IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Harshit Verma
April 20, 2024

Outline

Executive Summary

- In this project, a complete set of different data science methodologies were employed to study the historical launch data of SpaceX using Falcon9 rockets. Data were collected through public SpaceX API and web scraping from the SpaceX Wikipedia page. After preprocessing the collected data, exploratory data analysis was employed using magic SQL commands and various visualization libraries in Python to understand the data better. After gathering relevant columns to be used as features for predictive model development, various classification models were developed and tested for optimal results based on their respective accuracy scores.
- Four machine learning models were produced: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors. All produced similar results with an accuracy rate of about 83.33%. All models over-predicted successful landings. More data is needed for better model determination and accuracy.

Introduction

- SpaceX advertises Falcon 9 rocket launches on its website at a cost of 62 million dollars; other providers cost upward of 165 million dollars each, and 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. This information can be used if SpaceY wants to bid against SpaceX for a rocket launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data Collection through SpaceX public API and web scraping of the SpaceX Wikipedia page
- Perform data wrangling
  - Classifying true landings as successful and unsuccessful otherwise
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Tuned models using GridSearchCV

Data Collection

- The data collection process involved a combination of API requests from Space X's public API and web scraping data from a table in Space X’s Wikipedia entry.
- The following slide will show the flowchart of data collection from API, and the one after will show the flowchart of data collection from web scraping.
- **SpaceX API Data Columns:**

  FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

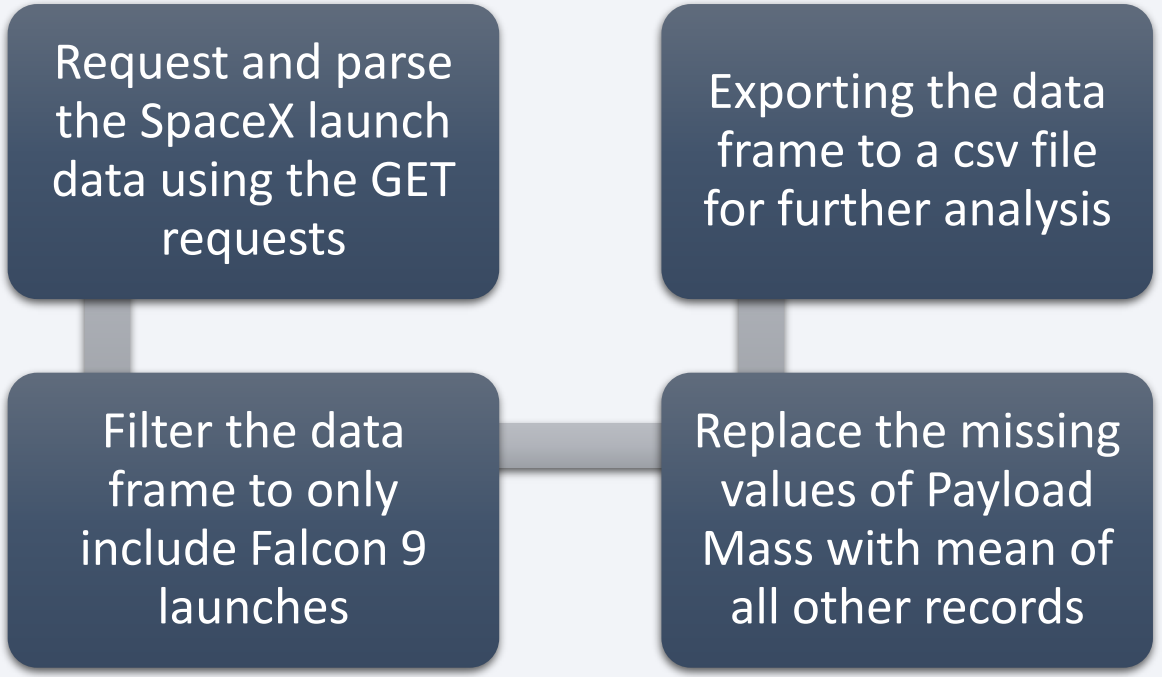
- **Wikipedia Webscrape Data Columns:**

  Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API

GitHub URL:

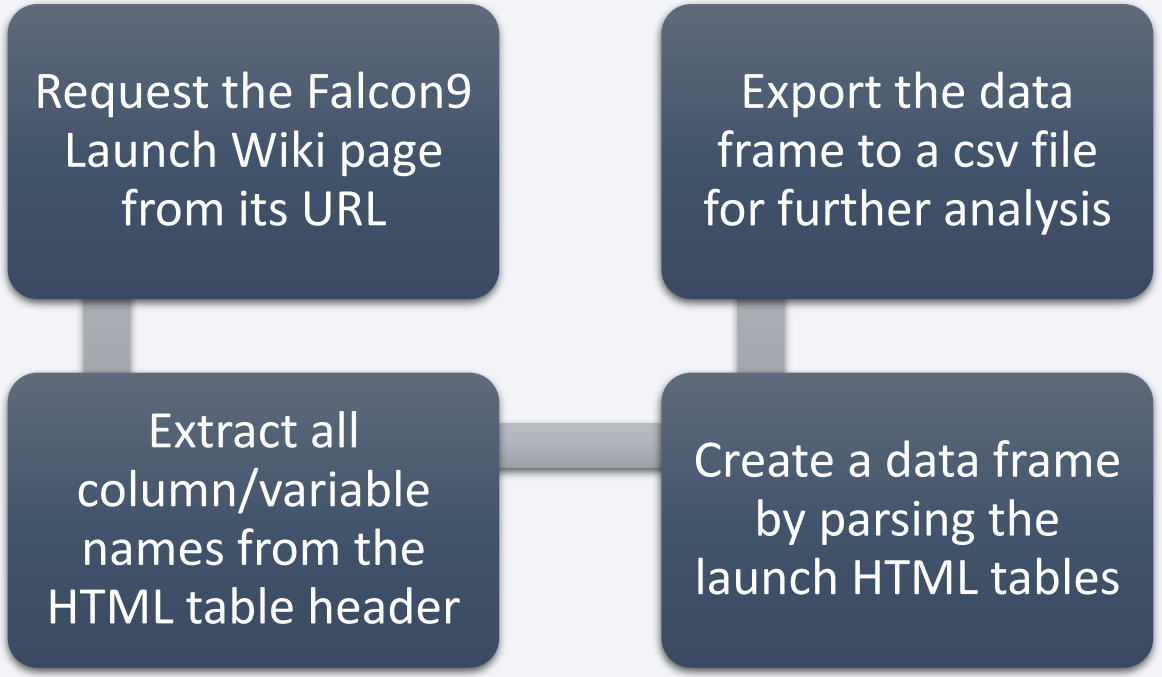
https://github.com/Harshit-Optimus0089/IBM_Data_Science_Capstone/blob/b6c0c33f19176419bbc743a953fa5f3e96eaa048/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

GitHub URL:

https://github.com/Harshit-Optimus0089/IBM_Data_Science_Capstone/blob/b6c0c33f19176419bbc743a953fa5f3e96eaa048/jupyter-labs-webscraping.ipynb



Data Wrangling

- Create a training label with landing outcomes where successful = 1 & failure = 0.
- The Outcome column has two components: ‘Mission Outcome’ and ‘Landing Location’
- New training label column ‘class’ with a value of 1 if ‘Mission Outcome’ is True and 0 otherwise.
- Value Mapping:

  True ASDS, True RTLS, & True Ocean – set to -> 1

  None None, False ASDS, None ASDS, False Ocean, False RTLS – set to -> 0

- GitHub URL:

  https://github.com/Harshit-Optimus0089/IBM_Data_Science_Capstone/blob/b6c0c33f19176419bbc743a953fa5f3e96eaa048/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Exploratory Data Analysis was performed on variables such as flight number, payload mass, launch site, orbit, class, and year.
- Plots Used:

  Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit vs. Success Rate, Flight Number vs. Orbit, Payload vs Orbit, and Success Yearly Trend

  Scatter plots, line charts, and bar plots were used to compare relationships between variables to decide if a relationship exists so that they could be used in training the machine learning model.
- GitHub URL:

  https://github.com/Harshit-Optimus0089/IBM_Data_Science_Capstone/blob/b6c0c33f19176419bbc743a953fa5f3e96eaa048/edadataviz.ipynb

EDA with SQL

- Loaded data set into IBM DB2 Database using sqlite3 library.
- Queried using SQL Python integration.
- Queries were made to get a better understanding of the dataset.
- Queried information about launch site names, mission outcomes, various payload sizes of customers and booster versions, and landing outcomes.
- GitHub URL:

  https://github.com/Harshit-Optimus0089/IBM_Data_Science_Capstone/blob/b6c0c33f19176419bbc743a953fa5f3e96eaa048/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Folium maps mark launch sites, successful and unsuccessful landings, and a proximity example to crucial locations, such as railway, highway, coast, city, etc.
- This allows us to understand why launch sites may be located where they are. Also visualizes successful landings relative to location.
- GitHub URL:

  https://github.com/Harshit-Optimus0089/IBM_Data_Science_Capstone/blob/b6c0c33f19176419bbc743a953fa5f3e96eaa048/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

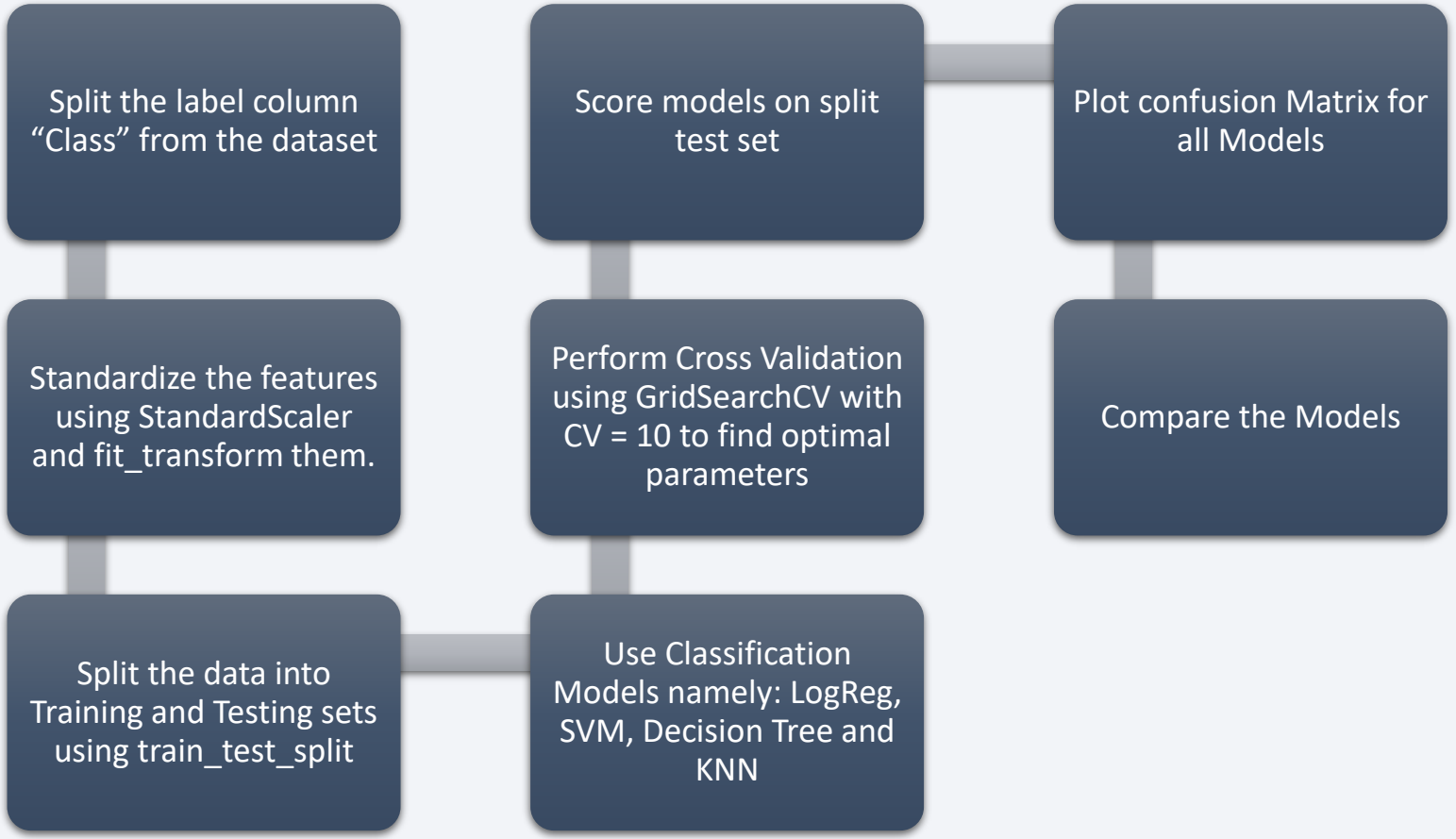
- Dashboard includes a pie chart and a scatter plot.
- A pie chart can be selected to show the distribution of successful landings across all launch sites and can be selected to show individual launch site success rates.
- Scatter plot takes two inputs: All sites or individual sites and payload mass on a slider between 0 and 10000 kg.
- The pie chart is used to visualize the launch site success rate.
- The scatter plot can help us see how success varies across launch sites, payload mass, and booster version categories.
- GitHub URL:

  https://github.com/Harshit-Optimus0089/IBM_Data_Science_Capstone/blob/d0f11f6cd0fa3c388c48797c4cb9e443890ff712/spacex_dash_app.py

Predictive Analysis (Classification)

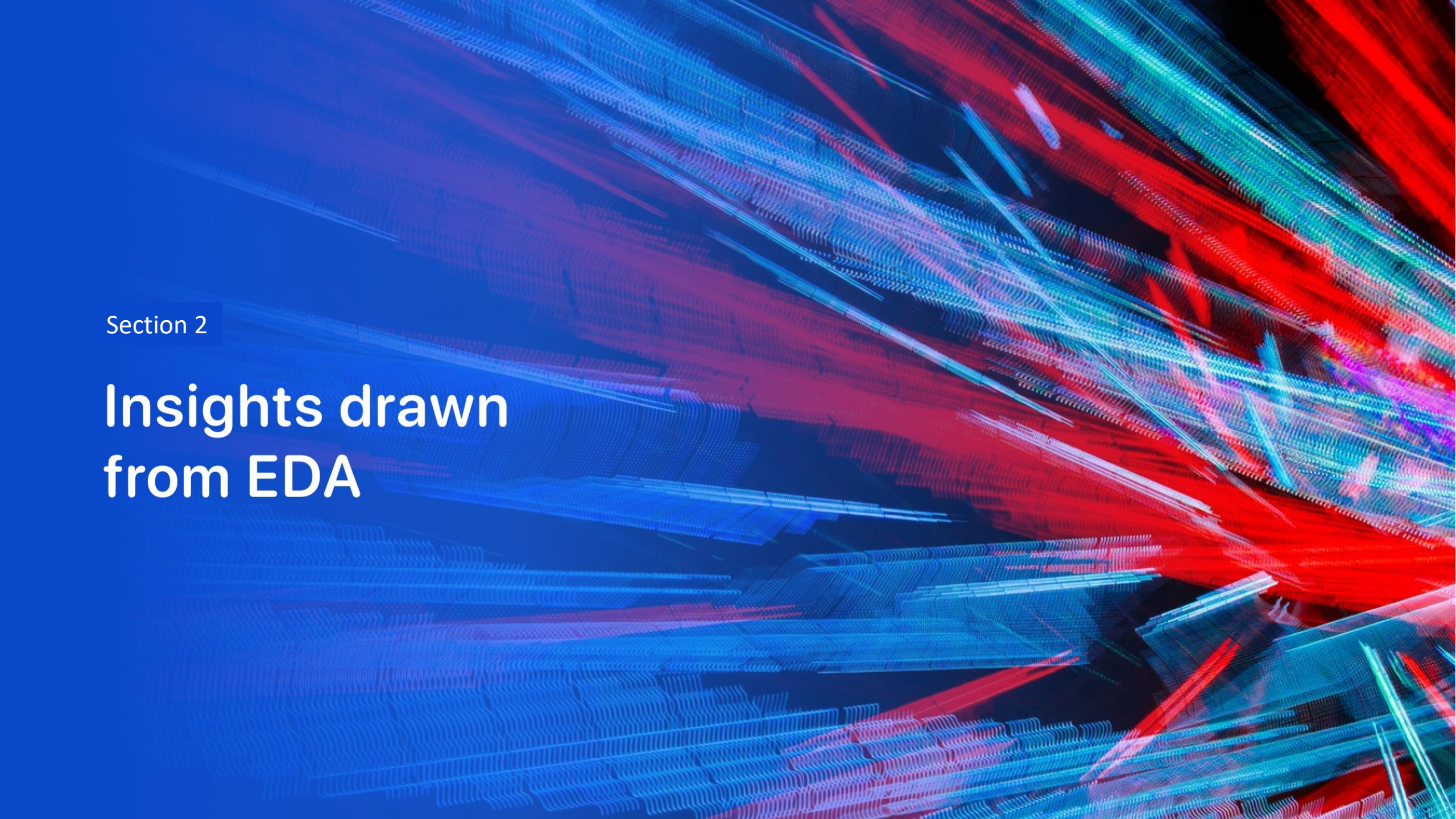
- GitHub URL:

  https://github.com/Harshit-Optimus0089/IBM_Data_Science_Capstone/blob/d0f11f6cd0fa3c388c48797c4cb9e443890ff712/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

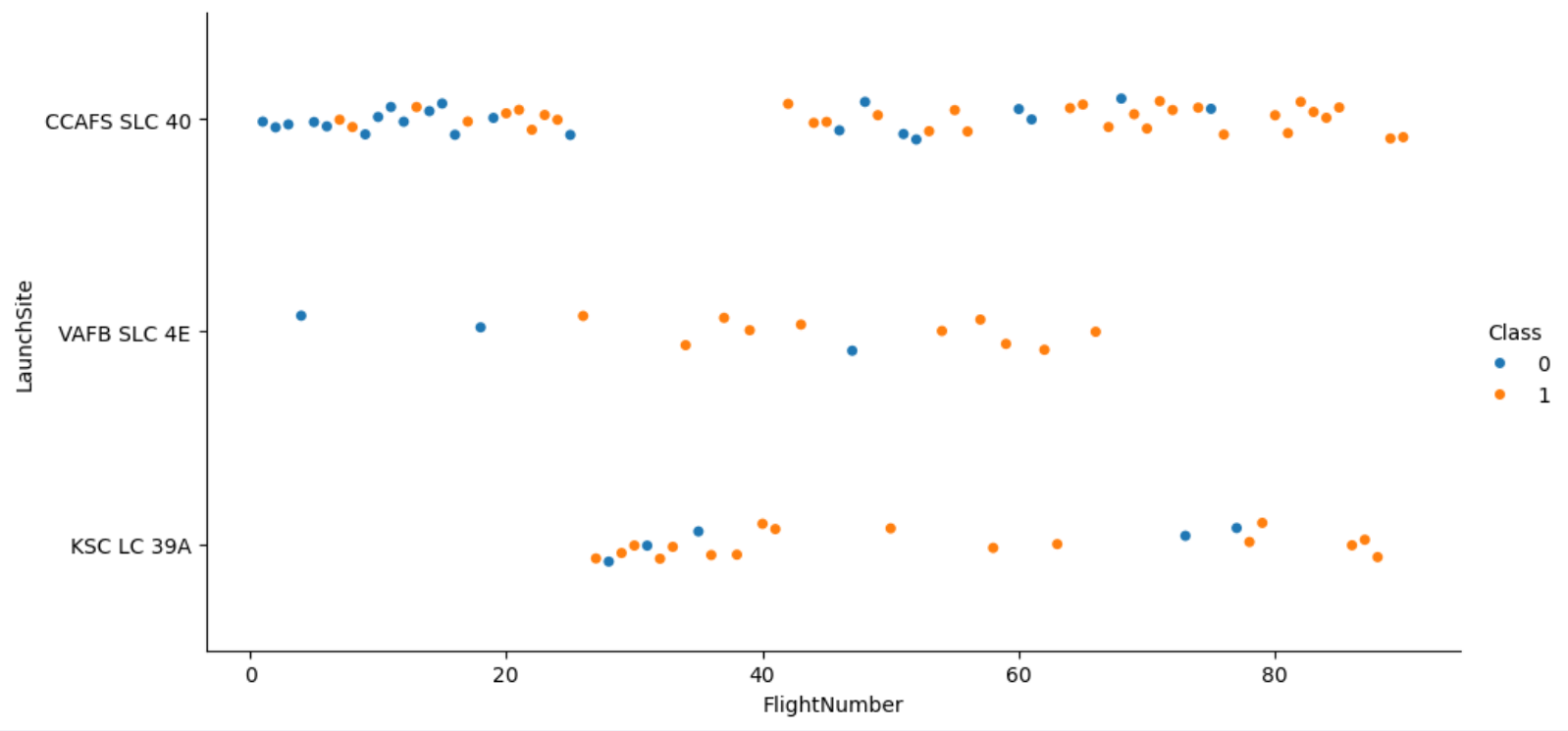


Results

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


Section 2
Insights drawn from EDA

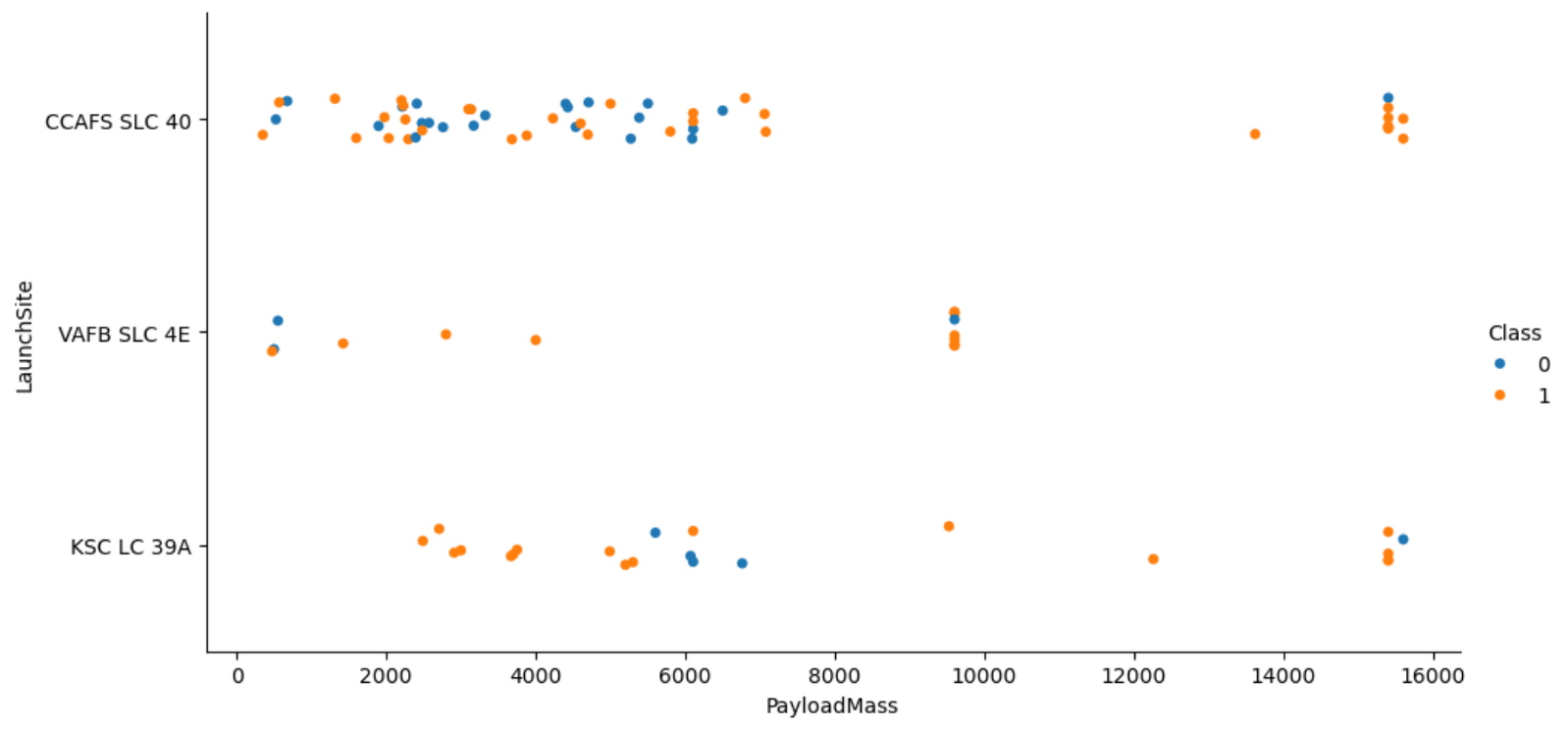
Flight Number vs. Launch Site



The orange indicates a successful launch and the blue indicates a failed launch

- Graphic suggests an increase in success rate over time.
- The success rate significantly increased after 20 flights, which shows the technology readiness threshold.
- CCAFS appears to be the main launch site, with the most volume.

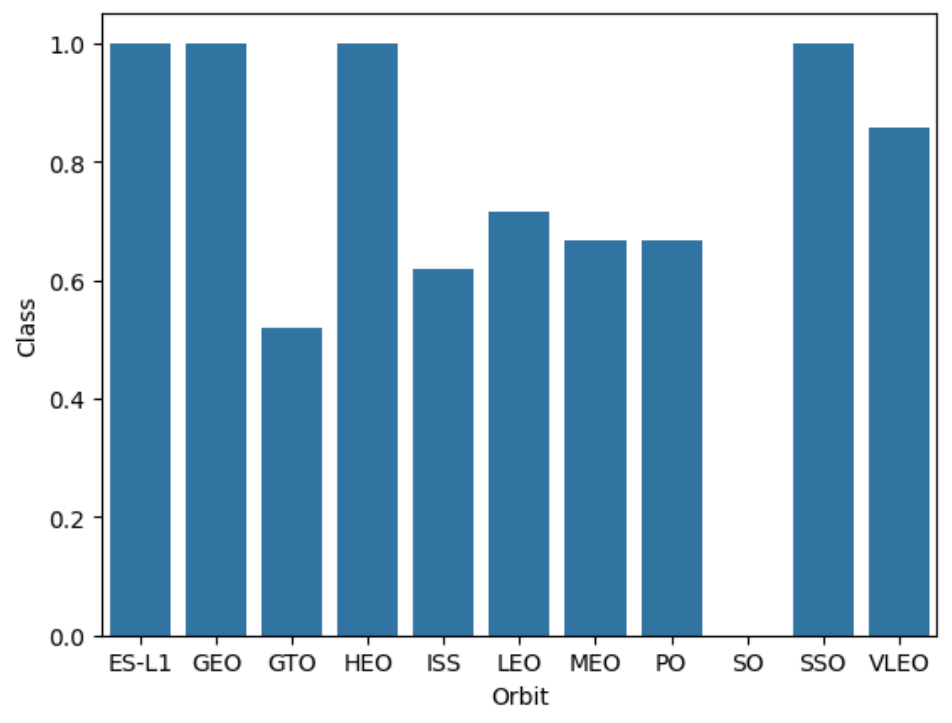
Payload vs. Launch Site



The orange indicates a successful launch and the blue indicates a failed launch

- Payload mass appears to fall mostly between 0-6000 kg.
- Different launch sites also seem to be used for different payload mass

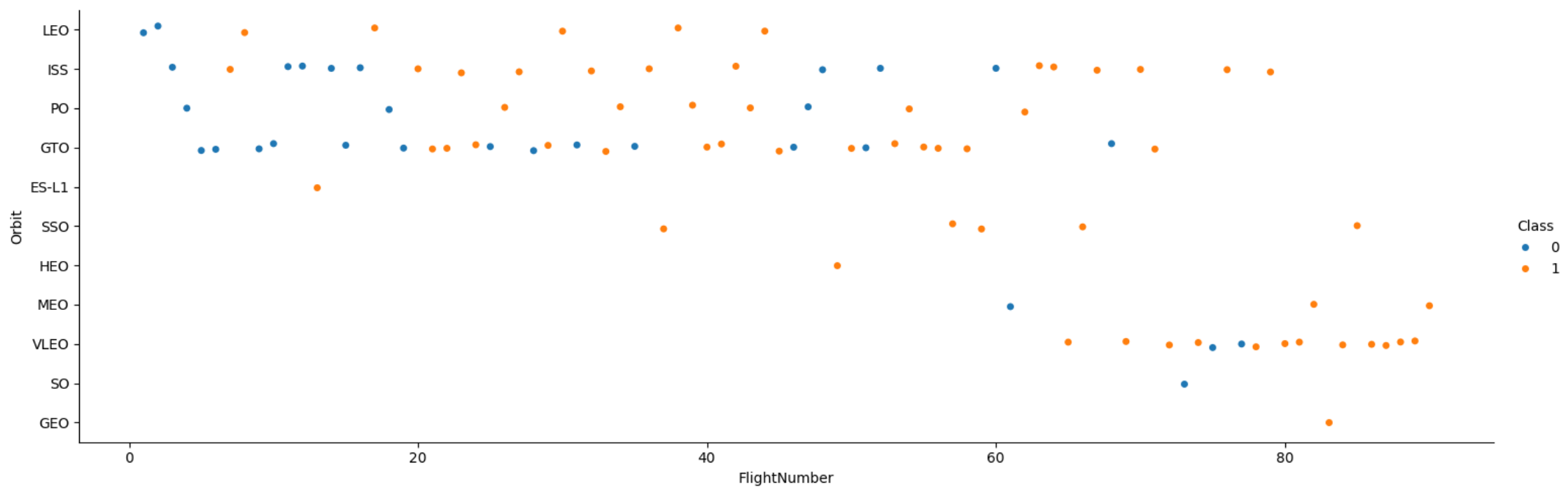
Success Rate vs. Orbit Type



Success Rate Scale with
0 as 0%
0.6 as 60%
1 as 100%

- ES-L1 (1), GEO (1), and HEO (1) have a 100% success rate (sample sizes in parenthesis), and SSO (5) has a 100% success rate.
- VLEO (14) has a decent success rate and attempts.
- SO (1) has 0% success rate
- GTO (27) has around a 50% success rate but the largest sample.

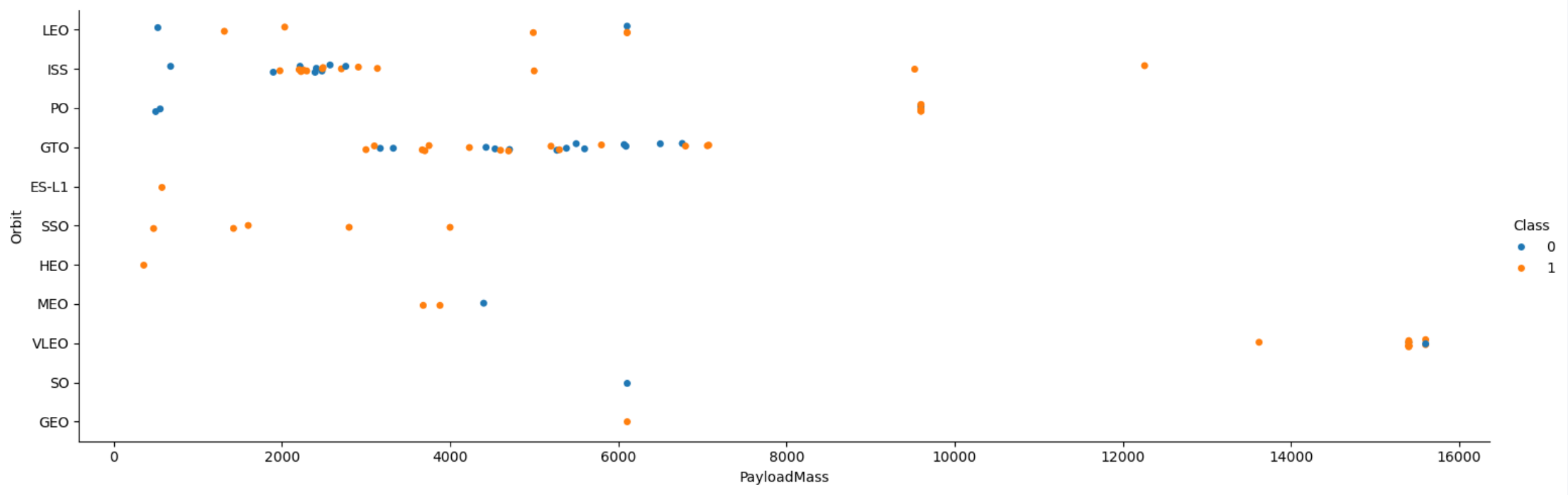
Flight Number vs. Orbit Type



The orange indicates a successful launch and the blue indicates a failed launch

- Launch Orbit preferences changed over Flight Number.
- Launch Outcome seems to correlate with this preference.
- SpaceX started with LEO orbits which saw moderate success LEO and returned to VLEO in recent launches
- SpaceX appears to perform better in lower orbits or Sun-synchronous orbits

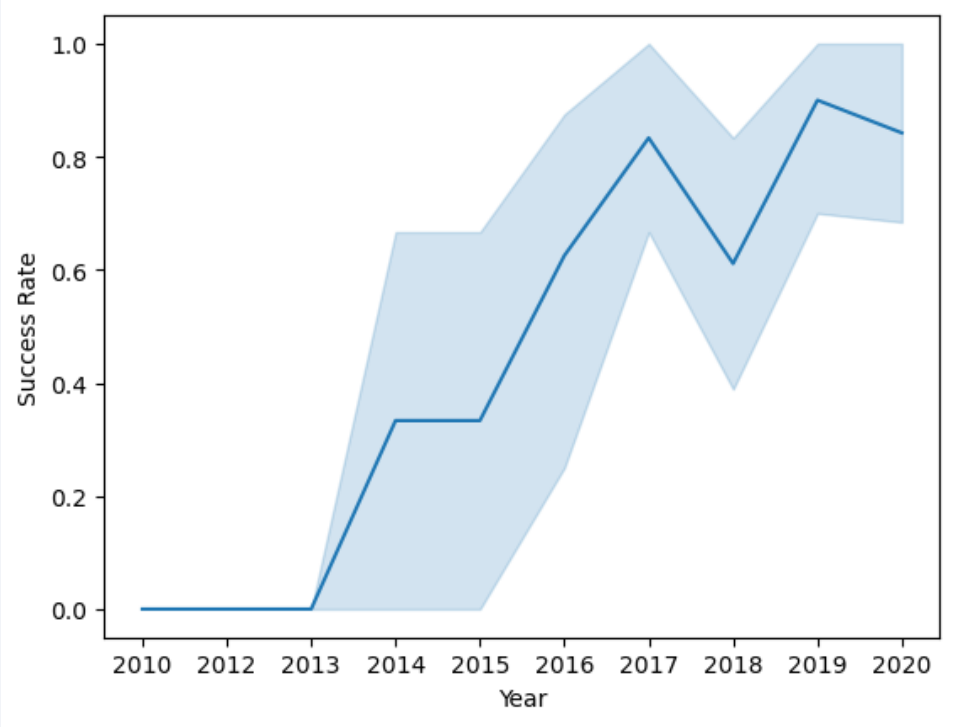
Payload vs. Orbit Type



The orange indicates a successful launch and the blue indicates a failed launch

- Payload mass seems to correlate with orbit
- LEO and SSO seem to have relatively low payload mass
- The other most successful orbit VLEO only has payload mass values in the higher end of the range

Launch Success Yearly Trend



95% confidence interval
(light blue shading)

- Success generally increased over time since 2013 with a slight dip in 2018
- Success in recent years at around 80%

All Launch Site Names

```
[9]: %sql SELECT DISTINCT "Launch_Site" FROM SPACEXTBL
 * sqlite:///my_data1.db
Done.
```

[9]:

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

- Query unique launch site names from the database.
- CCAFS SLC-40 and CCAFS LC-40 likely all represent the same launch site with data entry errors.
- CCAFS LC-40 was the previous name. Likely only 3 unique launch site values: CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E

Launch Site Names Begin with 'CCA'

```
[10]: %sql SELECT * FROM SPACEXTBL WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5
```

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

[10]:

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

```
[11]: %sql select SUM(PAYLOAD_MASS__KG_) AS 'TOTAL_PAYLOAD_BY_NASA(CRS)' from SPACEXTBL where "Customer" == 'NASA (CRS)'

 * sqlite:///my_data1.db
Done.
[11]:
```

TOTAL_PAYLOAD_BY_NASA(CRS)
45596

- This query sums the total payload mass in kg where NASA was the customer.
- CRS stands for Commercial Resupply Services, which indicates that these payloads were sent to the International Space Station (ISS).

Average Payload Mass by F9 v1.1

```
[16]: %sql select AVG(PAYLOAD_MASS__KG_) from SPACEXTBL where "Booster_Version" LIKE "F9 v1.1%"
 * sqlite:///my_data1.db
Done.
[16]: AVG(PAYLOAD_MASS__KG_)
      2534.6666666666665
```

- This query calculates the average payload mass or launches which used booster version F9 v1.1
- Average payload mass of F9 1.1 is on the low end of our payload mass range

First Successful Ground Landing Date

```
[18]: %sql select MIN("Date") as FIRST_SUCCESSFUL_GROUND_PAD_LANDING from SPACEXTBL where "Landing_Outcome" == "Success (ground pad)"
 * sqlite:///my_data1.db
Done.
[18]: FIRST_SUCCESSFUL_GROUND_PAD_LANDING
                                2015-12-22
```

- This query returns the first successful ground pad landing date.
- The first ground pad landing wasn’t until the end of 2015.
- Successful landings, in general, appear starting in 2014.

Successful Drone Ship Landing with Payload between 4000 and 6000

[31]:
```
%sql select "Booster_Version" from SPACEXTBL where "Landing_Outcome" == "Success (drone ship)" and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000
```

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

[31]:

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

- This query returns the four booster versions with successful drone ship landings and a payload mass between 4000 and 6000 non - inclusive.

Total Number of Successful and Failure Mission Outcomes

```
[63]: #%sql select (select COUNT(*) from SPACEXTBL where "Mission_Outcome" LIKE "Success%") as "Succe
      %sql select "Mission_Outcome", COUNT(*) as COUNT from SPACEXTBL GROUP BY "Mission_Outcome"
```

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

[63]:

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

- This query returns a count of each mission outcome.
- SpaceX appears to achieve its mission outcome nearly 99% of the time.
- This means that most of the landing failures are intended.
- Interestingly, one launch has an unclear payload status; unfortunately, one failed in flight.

Boosters Carried Maximum Payload

```
[20]: %sql select * from SPACEXTBL
      %sql select "Booster_Version", PAYLOAD_MASS__KG_ from SPACEXTBL where PAYLOAD_MASS__KG_ == (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

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

[20]:

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

- This query returns the booster versions that carried the highest payload mass of 15600 kg.
- These booster versions are very similar, and all are of the F9 B5 B10xx.x variety.
- This likely indicates that payload mass correlates with the booster version that is used.

2015 Failed Drone Ship Landing Records

```
[21]: %sql select substr(Date, 6,2) as MONTH, "Landing_Outcome", "Booster_Version", PAYLOAD_MASS__KG_, "Launch_Site" from SPACEXTBL where  substr(Date,0,5)='2015' and "Landing_Outcome" == "Failure (drone ship)"
 * sqlite:///my_data1.db
Done.
```

[21]:

MONTH	Landing_Outcome	Booster_Version	PAYLOAD_MASS__KG_	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	2395	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	1898	CCAFS LC-40

- This query returns the Month, Landing Outcome, Booster Version, Payload Mass (kg), and Launch site of 2015 launches where stage 1 failed to land on a drone ship.
- There were two such occurrences.

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

```
[75]: %sql select "Landing_Outcome", COUNT(*) as COUNT from SPACEXTBL where "Date" BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY "Landing_Outcome" ORDER BY COUNT desc
 * sqlite:///my_data1.db
Done.
```

[75]:

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

- This query returns a list of successful landings between 2010-06-04 and 2017-03-20 inclusively.
- There are two types of successful landing outcomes: drone ship and ground pad landings.
- There were 8 successful landings in total during this time period

Section 3

Launch Sites Proximities Analysis

Launch Site Locations

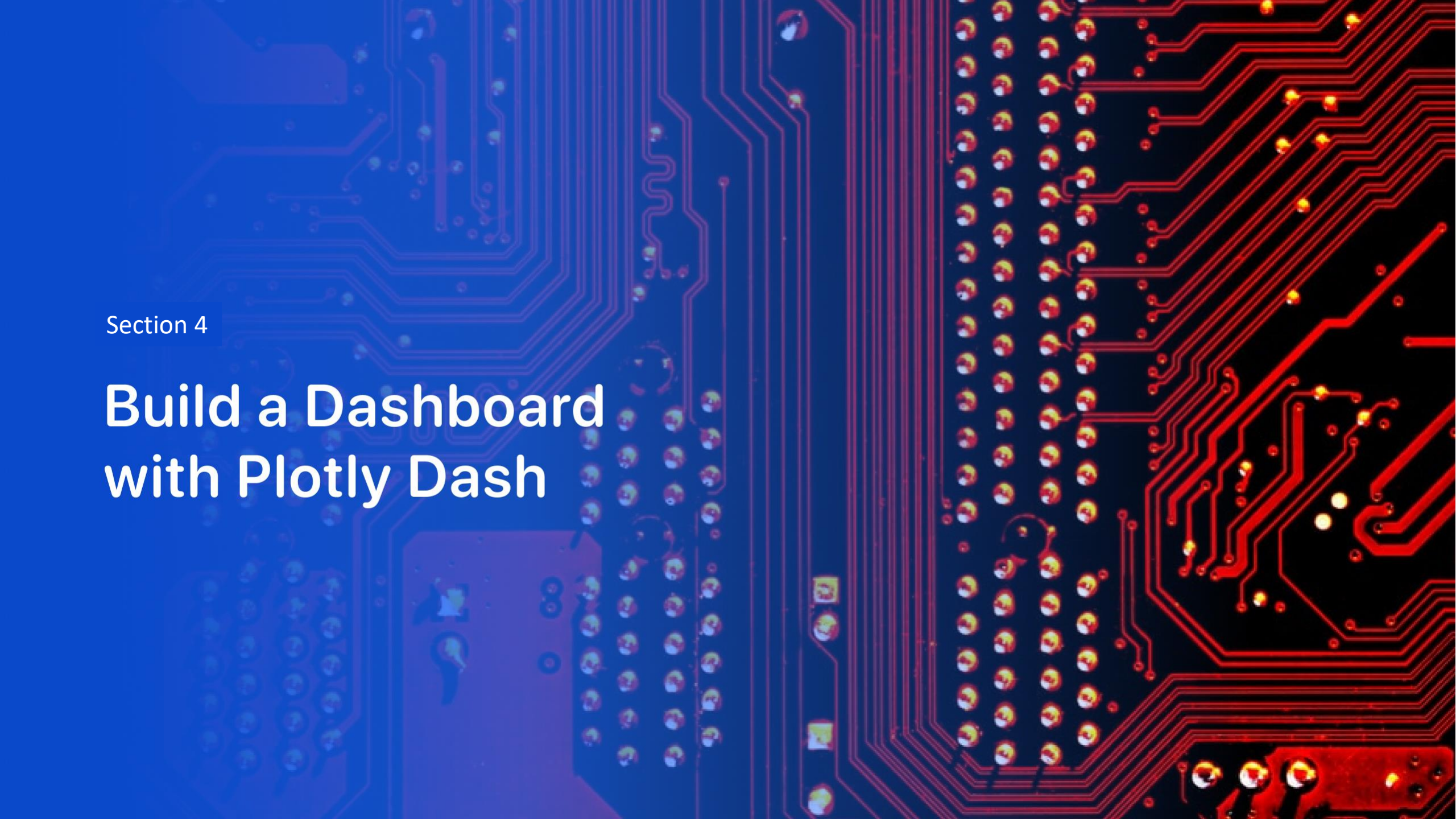
- The left map shows all launch sites relative US map. The right map shows the two Florida launch sites since they are very close to each other. All launch sites are near the ocean.

Colour-Coded Launch Markers

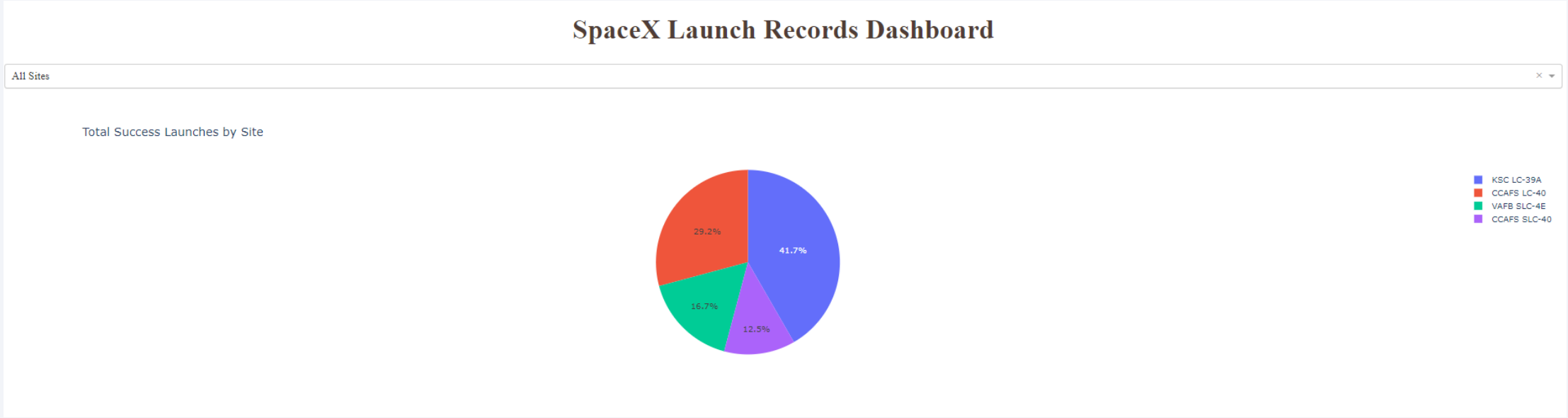
- Clusters on the Folium map can be clicked on to display each successful landing (green icon) and failed landing (red icon). In this example, CCAFS LC 40 shows 7 successful landings and 19 failed landings.

Key Location Proximities

- Using KSC LC-39A as an example, launch sites are very close to railways for large parts and supply transportation. Launch sites are close to highways for human and supply transport. Launch sites are also close to coasts and relatively far from cities so that launch failures can land in the sea to avoid rockets falling in densely populated areas.

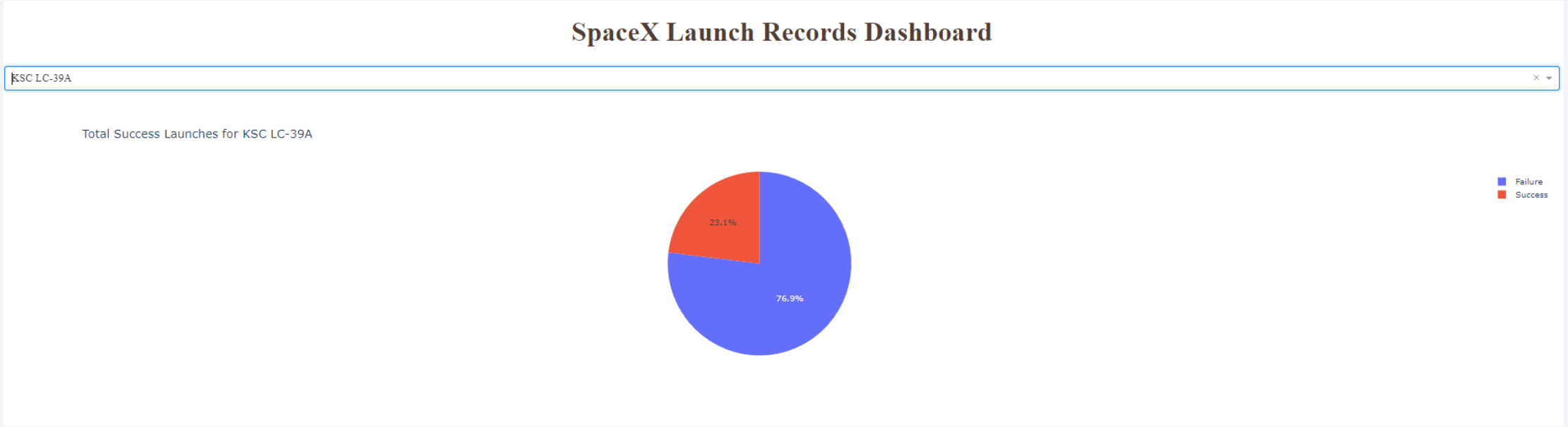

Section 4
Build a Dashboard with Plotly Dash

Successful Launches Across Launch Sites



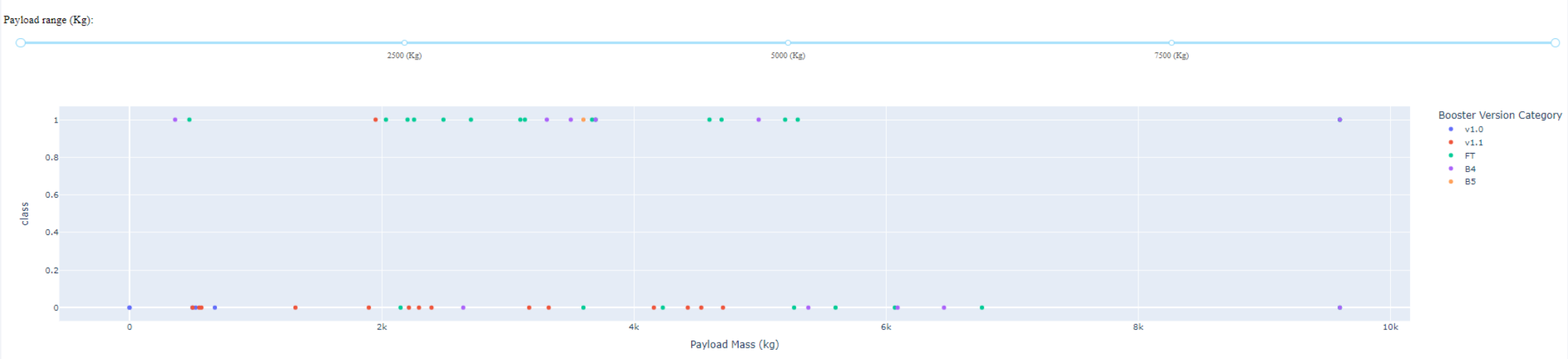
- This is the distribution of successful landings across all launch sites. CCAFS LC-40 is the old name of CCAFS SLC-40, so CCAFS and KSC have the same amount of successful landings, but a majority of the successful landings were performed before the name change. VAFB has the smallest share of successful landings. This may be due to the smaller sample and the increased difficulty of launching on the West Coast.

Highest Success Rate Launch Site



- KSC LC-39A has the highest success rate with 10 successful landings and 3 failed landings

<Dashboard Screenshot 3>

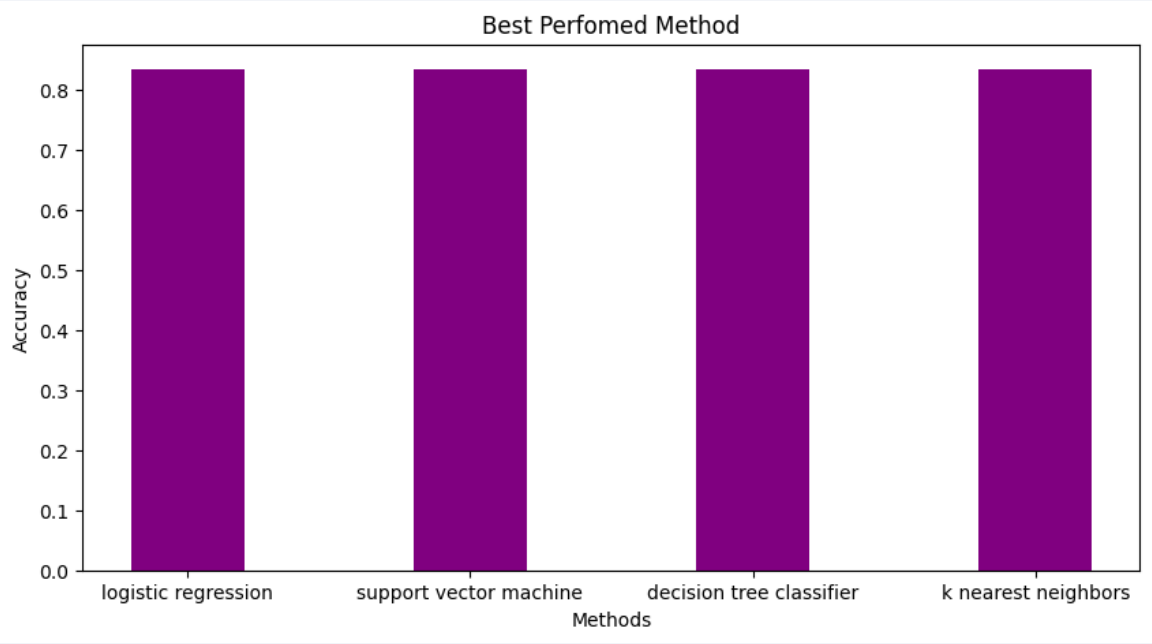


- Plotly dashboard has a Payload range selector. However, this is set from 0-10000 instead of the max Payload 15600. Class indicates 1 for successful landing and 0 for failure. The scatter plot also accounts for the booster version category in colour and the number of launches in point size. In this particular range of 0 to 6000, interestingly there are two failed landings with payloads of zero kg.

Section 5

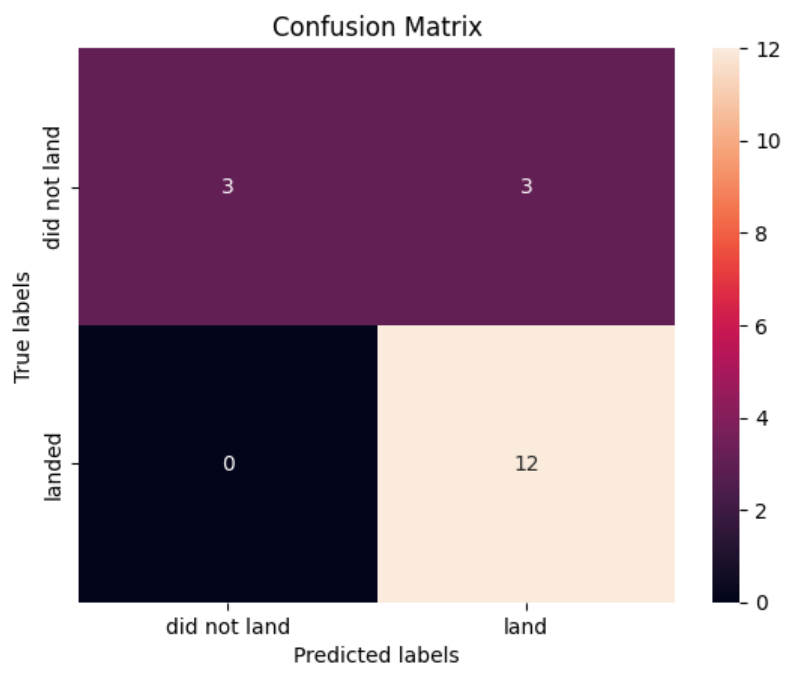
Predictive Analysis (Classification)

Classification Accuracy



- All models had virtually the same accuracy on the test set at 83.33% accuracy. It should be noted that the test size is small at only a sample size of 18.
- This can cause significant variance in accuracy results, such as those in Decision Tree Classifier model in repeated runs.
- We likely need more data to determine the best model.

Confusion Matrix



- Since all models performed the same for the test set, the confusion matrix is the same across all models. The models predicted 12 successful landings when the actual label was successful landing.
- The models predicted 3 unsuccessful landings when the actual label was unsuccessful landing.
- The models predicted 3 successful landings when the valid label was unsuccessful landings (false positives). Our models overpredict successful landings.

Conclusions

- Our task: to develop a machine learning model for Space Y which wants to bid against SpaceX
- The goal of the model is to predict when Stage 1 will successfully land to save ~$100 million USD
- Used data from a public SpaceX API and web scraping SpaceX Wikipedia page
- Created data labels and stored data in a DB2 SQL database
- Created a dashboard for visualization
- We created a machine-learning model with an accuracy of 83%
- Allon Mask of SpaceY can use this model to predict with relatively high accuracy whether a launch will have a successful Stage 1 landing before launch to determine whether the launch should be made or not
- If possible, more data should be collected to determine the best machine learning model better and improve accuracy

Appendix

- SpaceX database CSV file link:

  https://github.com/Harshit-Optimus0089/IBM_Data_Science_Capstone/blob/7ca0f11840aaf999e43eb1bbacd18eabc06ef633/Spacex.csv

- GitHub Repository link:

  https://github.com/Harshit-Optimus0089/IBM_Data_Science_Capstone

Thank you!