Winning Space Race with Data Science

Bandi Om Sai Sashank
5 April, 2024

Outline

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

Executive Summary

- Summary of methodologies
    - Utilized the SpaceX public API to get the details of all launches.
    - Used Exploratory Data Analysis(EDA) to get interactive visualization of the data.
    - Found the best hyperparameters for training LR, SVM and Decision Trees.
- Summary of all results
    - The best hyperparameters for
        - LR are C = 1 and penalty = l2 for solver lbfgs.
        - SVM are C = 1.0 and gamma = 0.03162277660168379 with sigmoid kernel.
        - Decision Tree are criterion = gini, max_depth = 2, max_features = sqrt, min_samples_leaf = 2, min_samples_split = 10, splitter = best.

Introduction

- Project background and context
  - This project is intended to collect data from SpaceX API, analyse the data, interactively visualize it.
- Problems you want to find answers
  - To find the best hyperparameters for Decision tree, SVM and Logistic Regression.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data is collected from the SpaceX public API.
- Perform data wrangling
  - Determine the important columns for training a machine learning model.
  - Convert string columns into categorical columns.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection

- Describe how data sets were collected.
  - The datasets are collected by contacting the SpaceX API.
  - The API resides at https://api.spacexdata.com
  - Various routes in the API provide various data about launches and launch sites.

	static_fire_date_utc	static_fire_date_unix	net	window	rocket	success	failures	details	crew	ships	capsules	pay
0	2006-03-17T00:00:00.000Z	1.142554e+09	False	0.0	5e9d0d95eda69955f709d1eb	False	[{'time': 33, 'altitude': None, 'reason': 'merlin engine failure'}]	Engine failure at 33 seconds and loss of vehicle	[]	[]	[]	[5eb0e4b5b6c3bb0006ee
							[{'time': 301, 'altitude':	Successful first stage burn and transition to second stage, maximum				

Data Collection – SpaceX API

- The completed Jupyter Notebook is at https://github.com/Sashank999/edX-ML-Capstone/blob/master/jupyter-labs-spacex-data-collection-api.ipynb

GET Past Launches

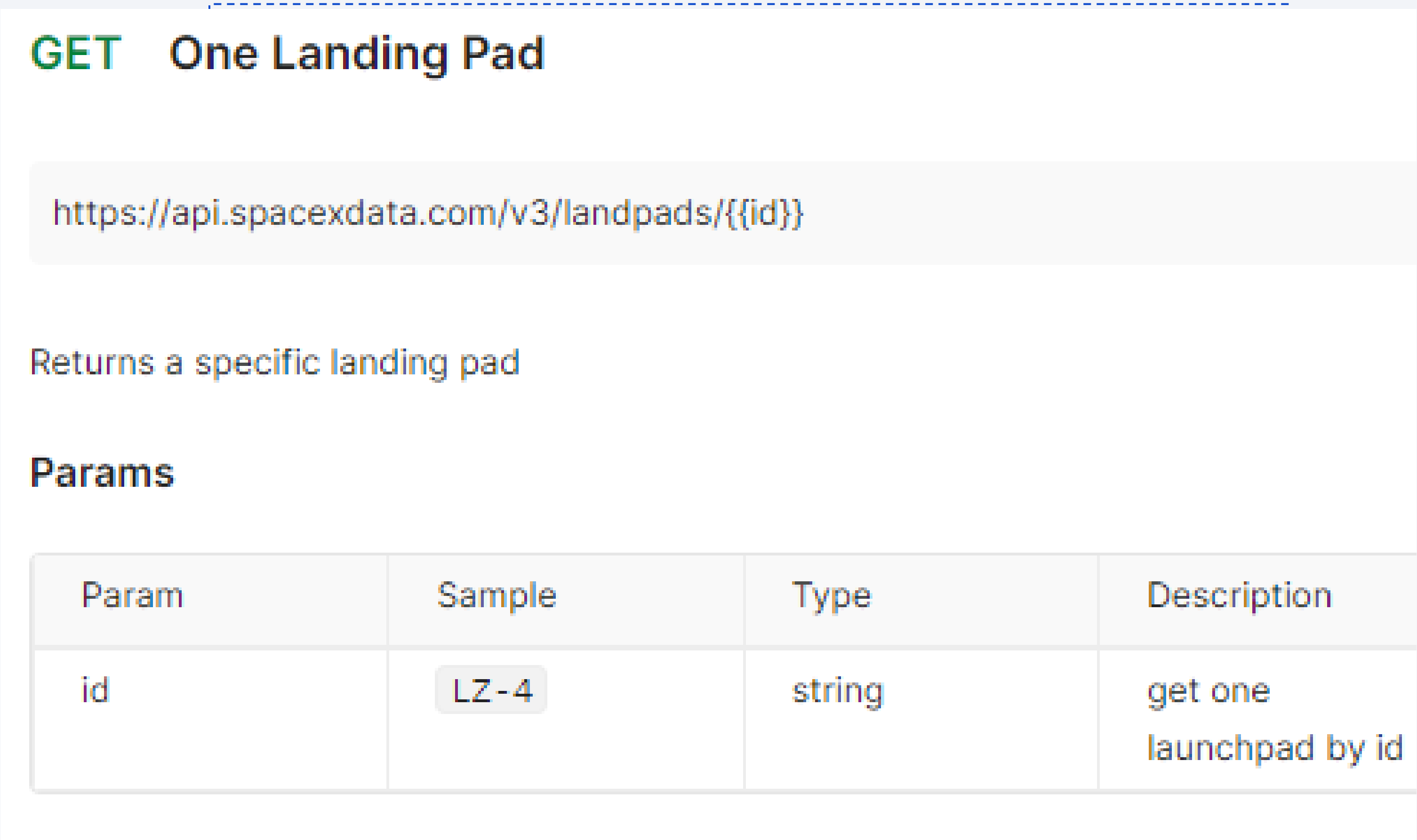
https://api.spacexdata.com/v3/launches/past

Optional Querystrings

Param	Sample	Type	Description
flight_id	`5a9fc479ab70786ba5a1eaaa`	string	Filter launches by mongo document id
start/end	`start=2017-06-22&end=2017-06-25`	valid JavaScript date format	Include both to sort by date range
flight_number	`60`	integer	Filter by flight number

Data Collection - Scraping

- The completed Jupyter Notebook is at https://github.com/Sashank999/edX-ML-Capstone/blob/master/jupyter-labs-spacex-data-collection-api.ipynb



GET One Landing Pad

https://api.spacexdata.com/v3/landpads/{{id}}

Returns a specific landing pad

Params

Param	Sample	Type	Description
id	LZ-4	string	get one launchpad by id

Data Wrangling

- The data is first analyzed to create a new column that displays whether the specific launch was successful or not.
- Next, all the value counts of columns Orbit, LaunchSite and Outcome are analysed.
- The completed Jupyter Notebook is at https://github.com/Sashank999/edX-ML-Capstone/blob/master/labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

EDA with Data Visualization

- The graphs plotted are:
  - Scatter plot with Flight Number and Launch Site
  - Scatter plot with Payload Mass and Launch Site
  - Bar plot with Orbit and Success Rate
  - Scatter plot with Flight Number and Orbit
  - Scatter plot with Payload Mass and Orbit
  - Line plot of Yearly Success Rate of launches
- The GitHub URL of your completed EDA with data visualization notebook is https://github.com/Sashank999/edX-ML-Capstone/blob/master/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- The SQL queries performed are:
    - select distinct "Launch_Site" from SPACEXTABLE
    - select * from SPACEXTABLE where "Launch_Site" like 'KSC%' limit 5
    - select SUM("PAYLOAD_MASS__KG_") from SPACEXTABLE where "Customer" like '%NASA%‘
    - select avg("PAYLOAD_MASS__KG_") from SPACEXTABLE where "Booster_Version" = 'F9 v1.1‘
    - select min("Date") from SPACEXTABLE where "Landing_Outcome" like '%Success%‘
    - select distinct "Booster_Version" from SPACEXTABLE where "PAYLOAD_MASS__KG_" > 4000 and "PAYLOAD_MASS__KG_" < 6000 and "Landing_Outcome" = 'Success (ground pad)'
- The GitHub URL of your completed EDA with SQL notebook is https://github.com/Sashank999/edX-ML-Capstone/blob/master/jupyter-labs-eda-sql-edx_sqllite.ipynb

Build an Interactive Map with Folium

- All launch sites in the given data are marked on the data to get an overview of the launch sites.
- Marked the sites where successful launches happened.
- Marked the sites where failed launches happened.
- The GitHub URL of your completed interactive map with Folium map is at https://github.com/Sashank999/edX-ML-Capstone/blob/master/lab_jupyter_launch_site_location.jupyterlite.ipynb

Predictive Analysis (Classification)

- The best parameters for models are calcualted by using the GridSearchCV class provided by the scikit-learn Python library.
- Every model is trained with an exhaustive list of hyperparameters to find the best parameter values.
- The GitHub URL of your completed predictive analysis lab is at https://github.com/Sashank999/edX-ML-Capstone/blob/master/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

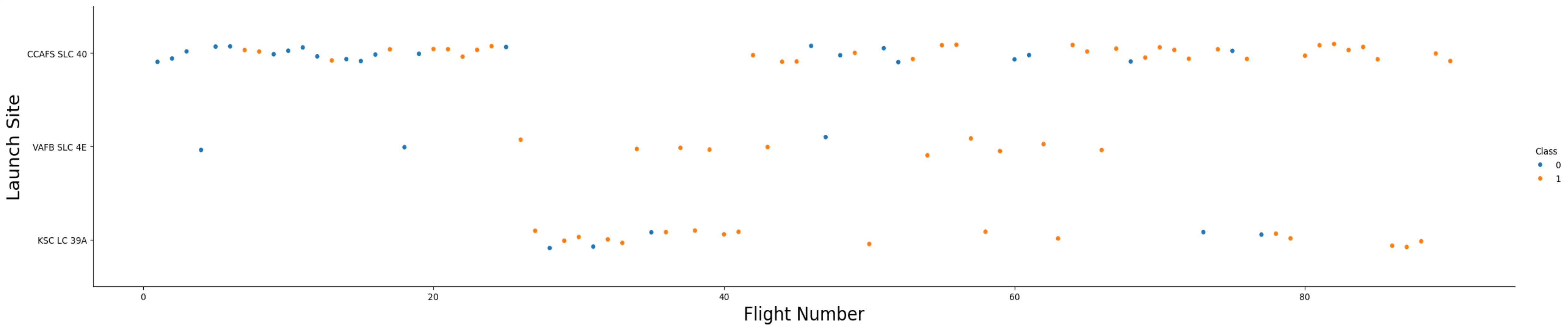
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

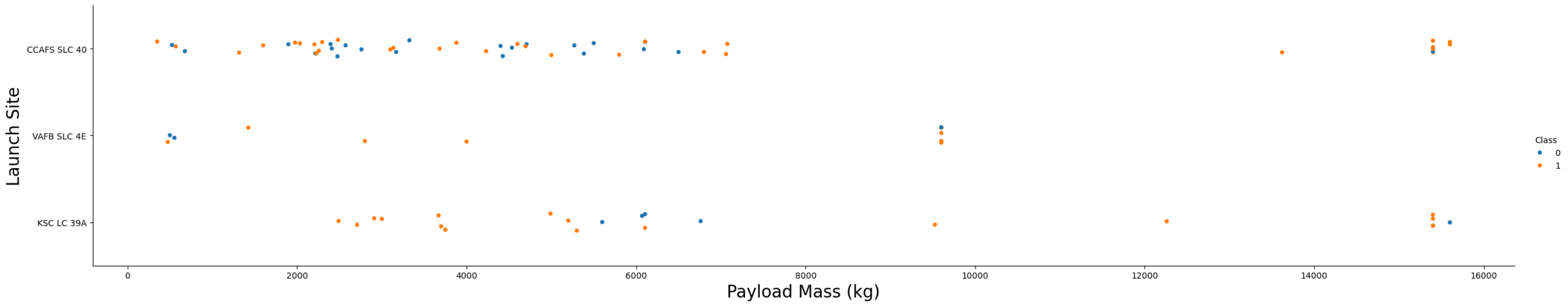
Flight Number vs. Launch Site

- Scatter plot of Flight Number vs. Launch Site



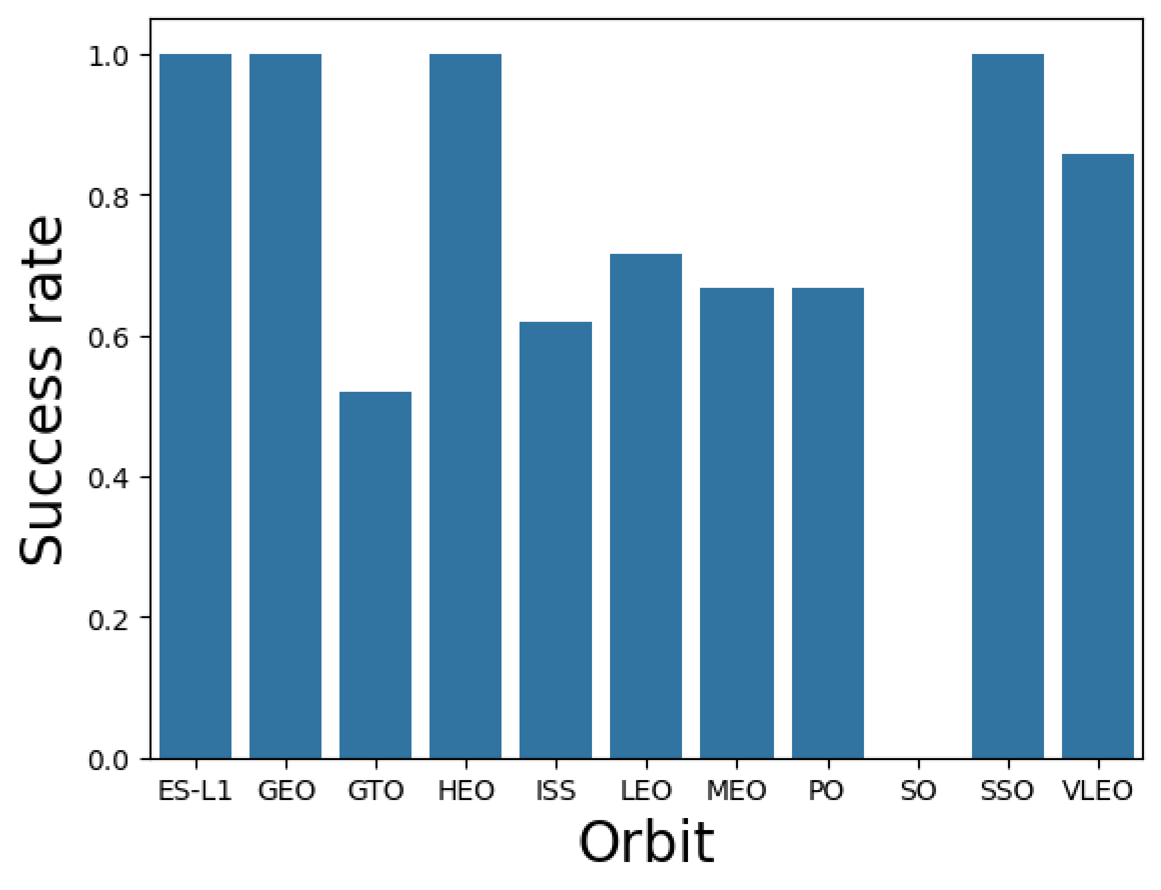
Payload vs. Launch Site

- Scatter plot of Payload vs. Launch Site



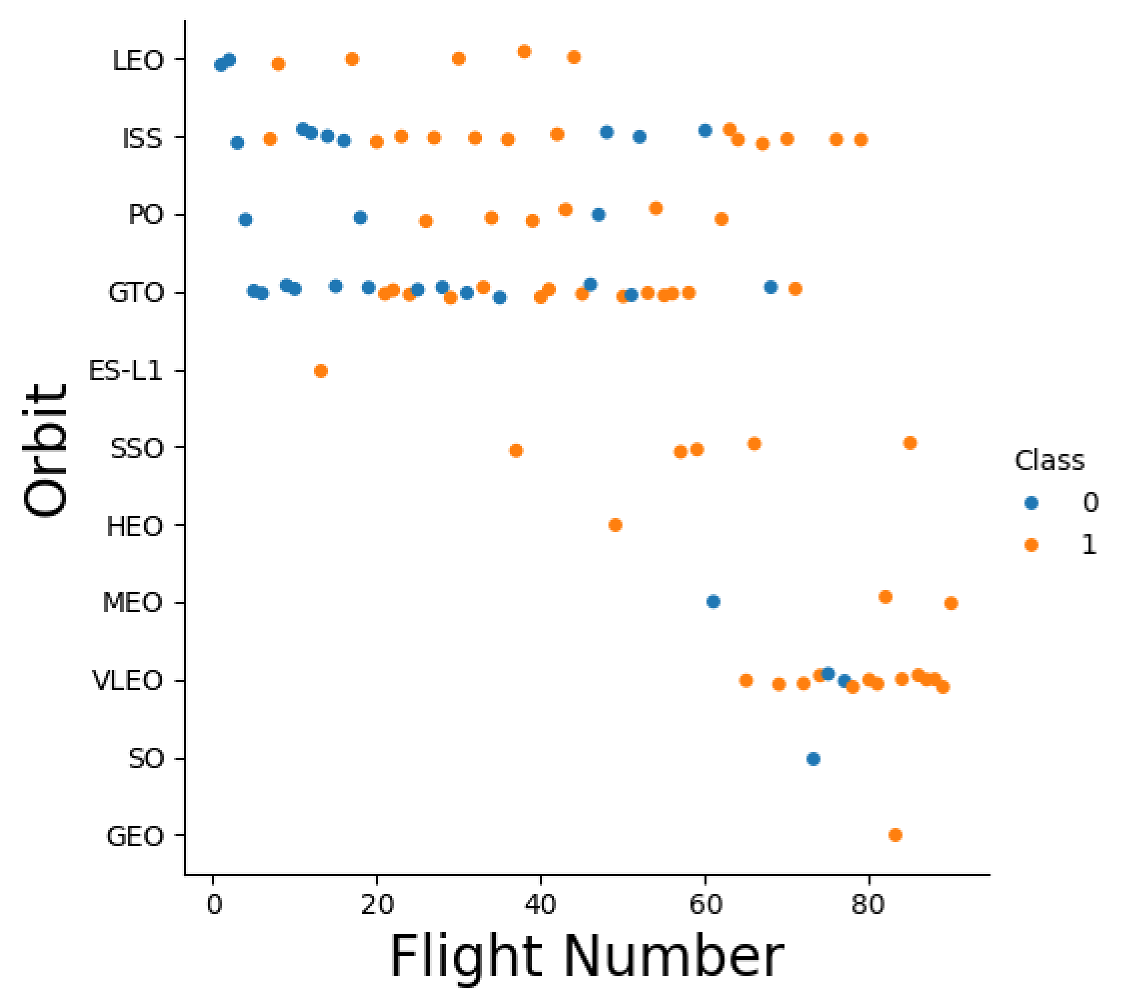
Success Rate vs. Orbit Type

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



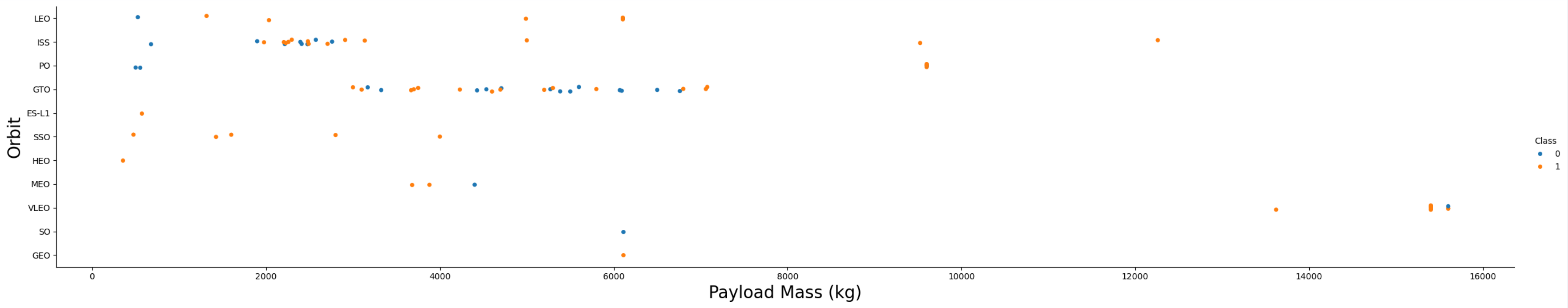
Flight Number vs. Orbit Type

- Scatter plot of Flight number vs. Orbit type



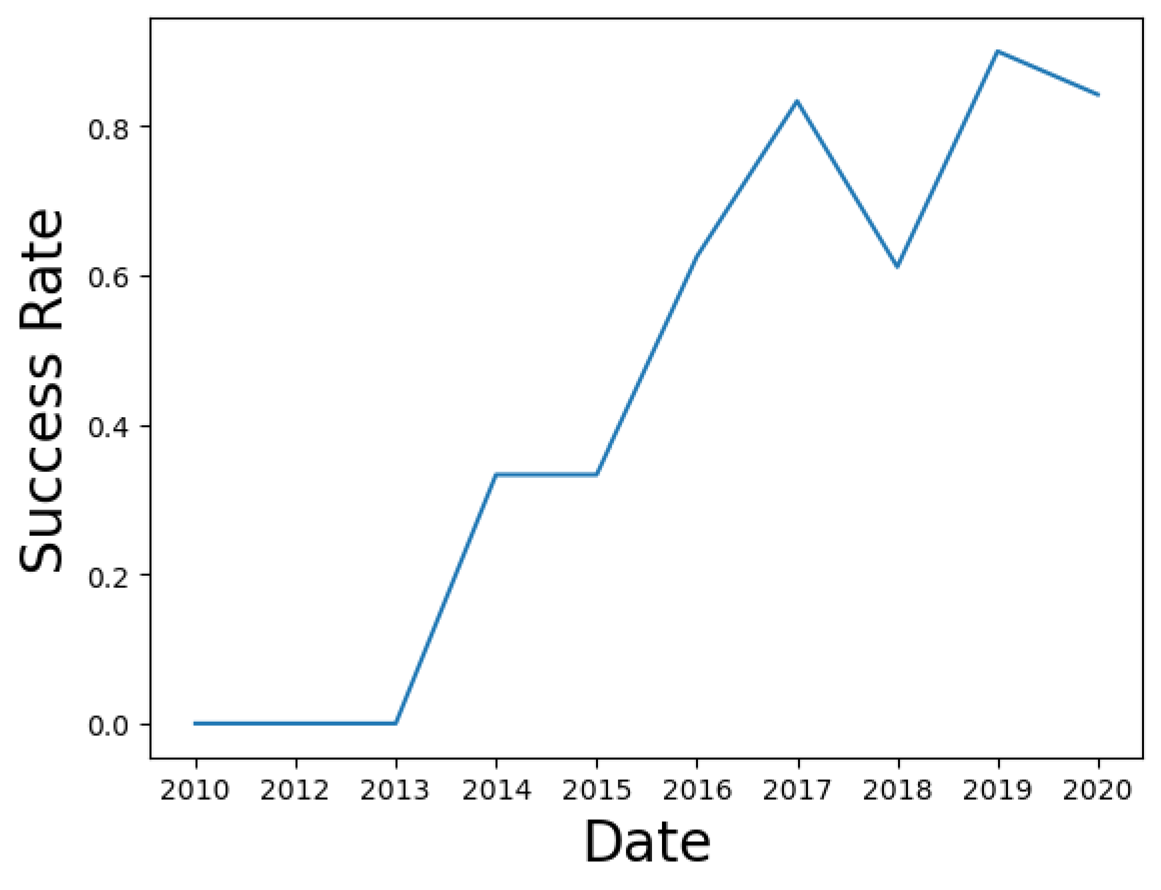
Payload vs. Orbit Type

- Show a scatter point of payload vs. orbit type



Launch Success Yearly Trend

- Show a line chart of yearly average success rate



All Launch Site Names

- Names of the unique launch sites are:
  - CCAFS LC-40
  - VAFB SLC-4E
  - KSC LC-39A
  - CCAFS SLC-40
- Query:
  - select distinct "Launch_Site" from SPACEXTABLE

Launch Site Names Begin with 'KSC'

- 5 records where launch sites’ names start with `KSC`

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-03-16	6:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-05-01	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
2017-05-15	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

- Query:
  - select * from SPACEXTABLE where "Launch_Site" like 'KSC%' limit 5

Total Payload Mass

- The total payload carried by boosters from NASA
    - 107010 kg
- Query:
    - select SUM("PAYLOAD_MASS__KG_") from SPACEXTABLE where "Customer" like '%NASA%'

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1
  - 2928.4
- Query:
  - select avg("PAYLOAD_MASS__KG_") from SPACEXTABLE where "Booster_Version" = 'F9 v1.1'

First Successful Ground Landing Date

- The dates of the first successful landing outcome on drone ship.

```
%sql select min("Date") from SPACEXTABLE where "Landing_Outcome" like '%Success%'
```

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

min("Date")
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 B1032.1
F9 B4 B1040.1
F9 B4 B1043.1

- Query:
  - select distinct "Booster_Version" from SPACEXTABLE where "PAYLOAD_MASS__KG_" > 4000 and "PAYLOAD_MASS__KG_" < 6000 and "Landing_Outcome" = 'Success (ground pad)'

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes

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

- Query:
  - select "Mission_Outcome", count("Mission_Outcome") from SPACEXTABLE group by "Mission_Outcome"

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass
- Query:
  - select distinct "Booster_Version" from SPACEXTABLE where "PAYLOAD_MASS__KG_" = (select max("PAYLOAD_MASS__KG_") from SPACEXTABLE)

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

2017 Launch Records

- The records which will display the month names, succesful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017

Month_Name	Landing_Outcome	Booster_Version	Launch_Site
02	Success (ground pad)	F9 FT B1031.1	KSC LC-39A
05	Success (ground pad)	F9 FT B1032.1	KSC LC-39A
06	Success (ground pad)	F9 FT B1035.1	KSC LC-39A
08	Success (ground pad)	F9 B4 B1039.1	KSC LC-39A
09	Success (ground pad)	F9 B4 B1040.1	KSC LC-39A
12	Success (ground pad)	F9 FT B1035.2	CCAFS SLC-40

- Query:
  - %sql select STRFTIME('%m', "Date") as "Month_Name", "Landing_Outcome", "Booster_Version", "Launch_Site" from SPACEXTABLE \ where substr("Date", 0, 5) = '2017' and "Landing_Outcome" = 'Success (ground pad)'

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

- The counts 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

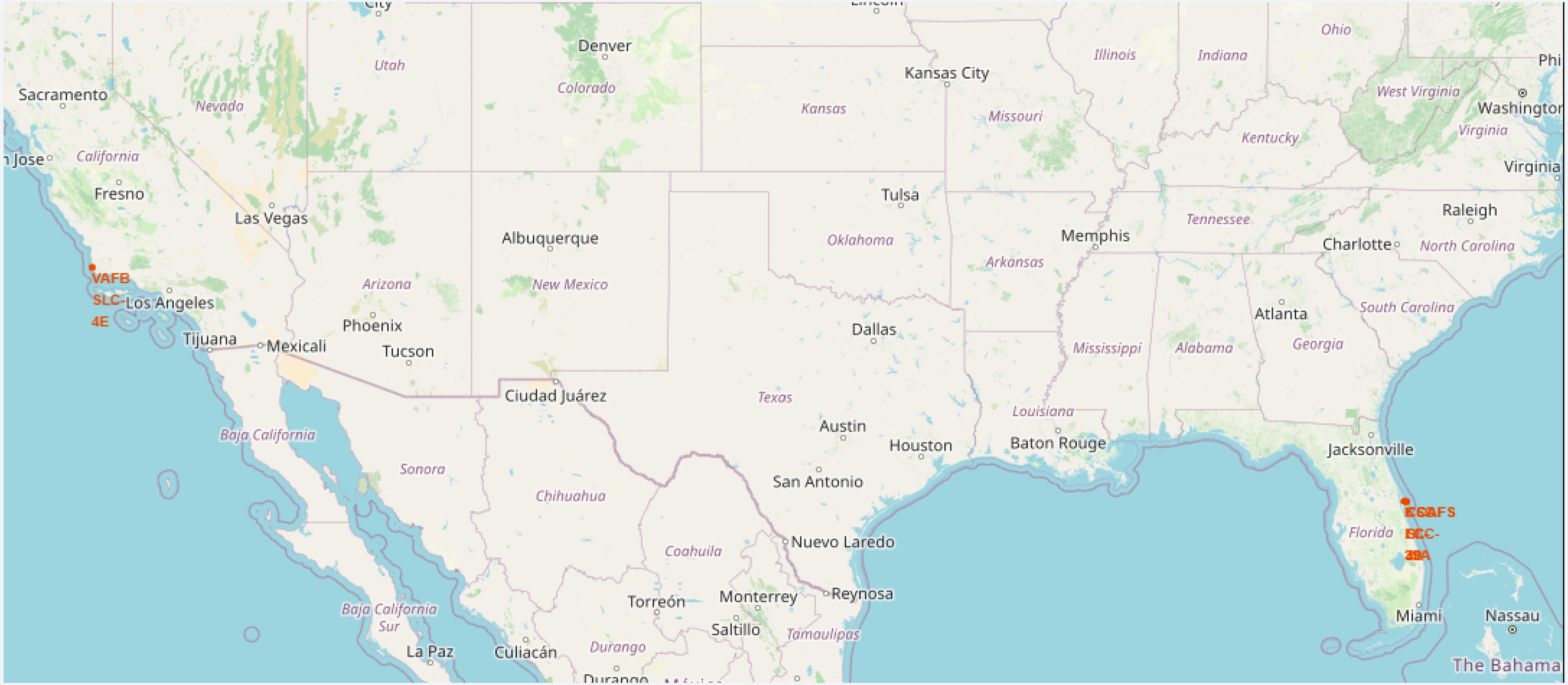
Landing_Outcome	C
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

- Query:
  - select "Landing_Outcome", count("Landing_Outcome") as "C" from SPACEXTABLE\
    where "Date" >= '2010-06-04' and "Date" <= '2017-03-20' \
    group by "Landing_Outcome" \ order by "C" DESC

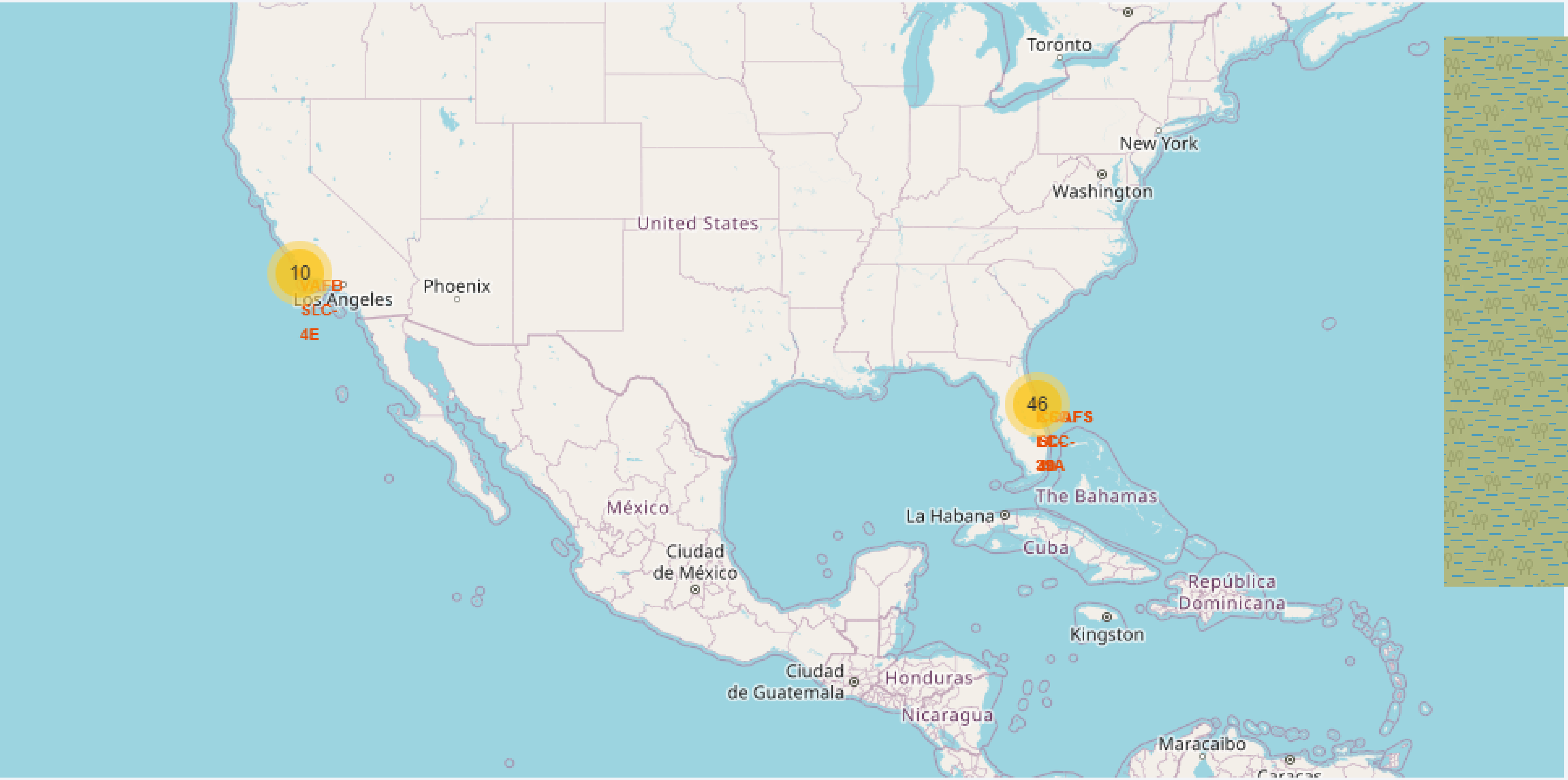
Section 3

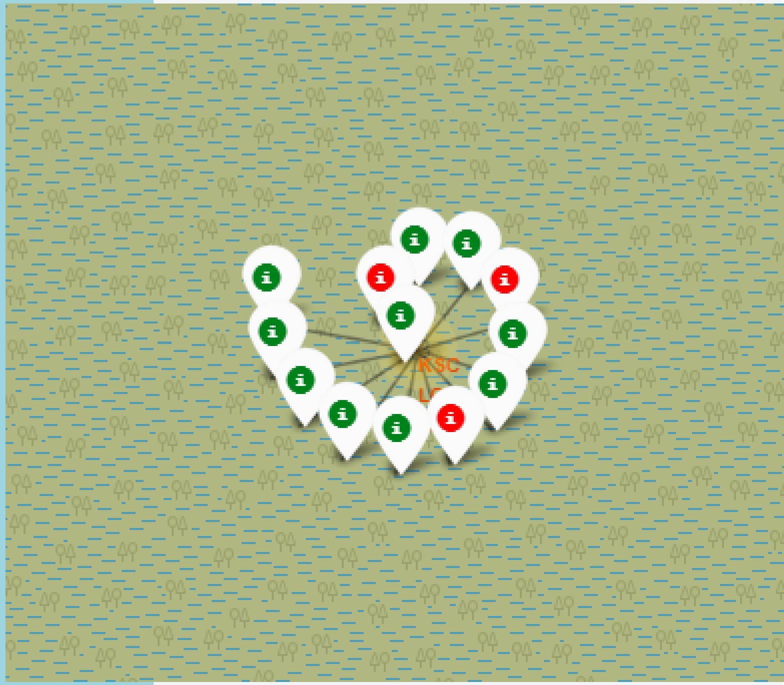
Launch Sites Proximities Analysis

Map of All Launch Sites



Launch Outcomes





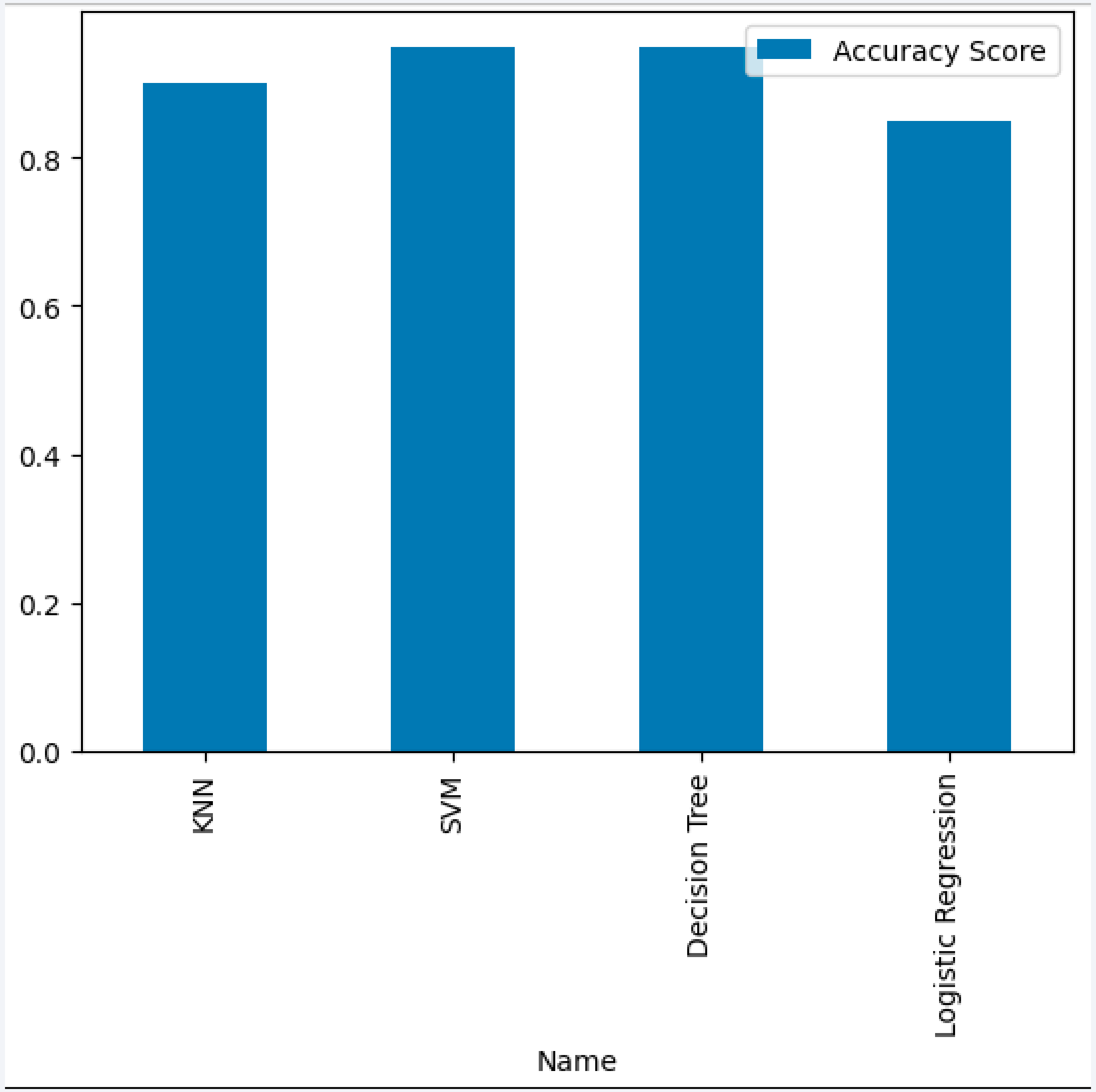
Launch Site Proximities

Section 5

Predictive Analysis (Classification)

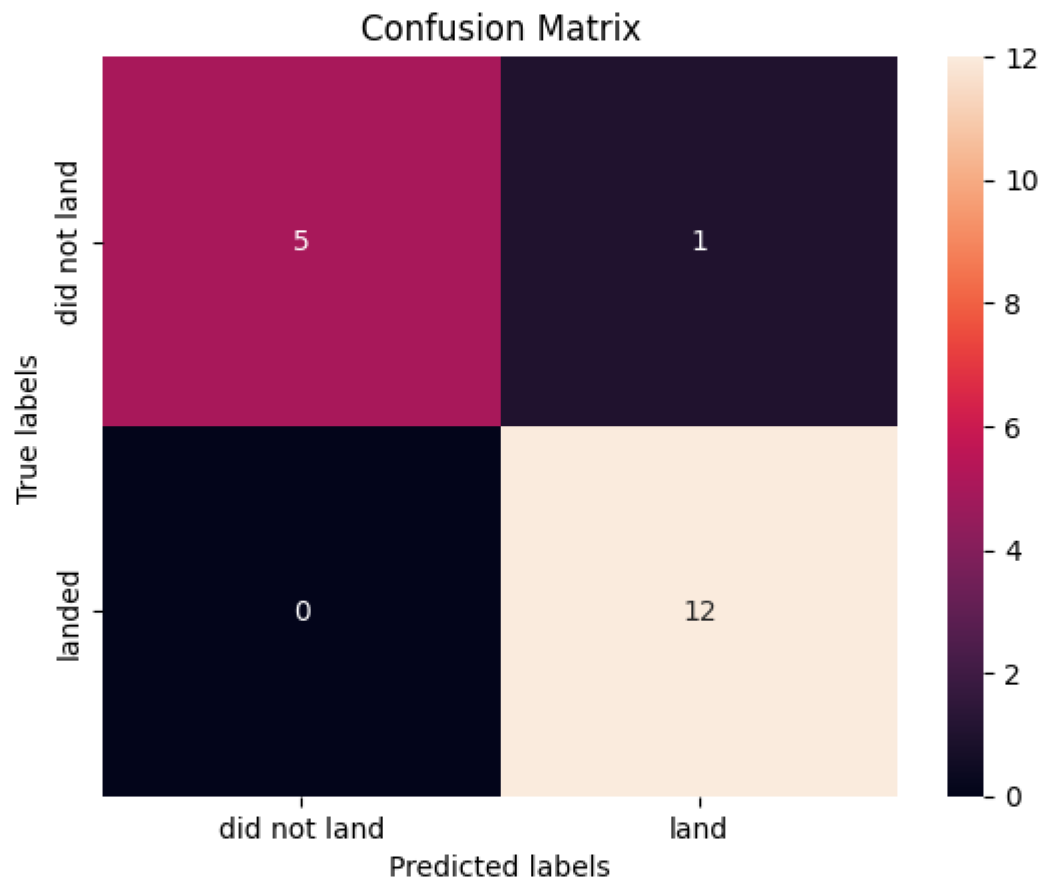
Classification Accuracy

- The most accurate models are SVM and Decision tree.

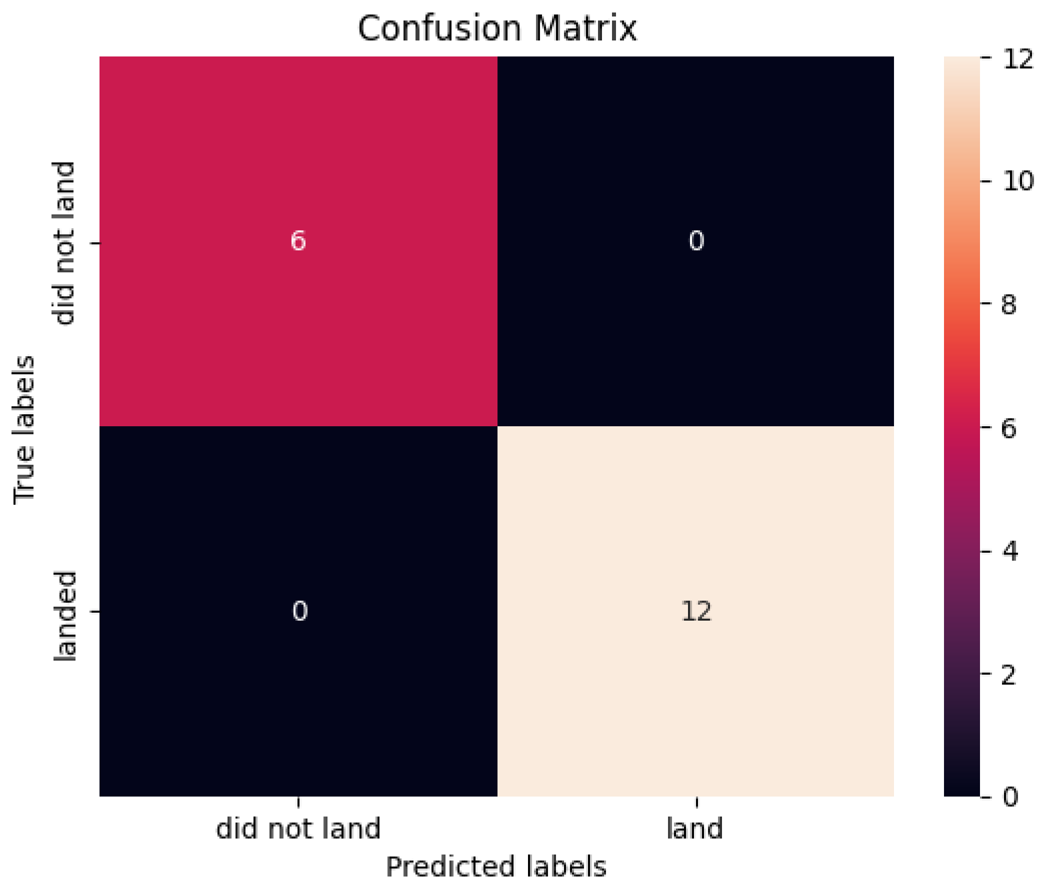


Confusion Matrix

SVM Confusion Matrix



Decision Tree Confusion Matrix



Thank you!