IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Khulasande Gumede
7 Jan 2025

Outline

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

Executive Summary

- This project involves data collection, data wrangling, exploratory data analysis, interactive visual analytics and predictive analysis for the successful landing of Falcon 9 first stage.
- The prediction achieves over 83% of accuracy.

Introduction

- We will predict if the Falcon 9 first stage will land successfully
- SpaceX 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 SpaceX 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 SpaceX for a rocket launch.

Section 1

Methodology

Methodology

Executive Summary

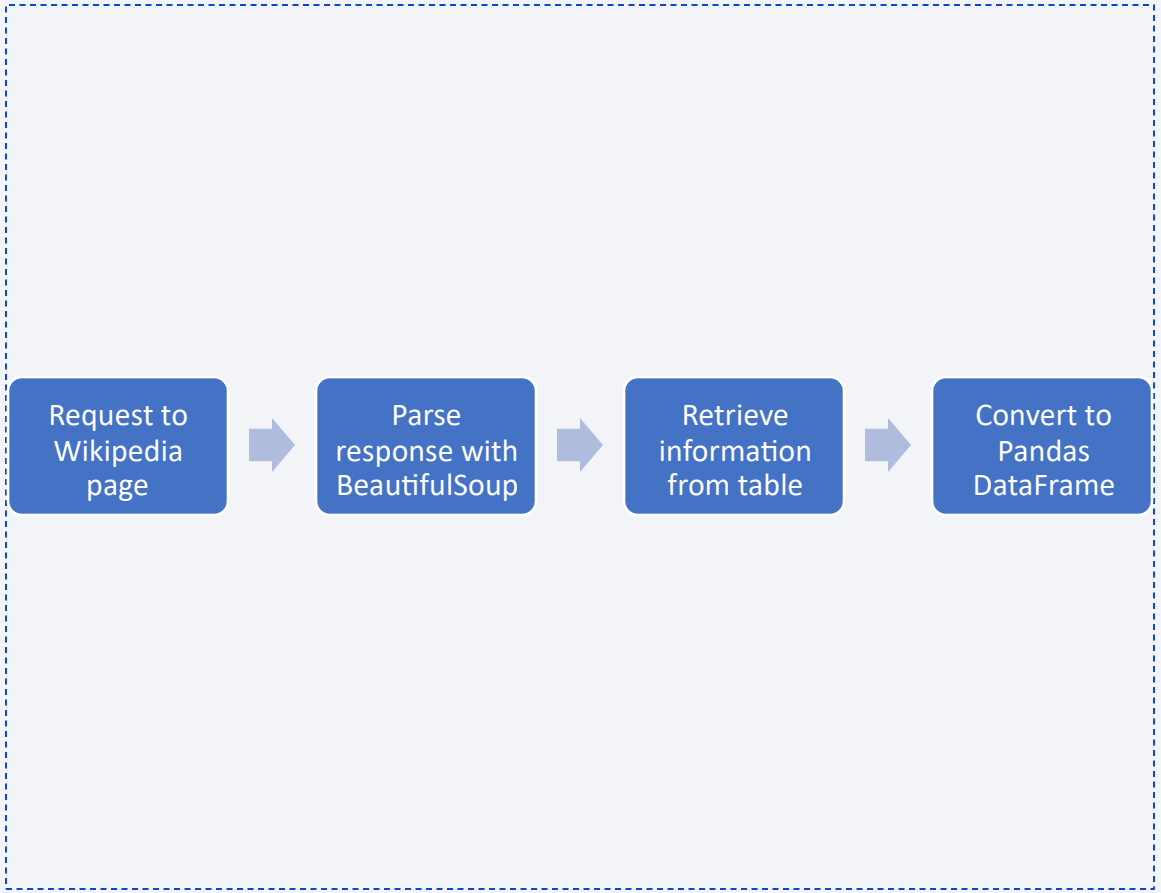
- Data collection methodology:
  - Request to SpaceX API
  - Web Scraping from Wikipedia using BeautifulSoup
- Perform data wrangling
  - Convert the Landing Outcome into a classification variable of zero or one using Pandas
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Try different models and tune with GridSearchCV

Data Collection – SpaceX API

- Request to SpaceX API and convert the JSON response to Pandas DataFrame
- https://github.com/harrycch/applied_data_science_capstone/blob/61d664cc59f806d6641cb47705cbfe1f4dab1c5a/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

- Request to Wikipedia page and retrieve the response using BeautifulSoup and convert to Pandas DataFrame
- https://github.com/harrycch/applied_data_science_capstone/blob/61d664cc59f806d6641cb47705cbfe1f4dab1c5a/jupyter-labs-webscraping.ipynb



Data Wrangling

- Convert Landing Outcome to a classification variable
- https://github.com/harrycch/applied_data_science_capstone/blob/61d664cc59f806d6641cb47705cbfe1f4dab1c5a/labs-jupyter-spacex-Data%20wrangling.ipynb

Retreive all unique values of Landing Outcome → Identify which are bad outcomes → Assign a classification variable with 1 representing good outcome and 0 representing bad outcome

EDA with Data Visualization

- Scatter Plots to see the relationship between:
  - Flight Number vs Pay Load Mass
  - Flight Number vs Launch Site
  - Pay Load Mass vs Launch Site
  - Flight Number vs Orbit
  - Pay Load Mass vs Orbit
- Bar chart to see the relationship between the Orbit and Success Rate
- Line chart to see the relationship between the year and Success Rate
- https://github.com/harrycch/applied_data_science_capstone/blob/4acfdcec6304ae8682309152e39e0949bfc30341/edadataviz.ipynb

EDA with SQL

- SQL queries performed
  - select distinct "Launch_Site" from SPACEXTABLE
  - select * from SPACEXTABLE where " Launch_Site" like 'CCA%' limit 5
  - select sum("PAYLOAD_MASS__KG_") as "Total Payload Mass" from SPACEXTABLE where "Customer" is "NASA (CRS)"
  - select avg("PAYLOAD_MASS__KG_") as "Average Payload Mass" from SPACEXTABLE where " Booster_Version" like "F9 v1.1%"
  - select min(Date) from SPACEXTABLE where " Landing_Outcome" is "Success (ground pad)"
  - select "Booster_Version" from SPACEXTABLE where " Landing_Outcome" is "Success (drone ship)" and "PAYLOAD_MASS__KG_" between 4000 and 6000
  - select "Mission_Outcome", count("Mission_Outcome") from SPACEXTABLE group by " Mission_Outcome"
  - select "Booster_Version" from SPACEXTABLE where "PAYLOAD_MASS__KG_" = (Select max("PAYLOAD_MASS__KG_") from SPACEXTABLE)
  - select substr(Date,6,2) as "Month Name", "Landing_Outcome", "Booster_Version", "Launch_Site" from SPACEXTABLE where substr(Date,0,5)='2015' and "Landing_Outcome" is "Failure (drone ship)"
  - select "Landing_Outcome", count("Landing_Outcome") from SPACEXTABLE where Date between '2010 -06-04' and '2017-03-20' group by "Landing_Outcome" order by count("Landing_Outcome") desc
- https://github.com/harrycch/applied_data_science_capstone/blob/128198633b0fbdbb2899fd3b49cd49f214841028/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Circle and Marker added to the following to visualize the location of launch sites
  - NASA Center
  - All Launch SItes
- Added Marker Cluster with green and red markers indicating if the launch was succeeded or failed
- Added polyline and marker to indicate the distance between a launch site and coastline, railway, highway and city
- https://github.com/harrycch/applied_data_science_capstone/blob/d34a63d05d94d472a65aeff719e90e3b532e3eb9/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Pie Charts for Total Success Launches by Site
  - Also able to display the pie chart for each site
- Scatter Plot for Correlation between Payload and Launch Site
  - Also able to adjust the payload range for display
- https://github.com/harrycch/applied_data_science_capstone/blob/1636461c250f7b6edf6941aae37bc18cc1d5043f/spacex_dash_app.py

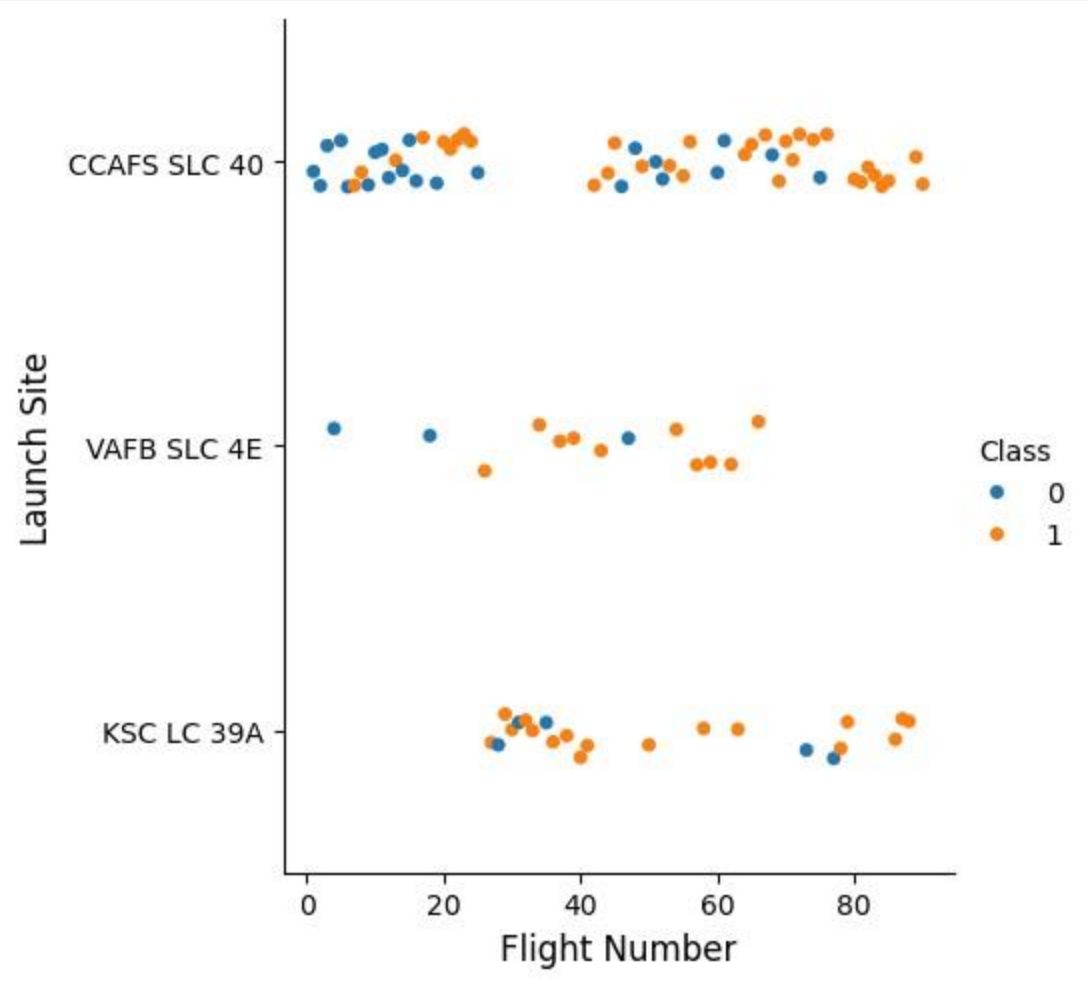
Predictive Analysis (Classification)

- Tune parameters of each model by GridSearchCV
- Try with the following models:
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree
  - K Nearest Neighbours
- https://github.com/harrycch/applied_data_science_capstone/blob/6a0f1d15c1d7a76073eb3c60d1c1f0a8891a6fb1/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

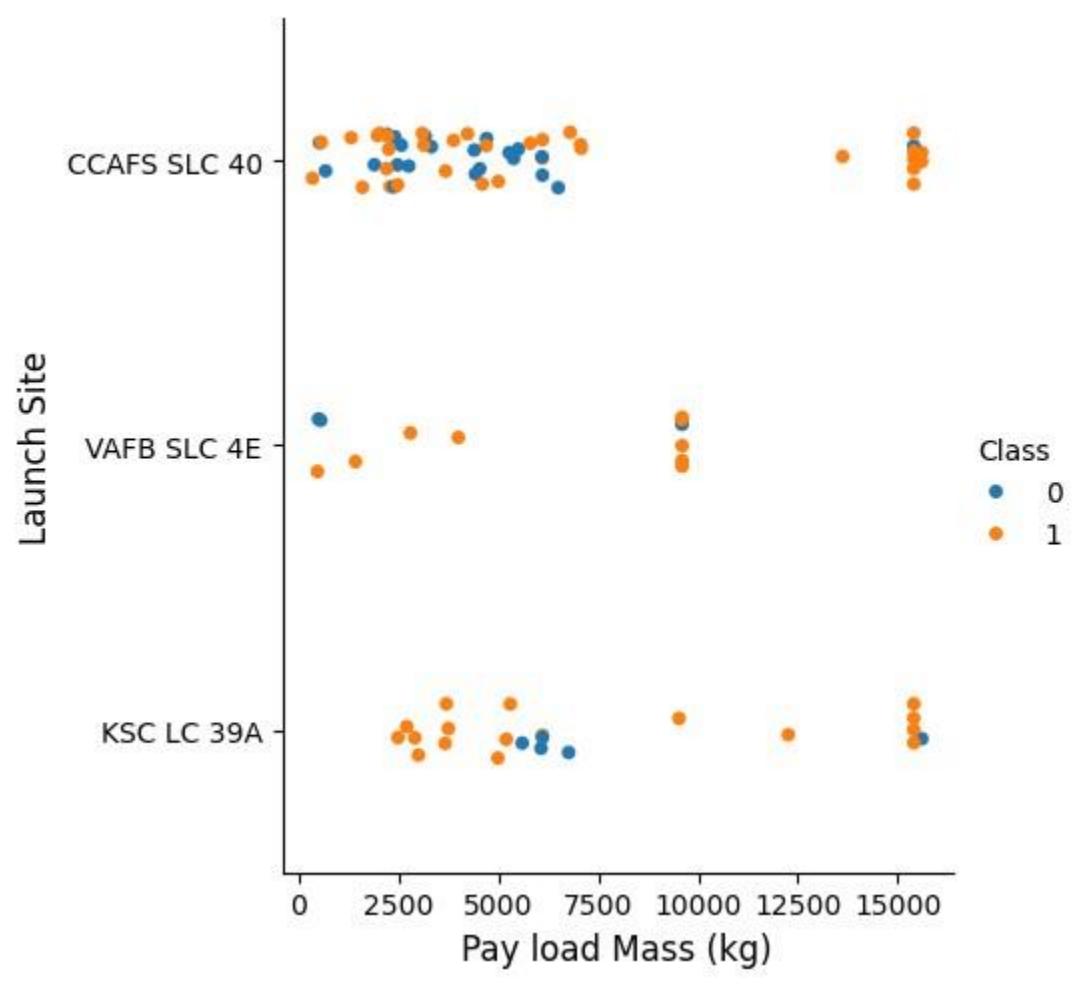
Section 2

Insights drawn from EDA

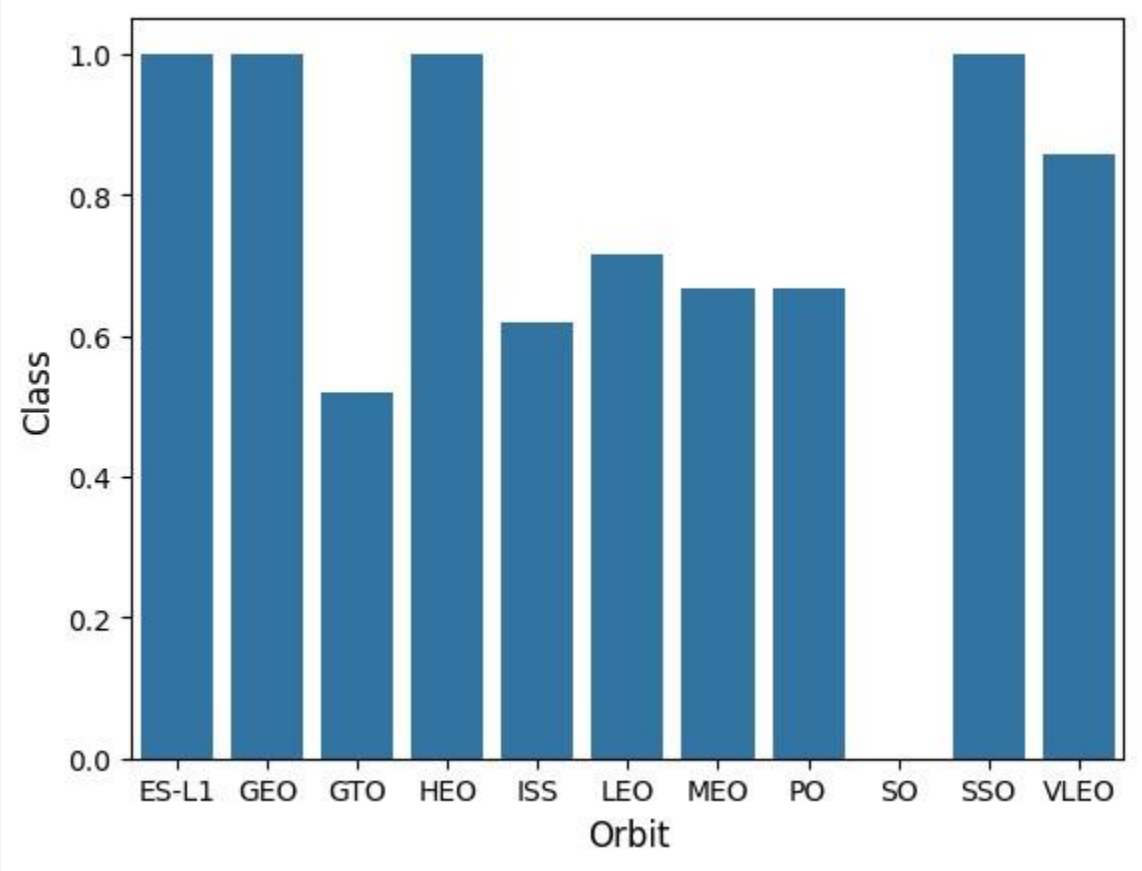
Flight Number vs. Launch Site



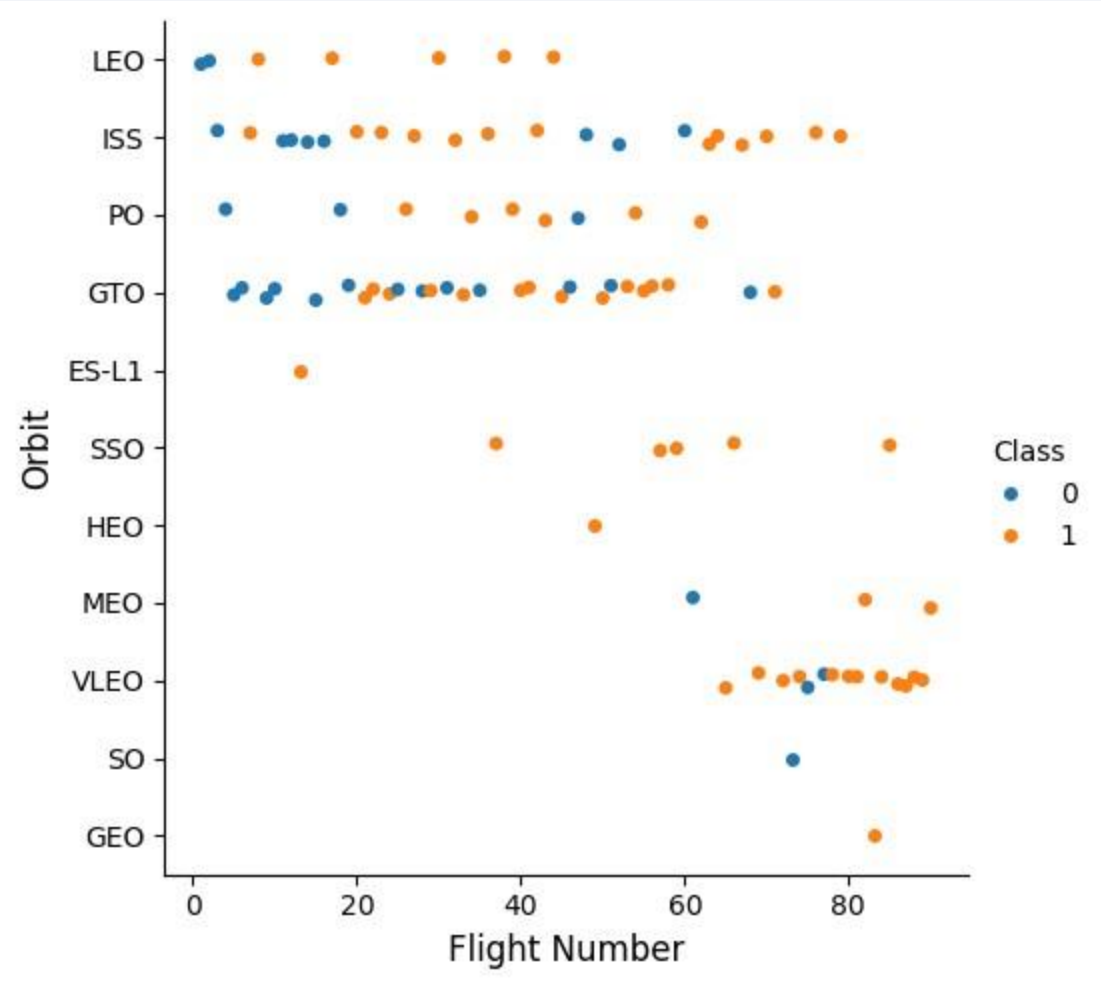
Payload vs. Launch Site



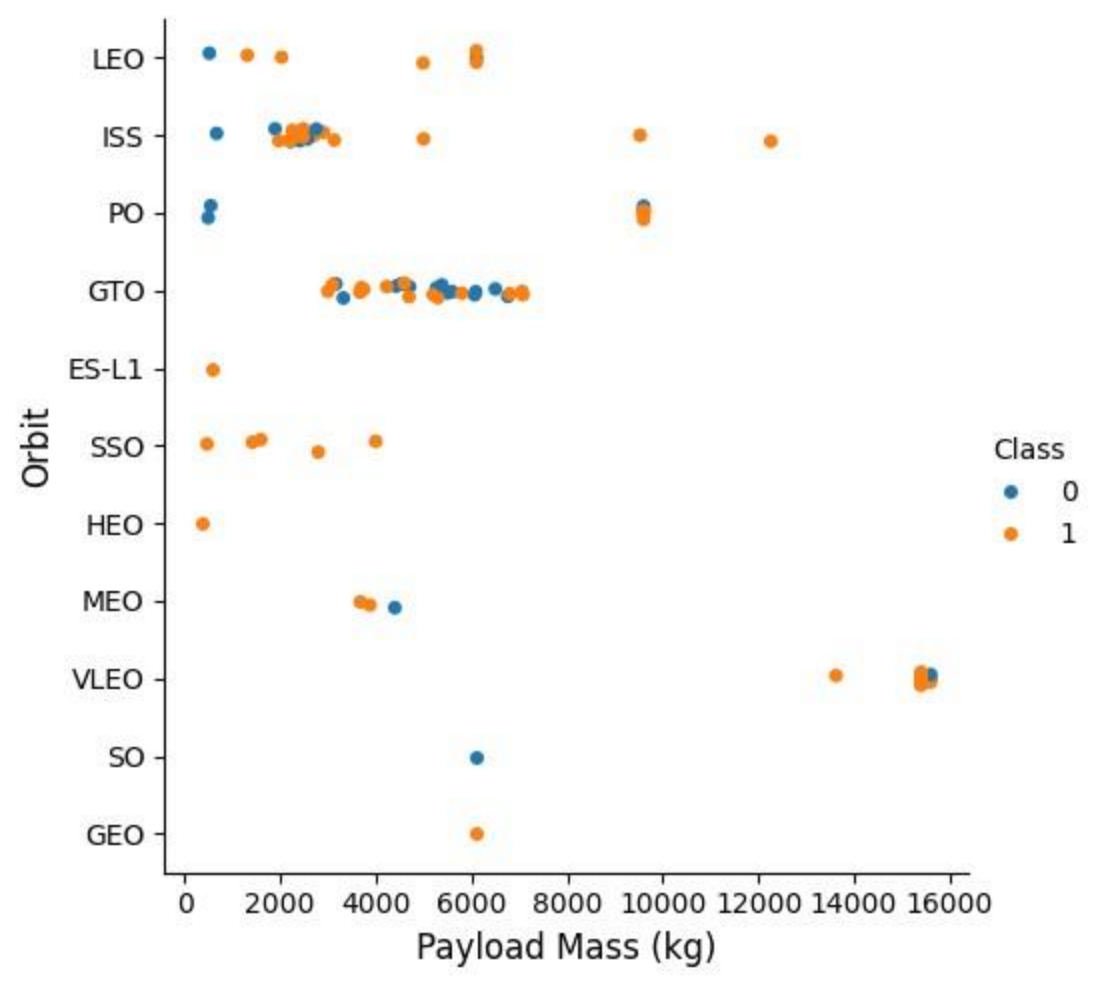
Success Rate vs. Orbit Type



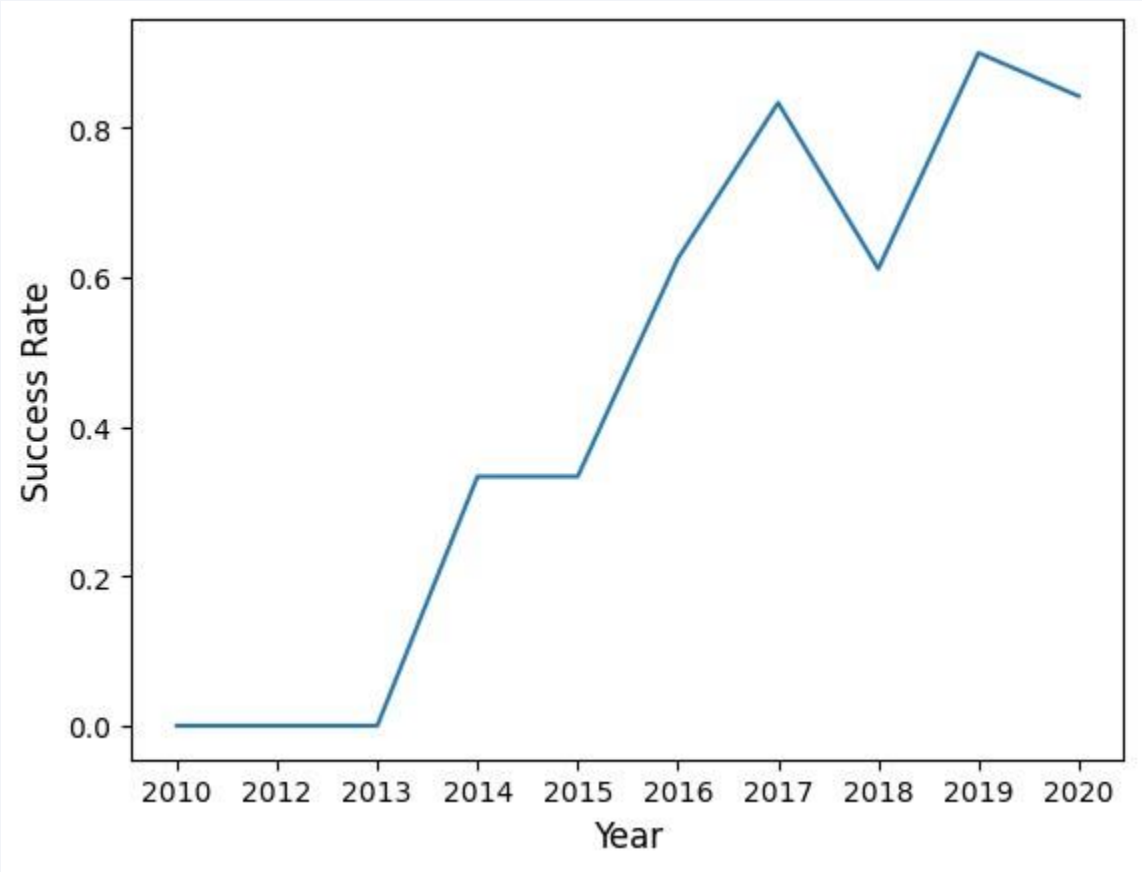
Flight Number vs. Orbit Type



Payload vs. Orbit Type



Launch Success Yearly Trend



All Launch Site Names

In [12]:
```
%sql select distinct "Launch_Site" from SPACEXTABLE
```

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

Out[12]:

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

Launch Site Names Begin with 'CCA'

In [13]:
```
%sql select * from SPACEXTABLE where "Launch_Site" like 'CCA%' limit 5
```

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

Out[13]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outc
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (paracl
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 (paracl
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No atte
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No atte
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No atte

Total Payload Mass

```
In [17]: %sql select sum("PAYLOAD_MASS__KG_") as "Total Payload Mass" from SPACEXTABLE where "Customer" is "NASA (CR
```

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

Out[17]:

Total Payload Mass
45596

Average Payload Mass by F9 v1.1

```
In [20]: : avg("PAYLOAD_MASS__KG_") as "Average Payload Mass" from SPACEXTABLE where "Booster_Version" like "F9 v1.1%"
```

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

Out[20]:

Average Payload Mass
2534.6666666666665

First Successful Ground Landing Date

```
In [27]: %sql select min(Date) from SPACEXTABLE where "Landing_Outcome" is "Success (ground pad)"

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

Out[27]:

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

In [30]:
```
%sql select "Booster_Version" from SPACEXTABLE where "Landing_Outcome" is "Success (drone ship)" and "PAYLO
```

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

Out[30]:

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

Total Number of Successful and Failure Mission Outcomes

In [34]:
```
%sql select "Mission_Outcome", count("Mission_Outcome") from SPACEXTABLE group by "Mission_Outcome"
```

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

Out[34]:

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

Boosters Carried Maximum Payload

In [35]:
```
%sql select "Booster_Version" from SPACEXTABLE where "PAYLOAD_MASS__KG_" = (Select max("PAYLOAD_MASS__KG_")
```

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

Out[35]:

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

In [40]:
```
%sql select substr(Date,6,2) as "Month Name", "Landing_Outcome", "Booster_Version", "Launch_Site" from SPAC
```

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

Out[40]:

Month Name	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

In [46]:
```
%sql select "Landing_Outcome", count("Landing_Outcome") from SPACEXTABLE where Date between '2010-06-04' an
```

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

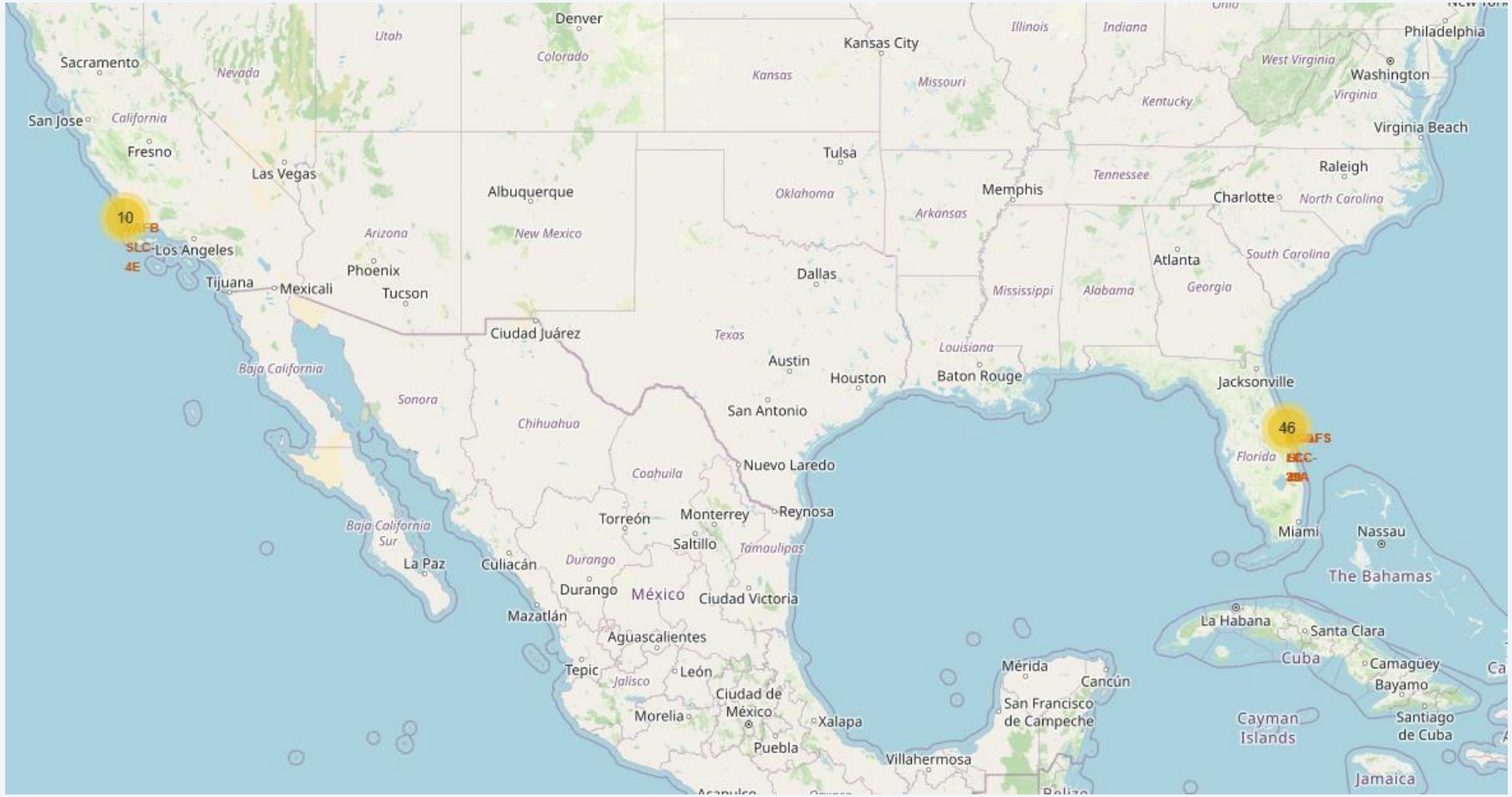
Out[46]:

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

Section 4

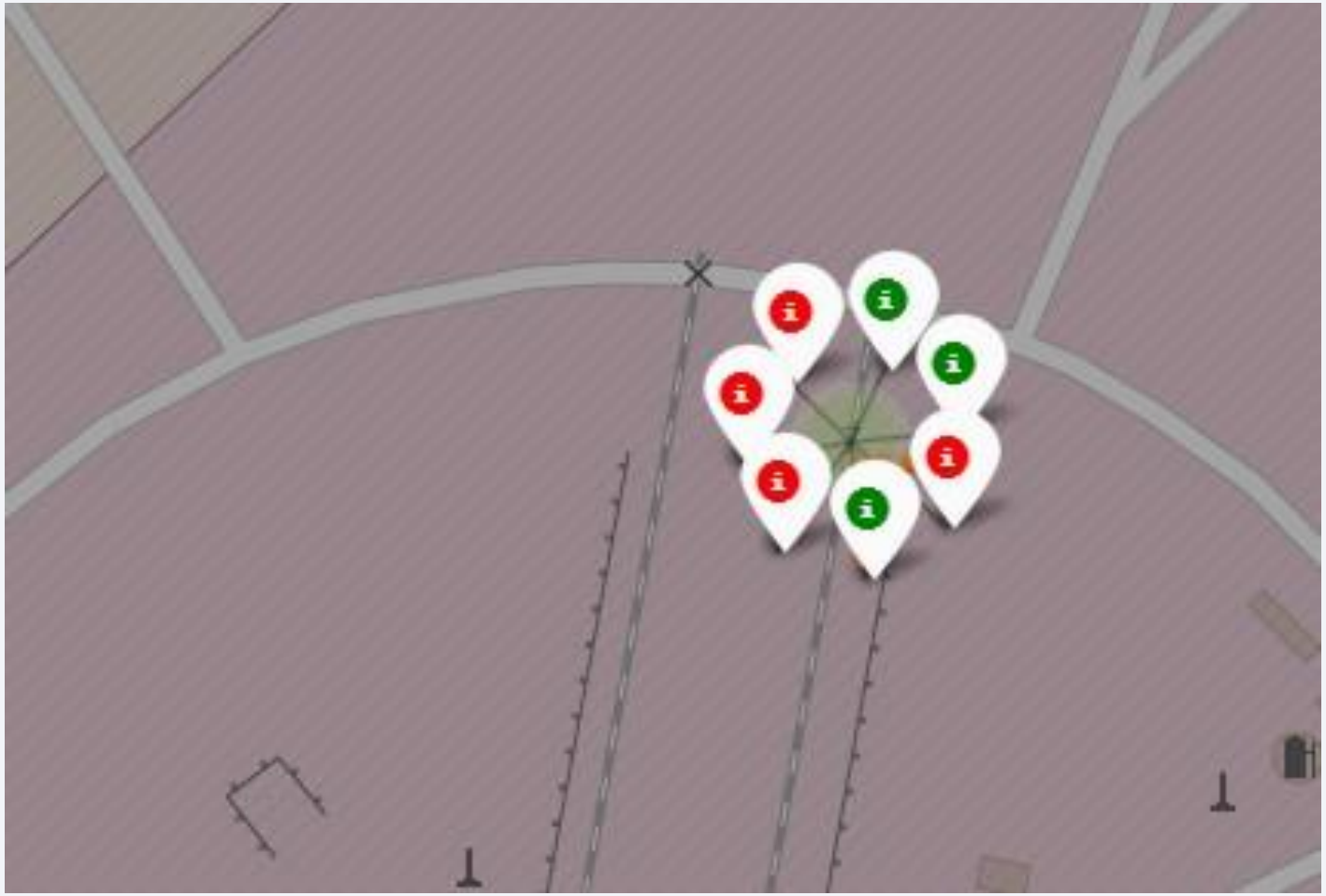
Launch Sites Proximities Analysis

All Launch Sites

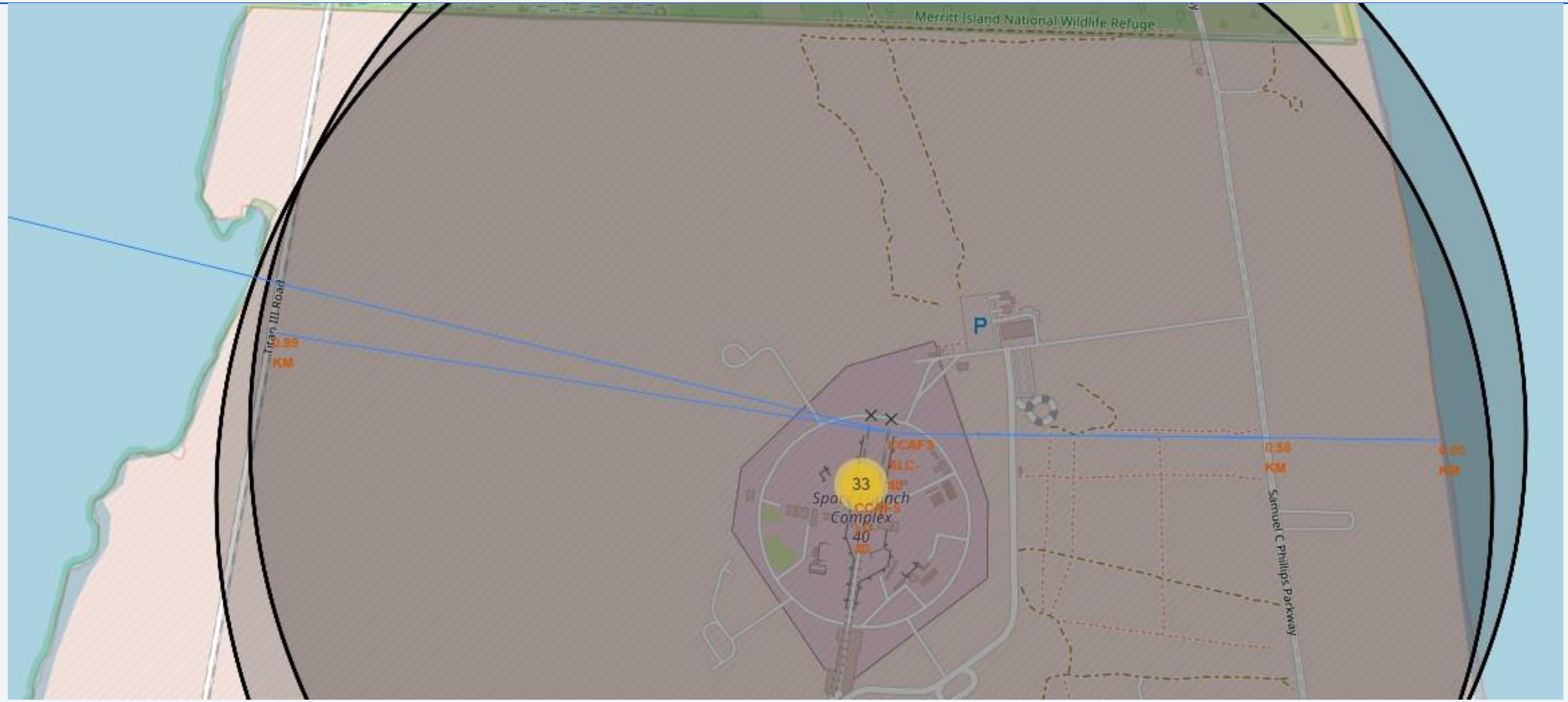


Success or Failure Launch

- Green: Success
- Red: Failure



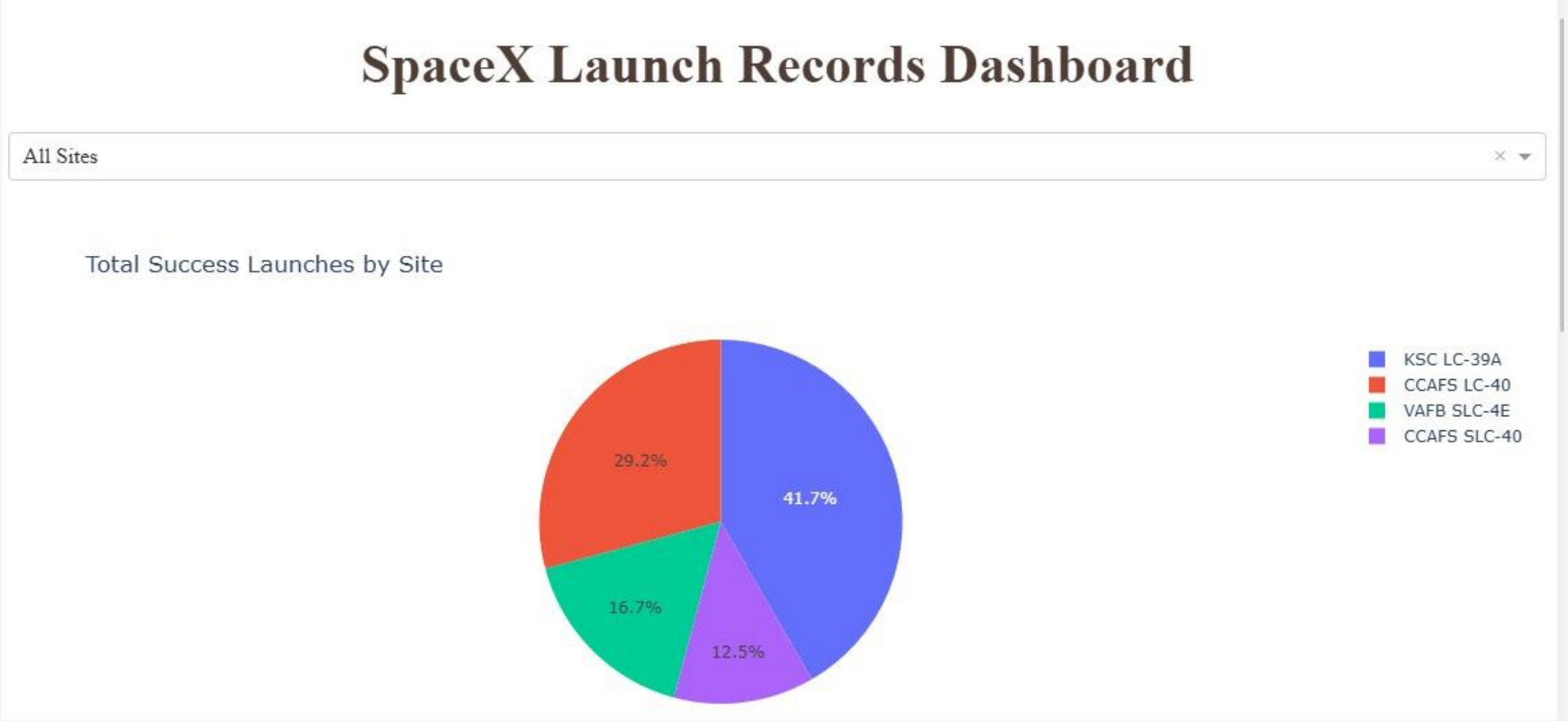
Distance from Launch Site to Coastline, Railway, Highway and City



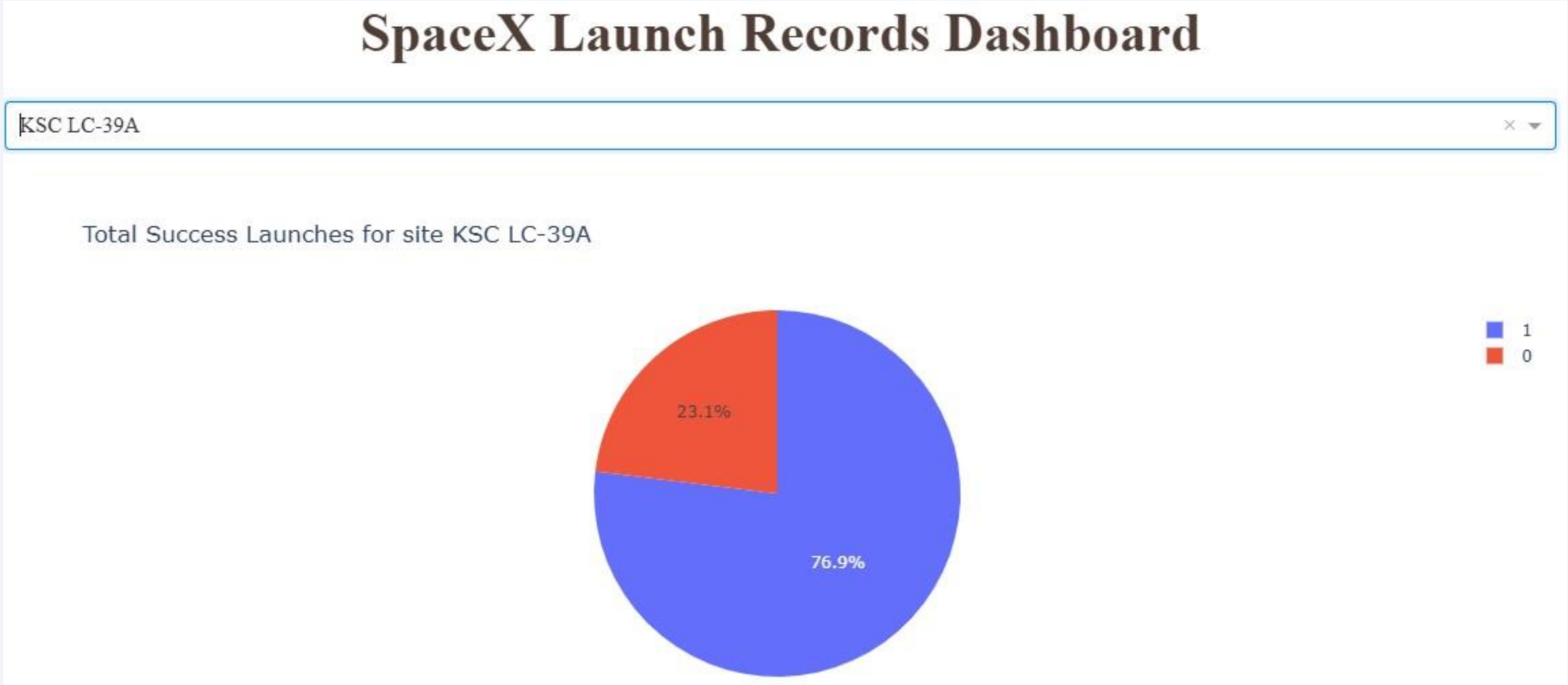
Section 5

Build a Dashboard with Plotly Dash

Dashboard: Total Success Launches by Site



Dashboard: Piechart for Launch Site with Highest Launch Success Ratio

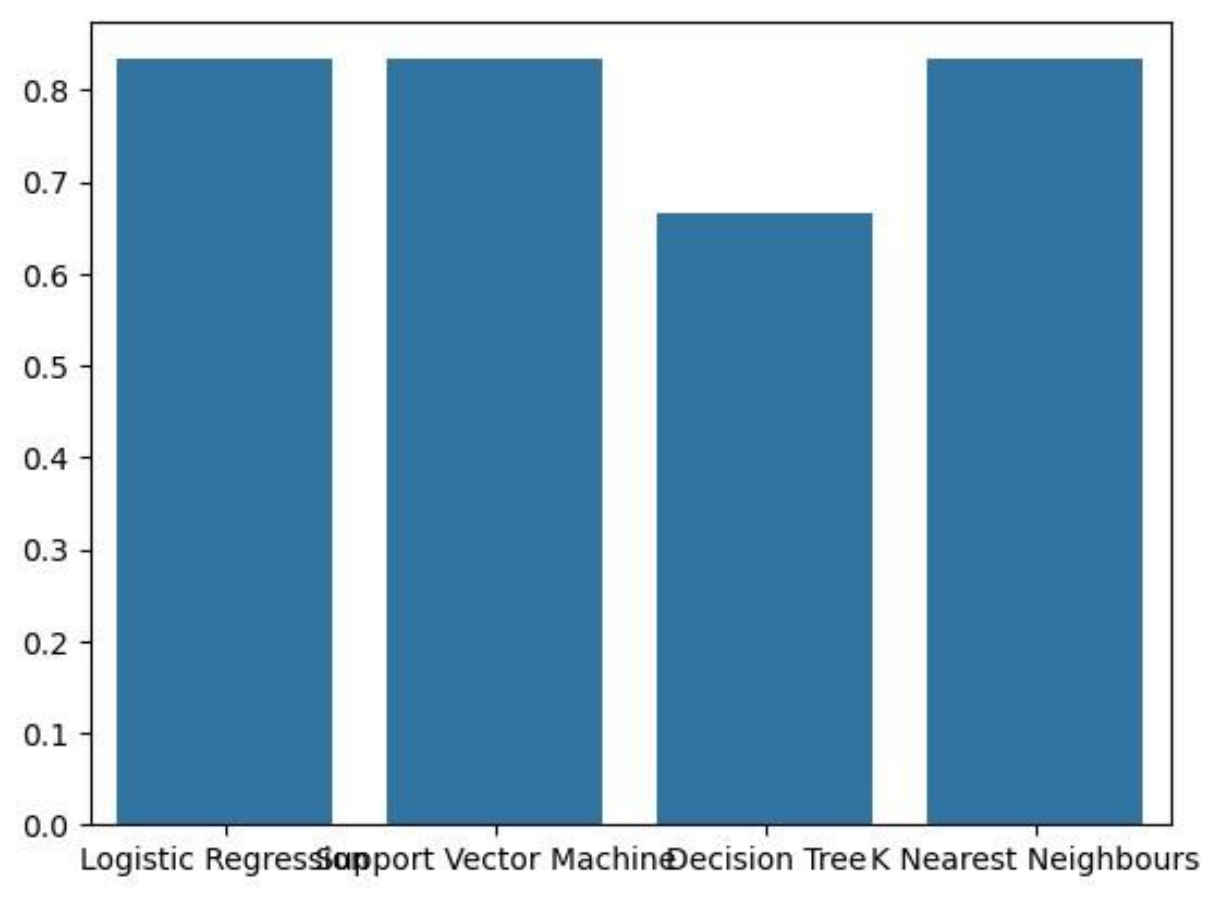


Dashboard: Payload vs Launch Outcome for sites

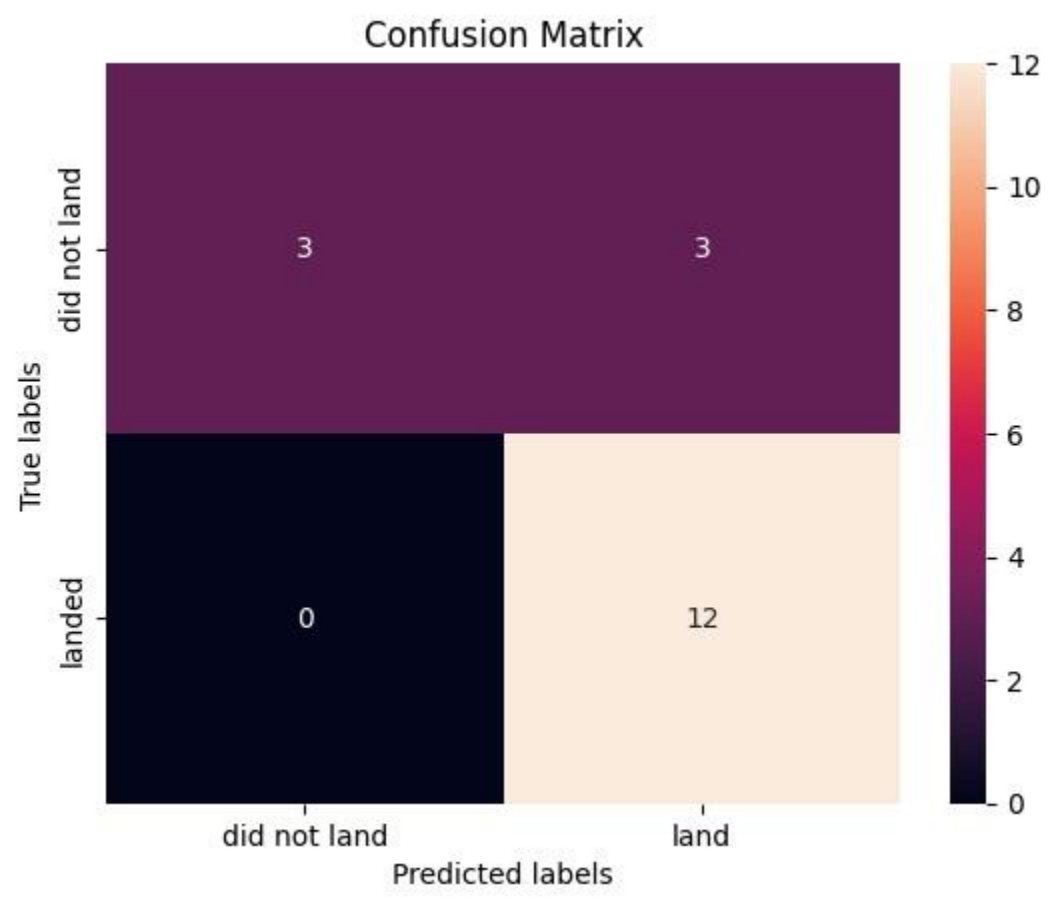
Section 6

Predictive Analysis (Classification)

Classification Accuracy



Confusion Matrix of Best performing model (Logistic Regression)



Thank you!