IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Hassan A. Baabdullah
Wednesday, January 20, 2024

Outline

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

Executive Summary

- Summary of Used Methodologies:
  - Using public information to predict if SpaceX will reuse the first stage.
  - Collecting data from SpaceX REST API and using this data to predict whether SpaceX will attempt to land a rocket or not.
  - Performing needed data wrangling to analyze and creating models.
- Summary of Results:
  - Different launch sites have different success rates. As a result, they can be used to help determine if the first stage will land successfully.

Introduction

- Project Background:
  - The commercial space age is here, with companies making space travel affordable for everyone. SpaceX leads the industry with accomplishments such as sending spacecraft to the International Space Station, launching Starlink satellite internet constellation, and conducting manned missions. SpaceX's cost-effective approach is due to reusable rocket launches, with Falcon 9 launches priced at $62 million, compared to other providers' $165 million. The first stage of Falcon 9 is critical and can be recovered, but its successful reuse is uncertain. As a data scientists for Space Y, our task is to gather information about SpaceX, create dashboards, and train a machine learning model to predict the reuse of the first stage.

Introduction

- Problems we are going to solve:
  - Understand the technical issues that cause fail launches.
  - Determine the best launching sites and circumstances.
  - Determine the price of each launch.
  - Determine if SpaceX will reuse the first stage.
  - Instead of using rocket science to determine if the first stage will land successfully, we will train a machine learning model and use public information to predict if SpaceX will reuse the first stage..

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Public information from List of "Falcon 9 and Falcon Heavy launches" (Wikipedia).
  - SpaceX REST API.
- Perform data wrangling
  - We will convert the [Outcome] to be classified, 0 (did not land) and 1 (did land).
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Using 'Exploratory Data Analysis' and observing success rate results.
  - In addition, we will determine what attributes are correlated with successful landings.

Data Collection

- We will use public information about *Falcon 9* rocket, which is available on Wikipedia page (List of Falcon 9 and Falcon Heavy launches) at https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches .

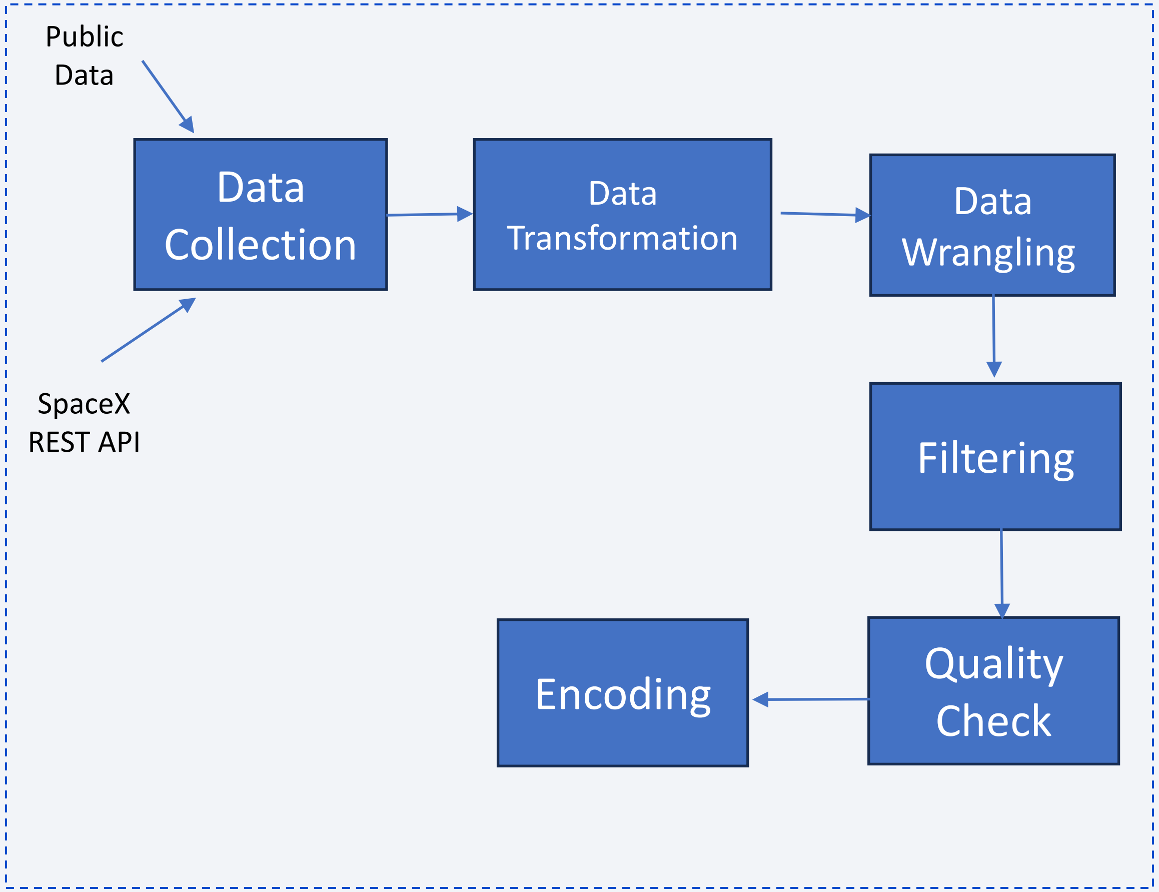
- Key Phrases:
  - Data Collection Source: SpaceX launch data gathered from the SpaceX REST API.
  - API Endpoint: The SpaceX REST API endpoints start with api.spacexdata.com/v4/ and include /capsules, /cores, and specifically /launches/past.
  - Data Transformation: JSON data from the API will be converted to a dataframe using the json_normalize function.
  - Data Sources: Web scraping of Wiki pages to obtain Falcon 9 launch data.

Data Collection (Continued..)

- Key Phrases (Continued...):
  - Data Wrangling: Wrangling data using an API, sampling data, and dealing with null values.
  - Filtering Data: Filtering out Falcon 1 launches from the dataset.
  - Handling Null Values: Calculating the mean of PayloadMass data and replacing null values with the mean.
  - Data Quality: Dealing with NULL values in PayloadMass and leaving the column LandingPad with NULL values for one hot encoding.

Data Collection – SpaceX API

- Data Collection Process: Gather data from the SpaceX REST API using the specific endpoint for past launches.
- Data Transformation Process: Convert JSON data to a DataFrame using the json_normalize function.
- Additional Data Source: Web scraping of Wiki pages to obtain Falcon 9 launch data.
- Data Wrangling Process: Wrangling data using an API, sampling data, and handling null values.
- Filtering Process: Filtering out Falcon 1 launches from the dataset.
- Data Quality Check: Calculating the mean of PayloadMass data and replacing null values with the mean.
- Data Encoding: Handling NULL values in LandingPad column using one hot encoding.

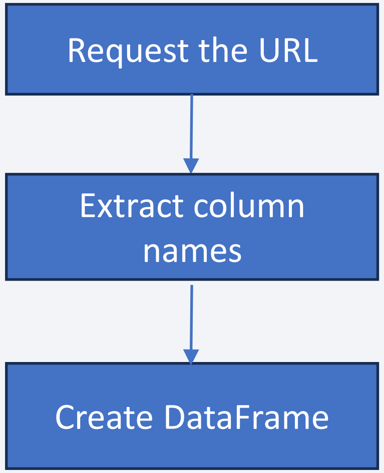


Data Collection - Scraping

- Web scraping process:
  - Using BeautifulSoup Library.
  - Extract a Falcon 9 launch records HTML table from Wikipedia
  - Parse the table and convert it into a Pandas data frame
- GitHub Repository URL: python-applied-data-science-capstone/jupyter-labs-webscraping.ipynb at master · hsnapps/python-applied-data-science-capstone (github.com)

Using the Wikipedia Page "List of Falcon 9 and Falcon Heavy launches" at

https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

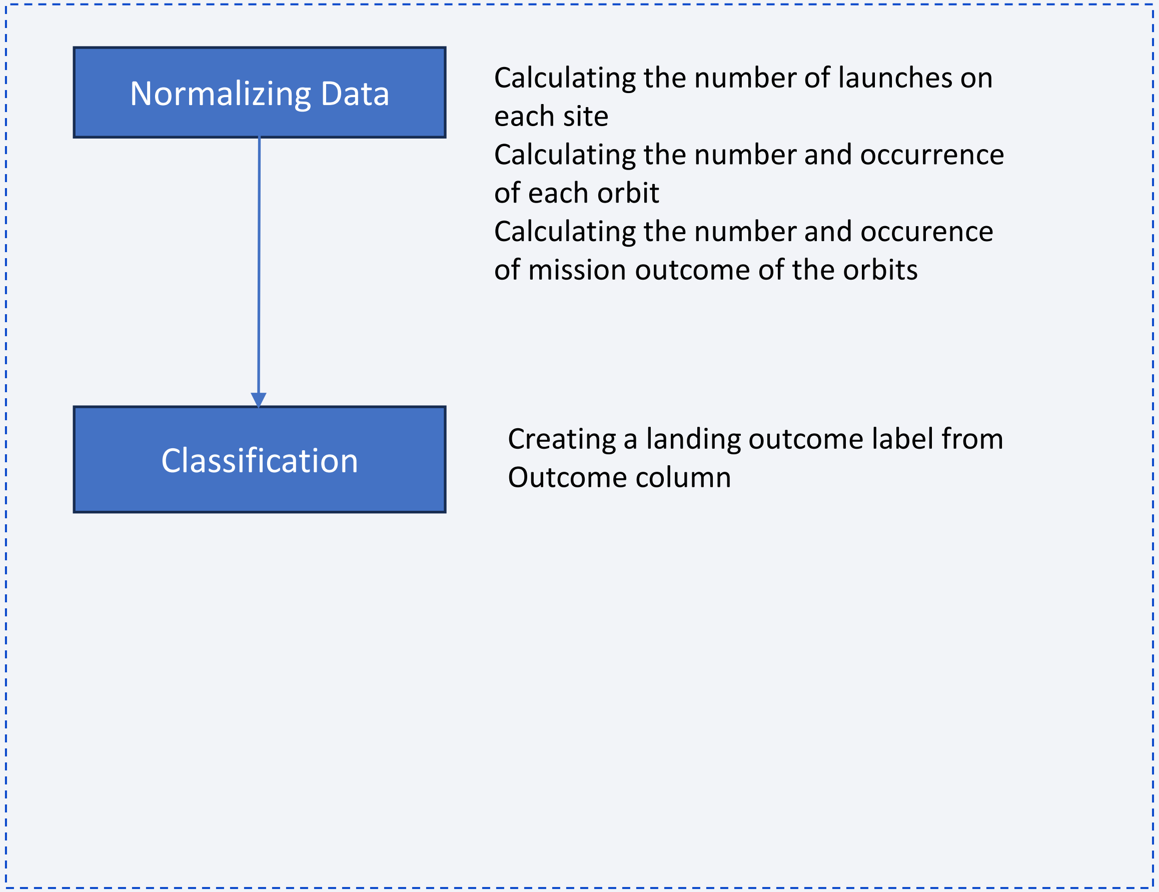


Extract column names from HTML table header

Create a data frame by parsing the launch HTML tables

Data Wrangling

- Normalizing Data :
  - Calculating the number of launches on each site.
  - Calculating the number and occurrence of each orbit.
  - Calculating the number and occurence of mission outcome of the orbits.
- Classification:
  - Creating a landing outcome label from Outcome column
- GitHub Repository URL: [python-applied-data-science-capstone/labs-jupyter-spacex-Data wrangling.ipynb at master · hsnapps/python-applied-data-science-capstone (github.com)]



Normalizing Data

Calculating the number of launches on each site
Calculating the number and occurrence of each orbit
Calculating the number and occurence of mission outcome of the orbits

Classification

Creating a landing outcome label from Outcome column

EDA with Data Visualization

- Plotted Charts:
  1. Scatter chart to indicate the continuous launch attempts between *FlightNumber* and *Payload*.
  2. Scatter chart to visualize the relationship between *FlightNumber* and *LaunchSite*.
  3. Scatter chart to visualize the relationship between *FlightNumber* and PayloadMass.
  4. Bar chart to visualize the relationship between success rate of each orbit type.
  5. Scatter chart to visualize the relationship between FlightNumber and Orbit type.
  6. Line chart to visualize the launch success yearly trend.
- GitHub Repository URL:

  python-applied-data-science-capstone/jupyter-labs-eda-dataviz.ipynb at master · hsnapps/python-applied-data-science-capstone (github.com)

EDA with SQL

- SQL Queries:
  - Creating SPACEXTABLE table to remove blank rows from table:
    ```
    CREATE TABLE SPACEXTABLE AS SELECT * FROM SPACEXTBL WHERE "DATE" IS NOT NULL
    ```
  - Display the names of the unique launch sites in the space mission:
    ```
    SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE
    ```
  - Display 5 records where launch sites begin with the string 'CCA':
    ```
    SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5
    ```
  - Display the total payload mass carried by boosters launched by NASA (CRS):
    ```
    SELECT SUM(PAYLOAD_MASS__KG_) AS SUM_PAYLOAD_MASS_NASA_CRS FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)'
    ```
  - Display average payload mass carried by booster version F9 v1.1:
    ```
    SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_F9_v1_1 FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1'
    ```
  - List the date when the first successful landing outcome in ground pad was achieved:
    ```
    SELECT MIN("Date") AS FIRST_SUCCESS_GROUND FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)'
    ```

EDA with SQL

- 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:

    ```
    SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)' AND
    PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
    ```

  - List the total number of successful and failure mission outcomes:

    ```
    SELECT Mission_Outcome, COUNT(*) AS "Count" FROM SPACEXTABLE GROUP BY Mission_Outcome
    ```

  - List the names of the booster_versions which have carried the maximum payload mass. Use a subquery:

    ```
    SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ IN(SELECT MAX(PAYLOAD_MASS__KG_) FROM
    SPACEXTABLE)
    ```

  - 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:

    ```
    SELECT SUBSTR("Date", 6, 2) AS 'Month', SUBSTR("Date", 0, 5) AS 'Year', Landing_Outcome FROM SPACEXTABLE WHERE
    Landing_Outcome LIKE 'Failure%' AND "Year" = '2015'
    ```

EDA with SQL

- SQL Queries:

  ○ 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:

    ```
    SELECT Landing_Outcome, COUNT(*) AS cnt FROM SPACEXTABLE WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY cnt DESC
    ```

- GitHub Repository URL:

  - python-applied-data-science-capstone/jupyter-labs-eda-sql-coursera_sqllite.ipynb at master · hsnapps/python-applied-data-science-capstone (github.com)

Build an Interactive Map with Folium

- Map Objects Used:
  - Markers (folium.map.Marker) on all launch sites: *To show the locations of launch stations on a map.*
  - Highlighting circles (folium.Circle): *To highlight launch locations.*
  - Markers on the success/failed launches for each site on the map: *To distinguish the success/failure launch sites.*
  - Distance lines (folium.PolyLine) between a launch site to its proximities: *To display the distance between the launch site and the nearest milestones.*
- GitHub Repository URL:
  - python-applied-data-science-capstone/lab_jupyter_launch_site_location.ipynb at master · hsnapps/python-applied-data-science-capstone (github.com)

Build a Dashboard with Plotly Dash

- Graphs and plots used in the Dashboard:
    - Pie chart interacts with dropdown component:
        - The dropdown contains the selection list of "ALL" followed by launching sites.
        - When the user selects "ALL" the pie charts displays the total attempts of each site.
        - When the user selects a site name the pie chart displays the success/failure attempts of the site.
    - Scatter chart displays the correlation between payload mass and launching attempts, interacts with range-slide component and dropdown component:
        - The user can specify the "*Payload Range (Kg)*" ranges from the range-slide.
        - If the dropdown value is "ALL", the chart displays correlation for all sites, otherwise it displays the correlation for the selects site.
- GitHub Repository URL:
    - python-applied-data-science-capstone/spacex_dash_app.py at master · hsnapps/python-applied-data-science-capstone (github.com)

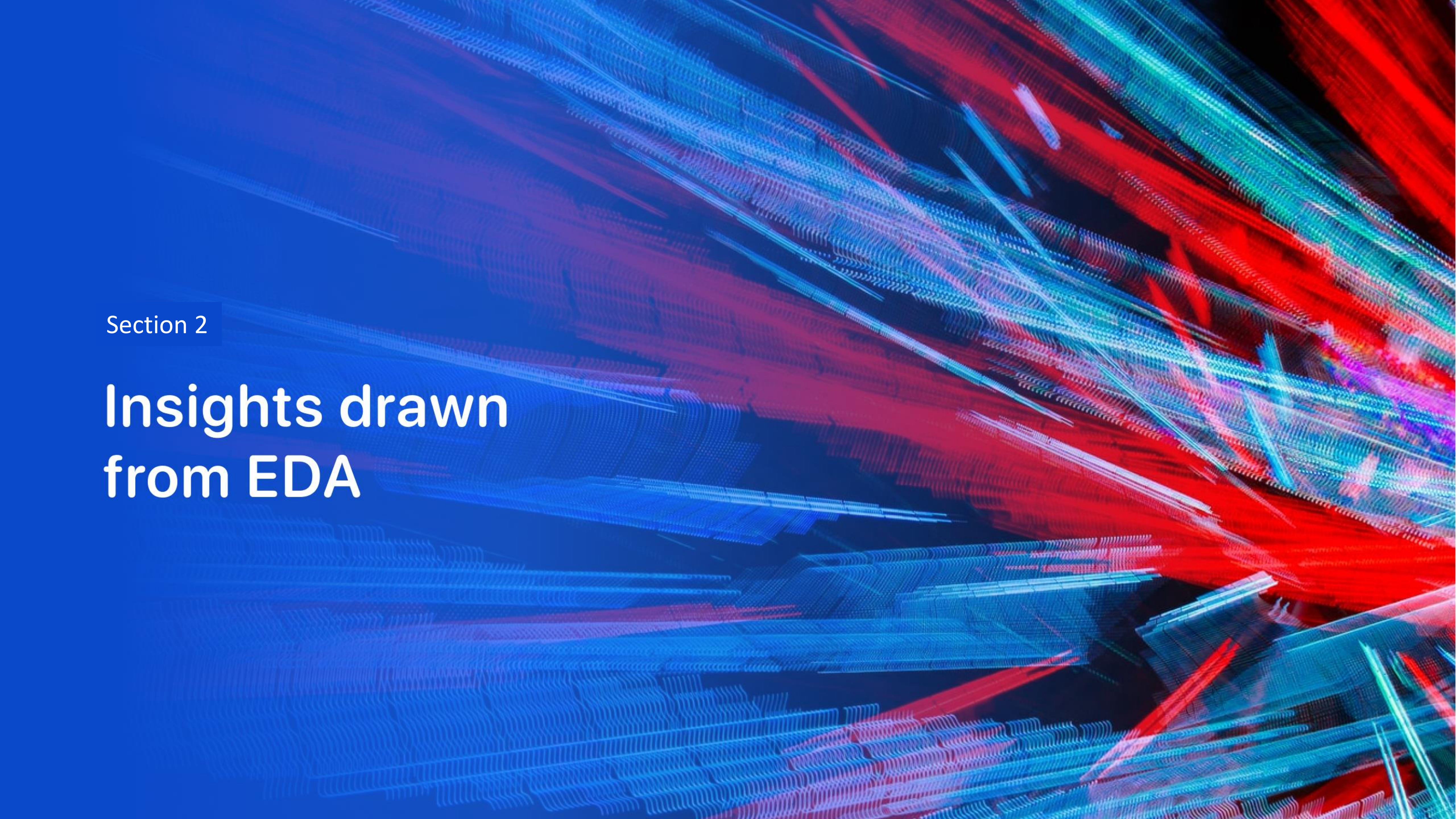
Predictive Analysis (Classification)

- Normalize the data for X and Y by using python libraries such as NumPy and scikit-learn.
- Split the training and testing samples using train_test_split from scikit-learn.
- Creating a logistic regression and GridSearchCV to perform hyperparameter tuning then calculating the accuracy.
- Creating SVM object and GridSearchCV object to find the best parameters from the dictionary parameters then calculating the accuracy.
- Creating a decision tree classifier object and GridSearchCV object to to find the best parameters from the dictionary parameters then calculating the accuracy.
- Create a KNN object and GridSearchCV object to find the best parameters from the dictionary parameters then calculating the accuracy.
- Finding the method performs best

Results

- Exploratory data analysis results:
  - Different launch sites have different success rates.
  - Success rate since 2013 has improved.
  - CCAFS LC-40 has a success rate of 60%, while KSC LC-39A and VAFB SLC 4E have a success rate of around 77%.
  - If the mass is above 10,000 kg the success rate of CCAFS LC-40 is 100%.
- Predictive analysis results (✷ The best accuracy):

Model	Accuracy
Logistic Regression	0.8464285714285713
Support Vector Machine	0.8732142857142856
K-Nearest Neighbors	0.8482142857142858
Decision Tree Accuracy ✷	0.8482142857142856

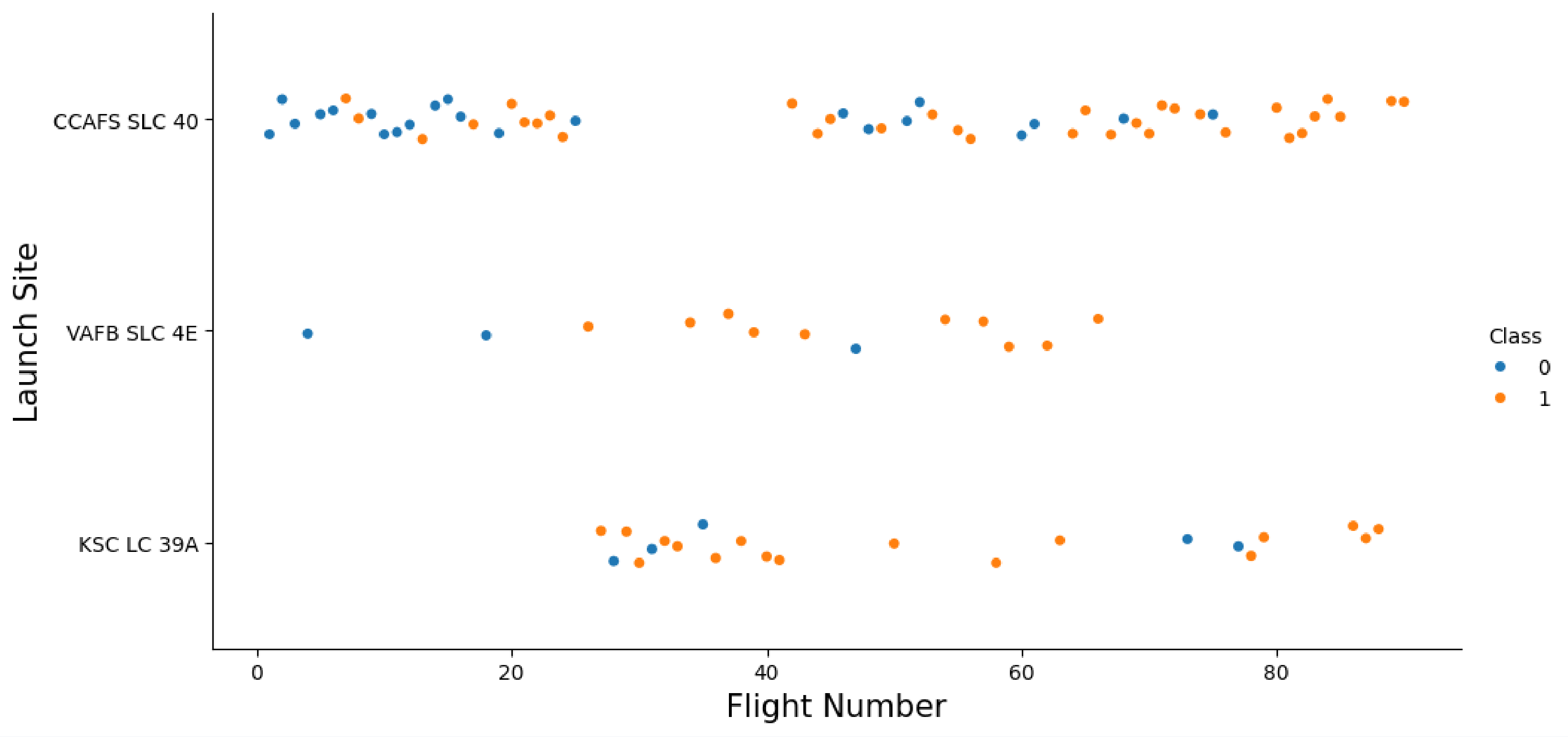


Section 2

Insights drawn from EDA

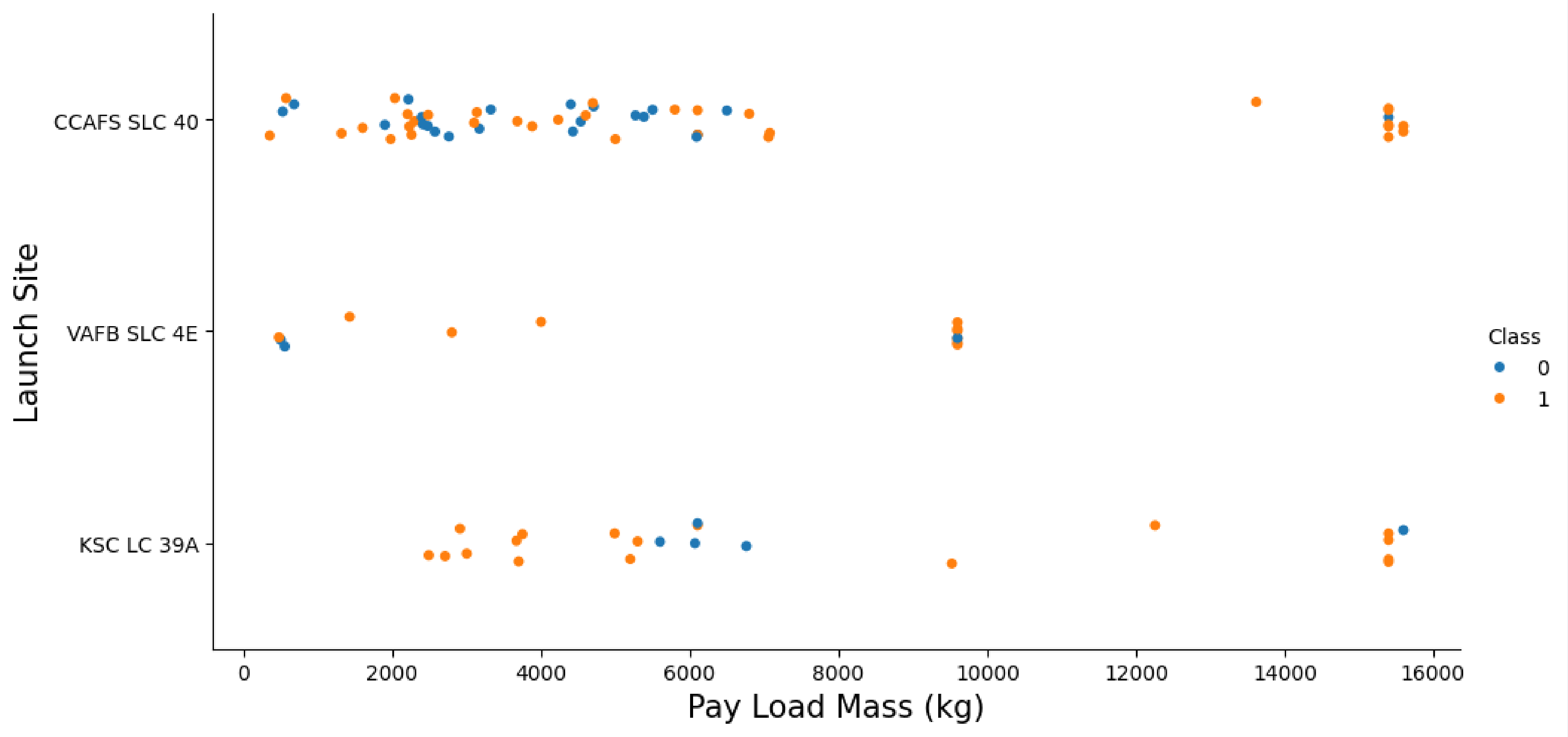
Flight Number vs. Launch Site

- Success rate increases in the late flights.



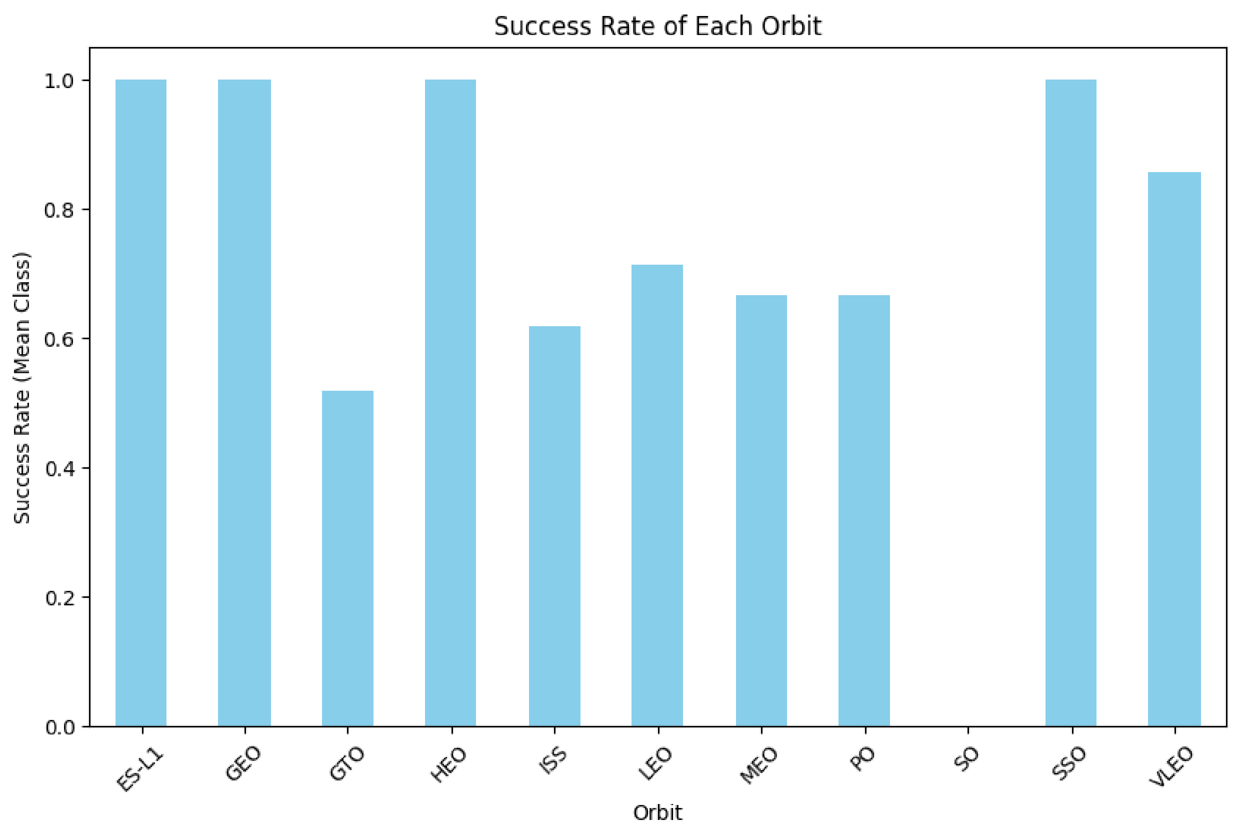
Payload vs. Launch Site

- If the mass is above 10,000 kg the success rate of CCAFS LC-40 is 100%.



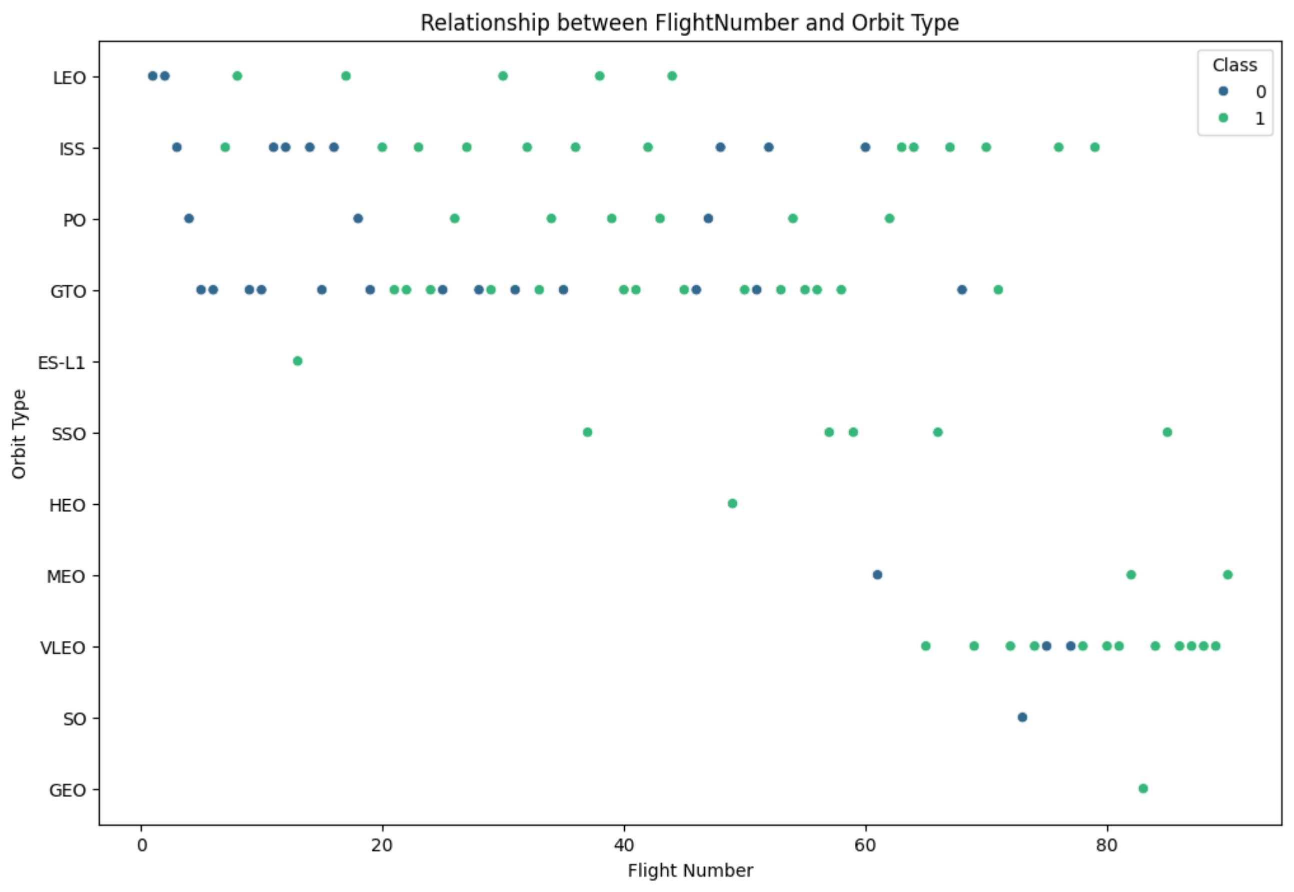
Success Rate vs. Orbit Type

- Orbits ES-L1, GEO, HEO and SSO have the best success rate.



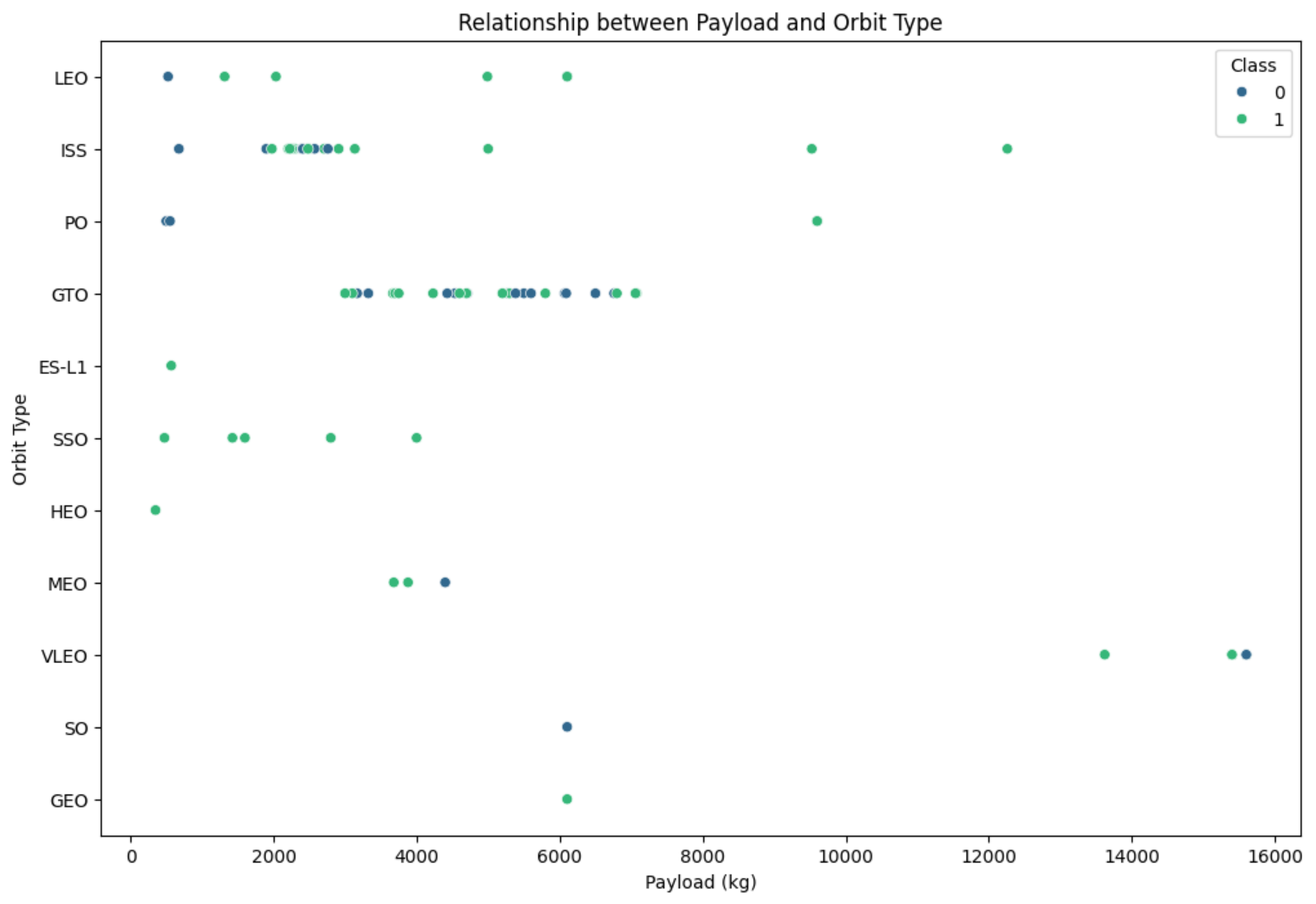
Flight Number vs. Orbit Type

- The LEO orbit the Success appears related to the number of flights.
- No clear relationship between flight number when in GTO orbit.



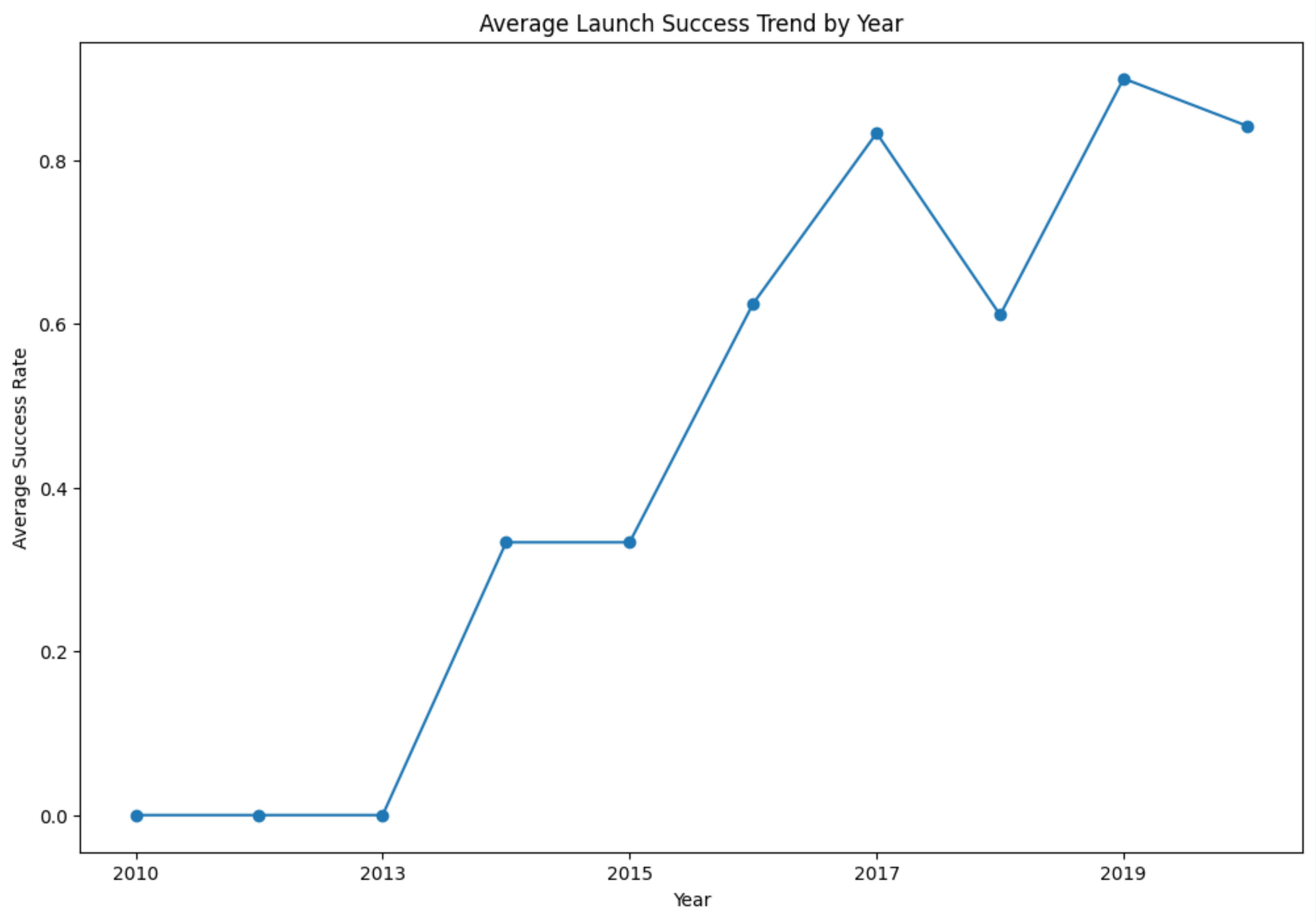
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- We cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.



Launch Success Yearly Trend

- The success rate since 2013 kept increasing till 2017 (stable in 2014) and after 2015 it started increasing.



All Launch Site Names

- Query

```
SELECT DISTINCT Launch_Site FROM
SPACEXTABLE
```

- Explanation

Displays the names of the unique launch sites in the space mission

- Result

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

Launch Site Names Begin with 'CCA'

- Query

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

- Result

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

- Explanation

Displays 5 records where launch sites begin with the string 'CCA'

Total Payload Mass

- Query

```
SELECT SUM(PAYLOAD_MASS__KG_) AS SUM_PAYLOAD_MASS_NASA_CRS FROM SPACEXTABLE
WHERE Customer = 'NASA (CRS)'
```

- Result

SUM_PAYLOAD_MASS_NASA_CRS
45596

- Explanation

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

Average Payload Mass by F9 v1.1

- Query

```
SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_F9_v1_1 FROM SPACEXTABLE WHERE Booster_Version =
'F9 v1.1'
```

- Result

AVG_F9_v1_1
2928.4

- Explanation

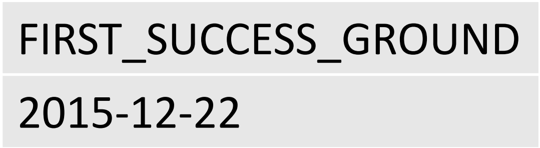
Displays average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date

- Query

```
SELECT MIN("Date") AS FIRST_SUCCESS_GROUND FROM SPACEXTABLE WHERE Landing_Outcome =
'Success (ground pad)'
```

- Result



FIRST_SUCCESS_GROUND
2015-12-22

- Explanation

Lists the date when the first successful landing outcome in ground pad was acheived.

Successful Drone Ship Landing with Payload between 4000 and 6000

- Query

```
SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone
ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
```

- Result

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

- Explanation

Lists the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.

Total Number of Successful and Failure Mission Outcomes

- Query

```
SELECT Mission_Outcome, COUNT(*) AS "Count" FROM SPACEXTABLE GROUP BY Mission_Outcome
```

- Result

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

- Explanation

Lists the total number of successful and failure mission outcomes.

Boosters Carried Maximum Payload

- Query

```
SELECT DISTINCT Booster_Version FROM
SPACEXTABLE WHERE
PAYLOAD_MASS__KG_ IN(SELECT
MAX(PAYLOAD_MASS__KG_) FROM
SPACEXTABLE)
```

- Explanation

Lists the names of the booster_versions which have carried the maximum payload mass. Use a subquery.

- Result

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

- Query

```
SELECT SUBSTR("Date", 6, 2) AS 'Month', SUBSTR("Date", 0, 5) AS 'Year', Landing_Outcome FROM
SPACEXTABLE WHERE Landing_Outcome LIKE 'Failure%' AND "Year" = '2015'
```

- Result

Month	Year	Landing_Outcome
01	2015	Failure (drone ship)
04	2015	Failure (drone ship)

- Explanation

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

- Query

```
SELECT Landing_Outcome, COUNT(*) AS
cnt FROM SPACEXTABLE WHERE "Date"
BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome ORDER BY
cnt DESC
```

- Explanation

Ranking 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.

- Result

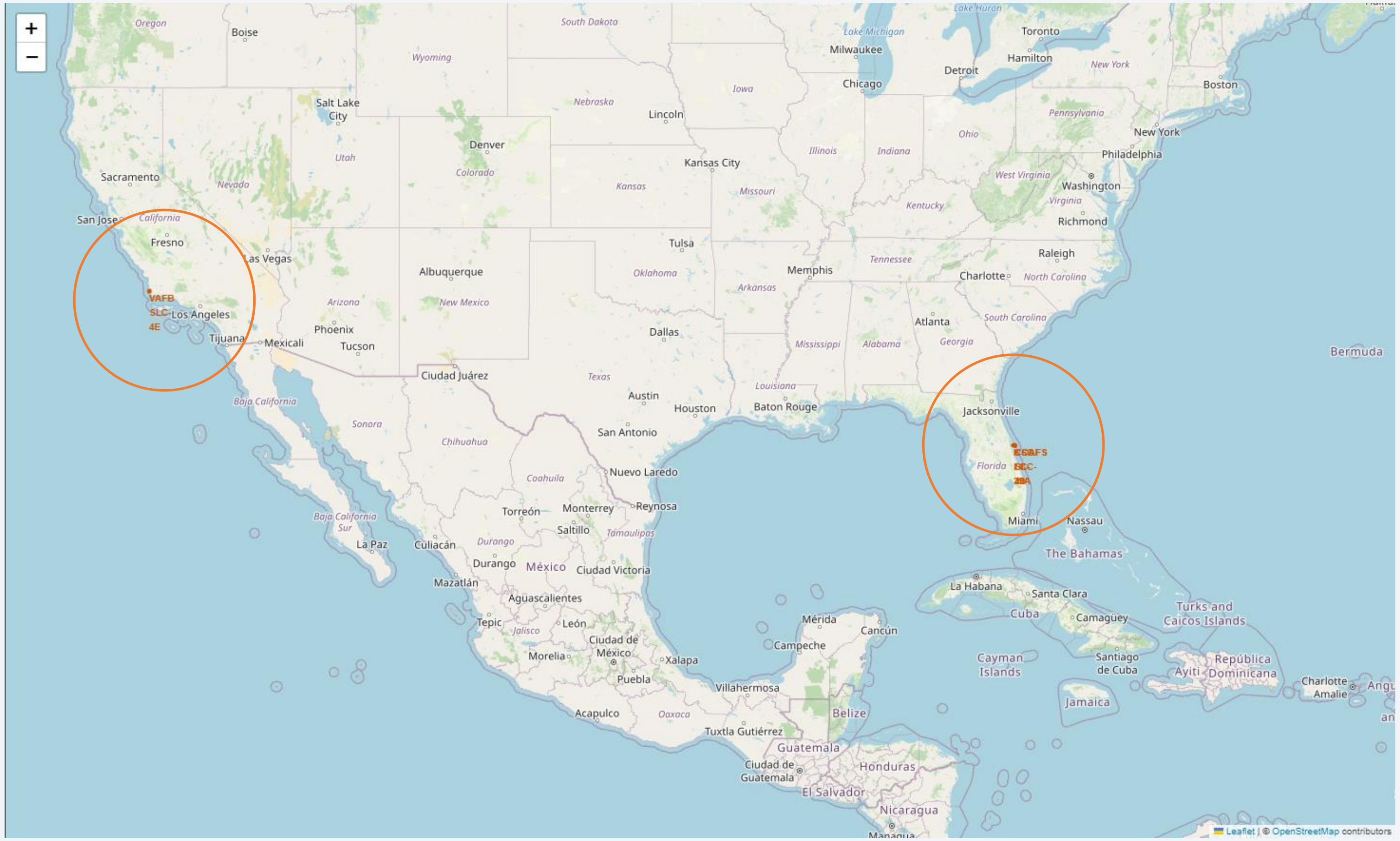
Landing_Outcome	Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

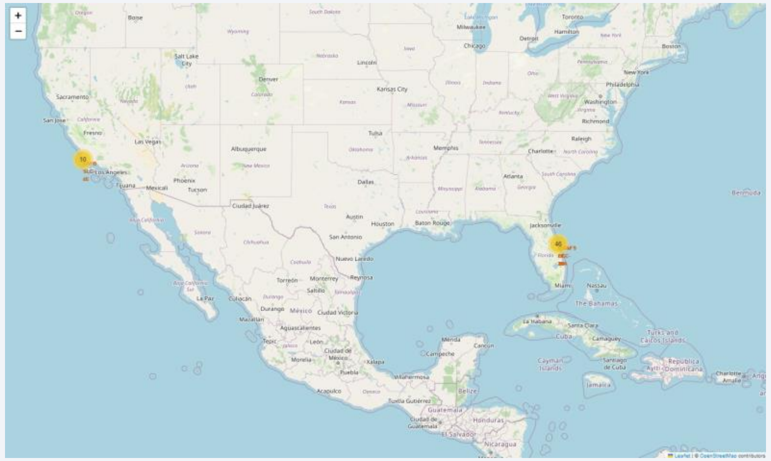
Launch Sites Location

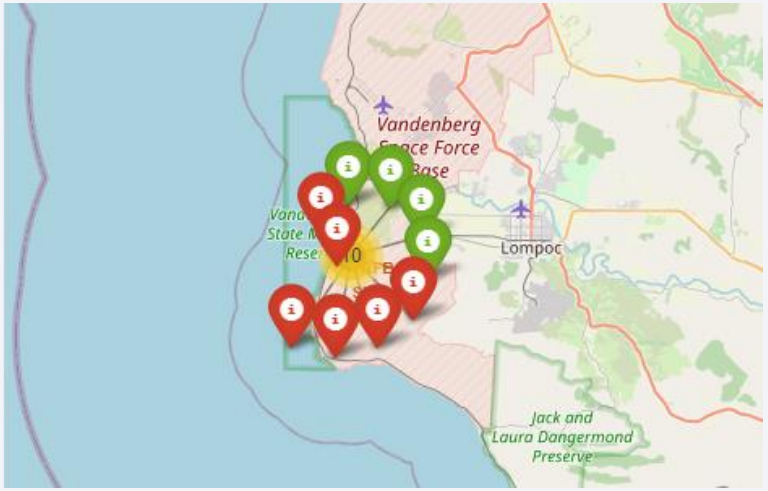
- In this map we can locate launch sites.
- We notice that all sites are close to the coasts.
- All the sites located between the latitudes 28 and 35, so they are not so close to the equator.



Success/Failed Launches For Each Site On The Map

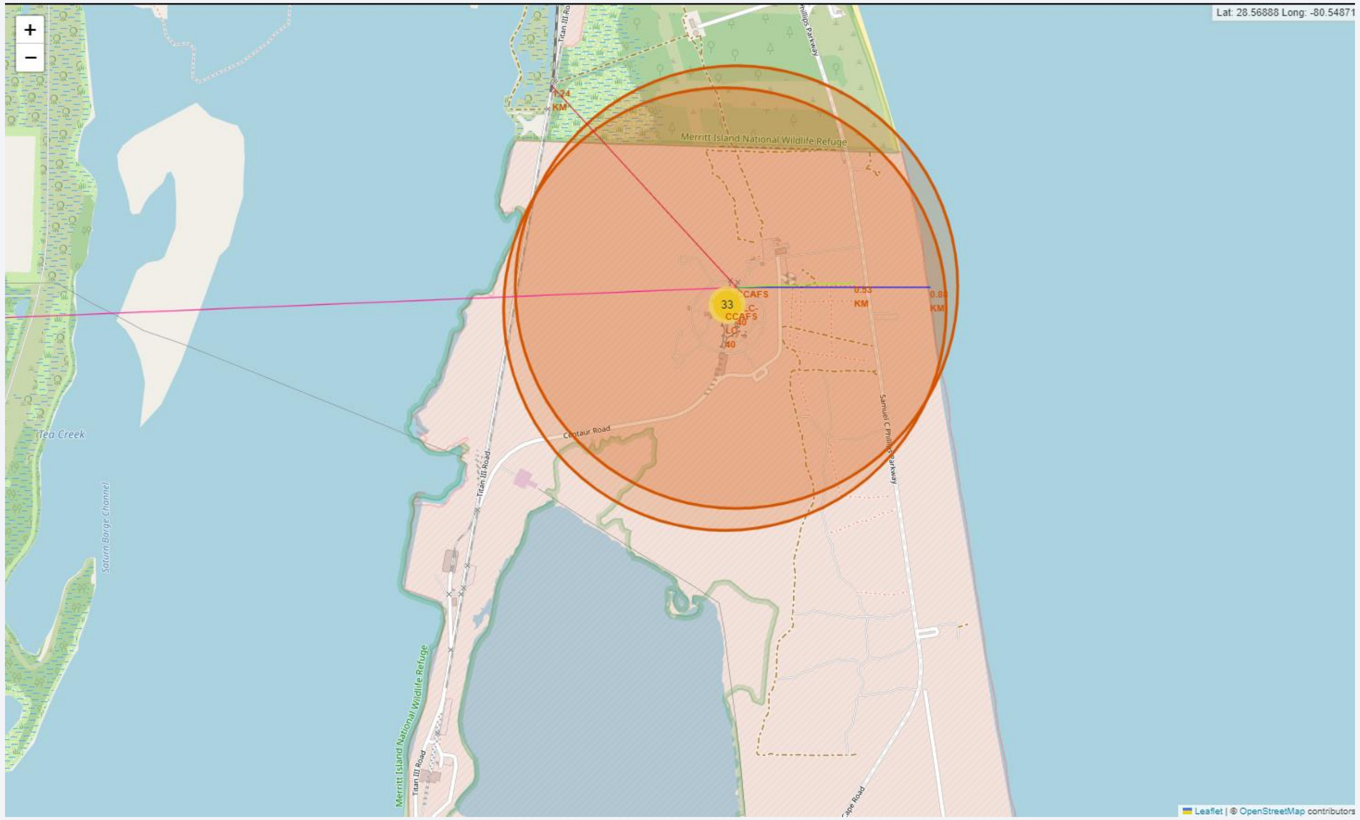
- In this interactive map we can see the locations of launch sites clickable and displays the Success/Failed launches of each site.



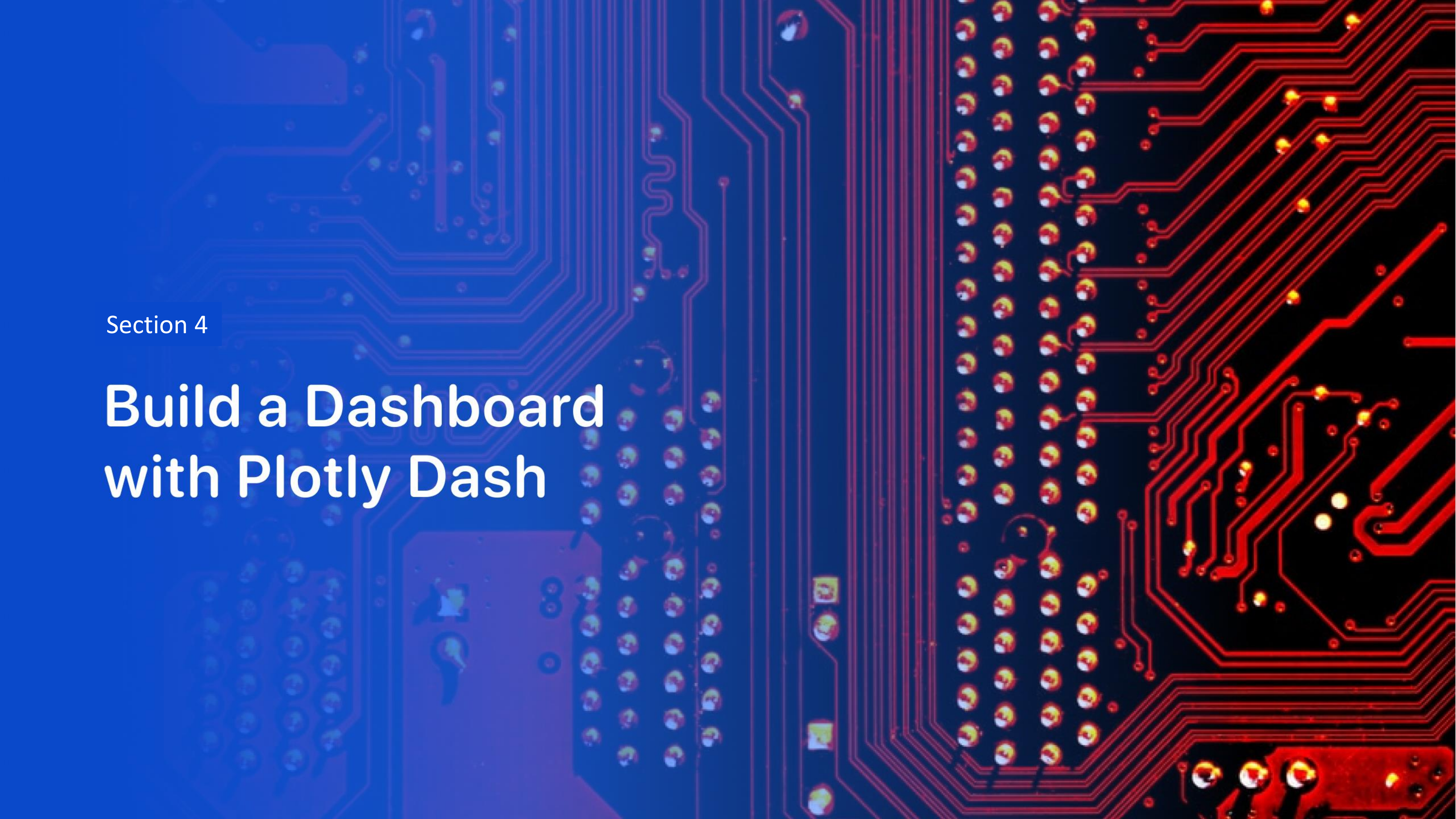


Distances Between A Launch Site And Proximities

- In this map we can see the distance between a sample launch site (CCAFS SLC-40) and some proximities.
- Although the site is quite close to coastline and railway, however, the distance between it and the nearest city is considered appropriate.



Lat: 28.56888 Long: -80.54871
+
−
1.24 KM
Merritt Island National Wildlife Refuge
33
CCAFS
LC-
40
CCAFS
LC-
40
0.53 KM
KM
Titan III Road
Centaur Road
Samuel C Phillips Parkway
Tea Creek
Saturn Barge Channel
Merritt Island National Wildlife Refuge
Leaflet | © OpenStreetMap contributors

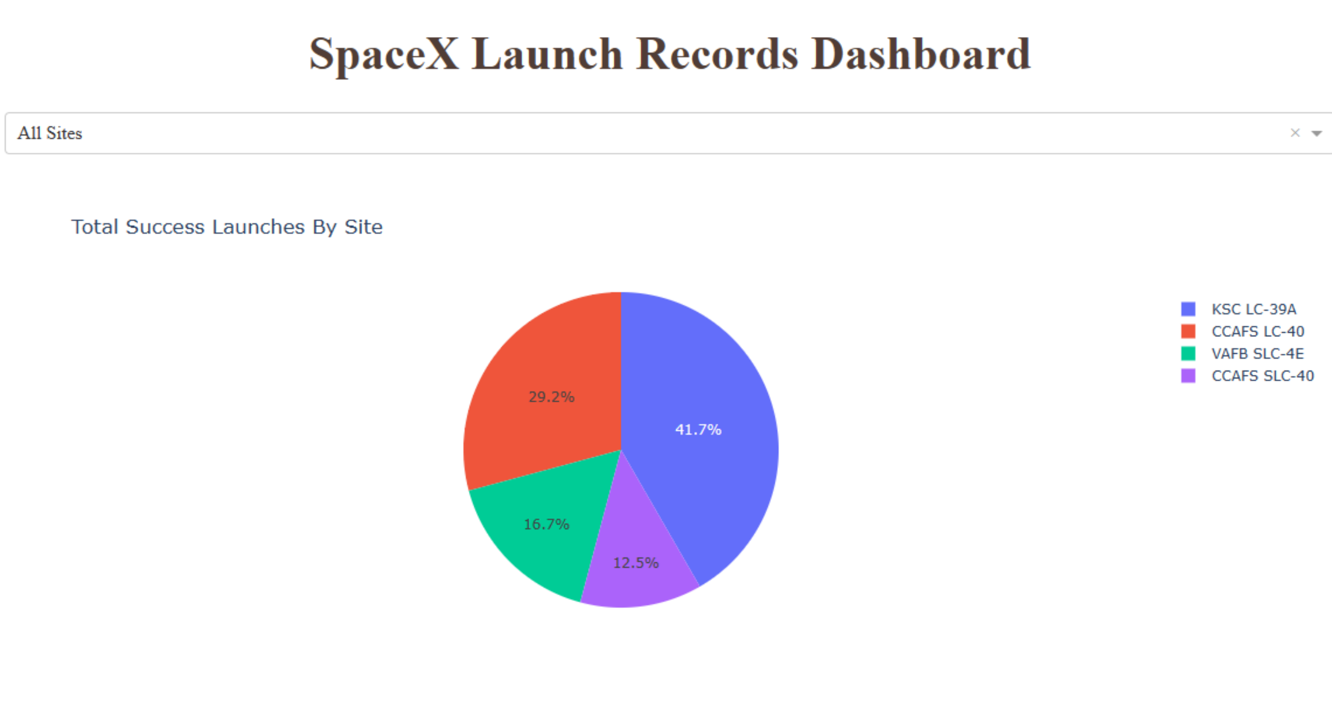


Section 4

Build a Dashboard with Plotly Dash

Launch Sites Interactive Dashboard

- In the interactive dashboard, we can see a **Pie** chart displays total launches of each site.
- The chart legend displays site’s name and its color on the chart.
- The dropdown above the pie chart can be used to select a specific launch site, then the chart changes to display the success launces of the site.

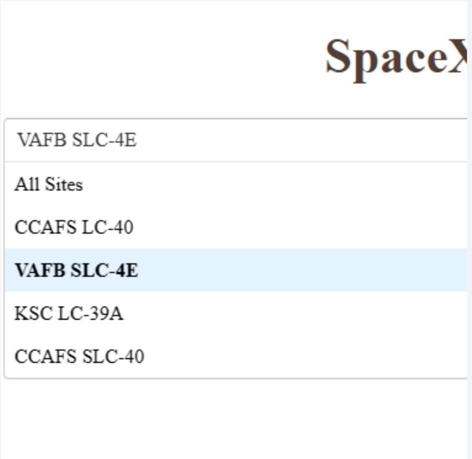


SpaceX Launch Records Dashboard

All Sites

Launch Sites Interactive Dashboard

- As seen on the screenshot, when we select a specific launch site the chart display the success launces of the site.



SpaceX

VAFB SLC-4E

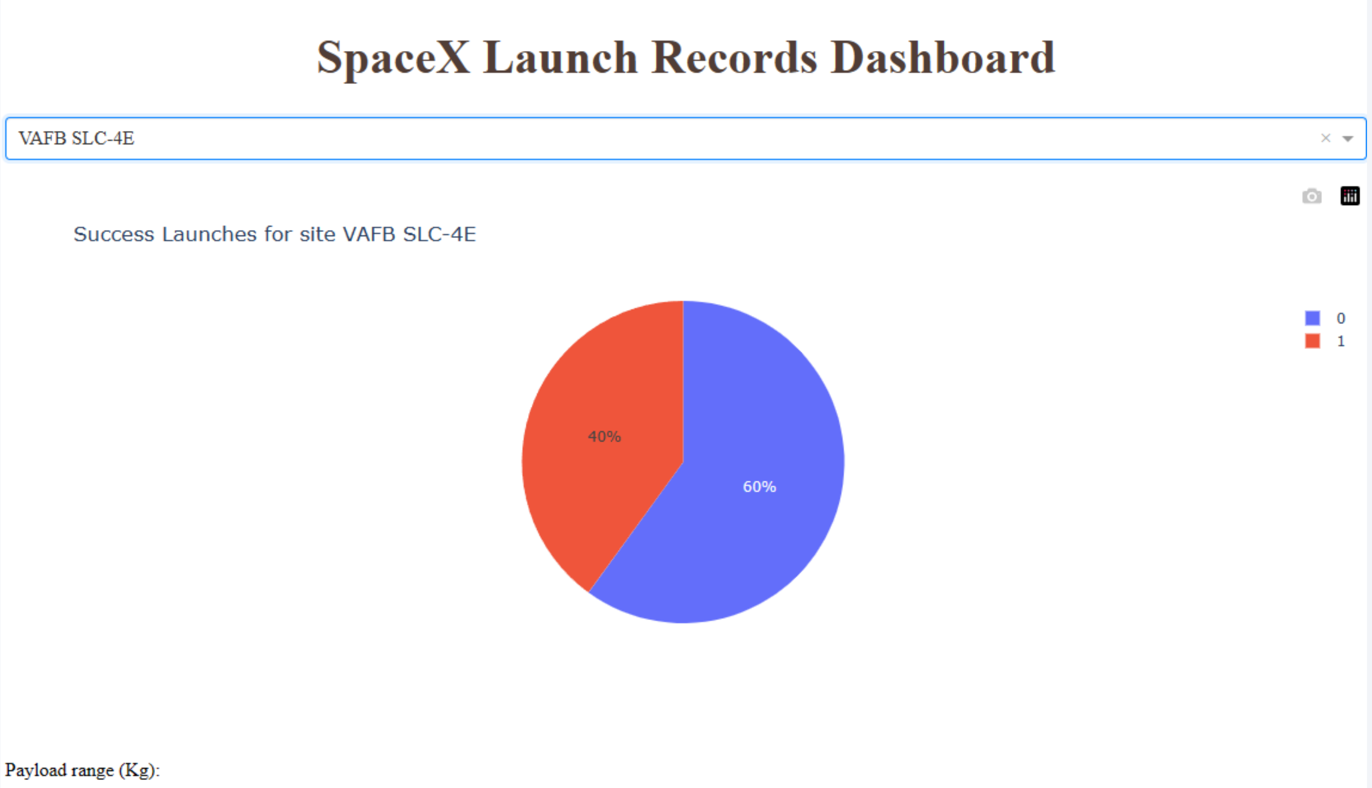
All Sites

CCAFS LC-40

VAFB SLC-4E

KSC LC-39A

CCAFS SLC-40



SpaceX Launch Records Dashboard

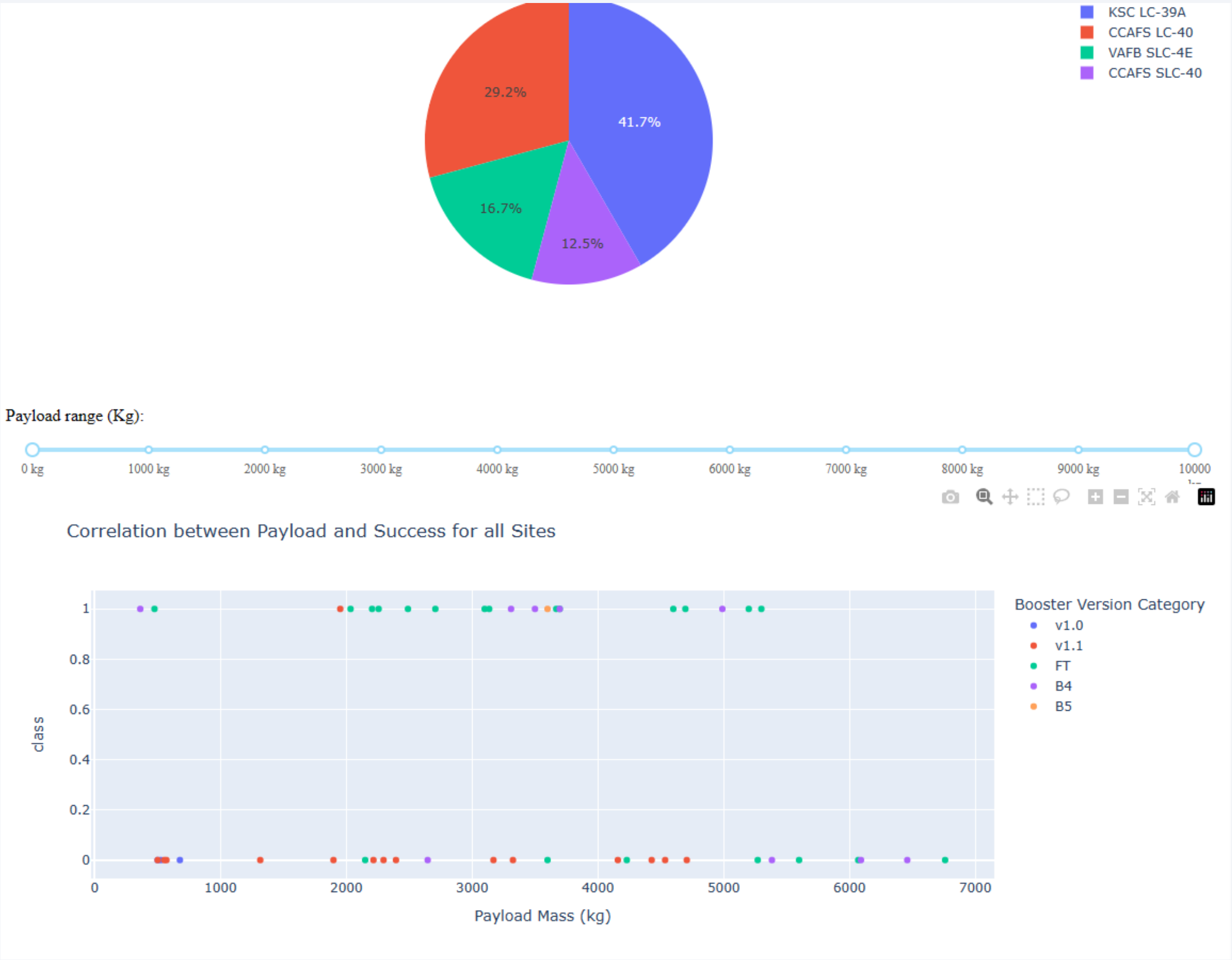
VAFB SLC-4E

Success Launches for site VAFB SLC-4E

Payload range (Kg):

Launch Sites Interactive Dashboard

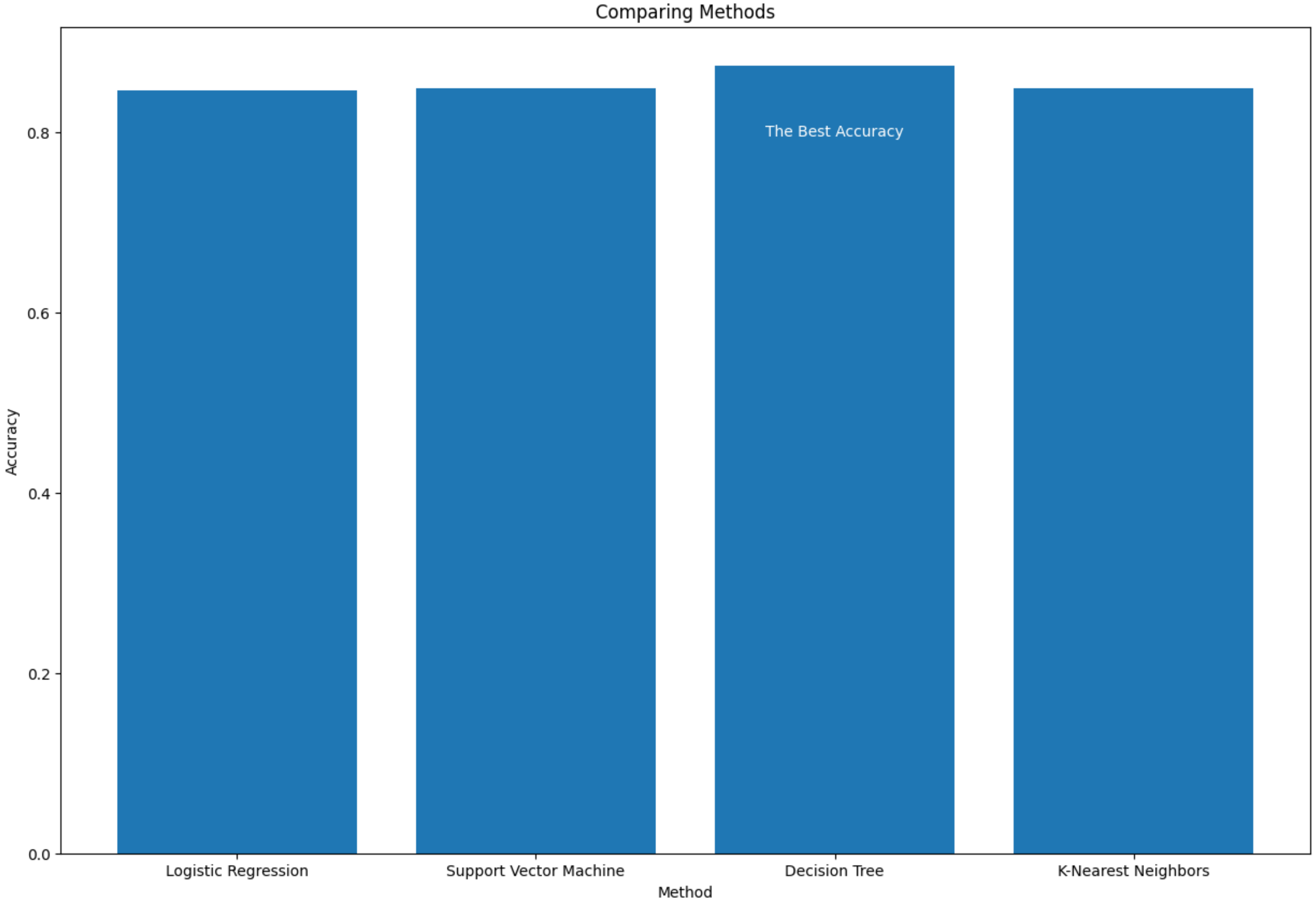
- Below the pie chart, we can see a scatter chart that interacts with the above dropdown and a range-slide component represents the payload range in kilogram.



Section 5

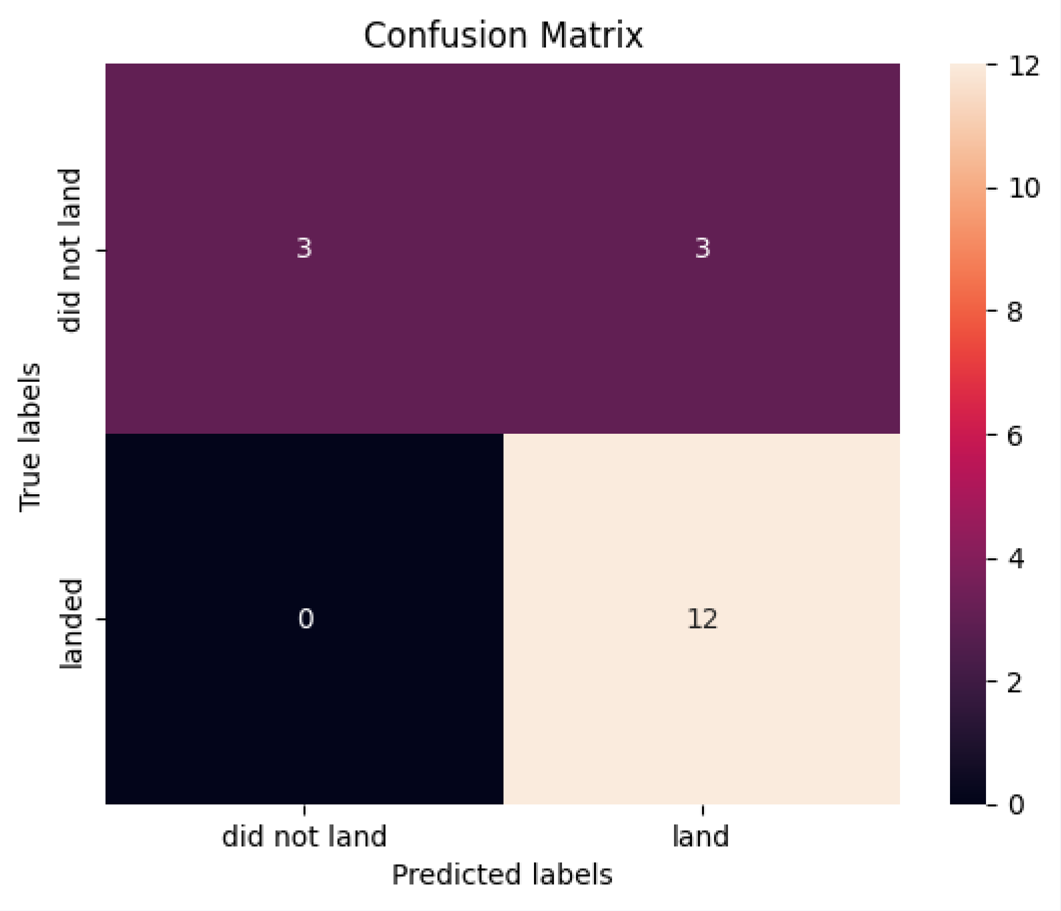
Predictive Analysis (Classification)

Classification Accuracy



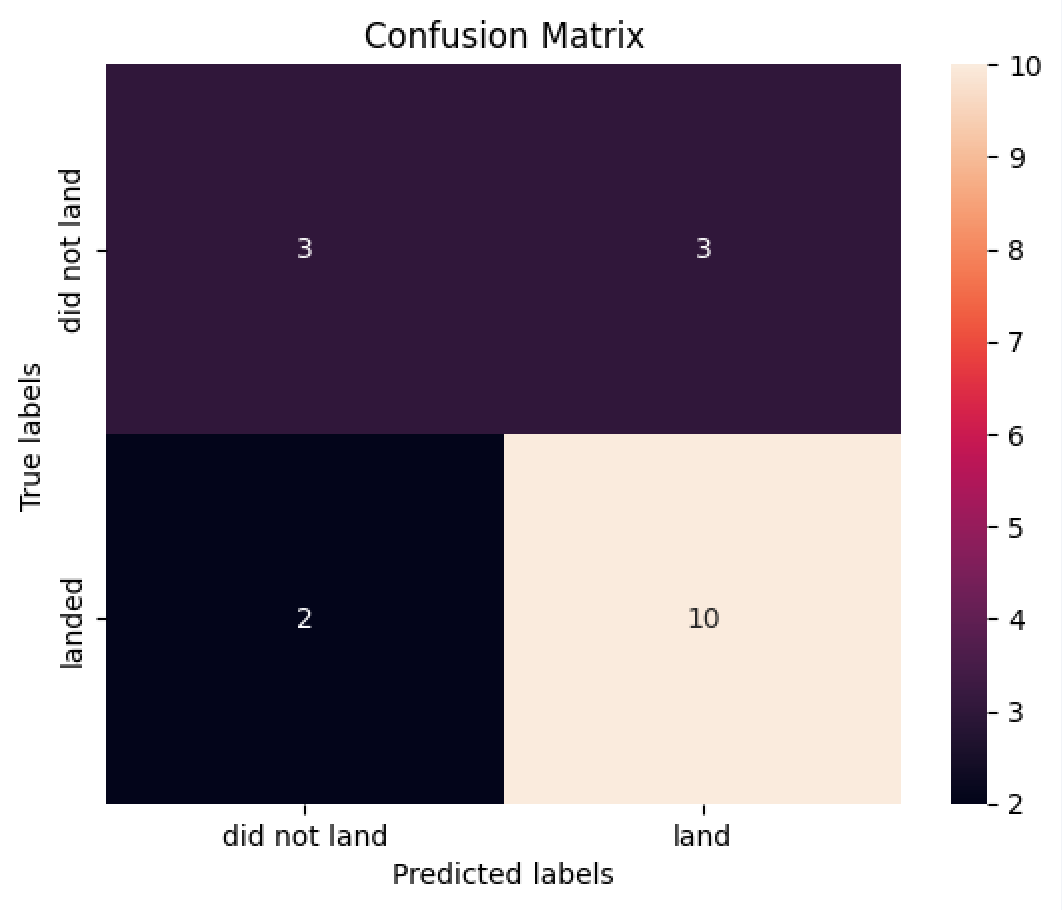
Confusion Matrix

- The Confusion Matrix of the models *Logistic Regression, Support Vector Machine (SVM)* and *K-Nearest Neighbors (KNN)*.
- The figure shows similarity of accuracy for the three models.



Confusion Matrix

- The Confusion Matrix of the ***Decision Tree*** model.
- The Decision Tree models gives us the best accuracy in prediction.



Conclusions

- At the end of this study, we can say that the price of launch is predictable.
- The most important parameters to predict are *Payload Mass* , *Orbit* and *Launch Site*.
- The *Tree Decision* model is the best to use for predicting the price of launch with accuracy % 84.82.
- There is a positive correlation between the payload mass and the success rate.
- The best orbits for the rocket are: SSO, VLEO, HEO, GEO and LEO.

Appendix

- Datasets used in the project:
  1. SpaceX public API: https://api.spacexdata.com/v4/
  2. List of Falcon 9 and Falcon Heavy launches on Wikipedia: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

- Project repository on github.com: https://github.com/hsnapps/python-applied-data-science-capstone

Thank you!