IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Yvonne Booth
27 December 2024

Outline

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

Executive Summary

The objective of this project was to predict if the Falcon 9 first stage will land successfully, using historical SpaceX data. This was achieved using data science techniques detailed in this presentation.

Summary of methodologies

- Data Collection with API
- Data Collection with Web Scraping
- Data Wrangling
- Exploratory Data Analysis using SQL
- Exploratory Data Analysis using Pandas and Matplotlib
- Interactive Visual Analytics with Folium lab
- Interactive Dashboard with Plotly Dash
- Machine Learning Predictive Analysis

Summary of all results

- Insights drawn from exploratory data analysis
- Launch sites proximity analysis
- Insights from an Interactive Dashboard
- Predictive Analysis (Classification)

Introduction

The commercial space age is here, with SpaceX, arguably the most accomplished company, successfully sending spacecraft to the International Space Station, and manned missions and the Starlink internet constellation to space.

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. The significant difference is because SpaceX can reuse the first stage.

The ultimate objective of this project is to predict if the Falcon 9 first stage will land successfully. By determining if the first stage will land, the cost of a launch can be determined. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

To achieve the objective the following tasks are required:

- Determine the price of each launch by gathering information about Space X and creating dashboards;
- Determine if SpaceX will reuse the first stage; and
- Train a machine learning model and use public information to predict if SpaceX will reuse the first stage.

Section 1

Methodology

Methodology

Summary

- Data collection methodology:

  SpaceX launch data was gathered from the SpaceX REST API and by web scraping related Wiki pages.

- Perform data wrangling and preliminary Exploratory Data Analysis (EDA)

  EDA was used to find patterns in the data and determine what would be the label for training supervised models.

- Perform EDA using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

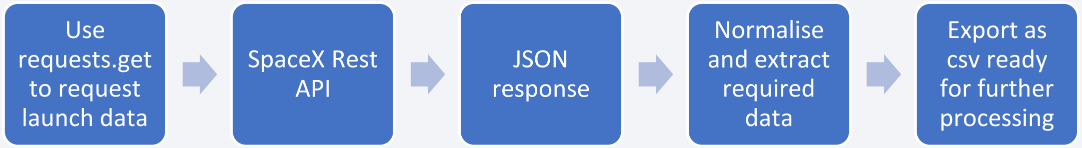
  The model with the best accuracy was determined using the training data. Logistic Regression, Support Vector machines, Decision Tree Classifier, and K-nearest neighbors were tested.

Data Collection – SpaceX API

- SpaceX launch data was requested from the SpaceX Rest API
- The JSON response was decoded and turned into a Pandas dataframe using .json_normalize()
- A subset of the data was extracted.
- The BoosterVersion column was filtered so that only Falcon 9 launches remained in the dataframe.
- A csv was exported of the resulting data.

[IBM-Data-Science-Capstone-Project/jupyter-labs-spacex-data-collection-api.ipynb at main · yvnb1/IBM-Data-Science-Capstone-Project]()

Flowchart of data collection process using Rest API



Use requests.get to request launch data

SpaceX Rest API

JSON response

Normalise and extract required data

Export as csv ready for further processing

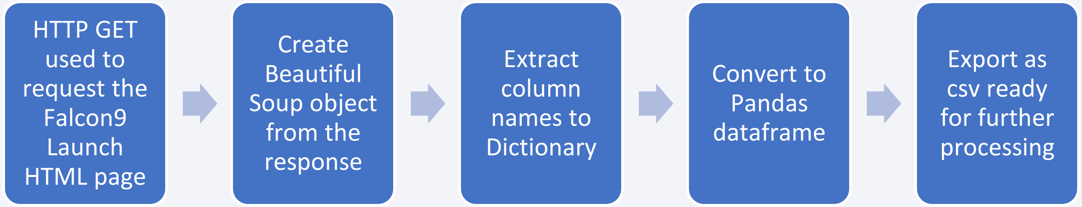
Data Collection - Scraping

Yvonne Booth 27/12/24

- Falcon 9 launch records HTML table was extracted from Wikipedia.
- A Beautiful Soup object was created from the HTML response.
- The table was parsed and converted into a Pandas data frame.
- A csv was exported of the resulting data.

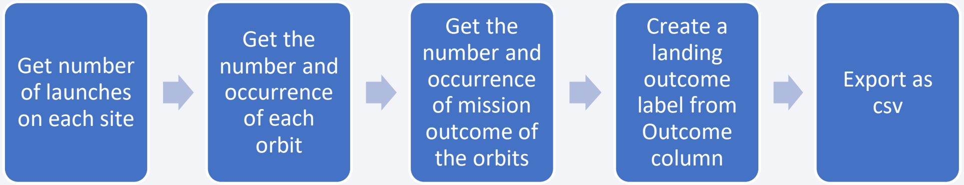
IBM-Data-Science-Capstone-Project/jupyter-labs-webscraping.ipynb at main · yvnb1/IBM-Data-Science-Capstone-Project

Flowchart of data collection process using web scraping


HTTP GET used to request the Falcon9 Launch HTML page
Create Beautiful Soup object from the response
Extract column names to Dictionary
Convert to Pandas dataframe
Export as csv ready for further processing

Data Wrangling

- Using the csv data saved in Data Collection, value_counts() on the column LaunchSite was used to determine the number of launches on each site.
- The method .value_counts() was used to determine the number and occurrence of each orbit in the column Orbit.
- The method .value_counts() was used on the column Outcome to determine the number of landing outcomes.
- The landing outcomes were given a label (Class) of 0 where the outcome was a bad landing and 1 where the outcome was a good landing.


Get number of launches on each site → Get the number and occurrence of each orbit → Get the number and occurrence of mission outcome of the orbits → Create a landing outcome label from Outcome column → Export as csv

IBM-Data-Science-Capstone-Project/labs-jupyter-spacex-Data_wrangling.ipynb at main · yvnb1/IBM-Data-Science-Capstone-Project

Labels (Class) assigned to each landing outcome.

Landing outcome code	Landing outcome description	Class
True Ocean	the mission outcome was successfully landed to a specific region of the ocean	1
False Ocean	the mission outcome was unsuccessfully landed to a specific region of the ocean	0
True RTLS	the mission outcome was successfully landed to a ground pad	1
False RTLS	the mission outcome was unsuccessfully landed to a ground pad	0
True ASDS	the mission outcome was successfully landed to a drone ship	1
False ASDS	the mission outcome was unsuccessfully landed to a drone ship	0
None ASDS and None None	these represent a failure to land	0

EDA with Data Visualization

This charts listed below were created using Exploratory Data Analysis using Pandas and Matplotlib on the dataset.

Scatter Plots

- Relationship between Flight Number and Launch Site
- Relationship between Payload Mass and Launch Site
- Relationship between Flight Number and Orbit Type
- Relationship between Payload Mass and Orbit type

Bar Chart

- Success rates of each Orbit Type were visualized to identify which orbits have the highest success rates.

Line Chart

- Launch success yearly trend to show how success rates changed over the years.

For more information on orbit types see https://earthobservatory.nasa.gov/features/OrbitsCatalog

See IBM-Data-Science-Capstone-Project/edadataviz.ipynb at main · yvnb1/IBM-Data-Science-Capstone-Project for the completed EDA with data visualization notebook.

EDA with SQL

The SpaceX dataset was loaded into the corresponding table in a Db2 database before executing SQL queries in order to complete the assignment tasks below.

- 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).
- Display average payload mass carried by booster version F9 v1.1.
- List the date when the first successful landing outcome in ground pad was acheived.
- 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.
- 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.
- 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.

IBM-Data-Science-Capstone-Project/jupyter-labs-eda-sql-coursera_sqllite.ipynb at main · yvnb1/IBM-Data-Science-Capstone-Project

Build an Interactive Map with Folium

A map was created with Folium and launch sites added using the coordinates in the database. folium.Circle and folium.Marker were used to create the markers showing the launch sites.

Successes and failures of launches for each site were added to the map using a MarkerCluster object to simplify visualization of the multiple launches at each site. Successes were assigned Class 1 with a green marker and Failures were assigned Class 0 with a red marker to make it easy for the user to visually see if launches at each site were successful or not.

Distances from the chosen launch site to marked proximities (closest coastline, city, railway, and highway) were calculated with polylines added to show the distance calculated.

GitHub URL of completed interactive map with Folium map:
[IBM-Data-Science-Capstone-Project/lab_jupyter_launch_site_location.ipynb at main · yvnb1/IBM-Data-Science-Capstone-Project]

Build a Dashboard with Plotly Dash

Yvonne Booth 27/12/24

A Plotly Dash application was built to perform interactive visual analytics on SpaceX launch data in real-time.

A drop-down input component was added to allow the user to choose one or all launch sites.

A pie chart based on the user input was added using a callback function, to show the percentage of successful launches in the case of all sites and the percentages of successful and failed launches in the case of individual sites.

A scatter plot was added using a callback function to show the relationship between payload and launch outcome.

A range slider was added to allow the user to interact with the pie chart and scatter plot and to easily and instantaneously visualize the effect of change in payload.

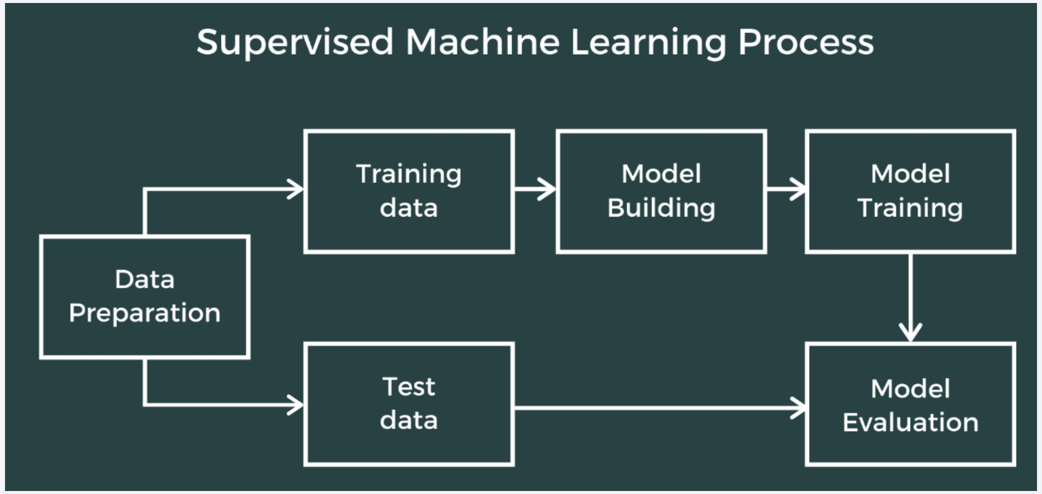
GitHub URL of completed Plotly Dash lab
IBM-Data-Science-Capstone-Project/spacex_dash_app.py at main · yvnb1/IBM-Data-Science-Capstone-Project

Predictive Analysis (Classification)

Yvonne Booth 27/12/24

Machine learning was used to determine if the first stage of Falcon 9 will land successfully. Data was split into training data and test data to find the best Hyperparameter for Support Vector Machine(SVM), Classification Trees, K-nearest neighbors and Logistic Regression. The method that performed best using test data was then found. The steps are summarized below, with a general flowchart of the process illustrated in the figure below.

- Data was standardized and transformed.
- Data was split into training and testing data.
- Each model was trained and Grid Search performed to find the hyperparameters that allowed each algorithm to perform best.
- Logistic Regression, Support Vector machines, Decision Tree Classifier, and K-nearest neighbors were tested.
- The confusion matrix was output for each algorithm.
- Using the best hyperparameter values, the model with the best accuracy was determined using the training data.


Supervised Machine Learning Process

GitHub URL of the completed predictive analysis lab
[IBM-Data-Science-Capstone-Project/SpaceX_Machine Learning Prediction_Part_5.ipynb at main · yvnb1/IBM-Data-Science-Capstone-Project]

Results

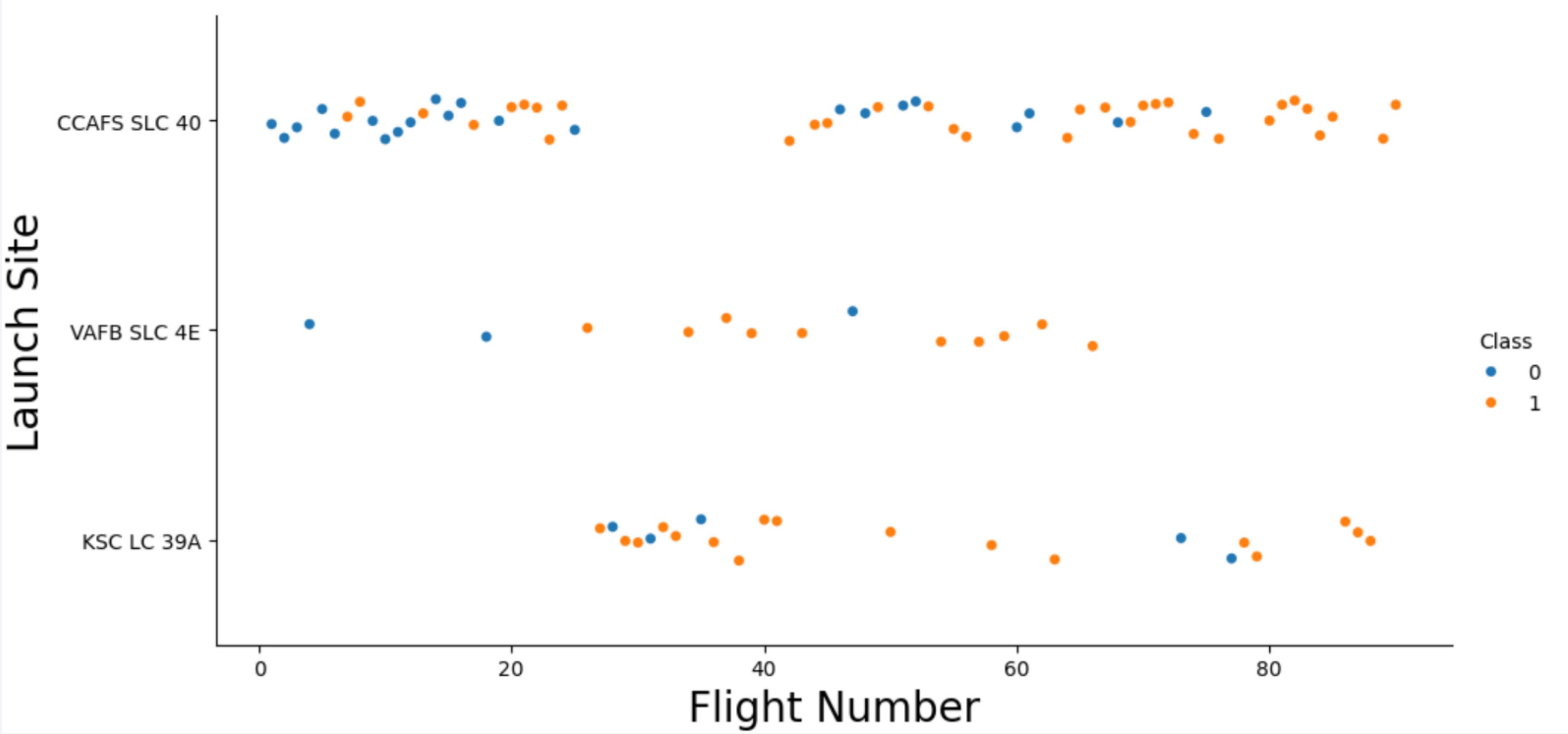
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Yvonne Booth 27/12/24

Section 2

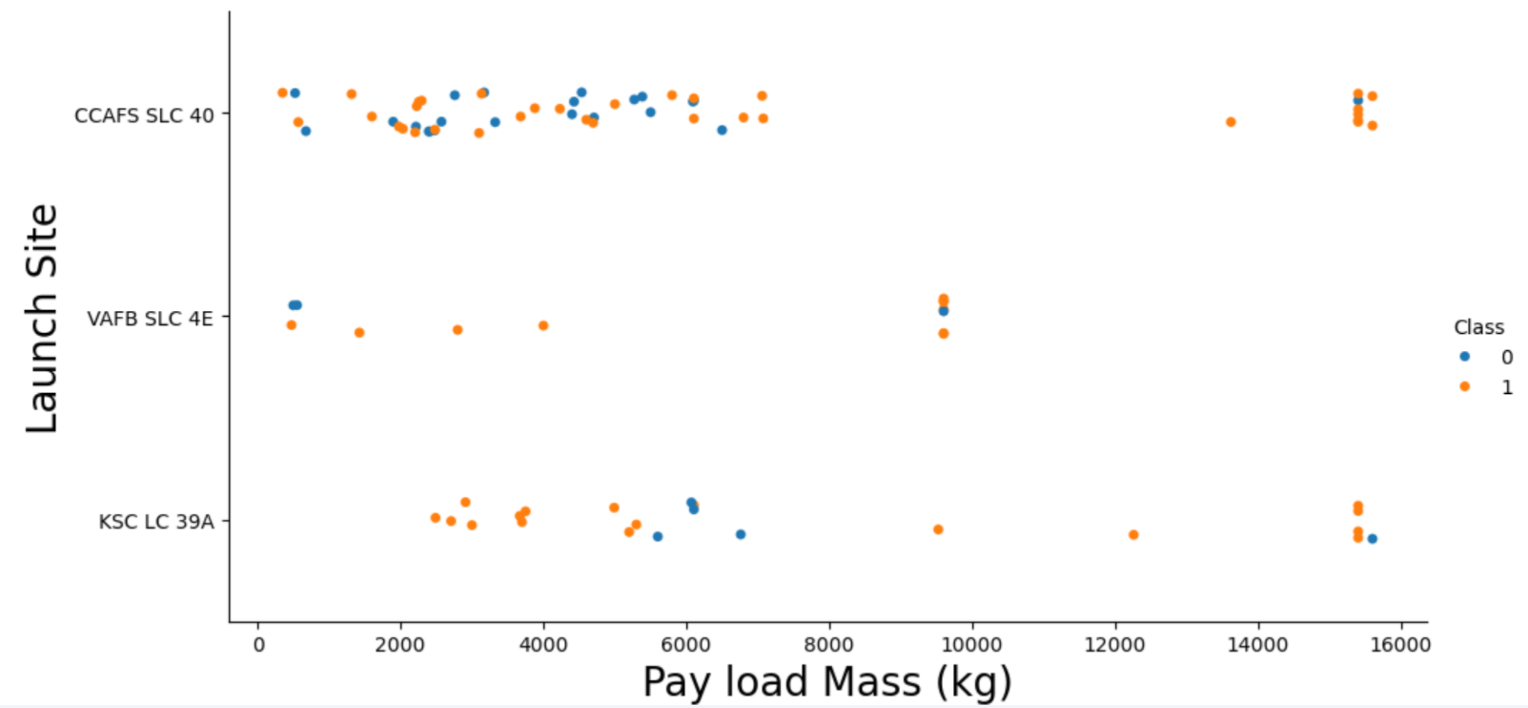
Insights drawn from EDA

Flight Number vs. Launch Site



The scatter plot shows that as flight number increases and experience is gained, the number of successful launches increases.

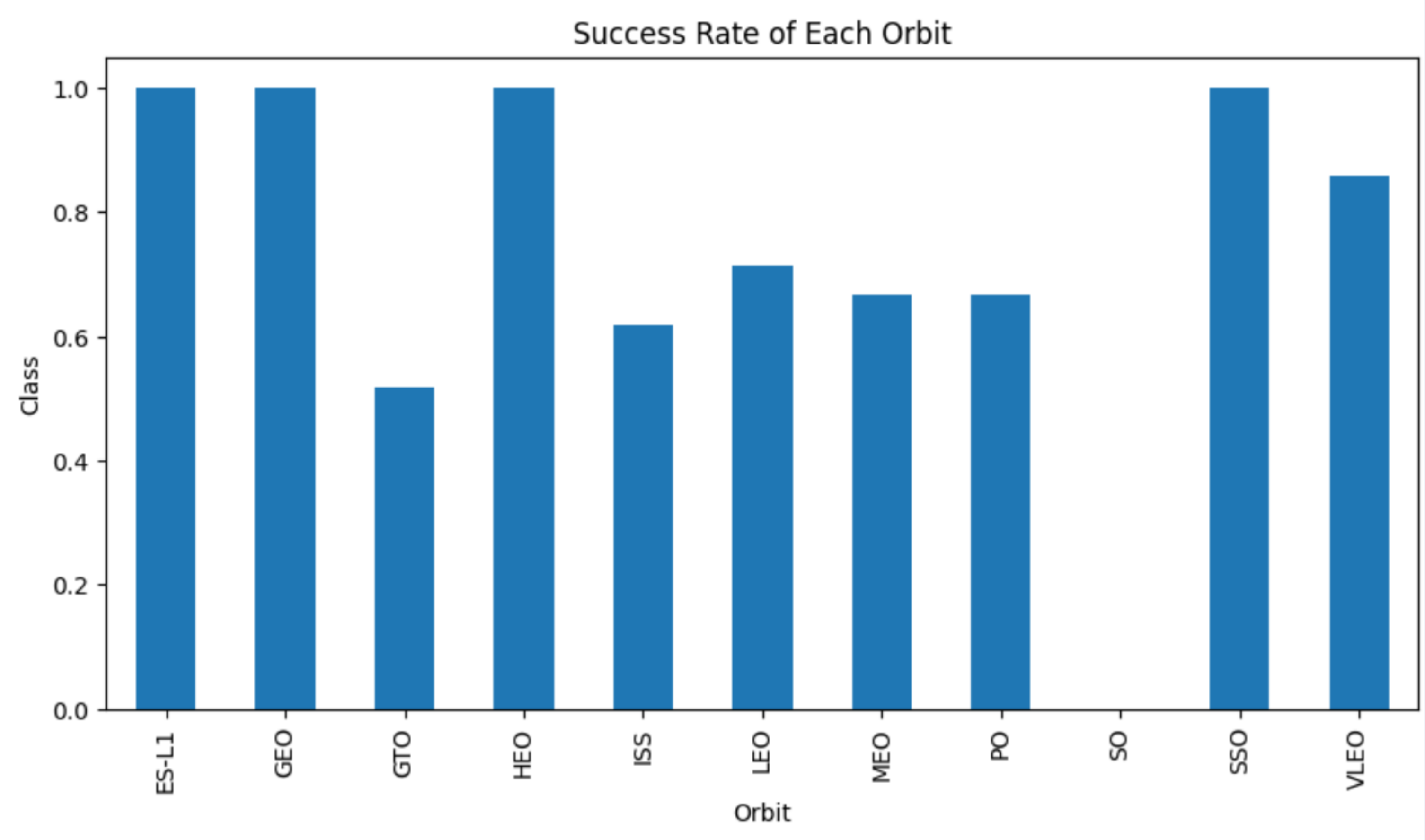
Payload vs. Launch Site



The scatter plot shows that the VAFB SLC 4E launch site had no rockets launched with a payload mass greater than 10000kg. At site CCAFS SLC 40 the number of successful launches appears to increase with increased payload. There is no clear relationship between launch outcome and payload mass at the other sites.

Success Rate vs. Orbit Type

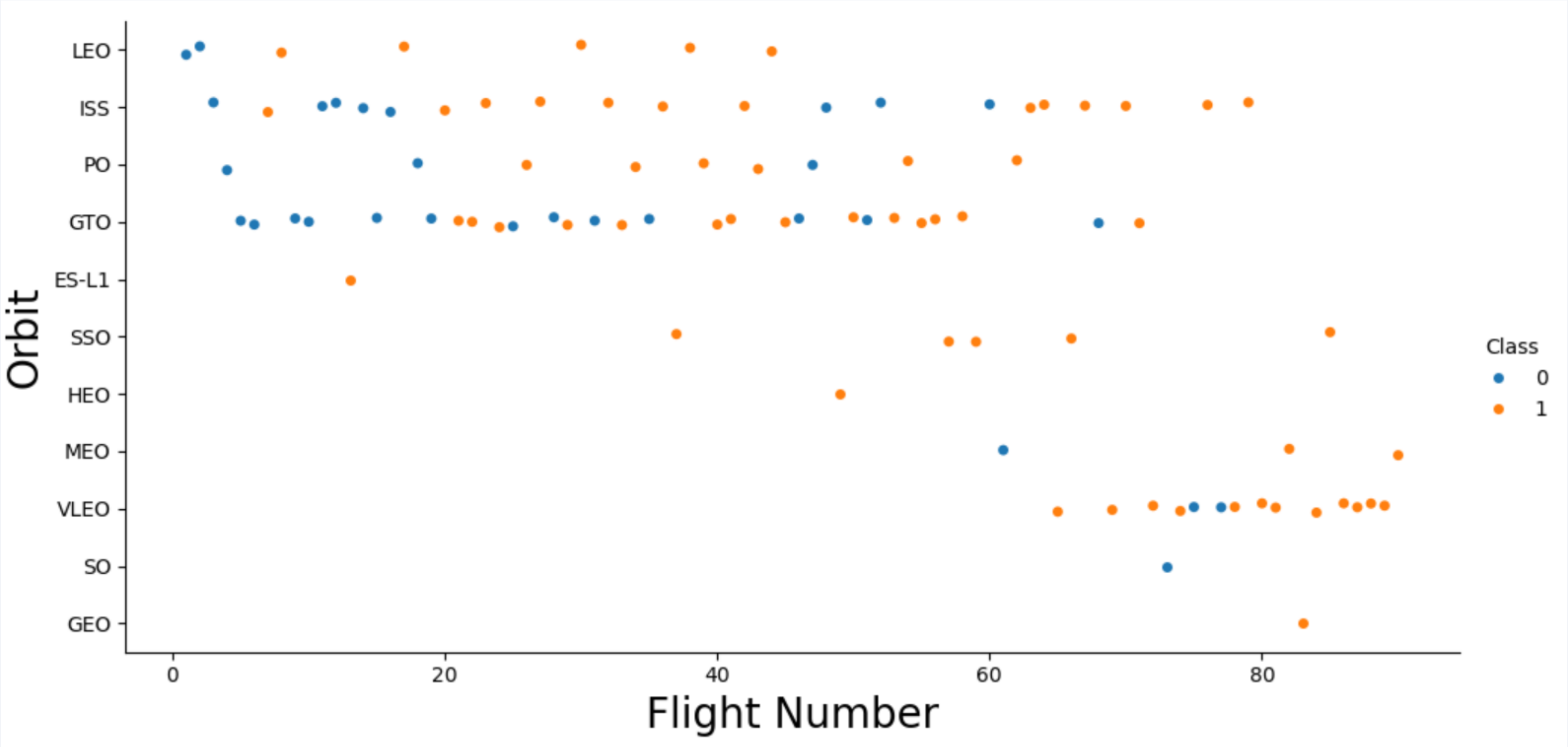
Yvonne Booth 27/12/24



The bar chart shows the most successful orbits being ES-L1, GEO, HEO and SSO.

Flight Number vs. Orbit Type

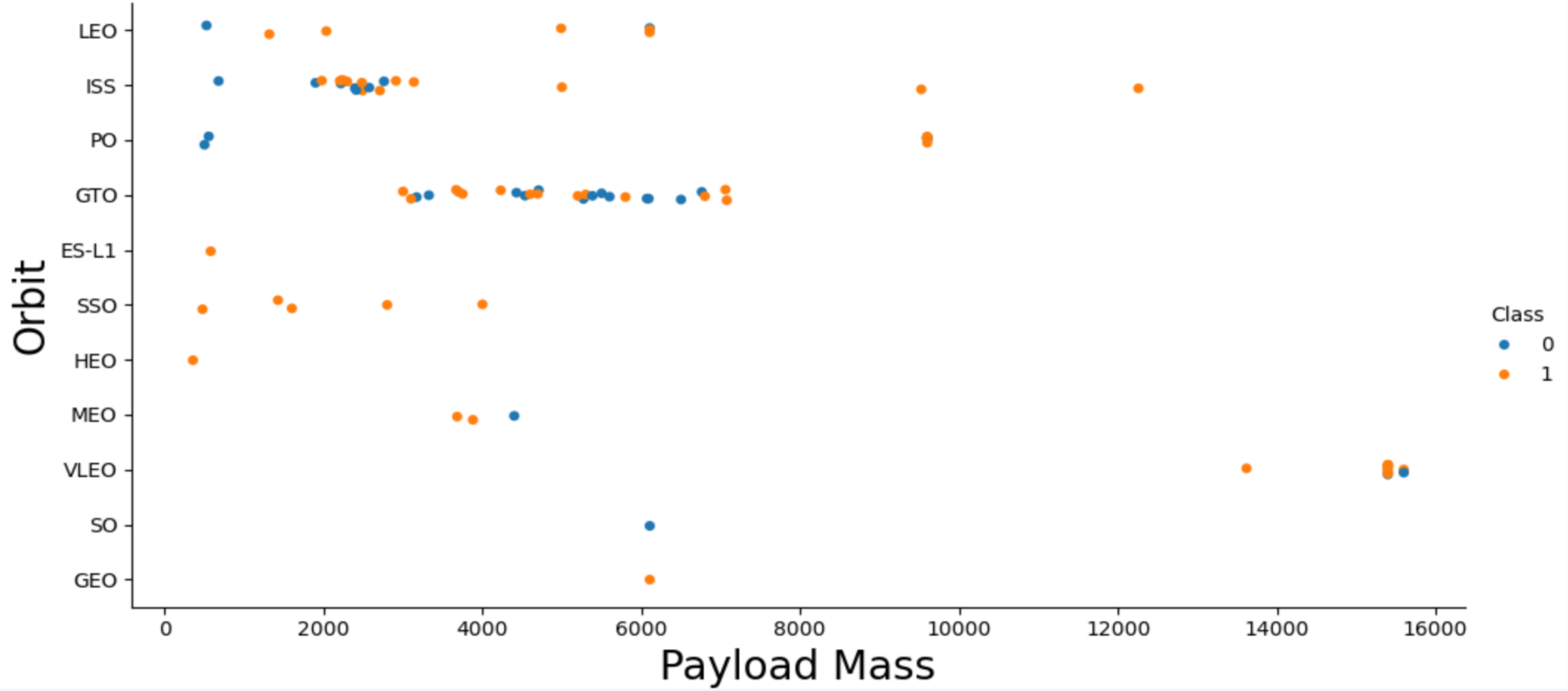
Yvonne Booth 27/12/24



The scatter plot shows that success of launches to the LEO orbit seem to be related to the number of flights with more success as flights increase. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.

Payload vs. Orbit Type

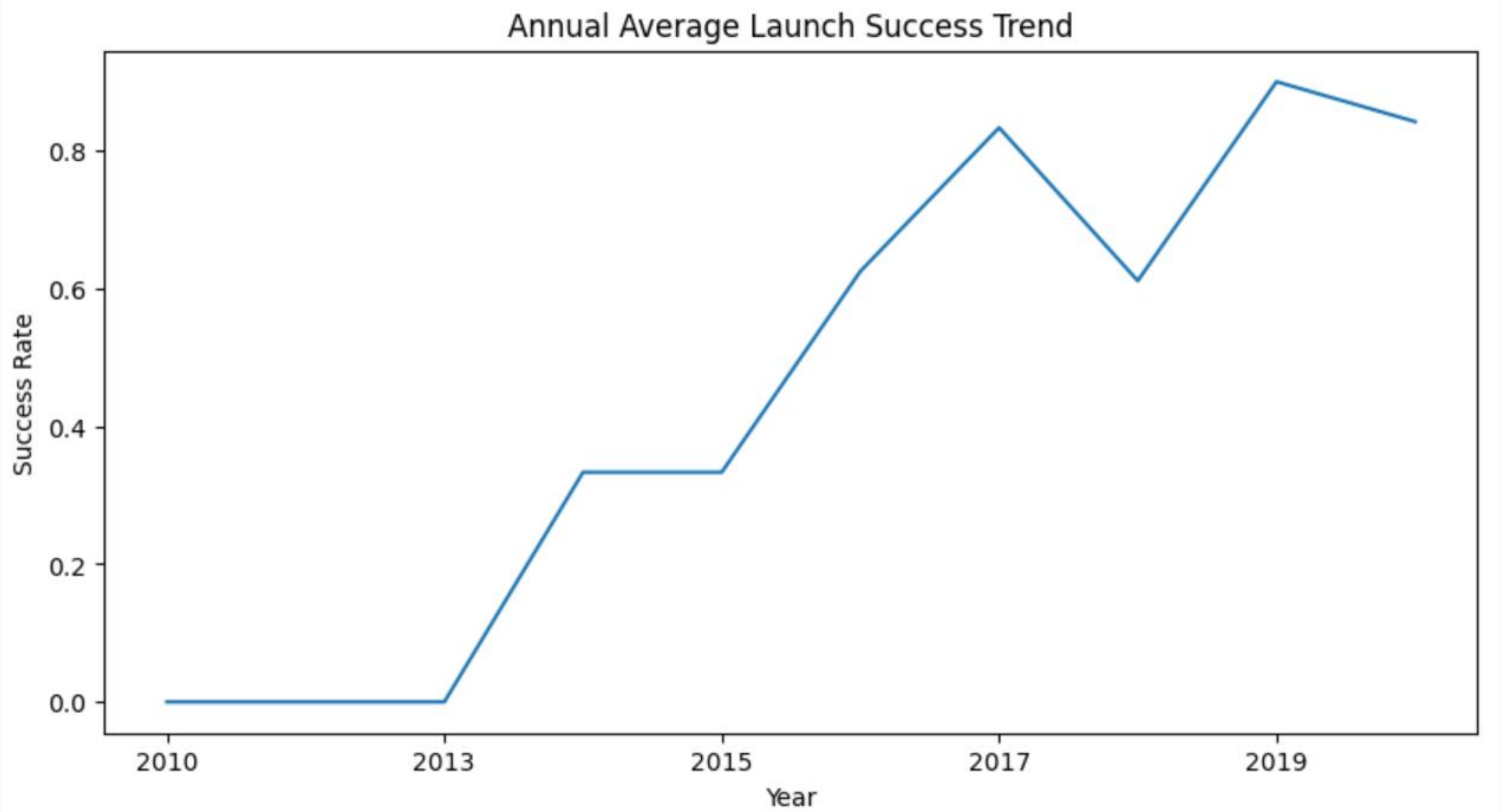
Yvonne Booth 27/12/24



The scatter plot shows with heavy payloads the successful landings are more for Polar, LEO and ISS orbits. However, for GTO, there is no apparent relationship between payload and success of landings.

Launch Success Yearly Trend

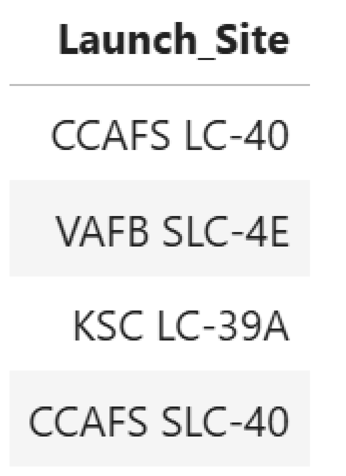
Yvonne Booth 27/12/24



There is an increasing trend in the success rate with time, which is expected as experience increases and lessons are learned from failures.

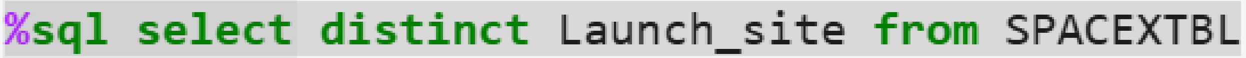
All Launch Site Names

The unique launch site names in the database are as follows:



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

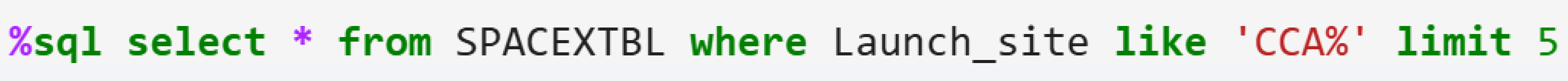
The sites were found using SQL query below which selects all distinct values for launch site from the data table.



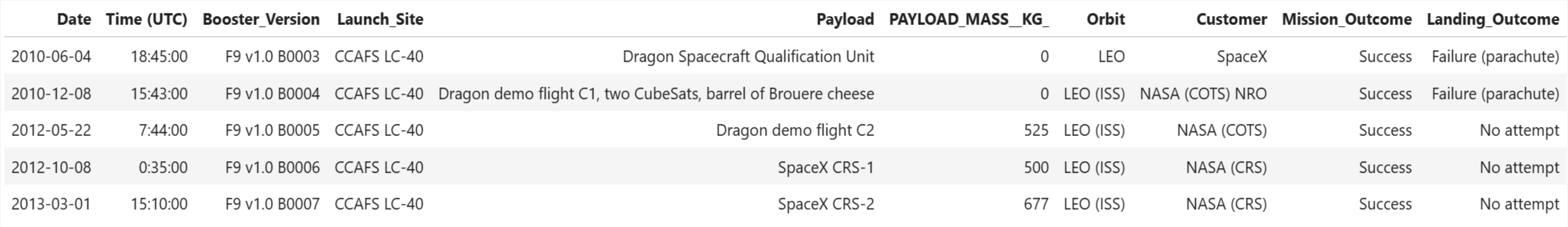
```
%sql select distinct Launch_site from SPACEXTBL
```

Launch Site Names Begin with 'CCA'

5 records where launch sites begin with `CCA` were found using the SQL query below which selects all entries from the table where the launch site begins “CCA”. The results are limited to 5 records and shown in the table.



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

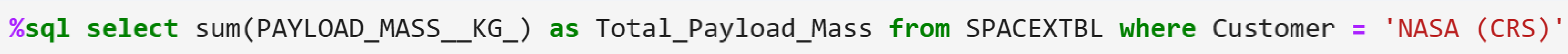


Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight 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

Total Payload Mass

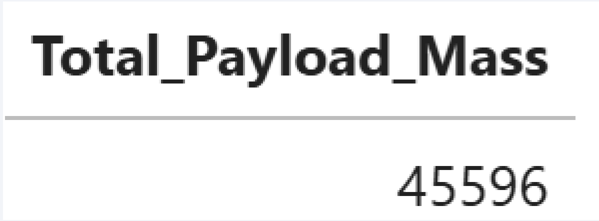
Yvonne Booth 27/12/24

The total payload carried by boosters from NASA was calculated by the SQL query below which sums the payload mass from the table where the customer is NASA.



```
%sql select sum(PAYLOAD_MASS__KG_) as Total_Payload_Mass from SPACEXTBL where Customer = 'NASA (CRS)'
```

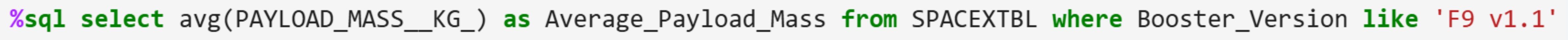
The result in kg is



Total_Payload_Mass
45596

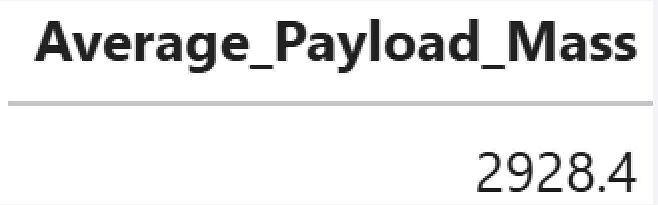
Average Payload Mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1 was calculated using the SQL query below which calculates the average payload mass from the table where the booster version is F9 v1.1.



```
%sql select avg(PAYLOAD_MASS__KG_) as Average_Payload_Mass from SPACEXTBL where Booster_Version like 'F9 v1.1'
```

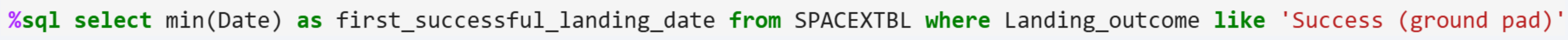
The result in kg is



Average_Payload_Mass
2928.4

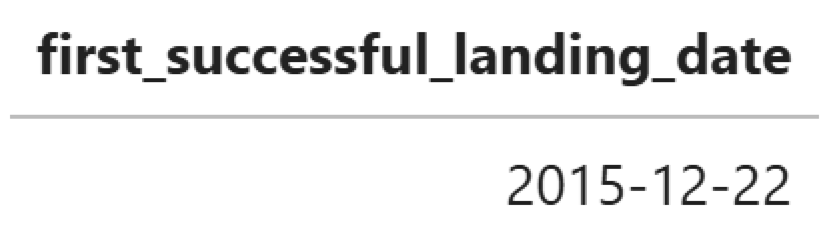
First Successful Ground Landing Date

The dates of the first successful landing outcome on a ground pad were determined using the SQL query below which selects the minimum date where the landing outcome is “Success (ground pad”).



```
%sql select min(Date) as first_successful_landing_date from SPACEXTBL where Landing_outcome like 'Success (ground pad)'
```

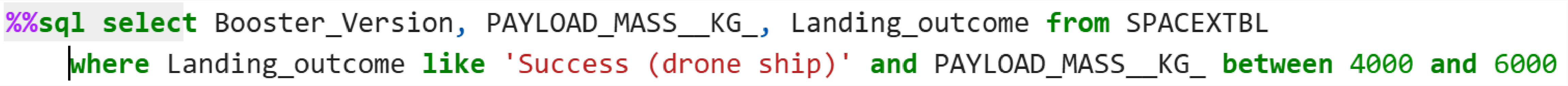
The result is



first_successful_landing_date
2015-12-22

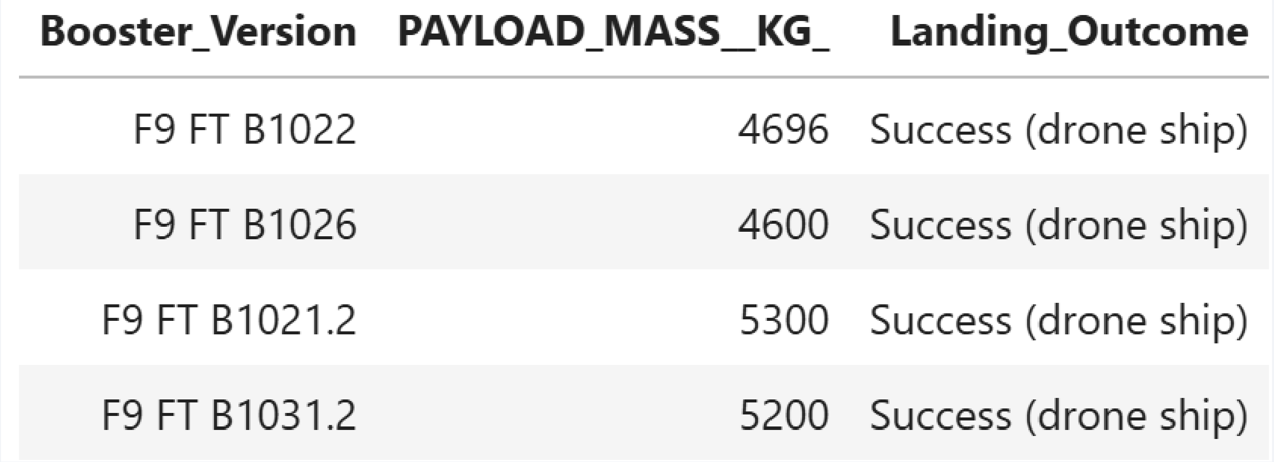
Successful Drone Ship Landing with Payload between 4000 and 6000

The names of boosters which have successfully landed on a drone ship and had payload mass greater than 4000kg but less than 6000kg was determined using the SQL query below which selects booster version, payload mass and landing outcome where the outcome is “Success (drone ship)” and payload is between 4000kg and 60000kg.



```
%%sql select Booster_Version, PAYLOAD_MASS__KG_, Landing_outcome from SPACEXTBL
    where Landing_outcome like 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000
```

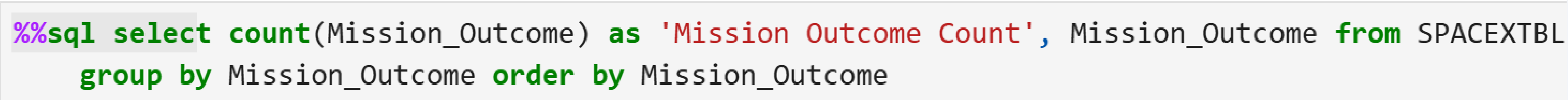
The result is



Booster_Version	PAYLOAD_MASS__KG_	Landing_Outcome
F9 FT B1022	4696	Success (drone ship)
F9 FT B1026	4600	Success (drone ship)
F9 FT B1021.2	5300	Success (drone ship)
F9 FT B1031.2	5200	Success (drone ship)

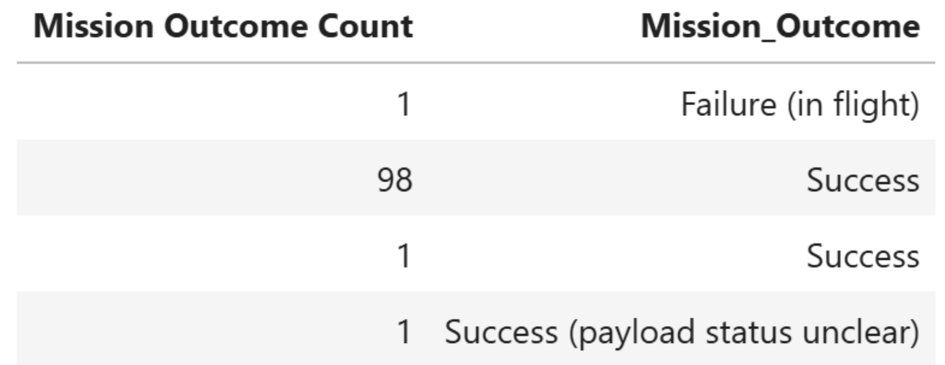
Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failed mission outcomes was determined with the SQL query below which counts the mission outcomes from the table, grouped by the outcome and presented in alphabetical order of Mission_Outcome.



```
%%sql select count(Mission_Outcome) as 'Mission Outcome Count', Mission_Outcome from SPACEXTBL
    group by Mission_Outcome order by Mission_Outcome
```

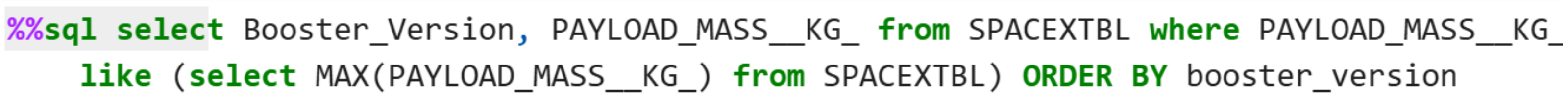
The result is



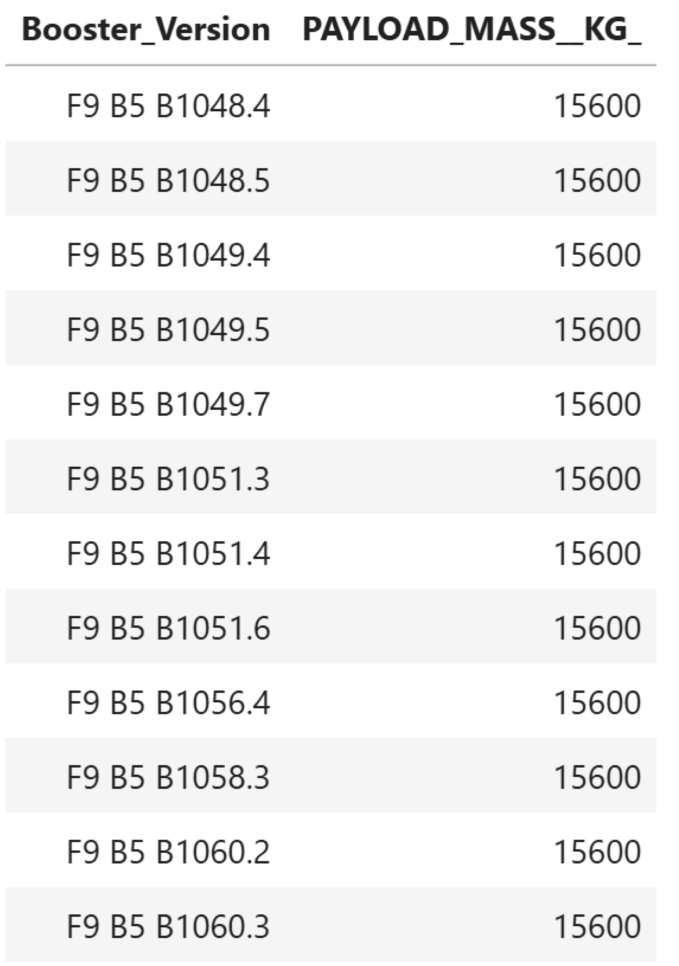
Mission Outcome Count	Mission_Outcome
1	Failure (in flight)
98	Success
1	Success
1	Success (payload status unclear)

Boosters Carried Maximum Payload

The names of the booster which have carried the maximum payload mass was determined using the SQL query below which selects the booster version and payload mass where payload mass is the maximum value in the table. The results are ordered by booster version.



```
%%sql select Booster_Version, PAYLOAD_MASS__KG_ from SPACEXTBL where PAYLOAD_MASS__KG_
    like (select MAX(PAYLOAD_MASS__KG_) from SPACEXTBL) ORDER BY booster_version
```



Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

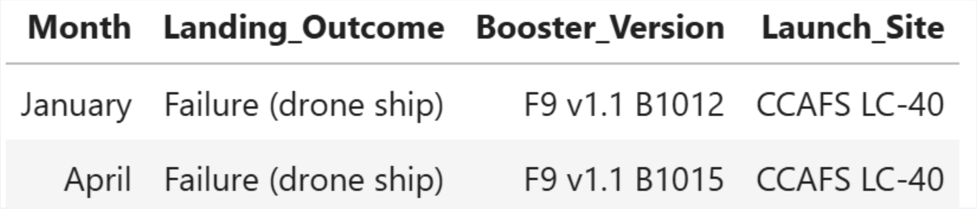
2015 Launch Records

The failed landing_outcomes in drone ship, their booster versions, and launch site names for year 2015 were retrieved using the SQL query opposite where the landing outcome was “Failure (drone ship”).

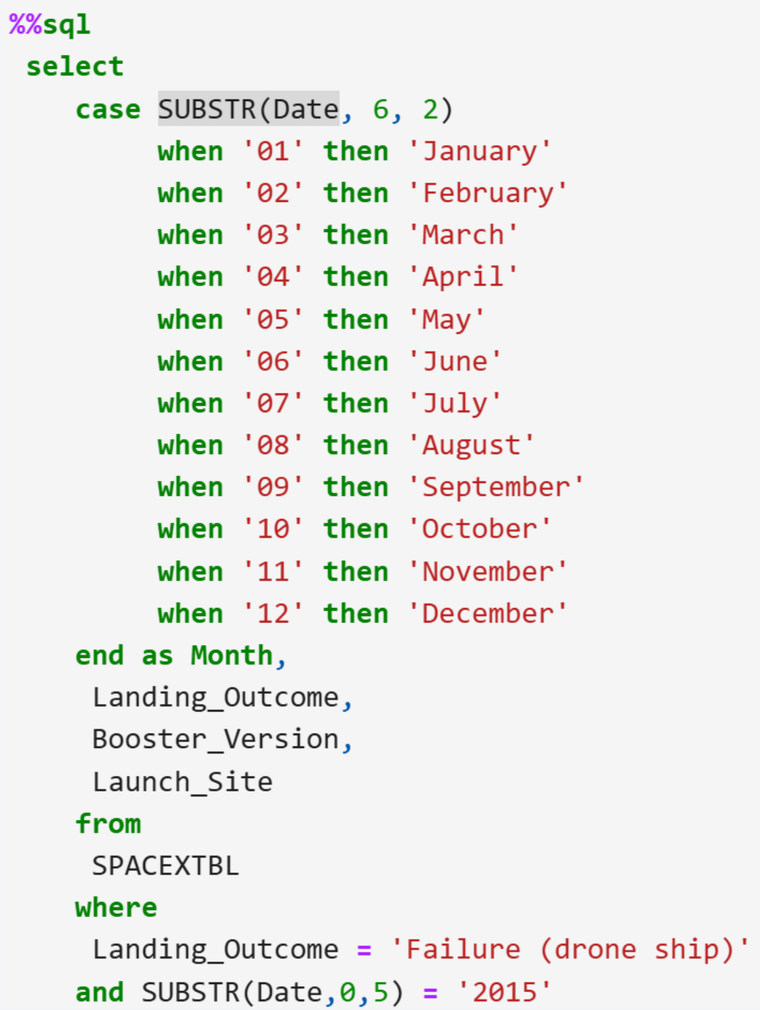
The code substr(Date, 6,2) as month was used to get the months and substr(Date,0,5)='2015’ was used to get the year as SQLLite does not support month names.

Note: SUBSTR extracts a substring from a string, starting at a specified position and with an optional length.

The result is



Month	Landing_Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

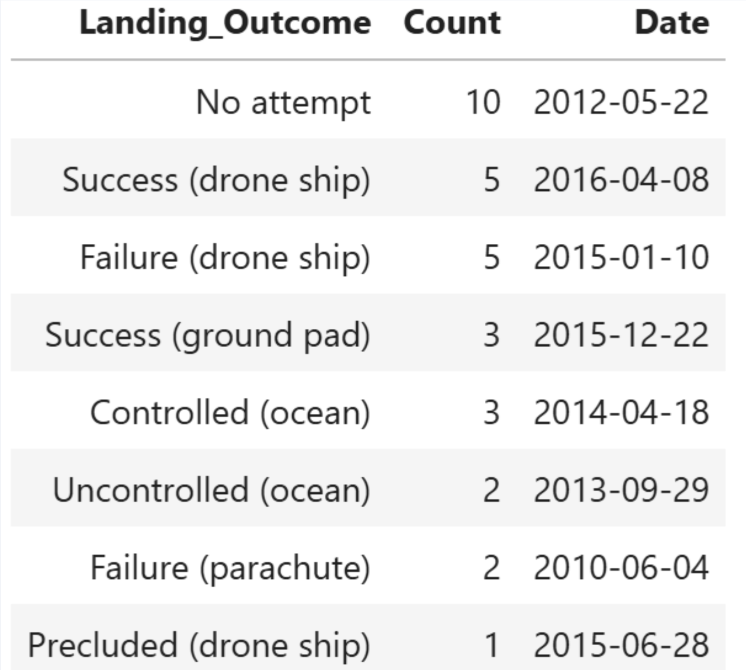


```
%%sql
 select
    case SUBSTR(Date, 6, 2)
         when '01' then 'January'
         when '02' then 'February'
         when '03' then 'March'
         when '04' then 'April'
         when '05' then 'May'
         when '06' then 'June'
         when '07' then 'July'
         when '08' then 'August'
         when '09' then 'September'
         when '10' then 'October'
         when '11' then 'November'
         when '12' then 'December'
    end as Month,
     Landing_Outcome,
     Booster_Version,
     Launch_Site
    from
     SPACEXTBL
    where
     Landing_Outcome = 'Failure (drone ship)'
    and SUBSTR(Date,0,5) = '2015'
```

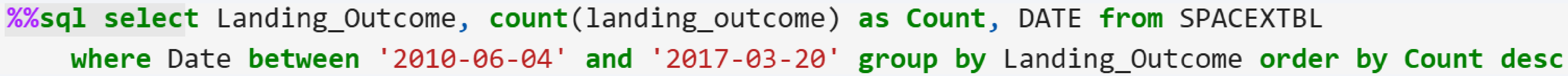
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

The count of the different landing outcomes between the date 2010-06-04 and 2017-03-20, ranked in descending order was determined by the SQL query below.

The query selected landing outcome, count of the landing outcomes and date from the table for the specified dates, grouped by the outcome and ordered by the count in descending order.



Landing_Outcome	Count	Date
No attempt	10	2012-05-22
Success (drone ship)	5	2016-04-08
Failure (drone ship)	5	2015-01-10
Success (ground pad)	3	2015-12-22
Controlled (ocean)	3	2014-04-18
Uncontrolled (ocean)	2	2013-09-29
Failure (parachute)	2	2010-06-04
Precluded (drone ship)	1	2015-06-28



```
%%sql select Landing_Outcome, count(landing_outcome) as Count, DATE from SPACEXTBL
    where Date between '2010-06-04' and '2017-03-20' group by Landing_Outcome order by Count desc
```

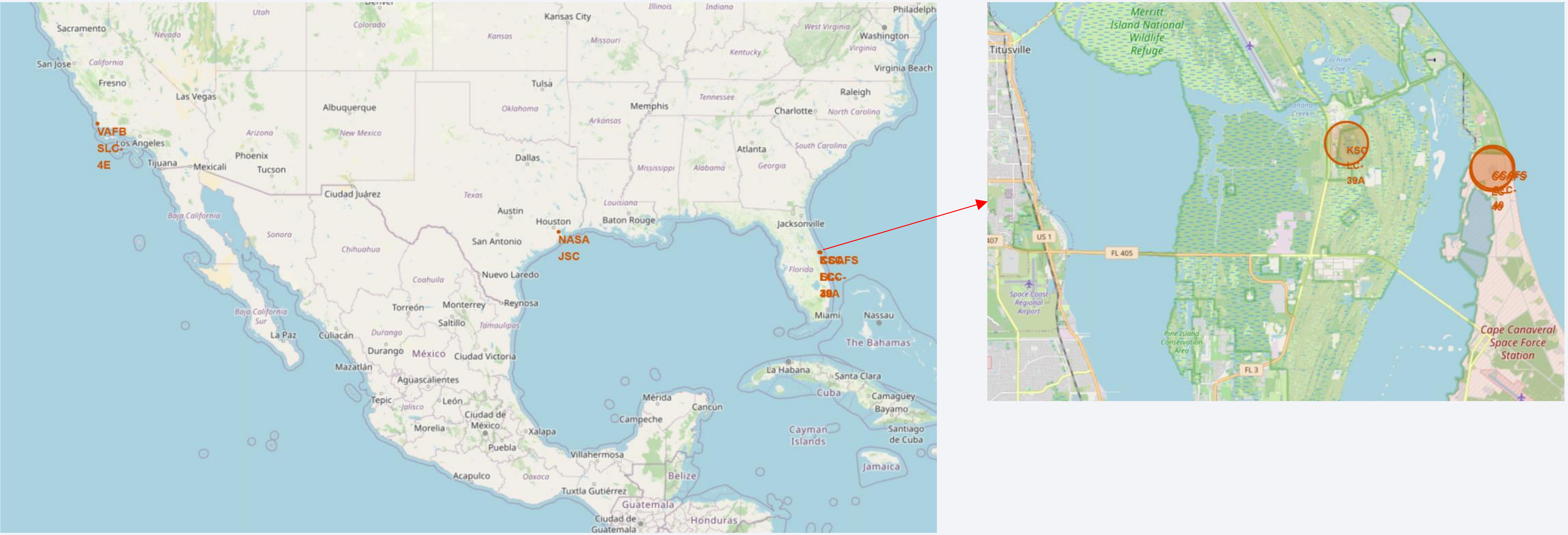
Yvonne Booth 27/12/24

Section 3

Launch Sites Proximities Analysis

Locations of all Launch Sites

Yvonne Booth 27/12/24



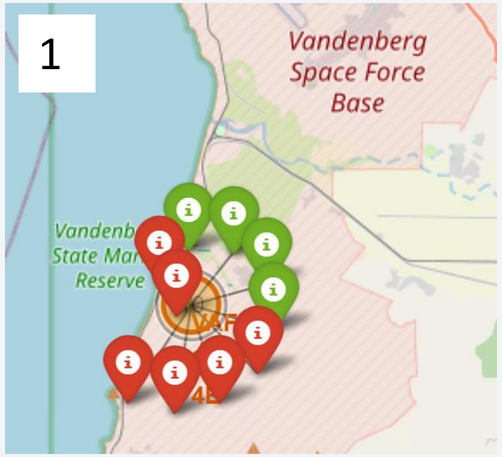
Locations of SpaceX launch sites in relation to the NASA Johnson Space Center at Houston, Texas :

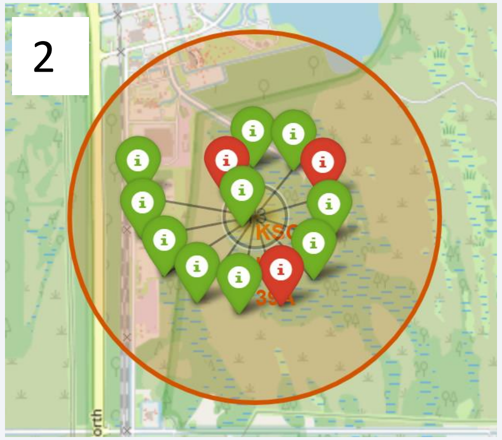
VAFB SLC 4E (California)
KSC LC-39A, CCAFS SLC-40 and CCAFS LC-40 (Florida)

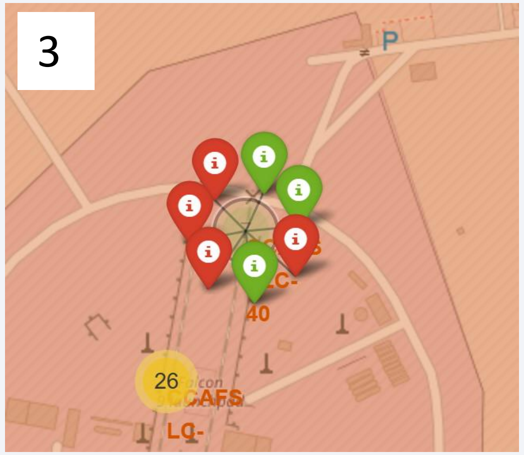
Launch Successes and Failures by Site

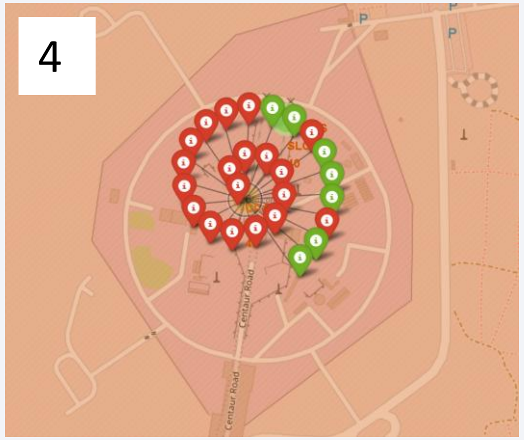
The figures opposite show visualisations representing launch failures (red) and successes (green) for each launch site.

1. VAFB
2. KSC LC-39A
3. CCAFS SLC-40
4. CCAFS LC-40


1


2


3

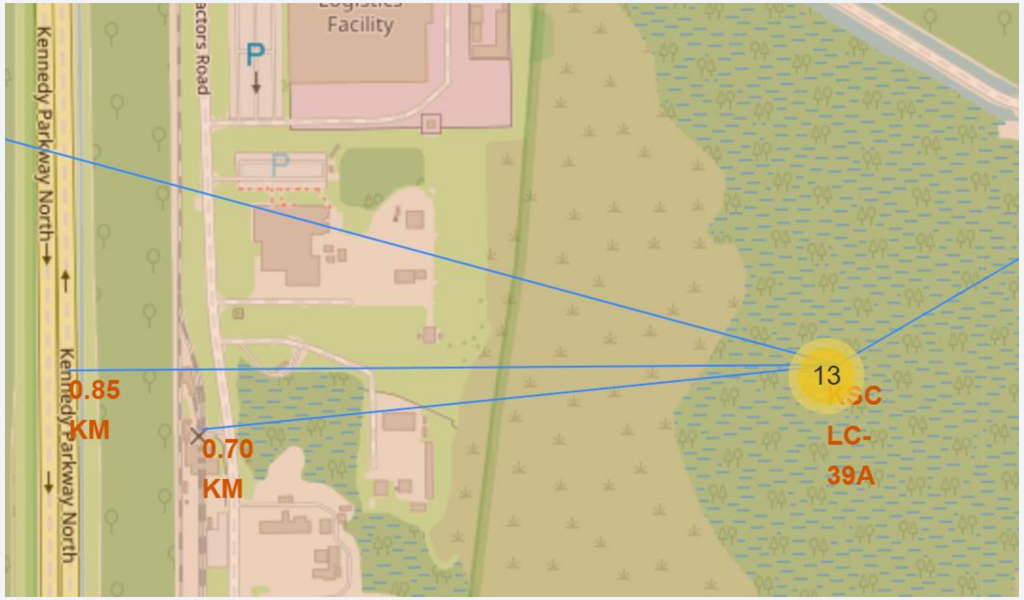

4

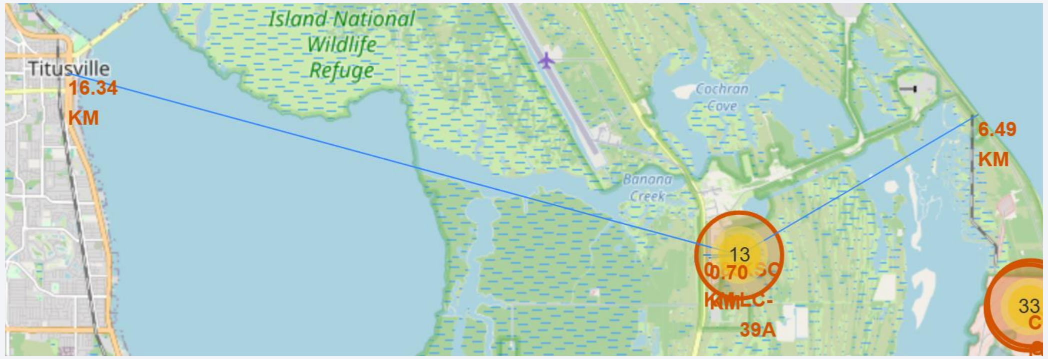
Launch Site KSC LC 39A Proximities

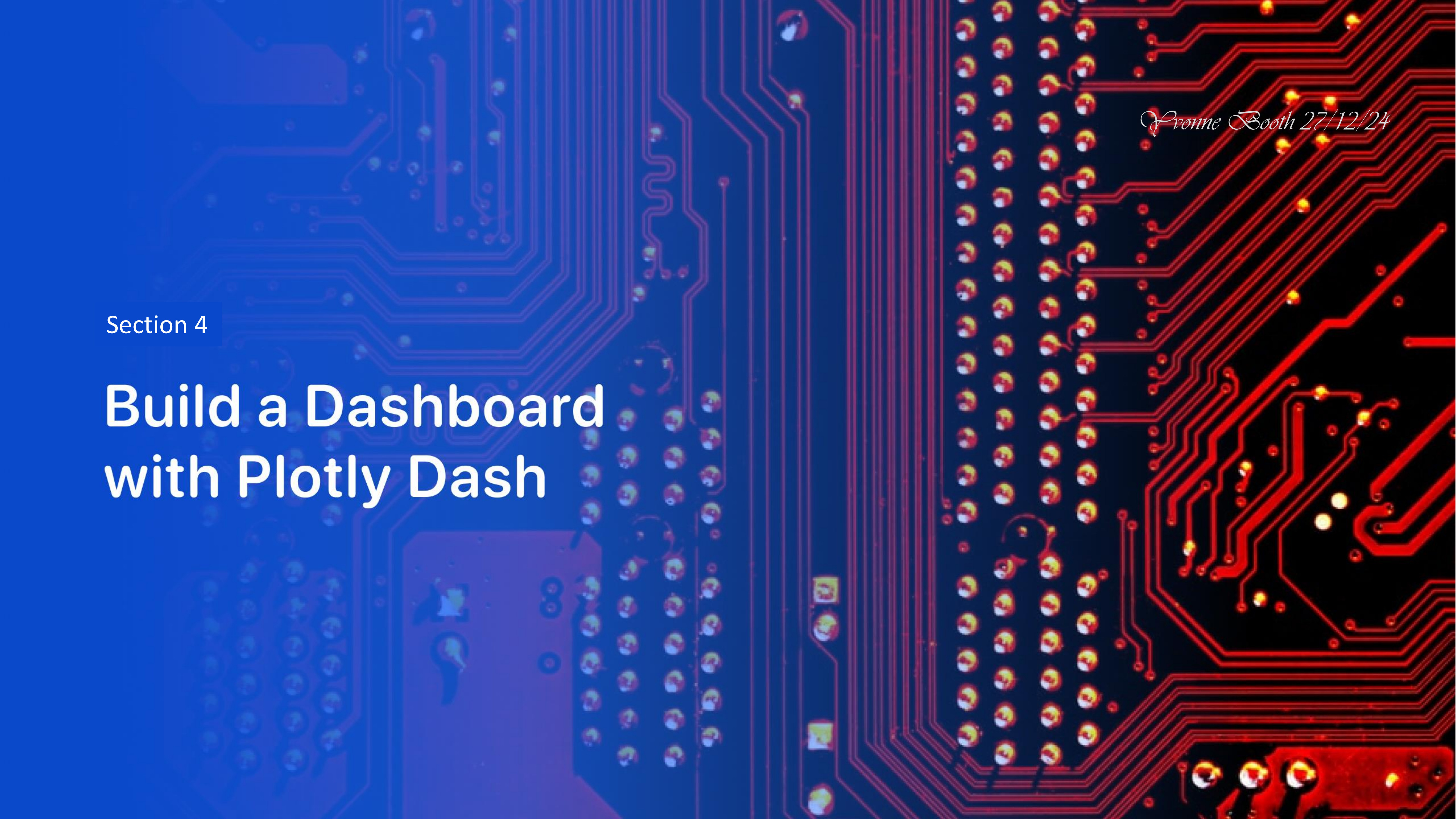
Launch site KSC LC 39A in Florida was selected to explore proximities to railway, highway, coastline, and city with distance calculated and displayed as shown in the figures opposite.

The top figure shows that the launch site is in very close proximity to a highway (Kennedy Parkway North) and railway with distances of 0.85km and 0.7km respectively. The bottom figure shows that the site is 6.49km from the coastline and the closest city is Titusville which is 16.34km away.

It makes sense that the site would be well connected by road and rail for personnel transfer and transporting goods, but far enough from the city to limit impact of noise and allow for a secure space around the site. Having the site close to the coast and away from built up areas also limits impact from debris should a launch fail.







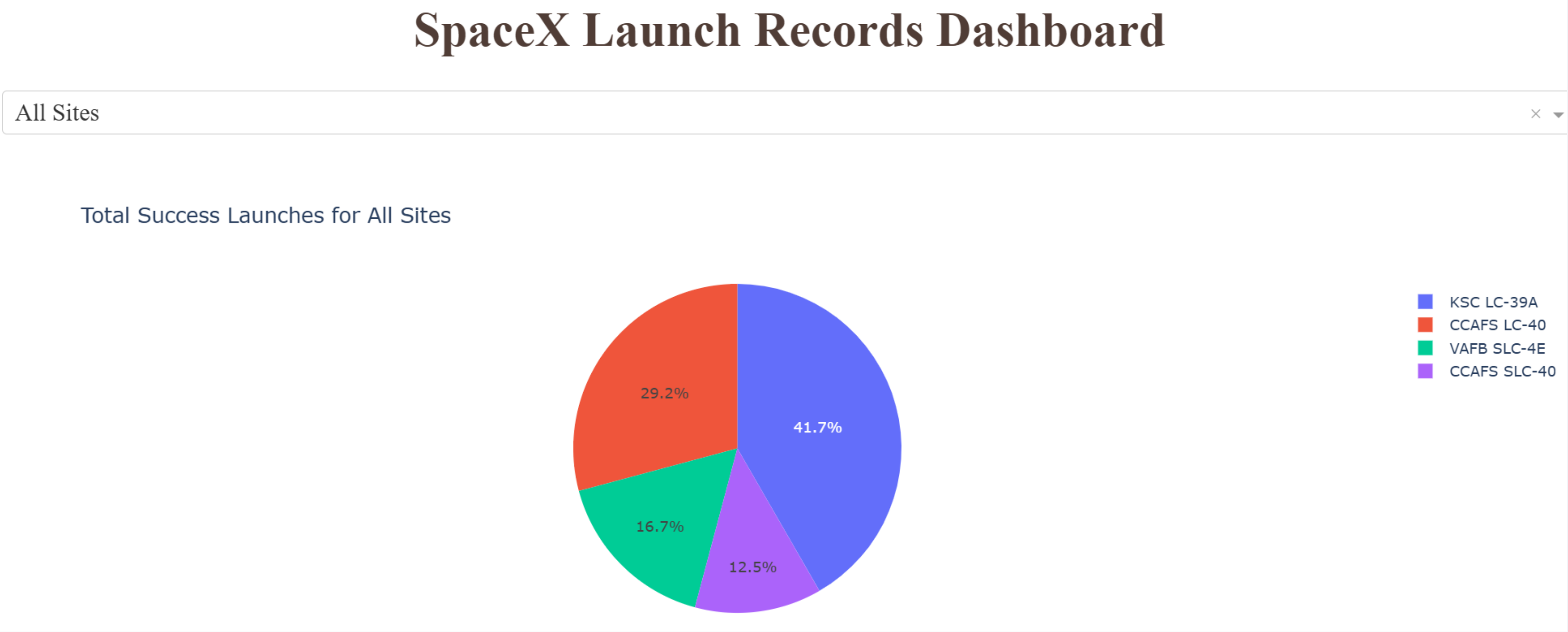
Yvonne Booth 27/12/24

Section 4

Build a Dashboard with Plotly Dash

Launch Success for All Sites

Yvonne Booth 27/12/24



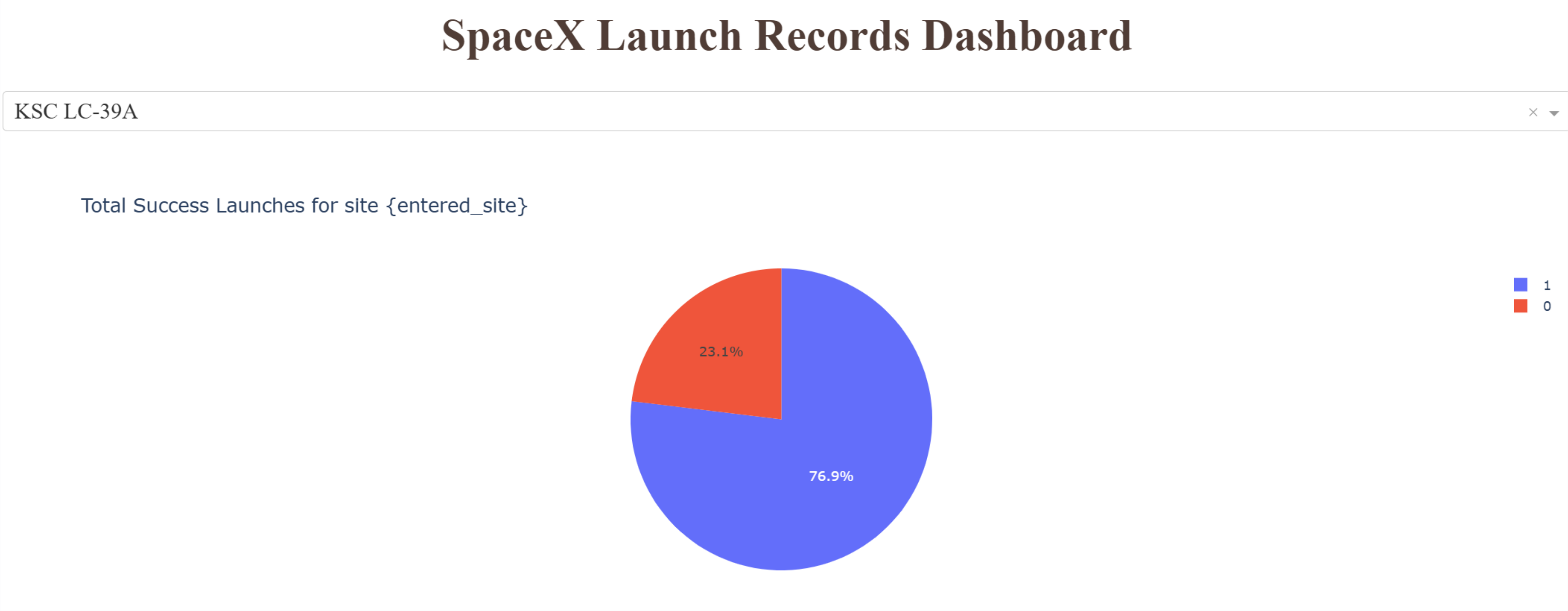
SpaceX Launch Records Dashboard

All Sites

KSC LC-39A had the highest proportion of successful launches (41.7%).

CCAFS SLC-40 had the lowest proportion of successful launches (12.5%).

Launch Outcomes for KSC LC-39A



SpaceX Launch Records Dashboard

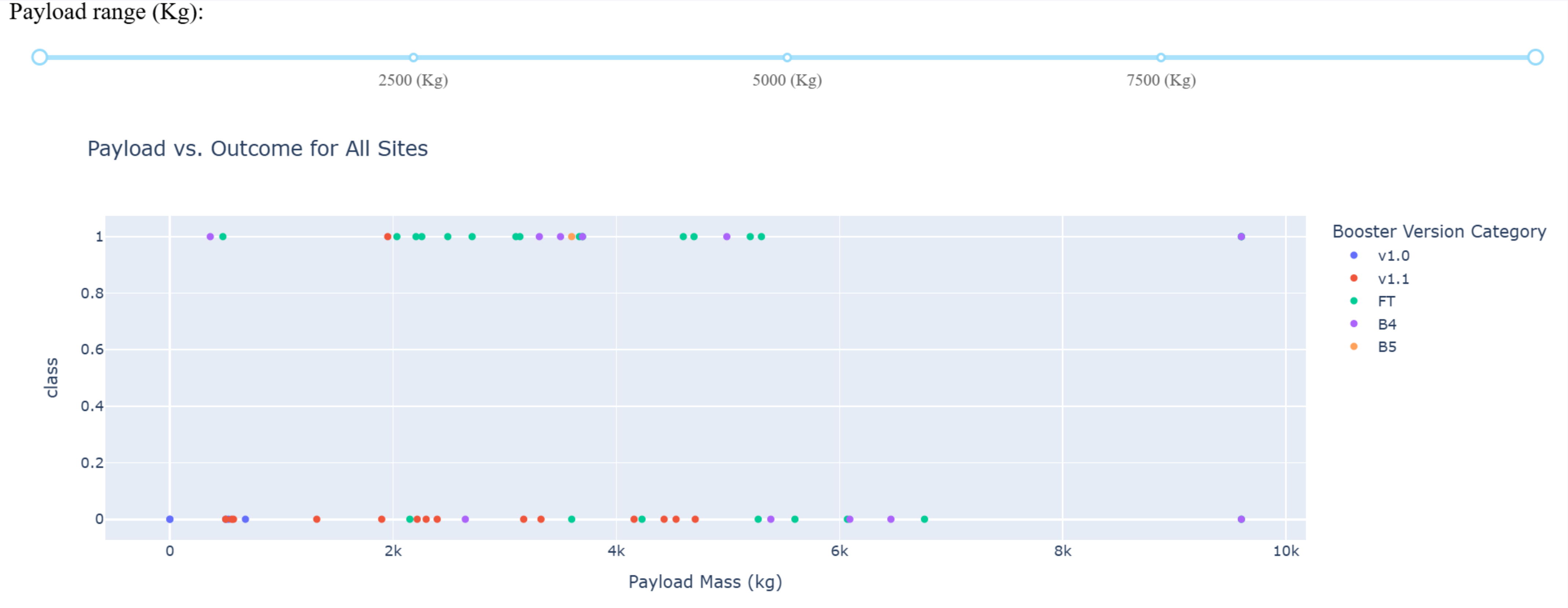
KSC LC-39A

Total Success Launches for site {entered_site}

Launch site KSC LC-39A had the highest launch success rate (Class 1) with a 76.9% success rate and 23.1% failure rate (Class 0).

Launch Outcomes based on Payload Mass for each Booster Version

Yvonne Booth 27/12/24



The plot shows that overall, FT booster version has the most successful launches.

Only booster version B4 has been launched with payloads over 8000kg with one successful and one failed launch.

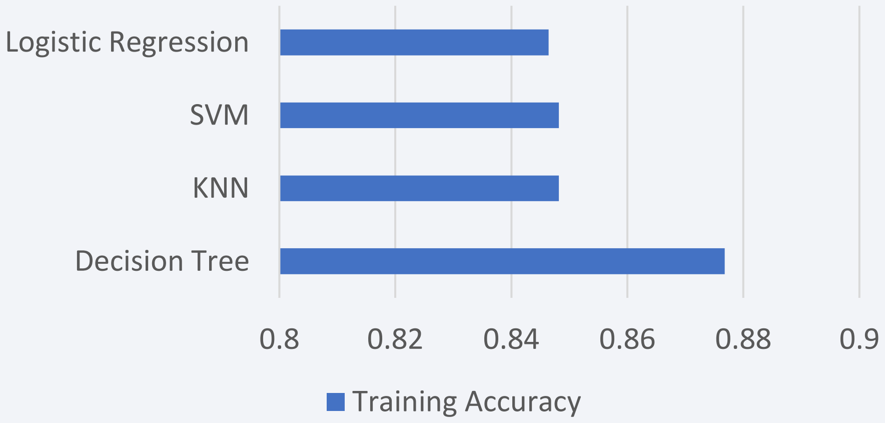
Yvonne Booth 27/12/24

Section 5

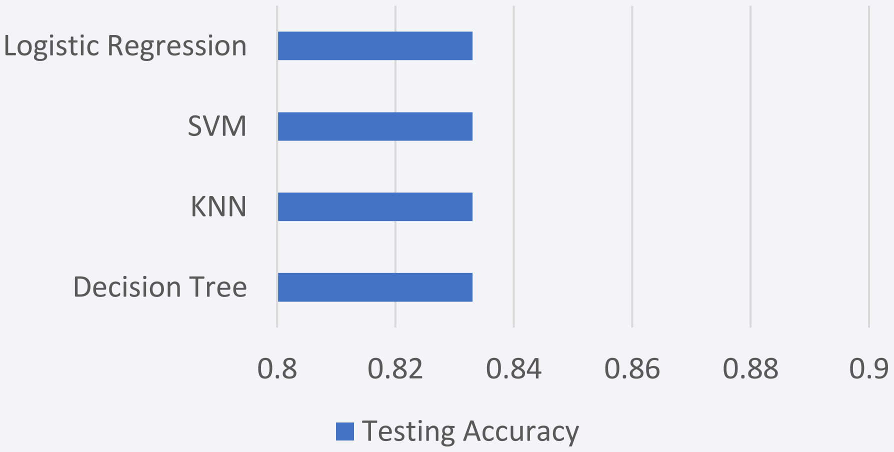
Predictive Analysis (Classification)

Classification Accuracy

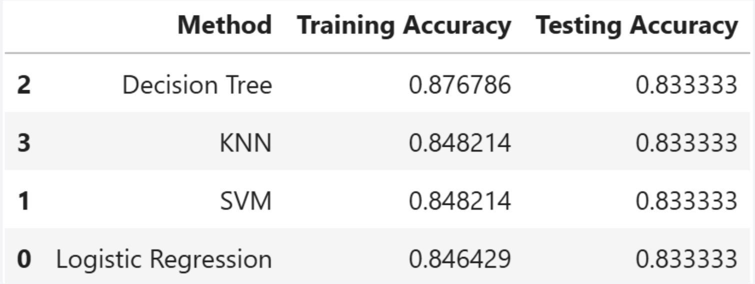
Training accuracy for all built classification models



Testing accuracy for all built classification models



Testing accuracy is 0.8333 for all models.
The training accuracy plot above shows that the Decision Tree is the most accurate with a value of 0.877. All values are shown in the table opposite.



	Method	Training Accuracy	Testing Accuracy
2	Decision Tree	0.876786	0.833333
3	KNN	0.848214	0.833333
1	SVM	0.848214	0.833333
0	Logistic Regression	0.846429	0.833333

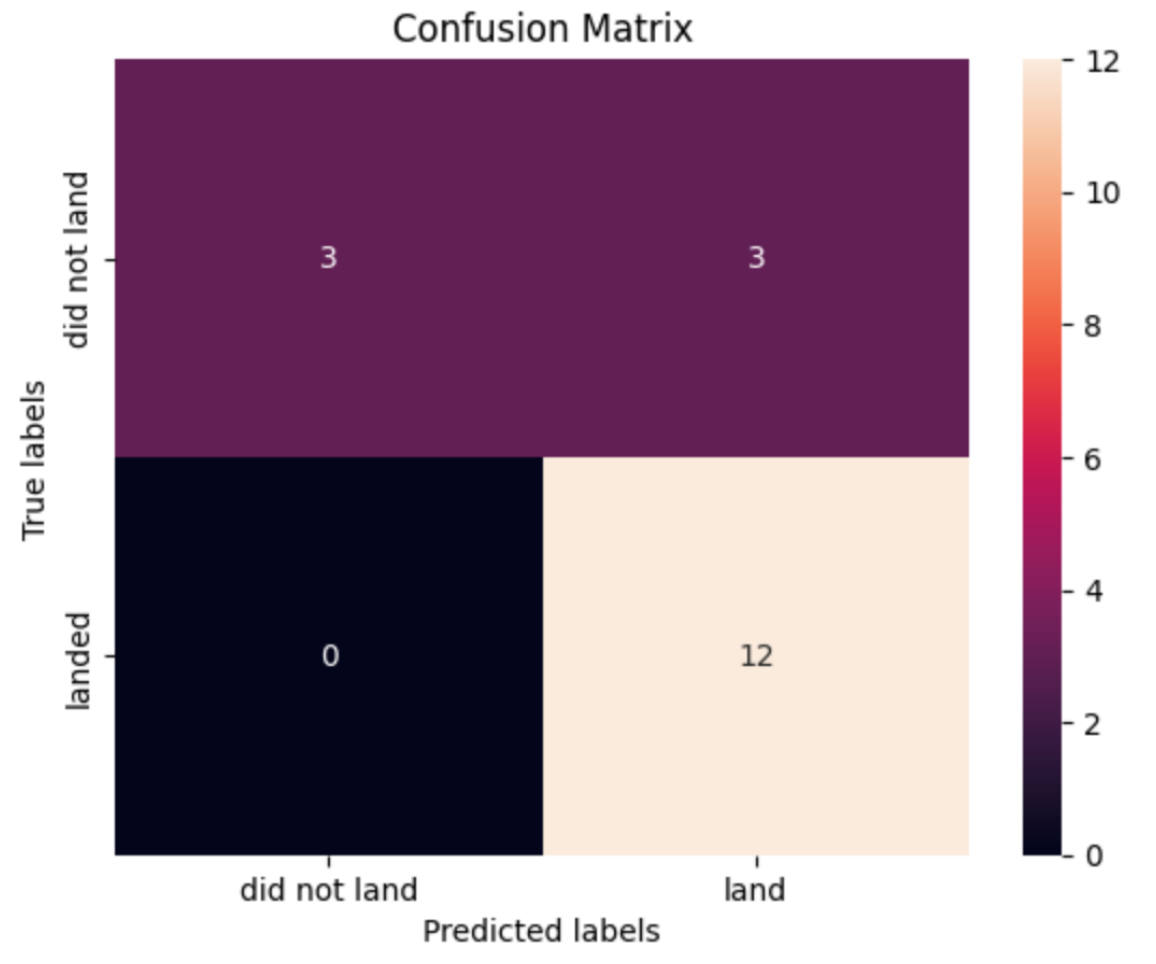
Confusion Matrix

The Confusion Matrix for the best performing model (Decision Tree) is shown opposite.

Twelve out of fifteen of the successful landings were predicted correctly as shown by the bottom right square (true positive).

Three failed landings were predicted correctly as shown by the top left square (true negative).

The model incorrectly predicted three landings as successes, but in reality these were failed landings as shown in the top right corner (false positive).



Conclusions

Flight Numbers

- As flight numbers increase with time, experience is gained and lessons are learned, which is reflected in an increased number of successful launches.

Payload

- At site CCAFS SLC 40 the number of successful launches increases with increased payload, but no clear relationship is seen at other launch sites between payload and success.
- The success for massive payloads (over 4000kg) is generally lower than that for low payloads.
- Only booster version B4 has been launched with payloads over 8000kg with one successful and one failed launch. There is not enough data to determine any relationship for these much heavier payloads.

Orbit

- Orbit types ES-L1, GEO, HEO, and SSO, have the highest (100%) success rate, although GEO, HEO, and ES-L1 only have 1 launch each. Further launches and more data is needed to determine if there is a general relationship between orbit type and success.
- Success of launches to the LEO orbit seem to be related to the number of flights with more success as flights increase.
- Launches to the Polar, LEO and ISS orbits appear to have more success with higher payloads, but other factors not taken into account in the plots such as flight number and launch site may be influencing this apparent relationship.
- VLEO (Very Low Earth Orbit) launches that have less distance to travel are associated with heavier payloads.

Launch Site

- CCAFS SLC-40 launch site in Florida had the lowest proportion of successful launches (12.5%).
- KSC LC-39A in Florida had the highest proportion of successful launches of all sites (41.7%), with a success rate of 76.9%. This cannot be explained by the data available in this project, but further research shows this site underwent refurbishment which started in 2013. This may be a factor.

Booster Version

- Overall, the FT booster version had the most successful launches. This was an upgrade from the v1.1 booster so it would be expected that it would be more successful than the earlier versions.

Classification Model

- The Decision Tree model correctly predicted 15 out of 18 landing outcomes.
- The Decision Tree classification model with a training accuracy of 0.877 is more accurate than the SVM and Logistic Regression models run.

Appendix

Jupyter Notebooks

Assosciated Jupyter Notebooks with code and results can be found by following the link below.

https://github.com/yvnb1/IBM-Data-Science-Capstone-Project

Thank you!