IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ayotunde Oyewole
June 10, 2024

Outline

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

Executive Summary

- Summary of methodologies

This project was an analysis of SpaceX Falcon 9 data to predict whether the its first stage would land successfully. The project started with a detailed exploration and a creation of the analytic base table using pandas. Interesting visuals were created as part of the project that helped to identify relevant insights about Falcon 9 launches, conditions favorable to successful landing etc.

- Summary of all results

The culmination of the project was the creation of four classification models – KNN, SVM, Logistic regression, and decision tree. The decision tree classifier had the best result, with an accuracy score of 88.8% on test data.

Introduction

- Project background and context

Space X 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 Space X 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 space X for a rocket launch.

- Problems you want to find answers

The main problem that this project attempts to solve is to help space exploration companies to determine if conditions are right for their spaceship to land successfully. If they can optimize the recovery and successful landing of their spaceship, it will result in significant cost savings.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

The project started with collection of data from the SpaceX API. The data for falcon 9 was filtered out and parsed into a Pandas data frame; and then cleaned.

- Perform data wrangling

The training labels were created with 1 indicating a successful launch and 0 meaning unsuccessful.

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Classification algorithms from sklearn were used to build the models. All four models were evaluated and compared based on their accuracy scores.

Data Collection – SpaceX API

A get request was made to SpaceX's rest API. The content of the request was then parsed into a pandas data frame using the json normalize function. The required data was filtered out by column and dates.

The notebook on data collection is available at https://github.com/asoyewole/IBM-Data-Science-Capstone-Project/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Get request – json normalize – data filtering (date, columns.

Data Wrangling

- Null records were identified on the landing pad column and handled. Subsequently, a column was engineered which classified bad landing outcomes from positive outcomes.

- The notebook for the data wrangling is available at https://github.com/asoyewole/IBM-Data-Science-Capstone-Project/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Pairing up matplotlib and seaborn, some plots were developed to help understand the data. The relationship between flight number and launch site was plotted as a scatterplot. The top orbit types based on average class was identified using a horizontal bar chart. Aside other scatterplots, one other important visual was the trend of success rate by year which showed a continuous growth with few troughs.
- The notebook for the data visualization is available at https://github.com/asoyewole/IBM-Data-Science-Capstone-Project/blob/main/edadataviz.ipynb

EDA with SQL

- Using SQL, a detailed exploration of the data was completed. Items like unique names of launch sites, total number of landing outcomes for each landing outcome, booster version that have the maximum payload mass etc. Screenshots of some sample queries are presented below.
- The notebook for the SQL data exploration is available at https://github.com/asoyewole/IBM-Data-Science-Capstone-Project/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
```

```
%sql SELECT Mission_Outcome, COUNT (Mission_Outcome) FROM SPACEXTABLE GROUP BY Mission_Outcome;
```

```
SELECT Landing_Outcome, COUNT(Landing_Outcome) FROM spacextable GROUP BY Landing_Outcome order by COUNT(Landing_Outcome) DESC
```

Build an Interactive Map with Folium

- A number of interactive maps were designed to explore the locations of the launch sites and how the locations and geographic proximities affect or impact successful landing.
- The notebook for the interactive map designed with folium is available at https://github.com/asoyewole/IBM-Data-Science-Capstone-Project/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

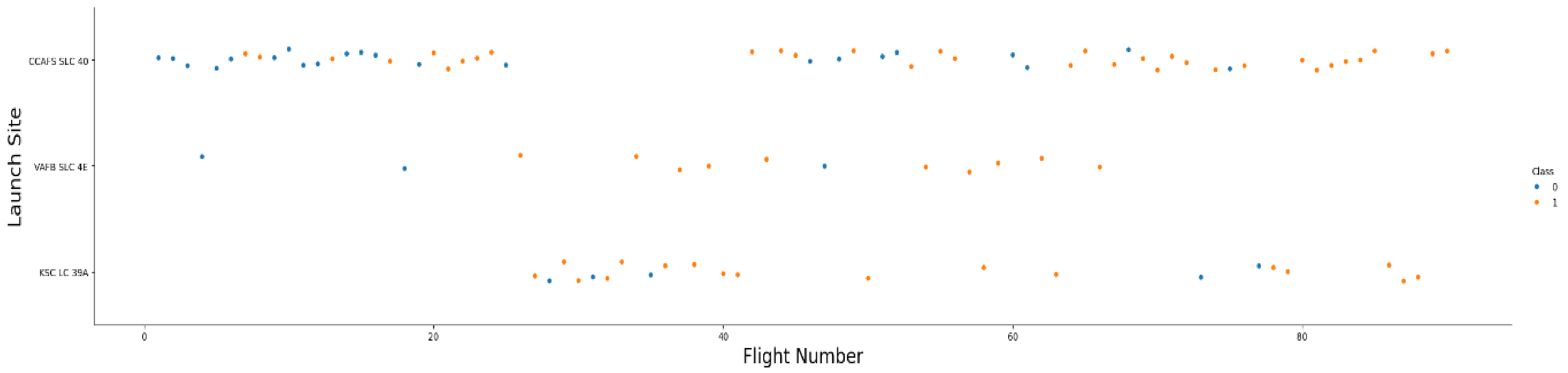
- Two interactive plots were included on the dash app – a pie chart showing the ratios of successful launches for each site, and a scatterplot that shows the interaction between the payload mass and the class.
- The python script for building the dash interactive app is available at https://github.com/asoyewole/IBM-Data-Science-Capstone-Project/blob/main/spacex_dash_app%20(1).py

Predictive Analysis (Classification)

- The data was split into features and target, and then scaled as is required for classification models. Both the features and the target was split into training and testing data. A grid search algorithm was used to tune the hyperparameters for each model. Logistic regression, KNN, SVM, and Decision Tree models were each fitted to the training data, and then tested using the test data.
- The notebook for the classification analysis is available at https://github.com/asoyewole/IBM-Data-Science-Capstone-Project/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

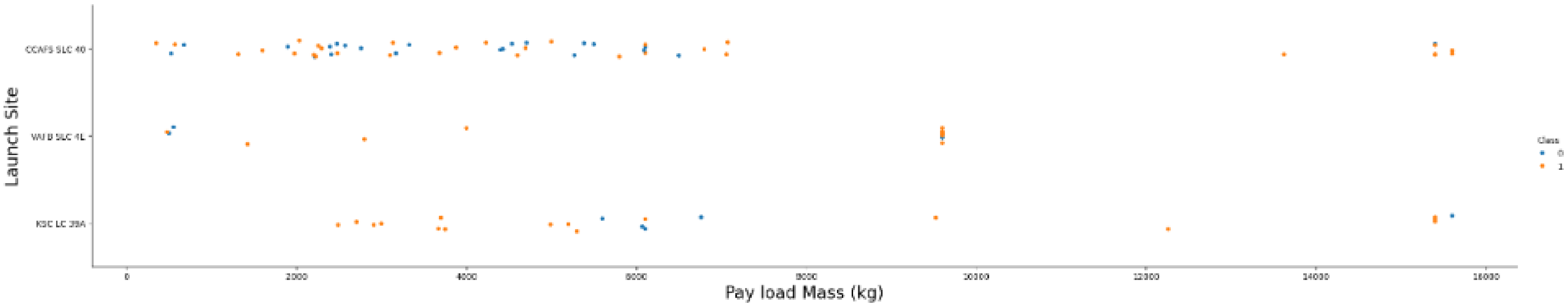
Section 2

Insights drawn from EDA



Flight Number vs. Launch Site

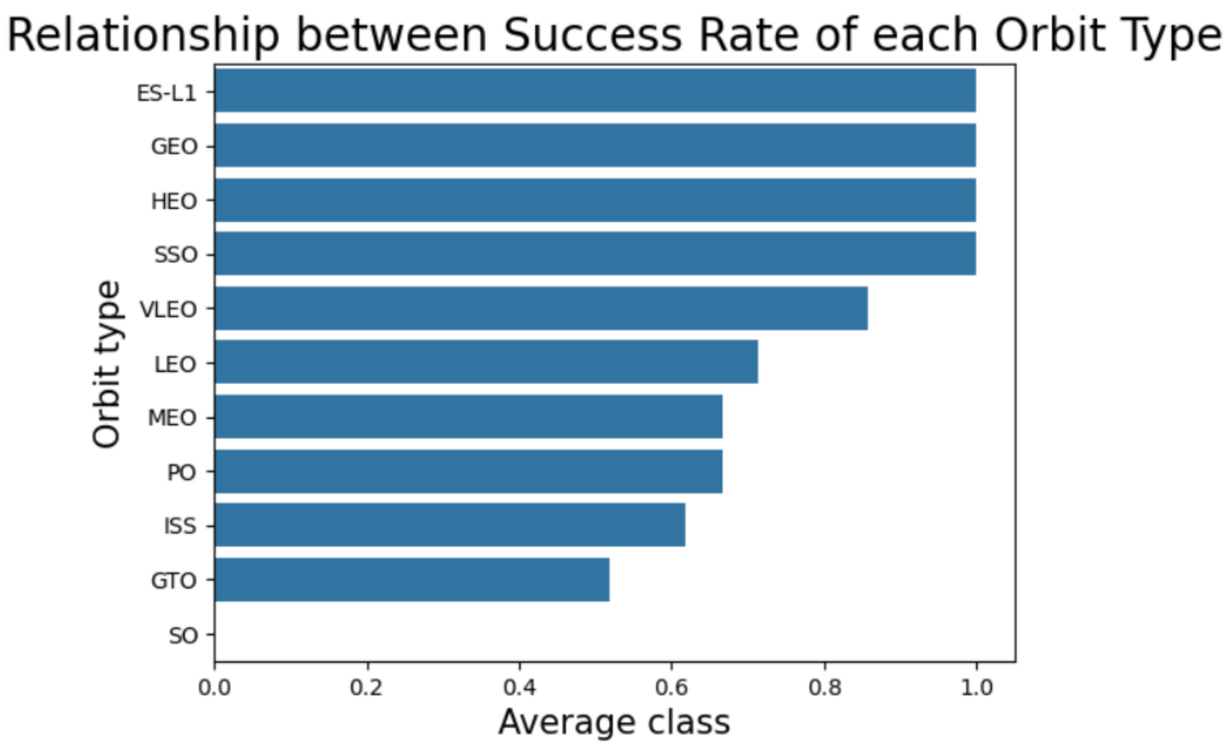
- There were more of class 0 launches at the CCAFS SLC 40 while class 1 launches are spread out between all three launch sites. VAFB SLC 4E had fewer flights compared to the other sites. In a similar vein, CCAFS SLC 40 had the most number of flights.



Payload vs. Launch Site

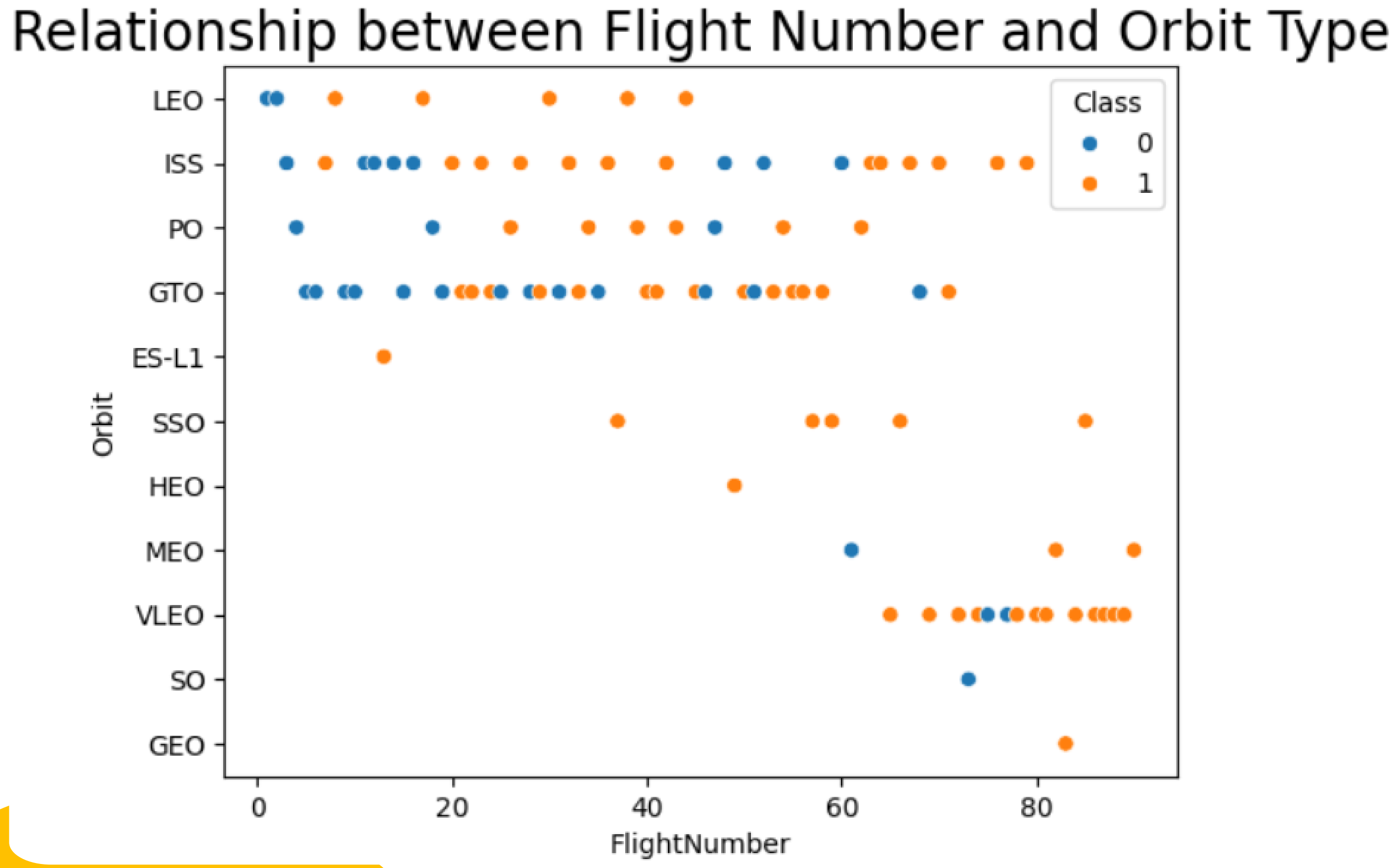
- It is notable here that there are no rockets launched for heavy pay load greater than 10 thousand kg for VAFB-SLC launch site.

Success Rate vs. Orbit Type



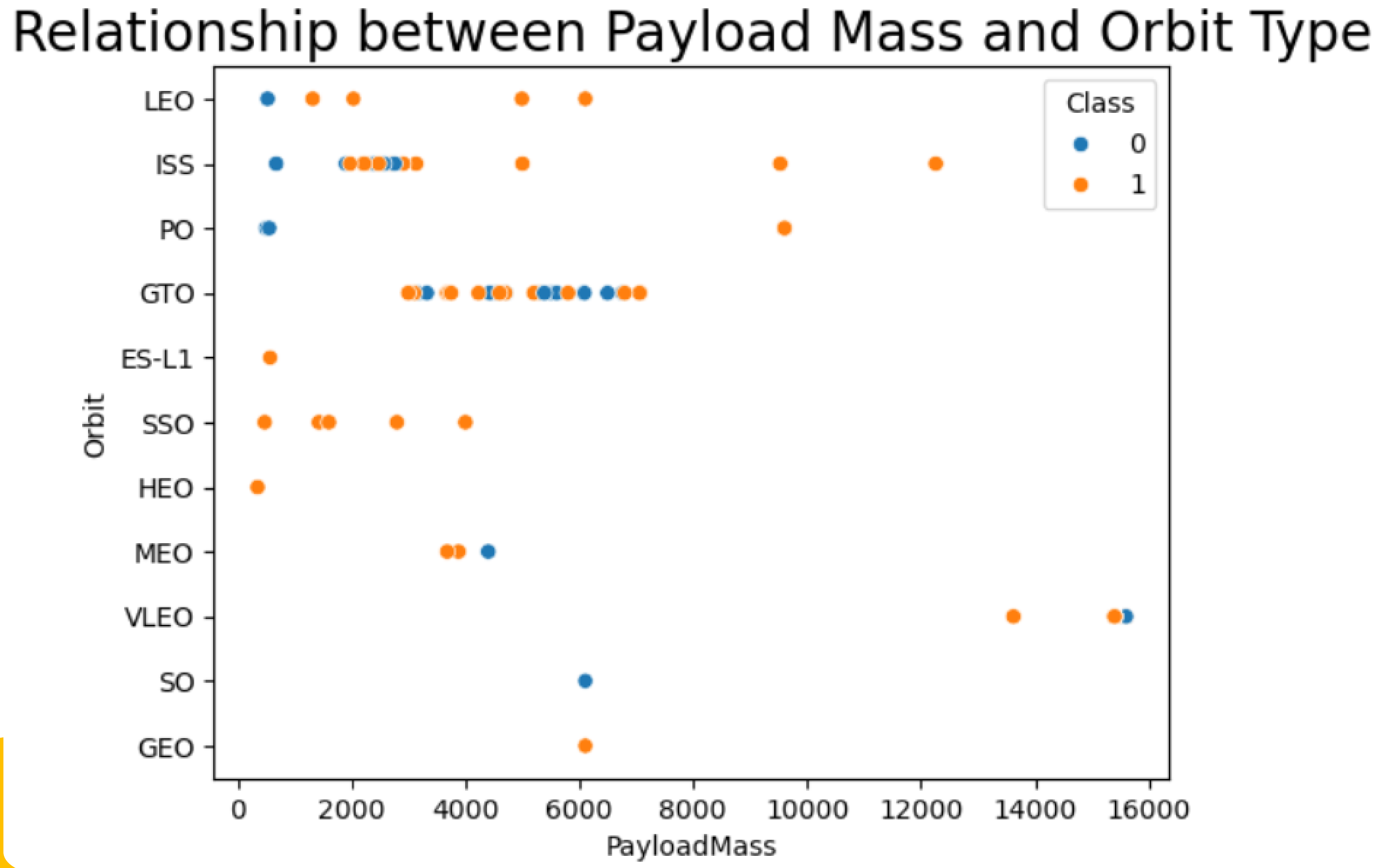
- ES-L1, GEO, HEO and SSO all had 100% success rates. On the other hand, SO, recorded no success at all while MEO, PO, ISS and GTO had success rates below 0.7. LEO and VLEO had successes between 0.7 and 0.9.

Flight Number vs. Orbit Type



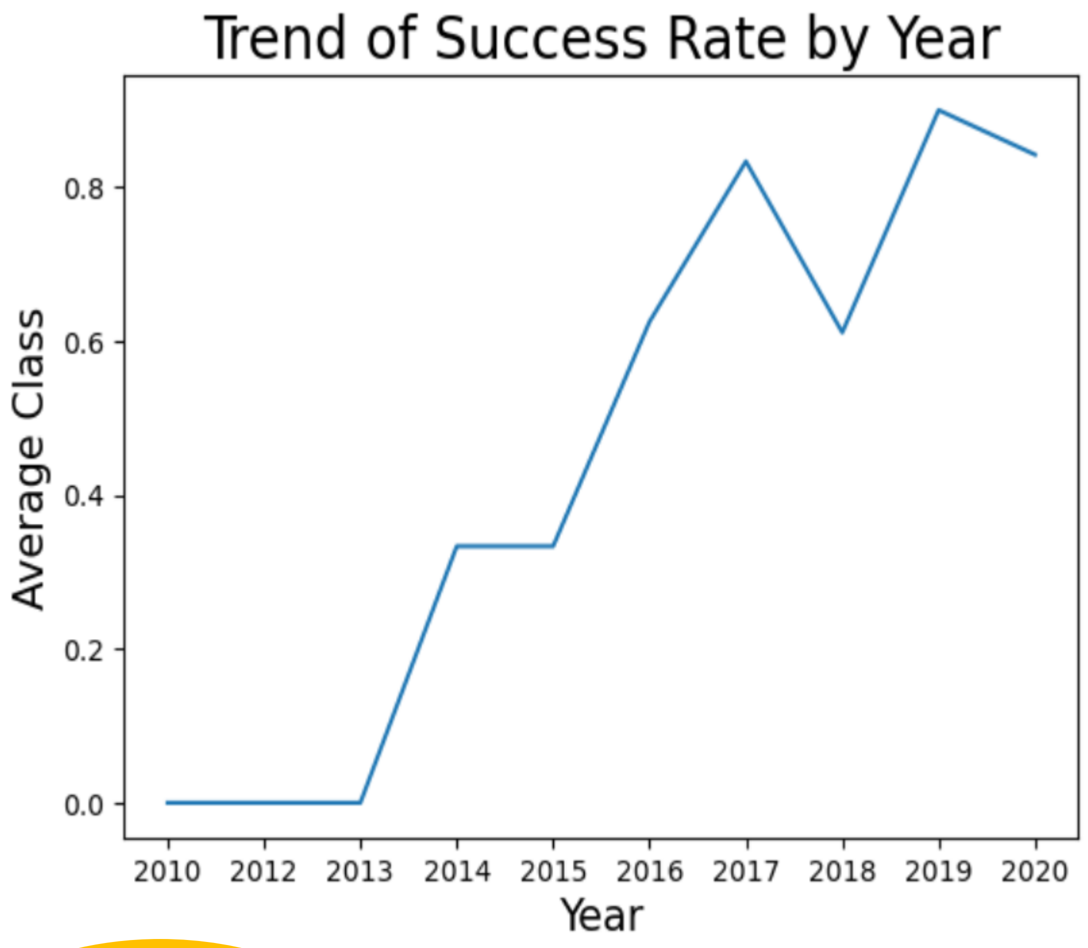
- the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend



- the sucess rate since 2013 kept increasing till 2020

All Launch Site Names

- The table shows the names of all the launch sites.

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

Launch Site Names Begin with 'CCA'

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

SUM(PAYLOAD_MASS__KG_)
45596

Total Payload Mass

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

Average Payload Mass by F9 v1.1

MIN(DATE)
2015-12-22

First Successful Ground Landing Date

Successful Drone Ship Landing with Payload between 4000 and 6000

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

Total Number of Successful and Failure Mission Outcomes

Out[16]:

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

Boosters Carried Maximum 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

2015 Launch Records

substr(Date, 6,2)	Landing_Outcome
06	Failure (parachute)
12	Failure (parachute)
01	Failure (drone ship)
04	Failure (drone ship)
01	Failure (drone ship)
03	Failure (drone ship)
06	Failure (drone ship)
12	Failure
02	Failure
03	Failure

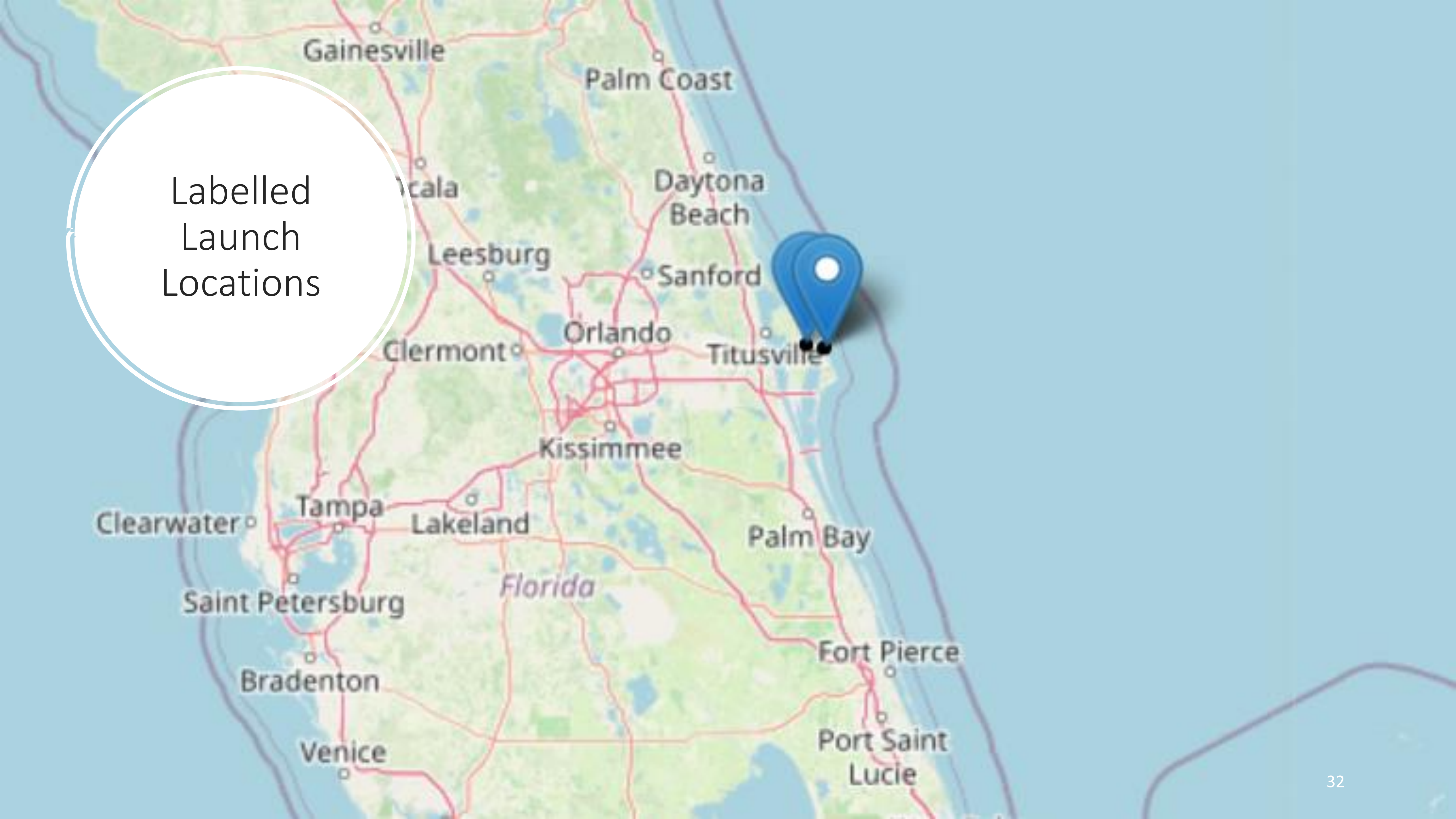
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

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

Section 3

Launch Sites Proximities Analysis

Labelled Launch Locations



Launch Map with Proximities

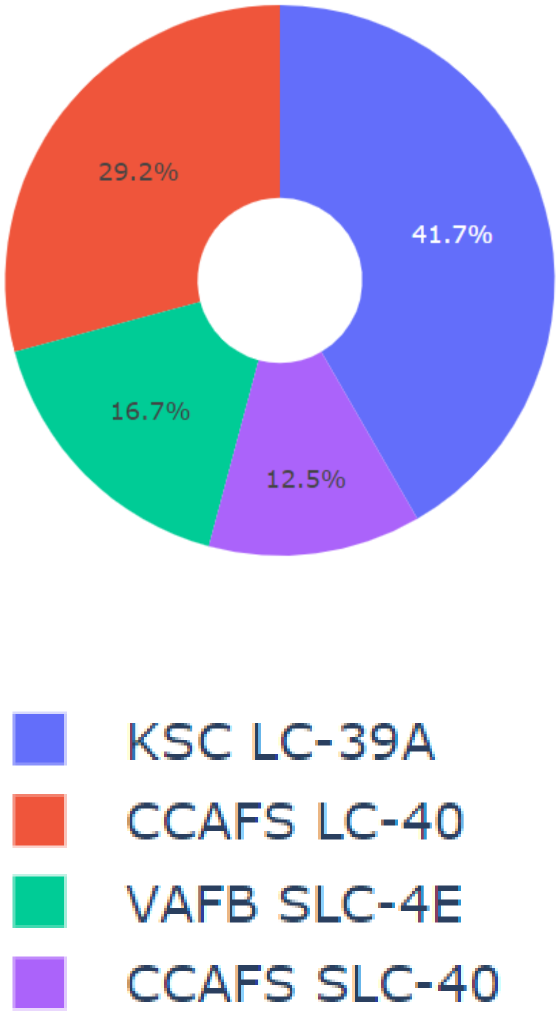
Successful Launches for Sites

Section 4

Build a Dashboard with Plotly Dash

Total Successful Launches of all Sites

- KSC LC-39A had more than a third of the successful launches. Only 12.5% were successful in CCAFS SLC-40.



Relationship between Payload and Launch Outcome

- The plot shows the relationship between payload and launch outcomes.

Section 5

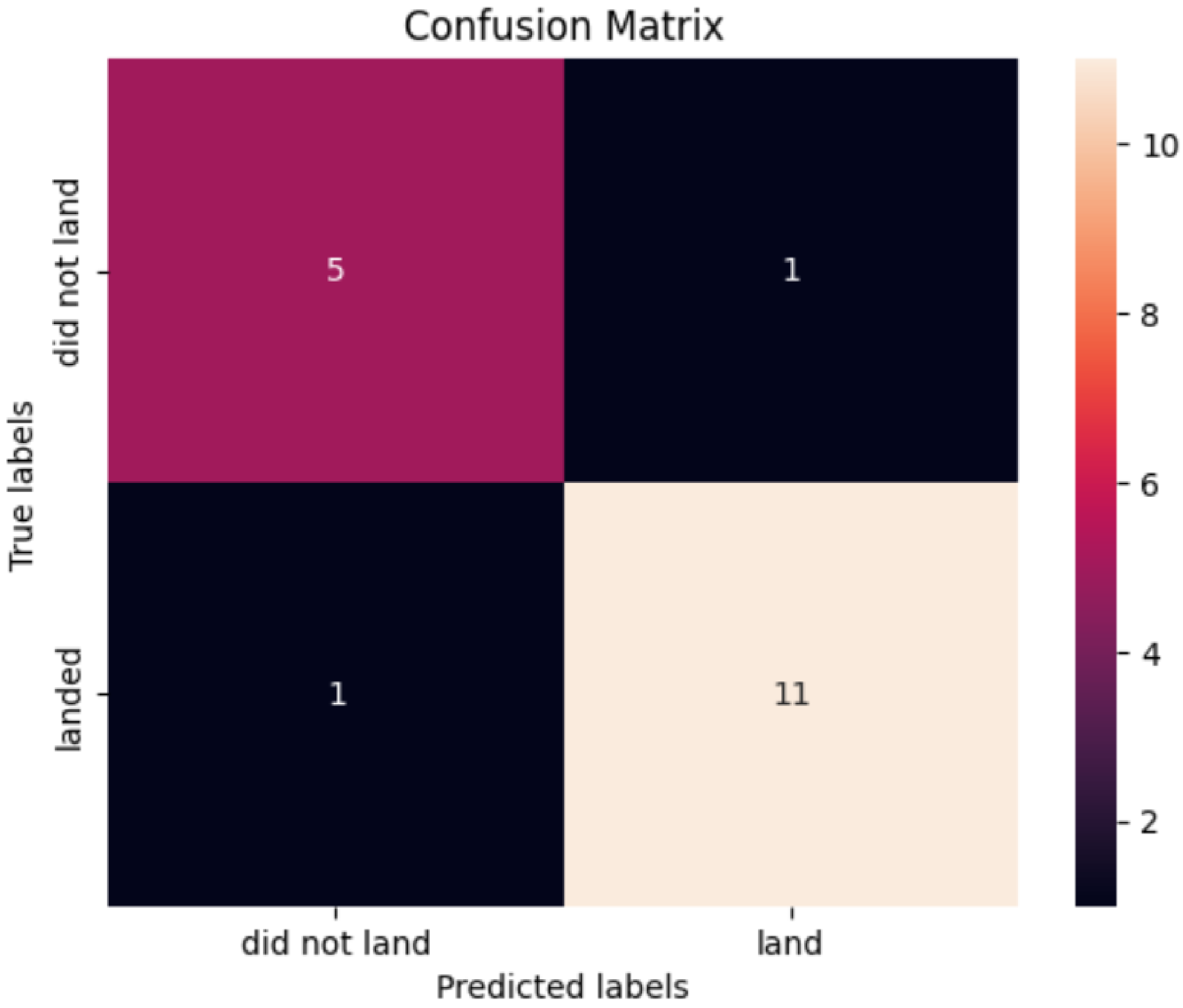
Predictive Analysis (Classification)

Classification Accuracy

Models	Accuracy
KNN	83.33%
Decision Tree	**88.89%**
SVM	83.33%
Logistic Regression	83.33%

Confusion Matrix

The decision tree model was the best performing. The confusion matrix shows that it correctly classified 5 unsuccessful landings, and 11 successful landings. Only 1 each were misclassified and false positives and false negatives respectively.



Conclusions

- This project involved the demonstration of a gamut of data science tools including SQL, python, visualization with folium, and dash etc.
- The final phase of the project was the development of four machine learning models to classify the outcomes of space ship landing attempts.
- The decision tree model was the best performing model for this project.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!