IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Gabriela Ribeiro
01/11/2023

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies
  - Data collection
  - Data wrangling
  - EDA with data visualization
  - Building an interactive map with Folium
  - Building a Dashboard with Plotly Dash
  - Predictive analysis (Classification)
- Summary of all results
  - EDA results
  - Interactive analytics
  - Predictive analysis

Introduction

Project background and context

- In this project, we're trying to figure out if we can predict whether the first stage of SpaceX's Falcon 9 rocket will land successfully.
- This prediction is essential because it affects the cost of rocket launches. If we succeed, it can save a lot of money. This information could also be valuable for other companies looking to compete with SpaceX.
- We're going to use data analysis tools to tackle this real-world problem and find useful insights.

Problems you want to find answers

- Can we predict the successful landing of the Falcon 9 rocket's first stage?
- How can this prediction impact the cost of a rocket launch, making it more cost-effective?
- What are the implications of this prediction for potential competitors in the rocket launch industry?
- How can data analysis tools and methodologies be applied to address these real-world business challenges and provide actionable insights?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Rest API
  - WEB Scrapping
- Perform data wrangling
  - One Hot Encoding data fields for Machine Learning and data cleaning irrelevant columns and null values.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Logistic Regression, K-nearest neighbors , Support Vector Machine, Decision Tree

Data Collection

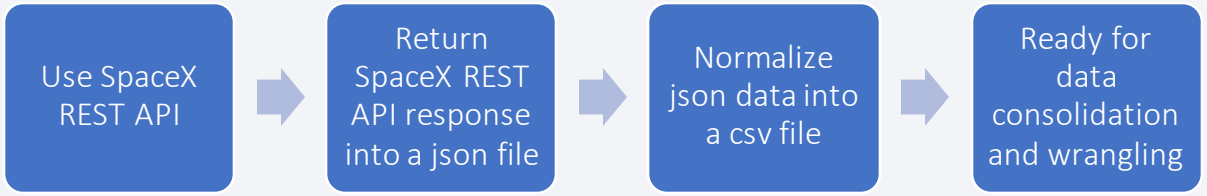
- The SpaceX REST API will give us data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.

- The SpaceX REST API endpoints, or URL, starts with api.spacexdata.com/v4/. We have the different end points, for example: /capsules and /cores. We will be working with the endpoint api.spacexdata.com/v4/launches/past.

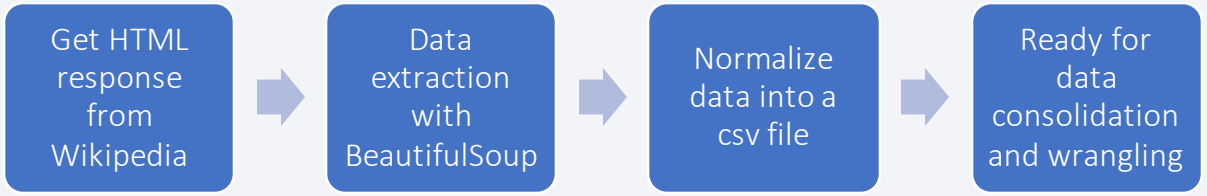
- Another popular data source for obtaining Falcon Launch data is web scraping related Wiki pages.

- Then we need to parse the data from those tables and convert them into a Pandas data frame for further visualization and analysis.

SpaceX REST API



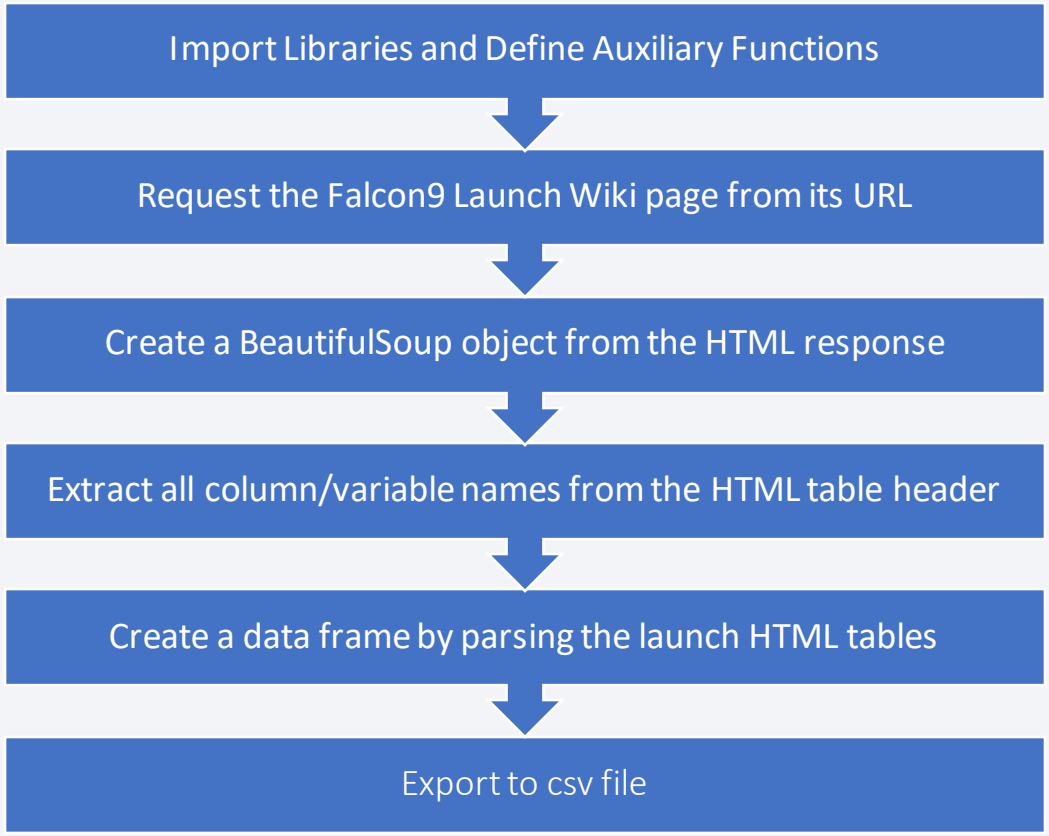
Web Scrapping



Data Collection – SpaceX API

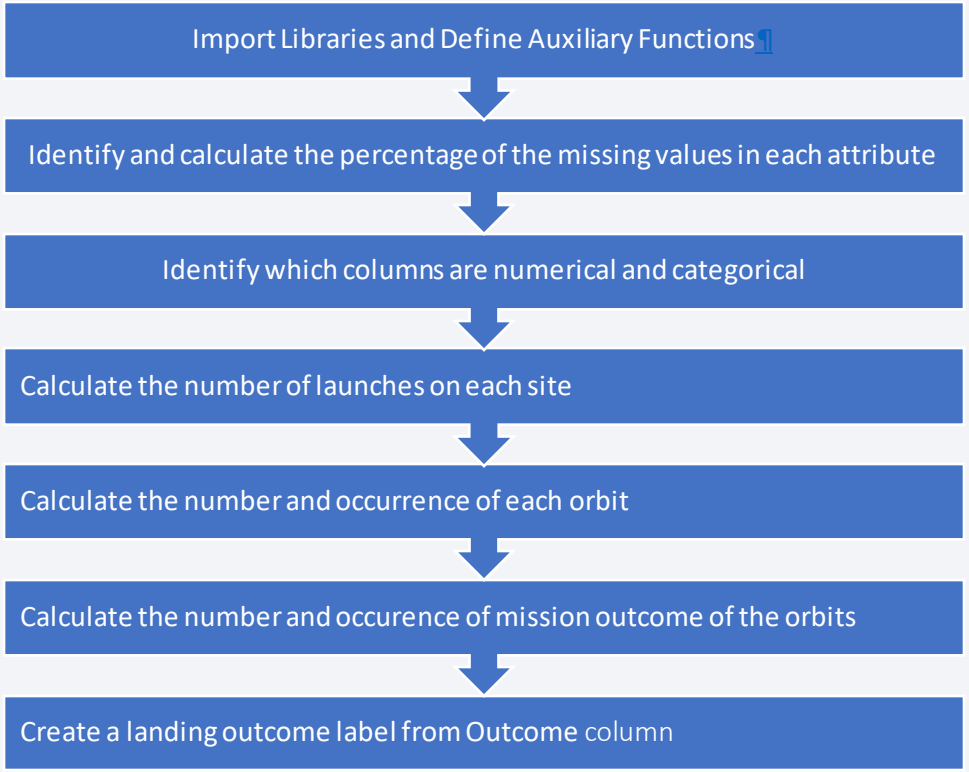
Github : https://github.com/gabrielaRibeiro1/DataScienceCapstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping



Github : https://github.com/gabrielaRibeiro1/DataScienceCapstone/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling



Github: https://github.com/gabrielaRibeiro1/DataScienceCapstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Visualize the relationship between Flight Number and Launch Site
- Visualize the relationship between Payload and Launch Site
- Visualize the relationship between success rate of each orbit type
- Visualize the relationship between FlightNumber and Orbit type
- Visualize the relationship between Payload and Orbit type
- Visualize the launch success yearly trend

Github: https://github.com/gabrielaRibeiro1/DataScienceCapstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with SQL

- **Query 1: Display the names of the unique launch sites in the space mission**
- SQL Query: SELECT DISTINCT Launch_Site FROM SPACEXTBL;
- Purpose: To list the unique launch sites in the dataset.

Github: https://github.com/gabrielaRibeiro1/DataScienceCapstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

- **Query 2: Display 5 records where launch sites begin with the string 'CCA'**
- SQL Query: SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
- Purpose: To retrieve 5 records where the launch site name starts with 'CCA'.

- **Query 3: Display the total payload mass carried by boosters launched by NASA (CRS)**
- SQL Query: SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE PAYLOAD LIKE '%CRS%';
- Purpose: To calculate the total payload mass of boosters launched by NASA in CRS missions.

- **Query 4: Display average payload mass carried by booster version F9 v1.1**
- SQL Query: SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_version LIKE '%F9 v1.1%';
- Purpose: To calculate the average payload mass carried by booster version F9 v1.1.

EDA with SQL

- **Query 5: List the date when the first successful landing outcome in ground pad was achieved.**
- SQL Query: SELECT DATE FROM SPACEXTABLE WHERE LANDING_OUTCOME LIKE 'Success (ground pad)' ORDER BY DATE ASC LIMIT 1;
- Purpose: To find the date of the first successful landing on a ground pad.

- **Query 6: List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000**
- SQL Query: SELECT Booster_version FROM SPACEXTABLE WHERE LANDING_OUTCOME LIKE 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
- Purpose: To retrieve the names of boosters with successful landings on a drone ship and specific payload mass range.

- **Query 7: List the total number of successful and failure mission outcomes**
- SQL Query: SELECT MISSION_OUTCOME, COUNT(*) AS total_number FROM SPACEXTBL GROUP BY MISSION_OUTCOME;
- Purpose: To count and categorize mission outcomes as successful or failure.

EDA with SQL

- **Query 8: List the names of the booster_versions which have carried the maximum payload mass. Use a subquery**
- SQL Query: SELECT BOOSTER_VERSION FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);
- Purpose: To find the names of booster versions that carried the maximum payload mass.

- **Query 9: List the records which will display the month names, failure landing outcomes in drone ship, booster versions, launch_site for the months in the year 2015.**
- SQL Query: SELECT SUBSTR(Date, 6, 2) AS Month, DATE, LANDING_OUTCOME, BOOSTER_VERSION, Launch_Site FROM SPACEXTABLE WHERE SUBSTR(Date, 0, 5) = '2015' AND LANDING_OUTCOME LIKE '%drone ship%' AND SUBSTR(Date, 6, 2) IS NOT NULL ORDER BY Month;
- Purpose: To retrieve records for specific months in 2015, showing month names, failure landing outcomes in drone ships, booster versions, and launch sites.

- **Query 10: 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 Query: SELECT LANDING_OUTCOME, COUNT(*) AS Outcome_Count FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' AND LANDING_OUTCOME IN ('Failure (drone ship)', 'Success (ground pad)') GROUP BY LANDING_OUTCOME ORDER BY Outcome_Count DESC;
- Purpose: To rank the count of landing outcomes between specific dates in descending order for specific landing outcomes.

Build an Interactive Map with Folium

- In the Folium map, several map objects were created and added to enhance the visualization and provide additional information to the viewer:
- **Markers**: Markers were added to indicate the specific launch sites (e.g., CCAFS LC-40, VAFB SLC-4E, KSC LC-39A) on the map. These markers help users quickly identify the launch locations and their geographical positions.
- **Circles**: Circles were used to represent the flight paths or coverage areas of SpaceX missions. For instance, circles around specific launch sites symbolize the reach of the missions launched from those locations. These circles provide a visual representation of mission coverage.
- **Lines**: Lines were utilized to represent the flight paths or trajectories of the missions. These lines help users understand the path taken by the spacecraft during the mission.
- **Popup Labels**: Popup labels were added to markers, circles, and lines to display additional information when users click on these map objects. The popups include details such as the launch site name, mission outcomes, and mission dates.
- The objects were added to the Folium map to provide a clear and informative representation of SpaceX missions. Users can easily identify launch sites, view mission trajectories, and access mission-specific information by interacting with the markers and popups. This enhances the user experience and facilitates the understanding of the dataset.

Github: https://github.com/gabrielaRibeiro1/DataScienceCapstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

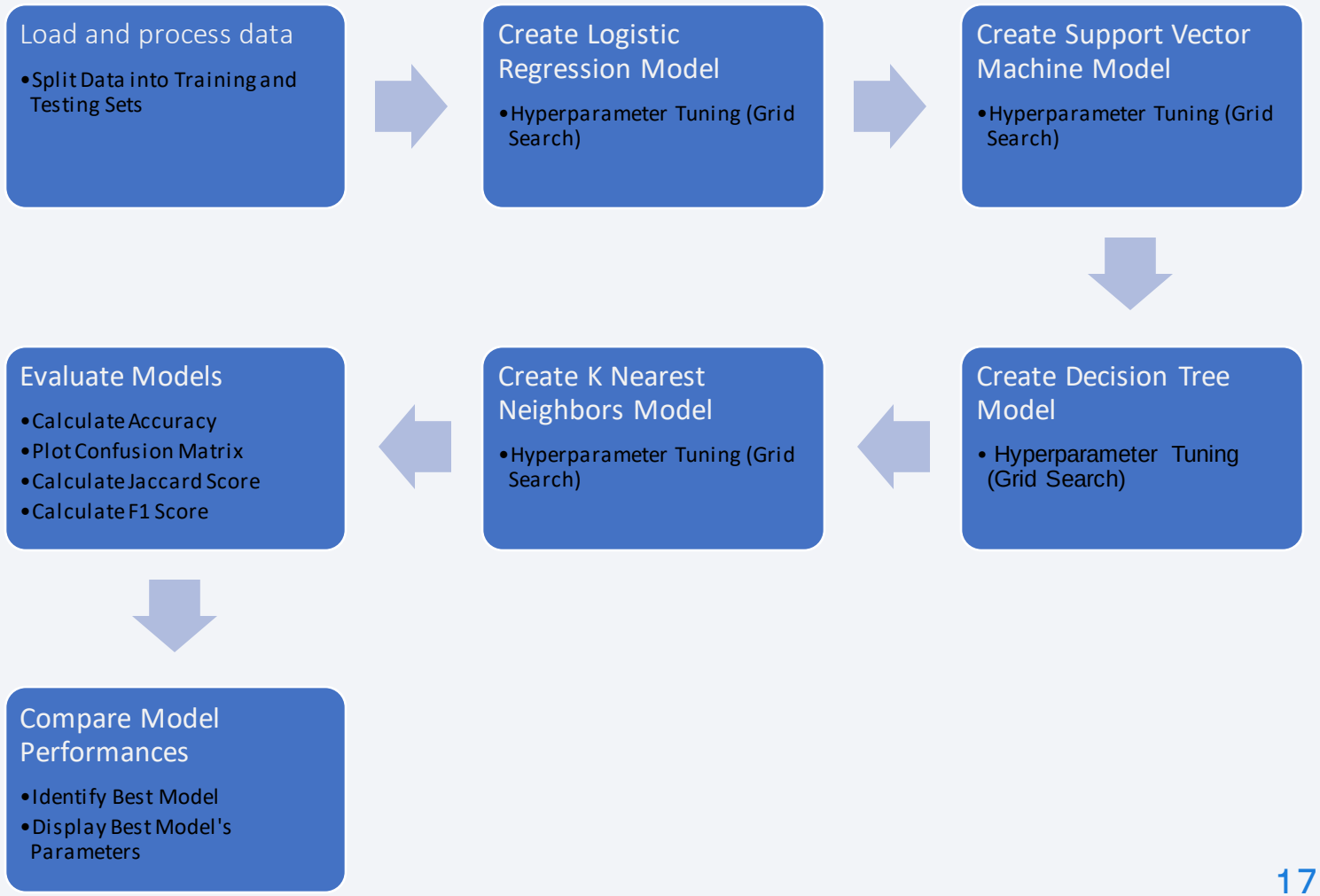
Build a Dashboard with Plotly Dash

- In the Plotly Dash dashboard application, several plots/graphs and interactions were added to facilitate interactive visual analytics of SpaceX launch data:
- **Launch Site Dropdown**: A dropdown menu was added to select different launch sites. This input component allows users to filter data by launch site and view relevant information.
- **Success Pie Chart**: A pie chart was included to display the distribution of successful and failed launches. Users can view the total success launches for all sites or focus on a specific site to see the success and failure counts.
- **Payload Range Slider**: A range slider input was introduced to select a payload range. Users can customize the payload range, which is a key variable for analysis.
- **Success-Payload Scatter Chart**: A scatter plot was created to visualize the relationship between payload mass and mission outcomes. Users can see if there's a correlation between payload and success or failure. Additionally, the booster version is color-labeled to observe mission outcomes with different boosters.
- The interactions and components were added to provide users with a dynamic and informative dashboard that helps answer questions related to SpaceX launch data. By selecting launch sites, payload ranges, and observing scatter plots and pie charts, users can gain insights into the largest successful launches, launch success rates, payload range success rates, and the F9 booster versions with the highest success rates.

Github: https://github.com/gabrielaRibeiro1/DataScienceCapstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- The final result was the Decision Tree model with the specified hyperparameters as the best-performing classification model.
- The Decision Tree model was found to be the best-performing model with an accuracy of 0.8732, and its best parameters include criterion: 'entropy', max depth: 12, max features: 'sqrt', min samples leaf: 4, min samples split: 2, and splitter: 'best'.
- Github: https://github.com/gabrielaRibeiro1/DataScienceCapstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

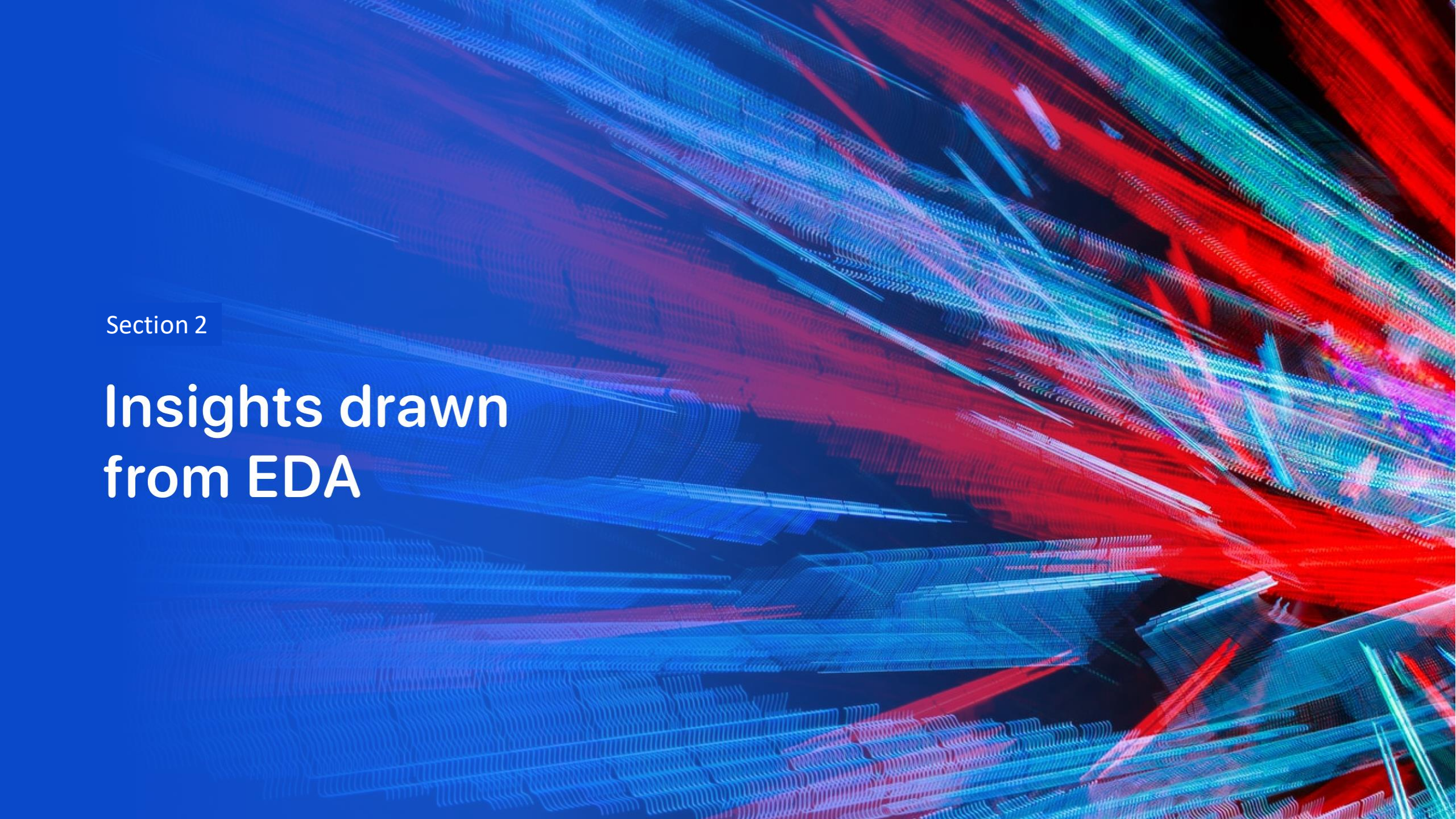


Results

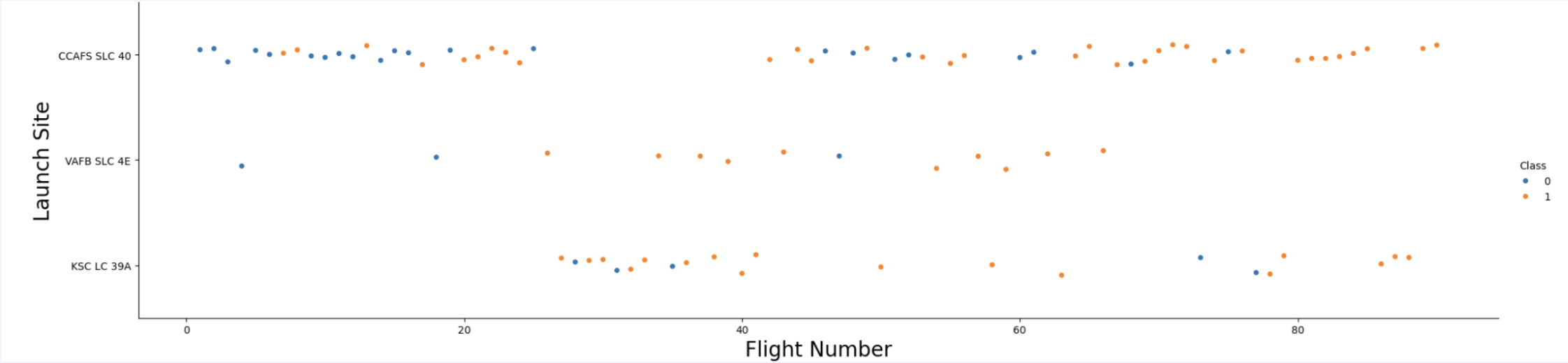
- The Decision Tree Model is the best-performing classification model.
- Low weighted payloads perform better than the heavier payloads.
- The success rates for SpaceX launches is directly proportional time in years.
- KSC LC 39A had the most successful launches from all the sites.
- Orbit GEO, HEO ,SSO ,ESL 1has the best success rate.

Section 2

Insights drawn from EDA

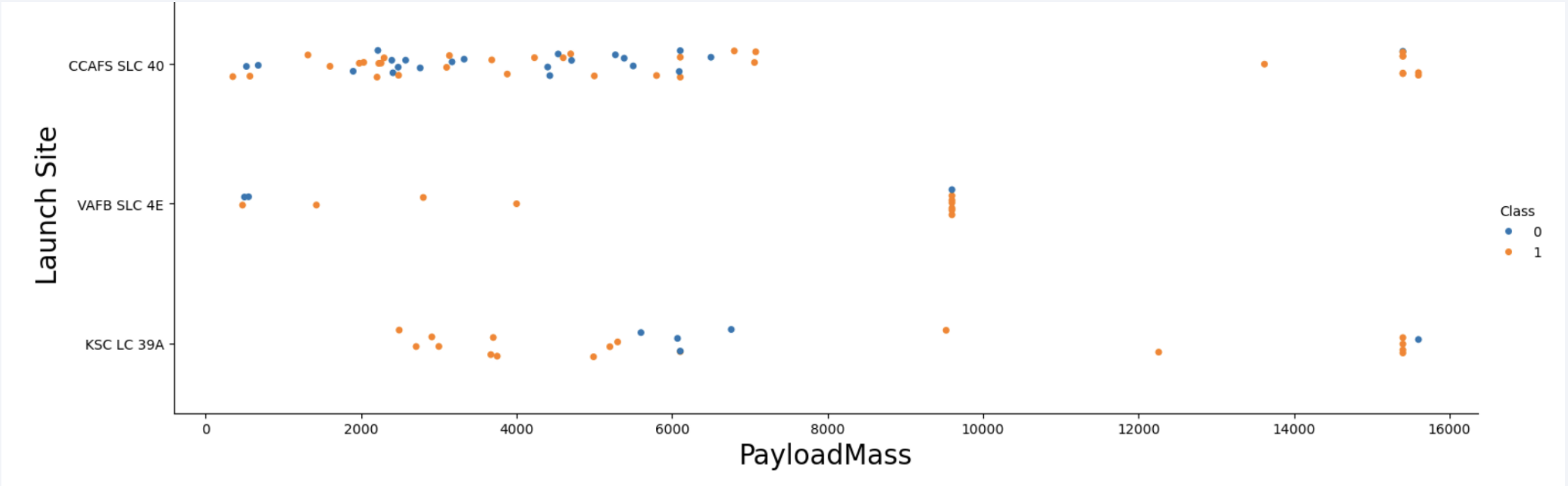


Flight Number vs. Launch Site



The success rate for launch site increases with the increasing of flight numbers.

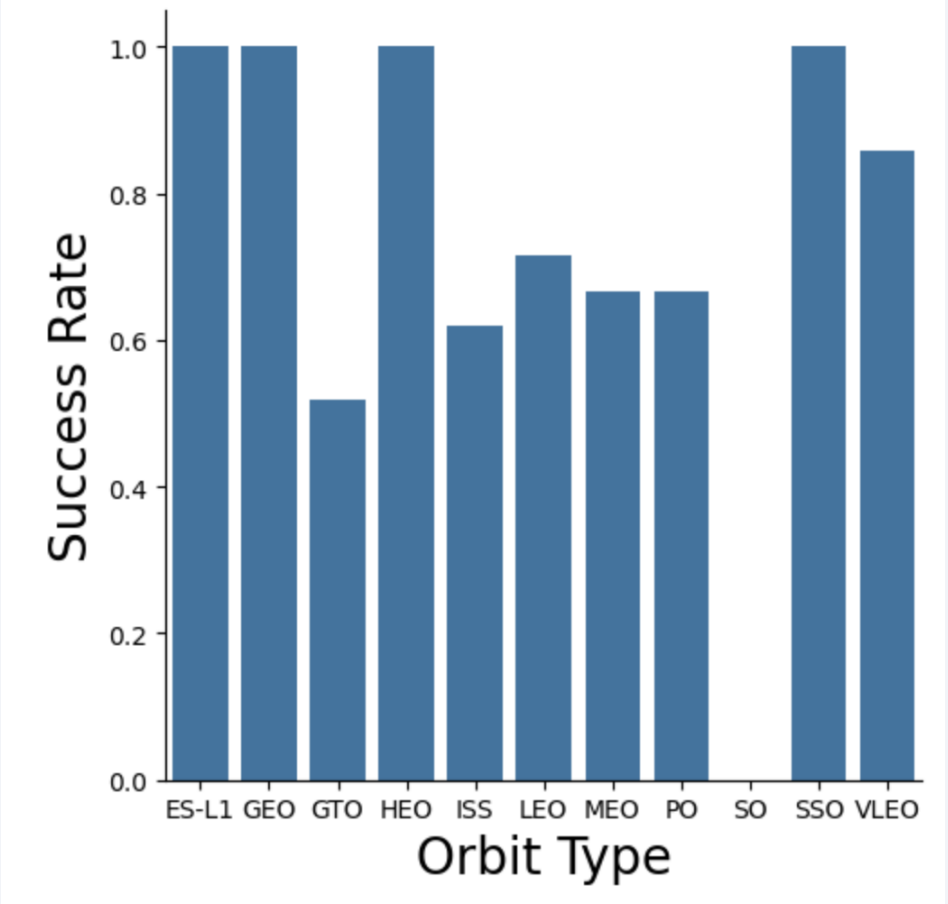
Payload vs. Launch Site



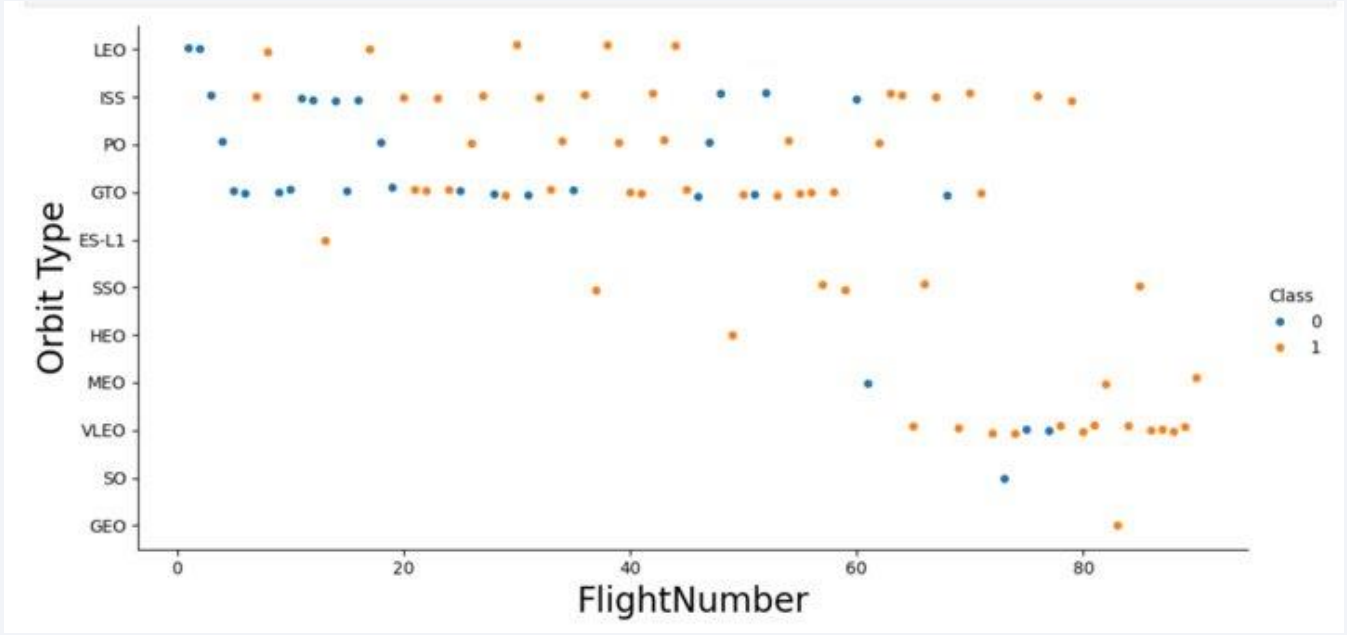
The success rate for launch site increases with the increasing of payload mass.

Success Rate vs. Orbit Type

- The orbit types with 100% success rate are :
  ES-L1, GEO , ISS and SSO.
  The orbit type with 0% rate success is : SO.

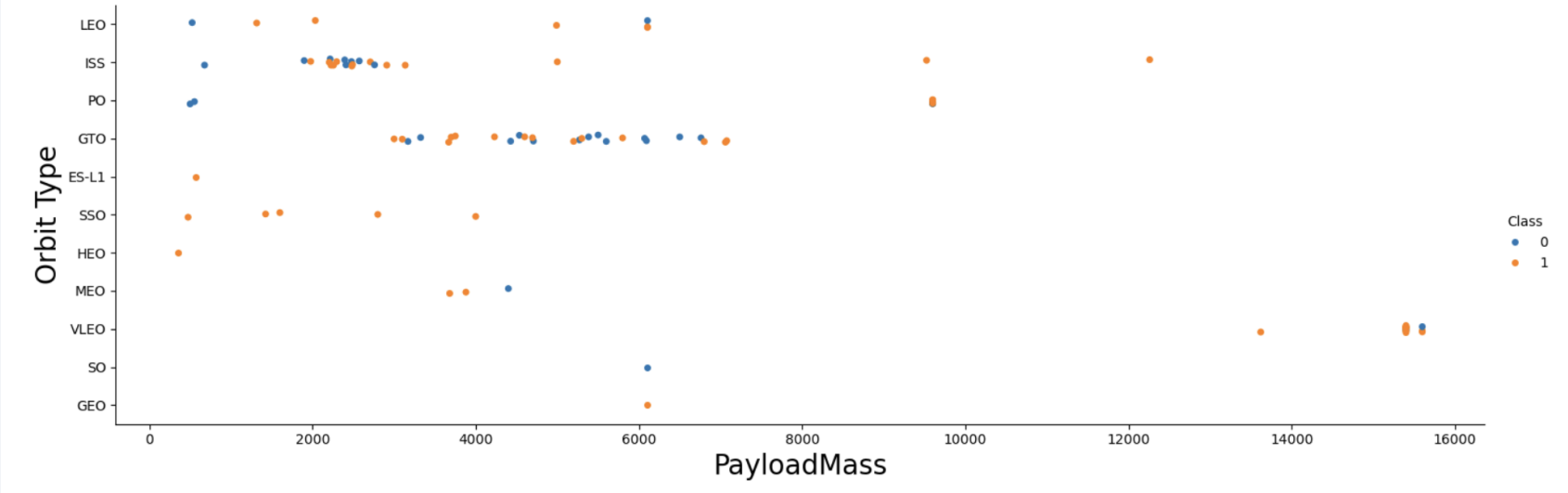


Flight Number vs. Orbit Type



- It doesn't seem to exist a proper relationship between flight number and GTO .

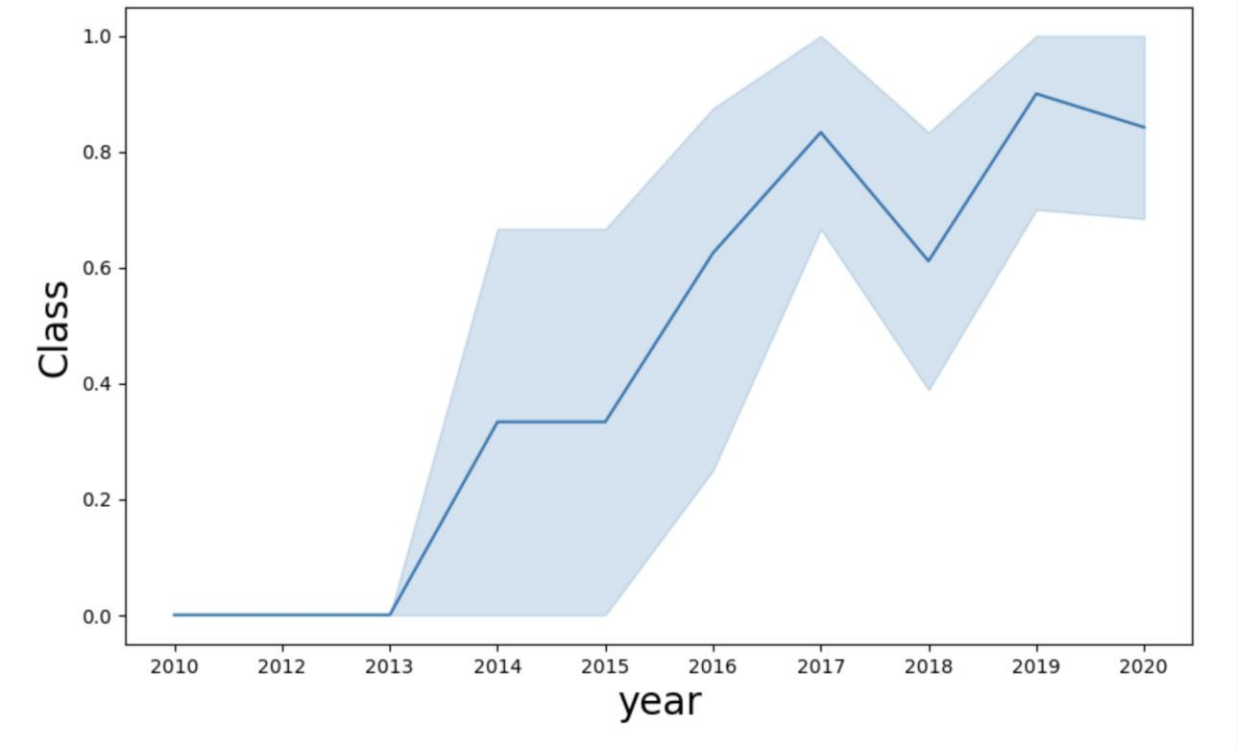
Payload vs. Orbit Type



- The payload mass between 2000kg and 3000kg affects significantly ISS.
- The payload mass between 7000kg and 10000kg affects significantly GTO.

Launch Success Yearly Trend

- Since 2013 there was a big increase in success rate.



All Launch Site Names

```
]: query = "SELECT distinct Launch_Site FROM SPACEXTBL"
   cur.execute(query)

   launch_sites = cur.fetchall()
   for site in launch_sites:
       print(site[0])
```

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

- Purpose: To list the unique launch sites in the dataset.

Launch Site Names Begin with 'CCA'

Task 2

Display 5 records where launch sites begin with the string '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
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- Purpose: To retrieve 5 records where the launch site name starts with 'CCA'.

Total Payload Mass

Task 3

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

```
# Task 3: Display the total payload mass carried by boosters launched by NASA (CRS)
query = "SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE PAYLOAD LIKE '%CRS%'"

# Execute the query
cur.execute(query)

# Fetch the result
total_payload_mass = cur.fetchone()[0]

# Display the total payload mass
print("Total payload mass carried by boosters launched by NASA (CRS):", total_payload_mass, "kg")
```

```
Total payload mass carried by boosters launched by NASA (CRS): 111268 kg
```

- Purpose: To calculate the total payload mass of boosters launched by NASA in CRS missions.

Average Payload Mass by F9 v1.1

Task 4

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

```
query = "SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_version LIKE '%F9 v1.1%'"

# Execute the query
cur.execute(query)

# Fetch the result
average_payload_mass = cur.fetchone()[0]

# Display the total payload mass
print("Average payload mass carried by boosters launched by Booster (F9 v1.1): {:.2f} kg".format(average_payload_mass))
```

```
Average payload mass carried by boosters launched by Booster (F9 v1.1): 2534.67 kg
```

- Purpose: To calculate the average payload mass carried by booster version F9 v1.1.

First Successful Ground Landing Date

Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
query = "SELECT DATE FROM SPACEXTABLE WHERE LANDING_OUTCOME LIKE 'Success (ground pad)' ORDER BY DATE ASC LIMIT 1"

# Execute the query
cur.execute(query)

date = cur.fetchone()[0]

print("Date: ", date)
```

```
Date:  2015-12-22
```

- Purpose: To find the date of the first successful landing on a ground pad.

Successful Drone Ship Landing with Payload between 4000 and 6000

Task 6

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

```
query = "SELECT Booster_version FROM SPACEXTABLE WHERE LANDING_OUTCOME LIKE 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 600
cur.execute(query)

results = cur.fetchall()
for row in results:
    print(row)
```

```
('F9 FT B1022',)
('F9 FT B1026',)
('F9 FT  B1021.2',)
('F9 FT  B1031.2',)
```

- Purpose: To retrieve the names of boosters with successful landings on a drone ship and specific payload mass range.

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

```
%sql SELECT MISSION_OUTCOME, COUNT(*) as total_number FROM SPACEXTBL GROUP BY MISSION_OUTCOME;
```

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

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

- Purpose: To count and categorize mission outcomes as successful or failure.

Boosters Carried Maximum Payload

Task 8

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

```
7]: %sql SELECT BOOSTER_VERSION FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);
```

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

7]:

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

- Purpose: To find the names of booster versions that carried the maximum payload mass.

2015 Launch Records

Task 9

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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
]: %sql SELECT * FROM SPACEXTBL WHERE LANDING_OUTCOME LIKE 'SUCCESS%' AND (DATE BETWEEN '2015-01-01' AND '2015-12-31') ORDER BY DATE DESC
 * sqlite:///my_data1.db
Done.
```

time_utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing_outcome
14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
17:54:00	F9 FT B1029.1	VAFB SLC-4E	Iridium NEXT 1	9600	Polar LEO	Iridium Communications	Success	Success (drone ship)
05:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
04:45:00	F9 FT B1025.1	CCAFS LC-40	SpaceX CRS-9	2257	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
21:39:00	F9 FT B1023.1	CCAFS LC-40	Thaicom 8	3100	GTO	Thaicom	Success	Success (drone ship)

- Purpose: To retrieve records for specific months in 2015, showing month names, failure landing outcomes in drone ships, booster versions, and launch sites.

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

Task 10

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(*) AS Outcome_CountFROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
AND LANDING_OUTCOME IN ('Failure (drone ship)', 'Success (ground pad)') GROUP BY LANDING_OUTCOMEORDER BY Outcome_Count DESC;
```

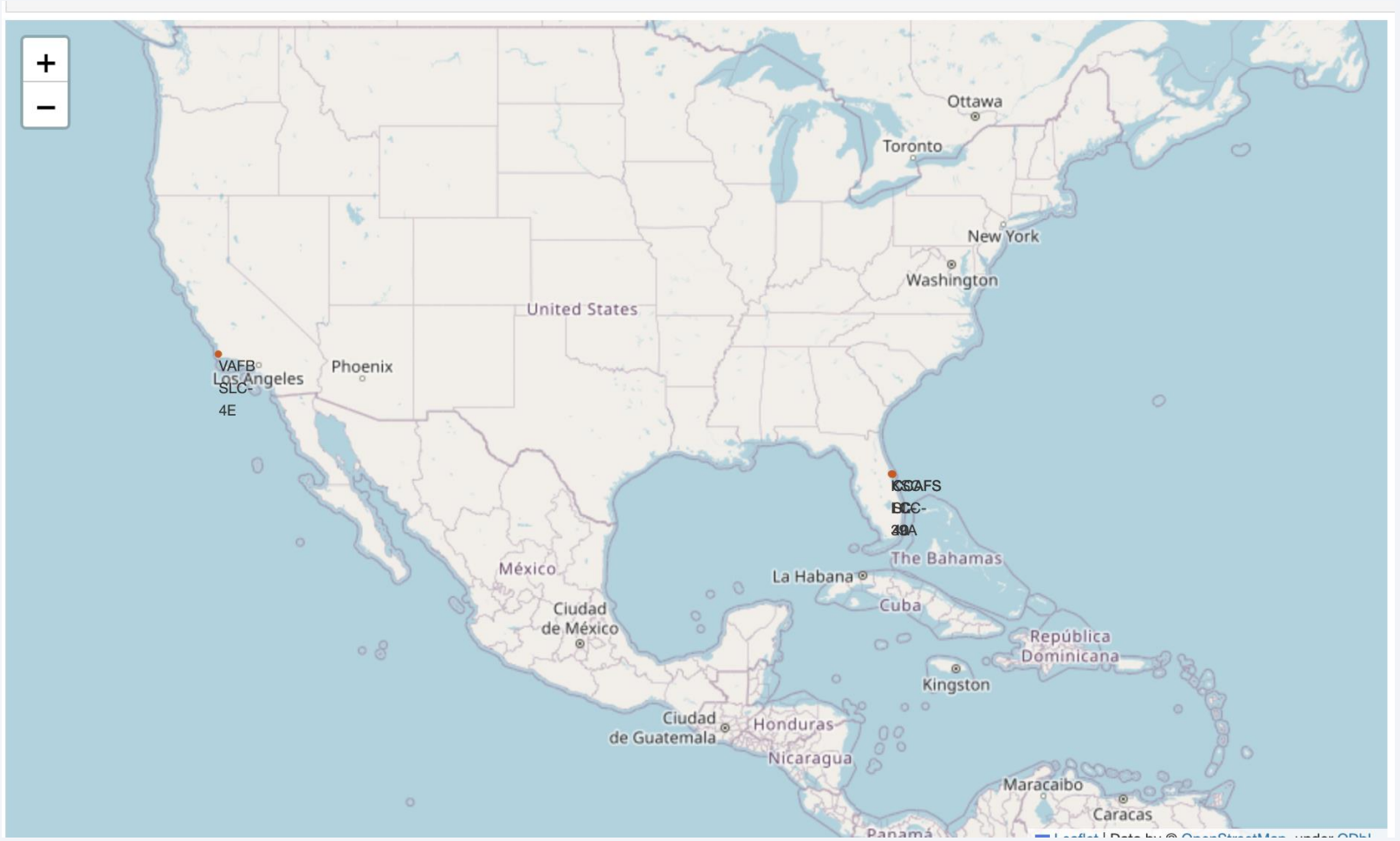
2016-05-27	21:39:00	F9 FT B1023.1	CCAFS LC-40	Thaicom 8	3100	GTO	Thaicom	Success	Success (drone ship)
2016-05-06	05:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2016-04-08	20:43:00	F9 FT B1021.1	CCAFS LC-40	SpaceX CRS-8	3136	LEO (ISS)	NASA (CRS)	Success	Success (drone ship)
2015-12-22	01:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)

- Purpose: To rank the count of landing outcomes between specific dates in descending order for specific landing outcomes.

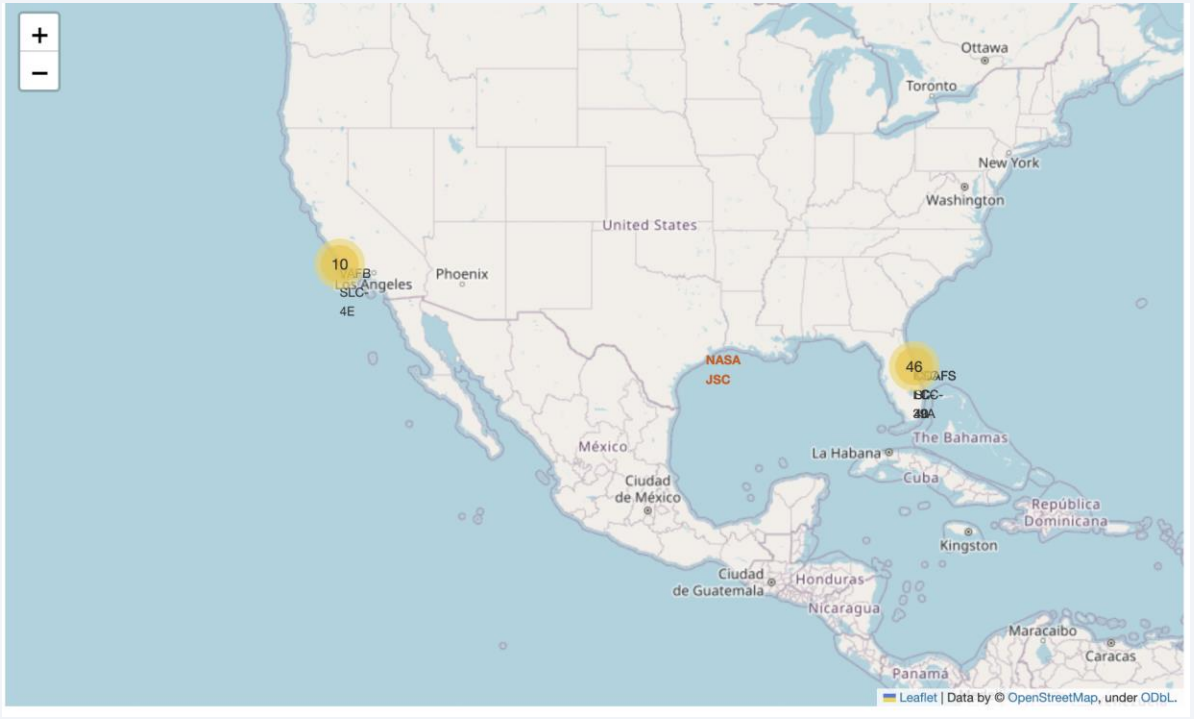

Section 3
Launch Sites
Proximities Analysis

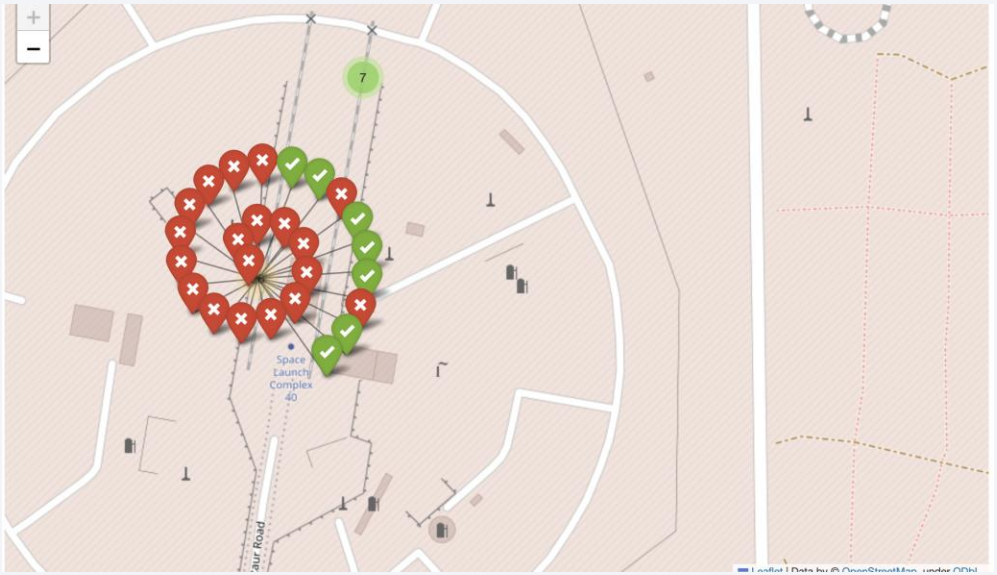
All Launch Sites' Location Markers

- The south area might be a proper area for the sites.



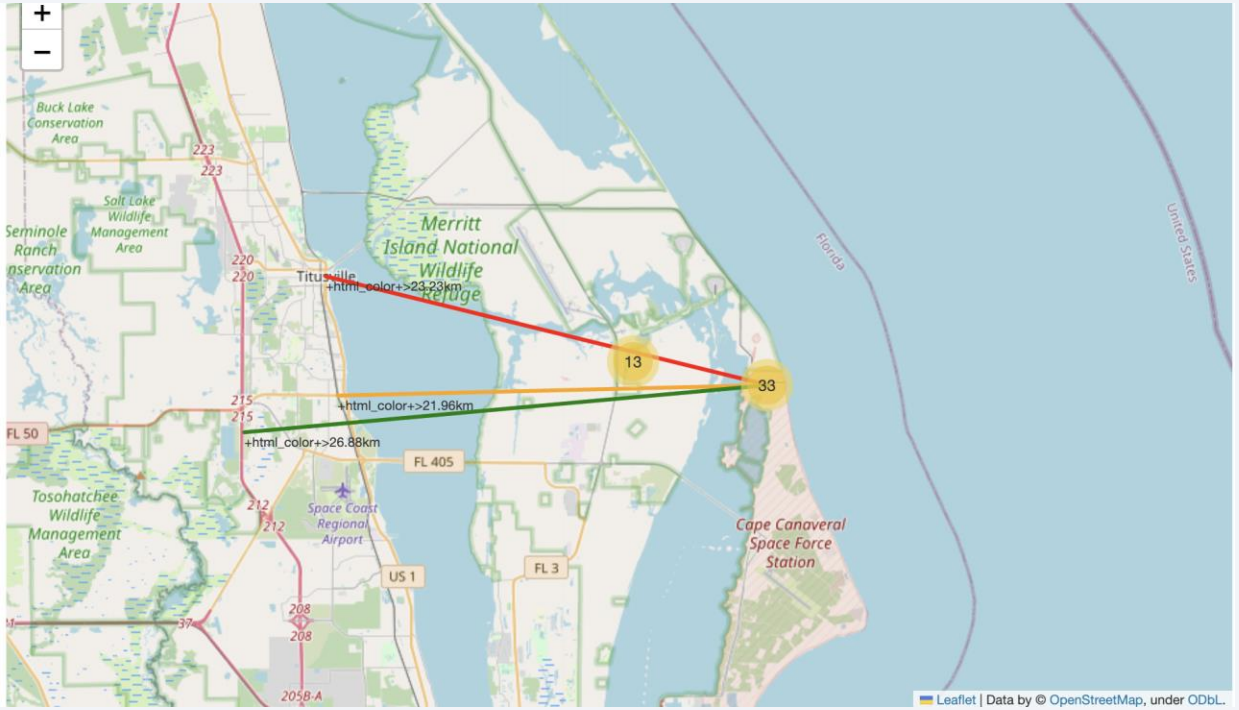
Color-labeled Launch Outcomes



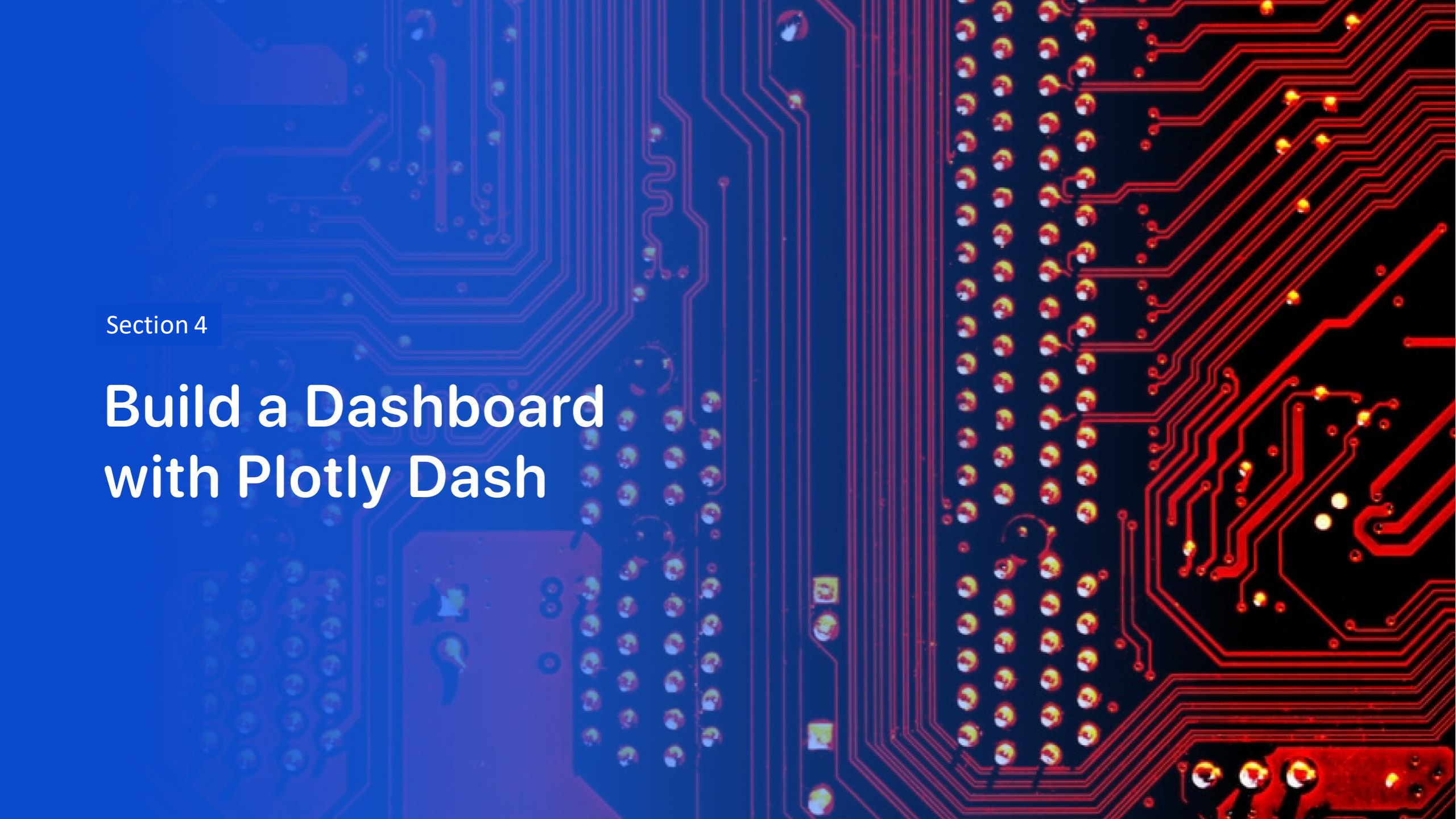


- Color icons displays better the rate of success by area.

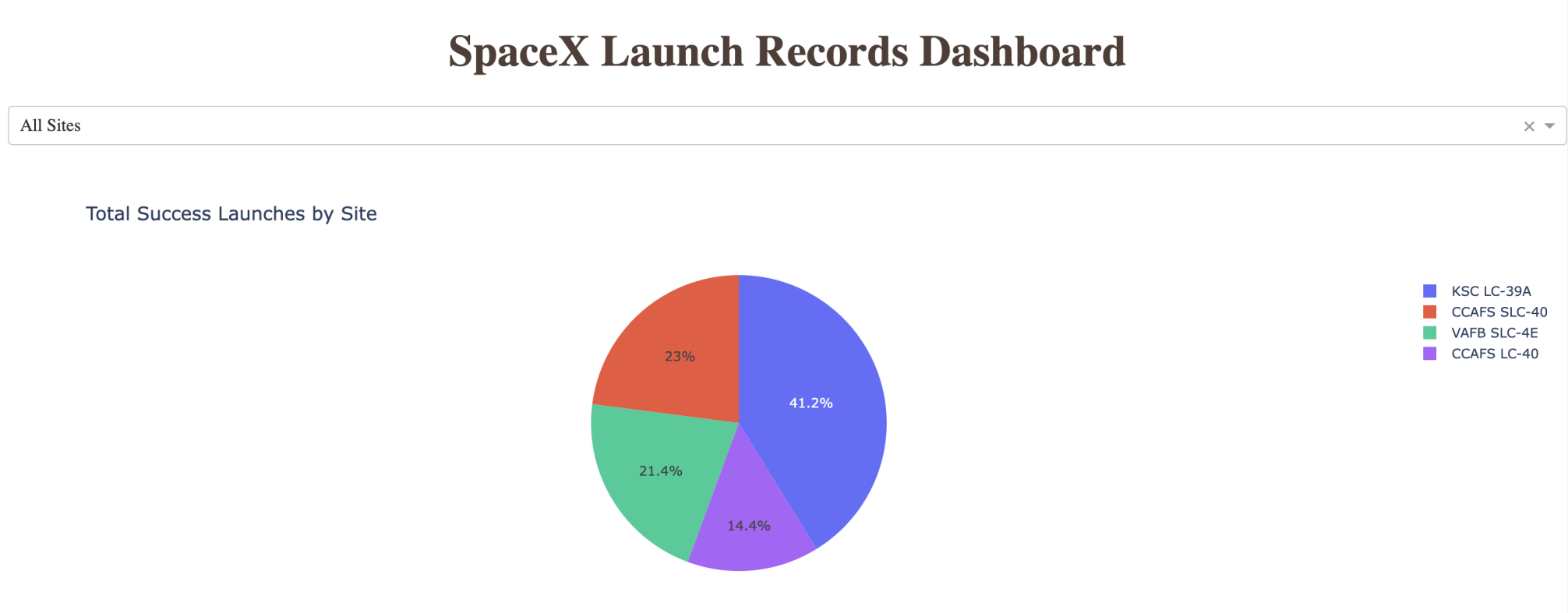
Launch Sites to its Proximities



All distances from launch sites to its proximities , they weren't far away from railway tracks.

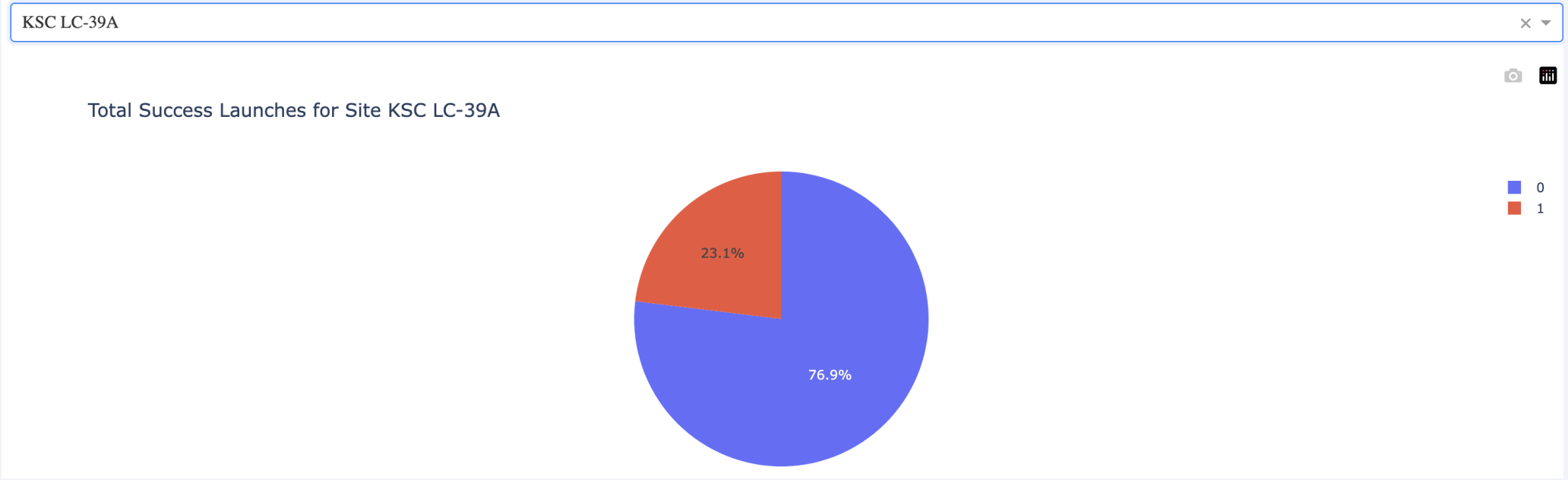

Section 4
Build a Dashboard with Plotly Dash

Total success launches by site



- KSA LC-39A has the highest success score with 41.2%
- CCAFS SLC-40 comes next with 23%
- Finally, VAFB SLC-4E and CCAFS LC-40 with 21.4% and 14.4%

Total success launches for Site KSC LC-39A



- KSC LC-39A has the highest score with 76.9% with payload range of 2000kg to 10000 kg and FT Booster version has the highest score.

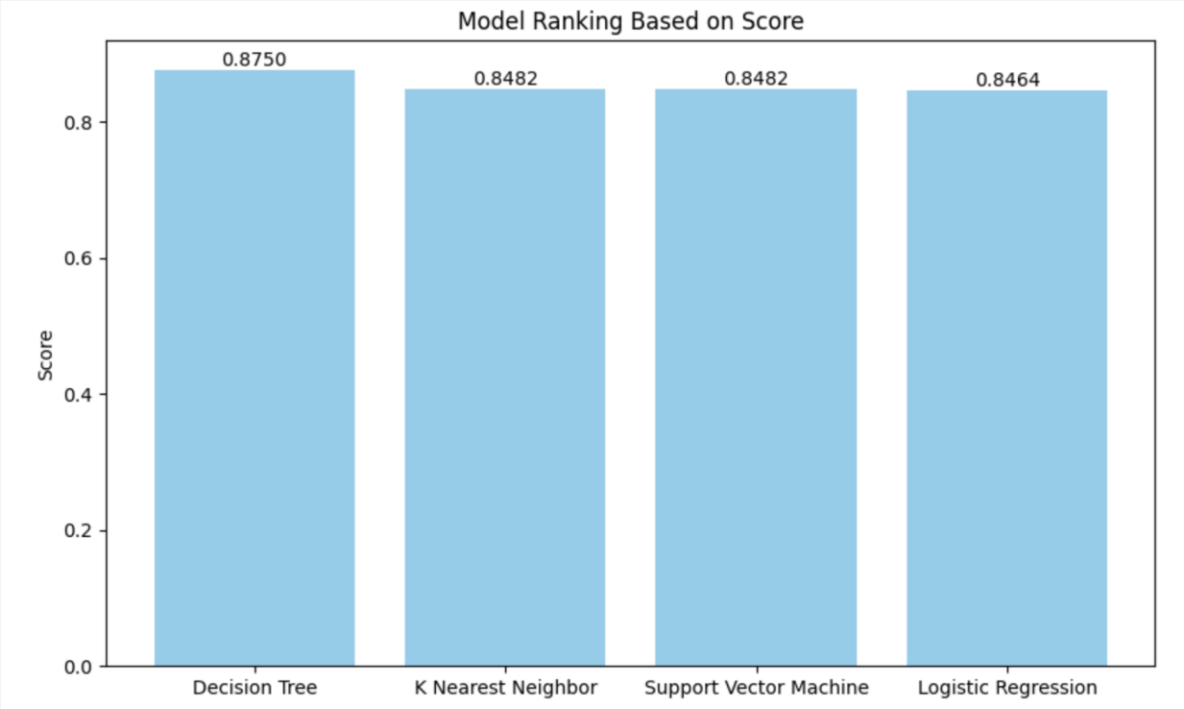
Correlation Between Payload and Success for All Sites

- The success rates for low weighted payloads is higher than the heavy weighted payloads

Section 5

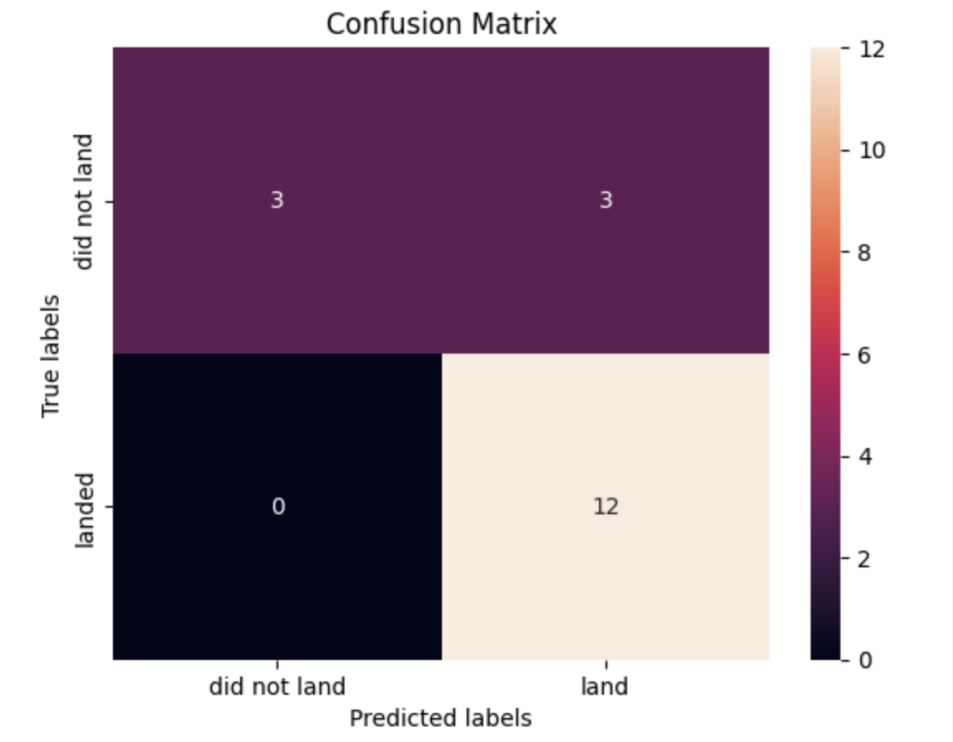
Predictive Analysis (Classification)

Classification Accuracy



- The best model is Decision Tree with the highest accuracy with 0.8750 , then the other models have an accuracy of 0.84

Confusion Matrix



Conclusions

- **Model Performance:** The models performed similarly on the test set with the decision tree model slightly outperforming
- **Equator:** Most of the launch sites are near the equator for an additional natural boost - due to the rotational speed of earth - which helps save the cost of putting in extra fuel and boosters
- **Coast:** All the launch sites are close to the coast
- **Launch Success:** Increases over time
- **KSC LC-39A:** Has the highest success rate among launch sites. Has a 100% success rate for launches less than 5,500 kg
- **Orbits:** ES-L1, GEO, HEO, and SSO have a 100% success rate
- **Payload Mass:** Across all launch sites, the higher the payload mass (kg), the higher the success rate

Appendix

Github Repository : https://github.com/gabrielaRibeiro1/DataScienceCapstone

Thank you!