Data Science Capstone project

Aadarsh Agarwal

21th August 2021

Outline

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

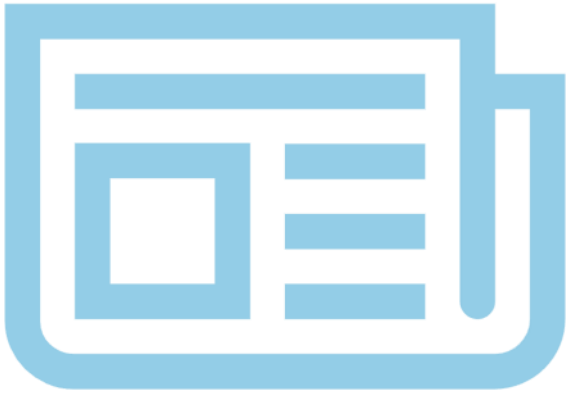
Executive Summary

- The Falcon 9 launch data is collected from two chief sources namely SpaceX API and Wiki Pages using web scraping.
- It is then refined by converting it into pandas dataframe and transforming all categorical values to numerical ones and accounting for null values.
- After the data wrangling exploratory analysis is performed using data visualization techniques and SQL. A dashboard is also constructed to better visualize the data.
- After understanding the relationship between various features of the dataset, various classification models are used to find the optimum one.
- Finally, the optimum classification model is used to predict whether the first stage will land successfully or not.

Introduction

- Project background and context
  - SpaceX is the most successful company of the commercial space age, making space travel affordable. The prime reason behind SpaceX's success is that its rocket launches are relatively inexpensive. This is because SpaceX can reuse the first stage.
  - In order to determine the cost of a launch, the reusability of the first stage needs to be determined and therefore its landing.
- Problems you want to find answers
  - Whether the first stage can be reused or not?
  - Whether the landing will be successful or not?

Methodology



- Data collection methodology:
  - Describe how data were collected
- Perform data wrangling
  - Describe how data were processed
- 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

Methodology

Data collection

- SpaceX launch data is gathered the SpaceX REST API using different endpoints
- Web Scraping Wiki pages for obtaining Falcon 9 Launch data

Data collection – SpaceX API

- The SpaceX REST API endpoints, or URL, starts with api.spacexdata.com/v4/
- The different end points, for example, are, /capsules and /cores
- The endpoint used is api.spacexdata.com/v4/launches/past
- This URL is used to target a specific endpoint of the API to get past launch data
- GitHub Url: https://github.com/Adarsh9904/Applied-Data-Science-Capstone/blob/master/Data%20Collection%20API.ipynb

Data collection – Web scraping

- Use the Python BeautifulSoup package to web scrape some HTML tables that contain valuable Falcon 9 launch records
- Parse the data from those tables and convert them into a Pandas data frame for further visualization and analysis
- Transform this raw data into a clean dataset which provides meaningful data on the situation
- GitHub URL: https://github.com/Adarsh9904/Applied-Data-Science-Capstone/blob/master/Data%20Collection%20with%20Web%20Scraping.ipynb

Data wrangling

- Perform a get request using the requests library to obtain the launch data, which will the be used to get the data from the API
- It can be viewed by calling the .json() method
- To convert this JSON to a dataframe, the json_normalize function is used which will “normalize” the structured json data into a flat table
- In the data, rockets have been represented by an identification number and not actual data
- Use the API again targeting another endpoint to gather specific data for each ID number
- The data will be stored in lists and will be used to create the dataset
- Filter the dataset to remove Falcon 1 data as only Falcon 9 data is required
- Finally replace the null values for Payload Mass with its mean

EDA with data visualization

- Scatter plots for categorical features were plotted using sns.catplot. The following charts were plotted:
  - FlightNumber vs PayloadMass
  - FlightNumber vs LaunchSite
  - PayloadMass vs LaunchSite
  - Orbit vs FlightNumber
  - PayloadMass vs Orbit
- The scatter plots helped determining the relationship between various parameters and their combined influence on the success rate.
- The bar chart for different orbit's success rate showed GEO and SSO had maximum success rates.
- Finally, a line chart showed that success rate has increased over the years.

- GitHub URL: https://github.com/Adarsh9904/Applied-Data-Science-Capstone/blob/master/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

- The performed SQL queries are as follows:
  - select unique(launch_site) from spacextbl;
  - select * from spacextbl where launch_site like 'CCA%' limit 5;
  - select sum(payload_mass__kg_) as total_payload_mass from spacextbl group by customer having customer='NASA (CRS)';
  - select avg(payload_mass__kg_) as average_payload_mass from spacextbl group by booster_version having booster_version = 'F9 v1.1';
  - select min(date) as first_successful_landing from spacextbl where "Landing _Outcome" = 'Success (ground pad)';
  - select unique(booster_version) as booster_names from spacextbl where "Landing _Outcome" = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000;
  - select count(mission_outcome) as succesful_outcomes from spacextbl where mission_outcome like '%Success%';
    select count(mission_outcome) as failed_outcomes from spacextbl where mission_outcome like '%Failure%';
    select mission_outcome, count(mission_outcome) from spacextbl group by mission_outcome;
  - select unique(booster_version) from spacextbl where payload_mass__kg_ = (select max(payload_mass__kg_) from spacextbl);
  - select monthname(date) as "Month", "Landing _Outcome", booster_version, launch_site from spacextbl where year(date)='2015' and "Landing _Outcome" = 'Failure (drone ship)';
  - select date, count("Landing _Outcome") as successful_landing_outcomes from spacextbl group by date having date between '2010-06-04' and '2017-03-20' order by date desc;
- GitHub URL: https://github.com/Adarsh9904/Applied-Data-Science-Capstone/blob/master/EDA%20with%20SQL.ipynb

Build an interactive map with Folium

- Location markers for all launch sites were added to the map
- Circle object was also added for each launch site
- Marker clusters were added to group the large number of launch sites together on the map
- Finally, a polyline was drawn between a selected launch site and coastline indicating the distance
- GitHub URL: https://github.com/Adarsh9904/Applied-Data-Science-Capstone/blob/master/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

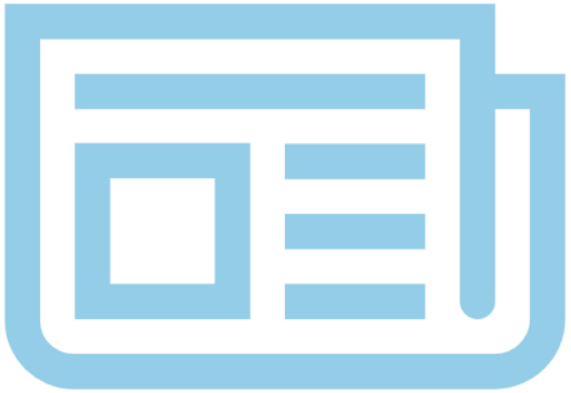
Build a Dashboard with Plotly Dash

- The pie chart is connected to the drop down menu.
- If all sites are selected the pie chart shows relative success of all Launch Sites.
- If a particular site is selected the pie chart shows successful and non-successful landings for that particular site.
- The scatter plot is connected to both the drop down menu and the range slider.
- If all sites are selected it shows the Payload vs Launch Outcome scatter plot for all sites where the payload range can be adjusted by the range slider.
- If a particular site is selected it shows the Payload vs Launch Outcome scatter plot for that particular site, where the payload range can be adjusted by the range slider.
- GitHub URL: https://github.com/Adarsh9904/Applied-Data-Science-Capstone/blob/master/spacex_dash_app.py

Predictive analysis (Classification)

- Preprocessing to standardize our data
- Train_test_split split the data into training and testing data,
- Train the model and perform Grid Search, to find the hyperparameters that allow a given algorithm to perform best
- Using the best hyperparameter values, determine the model with the best accuracy using the training data
- Test for Logistic Regression, Support Vector machines, Decision Tree Classifier, and K-nearest neighbors
- Finally output the confusion matrix.

Results

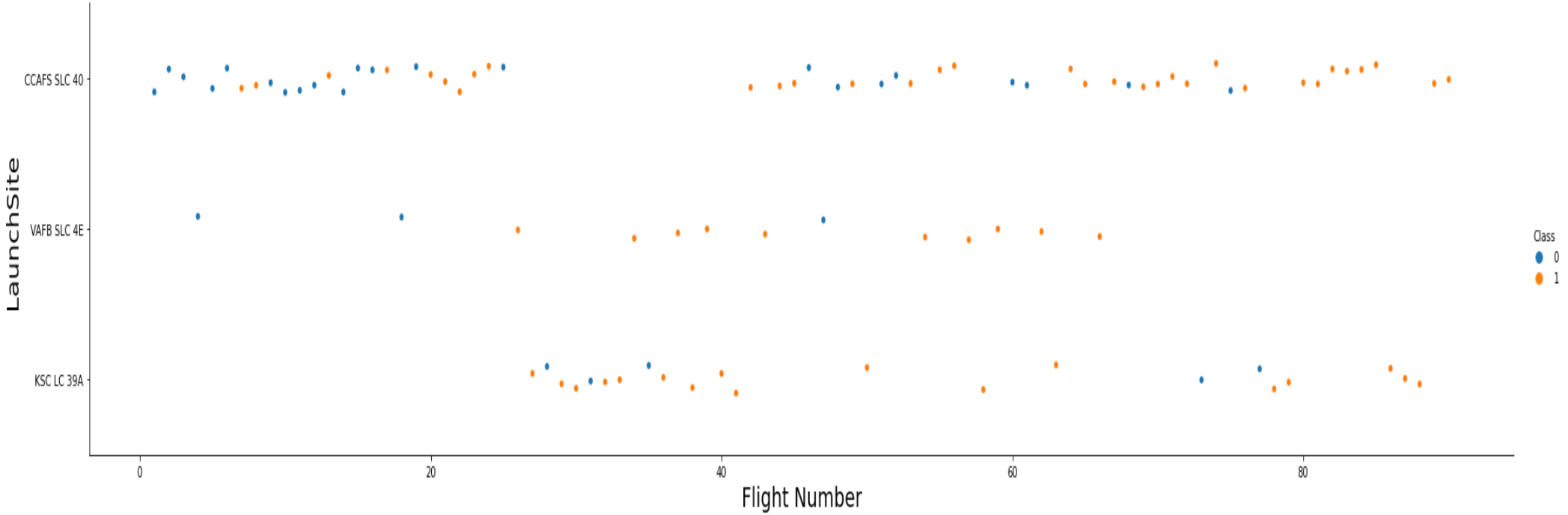


- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

EDA with Visualization

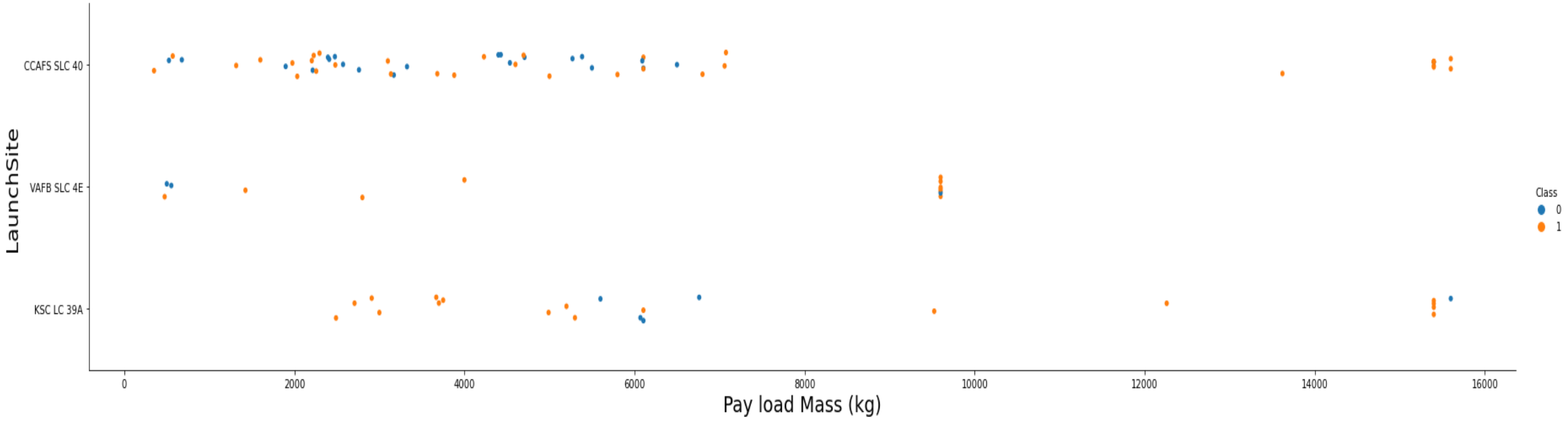
Flight Number vs. Launch Site

The number of success landings seems to increase with flight number. Also CCAFS LC-40 seems to lower success rate as compared to KSC LC-39A and VAFB SLC 4E.



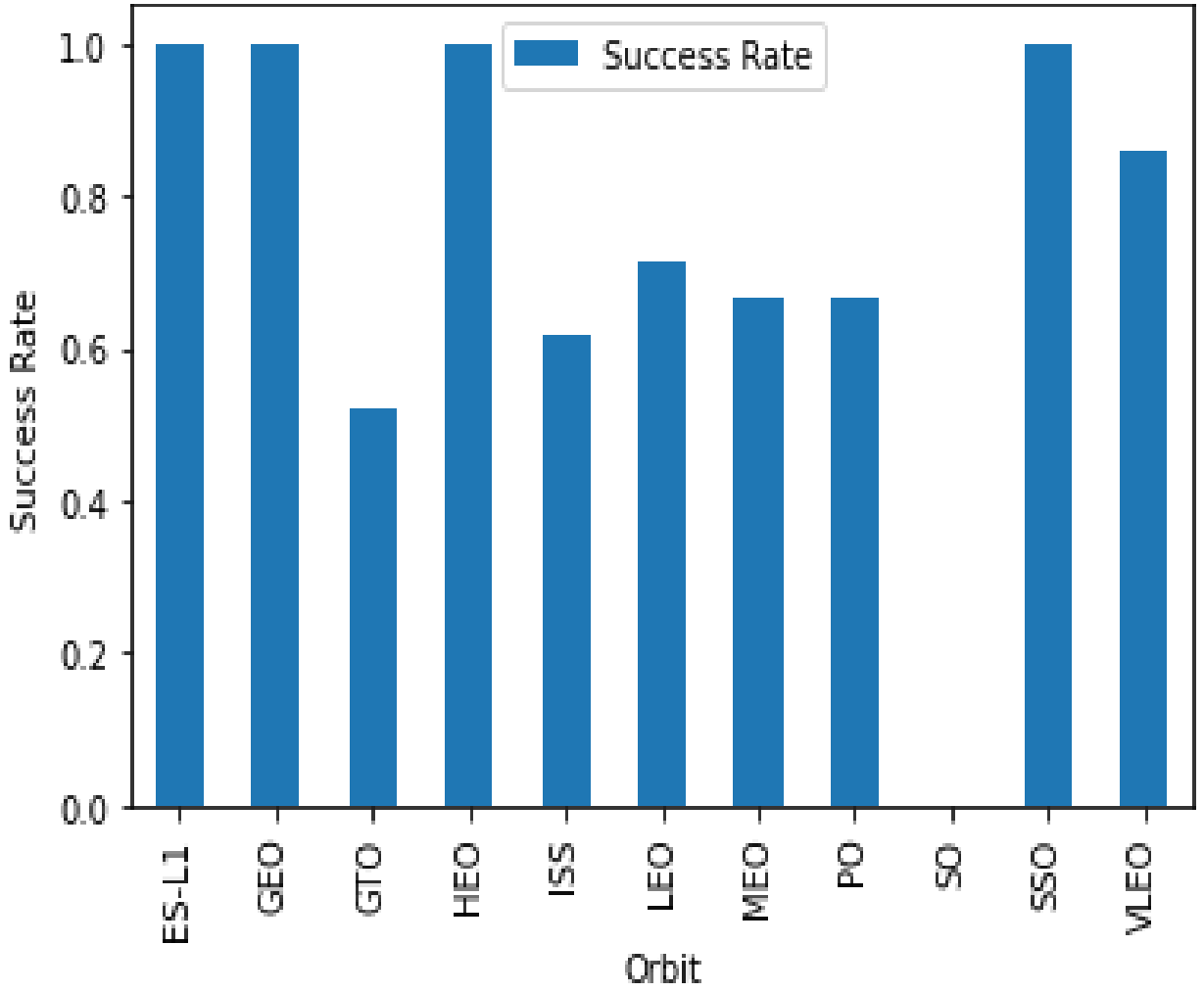
Payload vs. Launch Site

The success rate for VAFB SLC 4E and KSC LC 39A increases with increase in payload mass whereas there is no such pattern for CCAFS SLC 40



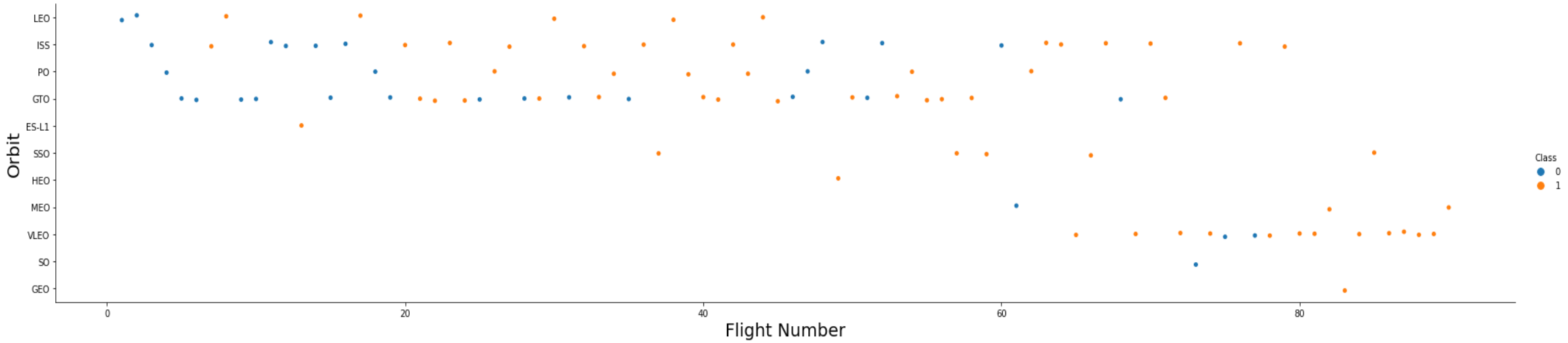
Success rate vs. Orbit type

- The success rate of ES-L1, GEO, HEO and SSO is maximum and had all successful landings whereas for SO, success is 0 and hence it had no successful landings.
- All the other orbits have their success rate in between them.



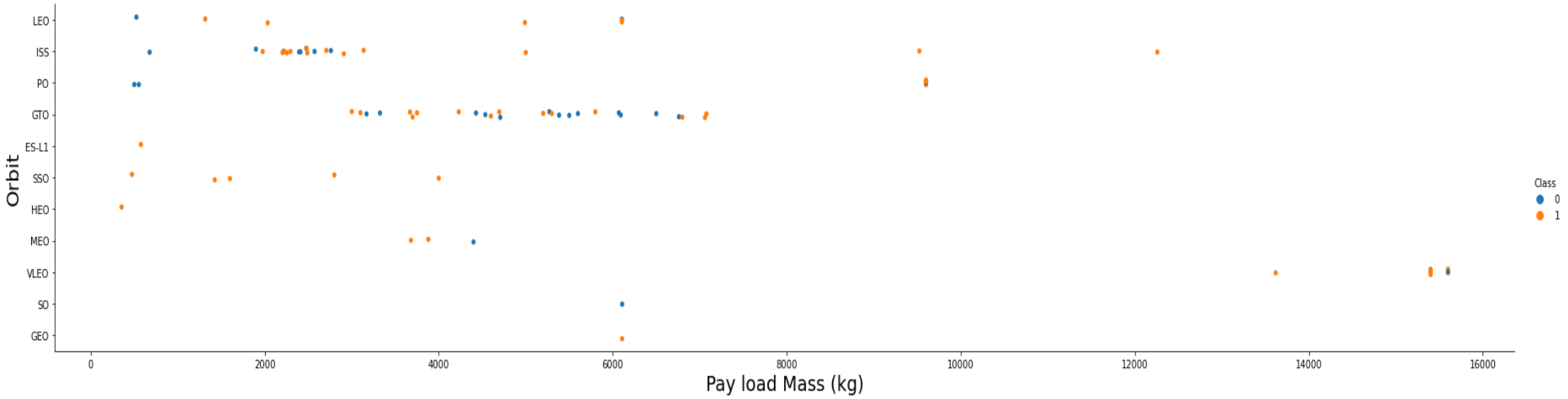
Flight Number vs. Orbit type

It can be seen that for LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.



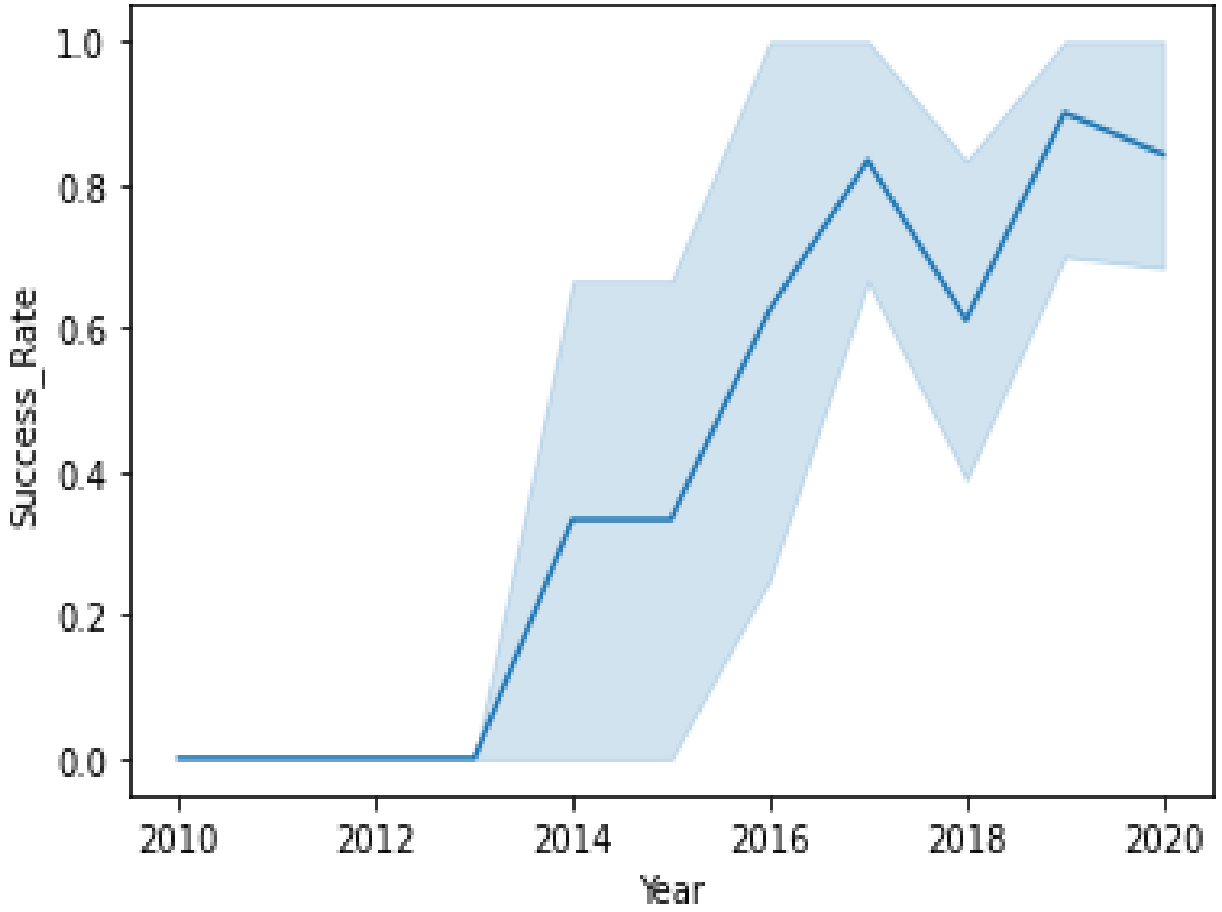
Payload vs. Orbit type

It is observed that Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.



Launch success yearly trend

It can be observed that the success rate since 2013 kept increasing till 2020



EDA with SQL

All launch site names

This query selects all the unique launch site names from the dataset

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

In [10]:
```
%%sql
select unique(launch_site) from spacextbl;
```

```
 * ibm_db_sa://vnm12624:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[10]:

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

Launch site names begin with `CCA`

This query shows records for all those launch sites whose names begin with “CCA”.

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

In [12]:
```
%sql select * from spacextbl where launch_site like 'CCA%' limit 5;
```

```
 * ibm_db_sa://vnm12624:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[12]:

DATE	Time (UTC)	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	Landing _Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	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)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total payload mass

This query calculates the total payload mass carried by boosters launched by NASA (CRS).

Display the total payload mass carried by boosters launched by NASA (CRS)

In [14]:
```
%%sql
select sum(payload_mass__kg_) as total_payload_mass from spacextbl
group by customer
having customer='NASA (CRS)';
```

```
 * ibm_db_sa://vnm12624:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[14]:

total_payload_mass
45596

Average payload mass by F9 v1.1

This query displays the average payload mass carried by booster version F9 v1.1

Display average payload mass carried by booster version F9 v1.1

In [16]:
```
%%sql
select avg(payload_mass__kg_) as average_payload_mass from spacextbl
group by booster_version
having booster_version = 'F9 v1.1';
```

```
 * ibm_db_sa://vnm12624:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[16]:

average_payload_mass
2928

First successful ground landing date

This query lists the date when the first successful landing outcome in ground pad was achieved.

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

In [38]:
```
%%sql
select min(date) as first_successful_landing from spacextbl
where "Landing _Outcome" = 'Success (ground pad)';
```

```
 * ibm_db_sa://vnm12624:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[38]:

first_successful_landing
2015-12-22

Successful drone ship landing with payload between 4000 and 6000

This query displays the names of boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

In [42]:
```
%%sql
select unique(booster_version) as booster_names from spacextbl
where "Landing _Outcome" = 'Success (drone ship)' and
payload_mass__kg_ between 4000 and 6000;
```

```
 * ibm_db_sa://vnm12624:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[42]:

booster_names
F9 FT B1021.2
F9 FT B1031.2
F9 FT B1022
F9 FT B1026

Total number of successful and failure mission outcomes

This query calculates and displays the total number of successful and failure mission outcomes

List the total number of successful and failure mission outcomes

In [53]:
```
%%sql

select count(mission_outcome) as succesful_outcomes from spacextbl
where mission_outcome like '%Success%';
```

```
 * ibm_db_sa://vnm12624:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[53]:

succesful_outcomes
100

In [54]:
```
%%sql

select count(mission_outcome) as failed_outcomes from spacextbl
where mission_outcome like '%Failure%';
```

```
 * ibm_db_sa://vnm12624:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[54]:

failed_outcomes
1

In [56]:
```
%sql select mission_outcome, count(mission_outcome) from spacextbl group by mission_outcome;
```

```
 * ibm_db_sa://vnm12624:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[56]:

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters carried maximum **payload**

This query lists the names of the booster which have carried the maximum payload mass.

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

In [58]:
```
%%sql
select unique(booster_version) from spacextbl
where payload_mass__kg_ = (select max(payload_mass__kg_) from spacextbl);
```

```
 * ibm_db_sa://vnm12624:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[58]:

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

2015 launch records

This query displays the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015

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

In [68]:
```
%%sql
select monthname(date) as "Month", "Landing _Outcome", booster_version, launch_site from spacextbl
where year(date)='2015' and "Landing _Outcome" = 'Failure (drone ship)';
```

```
 * ibm_db_sa://vnm12624:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[68]:

Month	Landing _Outcome	booster_version	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Rank success count between 2010-06-04 and 2017-03-20

This query ranks the count of successful landing outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

Rank the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

In [86]:
```
%%sql
select date, count("Landing _Outcome")
as successful_landing_outcomes from spacextbl
group by date
having date between '2010-06-04' and '2017-03-20'
order by date desc;
```

```
 * ibm_db_sa://vnm12624:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

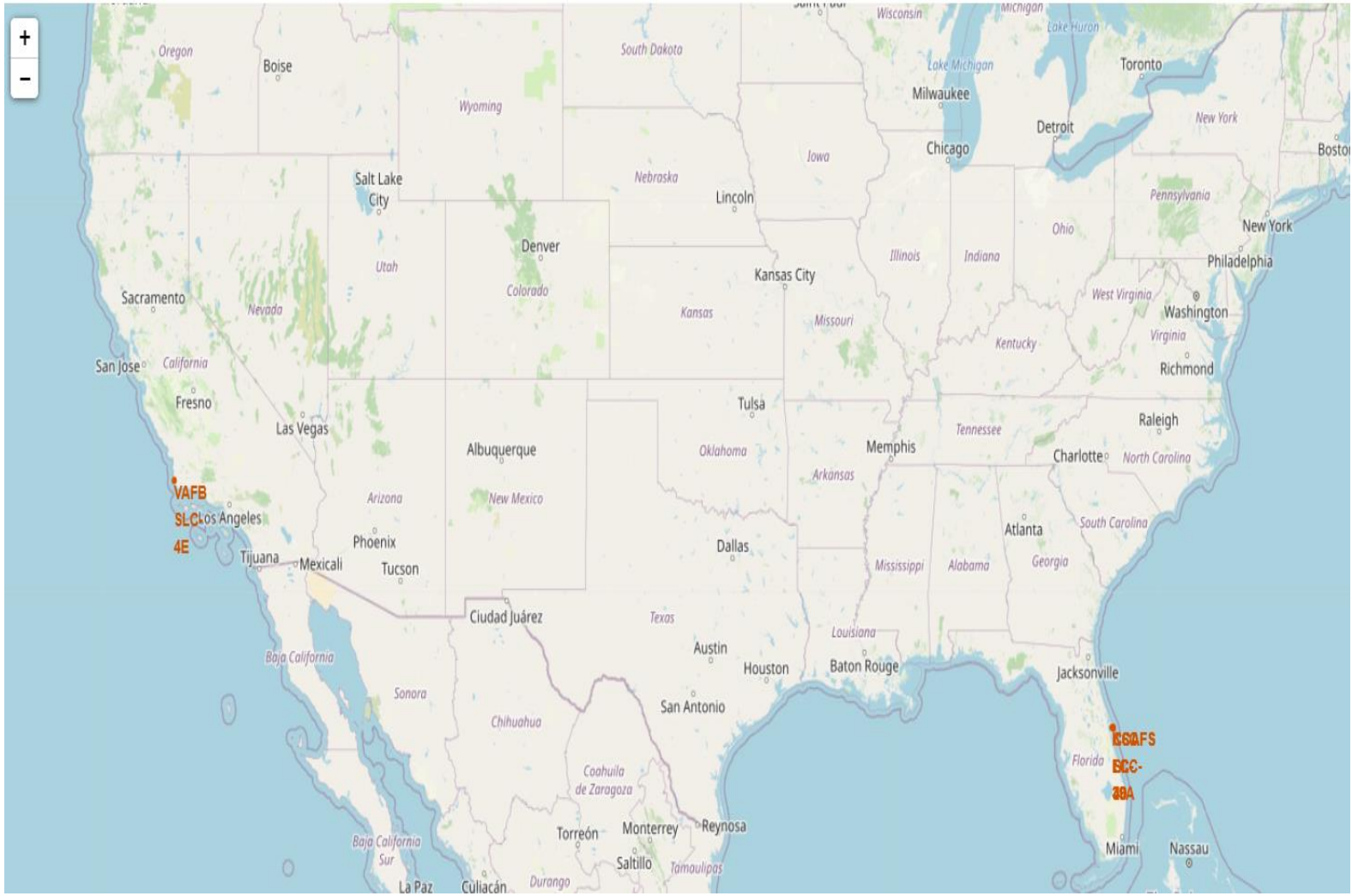
Out[86]:

DATE	successful_landing_outcomes
2017-03-16	1
2017-02-19	1
2017-01-14	1
2016-08-14	1
2016-07-18	1
2016-06-15	1
2016-05-27	1
2016-05-06	1
2016-04-08	1
2016-03-04	1
2016-01-17	1
2015-12-22	1
2015-06-28	1
2015-04-27	1
2015-04-14	1
2015-03-02	1
2015-02-11	1
2015-01-10	1
2014-09-21	1
2014-09-07	1
2014-08-05	1
2014-07-14	1
2014-04-18	1
2014-01-06	1
2013-12-03	1
2013-09-29	1
2013-03-01	1
2012-10-08	1
2012-05-22	1
2010-12-08	1
2010-06-04	1

Interactive map with Folium

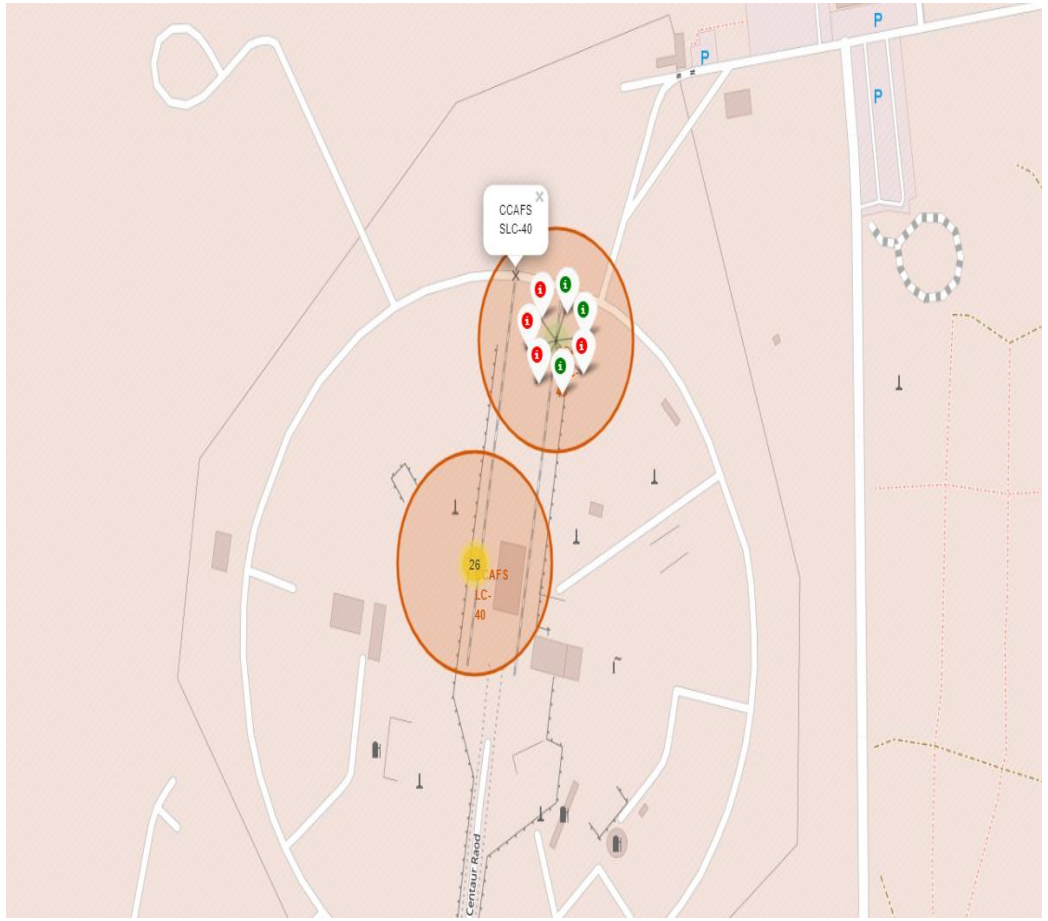
Launch Sites Locations

- This map shows the locations of the various launch sites.
- It can be observed that all the launch sites are in the coastal areas.



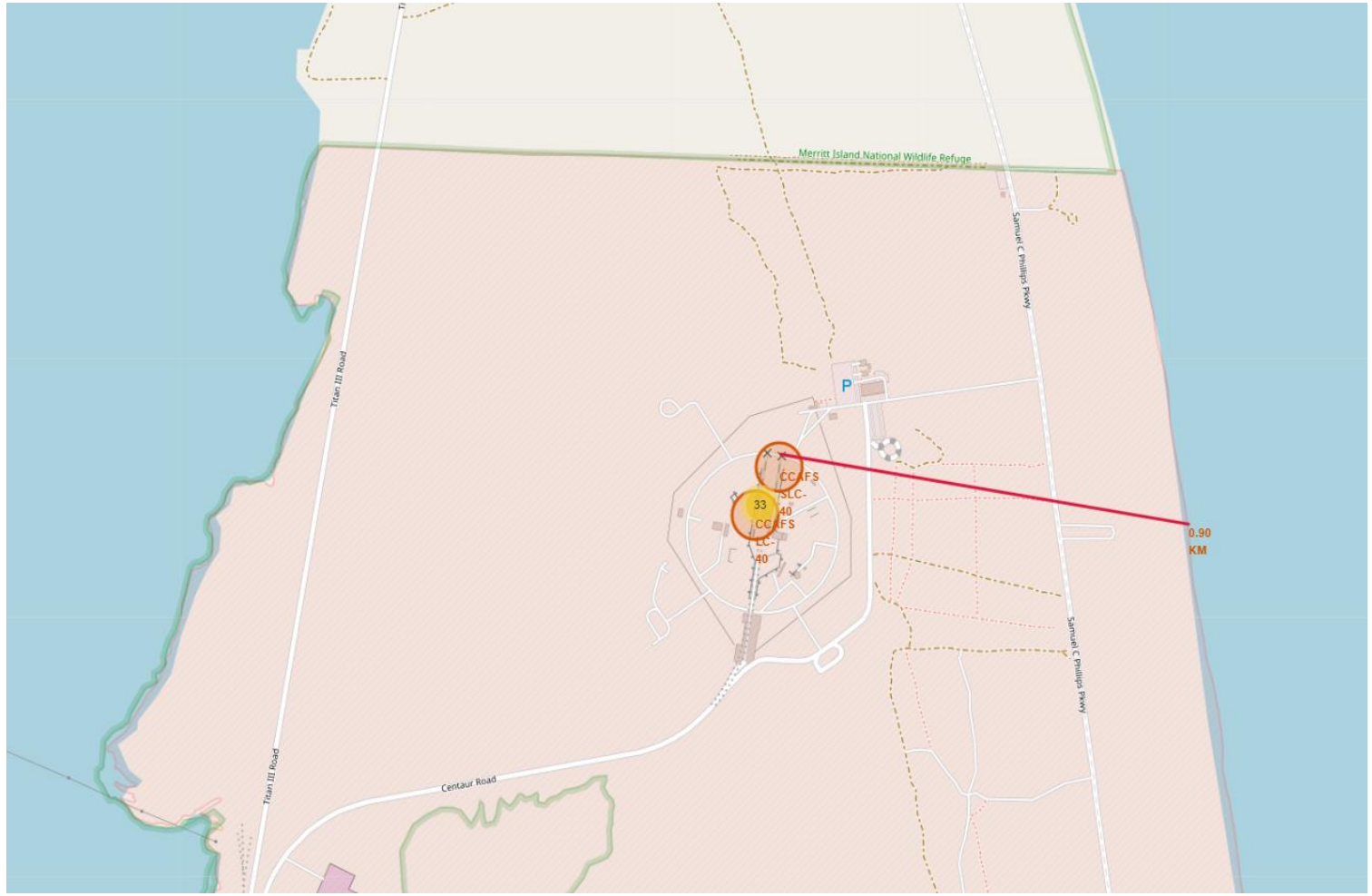
Color Coded Markers For Launch Sites

- This map shows zoomed in marker clusters. The markers in these clusters are color coded for success rate with successful landings indicated with green and failures indicated with red.
- This better helps to get an overview of successful landings at a launch site.



Distance Line Between Launch Site and Coastline

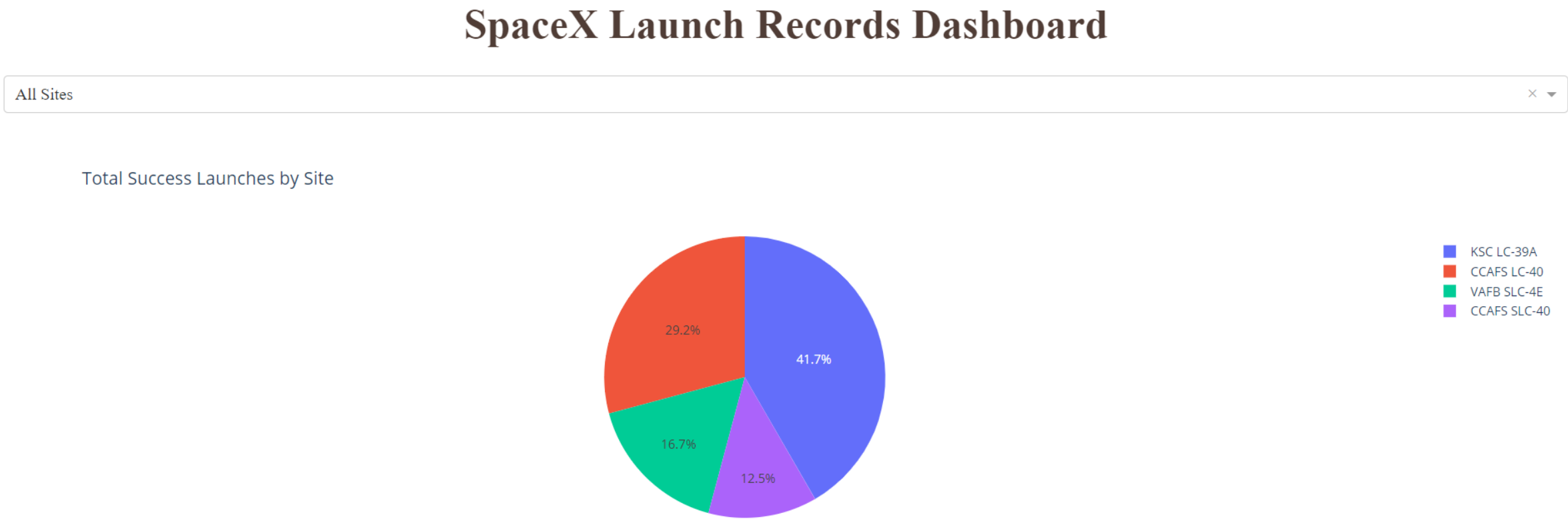
- This map shows that the approximate distance between launch site and coastline is 0.9 KM.
- The presence of nearby natural and man made geographical features help infer if it has any influence on site's success rate.



Build a Dashboard with Plotly Dash

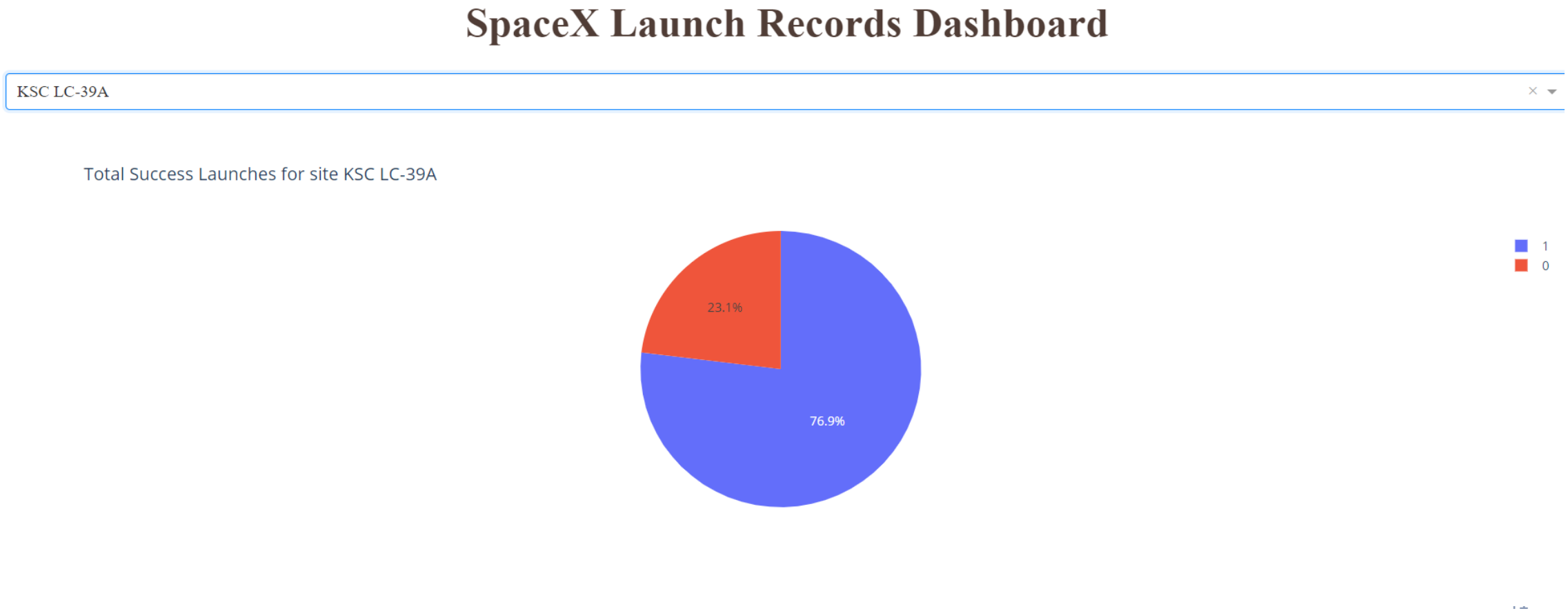
Relative Success of Launch Sites

This pie chart clearly shows that launch site KSC LC-39A has highest success rate among all the launch sites. The share of all launch sites is also shown.



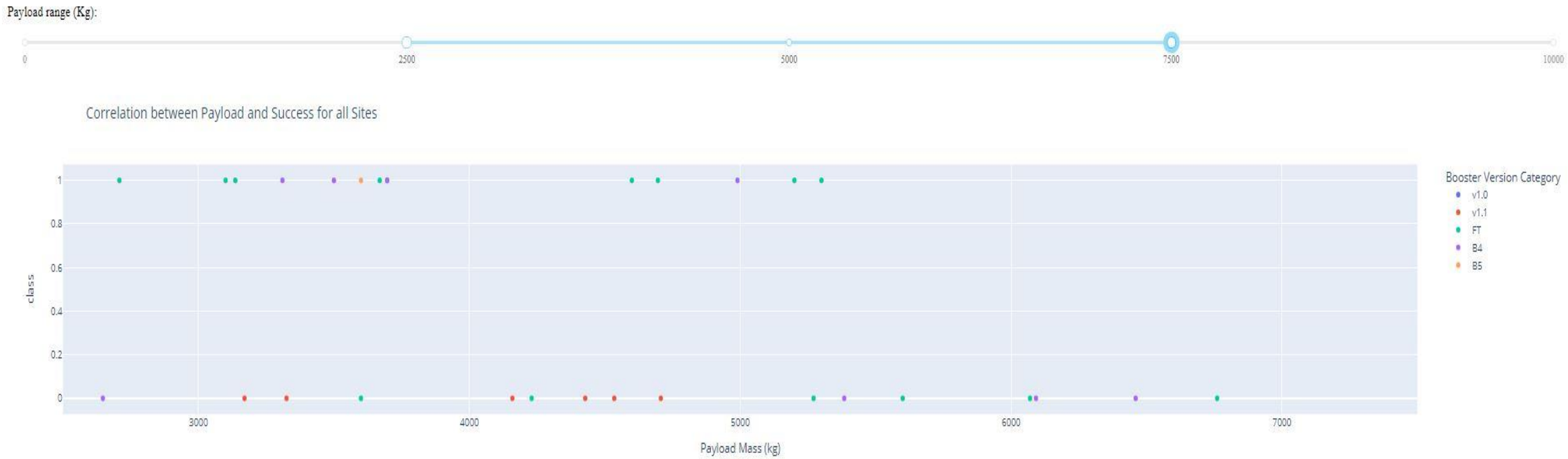
Successful Landings for KSC LC-39A

As inferred from the previous figure launch site KSC LC-39A has the maximum successful landings. The pie chart below shows the how many of its total landings were successful.



Payload vs Launch Outcome with Range Slider

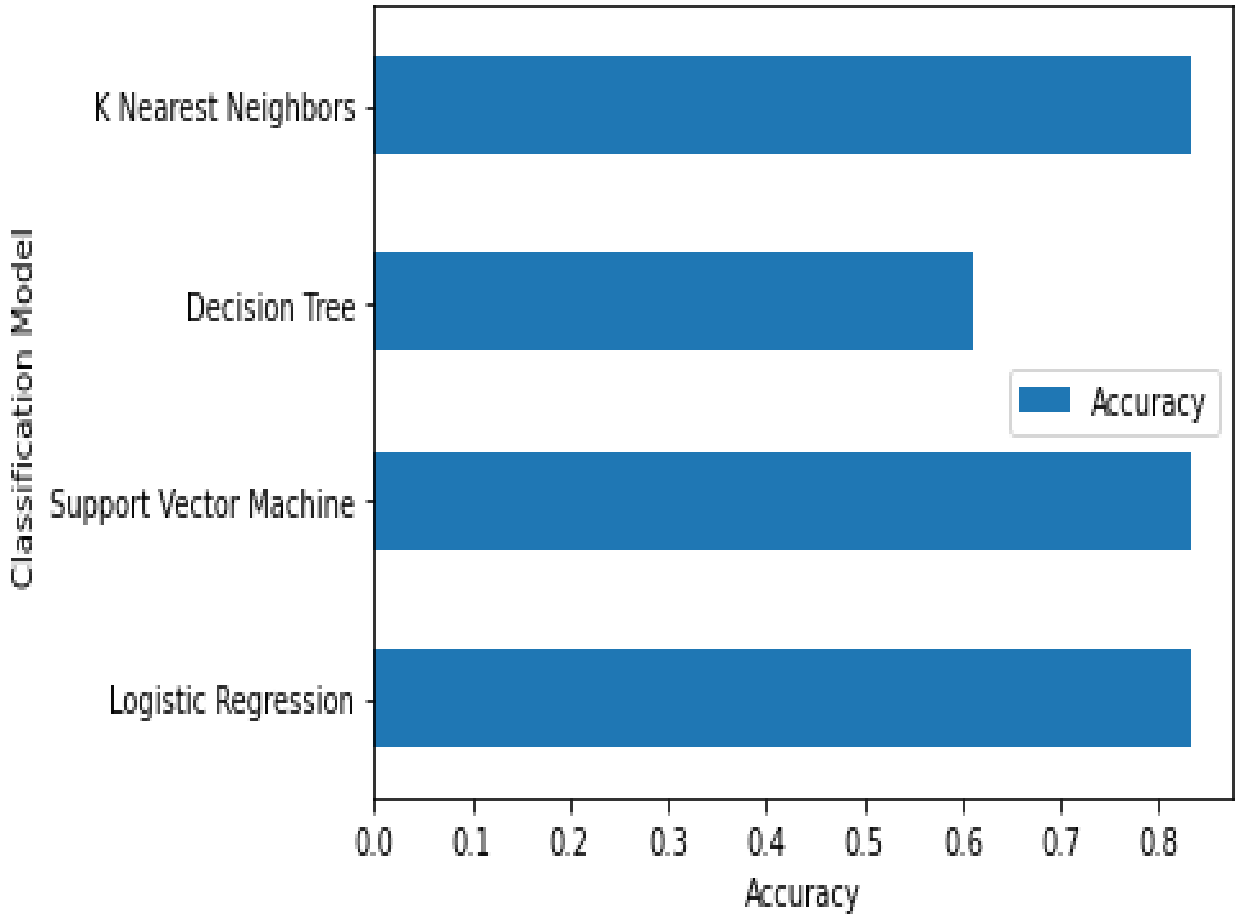
The plot below shows the correlation between payload mass and success for all sites where the selected payload mass range is 2500 to 7500 kg.



Predictive analysis (Classification)

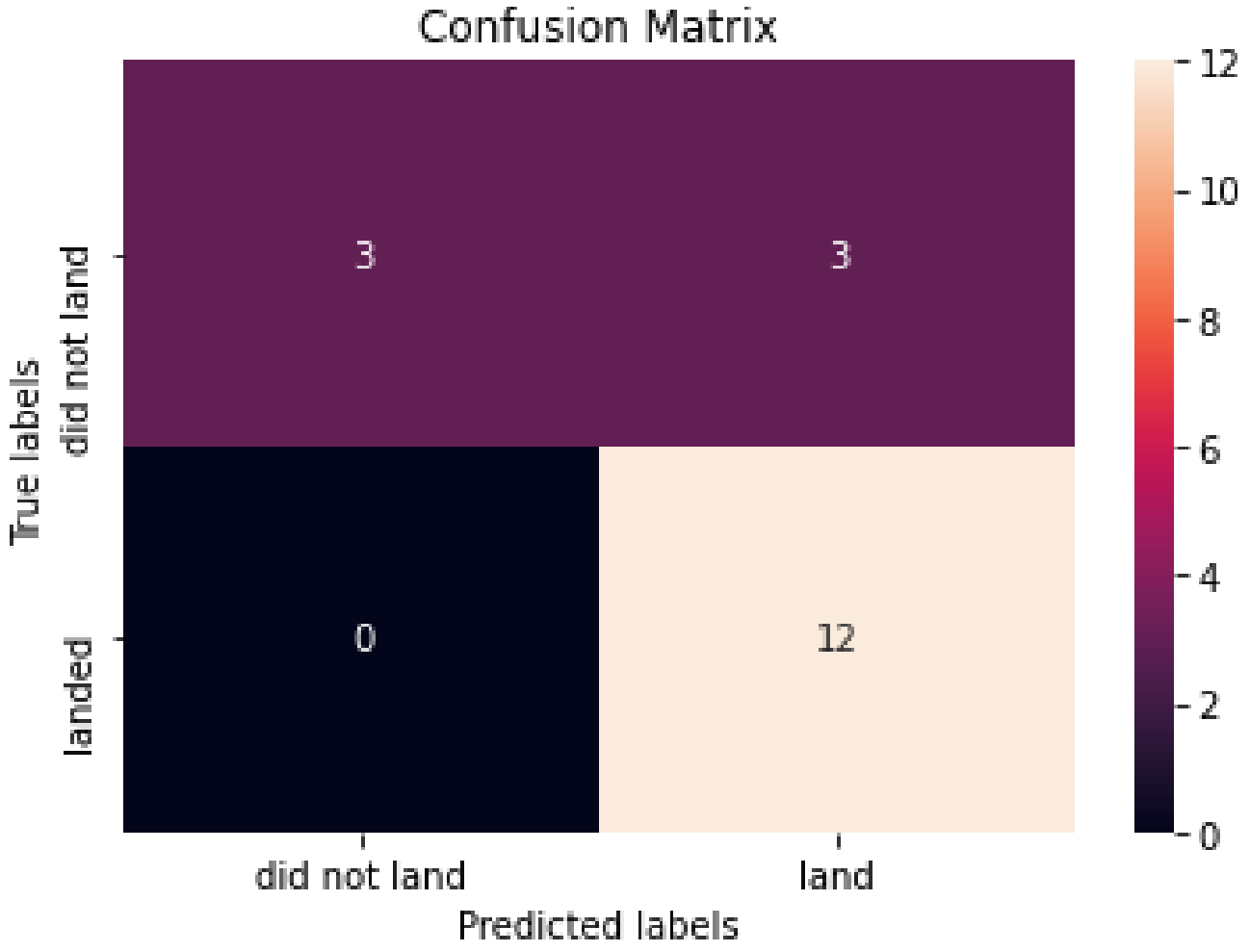
Classification Accuracy

All the models except for decision tree have an accuracy of 0.83 which also happens to be the maximum.



Confusion Matrix

- This is the confusion matrix for the model with maximum accuracy.
- From the matrix it can be inferred that the model predicts false positives.
- It means it predicts 3 of the 18 stage 1 will land successfully which in reality wasn't the case.



CONCLUSION

- The model developed is able to predict whether the given rocket’s first stage will land successfully or not with an accuracy of 0.83 on the test data set.
- This can help determine whether those first stage can be reused or not which in turn can help estimate the cost of the launch.
- Using this data, the company SpaceY can strategize accordingly to compete with SpaceX.

APPENDIX

The following is a link to GitHub repository which contains all the code and data utilized in this report.

https://github.com/Adarsh9904/Applied-Data-Science-Capstone.git