IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Panagiotis Prassas>
<27 July 2024>

Outline

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

Executive Summary

- Summary of methodologies:
  - Data Collection with API & Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis (EDA) using SQL
  - Exploratory Data Analysis (EDA) with Data Visualization
  - Interactive Visual Analytics with Folium
  - Building an Interactive Dashboard with Plotly Dash
  - Machine Learning (ML) Prediction
- Summary of all results
  - It was possible to collected valuable data from public sources
  - EDA allowed to identify which of the features are the best ones in order to predict success of landings
  - ML Prediction results showed which model is the best in order to predict which characteristics are crucial to drive this opportunity by the best way, by using all the data that have been collected

Introduction

- Project background and context

  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. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch

  So the main objective is to evaluate the viability of a new company named SpaceY to compete with SpaceX

- Problems you want to find answers
  - How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?
  - What is the best algorithm that can be used for binary classification in this case?
  - Where is the best place in order to perform the launches

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from 2 public sources:
    - SpaceX API (https://api.spacexdata.com/v4/rockets/)
    - Web Scraping (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)
- Perform data wrangling
  - Collected data structured based on a landing outcome after analyzing features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash

Methodology

Executive Summary

- Perform predictive analysis using classification models
    - Collected Data after Wrangling, EDA and Visualization, then it was divided in training and test datasets and was evaluated by 4 different classification models (Logistic Regression, SVM, KNN, Decision Tree) and each accuracy of these models was evaluated by using different combinations of parameters

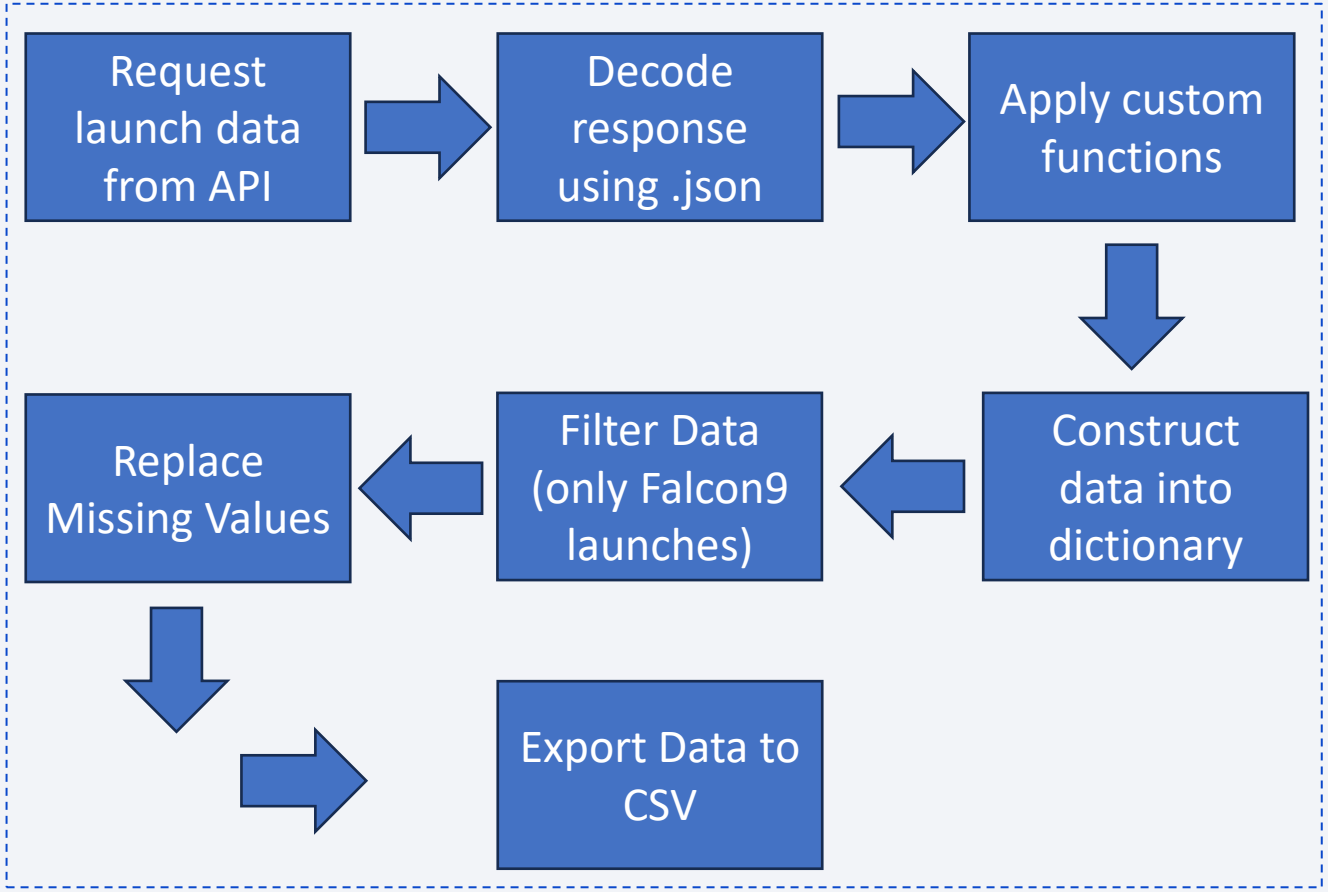
Data Collection

- Data was collected from :
  - SpaceX: https://api.spacexdata.com/v4/rockets/ (API used)
  - Wikipedia: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches (Web Scraping used)

Data Collection – SpaceX API

- SpaceX provides a public API from where data can be obtained(https://api.spacexdata.com/v4/rockets/)

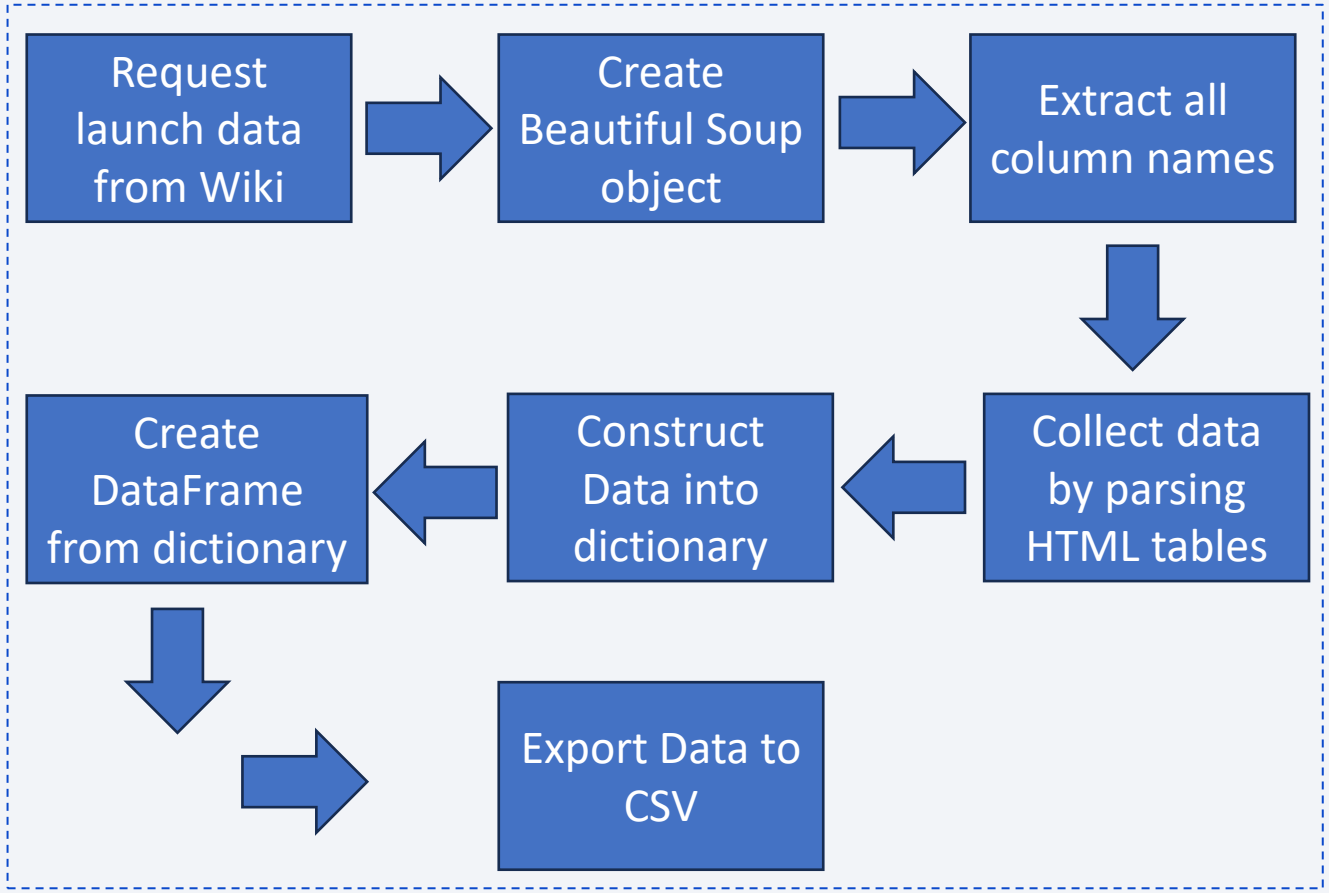
- Link to the notebook: https://github.com/PanagiotisPrassas/IBM_Applied_Data_Science_Capstone/blob/main/01.Data%20Collection%20API.ipynb



Data Collection – Web Scraping

- Wikipedia also provides data (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)

- Link to the notebook: https://github.com/PanagiotisPrassas/IBM_Applied_Data_Science_Capstone/blob/main/02.Data%20Collection%20with%20Web%20Scraping.ipynb



Data Wrangling

- We performed Exploratory Data Analysis (EDA) to the Data and we determined the training labels
- We calculated the number of the launches of each site and the number of each orbits
- After analyzing features, we created a landing outcome label and export all results to a CSV

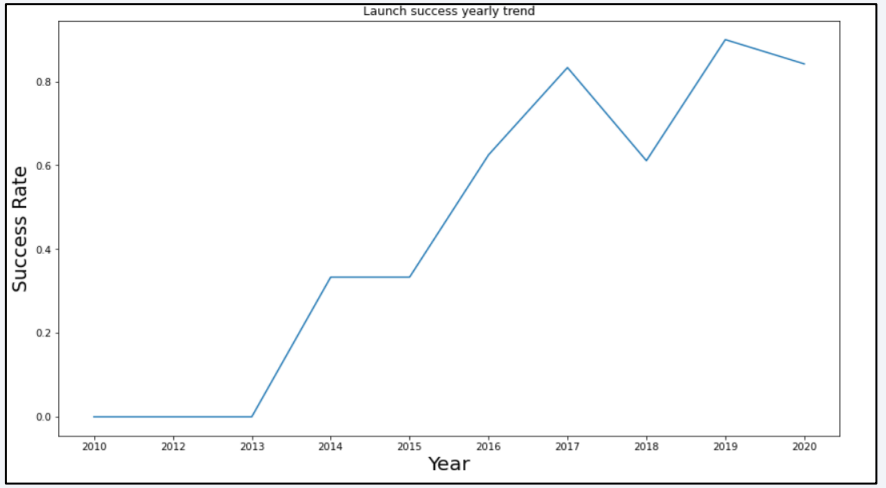
- Link to the notebook :

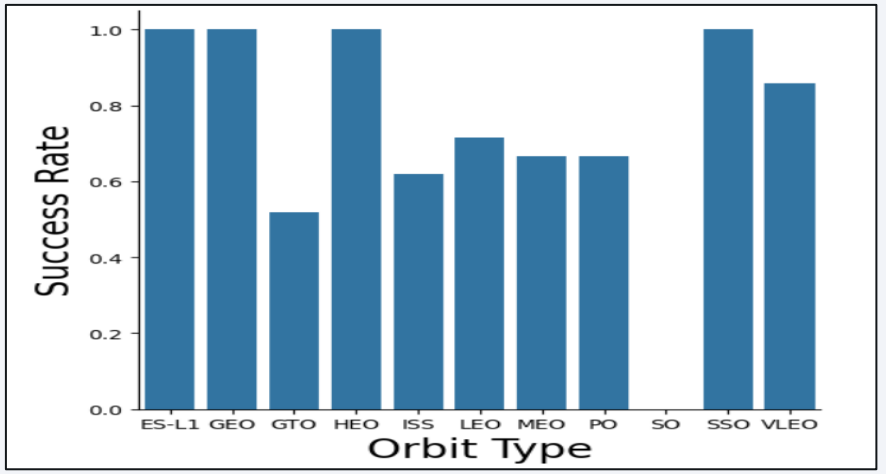
https://github.com/PanagiotisPrassas/IBM_Applied_Data_Science_Capstone/blob/main/03.Data%20Wrangling.ipynb

EDA with Data Visualization

- We performed EDA to the data by visualizing the relationships between the features such as flight number, launch site, payload, success rate of each orbit type and launch success yearly trend

- Link to the notebook : https://github.com/PanagiotisPrassas/IBM_Applied_Data_Science_Capstone/blob/main/05.Exploratory%20Data%20Analysis%20with%20Data%20Visualization.ipynb





EDA with SQL

- We downloaded the dataset and stored it in database table
- We performed SQL Queries to find out:
  - The names of the unique launch sites in the space mission
  - 5 records where launch sites begin with the string 'CCA'
  - The total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1
  - The date when the first successful landing outcome in ground pad was achieved
  - The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - The total number of successful and failure mission outcomes
  - The names of the booster versions which have carried the maximum payload mass

EDA with SQL

- The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Ranking 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

- Link to the notebook : https://github.com/PanagiotisPrassas/IBM_Applied_Data_Science_Capstone/blob/main/04.Exploratory%20Data%20Analysis%20with%20SQL.ipynb

Build an Interactive Map with Folium

- We marked all launch sites, and added map objects such as markers, circles, lines to mark the success or failure of launches for each site on the folium map.
- We assigned the feature launch outcomes (failure or success) to class 0 and 1.i.e., 0 for failure, and 1 for success.
- Using the color-labeled marker clusters, we identified which launch sites have relatively high success rate.

- Link to the notebook : https://github.com/PanagiotisPrassas/IBM_Applied_Data_Science_Capstone/blob/main/06.Interactive%20Visual%20Analystics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- We built an interactive dashboard with Plotly dash
- We plotted pie charts showing the total launches by a certain sites
- We plotted scatter graph showing the relationship with Outcome and Payload Mass (Kg) for the different booster version.

- Link to the notebook : https://github.com/PanagiotisPrassas/IBM_Applied_Data_Science_Capstone/blob/main/07.Interactive%20Dashboard%20with%20Ploty%20Dash.py

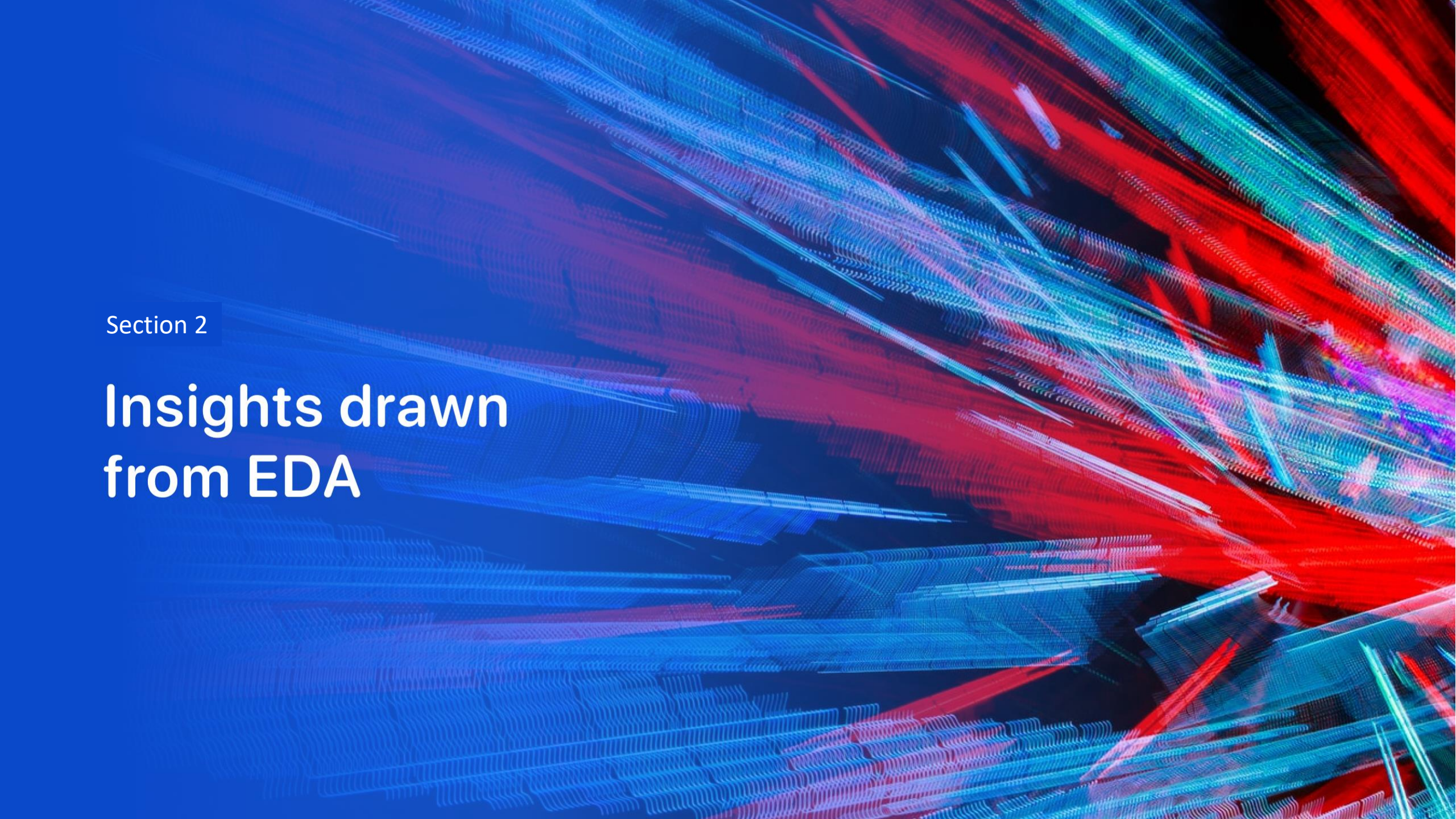
Predictive Analysis (Classification)

- We loaded the data using NumPy and Pandas, transformed the data, split our data into training and testing
- We built different machine learning models and tune different hyperparameters using GridSearchCV
- We used accuracy as the metric for our model, improved the model using feature engineering and algorithm tuning
- We found the best performing classification model

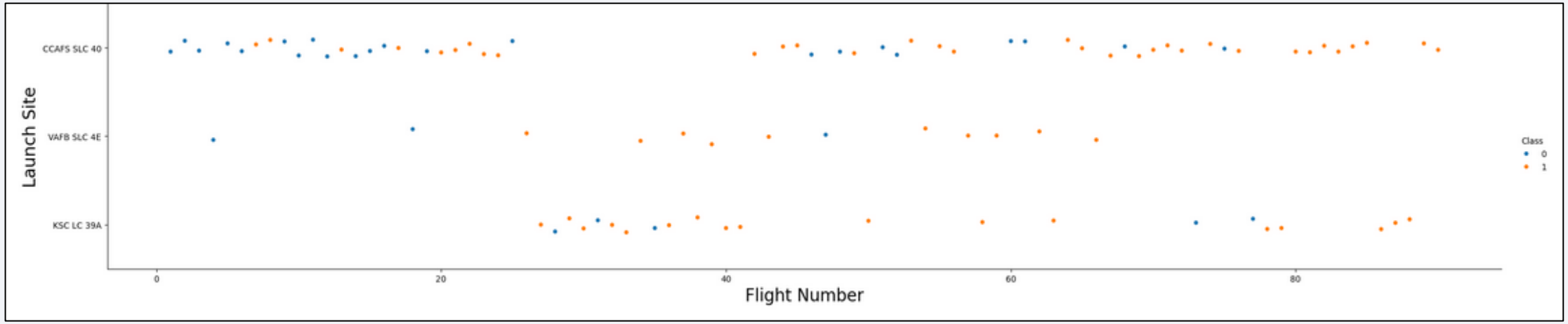
- Link to the notebook : https://github.com/PanagiotisPrassas/IBM_Applied_Data_Science_Capstone/blob/main/08.Machine%20Learning%20Prediction%20Part.ipynb

Results

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


Section 2
Insights drawn from EDA

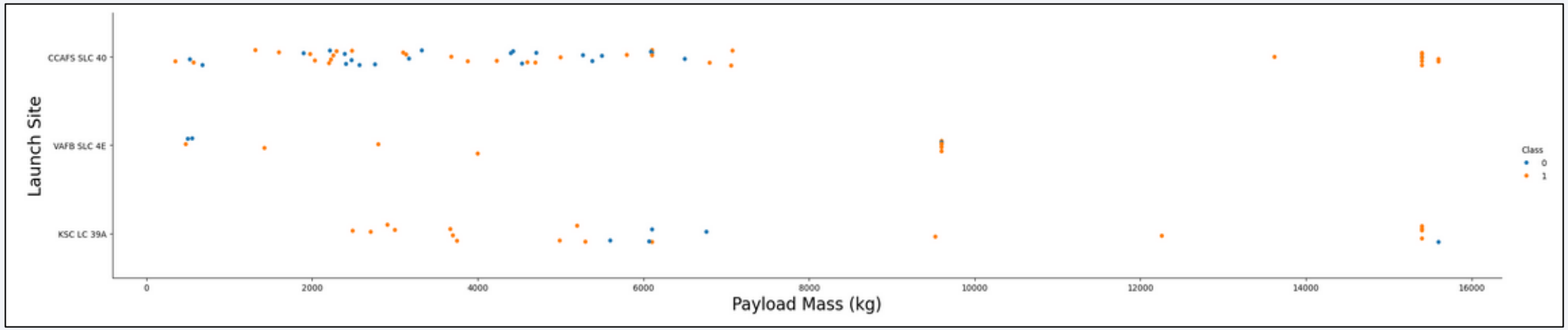
Flight Number vs. Launch Site



Explanation:

- The earliest flights all failed while the latest flights all succeeded
- The CCAFS SLC 40 launch site has about a half of all launches
- VAFB SLC 4E and KSC LC 39A have higher success rates
- It can be assumed that each new launch has a higher rate of success

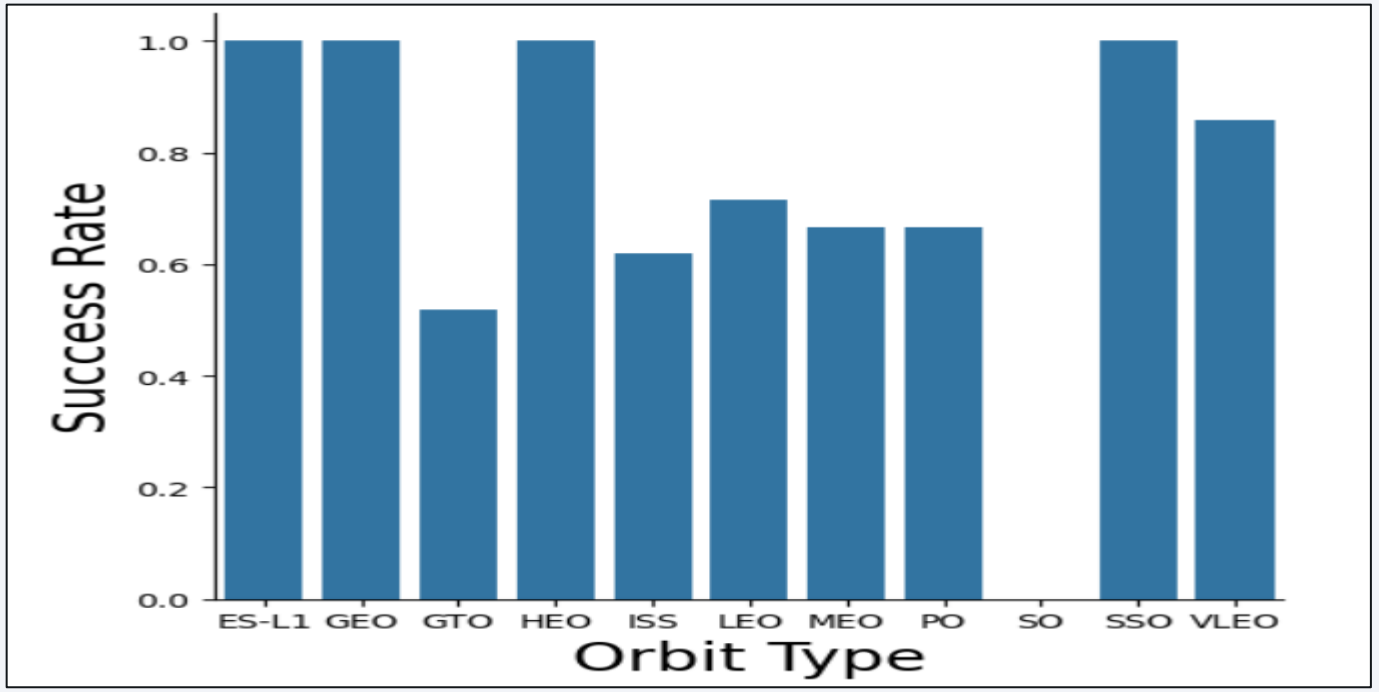
Payload vs. Launch Site



Explanation:

- For every launch site the higher the payload mass, the higher the success rate
- Most of the launches with payload mass over 7000 kg were successful
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too

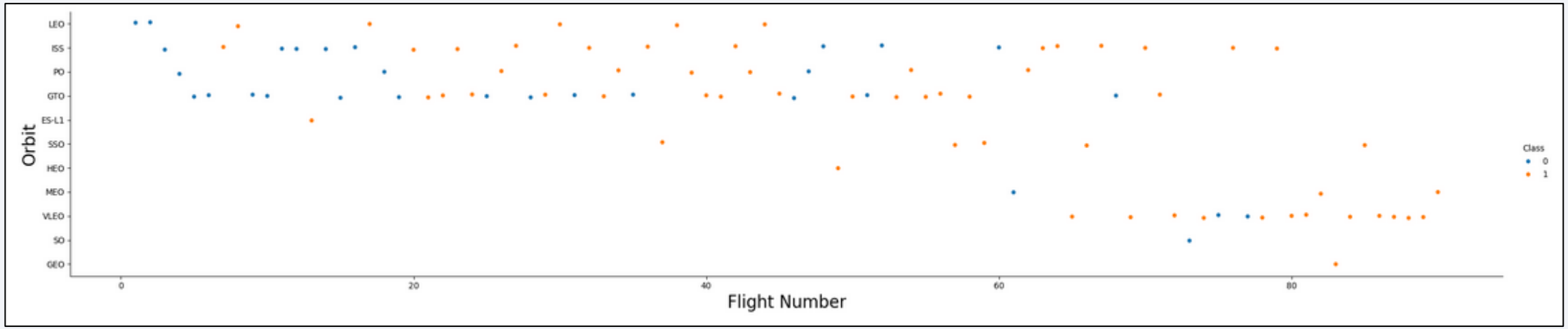
Success Rate vs. Orbit Type



Explanation:

- Orbits with 100% success rate: - ES-L1, GEO, HEO, SSO
- Orbits with success rate between 50% and 85%: - GTO, ISS, LEO, MEO, PO
- Orbits with 0% success rate: - SO

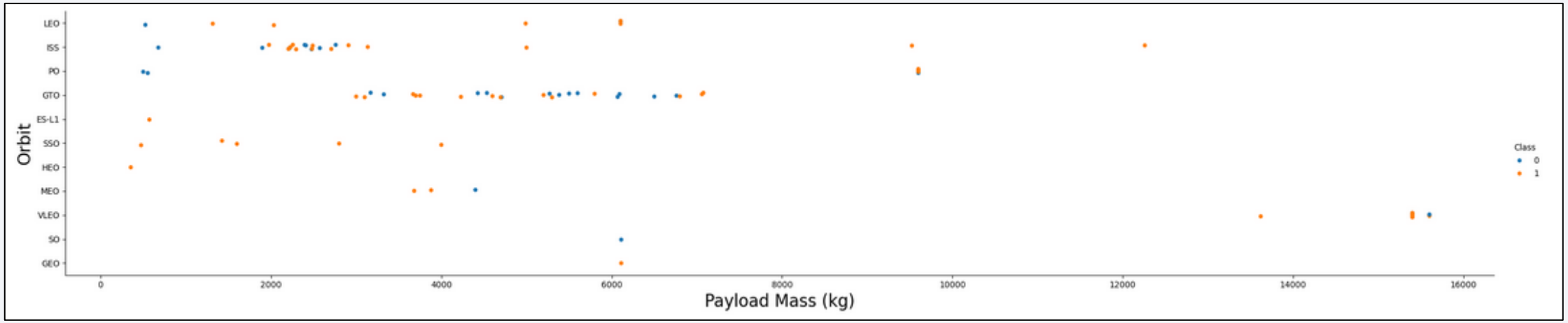
Flight Number vs. Orbit Type



Explanation:

- In the LEO orbit, success is related to the number of flights whereas in the GTO orbit, there is no relationship between flight number and the orbit.

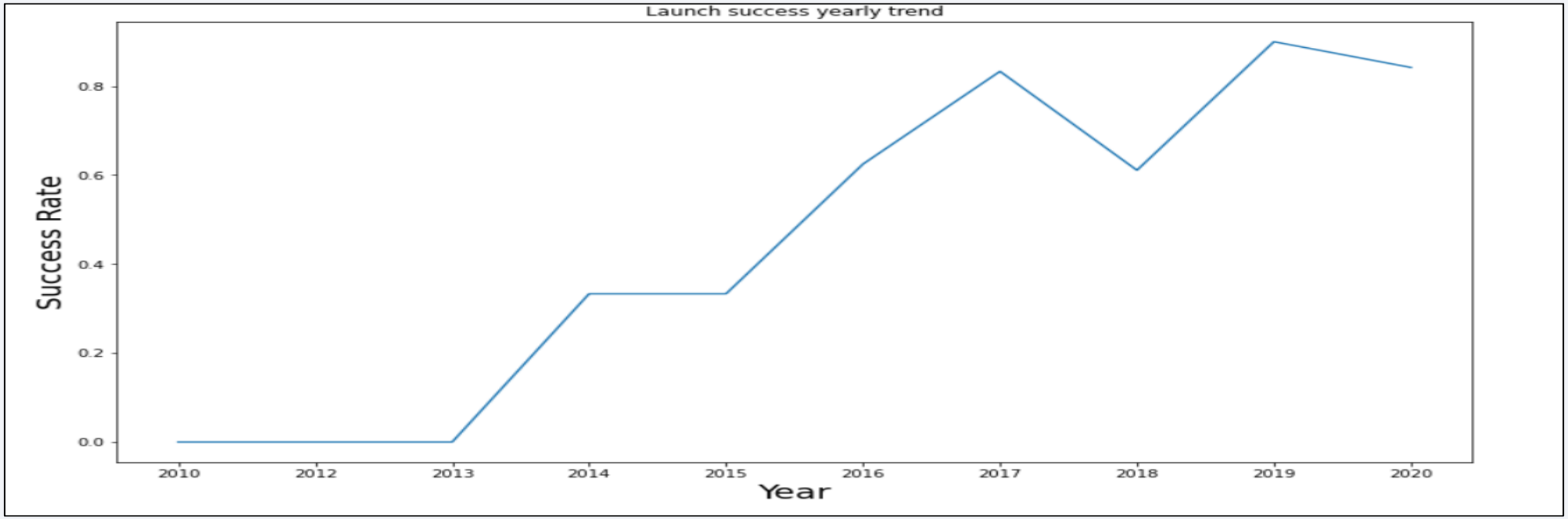
Payload vs. Orbit Type



Explanation:

- With heavy payloads, the successful landing are more for PO, LEO and ISS orbits

Launch Success Yearly Trend



Explanation:

- The success rate since 2013 kept increasing till 2020

All Launch Site Names

- We used DISTINCT in order to show all unique launch sites

```
%%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'

- We performed a SQL query to display 5 records where launch sites begin with the string ‘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
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success
2010-08-12	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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success

Total Payload Mass

- We performed a SQL query to display the total mass carried by boosters launched by NASA (CRS)

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE
WHERE Customer = 'NASA (CRS)';
```

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- We performed a SQL query to display the average payload mass carried by booster version F9 v1.1

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

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

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- We performed a SQL query to list the date when the first successful landing outcome in group pad was achieved

```
%%sql
SELECT min(Date) FROM SPACEXTABLE
WHERE UPPER(Landing_Outcome) LIKE '%SUCCESS%GROUND%PAD%';
```

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

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- We 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

```
%%sql
SELECT Booster_Version FROM SPACEXTABLE
WHERE UPPER(Landing_Outcome) LIKE '%SUCCESS%DRONE_SHIP%'
AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

```
 * 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

- We performed a SQL Query in order to calculate the total number of successful and failure mission outcomes

```
%%sql
SELECT Mission_Outcome, COUNT(*) AS NUMBER FROM SPACEXTABLE
GROUP BY Mission_Outcome;
```

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

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

Boosters Carried Maximum Payload

- We performed a SQL Query and Subquery in order to List the names of the booster versions which have carried the maximum payload mass

```
%%sql
SELECT Booster_Version FROM SPACEXTABLE
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);
```

```
 * 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

- We performed a SQL Query in order to list the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%%sql
SELECT SUBSTR(Date, 6, 2) as MONTH, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE
WHERE UPPER(Landing_Outcome) LIKE '%FAILURE%DRONE%SHIP%';
```

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

MONTH	Landing_Outcome	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
01	Failure (drone ship)	F9 v1.1 B1017	VAFB SLC-4E
04	Failure (drone ship)	F9 FT B1020	CCAFS LC-40
06	Failure (drone ship)	F9 FT B1024	CCAFS LC-40

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

- We performed a SQL Query in order to 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(Landing_Outcome) AS COUNT FROM SPACEXTABLE
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY COUNT DESC;
```

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

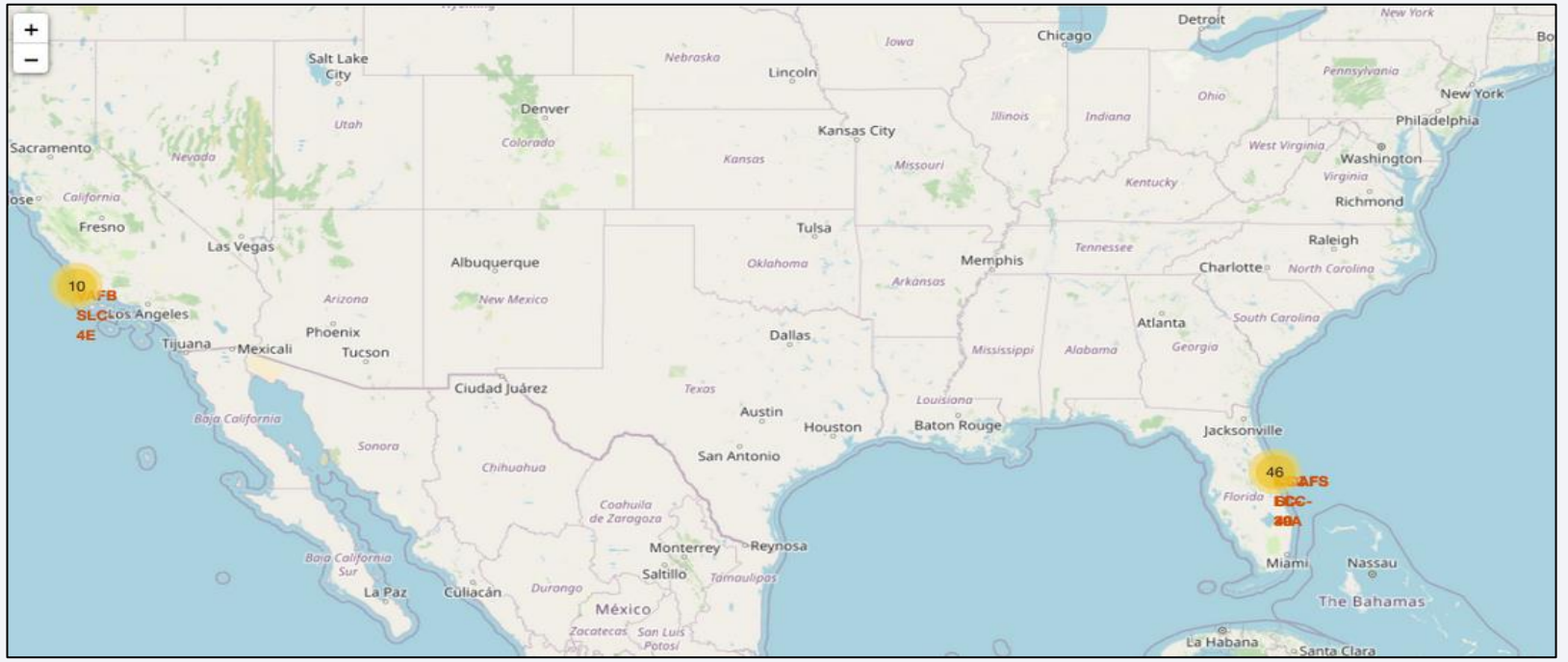
Landing_Outcome	COUNT
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

Section 3

Launch Sites Proximities Analysis

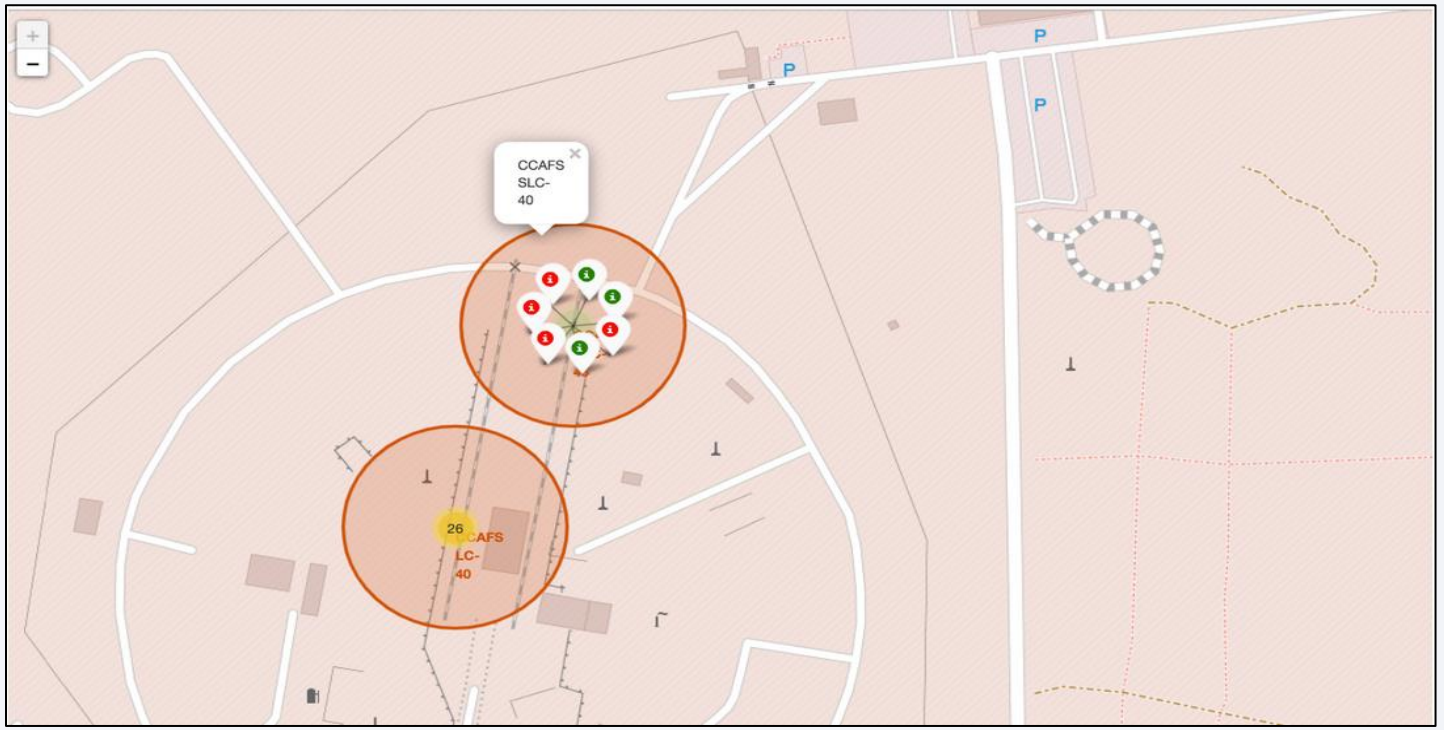
All launch sites location markers on global map

- All launch sites are in very close proximity to the coast, while launching rockets towards the ocean it minimizes the risk of having any debris dropping or exploding near people.



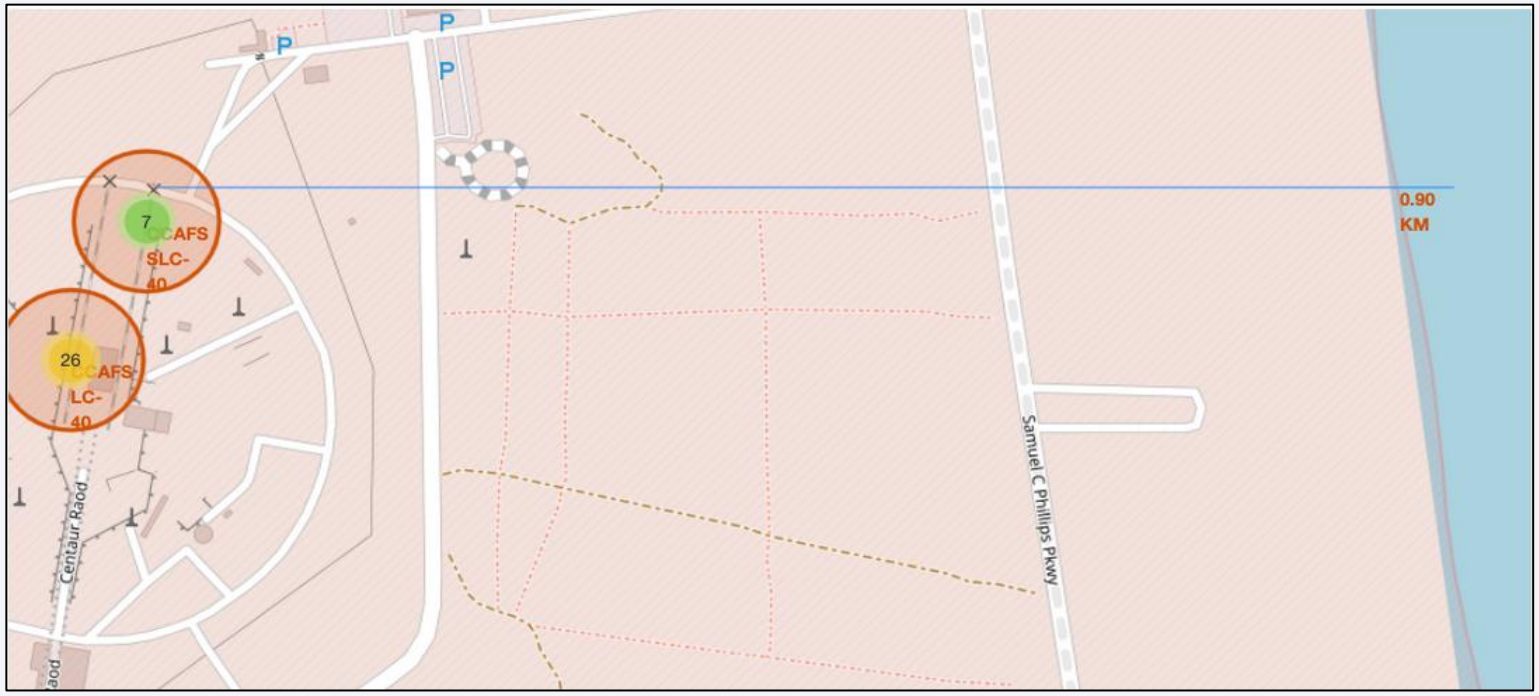
Color-labeled launch records on the map

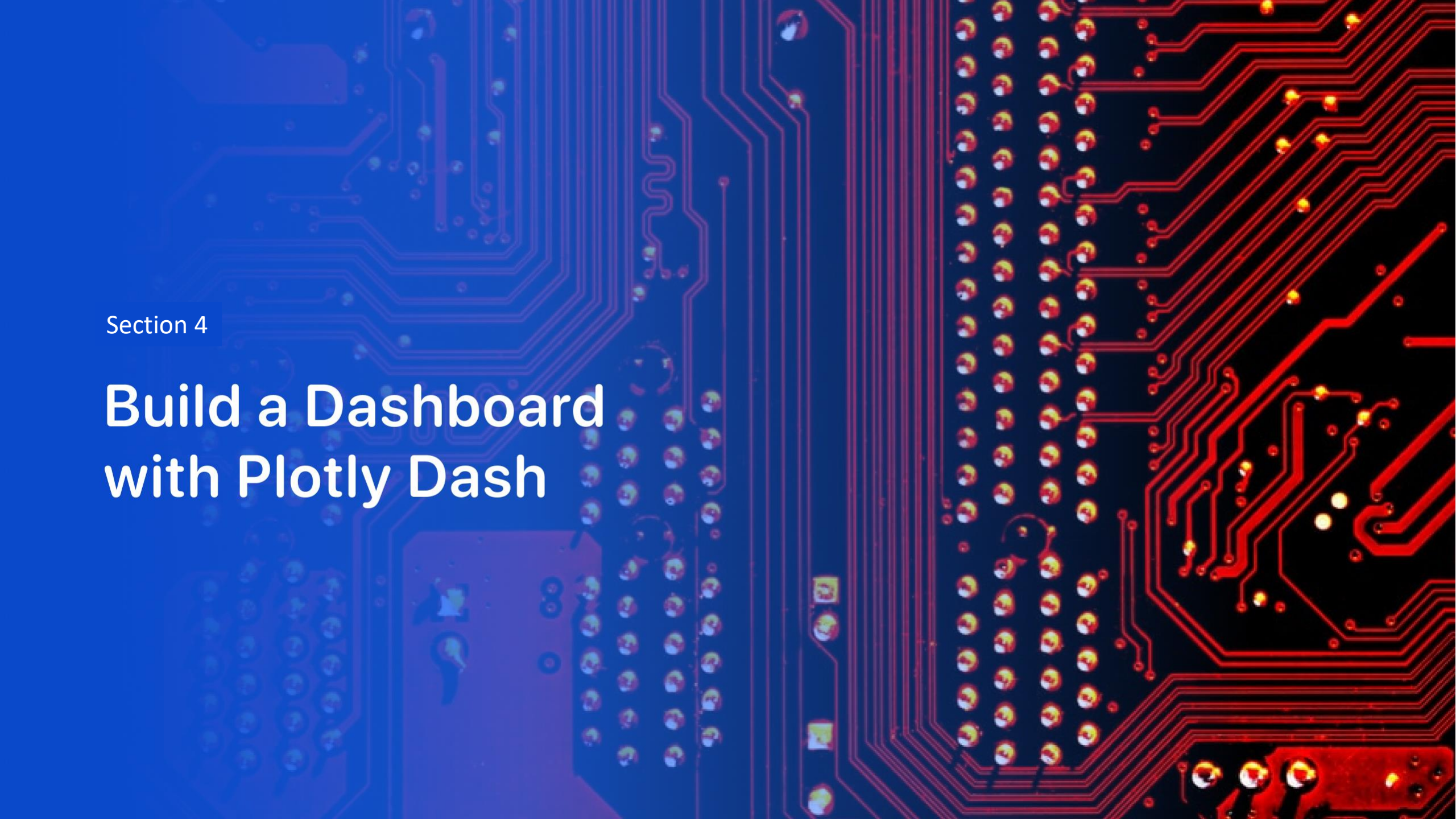
- From the color-labeled markers we should be able to easily identify which launch sites have relatively high success rates :
  - Green Marker = Successful Launch
  - Red Marker = Failed Launch



Launch Site distance to coast

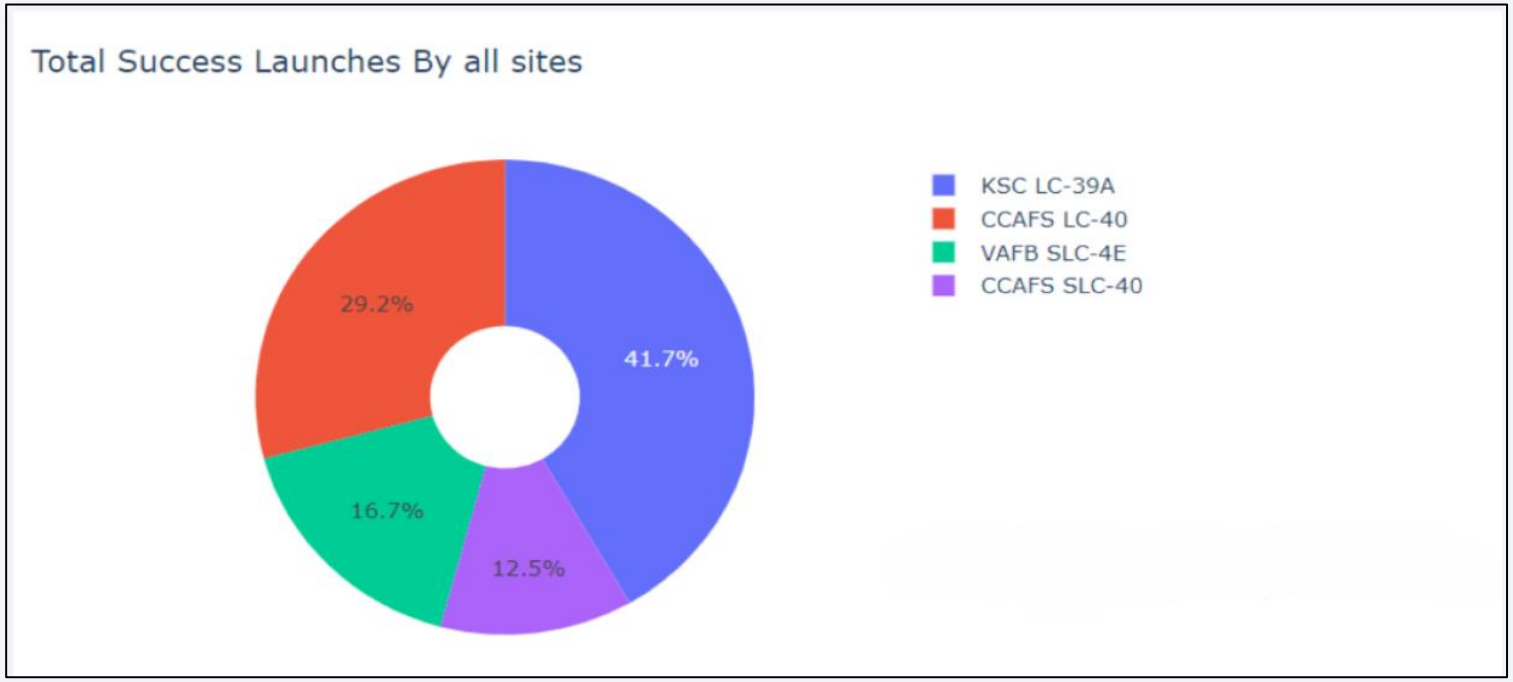
- From the visual analysis we can see that the AFS SLC-40 launch site is 0.90km away from the coast




Section 4
Build a Dashboard with Plotly Dash

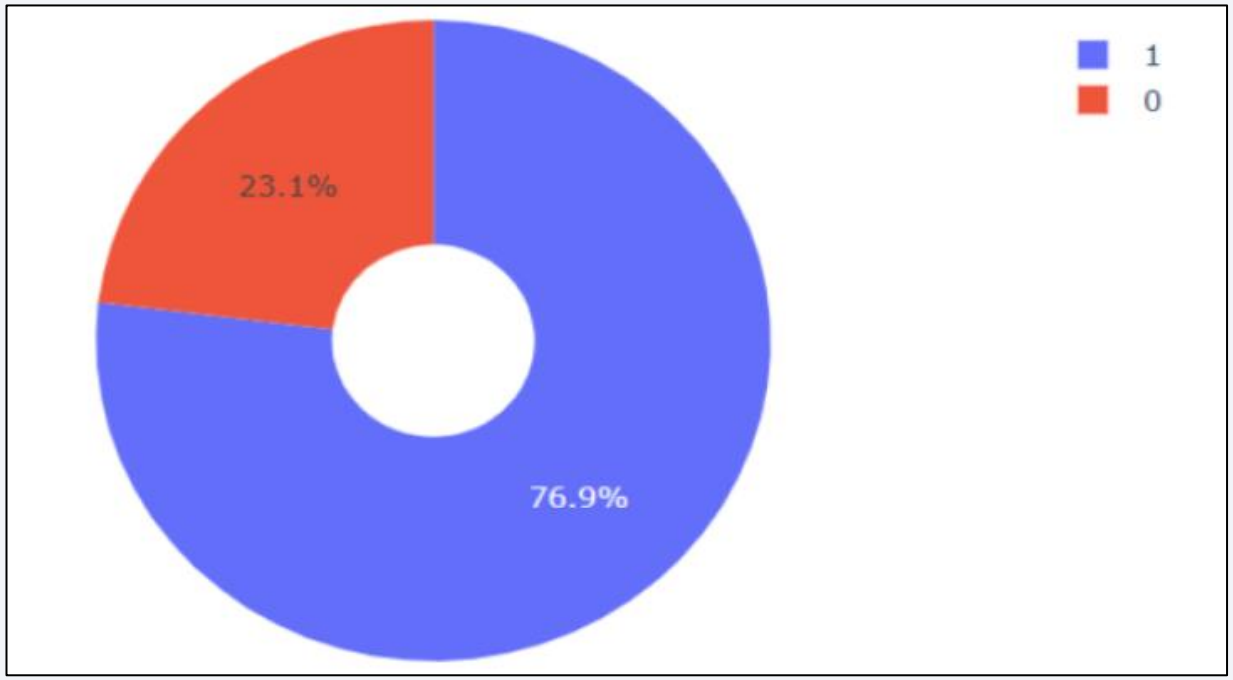
Total Success Launches By all sites

- We can see that the launch site with the most successful launches is the KSC LC-39A



Launch site with the highest launch success ratio

- KSC LC-39A, the launch site with the highest launch success ratio, achieved a 76.9% success rate along with a 23.1% failure rate



Payload vs. Launch Outcome scatter plot for all sites

- Scatter plots show that payloads between 2000 and 5500 kg have the highest success rate

Section 5

Predictive Analysis (Classification)

Classification Accuracy

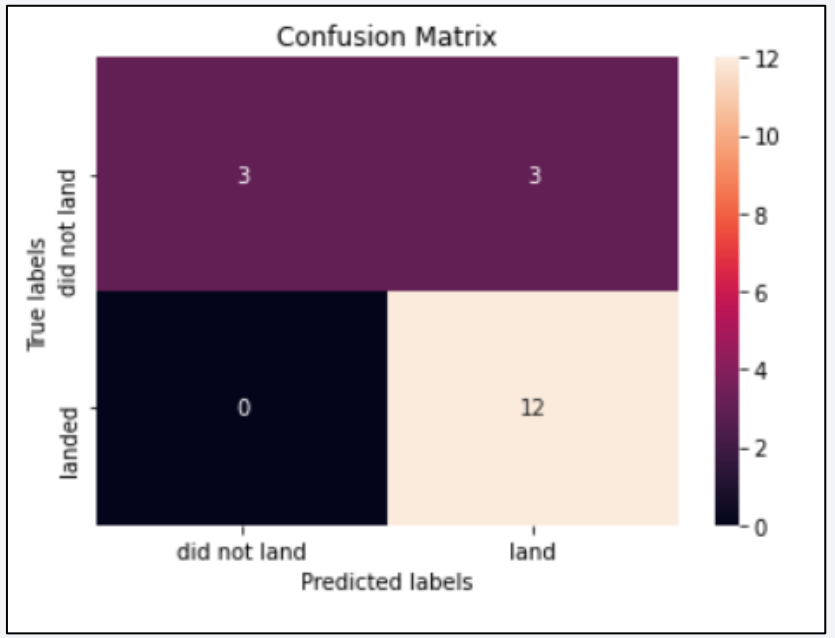
- All of the models have the same accuracy, so we performed GridSearchCV best scores in order to rank them

```
# Since their accuracies are all the same, we pick based on their best scores
algo_score = {'Logistic regresssion': [logreg_cv.best_score_], 'SVM': [svm_cv.best_score_], 'Decision tree': [tree_cv.best_s
df = pd.DataFrame.from_dict(algo_score, orient='index', columns=['Best scores'])
df
```

	Best scores
Logistic regresssion	0.846429
SVM	0.848214
Decision tree	0.889286
KNN	0.848214

Confusion Matrix

- In this visualization we show the confusion matrix of the best performing model (Decision Tree model). It visualizes that the classifier can distinguish between the different classes and the main problem is that the false positives (unsuccessful landings) are marked as successful landing by the classifier



Conclusions

Based on the methods used on the Data such as Wrangling, Visualization, EDA and ML we can present some interesting conclusions :

- In this project, we tried to predict if the first stage of a given Falcon 9 launch will land in order to determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch. Each feature of a Falcon 9 launch, such as its payload mass or orbit type, may affect the mission outcome in a certain way such as :
  - The larger the flight amount at a launch site, the greater the success rate at a launch site.
  - Launch success rate started to increase in 2013 till 2020.
  - Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
  - KSC LC-39A had the most successful launches of any sites
  - The Decision tree classifier is the best machine learning algorithm for this task.

Appendix

- Source code and notebooks links from GitHub : https://github.com/PanagiotisPrassas/IBM_Applied_Data_Science_Capstone/tree/main

Thank you!