IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Shamika Konara
08/07/2024

Outline

Executive Summary

- In this capstone project, we aim to predict whether the SpaceX Falcon 9 first stage will land successfully using various machine learning classification algorithms.
- The primary steps in this project include:
  - Data collection, wrangling, and formatting
  - Exploratory data analysis
  - Interactive data visualization
  - Machine learning prediction
- Our analysis indicates that certain features of the rocket launches correlate with the launch outcomes, i.e., success or failure. Based on our findings, we conclude that the decision tree algorithm may be the most effective for predicting the successful landing of the Falcon 9 first stage.

Introduction

- In this capstone project, we aim to predict whether the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches on its website at a cost of 62 million dollars, significantly lower than the 165 million dollars charged by other providers. This cost efficiency is largely due to SpaceX's ability to reuse the first stage of the rocket. By accurately predicting the landing success of the first stage, we can better estimate the cost of a launch.
- This information could be valuable for alternate companies looking to bid against SpaceX for rocket launch contracts.It is important to note that many unsuccessful landings are intentional, as SpaceX occasionally opts for a controlled landing in the ocean.
- The primary question we seek to answer is: given a set of features about a Falcon 9 rocket launch, including payload mass, orbit type, launch site, and others, can we predict whether the first stage of the rocket will land successfully?

Section 1

Methodology

Methodology

- The overall methodology includes:
    1. Data collection, wrangling, and formatting, using:
        - SpaceX API
        - Web scraping
    2. Exploratory data analysis (EDA), using:
        - Pandas and NumPy
        - SQL
    3. Data visualization, using:
        - Matplotlib and Seaborn
        - Folium
        - Dash
    4. Machine learning prediction, using
        - Logistic regression
        - Support vector machine (SVM)
        - Decision tree
        - K-nearest neighbors (KNN)

Data Collection – SpaceX API

- The API utilized for this project is https://api.spacexdata.com/v4/rockets/, which provides detailed data on various rocket launches conducted by SpaceX. For the purposes of this analysis, the data has been filtered to include only Falcon 9 launches.
- To handle missing values, each missing entry is replaced with the mean of the respective column. After processing, the dataset comprises 90 rows (instances) and 17 columns (features). The image below displays the first few rows of the filtered and cleaned data:

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.561857
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.561857
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.632093
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577366	28.561857

Data Collection - Scraping

- Web scraping
  - The data is scraped from https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
  - The website contains only the data about Falcon 9 launches.
  - We end up with 121 rows or instances and 11 columns or features. The picture below shows the first few rows of the data:

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10

Data Wrangling

- The data is later processed so that there are no missing entries and categorical features are encoded using one-hot encoding.
- An extra column called ‘Class’ is also added to the data frame. The column ‘Class’ contains 0 if a given launch is failed and 1 if it is successful.
- In the end, we end up with 90 rows or instances and 83 columns or features.

EDA with Data Visualization

Pandas and NumPy

Functions from the Pandas and NumPy libraries are utilized to derive basic information about the collected data, including:

- The number of launches at each launch site
- The frequency of each orbit type
- The number and frequency of each mission outcome

SQL

SQL queries are used to extract and analyze specific details from the data, such as:

- The names of the unique launch sites involved in the space missions
- The total payload mass carried by boosters launched by NASA (CRS)
- The average payload mass carried by booster version F9 v1.1

Matplotlib and Seaborn

Functions from the Matplotlib and Seaborn libraries are used to visualize the data through scatterplots, bar charts, and line charts. The plots and charts are used to understand more about the relationships between several features, such as:

- The relationship between flight number and launch site
- The relationship between payload mass and launch site
- The relationship between success rate and orbit type

EDA with Data Visualization

Folium

Functions from the Folium libraries are used to visualize the data through interactive maps. The Folium library is used to:

- Mark all launch sites on a map
- Mark the succeeded launches and failed launches for each site on the map
- Mark the distances between a launch site to its proximities such as the nearest city, railway, or highway

Dash

Functions from Dash are used to generate an interactive site where we can toggle the input using a dropdown menu and a range slider.

Using a pie chart and a scatterplot, the interactive site shows:

- The total success launches from each launch site
- The correlation between payload mass and mission outcome (success or failure) for each launch site

EDA with Data Visualization

Folium

Functions from the Folium libraries are used to visualize the data through interactive maps. The Folium library is used to:

- Mark all launch sites on a map
- Mark the succeeded launches and failed launches for each site on the map
- Mark the distances between a launch site to its proximities such as the nearest city, railway, or highway

Dash

Functions from Dash are used to generate an interactive site where we can toggle the input using a dropdown menu and a range slider.

Using a pie chart and a scatterplot, the interactive site shows:

- The total success launches from each launch site
- The correlation between payload mass and mission outcome (success or failure) for each launch site

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
- Explain why you added those objects
- Add the GitHub URL of your completed interactive map with Folium map, as an external reference and peer-review purpose

Results

- The results are split into 5 sections:
    - SQL (EDA with SQL)
    - Matplotlib and Seaborn (EDA with Visualization)
    - Folium
    - Dash
    - Predictive Analysis
- In all of the graphs that follow, class 0 represents a failed launch outcome while class 1 represents a successful launch outcome.

Results

- The names of the unique launch sites in the space mission

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

- 5 records where launch sites 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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

Results

- The total payload mass carried by boosters launched by NASA (CRS)

Total payload mass by NASA (CRS)
45596

- The average payload mass carried by booster version F9 v1.1

Average payload mass by Booster Version F9 v1.1
2928

- The date when the first successful landing outcome in ground pad was achieved

Date of first successful landing outcome in ground pad
2015-12-22

Results

- The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

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

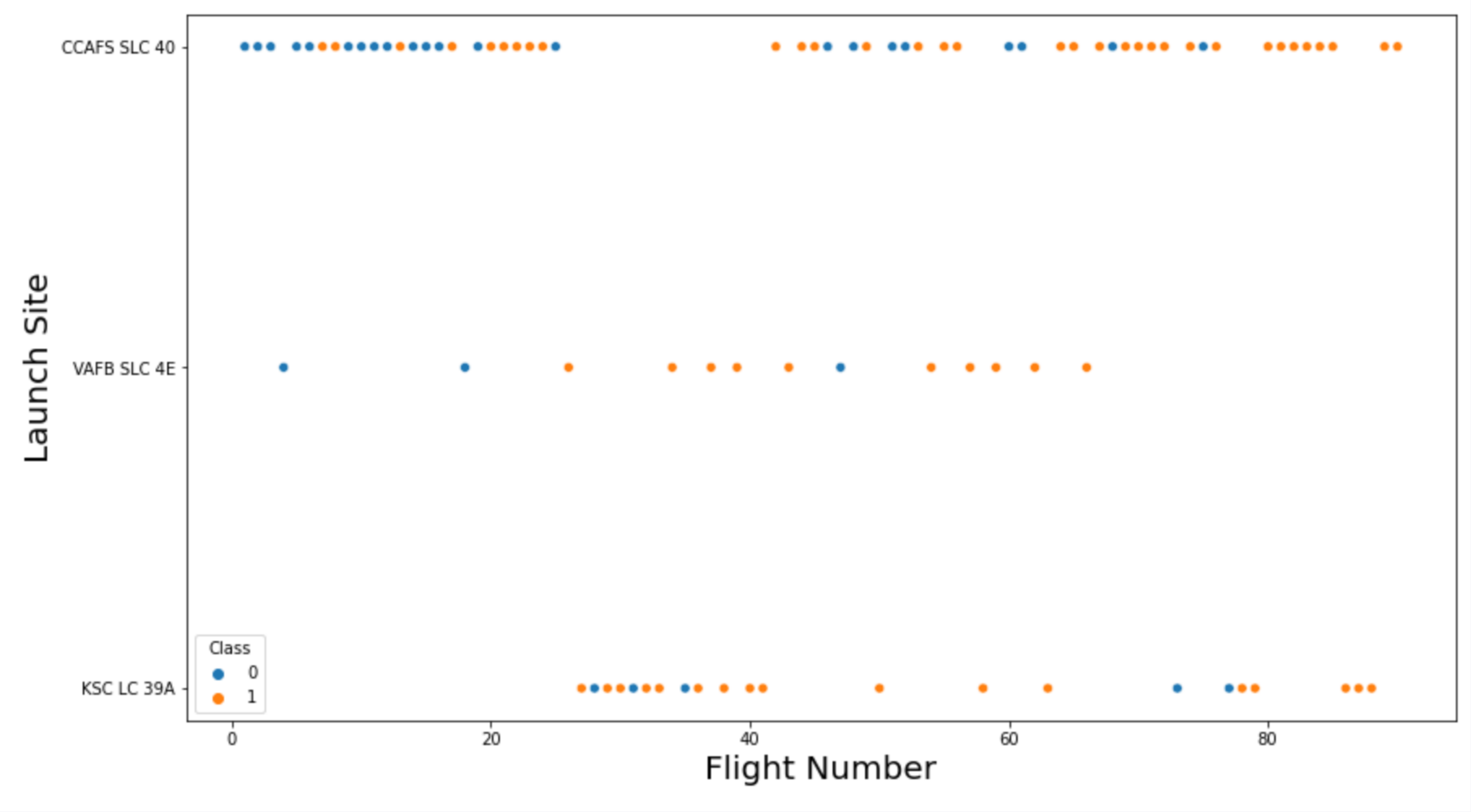
- The total number of successful and failure mission outcomes

number_of_success_outcomes	number_of_failure_outcomes
100	1

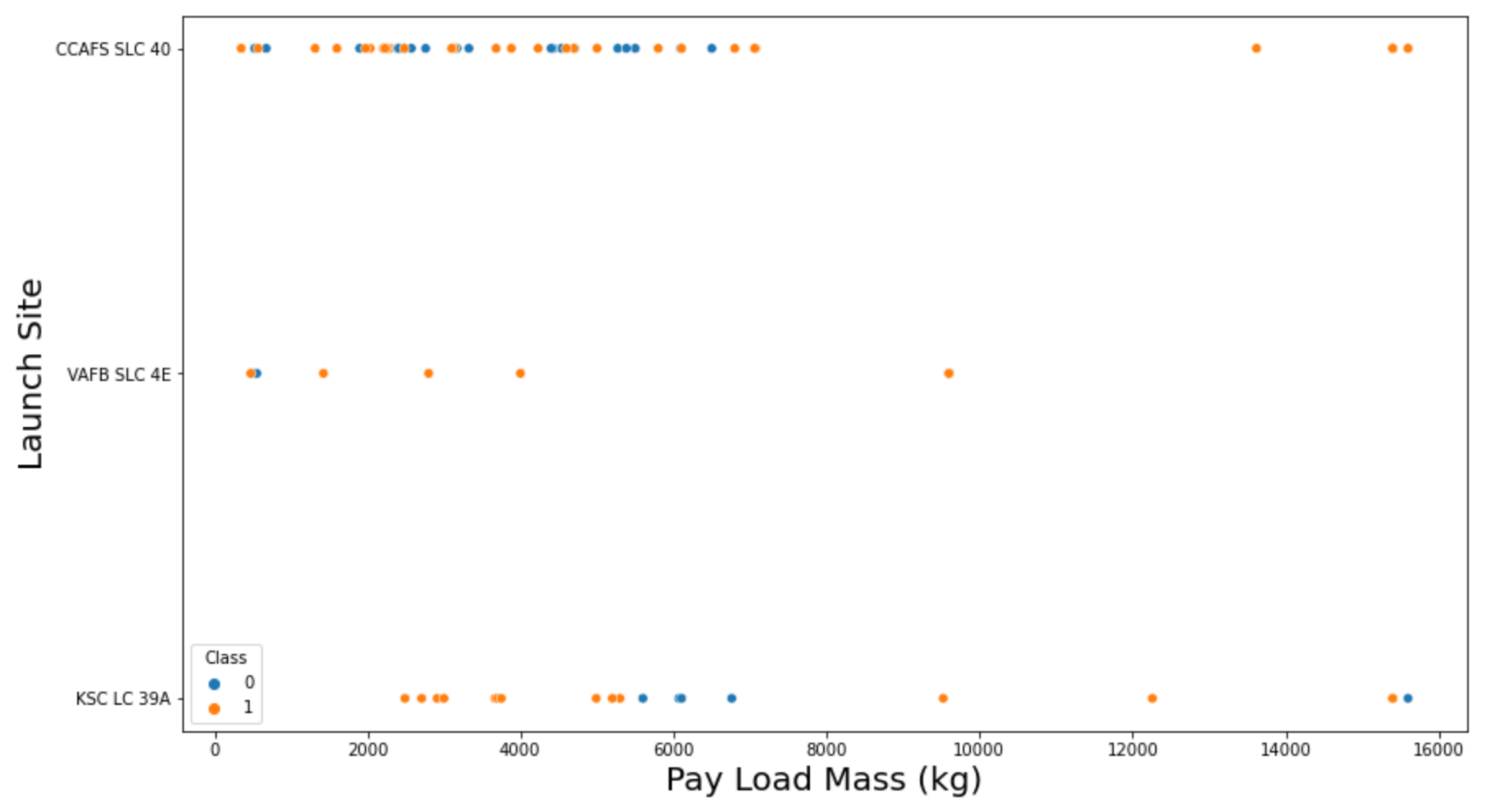
Section 2

Insights drawn from EDA

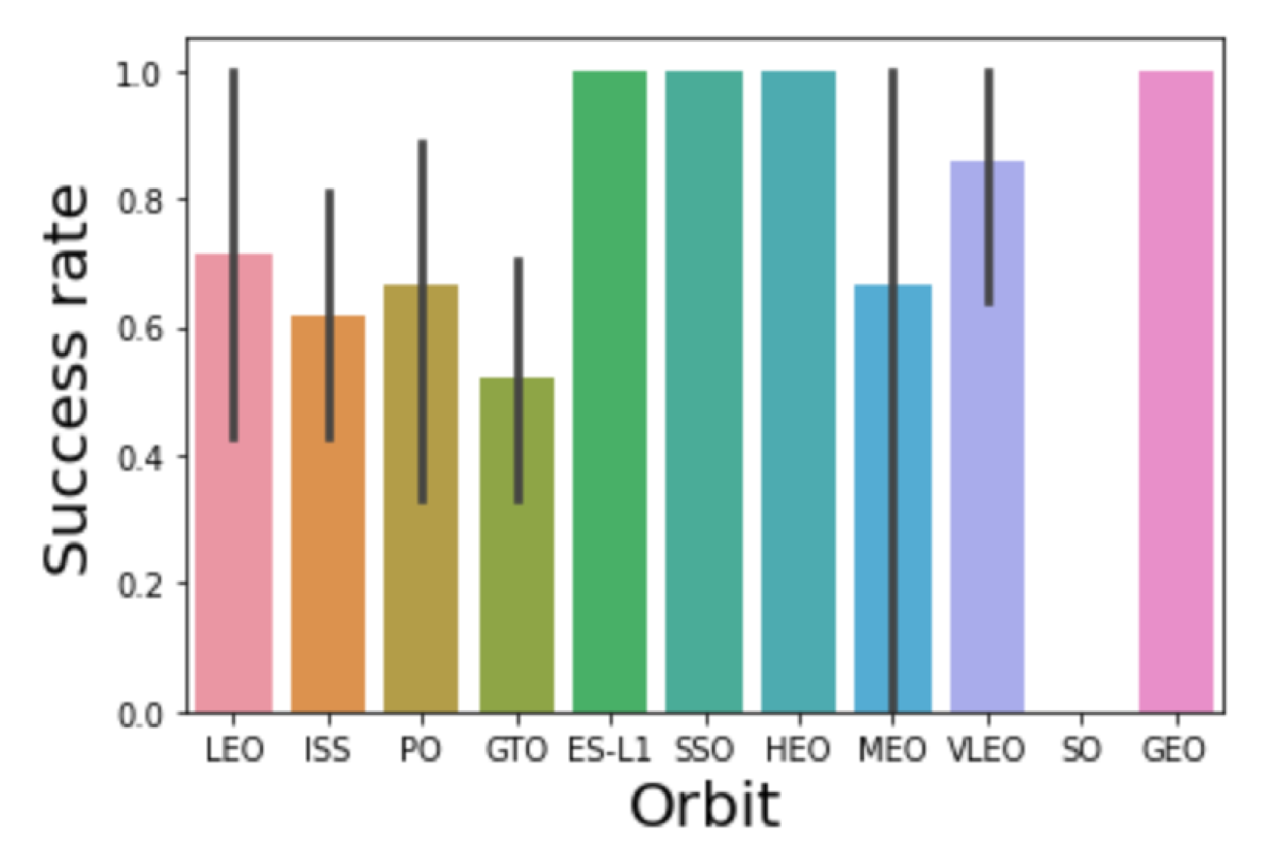
Flight Number vs. Launch Site



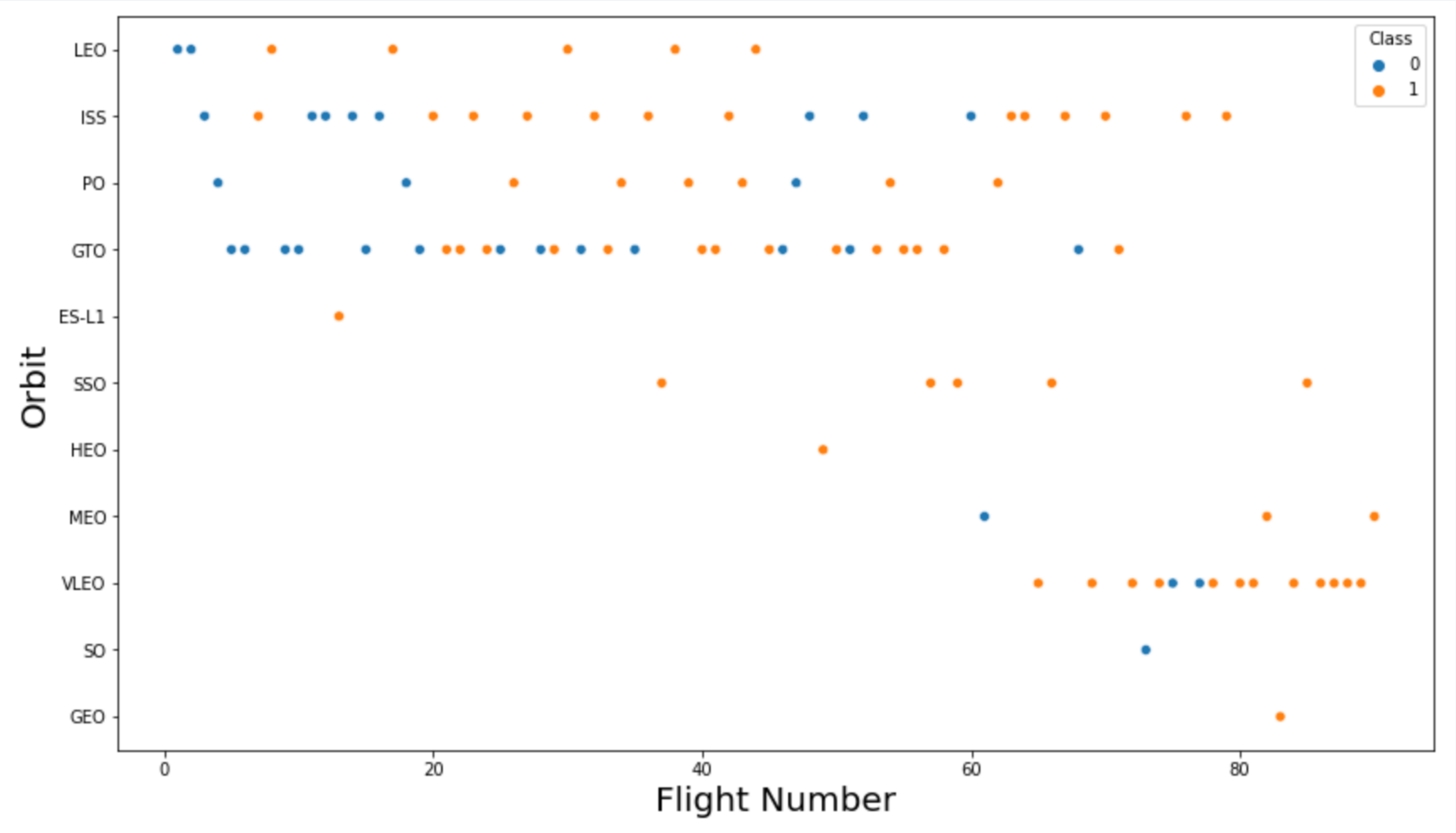
Payload vs. Launch Site



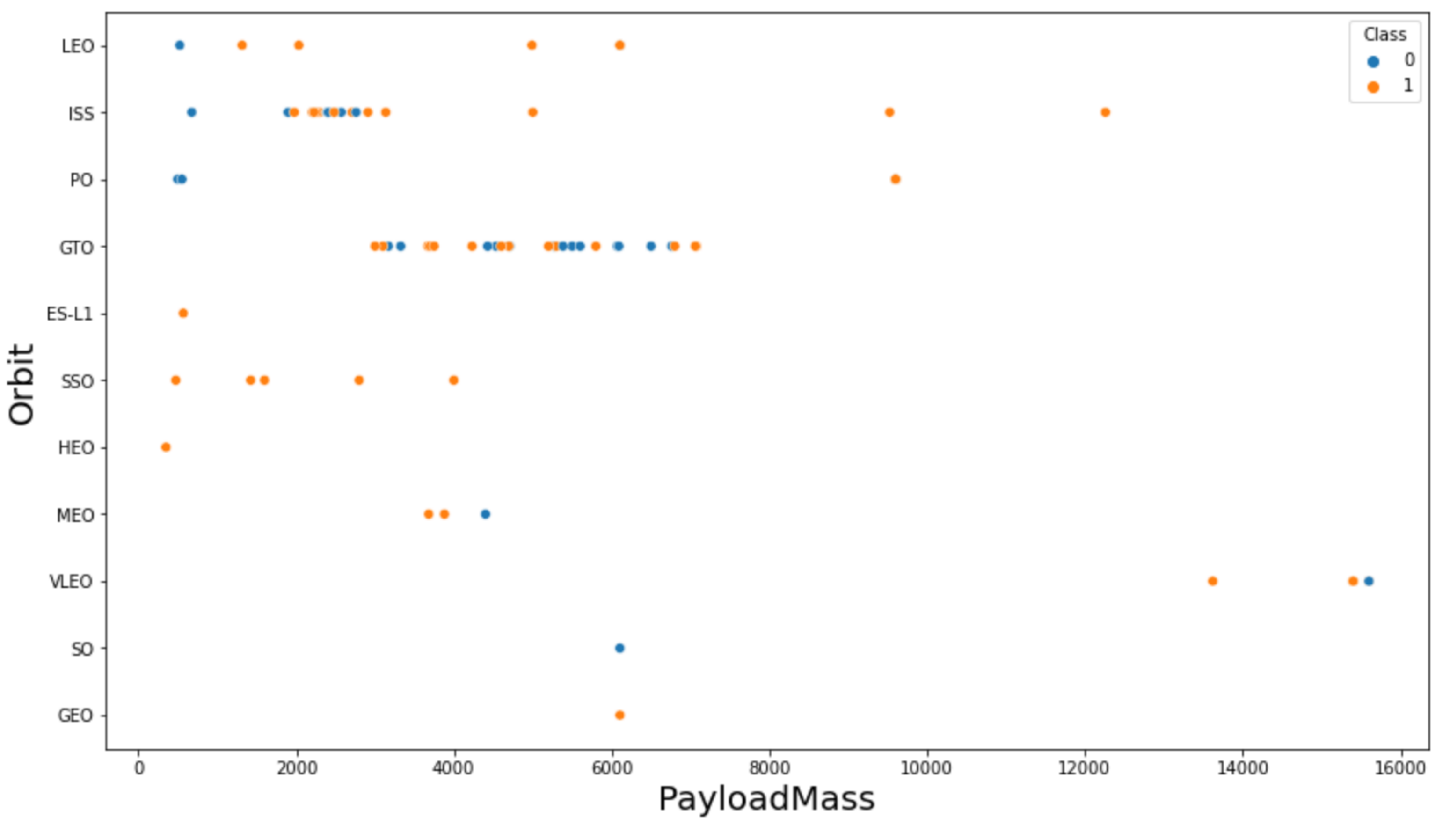
Success Rate vs. Orbit Type



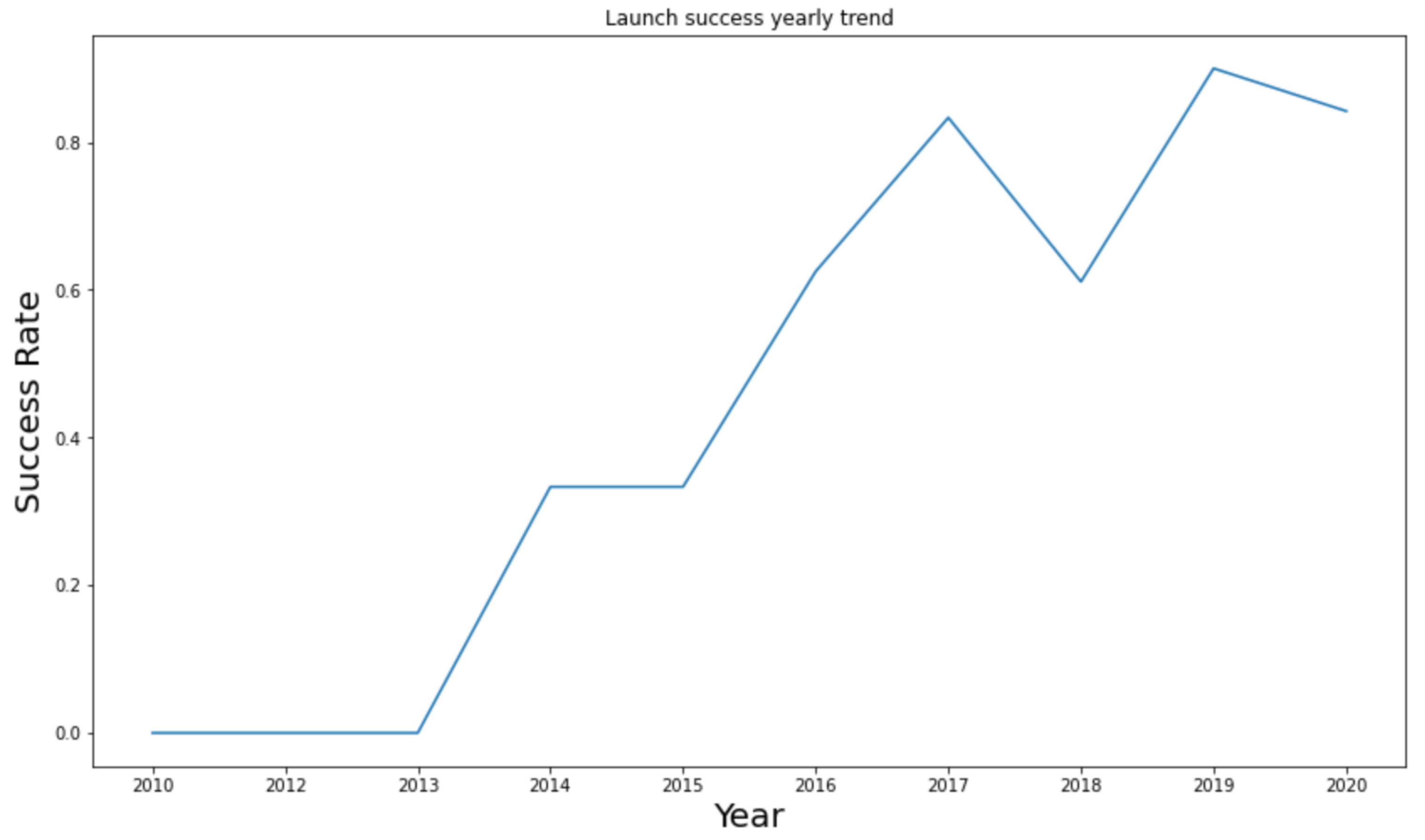
Flight Number vs. Orbit Type



Payload vs. Orbit Type



Launch Success Yearly Trend



All Launch Site Names

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

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

Total payload mass by NASA (CRS)
45596

Average Payload Mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1

Average payload mass by Booster Version F9 v1.1
2928

First Successful Ground Landing Date

- The date when the first successful landing outcome in ground pad was achieved

Date of first successful landing outcome in ground pad
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 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

- The total number of successful and failure mission outcomes

number_of_success_outcomes	number_of_failure_outcomes
100	1

Boosters Carried Maximum Payload

- The names of the booster versions which have carried the maximum payload mass

booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 Launch Records

- The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

DATE	booster_version	launch_site
2015-01-10	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	F9 v1.1 B1015	CCAFS LC-40

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

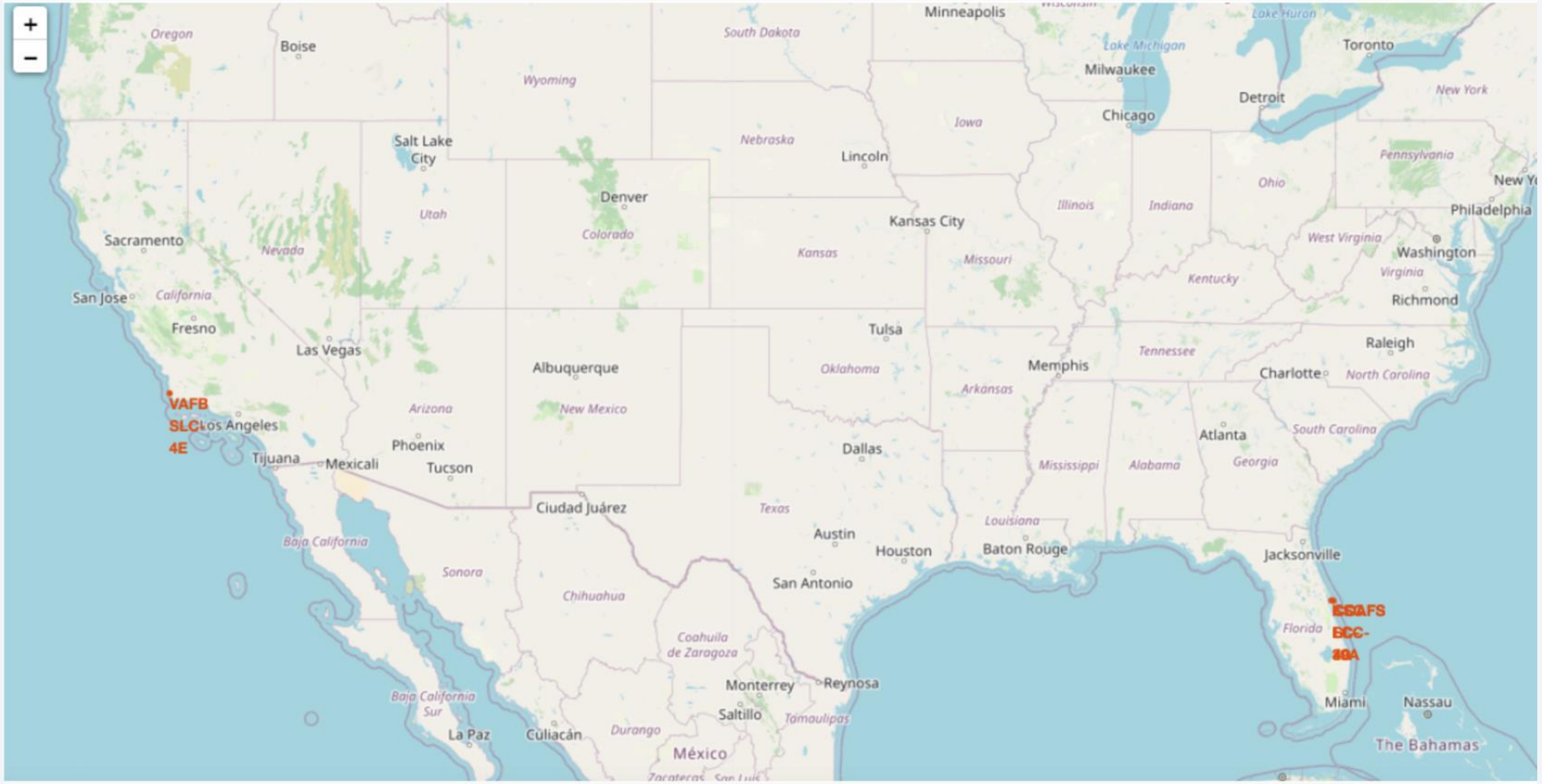
- The count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

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

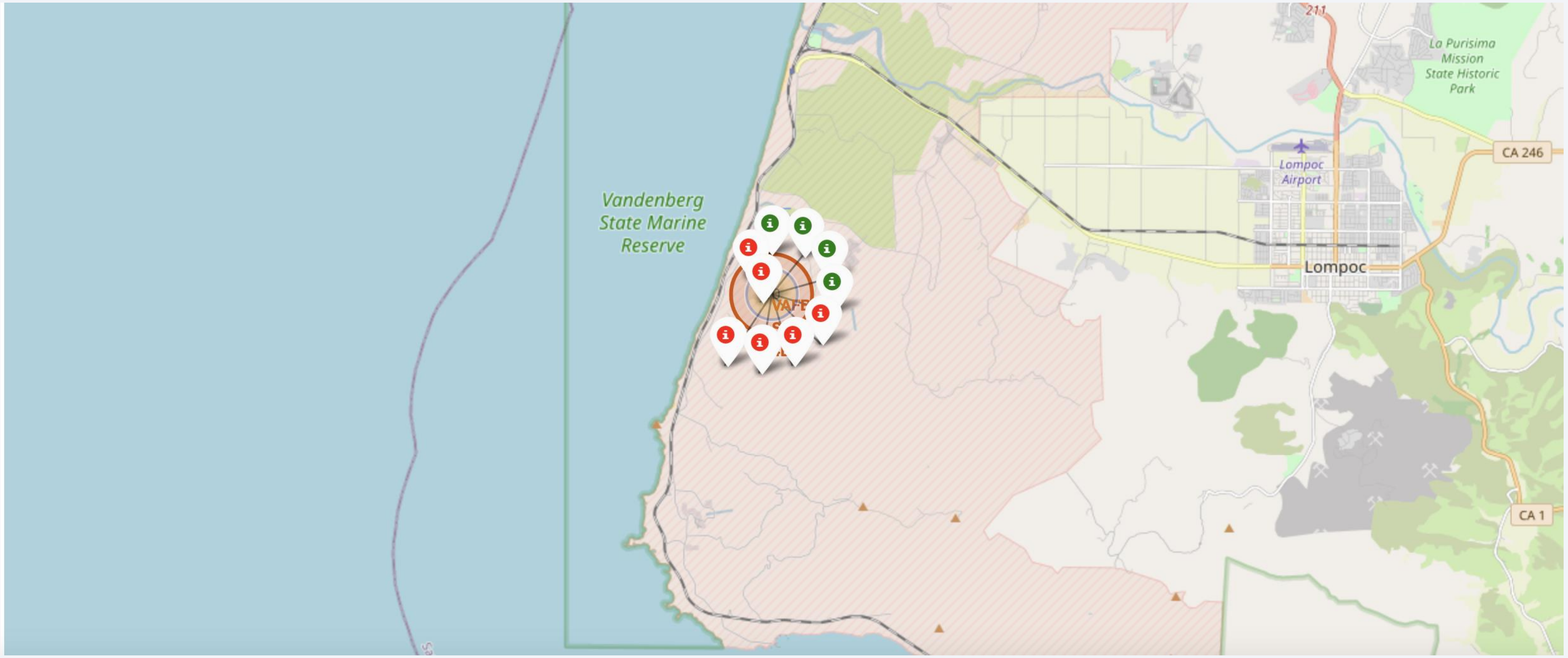
Section 3

Launch Sites Proximities Analysis

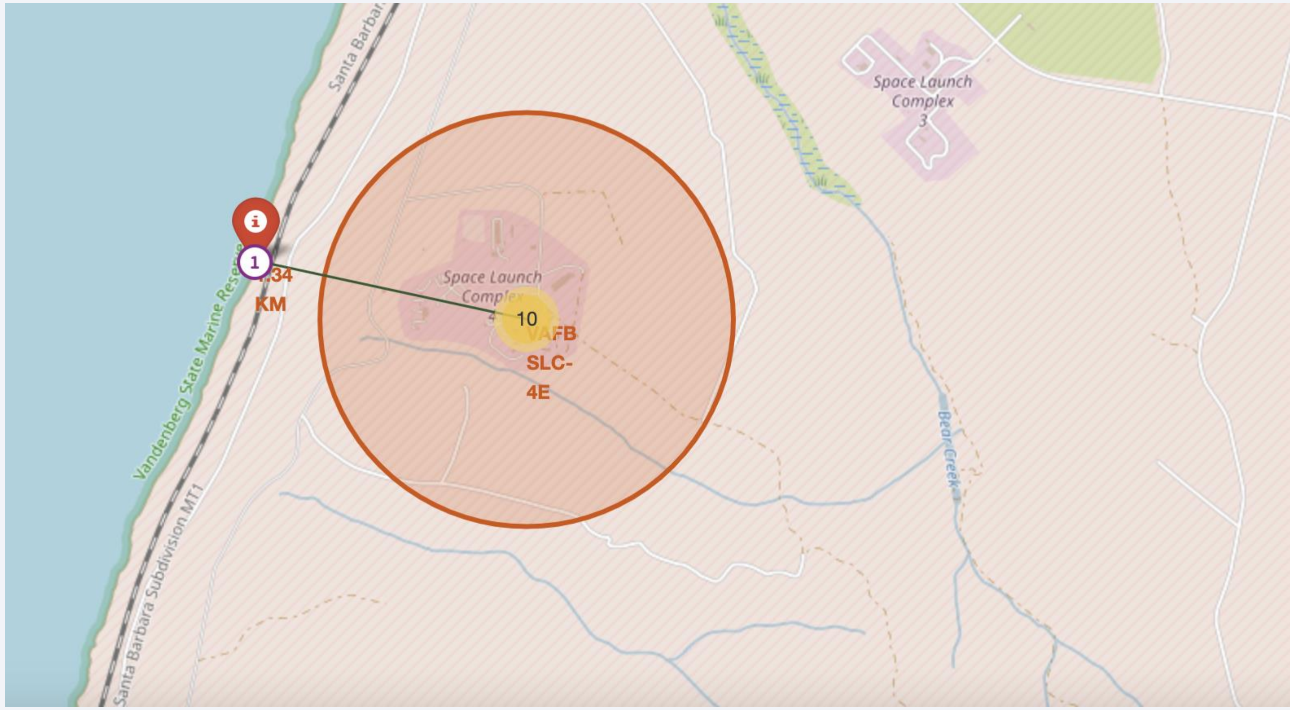
<Folium Map Screenshot 1>



<Folium Map Screenshot 2>



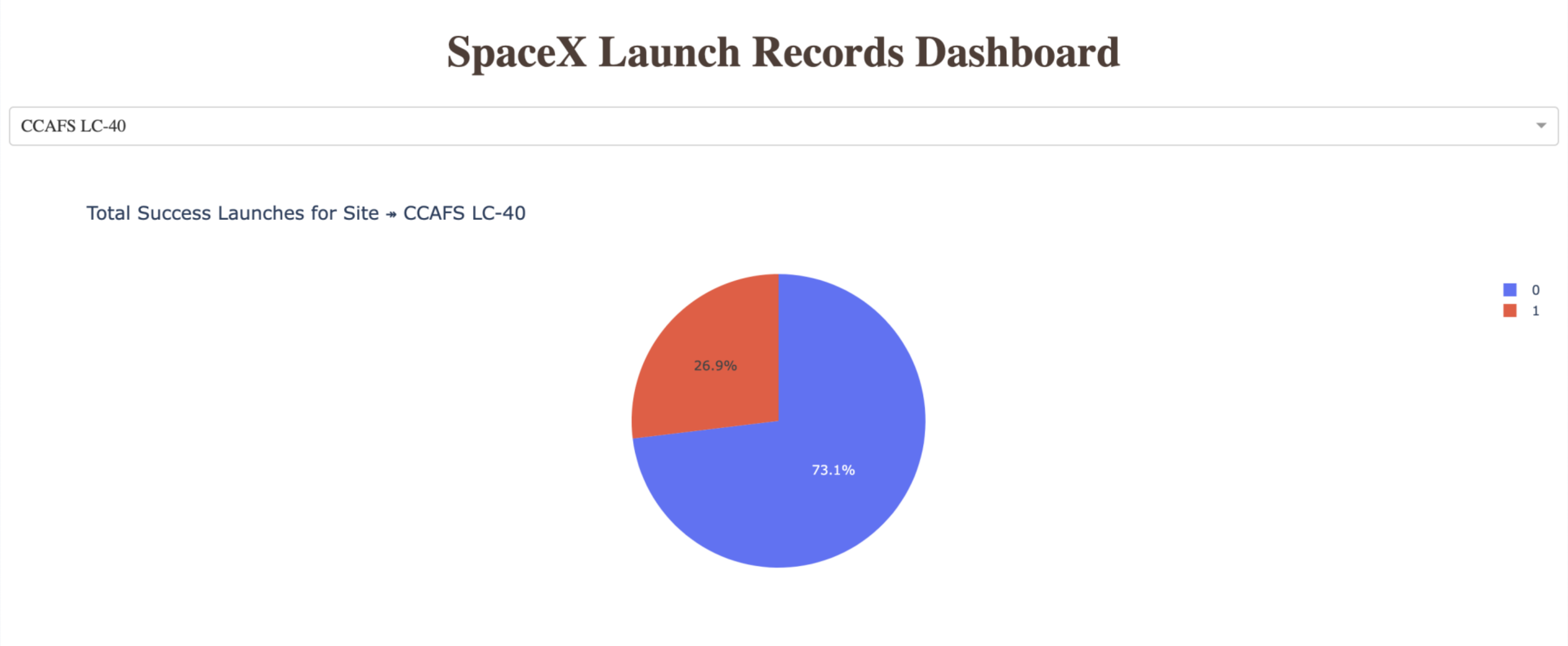
<Folium Map Screenshot 3>



Section 4

Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>



<Dashboard Screenshot 2>

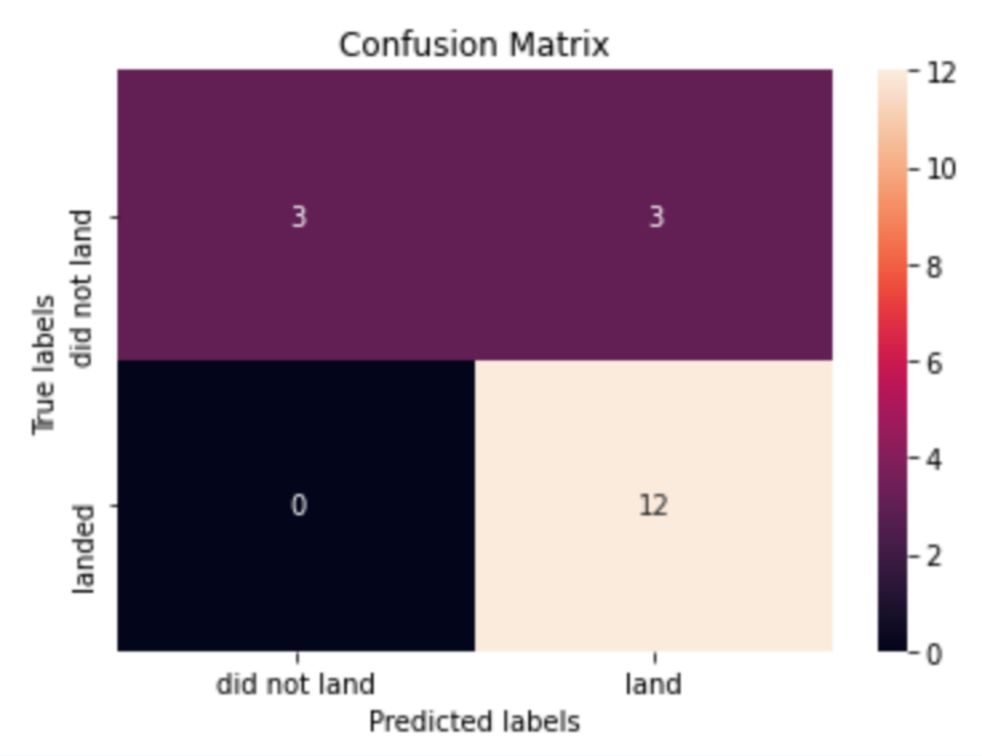
<Dashboard Screenshot 3>

Section 5

Predictive Analysis (Classification)

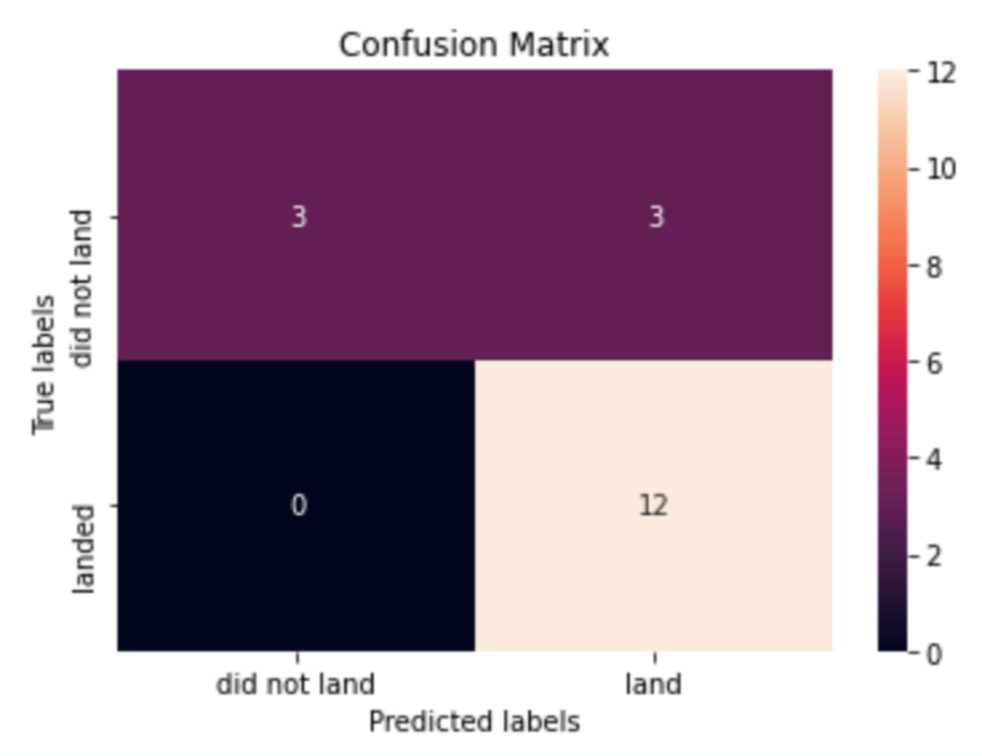
Classification Accuracy

- Logistic regression
  - GridSearchCV best score: 0.8464285714285713
  - Accuracy score on test set: 0.8333333333333334
  - Confusion matrix:



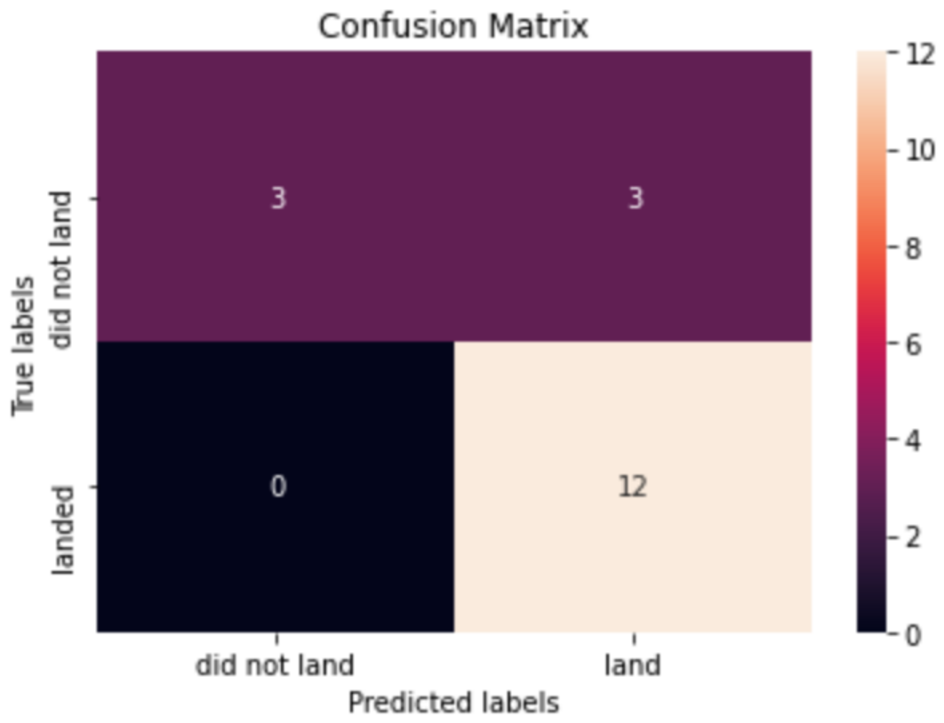
Confusion Matrix

- Support vector machine (SVM)
  - GridSearchCV best score: 0.8482142857142856
  - Accuracy score on test set: 0.8333333333333334
  - Confusion matrix:



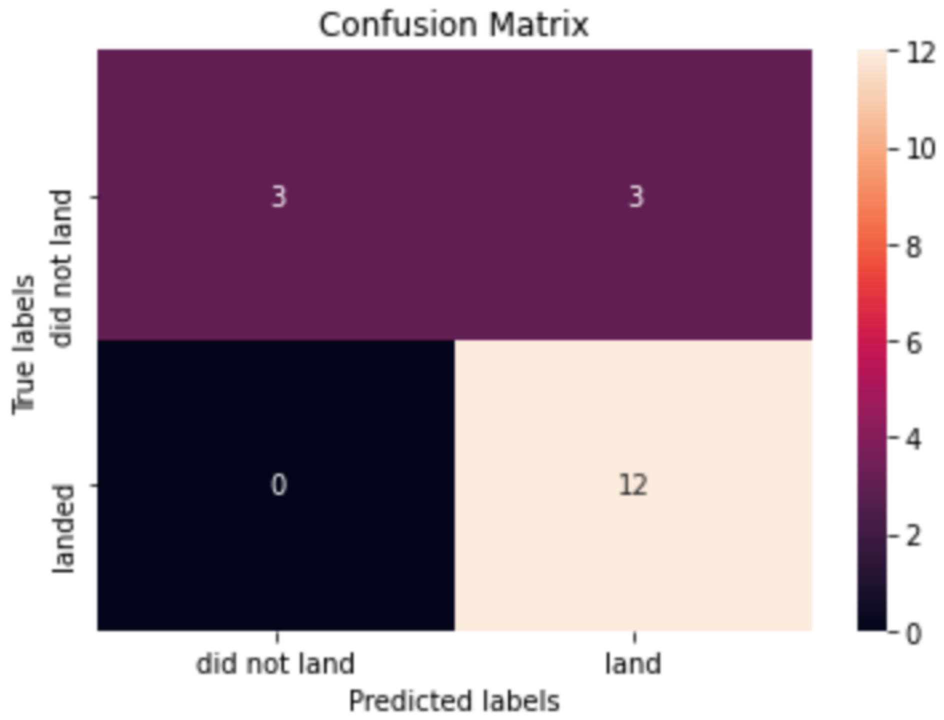
Confusion Matrix

- Decision tree
  - GridSearchCV best score: 0.8892857142857142
  - Accuracy score on test set: 0.8333333333333334
  - Confusion matrix:



Confusion Matrix

- K nearest neighbors (KNN)
  - GridSearchCV best score: 0.8482142857142858
  - Accuracy score on test set: 0.8333333333333334
  - Confusion matrix:



Conclusions

•In this project, we aim to predict whether the first stage of a Falcon 9 launch will land successfully, which helps determine the cost of a launch.
•Various features of a Falcon 9 launch, such as payload mass and orbit type, may influence the mission outcome.
•Several machine learning algorithms are applied to past Falcon 9 launch data to create predictive models.
•Among the four machine learning algorithms used, the decision tree algorithm produced the most accurate predictive model.

Thank you!