IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Rahul Rathore

13th August, 2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

This project help predicting if the SpaceX Falcon 9 first stage will land successfully using several machine learning algorithms.

- The project includes various steps as mentioned below:
  - Data collection, wrangling, and formatting
  - Exploratory data analysis
  - Interactive data visualization
  - Machine learning prediction
- Visualizations shows how landing outcome is correlated to the location of launch sites, payload mass and various other factors
- We found that almost all classification models have similar out of sample accuracy, however Decision Trees model also reflect high training accuracy.

Introduction

- 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 want to know correlation of outcome of launch on certain factors, so that we can find the more favorable outcome basis the factors, like Launch site, Payload mass, Booster version, etc.

Methodology

Methodology

- Data collection methodology:
  - Data is collected using Request to the SpaceX API and Web scrapping Falcon 9 Wikipedia web page.
- Perform data wrangling
  - Data is wrangled using python’s pandas and numpy libraries.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Whole data is standardized and then split into train and test data for the evaluation, the best hyperparameters are determined using GridSearchCV class and evaluated using score method.

Data Collection – SpaceX API

- The API endpoint is https://api.spacexdata.com/v4/launches/past
- The whole data is converted in json and then into dataframe using pandas json_normalize method.
- More api calls were made to get the values of booster version, launch site, payload mass and more.
- All the information is stored in launch_dict variable in dictionary format and then converted into dataframe using pandas's from_dict method.

Data Collection – SpaceX API

- Resulted Dataframe is filtered with Falcon 9 data and missing values are replaced with the mean of column.
- Final Dataframe is 90 rows × 17 columns in dimension.
- Here github file link: https://github.com/rathore793/IBM_capstone_project/blob/main/data_collection_api.ipynb

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.561857
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.561857
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.632093
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577366	28.561857

Data Collection - Web Scraping

- Data is collected using python module BeautifulSoup, it is a web scrapping module used to scrap below Wikipedia page: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- Soup object is created and various methods are applied in order to extract the required information from the web page.
- launch_dict is created with the column names as keys.

Data Collection - Web Scraping

- The dictionary is converted into dataframe using pandas.
- Final Dataframe is 120 rows × 11 columns in dimension.
- Here github file link: https://github.com/rathore793/IBM_capstone_project/blob/main/data_collection_web_scraping.ipynb

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10

Data Wrangling

- Further the data is analyzed for missing values or different data type of the columns.
- Then data is analyzed basis the different launch sites, Orbits and launching outcomes.
- Another column is added with name class which indicate the outcome of the launch, Success is 1 and Failure is 0.
- Here github file link: https://github.com/rathore793/IBM_capstone_project/blob/main/EDA.ipynb

EDA with Data Visualization

- Exploratory Data Analysis is performed using pandas and matplotlib.
- Scatter plot is created for visualize correlation between different features and launch outcomes.
- Bar plot is created to visualize the orbit relation with the success rate.
- Line plot is created to visualize the change in success rate over the year.
- Here’s the github link: https://github.com/rathore793/IBM_capstone_project/blob/main/EDA_with_visualization.ipynb

EDA with SQL

- SQL is used to find the insights of the data as mentioned below:
    - To find the names of all different sites.
    - Which booster version carried the maximum payload.
    - What is the date of first successful launch.
    - Count of total number of successful and failed outcomes.
    - Name of Booster version and Sites with failed outcomes.
    - Average payload mass carried by specific booster version, etc.
- Here's Github link: https://github.com/rathore793/IBM_capstone_project/blob/main/EDA_using_SQL.ipynb

Build an Interactive Map with Folium

- All sites are marked on the map with latitudes and longitudes with circle and marker as the name of the launch sites.
- Marker cluster is created with color coding, as red showing failed outcome and green as successful.
- Distance is calculated with its proximities such as coastline, railway line, city and highway, to check whether the launch sites are kept at the safe distance from proximities.
- Here’s github link: https://github.com/rathore793/IBM_capstone_project/blob/main/Interactive%20Visual%20Analytics.ipynb

Build a Dashboard with Plotly Dash

- We have created a dashboard with a dropdown menu to select the site name, range slider to select the range of payload mass, a pie chart to show success rate and scatter plot to show successful outcome basis the payload mass, booster version and name of the site.
- All site data is show by default so that we can compare the success rate of all the sites.
- Scatter plot helps in knowing what payload mass is having highest success rate for particular site.
- Here’s the github link for its python script: https://github.com/rathore793/IBM_capstone_project/blob/main/spacex_dash_app.py

Predictive Analysis

- Data is loaded, standardized and split into training and testing dataset.
- Objects were created for Logistic Regression, Support Vector Machine, Decision Tree and K Nearest Neighbor.
- Parameter variable also created for each method to find the best parameters.
- GridSearchCV object is created with parameter as method object, their parameter and CV.

Predictive Analysis

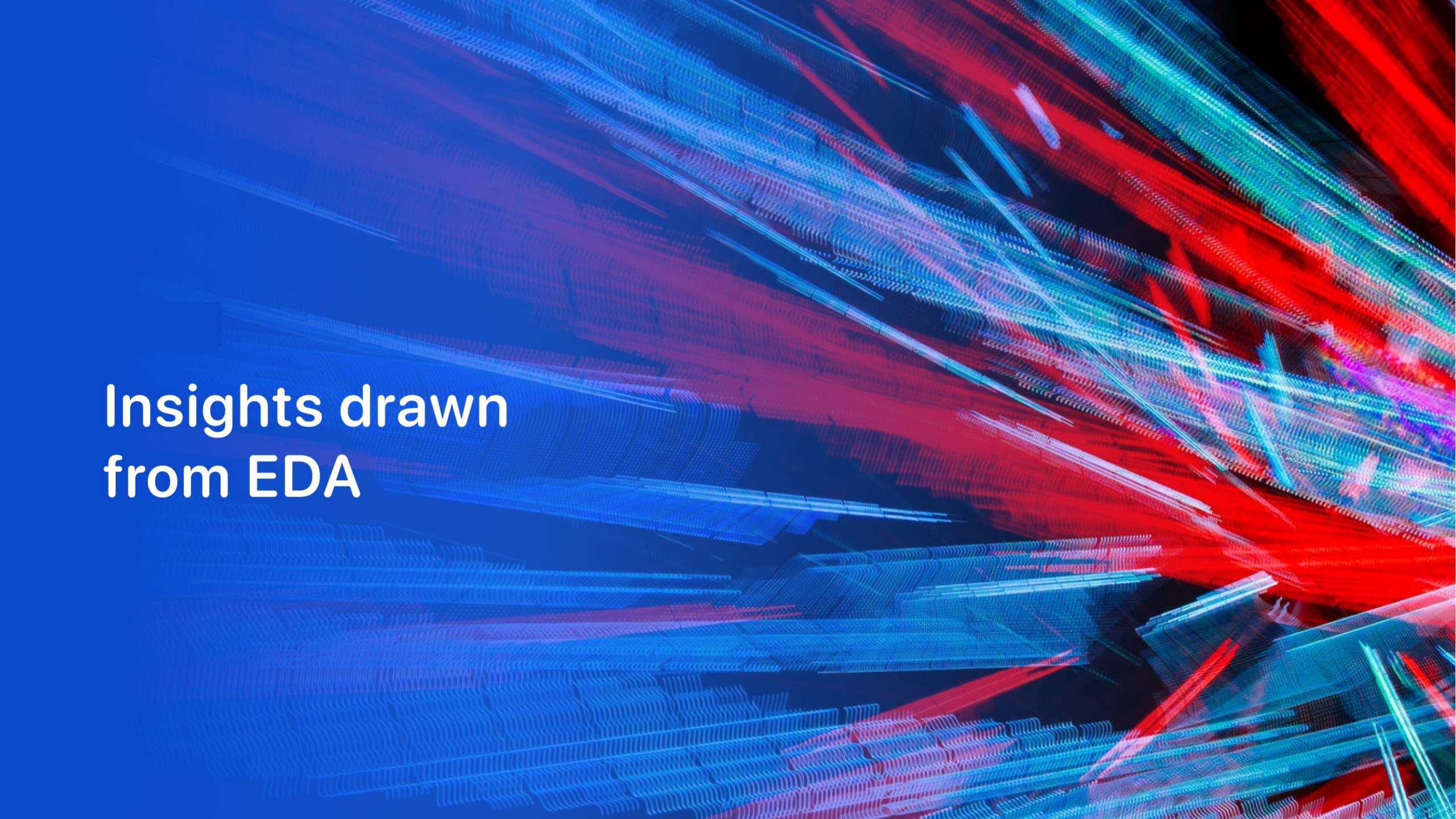
- GridSearchCV object is fitted with training input and output.
- Best Parameters are found using best_params_ method.
- Accuracy is checked using score method.
- Confusion matrix function is also created to visualize it for each classification method.
- Basis the training and testing accuracy best model is found.
- Here’s the link for github: https://github.com/rathore793/IBM_capstone_project/blob/main/predictive_analysis.ipynb

Results

The Project have various results basis SQL, visualization, folium, dash and predictive analysis, we will show each results in upcoming sections under below titles:

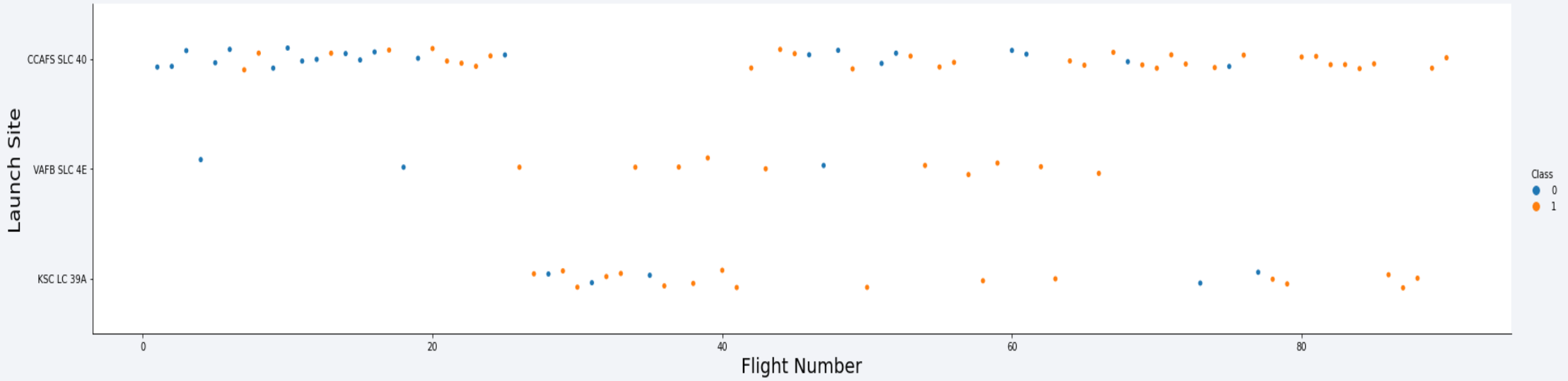
- Insights drawn from EDA
- Launch Sites proximities analysis
- Build Dashboard with plotly dash
- Predictive Analysis

Insights drawn from EDA



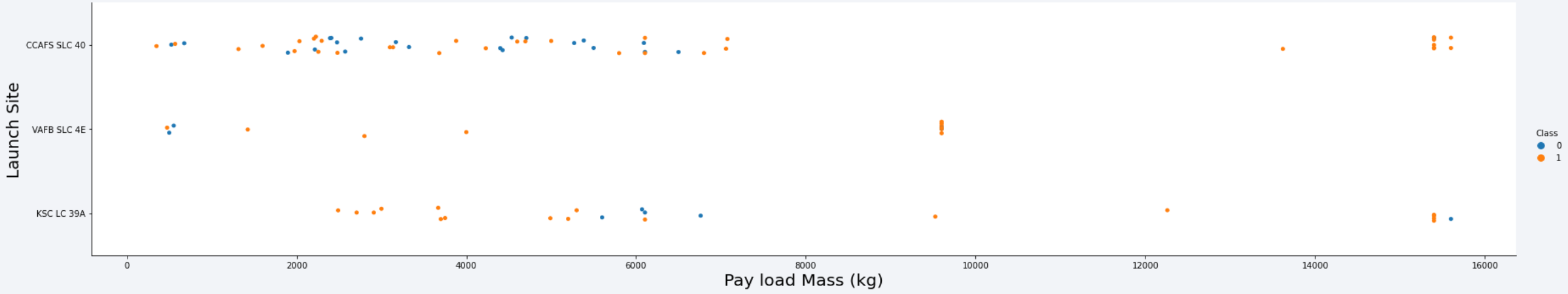
Flight Number vs. Launch Site

- The Relation between Flight Number and Launch Site shows that with increase in number of flight number the rate of success also increased.
- VAFB SLC 4E have lowest number of flight number but the success rate is high.



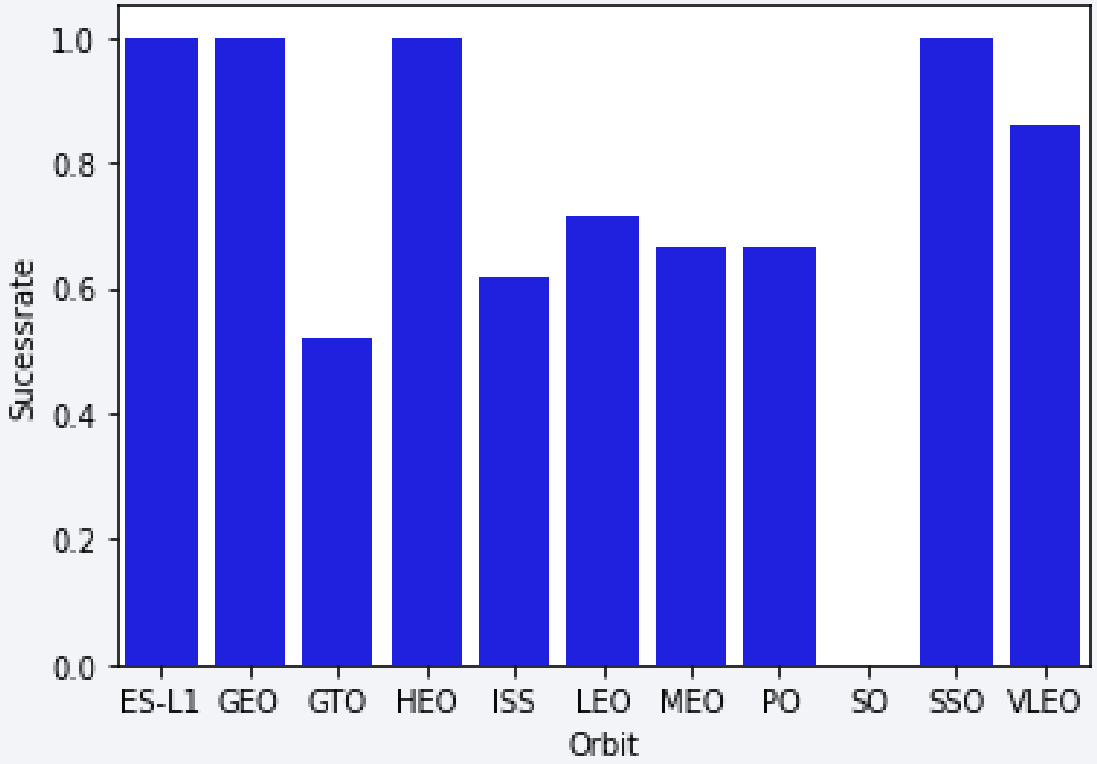
Payload vs. Launch Site

- The relation between Payload and Launch Site is shown below.
- Site CCAFS SLC 40 shows for payload mass greater than 14000kg the success rate is 100%.
- Site KSC LC 39A shows for payload mass less than 5000kg the success rate is 100%.
- Site VAFB SLC 4E shows for payload between 8-10000 kg the success rate is 100%.



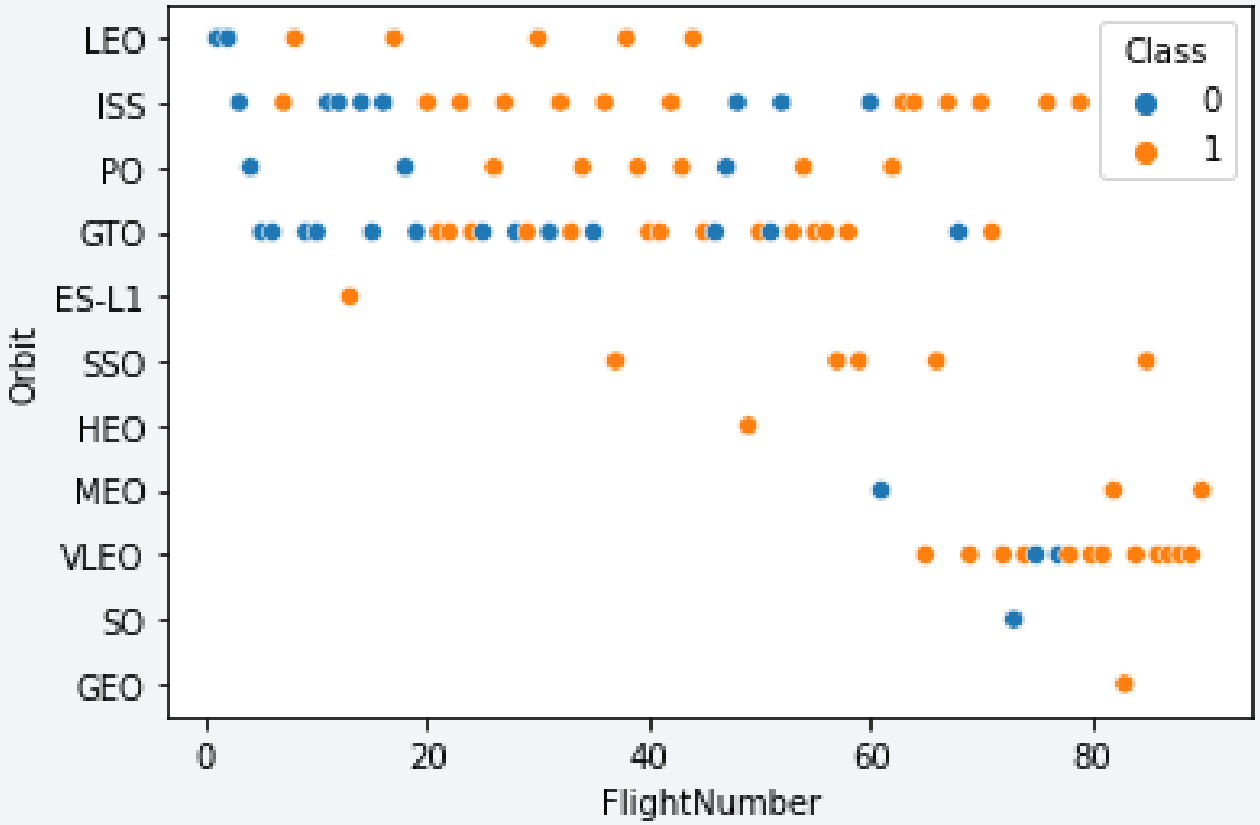
Success Rate vs. Orbit Type

- The Relationship shows that for the Orbit ES-L1, GEO, HEO and SSO the Success rate is 100%
- LEO and VLEO have Success rate between 70-80%.
- Rest all have success rate between 50 and 60%.



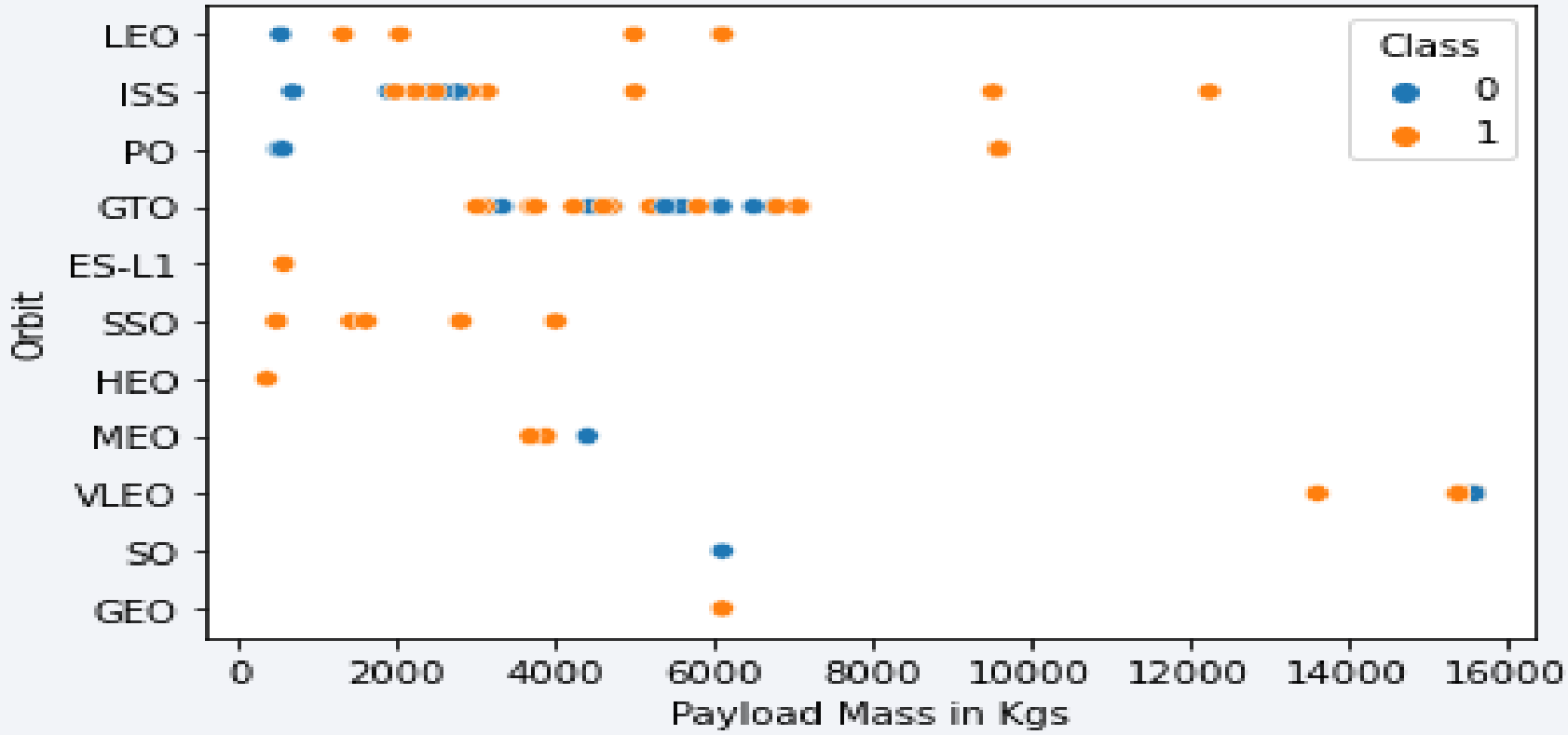
Flight Number vs. Orbit Type

- LEO shows that with increase in flight number the success rate is increasing.
- GEO do not have any relationship with Flight Number



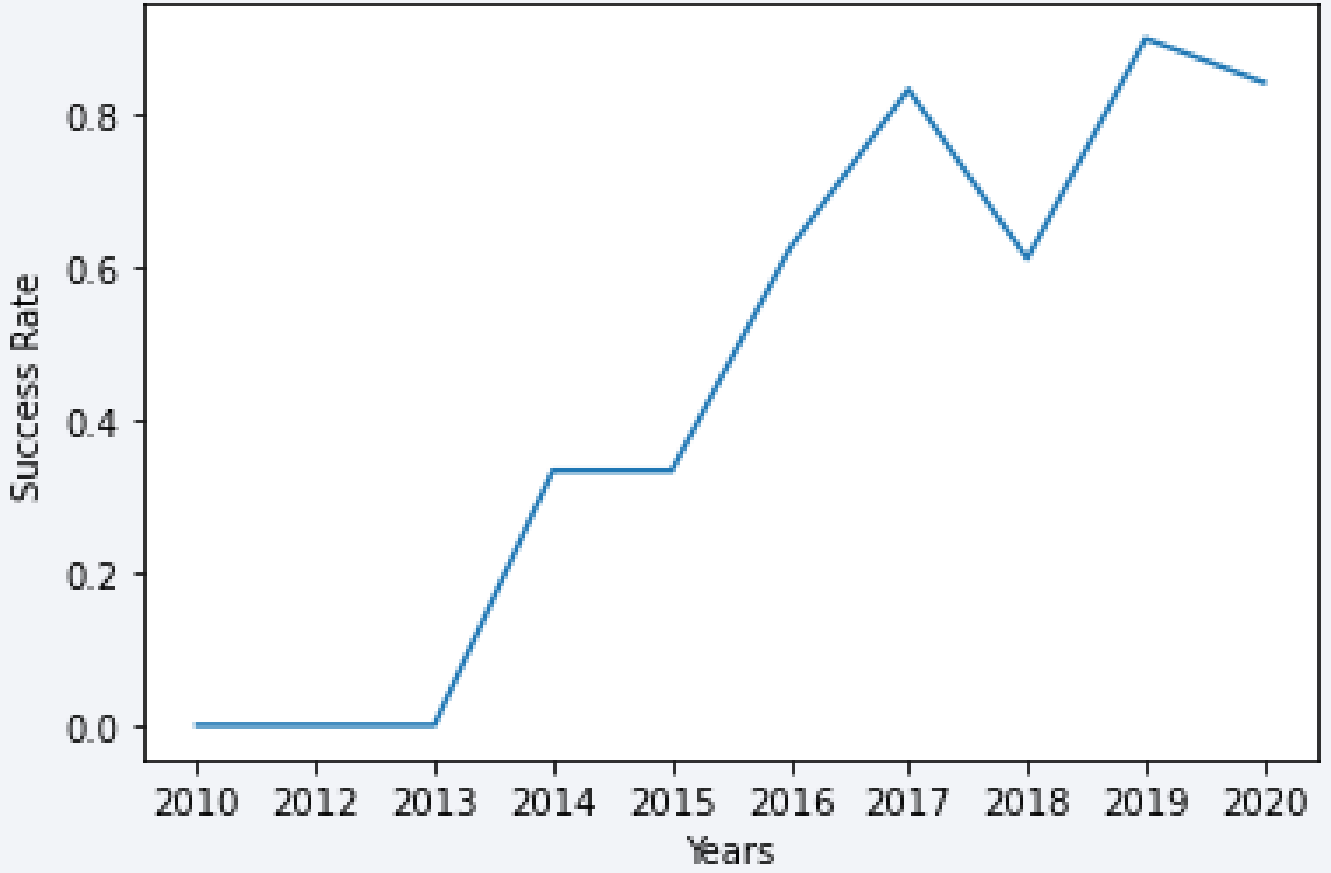
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS



Launch Success Yearly Trend

- With the help of screenshot we can verify that after year 2013 the success rate of launch is gradually increasing.



All Launch Site Names

- Query to find all distinct launch sites is given below:

```
%SQL SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL
```

- There are five 5 distinct launch sites

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

Launch Site Names Begin with 'CCA'

- Query to find all launch sites begin with ‘CCA’ is given below:

```
%SQL SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5
```

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

- Query to find Total Payload mass carried by NASA is below:

```
%SQL SELECT SUM(payload_mass__kg_) AS total_mass FROM SPACEXTBL WHERE payload LIKE '%CRS%'
```

- Total Mass carried by NASA payload is 111268 kgs

total_mass
111268

Average Payload Mass by F9 v1.1

- Query to find the average payload mass carried by booster version F9 v1.1 is below:

```
%SQL SELECT AVG(PAYLOAD_MASS__KG_) AS AVERAG_PAYLOAD_MASS FROM SPACEXTBL
WHERE BOOSTER_VERSION = 'F9 V1.1‘
```

- Average Payload mass carried by booster version F9 v1.1 is 2928 kg.

averag_payload_mass
2928

First Successful Ground Landing Date

- Query to find the first successful landing outcome on ground pad is given below:

```
%SQL SELECT MIN(DATE) FROM SPACEXTBL WHERE LANDING__OUTCOME = 'SUCCESS (GROUND PAD)‘
```

- Date of first Successful landing outcome is 22th December, 2015.

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List Query to find the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 is given below:
- %sql select booster_version from SPACEXTBL where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4001 and 5999

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

Total Number of Successful and Failure Mission Outcomes

- Query to find the total number of successful and failure mission outcomes is given below:

```
%sql select count(mission_outcome), mission_outcome from SPACEXTBL group by mission_outcome
```

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

Boosters Carried Maximum Payload

- Query to find the names of the booster which have carried the maximum payload mass is given below:

```
%sql select booster_version from SPACEXTBL
where payload_mass__kg_ = (select
max(payload_mass__kg_) from SPACEXTBL)
```

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

2015 Launch Records

- Query to find the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 is given below:

```
%sql select booster_version, launch_site from SPACEXTBL where
landing__outcome ='Failure (drone ship)' and year(DATE) = 2015
```

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 to 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 is given below:

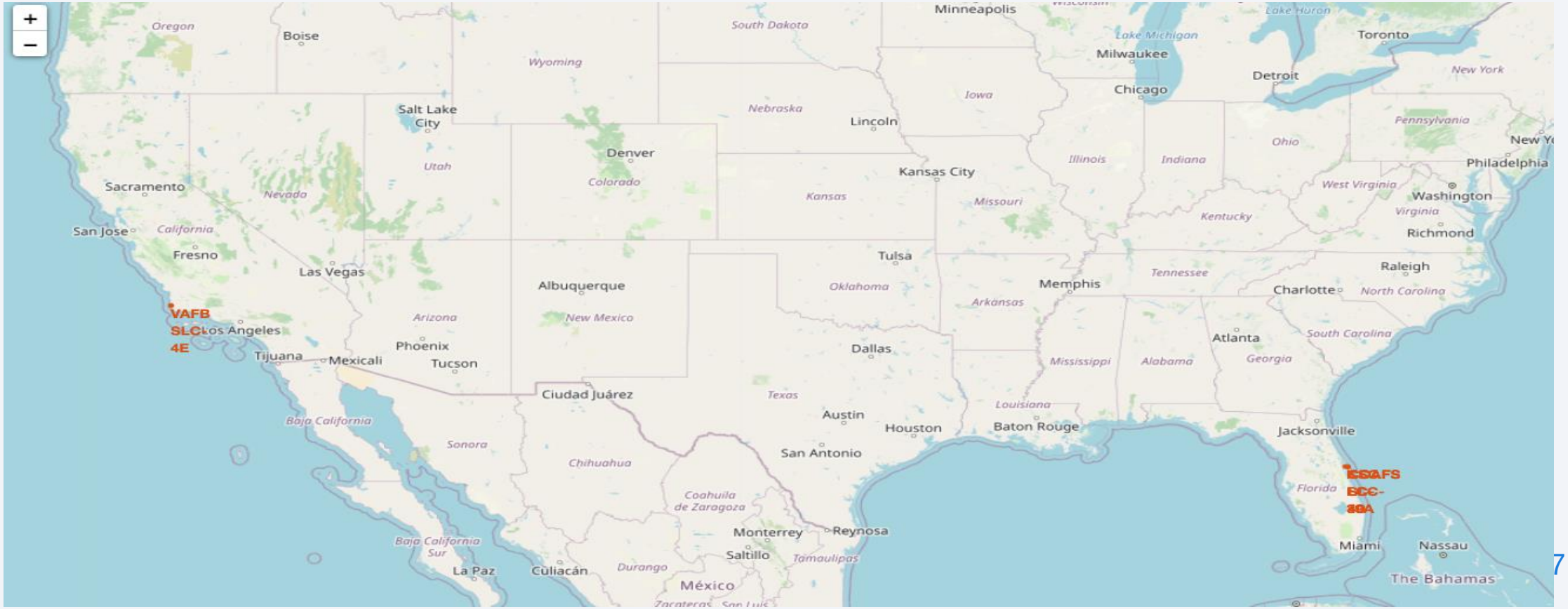
```
%sql select count(landing__outcome),
landing__outcome from SPACEXTBL where date
between '2010-06-04' and '2017-03-20' group by
landing__outcome order by
count(landing__outcome) desc
```

1	landing__outcome
10	No attempt
5	Failure (drone ship)
5	Success (drone ship)
3	Controlled (ocean)
3	Success (ground pad)
2	Failure (parachute)
2	Uncontrolled (ocean)
1	Precluded (drone ship)

Launch Sites Proximities Analysis

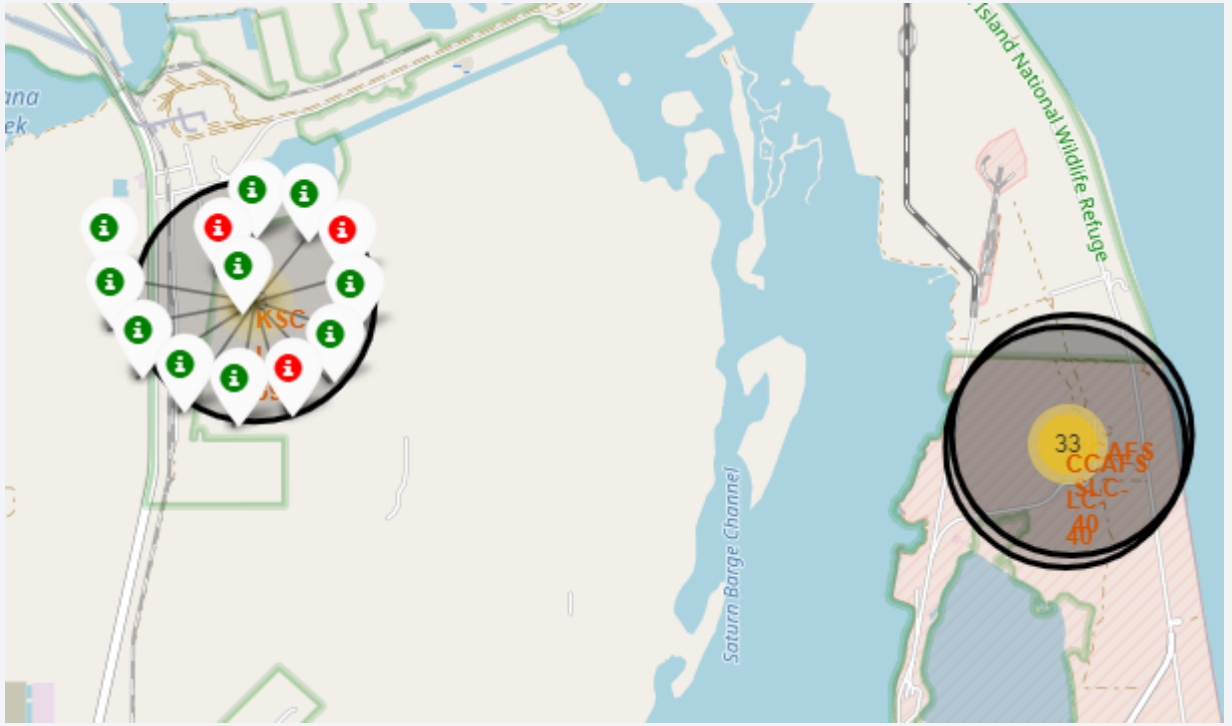
All Launch Sites on Map

- 3 out of 4 sites are in same state, all sites are near to coastline.



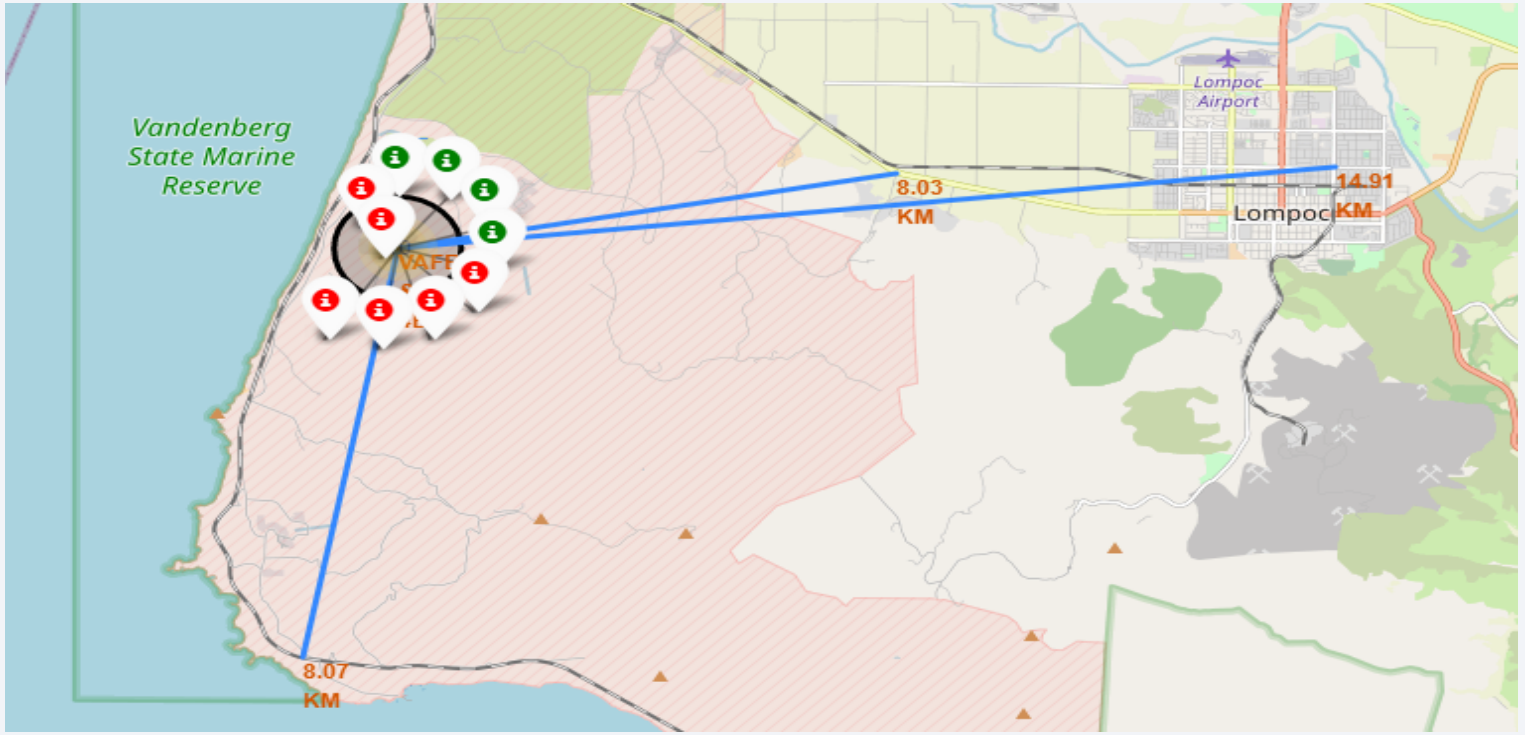
Success/Failed launches on each site on Map

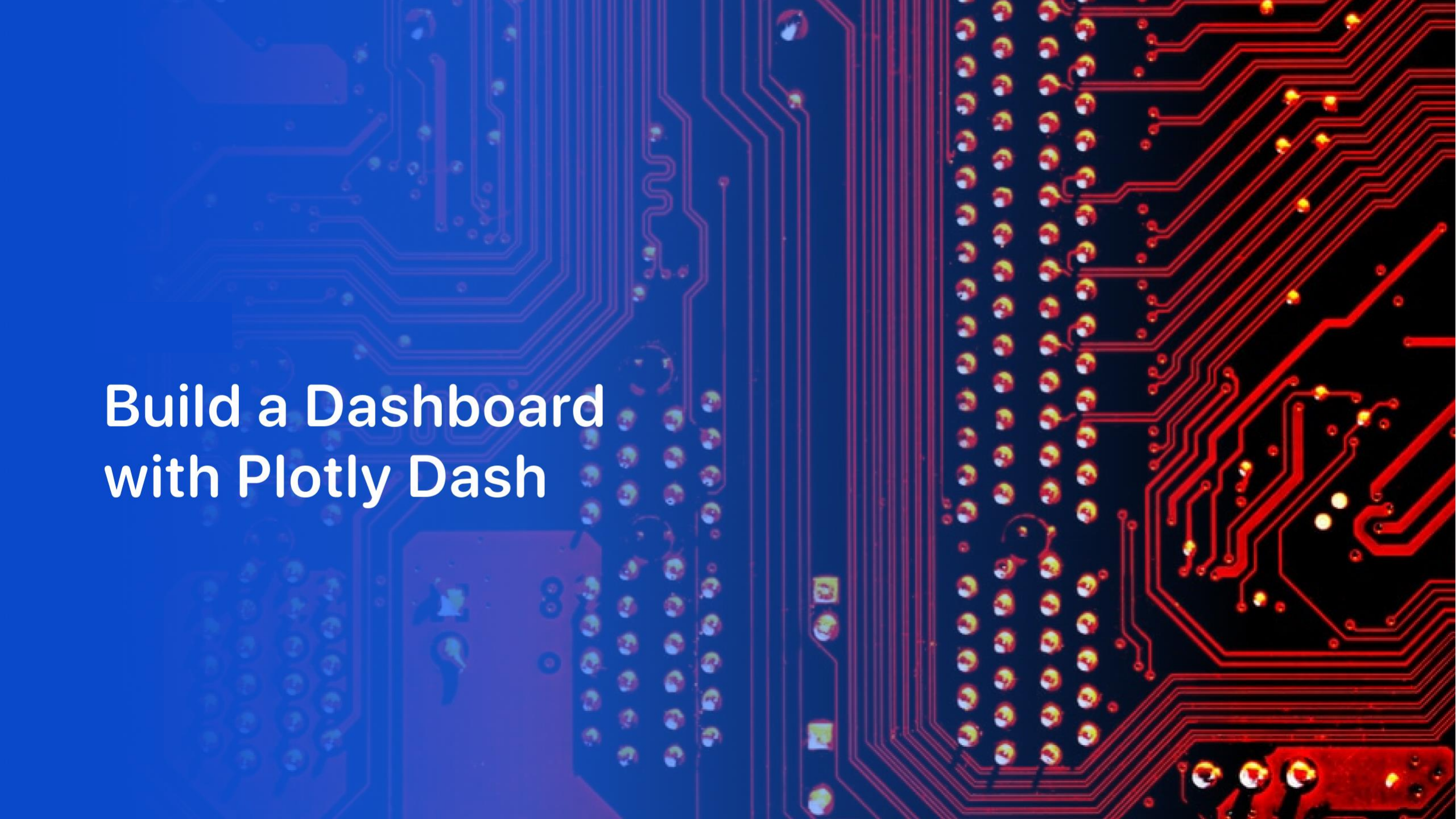
- Marker cluster are made of Failed and Success outcomes, where red marker mean failed and green mean successful outcome.



Proximities Distance from Launch Site

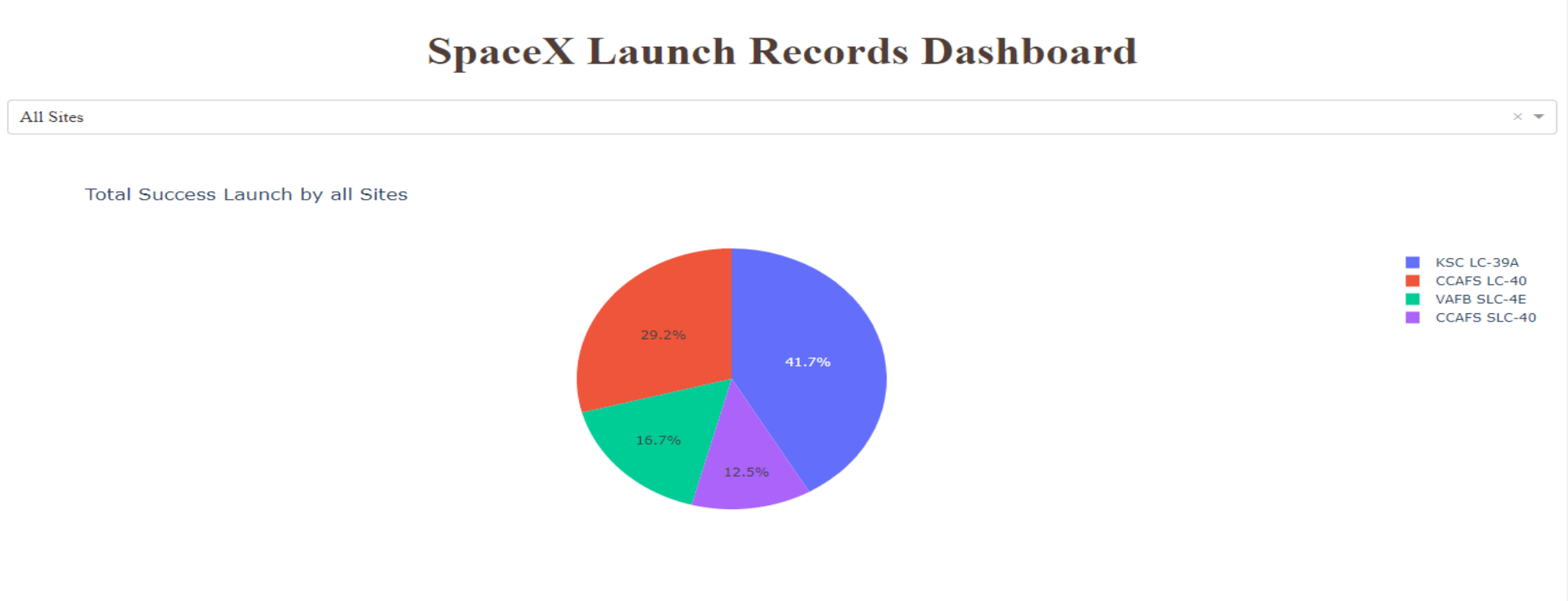
- Distance is calculated between proximities and Launch Site and the line is drawn using Polyline.
- All the Launch sites are kept at a safe distance from all the proximities.




Build a Dashboard
with Plotly Dash

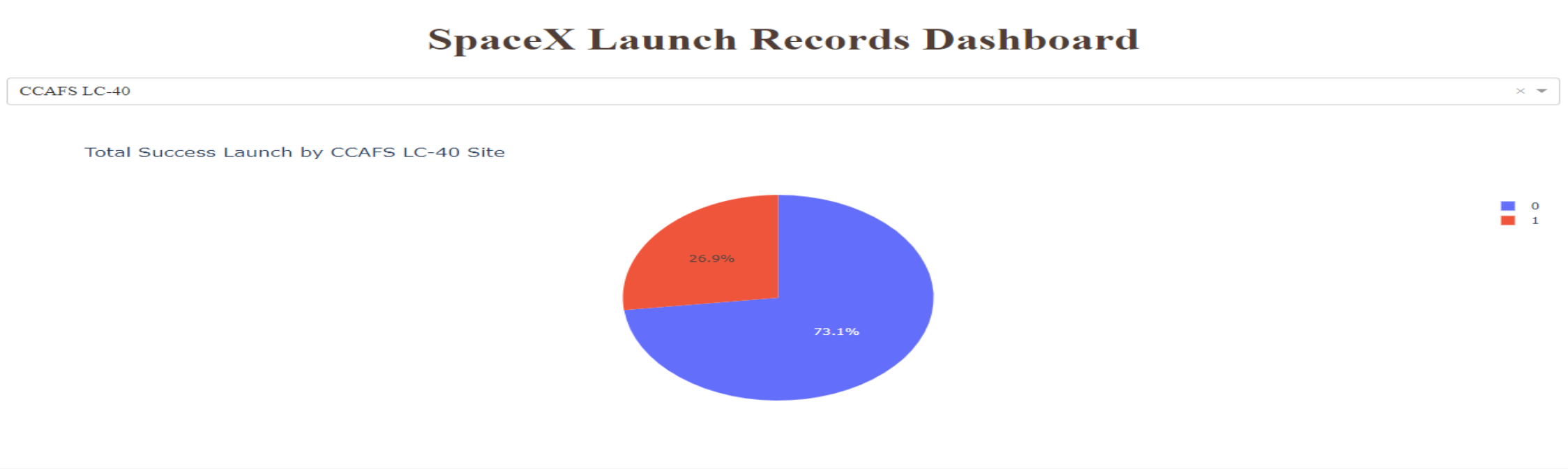
Launch Success Count for All Sites

- Launch Site KSC LC 39A is having most Success count, with 41.7% Success count, followed by CCAFS SLC 40 with 29.2%, VAFB SLC 4E with 16.7%.



Launch Site With Highest Success Ratio

- CCAFS LC-40 is having highest Success Ratio among all the other launch site with 41.7%.
- The success rate of CCAFS LC-40 is only 26.9%.



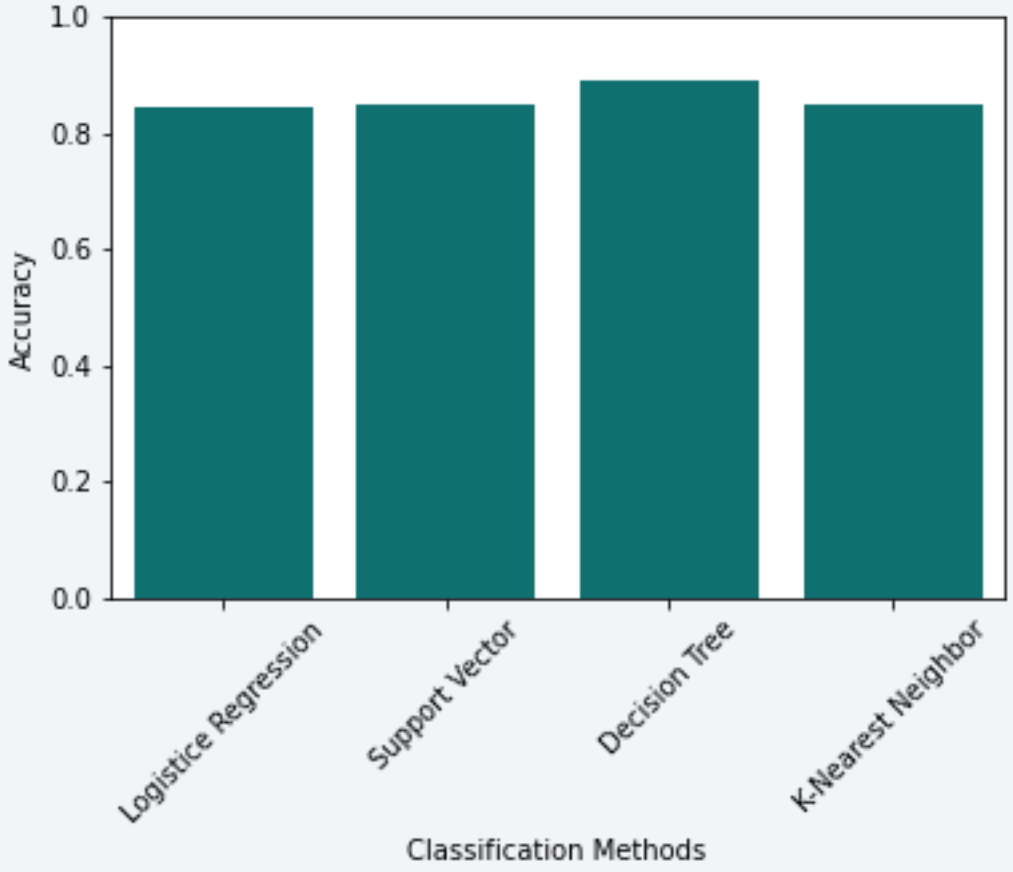
Payload vs Launch Outcome for all sites

- Booster version FT is better option for the Payload mass between 2000-5000 kg.
- Booster version B4 is good fit for payload mass less 4000 kg.

Predictive Analysis (Classification)

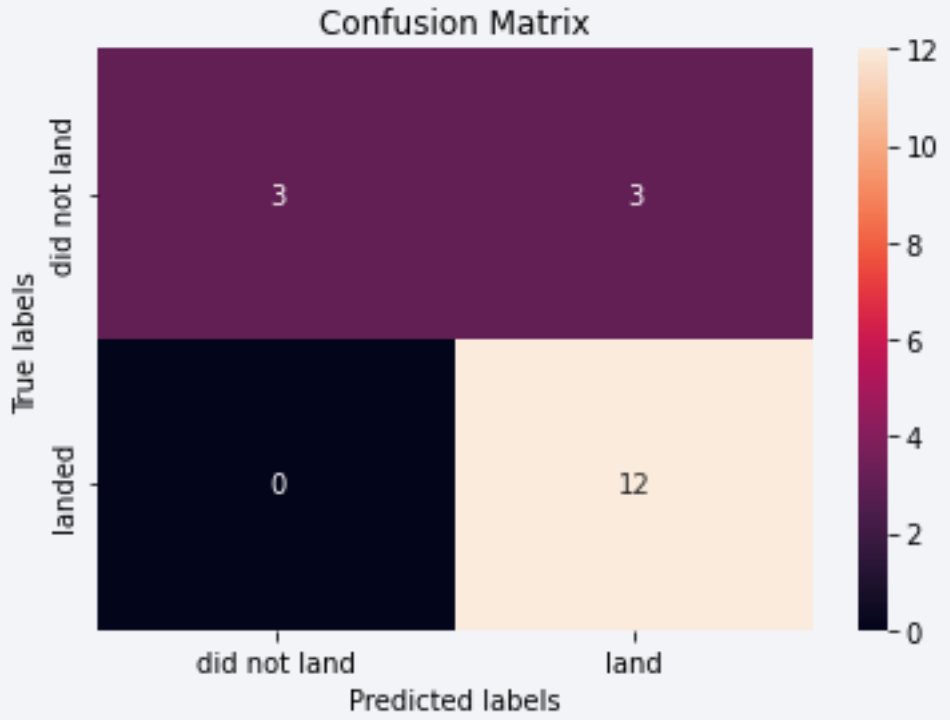
Classification Accuracy

- Best Score accuracy is highest for Decision Tree which is 88.9%.
- Testing Accuracy of all models is found to be equal.
- Training Accuracy of SVM model is highest with 88%



Confusion Matrix

- Confusion Matrix of Decision Tree is shown below with some observations.
- Out of 18 samples 12 times it was predicted successful land which actually landed.
- 3 time predicted landed however it did not landed.
- 3 times it was predicted did not land and it did not landed.



Conclusions

- The Project help us understanding the correlation of outcome of Launch on various factors including site location, booster version, payload mass etc.
- Site CCAFS SLC 40 is good option for payload mass greater than 14000kg, whereas site KSC LC 39A is better payload mass less than 5000kg.
- Site KSC LC 39A is having the highest success ration among all the sites.
- Booster version FT is better option for the Payload mass between 2000-5000 kg, whereas Booster version B4 is good fit for payload mass less 4000 kg.
- The predictive model produced by decision tree algorithm performed the best among the 4 machine learning algorithms employed with 83% out of sample accuracy.

Thank you!