IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Shraddha Patil
13th September 2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive Analytics result

Introduction

- Project background and context

  Space X 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 Space X can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against space X for a rocket launch. This goal of the project is to create a machine learning pipeline to predict if the first stage will land successfully.

- Problems you want to find answers

  - What factors determine if the rocket will land successfully?
  - The interaction amongst various features that determine the success rate of a successful landing.
  - What operating conditions needs to be in place to ensure a successful landing program.

Section 1

Methodology

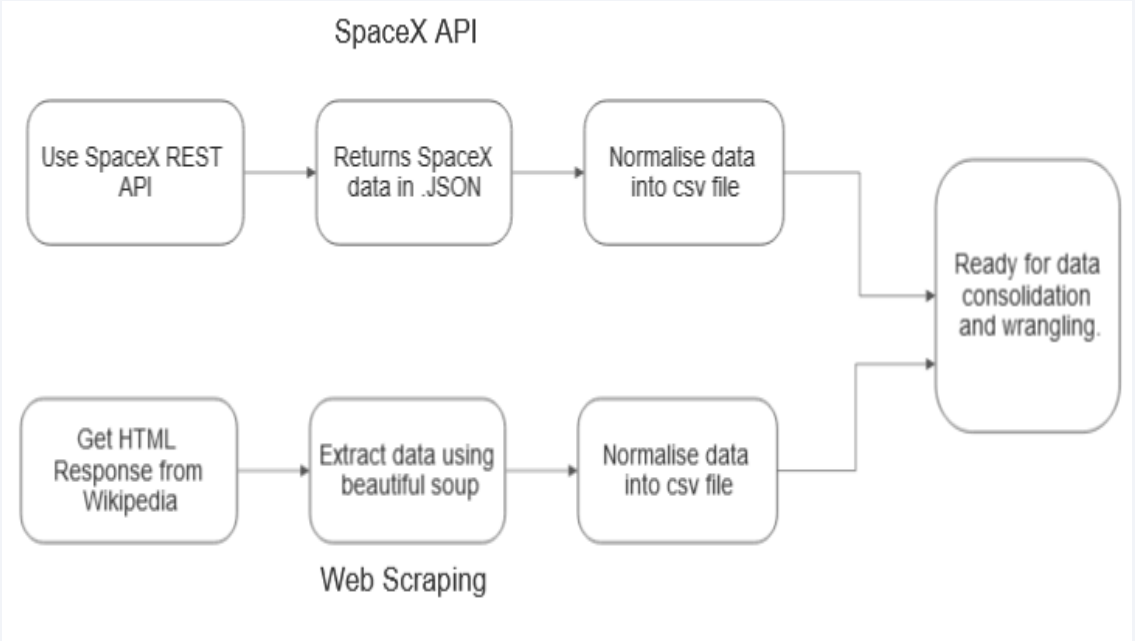
Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX API and web scraping from Wikipedia.
- Perform data wrangling
  - One hot encoding data fields for machine learning and dropping irrelevant columns(Transforming data and Machine learning)
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - LR, KNN, SVM, DT models have been built and evaluated for best classifier.

Data Collection

- The following datasets was collected:
  - SpaceX launch data that is gathered from the SpaceX REST API.
  - This API will give us data about launches, information about the rocket used for launch, payload delivered, launch specifications, landing specification and landing outcome.
  - The SpaceX REST API endpoints, or URL starts with apispacexdata.com/v4/.
  - Another popular data source for obtaining Falcon 9 Launch data is web scrapping Wikipedia using BeautifulSoup.

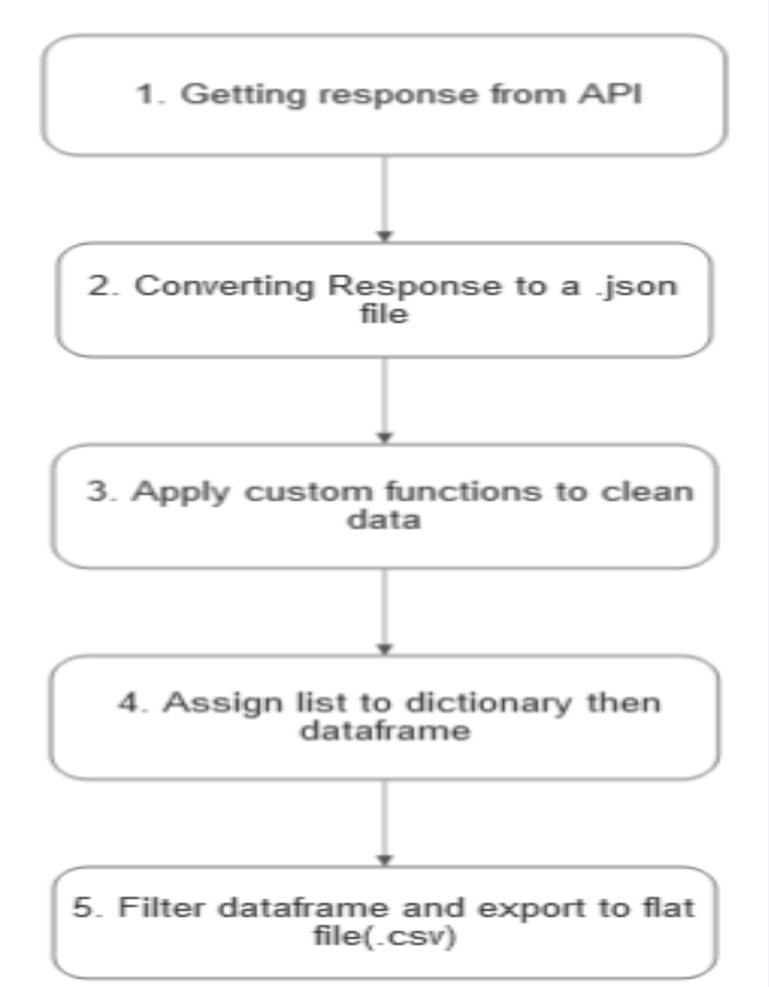


Data Collection – SpaceX API

- Data collection with SpaceX REST calls

- The GitHub link to the notebook is

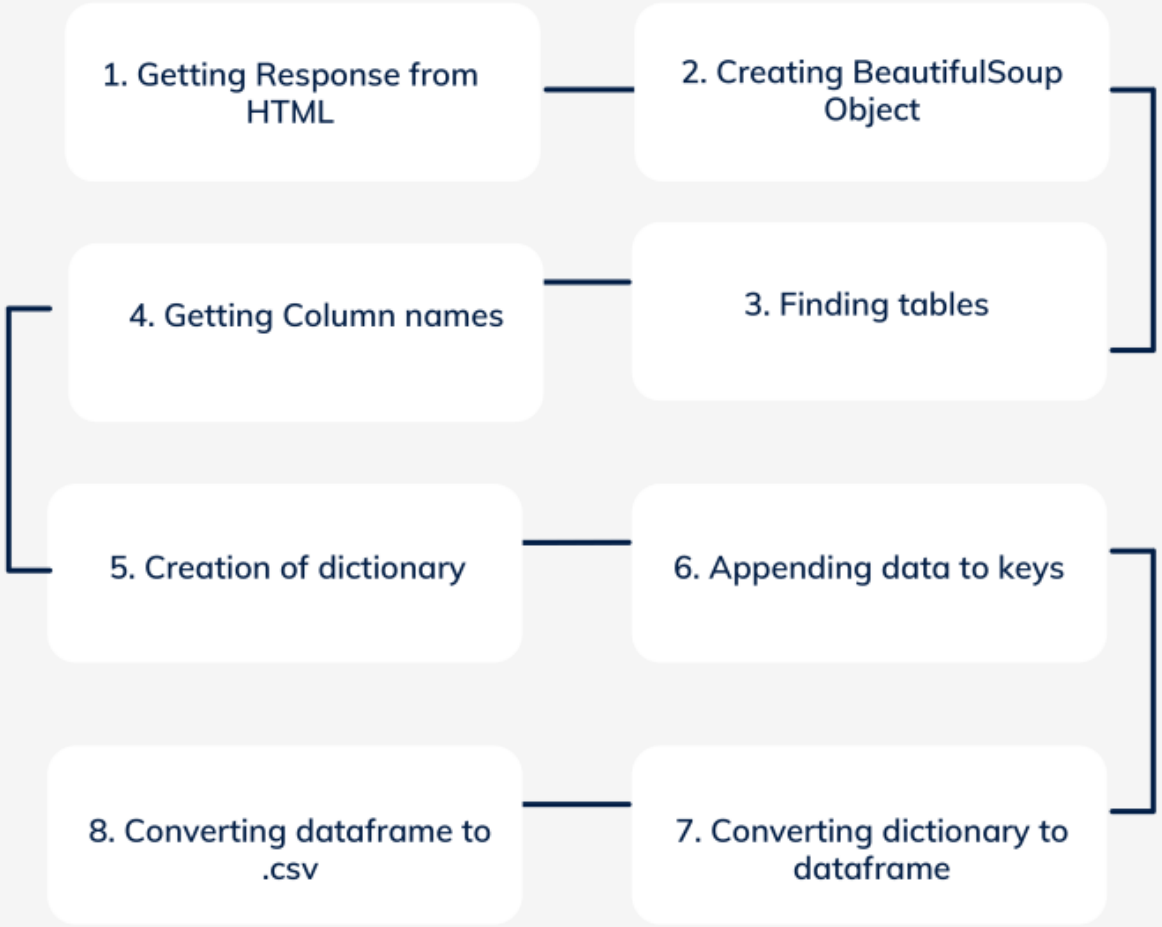
https://github.com/ShraddhaPatil15/Applied-Data-Science-Capstone-Project/blob/main/Collecting%20the%20data.ipynb



Data Collection - Scraping

- We applied web scrapping to webscrap Falcon 9 launch records with BeautifulSoup
- We parsed the table and converted it into a pandas dataframe.
- The GitHub link to the notebook is

https://github.com/ShraddhaPatil15/Applied-Data-Science-Capstone-Project/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb



Data Wrangling

- Data wrangling is the process of cleaning and unifying messy and complex data sets for easy access and analysis.
- Here we mainly convert those outcomes into Training labels with landing outcomes where successful = 1 and failure = 0.
- We will first calculate the number of launches on each site, then calculate the number and occurrence of mission outcome per orbit type.
- created landing outcome label from outcome column and exported the results to .csv file.
- The GitHub link to the notebook is

https://github.com/ShraddhaPatil15/Applied-Data-Science-Capstone-Project/blob/main/Data%20wrangling.ipynb

Data Wrangling

1. Calculate number of launches at each side

```
df.LaunchSite.value_counts()
```

```
CCAFS SLC 40     55
KSC LC 39A       22
VAFB SLC 4E      13
```

2. Calculate number and occurrence of each orbit

```
df.Orbit.value_counts()
```

```
GTO      27
ISS      21
VLEO     14
PO        9
LEO       7
SSO       5
MEO       3
ES-L1     1
HEO       1
SO        1
GEO       1
```

3. Calculate number and occurrence of mission outcome per orbit

```
landing_outcomes = df.Outcome.value_counts()
landing_outcomes
```

```
True ASDS      41
None None      19
True RTLS      14
False ASDS      6
True Ocean      5
False Ocean     2
None ASDS       2
False RTLS      1
```

4. Create landing outcome label from Outcome column

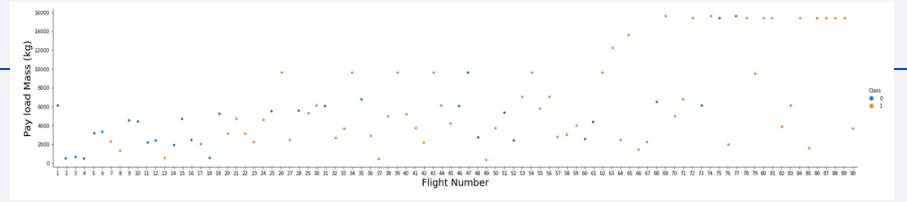
```
# landing_class = 0 if bad_outcome
df['class'] = df['Ou'].apply(lambda x: 'value if condition is met' if x condition else 'value if
```

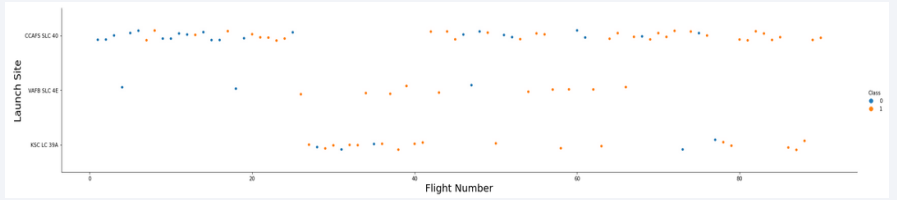
5. Export dataset as .csv

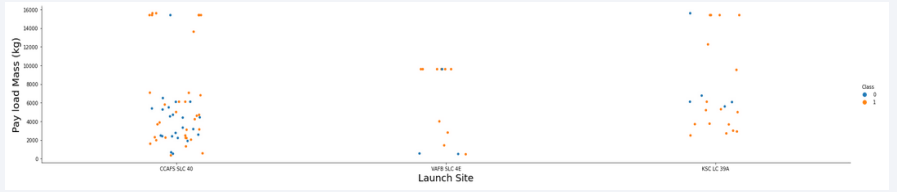
```
df.to_csv("dataset_part\_2.csv", index=False)
```

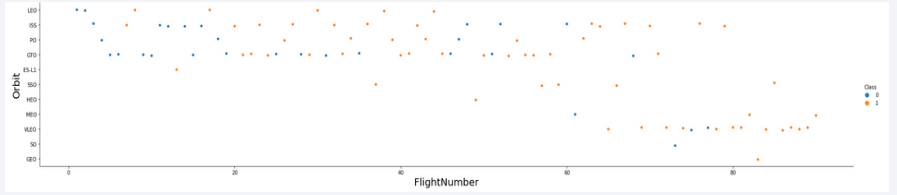
EDA with Data Visualization

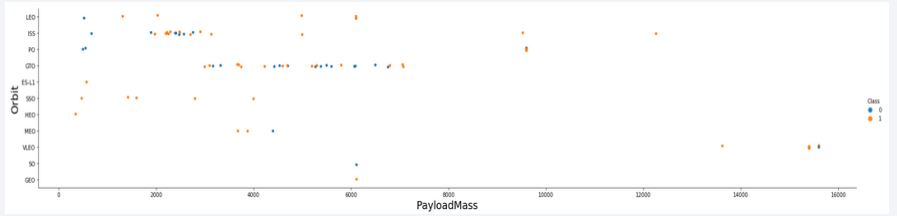
- Exploratory Data Analysis performed on variables Flight Number, Payload Mass, Launch Site, Orbit, Class and Year.
- The GitHub link to the notebook is https://github.com/ShraddhaPatil15/Applied-Data-Science-Capstone Project/blob/main/EDA%20with%20Visualization%20lab.ipynb
- Scatter Graphs Drawn:
  - Payload and Flight Number
  - Flight Number and launch Site
  - Payload and Launch Side
  - Flight Number and Orbit Type
  - Payload and Orbit Type











EDA with Data Visualization

- Bar Graphs Drawn:

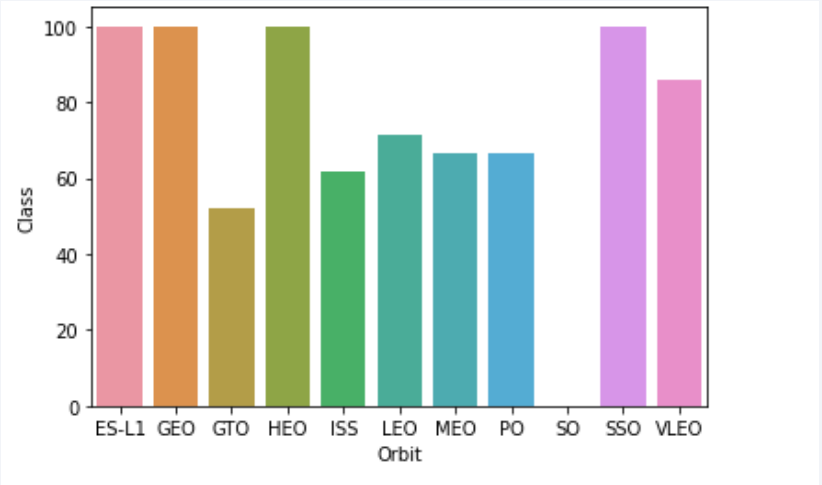
  Success Rate Vs. Orbit Type

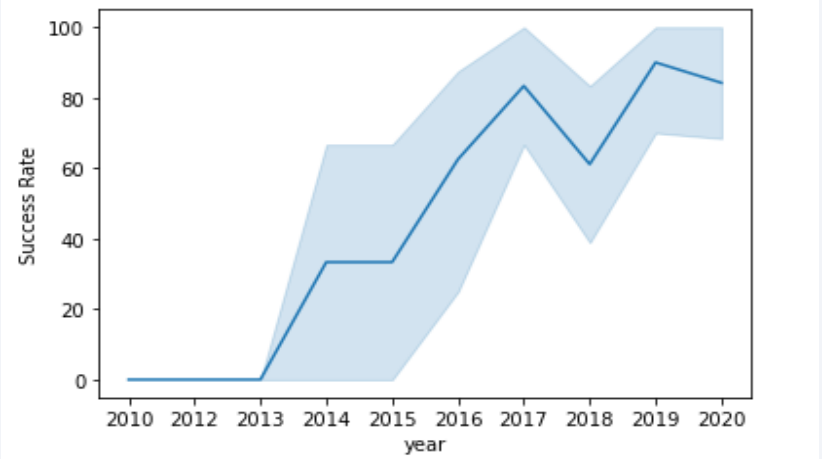
  Bar graphs are easiest to interpret a relation between attributes. Via this bar graph we can easily determine which orbits have highest probability of success.

- Line Graphs Drawn:

  Launch Success Yearly Trend

  Line graph are useful in showing trends clearly and aid in prediction for future.





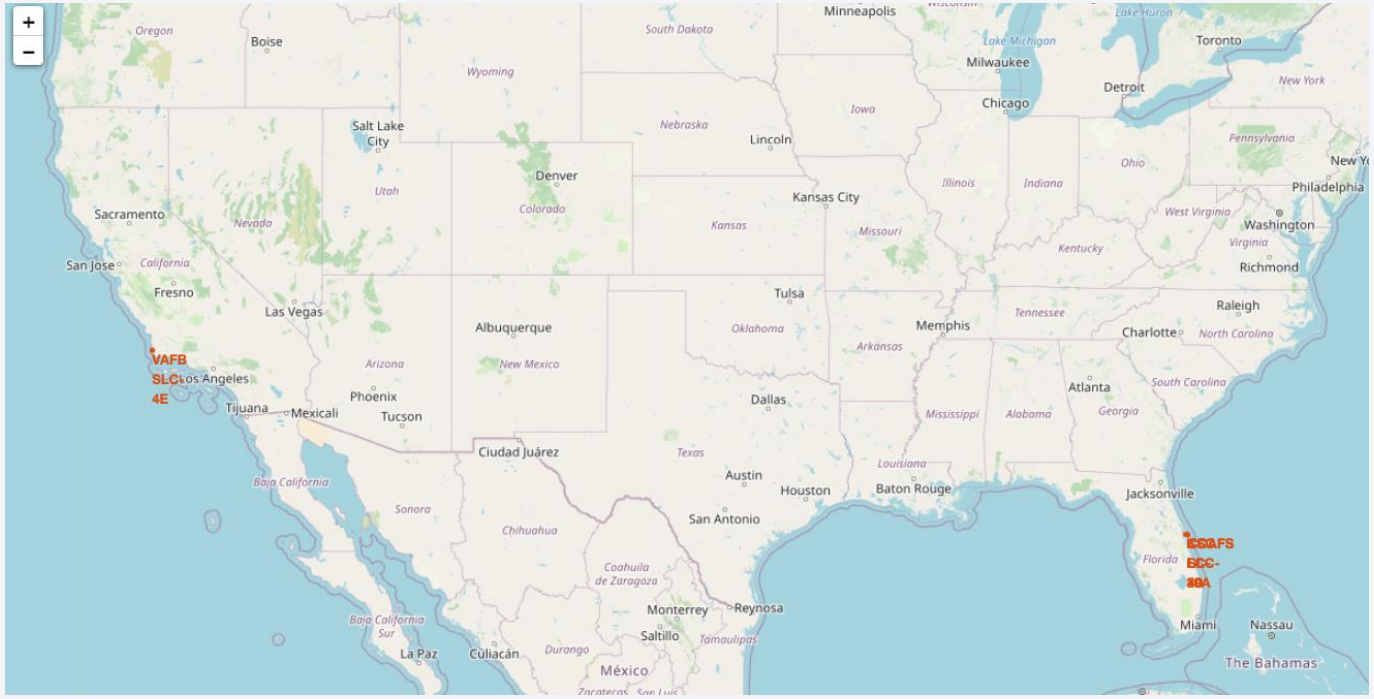
EDA with SQL

We performed SQL queries to gather information from given dataset:

- **Display the names of the unique launch sites in the space mission**
- **Display 5 records where launch sites begin with the string 'CCA'**
- **Display the total payload mass carried by boosters launched by NASA (CRS)**
- **Display average payload mass carried by booster version F9 v1.1**
- **List the date when the first successful landing outcome in ground pad was achieved.**
- **List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000**
- **List the total number of successful and failure mission outcomes**
- **List the names of the booster_versions which have carried the maximum payload mass.**
- **List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015**
- **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**

The GitHub link to the notebook is https://github.com/ShraddhaPatil15/Applied-Data-Science-Capstone-Project/blob/main/EDA%20with%20SQL%20lab.ipynb

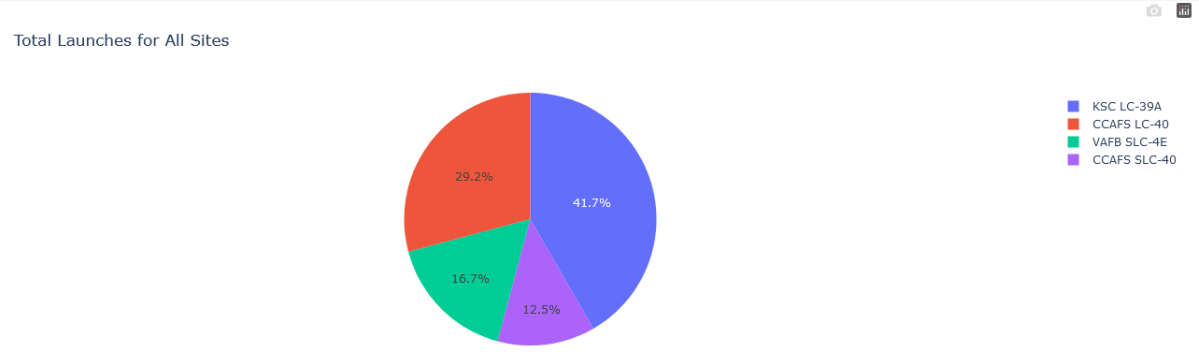
Build an Interactive Map with Folium

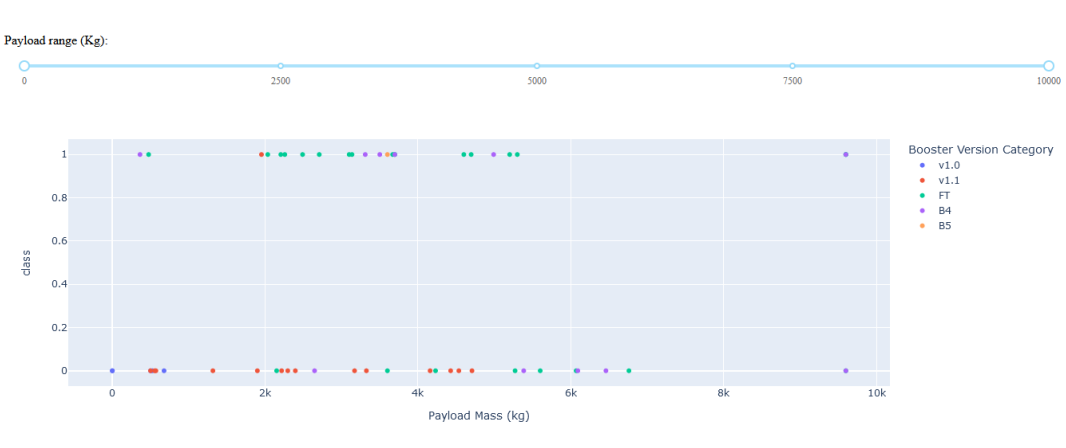


- Map markers have been added to the map with aim to finding an optimal location for building a launch site
- The GitHub link to the notebook is https://github.com/ShraddhaPatil15/Applied-Data-Science-Capstone-Project/blob/main/the%20Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

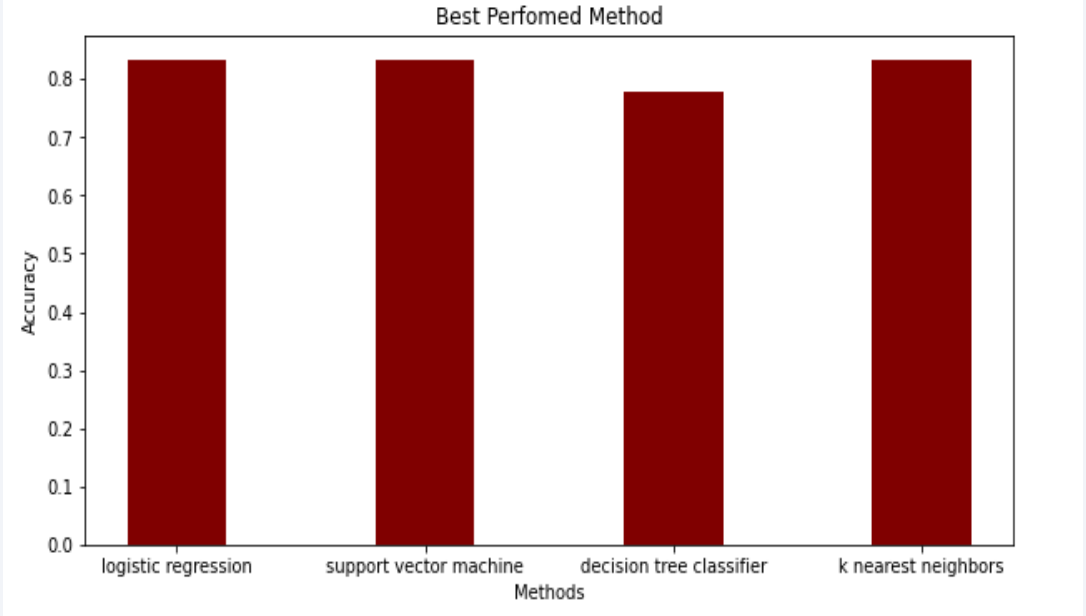
- We built an interactive dashboard with Plotly dash
- We plotted pie charts showing the total launches by a certain sites
- We plotted scatter graph showing the relationship with Outcome and Payload Mass (Kg) for the different booster version.
- The GitHub link to the notebook is https://github.com/ShraddhaPatil15/Applied-Data-Science-Capstone-Project/blob/main/spacex_dash_app.py





Predictive Analysis (Classification)

- We loaded the data using numpy and pandas, transformed the data, split our data into training and testing.
- We built different machine learning models and tune different hyper parameters using GridSearchCV.
- We used accuracy as the metric for our model, improved the model using feature engineering and algorithm tuning.
- We found the best performing classification model.
- The GitHub link to the notebook is https://github.com/ShraddhaPatil15/Applied-Data-Science-Capstone-Project/blob/main/SpaceX_Machine%20Learning%20Prediction.ipynb

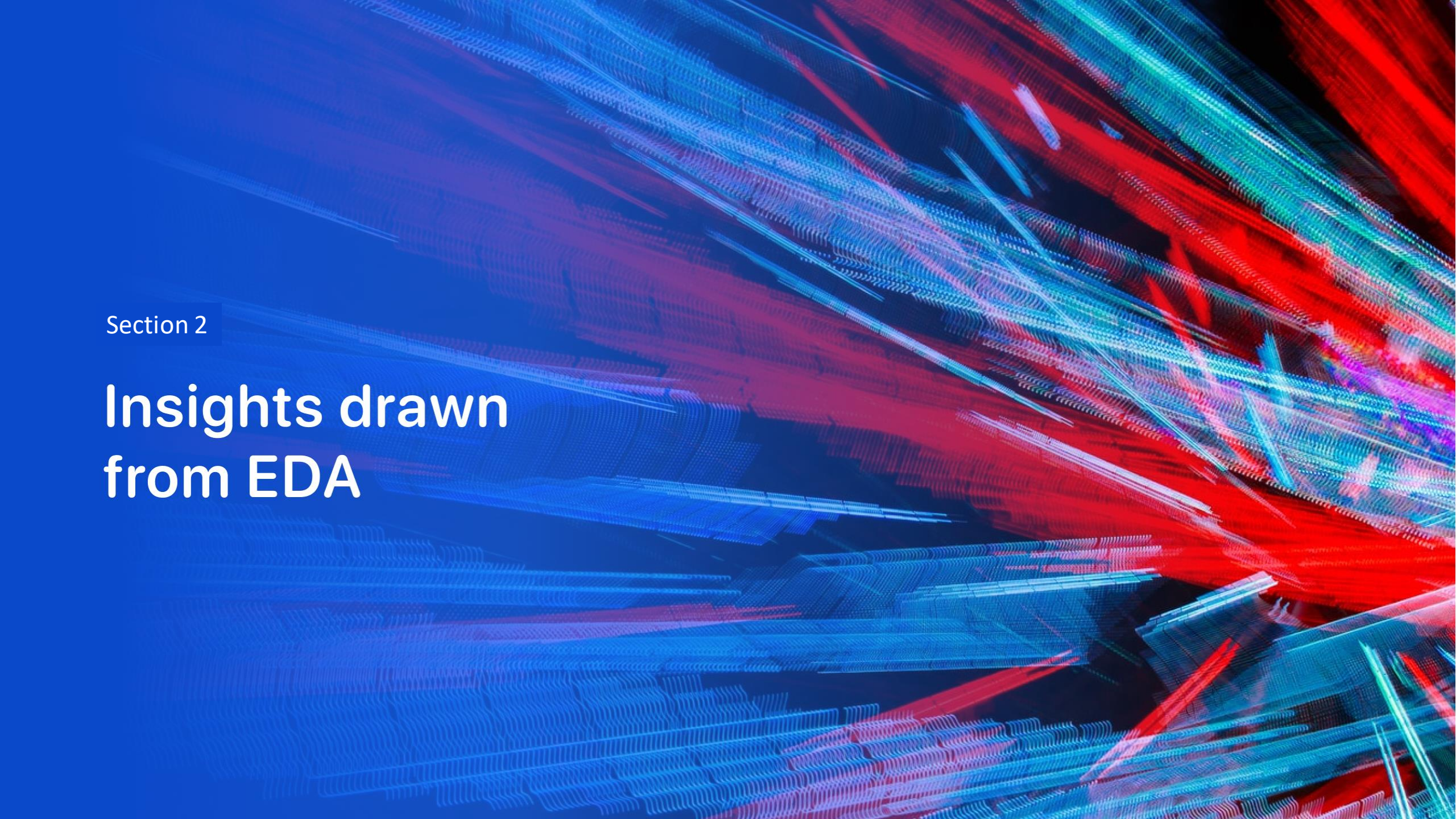


Results

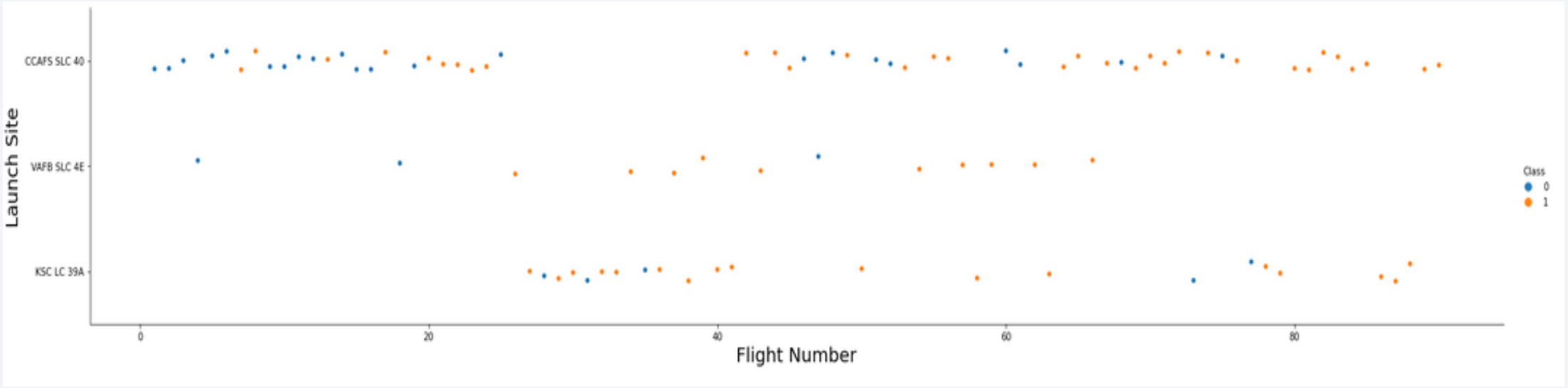
- The SVM, KNN and Logistic Regression models are the best in terms of prediction accuracy for this dataset.
- Low weighted payloads performs better than the heavier payloads.
- The success rates for SpaceX launches is directly proportional time in years they will eventually perfect the launches.
- KSC LC 39A had the most Successful launches from all the sites.
- Orbit GEO, HEO, SSO, ES L1 has the best Success Rate.

Section 2

Insights drawn from EDA

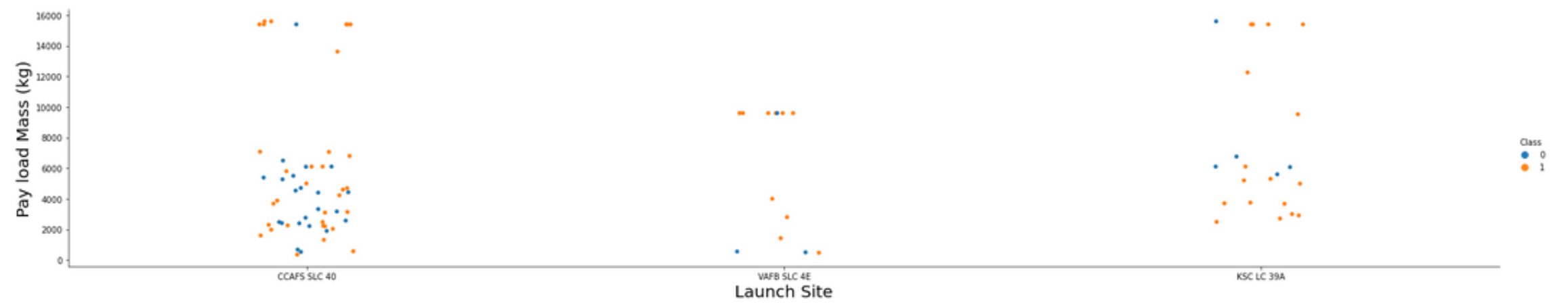


Flight Number vs. Launch Site



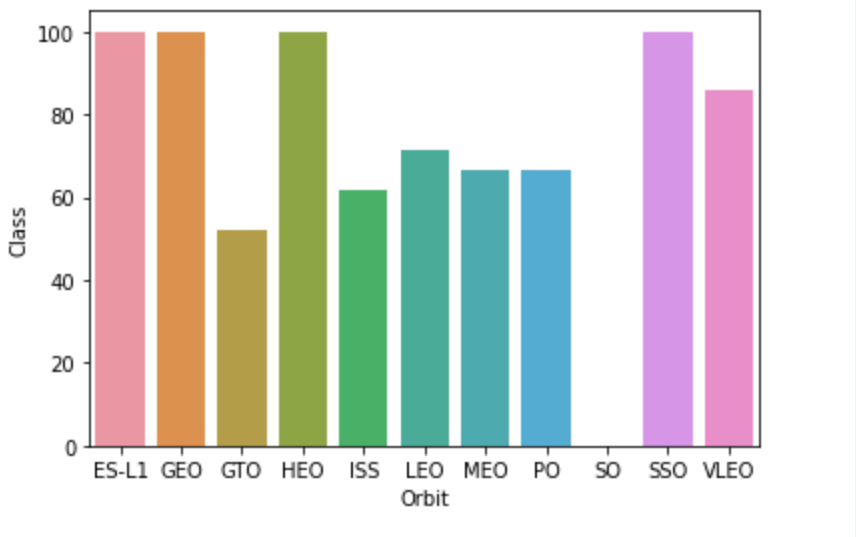
- Launches from the site of CCAFS SLC 40 are significantly higher than launches from other sites.
- With the Increase of flight number the success rate is increasing as well in the Launch sites

Payload vs. Launch Site



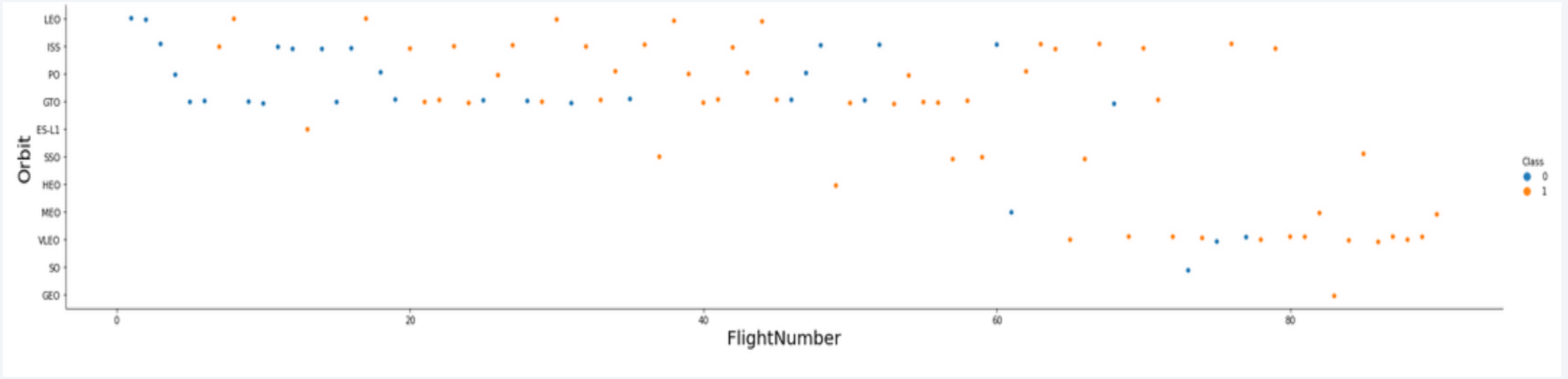
- The majority of Pay Loads with lower Mass have been launched from CCAFS SLC 40.

Success Rate vs. Orbit Type



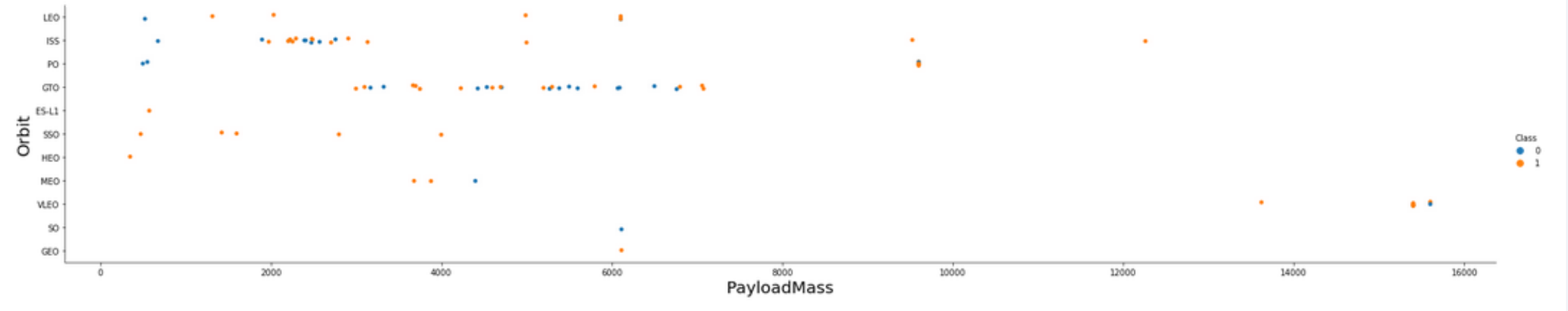
- The orbit types GEO, HEO, SSO, ES L1 are among the highest success rate.

Flight Number vs. Orbit Type



- It's hard to tell anything here, but we can say there is no actual relationship between flight number and orbit.

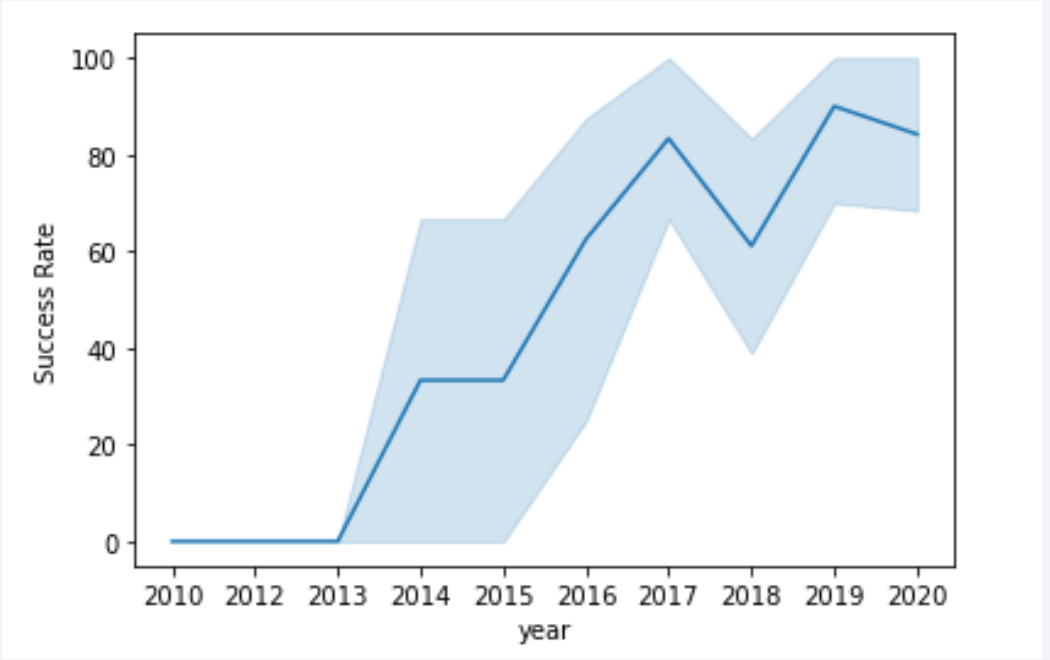
Payload vs. Orbit Type



- First thing to see is how the Pay load mass between 2000 and 3000 is affecting ISS.
- Similarly Pay load mass between 3000 and 7000 is affecting GTO.

Launch Success Yearly Trend

- Since the year 2013, there was a massive increase in success rate. However, It dropped little in 2018 but later it got stronger than before.



All Launch Site Names

- We can get the unique values by using "DISTINCT"

```
%sql select DISTINCT LAUNCH_SITE from SPACEXTBL;
```

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

Launch Site Names Begin with 'CCA'

- We can get only 5 rows by using "LIMIT"

```
%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

- We can get the sum of all values by using "SUM"

```
%sql select sum(payload_mass__kg_) as sum from SPACEXTBL where customer like 'NASA (CRS)'
```

SUM
45596

Average Payload Mass by F9 v1.1

- We can get the average of all values by using "AVG"

```
%sql select avg(payload_mass__kg_) as Average from SPACEXTBL where booster_version like 'F9 v1.1%'
```

average
2534

First Successful Ground Landing Date

- We can get the first successful ground landing date by using "MIN", Because first date is same with the minimum date.

```
%sql select min(date) as Date from SPACEXTBL where mission_outcome like 'Success'
```

DATE
2010-06-04

Successful Drone Ship Landing with Payload between 4000 and 6000

- We used the WHERE clause to filter for boosters which have successfully landed on drone ship and applied the AND condition to determine successful landing with payload mass greater than 4000 but less than 6000

```
%sql SELECT BOOSTER_VERSION FROM SPACEX WHERE LANDING__OUTCOME = 'Success (drone ship)' \
AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;
```

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

Total Number of Successful and Failure Mission Outcomes

- We used “count” to filter for WHERE Mission Outcome was a success or a failure.

```
%sql SELECT mission_outcome, count(*) as Count FROM SPACEXTBL GROUP by mission_outcome ORDER BY mission_outcome
```

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

Boosters Carried Maximum Payload

- We determined the booster that have carried the maximum payload using a subquery in the WHERE clause and the MAX() function.

```
%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

- We used a combinations of the WHERE clause, LIKE, AND conditions to filter for failed landing outcomes in drone ship, their booster versions, and launch site names for year 2015

```
%sql select MONTHNAME(DATE) as Month, landing__outcome, booster_version, launch_site from SPACEXTBL where DATE like '2015%' AND landing__outcome like 'Failure (drone ship)
```

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

- We selected Landing outcomes and the COUNT of landing outcomes from the data and used the WHERE clause to filter for landing outcomes BETWEEN 2010-06-04 to 2010-03-20.
- We applied the GROUP BY clause to group the landing outcomes and the ORDER BY clause to order the grouped landing outcome in descending order.

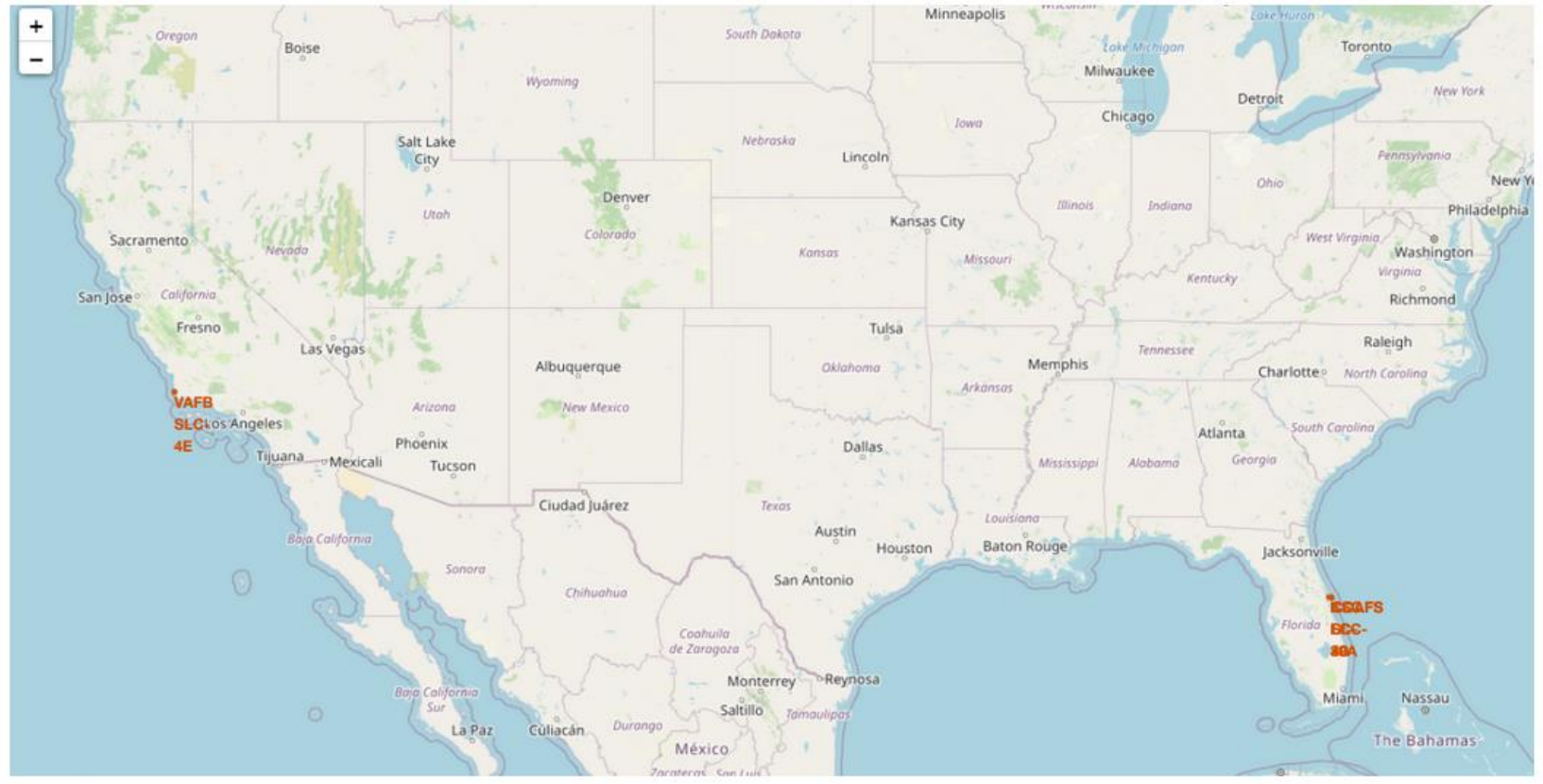
```
%sql select landing__outcome, count(*) as count from SPACEXTBL where Date >= '2010-06-04' AND Date <= '2017-03-20' GROUP by landing__outcome ORDER BY count Desc
```

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


Section 3
Launch Sites
Proximities Analysis

Location of all the Launch Sites

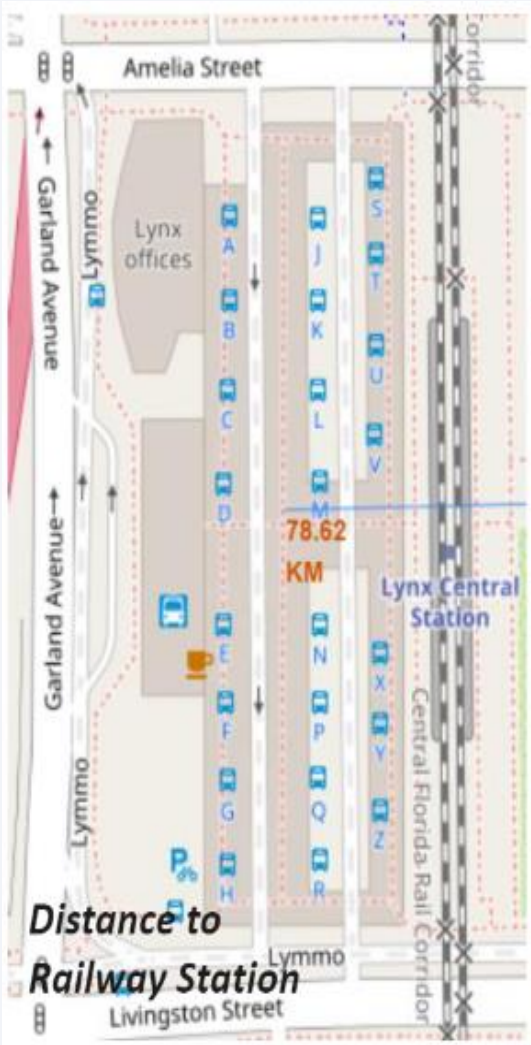
- We can see that all the SpaceX launch sites are located inside the United States



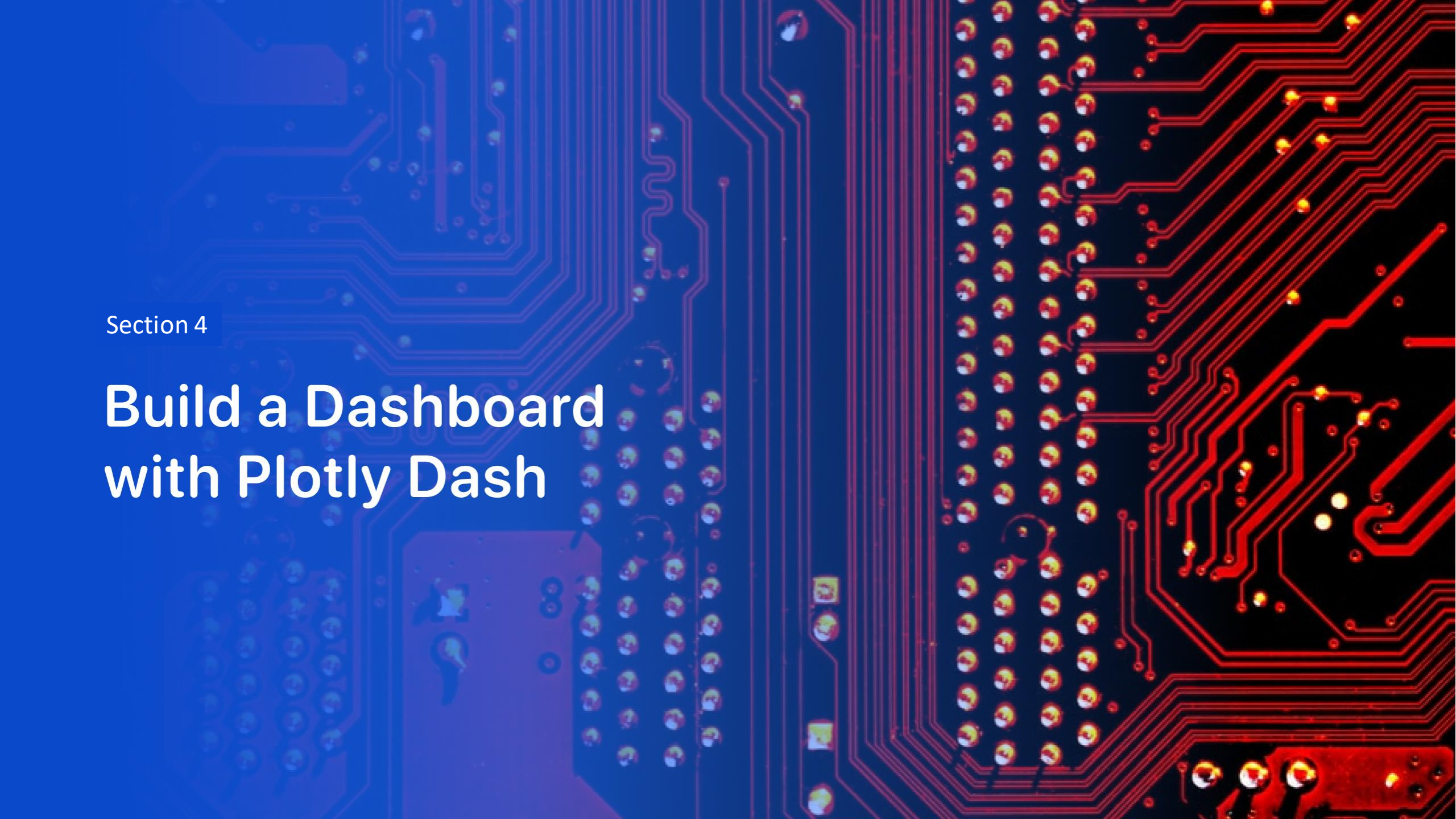
Markers showing launch sites with color labels

- Green Marker shows successful Launches and RED Marker shows failures.

Launch Sites Distance to Landmarks

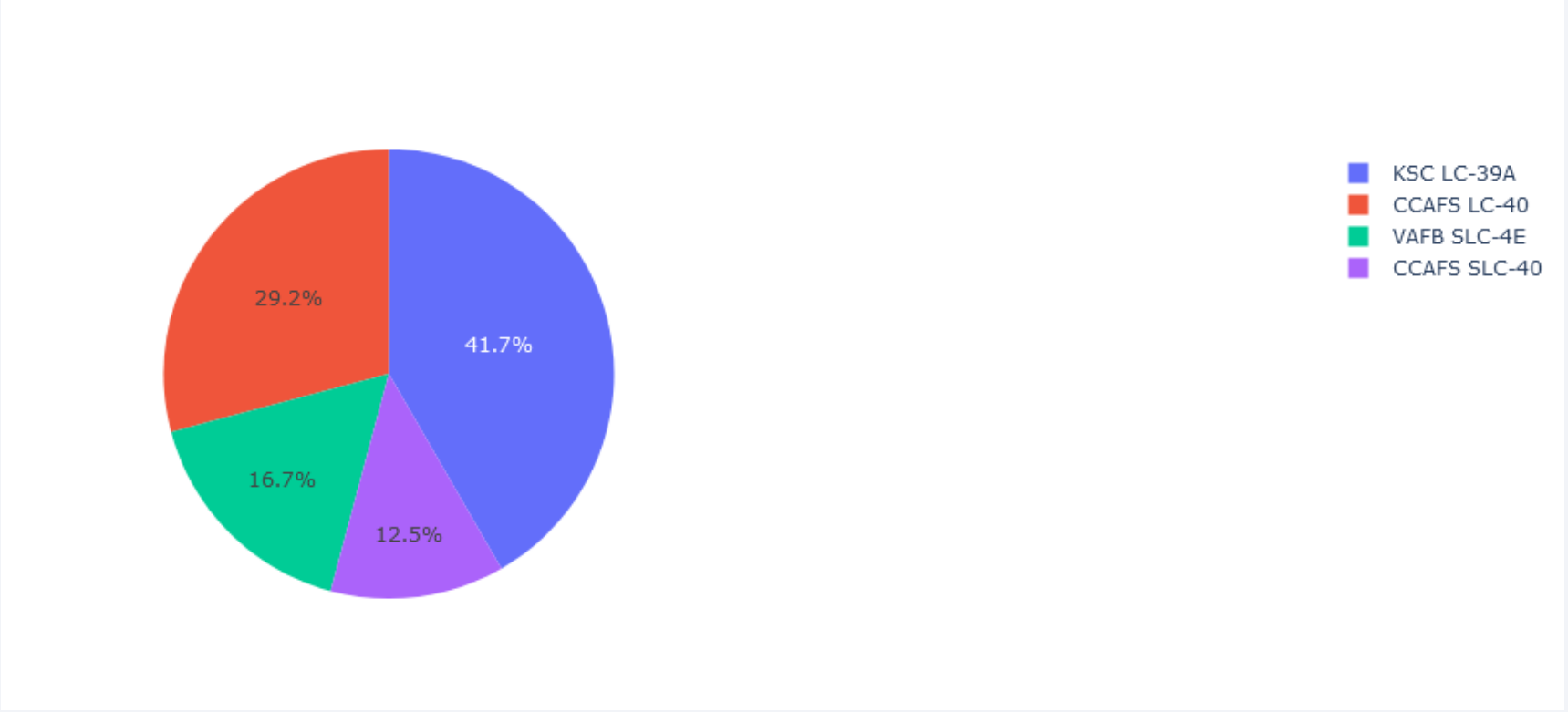


- Are launch sites in close proximity to railways? No
- Are launch sites in close proximity to highways? No
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes


Section 4
Build a Dashboard with Plotly Dash

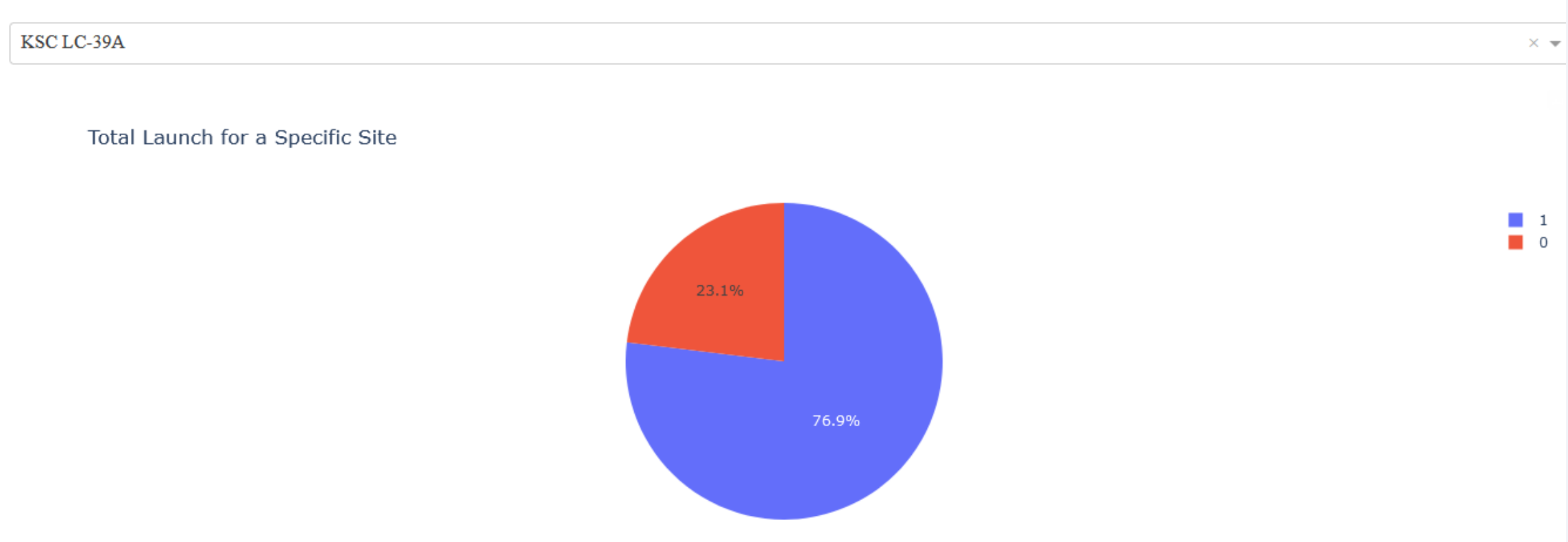
The success percentage by each sites.

- We can see that KSC LC-39A has the most successful launches from all the sites.



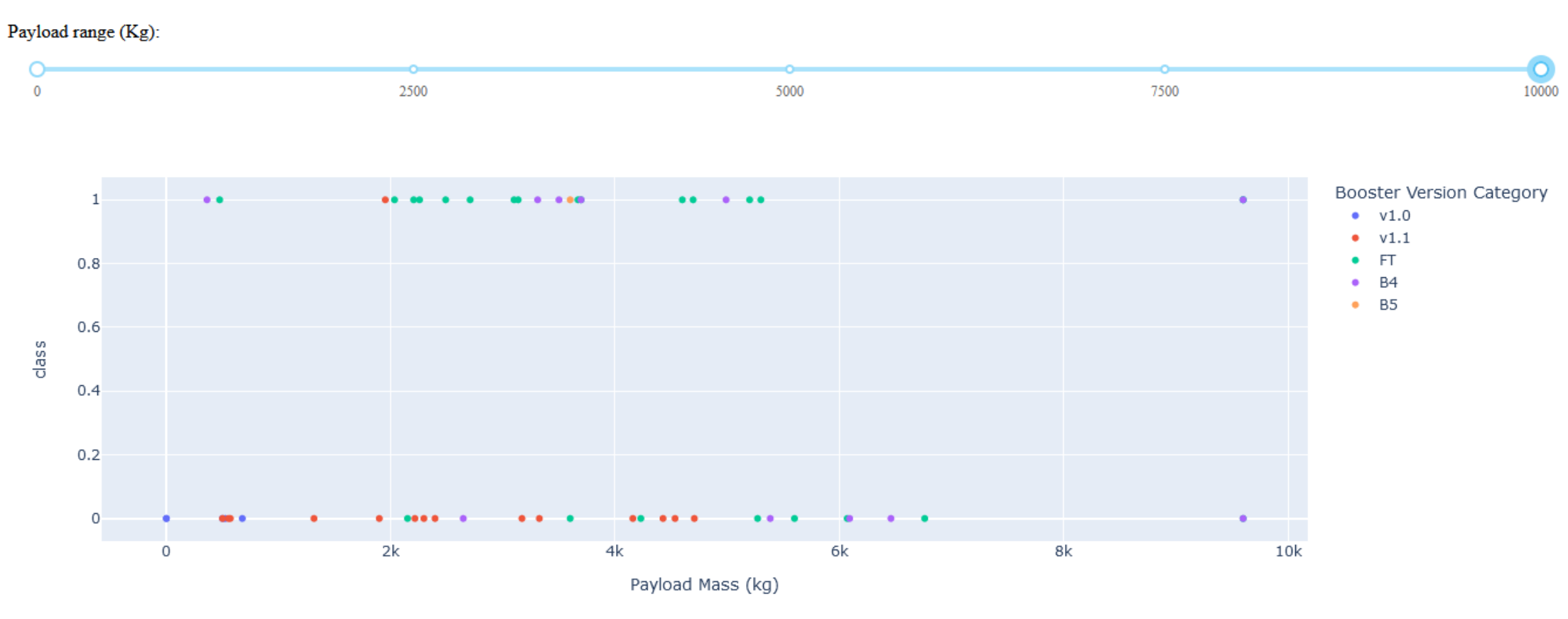
The highest launch-success ratio: KSC LC-39A

- KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate.



Payload Vs Launch Outcome Scatter Plot

- We can see that all the success rate for low weighted payload is higher than heavy weighted payload.

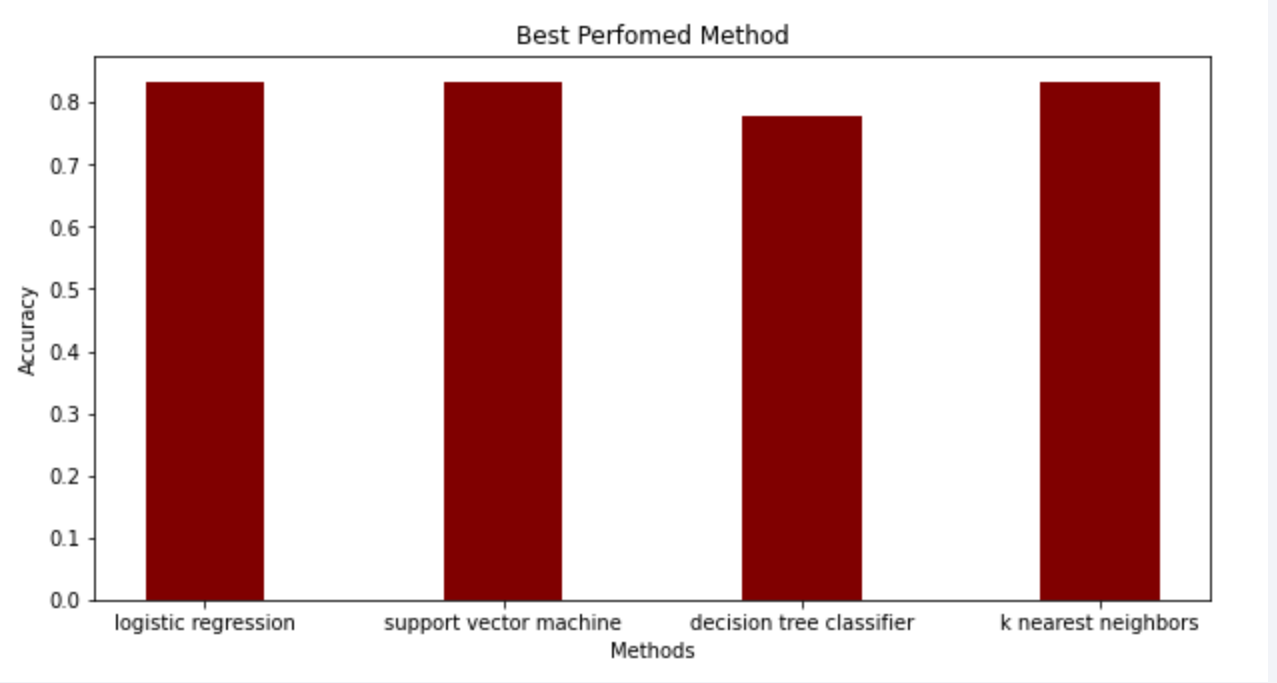


Section 5

Predictive Analysis (Classification)

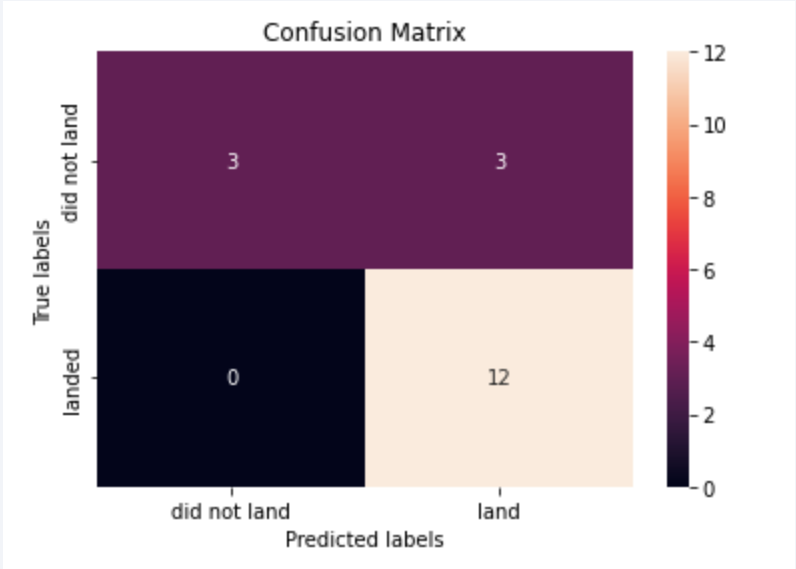
Classification Accuracy

- All models had virtually the same accuracy on the test set at 83.33% accuracy.
- This can cause large variance in accuracy results, such as those in Decision Tree Classifier model in repeated runs.
- We likely need more data to determine the best model.



Confusion Matrix

- Since all models performed the same for the test set, the confusion matrix is the same across all models. The models predicted 12 successful landings when the true label was successful landing.
- The models predicted 3 unsuccessful landings when the true label was unsuccessful landing.
- The models predicted 3 successful landings when the true label was unsuccessful landings (false positives). Our models over predict successful landings.



Conclusions

- We can conclude that:
  - The Tree Classifier Algorithm is the best Machine Learning approach for this dataset.
  - The low weighted payloads (which define as 4000kg and below) performed better than the heavy weighted payloads.
  - Starting from the year 2013, the success rate for SpaceX launches is increased, directly proportional time in years to 2020, which it will eventually perfect the launches in the future.
  - KSC LC-39A have the most successful launches of any sites; 76.9%
  - SSO orbit have the most success rate; 100% and more than 1 occurrence.

Appendix

- GitHub repository url: https://github.com/ShraddhaPatil15/Applied-Data-Science-Capstone-Project

Thank you!