IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<ZHEJUN HUANG>
<18/06/2025>

Outline

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

Executive Summary

- Summary of methodologies

  Data Collection;

  Data Wrangling;

  Perform EDA with Data Visualization and SQL;

  Perform interactive visual analytics using Folium and Plotly Dash;

  Perform predictive analysis using classification models

- Summary of all results

  Key Factors Influencing SpaceX First Stage Landing Success: **Flight number, booster version, launch site, payload mass, and orbit type**

Introduction

- ## Project background and context

- 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. In this lab, you will collect and make sure the data is in the correct format from an API. The following is an example of a successful and launch. This study is to investigate the factors affecting the success of a launch and predict if the Falcon 9 first stage will land successfully.

- ## Problems you want to find answers

What affects the success of a launch?

How to predict if the Falcon 9 first stage will land successfully based on historical data?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- 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 Collection – SpaceX API

  The dataset is collected from the SpaceX API using requests.

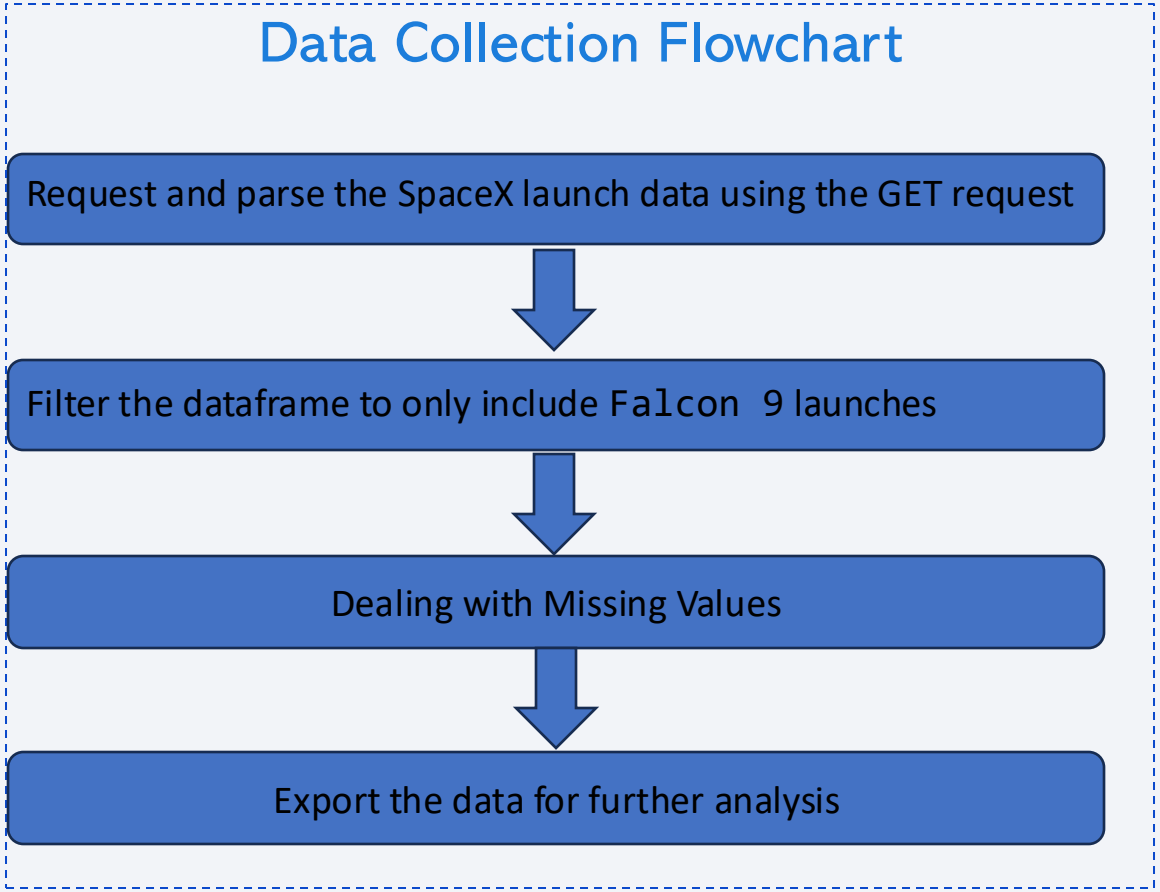
- ## Data Collection - Scraping

  Request the Falcon9 Launch Wiki page from its URL using web scraping.

Data Collection – SpaceX API

- Data Collection Flowchart is shown on the right

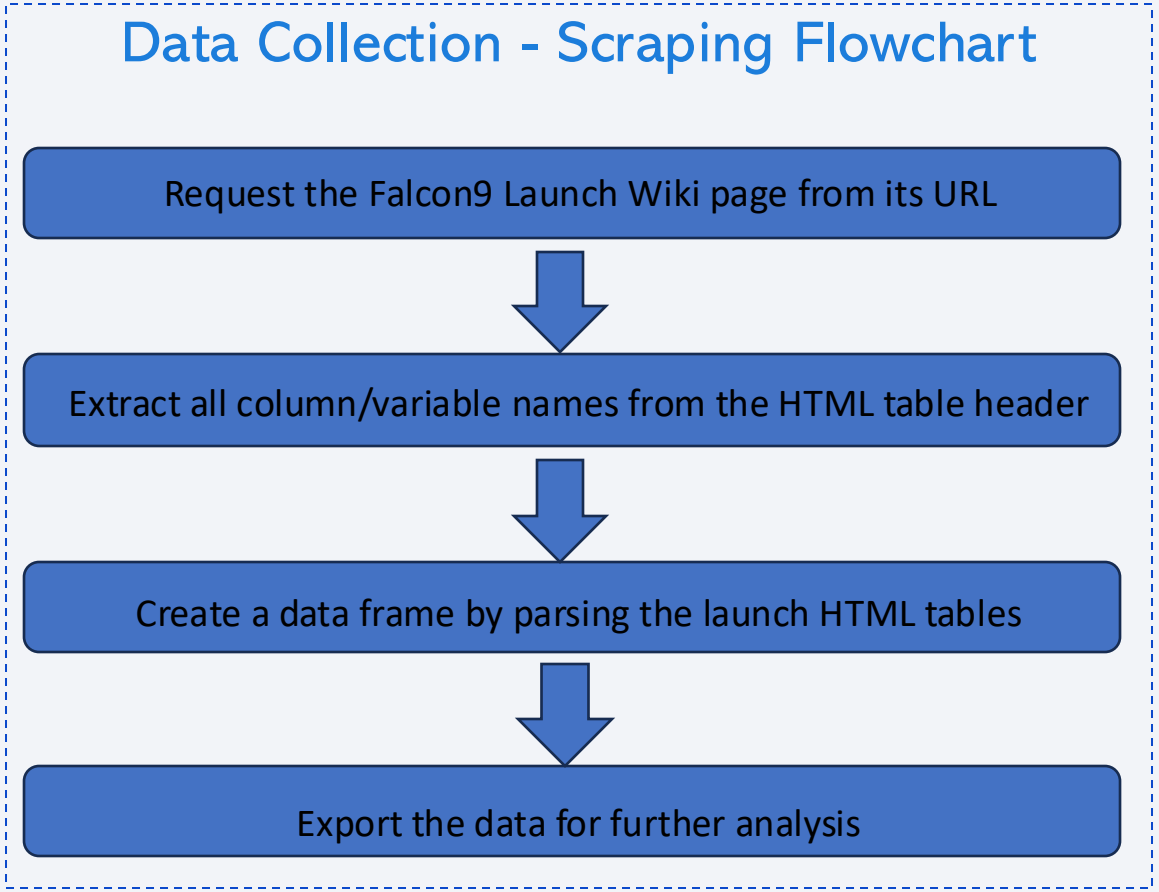
- The GitHub URL of the completed SpaceX API calls notebook is https://github.com/huangzj2025/DataSciencCapstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- The web scraping process using key phrases and flowcharts is shown on the right

- The GitHub URL of the completed web scraping notebook is https://github.com/huangzj2025/DataSciencCapstone/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

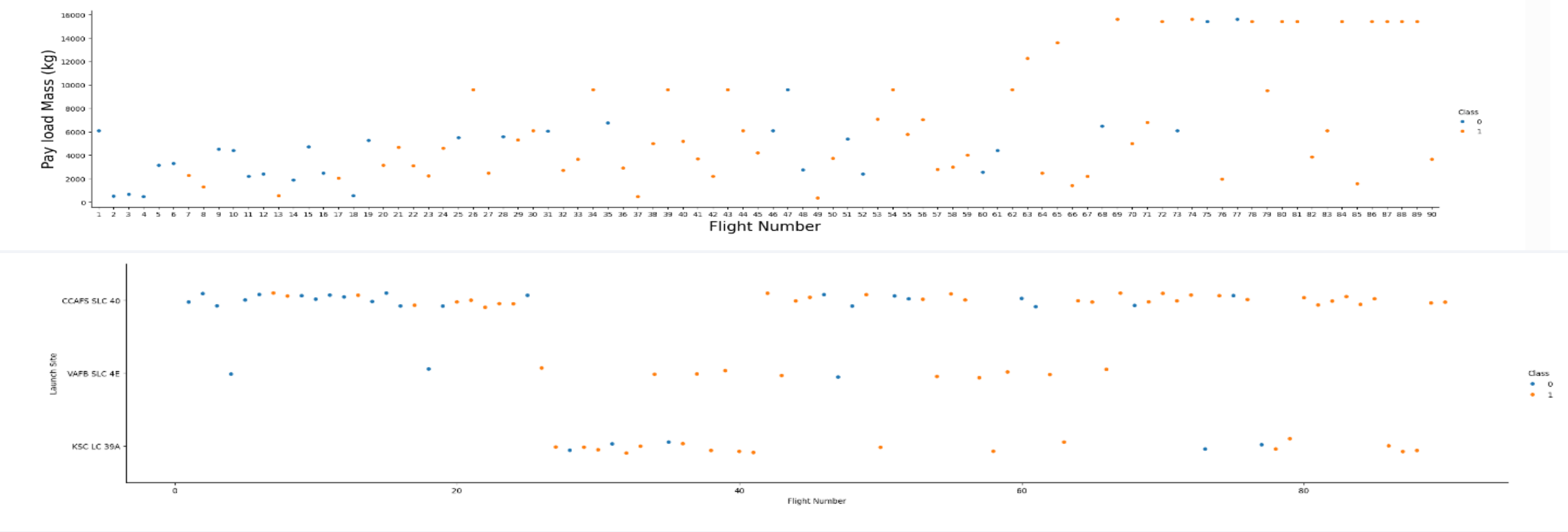
➢ Data Wrangling Process

- Import Libraries and Define Auxiliary Functions
- Calculate the number of launches on each site
- Calculate the number and occurrence of each orbit
- Calculate the number and occurence of mission outcome of the orbits
- Create a landing outcome label from Outcome column
- Export the data for further analysis

➢ https://github.com/huangzj2025/DataSciencCapstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- SpaceX Significantly Improves Launch Reliability and Payload Capacity



- https://github.com/huangzj2025/DataSciencCapstone/blob/main/edadataviz.ipynb

EDA with SQL

➢Summarize the SQL queries

- Connect to the database
- 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)
- 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 all the booster_versions that 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/huangzj2025/DataSciencCapstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Summarize
- Create a **folium Map** object markers to show the location of the NASA Johnson Space Center at Houston, Texas
- Add **folium.Circle** to add a highlighted circle area with a text label on launch sites
- Use **MarkerCluster** object to Mark the success/failed launches for each site on the map
- Use **PolyLine** between a launch site to the selected point to demonstrate the distance inbetween

- https://github.com/huangzj2025/DataSciencCapstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Summarize

Add a dropdown list to enable Launch Site selection to show the relationship between success and launch site.

Add a pie chart to show the total successful launches count for all sites

Add a slider to select payload range to show the effect of payload on the success of a launch

Add a scatter chart to show the correlation between payload and launch success

- https://github.com/huangzj2025/DataSciencCapstone/blob/main/spacex-dash-app.py

Predictive Analysis (Classification)

- Summarize of model development process

1) Create a column for the class storing the target object Y
2) Standardize the data characterizing features X
3) Split into training data and test data using train_test_split
4) Find best Hyperparameter for SVM, Classification Trees and Logistic Regression using GridSearchCV
5) Find the method performs best using test data

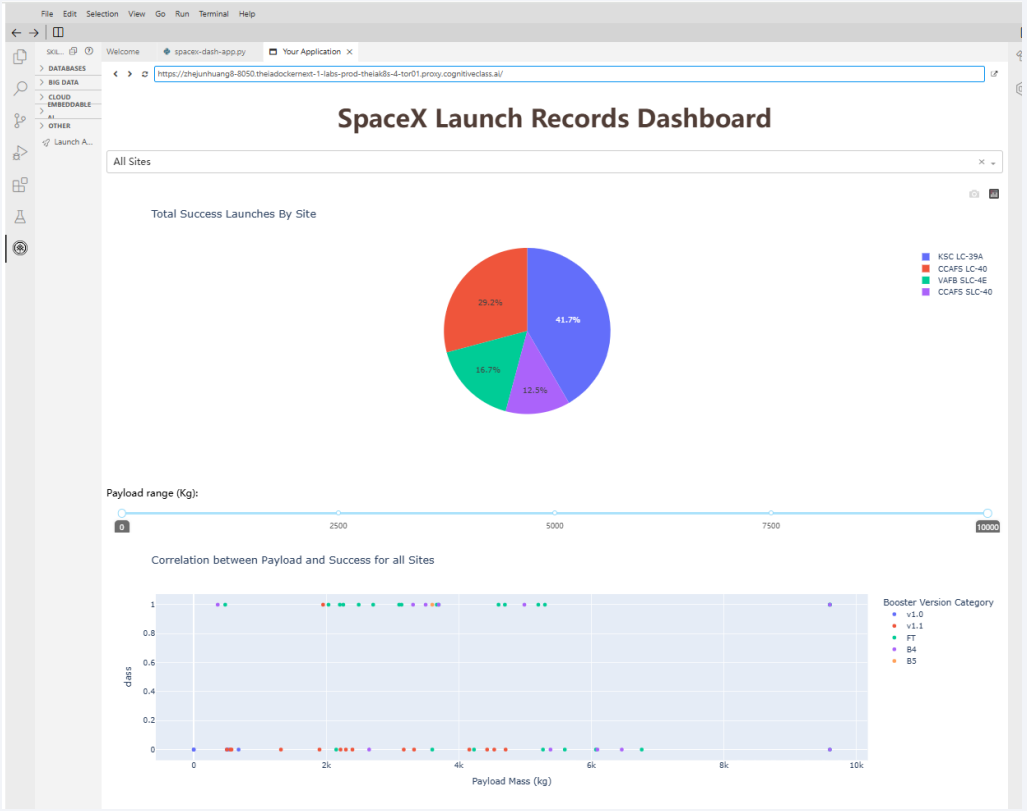
- https://github.com/huangzj2025/DataSciencCapstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

- Exploratory data analysis results:

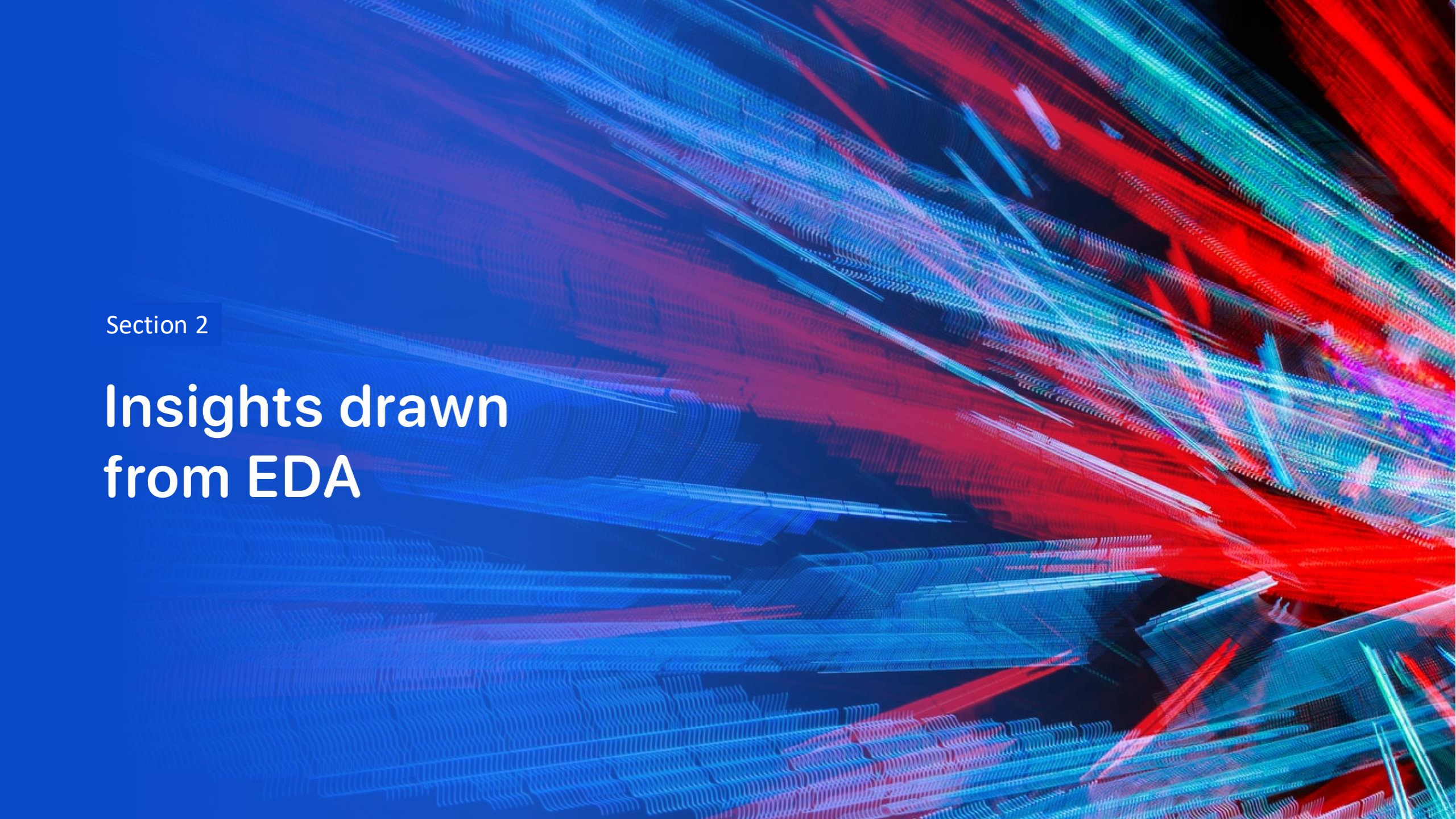
  SpaceX Significantly Improves Launch Reliability and Payload Capacity

- Interactive analytics demo in screenshots



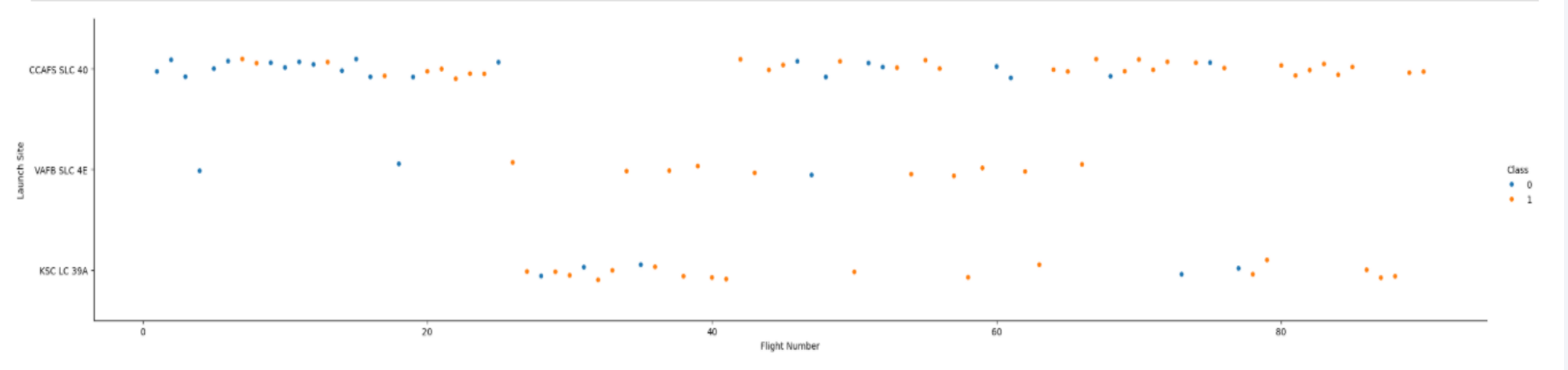
- Predictive analysis results

Meachine learning models predict that the accuracy of the rate of successful landing is over 80%, and the launch task using SpaceX is likely to succeed.


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

- Show a scatter plot of Flight Number vs. Launch Site



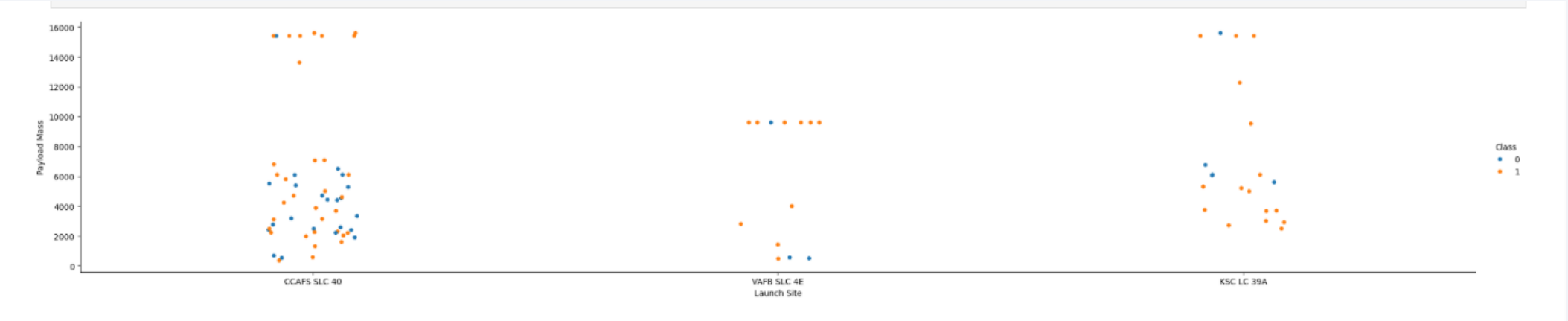
- Show the screenshot of the scatter plot with explanations

With the increase of the flight number, the success rate of all the three launch sites has improved.

VAFB SLC 4E has the highest success rate, followed by KSC LC39A and CCAFS SLC 40.

Payload vs. Launch Site

- Show a scatter plot of Payload vs. Launch Site



- Show the screenshot of the scatter plot with explanations

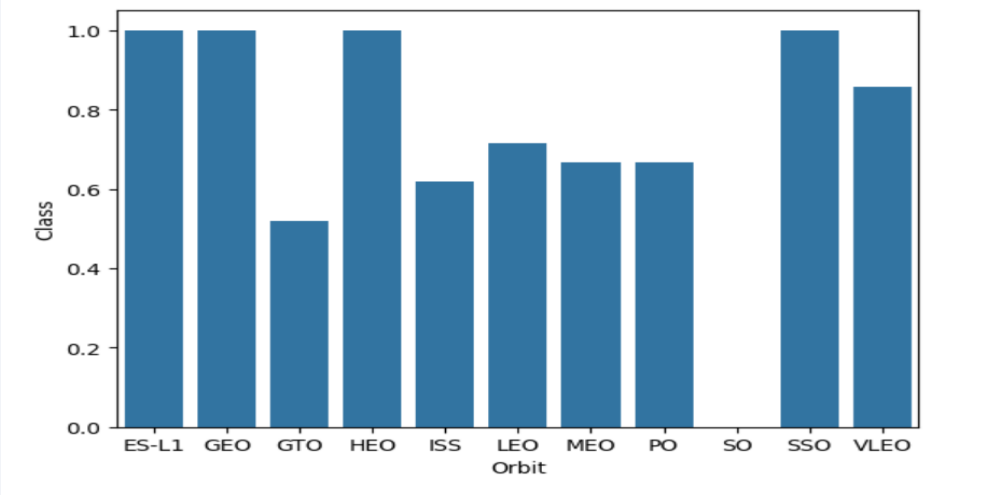
VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

CCAFS SLC 40 launchsite has no rockets launched for payload mass between 8000 and 12000.

The success rate of CCAFS SLC 40 with heavypayload mass is higher than that with low payload mass.

Success Rate vs. Orbit Type

- Show a bar chart for the success rate of each orbit type



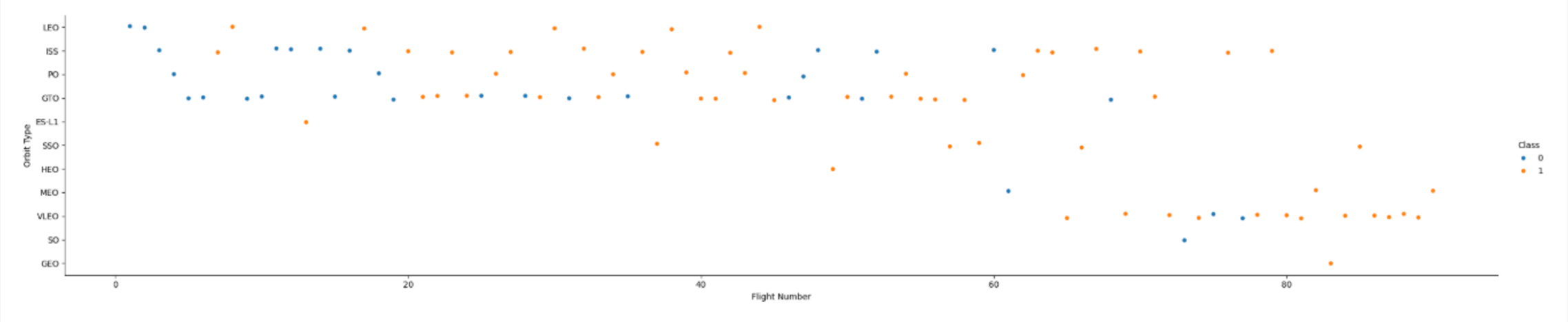
- Show the screenshot of the scatter plot with explanations

ES-L1, GEO, HEO, SSO have the highest success rates.

There are no rockets that go for SO orbit.

Flight Number vs. Orbit Type

- Show a scatter point of Flight number vs. Orbit type



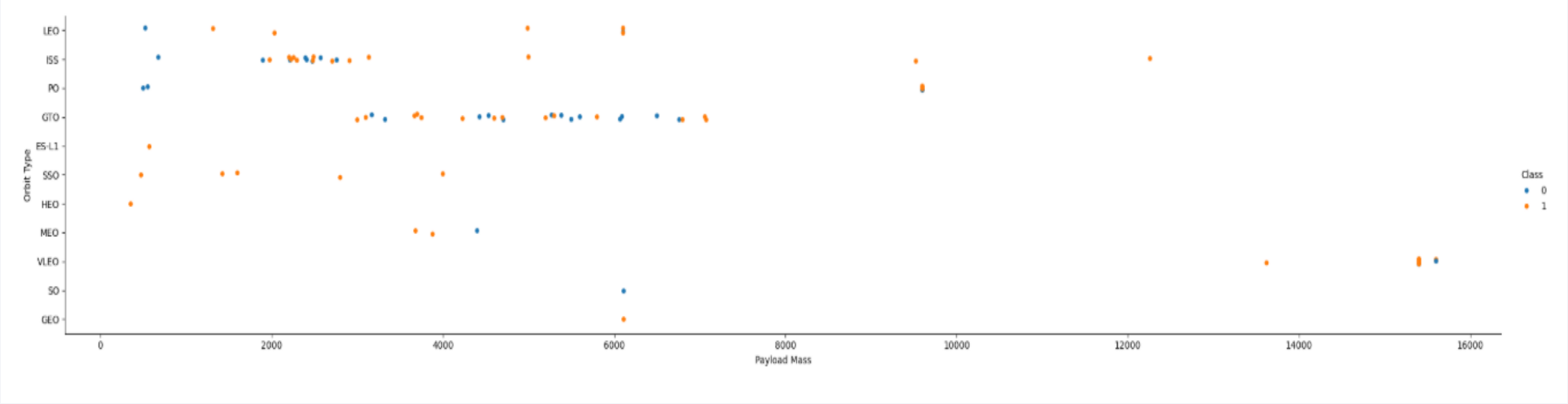
- Show the screenshot of the scatter plot with explanations

In the LEO orbit, success seems to be related to the number of flights.

In the GTO orbit, there appears to be no relationship between flight number and success.

Payload vs. Orbit Type

- Show a scatter point of payload vs. orbit type



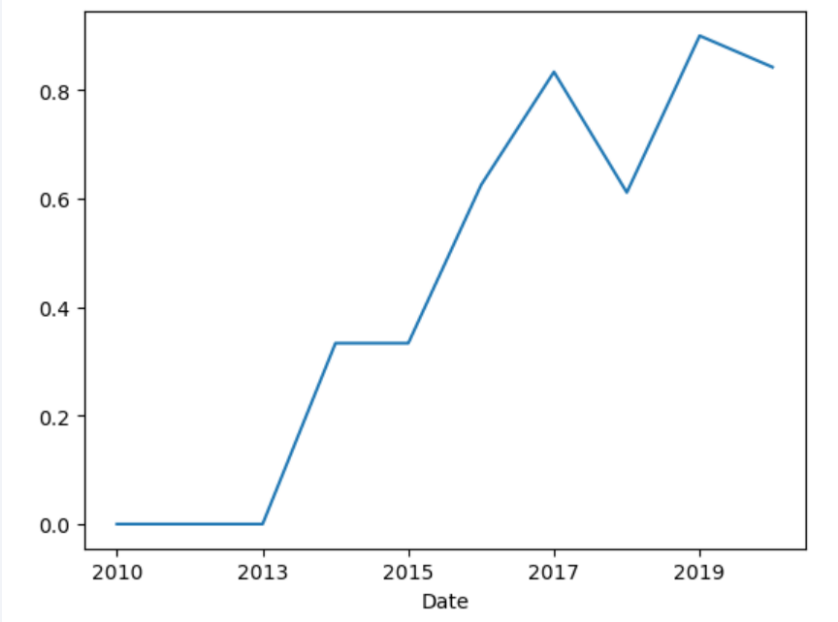
- Show the screenshot of the scatter plot with explanations

With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.

However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend

- Show a line chart of yearly average success rate



- Show the screenshot of the scatter plot with explanations

The sucess rate since 2013 kept increasing from 0 to over 80% till 2020.

All Launch Site Names

- Find the names of the unique launch sites
- `%sql Select Distinct Launch_Site from SPACEXTABLE;`
- Present your query result with a short explanation here

Select distinctive name in the column Launch_Site from the Table SPACEXTABLE

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- `%sql Select Launch_Site from SPACEXTABLE where Launch_Site like 'CCA%' Limit 5;`
- Present your query result with a short explanation here

Select the Launch Site with the site name starting with 'CCA' and display the first 5 records.

```
In [16]: %sql Select Launch_Site from SPACEXTABLE where Launch_Site like 'CCA%' Limit 5;
 * sqlite:///my_data1.db
Done.
```

Out[16]:

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- `%sql Select SUM(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer = 'NASA (CRS)'`
- Present your query result with a short explanation here

Filter the data by the customer is NASA, then select the payload column by applying sum to calculate the total payload.

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

```
In [26]: %sql Select SUM(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer = 'NASA (CRS)'
 * sqlite:///my_data1.db
Done.
Out[26]: SUM(PAYLOAD_MASS__KG_)
         45596
```

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- `%sql Select AVG(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version = 'F9 v1.1'`
- Present your query result with a short explanation here

  Filter data by using booster version, then average the payload column.

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

```
In [25]: %sql Select AVG(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version = 'F9 v1.1'
```

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

Out[25]:

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- `%sql Select Min(Date) from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)'`
- Present your query result with a short explanation here
- Filter the data by the landing outcome on ground pad, then use .min() function on the Date to find the first date

```
In [27]: %sql Select Min(Date) from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)'
 * sqlite:///my_data1.db
Done.
```

Out[27]:

Min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- `%sql Select Booster_Version from SPACEXTABLE where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000`
-

Present your query result with a short explanation here

Filter the data by using the conditions on payload, then select the booster version to print the list of the names of boosters.

Out[28]:

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- `%sql Select Mission_Outcome, Count(*) from SPACEXTABLE Group by Mission_Outcome`
- Present your query result with a short explanation here

Group the data using Mission_Outcome.

```
%sql Select Mission_Outcome, Count(*) from SPACEXTABLE Group by Mission_Outcome
```

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

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- `%sql Select Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);`
- Present your query result with a short explanation here

Filter data by using subquery to find whose payload is the maximum, then select Booster_Version to get the list of the names.

Out[36]:

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

- `%sql Select substr(Date,6,2) as Month, Booster_Version, Launch_Site from SPACEXTABLE where substr(Date,0,5)='2015' and Landing_Outcome='Failure (drone ship)';`

- Present your query result with a short explanation here

Filter the data by using year 2015 and the landing outcomes, then select the required information from the table.

Out[37]:

Month	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	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 Landing_Outcome, COUNT(Landing_Outcome) AS Outcome_Count
FROM SPACEXTABLE
WHERE Date >= '2010-06-04' AND Date <= '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY Outcome_Count DESC;
```

Out[41]:

Landing_Outcome	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

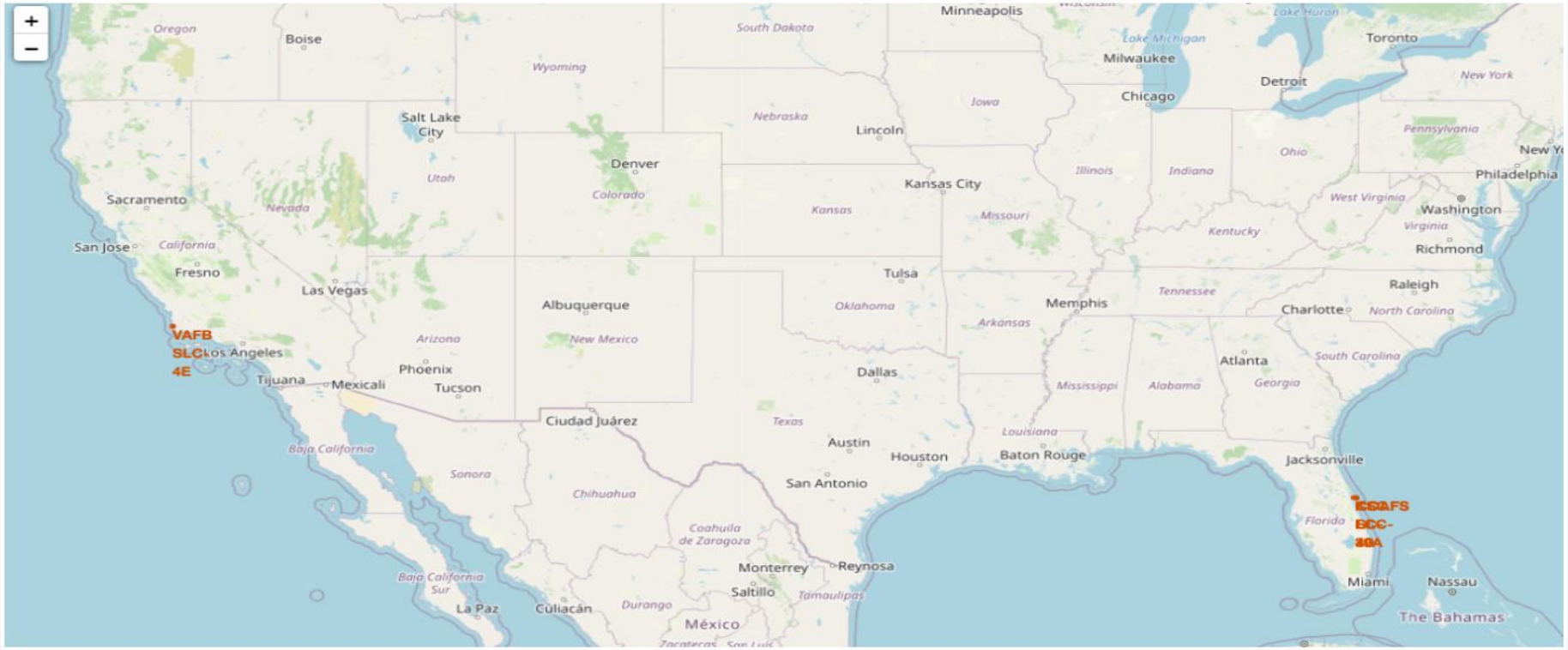
- Present your query result with a short explanation here

Filter the data using the date conditions, then group the data by 'Landing_Outcome' order by desc

Section 3

Launch Sites Proximities Analysis

<SpaceX Launch Site>

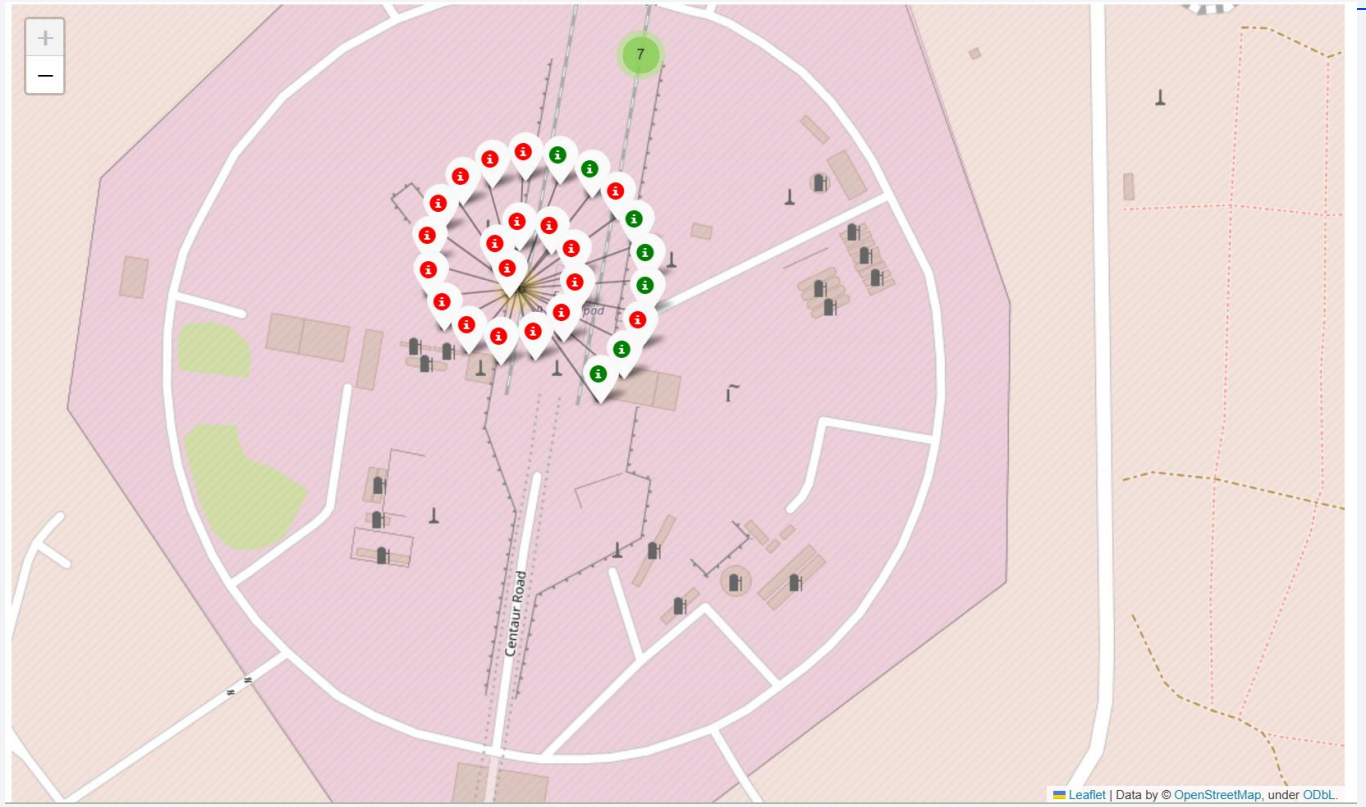


- Explain the important elements and findings on the screenshot

There are two clusters of SpaceX launch sites, one in the east coaster and one in the west coaster.

All launch sites are not in proximity to the Equator line, but they are in very close proximity to the coast.

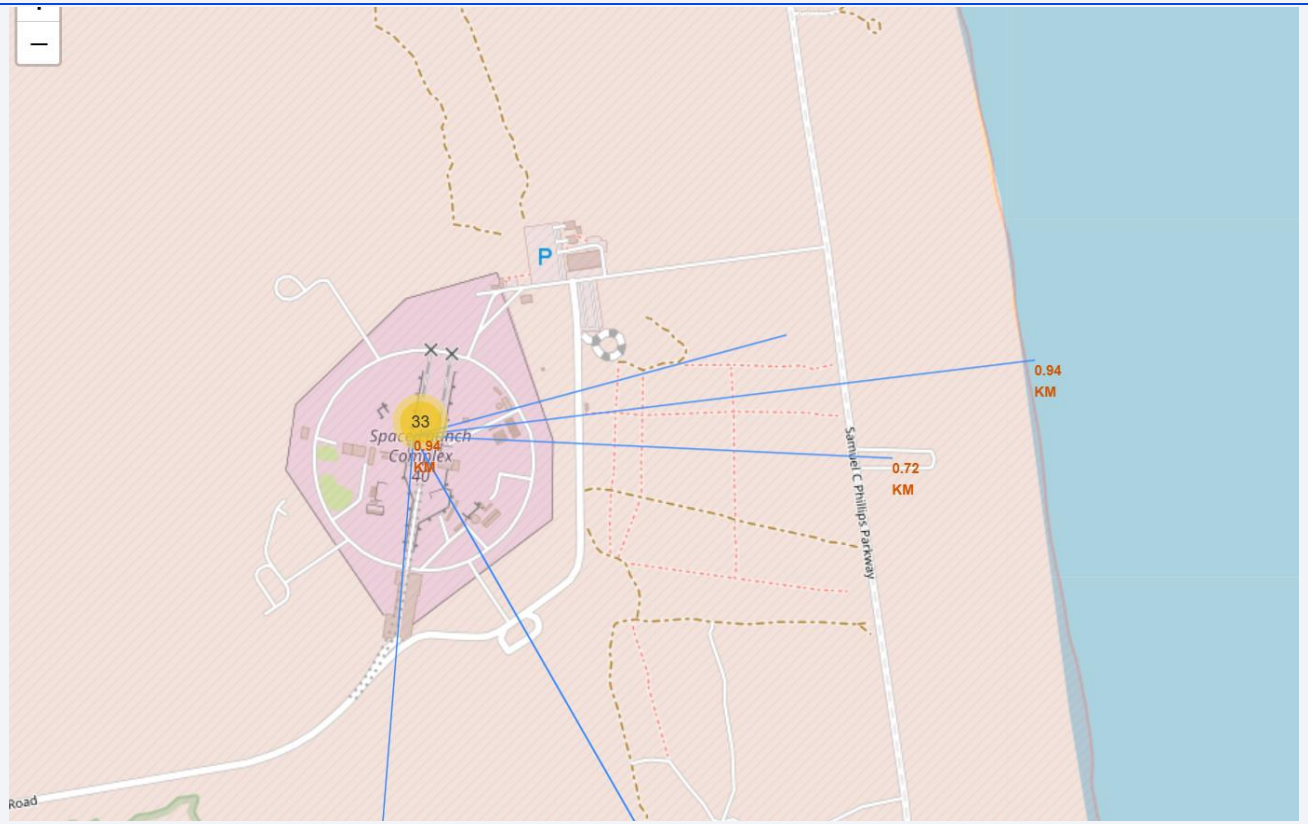
<SpaceX Launch Outcomes>



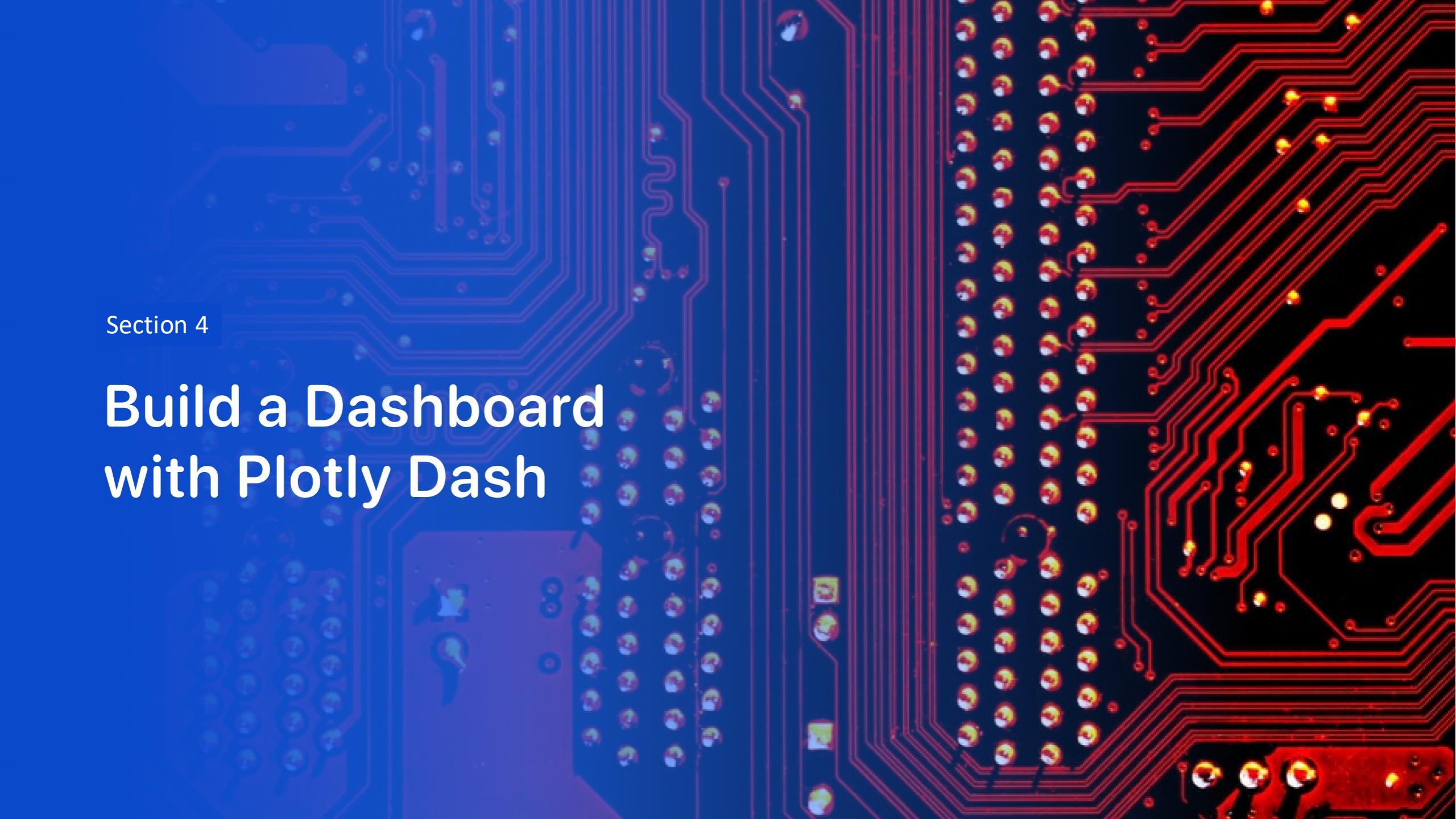
We can easily identify which launch sites have relatively high success rates by zooming in the launch site area.

For example, in CCAFS LC40, there are 7 successes and 19 failures as shown in the above graph.

<SpaceX Launch Site Distance>

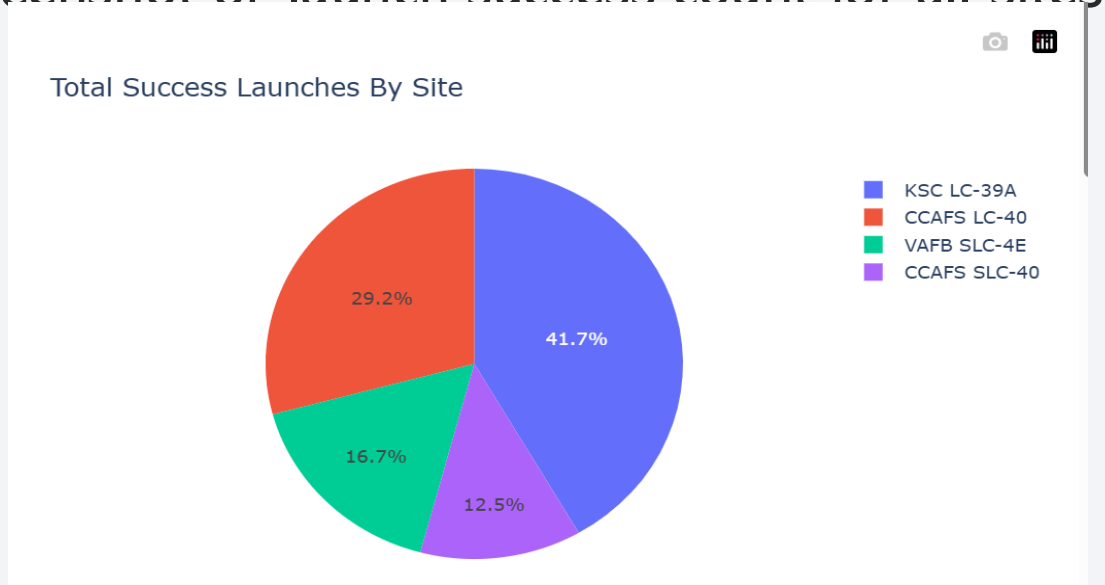


- The launch site in west coast is an idea launch site, which has the smallest distance to railway, highway and coastline as shown in the above graph.


Section 4
Build a Dashboard with Plotly Dash

<Total Success Launches By Site>

- Show the screenshot of launch success count for all sites, in a piechart



- Explain the important elements and findings on the screenshot

KSC LC-39A has the most launch success, accounting for 41.7%, followed by CCAFS LC-40 and VAFB SLC-4E.

CCAFS SLC-40 has the least launch success, around 12.5%.

<Total Success Launches for Site KSC LC-39A>

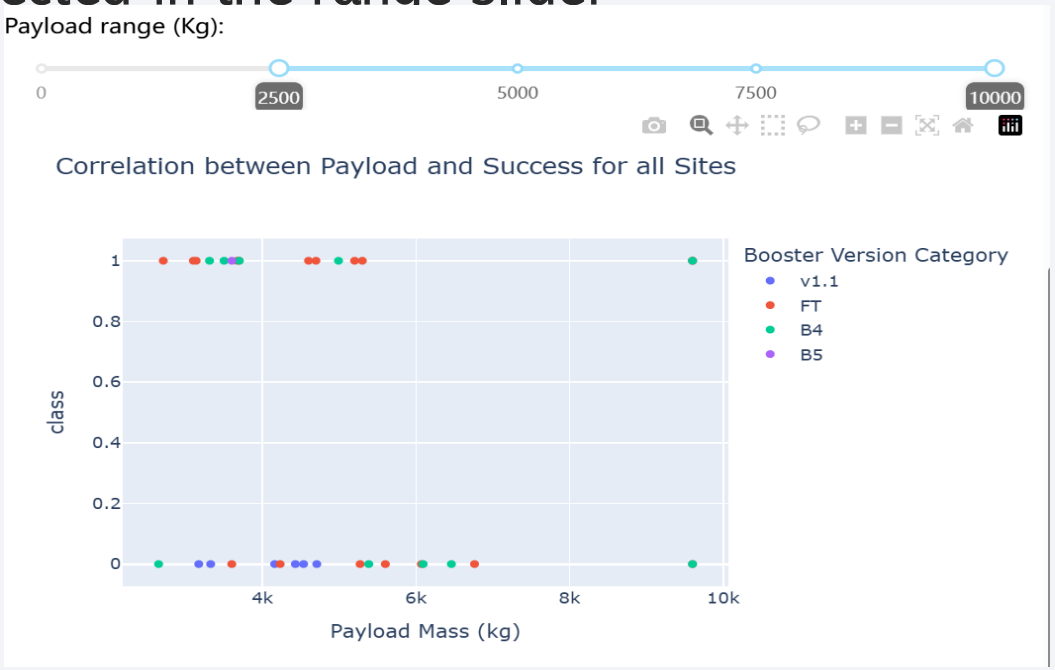
- Show the screenshot of the piechart for the launch site with highest launch success ratio

- Explain the important elements and findings on the screenshot

The site KSC LC-39A has the highest launch success ratio, almost 77%.

<Correlation between Success and Payload>

- Show screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider



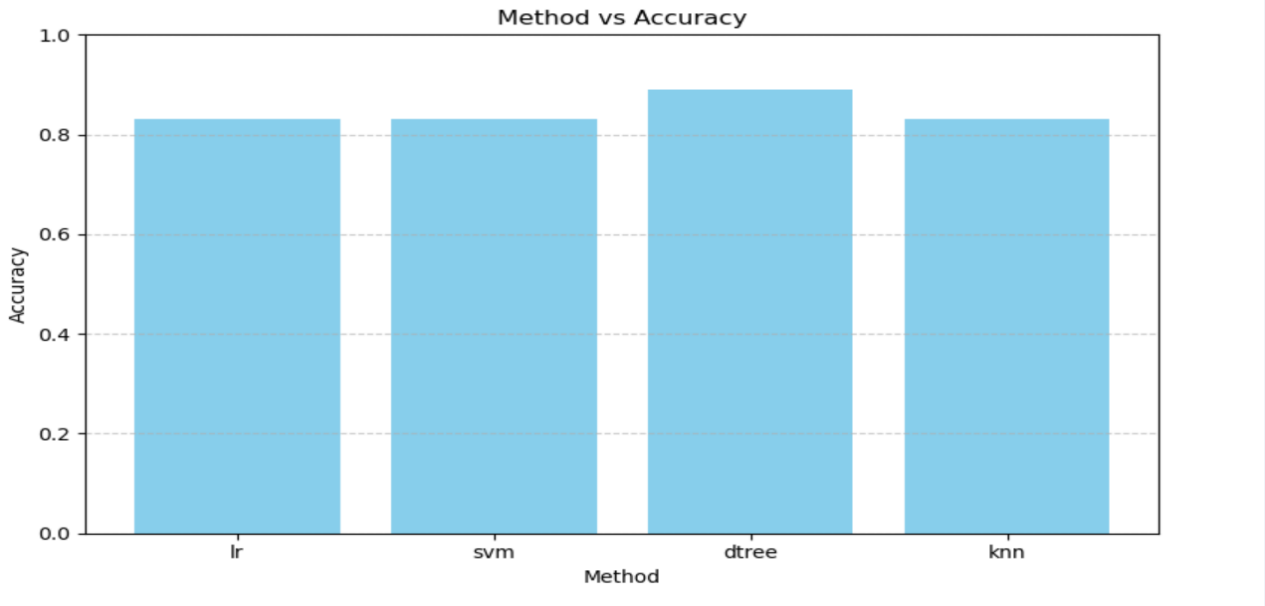
- Falcon FT has the highest launch success, but its payload is less than 8k.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

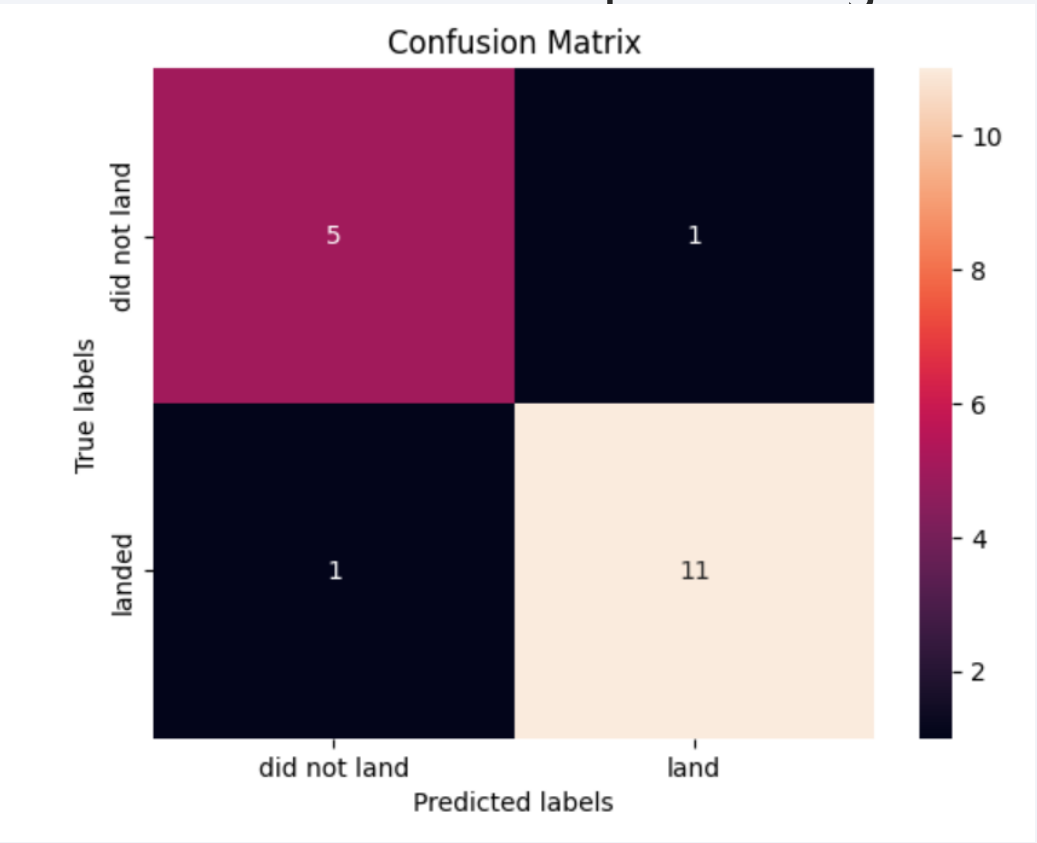
- Visualize the built model accuracy for all built classification models, in a bar chart



- Find which model has the highest classification accuracy
- The decision tree has the highest classification accuracy.

Confusion Matrix

- Show the confusion matrix of the best performing model with an explanation



- There are only 1 false positive and 1 false negative.

Conclusions

- There are several features affecting the success of launch, including Payload mass, Launch site, Booster version, Orbit type and Flight number.
- With the increasing of flight number, the success rate has improved a lot.
- Newer booster versions (e.g., falcon 9) are more reliable and have higher landing success rates.
- Different sites offer varying recovery conditions. Land-based landings are typically more successful than drone ship landings.
- Heavier payloads seem to have a little effect on the chances of a successful landing.

Appendix

- https://github.com/huangzj2025/DataSciencCapstone/tree/main

Thank you!