IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Michael-Ange Kakudji
21 October 2024

Outline

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

Executive Summary

- Summary of methodologies
  - Methodologies involved included data collection and cleaning, EDA, data visualization and model selection.
- Summary of all results
  - The best models selected for prediction included the Logistic Regression, SVM and K-Nearest Neighbours models with the highest F1-Scores at 89%.

Introduction

- Space X advertises Falcon 9 rocket launches for a fraction of the cost compared to its competitors, much of the savings is due to Space X reusing the first stage.
- The objective is to predict whether the first stage will land, and therefore the cost of a launch can be determined.
- This information can be used if an alternate company wants to bid against Space X for a rocket launch.
- ML models and data visualizations are to be utilized to determine the optimal prediction system to be used.

Section 1

Methodology

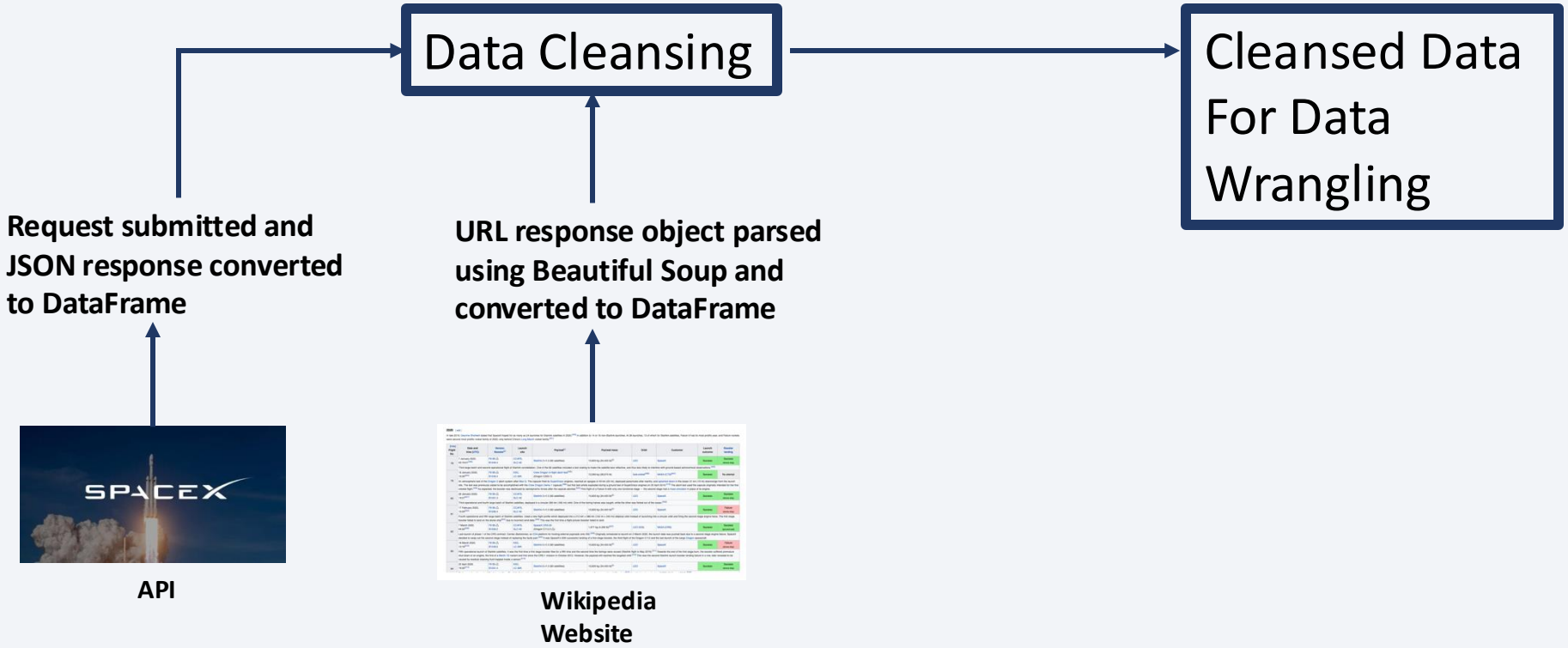
Methodology

Executive Summary

- Data collection methodology:
  - The data was collected through SpaceX API, web scraping from Wikipedia page and parsed using Beautiful Soup.
- Perform data wrangling
  - Exploratory Data Analysis methodologies and metrics were utilized to obtain a high-level better understanding of the data and create training labels such as landing outcome.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Python ML libraries and models were utilized in conjunction with the GridSearch function to determine the optimal tuning parameters, and models accuracy score recorded.

Data Collection

- The data was collected by submitting a request to the SpaceX API which included the rocket data, followed by web scraping from the publicly available Wikipedia information which included landing outcomes, and the data was parsed accordingly using Beautiful Soup.



Data Collection – SpaceX API

DataFrame created from SpaceX API response.

[36]:

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.561857
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.561857
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.632093
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577366	28.561857

GitHub URL: https://github.com/Michael-Ange917/Applied-Data-Science-Capstone-Project/blob/main/1.%20Hands-On%20Lab_Data%20Collection.ipynb

API Response object created in Python using Get request.

Response decoded as JSON and converted into Pandas DataFrame using .json_normalize()

Data slicing and cleansing to obtain the final usable DataFrame.

API

Data Collection - Scraping

DataFrame created from web scraping Wikipedia List of Falcon 9 and Falcon Heavy launches webpage.

[43]:

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.07B0003.18	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.07B0004.18	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.07B0005.18	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.07B0006.18	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.07B0007.18	No attempt\n	1 March 2013	15:10

GitHub URL: https://github.com/Michael-Ange917/Applied-Data-Science-Capstone-Project/blob/main/2.%20Hands-On%20Lab_Web%20Scraping.ipynb

Website object created in Python using Get request from URL.

Data response parsed using Beautiful Soup and content added to dictionaries using functions based on for loops. Dictionary converted to DataFrame.

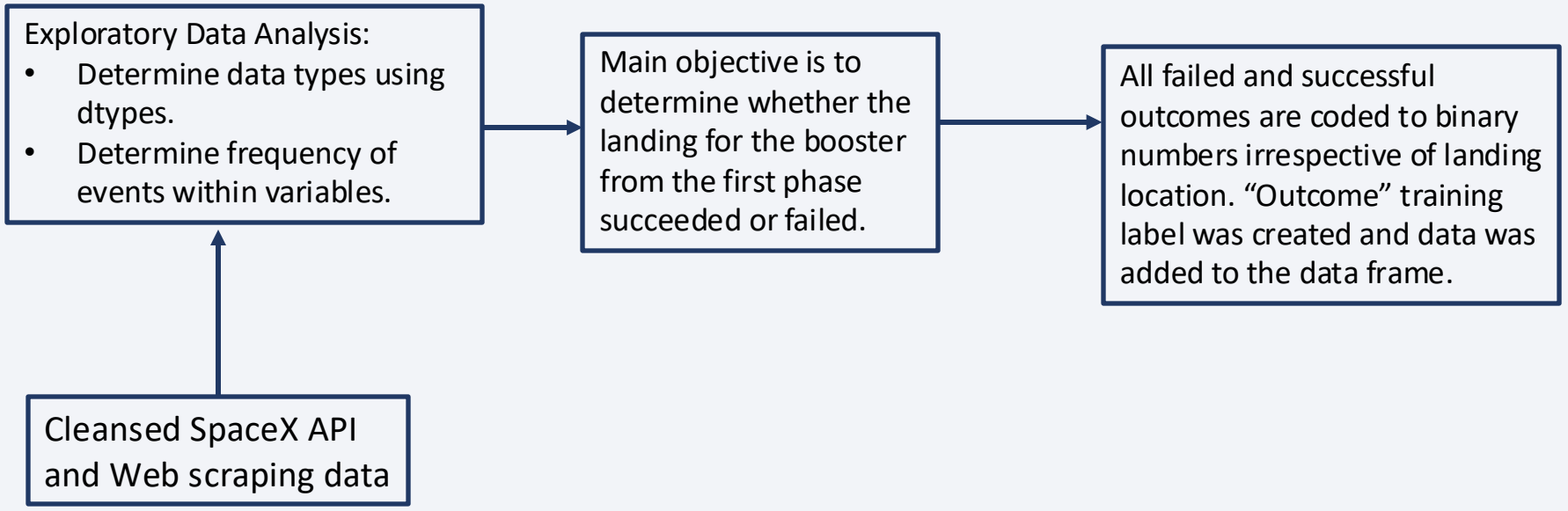
Data slicing and cleansing to obtain the final usable DataFrame.

Wikipedia Website

Data Wrangling

Exploratory Data Analysis methodologies and metrics were utilized to obtain a high-level better understanding of the data and create training labels such as landing outcome.

GitHub URL: https://github.com/Michael-Ange917/Applied-Data-Science-Capstone-Project/blob/main/3.%20Hands-On%20Lab_Data%20Wrangling.ipynb



EDA with Data Visualization

- The following plots were created for EDA purposes with the objective of visualizing the correlation between various variables:
  - Categorical Scatter Plot
  - Scatter Plot
  - Bar Chart
  - Line Chart

GitHub URL: https://github.com/Michael-Ange917/Applied-Data-Science-Capstone-Project/blob/main/4.%20Hands-On-Lab_EDA%20With%20Visualization%20Lab.ipynb

EDA with SQL

- SQL queries were executed to obtain visibility and extract the following data:
  - The names of the unique launch sites in the space mission.
  - The 5 records where launch sites begin with the string 'CCA'.
  - The total payload mass carried by boosters launched by NASA (CRS).
  - The average payload mass carried by booster version F9 v1.1.
  - Determine the date of when the first successful landing outcome on a ground pad was achieved.
  - The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
  - The total number of successful and failure mission outcomes.
  - The names of the booster versions which have carried the maximum payload mass.
  - The records which will display the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015.
  - The count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.

GitHub URL: https://github.com/Michael-Ange917/Applied-Data-Science-Capstone-Project/blob/main/5.%20Hands-On%20Lab_Complete%20EDA%20With%20SQL.ipynb

Build an Interactive Map with Folium

Markers, circles and lines were embedded in the Folium maps to be able to indicate and group the different launch locations. The icons would also indicate the unique mission landing outcomes and the distances of specific launch locations from nearby amenities and coastline.

GitHub URL: https://github.com/Michael-Ange917/Applied-Data-Science-Capstone-Project/blob/main/6.%20Hands-On-Lab_Interactive%20Visual%20Analytics%20With%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

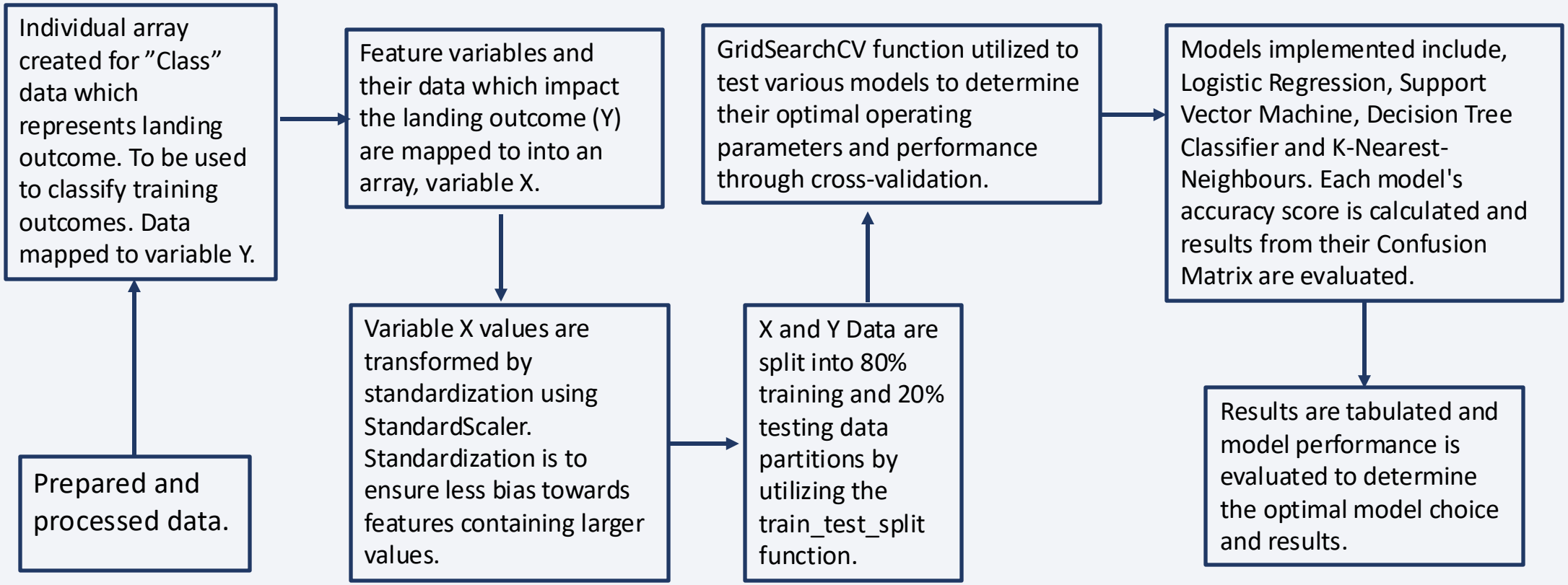
- Interactive dashboards were created using Plotly Dash which included a pie chart and scatter plot.
- The objective of the pie chart was to determine the launch outcome success rate for the collective and individual regions.
- The objective of the scatter plot was determine the correlation between the Payload Mass (Kg) and the successful or failed Landing Outcome per Booster Version implemented.

GitHub URL: https://github.com/Michael-Ange917/Applied-Data-Science-Capstone-Project/blob/main/7.%20spacex_dash_app.py

Predictive Analysis (Classification)

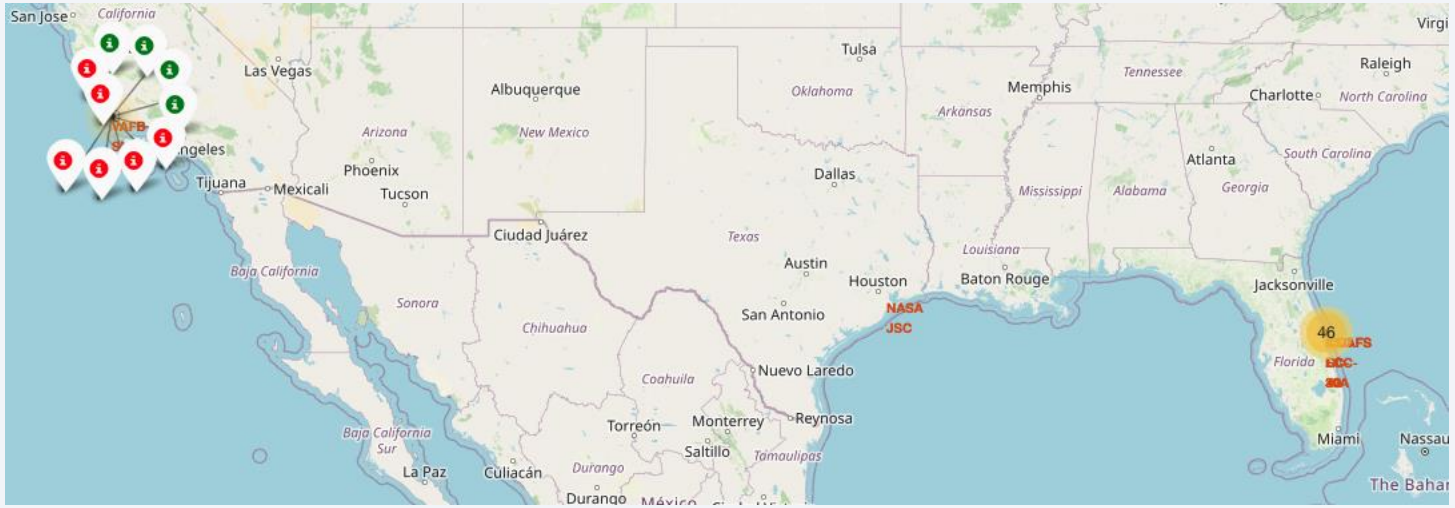
Python ML libraries and models were utilized in conjunction with the GridSearchCV function to determine the model's optimal tuning parameters, accuracy score and Confusion Matrix.

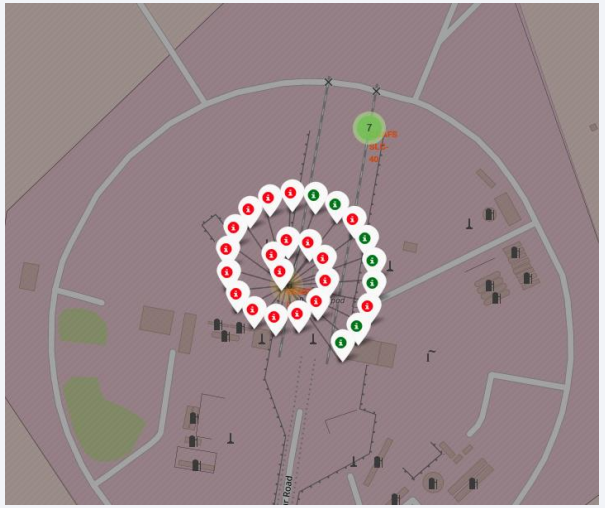
GitHub URL: https://github.com/Michael-Ange917/Applied-Data-Science-Capstone-Project/blob/main/8.%20SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



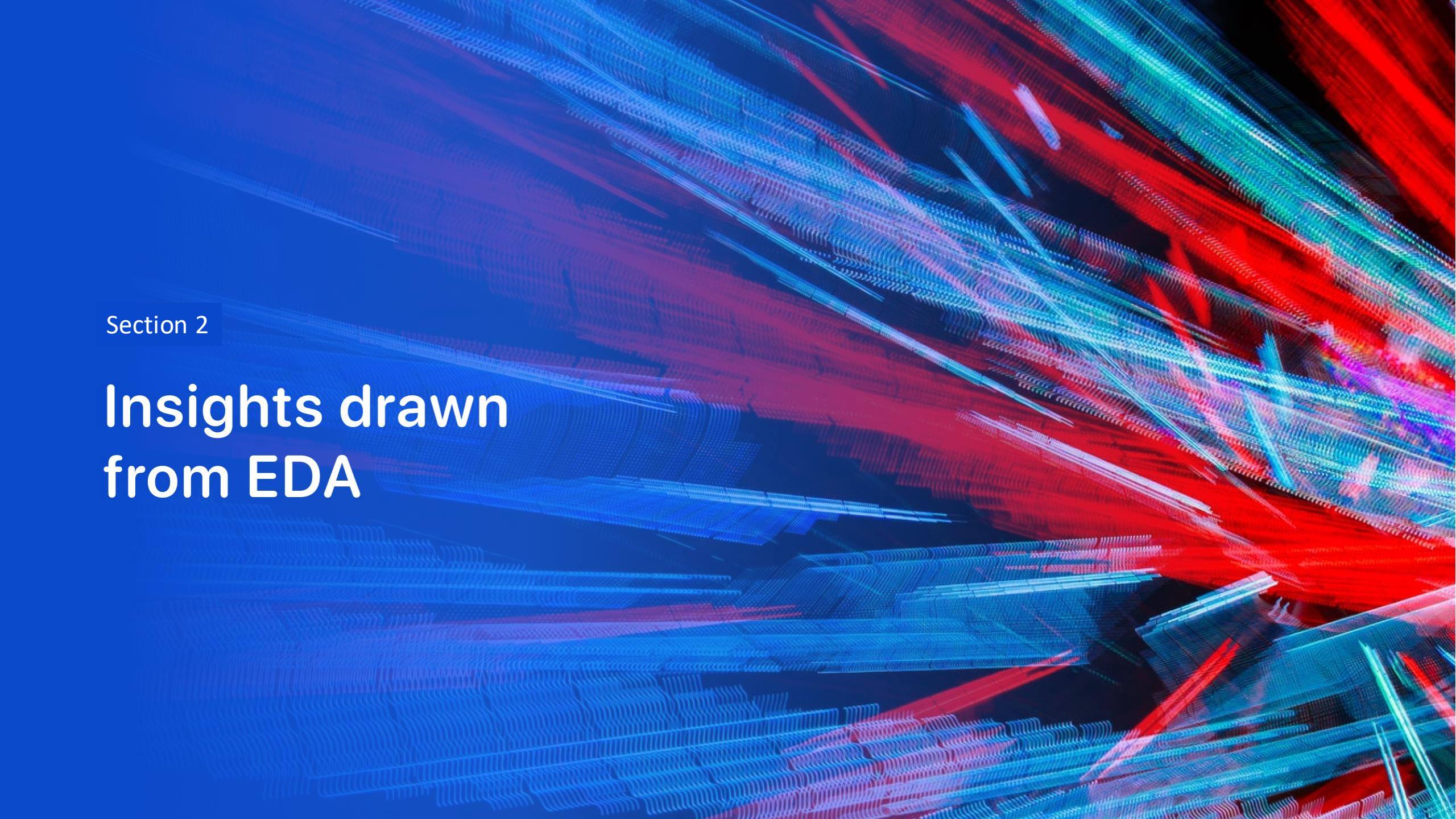
Results

- It was observed that there were more successful landing outcomes for launch locations CCAFS and KSC with the increase in the number of flights.

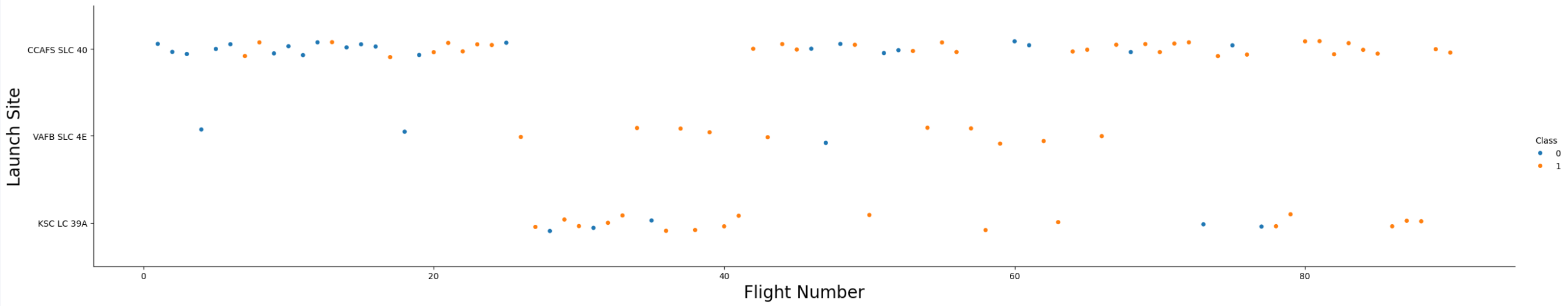




- The best models were chosen based on their confusion matrix results.
- The Logistic Regression, SVM and K-Nearest Neighbours had the highest F1 Score at 89%.
- Either of the three models can be used for predictions, future work can be done to narrow down the best-performing model and optimize results.


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

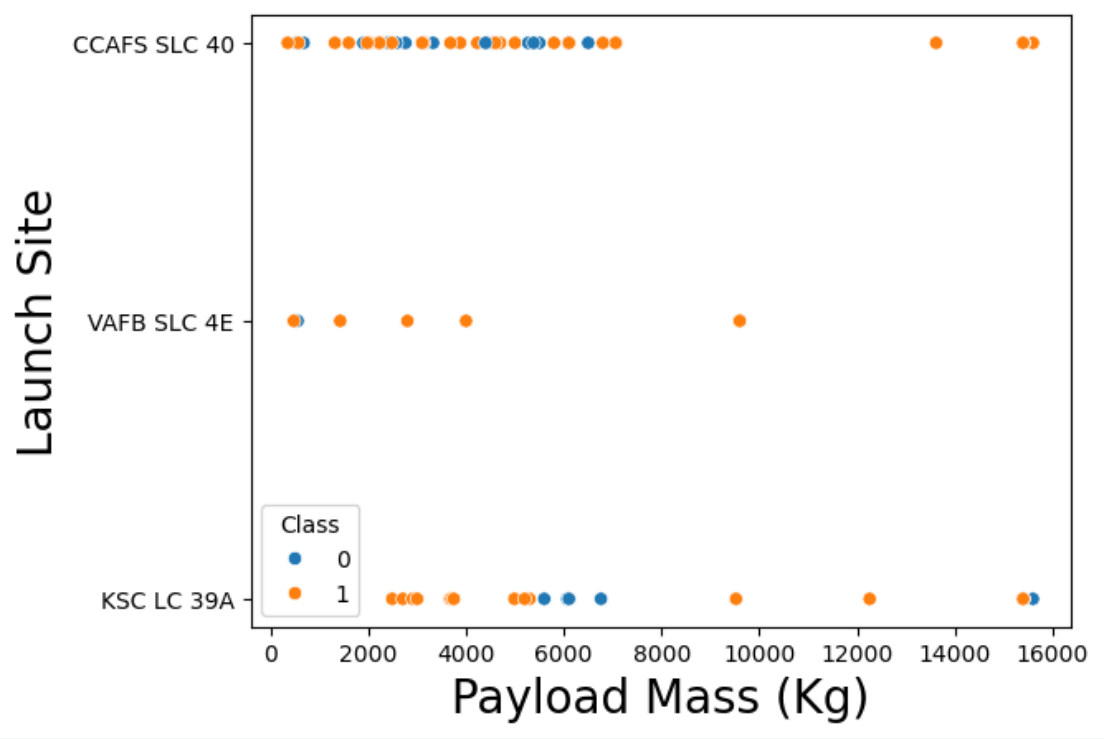


From the above plot, it is observed that there is more uniformity regarding successful landing outcomes for launch locations CCAFS and KSC with the increase of the number of flights.

Based on the data there were no launches for launch site VAFB beyond approximately flight number 70, however, there also seems to be a pattern of increased successful landings with the increase of the number of flights.

CCAFS has the highest amount of flights recorded.

Payload vs. Launch Site

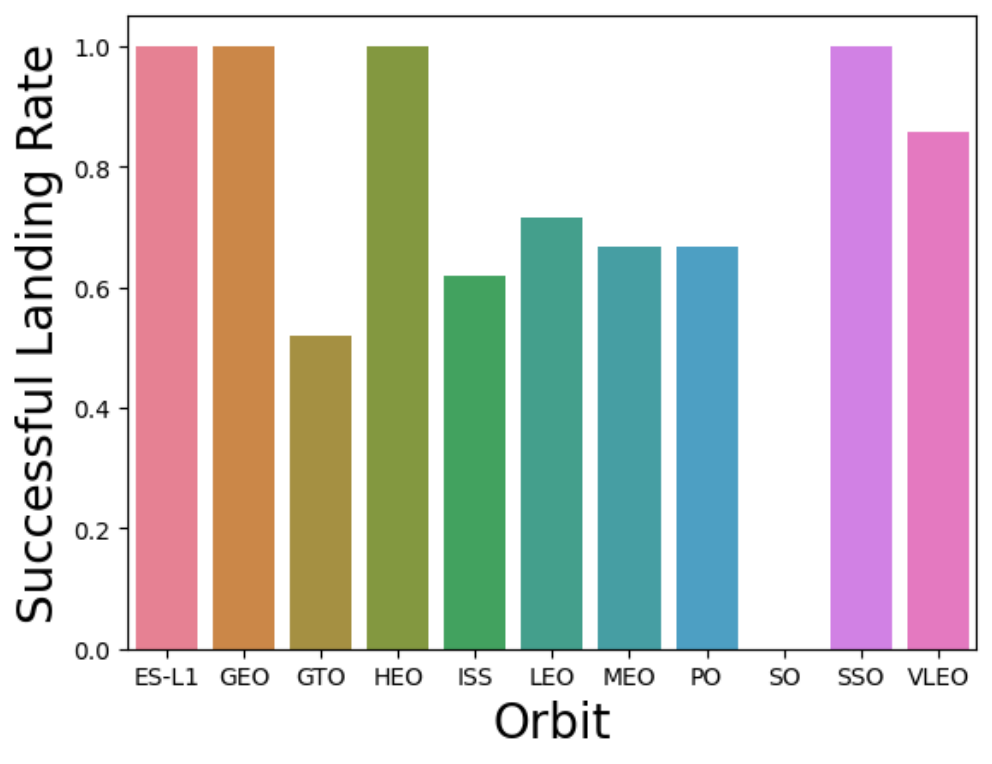


From the above plot, it is observed that there are no flights recorded for the VAFB launch site beyond a (heavy) payload mass of 10,000 kg.

CCAFS launch site observations include a higher successful landing outcome with the increased payload mass beyond 6000kg.

KSC launch site results are inconclusive, with no apparent clear trend.

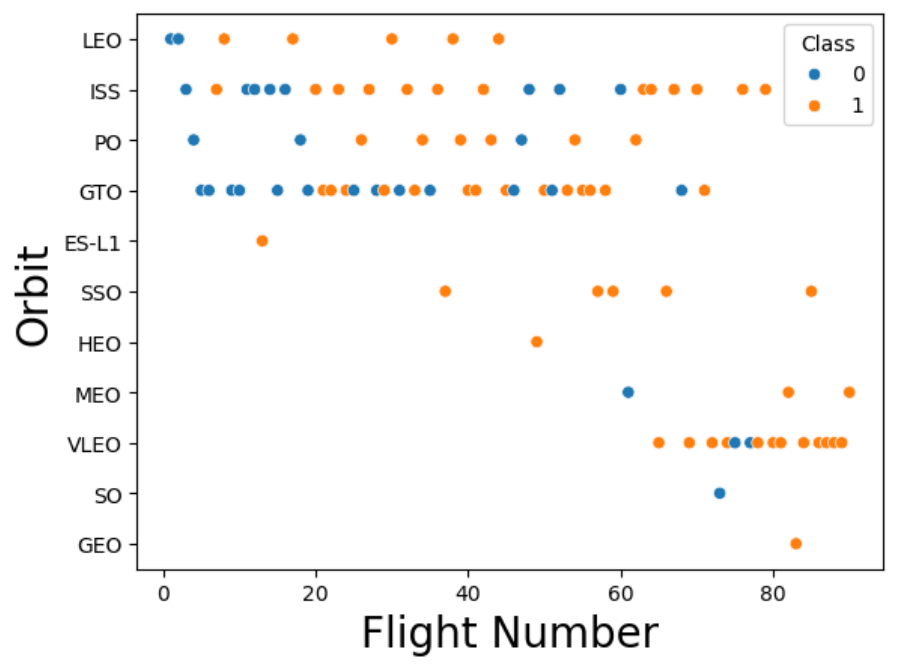
Success Rate vs. Orbit Type



From the image above, the orbits with a 100% successful landing rate include, ES-L1, GEO, HEO and SSO.

The orbit with the lowest successful landing rate is GTO.

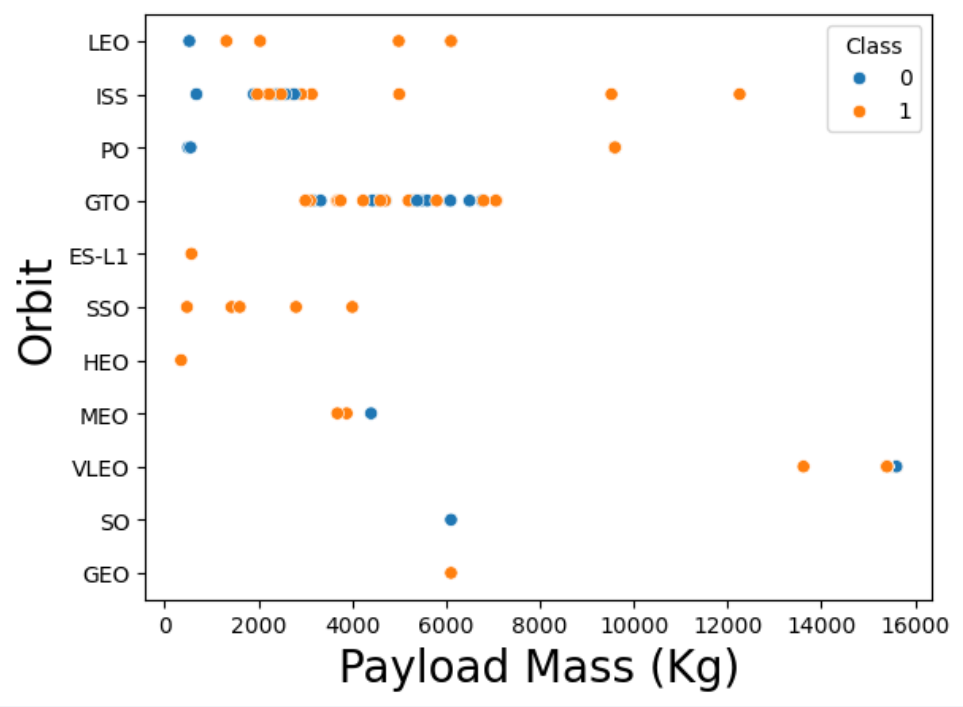
Flight Number vs. Orbit Type



From the above plot, the orbits it is observed that LEO, ISS, SSO and VLEO show an increase in successful landings with an increase in the number of flights.

The remaining orbits' results appear to be inconclusive, particularly GTO.

Payload vs. Orbit Type

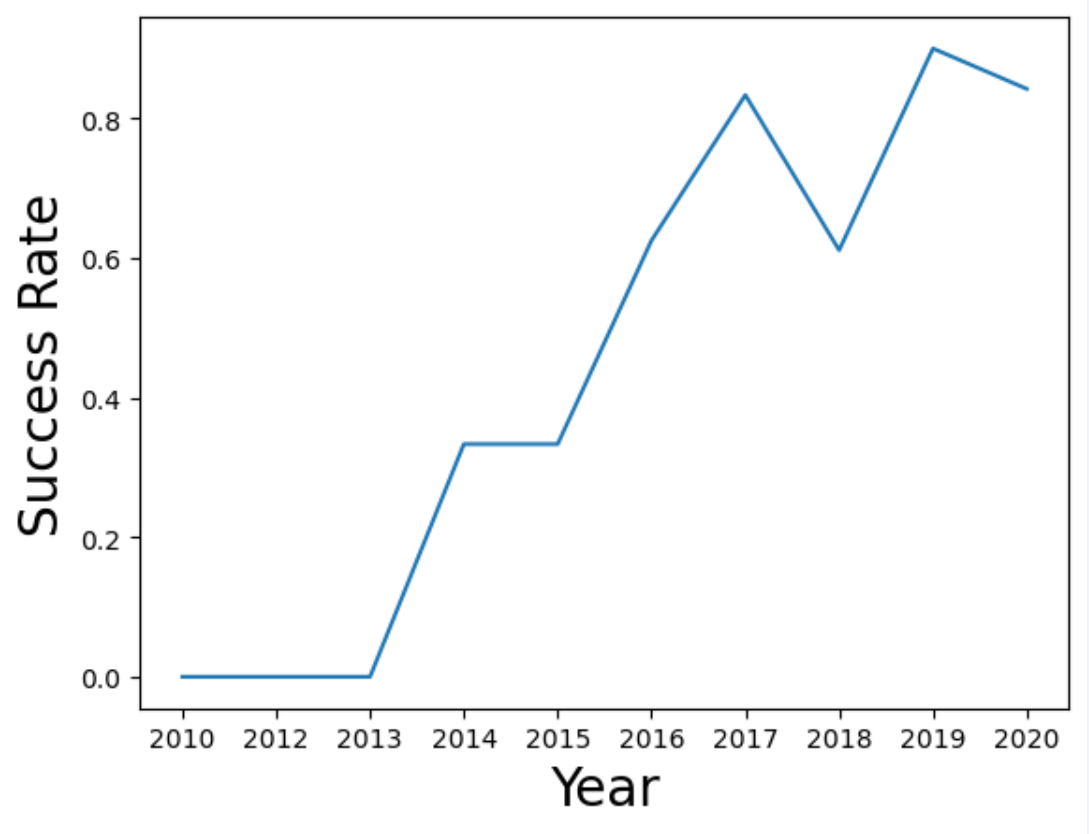


From the above plot, the orbits it is observed that LEO, ISS and SSO show an increase in successful landings with an increase in payload mass.

The remaining orbits' results appear to be inconclusive, particularly GTO.

The VLEO orbit tests include the highest payload masses, approximately higher than 13,000 kg.

Launch Success Yearly Trend



The graph indicates a clear increase in successful landings in correlation to an increase in years.

There is an observed stagnation between 2014 to 2015 and observed decreases between periods of 2017 to 2018 and 2019 to 2020.

All Launch Site Names

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

The image above displays the distinct launch sites utilized by SpaceX, determined to facilitate aggregating and disaggregating data and results obtained at various granularities.

Launch Site Names Begin with 'CCA'

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

The image above displays the results for launch sites beginning with “CCA”, indicated for filtering purposes of the records for specific launch sites.

Total Payload Mass

Total Payload Mass (kg)
45596

The image above displays the results for the total payload mass for all NASA launches, which was queried to potentially evaluate requirements per customer.

Average Payload Mass by F9 v1.1

Average Payload Mass (kg)
2534.67

The image above displays the results for the total payload mass for the F9 v1.1 booster version, which was queried to potentially evaluate performance per booster or their requirements.

First Successful Ground Landing Date

MIN(Date)
2015-12-22

The image above displays the results for the date when the first successful landing outcome on the ground pad was achieved, which was queried to potentially track any changes that were embedded in the latter missions.

Successful Drone Ship Landing with Payload between 4000 and 6000

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

The image above displays the name of the boosters which had successfully landed on the drone ship and had a payload mass greater than 4000 but less than 6000 kg, which was queried to potentially track any changes that were embedded in the latter missions.

Total Number of Successful and Failure Mission Outcomes

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

The image above displays the total number of successful and failed landing outcomes within the data set containing 101 records, it is observed that the successful landings accounted for 99% of the data set.

Boosters Carried Maximum Payload

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

The image to the left displays the all of the different booster versions that carried the maximum payload recorded. The relevant data could be valuable for future similar objectives.

2015 Launch Records

Month Name	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

The image above displays the list of the failed landing_outcomes in drone ships, their booster versions, and launch site names for in year 2015, which was queried to potentially track any of the variables that affected the landing outcome.

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

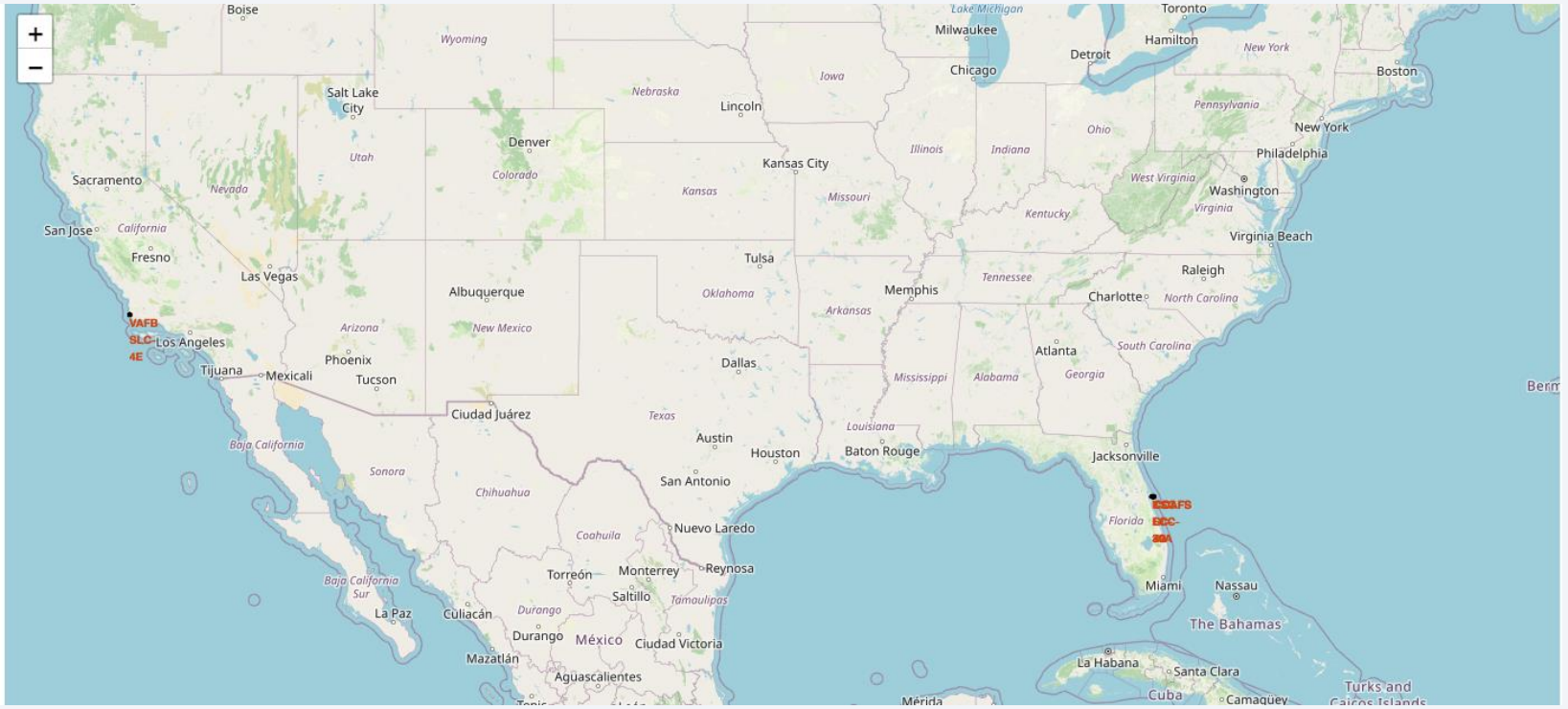
Landing_Outcome	Count
Failure (parachute)	30
Precluded (drone ship)	1

The image above displays the ranked 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, which was queried to determine the landing outcomes within the given period and determine the predominant outcomes.

Section 3

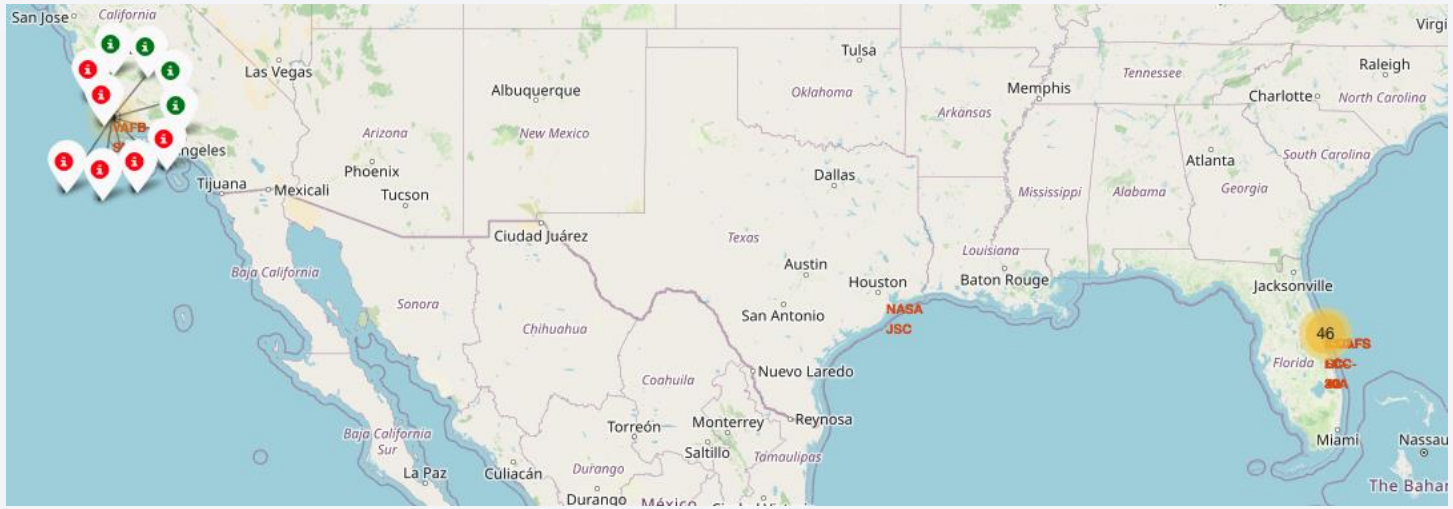
Launch Sites Proximities Analysis

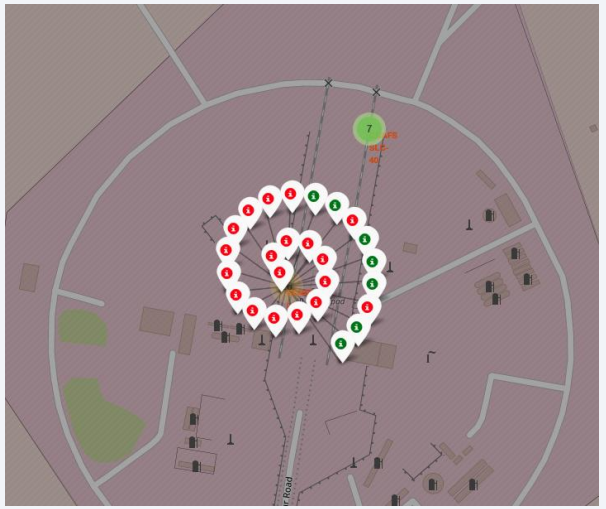
USA Folium Map with marked launch sites.



The image above shows a screenshot of the USA on a Folium map and indicates the marked launch sites used by SpaceX, which is the high-level geographical representation of the granular data available.

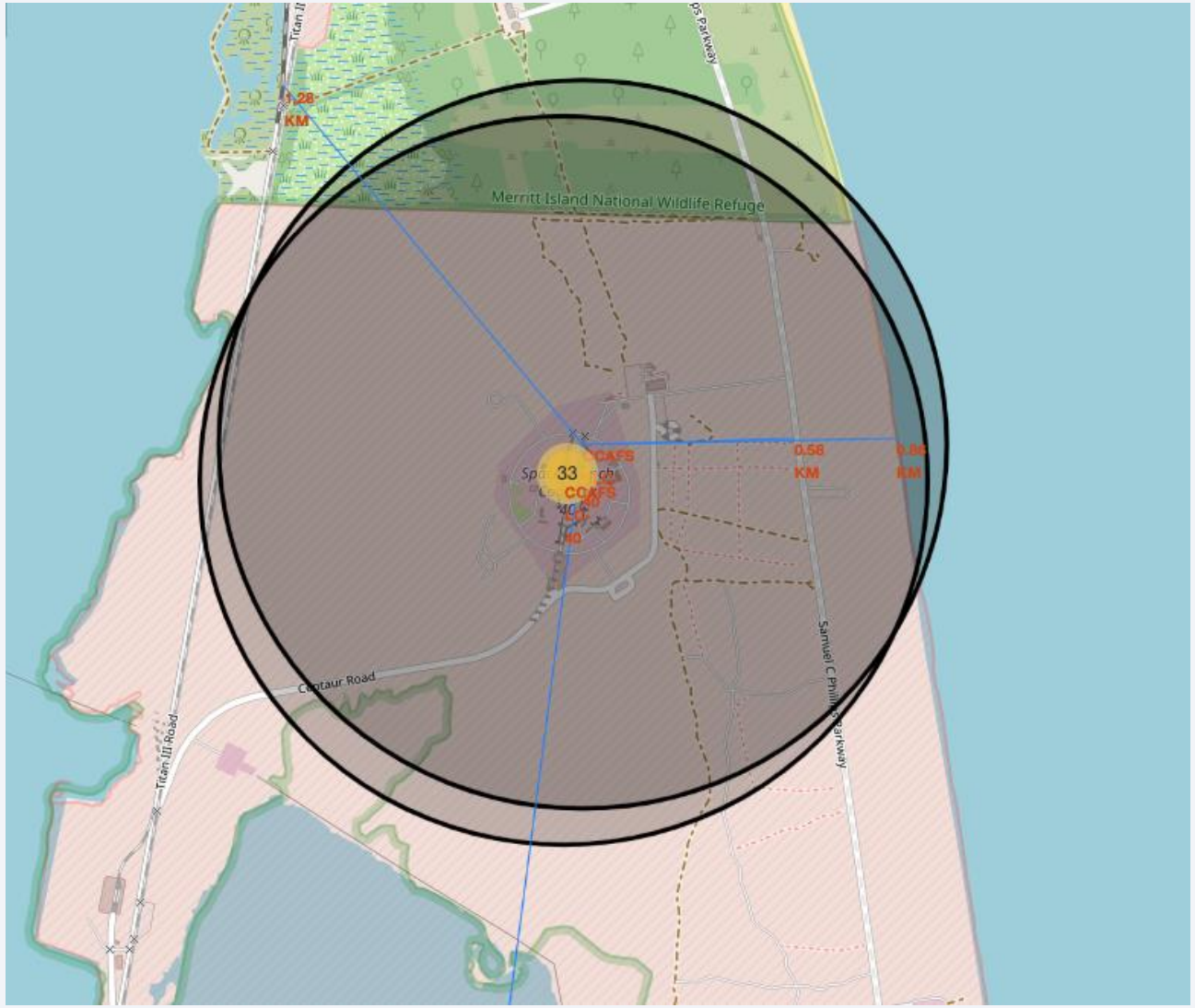
USA Folium Map with marked colour - coded launch outcomes.





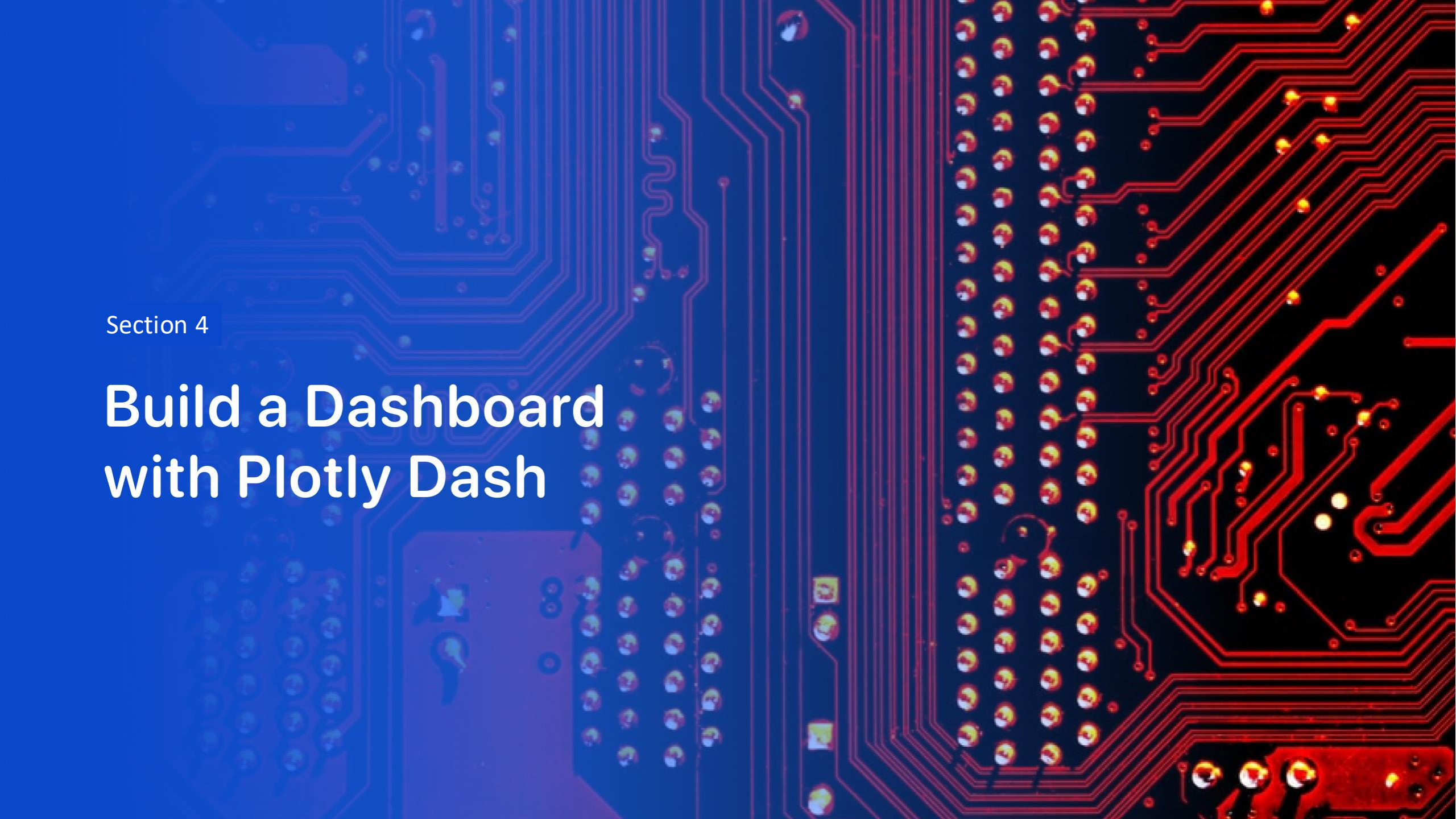
The images above shows screenshots of the USA on a Folium map and indicate the marked launch sites used and their colour-coded launch outcomes, which provides a graphical high-level view of the launch site with the most successful landing outcomes.

USA Folium Map with a selected launch site and its proximity to amenities.

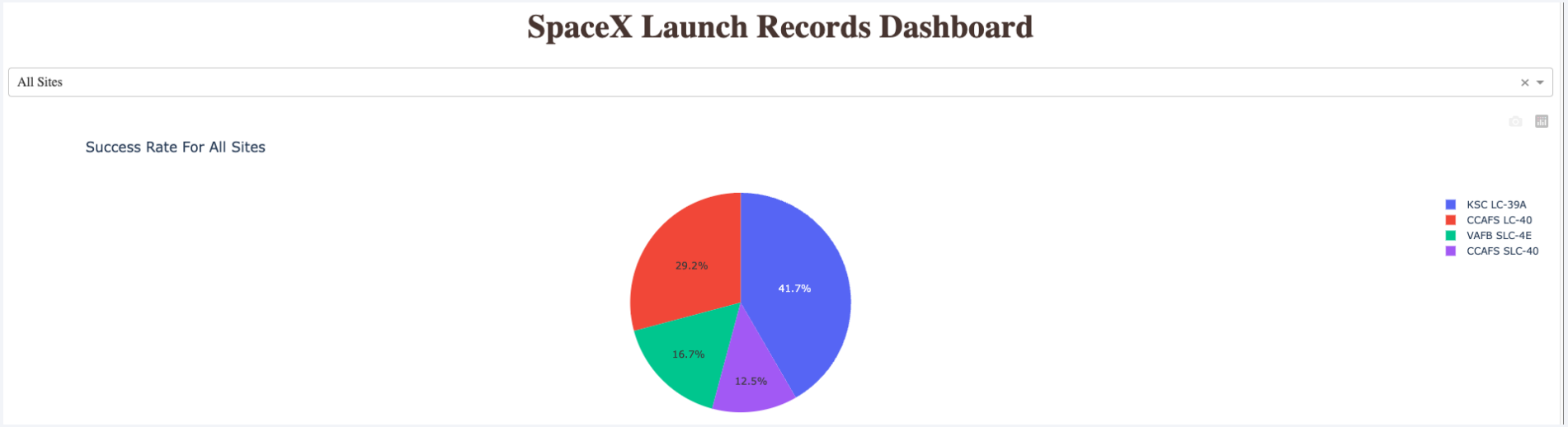


The image to the left shows a screenshot of the USA on a Folium map and indicates a selected launch site to its proximities such as railway, highway and coastline, with distance calculated and displayed.

The data can be used for optimizing the selection of a launch side based on its proximities to certain amenities.

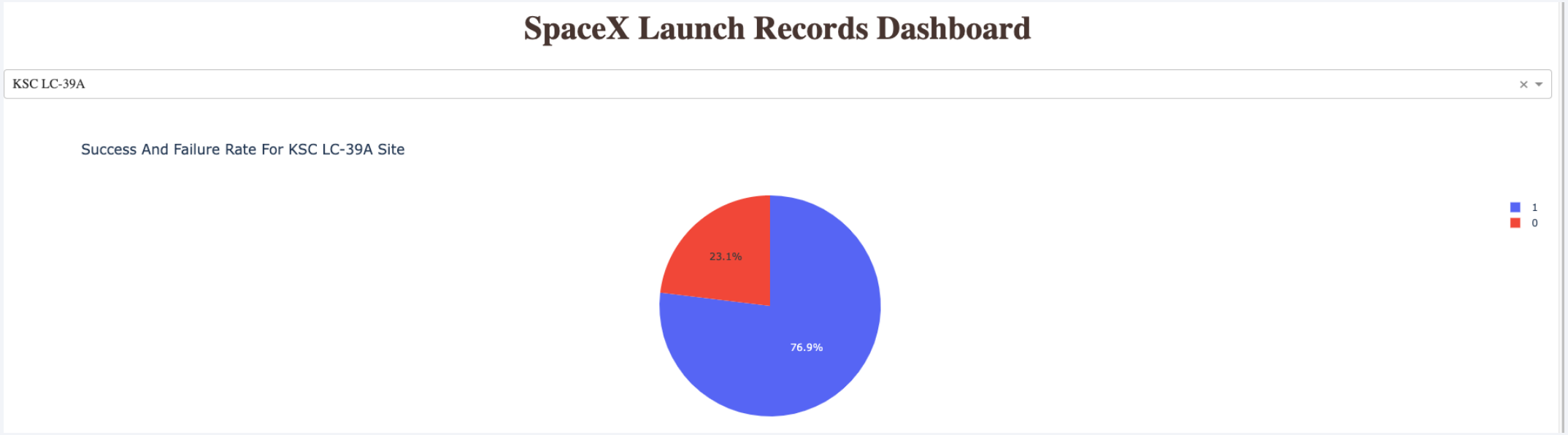

Section 4
Build a Dashboard with Plotly Dash

Pie chart with success ratio per launch site.



The image above indicates the results of a pie chart consisting of the contribution of the ratio of successful landing outcomes per launch site. It is observed that the launch site with the highest ratio of successful outcomes is KSC LC with an approximately 42% contribution.

Pie chart for KSC LC launch site success ratio.



The image above indicates the results of a pie chart consisting of the contribution of the ratio of successful landing outcomes for the KSC LC launch site, which had the highest success ratio. It is observed that the launch site has a success ratio of approximately 77% contribution.

payload range or booster version have the largest success rate, etc.

Scatter plot for the relationship between Payload Mass (kg) and Success Rate for All Sites.

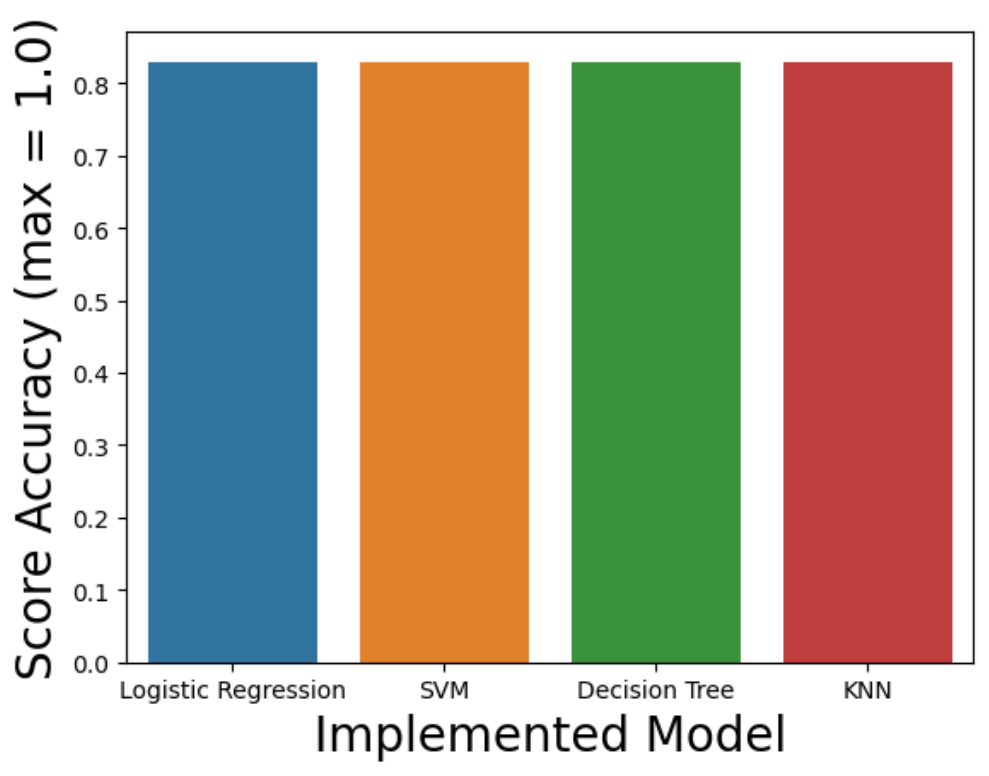
The images to the left show screenshots of the scatter generated indicating the relationship between Payload Mass (kg) and Success Rate for All Sites. The payload mass in the first image is varied between 0 to 10,000 kg and in the second image between 2500 to 7500 kg.

It is observed that FT is the best - performing booster version under both conditions. It is also observed that the v1.0 booster version was not used within the second scope of conditions.

Section 5

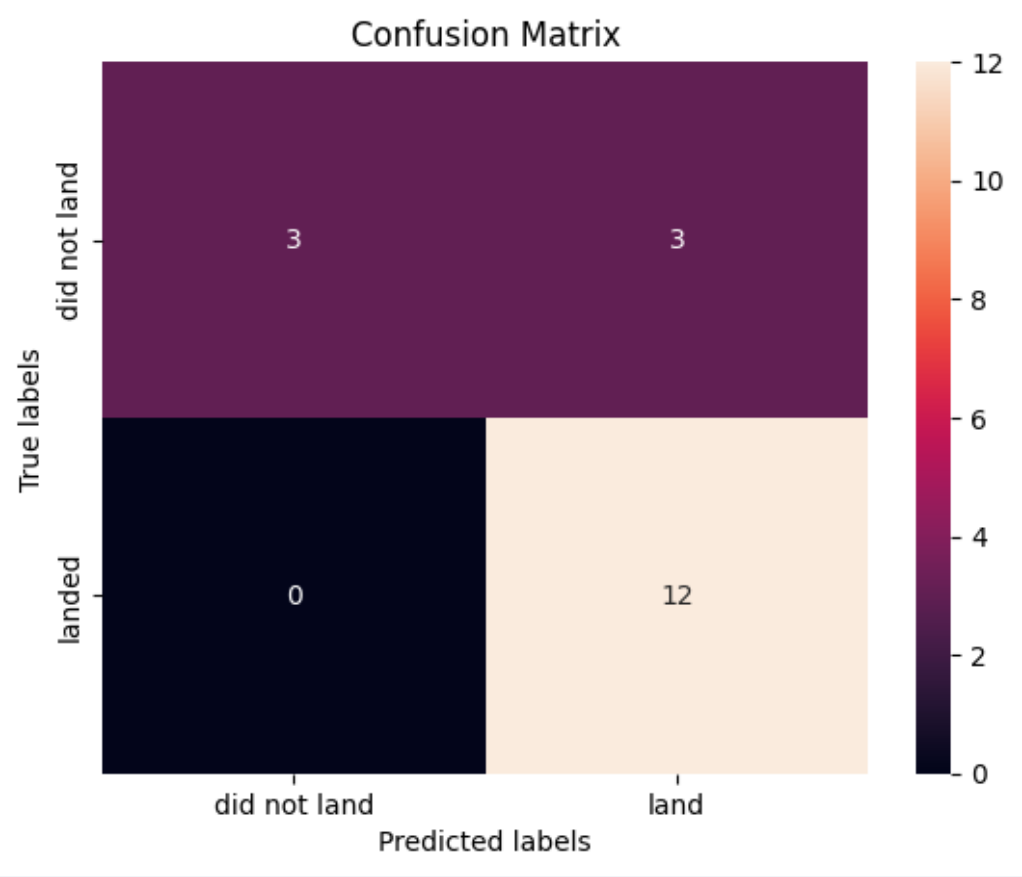
Predictive Analysis (Classification)

Classification Accuracy



- The above chart indicates that the accuracy calculated is the same at 83% across all the models.
- Confusion matrix to be used to determine the best - performing model.

Confusion Matrix



Precision: 0.8
Recall: 1
F1 Score: 0.89

- The above confusion matrix represents the identical results obtained for the Logistic Regression, SVM and K-Nearest Neighbours models. The obtained F1-Scores indicate the best performing models at 89% score.
- The Decision Tree Classifier model obtained an F1-Score of 88% which is marginally lower.

Conclusions

- The four models tested for predicting the landing outcome had identical score accuracies at 83%.
- The best models were chosen based on their confusion matrix results.
- The Logistic Regression, SVM and K-Nearest Neighbours had the highest F1 Score at 89%.
- Either of the three models can be used for predictions, future work can be done to narrow down the best - performing model and optimize results.

Appendix

- F1 Score calculation whereby:
- TP – True Positive
- FN – False Negative
- FP – False Positive
- TN – True Negative

```
TP = 11
FN = 1
FP = 2
TN = 4

Precision = (TP/(TP+FP))
Recall = TP/(TP+FN)
f1_score = (2*Precision*Recall) / (Precision + Recall)

print(f"Precision: {Precision}")
print(f"Recall: {Recall}")
print(f"F1 Score: {f1_score}")
```

Thank you!