IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

HENRY GYARTENG-MENSAH
04/12/2021

Outline

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

Executive Summary

- The research seek to predict the outcome of a spaceX booster landing successfully or not at the first stage of launch. To perform various analysis, the data was collected using spaceX's API. The API gives data with various attributes that can be used for the prediction of the landing or otherwise of the FALCON 9 launch. The collected data from is then passed into a pandas data frame for analysis. Various exploratory data analytics, predictive analytics and visualizations were performed during the research. From the analysis it was evident that, launch site VAFB SLC 4E, there rate of success for rockets launched for payloads 1000kg to 10000kg is very high. the orbit types: SSO,HEO,GEO and ES-11 has the highest success rate of landing with a mean class of 1. Whiles rockets in the orbit SO has a higher bad outcome of landing.

Introduction

- Project background and context

SpaceX saves the cost of using a rocket through it's rocket which can be reused. Other competitors spend as much as USD 165 million for a rocket launch whiles SpaceX spends as low as USD 62 million due to its Falcon 9 rocket reusability at the first stage.

- Problems you want to find answers

The research seeks to predict how successful Falcon 9 rocket launches landed at various sites during their first stage. This will be achieved using various exploratory data analysis, predictive analytics and visualizations.

Section 1

Methodology

Methodology

Executive Summary

- **Data collection:**
  - Data was collected with using the SpaceX's REST API. The API gives data with various attributes that can be used for the prediction of the landing or otherwise of the FALCON 9 launch. The collected data from is then passed into a pandas data frame for analysis.
- **Data wrangling**
  - During this stage, status of a booster landing successful or not were coded as 1 or 0 respectively.
- The researcher also performed an exploratory data analysis (EDA) using visualization and SQL and also performed interactive visual analytics using Folium and Plotly Dash.The researcher then performed a predictive analytics by using various machine learning techniques (SVM, Classification Trees,KNN and logistic regression). The data to be used by the model were split into train and test data to test the best hyperparameters.

Data Collection

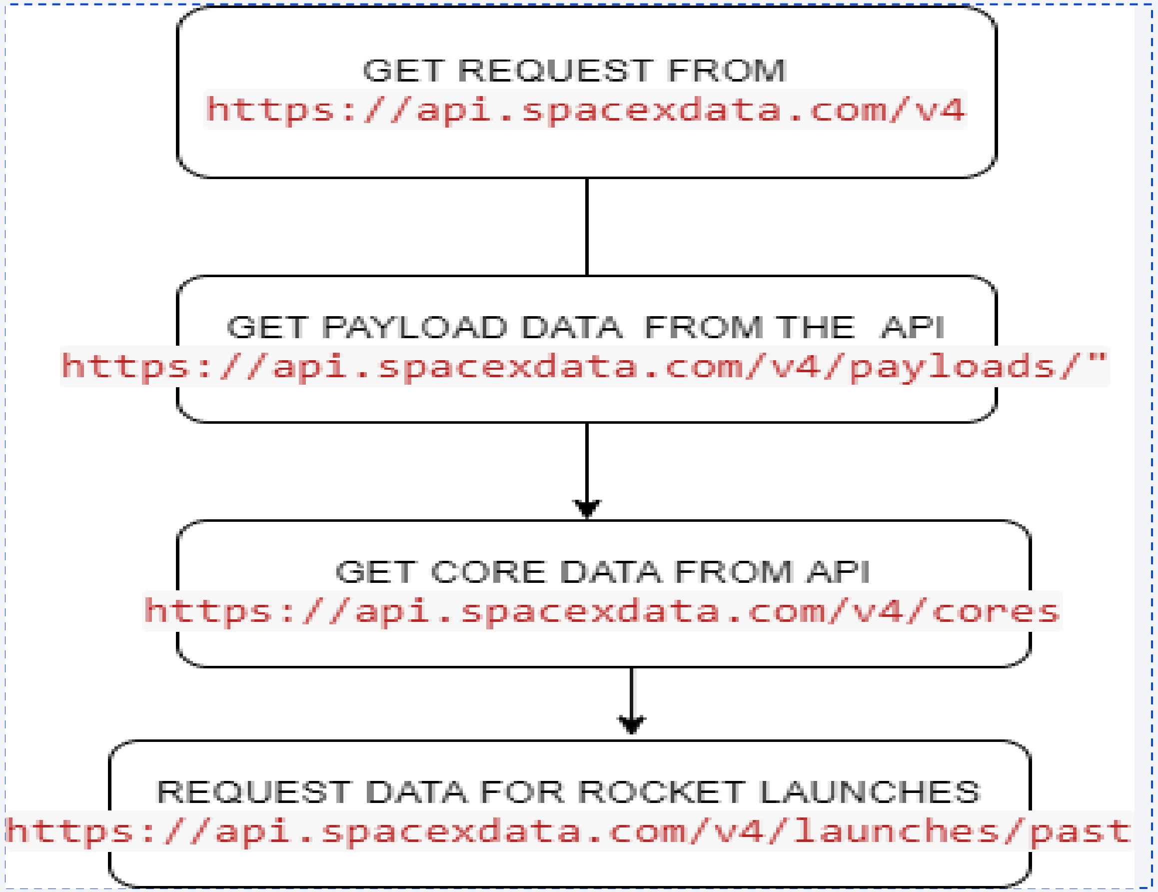
- Describe how data sets were collected.
- Data was collected with using the SpaceX’s REST API. The API gives data with various attributes that can be used for the prediction of the landing or otherwise of the FALCON 9 launch. The collected data from is then passed into a pandas data frame for analysis. The flowchart of the various API calls are given in the section below:
- The data collection process are presented using flowcharts in the slides below:

Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts

- Add the GitHub URL of the completed SpaceX API calls notebook (must include completed code cell and outcome cell), as an external reference and peer-review purpose

- Github URL: https://github.com/henalytics/Capstone/blob/master/SPACEX_DATA_COLLECTION.ipynb

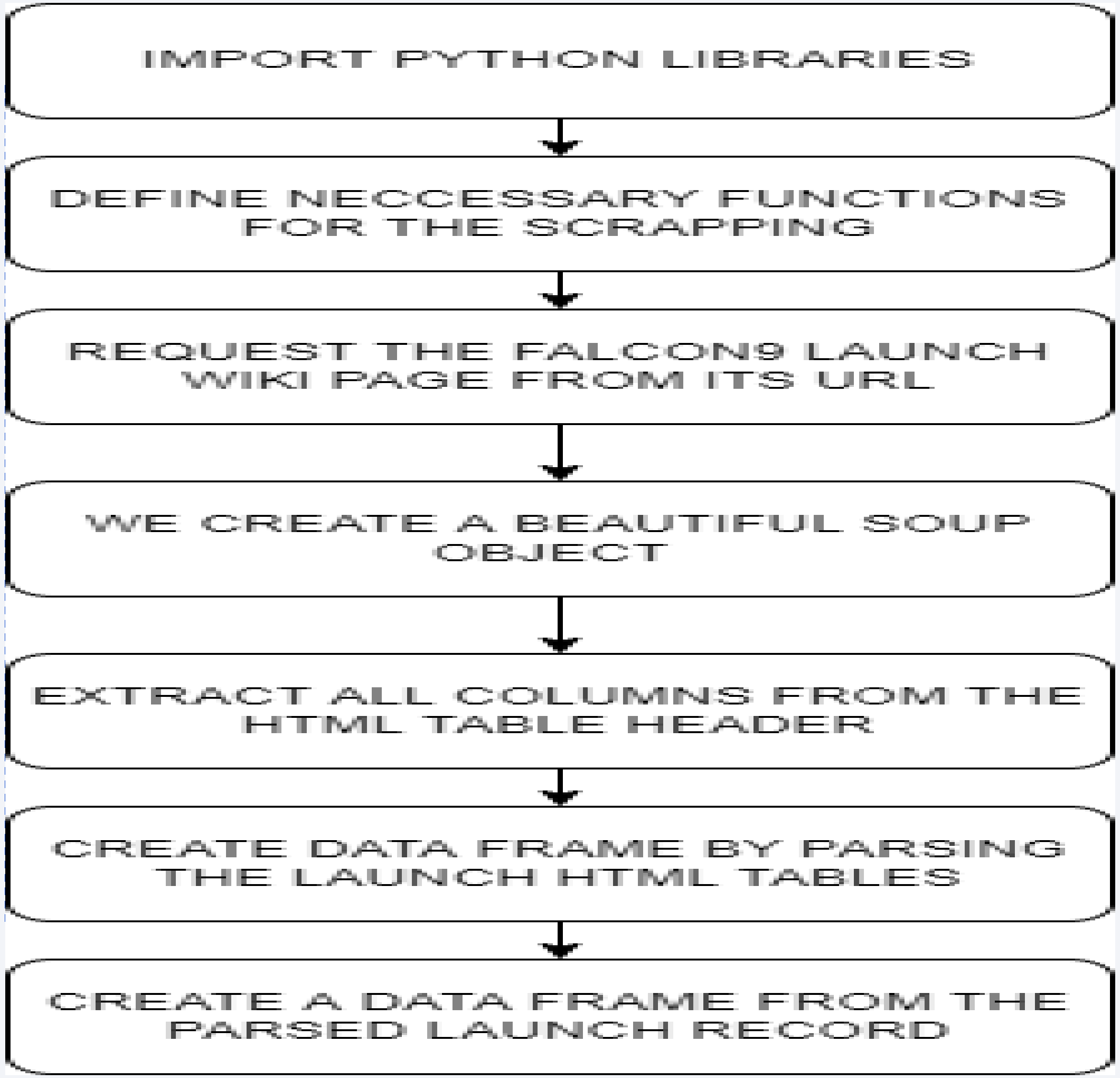


Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts

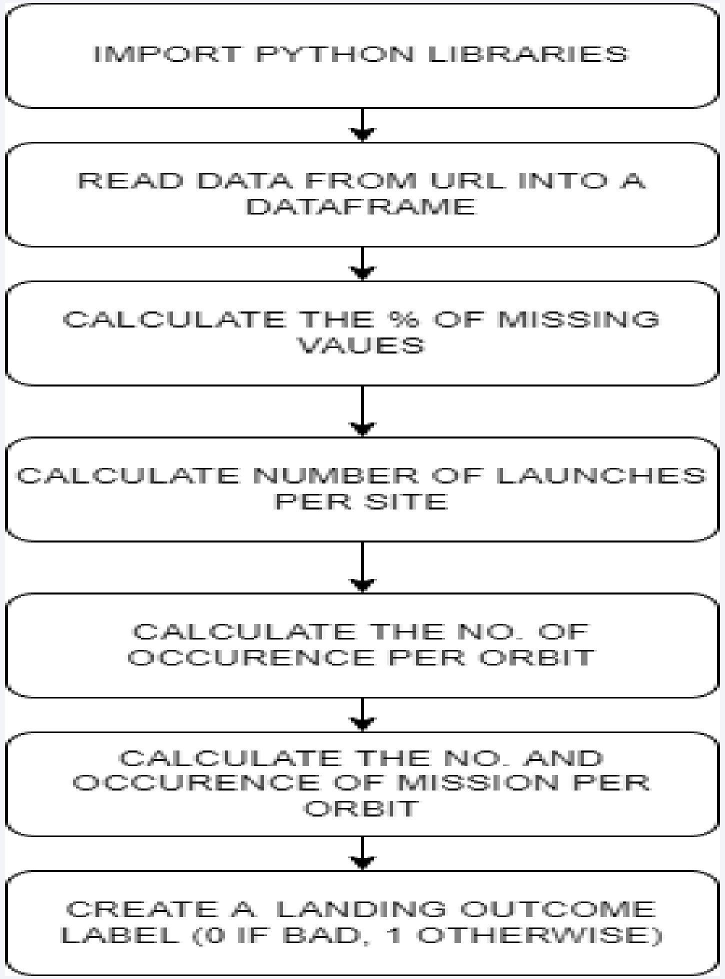
- Add the GitHub URL of the completed web scraping notebook, as an external reference and peer-review purpose

- URL: https://github.com/henalytics/Capstone/blob/master/WEBSCRAPPING.ipynb



Data Wrangling

- At the data wrangling stage, the following process were followed:
- The required python libraries needed for wrangling are imported.
- Data is the loaded inro a dataframe for necessary wrangling.
- The number of missing values in the dataframe are calculated
- The number of launches for each sites are calculated
- The number of occurrence for each orbit is also calculated.
- The mission outcome is also calculated. The various outcomes are grouped into bad outcomes and good outcomes
- The bad outcomes are labelled 0 and 1 otherwise. This is then assigned to the variable class and loaded into the dataframe.
- https://github.com/henalytics/Capstone/blob/master/EXPLORATORY_DATA_ANALYSIS.ipynb



EDA with Data Visualization

- To achieve the goal of the study, some exploratory data analysis and visualizations were performed. The Charts used at this stage are given below:
- Scatter plot (Catplot) : This was used to determine how the variables: Flight number and payload mass' effect on the landing outcome. The relationship between flight number and launch site were also determined .
- Bar Chart: This was also used to determine which orbits have the highest success rate.
- Line Chart: It is used to get average success trend over the years.
- Url for the EDA and data visualization: Capstone/MATPLOTLIB_EDA.ipynb at master · henalytics/Capstone (github.com)

EDA with SQL

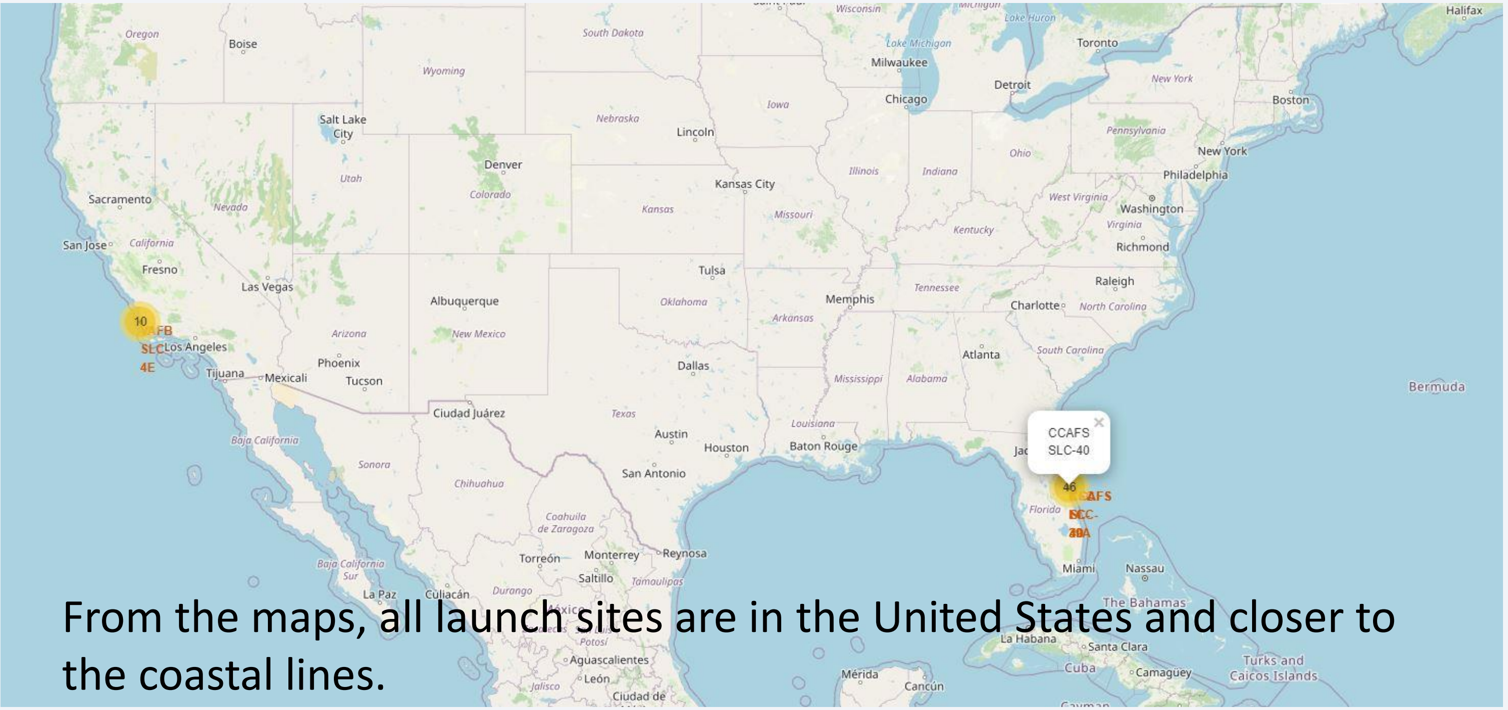
- The following are the SQL queries used”
- select distinct Launch_Site from SPACEXTBL (to display the names of the unique launch sites)
- select Launch_Site from SPACEXTBL where Launch_Site like 'CCA%' limit 5 (Display 5 records where launch sites begin with the string ‘CCA)
- select Customer,sum(PAYLOAD_MASS__KG_) as PAYLOAD from SPACEXTBL where Customer='NASA (CRS)' GROUP BY Customer; (the total payload mass carried by boosters launched by NASA(CRS).
- select Booster_Version,avg(PAYLOAD_MASS__KG_) as avgpayload from SPACEXTBL where Booster_Version='F9 v1.1' GROUP BY Booster_Version; (Average payload carried by booster F9 v1.1)
- SELECT landing__outcome,MIN(Date) as date_min from SPACEXTBL where landing__outcome='Success (ground pad)' GROUP BY landing__outcome; (first date a successful ground landing outcome was achieved)

EDA with SQL

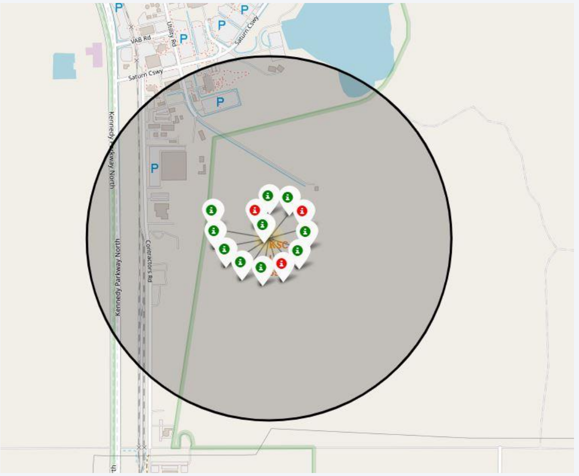
- The following are the SQL queries used”
- select Booster_Version,PAYLOAD_MASS__KG_ from SPACEXTBL where landing__outcome='Success (drone ship)' and PAYLOAD_MASS__KG_>4000 and PAYLOAD_MASS__KG_<6000 (boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000)
- select Mission_Outcome,count(Mission_Outcome) as total from SPACEXTBL GROUP BY Mission_Outcome; (the total number of successful and failure outcomes)
- select Booster_Version,PAYLOAD_MASS__KG_ from SPACEXTBL where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
- select Booster_Version,Launch_Site from SPACEXTBL where landing__outcome='Failure (drone ship)' and extract (YEAR from Date)='2015’
- select date,landing__outcome,count(landing__outcome) as countlan from SPACEXTBL where Date between '2010-06-04' AND '2017-03-20' and landing__outcome in ('Failure (drone ship)','Success (ground pad)') group by date,landing__outcome order by date desc
- select unique(Launch_Site) from SPACEXTBL

Reference: https://github.com/henalytics/Capstone/blob/master/EDA_SQL.ipynb

Build an Interactive Map with Folium



From the maps, all launch sites are in the United States and closer to the coastal lines.



Green Markers shows successful landings

Red Markers shows otherwise

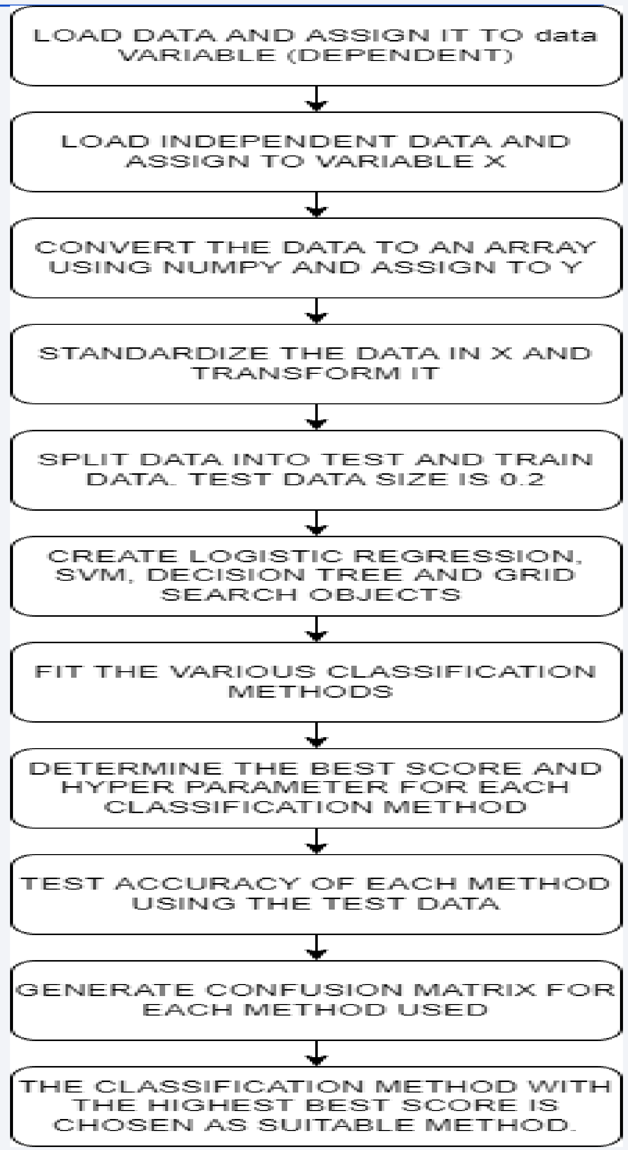
- Reference: [Capstone/INTERACTIVE DASHBOARD.ipynb at master · henalytics/Capstone (github.com)]

Build a Dashboard with Plotly Dash

- An interactive Dashboard consisting of the following were developed:
- The dashboard has a dropdown for the various launch sites, a pie chart that gives the distribution of success launch sites, a slider which selects the various payload(kg), a scatter plot which gives the relationship between the payload and landing success.
- A change in the slider range causes a change in the scatter plot that is linked.
- Again, the options selected at the dropdown interacts with its corresponding pie chart. The reference to the dashboard is give below:
- Reference: https://github.com/henalytics/Capstone/blob/master/spacex_dash_app.py

Predictive Analysis (Classification)

- The predictive analysis started with creating an array of the dependent variable and assigning it to a variable. The independent variables were also standardized. The data was later split into test and train dataset with a test size of 20%. A logistic regression,support vector machines, decision tree and KNN objects were created using the GridSearchCV object. The training data was then fitted to determine the best parameters for each classification model with their corresponding accuracy. The test data set was later used to determine the accuracy of each model. In determining the best performing classification model, the model with the highest accuracy was selected.
- The flowchart is given below:
- Reference: https://github.com/henalytics/Capstone/blob/master/PREDICTIVE_ANALYTICS_SPACEX%20(1).ipynb



Results

- **EXPLORATORY DATA ANALYSIS**

From various exploratory data analysis, the research revealed that for all sites, flights above 20 has a higher success rate. In view of that, the rate of success increase with increasing number of flights. For all sites, there is a higher success rate for rockets launched for payloads greater than 8000kg. For launch site VAFB SLC 4E, there rate of success for rockets launched for payloads 1000kg to 10000kg is very high. Again, the orbit types: SSO,HEO,GEO and ES-11 has the highest success rate of landing with a mean class of 1. Whiles rockets in the orbit SO has a higher bad outcome of landing. In the LEO orbit the Success rate appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit. With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS. But for GTO it is difficult to distinguish since the number of positive and negative outcomes are fairly same. The rate of success for a booster landing started increasing after the year 2013. This rate of success kept increasing till the year 2020.

Results

- **INTERACTIVE ANALYTICS**

The following results were revealed from the interactive dashboard: KSC LC-39A site has the largest successful launches compared to the other sites. The proportion of success to other launch site is 41.7%. The total number of successful launches in the site KSC LC-39A was 76.9%. This means that for every launch, there is a 76.9% chance that the rocket will land successfully and a 23.1% chance that it will fail one way or the other. payload booster version FT has the highest success rate for a given payload range. The booster in the payload mass range 2000 to 6000 (kg) has a high success rate when compared to other versions.

- **PREDICTIVE ANALYSIS**

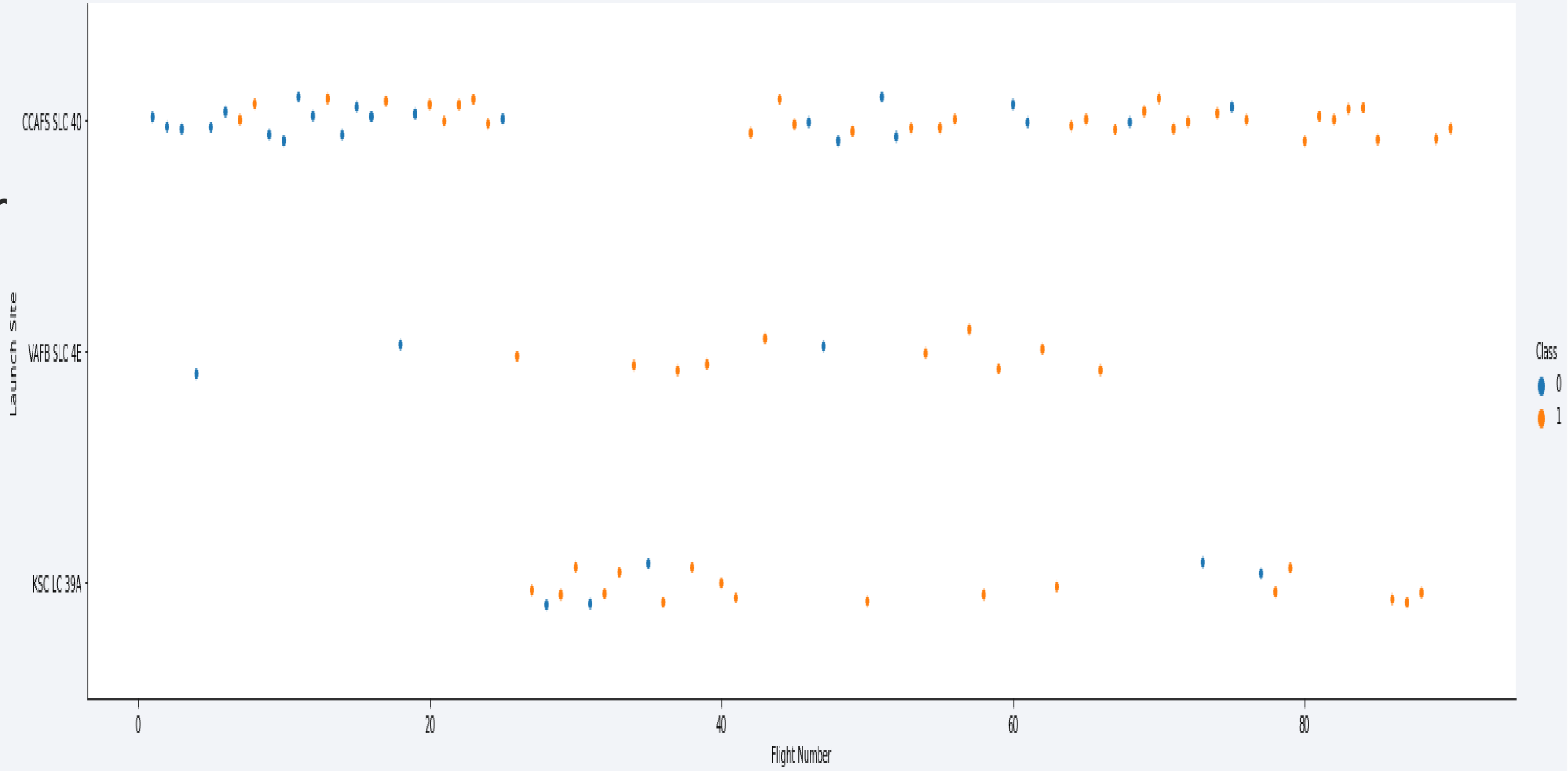
Testing the various classification models shows that, the decision tree classification algorithm performed best with a best score of 0.875 when compared to other classification methods (KNN,SVM,Logistic Regression). The accuracy score of the decision tree classifier was 0.944 when the test data was used to access the accuracy of the classifier.

Section 2

Insights drawn from EDA

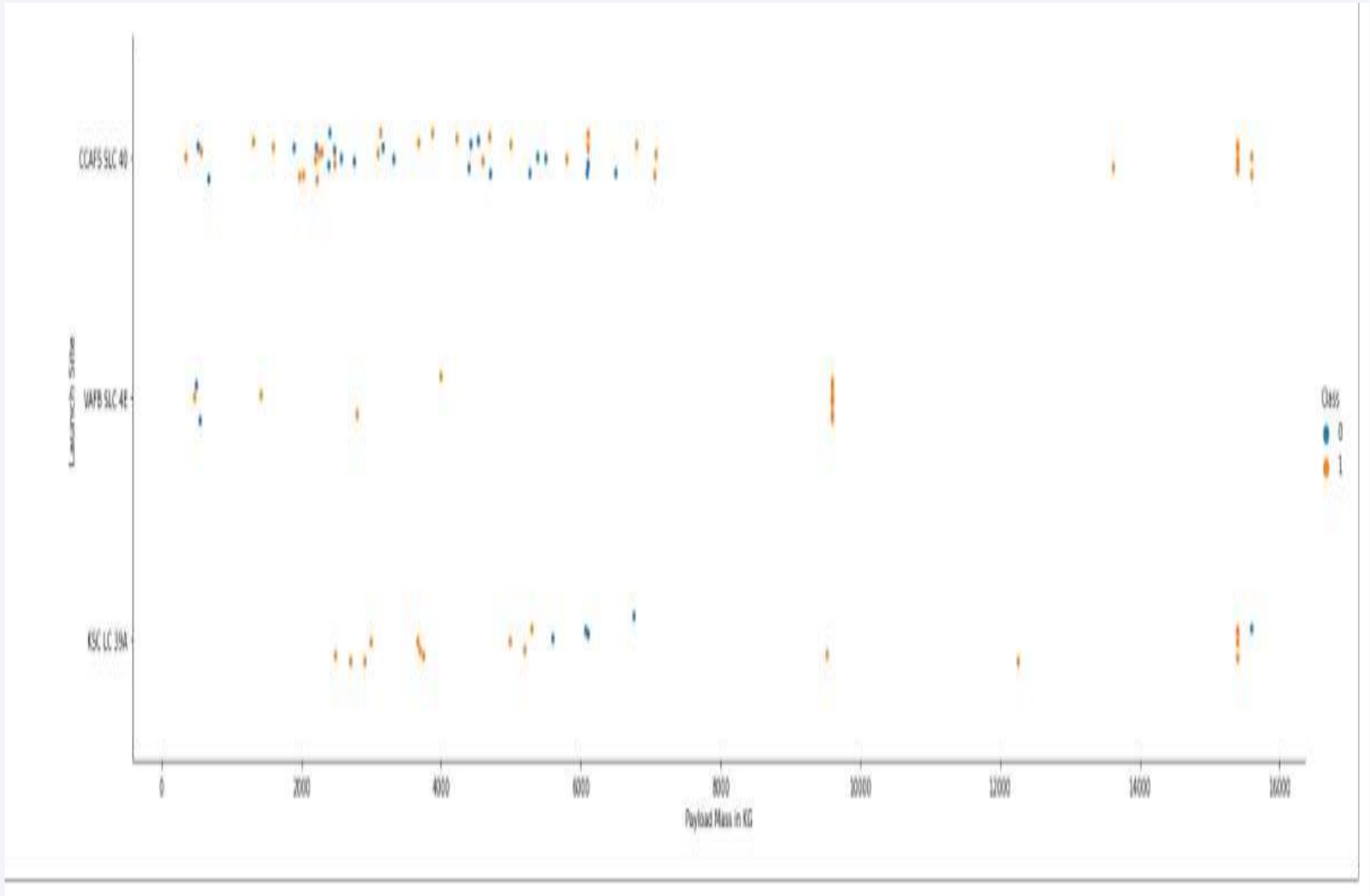
Flight Number vs. Launch Site

- From the graph, it can be seen that the for all sites, flights above 20 has a higher success rate. In view of that, the rate of success increase with increasing number of flights. That is there is a positive relationship with the number of flights and the rate of success.



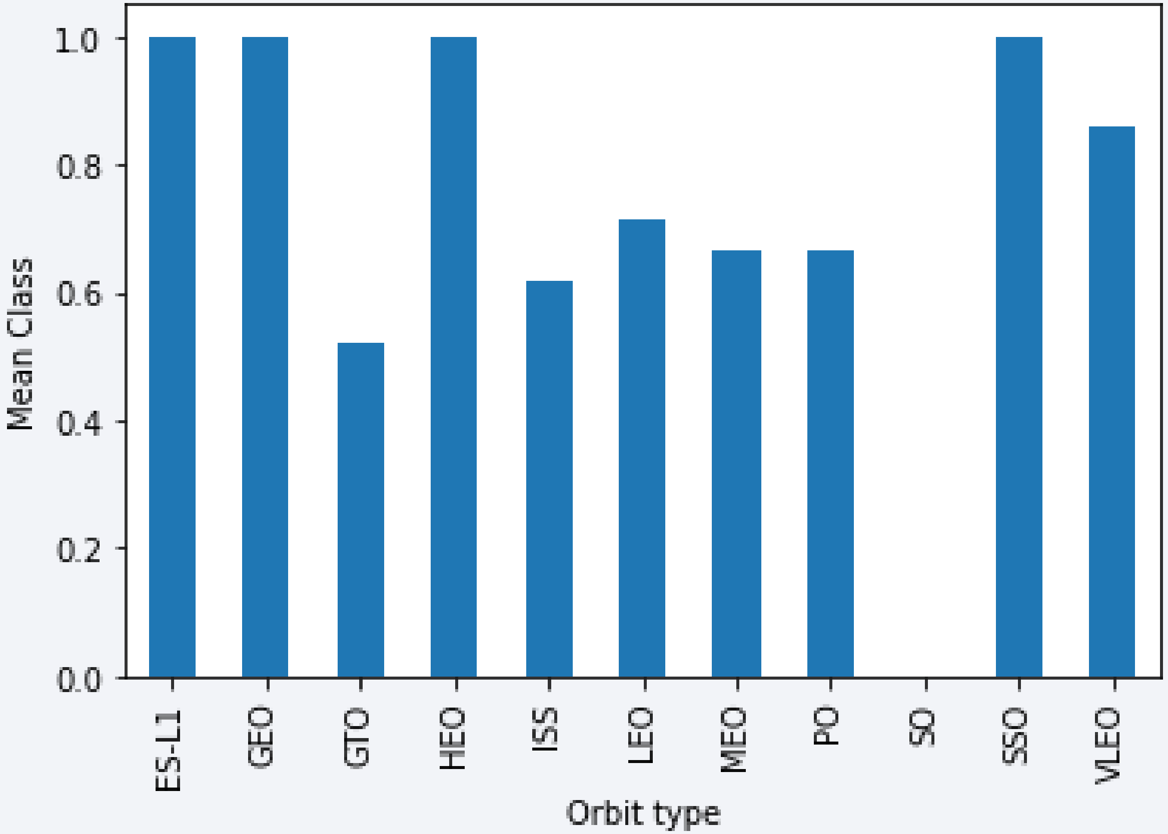
Payload vs. Launch Site

- For all sites, there is a higher success rate for rockets launched for payloads greater than 8000kg. For launch site VAFB SLC 4E, there rate of success for rockets launched for payloads 1000kg to 10000kg is very high.



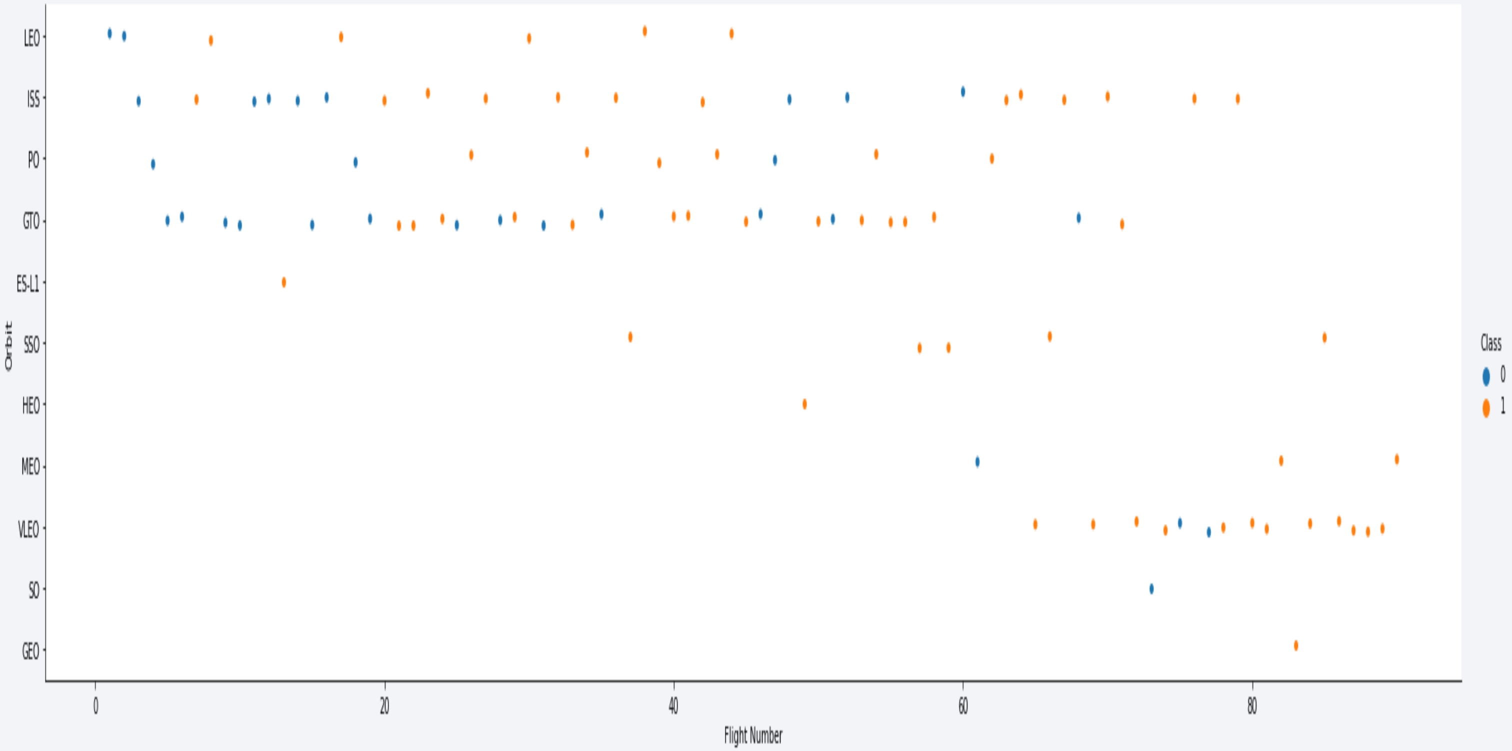
Success Rate vs. Orbit Type

- From the graph, it can be seen that rockets launched in the orbit types: SSO,HEO,GEO and ES-11 has the highest success rate of landing with a mean class of 1. Whiles rockets in the orbit SO has a higher bad outcome of landing.



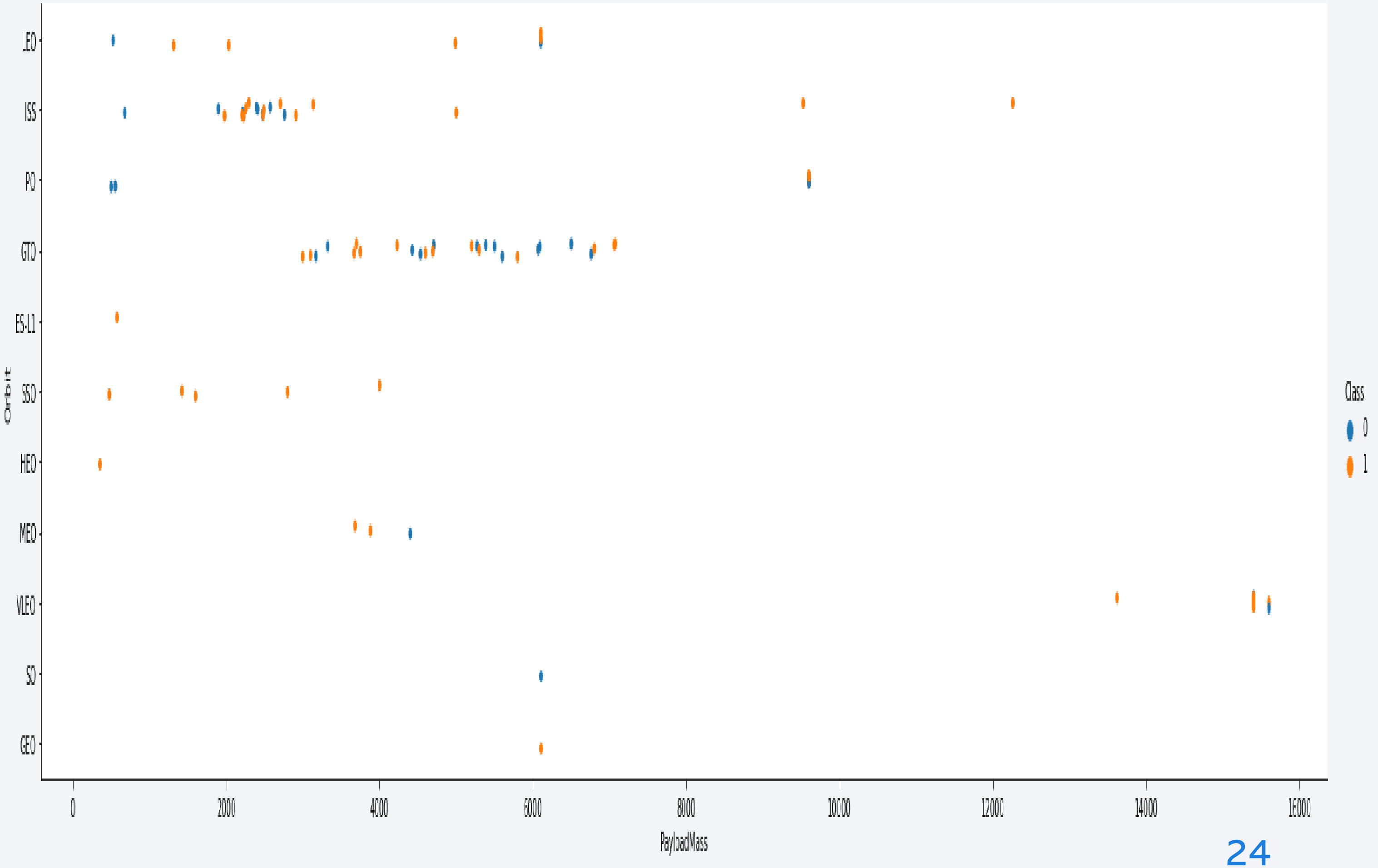
Flight Number vs. Orbit Type

- In the LEO orbit the Success rate appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.



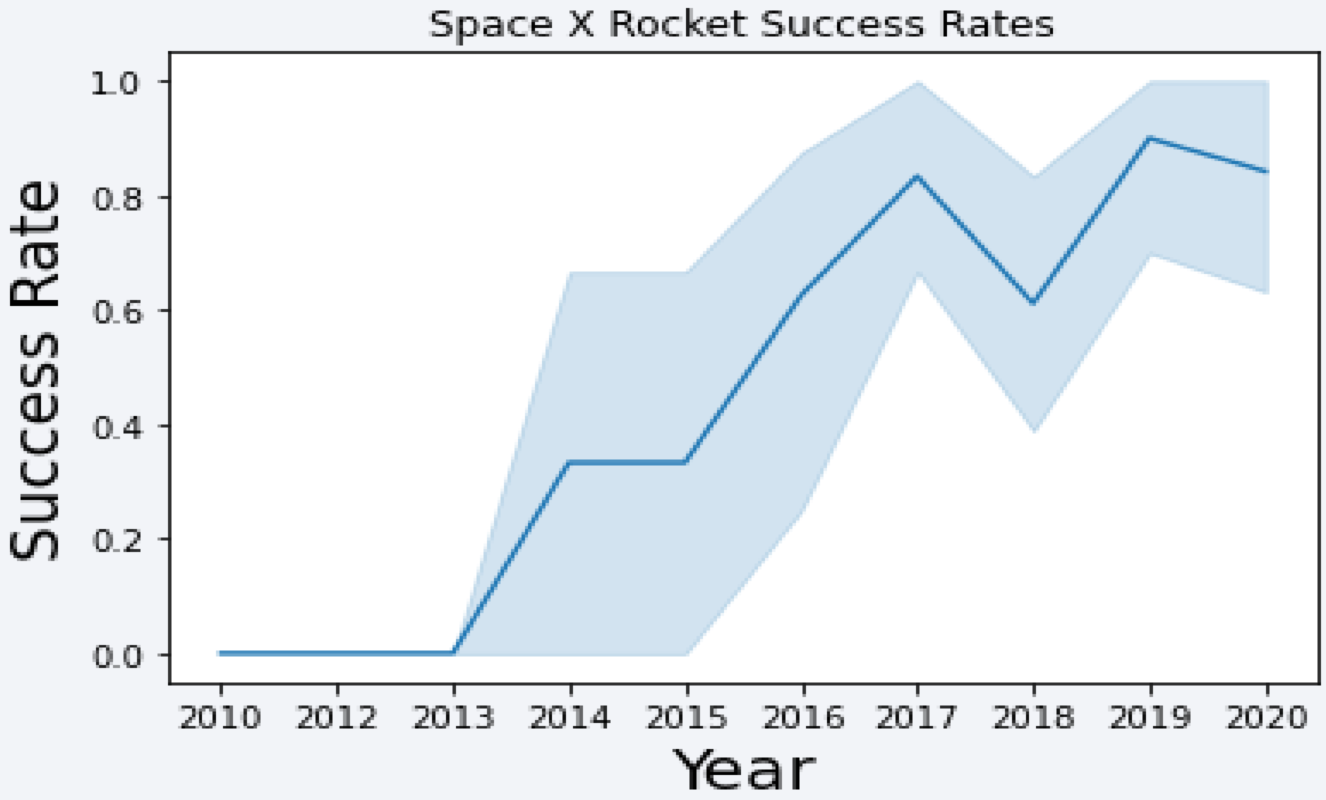
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS. But for GTO it is difficult to distinguish since the number of positive and negative outcomes are fairly same.



Launch Success Yearly Trend

- The rate of success for a booster landing started increasing after the year 2013. This rate of success kept increasing till the year 2020.



All Launch Site Names

- **QUERY:** select distinct Launch_Site from SPACEXTBL;
- This distinct statement enables the data scientist select all launch site names uniquely without any duplication.

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

Launch Site Names Begin with 'CCA'

- The five launch sites that begin with CCA was retrieved using the following query
- **QUERY:** select Launch_Site from SPACEXTBL where Launch_Site like 'CCA%' limit 5;
- The launch sites were selected from the table where each launch sites contains the keyword CCA.

launch_site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass carried by boosters launched by NASA (CRS)

- **QUERY:** select Customer,sum(PAYLOAD_MASS__KG_) as PAYLOAD from SPACEXTBL where Customer='NASA (CRS)' GROUP BY Customer;
- The above query calculates the sum of payload mass with NASA (CRS) as the customer. The results are given below :

customer	payload
NASA (CRS)	45596

Average Payload Mass by F9 v1.1

- **QUERY:** select Booster_Version,avg(PAYLOAD_MASS__KG_) as avgpayload from SPACEXTBL where Booster_Version='F9 v1.1' GROUP BY Booster_Version;
- The query above calculates the average payload mass carried by booster F9 v1.1 by grouping all booster versions. The average payload calculated is then assigned to the variable avgpayload.

booster_version	avgpayload
F9 v1.1	2928

First Successful Ground Landing Date

- **QUERY:** SELECT landing__outcome,MIN(Date) as date_min from SPACEXTBL where landing__outcome='Success (ground pad)' GROUP BY landing__outcome;
- The landing outcome and the first date (minimum date) from the date field were selected from the table and renamed as date_min for all landing outcomes that contains the keyword “Success (ground pad)”. The resulting outcome is then grouped by the landing outcome field. The result is given below:

landing__outcome	date_min
Success (ground pad)	2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- **QUERY:** select Booster_Version,PAYLOAD_MASS__KG_ from SPACEXTBL where landing__outcome='Success (drone ship)' and PAYLOAD_MASS__KG_>4000 and PAYLOAD_MASS__KG_<6000;
- The query selects the booster version, and payload with mass between 4000kg to 6000kg for boosters which have success landing on a drone ship. The results is shown below:

booster_version	payload_mass__kg_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

Total Number of Successful and Failure Mission Outcomes

- **QUERY:** select Mission_Outcome,count(Mission_Outcome) as total from SPACEXTBL GROUP BY Mission_Outcome;
- The query above counts the number of mission outcomes and then groups them as being successful or otherwise. The result is given below:

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

Boosters Carried Maximum Payload

- **QUERY:** select Booster_Version,PAYLOAD_MASS__KG_ from SPACEXTBL where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
- The query above selects the booster version,payload from the table with their mass equal to the maximum payload mass(Kg). The result is shown below:

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

- **QUERY:** select Booster_Version,Launch_Site from SPACEXTBL where landing__outcome='Failure (drone ship)' and extract (YEAR from Date)='2015'
- The query selects the booster version and launch site from the table for landing outcomes that were considered a failure on a drone ship for the year 2015. The year was extracted from the date field where the format was in dd/mm/yyyy. The result is given below:

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

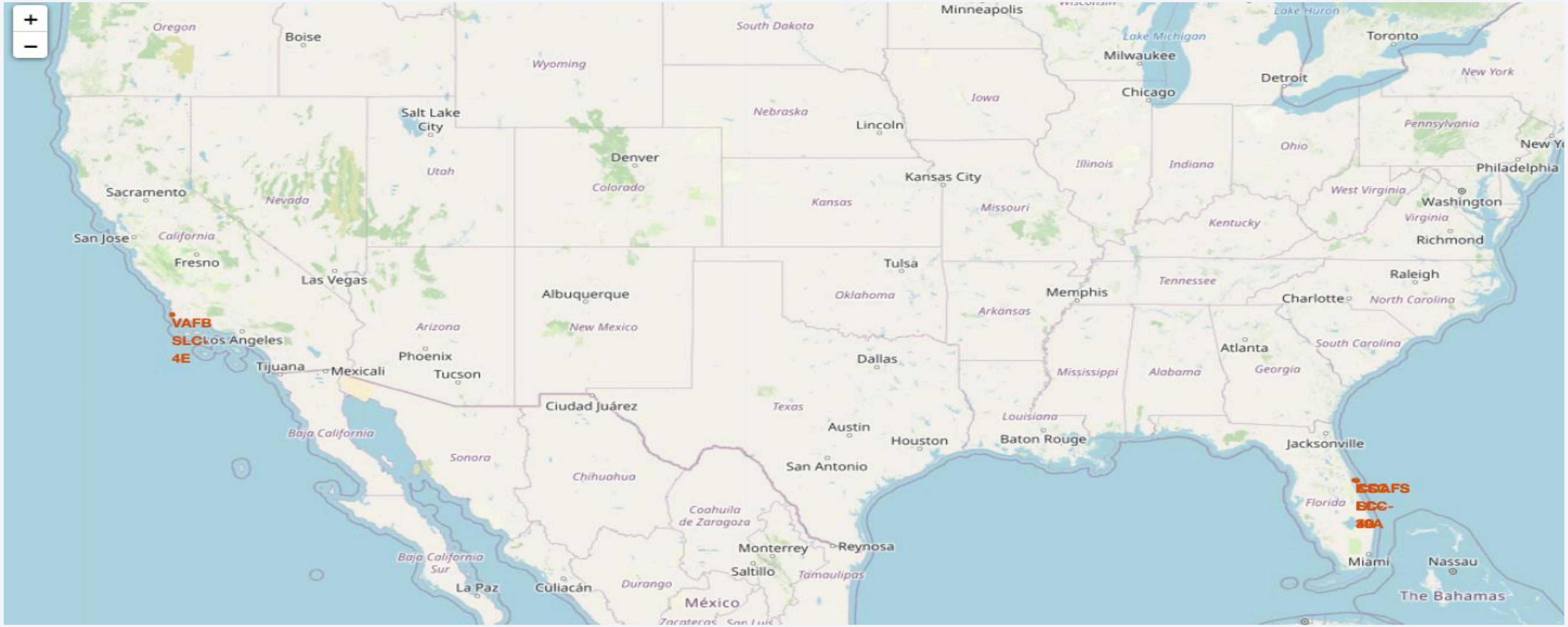
- **QUERY:** select date,landing__outcome,count(landing__outcome) as countlan from SPACEXTBL where Date between '2010-06-04' AND '2017-03-20' and landing__outcome in ('Failure (drone ship)','Success (ground pad)') group by date,landing__outcome order by date desc
- The above query gives the total landing outcomes for the period (04-06-2010 to 20-03-2017 for various landing areas for each date given.

DATE	landing__outcome	countlan
2017-02-19	Success (ground pad)	1
2016-07-18	Success (ground pad)	1
2016-06-15	Failure (drone ship)	1
2016-03-04	Failure (drone ship)	1
2016-01-17	Failure (drone ship)	1
2015-12-22	Success (ground pad)	1
2015-04-14	Failure (drone ship)	1
2015-01-10	Failure (drone ship)	1

Section 4

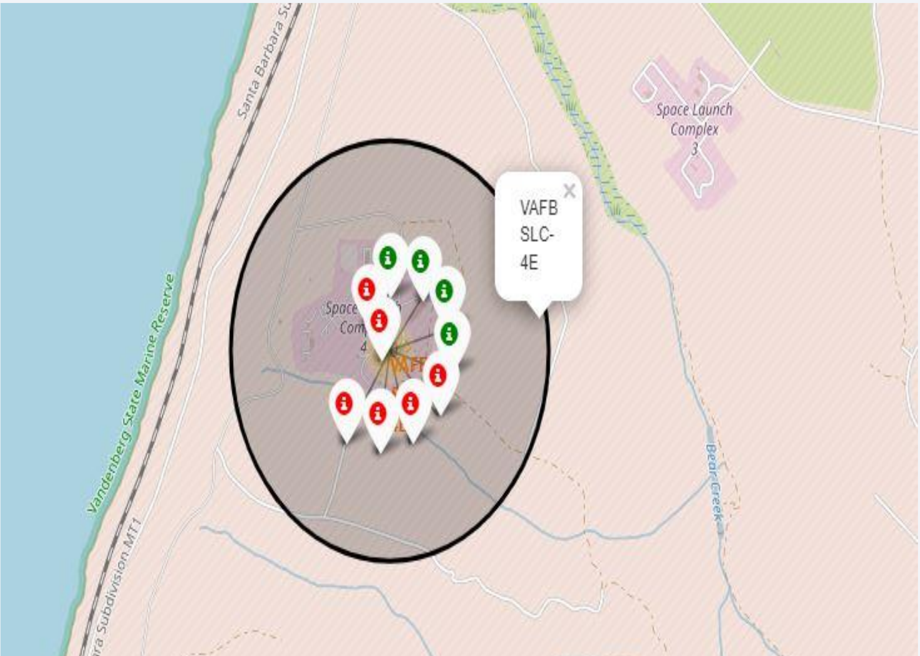
Launch Sites Proximities Analysis

ALL LAUNCH SITES ON THE MAP

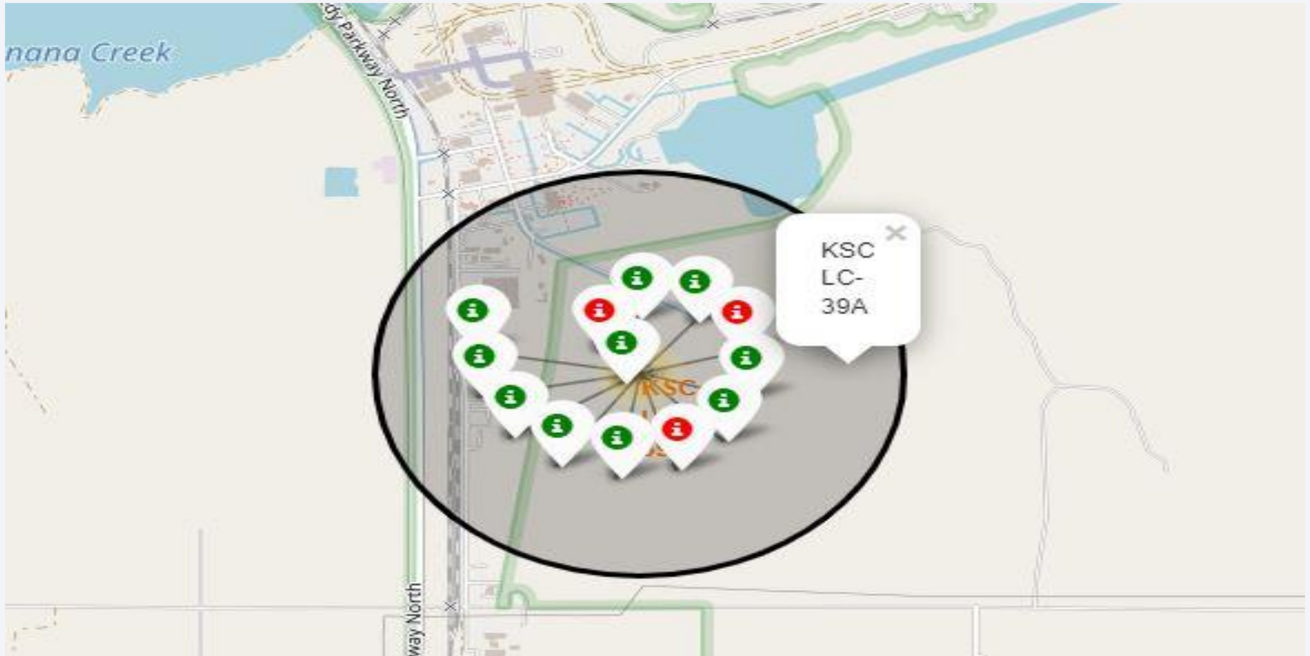


- It can be seen from the map that, all the launch sites are closer to the coast and some are closer to the city of Florida

SUCCESS OR FAILED LAUNCHES FOR EACH SITES ON THE MAP



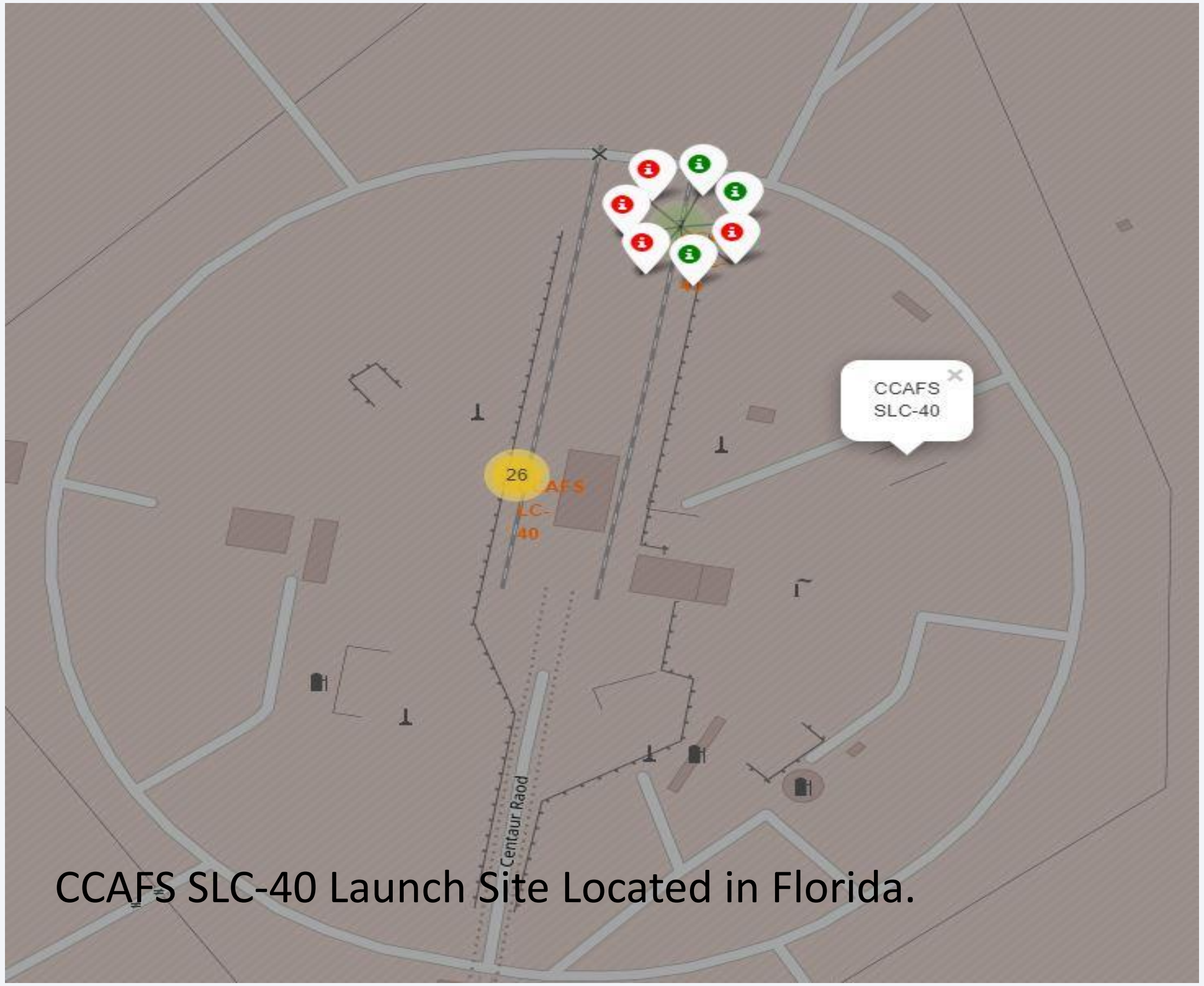
VAFB SLC – 4E Launch Site closer to Vandenberg State Marine reserve

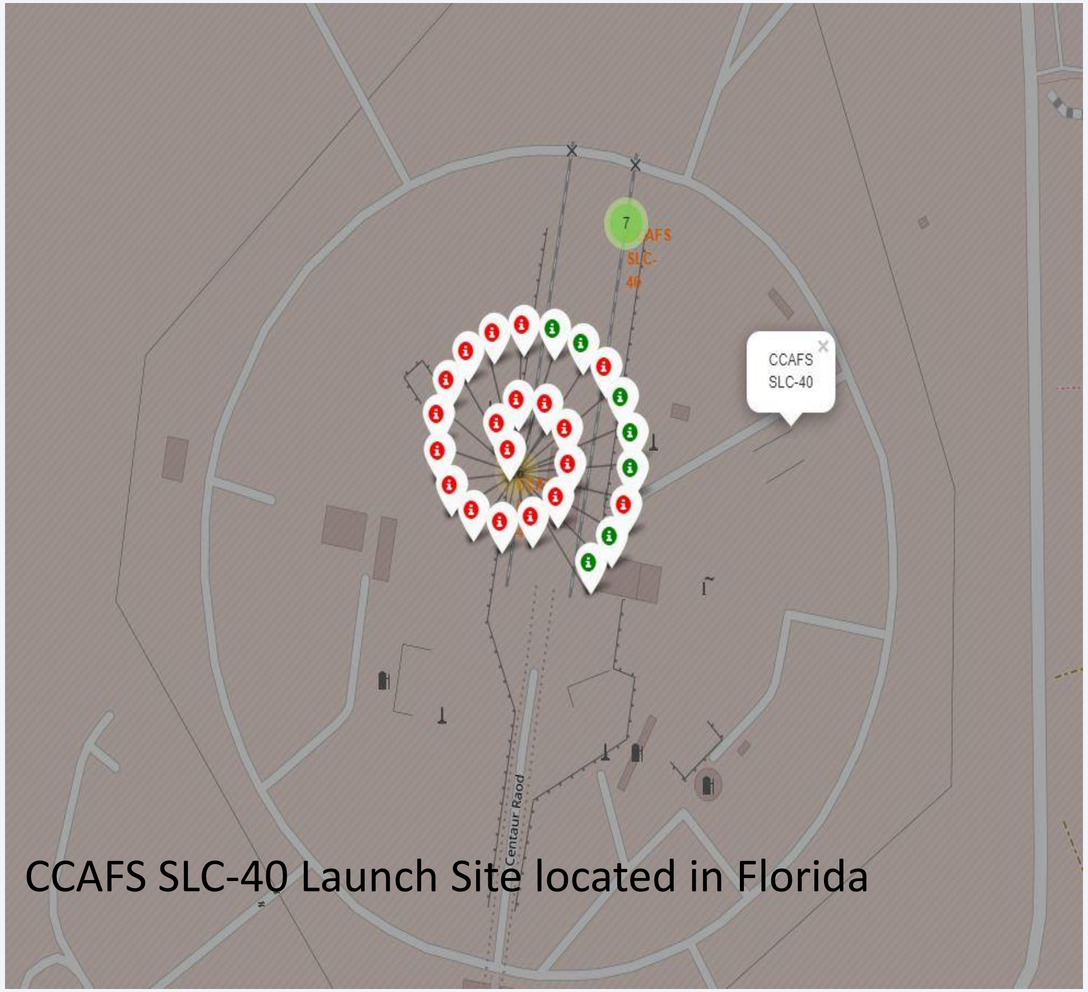


KSC LC-39A Launch Site closer to Vandenberg State Marine reserve

Green makers shows successful landing and Red markers shows unsuccessful landing.

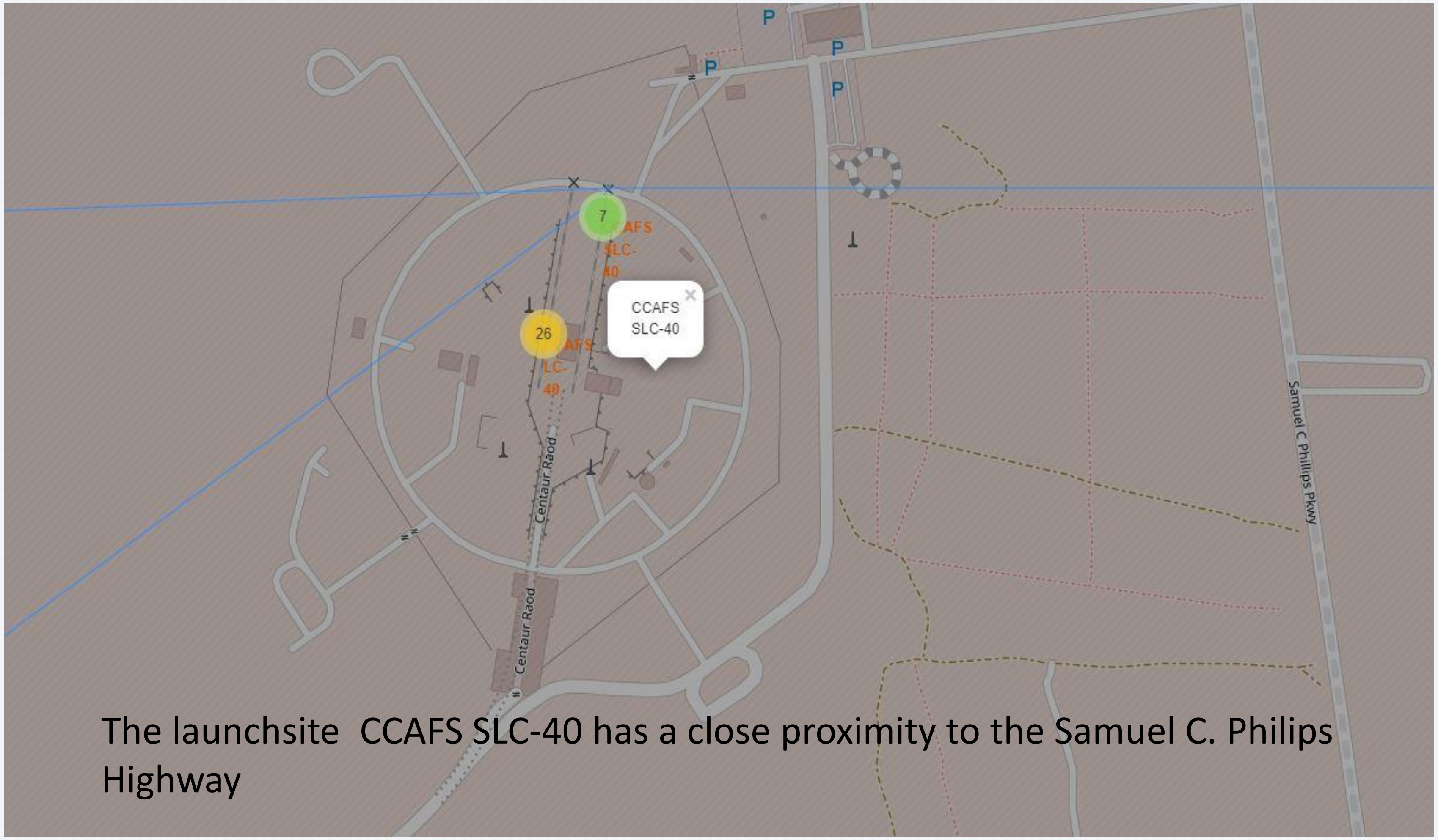
SUCCESS OR FAILED LAUNCHES FOR EACH SITES ON THE MAP





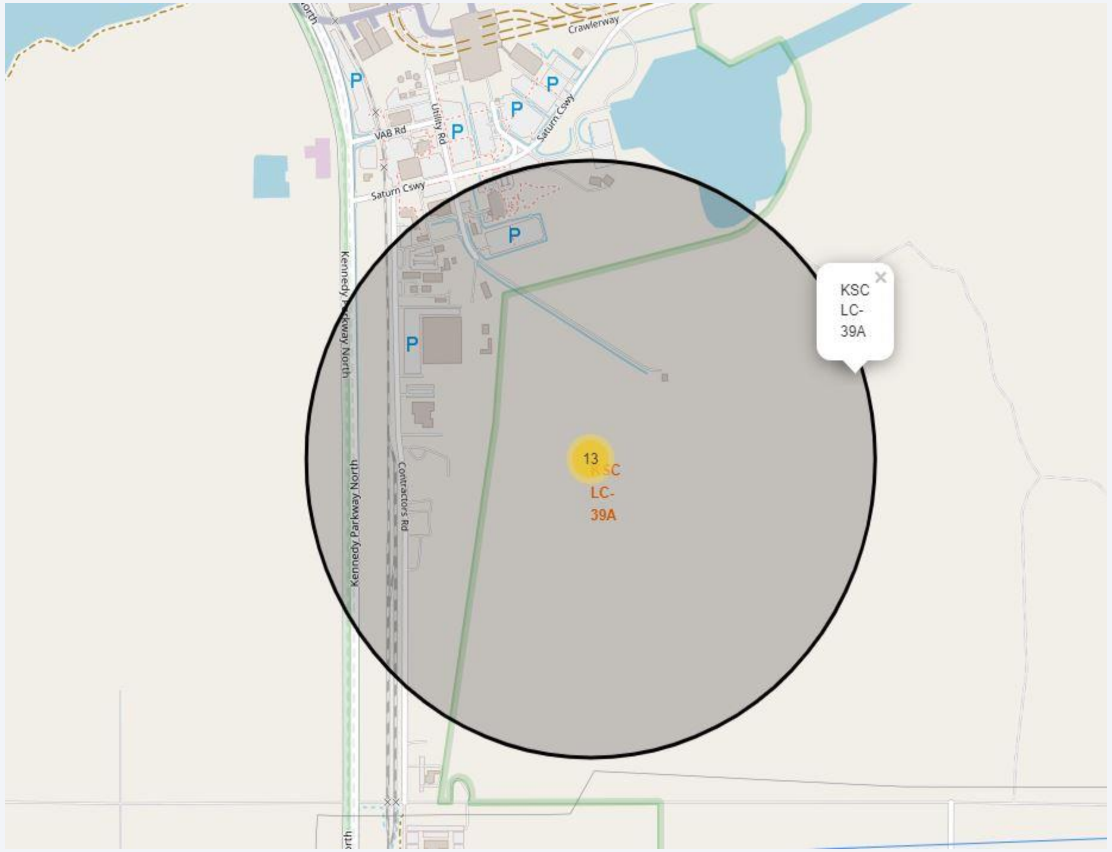
Green makers shows successful landing and Red markers shows unsuccessful landing.

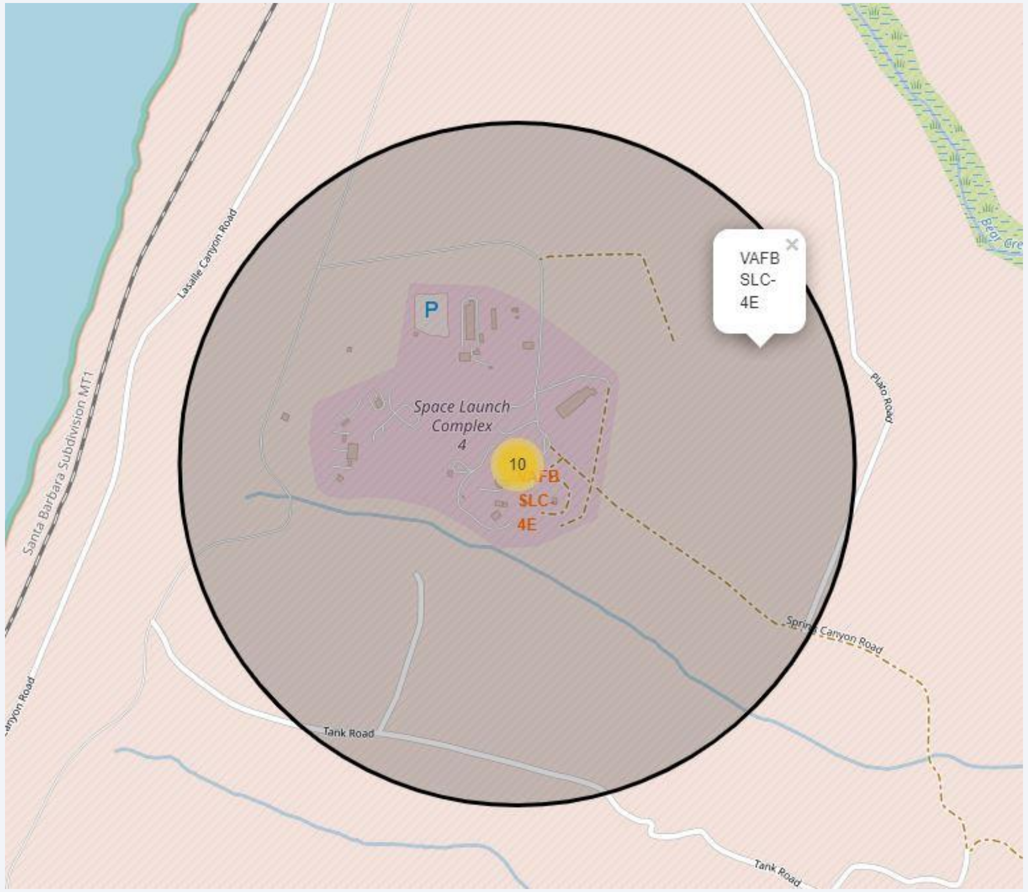
PROXIMITY OF LAUNCH SITES TO HIGHWAY



The launchsite CCAFS SLC-40 has a close proximity to the Samuel C. Philips Highway

PROXIMITY OF LAUNCH SITES TO RAILWAY





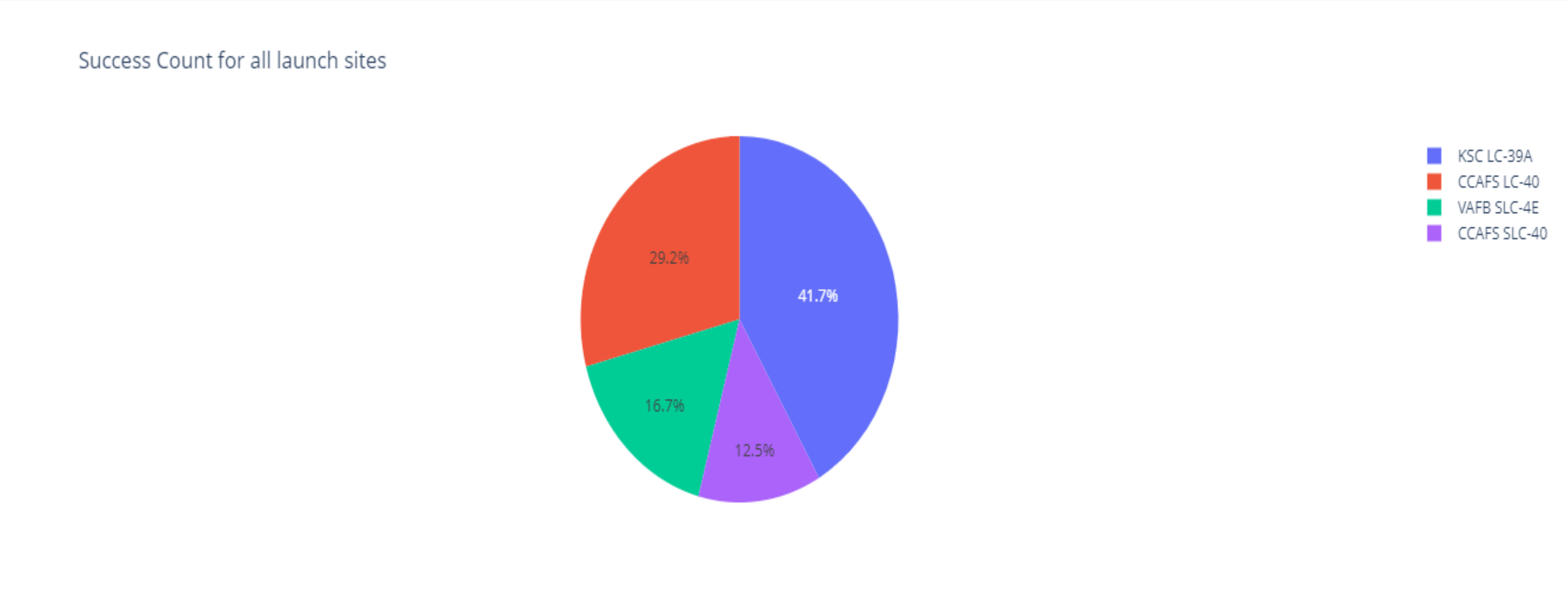
The sites KSC LC-39A and VAFB SLC-4E are close to the rail line .

Section 5

Build a Dashboard with Plotly Dash

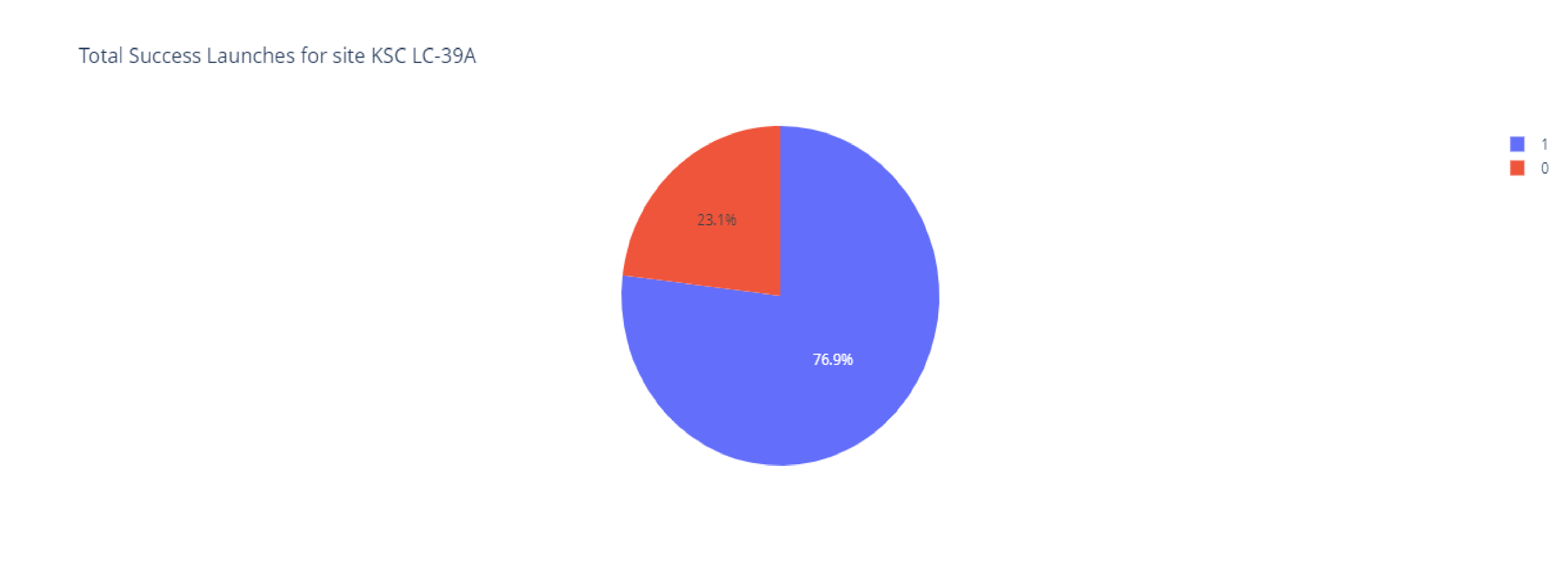
INTERACTIVE DASHBOARD FOR SPACEX

It can be seen from the diagram below that, KSC LC-39A site has the largest successful launches compared to the other sites. The proportion of success is 41.7%.



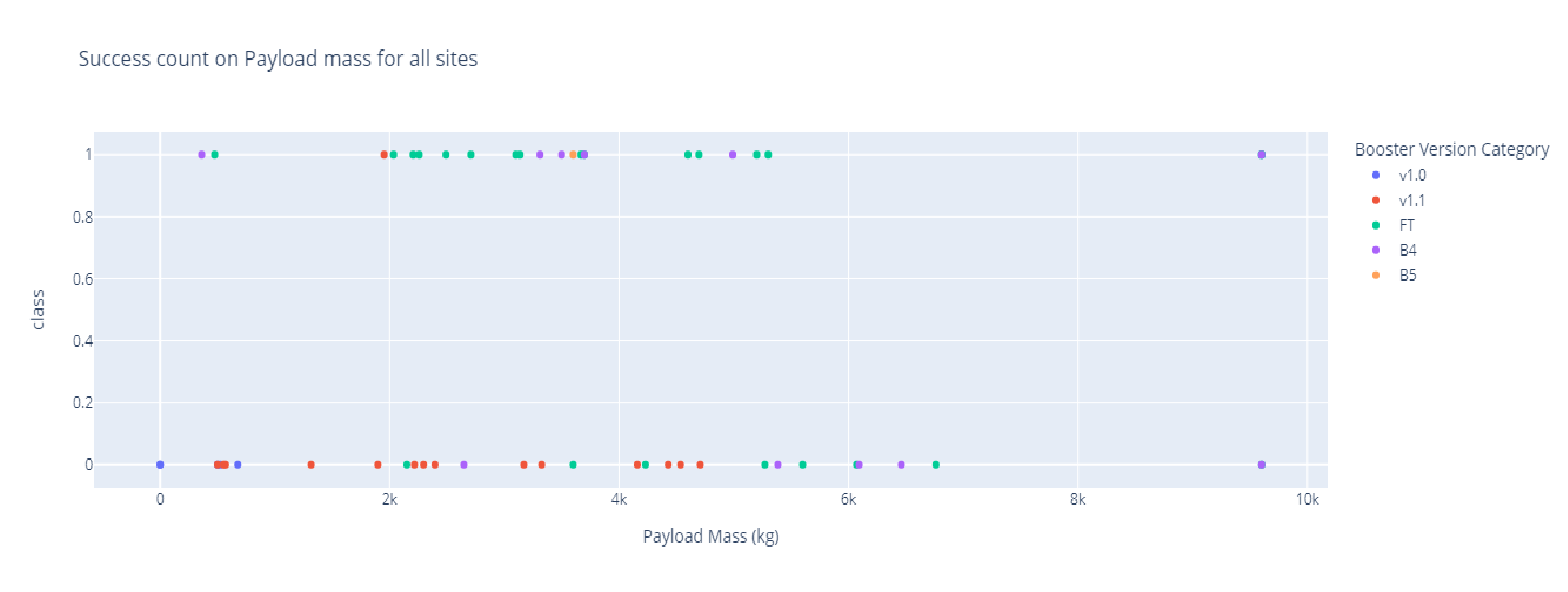
DASHBOARD LAUNCH SITE - KSC LC-39A

The total number of successful launches in the site KSC LC-39A was 76.9%. This means that for every launch, there is a 76.9% chance that the rocket will land successfully and a 23.1% chance that it will fail one way or the other. This is evident in the graph below:



INTERACTIVE DASHBOARD FOR SPACEX

- It can be seen that, payload booster version FT has the highest success rate for a given payload range. The booster in the payload mass range 2k to 6k (kg) has a high success rate when compared to other versions.

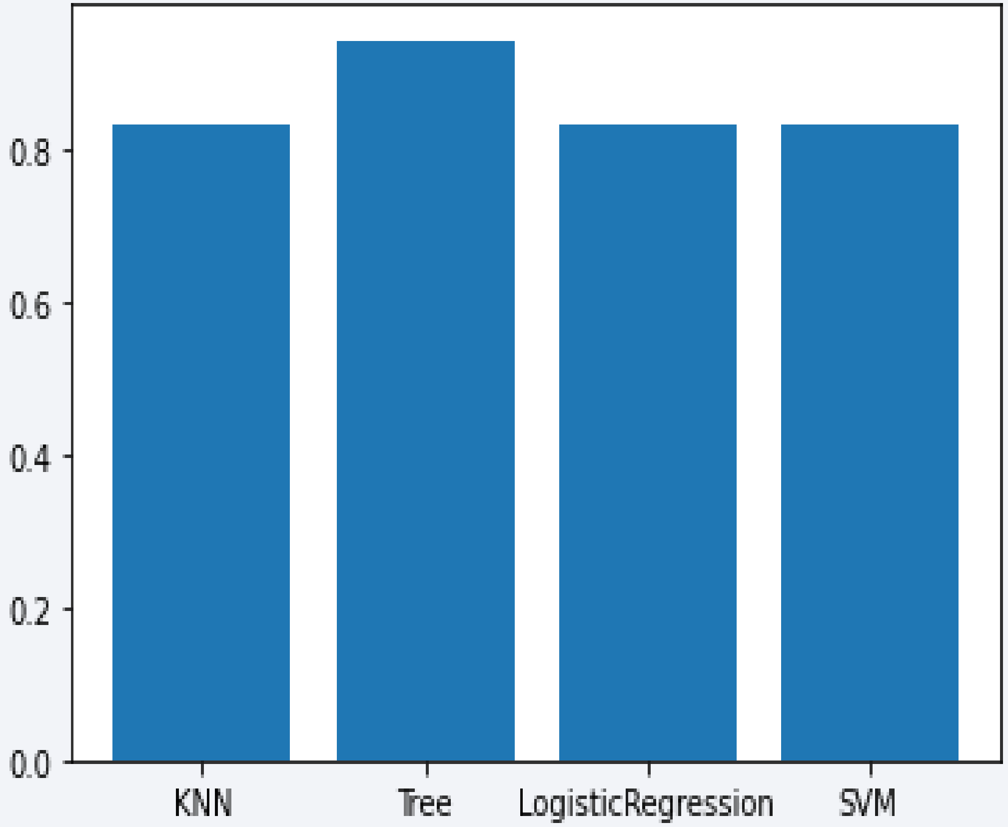


Section 6

Predictive Analysis (Classification)

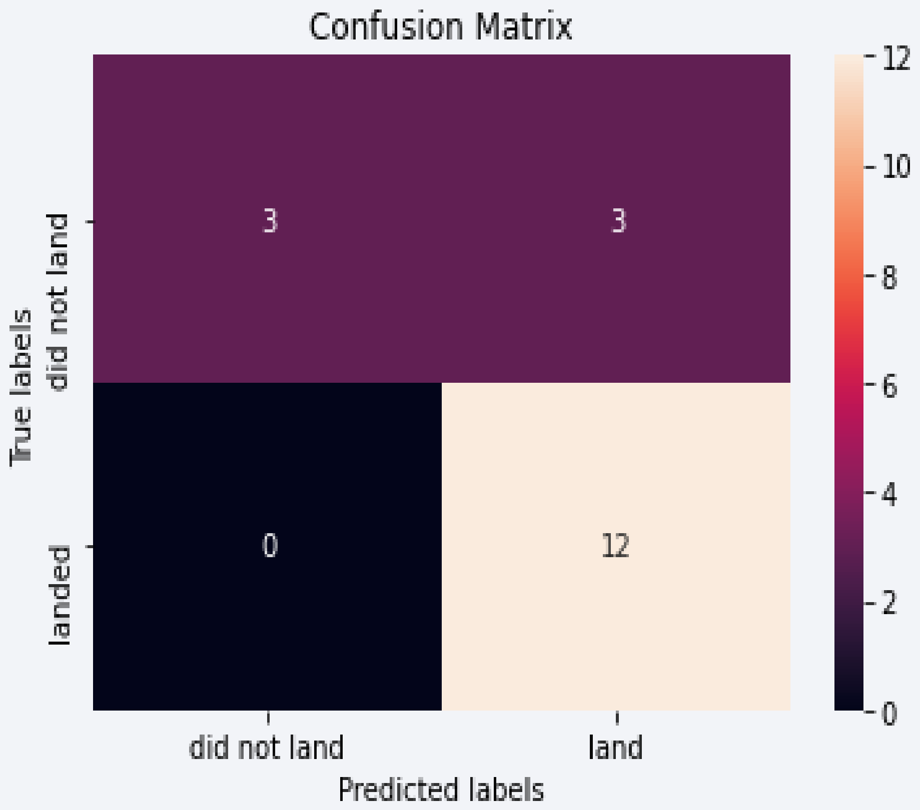
Classification Accuracy

- The bar chart shows the classification accuracy for the various methods. From the graph, the decision tree algorithm yielded the highest classification accuracy.



Confusion Matrix – DECISION TREE

- From the diagram, the confusion matrix of the decision tree can distinguish between different classes with a higher true negatives.



Conclusions

- The best classification method for the predictive analytics is the decision tree classifier.
- Most of the launch sites were closer to coastlines and further away from cities.
- The success rate of landing increased as the year increases.
- For launch site VAFB SLC 4E, there rate of success for rockets launched for payloads 1000kg to 10000kg is very high.
- the orbit types: SSO,HEO,GEO and ES-11 has the highest success rate of landing with a mean class of 1.
- Rockets in the orbit SO has a higher bad outcome of landing.

Appendix

Task 1

Display the names of the unique launch sites in the space mission

In [7]:
```
%%sql
select distinct Launch_Site from SPACEXTBL;
```

```
 * ibm_db_sa://ktd71126:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875/bludb
Done.
```

Out[7]:

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

Task 2

Display 5 records where launch sites begin with the string 'CCA'

In [8]:
```
%%sql
select Launch_Site from SPACEXTBL where Launch_Site like 'CCA%' limit 5;
```

```
 * ibm_db_sa://ktd71126:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875/bludb
Done.
```

Out[8]:

launch_site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Appendix

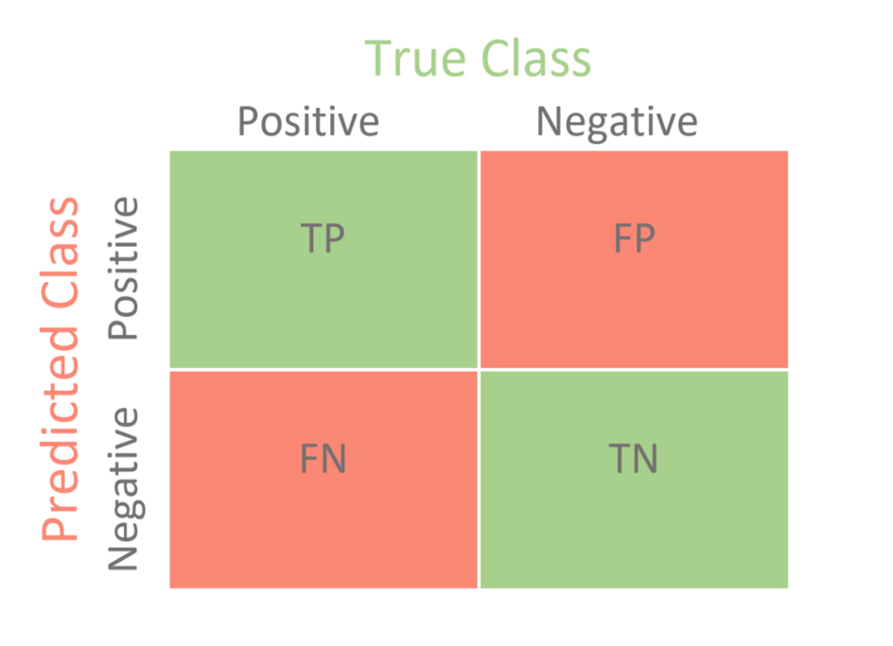
In [71]:

```python
# Function to assign color to launch outcome
def assign_marker_color(launch_outcome):
    if launch_outcome == 1:
        return 'green'
    else:
        return 'red'

spacex_df['marker_color'] = spacex_df['class'].apply(assign_marker_color)
spacex_df.tail(10)
```

Out[71]:

	Launch Site	Lat	Long	class	marker_color
46	KSC LC-39A	28.573255	-80.646895	1	green
47	KSC LC-39A	28.573255	-80.646895	1	green
48	KSC LC-39A	28.573255	-80.646895	1	green
49	CCAFS SLC-40	28.563197	-80.576820	1	green
50	CCAFS SLC-40	28.563197	-80.576820	1	green
51	CCAFS SLC-40	28.563197	-80.576820	0	red
52	CCAFS SLC-40	28.563197	-80.576820	0	red
53	CCAFS SLC-40	28.563197	-80.576820	0	red
54	CCAFS SLC-40	28.563197	-80.576820	1	green
55	CCAFS SLC-40	28.563197	-80.576820	0	red



Thank you!