IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ewerthon Jose Kutz
Data Scientist
2024-09-05
Brazil

Outline

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

Executive Summary

Purpose

This project aims to understand the different aspects of **costs** in space rocket launches. Its goals were to:

- Explore and analyze the data from SpaceX launches to gain insights.
- Predict the success or failure of a rocket landing.

Methodology

A Data Mining project was executed, involving data collection, data wrangling, exploratory data analysis, data visualization and machine learning modelling.

The main technologies employed in the project were Python and SQL, utilizing notebooks, API calls, Web Scrapping and libraries such as Pandas, Numpy, Matplotlib, Folium, Dash, and Scikit-Learn.

Results

The exploration uncovered insights about the most important features related to the landing success and, thus, the cost of launches. That enabled the creation of a machine learning model with **>80%** of accuracy on predicting success landings.

Introduction

Background and Context

Many companies are already operating commercial space travels, with SpaceX being the most successful of them. Their low cost, coming primarily from the fact that they can land the first stage of the rocket and reutilize it, is their main competitive edge.

This project was born from the interest to understand the different aspects and features related to this landing process and explore the possibility to predict if a landing will be successful or not.

Problems to be explored

- What data related to SpaceX launches and landings can we collect?
- What is the quality of this data? Does it need to be cleaned?
- What are the most important features related to the landing success?
- Is it possible to predict the success of the landing via machine learning modelling?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
- Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

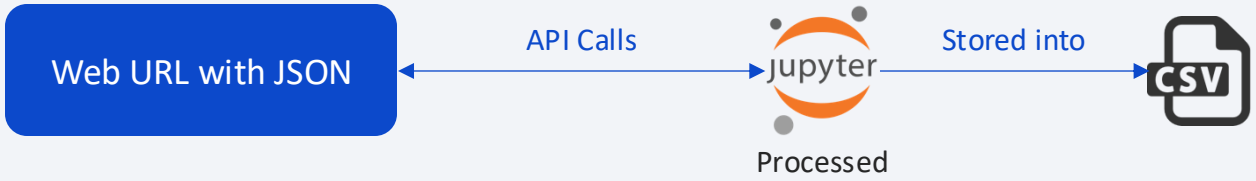
Data Collection

- Data sets were collected from **two** sources: a json from a web URL and an HTML from a Wikipedia page. Both contains data about rocket spaces launches (Falcon 9 and Falcon Heavy).
- Python was used for both cases. API calls via requests library to obtain data from the URLs, BautifulSoup to parse HTML data and Pandas to load the data sets as data frames.

Data Collection – SpaceX API

Jupyter Notebook Link

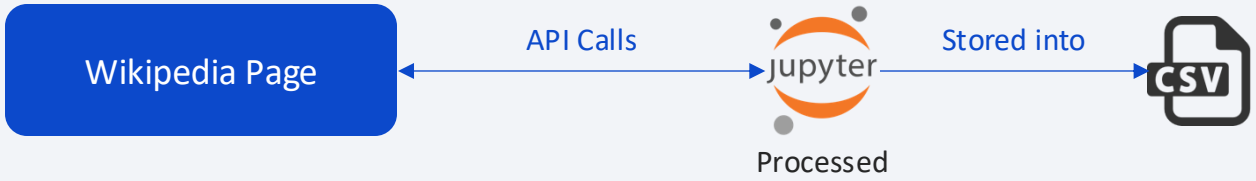
Data source 1 (JSON): Jupyter notebook with Python Code was used to make API calls via requests library, store data in a Pandas DataFrame and save it to the “data/processed” folder as a CSV. Flowchart below:



Data Collection - Scraping

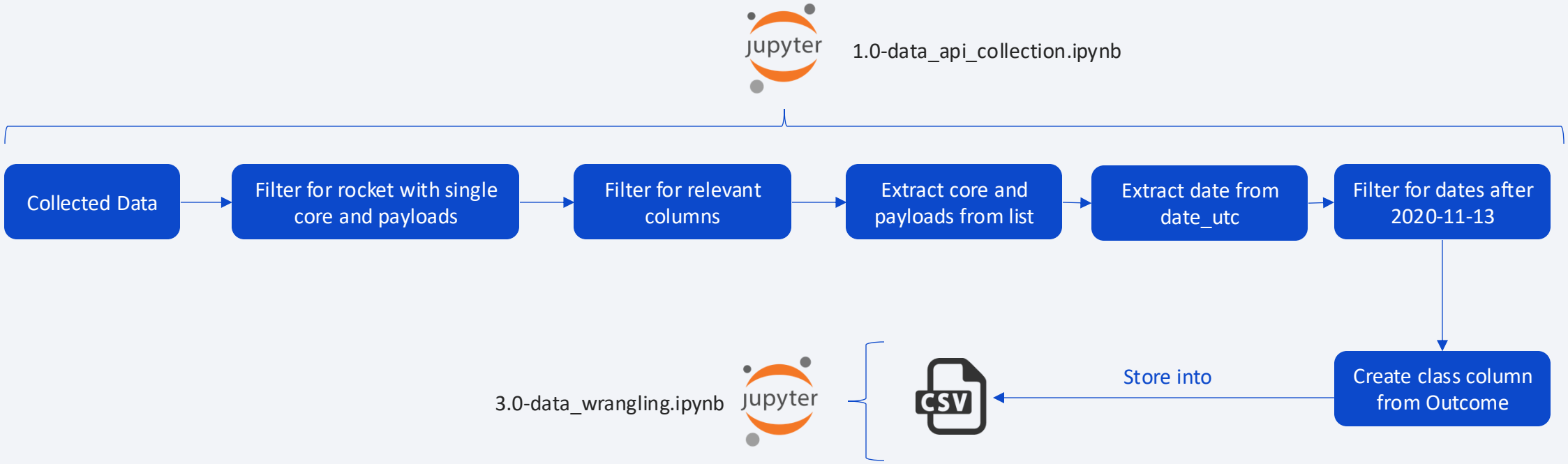
Jupyter Notebook Link

Data source 2 (HTML table): Jupyter notebook with Python Code was used to make API calls via requests library to a Wikipedia page, then data was parsd with BeautifulSoup and data stored in a Pandas DataFrame, then saved to the “data/processed” folder as a CSV. Flowchart below:



Data Wrangling

Data was wrangled on Jupyter notebook with Python Code and, then saved to the “data/processed” folder as a CSV. Flowchart below:



EDA with Data Visualization

Jupyter Notebook Link

6 charts were developed to analyze the correlation between features and target (successful landing).

- Strip plots were used to study correlation between categorical and numerical features vs the target.
- Bar charts were used to visualize categorical features vs target, where the target was represented as the success rate (mean of class outcome).
- Time series charts were used to study the success rate over the years.

EDA with SQL

Jupyter Notebook Link

- Queries were performed using sqlalchemy and sqlite via Jupyter Notebooks.
- Simple functions were required (DISTINCT, LIKE, SUM, AVG, MIN, GROUP BY, COUNT).
- Simply WHERE clauses and subqueries were also required.

Build an Interactive Map with Folium

Jupyter Notebook Link

- Circles and markers were created to mark NASA space center and all launching sites from the data set in the map.
- Cluster markers were created to represent successful and unsuccessful launches for each site.
- Lines were added to the map to show the near infrastructure from launching sites (closest railway, highway, city, and coastline).
- The object were added to highlight different aspects of the launching sites: chosen location, infrastructure proximity and how this is related to launches outcomes.

Build a Dashboard with Plotly Dash

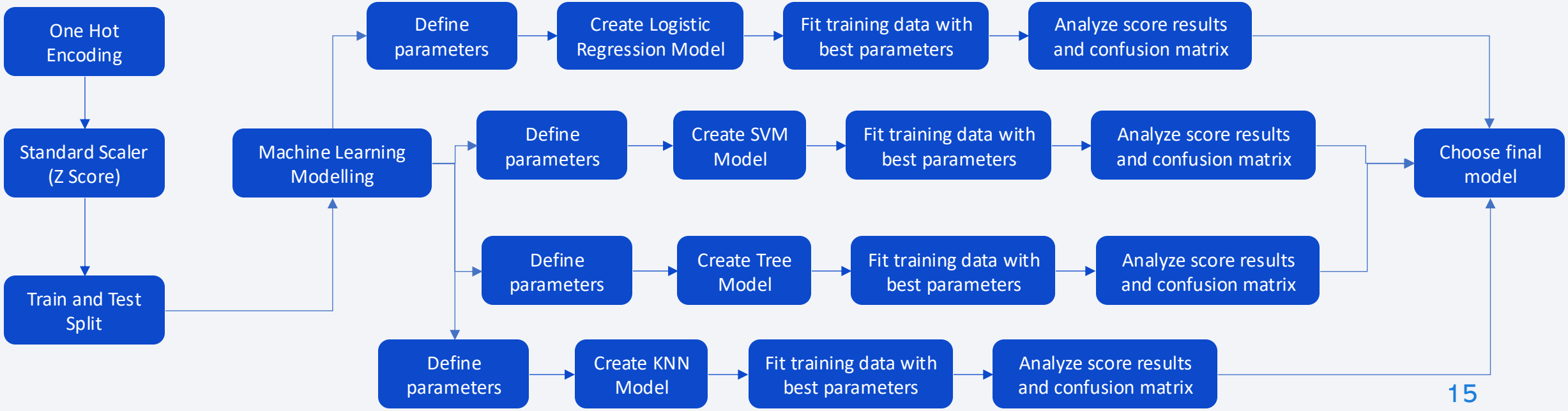
Jupyter Notebook Link

- A pie chart was inserted into the dashboard to enable a visual analysis on the successfulness of launches (overall and per site).
- A scatter plot was used to present the success count on Payload mass (overall and per site).
- Dropdown lists were used for choosing all sites or a specific site for the analysis and a value slide to choose the value range to the Payload mass on the scatter plot.
- These plots and interactions allows the user to customize the visualization according to their analytic needs on the success or failure of launches.

Predictive Analysis (Classification)

Jupyter Notebook Link

- Data was preprocessed with One Hot Encoding and Standardization (Z Score) techniques and split in training and testing data sets on scikit-learn.
- Four different models were evaluated: Logistic Regression, SVM, Decision Trees, and KNN. Each model went on hyperparameter tunning with GridSearchCV.
- The best classification model was determined using their **accuracy** score and confusion matrix on the test set.

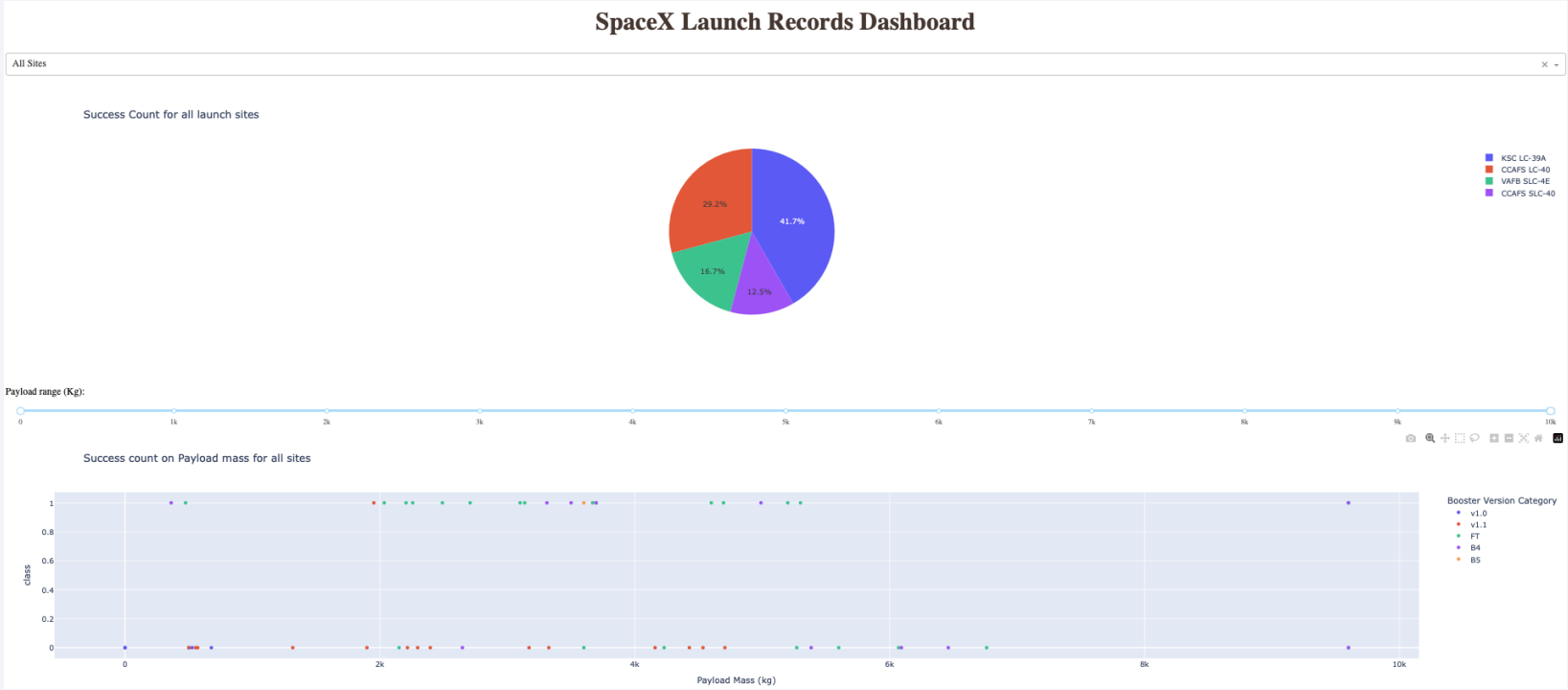


Results

- Exploratory Data Analysis:
  - As Flight Number and Payload Mass increases, success rate also increases.
  - Launch Site CCAFS SLC 40 has most of the lower Flight Numbers, also concentrating less successful landing.
  - Launch Site VAFB-SLC does not have launches with heavy payload mass (greater than 10k).
  - Success Rate has a high variance across different Orbit Types. SO has a 0% success Rate and five have 100% (ES-L1, GEO, HEO, SSO, and VLEO). Some of these Orbit Type results are also influenced by other features, such as the lower Flight Numbers in LEO (71% success rate).
  - Success Rate is increasing over the years, indicating that accumulating know-how and implementing new technologies positively influences the success of landings.

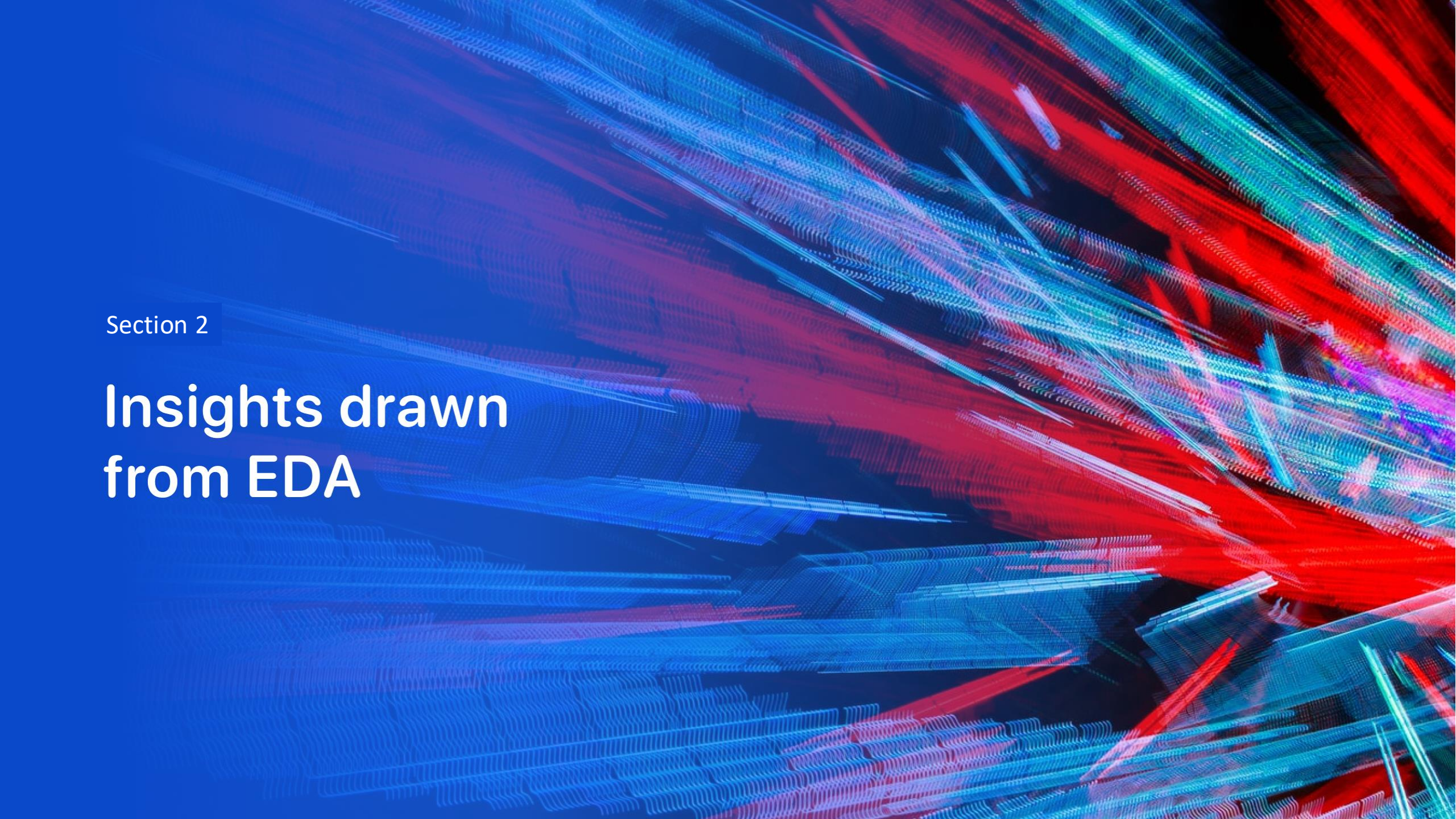
Results

- Interactive analytics demo:



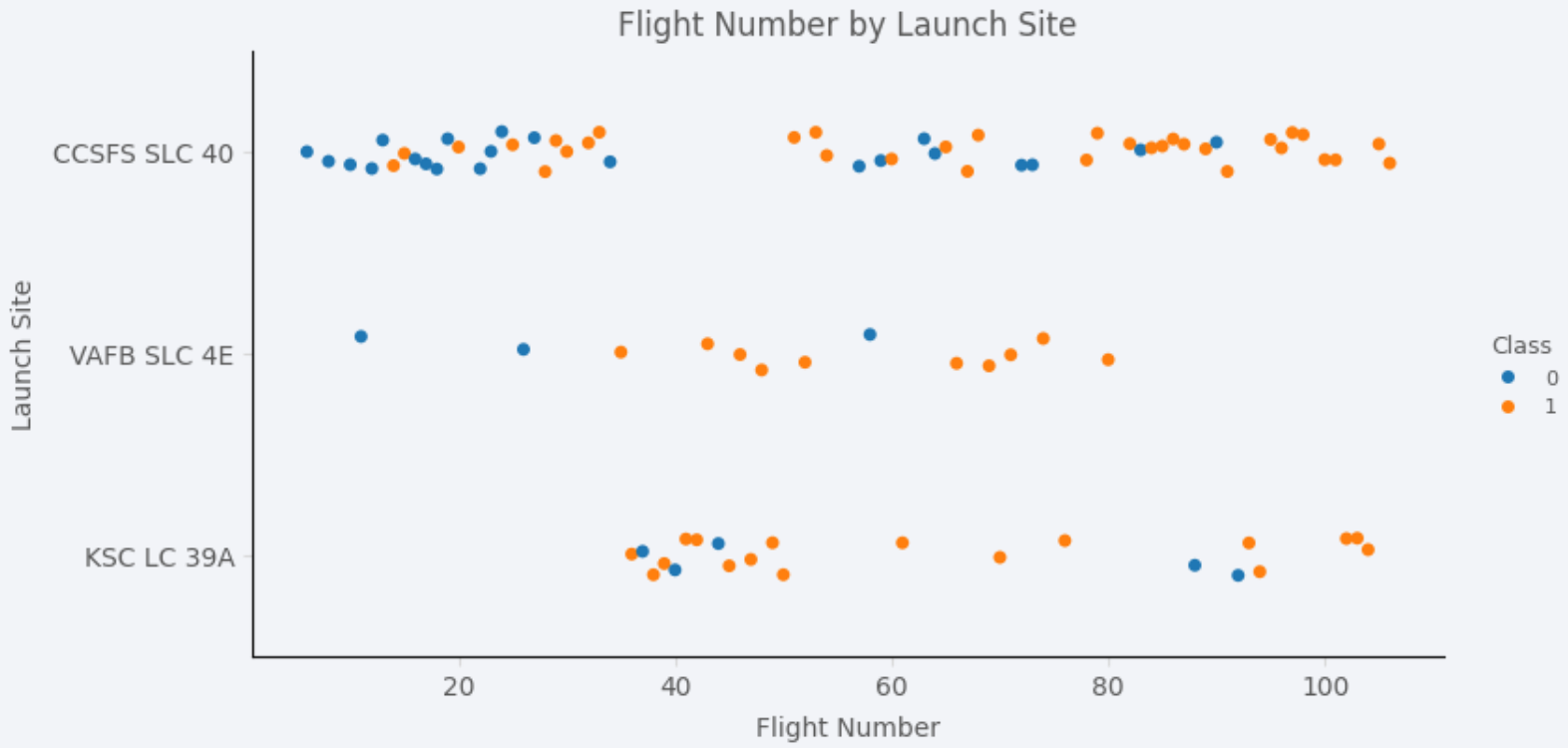
Results

- Predictive analysis results:
  - Models were evaluated using accuracy score, which is the percentage of correct predictions of the model.
  - Results are very similar across different models and hyperparameters tuning – from 83-88% of accuracy on unseen data.
  - Most errors in predictions are False Positives: when the model predicts a successful landing, but an unsuccessful one happened.
  - The best result comes from the Decision Tree Model with tuned hyperparameters: 88% of accuracy and predicting the correct outcome in 16 of the 18 landings on unseen data.


Section 2
Insights drawn from EDA

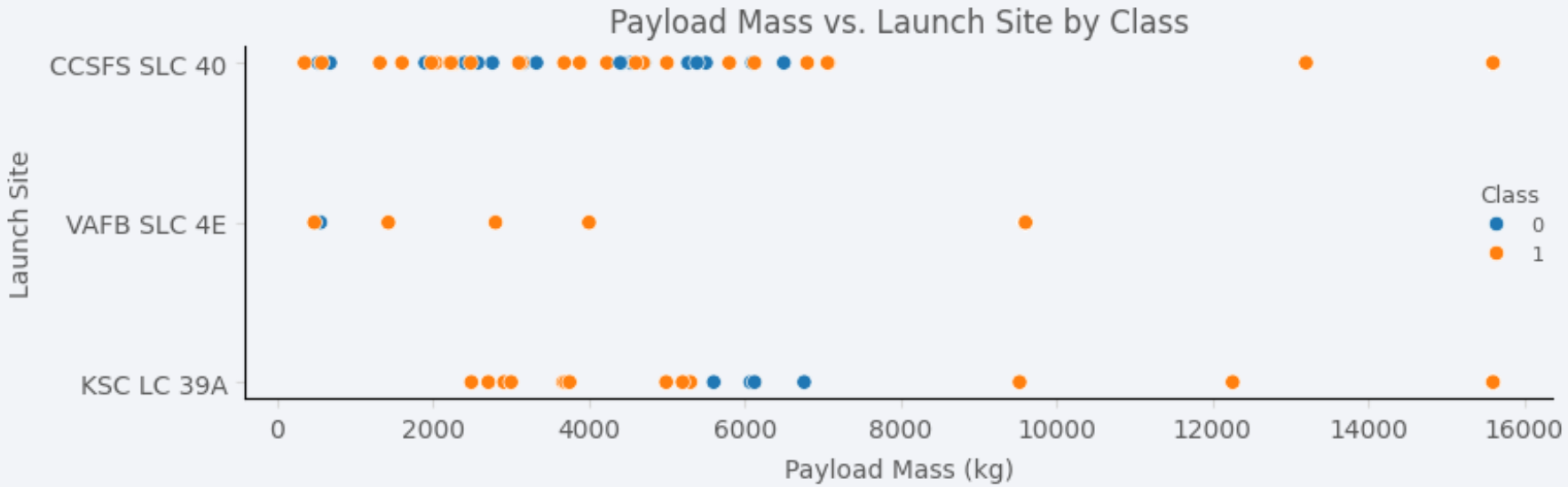
Flight Number vs. Launch Site

- As Flight Number and Payload Mass increases, success rate also increases.
- Launch Site CCAFS SLC 40 has most of the lower Flight Numbers, also concentrating less successful landing.



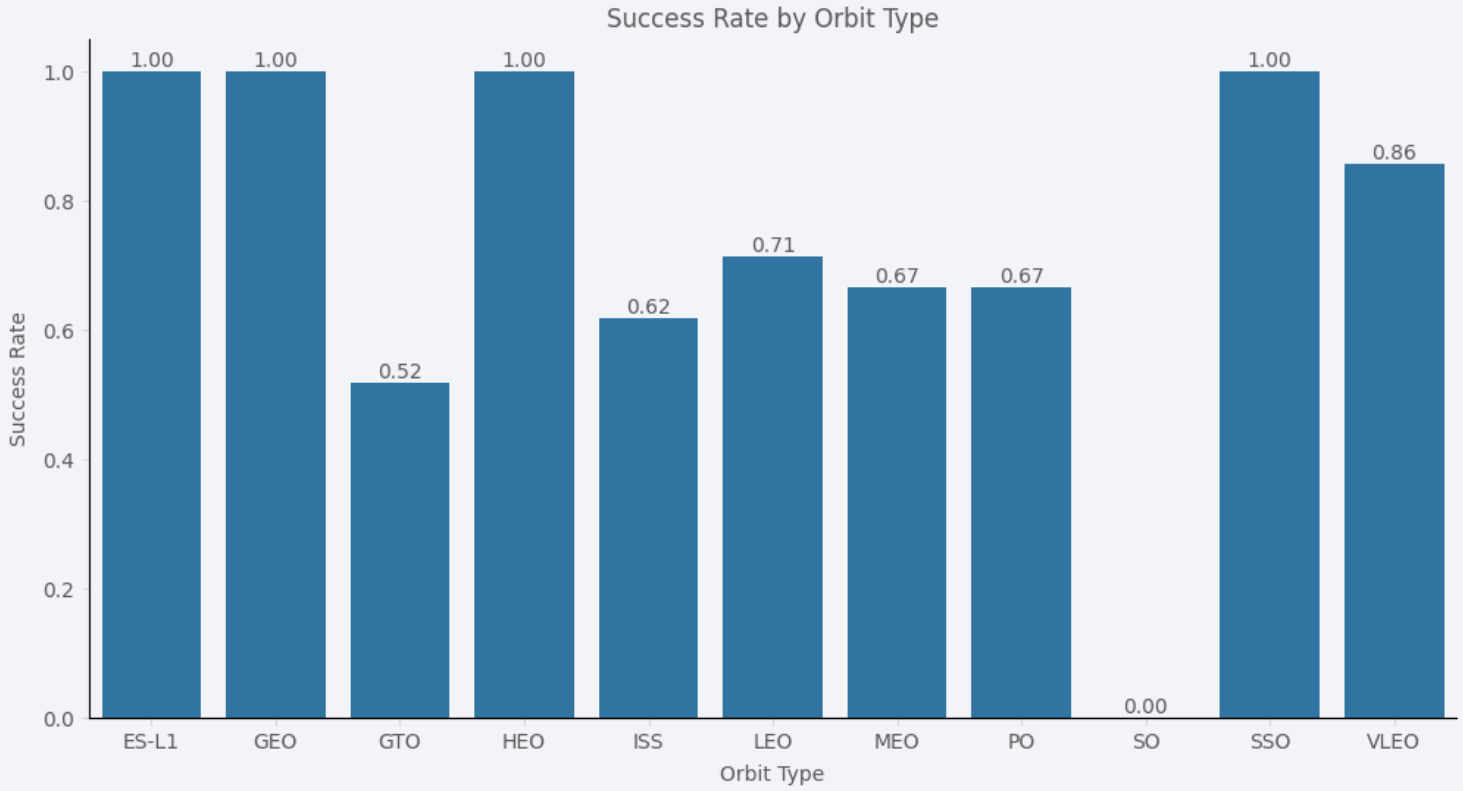
Payload vs. Launch Site

- Launch Site VAFB-SLC does not have launches with heavy payload mass (greater than 10k) and high success rate.
- Launch Site CCSFS SLC 40 has the most launches.



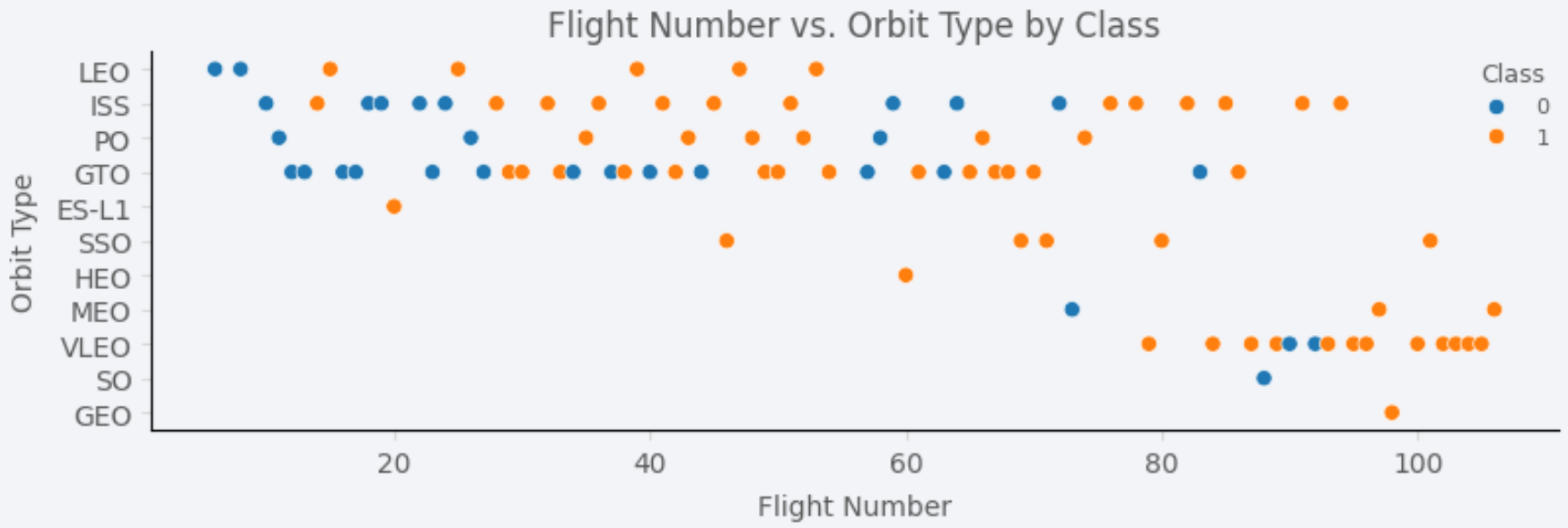
Success Rate vs. Orbit Type

- Success Rate has a high variance across different Orbit Types.
- SO has a 0% success Rate and five Orbit Types have 100% (ES-L1, GEO, HEO, SSO, and VLEO).



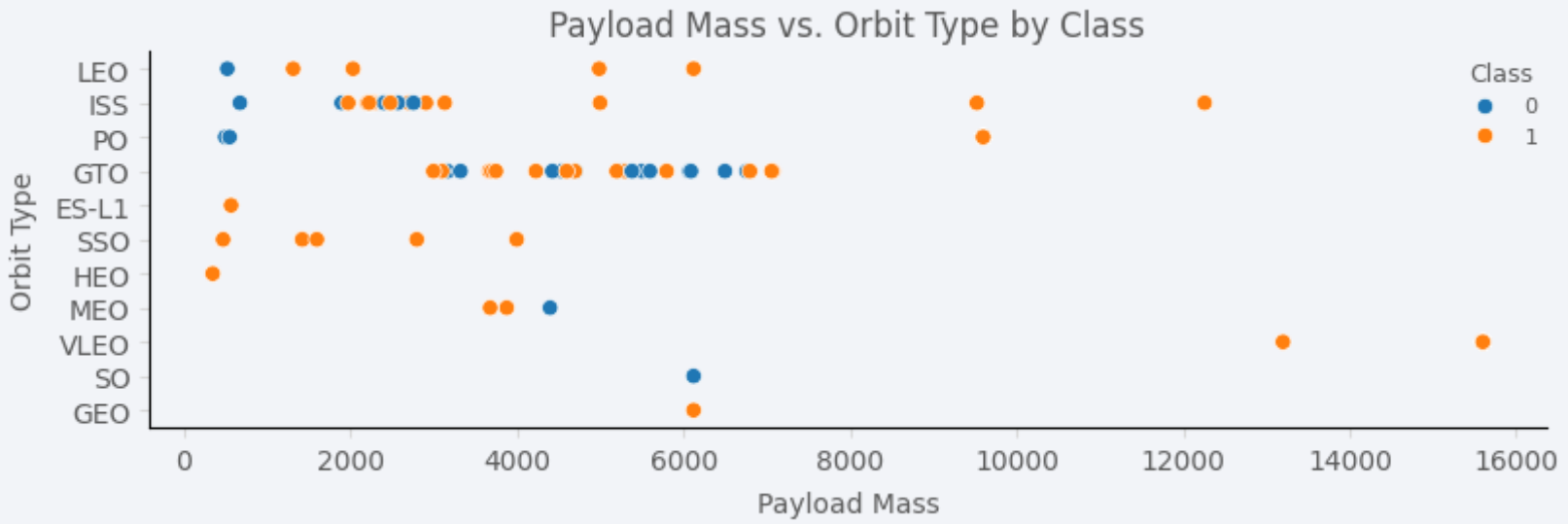
Flight Number vs. Orbit Type

- Orbit Types success rate are influenced by Flight Numbers.
- LEO, ISS, PO, and GTO had some of the lower (older) Flight Numbers that influenced their success on landings.



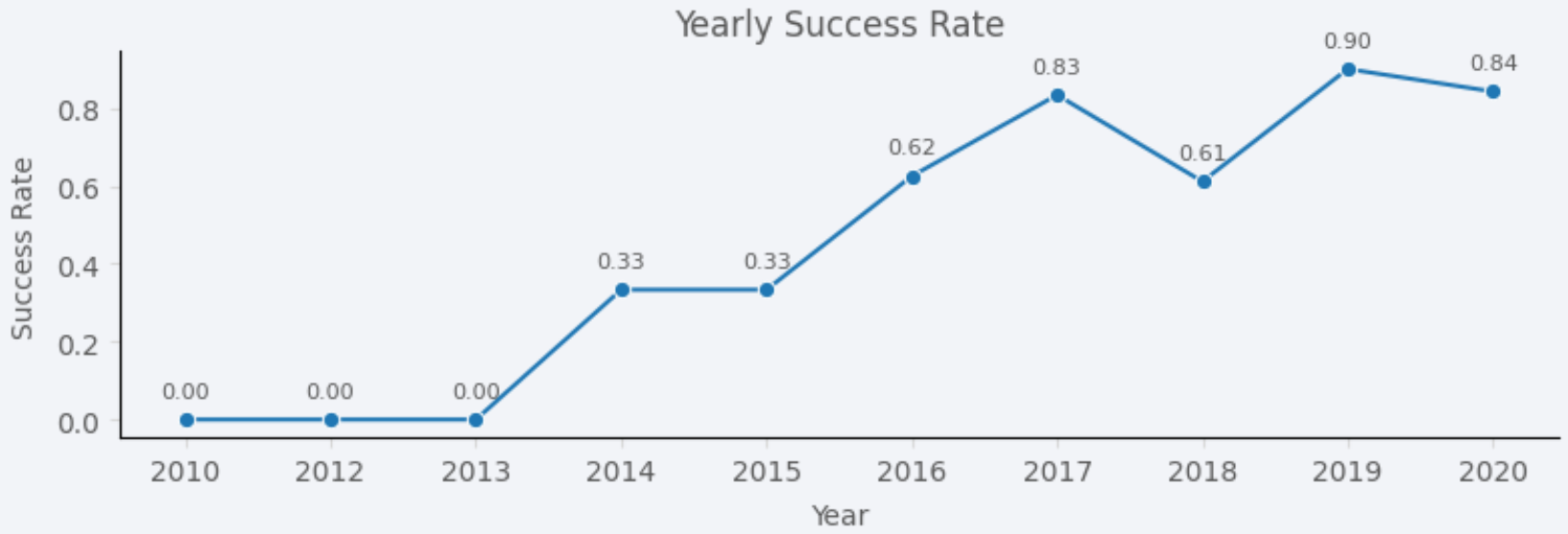
Payload vs. Orbit Type

- Some Orbit Types are applicable for lower Payload Mass.
- The chart reinforces how high Payload Mass seems to be correlated with higher success rate on landings.



Launch Success Yearly Trend

- Success Rate is increasing over the years, indicating that accumulating know-how and implementing new technologies positively influences the success of landings.



All Launch Site Names

- 4 Launch Sites were identified.

```
%sql SELECT DISTINCT Launch_Site FROM spacextable
```

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

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

Launch Site Names Begin with 'CCA'

- LIKE function was used to identify launch sites starting with ‘CCA’.
- The returned table was limited to five rows

```
%sql SELECT * FROM spacextable WHERE Launch_Site LIKE "CCA%" LIMIT 5
```

```
 * sqlite:///spacextable.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

- 45.5 tons of Payload Mass was carried by boosters launched by NASA (CRS).

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM spacextable WHERE Customer = "NASA (CRS)"
```

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- 2.5 tons is the average Payload Mass of F9 v1.1 boosters.
- The query included all F9 v1.1 models.

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM spacextable WHERE Booster_Version LIKE "F9 v1.1%"
```

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

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

- First success in ground pad was achieved in 2015-12-22.

```
%sql SELECT MIN(Date) FROM spacextable WHERE Landing_Outcome = "Success (ground pad)"
```

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

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The four boosters listed in the result of the query had a success in drone ship on payloads between 4k and 6k, with one occurrence each.

```
%sql SELECT Booster_Version, COUNT(*) AS occurrences \
FROM spacextable \
WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND Landing_Outcome = "Success (drone ship)" \
GROUP BY Booster_Version
```

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

Booster_Version	occurrences
F9 FT B1021.2	1
F9 FT B1031.2	1
F9 FT B1022	1
F9 FT B1026	1

Total Number of Successful and Failure Mission Outcomes

- One mission failed during flight. The rest are all successful.
- Data could be further processed to binarize the categories in “Failure” or “Success” only.

```
%sql SELECT Mission_Outcome, COUNT(*) AS occurrences \
FROM spacextable \
GROUP BY Mission_Outcome
```

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

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

Boosters Carried Maximum Payload

- Subquery was used to get the maximum payload mass.
- 12 boosters carried its maximum.

```
%sql SELECT Booster_Version \
FROM spacextable \
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM spacextable)
```

```
 * sqlite:///spacextable.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

- Two failures on drone shipping were recorded in 2015, both on CCAFS LC-40.

```
%sql SELECT substr(Date, 6, 2) as Month_Name, Booster_version, Launch_Site, Landing_Outcome \
FROM SPACEXTBL \
where Landing_Outcome = "Failure (drone ship)" and SUBSTR(Date, 0, 5)= "2015"
```

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

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

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

- GROUP BY and ORDER BY clauses were used to get the results.
- Top 1 is not-attempted landings.

```
%sql SELECT Landing_Outcome, COUNT(*) as occurrences \
FROM SPACEXTBL \
WHERE Date BETWEEN "2010-06-04" AND "2017-03-20" \
GROUP BY Landing_Outcome \
ORDER BY occurrences DESC
```

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

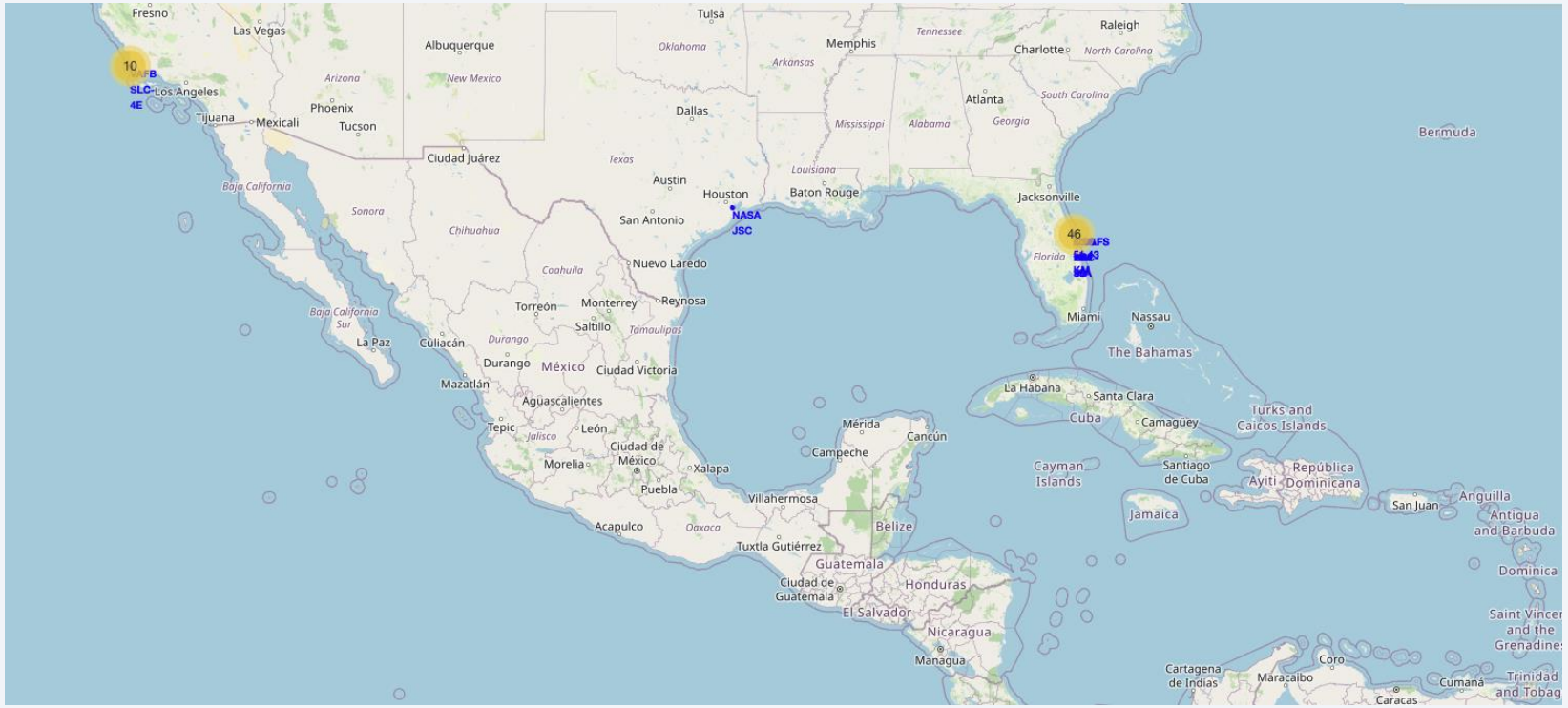
Landing_Outcome	occurrences
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

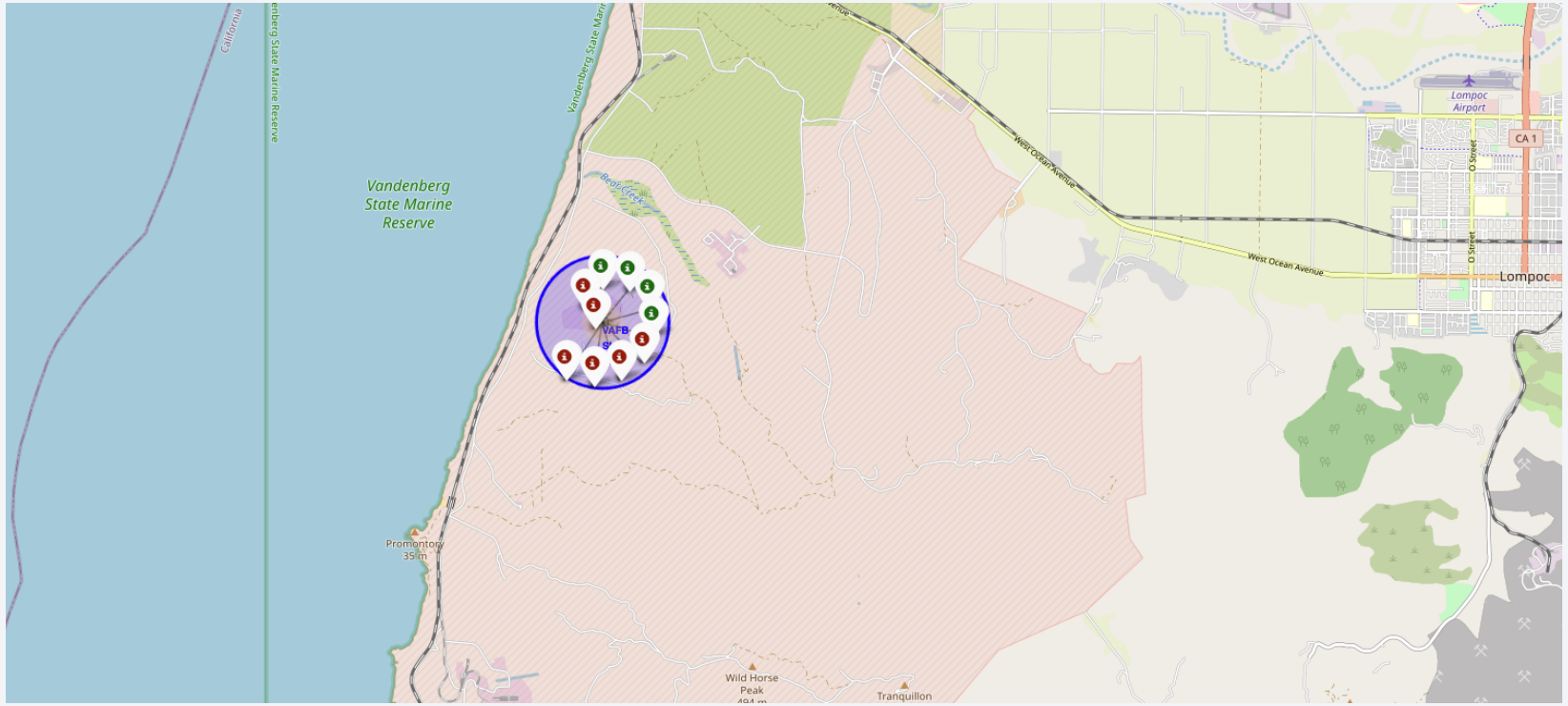
Launch Sites across Global Map

- Launch Sites are in the South of USA (closer to Equator Line).
- They are also close to the coast and three of them are in Cape Canaveral.



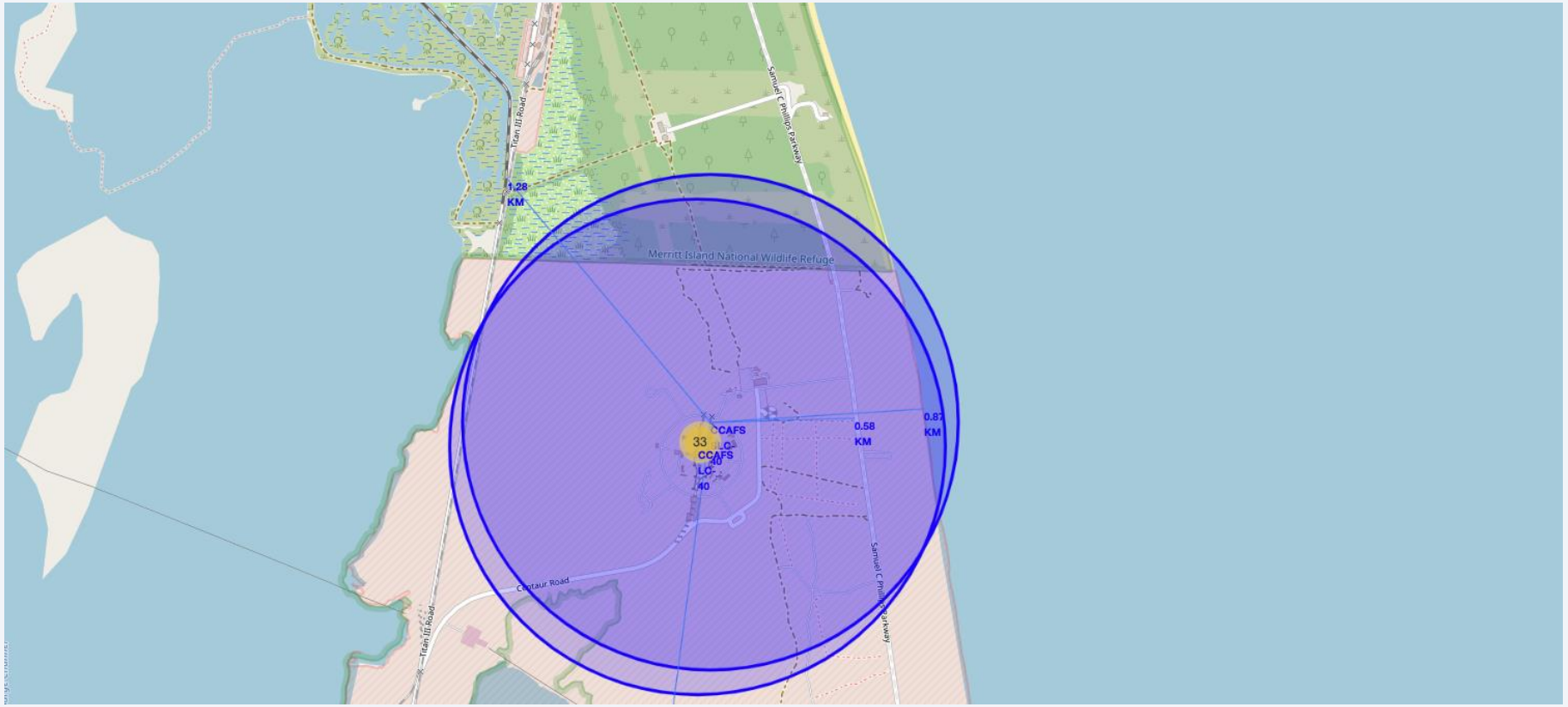
Launches with Successful Landings

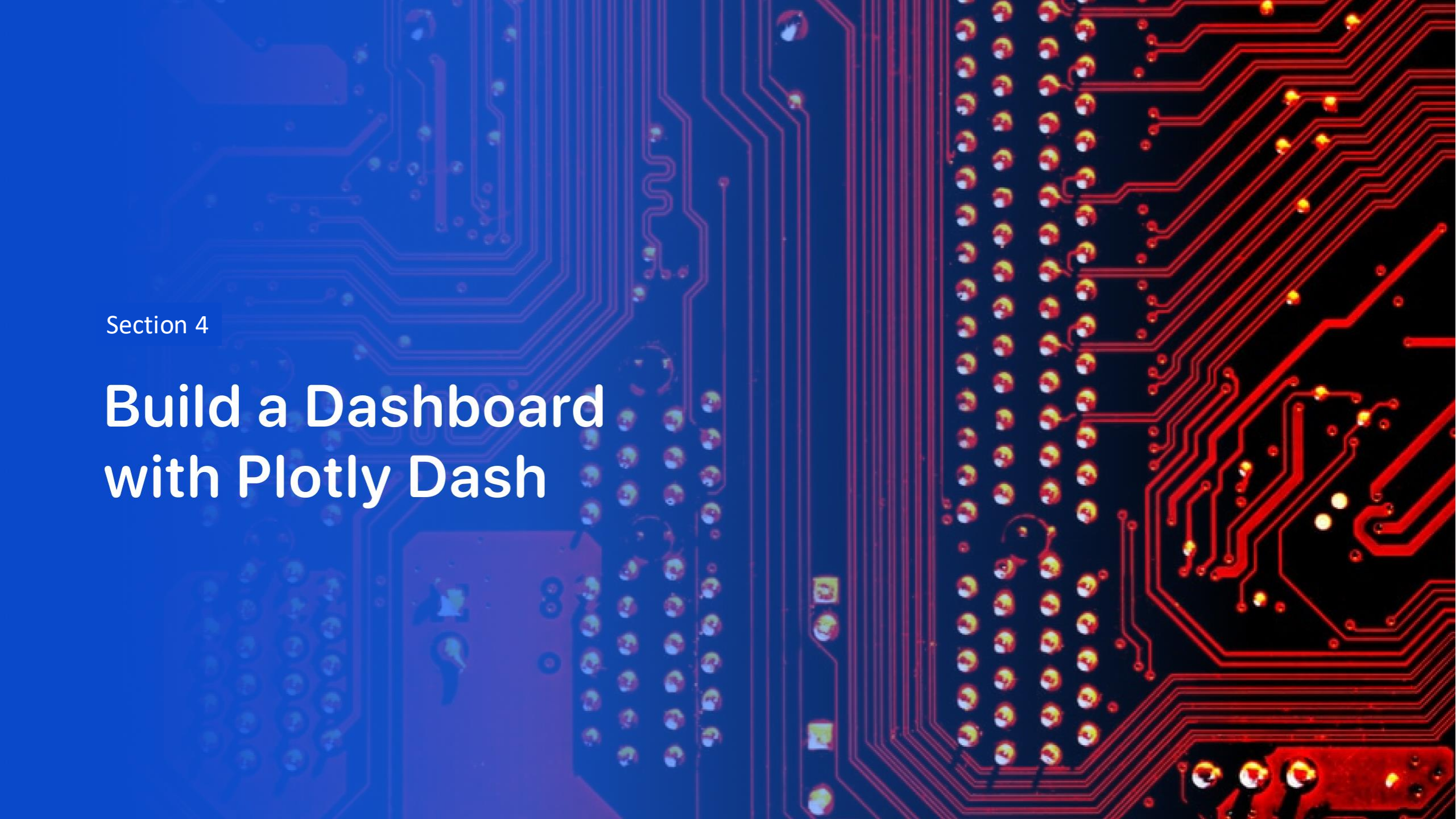
- A Marker cluster was created to color code the outcome of the landings for all launch sites.
- This gives a quick overview of the success rate of each site.



Launch Sites Infrastructure

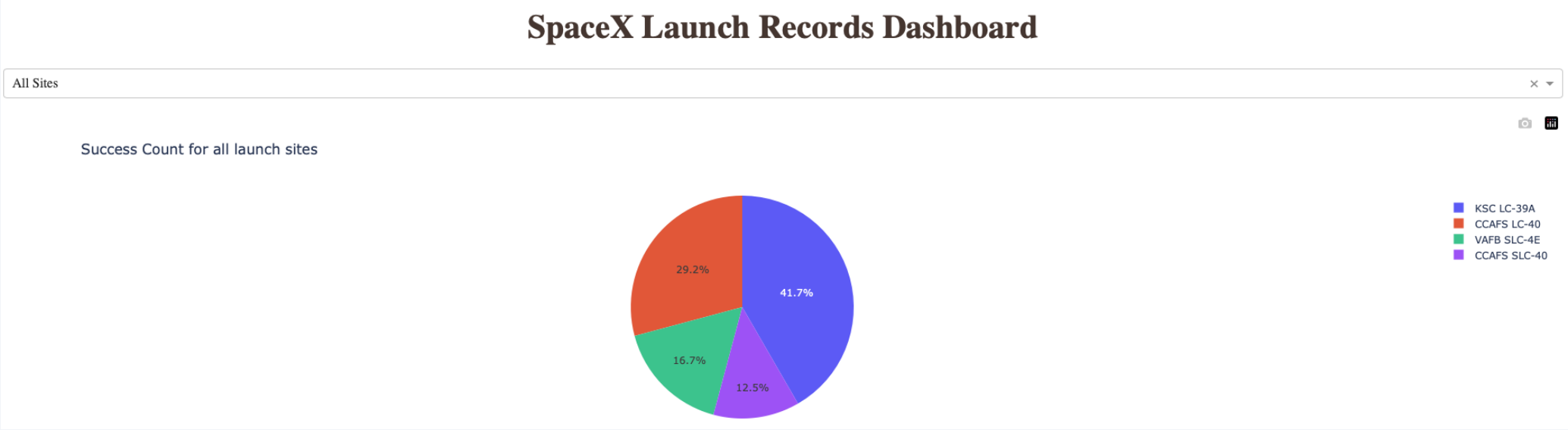
- Coastline, Highway and Railway are all within 1 mile of CCAFS SLC-40 Launch Site. The city is 50km from there.
- That indicates the strategic approach to infrastructure for the launch sites.




Section 4
Build a Dashboard with Plotly Dash

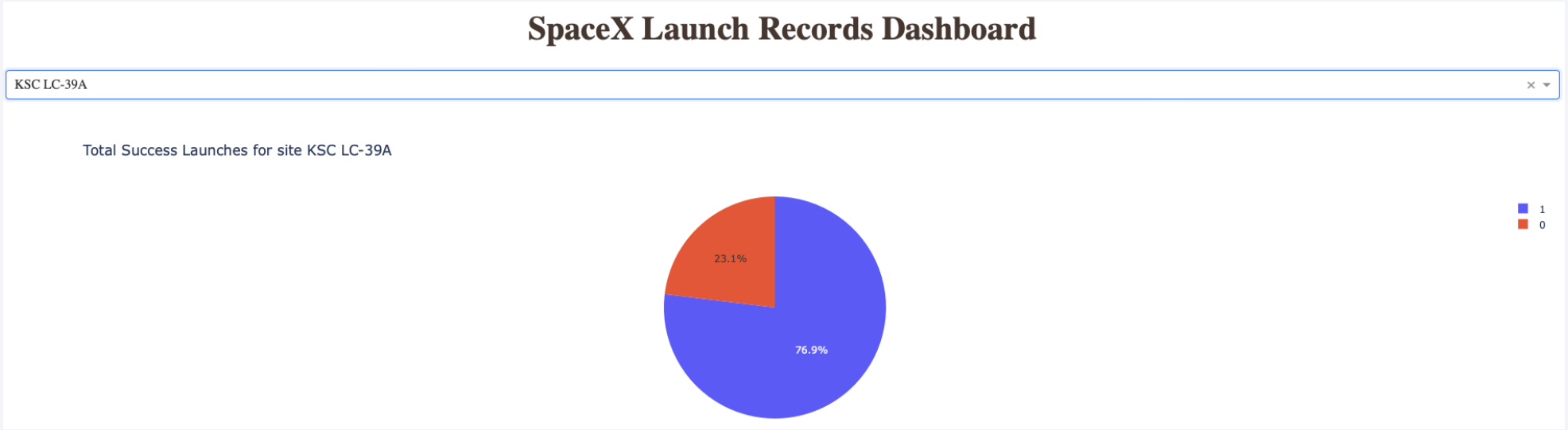
Total Success Launches by Site

- Most successful launches come from KSC LC-39A.
- The distribution is quite even for the four different sites. The number of attempted compared to successful launches is required for a deeper analysis.



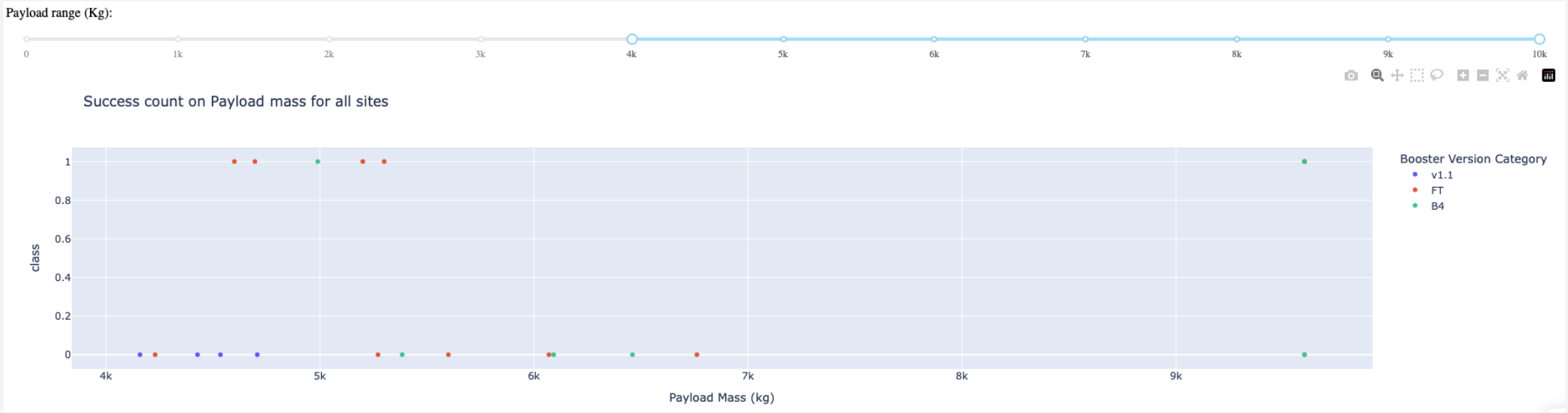
Highest Successful Launches Site

- KSC LC-39A had 13 launches, with 10 being successful.
- This launch site can be considered a benchmark to all others.



Payload vs Launch Outcome by Booster Version

- With a payload over 4k, there are a higher percentage of failures for all sites.
- Booster version v1.1 did not have a single success.

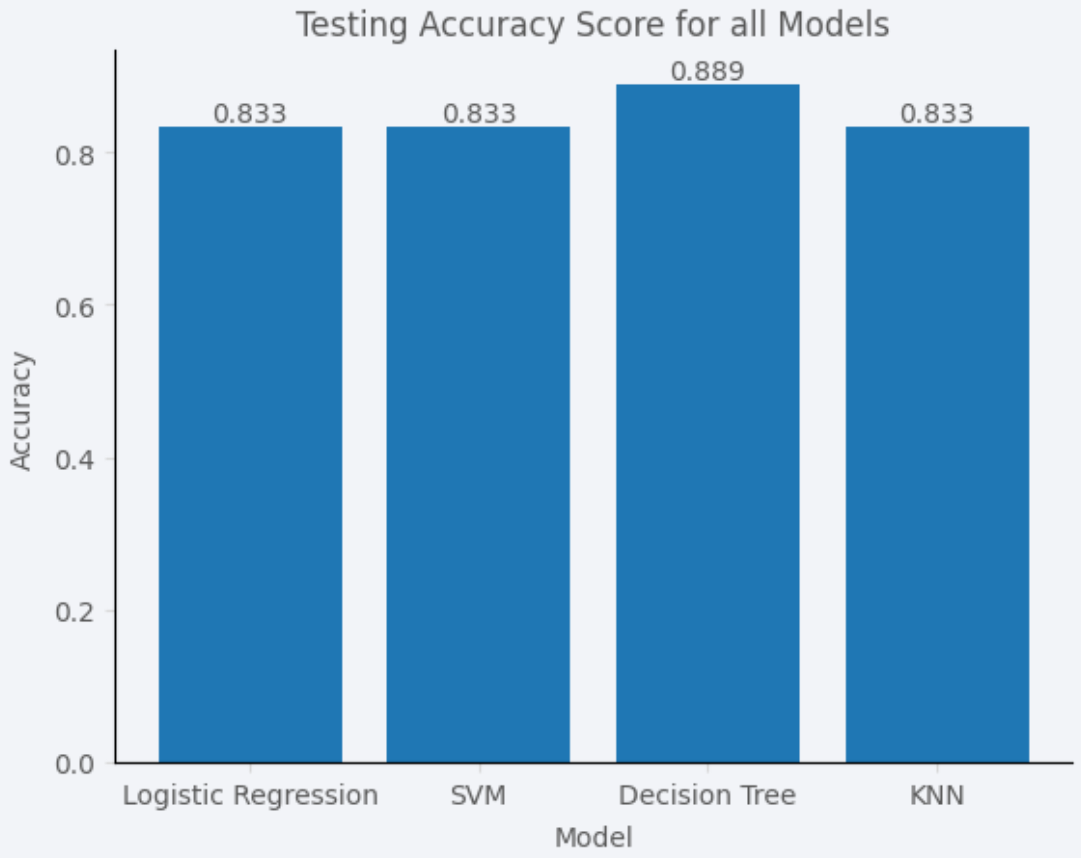


Section 5

Predictive Analysis (Classification)

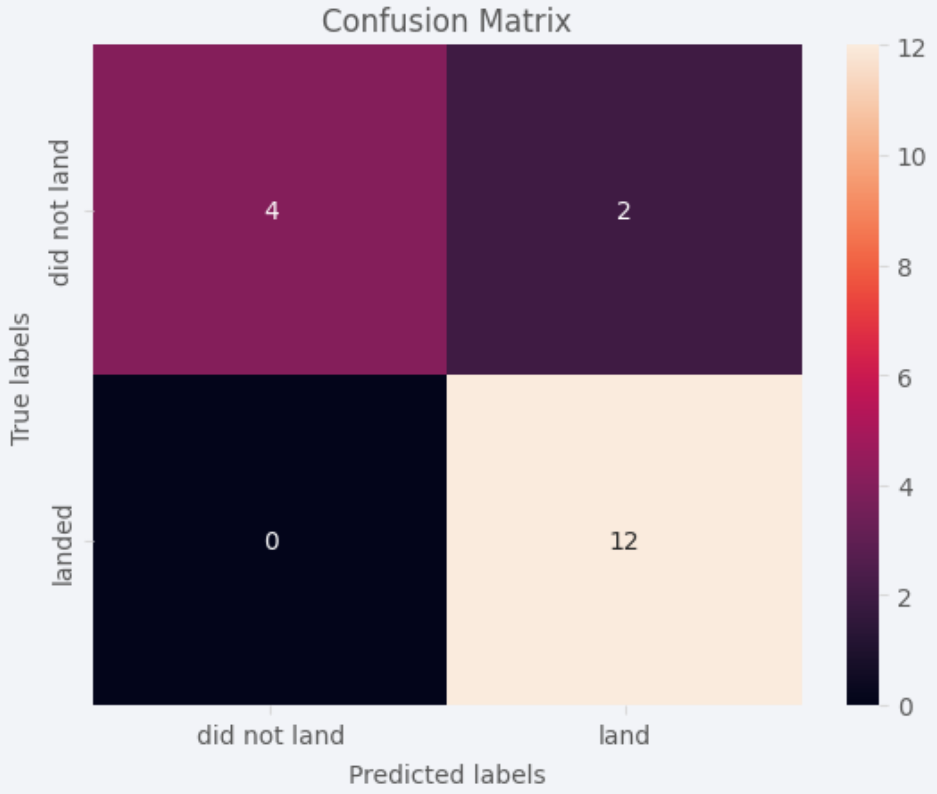
Classification Accuracy

- Decision Tree outperformed all other models on testing dataset. Thus, was chosen as final model.



Confusion Matrix

- Model achieved 88% of accuracy, predicting the correct outcome in 16 of the 18 landings on unseen data.
- The two errors are False Positives: when the model predicts a successful landing, but an unsuccessful one happened.



Conclusions

- The model had a satisfactory performance, with 88% accuracy in its first iteration – indicating a potential to go into production and deliver value.
- The EDA presented a complete overview for any kind of stakeholders, enabling the problem to be viewed from different perspectives, including technical and infrastructure aspects.
- Machine Learning Explainability techniques can be employed to understand the features even better, leading to further development of the model.
- Having more data for training and validation is also important - the model was validated using only 18 records.
- Other scores for validation and more complex machine learning models could be used – the model still has a high error on False Positives.

Appendix

- GitHub project folder: https://github.com/ewerthonk/ml-spacey.
- Notebooks on GitHub are not exact copies from the one from the labs. They were developed from ground and may have different architectures compared to the labs.
- All data used in the notebooks and python files is stored on “/data” folder on the GitHub project.

Thank you!