IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Grey>
<2024.8>

Outline

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

Executive Summary

We will predict if the SpaceX Falcon 9 first stage will land successfully using several machine learning classification algorithms.

The main process:

- Data collection, wrangling
- Exploratory data analysis using SQL
- Interactive data visualization
- Machine learning prediction

We obtain some preliminary insights about how each important variable would affect the success rate.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

In this capstone project, we aim to predict the successful landing of the Falcon 9 first stage. SpaceX advertises its Falcon 9 rocket launches on its website at a cost of $62 million; in contrast, other providers charge upwards of $165 million per launch, with much of the savings attributed to SpaceX's ability to reuse the first stage.

Consequently, if we can ascertain whether the first stage will land successfully, we can also estimate the overall cost of a launch. This information could be valuable for alternative companies seeking to compete against SpaceX for rocket launch contracts. Most unsuccessful landings are premeditated; occasionally, SpaceX conducts controlled ocean landings as part of their strategy. The primary question we seek to address is: given a specific set of features related to a Falcon 9 rocket launch—including payload mass, orbit type, and launch site—will the first stage successfully achieve landing?"

Section 1

Methodology

Methodology

- Data collection methodology:
  - SpaceX API
  - Web scraping
- Perform data wrangling
  - Pandas and NumPy
  - SQL
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
  - Matplotlib and Seaborn
  - Folium
  - Dash
- Perform predictive analysis using classification models,using:
  - Logistic regression
  - Support vector machine (SVM)
  - Decision tree
  - K–nearest neighbors (KNN)

Data Collection

Data Collection – SpaceX API

The API used is https://api.spacexdata.com/v4/rockets/. The API provides data about many types of rocket launches done by SpaceX, the data is therefore filtered to include only Falcon 9 launches.

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	Gr
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	

calls here

GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	La
False	False	False	None	1.0	0	B0003	-80.577366	28.
False	False	False	None	1.0	0	B0005	-80.577366	28.
False	False	False	None	1.0	0	B0007	-80.577366	28.
False	False	False	None	1.0	0	B1003	-120.610829	34.6
False	False	False	None	1.0	0	B1004	-80.577366	28.

Perlink for reference:

https://github.com/grey-0423/Applied-Data-Science-Capstone/blob/a48231e89f266038c9fa40f58dbca64f0eb198be/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

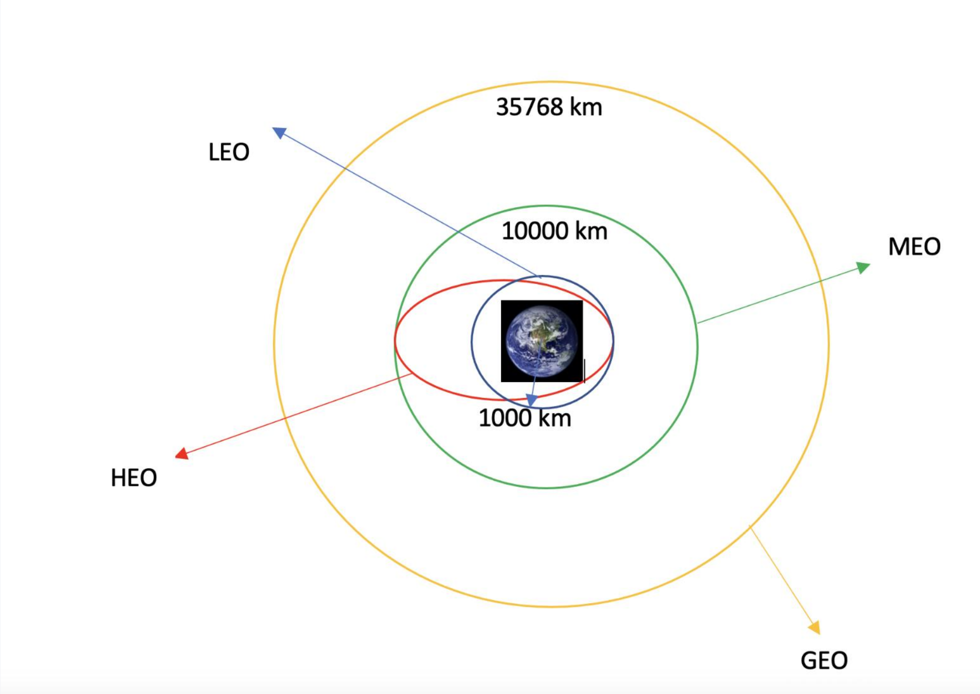
The data is scraped from https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
The website contains only the data about Falcon 9 launches.
We end up with 121 rows or instances and 11 columns or features.

```
[14]: extracted_row = 0
#Extract each table
for table_number,table in enumerate(soup.find_all('table',"wikitable plainrowheaders collapsible")):
   # get table row
    for rows in table.find_all("tr"):
        #c E114 indentation is not a multiple of 4 eading is as number corresponding to launch a number
        if (comment) (pycodestyle)
            if rows.th.string:
                flight_number=rows.th.string.strip()
                flag=flight_number.isdigit()
        else:
            flag=False
        #get table element
        row=rows.find_all('td')
        #if it is number save cells in a dictonary
        if flag:
            extracted_row += 1
            # Flight Number value
            # TODO: Append the flight_number into launch_dict with key `Flight No.`
            launch_dict['Flight No.'].append(flight_number)
            print(flight_number)
            #print(flight_number)
            datatimelist=date_time(row[0])

            # Date value
            # TODO: Append the date into launch_dict with key `Date`
            date = datatimelist[0].strip(',')
            #print(date)
            launch_dict['Date'].append(date)
            print(date)

            # Time value
            # TODO: Append the time into launch_dict with key `Time`
            time = datatimelist[1]
```

Data Wrangling

- The dataset is subsequently processed to ensure the absence of missing entries, and categorical features are encoded utilizing one-hot encoding techniques. Additionally, a new column labeled ‘Class’ is incorporated into the data frame, where this column indicates a value of 0 for failed launches and 1 for successful ones. Ultimately, we arrive at a total of 90 rows or instances and 83 columns or features.

https://github.com/grey-0423/Applied-Data-Science-Capstone/blob/df4acfc888a78d0c645cdf12c1e0e493168c6918/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

Pandas and NumPy	Dash	Matplotlib and Seaborn	Folium
• The number of launches on each launch site • The number of occurrence of each orbit • The number and occurrence of each mission outcome	• 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	• visualize the data through scatterplots, bar charts, and line charts. • understand more about the relationships between several features	• 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

https://github.com/grey-0423/Applied-Data-Science-Capstone/blob/a48231e89f266038c9fa40f58dbca64f0eb198be/visualization.ipynb

EDA with SQL

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

https://github.com/grey-0423/Applied-Data-Science-Capstone/blob/a48231e89f266038c9fa40f58dbca64f0eb198be/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Done.

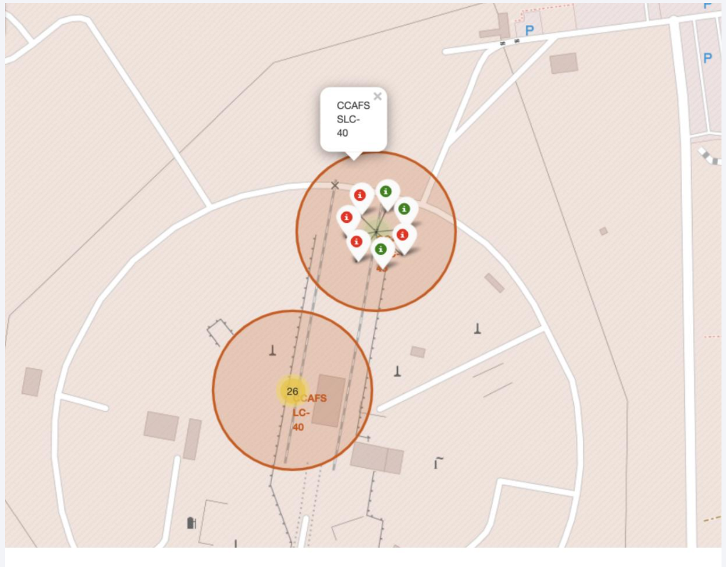
15]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Cu
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	
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)	
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	

Build an Interactive Map with 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



—Be able to easily identify which launch sites have relatively high success rates.

https://github.com/grey-0423/Applied-Data-Science-Capstone/blob/a48231e89f266038c9fa40f58dbca64f0eb198be/lab_jupyter_launch_site_location.ipynb

Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model
- You need present your model development process using key phrases and flowchart
- Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose

Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

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

- 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

Results

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

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

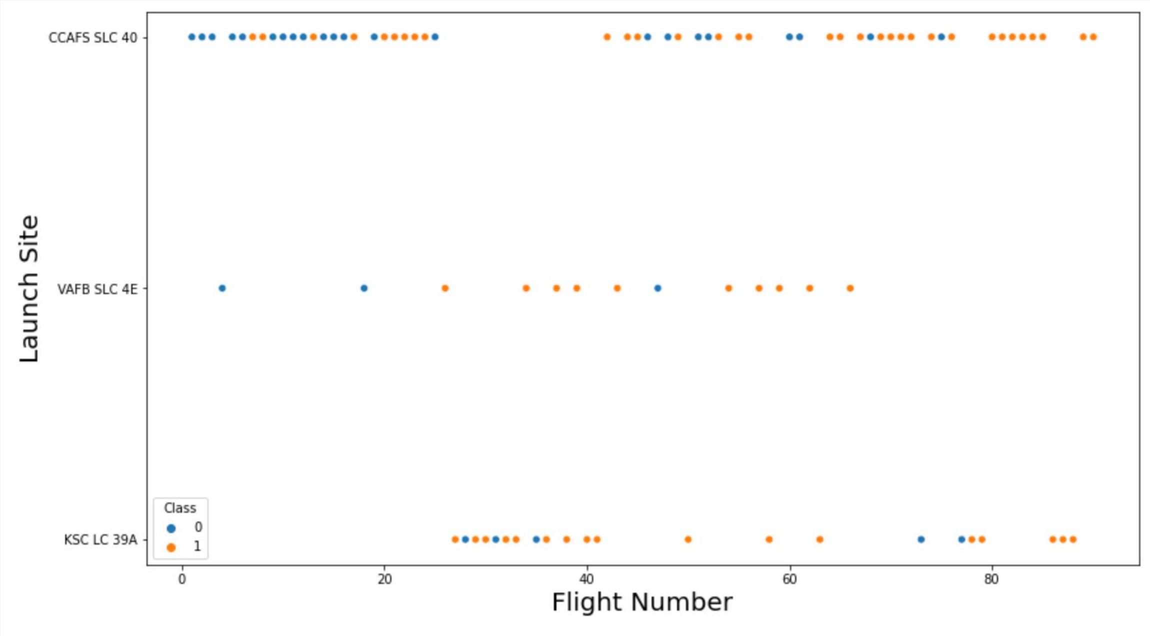
landing__outcome	landing_count
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 2

Insights drawn from EDA

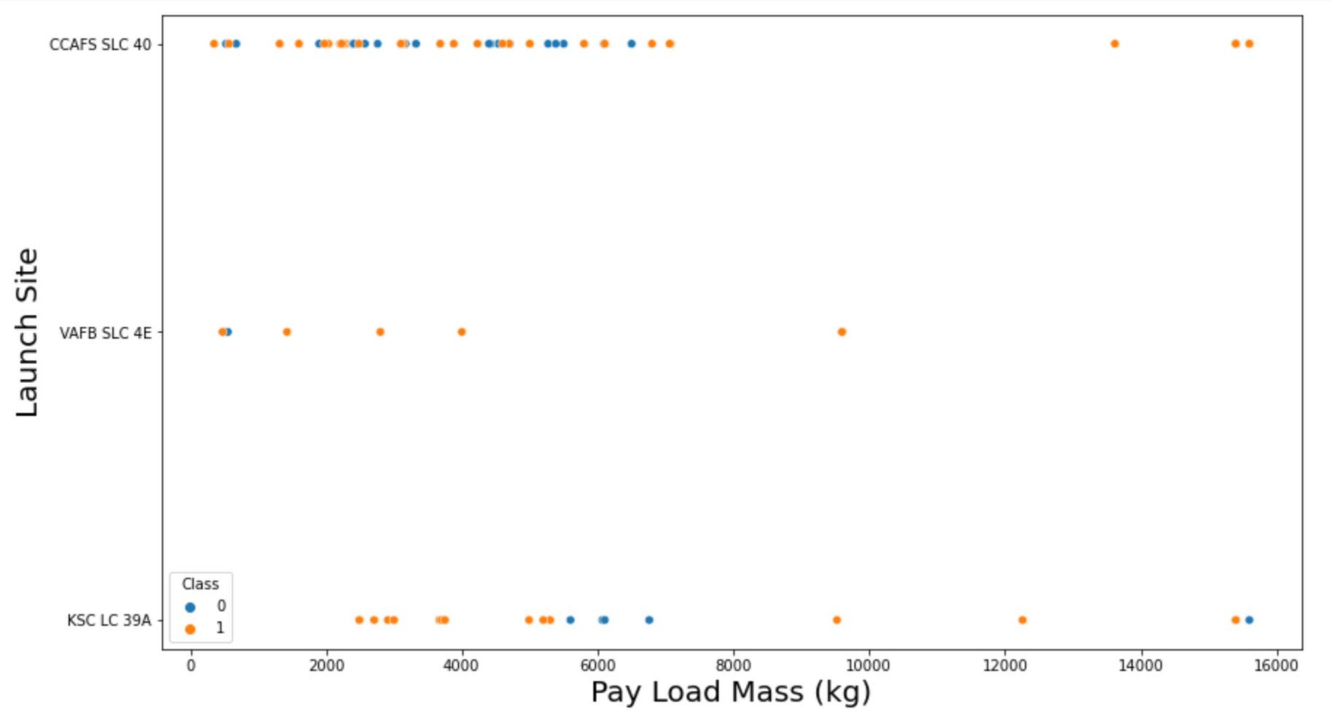
Flight Number vs. Launch Site

- Show a scatter plot of Flight Number vs. Launch Site



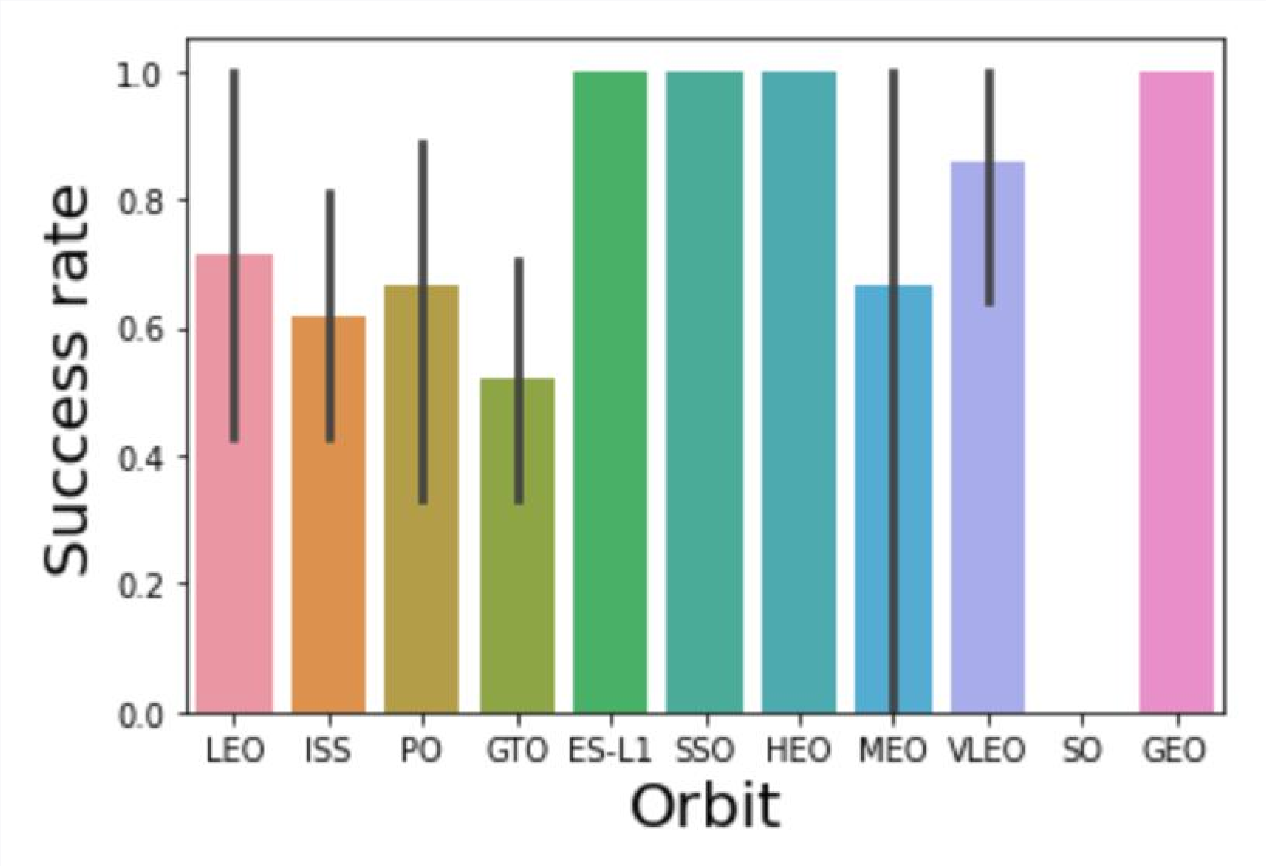
Payload vs. Launch Site

- Show a scatter plot
  of Payload vs. Launch Site



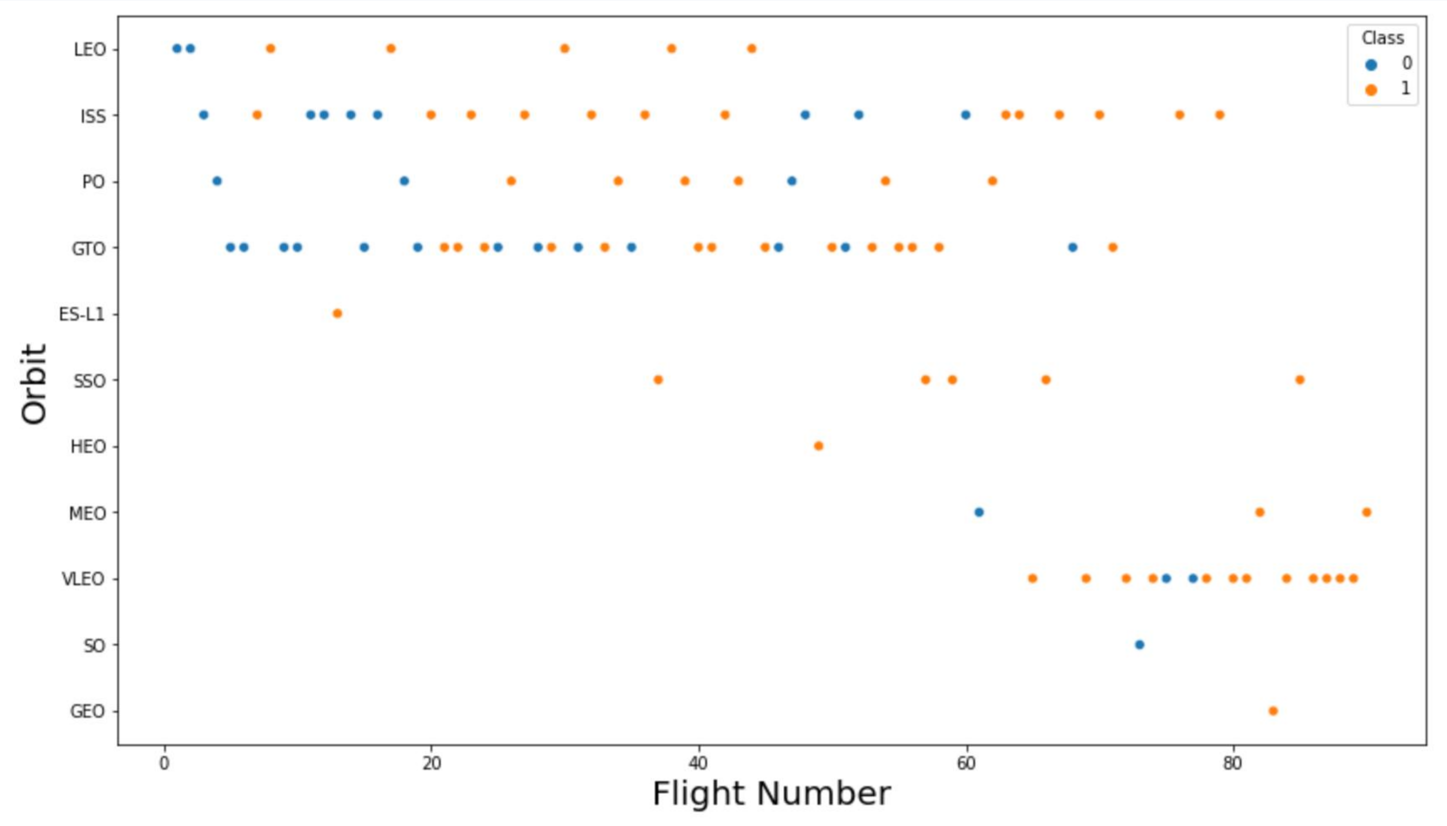
Success Rate vs. Orbit Type

- Show a bar chart for the success rate of each orbit type



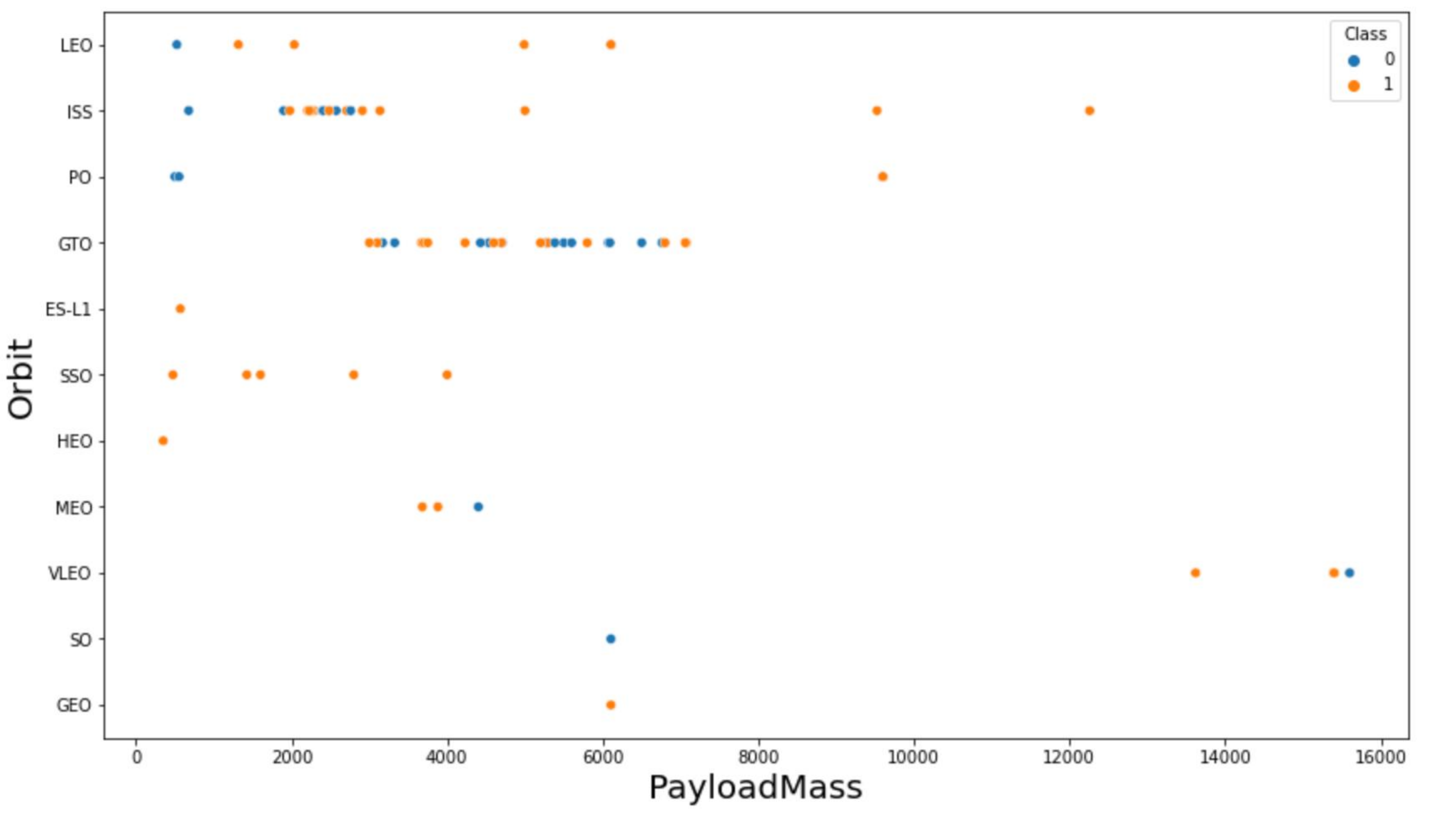
Flight Number vs. Orbit Type

- Show a scatter point of Flight number vs. Orbit type



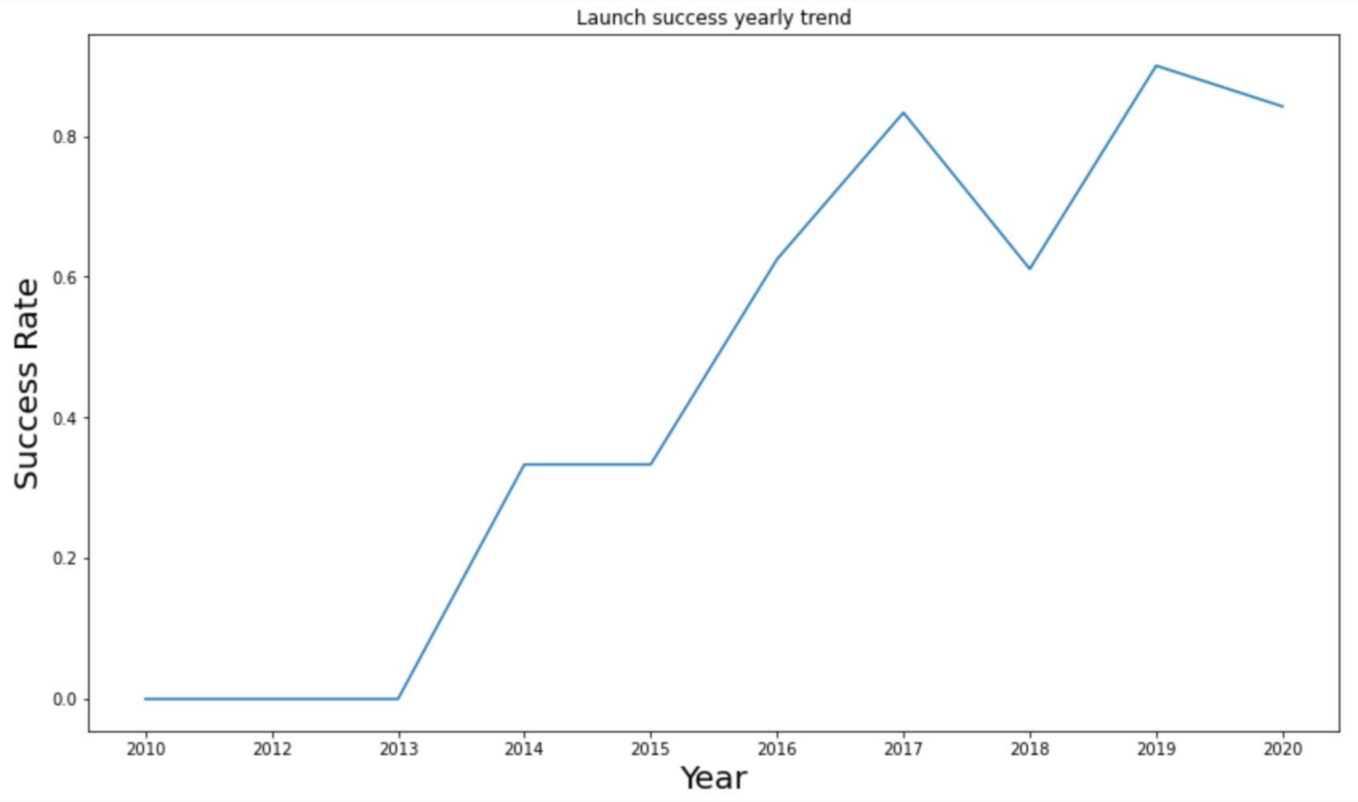
Payload vs. Orbit Type

- Show a scatter point of payload vs. orbit type



Launch Success Yearly Trend

- Show a line chart of yearly average success rate



This chart shows the increasing trend of SpaceX launch success, which, despite setbacks in 2018, has since returned to a high success rate.

Launch Site Names Begin with 'CCA'

The names of the unique launch sites

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

Total Payload Mass

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

Total payload mass by NASA (CRS)
45596

Average Payload Mass by F9 v1.1

- Calculate 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

- Find the dates of the first successful landing outcome on ground pad

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

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had 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

- Calculate the total number of successful and failure mission outcomes

number_of_success_outcomes	number_of_failure_outcomes
100	1

Boosters Carried Maximum Payload

- List the names of the booster 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

- List 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

- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) 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

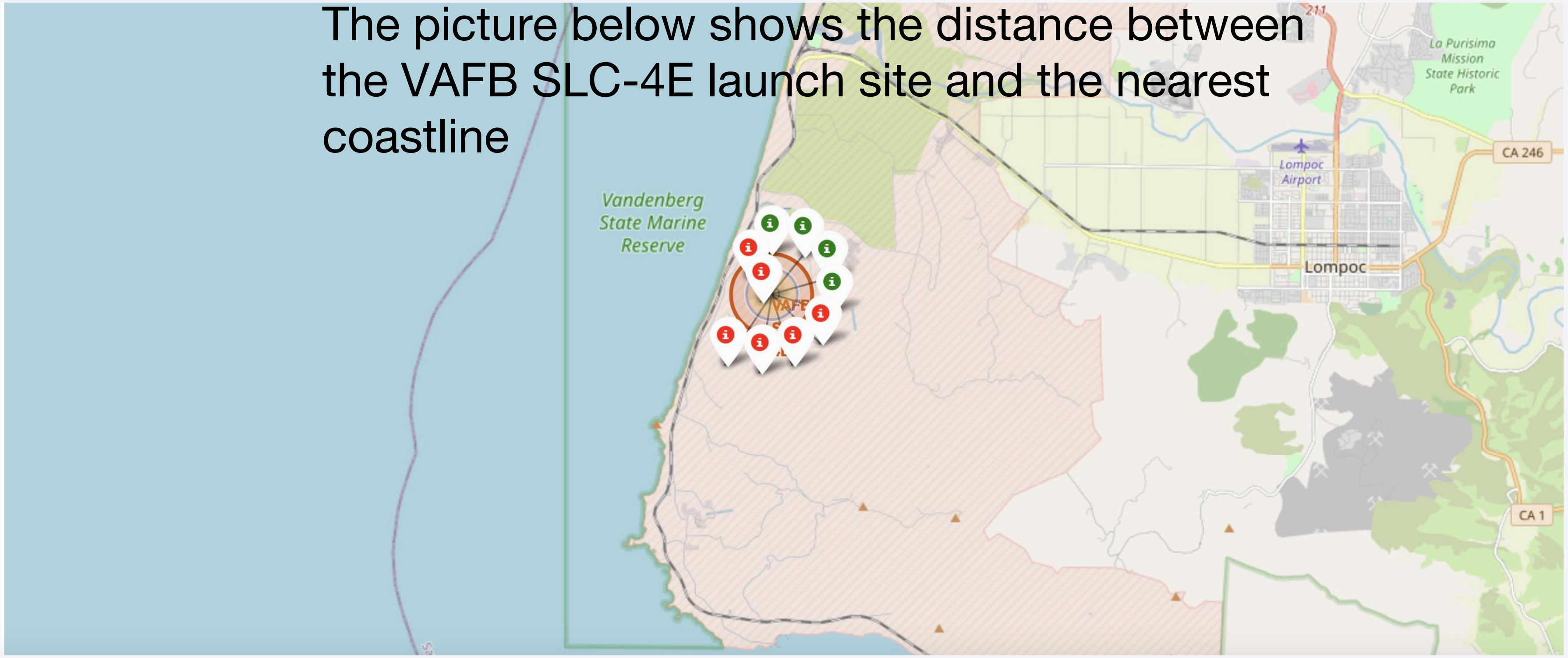
Section 3

Launch Sites Proximities Analysis

<Folium Map 1>

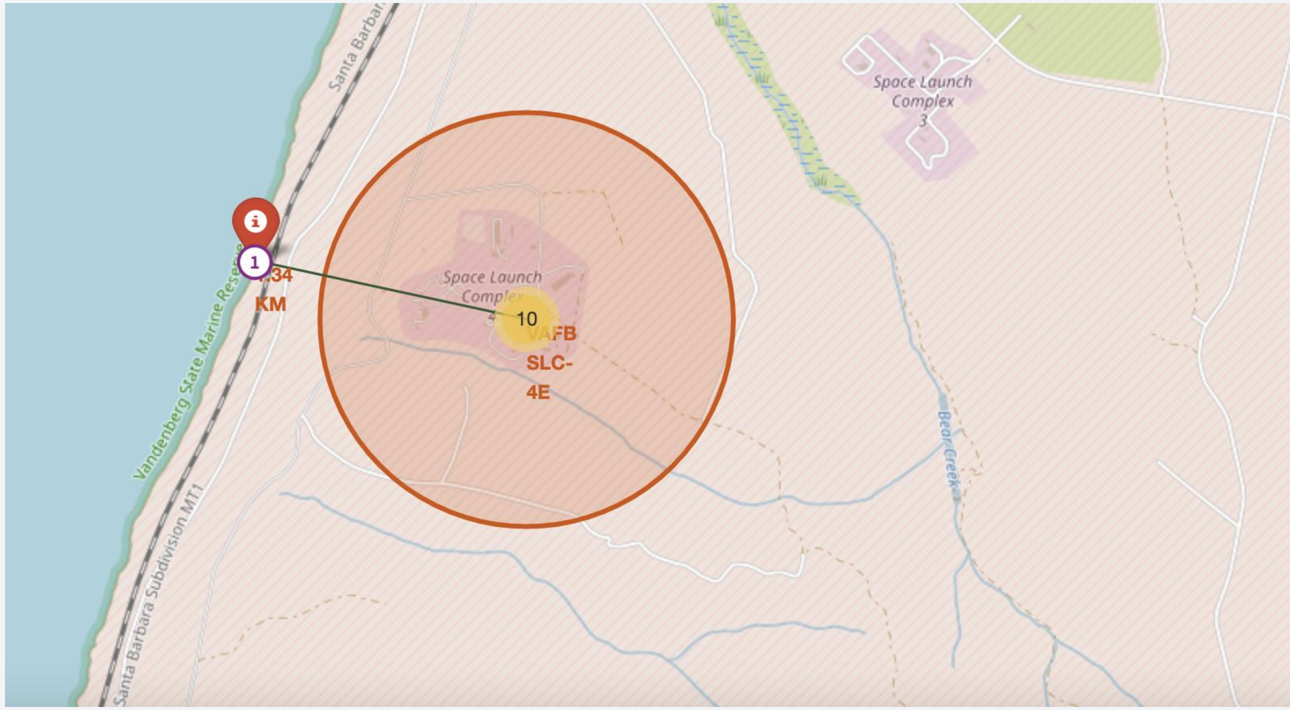
The successful and unsuccessful launches for each site are displayed on the map. By zooming in on a specific launch site, we can observe green and red markers; each green marker signifies a successful launch, while each red marker indicates a failed launch

The picture below shows the distance between the VAFB SLC-4E launch site and the nearest coastline



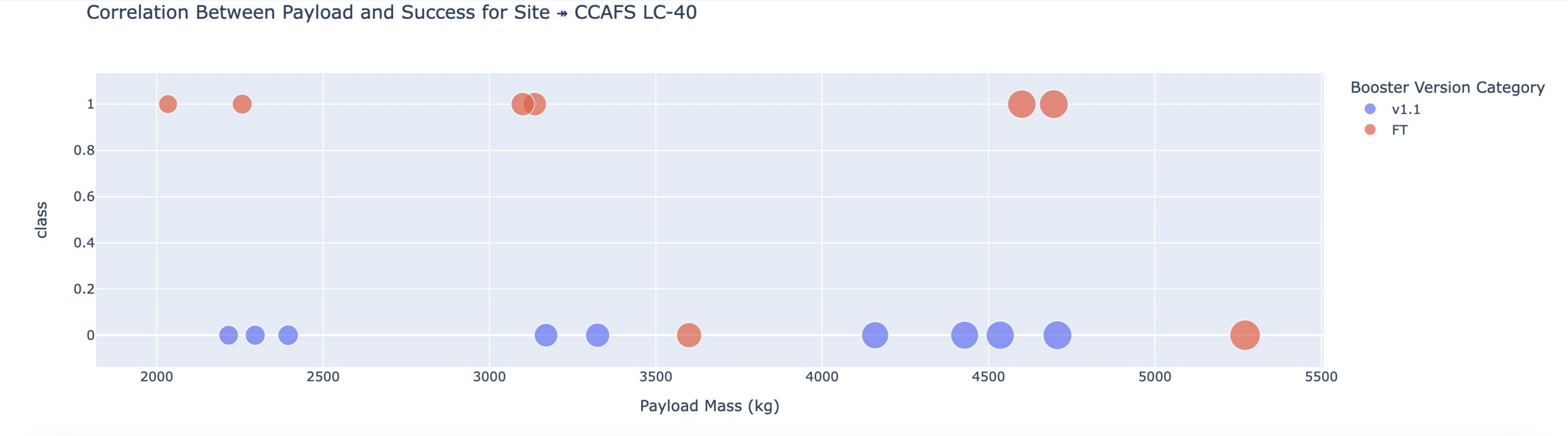
<Folium Map 2>

The picture below shows the distance between the VAFB SLC-4E launch site and the nearest coastline



<Folium Map 3>

The picture below shows a scatterplot when the payload mass range is set to be from 2000kg to 8000kg.
v1.1 that failed launches clustered around roughly 2300kg,and 4200-4700kg



Section 4

Build a Dashboard with Plotly Dash

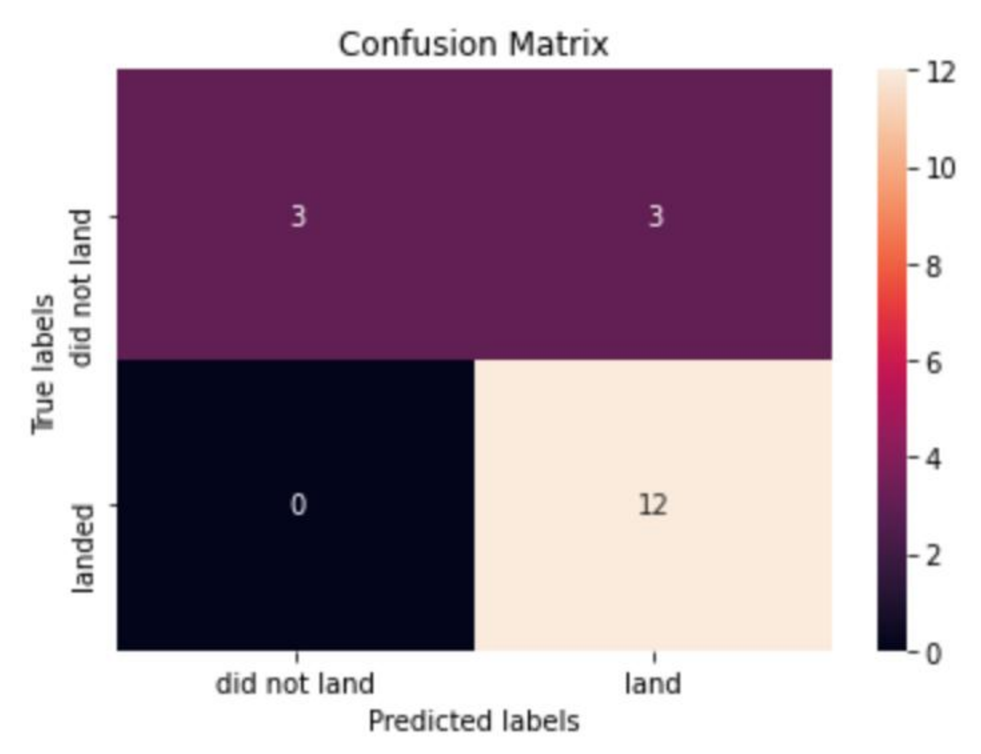
<1>

Logistic regression

GridSearchCV best score: 0.8464285714285713

Accuracy score on test set: 0.8333333333333334

Confusion matrix:

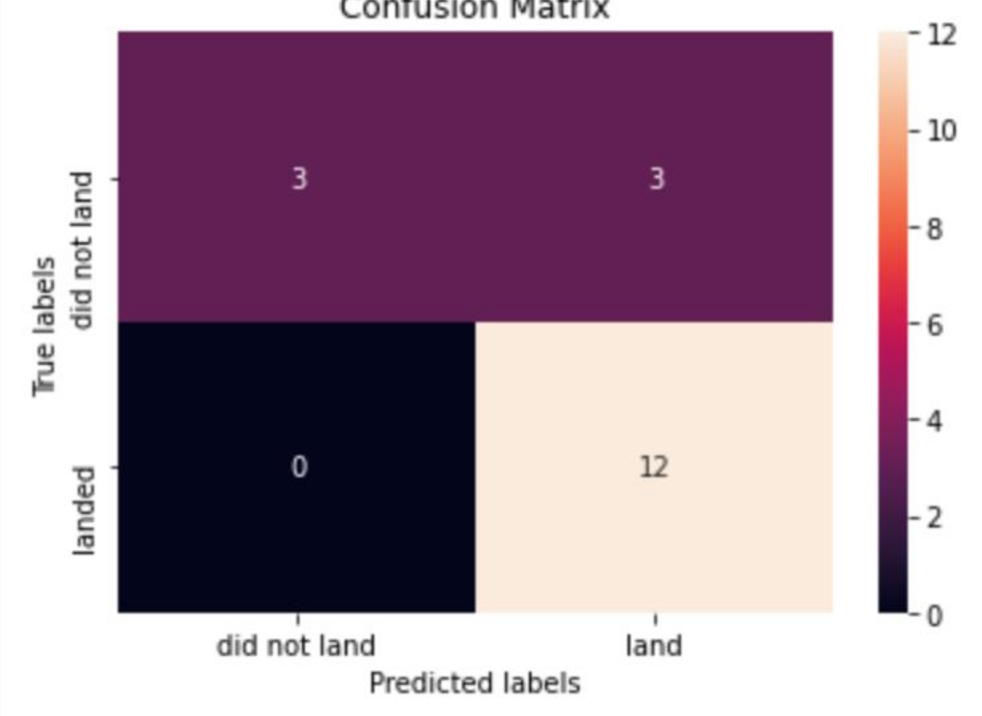


Support vector machine (SVM)

GridSearchCV best score: 0.8482142857142856

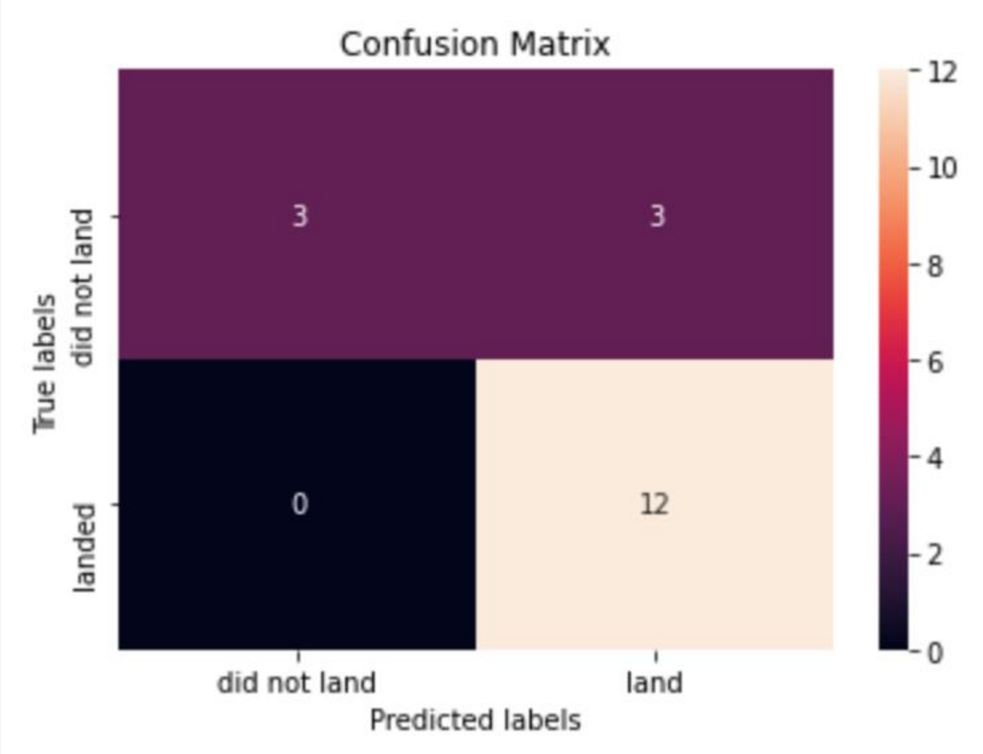
Accuracy score on test set: 0.8333333333333334

Confusion matrix:

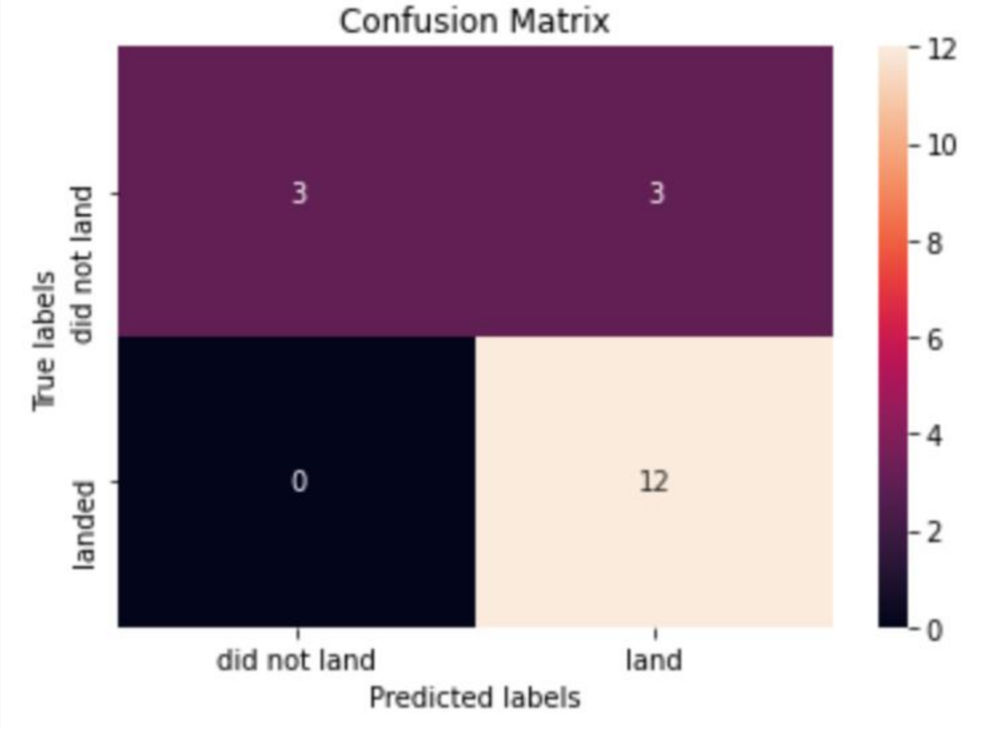


< 2>

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



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



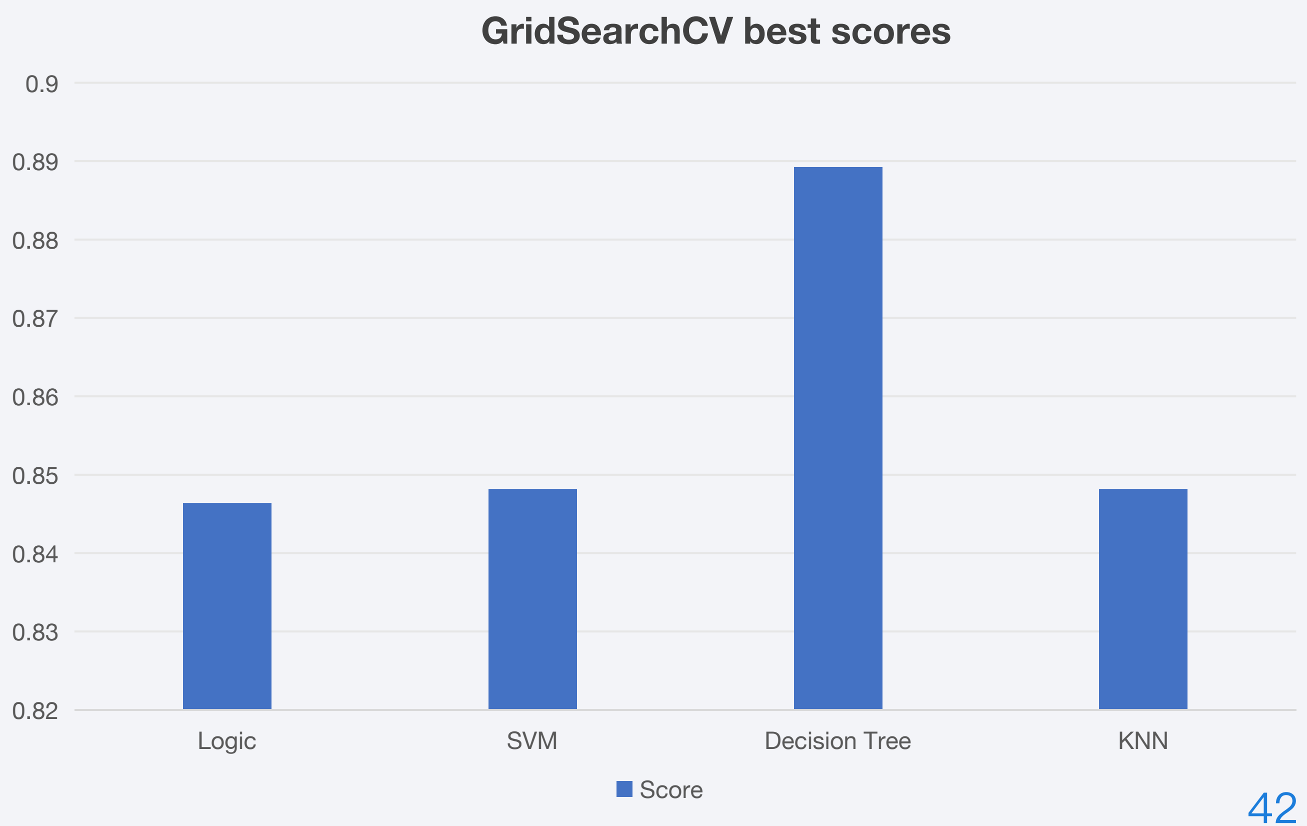
Section 5

Predictive Analysis (Classification)

Classification Accuracy

They all share the same accuracy score and confusion matrix: 0.8333334,we put their GridSearchCV best scores are used to rank them instead.

Decision tree has the best score



Conclusions

It is evident that certain features may correlate with mission outcomes in various ways. For instance, missions involving heavy payloads tend to exhibit higher successful landing rates for orbit types such as Polar, LEO, and ISS.

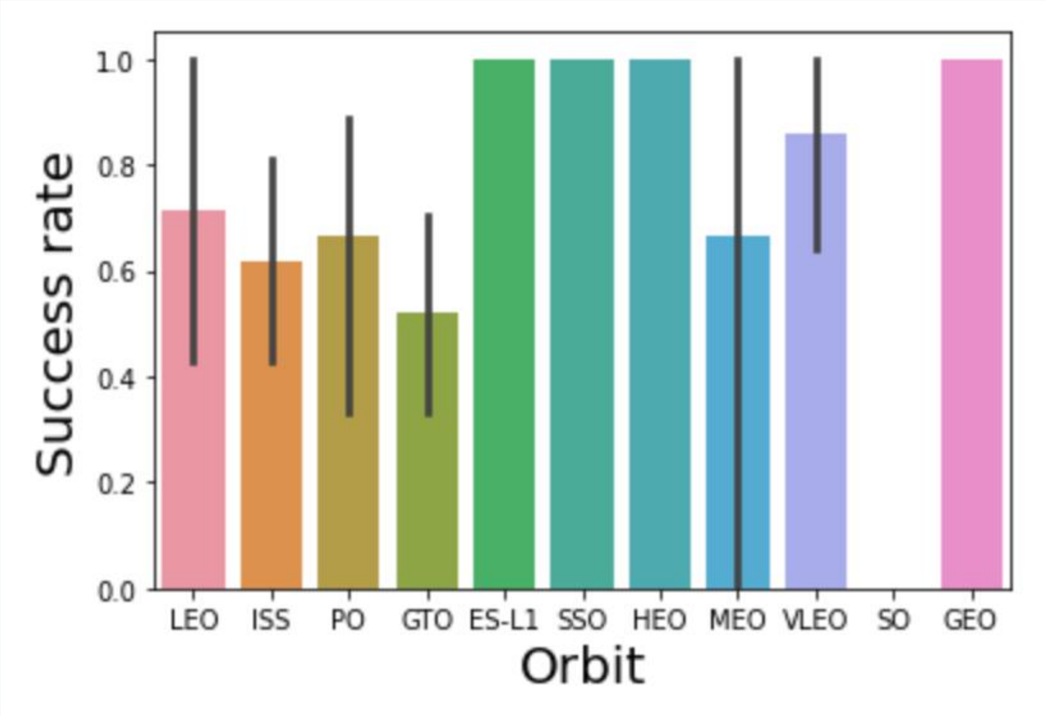
Conversely, for GTO missions, the distinction is less clear as both successful and unsuccessful landings are present. Consequently, each feature appears to exert a specific influence on the final mission outcome.

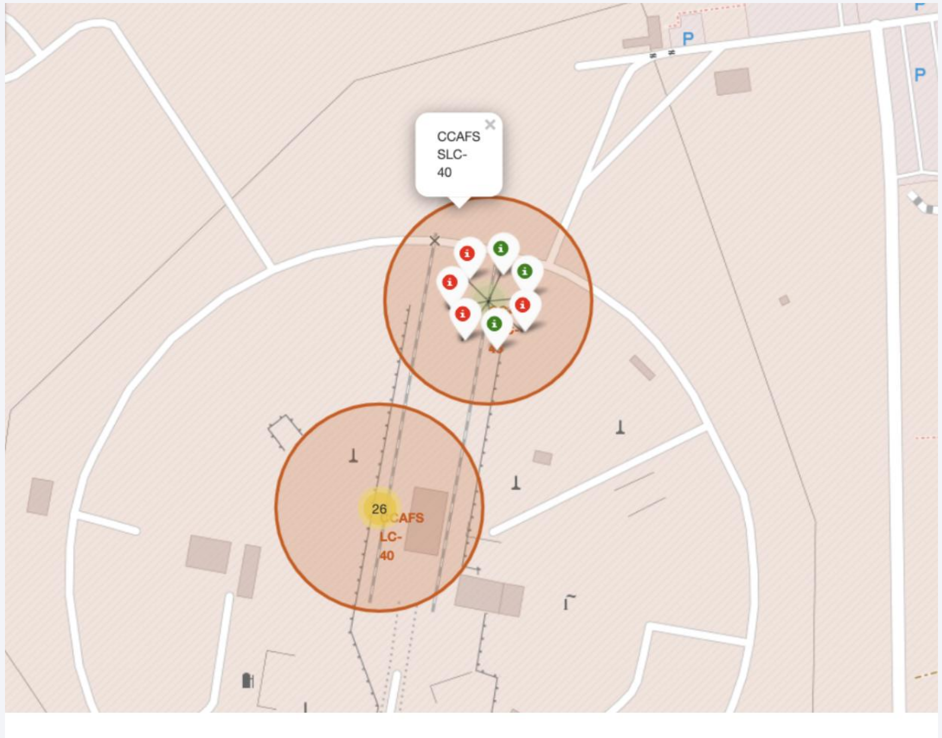
The precise mechanisms by which these features affect the mission results remain challenging to elucidate.

Nevertheless, we can employ machine learning algorithms to analyze historical data patterns and predict the likelihood of mission success based on the identified features.

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!