IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ana Rodrigues
06/08/2024

Outline

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

Executive Summary

This project aims to predict the conditions for a successful landing of the Falcon 9 first stage. For this, the following steps were applied:

Summary of methodologies

- Collecting relevant data using the SpaceX REST API and web scraping techniques
- Data wrangling of requested data
- Exploratory Data Analysis (EDA) with SQL
- Exploratory Data Analysis with Pandas and Matplolib
- Interactive Visual Analytics and dashboard using Folium and Plotly
- Predictive analysis (classification: logistic regression, support vector machines (SVM), decision trees, k-nearest neighbours (knn))

Summary of results

- All four models performed similarly, with accuracy around 80%
- Launching sites are usually near the equator and close to the coast

Introduction

Background

Space exploration is a multi-million dollar industry, and as such, any investment is the result of complex and calculated decisions. An interesting example is the launch of SpaceX Falcon 9 rockets which represents a cost of 62 million dollars, while other providers cost upward of 165 million dollars each. The difference in savings is due to SpaceX ability to reuse the first stage. Knowing this, a competitor could determine if the first stage will land, and so determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

Use case

- What are the key factors influencing a successful landing?
- Evaluate the success rate of landing
- Find the best predictive model

Section 1

Methodology

Methodology

Executive Summary

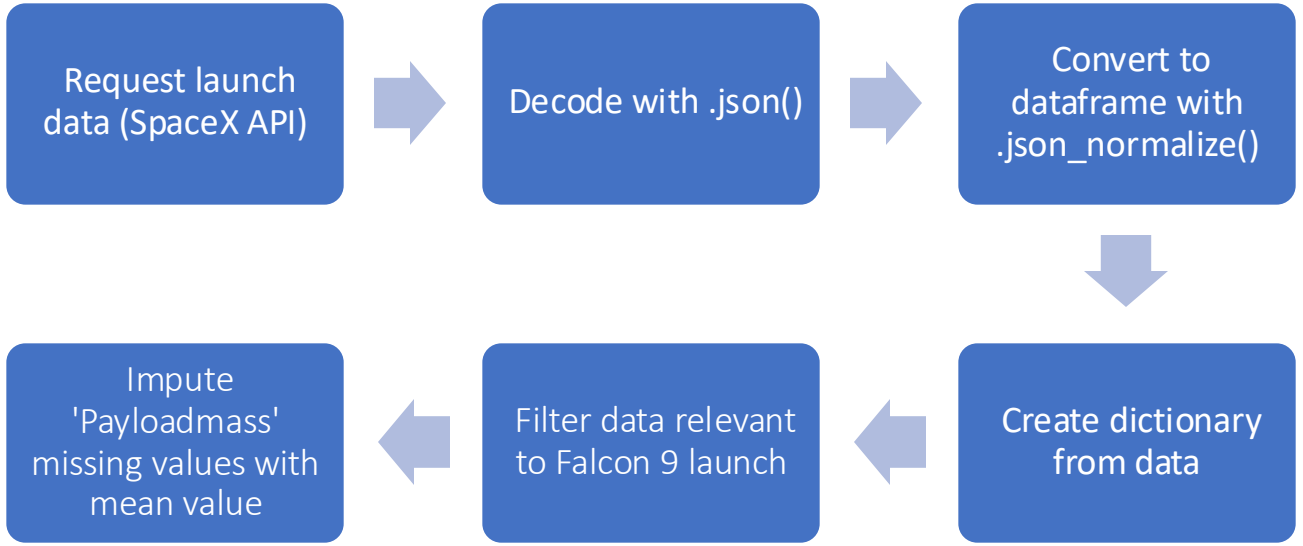
- Data collection methodology:
  - Data was collected with the combined use of SpaceX REST API and web scraping from Wikipedia
- Perform data wrangling
  - Data wrangling involved filtering data, imputation of missing values, and one-hot encoding of categorical variables
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building, tuning, and evaluating classification models (logistic regression, SVM, decision trees, knn)

Data Collection

Datasets were collected through…

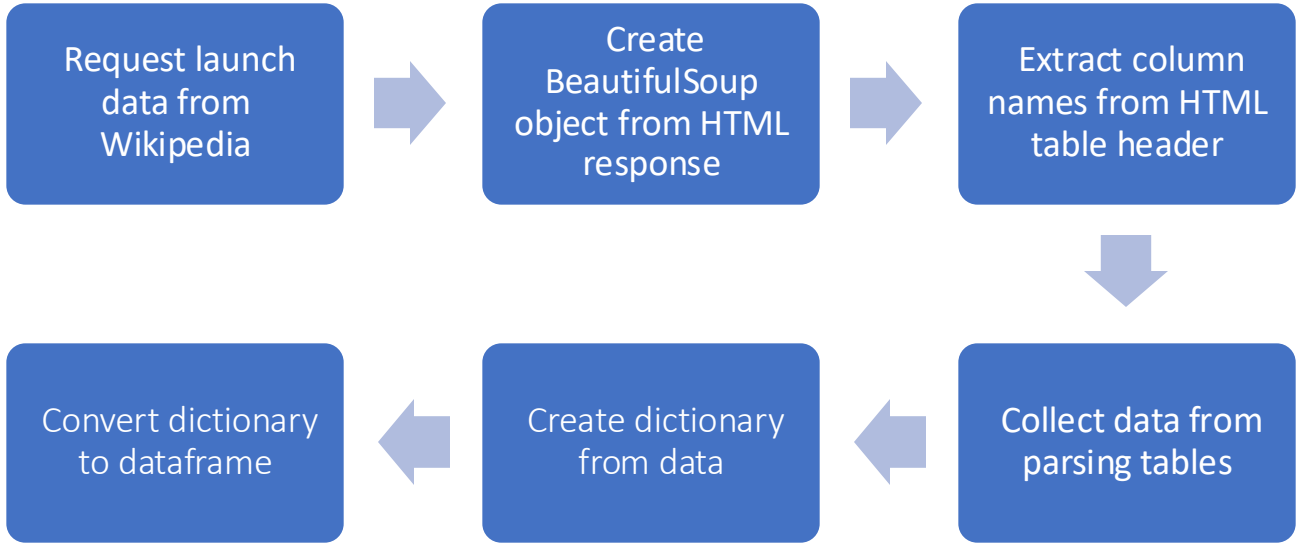
- **the SpaceX REST API:**
  - Requested data from API
  - Decoded information and transformed it into a dataframe
  - Filtered relevant information regarding Falcon 9 and proceeded to handle missing values
- **and web scraping from Wikipedia:**
  - Extracted data from Wikipedia
  - Created BeautifulSoup object with data
  - Converted to dataframe

Data Collection – SpaceX API



https://github.com/APARodrigues/ds-portfolio/blob/d81085b3d5b6230617bf028e2f76eefdcbfa3526/IBM-DS/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping



https://github.com/APARodrigues/ds-portfolio/blob/d81085b3d5b6230617bf028e2f76eefdcbfa3526/IBM-DS/jupyter-labs-spacex-data-collection-api.ipynb

Data Wrangling

Steps to data wrangling…

- **Exploratory data analysis**
  - Determined training labels for the landing outcome (successful/failure)
  - Calculated number of launches from each site, number and occurrence of each orbit, and number and occurrence of mission outcome by orbit
- **Landing outcome**
  - Created landing outcome label from Outcome column
  - Exported information to csv

https://github.com/APARodrigues/ds-portfolio/blob/d81085b3d5b6230617bf028e2f76eefdcbfa3526/IBM-DS/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

Charts:

- Flight Number vs Launch Site
- Payload Mass vs Launch Site
- Success rate of each orbit type
- Flight Number vs Orbit type
- Payload Mass vs Orbit type
- Launch success yearly trend
  - ➢ These allowed a proper analysis of the relationship between variables, so as to discern what features to use in the machine learning models

One-hot encoding & Type formatting

Created dummy variables to categorical columns, to be used in the predictive models, as well as standardized the type of numerical features

https://github.com/APARodrigues/ds-portfolio/blob/b7b5b4aa3a7b9983e5c4b3d6d94a60ee4214e3c4/IBM-DS/edadataviz.ipynb

EDA with SQL

Display:

- Names of the unique launch sites in the space mission
- Five records where launch sites begin with the string 'CCA'
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1

List:

- Date when the first successful landing outcome in ground pad was achieved
- Names of boosters which have success in drone ship and have payload mass between 4000 and 6000
- Total number of successful and failure mission outcomes
- Names of the booster versions which have carried the maximum payload mass
- Records with the month, failure landing_outcomes in drone ship, booster versions, and launch_site for the year 2015

Rank the count of landing outcomes between 2010-06-04 and 2017-03-20, in descending order

https://github.com/APARodrigues/ds-portfolio/blob/802fcf8a9411895a22cf918c6612a65548c5499d/IBM-DS/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Mark all launch sites on a map: *overview of launch site geographic distribution*

- Created a circle at NASA Johnson Space Center's coordinate with a popup label showing its name
- Added red circles for each launch site

Mark the success/failed launches for each site on the map: *easily identify which launch sites have higher success rate*

- Created markers for all launch records
- Added colored markers depending on whether the launch was successful (green) or a failure (red)

Calculate the distances between a launch site to its proximities:

- Added colored lines to show the distance between launch sites and their proximity to points of interest such as railways, cities, coastline, and highways

https://github.com/APARodrigues/ds-portfolio/blob/802fcf8a9411895a22cf918c6612a65548c5499d/IBM-DS/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

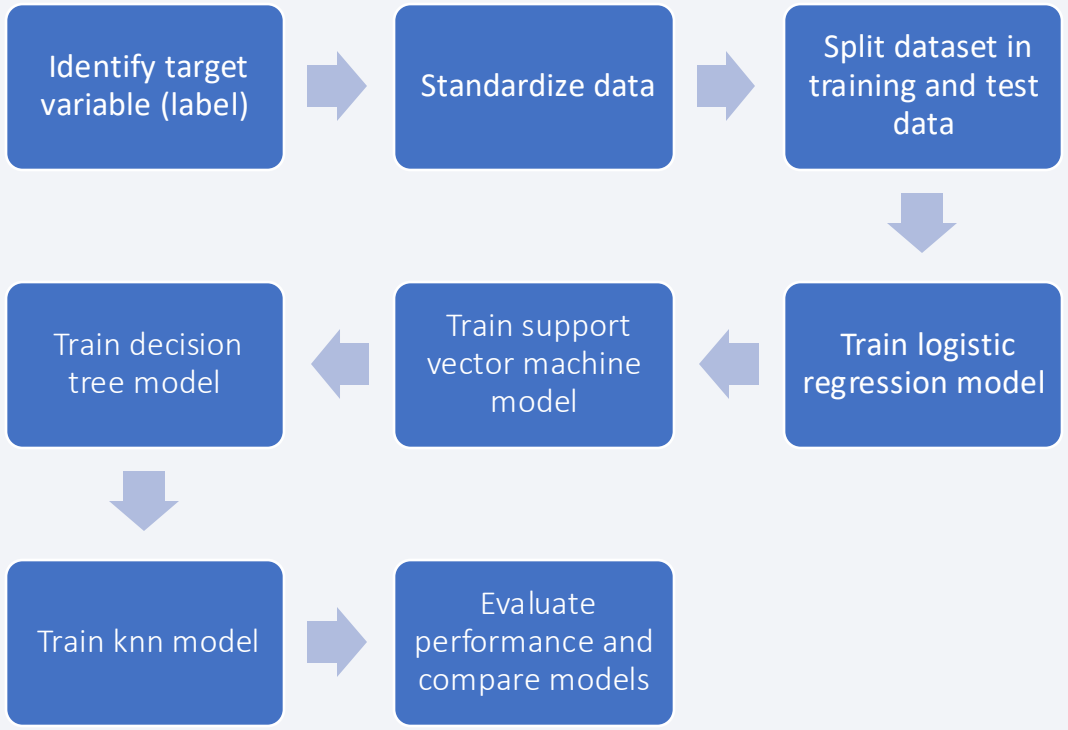
- **Launch Site Drop-down:** choose launch site to analyze
- **Pie chart:** visualizing launch success counts
- **Range Slider to Select Payload:** analyze if a variable payload is correlated to mission outcome
- **Scatter plot with payload vs launch outcome:** visually observe how payload may be correlated with mission outcomes for selected site

https://github.com/APARodrigues/ds-portfolio/blob/802fcf8a9411895a22cf918c6612a65548c5499d/IBM-DS/spacex_dash_app.py

Predictive Analysis (Classification)

Optimize models: *tune models using grid search of hyperparameters*

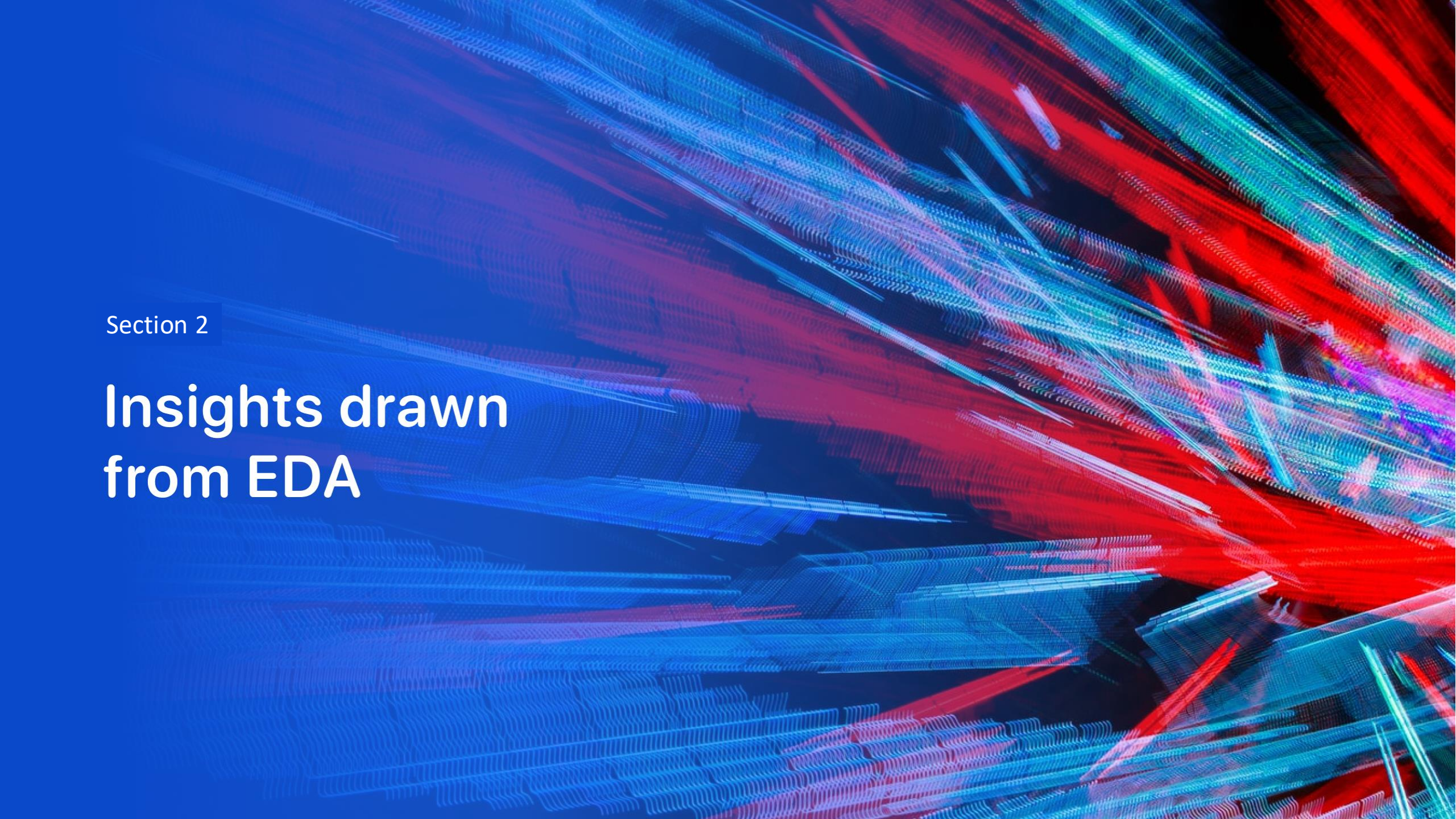
Evaluate and compare models: *use performance metrics, such as accuracy and confusion matrix, to compare models and choose best model*



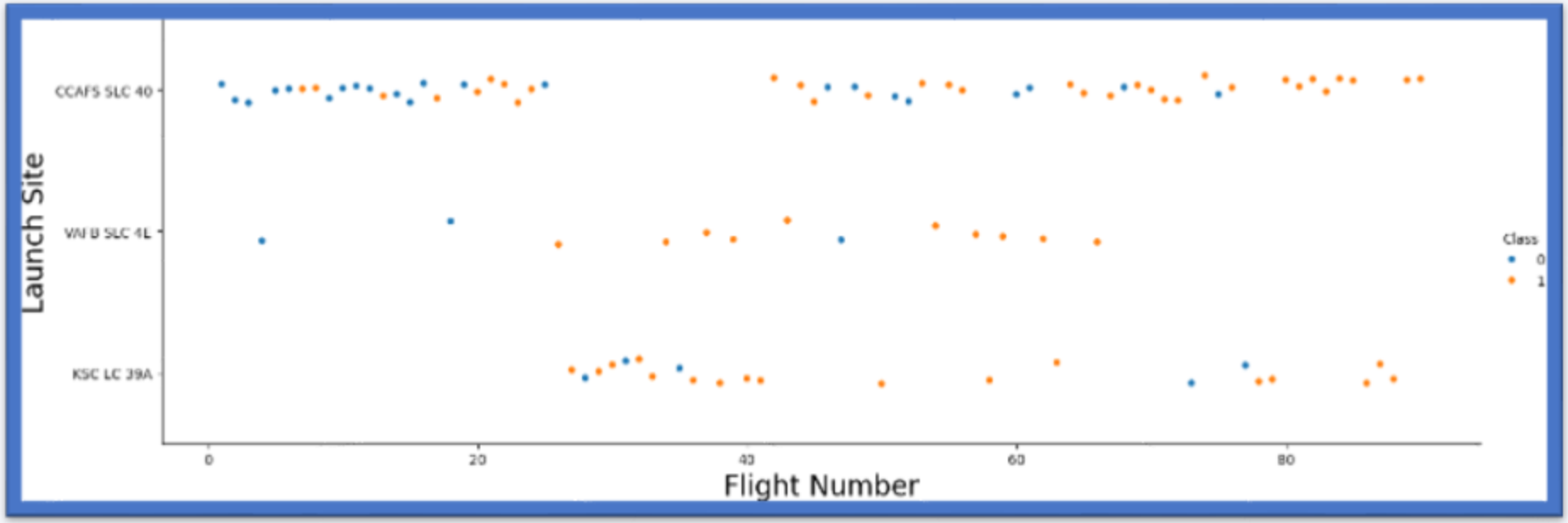
https://github.com/APARodrigues/ds-portfolio/blob/802fcf8a9411895a22cf918c6612a65548c5499d/IBM-DS/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

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

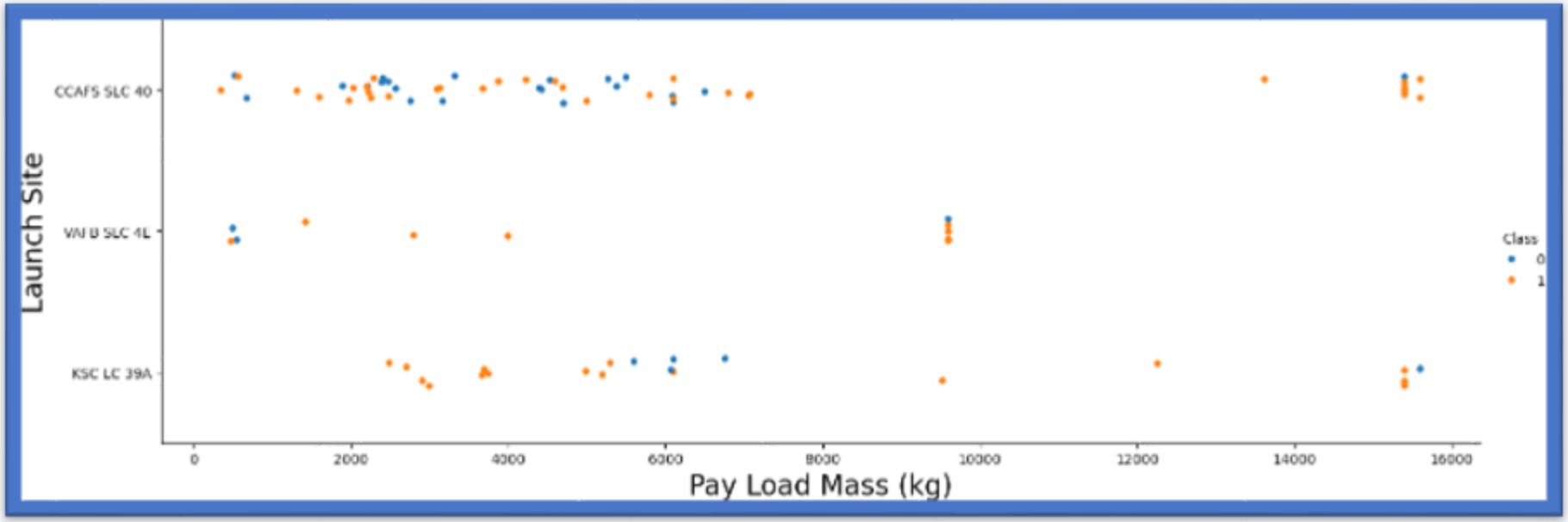

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



This shows an increase in success rate (orange = success, blue = failure) with the number of flights launched at each site.

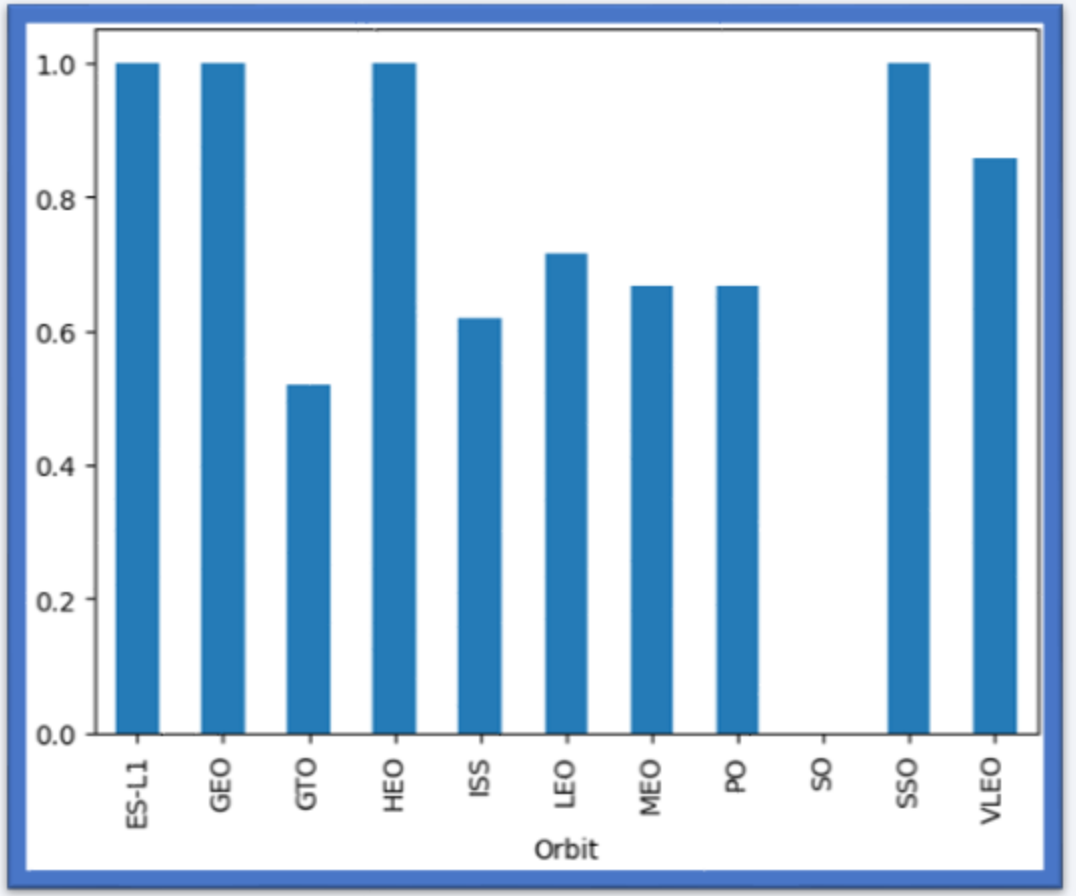
Payload vs. Launch Site



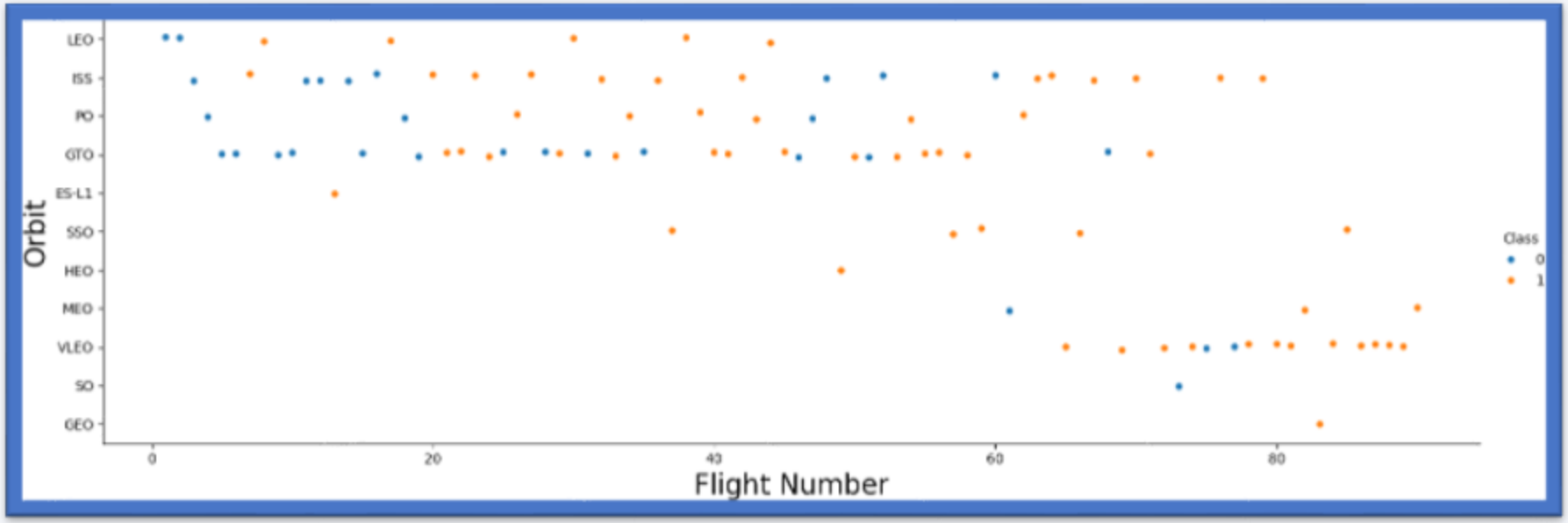
This shows an increase in success rate (orange = success, blue = failure) with higher payloads, although payload mass seems to be most frequent below 7000 kg.

Success Rate vs. Orbit Type

- **Orbits with 100% success rate:** ES-11, GEO, HEO, SSO
- **Orbits with success rate in the range 50-90%:** GTO, ISS, LEO, MEO, PO, VLEO
- **Orbits with 0% success rate:** SO

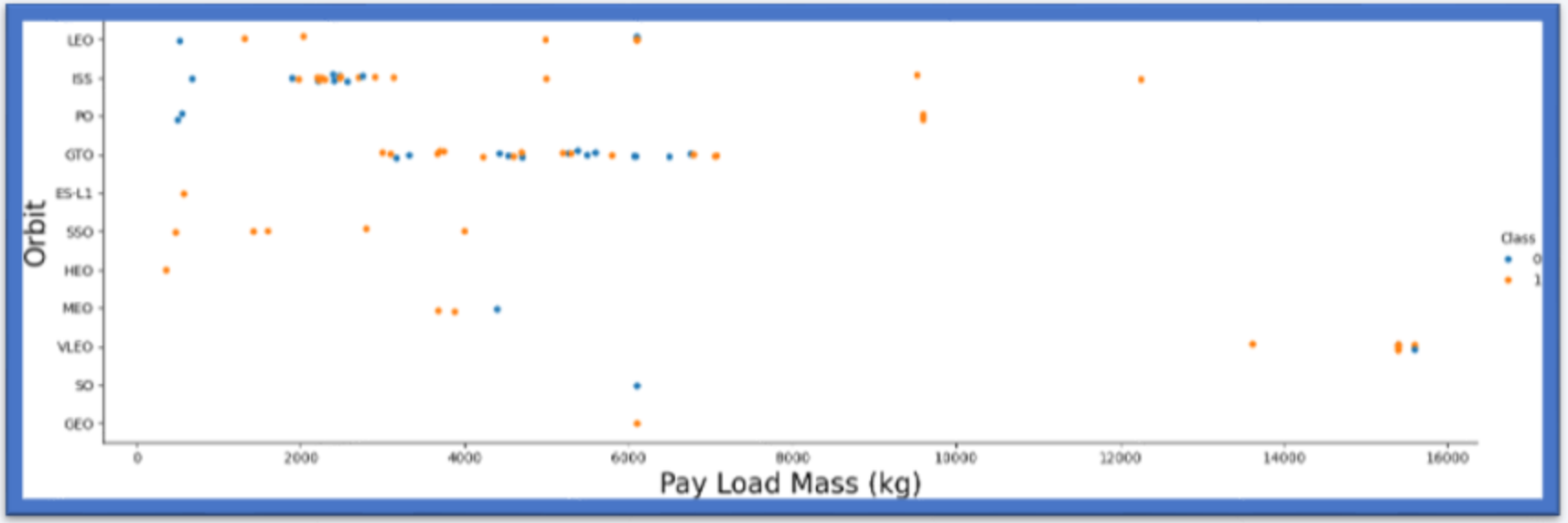


Flight Number vs. Orbit Type



With higher flight numbers we can see a broader choice or orbits, in particular VLEO. This also shows that the preferred orbits are ISS, PO, GTO.

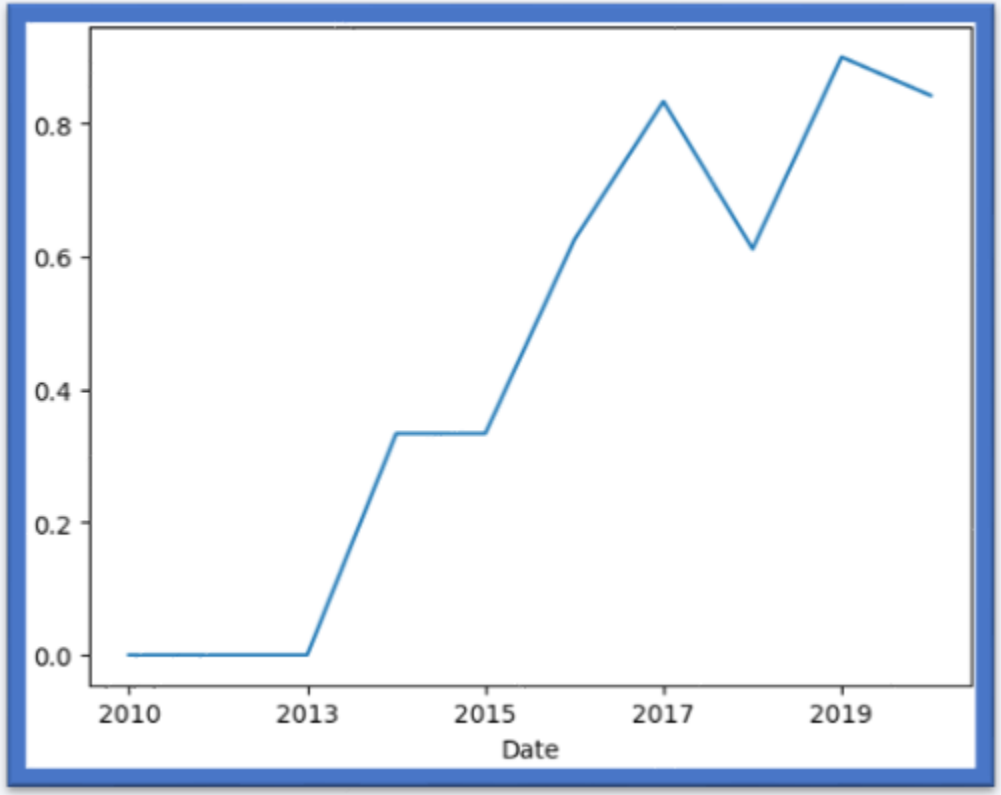
Payload vs. Orbit Type



Higher payloads are used mostly on orbits VLEO, PO, and ISS.

Launch Success Yearly Trend

- Success rate has seen an overall improvement since 2013
- Between 2014 and 2015 the success rate reached a plateau
- In 2017-2018 and 2019-2020 the success rate saw a small decline



All Launch Site Names

Query the DISTINCT launch sites available in SpaceX data

```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE
```

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

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

Launch Site Names Begin with 'CCA'

SELECT the five entries WHERE the launch site names begins with '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

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

```
%sql SELECT sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer like '%NASA%'
```

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

sum(PAYLOAD_MASS__KG_)
107010

Average Payload Mass by F9 v1.1

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

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version like '%F9 v1.1%'
```

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

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

List the date when the first successful landing outcome in ground pad was achieved

```
%sql select min(date) as Date from SPACEXTABLE where mission_outcome like 'Success'
```

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

Date
2010-06-04

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
%sql select booster_version from SPACEXTABLE where (Mission_Outcome like 'Success') AND (payload_mass__kg_ BETWEEN 4000 AND
```

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

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

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
%sql SELECT mission_outcome, count(*) as Count FROM SPACEXTABLE GROUP by mission_outcome ORDER BY mission_outcome
```

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

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

Boosters Carried Maximum Payload

List the names of the booster_versions which have carried the maximum payload mass

```
maxm = %sql select max(payload_mass__kg_) from SPACEXTABLE
maxv = maxm[0][0]
%sql select booster_version from SPACEXTABLE where payload_mass__kg_=(select max(payload_mass__kg_) from SPACEXTABLE)
```

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

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

```
%sql SELECT substr(Date, 6, 2) AS Month, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Failure (drone ship)' AND substr(Date,0,5)='2015' ORDER BY Month;
```

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

Month	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

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

Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

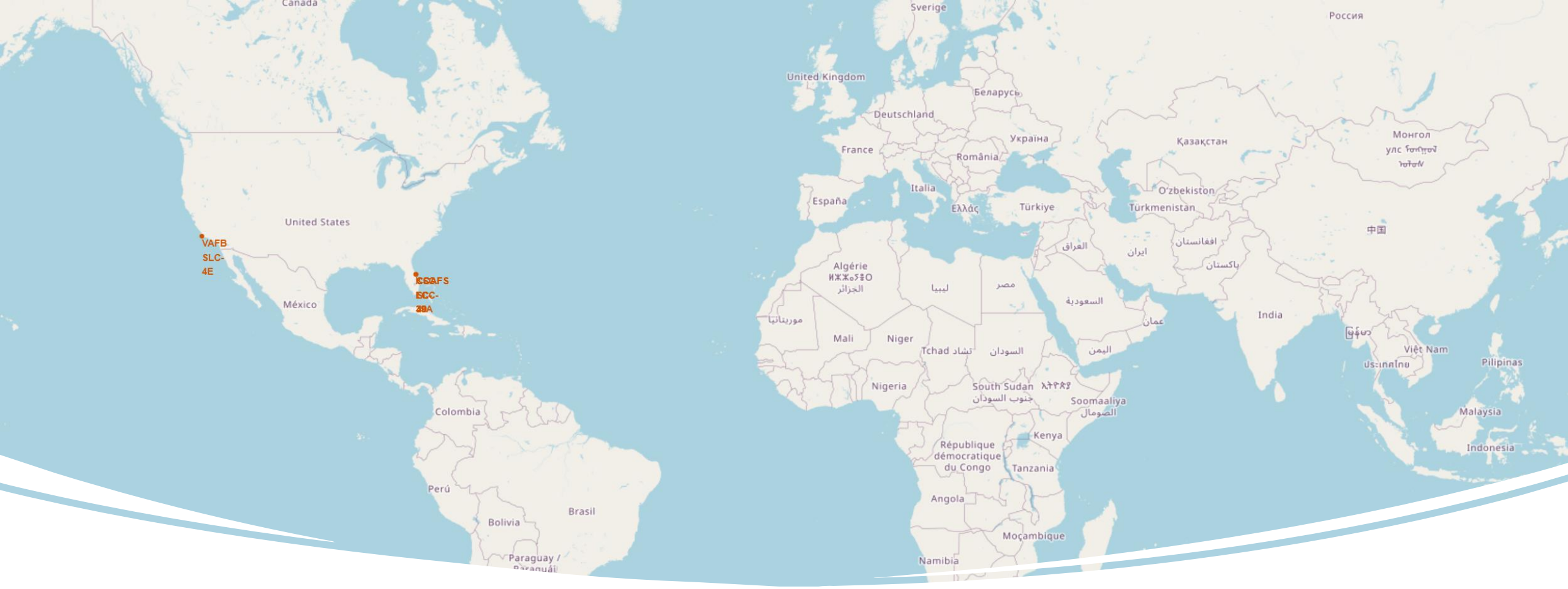
```
%sql select landing_outcome, count(*) as count from SPACEXTABLE where Date >= '2010-06-04' AND Date <= '2017-03-20'  GROUP by landing_outcome ORDER BY count Desc
```

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

Landing_Outcome	count
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

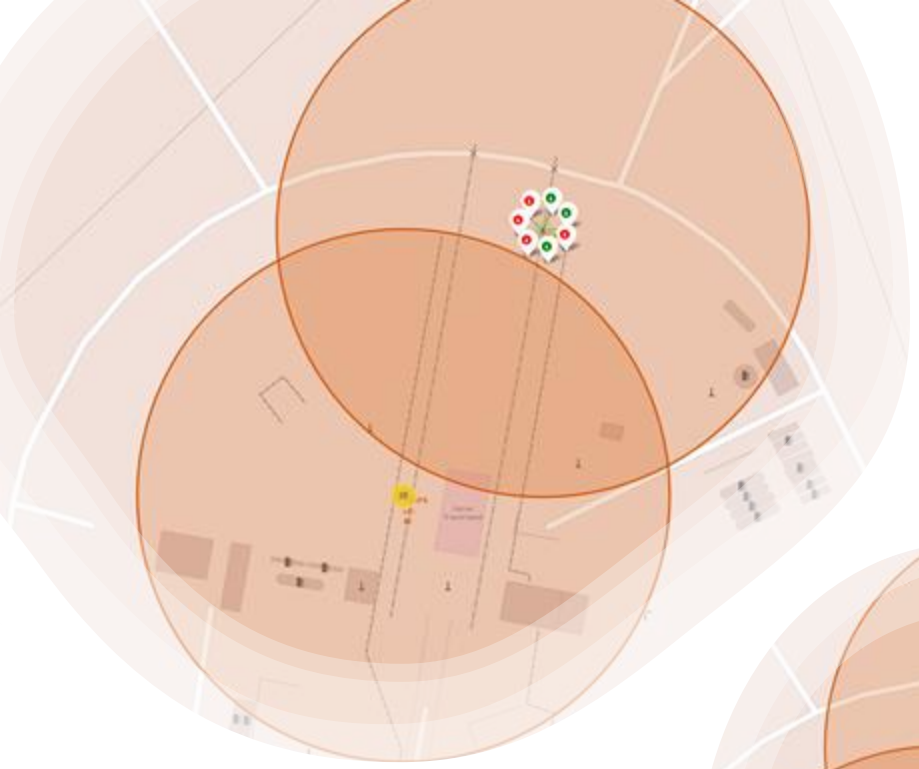
Section 3

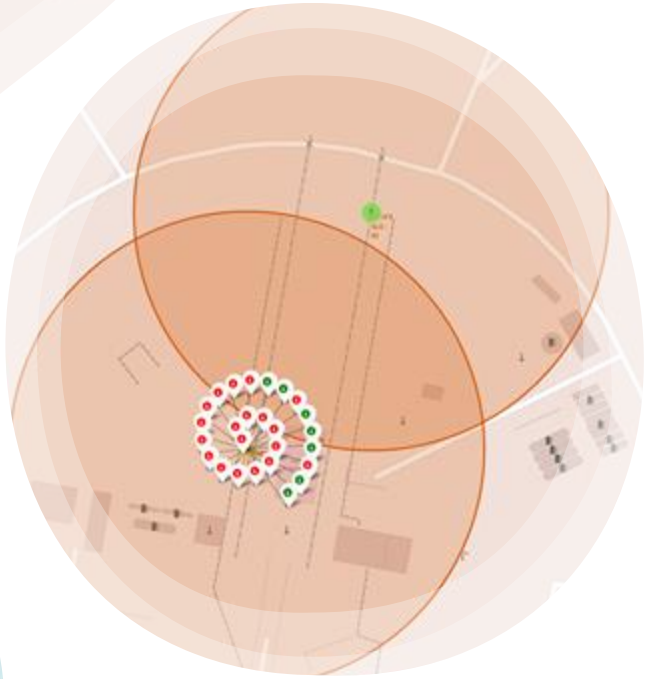
Launch Sites Proximities Analysis



Global launch site overview

- SpaceX uses launch sites in US territory
- Launch sites are near the coastline and near the Equator





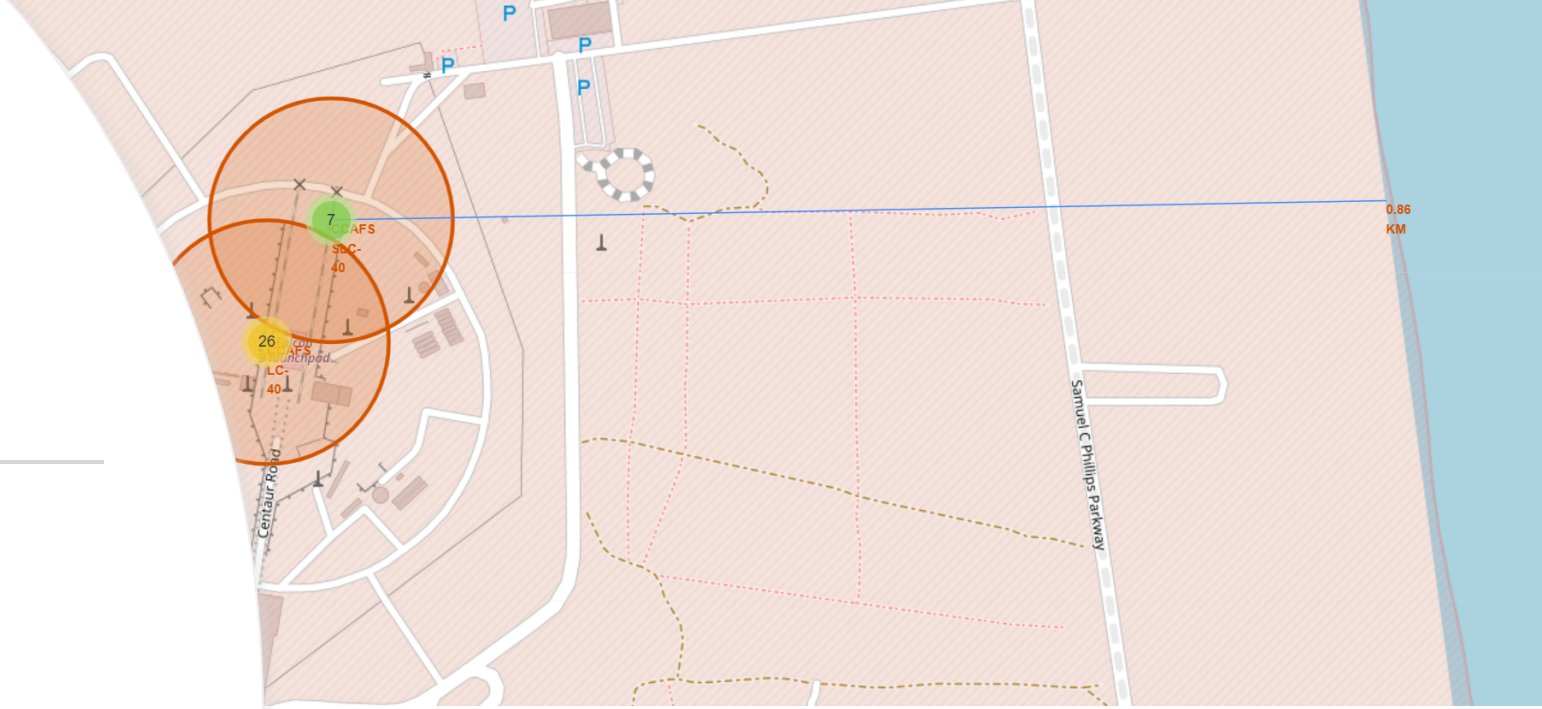
Launch Outcomes

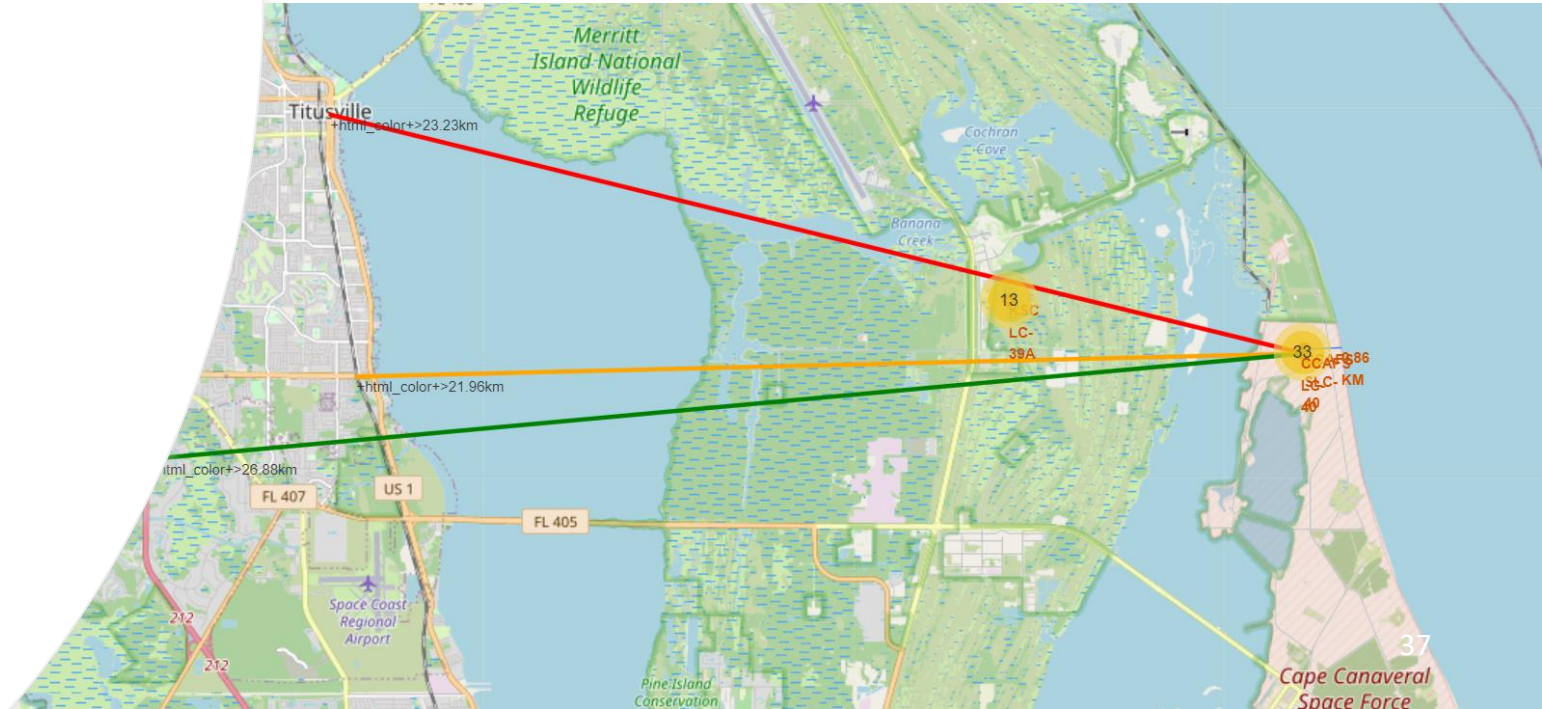
- Green markers = successful outcome
- Red markers = unsuccessful outcome
- KSC LC-39A has a high success rate

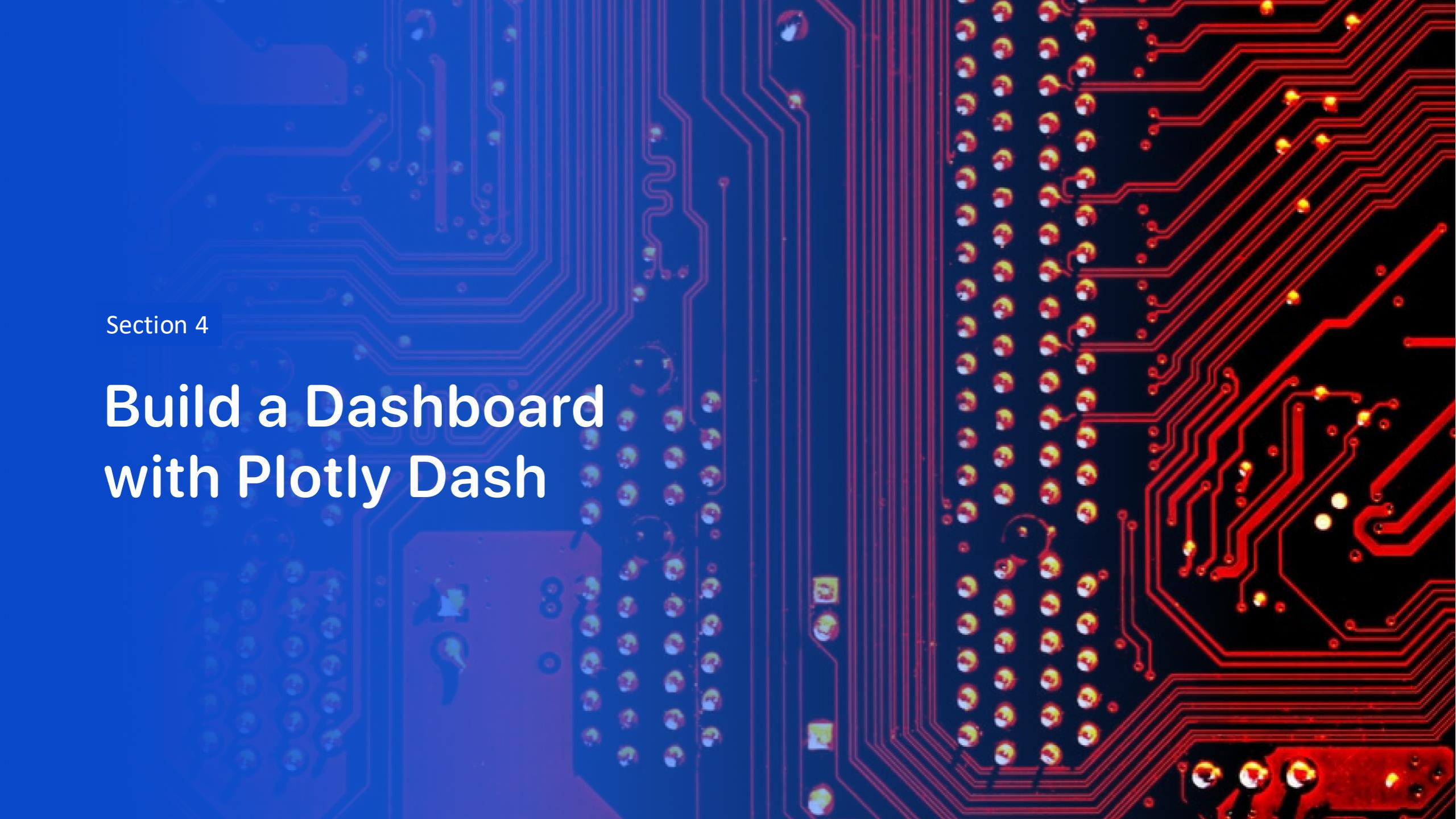
Proximity to points of interest

Launch sites

- Are not close to railways or highways
- Are at a certain distance from cities
- Are close to the coast

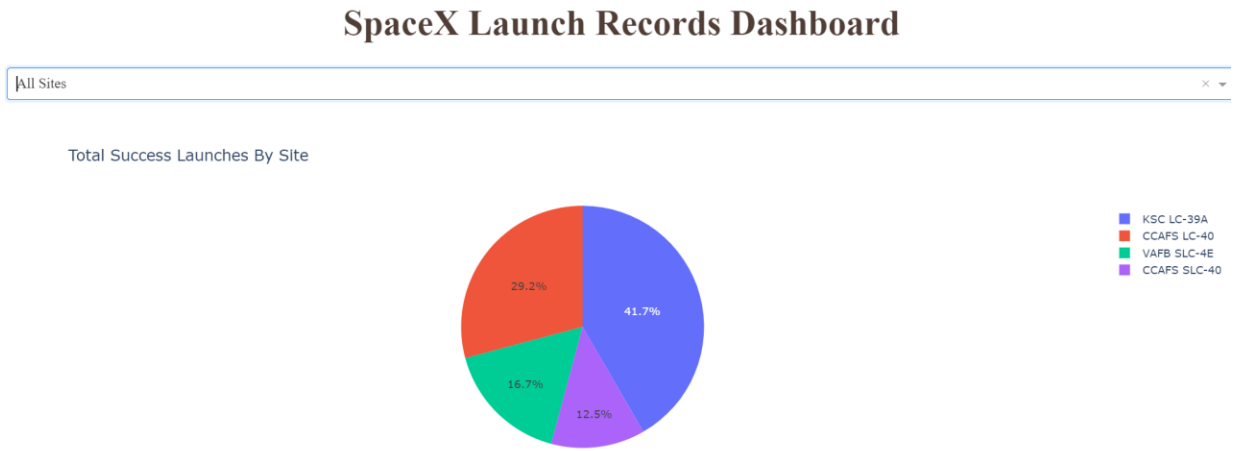





Section 4
Build a Dashboard with Plotly Dash

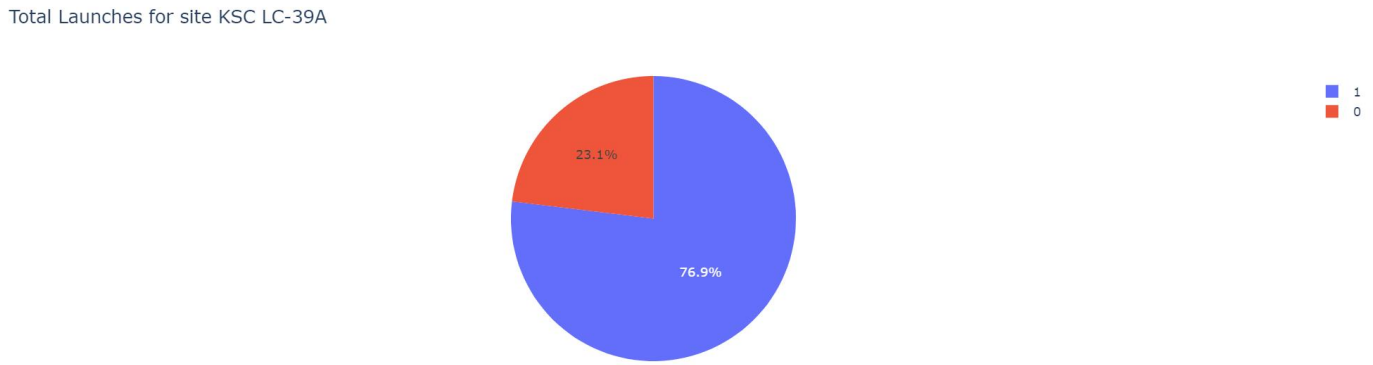
Total success launches by site

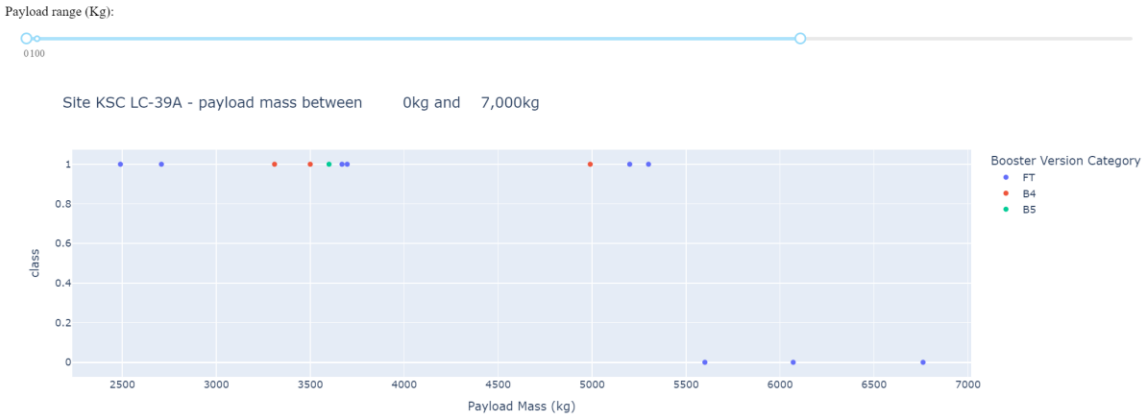
- KSC LC-39A has the highest launch success rate with 41.7%
- CCAFS SLC-40 has the lowest launch success rate with 12.5%

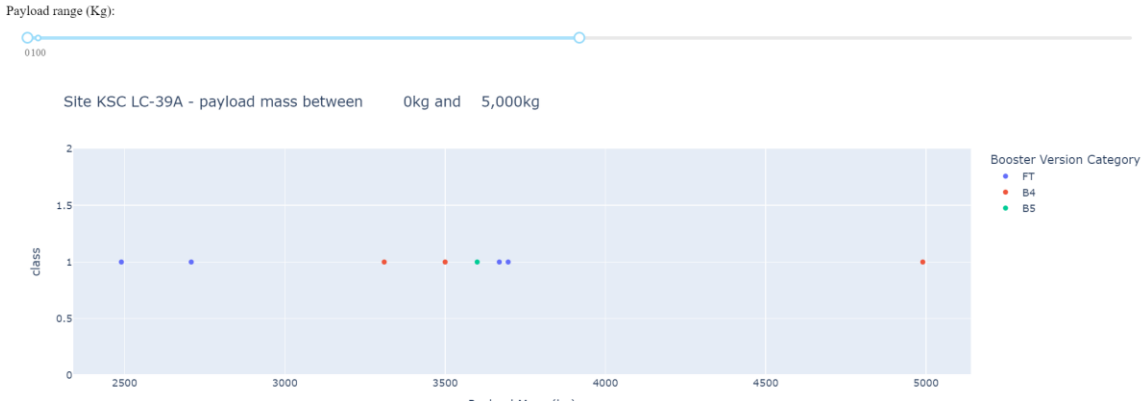


KSC LC-39A launches

- Looking at the launch site with highest success rate, it shows that 76.9% of its total launches was successful
- Only 23.1% of total launches at KSC LC-39 was unsuccessful







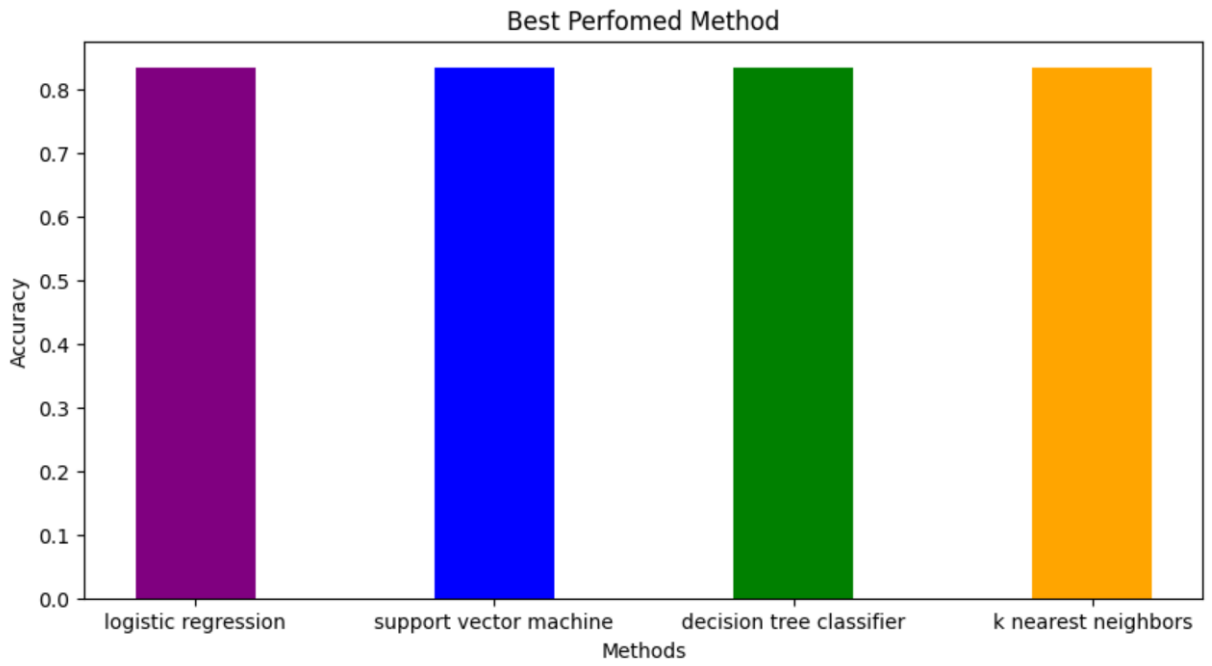
Launch outcome vs Payload mass

- Class 0 = failure; class 1 = success
- Payloads below 5500kg have a higher chance of success

Section 5

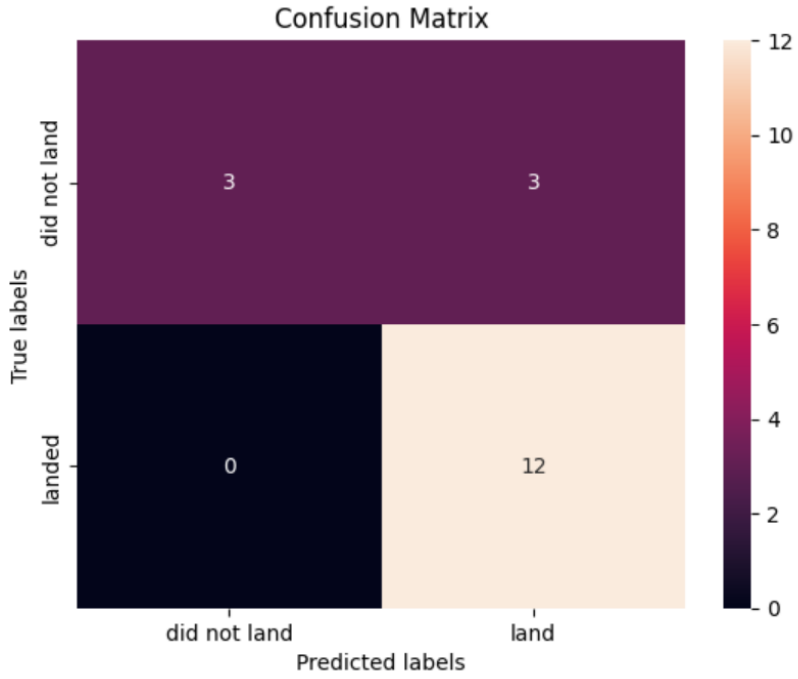
Predictive Analysis (Classification)

Classification Accuracy



- Every model performed essentially at the same level of accuracy (83.3%)
- This similarity is most likely due to the dataset size being too small
- It is useful to look to other metrics to choose the best model

Confusion Matrix



- Considering the accuracy of the models, the confusion matrix is essentially the same, with only 3 missed predictions

Conclusions

- Considering the main goals of this project:
  - What are the key factors influencing a successful landing?
  - Evaluate the success rate of landing
  - Find the best predictive model

- Some factors are relevant to the successful outcome of the stage landing: launch site location, payload mass, booster version, orbit type

- Success rate of landing has seen an overall improvement since 2013, albeit a couple of declines; the highest success rate reaches almost 42% in one of the launch sites (KSC LC-39)

- All models performed similarly with this dataset, with an accuracy of around 83%

Notes:

- In order to improve model performance and quality of results the dataset needs to be extended, other models could be used, some additional feature engineering could be interesting

Thank you!