IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sushil Kumar Mondal
August 13, 2025

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
  - Interactive Visual Analytics with Plotly Dash
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive Analytics result

Introduction

- Project background and context

  SpaceX lists Falcon 9 rocket launches on its website at a cost of $62 million, compared to over $165 million for other providers. Much of this cost advantage comes from SpaceX's ability to reuse the rocket's first stage. Therefore, predicting whether the first stage will land successfully allows us to estimate launch costs. Such information could be valuable for alternative companies looking to compete with SpaceX in bidding for launch contracts. The goal of this project is to develop a machine learning pipeline that predicts the likelihood of a successful first-stage landing.

- Problems you want to find answers
  - What factors determine if the rocket will land successfully?
  - The interaction amongst various features that determine the success rate of a successful landing.
  - What operating conditions needs to be in place to ensure a successful landing program.

Section 1

Methodology

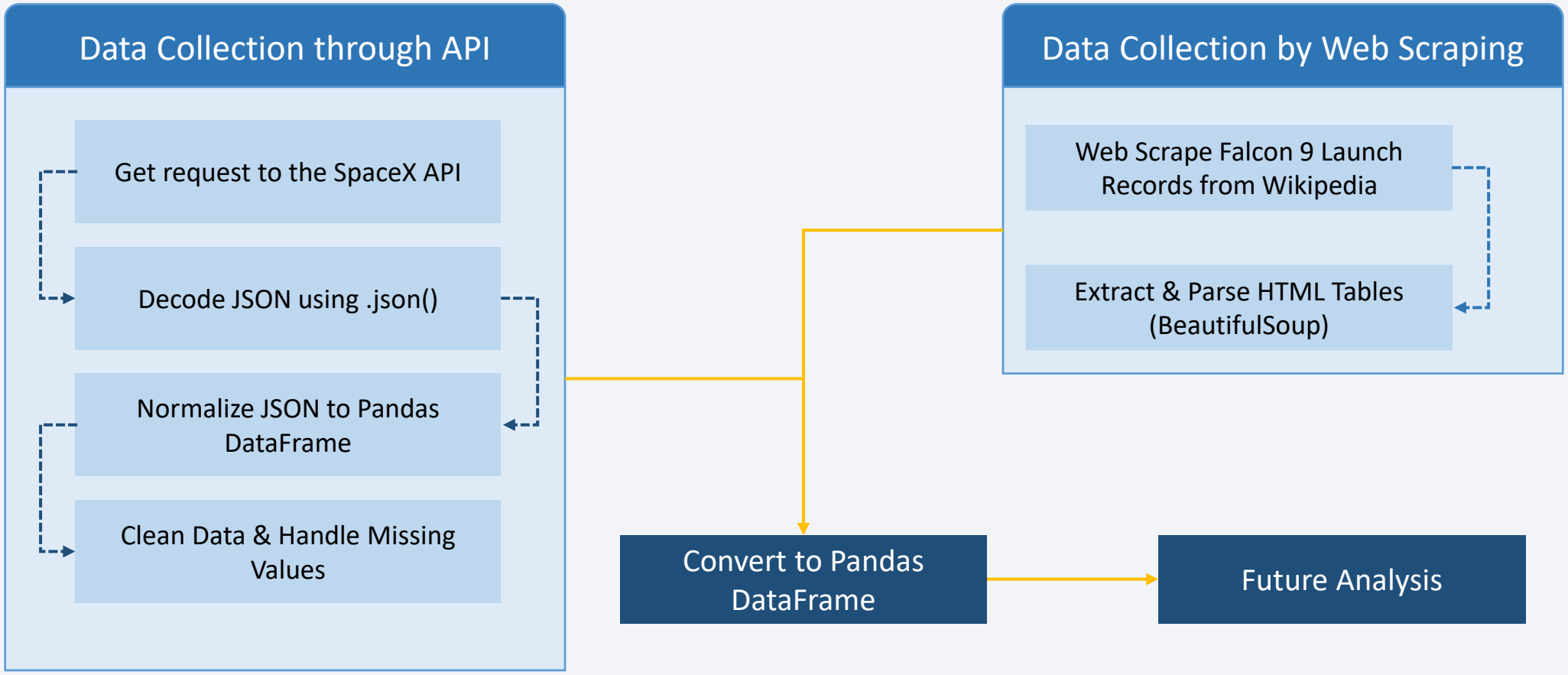
Methodology

Executive Summary

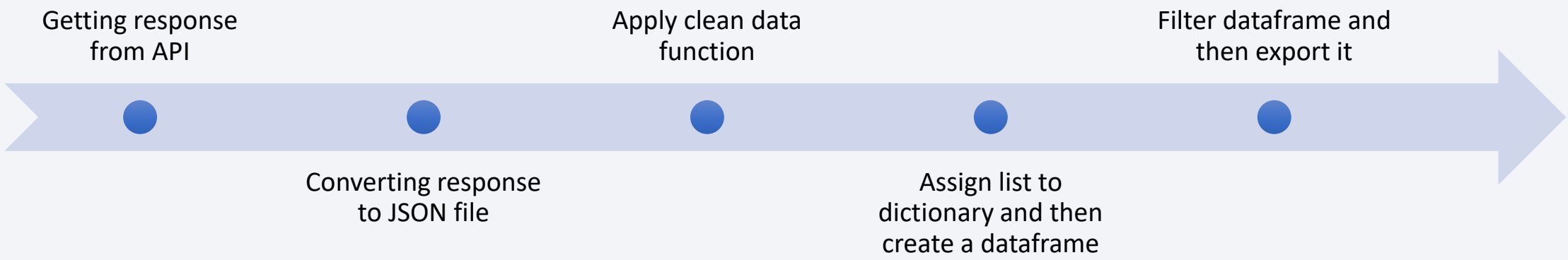
- Data collection methodology:
  - Data was collected using SpaceX API and web scraping from Wikipedia.
- Perform data wrangling
  - One-hot encoding was applied to categorical features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Built by Logistic Regression, Decision Tree Classifier, SVC, and KNN
  - Tuned by using GridSearchCV
  - Evaluated by confusion matrix

Data Collection

- Describe how data sets were collected.



Data Collection – SpaceX API



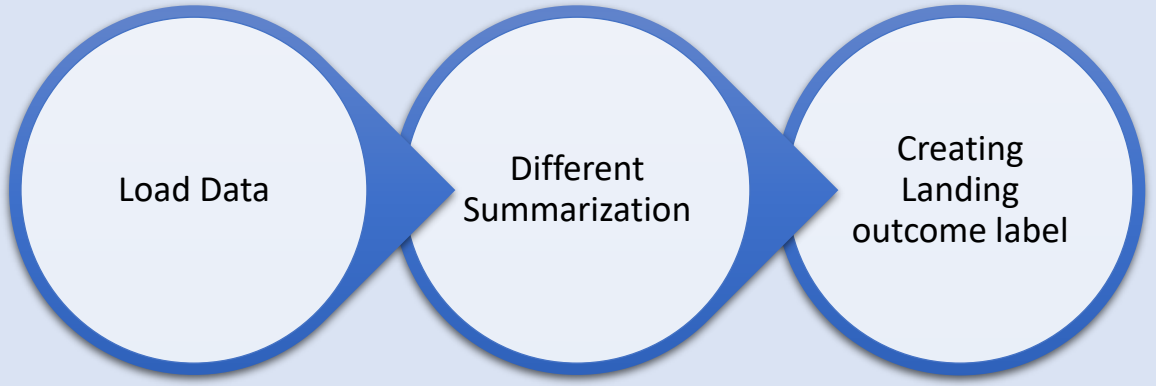
GitHub Link: Applied-Data-Science-Capstone-Coursera/Complete the Data Collection API Lab.ipynb at main · MacLover1984/Applied-Data-Science-Capstone-Coursera

Data Collection - Scraping

GitHub URL: Applied-Data-Science-Capstone-Coursera/Complete the Data Collection with Web Scraping lab.ipynb at main · MacLover1984/Applied-Data-Science-Capstone-Coursera

Data Wrangling

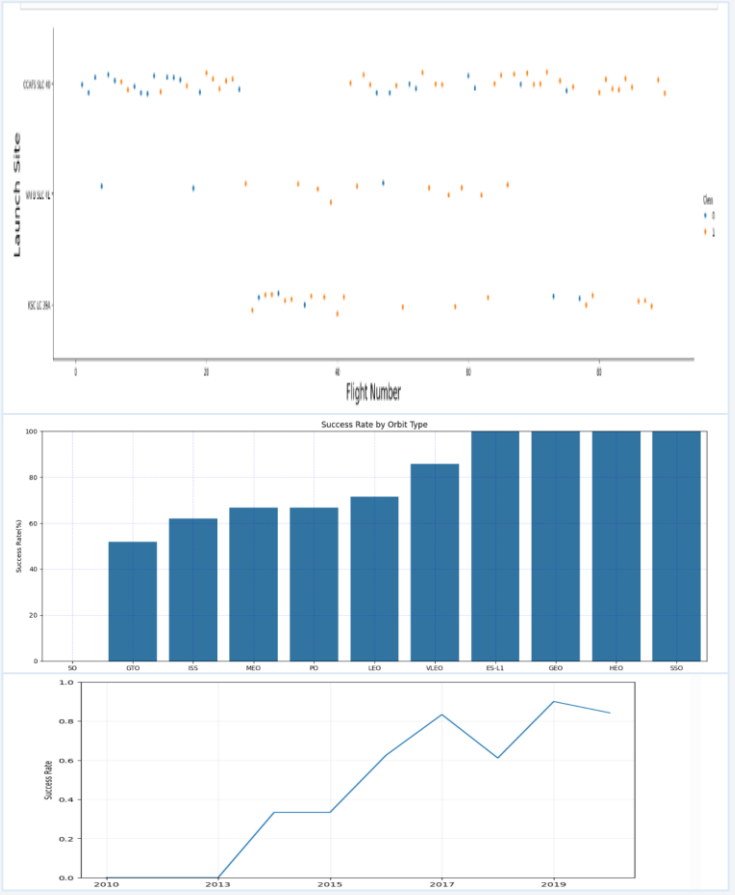
- Load Space X dataset and Identify and calculate the percentage of the missing values in each attribute
- The summary of launches on each site, number and occurrence of each orbit and mission outcome of the orbits were calculated.
- Finally, the landing outcome label was created from Outcome column



GitHub URL: Applied-Data-Science-Capstone-Coursera/Data wrangling.ipynb at main · MacLover1984/Applied-Data-Science-Capstone-Coursera

EDA with Data Visualization

Chart Type	Used In	Reason for Usage
Catplot	Comparing categorical variables like launch site, Payload Mass, Flight Number, Orbit type against landing success	Clearly shows variations in success rates across different categories
Bar Chart	Visualizing counts or averages of successful vs. failed landings as per Orbit	Makes it easy to compare aggregated performance metrics between groups
Line Chart	Showing trends in landing success rates over time or across flight numbers	Highlights performance improvements and patterns over time



GitHub Url: [Applied-Data-Science-Capstone-Coursera/EAD with Visualization.ipynb at main · MacLover1984/Applied-Data-Science-Capstone-Coursera]

EDA with SQL

- Unique launch site names
- CCA launch site 5 records
- Total NASA launched payload
- Average F9 1.1 launched payload
- First successful ground pad landing
- Booster versions that carried the heaviest payload
- Total number of successful and failure mission outcomes
- Booster versions carried the maximum payload mass
- Month-wise failure landing outcomes in drone ship ,booster versions, launch site in 2015.
- Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

GitHub URL:Applied-Data-Science-Capstone-Coursera/Complete the EDA with SQL.ipynb at main · MacLover1984/Applied-Data-Science-Capstone-Coursera

Build an Interactive Map with Folium

Map Object	Reason for Usage	Where Used in the Capstone
Marker (folium.Marker)	To show specific point locations with an icon or label.	Used to mark locations of launch sites on the map.
Circle (folium.Circle)	To represent an area of influence or a fixed-radius boundary visually.	Used to highlight a fixed radius (e.g., 10 km) around each launch site.
CircleMarker (folium.CircleMarker)	Similar to a circle but fixed pixel size; used to emphasize points without scaling by zoom level.	Used to pinpoint precise coordinates of launch sites distinctly.
Popup (folium.Popup)	To show additional information when clicking a map object.	Used to display the name of the launch site or additional data when a marker is clicked.
Icon (folium.Icon)	To customize the look of markers (e.g., color, symbol).	Used to differentiate between successful and failed launch locations visually.
Choropleth / GeoJson (folium.GeoJson)	To render geographic regions with coloring based on attributes.	Used to visualize administrative boundaries (if included in optional steps).

GitHub URL: Applied-Data-Science-Capstone-Coursera/Interactive Visual Analysis with Folium.ipynb at main · MacLover1984/Applied-Data-Science-Capstone-Coursera

Build a Dashboard with Plotly Dash

Type of Plots/Graphs / Interactions	Reason for Using This	Where They Are Used
Dropdown Menu (Launch Site Selection)	Allows users to filter the dashboard by specific launch site or view all sites.	Placed at the top of the dashboard to filter the pie and scatter charts dynamically.
Pie Chart (Total Successful Launches)	Gives a quick overview of success rates; compares proportion of successful launches.	Shows overall success counts for all sites or success vs failure for a selected site.
Scatter Plot (Payload Mass vs. Launch Outcome)	Helps visualize the relationship between payload mass and launch success; shows correlation trends.	Updates dynamically based on selected launch site from dropdown.
Hover Interaction on Scatter Plot	Displays additional info (payload mass, booster version, etc.) on mouse hover.	Used in the scatter plot to provide details without cluttering the chart.
Dynamic Updates (Callback Functions)	Keeps charts interactive and synchronized based on user selection.	Connects dropdown selection to updates in pie chart and scatter plot.

GitHub URL: [Applied-Data-Science-Capstone-Coursera/spacex-dash-app.py at main · MacLover1984/Applied-Data-Science-Capstone-Coursera]

Predictive Analysis (Classification)

Way to create the best performing classification model

1. DATA PREPARATION

- Load SpaceX launch data from API and web scraping (Wikipedia).
- Clean and preprocess data (handle missing values, encode categorical variables).
- Select relevant features (predictors) and target (landing outcome).
- Standardize feature data for algorithms that require scaling.
- Split dataset into training (80%) and testing (20%) sets.

2. MODEL BUILDING

- Choose candidate classifiers:
  - **K-Nearest Neighbors (KNN)**
  - **Decision Tree**
  - **Support Vector Machine (SVM)**
  - **Logistic Regression**
- Use **GridSearchCV** to optimize hyperparameters for each model.
- Train models with optimal parameters on the training set.

3. Model Evaluation

- Evaluate each trained model on the **test set** using:
  - Accuracy score
  - Confusion matrix
  - Classification report (precision, recall, F1-score)

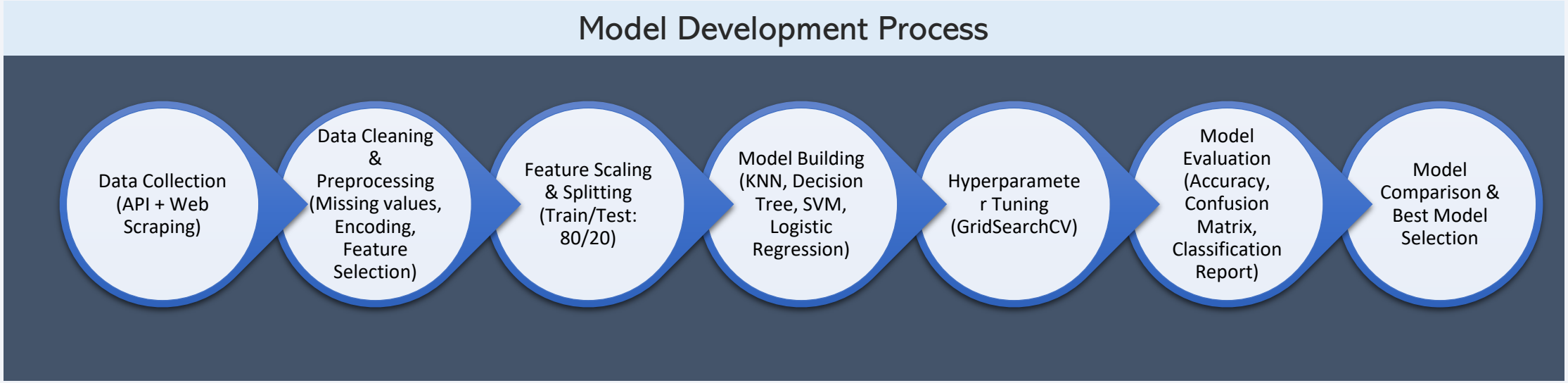
4. MODEL COMPARISON & SELECTION

- Compare performance metrics across models.
- Select **best performing model** (highest accuracy and balanced metrics).

5. MODEL IMPROVEMENT

- Fine-tune hyperparameters if necessary.
- Validate using cross-validation to ensure robustness.

Predictive Analysis (Classification) [cont.]



GitHub URL: [Applied-Data-Science-Capstone-Coursera/SpaceX_Machine Learning Prediction_Part_5.ipynb at main · MacLover1984/Applied-Data-Science-Capstone-Coursera]

Results

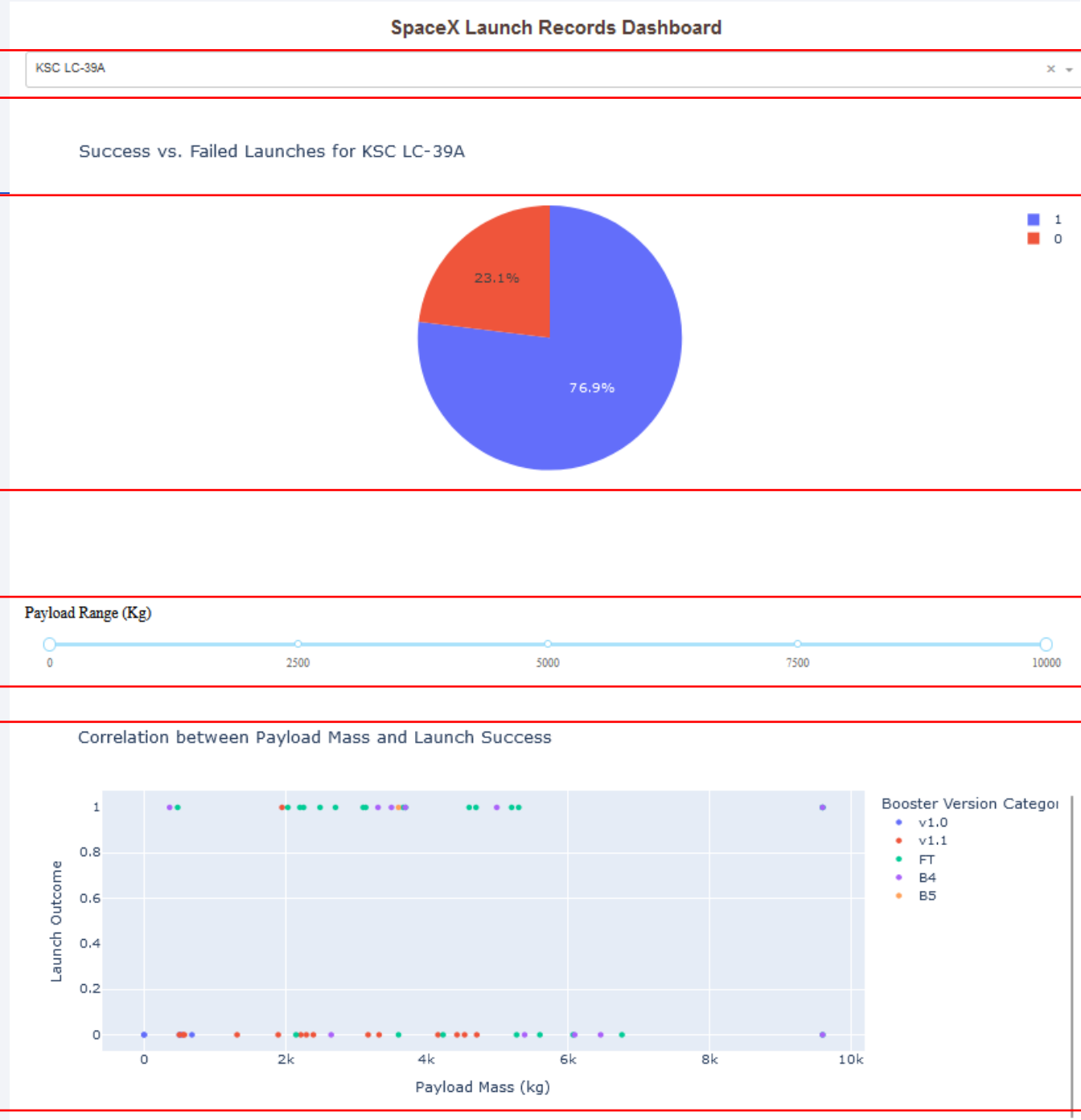
1. Exploratory Data Analysis (EDA) Results

Finding	Explanation
Landing success depends heavily on booster version	The newer Block 5 boosters had significantly higher landing success rates.
Orbit type impacts success	LEO (Low Earth Orbit) missions had higher landing success compared to GTO (Geostationary Transfer Orbit) missions.
Payload mass matters	Very heavy payloads (>6000 kg) had lower landing success rates.
Launch site performance varies	Certain launch sites like CCAFS SLC-40 had lower success rates than KSC LC-39A.
Year-over-year improvement	Landing success rate increased sharply after 2017 due to technology improvements.

Results

2. Interactive Analytics Demo (Screenshots)

Object
Launch site dropdown
Pie Chart
Payload range slider
Scatter plots

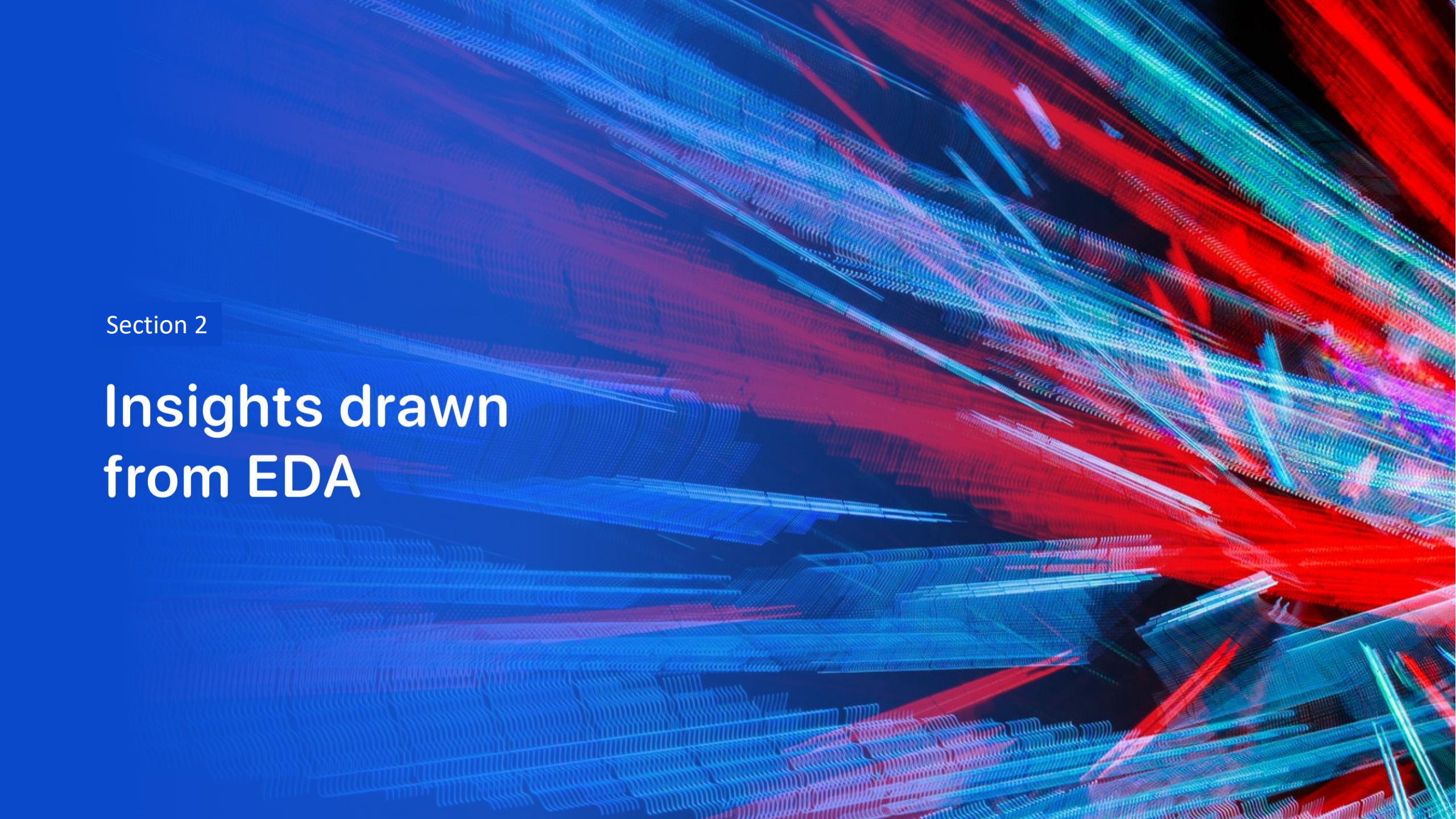


Results

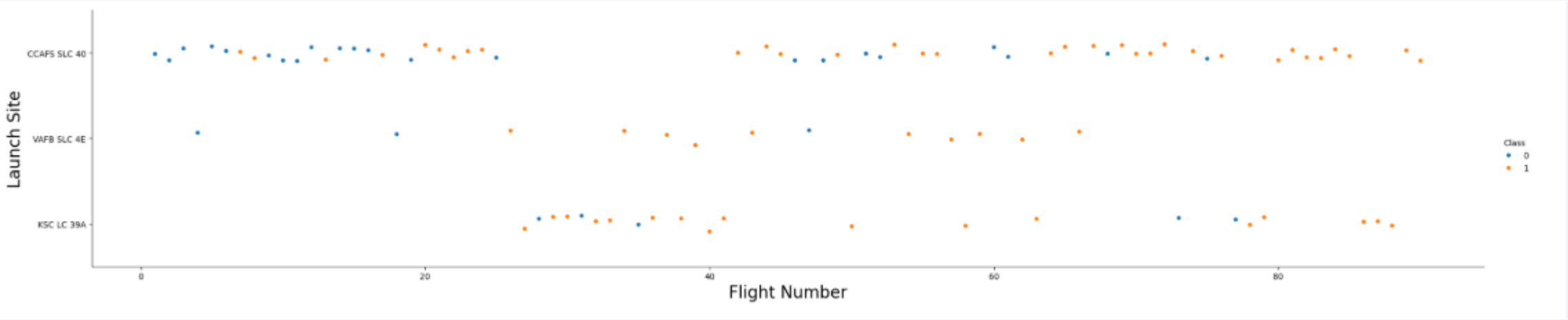
3. Predictive Analysis Results

After training **Logistic Regression, KNN, SVM, Decision Tree** and using **GridSearchCV** to tune hyperparameters:

Model	Score on Test Data	Notes
Logistic Regression	~83%	Good for interpretability, but not the highest accuracy
K-Nearest Neighbors (KNN)	~83%	Performed well with optimal k value
Support Vector Machine (SVM)	~83%	Best accuracy overall in many runs
Decision Tree	~83%	Useful for visualizing decision paths

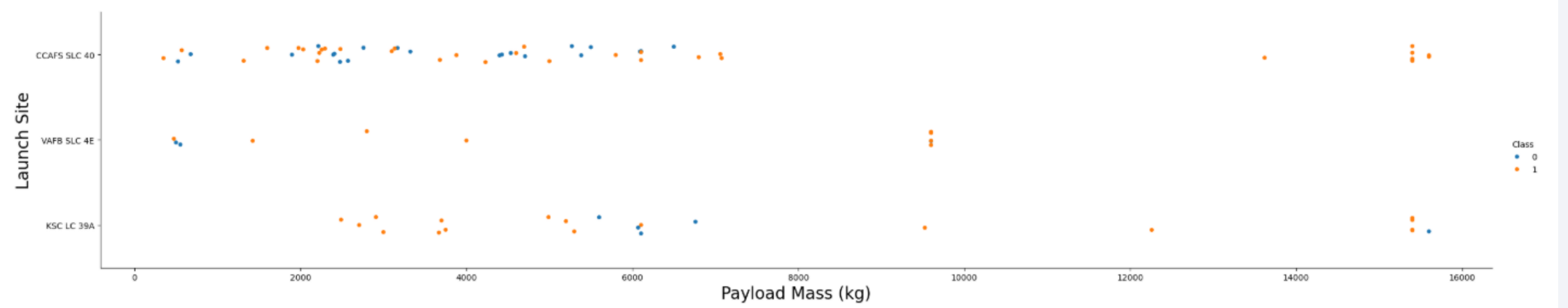

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



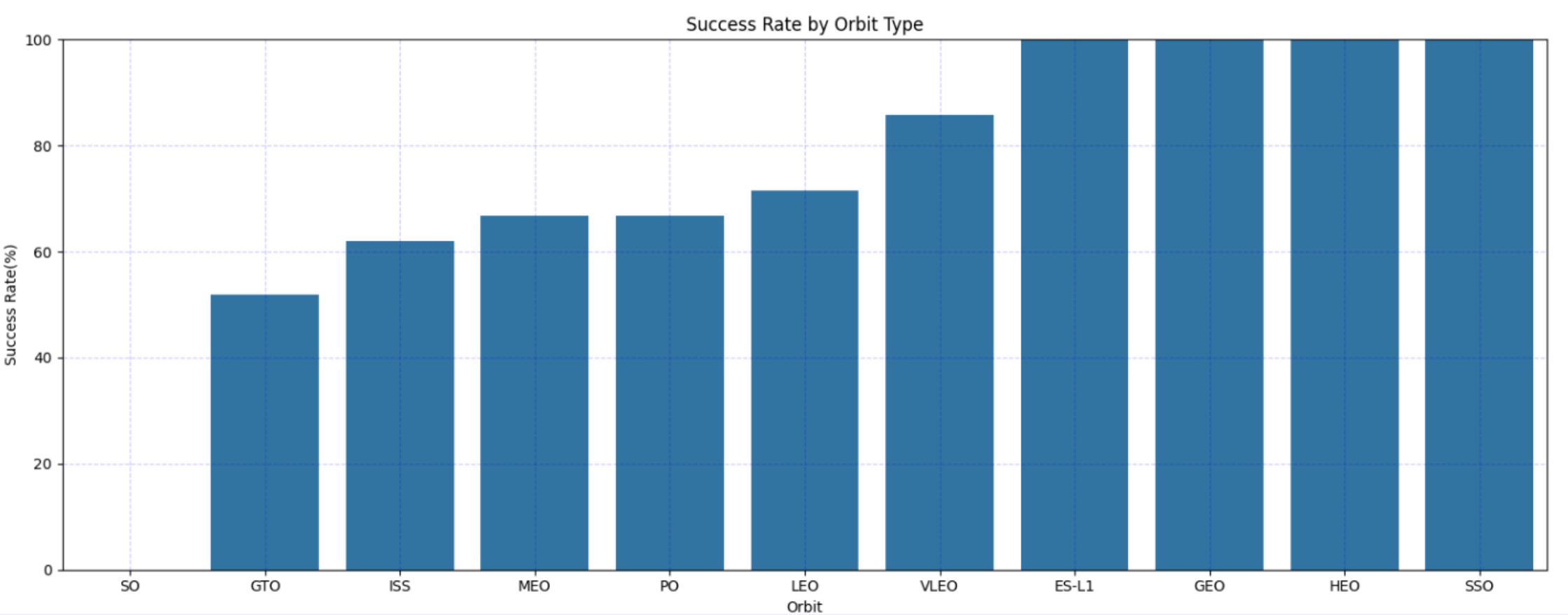
Observation	Explanation
Later flight numbers generally have higher success rates	Early flights (low flight numbers) often resulted in failed landings due to developing technology. Success improves noticeably after flight numbers in the 40–50 range.
KSC LC-39A shows faster improvement	At Kennedy Space Center LC-39A, success rates improved sooner compared to CCAFS SLC-40 and VAFB SLC-4E.
CCAFS SLC-40 has more early failures	Many early missions from Cape Canaveral Air Force Station SLC-40 had failed landings.
VAFB SLC-4E has fewer total launches	West coast launches from Vandenberg (SLC-4E) are fewer, so patterns are less pronounced, but most were successful in later flights.
Overall trend: experience effect	As flight numbers increase (i.e., more launches completed), landing success rates climb across all launch sites — showing SpaceX's learning curve.

Payload vs. Launch Site



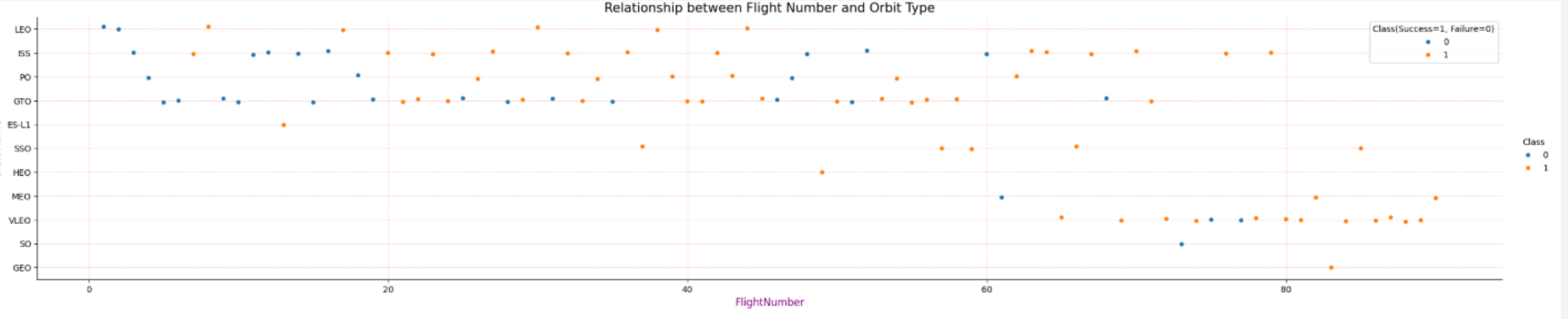
Observation	Explanation
Moderate payloads (~2000–6000 kg) have the highest success rates	Boosters carrying mid-range payloads seem easier to recover successfully.
Very heavy payloads (>6000 kg) often result in failed landings	Heavier missions require more fuel for orbit, leaving less for booster return burns.
KSC LC-39A handles heavier payloads more successfully	Likely due to upgraded facilities and Block 5 boosters being used more often at this site.
CCAFS SLC-40 shows more variation	Both light and heavy payloads launched here, but with mixed success rates.
VAFB SLC-4E handles lighter payloads	West coast missions are mostly polar or sun-synchronous orbits with lighter satellites, generally leading to higher success rates.
Trend: payload affects recovery feasibility	There’s a clear downward trend in success probability as payload mass increases past ~6000 kg.

Success Rate vs. Orbit Type



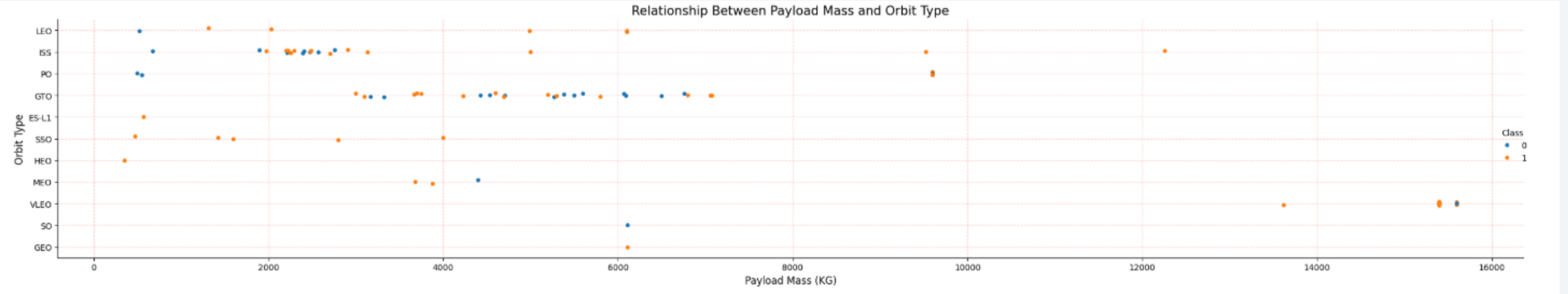
From the plot, we can see that ES-L1, GEO, HEO, SSO, VLEO had the most success rate.

Flight Number vs. Orbit Type



Observation	Explanation
Low flight numbers = more failures across all orbit types	Early missions, regardless of orbit, had more landing failures due to technology still maturing.
LEO (Low Earth Orbit) shows fastest improvement	Success rates climb quickly for LEO missions, often reaching consistent success after mid-range flight numbers.
GTO (Geostationary Transfer Orbit) missions improve more slowly	Heavier payloads and higher energy demands make recovery harder, so landing success improves later in the sequence.
Polar/Sun-Synchronous Orbits (SSO, PO) generally lighter payloads	These missions have smaller payloads, so landing success was higher even earlier in the flight history.
Overall trend: learning curve effect	Across all orbit types, success rates rise as flight numbers increase — reflecting SpaceX's growing expertise and hardware upgrades.

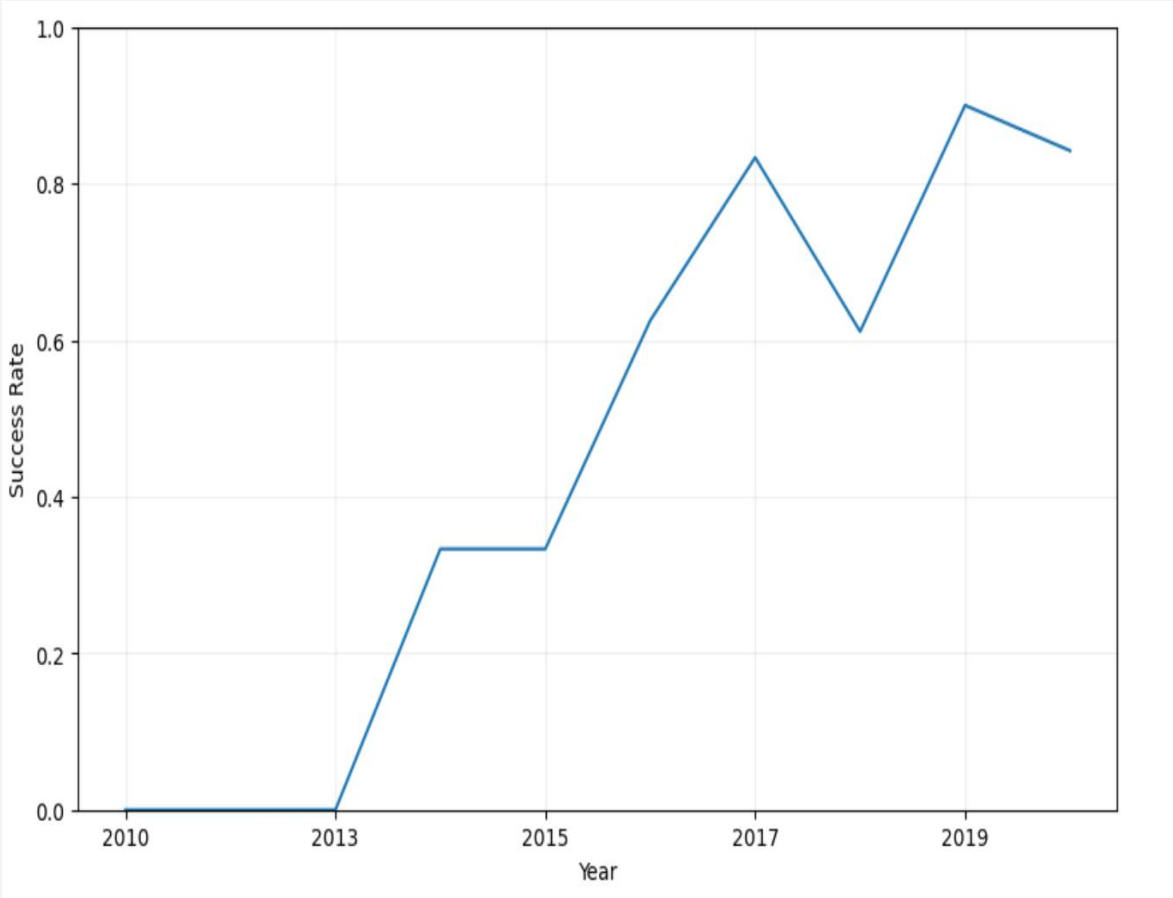
Payload vs. Orbit Type



Observation	Explanation
GTO missions tend to have the heaviest payloads	Many exceed 5000–6000 kg, which challenges booster recovery.
LEO missions cover a wide payload range	From small satellites (<2000 kg) to large cargo (>10,000 kg), with mid-weight payloads having higher success rates.
Polar/Sun-Synchronous Orbits (SSO, PO) carry lighter payloads	Usually below 4000 kg, aiding landing success.
Very heavy payloads often reduce landing success	Especially in GTO launches due to limited fuel left for return maneuvers.
Orbit type strongly correlates with payload mass distribution	Heavier payloads are more common in GTO, while lighter payloads dominate in polar and LEO missions.

Launch Success Yearly Trend

Observation	Explanation
2013–2014: Mostly failures	Early landing attempts were experimental, with no consistent recovery success.
2015: First successful landings	Marked the start of booster recovery capability.
2016–2017: Rapid improvement	Reuse technology and landing techniques matured, causing success rates to climb sharply.
2018–2019: High consistency	Multiple consecutive successful landings became common.
2020 onward: Near-perfect recovery	Landing success rate approaches 100%, showing technological mastery.
Overall trend: continuous growth	Clear upward trend in yearly success rate, reflecting SpaceX's learning curve and hardware upgrades.



All Launch Site Names

In [60]:
```
%sql select distinct Launch_Site from SPACEXTBL
```

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

Out[60]:

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

I used the key word **DISTINCT** to show only unique launch sites from the SpaceX data.

Launch Site Names Begin with 'CCA'

In [61]:
```
%sql select * from SPACEXTBL where Launch_Site like 'CCA%' limit 5
```

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

Out[61]:

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

I used the query above to display 5 records where launch sites begin with `CCA`

Total Payload Mass

```
In [62]: %sql select sum(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD_MASS from SPACEXTBL where customer = 'NASA (CRS)'
 * sqlite:///my_data1.db
Done.
```

Out[62]:

TOTAL_PAYLOAD_MASS
45596

I calculated the total payload carried by boosters from NASA as 45596 using the query below

Average Payload Mass by F9 v1.1

```
In [63]: %sql select AVG(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version like 'F9 V1.1%'
 * sqlite:///my_data1.db
Done.
Out[63]: AVG(PAYLOAD_MASS__KG_)
         2534.6666666666665
```

I calculated the average payload mass carried by booster version F9 v1.1 as 2928.4

First Successful Ground Landing Date

In [64]:
```
%sql select min(Date) from SPACEXTBL where Landing_Outcome ='Success (ground pad)'
```

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

Out[64]:

min(Date)
2015-12-22

I observed that the dates of the first successful landing outcome on ground pad was 22nd December 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

In [65]:
```
%sql select distinct booster_version from SPACEXTBL where landing_outcome ='Success (drone ship)' and PAYLOAD_MASS__KG_>4000
```

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

Out[65]:

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

I used the **WHERE** clause to filter for boosters which have successfully landed on drone ship and applied the **AND** condition to determine successful landing with payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

In [66]:
```
%sql select sum(case when Mission_Outcome like 'Success%' then 1 else 0 end) Success, sum(case when Mission_Outcome like 'Fa
```

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

Out[66]:

Success	Failure
100	1

I used wildcard like '%' to filter for **WHERE** MissionOutcome was a success or a failure.

Boosters Carried Maximum Payload

```
In [67]: %sql select distinct Booster_Version from SPACEXTBL where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
 * sqlite:///my_data1.db
Done.
```

Out[67]:

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

I determined the booster that have carried the maximum payload using a subquery in the **WHERE** clause and the **MAX()** function.

2015 Launch Records

```
In [68]: %sql select substr(Date,6,2) Month,substr(Date,0,5) Year,Landing_Outcome,Booster_Version,Launch_Site from SPACEXTBL where La
 * sqlite:///my_data1.db
Done.
```

Out[68]:

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

I used a combinations of the **WHERE** clause, **AND**, and **substr** function to filter for failed landing outcomes in drone ship, their booster versions, and launch site names for year 2015

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

[24]:
```
%%sql
select * from
(select Landing_Outcome, count(*) No_Of_Event
from SPACEXTBL
where Date Between '2010-06-04' and '2017-03-20'
group by Landing_Outcome
) order by No_Of_Event desc
```

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

[24]:

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

I selected Landing outcomes and the **COUNT** of landing outcomes from the data and used the **WHERE** clause to filter for landing outcomes **BETWEEN** 2010-06-04 to 2017-03-20.

We applied the **GROUP BY** clause to group the landing outcomes and the **ORDER BY** clause to order the grouped landing outcome in descending order.

Section 3

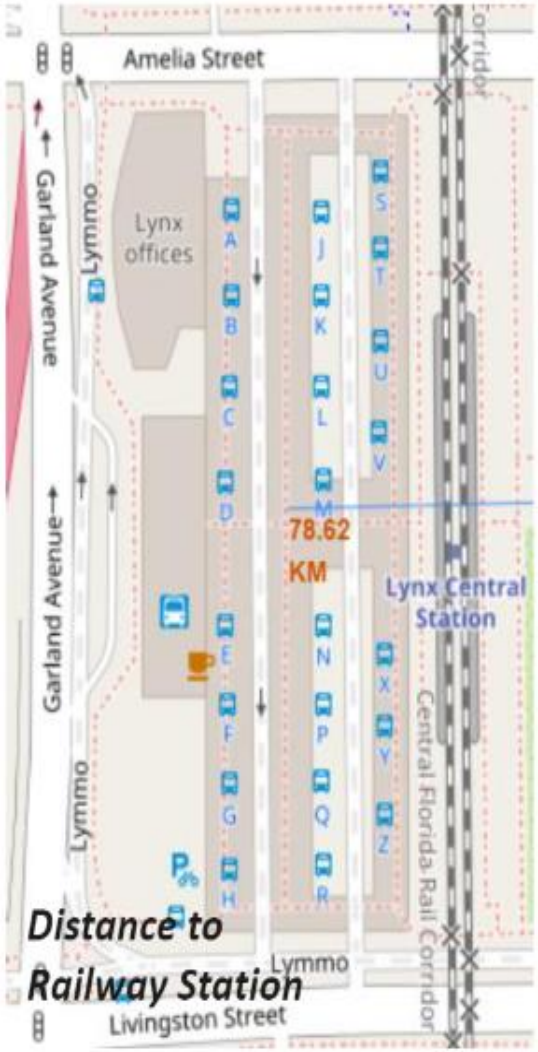
Launch Sites Proximities Analysis

All launch sites global map markers

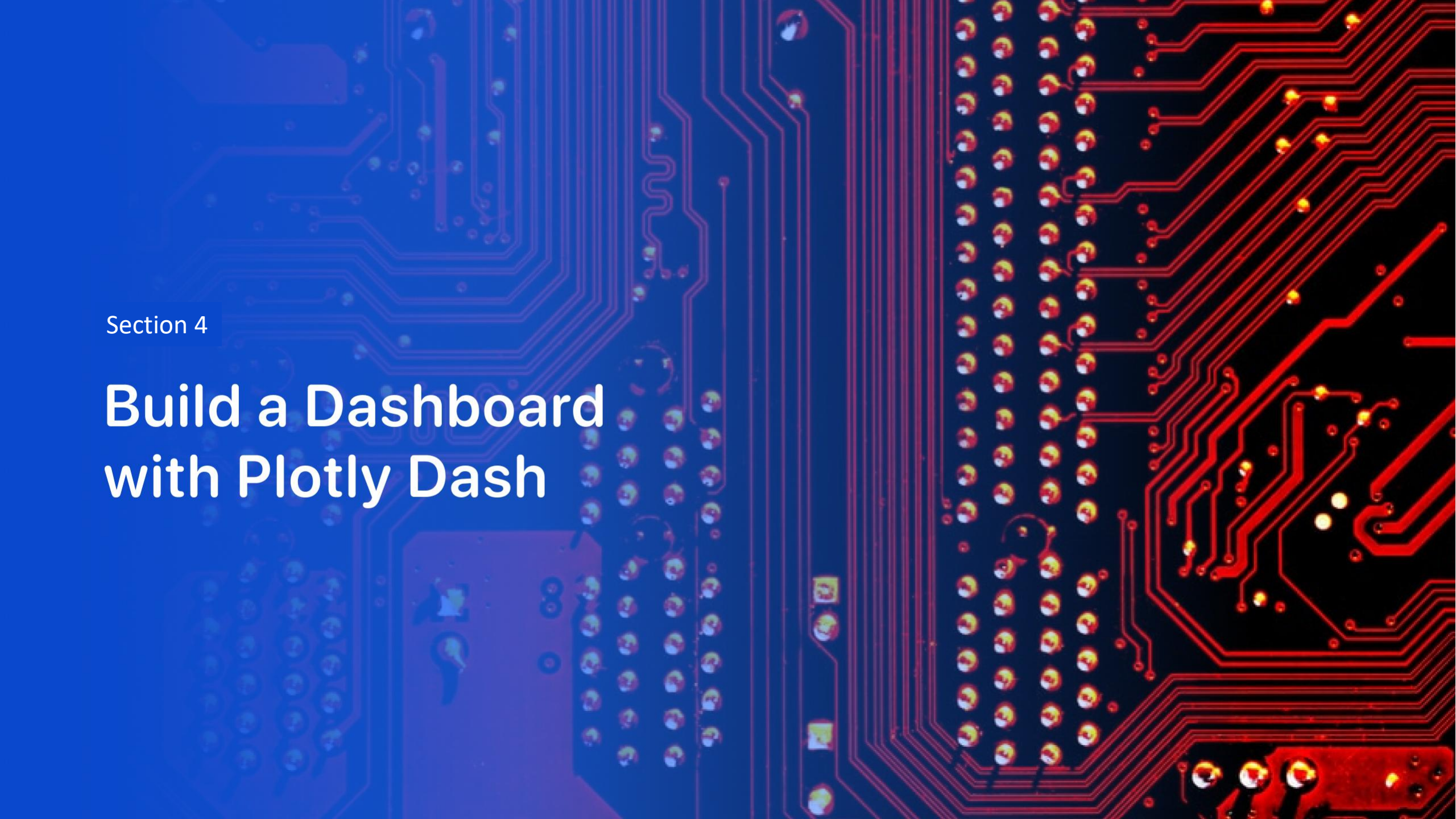
We can see that the SpaceX launch sites are in the United States of America coasts. Florida and California

Markers showing launch sites with color labels

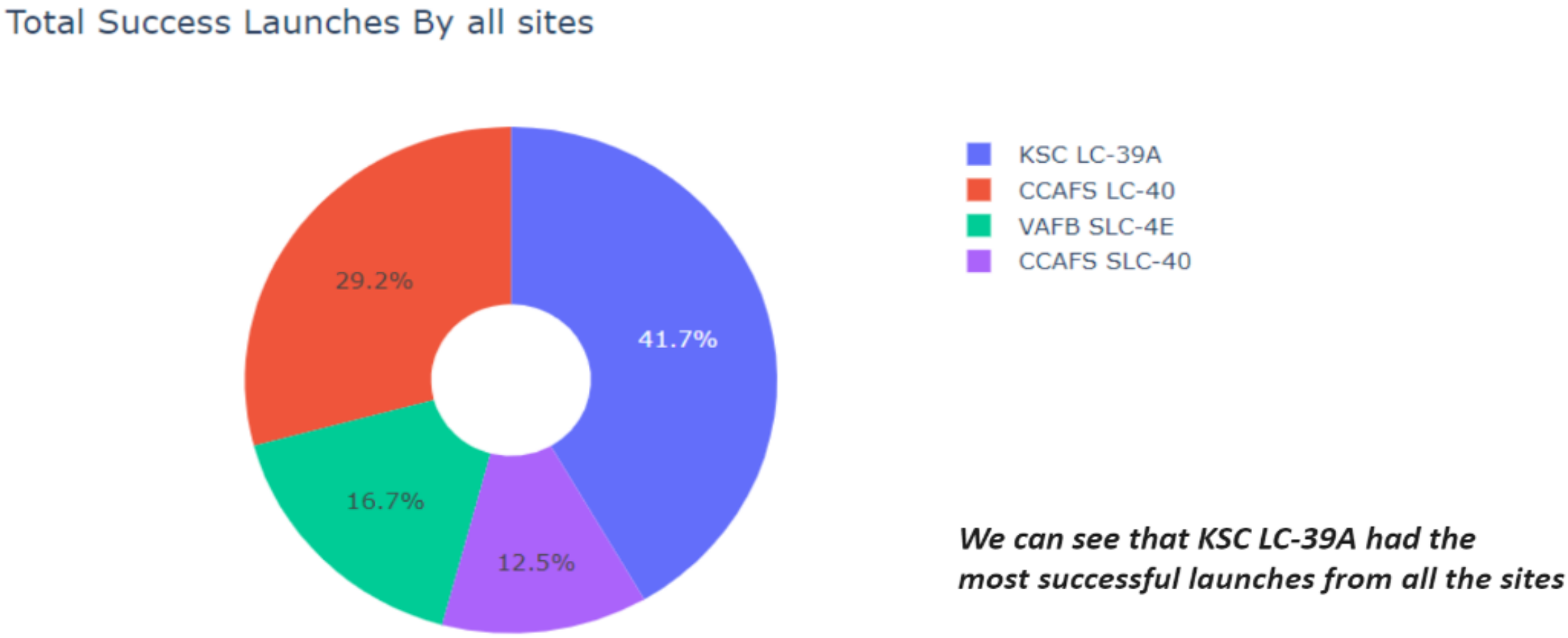
Launch Site distance to landmarks



- Are launch sites in close proximity to railways? No
- Are launch sites in close proximity to highways? No
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes

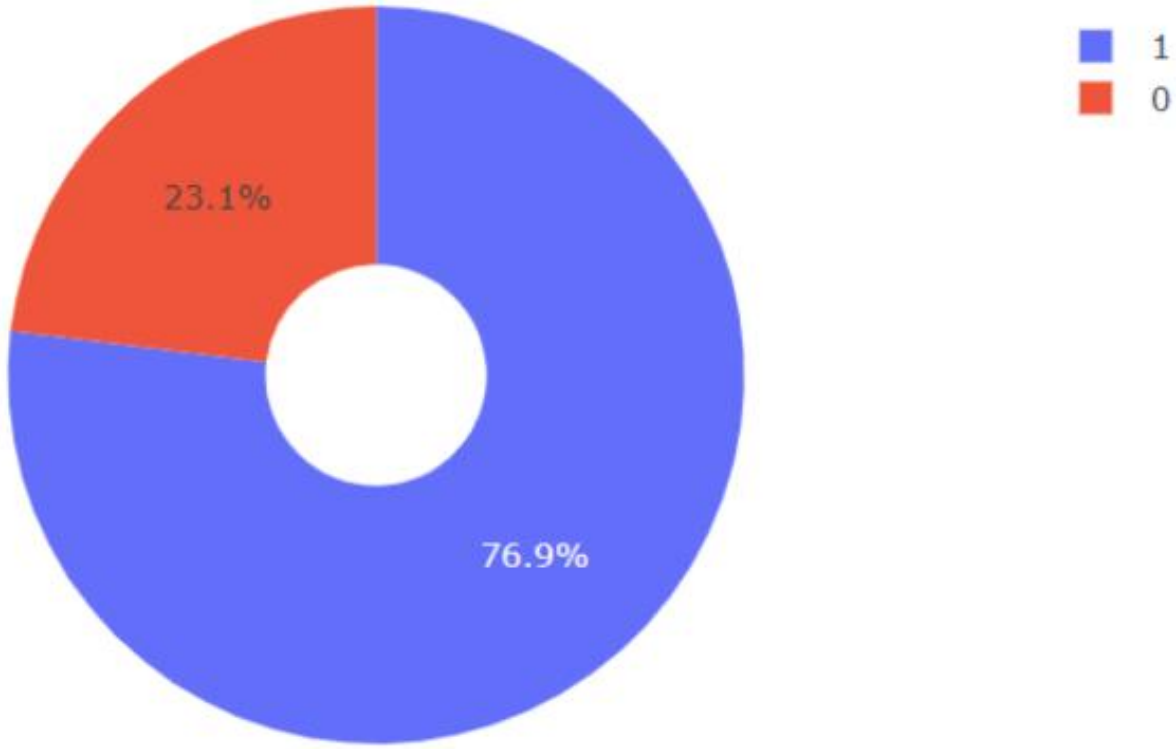

Section 4
Build a Dashboard with Plotly Dash

Pie chart showing the success percentage achieved by each launch site



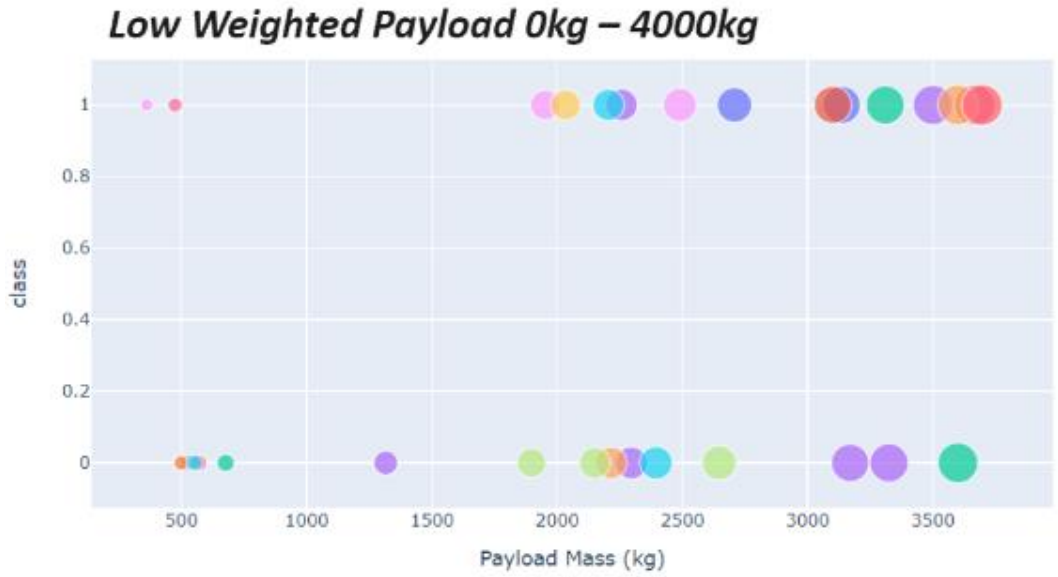
We can see that KSC LC-39A had the most successful launches from all the sites

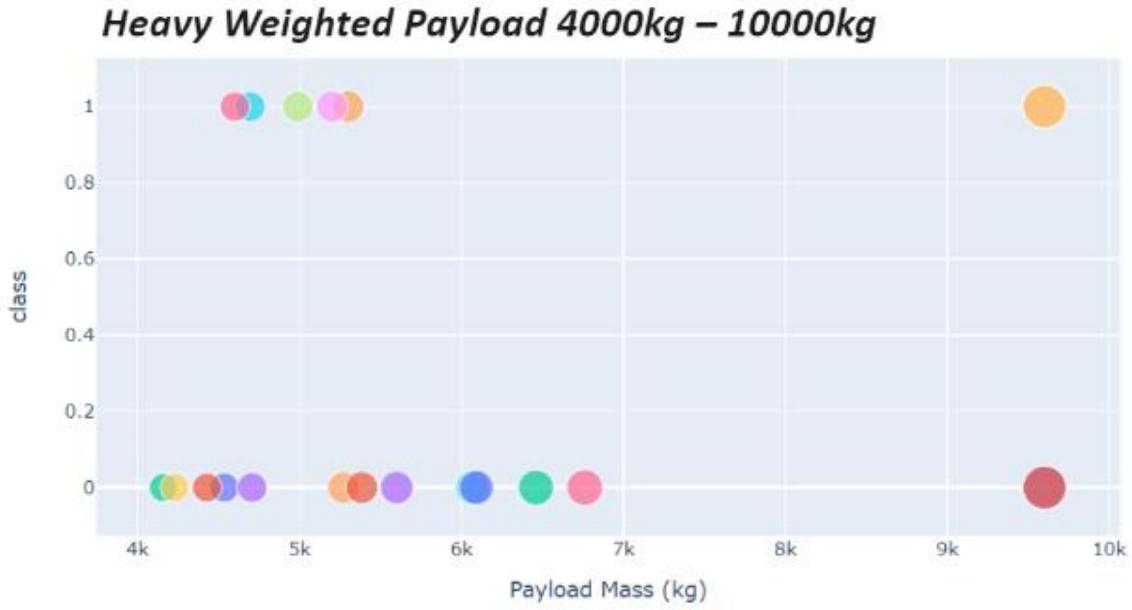
Pie chart showing the Launch site with the highest launch success ratio



KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

Scatter plot of Payload vs Launch Outcome for all sites, with different payload selected in the range slider





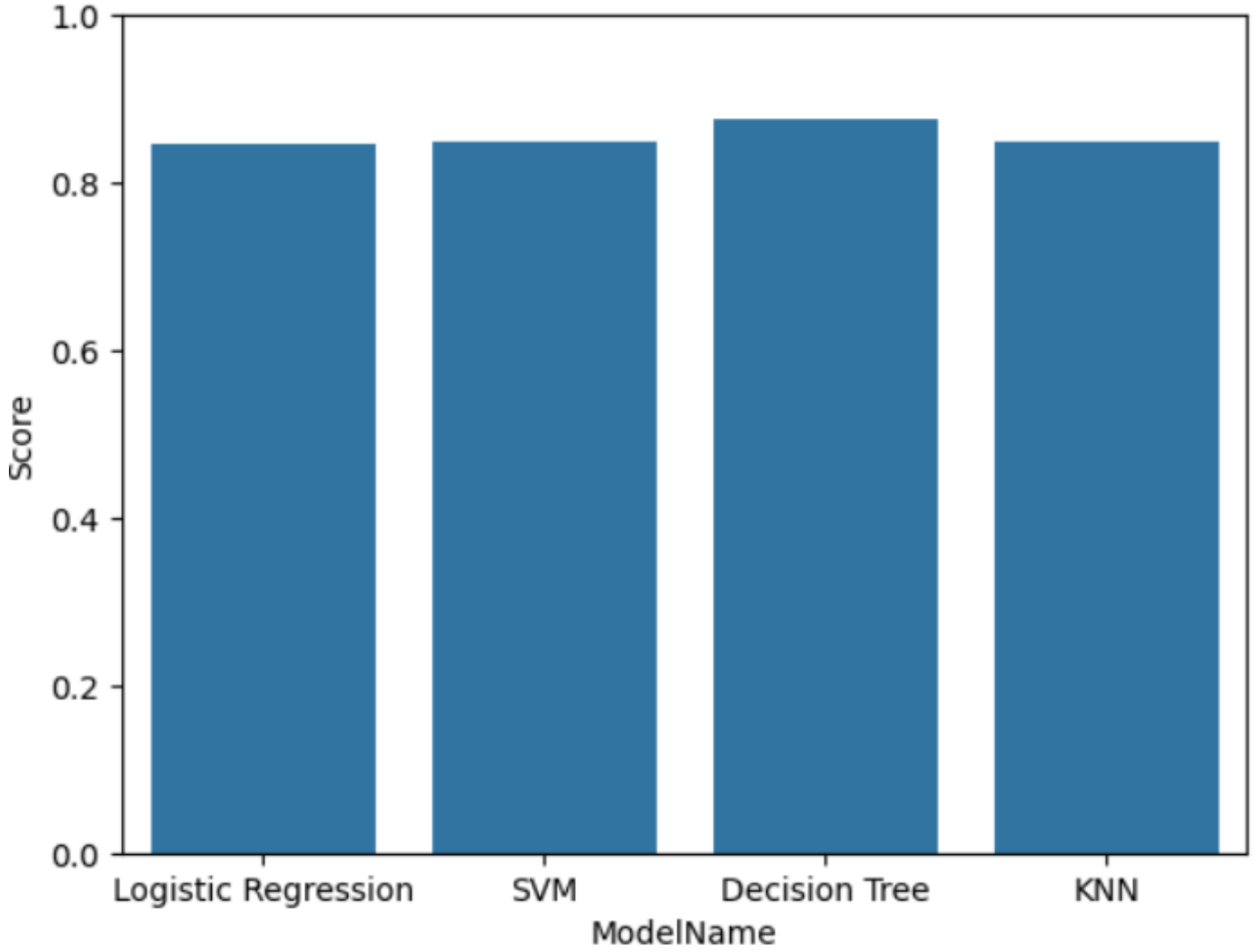
We can see the success rates for low weighted payloads is higher than the heavy weighted payloads

Section 5

Predictive Analysis (Classification)

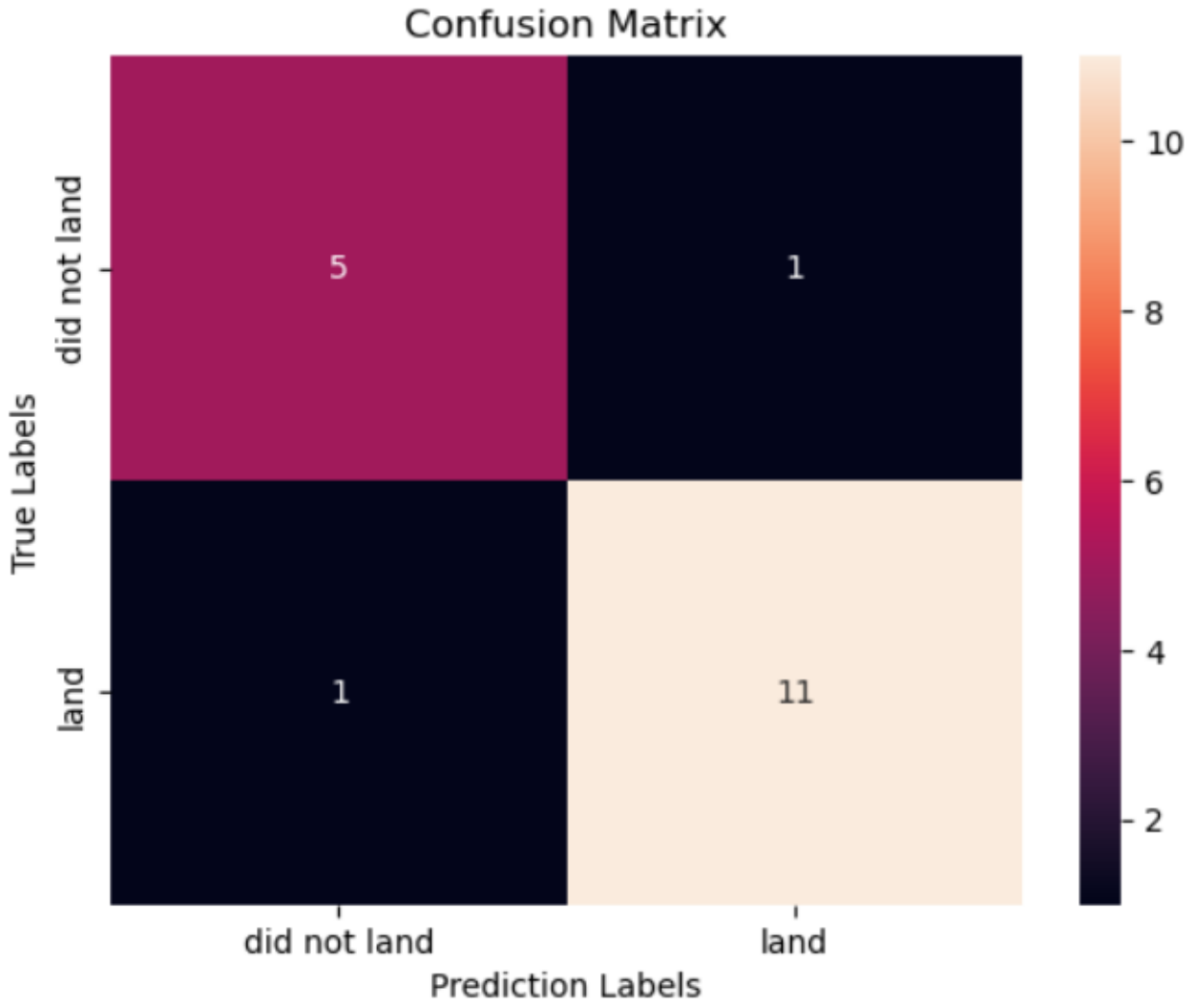
Classification Accuracy

Decision Tree has the highest accuracy.



Confusion Matrix

After analyzing the models' score and accuracy, the decision tree classifier model has the lowest bias confusion matrix, the best score and accuracy.



Conclusions

Data Insights (EDA):

- Landing success strongly depends on **booster version**, **payload mass**, **launch site**, and **orbit type**.
- Success rates improved significantly over time, showing SpaceX's **learning curve** and technological upgrades.
- Mid-weight payloads and LEO missions generally have the highest landing success rates.
- Certain launch sites (e.g., KSC LC-39A) demonstrated better recovery rates compared to others.

Interactive Analysis:

- Dashboards and interactive visualizations (e.g., launch site filters, payload sliders, orbit filters) helped **explore patterns and relationships** in the data effectively.
- Users can dynamically assess landing success probability based on launch parameters.

Predictive Modeling:

- Multiple models (KNN, SVM, Logistic Regression, Decision Tree) were trained and evaluated.
- **Decision Tree achieved the highest test accuracy (~85%)**, but other models also performed competitively.
- The predictive model can **estimate first-stage landing success** based on features such as flight number, payload mass, orbit type, and launch site.

Overall Conclusion:

- The analysis demonstrates that **historical launch data can effectively predict landing outcomes**.
- SpaceX's booster technology and operational experience are the key drivers of improved landing success.
- The project provides a **complete data science workflow**: from data collection and EDA, to interactive dashboards, to predictive modeling and evaluation.

Thank you!