IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Eric Rivera
November 11, 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 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.
- Problems you want to find answers
  - We will determine whether the First Stage of SpaceX launches can be recovered or crash.
  - We will determine the price of each launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected with information from Wikipedia and various API.
- 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

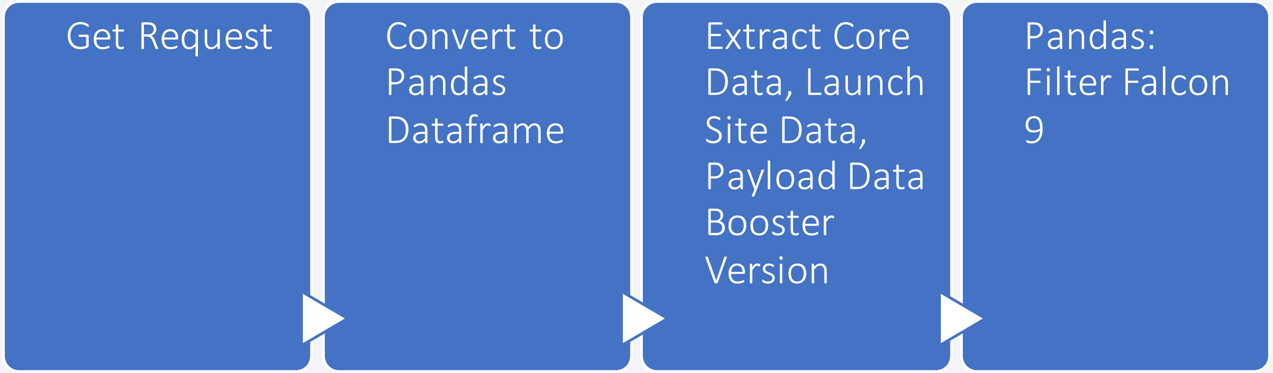
Data Collection

- The data was collected using various methods
  - Data collection was done using get request to the SpaceX API.
  - Next, we decoded the response content as a Json using .json() function call and turn it into a pandas dataframe using .json_normalize().
  - We then cleaned the data, checked for missing values and fill in missing values where necessary.
  - In addition, we performed web scraping from Wikipedia for Falcon 9 launch records with BeautifulSoup.
  - The objective was to extract the launch records as HTML table, parse the table and convert it to a pandas dataframe for future analysis.

Data Collection – SpaceX API

- We used the Get request to the SpaceX API to collect data, clean the requested data and did some basic data wrangling and formatting.

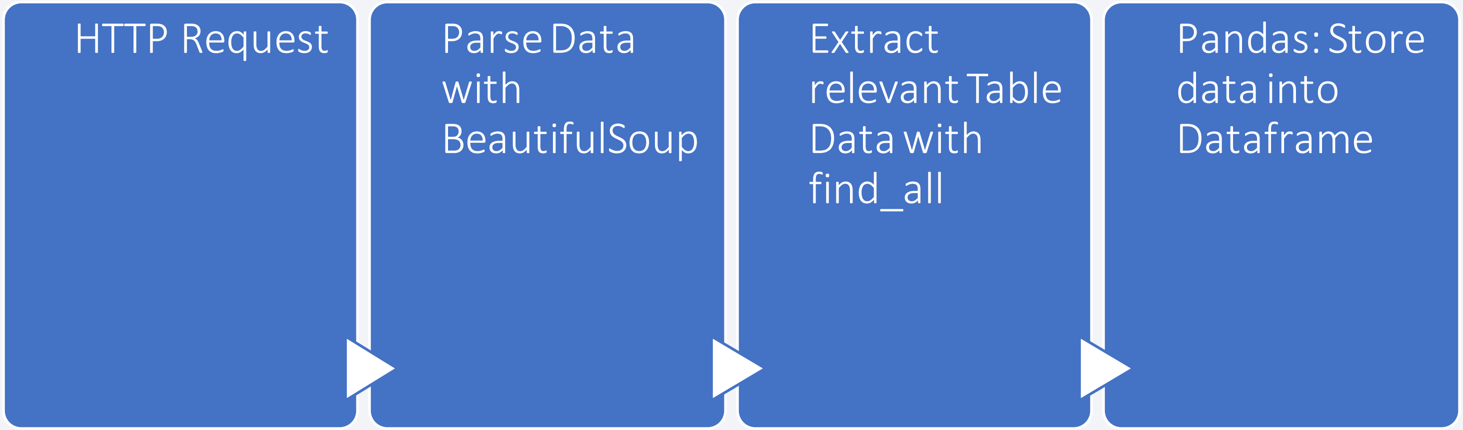
- SpaceX API calls notebook:
  - https://github.com/ericrivera81/SpaceX-Capstone/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb



Data Collection - Scraping

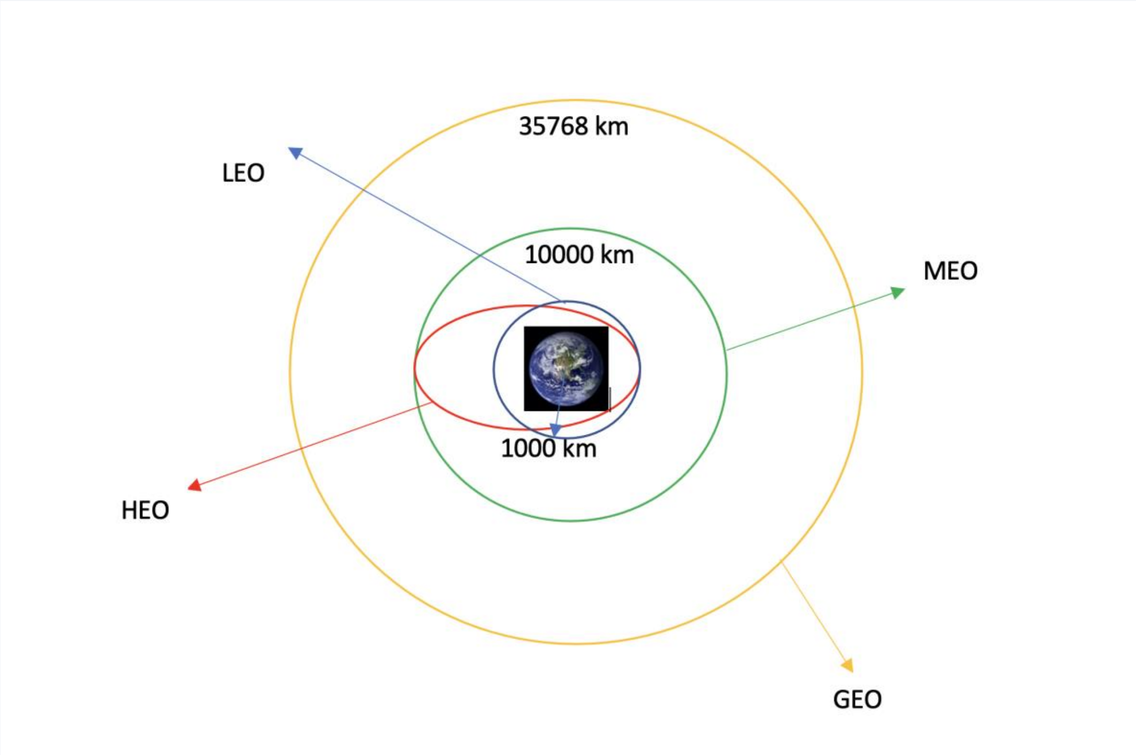
- Present your web scraping process using key phrases and flowcharts

- Web Scraping notebook: https://github.com/ericrivera81/SpaceX-Capstone/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb



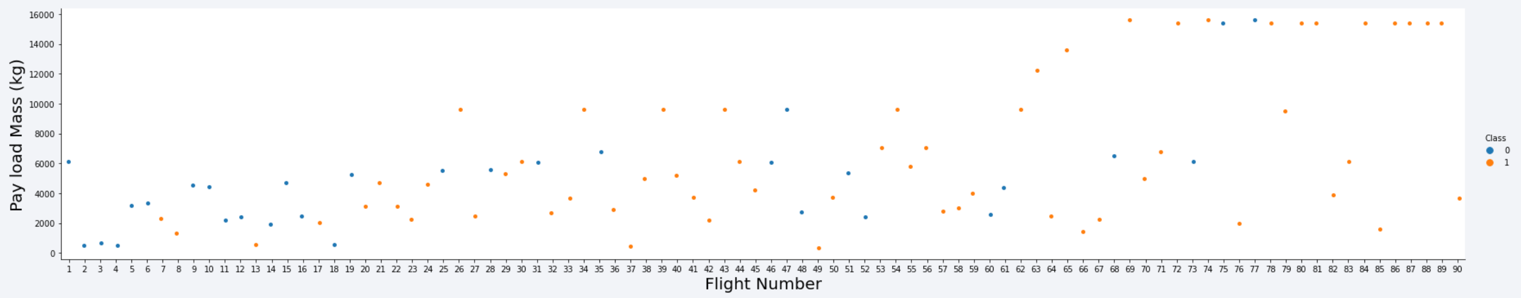
Data Wrangling

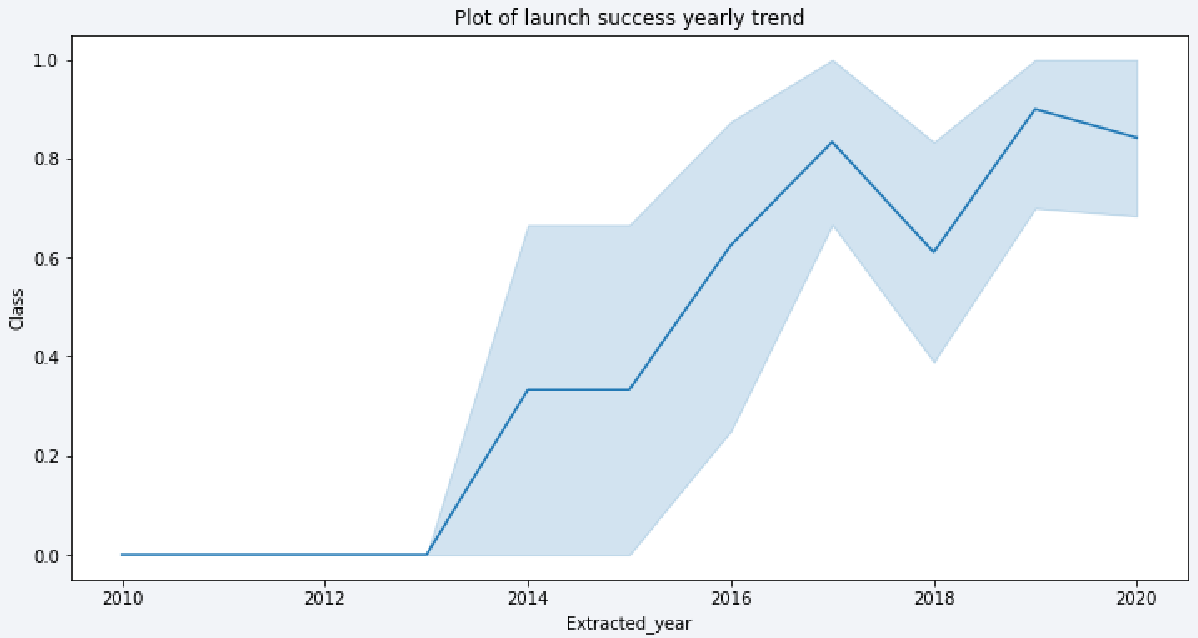
- Performed exploratory data and determined training labels.
- Calculated the number of launches at each site and the number and occurrence of each orbit.
- Created landing outcome label from outcome colomn and exported the results to csv.
- Data Wrangling Notebook: https://github.com/ericrivera81/SpaceX-Capstone/blob/main/Data%20Wrangling.ipynb



EDA with Data Visualization

- We performed exploratory Data Analysis and Feature Engineering using Pandas and Matplotlib
- Exploratory Data Analysis
- Preparing Data Feature Engineering
- EDA with Data Visualization Notebook: https://github.com/ericrivera81/SpaceX-Capstone/blob/main/EDA%20with%20Data%20Visualization.ipynb





EDA with SQL

- We applied EDA with SQL to get data insight. We wrote queries to find out for the instance:
  - The names of unique launch sites
  - Total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1
  - Total number of successful and failed mission outcomes
  - Failed landing outcomes, booster version and launch site names
- EDA with SQL Notebook: https://github.com/ericrivera81/SpaceX-Capstone/blob/main/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

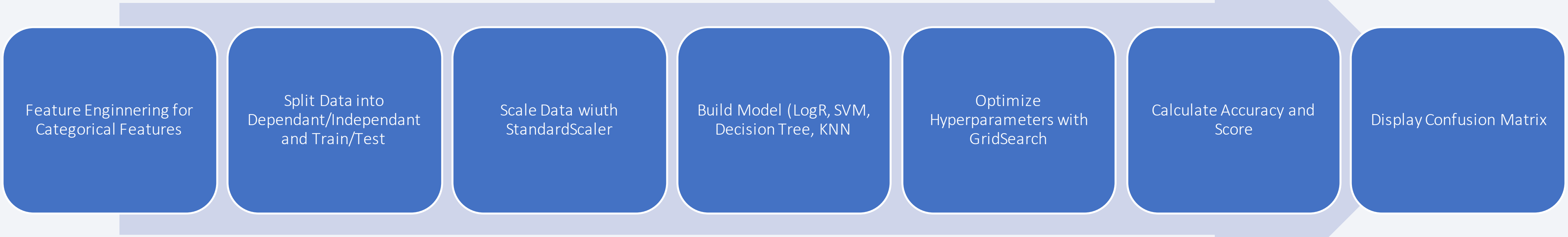
- Map Objects:
  - Edged circles with 1000m radius from Launch Sites
  - Markers for labeling all objects
  - Marker Cluster for creating a bunch of markers around launch sites to indicate success of failure of first stage launches
  - Lines to measure the distances between launch sites and the nearest city
- Interactive Map with Folium Notebook: https://github.com/ericrivera81/SpaceX-Capstone/blob/main/module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- We built a dashboard using Plotly dash
- We plotted pie charts showing the total launches by specific sites
- We also plotted a graph showing the relationship with Outcome and Payload Mass for the different booster versions
- Dashboard with Plotly App: https://github.com/ericrivera81/SpaceX-Capstone/blob/main/app.py

Predictive Analysis (Classification)

- We loaded the data using NumPy and Pandas, transformed the data, and split the data into training and testing



- Machine Learning Prediction Notebook: https://github.com/ericrivera81/SpaceX-Capstone/blob/main/module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

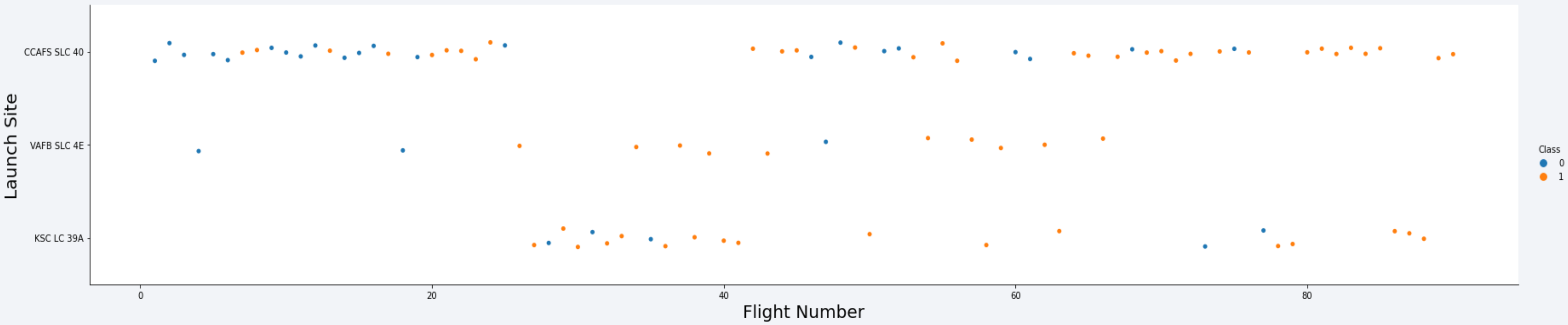
Results

- Exploratory data analysis results
  - Launch success rates increases over time
  - Higher success rates for higher orbits
- Interactive analytics demo in screenshots
  - Higher success rates for higher payload mass
- Predictive analysis results
  - Best results with Logistic Regression and Support Vector Machine

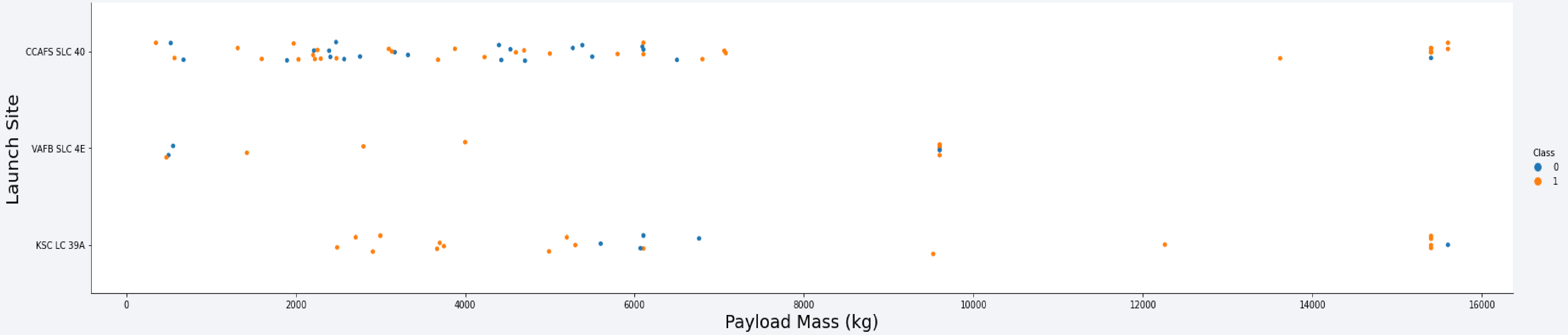
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site



Payload vs. Launch Site

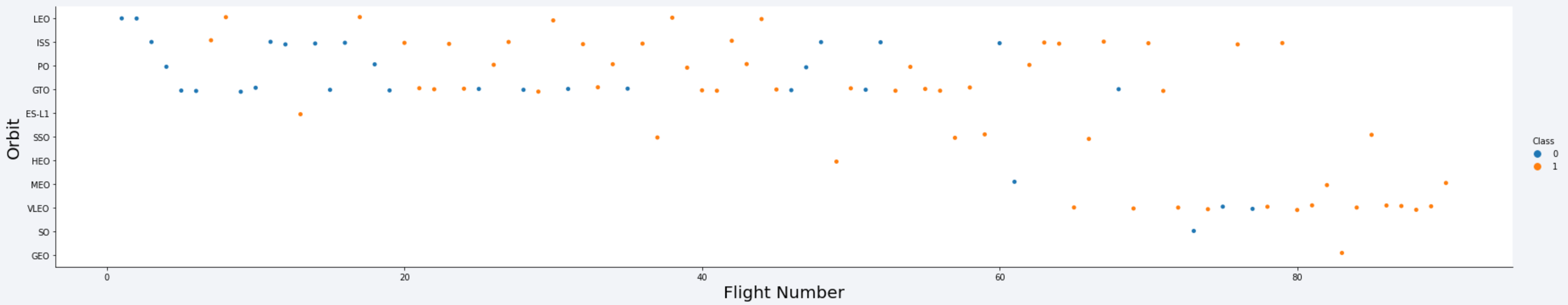


- Payloads that approach MAX(Payload) tended to launch from CCAFS SLC 40 & KSC LC 39A
- Payloads less than 8000 kg tended to fail at a higher rate when launched from CCAFS SLC 40, plausibly due to that launch site being used for R&D versus the other two launch sites used with less failure-tolerant payloads.

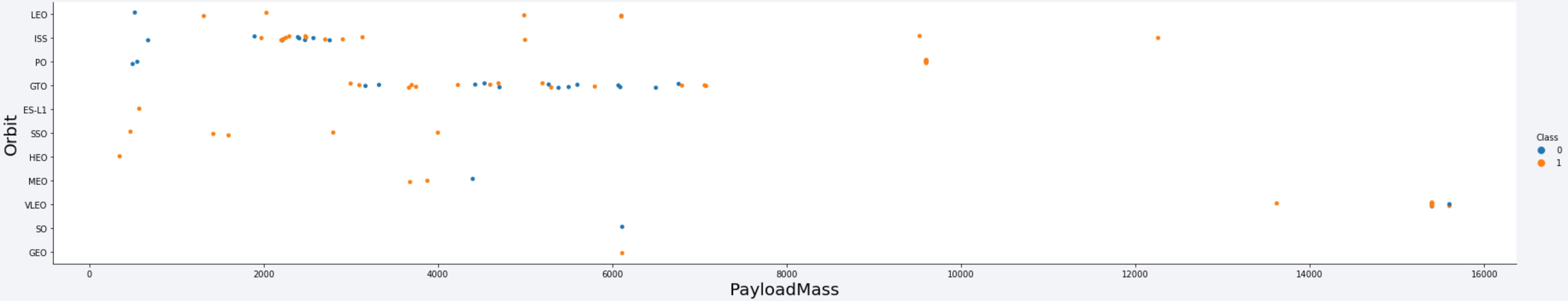
Success Rate vs. Orbit Type

- The greater the payload mass for launch site CCAFS SLC 40, the higher the success rate for the orbit.

Flight Number vs. Orbit Type

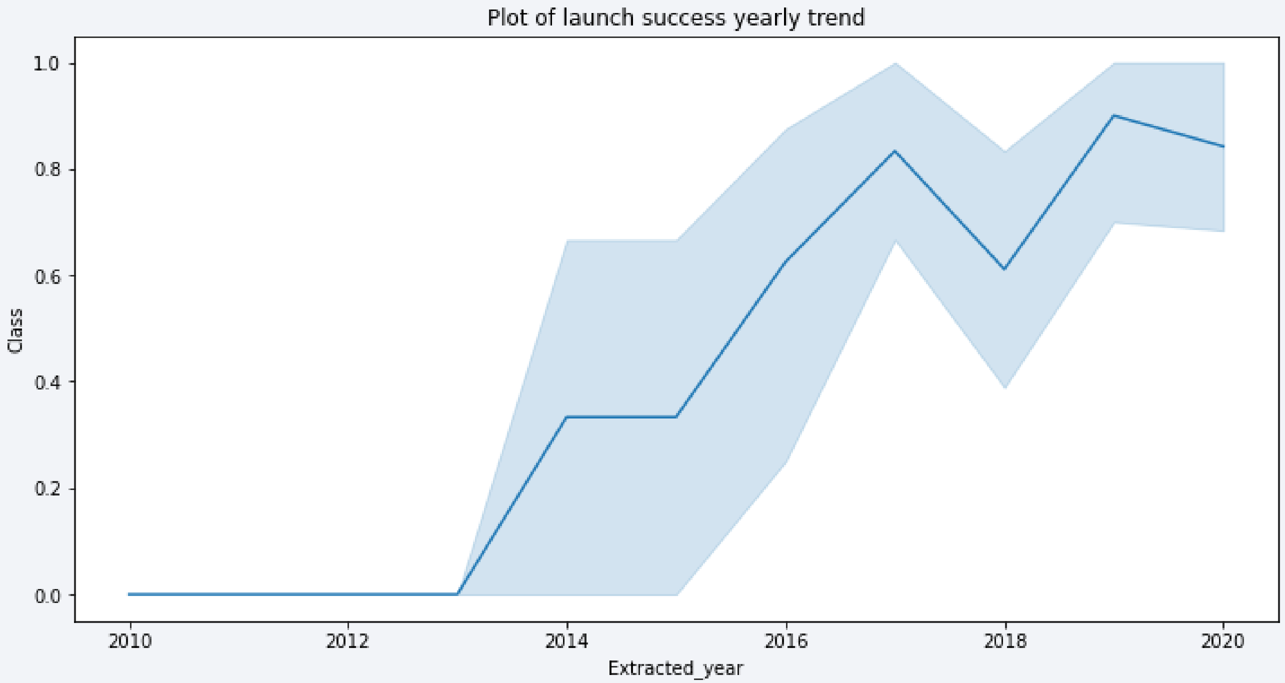


Payload vs. Orbit Type



Launch Success Yearly Trend

- Success rates have increased since 2013 to 2020.



All Launch Site Names

- CCA – Cape Canaveral Launch Center
- KSC – Kennedy Space Center
- VAFB – Vandenburg Air Force Base

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites' names start with `CCA`

In [11]:

```
%sql SELECT LAUNCH_SITE from SPACEXTBL where (LAUNCH_SITE) LIKE 'CCA%' LIMIT 5;
```

```
 * ibm_db_sa://yfv48336:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32716/bludb
Done.
```

Out[11]:

launch_site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

Display the total payload mass carried by boosters launched by NASA (CRS)

```
In [29]: %sql select sum(PAYLOAD_MASS__KG_) as payloadmass from SPACEXTBL where Customer LIKE 'NASA (CRS)';

 * ibm_db_sa://yfv48336:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32716/bludb
Done.
```

Out[29]:

payloadmass
22007

Average Payload Mass by F9 v1.1

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

```
%sql select avg(PAYLOAD_MASS__KG_) as payloadmass from SPACEXTBL
```

```
 * ibm_db_sa://yfv48336:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32716/bludb
Done.
```

payloadmass
5692

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on drone ship.Present your query result with a short explanation here

```
%sql select min(DATE) from SPACEXTBL;

 * ibm_db_sa://yfv48336:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32716/bludb
Done.
```

1
2010-04-06

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

```
%sql select BOOSTER_VERSION from SPACEXTBL where LANDING__OUTCOME='Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000;

 * ibm_db_sa://yfv48336:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32716/bludb
Done.
```

booster_version
F9 FT B1022
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

```
%sql select count(MISSION_OUTCOME) as missionoutcomes from SPACEXTBL GROUP BY MISSION_OUTCOME;
```

```
 * ibm_db_sa://yfv48336:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32716/bludb
Done.
```

missionoutcomes
44
1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

```
%sql select BOOSTER_VERSION as boosterversion from SPACEXTBL where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTBL);
```

```
 * ibm_db_sa://yfv48336:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32716/bludb
Done.
```

boosterversion
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3

2015 Launch Records

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

```
%sql SELECT MONTH(DATE),MISSION_OUTCOME,BOOSTER_VERSION,LAUNCH_SITE FROM SPACEXTBL where EXTRACT(YEAR FROM DATE)='2015';
```

```
 * ibm_db_sa://yfv48336:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32716/bludb
Done.
```

1	mission_outcome	booster_version	launch_site
10	Success	F9 v1.1 B1012	CCAFS LC-40
11	Success	F9 v1.1 B1013	CCAFS LC-40
2	Success	F9 v1.1 B1014	CCAFS LC-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

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

```
%sql SELECT LANDING__OUTCOME FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' ORDER BY DATE DESC;
```

```
 * ibm_db_sa://yfv48336:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32716/bludb
Done.
```

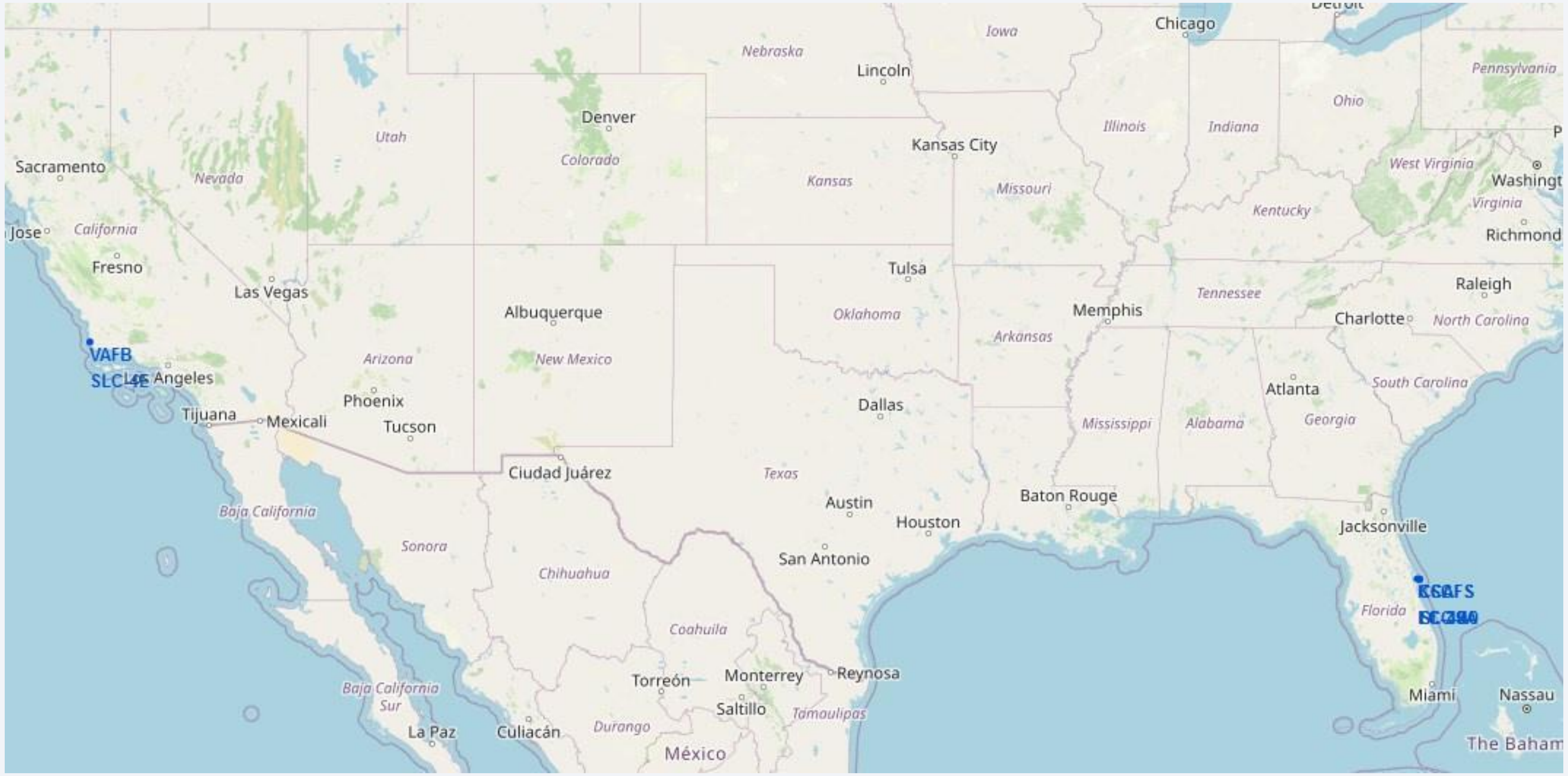
landing__outcome
Success (ground pad)
Success (ground pad)
Success (drone ship)
Success (drone ship)
Failure (drone ship)
Controlled (ocean)
Failure (drone ship)
No attempt
No attempt
No attempt
No attempt
No attempt
No attempt
No attempt
Failure (parachute)

Section 3

Launch Sites Proximities Analysis

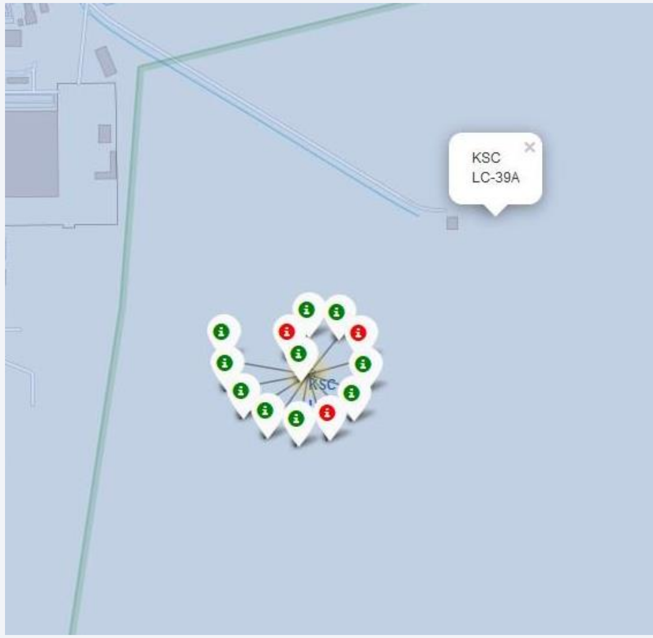
Folium Map Launch Sites

- Launch sites are located in California (West) and Florida (East)

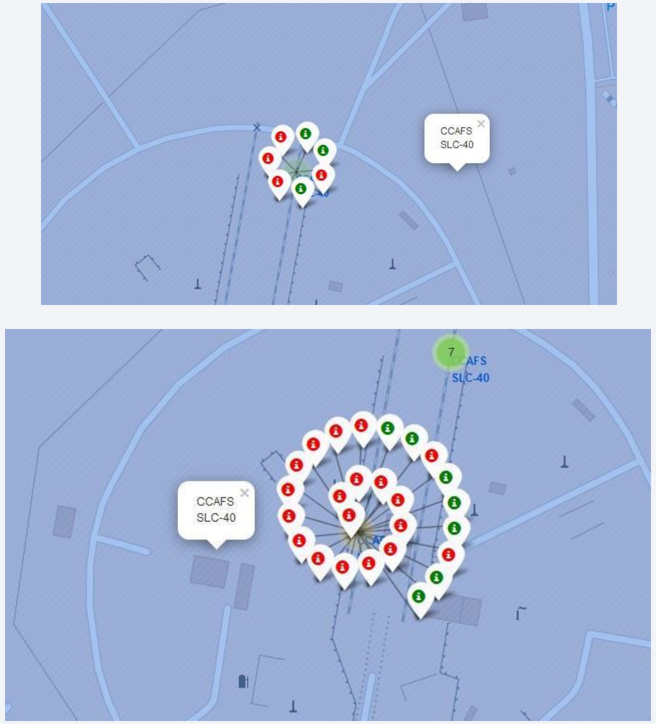


Launch Site Locations' Success and Failures

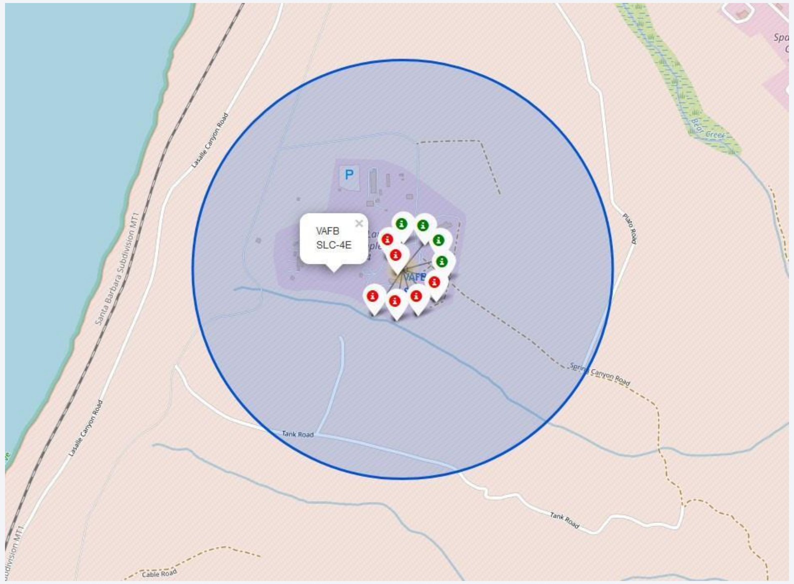
- Green indicates successful launches, meanwhile red means failures



Kennedy Space Center, Florida



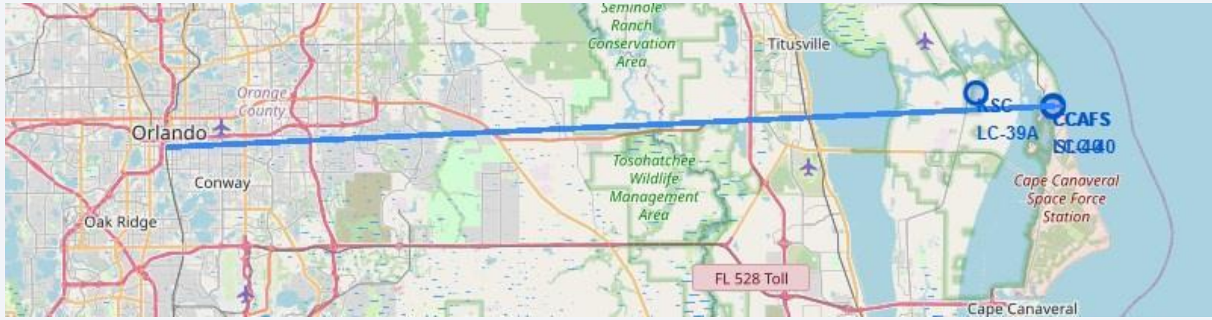
Cape Canaveral, Florida

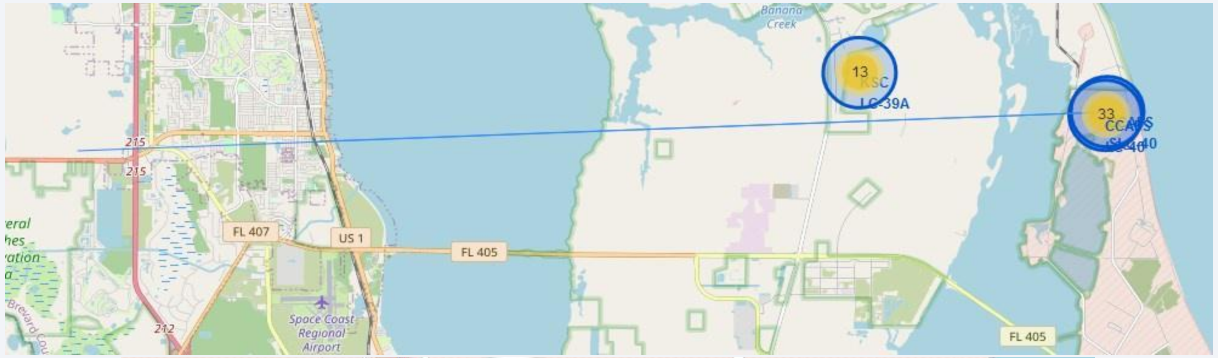


Vandenburg Air Force Base, California

Launch Site Distances to Landmarks

- Cape Canaveral Sites are near Orlando (78.4 KM)
- Launch site to major highway and trains (28.9 KM)
- Launch site to coast (0.9 KM)
- We can determine that major cities, highways and trainlines are safely away from a launch site.
- Coast lines are close and within view of the launch sites.

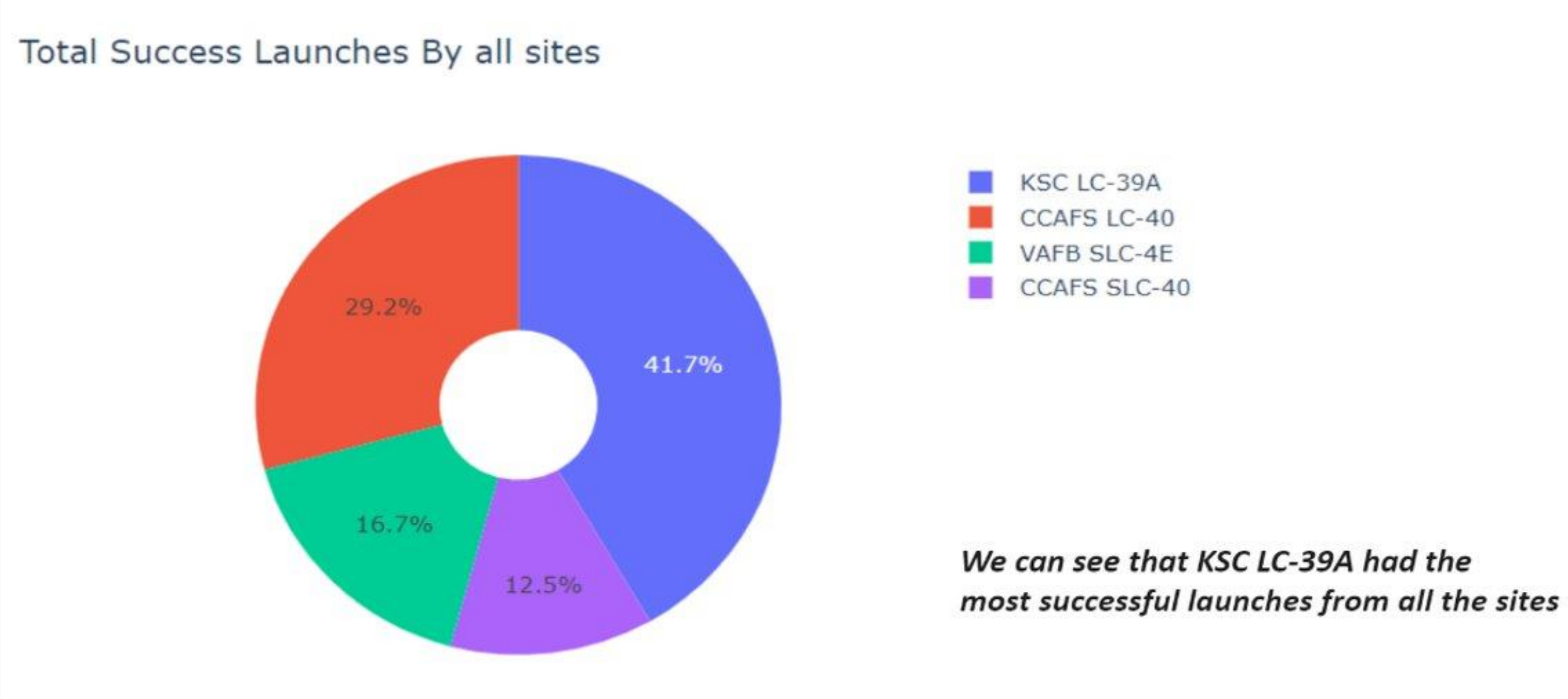




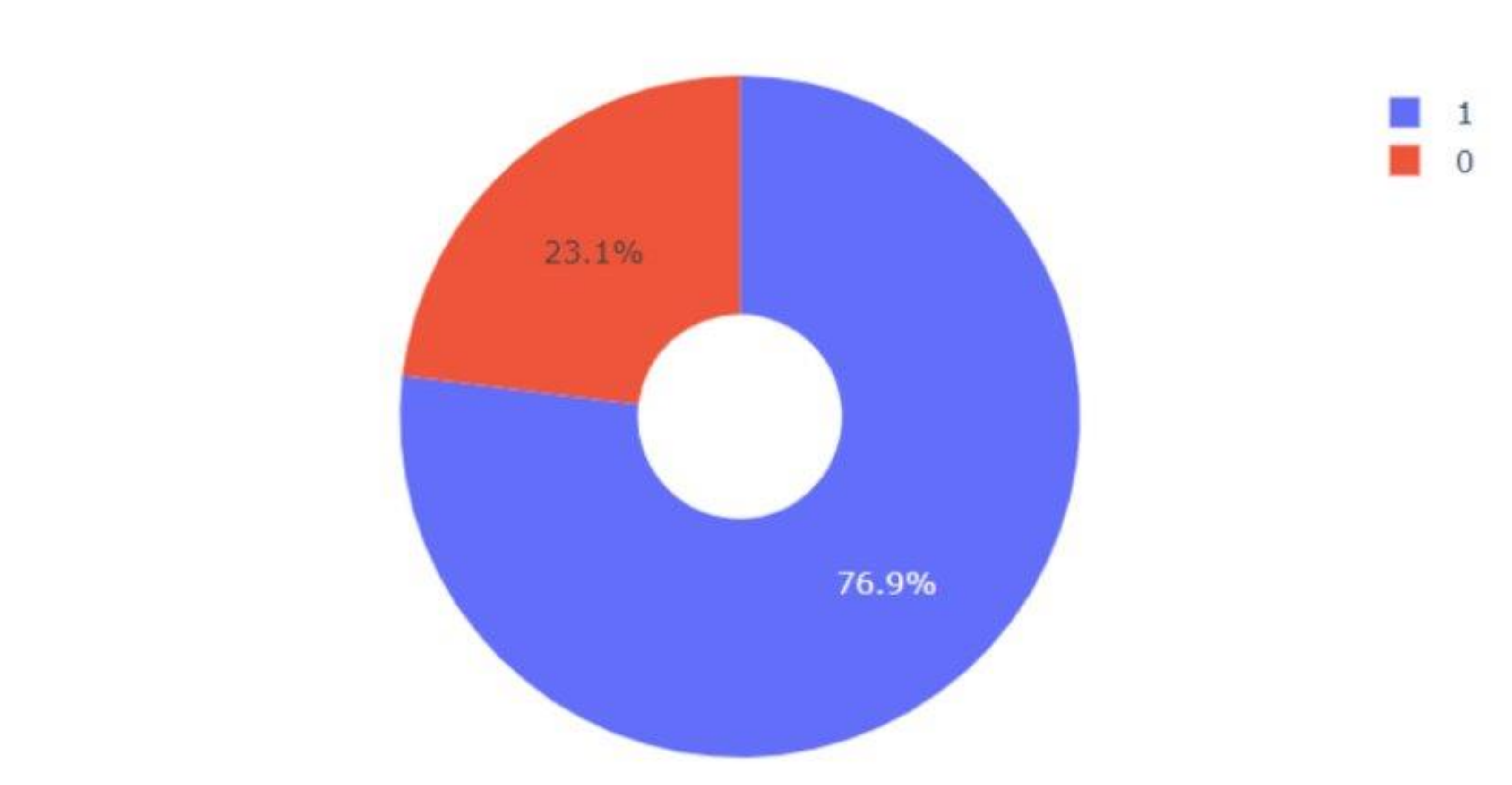
Section 4

Build a Dashboard with Plotly Dash

Total Launch Success Chart

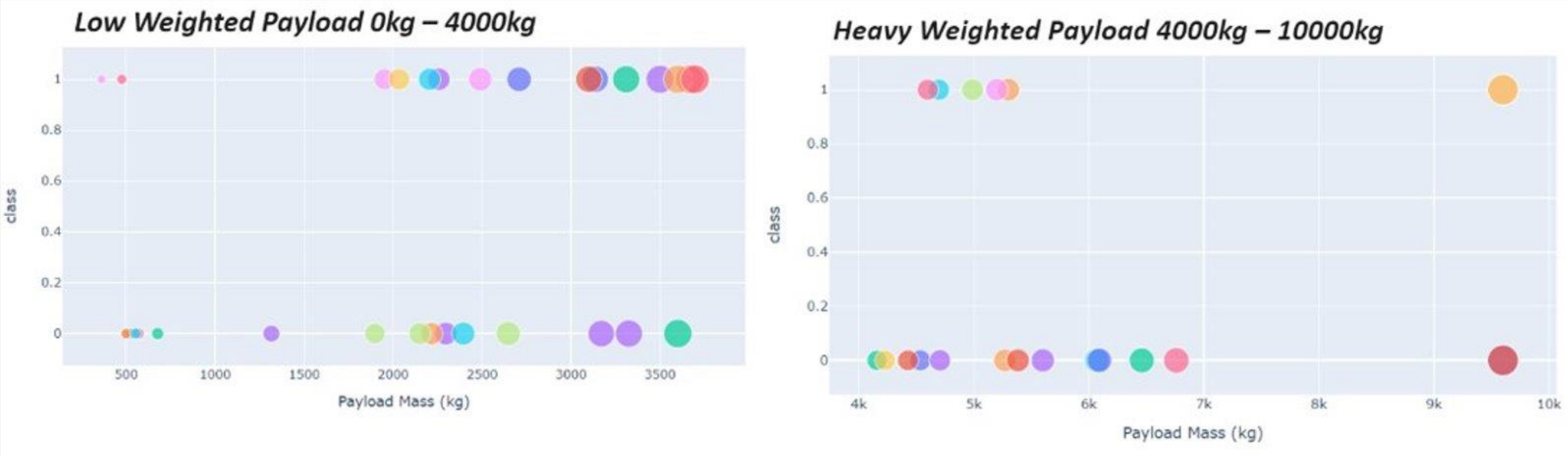


Launch Site with Highest Launch Success



KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

Payload vs Launch Outcome for All Sites



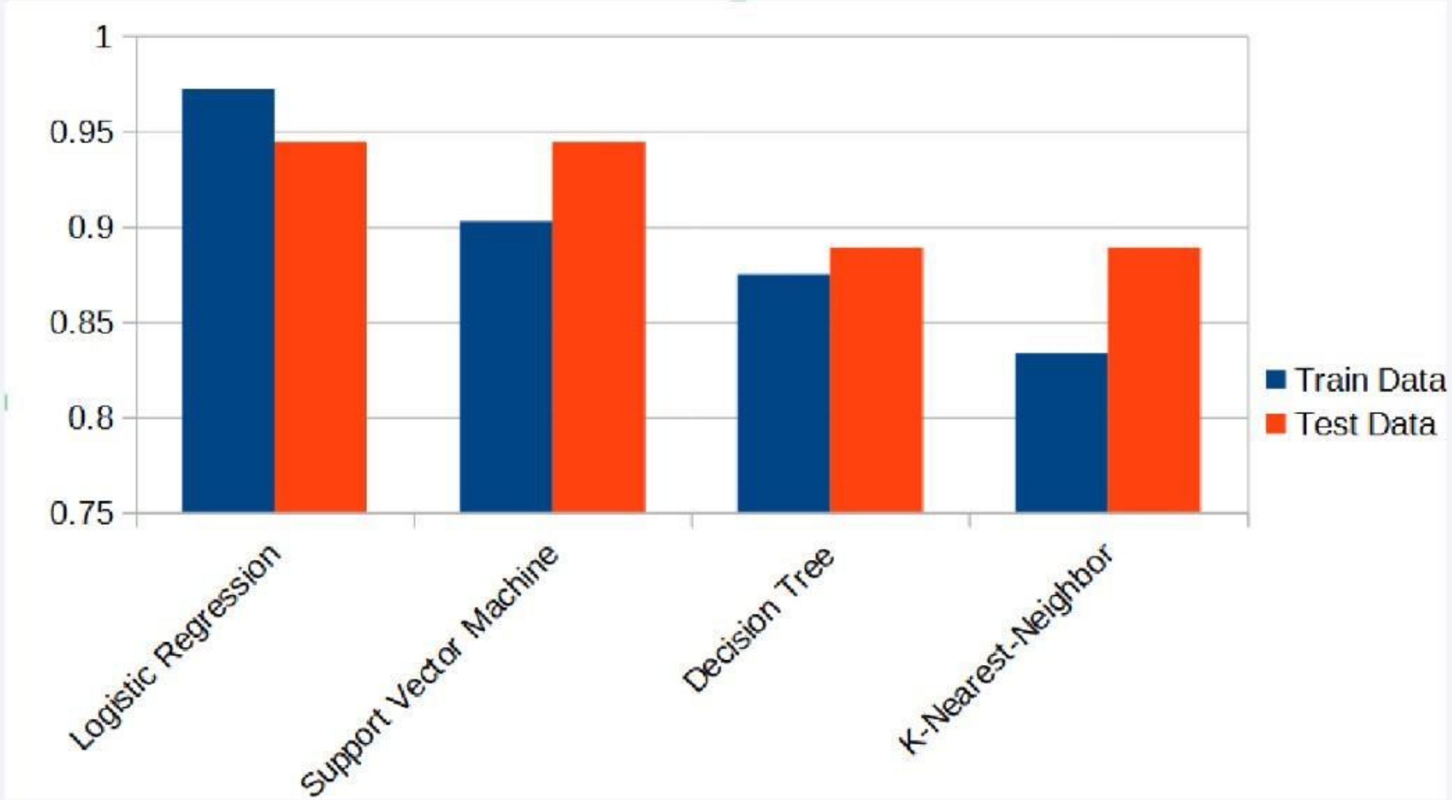
We can see the success rates for low weighted payloads is higher than the heavy weighted payloads

Section 5

Predictive Analysis (Classification)

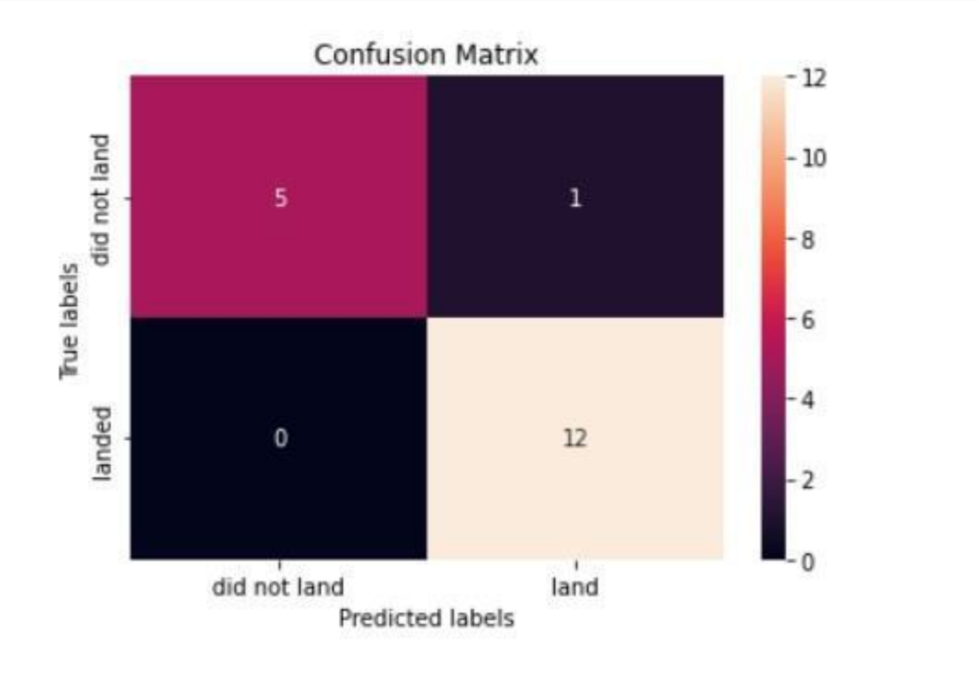
Classification Accuracy

- Logistic Regression has the best results for train data, as well as test data.



Confusion Matrix

- True Positives: 12
- True Negatives: 5
- False Positives: 1
- False Negatives: 0



Conclusions

- The more launches at a specific site, the more successful future launches will be
- Launch success had increased from 2013 until 2020
- We have no false negatives
- Kennedy Space Center is the most successful launch site

Thank you!