IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Rayen Ben Soltane
28/12/2022

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results

The Data Science methodology involving data collection, data wrangling, exploratory data analysis, data visualization, model development, model evaluation, and reporting your results to stakeholders.

Introduction

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, much of the savings is because SpaceX can reuse the first stage.

Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

The main goal of this project was to predict if the Falcon 9 first stage will land successfully.

I have been tasked with predicting if the first stage of the SpaceX Falcon 9 rocket will land successfully. With the help of my Data Science findings and models, the competing startup I have been hired by can make more informed bids against SpaceX for a rocket launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The first sept is collecting data. For that we need to request rocket launch data from SpaceX API.
- Perform data wrangling
  - We can see below that some of the rows are missing values in our dataset. Multiple methods exist to replace these empty value cells. In our project we use the mean value.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- Different endpoints are associated to this API such as ‘capsules’, ‘cores’ and ‘launches’. This API will give us data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.
- We will perform a get request using the requests library to obtain the launch data, which we will use to get the data from the API. This result can be viewed by calling the .json() method. Now we decode the response content as a Json using .json() and turn it into a Pandas dataframe using .json_normalize().

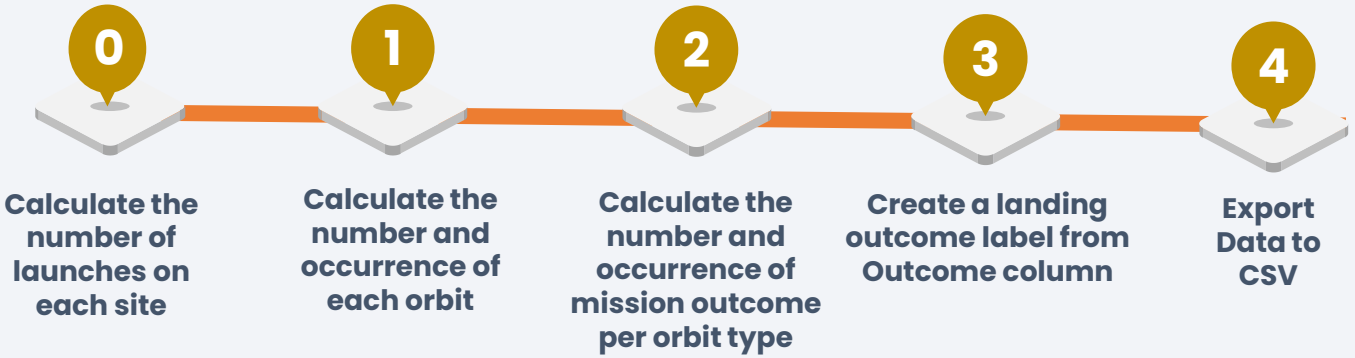
Data Collection – SpaceX API

- We used the get request to the SpaceX API to collect data, clean the requested data
- The link to the notebook is the following : https://github.com/Ray-BaToDs/IBM-Applied-Data-Science-Capstone/blob/master/Week%201-%20Mission%201%20-Data%20Collection%20API%20Lab.ipynb

Data Collection - Scraping

- We applied web scrapping to webscrap Falcon 9 launch records with BeautifulSoup
- We parsed the table and converted it into a pandas dataframe.
- The link to the notebook is the following : https://github.com/Ray-BaToDs/IBM-Applied-Data-Science-Capstone/blob/master/Week%201-%20Mission%202%20-jupyter-labs-webscraping.ipynb

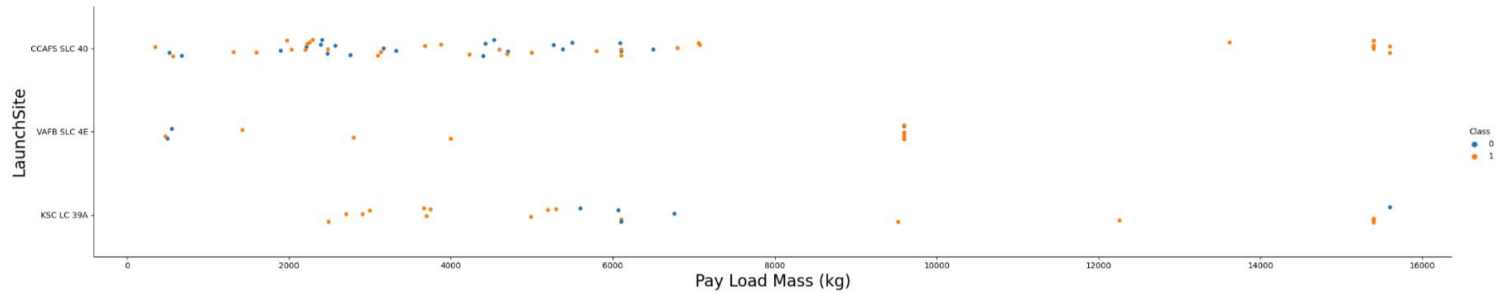
Data Wrangling



- The main goal of this process is to Perform exploratory Data Analysis and determine Training Labels. For that we need to convert landing outcomes into Training Labels with 1 meaning the booster successfully landed and 0 meaning it was unsuccessful.
- The link to the notebook is the following : https://github.com/Ray-BaToDs/IBM-Applied-Data-Science-Capstone/blob/master/Week%201-%20Mission%203%20-IBM-DS0321EN-SkillsNetwork_labs_module_1_L3_labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

EDA with Data Visualization

- We explored the data by visualizing the relationship between flight number and launch Site, payload and launch site, success rate of each orbit type, flight number and orbit type, the launch success yearly trend. Multiple pertinent observations were noted from these visualizations.
- The link to the notebook is the following : https://github.com/Ray-BaToDs/IBM-Applied-Data-Science-Capstone/blob/master/Week%202-%20Mission%202%20-IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb



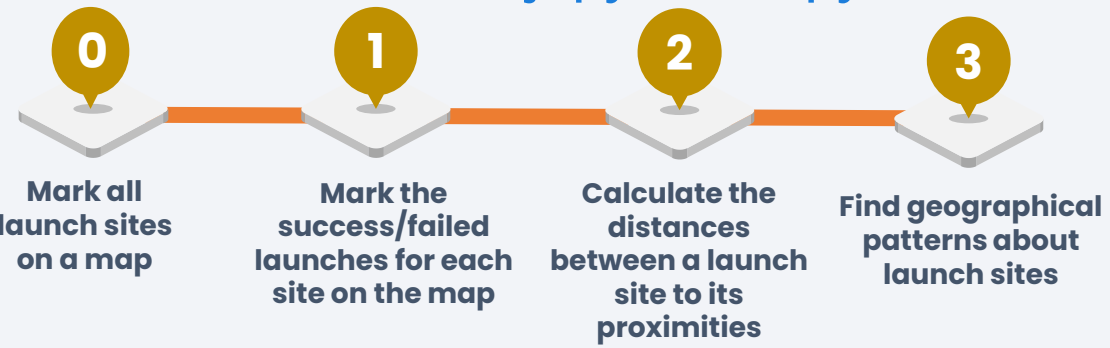
Now if you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

EDA with SQL

- In the context of EDA using SQL we performed multiple SQL queries such as :
  - Names of unique launch sites in the space mission.
  - Total payload mass carried by boosters launched by NASA
  - Average payload mass carried by booster version F9 v1.1
  - Total number of successful and failure mission outcomes
  - Failed landing outcomes in drone ship, their booster version and launch site names.
- We used SQLite to connect to our SpaceX dataset database.
- The link to the notebook is the following : https://github.com/Ray-BaToDs/IBM-Applied-Data-Science-Capstone/blob/master/Week%202-%20Mission%201%20-jupyter-labs-eda-sql-coursera_sqllite.ipynb

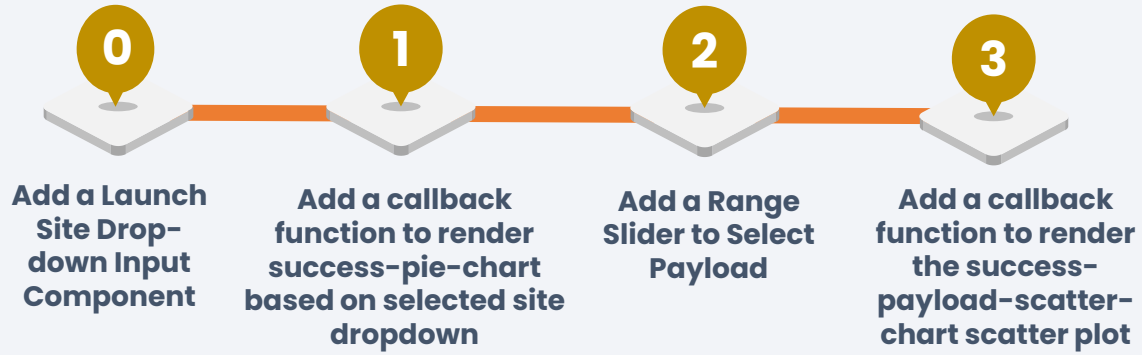
Build an Interactive Map with Folium

- On our folium map we added objects such as markers, circles, lines to display successful and unsuccessful launches for the registered sites. Circles correspond to sites, markers allowed us to cluster the different outcomes on a unique site, lines showed the distance between the launch site and its proximities.
- Observations came out of this map studies. We understood that a launch site could be close to a highway or railroad, however there is an important distance to the closest city, to protect civilians.
- The link to the notebook is the following : https://github.com/Ray-BaToDs/IBM-Applied-Data-Science-Capstone/blob/master/Week%203-%20Mission%201%20-lab_jupyter_launch_site_location.jupyterlite.ipynb



Build a Dashboard with Plotly Dash

- The dashboard we built includes interactions such as a range slider and a Drop_down Menu.
- In our dashboard we also display plots such as pie charts and scatter graphs to show the total launches per site and the relationship between Outcome and Payload Mass.
- The link to the notebook is the following : https://github.com/Ray-BaToDs/IBM-Applied-Data-Science-Capstone/blob/master/Week%203-%20Mission%202%20-spacex_dash_app.py

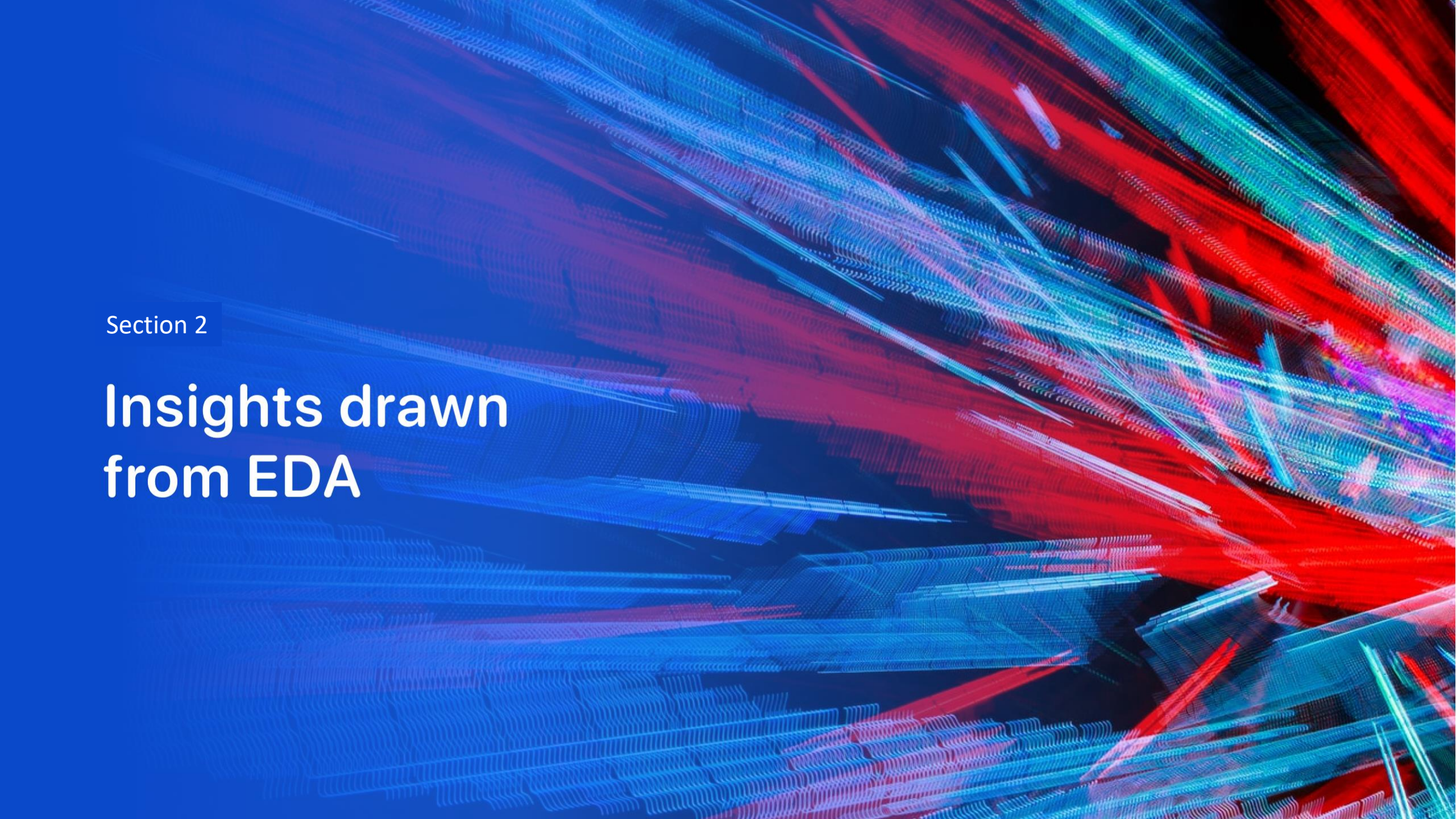


Predictive Analysis (Classification)

- In this step, we created a machine learning pipeline to predict if the first stage will land given the data from the preceding labs.
- We begin by performing exploratory Data Analysis. We Standardize the data and Split into training data and test data. Once our data is ready to be put to work, then we find the best Hyperparameter for SVM, Classification Trees and Logistic Regression.
- Models are trained and hyperparameters are selected using the function GridSearchCV.
- After having tested the different models, we compared the scores of each model to find the best one.
- The link to the notebook is the following : https://github.com/Ray-BaToDs/IBM-Applied-Data-Science-Capstone/blob/master/Week%204-%20Mission%201%20-SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

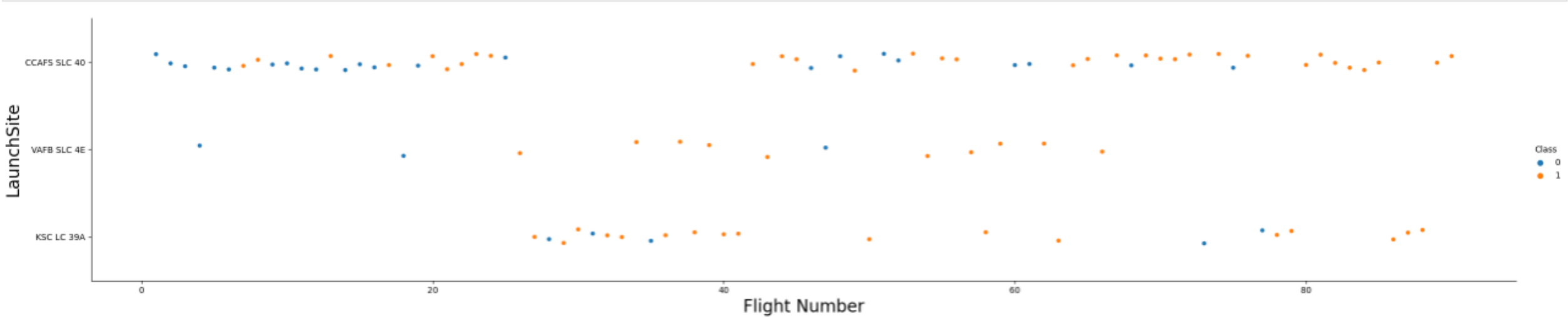
Results

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


Section 2
Insights drawn from EDA

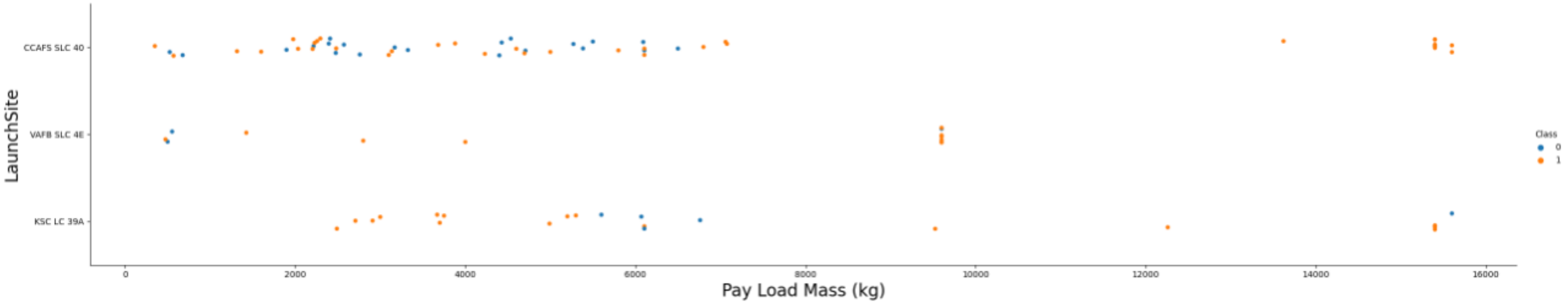
Flight Number vs. Launch Site

- Here we can see that the more a site experiences launches, the greater the chance of success of the mission



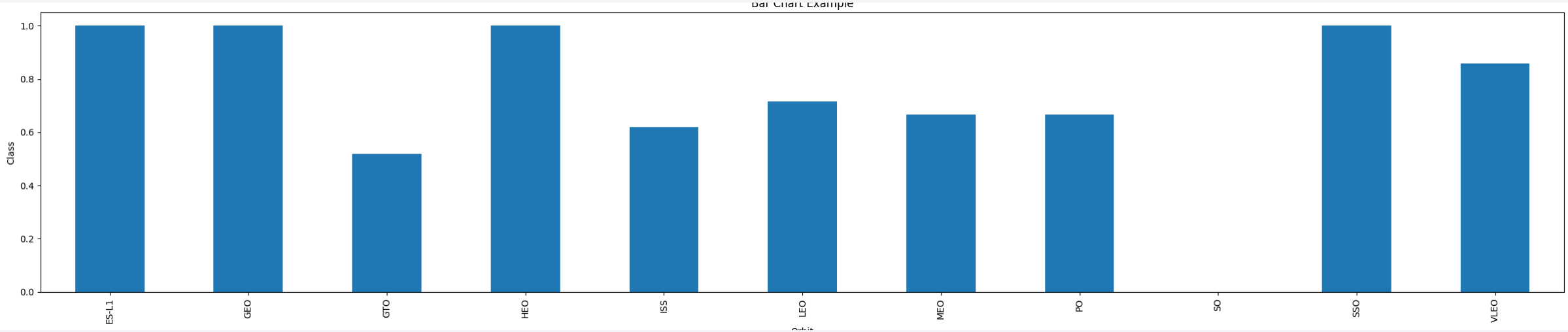
Payload vs. Launch Site

- Here we can see that the greater the payload mass for launch site CCAFS SLC 40, the greater the chance of success of the mission.



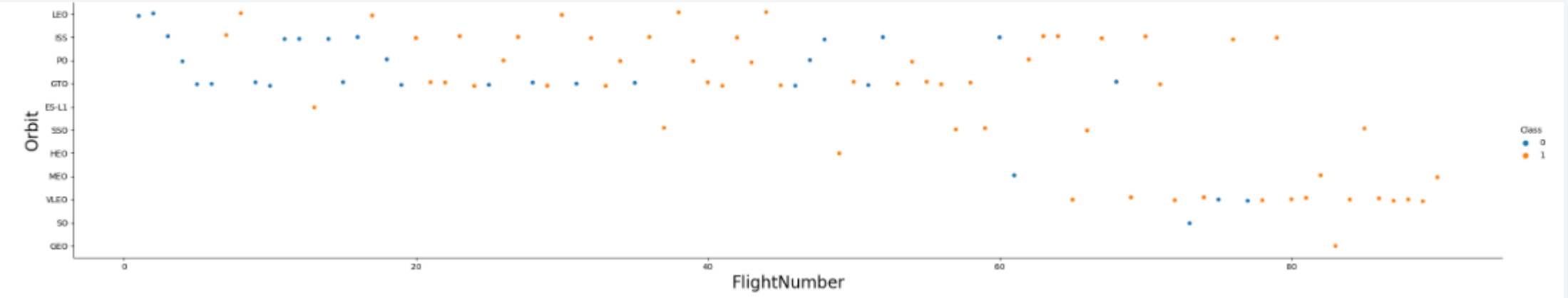
Success Rate vs. Orbit Type

- Here we can see that ES-L1, GEO, HEO, SSO, VLEO had the most success rate.



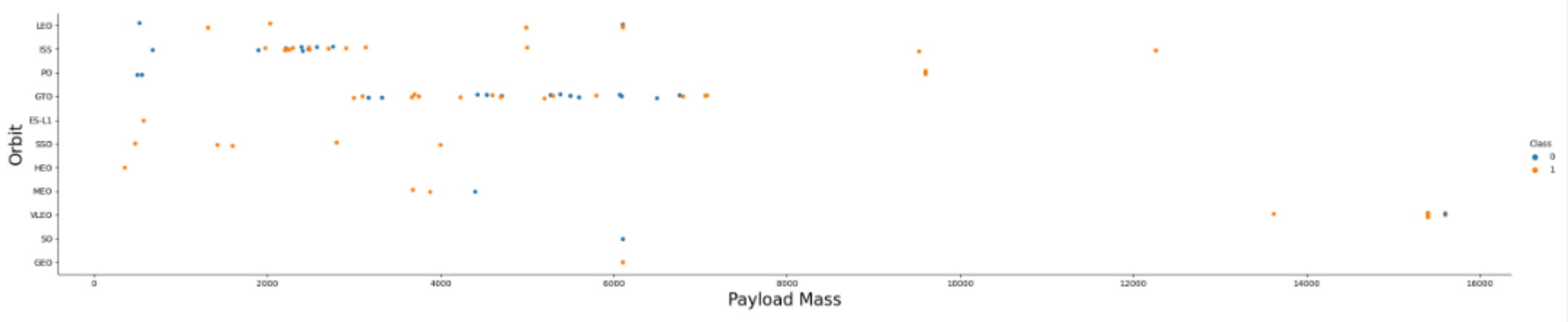
Flight Number vs. Orbit Type

- Here we can see that the LEO orbit, success is related to the number of flights whereas in the GTO orbit, there is no relationship between flight number and the orbit.



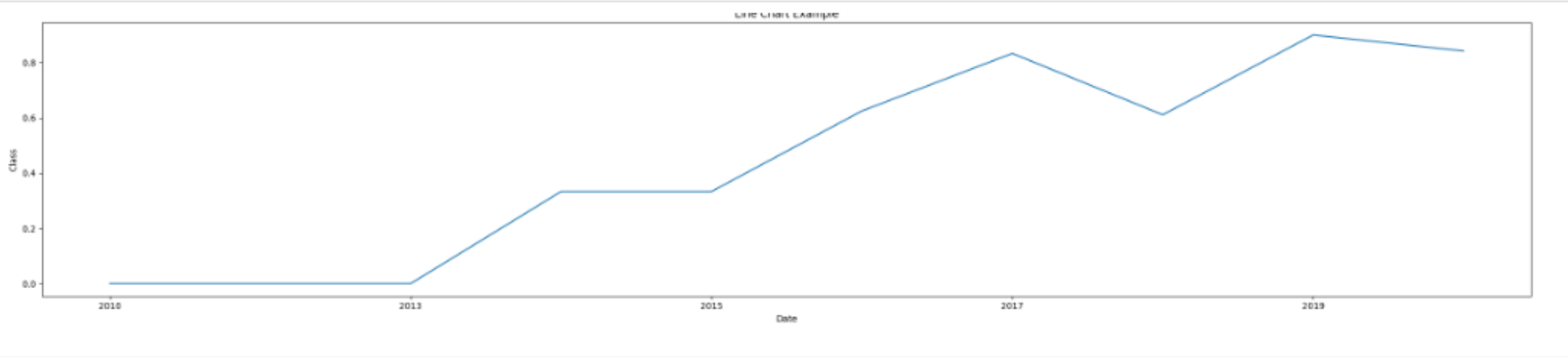
Payload vs. Orbit Type

- Here we can see that with important payloads, successful landings are better distributed towards PO, LEO and ISS orbits.



Launch Success Yearly Trend

- Here we can see that from 2013, the success rate of the missions kept increase year after year.



All Launch Site Names

```
%sql SELECT DISTINCT Launch_site FROM SPACEXTBL;
```

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

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

Launch Site Names Begin with 'CCA'

```
%sql SELECT * FROM SPACEXTBL WHERE Launch_site LIKE 'CCA%' LIMIT 5;
```

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

Done.
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer LIKE 'NASA%';
```

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

SUM(PAYLOAD_MASS__KG_)
99980

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version LIKE 'F9 v1.1%';
```

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

Done.
```

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

```
%sql SELECT min(date),"Launch_Site","Mission_Outcome" FROM SPACEXTBL WHERE "Landing _Outcome" LIKE 'Success%';
```

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

Done.
```

min(date)	Launch_Site	Mission_Outcome
01-05-2017	KSC LC-39A	Success

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT Booster_Version FROM SPACEXTBL 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

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT COUNT(CASE WHEN Mission_Outcome LIKE 'Success%' THEN 1 END) AS num_success,COUNT(CASE WHEN Mission_Outcome LIKE 'Failure%' THEN 1 END) AS num_failure FROM SPACEXTBL;
 * sqlite:///my_data1.db
Done.
```

num_success	num_failure
100	1

Boosters Carried Maximum Payload

```
%sql SELECT "Booster_Version" FROM SPACEXTBL WHERE "PAYLOAD_MASS__KG_" = (SELECT max(PAYLOAD_MASS__KG_) FROM SPACEXTBL) ORDER BY "Booster_Version" asc;
 * sqlite:///my_data1.db
Done.
```

Booster_Version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5

2015 Launch Records

```
%sql SELECT substr(Date, 4, 2) as Month,substr(Date,7,4) as Year,"Booster_Version","Landing _Outcome","Launch_Site" FROM SPACEXTBL WHERE "Landing _Outcome"='Failure (drone ship)' and substr(Date,7,4)='20
```

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

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

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

```
%sql SELECT RANK() OVER (ORDER BY "Landing _Outcome" ASC) as Rank,"Landing _Outcome",COUNT(*) FROM SPACEXTBL WHERE "Landing _Outcome" LIKE 'Success%' and "date" BETWEEN '04-06-2010' AND '20-03-2017' GROU
```

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

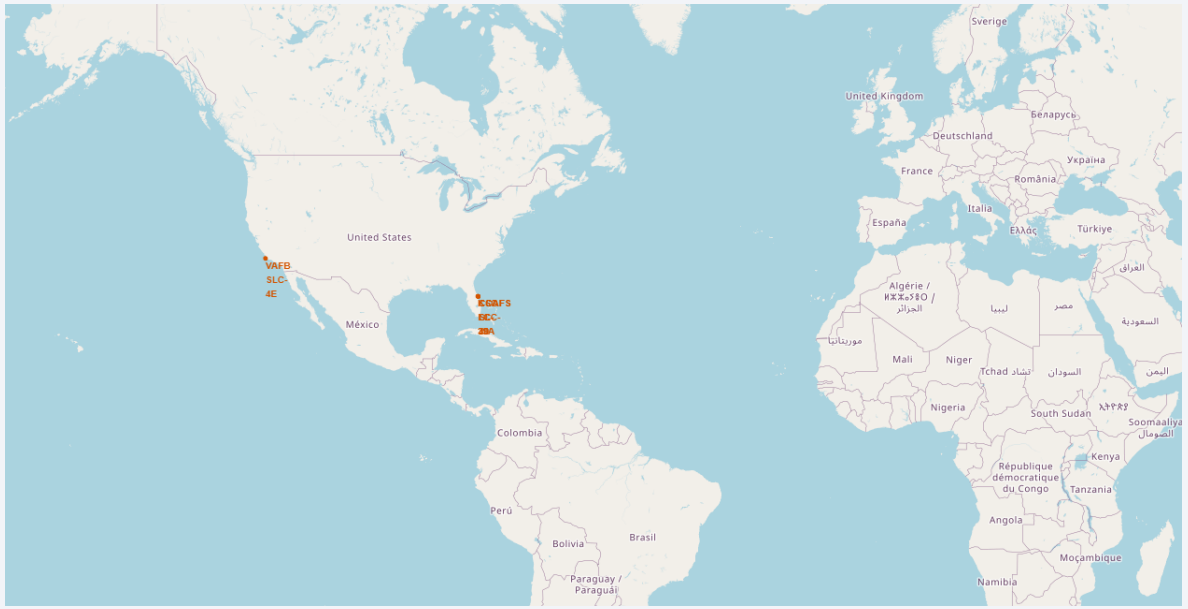
Rank	Landing _Outcome	COUNT(*)
1	Success	20
2	Success (drone ship)	8
3	Success (ground pad)	6

Section 3

Launch Sites Proximities Analysis

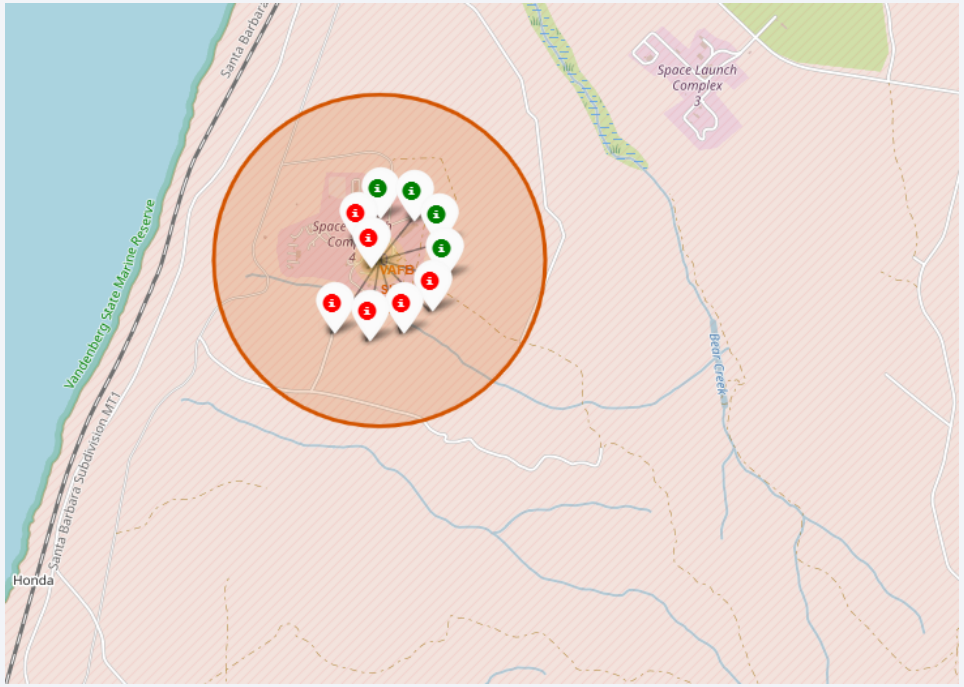
All Launch Sites

- By adding circle markers to the Folium Map, we can see that the Launch Sites are located at the United States coasts.



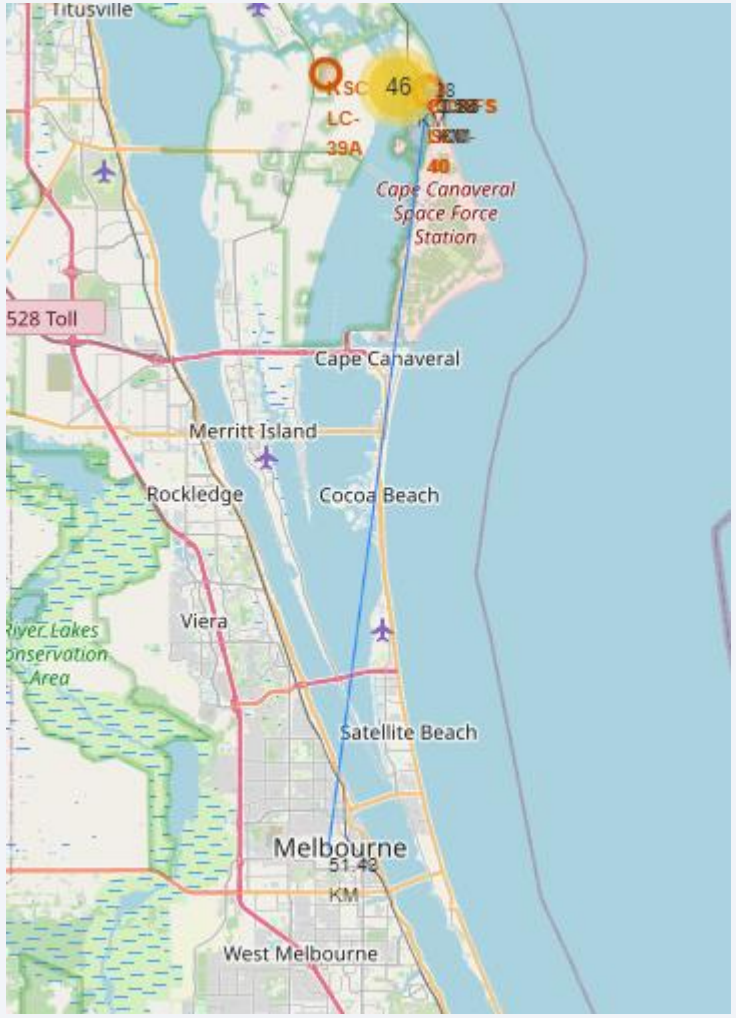
Launch Site Cluster

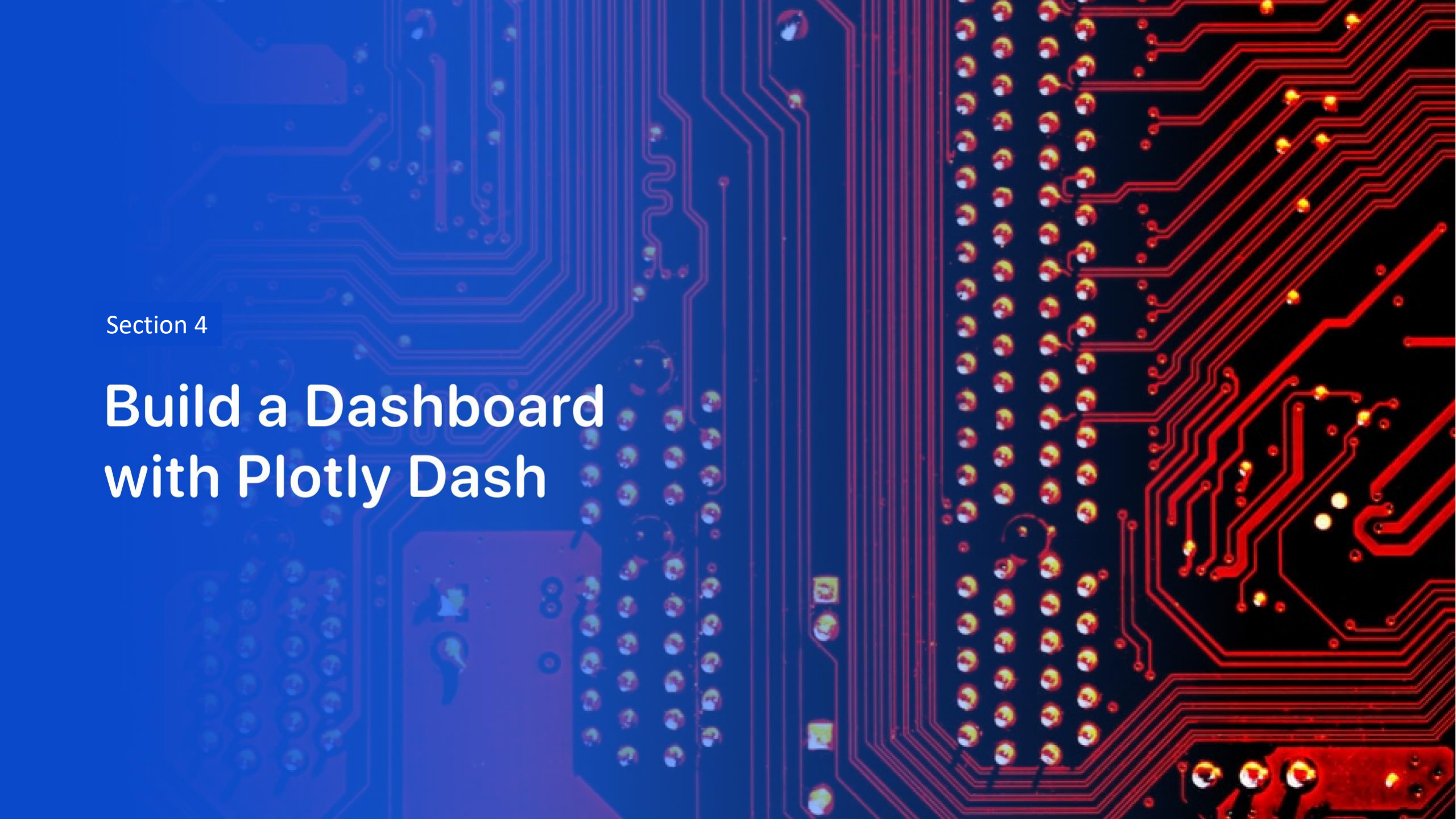
- Red markers indicate that the launch was unsuccessful, and the green markers indicate a successful launch.



Distance measuring from Launch Site

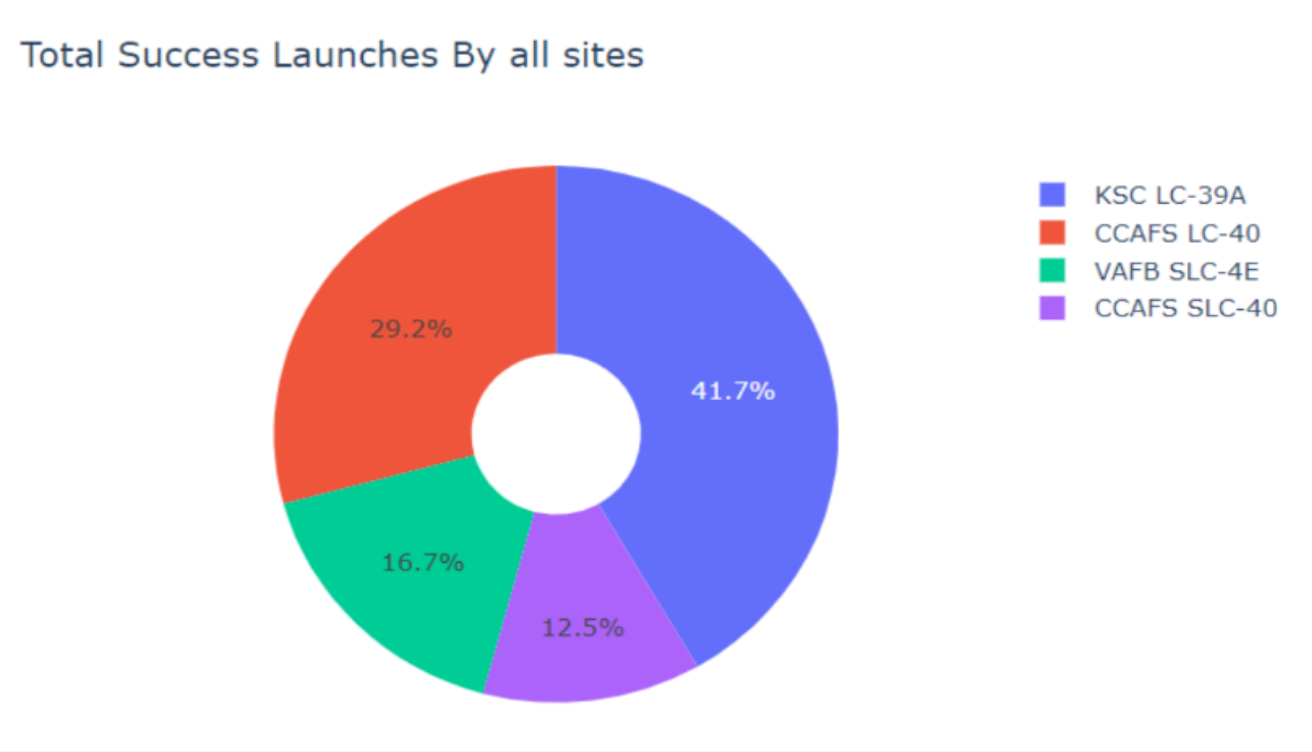
- We add lines to show the distance between the Launch Site and proximities such as the closest City or the closest Highway.




Section 4
Build a Dashboard with Plotly Dash

Pie chart of the successful launches by all sites

- It is clear that the ‘KSC LC-39A’ site has the most successful launches from all the sites



Section 5

Predictive Analysis (Classification)

Classification Accuracy

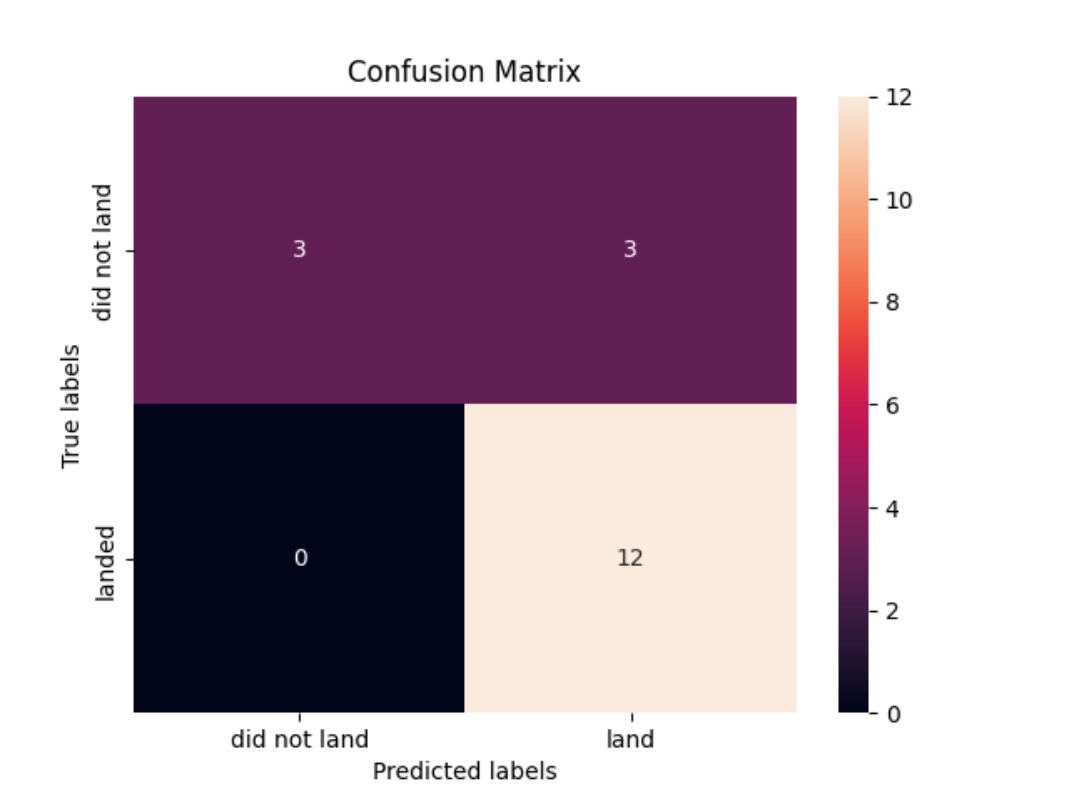
- We compare every model's accuracy, and we can conclude that the tree classifier is the most accurate model for our subject.

```
print("tuned hpyerparameters :(best parameters) ",knn_cv.best_params_)
print("KNN accuracy :",knn_cv.best_score_)
print("Logistic Regression accuracy :",logreg_cv.best_score_)
print("SVM accuracy :",svm_cv.best_score_)
print("Tree Classifier accuracy :",tree_cv.best_score_)
```

```
tuned hpyerparameters :(best parameters)  {'algorithm': 'auto', 'n_neighbors': 10, 'p': 1}
KNN accuracy : 0.8482142857142858
Logistic Regression accuracy : 0.8464285714285713
SVM accuracy : 0.8482142857142856
Tree Classifier accuracy : 0.8767857142857143
```

Confusion Matrix

- This is the confusion matrix of the decision tree classifier. We can see that we have 3 false positives, meaning the model classes landings as successful when the landing was unsuccessful.



Conclusions

Through exploratory data analysis results, interactive analytics and predictive analysis results, we were able to extract many observations and conclusions:

- The more a launch site had experience with flights, the more they learn from their mistakes and succeed in their latest flights.
- Launch success rates have been increasing year after year, this may be due to better technology and greater experience.
- Some orbits are safer than others. We note that ES-L1, GEO, HEO, SSO, VLEO had the best success rates.
- After having tested various ML models we can conclude that the Decision tree classifier is the best machine learning algorithm for this task. We can reuse this model to predict whether the future missions will succeed or not, that was the target of our project.

Thank you!