Winning Space Race with Data Science

Sagar Varandekar
18-04-2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection
  - Data wrangling
  - Exploratory Data Analysis with data visualization
  - Exploratory Data Analysis with SQL
  - Building an interactive map with Folium
  - Building a Dashboard with Plotly Dash
  - Predictive analysis (Classification)

Executive Summary

- Summary of Results
  - Launch success rate since 2013 kept increasing till 2020 with successful landing rate about 50%.
  - Most part of recent launches were made from CCAFS LC-40 with a success rate of 73.1% and from KSC LC-39A with a success rate of 76.9%.
  - Launch sites are located as close as possible to the Equator line and in very close proximity to the coast, railway and highway. On the contrary, they maintain a certain distance to the cities.
  - Most part of the launches are carried out with a payload mass which varies from 2000 to 7000 kg with a high success rate. Heavy payload missions (payload mass > 8000 kg) have a good success rate.
  - Most recent missions use a VLEO orbit and present a high successful rate of 86%.
  - Boosters F9 FT have successfully landed on drone ship.
  - Models parameters are optimized.
  - Logistic regression, SVM, KNN and Decision Tree provide accurate predictions on the landing success. They should be improved to avoid the false positives which would impact directly on the final cost of the mission.

Introduction

- **Project background and context**

  We predicted if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

- **Common problems that needed solving**
  - How different conditions and locations influence landing success.
  - The influence of different rocket variables on the success rate of a rocket landing.
  - What conditions does SpaceX have to achieve to get the best results and ensure the best rocket success landing rate.

Section 1

Methodology

Methodology

Executive Summary

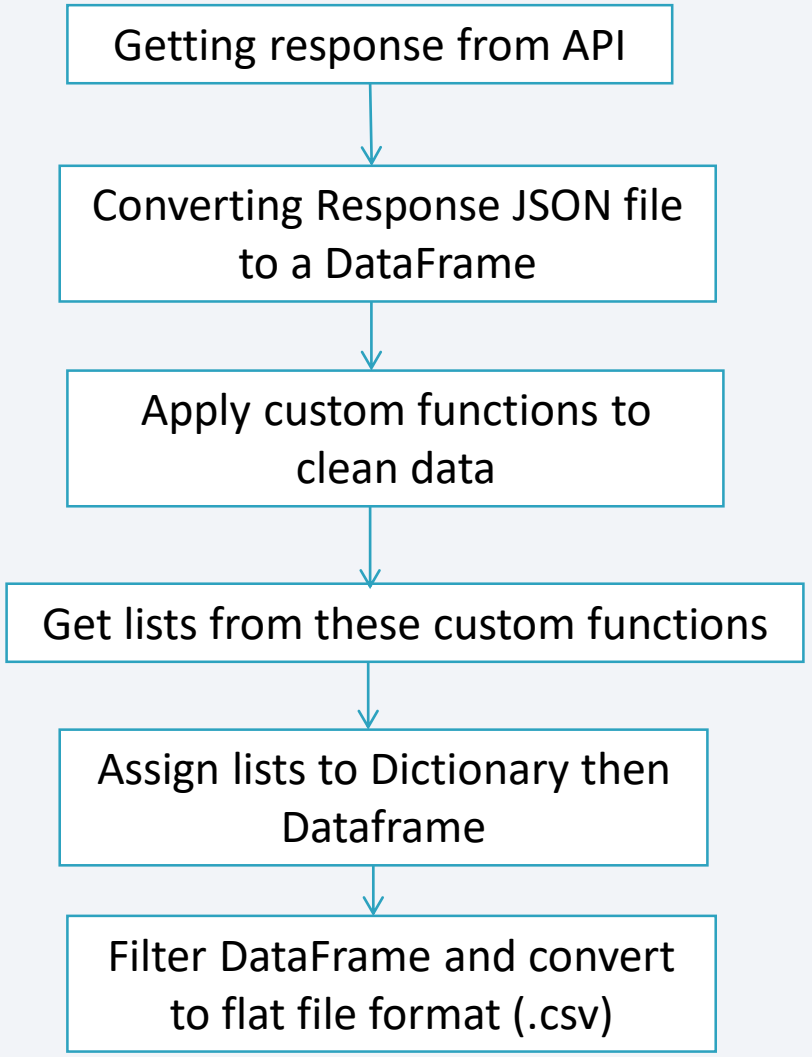
- Data collection methodology:
  - Required data was retrieved from SpaceX API and web scraped to collect Falcon 9 historical launch records from a Wikipedia page.
- Performed data wrangling
  - Imputed missing values with mean and other appropriate methods. One Hot Encoding data fields for Machine Learning and dropping irrelevant columns
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
  - Different models were trained and various hyper-parameter values tuned. Model with highest accuracy amongst these models with tuned hyper-parameters was selected.

Data Collection

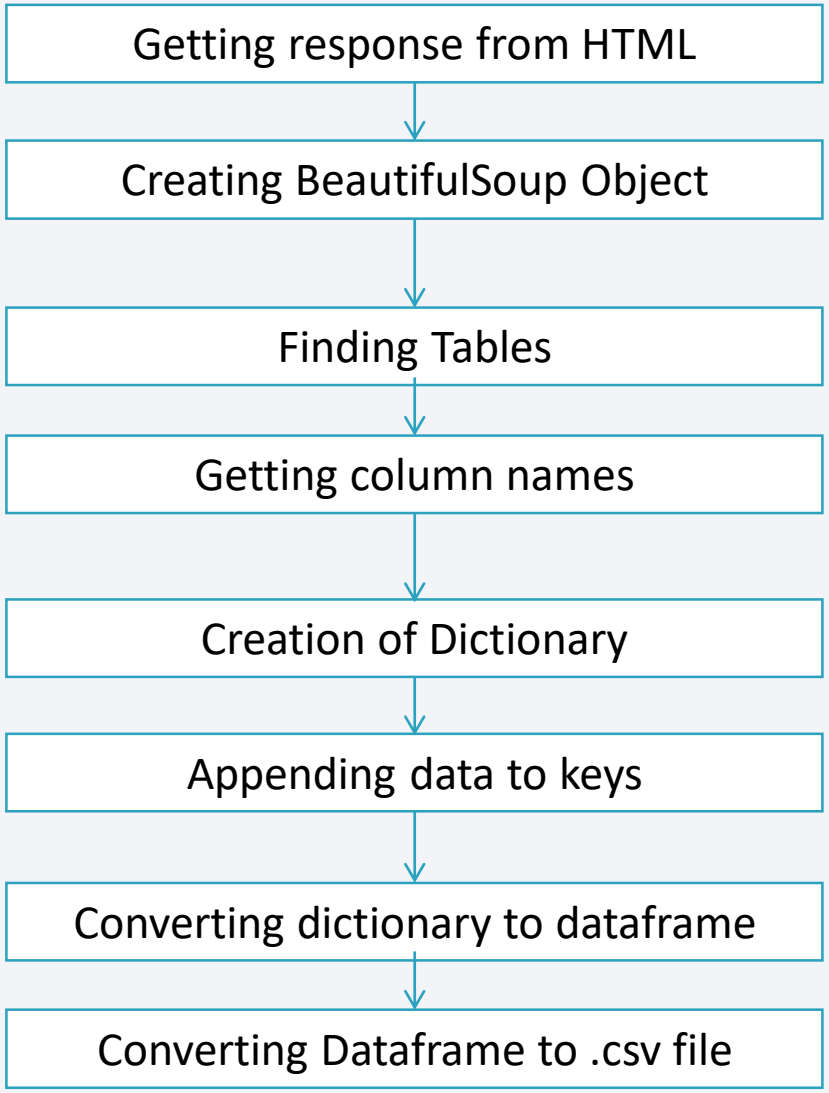
- Required data was retrieved from SpaceX API by making a get request to the SpaceX API.
- Data was also web scraped to collect Falcon 9 historical launch records from a Wikipedia page titled ‘list of Falcon 9 and Falcon heavy launches’ using python BeautifulSoup library.
- Our goal is to use this data to predict whether SpaceX will attempt to land rocket first stage or not.

Data Collection – SpaceX API

GitHubURL



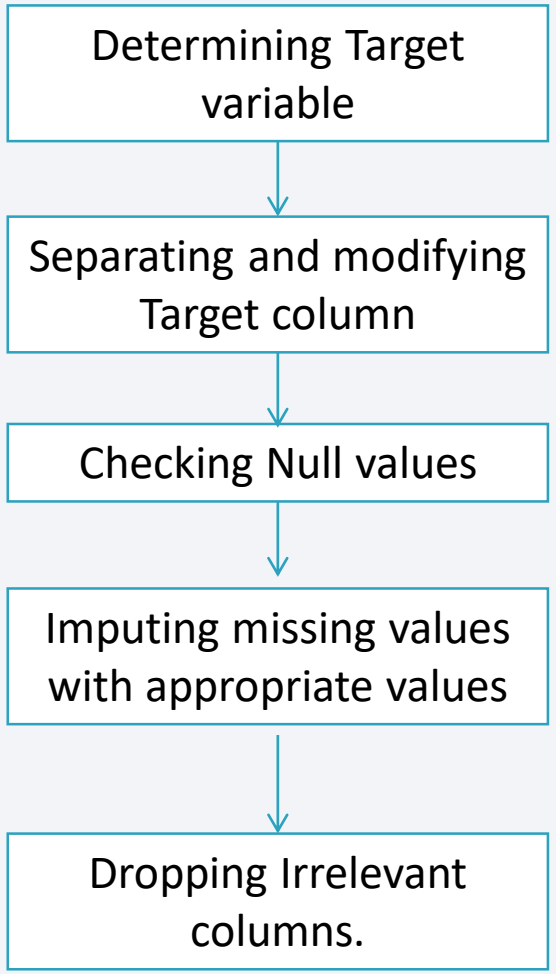
Data Collection - Scraping



GitHubURL

Data Wrangling

- Retrieved data was cleaned and transformed so that it can be used for modeling.
- Landing outcome column was modified to have only two values 1 for all kinds of success and 0 for all kinds of failure.
- Landing outcome set as the Target feature.
- Missing values in Payload column were replaced with the mean value.
- Irrelevant columns were dropped.
- Columns having categorical values were one hot encoded and column was created for each categorical value dropping the original column.



GitHubURL

EDA with Data Visualization

Scatter Graphs being drawn:

- Flight Number VS. Payload Mass
- Flight Number VS. Launch Site
- Payload VS. Launch Site
- Orbit VS. Flight Number
- Payload VS. Orbit Type
- Orbit VS. Payload Mass

The scatter graphs shows how successful and failed landing outcomes were distributed against different parameters.

GitHubURL

Bar Graph being drawn:

Mean VS. Orbit

The Bar graph shows the success rate for different orbits.

Line Graph being drawn:

Success Rate VS. Year

The line graph shows the trend of success rate with each passing year.

EDA with SQL

- Displayed the names of the **unique launch sites** in the space mission using **DISTINCT** command.
- Displayed 5 records where launch sites **begin with** the string **'CCA‘** using WHERE and **LIKE** command.
- Displayed the **total payload** mass carried by boosters launched by NASA (CRS) using **SUM** and WHERE command.
- Displayed **average payload** mass carried by booster version F9 v1.1 using **AVG** command.
- Listed the date when **first successful landing** outcome in ground pad was achieved using **MIN** command.
- Listed the names of the boosters which have success in drone ship and have payload mass **greater than 4000 but less than 6000** using WHERE and **BETWEEN** command.
- Listed the **total number** of successful and failure mission outcomes using **COUNT** command.
- Listed the names of the booster_versions which have carried the **maximum payload** mass using **MAX** command in a **subquery**.
- Listed the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 using WHERE and AND command.
- **Ranked** the **count of landing outcomes** between the date 2010-06-04 and 2017-03-20, in descending order using **GROUPBY** and **ORDERBY** command.

GitHubURL

Build an Interactive Map with Folium

- To visualize the Launch Data into an interactive map. We took the Latitude and Longitude Coordinates at each launch site and added a Circle Marker around each launch site with a label of the name of the launch site.
- We assigned the dataframe launch_outcomes(failures, successes) to classes 0 and 1 with Green and Red markers on the map in a MarkerCluster() to show successful/failure launches for each launch site.
- Using Haversine's formula we calculated the distance from the Launch Site to various landmarks to find various trends about what is around the Launch Site to measure patterns. Lines are drawn on the map to measure distance to landmarks.

GitHubURL

Build a Dashboard with Plotly Dash

- Pie chart of Launch Success counts for all sites and success rate of launches for each site was created in the dashboard.
- Dropdown was added to the dashboard to select launch site.
- Scatter Plot of Payload Mass vs Landing outcome was plotted with a slider for payload given to select payload range. Colors to data points with different booster versions.
- Pie Chart for all sites showed launch site with highest count of launch success.
- Pie chart for each launch site showed respective success rate with highest success rate of site KSC LC-39A.

GitHubURL

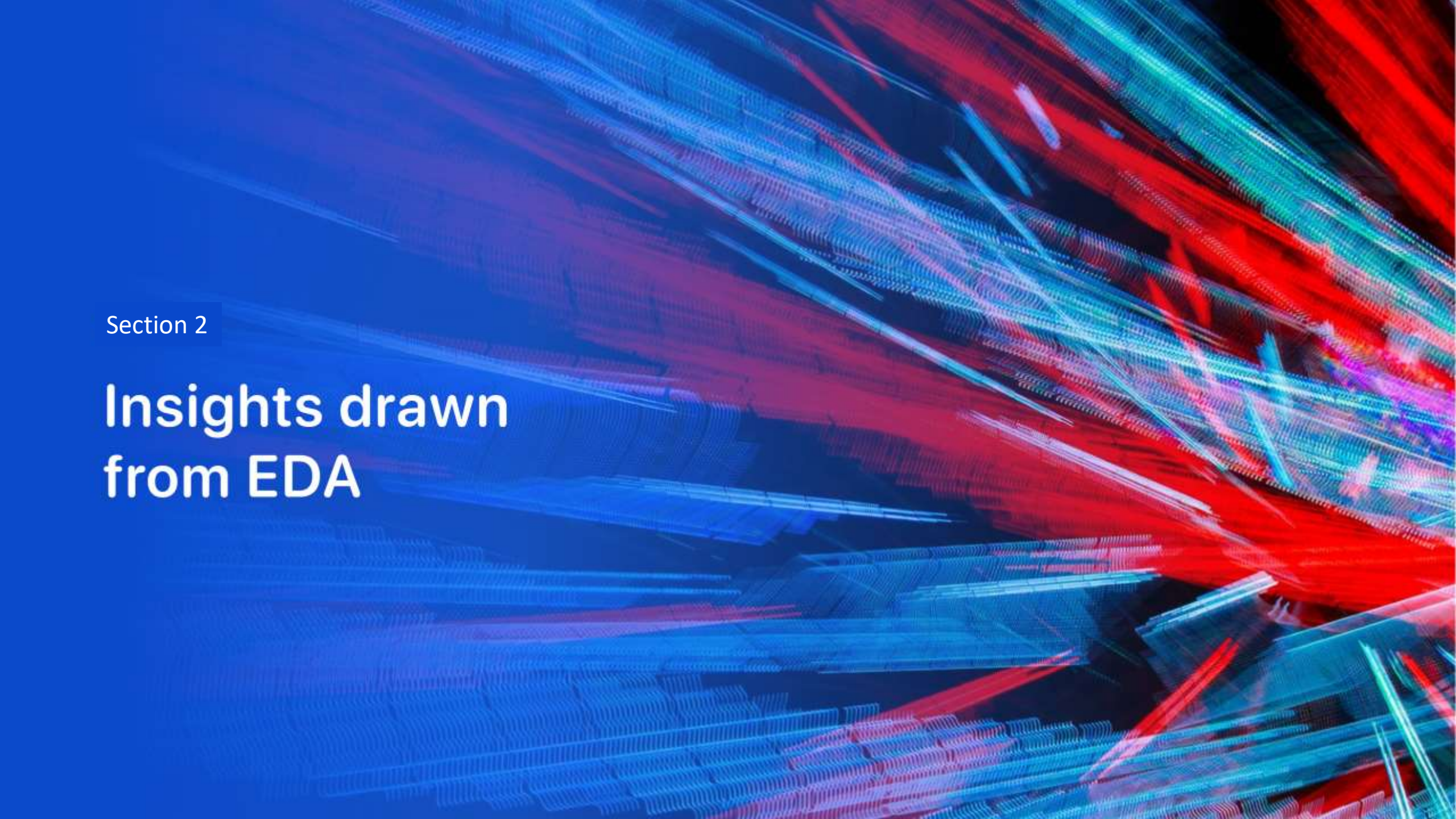
Predictive Analysis (Classification)

- **BUILDING MODEL**
- Load our dataset into NumPy and Pandas
- Transform Data
- Split our data into training and test data sets
- Decide which type of machine learning algorithms we want to use
- Set our parameters and algorithms to GridSearchCV
- Fit our datasets into the GridSearchCV objects and train our dataset.
- **EVALUATING MODEL**
- Check accuracy for each model
- Get tuned hyperparameters for each type of algorithms
- Plot Confusion Matrix
- **IMPROVING MODEL**
- Feature Engineering
- Algorithm Tuning
- **FINDING THE BEST PERFORMING CLASSIFICATION MODEL**
- The best model is the model with the best accuracy score on the test data.
- Accuracy score of all models was same on test data hence model with highest accuracy score on training data was selected.

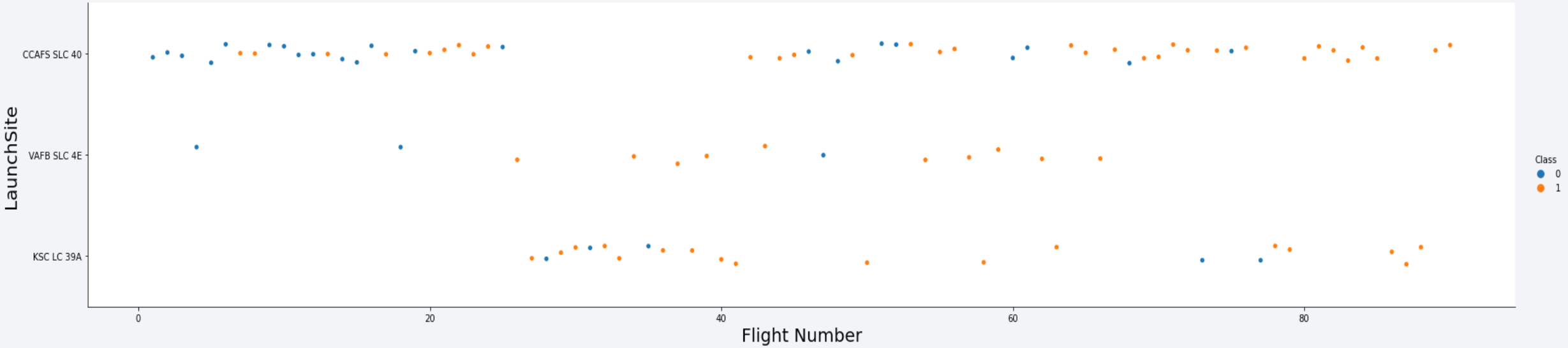
GitHubURL

Results

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


Section 2
Insights drawn
from EDA

Flight Number vs. Launch Site



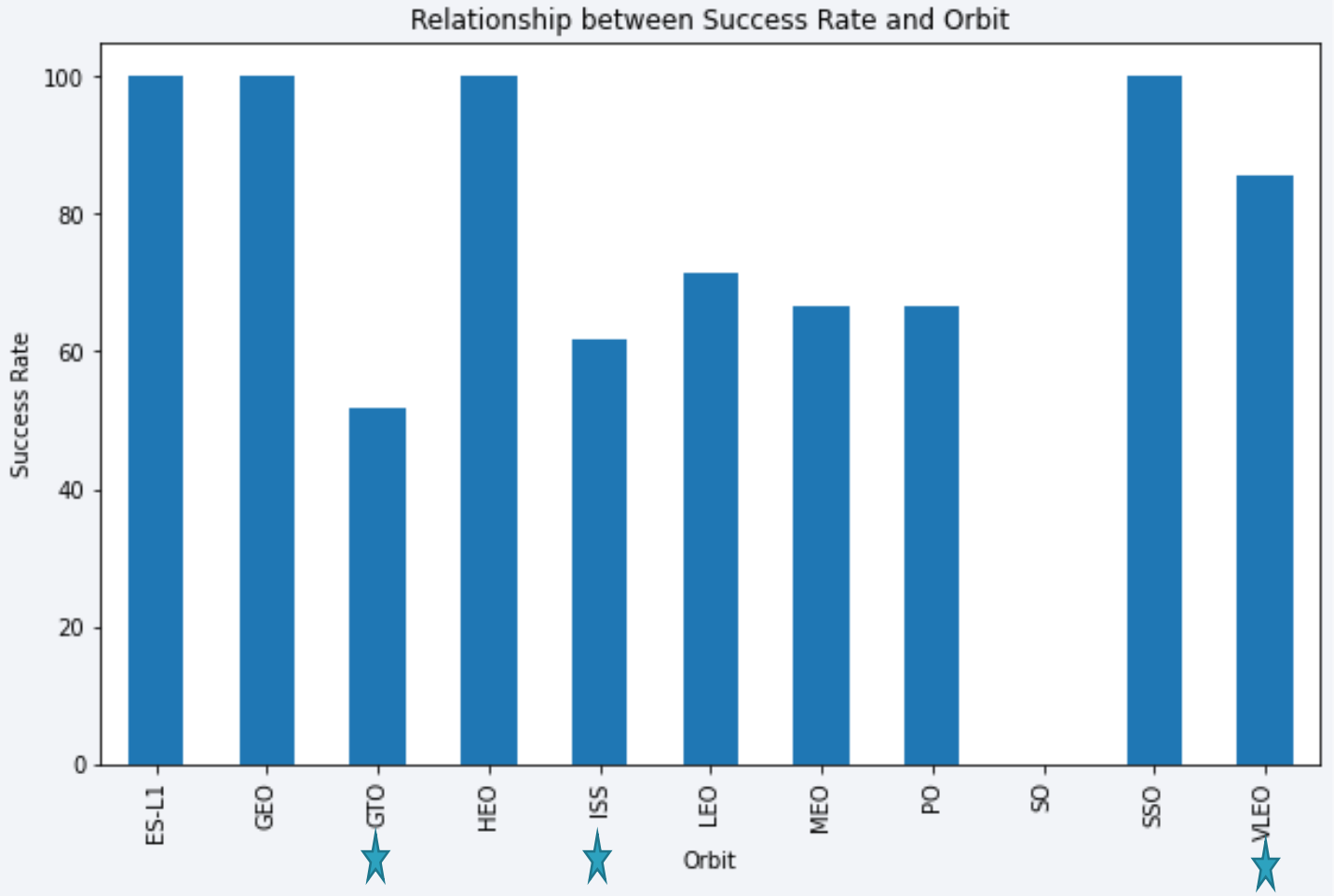
- Launch success rate has increased for all the launch sites from the first to the last launch.
- Most part of the launches were performed from CCAFS and KSC launch sites.

Payload vs. Launch Site

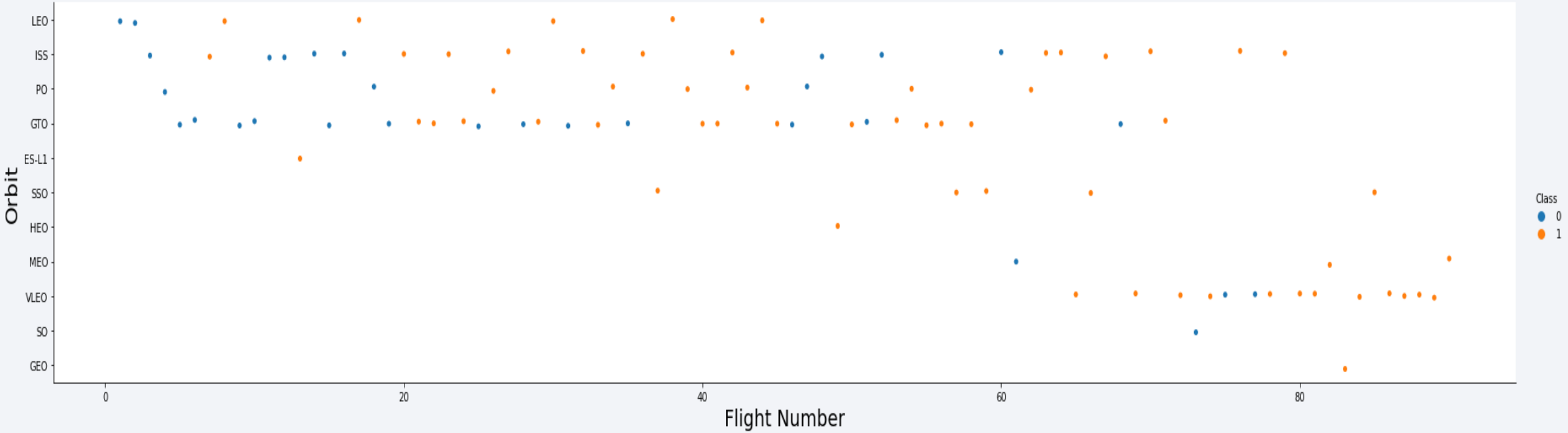
Launch Success Rate is higher for payload range greater than 8000 kg.

Success Rate vs. Orbit Type

- The orbits ES-L1, GEO, HEO and SSO have highest 100% sucess rate.
- No successful launch found for orbit type SO.
- The rest of orbit types present a success rate of around 60%.
- Number of launches per orbit type would complete the information provided by this bar chart since the higher the number of flights is the more representative (★) is the calculated success rate (see next slide).

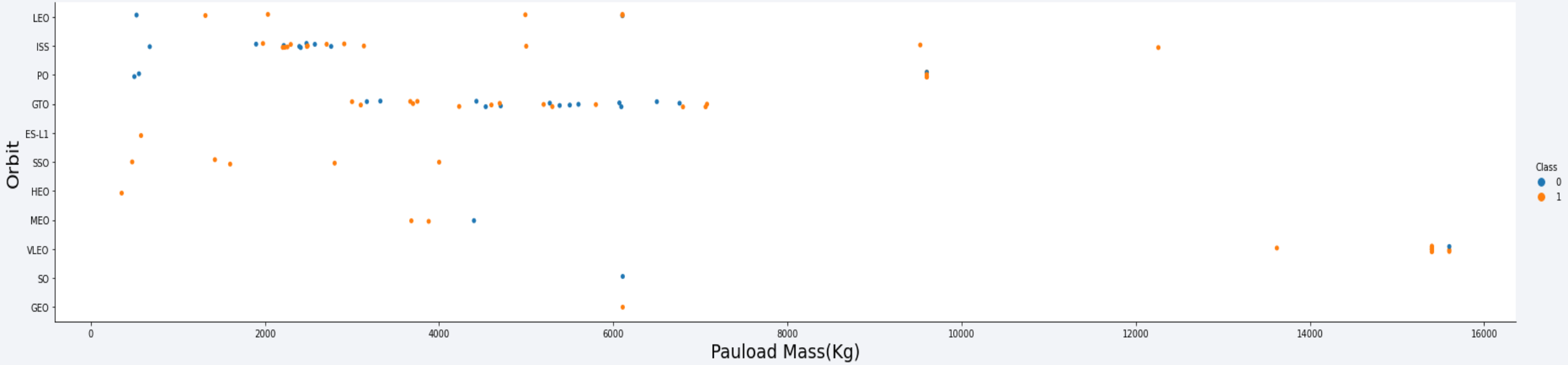


Flight Number vs. Orbit Type



- Higher number of launches correspond mainly to the orbit types ISS (21), GTO (27) and VLEO (14).
- In LEO orbit the success appears related to the number of flights.
- There seems to be no relationship between orbit type and flight number for GTO and ISS orbits.
- VLEO orbit type has been choosen for the most recent launches (from flight 65 on) and present a high successful rate of 86%.

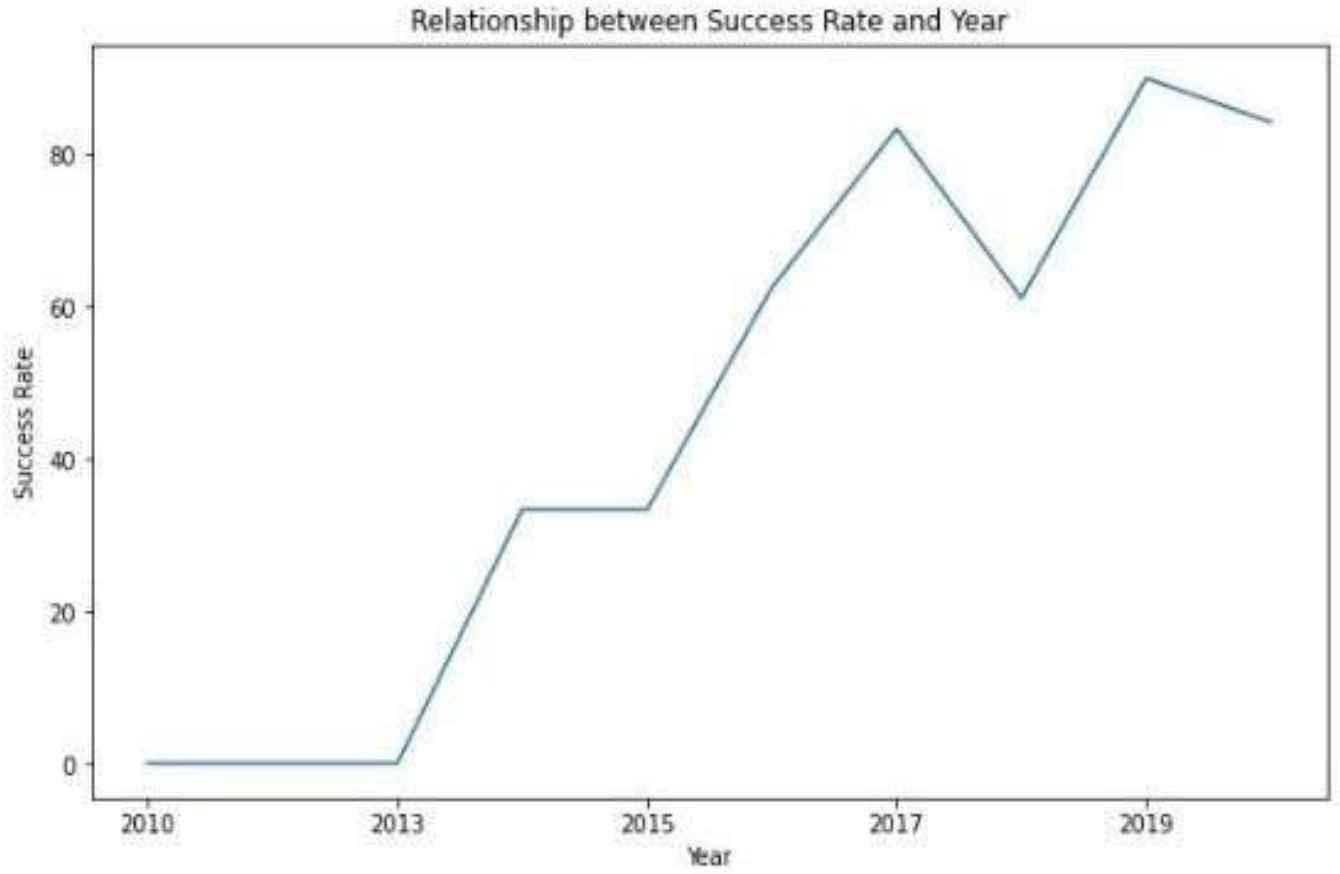
Payload vs. Orbit Type



Heavier payloads have a negative influence on GTO orbit and positive influence on LEO and ISS orbits.

Launch Success Yearly Trend

- The success rate since 2013 kept increasing till 2017.
- In 2018 the rate falls to around 60%, but it raised again to its maximum in 2019 and falls a bit agin in 2020.



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

All Launch Site Names

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

```
In [4]: %sql select unique(launch_site) from spacex
```

```
 * ibm_db_sa://zbx64872:***@824dfd4d-99de-440d-9991-629c01b38
Done.
```

Out[4]:

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

Launch Site Names Begin with 'CCA'

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

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

```
 * ibm_db_sa://zbx64872:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

Out[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 for NASA

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

In [6]:
```
%sql select sum(payload_mass__kg_) from spacex where customer='NASA (CRS)'
```

```
 * ibm_db_sa://zbx64872:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1
Done.
```

Out[6]:

1
45596

Average Payload Mass by Booster Version F9 v1.1

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

```
In [7]: %sql select avg(payload_mass__kg_) from spacex where booster_version like 'F9 v1.1%'

 * ibm_db_sa://zbx64872:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg
Done.

Out[7]:    1
           2534
```

First Successful Ground Landing Date

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

Hint:Use min function

In [8]:
```
%sql select min(DATE) as First_success_groundpad from spacex where landing__outcome='Success (ground pad)'
```

```
 * ibm_db_sa://zbx64872:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.c
Done.
```

Out[8]:

first_success_groundpad
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 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 [9]:
```
%sql select booster_version from spacex where (landing__outcome='Success (drone ship)' and payload_mass__kg_ between 4000 and 6000
```

```
 * ibm_db_sa://zbx64872:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

Out[9]:

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

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
In [10]: %sql select count(*) as mission_outcome_success from spacex where mission_outcome like 'Success%'
```

```
 * ibm_db_sa://zbx64872:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdor
Done.
```

Out[10]:

mission_outcome_success
100

```
In [11]: %sql select count(*) as mission_outcome_failure from spacex where mission_outcome like 'Failure%'
```

```
 * ibm_db_sa://zbx64872:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdor
Done.
```

Out[11]:

mission_outcome_failure
1

Boosters Carried Maximum Payload

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

In [12]:
```
%sql select booster_version from spacex where payload_mass__kg_=(select max(payload_mass__kg_) from spacex)
```

```
 * ibm_db_sa://zbx64872:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:
Done.
```

Out[12]:

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

Booster Version F9 B5 carried maximum payload.

2015 Launch Records of Failed landing outcome-Drone Ship

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

In [13]:
```
%sql select date, booster_version, launch_site, landing__outcome from spacex where (landing__outcome='Failure (drone ship)' and c
```

```
 * ibm_db_sa://zbx64872:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

Out[13]:

DATE	booster_version	launch_site	landing__outcome
2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

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

In [14]:
```
%sql  SELECT landing__outcome, COUNT(DATE) FROM spacex where DATE between '2010-06-04' and '2017-03-20' GROUP BY  landing__outcor
```

```
 * ibm_db_sa://zbx64872:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

Out[14]:

landing__outcome	2
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

Launch Sites marked on Map

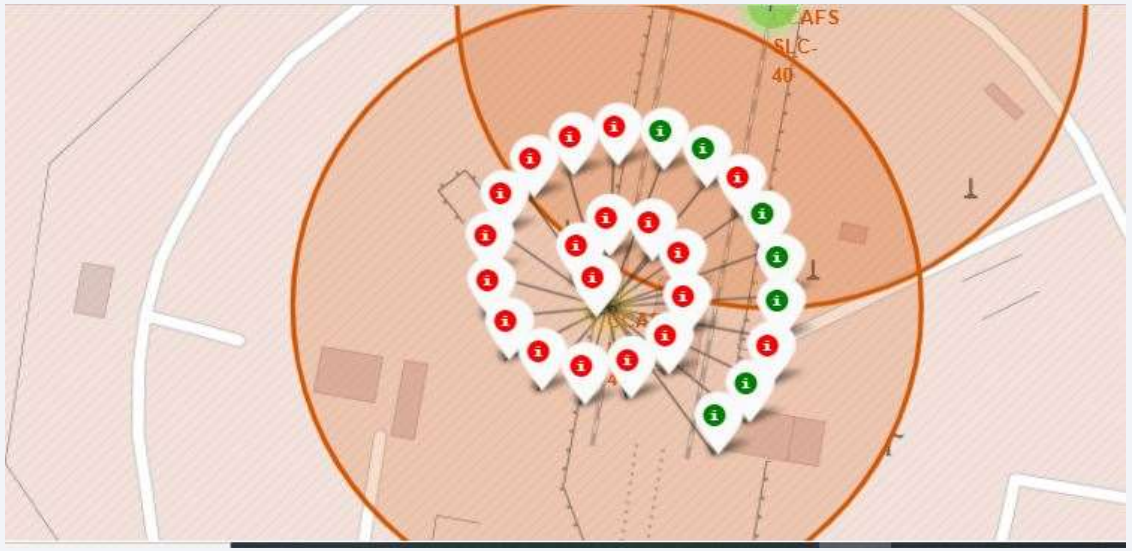
- Three launch sites are close to each other in Orlando, Florida and only one launch site is in California.
- All launch sites are in very close proximity to the coast so that first stage can be thrown to the sea after the take-off.

Success/failed launches clustered for each site marked on the map

MarkerCluster Object was used to mark success/failure of launch sites.

Success/failed launches for each site marked on the map

KSC LC-39A

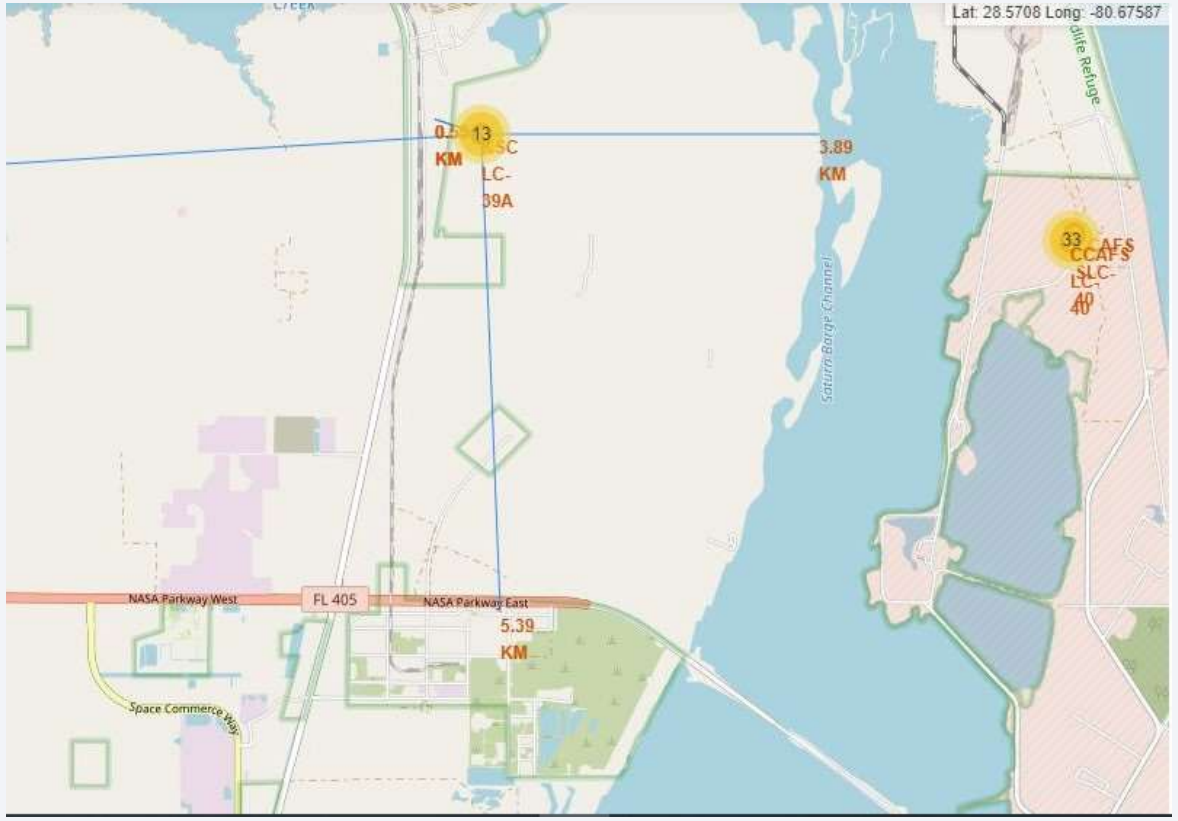


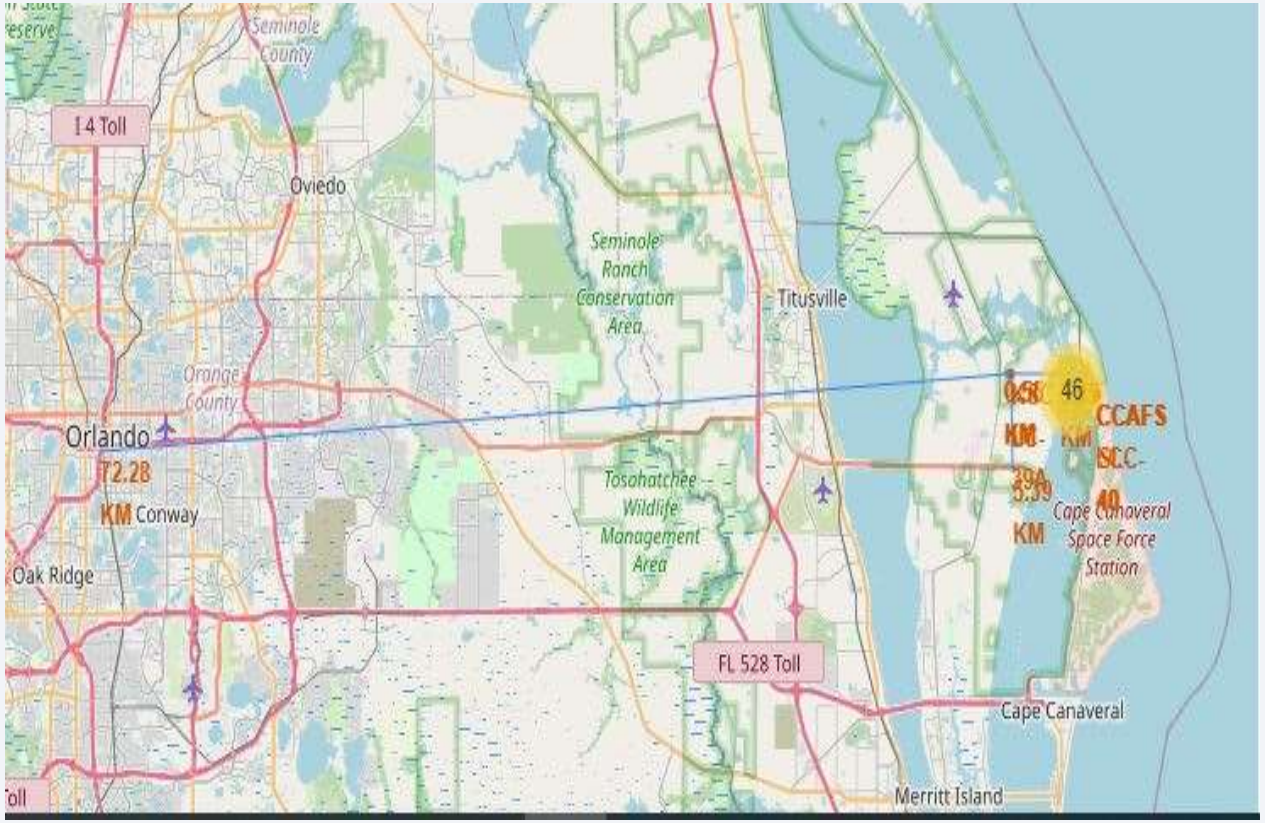
CCAFS LC-40

VAFB SLC-4E

Launch Site KSC LC-39A has highest success ratio.

Distances between a launch site to its proximities

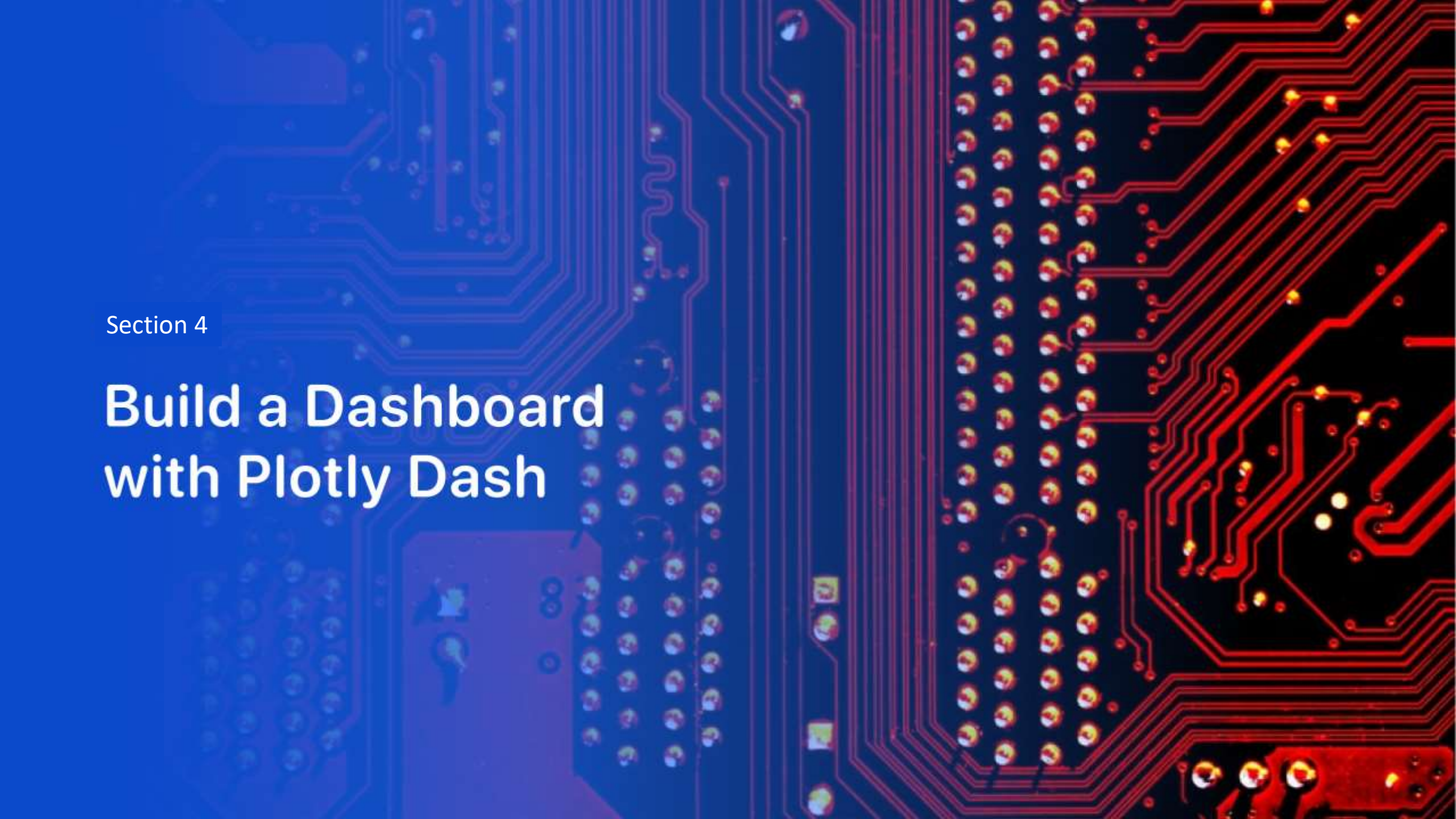




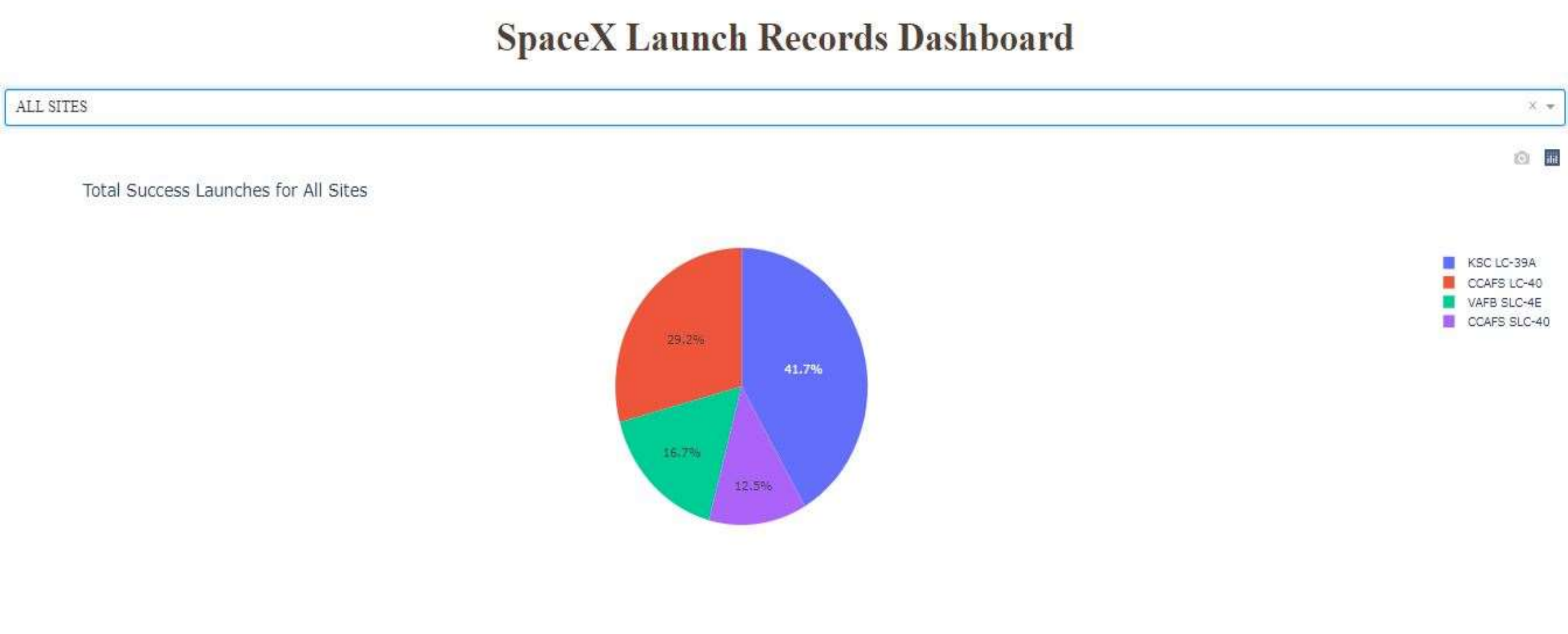
- Launch sites are close to railways and highways for logistic reasons.
- They are also close to the coastline so that first stage can be thrown to the sea after the take-off.
- On the contrary, they keep certain distance away from cities.

Section 4

Build a Dashboard with Plotly Dash

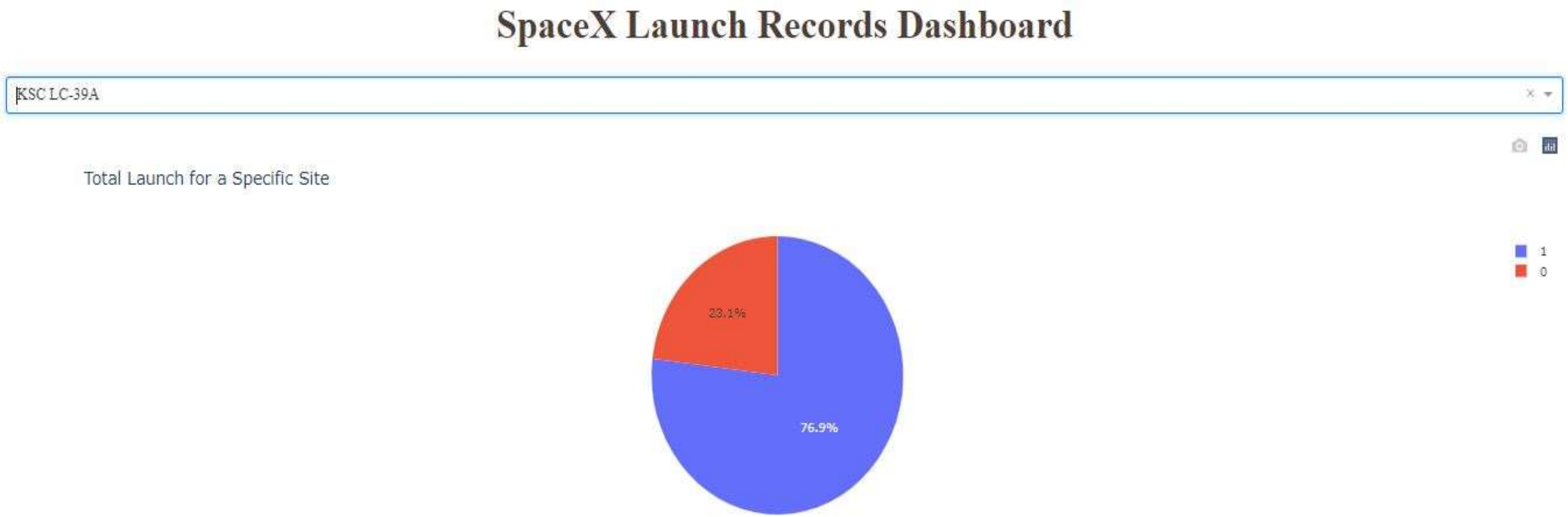


Launch Success Count for all sites Pie Chart



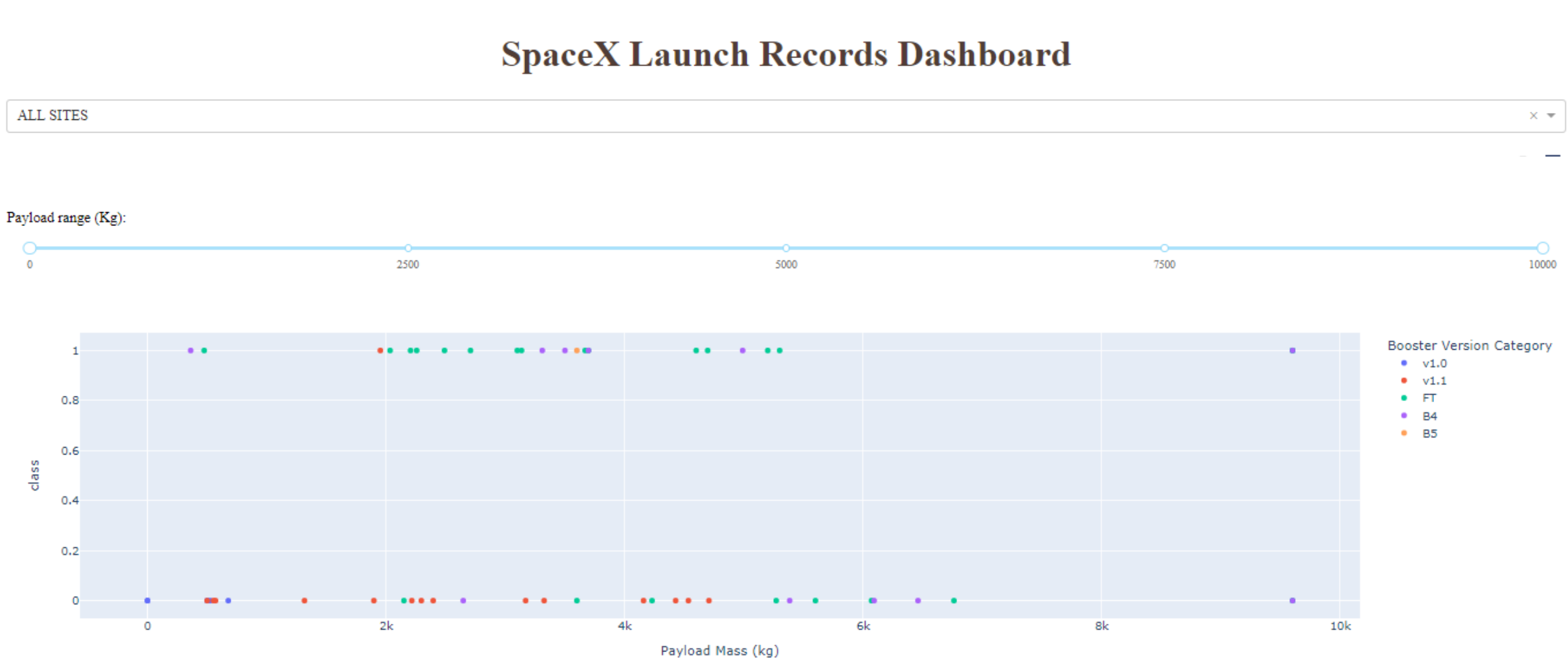
Launch Site KSC-LC-39A has highest count of successful launches.

The piechart for the launch site with highest launch success ratio



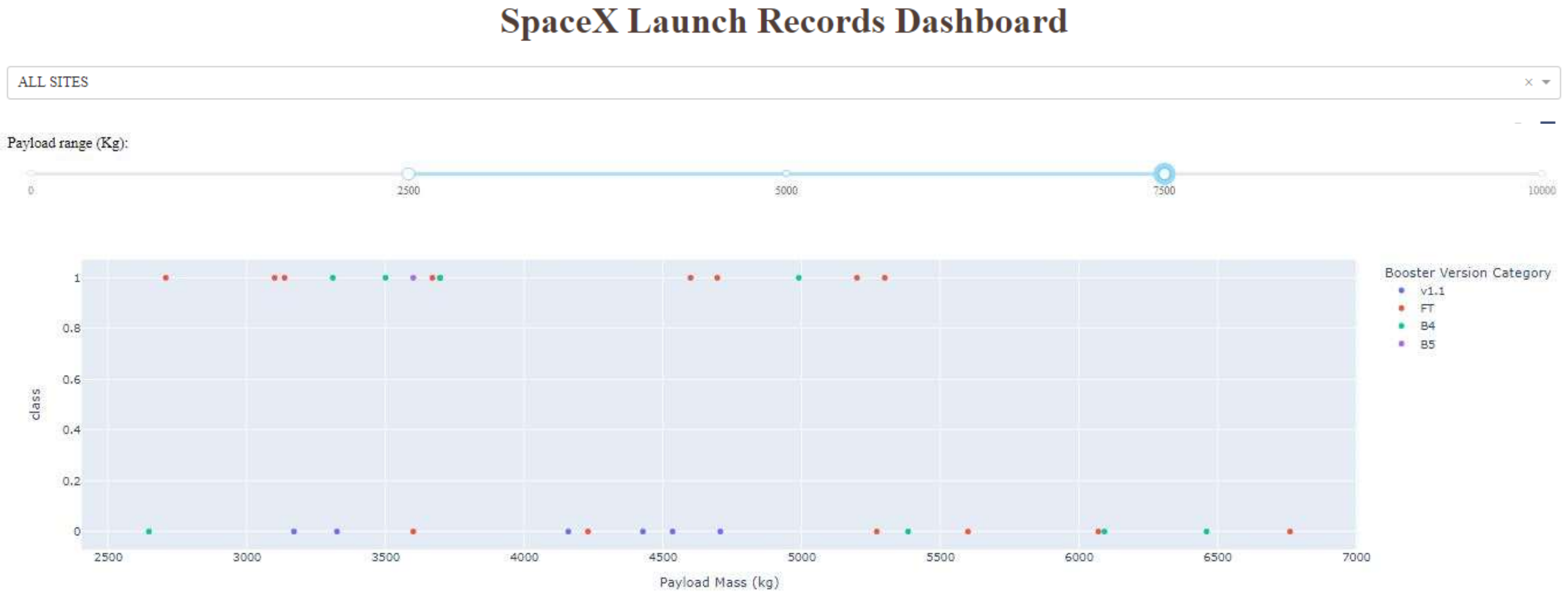
Launch Site KSC-LC-39A has highest success rate of 76.9%

Payload vs. Launch Outcome scatter plot for all sites



- A success rate of 0% is obtained for payload masses in the range 6000-7000 kg.
- Most part of the launches are carried out with a payload mass which varies from 2000 to 7000 kg

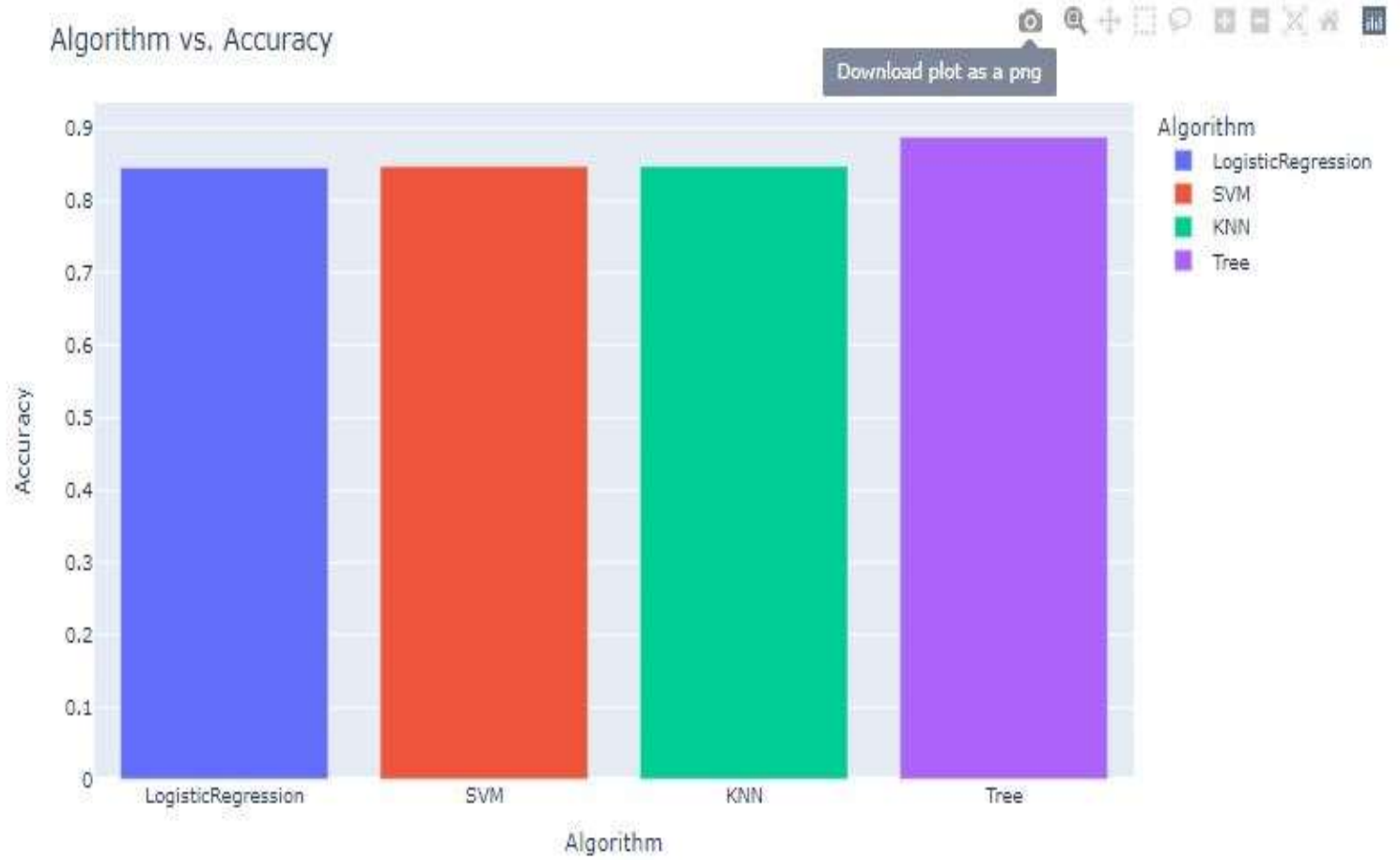
Payload vs. Launch Outcome scatter plot for all sites 2



Section 5

Predictive Analysis (Classification)

Classification Accuracy

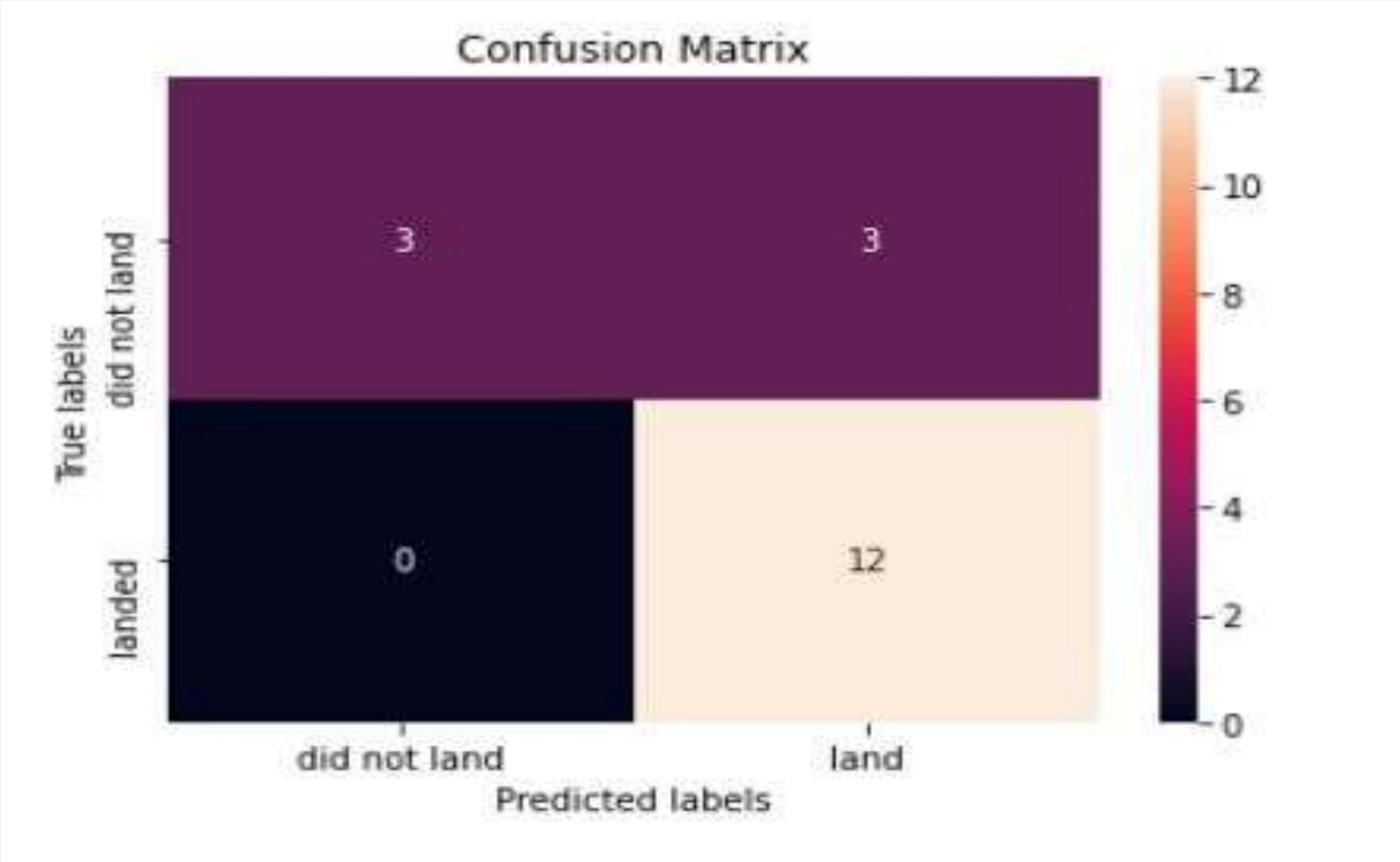


	Algorithm	Accuracy
0	LogisticRegression	0.846429
1	SVM	0.848214
2	KNN	0.848214
3	Tree	0.889286

Decision Tree Classification model has the highest accuracy on training data.

After selecting the best hyperparameters for the decision tree classifier using the validation data, we achieved 83.33% accuracy on the test data.

Confusion Matrix for Decision Tree Classifier



Examining the confusion matrix, we see that Tree can distinguish between the different classes. We see that the major problem is false positives.

Conclusions

- The Tree Classifier Algorithm is the best for classification for this dataset.
- Launch success rate since 2013 kept increasing till 2020.
- Success rate improves as number of flights increases.
- Most part of recent launches were made from CCAFS LC-40 with a success rate of 73.1% and from KSC LC-39A with a success rate of 76.9%.
- We can see that KSC LC-39A had the most successful launches from all the sites and also the highest success rate.
- Most part of the launches are carried out with a payload mass which varies from 2000 to 7000 kg with a good success rate. Heavy payload missions (payload mass > 8000 kg) have a higher success rate.
- Launch sites are located in very close proximity to the coast, railway and highway. On the contrary, they maintain a certain distance to the cities.
- Orbit GEO,HEO,SSO,ES-L1 has the best Success Rate.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!