IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Wouter R
2024/03/07

Outline

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

Executive Summary

- In this capstone project we will use machine learning classification algorithms to determine if the SpaceX Falcon9 will land successfully.
- The main methods used to achieve our goal are:
  * Data collection, wrangling and formatting
  * Exploratory data analysis
  * Interactive data analyzation
  * Machine learning prediction
- Conclusion

Introduction

- In this capstone, we will predict if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch.
- The main question that we are trying to answer in this capstone is: Will the first stage of the SpaceX Falcon g rocket landsuccesfully. We are using a dataset including features such as Payload Mass, Launch Site, Orbit Type etc. To answer this question.

Section 1

Methodology

Methodology

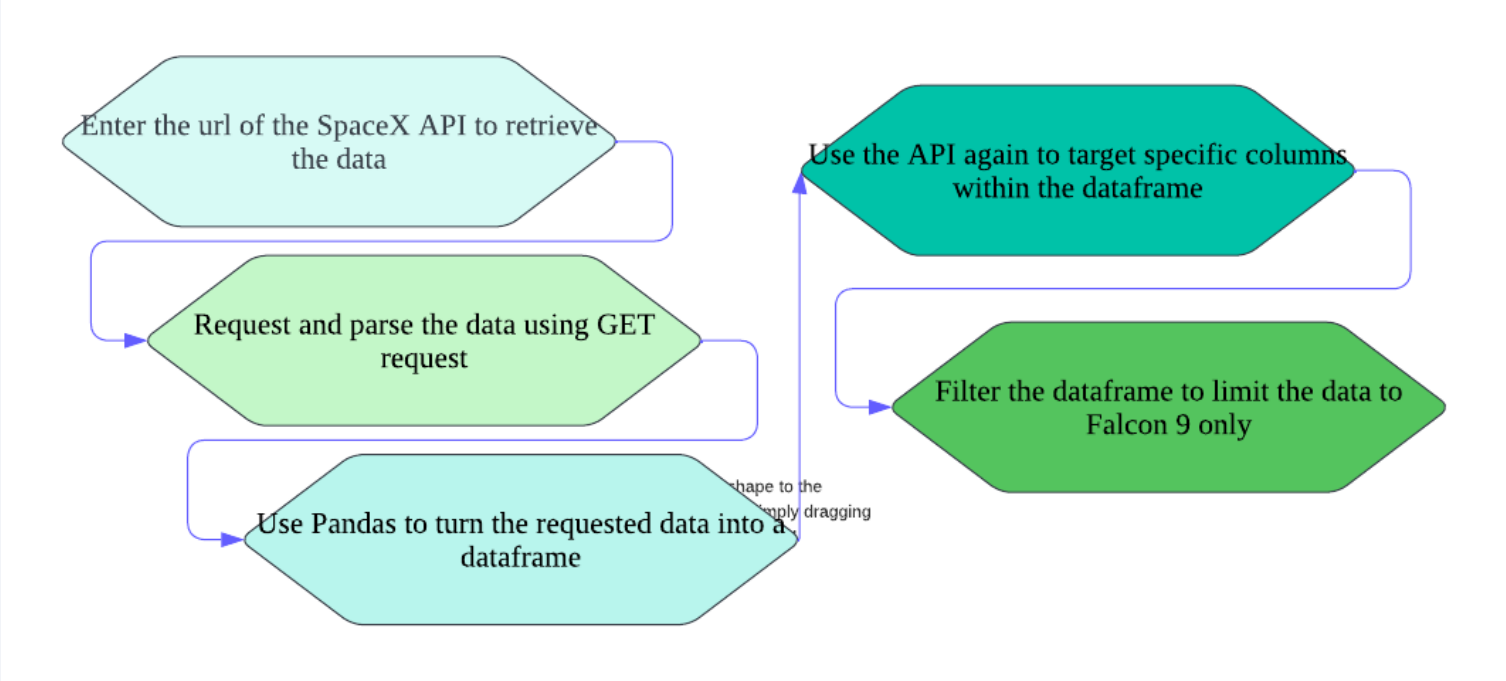
- Data collection, wrangling and formatting using:
  - SpaceX API
  - Webscraping
- Exploratory data analysis (EDA) using:
  - Pandas & Numpy,
  - SQL
- Interactive visual analytics using:
  - Matplotlib, Seaborn
  - Folium, Plotly
- Machine learning prediction using:
  - Logistic regression
  - Support Vector Machine (SVM)
  - Decision tree
  - K-nearest neighbors (KNN)

Data Collection

- The data sets are collected by using:
- SpaceX API
- Webscraping

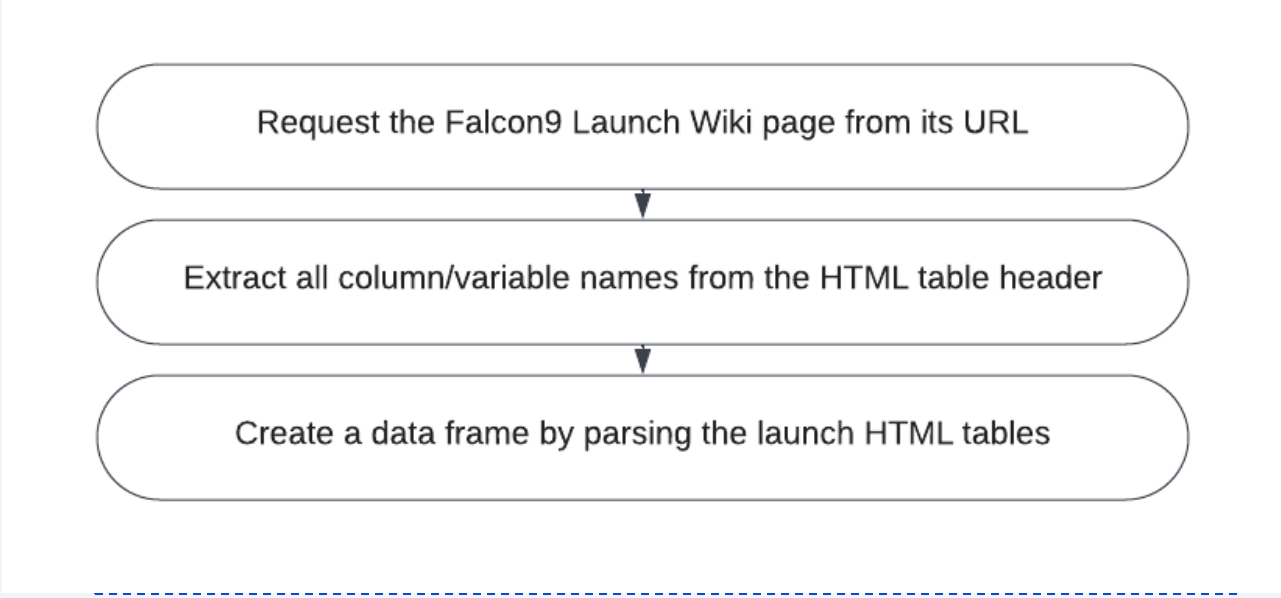
Data Collection – SpaceX API

- Used API: https://api.spacexdata.com/v4/launches/past
- The data provided this way include data from many launches, therefore they are filtered to include only the data for Falcon 9
- Missing data are replaced by the mean value of the values in that column
- For a closer inspection of the data see :https://github.com/Woutusa/IBM-AppliedDataScience-CapstoneProject/blob/main/01%20jupyter-labs-spacex-data-collection-api.ipynb



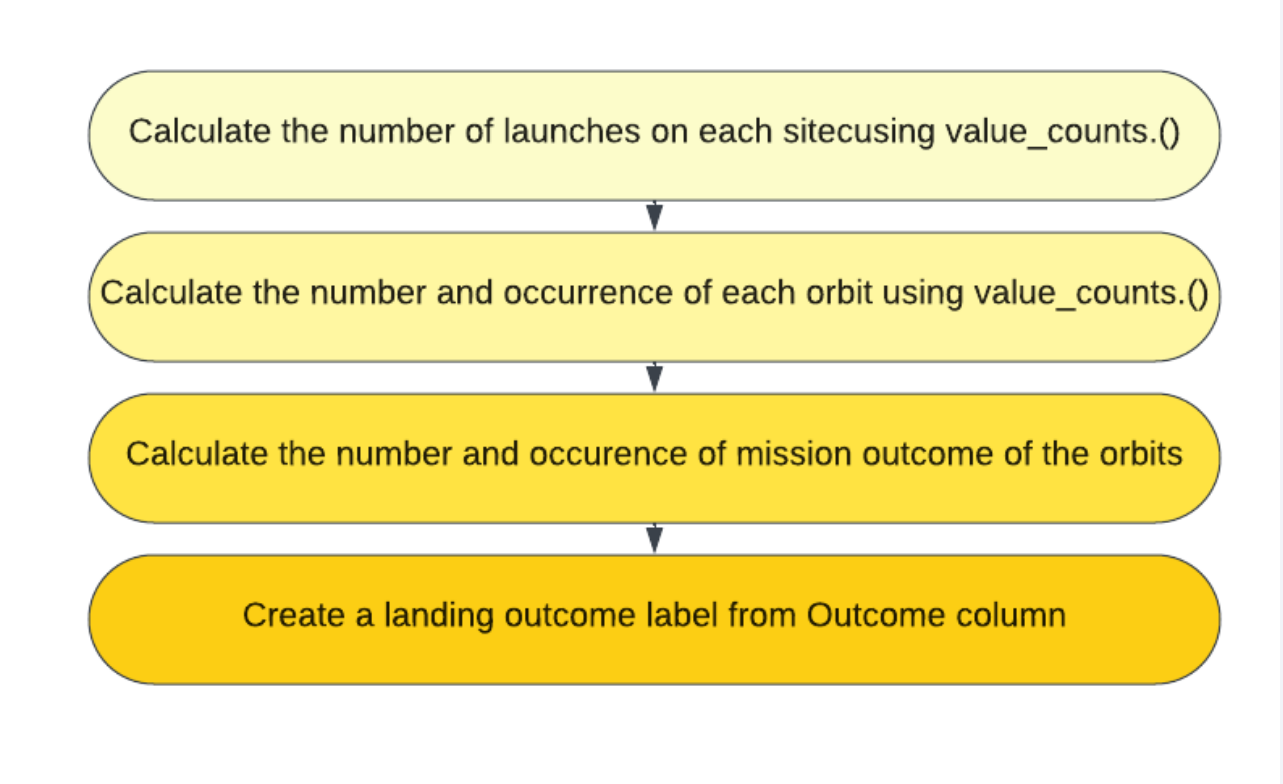
Data Collection – Scraping

- Web scrap Falcon 9 launch records with BeautifulSoup from: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- Parse the table and convert it into a Pandas data frame
- https://github.com/Woutusa/IBM-AppliedDataScience-CapstoneProject/blob/main/02%20jupyter-labs-webscraping.ipynb



Data Wrangling

- Identify and calculate the percentage of the missing values in each attribute
- Calculate number of launches, occurrence of each orbit and mission outcome of the orbits, using these data create a landing outcome column
- https://github.com/Woutusa/IBM-AppliedDataScience-CapstoneProject/blob/main/03%20labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

Charts used for Data Visualization:

- *Scatter Plot*: used to visualize relationship between: flight number & payload mass; flight number & launch sites; payload & launch sites; flight number & orbit type; payload & orbit type
- *Bar chart*: used to visualize the success rate of each orbit
- *Line Plot*: used to visualize relationship between success rate and date

- https://github.com/Woutusa/IBM-AppliedDataScience-CapstoneProject/blob/main/05%20jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first succesful landing outcome in ground pad was acheived.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
- 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.
- 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.

- https://github.com/Woutusa/IBM-AppliedDataScience-CapstoneProject/blob/main/04%20jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Marked all launch sites on a map
- Marked the success/failed launches for each site on the map to enhance visualization of which launch sites have a high success rate
- Calculated the distances between a launch site to its nearest landmarks like railroads, coastlines, highways,...

- https://github.com/Woutusa/IBM-AppliedDataScience-CapstoneProject/blob/main/06%20lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

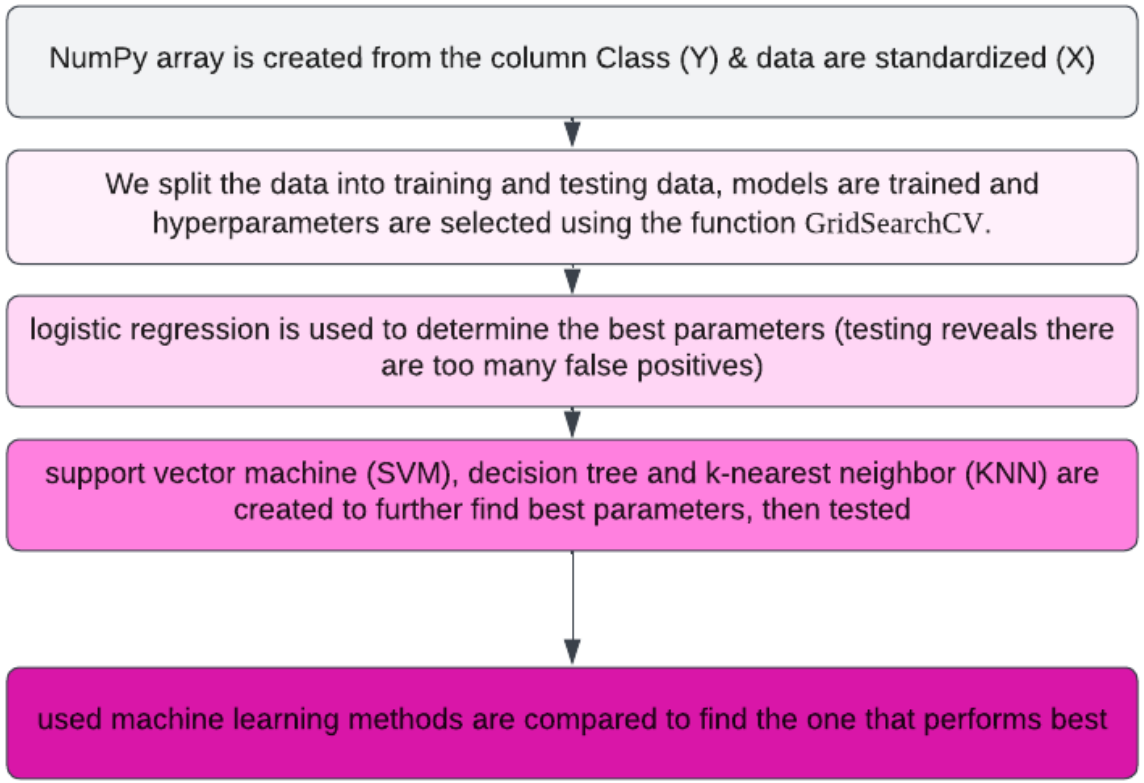
- Used pie charts and scatter plots to visualize the launch records of SpaceX
- Charts display launch success rate by launch site
- Provide inside in influences on launch success rate by factors like payloadmass and booster versions

- https://github.com/Woutusa/IBM-AppliedDataScience-CapstoneProject/blob/main/07%20Ploty%20Dash%20Interactive%20Dashboard

Predictive Analysis (Classification)

- SciKit-Learn allows us to split our data into training and testing data to find the best performing classification model

- https://github.com/Woutusa/IBM-AppliedDataScience-CapstoneProject/blob/main/08%20SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

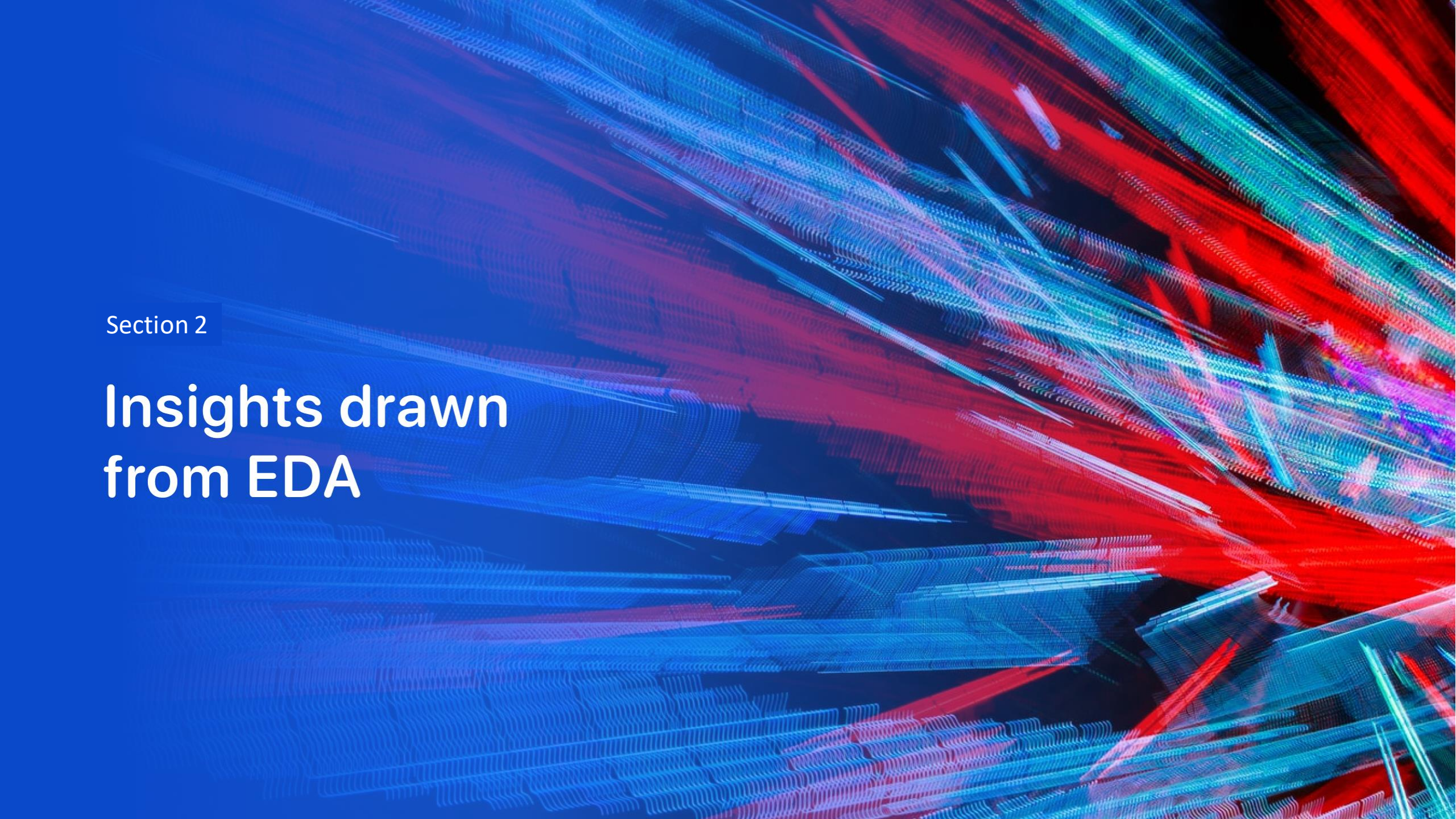


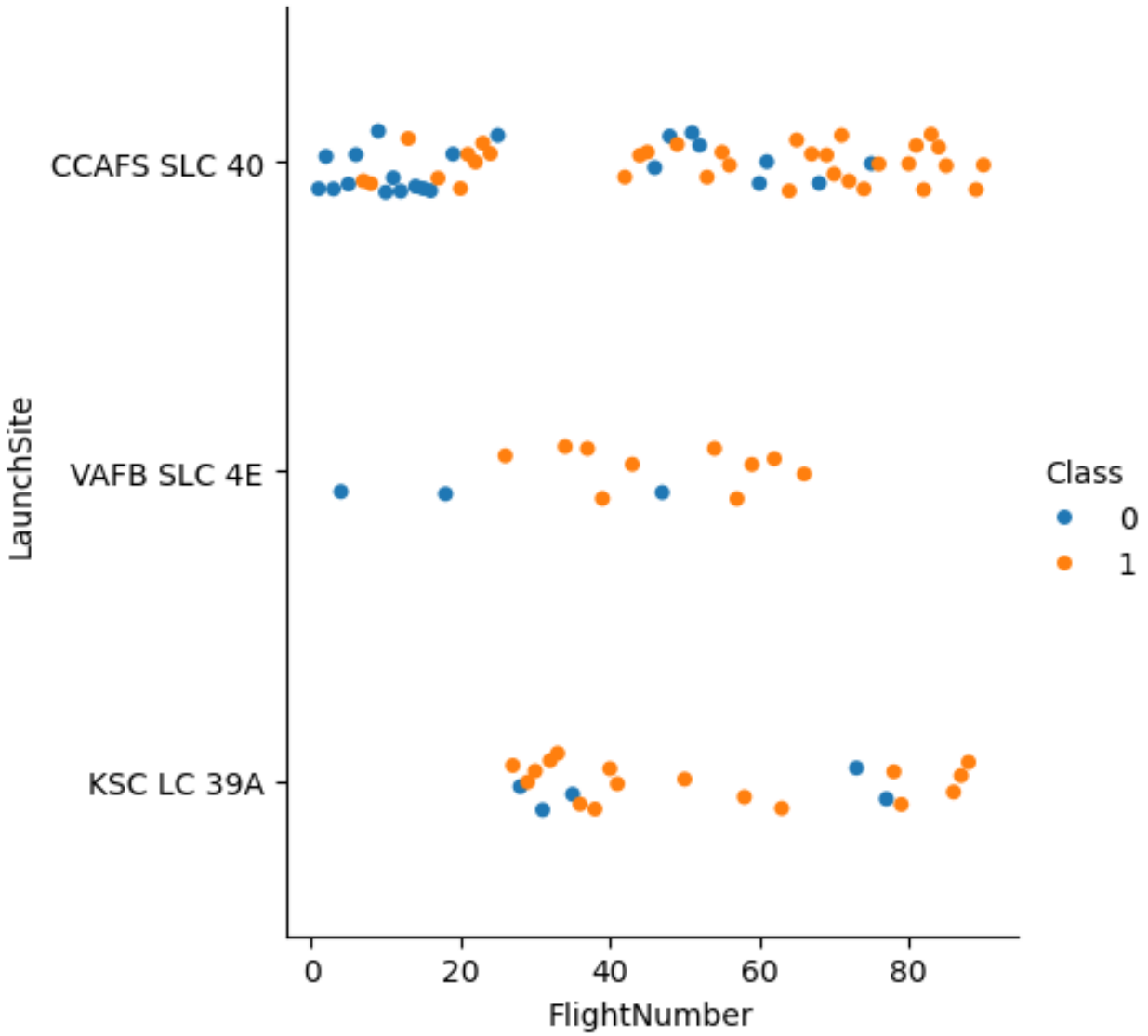
Results

- Exploratory data analysis showed us that there was an increase in successful launches after 2013.
- It also showed a correlation between successful landings and flight number.
- Launch sites are aligned along the coast line, this makes landings in the water a possibility.
- The allocation of launch sites near highways and railroads makes for a facilitation of transport.
- The machine learning methods were able to predict the success rate of landings with a 83.334% accuracy ratio.

Section 2

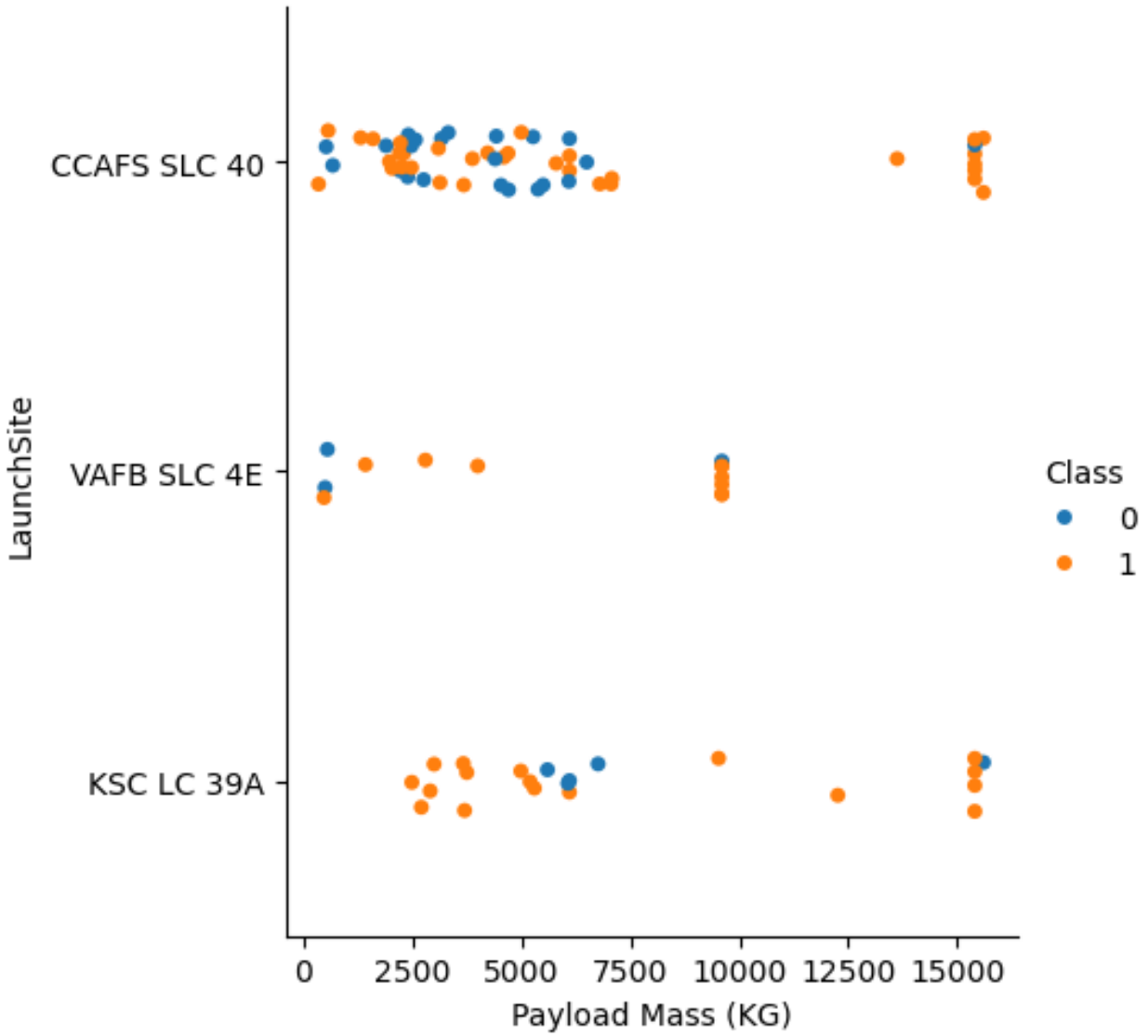
Insights drawn from EDA





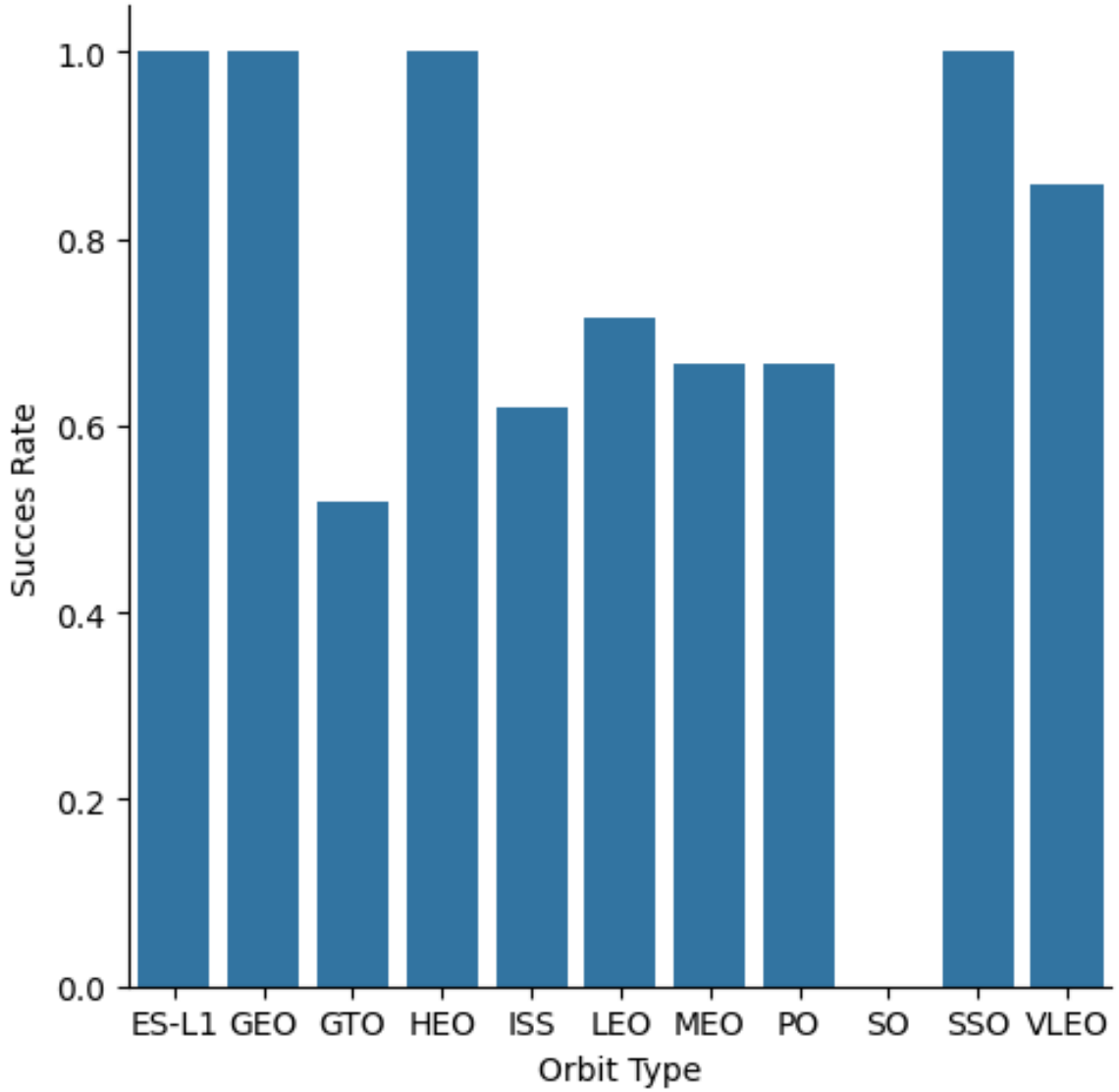
Flight Number vs. Launch Site

- The site CCAFS SLC 40 has the largest number of launches.
- The higher the flight number, the higher the rate of success.



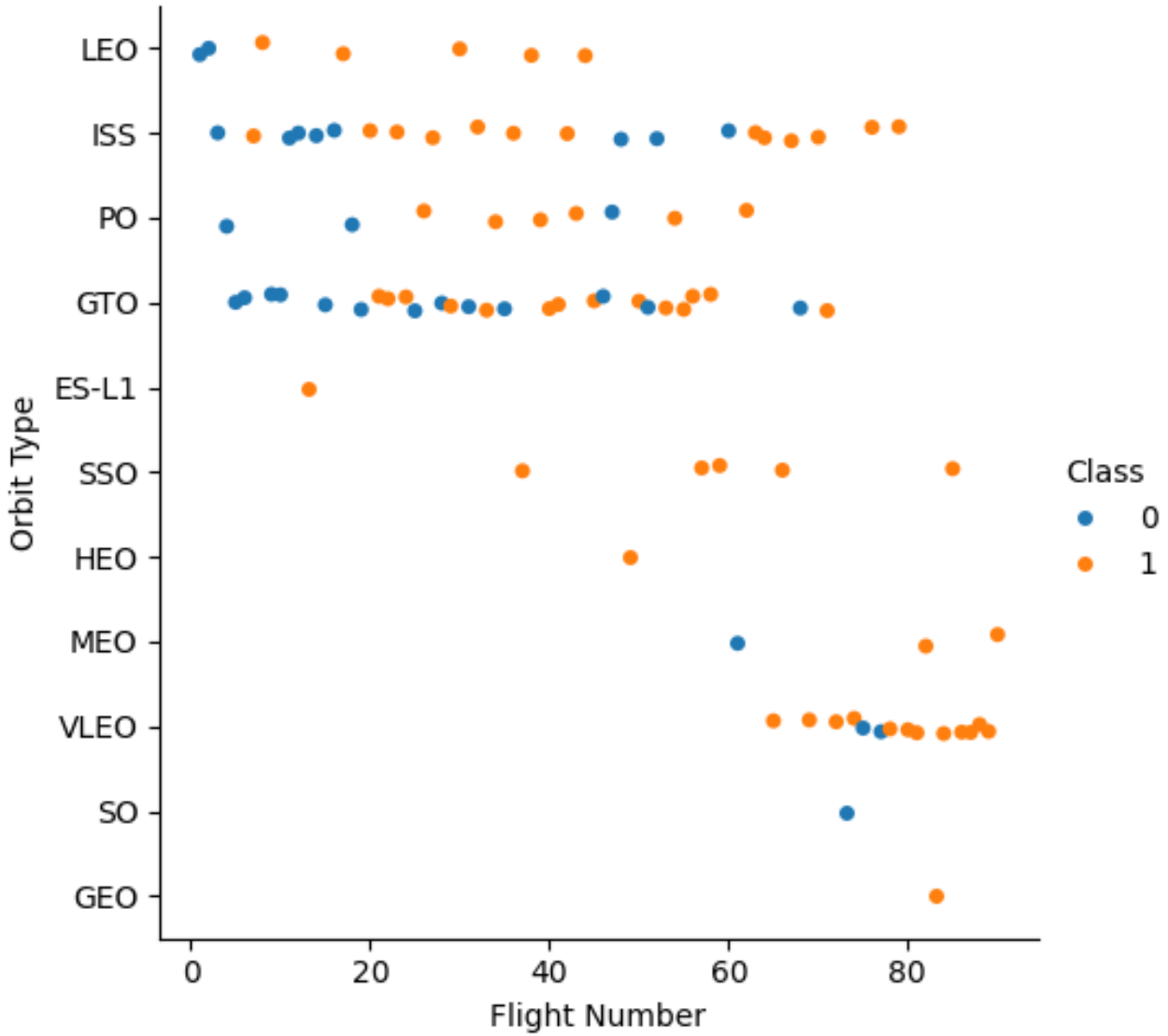
Payload vs. Launch Site

- At the CCAFS SLC 40 launch site there are less launches with a payload mass of over 7500kg, but they are relatively more successful.
- At the VAFB SLC 4E site there are no launches with a payload mass larger than 10 tons.



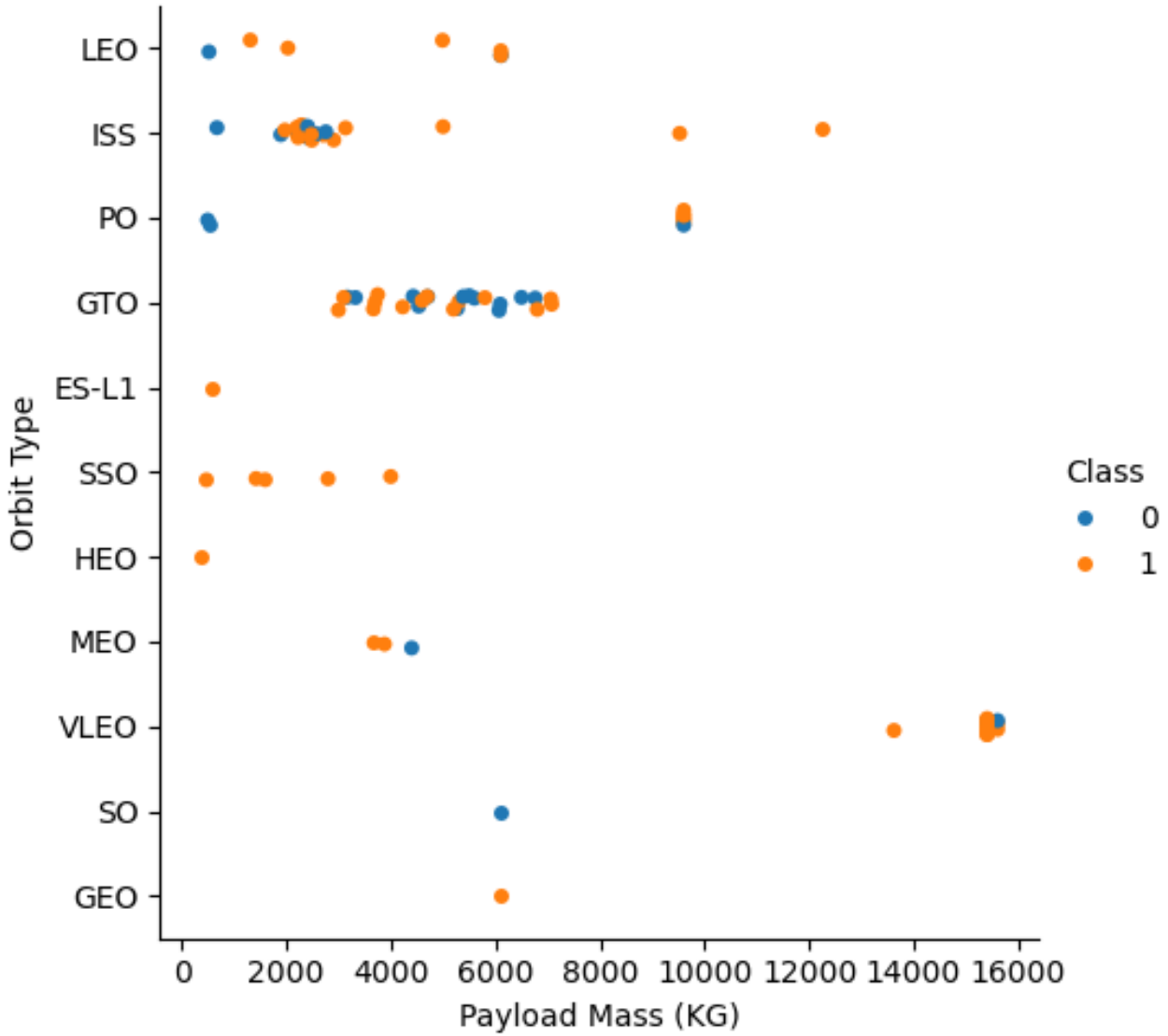
Success Rate vs. Orbit Type

- The orbit types with the highest success rates are:
  - ES-L1
  - GEO
  - HEO
  - SSO



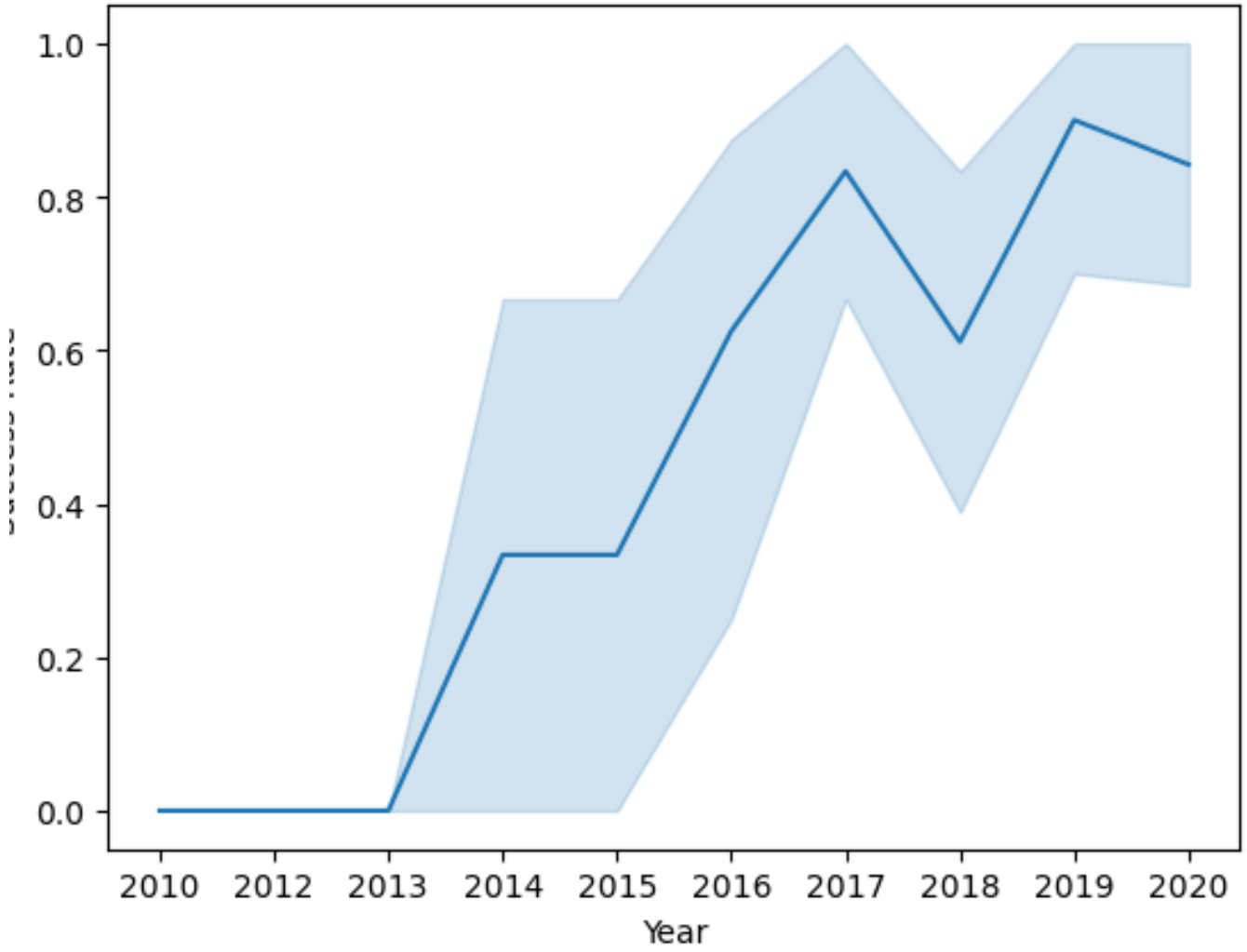
Flight Number vs. Orbit Type

- For LEO and PO orbits the success appears related to the number of flights
- there seems to be no relationship between flight number when in GTO orbit



Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- For GTO we cannot make this distinction as both successful and unsuccessful missions are present.



Launch Success Yearly Trend

Notice that the success rate kept increasing between 2013 and 2020.

All Launch Site Names

- Below are the names of the launch sites and their respective latitudes and longitudes, these were taken from spacex_launch_geo.csv and used to plot the launch sites' location on the map.

Launch Site	Latitude	Longitude
CCAFS LC-40	28.562302	-80.577356
CCAFS SLC-40	28.563197	-80.576820
KSC LC-39A	28.573255	-80.646895
VAFB SLC-4E	34.632834	-120.610745

Launch Site Names That Begin with 'CCA'

The SQL code shown to the right reveals 5 launch site names that begin with the letters 'CCA'. Note that use of these sites is not exclusive to SpaceX and that other space companies like NASA are also making use of these sites.

Display 5 records where launch sites begin with the string 'CCA'

```
%sql SELECT * FROM SPACEXTBL 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
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success
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
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success

Total Payload Mass

- The below SQL code calculates the total payload mass carried by boosters launched by NASA.

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE CUSTOMER = 'NASA(CRS)';
```

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

SUM(PAYLOAD_MASS__KG_)
None

Average Payload Mass by F9 v1.1

- The below SQL code calculates the average payload mass carried by booster F9 v1.1 from SpaceX.
- The average payload mass is 2928.4 kg.

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1' ;
```

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

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- The below SQL code calculates the date of the first successful landing date.
- The first successful landing date was

```
%sql SELECT MIN(DATE) FROM SPACEXTBL 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

```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ B
```

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

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

The SQL code calculates and shows the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

The SQL code calculates and shows the total number of successful and failure mission outcomes.

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

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

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

Boosters Carried Maximum Payload

- This list shows the names of the booster which have carried the maximum payload mass, on top is the SQL code thatled to this result.

```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL \
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

```
 * 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

This list shows the failed landing outcomes in drone ship, their booster versions, and launch site names for the year 2015.

```
%sql SELECT substr(Date,6,2)as month, BOOSTER_VERSION,LAUNCH_SITE,LANDING_OUTCOME\
FROM SPACEXTBL \
WHERE LANDING_OUTCOME = 'Failure (drone ship)' and substr(Date,0,5)='2015'
```

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

month	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)

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

- This code provides us with a list that has the count of all landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

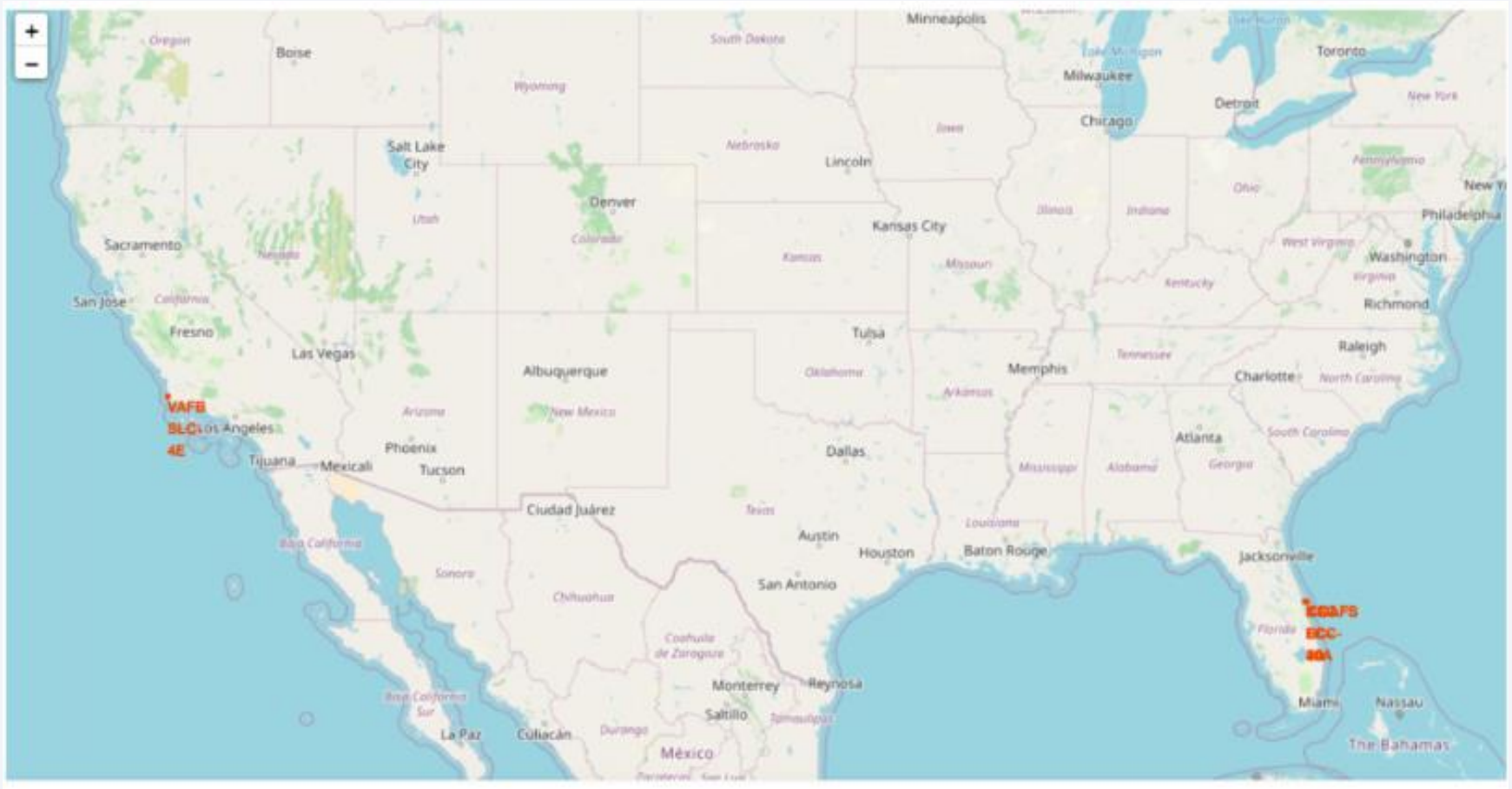
```
%sql SELECT LANDING_OUTCOME, COUNT(*)AS COUNT_OUTCOMES\
FROM SPACEXTBL\
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'\
GROUP BY LANDING_OUTCOME \
ORDER BY COUNT_OUTCOMES DESC
```

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

Landing_Outcome	COUNT_OUTCOMES
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

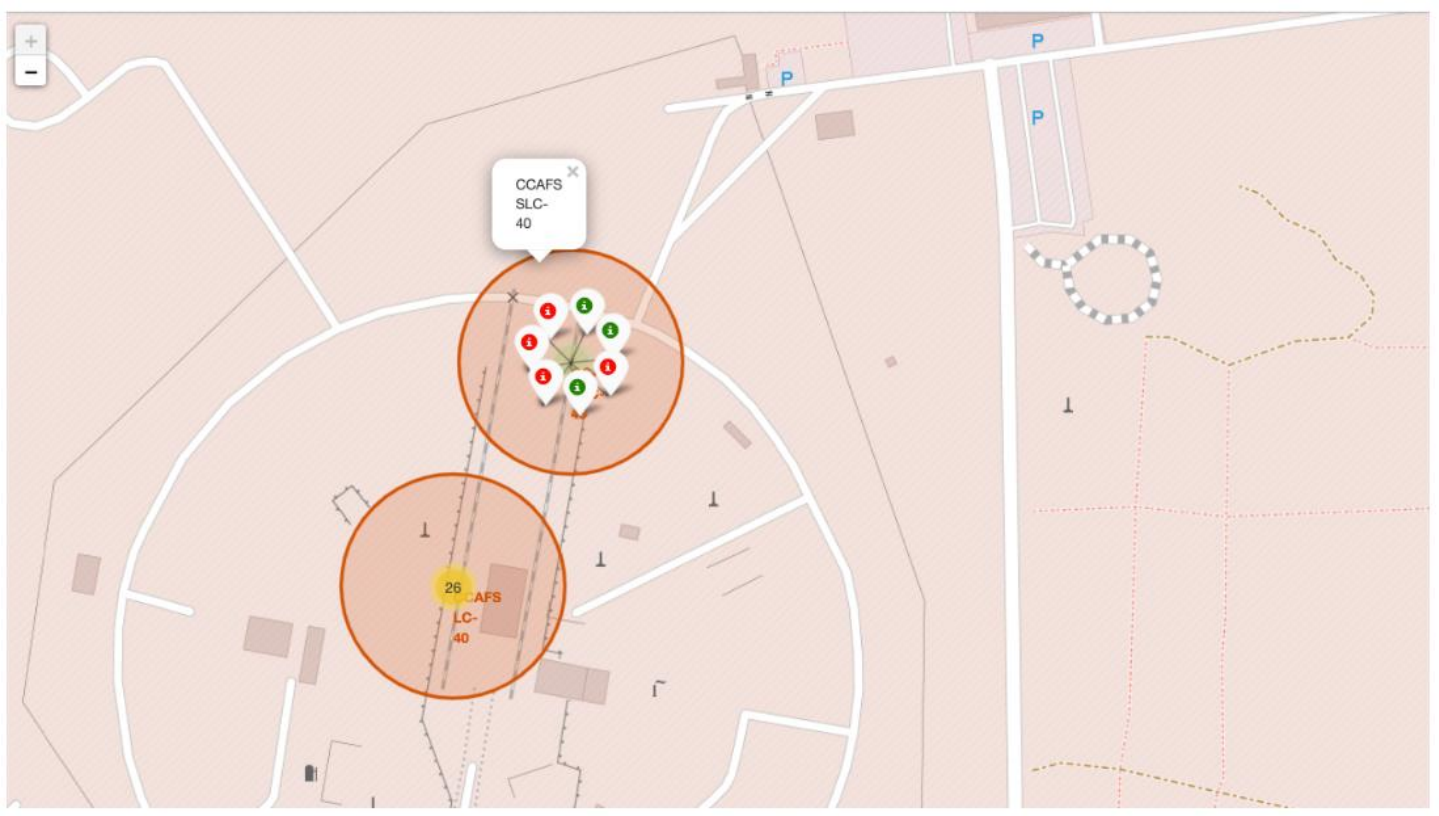

Section 3
Launch Sites
Proximities Analysis

Map showing all launch sites



- This map shows us that all launch sites are in close proximity to the coast line.

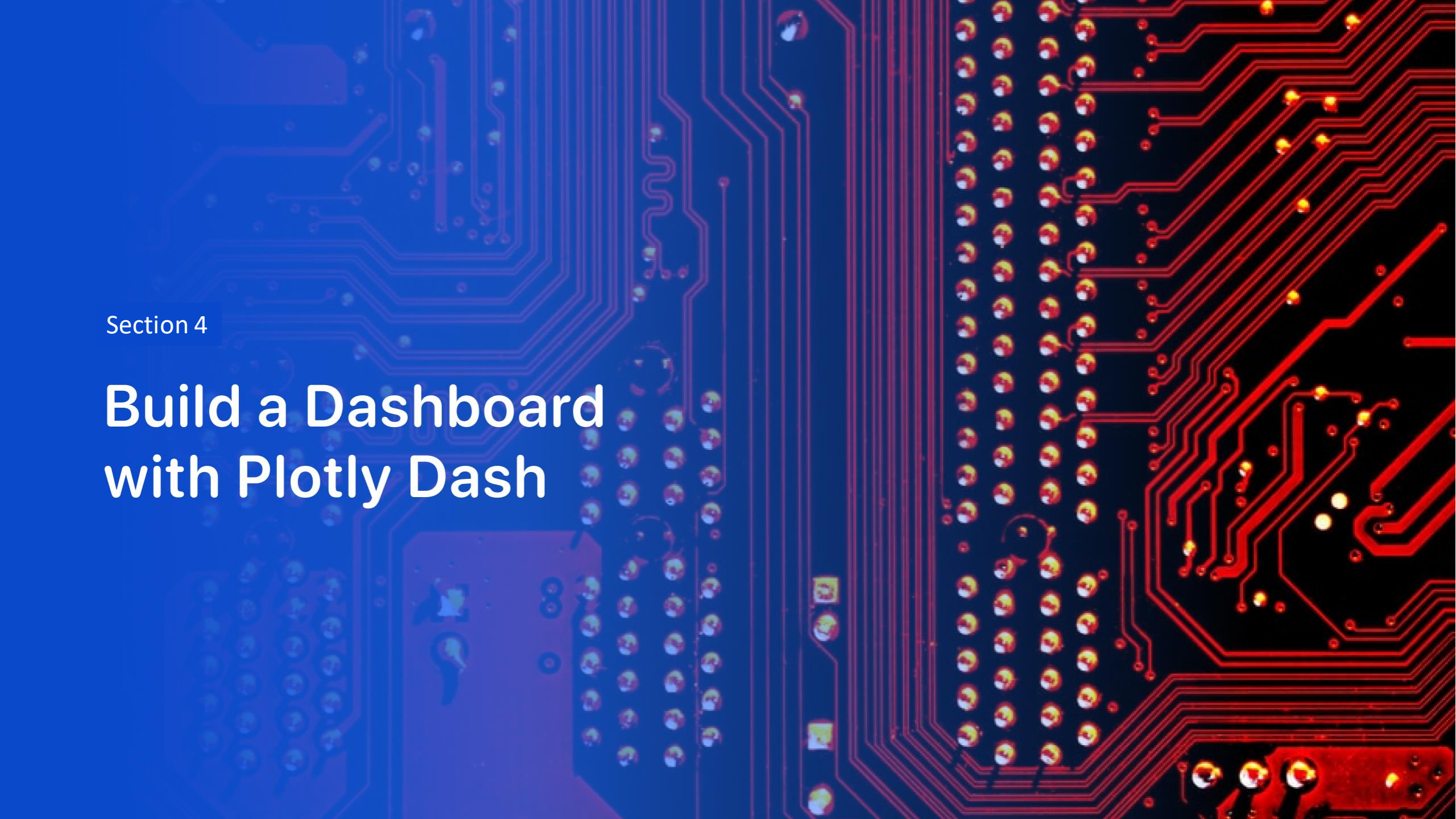
Launch Site Detail



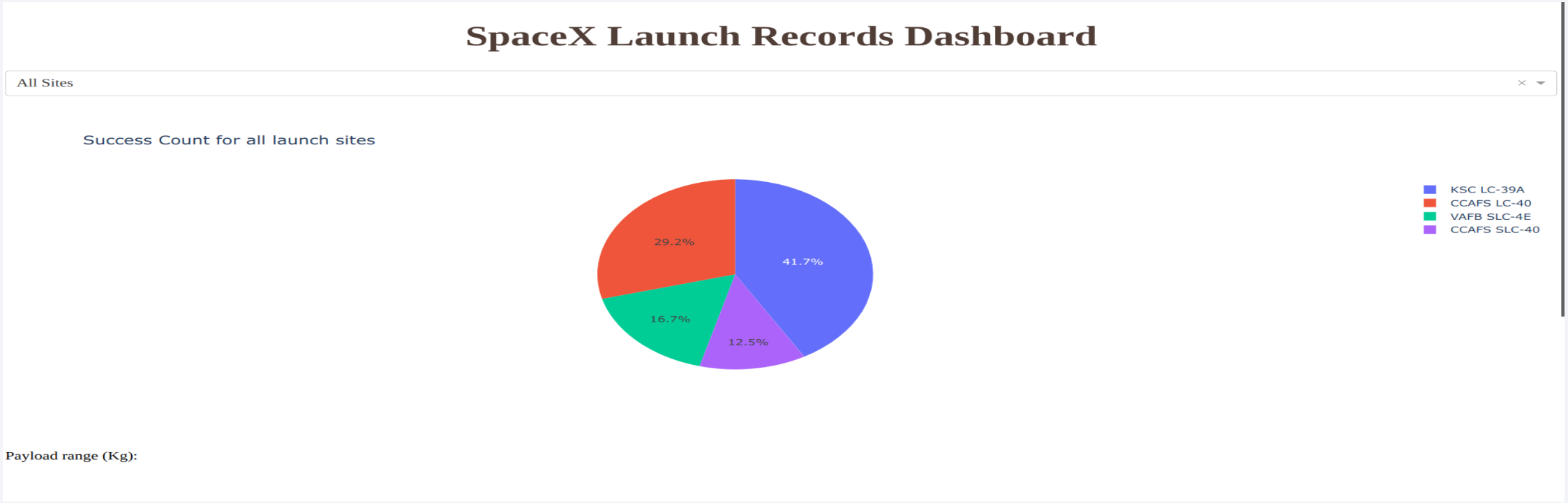
- The green and red markings show the success or failure landings from this site.
- These markings help us to easily identify which launch sites have relatively high success rates.

Launch Site Proximities

- This map shows a launch site that has a railway, highways, and a coastline all in close proximity.


Section 4
Build a Dashboard with Plotly Dash

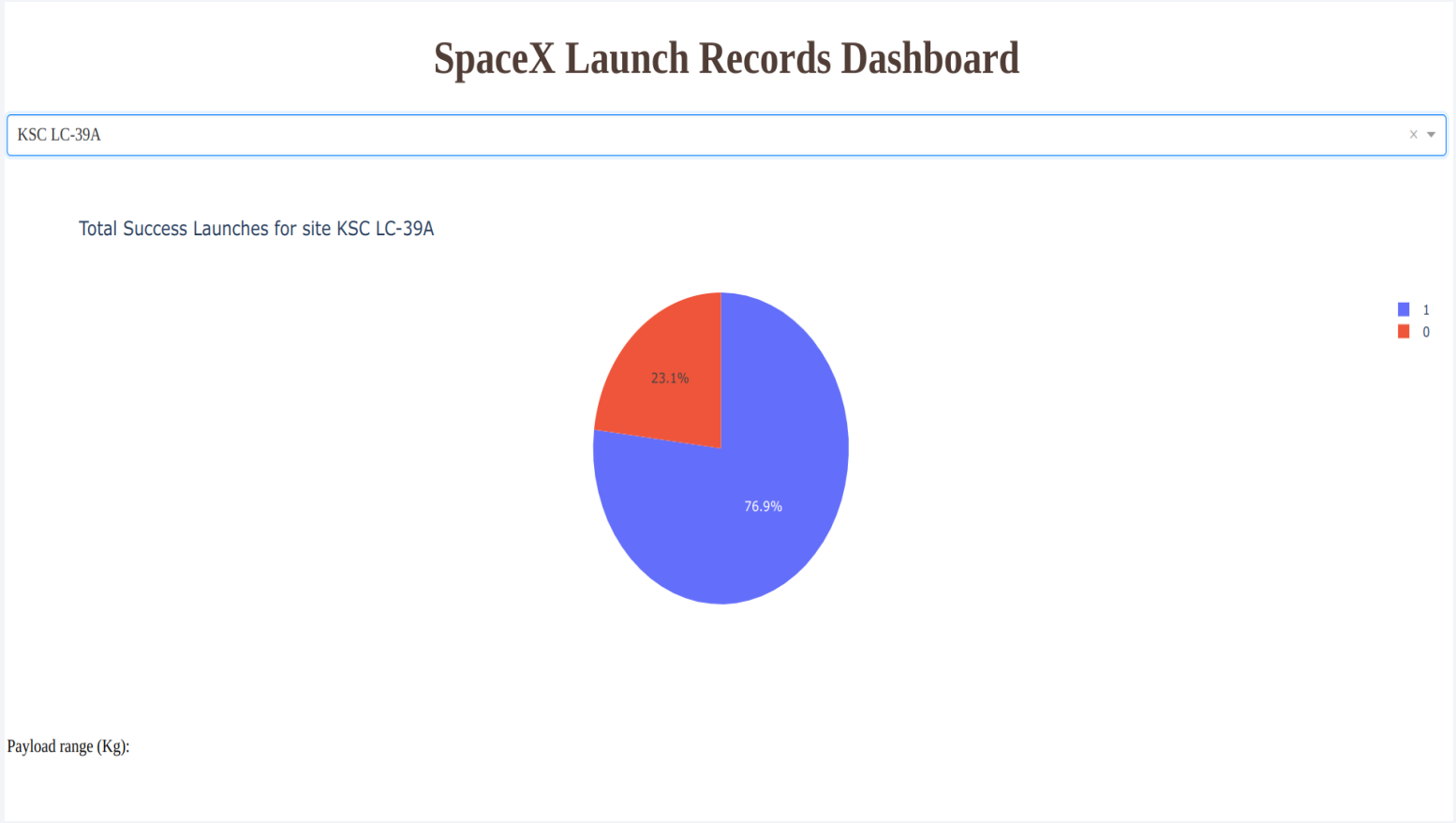
Launch Successes on All Sites



- Success rate of Space X launches for all sites

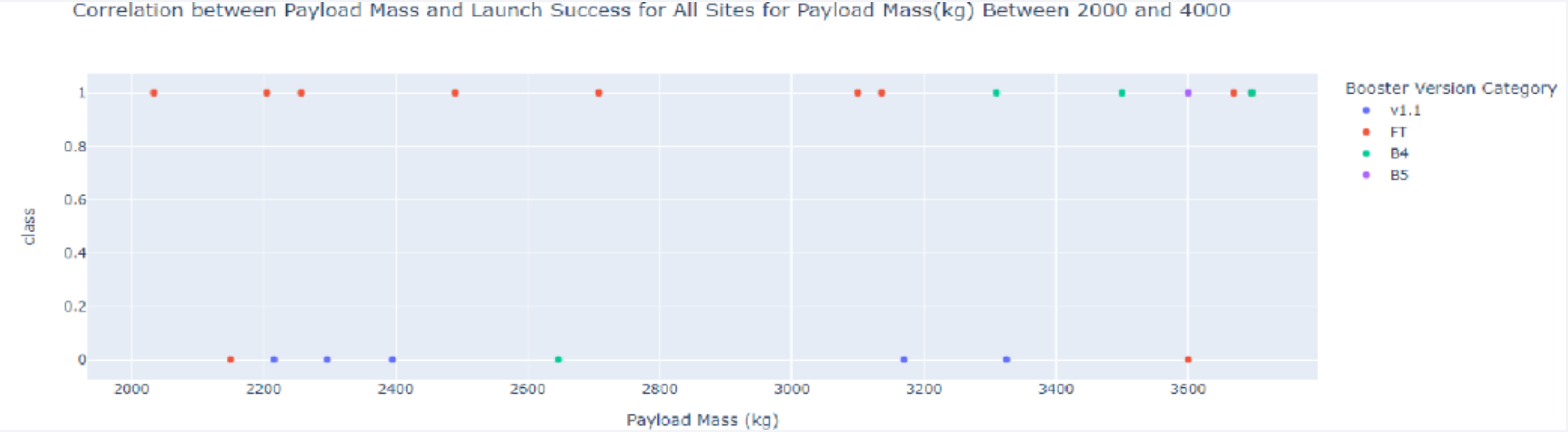
- Site KSC LC-39A has the highest success rate according to this chart

Launch Site with Highest Success Ratio



- Piechart showing the launch site with highest launch success ratio: KSC LC-39A
- This site has a 76.9% success ratio, only 23.1% of launches fail

Payload Mass & Launch Success



- A payload mass between 2000kg and 4000kg proves to be the most successful

Section 5

Predictive Analysis (Classification)

Classification Accuracy

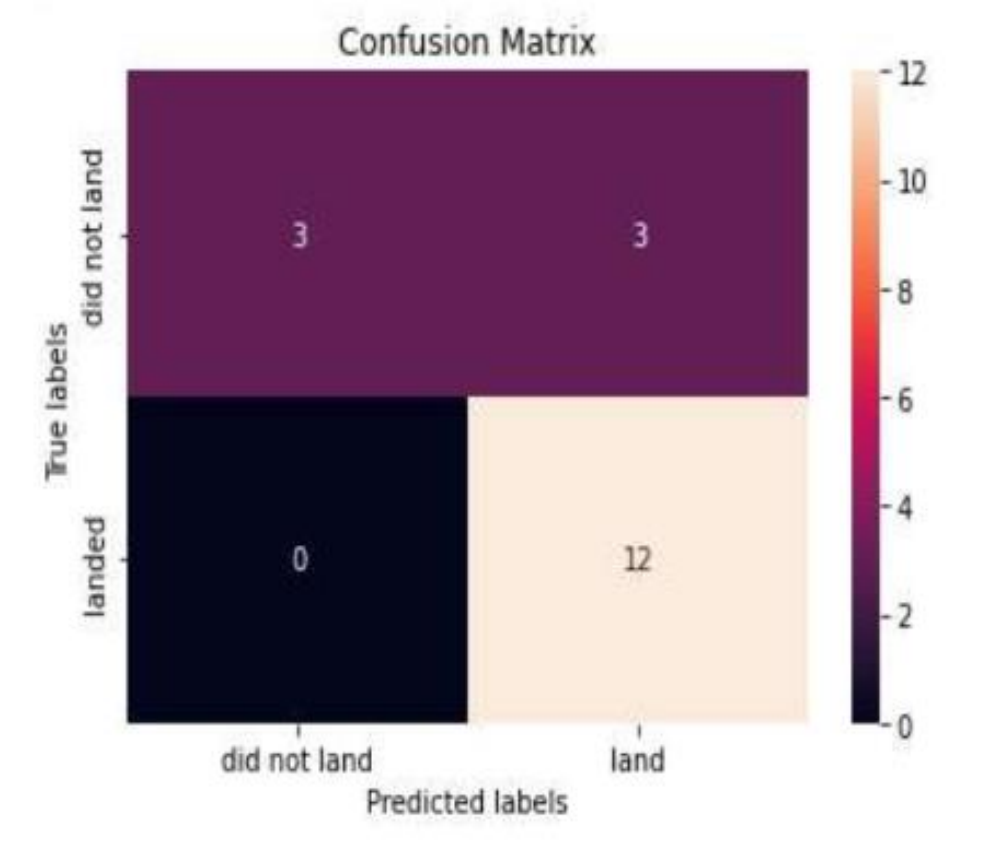
- We compared logistic regression, SVM, decision tree and KNN classification models
- In comparing the different classification models regarding built model accuracy, we conclude that all models have the same outcome: an 83.33% accuracy ratio

```
compare_models = [logreg_cv, svm_cv, tree_cv, knn_cv]

all_scores = [model.score(X_test,Y_test) for model in compare_models]
all_scores
```

```
[0.8333333333333334,
 0.8333333333333334,
 0.8333333333333334,
 0.8333333333333334]
```

Confusion Matrix



- The confusion matrix of the logistic regression model shows that the model only failed to predict 3 landings

Conclusions

- In order to determine the cost of a launch, we tried to predict if the first landing of Space X Falcon 9 launch would be successful.
- Using Space X' information we concluded that these factors highly contribute to the success of a launch:
  - SIte KSC LC-39A has the highest success rate out of all the observed sites
  - From 2013 onwards the success rate increased drastically
  - This increase is also related to the amount of flight numbers
- In order to produce a predictive statement, various models of machine learning were applied.
- The success rate of all employed machine learning models is 83.33%

Thank you!