IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Chiranjeevi Nalapalu
20.01.2024

Outline

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

Executive Summary

- Data collection methodology was Using API's and Webscraping. Performed data wrangling using Python packages Pandas and NumPy. Performed exploratory data analysis (EDA) using visualization and SQL. Performed interactive visual analytics using Folium and Plotly Dash. Performed predictive analysis using classification models
- Flights gradually increased their success rate with time. Higher payload mass seems more successful. ES-L1, GEO, HEO and SSO orbits have highest success rates/ VLEO orbits have highest Payload mass. Success rate of launches increases year after year. Booster version category FT is the most successful for lower payloads but less successful for higher payload masses. KSC LC-39A has the highest percentage of successful launches. Logistic the best performing model as it has an accuracy of 0.9444

Introduction

- 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.
- Objectives
  - To determine the price of each launch.
  - To determine if SpaceX will reuse the first stage

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using API's ans Webscraping
- Perform data wrangling
  - Using Python packages Pandas and NumPy
- 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 was collected via 2 methods.
    1. Requesting and cleaning data from the SpaceX API
    2. Extracting Falcon 9 HTML tables from Wikipedia and converting it to a Data Frame

Data Collection – SpaceX API

1. Request and parse the SpaceX launch data using the GET request
2. Filter the dataframe to only include Falcon 9 launches
3. Dealing with Missing Values

- https://github.com/ChiranNala/Applied-Data-Science-Capstone/blob/55cf767746cbc4f740e5bf8a0983473facd4f156/Lab1_Collecting_the_data.ipynb

Data Collection - Scraping

1. Request the Falcon9 Launch Wiki page from its URL
2. Extract all column/variable names from the HTML table header
3. Create a data frame by parsing the launch HTML tables

- https://github.com/ChiranNala/Applied-Data-Science-Capstone/blob/55cf767746cbc4f740e5bf8a0983473facd4f156/Web_scraping_Falcon_9_and_Falcon_Heavy.ipynb

Data Wrangling

- Data Wrangling was done with Python packages Pandas and NumPy
  - Calculate the number of launches on each site
  - Calculate the number and occurrence of each orbit
  - Create a landing outcome label from Outcome column

- https://github.com/ChiranNala/Applied-Data-Science-Capstone/blob/55cf767746cbc4f740e5bf8a0983473facd4f156/Lab2_Data_wrangling.ipynb

EDA with Data Visualization

- Three bar plots, one scatter plot and one line plot were plotted
  - Bar plot 1: Visualize the relationship between success rate of each orbit type
  - Bar plot 2: Visualize the relationship between FlightNumber and Orbit type
  - Bar plot 3: Visualize the relationship between Payload and Orbit type
  - Scatter plot: Visualize the relationship between Payload and Orbit type
  - Line plot: Visualize the launch success yearly trend
- https://github.com/ChiranNala/Applied-Data-Science-Capstone/blob/55cf767746cbc4f740e5bf8a0983473facd4f156/Assignment_Exploring_and_Preparing_Data.ipynb

EDA with SQL

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first succesful landing outcome in ground pad was acheived.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
- List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
- 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.
- https://github.com/ChiranNala/Applied-Data-Science-Capstone/blob/55cf767746cbc4f740e5bf8a0983473facd4f156/SQL_Notebook_for_Peer_Assignment.ipynb

Build an Interactive Map with Folium

- Mark all launch sites on a map using Circles and Markers
- Mark the success/failed launches for each site on the map using MarkerCluster
- Calculate the distances between a launch site to its proximities

- To better visualize data spatially on a map

- https://github.com/ChiranNala/Applied-Data-Science-Capstone/blob/299f443b4997fd040a2eeae73df48c79d4fa0e83/Launch_Sites_Locations_Analysis_with_Folium.ipynb

Build a Dashboard with Plotly Dash

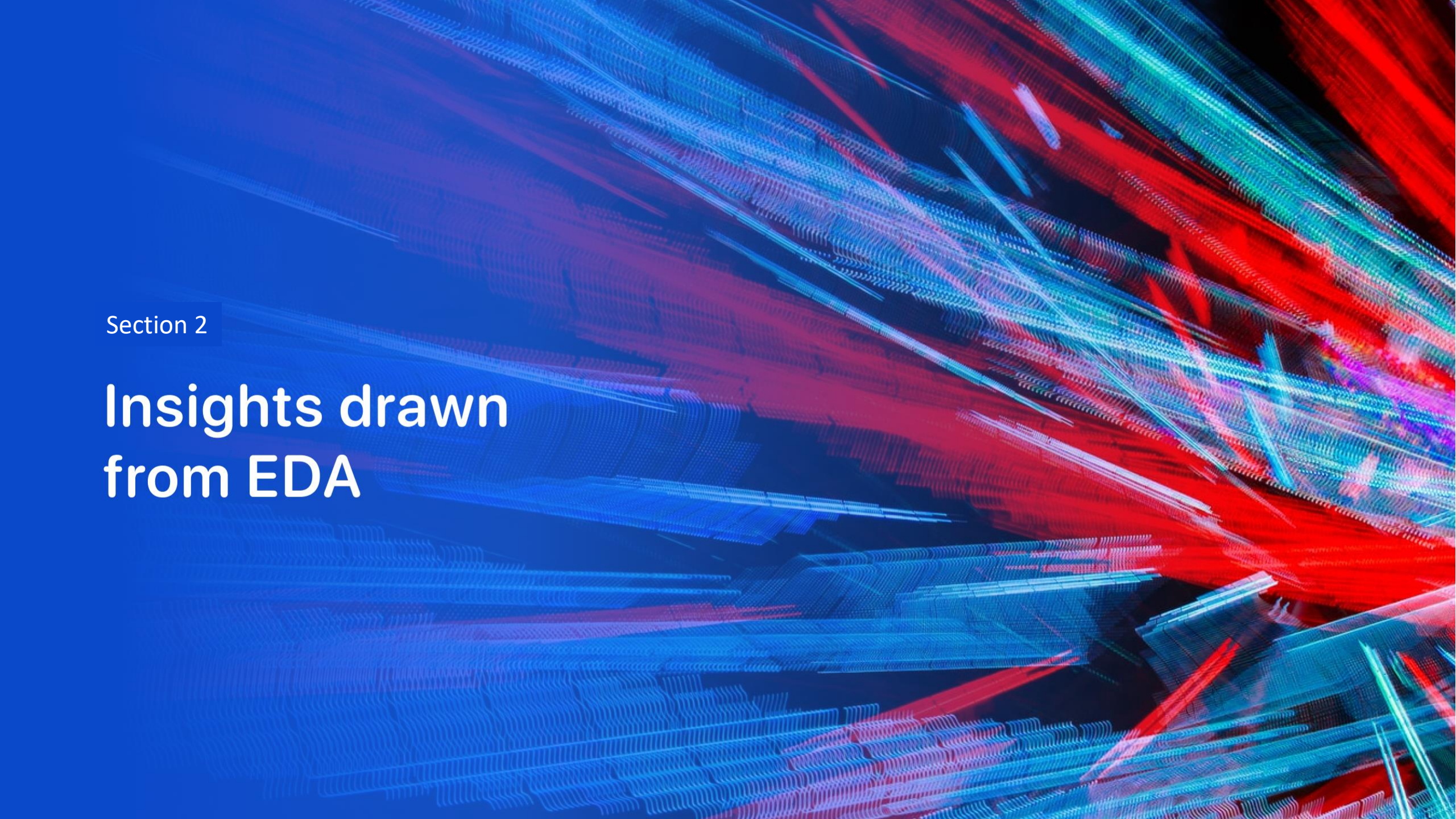
- Pie charts and a scatter plot were created to answer the following questions
- Which site has the largest successful launches?
- Which site has the highest launch success rate?
- Which payload range(s) has the highest launch success rate?
- Which payload range(s) has the lowest launch success rate?
- Which F9 Booster version (v1.0, v1.1, FT, B4, B5, etc.) has the highest launch success rate?
- https://github.com/ChiranNala/Applied-Data-Science-Capstone/blob/a0029be32fd9edeb4a4a07c4972567f7559f02e7/spacex_dash_app%20(1).py

Predictive Analysis (Classification)

- Performed exploratory Data Analysis and determine Training Labels
- Created a column for the class
- Standardized the data
- Split into training data and test data
- Found best Hyperparameter for SVM, Classification Trees and Logistic Regression
- Found the method performs best using test data
- https://github.com/ChiranNala/Applied-Data-Science-Capstone/blob/a0029be32fd9edeb4a4a07c4972567f7559f02e7/Assignment_Machine_Learning_Prediction.ipynb

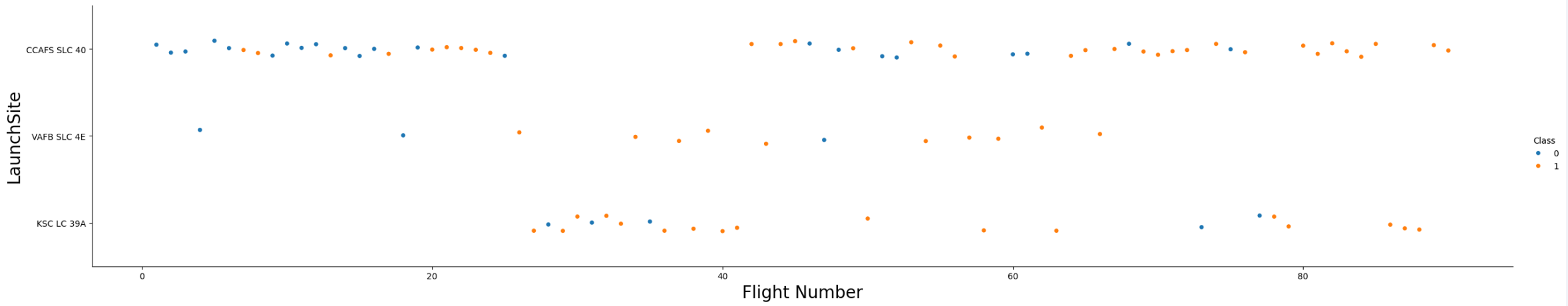
Results

- Exploratory data analysis results
  - Flights gradually increased their success rate with time
  - Higher payload mass seems more successful
  - ES-L1, GEO, HEO and SSO orbits have highest success rates
  - VLEO orbits have highest Payload mass
  - Success rates increase with time
- Predictive analysis results
  - Logistic regression, SVM and KNN show the same highest accuracy of 0.944
  - Decision tree shows lowest accuracy with accuracy 0.833


Section 2
Insights drawn from EDA

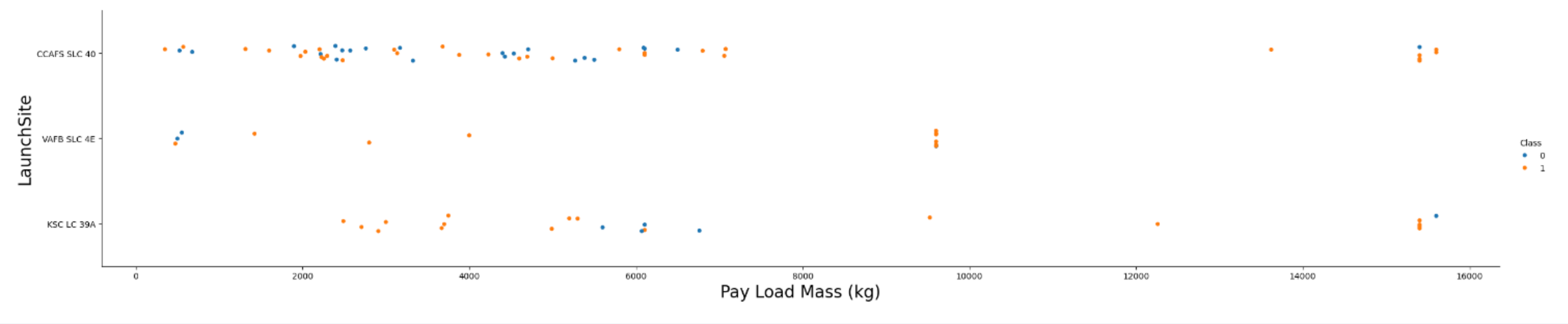
Flight Number vs. Launch Site

- Scatter plot of Flight Number vs. Launch Site
- We can see that initial flights were unsuccessful and were from launch site CCAFS SLS 40 and VAFB SLS 4E and none from KSC IC 39A
- In later flights we can see they are majorly successful and from Launch site and are from CCAFS SLS 40 and KSC IC 39A only



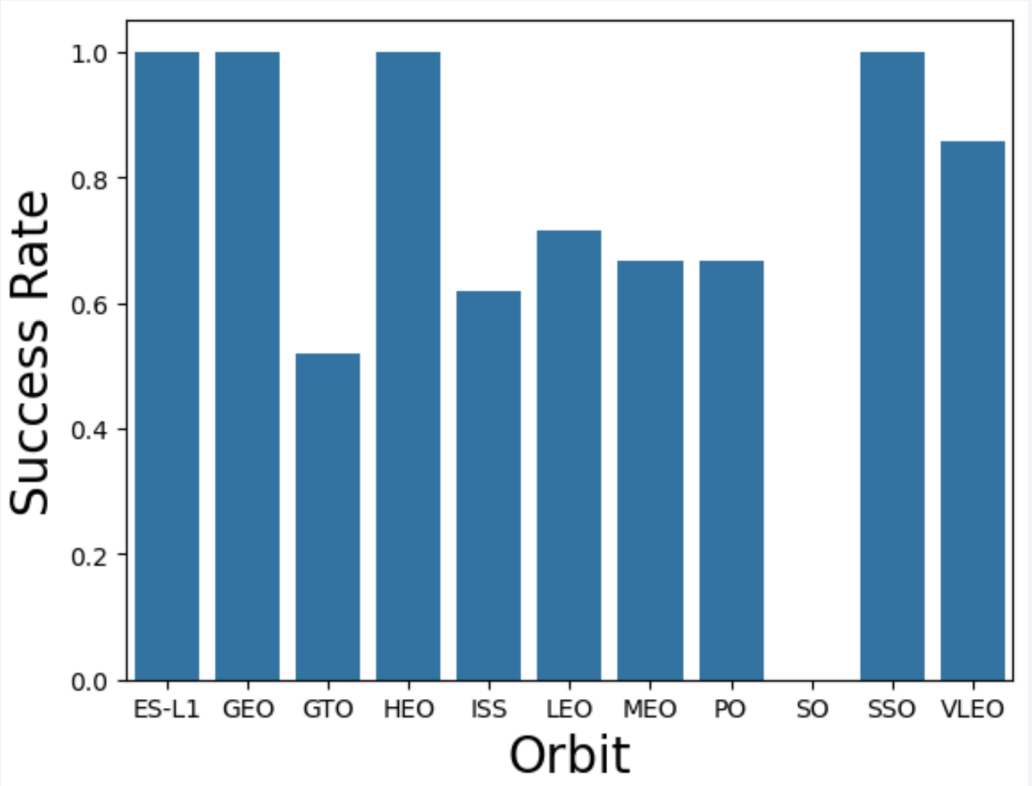
Payload vs. Launch Site

- Scatter plot of Payload vs. Launch Site
- Larger payloads are more successful and are launched from CCAFS SLS 40 and KSC IC 39A
- Smaller payloads are have a lower success rate and are launched from CCAFS SLS 40 and VAFB SLS 4E



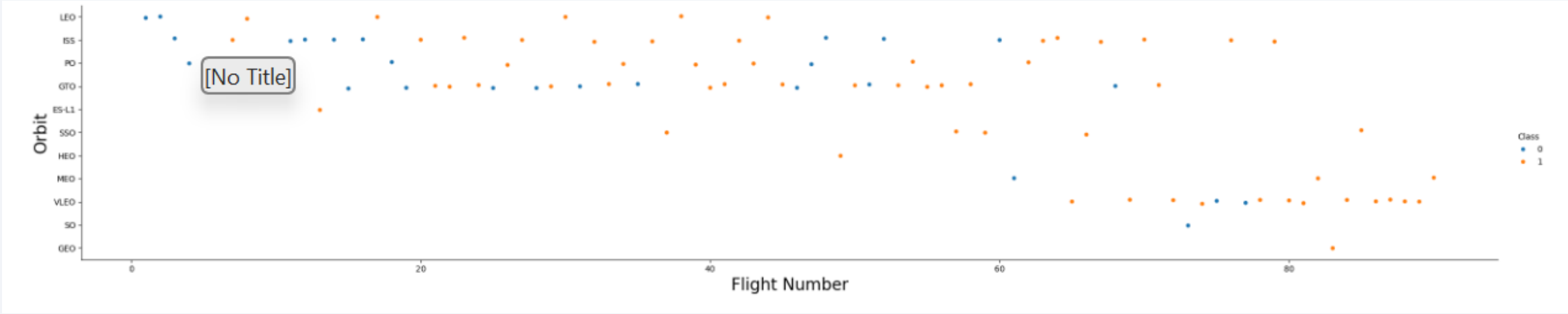
Success Rate vs. Orbit Type

- Bar chart for the success rate of each orbit type
- ES-L1, GEO, HEO and SSO orbits have highest success rates
- SO, GTO and ISS have the lowest



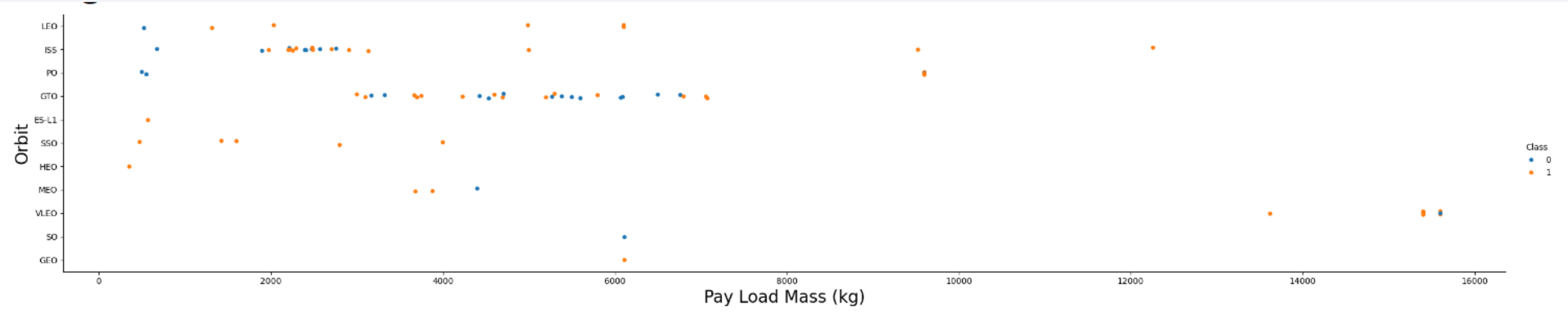
Flight Number vs. Orbit Type

- Show a scatter point of Flight number vs. Orbit type
- Higher flight numbers are more successful than lower ones.
- Large variation of success rate for lower flight numbers



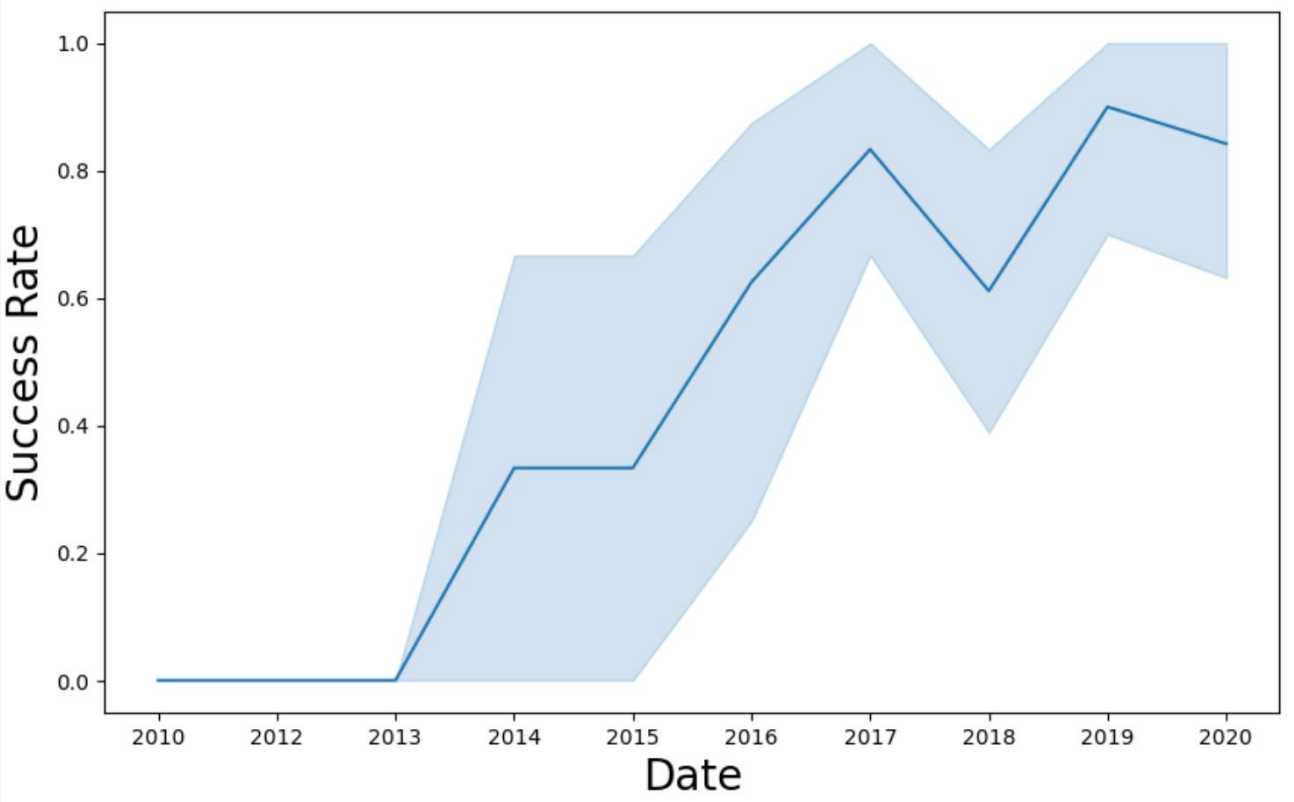
Payload vs. Orbit Type

- Scatter point of payload vs. orbit type
- Larger payloads are sent into orbit VLEO and PO
- Larger payloads are more successful especially above 800kg



Launch Success Yearly Trend

- Line chart of yearly average success rate
- Success rate increases year after year
- Larger variation in success rate in early years but becomes less in later years
- A dip of 20% percent seen from 2017 to 2019



All Launch Site Names

- Used DISTINCT command

```
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE
```

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

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

Launch Site Names Begin with 'CCA'

- Used LIKE command

```
%sql SELECT * FROM SPACEXTABLE 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

- Used SUM and WHERE command

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)'
```

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Used AVG and WHERE command

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1'
```

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

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- Used MIN and WHERE command

```
%sql SELECT MIN(DATE) FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)'
```

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

MIN(DATE)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Used WHERE and BETWEEN command

```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 40
```

```
 * 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

- Used COUNT and sub-querry

```
%sql SELECT COUNT(*) FROM SPACEXTABLE WHERE Mission_Outcome = (SELECT Mission_Outcome FROM SPACEXTABLE WHERE Mission_Outcome
```

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

COUNT(*)
100

Boosters Carried Maximum Payload

- Used SELECT and MAX command with sub-querry

```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
```

```
 * 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

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

- Present your query result with a short explanation here

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

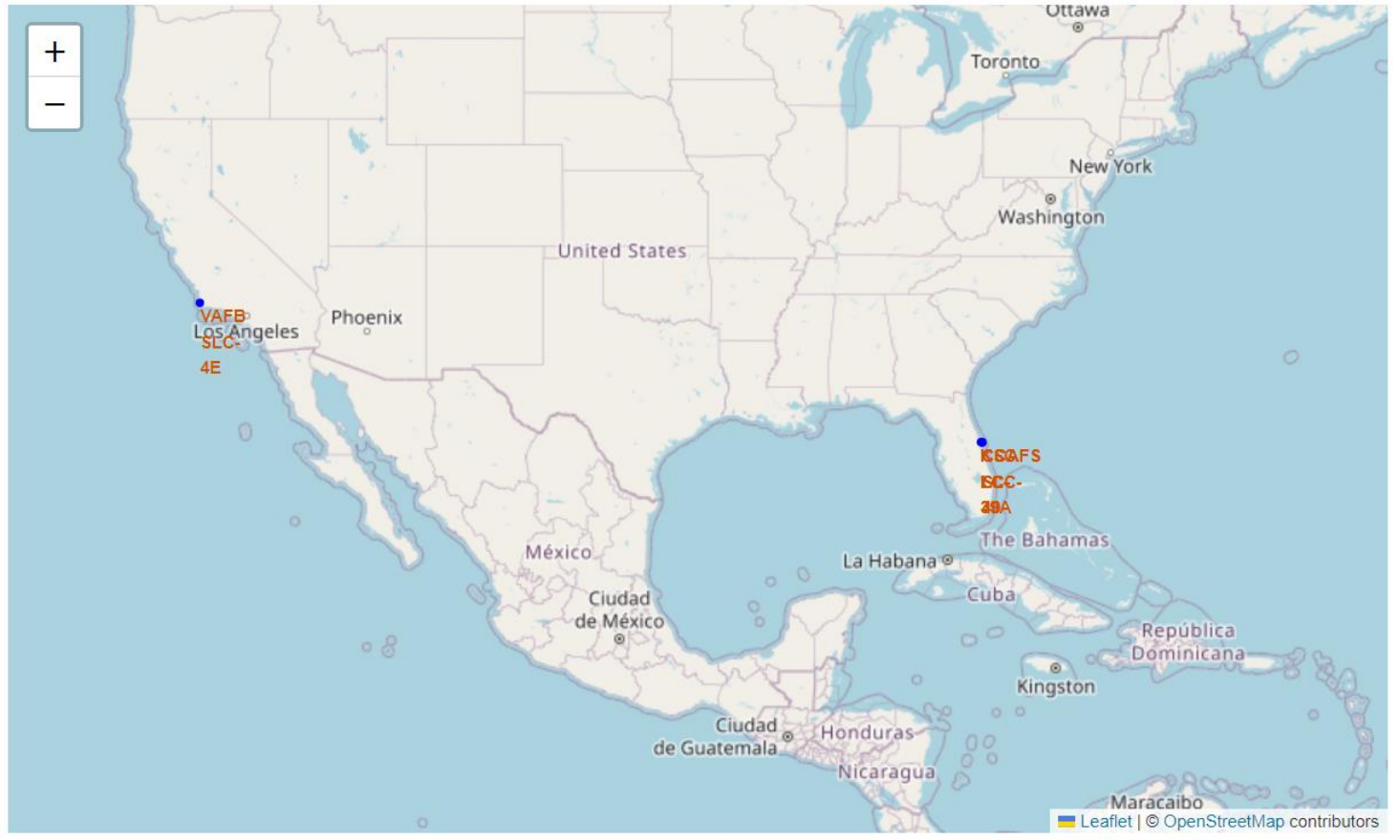
- Present your query result with a short explanation here

Section 3

Launch Sites Proximities Analysis

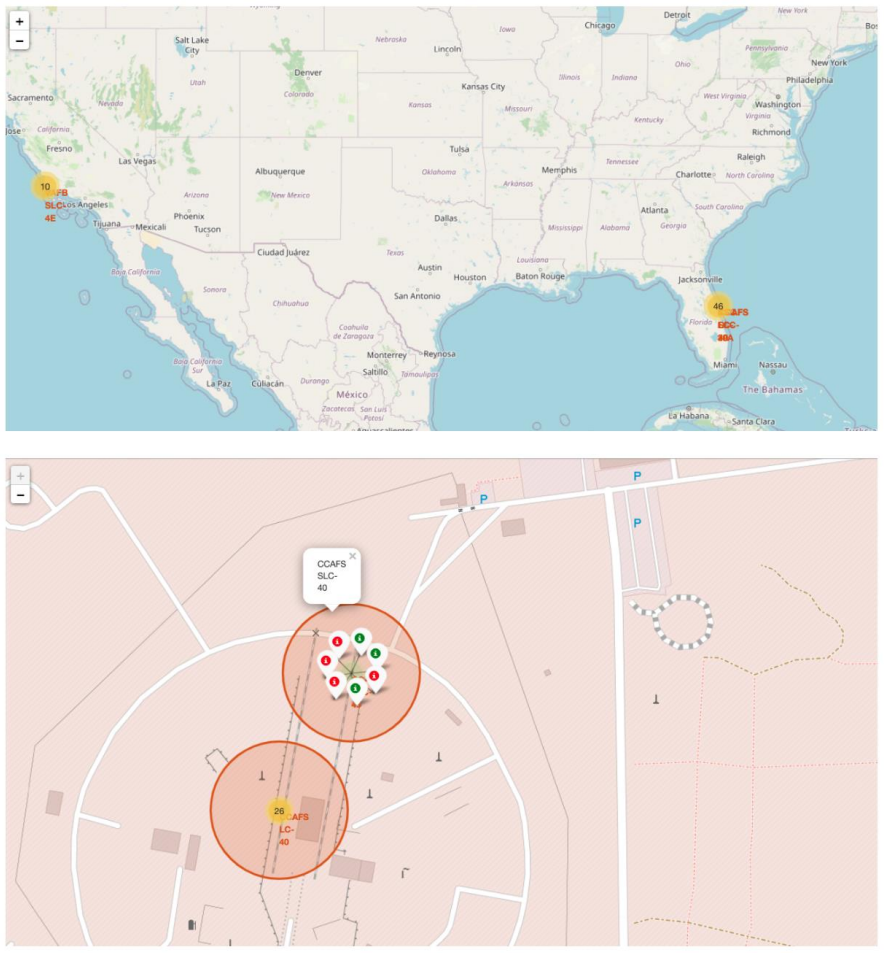
Launch Site Locations

- Launch Site of Space X Flights



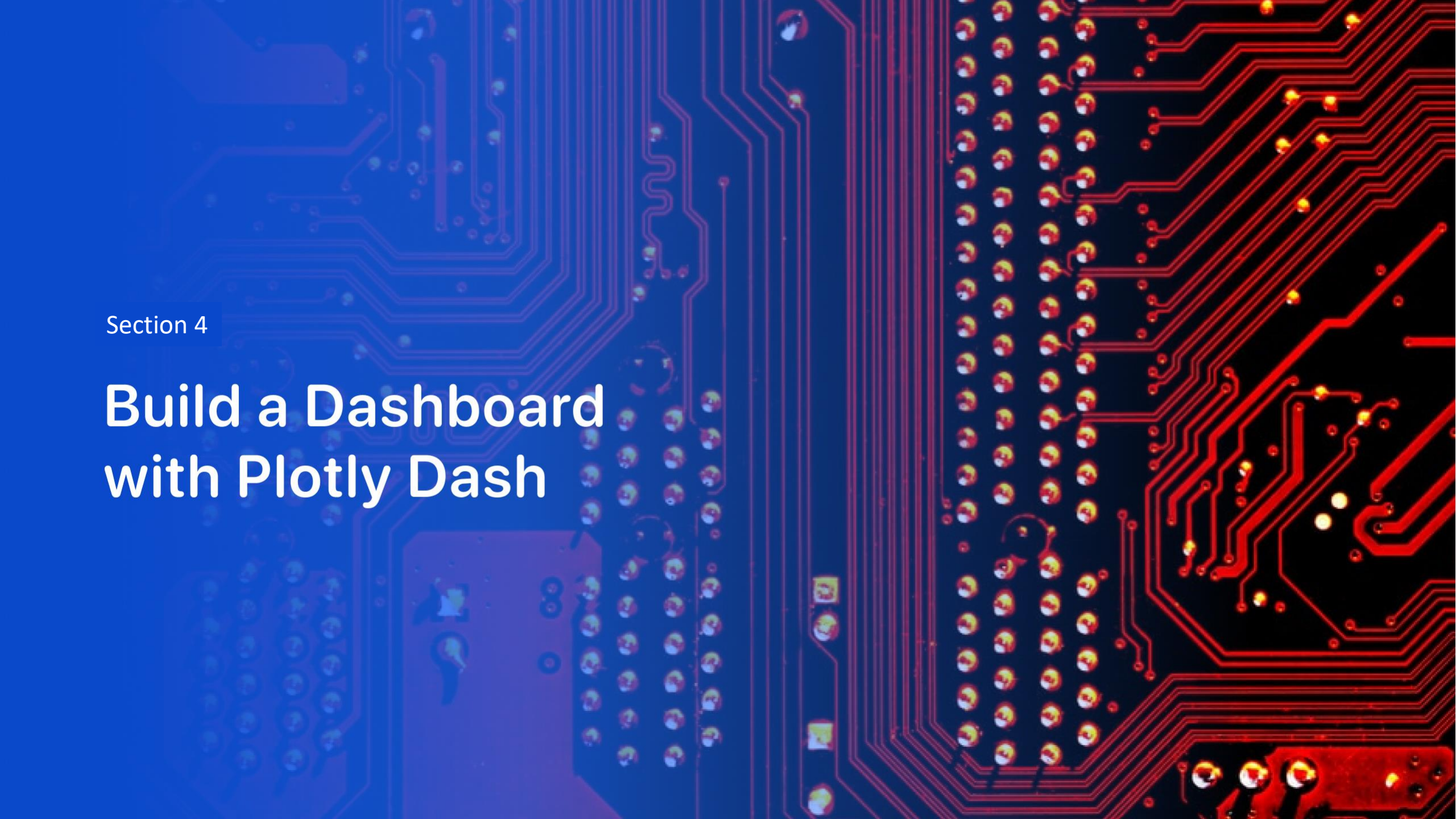
Success/failed launches for each site on the map

- In depth of Launch Site with Markers indicating successful and failed flights



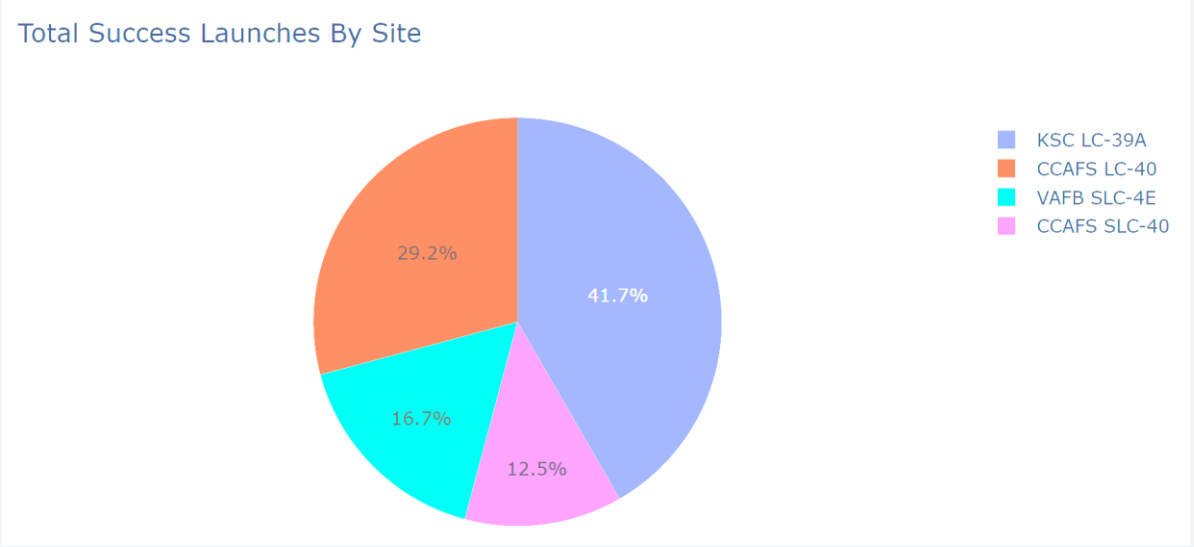
Calculate the distances between a launch site to its proximities

- Interactive map to calculate the distances between a launch site to its proximities


Section 4
Build a Dashboard with Plotly Dash

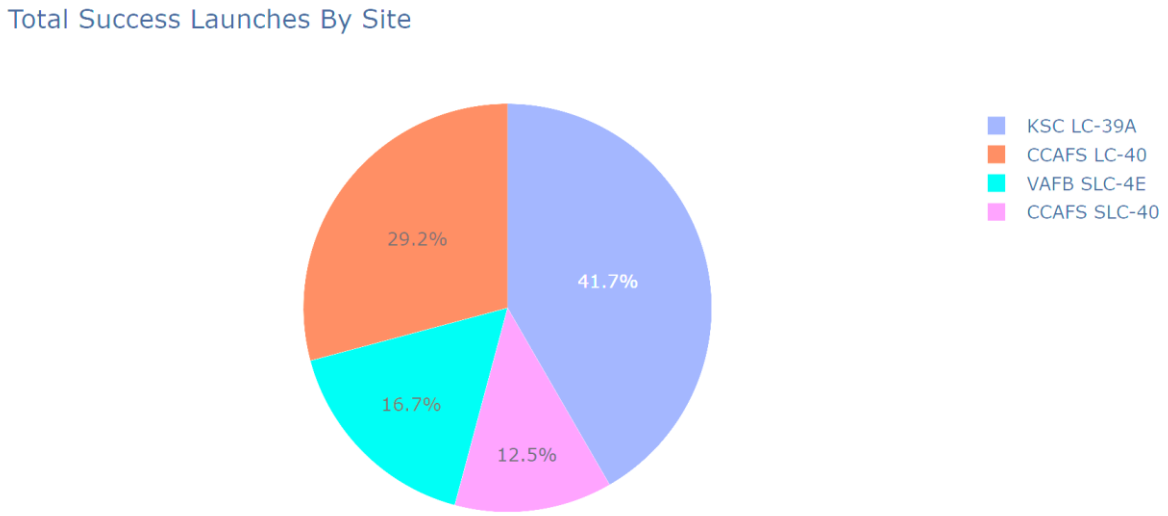
Success count for Launch sites

- KSC LC-39A has the highest count of successful launches
- CCAFS SLC-40 has the lowest



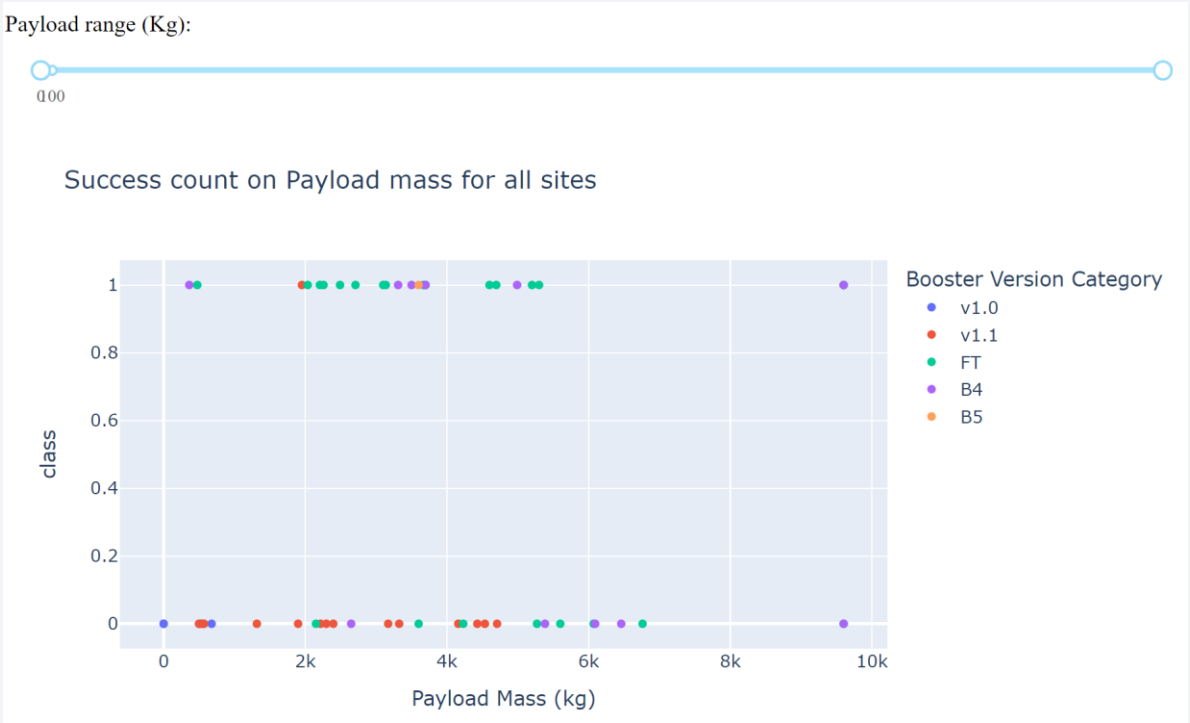
Success percentage for Launch sites

- KSC LC-39A has the highest percentage of successful launches
- CCAFS SLC-40 has the lowest



Payload mass vs Booster Version category

- Booster version category FT is the most successful for lower payloads but less successful for higher payload masses
- Booster version category v1.0 is the least successful

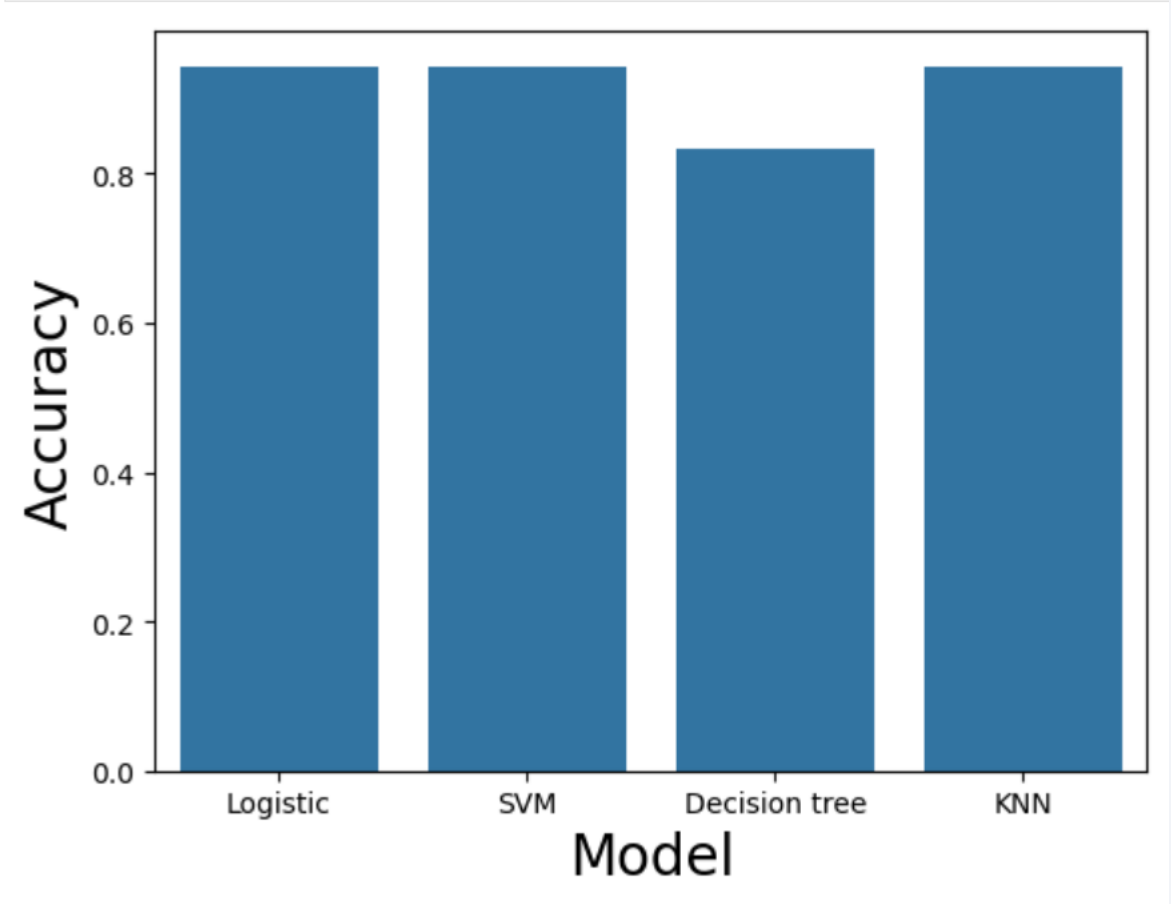


Section 5

Predictive Analysis (Classification)

Classification Accuracy

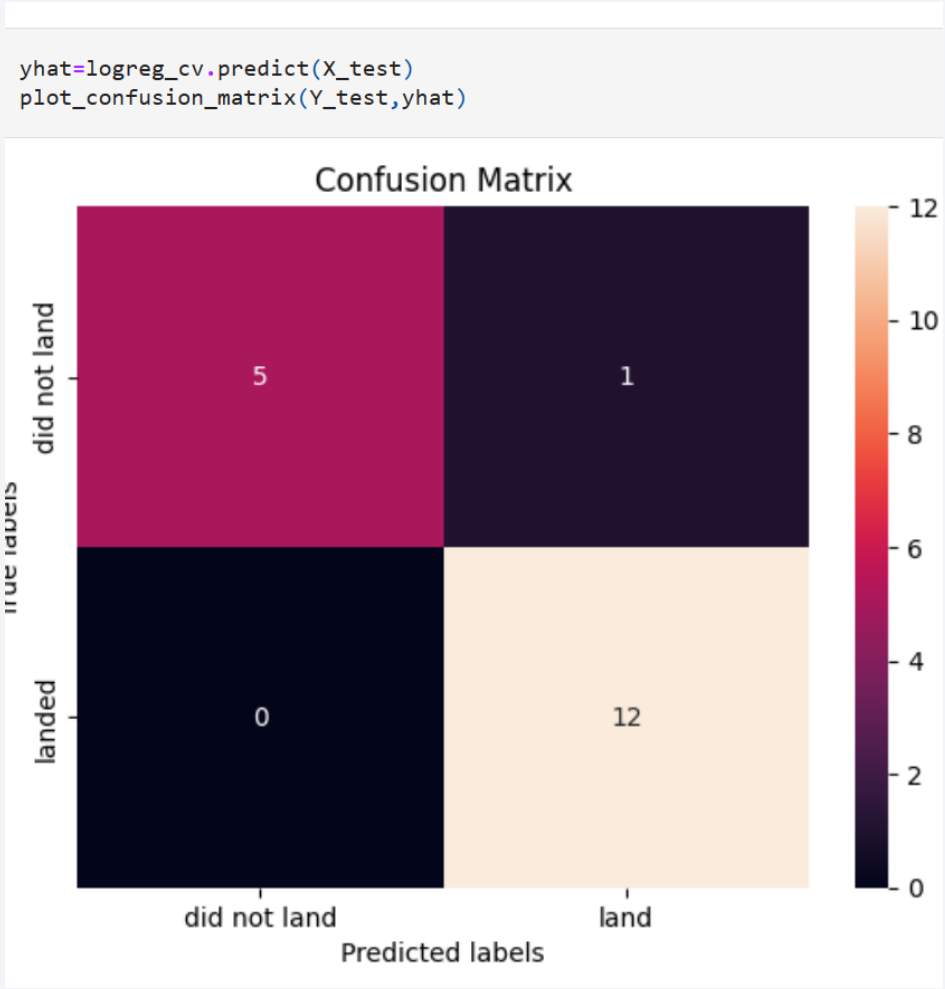
- Logistic, SVM and KNN model has the highest classification accuracy



Confusion Matrix

- Logistic the best performing model as it has an accuracy of 0.9444
- It is a simpler model

```
yhat=logreg_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```



Conclusions

- Flights gradually increased their success rate with time
- Higher payload mass seems more successful
- ES-L1, GEO, HEO and SSO orbits have highest success rates
- VLEO orbits have highest Payload mass
- Success rate of launches increases year after year
- Booster version category FT is the most successful for lower payloads but less successful for higher payload masses
- KSC LC-39A has the highest percentage of successful launches
- Logistic the best performing model as it has an accuracy of 0.9444

Thank you!