IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Franco Kees
21 July 2023

Outline

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

Executive Summary

- The aim was to investigate the factors that affected the success of SpaceX missions, such as the rocket design, payload, launch site and landing method
- SpaceX mission data was obtained from the SpaceX public API and from SpaceX launch data on Wikipedia
- Various machine learning classification models were trained and tested to find the most accurate predictor of launch success
- Github repository

Introduction

- We will predict if the Falcon 9 first stage will land successfully.
- Therefore if we can determine if the first stage will land, we can determine the cost of a launch.

Section 1

Methodology

Methodology

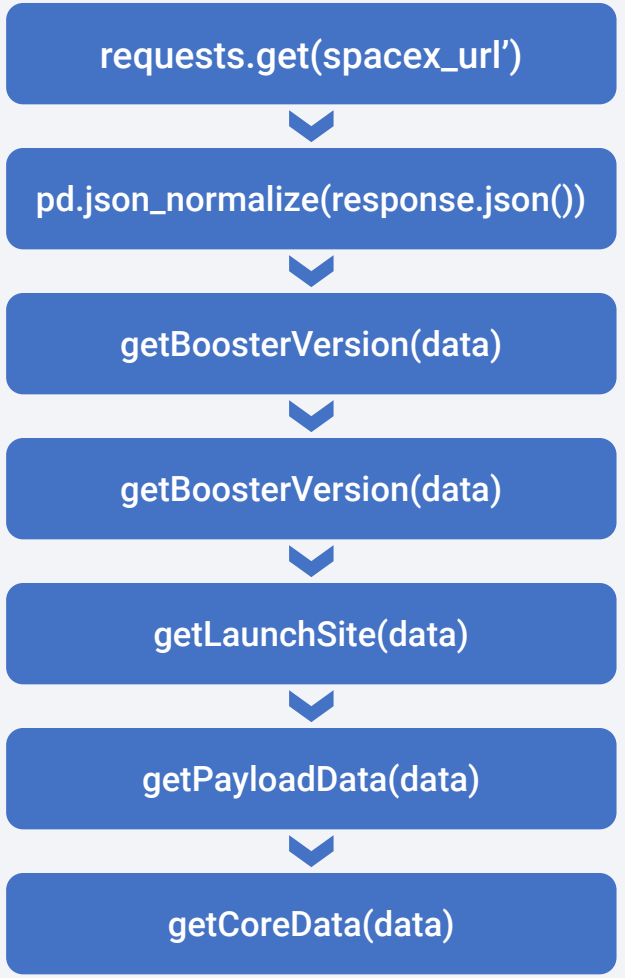
Executive Summary

- Data collection methodology:
  - SpaceX API and web scraping Wikipedia
- Perform data wrangling
  - Dealt with missing values and created landing category variable
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Grid search to train, test and tune classification models

Data Collection – SpaceX API

SpaceX REST API requests are:

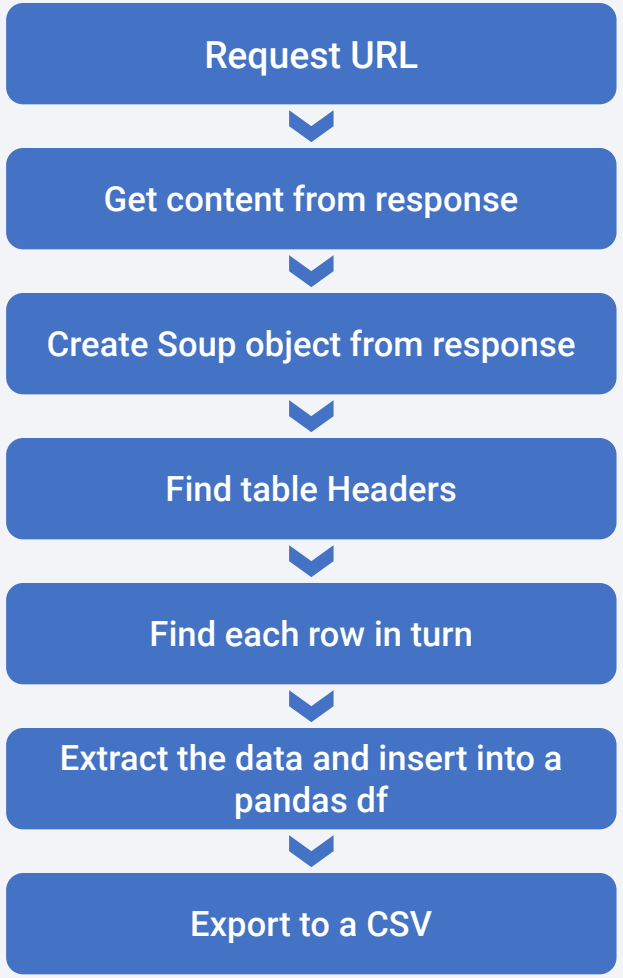
- Past launch data is requested from: https://api.spacexdata.com/v4/launches/past
- Booster data is gained from a rocket type request: https://api.spacexdata.com/v4/rockets/
- Launchpad name, longitude and latitude comes from a launch site request: https://api.spacexdata.com/v4/launchpads/
- Payload data comes from a payloads request: https://api.spacexdata.com/v4/payloads/
- Specific rocket core information comes from a core request: https://api.spacexdata.com/v4/cores/
- Falcon 9 data was exported to a CSV file



GITHUB: DataScience/Lab_1/DataCollection.ipynb at main (github.com)

Data Collection - Scraping

- Used the requests library to scrape data from https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- Used BeautifulSoup to parse the content returned in the response
- The parsed data was exported to a CSV file

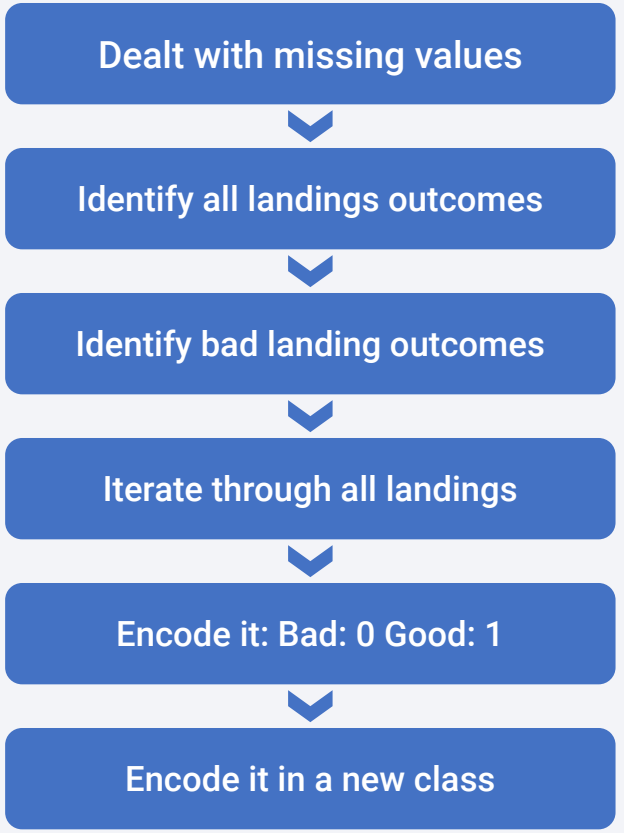


GITHUB: DataScience/Lab_1/DataWebscraping.ipynb at main

Data Wrangling

- A classification variable ("class") was created to encode the landing outcome as either 0 (bad) or 1 (good). The variable is the target variable that the classification algorithm will need to predict.

GITHUB: [DataScience/Lab_1/DataWrangling.ipynb at main]



EDA with Data Visualization

- Conducted EDA with Matplotlib and Feature Engineering with Pandas.
- Explored Flight number, Launch site, Payload mass, and orbit type relationships using scatter plots, bar graphs, and line graphs.
- Color-coded visualizations by "class" for launch outcome visibility.
- Employed one-hot-encoding and cast data to float64 for analysis.
- Gained valuable insights for informed decisions.

GITHUB: DataScience/Lab_2/edaDataViz.ipynb at main

EDA with SQL

- SQL queries:

  - SELECT DISTINCT Launch_Site from SPACEXTBL;
  - SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
  - SELECT SUM(PAYLOAD_MASS__KG_) AS Total_Payload_Mass_Kg FROM SPACEXTBL WHERE Customer = 'NASA (CRS)';
  - SELECT AVG(PAYLOAD_MASS__KG_) AS Total_Payload_Mass_Kg FROM SPACEXTBL WHERE Booster_Version = 'F9 v1.1';
  - SELECT MIN(Date) AS First_Successful_Landing_Date FROM SPACEXTBL WHERE Landing_Outcome ="Success (ground pad)";
  - SELECT DISTINCT Booster_Version AS Bts FROM SPACEXTBL WHERE Landing_Outcome = "Success (drone ship)" AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;
  - SELECT Mission_Outcome, COUNT(*) AS Total_Count FROM SPACEXTBL GROUP BY Mission_Outcome;
  - SELECT DISTINCT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
  - SELECT CASE WHEN substr(Date, 4, 2) = '01' THEN 'January'WHEN substr(Date, 4, 2) = '02' THEN 'February'WHEN substr(Date, 4, 2) = '03' THEN 'March'WHEN substr(Date, 4, 2) = '04' THEN 'April'WHEN substr(Date, 4, 2) = '05' THEN 'May'WHEN substr(Date, 4, 2) = '06' THEN 'June'WHEN substr(Date, 4, 2) = '07' THEN 'July'WHEN substr(Date, 4, 2) = '08' THEN 'August'WHEN substr(Date, 4, 2) = '09' THEN 'September'WHEN substr(Date, 4, 2) = '10' THEN 'October'WHEN substr(Date, 4, 2) = '11' THEN 'November' WHEN substr(Date, 4, 2) = '12' THEN 'December' END AS Month, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTBL WHERE substr(Date, 7, 4) = '2015' AND Landing_Outcome = 'Failure (drone ship)';

  - **GITHUB**: DataScience/Lab_2/edaSQL.ipynb at main · etziok/DataScience (github.com)

Build an Interactive Map with Folium

- Circles with Markers were added to indicate launch locations
- A MarkerCluster was added for each site to visualize the launch outcomes.
- A line was added to show the proximity of the two (Launch site and coast).
- **GITHUB:** DataScience/Lab_3/foliumMap.ipynb at main

Build a Dashboard with Plotly Dash

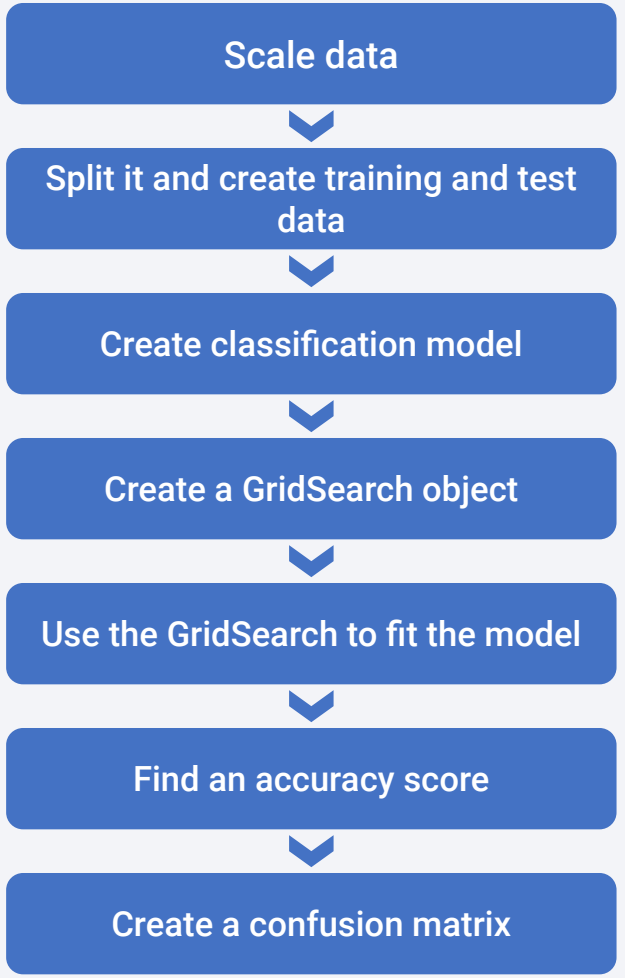
- An interactive dashboard was created to explore the impact of launch site, payload mass, and booster type on launch outcomes (success or failure).
- Launch site could be selected from a dropdown menu, and payload mass could be adjusted using a slider control.
- A pie chart displayed either the overall success rate for all launch sites or the distribution of successful and failed launches for a specific site.
- A scatter chart illustrated how launch outcomes varied based on the selected site and payload range, with data points color-coded by booster type.
- **GITHUB:** DataScience/Lab_3/spacex_dash_app.py at main

Predictive Analysis (Classification)

We use:

- Logistic regression
- Support vector machines
- Decision tree
- K-nearest neighbors

- **GITHUB:** DataScience/Lab_4/machineLearningPrediction.jupyterlite.ipynb

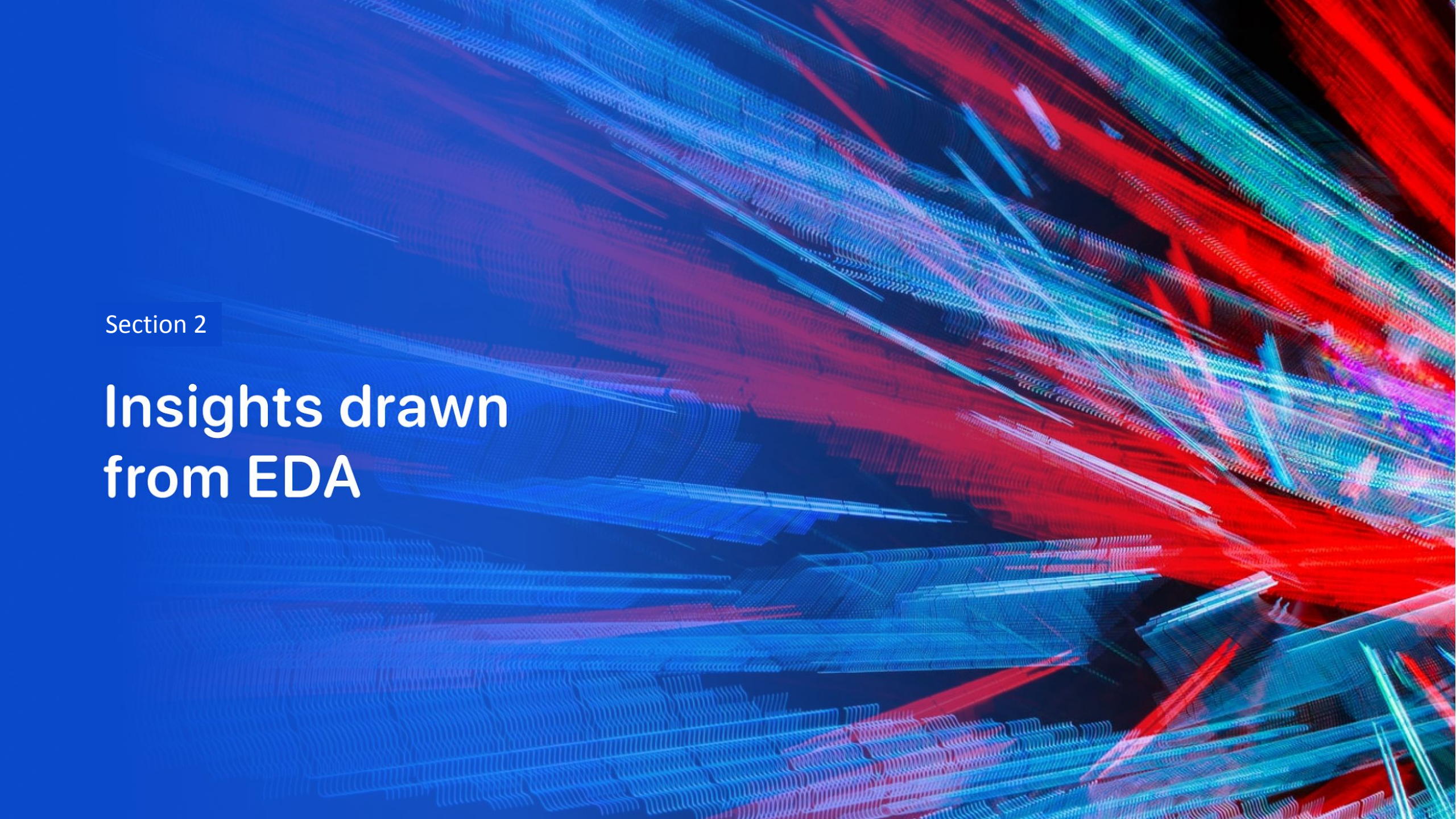


Results

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

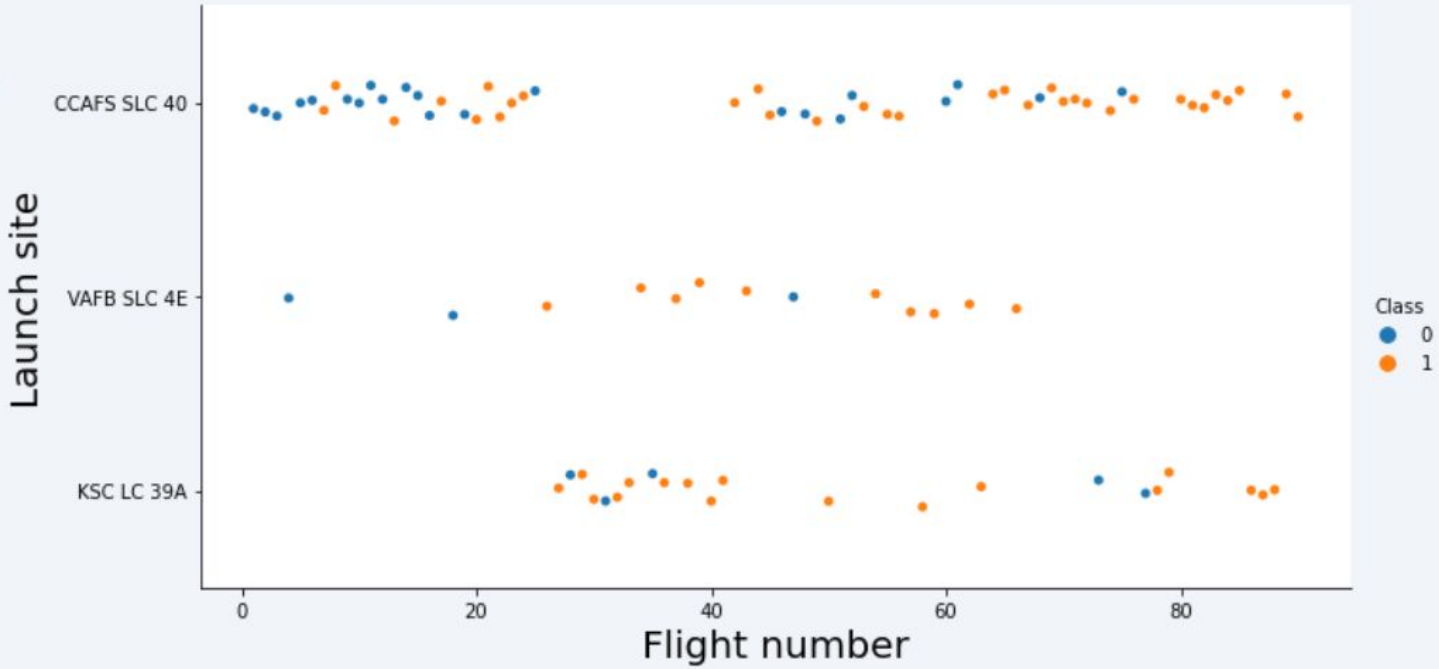
Section 2

Insights drawn from EDA



Flight Number vs. Launch Site

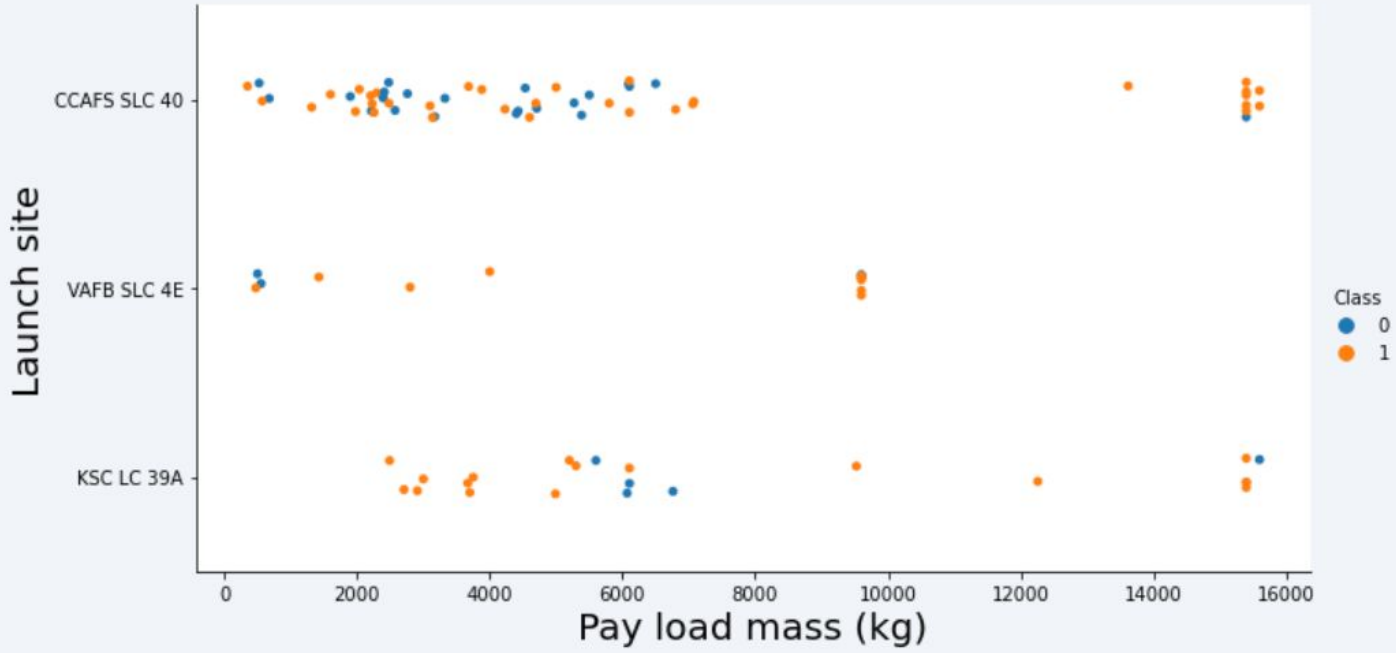
- Scatter plot “Flight Number vs. Launch Site” (colored by outcome: blue is bad and orange is good)
- CCAFS conducted the most tests, while VAFB and KSC boast higher success rates, likely due to their extensive use in later flights, benefiting from improved reliability.



Payload vs. Launch Site

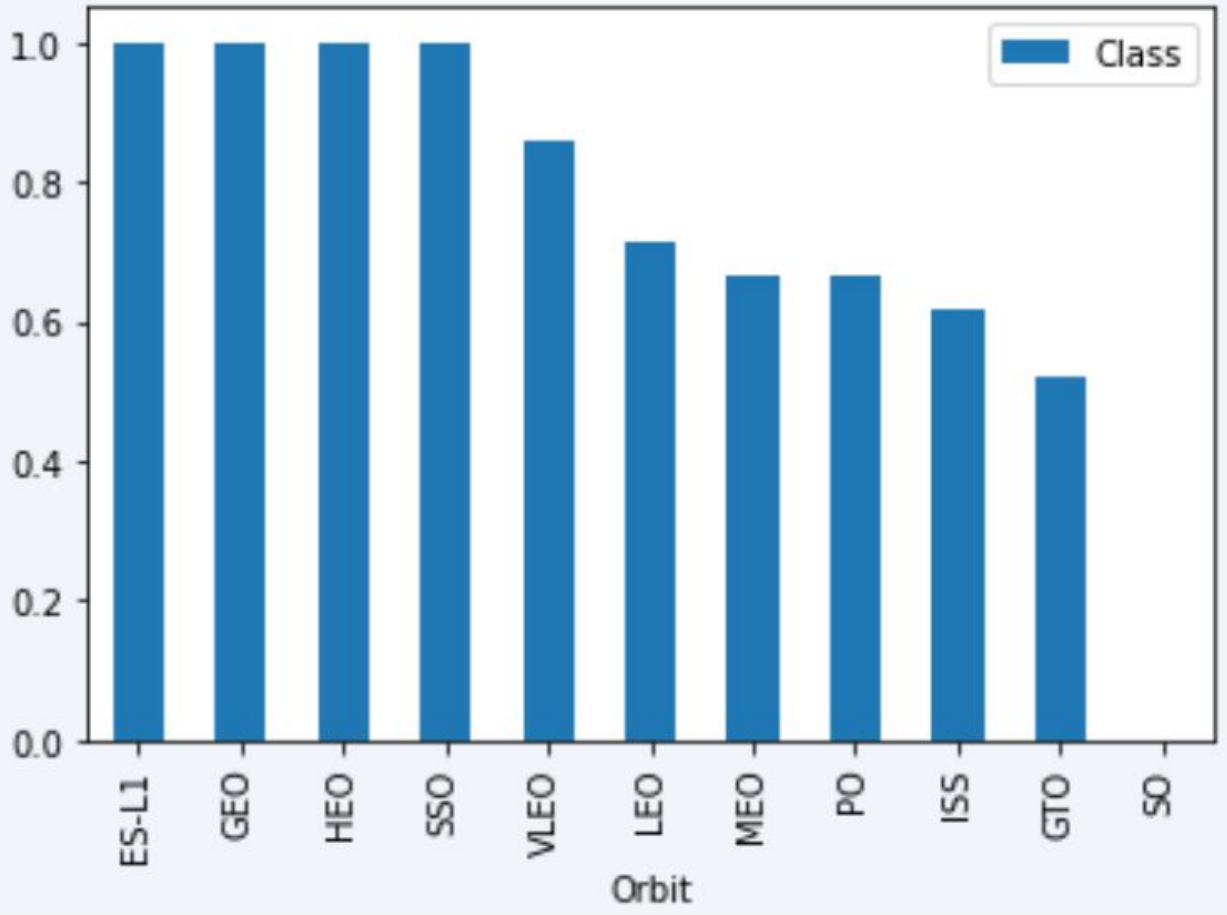
- Scatter plot of ”Payload vs. Launch Site” (blue is bad and orange is good)

- VAFB... has not had many heavy flights.
- Launches with higher payloads were more successful.



Success Rate vs. Orbit Type

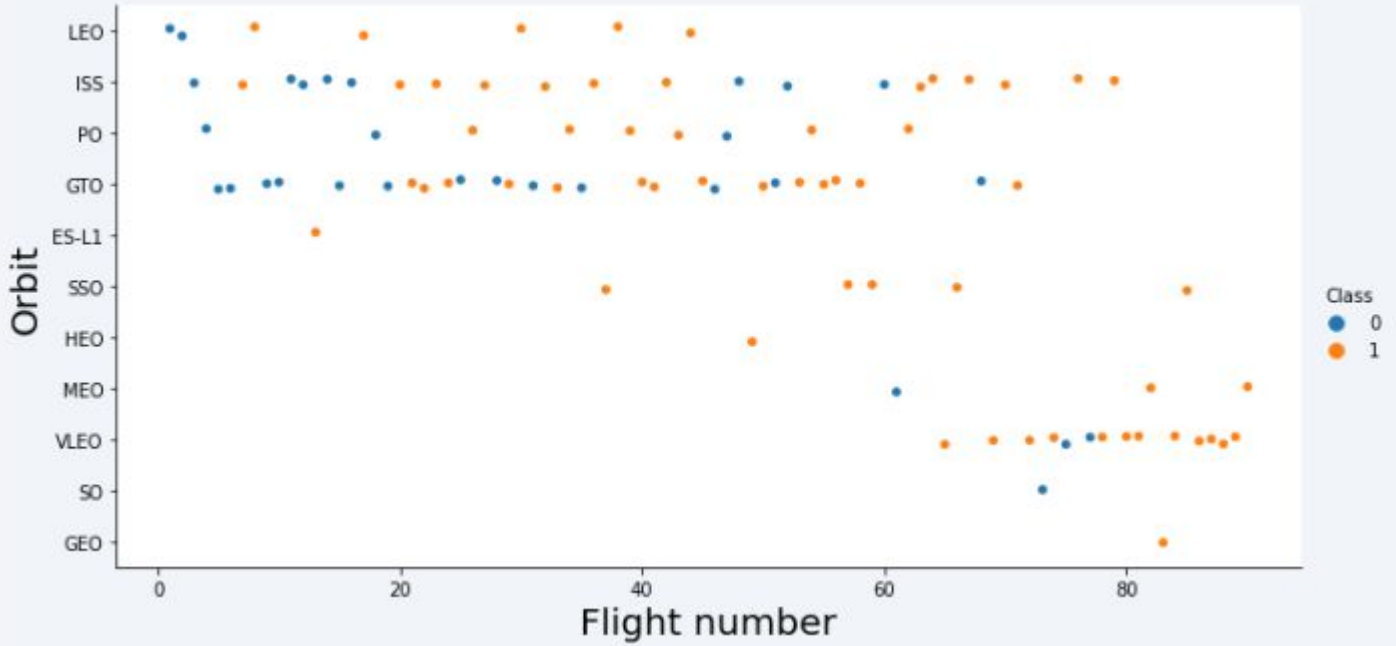
- Bar chart for the success rate of each orbit type
- Higher orbits (GEO, HEO) seem to have higher rates of success.



Flight Number vs. Orbit Type

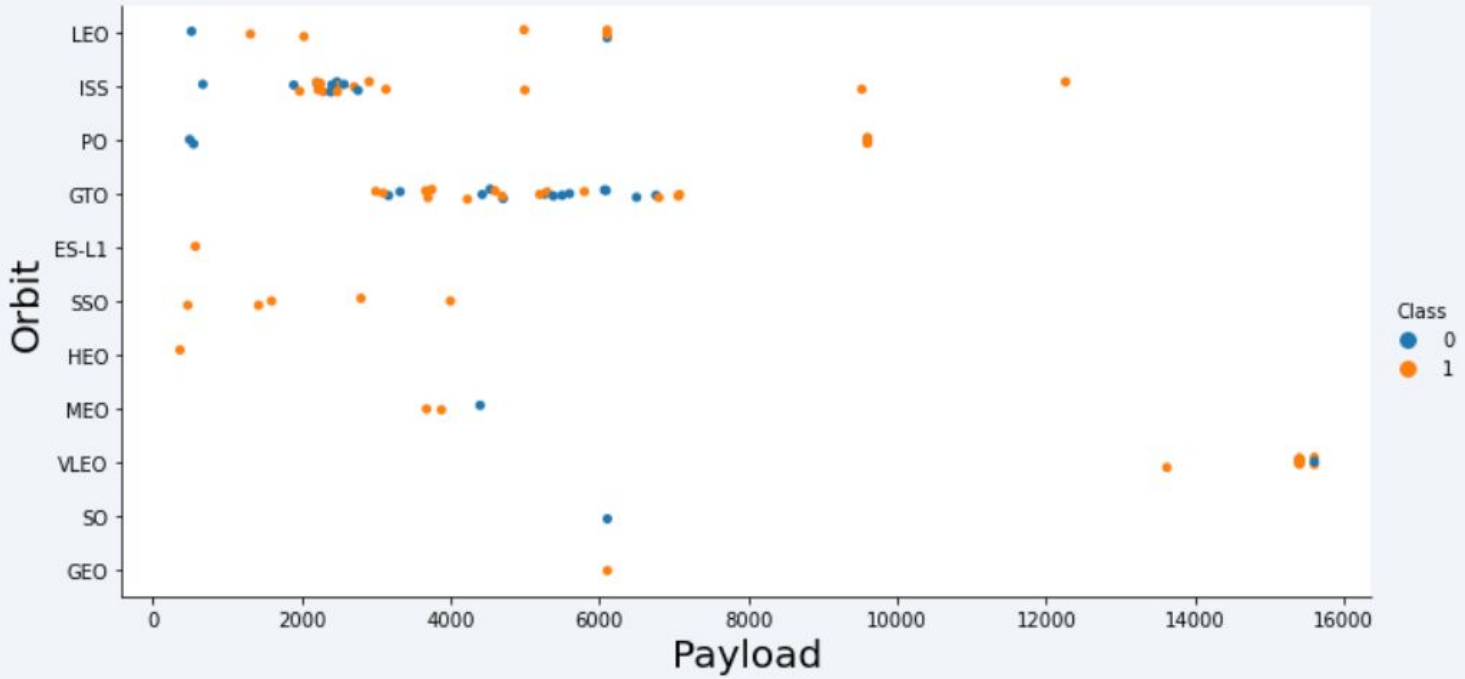
- Scatter point of “Flight number vs. Orbit type”(blue is bad and orange is good)

- The outcomes have tended to improve across all orbits as the number of flights increased.



Payload vs. Orbit Type

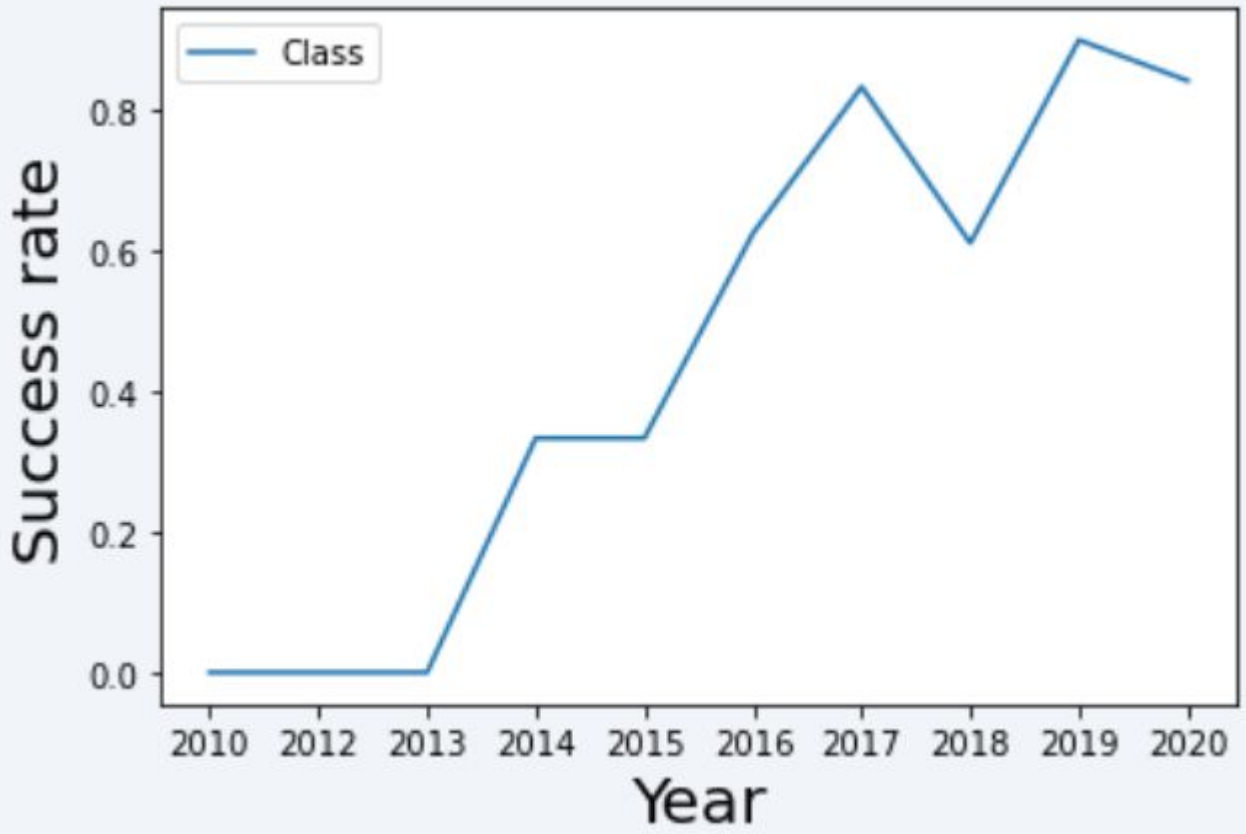
- Scatter point of “Payload vs. Orbit type”(blue is bad and orange is good)
- There are no clear patterns of increased success with increased payload for any given orbit.
- SSO orbit is appears consistently successful



Launch Success Yearly Trend

- Line chart of yearly average success rate

- The success rate has been steadily increasing. 2018 was a exception



All Launch Site Names

Unique launch sites names

- SELECT DISTINCT Launch_Site FROM SPACEXTBL

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

Launch Site Names Begin with 'CCA'

5 records where launch sites begin with `CCA`

- 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 _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

Total payload carried by boosters from NASA

- SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer = 'NASA (CRS)'
- SUM calculates the total payload value for records filtered by Customer using the condition specified in the WHERE command.

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Similar method to the last slide but using AVG and another filter.

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

Find the dates of the first successful landing outcome on ground pad

- SELECT MIN(substr(Date,7,4)||'-'||substr(Date,4,2)||'-'||substr(Date,1,2)) AS FIRST_DATE FROM SPACEXTBL WHERE "Landing_Outcome" = 'Success (ground pad)'
- In SQLLite, dates are stored as strings, not as a separate date type. To correctly order and find the earliest date, you need to use the substr command and specific patterns before using the MIN command.

FIRST_DATE
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

- SELECT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND "Landing _Outcome" = 'Success (drone ship)'
- We use BETWEEN for filtering records

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

Total Number of Successful and Failure Mission Outcomes

Calculate the total number of successful and failure mission outcomes

- SELECT Mission_Outcome, COUNT(*) FROM SPACEXTBL GROUP BY Mission_Outcome
- GROUP BY aggregates the data according to the variable

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 which have carried the maximum payload mass

- SELECT DISTINCT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
- A subquery finds the maximum payload mass [using MAX()] across all records, which then becomes the filter to select all booster versions that have carried that mass.

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

- SELECT substr(Date,7,4) AS YEAR, substr(Date,4,2) AS MONTH, "Landing _Outcome", Booster_Version, Launch_Site FROM SPACEXTBL WHERE substr(Date,7,4)='2015' AND "Landing _Outcome" = 'Failure (drone ship)'
- The AND condition is used to create two filters

YEAR	MONTH	Landing _Outcome	Booster_Version	Launch_Site
2015	01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015	04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

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

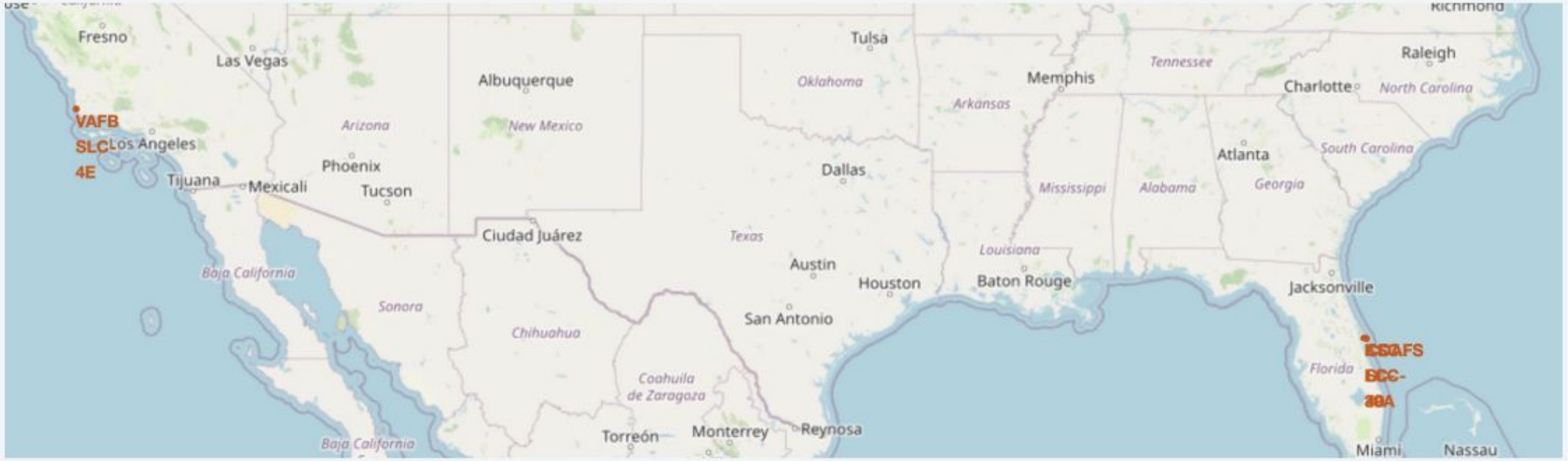
- SELECT "Landing _Outcome", Count(*) FROM SPACEXTBL WHERE "Landing _Outcome" IN ('Success','Success (drone ship)','Success (ground pad)') AND substr(Date,7,4)||substr(Date,4,2)||substr(Date,1,2) BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY "Landing _Outcome" ORDER BY (Count(*)) DESC;
- Additional SQLLite date handling, DISTINCT, filtering with BETWEEN, and outcome aggregation using GROUP.

Landing _Outcome	Count(*)
Success (drone ship)	4
Success (ground pad)	2


Section 3
Launch Sites
Proximities Analysis

Launch site map

- The launch sites

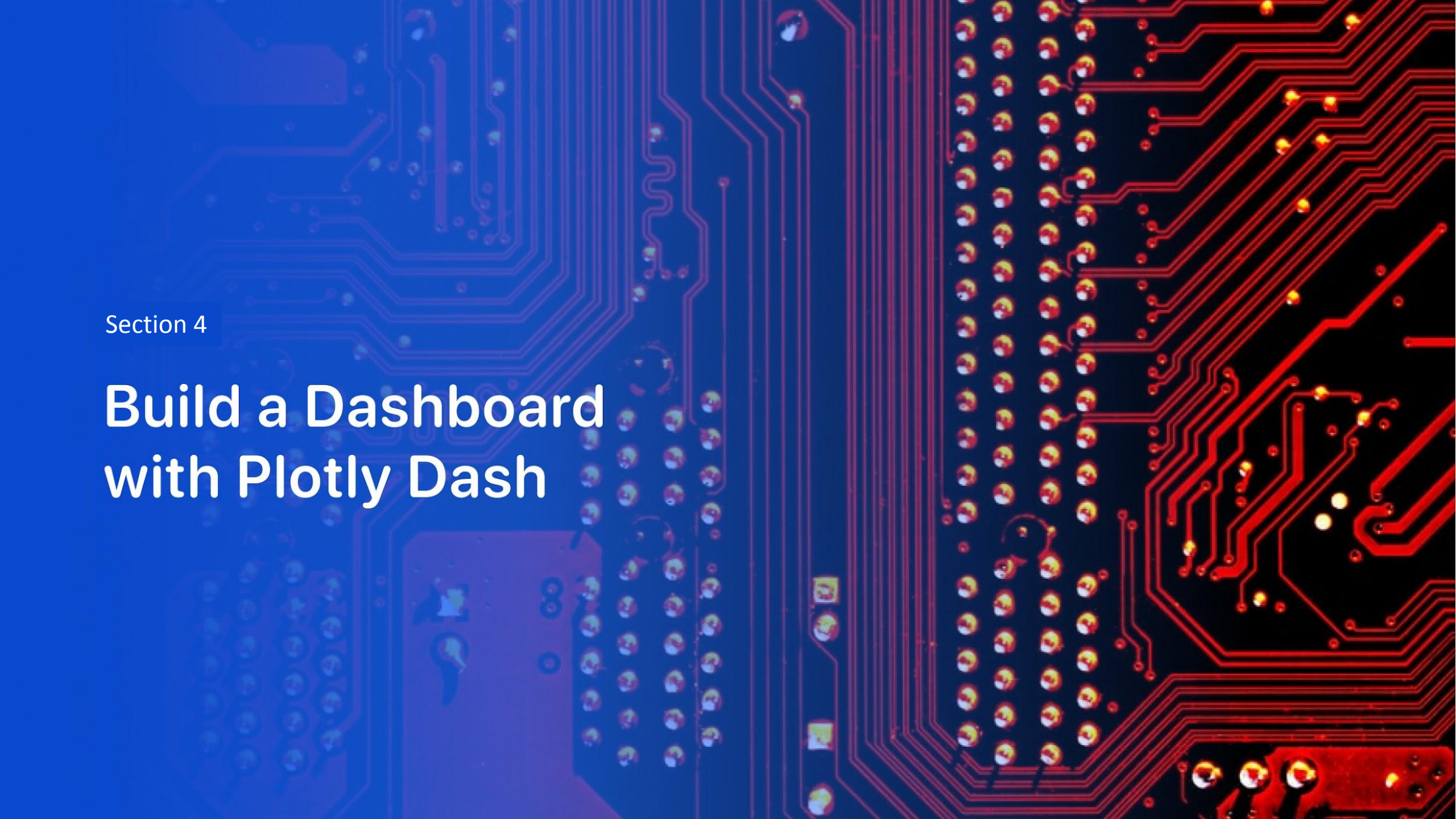


Launch site outcomes

- Clustered markers displaying launch outcomes for launch site KCS.

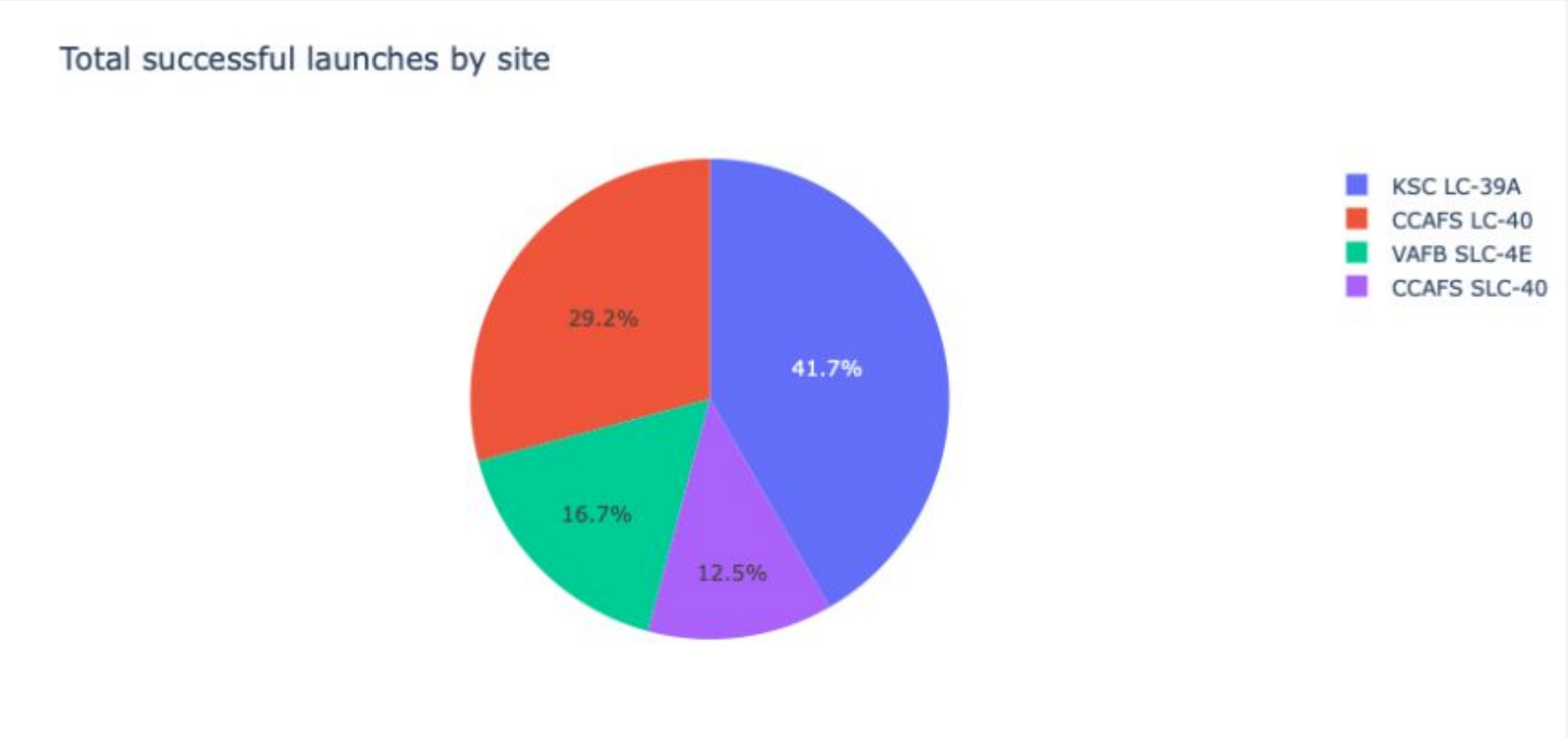
Proximity of launch site to coastal

- Knowing the distance between a launch site and other land/coastal features could be crucial.


Section 4
Build a Dashboard with Plotly Dash

Launch successes

- KSC has witnessed a higher number of successful launches.

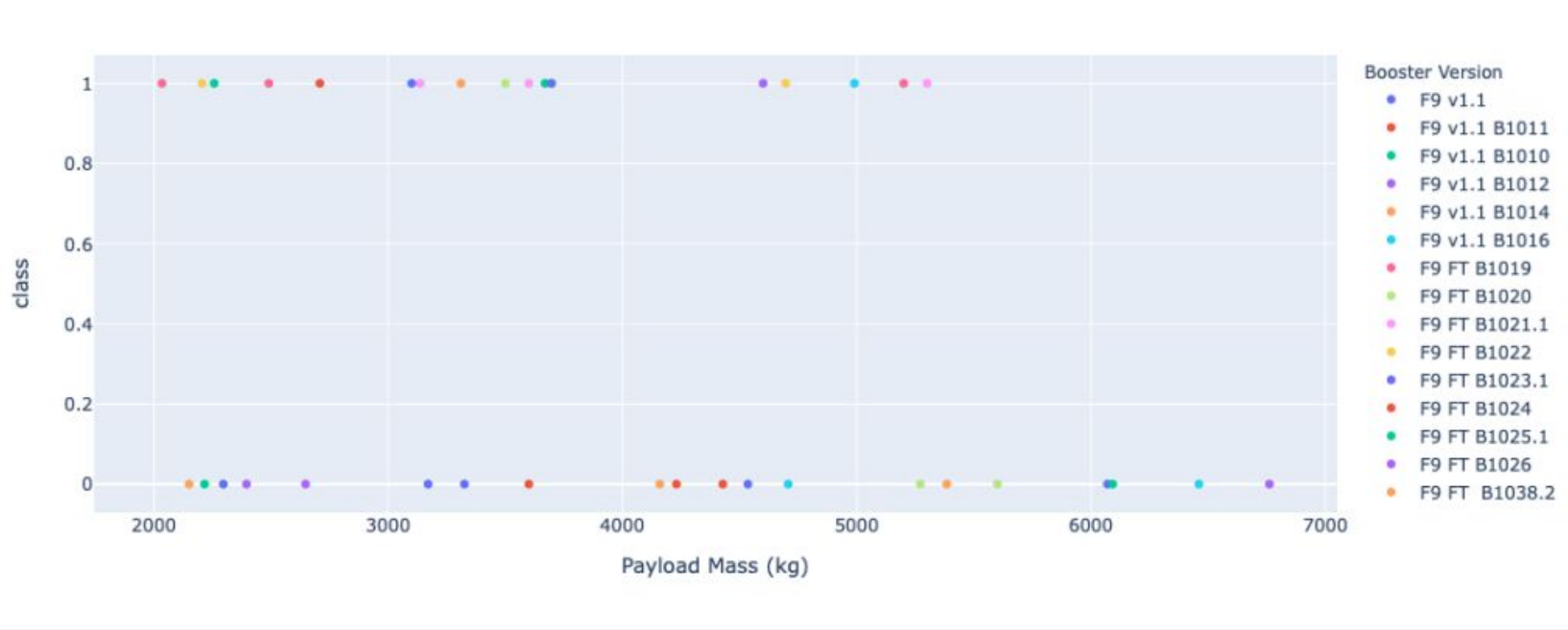


Launch site with the highest success rate

- The KSC launch site boasts the highest success rate, which indicates that its proportion of successful launches in the previous graph is not solely due to the overall number of launches.

Payload vs. Launch Outcome

- To gain better insights from this data, further drilling down is necessary, such as re-categorizing booster types into v1.1 and FT.

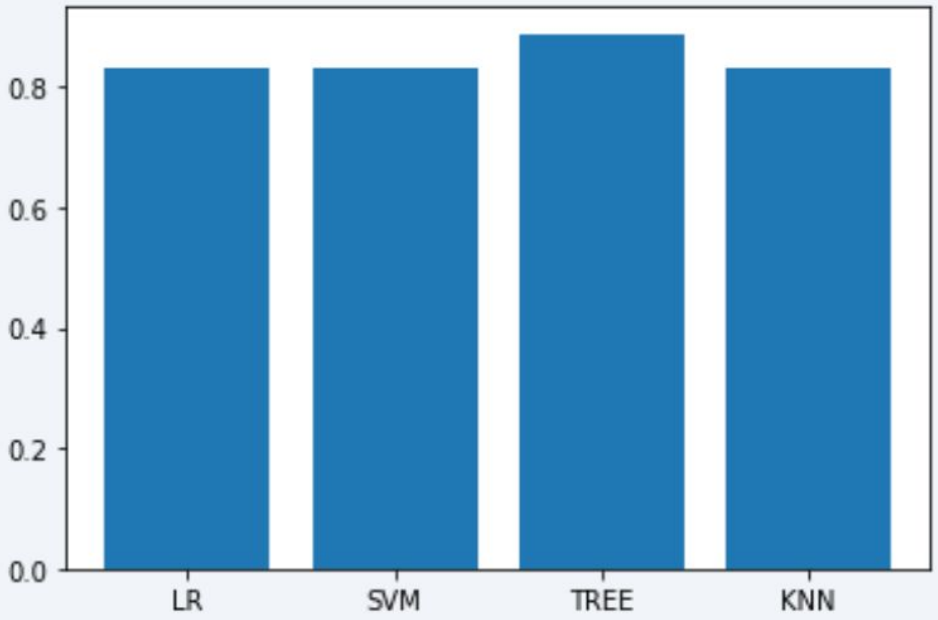


Section 5

Predictive Analysis (Classification)

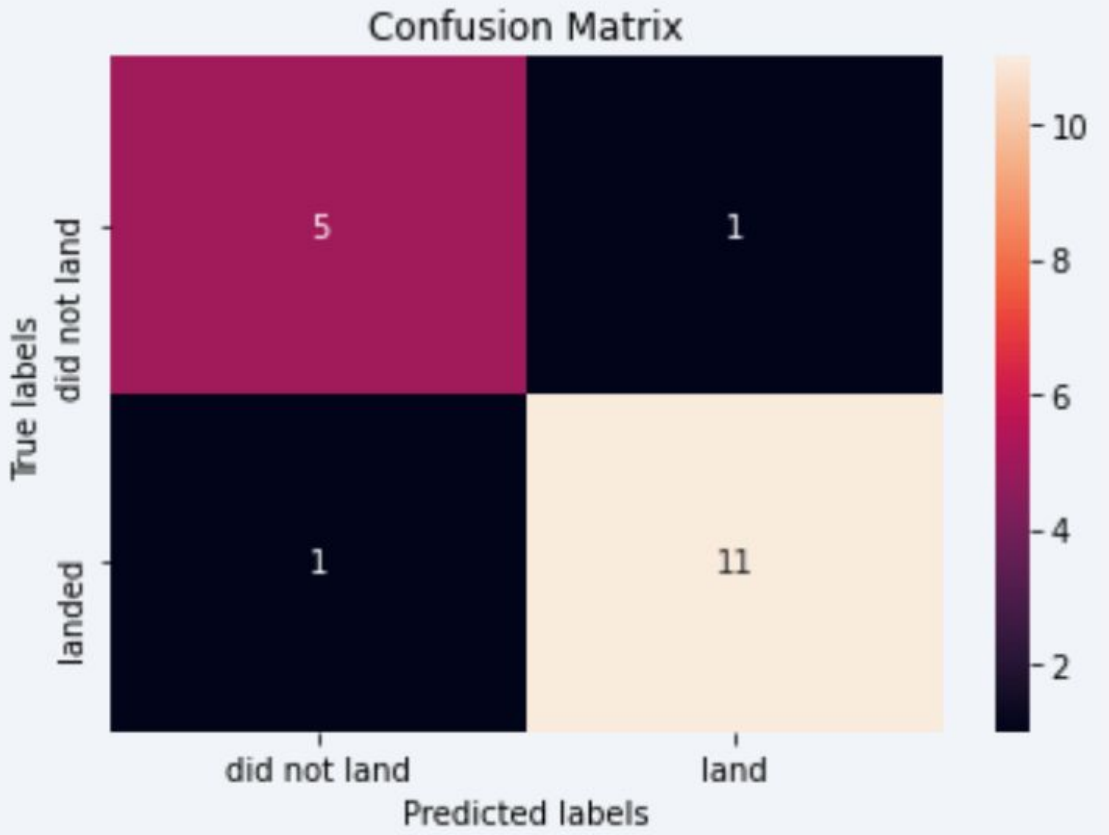
Classification Accuracy

- Model accuracy for each classification models
- The tree classification model showed the highest accuracy at 0.89



Confusion Matrix

- The decision tree demonstrated the lowest combined false positives and false negatives (only one of each, totaling two), making it the best model in that regard.



Conclusions

- Launch success rates show a positive increase over time.
- The KSC launch site exhibits the most favorable launch outcomes.
- A decision tree is the optimal classifier for predicting launch outcomes using information such as launch site, payload, orbit, booster types, etc.

Appendix

- I don't think I need to add anything else here, as everything has been included (such as the links, etc.) in their respective sections.

Thank you!