DATA SCIENCE CAPSTONE PROJECT: SPACE Y

PREPARED BY MELANIE CHAN

OUTLINE

- EXECUTIVE SUMMARY
- INTRODUCTION
- METHODOLOGY
- RESULTS
- CONCLUSION
- APPENDIX

EXECUTIVE SUMMARY

This project is conducted to determine the cost of each launch of SPACE X and whether SPACE X will reuse the first stage. This information can be used for Space Y to bid against SpaceX for a rocket launch. The project utilized SPACE X's data obtained from API and webs craping to pull out insights to find specific features to maximize the success rate of launches and landing of the first stage of a rocket for the upcoming new company, SPACE Y.

This report highlights the important features and results that allow SPACE X to have a 100% mission outcome and factors to have its first stage to land successfully. Visual representation of the different launch sites is also displayed to show its geographical features and relationship with the success landing rates. Lastly, the report highlights the best prediction models to predict success landings.

As a result, a few recommendations were reported on the best features and launch sites for SPACE Y based on the insights obtained by SPACE X.

INTRODUCTION

PROJECT SCENARIO & OVERVIEW

With the rise of commercial space, companies are making space travel affordable for everyone. Space Y is new company that aims to compete with Space X. Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage. However, the first stage sometimes does not land, crash or is sacrificed by Falcon X.

Objective of this project is to determine the cost of each launch of SPACE X and whether SPACE X will reuse the first stage. This information can be used for Space Y to bid against SpaceX for a rocket launch.

It is required to find the optimum way to estimate the total cost of each launch and make predictions successful landings of the first stage.

METHODOLOGY

EXECUTIVE SUMMARY

- Data Collection methodology:
  - Data from Space X was obtained from 2 sources:
    - Space X API (https://api.spacexdata.com/v4/rockets/)
    - Web Scraping from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)
- Perform Data Wrangling
  - Collected data was enriched and preprocessed by creating a landing outcome label based on insights after summarizing and analyzing features

METHODOLOGY

EXECUTIVE SUMMARY

- Perform Exploratory Data Analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Four different classification models were used and evaluated to provide predictions with the highest accuracy

DATA COLLECTION - SPACE X API

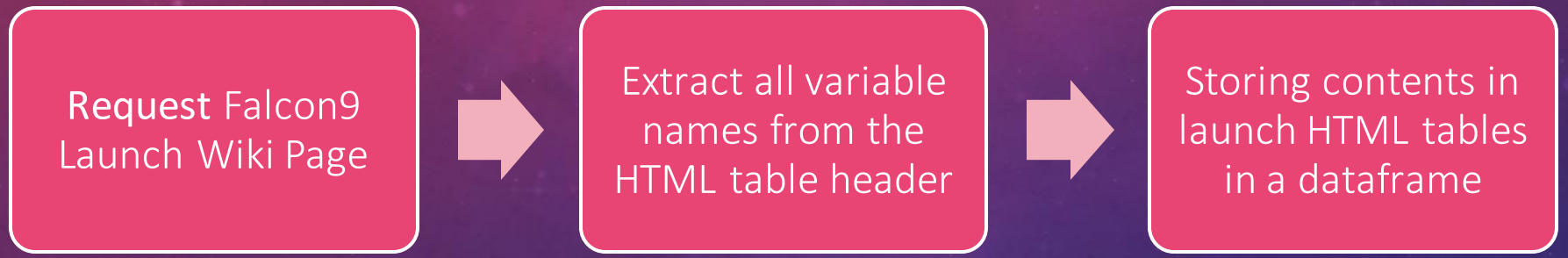
- SpaceX offers a public API where data can be obtained and used
- API was used according to the flowchart shown below:

- Source code can be found here:

  https://github.com/melchanpc/datasciencerepo/blob/main/applieddatascience/jupyter-labs-spacex-data-collection-api.ipynb

DATA COLLECTION - WEB SCRAPING

- Data from SpaceX launches can also be obtained from Wikipedoa
- Data are downloaded from Wikipedia according to the flowchart below:



- Source code can be found here:

  https://github.com/melchanpc/datasciencerepo/blob/main/applieddatascience/jupyter-labs-webscraping.ipynb

DATA WRANGLING

- In the dataset, there are several scenarios where booster land successfully at different regions or land unsuccessfully at different regions. A few significant scenarios are True Ocean, True RTLS and True ASDS represented that the booster landed successfully to a specific region of ocean, ground pad and drone ship respectively whereas False Ocean, False RTLS and False ASDS represented that the booster landed unsuccessfully to a specific of ocean, group pad and drone ship respectively.
- Summarizations of the dataset consisted of the number of launches on each site, the number and occurrences of each orbit and the number and occurrences of mission outcome per orbit type.

DATA WRANGLING

- Outcomes stated were converted into training labels with 1 which meant the booster successfully landed or 0 which meant it was unsuccessful.

- Source code can be found here:

  https://github.com/melchanpc/datasciencerepo/blob/main/applieddatascience/IBM-DS0321EN-SkillsNetwork_labs_module_1_L3_labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

EDA WITH DATA VISUALIZATION

- To explore data, scatterplots, bar plots and line plots were used to visualize the relationship between pair of features and any trend:
  - Payload Mass vs. Flight Number, Launch Site vs. Flight Number, Launch Site vs. Payload Mass, Orbit vs Success Rate, Flight Number vs. Orbit Type, Payload Mass vs. Orbit Type and Success Rate Yearly Trend
- Feature engineering is performed to obtain some preliminary insights about how each important variable would affect the success rate
- Source code can be found here:

  https://github.com/melchanpc/datasciencerepo/blob/main/applieddatascience/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA WITH SQL

- The following queries were performed:
  - Displaying names of unique launch sites in the space mission
  - Displaying 5 records where launch sites begin with the string 'CCA'
  - Displaying the total payload mass carried by boosters launched by NASA (CRS)
  - Displaying the average payload mass carried by booster version F9 v1.1
  - Listing the date when the first successful landing outcome in group pad was achieved
  - Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - Listing the total number of successful and failure mission outcomes
  - Listing the names of the booster versions which have carried the maximum payload mass
  - Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
  - Ranking the count of landing outcomes between date 2010-06-04 and 2017-03-20 in descending order
- Source code can be found here:

  https://github.com/melchanpc/datasciencerepo/blob/main/applieddatascience/jupyter-labs-eda-sql-coursera_sqllite.ipynb

BUILD AN INTERACTIVE MAP WITH FOLIUM

- Markers, circles, lines and marker clusters were used in the map with Folium
  - Marker with circle, pop up label and text label of NASA Johnson Space Centre using its latitude and longitude coordinates as a start location
  - Markers with circle, pop up label and text label of all launch sites using their laitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts
  - Colored markers were added to indicate success and failed launches and marker cluster to identify launch sites with relatively high success rates
  - Colored lines were added to show distances between the launch sites and its proximities, eg. Railway, Highway, Coastline and closest city
- Source code can be found here:

  https://github.com/melchanpc/datasciencerepo/blob/main/applieddatascience/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

BUILD DASHBOARD WITH PLOTLY DASH

- The following graphs and plots were built to visualize the data:
  - Percentage of successful launches by site
  - Payload mass range
  - Correlation between payload and success for all sites
- The graphs and plots allowed to quickly analyze the relation between payloads and launch sites and help to identify the best site to launch according to the payload in one dashboard
- Source code can be found here:

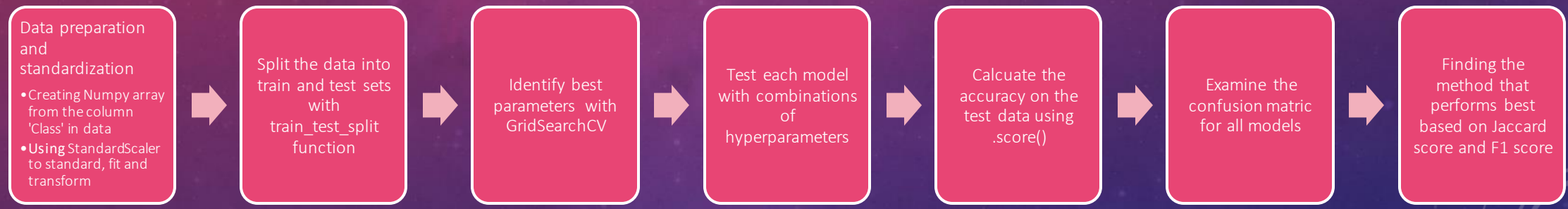
  https://github.com/melchanpc/datasciencerepo/blob/main/applieddatascience/spacex_dash_app.py

PREDICTIVE ANALYSIS (CLASSIFICATION)

- Four classifications models were compared
  - Logistic regression
  - Support vector machine
  - Decision tree
  - K nearest neighbors
- Accuracy for each model were calculated, confusion matrix were performed and the best method was determined based on their Jaccard Score and F1 score.

PREDICTIVE ANALYSIS (CLASSIFICATION)

- Technical methodology is shown in the flowchart below



- Source code can be found here:

  https://github.com/melchanpc/datasciencerepo/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

RESULTS

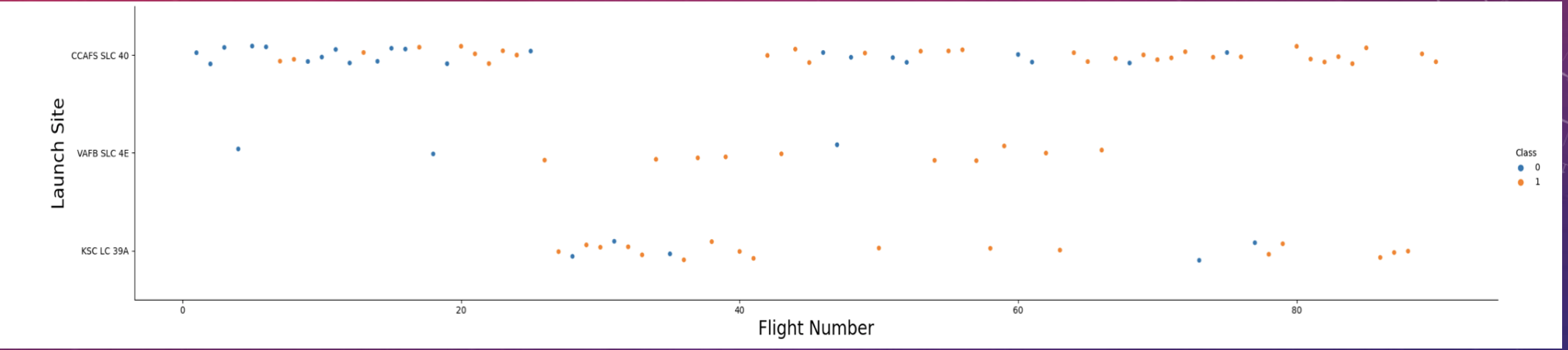
THE FOLLOWING SLIDES PROVIDES INSIGHTS FROM THE DATASET AND DISPLAYED IN 4 SECTIONS:

- INSIGHT DRAWN FROM EXPLORATORY DESCRIPTIVE ANALYSIS (EDA)
  - EDA WITH VISUALIZATION
  - EDA WITH SQL
- LAUNCH SITES PROXIMITIES ANALYSIS
- DASHBOARD WITH PLOTLY DASH
- PREDICTIVE ANALYSIS (CLASSIFICATION)

INSIGHT DRAWN FROM EXPLORATORY DESCRIPTIVE ANALYSIS (EDA)

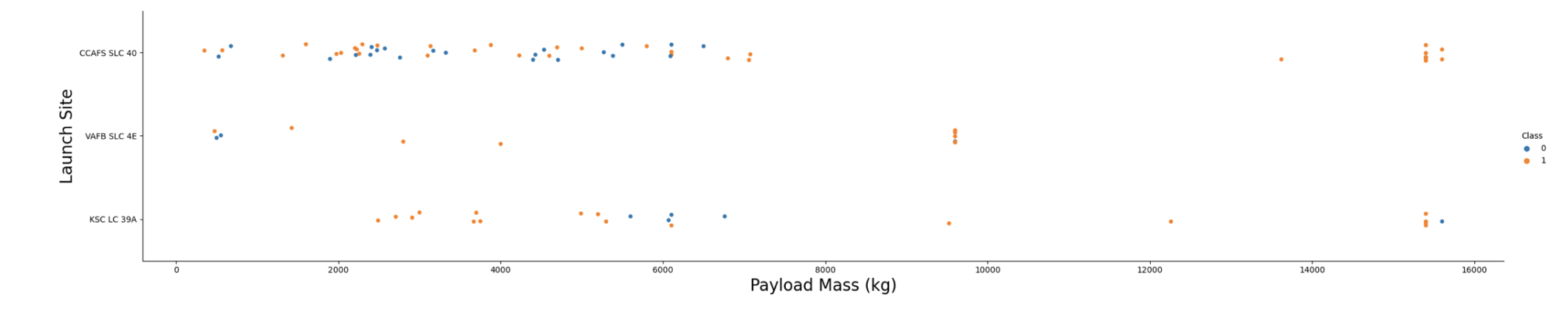
EDA WITH VISUALIZATION

FLIGHT NUMBER VS. LAUNCH SITE

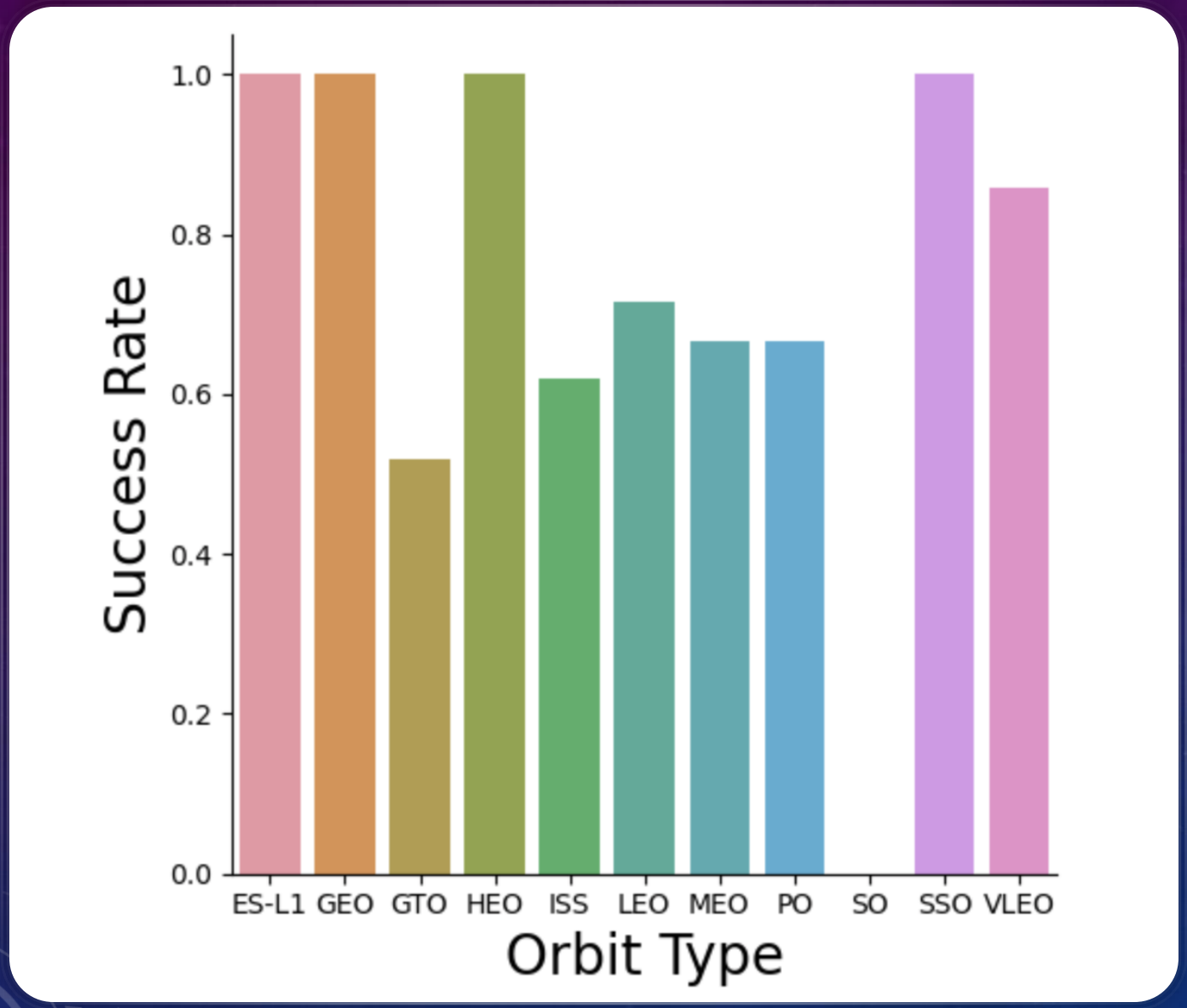


- From the plot above, it is shown that as the flight number gets larger, the success rate for the first stage to land increases for all sites.
- It can assumed that each new launch incur a higher rate of success
- VAFB SLC 4E and KSC LC 39A have more consistent number of success rate as the flight number increases
- The earliest launch done by CCAFS SLC 40 had more failed landing than successful ones but the successful attempts gradually increased as the flight number increased.
- The geographical characteristics of the launch site may possibly affect its success rate eg. Wind speed, weather

PAYLOAD MASS VS. LAUNCH SITE



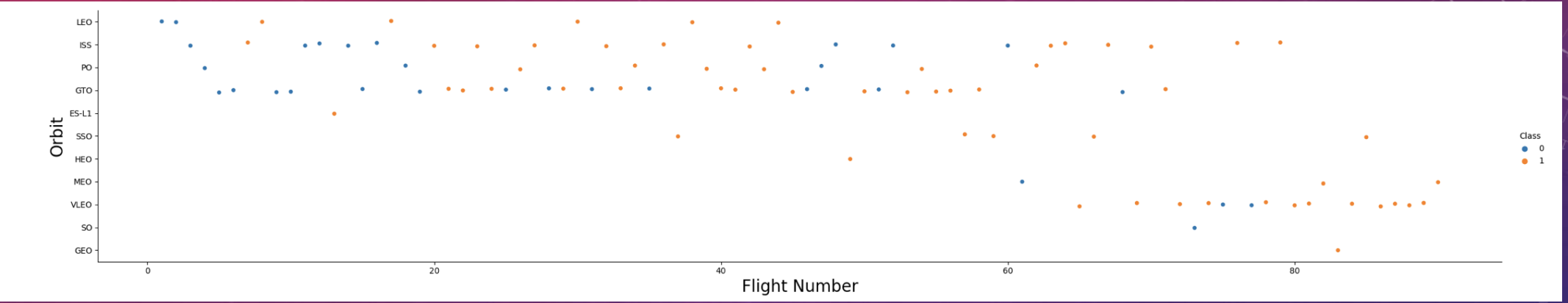
- From the plot above, CCAFS SLC 40 is more favorable launch site as there are rockets launched for heavy payload mass(greater than 10000) and first stage landed successfully.
- KSC LC 39A seemed to be a suitable site for rockets with payload between 2000kg to 5500kg as their success rate is at 100%



SUCCESS RATE VS. ORBIT TYPE

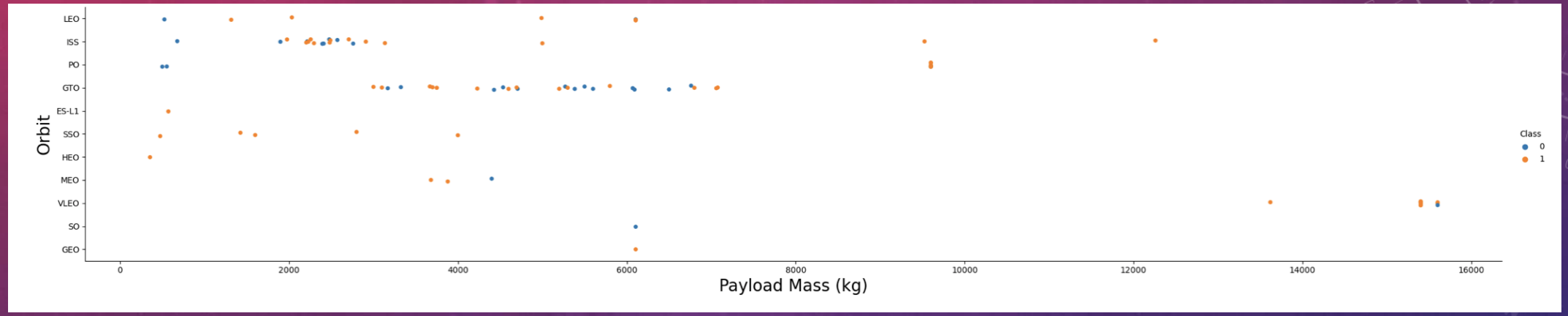
- ES-L1, GEO, HEO and SSO are the best orbits to launch rockets as it has 100% success rate for the first stage to land.
- VLEO would be a secondary option as the success rate is at 85%.

FLIGHT NUMBER VS. ORBIT TYPE



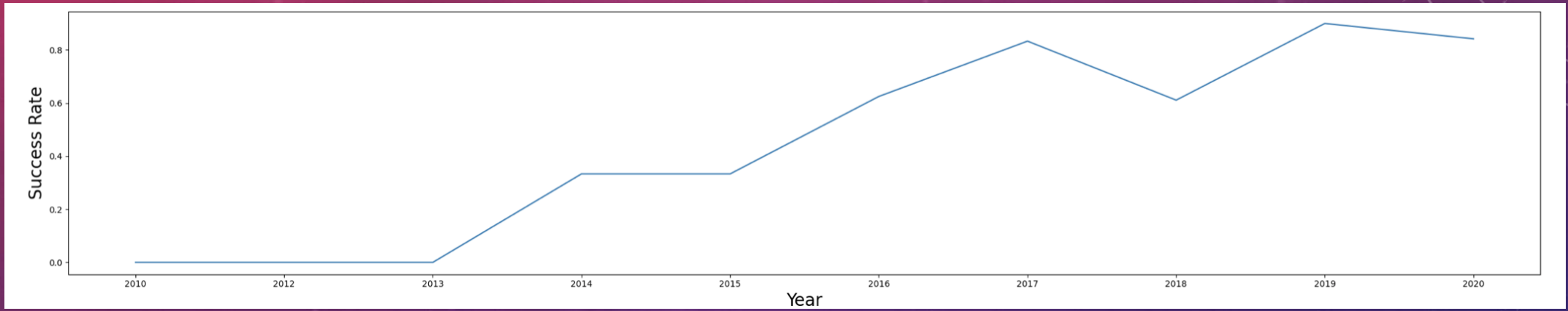
- There seems to be no relationship between flight number when in GTO orbit
- In the Leo Orbit, the success rate seems to be increased as the flight number increased

PAYLOAD MASS VS. ORBIT TYPE



- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- However, for GTO it cannot be distinguished this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

YEARLY TREND OF LAUNCH SUCCESS RATE



- From the plot above, the success rate since 2013 kept increasing till 2020.

INSIGHT DRAWN FROM EXPLORATORY DESCRIPTIVE ANALYSIS (EDA)

EDA WITH SQL

LAUNCH SITE NAMES

```
%sql select distinct launch_site from SPACEXDATASET;
```

```
 * ibm_db_sa://phb02330:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
   sqlite:///my_data1.db
Done.
```

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

- The above query showed the available launch sites in the dataset.

LAUNCH SITE THAT STARTS 'CCA'

```
%sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5;
```

```
 * ibm_db_sa://phb02330:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
   sqlite:///my_data1.db
Done.
```

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

- 2 out of 5 samples shown for CCAFS LC-40 have failure in the landing outcome.

TOTAL PAYLOAD MASS

```
%sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)';
```

```
 * ibm_db_sa://phb02330:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
   sqlite:///my_data1.db
Done.
```

total_payload_mass
45596

- Total payload mass carried by boosters launched by NASA (CRS) is 45596 kg.

AVERAGE PAYLOAD MASS BY F9 V1.1

```
%sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXDATASET where booster_version like '%F9 v1.1%';
```

```
 * ibm_db_sa://phb02330:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
   sqlite:///my_data1.db
Done.
```

average_payload_mass
2534

- The average payload mass carried by booster version F9 v1.1 is 2534 kg.

FIRST SUCCESSFUL GROUND LANDING DATE

```
%sql select min(date) as first_successful_landing from SPACEXDATASET where landing__outcome = 'Success (ground pad)';
```

```
 * ibm_db_sa://phb02330:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
   sqlite:///my_data1.db
Done.
```

first_successful_landing
2015-12-22

- The first successful landing outcome in ground pad was achieved was 2015-12-22.

SUCCESFUL BOOSTER IN DRONE SHIP WITH PAYLOAD MASS GREATER THAN 4000KG AND LESS THAN 6000KG

```
%sql select booster_version from SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 600(
```

```
 * ibm_db_sa://phb02330:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
   sqlite:///my_data1.db
Done.
```

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

- Boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000 are the F9 FT versions.
- The specific versions are listed in the query shown above.

MISSION OUTCOMES

```
%sql select mission_outcome, count(*) as total_number from SPACEXDATASET group by mission_outcome;
```

```
 * ibm_db_sa://phb02330:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
   sqlite:///my_data1.db
Done.
```

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- The number of success mission outcomes is 99, the failure in flight is 1 and success mission outcome with payload status unclear is 1.

BOOSTERS THAT CARRIED MAXIMUM PAYLOAD

```
%sql select booster_version from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXDATASET);
```

```
 * ibm_db_sa://phb02330:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
   sqlite:///my_data1.db
Done.
```

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 query above show the list booster versions which have carried the maximum payload mass.

LAUNCH RECORDS FOR THE YEAR 2015

```
%%sql select monthname(date) as month, date, booster_version, launch_site, landing__outcome from SPACEXDATASET
      where landing__outcome = 'Failure (drone ship)' and year(date)=2015;
```

```
 * ibm_db_sa://phb02330:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
   sqlite:///my_data1.db
Done.
```

MONTH	DATE	booster_version	launch_site	landing__outcome
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

- For the year 2015, the failed landing outcome consists of F9 v1.1 B1012 and F9 v1.1 B1015 at the launch site of CCAFS LC-40.

LANDING OUTCOME BETWEEN 2010-06-04 AND 2017-03-20

```
%%sql select landing__outcome, count(*) as count_outcomes from SPACEXDATASET
      where date between '2010-06-04' and '2017-03-20'
      group by landing__outcome
      order by count_outcomes desc;
```

```
 * ibm_db_sa://phb02330:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
   sqlite:///my_data1.db
Done.
```

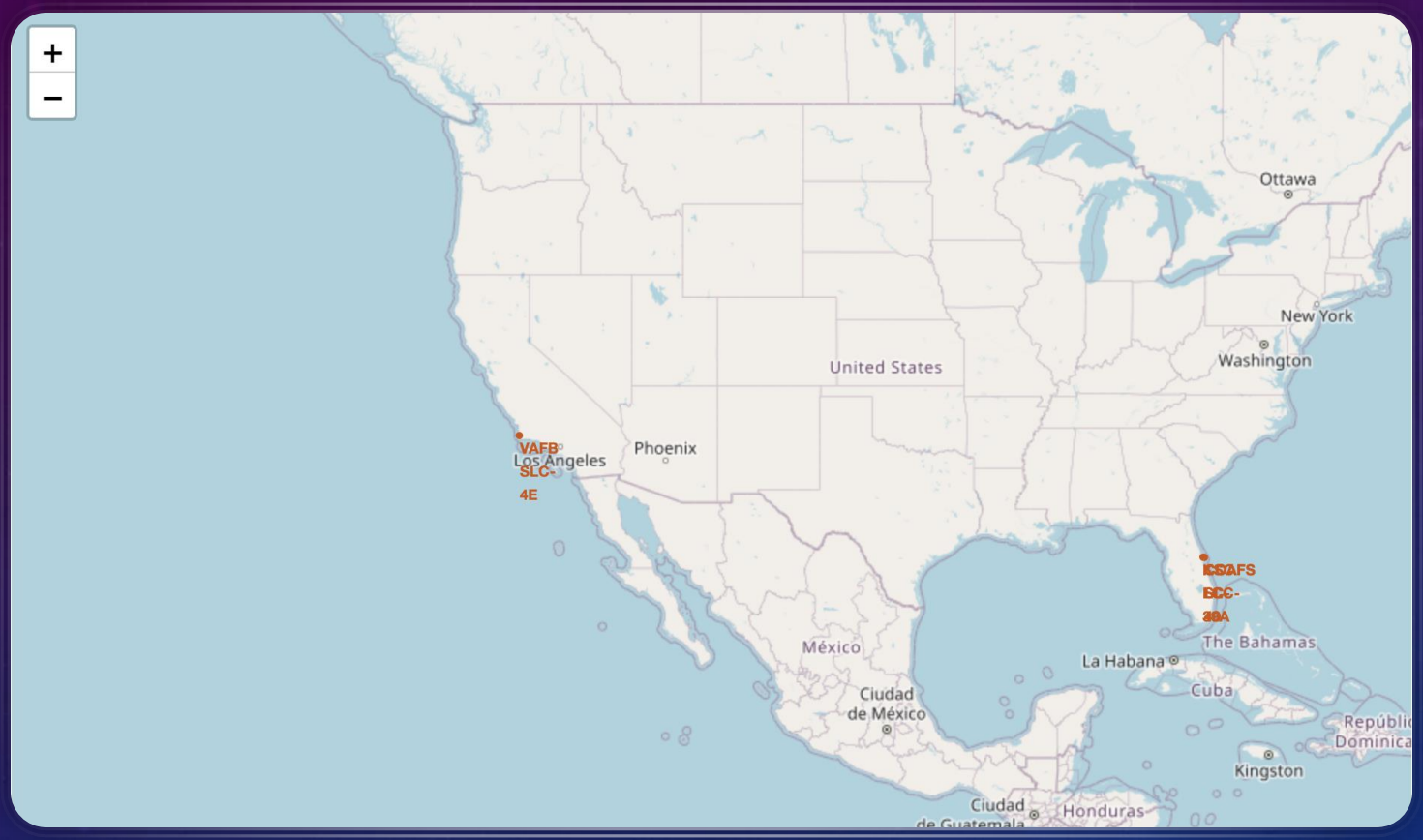
landing__outcome	count_outcomes
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

- In between 2010-06-04 and 2017-03-20, the highest number of outcomes is 'No attempt'. With further exploration, the rocket launch mission outcome was successful for all 'no attempt' landing outcome. More investigation is required for this landing outcome.

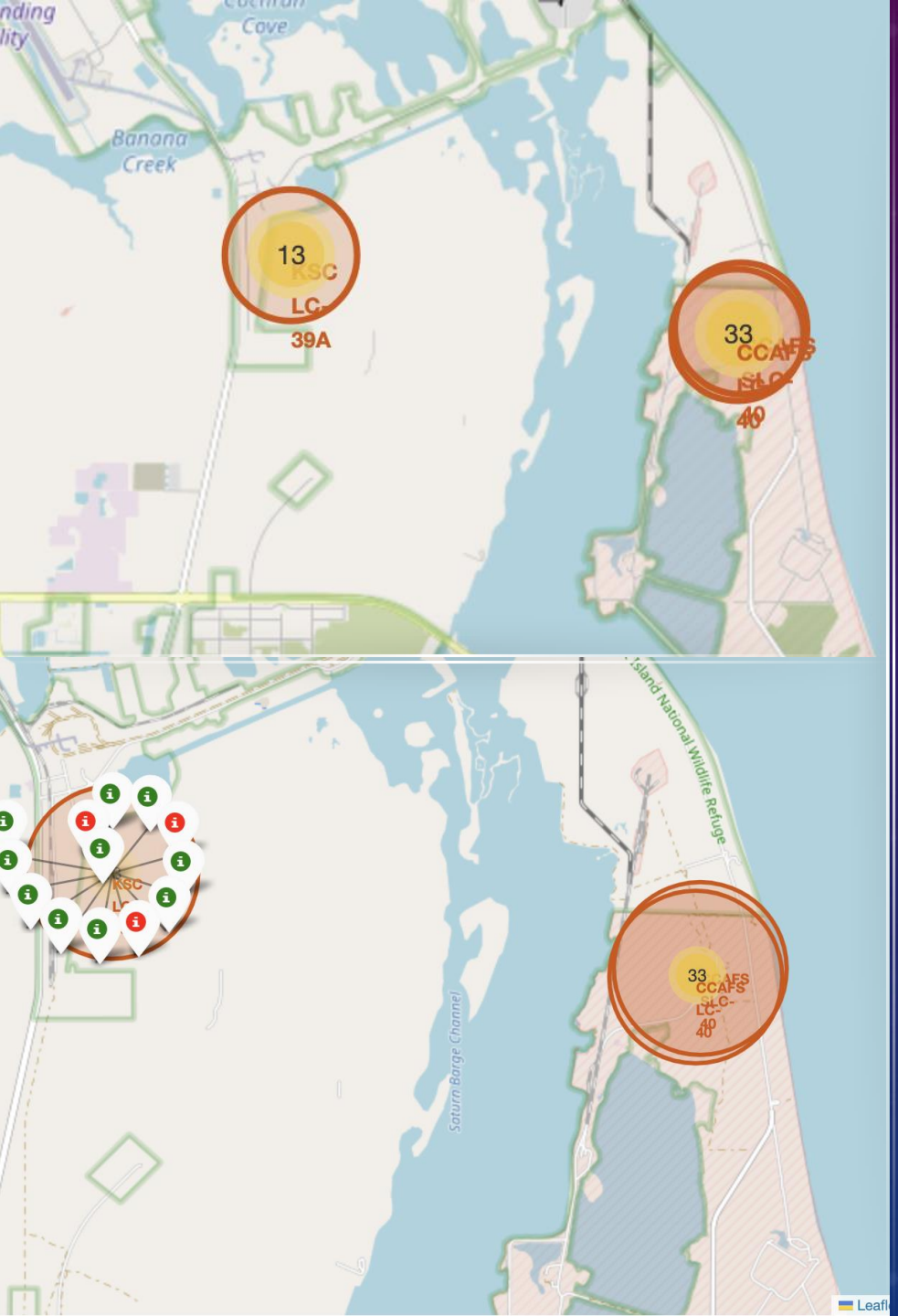
LAUNCH SITES PROXIMITIES ANALYSIS

THIS SECTION SHOWS THE GEOGRAPHICAL INSIGHTS OF ALL LAUNCH SITES AND DETAILS ON BEST LAUNCH SITE.

ALL LAUNCH SITES



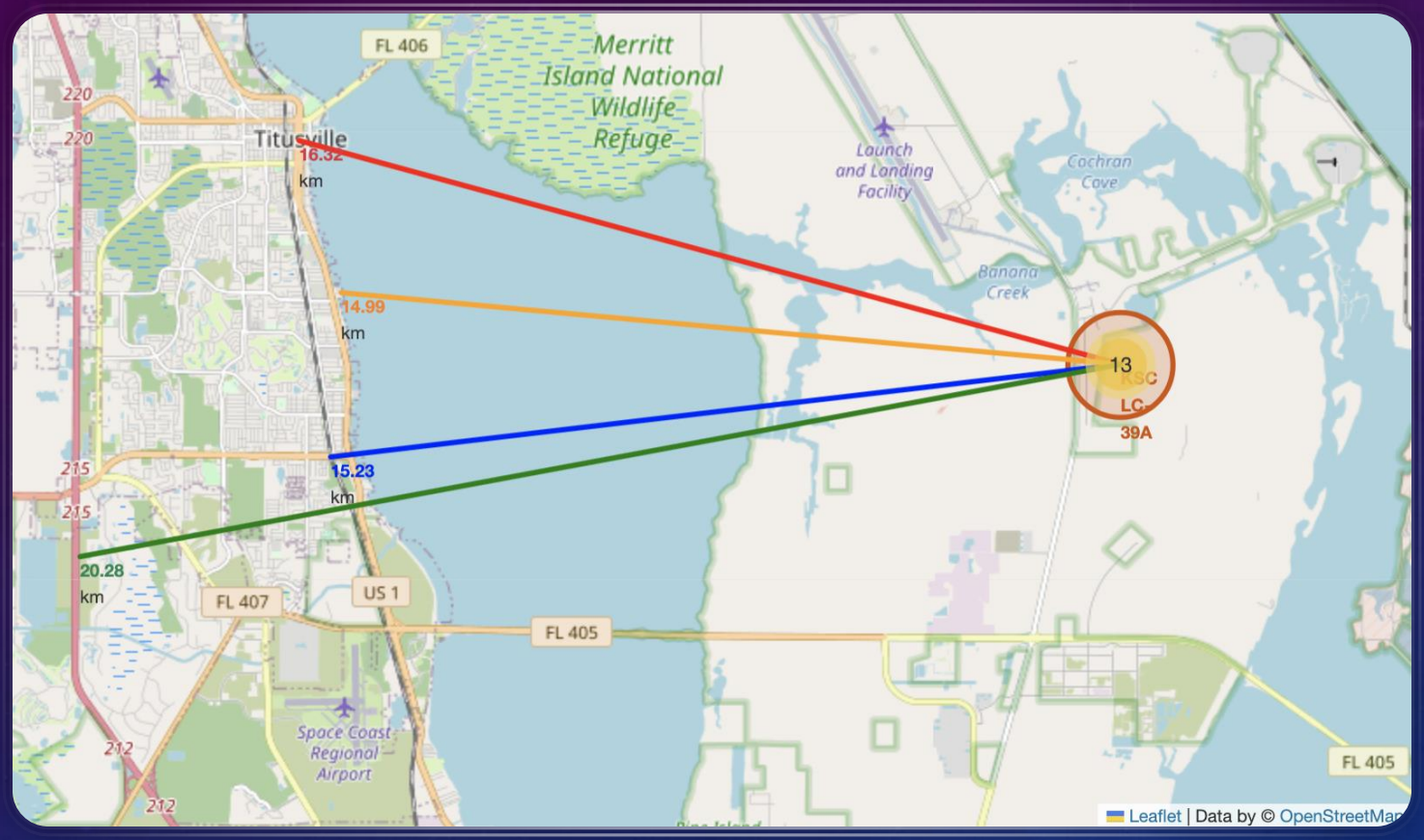
- All launch sites are in close proximity to the coast, while launching rockets towards the ocean it minimizes the risk of having any debris dropping or exploding near people on land.



LAUNCH OUTCOME BY SITE

- Launch Site KSC LC-39A has a very high success rate based on the high number of successful launches.
- From the colored-labeled markers on the image shown on the left, it can be determined which launch sites have relatively high success rates
  - Green Marker = Successful Launch
  - Red Marker = Failed Launch

DISTANCE FROM THE LAUNCH SITE & ITS PROXIMITIES

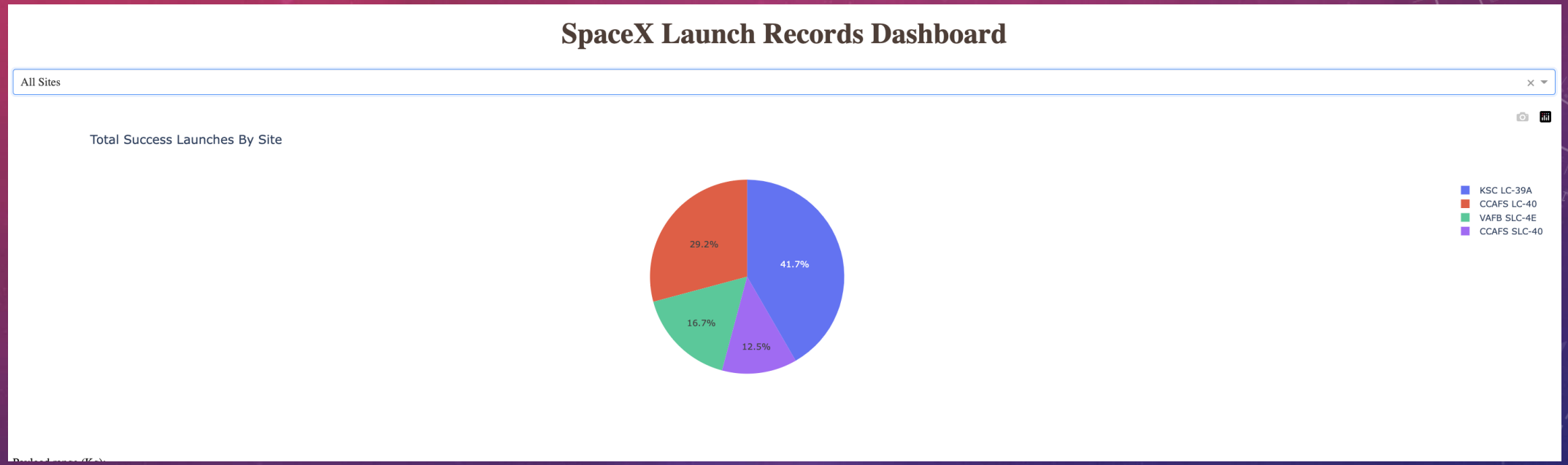


- From the visual map of the launch site KSC LC-39A shown on the left, it is observed that:
  - The site is relative close to railway with distance of 15.23km
  - The site is relatively close to highway with distance of 20.28km
  - The site is relatively close to coastline with distance of 14.99km
  - The nearest city from the site which is Titusville has a distance of 16.32km.
- Based on the Kennedy Space Centre (https://www.kennedyspacecenter.com/launches-and-events/events-calendar/see-a-rocket-launch) , a general distance from the launch site for public viewing is at 11km to 12km, it can be said KSC LC-39A is a safe distance from populated areas as the average distance is 16.6km away from populated areas.

DASHBOARD WITH PLOTLY DASH

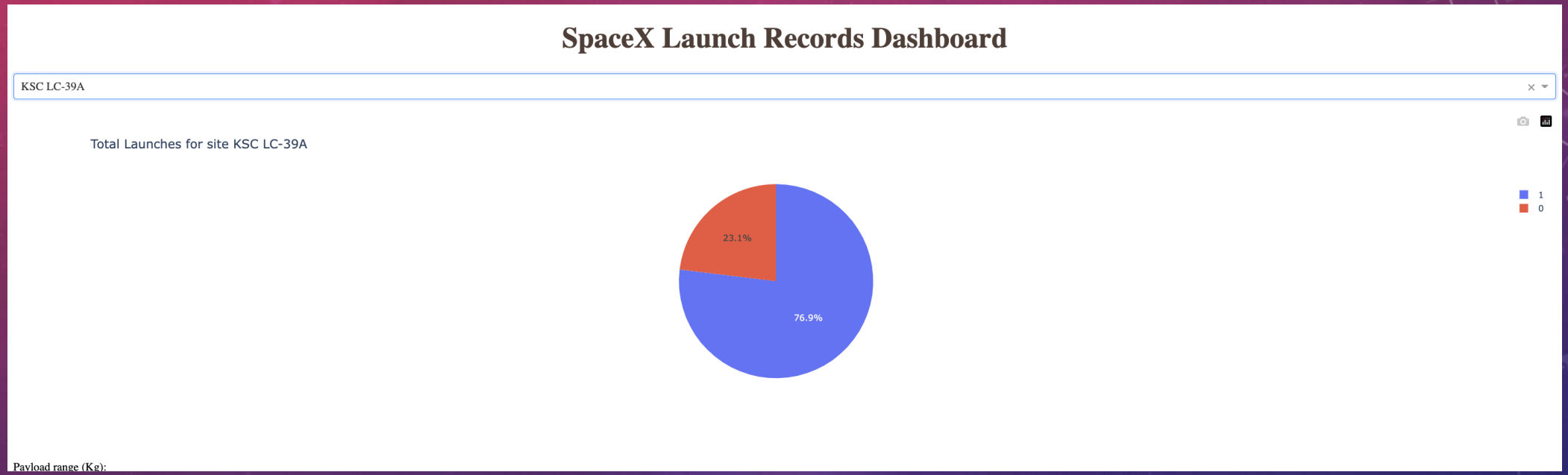
THIS SECTION SHOWS THE VISUAL REPRESENTATION OF THE RELATIONSHIP LAUNCH SITES, SUCCESS RATES AND ITS FEATURES.

LAUNCH SUCCESS DISTRIBUTION FOR ALL SITES



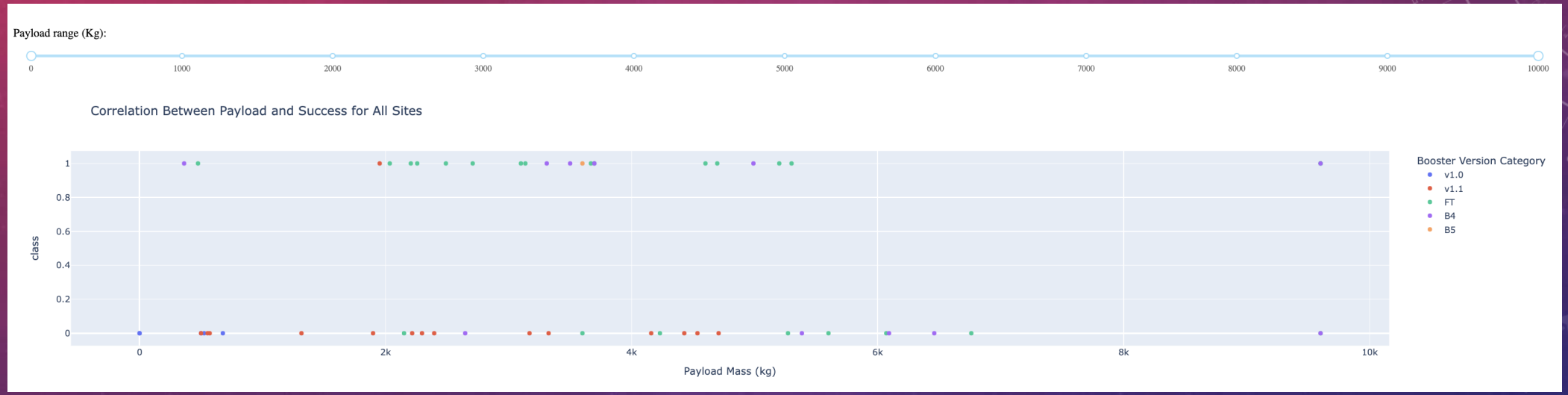
- Based on the pie chart that shows the total success launches by site, it is shown that KSC LC-39A has the highest successful launches.

SUCCESS AND FAILURE RATE FOR KSC LC-39A



- Based on pie chart that shows the distribution of successful and unsuccessful launches for KSC LC-39A, it is shown that KSC LC-39A has the largest successful launches with 77% of success rate and only 23.1% of failure rate.

PAYLOAD MASS VS. LAUNCH OUTCOME FOR ALL SITES



- The chart shows the correlation between payload and success for all sites, it can be proven that payloads between 2000kg and 6000kg have the highest success rate
- Majority of FT boosters shows highest launch success rate, especially in the payload range of 2000kg to 6000kg (class = 1)

PREDICTIVE ANALYSIS (CLASSIFICATION)

THIS SECTION SHOWS THE BEST MODEL FOR PREDICTION BASED ON ACCURACY METRICES

ACCURACY SCORES

- Four classification models were tested on a test set of sample size 18 and whole data set
- Based on the scores of the test set, no models can be confirmed the best as they have the similar scores. This may be due to the small test sample size.
- When the models are tested with the whole dataset, it is shown that Decision Tree model is the best method to use as it has the highest Jaccard, F1 and accuracy scores.

Accuracies scores from Test sets for all models

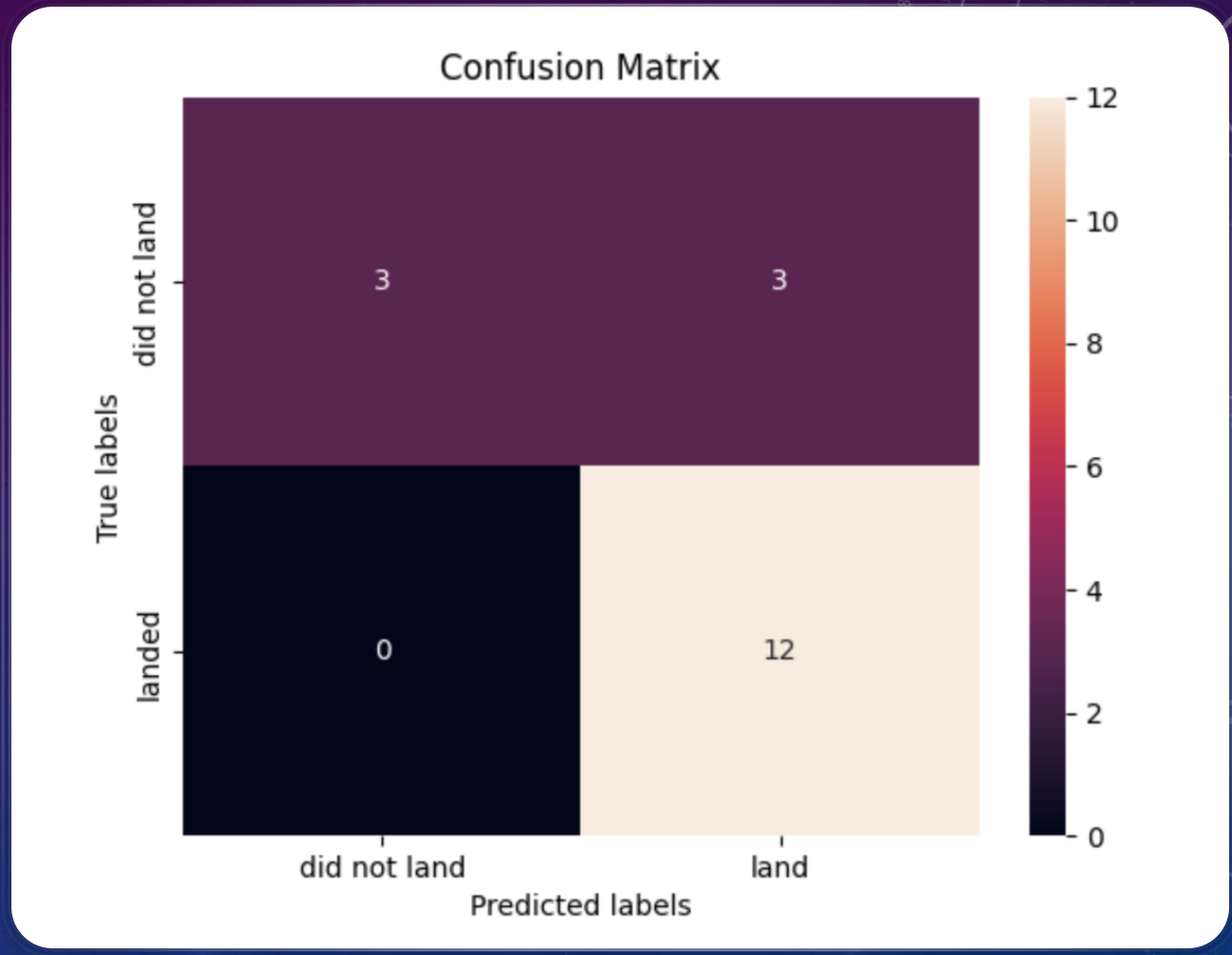
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

Accuracies scores from whole dataset for all models

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.819444	0.819444
F1_Score	0.909091	0.916031	0.900763	0.900763
Accuracy	0.866667	0.877778	0.855556	0.855556

CONFUSION MATRIX

- The confusion matrix shows the summary of the prediction results on the test dataset for all models
- All the models have the same confusion matrix.
- From the plot shown on the left, the models proves its accuracy by showing large numbers of true positive and true negative.
- However, there is an issue on the number of false positives that needs to be considered in future modelling.



CONCLUSION

- For SPACE Y to bid against SPACE X, it is recommended for the rockets or launches to take the following into considerations:
  - The rocket should have low payload mass, somewhere between 2000kg to under 6000kg. This is because this range of payload mass shows higher success rate that larger payload mass (over 7000kg)
  - Orbits ES-L1, GEO, HEO and SSO are good options to launch as these orbit types have 100% success rate.
  - Launch sites closer to the coastal area and at least 16km away from populated areas is ideal to avoid any possible fallen debris.
  - The launch site KSC LC 39A would be optimum launch site as it showed this highest success rate of the other launch site.
  - If the launches is designed with heavy payload mass, CCAFS SLC 40 is more favorable launch site as there are rockets launched for heavy payload mass(greater than 10000) and first stage landed successfully.
  - Decision Tree Classifier can be used to predict successful landings with the features of the rockets designed for Space Y. This is because it has the highest accuracy results compared to other models

APPENDIX

- https://github.com/melchanpc/datasciencerepo/blob/main/applieddatascience/jupyter-labs-spacex-data-collection-api.ipynb
- https://github.com/melchanpc/datasciencerepo/blob/main/applieddatascience/jupyter-labs-webscraping.ipynb
- https://github.com/melchanpc/datasciencerepo/blob/main/applieddatascience/IBM-DS0321EN-SkillsNetwork_labs_module_1_L3_labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb
- https://github.com/melchanpc/datasciencerepo/blob/main/applieddatascience/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb
- https://github.com/melchanpc/datasciencerepo/blob/main/applieddatascience/jupyter-labs-eda-sql-coursera_sqllite.ipynb
- https://github.com/melchanpc/datasciencerepo/blob/main/applieddatascience/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb
- https://github.com/melchanpc/datasciencerepo/blob/main/applieddatascience/spacex_dash_app.py
- https://github.com/melchanpc/datasciencerepo/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

THANK YOU!