IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

IBRAHIM HAKIM

7 January,2023

Outline

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

Executive Summary

This project focused primarily on predicting the class of rocket launch, viz 1 or 0.

The Data for the project was obtained from two main sources. Most of the data was from REST API of SpaceX which can be obtain from different endpoints. Another source was from then Wikipedia web page hosting several about Falcon 9 and Falcon heavy launches. The data from this source was obtained using requests from python module and parsed using the Beautiful Soup() objects.

Additionally, The data was processed and cleaned after it was stored into a pandas data frame to remove missing and incorrect data formats or types. Several exploratory data analysis was carried out to check for the significance of each feature to the project.

Executive Summary

- From the analysis,it seen that among all location,the site CCAFS SLC-40 had the highest success ratio of about 57.1% though KSC LC-39A had the overall success rate of 41.7% in total. Most of the launch site are in close proximity to highways and coastline but far away from the cities.
- The analysis also revealed that lower payload masses ranging between 1000-4000kg is associated with higher landing success rate, with payload mass of above 5000kg recording low success rates. A couple of launches made to orbits types such as GEO,HEO and SSO showed a very high landing success rate.

Executive Summary

- Since SpaceX started it rocket launches in 2010, The yearly trend of success rate has continued to increase. There was a sharp increase from 2013 continuously and reaching it peak in 2017.
- Lastly, several classification models were built through data modelling pipeline from which the best model was chosen. Though all the models performed appreciably well, the model which performed best was Decision tree model which was able to accurately classify 15 out of 18 records of the evaluation datasets.

Introduction

- SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers such as Virgin Galactic cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore if it can be determine that the first stage of Falcon 9 rocket launch 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. This project will focus on predicting whether the first stage of rocket launch by SpaceX will land successfully.
- The projects seeks to answer a wide variety of questions amongst which the most pressing ones are:

Introduction

- What features the affects probability that the first stage will land successfully?
- What sites were Launches made and what land features are in close proximity to them?
- What site location has the highest landing success rate?
- What Booster Version has the highest success rate?
- What payload mass has the highest success rate?
- What is the yearly trend of success rate since SpaceX launches started?

Section 1

Methodology

Methodology

Executive Summary

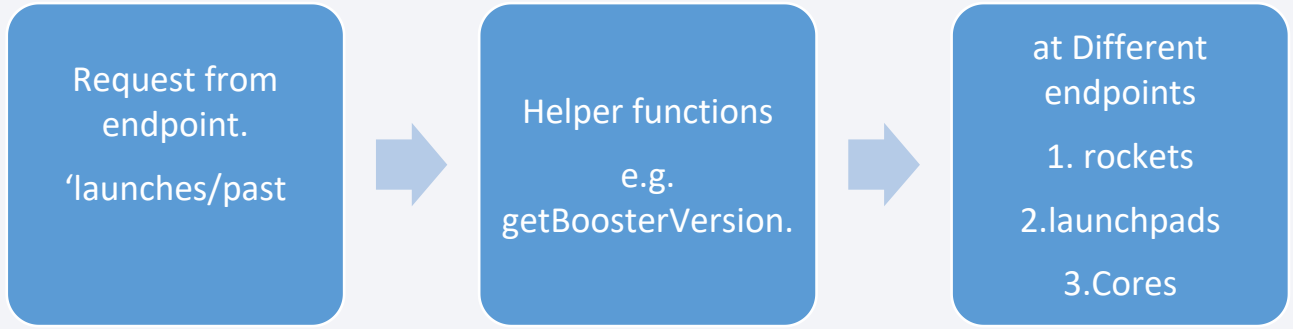
- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytic using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- Describe how data sets were collected.
- You need to present your data collection process use key phrases and flowcharts

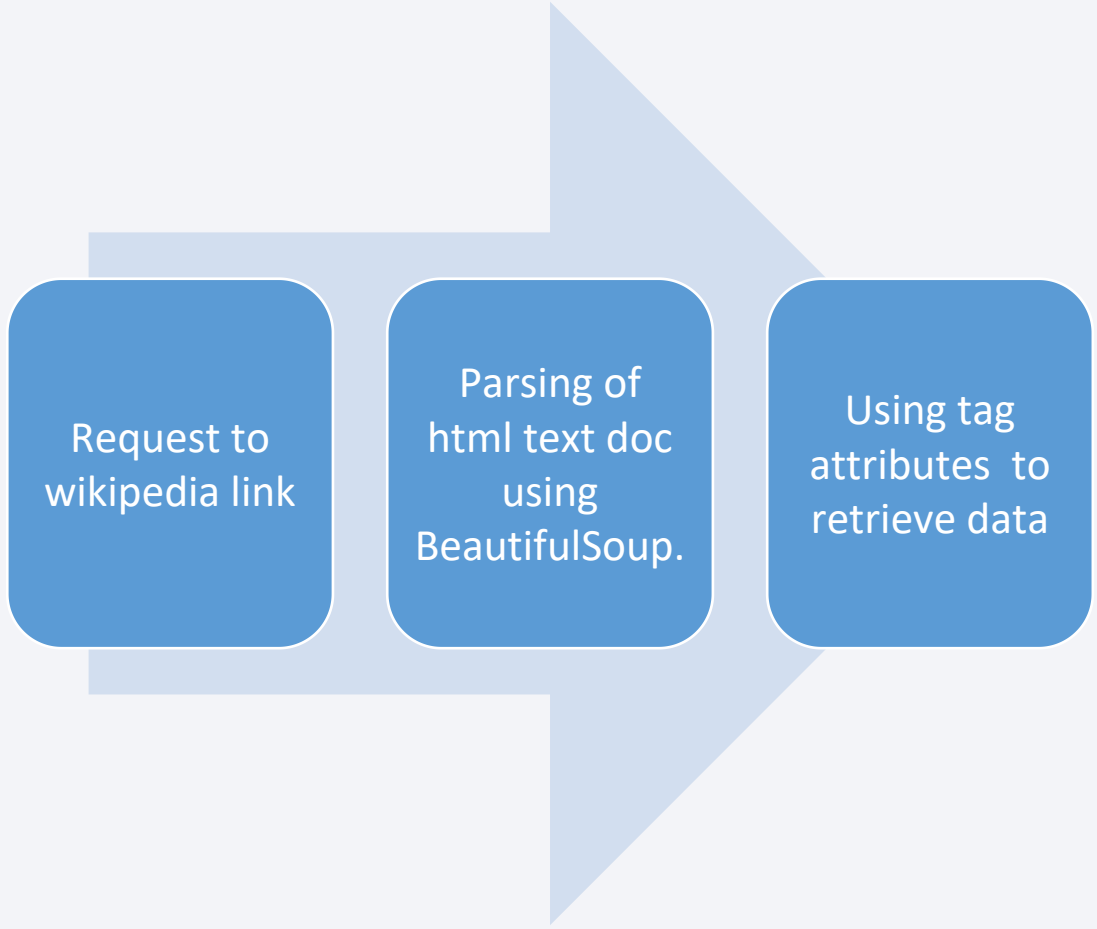
Data Collection – SpaceX API

- The Data for this project was collected through SpaceX API calls using the get request function. The API call was made through the endpoint 'launches/past' of the base url https://api.spacexdata.com/v4/ . Much of the data where Id's and therefore helper functions has to be defined to acquire the most useful data from different endpoints of the .
- The data was then cleaned by the helper functions as showed in the flowchart. The detailed process of the data collection is saved in this notebook link.



Data Collection – Scraping

- Additionally, some of the data was also collected through web scraping of Falcon9 and Falcon Heavy launch records on Wikipedia. The html text obtained was parsed using python BeautifulSoup function to acquire the desired tabular data
- Several similar python functions where used to extract the data for each row and column. The data was also queried to retrieve on data on Falcon 9 launches
- The whole web scraping process is can be accessed for review in the git hub repository link



Data Wrangling

- The data obtained has few empty rows and mismatched data types. Most of data were of objects,floats and boolean values.
- The Data obtained where cleaned to remove empty rows from the data. The rows of Payload mass was replaced with the mean value of data column since it is a numeric column. That of the landing pads which has the highest number of null values were left empty.
- an additional column (Class i.e 1 or 0) was added to accurately understand the true class of each record. The class was obtained from the column consisting of values(True ASDS, False ASDS,True RTLS, False RTLS True Ocean, False Ocean etc)
- A lot of the wrangling process was an exploratory Data analysis to understand the features of the data. From which was observed that SpaceX had a success rate of 66.67% of the first stage landing.
- The notebook containing the step-by-step process of data wrangling process can be accessed through this link

EDA with Data Visualization

- The Exploratory Data analysis with Data visualization focused primarily on scatter, bar and line plots.
- The scatter plots where used to understand the structure of numerical features based on success rate of a particular launch. This plot focused on launch site, payload mass,flight numbers with highest success rate whilst the line plot was used to visualize the trend of success rate in launch years. Moreover, The bar charts was used to understand the orbit types with the highest success rate.
- The GitHub URL of the completed EDA with data visualization notebook can be accessed here.

EDA with SQL

- An additional exploratory Data Analysis was carried out using SQL.
- The Exploratory Data Analysis with SQL focused on queries to examine the most prominent features within the data. The exploration were mainly select queries to;
  - Display unique launch site.
  - Launch sites begging with special string i.e 'CCA.
  - Total Payload mass with a specific Spacecraft Customer.
  - Average Payload mass that a specific Booster Version had
  - Date when the first landing success was achieved.
  - Rank of Landing Outcomes that Occurred between specified dates.

- Access the GitHub URL of the completed EDA with SQL notebook, as an external reference here

Build an Interactive Map with Folium

- The visual in this section is a map with a start location set to NASA Johnson Space Center. Marker clusters have being added consequently to each launch site with each marker identifying the class of the launch at that launch site as either successful (green) or otherwise (red). Also, Distances where calculated and poly line where added in order to determine the distance between the launch site to some important land marks such as railway, highway, coastline or city. This helped to determine the which land mark is in close proximity to this landmarks.
- Generally the map features added to the map were marker clusters, markers with popups, and poly lines.
- The GitHub URL of the completed interactive map with Folium, as an external reference can be accessed here

Build a Dashboard with Plotly Dash

- An additional Dashboard was created to better understand the launch site features. The visuals or components added were; Pie chart, Scatter plot, drop down option picker and a range slider.
- The pie chart was used to show the count of Class of each launch site whiles the scatter plot was used to show trends of how the class of a launch changes with changes in payload mass differentiated by Booster versions. This graphs where accustomed to interactive features such as the drop down to drill through different launch sites whiles the range slider offered the opportunity of selecting ranges of payload mass. This selection is important to view the range with the highest landing outcome.
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose

Predictive Analysis (Classification)

- With the available cleaned data, several classification models were built. Amongst the models built were k-nearest neighbor, support vector machines, regression tree classifier and logistic regression. The data was first transform using standard scalar so that the data point are within the same range. In order to ensure the models built had high out-of-sample accuracy, the data was split into training and testing datasets using train_test_split function of the sklearn. Due to parameter tuning involved in most classification models, The GridSearchCV function from the sklearn.model_selection was used to determine the best hyperparamters in each classification model built. This cross validation also speedup the model building process since it provided pipeline through which different parameters can be selected easily to fit the model.
- The score() method of the GridSearchCV was primarily used to evaluate the accuracy of the models on the evaluation data.
- The GitHub URL of the completed predictive analysis lab, as an external reference can be accessed here.

Results

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

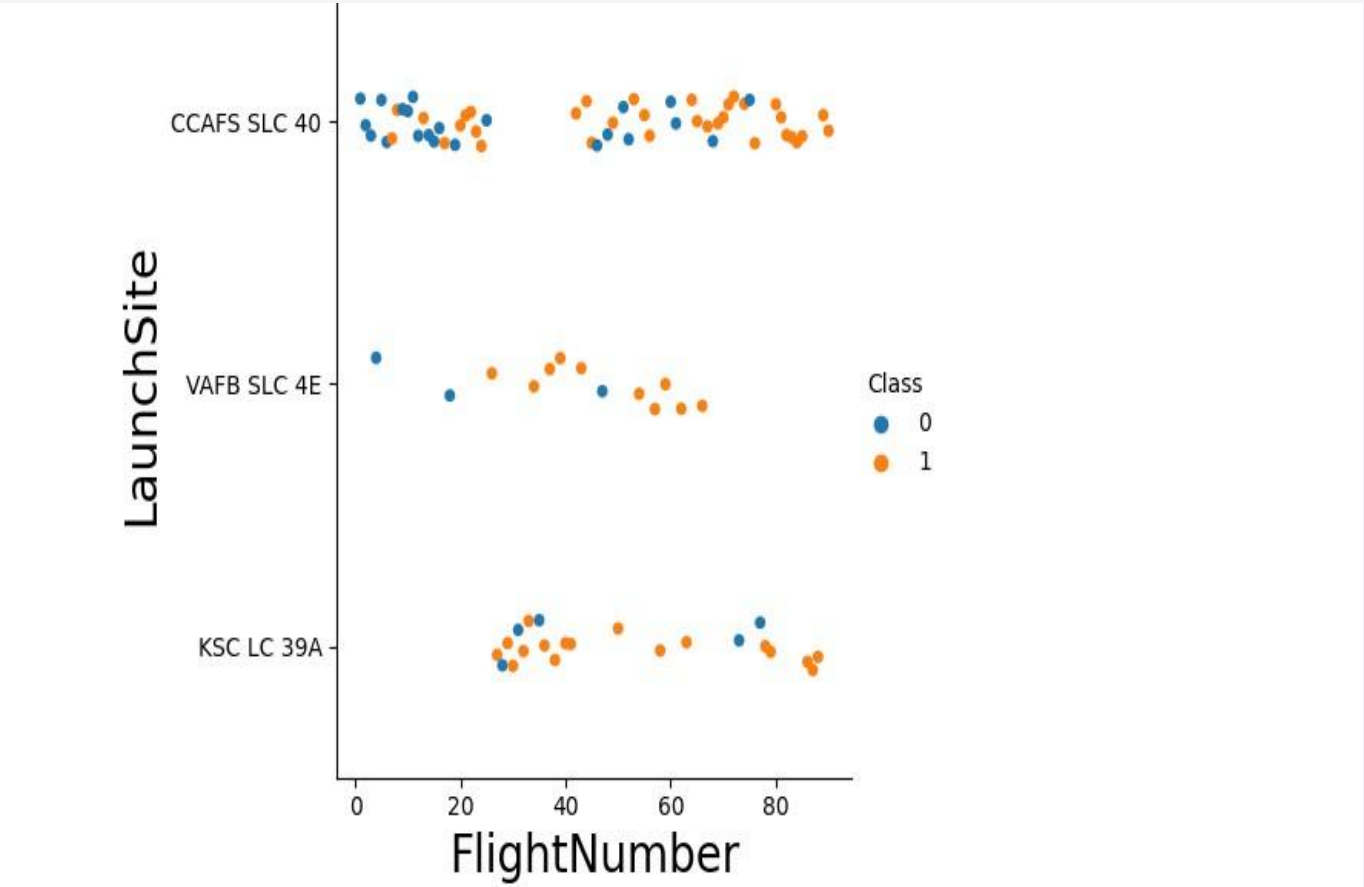
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

It is seen from the scatter plot that the launch site with a highest success rate is 'VAFB SLC 4E', followed by 'KSC LC 39A'.

Consequently the launch site 'CCAFS SLC 40' has a high proportion of unsuccessful outcomes compared to any other launch site.



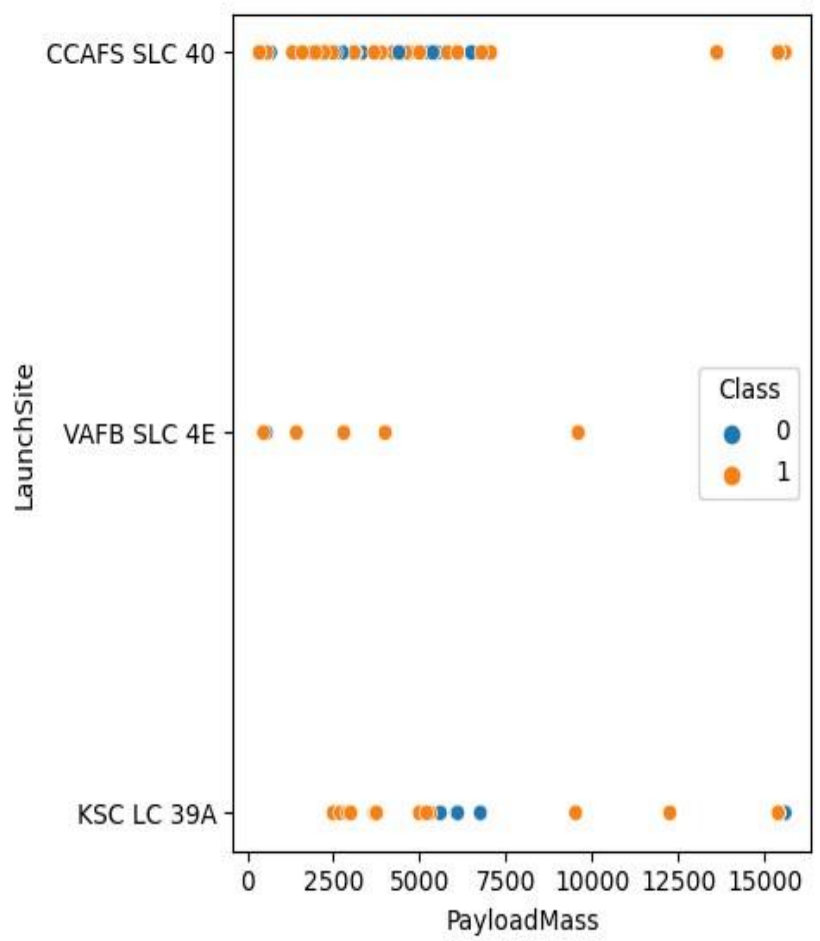
Payload vs. Launch Site

Different payload mass has different success rate.

It can be seen that for all launch site, most success rate was recorded at a payload mass less than 4000kg.

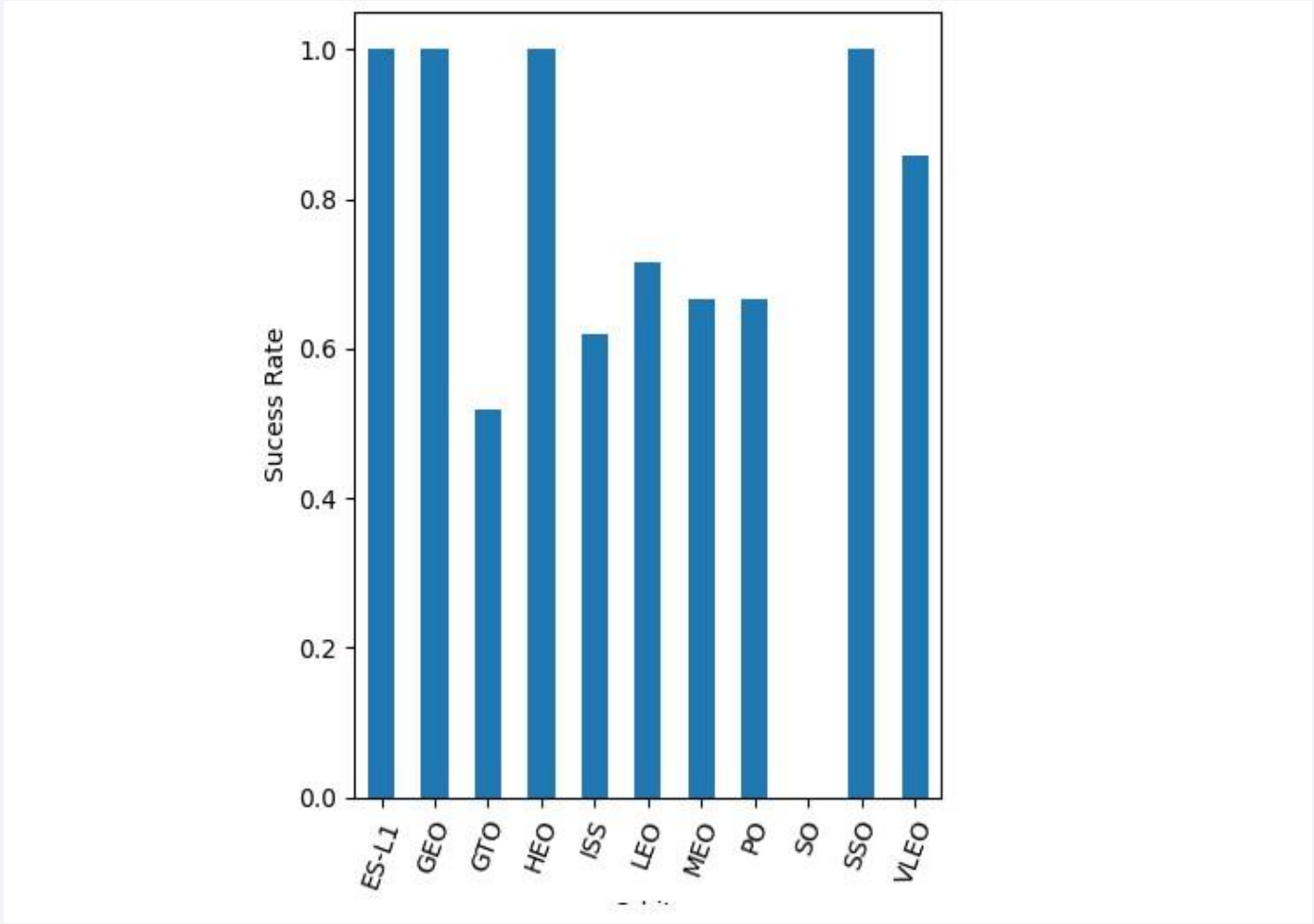
It will be evident enough to explore success rate across different ranges of payload mass.

It is also seen that the launch site 'VAFB SLC 4E' which had launches below 5000kg had a very high launch success rate



Success Rate vs. Orbit Type

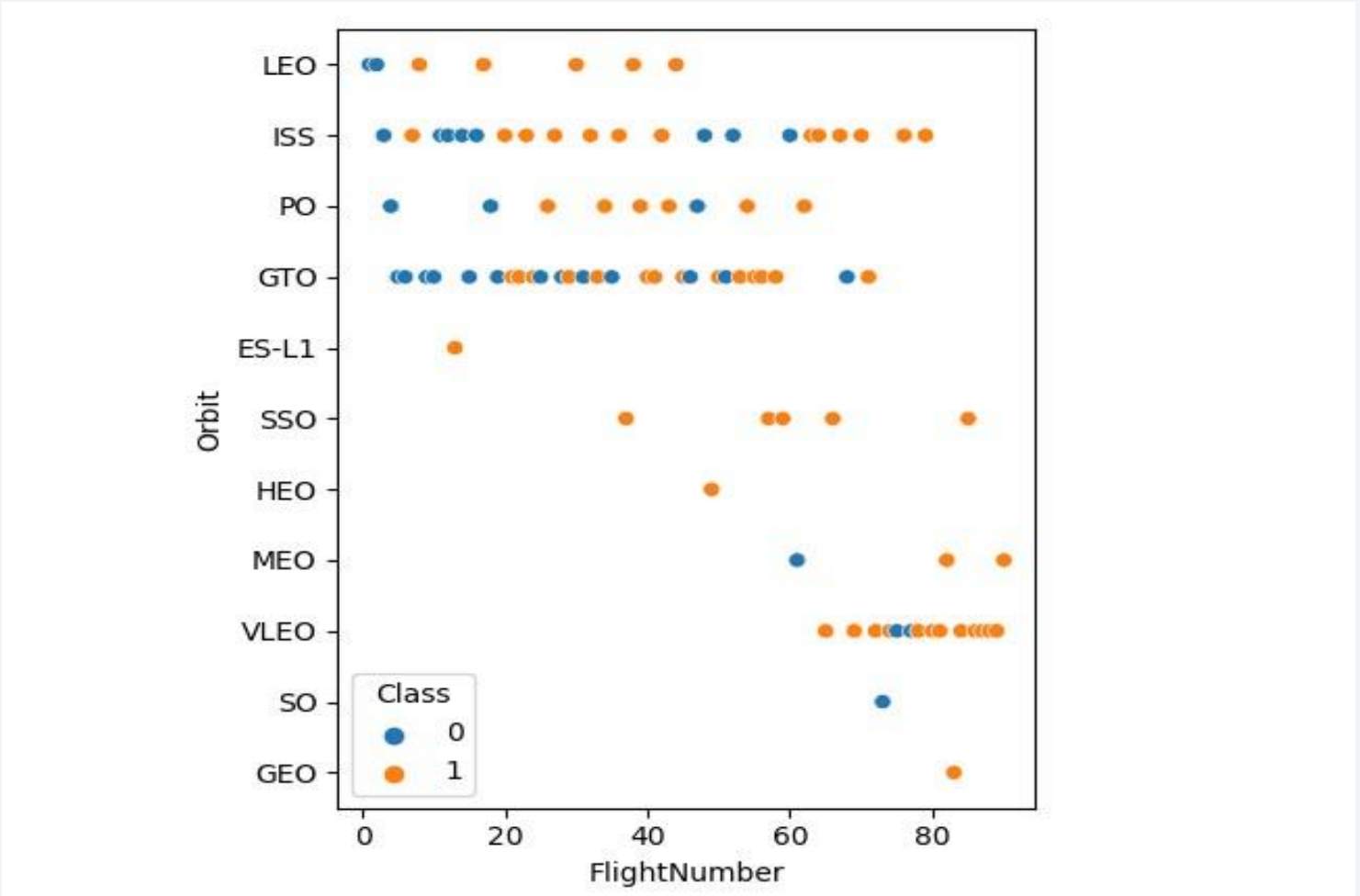
Different launches are made to different orbits in space. This orbits differ by their distances to the earths equator. This graph depicts that, the orbits ES-L1, GEO,HEO and SSO had the highest landing success rate almost close to 100%.



Flight Number vs. Orbit Type

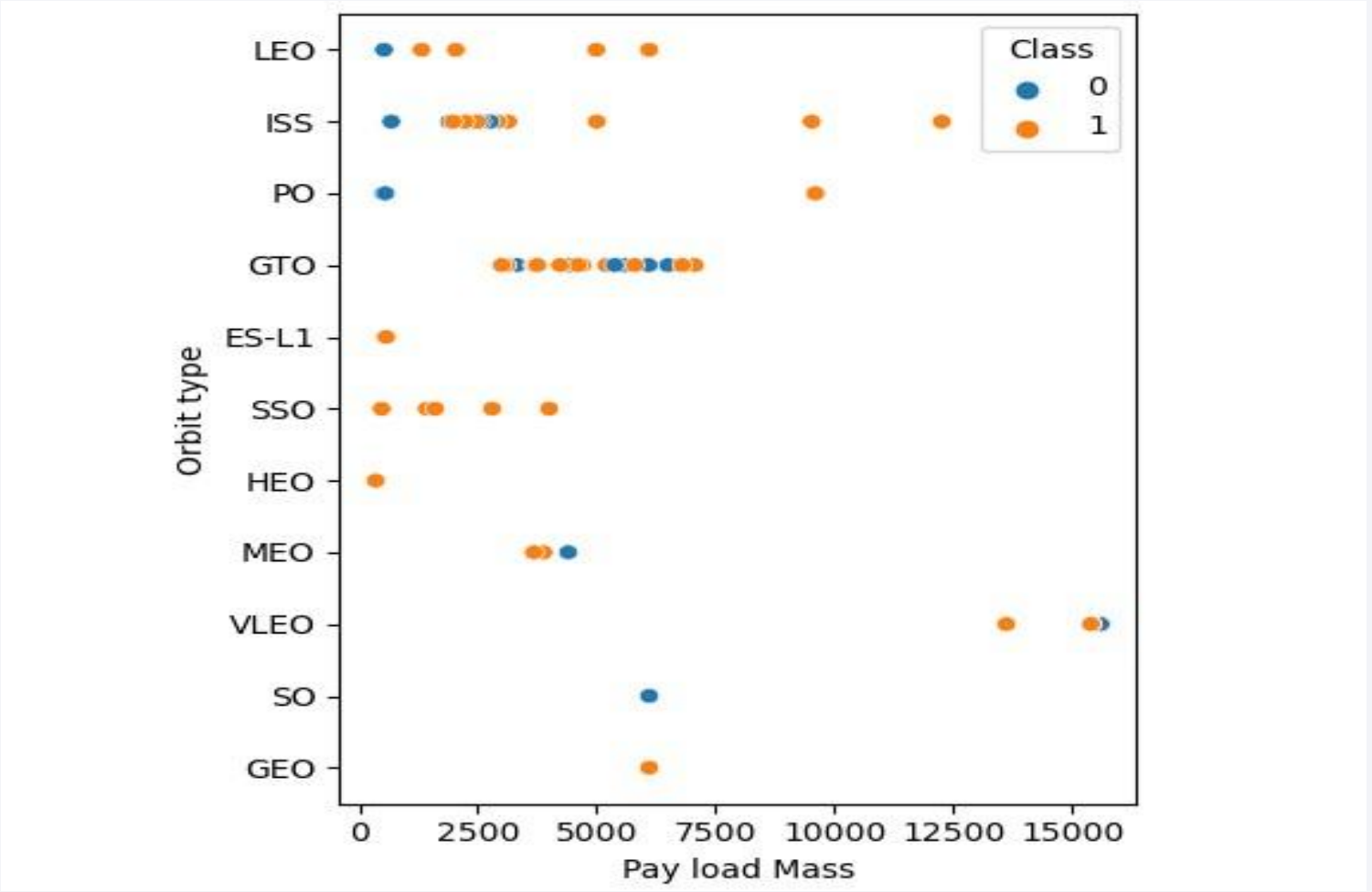
With an increasing Flight number, subsequent launches made to orbit type such as SSO, VLEO had a high landing success rate.

Most unsuccessful landing outcomes from orbits such as ISS, GTO occurred in launches carried earlier with flight numbering below 40



Payload vs. OrbitType

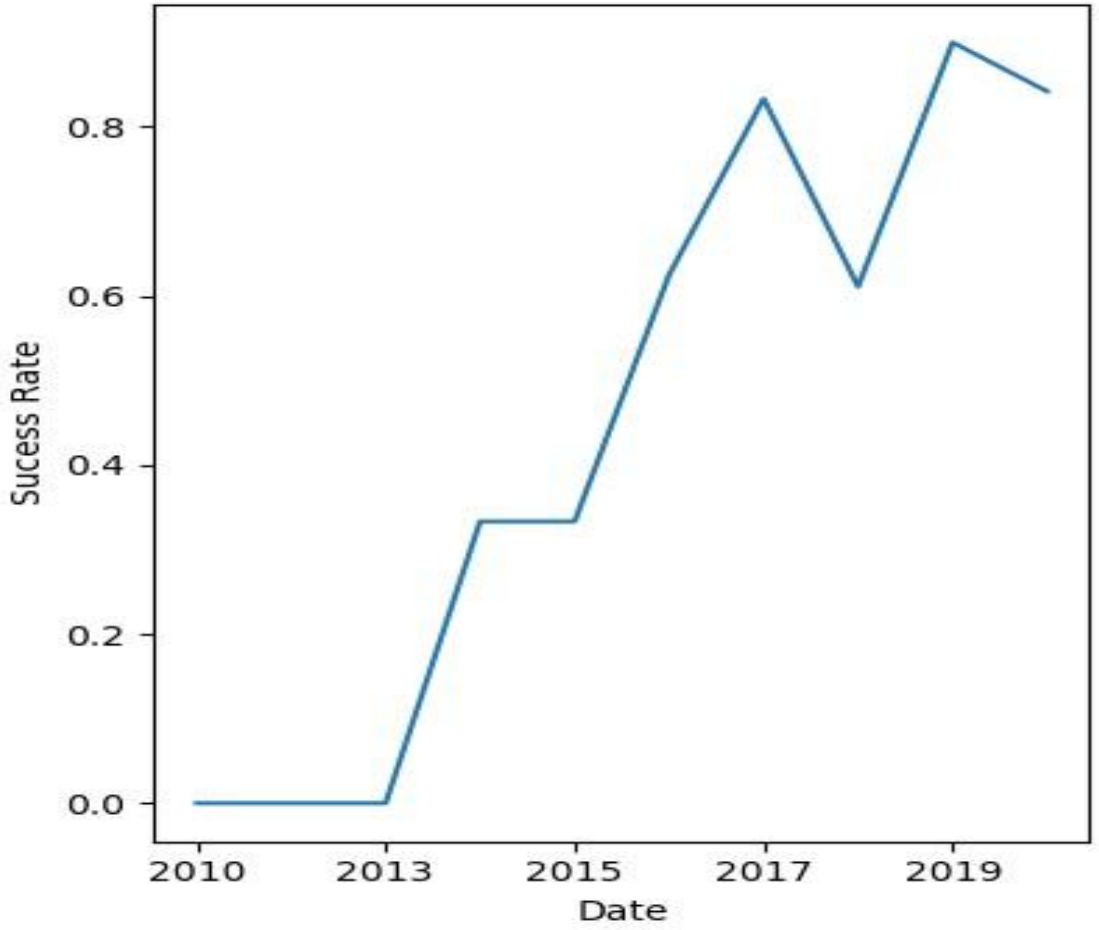
Though there is less inference that can be drawn from this plot, it is seen that launches to orbit SSO with payload mass below 5000kg had a high success rate.



Launch Success Yearly Trend

With increasing number of launches, SpaceX had increase in landing success rate.

It can be seen from the graph that, the success rate started to increase rapidly from 2013 continuously until 2018 where it declined.



All Launch Site Names

SpaceX launch site of their rockets were at the following

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

The launch site are coded locations around NASA Johnson Space Center

Launch Site Names Begin with 'CCA'

- This are launch records beginning with 'CCA' limited to only the first five records from derived SpaceX datasets

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

The total payload mass carried by boosters from NASA is 45,596kg.

- This means that all launches made by NASA (CRS) weigh below 5000kg in payload mass of the spacecraft.

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2928.4kg

This is a query result from SpaceX Table in SQL.

First Successful Ground Landing Date

- Many launches had successful landing outcome either by ship drone or otherwise. The dates of the first successful landing outcome on ground pad occurred in 1st May,2017
- This is the result of query result from SQL attached for references.

Successful Drone Ship Landing with Payload between 4000 and 6000

- Furthermore, different boosters have different payload mass which greatly affects the landing outcome of the spaceship. The List of names of boosters which had successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are as follow;
  - F9 FT B1022
  - F9 FT B1026
  - F9 FT B1021.2
  - F9 FT B1031.2

All of which are Falcon 9 rocket launches.

Total Number of Successful and FailureMission Outcomes

- Launch missions may be successful or otherwise. In all the total number of launches with a successful mission is 100 with only 1 unsuccessful outcome.

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

Boosters Carried MaximumPayload

- The list of names of the booster which have carried the maximum payload mass are :

1. F9 B5 B1048.4
2. F9 B5 B1049.4
3. F9 B5 B1051.3
4. F9 B5 B1056.4
5. F9 B5 B1048.5
6. F9 B5 B1051.4
7. F9 B5 B1049.5
8. F9 B5 B1060.2
9. F9 B5 B1058.3
10. F9 B5 B1051.6
11. F9 B5 B1060.3
12. F9 B5 B1049.7

2015 Launch Records

- The List of failedlanding_outcomes in drone ship, their booster versions, andlaunch site names for in year 2015 are displayed in the image

substr(Date,4,2)	Failure landing_outcome(drone ship)	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
02	Failure (drone ship)	F9 v1.1 B1013	CCAFS LC-40
03	Failure (drone ship)	F9 v1.1 B1014	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1016	CCAFS LC-40
06	Failure (drone ship)	F9 v1.1 B1018	CCAFS LC-40
12	Failure (drone ship)	F9 FT B1019	CCAFS LC-40

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

- The rank of 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 is depicted in the diagram.

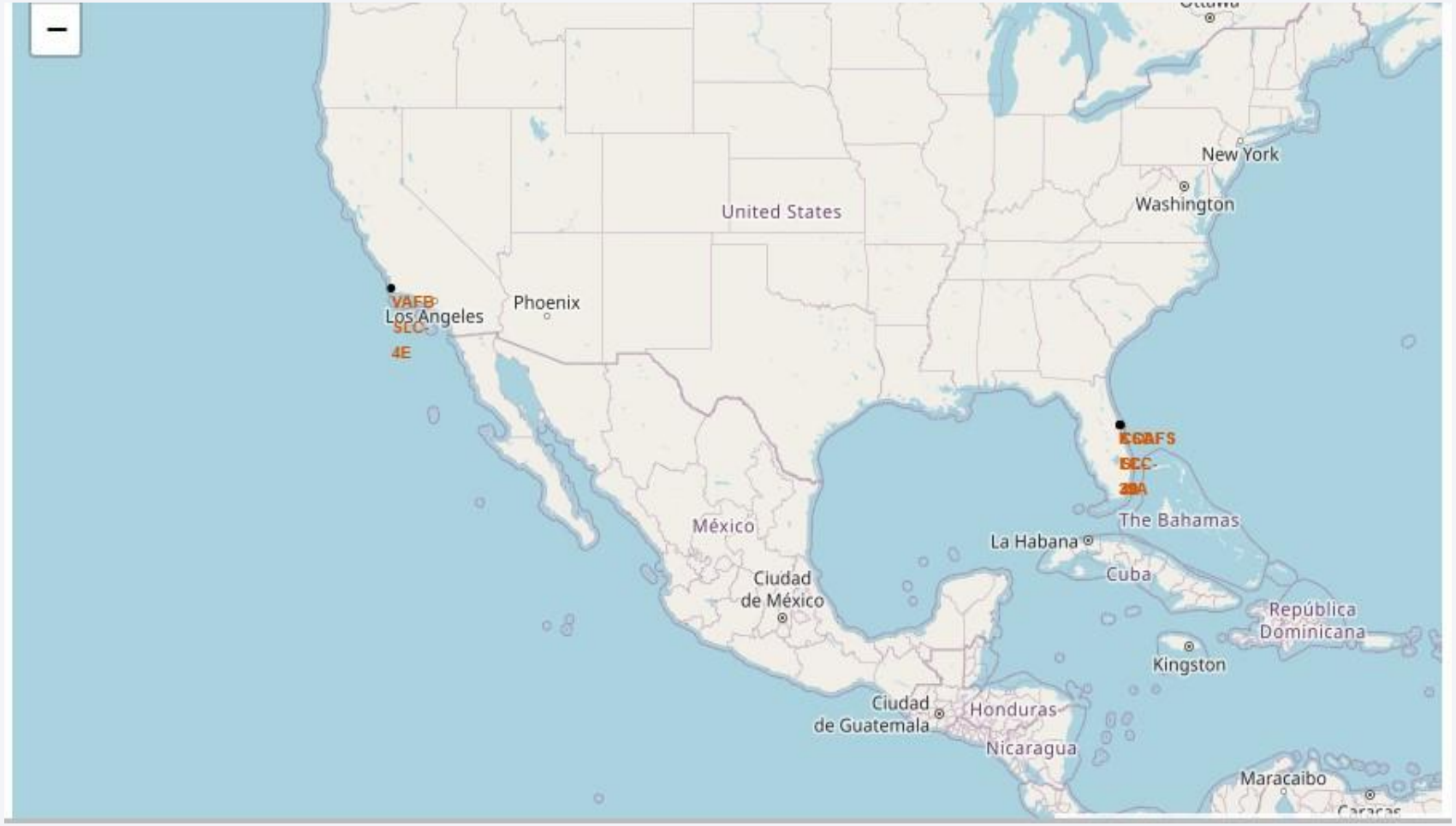
Landing _Outcome	COUNTS
Failure (parachute)	57

Section 3

Launch Sites Proximities Analysis

LAUNCH SITE LOCATIONS

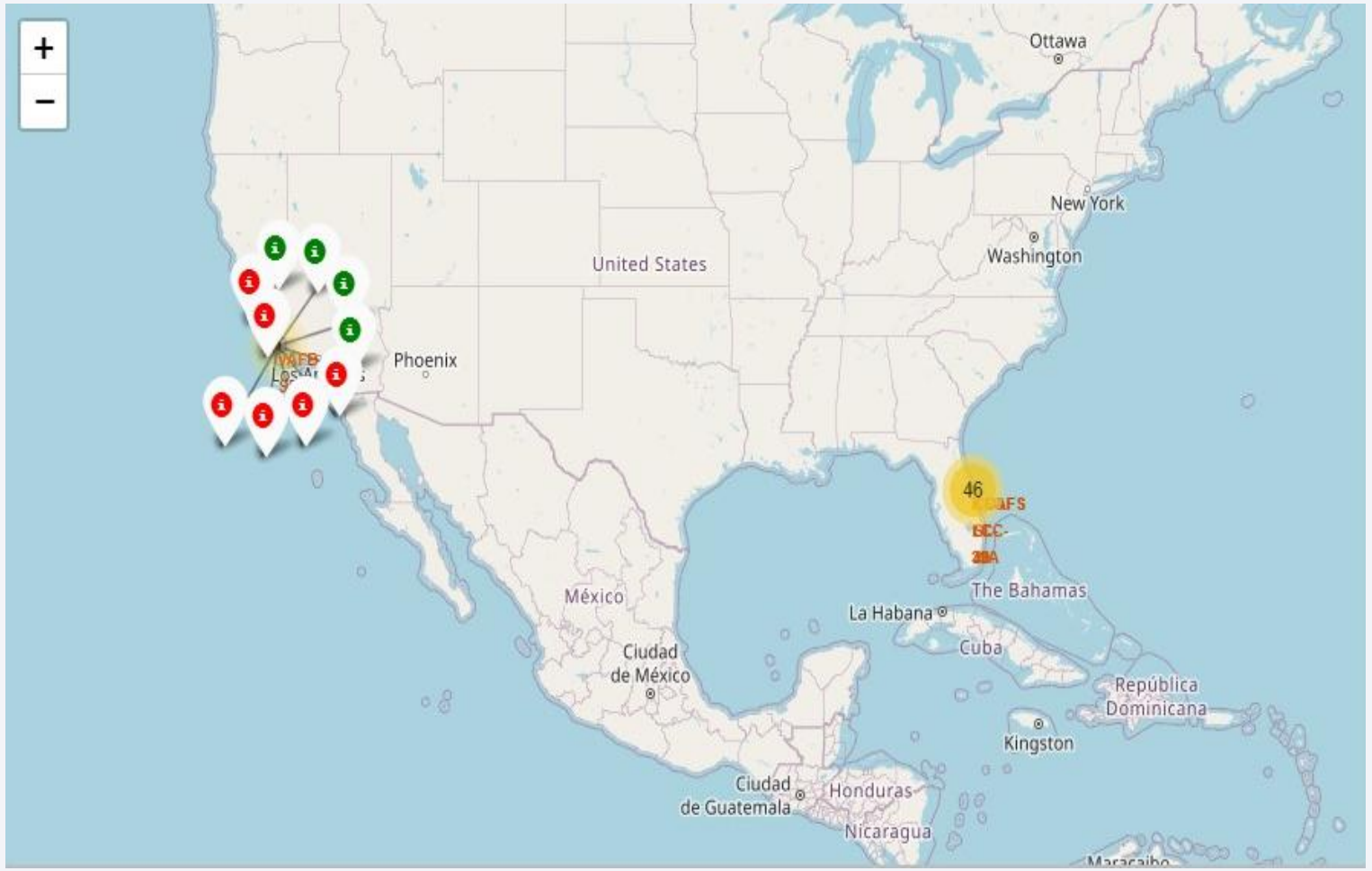
From the map, it is seen that the launch site are wide apart from each other with only few starting with similar coded names being in close proximity. It can also be seen that the launch site are near to the coastline.



LAUNCH SITE LANDING OUTCOME

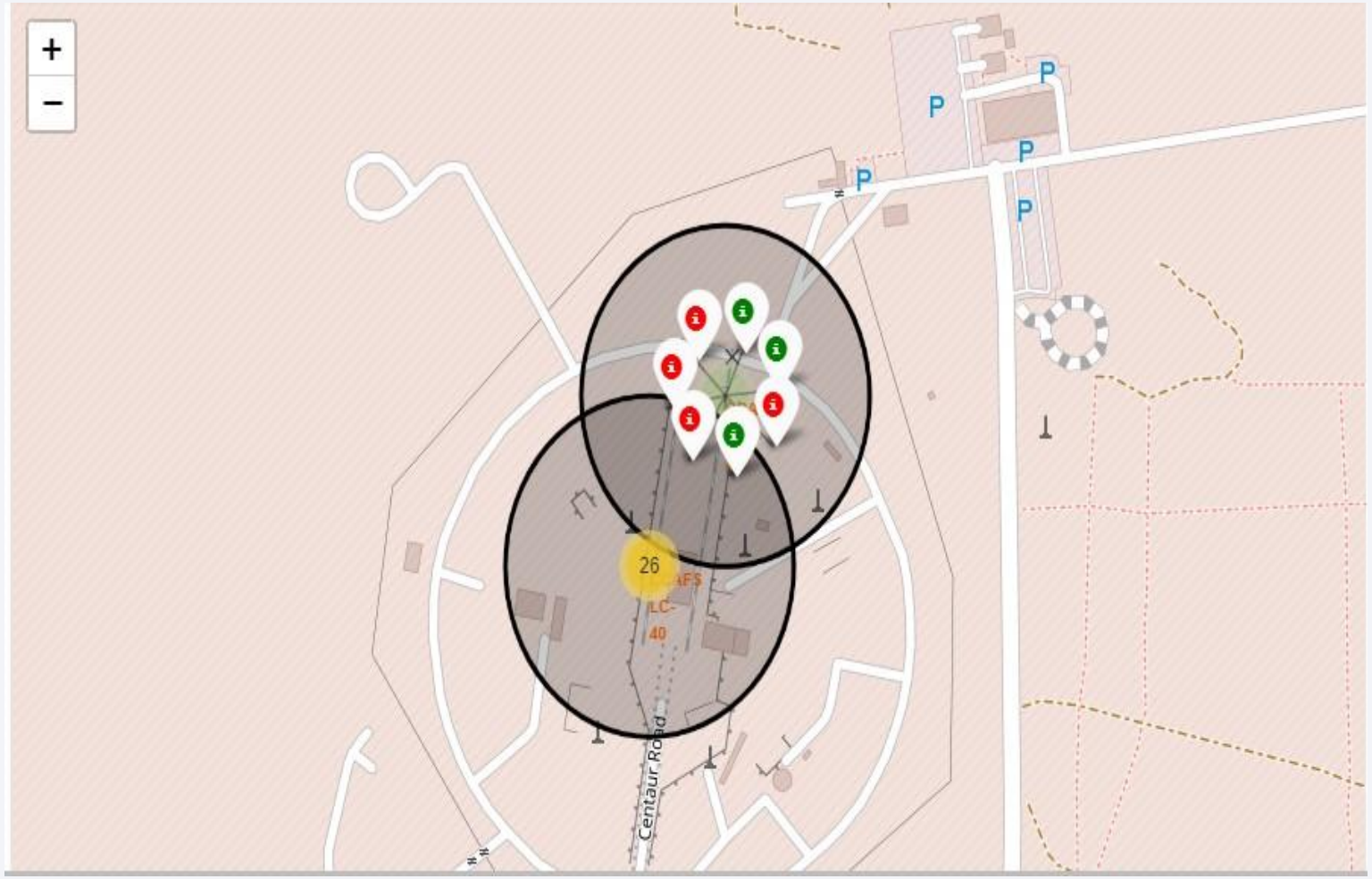
Exploring the map features further, The color coded features at each launch sites a markers added to track the landing outcomes of launches at a particular site. Green for successful landing and red otherwise.

This map therefore shows a low successful landing at launch site VAFB SLC-4E.



LAUNCH SITE LANDING OUTCOME

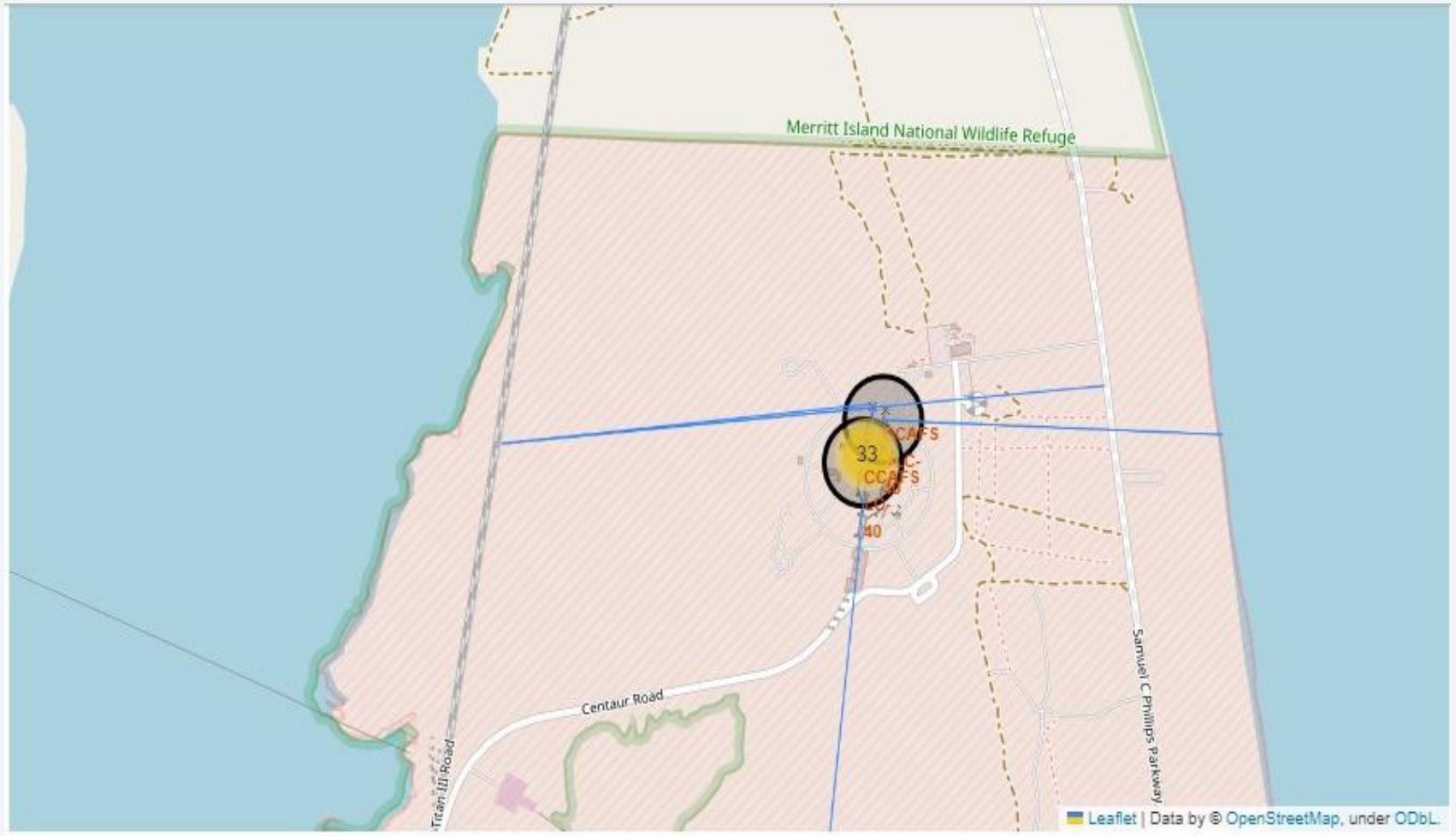
- The same can be said of this map feature but with higher successful landing outcomes compared to the above.



LANDING SITE PROXIMITY

From the map it is seen that the launch site is closest to highway with a distance measuring about 0.60km. Other land features in close proximity to a launch site is coastline (0.88km).

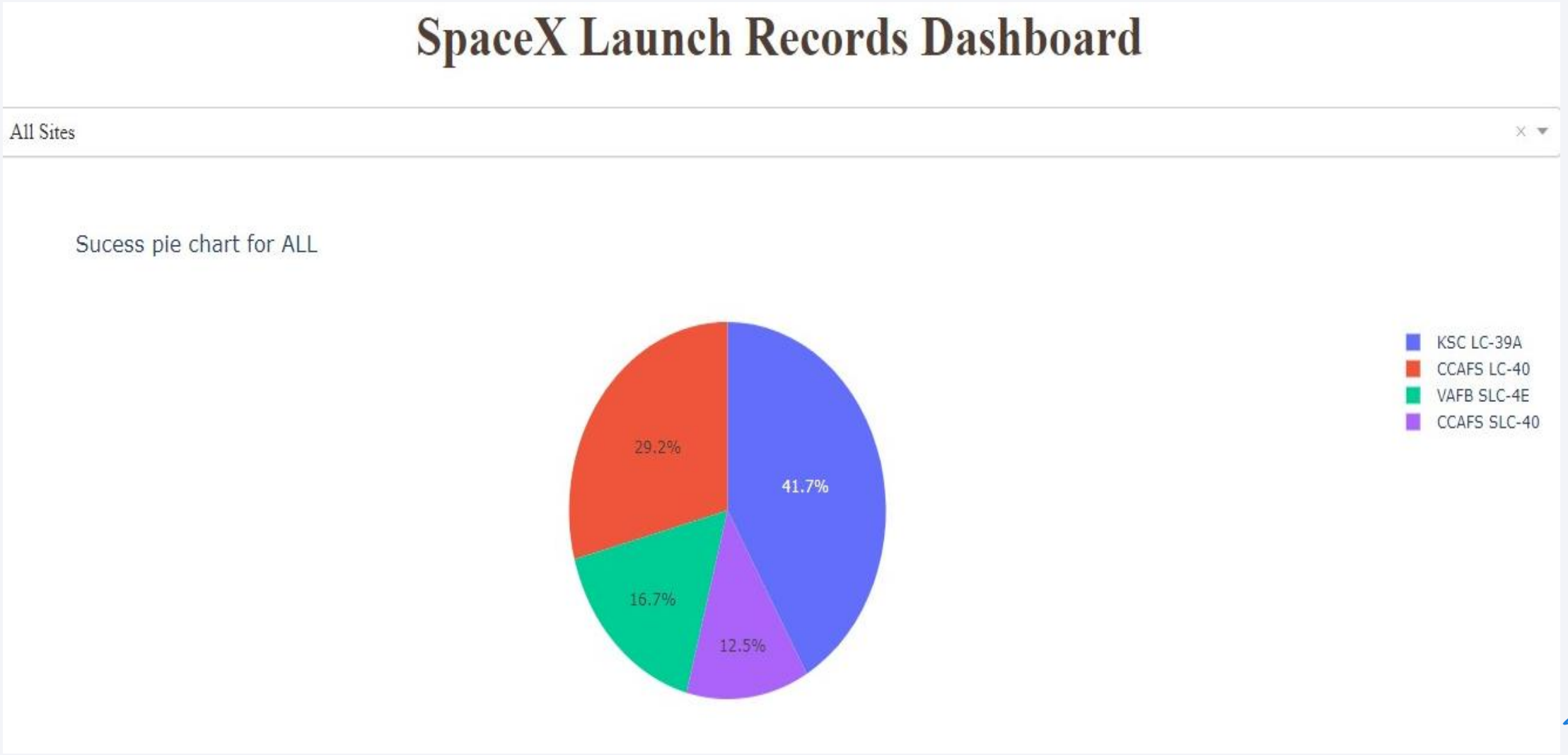
This help to answer a question of what land feature is in close proximity to a launch site.



Section 4

Build a Dashboard with Plotly Dash

LAUNCH SITES SUCCESS COUNTS



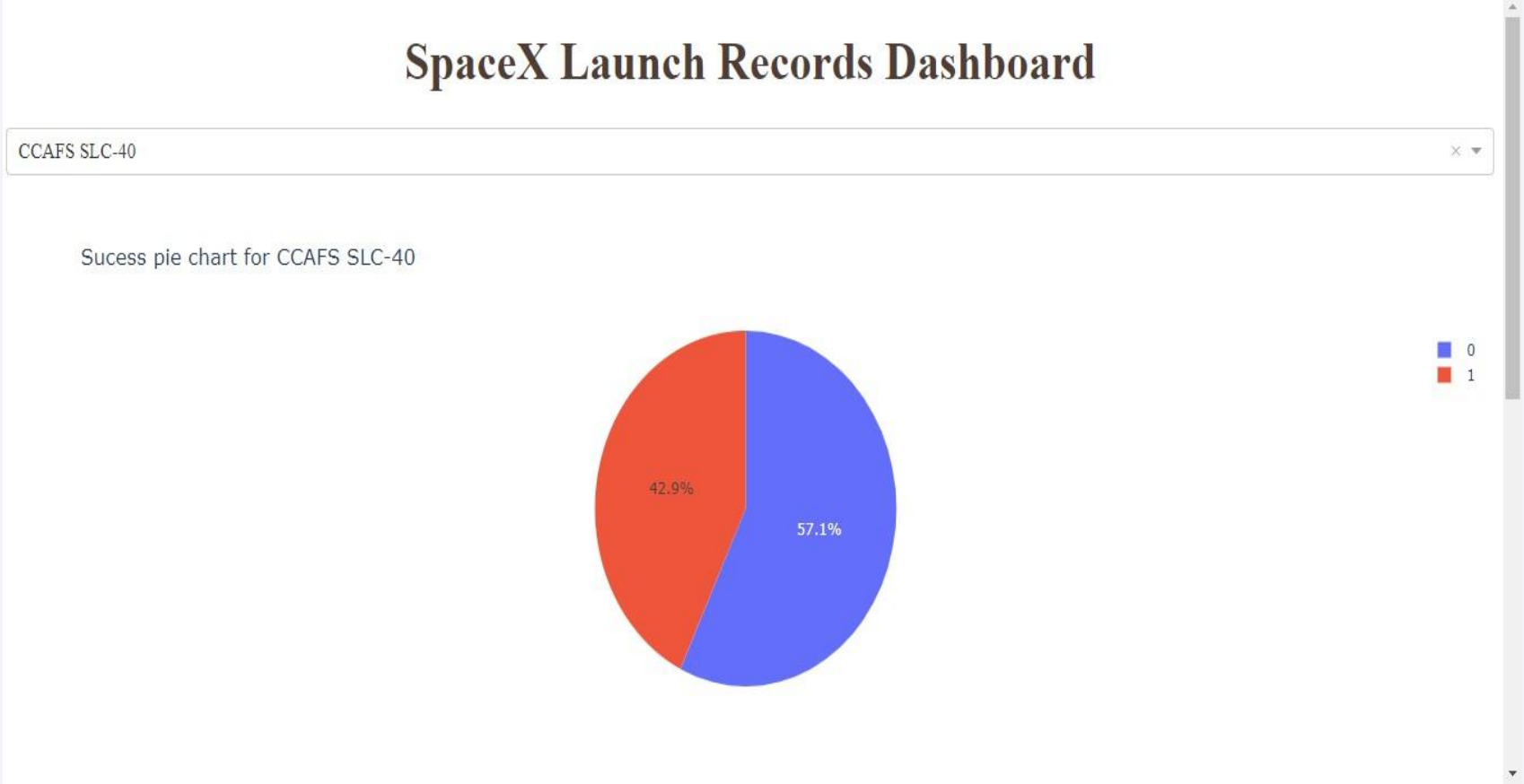
LAUNCH SITES SUCCESS COUNTS

Further exploring the Dashboard shows that the launch site with the highest success rate of 41.7% is 'KSC LC-39A'.

- The second is CCAFS CS-40 with a percentage of 29.2%, followed by VAFB SLC-4E and then CCFS SLC-40.
- This show the overall success rate across all launch sites and may differ by success ratio per a launch site as was depicted in the map.

LAUNCH SITE SUCCESS RATIO

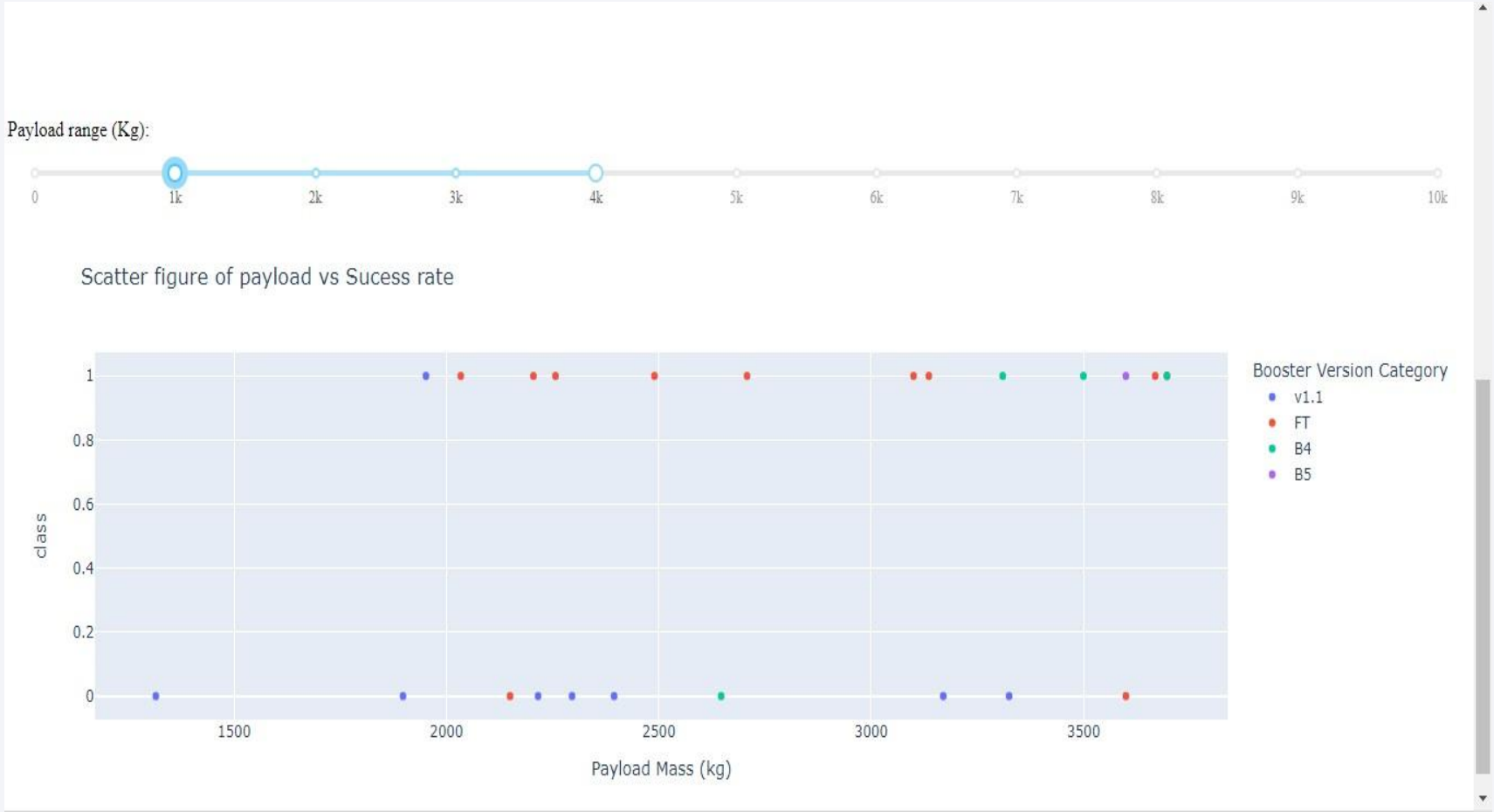
Again, as was explained above, the launch site with the highest success ratio is CCFS SLC-40. Though it is the launch site with the least overall successful landing outcome, it recorded the highest success ratio and may be due to the number of launches that site.



LAUNCH OUTCOME VRS PAYLOAD STRUCTURE

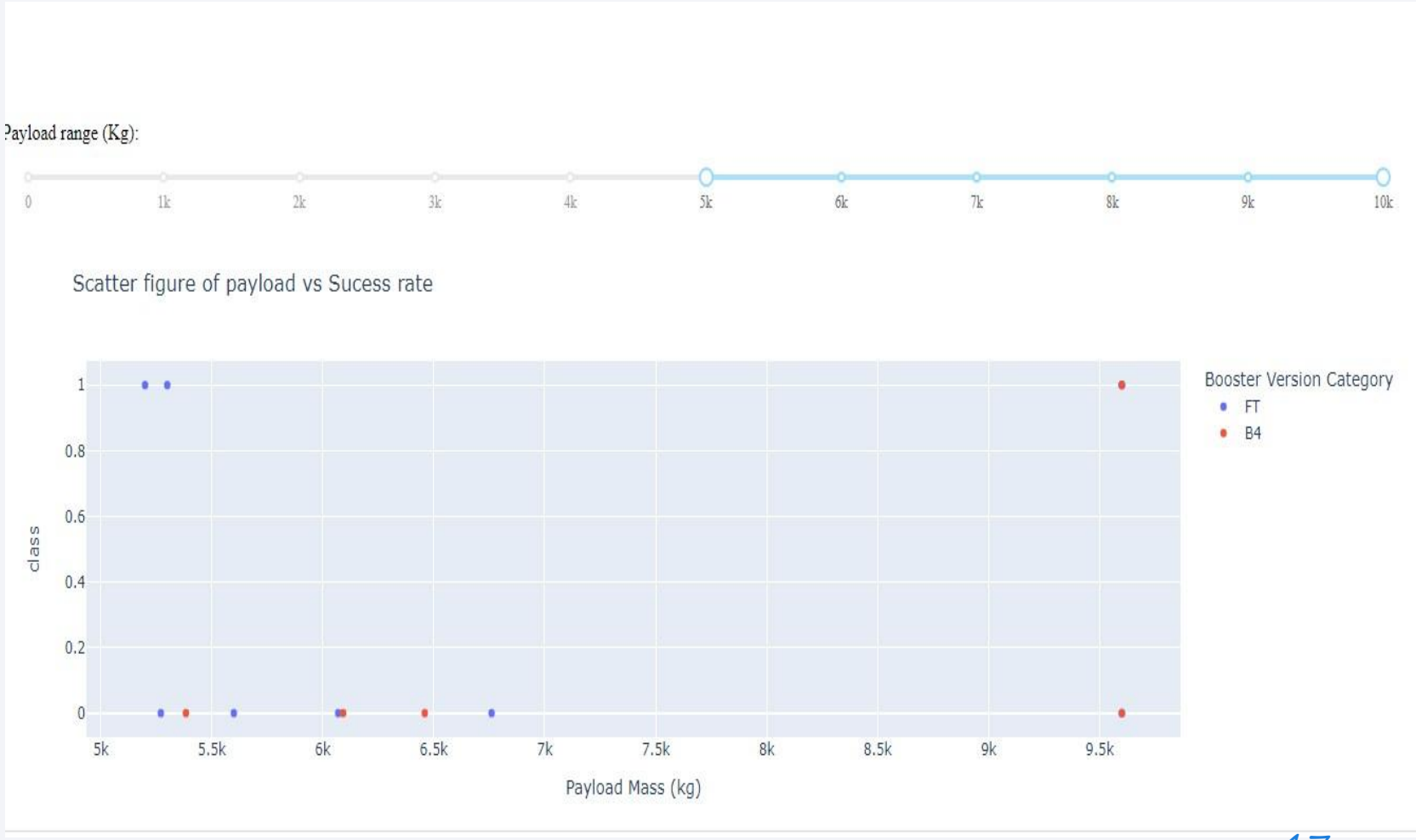
Determining the payload mass with the highest success rate is crucial.

From the graph, it is seen that payload mass of between 1000-4000kg has the highest success rate. This shows that rockets of payload mass equal to or less than 5000kg has high success rate.



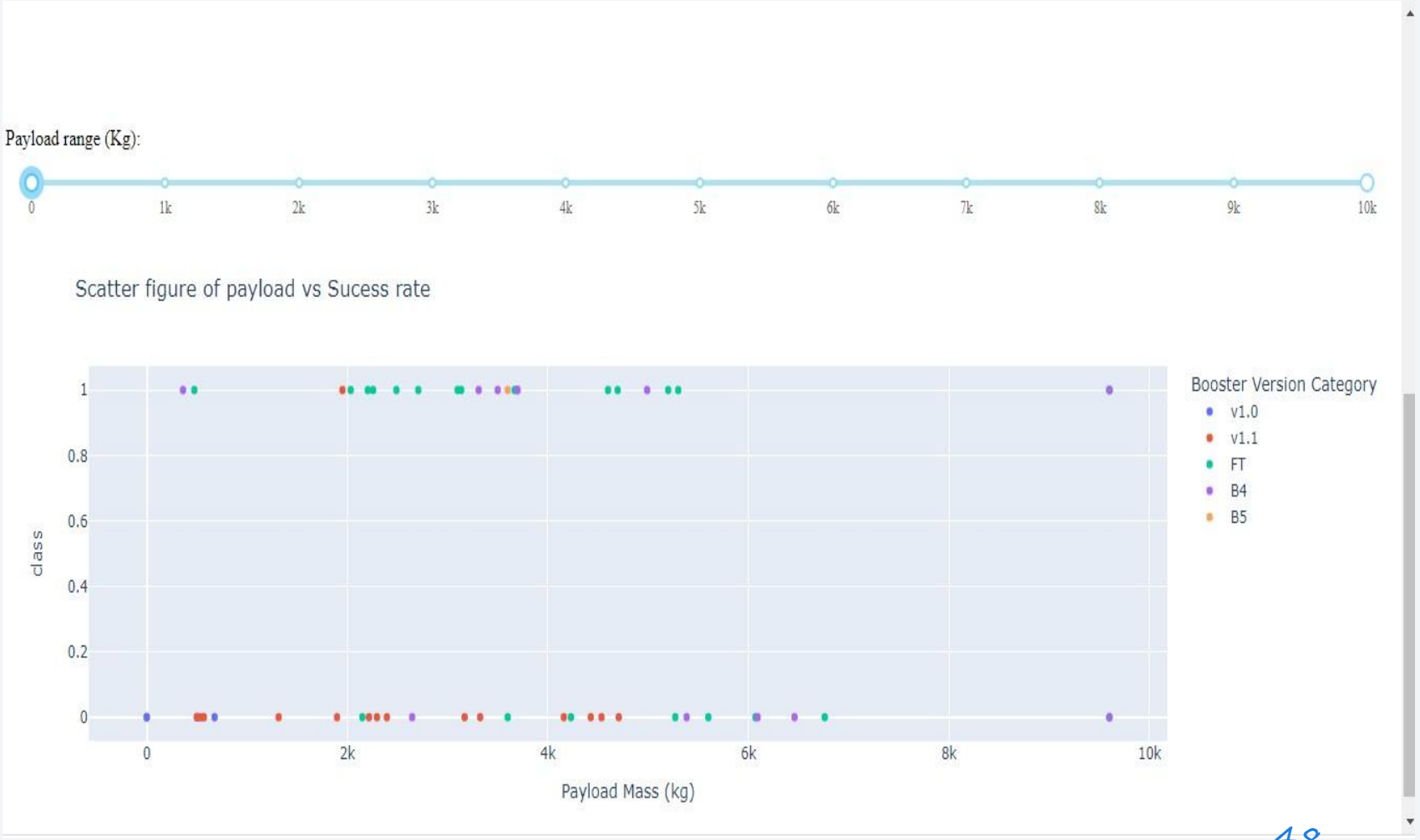
LAUNCH OUTCOME VRS PAYLOAD STRUCTURE

- It is also evident from the graph that payload mass equal to and greater than 5000kg has a low success rate.



BOOSTER VERSION SUCCESS RATE

- Moreover the Booster version category with the highest success ratio is FT. This means rockets with this booster version may land successfully.

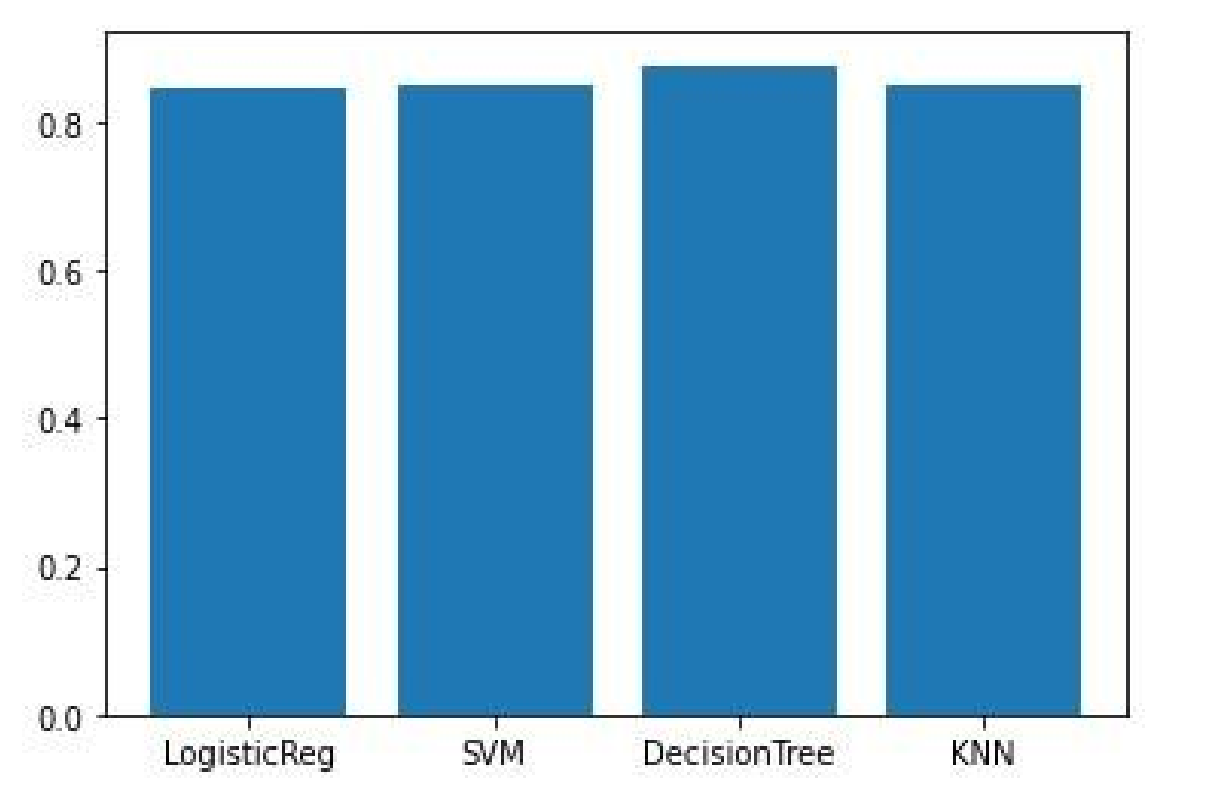


Section 5

Predictive Analysis (Classification)

Classification Accuracy

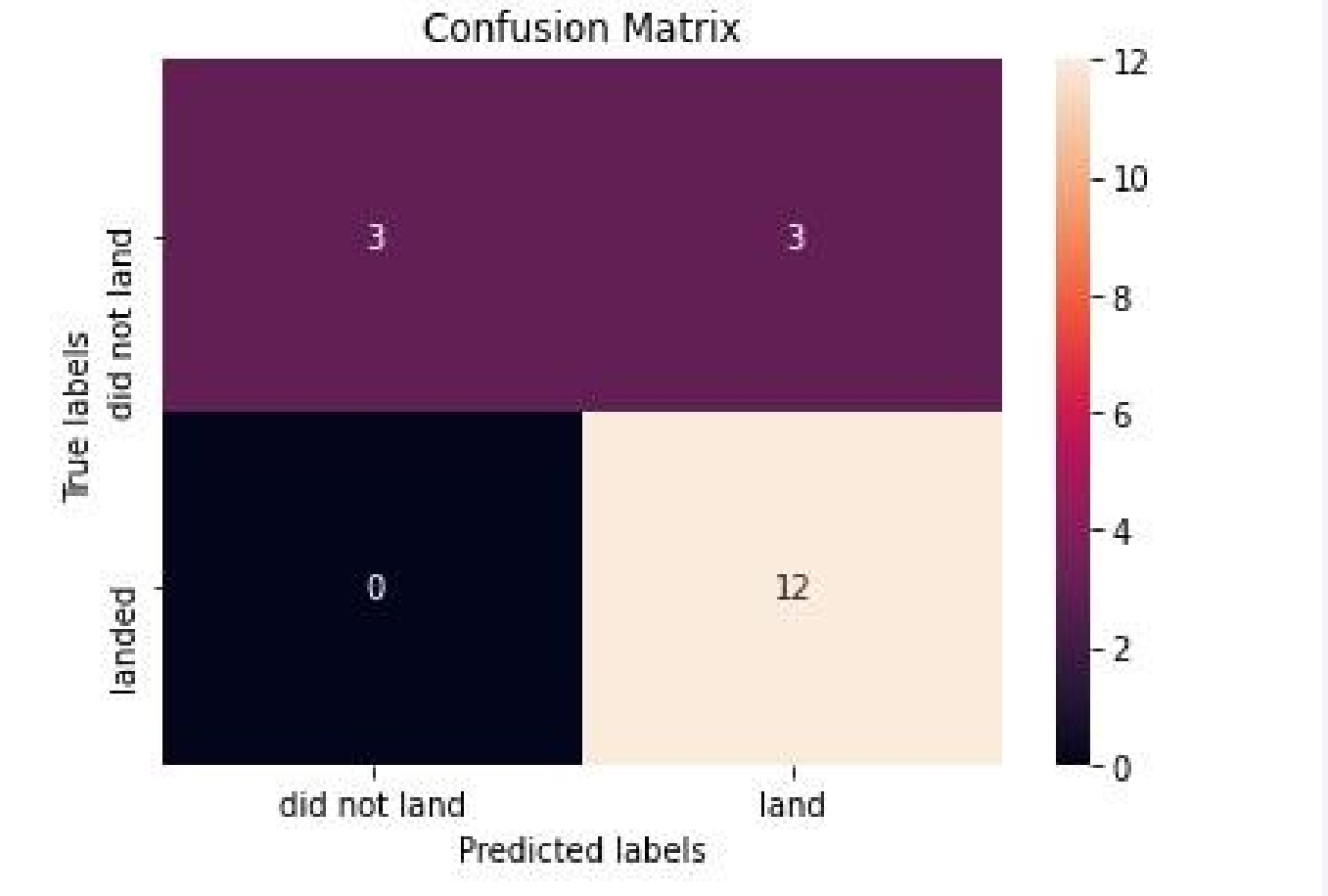
Though all the model performed preferably well with an accuracy of 80%, the model that performed best with higher accuracy as can be seen from the graph was Decision Tree classifier. The model had an accuracy of 88 percent.



Confusion Matrix

The model was able to accurately classify 15 out of the 18 records of the evaluation data correctly.

This shows that the model had a higher out of sample accuracy and will work well on new data.



Conclusions

- Though several features affects the probability of the first stage to launch, the features that most affects the success rate are, the payload mass, Booster Version, orbit types synchronous to the earth.
- Launches were made at CCAFS LC-40, VAFB SLC-4E,KSC LC-39A and CCAFS SLC-40 which are in close proximity to coastline and highway. Launch site are chosen such that they are away from cities.
- Also, The launch site with highest success ratio is CCAFS SLC-40 though it is the site with fewer launches.
- Amongst the Falcon 9 rocket launches with several booster version, the category of the versions with the highest success ratio is the FT.

Conclusions

- Furthermore, since SpaceX started rocket launches in 2010, The landing success rate continued to rise reaching it peak in 2017. With this success SpaceX rocket launches will continue to be cheaper.
- Among several features of the first stage of Falcon 9 which greatly affected it landing outcomes were the payload mass. Higher payload of above 5000kg led to a very low success rate. The range of payload with the highest success rate is between 1000kg to 4000kg. This means that lower payload masses leads to higher success rate
- Several classification models were built through a model pipeline using the GridSearchCV. Among all these models the best model was Decision Tree Classifier which was able to accurately classify 15 out of 18 records of the evaluation Data. This model had an accuracy score of about 88 percent on the training data set.

Appendix

- Success Rate: is the proportion of successful landing outcomes. Higher proportion of 1's from the 'CLASS' column of the dataset used means higher success rate and 0's otherwise.
- You can view the code and other projects processes from the repository link

Thank you!