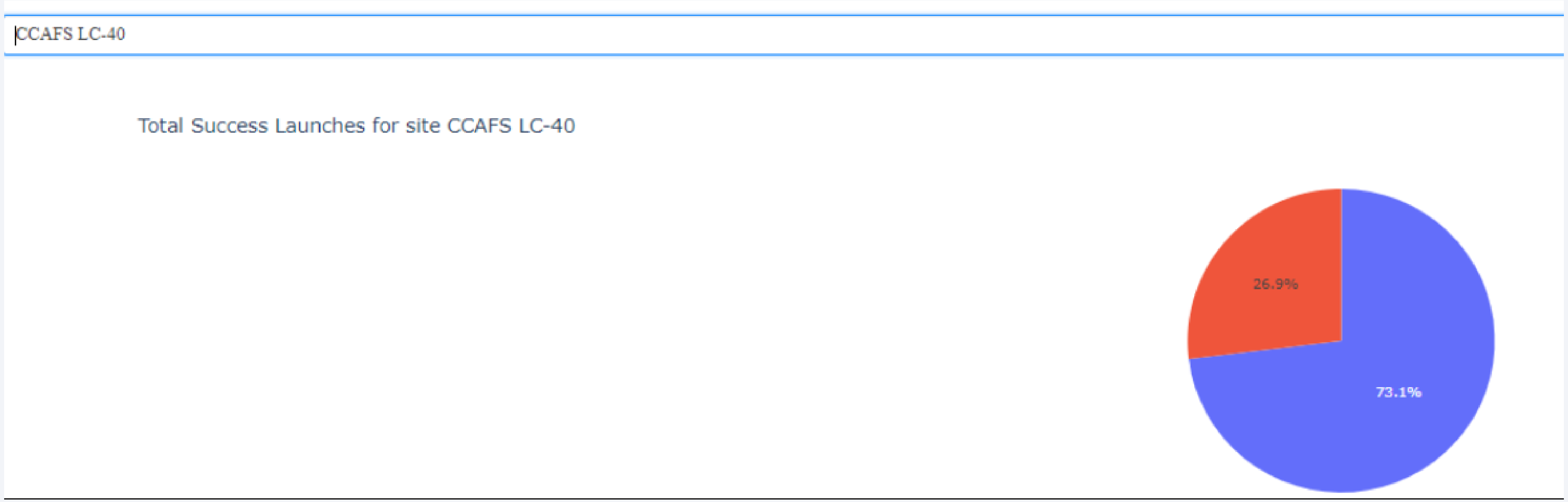
Dashboard: Total Success Launches By Site

Overall, the launch site with the most successful launches and landings is clearly the KSC LC-39A launch site as 41% of the successful Space X landings have been launched from there.



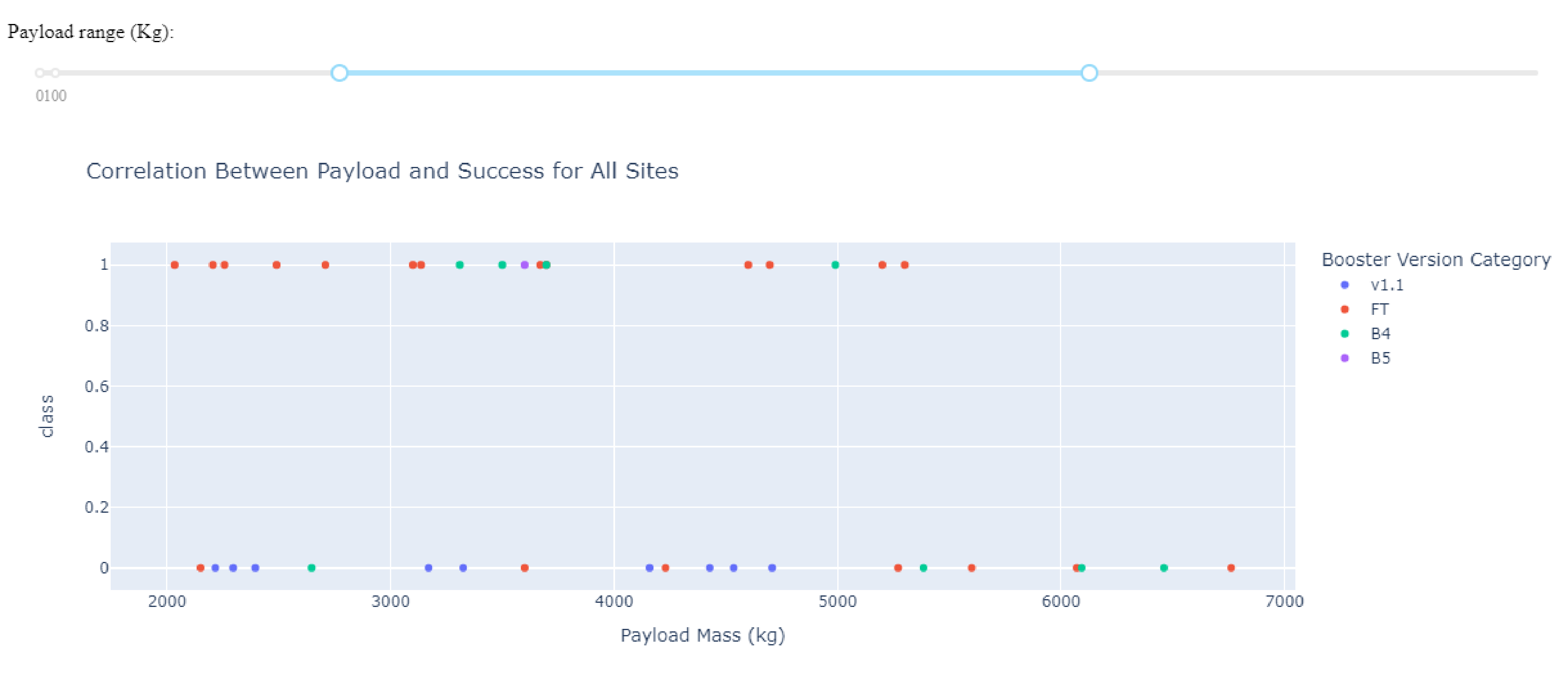
Dashboard: Most Successful Launch Site

Based on the pie chart, 73% of the launches were successful, which is an impressive amount for just one launch site.



Dashboard: Does Payload affect Launch Outcome?

- The graph below indicates that successful landings occurred for smaller payloads for the FT Booster Version. It also shows that the v1.1 Boosters consistently had failed landings for the payloads in this range.

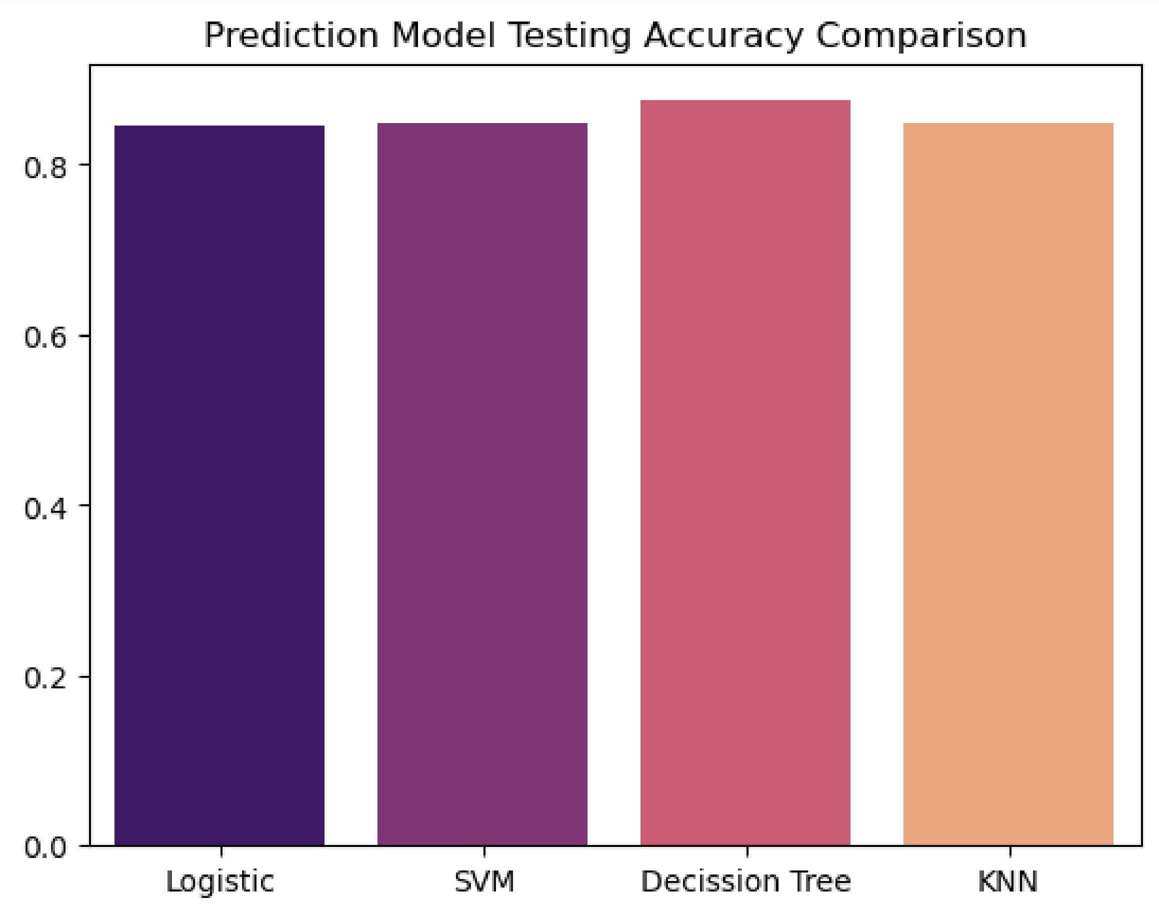


Section 5

Predictive Analysis (Classification)

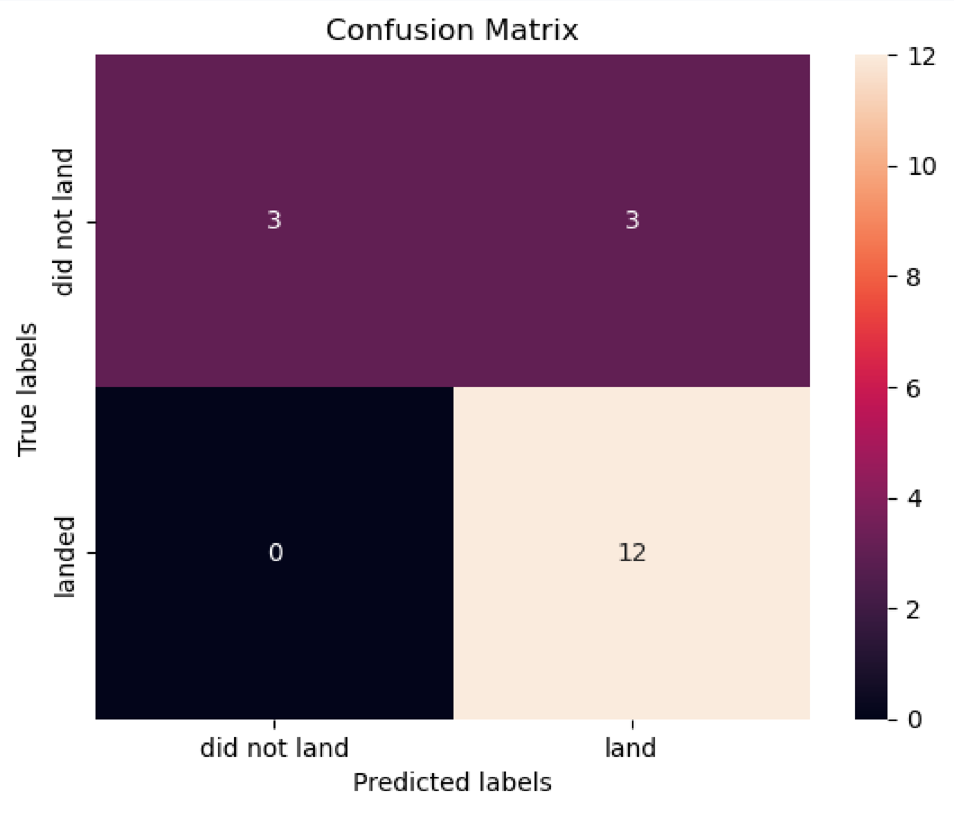
Classification Accuracy

Although testing indicated that all models had similar accuracy scores, the Decision Tree Classification model was the leader in test data accuracy for the tuned hyperparameters.



Confusion Matrix

The Confusion Matrix for the highest performing Decision Tree Classification model was the same as other models' Confusion matrix. They all performed very well with predicting true landing and mission successes, but struggled slight with predicting mission failures (False Positives.



Conclusions

The following features do have an impact on whether a SpaceX Falcon 9 launch and landing is successful.

- Launch Site
- Payload Mass
- Rocket Booster used

The Decision Tree Classification model takes into account the above and can be used to make pretty accurate predictions on whether a future launch will be successful or not.

Appendix

- All Python code snippets, SQL queries, charts, Notebook outputs, and data sets used during this project can be found on Github.

https://github.com/Tigrisriver/IBM-Date-Science/tree/main

Thank you!