IBM Developer
SKILLS NETWORK

WINNING SPACE RACE WITH DATA SCIENCE

SPACEX PROJECT

OGUZHAN ICER 24/06/2023
HTTPS://GITHUB.COM/OZIMAN-BOOP/IBM_SPACEX_CAPSTONE_PROJECT

OUTLINE

EXECUTIVE SUMMARY

- The SpaceX Data Science Project involved the collection of data from both the SpaceX REST API and SpaceX Wikipedia page. The data was utilized to analyze and advance space technology and exploration. The project focused on identifying successful landings by creating a "Class" column as labels.
- Exploratory Data Analysis (EDA) was performed using SQL queries, visualization techniques, folium maps, and interactive dashboards. This allowed for a comprehensive examination of the data, uncovering relationships, patterns, and trends. Relevant columns were selected as features for further analysis.
- To prepare the data for machine learning algorithms, One Hot Encoding was applied to convert categorical variables into binary format. The data was then standardized to ensure consistency across different features. GridSearchCV was utilized to determine the optimal parameters for the machine learning models.
- Four different models were implemented: Logistic Regression, Support Vector Machine (SVM), Decision Tree Classifier, and K Nearest Neighbors. Interestingly, all four models produced similar results, with an accuracy rate of approximately 83.33%. However, it was observed that all models tended to over-predict successful landings.
- These findings highlight the need for additional data to improve the model's accuracy and predictive capabilities. Obtaining more comprehensive and diverse data could lead to better model determination and accuracy. Overall, the project demonstrated the application of data science techniques in analyzing SpaceX data and their significance in advancing space technology and exploration.

INTRODUCTION

In this capstone project, our objective is to predict the successful recovery of the Falcon 9 rocket's first stage. This prediction holds significant importance as it directly impacts the cost of a rocket launch. SpaceX, a leading aerospace company, advertises Falcon 9 rocket launches on its website, offering a price of 62 million dollars, whereas other providers charge significantly higher, around 165 million dollars per launch. The key factor contributing to this cost difference is SpaceX's ability to successfully reuse the first stage of the Falcon 9 rocket.

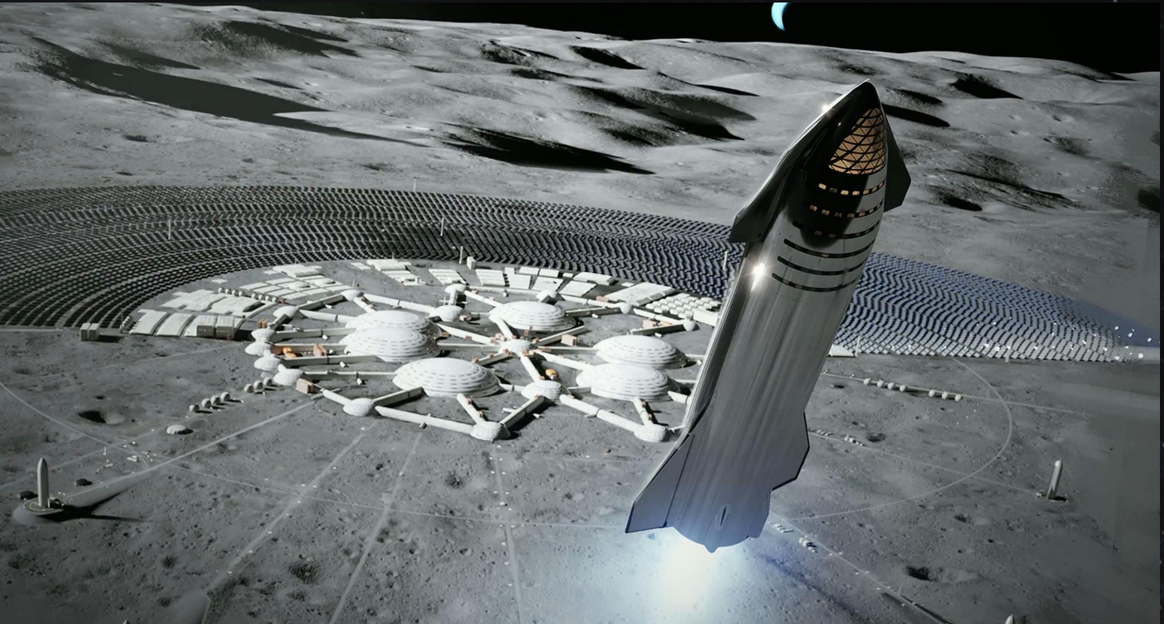
By accurately determining whether the first stage will land successfully, we can effectively estimate the cost of a launch. This information becomes crucial when an alternate company, such as Space Y, wants to compete with SpaceX for a rocket launch contract. Having insights into the likelihood of a successful stage 1 recovery enables Space Y to make informed decisions and potentially offer competitive bids against SpaceX.

To accomplish this task, we will leverage machine learning techniques and develop a robust prediction model. By analyzing historical data related to Falcon 9 rocket launches and the outcomes of their first stage recoveries, we aim to build a reliable model that can predict the success or failure of stage 1 recovery accurately.

Through this capstone project, we will equip Space Y with the necessary tools and insights to make strategic and cost-effective decisions in the highly competitive space industry. By leveraging machine learning and predictive modeling, we will contribute to enhancing Space Y's competitiveness and drive advancements in space exploration and technology.

Section 1

METHODOLOGY



Executive Summary

- Data collection methodology:
  - Gathered data from SpaceX public API and by scrapping SpaceX Wikipedia page
- Perform data wrangling
  - Classifying true landings as successful and unsuccessful otherwise
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - We tuned the models using GridSearchCV

Data Collection

Collection

Data sets were collected from Space X API (https://api.spacexdata.com/v4/rockets/)

and from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches), using web scraping technics.

Data Collection – SpaceX API

- SpaceX offers a public API from where data can be obtained and then used;
- This API was used according to the flowchart beside and then data is persisted.

Request API and parse the SpaceX launch data → Filter data to only include Falcon 9 launches → Deal with Missing Values

https://github.com/oziman-boop/IBM_SPACEX_CAPSTONE_PROJECT/blob/main/jupyter_labs_spacex_data_collection_api.ipynb

Data Collection -Scraping

- Data from SpaceX launches can also be obtained from Wikipedia;
- •Data are downloaded from Wikipedia according to the flowchart and then persisted.

Request the Falcon9 Launch Wiki Page → Extract al column/variable names from the HTML table header → Create a data frame by parsing the launch HTML tables

https://github.com/oziman-boop/IBM_SPACEX_CAPSTONE_PROJECT/blob/main/jupyter_labs_webscraping.ipynb

Data Wrangling

- Initially some Exploratory Data Analysis (EDA) was performed on the dataset.
- •Then the summaries launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type were calculated.
- •Finally, the landing outcome label was created from Outcome column.

https://github.com/oziman-boop/IBM_SPACEX_CAPSTONE_PROJECT/blob/main/IBM_DS0321EN_SkillsNetwork_labs_module_1_L3_labs_jupyter_spacex_data_wrangling_jupyterlite_jupyterlite.ipynb

EDA with Data Visualization

Exploratory Data Analysis performed on variables Flight Number, Payload Mass, Launch Site, Orbit, Class and Year.

Plots Used:

Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit vs. Success Rate, Flight Number vs. Orbit, Payload vs Orbit, and Success Yearly Trend.

Scatter plots, line charts, and bar plots were used to compare relationships between variables to

decide if a relationship exists so that they could be used in training the machine learning model.

GitHub URL: https://github.com/oziman-boop/IBM_SPACEX_CAPSTONE_PROJECT/blob/main/IBM_DS0321EN_SkillsNetwork_labs_module_2_jupyter_labs_eda_dataviz_ipynb_jupyterlite.ipynb

EDA with SQL

Loaded data set into IBM DB2 Database.

Queried using SQL Python integration.

Queries were made to get a better understanding of the dataset.

Queried information about launch site names, mission outcomes, various pay load sizes of customers and booster versions, and landing outcomes

GitHub URL: https://github.com/oziman-boop/IBM_SPACEX_CAPSTONE_PROJECT/blob/main/jupyter_labs_eda_sql_coursera_sqllite.ipynb

Build an Interactive Map with Folium

Folium maps mark Launch Sites, successful and unsuccessful landings, and a proximity example to key locations: Railway, Highway, Coast, and City.

This allows us to understand why launch sites may be located where they are. Also visualizes successful landings relative to location.

GitHub URL:

https://github.com/oziman-boop/IBM_SPACEX_CAPSTONE_PROJECT/blob/main/SpaceX_Generating_Maps_with_Folium-2.ipynb

Build a Dashboard with PlotlyDash

The following graphs and plots were used to visualize data

- Percentage of launches by site
- Payload range
- This combination allowed to quickly analyze

the relation between payloads and launch sites, helping to identify where is best place to launch according to payloads.

https://github.com/oziman-boop/IBM_SPACEX_CAPSTONE_PROJECT/blob/main/spacex_dash_app%20(1).py

Predictive Analysis (Classification)

Four classification models were compared: logistic regression, support vector machine, decision tree and k nearest neighbors.

Data preparation and standardization → Test of each model with combinations of hyperparameters → Comparison of results

https://github.com/oziman-boop/IBM_SPACEX_CAPSTONE_PROJECT/blob/main/IBM_DS0321EN_SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5_jupyterlite_2.ipynb

Results

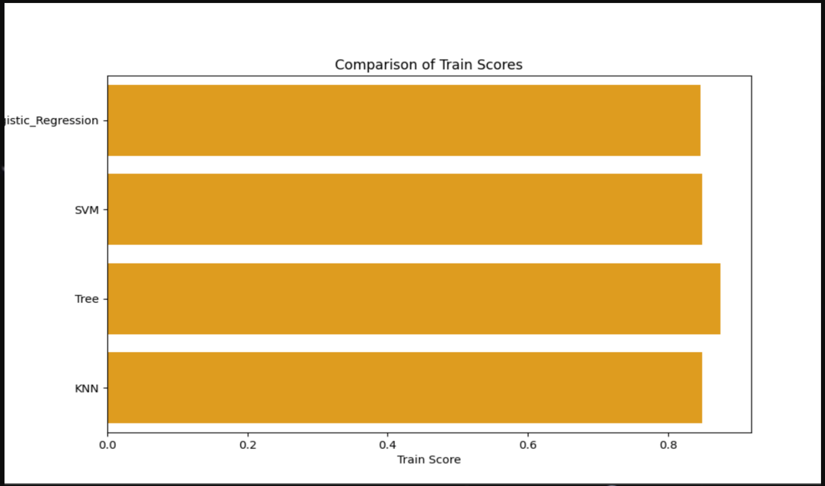
- Exploratory data analysis results:
- Space X uses 4 different launch sites;
- The first launches were done to Space X itself and NASA;
- The average payload of F9 v1.1 booster is 2,928 kg;
- The first success landing outcome happened in 2015 fiver year after the first launch;
- Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average;
- Almost 100% of mission outcomes were successful;
- Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015;
- The number of landing outcomes became as better as years passed.

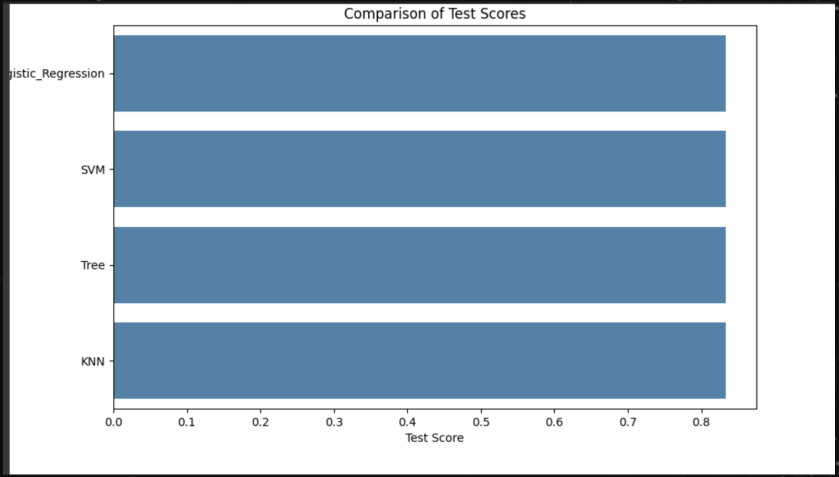
Results

- Folium maps mark Launch Sites, successful and unsuccessful landings, and show distances to key locations: Railway, Highway, Coast, and City.
- This allows us to understand why launch sites may be located where they are. Also visualizes successful landings relative to location.

Results

Predictive Analysis showed that Decision Tree Classifier is the best model to predict successful landings, having accuracy close to 87% however accuracy for test data is same over four models by 83.33%.





Section 2

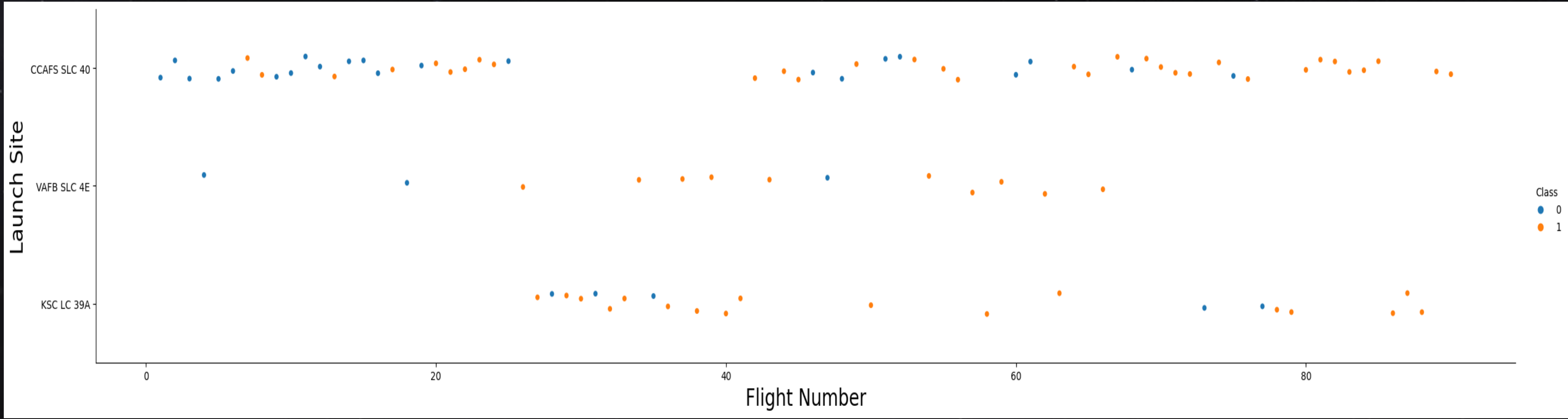
INSIGHT DRAWN FROM EDA

Flight Number vs. Launch Site

Blue indicates successful launch and orange indicates unsuccessful launch.

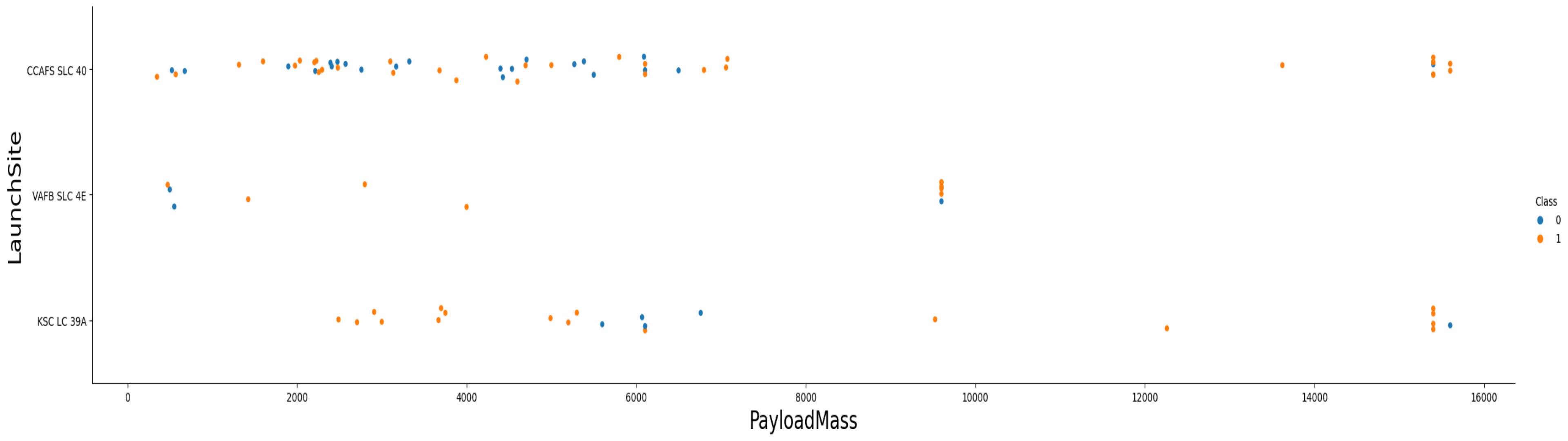
Graphic suggests an increase in success rate over time (indicated in Flight Number). Likely a big breakthrough around flight 20 which significantly increased success rate.

CCAFS appears to be the main launch site as it has the most volume.



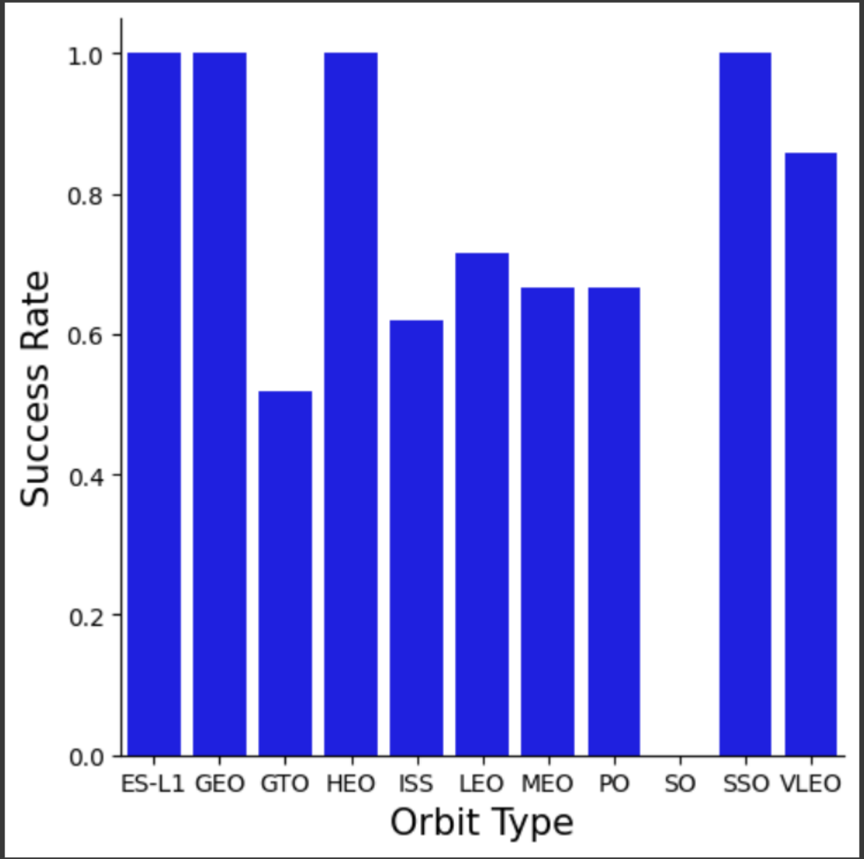
Payload vs. Launch Site

Blue indicates successful launch and orange indicates unsuccessful launch. Payload mass appears to fall mostly between 0-6000 kg. Different launch sites also seem to use different payload mass.



Success Rate vs. Orbit Type

- ES-L1, GEO, HEO have 100% success rate, each of them has one sample.
- SSO has 100% success rate with 5 samples.
- VLEO has 85% success rate.
- SO has 0% success rate.
- GTO has the around 50% success rate but it has 27 samples.



Flight Number vs. Orbit Type

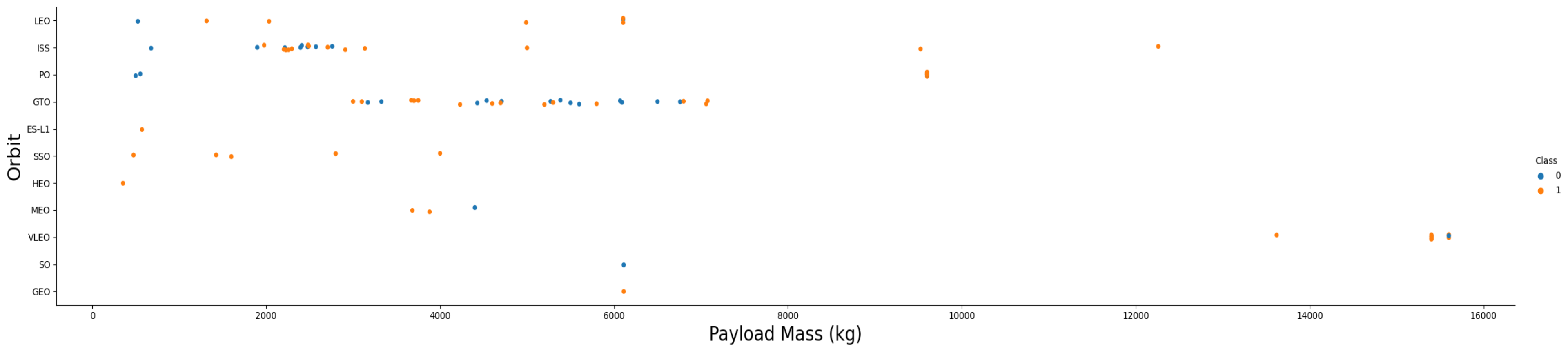
Blue indicates successful launch and orange indicates unsuccessful launch.

- Launch Orbit preferences changed over Flight Number.
- Launch Outcome seems to correlate with this preference.
- SpaceX started with LEO orbits which saw moderate success LEO and returned to VLEO in recent launches
- SpaceX appears to perform better in lower orbits or Sun-synchronous orbits

Payload vs. Orbit Type

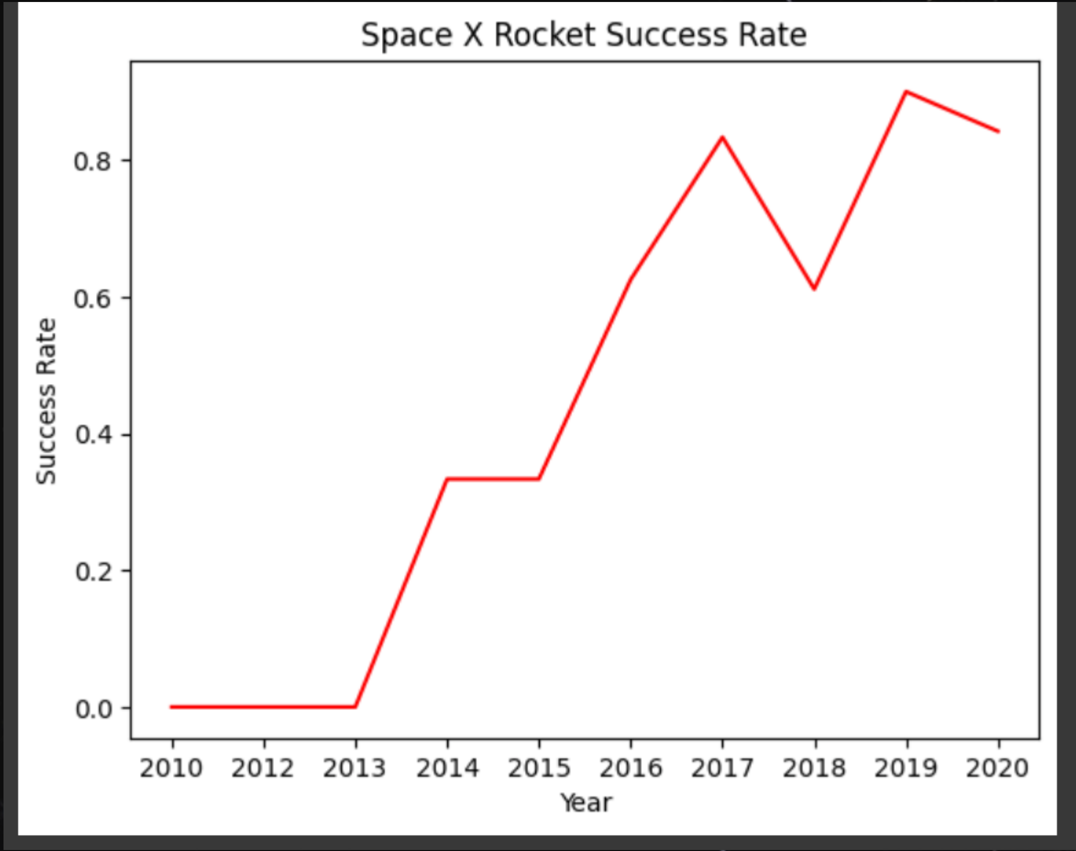
Blue indicates successful launch and orange indicates unsuccessful launch.

- Payload mass seems to correlate with orbit
- LEO and SSO seem to have relatively low payload mass
- The other most successful orbit VLEO only has payload mass values in the higher end of the range



Launch Success Yearly Trend

- Success generally increases over time since 2013 with a slight dip in 2018
- Success in recent years at around 82%



All Launch Site Names

As We can see, there are 4 different launch sites in the dataset and also there is a none launch site value.

```
%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
None

Launch Site Names Begin with 'CCA'

```
[34] %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
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

In that DF, we can see the firs five Launch Sites which starts with 'CCA'.

Total Payload Mass

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD_MASS FROM SPACEXTBL WHERE CUSTOMER = "NASA (CRS)"
```

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

TOTAL_PAYLOAD_MASS
45596.0

We found the total Payload Mass with using SUM command in sql. It's shown us the total Payload Mass is 45596.0

Average Payload Mass by F9 v1.1

```
[38] %sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVERAGE_PAYLOAD_MASS FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1'

 * sqlite:///my_data1.db
Done.
AVERAGE_PAYLOAD_MASS
2928.4
```

We found the average Payload Mass with using AVG command in sql. It's shown us average Payload Mass is 2928.4

First Successful Ground Landing Date

```
%sql SELECT DATE AS FIRST_SUCCESS_GROUND_PAD FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (ground pad)'
```

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

FIRST_SUCCESS_GROUND_PAD
22/12/2015
18/07/2016
19/02/2017
05/01/2017
06/03/2017
14/08/2017
09/07/2017
15/12/2017
01/08/2018

As we can see in the list, the first date of succesful ground landing date is '22/12/2015'

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

We found the correct booster version which its payload between 4000 and 6000 that land on a drone ship succesfully.

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT mission_outcome, COUNT(*) as total_number from SPACEXTBL GROUP BY mission_outcome
```

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

Mission_Outcome	total_number
None	898
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

As a result, there is 100 succesful and 1 failure mission outcomes.

Boosters Carried Maximum Payload

```
%sql SELECT booster_version FROM SPACEXTBL WHERE payload_mass__kg_ = (select max(payload_mass__kg_) FROM SPACEXTBL)
```

```
 * 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

There are 12 different booster version which have carried maximum payload mass.

2015 Launch Records

```
%sql SELECT CASE substr(Date, 4, 2) WHEN '01' THEN 'January' WHEN '02' THEN 'February' WHEN '03' THEN 'March' WHEN '04' THEN 'April' WHEN '05' THEN 'May' WHEN '06' THEN 'Ju
```

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

Month	Landing_Outcome	Booster_Version	Launch_Site
October	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
June	Precluded (drone ship)	F9 v1.1 B1018	CCAFS LC-40

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

```
%sql SELECT landing_outcome, COUNT(landing_outcome) AS count_of_outcome FROM (select landing_outcome FROM SPACEXTBL WHERE date > '04/06/2010' and date <'20/03/2017')GROUP BY l
```

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

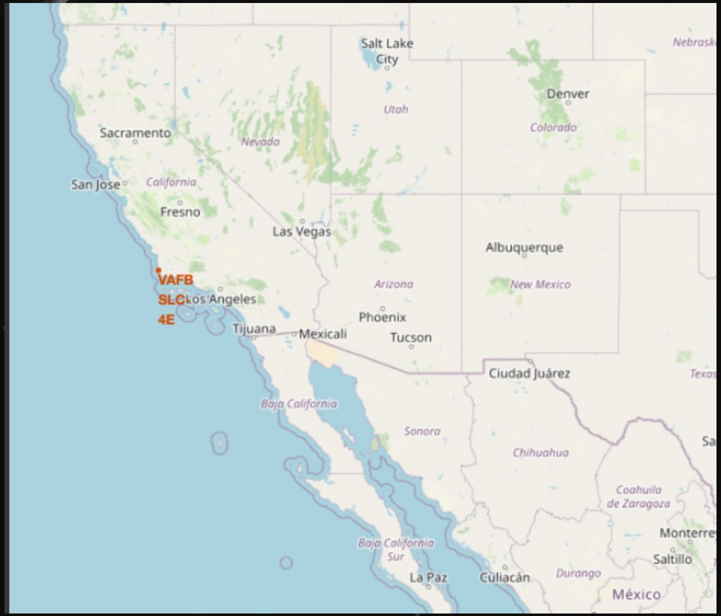
landing_outcome	count_of_outcome
Success	20
No attempt	9
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1

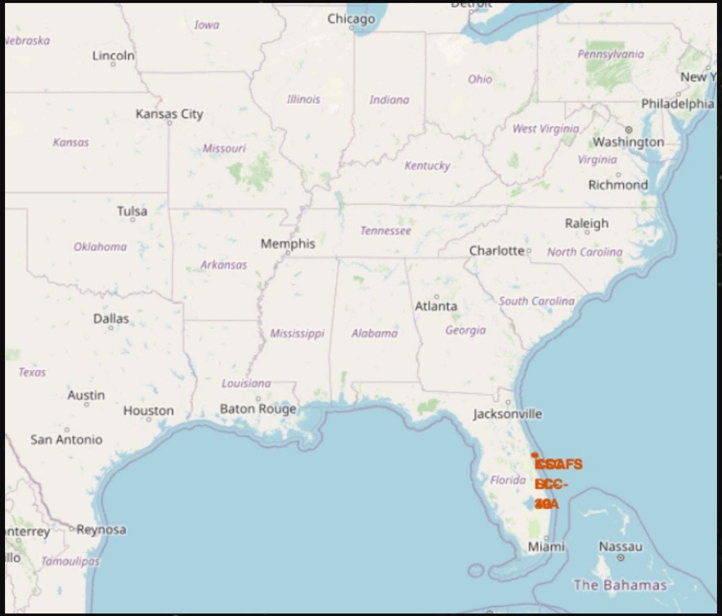
Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

Section 3

Launch sites proximities Analysis

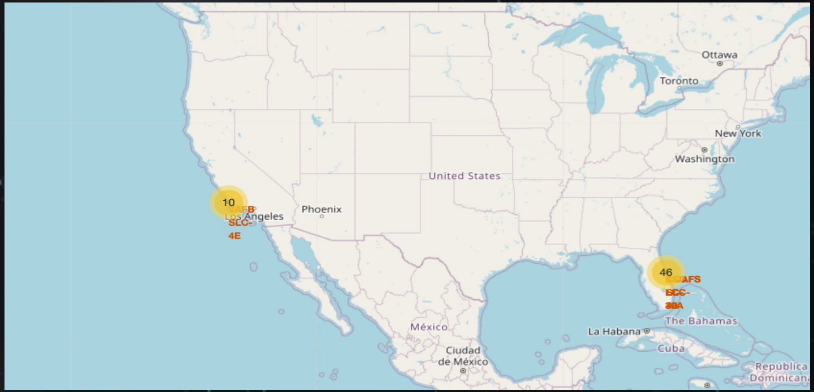
Launch Site Locations

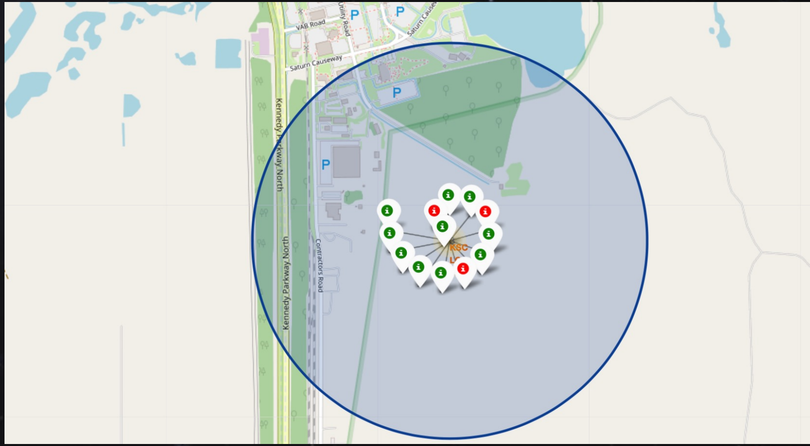




In the map, we can see all launch sites are close to ocean and they are near to Ecuador line.

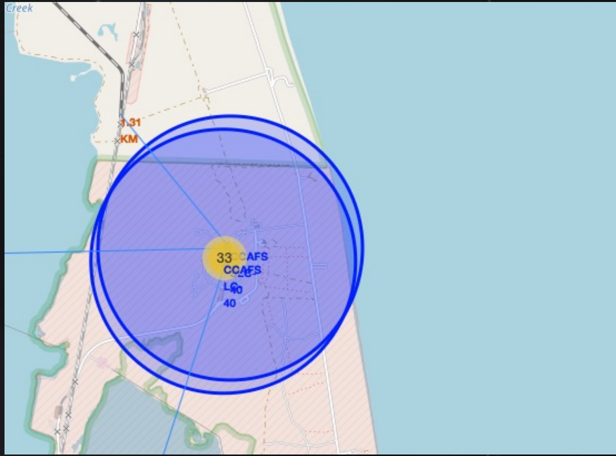
Color Coded Launch Markers

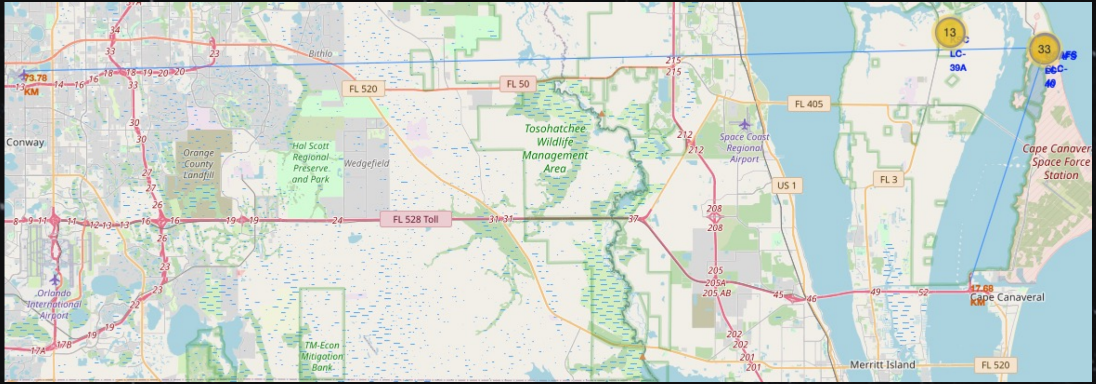




Clusters on Folium map can be clicked on to display each successful landing (green icon) and failed landing (red icon). In this example KSCLC 39-A shows 10 successful landings and 3 failed landings.

KEY LOCATION PROXIMITIES





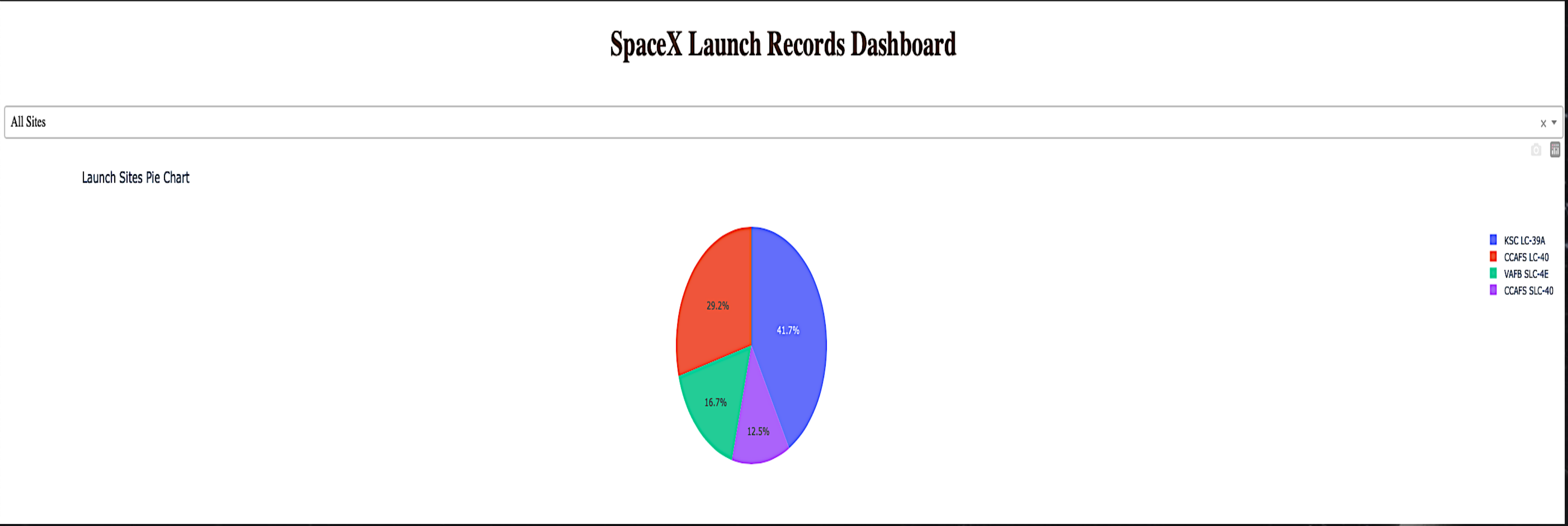
We can see from photo that the distance between launch site and railway is 1.31 km.

We can see distances of launch site to key locations such as railway, highway and city

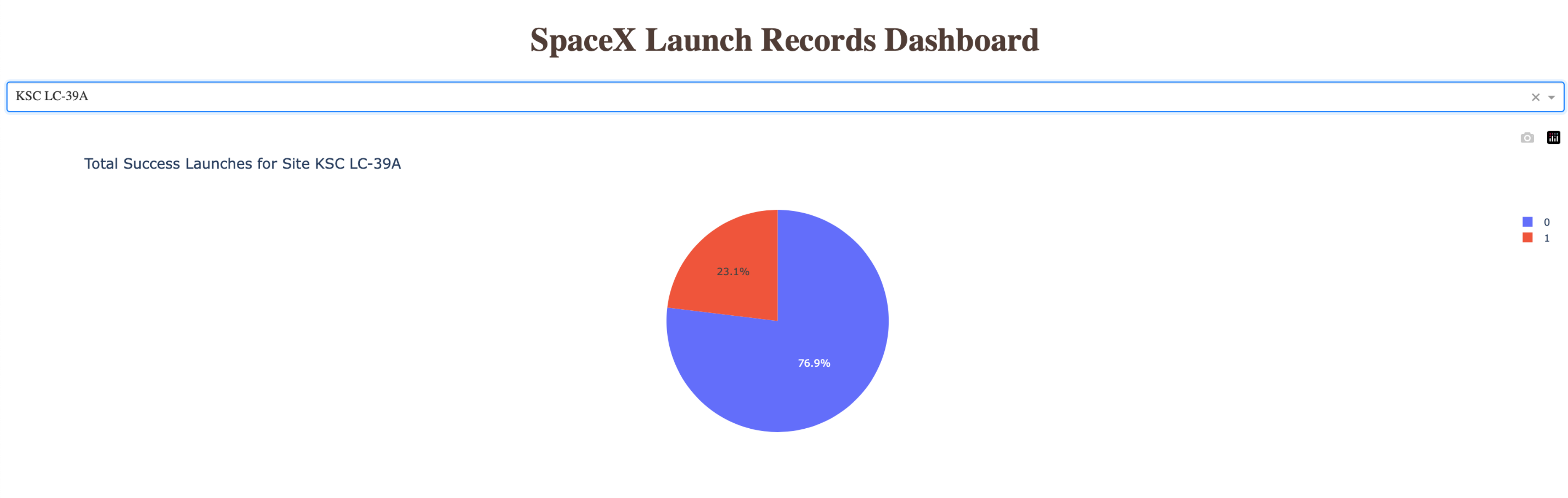
SECTION 4

BUILD A DASHBOARD WITH PLOTLY DASH

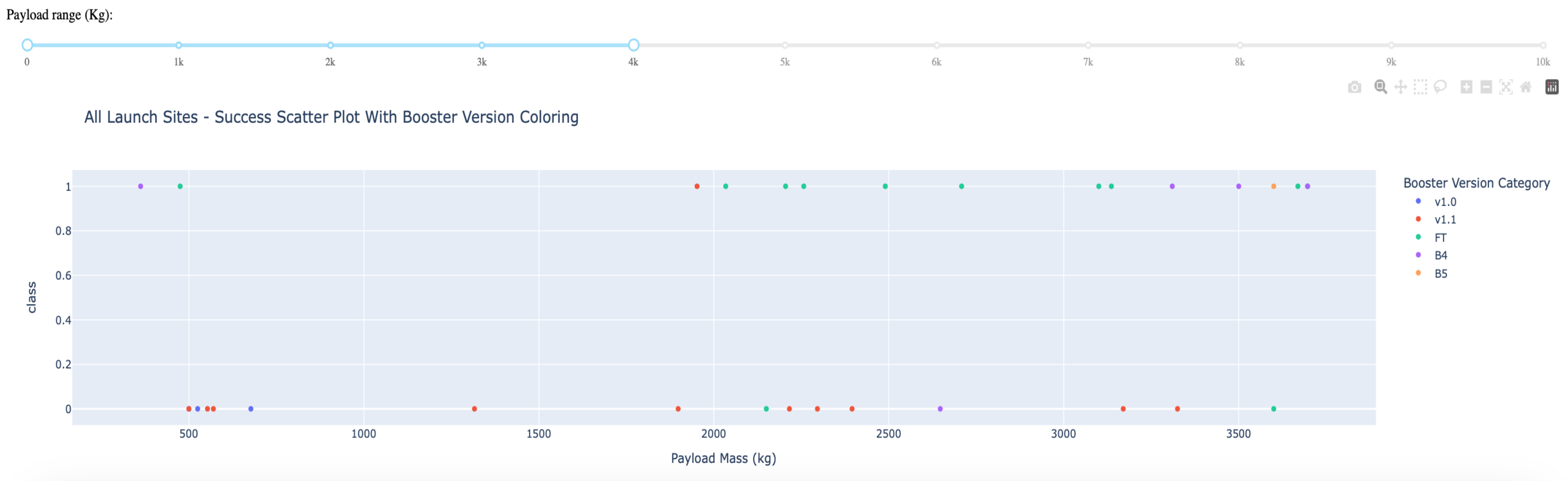
SUCCESSFUL LAUNCHES BY ALL SITES



LAUNCH SITE WITH HIGHEST SUCCESS RATE

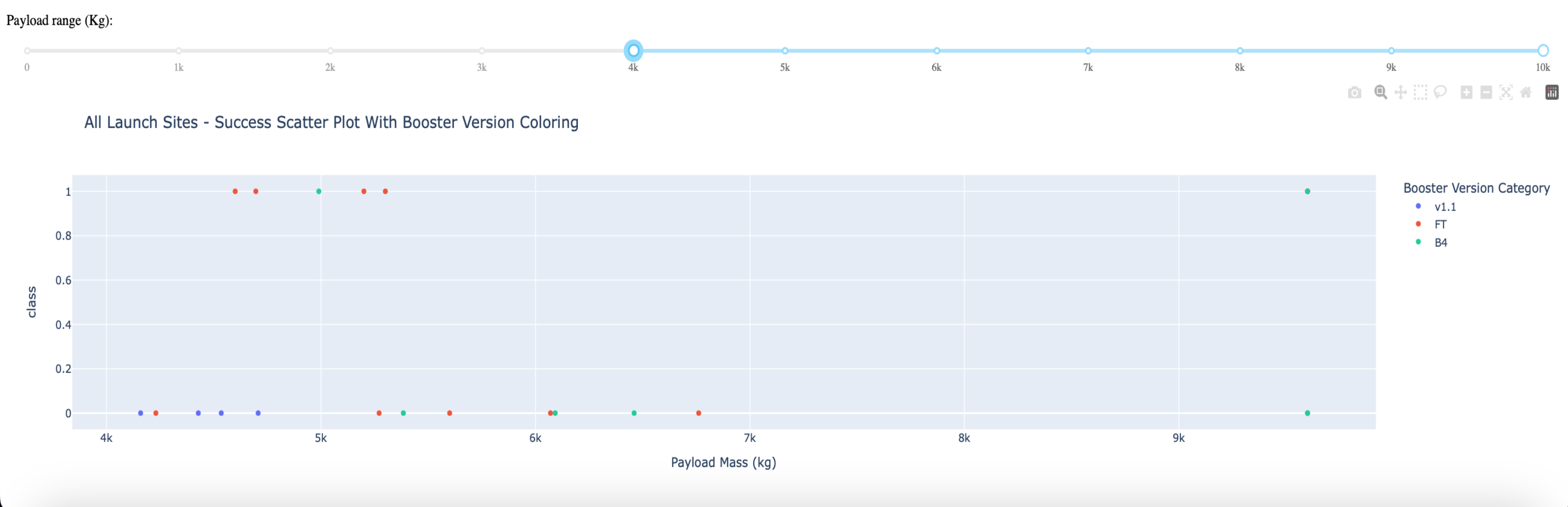


PAYLOAD AND SUCCESS RATE



AS WE CAN SEE UNDER 4000 KG PAYLOAD AND FT BOOSTER IS THE MOST SUCCESSFUL COMBINATION

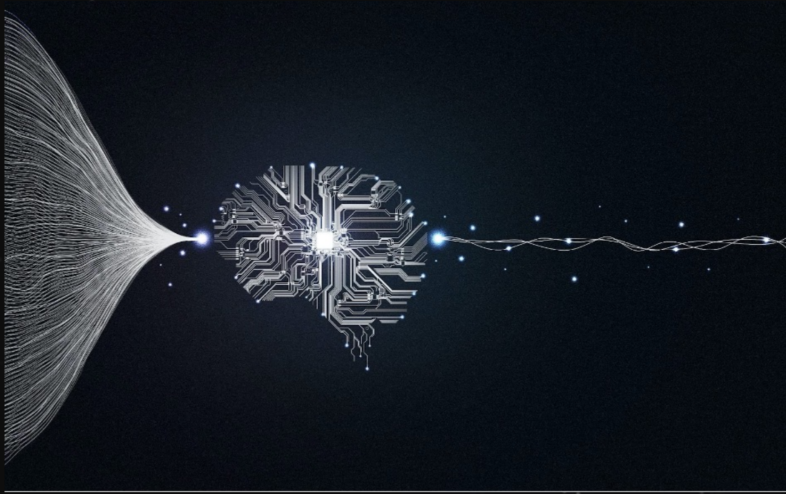
SUCCESS RATE IN HIGH PAYLOADS



THERE IS NOT SO MUCH ATTEMPS ABOVE 7000 PAYLOADS FOR ALL BOOSTER VERSIONS

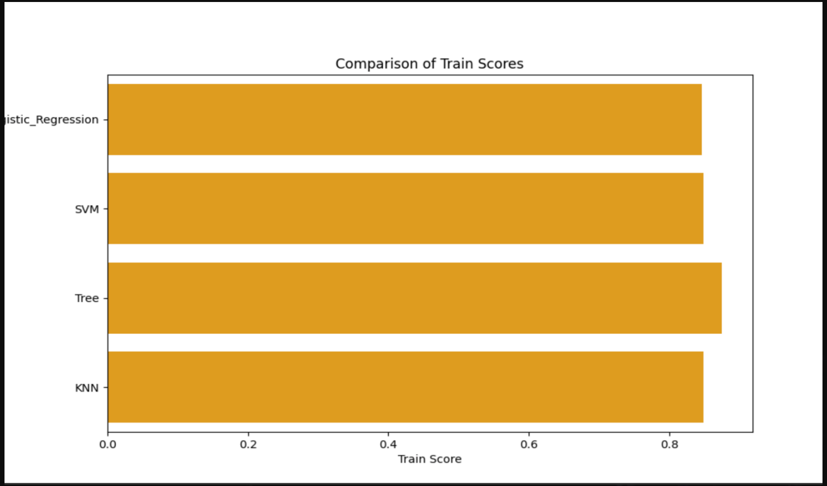
SECTION 5

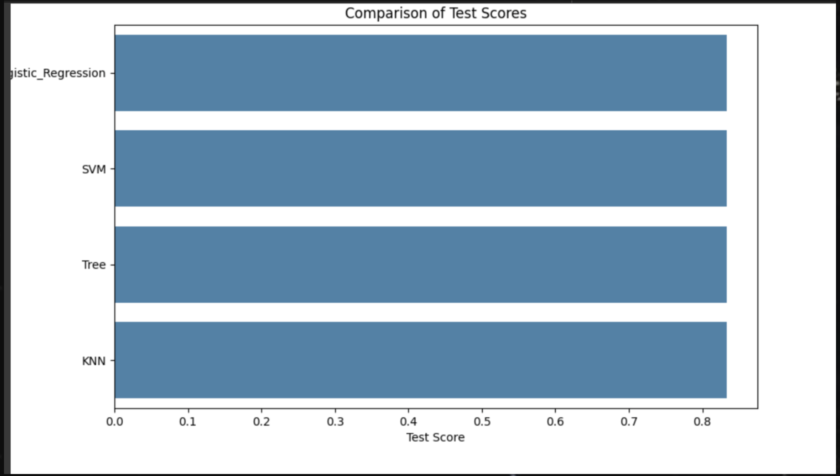
PREDICTIVE ANALYSIS(CLASSIFICATION)



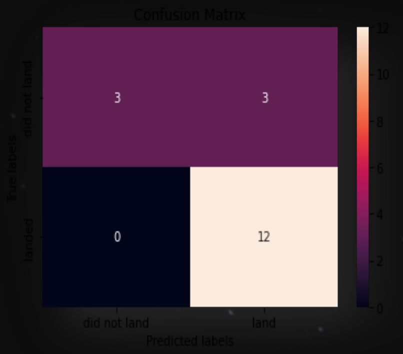
Classification accuracy

- Four classification models were tested, and their accuracies are plotted beside;
- The model with the highest classification accuracy is Decision Tree Classifier, which has accuracies over than 87%.



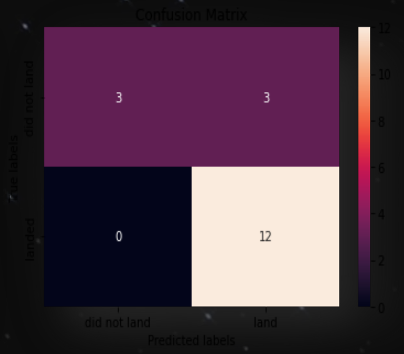


Confusion matrixs of all models



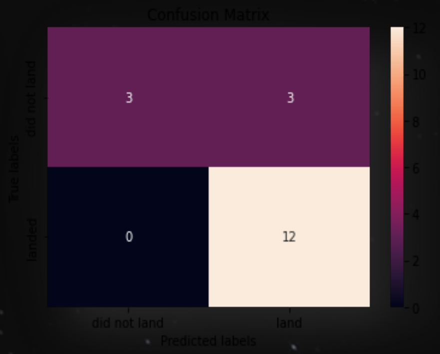
Logaritmic Regression

Accuracy: 83.33%



SVM

Accuracy: 83.33%



Decision Tree

Accuracy: 83.33%

KNN

Accuracy: 83.33%

Since all models performed the same for the test set, the confusion matrix is the same for all models. The model predicted 12 successful landings when the true label was successful landing. The models predicted 3 unsuccessful landings when the true label was unsuccessful landing. The models predicted 3 successful landings when the true label was unsuccessful landings(false positives.)

CONCLUSIONS

- FOR THIS DATASET THE DECISION TREE MODEL IS THE MOST SUCCESSFUL MODEL COMPARED TO OTHER MODELS
- UNDER PAYLOADS SUCH AS UNDER 6000 KG IS MORE SUCCESSFUL OTHER THAN HIGHER PAYLOADS
- KSC LC-39A HAS HIGHEST SUCCESS RATE
- THE RATE OF SUCCESSFUL LANDINGS ARE IMPROVING OVER THE YEARS
- SSO ORBIT HAS THE HIGHEST SUCCESS RATE

APPENDIX

. We could use additional classification models in order to using only 4 models to choose most successful model such as XGBoost, Naïve Bayes, Stoshastic Gradient Descent, Random Forest models.

Thank you!