IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Marius Kuate
25.04.2024

Outline

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

Executive Summary

- **Summary of methodologies**
  - Data wrangling, exploratory analysis, feature selection Utilizing SpaceX REST API data
  - For Classification: Trainning and evaluation of KNN, SVM, Decision Tree, Logistic Regression
- **Summary of all results**
  - The KNN model achieved the highest accuracy 0.94. follow by SVM with an accuracy of 0.83

Introduction

- **SpaceX's transformative impact**
  - Affordable SpaceX Travel through innovative reuse of rocket components
  - Space Y's aspiration to compete with SpaceX
- **Optimizing launch cost competitiveness**
  - Tasked with determining launch pricing and predicting rocket reuse
  - Understanding factors influencing SpaceX's reuse decisions
  - Forecasting first stage reusability with machine learning

Section 1

Methodology

Methodology

Executive Summary

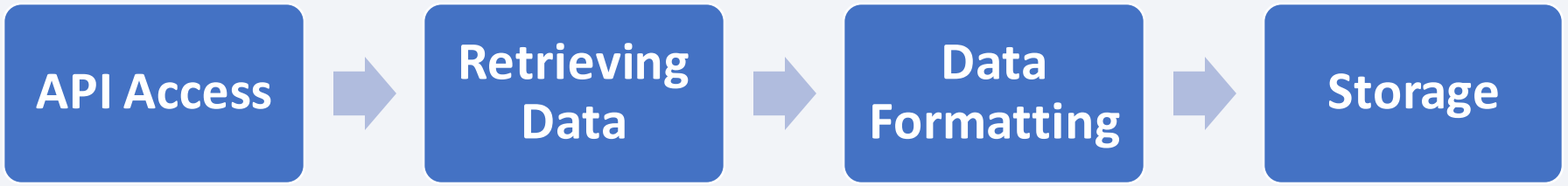
- **Data collection methodology**
  - Utilized SpaceX REST API (api.spacexdata.com/v4/) endpoints to retrieve relevant launch data and extract with HTPP requests
  - Captured data in structured format (.csv) for further analysis
- **Perform data wrangling**
  - Addressed missing or inconsistent values, ensuring data integrity
  - Create label column based on the outcome of the land
- **Perform exploratory data analysis (EDA) using visualization and SQL**

Methodology

Executive Summary

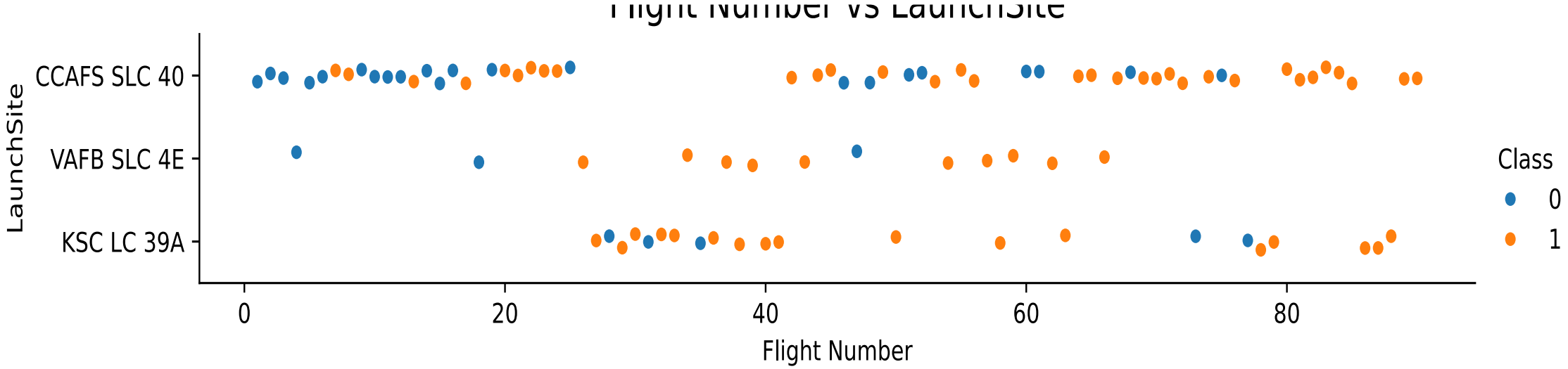
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Model Selection: Logistic Regression, KNN, SVM, TreeClassifer
  - Split data in training and test dataset
  - Fit each model to the training data and utlize grid search to find optimal hyperparameters for each model, improving performance
  - Evaluate by using accuracy and confusion matrix

Data Collection

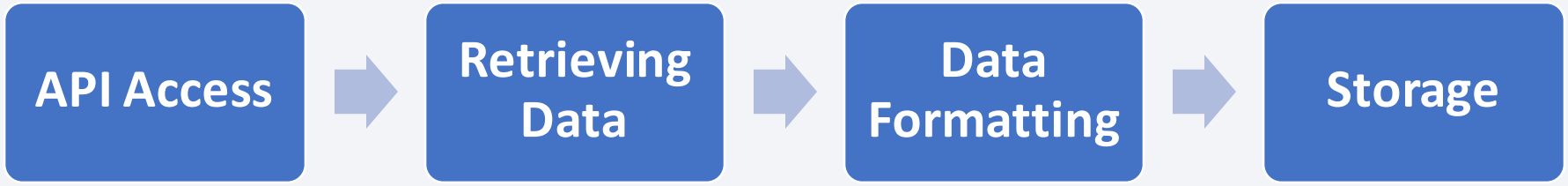


Flight Number vs. Launch Site

- A large part of the rockets on site CCAFS with a number flights inferior to 20 did not have a correct landing.
- Rockets launched on site KSC LC 39A have min 25 continuous launch attemps



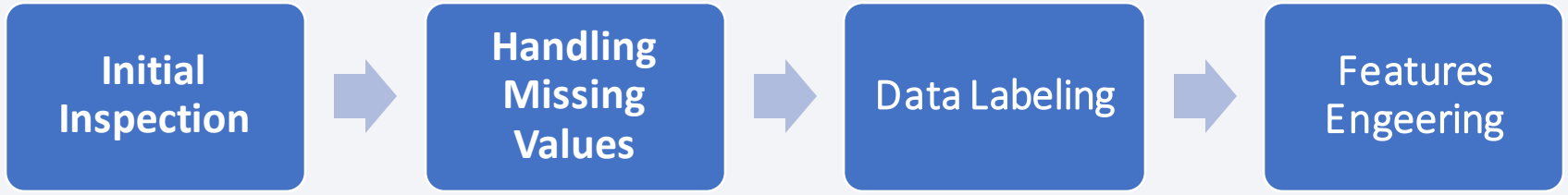
Data Collection – SpaceX API



GitHub URL of the completed SpaceX API calls:

https://github.com/kuatemarius/hello_data_science/blob/main/Data%20Science/projet_spaceY/jupyter-labs-spacex-Data%20Collection-api.ipynb

Data Wrangling and Preparation



GitHub URL for data wrangling:

https://github.com/kuatemarius/hello_data_science/blob/main/Data%20Science/projet_spaceY/labs-jupyter-spacex-Data%20Wrangling.ipynb

GitHub URL For Features Engeering : https://github.com/kuatemarius/hello_data_science/blob/main/Data%20Science/projet_spaceY/jupyter-labs-eda-dataviz.ipynb

EDA with Data Visualization

- Relationship between the the continuous launch attempts and the payload of rockets
- Relationship between the the number of launch attempts and different launch sites
- Relationsip between the payload and the launch site of rockets

GitHub URL: https://github.com/kuatemarius/hello_data_science/blob/main/Data%20Science/projet_spaceY/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- **Data Retrieval**
  - `%sql select distinct Launch_Site from SPACEXTABLE`
- **Data Filtering**
  - `%sql select Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ <6000`
  - `%sql select * from SPACEXTABLE where Launch_Site like "CCA%" LIMIT 5`
  - `%sql select * from SPACEXTABLE where Launch_Site like "CCA%" LIMIT 5`
- **Data Aggregation**
  - `%sql select SUM(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer = "NASA (CRS)"`
  - `%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version = "F9 v1.1"`
  - `%sql select min(Date) from SPACEXTABLE where Mission_Outcome = "Success"`

EDA with SQL

- **Grouping**
  - %sql SELECT COUNT(*), Mission_Outcome FROM SPACEXTABLE GROUP BY Mission_Outcome;
  - %sql SELECT Landing_Outcome, COUNT(Landing_Outcome) AS Outcome_Count FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY Outcome_Count DESC;

- GitHub URL: https://github.com/kuatemarius/hello_data_science/blob/main/Data%20Science/projet_spaceY/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build a Dashboard with Plotly Dash

- Total success of Lauches on all sites and on each different site: To see which launch site the the best success rate
- Correlation between Payload and success for all sites: To see the impact of rocket's payload on the outcome of the launch.

GitHub URL of completed Plotly Dash:

https://github.com/kuatemarius/hello_data_science/blob/main/Data%20Science/projet_spaceY/spacex_dash_app.py

Predictive Analysis (Classification)

GitHub URL: https://github.com/kuatemarius/hello_data_science/blob/main/Data%20Science/projet_spaceY/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

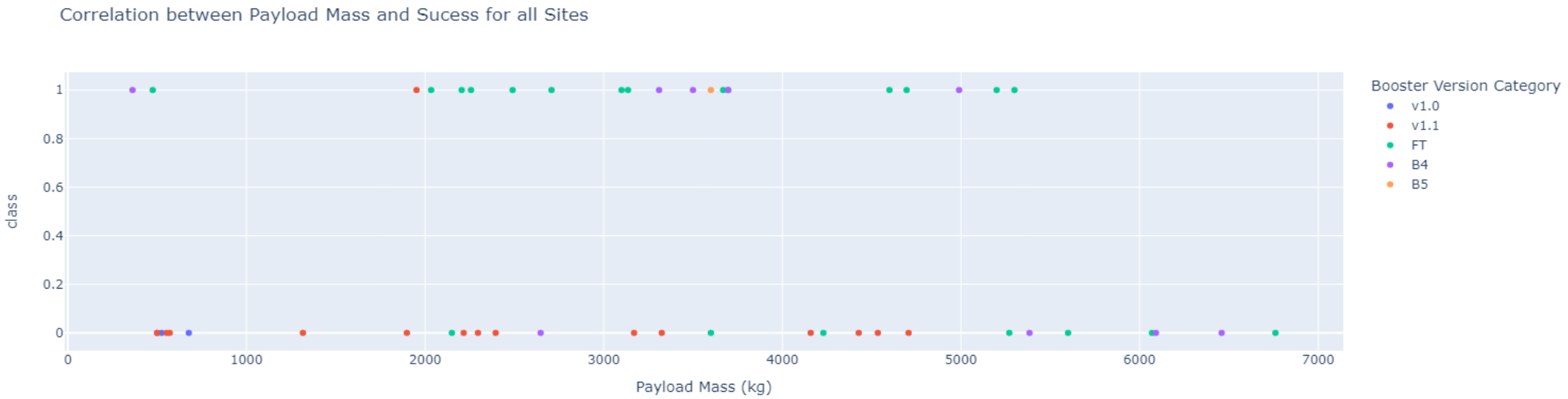
- Predictive analysis results
- Accaray:
- KNN: 0.94
- SVM: 0.83
- Logistic Regression: 0.80
- TreeClassifer: 0.722

- Exploratory data analysis results
- Interactive analytics demo in screenshots

Total sucess Launches on all sites

Results

From the dashboard, on the figure "Total success Launches on all sites "we can note that the launch site KSC LC 39-A hast the best success rate compared than other site.

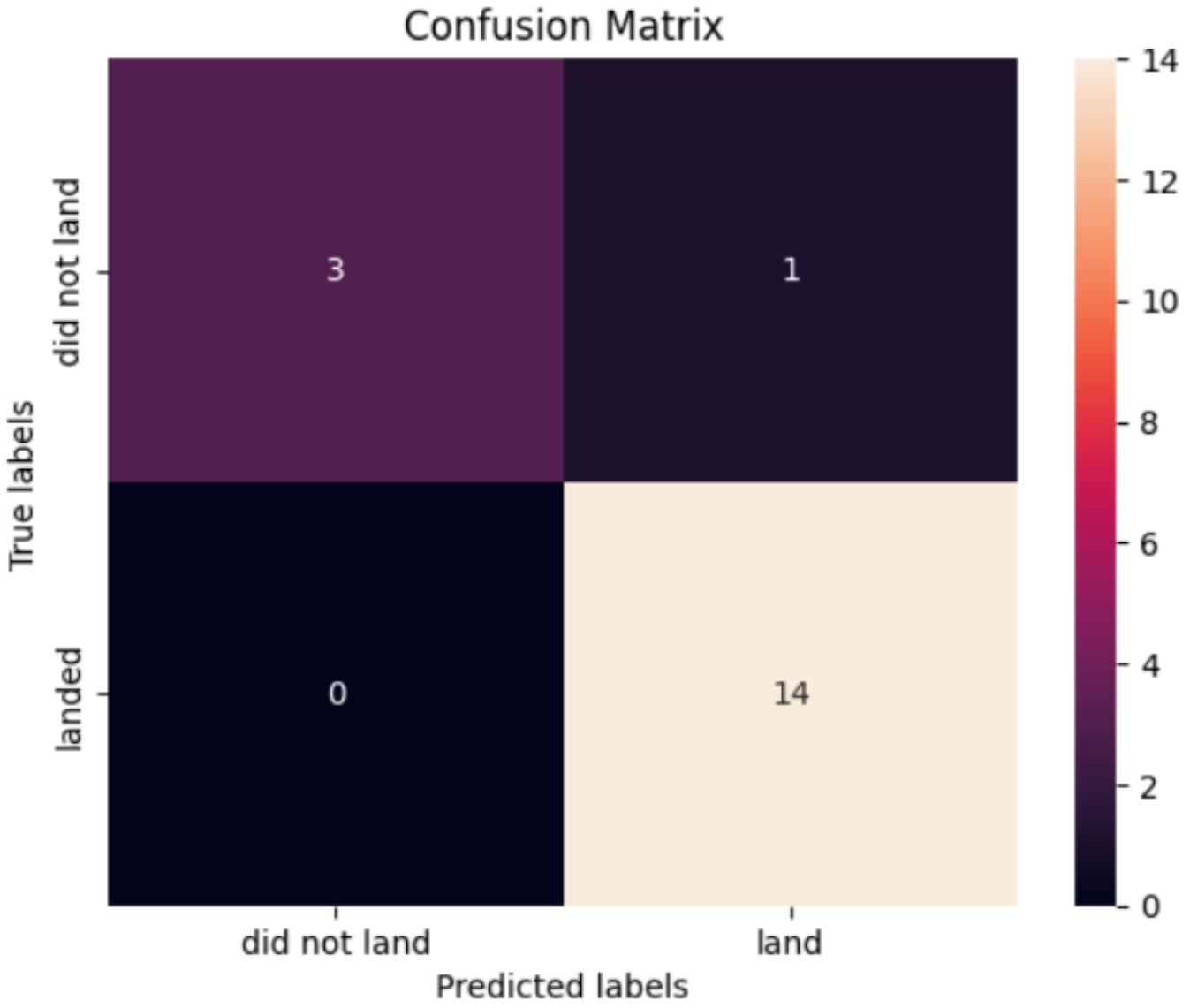


Results

On the Figure " Correlations between payload and success" we can note the rockets with less payload may have a better success rate than rockets with higher payload.

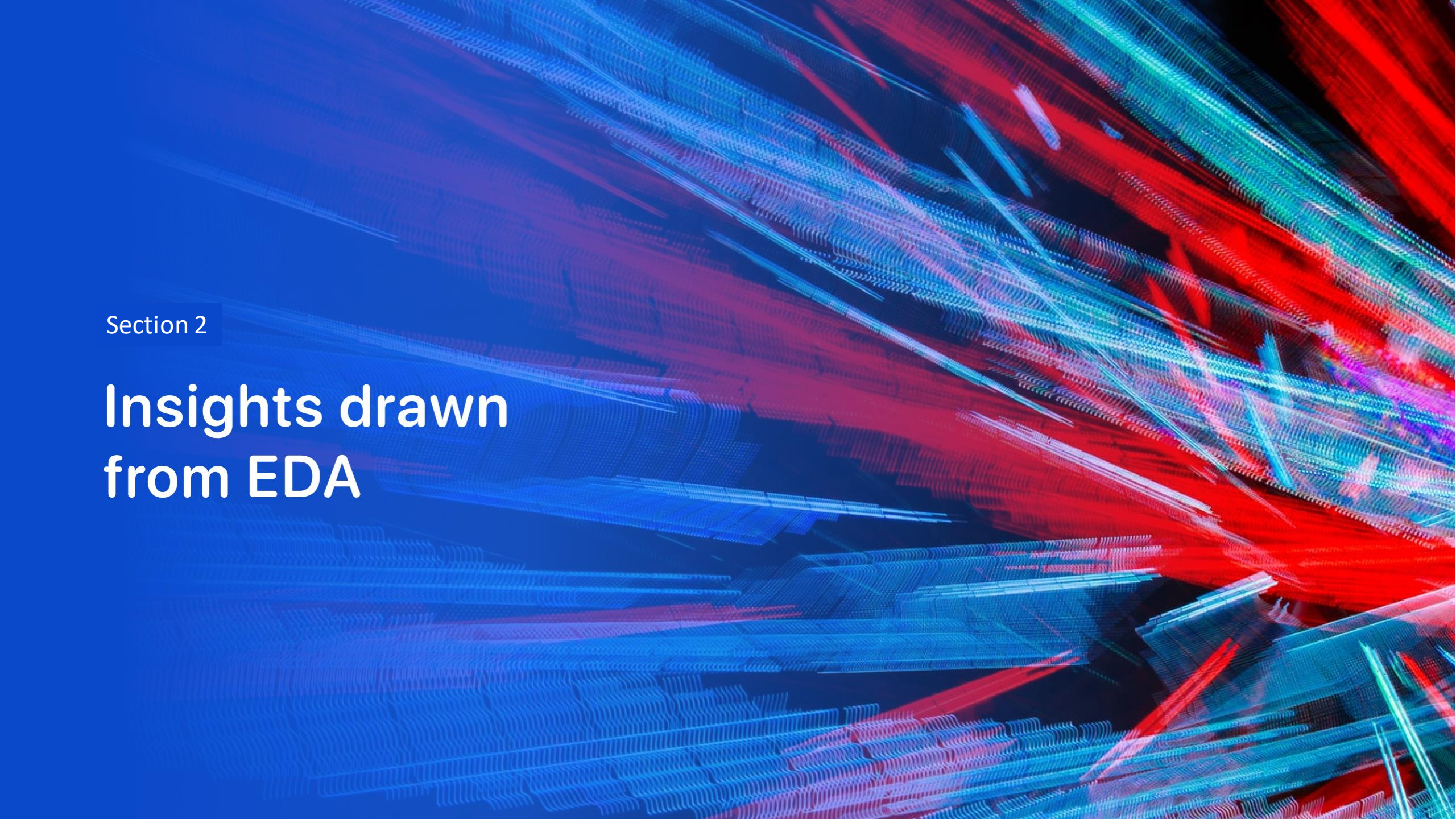
Results

- Predictive analysis (Accuracy) results
  - KNN: 0.94
  - SVM: 0.83
  - Logistic Regression: 0.80
  - TreeClassifer: 0.722



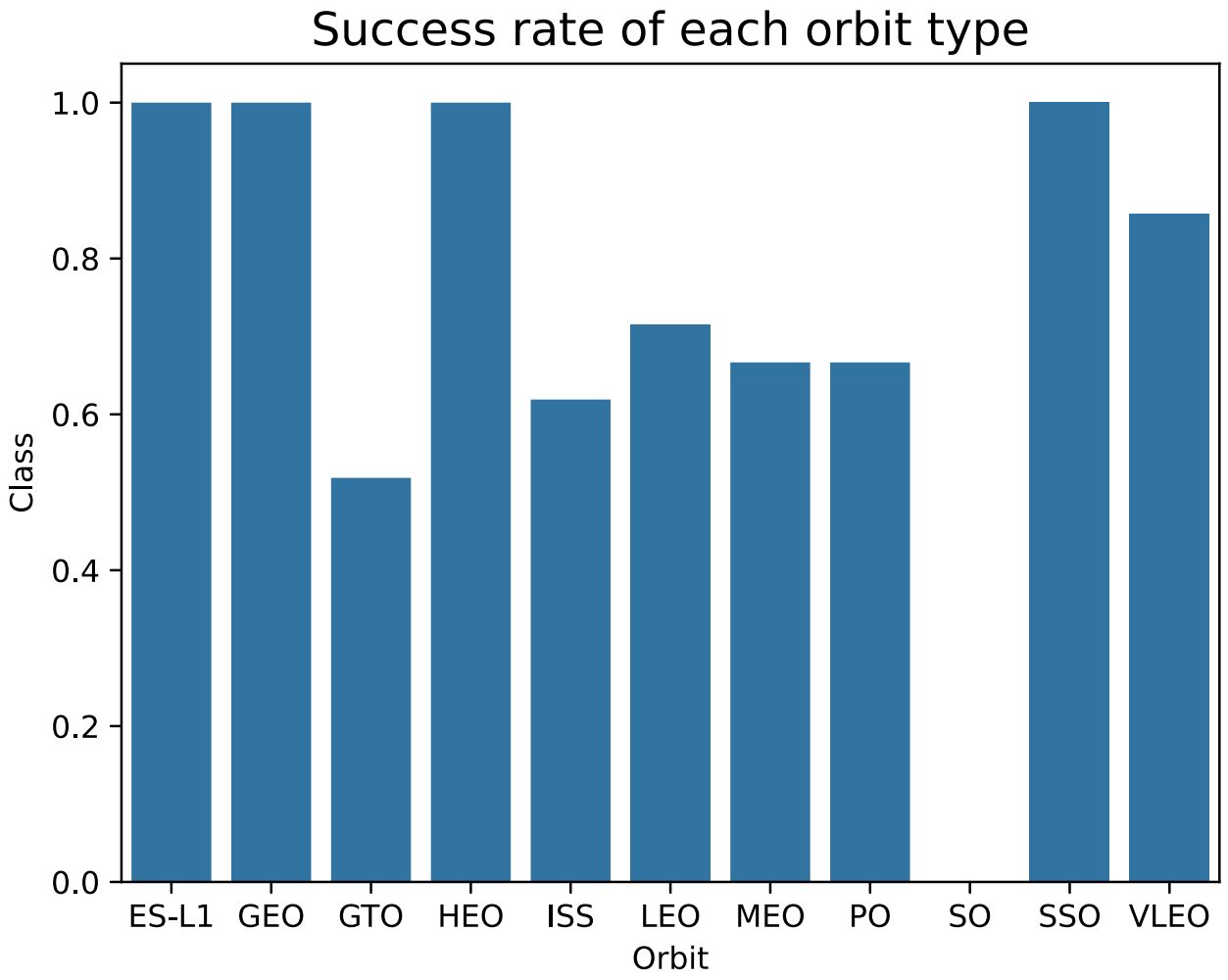
Section 2

Insights drawn from EDA



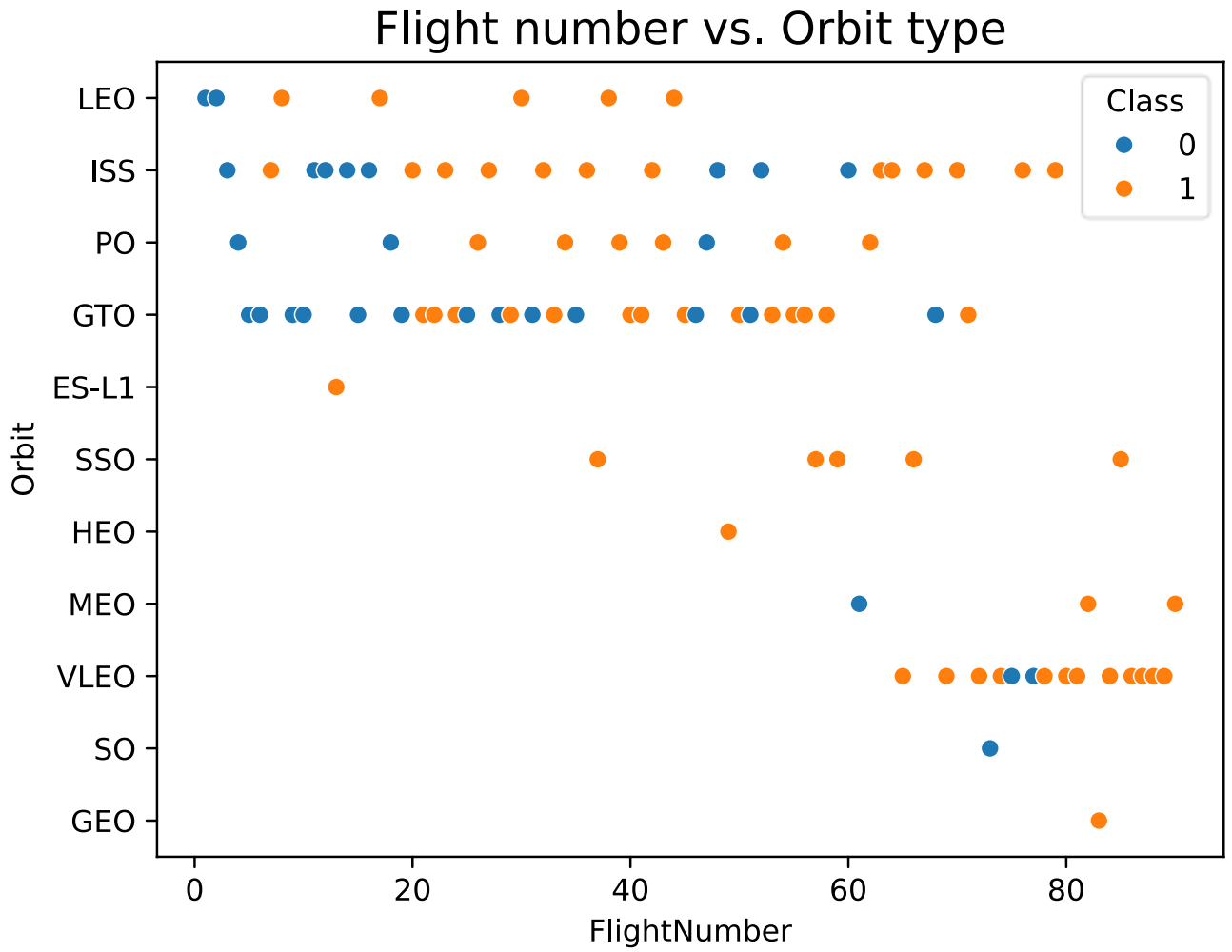
Success Rate vs. Orbit Type

- Orbit as ES-L1, GEO, HEO and SSO hast the best launch success rate.
- Launch on orbit SO always fail



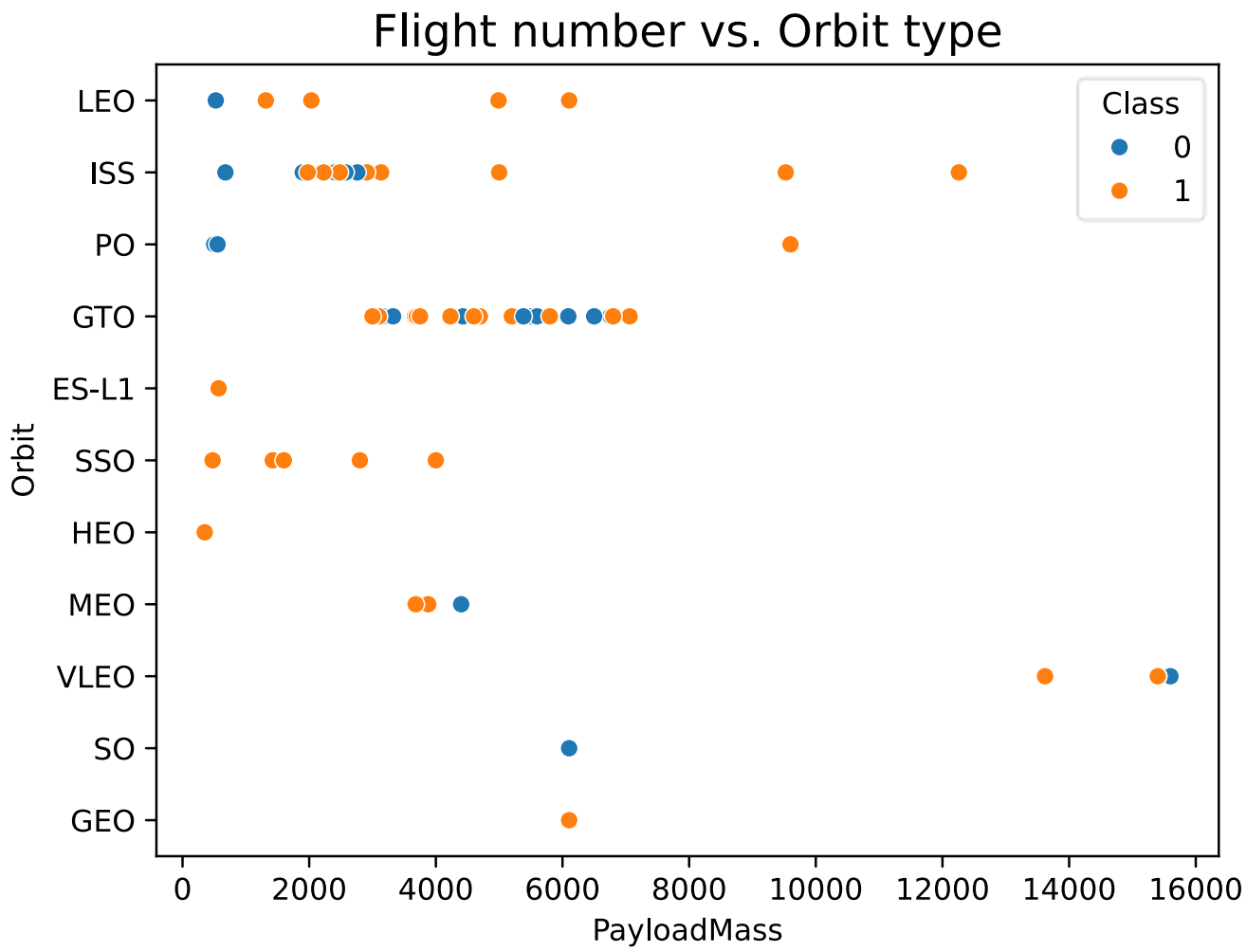
Flight Number vs. Orbit Type

- By LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.



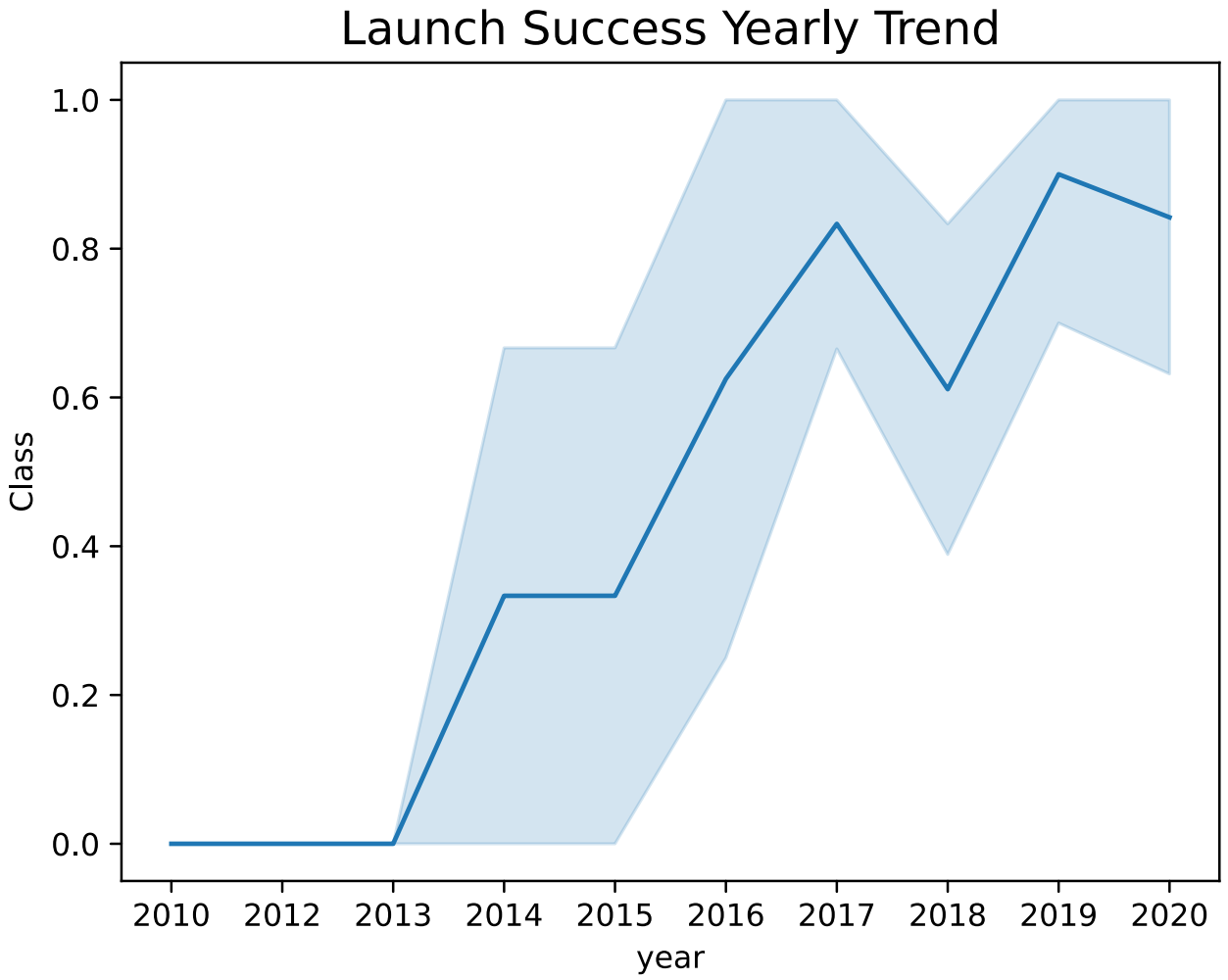
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.



Launch Success Yearly Trend

- We can observe that the success rate since 2013 kept increasing till 2017 (stable in 2014) and after 2015 it started increasing.



All Launch Site Names

- %sql select distinct Launch_Site from SPACEXTABLE
- selects unique values from the "Launch_Site" column in the "SPACEXTABLE" table.

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

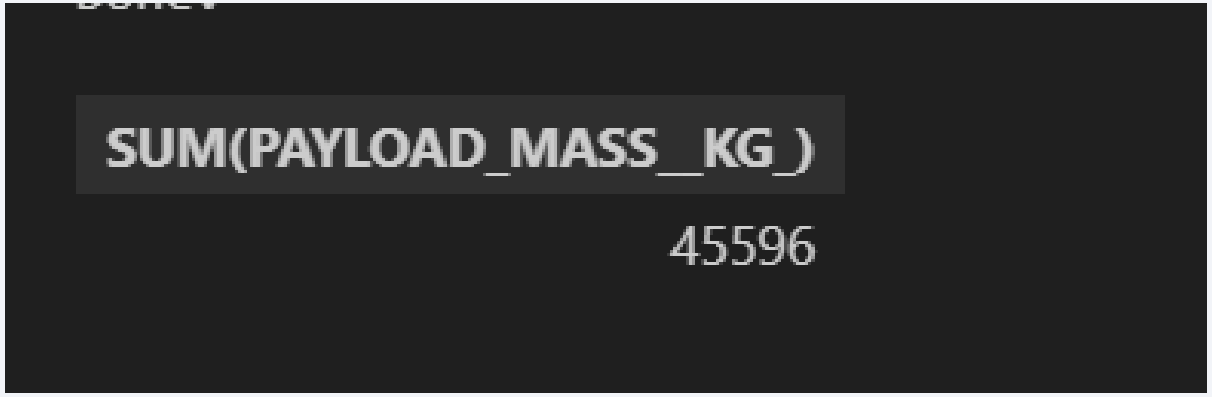
Launch Site Names Begin with 'CCA'

- %sql select * from SPACEXTABLE where Launch_Site like "CCA%" LIMIT 5
- retrieves the first 5 rows from the "SPACEXTABLE" table where the "Launch_Site" column starts with "CCA". The "%" symbol acts as a wildcard, allowing for partial matches.

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

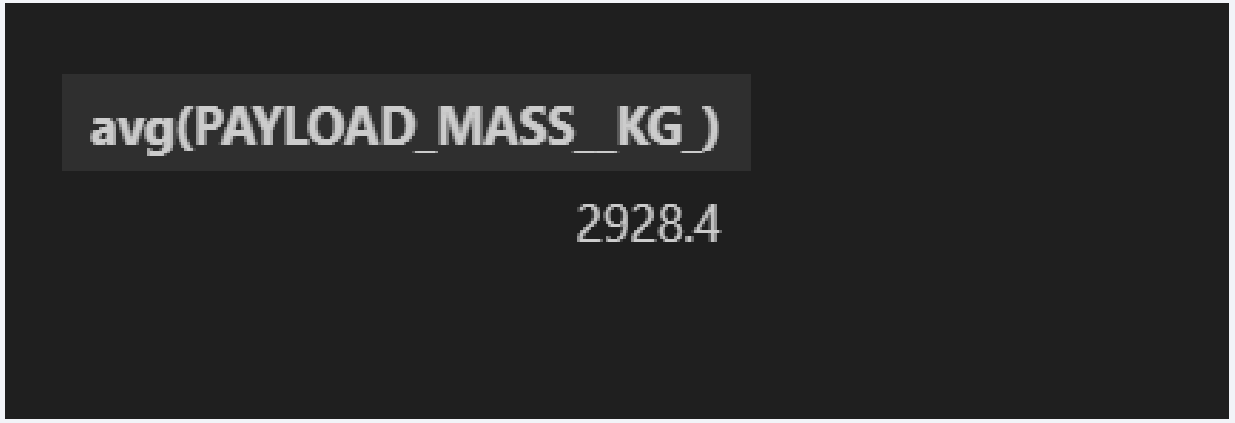
Total Payload Mass

- %sql select SUM(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer = "NASA (CRS)"
- query calculates the sum of the "PAYLOAD_MASS__KG_" column from the "SPACEXTABLE" table where the "Customer" column is "NASA (CRS)". It provides the total payload mass delivered to NASA (CRS) program by SpaceX



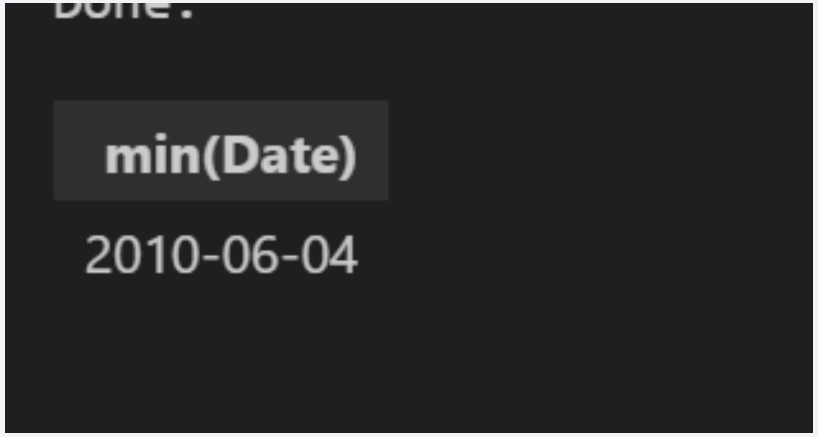
Average Payload Mass by F9 v1.1

- %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version = "F9 v1.1"
- query calculates the average payload mass (in kilograms) for launches where the "Booster_Version" column is "F9 v1.1" in the "SPACEXTABLE" table. It provides the average payload mass for launches using the Falcon 9 version 1.1 booster.



First Successful Ground Landing Date

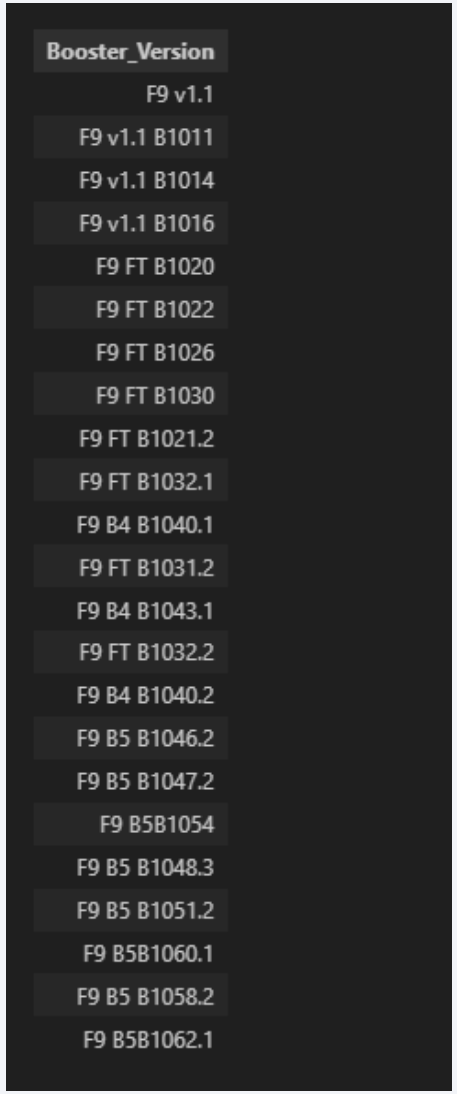
- %sql select min(Date) from SPACEXTABLE where Mission_Outcome = "Success"
- query retrieves the earliest date of successful missions from the "SPACEXTABLE" table, where the "Mission_Outcome" column is labeled as "Success". It provides the date of the first successful mission recorded in the dataset.



Successful Drone Ship Landing with Payload between 4000 and 6000

- %sql select Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ <6000

- query selects the "Booster_Version" from the "SPACEXTABLE" table where the "PAYLOAD_MASS__KG_" column is greater than 4000 kg and less than 6000 kg. It retrieves the booster versions used for payloads within this specified mass range.



Booster_Version
F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1014
F9 v1.1 B1016
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1030
F9 FT B1021.2
F9 FT B1032.1
F9 B4 B1040.1
F9 FT B1031.2
F9 B4 B1043.1
F9 FT B1032.2
F9 B4 B1040.2
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5B1054
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5B1060.1
F9 B5 B1058.2
F9 B5B1062.1

Total Number of Successful and Failure Mission Outcomes

- %sql SELECT COUNT(*), Mission_Outcome FROM SPACEXTABLE GROUP BY Mission_Outcome;

- query counts the number of occurrences of each unique value in the "Mission_Outcome" column from the "SPACEXTABLE" table and groups them by their respective outcomes. It provides a summary of the frequency of different mission outcomes recorded in the dataset.

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

Boosters Carried Maximum Payload

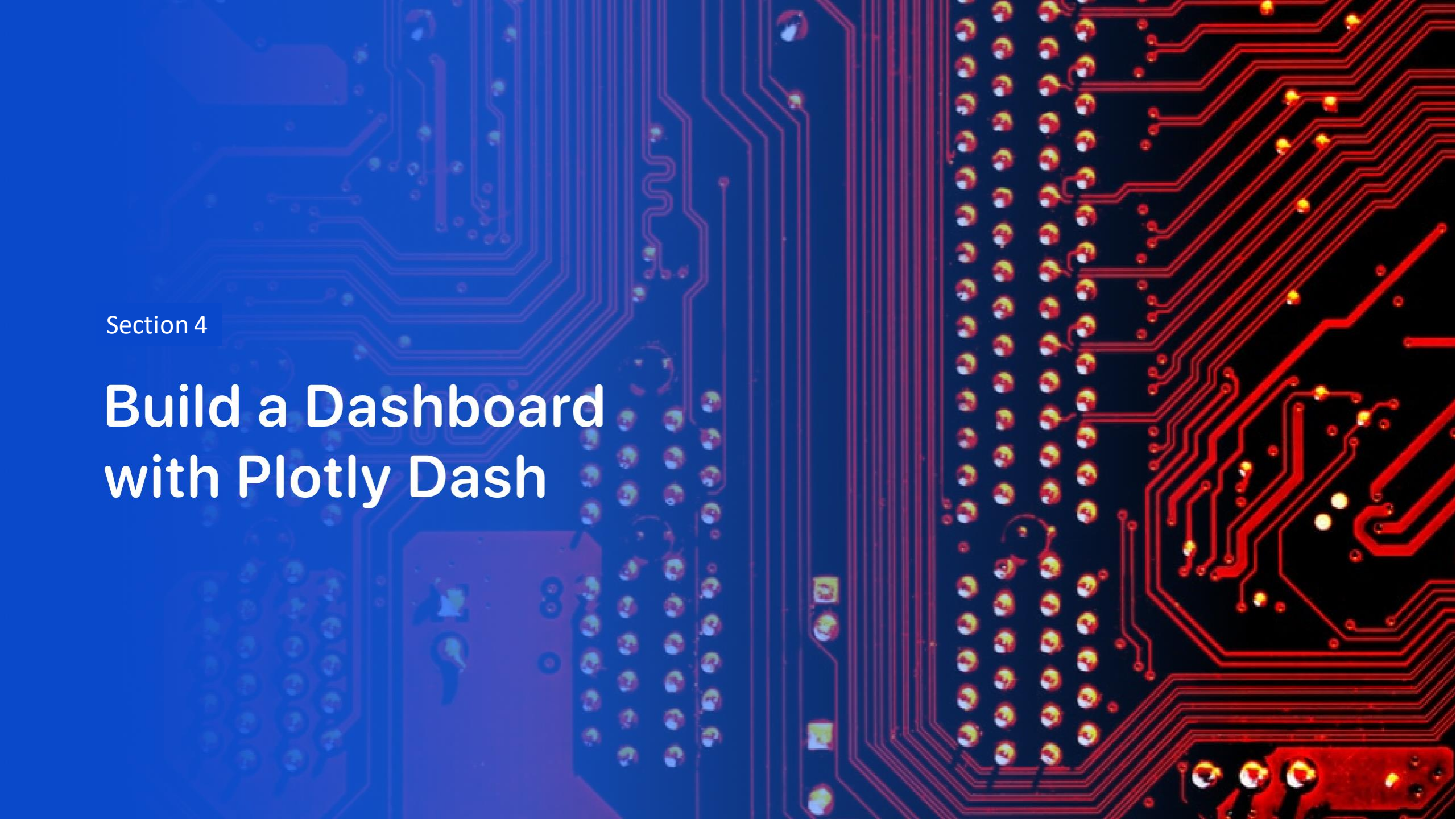
- %sql SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE Payload_Mass__kg_ = (SELECT MAX(Payload_Mass__kg_) FROM YourTableName);
- query selects unique values from the "Booster_Version" column in the "SPACEXTABLE" table where the payload mass is equal to the maximum payload mass found in the table. It retrieves the booster versions associated with the heaviest payloads recorded in the dataset. Please replace "YourTableName" with the actual name of your table

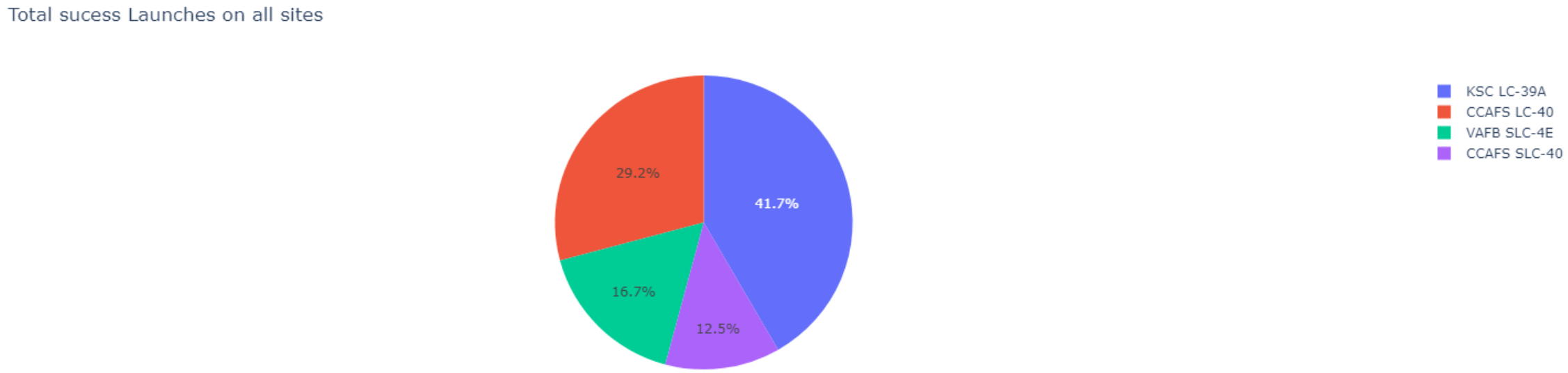
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- %sql SELECT Landing_Outcome, COUNT(Landing_Outcome) AS Outcome_Count FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY Outcome_Count DESC;

- query retrieves the count of each unique value in the "Landing_Outcome" column from the "SPACEXTABLE" table, filtered by a specified date range. The results are grouped by "Landing_Outcome" and ordered by the count of each outcome in descending order

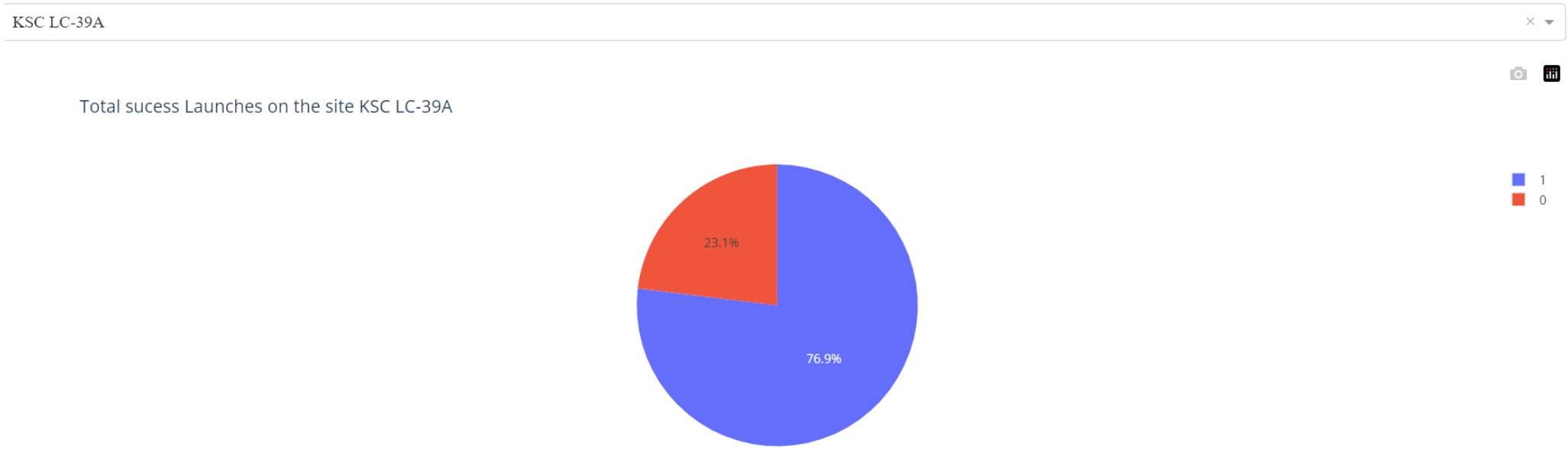
Landing_Outcome	Outcome_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1


Section 4
Build a Dashboard with Plotly Dash



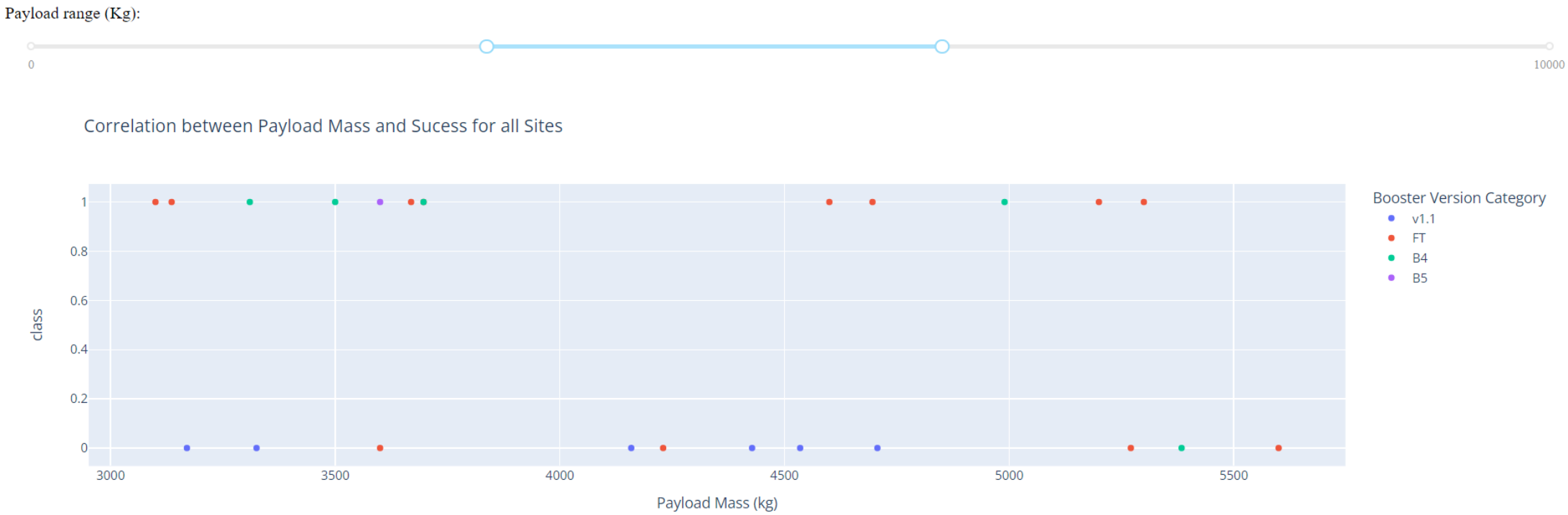
Total success on all sites

From the dashboard, on the figure "Total success Launches on all sites "we can note that the launch site KSC LC 39-A hast the best success rate compared than other site.



Total succesison the site KSC LC

- This site has 76% of launch success



Payload Mass vs Success for all sites

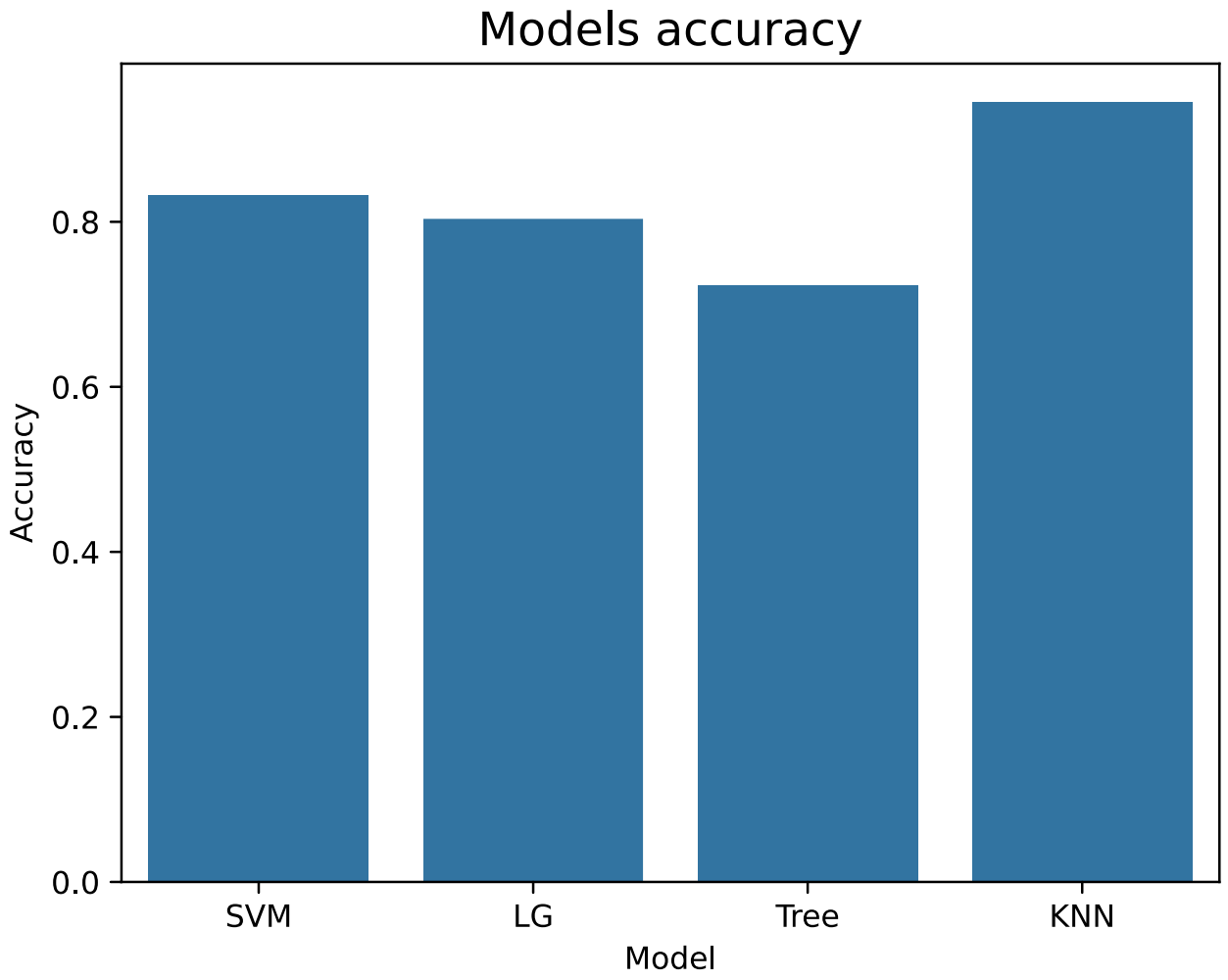
- For a paylaod between 3000 and 6000 for all sites, the success rate by heavier Rockets loock significanly smaller than rockets with smaller weight. I seem to be a correlation between the payload mass and a successful launch

Section 5

Predictive Analysis (Classification)

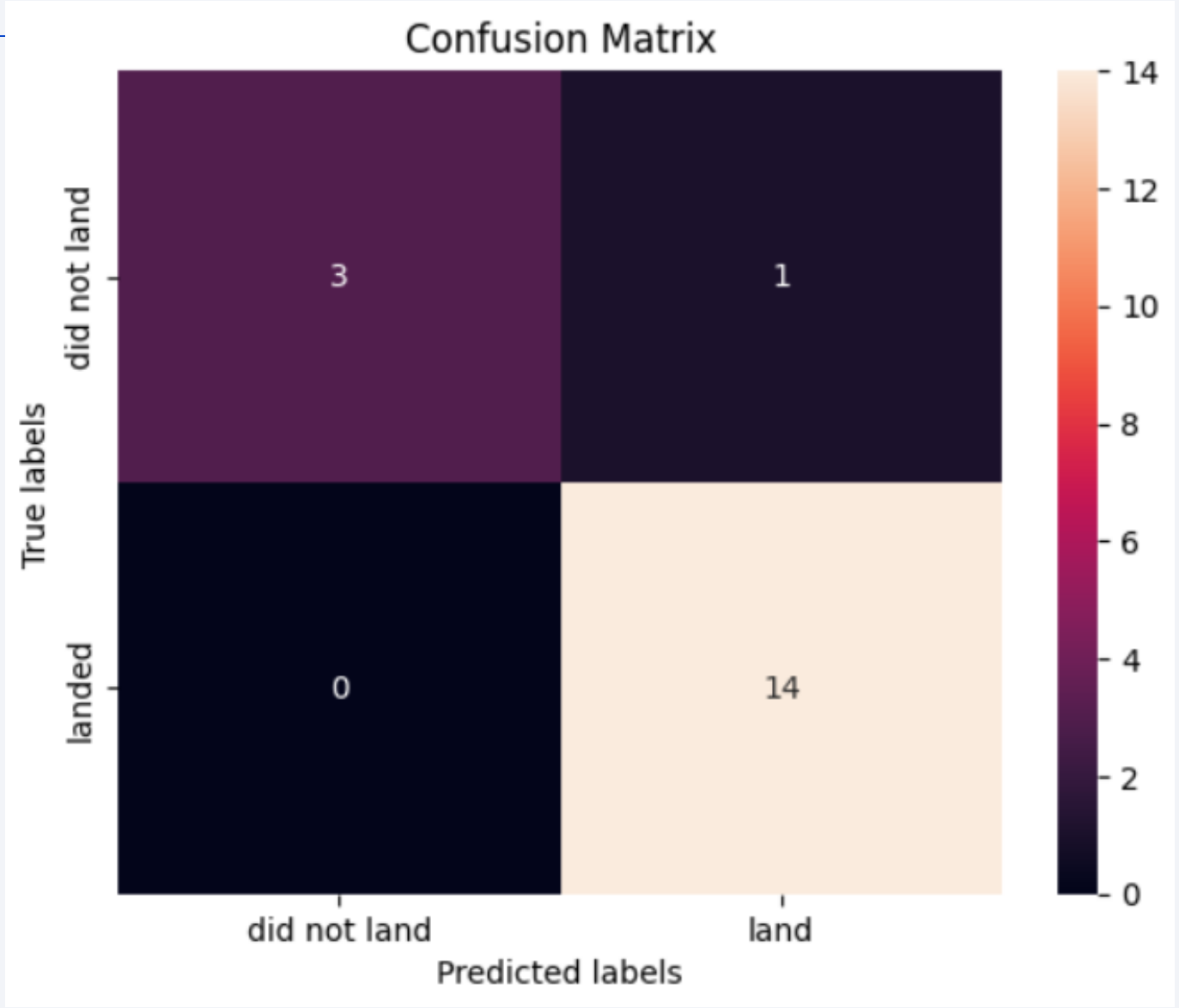
Classification Accuracy

- KNN hast the highest accuray



Confusion Matrix

- We have 18 Test entries. 14 are from the true class 1(landed succesfully) and 4 are from thes class 0(did not landed sucessflly.
- KNN hast a very high true positive rate(=1) because it predictid well all postive entries(14 entries)
- KNN predict bad one entry from the class 0. That is why it True Negative rate is less than 1



Conclusions

- KSC LC 39-A hast the best success rate compared than other site
- For all sites, the success rate by smaller Rockets are significanly higher than rockets with bigger weight.
- To predict if a launch will be succesfull, KNN is the best model with an accuray of 0.9

Thank you!