IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Alejandra Marín Olavarría
20 de Mayo, 2025

Outline

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

Executive Summary

- SpaceX has become a global leader in affordable and reliable space launches. With a significant cost reduction compared to competitors (from $165M to $62M per launch), agencies like NASA have partnered with them.
- In this report, I’m taking the role of a Data Scientis, this project imagines a new competitor, SpaceY, founded by billionaire Allon Musk. As a data scientist for SpaceY, the goal was to explore whether SpaceX's success could be rivaled using data science. The project journey includes data collection, preprocessing, analysis, predictive modeling, and dashboard creation.

Introduction

- The space race has been ongoing since 1957, with costs remaining extremely high. SpaceX disrupted the market by introducing reusable first-stage rockets, making missions far more cost-effective.

- This presentation dives into identifying which factors (like orbit, payload, and launch site) influence the success of first-stage landings, using data science methods provided by IBM.

Section 1

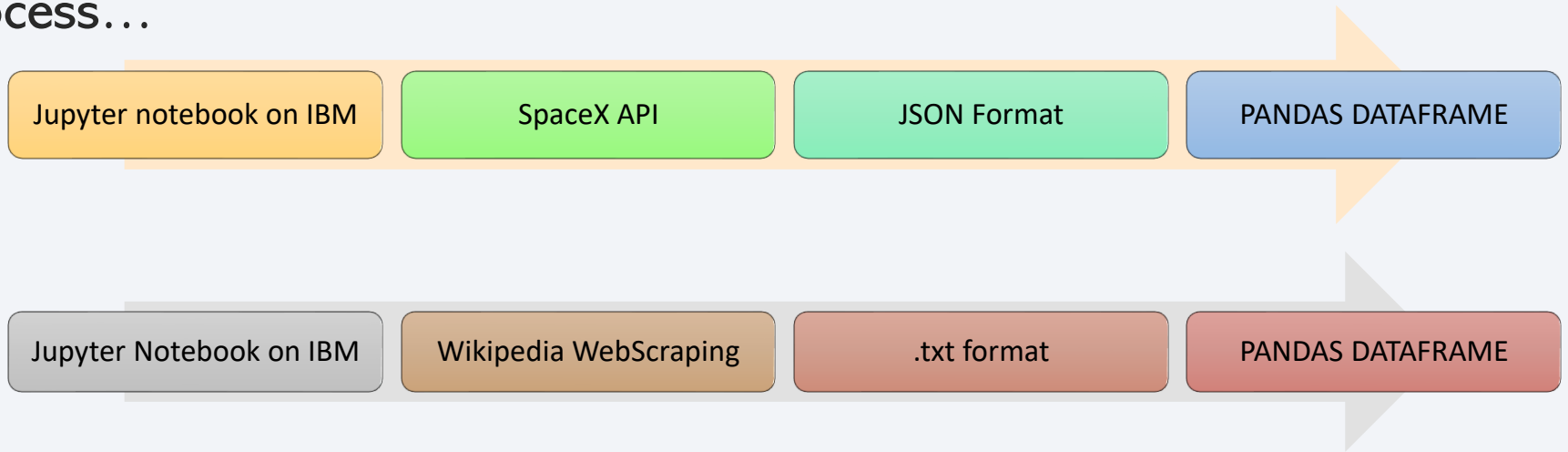
Methodology

Methodology

- Data collection methodology: Data was sourced from the SpaceX public API and through scraping Wikipedia.
- Perform data wrangling: Using Pandas and NumPy, we cleaned, normalized, and encoded data.
- Perform exploratory data analysis (EDA): Visuals via Matplotlib/Seaborn and SQL queries were used to understand key variables.
- Perform interactive visual analytics: Implemented through Folium maps and Plotly Dash dashboards.
- Perform predictive analysis using classification models: Built classification models to predict landing success, followed by model selection via GridSearchCV.

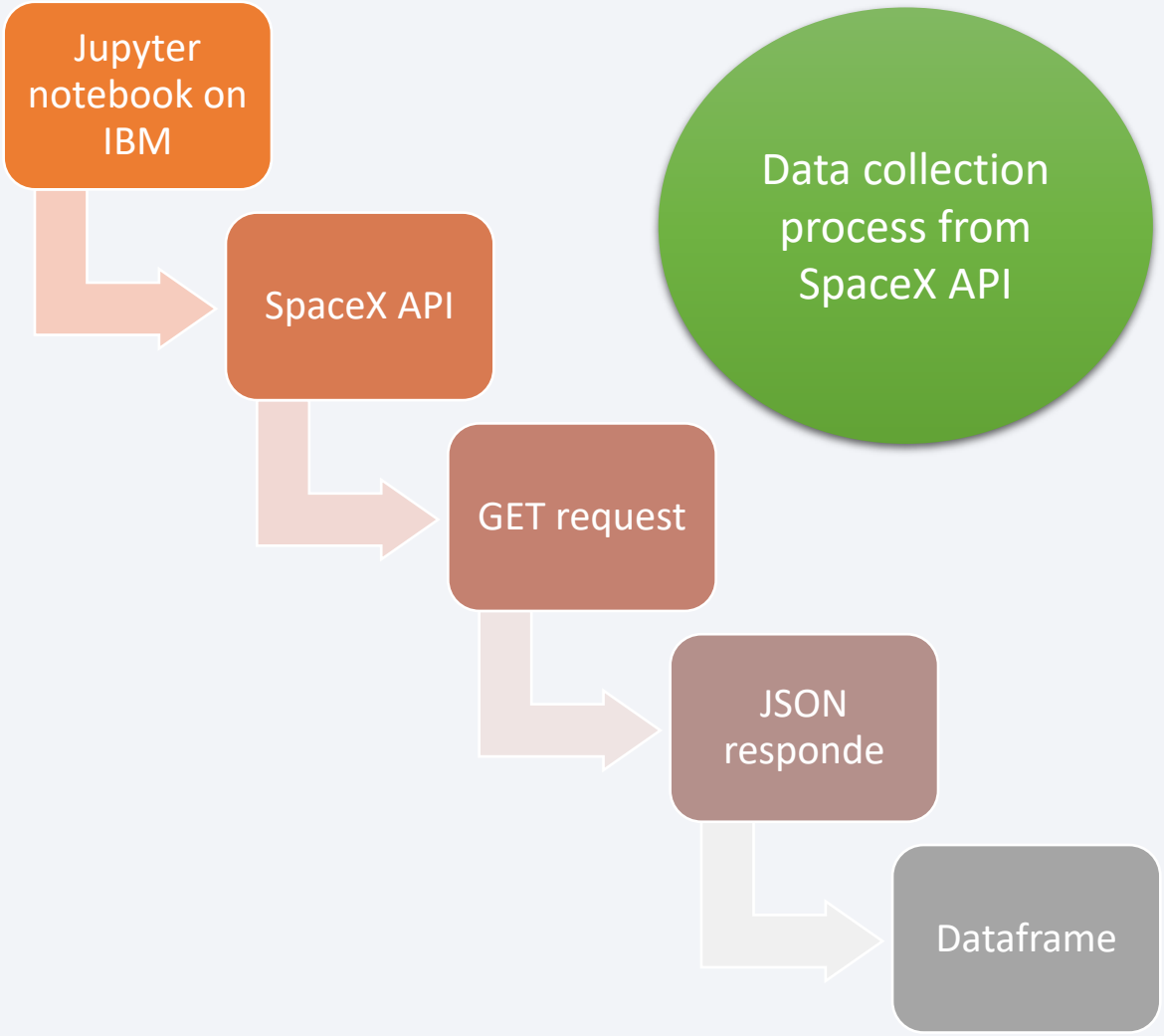
Data Collection

- The data was collected from 2 main sources:
  - SpaceX API and Wikipedia. The first one is an open source REST API for launch, rocket, core, starlink, launchpad and landing pad data. Meanwhile, Wikipedia is a known free online encyclopedia, created and edited by many people as volunteers, all these people around the world and hosted also by the Wikimedia Foundation
- The process…



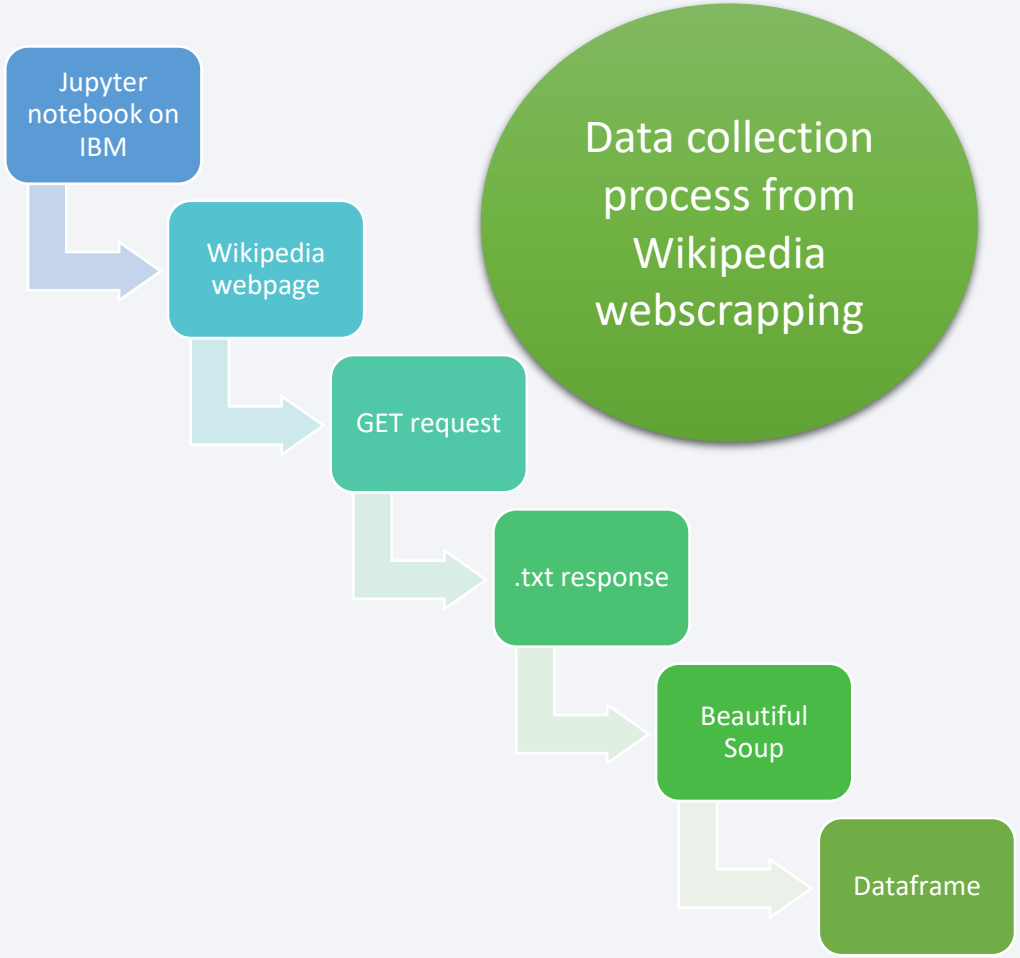
Data Collection – SpaceX API

- I began collecting data from the SpaceX API by importing essential libraries like pandas, numpy, and requests. After setting up a GET request to the API endpoint, the response was received in JSON format and subsequently converted into a pandas DataFrame for analysis.
- GitHub URL of the completed SpaceX API calls notebook click here



Data Collection - Scraping

- To collect additional data on Falcon 9 landings, I performed web scraping on a specific Wikipedia page using the requests and BeautifulSoup libraries. First, an HTTP GET request was sent to retrieve the webpage content in plain text format. This raw HTML content was then parsed using BeautifulSoup to navigate and extract relevant tables containing information about launch dates, payloads, and landing outcomes. Once the desired table elements were identified, the data was systematically extracted and cleaned. Finally, the structured data was transformed into a pandas DataFrame, making it ready for further analysis and integration with other datasets.



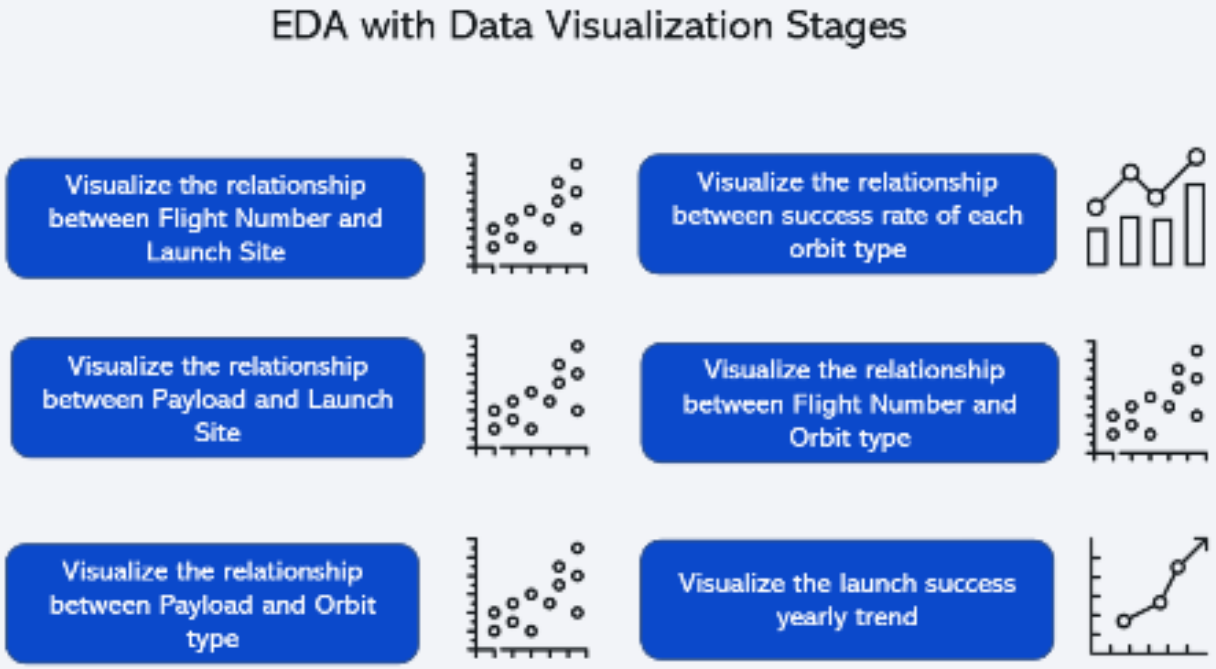
Data Wrangling

At this stage, we imported the previously collected datasets into our environment using pandas and numpy. The goal was to inspect the DataFrame and begin preparing it for machine learning. We started by cleaning the data — handling missing values, removing irrelevant features, and ensuring consistency — then selected the most relevant variables that could serve as strong predictors for training a machine learning model.

Loaded and combined datasets	Assessed missing values and handled nulls	Distinguisehd categorical vs numerical columns	Engineered new features (a.e landing success label).	Counted launches per site, orbits used and created custom outcome fields

EDA with Data Visualization

- During this phase, we finalized our Exploratory Data Analysis (EDA) by exploring the relationships between input features and the target variable using various visualization tools, including Seaborn and Matplotlib. Additionally, we carried out feature engineering by transforming categorical variables into numerical representations through one-hot encoding, enabling their use in machine learning models.



EDA with SQL

- Using SQL we explored:

SQL Quires

- 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).
- List the date when the first successful landing outcome in ground pad was achieved.
- 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 failed landing outcomes in drone ship, their booster versions, and launch site names for the 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.

Build an Interactive Map with Folium

- In this stage we used folium to represent our dataframe.
- We started our interactive map by drawing 4 circles on 4 different sites, those belongs to Falcon 9 launches and have this information in there
- We put markers on the same sites to represent right and failed first stage rockets return using markes objects
- And finally, calculated the distances between the launch site to its proximities
  - Closest city
  - Coastline
  - Highway
- Then we drew polylines to represent them

LAUNCH SITE	LAT	LONG
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	-120.610746

Build a Dashboard with Plotly Dash

Building an Interactive Dashboard with Plotly Dash

1- We added a dropdown list to enable Launch Site selection including the following options:

All Sites, CCAFS LC-40, CCAFS SLC-40, VAFB SLC-4E, KSC LC-39A

2- we added a pie chart to show the total successful launches count for all sites

3- we added a slider to select payload which ranges from 0 -10000

4- finally we added a scatter chart to show the correlation between payload and launch success

GitHub URL of the completed Building an Interactive Dashboard with Plotly Dash notebook, Click Here

Predictive Analysis (Classification)

Machine Learning Stages:

1- Importing the required libraries.

2- Loading the cleaned data.

3- Standardizing the data to prevent the bias.

4- splitting the data into 20% for testing data and 80% training data.

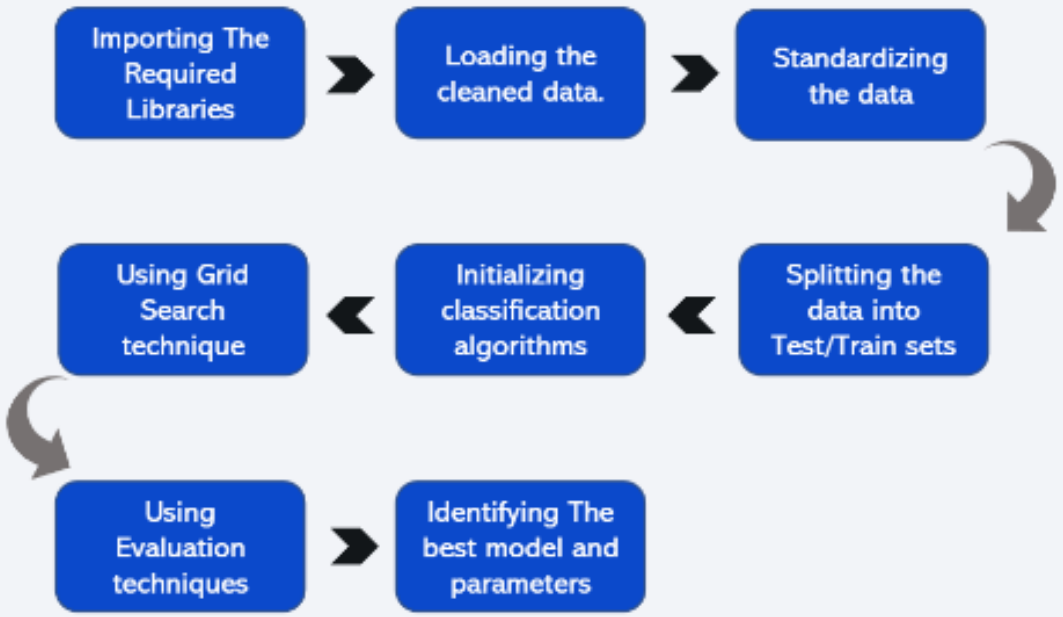
5- Initializing 4 different classification algorithms:

- Logistic Regression (LR)
- Support Vector Machine (SVM)
- Decision Tree (DT)
- K nearest neighbors (KNN)

6- Using Grid Search technique to find the best parameters

7- Using Evaluation techniques including, Confusion matrix , F1 score, Jaccard Score for the purpose of using the best

Machine Learning Pipelines

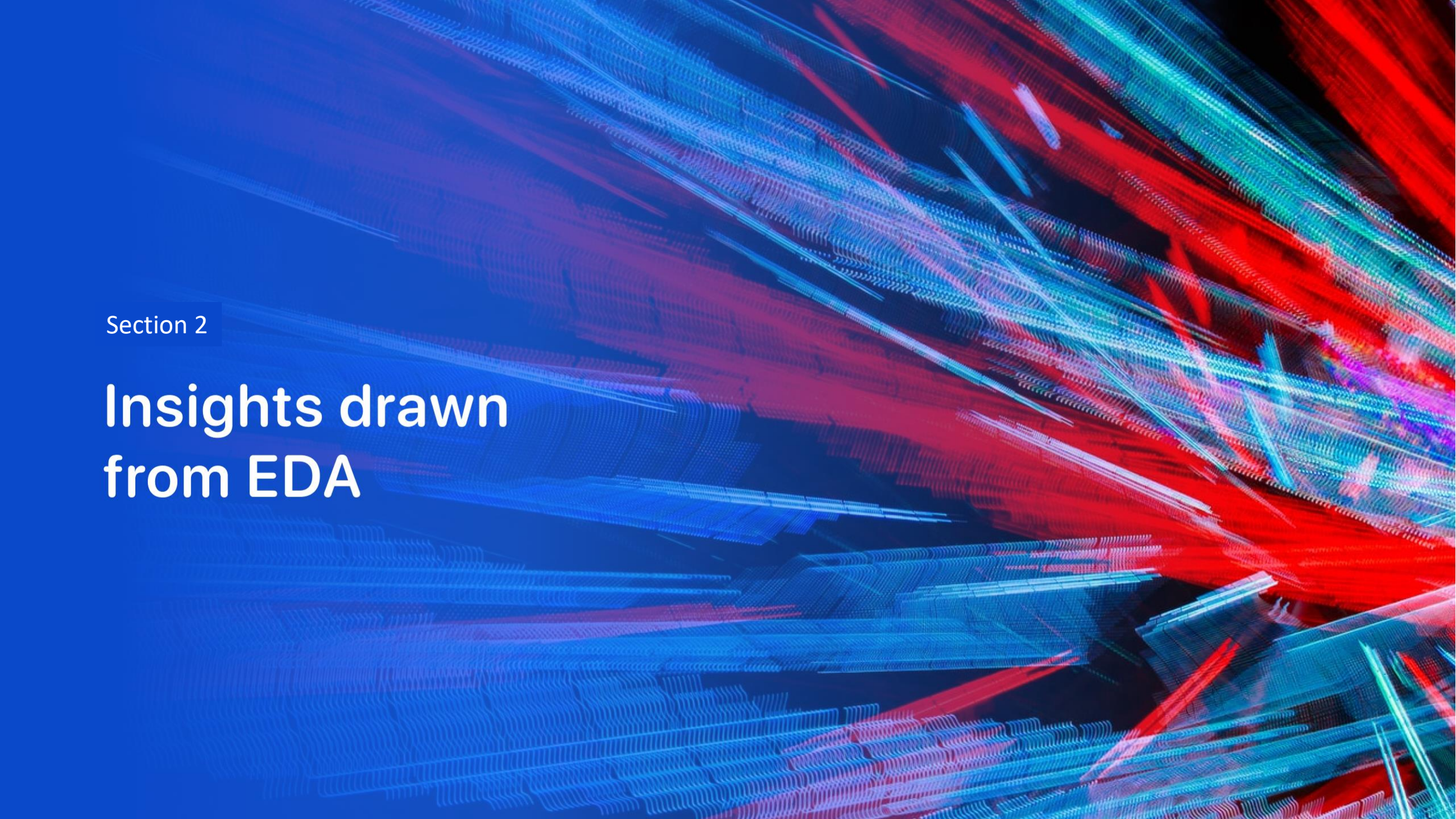


Results

EXPLORATORY DATA ANALYSIS RESULTS

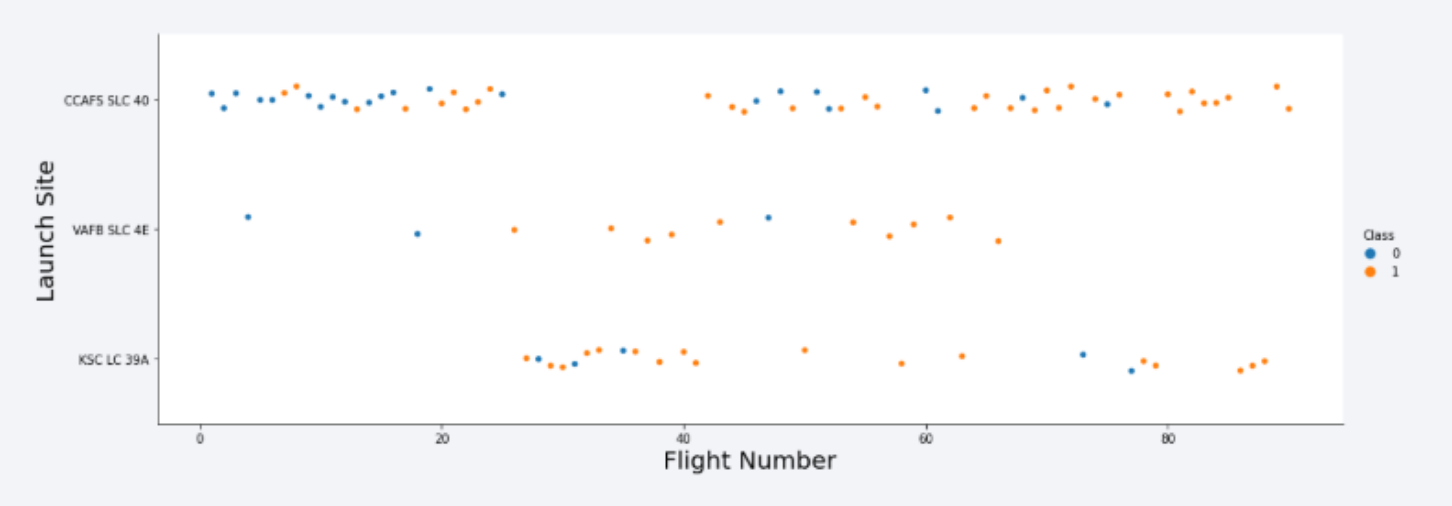
INTERACTIVE ANALYTICS DEMO IN SCREENSHOTS

PREDICTIVE ANALYSIS RESULTS


Section 2
Insights drawn from EDA

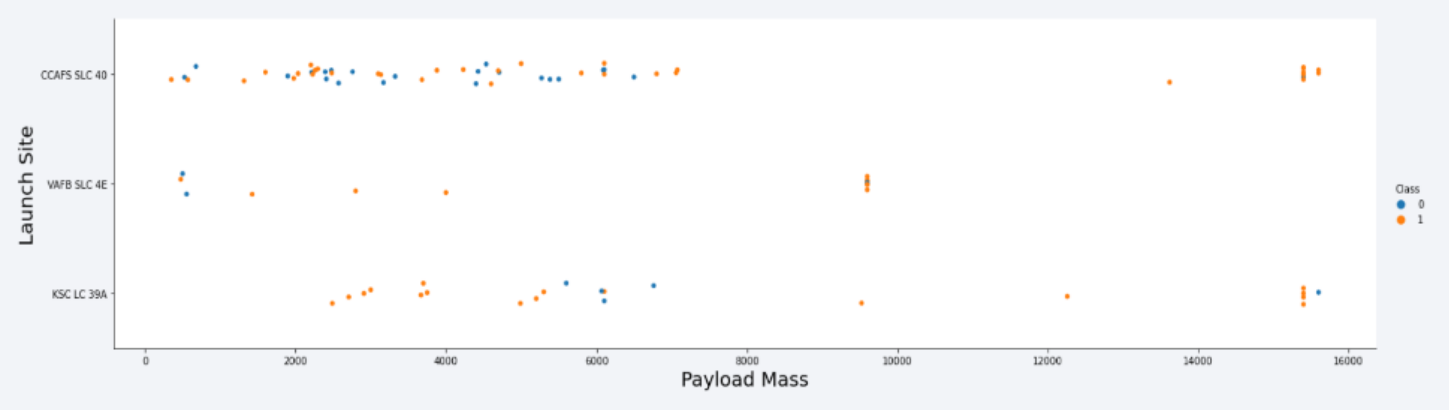
Flight Number vs. Launch Site

- 1.- ccafs slc40: the most usable site for launching spaceX rockets, with 55 trials, 33 of them successful, 22 failed → 60% success rate
- 2.- vafb slc4E: least usable site, 13 trials: 10 of them successful, 3 failed → 77% success rate
- 3.- vafb slc4E: moderate site, 22 trials, 17 successful and 5 failed → 77% success rate.



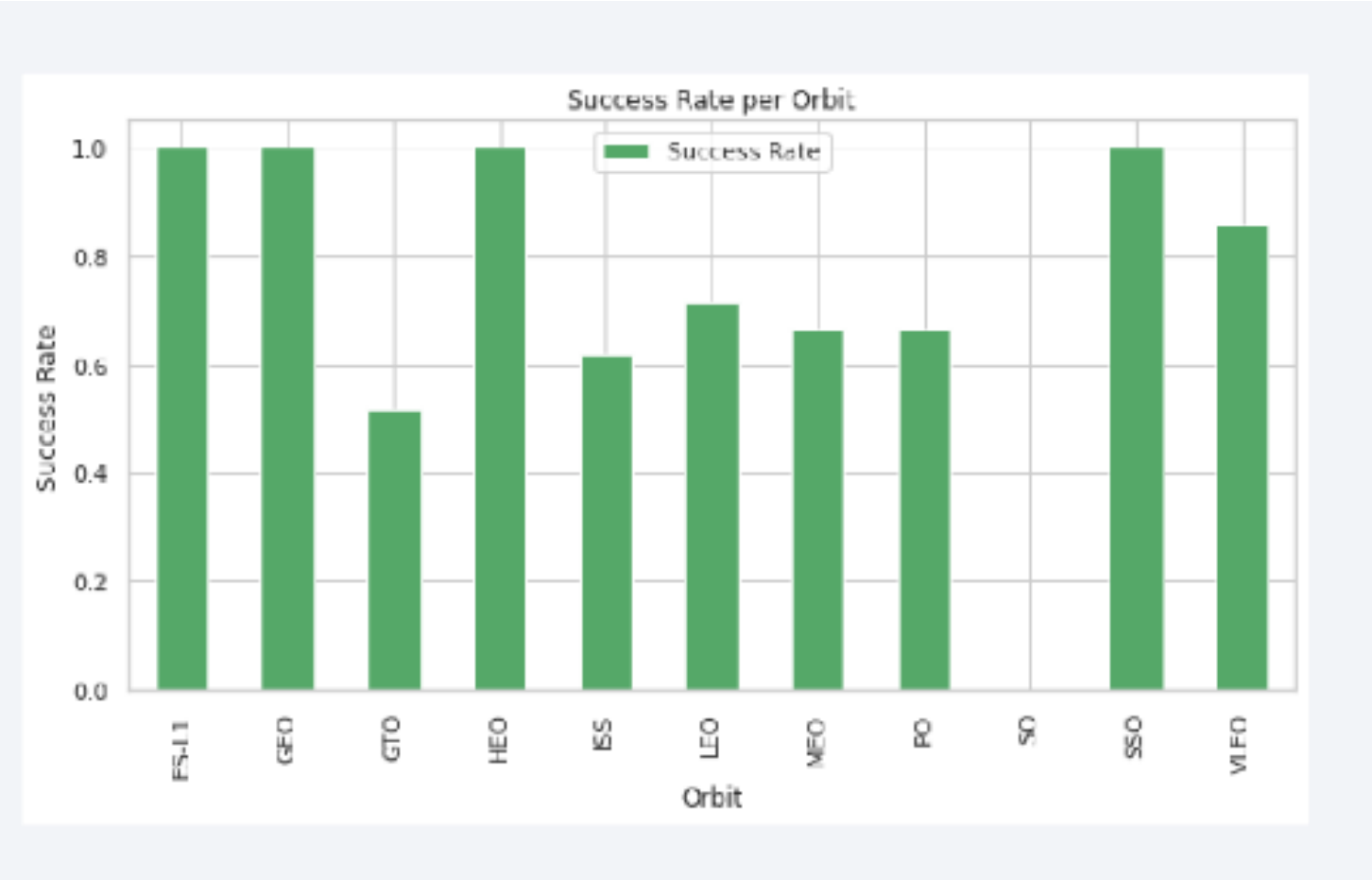
Payload vs. Launch Site

- Plot explanation: There’s no strong relationship between payload mass vs success first stage return.

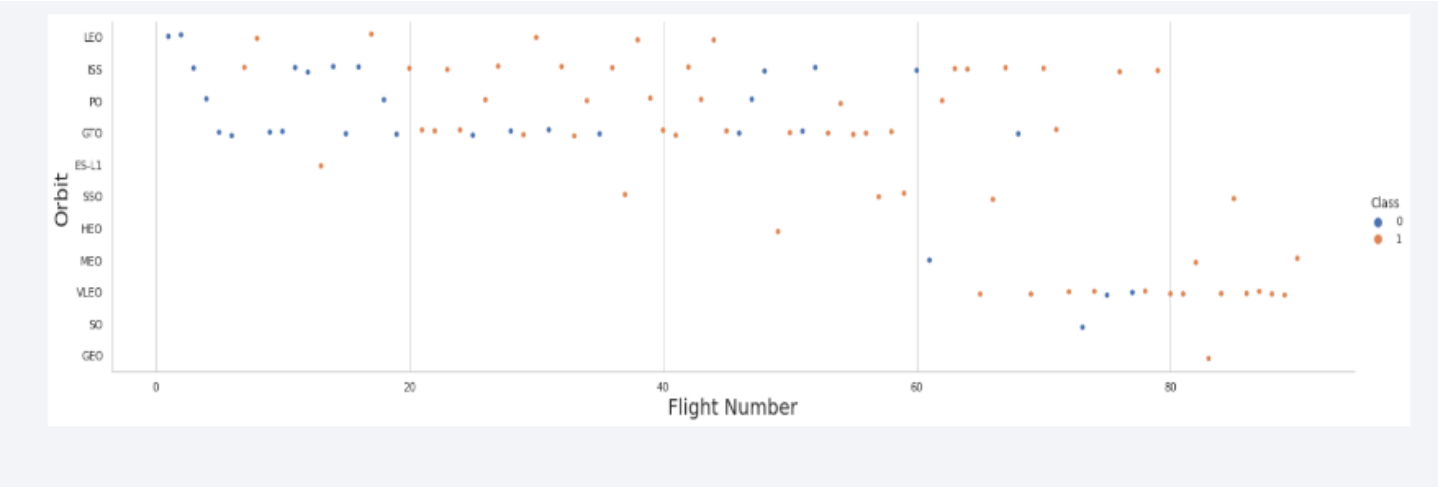


Success Rate vs. Orbit Type

- Bar plot explanation:
- Best orbits are ES.L1, HEO, GEO and SSO in terms of successful 1st stage returns.
- Worst orbit: GTO

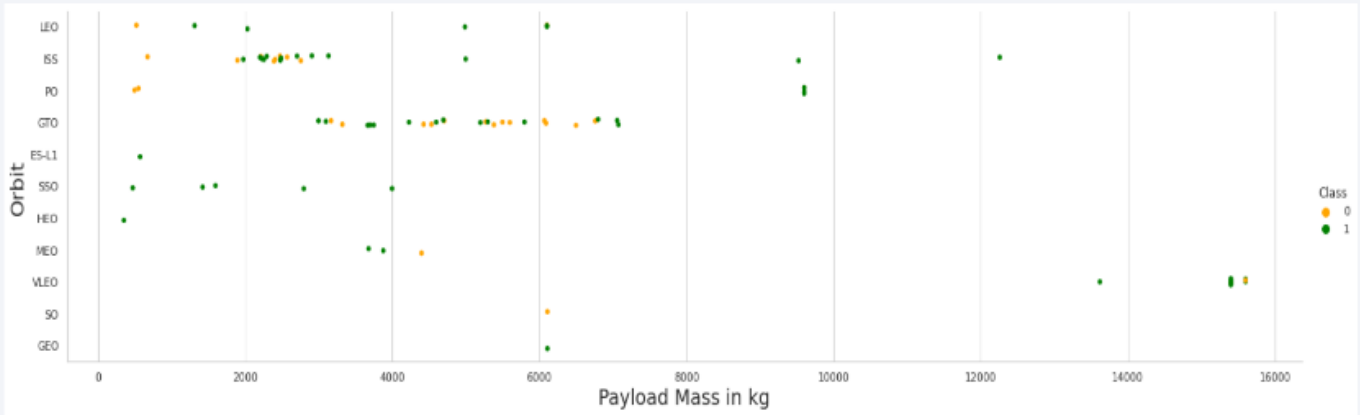


Flight Number vs. Orbit Type



In LEO orbit, the success is related to the number of flights

Payload vs. Orbit Type



We observe that Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.

Launch Success Yearly Trend

The success rate since 2013 kept increasing until the end of 2020.

All Launch Site Names

Display the names of the unique launch sites in the space mission

```
%sql select distinct launch_site from SPACEXTBL
```

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

As shown above we have 4 distinct sites for rockets launches listed in the Table above.

Launch Site Names Begin with 'CCA'

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

```
%sql select * from SPACEXTBL where launch_site like 'CCA%' limit 5;
```

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcom
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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

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

```
%sql select sum(payload_mass__kg_) from SPACEXTBL where customer = 'NASA (CRS)';
```

1
45596

The total amount of payload that moved to the outer space by NASA through SpaceX rockets equals to 45596 Kg which is equal to 50.261 Us ton.

Average Payload Mass by F9 v1.1

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

```
%sql select avg(payload_mass__kg_) as avg_mass_F9 from SPACEXTBL where booster_version = 'F9 v1.1'
```

avg_mass_f9
2928

The average payload mass carried by booster version F9 v1.1 is 2928 kg

First Successful Ground Landing Date

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

Hint:Use min function

```
%sql select min(DATE) from SPACEXTBL where landing__outcome = 'Success (ground pad)'
```

1
2015-12-22

Date of the first successful landing outcome on ground pad was in 22-12-2015

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 SPACEXTBL\
where (landing__outcome = 'Success (drone ship)' and (payload_mass__kg_ > 4000 and  payload_mass__kg_ > 6000));
```

booster_version
F9 FT B1029.1
F9 FT B1036.1
F9 B4 B1041.1

The boosters which have success in drone ship landing with payload between 4000 and 6000 kg are :

- F9 FT B1029.1
- F9 FT B1036.1
- F9 B4 B1041.1

Total Number of Successful and Failure Mission Outcomes

- We clearly see the success rate mission outcomes is the most dominant we have, only 1 mission were failed while we have 99 successful ones.

List the total number of successful and failure mission outcomes

```
%sql select mission_outcome, count(mission_outcome) as counts from SPACEXTBL GROUP BY mission_outcome
```

mission_outcome	counts
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

The booster versions who carry the maximum payload start w/F9 B5 and ranges from B1048 to B1060.

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%sql select distinct booster_version from SPACEXTBL\
where payload_mass__kg_ in (select max(payload_mass__kg_) from SPACEXTBL);
```

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

2015 Launch Records

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%sql select landing__outcome, booster_version, launch_site from SPACEXTBL\
where (landing__outcome = 'Failure (drone ship)' and date like '2015%')
```

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

We have two failed landing in 2015 on a drone ship which both in the same site, CCAFS LC-40 and with same booster version F9 v1.1

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

```
%sql select landing__outcome, count(*) as counts_of_landing_outcomes from SPACEXTBL\
where DATE between '2010-06-04' and  '2017-03-20' group by landing__outcome\
order by count(landing__outcome) desc
```

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

Section 3

Launch Sites Proximities Analysis

Folium Map: Chosen Launch sites

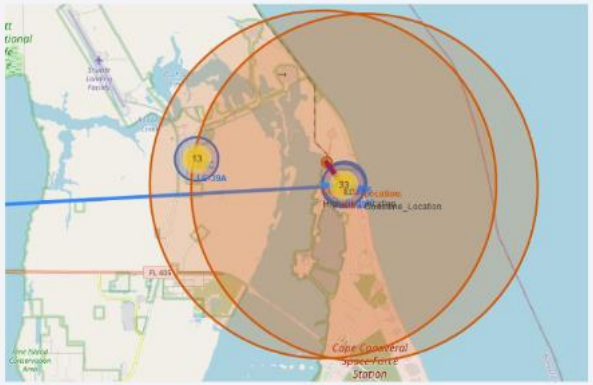
- All locations are near coast and Ecuador line
- SpaceX focus on locations that are close to water for the purpose of avoiding accidents.
- Launch sites are distributed in 2 states: Florida, California

Folium Map: The success rate for each location

- Green marker → successful return
- Red marker → failed return

- We can easily identify which launch sites have high success rates.

Folium Map: Closest Proximities to CCAFS LC-40

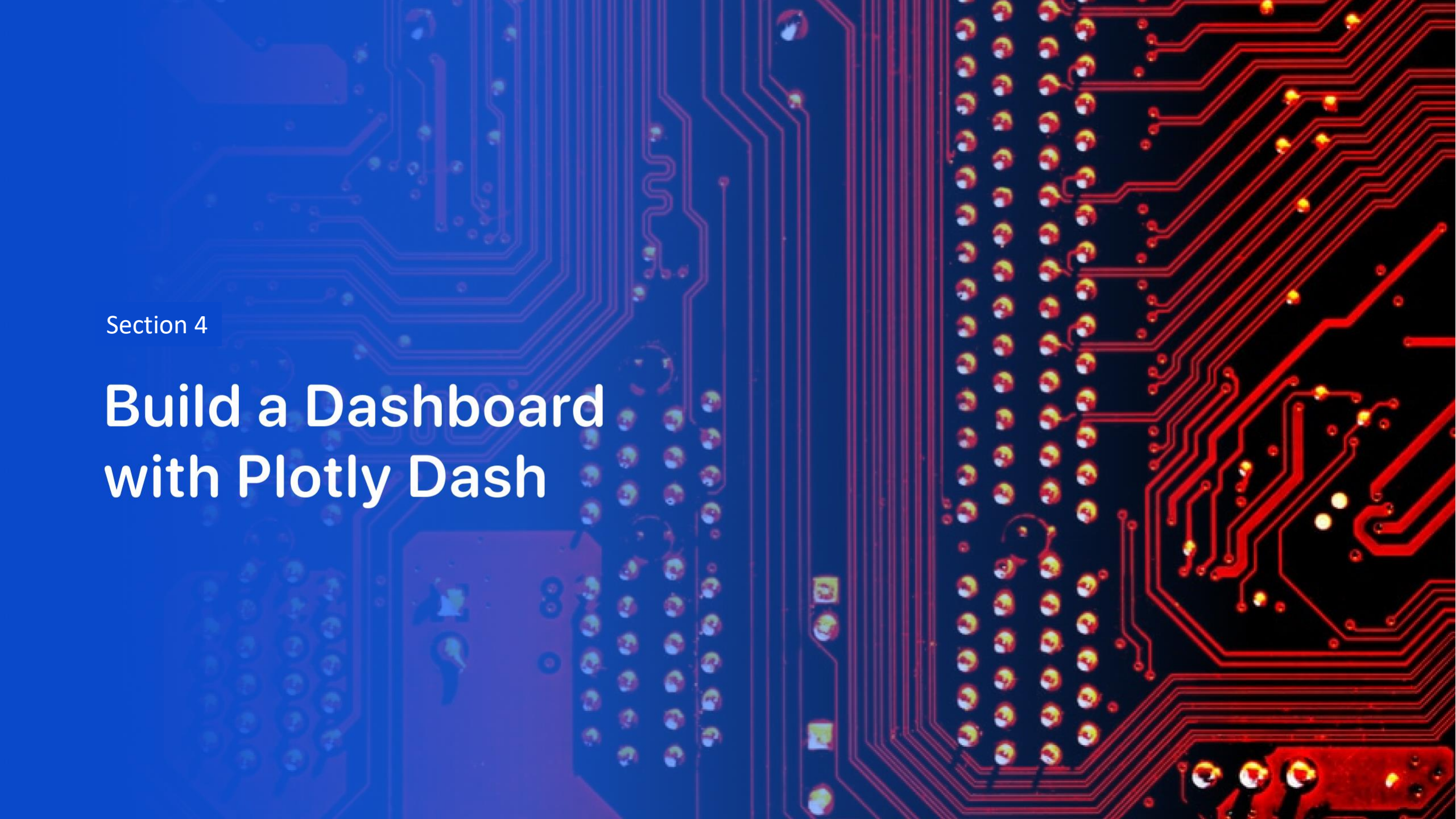


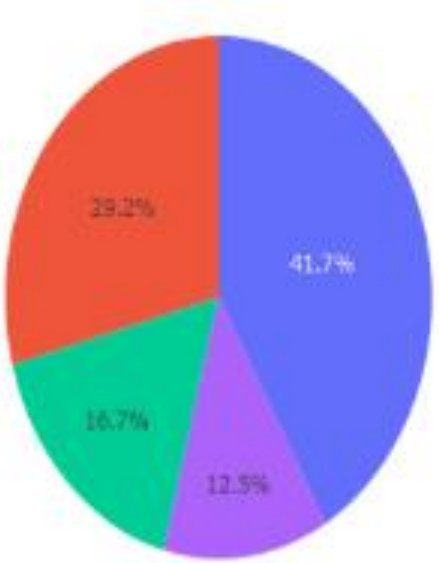
Proximities Cordinates

	Location	Lat	Long
0	Orlando_Location	28.52300	-81.38260
1	Coastline_Location	28.56146	-80.56746
2	Highway_Location	28.56270	-80.58703

we calculated the distances between the launch site (CCAFS LC-40) to its proximities

Orlando City Distance ≈ 78.8 Km,
Coastline Distance ≈ 0.97 Km,
Highway Distance ≈ 0.95Km

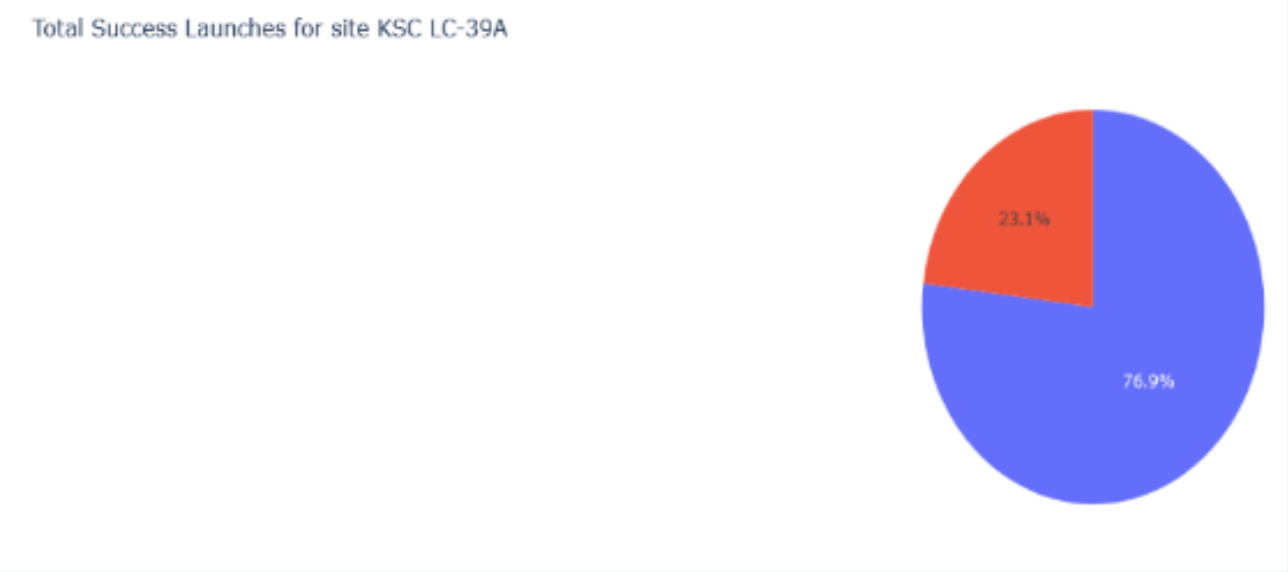

Section 4
Build a Dashboard with Plotly Dash



Succeded Launches for all sites.

- It shows the success % for every site in terms of 1st return
  - Best site: KSC LC-39A → 41.7% successful
  - Worst site: CCAFS SLC-40 → 12.5% success rate.

For KSC LC-39A



- Total succeded launches
  - 76.9% successful missions
  - 23.1% failed missions

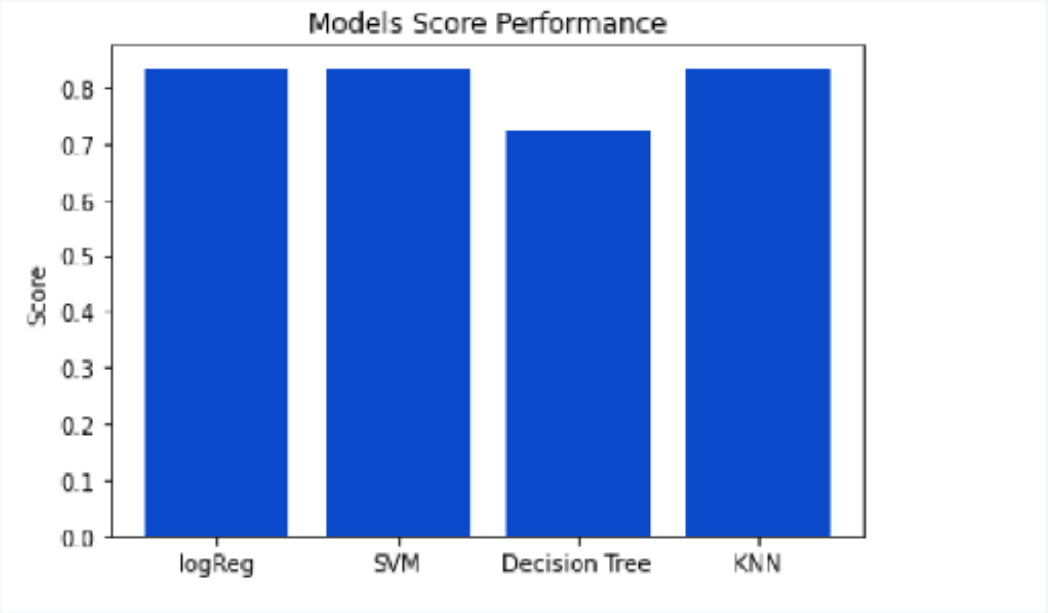
Scatter plot: Launch outcome based on payload mass

- We infer that the success rate per booster version for example, the payload mass <4000 kg are more likely to be successful

Section 5

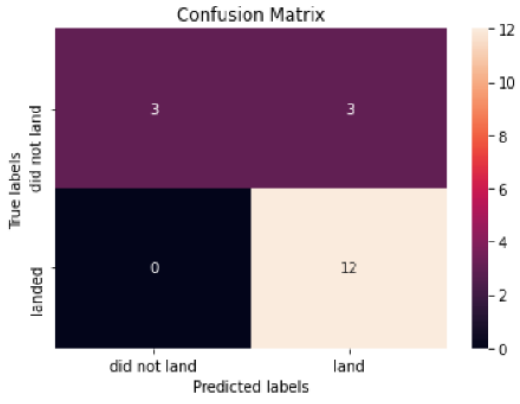
Predictive Analysis (Classification)

Classification Accuracy



Logistic Regression , SVM and KNN has the same performance where the Jaccard Score is same : 0.8
Where Decision Tree has the worst performance compared to other models.

Confusion Matrix




```
Logistic Regression:
Jaccard Score of = 0.8
F1 Score = 0.7777777777777778
====================================
SVM:
Jaccard Score of = 0.8
F1 Score = 0.7777777777777778
====================================
Decision Tree:
Jaccard Score of = 0.6666666666666666
F1 Score = 0.6727272727272727
====================================
KNN:
Jaccard Score of = 0.8
F1 Score = 0.7777777777777778
====================================
```

Confusion Matrix

Logistic Regression , SVM and KNN have the same confusion matrix and results:

- True Positive = 12
- False Positive = 0
- True Negative = 3
- False Negative = 3

Conclusions

Reusable first-stage returns significantly reduce costs.

A range of features affect landing success.

Launch sites are selected strategically for geography and logistics.

Orbit and booster version are crucial to predicting outcomes.

SpaceX's success is increasing steadily — understanding their data gives new players like SpaceY a strategic advantage.

Thank you!