Winning Space Race with Data Science

Author: Kiran Ravi
14/08/2024

Outline

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

Executive Summary

This report sets out to explore three main questions: can we predict whether a Falcon 9 first stage landing attempt will be successful, what factors affect this prediction, and what classification model most accurately predicts the outcome.

The methodology to approach this question consisted of a few distinct steps. First, historic data collection, data wrangling, and exploratory data analysis were conducted using SQL and Python. Then interactive maps and dashboards were made using Folium and Dash to gain greater insights into the launch data. Finally, different classification models were constructed and compared to see which would perform best in predicting the landing outcome.

Of the four models tested, it was found that the SVM most accurately estimated the landing outcome, with an accuracy of 87.78%. Many factors seemed to have some effect on the outcome, but payload range, landing site, and orbit type all had clear relations with landing outcome.

Introduction – Project Background

- With the emergence of the commercial space travel, cost saving is crucial to remain competitive in an extremely expensive market
- One of the main commerical space travel providers are SpaceX, who have managed to gain a competitive edge by reusing the first stage of their Falcon 9 rockets
- This allows SpaceX to reduce their costs to **62 million dollars** per launch as compared to other providers whose costs can be upwards of **165 million dollars**.
- In order to remain competitive, SpaceX needs to ensure that the first stage of their rocket will successfully land so it can be reused

Introduction - Business Questions

1. Can we determine whether a Falcon 9 first stage landing attempt will be successful based on historic data ?

2. What factors affect this outcome?

3. What model most accurately predicts this outcome?

Section 1

Methodology

Methodology – Executive Summary

- Data collection through:
  - SpaceX REST API
  - Historic launch data scraped from Wikipedia
- Exploratory data analysis (EDA) using:
  - Numpy and Pandas
  - SQL
- Data Visualisation using:
  - Matplotlib and Seaborn - for informative plots on launch stats
  - Folium – for interactive maps giving information on launch sites
  - Dash and Plotly – for interactive dashboards on launch data

Methodology – Executive Summary

- Machine Learning Prediction
  - Models: logistic regression, support vector machine, decision tree, k-nearest neighbours
  - Model parameters tuning done using GridSearchCV
  - Model accuracy validated using confusion matrices, accuracy scores, f1-scores, and jaccard scores

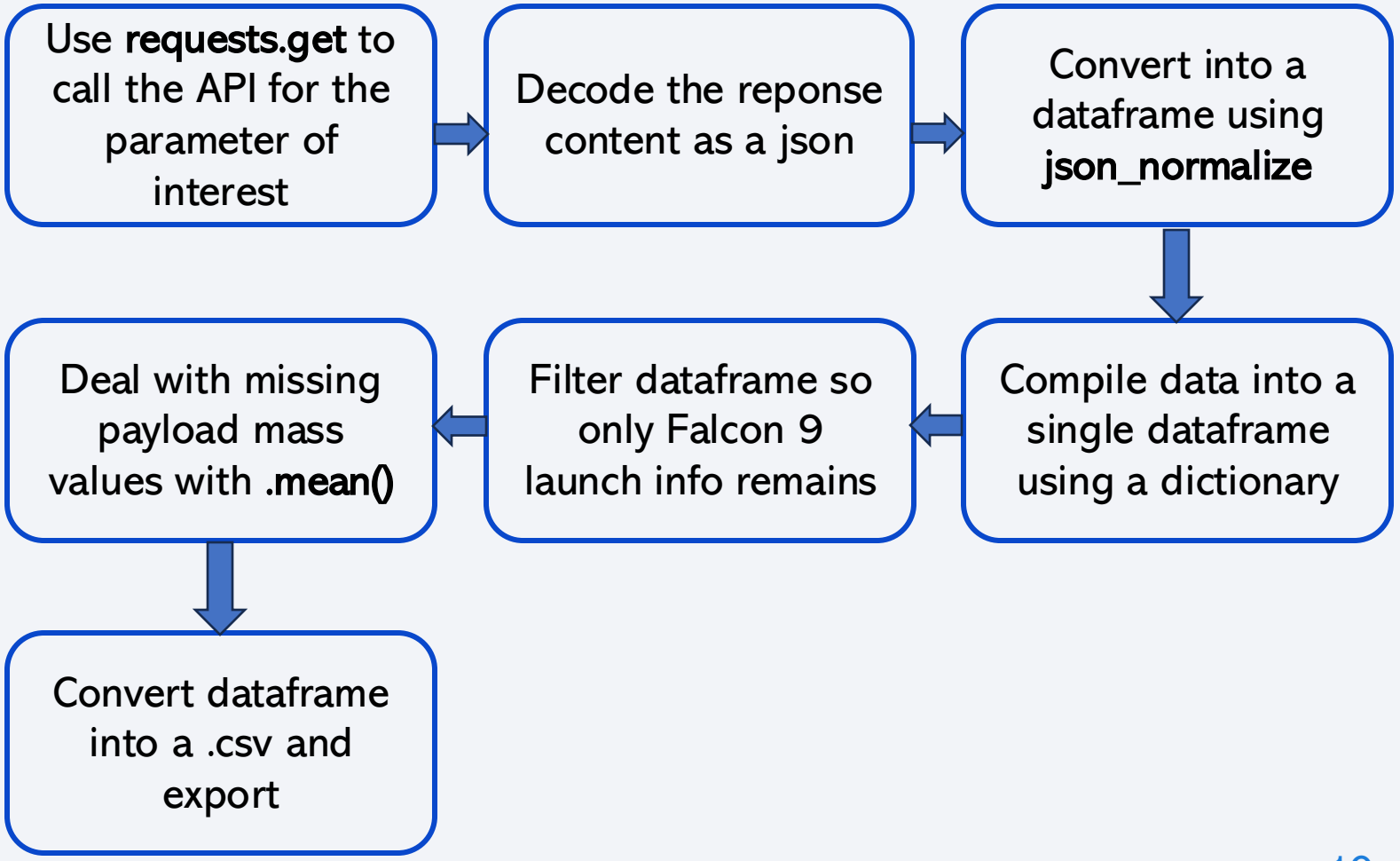
Data Collection

- SpaceX launch data was collected through two sources:
  1. The SpaceX REST API
  2. Launch data scraped directly from the SpaceX Falcon 9 launches Wikipedia page
- Both sources are required to give a complete data set that can be used for landing outcome prediction
- Once combined, the following data from each launch was available:

- Flight Number
- Launch Date
- Booster Version
- Payload Mass
- Orbit Type
- Launch Site
- Launch Outcome
- Number of flights
- Grid Fins (Y/N)
- Reused (Y/N)
- Reused Count
- Legs (Y/N)
- Landing Pad (Y/N)
- Block (Y/N)
- Serial Number
- Longitude of Launch
- Latitude of Launch

Data Collection – SpaceX API

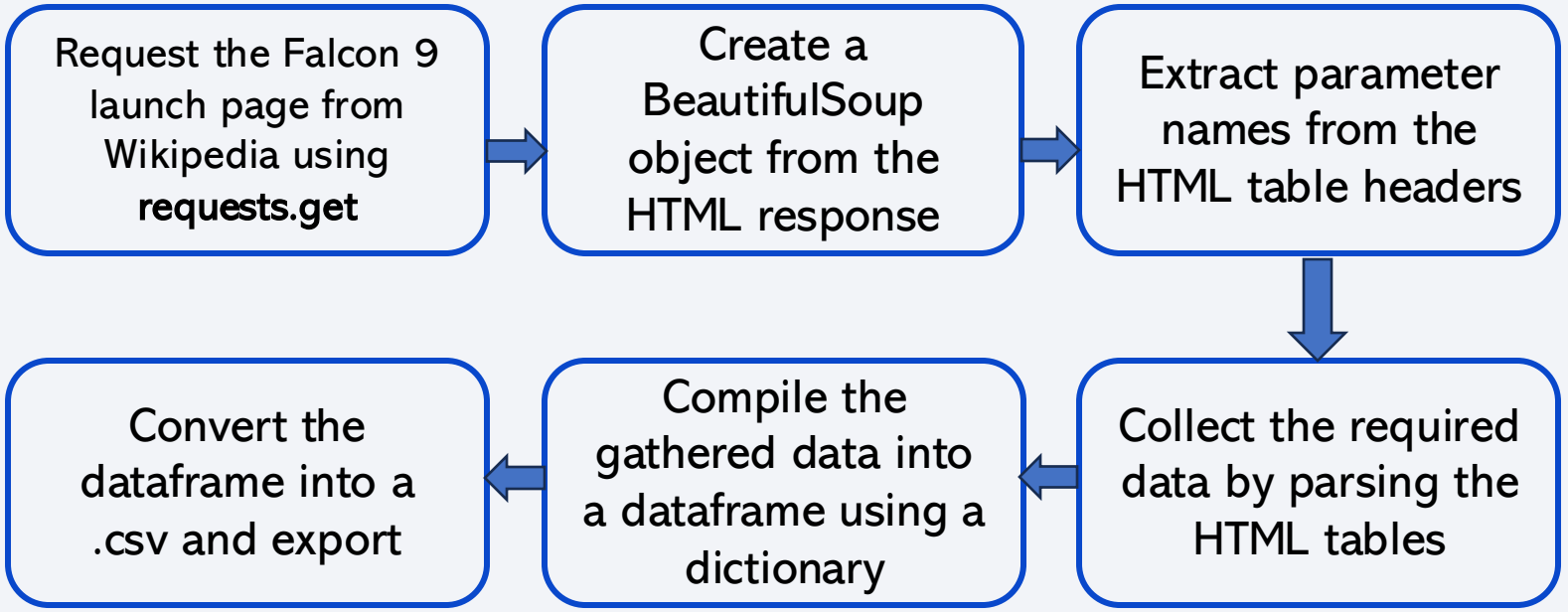
- Data collection from the SpaceX API was conducted as shown in the flowchart



API data collection notebook (GitHub)

Data Collection – Web Scraping

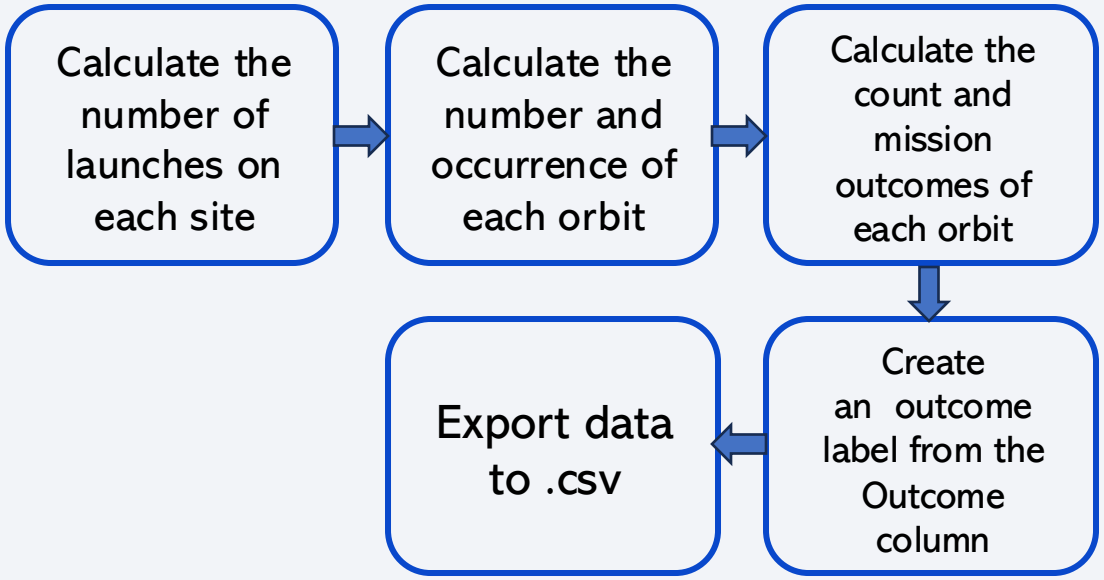
- Data collection from the SpaceX Falcon 9 launches Wikipedia page was conducted as shown in the flowchart



Web scraping data collection notebook (GitHub)

Data Wrangling

- To wrangle the data, some EDA needed to be performed, as shown in the flowchart
- Once EDA was performed, it was determined that the various landing outcomes could be split into two training labels
  - First half of outcome – whether landing was successful or not
  - Second half of outcome – where the landing was attempted



Landing outcome label	Landing Outcome
Successful (1)	True ASDS True RTLS True Ocean
Unsuccessful (0)	None None False ASDS False Ocean None ASDS False RTLS

Data wrangling notebook (GitHub)

EDA with SQL

The following data was queried using SQL as part of EDA:

- Unique launch site names
- 5 records where the launch site beings with the string 'CCA'
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- Date when first successful landing outcome in ground pad was achieved
- Names of boosters which have success in drone ship landings, and have a payload mass between 4000 and 6000 kg
- Total number of successful and failure mission outcomes
- Booster versions which have carried the maximum payload mass
- Drone ship landing failures with booster version and launch sites for the months in 2015
- Ranking landing outcomes by count in descending order between 2010-06-04 and 2017-03-20

SQL EDA notebook (GitHub)

EDA with Data Visualization

The following charts were plotted as part of EDA:

- Flight Number vs. Payload Mass (scatter)
- Flight Number vs. Launch Site (scatter)
- Payload Mass vs. Launch Site (scatter)
- Success rate of each Orbit Type (bar)
- Flight Number vs. Orbit Type (scatter)
- Payload Mass vs. Orbit Type (scatter)
- Average Success Rate per Year (line)

- **Scatter plots** were used to establish trends between models. If a trend is successfully identified, it can be used in the machine learning model for prediction
- **Bar charts** allow for visual comparison of values between discrete categories, such as success rate per launch site
- **Line charts** allow for easy visualisation of trends over a continuous variable, such as time

Data Visualization EDA notebook (GitHub)

Launch Sites Map - Folium

- Markers for all launch sites:
  - Added circle markers with popup and text labels to mark the listed launch sites based on coordinates
  - Locations marked to show the proximity of launch sites to the equator and to the coast
- Markers for successful/failed launches:
  - Added coloured markers within marker clusters at each launch site to indicate successful and failed launches
  - This allowed visual identification of which launch site had the most successful launches
- Distance between launch site and proximities:
  - Added lines to indicate distances between launch site and the closest coastline, railway, road, and city

Launch Sites

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

Folium Maps Notebook (GitHub)

Build a Dashboard with Plotly Dash

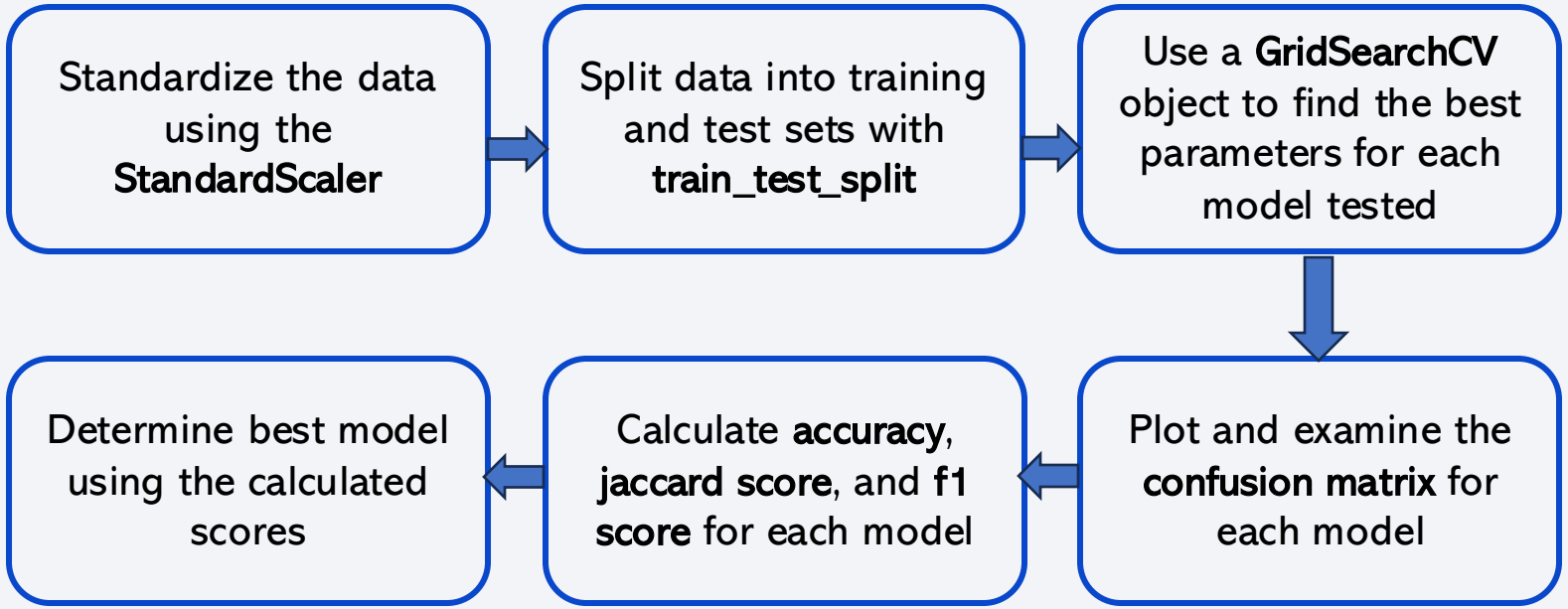
- Two interactive elements to filter data as required:
  - Dropdown allowing display of (1) data from all sites, or (2) from a specific site
  - Slider allowing filtering of payload mass range displayed on scatter plot
- Two plots showing filtered data:
  - **Pie chart** displaying either
    - Proportion of successful launches from all sites (1)
    - Proportion of successful and unsuccessful launches from a specific site (2)
  - **Scatter plot** displaying the relationship between specified payload mass range and success rate for different booster versions

Dashboard Script (GitHub)

Predictive Analysis (Classification)

- Column containing landing outcome labels ("Class") was used as the classifier data

- All model parameters were optimised using GridSearch with a cv value of 10



Models Tested

- Logistic Regression (LR)
- Support Vector Machine (SVM)
- Decision Tree (DT)
- K-Nearest Neighbour (KNN)

Predictive Analysis Notebook (GitHub)

Results Overview

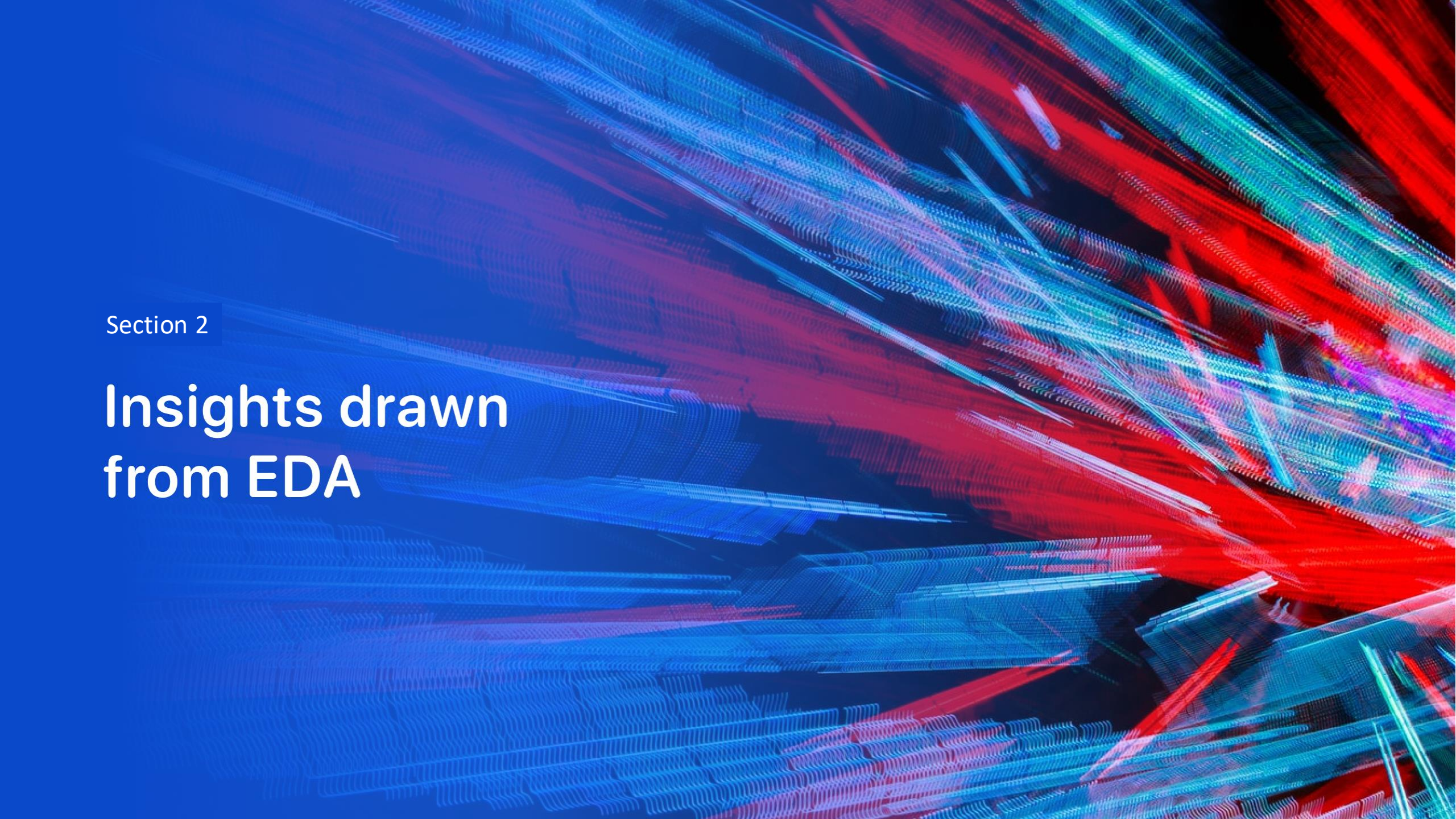
Now that the Methodology has been outlined, the results will be presented as follows:

Section 2: Insights Drawn from EDA

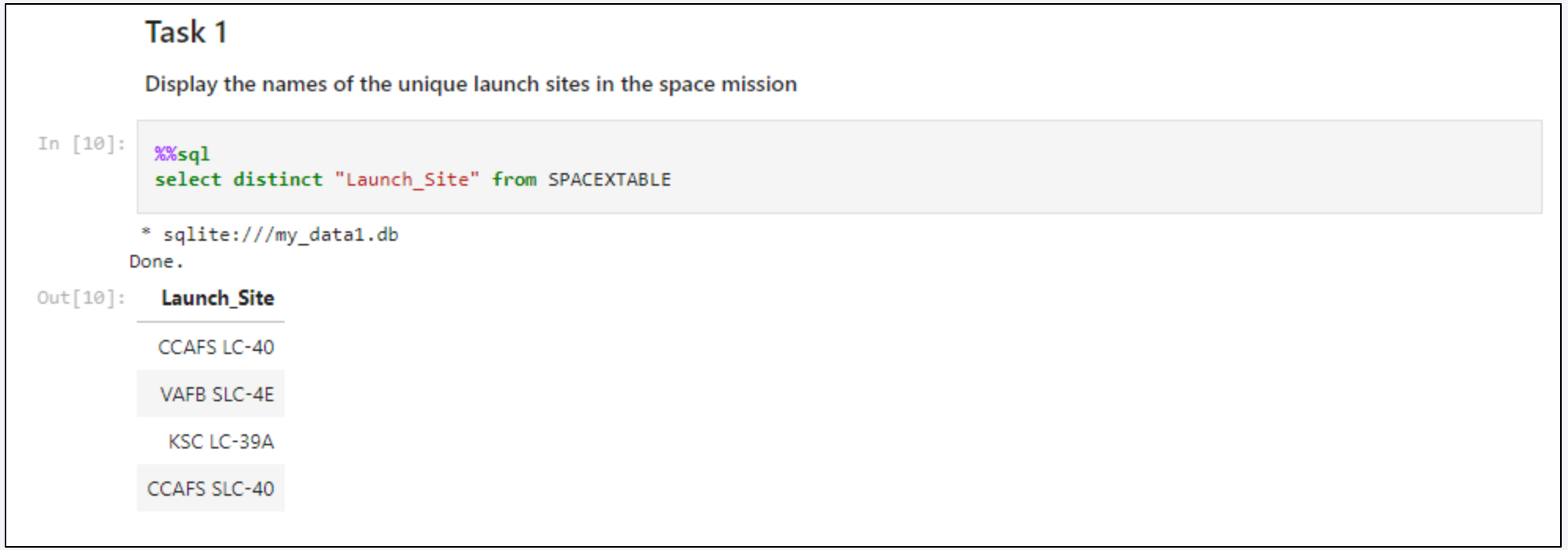
Section 3: Launch Sites Proximity Analysis

Section 4: Plotly Dashboard Analysis

Section 5: Predictive Analysis (Classification)


Section 2
Insights drawn from EDA

SQL - All Launch Site Names



Task 1

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

In [10]:
```
%%sql
select distinct "Launch_Site" from SPACEXTABLE
```

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

Out[10]:

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

- Queries database for the unique launch site names

SQL - Launch Site Names Begin with 'CCA'

- Queries database for 5 records where the launch site begins with 'CCA'

Task 2

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

In [11]:
```
%%sql
select * from SPACEXTABLE where "Launch_Site" Like "CCA%" LIMIT 5
```

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

Out[11]:

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

SQL - Total Payload Mass

Task 3

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

In [12]:
```
%%sql
select sum("PAYLOAD_MASS__KG_") as "Total Payload Mass" From SPACEXTABLE where Customer =  "NASA (CRS)"
```

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

Out[12]:

Total Payload Mass
45596

- Queries database and calculates the total payload carried by boosters from NASA (CRS)

SQL - Average Payload Mass by F9 v1.1

Task 4

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

In [13]:
```
%%sql
select avg("PAYLOAD_MASS__KG_") as "Mean Payload Mass" from SPACEXTABLE where "Booster_Version" Like "F9 v1.1%"
```

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

Out[13]:

Mean Payload Mass
2534.6666666666665

- Queries database and calculates the average payload mass carried by booster version F9 v1.1

SQL - First Successful Ground Landing Date

Task 5

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

Hint:Use min function

In [14]:
```
%%sql
select min(Date) as "First Successful Landing" from SPACEXTABLE where "Landing_Outcome" = "Success (ground pad)"
```

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

Out[14]:

First Successful Landing
2015-12-22

- Queries database for the dates of the first successful landing outcome on ground pad

SQL - Successful Drone Ship Landing with Payload between 4000 and 6000

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

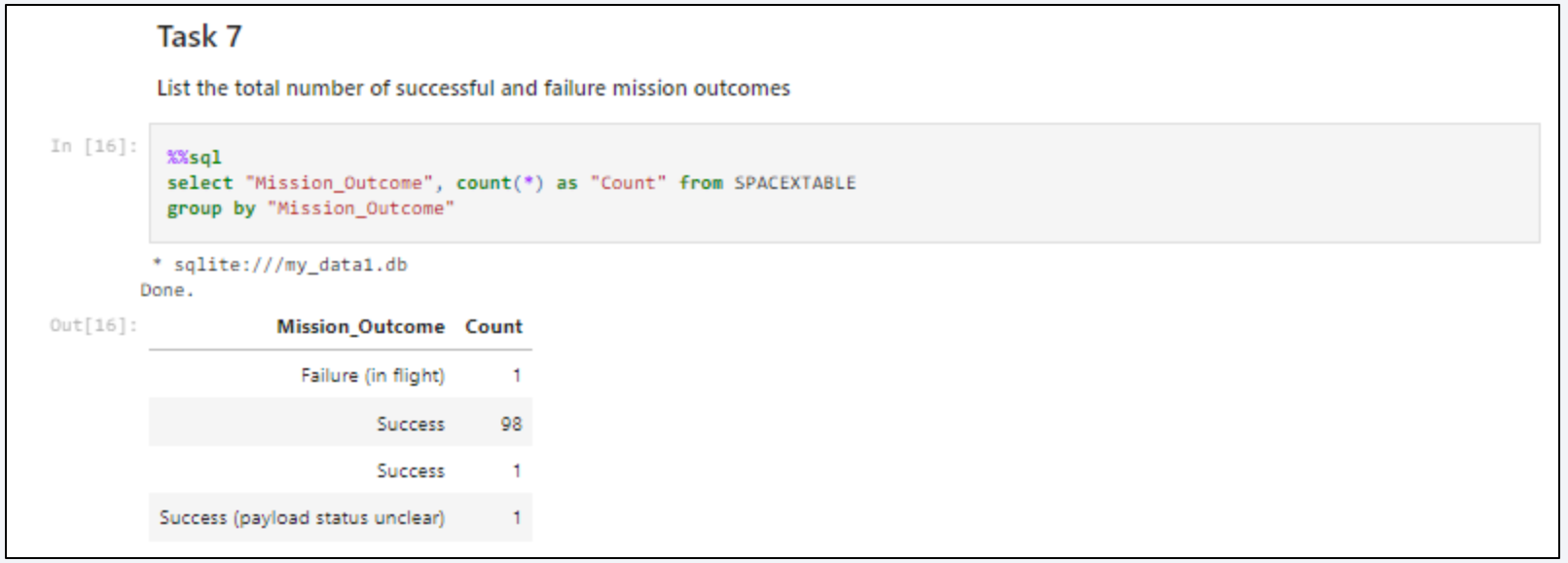
```
[15]
%%sql
select "Booster_Version" from SPACEXTABLE where "Landing_Outcome" = "Success (drone ship)" and "PAYLOAD_MASS__KG_" between 4000 and 6000
```

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

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

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

SQL - Total Number of Successful and Failure Mission Outcomes



Task 7

List the total number of successful and failure mission outcomes

In [16]:

```
%%sql
select "Mission_Outcome", count(*) as "Count" from SPACEXTABLE
group by "Mission_Outcome"
```

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

Out[16]:

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

- Query database and calculate the total number of successful and failure mission outcomes

SQL - Boosters Carried Maximum Payload

- Query database and list the names of the booster which have carried the maximum payload mass

Task 8

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

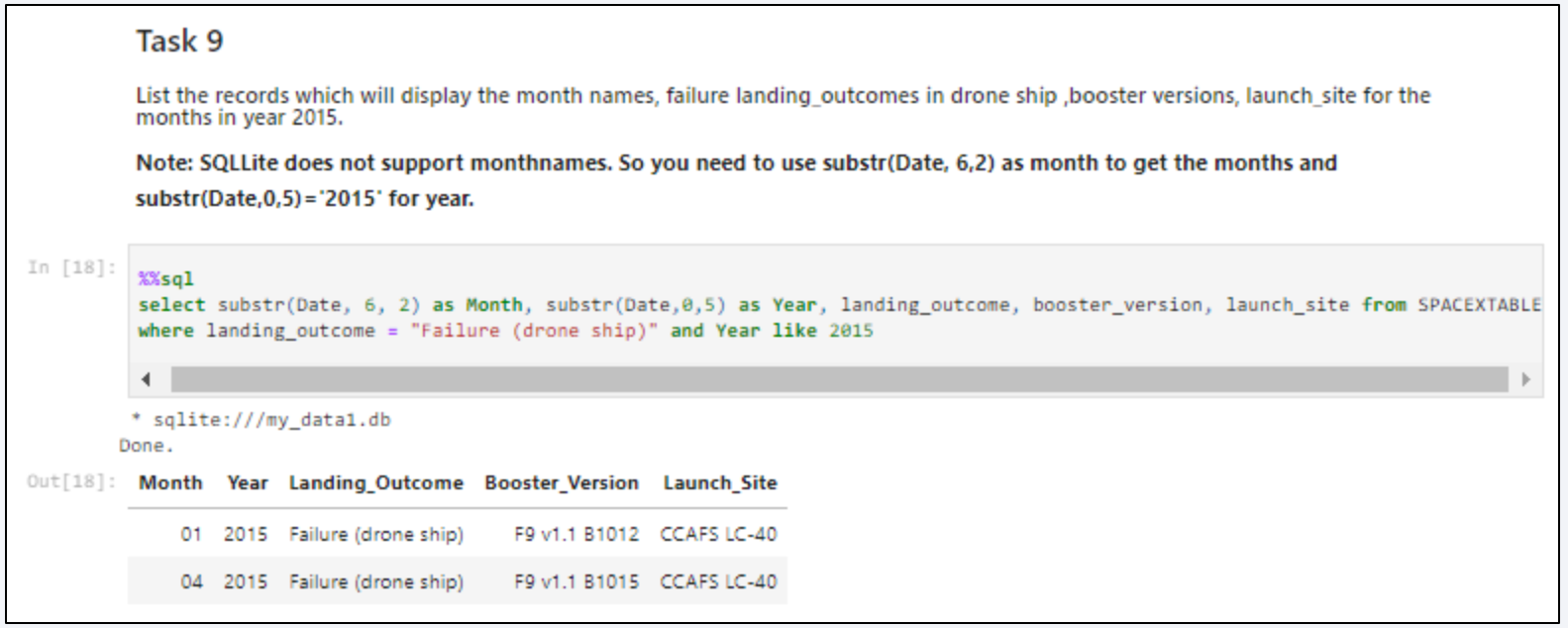
In [17]:
```
%%sql
select booster_version from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)
```

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

Out[17]:

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

SQL - 2015 Launch Records



Task 9

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

In [18]:
```
%%sql
select substr(Date, 6, 2) as Month, substr(Date,0,5) as Year, landing_outcome, booster_version, launch_site from SPACEXTABLE
where landing_outcome = "Failure (drone ship)" and Year like 2015
```

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

Out[18]:

Month	Year	Landing_Outcome	Booster_Version	Launch_Site
01	2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- Query database and list the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

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

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

Task 10

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.

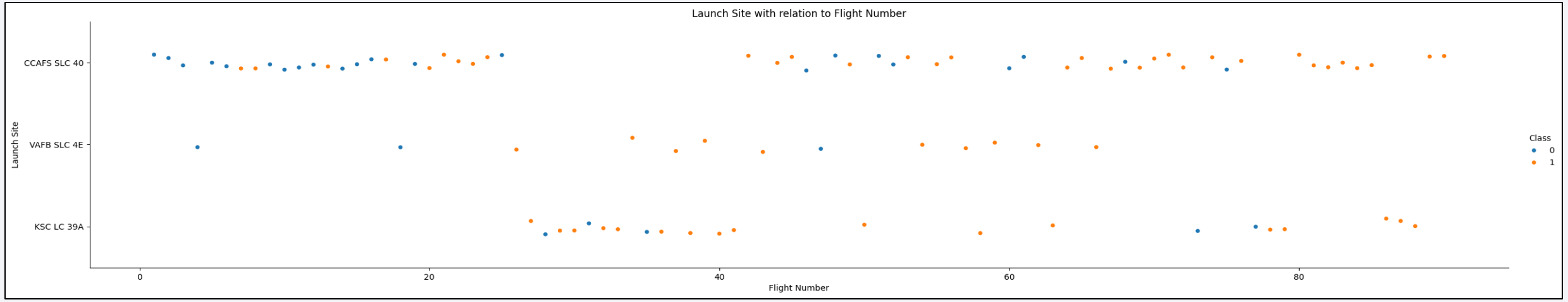
In [20]:
```
%%sql
select "Landing_outcome", count(*) as "Count" from SPACEXTABLE
where Date between "2010-06-04" and "2017-03-20"
group by landing_outcome
order by "Count" desc
```

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

Out[20]:

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

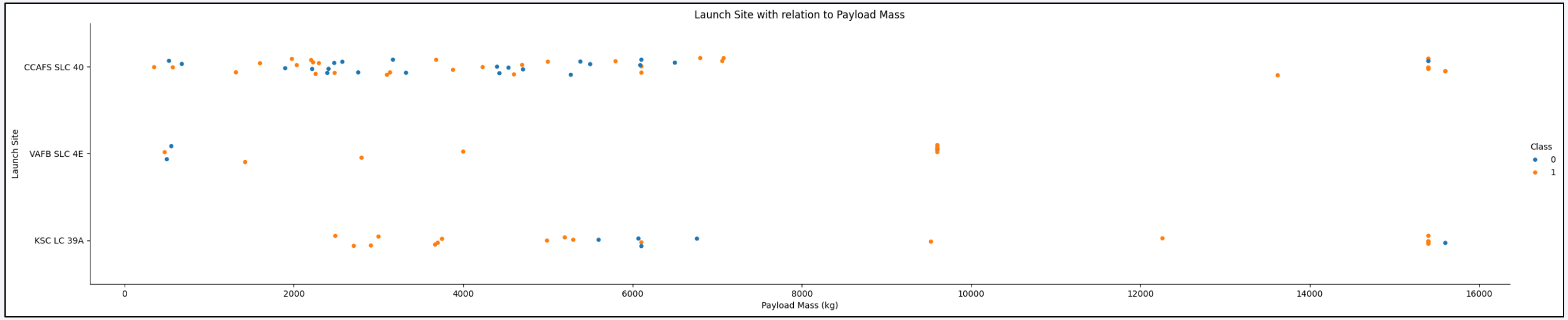
Visualisation - Flight Number vs. Launch Site



The following insights can be gained from this scatter plot:

- Landing success rate has increased over time
- KSC LC 39A has a higher proportional success rate than the other sites
- CCAFS SLC 40 has been operating the longest and has the most launches of all sites

Visualisation - Payload vs. Launch Site



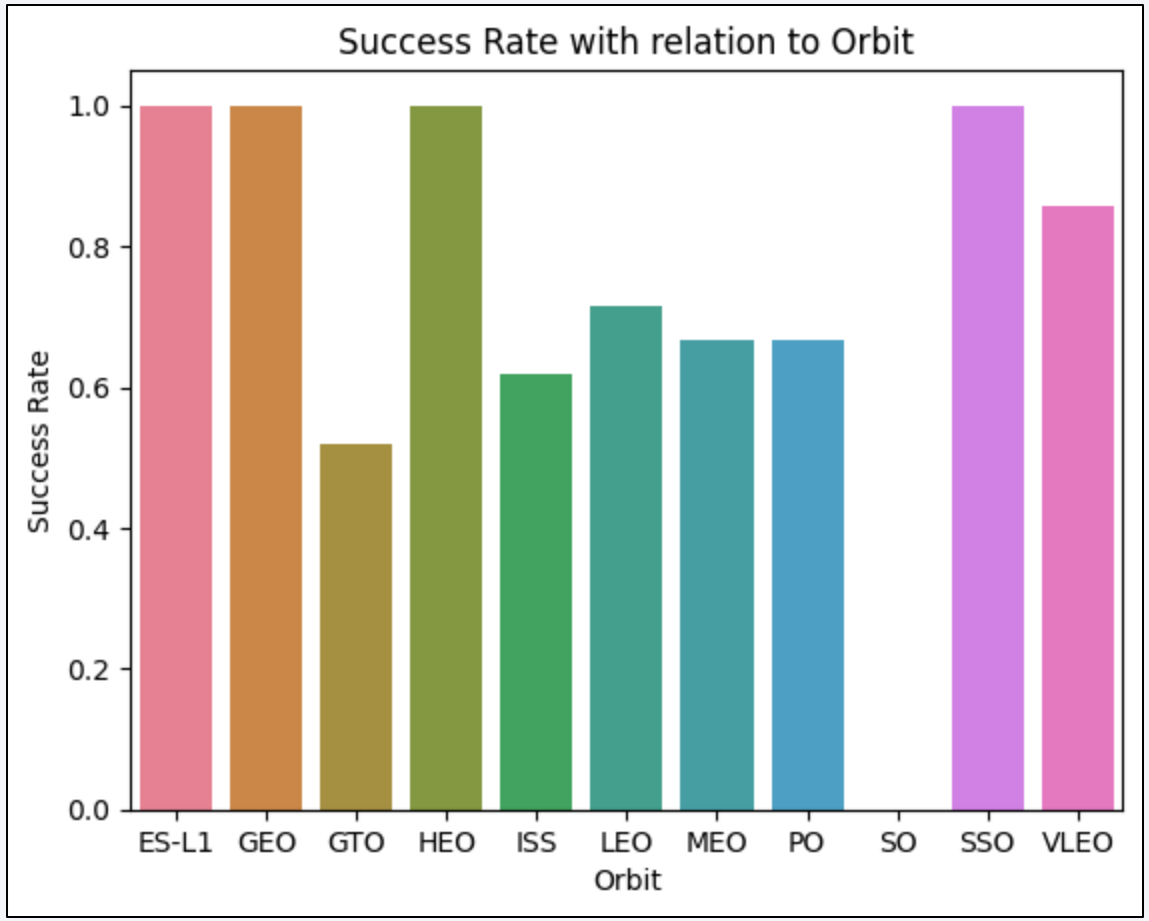
The following insights can be gained from this scatter plot:

- Payload range of 2000-6000 kg has the highest success rate
- Higher payload launches were less common, but have a higher proportional success rate
- Higher payload launches only occur at CCAFS SLC 40 and KSC LC 39A

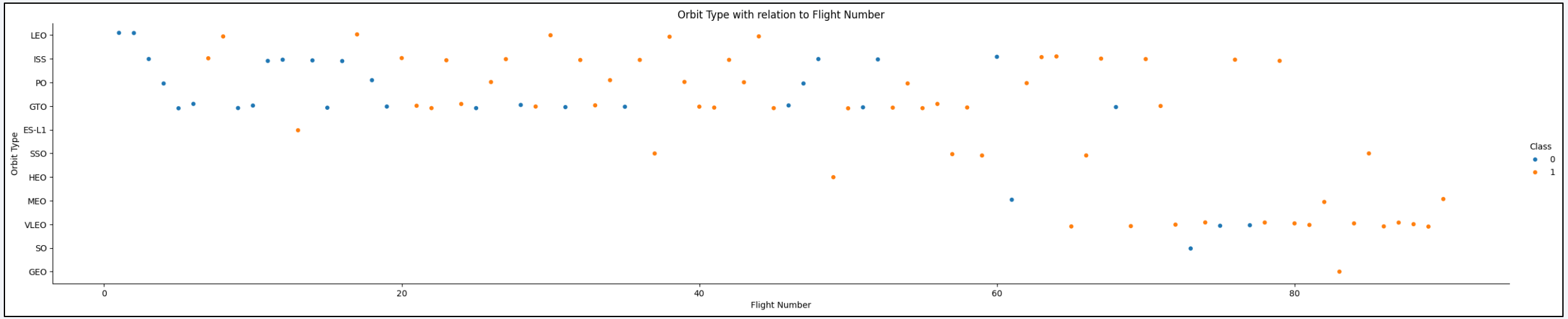
Visualisation - Success Rate vs. Orbit Type

The following insights can be gained from this bar plot:

- ES-L1, GEO, HEO, and SSO orbits all have a mean success rate of 100%
- SO orbit has a mean success rate of 0%
  - Note only 1 SO launch has been attempted
- All other launch types have a mean success rate of over 50%



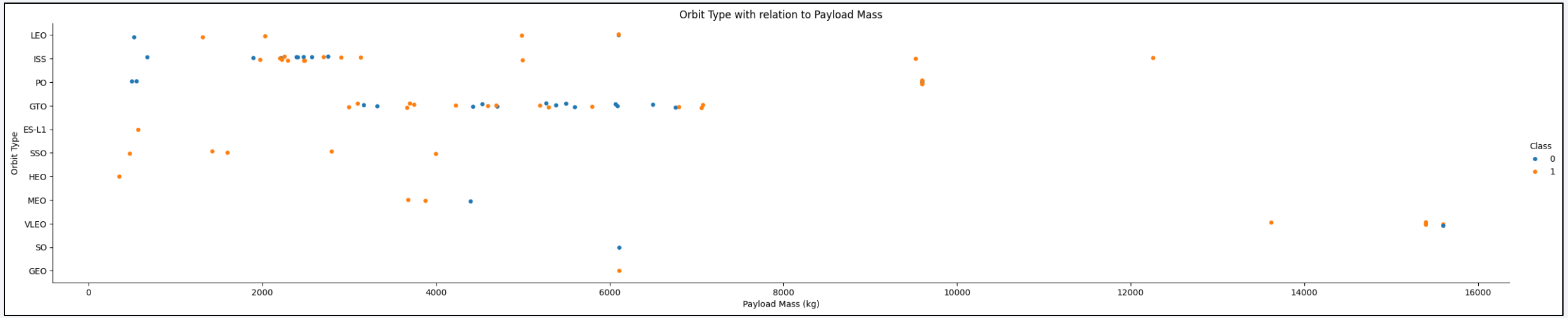
Visualisation - Flight Number vs. Orbit Type



The following insights can be gained from this scatter plot:

- Over time, attempts at the same orbit type tend to yield more successful landings
- Certain launch types (such as GEO, SO, HEO) have only been attempted once, and so their mean success rates on the previous bar chart may not be fully representative

Visualisation - Payload vs. Orbit Type



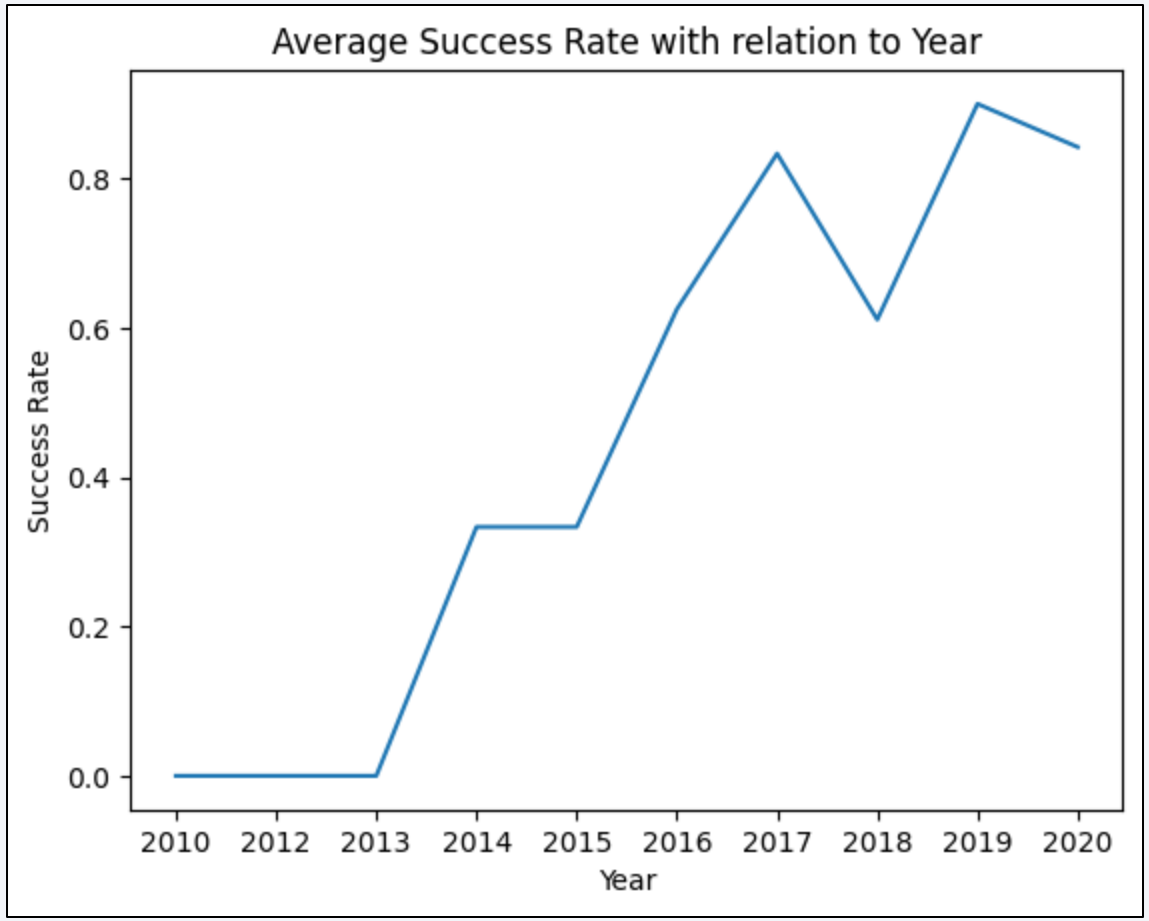
The following insights can be gained from this scatter plot:

- GTO orbits tend to fail with heavier payloads
- ISS and LEO orbits have a higher success rate with heavier payloads

Visualisation - Launch Success Yearly Trend

The following insights can be gained from this line plot:

- As stated in relation to the first scatter plot, success rate of landings has increased with time
- Improvements to landing rate have been fairly similar year on year with the exception of the following periods:
  - 2014 to 2015 (no improvement)
  - 2017 to 2018 (decrease in success rate)
  - 2019 to 2020 (decrease in success rate)



Section 3

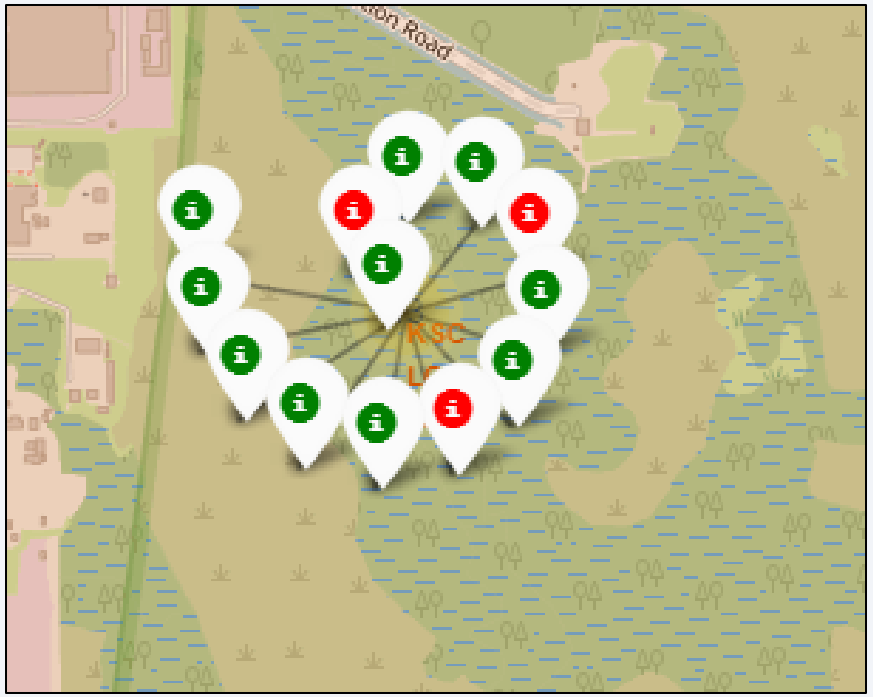
Launch Sites Proximities Analysis

Folium Maps – Launch Site Locations

- This screenshot displays the positioning of the launch sites
- Two key details can be found with relation to launch site positioning:
  - **Launch sites are located close to the equator.** Launching from as close to the equator as possible allows the ship to get the greatest launch speed possible, as the earth is spinning fastest around the equator
  - **Launch sites are located close to the coast.** This is done so that any damage to surrounding land masses are minimised in the event of an unsuccessful launch/landing.

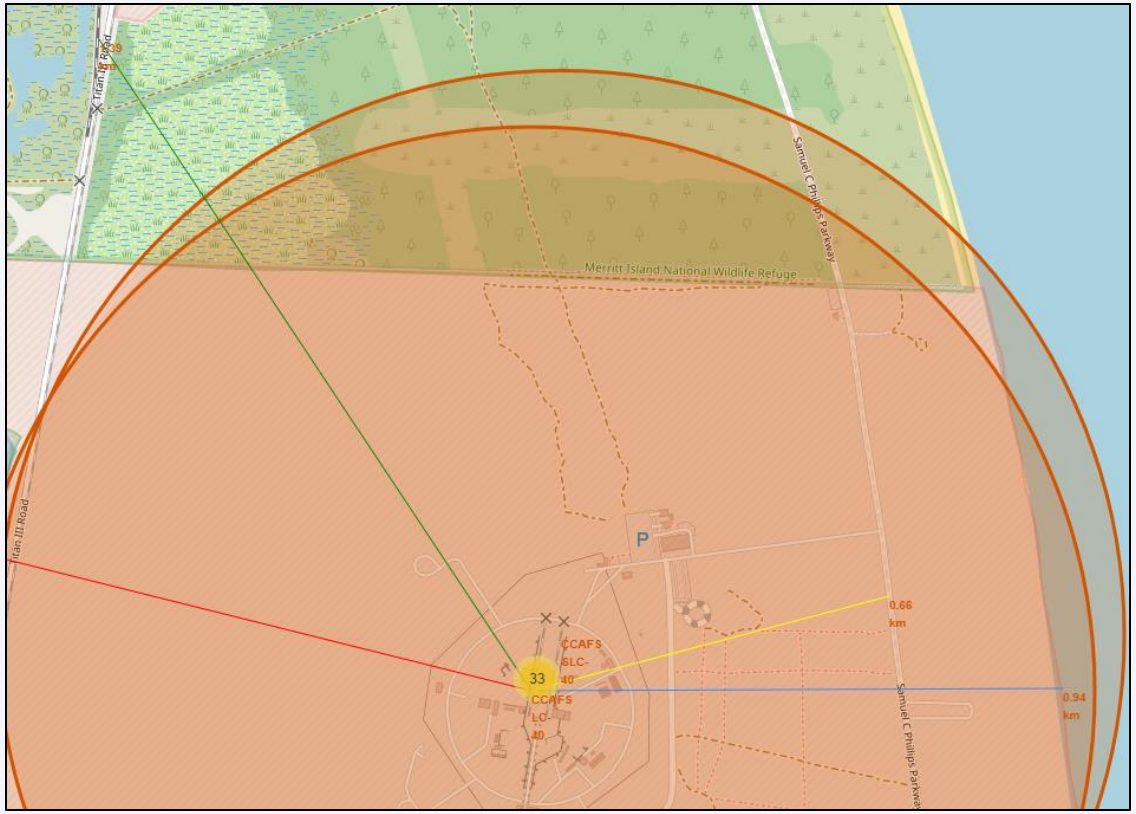
Folium Maps – Launch Clusters

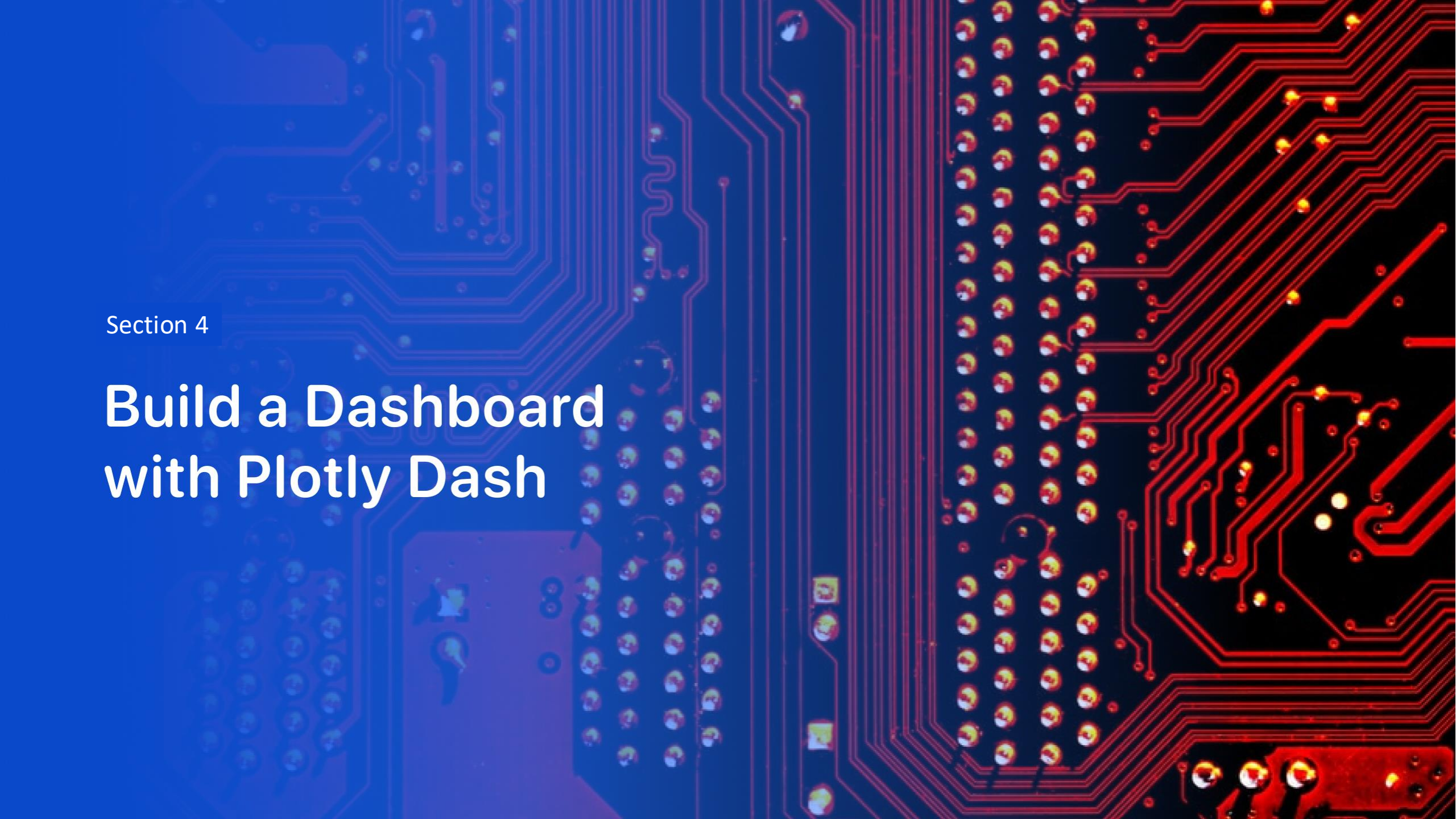
- Clusters showcasing successful and unsuccessful landings at each site (colour-coded)
  - Successful landings are indicated by Green markers
  - Unsuccessful landings are indicated by Red markers
- Launch site KSC LC-39A has the highest success rate for landings



Folium Maps - Proximities

- The distance between a launch site and its proximities (nearest coast, city, railway and road) can be calculated using the respective coordinates
- CCAFS LC-40 is within close proximity of a coast, a railway and a road (blue, green and yellow lines), while the nearest city, Titusville, is much further away (red line, not visible in screenshot)
    - Rail and road proximities are important for accessibility to the launch site
    - Keeping a good distance from the nearest city, and being close to the coast, ensure that damage is mitigated if the launch/landing fails

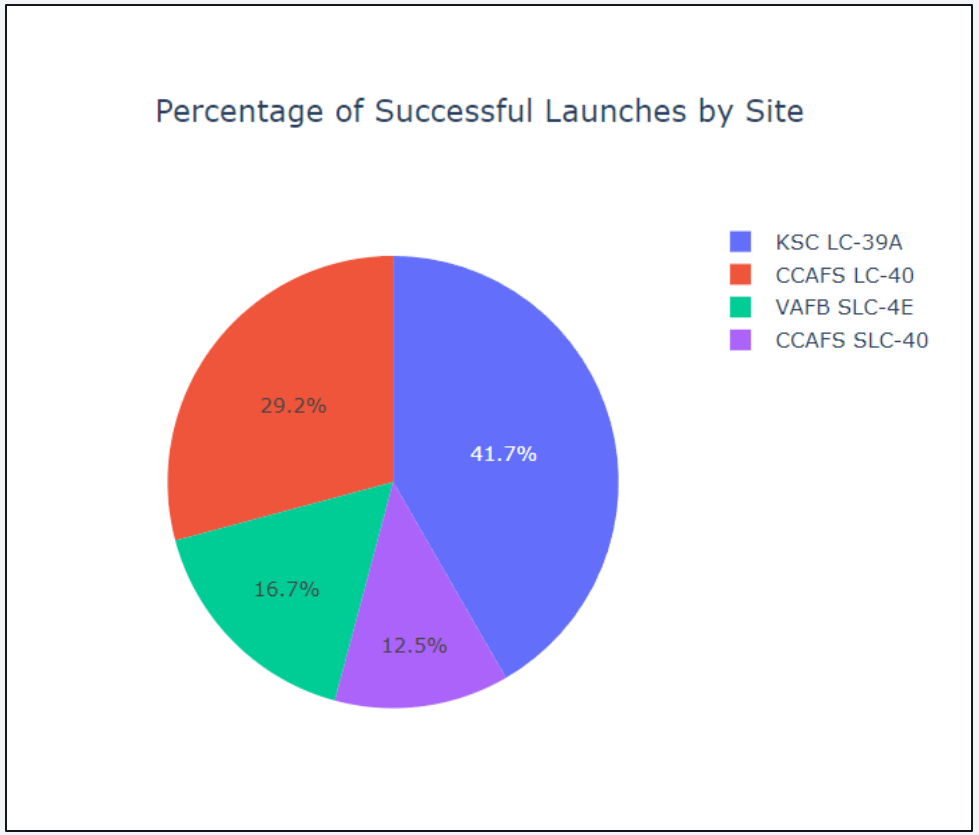



Section 4
Build a Dashboard with Plotly Dash

Dashboard – Successful Landings by Site

- The pie chart from the dashboard allows for direct comparison of the proportion of successful landings from all launch sites

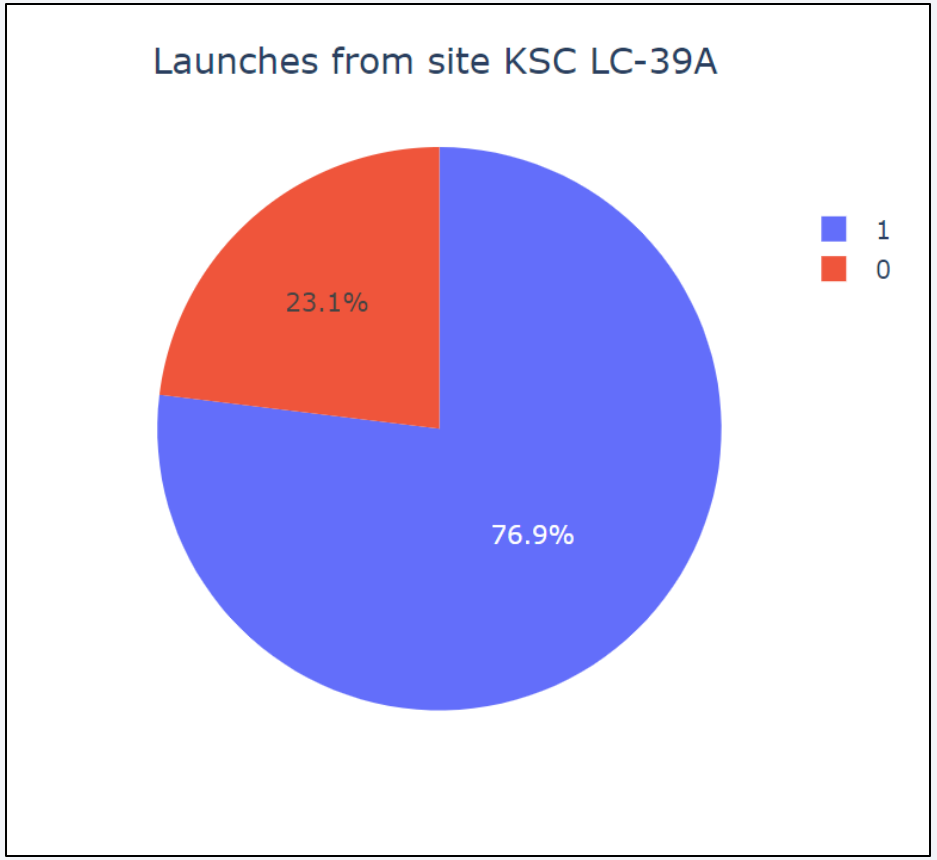
- KSC LC 39-A has the highest proportion of successful landings (as seen previously on the map), with 41.7% of all successful landings coming from this site



Dashboard – KSC LC-39A

- The dropdown menu on the dashboard can be used to choose a single site, giving a breakdown of the successful and unsuccessful landings from a particular site

- For KSC LC-39A, which we established had the highest success rate of all the sites, we can see that over 75% of landings from this site have been successful
  - 10 successful landings
  - 3 failed landings



Dashboard – Payload Mass

- The scatter plot shows the relationship between the payload mass and the launch success rate, as well as giving an insight into how the booster version affects the launch outcome

- The highest success rate occurs between 2000 and 6000 kg payload mass, with most successes using the FT booster version

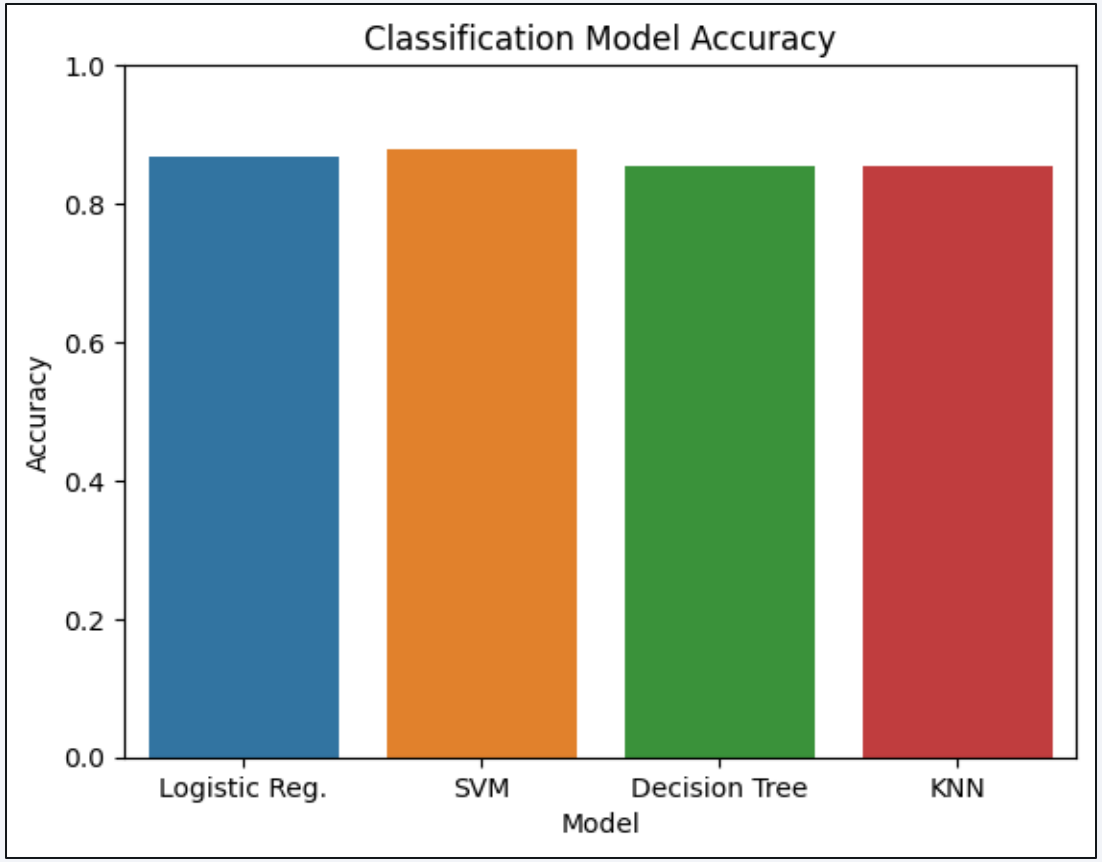
Section 5

Predictive Analysis (Classification)

Classification Accuracy

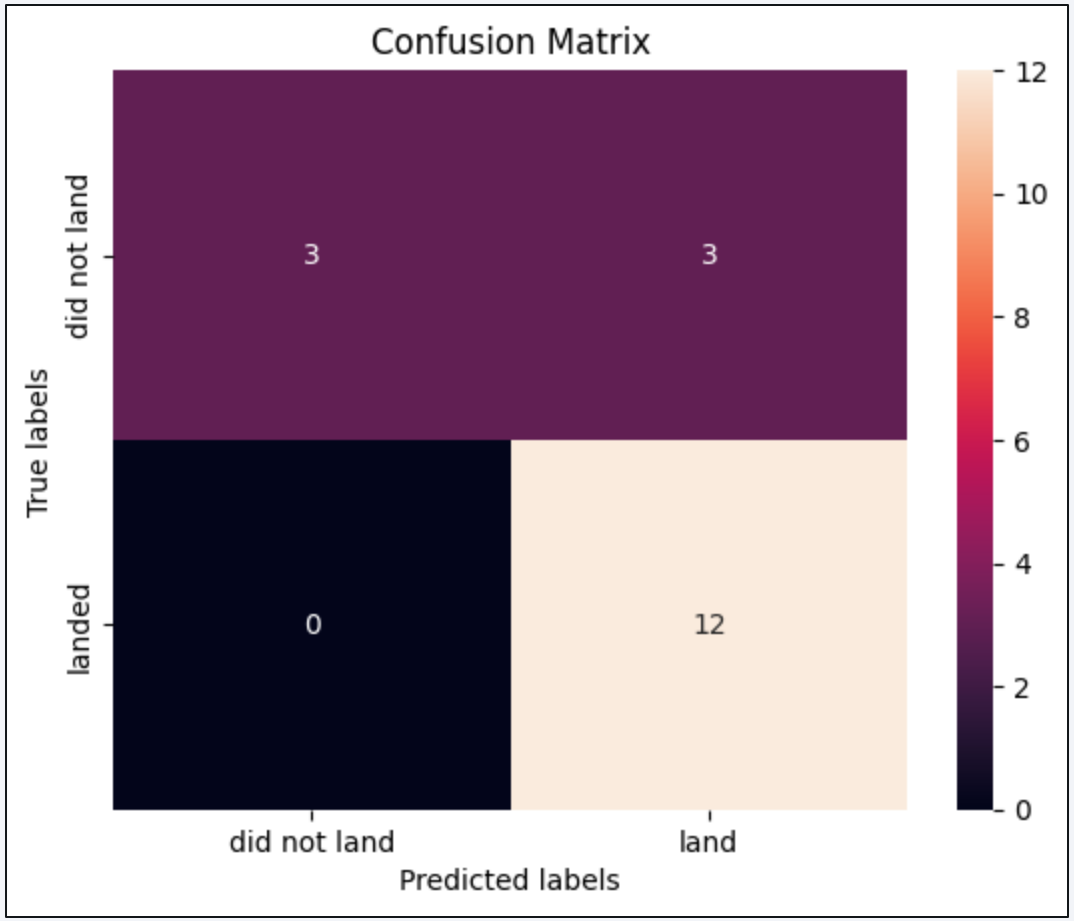
- The bar plot shows that the SVM showed the highest accuracy score of all the models

- The table also shows that the SVM outperformed the other models in Jaccard and F1 score metrics as well
  - These scores can be seen in Appendix A



Confusion Matrix

- The confusion matrix for the SVM is shown on the right
- All cases in which the rocket landed were predicted correctly (True Positive), with no False negatives being predicted
- The model did not correctly predict cases in which the rocket would not land, with 50% of unsuccessful landings being flagged as False Positives



Conclusions

From the results gathered, we can answer the business questions posed at the beginning of this report:

- Landing success rate has increased generally over time
  - This suggests an increase in experience and expertise
- KSC LC-39A has the highest landing success rate of all sites
  - 76.9% of landings at this site have been successful
- The optimum payload range for a successful landing is 2000 – 6000 kg
  - The FT booster version had the greatest contribution to successes in this payload range

Conclusions

- ES-L1, GEO, HEO, and SSO orbits all have a 100% mean landing success rate
- The Support Vector Machine (SVM) classification model had the highest accuracy when predicting successful/unsuccessful landings
  - SVM has an accuracy score of 87.78% (rounded to two decimal places)

Thank you!

Appendix A – Prediction Model Score Table

	Accuracy	Jaccard	F1
Logistic Reg.	0.866667	0.833333	0.909091
SVM	0.877778	0.845070	0.916031
Decision Tree	0.855556	0.819444	0.900763
KNN	0.855556	0.819444	0.900763