IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Kamolchanok
Sirikulwattananon
17-Feb-2025

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results

Introduction

- Project background and context
- Problems you want to find answers

Section 1

Methodology

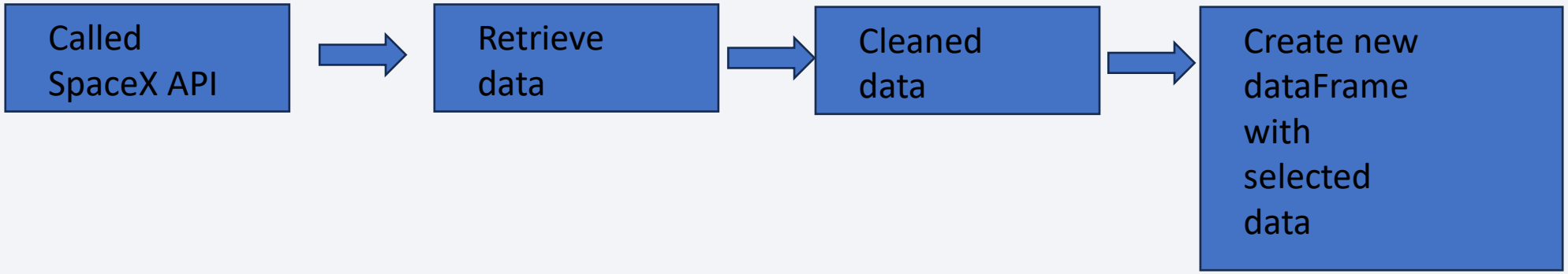
Methodology

Executive Summary

- Data collection methodology:
  - Called SpaceX API to collected data
- Perform data wrangling
  - Cleaning nan value and normalize data, selected columns we needed
- 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 GridSearchCV to help selected best models from LogisticRegression, SVC, DecisionTreeClassifier and KNeighborsClassifier

Data Collection

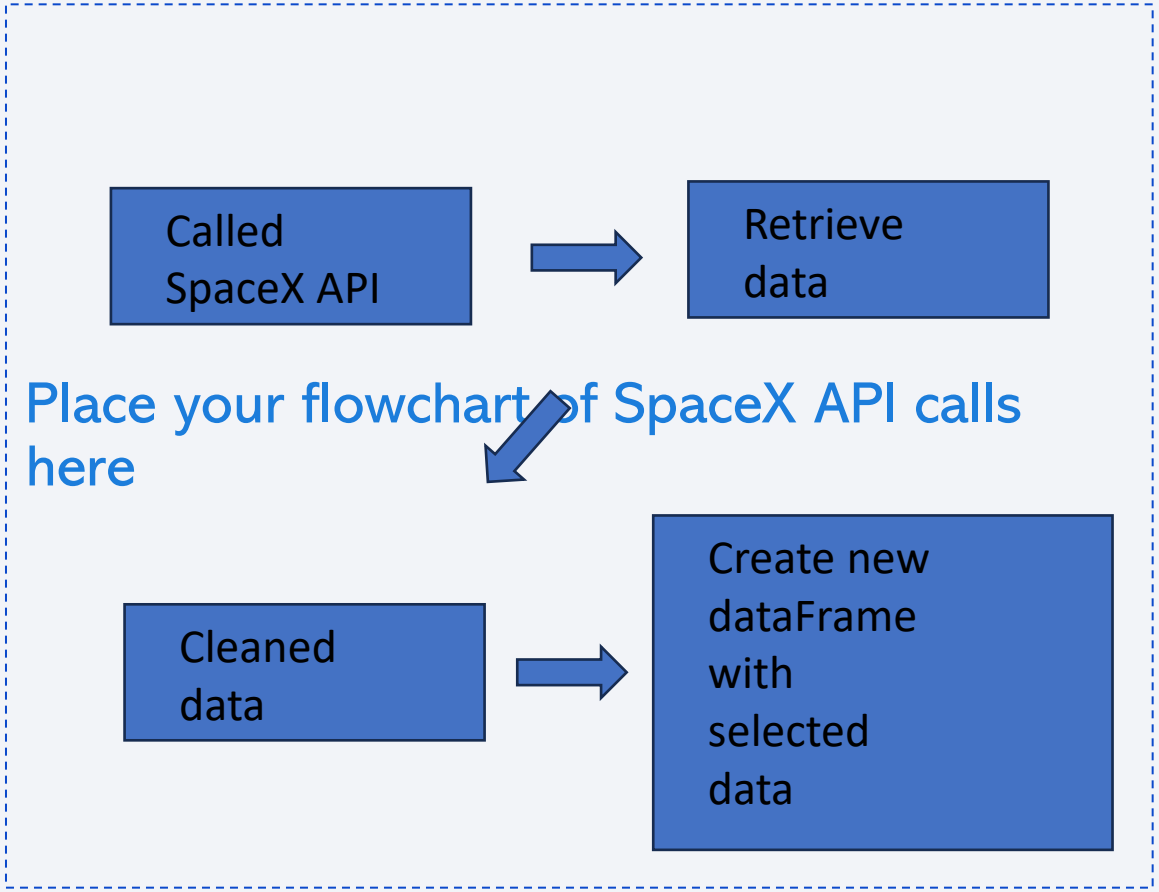
- The data collection process involves making an API request to the SpaceX API to retrieve past launch data. Then cleaned data by dropped data that have NaN and normalize data. Finally, Selected data that related to our target.



Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts

- Add the GitHub URL of the completed SpaceX API calls notebook (must include completed code cell and outcome cell), as an external reference and peer-review purpose
- Reference link: https://github.com/Kamolchanok-S/Applied-Data-Science-Capstone-IBM-Data-Science/blob/8454e1fe7f44ad76832e36f92928d7f4c3f075cf/1%20Hands-on%20Lab%20Complete%20the%20Data%20Collection%20API%20Lab.ipynb



Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts

- Add the GitHub URL of the completed web scraping notebook, as an external reference and peer-review purpose

- Reference link: https://github.com/Kamolchanok-S/Applied-Data-Science-Capstone-IBM-Data-Science/blob/8454e1fe7f44ad76832e36f92928d7f4c3f075cf/2%20Hands-on%20Lab%20Complete%20the%20Data%20Collection%20with%20Web%20Scraping%20lab.ipynb

- Web scraping Falcon 9 and Falcon Heavy Launches Records from Wikipedia
- Request the Falcon9 Launch Wiki page from its URL
- Extract a Falcon 9 launch records HTML table from Wikipedia
- Parse the table and convert it into a Pandas data frame
- Extract all column/variable names from the HTML table header
- Create a data frame by parsing the launch HTML tables
- Export to CSV

Data Wrangling

- Data Collection: Collect data from the Wikipedia
- Web Scraping: Use BeautifulSoup to perform web scraping and extract the HTML table
- HTML Table Extraction: Extract the relevant HTML table containing the launch records
- Data Parsing: Parse the HTML table to extract individual data element.
- Column Extraction: Extract column names and data values from the parsed HTML table and save to dataFrame
- Data Cleaning and Data Transformation: Clean the missing data, removing irrelevant columns, and standardizing formats
- Export to CSV: Export data to CSV file

Reference link: https://github.com/Kamolchanok-S/Applied-Data-Science-Capstone-IBM-Data-Science/blob/8454e1fe7f44ad76832e36f92928d7f4c3f075cf/3%20Hands-on%20Lab%20Data%20Wrangling.ipynb

EDA with Data Visualization

- Bar chart: Show landing outcome frequency.
- Pie chart: Display landing success ratio.
- Scatter plot: Explore payload mass vs. landing outcome.
- Line chart: Track landing success over time.

Reference link: https://github.com/Kamolchanok-S/Applied-Data-Science-Capstone-IBM-Data-Science/blob/d4539ec1395321254426b6d4492246355059ec3d/2%20EDA%20with%20Visualization%20Lab.ipynb

EDA with SQL

- Task 1: Retrieve all columns from the SpaceX dataset to inspect the raw data.
- Task 2: Count the total number of launches to understand the dataset size.
- Task 3: Filter and count successful landings to analyze landing success rates.
- Task 4: Group data by launch site to compare launch frequencies.
- Task 5: Sort payload masses to identify the heaviest and lightest payloads.
- Task 6: Calculate the average payload mass for each launch site.
- Task 7: Identify unique payloads to understand the variety of missions.
- Task 8: Filter launches by specific dates to analyze trends over time.
- Task 9: Categorize landings as successful or unsuccessful using CASE statements.
- Task 10: Summarize total payload mass delivered by each launch site.

Reference link: https://github.com/Kamolchanok-S/Applied-Data-Science-Capstone-IBM-Data-Science/blob/8454e1fe7f44ad76832e36f92928d7f4c3f075cf/1%20Hands-on%20Lab%20Complete%20the%20EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- Markers: Pinpoint launch sites.
- Circles: Highlight launch areas.
- Lines: Show flight paths.
- Polygons: Outline zones.
- Popups: Display launch details.

Reference link: https://github.com/Kamolchanok-S/Applied-Data-Science-Capstone-IBM-Data-Science/blob/8454e1fe7f44ad76832e36f92928d7f4c3f075cf/1%20Hands-on%20Lab%20Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

- Scatter Plot: Payload mass vs. outcome – Analyze landing success.
- Pie Chart: Landing success ratio – Show outcome distribution.
- Dropdowns: Filter by site/payload – Enable data exploration.
- Slider: Adjust payload range – Focus on specific data.

Reference link : https://github.com/Kamolchanok-S/Applied-Data-Science-Capstone-IBM-Data-Science/blob/8454e1fe7f44ad76832e36f92928d7f4c3f075cf/Dash/spacex_dash_app.py

Predictive Analysis (Classification)

- Process:
  - Data preparation (loading, preprocessing, and splitting).
  - Model selection (Logistic Regression, SVM, Decision Tree, KNN).
  - Hyperparameter tuning using GridSearchCV.
  - Model evaluation (accuracy and confusion matrix).
  - Model comparison to identify the best-performing model.
- Result:
  - The best-performing model in the notebook is the Decision Tree Classifier, as determined by the accuracy and confusion matrix analysis.

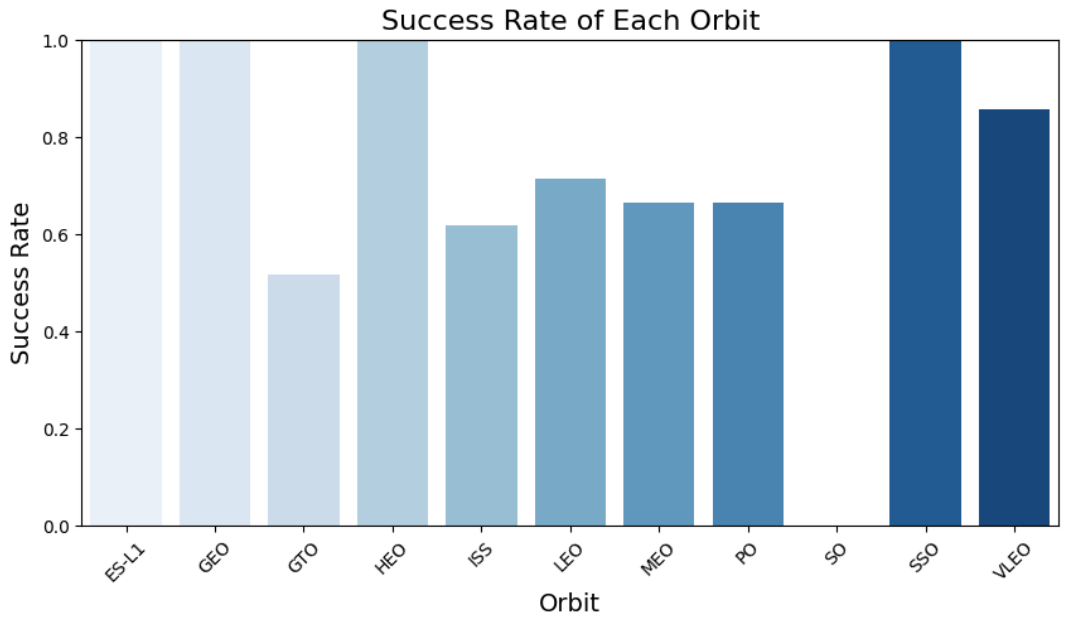
Reference link: https://github.com/Kamolchanok-S/Applied-Data-Science-Capstone-IBM-Data-Science/blob/8454e1fe7f44ad76832e36f92928d7f4c3f075cf/Hands-on%20Lab%20Complete%20the%20Machine%20Learning%20Prediction%20lab.ipynb

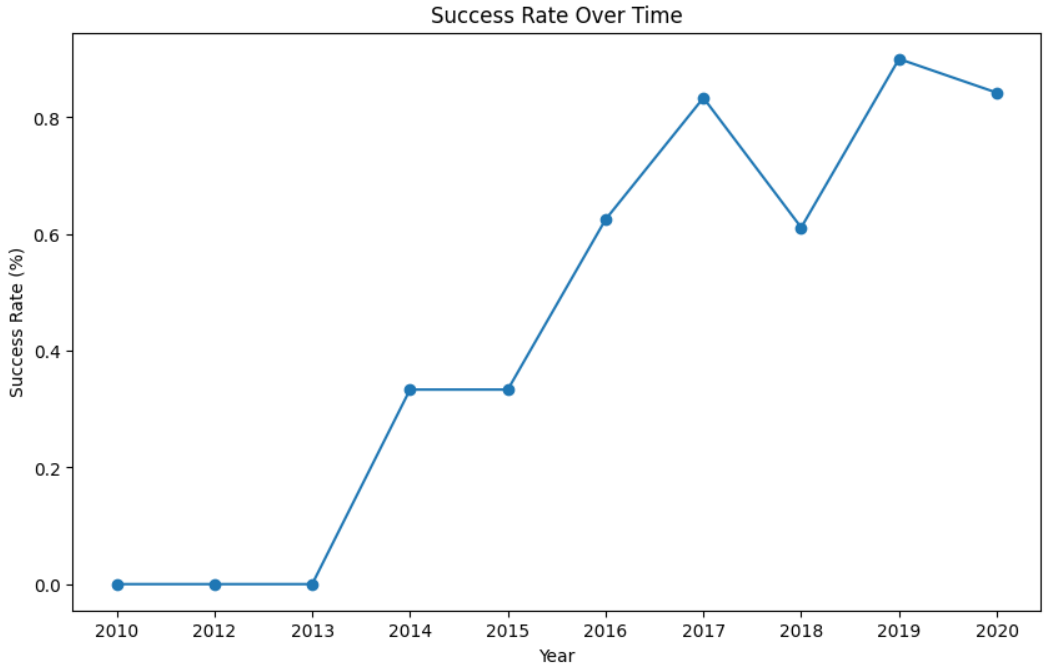
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Results

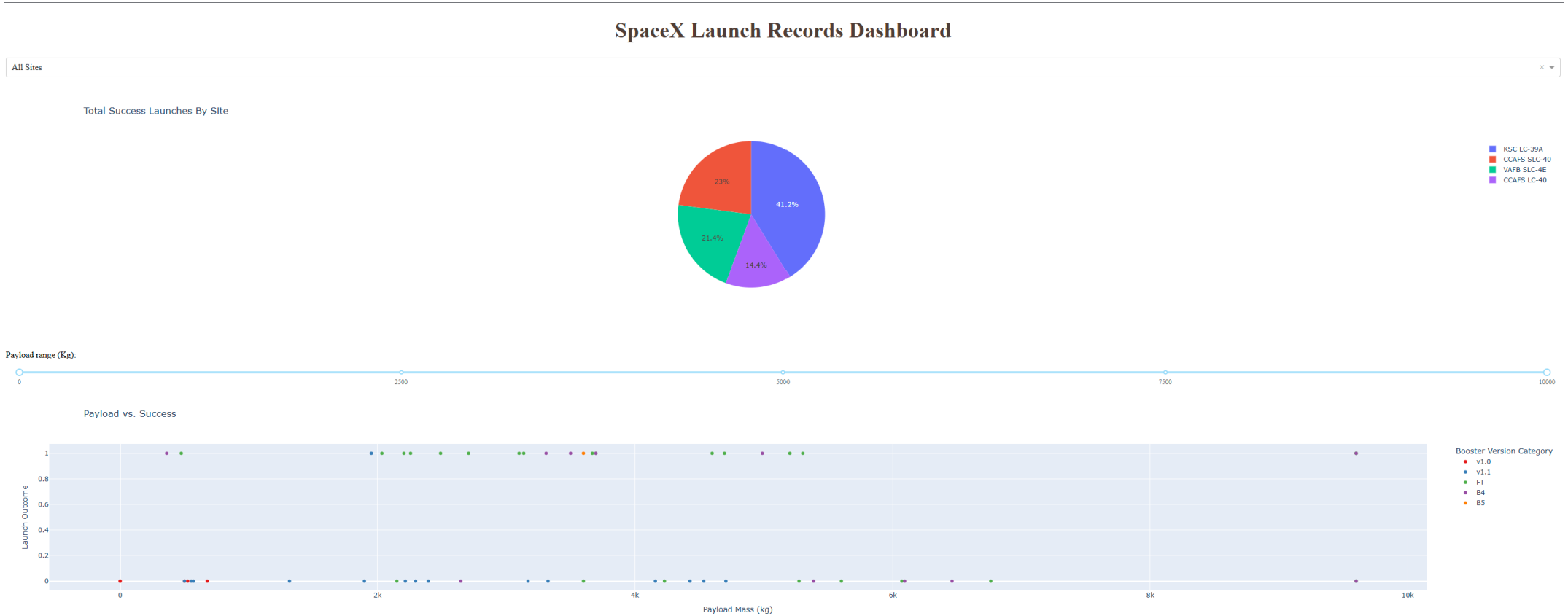
- Exploratory data analysis results





Results

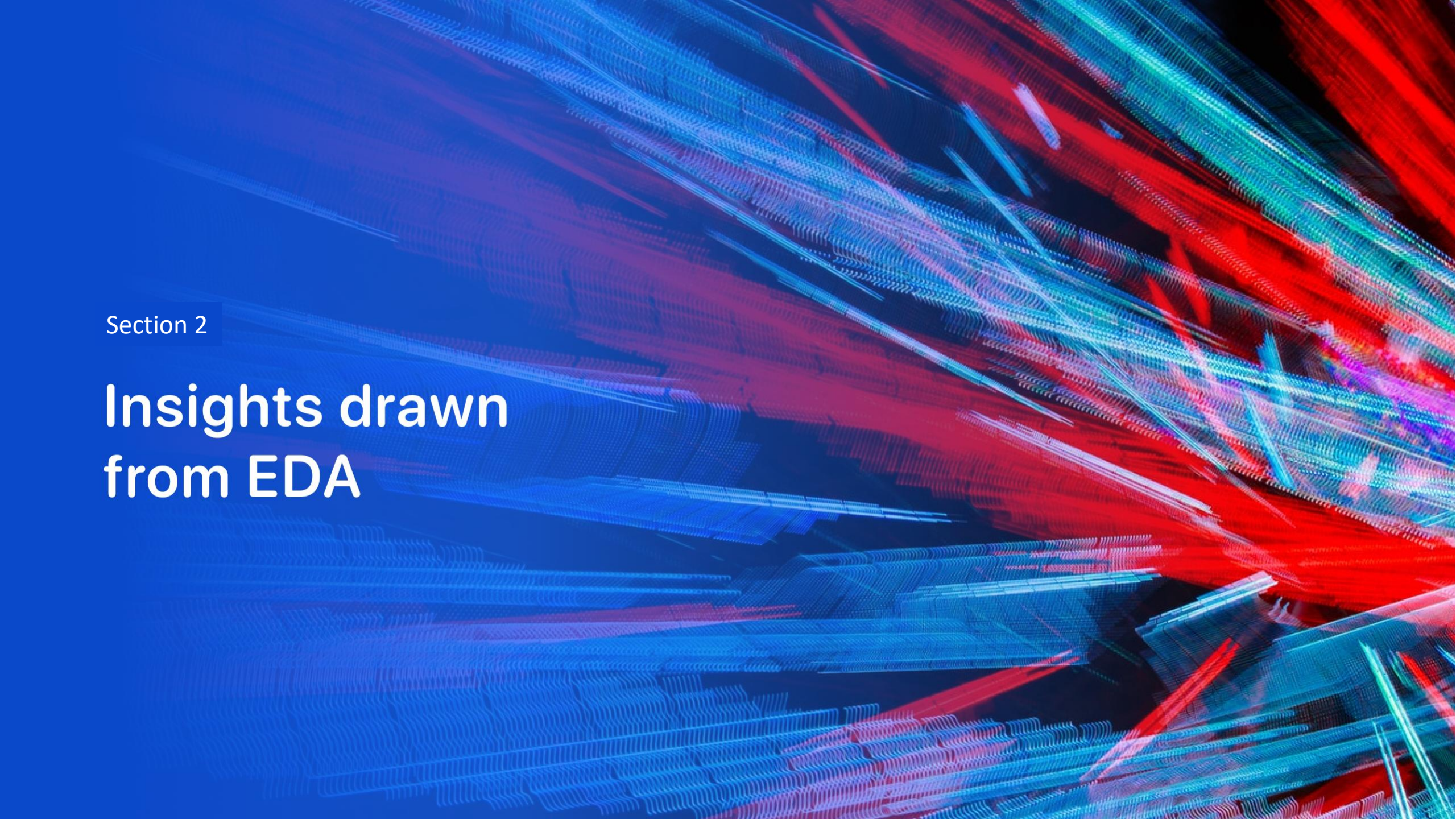
- Interactive analytics demo in screenshots



Results

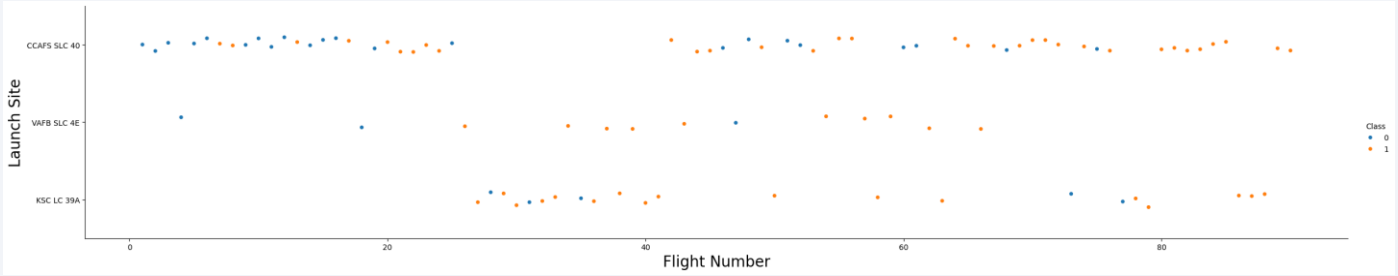
- Predictive analysis results
- Logistic Regression achieved a training and test accuracy of 83.33%. Its confusion matrix showed 12 true positives, 3 false positives, 1 true negative, and 2 false negatives. The best hyperparameters were C=1, penalty='l2', and solver='lbfgs’
- Support Vector Machine (SVM) also achieved a training and test accuracy of 83.33%. Its confusion matrix matched Logistic Regression, with 12 true positives, 3 false positives, 1 true negative, and 2 false negatives. The best hyperparameters were C=1, gamma=0.01, and kernel='sigmoid’
- Decision Tree performed the best, with a perfect training accuracy of 100% and a test accuracy of 94.44%. Its confusion matrix showed 13 true positives, 1 false positive, 3 true negatives, and 1 false negative. The best hyperparameters were criterion='gini', splitter='best', max_depth=18, max_features='auto', min_samples_leaf=1, and min_samples_split=2
- K-Nearest Neighbors (KNN) achieved a training and test accuracy of 83.33%. Its confusion matrix showed 12 true positives, 3 false positives, 1 true negative, and 2 false negatives. The best hyperparameters were n_neighbors=10, algorithm='auto', and p=1

Conclusion: The Decision Tree model is the best-performing model, with the highest test accuracy (94.44%) and the lowest number of misclassifications. It outperformed Logistic Regression, SVM, and KNN


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

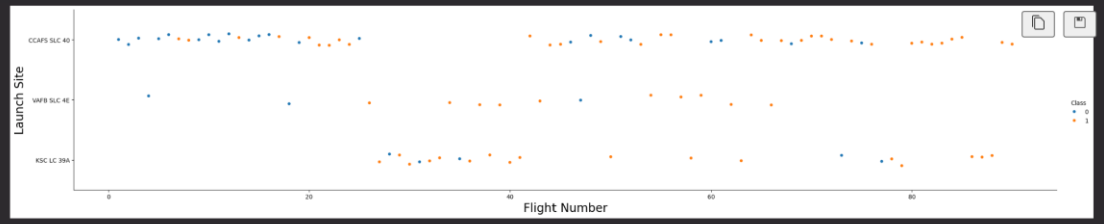
- Show a scatter plot of Flight Number vs. Launch Site



- Show the screenshot of the scatter plot with explanations

```
sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()
```

Python



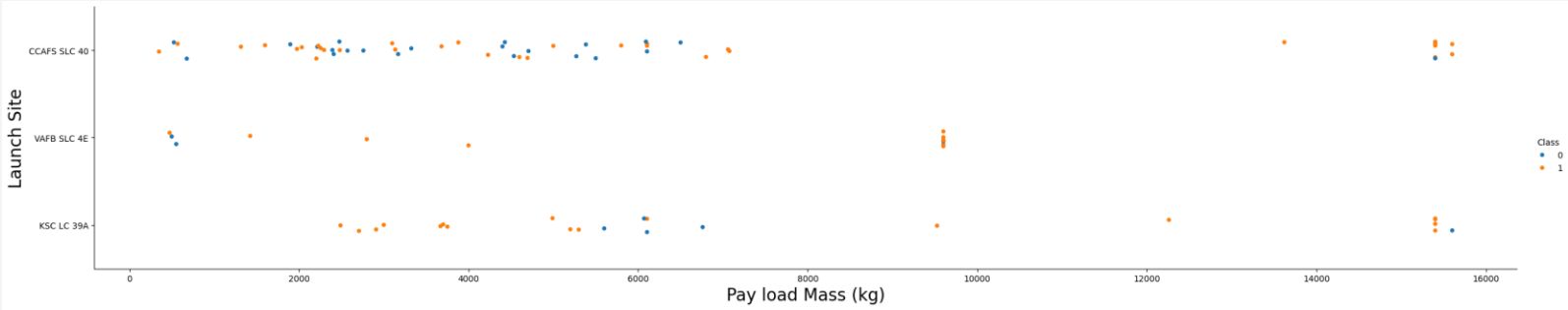
Now try to explain the patterns you found in the Flight Number vs. Launch Site scatter point plots.

Higher Flight Numbers correlate with better landing success across all sites. While

- CCAFS SLC-40: Mixed results, but success improves over time.
- VAFB SLC-4E: High success rate, especially in later flights.
- KSC LC-39A: Consistent success, particularly in higher Flight Numbers.

Payload vs. Launch Site

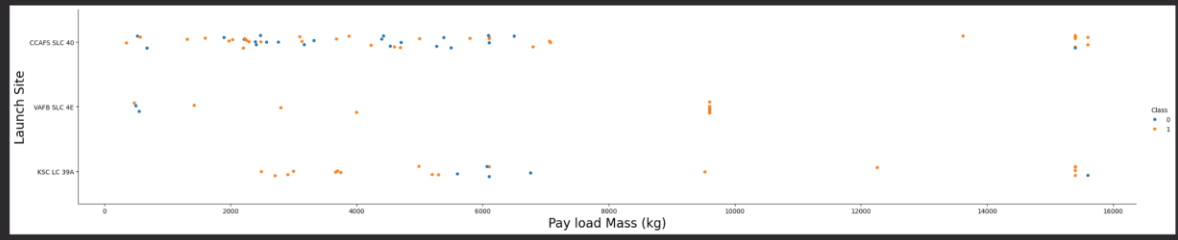
- Show a scatter plot of Payload vs. Launch Site



- Show the screenshot of the scatter plot with explanations

```
sns.catplot(y="LaunchSite", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("Pay load Mass (kg)",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()
```

Python



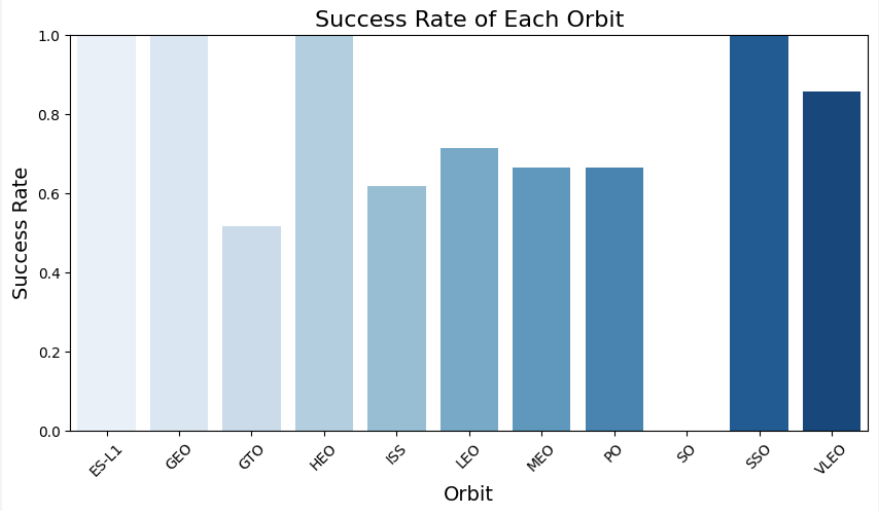
Now if you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

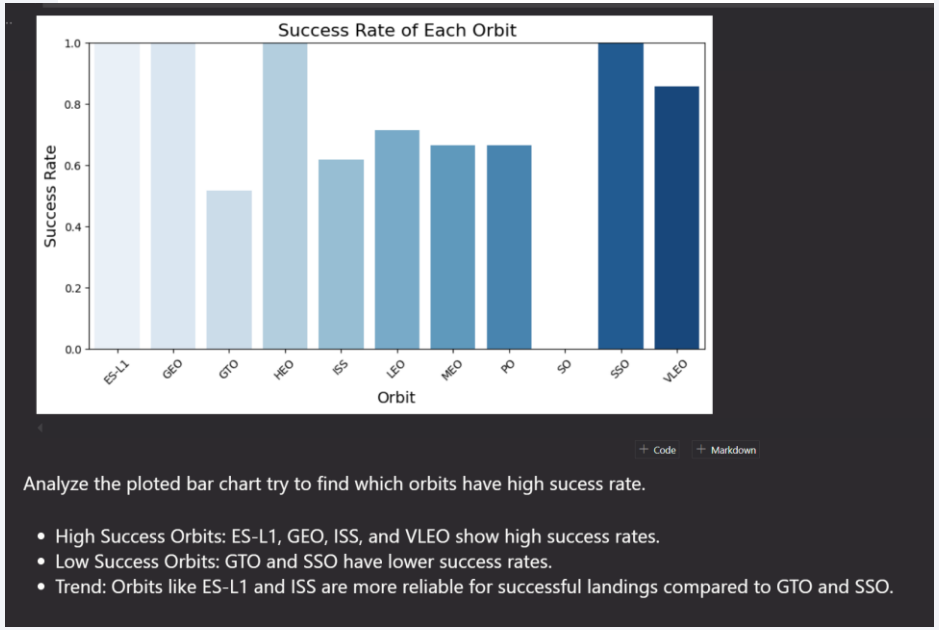
VAFB SLC-4E: No heavy payloads (>10,000 kg). CCAFS SLC-40: Handles all payload sizes, mixed success. KSC LC-39A: High success with heavy payloads. Trend: Heavy payloads launched from CCAFS SLC-40 and KSC LC-39A; VAFB SLC-4E for lighter payloads.

Success Rate vs. Orbit Type

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

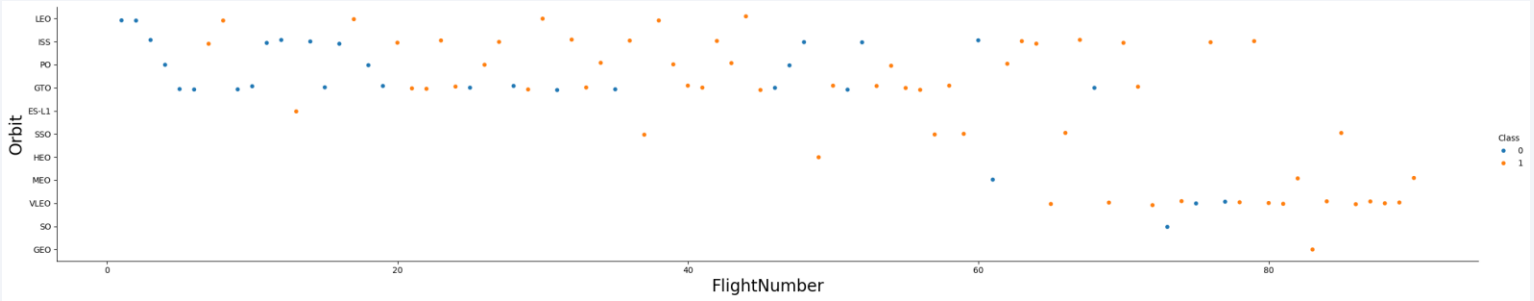
- Show the screenshot of the scatter plot with explanations



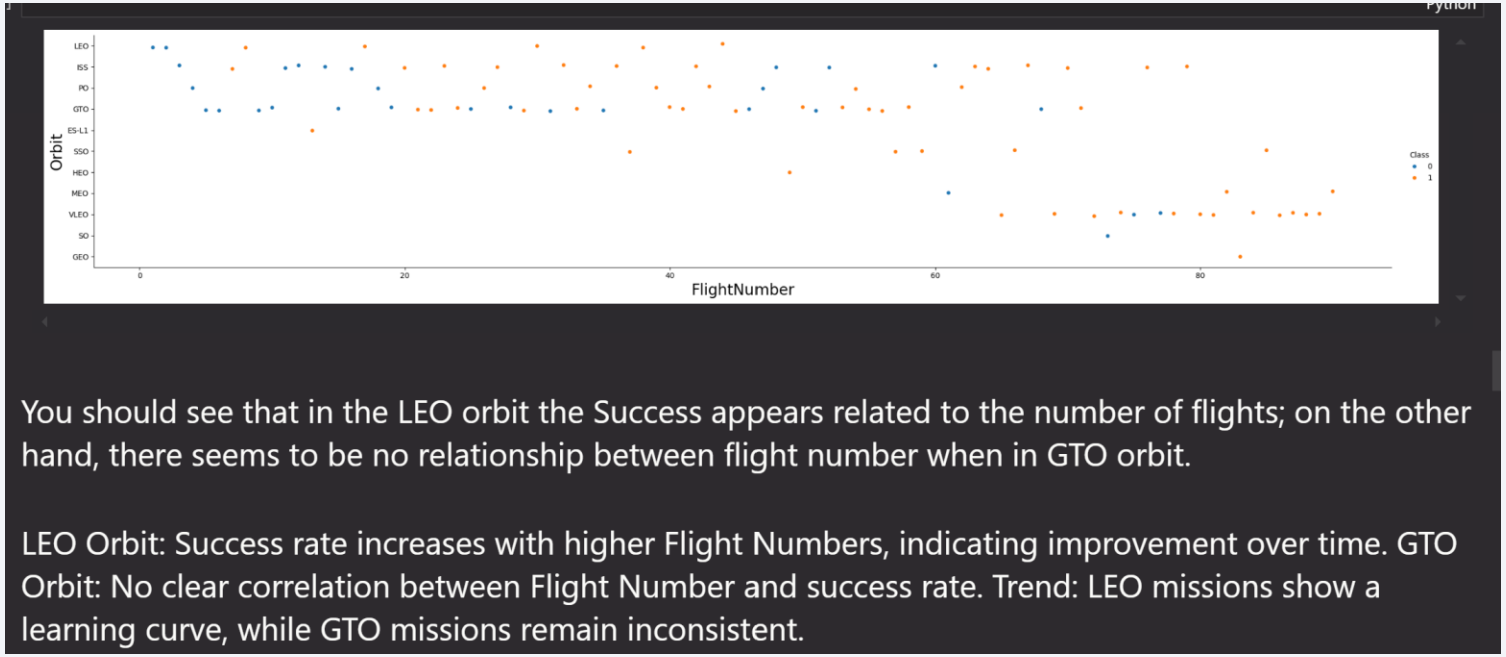


Flight Number vs. Orbit Type

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

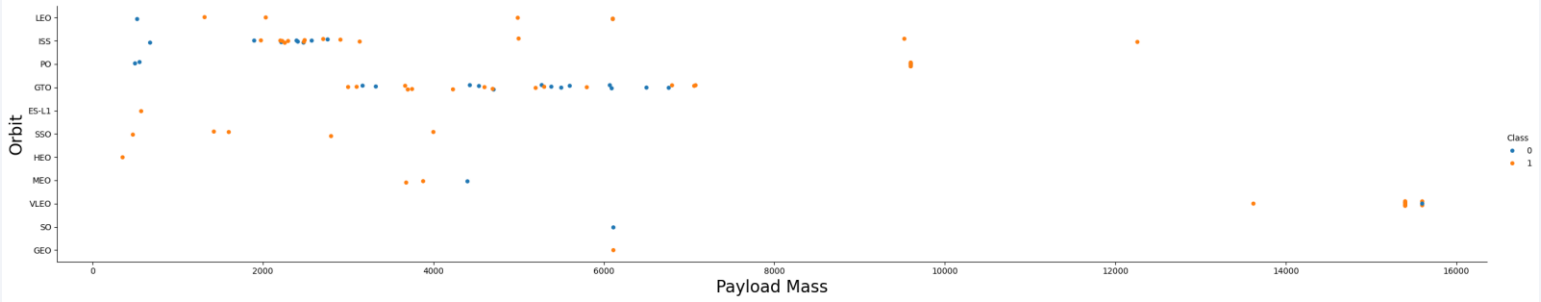


- Show the screenshot of the scatter plot with explanations



Payload vs. Orbit Type

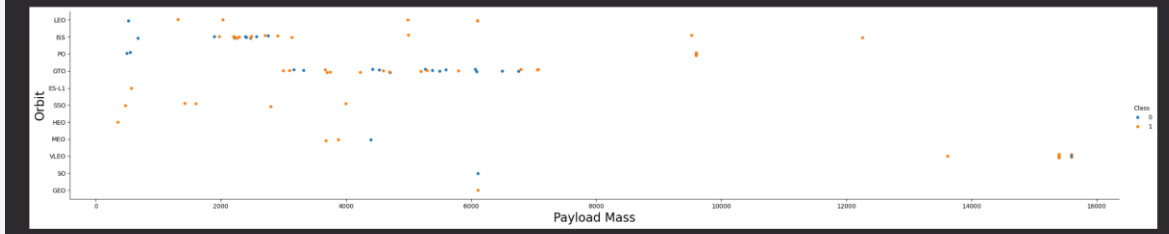
- Show a scatter point of payload vs. orbit type



- Show the screenshot of the scatter plot with explanations

```
sns.catplot(y="Orbit", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("Payload Mass",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```

Python



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

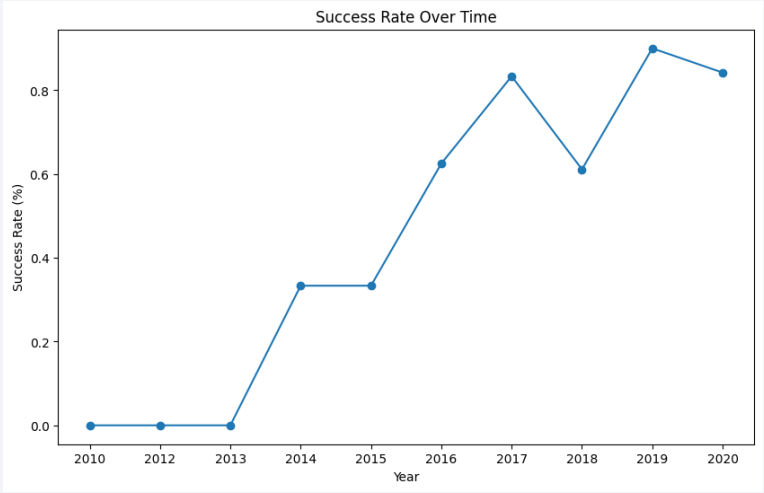
However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

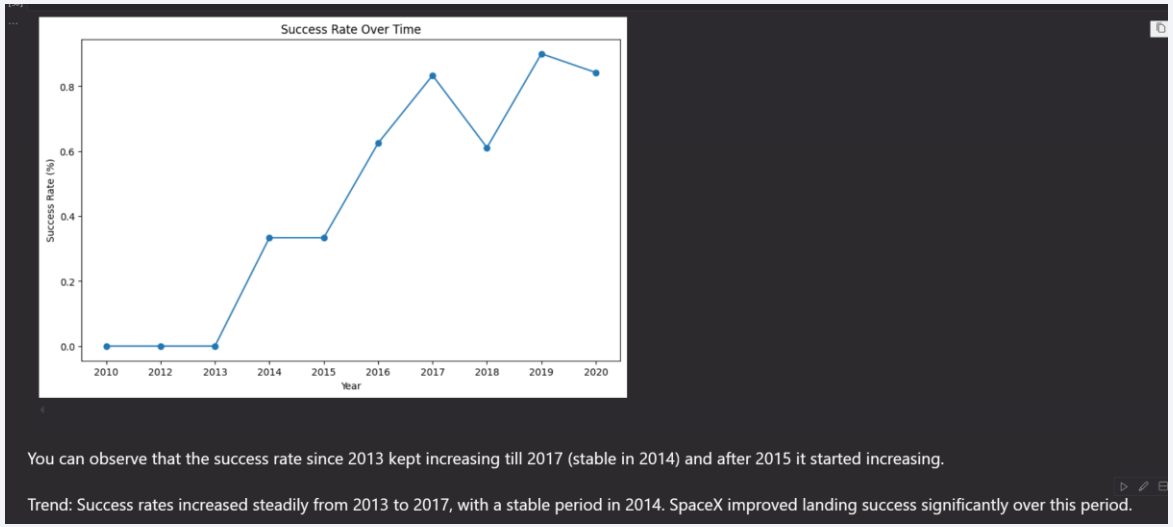
Polar, LEO, ISS: High success rates with heavy payloads. GTO: Mixed success rates, making it harder to distinguish trends. Trend: Heavy payloads are more successful in Polar, LEO, and ISS orbits compared to GTO.

Launch Success Yearly Trend

- Show a line chart of yearly average success rate

- Show the screenshot of the scatter plot with explanations





All Launch Site Names

- Find the names of the unique launch sites
- There are 4 launch sites in this data

```
%sql select DISTINCT Launch_Site from SPACEXTABLE
```

```
15]
 * sqlite:///my_data1.db
Done.
```

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`
- There are first 5 records of CCA in this data

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
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- The total payload carried by boosters from NASA is 45596

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

```
%sql select SUM(PAYLOAD_MASS__KG_) from SPACEXTABLE WHERE Customer = 'NASA (CRS)'
```

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- The average of payload mass is 2534.6666666666665

Display average payload mass carried by booster version 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

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- The first date of successful landing outcome on ground pad is 2015-12-22

```
%sql select min(Date) from SPACEXTABLE WHERE Landing_Outcome == 'Success (ground pad)'
```

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

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

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

```
%sql select DISTINCT Booster_Version from SPACEXTABLE WHERE Landing_Outcome == 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000
```

[23]

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

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
- There are 99 successful mission and 1 failure mission

```
%sql SELECT DISTINCT Mission_Outcome, COUNT(Mission_Outcome) AS outcome_count FROM SPACEXTABLE GROUP BY Mission_Outcome
```

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

Mission_Outcome	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
- There are 12 boosters which have carried the maximum payload mass

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

```
%sql select DISTINCT Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)
```

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

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

- There are 2 failed landing_outcomes in drone ship

```
%sql select substr(Date,6,2) as Month, * from SPACEXTABLE where substr(Date,1,4) = '2015' and Landing_Outcome = 'Failure (drone ship)'
```

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

Month	Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
01	2015-01-10	9:47:00	F9 v1.1 B1012	CCAFS LC-40	SpaceX CRS-5	2395	LEO (ISS)	NASA (CRS)	Success	Failure (drone ship)
04	2015-04-14	20:10:00	F9 v1.1 B1015	CCAFS LC-40	SpaceX CRS-6	1898	LEO (ISS)	NASA (CRS)	Success	Failure (drone ship)

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
- There are 8 Landing_Outcome

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 DISTINCT Landing_Outcome, COUNT(Landing_Outcome) from SPACEXTABLE WHERE Date >= '2010-06-04' AND Date <= '2017-03-20' GROUP BY Landing_Outcome ORDER BY COUNT(Landing_Outcome) DESC
```

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

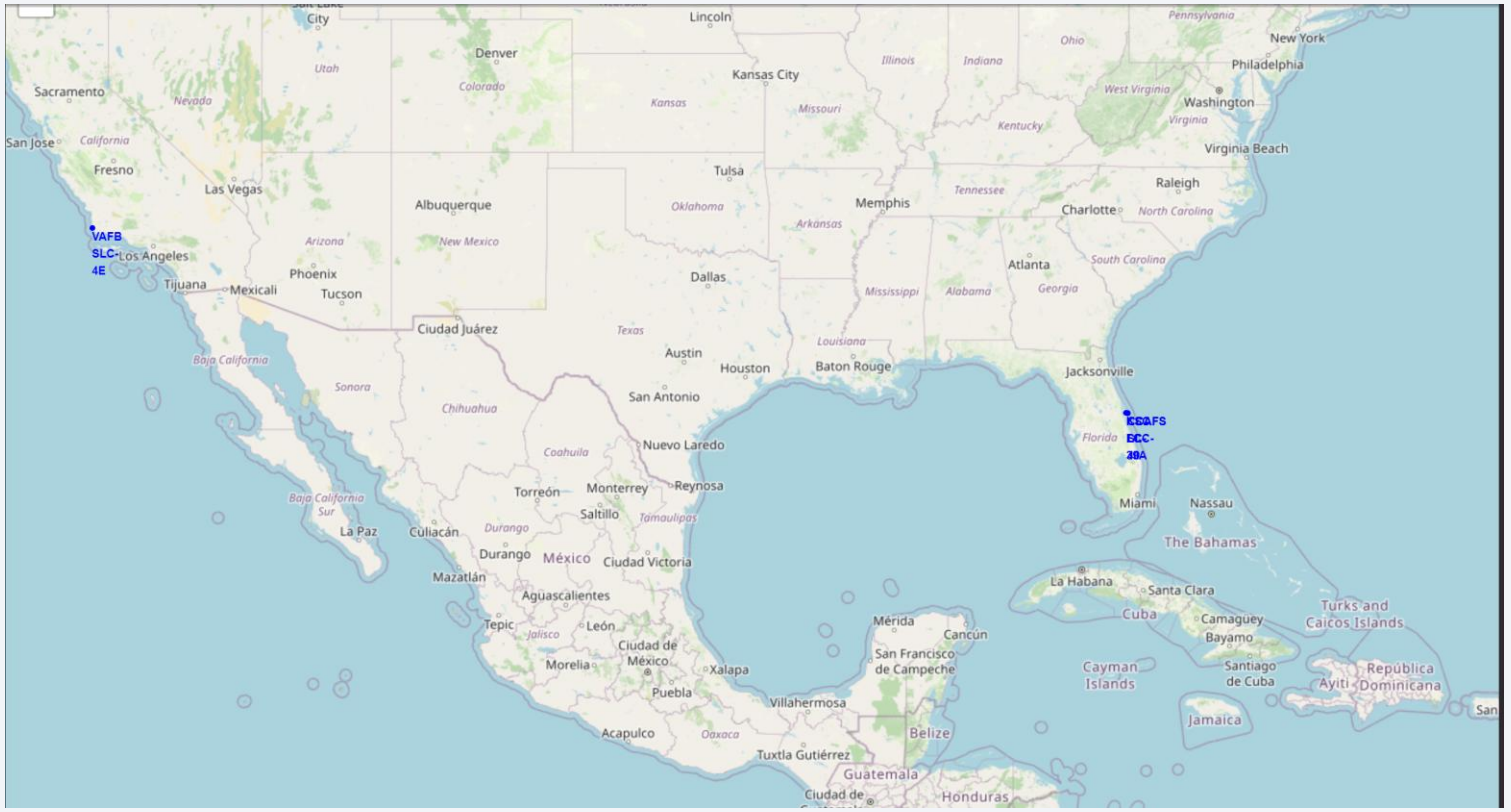
Landing_Outcome	COUNT(Landing_Outcome)
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

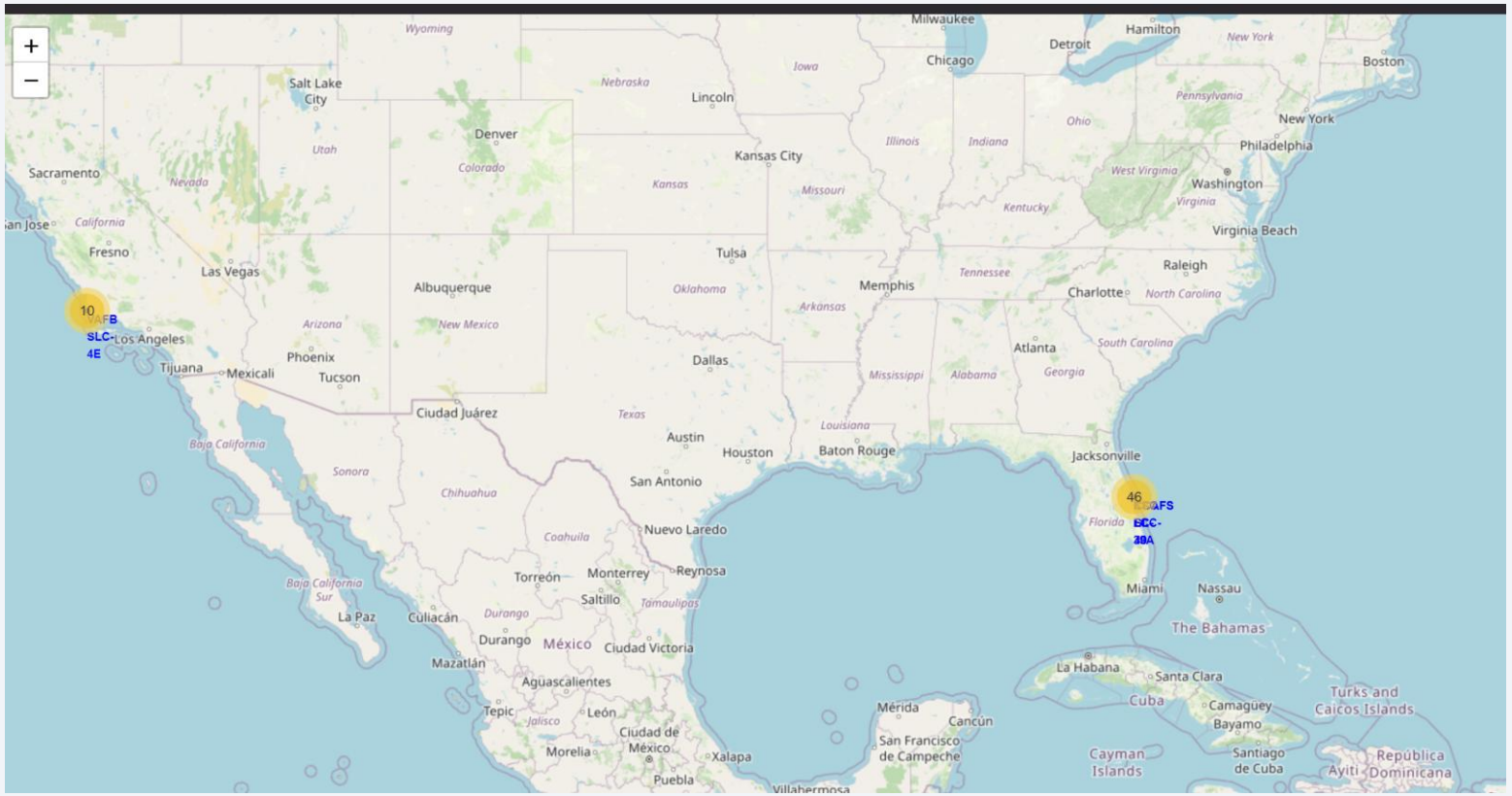
All Launch Sites' Location

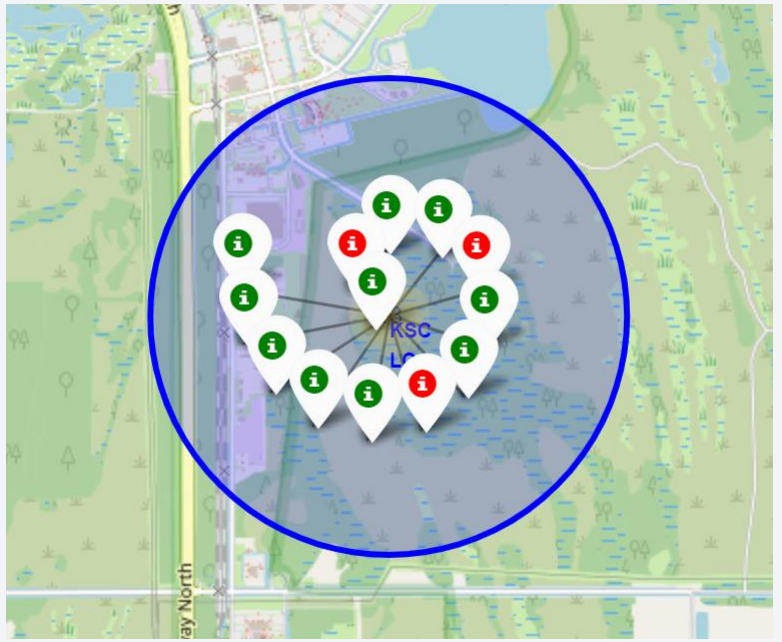
- There is only one launch site on west coast, while 3 sites are on east coast



The Success/Failed Launches

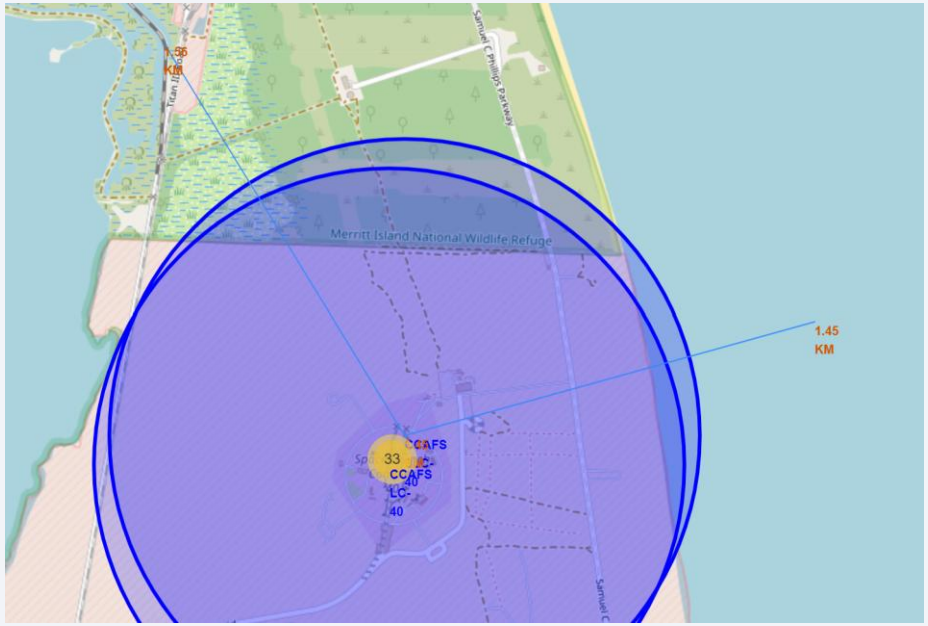
- KSC LC-39A site have the highest success rate of launches.

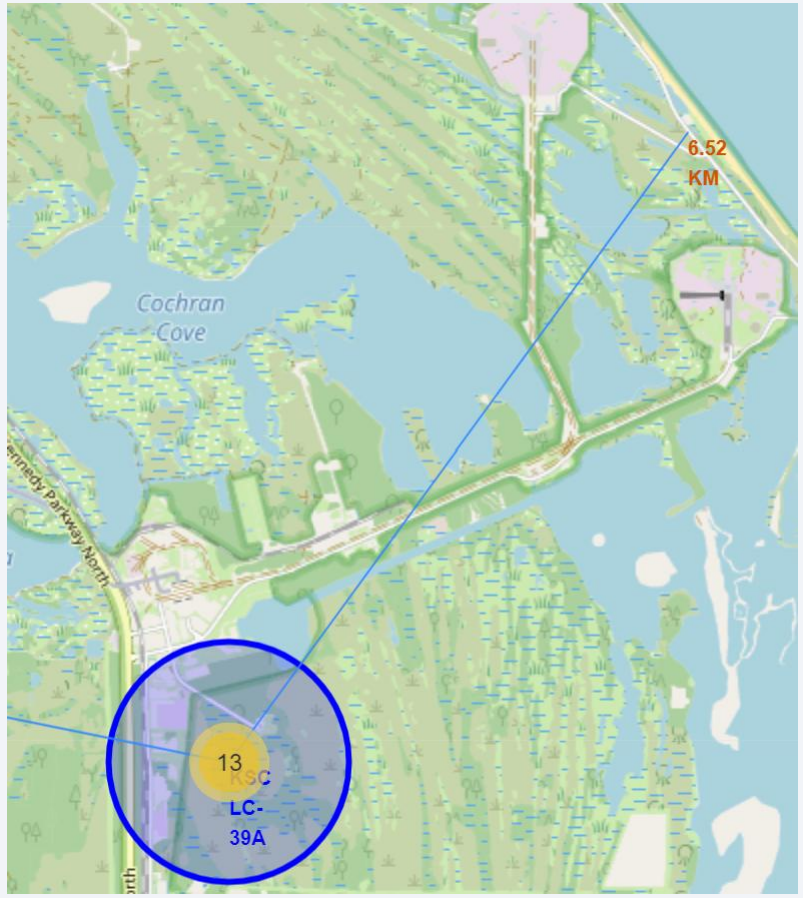


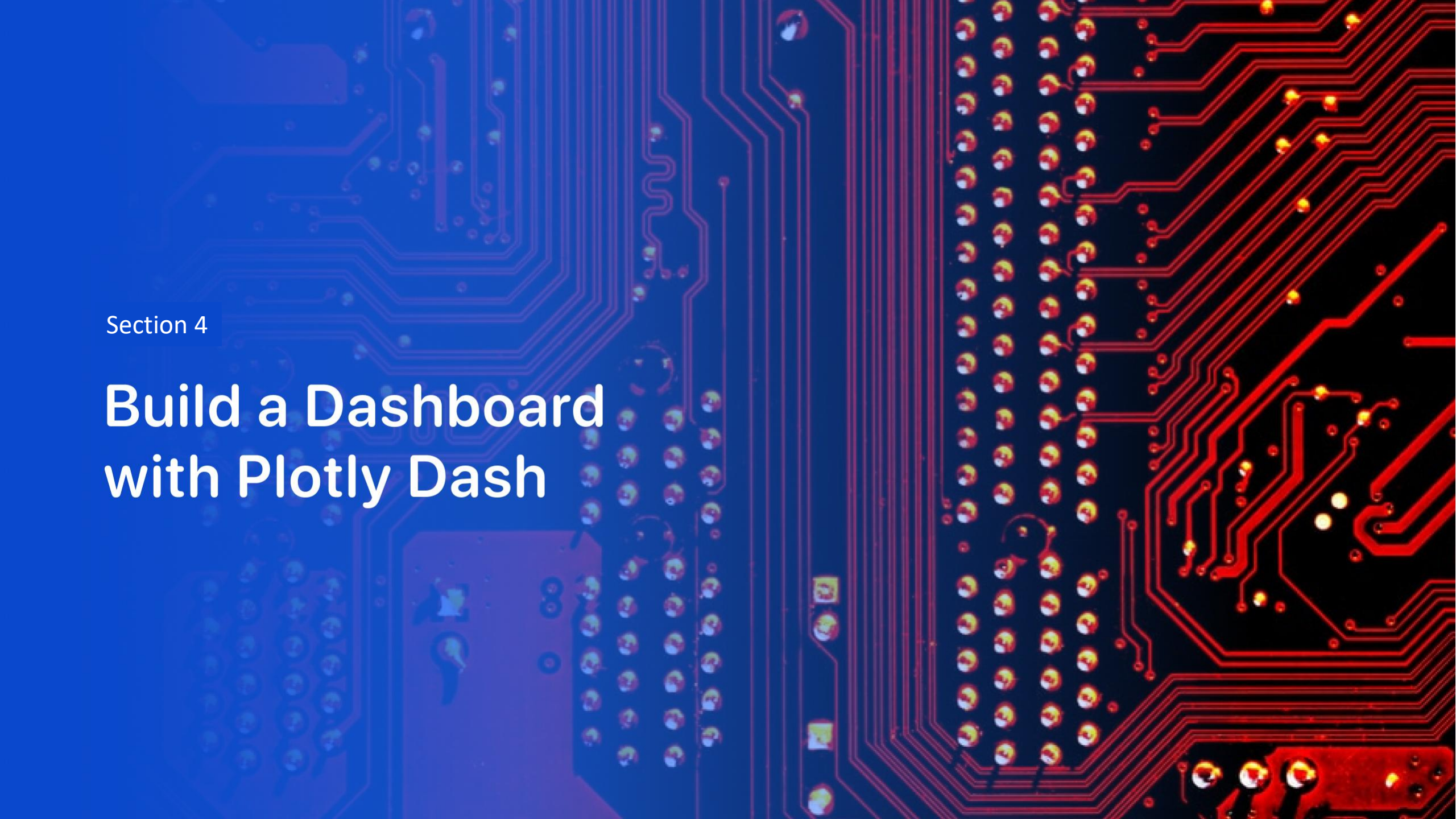


Proximities railway, highway and coastline with distance

- SpaceX launch sites, their success rates, and geographical relationships with key features like coastlines and cities.

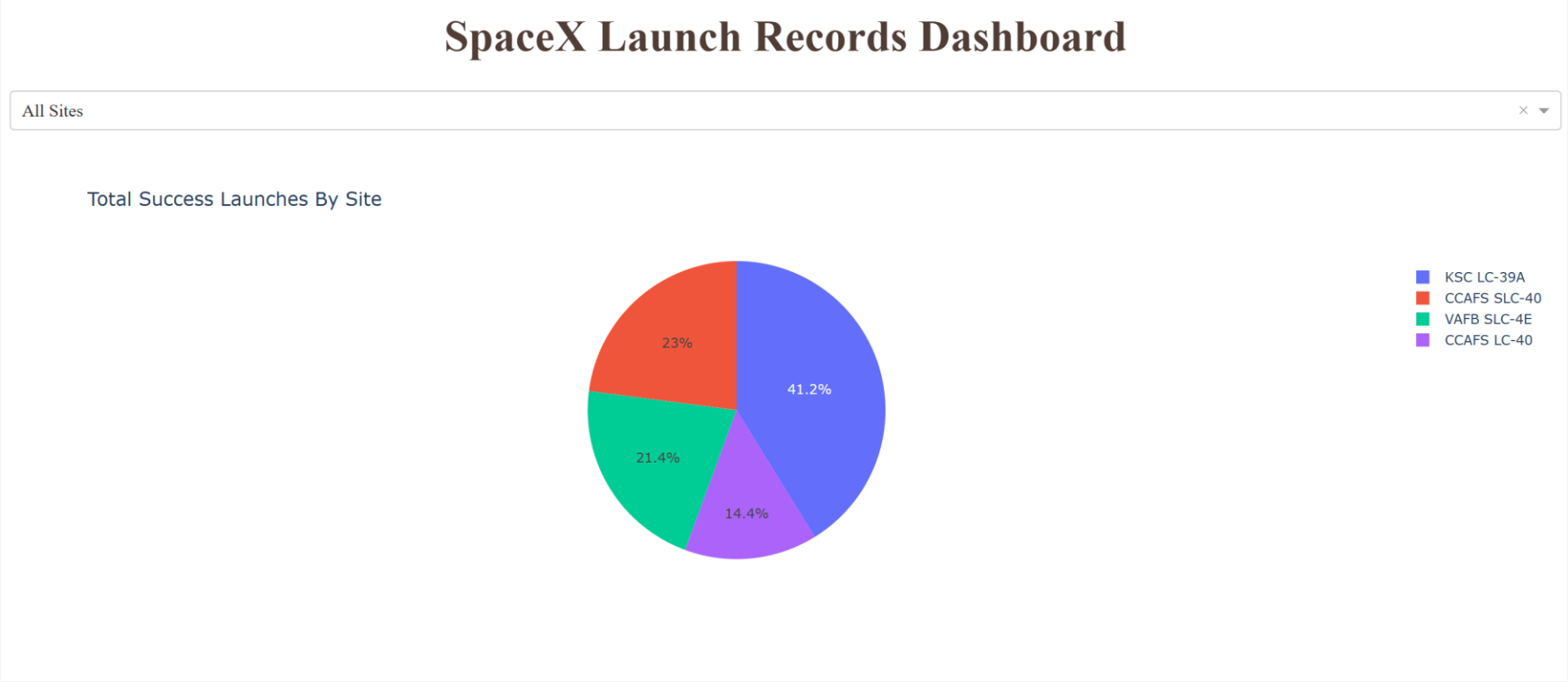





Section 4
Build a Dashboard with Plotly Dash

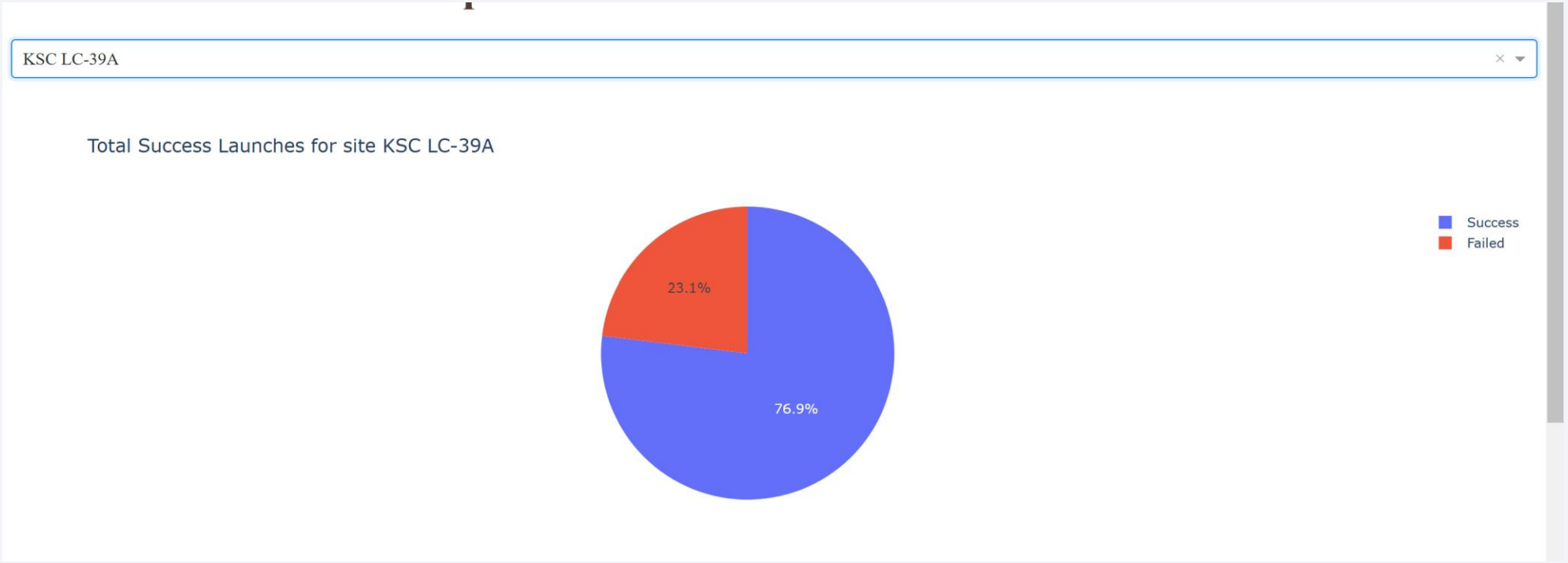
Total Success Launches by Sites

- KSC LC-39A have the highest success rate at 41.2% while lowest score 14.4% come from CCAFS LC-40



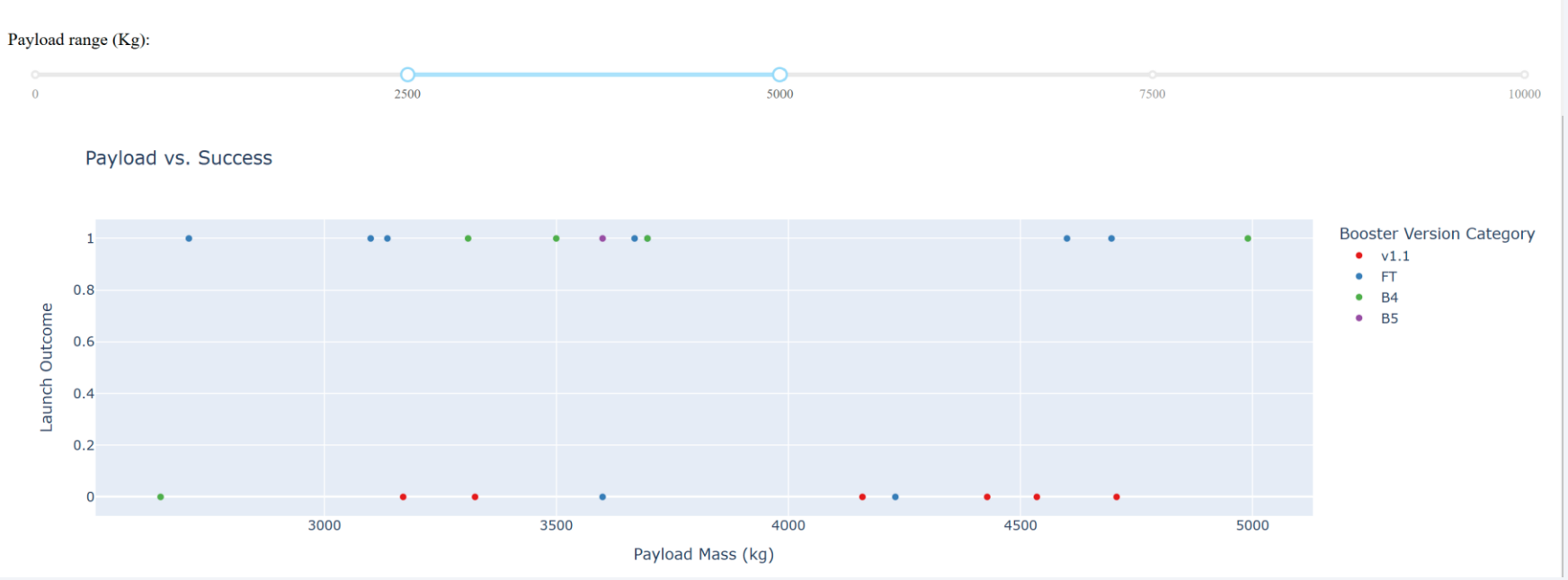
The total success launch for site KSC LC-39A

- The success rate of this site is 76.9%, along with failed rate 23.1%



The success outcome with Booster Version from Payload Mass

- With payload between 2500 and 5000 kg, Booster version v1.1 is likely to be fail.

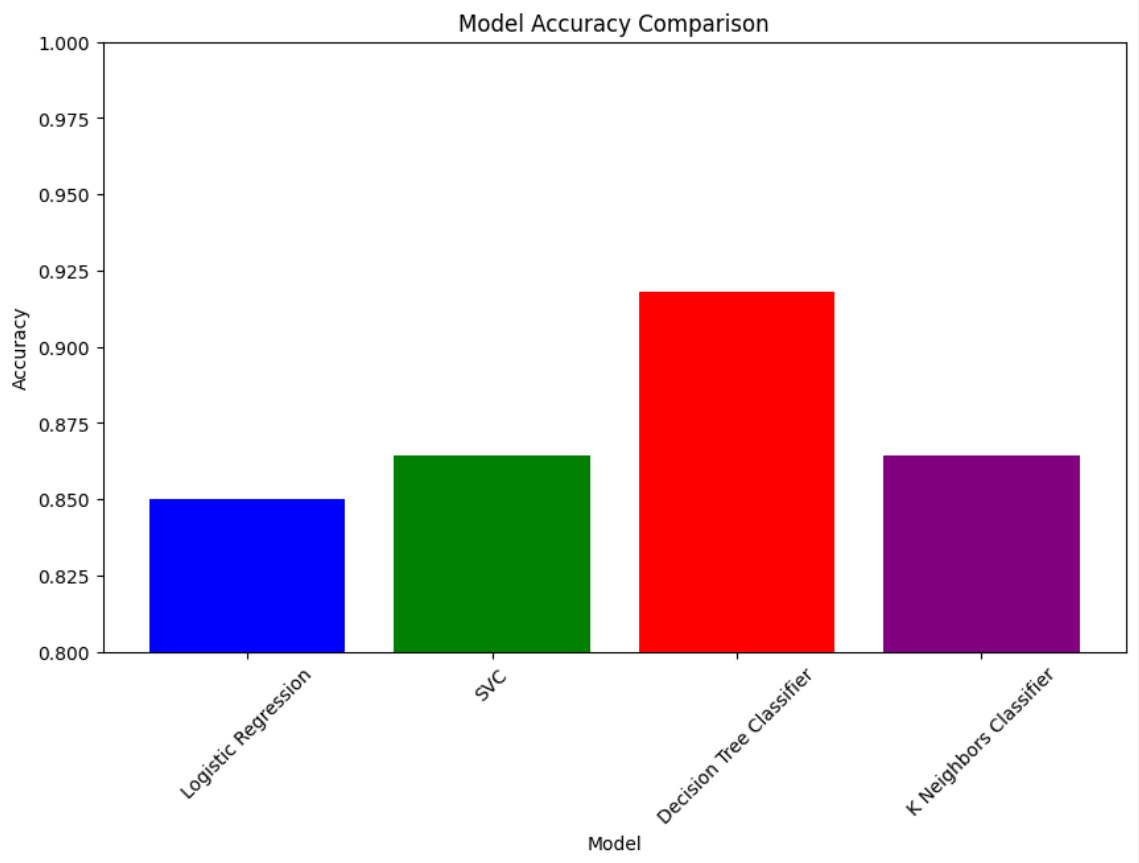


Section 5

Predictive Analysis (Classification)

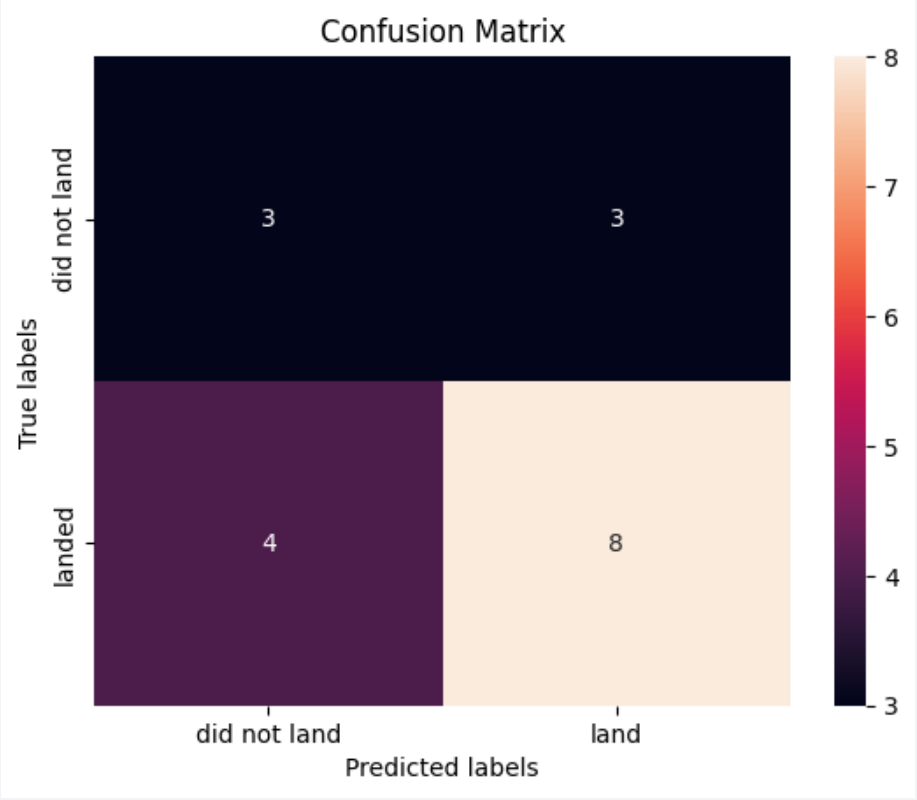
Classification Accuracy

- Decision Tree Classifier is the best model



Confusion Matrix

- The true prediction of landing is 8 out of 12
- While failed landing is 50% correct



Conclusions

- Successful Data AnalysisData was collected via SpaceX API and web scraping.
- Data cleaning and exploratory analysis provided valuable insights into launch outcomes
- Interactive and Predictive Insights:
  - Interactive visualizations (Folium, Plotly Dash) helped explore trends
  - Machine learning models were applied to predict landing outcomes
  - Best Model IdentifiedDecision Tree Classifier performed best with 94.44% test accuracy.Outperformed Logistic Regression, SVM, and KNN in predictive accuracy
- Key Findings:
  - KSC LC-39A had the highest launch success rate
  - Booster version v1.1 showed lower success for payloads between 2500-5000 kg

Appendix

- Create Bar chart for model accuracy

```
result = {'Model': ['Logistic Regression', 'SVC', 'Decision Tree Classifier', 'K Neighbors Classifier'],
          'Accuracy': [0.8500, 0.8643, 0.9179, 0.8643]}
result_df = pd.DataFrame(result)
result_df
# Plot bar chart
plt.figure(figsize=(10, 6))
plt.bar(result_df['Model'], result_df['Accuracy'], color=['blue', 'green', 'red', 'purple'])

# Customize plot
plt.xlabel('Model')
plt.ylabel('Accuracy')
plt.title('Model Accuracy Comparison')
plt.ylim(0.8, 1.0)  # Set y-axis limits for better visualization
plt.xticks(rotation=45)  # Rotate x labels for better readability

# Show plot
plt.show()
```

Thank you!