IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Seid Mohammed Adem

03.12.2023

Outline

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

Executive Summary

- Methodologies used for the report
  - Data collection using Web Scraping and API
  - Data Wrangling
  - Exploratory Data Analysis and Visualization with SQL and Folium
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis helps to identify best features
  - Machine Learning Prediction used to predict according to the best featuers

Introduction

- SpaceX disrupt the space industry by reducing the cost from 165 to 62 million dollars
- The aim of this project is to develop a machine learning pipeline to predict the landing outcome of the first stage.
- The problems that will be answered:
  - The best way of prediction to estimate the total cost of launches
  - To identify the best place of launches

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using web scrapping and using SpaceX API
- Perform data wrangling
  - Collected data was prepared and checked from errors and missing values by creating landing outcome label on the outcome data
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - The were normalized, divided in training and test set and the n by four different classification models. Based on the accuracy of each model was evaluted

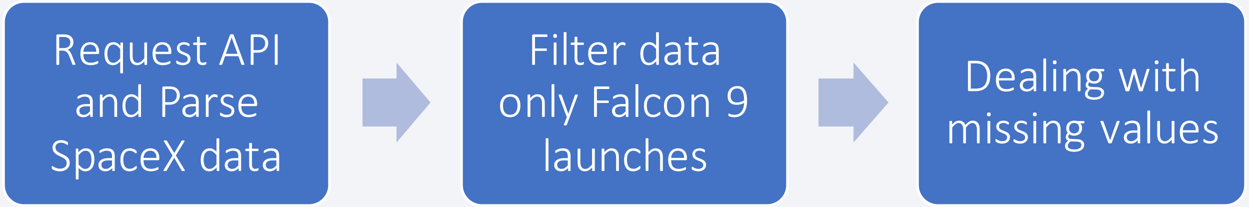
Data Collection

- Data set were collected :
    - SpaceX API
    - Using web scraping
- Data collection process use key phrases and flowcharts

Data Collection – SpaceX API

- SpaceX offers a public API from where data can be obtained

- Please have a look on my github (https://github.com/Seid-M-Adem/Applied-Data-Science-Capstone/blob/main/data-collection-api.ipynb)



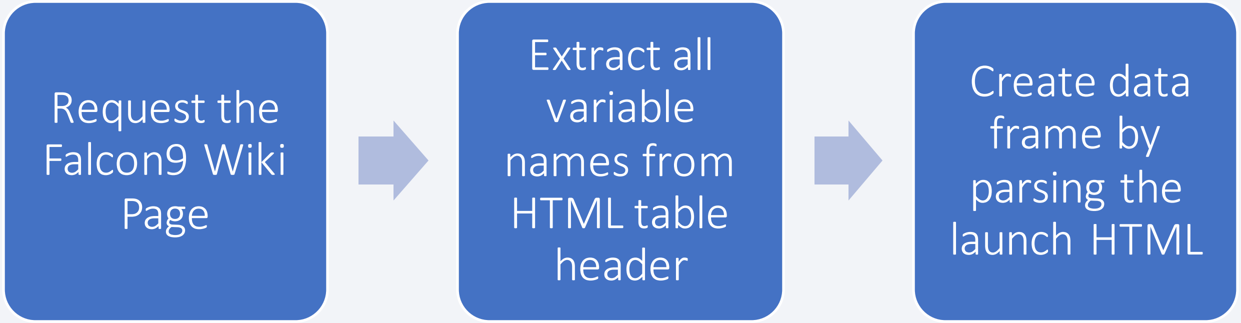
Request API and Parse SpaceX data

Filter data only Falcon 9 launches

Dealing with missing values

Data Collection - Scraping

- Data from SpaceX launches can also be obtained from Wikipedia.
- The data are scraped as the flowchart.
- Have a look the detail in my GitHub: https://github.com/Seid-M-Adem/Applied-Data-Science-Capstone/blob/main/webscraping-3.ipynb



Request the Falcon9 Wiki Page

Extract all variable names from HTML table header

Create data frame by parsing the launch HTML

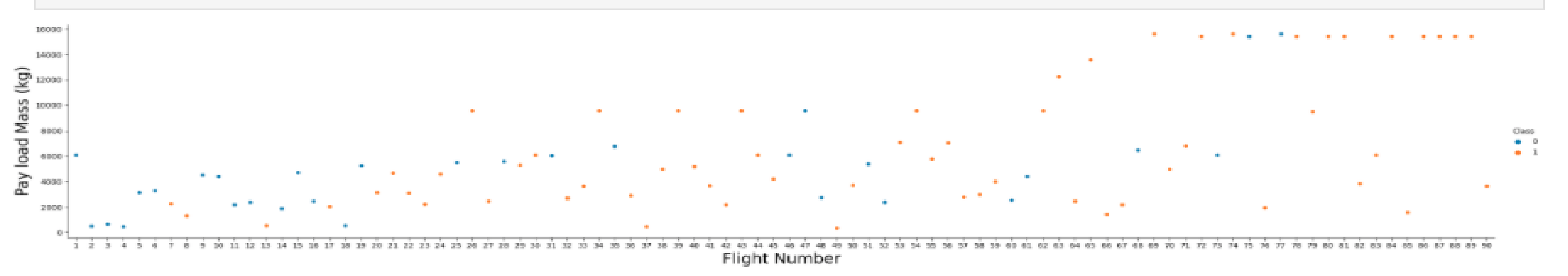
Data Wrangling

- Exploratory data analysis was performed to see the data
- Then the summaries launches per site, occurrences of each orbit and mission outcomes per orbit type were calculated
- The landing outcome label created from the outcome column
- See the details on my GitHub: https://github.com/Seid-M-Adem/Applied-Data-Science-Capstone/blob/main/data_wrangling_jupyterlite.jupyterlite.ipynb

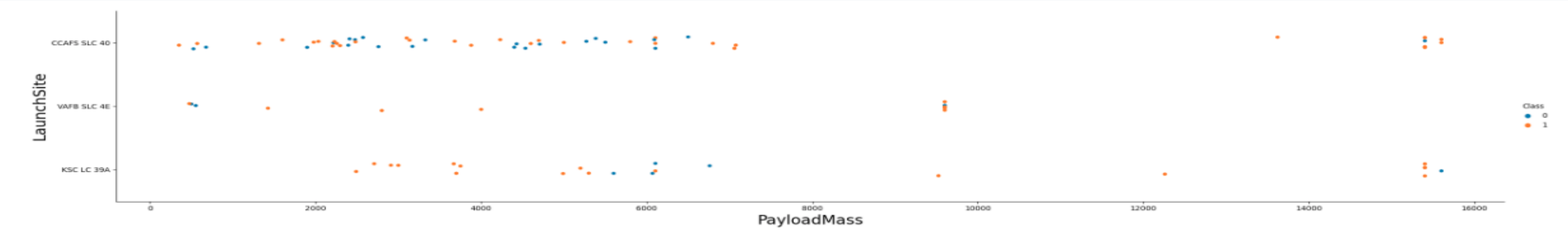
EDA using Dataset
Summaries of mission
Landing Outcome label

EDA with Data Visualization

- We plot out the FlightNumber vs. PayloadMassand and we see that as the flight number increases, the first stage is more likely to land successfully.



- We observe Payload Vs. Launch Site scatter point chart



- See more detail on my GitHub: https://github.com/Seid-M-Adem/Applied-Data-Science-Capstone/blob/main/eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- Summarize the SQL queries performed

  - Display the names of the unique launch sites: **%sql** select distinct launch_site from SPACEXTBL;
  - Display 5 records where launch sites with the string 'CCA': **%sql** SELECT * FROM SPACEXTBL WHERE launch_site like 'CCA%' limit 5;
  - Display the total payload mass carried by boosters launched by NASA (CRS): **%sql** select sum(payload_mass__kg_) as total_payload_mass from SPACEXTBL where customer = 'NASA (CRS)';
  - Display average payload mass carried by booster version F9 v1.1: **%sql** SELECT AVG(PAYLOAD_MASS__KG_) AS AVERAGE_PAYLOAD FROM SPACEXTBL WHERE Booster_Version like 'F9 v1.1
  - List the date when the first succesful landing outcome in ground pad was achieved: **%sql** SELECT (DATE) AS SUCCESS_GP FROM SPACEXTBL WHERE Landing_Outcome = 'Success (ground pad)';
  - List of the boosters in drone ship that have mass > 4000 but < 6000: **%sql** SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND Landing_Outcome = 'Success (drone ship)';
  - List the total number of successful and failure mission outcomes
  - List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
  - List the records that display the month, failure outcomes in drone ship ,booster versions, launch_site in year 2015.

EDA with SQL

- Summarize the SQL queries performed
  - List the total number of successful and failure mission outcomes
  - List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
  - 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.

  - See details on my GitHub: https://github.com/Seid-M-Adem/Applied-Data-Science-Capstone/blob/main/eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- **Interactive visual analytics using Folium by using the launch data.** For example latitude and longitude coordinates at each site by labeling the name of the launch site a circle marker is added.
- We assigned the outcome failure and success to class 0 and 1 with red and green respectively.

To calculated the distance of the launch sites to various landmark, we used the Haversine's formula for example to answer the questions like:

- How close the launch sites with railways, highways and coastlines?
- How close the launch sites with nearby cities?

From: https://github.com/Seid-M-Adem/Applied-Data-Science-Capstone/blob/main/Folium.zip

Hint: Please download and unzip to see

Build a Dashboard with Plotly Dash

- We built an interactive dashboard with Plotly dash which allowing the user to see the percentage of the launches by site and payload range as we need .
- The ploted pie charts showing the total launches by a certain sites.
- The plotted scatter graph showing the relationship with Outcome and Payload
- It will help to identify where is the best place to launch according to payloads.

Source code: https://github.com/Seid-M-Adem/Applied-Data-Science-Capstone/blob/main/dash_interactivity.py

Predictive Analysis (Classification)

Building the Model

- Load the dataset
- Transform the data and then split it
- Decide which type of ML to use
- Set the parameters and algorithms to GridSearchCV and fit it to dataset

Evaluating the Model

- Check the accuracy for each model
- Get tuned hyperparameters for each type of algorithms
- Plot the confusion matrix

Improving the Mpdel

- Use Feature Engineering and Algorithm Tuning

Find the Best Model

- The model with the best accuracy score will be the best performance model

- Source code: https://github.com/Seid-M-Adem/Applied-Data-Science-Capstone/blob/main/Prediction%20Analysis.pdf
- Hint : click more

Results

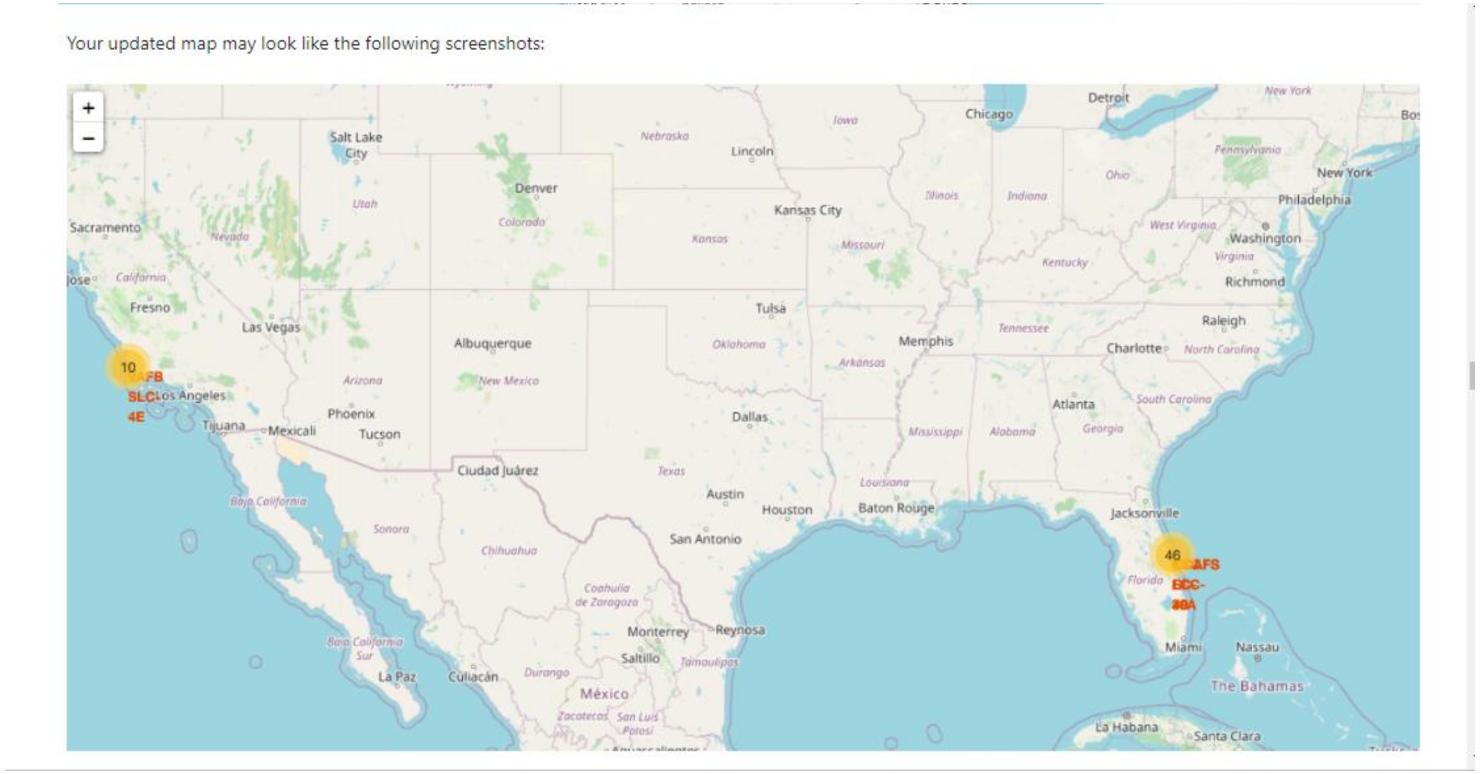
Exploratory data analysis results:

- Space X uses 4 different launch sites;
- The first launches were done to Space X itself and NASA;
- The average payload of F9 v1.1 booster is 2,928 kg;
- The first success landing outcome happened in 2015 fiver year after the first launch;
- Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average;
- Almost all of the mission outcomes were successful;
- Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015;
- The number of landing outcomes became as better as years passed.

Results

Interactive analytics help us to identify the launch sites used to be in
safe places, near sea, for example and have a good logistic infrastructure around.

- Most launches happens at east cost launch sites.



Your updated map may look like the following screenshots:

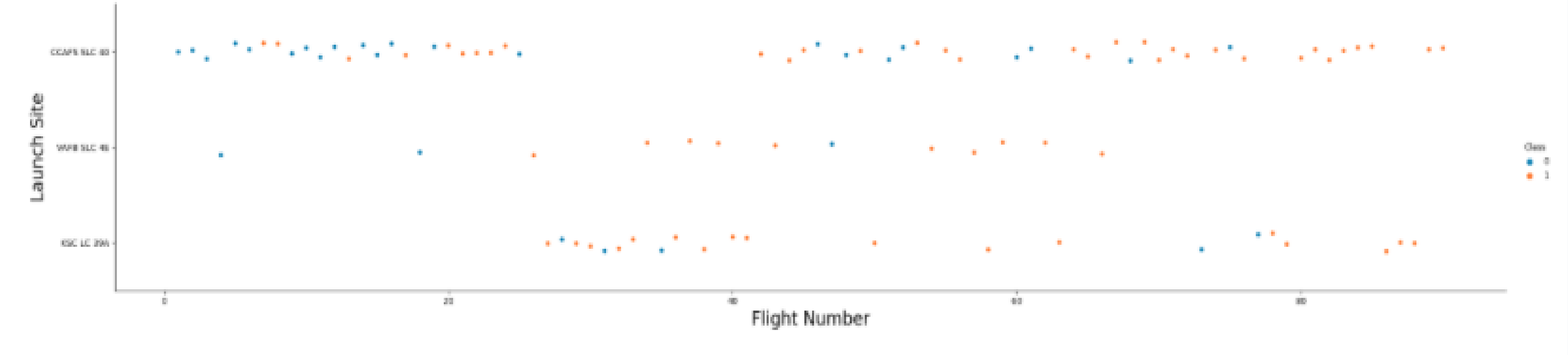
Results

- Based on the Predictive, Decision Tree Classifier is the best model to predict successful landings, having accuracy over 87% and accuracy for test data over 94%.

Section 2

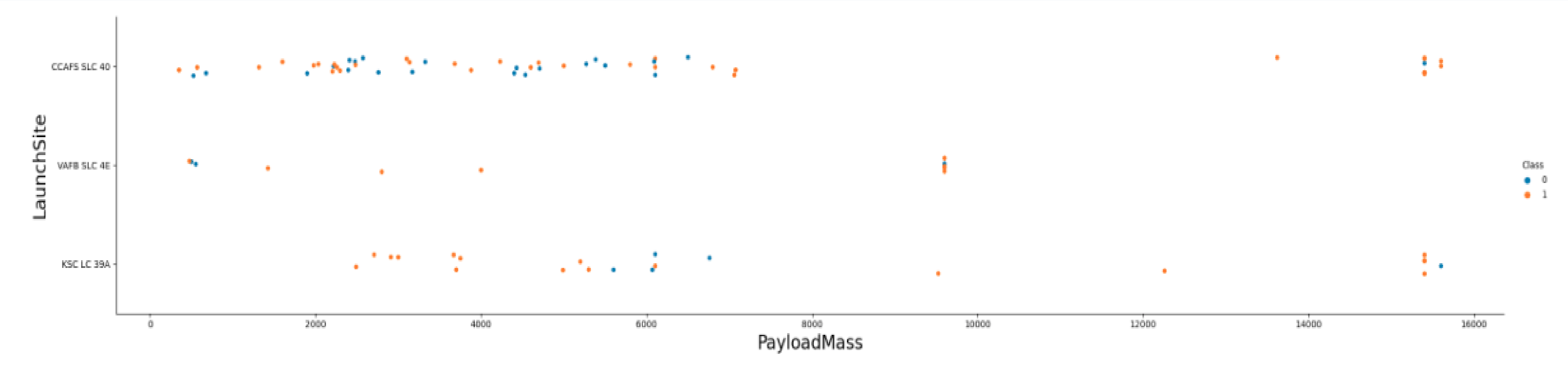
Insights drawn from EDA

Flight Number vs. Launch Site



- As we on the above plot, the best launch site nowadays is CCAF5 SLC 40, where most of recent launches were successful;
- In addition we can see that the second place VAFB SLC 4E and third place KSC LC 39A;
- The general success rate improved over time.

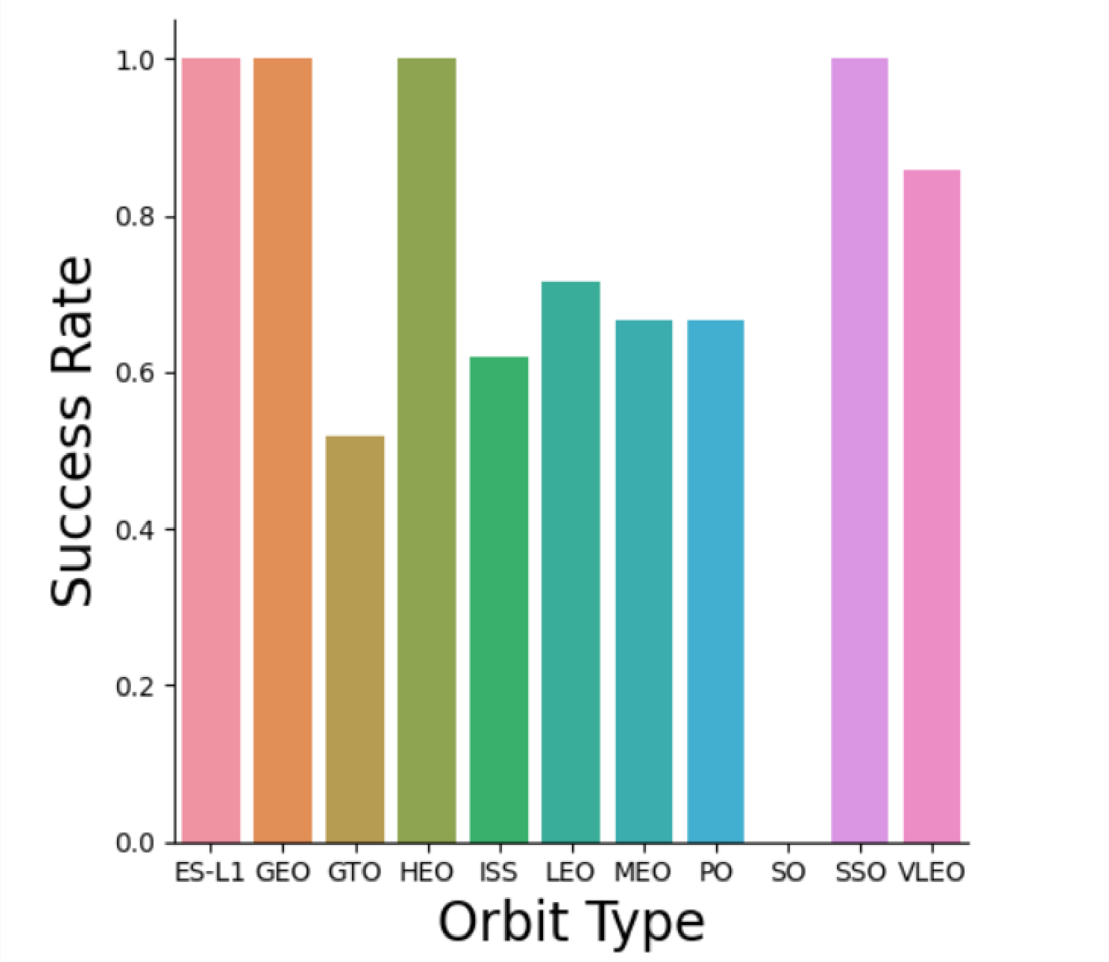
Payload vs. Launch Site



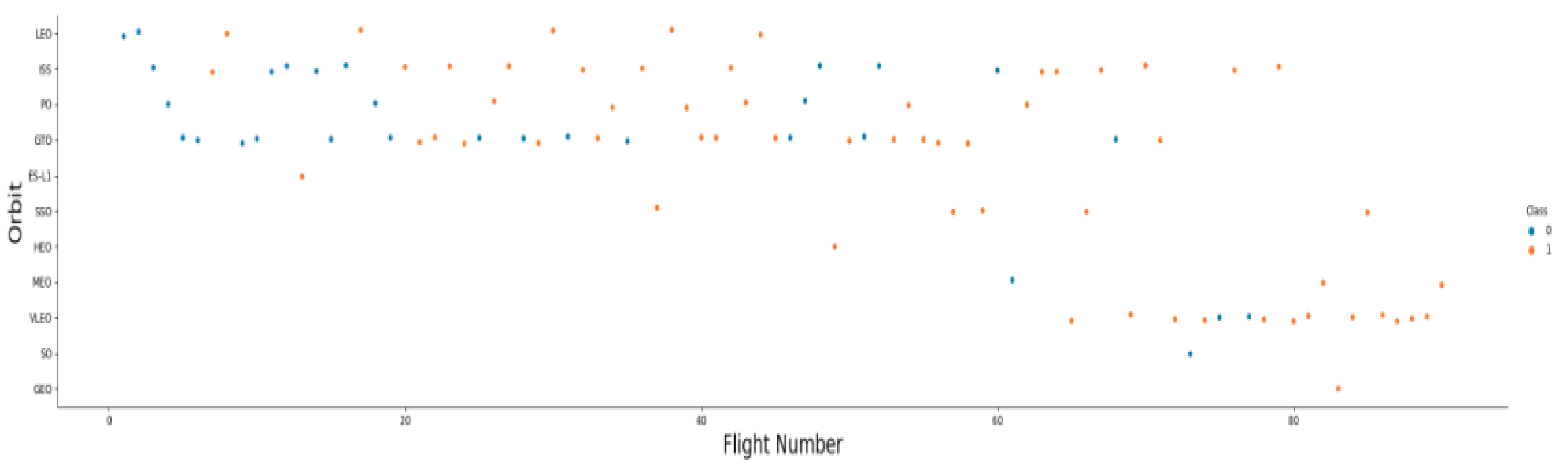
- Payloads over 9,000kg have excellent success rate
- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.

Success Rate vs. Orbit Type

- The highest success rates happens to orbits:
- ES-L1;
- GEO;
- HEO; and
- SSO

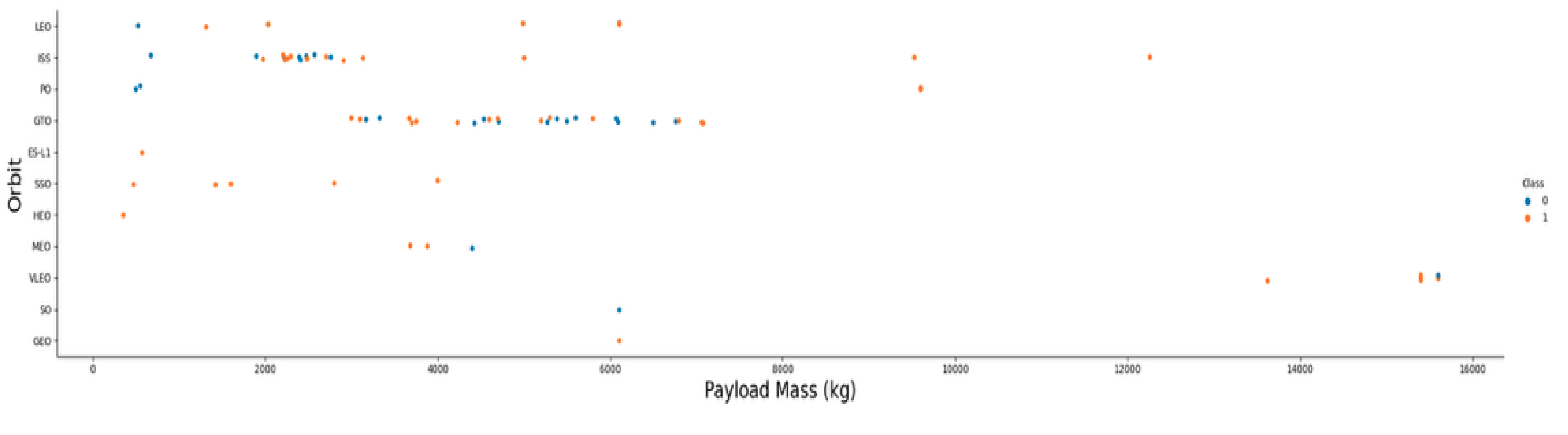


Flight Number vs. Orbit Type



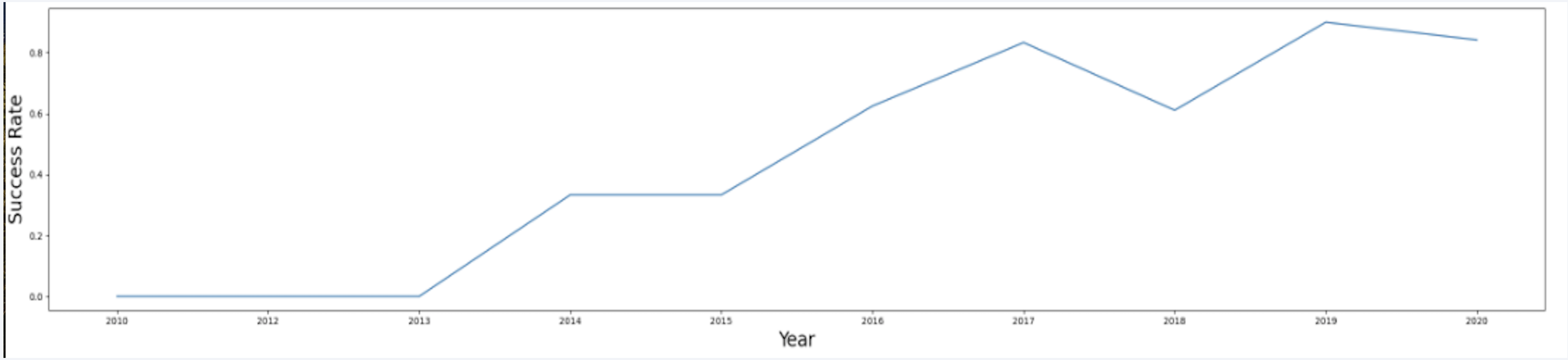
- Success rate improved over time to almost all orbits
- VLEO is the most frequently used orbit

Payload vs. Orbit Type



- ISS orbit has the widest range of payload and a good rate of success;
- There are few launches to the orbits SO and GEO
- We cannot see relation between payload and success rate to orbit GTO

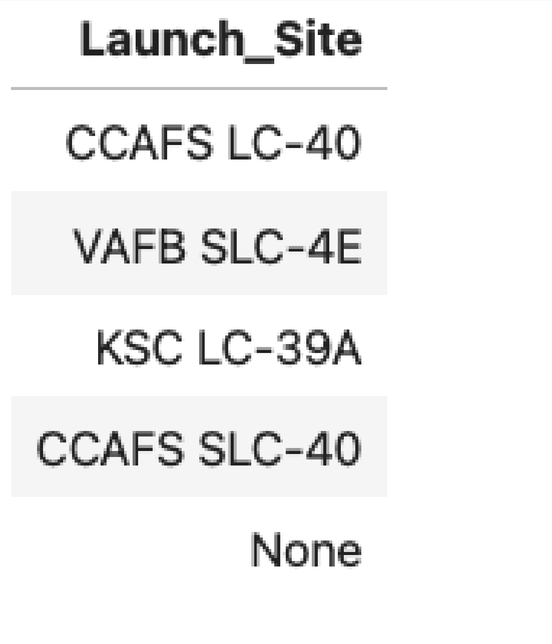
Launch Success Yearly Trend



- Success rate is increasing from 2013 until 2020
- The first three years were a period of development

All Launch Site Names

- **There are four launch sites:**

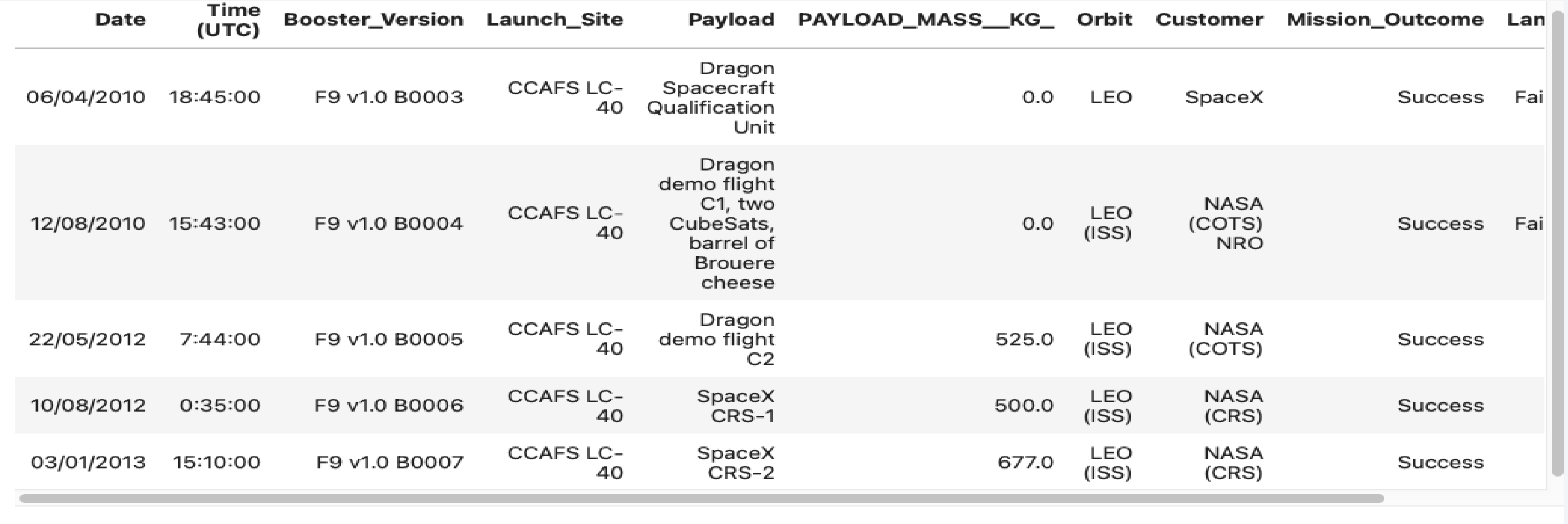


Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

- **The query used: %sql** select distinct launch_site from SPACEXTBL;

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`

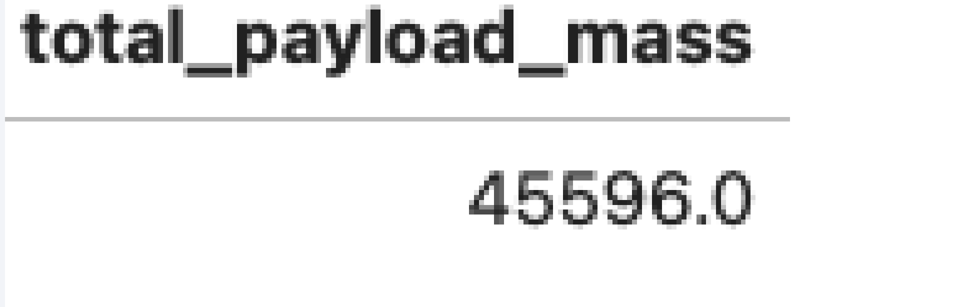


Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Lan
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Fai
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Fai
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	

- The query used: **%sql SELECT * FROM SPACEXTBL WHERE launch_site like 'CCA%' limit 5;**

Total Payload Mass

- The total payload carried by boosters from NASA



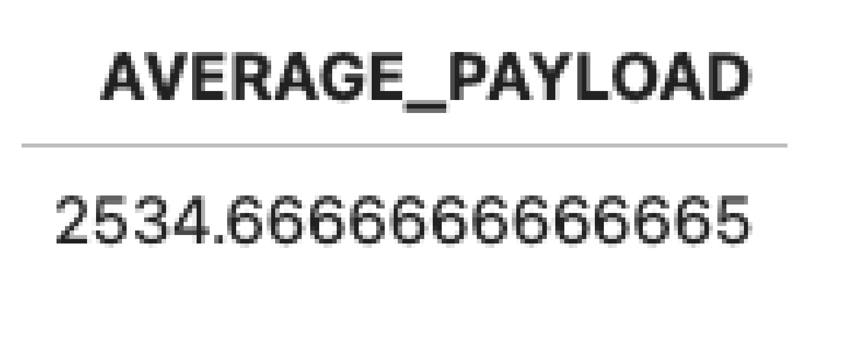
total_payload_mass
45596.0

- Total payload calculated by summing all payloads whose codes contain ‘CRS’, which corresponds to NASA using the following query:

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

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1

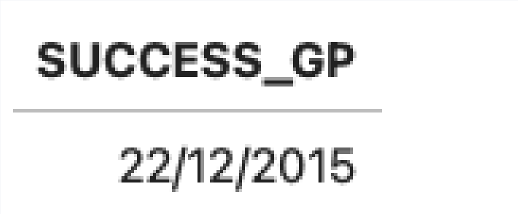


AVERAGE_PAYLOAD
2534.6666666666665

- The query used for the above result: **%sql** SELECT AVG(PAYLOAD_MASS__KG_) AS AVERAGE_PAYLOAD FROM SPACEXTBL WHERE Booster_Version like 'F9 v1.1%';

First Successful Ground Landing Date

- The first successful landing outcome on ground pad:

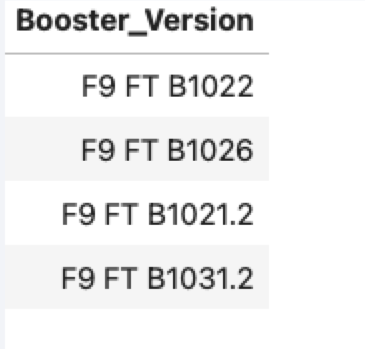


SUCCESS_GP
22/12/2015

- Getting the minimum value for date from the query

Successful Drone Ship Landing with Payload between 4000 and 6000

- List of boosters names which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are:



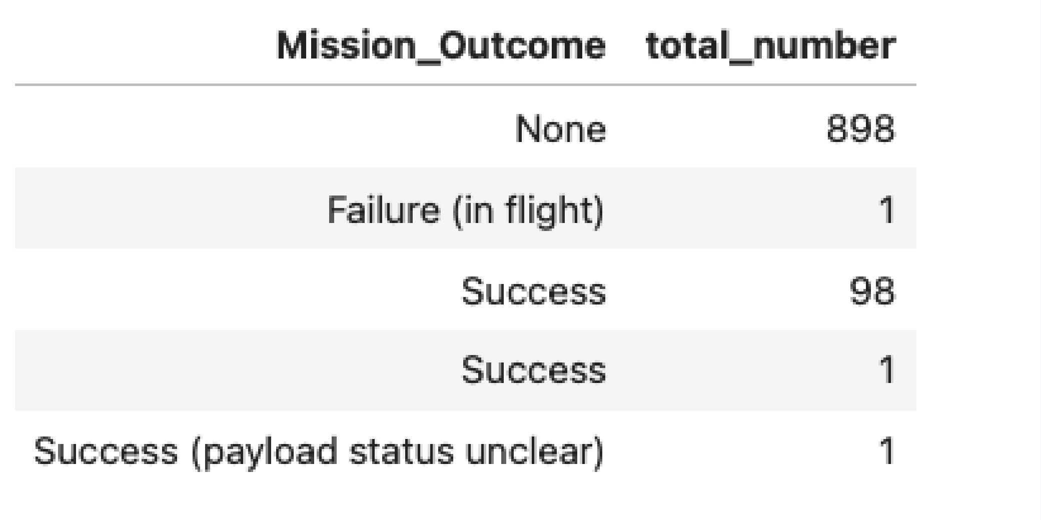
Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

- The query used for the above result:

- **%sql** SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND Landing_Outcome = 'Success (drone ship)';

Total Number of Successful and Failure Mission Outcomes

- The number of successful and failure mission outcomes:

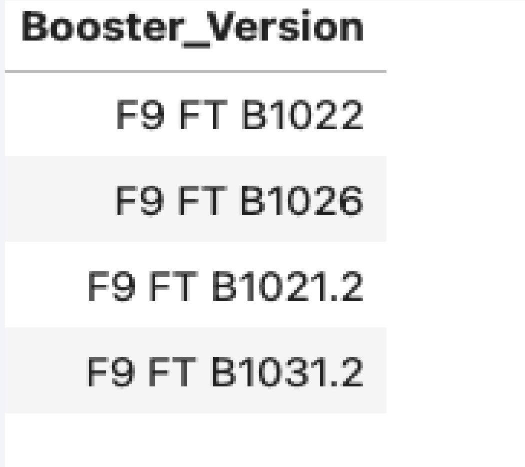


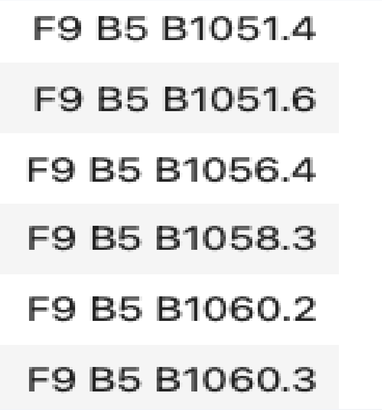
Mission_Outcome	total_number
None	898
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- The query used for the above result: sql select mission_outcome, count(*) **as** total_number **from** SPACEXTBL group by mission_outcome;

Boosters Carried Maximum Payload

- The list of the booster names which have carried the maximum payload mass are:



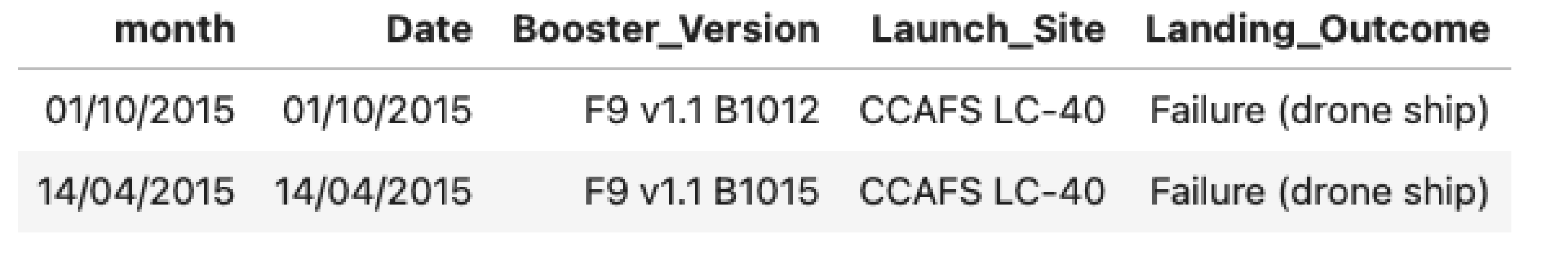


Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

- The query used for the above result: **%sql** SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL) ORDER BY BOOSTER_VERSION;

2015 Launch Records

- Failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015:

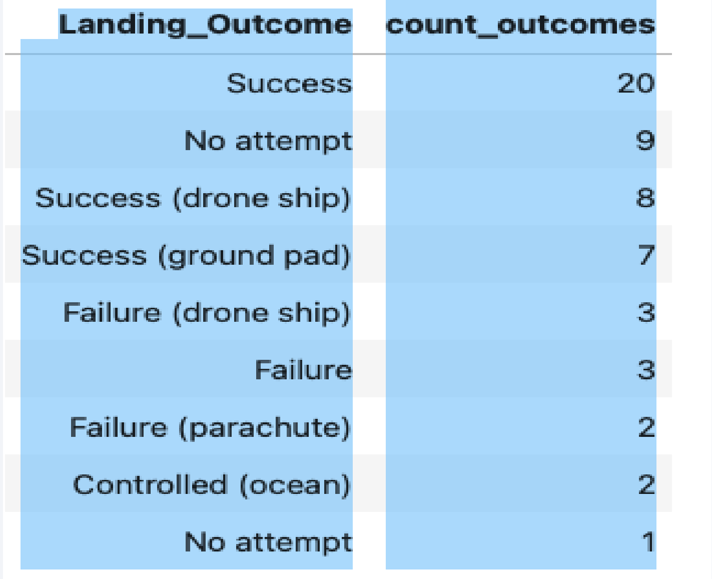


month	Date	Booster_Version	Launch_Site	Landing_Outcome
01/10/2015	01/10/2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
14/04/2015	14/04/2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

- The query used for the above result: **%%sql select** date **as month**, date, booster_version, launch_site, Landing_Outcome **from** SPACEXTBL **where** Landing_Outcom

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

- The rank 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_outcomes
Success	20
No attempt	9
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1

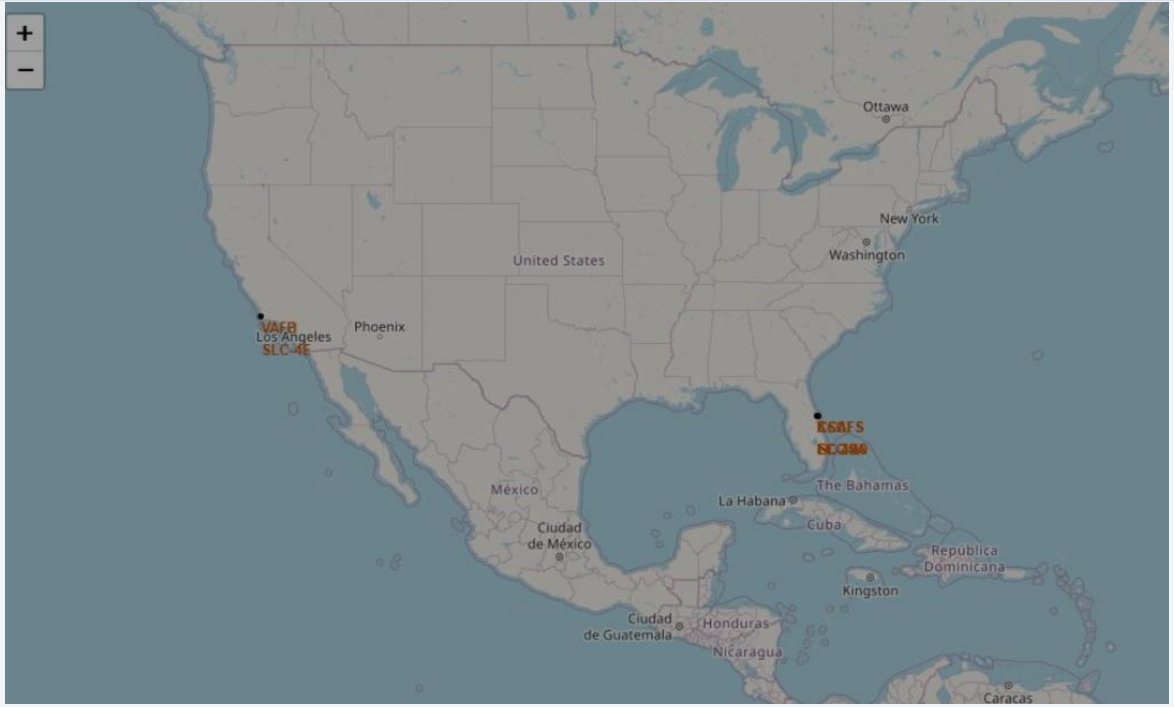
- The query used for the above result: **%%sql select Landing_Outcome, count(*) as count_outcomes from SPACEXTBL**

- **where date between '04/06/2010' and '20/03/2017'**
- **group by Landing_Outcome**
- **order by count_outcomes desc;**

Section 3

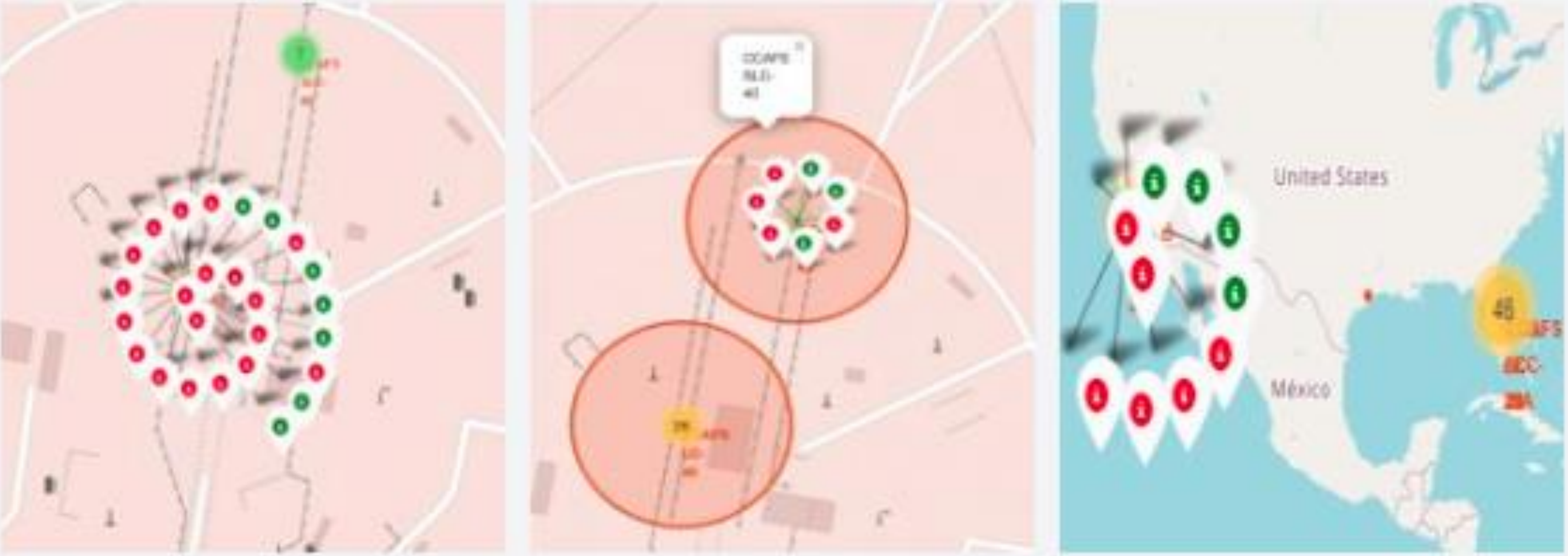
Launch Sites Proximities Analysis

Location of all launch sites



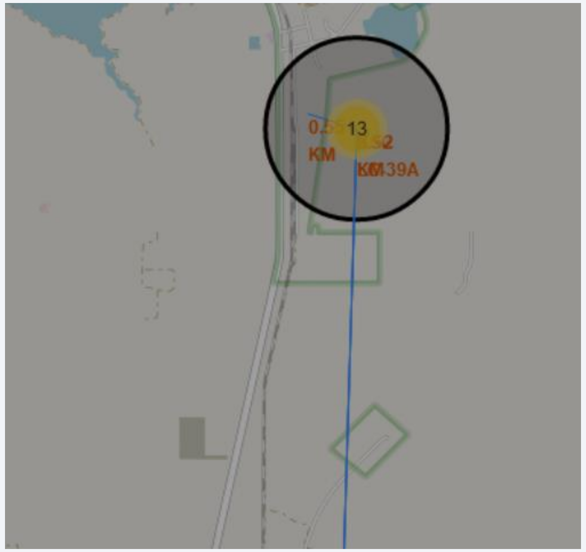
- Launch sites are near sea, probably by safety, but not too far from roads and railroads.

Colored markers to show the different launch site

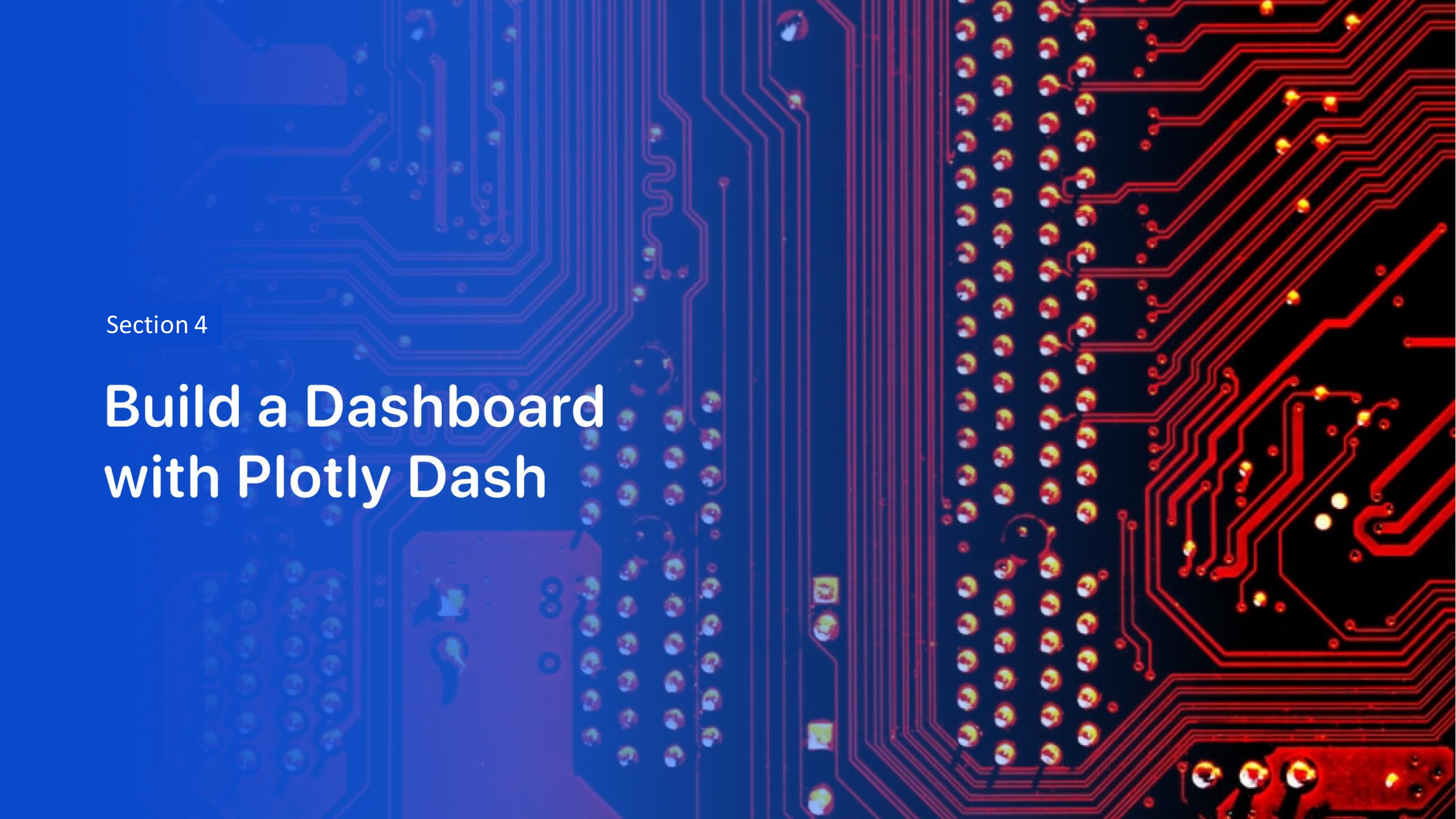


- Green markers indicate successful and red ones indicate failure

Launch sites with logistic and Safety



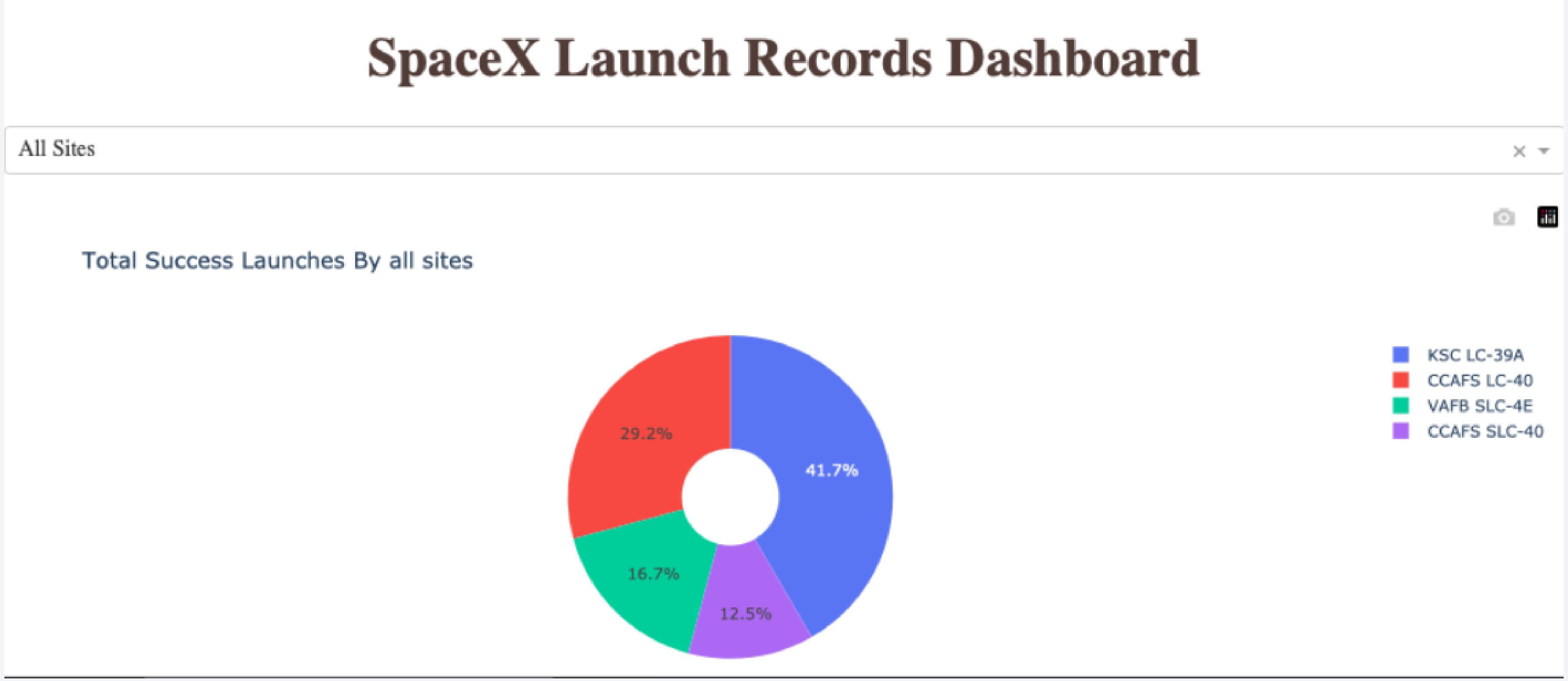
- Launch site KSC LC-39A has good logistics aspects, being near railroad and road and relatively far from inhabited areas.



Section 4

Build a Dashboard with Plotly Dash

Successful Launches Site



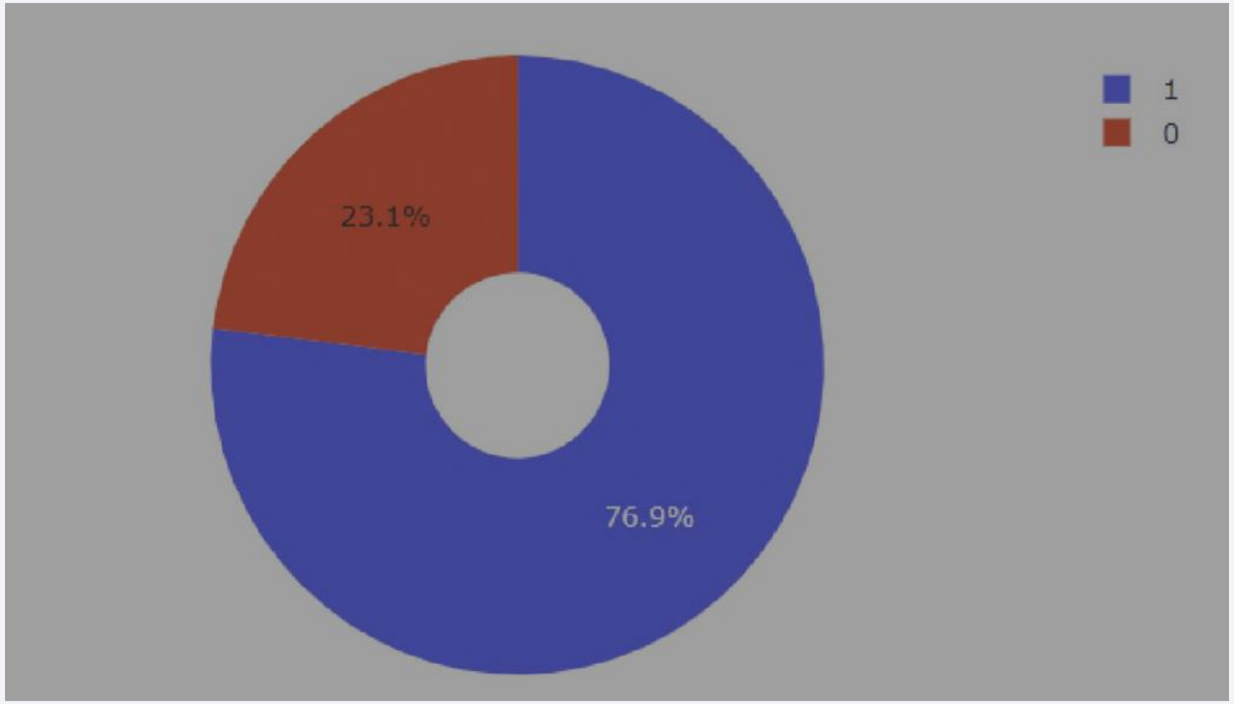
SpaceX Launch Records Dashboard

All Sites

Total Success Launches By all sites

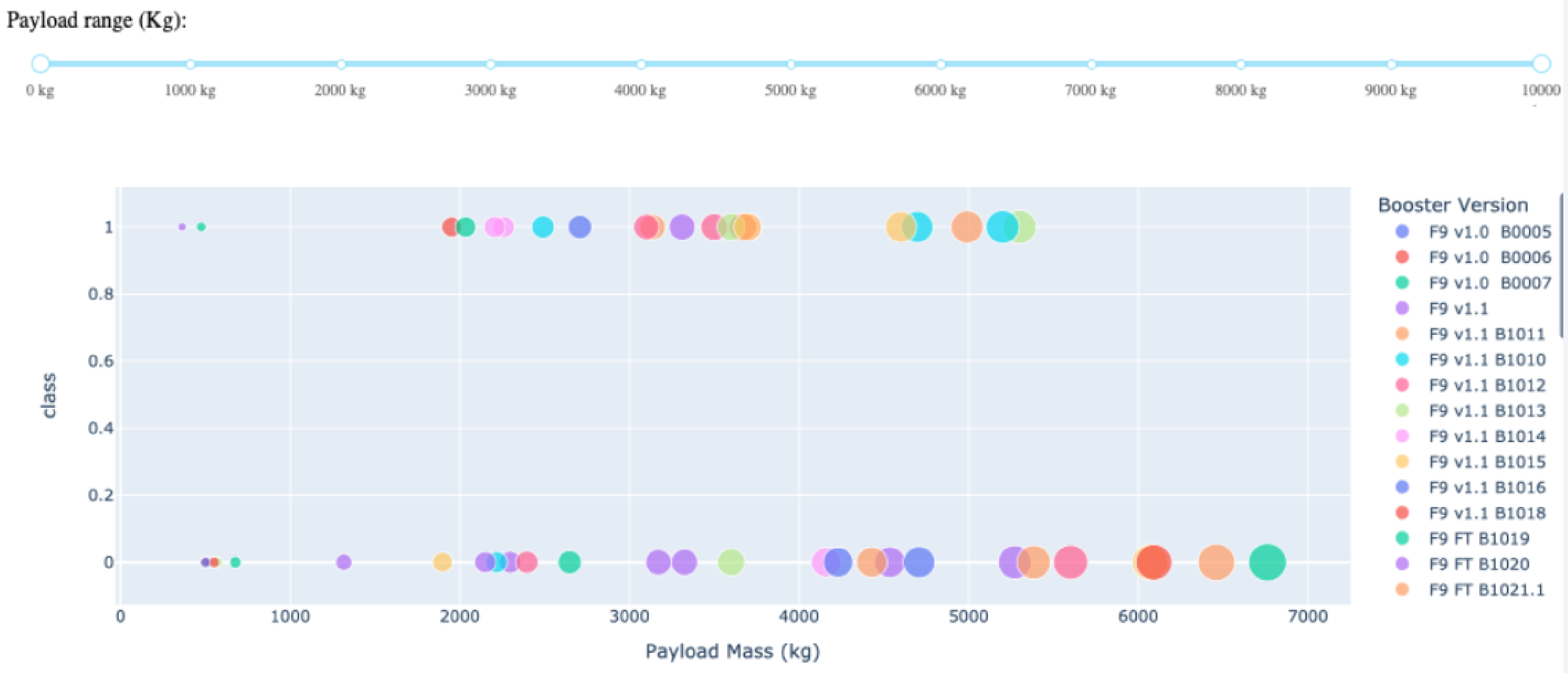
- KSC LC-39 have the most successful launches from all sites

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



- KSC LC-39A achieved a 76.9 success rate

Payload vs Launch Outcome Scatter Plot

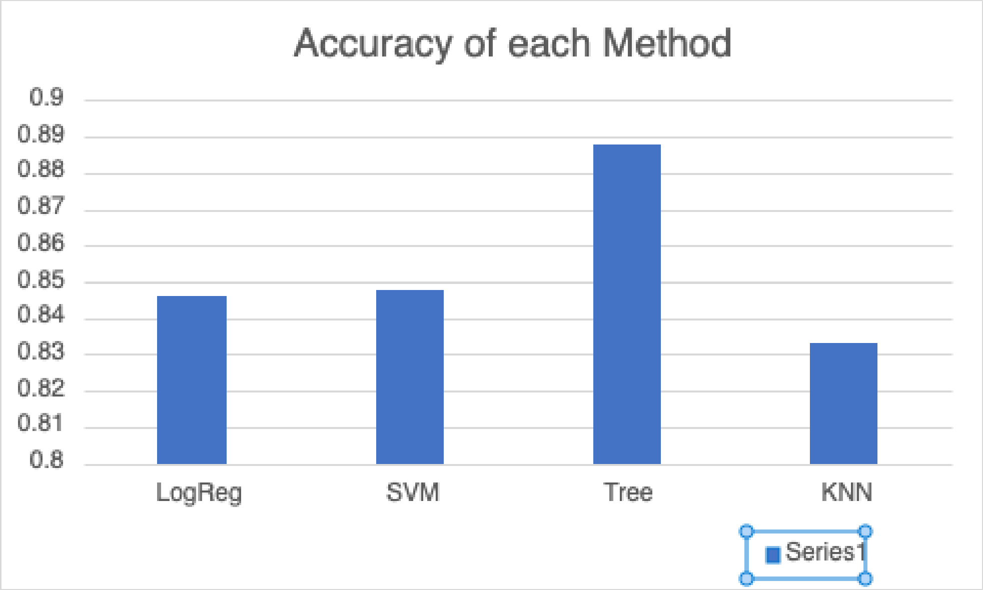


Payload range (Kg):

Section 5

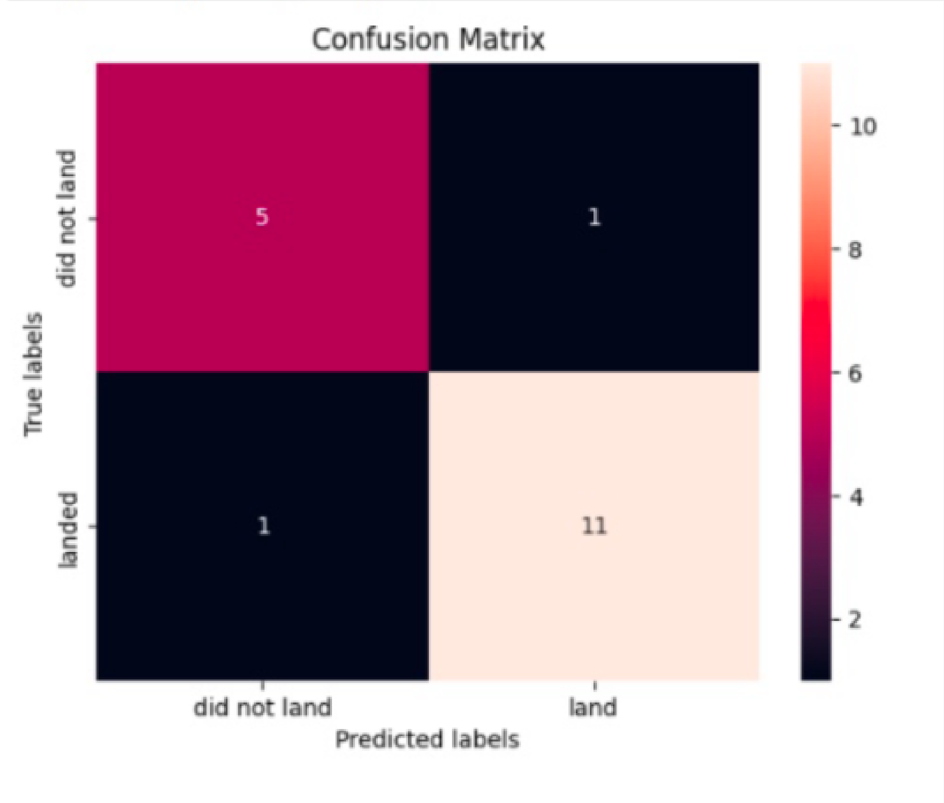
Predictive Analysis (Classification)

Classification Accuracy



- The model with the highest classification accuracy is Decision Tree which is 88%.

Confusion Matrix



- Confusion matrix of Decision Tree Classifier proves its accuracy by showing the big numbers of true positive and true negative compared to the false ones.

Conclusions

- We can conclude that:
- The Decision Tree Classifier Algorithm is the best Machine Learning approach for this dataset.
- The low weighted payloads performed better than the heavy weighted payloads.
- From 2013, the success rate for SpaceX launches is increased,
- KSC LC-39A is the most successful launches
- SSO orbit have the most success rate

Appendix

- GitHub : https://github.com/Seid-M-Adem/Applied-Data-Science-Capstone/tree/main

Thank you!