IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Kua Chin Hock
29-December-2021

Outline

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

Executive Summary

Summary of Methodologies

- Raw data obtained from SpaceX API.
- Data is preprocessed to remove null values, and unpivot with one hot encoding.
- Dashboard is built to visualize the correlations.
- Folium map is used to visualize the launch sites on map.
- Classification methods like KNN, decision tree, SVM, logistic regression were used.

Summary of Results

- Success rate has been high after year 2017.
- Optimum payload is between 2,000 kg to 4,000 kg.
- All launch sites are near to coastline, highway or railway.

Introduction

Project background and context

- We like to predict if Falcon 9 first stage will land successfully.
- If first stage rocket can be reused, it represents significant cost savings.

Problems to find answers

- We like to validate the statement above using data science method.
- This information allows us to bid against SpaceX.

Section 1

Methodology

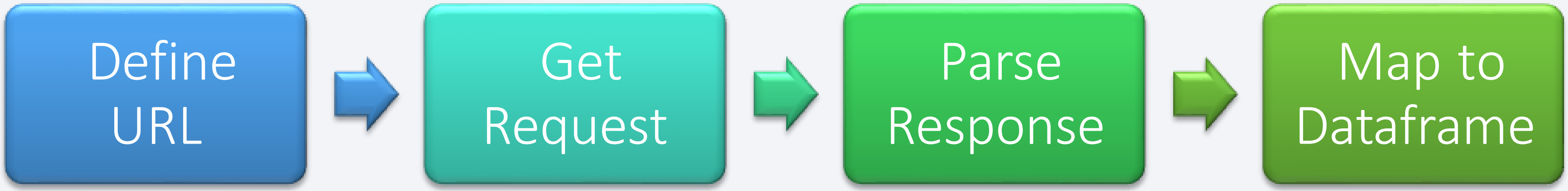
Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from https://api.spacexdata.com.
- Perform data wrangling
  - Data was processed by removing null values and transformed using one hot encoding.
- 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 models was build using grid search approach to find the optimum parameters on logistic regression, support vector machine, decision tree, k nearest neighbor. Models are evaluated using the accuracy score.

Data Collection

- The data sets were collected from:
  - https://api.spacexdata.com/v4/launches/past
  - https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- There are four major steps in the data collection process, namely define URL, perform get request, parse response, and mapping result to dataframe.



Data Collection – SpaceX API

GitHub URL @ [Lab 1 - Collecting the data.ipynb]

Flowchart of SpaceX API Calls

REST API Request

- Perform get request to SpaceX URL **https://api.spacexdata.com/v4/launches/past**.

Status Code

- Ensure response status code is 200.

Map to Pandas Dataframe

- Use json_normalize() to map JSON result to Python Pandas dataframe.

Check Data

- Inspect data quality

Filter Data

- Filter Falcon 9 launches only.
- Replace missing values with average.

Data Collection - Scraping

GitHub URL @ Lab 2 - Data Collection with Web Scraping.ipynb

Flowchart of Web Scraping

HTTP Get Request

- Download HTML using get request from URL https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches.

Parse HTML

- Parse the HTML result using BeautifulSoup.

Ensure Correct Data

- Print page title to ensure correct data.

Extract Table

- Extract the data from the Falcon 9 table.
- Iterate rows to get data.

Data Wrangling

GitHub URL @ [Lab 3 - Data Wrangling.ipynb]

Read Data

- Read raw data into Pandas Dataframe.
- Check data types and null count.

Identify Launch Site

- Find unique launch site in the data.

Occurrence of Orbit

- Calculate occurrence of each orbit type.

Create landing outcome label

- Label landing outcome as success or failure.
- Add the result to "Class" column.

EDA with Data Visualization

GitHub URL @ Lab 5 - Exploratory Visualization.ipynb

Orbit – Class Bar Chart

- To analyze which orbits have high success rate.

Flight Number – Launch Site Scatter Plot

- To analyze which flight numbers and launch site have high success rate.

Payload Mass (kg) - Orbit Scatterplot

- To analyze payload mass and orbit on success rate.

Year – Avg Class Lineplot

- To analyze relationship on success rate by year.

EDA with SQL

GitHub URL @ [Lab 4 - Exploratory Data Analysis.ipynb]

Find names of unique launch sites.

- select distinct(LAUNCH_SITE) from SPACEXDATASET

Find launch sites being with 'CCA'.

- select * from SPACEXDATASET where LAUNCH_SITE like 'CCA%' limit 5;

Find total payload mass carried by NASA (CRS).

- select sum(PAYLOAD_MASS__KG_) from SPACEXDATASET where CUSTOMER = 'NASA (CRS)';

Find booster versions with maximum payload mass.

- **select distinct**(BOOSTER_VERSION) **from** SPACEXDATASET where PAYLOAD_MASS__KG_ **=** (**select max**(PAYLOAD_MASS__KG_) **from** SPACEXDATASET);

Build an Interactive Map with Folium

GitHub URL @ [Lab 6 - Data Visualization with Folium.ipynb]

- 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

Build a Dashboard with Plotly Dash

GitHub URL @ [Lab 7 - Dash App.py]

- Summarize what plots/graphs and interactions you have added to a dashboard
- Explain why you added those plots and interactions
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose

Predictive Analysis (Classification)

GitHub URL @ [Lab 8 - Machine Learning Prediction.ipynb]

- 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

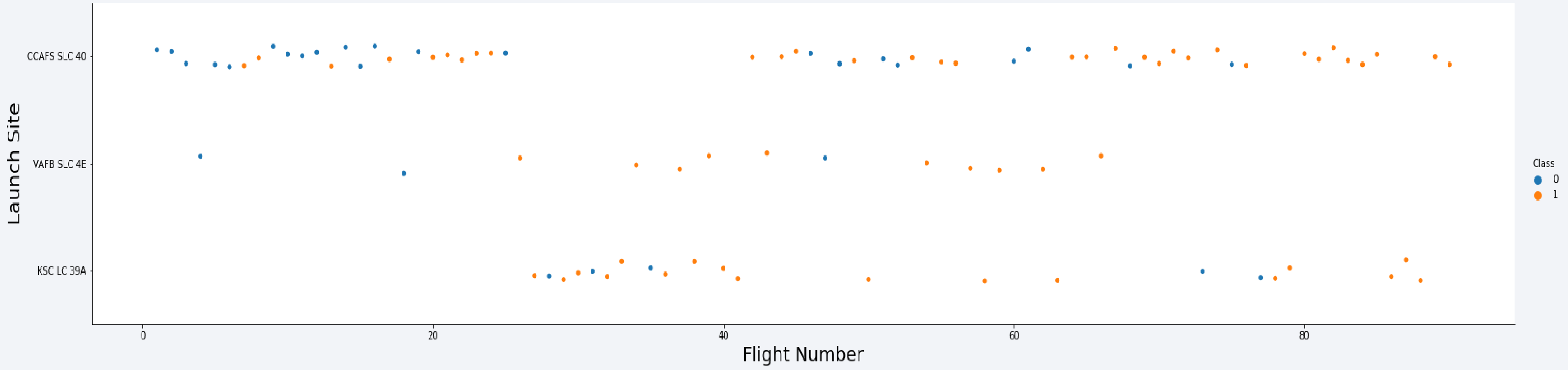
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

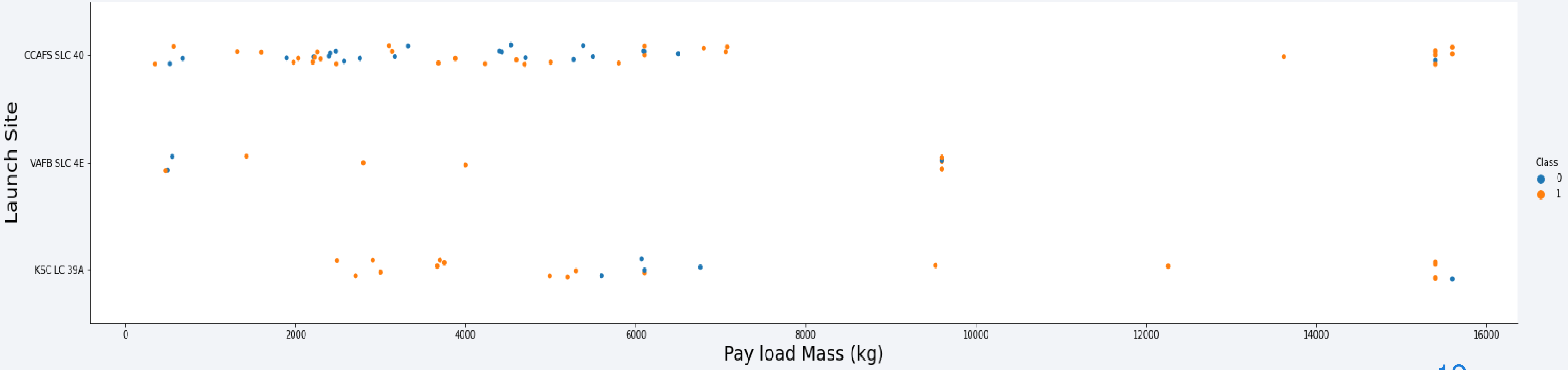
Flight Number vs. Launch Site

- Launch Site CCAFS SLC 40 has the most flight numbers.
- Success rate increases with the flight number for all launch sites.



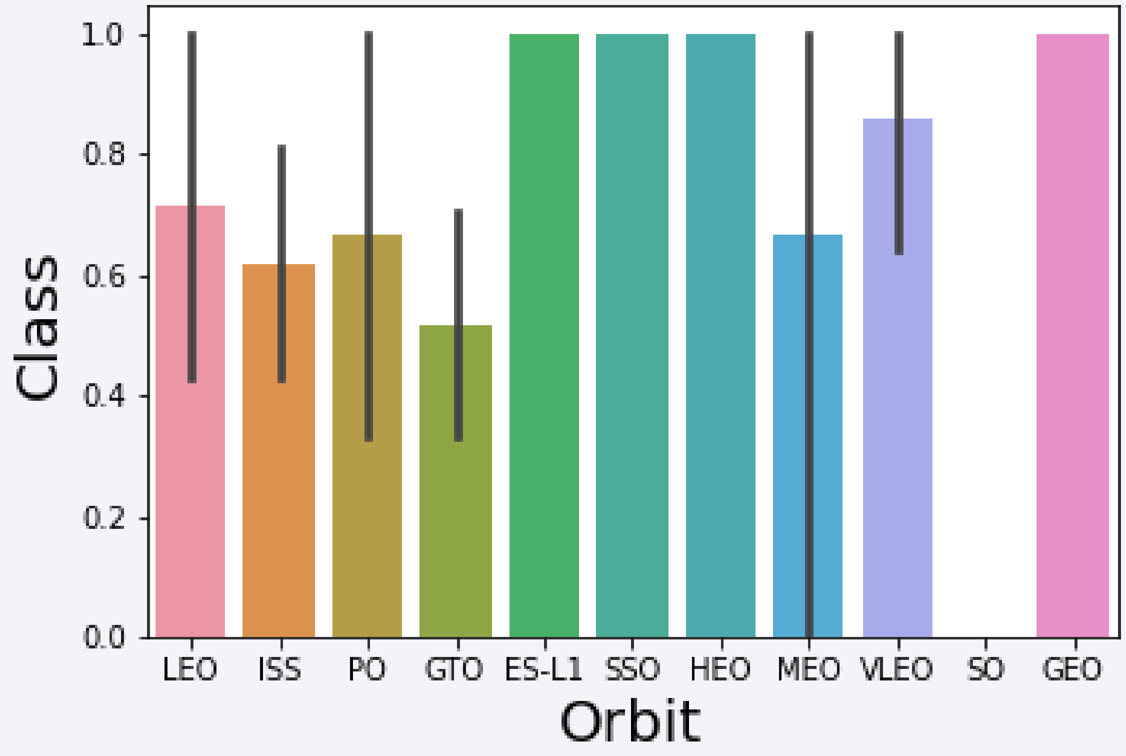
Payload vs. Launch Site

- Most payload masses are below 8000 kg.
- Higher success rate is observed on lower payload compared to higher payload.



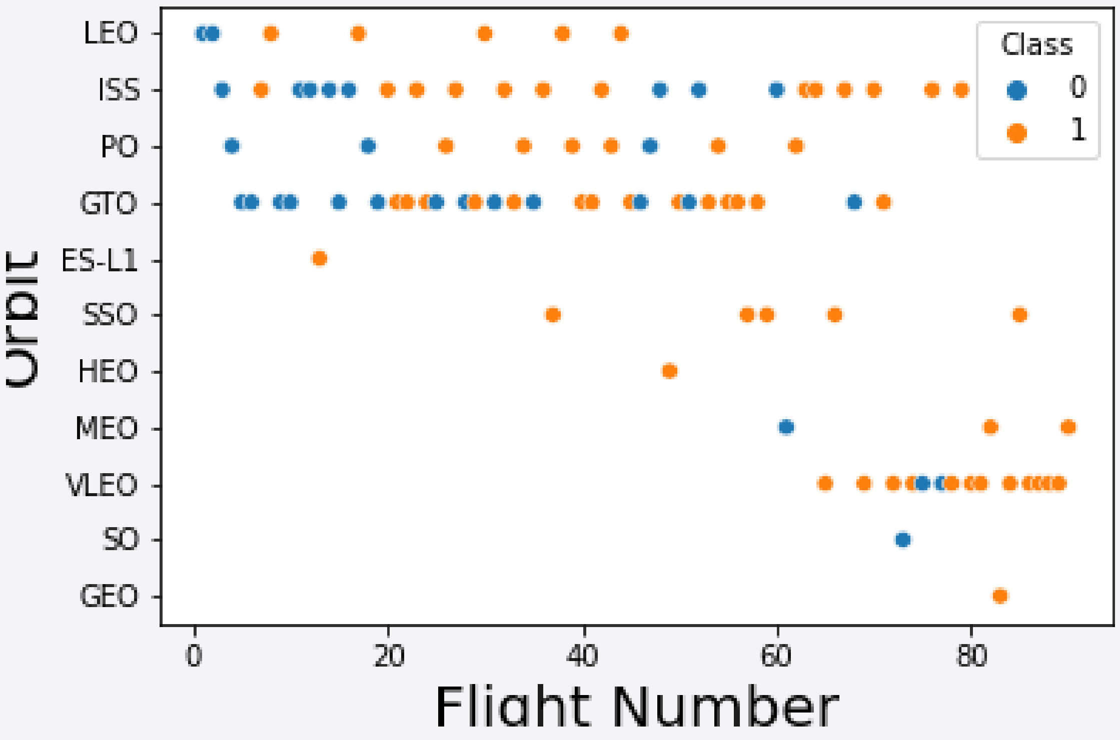
Success Rate vs. Orbit Type

- Orbit types for ES-L1, SSO, HEO, and GEO have 100% success rate.
- Orbit type GTO has the lowest success rate at 50%.



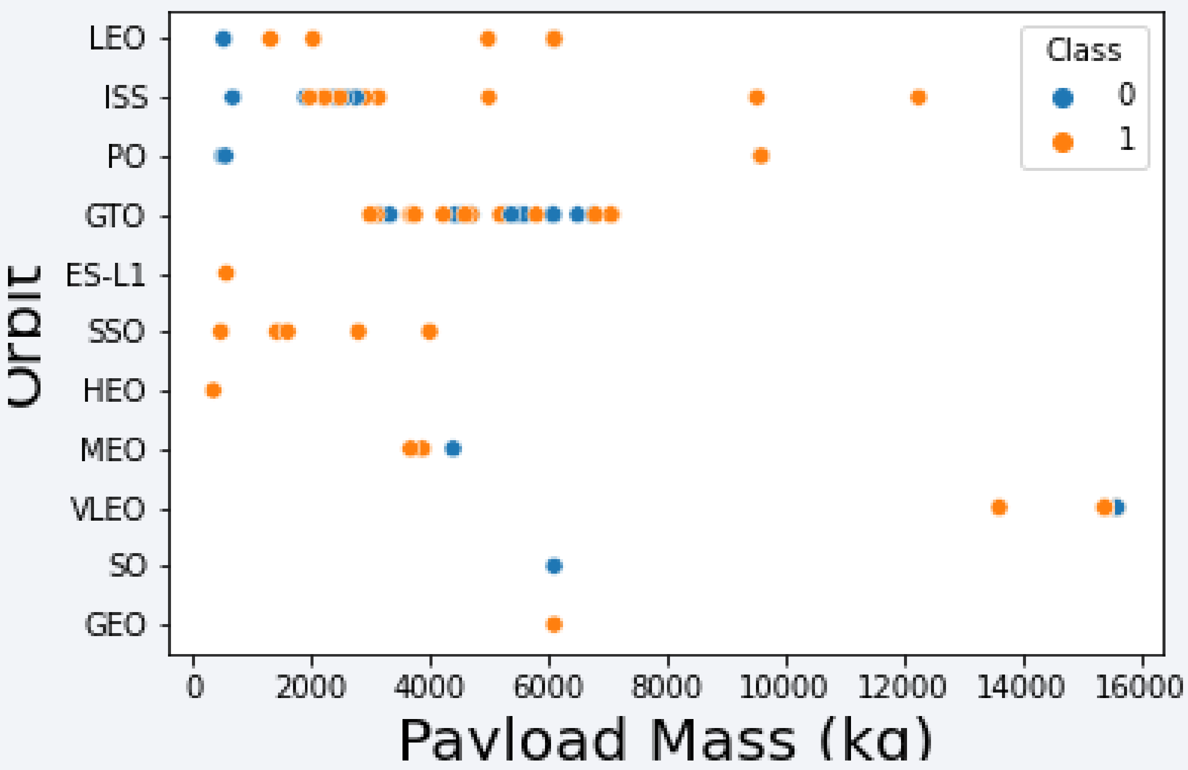
Flight Number vs. Orbit Type

- High success rate is observed for very low earth orbits (VLEO). It also has high flight number.
- There are more flights to the LEO, ISS, PO, and GTO orbits.



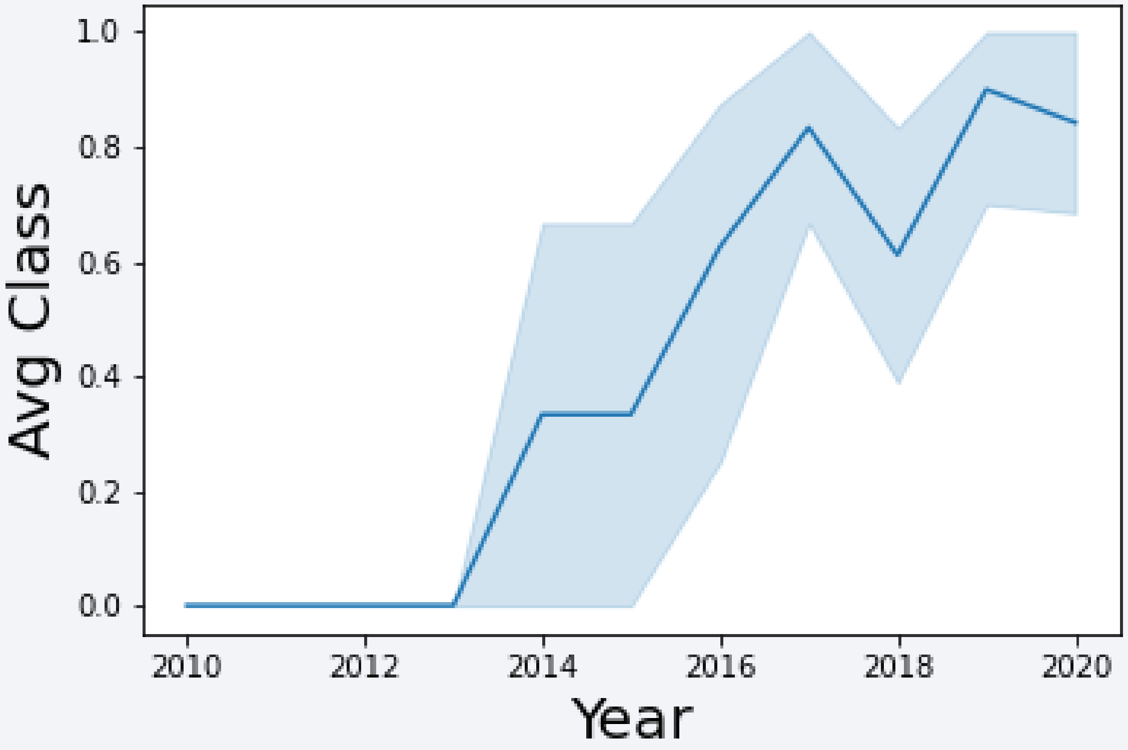
Payload vs. Orbit Type

- Most payload masses are below 8000 kg.
- There is a 100% success rate for all payload mass for the sun-synchronous orbit (SSO).



Launch Success Yearly Trend

- The average success rate increases over years.
- The success rate has been above 60% after year 2016.



All Launch Site Names

- There are four distinct launch sites in the data that we analyze.

No	Launch Site
1	CCAFS LC-40
2	CCAFS SLC-40
3	KSC LC-39A
4	VAFB SLC-4E

Launch Site Names Begin with 'CCA'

- We query the records where the 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

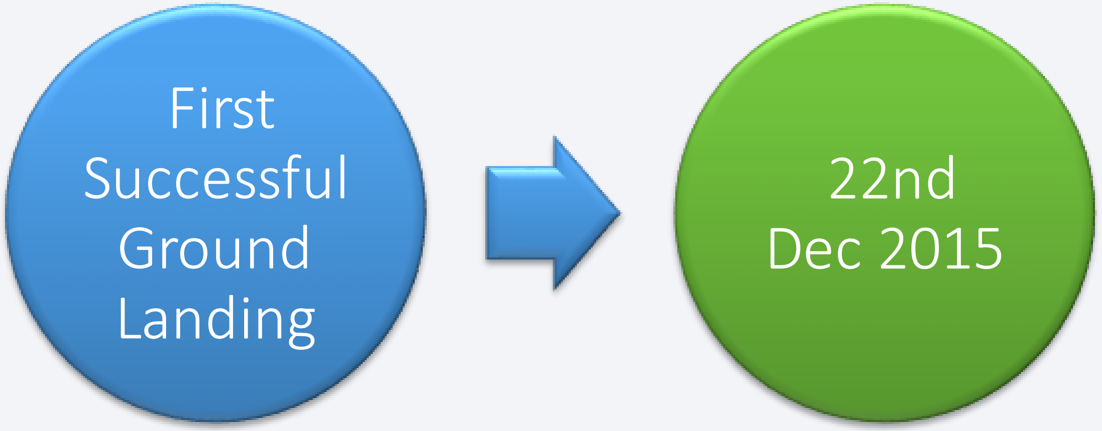
- The total payload mass carried by boosters launched by NASA (CRS) is 45,596 kg.

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2,928 kg.

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad was achieved on **22nd December 2015**.



Successful Drone Ship Landing with Payload between 4000 and 6000

- The table shows the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000.
- There are a total of four boosters achieving this criteria.

No	Booster Version
1	F9 FT B1021.2
2	F9 FT B1031.2
3	F9 FT B1022
4	F9 FT B1026

Total Number of Successful and Failure Mission Outcomes

- The total number of successful mission outcomes is 100, where one of the payload status is unclear.
- The total number of failure mission outcomes is 1.

Mission Outcome	Count
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- The maximum payload mass for all boosters is 15,600 kg.
- There is a total of 12 booster which have carried the maximum payload mass.

No	Booster Version
1	F9 B5 B1048.4
2	F9 B5 B1048.5
3	F9 B5 B1049.4
4	F9 B5 B1049.5
5	F9 B5 B1049.7
6	F9 B5 B1051.3
7	F9 B5 B1051.4
8	F9 B5 B1051.6
9	F9 B5 B1056.4
10	F9 B5 B1058.3
11	F9 B5 B1060.2
12	F9 B5 B1060.3

2015 Launch Records

- In 2015, there are two failed landing outcomes on drone ship.
- Their booster versions and launch sites are shown below.

Booster Version	Launch Site	Landing Outcome
F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

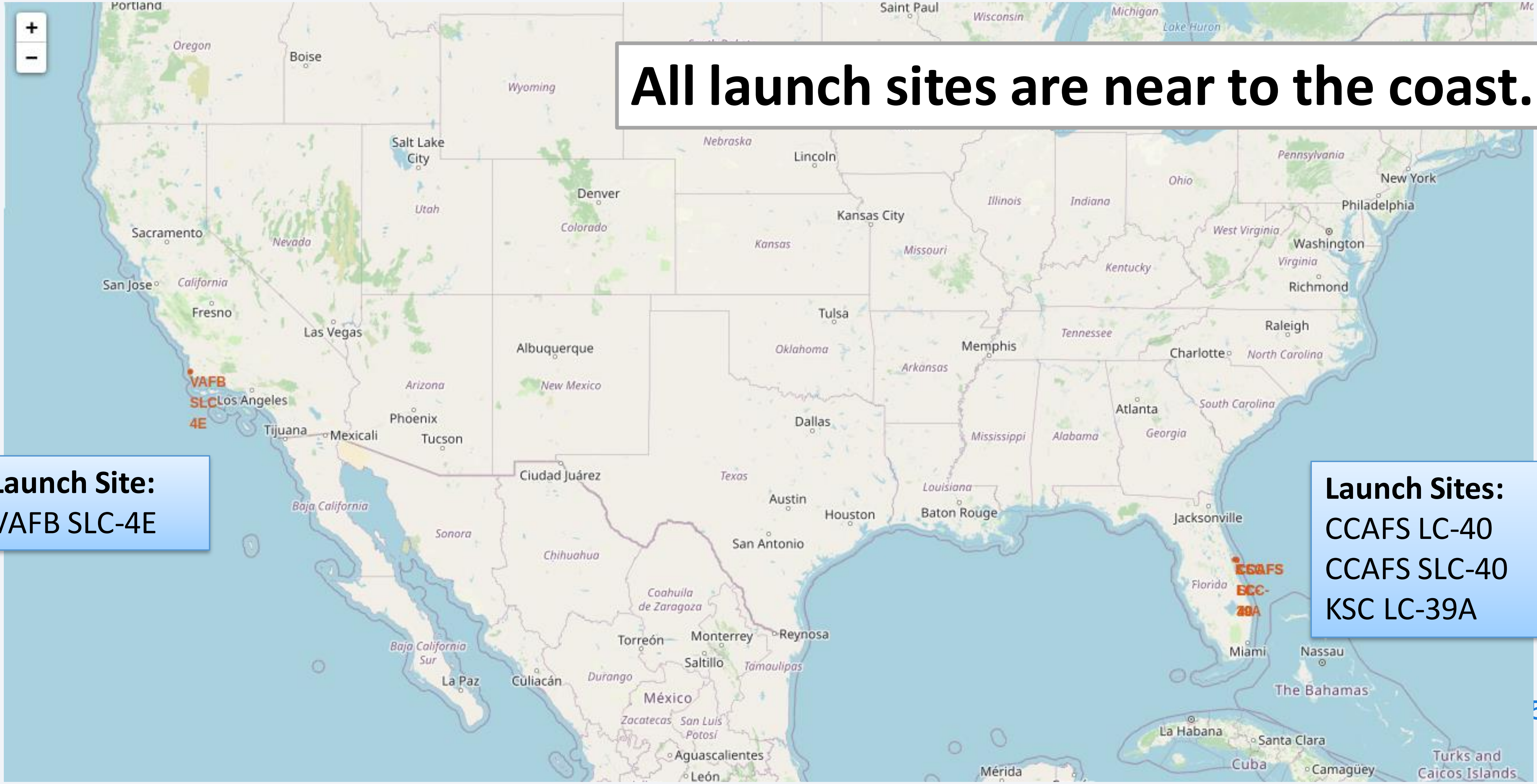
- The landing outcomes between 2010-06-04 and 2017-03-20 are ranked in descending order and result shown in the table.
- There are more failures and no attempt during this initial period.

Landing Outcome	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 4

Launch Sites Proximities Analysis

All Launch Site Locations



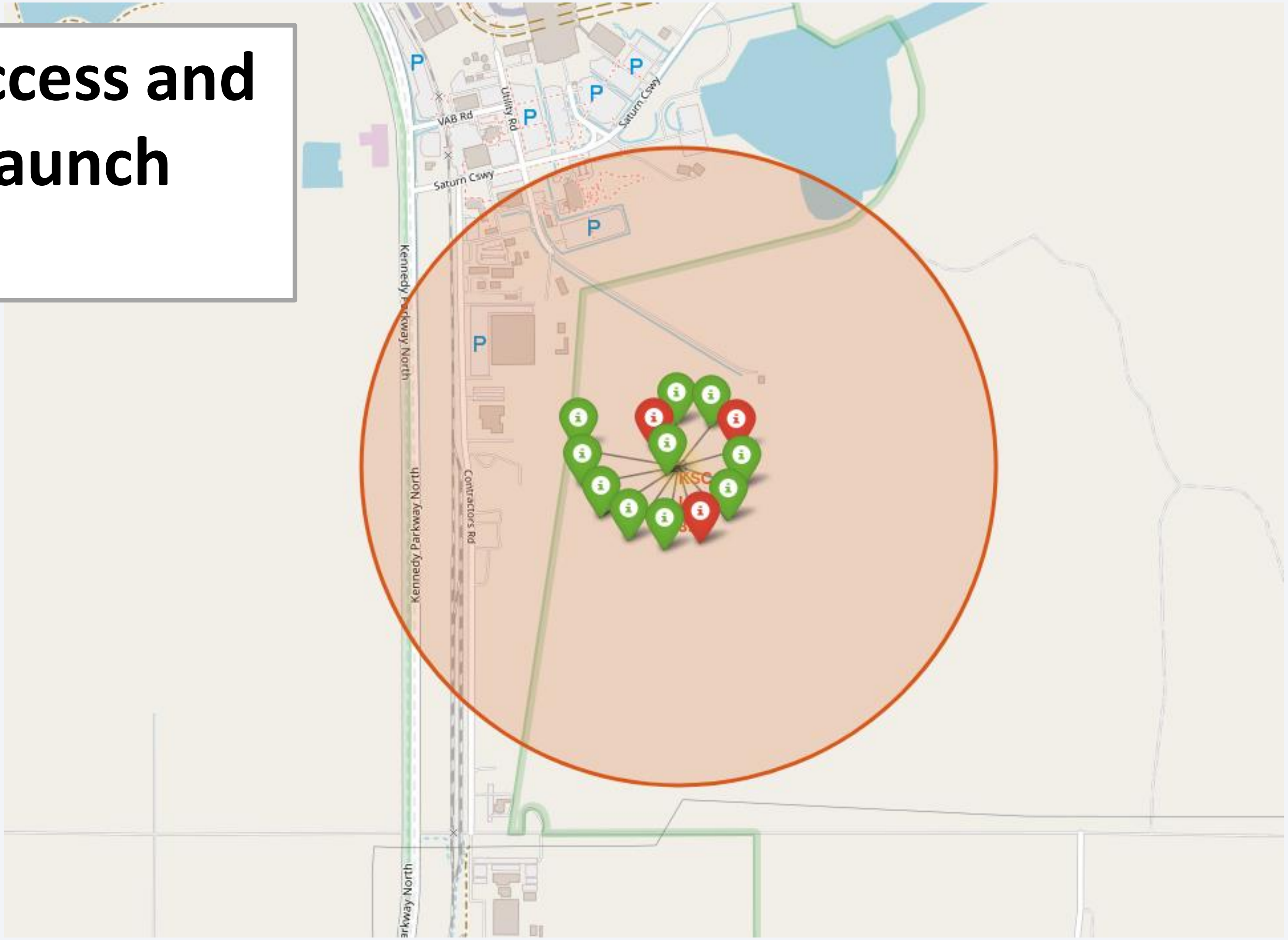
All launch sites are near to the coast.

Launch Site:
VAFB SLC-4E

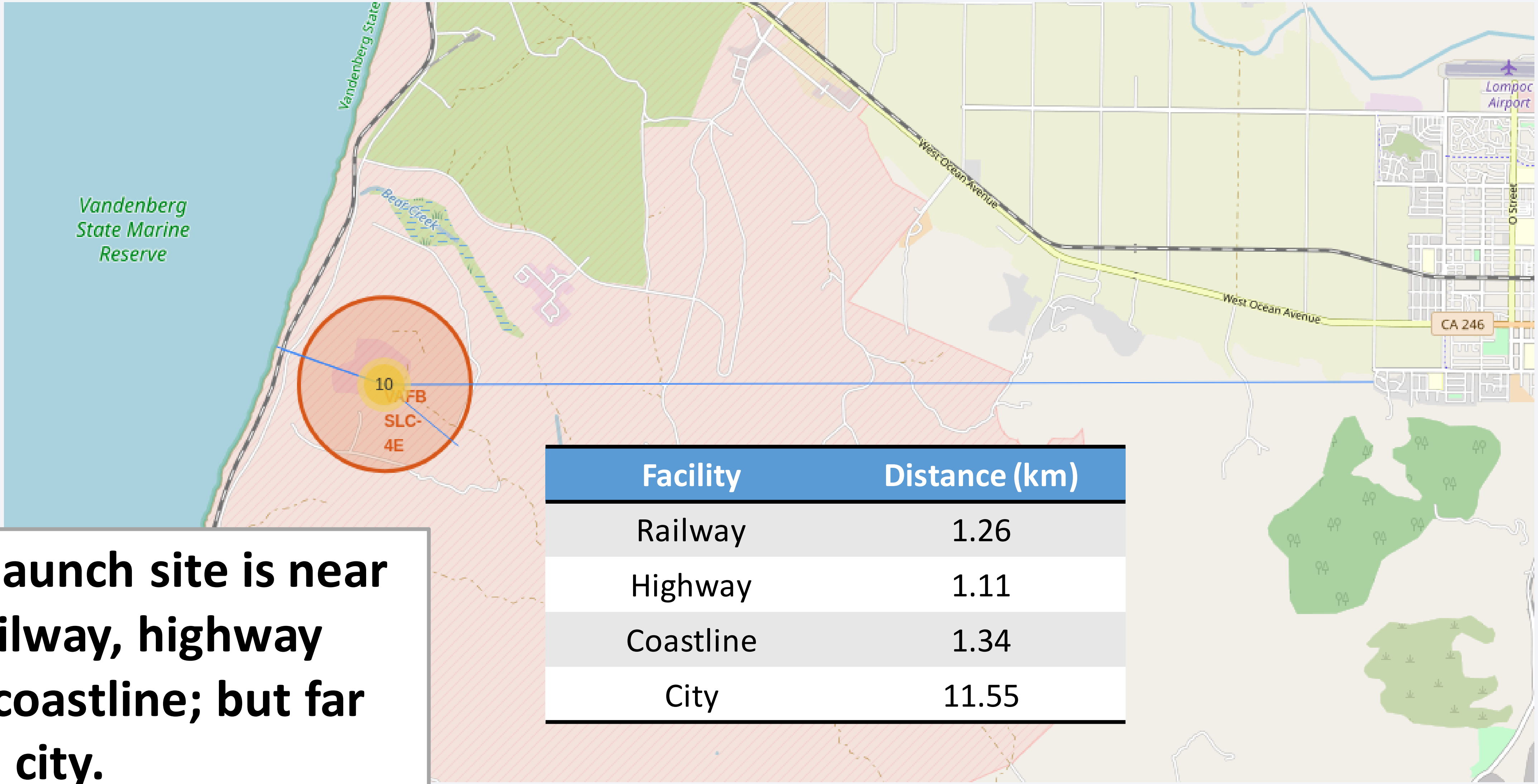
Launch Sites:
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A

Launch Outcomes for Site KSC LC-39A

There are 10 success and 3 failure at the launch site KSC LC-39A.



Proximities of Launch Site to Facilities



Facility	Distance (km)
Railway	1.26
Highway	1.11
Coastline	1.34
City	11.55

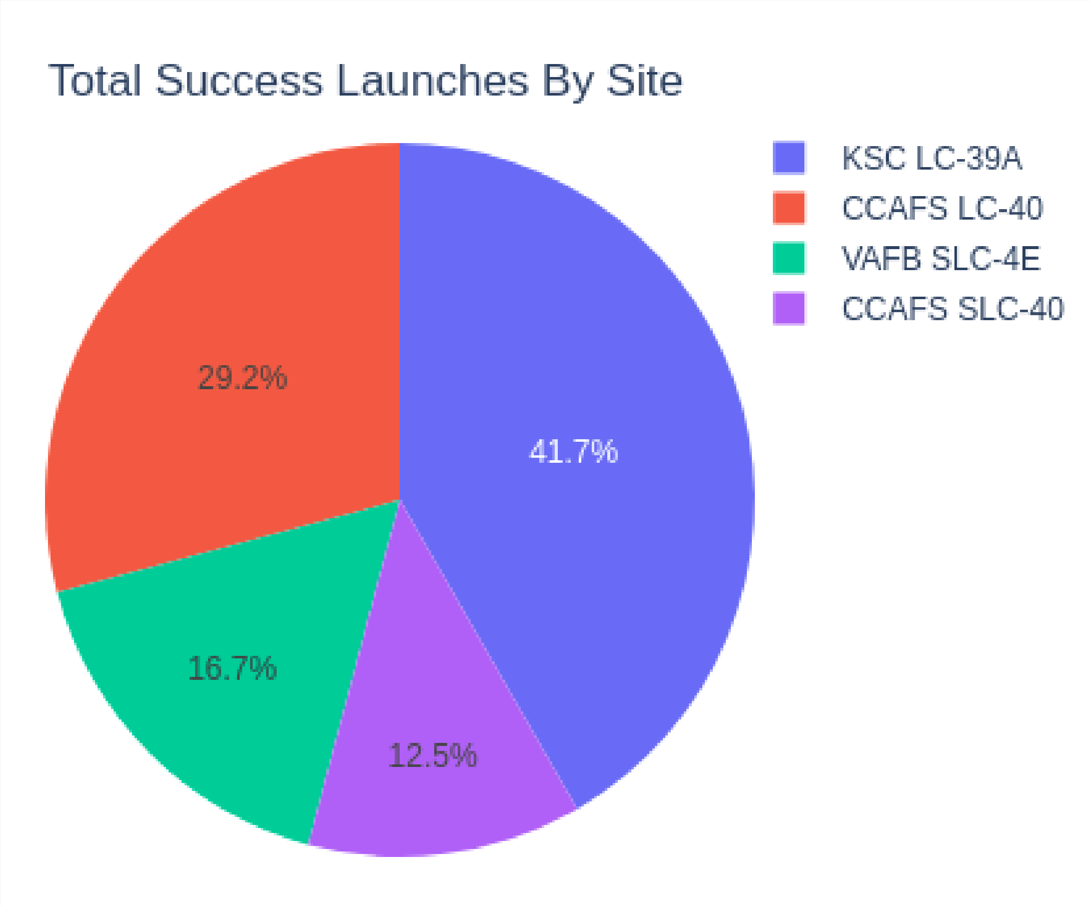
The launch site is near to railway, highway and coastline; but far from city.

Section 5

Build a Dashboard with Plotly Dash

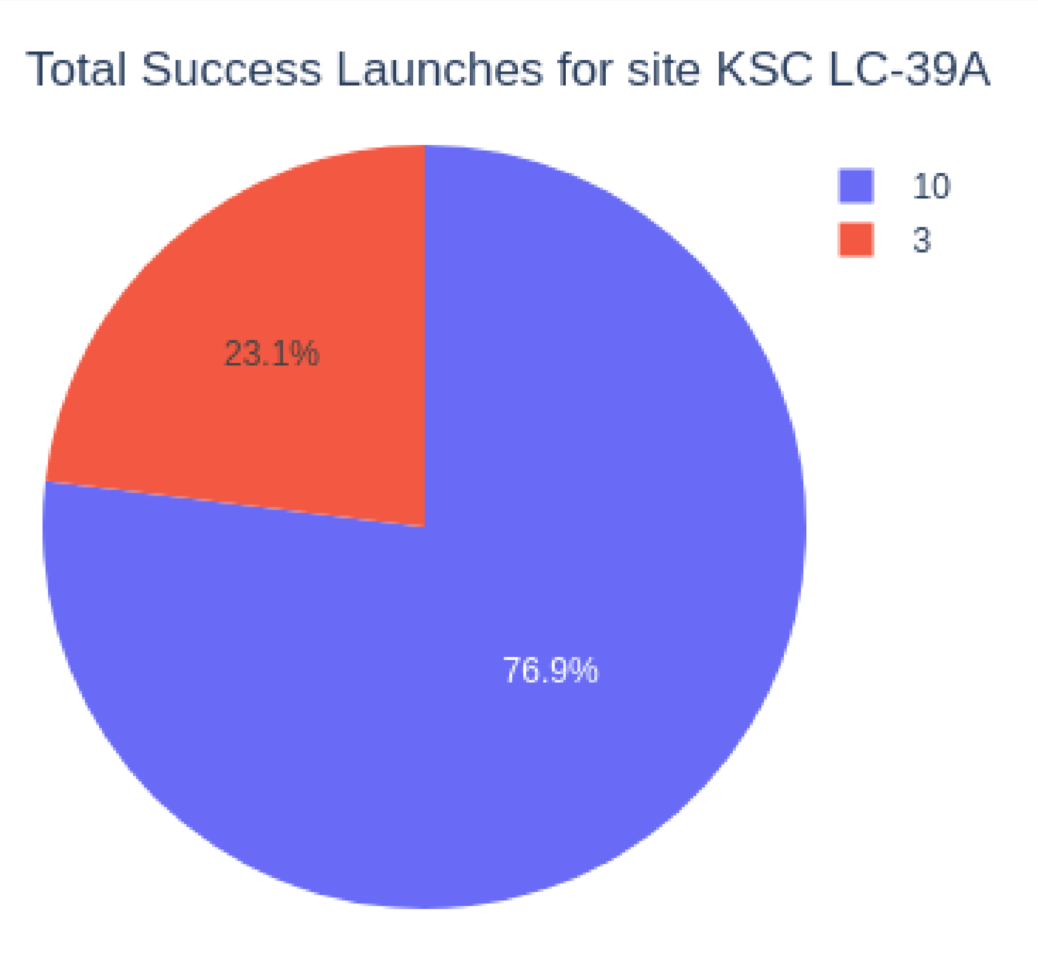
Success Rate for All Launch Sites

- Launch site KSC LC-39A has the highest success count compared to other launch sites.
- It accounted for 41.7% of all the success launches.



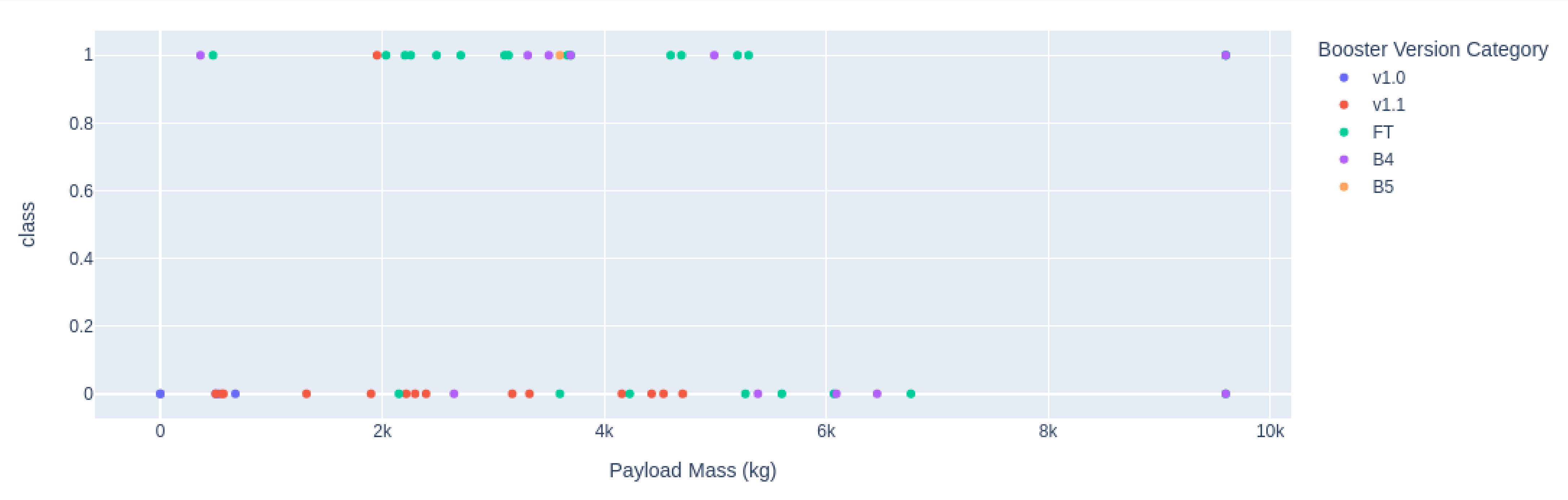
Success Rate for Launch Site KSC LC-39A

- KSC LC-39A has the highest launch success ratio.
- KSC LC-39A made 13 launches.
  - 10 success launches.
  - 3 failed launches.



Correlation between Payload and Launch Outcome

- Booster version category FT has the largest success rate.
- Within this FT booster category, payload range between 2,000 kg to 4,000 kg has a better success rate.

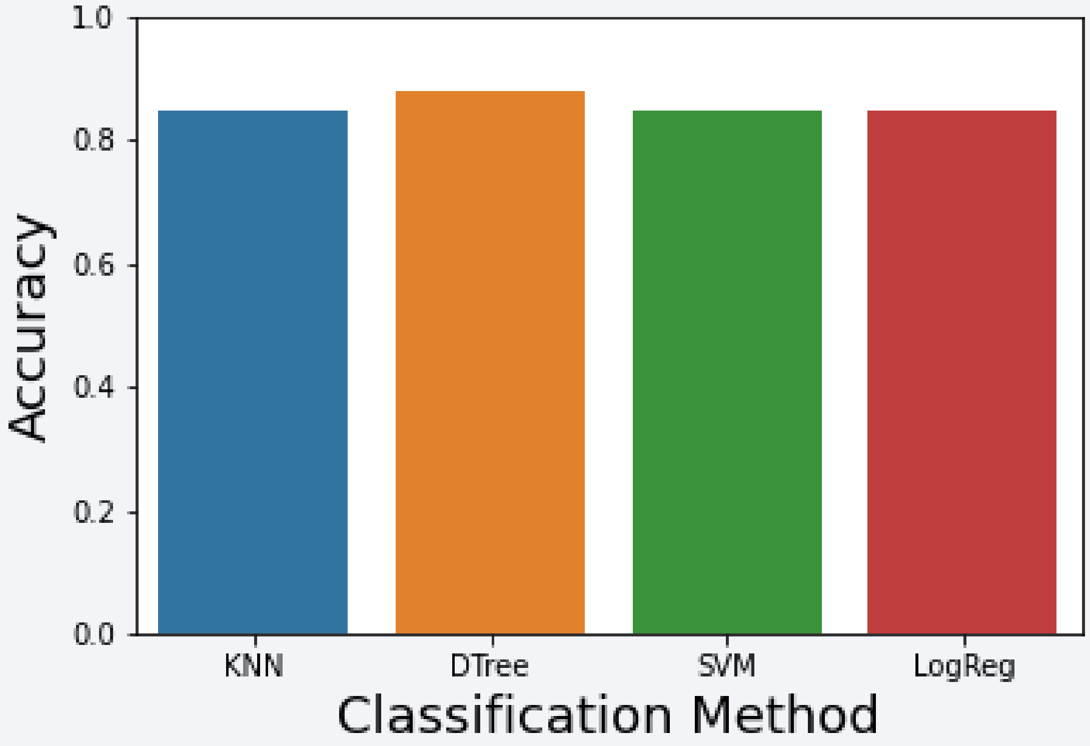


Section 6

Predictive Analysis (Classification)

Classification Accuracy

- The decision tree classification (DTree) method has the highest accuracy at 0.8768.
- The other three classification methods like K nearest neightbor (KNN), support vector machine (SVM), logistic regression (LogReg) have similar accuracy around 0.8482.



Confusion Matrix

- The confusion matrix for the decision tree classification method shows a high true positive result.
- There are 12 cases of landed true labels with correct prediction.
- There are 3 cases of did not land true labels with correct prediction.
- The incorrect predictions are 0 and 3 cases each.

Conclusions

- Success rate has been high after year 2017.
- Optimum payload is between 2,000 kg to 4,000 kg.
- All launch sites are near to coastline, highway or railway.
- Launch site KSC LC-39A has the highest success count
- Decision Tree classification method has the highest accuracy in predicting the landing outcome.

Appendix – Python Code Snippet for Dashboard

```python
# Import required libraries
import pandas as pd
import dash
#import dash_html_components as html
#import dash_core_components as dcc
from dash import html
from dash import dcc
from dash.dependencies import Input, Output
import plotly.express as px

# Read the airline data into pandas dataframe
spacex_df = pd.read_csv("spacex_launch_dash.csv")
max_payload = spacex_df['Payload Mass (kg)'].max()
min_payload = spacex_df['Payload Mass (kg)'].min()

# Create a dash application
app = dash.Dash(__name__)

# Create an app layout
app.layout = html.Div(children=[html.H1('SpaceX Launch Records Dashboard',
                                        style={'textAlign': 'center', 'color': '#503D36',
                                               'font-size': 40}),
                                # TASK 1: Add a dropdown list to enable Launch Site selection
                                # The default select value is for ALL sites
                                # dcc.Dropdown(id='site-dropdown',...)
                                html.Div(dcc.Dropdown(id='site-dropdown',
                                    options=[
                                        {'label': 'All Sites', 'value': 'ALL'},
                                        {'label': 'CCAFS LC-40', 'value': 'CCAFS LC-40'},
                                        {'label': 'CCAFS SLC-40', 'value': 'CCAFS SLC-40'},
                                        {'label': 'KSC LC-39A', 'value': 'KSC LC-39A'},
                                        {'label': 'VAFB SLC-4E', 'value': 'VAFB SLC-4E'}
                                    ],
                                    value='ALL',
                                    placeholder='Select a Launch Site here'
```

You have Maven installed on your system. Do you want to

Appendix – SQL Queries

```
%sql select max(PAYLOAD_MASS__KG_)  from SPACEXDATASET;

%%sql
select distinct(BOOSTER_VERSION)
from SPACEXDATASET
where PAYLOAD_MASS__KG_  = (select max(PAYLOAD_MASS__KG_)  from SPACEXDATASET);

%%sql
select BOOSTER_VERSION,  LAUNCH_SITE,  LANDING__OUTCOME
from SPACEXDATASET
where LANDING__OUTCOME  = 'Failure (drone ship)' and YEAR(DATE) = '2015';
```

Thank you!