IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Gulixian Shawuti
04/04/2024

Outline

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

Executive Summary

This project involved a multifaceted approach to analyze SpaceX launch data and predict the successful landing of first stages.

Methodology used in this project:

— data collection through an open-source SpaceX REST API and supplemented it with historical data obtained via web scraping.

—- EDA conducted using SQL, Matplotlib, and Pandas, we gained insights into the dataset. Visualizations were created using Folium and an interactive analytical dashboard was generated using Python's Plotly and Dash frameworks.

—-Trained three distinct machine learning models - Decision Tree Classifier, K-Nearest Neighbors, and Support Vector Machine (SVM)

Final result revealed that approximately 65-70% of all SpaceX launches typically achieve successful first stage landings. This comprehensive approach provides valuable insights into SpaceX's operational performance and the predictive capabilities of machine learning models in this domain.

Introduction

In recent decades, commercial spaceflight has materialized as a viable industry. Among the leading entities in this domain is SpaceX, widely regarded as exceptionally successful.

The objective is to evaluate Space Y to compete with Space X first stage rocket landings as a means to evaluate the relative expenses associated with each launch.

Desirable outcome:

—- The best way to estimate the total cost for launches, by predicting successful landings of the first stage of rockets;

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Used SpaceXREST API
  - Web scraping through Python
- Perform data wrangling

  Examined success rate by orbit type, launch site location and payload mass

Methodology

Executive Summary

- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Three ML models to do prediction and compare accuracy of each model evaluated using different combinations of parameters

Data Collection

- Describe how data sets were collected two main sources:

  1.From SpaceX REST API to get our data to analyse

  Utilized distinct endpoints to access data pertaining to variables such as landing status, payload mass, and the nature of the landing (e.g., drone ship or ocean).

  2. Web scraping on Wikipedia

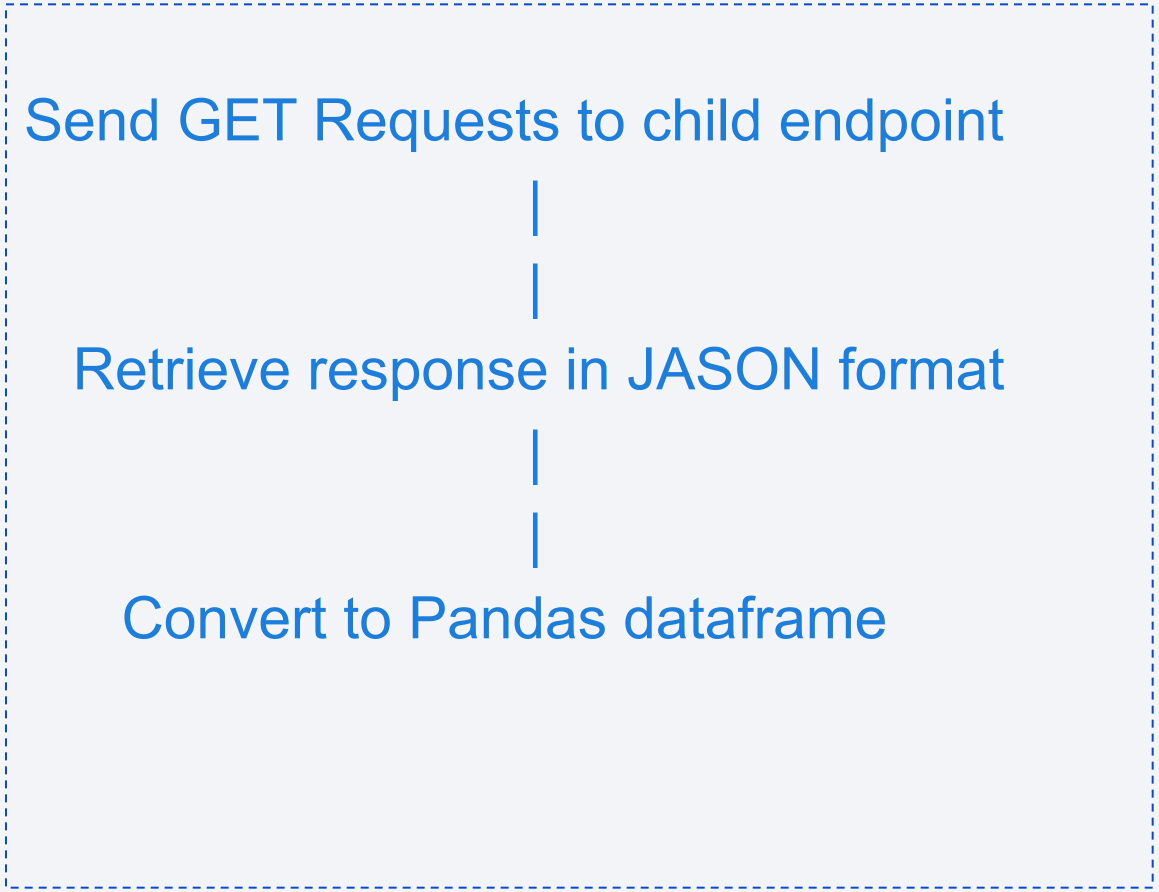
  The Wikipedia page dedicated to SpaceX Falcon 9 launches contains a wealth of pertinent information regarding launches completed up to the present time.

- You need to present your data collection process use key phrases and flowcharts

Data Collection – SpaceX API

- SpaceX offers a public API from where data can be obtained and then used
- This API was used according to the flowchart beside and then data is persisted.

- GitHub URL of the completed SpaceX API calls notebook https://github.com/Gulxan/testreposit/blob/main/jupyter-labs-webscraping.ipynb



Send GET Requests to child endpoint

Retrieve response in JASON format

Convert to Pandas dataframe

Data Wrangling

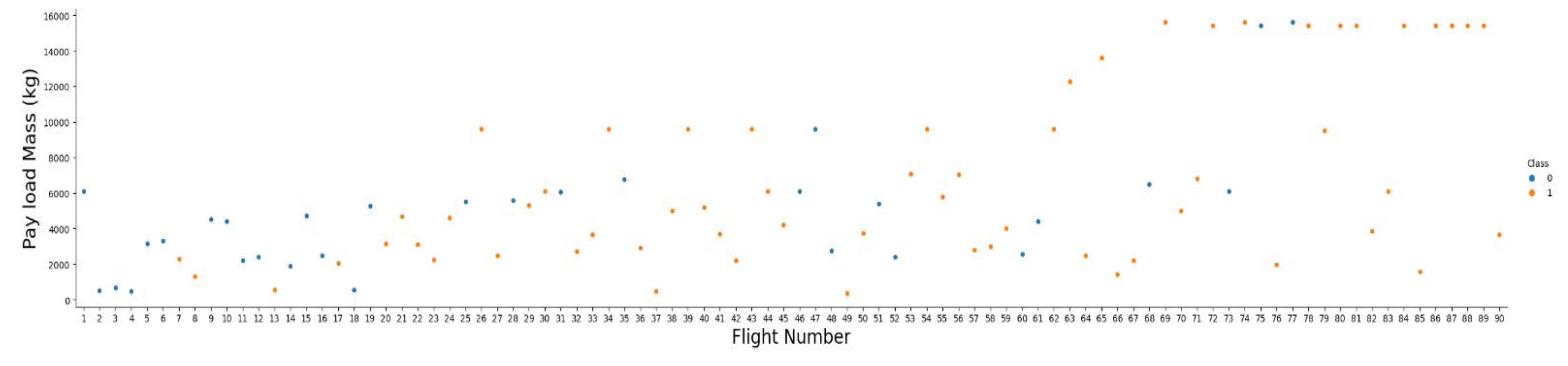
- EDA was performed on the dataset.
- Summaries launches per site,orbit and mission outcome per orbit type were calculated
- Landing_outcome label created

GitHub URL of the completed data wrangling related notebooks,

https://github.com/Gulxan/testreposit/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

EDA in particularly scatter plots and barplots was used to visualize what factors were most important in determining the success of a first-stage landing.



GitHub URL of the EDA with data visualization notebook,

https://github.com/Gulxan/testreposit/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- Unique launch site names in space mission and top 5 launch site names starts with ‘CCA’
- Total payload mass carried by boosters launched by NASA
- Average payload mass carried y booster version F9V1.1
- First successful landing outcome (Groundpad) achieved date
- Total number of successful and failure mission outcomes
- Booster version names which carried the max payload mass
- Failed landing outcomes in drone ship, their booster versions
- Rank of the count of landing outcomes

GitHub URL of the completed EDA with SQL notebook

https://github.com/Gulxan/testreposit/blob/main/jupyter-labs-eda-sql-edx_sqllite.ipynb

Build an Interactive Map with Folium

- This Folium was employed to create geographical visualizations of launch site distributions within the dataset.
- Each launch site was marked with an orange dot, with clustering predominantly observed in three main locations: Florida, Texas, and California.
- Furthermore, lines are used to indicate distances between two coordinates.

GitHub URL of the completed interactive map with Folium map

https://github.com/Gulxan/testreposit/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- The primary factors influencing landing statuses, particularly focusing on the launch site and payload mass of the respective rockets.
- Leveraging an interactive Plotly Dash dashboard, can effectively filter data based on these parameters, facilitating clearer visualization of success rates across various launch sites and payload mass ranges.

- GitHub URL of the completed Plotly Dash lab,

Predictive Analysis (Classification)

- For predictive Analysis four machine learning models typically applied in classification tasks, including K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression, and Decision Tree Classification.
- Following model training:

1. evaluated their performance by utilizing a distinct validation/test set with corresponding labels.
2. Subsequently, we assessed the accuracy of each model across epochs to ascertain the optimal performance.

Data preparation→Test of each model with combination of parameters—--> Final results

Github

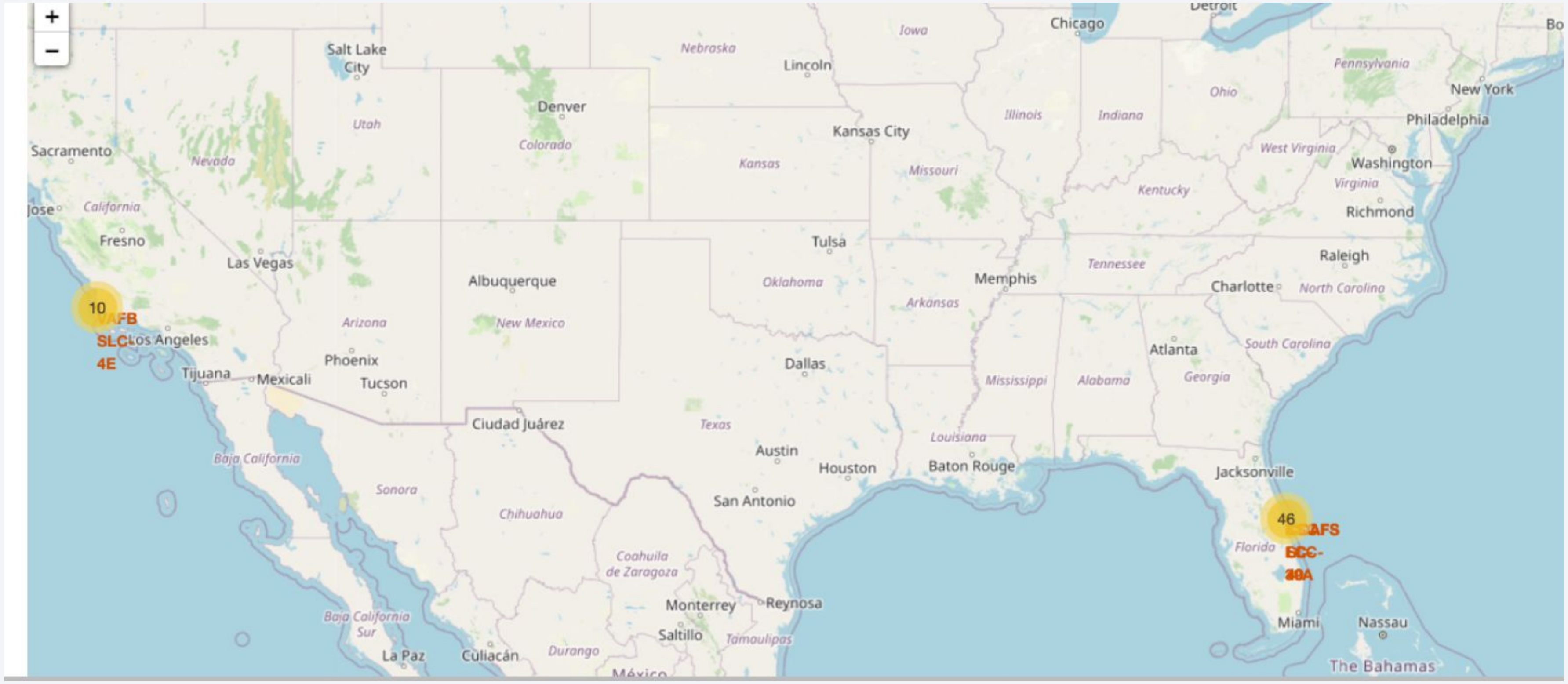
https://github.com/Gulxan/testreposit/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

- Exploratory data analysis results
- Space X uses 4 different launch sites
- The first launches were done to SpaceX itself and NASA
- The total payload mass booster is 4.5596kg
- 2016 is the year of success landing on drone ship out come
- The number of landing outcomes became as better as years goes

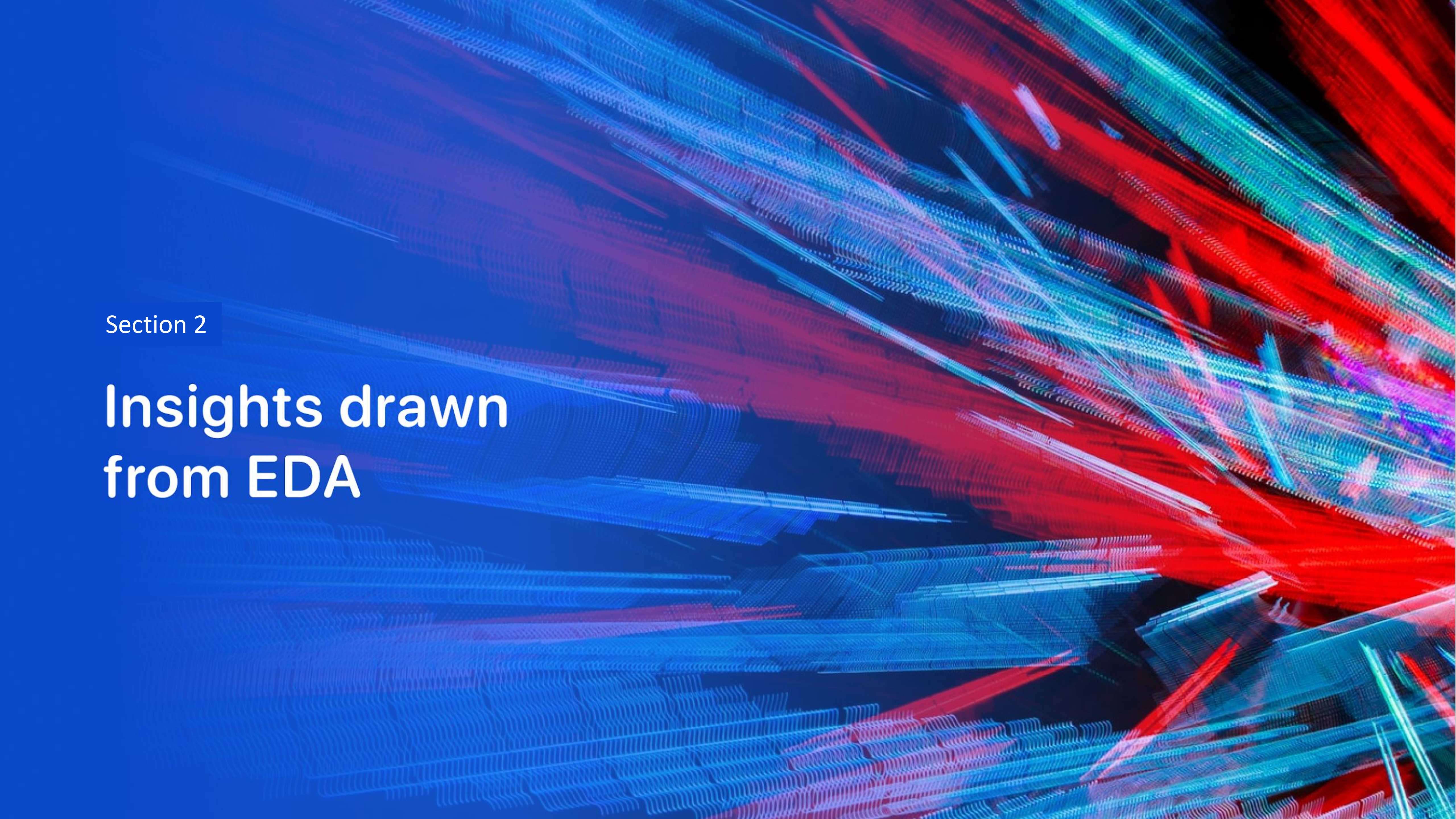
Results

- Using interactive analysis was possible to identify that launch sites to be near in highway ,railway and so on..



Results

Predictive Analysis showed that Decision Tree Classifier is the best model to predict successful landings

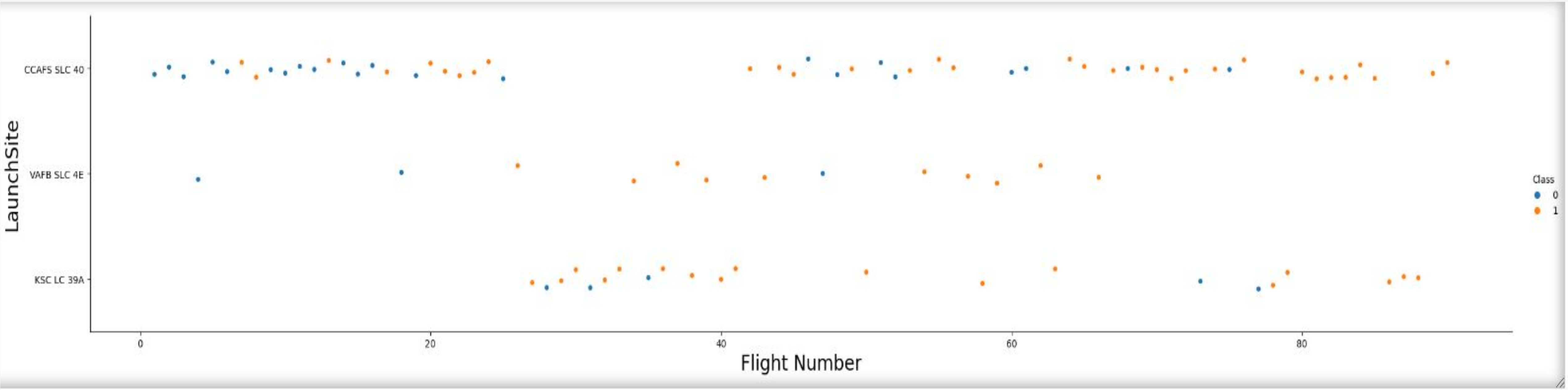


Section 2

Insights drawn from EDA

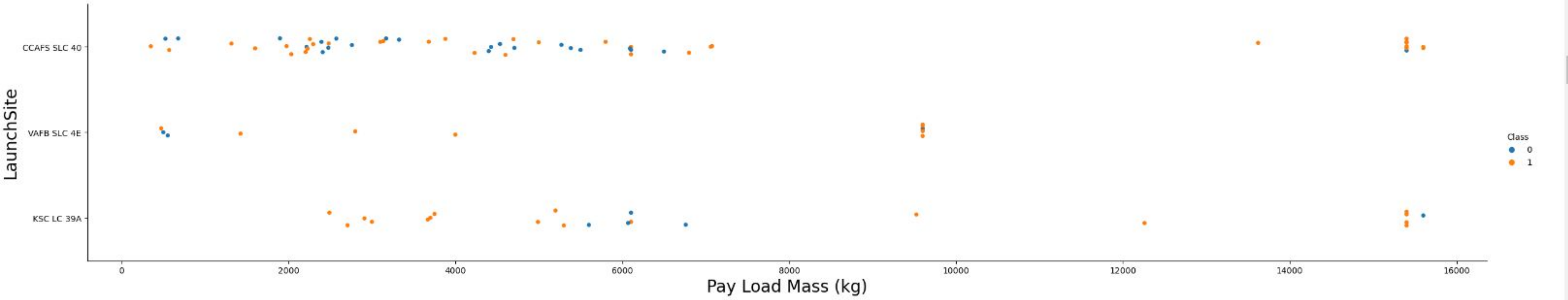
Flight Number vs. Launch Site

- Best Launch site is CCAF5 SLC 40, next VAFBSLC 4E and KSC LC 39A are the top 3.
- Over time, success rate is going upward direction



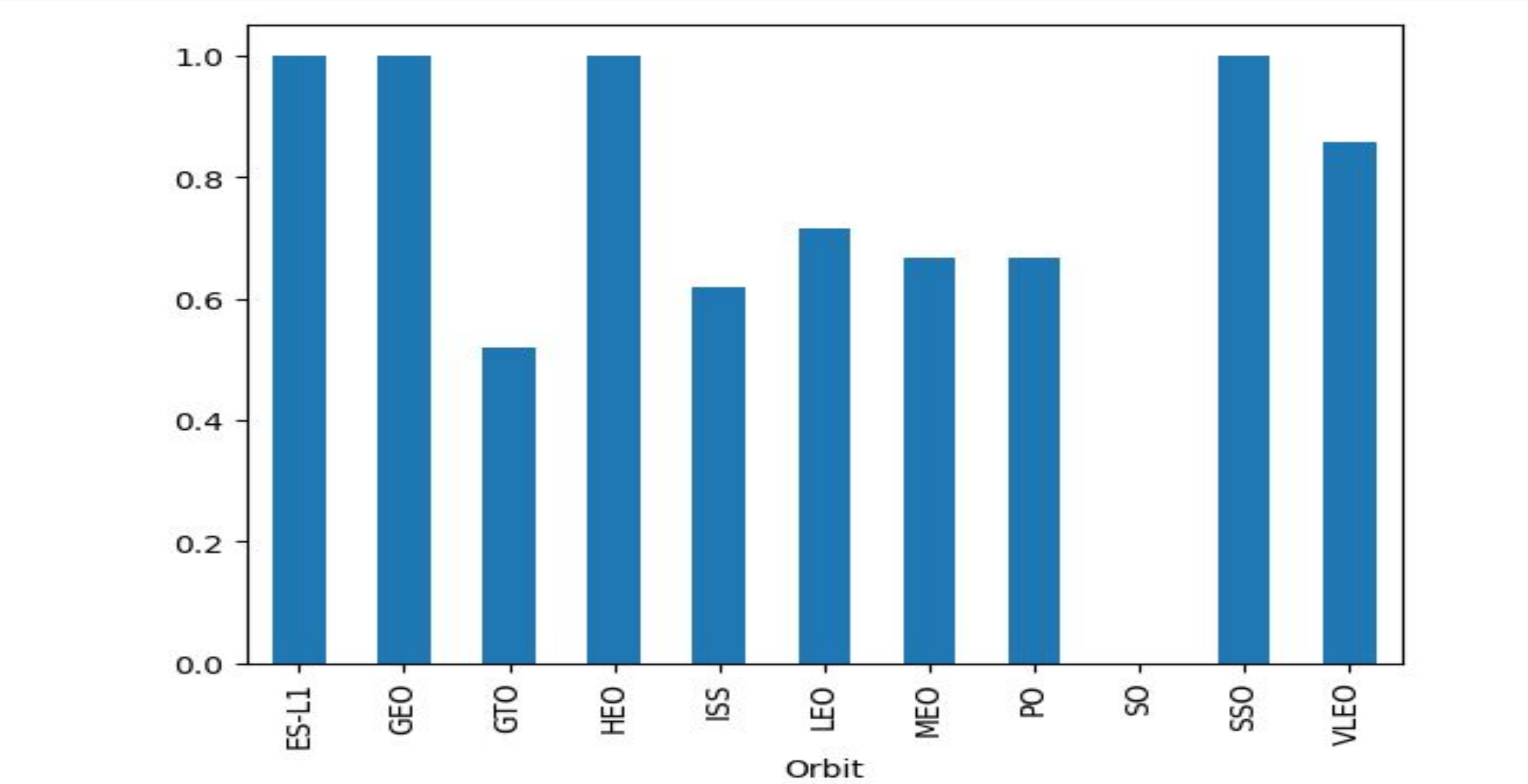
Payload vs. Launch Site

- Payloads over 9000 kg have good success rate
- Launch site CCAFS SLC 40 AND KSCLC39A are the possible sites for 12000 kg payloads



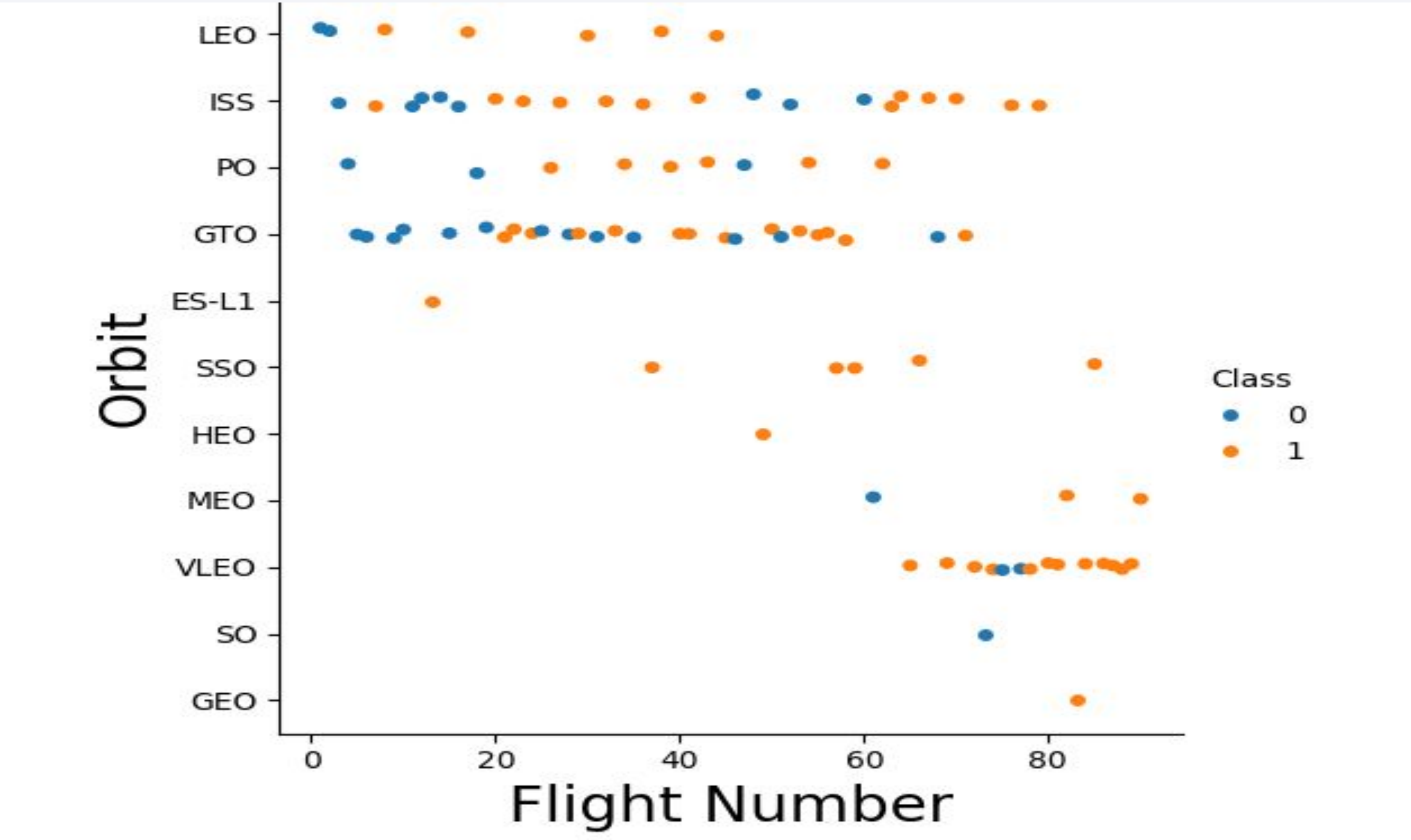
Success Rate vs. Orbit Type

- The biggest success rates happens to orbits are shown below



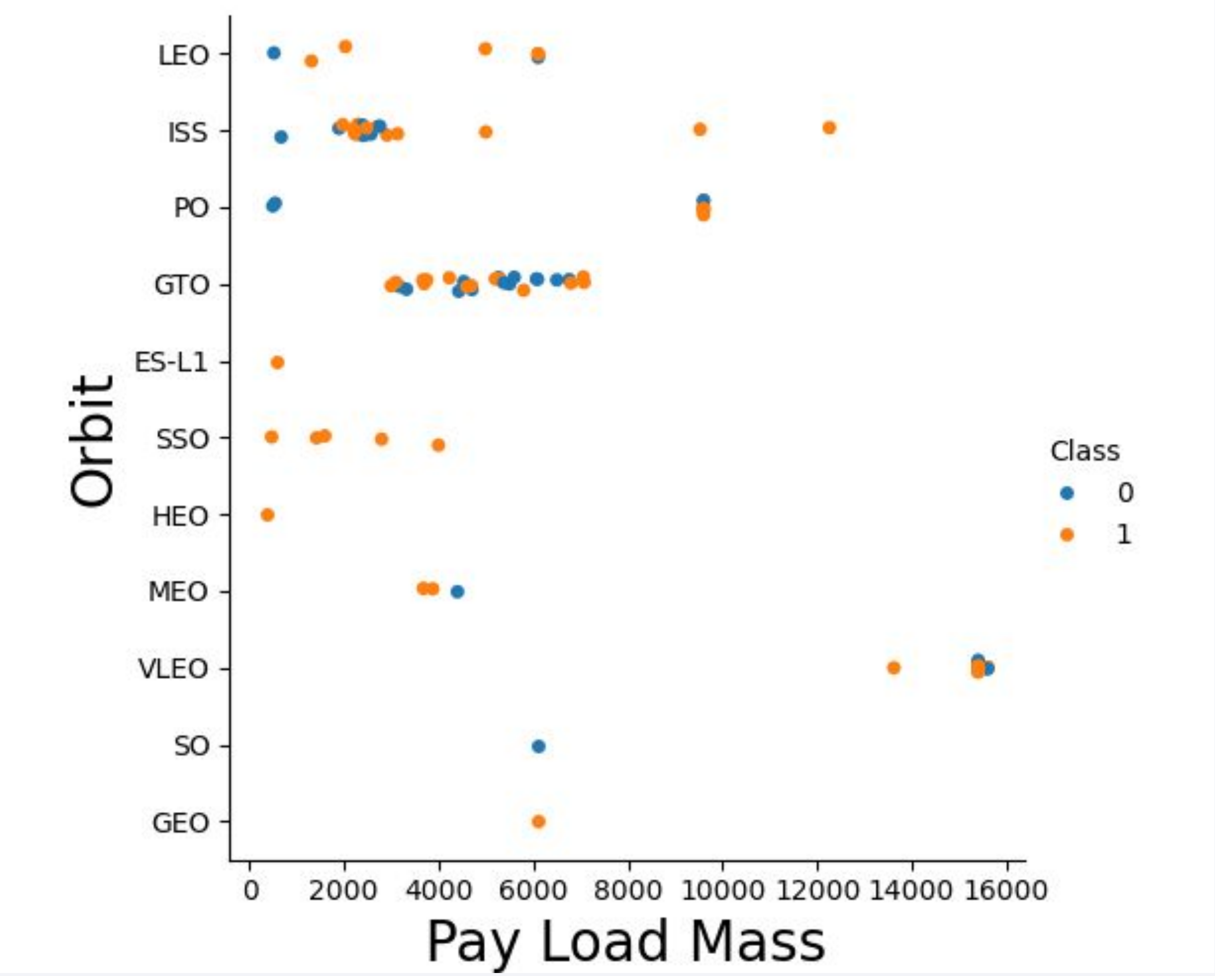
Flight Number vs. Orbit Type

- Below is clearly shows that success rate goes up over time.



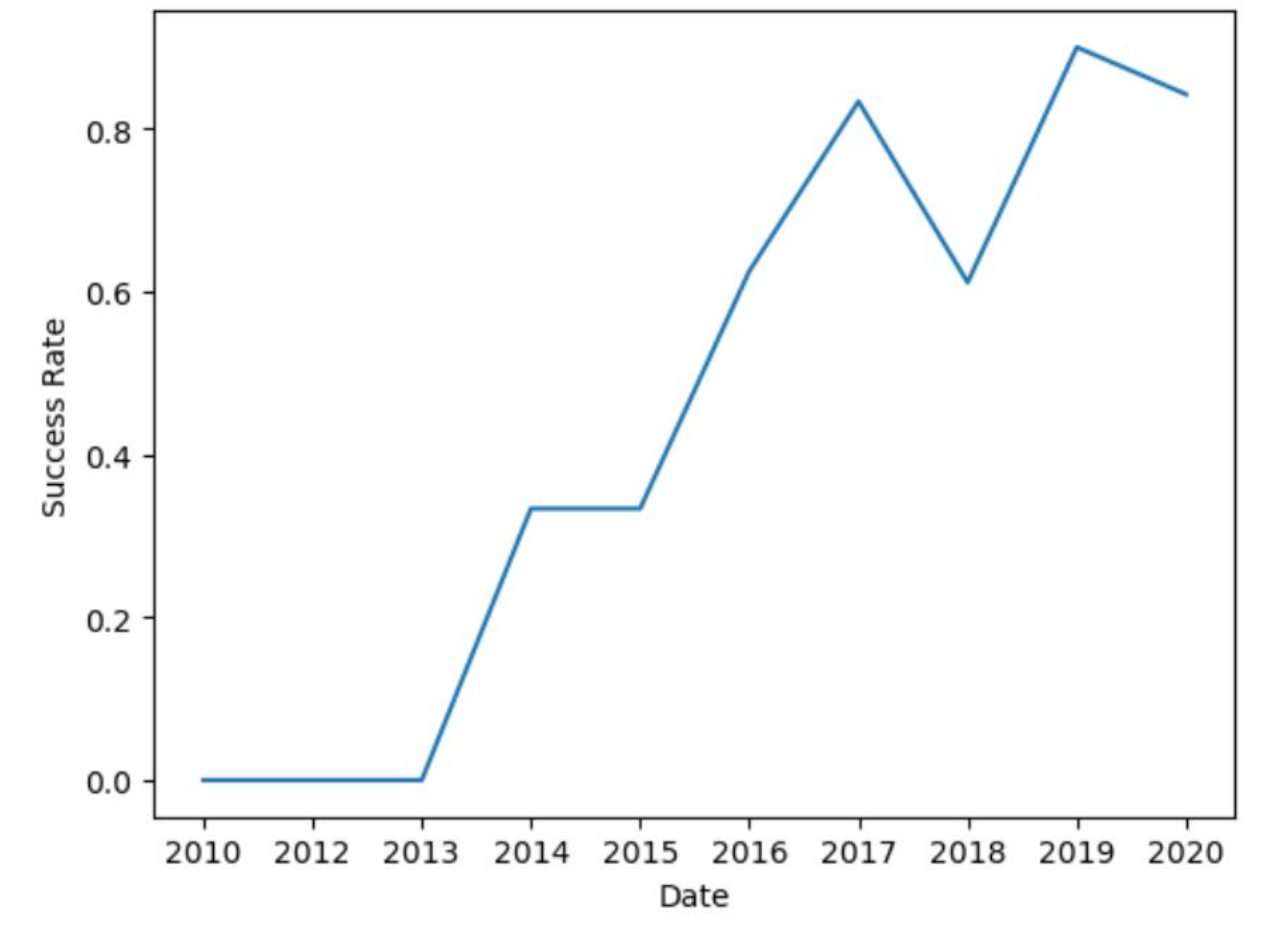
Payload vs. Orbit Type

- Show there are no relation between payload and success rate to orbit GTO



Launch Success Yearly Trend

- Shows that success rate started in 2013 upwards trend
- Major breakthrough in on 2015 onwards

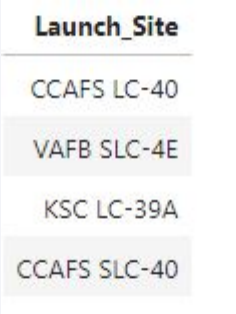


All Launch Site Names

SELECT DISTINCT Launch_Site FROM SPACEXTABLE;

The distinct launch sites include:

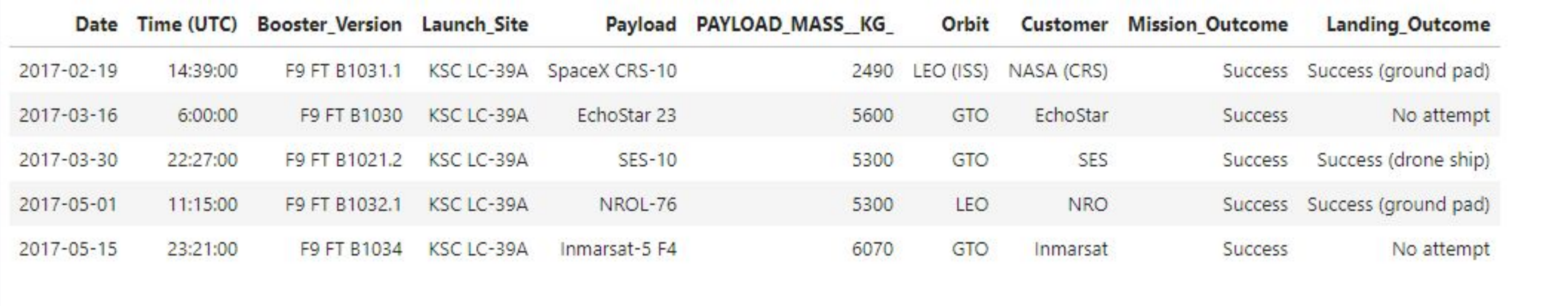
- CCAFS LC-40
- VAFB SLC-4E
- KSC LC-39A
- CCAFS SLC-40
- This was most evident through the following SQL query:



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

Launch Site Names Begin with 'KSC'

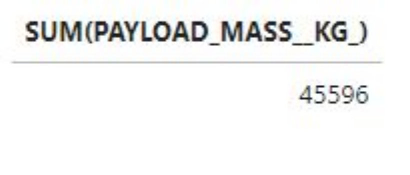
- Find 5 records where launch sites’ names start with `KSC`



Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-03-16	6:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-05-01	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
2017-05-15	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

Total Payload Mass

- Calculate the total payload carried by boosters from NASA



SUM(PAYLOAD_MASS__KG_)
45596

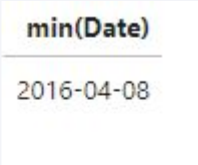
Total payload carried by boosters from NASA in 45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Present your query result with a short explanation here

First Successful Landing Outcome on droneship Date

- Find the dates of the first successful landing outcome on drone ship.



min(Date)
2016-04-08

The first successful landing outcome on drone ship is April 8 2016

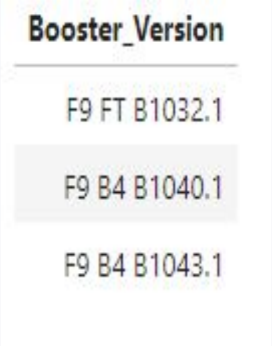
Successful ground pad Landing with Payload between 4000 and 6000

The boosters which have successfully landed on a ground pad and had payload mass greater than 4000 but less than 6000 kilograms include:

F9 FT B1032.1

F9 FT B4 B1040.1

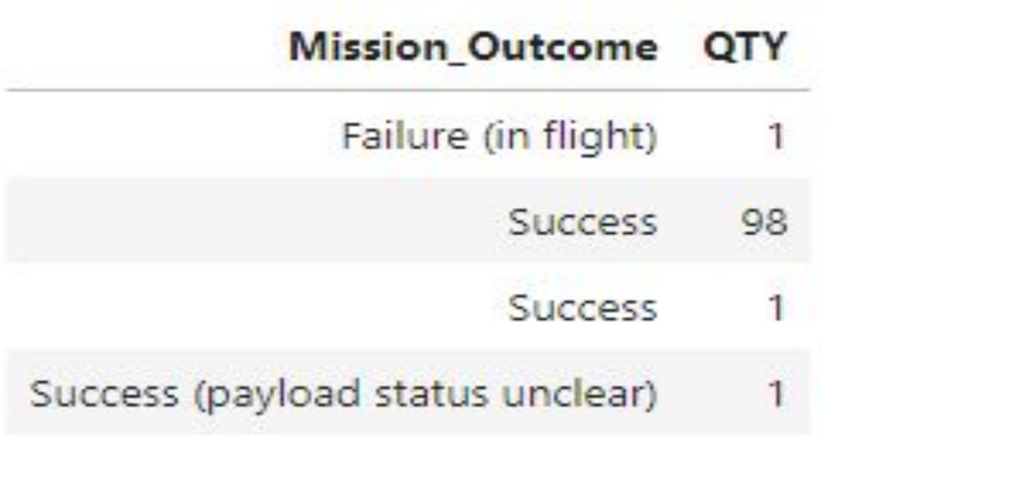
F9 FT B4 B1043.1



Booster_Version
F9 FT B1032.1
F9 B4 B1040.1
F9 B4 B1043.1

Total Number of Successful and Failure Mission Outcomes

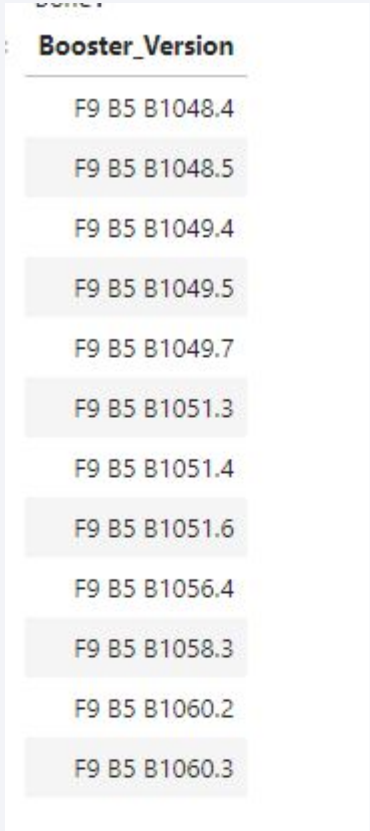
- Calculate the total number of successful and failure mission outcomes



Mission_Outcome	QTY
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

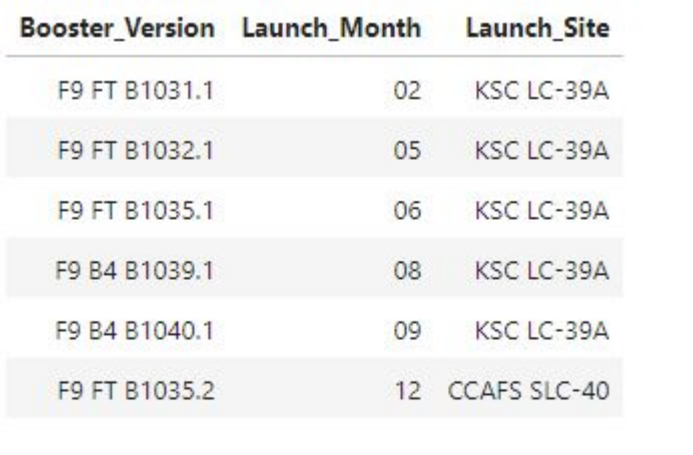
- List the names of the booster which have carried the maximum payload mass



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

2017 Launch Records

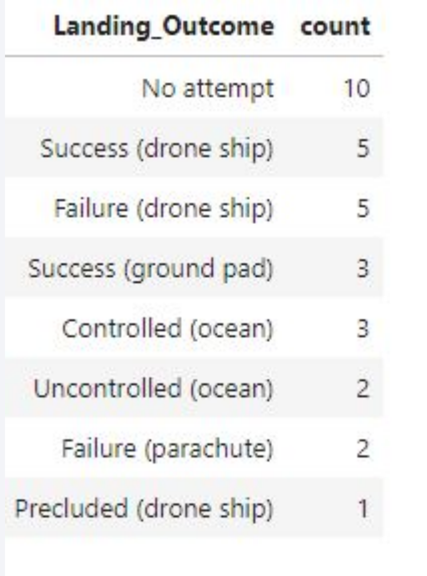
- List the records which will display the month names, successful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017



Booster_Version	Launch_Month	Launch_Site
F9 FT B1031.1	02	KSC LC-39A
F9 FT B1032.1	05	KSC LC-39A
F9 FT B1035.1	06	KSC LC-39A
F9 B4 B1039.1	08	KSC LC-39A
F9 B4 B1040.1	09	KSC LC-39A
F9 FT B1035.2	12	CCAFS SLC-40

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

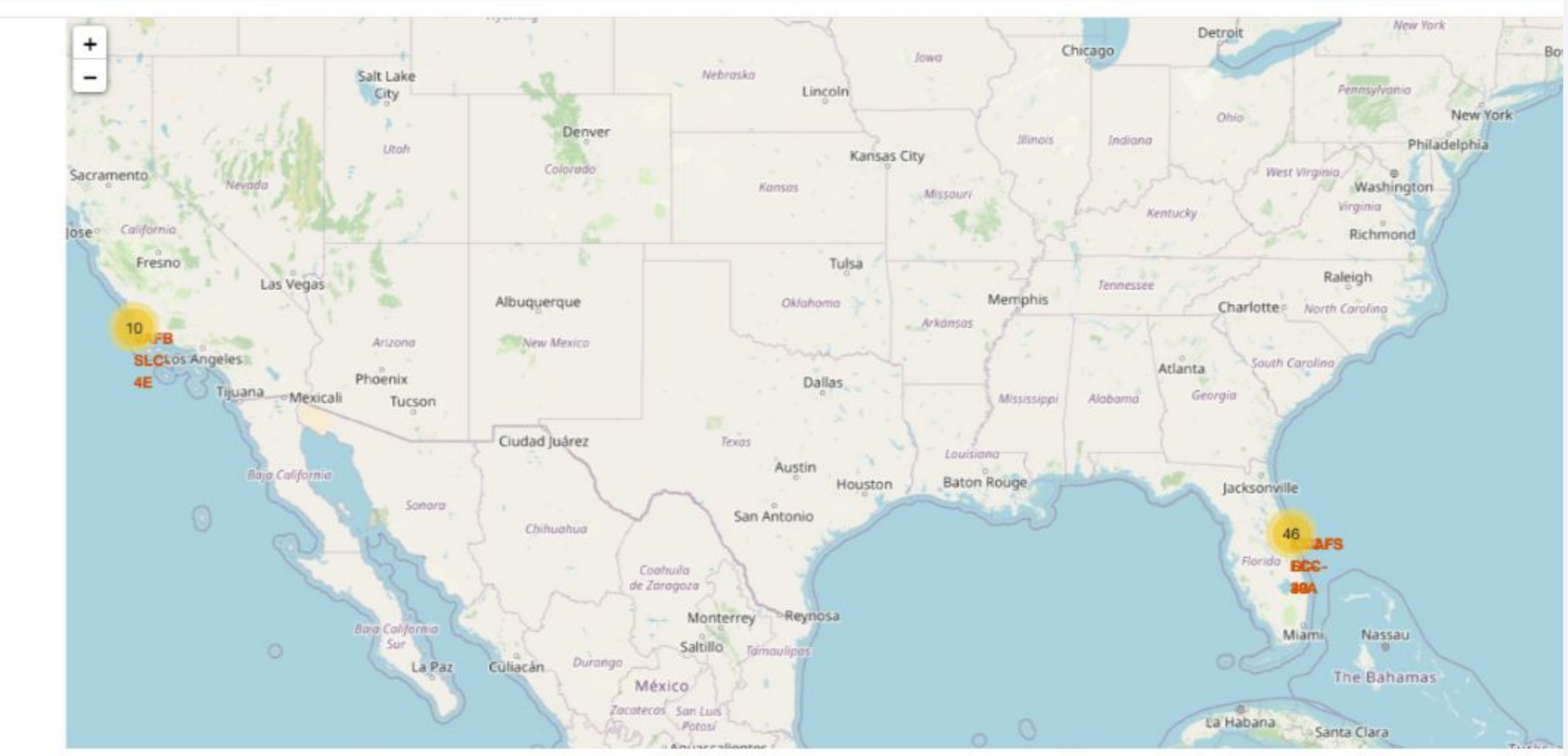


Landing_Outcome	count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

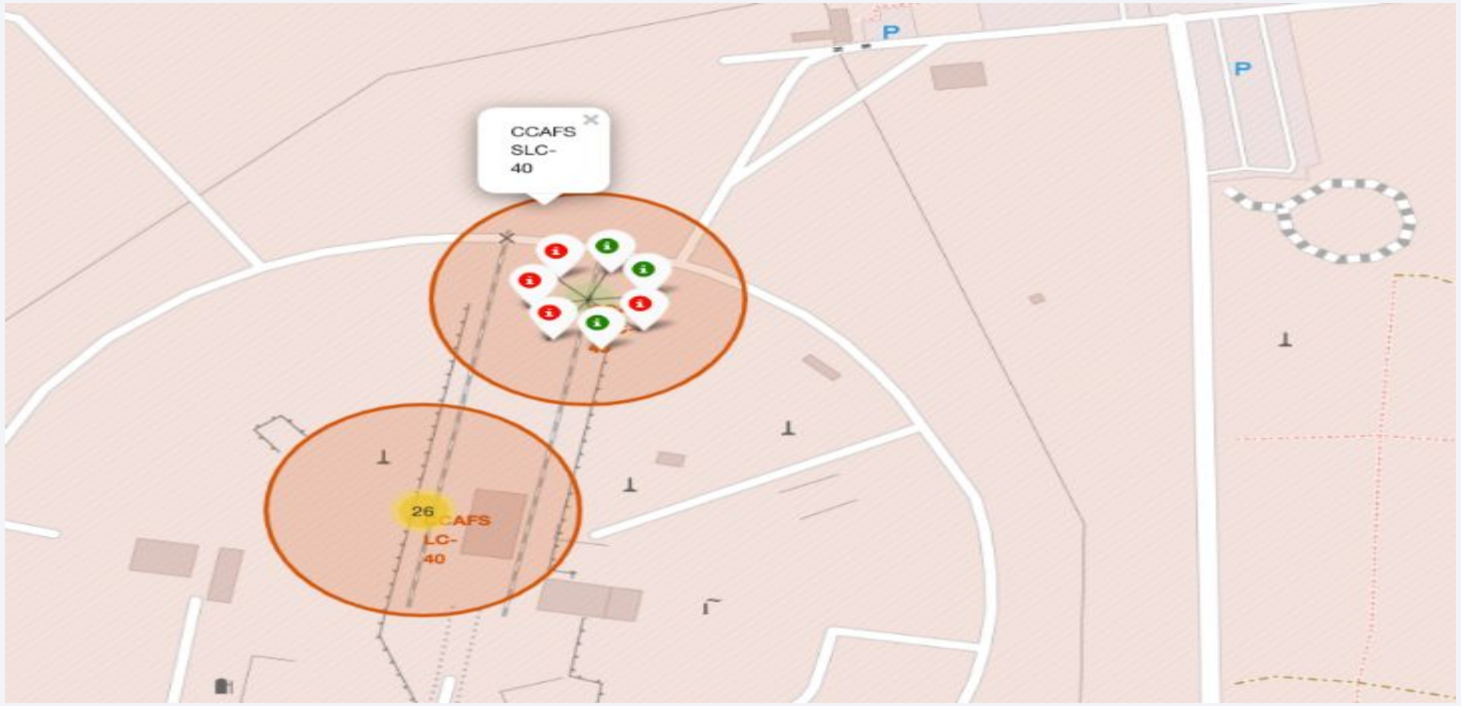
Section 3

Launch Sites Proximities Analysis

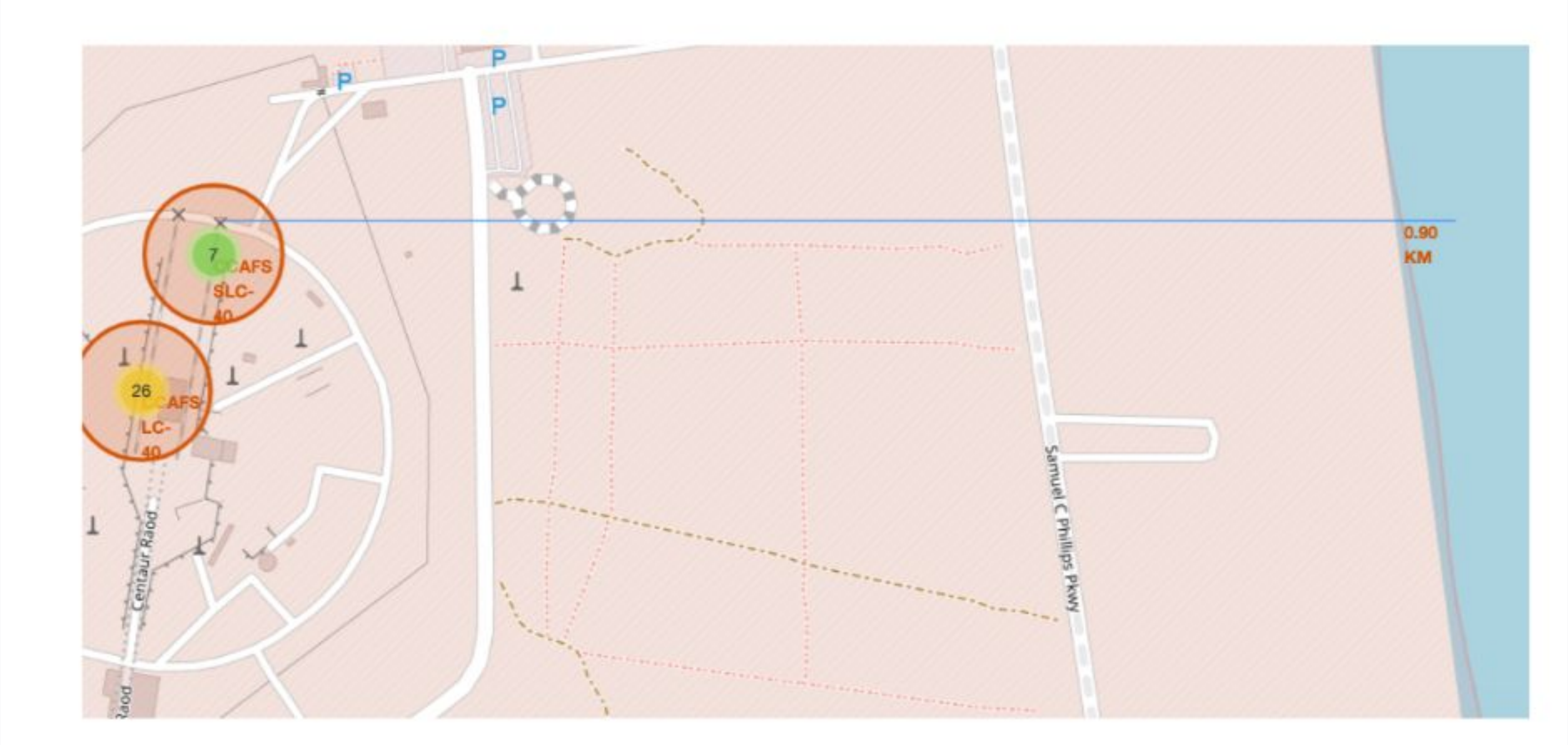
All Launch Sites

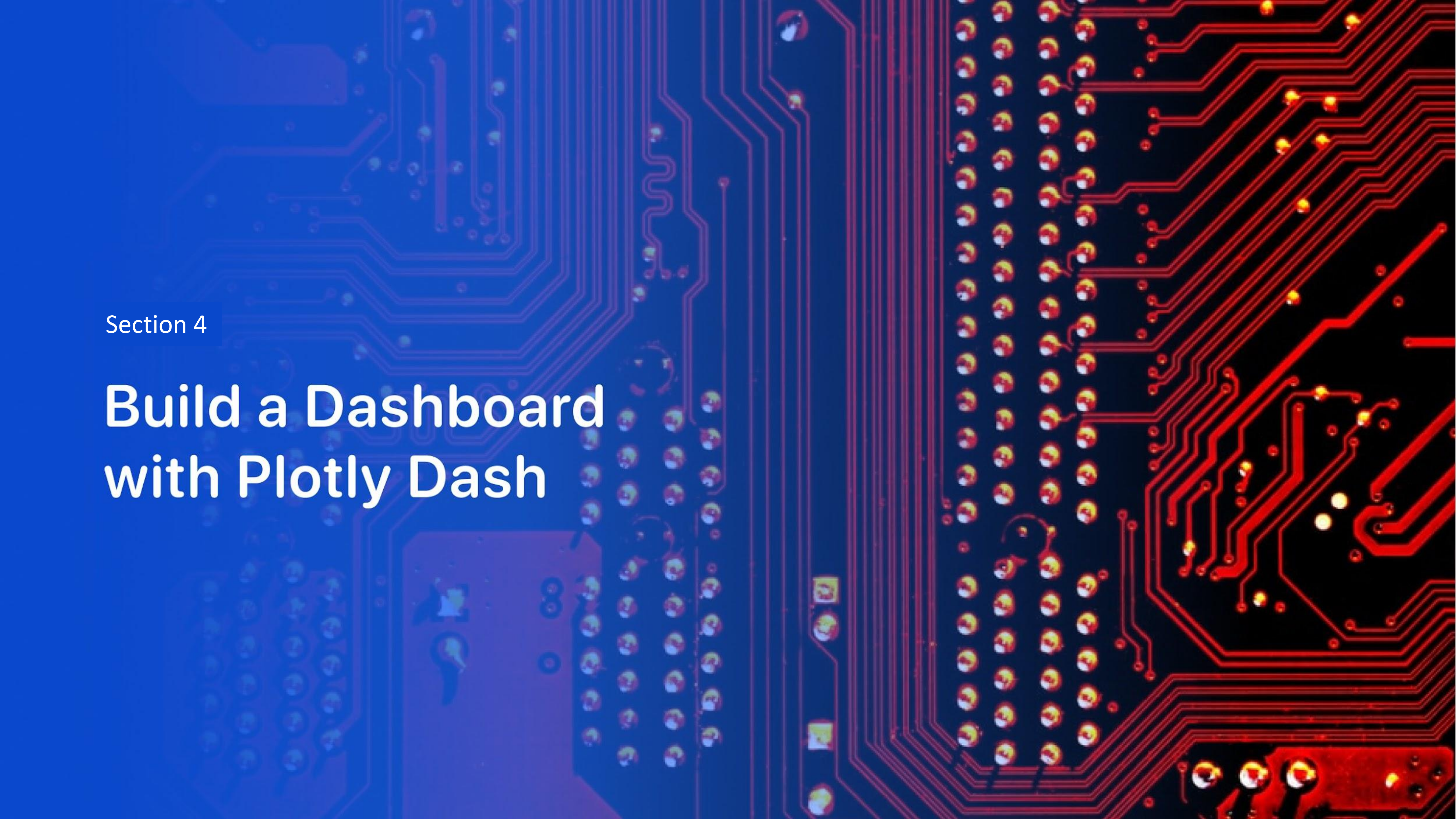


<Launch outcome



Launch sites and geographical advantages

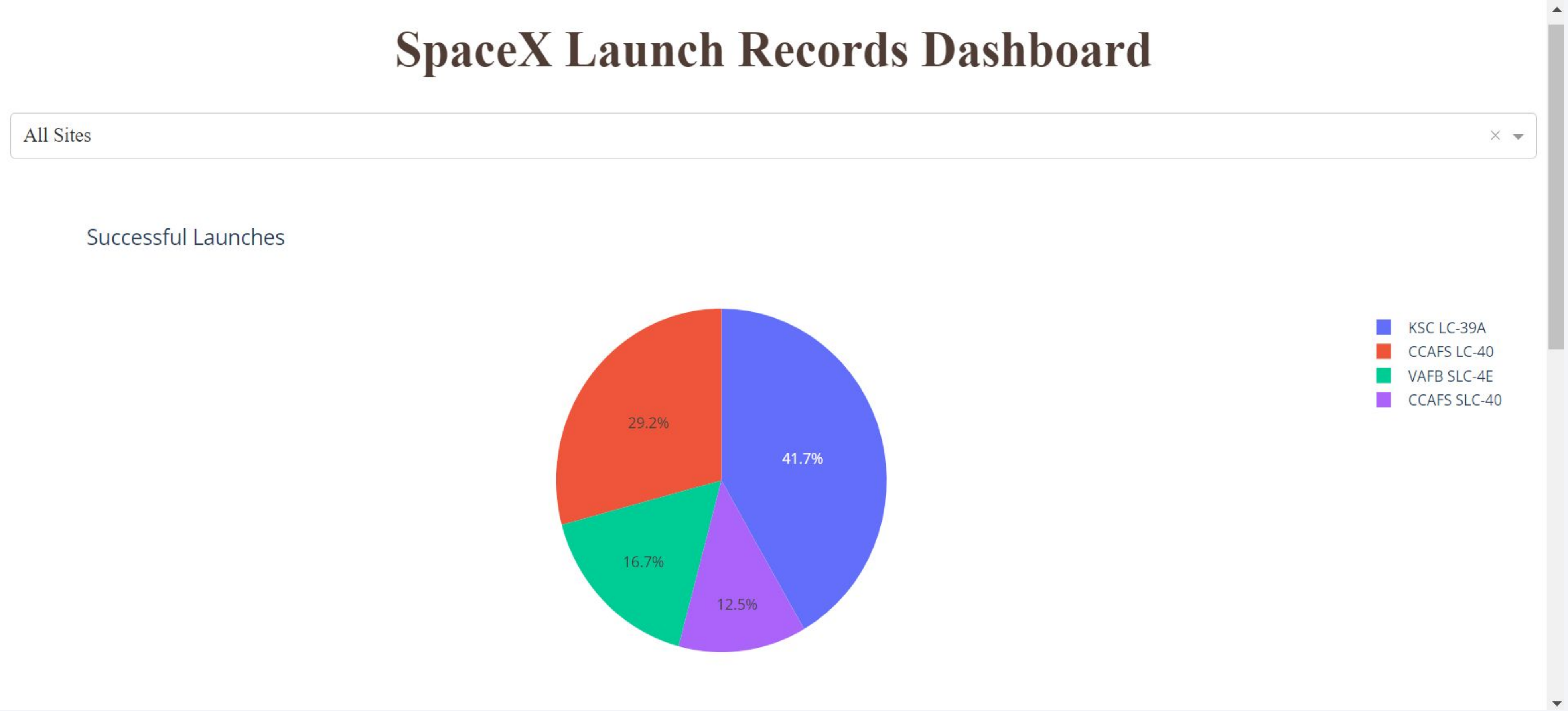




Section 4

Build a Dashboard with Plotly Dash

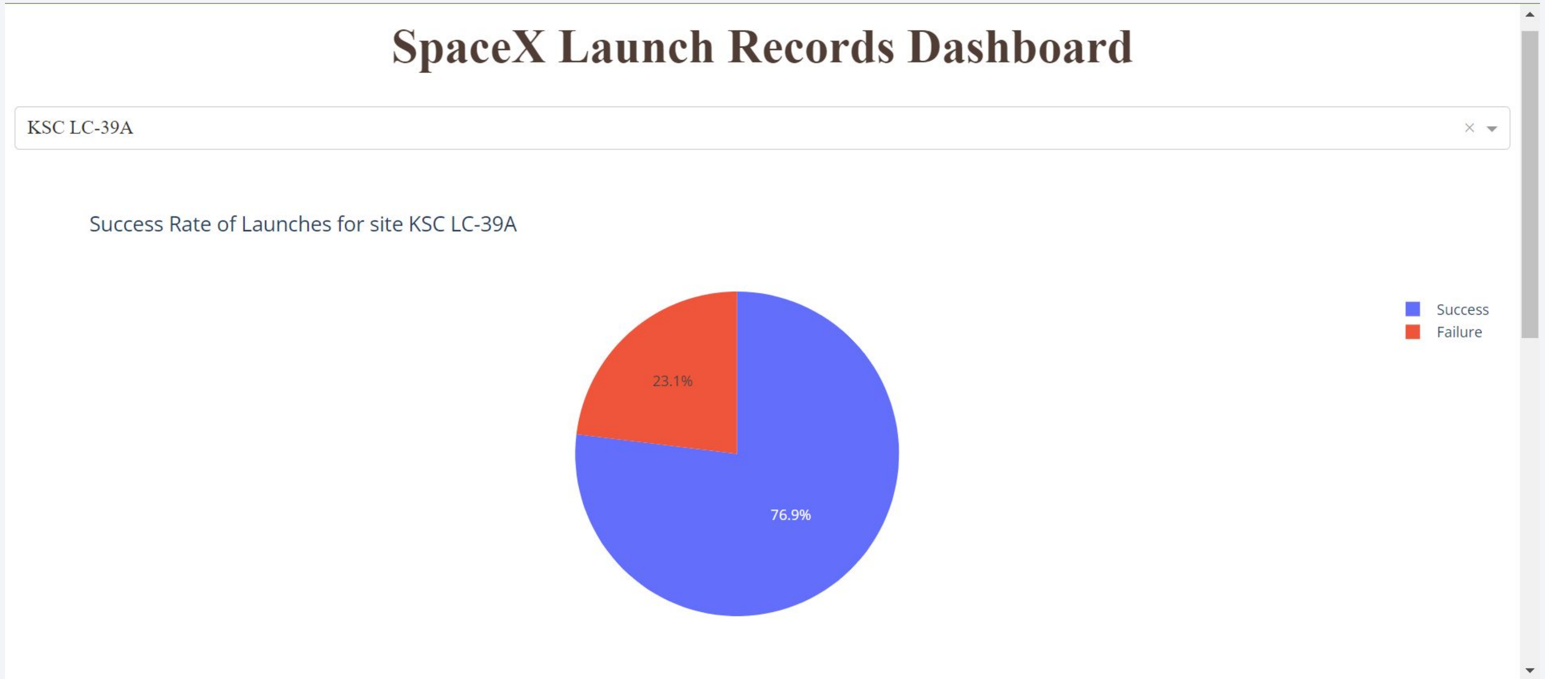
Successful Launches by site



SpaceX Launch Records Dashboard

Launch Success Ratio for KSC LC-39A

Success rate is 77%

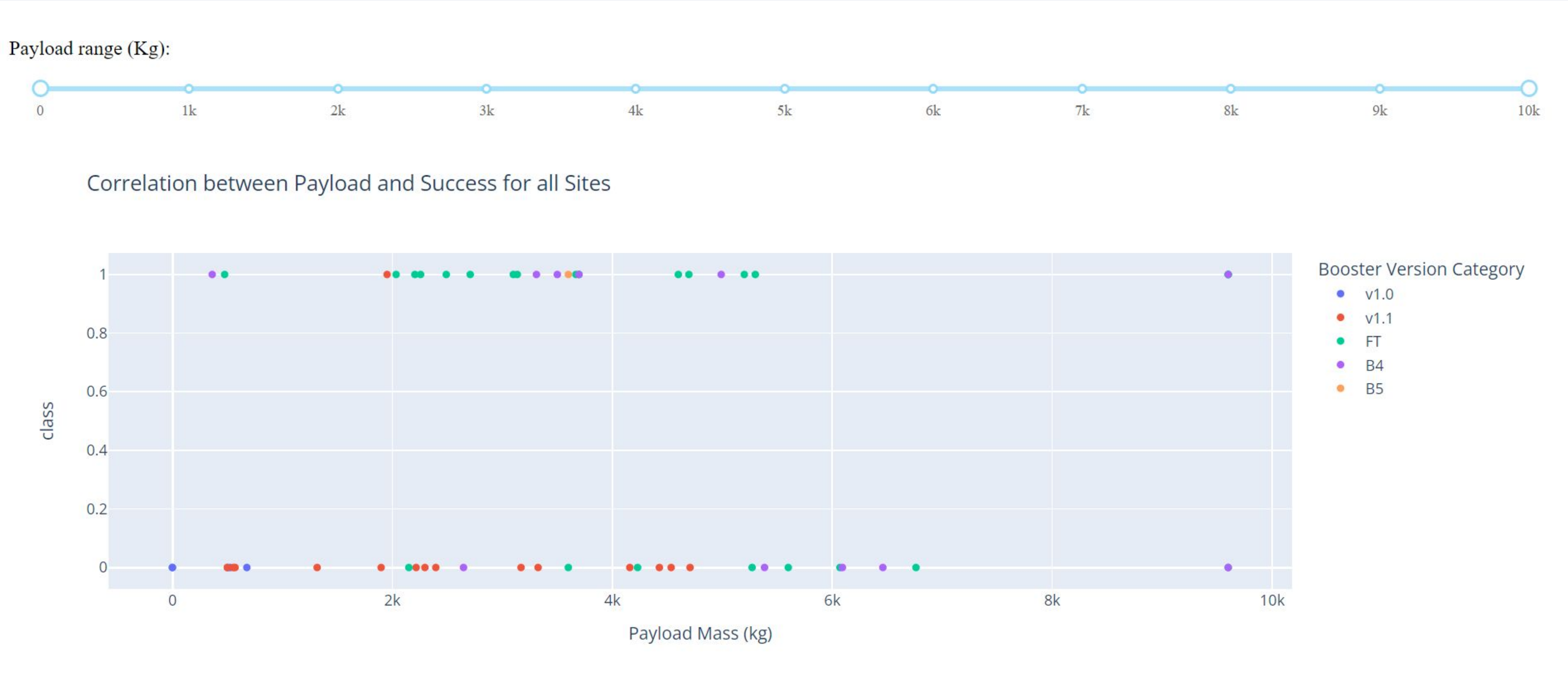


SpaceX Launch Records Dashboard

KSC LC-39A

Success Rate of Launches for site KSC LC-39A

Payload vs. Launch Outcome



Payload range (Kg):

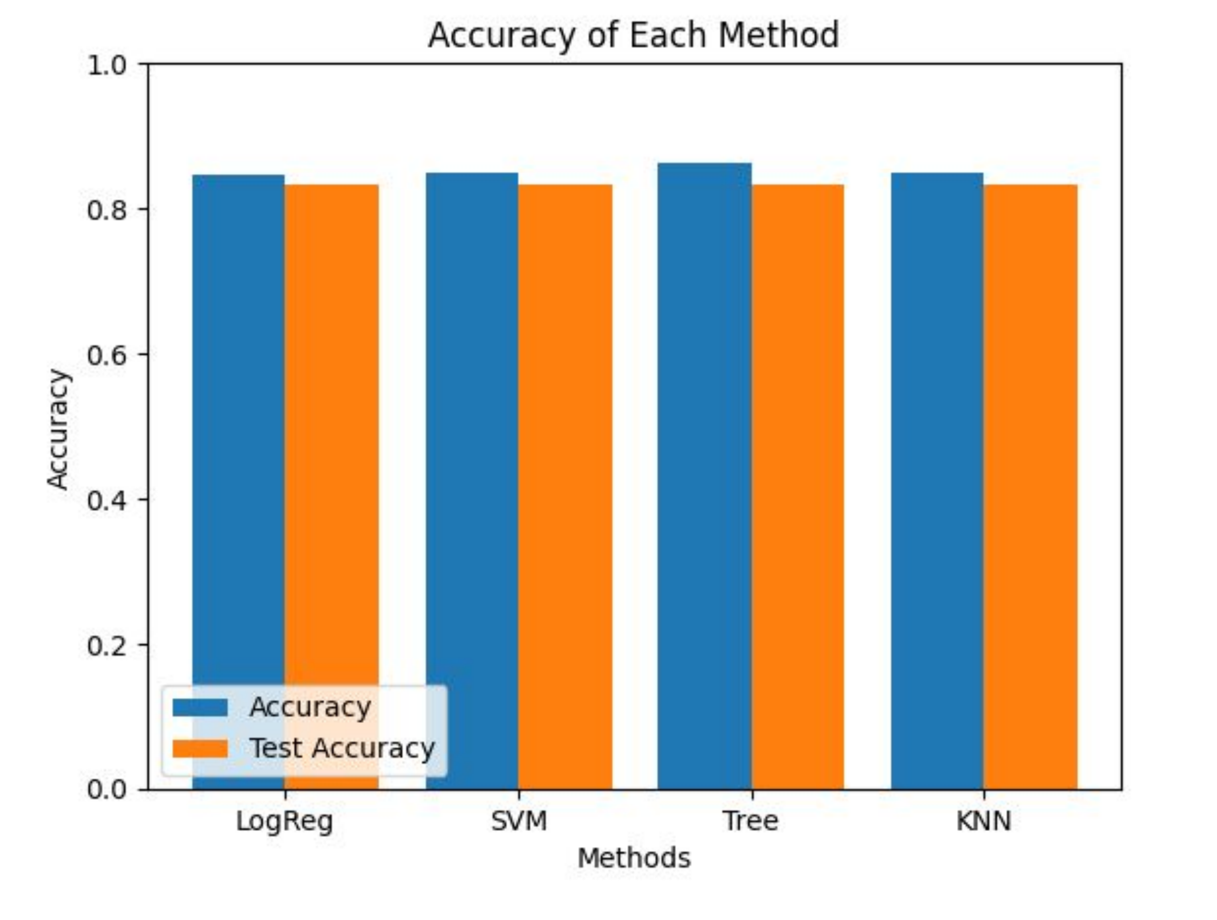
Correlation between Payload and Success for all Sites

Section 5

Predictive Analysis (Classification)

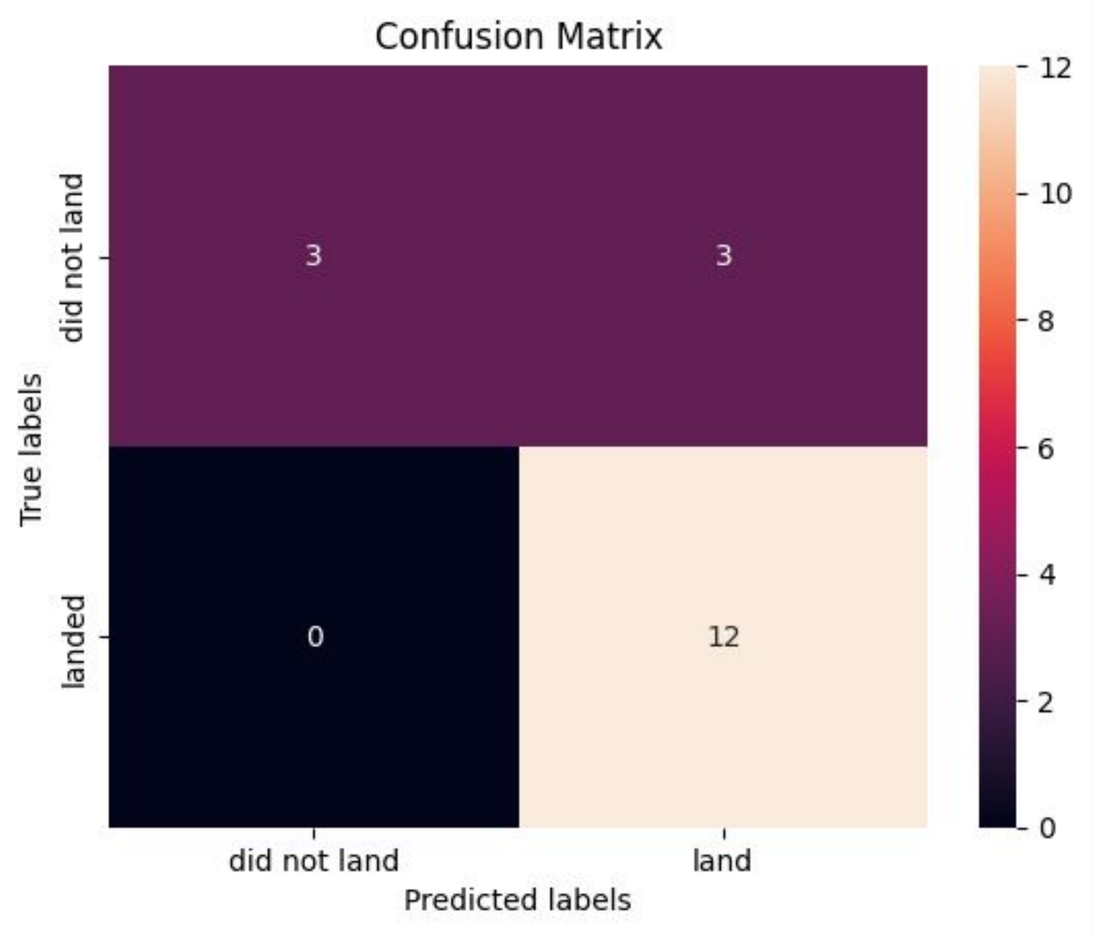
Classification Accuracy

- Visualize the built model accuracy for all built classification models, in a bar chart
  - The bar chart demonstrates the classification accuracy for the four machine learning models utilized for prediction.
  - It is clear that the Decision Tree Classifier performed slightly best on our dataset.



Confusion Matrix

- While the Decision Tree Classifier performed better, its weakness was false positives: predicting that a certain landing was successful when in fact it wasn’t.

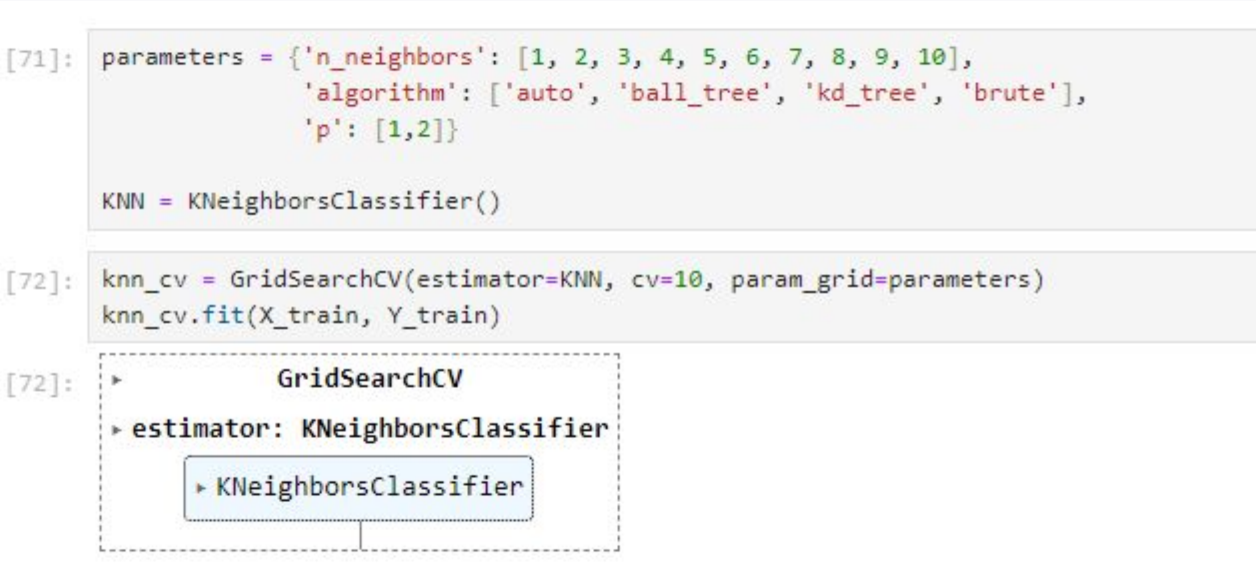


Conclusions

- Different data sources were analysed, refining conclusion along the process
- The best launch site is KSC LC-39A
- Max payload bigger than 8000 are more successful than less than 8000
- Overtime, success rate going up (2013-2020)
- Decision Tree is the most effective to predict successful landings.

All the demonstrated notebooks are stored on a master public GitHub repository, which can be accessed through this link.

Appendix



```
[71]: parameters = {'n_neighbors': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10],
                    'algorithm': ['auto', 'ball_tree', 'kd_tree', 'brute'],
                    'p': [1,2]}

      KNN = KNeighborsClassifier()
```

```
[72]: knn_cv = GridSearchCV(estimator=KNN, cv=10, param_grid=parameters)
      knn_cv.fit(X_train, Y_train)
```

```
[72]: ▸           GridSearchCV
      ▸ estimator: KNeighborsClassifier
            ▸ KNeighborsClassifier
```

Thank you!