IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Diego Andrés Monroy Cárdenas
10/08/23

Outline

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

Executive Summary

Data science project is developed to predict a successful firs-stage recovery of a rocket spaceship based on public data information of Space X company provided by their API and Wiki pages with launches recompiled information.

After a data collection and wrangling process, relationships between main features, needed to models training process and relevant information to stakeholders are investigated with an exploratory data analysis process and summarized using statistics plots representation with interactives dashboards. Logistic Regression, Support Machine Vectors, Classification trees and K-Nearst Neighbors models using Grid search methodology to tune hyperparameters are trained and tested to achieve prediction requirement. Finally, results evaluation models are presented and in appendix section key steps and methods used in this project are consigned.

Introduction

Several stakeholders are interested in create a new spaceships company to participate in the new private space exploration era. They has identified like a competitive advantage the technical capacity to recover the first-stage component of a rocket launch. This capacity not only save a lot of resources and investment but is already being exploited by market competitors.

One of this established competitors is Space X company. Using their public data, the main objective of stakeholders is to obtain a model to predict if a first-stage component is going to be recover successfully after a rocket launch. Also, there are interested in to learn another relevant technical information to the business-like types of boosters, payloads delivered, orbits achieved, launch locations, costumers, success outcomes and historical development of launch missions of Space X company.

This data science project is oriented to analysis which data available is relevant to be used as an input to build a binary classification model to predict the success or failure in first-stage recovery. Moreover, relationships of different data are presented to give a complete context of business to stakeholders.

Section 1

Methodology

Methodology

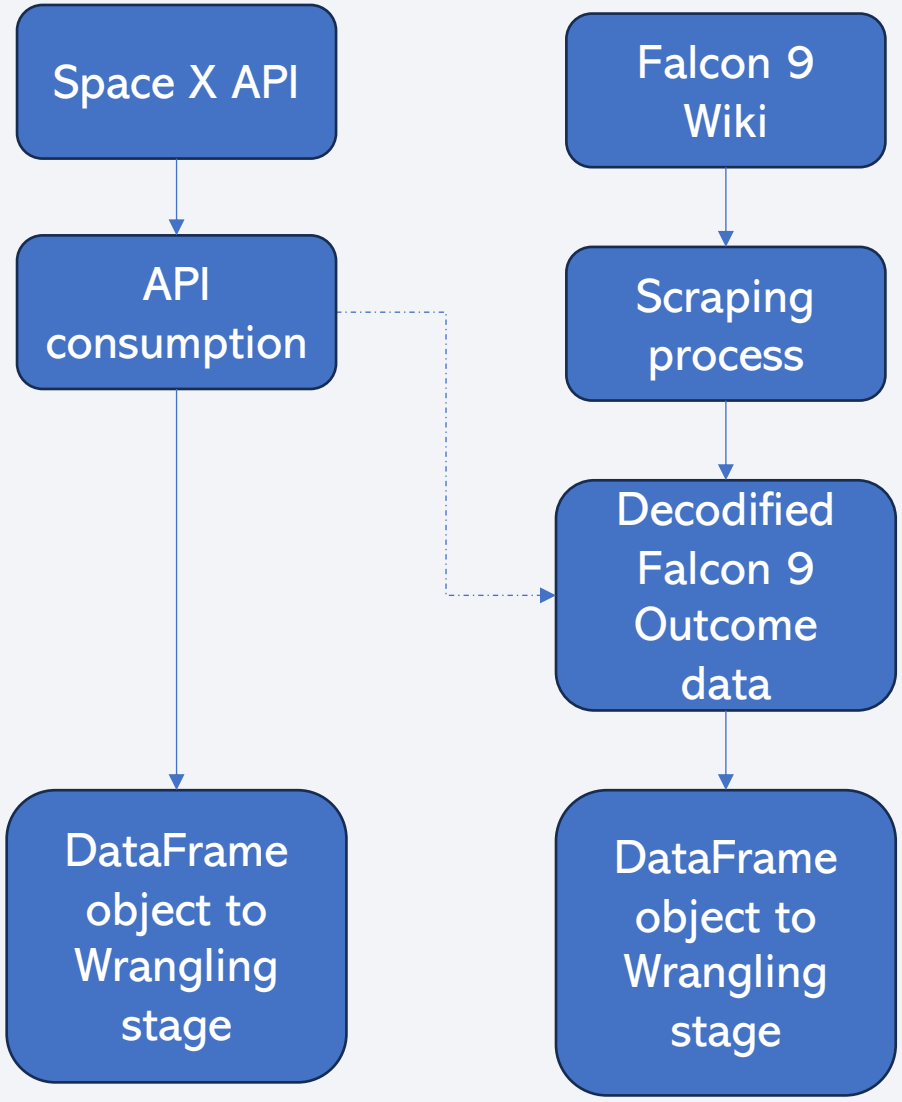
- Data collection
- Data wrangling
- Exploratory data analysis (EDA) using visualization and SQL
- Interactive visual analytics using Folium and Plotly Dash
- Predictive analysis using classification models

Data Collection

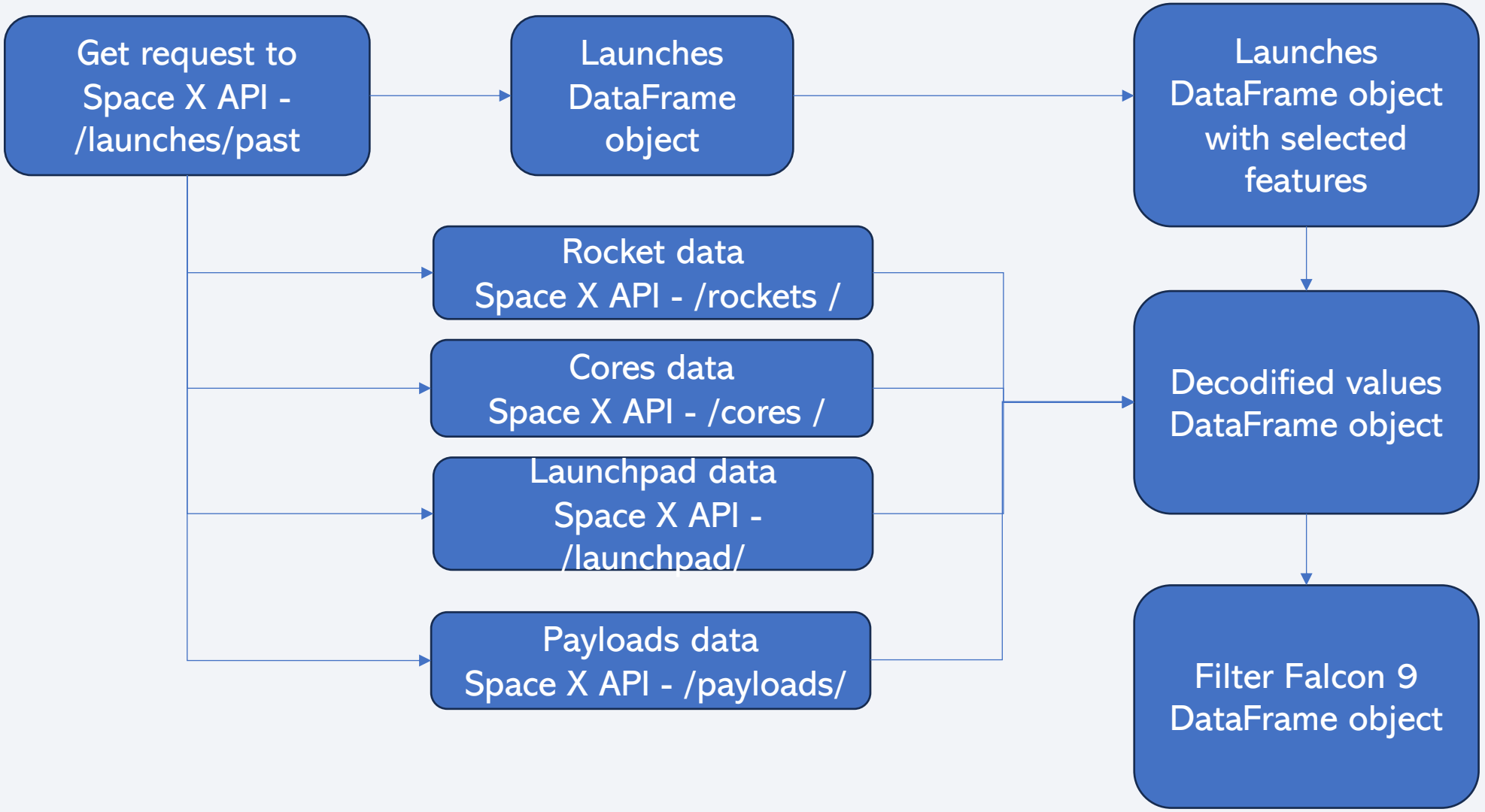
Data collection processes was perform using two different data sources:

- **SpaceX API:** This API has data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.
- **Wikipedia - List of Falcon 9 and Falcon Heavy launches:** This is a popular data source that collect Falcon 9 historical launch records. From it, is possible to extract the required information about the outcome of the Falcon 9 first stage.

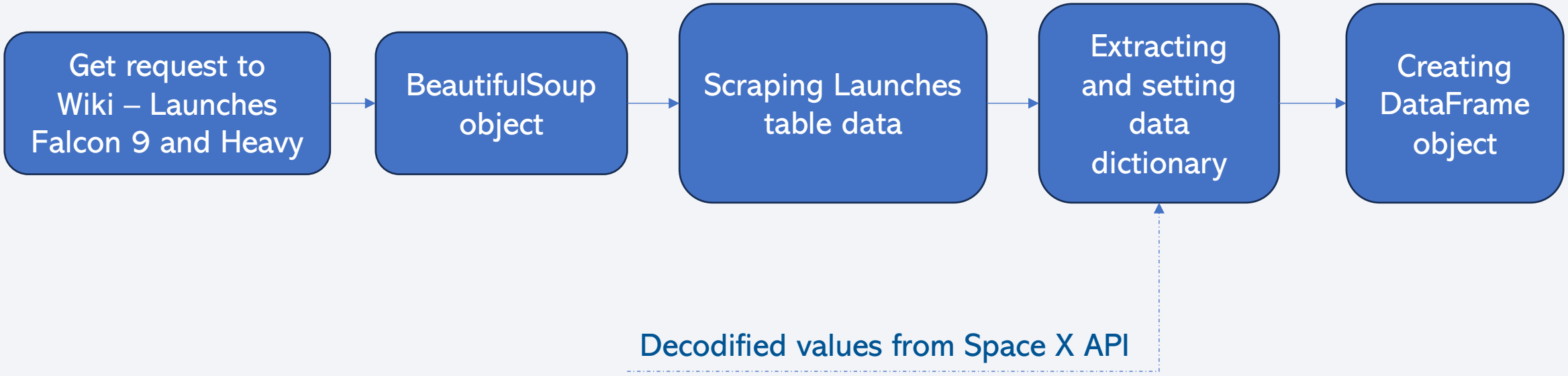
The data about Falcon 9 first stage outcome in wiki data source, have some codified information that must be formatted on a Wrangling stage. The information about this codes are available on the Space X API. In this way, data from both sources are combined and used to prepare data to further analysis.



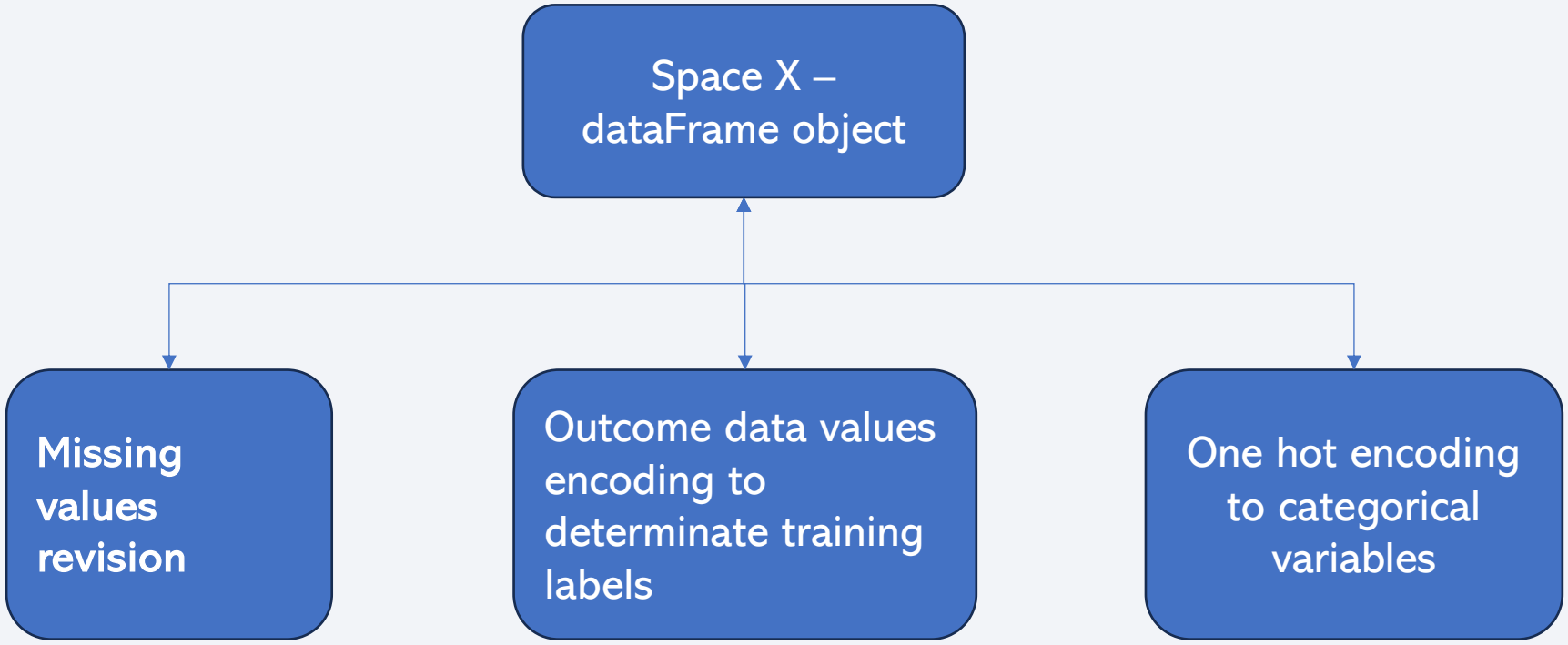
Data Collection – SpaceX API



Data Collection - Scraping



Data Wrangling



EDA with Data Visualization

The Exploratory Data Analysis with visualization were performed over the following features:

- **Flight Number VS Payload:** To understand how the continuous launch attempts would affect the launch outcome. On the scatter plot the outcome value is codified as color by each point.
- **Flight Number VS Launch Site:** To understand how the launch outcome is distributed by each launch site over all missions considered. On the scatter plot the outcome value is codified as color by each point.
- **Payload Mass VS Launch Site:** To understand how the different payload mass would affect the launch site selection. On the scatter plot the outcome value is codified as color by each point.
- **Orbit type VS Average Outcome :** To understand how the average launch outcome is distributed by each type of orbit type target.
- **Flight Number VS Orbit:** To understand how the continuous launch attempts to different target orbits would affect the launch outcome. On the scatter plot the outcome value is codified as color by each point.
- **Payload Mass VS Orbit :** To understand how the launch outcome is distributed by each orbit type target over all Payload Mass delivered. On the scatter plot the outcome value is codified as color by each point.
- **Flight Number VS Orbit:** To understand how the continuous launch attempts to different target orbits would affect the launch outcome. On the scatter plot the outcome value is codified as color by each point.
- **Date VS Success rate :** To understand how the success launch outcome tendency had varied over time.

EDA with SQL

The Exploratory Data Analysis with SQL were performed using the following queries:

- 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 succesful landing outcome in ground pad was achived.
- 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 records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months 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.

Build an Interactive Map with Folium

The launch success rate may depend on the location and key factors over the proximities of a launch site. Interactive visual analytics was used to discover those by analyzing the locations with `Folium` library. The following map objects were added to a map during the process:

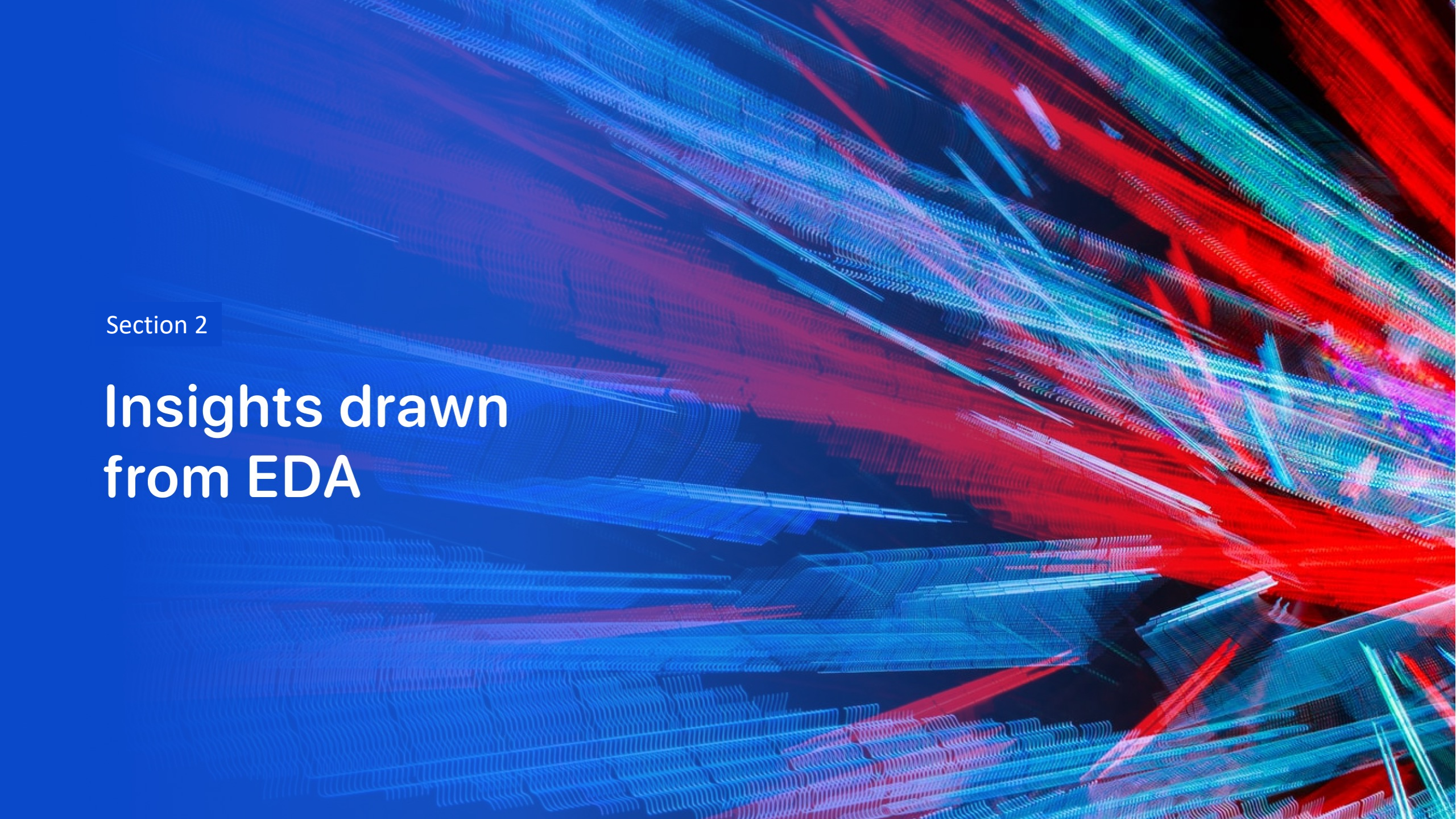
- Folium circles and markers over each launch site to visualize its geographical location.
- Folium marker cluster object over each site location in order to visualize the quantity of launch performed and and outcome values. (success and fail attempts).
- Folium marker and line objects to highlight distances from launch locations and key factors over the proximities like cities, railroads, proximities to coast, etc.

Build a Dashboard with Plotly Dash

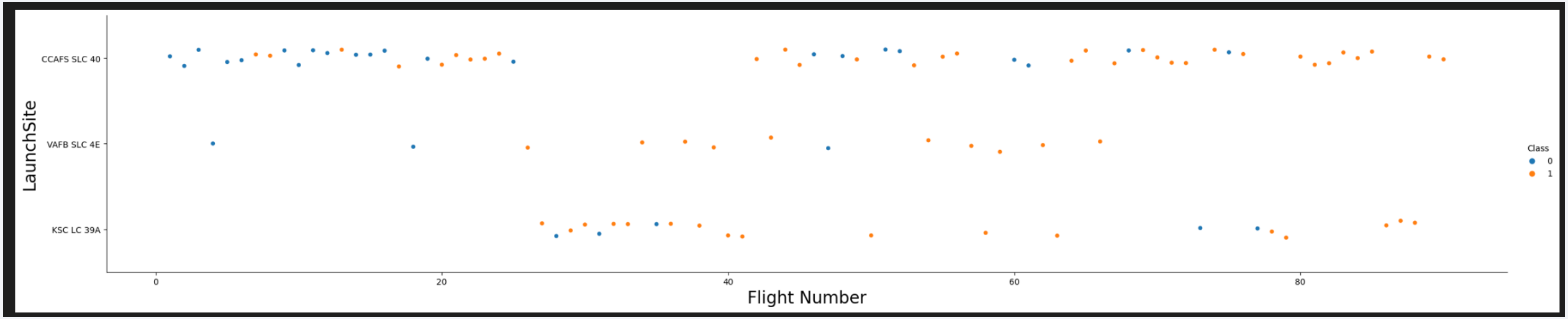
Interactive plots dashboard were created to visualize the relationship of some aspects like launch sites, payloads ranges, and boosters version performance against success rates. The following plots/graphs and interaction were added to the dashboard during the process:

- **Total success launches by site:** To understand how is the percentage of success outcomes over all launch sites.
- **Total success launches for site:** To understand how is the percentage of success outcomes by each site.
- **Correlation between Payload and success rate for all sites per booster version:** To understand how is the different booster version performance influence on success outcomes over a specific range of payload.
- **Correlation between Payload and success rate for site per booster version:** To understand how is the different booster version performance influence on success outcomes over a specific range of payload and launch site.

Predictive Analysis (Classification)

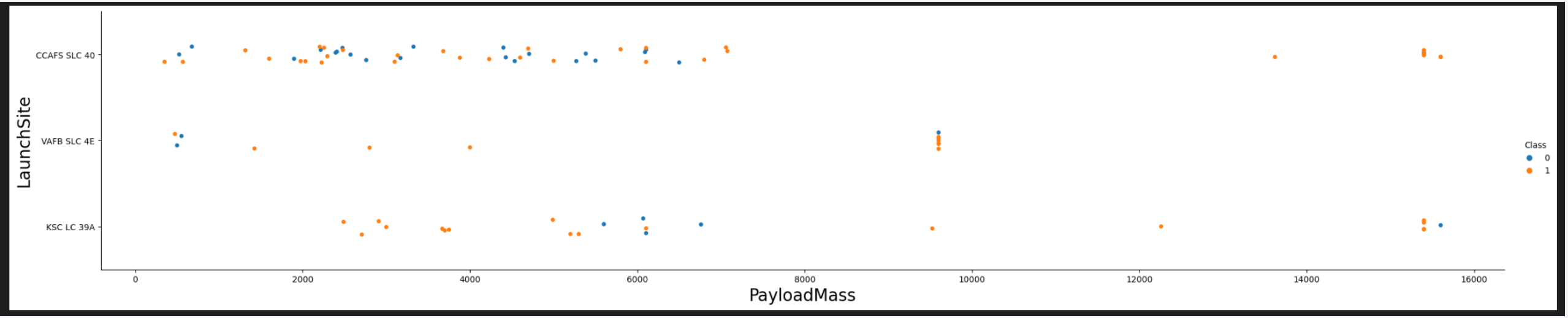

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



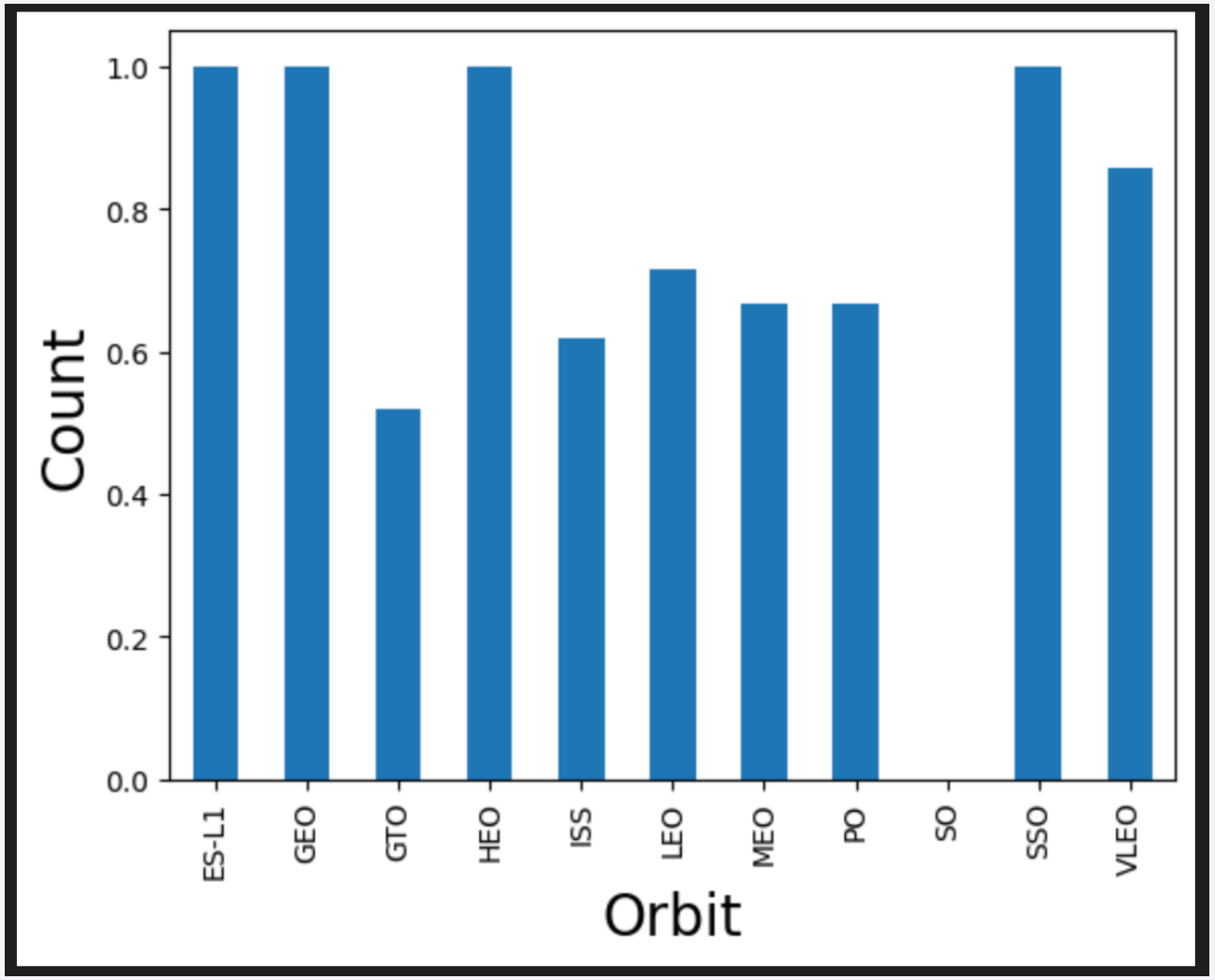
In the launch site CCAFS SLC 40 is possible to visualize that success outcomes has been increased by numbers of flights. The site WAFB SLC 4E has been used in a minor way but the most of launches has been successfully. Site KSC LC 39A has been the last site to be used to made launches.

Payload vs. Launch Site



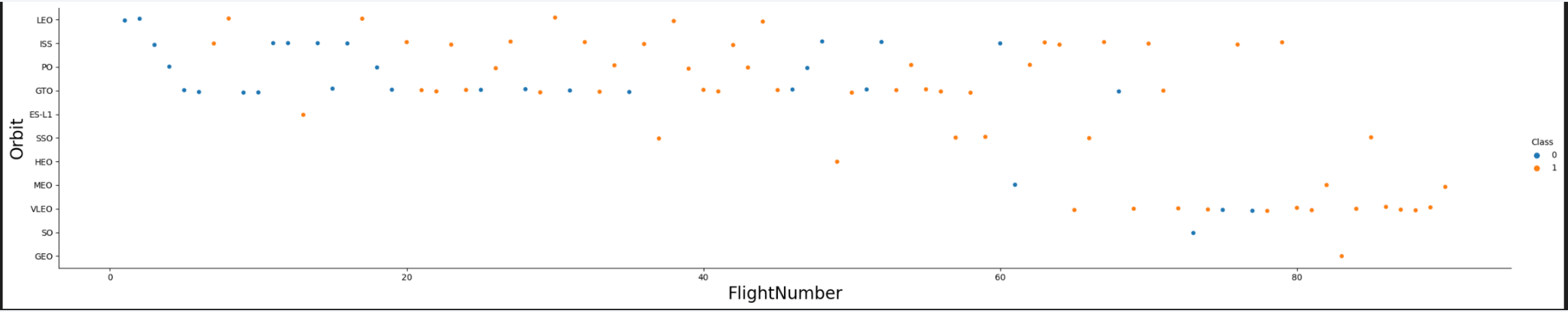
From the launch sites CCAFS SLC 40 and WAFB SLC 4E are been launched flights with the greater payload mass above from 10.000 Kg and this has been successfully.

Success Rate vs. Orbit Type



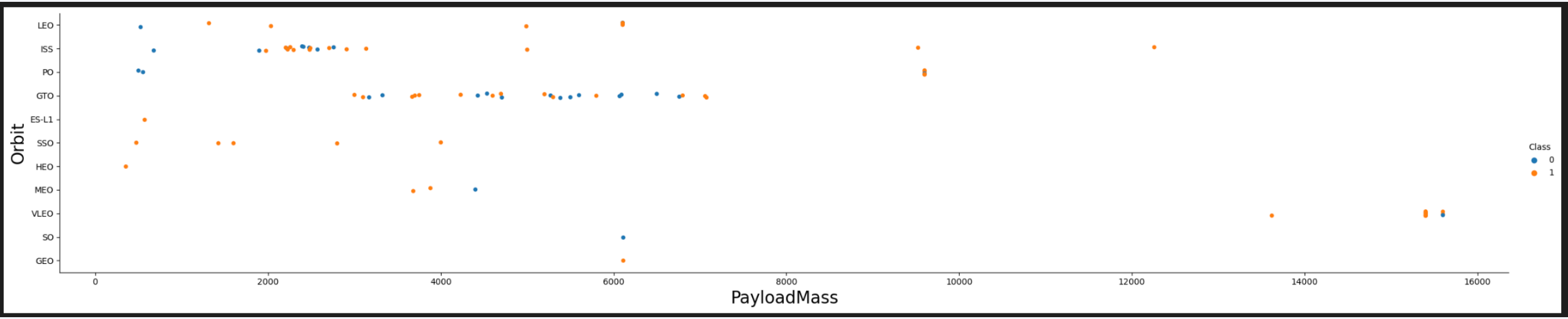
Launches to orbits ES-L1, GEO, SSO and HEO have the highest success rates, but from Flight Number vs. Orbit Type plot, is possible to visualize that only has been one launch to GEO orbit that was successful.

Flight Number vs. Orbit Type



Launches to GTO, PO, ISS and LEO orbits are the highest target destinations. Launches to other orbits has been few but with a high success outcome.

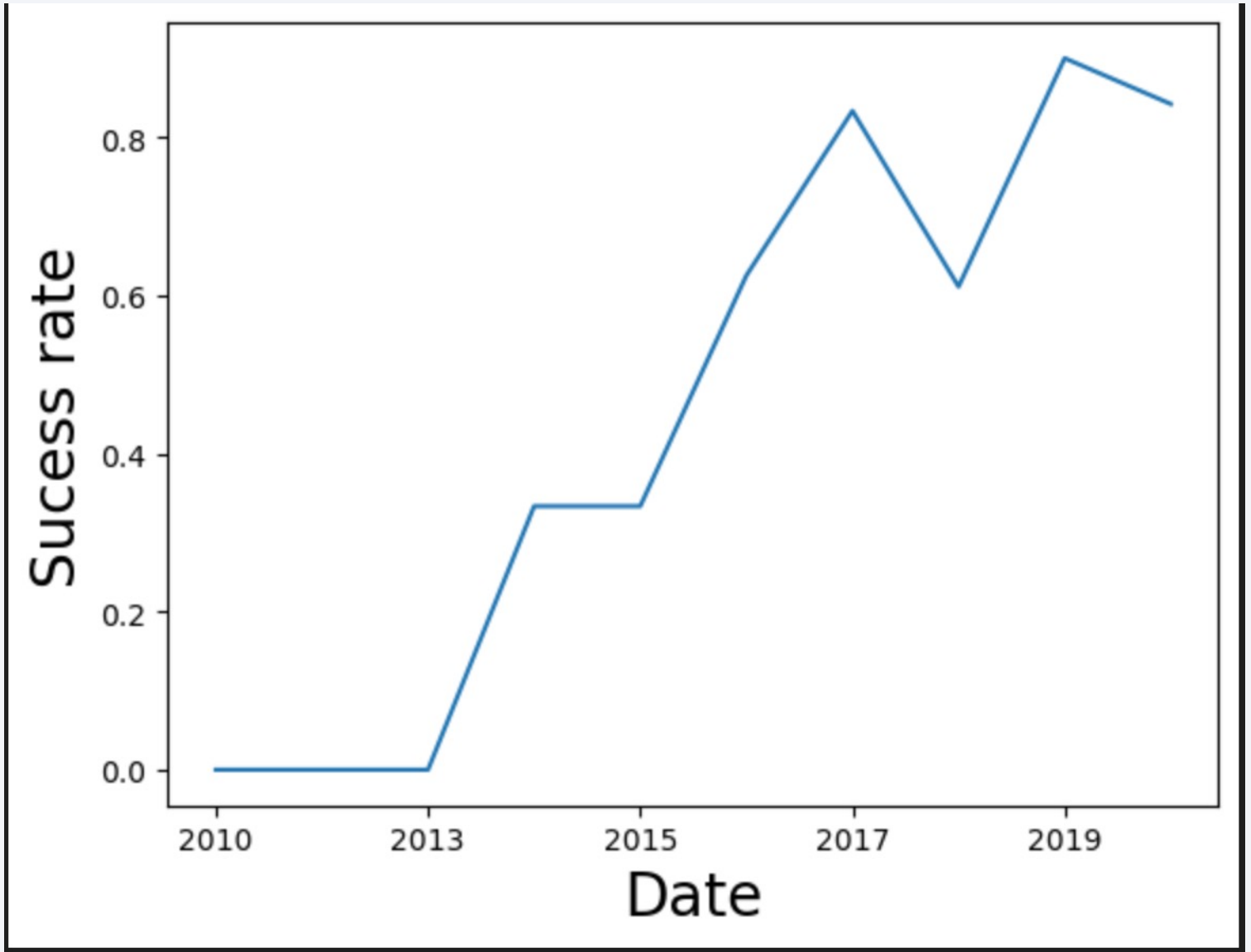
Payload vs. Orbit Type



Launches to GTO, and ISS orbits are the highest target destinations with payloads under 8000 Kg. VLEO orbit target has the highest payload value delivered.

Launch Success Yearly Trend

Only after 2013 launch success rate started to grow and this tendency has been maintained.



All Launch Site Names

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

On data set, there are register four places of launch.

Launch Site Names Begin with 'CCA'

On image are displayed five register with the string CCA in the launch site name.

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	
2010-08-12	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	
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	

Total Payload Mass

SUM(PAYLOAD_MASS__KG_)
45596

NASA customer has carried a total payload of 45,596 Kg on their launches.

Average Payload Mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1 is 2,928.4 Kg

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

MIN(Date)
2015-12-22

The first successful landing outcome on ground pad was in 2015/12/22

Successful Drone Ship Landing with Payload between 4000 and 6000

This is the list of names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 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

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

This is the total number of successful and failure mission outcomes.

Boosters Carried Maximum Payload

This is the list the names of the booster which have carried the maximum payload mass.

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

substr(Date,6,5)	Booster_Version	Launch_Site	Landing_Outcome
10-01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

This is the list of failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015.

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

Date	Landing_Outcome	CLO
2012-05-22	No attempt	10
2015-12-22	Success (ground pad)	5
2016-08-04	Success (drone ship)	5
2015-10-01	Failure (drone ship)	5
2014-04-18	Controlled (ocean)	3
2013-09-29	Uncontrolled (ocean)	2
2015-06-28	Precluded (drone ship)	1
2010-08-12	Failure (parachute)	1

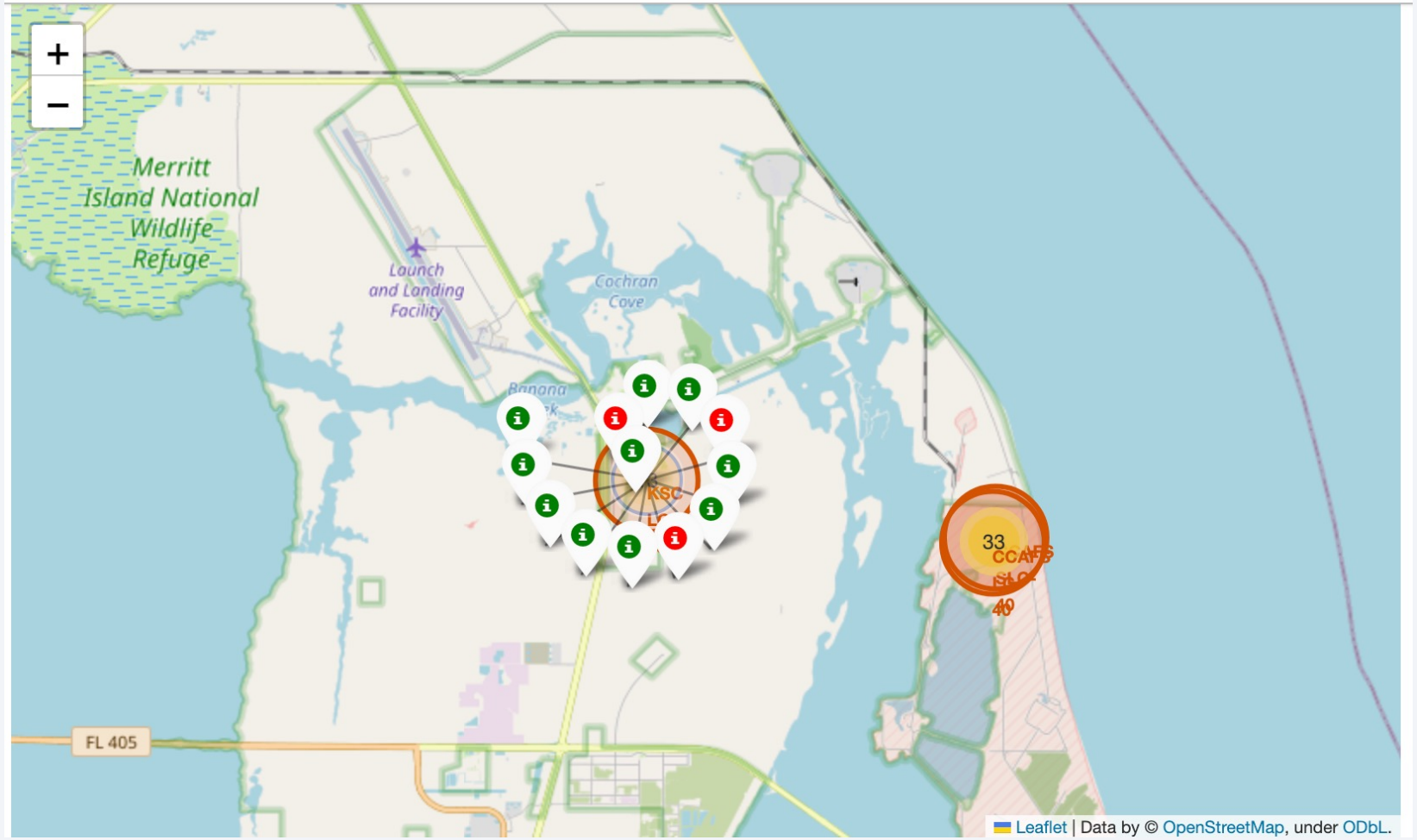
This is rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010/06/04 and 2017/03/20, in descending order.


Section 3
Launch Sites
Proximities Analysis

Launch site's locations

With red markers are denoted all launch site of Space X missions. All locations are based at south of E.E.U.U.

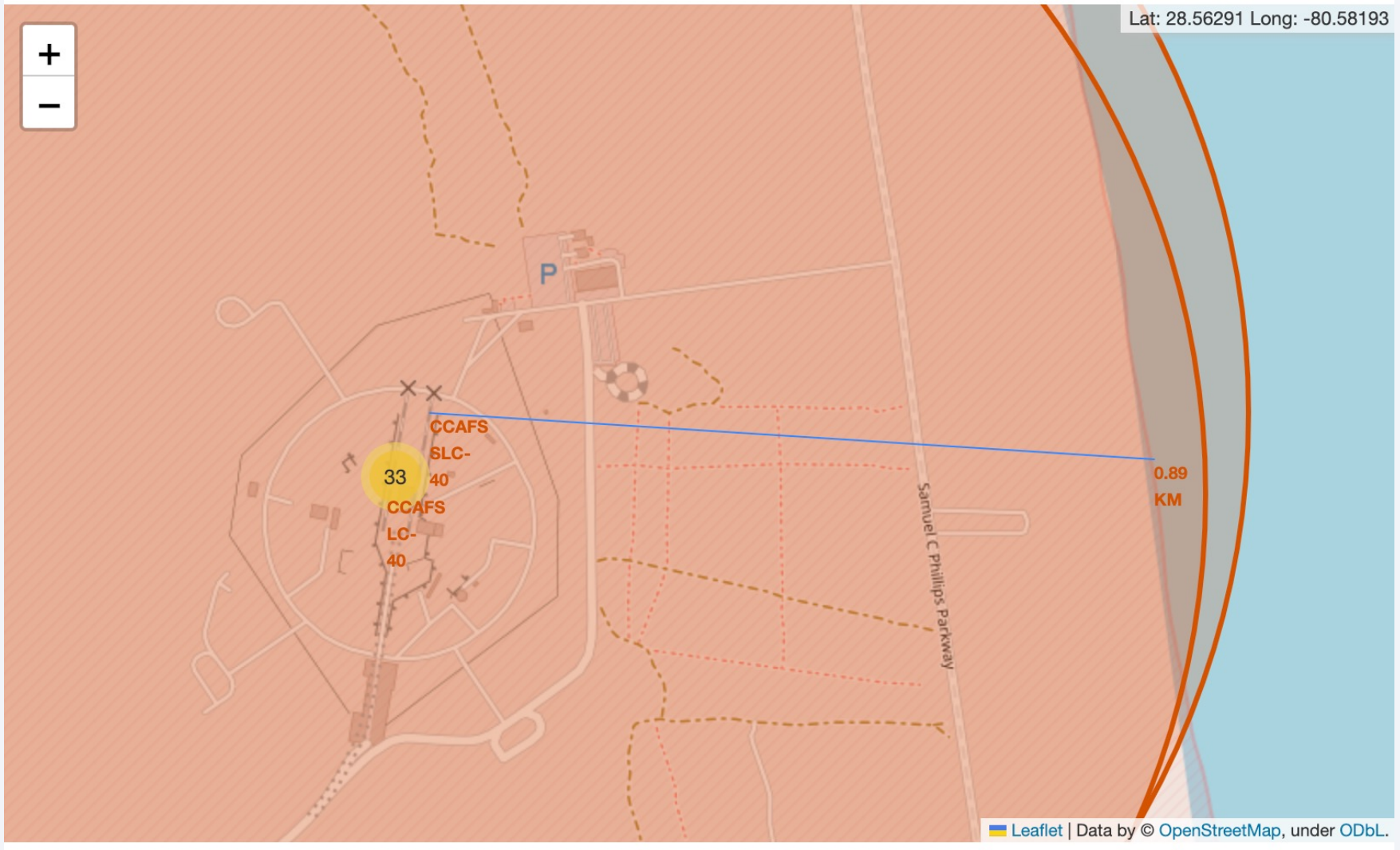
Success/failed launches for each site

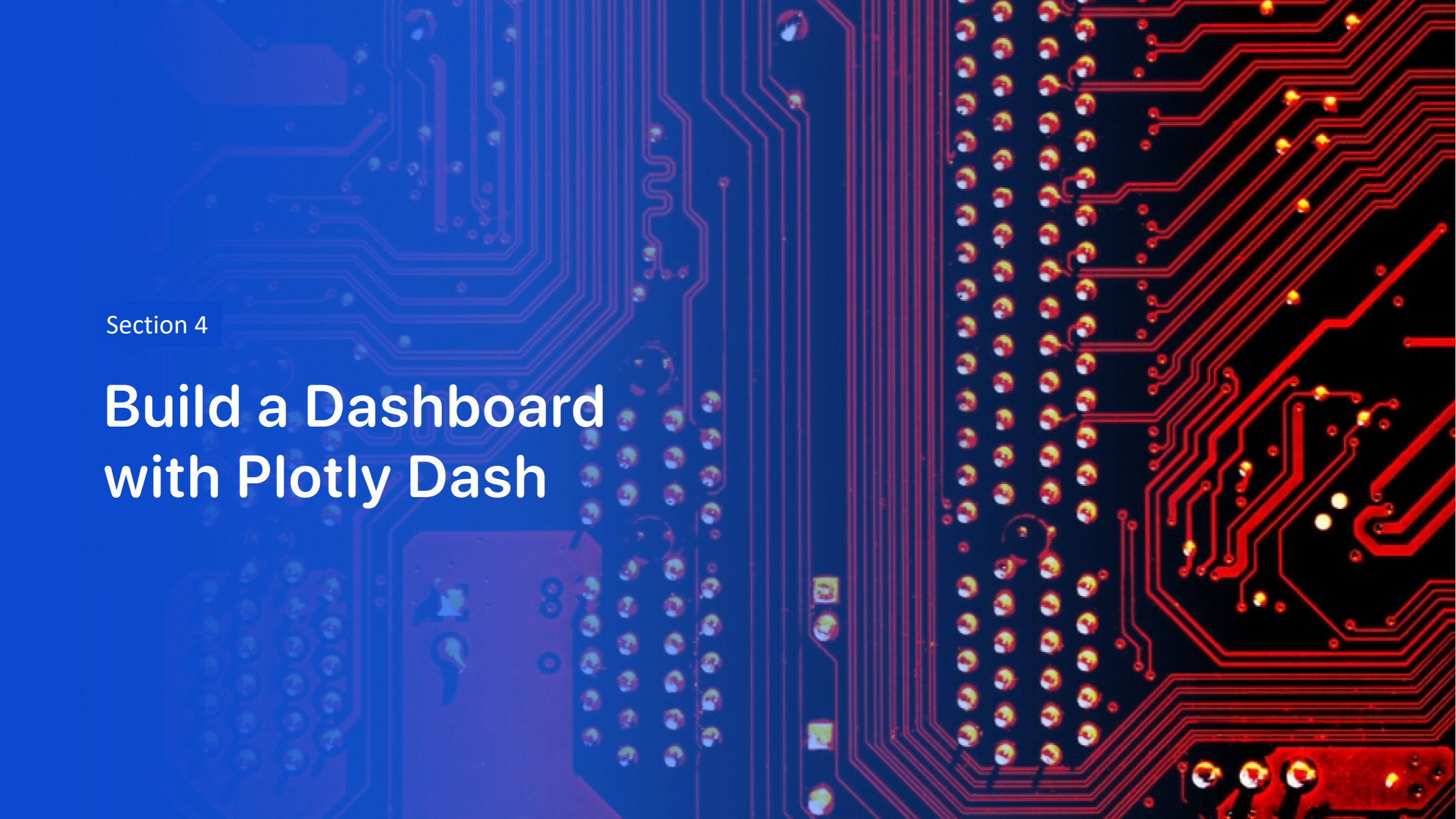


On the map are displayed color-labeled launch outcomes. The quantity of green labeled represent successful and the red ones are failed launches to the site KLC LC 39A.

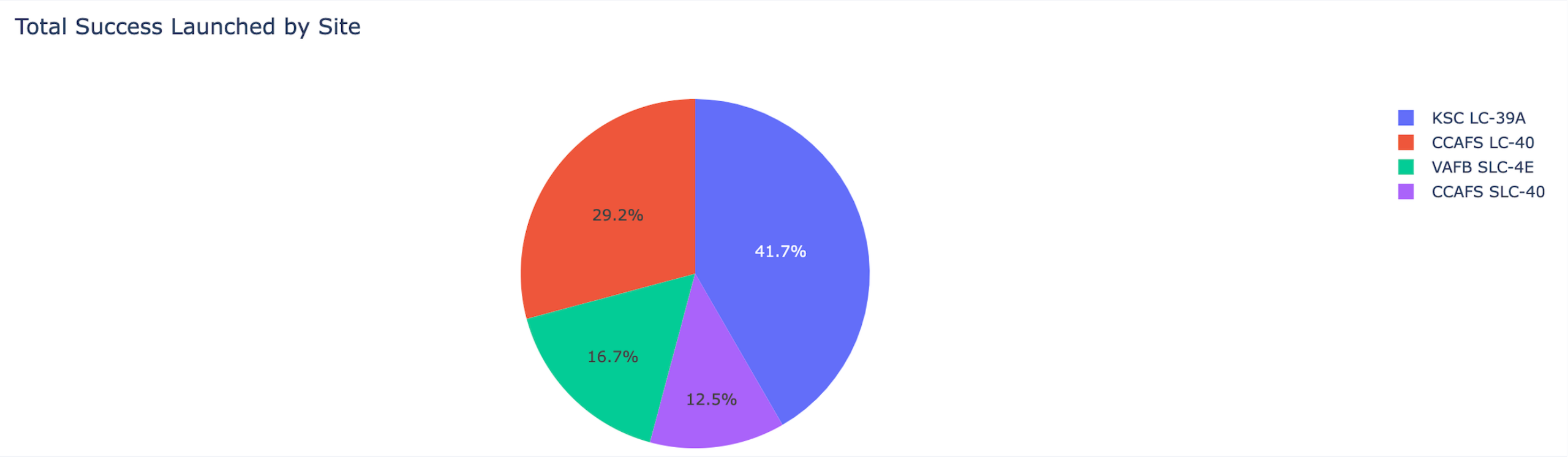
Distances between a launch site to its key proximities

On the map are displayed the distance of 0.98 Km from coastline to the launch sites CCAFS SLC 40 and CCAFS LC 40.



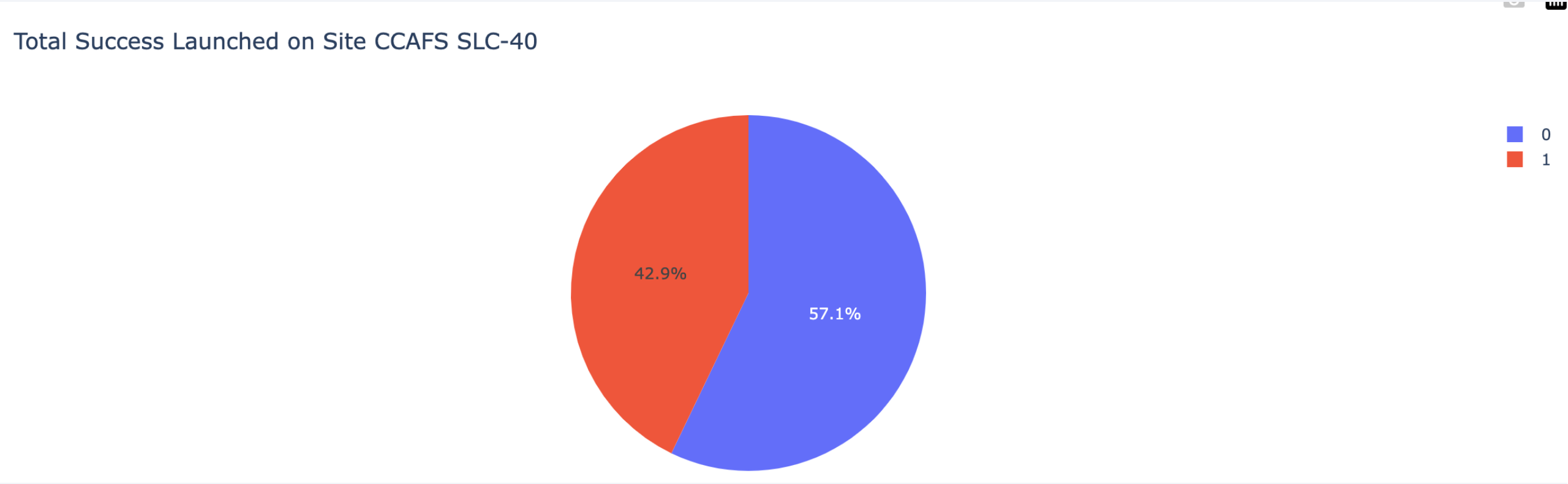

Section 4
Build a Dashboard with Plotly Dash

Launch success percentages by site



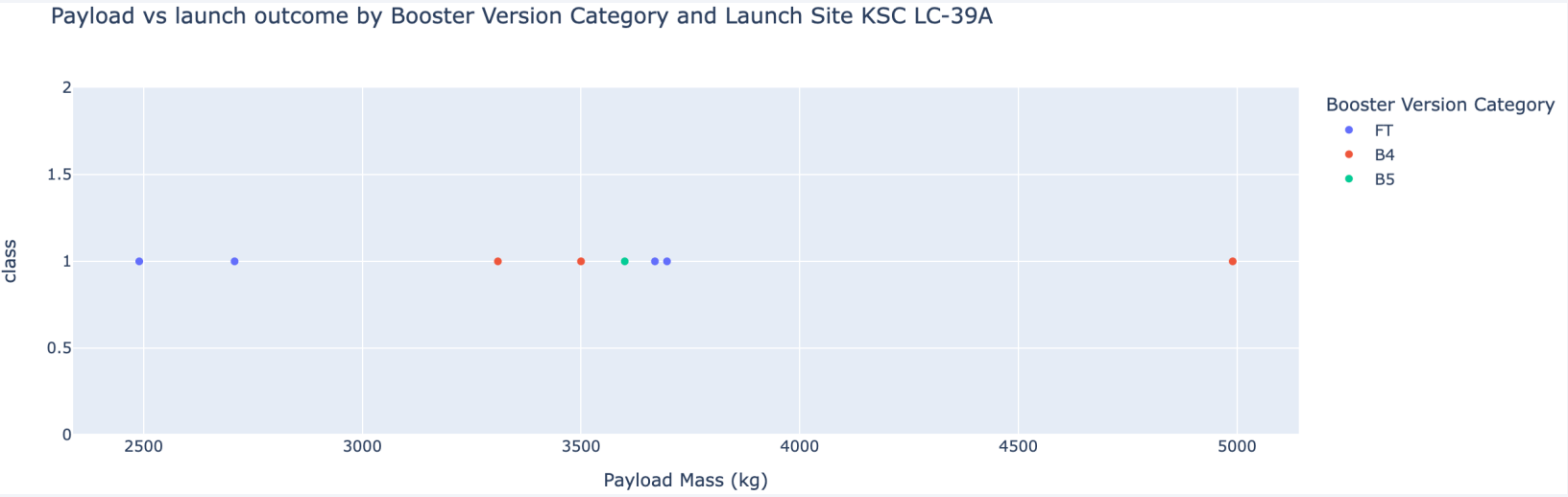
The launch site with the highest success percentage is KSC LC 39A with 41.7%.

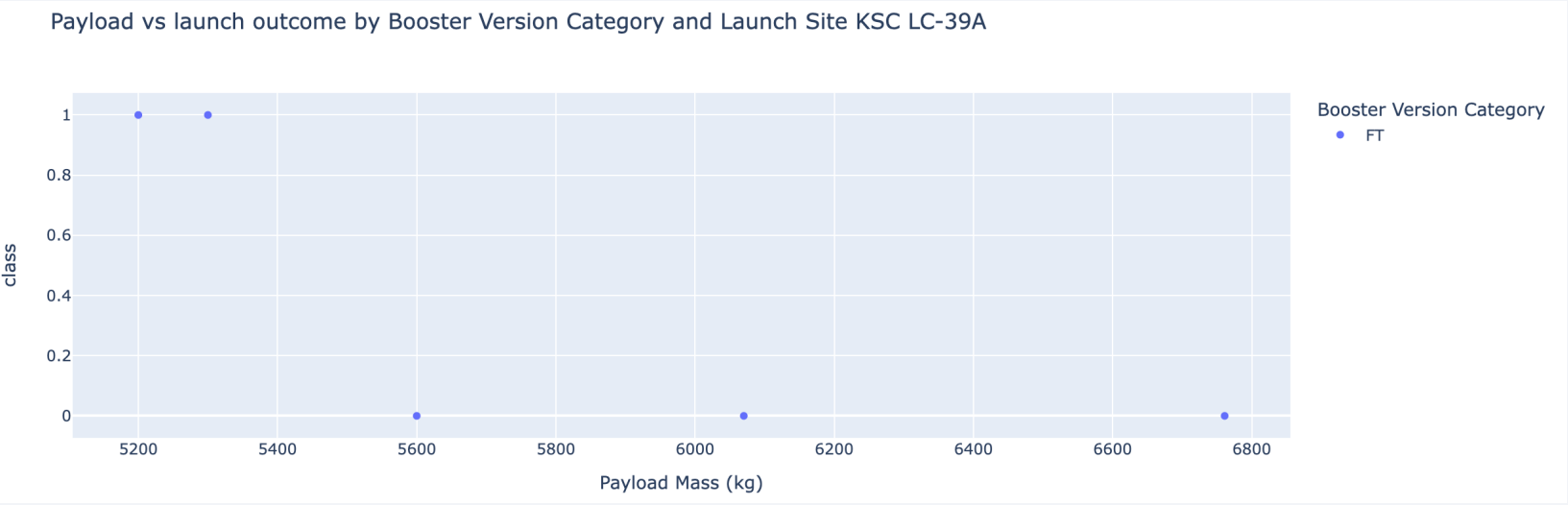
Launch site with highest launch success ratio



The launch site CCAFS SLC 40 has the highest success ratio is with 42.9%.

Launch outcome from KSC LC 39A per different payloads

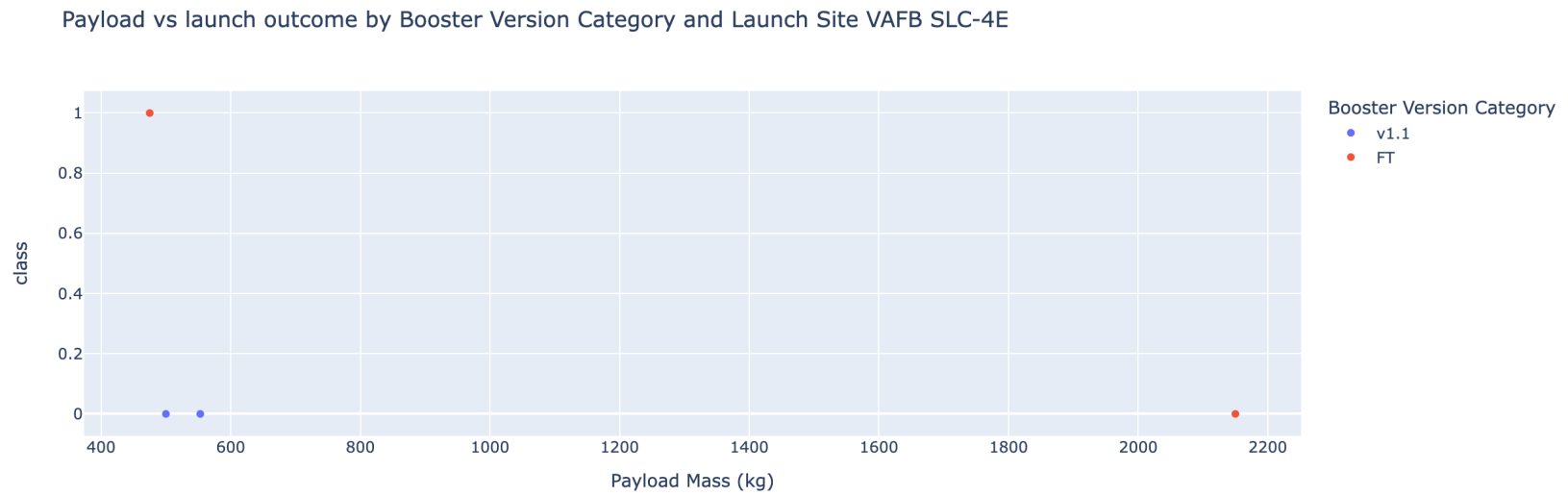


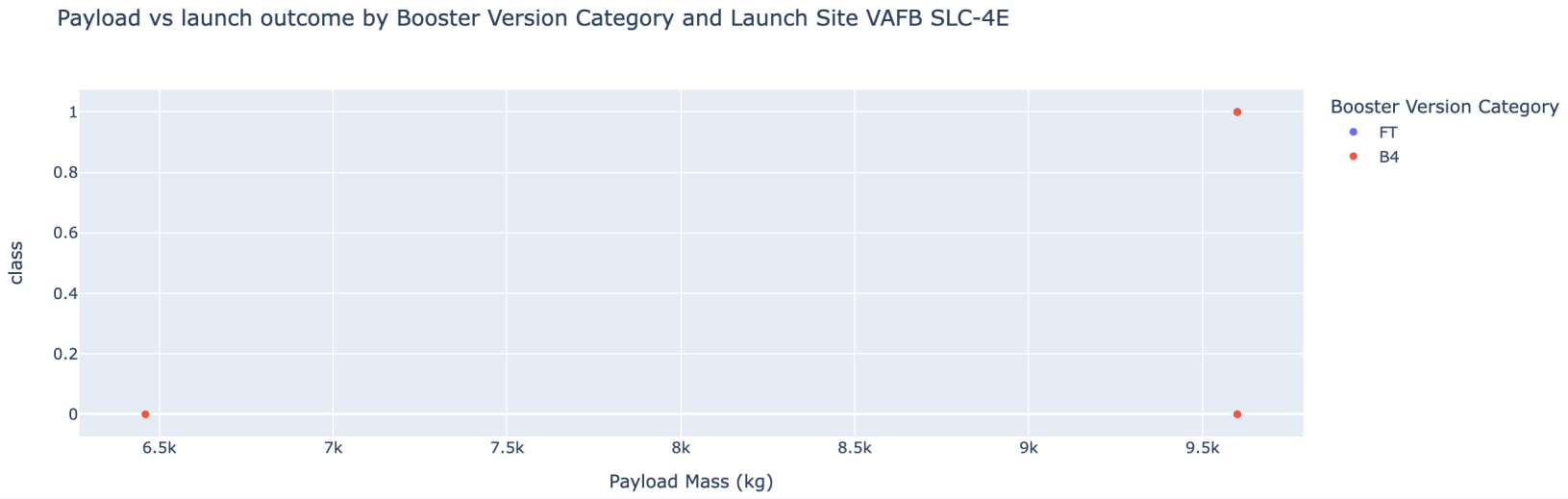


- Payload range for first figure from 0 to 5000 Kg.
- Payload range for second figure from 5000 to 10000 Kg.

Launch outcome from VAFB SLC 4E per different payloads

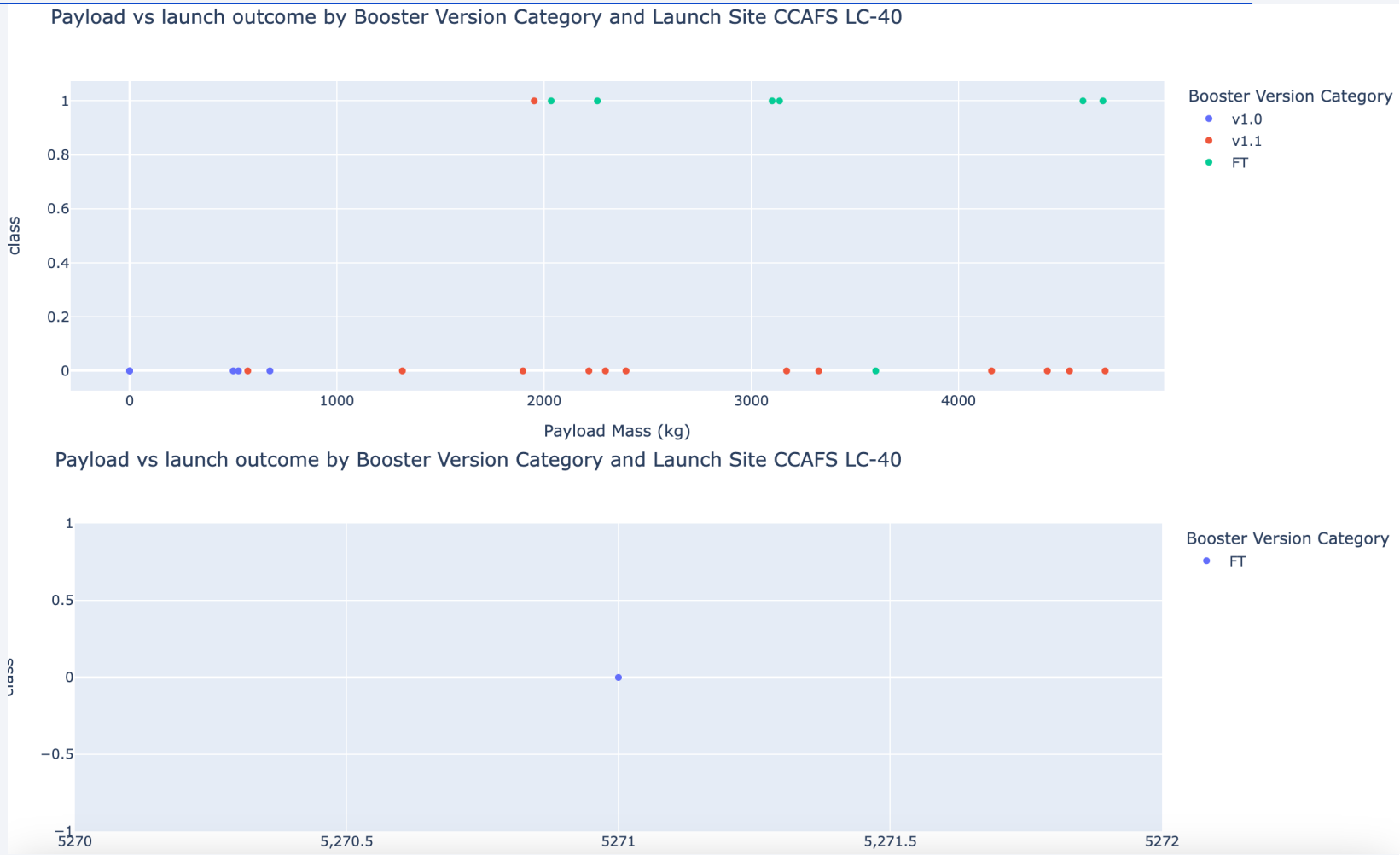
- Payload range for first figure from 0 to 5000 Kg.
- Payload range for second figure from 5000 to 10000 Kg.



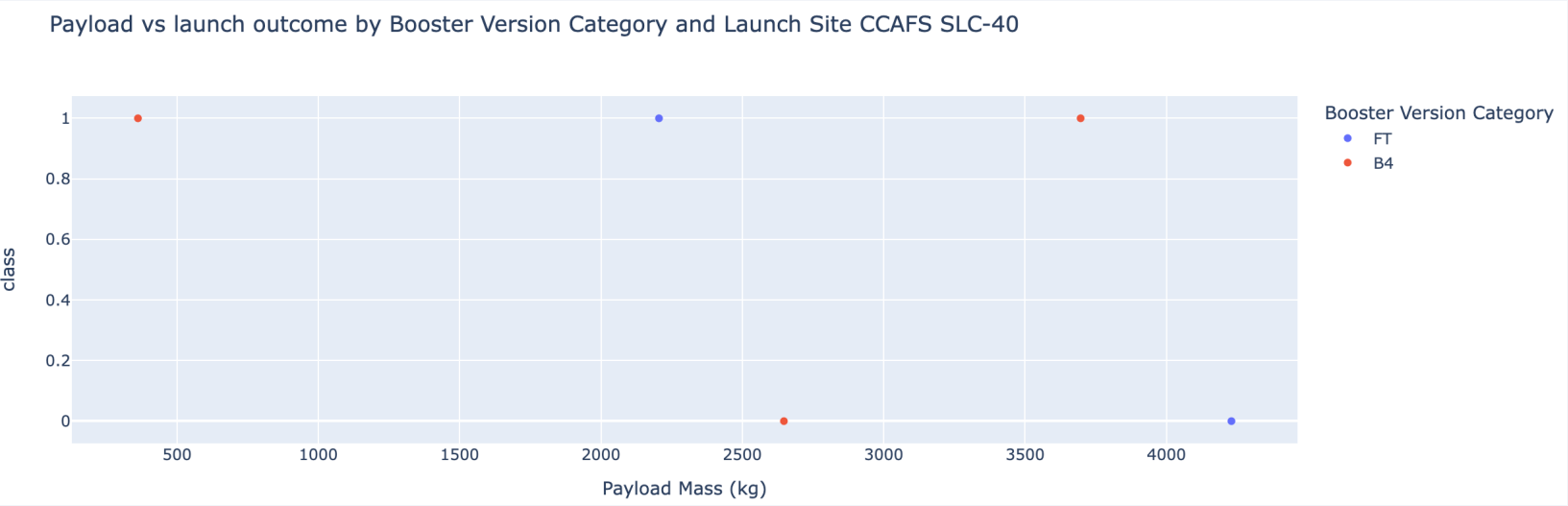


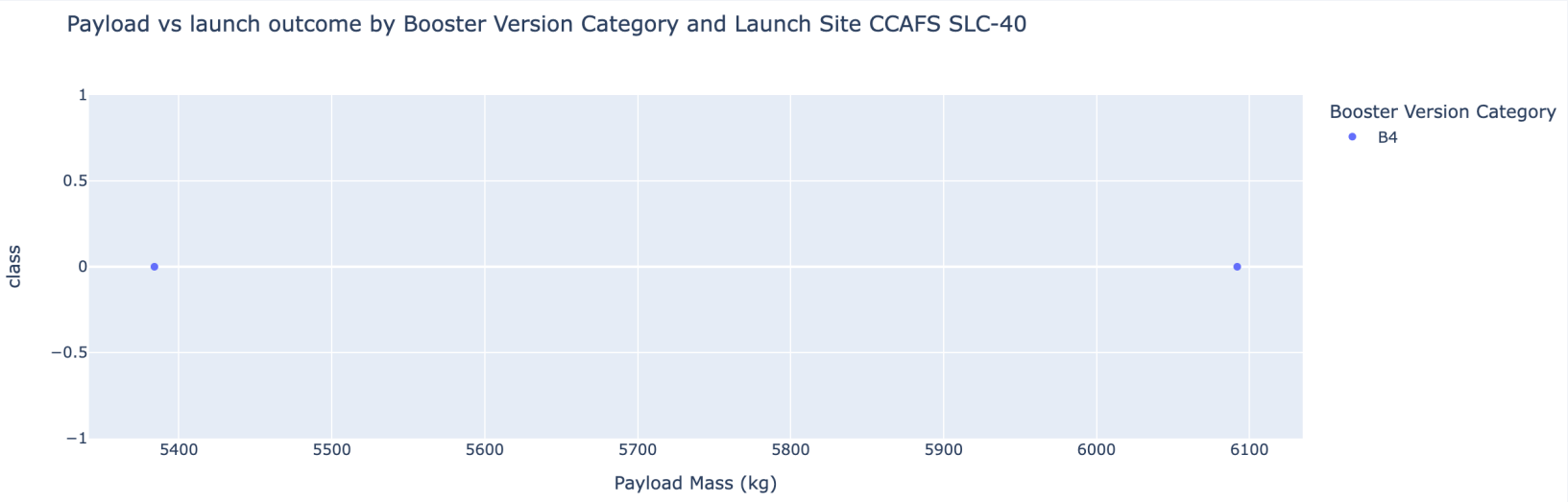
Launch outcome from CCAFS LC 40 per different payloads

- Payload range for first figure from 0 to 5000 Kg.
- Payload range for second figure from 5000 to 10000 Kg.



Launch outcome from CCAFS SLC 40 per different payloads



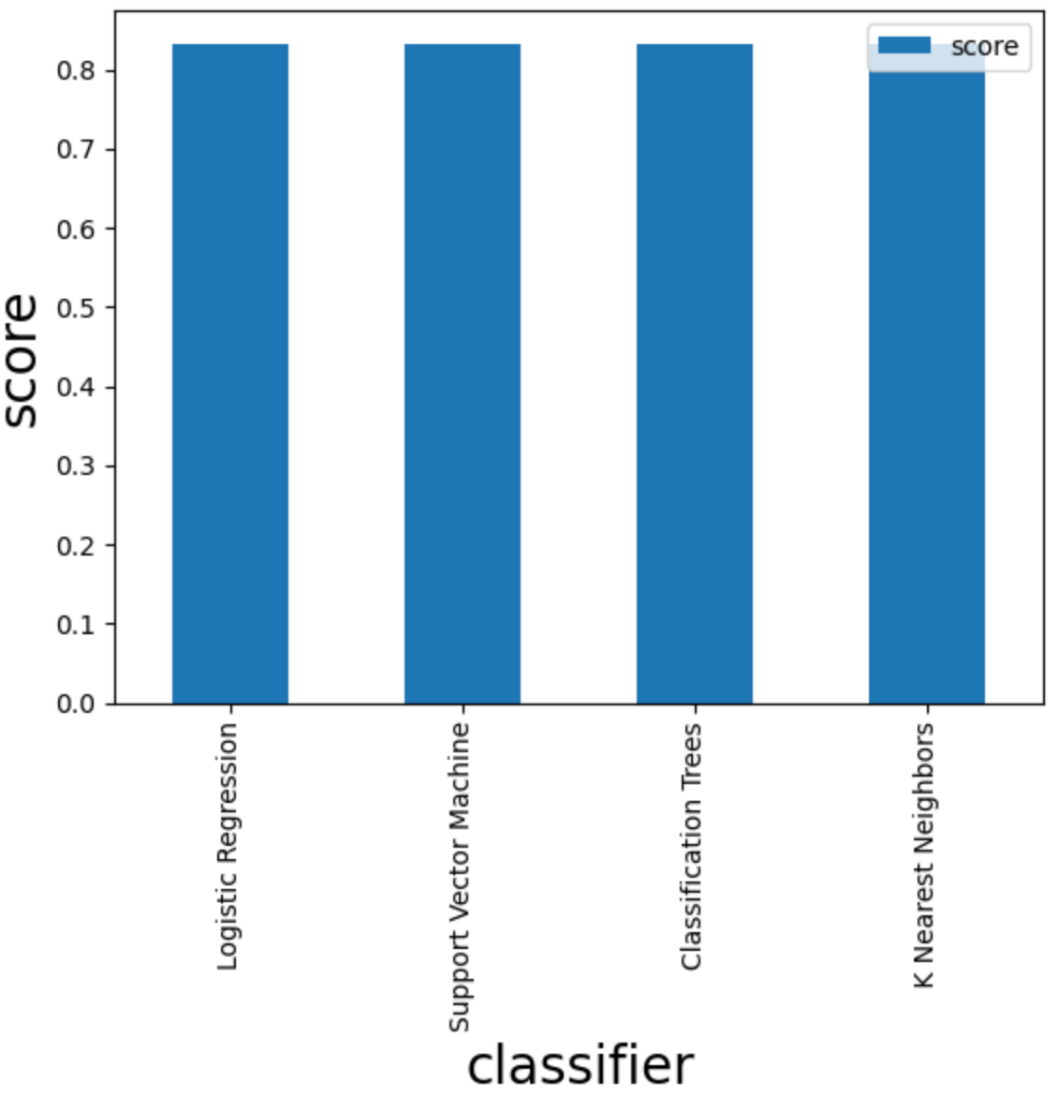


- Payload range for first figure from 0 to 5000 Kg.
- Payload range for second figure from 5000 to 10000 Kg.

Section 5

Predictive Analysis (Classification)

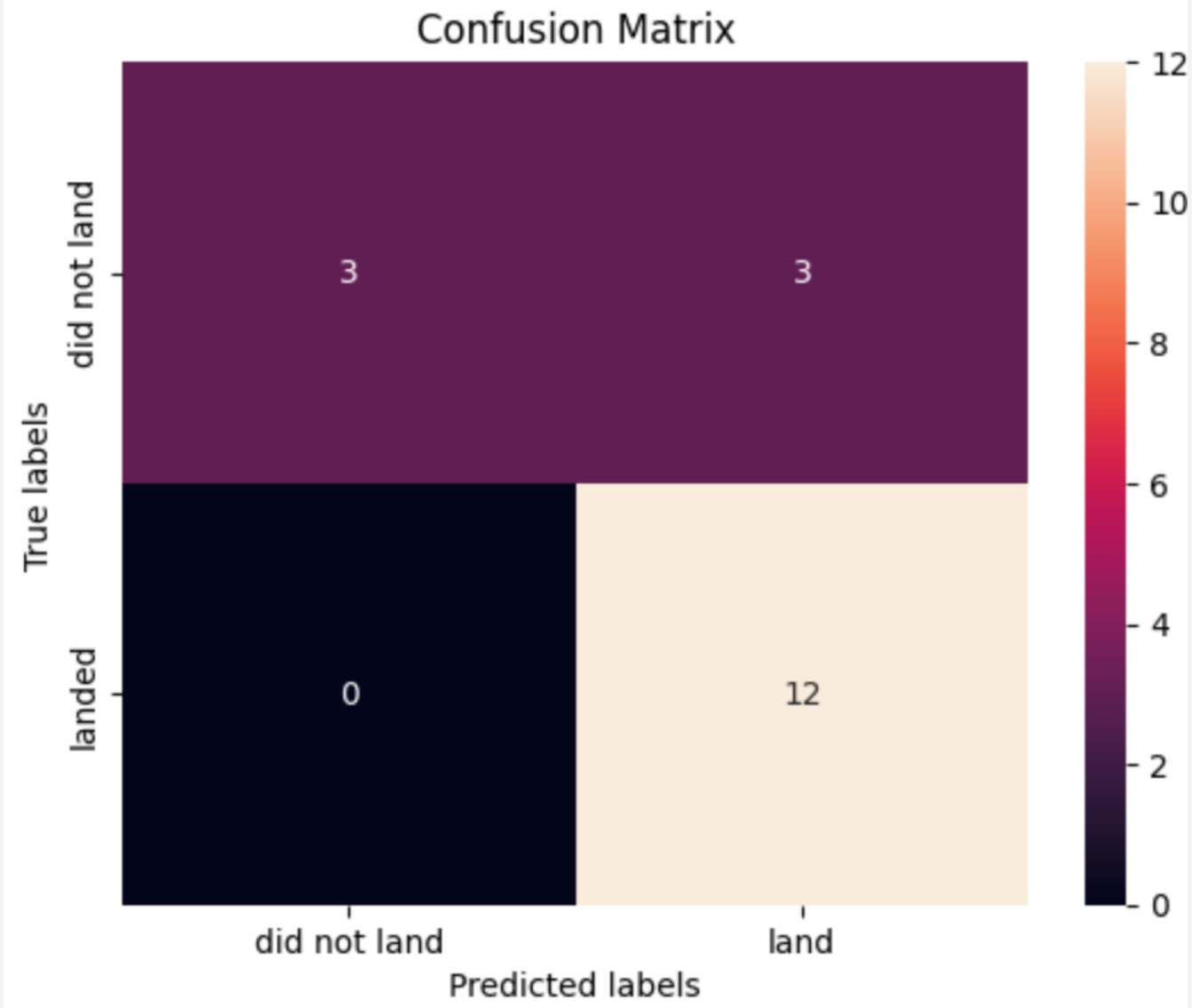
Classification Accuracy



All trained classification models have the same accuracy score to the given data set.

Confusion Matrix

All trained models have the same confusion matrix to the given data set.



Conclusions

- Any trained model can be used to predict a successful launch outcome with the same accuracy.
- EDA process using data visualization plots libraries (including interactive plots) and SQL queries allow to identify some insights and relationships about data that are out of the scope from the main objective of the analysis but can be useful to stakeholders.
- Geographical visualization of data allow to stakeholders give a location context to the insights and relationships founded in other stages.
- Data wrangling process must be oriented to prepare data to be used with the defined models to answer the main question of the analysis.

Appendix – SQL queries

Task 1: SELECT DISTINCT(Launch_Site) FROM SPACEXTABLE

Task 2: SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE "CCA%" LIMIT 5

Task 3: SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer = "NASA (CRS)"

Task 4: SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version = "F9 v1.1"

Task 5: SELECT MIN(Date) FROM SPACEXTABLE WHERE Landing_Outcome = "Success (ground pad)"

Task 6: SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = "Success (drone ship)" AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ <6000

Task 7: SELECT Mission_Outcome, COUNT(Mission_Outcome) FROM SPACEXTABLE GROUP BY Mission_Outcome

Task 8: SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ =(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)

Task 9: SELECT substr(Date,6,5), Booster_Version, Launch_Site, Landing_Outcome FROM SPACEXTABLE WHERE substr(Date,1,4)="2015" AND Landing_Outcome= "Failure (drone ship)"

Task 10: SELECT Date, Landing_Outcome, COUNT(Landing_Outcome) AS CLO FROM SPACEXTABLE WHERE DATE>"2010-06-04" AND DATE<"2017-03-20" GROUP BY Landing_Outcome ORDER BY CLO DESC

Appendix – Logistic regression model Grid Search parameters

```
parameters ={"C":[0.01,0.1,1],'penalty':['l2'], 'solver':['lbfgs']}# l1 lasso l2 ridge
lr=LogisticRegression()
logreg_cv = GridSearchCV(estimator = lr,
                         param_grid= parameters,
                         cv=10,# numero grupos de fold cross validations usados
                         n_jobs=-1
                        )
logreg_cv=logreg_cv.fit(X_train,Y_train)
```

Appendix – Support vector machine model Grid Search parameters

```
parameters = {'kernel':('linear', 'rbf','poly','rbf', 'sigmoid'),
              'C': np.logspace(-3, 3, 5),
              'gamma':np.logspace(-3, 3, 5)}
svm = SVC()
```

```
svm_cv = GridSearchCV(estimator = svm,
                      param_grid= parameters,
                      cv=10,# numero grupos de fold cross validations usados
                      n_jobs=-1
                     )
svm_cv=svm_cv.fit(X_train,Y_train)
```

Appendix – Tree regression model Grid Search parameters

```
parameters = {'criterion': ['gini', 'entropy'],
                'splitter': ['best', 'random'],
        'max_depth': [2, 4, 6, 8, 10, 12, 14, 16, 18],
        'max_features': ['sqrt'],
        'min_samples_leaf': [1, 2, 4],
            'min_samples_split': [2, 5, 10]
            }

tree = DecisionTreeClassifier()
```

```
tree_cv = GridSearchCV(estimator = tree,
                       param_grid= parameters,
                       cv=10,# numero grupos de fold cross validations usados
                       n_jobs=-1
                      )
tree_cv=tree_cv.fit(X_train,Y_train)
```

Appendix – K-Neighbors model Grid Search parameters

```
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()
```

```
knn_cv = GridSearchCV(estimator = KNN,
                      param_grid= parameters,
                      cv=10,# numero grupos de fold cross validations usados
                      n_jobs=-1
                     )
knn_cv=knn_cv.fit(X_train,Y_train)
```

Appendix – Notebooks external references

Data Collection – SpaceX API: https://github.com/dimonry/DATA_SCIENCE_CAPSTONE/blob/main/jupyter-labs-spacex-data-collection-api.ipynb
Data Collection – Wiki Scraping: https://github.com/dimonry/DATA_SCIENCE_CAPSTONE/blob/main/jupyter-labs-webscraping.ipynb
Data Wrangling: https://github.com/dimonry/DATA_SCIENCE_CAPSTONE/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_1_L3_labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb
EDA with Data Visualization: https://github.com/dimonry/DATA_SCIENCE_CAPSTONE/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb
EDA with SQL: https://github.com/dimonry/DATA_SCIENCE_CAPSTONE/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb
Interactive Map with Folium: https://github.com/dimonry/DATA_SCIENCE_CAPSTONE/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb
Dashboard with Plotly Dash: https://github.com/dimonry/DATA_SCIENCE_CAPSTONE/blob/main/spacex_dash_app.py
Predictive Analysis (Classification): https://github.com/dimonry/DATA_SCIENCE_CAPSTONE/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Thank you!