Winning Space Race with Data Science

Lapo Dini
09/11/2022

Outline

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

Executive Summary

Summary of methodologies

- Data collection via SpaceX API, and Web Scraping
- Exploratory Data Analysis (EDA) and Data Visualization
- EDA with SQL
- Interactive Map with Folium
- Dashboard with Plotly Dash
- Predictivce Analisys

Summary of all results

- Exploratory Data Analysis resulsts
- Interactive map and dashboard
- Predictive results

Introduction

- Project background and context

  With this project we want to predict the successful landing of the first stage of Falcon 9 rocket from SpaceX. On their website SpaceX say that the cost of a single launch is 62 million dollars, against competitors' costs that are about 165 million dollars. The main reason for this is that SapceX can reuse their first stage booster and cut the costs of a single launch.

  We can see this as a useful information for a company who wants to compete with SpaceX.

- Problems you want to find answers
  - What are the main characteristics of a successful landing?
  - How relationship between rocket variables affect the success of a landing?
  - What are the conditions that grant SpaceX a higher success rate?

Section 1

Methodology

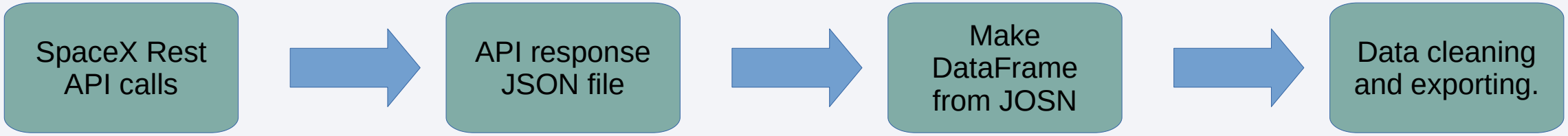
Methodology

Executive Summary

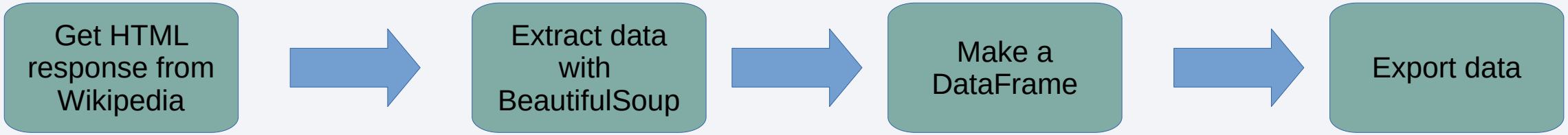
- Data collection methodology:
  - SpaceX API, Web Scraping
- Perform data wrangling
  - Filtered Falcon 9 data
  - Filled missing value in the dataset
  - Converting string variable in numbers.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Model Used: LogisticRegression, KNNeighbours, Supported Vector Machine, and Decision Tree.
  - Tuning with GridSearchCV.
  - Evaluation with Accuracy Score and Confusion Matrix.

Data Collection

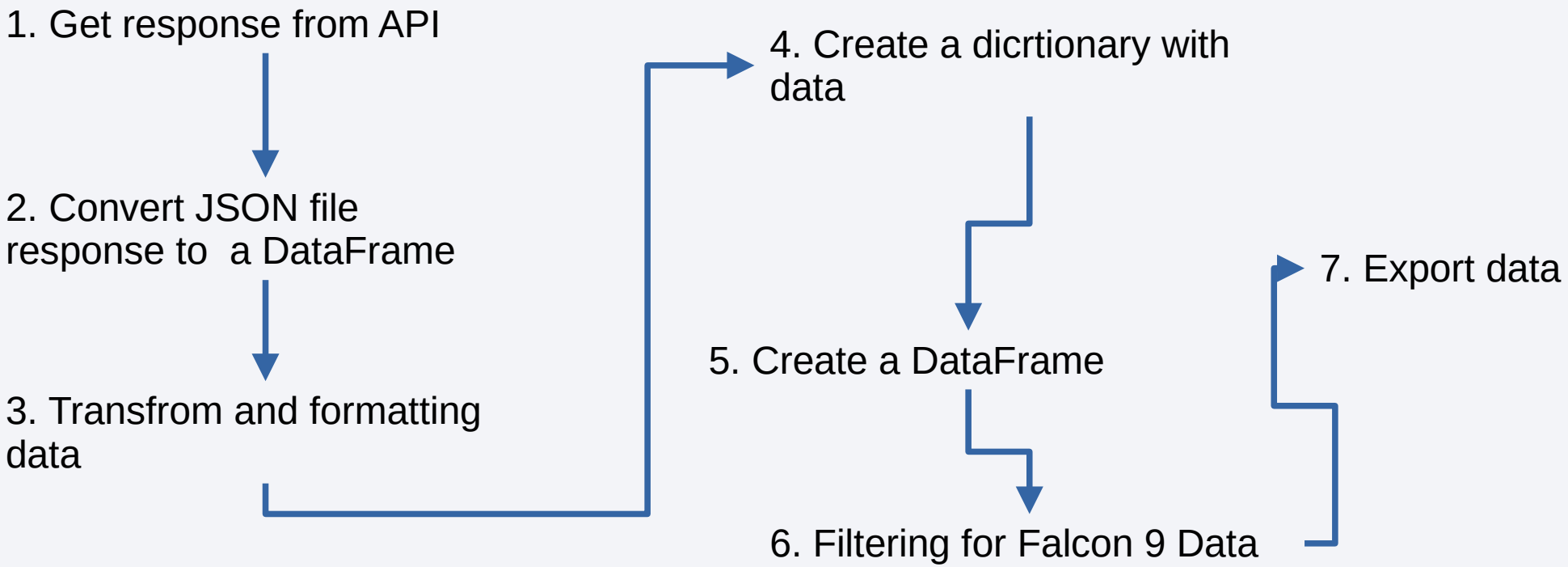
- Datasets are collected using SpaceX API and web scraping form Wikipedia.
  - SpaceX API informations: rocket, launch, payload information.



  - Wikipedia Informations: rocket, launches, landing, payload information.

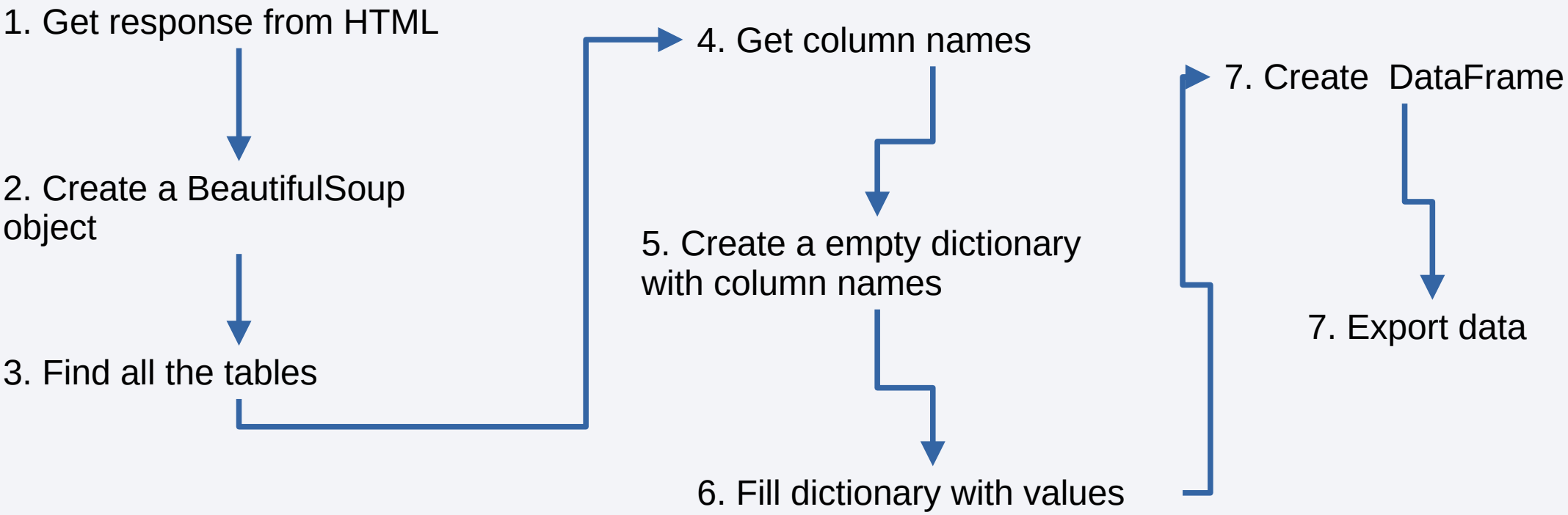


Data Collection – SpaceX API



Notebook with the entire process

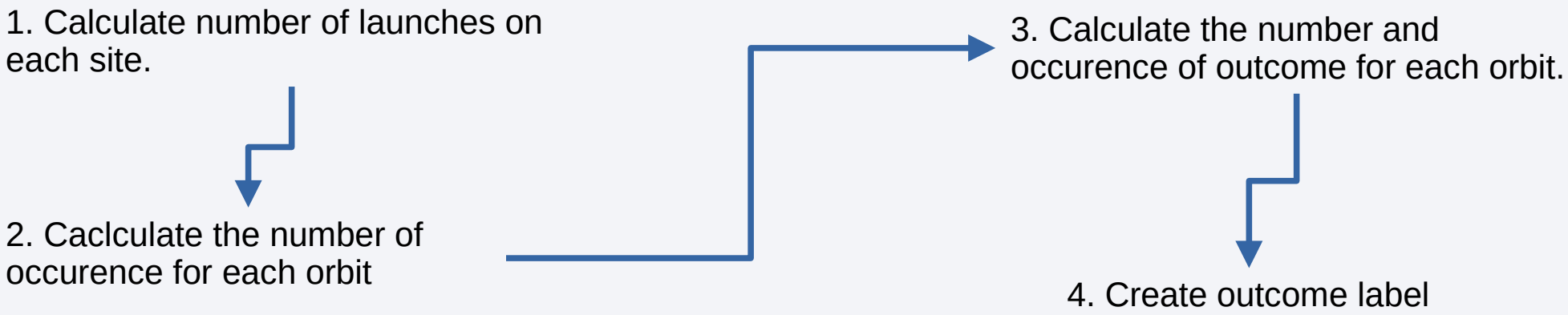
Data Collection – Scraping



Notebook with the entire process

Data Wrangling

- In the data set, there are several different cases where the booster did not land successfully.
  - True Ocean, True RTLS, True ASDS are positive outcomes (label 1)
  - False Ocean, False RTLS, True ASDS are negative outcomes (label 0)
- We need to transform categorical variables into numbers and fill missing values.



Notebook with the entire process

EDA with Data Visualization

Graphs presented in this report.

- **Scatter Graphs:**
  - Flight Number vs. Payload Mass
  - Flight Number vs. Launch Site
  - Payload Mass vs. Launch Site
  - Flight Number vs. Orbit Type
  - Payload Mass vs. Orbit Type

- **Bar Graphs:**
  - Rate of success for Orbit Type

Line Graphs:

- Success rate over Years

Notebook with the entire process

EDA with SQL

Queries Executed in the project:

- Name of launch sites.
- 5 records with name starting with CCA
- Total Payload Mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- Date when the first successful landing outcome in ground pad was acheived
- 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 failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Rank of landing outcomes between the date 2010-06-04 and 2017-03-20.

Build a Dashboard with Plotly Dash

Element added to the Dashboard:

- Dropdown with possibility to select a particular launch site or all sites.
- Pie chart representing:
  - Number of launches for each site, if ALL is selected.
  - Failed landings vs. Successful landings from the selected site.
- Range Slider to select a payload mass range
- Scatter plot that shows the relationship between payload mass and outcome for a selected site.

Explain why you added those plots and interactions

Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose

Predictive Analysis (Classification)

Data preparation:

- Load data
- Normalize data
- Split data into train and test sets

Model Preparation:

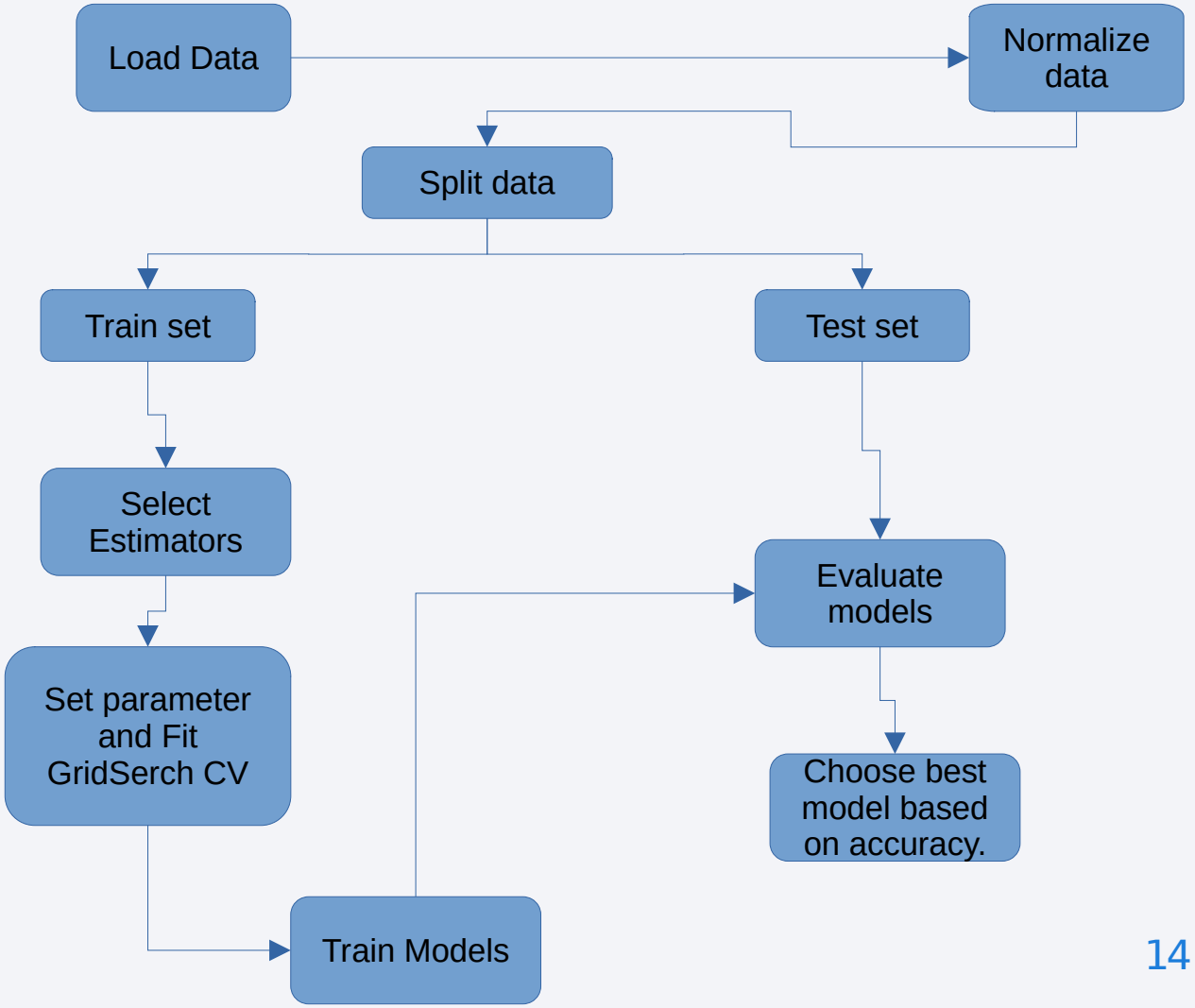
- Select model from Scikit-Learn libraries
- Set parameters for each model to GridSearchCV
- Fit GridSearchCV model to train data

Model Evaluation:

- Get best parameters for each model
- Get accuracy score on test set for each model
- Get confusion matrix for each model

Model Comparison:

- Get best model based on accuracy score



Notebook with the entire process

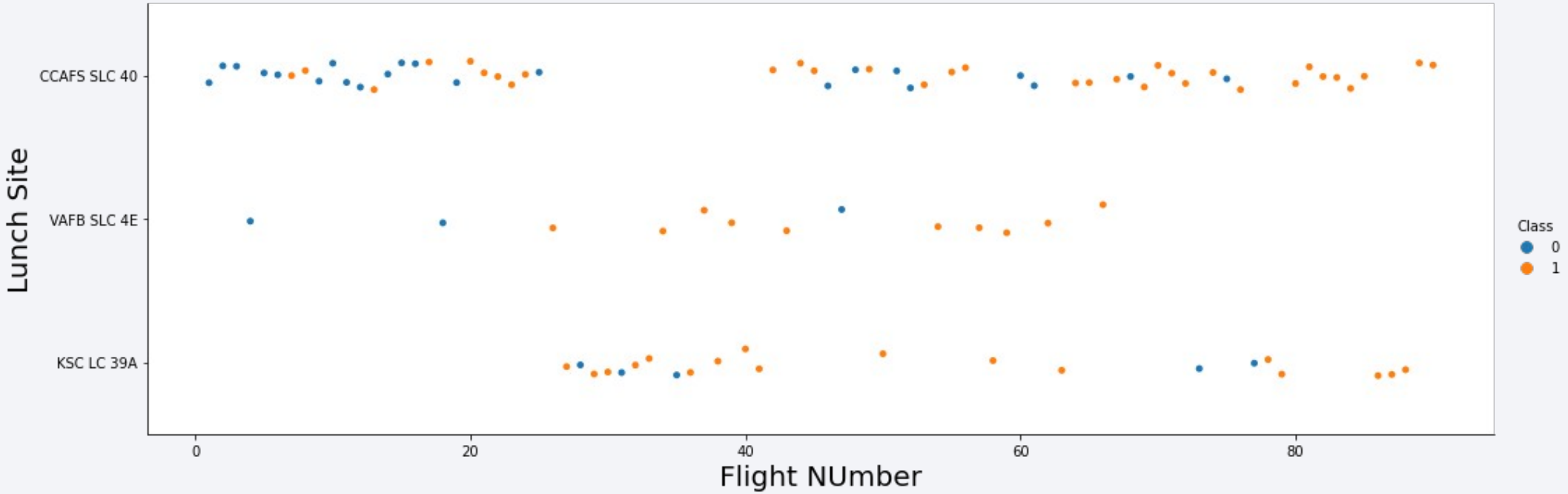
Results

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

Section 2

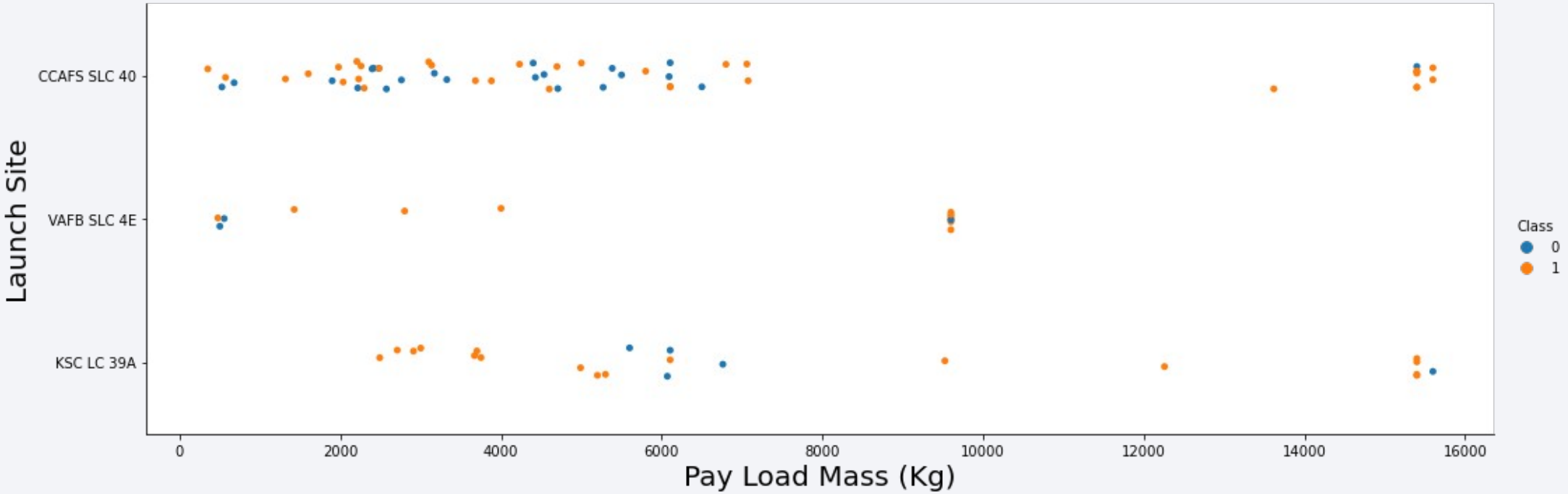
Insights drawn from EDA

Flight Number vs. Launch Site



- Strong correlation between Flight Number and outcome for each site
- Early Flight has more chances of a bad outcome, in particular for the launch site CCAFS SLC-40

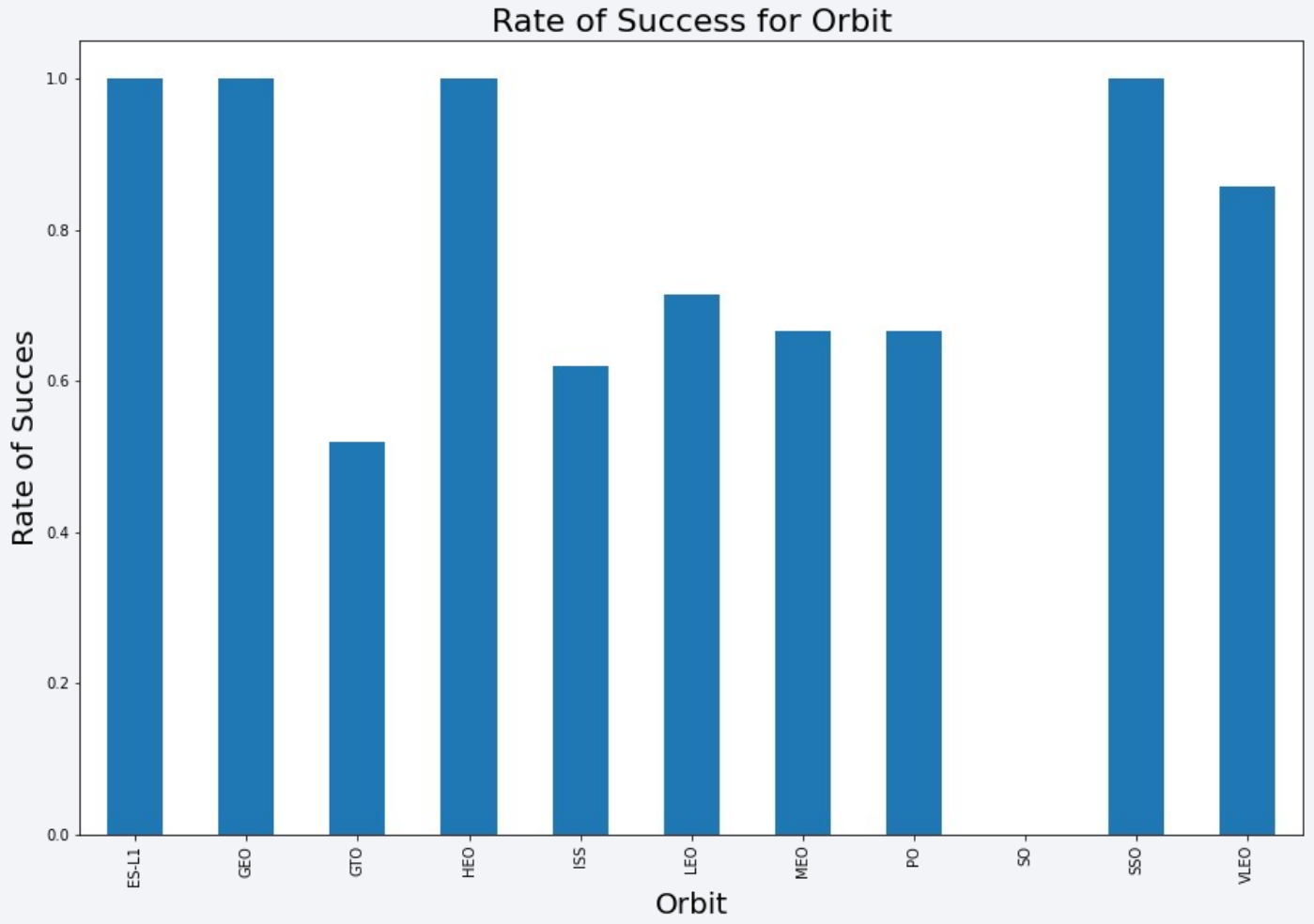
Payload vs. Launch Site



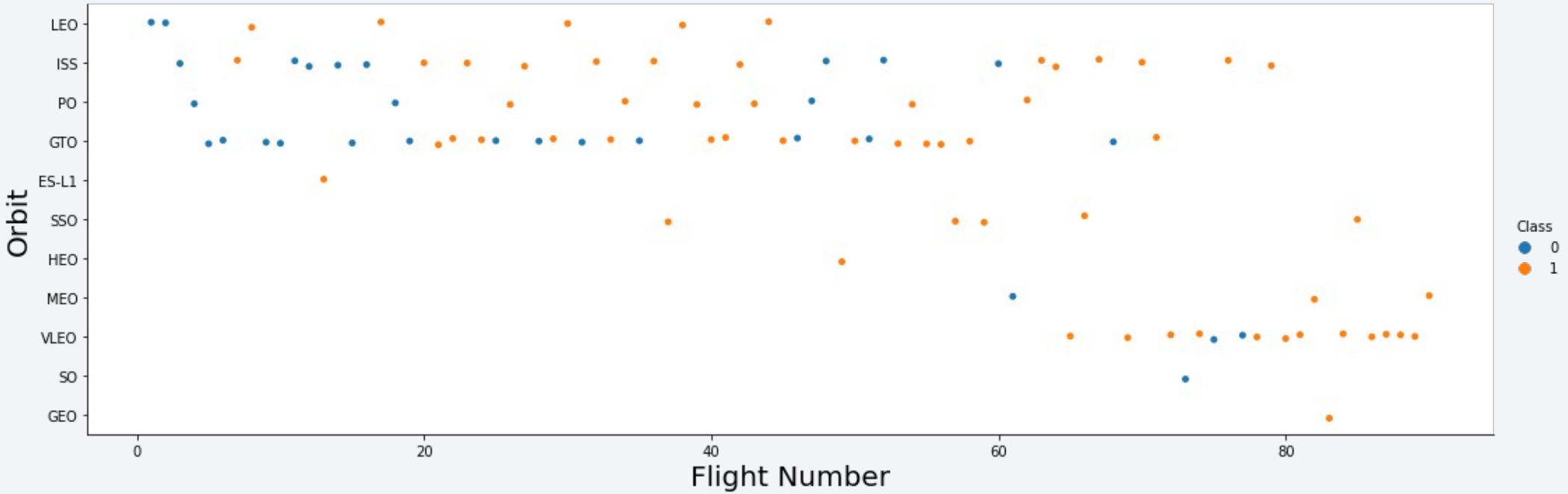
- VAFB SLC 4E has no heavy launches
- There is not clear correlation for the site CCAFS SLC 40
- KSC LC 39A has more success rate for light launches

Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, SSO orbits has the highest rate of success.
- No successful launches for SO orbits

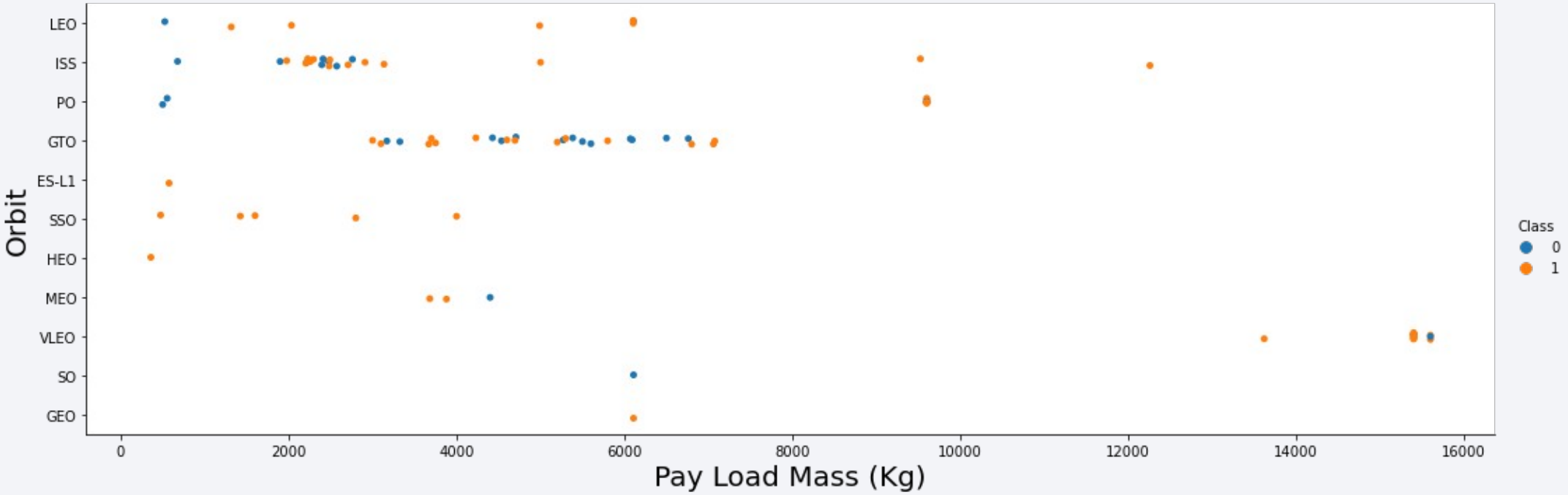


Flight Number vs. Orbit Type



- In the LEO orbit the Success appears related to the number of flights
- There seems to be no relationship between flight number when in GTO orbit

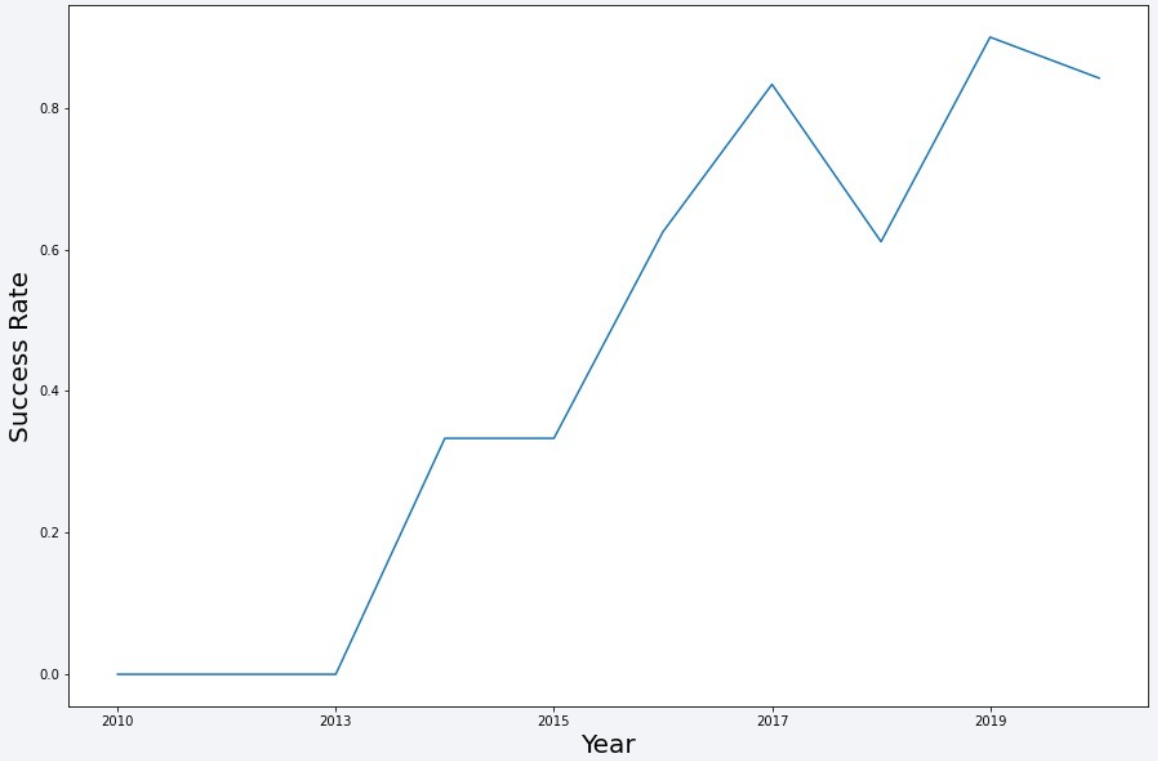
Payload vs. Orbit Type



- Heavy payload mass is more related to success for Polar, LEO, ISS
- No correlation between payload mass and outcome for GTO orbit
- No heavy launches for ES-L1, SSO, HEO, MEO

Launch Success Yearly Trend

- We can see from the graph that success rate increase from 2013 till 2020.
- Otherwise we can see some stops and slight decreasing during time.
- Positive trend



All Launch Site Names

SQL Query:

```
%%sql
SELECT unique(LAUNCH_SITE)
FROM SpaceX
```

Results:

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

Explanation:

We use UNIQUE() function on LAUNCH_SITE in SELECT statement to get a list of the distinct values form SpaceX table.

Launch Site Names Begin with 'CCA'

SQL Query:

```
%%sql
SELECT *
FROM SpaceX
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5
```

Results:

DATE	time_utc_	booster_version	launch_site	payload	payload_mass_kg_	orbit	customer	mission_outcome	landing_outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
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	Success	Failure (parachute)
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-12	22:41:00	F9 v1.1	CCAFS LC-40	SES-8	3170	GTO	SES	Success	No attempt

Explanation:

Here we use a wildcard in a WHERE clause to get all records with LAUNCH_SITE name that begin with CCA

Total Payload Mass

SQL Query:

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_)
FROM SpaceX
WHERE CUSTOMER = 'NASA (CRS)'
```

Results:

1
22007

Explanation:

With the SUM() function applied to PAYLOAD_MASS__KG_, we can get the total mass. The WHERE clause filter records to get information relative only to NASA (CRS).

Average Payload Mass by F9 v1.1

SQL Query:

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_)
FROM SpaceX
WHERE BOOSTER_VERSION LIKE 'F9 v1.1%'
```

Results:

1
3226

Explanation:

We applied the AVG() aggregation function to get the average payload mass from SpaceX table, then we filter the records with a WHERE clause and a wildcard to get the info of the Falcon 9 v1.1.

First Successful Ground Landing Date

SQL Query:

```
%%sql
SELECT MIN(DATE)
FROM SpaceX
WHERE LANDING__OUTCOME = 'Success (ground pad)'
```

Results:

1
2017-01-05

Explanation:

Here we want to find the first time where a landing on ground pad occurs. We use the MIN() function on the DATE column to get this info.

Successful Drone Ship Landing with Payload between 4000 and 6000

SQL Query:

```
%%sql
SELECT BOOSTER_VERSION
FROM (SELECT BOOSTER_VERSION, PAYLOAD_MASS__KG_
      FROM SpaceX
      WHERE LANDING__OUTCOME = 'Success (drone ship)')
WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
```

Results:

booster_version
F9 FT B1022
F9 FT B1031.2

Explanation:

Here we use a sub-query. Since we want to find only the booster version, first of all we filtered the results using a sub-query to get booster version with LANDING_OUTCOME that is a Success on a drone ship. After that we use the filtered information to get the booster version with payload mass between 4000 kg and 6000 kg.

Total Number of Successful and Failure Mission Outcomes

SQL Query:

```
%%sql
SELECT (SELECT COUNT(MISSION_OUTCOME)
        FROM SpaceX
        WHERE MISSION_OUTCOME LIKE 'Success%') AS Success_Landing,
    (SELECT COUNT(MISSION_OUTCOME)
    FROM SpaceX
    WHERE MISSION_OUTCOME LIKE 'Fail%') AS Fail_Landing
FROM SpaceX
LIMIT 1
```

Results:

success_landing	fail_landing
45	0

Explanation:

Here we used a sub-query directly in a SELECT statement and used aliases to get the info.

Boosters Carried Maximum Payload

SQL Query:

```
%%sql
SELECT BOOSTER_VERSION
FROM SpaceX
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_)
      FROM SpaceX)
```

Results:

booster_version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3

Explanation:

To find this information we used a sub-query in a WHERE clause, because we cannot pass aggregation functions in this type of clause.

2015 Launch Records

SQL Query:

```
%%sql
SELECT BOOSTER_VERSION, LAUNCH_SITE, LANDING__OUTCOME
FROM SpaceX
WHERE (LANDING__OUTCOME = 'Failure (drone ship)' AND YEAR(DATE) = 2015)
```

Results:

booster_version	launch_site	landing__outcome
F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)

Explanation:

Here we combined two conditions in a WHERE clause with an AND logical operator, to get records that satisfy both conditions.

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

SQL Query:

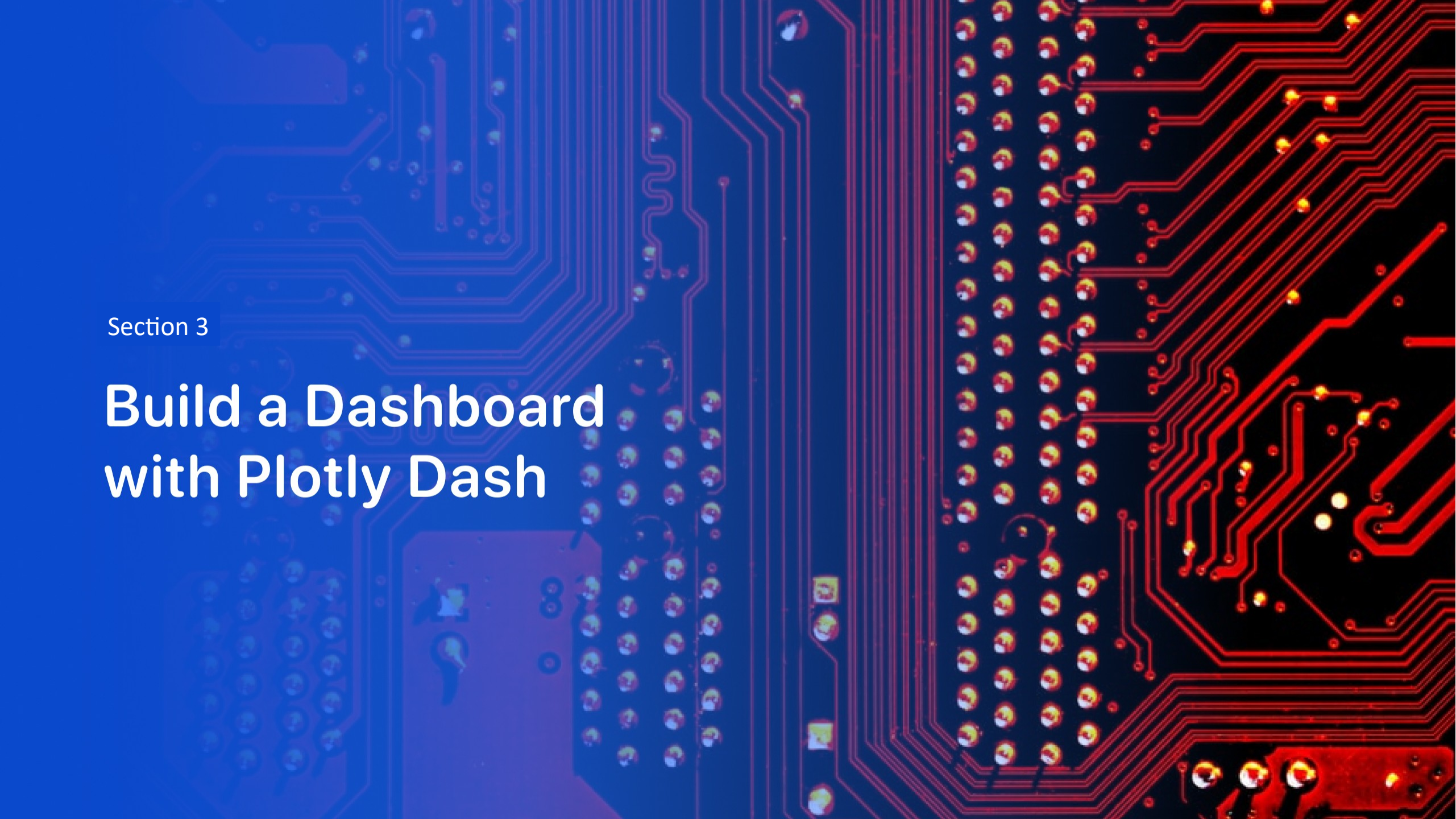
```
%%sql
SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS COUNT
FROM SpaceX
WHERE (DATE BETWEEN '2010-06-04' AND '2017-03-20')
GROUP BY LANDING__OUTCOME
```

Results:

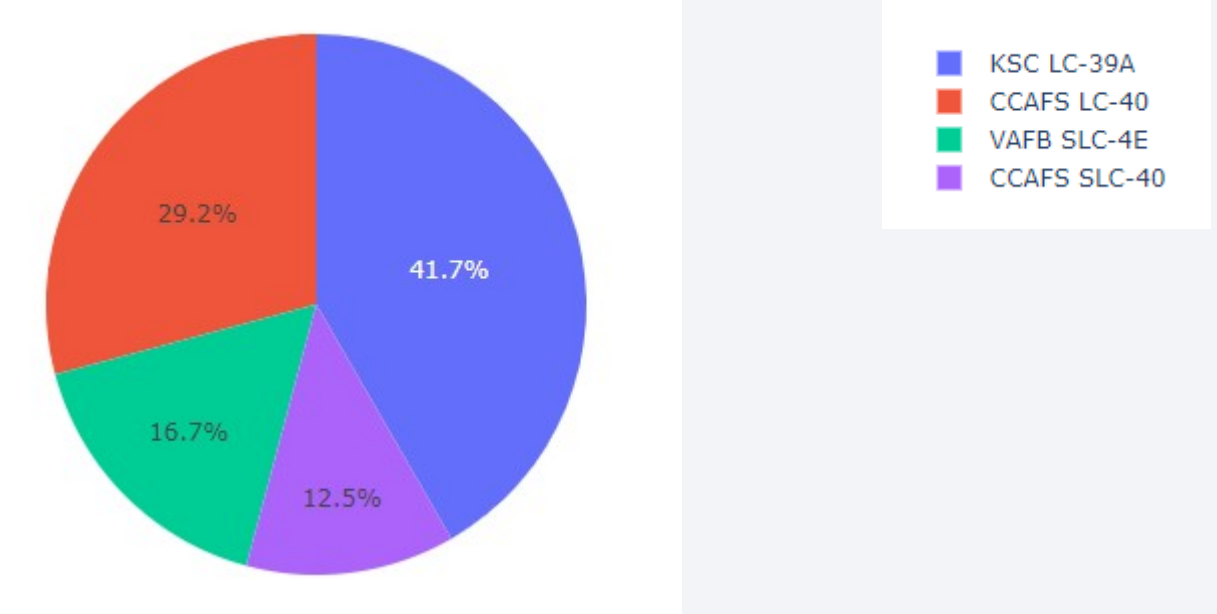
landing__outcome	COUNT
Controlled (ocean)	1
Failure (drone ship)	2
Failure (parachute)	1
No attempt	7
Success (drone ship)	2
Success (ground pad)	2

Explanation:

To obtain the total of landing outcome for each class we use the COUNT() function with the GRUOP BY clause to group the results for every possible landing outcome.

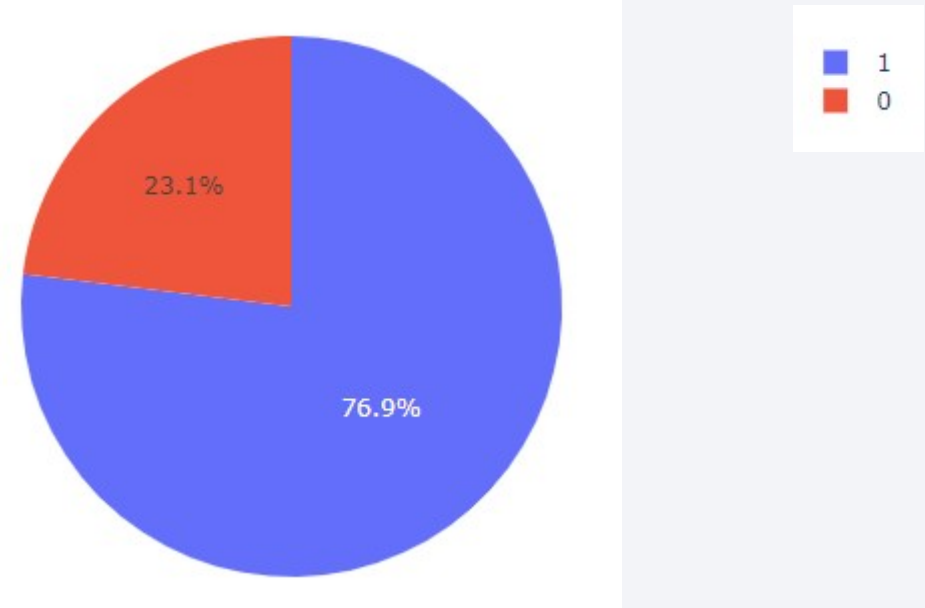

Section 3
Build a Dashboard with Plotly Dash

Launch Success for all Sites



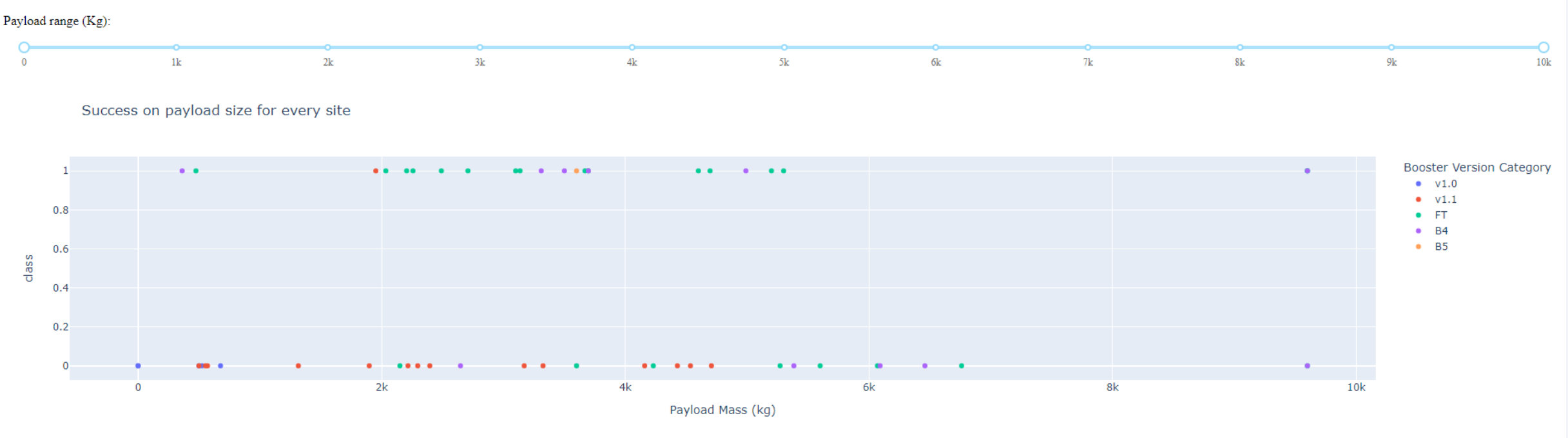
The chart shows launch success for all the site from the dashboard we have created. As we can see the majority of launches are from KSC LC-39A.

Success rate for KSC LC-39A



The chart show the success rate for the launch site KSC LC-39A. As we can see the blue slice represent the success of a launch and the red slice represent the failure. As shown before this site has the highest success rate.

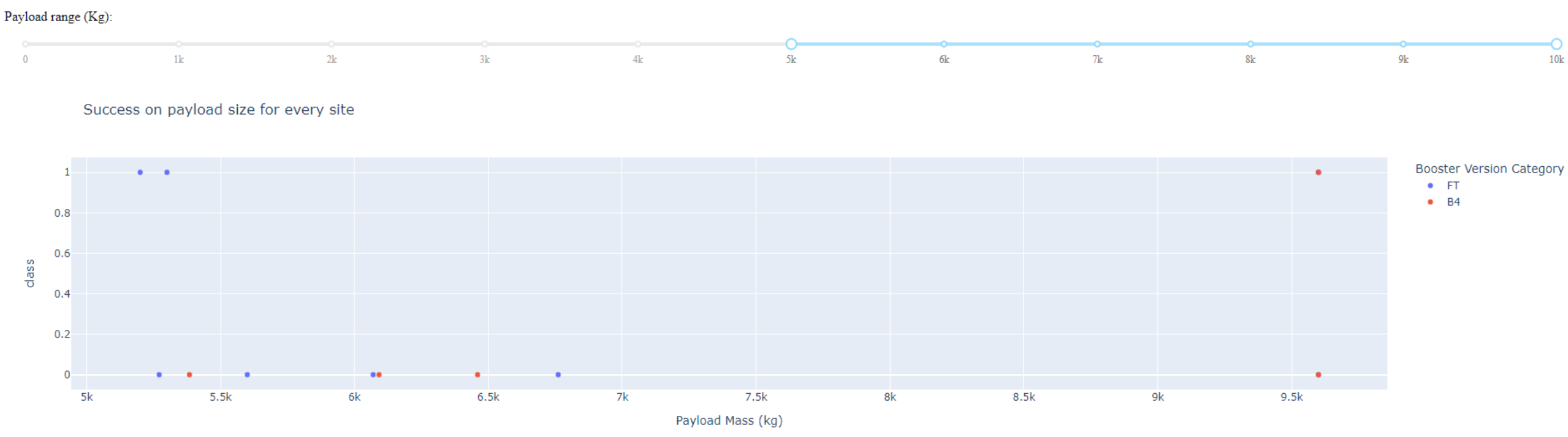
Payload vs. Launch Outcome (full range mass)



Here are shown the outcome for payload mass carried by different boosters. FT booster has a good rate of success if the payload is quite light, otherwise we can see that for heavy payloads outcome is more like to be a failure.

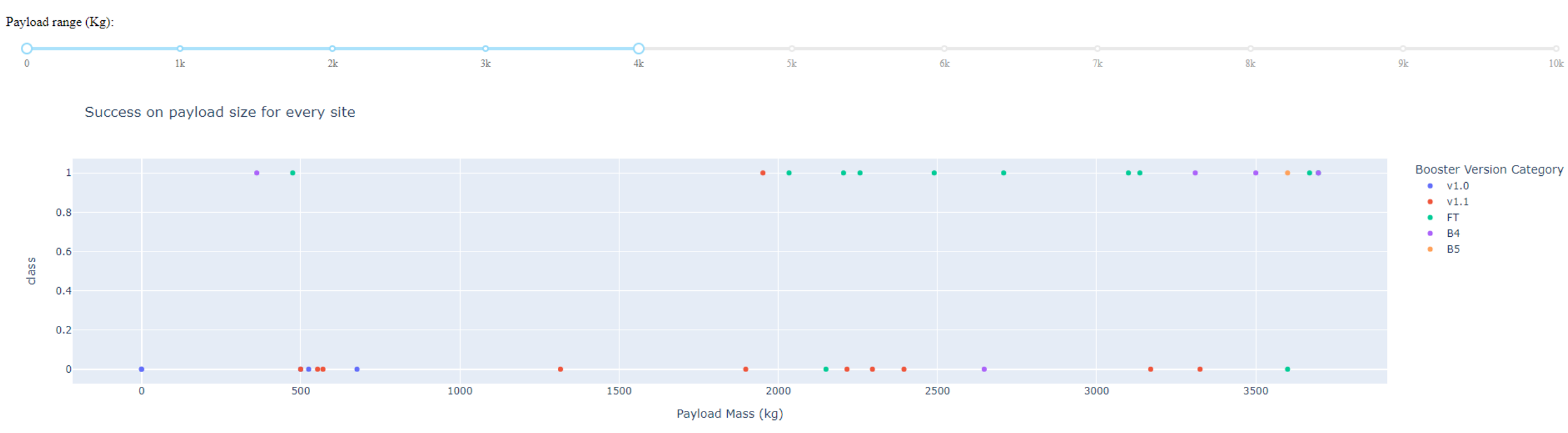
Notice that we don’t have many information about heavy payloads.

Payload vs. Launch Outcome (large masses)



We can see that we don’t have many heavy payload launches but FT booster has more chances to success for masses between 5000 kg and 5500 kg. On the other hand B4 booster has more success with heavy payloads.

Payload vs. Launch Outcome for small masses



We can see here that small payloads have more chances of success if they are carried by FT booster.

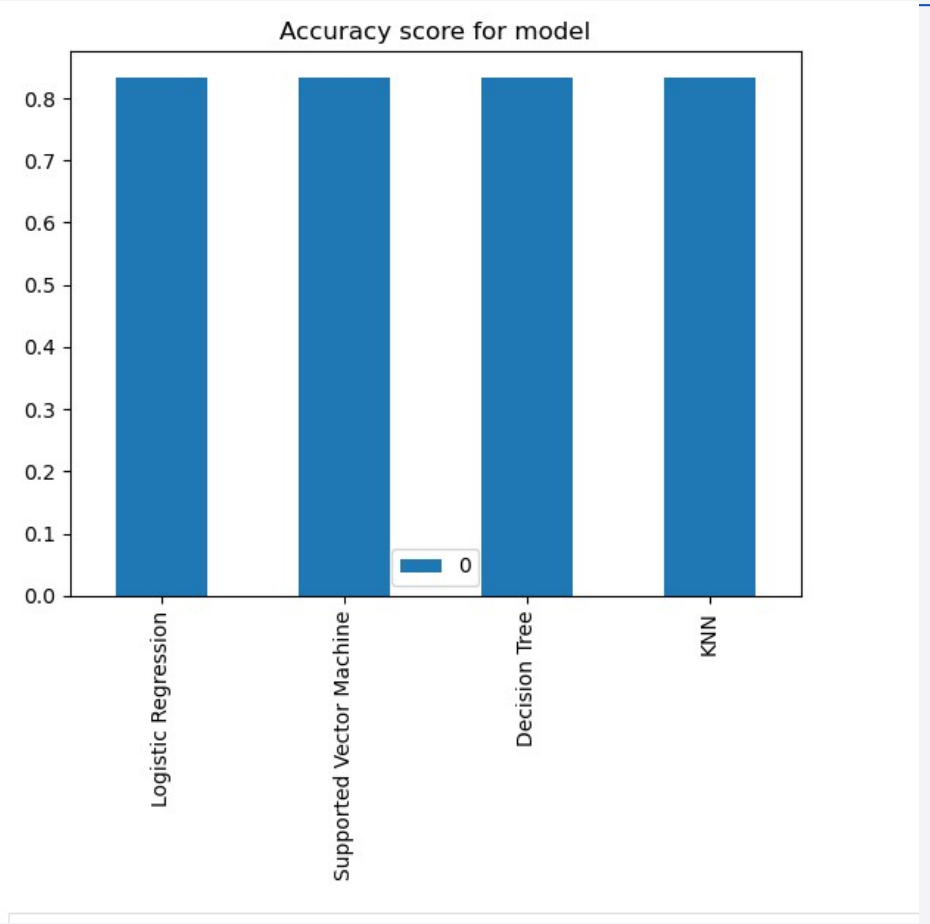
Section 4

Predictive Analysis (Classification)

Classification Accuracy

We cannot see any difference in accuracy from testing the model on test set: each model has the same accuracy score.

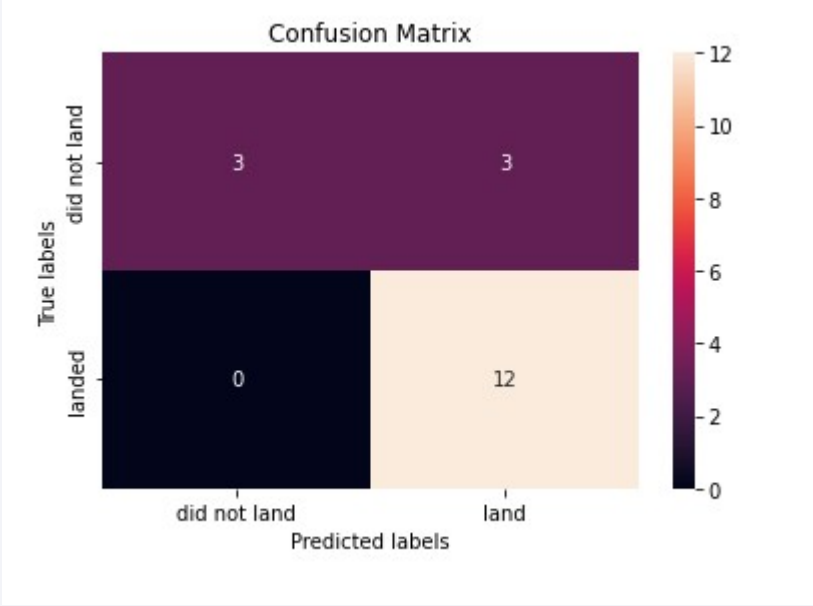
We choose Decision Tree because it performs better on train set.



Confusion Matrix

Confusion matrix for Decision Tree model:

- Predicted label vs. real label
- No false negative predicted
- Only 3 false positive predicted



Conclusions

- Many factors affect the outcome: Booste type, payload mass, launch site, orbit
- Number of previous launches affect the outcome: gaining knowledge from previous launches.
- Orbit type has a high impact on the outcome: GEO, HEO, SSO, ES L1 have the highest rate of success.
- Depending on orbit, payload mass can affect outcome: certain masses provide an higher rate of success on certain orbit
- We cannot say why some launch site are better than other: more data are necessary.
- Decision tree performs very well on this dataset.

Thank you!