APPLIED DATA SCIENCE CAPSTONE PROJECT

JANANI DHANDAYUTHAPANI

FEBRUARY 9 ,2023

OUTLINE

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

EXECUTIVE SUMMARY

- In this capstone, we will predict if the Falcon 9 first stage will land successfully. 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.
- We use various steps such as:
  - Data Collection,
  - Data wrangling,
  - Exploratory Data Analysis using SQL ,
  - Exploratory analysis using Pandas and Matplotlib,
  - Data Visualization with Folium,
  - Interactive Dashboard with Plotly Dash,
  - Predictive analysis using Machine learning models.
- After creating many models ,we conclude that the Decision tree is the best fit to predict the success or failure of the Falcon 9 Launch mission.

INTRODUCTION

- In the era of commercial space travel, SpaceX is among the most successful companies . This is possible because SpaceX makes rocket launches relatively inexpensive compared to others like Blue origin or Rocket lab.
- SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards 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.
- In order to determine the price of each launch we will gather information, create dashboards for our team and train machine learning models to predict if SpaceX will reuse the first stage.
- WILL THE FIRST STAGE OF THE SPACEX FALCON 9 ROCKET LAUNCH LAND SUCCESSFULLY ?

METHODOLOGY

- The overall methodology includes:
- Data collection, wrangling, and formatting, using:
  - SpaceX API
  - Web scraping
- Exploratory data analysis (EDA), using:
  - Pandas and NumPy
  - SQL
- Data visualization, using:
  - Matplotlib and Seaborn
  - Folium
  - Dash
- Machine learning prediction, using
  - Logistic regression
  - Support vector machine (SVM)
  - Decision tree
  - K-nearest neighbors (KNN)

RESULTS Data Collection and Data Wrangling Methodology

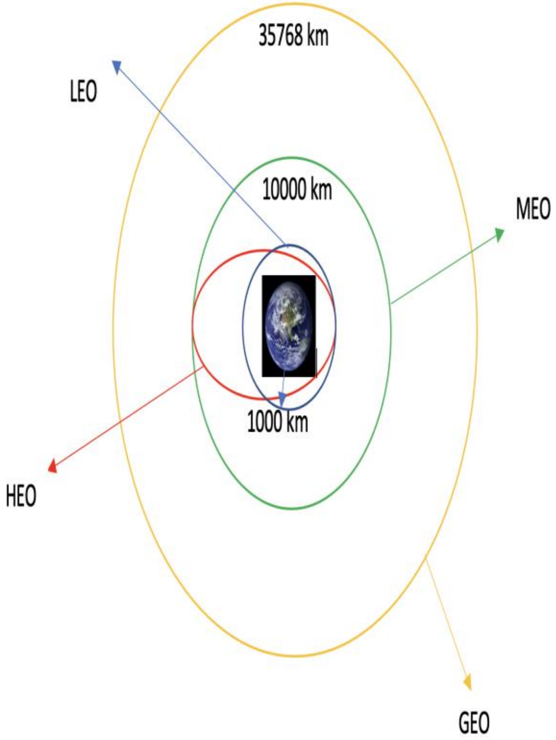
- **Libraries used : NumPy , Pandas, requests and Datetime**
- **We make a get request to the SpaceX API and then clean the data retrieved.**
- **We use the following URL to access the SpaceX launch Data :**
- **SpaceX API**
- **Steps taken:**
- **- we request and parse the SpaceX launch data using the GET REQUEST**
- **- we use json_normalize method to convert the Json file into a data frame**
- **-we get information from the data such as booster name, payload mass, orbit, name of launch site, latitude and longitude. We also learn about the type of landing, number of flights with a particular core ,etc.**
- **- we convert the data into a data frame**
- **- We filter the data frame to include only Falcon 9 launches**
- **- We replace missing values with the mean**

WEBSCRAPING

- Libraries used : sys, Pandas, unicodedata , re and BeautifulSoup from bs4
- We extract a Falcon9 historical launch data contained in a html table from wikipedia.
- We use the following page to scrape the data
- List Of Falcon9 and Falcon Heavy Launches
- We parse the table and load it into a pandas data frame.
- We have 121 rows and 11 columns of Falcon9 data only that we then clean up so that missing entries are replaced with mean and categorical data are encoded with one hot encoding.

DATA WRANGLING

- We performed exploratory data analysis and determined the training labels.
- We calculate the percentage of missing values and identify which columns are numerical or categorical.
- We calculated the number of launch sites, number of launches at each site using the value_counts() method.
- Each launch aims for a dedicated orbit. Some of the most common are shown in the picture.
- We created landing outcome label from outcome column and export the results to csv.

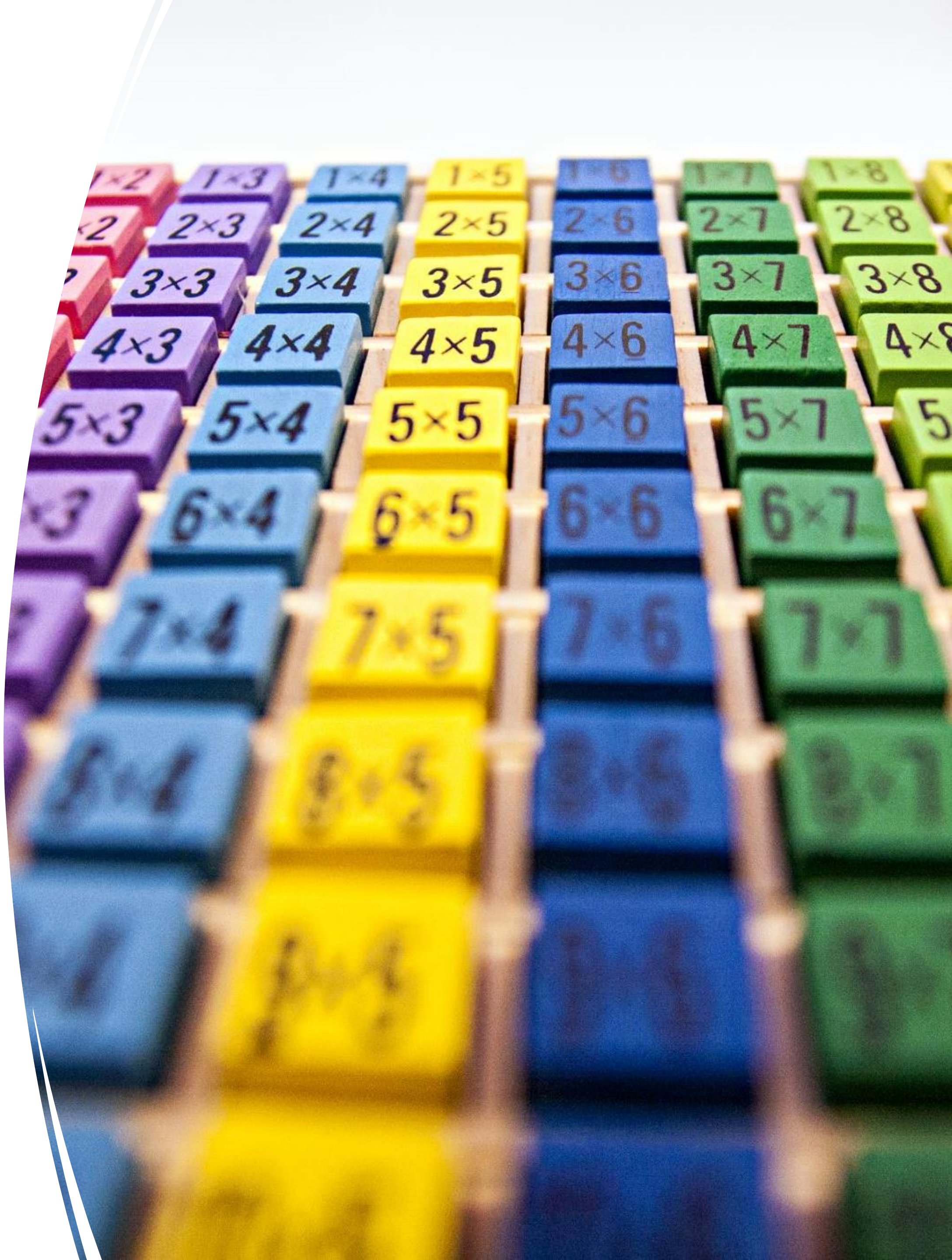


EDA WITH SQL

- We download the SpaceX dataset as a csv file and store it in a db2 instance on the cloud.
- We then establish a connection with the database and explore the data using SQL queries to gather information.
- We find unique launch sites in the space mission.
- We also calculate the total payload mass carried by boosters launched by NASA.
- The first successful landing outcome in the ground pad happened on 01-03 –2013.
- WE also find the booster versions that carried maximum payload mass.

EDA with Python libraries

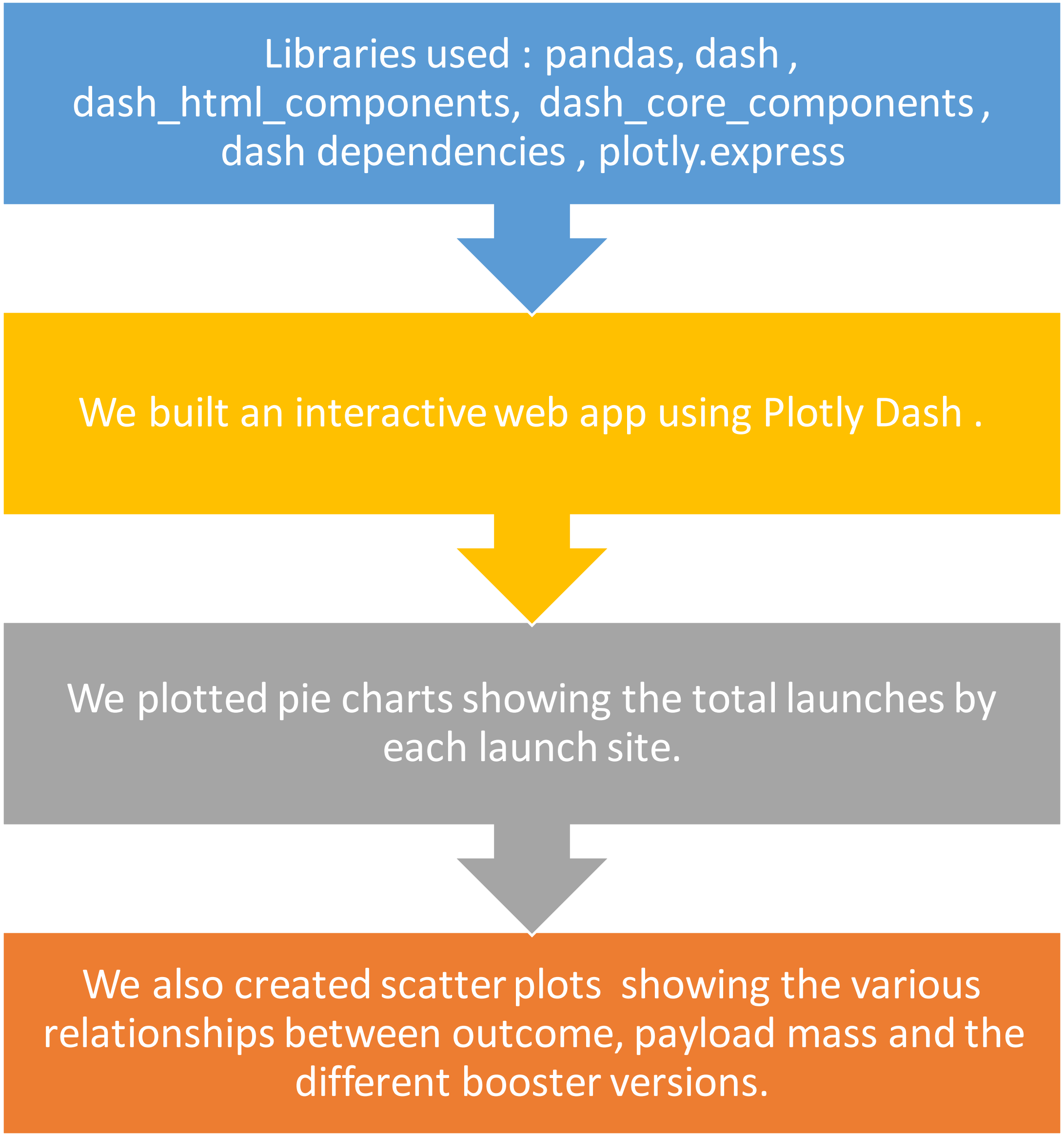
- Libraries used : pandas, NumPy, matplotlib and seaborn.
- First, we read the SpaceX dataset into a pandas data frame.
- Then we use seaborn and pyplot (the plotting interface of matplotlib) to visualize the relationship between various metrics.



Interactive Maps with Folium

- We used the Folium package to mark all the launch sites on a map with the latitude and longitude coordinates.
- We created a map object with the initial coordinates to be NASA Johnson Space center at Houston, Texas.
- Folium.Circle was used to create a highlighted circle area with text label around launch sites.
- If a launch was successful (i.e. if column 'class' = 1) we assigned a green marker and if a launch was a failure ('class' = 0) , a red marker was assigned.Thus, we know how many successful launches were done in each site.
- The distance of the launch sites from cities, railways, Highways and coastline were determined .
- From the Folium Maps Visualization, we found that the launch sites were in close proximity to roadways, railways and the coastline. But they were at a distance of around 50 km from the major city areas.

BUILD A DASHBOARD WITH PLOTLY DASH



Libraries used : pandas, dash , dash_html_components, dash_core_components , dash dependencies , plotly.express

We built an interactive web app using Plotly Dash .

We plotted pie charts showing the total launches by each launch site.

We also created scatter plots showing the various relationships between outcome, payload mass and the different booster versions.

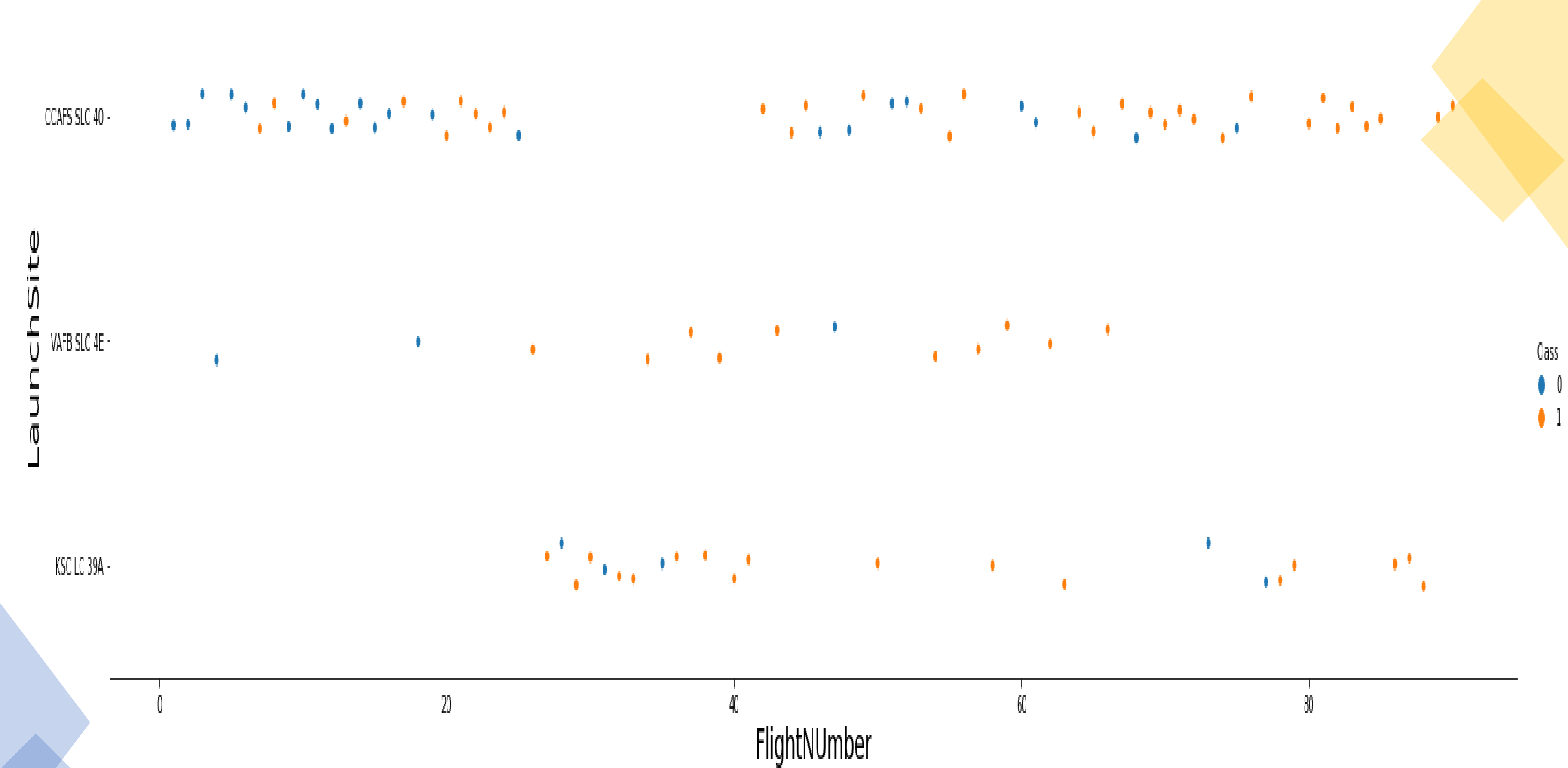
PREDICTIVE ANALYSIS – CLASSIFICATION

We load the data into pandas , transformed it into s suitable data frame format , replaced missing values with the mean and then split the data into testing and training data.

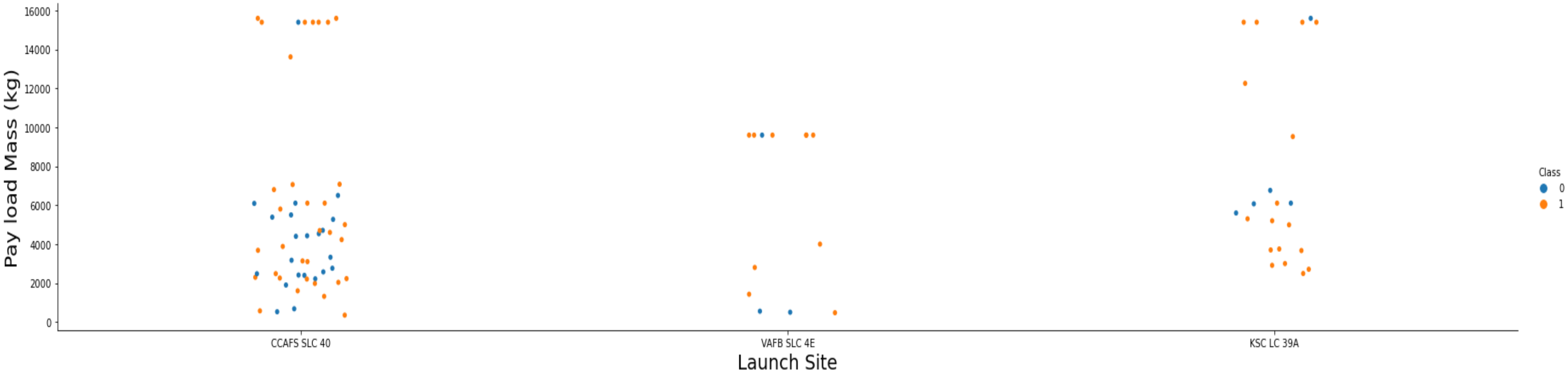
We trained different machine learning models and then adjusted the parameters (hyperparameter tuning) using GridSearchCv.

After fitting the various models, Decision tree algorithms were found to have greater accuracy based on out of sample data predictions.

Flight Number Vs Launch site

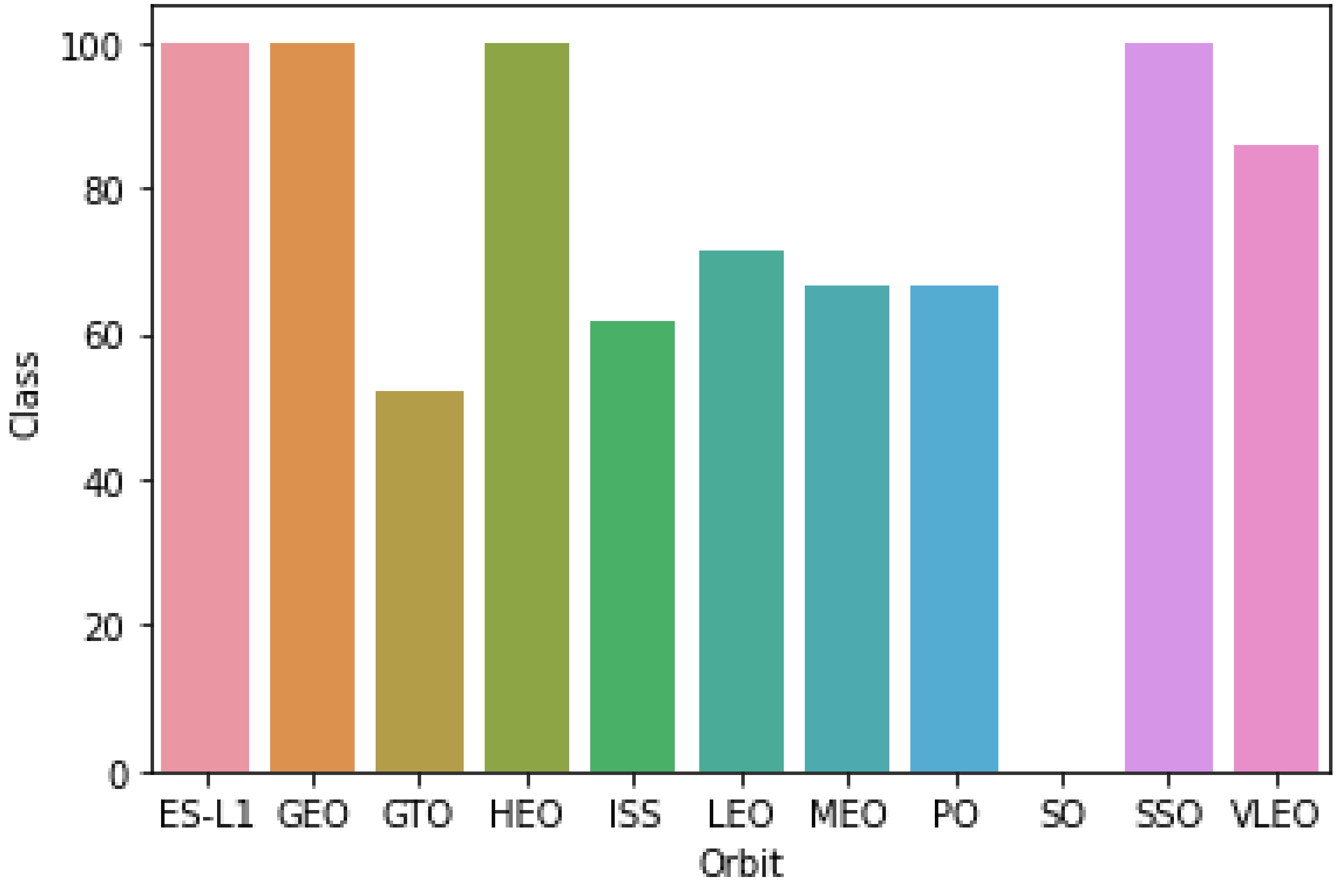


Payload Vs Launch site



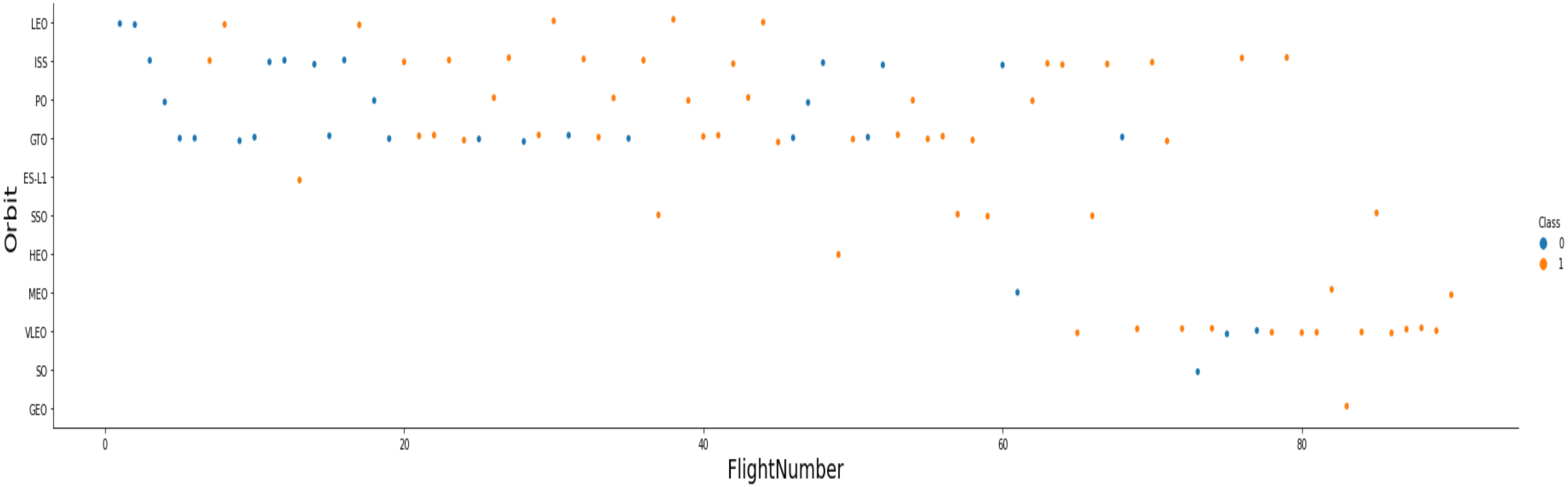
From the scatter plot, it is obvious that the greater the payload mass for the launch site CCAFS SLC40 the higher the success rate of the rocket.

Success rate Vs Orbit Type
ES-L1,GEO,HEO,SSO, VLEO has the high success rates.



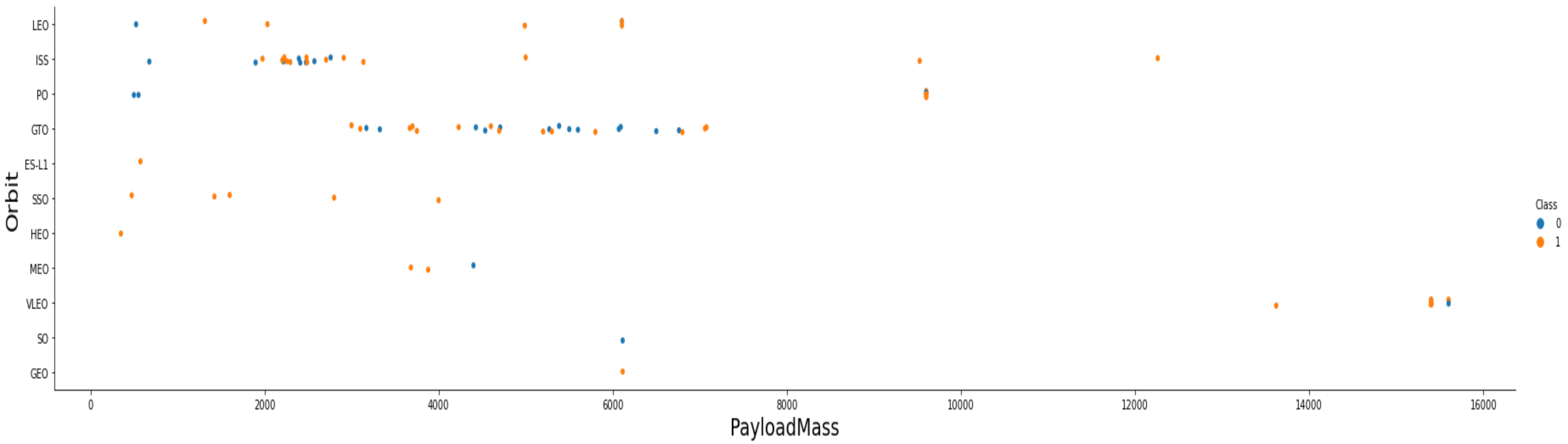
Flight Number Vs Orbit Type

- We observe that in the LEO orbit, success is related to the number of flights whereas in GTO there is no such relationship



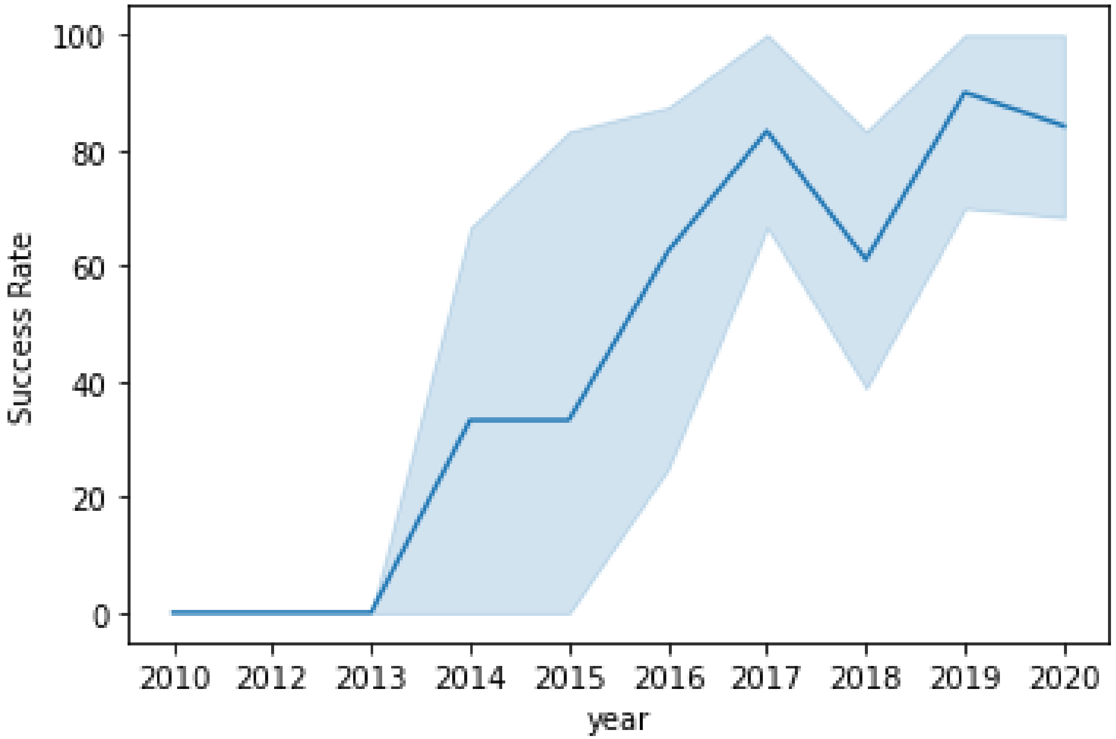
Payload mass vs Orbit

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

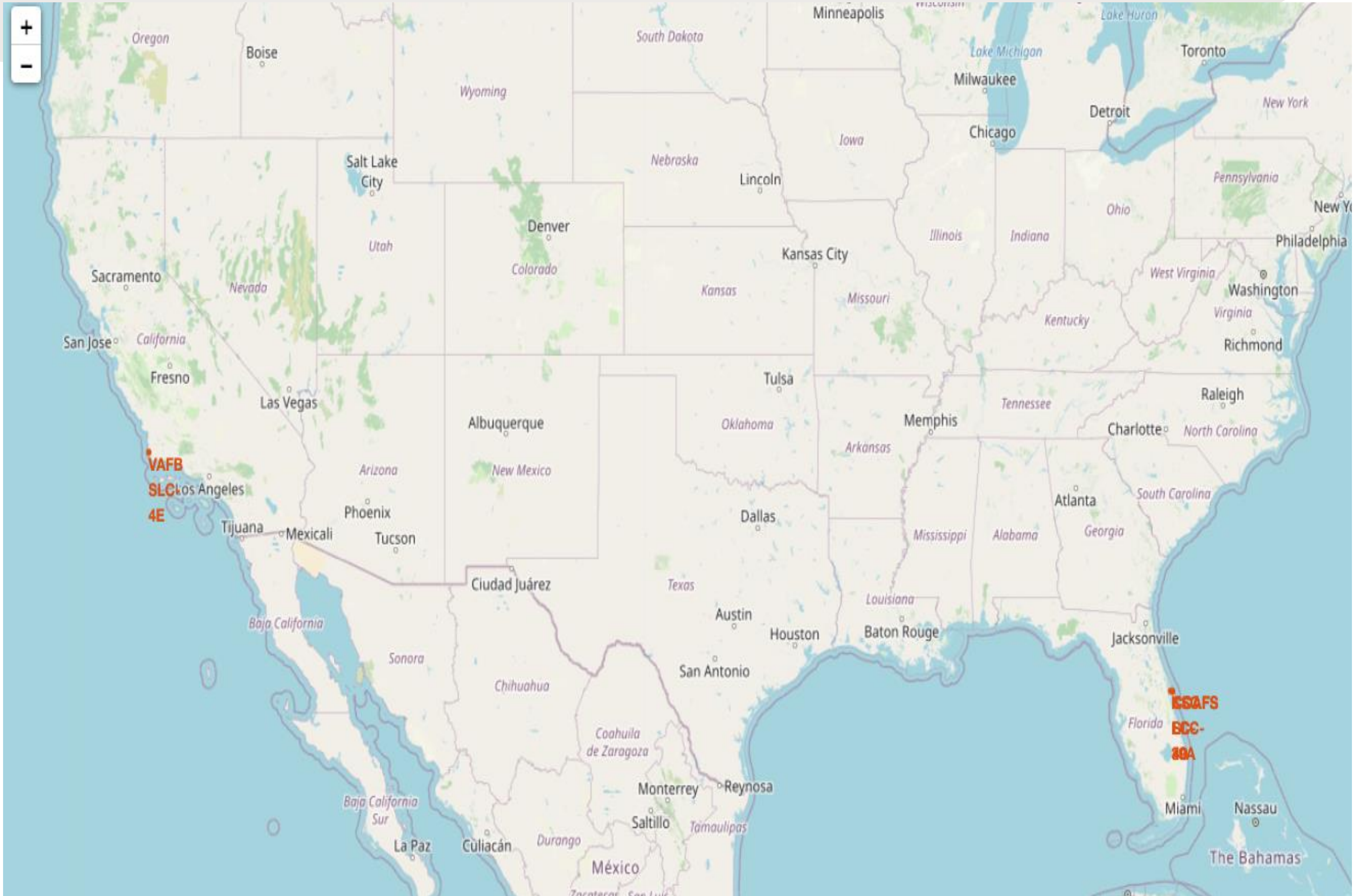


Launch Success Yearly Trend

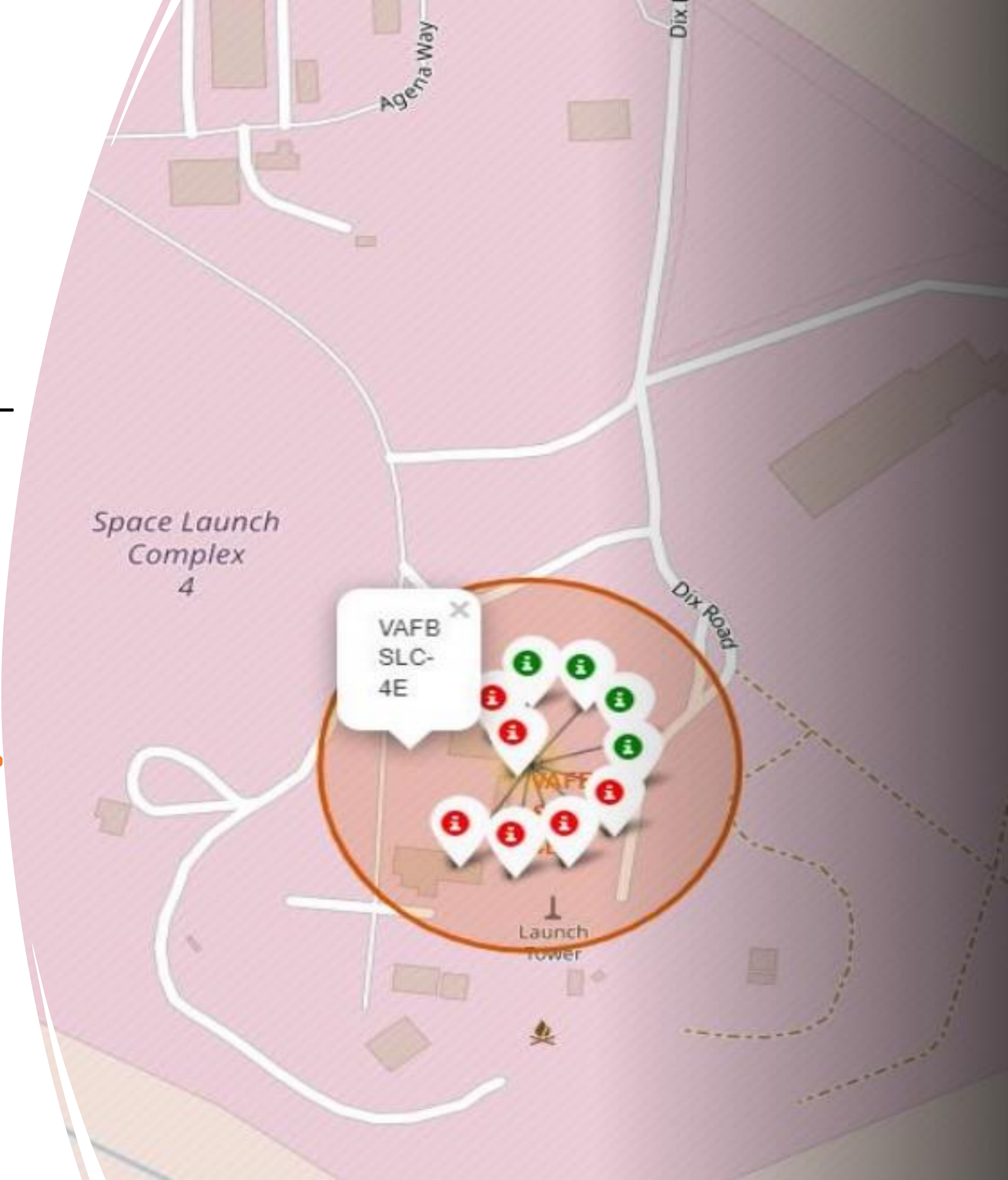
We can observe that the success rate since 2013 kept increasing till 2020.

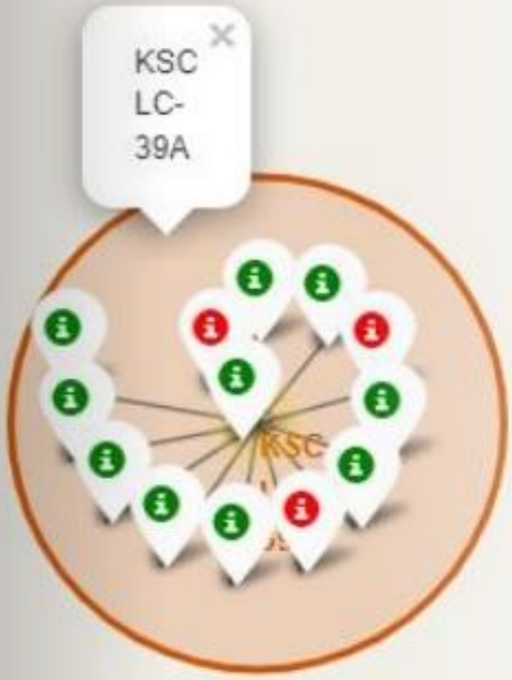


Map showing All Launch sites

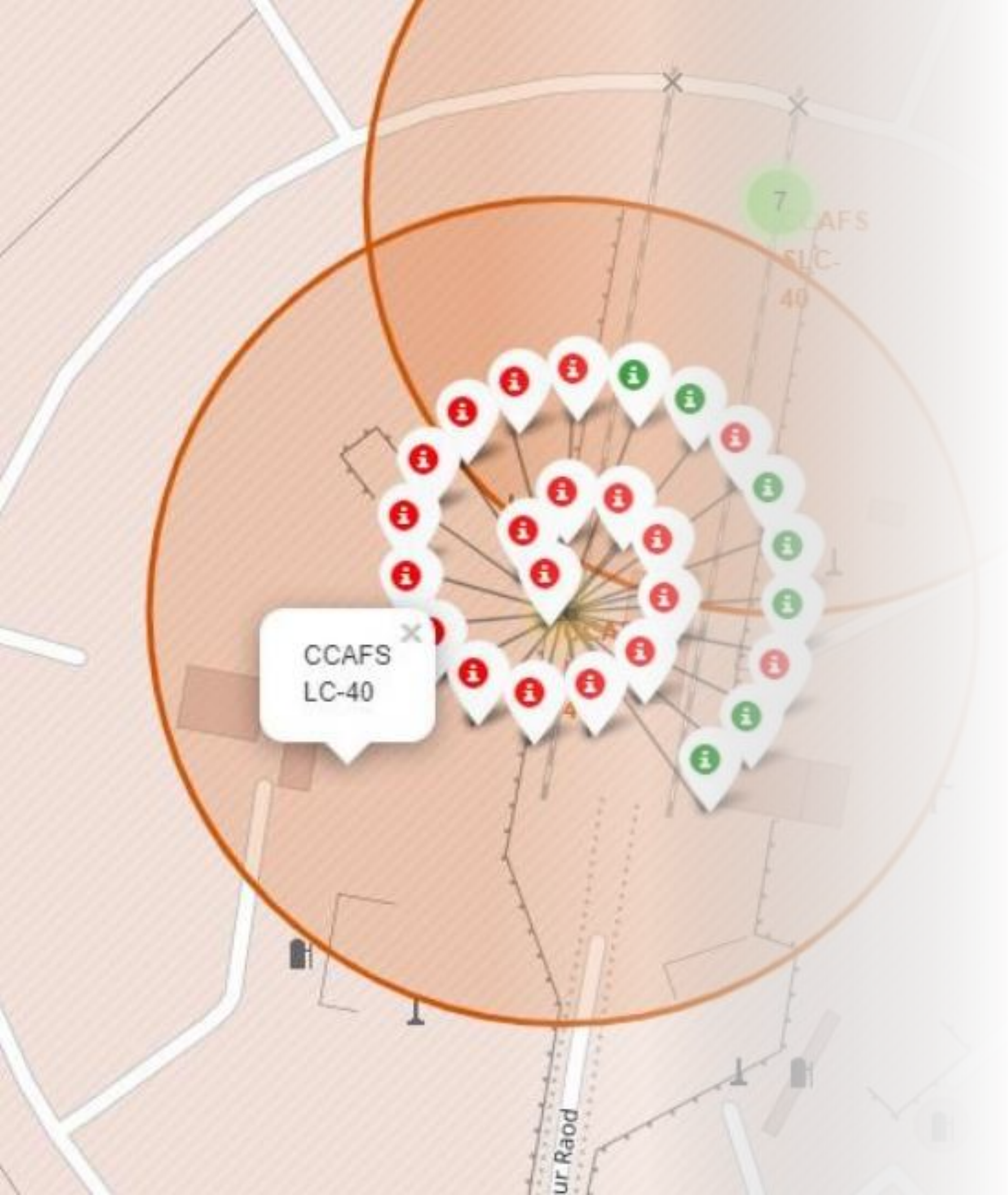


Out of a total of 10 launches at SpaceX VAFB SLC-4E launch site, ^ failed and 4 rockets landed successfully.

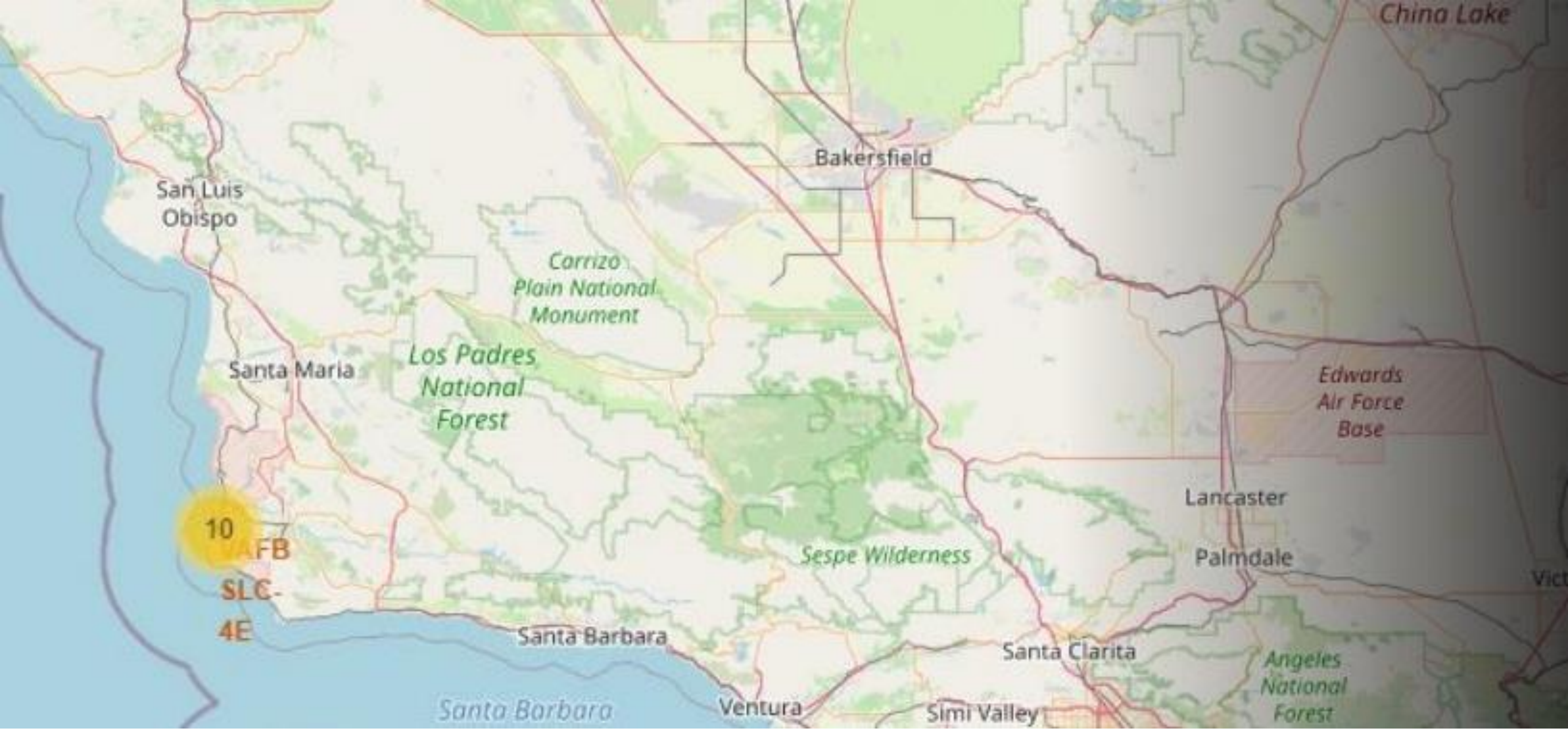




Out of a total of 14 launches at the SpaceX KSC LC39A Launch Site Merritt Island, 10 rockets landed and 4 failed to launch.



Out of a total of 26 launches at the SpaceX CCAFS LC_40 launch site, Only 7 landed and the other 19 rockets failed to land.

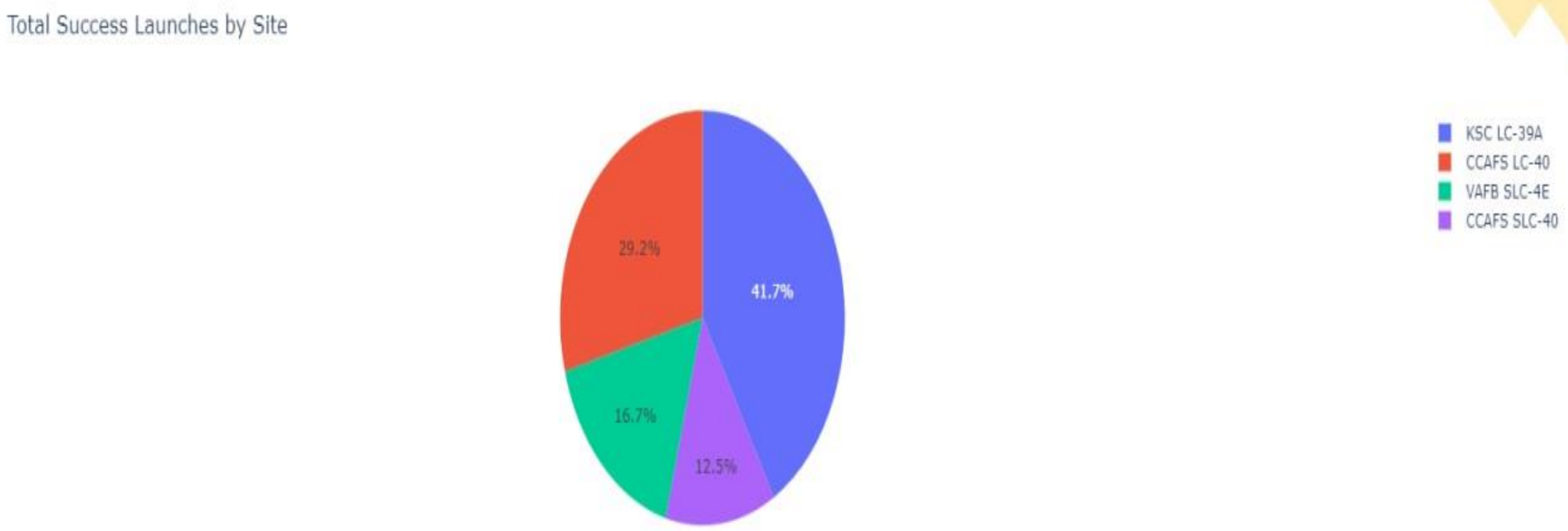


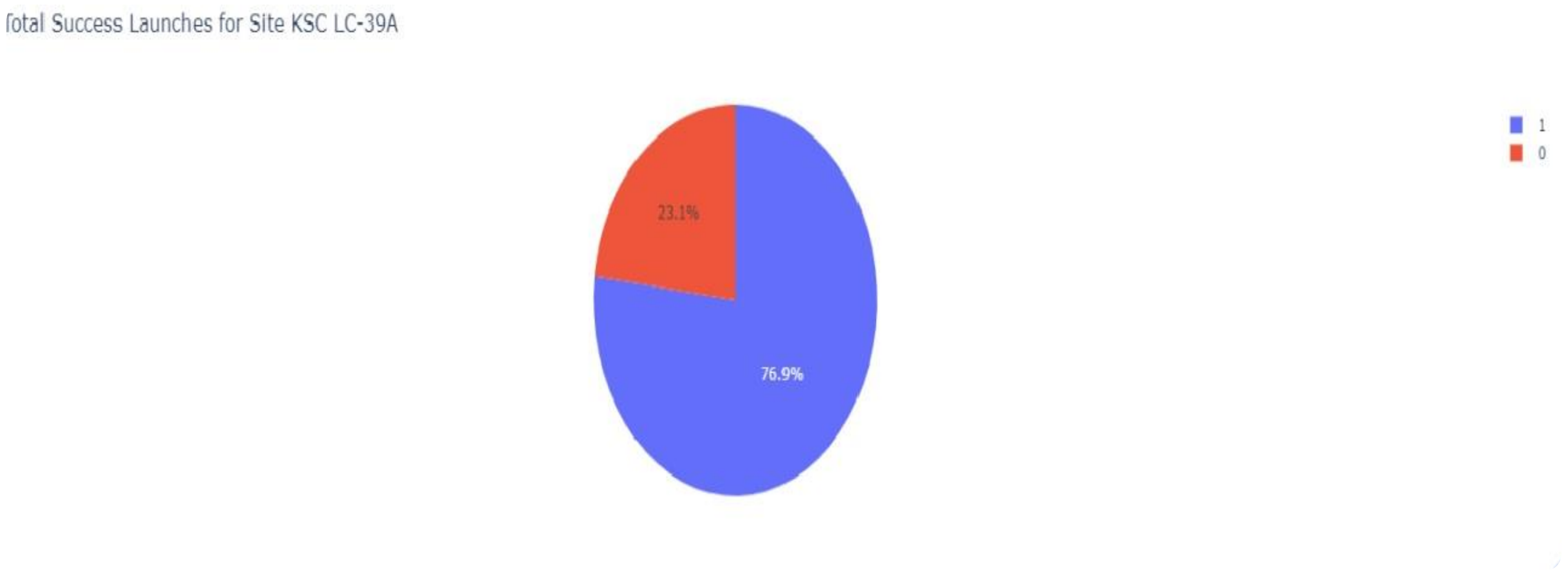
SPACEX VAFB SLC – 4E Launch Site

- The cluster markers indicate that 10 Falcon 9 launches have taken place in the launch site stationed off the coast of Santa Maria

Total Success Launches by Site

- The KSC LC-39A Launch site accounts for the largest percentage of the total number of successful landings at 41.7%



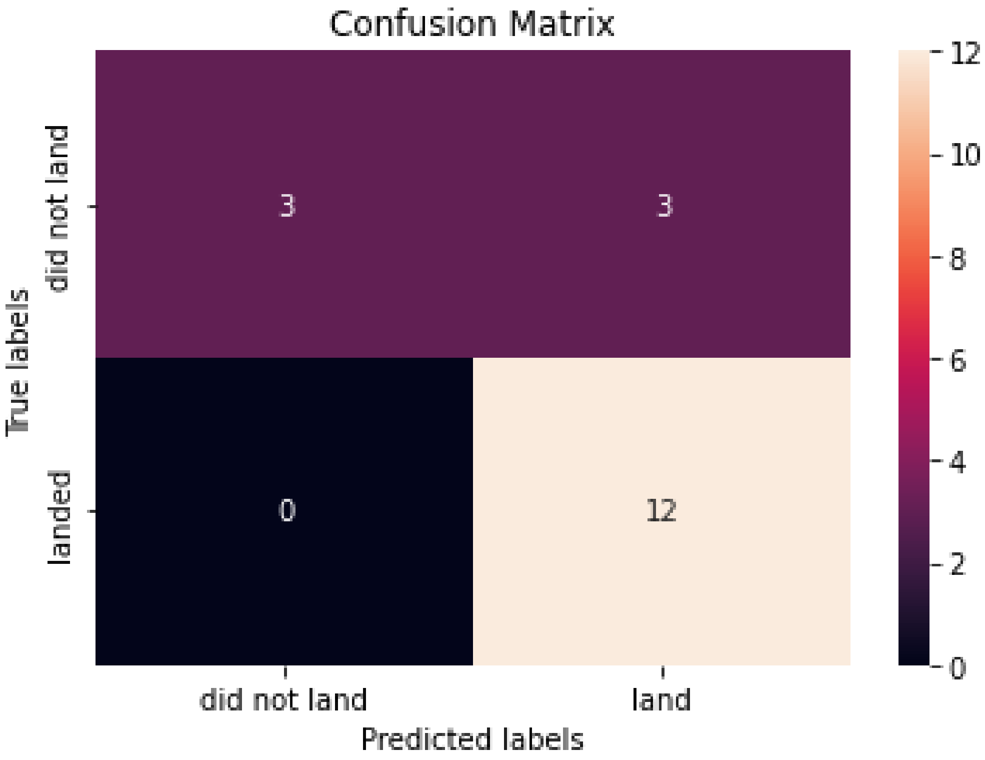


Launch Site with the highest probability of success

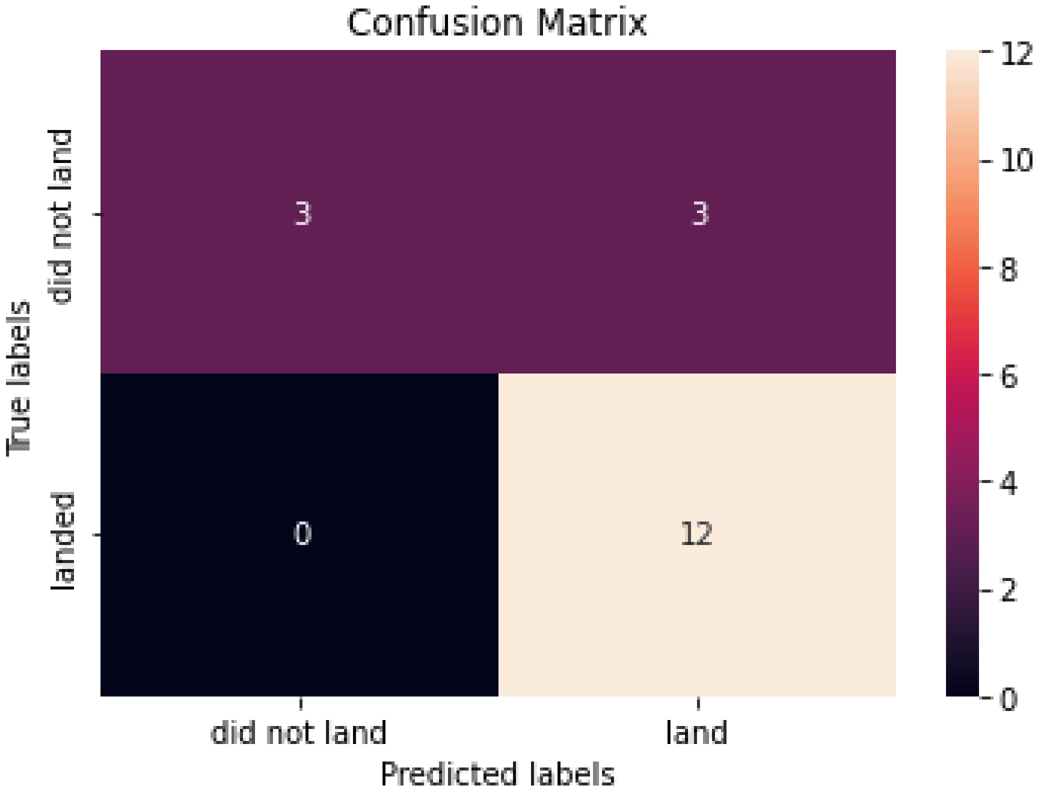
- The KSC LC-39A Launch Site also has the highest probability of success per launch
- 76.9% of all launches at the KSC LC-39A Site Land Successfully
- 23.1% of all launches at the KSC LC-39A Site Fail to Land

PREDICTIVE ANALYSIS- SUPPORT VECTOR MACHINES

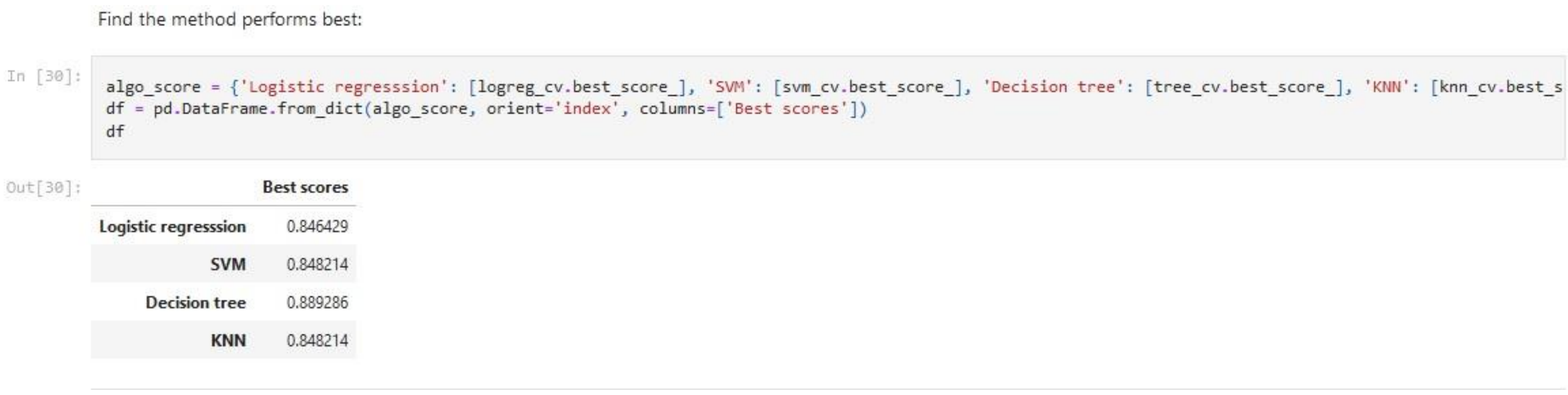
- The SVM algorithm had an accuracy of 83.33%



PREDICTIVE ANALYSIS- DECISION TREE CLASSIFIER



COMPARISON OF VARIOUS ALGORITHMS



Find the method performs best:

In [30]:

```
algo_score = {'Logistic regresssion': [logreg_cv.best_score_], 'SVM': [svm_cv.best_score_], 'Decision tree': [tree_cv.best_score_], 'KNN': [knn_cv.best_s
df = pd.DataFrame.from_dict(algo_score, orient='index', columns=['Best scores'])
df
```

Out[30]:

	Best scores
Logistic regresssion	0.846429
SVM	0.848214
Decision tree	0.889286
KNN	0.848214

CONCLUSION

- The probability of successful landings has increased every year however in recent years (2018-2020) the yearly rate of increase has declined with a maximum probability of success being about 0.65 in 2020.
- Rockets with a low payload mass are more likely to land successfully than the ones with a heavier payload mass.
- We were able to build a decision tree model that can predict the probability of Falcon 9 Rocket Stage 1 Landing Successfully with an 94.44% accuracy on our out of sample data and 87.5% accuracy on in sample data.
- The KSC LC-39A Launch Site has the highest probability of success per launch .
- The type of orbit required for the launch has an impact on the landing success, the ES-L1, SSO, HEO, and GEO orbits have the highest rate of success for landing .