Winning Space Race with Data Science

<Gerardo Svriz>
<2024-09-30>

Outline

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

Executive Summary

Summary of methodologies

- Collect data using SpaceX REST API and web scraping techniques
- Wrangle data to create success/fail outcome variable
- Explore data with data visualization techniques, considering the following factors: payload, launch site, flight number and yearly trend
- Analyze the data with SQL, calculating the following statistics: total payload, payload range for successful launches, and total # of successful and failed outcomes
- Explore launch site success rates and proximity to geographical markers
- Visualize the launch sites with the most success and successful payload ranges
- Build Models to predict landing outcomes using logistic regression, support vector machine (SVM), decision tree and K nearest neighbor (KNN)

Executive Summary

Summary of all results

Exploratory Data Analysis

- Launch success has improved over time
- KSC LC 39A has the highest success rate among landing sites
- Orbits ES L1, GEO, HEO and SSO have a 100% success rate

Visual Analytics

- Most launch sites are near the equator, and all are close to the coast
- Launch sites are far enough away from anything a failed launch can damage

Predictive Analytics

- Decision Tree model is the best predictive model for the dataset

Introduction

Project background and context

SpaceX, a pioneer in the space industry, seeks to make space travel more affordable for everyone. Its achievements include sending spacecraft to the International Space Station, deploying a constellation of satellites that provide global internet access, and conducting manned missions. This is made possible by reusing the first stage of the Falcon 9 rocket, which significantly reduces launch costs (approximately $62 million per mission). In comparison, other providers, who do not reuse this section, incur costs in excess of $165 million. In order to predict whether SpaceX or its competitors will be able to reuse the first stage, public data and machine learning models can be used to estimate the cost of each launch.

Section 1

Methodology

Methodology

Data collection methodology (API):

- Retrieve rocket launch data from the SpaceX API.
- Decode the response using .json() and convert it to a dataframe with .json_normalize().
- Use custom functions to gather additional launch information from the SpaceX API.
- Create a dictionary from the gathered data.
- Convert the dictionary into a dataframe.
- Filter the dataframe to include only Falcon 9 launches.
- Replace missing values in the "Payload Mass" field by calculating the mean.
- Export the data to a CSV file.

Methodology

Data collection methodology (Web Scraping):

- Request Falcon 9 launch data from Wikipedia.
- Create a BeautifulSoup object from the HTML response.
- Extract column names from the HTML table header.
- Collect data by parsing the HTML tables.
- Create a dictionary from the extracted data.
- Convert the dictionary into a dataframe.
- Export the data to a CSV file.

Methodology

Perform data wrangling

- Perform EDA and determine data labels
- Calculate:
  - # of launches for each site
  - # and occurrence of orbit
  - # and occurrence of mission outcome per orbit type
- Create binary landing outcome column (dependent variable)
- Export data to csv file

Methodology

EDA using SQL

Queries:

- 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.

Methodology

EDA using SQL

Queries:

- 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.

Methodology

EDA using visualization

Charts

- Flight Number vs. Payload
- Flight Number vs. Launch Site
- Payload Mass (kg) vs. Launch Site
- Payload Mass (kg) vs. Orbit type

Methodology

Perform interactive visual analytics using Folium and Plotly Dash

Maps using Folium.

- Markers showing the locations of launch sites.
- Colored markers representing the outcomes of launches.
- Measuring distances between launch sites and nearby locations.

Methodology

Perform predictive analysis using classification models

Dropdown Menu for Launch Sites

- Enables users to select either all launch sites or a specific launch site.

Pie Chart Displaying Launch Successes

- Provides a visual breakdown of successful and unsuccessful launches as a percentage of the total.

Payload Mass Range Slider

- Allows users to choose a specific payload mass range.

Scatter Plot of Payload Mass vs. Success Rate by Booster Version

- Displays the relationship between payload mass and launch success for different booster versions.

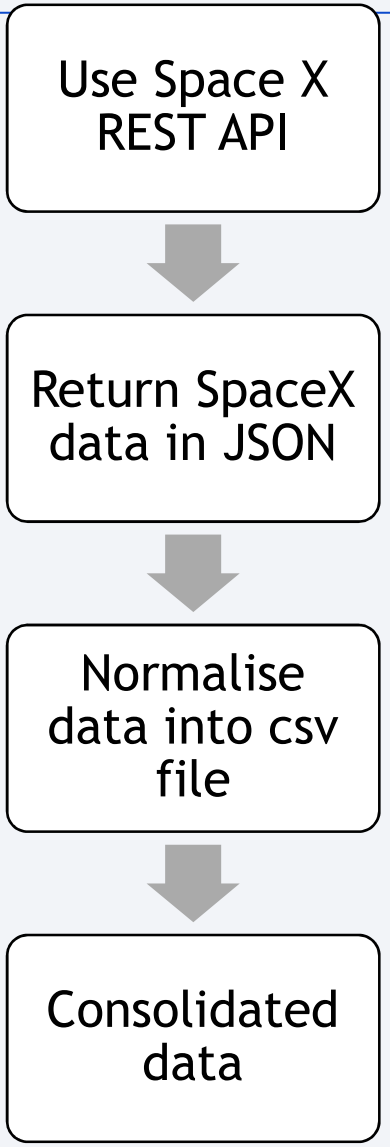
Methodology

Perform predictive analysis using classification models

- Create a NumPy array from the "Class" column.
- Standardize the dataset using StandardScaler by fitting and transforming the data.
- Split the dataset using train_test_split.
- Set up a GridSearchCV object with cv=10 for parameter tuning.
- Apply GridSearchCV on multiple algorithms: logistic regression (LogisticRegression()), support vector machine (SVC()), decision tree (DecisionTreeClassifier()), and k-nearest neighbors (KNeighborsClassifier()).
- Calculate the accuracy on the test set for all models using .score().
- Evaluate the confusion matrix for each model.
- Determine the best model based on Jaccard Score, F1 Score, and accuracy.

Data Collection – SpaceX API

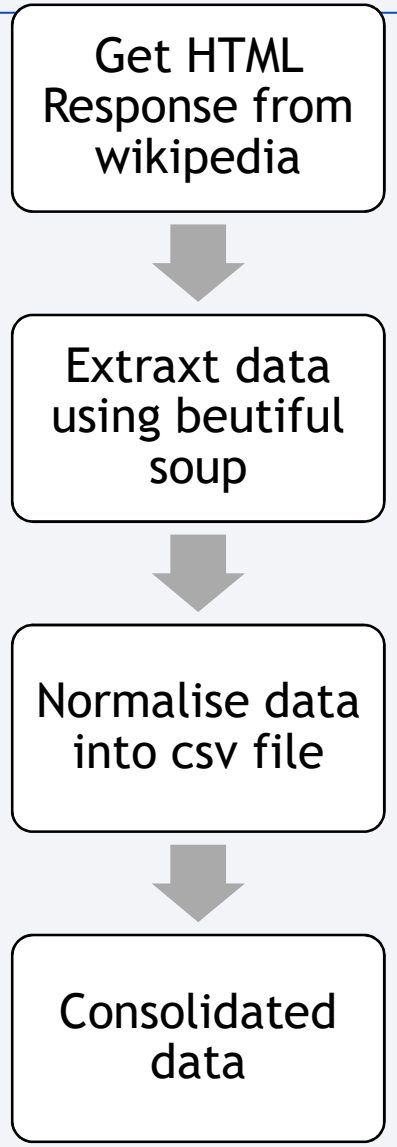
- SpaceX Launch Data: Collected from the SpaceX REST API, this dataset includes information on the rocket used, payload delivered, launch details, landing specifications, and outcomes.
- API Endpoints: The URLs for the SpaceX REST API begin with `api.spacexdata.com/v4/`.
- Jupyter Notebook on GitHub



Data Collection - Scraping

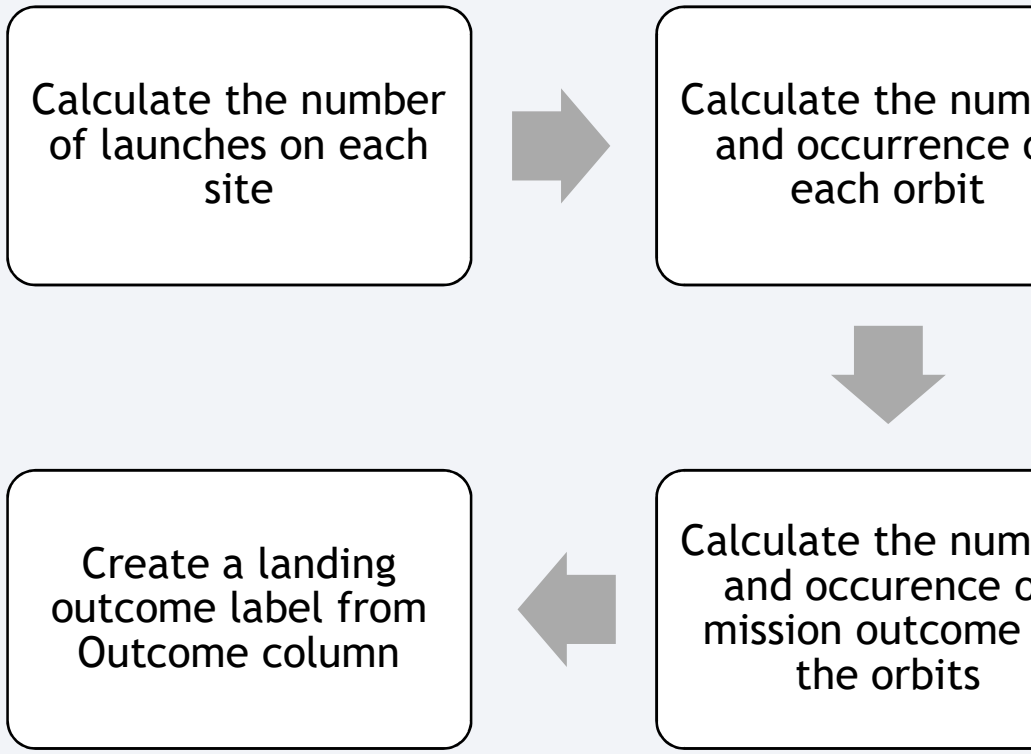
- Alternative Data Source: Falcon 9 launch data can also be obtained by web scraping Wikipedia with BeautifulSoup.

- Jupyter Notebook on GitHub

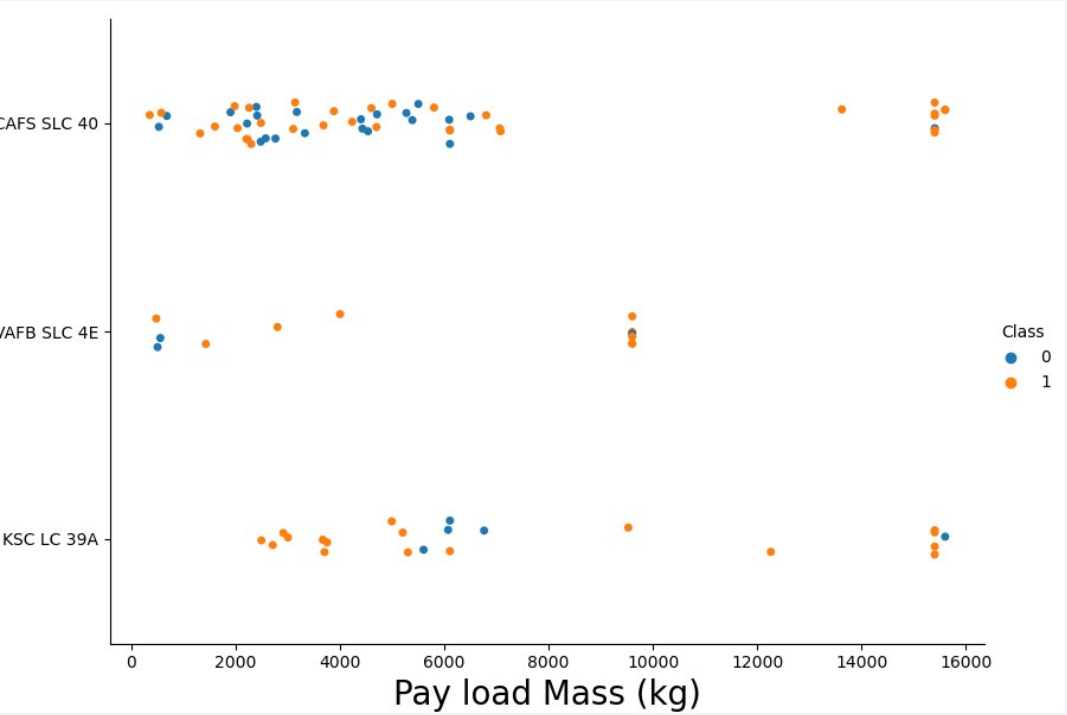


Data Wrangling

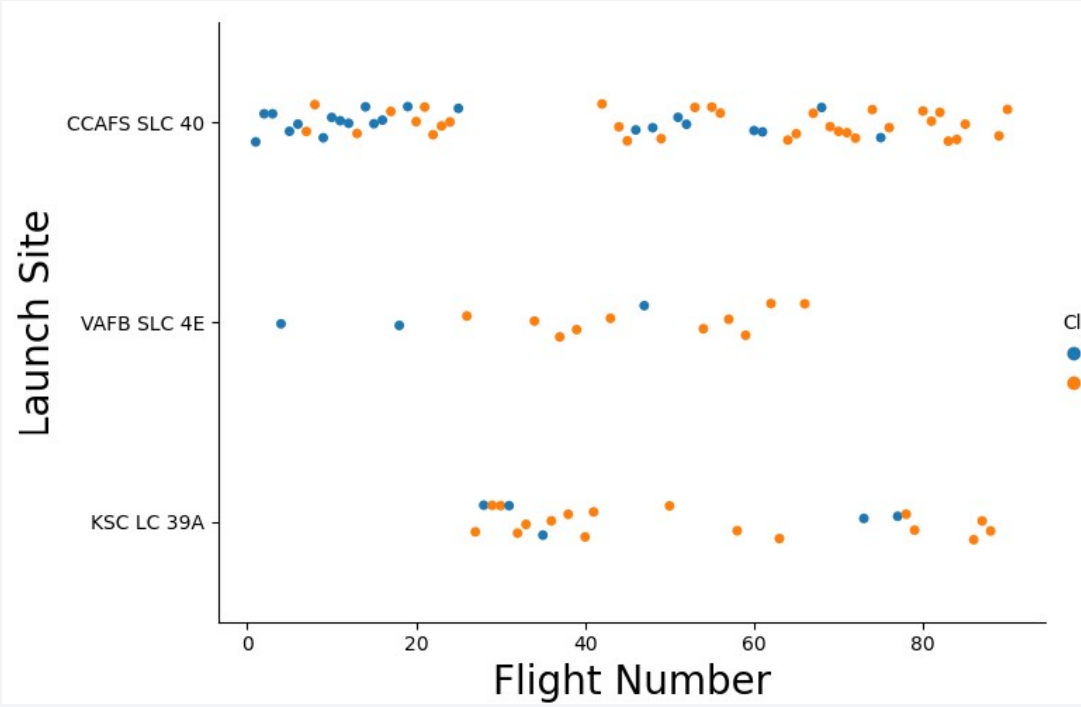
- I conducted exploratory data analysis to identify the training labels.
- I calculated the total number of launches at each site and analyzed the frequency of each orbit.
- I created landing outcome labels from the outcome column and exported the results to a CSV file.
- Jupyter Notebook on GitHub



EDA with Data Visualization



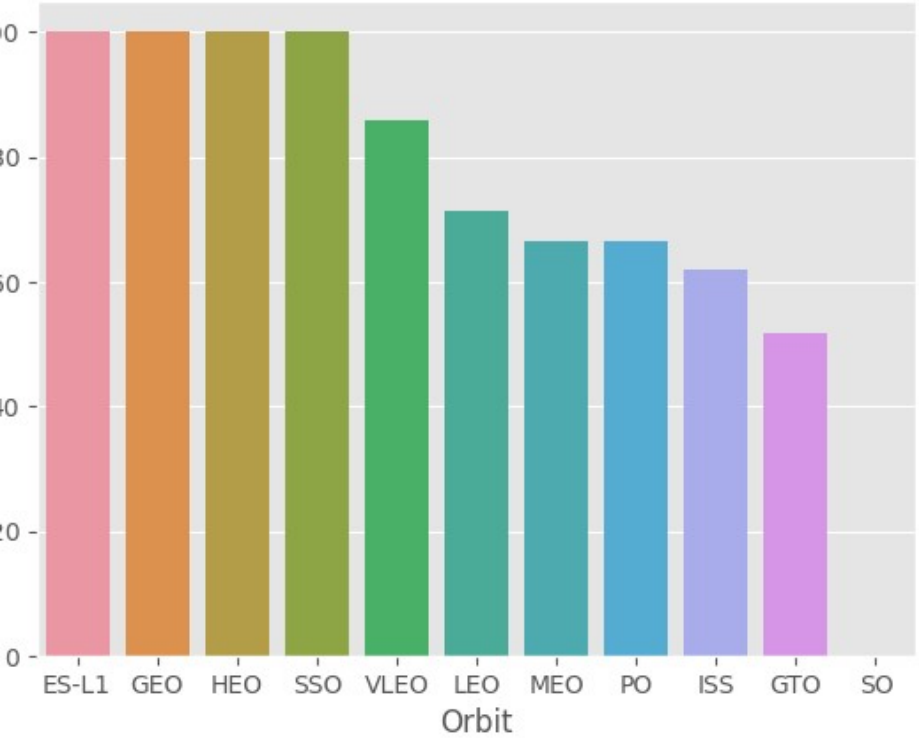
It is noted that no rockets with a heavy payload mass (over 10,000 g) were launched for the VAFB-SLC launch site.



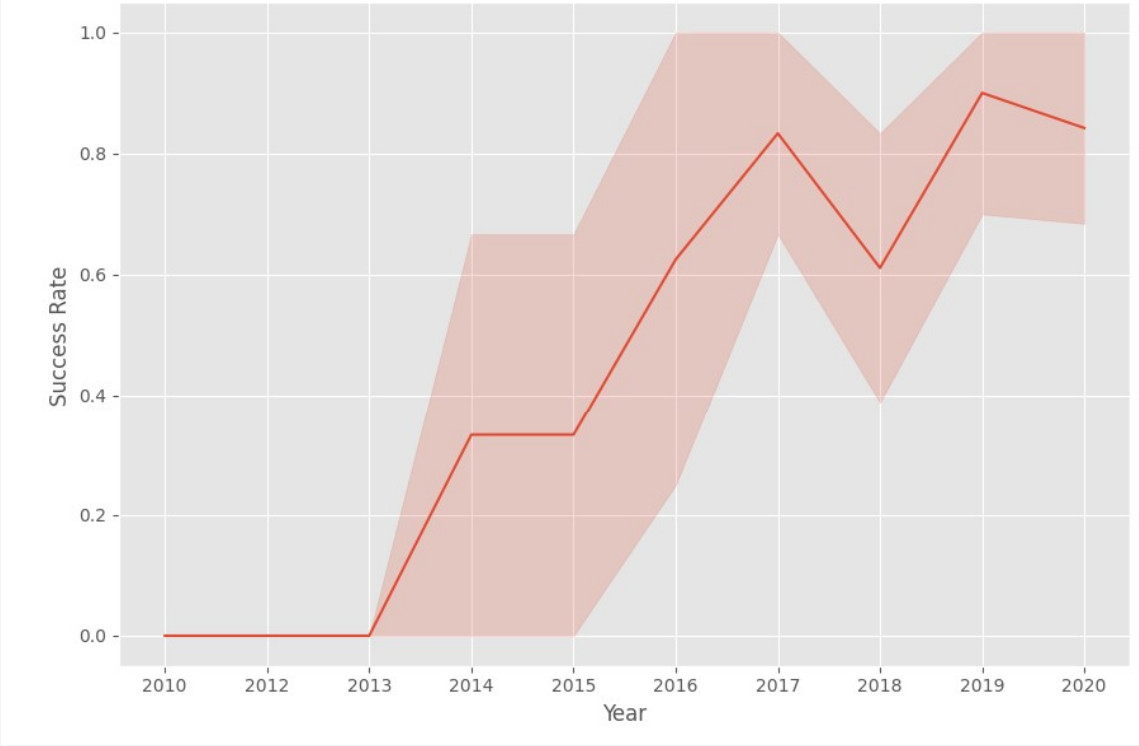
We can deduce that, as the flight number increases in each of the 3 launch sites, so does the success rate.

- [Jupyter Notebook on GitHub]

EDA with Data Visualization



ES-L1, GEO, HEO and SSO orbits have the highest success rates (100%)



The success rate since 2013 continued to increase until 2020.

- [Jupyter Notebook on GitHub]

EDA with SQL

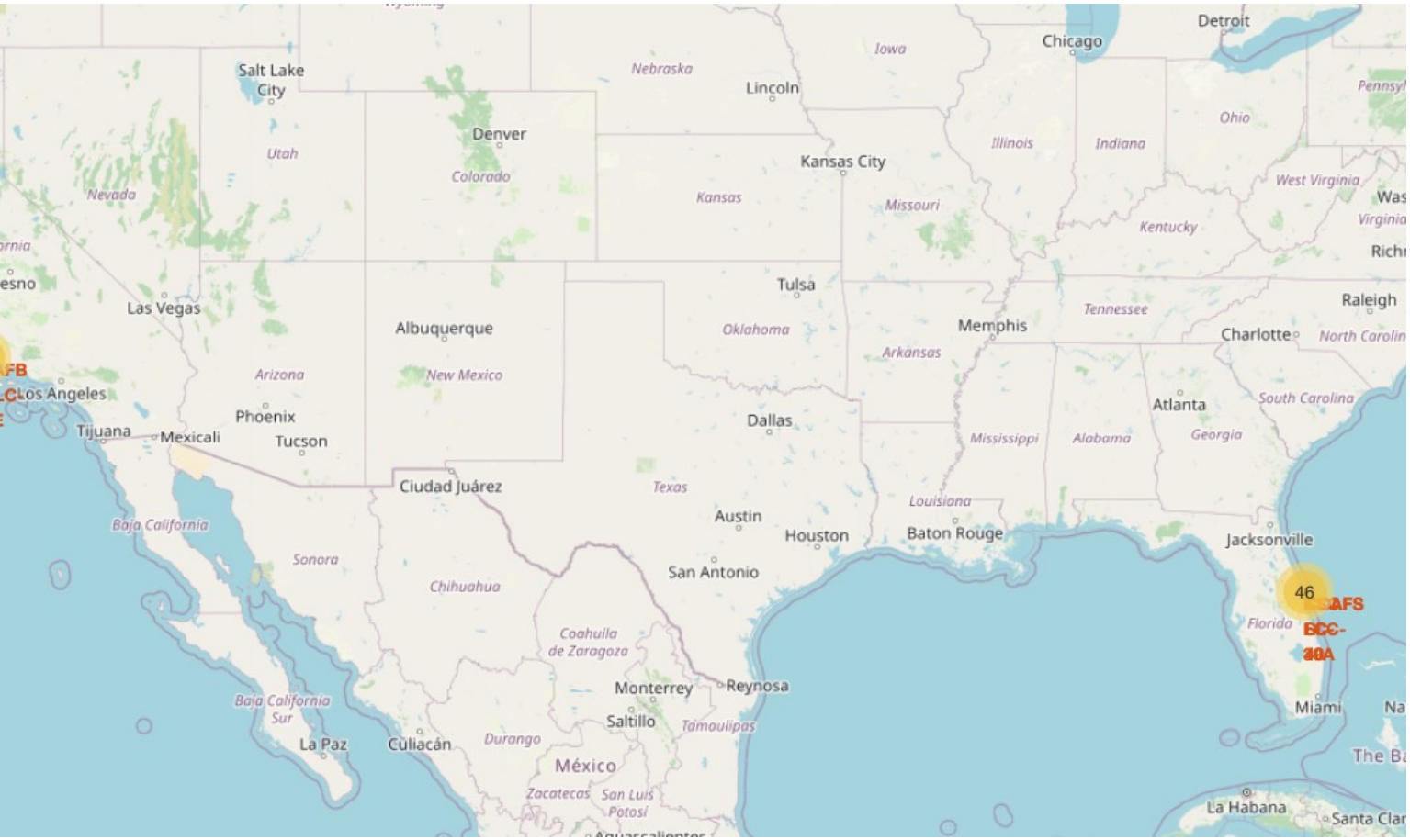
- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.
- List the names of the boosters 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.
- Jupyter Notebook on GitHub

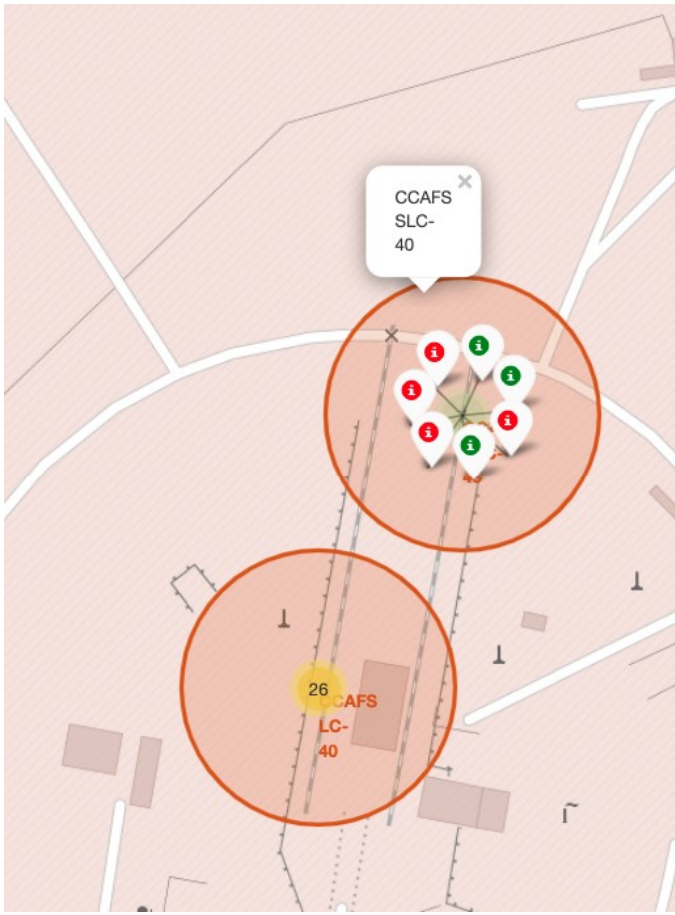
Build an Interactive Map with Folium

- folium.Marker(): It was used to create a marker on a map at specified coordinates to denote a location.
- folium.Circle(): It was used to draw a circle on the map to represent an area with a specified radius around a given point.
- folium.Icon(): It was used to customize the appearance of a marker by changing its color, icon, and other attributes.
- folium.PolyLine(): It was used to draw a line connecting a series of geographic points on the map.
- folium.plugins.AntPath(): It was used to create animated paths on the map to illustrate movement or routes.
- markerCluster(): It was used to group nearby markers into clusters to improve map readability and organization.

- Jupyter Notebook on GitHub

Build an Interactive Map with Folium





- [Jupyter Notebook on GitHub]

Predictive Analysis (Classification)

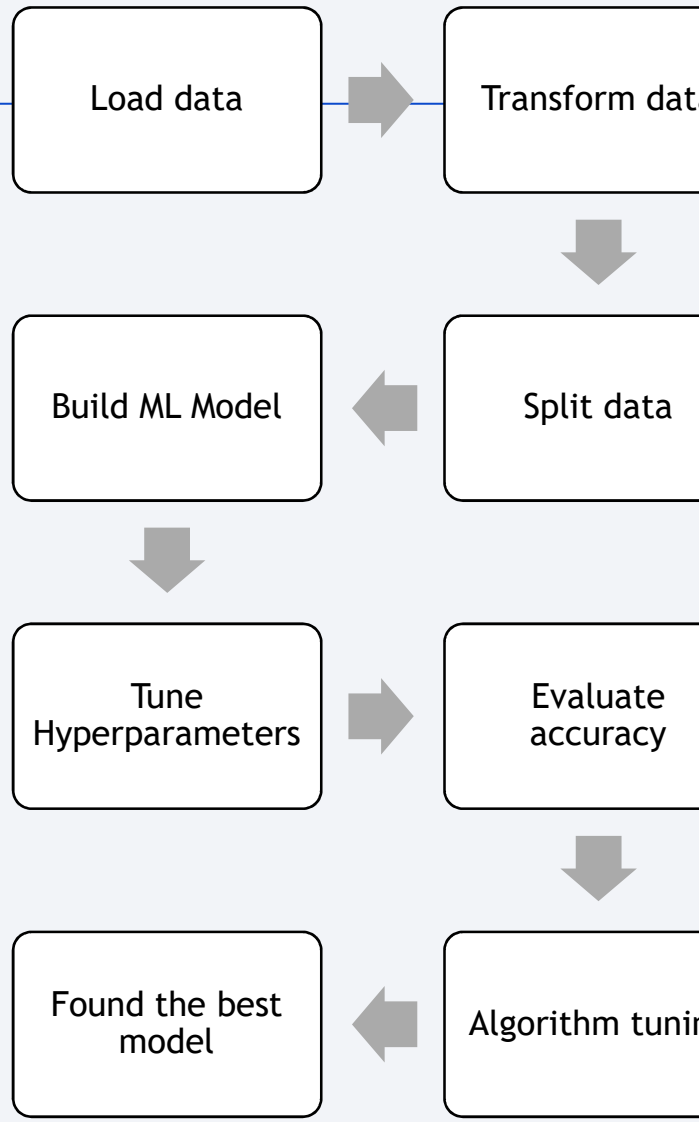
imported the data using NumPy and Pandas, ransformed it, and divided it into training and testing sets.

developed various machine learning models and optimized hyperparameters with GridSearchCV.

utilized accuracy as the evaluation metric, enhanced he model through feature engineering and algorithm uning, and ultimately identified the best-performing classification model.

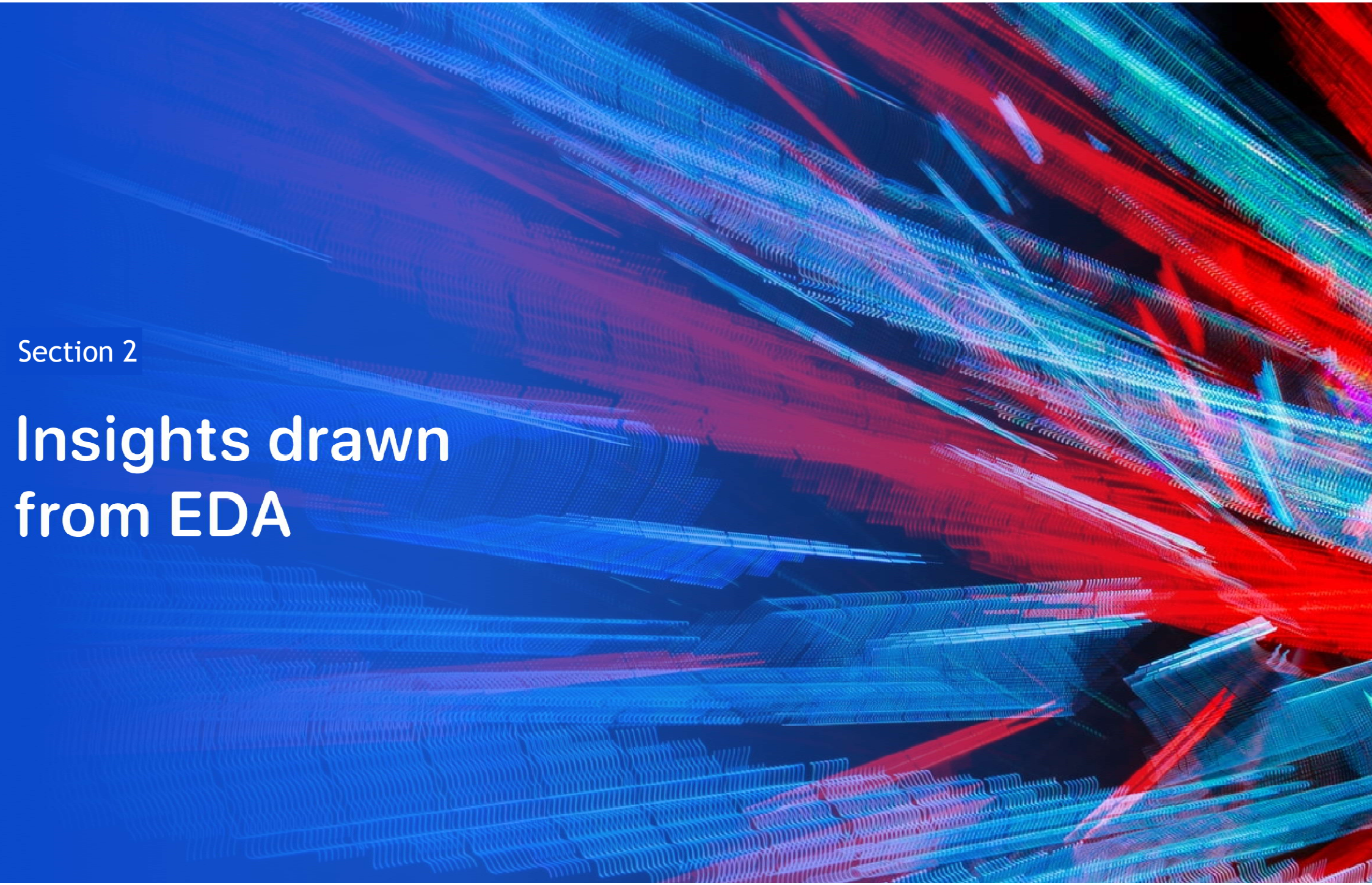
You need present your model development process using key phrases and flowchart

- Jupyter Notebook on GitHub



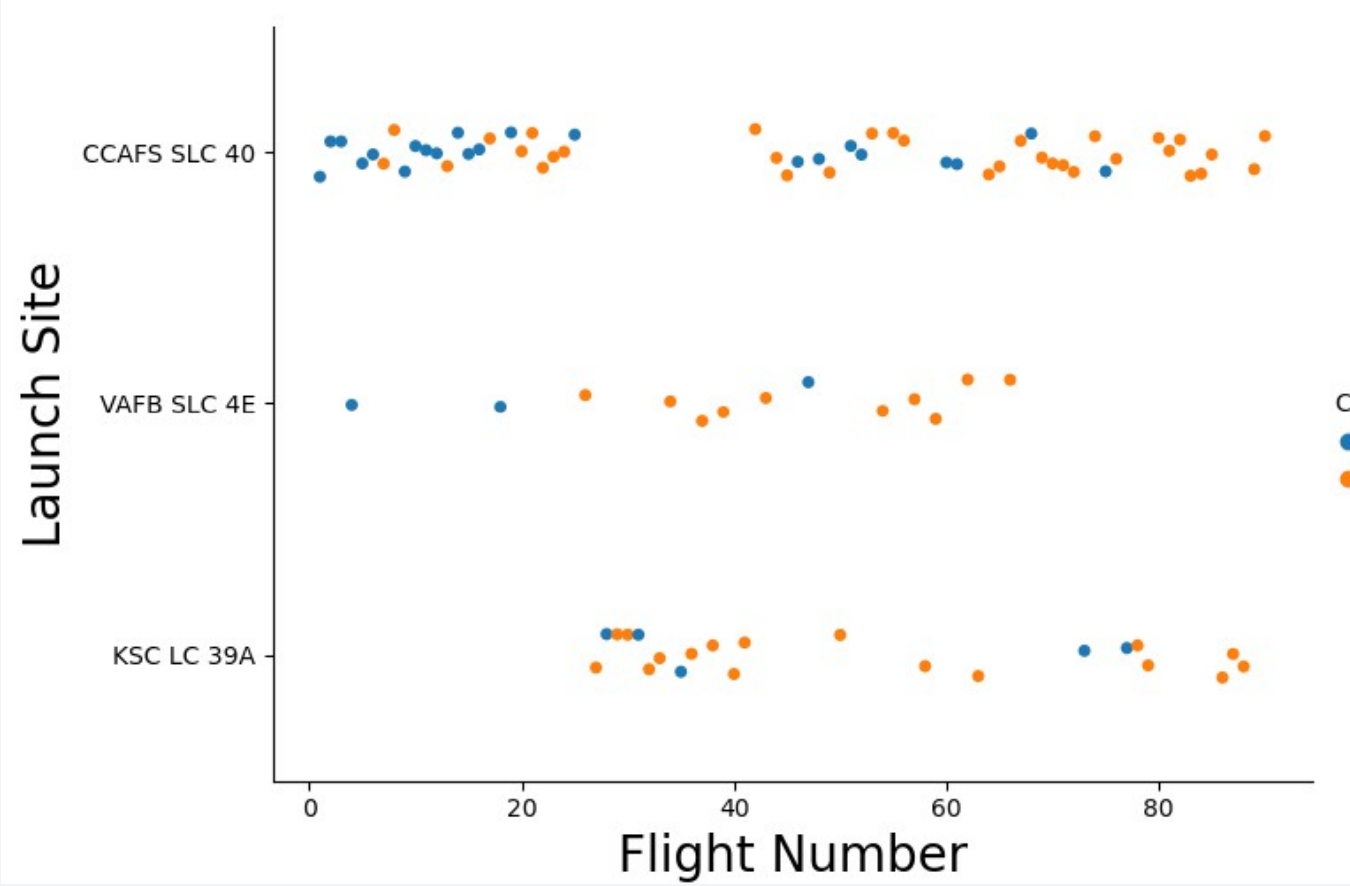
Results

- Launch success has improved over time
- KSC LC 39A has the highest success rate among landing sites
- Orbits ES L1, GEO, HEO and SSO have a 100% success rate
- Launch sites are far enough away from anything a failed launch can damage
- Decision Tree model is the best predictive model for the dataset


Section 2
Insights drawn
from EDA

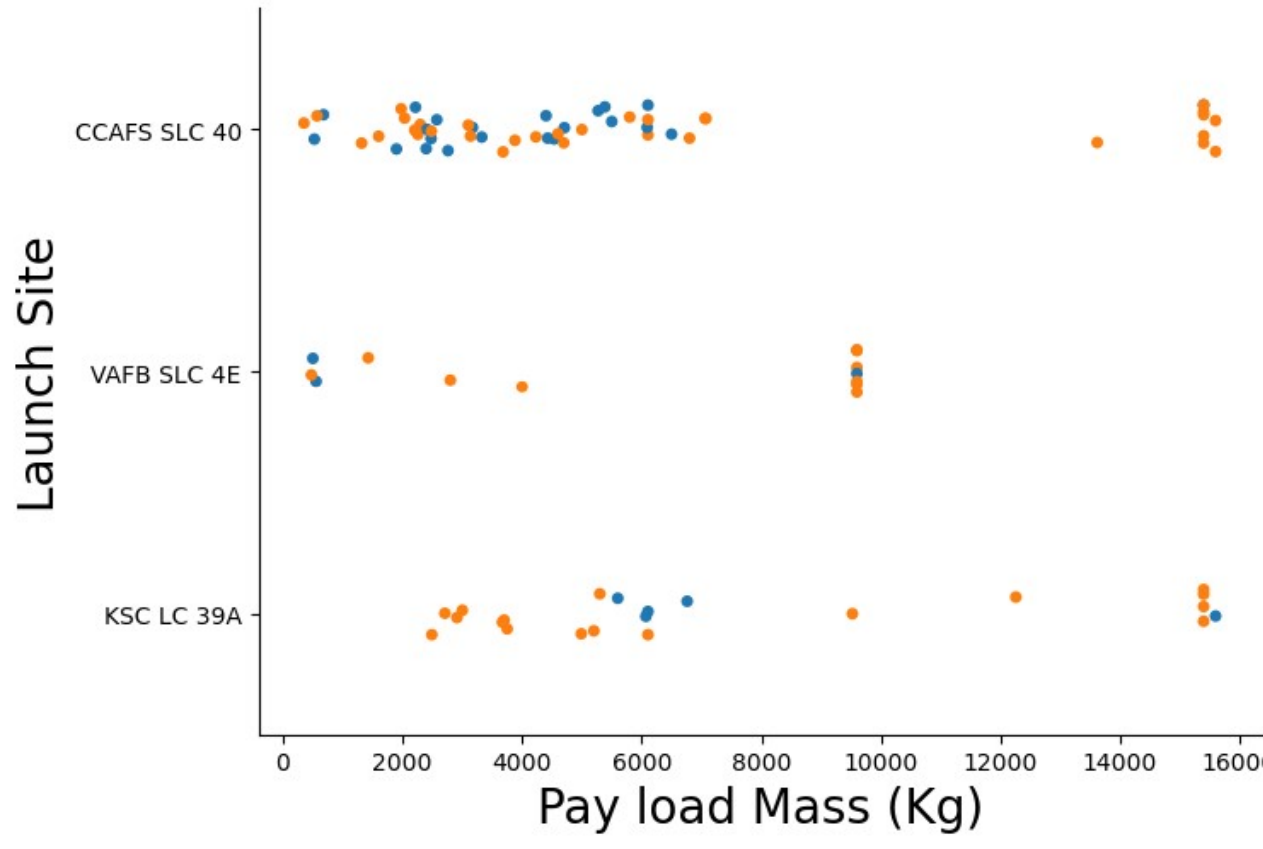
Flight Number vs. Launch Site

- The number of launches from CCAFS SLC 40 is significantly higher than the other locations.



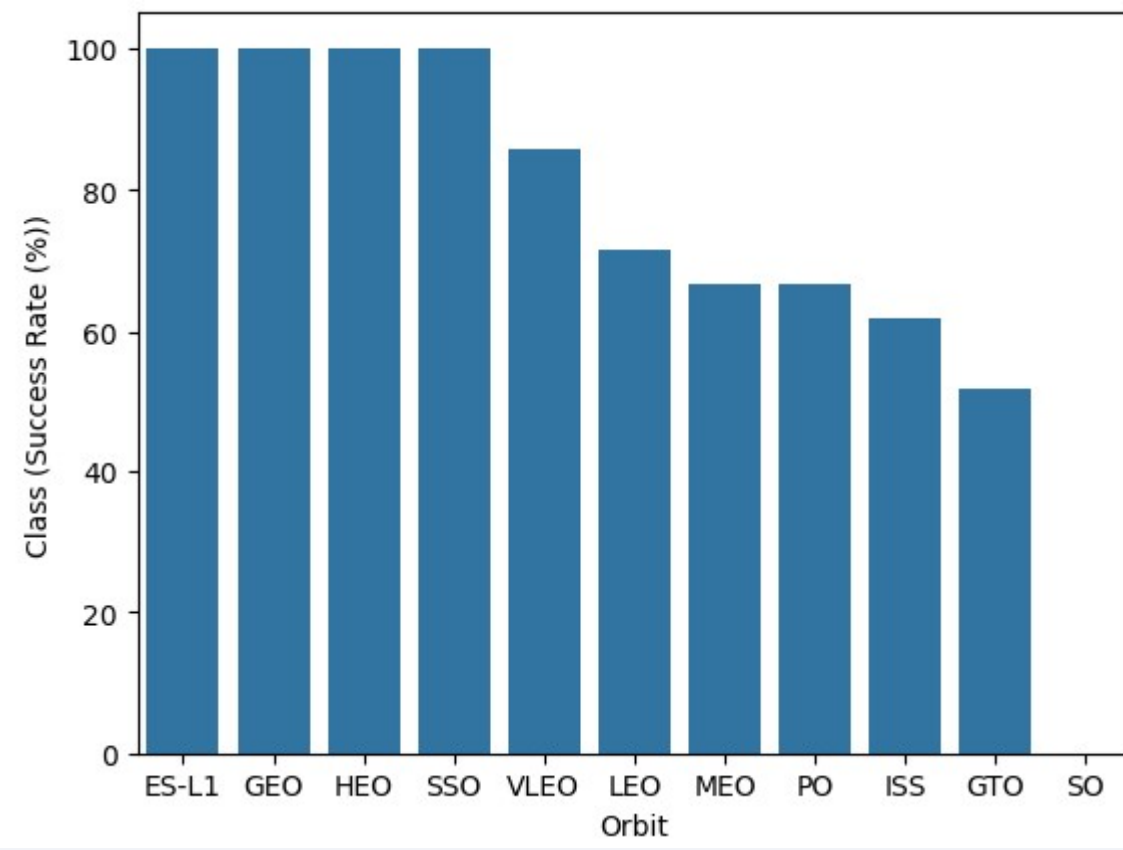
Payload vs. Launch Site

- No rockets with a heavy payload mass (greater than 10,000 kg) were launched for the VAFB-SLC launch site.



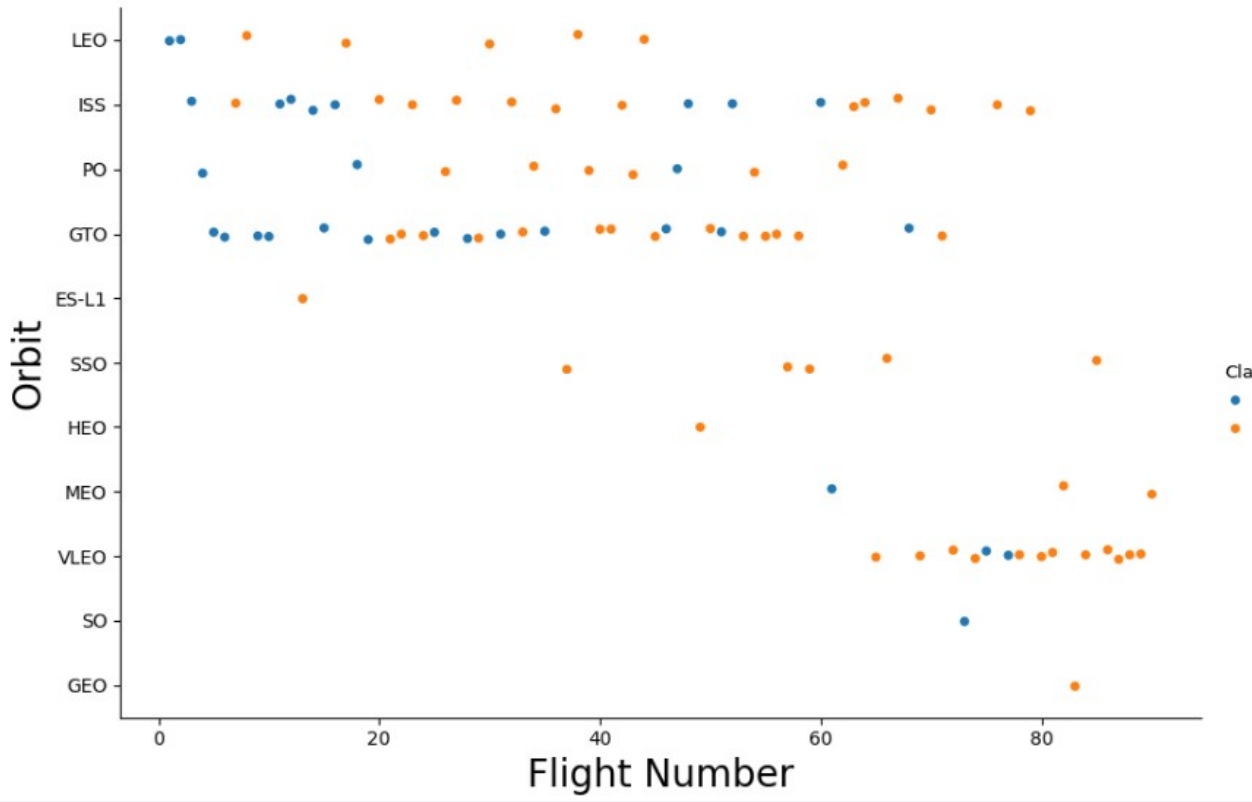
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO and SSO are orbits with the highest effectiveness success rates.



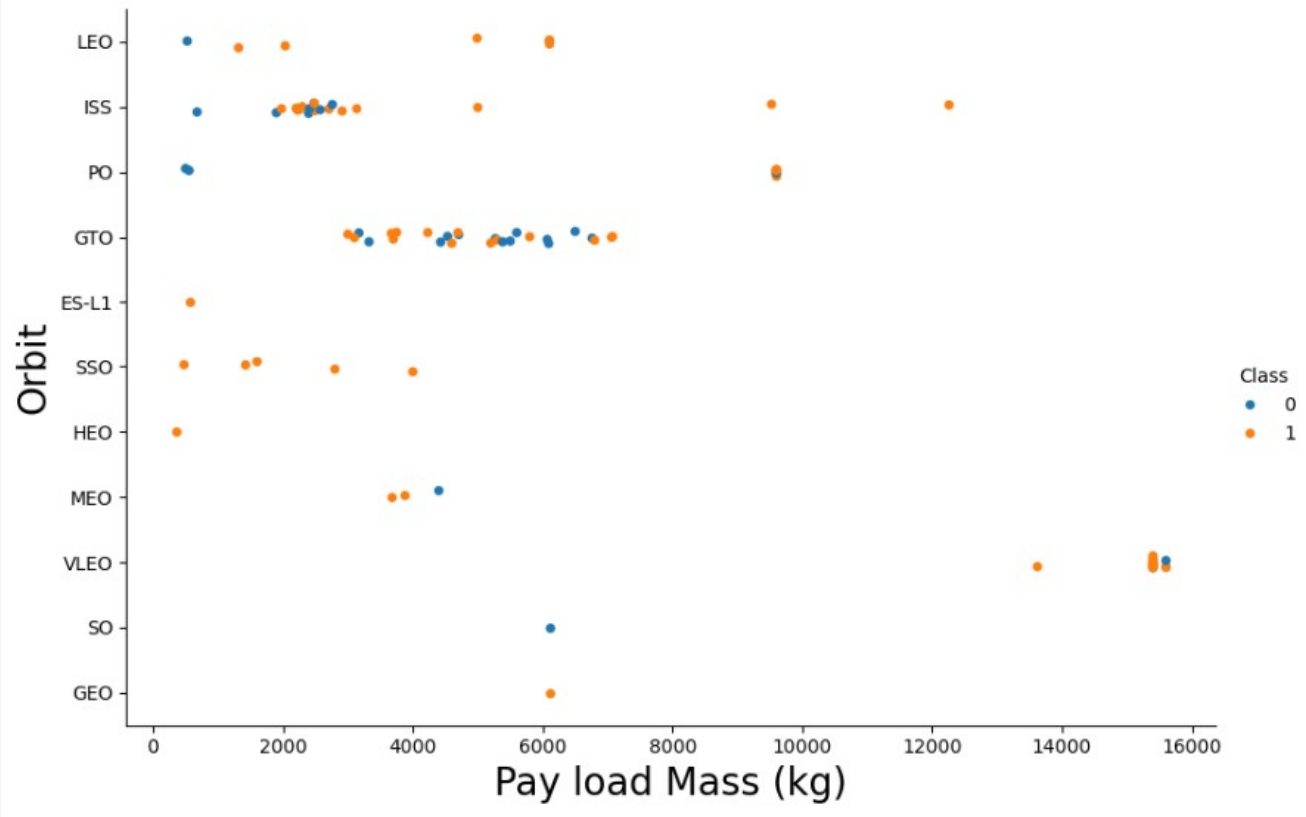
Flight Number vs. Orbit Type

- It seems that the number of flights has a relationship with the success rate of the flights.



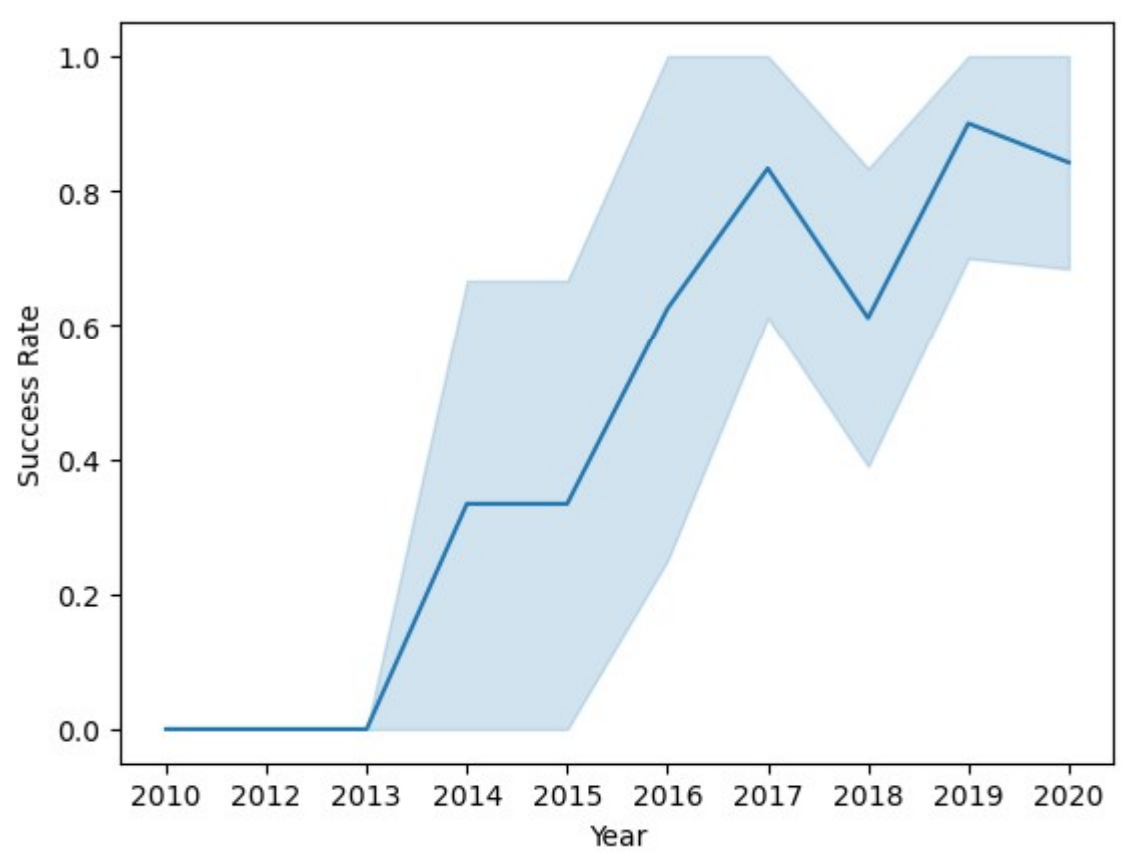
Payload vs. Orbit Type

- It can be observed that heavier payloads tend to have a higher success rate for landings in PO, LEO, and ISS orbits.



Launch Success Yearly Trend

- You can observe that the success rate has consistently risen from 2013 to 2020



All Launch Site Names

- **%sql** SELECT DISTINCT LAUNCH_SITE as "Launch_Sites" FROM SPACEXTBL;

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

Launch Site Names Begin with 'CCA'

- **%sql** select * from SPACEXTBL where launch_site like 'CCA%' limit 5;

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 Payload Mass(Kgs)", Customer FROM 'SPACEXTBL' WHERE Customer = 'NASA (CRS)';

Total Payload Mass(Kgs)	Customer
45596	NASA (CRS)

Average Payload Mass by F9 v1.1

- **%sql** SELECT AVG(PAYLOAD_MASS__KG_) as "Payload Mass Kgs", Customer, Booster_Version FROM 'SPACEXTBL' WHERE Booster_Version LIKE 'F9 v1.1%';

Payload Mass Kgs	Customer	Booster_Version
2534.6666666666665	MDA	F9 v1.1 B1003

First Successful Ground Landing Date

- **%sql** SELECT MIN(DATE) FROM 'SPACEXTBL' WHERE Landing_Outcome = 'Success (ground pad)'

MIN(DATE)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- **%sql** SELECT DISTINCT Booster_Version, Payload FROM SPACEXTBL WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;

Booster_Version	Payload
F9 FT B1022	JCSAT-14
F9 FT B1026	JCSAT-16
F9 FT B1021.2	SES-10
F9 FT B1031.2	SES-11 / EchoStar 105

Total Number of Successful and Failure Mission Outcomes

- **%sql** SELECT Mission_Outcome, COUNT(Mission_Outcome) as Total FROM SPACEXTBL GROUP BY Mission_Outcome;

Mission_Outcome	Total
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

```
• %sql SELECT Booster_Version,Payload, PAYLOAD_MASS__KG_
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (
  SELECT MAX(PAYLOAD_MASS__KG_)
  FROM SPACEXTBL
);
```

Booster_Version	Payload	PAYLOAD_MASS
F9 B5 B1048.4	Starlink 1 v1.0, SpaceX CRS-19	
F9 B5 B1049.4	Starlink 2 v1.0, Crew Dragon in-flight abort test	
F9 B5 B1051.3	Starlink 3 v1.0, Starlink 4 v1.0	
F9 B5 B1056.4	Starlink 4 v1.0, SpaceX CRS-20	
F9 B5 B1048.5	Starlink 5 v1.0, Starlink 6 v1.0	
F9 B5 B1051.4	Starlink 6 v1.0, Crew Dragon Demo-2	
F9 B5 B1049.5	Starlink 7 v1.0, Starlink 8 v1.0	
F9 B5 B1060.2	Starlink 11 v1.0, Starlink 12 v1.0	
F9 B5 B1058.3	Starlink 12 v1.0, Starlink 13 v1.0	
F9 B5 B1051.6	Starlink 13 v1.0, Starlink 14 v1.0	
F9 B5 B1060.3	Starlink 14 v1.0, GPS III-04	
F9 B5 B1049.7	Starlink 15 v1.0, SpaceX CRS-21	

2015 Launch Records

- **%sql** SELECT substr(Date,0,5), substr(Date,6, 2),Booster_Version, Launch_Site, Payload, PAYLOAD_MASS__KG_, Mission_Outcome, Landing_Outcome FROM SPACEXTBL WHERE substr(Date,0,5)='2015' AND Landing_Outcome = 'Failure (drone ship)';

substr(Date,0,5)	substr(Date,6, 2)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Mission_Outcome	Landing_Outcome
2015	01	F9 v1.1 B1012	CCAFS LC-40	SpaceX CRS-5	2395	Success	Failure (drone ship)
2015	04	F9 v1.1 B1015	CCAFS LC-40	SpaceX CRS-6	1898	Success	Failure (drone ship)

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

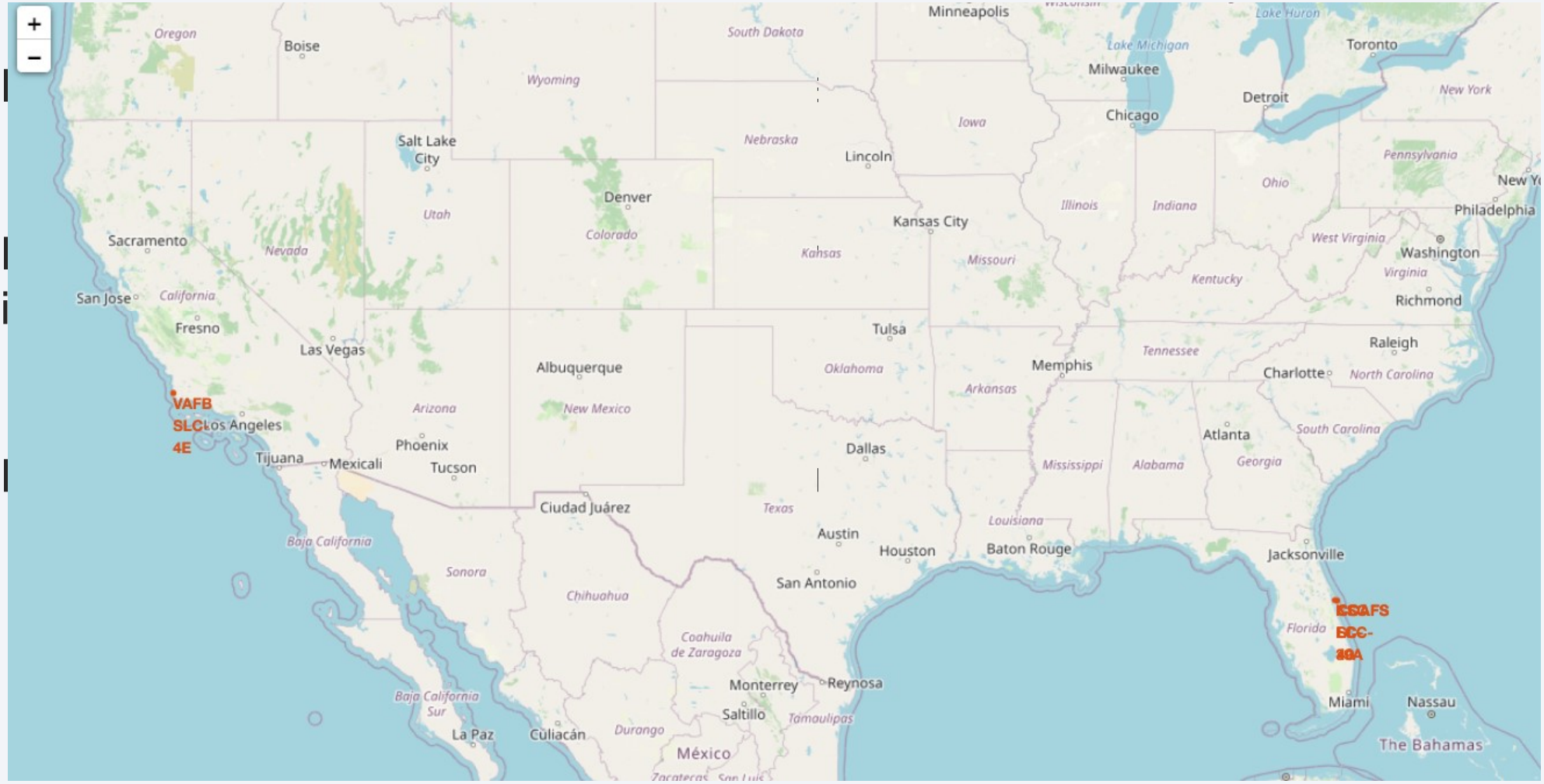
```
ql SELECT *
OM SPACEXTBL
HERE Landing_Outcome LIKE 'Success%'
ID (Date BETWEEN '2010-06-0
ID '2017-03-20')
RDER BY Date DESC;
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	La
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	
2017-01-14	17:54:00	F9 FT B1029.1	VAFB SLC-4E	Iridium NEXT 1	9600	Polar LEO	Iridium Communications	Success	
2016-08-14	5:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	
2016-07-18	4:45:00	F9 FT B1025.1	CCAFS LC-40	SpaceX CRS-9	2257	LEO (ISS)	NASA (CRS)	Success	
2016-05-27	21:39:00	F9 FT B1023.1	CCAFS LC-40	Thaicom 8	3100	GTO	Thaicom	Success	
2016-05-06	5:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	
2016-04-08	20:43:00	F9 FT B1021.1	CCAFS LC-40	SpaceX CRS-8	3136	LEO (ISS)	NASA (CRS)	Success	
2015-12-22	1:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	

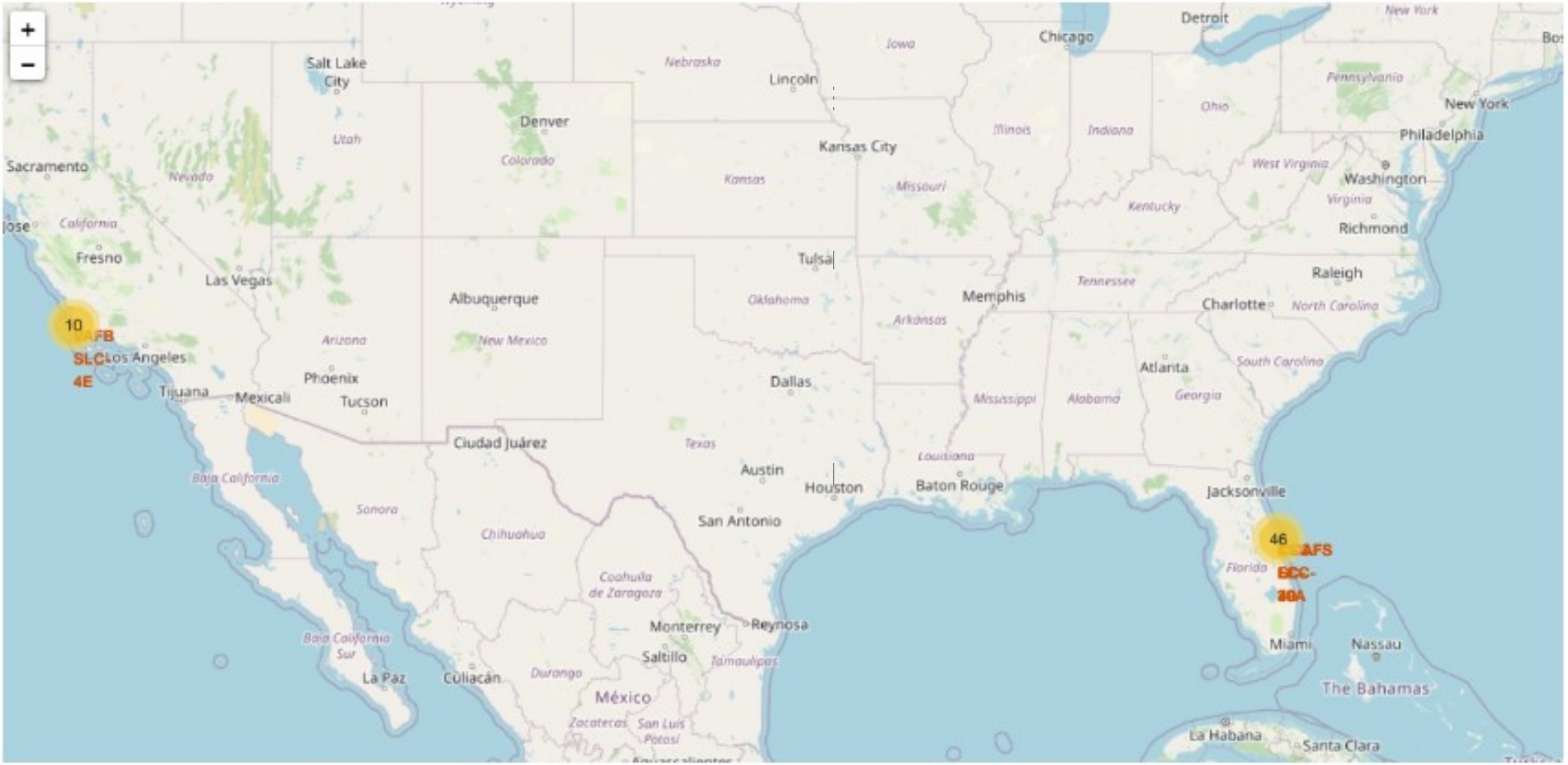
Section 3

Launch Sites Proximities Analysis

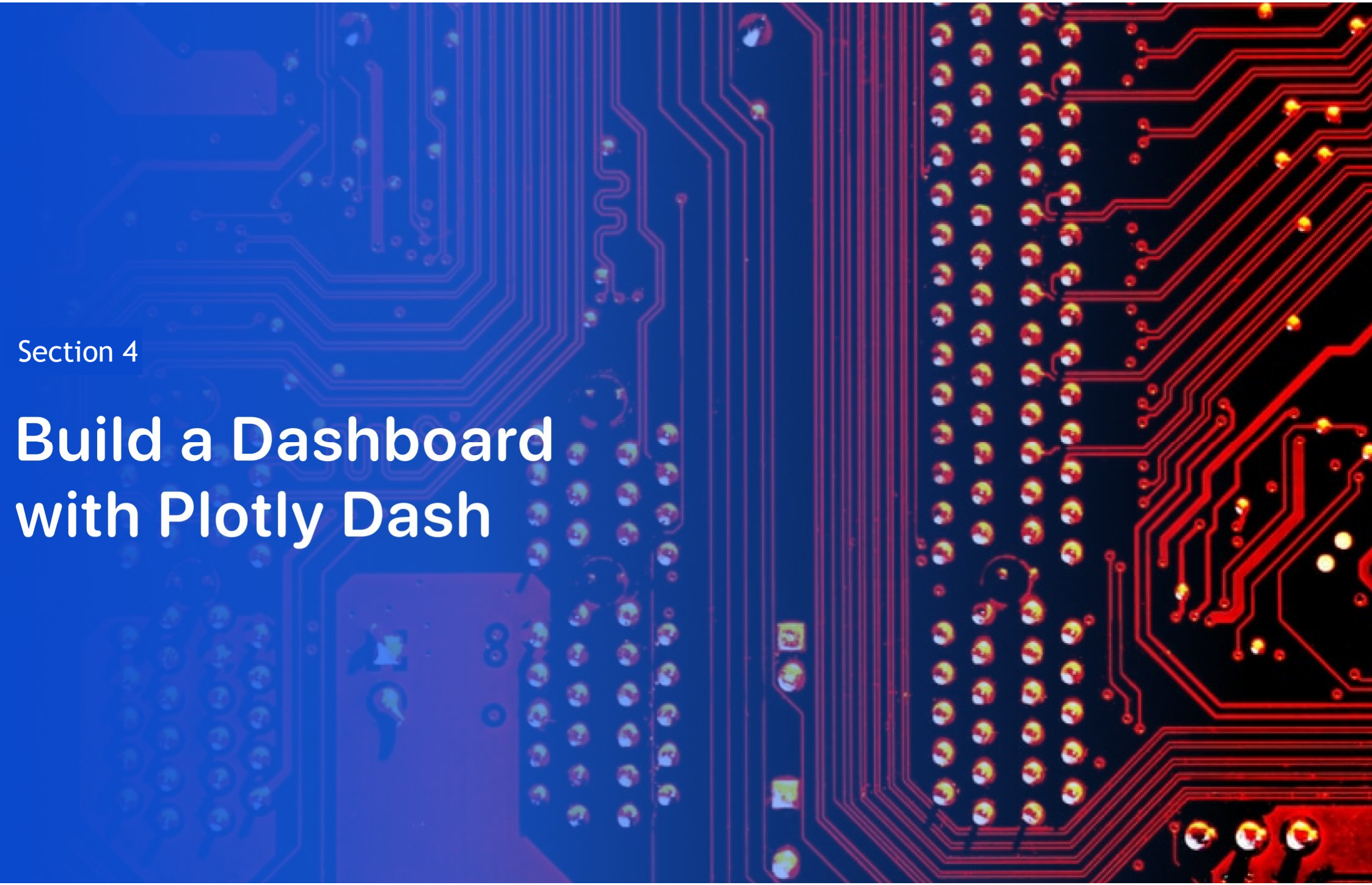
<Folium Map Screenshot 1>



<Folium Map Screenshot 2>



<Folium Map Screenshot 3>


Section 4
Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>

- Replace <Dashboard screenshot 1> title with an appropriate title
- Show the screenshot of launch success count for all sites, in a piechart
- Explain the important elements and findings on the screenshot

<Dashboard Screenshot 2>

- Replace <Dashboard screenshot 2> title with an appropriate title
- Show the screenshot of the piechart for the launch site with highest launch success ratio
- Explain the important elements and findings on the screenshot

<Dashboard Screenshot 3>

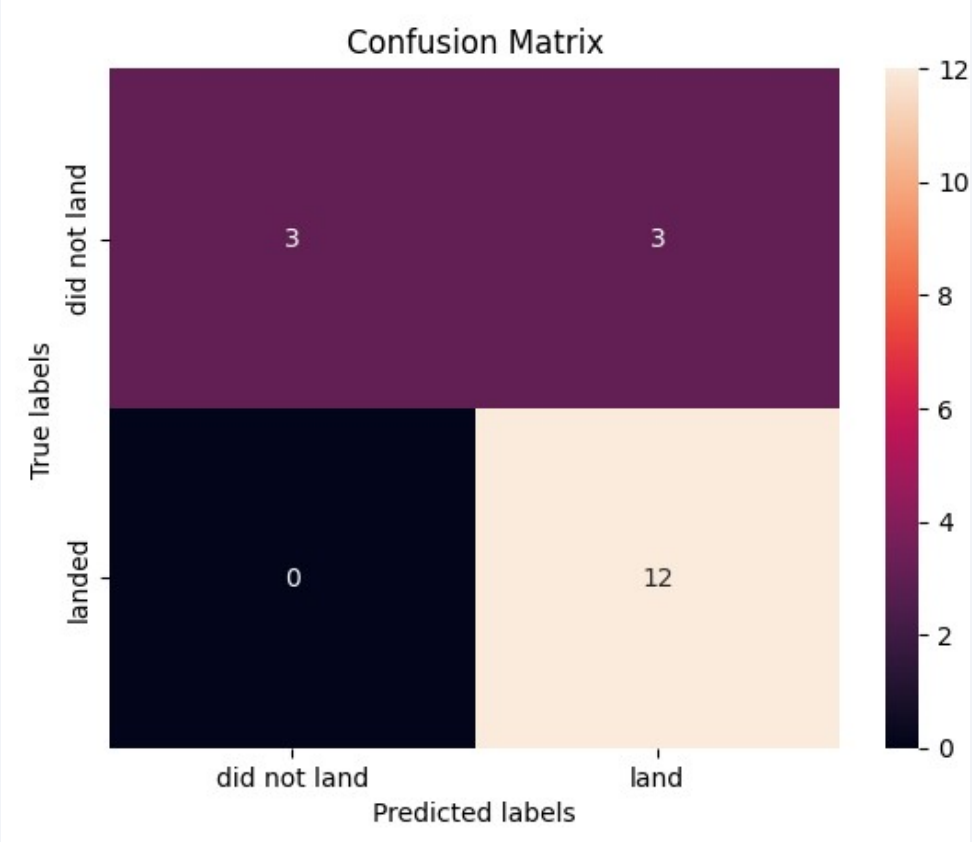
- Replace <Dashboard screenshot 3> title with an appropriate title
- Show screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider
- Explain the important elements and findings on the screenshot, such as which payload range or booster version have the largest success rate, etc.

Section 5

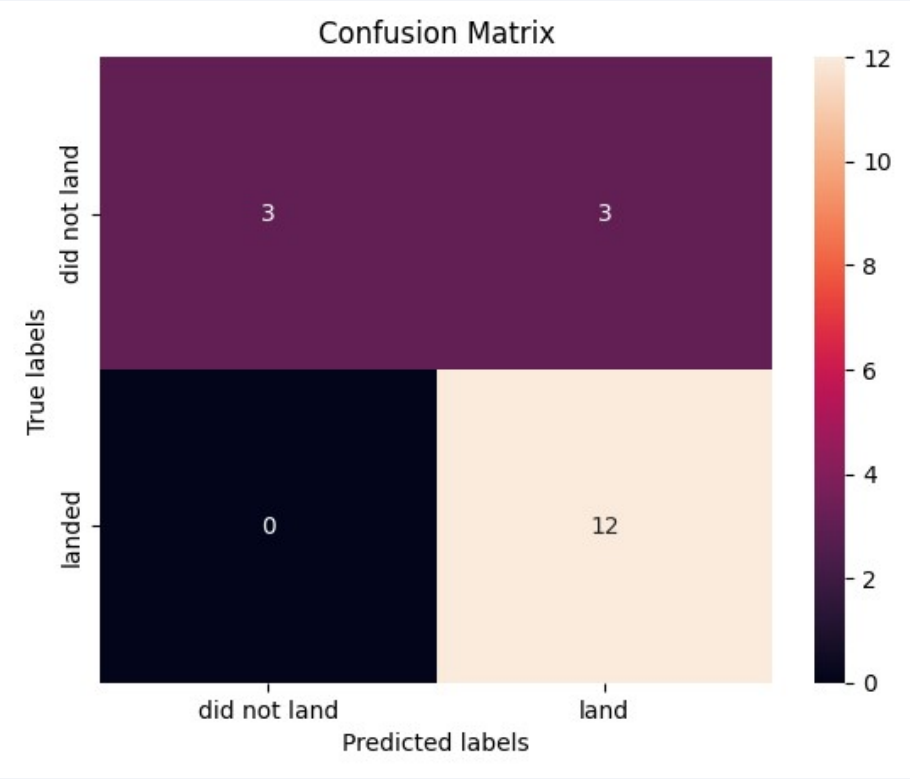
Predictive Analysis (Classification)

Confusion Matrix

Decision Tree



KNN



Conclusions

- The SVM, KNN, and Logistic Regression models demonstrated the highest prediction accuracy for this dataset.
- Lower-weight payloads performed better compared to heavier ones.
- The success rates of SpaceX launches have shown a direct positive correlation with time, indicating that launch performance is likely to improve over the years.
- The KSC LC 39A launch site recorded the highest number of successful launches compared to other sites.
- The orbits GEO, HEO, SSO, and ES L1 exhibited the best success rates.

Thank you!