IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Andres Chousal>
<5/24/2024>

Outline

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

Executive Summary

- In this presentation, we delve into the methodologies employed and the results obtained through our data analysis efforts. The primary objective was to extract actionable insights from the available datasets to inform decision-making and enhance business strategies.
- Data collection
- Data wrangling
- EDA with data visualization
- EDA with SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly Dash
- Predictive analysis (Classification)

Introduction

- SpaceX, under the visionary leadership of Elon Musk, has reshaped the aerospace industry with its pioneering missions and remarkable achievements. Central to SpaceX's legacy are its unparalleled success rates in both launching payloads into orbit and, perhaps equally impressive, landing reusable rockets back on Earth.
- Equally noteworthy is SpaceX's revolutionary approach to rocket reusability, epitomized by the successful landings of Falcon 9 boosters on drone ships or landing pads. These landings represent a monumental leap forward in the quest for cost-effective space travel, as they enable the reuse of expensive rocket components, thereby driving down the overall cost of access to space.
- We intend to analyze the patterns to determine what makes up a successful flight.

Section 1

Methodology

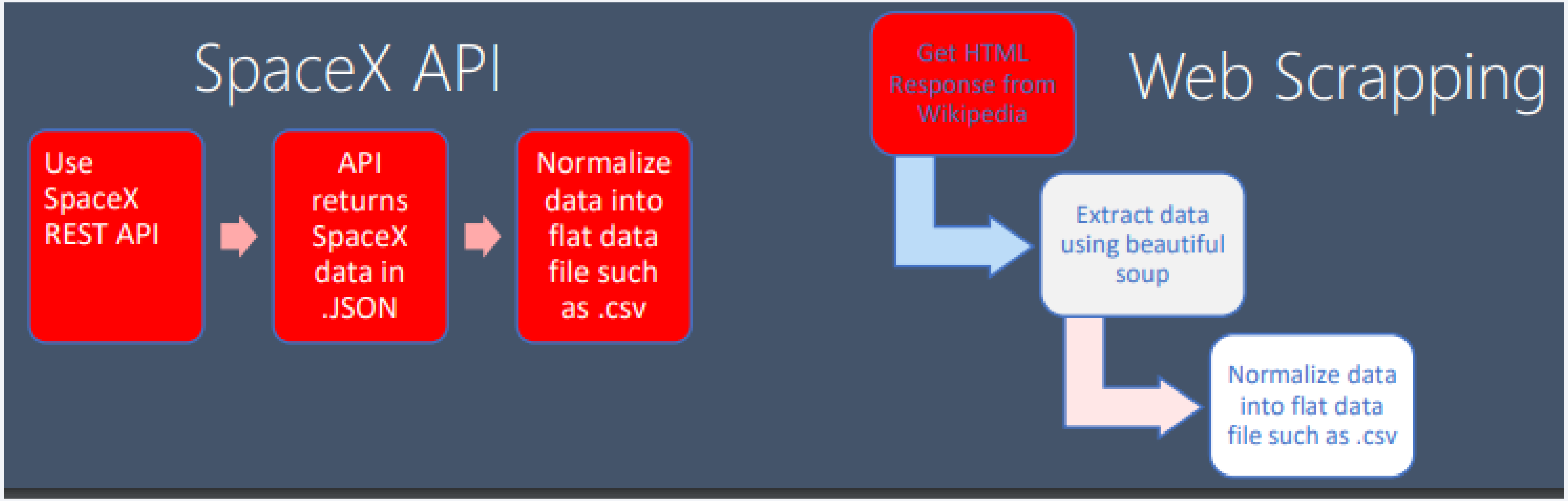
Methodology

Executive Summary

- Data collection methodology:
  - SpaceX Rest API
  - Web Scrapping
- Perform data wrangling
  - One Hot Encoding data fields
- 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

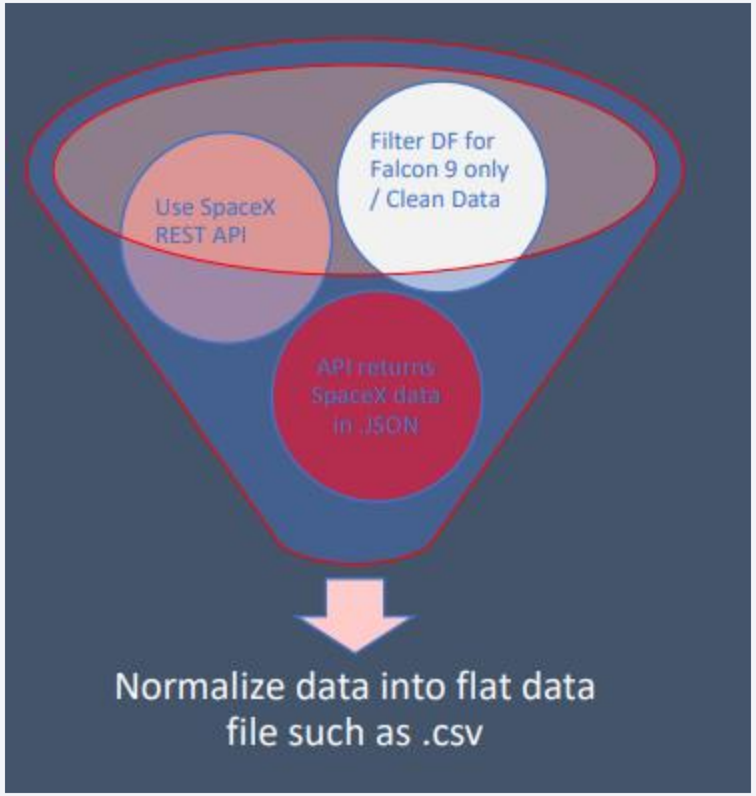
Data Collection

- Data sets were collected using SpaceX REST API
- Web scraping is another means to gathering data through Wikipedia



Data Collection – SpaceX API

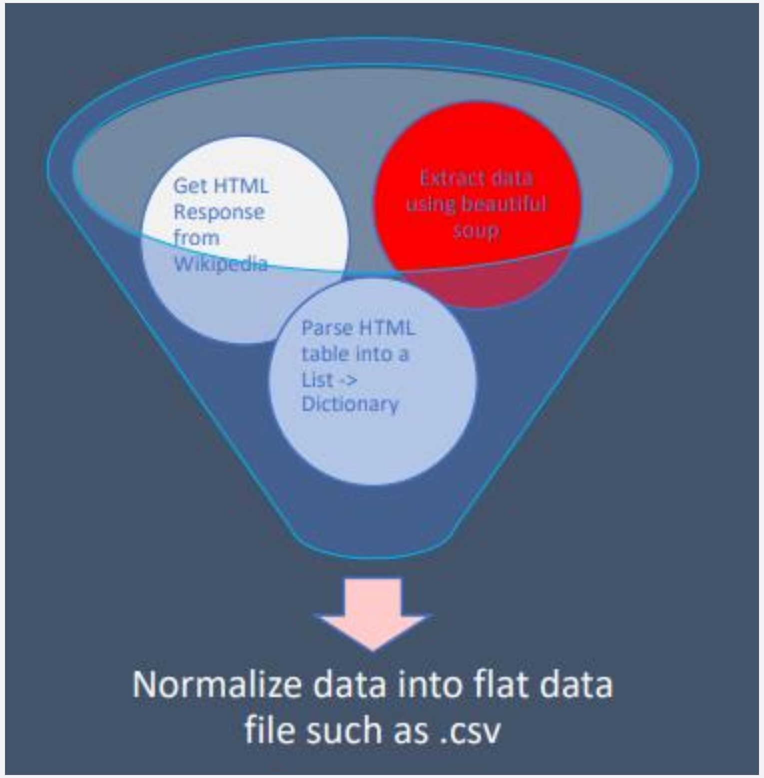
1. Get response from API
2. Convert response to a .json file
3. Apply Custom functions to clean data
4. Assign list to dictionary and dataframe
5. Filter data frame and export to .csv



GitHub link

Data Collection - Scraping

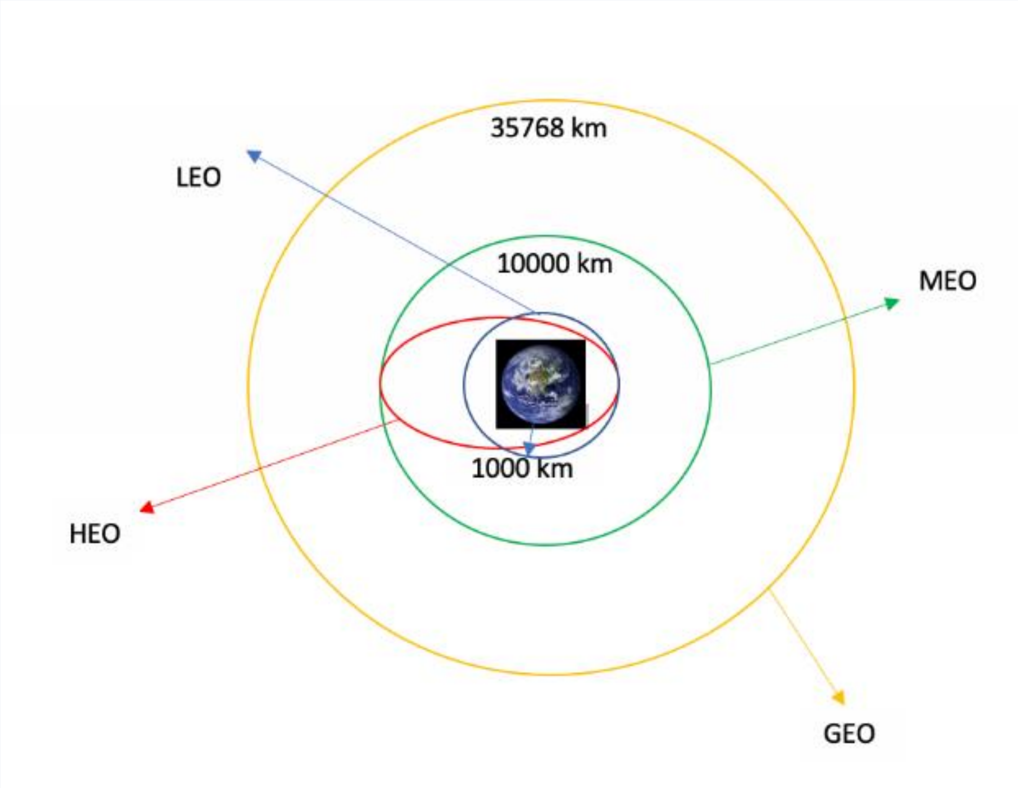
1. Getting Response from HTML
2. Creating BeautifulSoup Object
3. Finding tables
4. Getting column names
5. Creation of dictionary
6. Appending data to keys
7. Converting dictionary to dataframe
8. Dataframe to .CSV



[GitHub link]

Data Wrangling

- After loading the data collected, it is necessary to preprocess the information to fill in any missing values.
- Using the following data
  - Number of launches on each site
  - Number and occurrence of each orbit
- Calculate the number and occurrence of mission outcome of the orbits
- Create a landing outcome label from Outcome column (to a 0,1 class)



EDA with Data Visualization

- Scatter Graphs being drawn:
  - Flight Number VS. Payload Mass
  - Flight Number VS. Launch Site
  - Payload VS. Launch Site
  - Orbit VS. Flight Number
  - Payload VS. Orbit Type
  - Orbit VS. Payload Mass
- Bar Graph being drawn:
  - Mean VS. Orbit
- Line Graph being drawn:
  - Success Rate VS. Year

Graph type depends on the data's variables and desired relationship analysis.

[GitHub link]

EDA with SQL

- Performed SQL queries to gather information about the dataset.
- For example of some questions we were asked about the data we needed information about. Which we are using SQL queries to get the answers in the dataset :
    - Displaying the names of the unique launch sites in the space mission
    - Displaying 5 records where launch sites begin with the string CCA'
    - Displaying the total payload mass carried by boosters launched by NASA (CRS)
    - Displaying average payload mass carried by booster version F9 v1.1
    - Listing the date where the successful landing outcome in drone ship was achieved.
    - Listing the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000
    - Listing the total number of successful and failure mission outcomes
    - Listing the names of the booster_versions which have carried the maximum payload mass.
    - Listing the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
    - Ranking the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order

GitHub link

Build an Interactive Map with Folium

- To visualize the Launch Data into an interactive map. We took the Latitude and Longitude Coordinates at each launch site and added a Circle Marker around each launch site with a label of the name of the launch site.
- We assigned the dataframe launch_outcomes(failures, successes) to classes 0 and 1 with Green and Red markers on the map in a MarkerCluster()
- Using Haversine's formula we calculated the distance from the Launch Site to various landmarks to find various trends about what is around the Launch Site to measure patterns. Lines are drawn on the map to measure distance to landmarks

GitHub link

Build a Dashboard with Plotly Dash

- Plotly Dash helped us create an interactive app to better visualize information through custom graphs.
  - Pie chart with success launches by site
  - Scatter plot with variable payload mass

GitHub link

Predictive Analysis (Classification)

- **BUILDING MODEL**

Load our dataset into NumPy and Pandas

Transform Data

Split our data into training and test data sets

Check how many test samples we have

Decide which type of machine learning algorithms we want to use

Set our parameters and algorithms to GridSearchCV

Fit our datasets into the GridSearchCV objects and train our dataset.

- **EVALUATING MODEL**

Check accuracy for each model

Get tuned hyperparameters for each type of algorithms

Plot Confusion Matrix

- **IMPROVING MODEL**

Feature Engineering

Algorithm Tuning

- **FINDING THE BEST PERFORMING CLASSIFICATION MODEL**

The model with the best accuracy score wins the best performing model

In the notebook there is a dictionary of algorithms with scores at the bottom of the notebook.

GitHub link

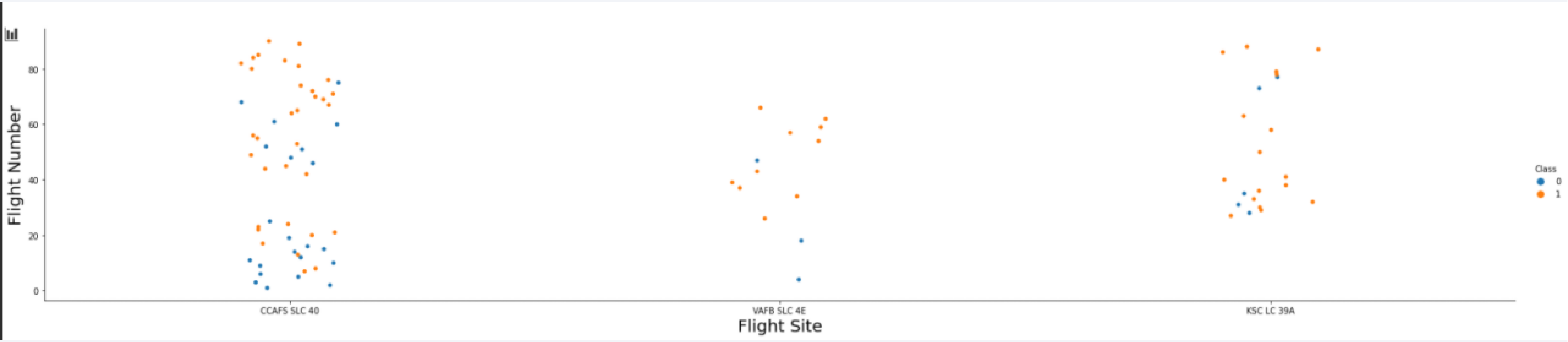
Results

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

Section 2

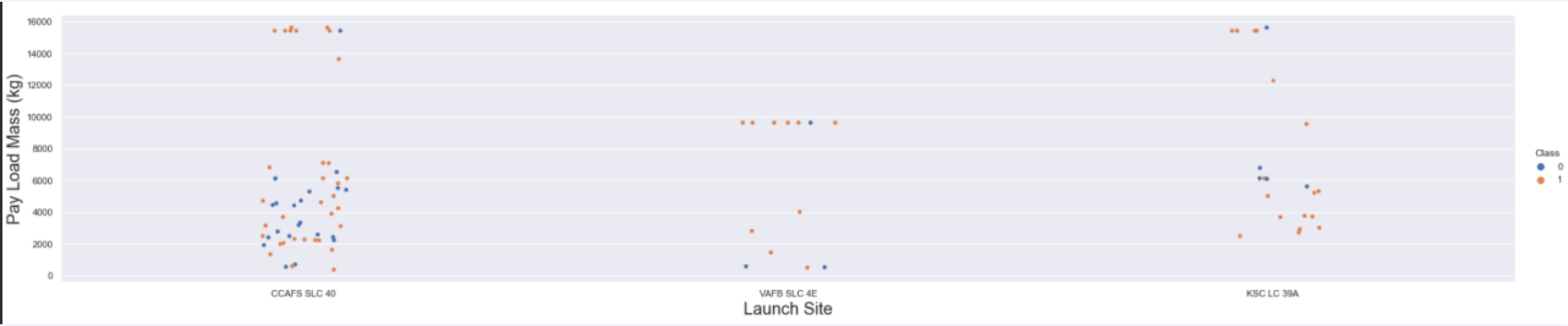
Insights drawn from EDA

Flight Number vs. Launch Site



- Flights on y axis vs Flight site on x axis

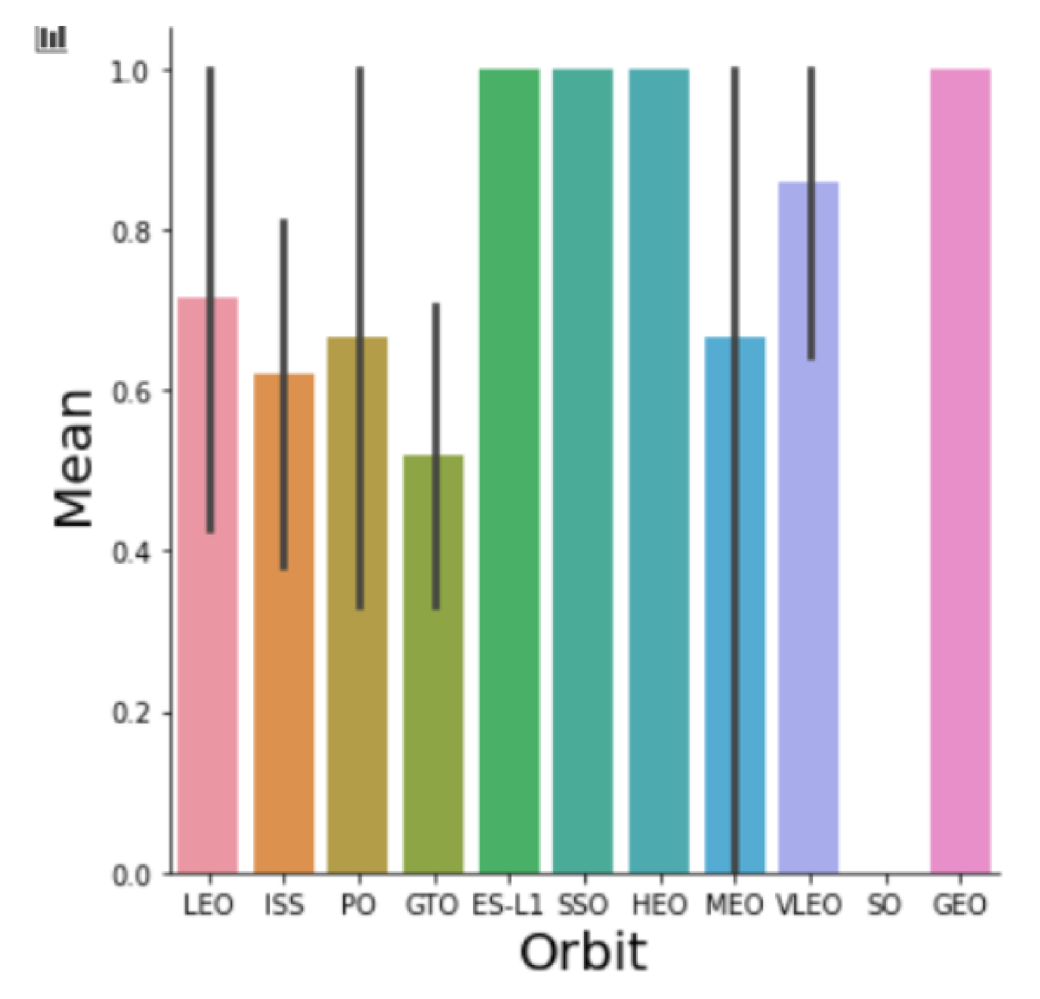
Payload vs. Launch Site



- Higher payload mass means more success

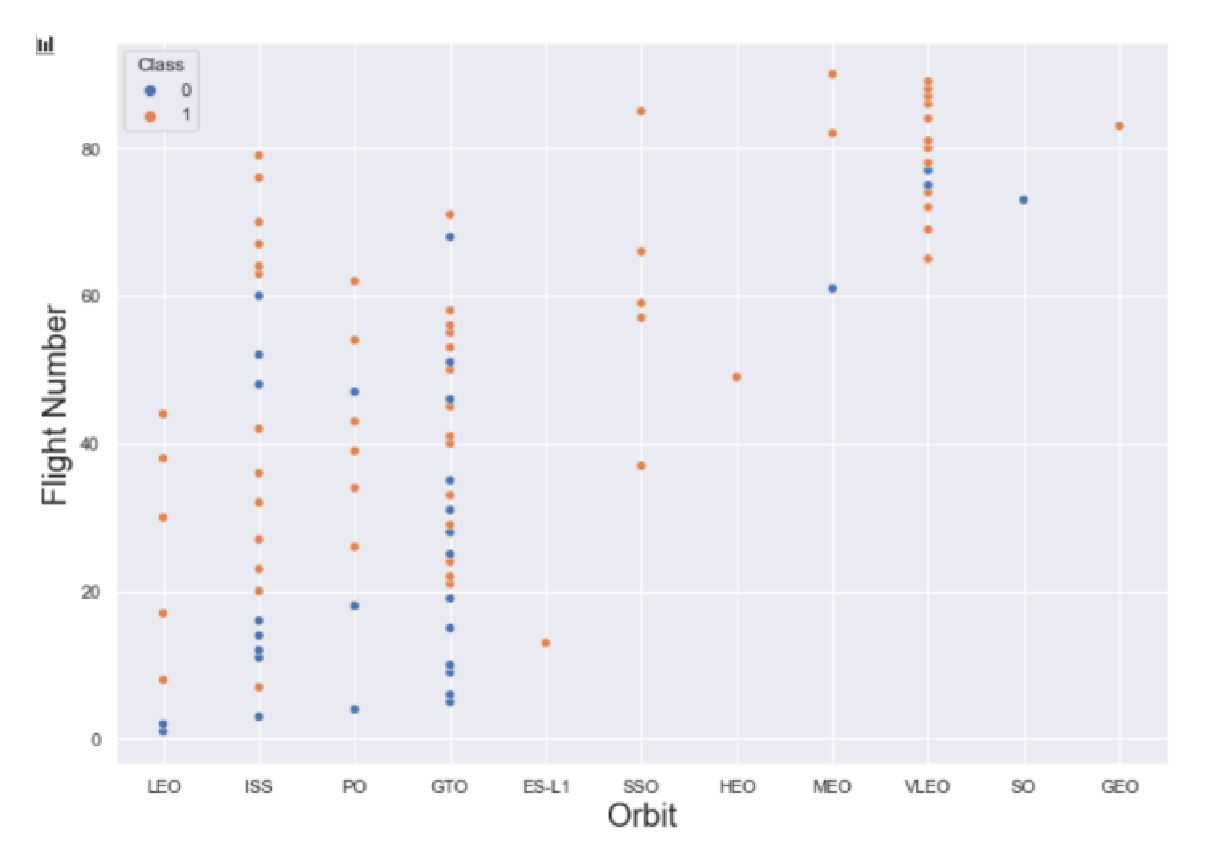
Success Rate vs. Orbit Type

- Orbit GEO,HEO,SSO,ES-L1 has the best Success Rate



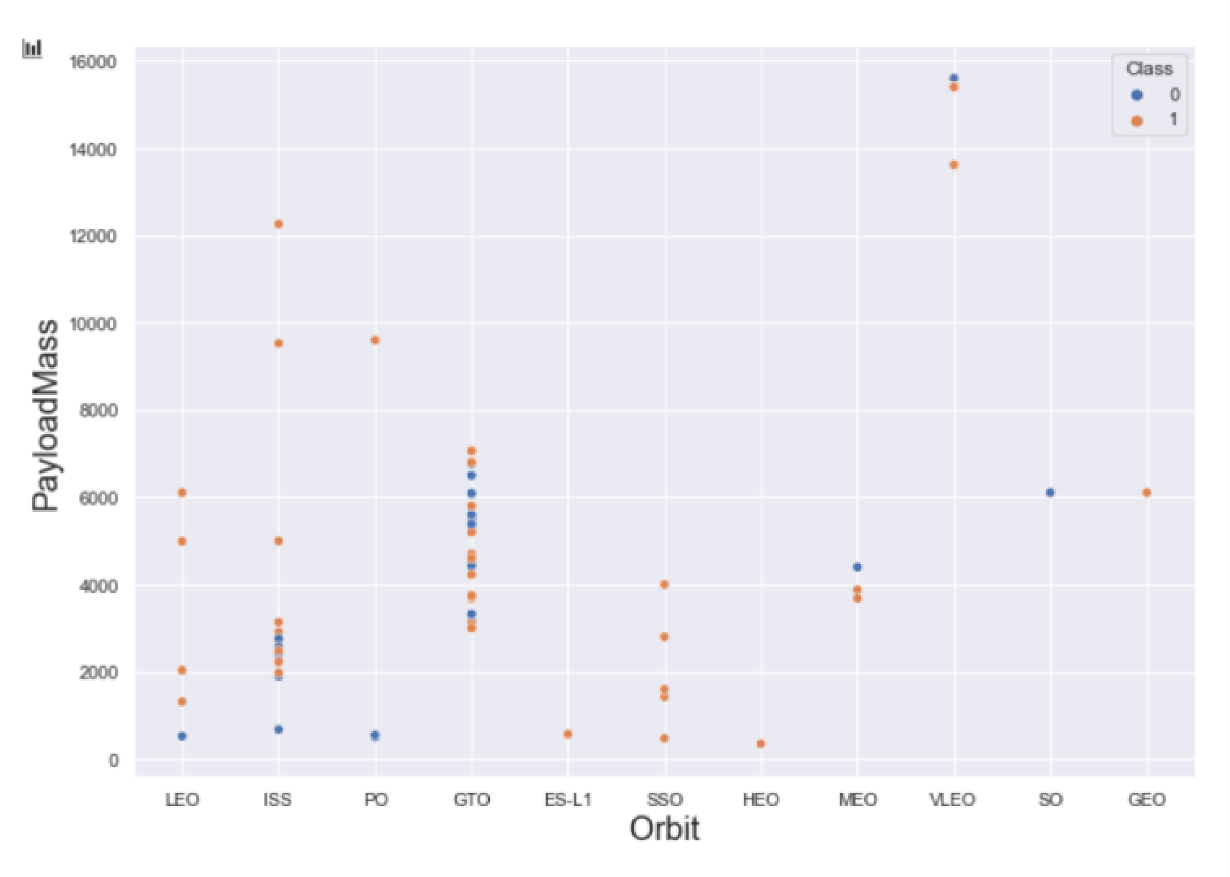
Flight Number vs. Orbit Type

- No clear correlation between flight number success and orbit type.



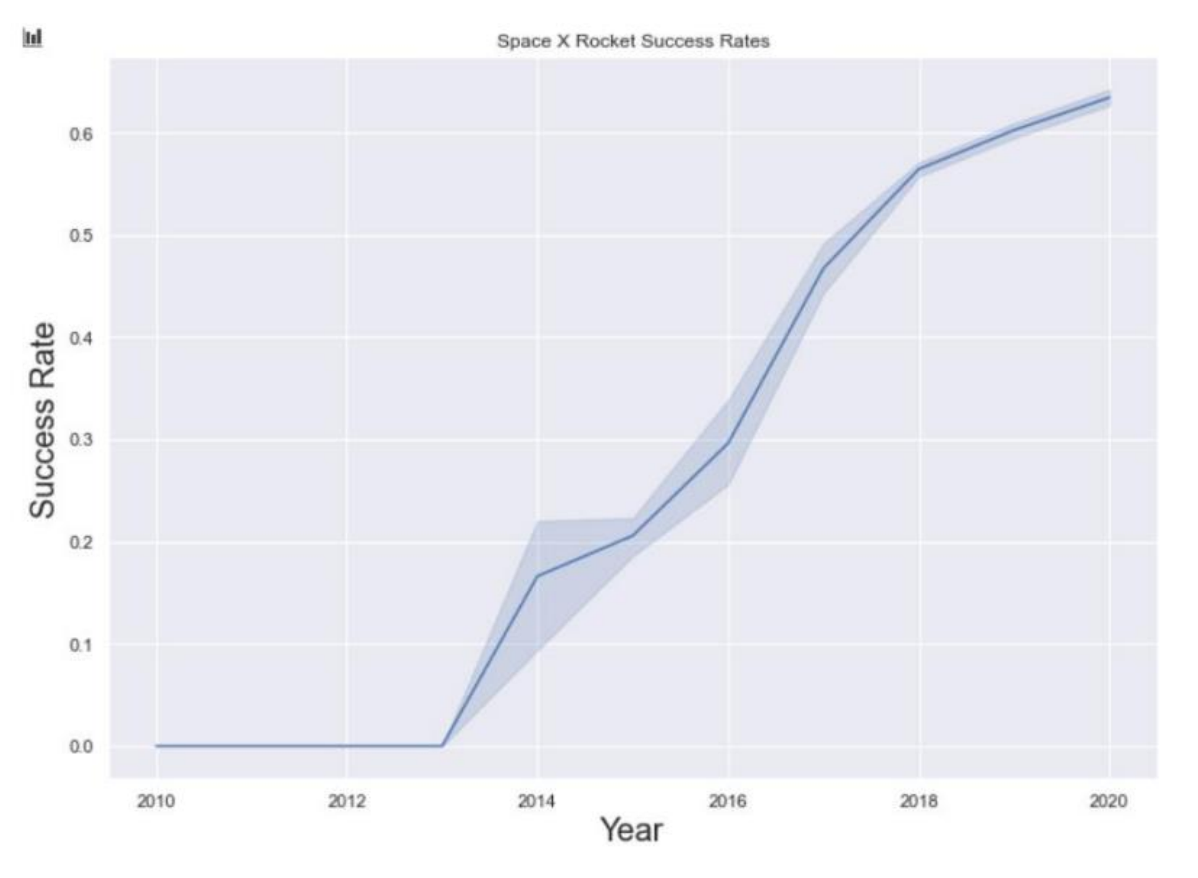
Payload vs. Orbit Type

- Some orbits experience no failure, such as SSO. GTO orbit seems to have high failure rate and not that much payload variation.



Launch Success Yearly Trend

- Success rate steadily increasing over the years.



Section 2.1

EDA with SQL

All Launch Site Names

- SELECT DISTINCT Launch_Site FROM SPACEXTABLE;
- Selects unique names from column launchsite from database

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

Launch Site Names Begin with 'CCA'

- SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
- Selects top 5 results from database, where launch site includes keyword 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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

- SELECT SUM(Payload_Mass__kg_) AS Total_Payload_Mass FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)';
- This is a sum of the column payload mass, presented as total payload mass from database, where customer is also NASA.

Total_Payload_Mass
45596

Average Payload Mass by F9 v1.1

- SELECT AVG(Payload_Mass__kg_) AS AVG_BOOSTER_F9 FROM SPACEXTABLE WHERE Booster_Version LIKE 'F9%’;
- Selects average value from column payload mass, presented as AVG_BOOSTER_F9 from database where booster version also contains key word F9

AVG_BOOSTER_F9
6138.287128712871

First Successful Ground Landing Date

- select min(Date) AS MINSUCDATE FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)’;
- Selects the minimum date, presented as MINSUCDATE from database, where Landing Outcome is equal to success.

MINSUCDATE
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

SELECT Booster_Version

FROM SPACEXTABLE

WHERE Landing_Outcome = 'Success (drone ship)'

AND Payload_Mass__kg_ > 4000

AND Payload_Mass__kg_ < 6000;

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

- Selects booster version from database where landing is a success and payload is between specified values.

Total Number of Successful and Failure Mission Outcomes

```sql
SELECT
    CASE
        WHEN Mission_Outcome LIKE 'Failure%' THEN 'Failure'
        ELSE Mission_Outcome
    END AS Outcome_Category,
    COUNT(*) AS Total_Outcomes
FROM
    SPACEXTABLE
WHERE
    Mission_Outcome IN ('Success', 'Failure') OR
    Mission_Outcome LIKE 'Failure%'
GROUP BY
    Outcome_Category;
```

Outcome_Category	Total_Outcomes
Failure	1
Success	98

- Counts all mission outcomes separating successes and failures and displays them in a table.

Boosters Carried Maximum Payload

```sql
SELECT Booster_Version
FROM SPACEXTABLE
WHERE Payload_Mass__kg_ = (
    SELECT MAX(Payload_Mass__kg_)
    FROM SPACEXTABLE
);
```

- Displays maximum payload caried by booster versions.

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

```sql
SELECT
    CASE substr(Date, 6, 2)
        WHEN '01' THEN 'January'
        WHEN '02' THEN 'February'
        WHEN '03' THEN 'March'
        WHEN '04' THEN 'April'
        WHEN '05' THEN 'May'
        WHEN '06' THEN 'June'
        WHEN '07' THEN 'July'
        WHEN '08' THEN 'August'
        WHEN '09' THEN 'September'
        WHEN '10' THEN 'October'
        WHEN '11' THEN 'November'
        WHEN '12' THEN 'December'
    END AS Month_Name,
    Landing_Outcome,
    Booster_Version,
    Launch_Site
FROM
    SPACEXTABLE
WHERE
    substr(Date, 0, 5) = '2015' AND
    Landing_Outcome LIKE 'Failure%' AND
    Landing_Outcome LIKE '%drone ship%';
```

Lists the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Month_Name	Landing_Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

```
SELECT Landing_Outcome, COUNT(*) AS Outcome_Count
FROM SPACEXTABLE
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY Outcome_Count DESC;
```

Landing_Outcome	Outcome_Count
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

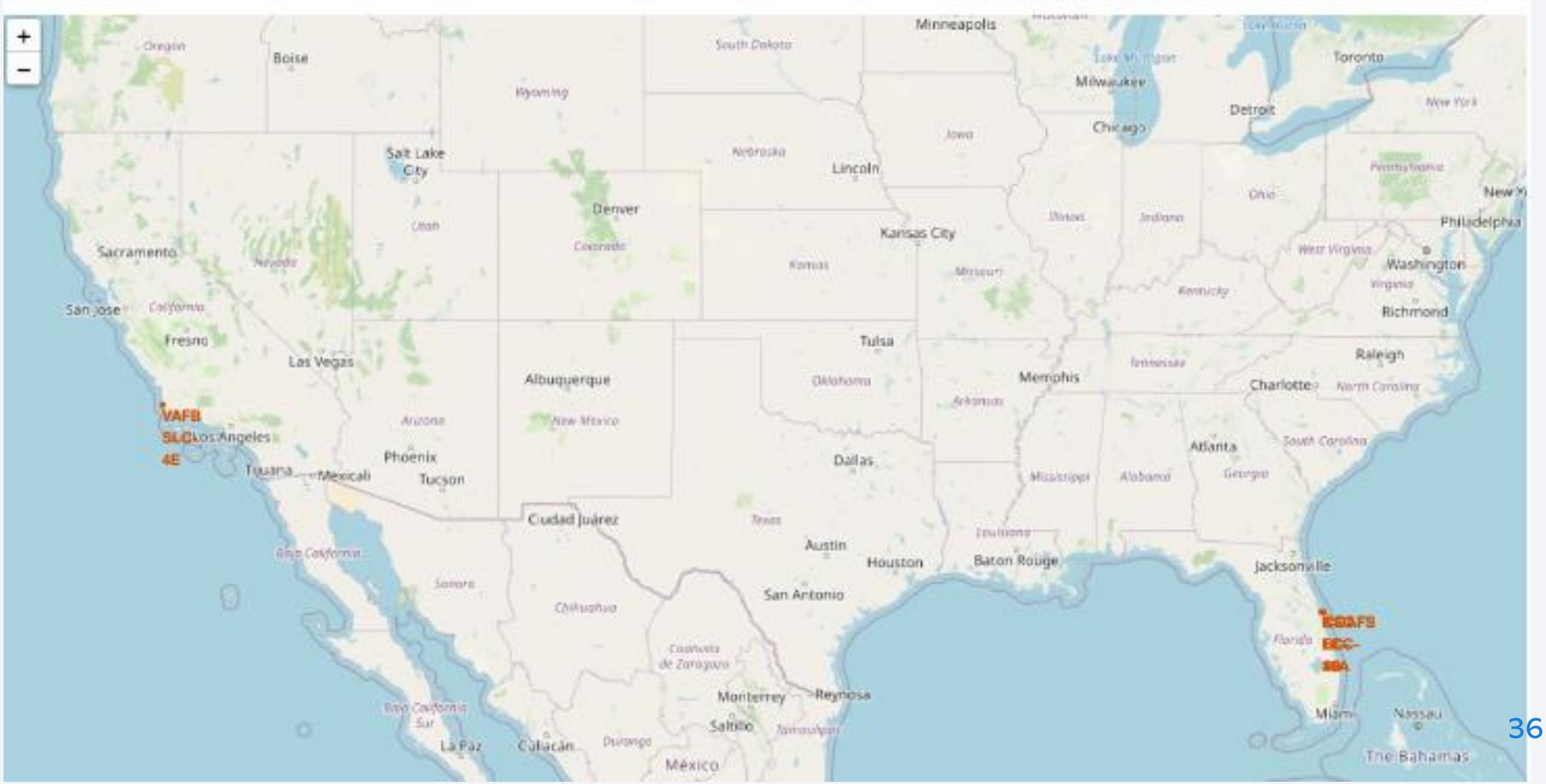
- Ranks the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

Section 3

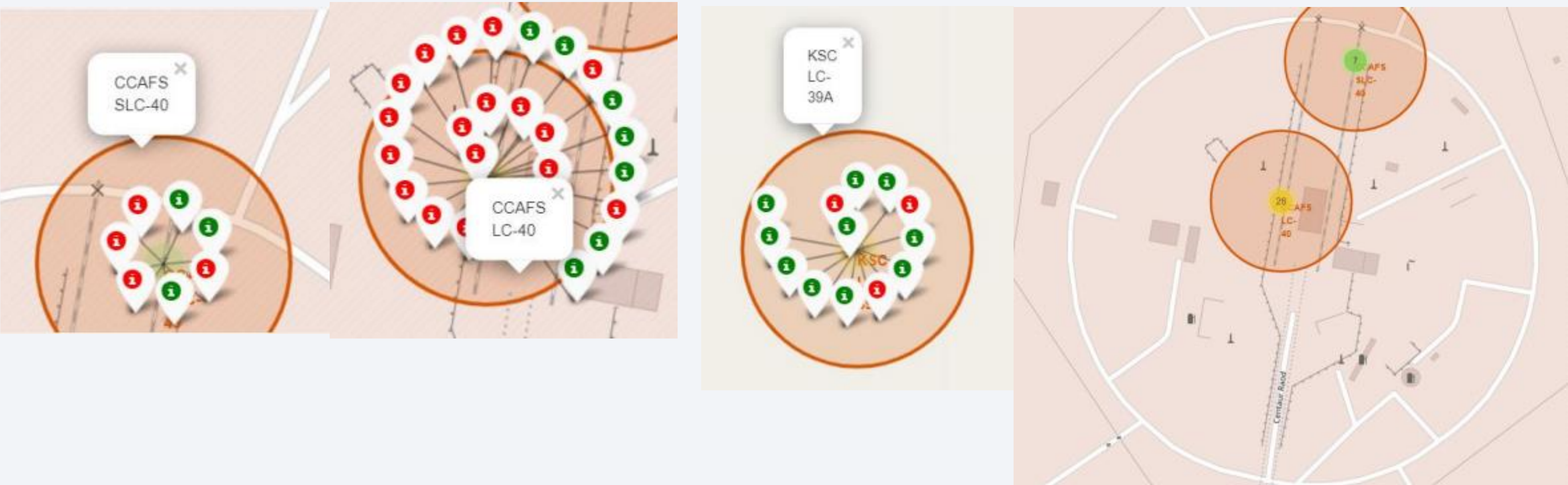
Launch Sites Proximities Analysis

Launch sites USA

We can see two launch sites on either coasts of the US



Labelled Markers



Color coded markers, featuring red for failure and green for success.

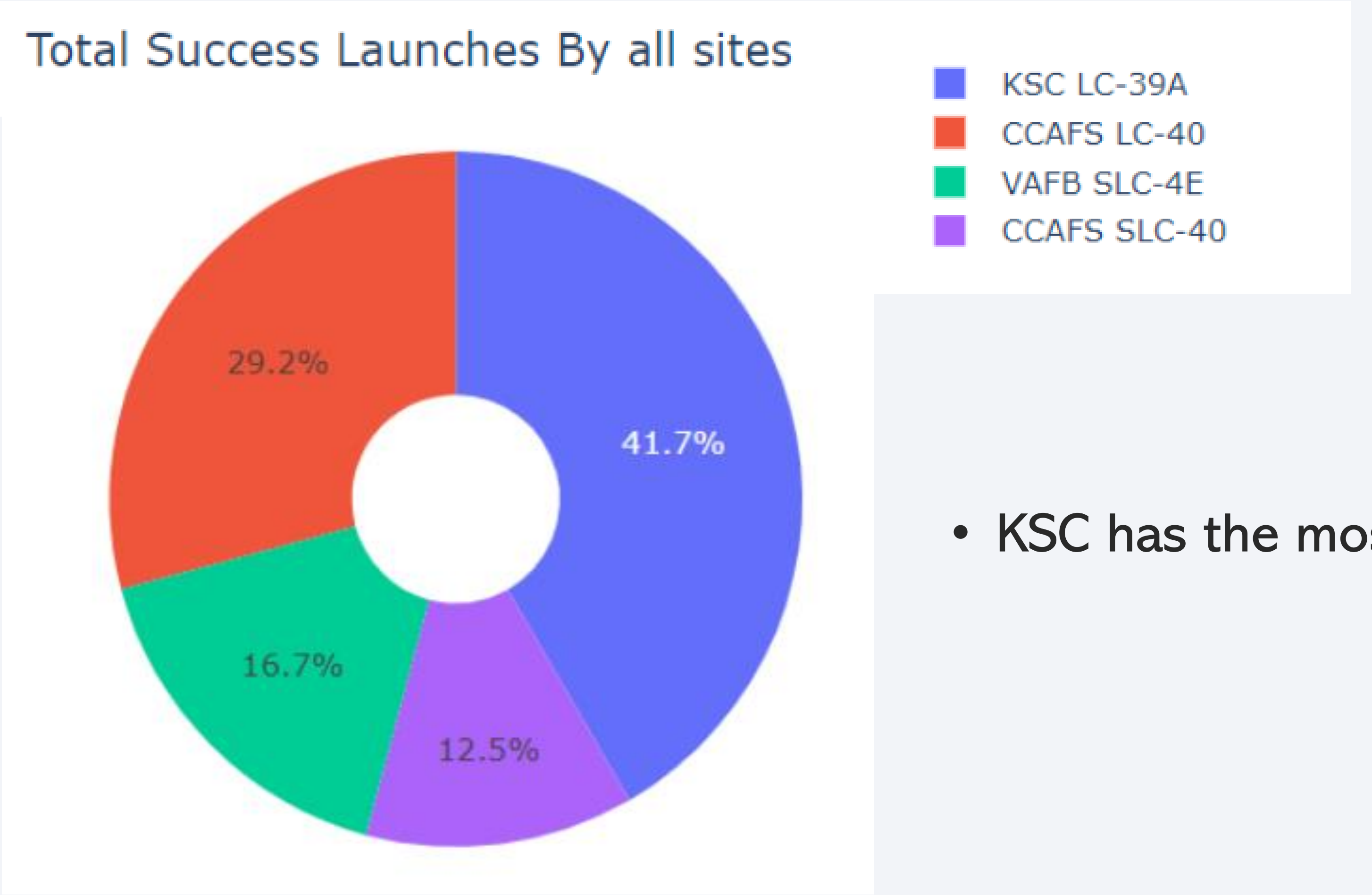
Launch site distances from landmarks

- •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 4

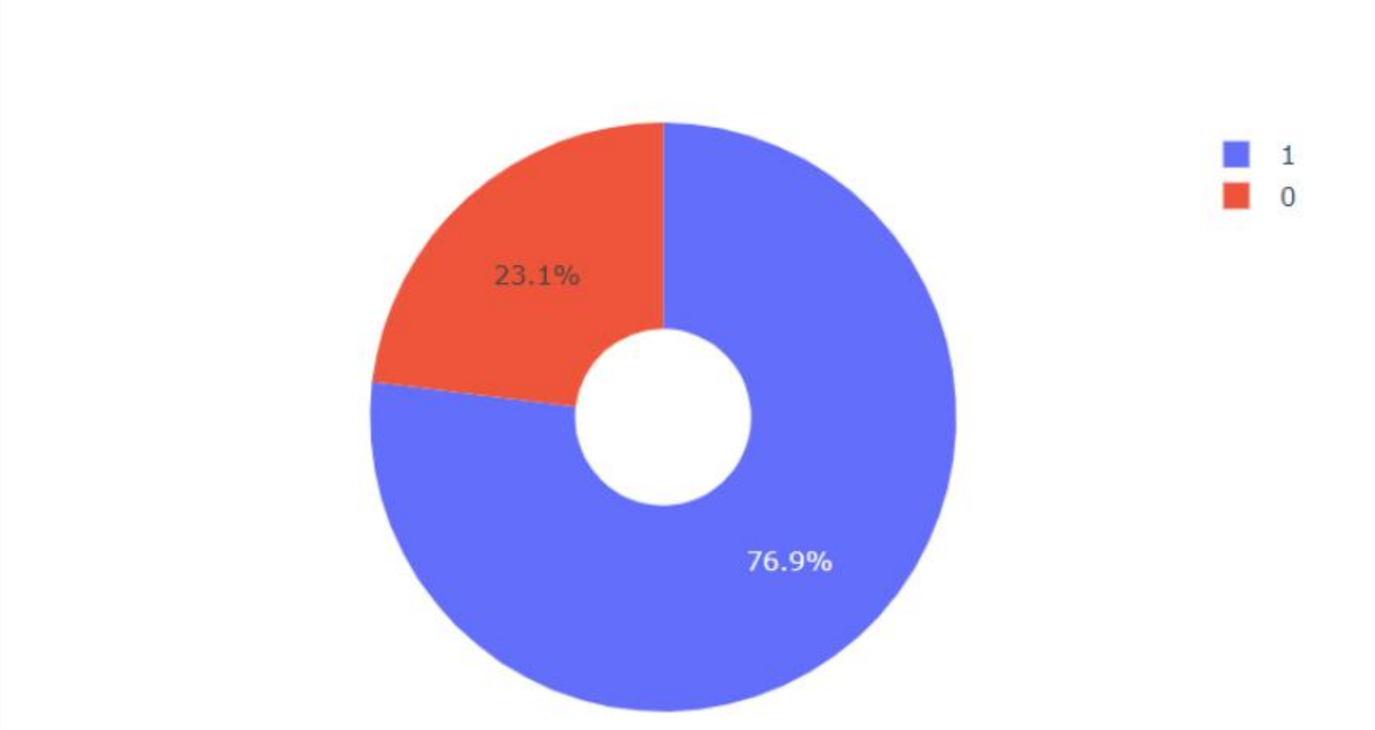
Build a Dashboard with Plotly Dash

Plotly pie chart



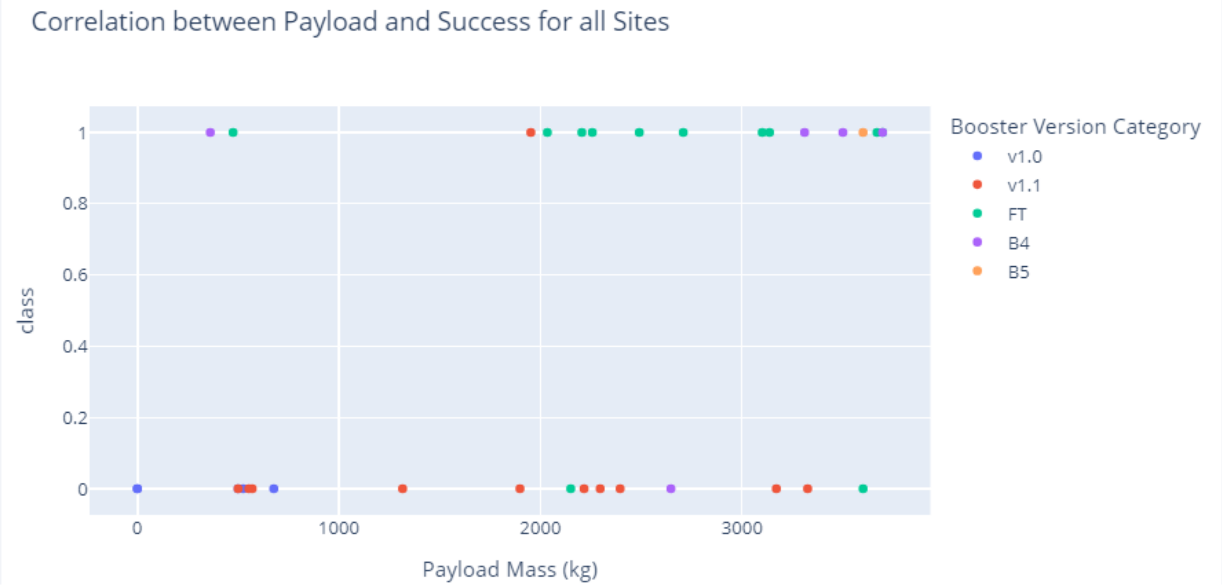
- KSC has the most successful launches

Plotly pie chart for highest success ratio

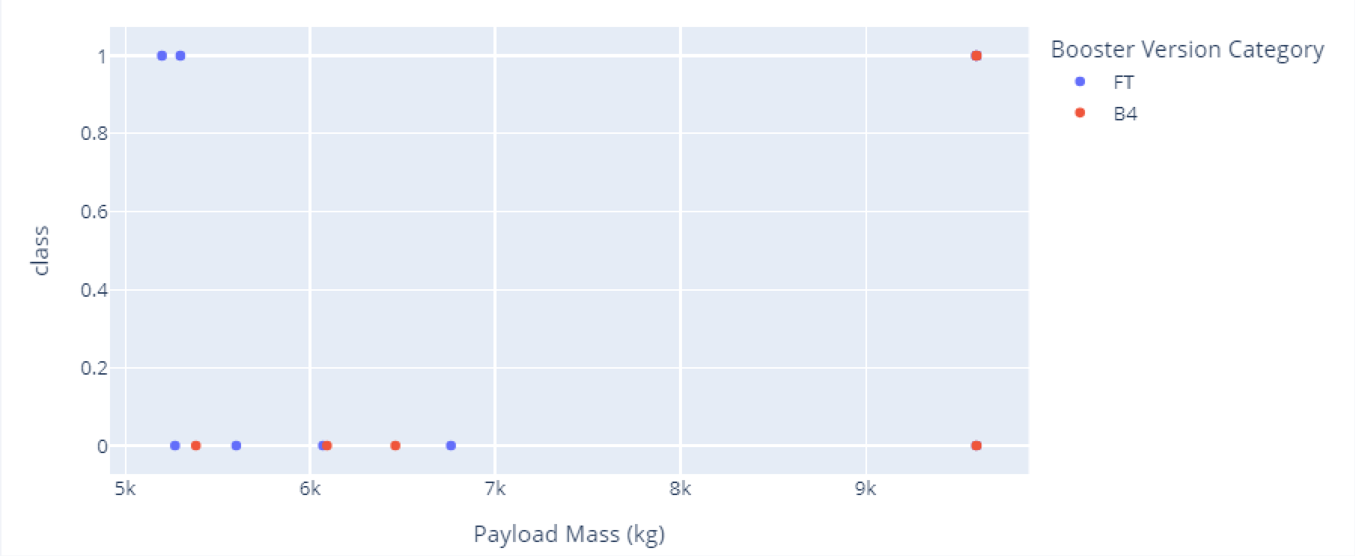


- KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slide



0-4000 range



- We can see the success rates for low weighted payloads is higher than the heavy weighted payloads and that the boosters used for heavy launches are fewer.

5000-10000 range

Section 5

Predictive Analysis (Classification)

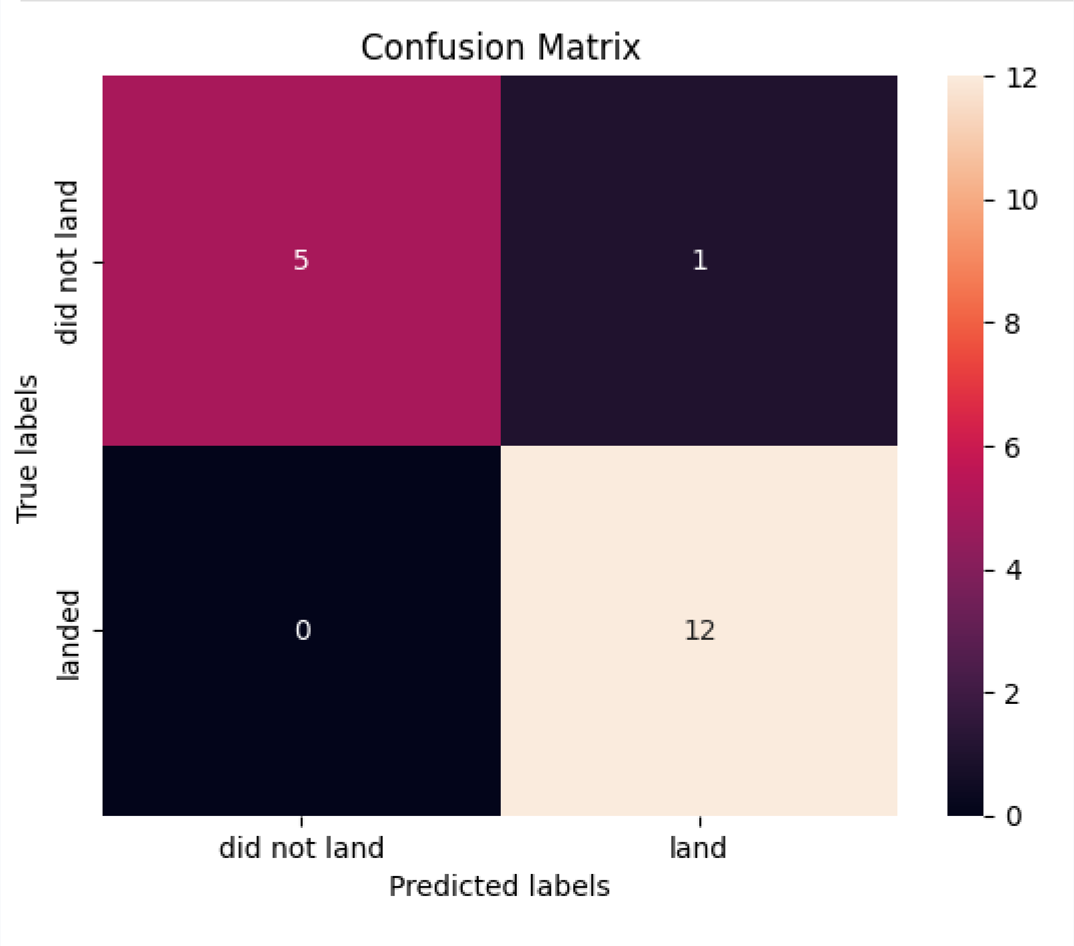
Classification Accuracy

- The tree CV has the best accuracy with a score of 0.93 and the following parameters

```
GridSearchCV(cv=10, estimator=DecisionTreeClassifier(),
             param_grid={'criterion': ['gini', 'entropy'],
                         'max_depth': [2, 4, 6, 8, 10, 12, 14, 16, 18],
                         'max_features': ['auto', 'sqrt'],
                         'min_samples_leaf': [1, 2, 4],
                         'min_samples_split': [2, 5, 10],
                         'splitter': ['best', 'random']})
```

Confusion Matrix

- This is a confusion matric for the tree decision classifier.
- It has 1 false positive



Conclusions

- The best algorithm for our dataset was Tree CV.
- Low weighted rockets may have better success rate during launches.
- Successful launches have been steadily increasing over the years

Thank you!