IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Name>
<Date>

Outline

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

Executive Summary

- Summary of methodologies
- Data Collection through API
- Data Collection with Web Scraping
- Data Wrangling
- Exploratory Data Analysis with SQL
- Exploratory Data Analysis with Data Visualization
- Interactive Visual Analytics with Folium
- Machine Learning Prediction
- Summary of all results
- Exploratory Data Analysis result
- Interactive analytics in screenshots
- Predictive Analytics result

Introduction

- Project background and context

SpaceX advertises Falcon 9 rocket launches at a cost of $62 million, while other providers charge upwards of $165 million per launch. Much of SpaceX's savings come from reusing the first stage of the rocket. Therefore, if we can predict whether the first stage will land successfully, we can estimate the cost of a launch. This information could be useful for another company looking to compete with SpaceX for a rocket launch bid. The goal of this project is to develop a machine learning pipeline to predict the successful landing of the first stage.

- Problems you want to find answers

What factors influence whether the rocket will land successfully?

- The interaction between various features that impact the likelihood of a successful landing.
- The specific operating conditions required to ensure a successful landing program.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API and Web scraping
- Perform data wrangling
  - Summarize and analyze features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data is normalized and divided to training and test sets. Different models were used and compared with variations of parameters

Data Collection

Describe how data sets were collected.

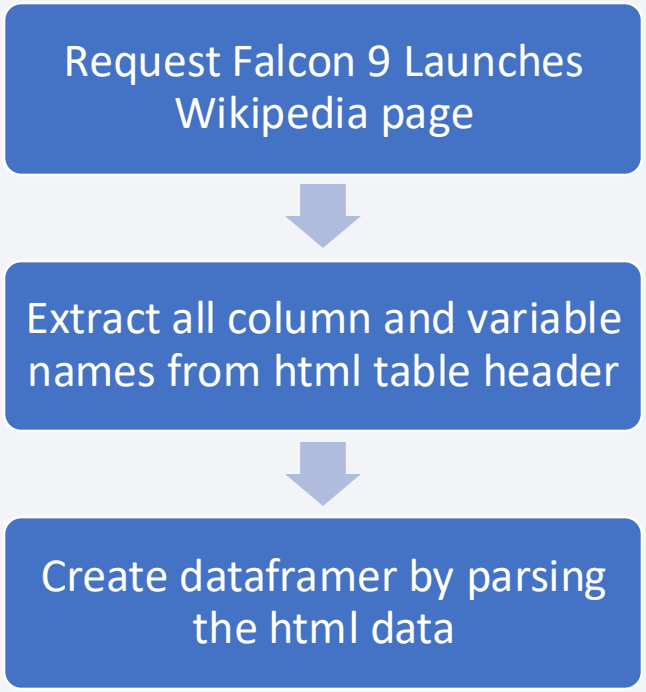
- Datasets were pulled from Space X API and Web-scraped from Wikipedia

Data Collection – SpaceX API

- SpaceX provides a public API that allows data retrieval, which is then processed following the steps outlined in the accompanying flowchart, and the data is stored afterward
- https://github.com/teetri/ibm-data-science-coursera/blob/main/Capstone/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

- Launch data from SpaceX can also be sourced from Wikipedia. Following the steps in the flowchart, the data is downloaded from Wikipedia and stored for future use
- https://github.com/teetri/ibm-data-science-coursera/blob/main/Capstone/jupyter-labs-webscraping.ipynb



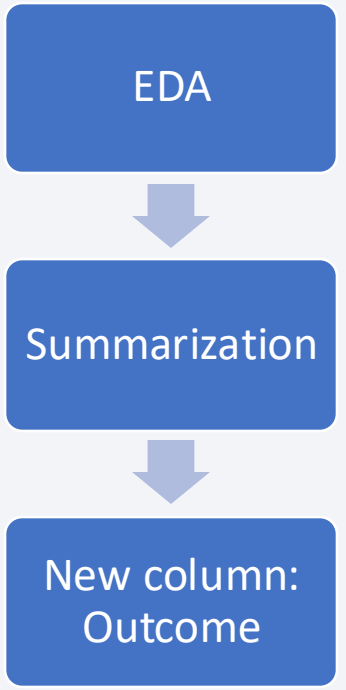
Data Wrangling

An initial Exploratory Data Analysis (EDA) was conducted on the dataset.

• A summary of launches per site, occurrences of each orbit, and mission outcomes per orbit type was calculated.

• Lastly, the landing outcome label was generated from the Outcome column.

- https://github.com/teetri/ibm-data-science-coursera/blob/main/Capstone/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- To explore data, scatterplots and bar plots were used to visualize the relationship between pair of features: Payload Mass X Flight Number, Launch Site X Flight Number, Launch Site X Payload Mass, Orbit and Flight Number, Payload and Orbit
- https://github.com/teetri/ibm-data-science-coursera/blob/main/Capstone/edadataviz.ipynb

EDA with SQL

- The following SQL queries were executed:
  - List of unique launch site names in space missions.
  - Top 5 launch sites starting with 'CCA'.
  - Total payload mass carried by boosters launched by NASA (CRS).
  - Average payload mass carried by the booster version F9 v1.1.
  - Date of the first successful landing outcome on a ground pad.
  - Boosters that successfully landed on a drone ship with payloads between 4000 and 6000 kg.
  - Total count of successful and failed mission outcomes.
  - Booster versions that carried the maximum payload mass.
  - Failed landing outcomes on drone ships in 2015, along with their booster versions and launch sites.
  - Ranking of landing outcomes (e.g., Failure on droneship or Success on ground pad) between June 4, 2010, and March 20, 2017.
- https://github.com/teetri/ibm-data-science-coursera/blob/main/Capstone/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Folium Maps utilized markers, circles, lines, and marker clusters:
  - Markers were used to represent locations, such as launch sites.
  - Circles highlighted areas around specific coordinates, like the NASA Johnson Space Center.
  - Marker clusters grouped events occurring at the same location, such as multiple launches from a launch site.
  - Lines connected coordinates to show distances between two points.
- https://github.com/teetri/ibm-data-science-coursera/blob/main/Capstone/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

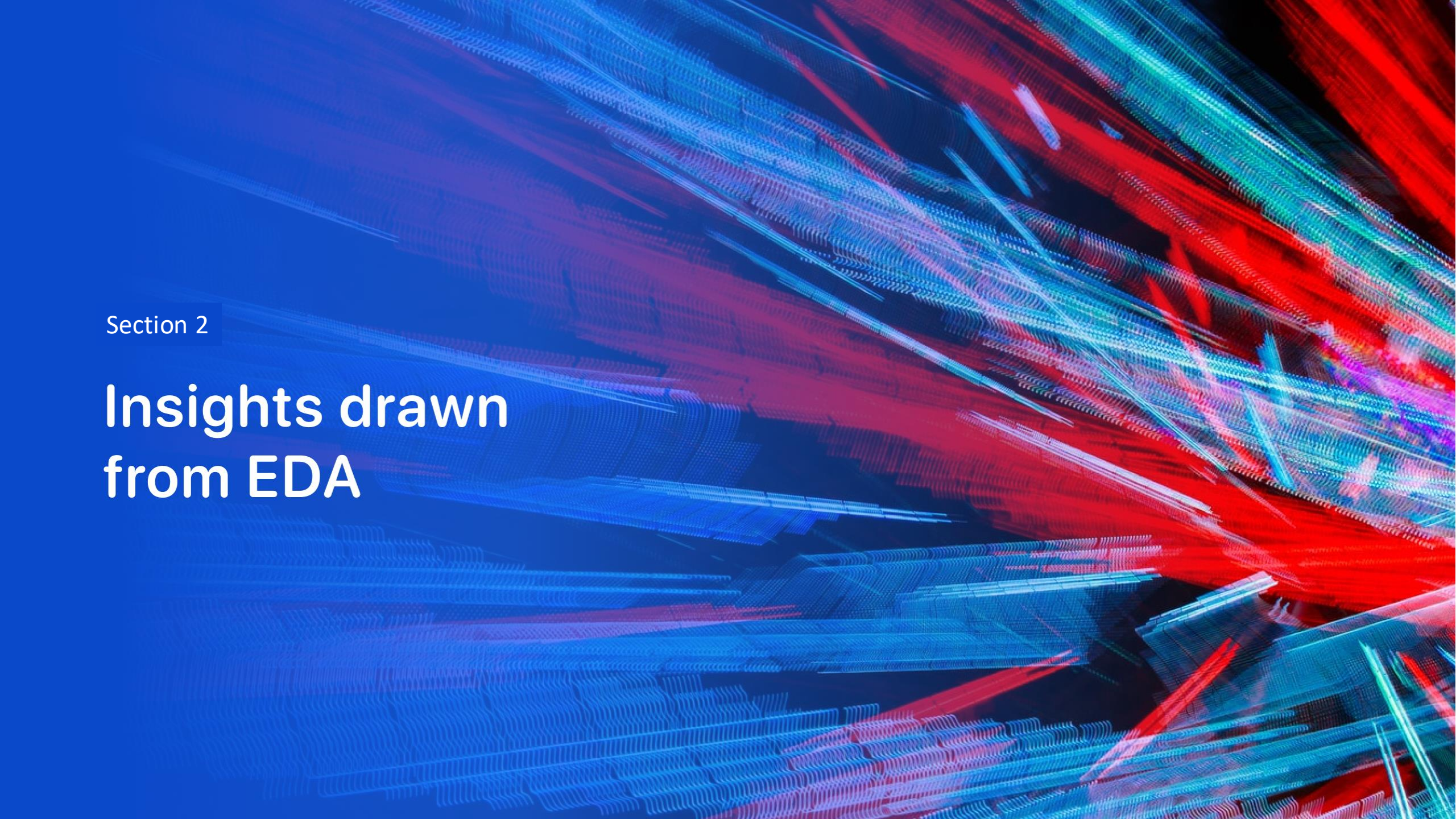
- developed an interactive dashboard using Plotly Dash:
  - Pie charts were created to display the total number of launches at specific sites.
  - A scatter plot was used to illustrate the relationship between Outcome and Payload Mass (Kg) across different booster versions.
- https://github.com/teetri/ibm-data-science-coursera/blob/main/Capstone/spacex_dash_app.py

Predictive Analysis (Classification)

- We loaded the data using NumPy and Pandas, transformed it, and then split it into training and testing sets.
  • Various machine learning models were built and hyperparameters were tuned using GridSearchCV.
  • Accuracy was used as the performance metric, and the model was refined through feature engineering and algorithm adjustments.
  • The best-performing classification model was identified.
- https://github.com/teetri/ibm-data-science-coursera/blob/main/Capstone/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

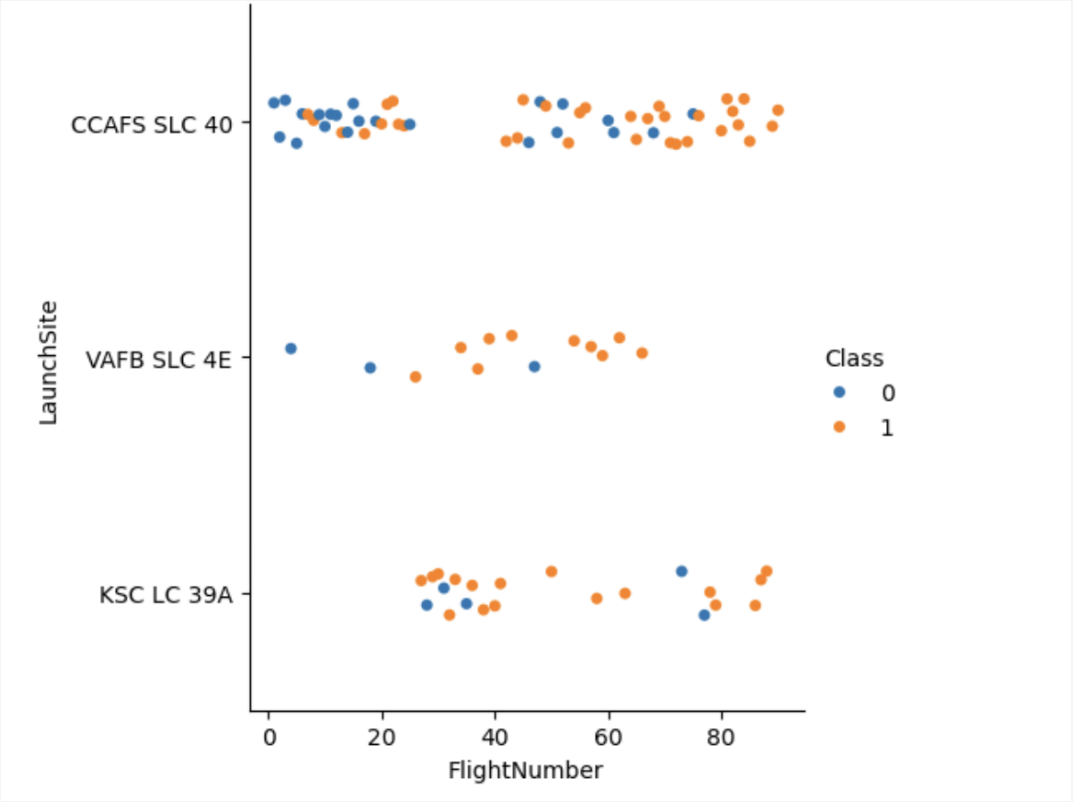
Results

- Exploratory Data Analysis results:
  - SpaceX operates from 4 distinct launch sites.
  - The initial launches were conducted for SpaceX itself and NASA.
  - The average payload for the F9 v1.1 booster is 2,928 kg.
  - The first successful landing outcome occurred in 2015, five years after the first launch.
  - Several Falcon 9 booster versions succeeded in landing on drone ships with payloads above the average.
  - Nearly 100% of mission outcomes were successful.
  - In 2015, two booster versions, F9 v1.1 B1012 and F9 v1.1 B1015, failed to land on drone ships.
  - Landing outcomes improved over the years.

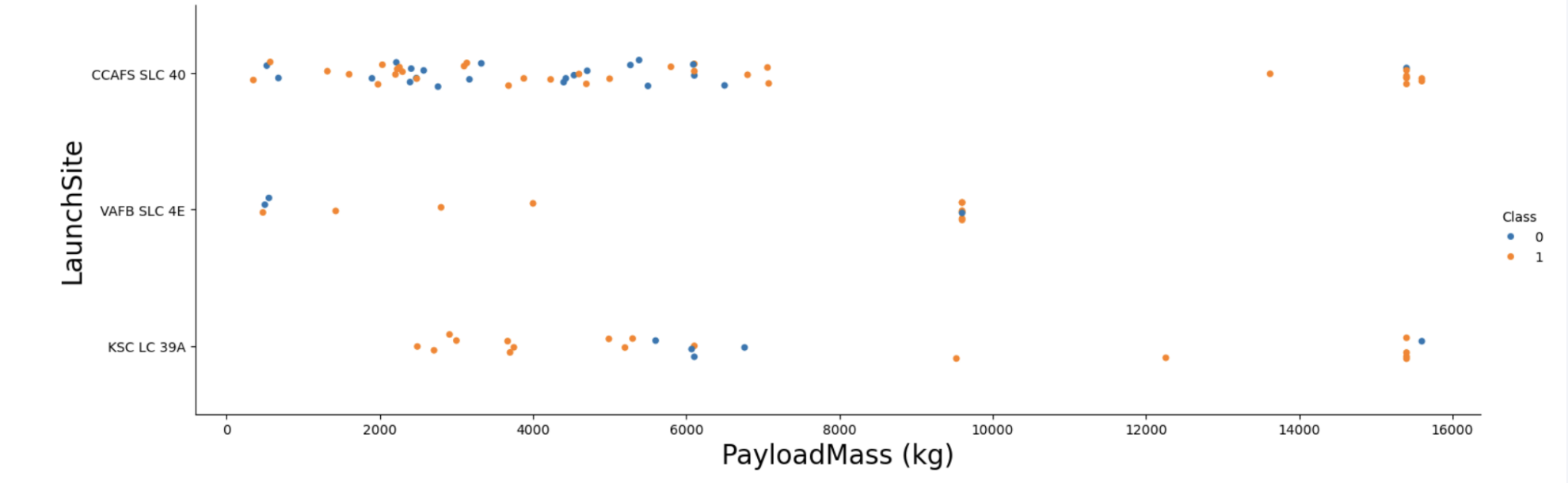

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

- The plot revealed that a higher number of flights at a launch site is associated with a greater success rate at that site.

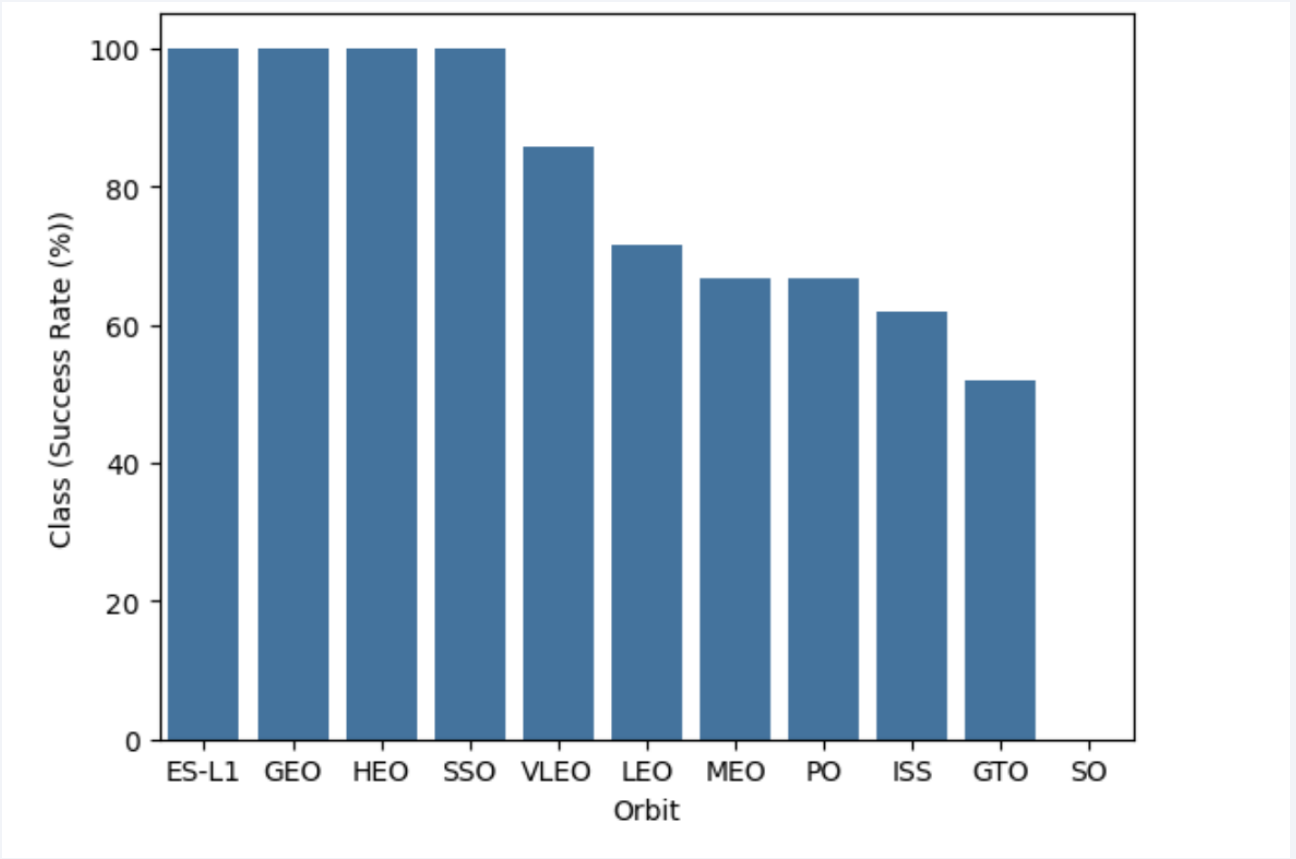


Payload vs. Launch Site



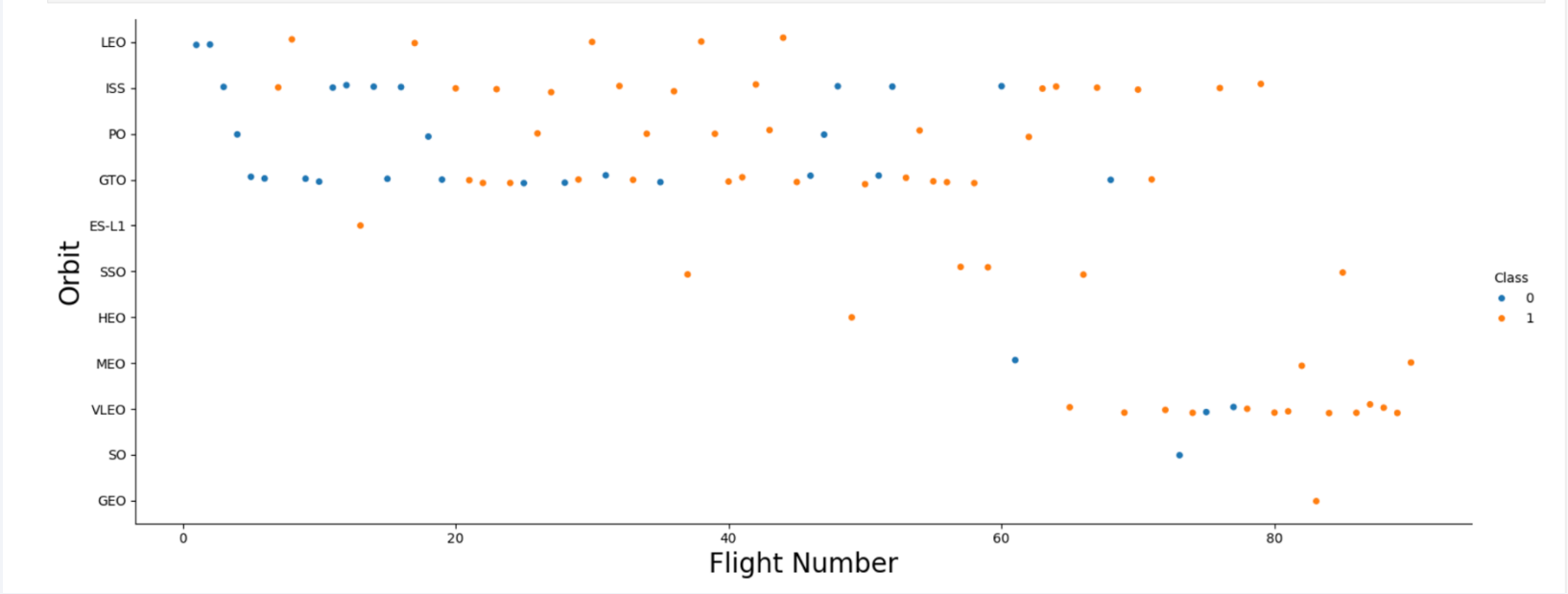
Success Rate vs. Orbit Type

- From the plot, we can see that ES-L1, GEO, HEO, SSO, VLEO had the most success rate.



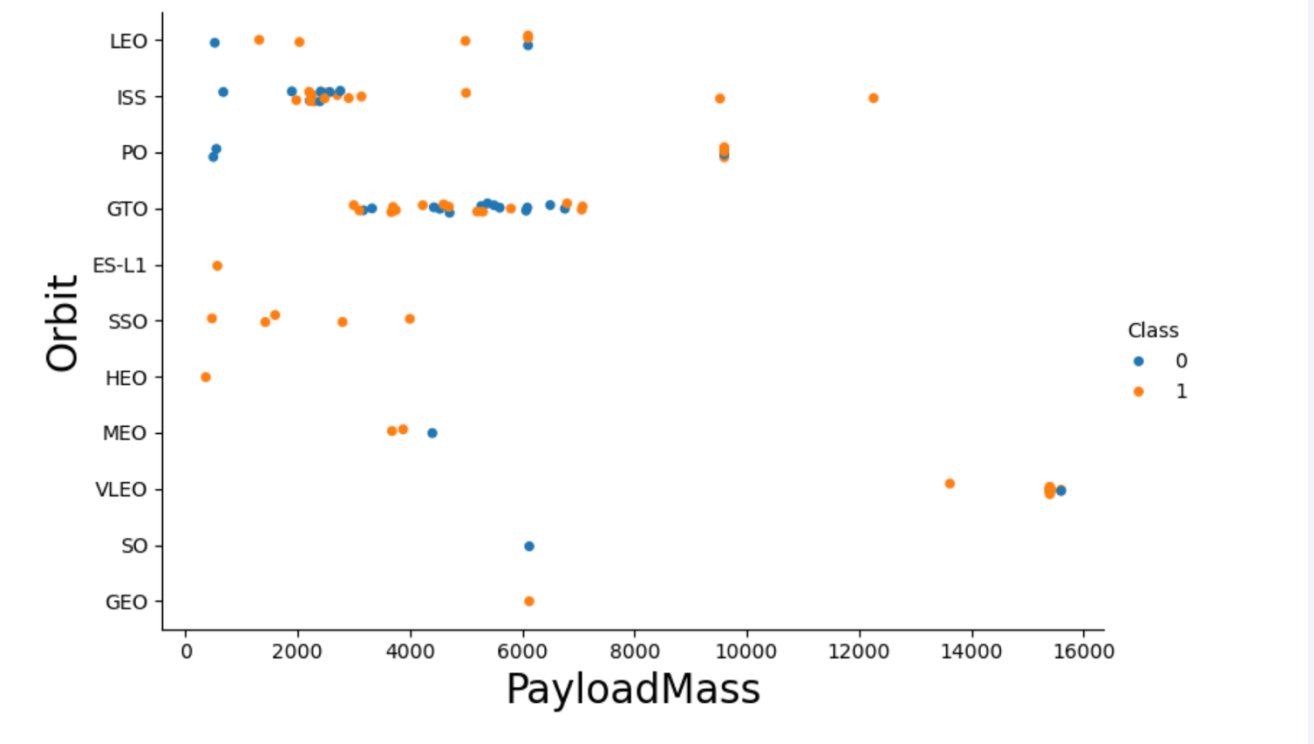
Flight Number vs. Orbit Type

- The plot below displays Flight Number versus Orbit Type. It shows that for the LEO orbit, success is correlated with the number of flights, whereas for the GTO orbit, there is no discernible relationship between flight number and succes



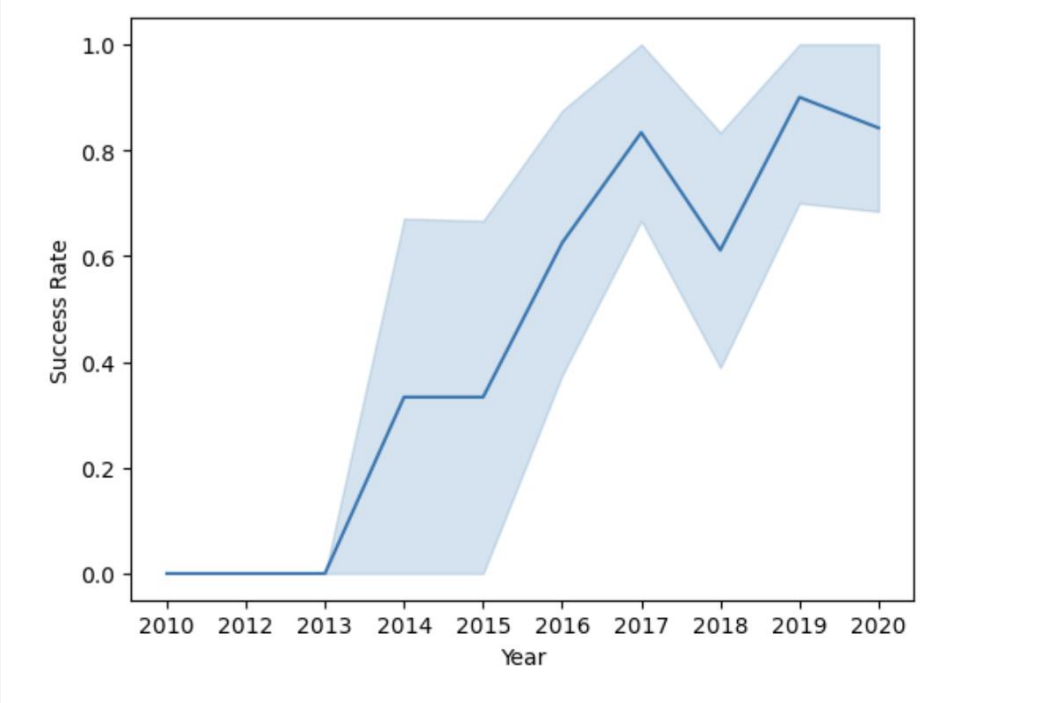
Payload vs. Orbit Type

- It is observed that heavier payloads are associated with a higher rate of successful landings for PO, LEO, and ISS orbits.



Launch Success Yearly Trend

- The plot shows that the success rate has been steadily increasing from 2013 up to 2020.



All Launch Site Names

- The keyword DISTINCT was used to display only the unique launch sites from the SpaceX data.

```
10]: %sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE;
 * sqlite:///my_data1.db
Done.
```

10]:

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

Launch Site Names Begin with 'CCA'

- The query was used to retrieve 5 records where the launch sites start 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
2010-06-04	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

- The total payload carried by NASA boosters was calculated as 45,596 using the query provided.

Task 3

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

```
%sql select sum(PAYLOAD_MASS__KG_) as TotalMass from SPACEXTABLE where Customer = 'NASA (CRS)'
```

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

TotalMass
45596

Average Payload Mass by F9 v1.1

- The average payload mass for the booster version F9 v1.1 was calculated as 2,928.4 kg.

Task 4

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

```
[21]: %sql select avg(PAYLOAD_MASS__KG_) as payloadmass from SPACEXTBL where Booster_Version = 'F9 v1.1';
 * sqlite:///my_data1.db
Done.
[21]:
```

payloadmass
2928.4

First Successful Ground Landing Date

- The date of the first successful landing outcome on a ground pad was observed to be December 22, 2015.

Task 5

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

Hint:Use min function

```
25]: %sql select min(DATE) from SPACEXTBL where Landing_Outcome = 'Success (ground pad)';
 * sqlite:///my_data1.db
Done.
```

25]:

min(DATE)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The WHERE clause was used to filter boosters that successfully landed on a drone ship, and the AND condition was applied to select those with a payload mass greater than 4,000 kg but less than 6,000 kg.

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

16]:
```
%sql select BOOSTER_VERSION from SPACEXTBL where LANDING_OUTCOME='Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000
```

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

16]:

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

Total Number of Successful and Failure Mission Outcomes

- The wildcard '%' was used in the WHERE clause to filter for records where the MissionOutcome was either a success or a failure.

Task 7

List the total number of successful and failure mission outcomes

[30]:
```
%sql select count(MISSION_OUTCOME) as success from SPACEXTBL where Mission_Outcome like 'Success%';
```

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

[30]:

success
100

[31]:
```
%sql select count(MISSION_OUTCOME) as failure from SPACEXTBL where Mission_Outcome like 'Failure%';
```

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

[31]:

failure
1

Boosters Carried Maximum Payload

- We identified the booster that carried the maximum payload by using a subquery in the WHERE clause along with the MAX() function.

Task 8

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

[18]:
```
%sql select BOOSTER_VERSION as boosterversion from SPACEXTBL where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTBL);
```

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

[18]:

boosterversion
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

- We used a combination of the WHERE clause, LIKE, AND, and BETWEEN conditions to filter for failed landing outcomes on drone ships, along with their booster versions and launch site names for the year 2015.

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.

```
[42]: %sql SELECT LANDING_OUTCOME,BOOSTER_VERSION,LAUNCH_SITE, date FROM SPACEXTBL where Landing_Outcome like 'Failure (Drone ship)'
```

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

[42]:

Landing_Outcome	Booster_Version	Launch_Site	Date
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	2015-01-10
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	2015-04-14

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

- We selected landing outcomes and their counts from the data, using the WHERE clause to filter outcomes between June 4, 2010, and March 20, 2017.
- • The GROUP BY clause was used to group the landing outcomes, and the ORDER BY clause sorted these groups in descending order.

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.

```
nding_outcome) FROM SPACEXTBL WHERE "DATE" BETWEEN '2010-06-04' AND '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

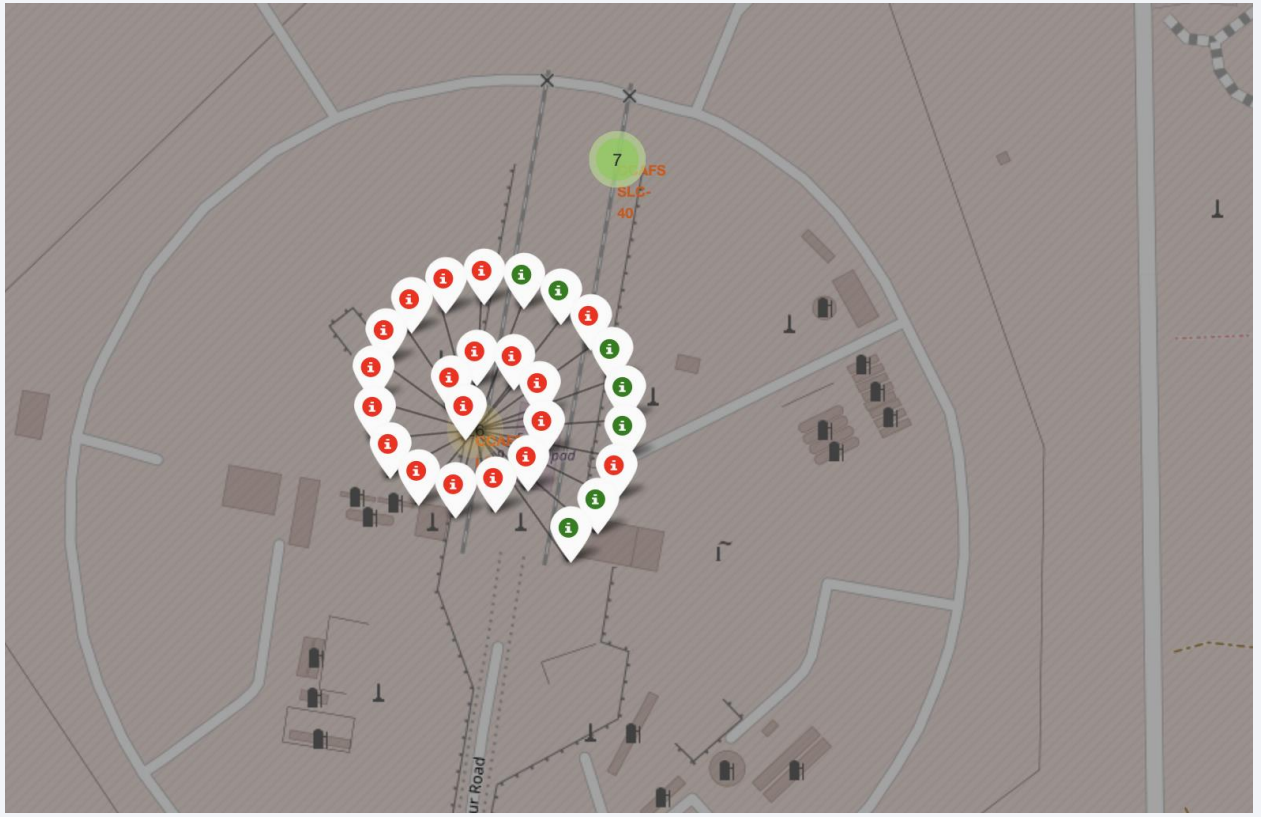
Launch Sites Proximities Analysis

All launch site global map

- All launch sites are in USA

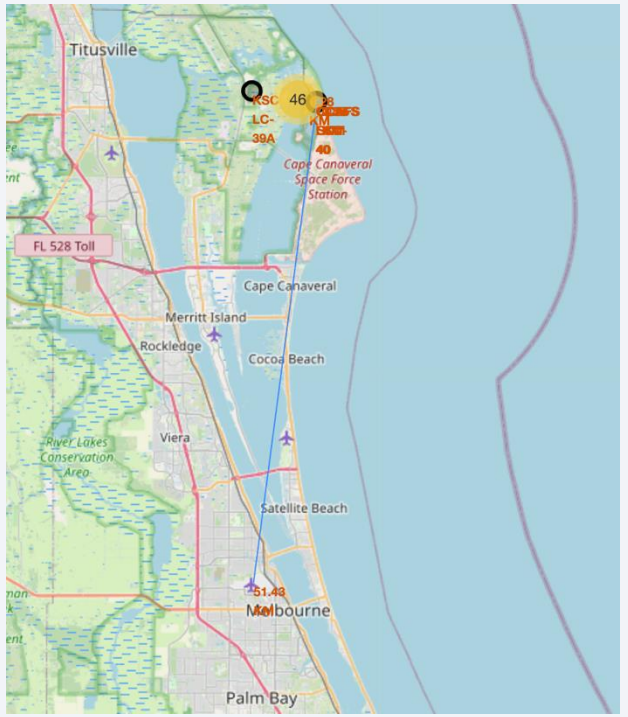
Launch sites with color labels

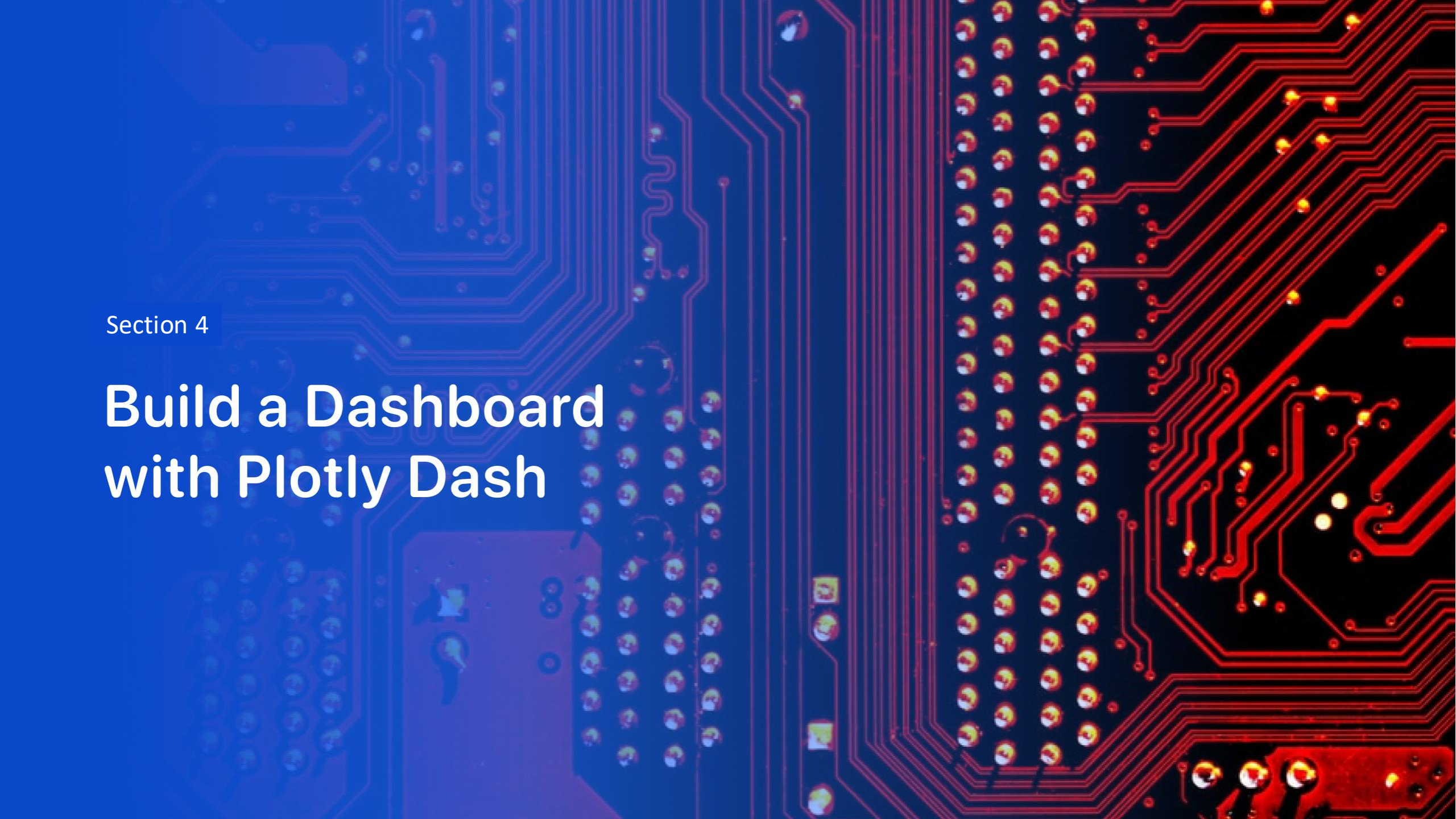
- Green-Success Red-Failure



Launch sites distance to landmarks

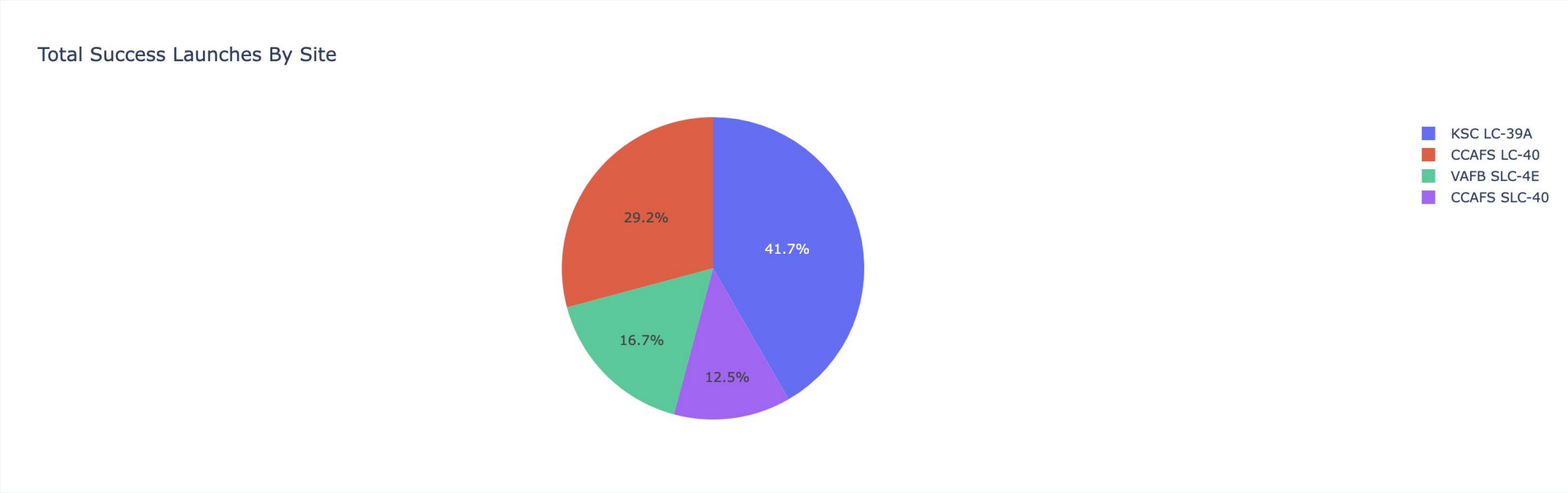
- We can see distance line from launch site to coastline, airport




Section 4
Build a Dashboard with Plotly Dash

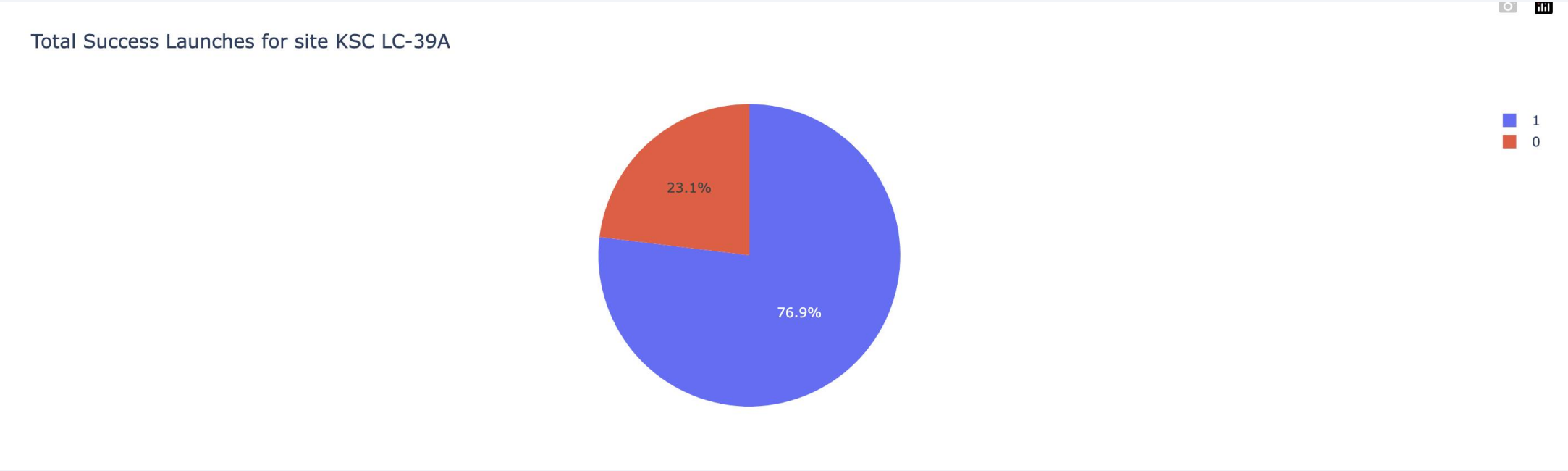
Pie chart showing the success percentage achieved by each launch site

- KSC LC-29A has the most successful launch

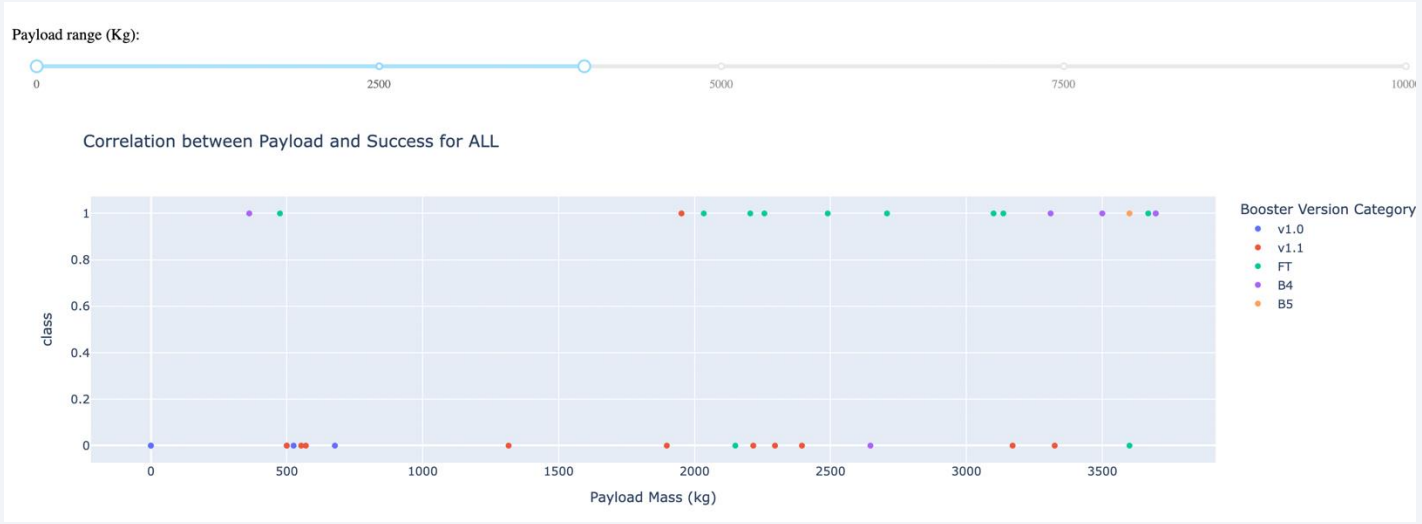


Pie chart showing the Launch site with the highest launch success ratio

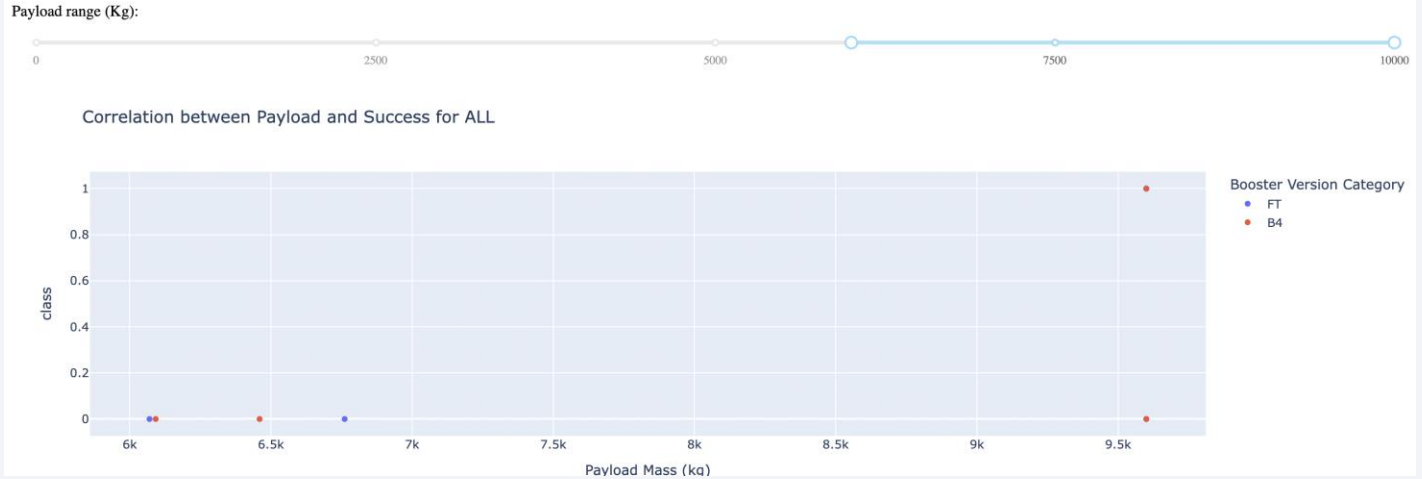
- KSC LC-39A has 76.9% success rate



Scatter plot of Payload vs Launch Outcome for all sites, with different payload selected



- 0-4000kg



- 6000-10000kg

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Decision Tree is the model with highest accuracy

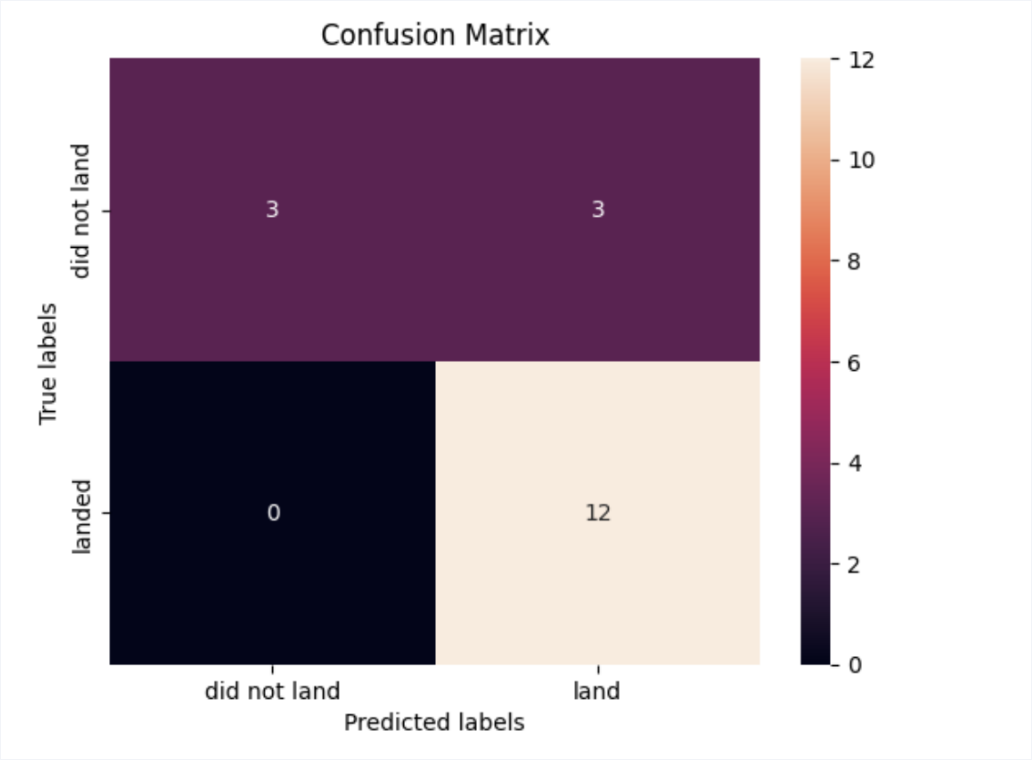
Find the method performs best:

```
# After comparing accuracy of above methods, they all preformed practically
# the same, except for tree which fit train data slightly better but test data worse.
models = {'LogisticRegression':logreg_cv.best_score_,
          'SupportVectorMachine': svm_cv.best_score_,
          'DecisionTree':tree_cv.best_score_,
          'KNeighbours':knn_cv.best_score_
         }
bestalgorithm = max(models, key=models.get)
print('Best model is', bestalgorithm,'with a score of', models[bestalgorithm])
if bestalgorithm == 'LogisticRegression':
    print('Best params is :', logreg_cv.best_params_)
if bestalgorithm == 'SupportVectorMachine':
    print('Best params is :', svm_cv.best_params_)
if bestalgorithm == 'DecisionTree':
    print('Best params is :', tree_cv.best_params_)
if bestalgorithm == 'KNeighbours':
    print('Best params is :', knn_cv.best_params_)
```

```
Best model is DecisionTree with a score of 0.8767857142857143
Best params is : {'criterion': 'gini', 'max_depth': 2, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples
_split': 10, 'splitter': 'best'}
```

Confusion Matrix

- The confusion matrix for the decision tree classifier indicates that the classifier is able to differentiate between different classes. However, a significant issue is the false positives, where unsuccessful landings are incorrectly classified as successful.



Conclusions

We can conclude that:

- A higher number of flights at a launch site is associated with a greater success rate at that site.
- The launch success rate began increasing in 2013 and continued to rise until 2020.
- The orbits ES-L1, GEO, HEO, SSO, and VLEO had the highest success rates.
- KSC LC-39A had the most successful launches among all sites.
- The Decision Tree classifier is the most effective machine learning algorithm for this task.

Appendix

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

Thank you!