# Winning Space Race with Data Science

Arshman Dixit
06-01-2024

# Outline

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

# Executive Summary

- Summary of methodologies
  - Data Collection using API
  - Data Collection using Web Scrapping
  - 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

  Space X 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 Space X can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. The goal of the project is to analyze and predict whether SpaceY can land its first stage of rocket to determine the cost of the launch using the Space X previous lauches.

- Problems you want to find answers
  - What factors effect the launch/landing of the rocket?
  - Relation of factors that are effecting the landing the most and how? (This will help in predicting the factors to take in account during operations of our rockets.)

Section 1

# Methodology

# Methodology

## Executive Summary

- Collect data using SpaceX REST API and web scrapping data from Wikipedia.
- Wrangle data – filtering, handling missing values and applying One-Hot Encoding to convert the categorical data to numerical for data analysis and prediction model.
- Exploratory Data Analysis with SQL to find the relation between factors effecting the success of landing the first stage of rocket. Also, finding new insights.
- Visualize the data using Folium and Dash for other stakeholders and ease of access to data.
- Modeling to predict landing outcome using classification techniques. Tune the model and find the best algorithms and its best parameter for a pipeline for future use.

# Data Collection – SpaceX API

**Request**
- Using Request Package get the byte response from SpaceX API
- Convert the Response to DataFrame

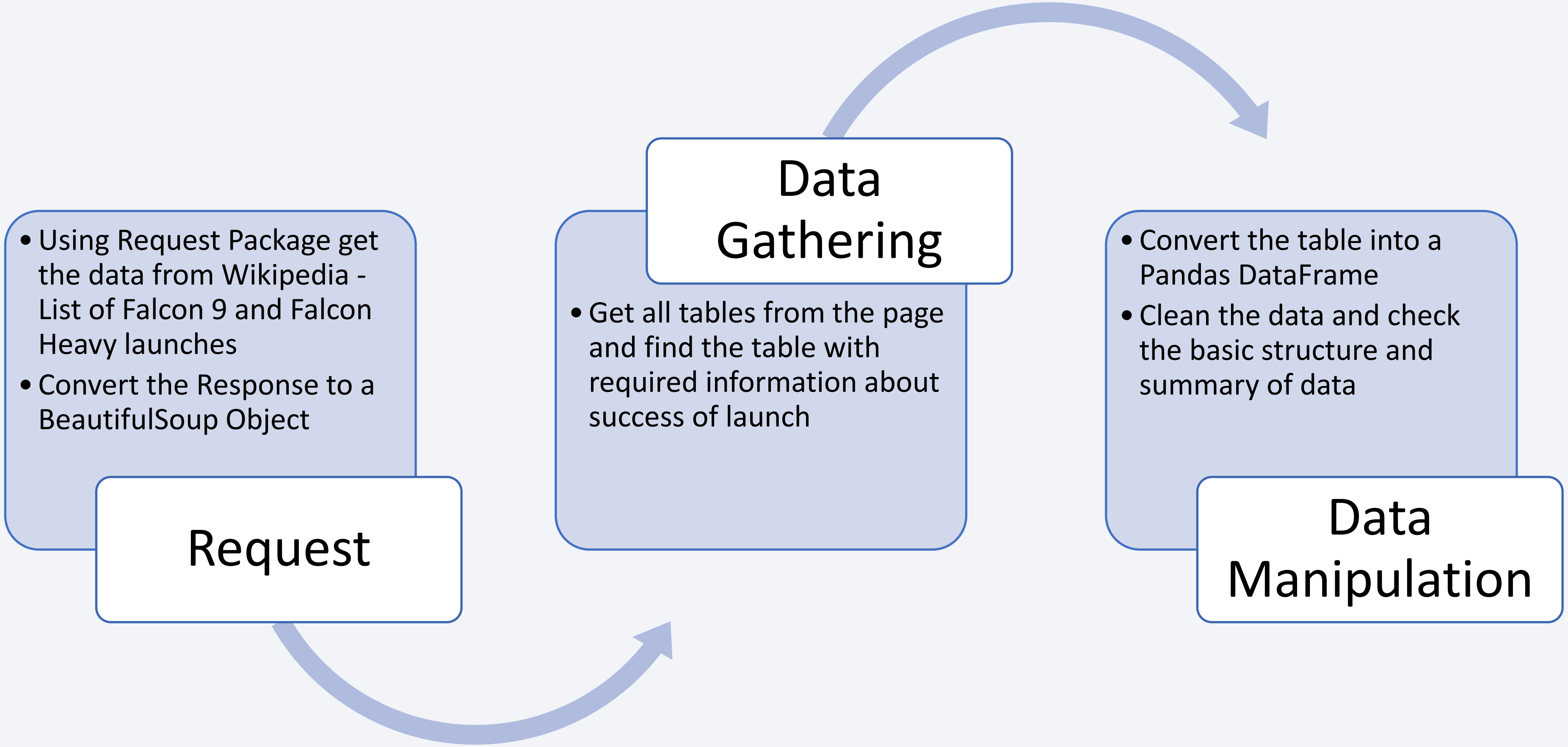
**Data Cleaning**
- Get the required columns and convert them to required data types

**Data Manipulation**
- Get the required meta data from SpaceX API
- Append and clean that to main DataFrame

**Data Wrangling**
- Get the required records for Falcon-9 only.
- Deal with the missing data

For Reference: [SpaceX API Data Collection Notebook]

# Data Collection - Scraping




For Reference – [Wikipedia Falcon 9 Data Scrapping]

# Data Wrangling

Steps:

- Perform EDA and determine the best training features
- Get a basic insight to extracted data and basic summary to each extracted feature
- Provide Insights:
  - # of launches from each Launch Site
  - # of launches in each orbit and number of mission in these orbit
- Create binary landing outcome feature for determining the outcomes from data

- Total number of outcomes:
  - True ASDS: Successful land on a drone ship
  - False ASDS: Failure to land on drone ship
  - True RTLS: Successful land on a land area
  - False RTLS: Failure to land on a land area
  - True Ocean: Successful land in ocean
  - False Ocean: Failure to land in ocean

For Reference – [Data Wrangling]

# EDA with Data Visualization

## Charts

- Flight Number vs Payload Mass
- Launch Site vs Flight Number
- Launch Site vs Payload Mass
- Outcome vs Orbit
- Orbit vs Flight Number
- Orbit vs Payload
- Launch Success Yearly Trend

## Analysis

- Viewing relationship between features to gain insights and gaining additional insight for model training for landing outcome
- Gaining insight how different factors play into or different categories work with landing outcome

For Reference – [Data Viz EDA]

# EDA with SQL

## Queries

- Display Unique Launch Sites
- First five records where Launch Site is in Cape Canaveral Air Force Station
- Total Payload Mass carried by boosters from NASA (CRS)
- Average Payload Mass carried by Falcon-9 v1.1
- First Successful ground landing
- Names of Boosters with successful drone ship landing
- Count of Landing outcomes
- Booster Names and versions with maximum payload mass
- Month and booster version with landing outcome – failure drone ship
- Count of landing outcomes between 04-06-2010 to 20-03-2017

For Reference – [EDA with SQL]

# Build an Interactive Map with Folium

## Markers

- Added NASA Johnson Space Center and Launch sites markers on map to provide a basic view of launch sites and distance from NASA Space Center
- Added colored marker for launch outcomes green – success and red – failure for the each launch site
- Also, added nearest city, highway, railway and coastline with distance to CCAFS Air Station

For Reference – Launch Site GeoViz

# Build a Dashboard with Plotly Dash

- Dropdown list with Launch Sites
  - Allow user to select All launch sites or a specific launch site
- Pie Chart showing Successful Launches
  - Using the dropdown a pie chart is created showing Successful launches against failed launches for specified launch site
- Slider for Payload Mass
  - Allow user to select payload mass range
- Scatter Chart Payload Mass vs Class
  - Scatter plot for Success rate for specified launch site at specified Payload Mass range

For Reference – [Dashboard SourceCode]

# Predictive Analysis (Classification)

- Convert the DataFrame into Numpy Array for model fitting
- Standardized the data for lower probability of feature favourism
- Split the data for training and testing the models
- Create a GridSearchCV for Cross Validation of different parameters for each model
- Find the best model and its parameter between:
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree Classifier
  - K- Nearest Neighbors
- Calculate the accuracy for each model with it’s best parameters
- Create a confusion matrix for evaluation of the models
- Identify the best model using Jaccard Index, F1-Score and Accuracy

# Results

## Executive Summary

- Exploratory data analysis results
  - Launch success improved
  - KSC LC-39A has the highest success rate
- Visual Analytics
  - Most launches are near equator and coast
  - Launch sites are far from cities, highway or railway (public infrastructure) but are close to Coastlines
- Predictive analysis results
  - Decision Tree Classifier is the best model to predict the landing outcome

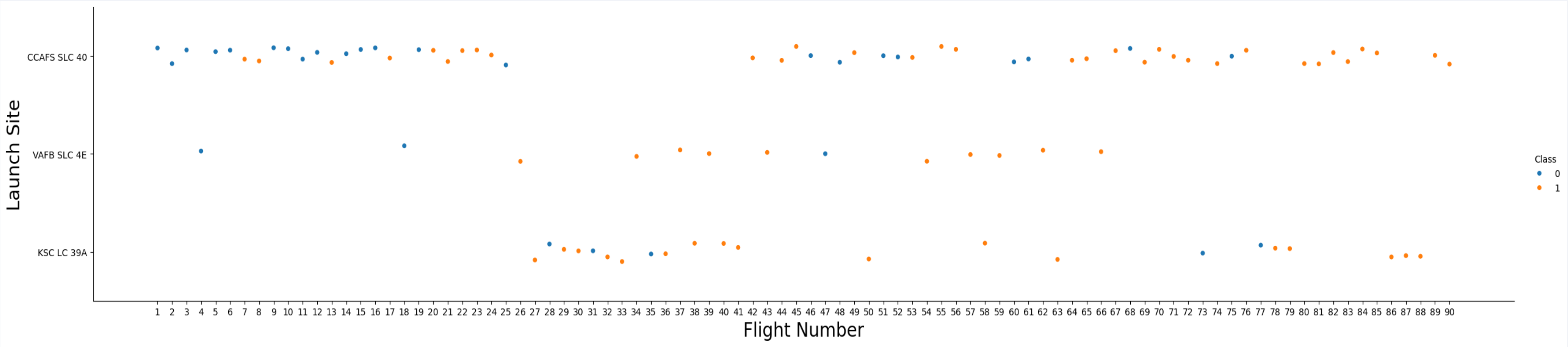
Section 2

# Insights drawn from EDA

# Flight Number vs. Launch Site

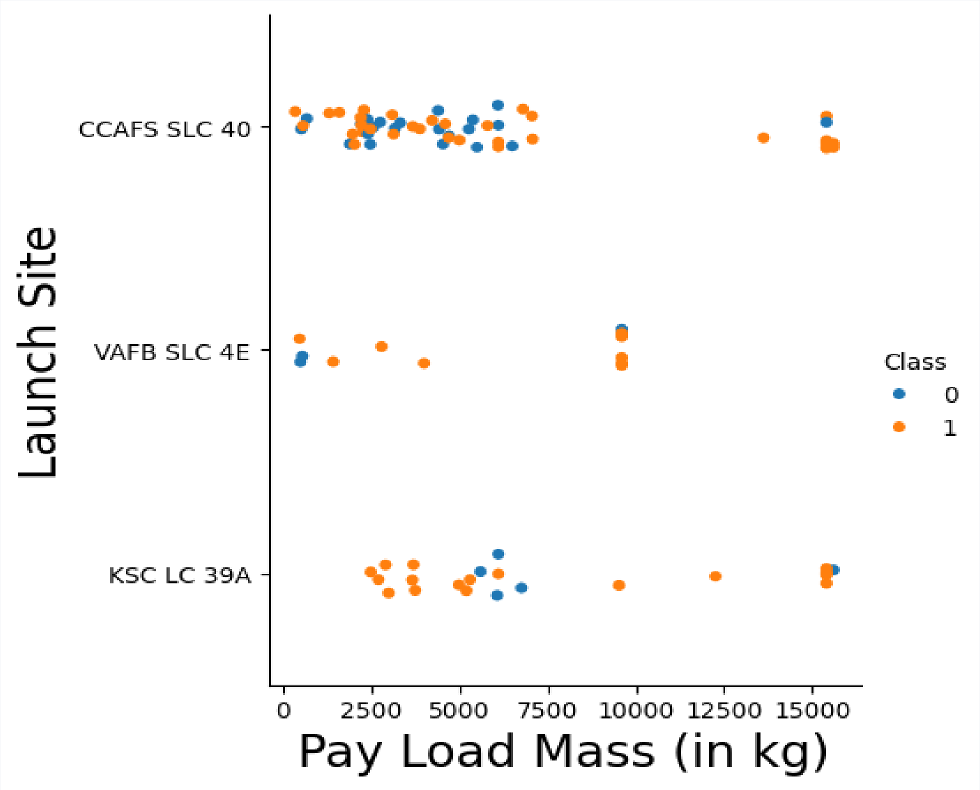
EDA

- Earlier flight had a lower success rate, whereas later the success rate increased
- CCAFS SLC-40 was most used launch site for Falcon-9 flights
- VAFS SLC 4E and KSC LC 39A have better success rates, but started hosting the launches after CCAFS SLC-40



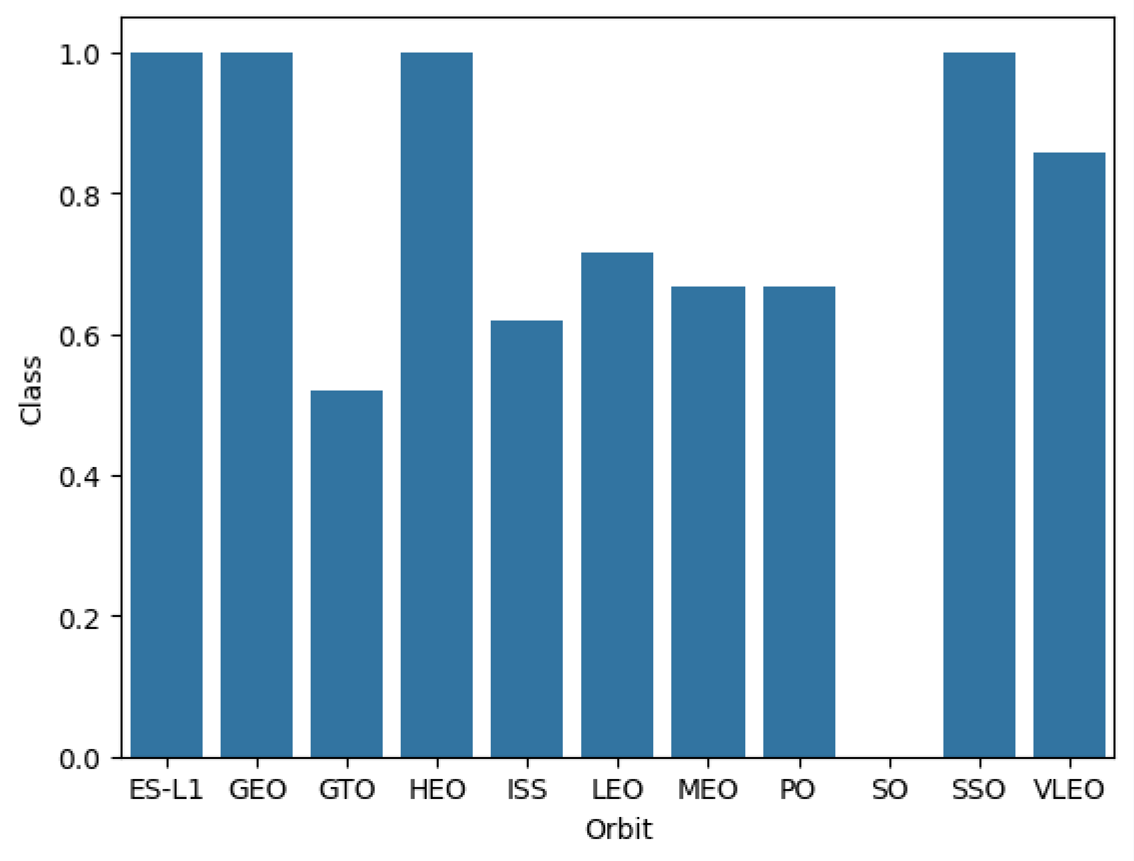
# Payload vs. Launch Site

- Most launches payload mass was between 1000kg to 10000kg
- Launch Success Rate is higher when payload mass is higher >10000kg
- VAFB SLC 4E flights took maximum of 10000kg of payload mass
- **Higher the payload mass, higher the chances of mission to be successful**



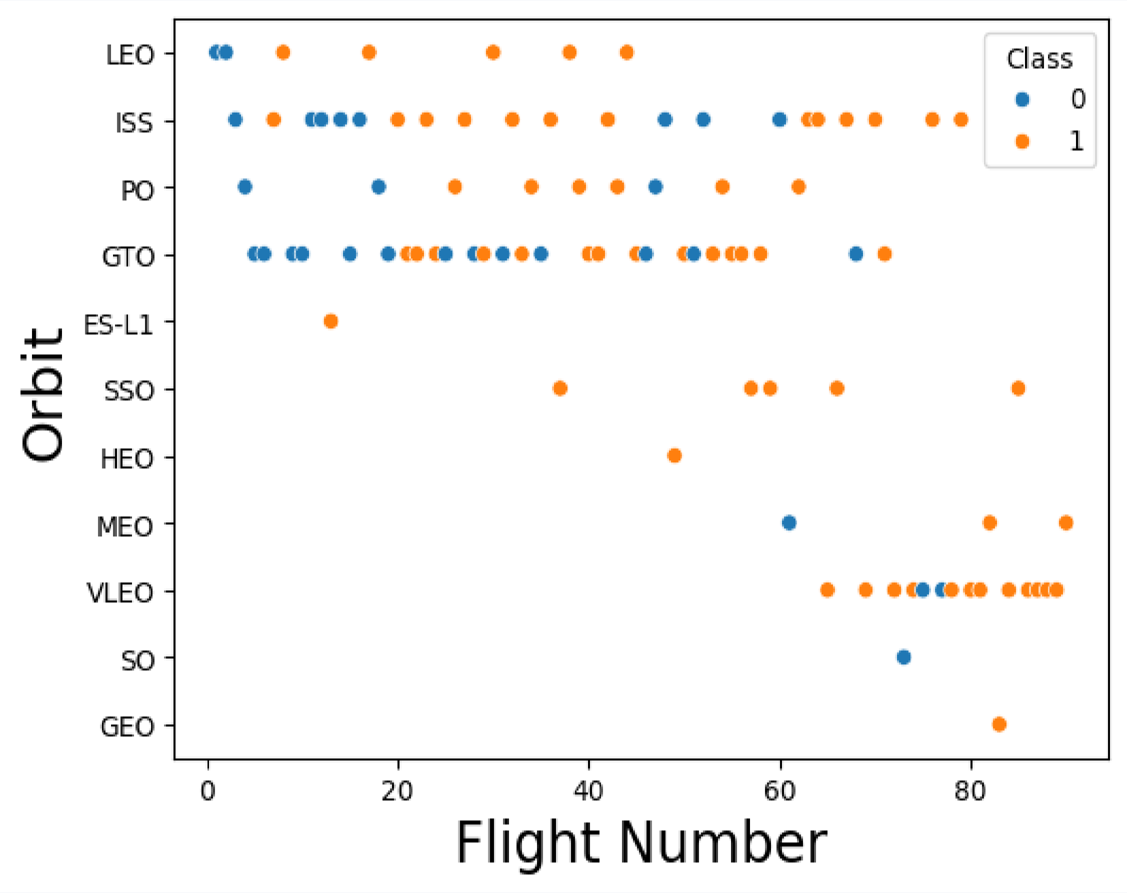
# Success Rate vs. Orbit Type

- Success Rate for ES-L1, GEO, HEO, SSO is 100%
- GTO, ISS, LEO, MEO, PO, VLEO saw few failed mission
- SO never saw any success mission



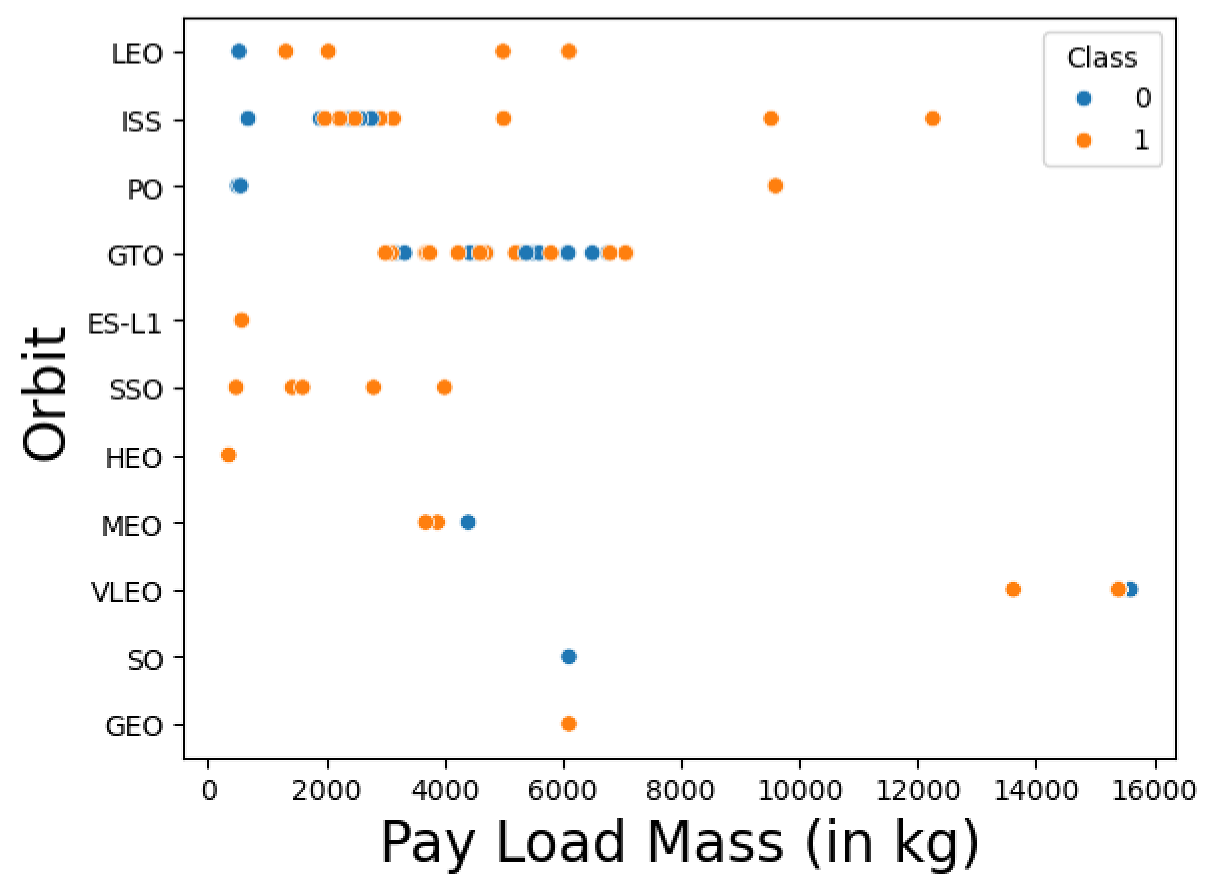
# Flight Number vs. Orbit Type

- GEO, SO, VLEO, MEO, HEO, SSO mission started in later stages
- Success rate improved in each orbit

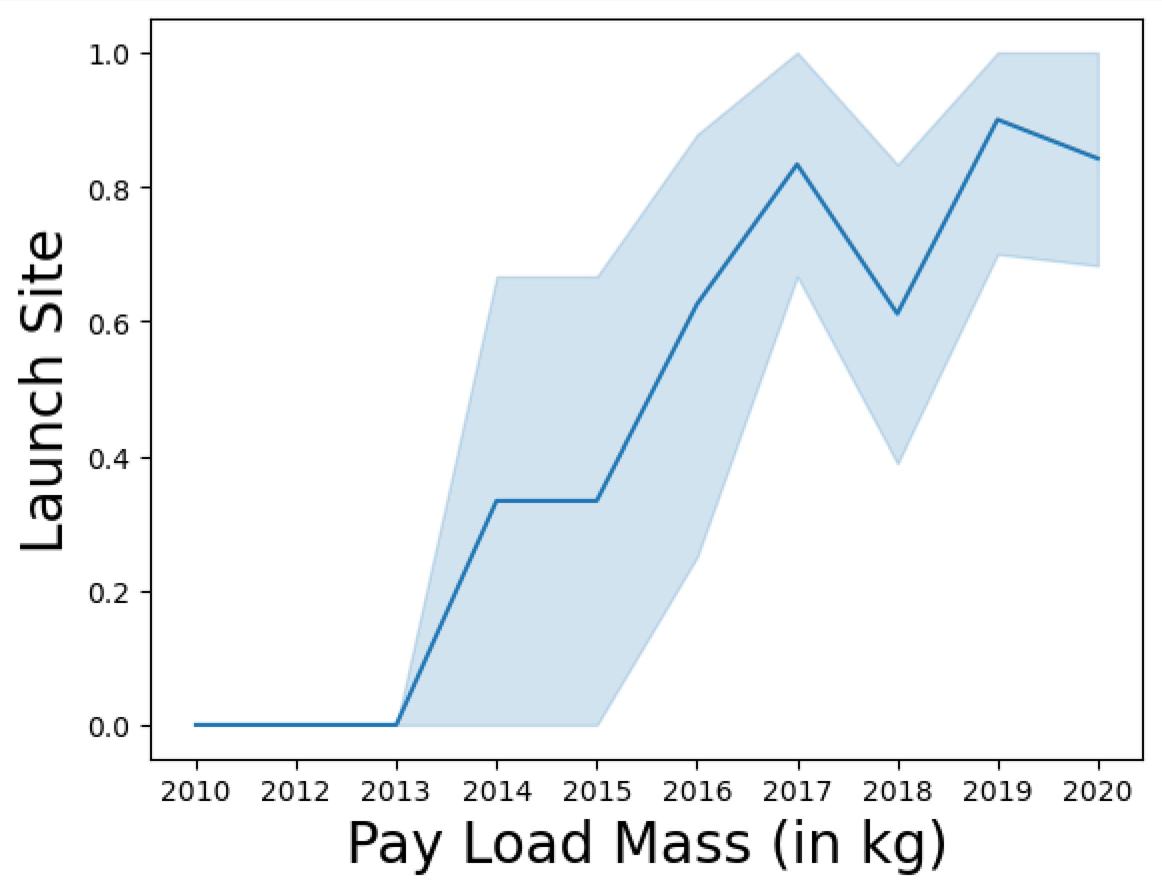


# Payload vs. Orbit Type

- ES-L1, SSO, HEO with 100% success rate took lower payload mass mission
- GTO flights payload mass is ranged between 3000 to 8000 kg
- LEO, ISS, PO saw improvement with increase in payload mass



# Launch Success Yearly Trend




- Trend show improvement/success started from year 2013

# All Launch Site Names

- From this query, we can find basic start to analysis by planning and marking different launch sites used for Falcon-9 flights:

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

```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTBL
```

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

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

```
%sql SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE '%CCA%' LIMIT 5
```

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

| 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

```
%%sql SELECT SUM(PAYLOAD_MASS__KG_) AS Total_NASAPayload_Mass
FROM SPACEXTBL
WHERE Customer LIKE '%CRS%'
```

✓ 0.0s

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

| Total_NASAPayload_Mass |
|---|
| 48213 |

- Total Payload Mass launched by NASA (CRS) is around 48213kg

# Average Payload Mass by F9 v1.1

- Average Payload Mass carried by Falcon 9 v1.1 is around 2534.67kg

```
%%sql SELECT AVG(PAYLOAD_MASS__KG_) AS Average_F9_Payload
FROM SPACEXTBL
WHERE Booster_Version LIKE '%F9 v1.1%'
```

✓ 0.0s

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

| Average_F9_Payload |
| --- |
| 2534.6666666666665 |

# First Successful Ground Landing Date

- First Successful landing on a ground pad was on 22 December, 2015

```
%%sql SELECT MIN(Date) AS First_Success_Land

FROM SPACEXTBL
WHERE Landing_Outcome LIKE '%Success (ground%'
```

✓ 0.0s

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

| First_Success_Land |
|---|
| 2015-12-22 |

# Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql SELECT Booster_version
FROM SPACEXTBL
WHERE Landing_Outcome LIKE '%Success%' AND Landing_Outcome LIKE '%Success (drone%' 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 Mission_Outcome, COUNT(Mission_Outcome) AS Count
FROM SPACEXTBL
GROUP BY Mission_Outcome
ORDER BY Count DESC
```

✓ 0.0s

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

| Mission_Outcome | Count |
|---:|---:|
| Success | 98 |
| Success (payload status unclear) | 1 |
| Success | 1 |
| Failure (in flight) | 1 |

# Boosters Carried Maximum Payload

```
%%sql
SELECT Booster_version
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

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

| 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

- Failed Landing Outcomes in year 2015 were both from CCAFS LC-40 launch site

**Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.**

```
%%sql
Select substr(Date,6,2) as Month, Landing_Outcome, Booster_version, Launch_site
FROM SPACEXTBL
WHERE Landing_Outcome LIKE "%Failure%" and Landing_Outcome LIKE "%drone%" AND
substr(Date,0,5) = '2015'
```

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

| Month | Landing_Outcome | Booster_Version | Launch_Site |
|---|---|---|---|
| 01 | Failure (drone ship) | F9 v1.1 B1012 | CCAFS LC-40 |
| 04 | Failure (drone ship) | F9 v1.1 B1015 | CCAFS LC-40 |

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

- Rank 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

```
%%sql
SELECT Mission_Outcome, COUNT(Mission_Outcome) AS Count
FROM SPACEXTBL
WHERE (Date>='2010-06-04' AND Date<='2017-03-20')
GROUP BY Mission_Outcome
ORDER BY Count DESC
```

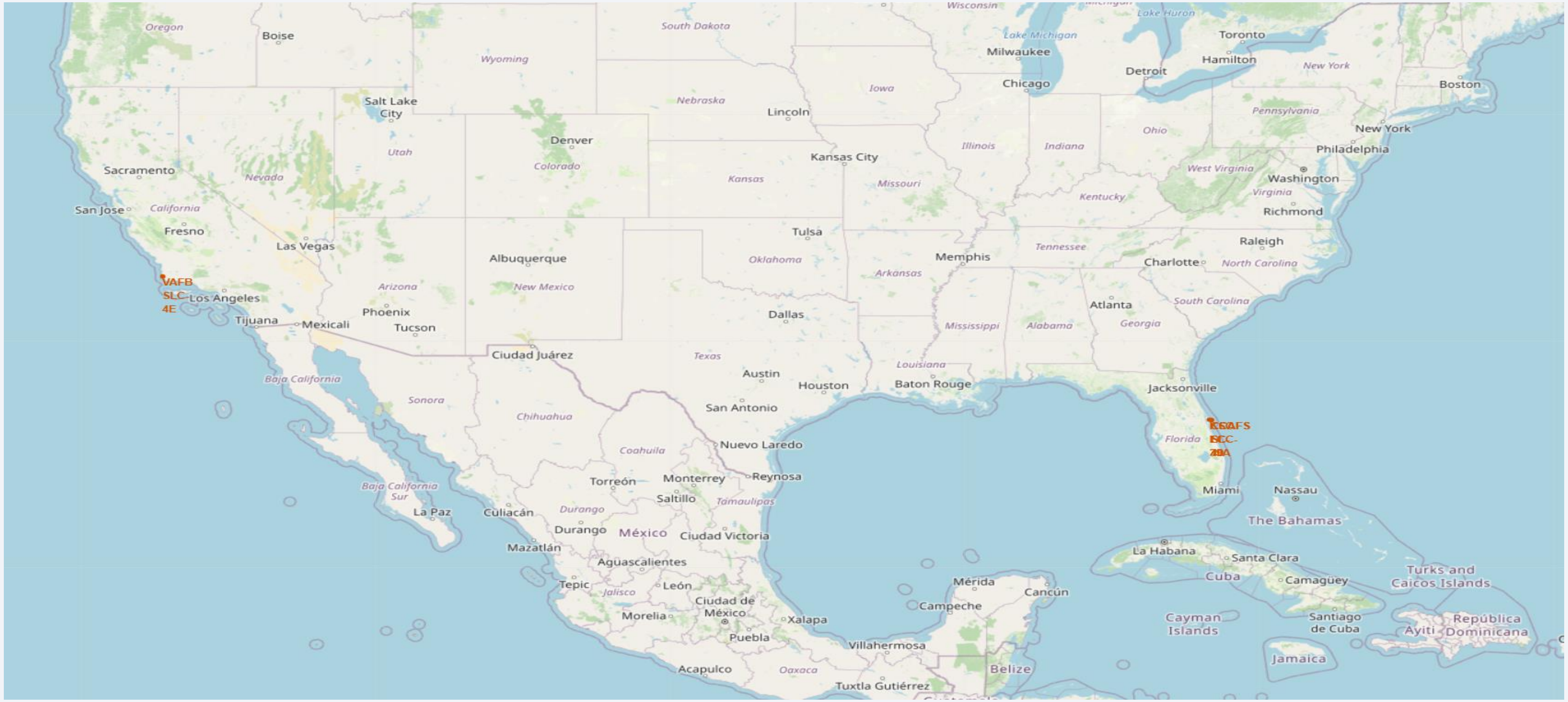
```
 * sqlite:///my_data1.db
Done.
```

| Mission_Outcome | Count |
|---|---|
| Success | 30 |
| Failure (in flight) | 1 |

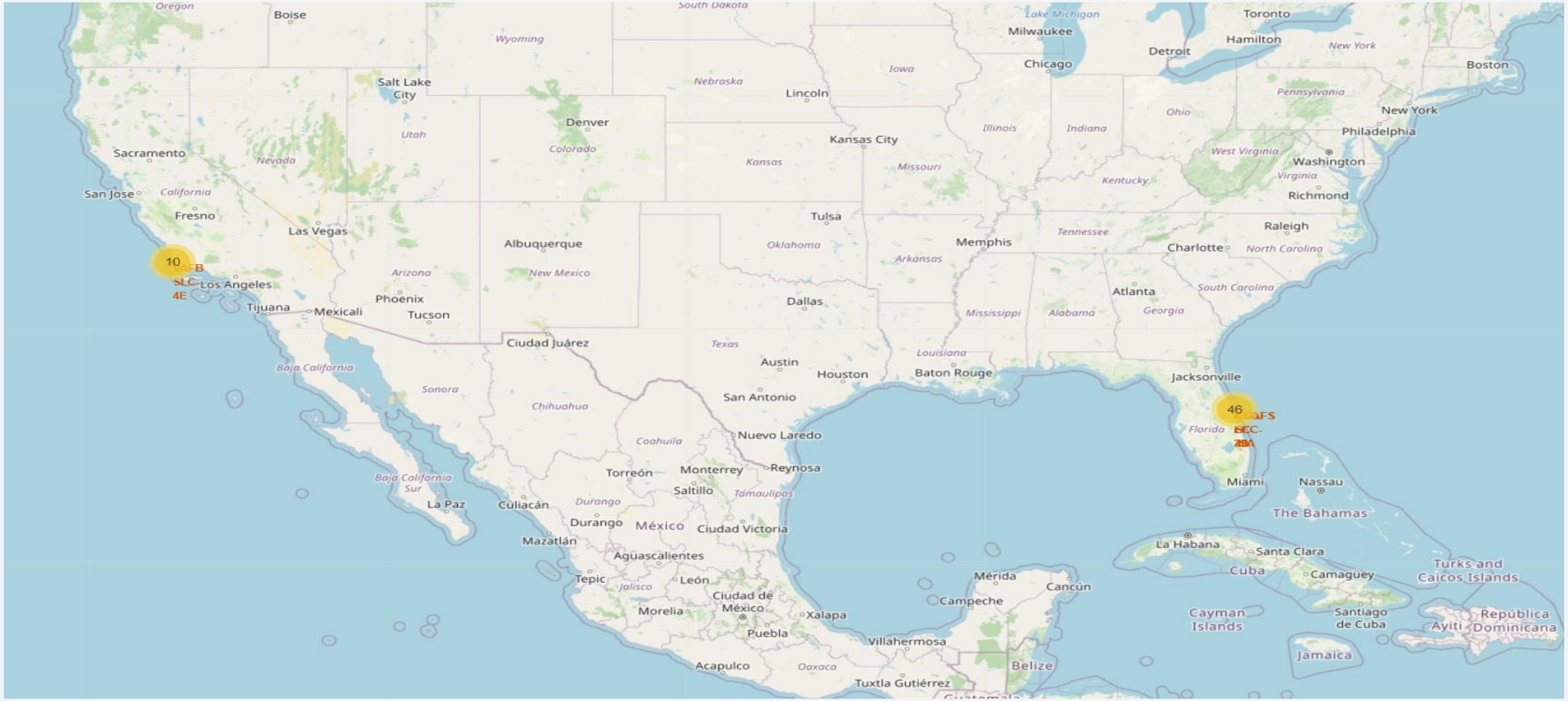
Section 3

# Launch Sites Proximities Analysis

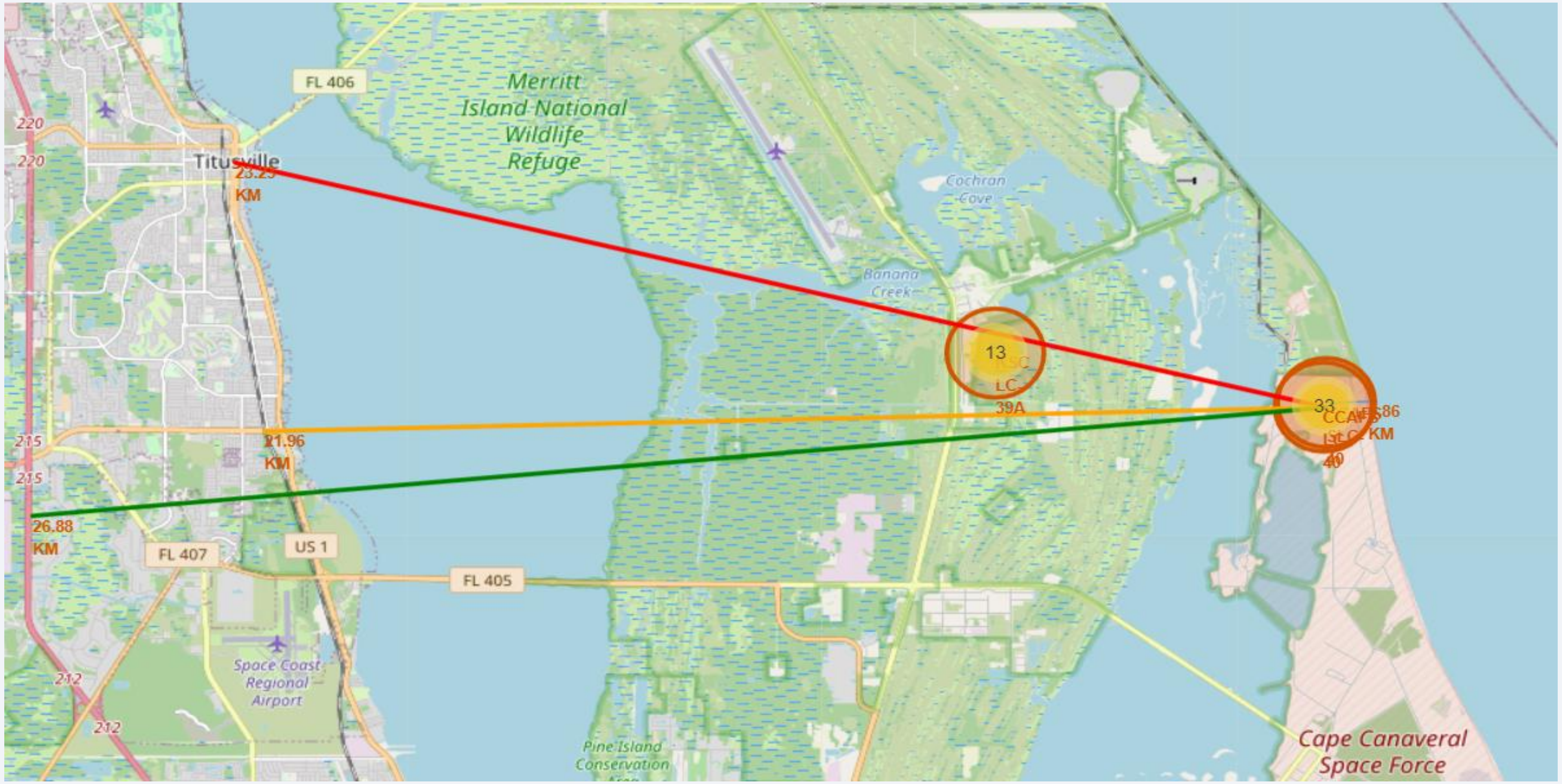
# Launch Sites



# Mission From Each Launch Sites



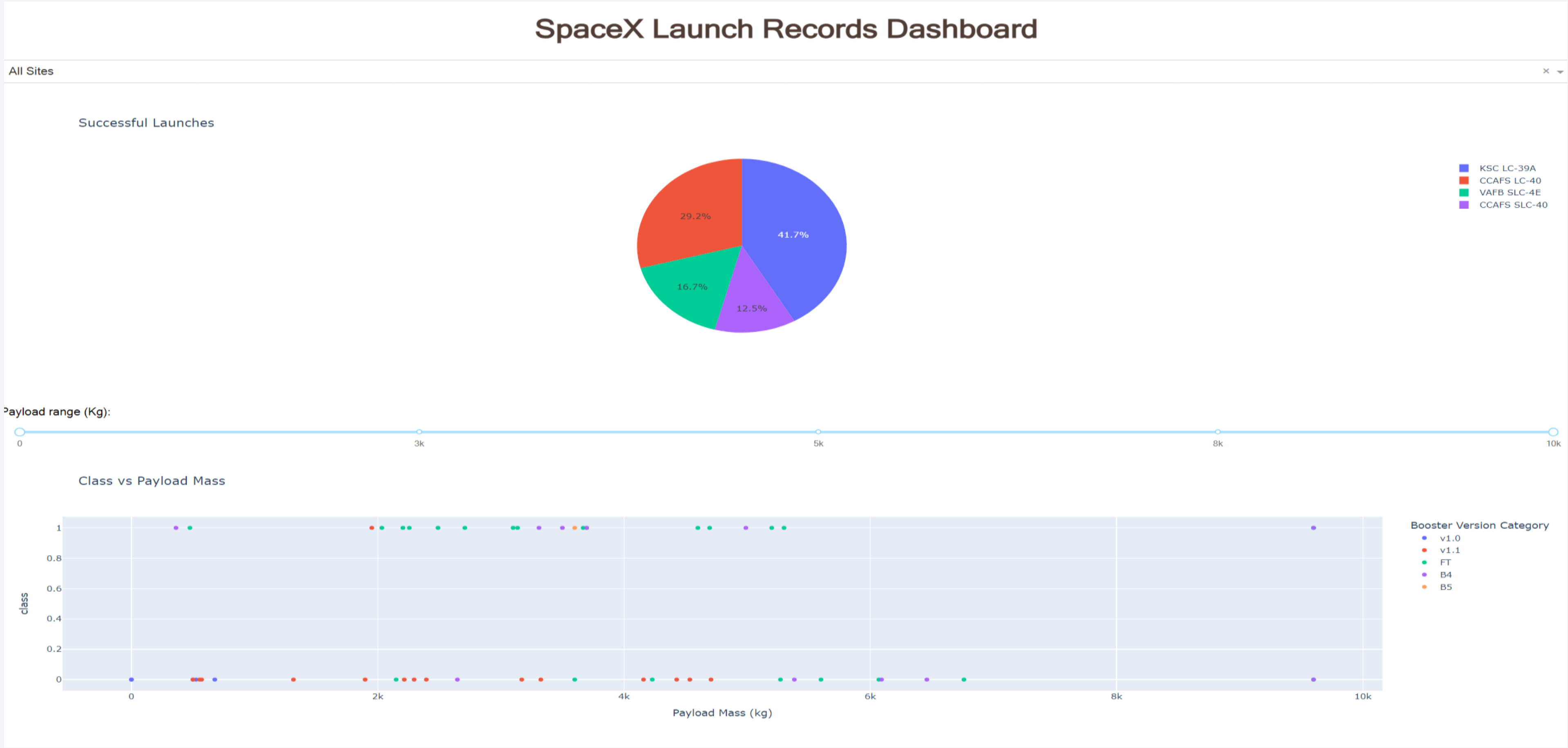
# Public Infrastructure from CCA



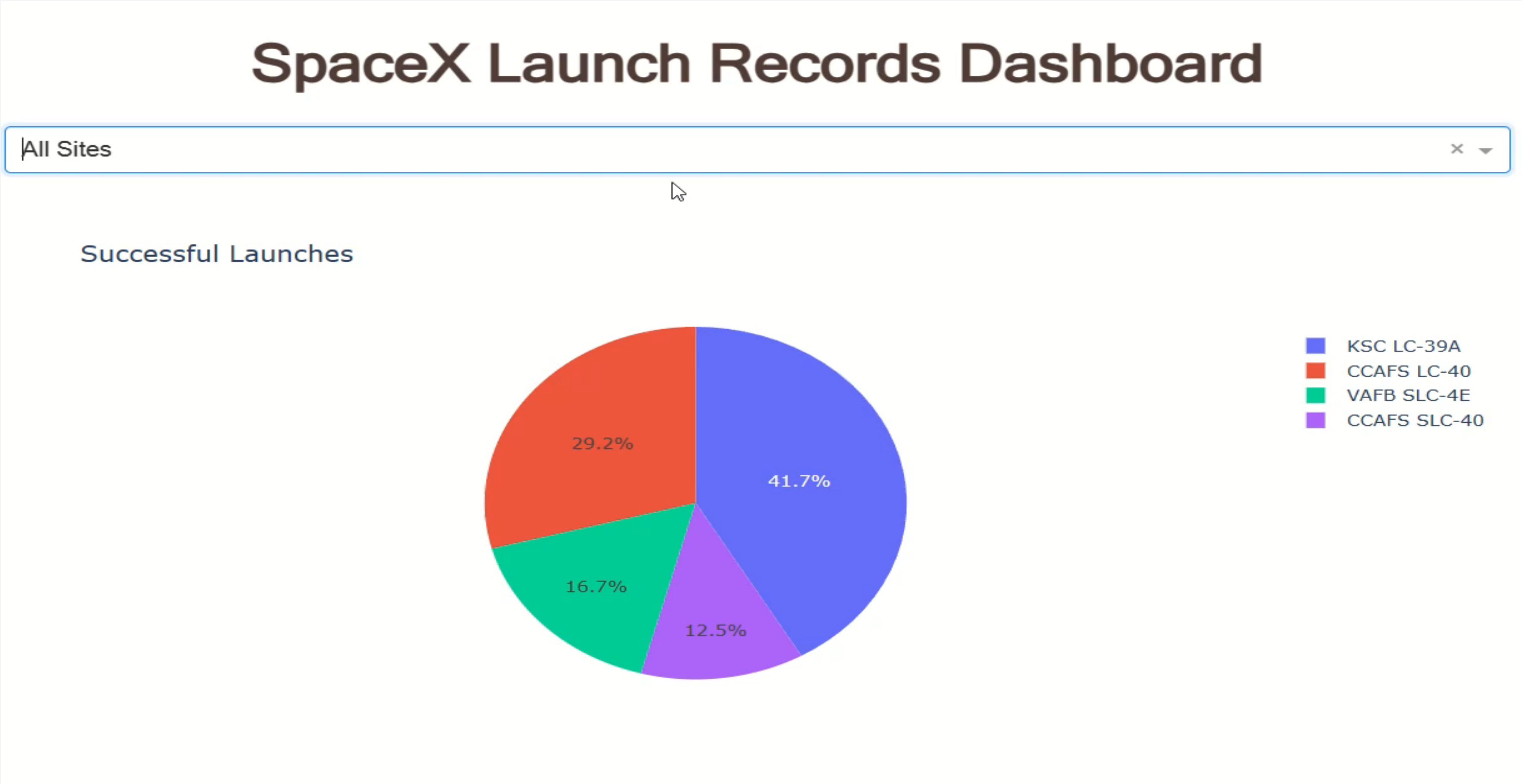
Section 4

# Build a Dashboard with Plotly Dash

# Dashboard Overview



# Pie Chart


SpaceX Launch Records Dashboard
All Sites
Successful Launches
29.2%
41.7%
16.7%
12.5%
KSC LC-39A
CCAFS LC-40
VAFB SLC-4E
CCAFS SLC-40

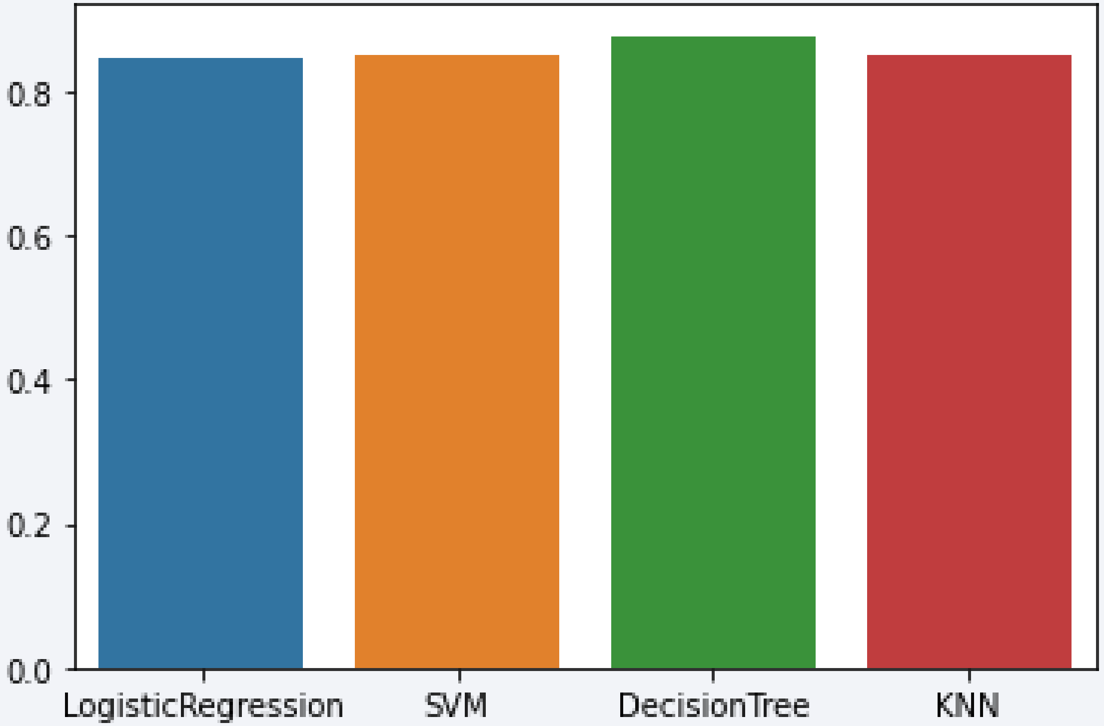
# Scatter Plot

Section 5

# Predictive Analysis (Classification)

# Classification Accuracy

- Decision outperformed other models by a slight margin and highest accuracy = 87.67%



# Confusion Matrix

Decision Tree correctly labelled 16 records whereas other only labeled 15 correctly

# Conclusions

- Model Performance: Decision Tree outperformed other models
- Most launches were close to equator
- Launch sites are close to coastline and far from public infrastructure
- Launch success improved overtime
- Few orbits such as ES-L1, GEO, HEO saw 100% success rate whereas SO didn't saw any success

# Appendix

Things to consider:

- Dataset: Dataset was quite skewed and may see different results once there is more data available
- Feature Analysis: Many features were not taken into account that may affect the failure or success of landing
- Many models that are available to use were not used in predictive analysis that may have better results

# Thank you!