IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Raphael Lawrence Farodoye
03/02/2026

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection via API, SQL and Web Scrapping
  - Data Wrangling and Analysis
  - Interactive Maps with Folium
  - Predictive Analysis for each classification model
- Summary of all results
  - Data Analysis along with interactive Visualizations
  - Best Model for Predictive Analysis

Introduction

- Project background and context

  We predicted if the Falcon 9 first stage will land successfully . 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 successfully. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

- Problems you want to find answers
  - With what factors will the rocket land successfully?
  - The effect of each relationship of rocket variables on outcome.
  - Conditions which will aid SpaceX have to achieve the best results.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Via SpaceX Rest API
  - Web Scrapping from Wikipedia
- Perform data wrangling
  - One hot encoding of categorical features
- Perform exploratory data analysis (EDA) using visualization and SQL
  - One hot encoding of categorical features
  - Scatterplot and bar graph to show patterns between features
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- Describe how data sets were collected.
- The data are collected using a get request to the Space X API.
- The dataset was converted to DataFrame.
- The dataset was filtered to include only Falcon 9 launches.
- The missing values in PayloadMass column are replaced with the mean.

Data Collection – SpaceX API

```
spacex_url="https://api.spacexdata.com/v4/launches/past"
```

```
In [8]:
response = requests.get(spacex_url)
```

Data Response from API

Convert json to DataFrame

Apply Custom functions to clean data

Filter DataFrame and export filer

```
# Call getLaunchSite
getLaunchSite(data)
```

```
# Call getPayloadData
getPayloadData(data)
```

```
# Call getCoreData
getCoreData(data)
```

```
# Use json_normalize meethod to convert th

data = pd.json_normalize(response.json())
```

```
data_falcon9 = df[df['BoosterVersion']!= 'Falcon 1']
```

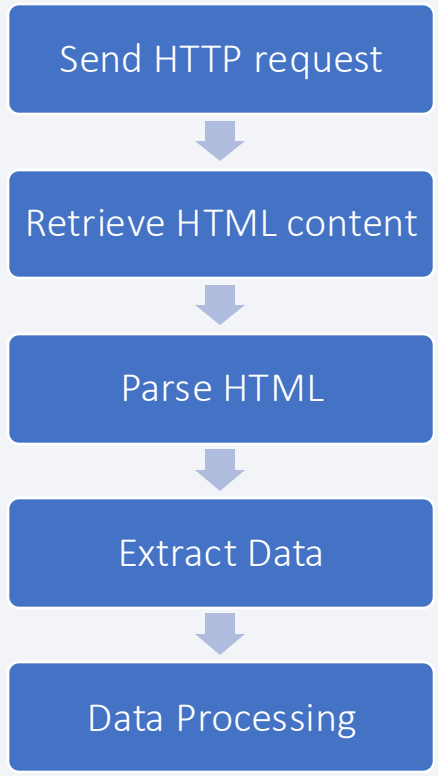
Now that we have removed some values we should reset the FlgihtNumber column

```
data_falcon9.loc[:,'FlightNumber'] = list(range(1, data_falcon9.shape[0]
data_falcon9
```

GITHUB URL

Data Collection - Scraping

- Send HTTP request: Access the desired website (e.g Wikipedia Page)
- Retrieve HTML content: Load the entire HTML structure of the webpage
- Parse HTML: Use libraries like BeatifulSoup to process the HTML data
- Extract data: Extract specific data (e.g., tables or text) from the HTML content
- Data Preprocessing: Convert the extracted data into a DataFrame, Clean and filter it

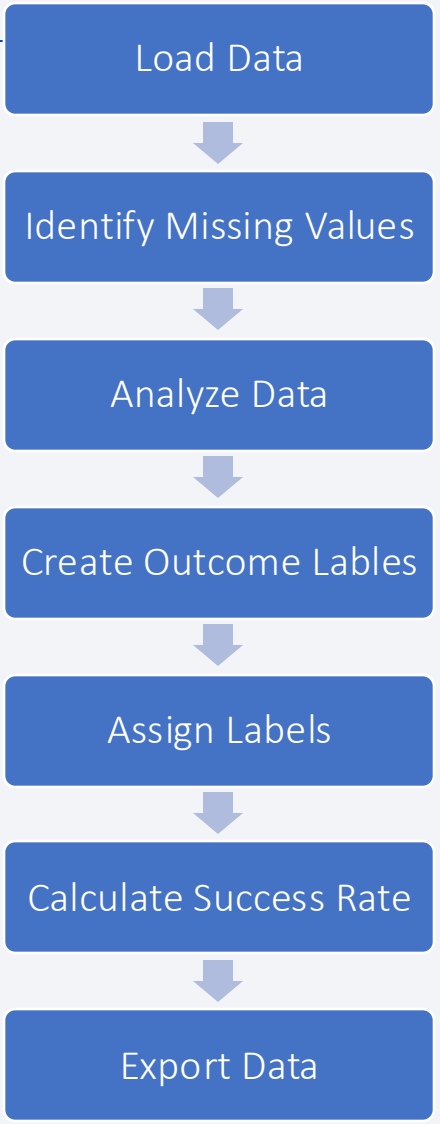


GITHUB URL

Data Wrangling

- Objective: Convert outcomes into training labels (1 = success, 0 = failure).
- Install Libraries: Use pandas and numpy for data manipulation.
- Load Data: Read SpaceX dataset with pd.read_csv().
- Identify Missing Values: Calculate percentage of missing values for each attribute.
- Data Analysis:
  - Determine number of launches at each site using value_counts().
  - Calculate occurrence of each orbit.
  - Analyze outcomes in the "Outcome" column.
- Categorize Outcomes:
  - Create a set of unsuccessful outcomes.
  - Assign labels (0 for failure, 1 for success) to a new "Class" column.
- Calculate Success Rate: Compute the average of the "Class" column to determine success rate.
- Export Data: Save the processed dataset as a CSV file for further analysis.

GITHUB URL



EDA with Data Visualization

- Sum Objective: Explore the SpaceX dataset to uncover patterns and insights related to launch success.
- Install Libraries: Utilize pandas, numpy, seaborn, and matplotlib for data manipulation and visualization.
- Load Data: Read the SpaceX dataset using pd.read_csv().
- Data Overview: Display the first few rows of the dataset to understand its structure and contents.
- Visualize Relationships:
  - Create scatter plots to show the relationship between FlightNumber and PayloadMass against launch outcomes.
  - Use bar charts to visualize the success rate for different orbits.
  - Analyze the success rate across different launch sites using categorical plots.
- Identify Patterns:
  - Assess how the number of flights impacts the success rate.
  - Observe trends in payload mass concerning successful landings.
- Export Insights: Save visualizations and analyses for presentation and reporting

GITHUB URL

EDA with SQL

- Objective Objective: Perform SQL queries on the SpaceX dataset to analyze various aspects of launch data.
- Connect to Database: Load the SpaceX dataset into a Db2 database for querying.
  - Task 1: Display unique launch sites from the dataset.
  - Task 2: Retrieve 5 records where launch sites begin with 'CCA'.
  - Task 3: Calculate total payload mass carried by boosters launched by NASA (CRS).
  - Task 4: Display average payload mass for booster version F9 v1.1.
  - Task 5: List the date of the first successful landing on a ground pad.
  - Task 6: List boosters with successful drone ship landings and payload mass between 4000 and 6000 kg.
  - Task 7: Count successful and failed mission outcomes.
  - Task 8: Find booster versions that carried the maximum payload mass using a subquery.
  - Task 9: List records with landing failures on a drone ship for 2015, including month names.
  - Task 10: Rank landing outcomes between specific dates in descending order

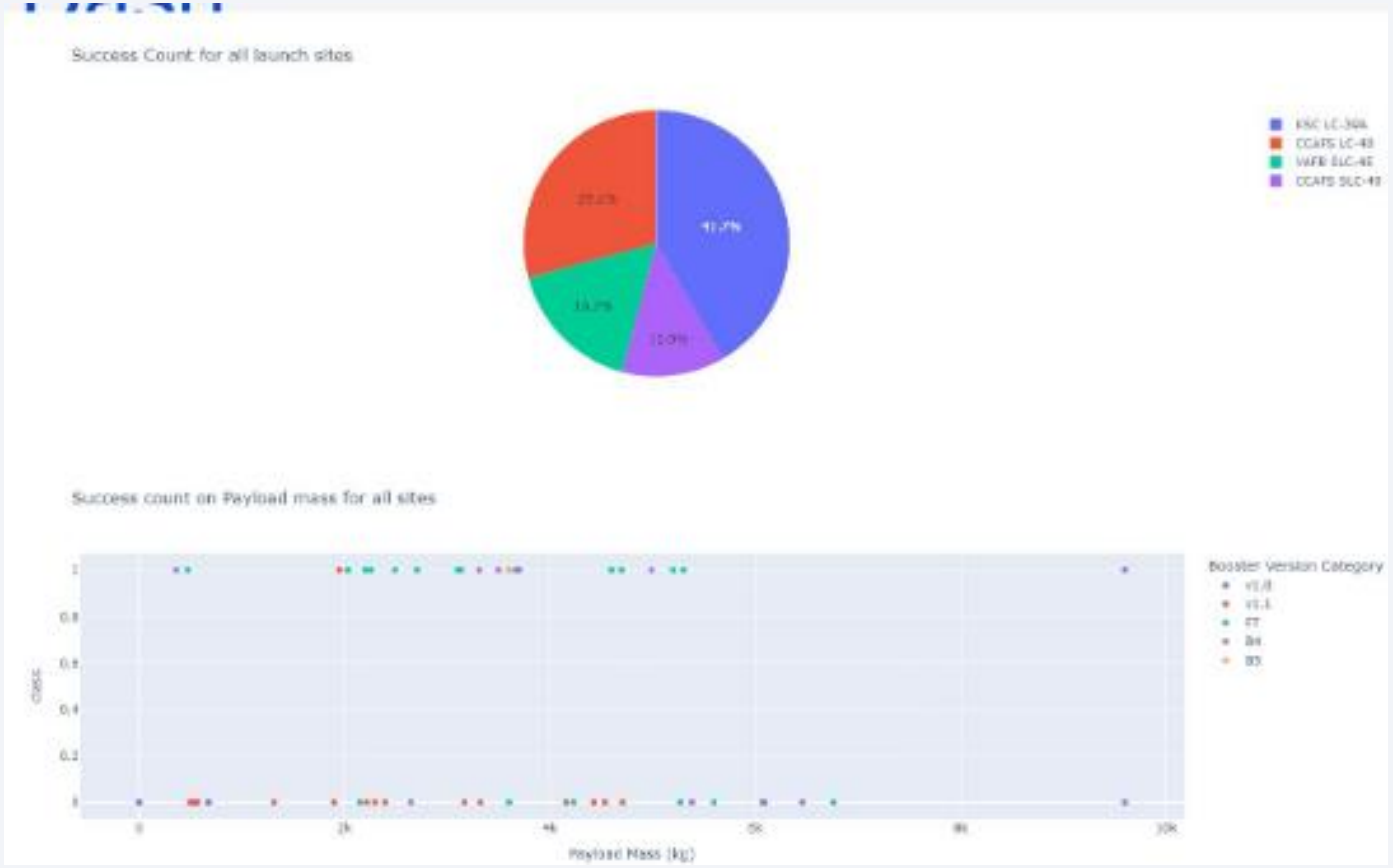
GITHUB URL

Build an Interactive Map with Folium

- Markers:
  - Added to indicate launch sites for Falcon 9 missions.
  - Provide interactive labels with details about each launch site.
- Circles:
  - Represented the area of influence around each launch site.
  - Help visualize the geographical context of launch operations.
- Popups:
  - Included with markers to display additional information about each launch, such as launch dates and success rates.
- Enhance user engagement and understanding of the data.

GITHUB URL

Build a Dashboard with Plotly Dash



Pie chart to show the success launches of all / each site.

Scatter plot to show the success launches of all / each site by payload mass

GITHUB URL

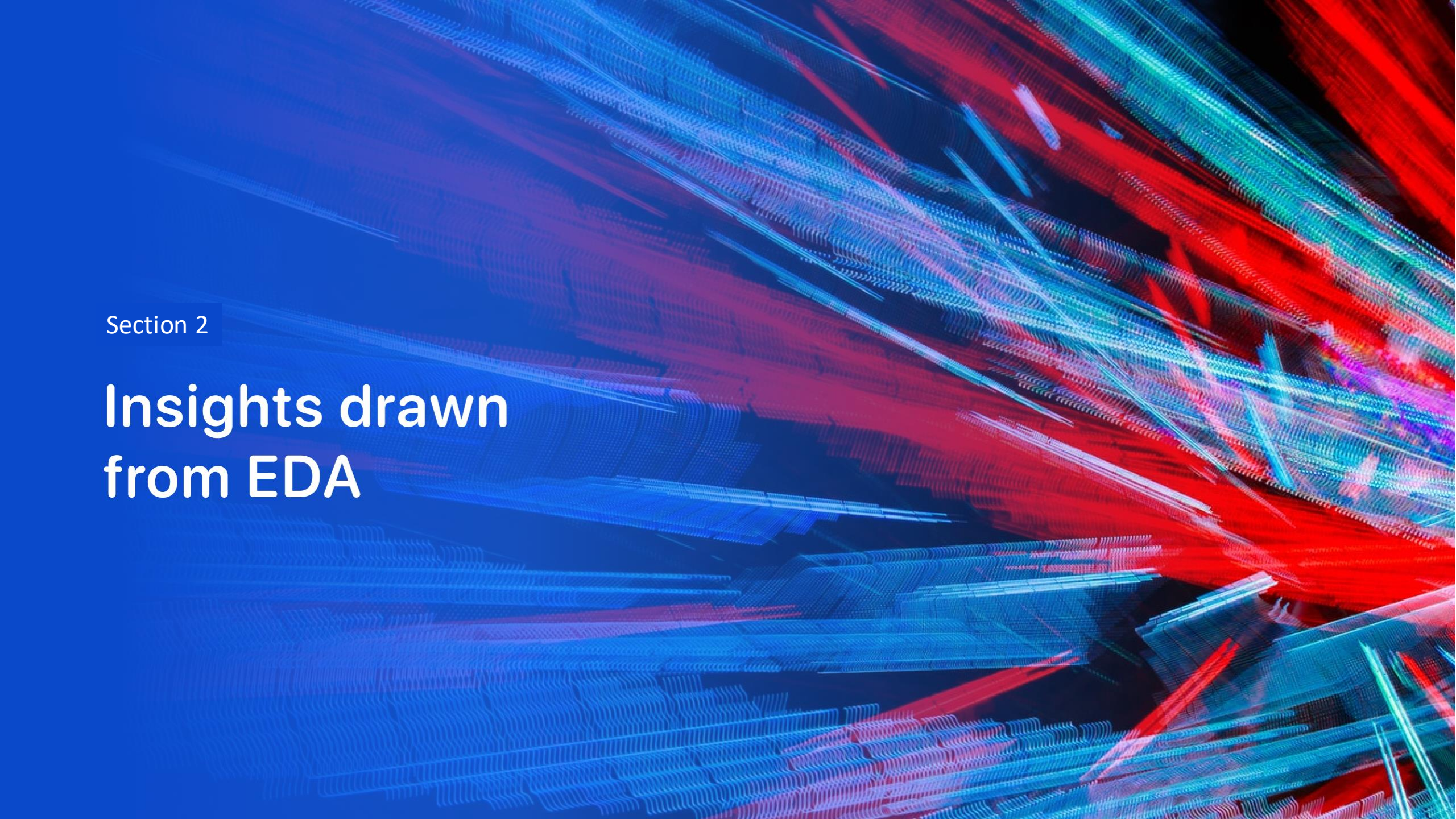
Predictive Analysis (Classification)

- Different models Logistic Regresssion (LR), Support Vector Machine (SVM), Decision Tree and K Nearest Neighbors (KNN) are trained and fitted with GridSearchCV for hyper parameter tunning, to find the best parameters.
- LR, SVM and KNN achieved an equal accuracy of 83,33 %

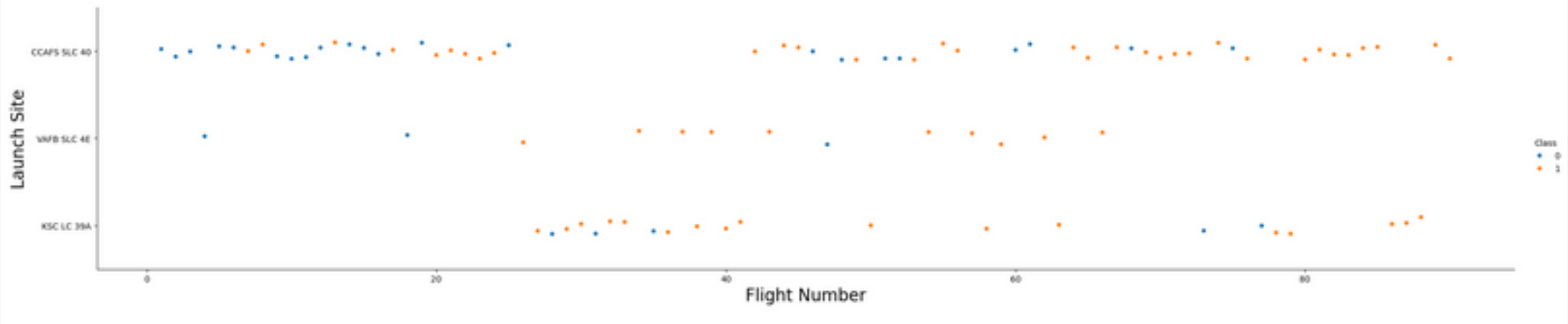
GITHUB URL

Results

- Exploratory data analysis results:
  - Space X uses 4 different launch sites;
  - Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average.
  - The number of landing outcomes became as better as years passed.
  - LR, SVM, KNN are the top-performing models for forecasting outcomes in this data
  - GEO, HEO, SSO, ES L1 orbit types exhibit the highest rates of successful launches

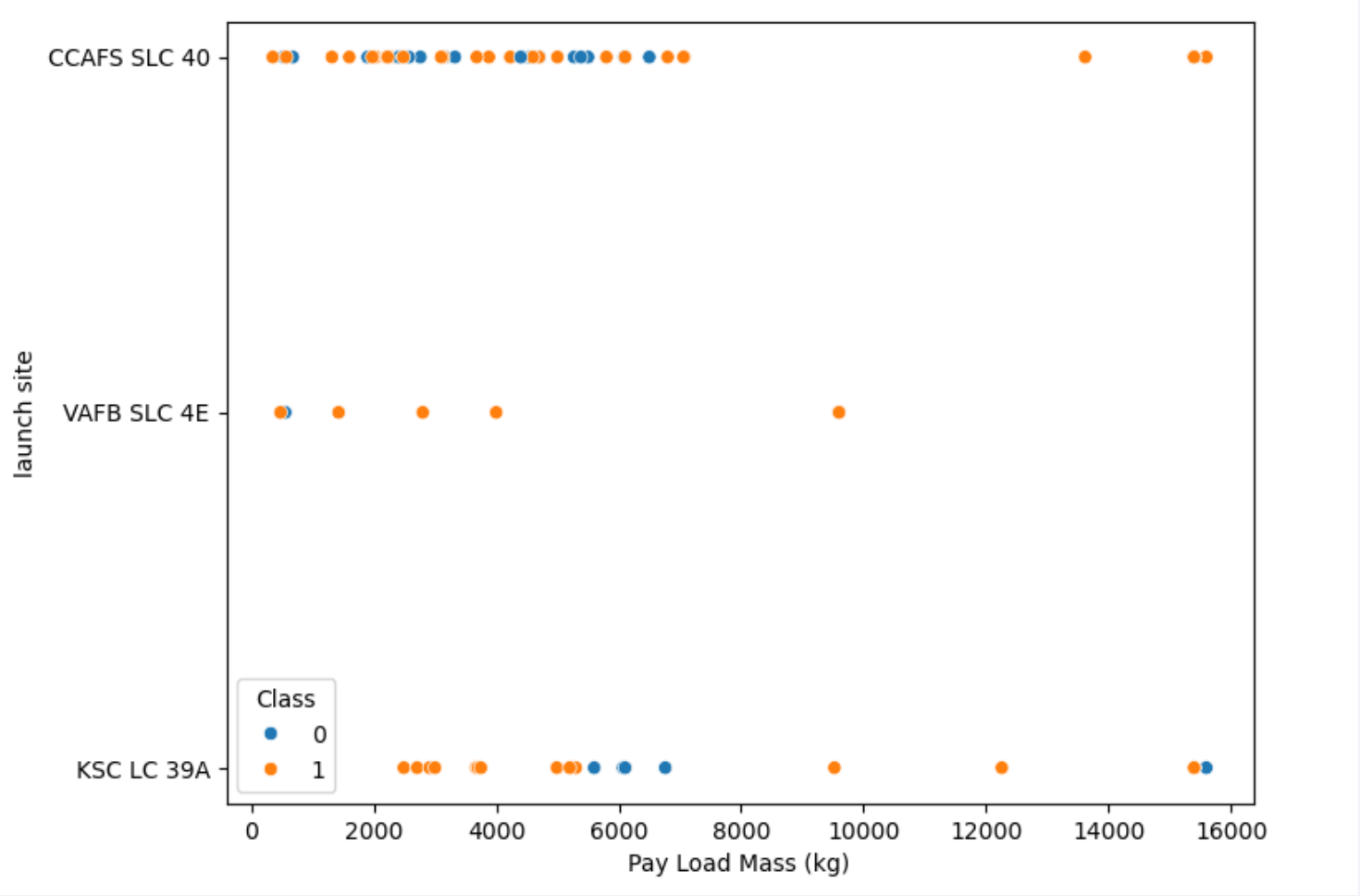

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



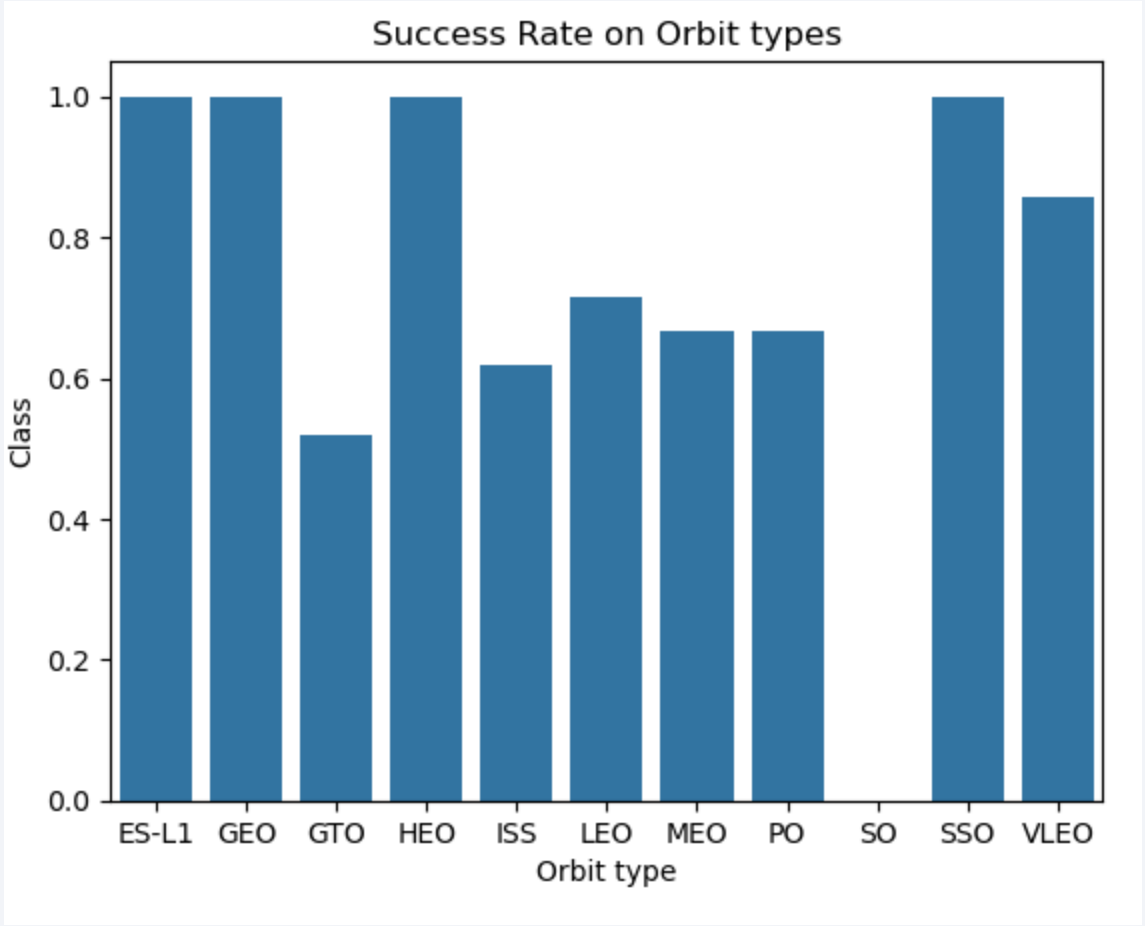
Total numbers of launches from launch site CCAFS SLC 40
Are significantly higher than the other launches sites

Payload vs. Launch Site



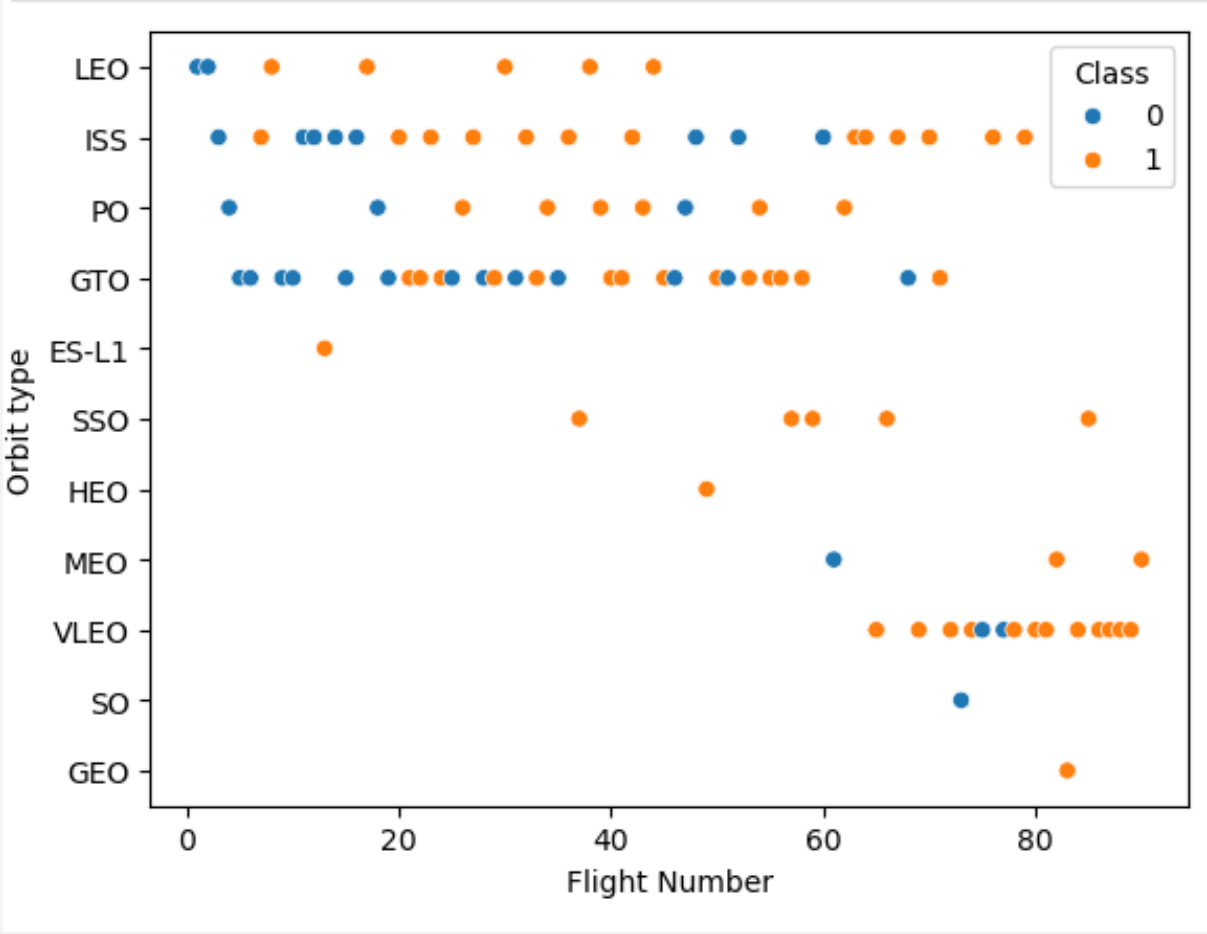
Payloads with lower mass are have more launches compared to those with higher mass across all three launch sites

Success Rate vs. Orbit Type



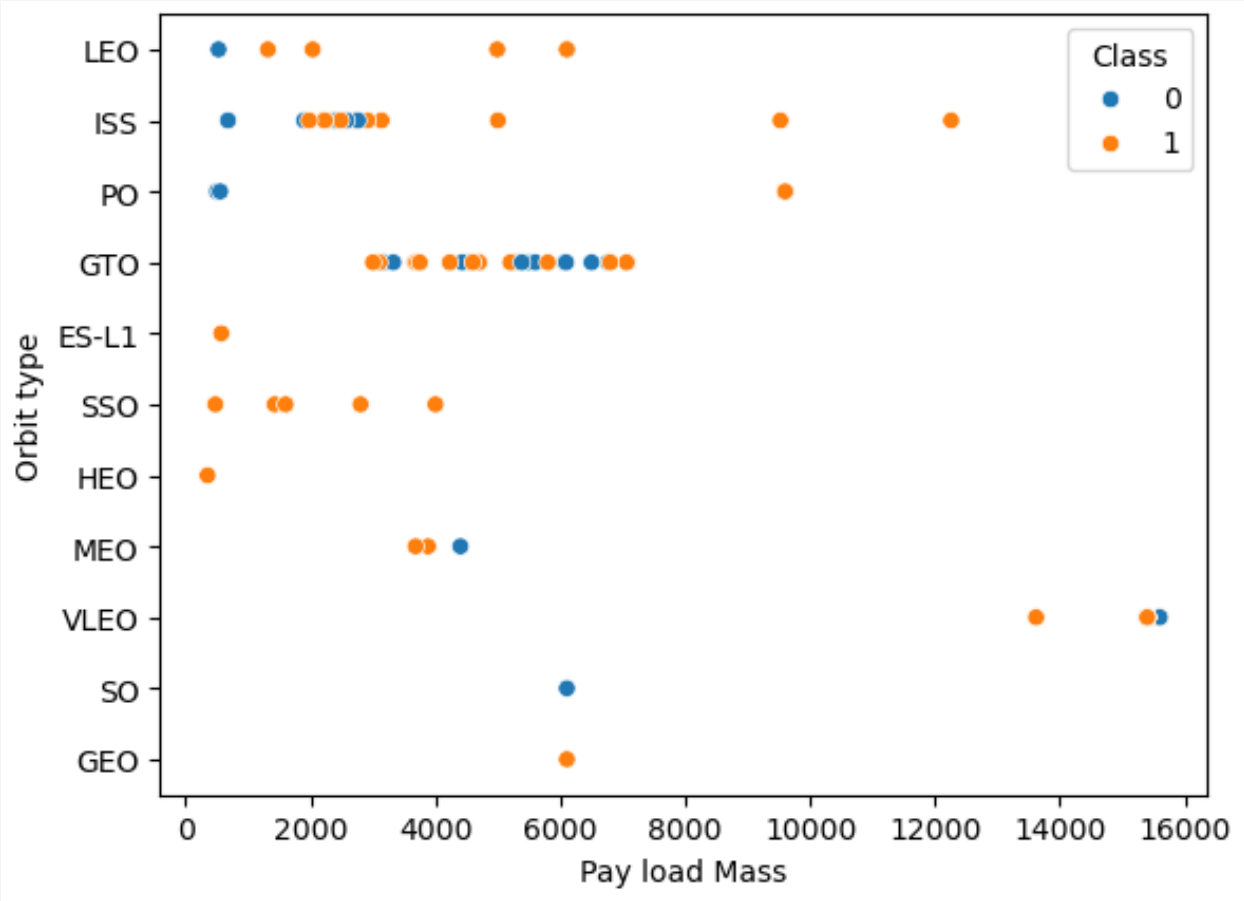
Orbit types ES-L1, GEO, HEO, SSO have the highest success rate among all.

Flight Number vs. Orbit Type



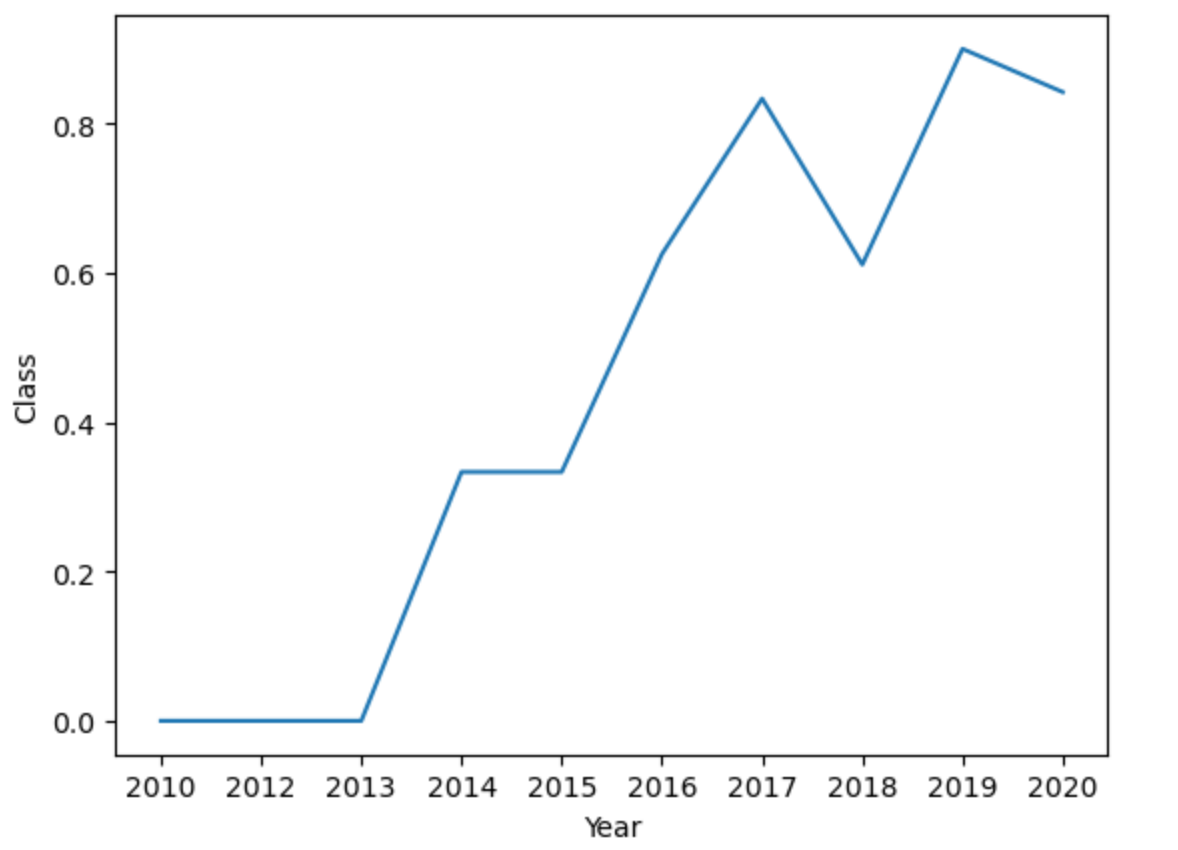
LEO, ISS, PO, GTO orbits have the most launches in the earlier years, but it slowly shifted to VLEO orbit in the later years

Payload vs. Orbit Type



Heavy payloads tend to have higher successful landing rates for PO, LEO and ISS orbit, gut GTO orbit success is less predictable with an almost equal mix of success and failures

Launch Success Yearly Trend



The success rate of launches kept increasing from 2013 and reduced in 2018 and spiked again, there are probably advancement in technologies or they tried different approaches

All Launch Site Names

```
%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

The rocket launches took place in 4 launch sites.

Launch Site Names Begin with 'CCA'

```
%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

Obtained 5 launch sites records than starts with "CCA"

Total Payload Mass

Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql SELECT SUM(PAYLOAD_MASS__KG_)  FROM SPACEXTBL WHERE CUSTOMER="NASA (CRS)";
```

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

SUM(PAYLOAD_MASS__KG_)
45596

Performed an SQL query to obtain the total payload mass carried by boosters launched by NASA (CRS)

Average Payload Mass by F9 v1.1

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

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

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

Performed an SQL query to calculate the average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date

```
%sql SELECT MIN(Date), * FROM SPACEXTABLE WHERE Landing_Outcome LIKE '%ground pad%';
```

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

MIN(Date)	Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2015-12-22	2015-12-22	1:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)

Performed an SQL query to find the date of the first successful landing outcome on ground pad

Successful Drone Ship Landing with Payload between 4000 and 6000

Booster_Version	Landing_Outcome
F9 FT B1032.1	Success (ground pad)
F9 B4 B1040.1	Success (ground pad)
F9 B4 B1043.1	Success (ground pad)

Performed an SQL query to list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT Mission_Outcome, Count(*) AS TOTAL FROM  SPACEXTABLE GROUP BY Mission_Outcome;
```

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

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

Performed an SQL query to calculate the total number of successful and failure mission outcomes

Boosters Carried Maximum Payload

```
%sql SELECT Booster_Version, (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE) AS MAX_PAYLOAD FROM SPACEXTABLE;
```

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

Booster_Version	MAX_PAYLOAD
F9 v1.0 B0003	15600
F9 v1.0 B0004	15600
F9 v1.0 B0005	15600
F9 v1.0 B0006	15600
F9 v1.0 B0007	15600
F9 v1.1 B1003	15600
F9 v1.1	15600
F9 v1.1	15600
F9 v1.1	15600
F9 v1.1	15600
F9 v1.1	15600

Performed an SQL query to list the names of the boosters which have carried the maximum payload mass

2015 Launch Records

```
%sql SELECT Booster_Version, (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE) AS MAX_PAYLOAD FROM SPACEXTABLE;
```

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

Booster_Version	MAX_PAYLOAD
F9 v1.0 B0003	15600
F9 v1.0 B0004	15600
F9 v1.0 B0005	15600
F9 v1.0 B0006	15600
F9 v1.0 B0007	15600
F9 v1.1 B1003	15600
F9 v1.1	15600
F9 v1.1	15600
F9 v1.1	15600
F9 v1.1	15600
F9 v1.1	15600

Performed an SQL query to list failed landing outcomes in drone ship, their booster versions, and launch sit names for the year 2015

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

```
%sql SELECT Landing_Outcome ,COUNT(*) AS Total FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORD
```

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

Landing_Outcome	Total
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

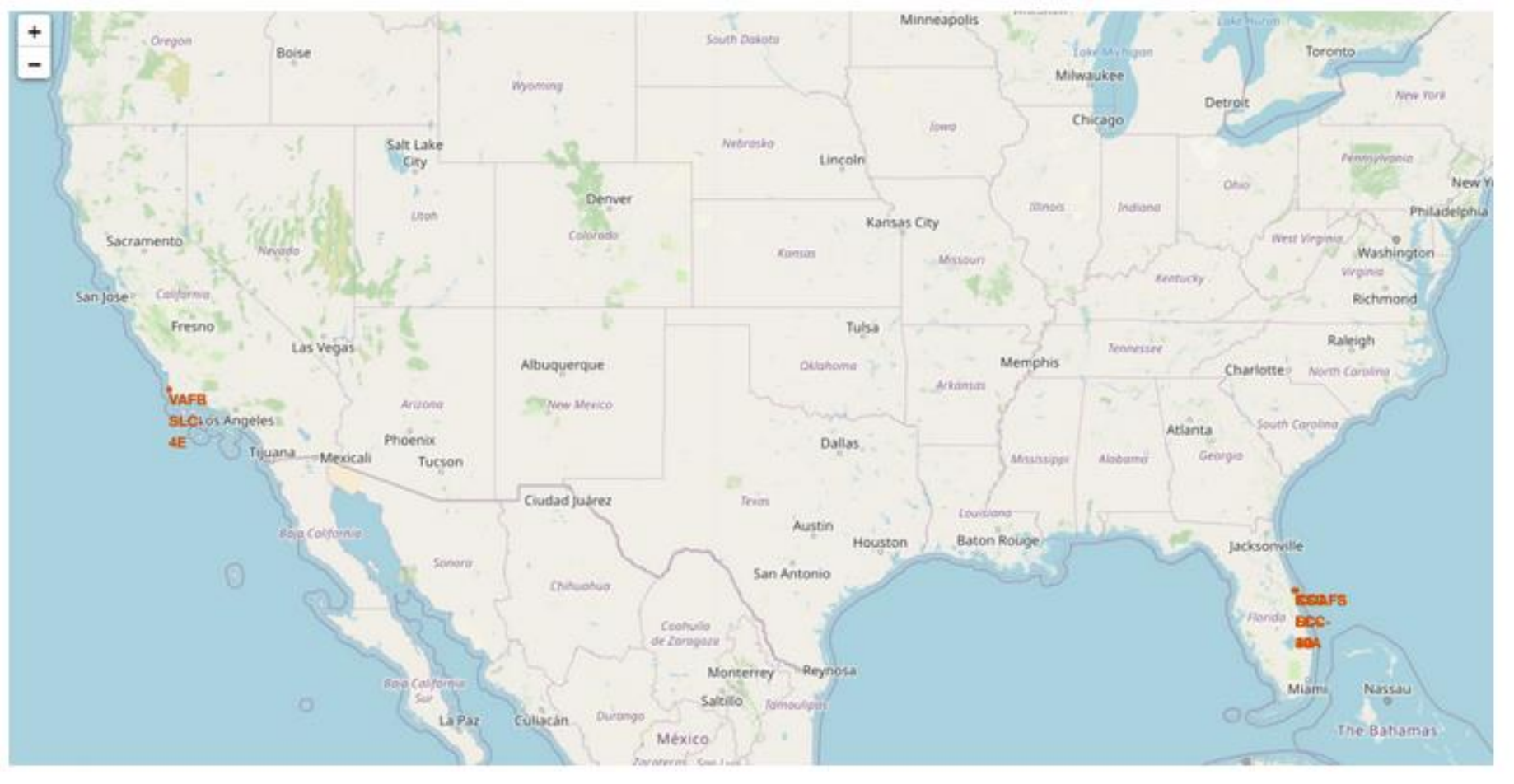
We selected Landing outcomes and the COUNT of landing outcomes from the data and used the WHERE clause to filter for landing outcomes BETWEEN 2010-06- 04 to 2010-03-20.

We applied the GROUP BY clause to group the landing outcomes and the ORDER BY clause to order the grouped landing outcome in descending order.

Section 3

Launch Sites Proximities Analysis

Marked Launch Sites



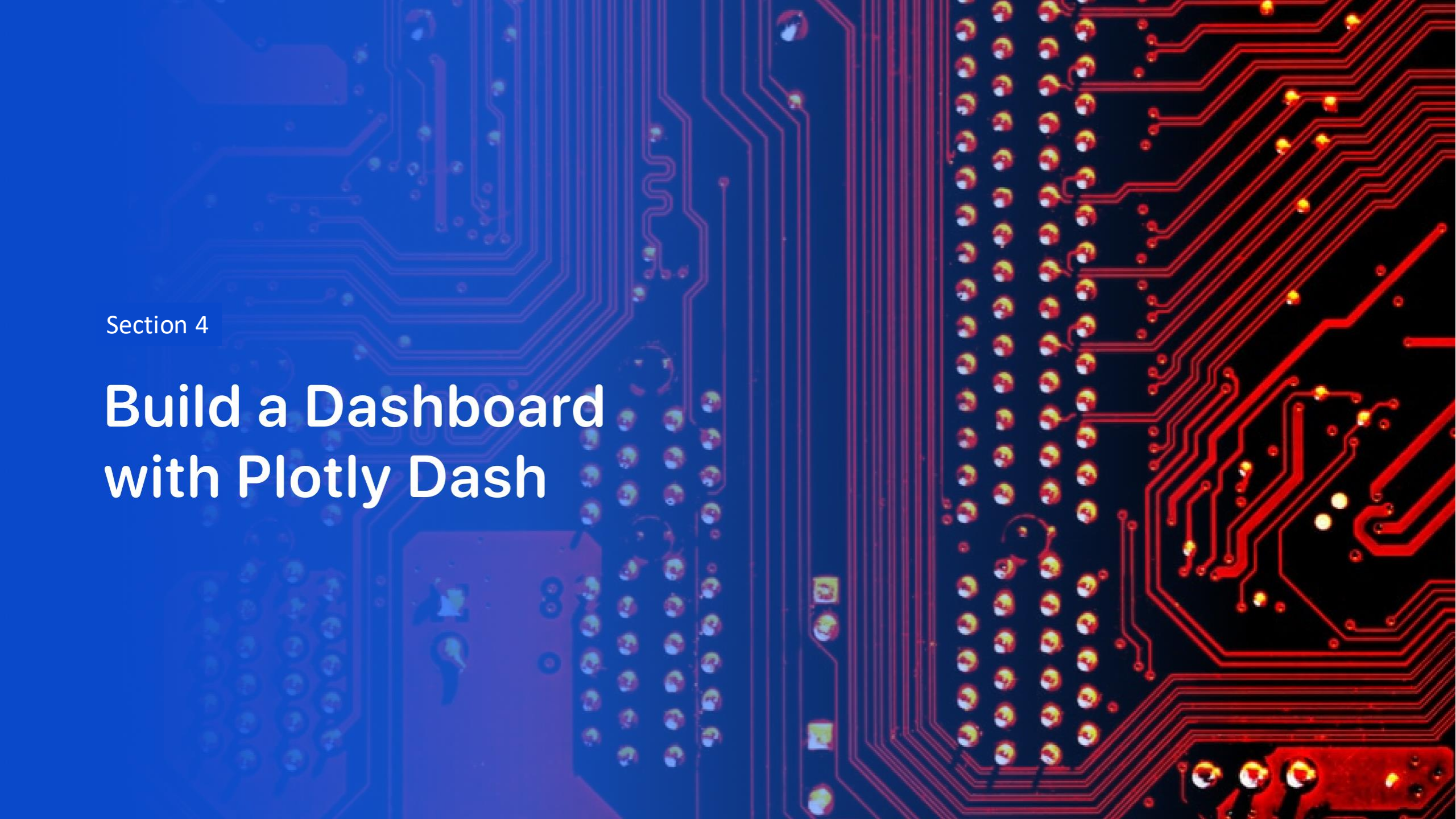
The launch sites are labelled by a marker with their names on the map

Launch Sites with high success rates

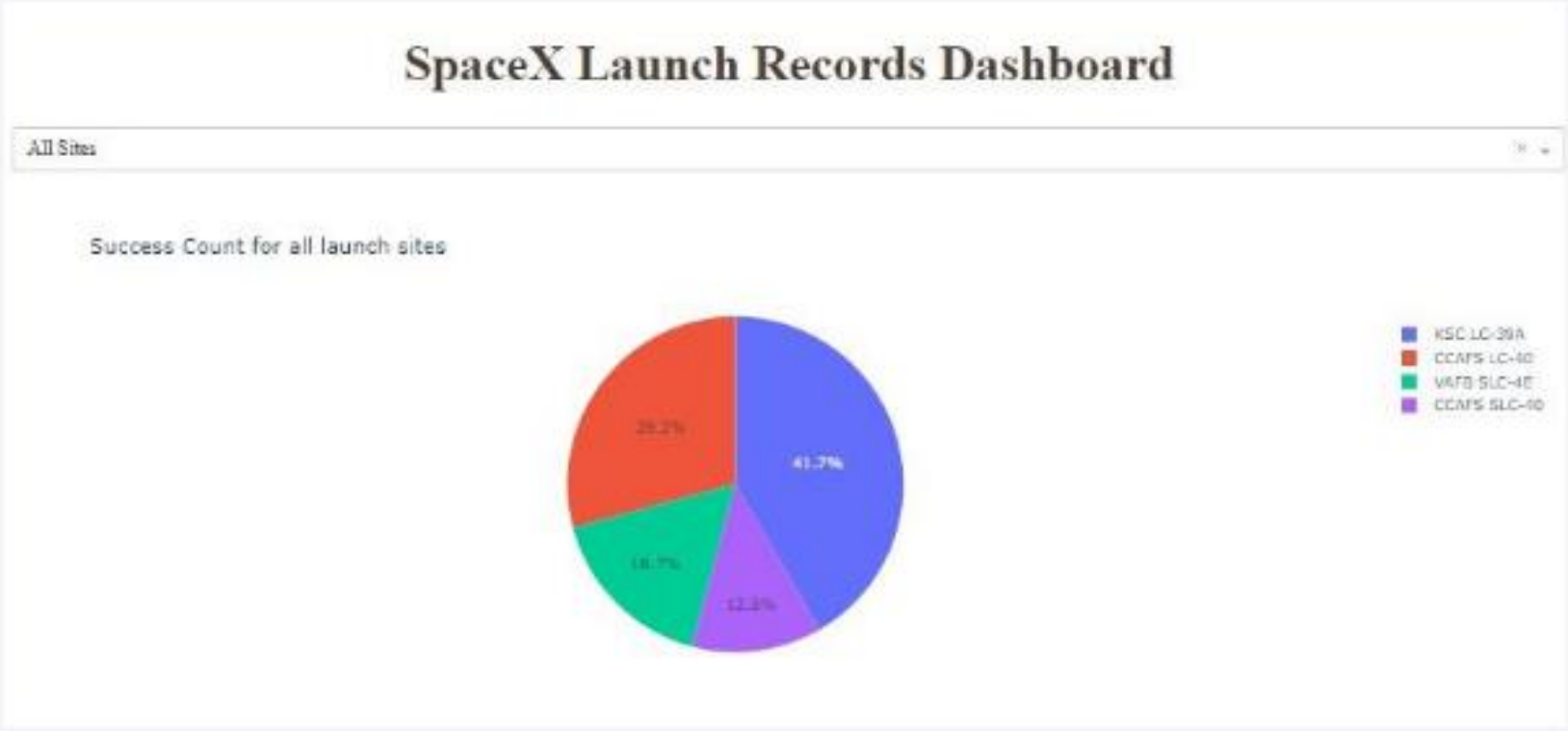
The launch records are grouped in clusters on the map, then labelled by green markers for successful launches, and red markers for failure

Distance between a launch site to its proximities

Draw a PolyLine between a launch site to the selected coastline point

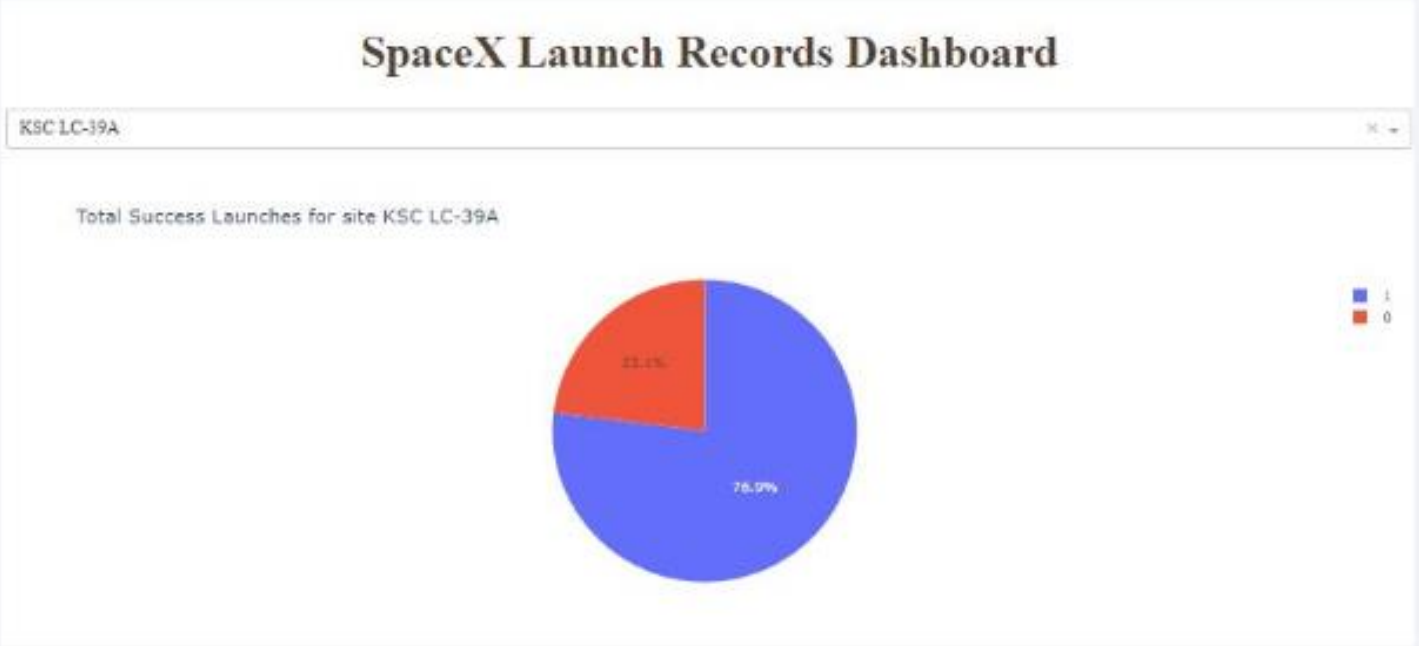

Section 4
Build a Dashboard with Plotly Dash

Total Success Launches for all sites



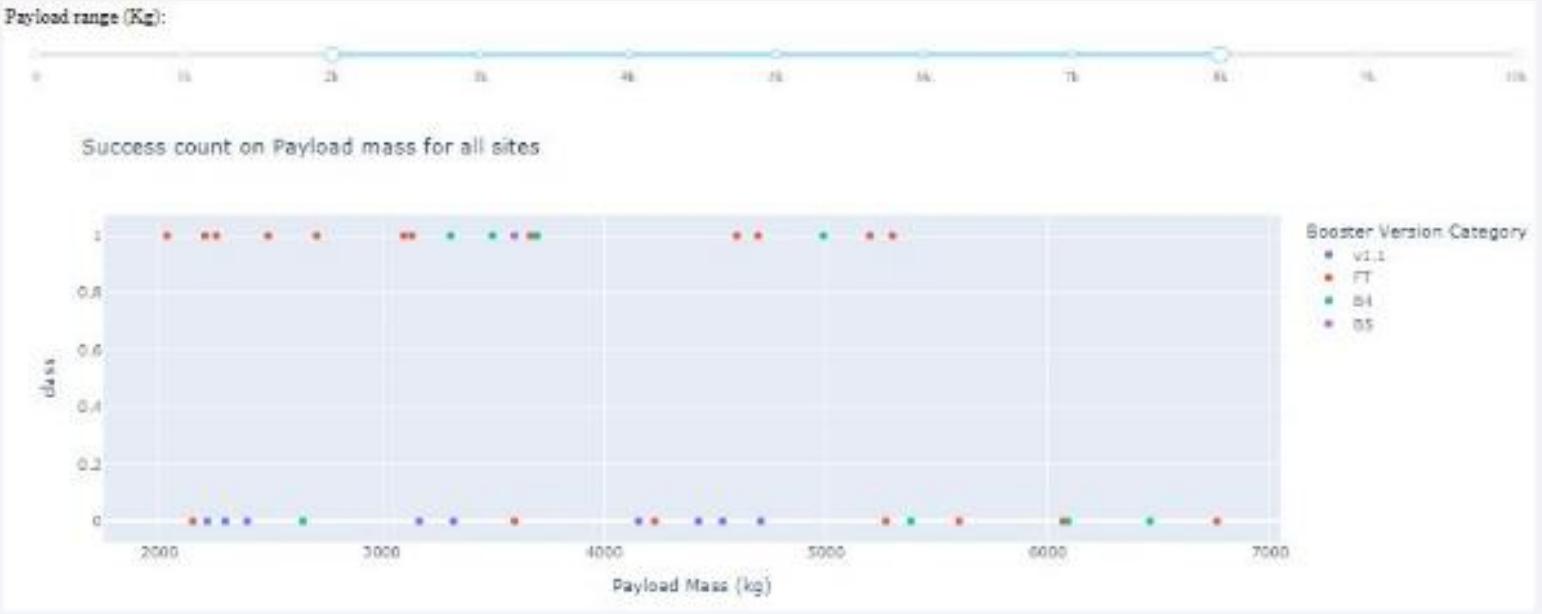
KSC LC-39A has the highest amount of success launches with 41.7% from the entire record, CCAFS SLC-40 has the lowest amount of success launches with only 12,5 %

Piechart of the launch site with the highest launch success



KSC LC-39A which is the launch site with highest amount of success, has a 76,9 success rate for the launches from its site, and 23,1 failure rate

Payload range with highest success launches

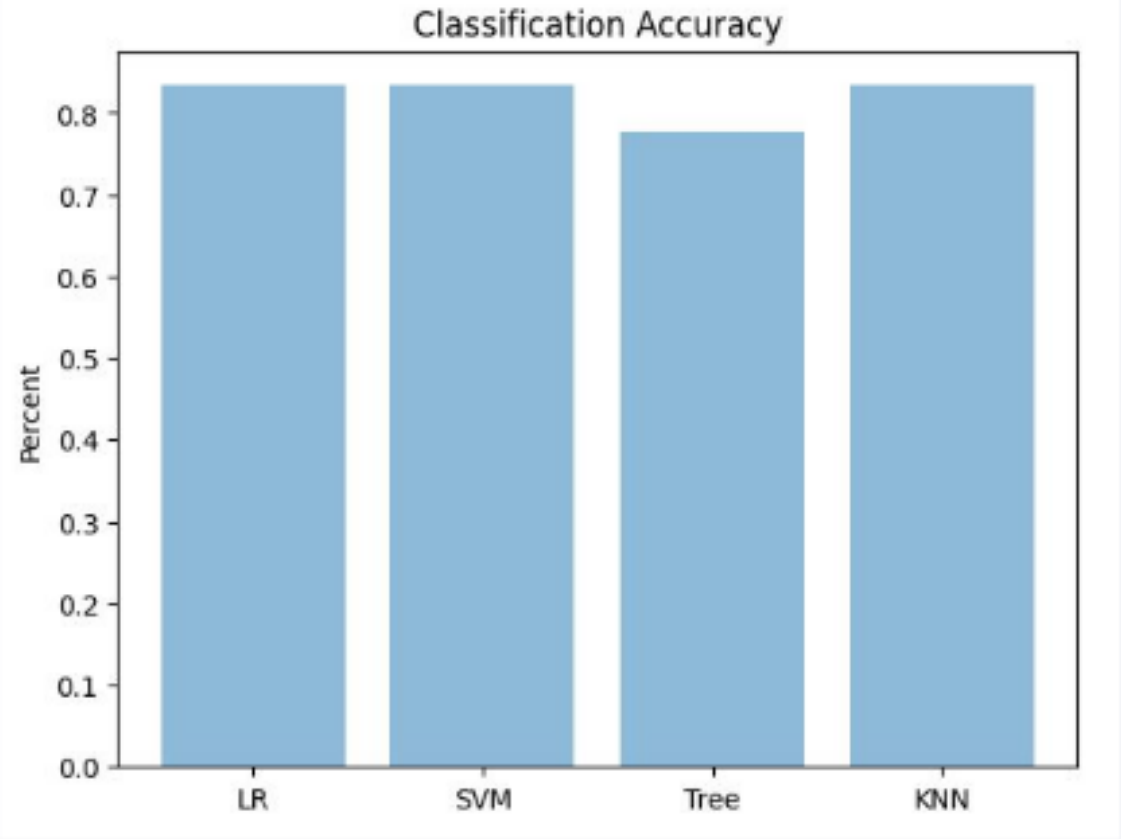


The payload range that has the highest success launches is between 2.000 and 4.000 kg, which can be seen the most number of plots in that range

Section 5

Predictive Analysis (Classification)

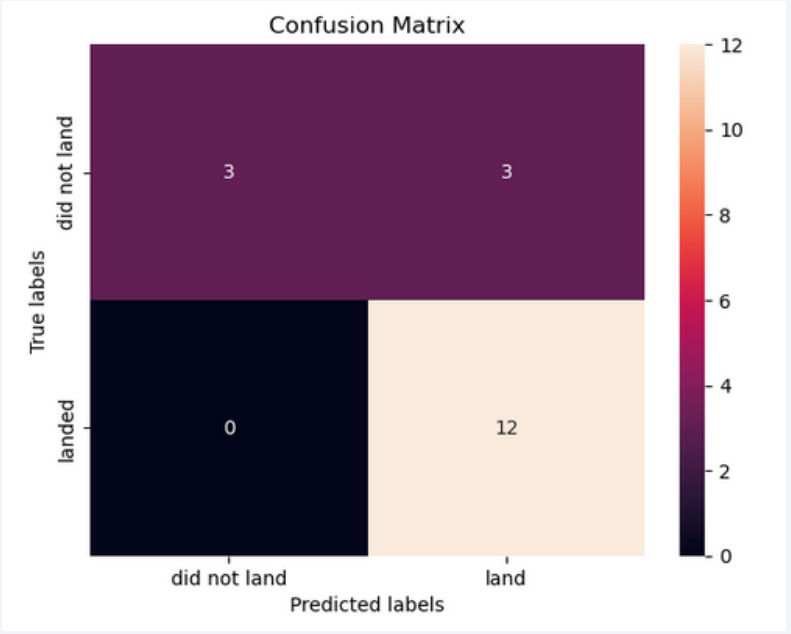
Classification Accuracy



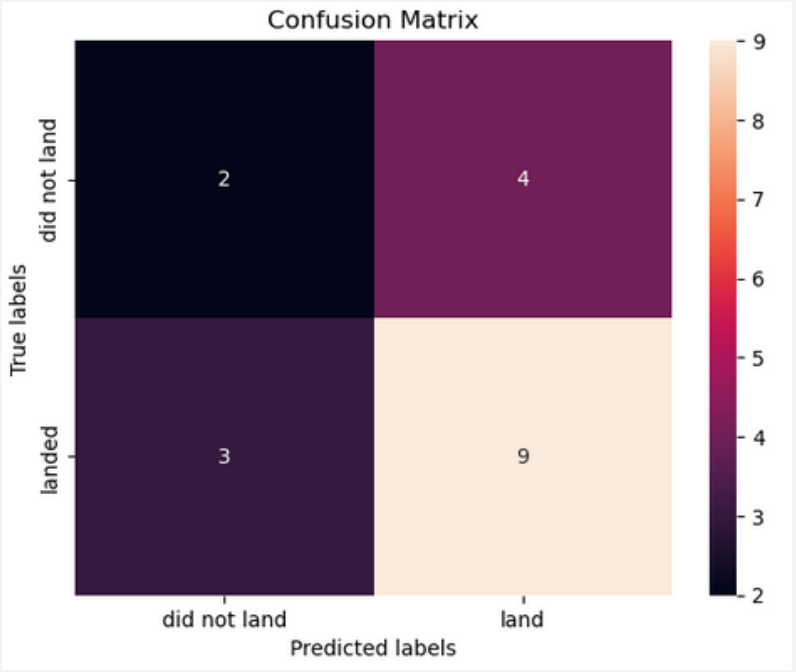
The best performed methods are: LR, SVM, KNN where all 3 achieved the highest accuracy of 83,33%

Confusion Matrix

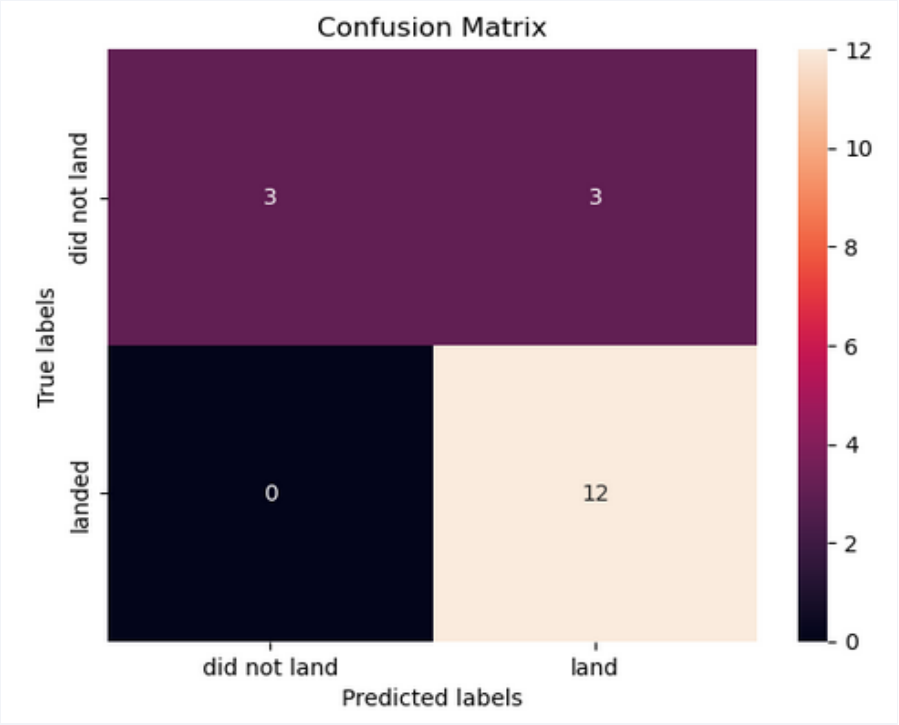
Logistic Regression Confusion Matrix



SVM Confusion Matrix



KNN Confusion Matrix



LR, SVM, KNN models have the same accuracy of 83,33% as displayed earlier, hence the same confusion matrix

Conclusions

- We can conclude that:
  - The larger the flight amount at a launch site, the greater the success rate at a launch site.
  - Launch success rate started to increase in 2013 till 2020.
  - Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
  - KSC LC-39A had the most successful launches of any sites.
  - The Decision tree classifier is the best machine learning algorithm for this task.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!