IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Kiarash Astanboos
3/21/2025

Outline

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

Executive Summary

- Summary of methodologies
  - Module 1
    - Data Collection through API
    - Data Collection with Web Scaping
    - Data Wrangling
  - Module 2
    - EDA with SQL
    - EDA with Visualization
  - Module 3
    - Interactive Visual analytics with Folium
    - Interactive Visual analytics with Plotly Dash
  - Module 4
    - Machine Learning Predection
- Summary of all results
  - EDA results
  - Interactive analytics results
  - Machine Learning reults

Introduction

Space Exploration Technologies Corp. commonly referred to as **SpaceX**, is an American space technology company headquartered at the Starbase development site near Brownsville, Texas. Since its founding in 2002, the company has made numerous advancements in rocket propulsion, reusable launch vehicles, human spaceflight and satellite constellation technology. 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.

Problems:

- Identifying the key factors that impact the success of a landing.
- Analyzing the relationship between different variables and their influence on landing outcomes.
- Determining the optimal conditions that maximize the probability of a successful landing.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from SpaceX API and web scraping from a Wikipedia page.
- Perform data wrangling
  - New features like class has been createtd. Also perfomed OHE for 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 collection is the process of gathering and measuring information from various sources to analyze and make data-driven decisions. It is a crucial step in data science, machine learning, and research, as it ensures that relevant, accurate, and sufficient data is available for analysis. We have used REST API and web scraping to collect the datasets.

- With REST API, we send a request to an API endpoint, which returns data in JSON format. Using pandas, we convert this JSON response into a structured DataFrame. The data is then cleaned and preprocessed for further analysis.

- For web scraping, we send an HTTP request to a webpage and extract its HTML content. Using BeautifulSoup, we parse the HTML and locate relevant data, such as tables. The extracted information is then structured into a pandas DataFrame for analysis.

Data Collection – SpaceX API

- Notebook on Github : Link

Get request for rocket launch data with request.get()

Use pd.json_normalize() to convert the json response to a dataframe

Clean the dataframe

Data Collection - Scraping

- Notebook on Github : Link

Get request for rocket launch data with request.get() to Falcon9 Launch Wiki page

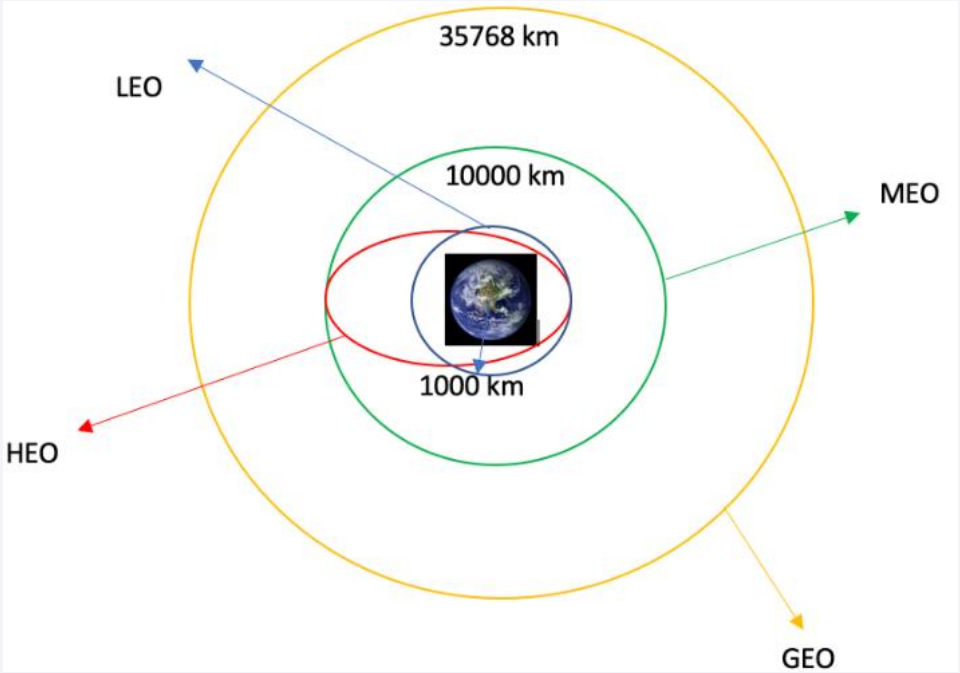
Create Beautiful object from the HTML response

Extract all column/variable names from the HTML table header

Create a data frame by parsing the launch HTML tables

Data Wrangling

- Notebook on Github : Link



Identify and calculate the percentage of missing values in each attribute and which columns are numerical and categorical

Calculate the number of launches on each site

Calculate the number and occurrence of each orbit

Calculate the number and occurence of mission outcome of the orbits

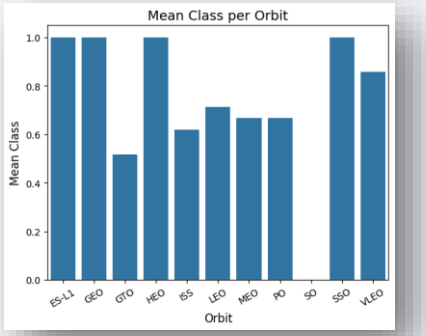
Create a landing outcome label from Outcome column

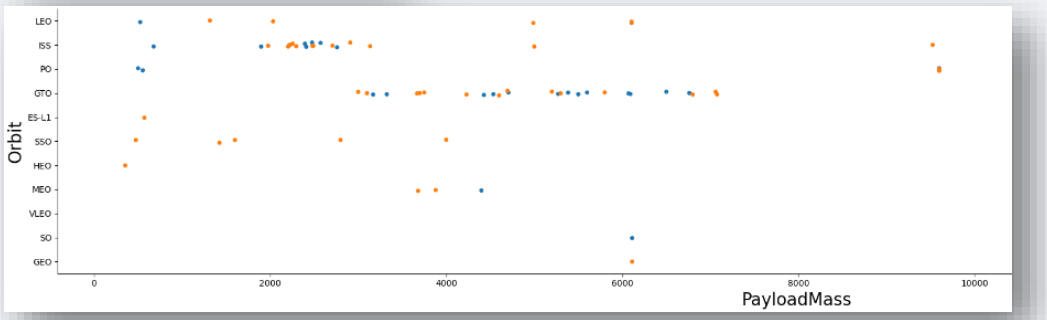
EDA with Data Visualization

We have used scatter plot and bar plot at the start to find relationship between attributes. Then we used feature engineering to choose the most important features.

Features that have been plotted:

- Flight Number and Launch site
- Payload and Launch Site
- Success Rate of each Orbit Type
- Flight Number and Orbit Type
- Payload and Flight Number





Notebook on Github : Link

EDA with SQL

The SQL queries that we performed to get better understanding of the dataset:

- Displaying the names of the launch sites.
- Displaying 5 records where launch sites begin with the string ‘CCA’.
- Displaying the total payload mass carried by booster launched by NASA (CRS).
- Displaying the average payload mass carried by booster version F9 v1.1.
- Listing the date when the first successful landing outcome in ground pad was achieved.
- Listing the names of the boosters which have success in drone ship 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 failed landing_outcomes in drone ship, their booster versions, and launch sites
- names for in year 2015.
- Rank the count of landing outcomes or success between the date 2010-06-04 and
- 2017-03-20, in descending order.

Notebook on Github : Link

Build an Interactive Map with Folium

To create an interactive map for visualizing launch data, we plotted each launch site using its latitude and longitude coordinates and marked them with labeled circle markers.

Next, we categorized the launch outcomes—success and failure—as classes 1 and 0, respectively, using green and red markers within a MarkerCluster().

Additionally, we applied the Haversine formula to calculate the distances from launch sites to key landmarks, helping answer questions such as:

- How close are the launch sites to railways, highways, and coastlines?
- How close are the launch sites to nearby cities?

Notebook on Github : Link

Build a Dashboard with Plotly Dash

- We built an interactive dashboard with Plotly dash.
- We plotted pie charts showing the total launches by each site and scatter graph showing the relationship with Outcome and Payload Mass (Kg) for the different booster version.
- The user can choose the site via dropdown and use slider to select payload range

Notebook on Github : Link

•Define a parameter grid for hyperparameter tuning.
•Use **GridSearch** and **cross-validation** to train and optimize the model.

Predictive Analysis (Classification)

Preproces Data:
- Create a column for the class
- Normalize Data
- Split data to training and test

Building models:
- Create model
- Define a parameter grid for hyperparameter tuning.
- Use GridSearch and cross-validation to train and optimize the model.

Evaluate and find the best model:
- Check accuracy
- Plot the confusion matrix
- Select the model with the highest accuracy

Notebook on Github : Link

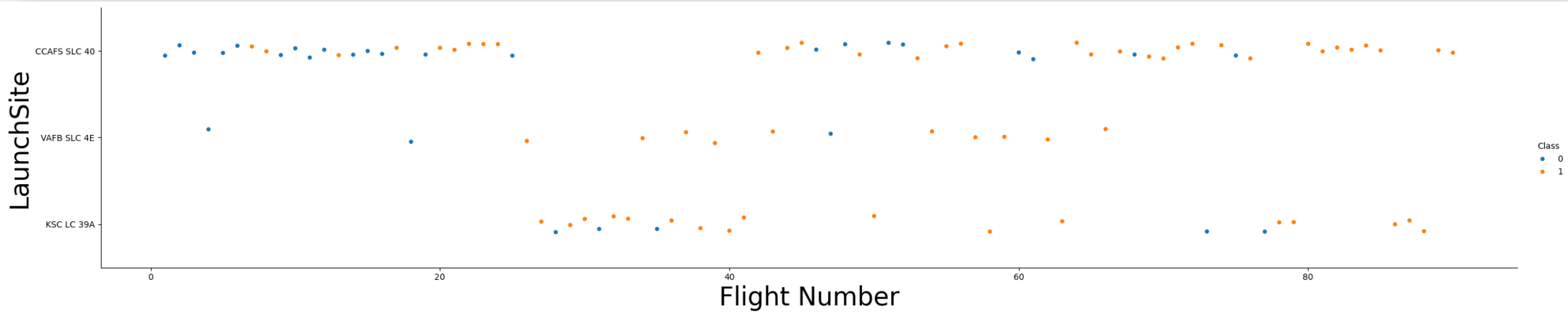
Results

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

Section 2

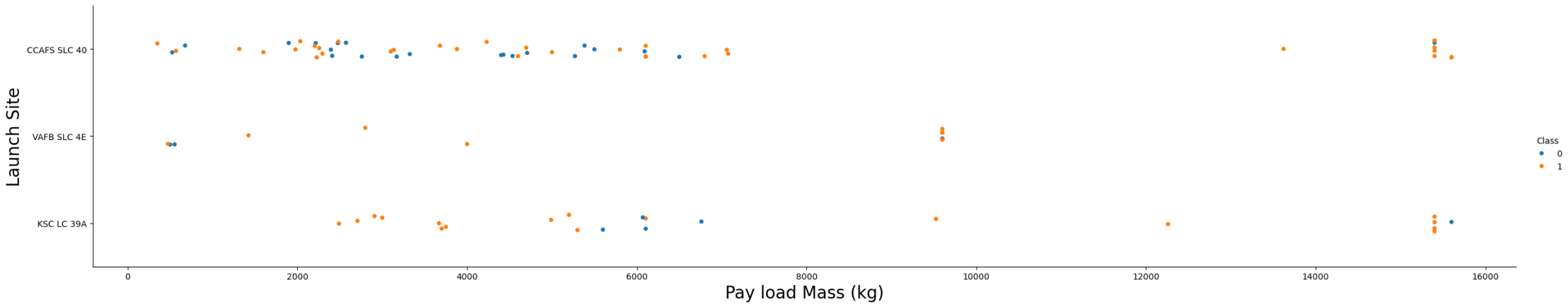
Insights drawn from EDA

Flight Number vs. Launch Site



This scatter plot shows that the success is more frequent in higher flight numbers.
Also the CCAFS site has the most flights.

Payload vs. Launch Site

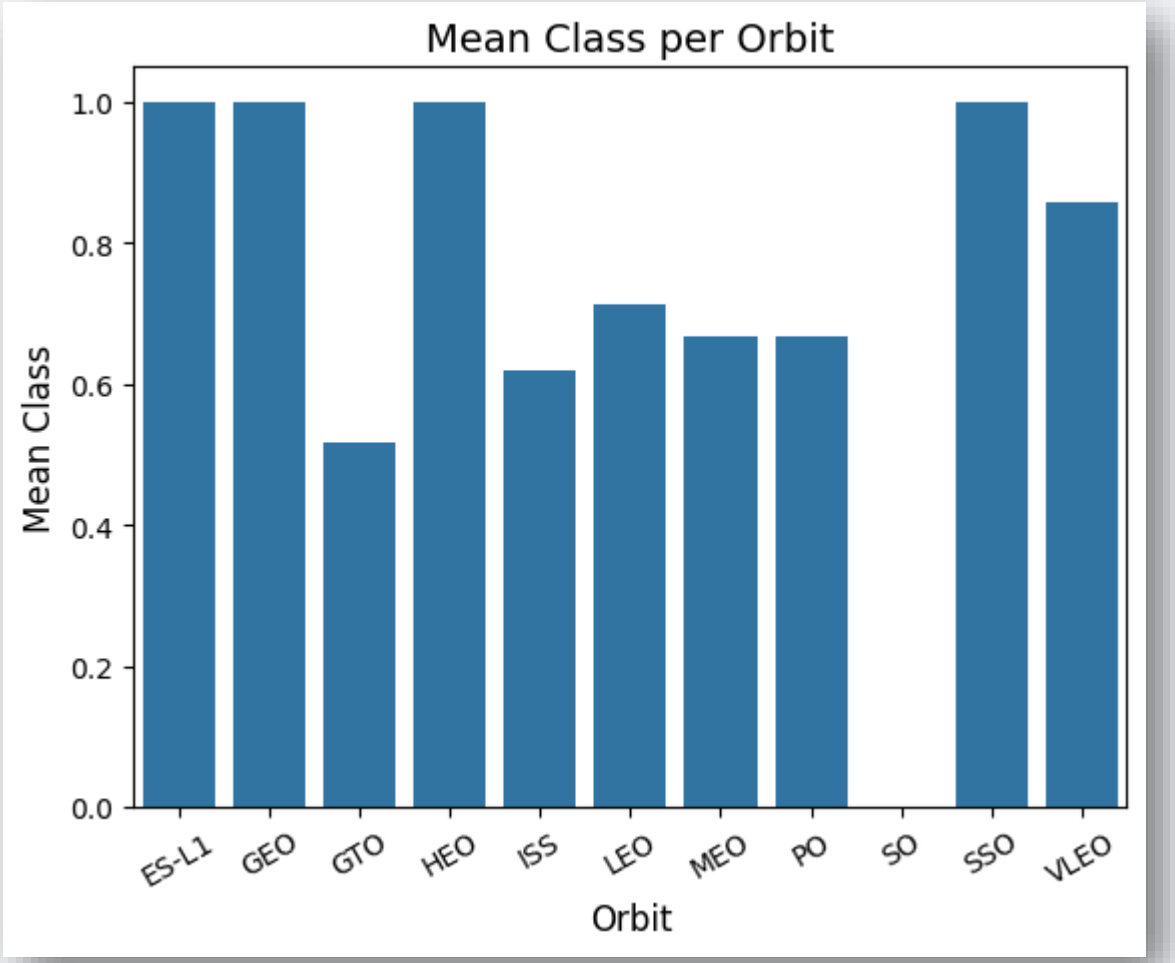


for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

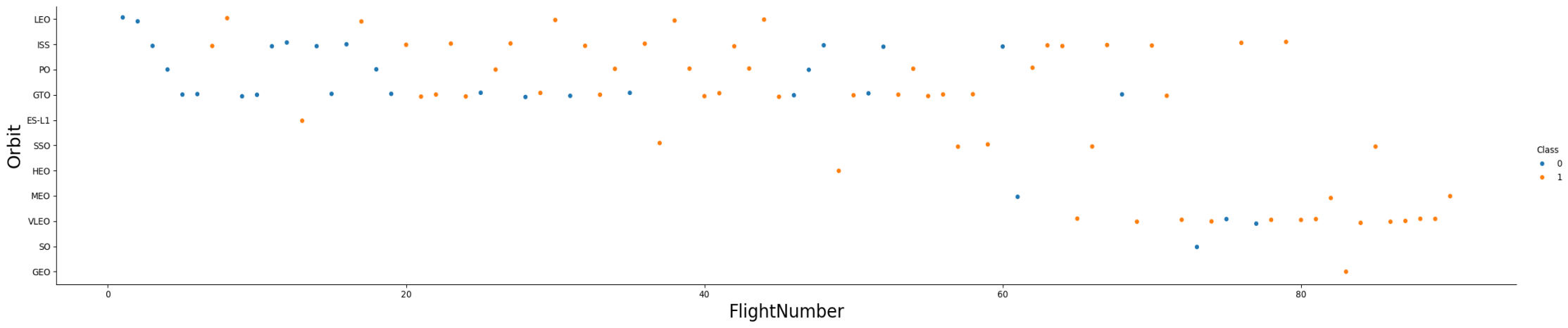
We can also see that a higher payload mass increases thabe chance of a successful landing.

Success Rate vs. Orbit Type

- some orbits has 100% success rate such as SSO, HEO, GEO AND ES-L1 while SO orbit produced 0% rate of success
  some of this orbits has only 1 occurrence such as GEO, SO, HEO and ES-L1 which mean we need more data to see pattern or trend before we draw any conclusion.

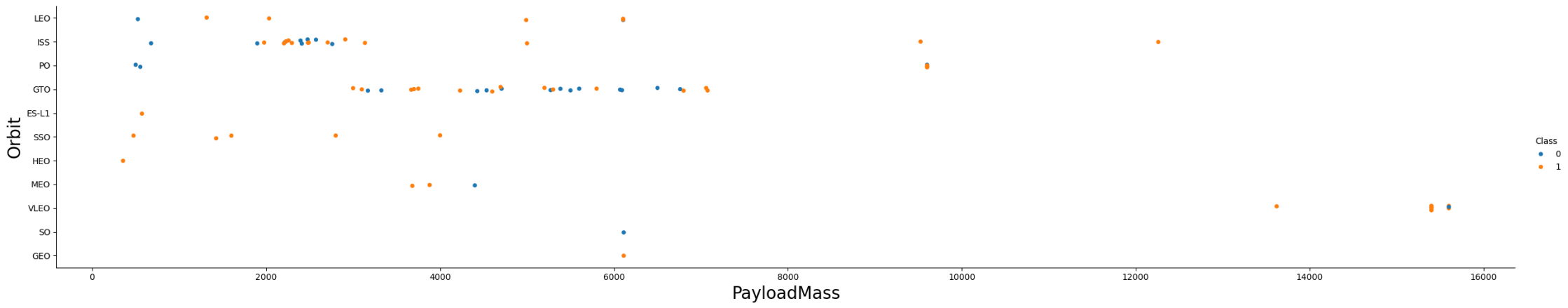


Flight Number vs. Orbit Type



- in the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.

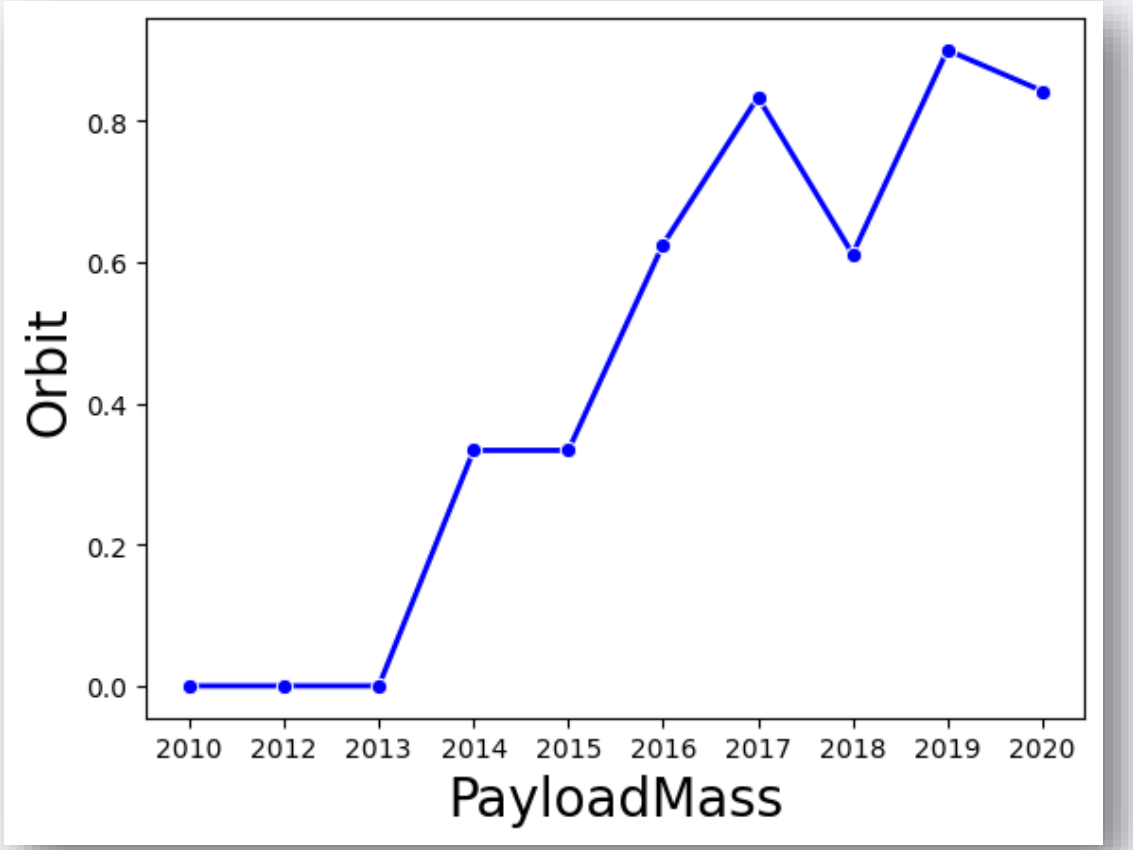
Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend

- the sucess rate since 2013 kept increasing till 2020



All Launch Site Names

```
%sql SELECT DISTINCT LAUNCH_SITE as "Launch_Sites" FROM SPACEXTBL;
```

✓ 0.0s

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

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

Launch Site Names Begin with 'CCA'

```
%sql SELECT LAUNCH_SITE FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

✓ 0.0s

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

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer LIKE 'NASA (CRS)%';
```

✓ 0.0s

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

SUM(PAYLOAD_MASS__KG_)
48213

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE BOOSTER_VERSION LIKE 'F9 v1.1%';
```

✓ 0.0s

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

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

```
%sql SELECT MIN(DATE) AS 'First Successful Landing' FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (ground pad)';
```

✓ 0.0s

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

First Successful Landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000
```

✓ 0.0s

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

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

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT COUNT(MISSION_OUTCOME) AS 'successful mission' FROM SPACEXTBL WHERE MISSION_OUTCOME LIKE 'Success%';
```

✓ 0.0s

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

successful mission
100

```
%sql SELECT COUNT(MISSION_OUTCOME) AS 'failure mission' FROM SPACEXTBL WHERE MISSION_OUTCOME LIKE 'Fail%'
```

✓ 0.0s

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

failure mission
1

Boosters Carried Maximum Payload

```
%sql SELECT DISTINCT BOOSTER_VERSION AS 'Booster Versions which carried the Maximum Payload Mass' FROM SPACEXTBL \
WHERE PAYLOAD_MASS__KG_ =(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

✓ 0.0s

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

Booster Versions which carried the Maximum Payload Mass
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 BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE DATE LIKE '2015-%' AND \
LANDING_OUTCOME = 'Failure (drone ship)';
```

✓ 0.0s

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

Booster_Version	Launch_Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

```
%sql SELECT LANDING_OUTCOME as 'Landing Outcome', COUNT(LANDING_OUTCOME) AS 'Total Count' FROM SPACEXTBL \
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' \
GROUP BY  LANDING_OUTCOME \
ORDER BY COUNT(LANDING_OUTCOME) DESC ;
```

✓ 0.0s

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

Landing Outcome	Total 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

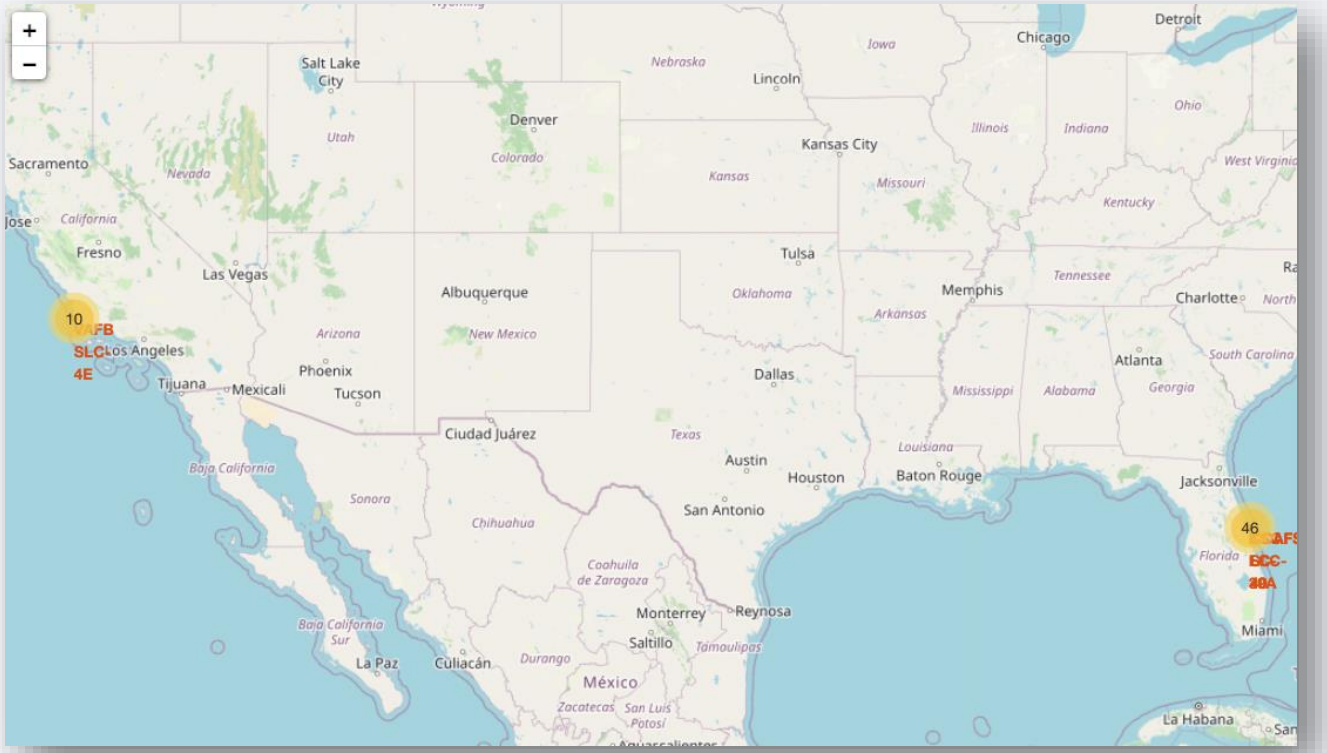
Section 3

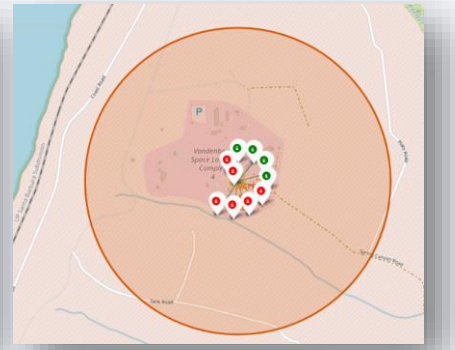
Launch Sites Proximities Analysis

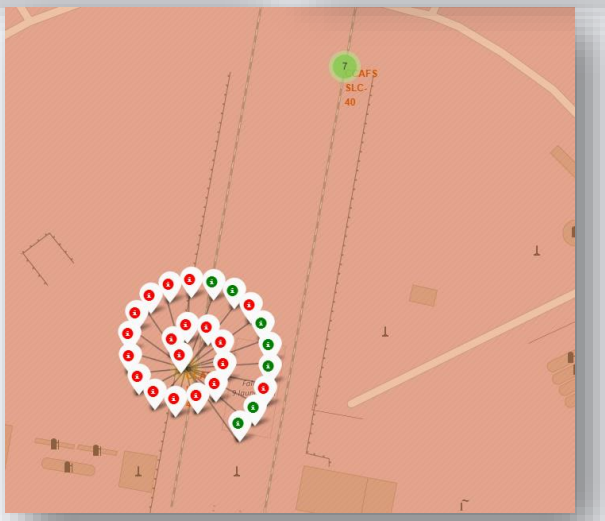
Location of launch sites

all the SpaceX launch sites are located inside the United States

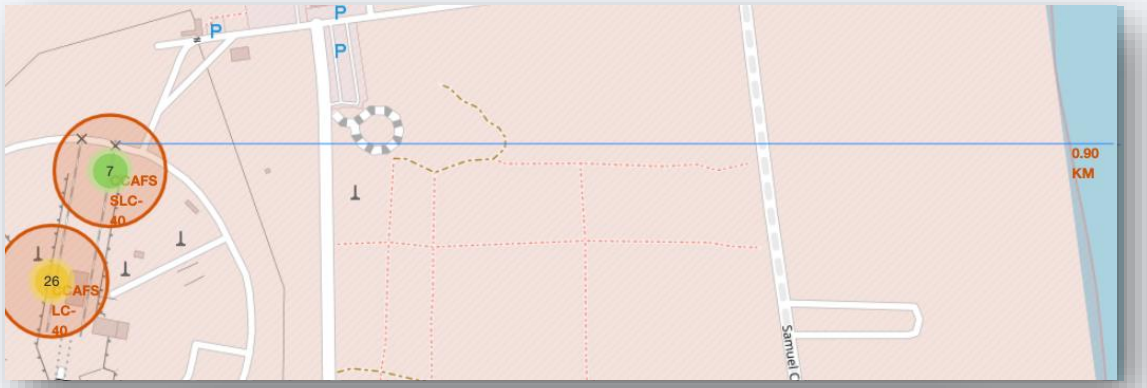
Color labeled launch outcomes

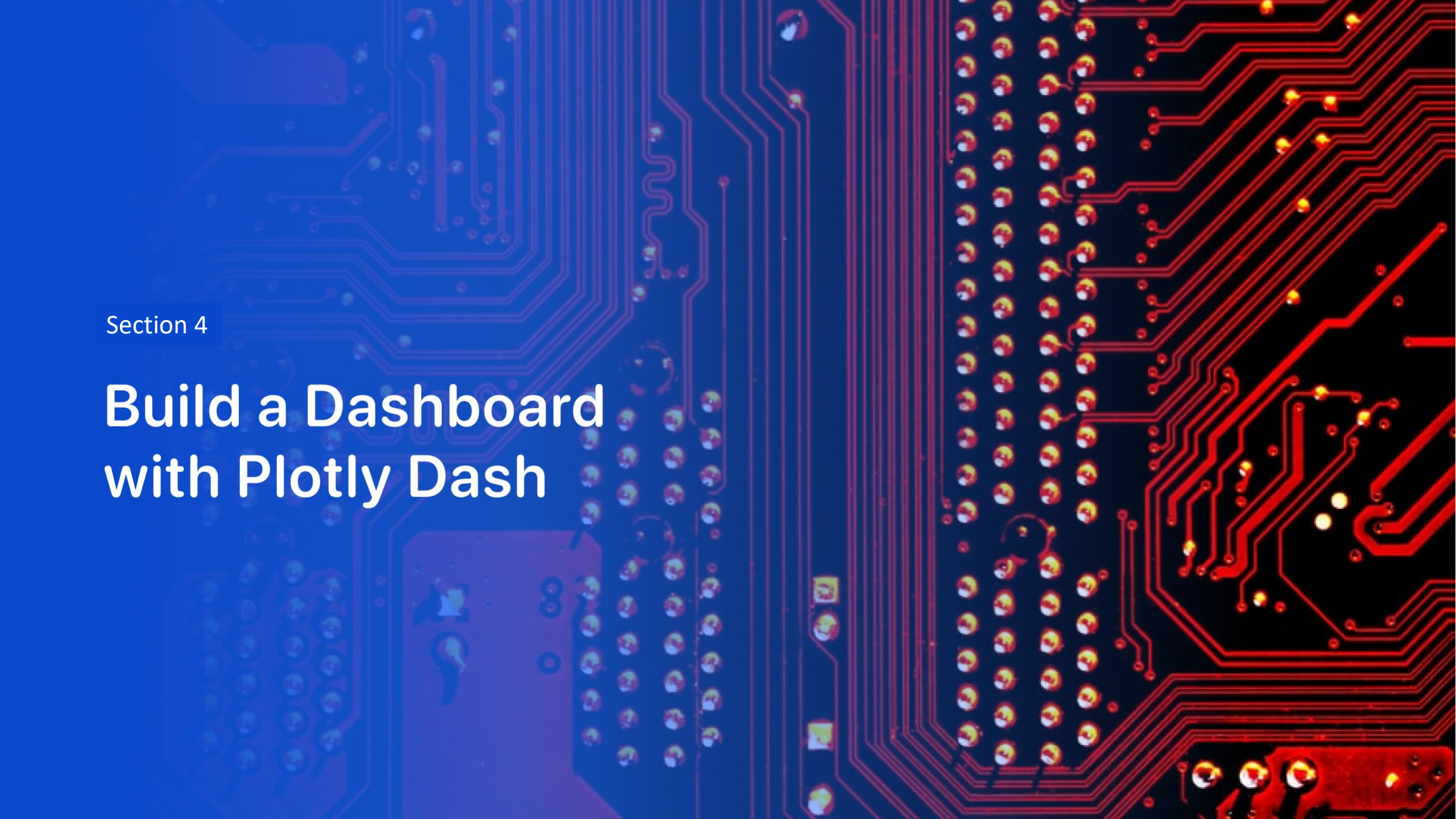






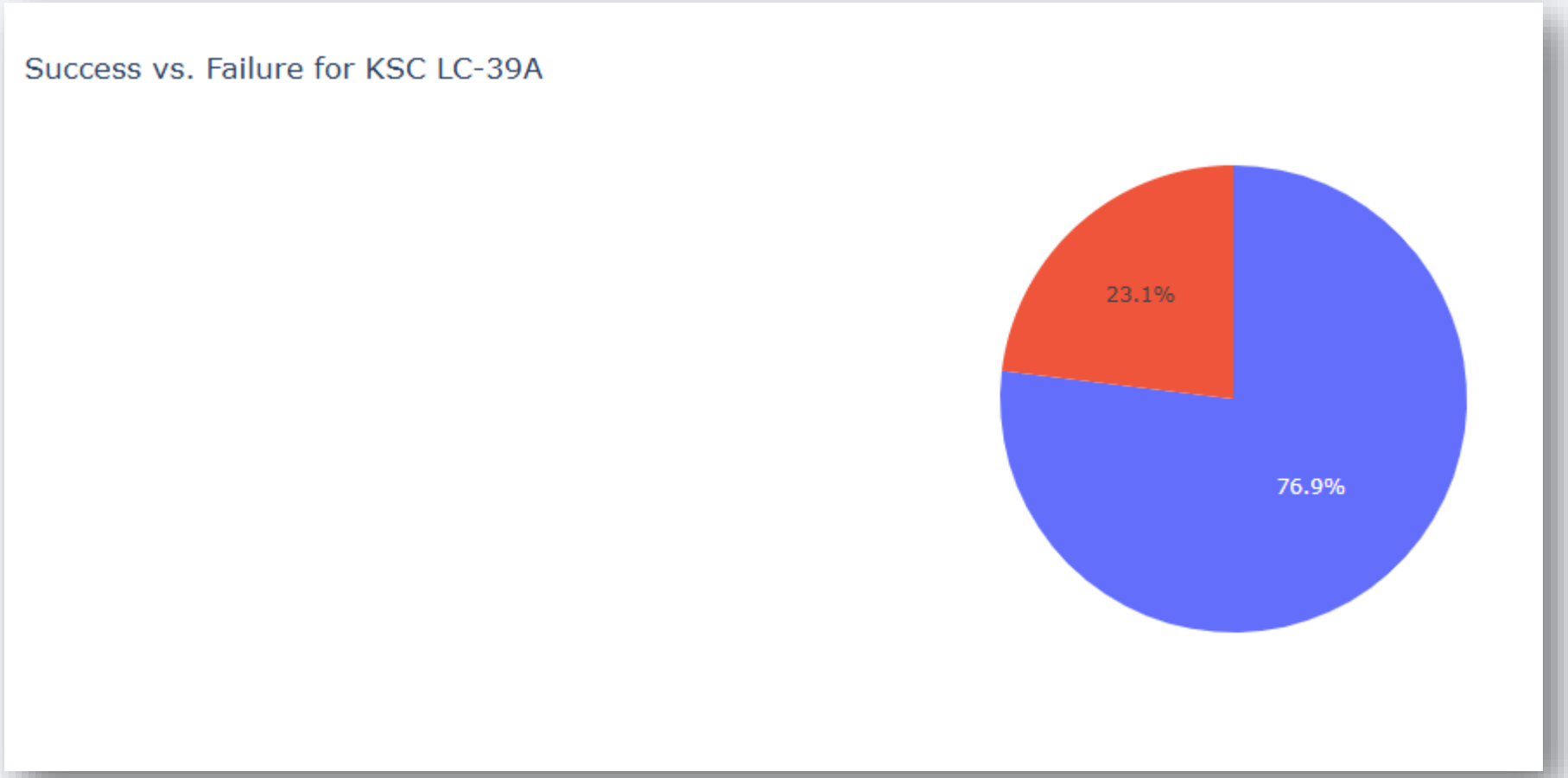
Launch Sites Distance to Landmarks




Section 4
Build a Dashboard with Plotly Dash

Total successful lanches for all sites

highest launch success



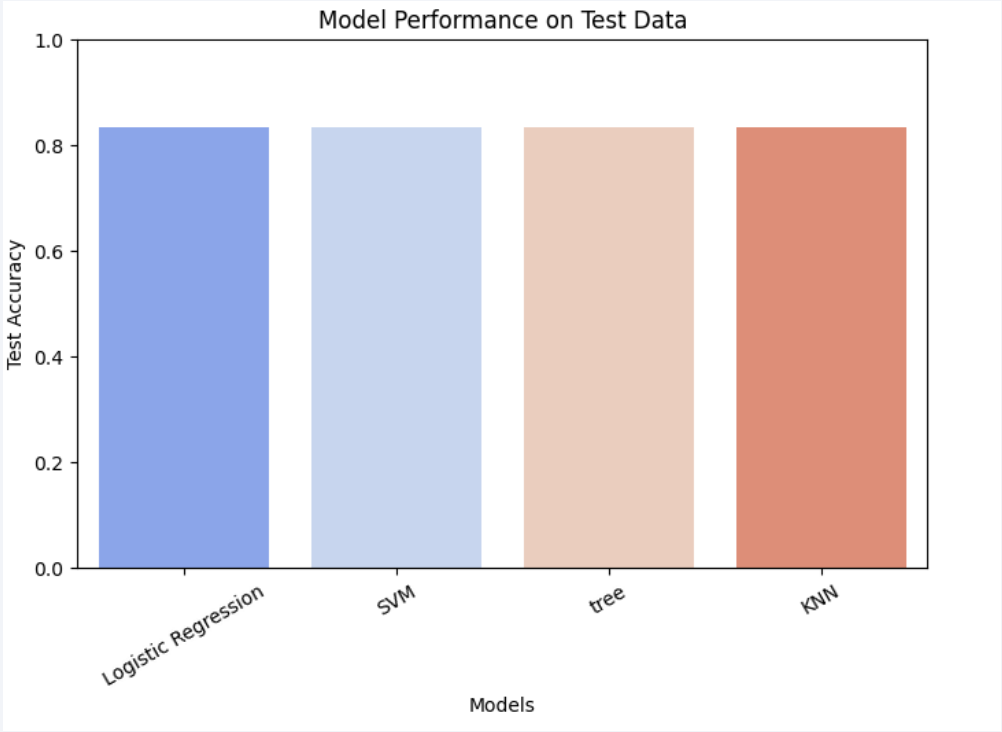
Payload vs Launch Outcome

Section 5

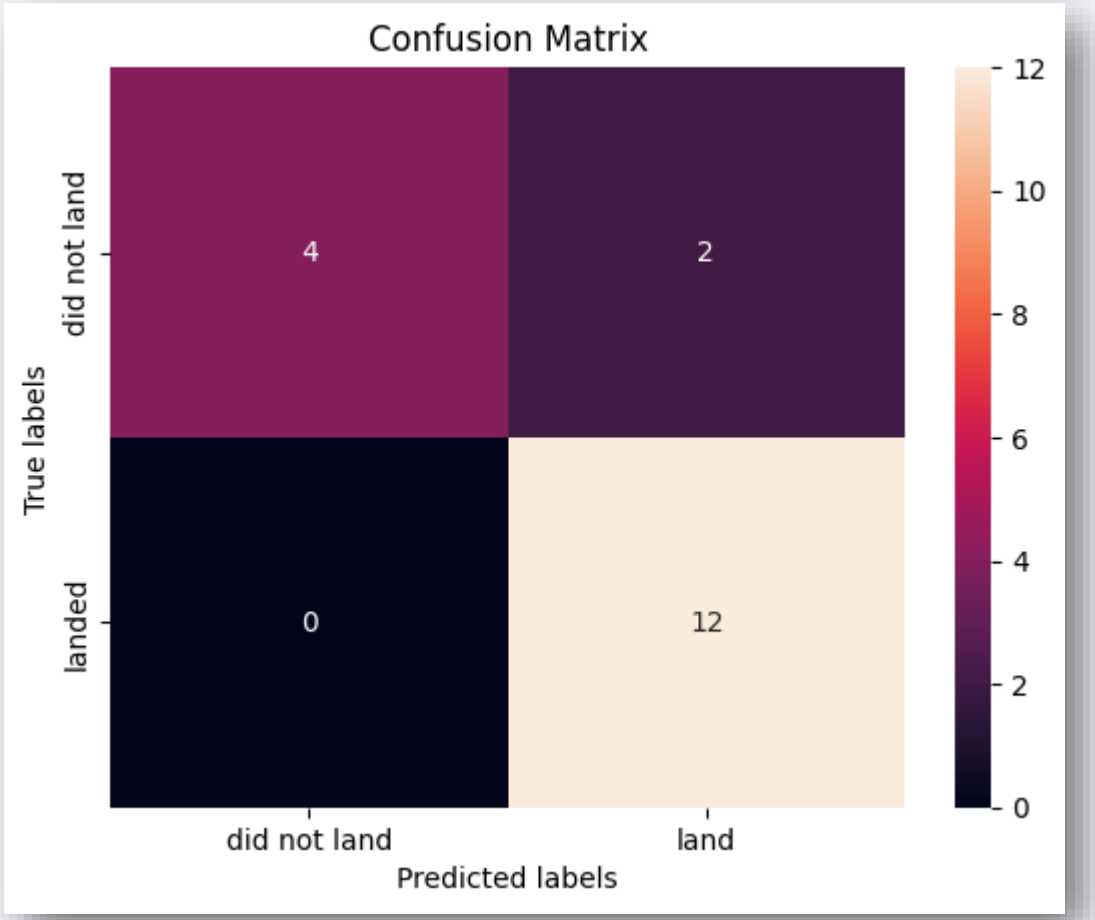
Predictive Analysis (Classification)

Classification Accuracy

- All models had same accuracy on test data but Decision Tree performed better on training data.



Confusion Matrix



The confusion matrix for the decision tree classifier shows that the classifier can distinguish between the different classes.

Conclusions

- The Tree Classifier Algorithm is the most effective machine learning approach for this dataset.
- Payloads weighing 4000 kg or less performed better than heavier payloads.
- Since 2013, the success rate of SpaceX launches has increased consistently, showing a direct correlation with time up to 2020, indicating continued improvements in future launches.
- KSC LC-39A has the highest success rate among all launch sites, with 76.9% successful launches.
- The SSO orbit has the highest success rate at 100%, with multiple successful launches.

Thank you!