IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Christopher Ereforokuma

4th February 2023

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results

Introduction

- Project background and context
- Problems you want to find answers

Section 1

Methodology

Methodology

Executive Summary

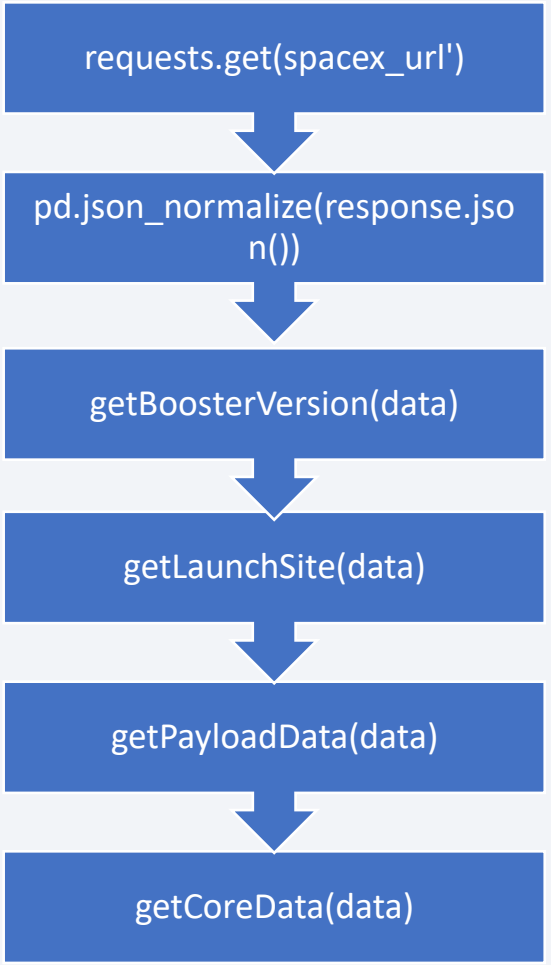
- Data collection methodology:
  - Data was collected by Web Scraping Wikipedia and using SpaceX API
- Perform data wrangling
  - Null data was removed and created a Landing Outcomes label
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection – SpaceX API

Data collection by SpaceX REST API was carried out as follows:

- Launch data was requested.
- Booster data was obtained from a rocket type request.
- Launchpad name, longitude and latitude came from a launch site request.
- Payload data came from a payloads request.
- Specific rocket core data was gotten from core request.

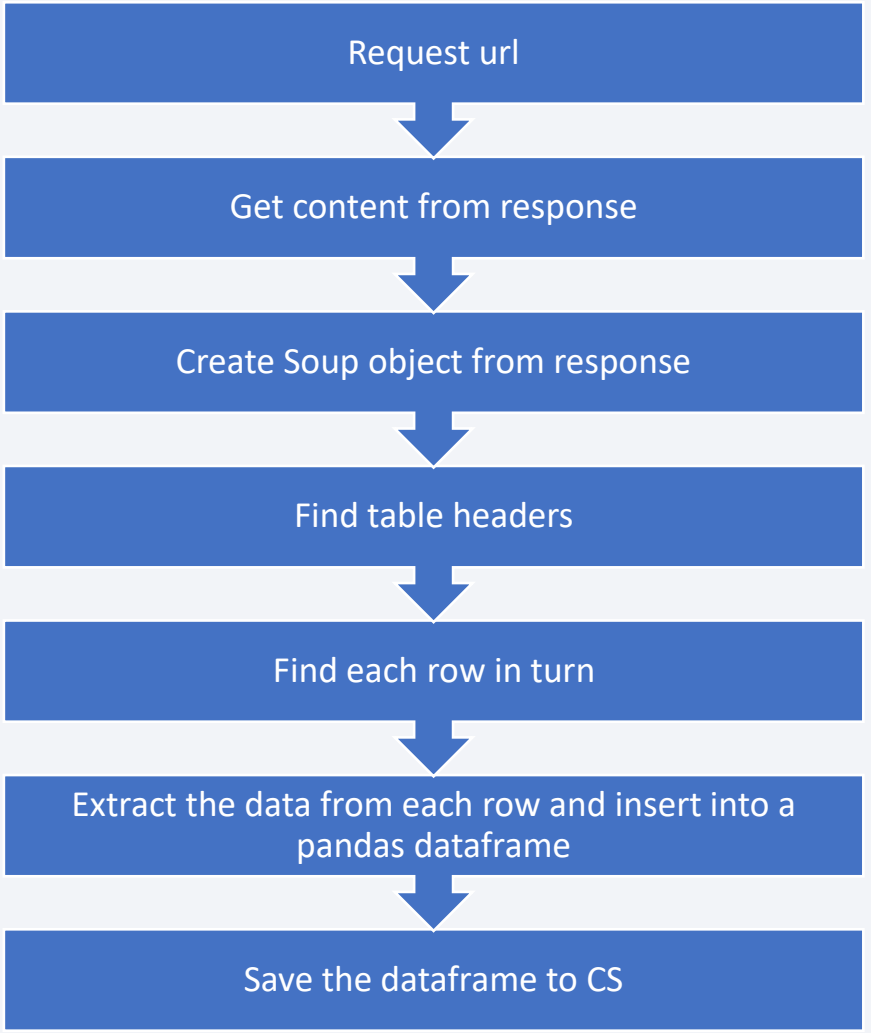
https://github.com/fero-chris94/Data-Science/blob/master/Data%20Collection.ipynb



Data Collection - Scraping

- Used the requests library to scrape data.
- Used BeautifulSoup to parse the content returned in the response
- The parsed data was added to a pandas dataframe and then exported to a CSV file

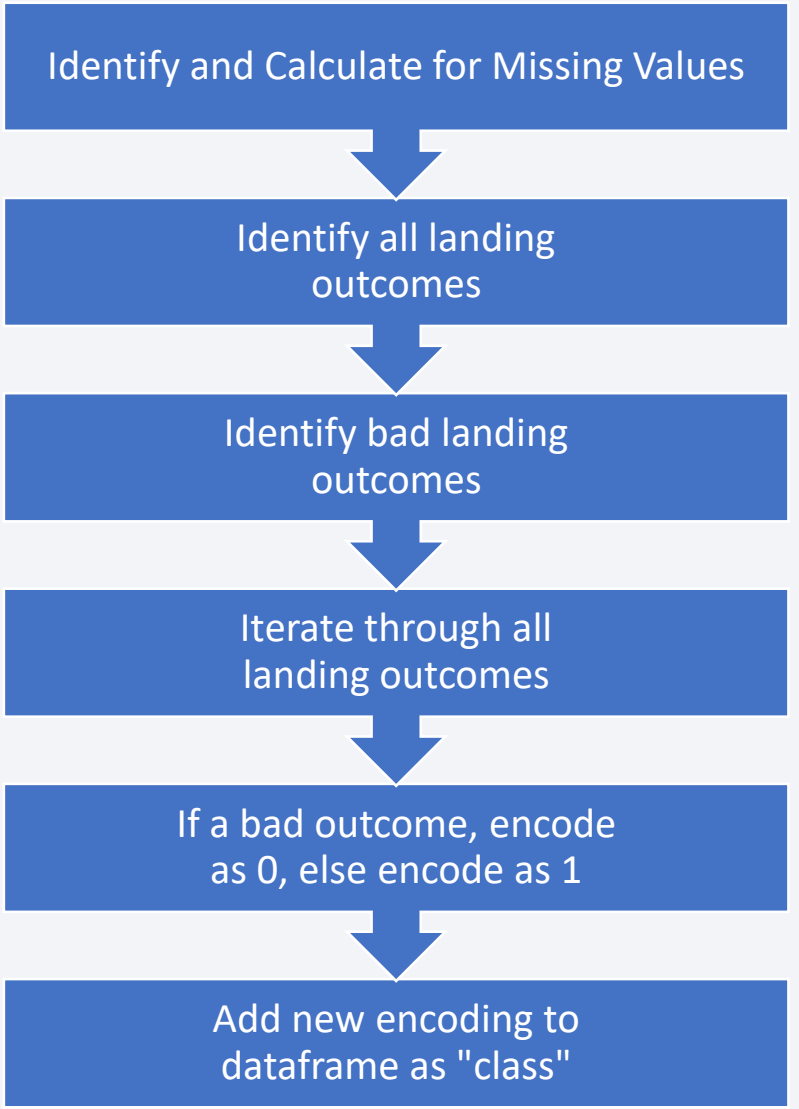
https://github.com/fero-chris94/Data-Science/blob/master/SpaceX%20Data%20Collection%20with%20Web%20Scraping.ipynb



Data Wrangling

- We Calculated the number of launches on each Site, number and occurrence of each Orbit.
- We Created a Variable named “class” to encrypt the “Landing Outcome” as either 0 or 1.

- https://github.com/fero-chris94/Data-Science/blob/master/Data%20Wrangling.ipynb



EDA with Data Visualization

- Exploratory Data Analysis was performed at this point
- Matplotlib Visualization techniques and Feature Engineering to call up plots for the relationships between, Flight number, Launch site, Payload mass and Orbit type.
- The plots included Scatter plots, Bar graphs and Line Graphs.
- All visualizations were color-coded by "class" so the effects of the variables and their relationships on the launch outcome were visible.
- One-hot-encoding was employed to turn categorical data to numeric data, suitable for the classification algorithms.
- https://github.com/fero-chris94/Data-Science/blob/master/Exploratory%20Data%20Analysis%20with%20Visualization.ipynb

EDA with SQL

- **The following queries were performed:**
  - SELECT DISTINCT Launch_Site from **SPACEXTBL**
  - SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5
  - SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer = 'NASA (CRS)'
  - SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version = 'F9 v1.1'
  - 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)'
  - SELECT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND "Landing _Outcome" = 'Success (drone ship)'
  - SELECT Mission_Outcome, 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 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)'
  - 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

- https://github.com/fero-chris94/Data-Science/blob/master/Exploratory%20Data%20Analysis%20with%20SQL.ipynb

Build an Interactive Map with Folium

- We added Circles and Markers to the interactive Folium Map to indicate Launch locations of SpaceX
- MarkerClusters were added to each site to visualize their Launch Outcomes.
- Lines were drawn between Launch Sites and the Coast, with label showing the distance added to show proximity between locations. These lines helped visualize distance to features like Highways and Railways etc. that are meant to be avoided.

- https://github.com/fero-chris94/Data-Science/blob/master/Interactive%20Visual%20Analytics%20with%20Folium%20(2).ipynb

Build a Dashboard with Plotly Dash

- An interactive dashboard was created to allow users to investigate the effects of launch site, payload mass and booster type on the launch outcome (good or bad).
- Launch site was selectable from a drop-down menu and a range of payload masses could be selected using a slider control.
- A pie chart showed either the successful outcome for all launch sites or the proportion of good and bad outcomes for any one selected launch site.
- A scatter chart showed how the launch outcome varied by the selected site and payload range and the data points were color-coded by booster type.

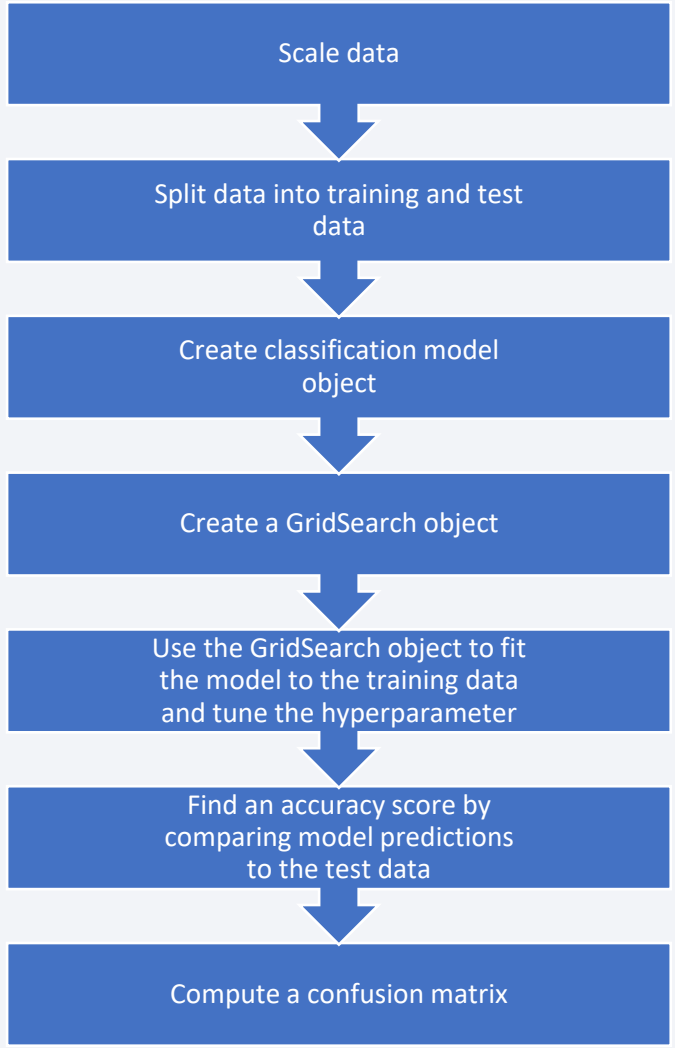
https://github.com/fero-chris94/Data-Science/blob/master/ibm-ds-pro-capstone.ipynb

Predictive Analysis (Classification)

- The Decision tree was the best performing classification with an accuracy score of 0.875. I came
- I came to this result after following the process outlined in the flowchart for the algorithms including:

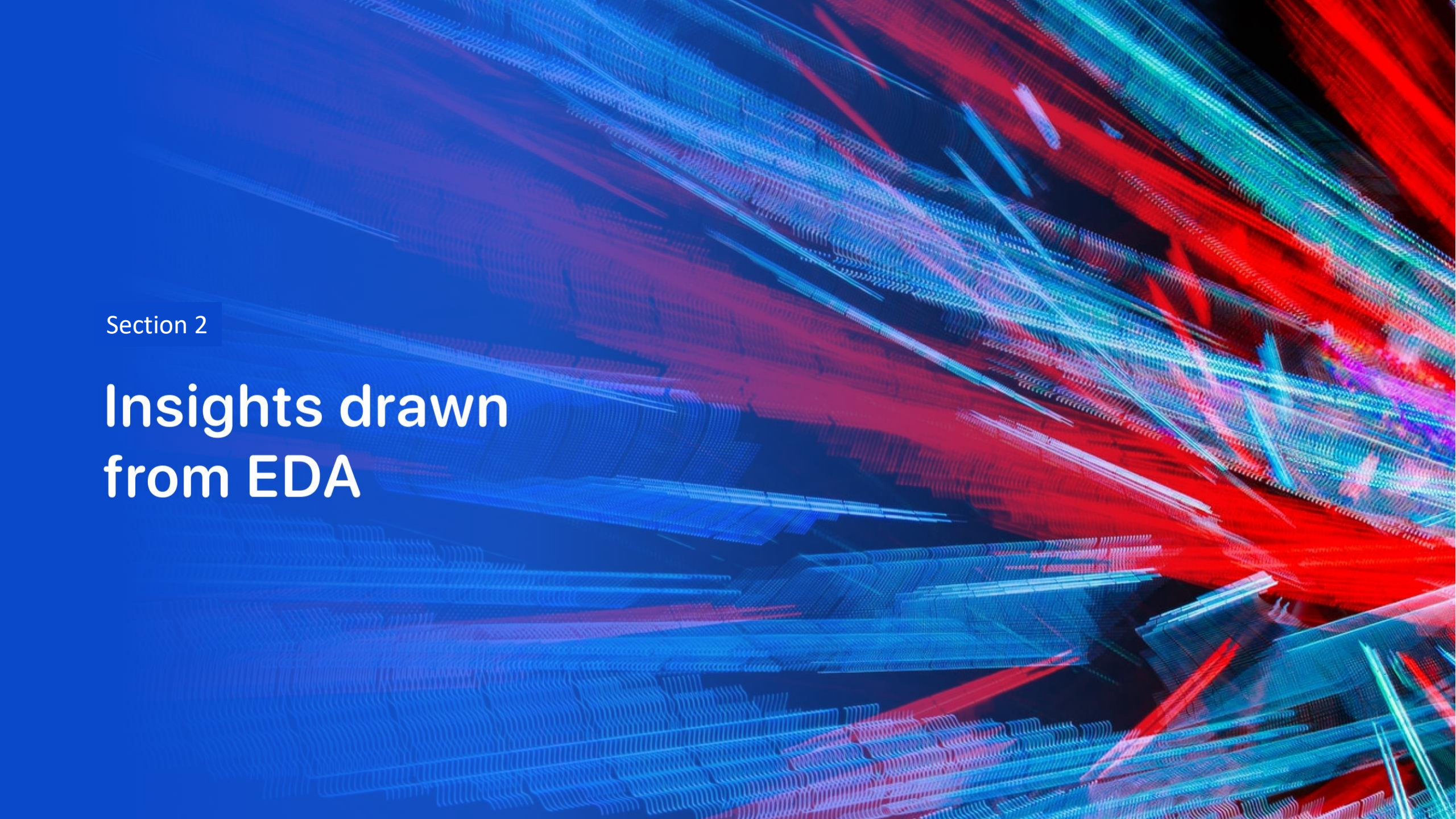
1. Logistic regression
2. Support vector machines
3. Decision tree
4. K-nearest neighbors

- https://github.com/fero-chris94/Data-Science/blob/master/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

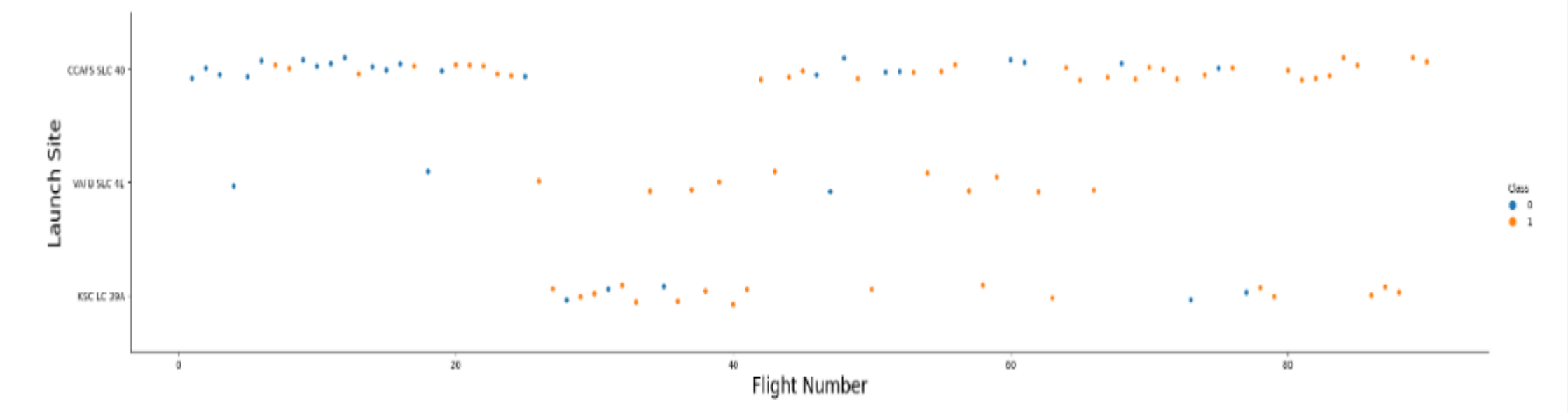


Results

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


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- Show a scatter plot of Flight Number vs. Launch Site

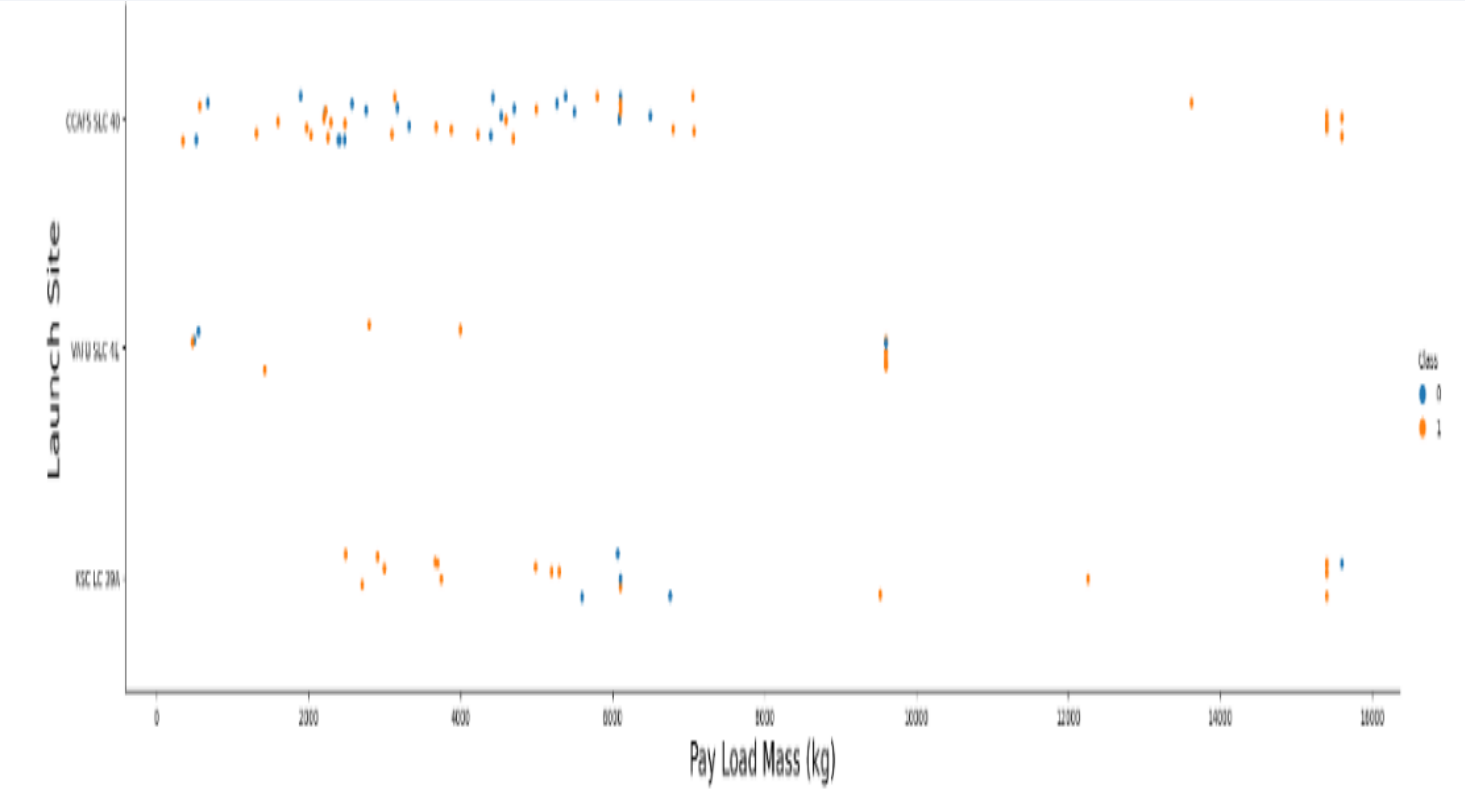
- Show the screenshot of the scatter plot with explanations

Payload vs. Launch Site

- Show a scatter plot of Payload vs. Launch Site

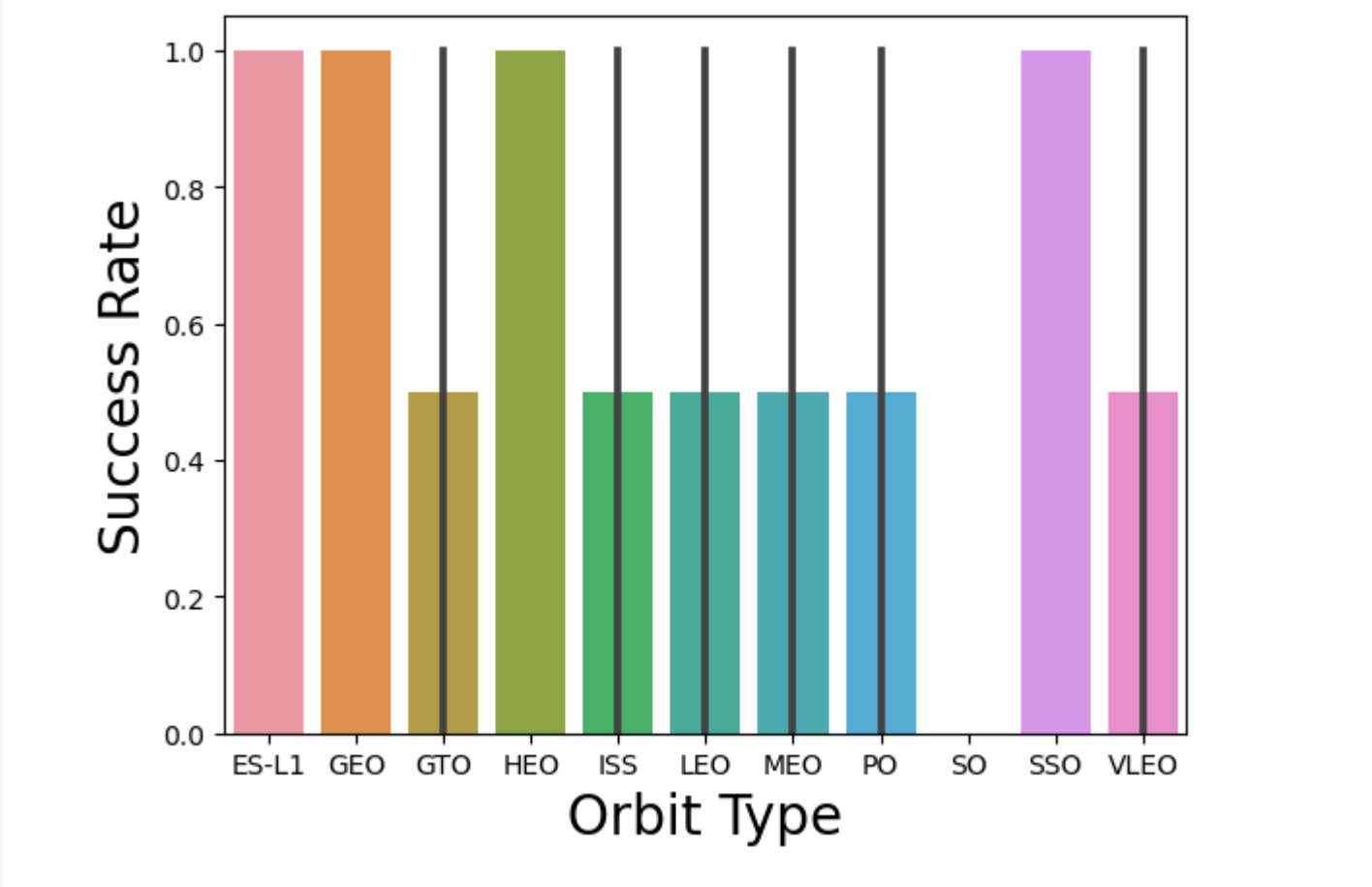
- VAFB... has not had many heavy flights.

- Launches with higher payloads were more successful (possibly because they were only carried out when the engineers knew that the reliability had improved and the extra cost of the payload would be safer)



Success Rate vs. Orbit Type

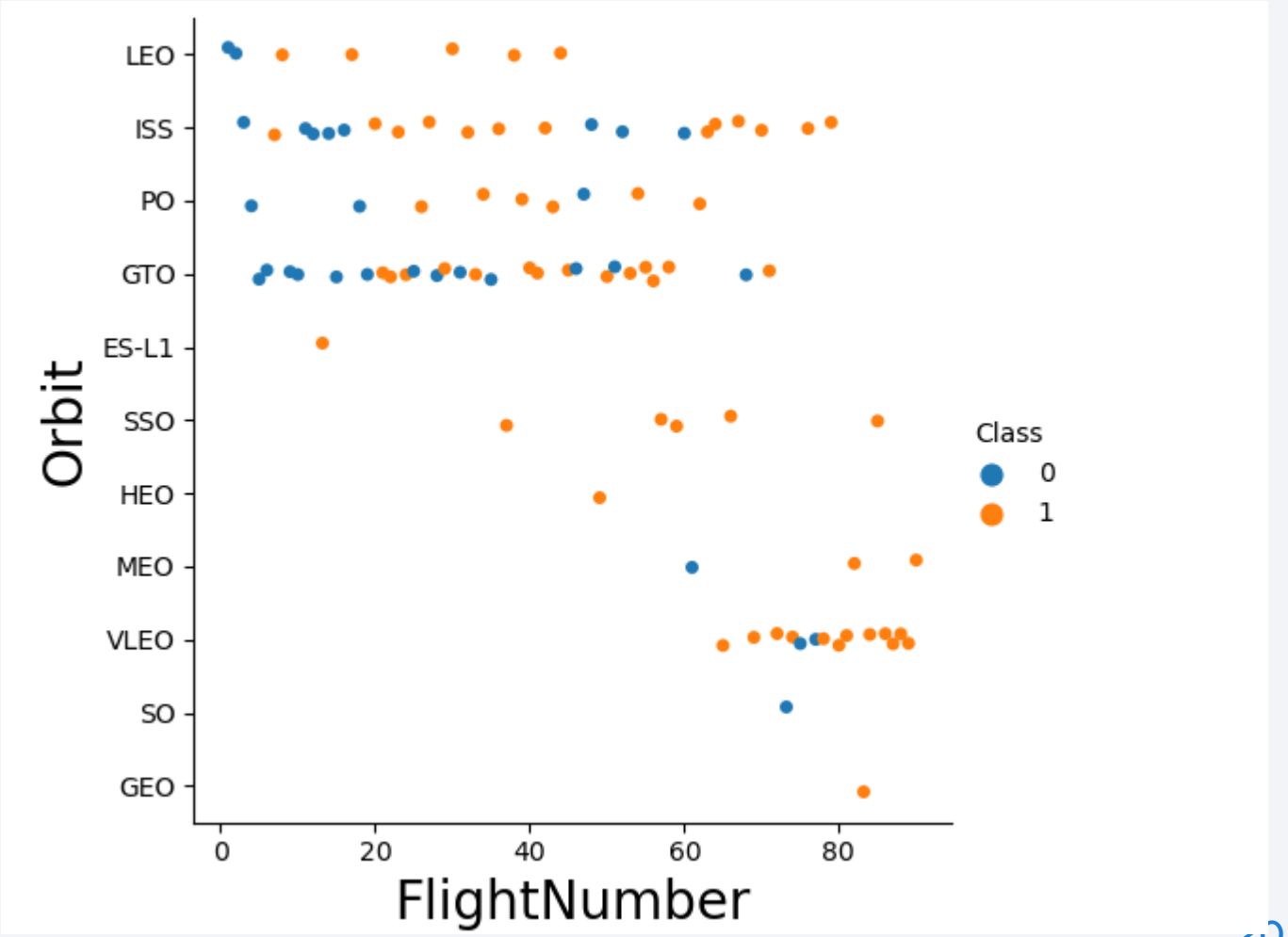
- Show a bar chart for the success rate of each orbit type
- Show the screenshot of the scatter plot with explanations



Flight Number vs. Orbit Type

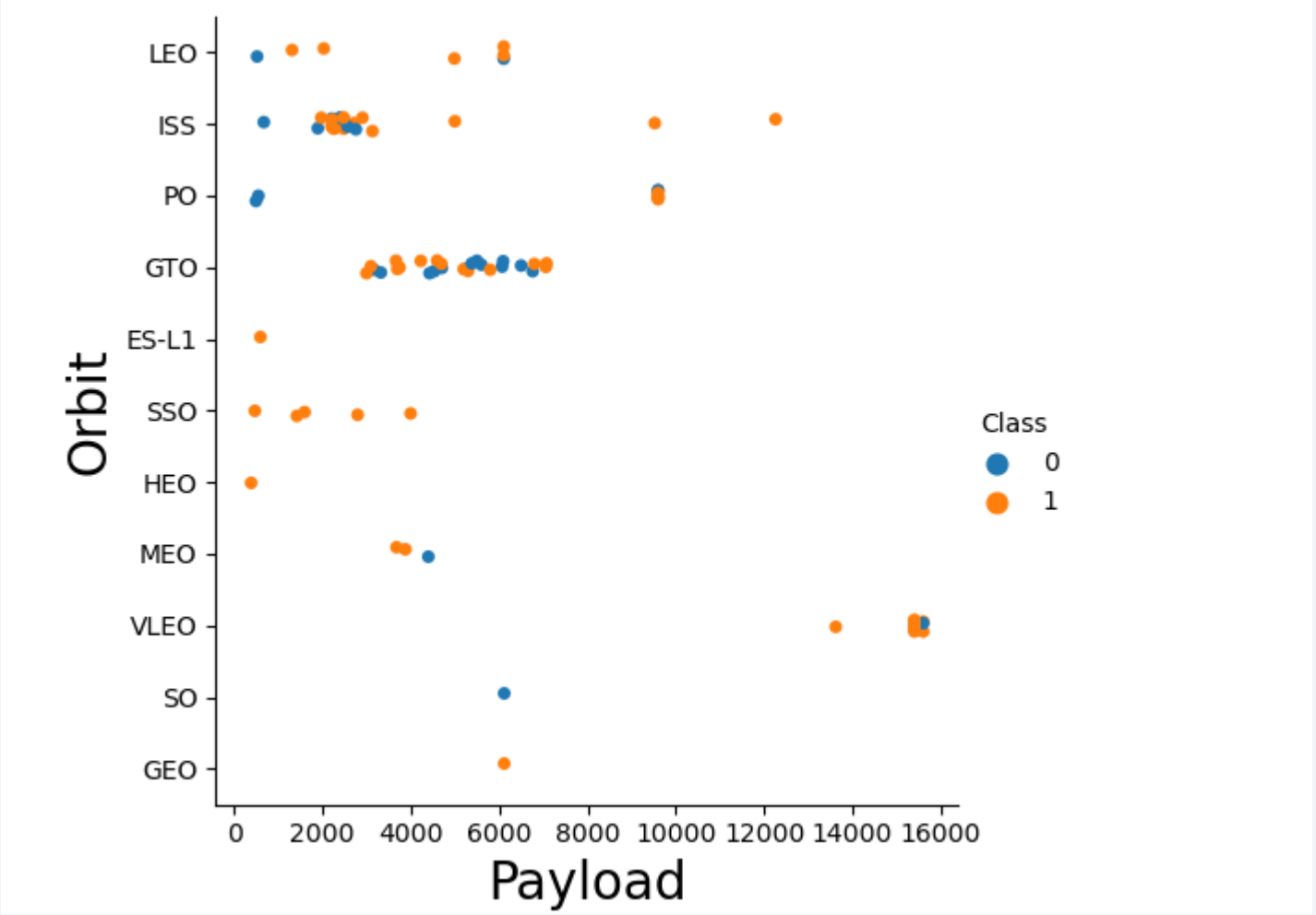
- Show a scatter point of Flight number vs. Orbit type

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



Payload vs. Orbit Type

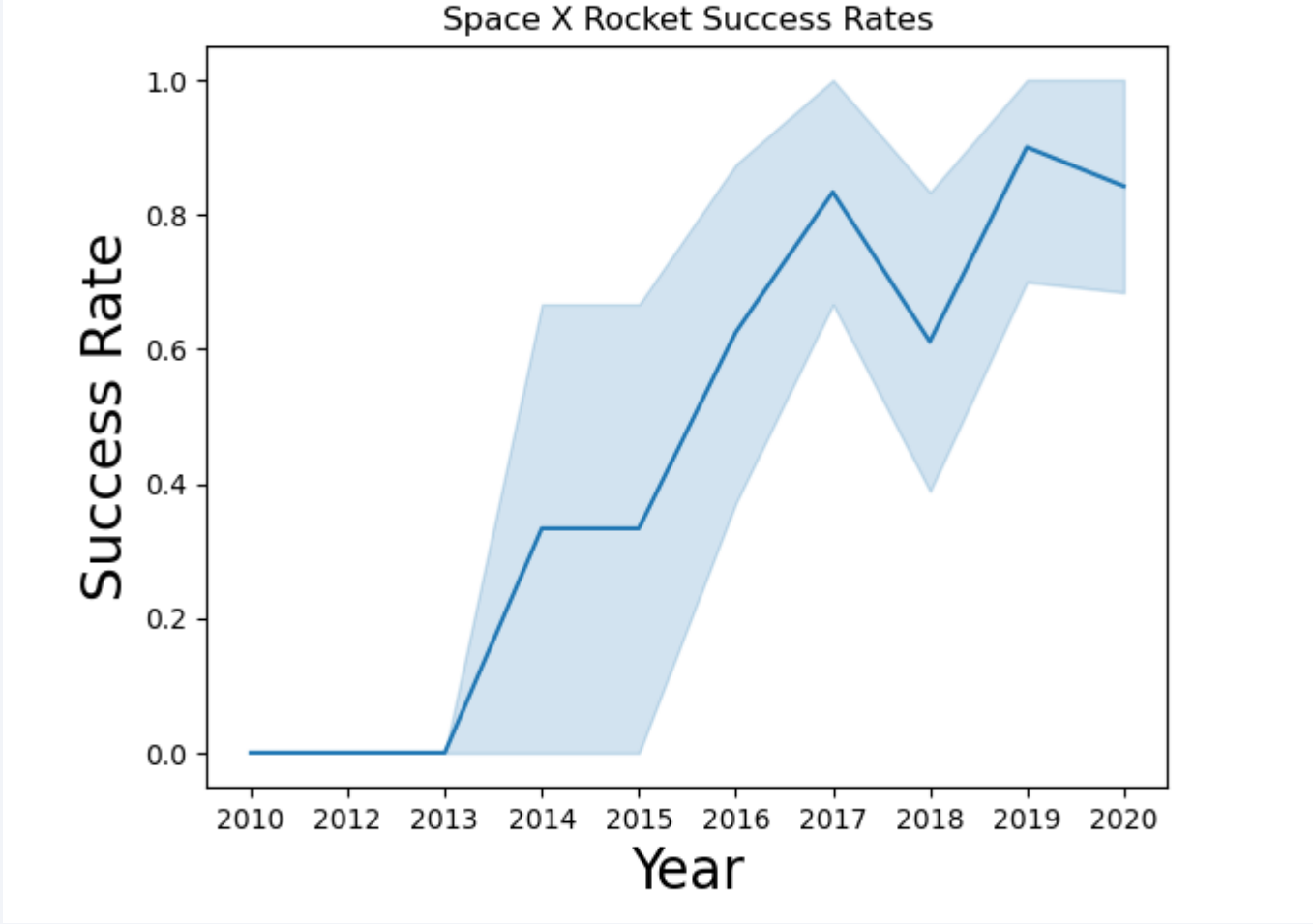
- Show a scatter point of payload vs. orbit type
- There are no clear patterns of increased success with increased payload for any given orbit.
  • SSO orbit is appears consistently successful (note that SSO = SO so actually it isn't 100% successful).



Launch Success Yearly Trend

- Show a line chart of yearly average success rate

- The success rate has been steadily increasing since 2013, although there was a dip around 2018.



All Launch Site Names

- select distinct(LAUNCH_Site) from SPACEXTBL;
- SELECT DISTINCT selects the unique values from the specified column.

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

Launch Site Names Begin with 'CCA'

- 5 Records where launch site name begins with CCA

  select * from SPACEXTBL where LAUNCH_SITE like 'CCA%' limit 5;

- *, selects all; WHERE filters the data based on the pattern specified by the LIKE

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

- The total payload carried by boosters from NASA

```
SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer =
'NASA (CRS)'
```

• SUM sums the payload values in the set of records filtered by Customer according to the condition specified in the WHERE command.

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1

```
SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version = 'F9 v1.1'
```

- Similar method to the last slide but using AVG and a different filter condition.

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad

  select min(DATE) from SPACEXTBL where [Landing _Outcome] = 'Success (ground pad)';

- Select min() selects the minimum value on the date column and where specifies the condition that it must be a success and ground pad.

Successful Drone Ship Landing with Payload between 4000 and 6000

- 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 [Landing_Outcome] = 'Success (drone ship)' \AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;
```

- The BETWEEN command is useful for filtering records by some range of values within a column

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

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes

```
select count(MISSION_OUTCOME) from SPACEXTBL where MISSION_OUTCOME
= 'Success' or MISSION_OUTCOME = 'Failure (in flight)'
```

- Count() returns the count of mission outcomes, and the where clause give condition for values returned.

count(MISSION_OUTCOME)
99

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass

  select BOOSTER_VERSION from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)

- A subquery is used to find the maximum payload mass across all records [using MAX()] and that returned value then becomes the filter with which 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 BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE
[LANDING _OUTCOME] = 'Failure (drone ship)' AND Date LIKE '%2015';
```

The Select command calls data from the columns Launch Site and Booster Versions, whilst the WHERE AND gives the two conditions.

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

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

- 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

```
SELECT [Landing _Outcome], COUNT([Landing_Outcome])
FROM  SpaceXTBL WHERE Date BETWEEN '04-06-2010' AND '20-03-2017'  GROUP BY [Landing _Outcome] ORDER BY COUNT([Landing _Outcome]) DESC ;
```

Landing _Outcome	COUNT([Landing _Outcome])
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

Section 3

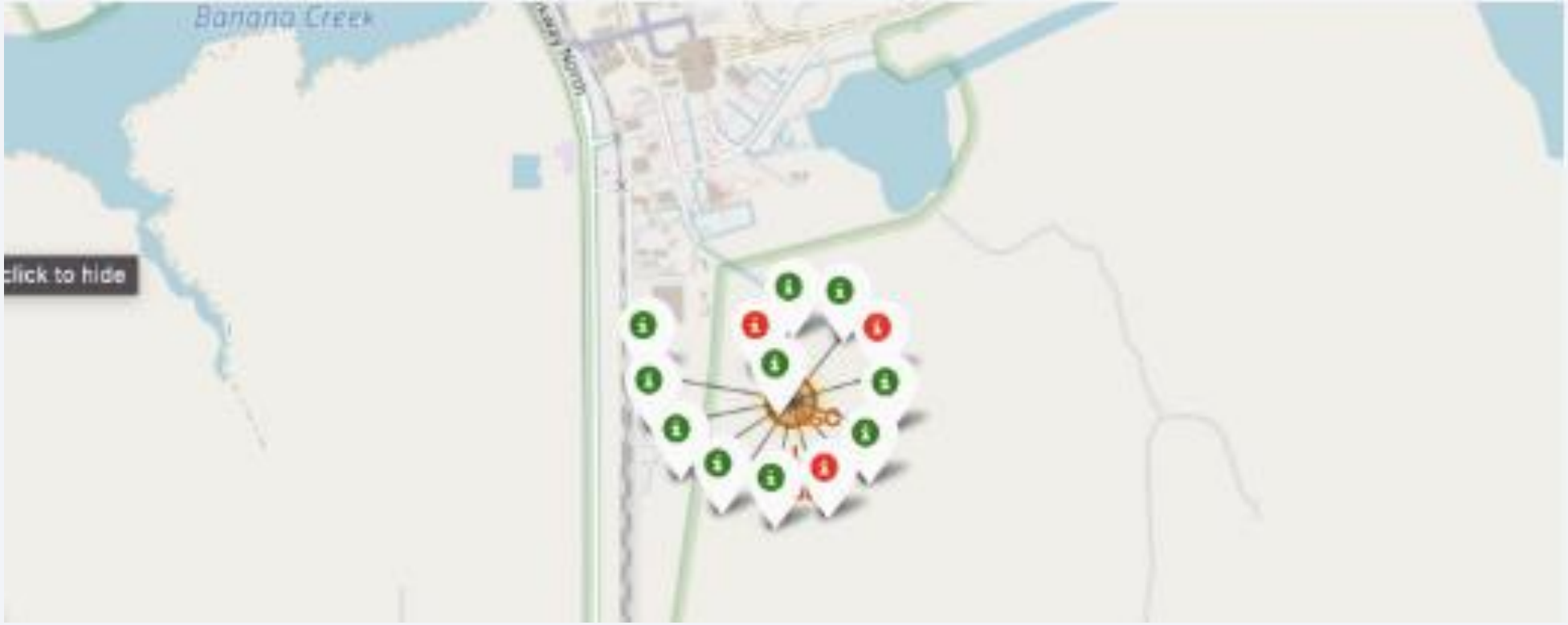
Launch Sites Proximities Analysis

Launch site locations

The Launch Site are Positioned in close proximity to Coasts of Major Oceans

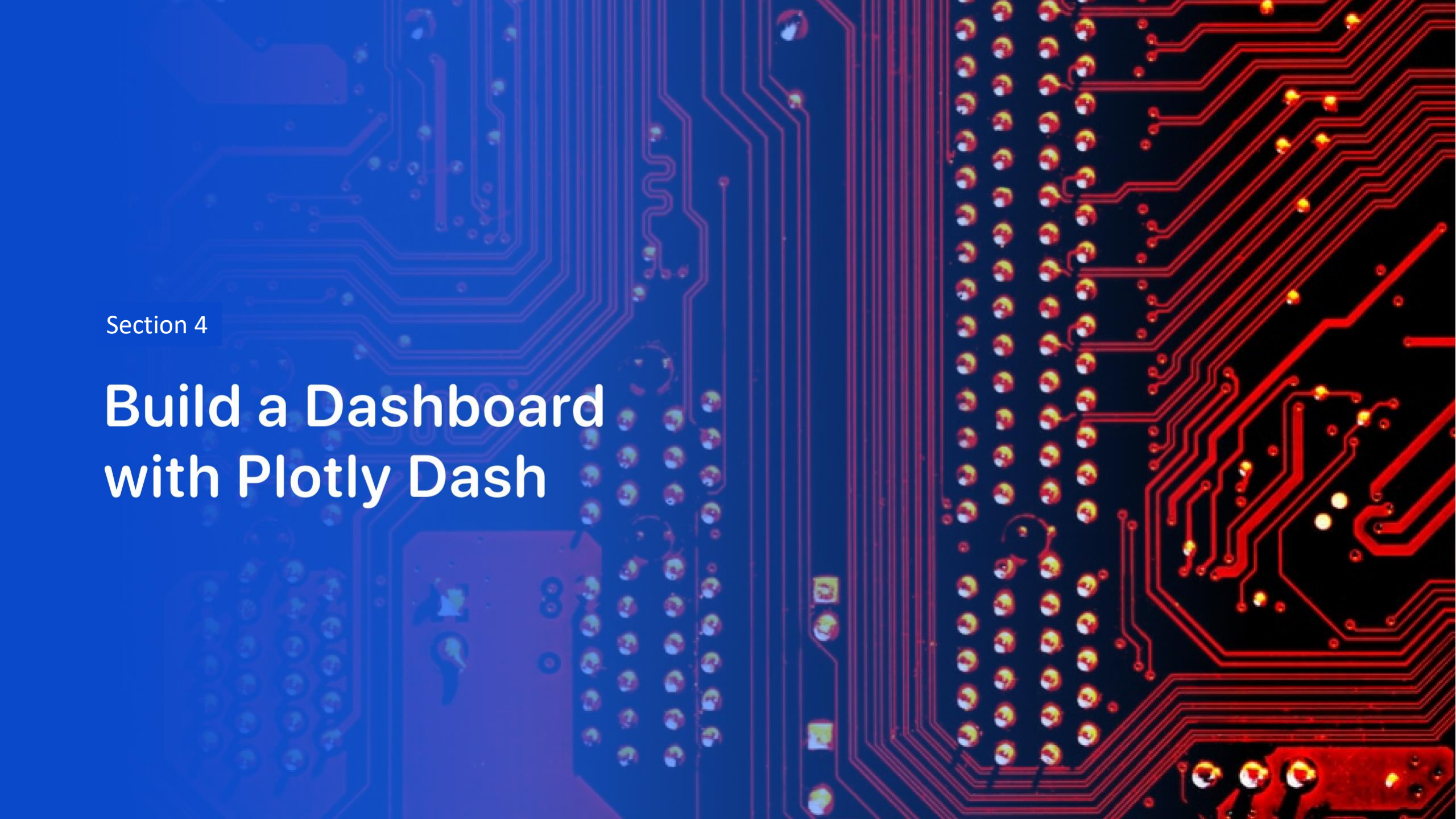
Launch site outcomes

We can Observed clustered set of markers. With the red indicating failed Launches and the Green indicating successful launches



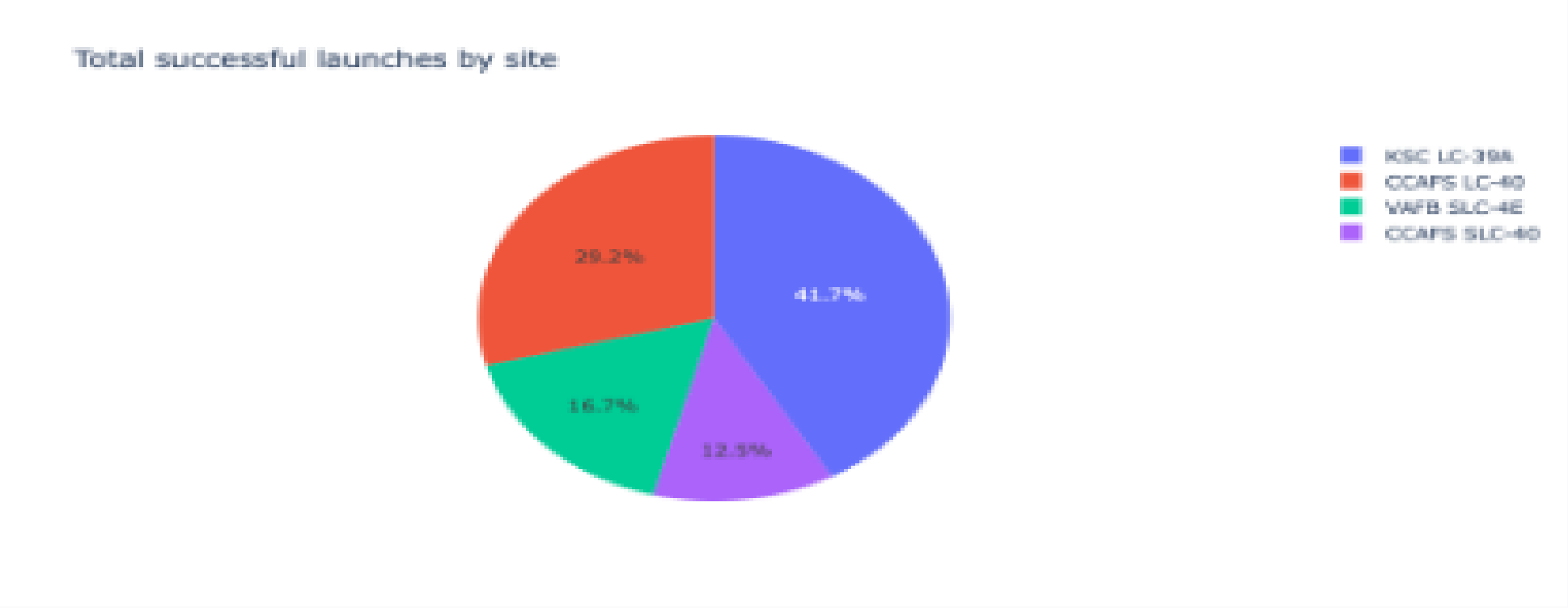
Proximity of launch site to other land/coastal features

- We can observe a line marking the proximity of a launch site to the Ocean, at the end of which is an indicator showing distance in kilometers.


Section 4
Build a Dashboard with Plotly Dash

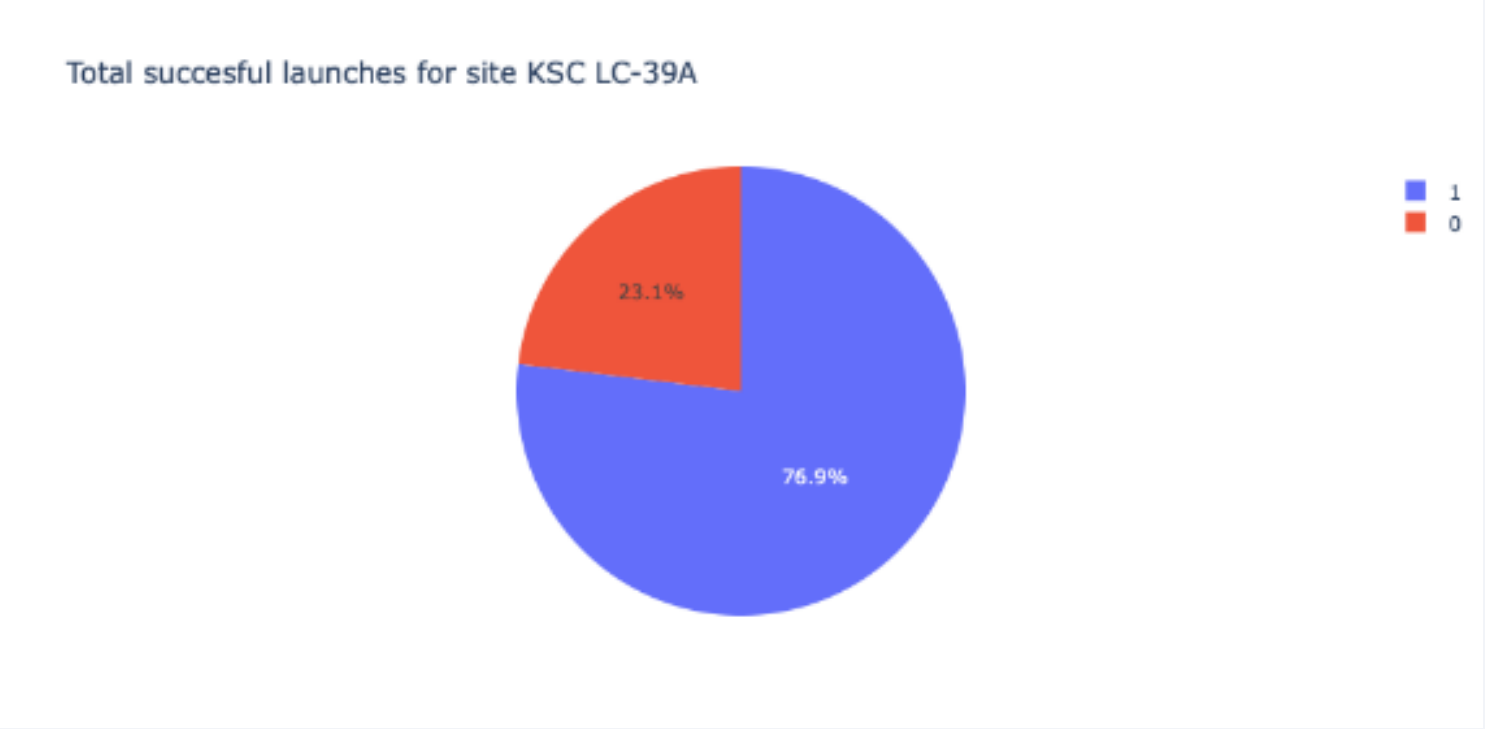
Launch successes across all sites

- Launch Site KSC LC-39A has the highest number of successful launches while CCAFS SLC-40 has the least.



Launch site with the highest success rate

- KSC LC-39A with the highest Success rate has 76.9% of its launches to be successes, whilst 23.1% were failures.



Payload (1000 – 7000 kg) vs. Launch Outcome for all sites

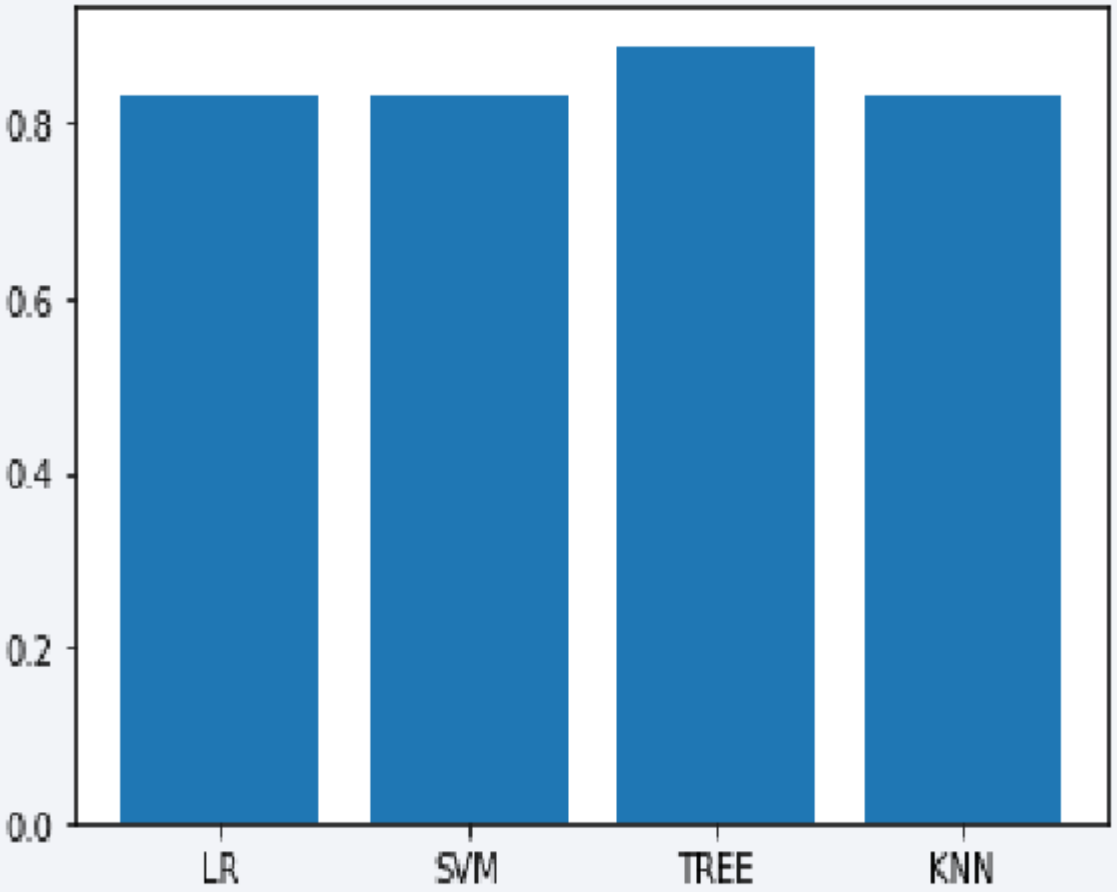
We can observe the absence of successful launches when the payload between 6000 and 7000 kilograms.

Section 5

Predictive Analysis (Classification)

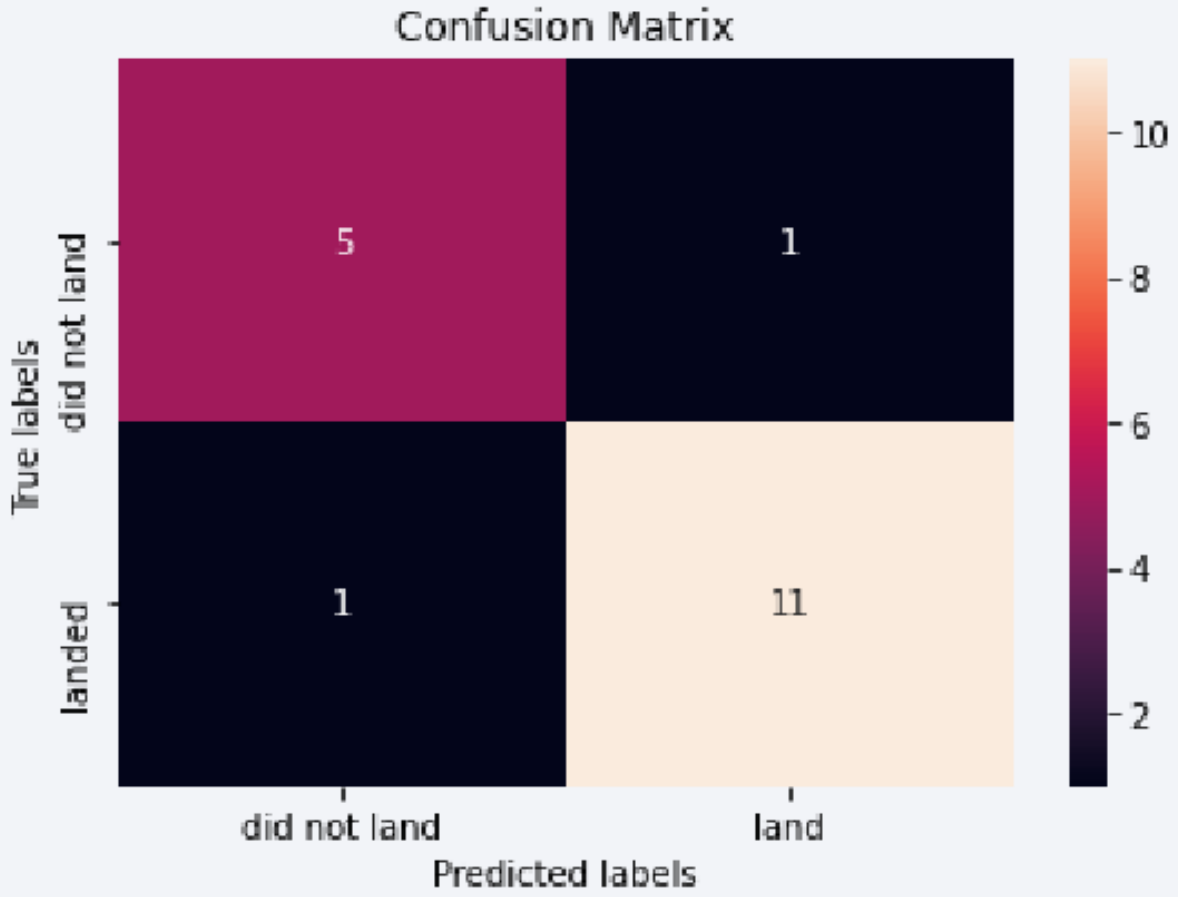
Classification Accuracy

- Model accuracy for each classification models
- The decision tree (TREE) classification model showed the highest accuracy at around 0.89



Confusion Matrix

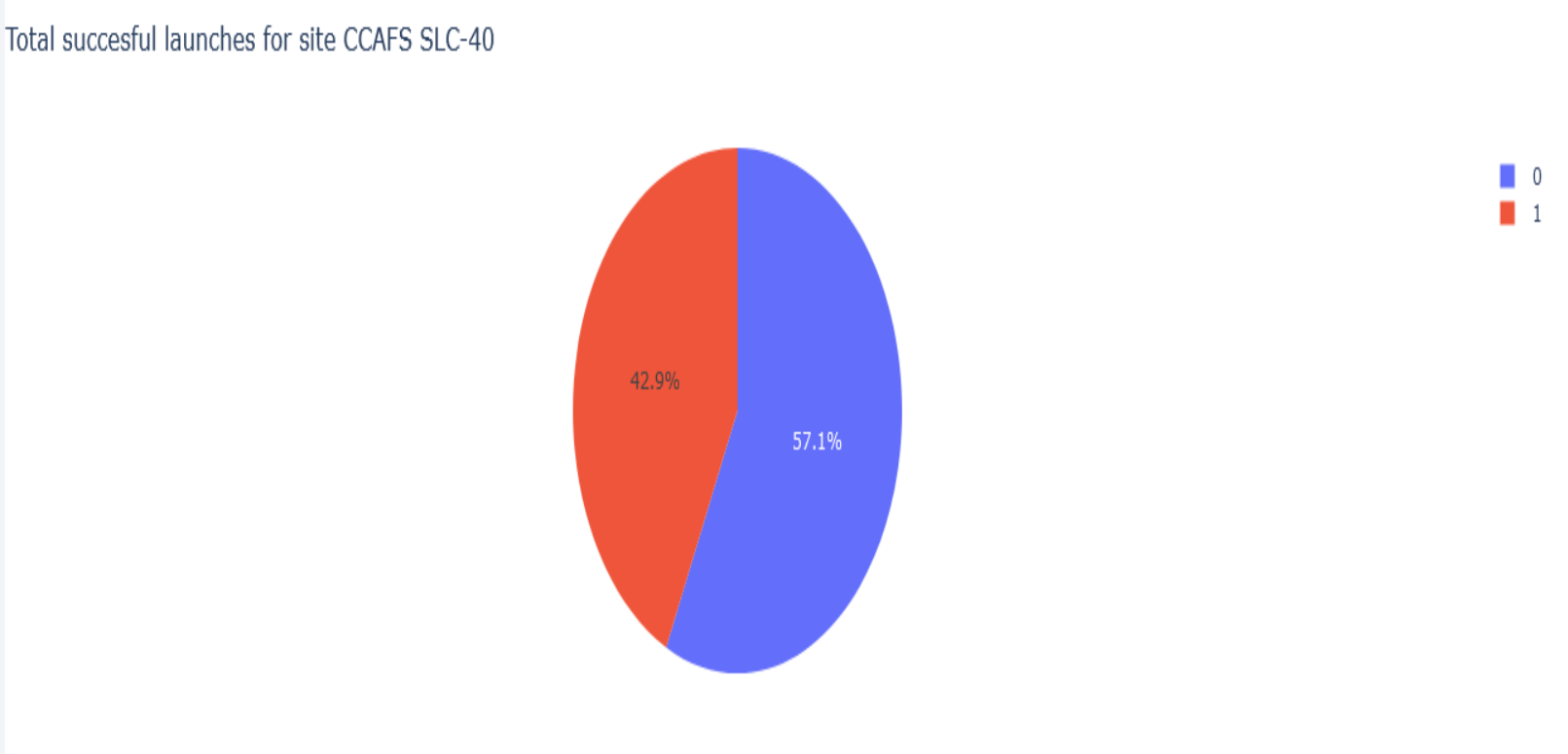
- The decision tree showed the fewest combined false positives and false negatives (one of each, so two in total) so, in that respect, was the best model.



Conclusions

- The Decision Tree has the highest accuracy for predicting Launch outcomes based on information such as the launch site, payload, orbit and booster types etc
- The KSC launch site has the best launch outcomes.
- Launch success rates increase with time.
- Payloads of over 6000kg are much less likely to have successful launches.

Appendix



Thank you!