IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Peng Le
19/02/2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive Analytics result

Introduction

- Project background and context
  - We are from a Company called Space Y, to compete with SpaceX
  - The project is to determine the winning strategy for Space Y against Space X
- Problems you want to find answers
  - We intend to find out the price of each launch and how to compete with SpaceX
  - This is done by using machine learning model to predict if SpaceX will reuse the first stage

Section 1

Methodology

Methodology

Executive Summary

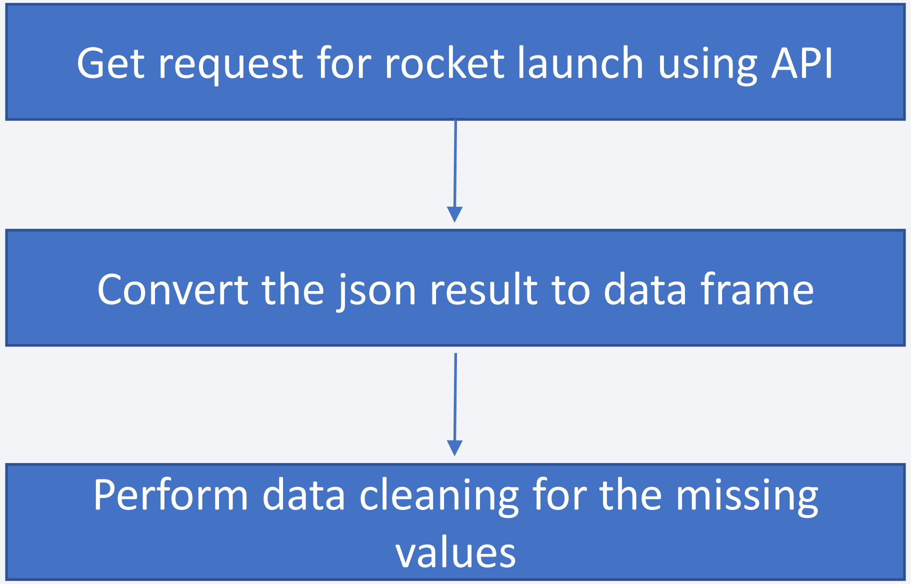
- Data collection methodology:
  - Data was collected using SpaceX REST API
  - Data is also collected from Wikipedia using Python BeautifulSoup Package
- Perform data wrangling
  - One hot encoding was applied to categorical features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- Data are collected using these key processes:
  - Frist, Data collection was done using get request to the SpaceX API
  - Next, the response content was decoded as a Json using json() function call and turn it into a pandas dataframe using json_normalize
  - The data was cleaned, checked for missing values and missing values were filled
  - Additional data were collected using web scraping from Wikipedia for Falcon 9 launch records with BeautifulSoup
  - The launch records as HTML table, parse the table and converted to a pandas dataframe for future analysis

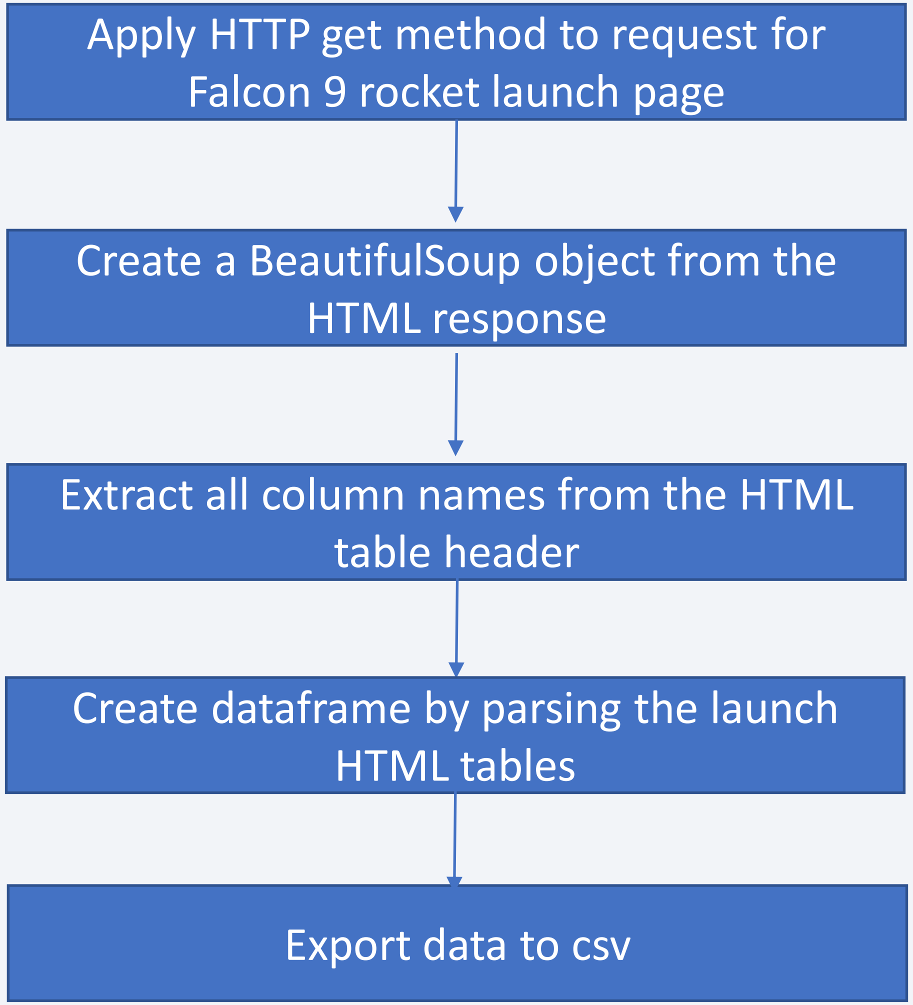
Data Collection – SpaceX API

- Get request to the SpaceX API to collect data, clean the requested data and did some basic data wrangling and formatting.
- Github URL: https://github.com/swatowpengle/Data-Science-and-IBM-Machine-Learning-Capstone-Project/blob/main/Complete%20the%20Data%20Collection%20API%20Lab.ipynb



Data Collection - Scraping

- Web scrapping is done to Falcon 9 launch records with BeautifulSoup
- The table was parsed and convert to dataframe
- GitHub URL: https://github.com/swatowpengle/Data-Science-and-IBM-Machine-Learning-Capstone-Project/blob/main/Complete%20the%20Data%20Collection%20with%20Web%20Scraping%20lab.ipynb



Data Wrangling

- Data are processed using the following steps:
  - Calculate the number of launches on each site
  - Calculate the number and occurrence of each orbit
  - Calculate the number and occurrence of mission outcome per orbit type
  - Create a landing outcome label from Outcome column
- GitHub URL: https://github.com/swatowpengle/Data-Science-and-IBM-Machine-Learning-Capstone-Project/blob/main/Complete%20the%20EDA%20lab.ipynb

EDA with Data Visualization

- The charts include
  - 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
- GitHub URL: https://github.com/swatowpengle/Data-Science-and-IBM-Machine-Learning-Capstone-Project/blob/main/Complete%20the%20EDA%20with%20Visualization.ipynb

EDA with SQL

- The SQL queries include
  - The names of the unique launch sites in the space mission
  - 5 records where launch sites begin with the string 'KSC'
  - Total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1
  - Date where the successful landing outcome in drone ship was achieved
  - The records which will display the month names, successful landing outcomes in ground pad ,booster versions, launchsite for the months in year 2017
  - Count of successful landingoutcomes between the date 2010-06-04 and 2017-03-20 in descending order.
- GitHub URL: https://github.com/swatowpengle/Data-Science-and-IBM-Machine-Learning-Capstone-Project/blob/main/jupyter-labs-eda-sql-edx.ipynb

Build an Interactive Map with Folium

- Sites are marked in the map
- The outcome of successes and failures at each site is visualized :
- The distances from the sites to railways, highways, coastlines and cities are calculated
- GitHub URL: https://github.com/swatowpengle/Data-Science-and-IBM-Machine-Learning-Capstone-Project/blob/main/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

- An interactive dashboard is built with Plotly dash
- Pie charts are plotted to show the total launches by a certain sites
- Scatter graph are ploted to the relationship with Outcome and Payload Mass (Kg) for the different booster version

Predictive Analysis (Classification)

- Data is loaded using numpy and pandas, transformed the data, split our data into training and testing
- Different machine learning models and tune different hyoerparameters using GridSearchCV
- Accuracy is used as the metric for our model to improve the model using feature engineering and algorithm tuning
- The best performing classification model is identified
- Github URL: https://github.com/swatowpengle/Data-Science-and-IBM-Machine-Learning-Capstone-Project/blob/main/Machine%20Learning%20Prediction.ipynb

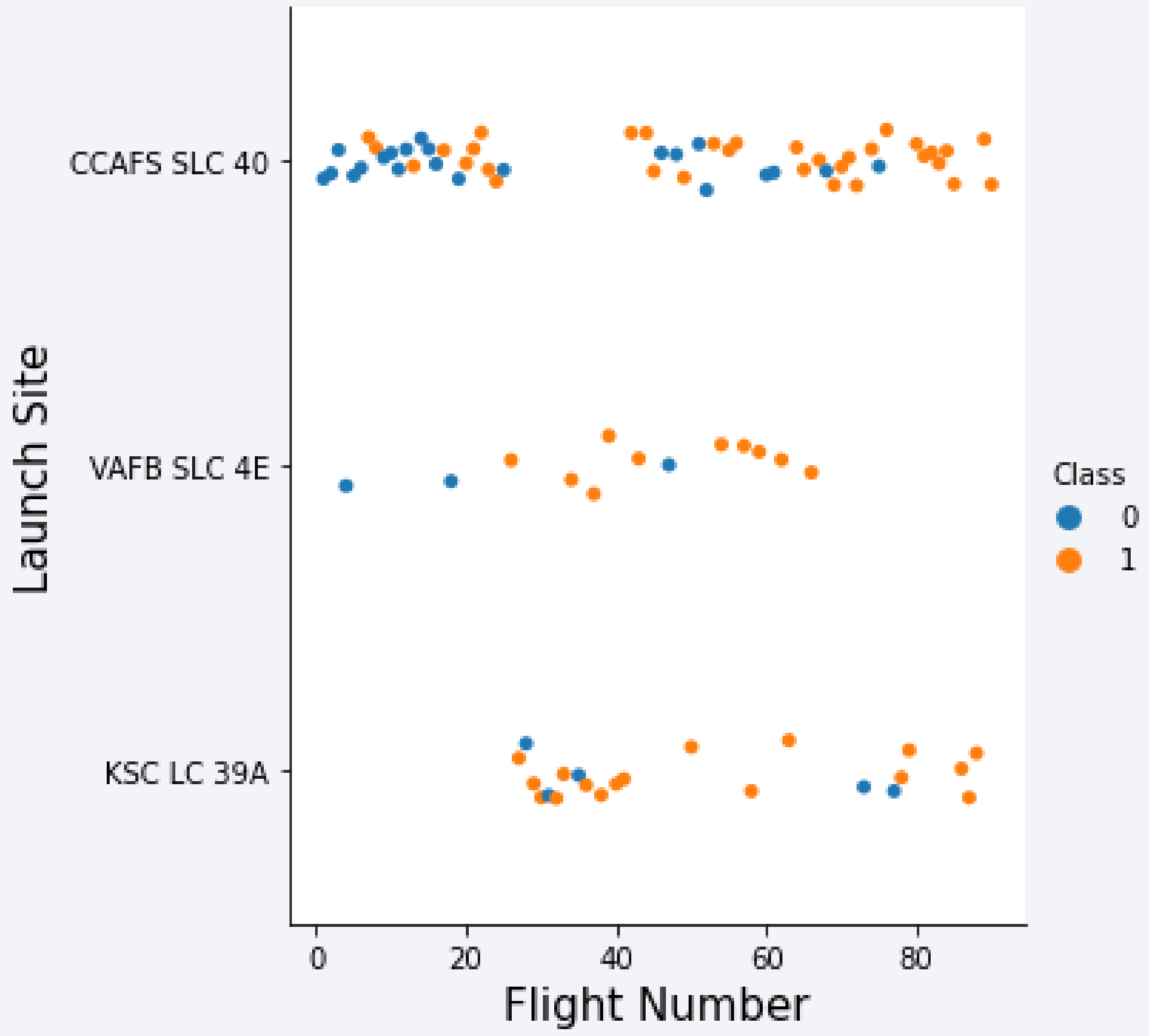
Results

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

Section 2

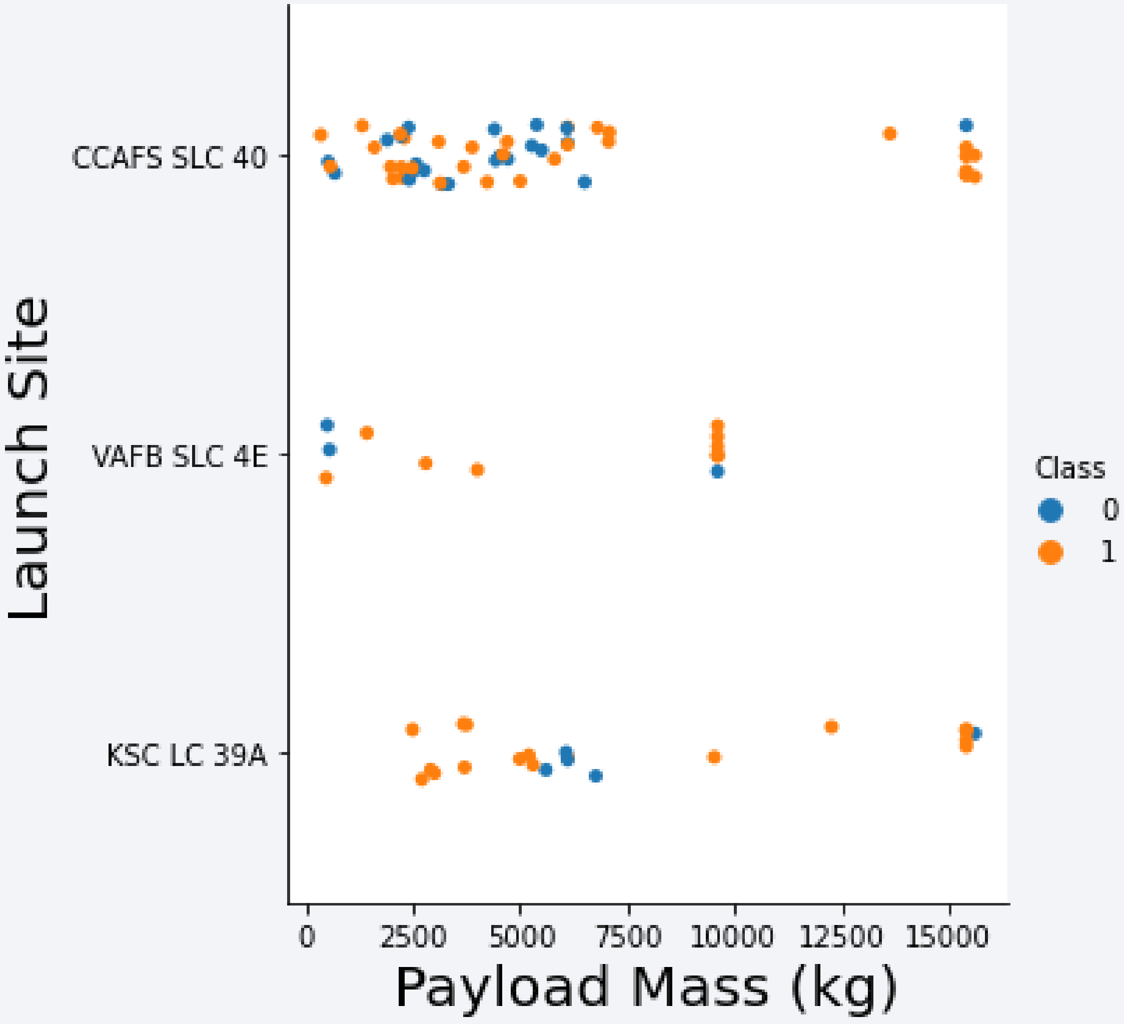
Insights drawn from EDA

Flight Number vs. Launch Site



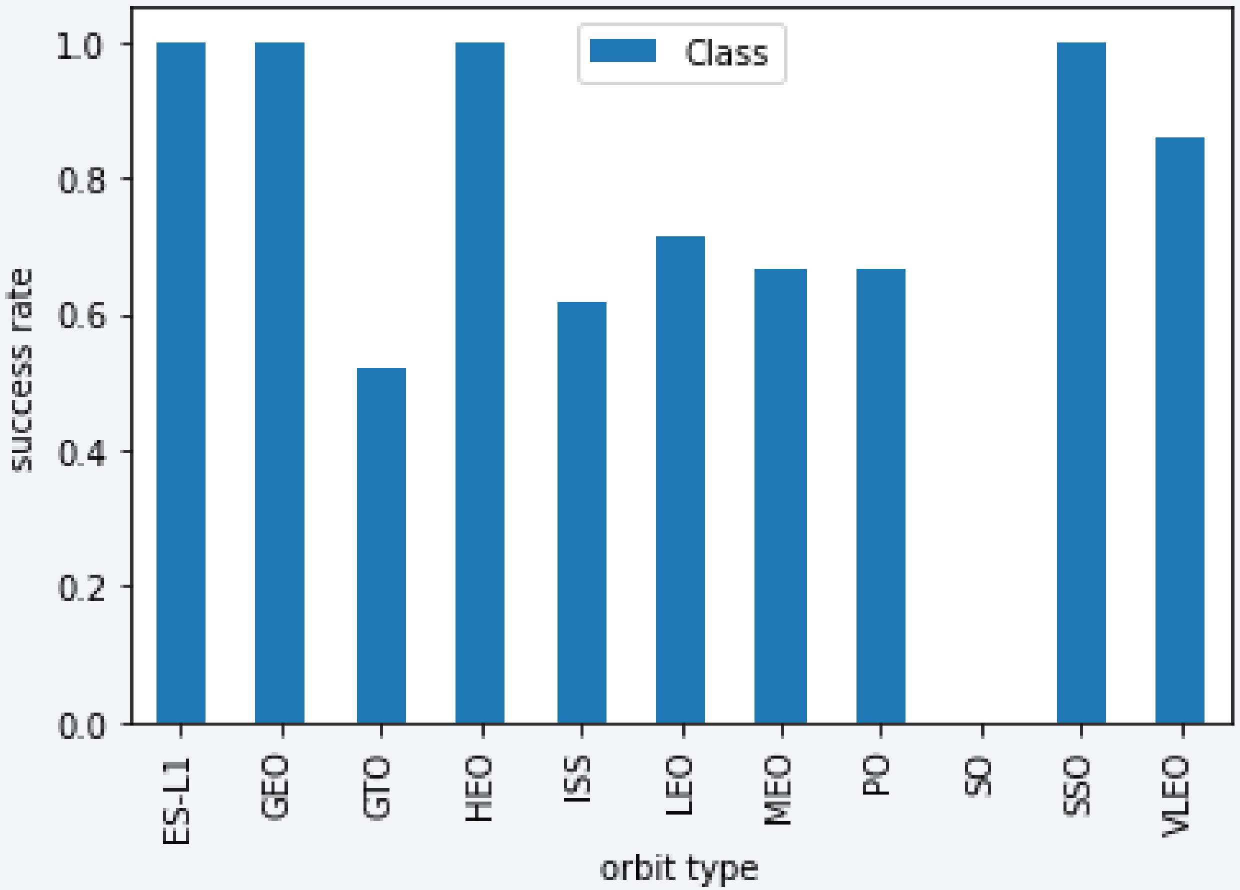
- Flight Number Vs Launch Site
  - X-axis: Flight Number
  - Y-axis: Launch Site
  - 0 (Blue Dot) represents failure
  - 1(Orange Dot) represents success
- The success of launching does not depend on the launching site as shown in the diagram

Payload vs. Launch Site



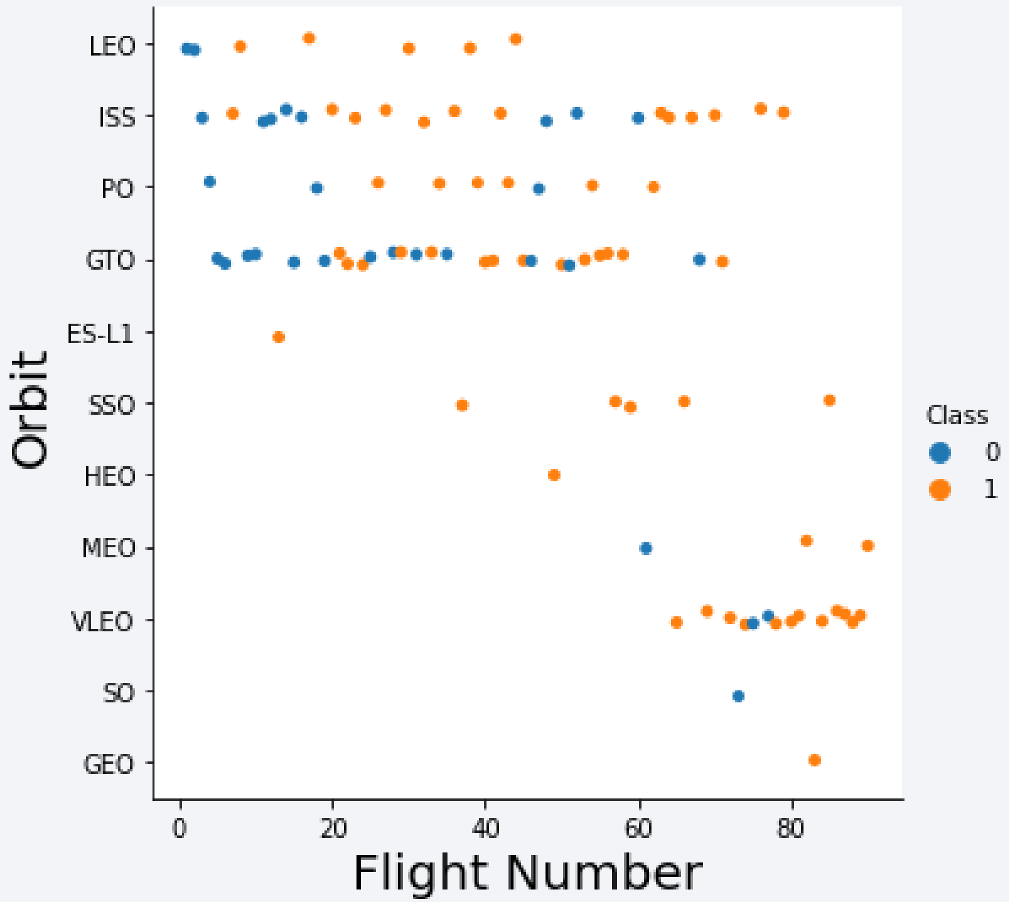
- Payload Vs Launch Site
  - X-axis: Payload Mass
  - Y-axis: Launch Site
  - 0 (Blue Dot) represents failure
  - 1(Orange Dot) represents success
- There is little correlation between payload and launch sites

Success Rate vs. Orbit Type



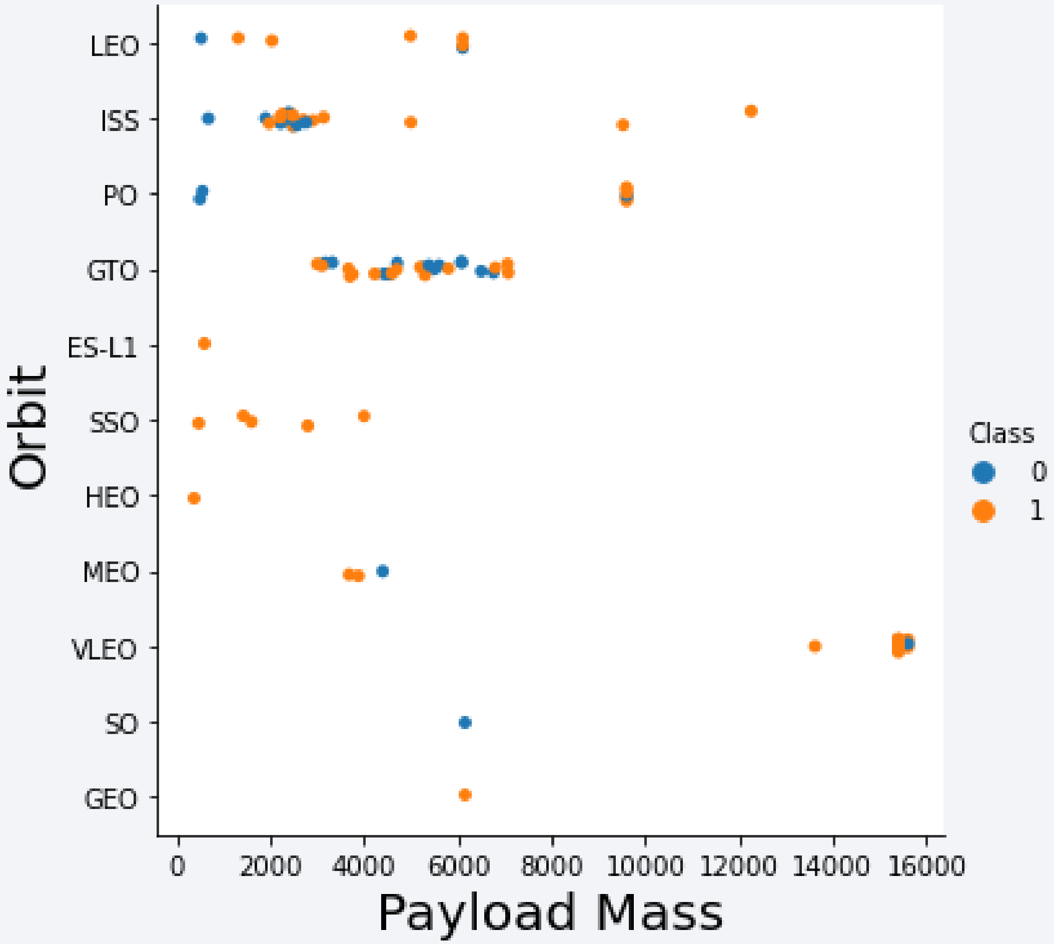
- Success Rate Vs Orbit Type
  - X-axis: Orbit Type
  - Y-axis: Success Rate
  - The height of bar represents success rate
  - 1 is maximum, 0 is minimum
- ES-L1, GEO, HEO and SSO have 100% success rate

Flight Number vs. Orbit Type



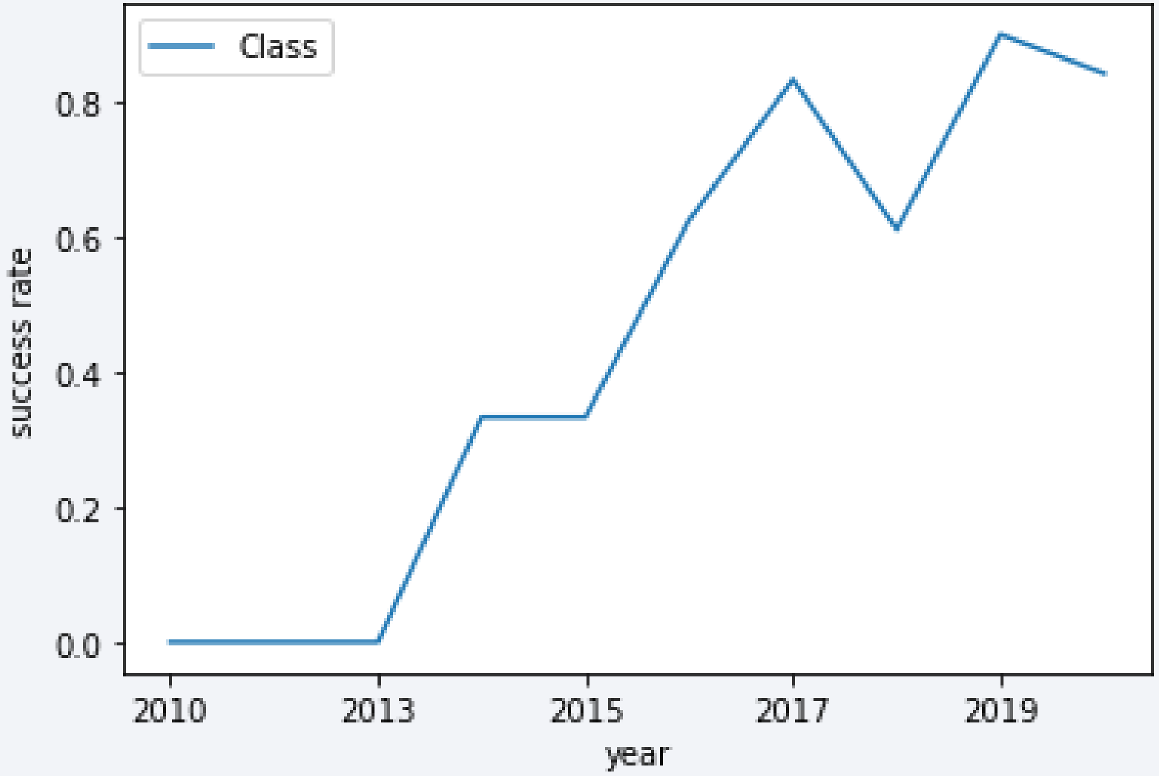
- Flight Number vs. Orbit Type
  - X-axis: Flight Number
  - Y-axis: Orbit Type
  - 0 (Blue Dot) represents failure
  - 1(Orange Dot) represents success

Payload vs. Orbit Type



- Show a scatter point of payload vs. orbit type
- Show the screenshot of the scatter plot with explanations

Launch Success Yearly Trend



- Launch Success Yearly Trend
  - X-axis: Year
  - Y-axis: Success rate
- Generally the success rate increases over years

All Launch Site Names

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

- SQL Query:

```
SELECT DISTINCT launch_site FROM
SPACEXTBL
```

- Screen shot of results are shown in the right

Launch Site Names Begin with 'KSC'

- SQL Query:

Display 5 records where launch sites begin with the string 'KSC'

```
5]: %sql SELECT * FROM SPACEXTBL where launch_site LIKE 'KSC%' LIMIT 5
```

- Screen shot of results are shown below

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	06:00:00	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

Total Payload Mass

- The total payload mass for different versions are shown in the screenshot at the right

booster_version	2
F9 B4 B1039.2	2647
F9 B4 B1039.1	3310
F9 B4 B1045.2	2697
F9 B5 B1056.2	2268
F9 B5 B1058.4	2972
F9 B5 B1059.2	1977
F9 B5B1050	2500
F9 B5B1056.1	2495
F9 FT B1035.2	2205
F9 FT B1021.1	3136
F9 FT B1025.1	2257
F9 FT B1031.1	2490
F9 FT B1035.1	2708
F9 v1.0 B0006	500
F9 v1.0 B0007	677
F9 v1.1	2296
F9 v1.1 B1010	2216
F9 v1.1 B1012	2395
F9 v1.1 B1015	1898
F9 v1.1 B1018	1952

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- The average payload mass is 2928

```
%sql SELECT AVG(payload_mass__kg_) FROM SPACEXTBL WHERE booster_version = 'F9 v1.1'
```

```
 * ibm_db_sa://dht81407:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
```

1
2928

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- The query is shown below

List the date where the succesful landing outcome in drone ship was acheived.

Hint:Use min function

```
%sql SELECT MIN(DATE) FROM SPACEXTBL WHERE landing__outcome = 'Success (drone ship)'
```

```
 * ibm_db_sa://dht81407:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
```

1
2016-04-08

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are shown at the right

booster_version
F9 B4 B1040.1
F9 B4 B1043.1
F9 FT B1032.1

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes
- There are 1 failure in flight, 99 successes and 1 success but the payload status not clear

mission_outcome	**2**
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- The name of the boosters and the payload mass are shown in the right

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

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 are shown in the right

MONTH	landing__outcome	booster_version	launch_site
2	Success (ground pad)	F9 FT B1031.1	KSC LC-39A
5	Success (ground pad)	F9 FT B1032.1	KSC LC-39A
6	Success (ground pad)	F9 FT B1035.1	KSC LC-39A
8	Success (ground pad)	F9 B4 B1039.1	KSC LC-39A
9	Success (ground pad)	F9 B4 B1040.1	KSC LC-39A
12	Success (ground pad)	F9 FT B1035.2	CCAFS SLC-40

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

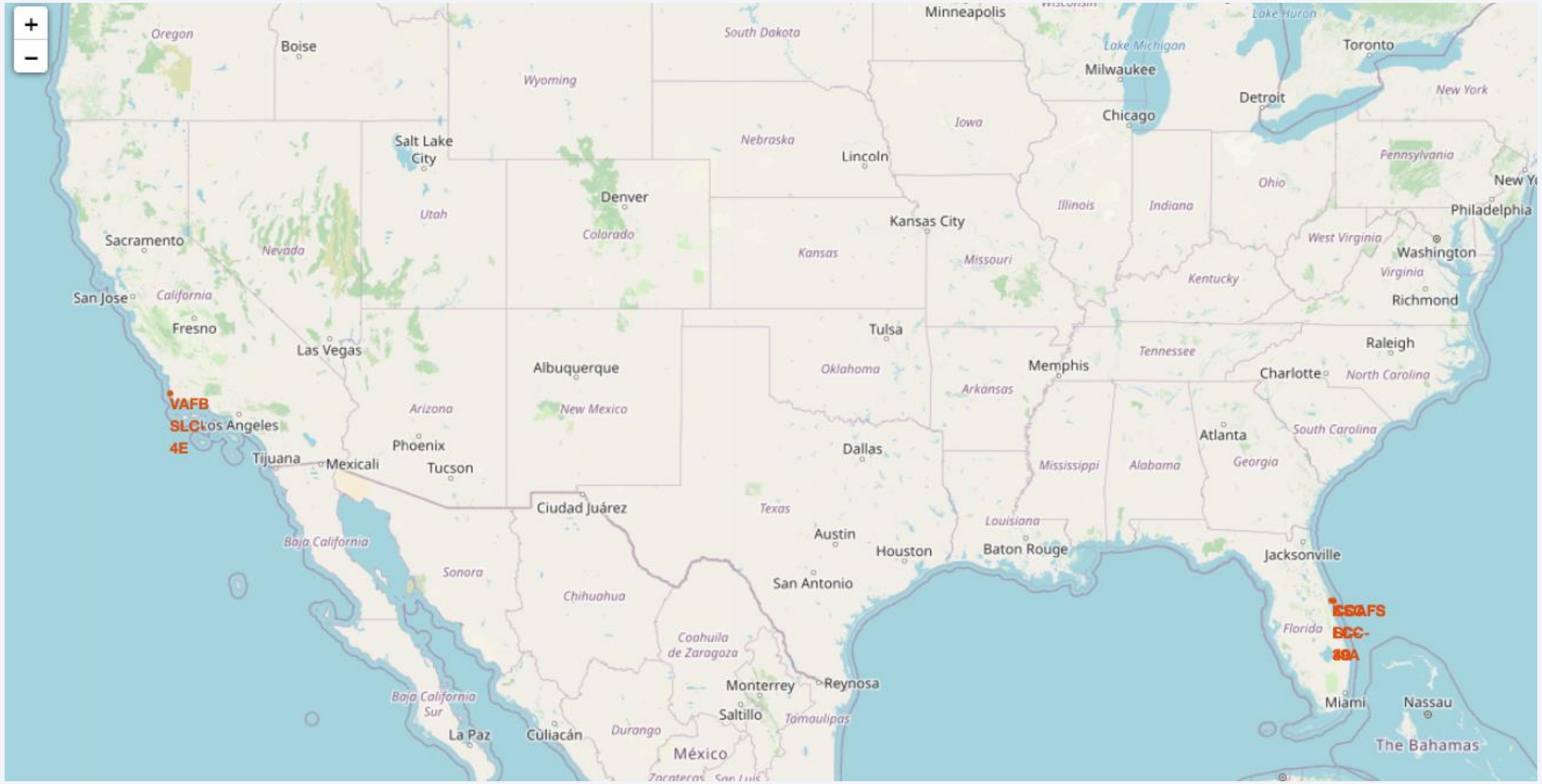
- Rank the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order

DATE	landing__outcome
2017-02-19	Success (ground pad)
2017-01-14	Success (drone ship)
2016-08-14	Success (drone ship)
2016-07-18	Success (ground pad)
2016-05-27	Success (drone ship)
2016-05-06	Success (drone ship)
2016-04-08	Success (drone ship)
2015-12-22	Success (ground pad)

Section 4

Launch Sites Proximities Analysis

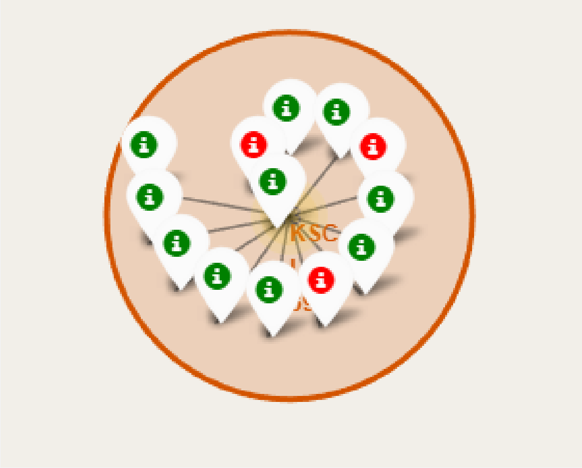
All Launch Sites

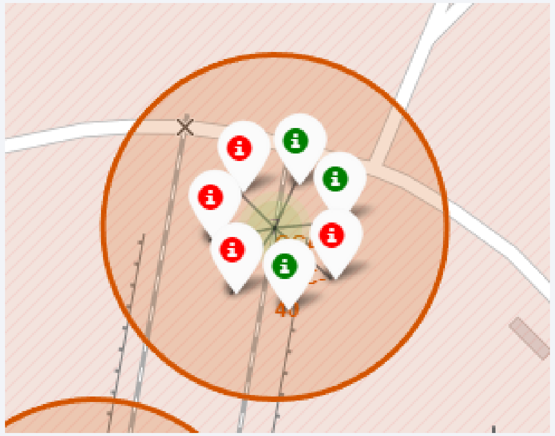


- All launch sites are in proximity to equator line and are in close proximity to the coast

Markers showing launch sites with color labels

Florida Launch Sites



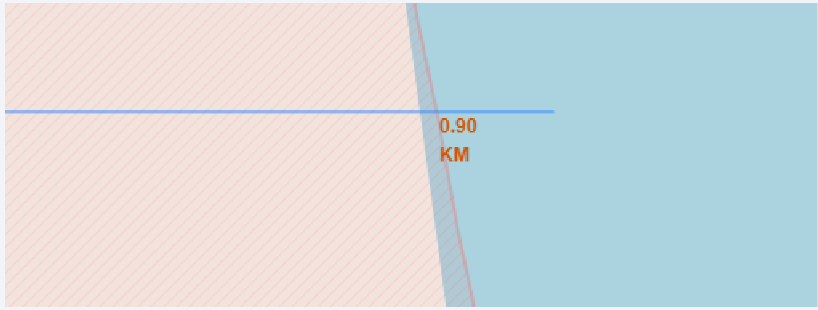


California Launch Sites

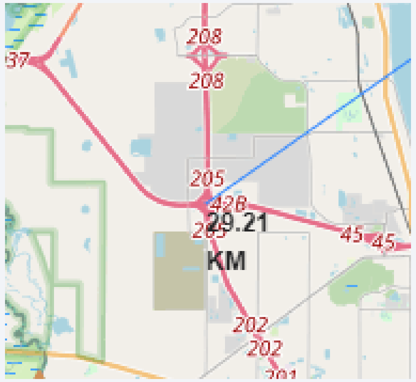
Green Markers: Success **Red Markers: Failure**

Launch Site distance to landmarks

Distance to Coast Line: 0.90km

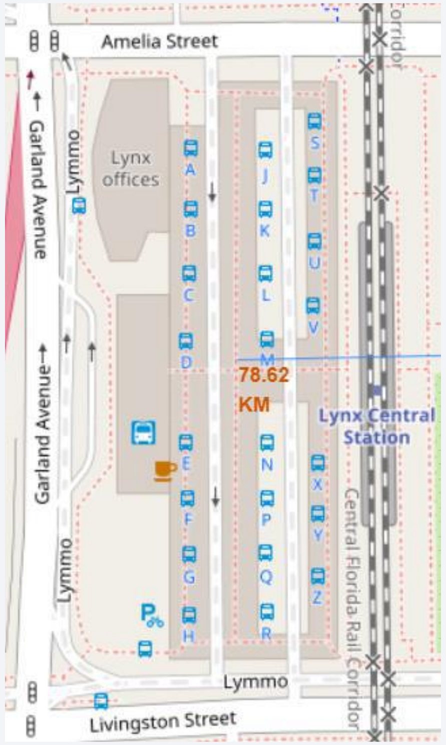


Distance to Highway: 29.21km



Distance to Florida City 78.45 km

Distance to Railway Station: 78.62km



Are launch sites in close proximity to railways? No

Are launch sites in close proximity to highways? No

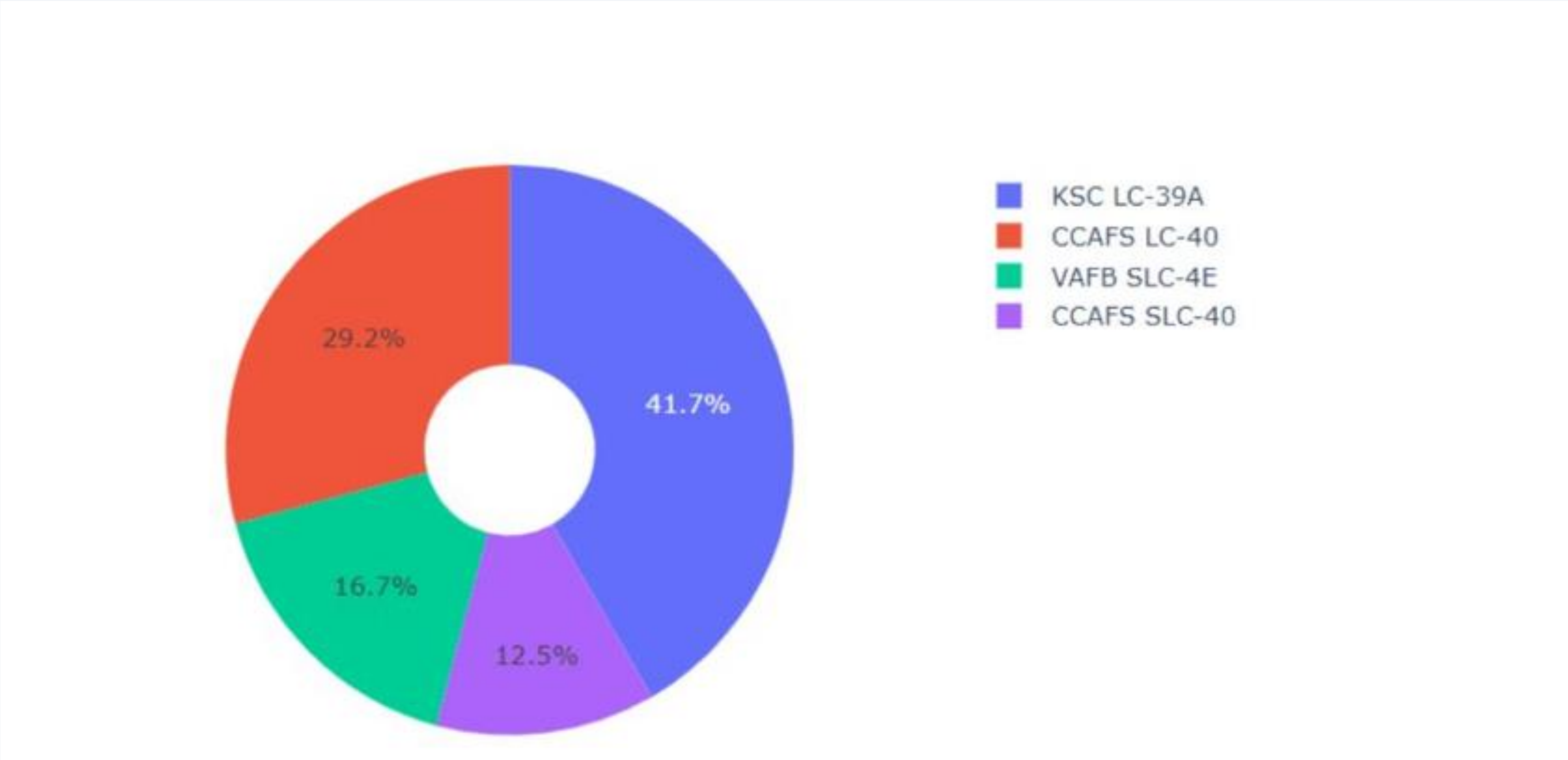
Are launch sites in close proximity to coastline? Yes

Do launch sites keep certain distance away from cities? Yes

Section 5

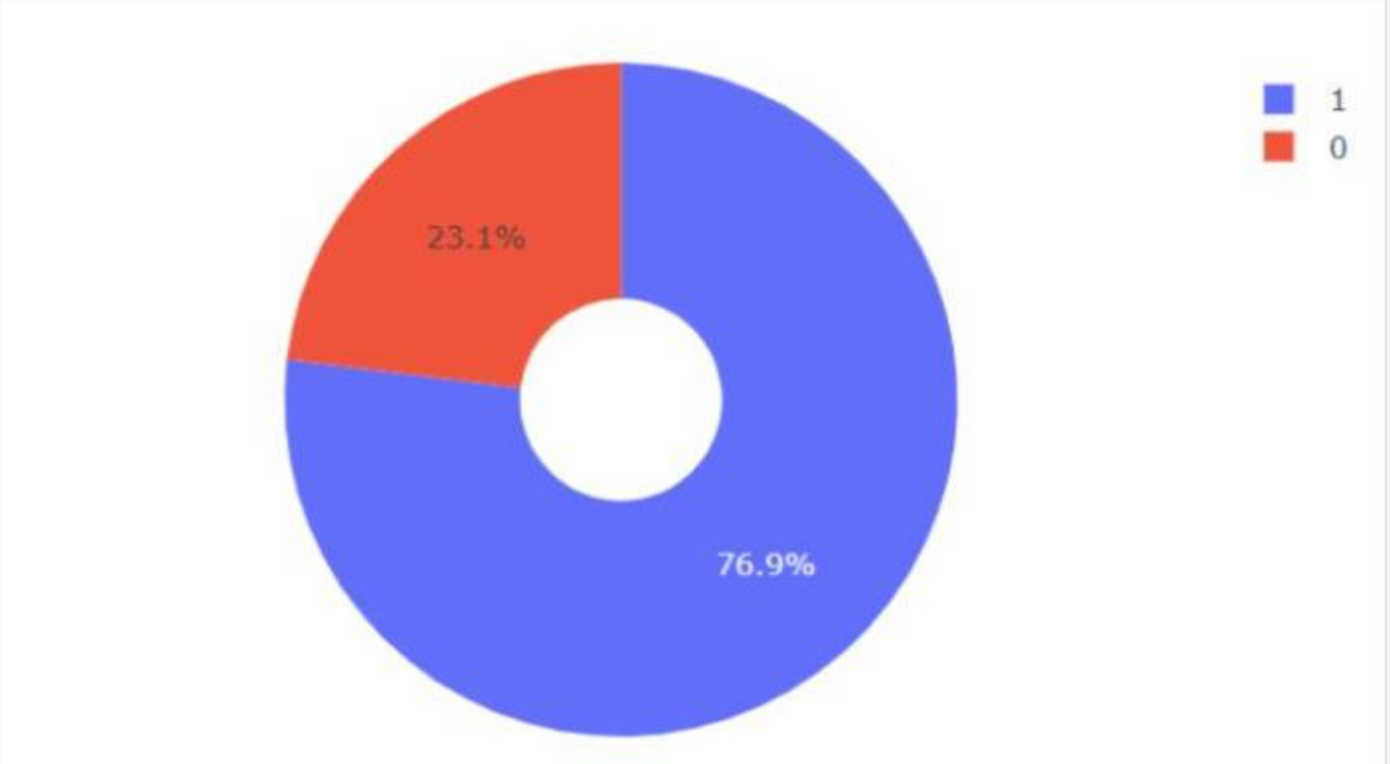
Build a Dashboard with Plotly Dash

Total Success Launch by All Sites



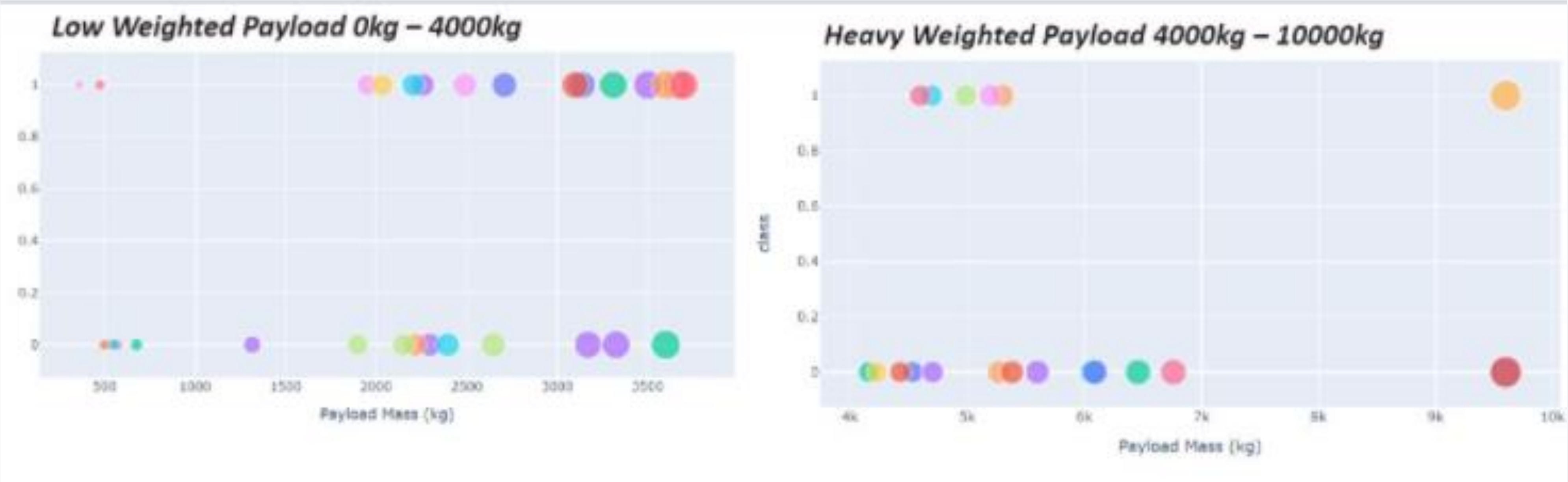
KSC LC-39A has the highest successful launch from all sites

Success Rate of KSC LC-39A



- The success rate of KSC LC-39A is 76.9%

Low Weighted Payload Vs Heavy Weighted Payload



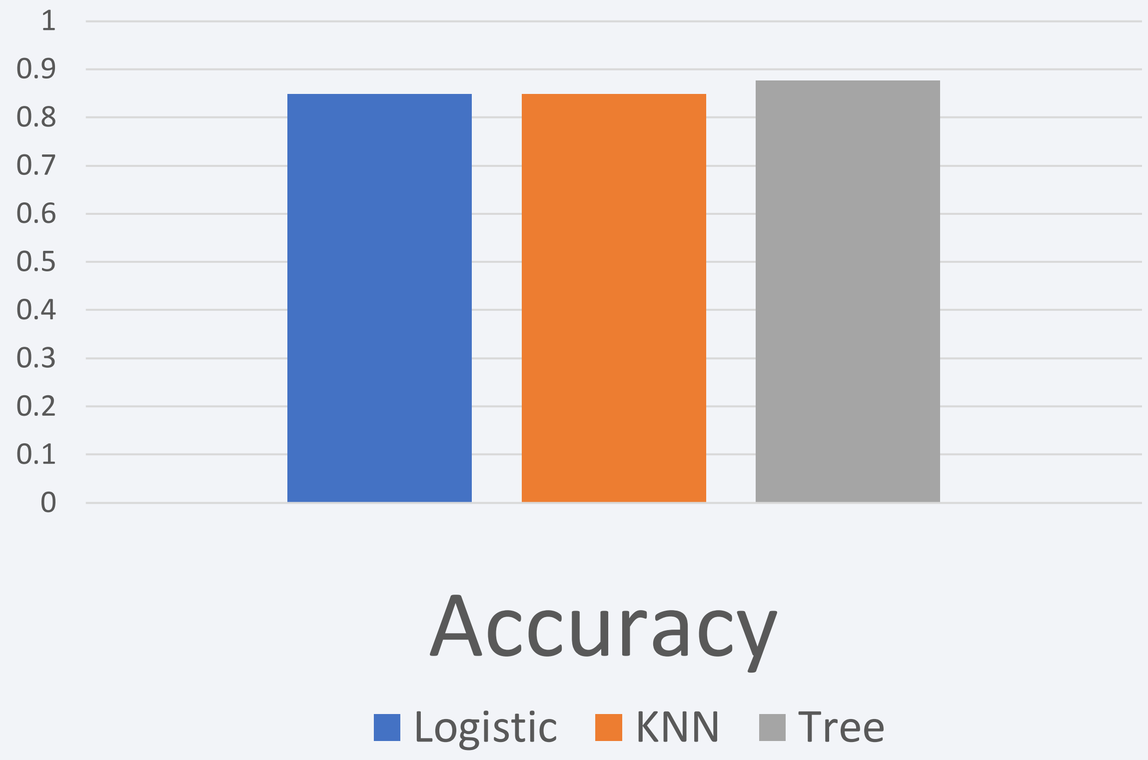
- Success rate of low weight payload is higher than that of heavy weighted ones

Section 6

Predictive Analysis (Classification)

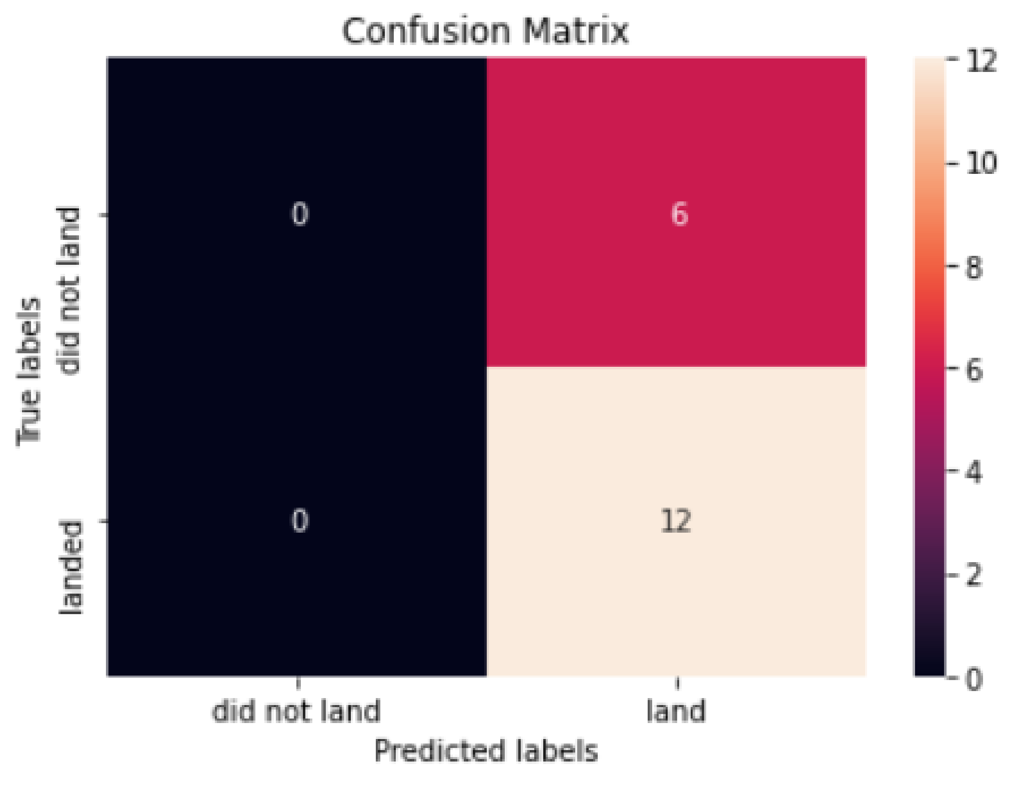
Classification Accuracy

- The decision tree classifier is the model with the highest classification accuracy



Confusion Matrix

- The best performing model is Tree
- Examining the confusion matrix, we see that Tree can distinguish between the different classes. But the major problem is false positives.



Conclusions

- The larger the flight amount at a launch site, the greater the success rate at a launch site.
- Launch success rate started to increase from 2013 to 2020.
- Orbits ES L1, GEO, HEO, SSO, VLEO had the most number of success rate.
- KSC LC 39A had the most successful launches of any sites.
- The Decision tree classifier is the best machine learning algorithm for this task.

Thank you!