Winning Space Race with Data Science

Michele Mogos
03/07/2022

Outline

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

Executive Summary

- *Summary of methodologies*
  - Data Collection (API, Web Scrapping)
  - Data Wrangling
  - Expl. Data Analysis (EDA) – Data Visualization & SQL
  - Interactive maps (Folium) & Dashboards (Plotly Dash)
  - Predictive Analysis
- *Summary of all results*
  - EDA results
  - Interactive maps (Folium) & Dashboards (Plotly Dash)
  - Predictive Analysis

Introduction

Project background and context

Falcon9 is a partially reusable medium-lift launch vehicle produced by SpaceX. The rocket has two stages, with the first stage capable of landing vertically. SpaceX claims that a Falcon9 launch cost 62M USD, while the cost claimed by the competition (traditional rockets) is 165M USD.

Problems you want to find answers

Can the rocket 1st stage re-usability be predicted and used for cost estimation purposes?
What is a successful/failed landing? What are the main characteristics?
What are the effects of these factors on the outcome?
What conditions influence the successful landing rate?

Section 1

Methodology

Methodology

Executive Summary

Data collection methodology:

SpaceX REST API, Web Scrapping (Wikipedia)

Perform data wrangling

Non-relevant data excluded, 1Hot encoding for models classification

Perform expl. data analysis (EDA) using visualization and SQL)

Perform interactive visual analytics using Folium and Plotly Dash

Perform predictive analysis using classification models

How to build, adjust and assess classification models

Data Collection

- The needful data were collected through API and Web scrapping
  - Key info retrieved via API: rocket, launchpad, payloads, cores, etc.

SpaceX REST API Calls → JSON file → DF creation → DF Export

  - Key info retrieved via Web Scrapping: launches, payload

HTML connection → Data Extraction (BeautifulSoup) → DF creation → DF Export

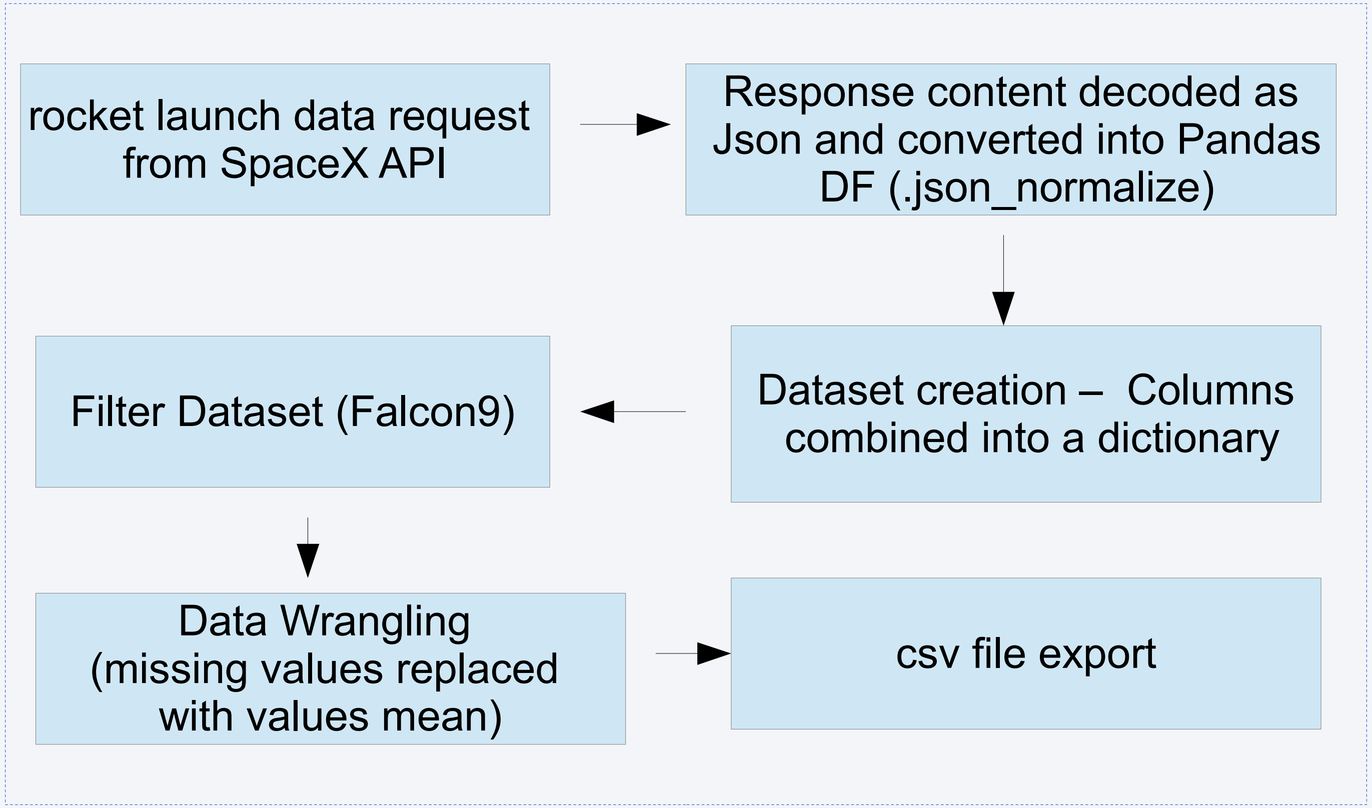
Data Sources:

https://api.spacexdata.com/v4

https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Data Collection – SpaceX API

- The following key data were collected:
- *Rocket*: name of the booster
- *Launchpad*: name/location of the launch site
- *Payload*: the payload mass and the target orbit
- *Cores*: landing outcome, type of landing, etc..



https://github.com/MMG2022/Capstone_Project/blob/main/Capstone_Data-collection_api.ipynb

Data Collection - Scraping

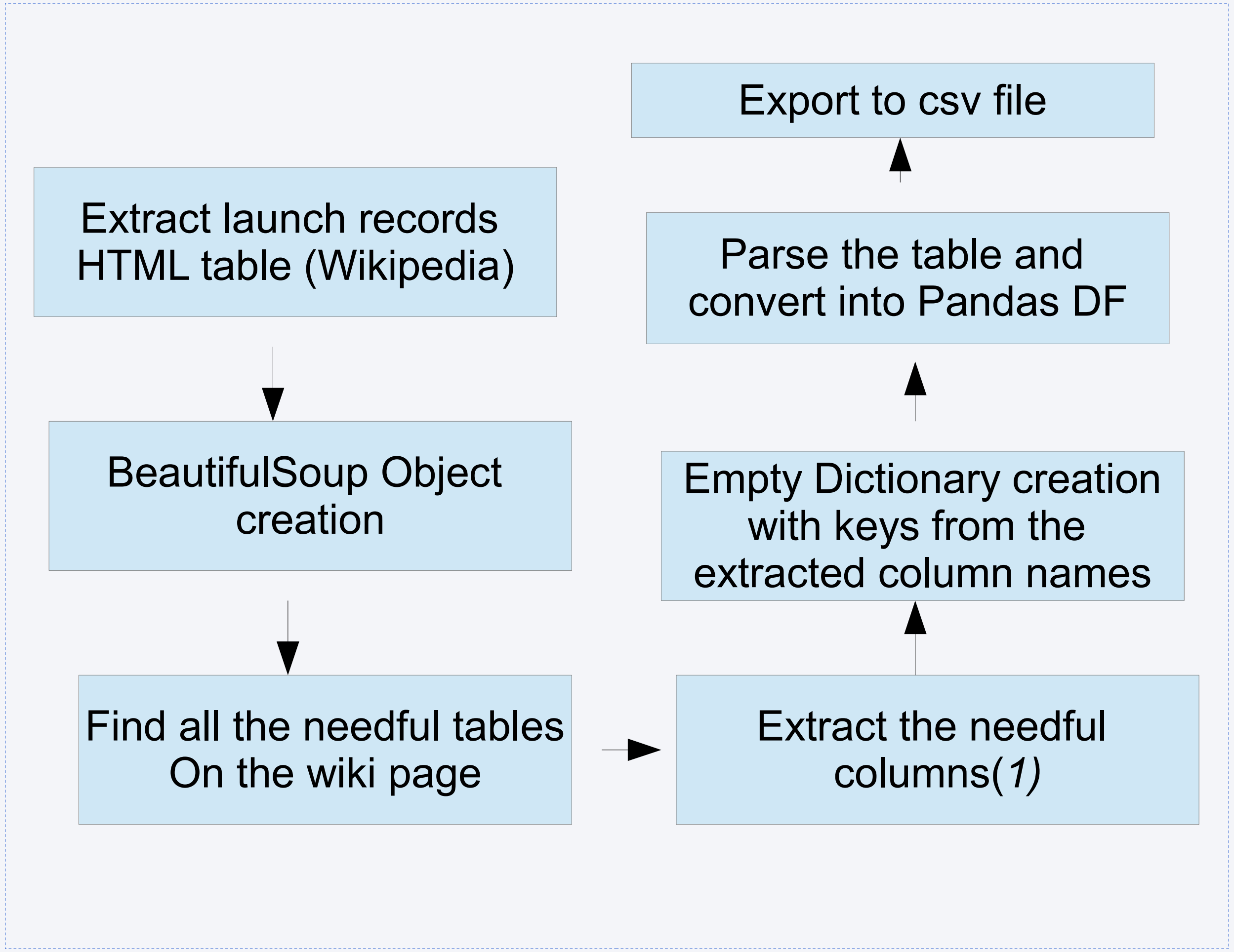
The following key data were collected *(1)*:

Flight No., Launch site, Payload, Payload mass, Orbit, Customer, Launch outcome

The following key data were also included:

Version Booster, Booster landing, Date,

https://github.com/MMG2022/Capstone_Project/blob/main/Capstone_WebScrapping.ipynb

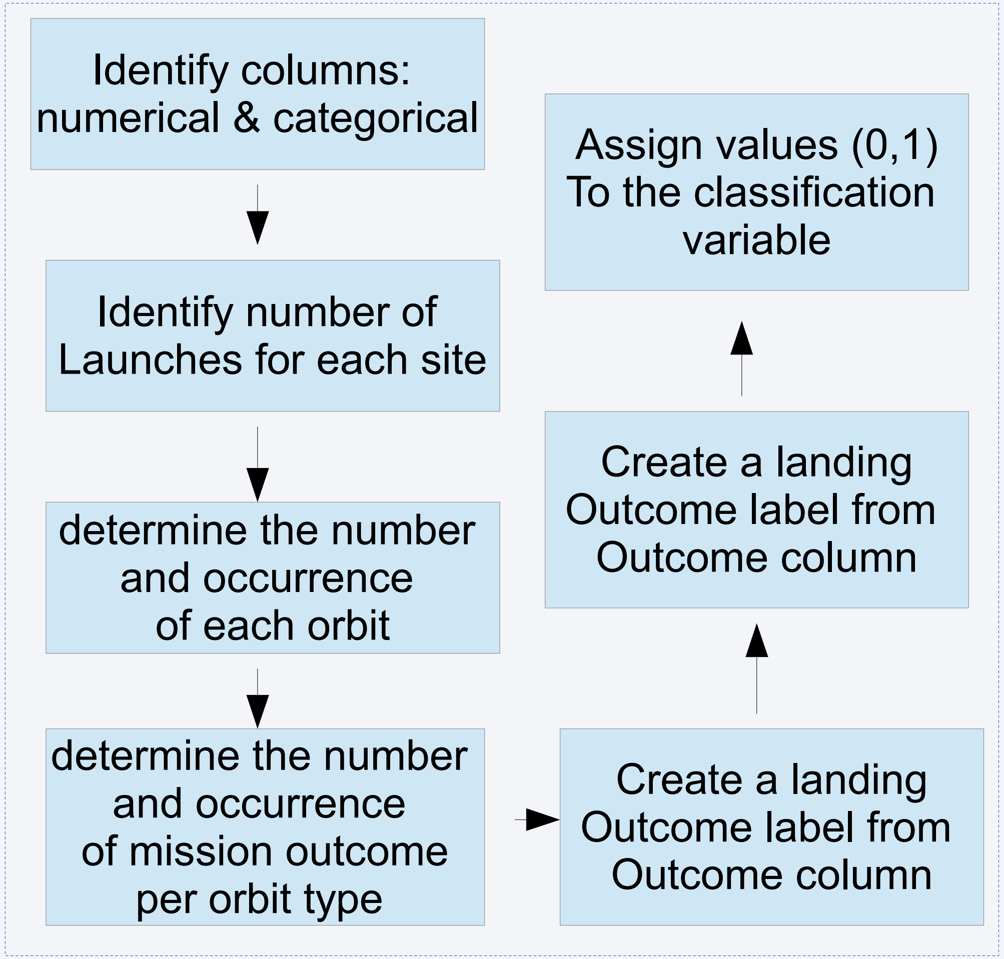


Data Wrangling

There are cases where the booster did *not* land successfully

- Successful Mission: *True Ocean, True RTLS, True ASDS*
- Failed Mission: *False Ocean, False RTLS, False ASDS*

- *Data Transformation (string var. Categorical var.):*
  - *1=Success, 0=Failure*



https://github.com/MMG2022/Capstone_Project/blob/main/Capstone_Data_Wrangling.ipynb

EDA with Data Visualization

Scatter plots: show relationships/correlation between variables

Flight # Vs. Payload Mass	**Payload Mass Vs. Launch Site**	**Payload Vs. Orbit type**
Flight # Vs. Launch Site	**Orbit Vs. Flight Number**	

Bar Graph: show relationships between numerical/categorical variables.

Success rate Vs. Orbit

Line Graphs: show data trends and behaviour

Success rate Vs. Year

https://github.com/MMG2022/Capstone_Project/blob/main/Capstone_EDA_Visualization.ipynb

EDA with SQL

- The following queries were executed
- 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 successful landing outcome in ground pad was achieved
- List the names of the boosters which have success in drone ship and have payload mass > 4000 but < 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. Use a sub-query
- 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 successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

https://github.com/MMG2022/Capstone_Project/blob/main/Capstone_EDA_SQL_2.ipynb

Build an Interactive Map with Folium

- The NASA Johnson Space Center (Huston, Texas) is represented in the Folium Map.
- Folium maps offer visuals that helps to better understand the launch sites locations, their surroundings and the number of successful/unsuccessful missions.
- The following elements can be observed:
  - Red circle and label @NASA JSC’s coordinate
  - Red circle and label at each launch site coordinate
  - Clusters to show common info
  - Markers showing successful (green) and unsuccessful (red) landings
  - Markers showing distance between launch site and selected locations

Build a Dashboard with Plotly Dash

- The following components were included in the Dashboard:
  - *Drop-down*: to select the launch site/s
  - *Pie Chart*: to show the successes/failures for the selected site/s
  - *Range slider*: payload mass selection (per-defined range)
  - *Scatter plot*: to show the relationship/correlation between variables

https://github.com/MMG2022/Capstone_Project/blob/main/Capstone_Dashboard_SpaceX.ipynb

Predictive Analysis (Classification)

The best classification model was identified in the following way:

- *Data preparation*:
  - Data-set load, Data Normalization, Data-set split (Train, Test)
- *Model preparation*:
  - Machine Learning algorithm selection, parameters settings to GrindSearchCV, GrindSearch Model training
- *Model assessment*:
  - Accuracy computation for each model, Confusion Matrix plotting
- *Model comparison*:
  - Accuracy of the different models comparison, selection of the model with the best accuracy

- https://github.com/MMG2022/Capstone_Project/blob/main/Capstone_ML.ipynb

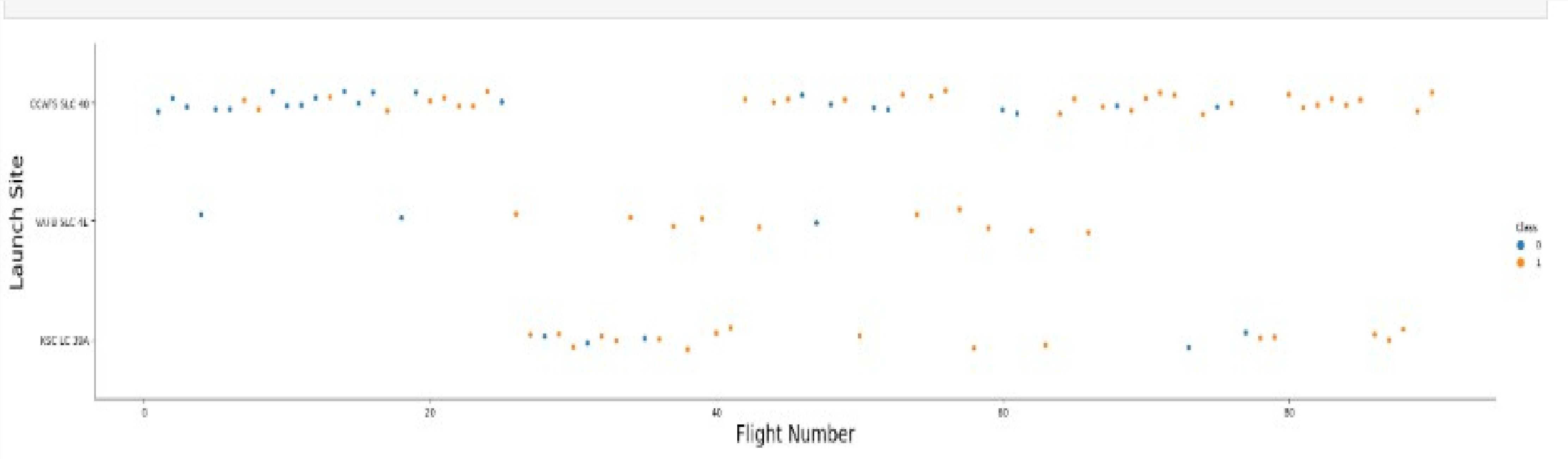
Results

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

Section 2

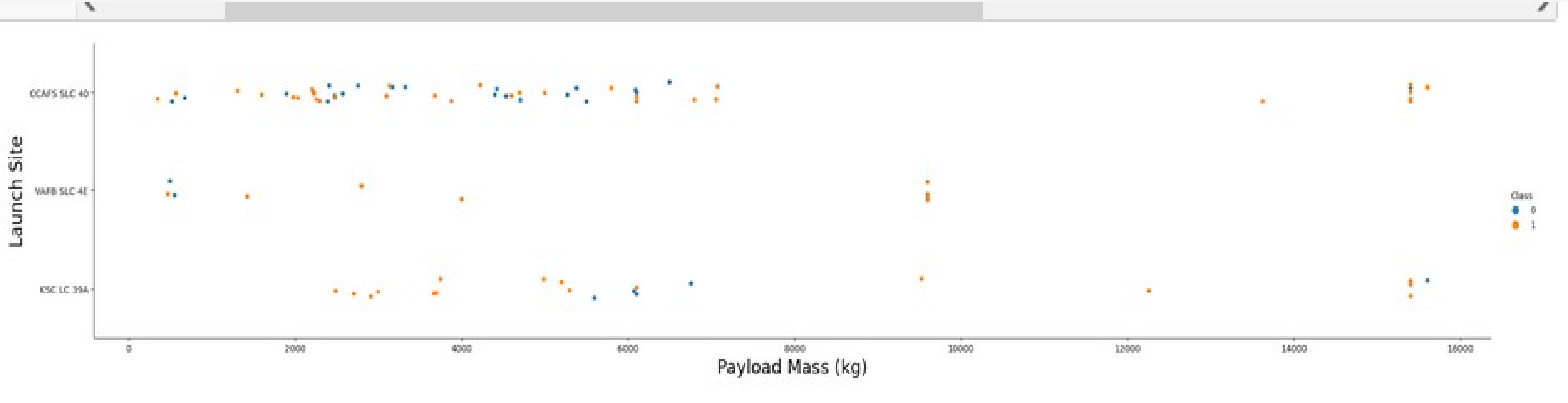
Insights drawn from EDA

Flight Number vs. Launch Site



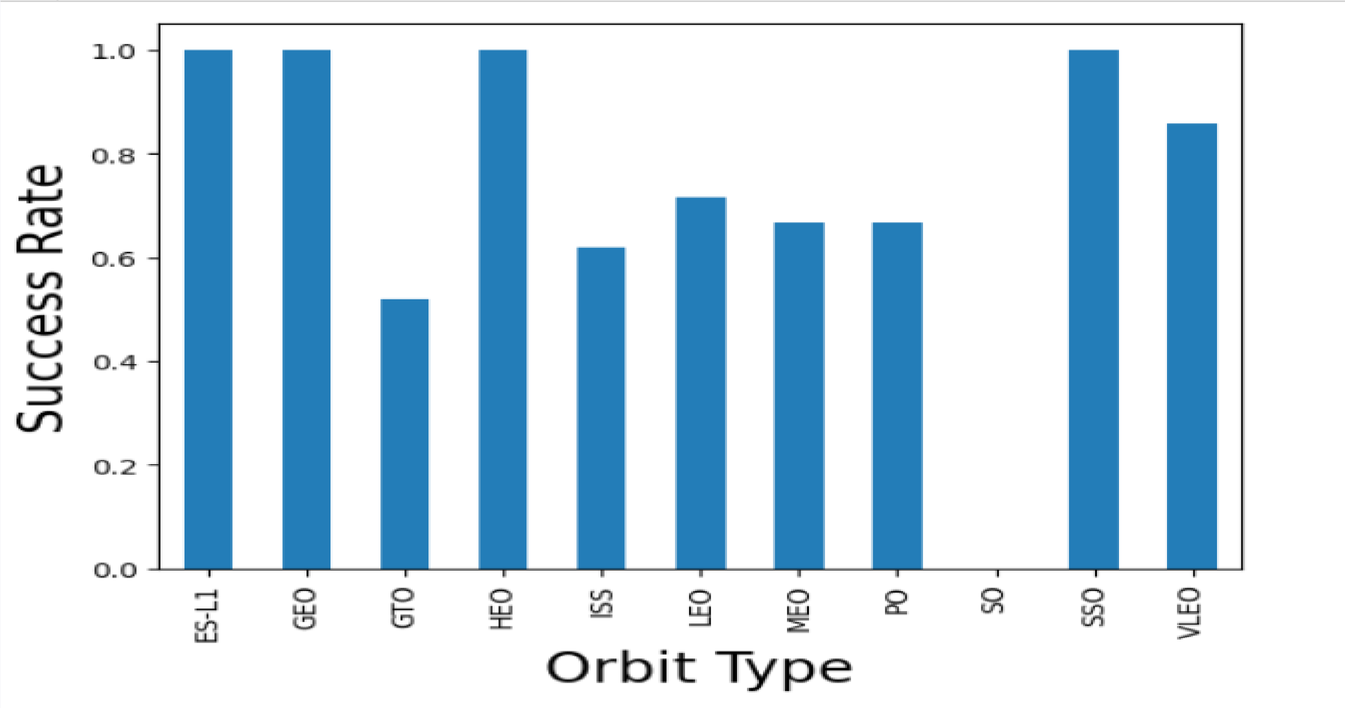
- From this graph we can note that the successful missions are positively correlated with the number of flight

Payload vs. Launch Site



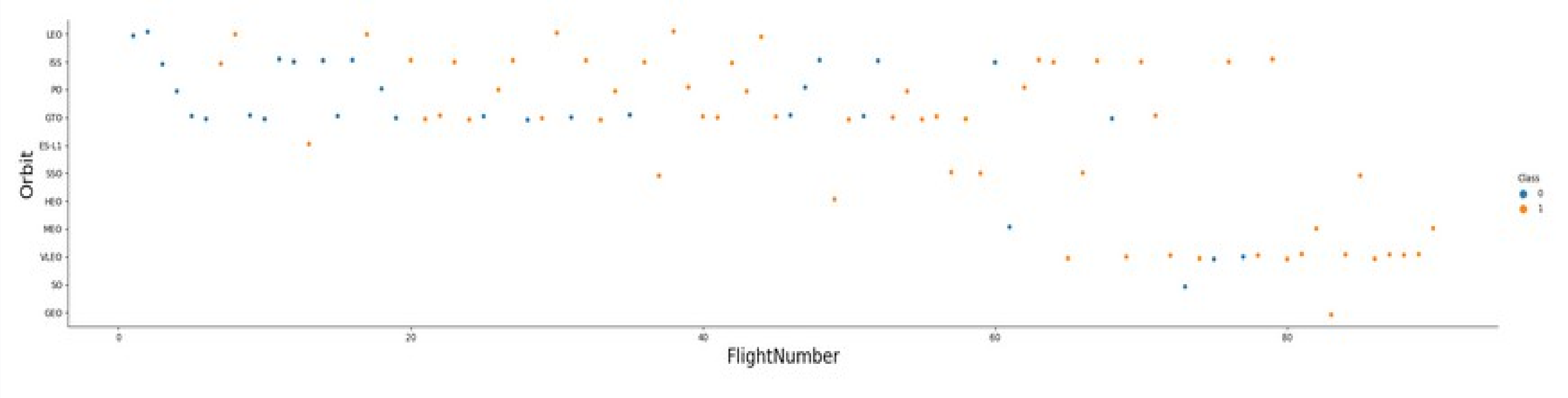
This graph seems to suggest that an heavy payload is correlated with successful missions at certain sites but not everywhere

Success Rate vs. Orbit Type



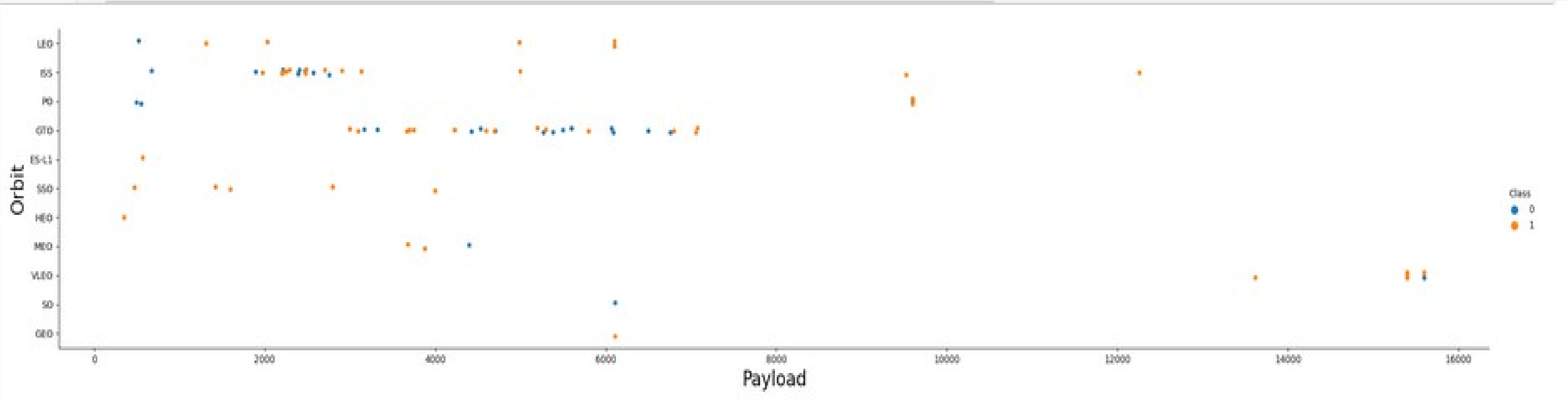
From this bar graph we can appreciate the relationship between success rate and orbit type. ES-L1, GEO, HEO and SSO show the best success rate

Flight Number vs. Orbit Type



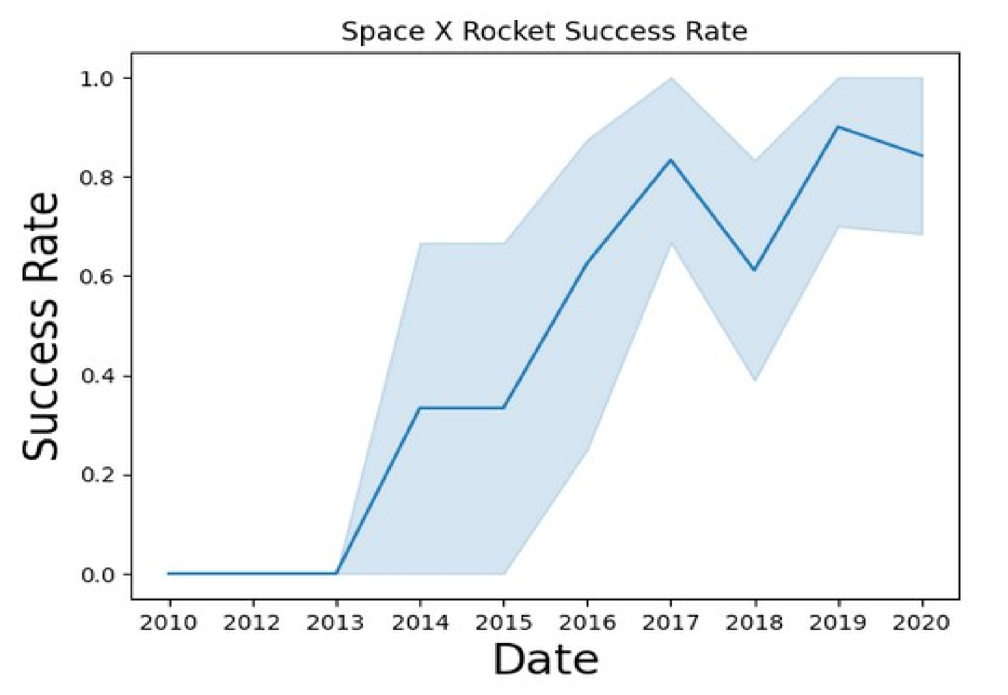
From this graph we can note a different degree of correlation between successful missions and number of flights for the different Orbits. For instance, while LEO shows a strong positive correlation, GTO does not show signs of correlation.

Payload vs. Orbit Type



From this graph we can note a different degree of correlation between successful missions and payload for the different Orbits. For instance, while LEO and ISS shows a positive correlation, GTO and the rest does not show significant signs of correlation.

Launch Success Yearly Trend



- From this graph we can observe a strong positive trend of the success rate as of 2013 with just one significant setback in 2018

All Launch Site Names

Display the names of the unique launch sites in the space mission

In [7]:

```
%%sql
SELECT DISTINCT LAUNCH_SITE
FROM SPACEXTBL;
```

```
 * sqlite:///my_data1.db
Done.
```

Out[7]:

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

- Using "DISTINCT" in the query results in the list of unique values for the selected field

Launch Site Names Begin with 'CCA'

Display 5 records where launch sites begin with the string 'CCA'

In [8]:

```
%%sql
SELECT LAUNCH_SITE
FROM SPACEXTBL
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5;
```

```
 * sqlite:///my_data1.db
Done.
```

Out[8]:

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Using “WHERE” followed by “LIKE” in the query results in the list of records containing the indicated string (CCA). “LIMIT” was used to limit the result to 5 records

Total Payload Mass

Display the total payload mass carried by boosters launched by NASA (CRS)

In [9]:

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_)
FROM SPACEXTBL
WHERE Customer = 'NASA (CRS)';
```

```
 * sqlite:///my_data1.db
Done.
```

Out[9]:

SUM(PAYLOAD_MASS__KG_)
45596

- Query used to display the the “PAYLOAD_MASS” of a specific customer: “NASA (CRS)”

Average Payload Mass by F9 v1.1

Display average payload mass carried by booster version F9 v1.1

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

```
 * sqlite:///my_data1.db
Done.
```

AVG(PAYLOAD_MASS__KG_)
2928.4

Query used to display the the average “PAYLOAD_MASS” of a specific booster: “F9 v1.1”

First Successful Ground Landing Date

```
[23]: %sql SELECT MIN("DATE") FROM SPACEXTBL WHERE "Landing _Outcome" = 'Success (ground pad)';

 * sqlite:///my_data1.db
Done.

[23]: MIN("DATE")
      01-05-2017
```

- Query used to display the date of the first successful landing outcome (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

In [14]:
```
%sql SELECT "BOOSTER_VERSION" FROM SPACEXTBL WHERE "LANDING _OUTCOME" = 'Success (drone ship)' \
AND "PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MASS__KG_" < 6000;
```

```
 * sqlite:///my_data1.db
Done.
```

Out[14]:

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

- Query used to display the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

In [15]:
```
%sql SELECT (SELECT COUNT("MISSION_OUTCOME") FROM SPACEXTBL WHERE "MISSION_OUTCOME" LIKE '%Success%') AS SUCCESS, \
(SELECT COUNT("MISSION_OUTCOME") FROM SPACEXTBL WHERE "MISSION_OUTCOME" LIKE '%Failure%') AS FAILURE
```

```
 * sqlite:///my_data1.db
Done.
```

Out[15]:

SUCCESS	FAILURE
100	1

- Query used to Calculate the total number of successful and failure mission outcomes.
  - The 1st sub-query counts the successful missions; the 2nd sub-query counts the unsuccessful missions.
  - "WHERE" followed by "LIKE" is used to filter the missions outcome
  - "COUNT" is used to count the filtered records

Boosters Carried Maximum Payload

Task 8

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

In [16]:

```
%sql SELECT DISTINCT "BOOSTER_VERSION" FROM SPACEXTBL \
WHERE "PAYLOAD_MASS__KG_" = (SELECT max("PAYLOAD_MASS__KG_") FROM SPACEXTBL)
```

```
 * sqlite:///my_data1.db
Done.
```

Out[16]:

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

- Query used to display the booster versions with the max payload mass.
  - A sub-query was used to show the max Payload Mass
  - “DISTINCT” was used to show only unique values(booster version) with the max payload mass

2015 Launch Records

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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

[17]:

```
%sql SELECT substr("DATE", 4, 2) AS MONTH, "BOOSTER_VERSION", "LAUNCH_SITE" FROM SPACEXTBL\
WHERE "LANDING _OUTCOME" = 'Failure (drone ship)' and substr("DATE",7,4) = '2015'
```

```
 * sqlite:///my_data1.db
Done.
```

ut[17]:

MONTH	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

- Query used to List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - SQL Lite does not support month names and therefore “SUBSTR” is used to elaborate the date

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

All Landings

```
%sql SELECT "LANDING _OUTCOME", COUNT("LANDING _OUTCOME") FROM SPACEXTBL\
WHERE "DATE" >= '04-06-2010' and "DATE" <= '20-03-2017' \
GROUP BY "LANDING _OUTCOME" \
ORDER BY COUNT("LANDING _OUTCOME") DESC ;
```

```
 * sqlite:///my_data1.db
Done.
```

Landing _Outcome	COUNT("LANDING _OUTCOME")
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

Successful Landings

Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

In [18]:

```
%sql SELECT "LANDING _OUTCOME", COUNT("LANDING _OUTCOME") FROM SPACEXTBL\
WHERE "DATE" >= '04-06-2010' and "DATE" <= '20-03-2017' and "LANDING _OUTCOME" LIKE '%Success%'\
GROUP BY "LANDING _OUTCOME" \
ORDER BY COUNT("LANDING _OUTCOME") DESC ;
```

```
 * sqlite:///my_data1.db
Done.
```

Out[18]:

Landing _Outcome	COUNT("LANDING _OUTCOME")
Success	20
Success (drone ship)	8
Success (ground pad)	6

Query used to 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.

- “GROUP BY” is used to group results by landing outcome
- “ORDER BY” and “DESCEND” are used to show the results in a descending order

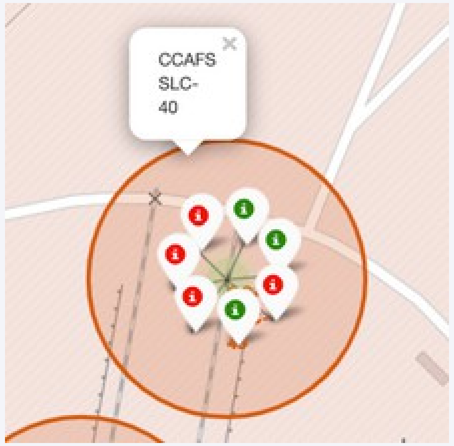
Section 3

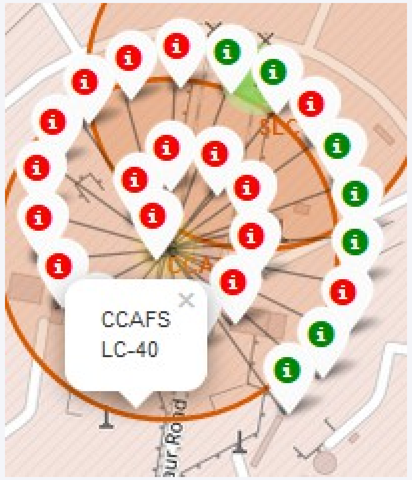
Launch Sites Proximities Analysis

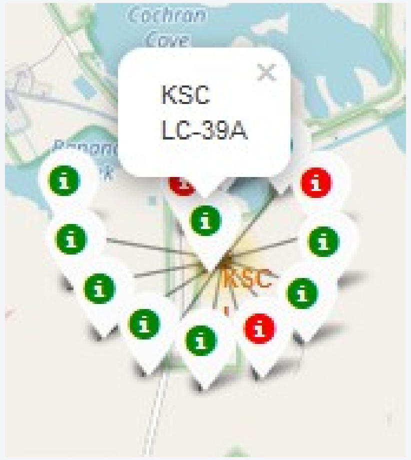
LAUNCH SITES – FOLIUM GLOBAL MAP

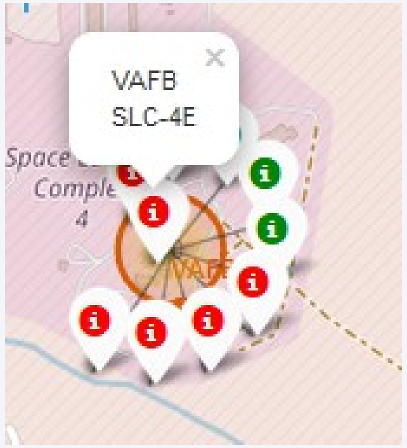
- The launch sites plotted on a global map shows their location on the coasts of the USA

Color Labelled Outcome Markers



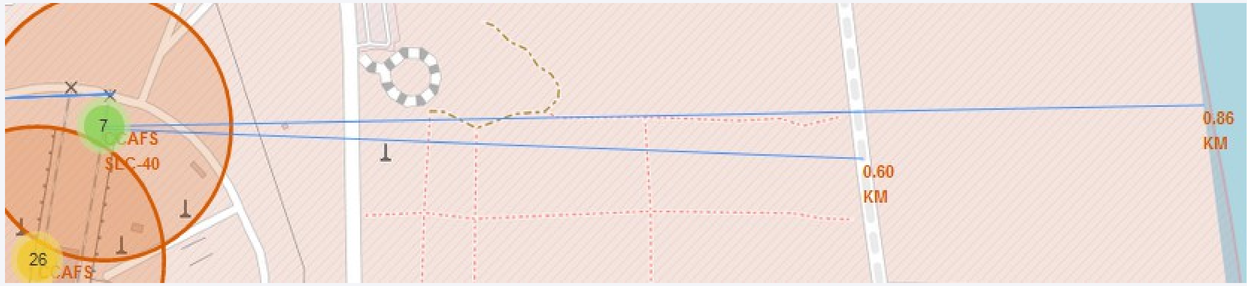






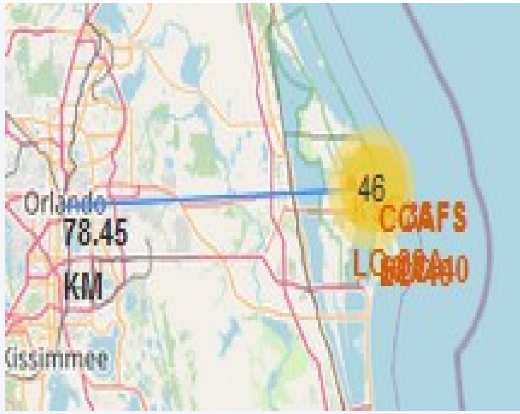
- Green markers indicates successful launches while red markers indicates unsuccessful launches.
- KSC LC-29A appear to be the site with the highest success rate

CCAFS SLC-40 Proximities




```
distance_highway = calculate_distance(launch_site_lat, launch_site_
print('distance_highway =',distance_highway, ' km')
distance_railroad = calculate_distance(launch_site_lat, launch_site_
print('distance_railroad =',distance_railroad, ' km')
distance_city = calculate_distance(launch_site_lat, launch_site_lon
print('distance_city =',distance_city, ' km')
```

```
distance_highway = 0.5834695366934144  km
distance_railroad = 1.2845344718142522  km
distance_city = 51.43416999517233  km
```




```
distance_highway = 0.5834695366934144  km
distance_railroad = 1.2845344718142522  km
distance_city = 51.43416999517233  km
```

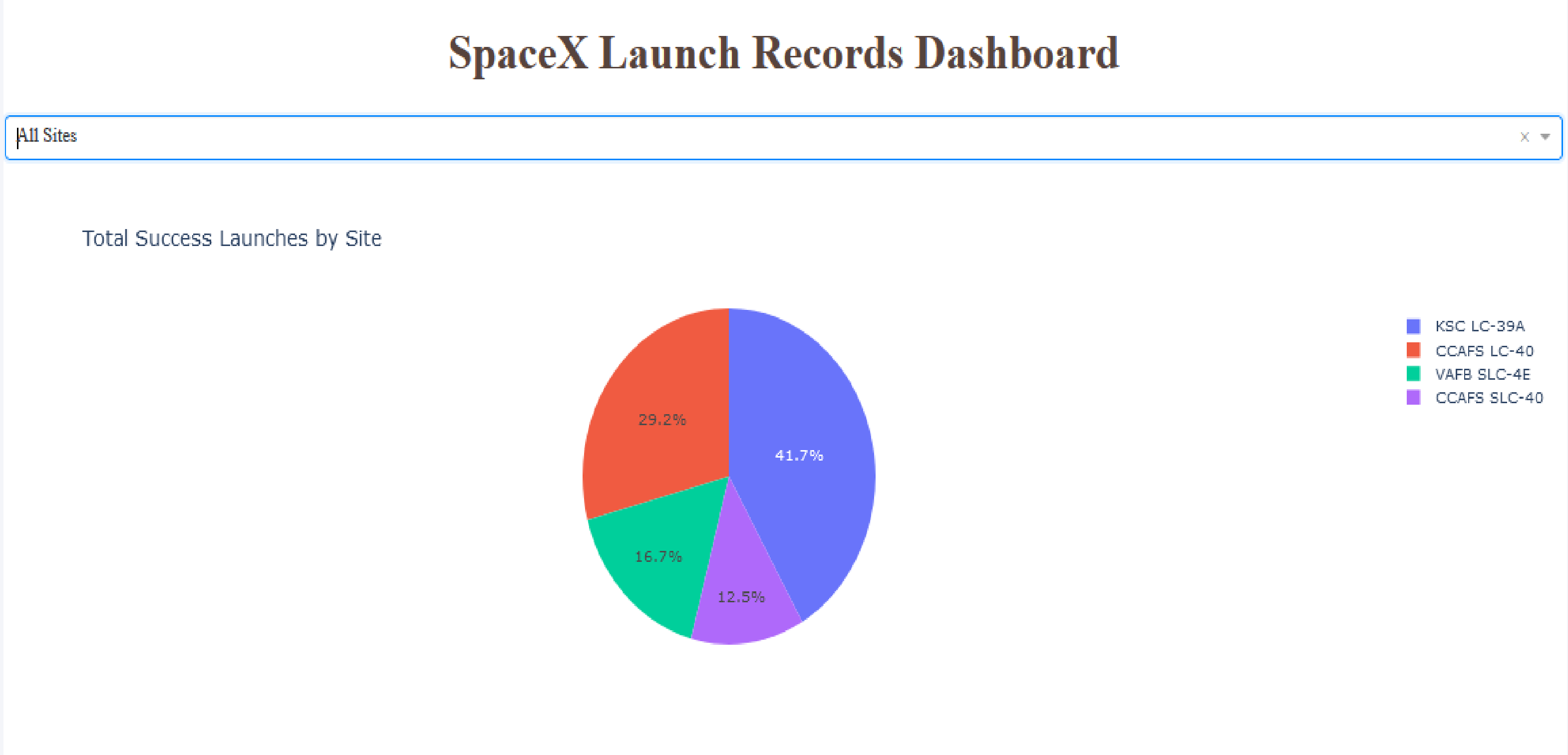

After you plot distance lines to the proximities, you can answer the following questions easily:

- Are launch sites in close proximity to railways? yes
- Are launch sites in close proximity to highways? yes
- Are launch sites in close proximity to coastline? yes
- Do launch sites keep certain distance away from cities? yes

Section 4

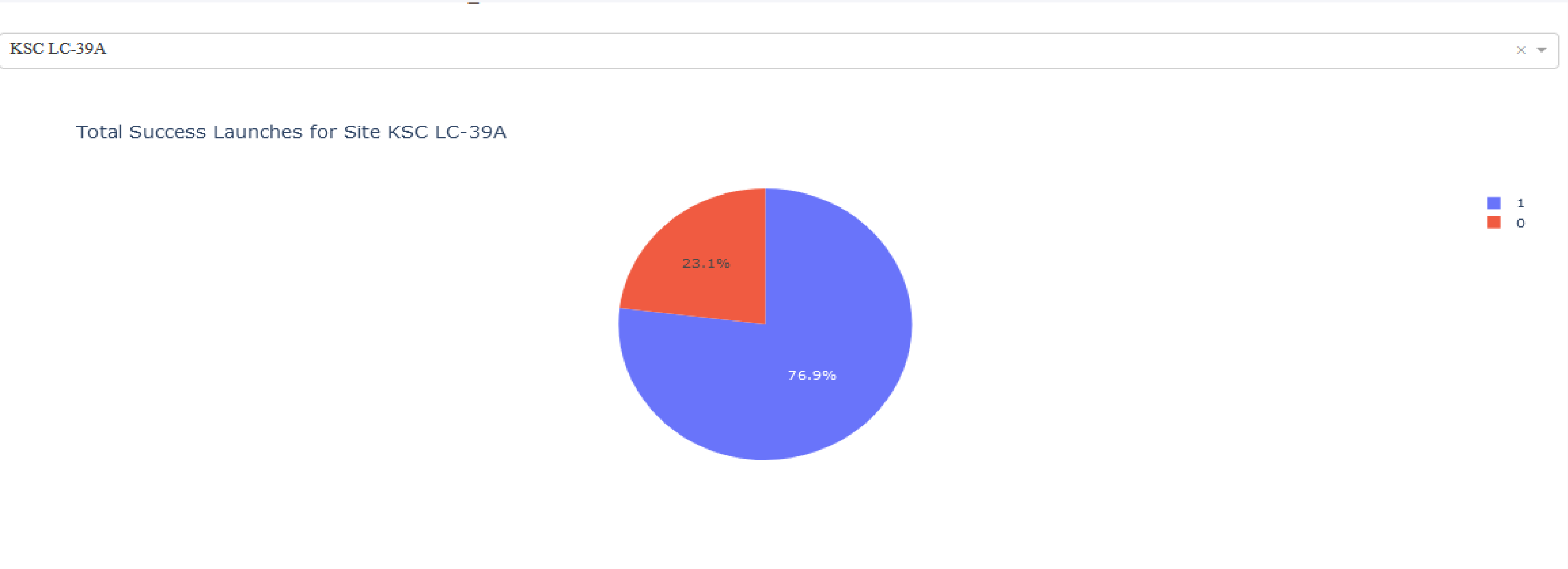
Build a Dashboard with Plotly Dash

Total Successful Launches by Site



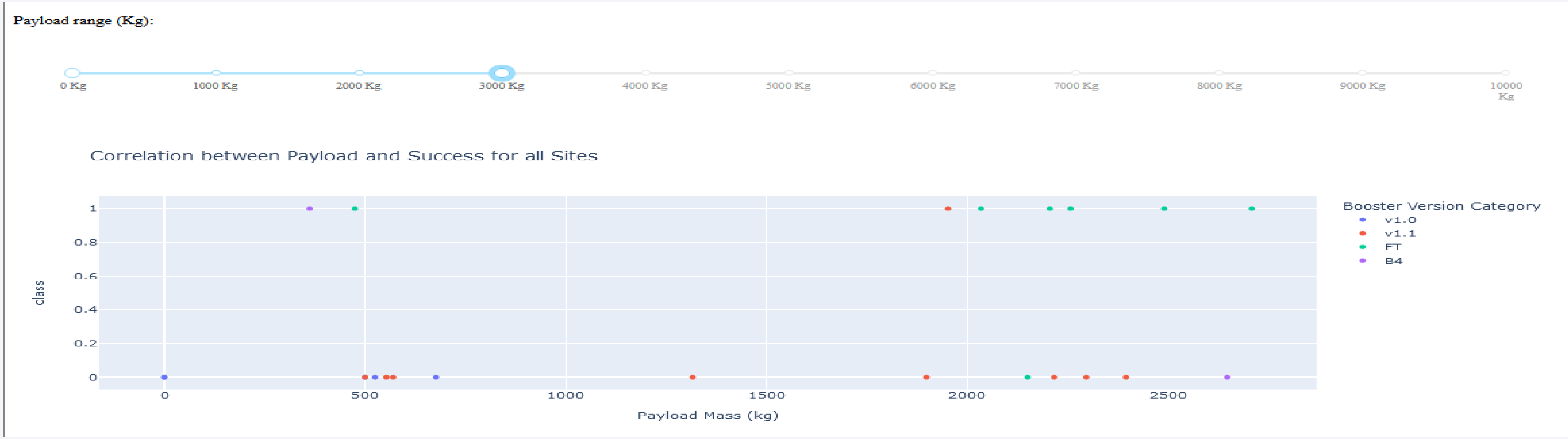
- With almost 42% Successful rate KSC LC-9A shows the best results

KSC LC-39A – Total Successful Launches



KSC LC-9A ranked 1st among launching sites. This result was achieved by scoring a <u>77% success rate</u>

Payload mass Vs. Outcome by Site (Low weight)

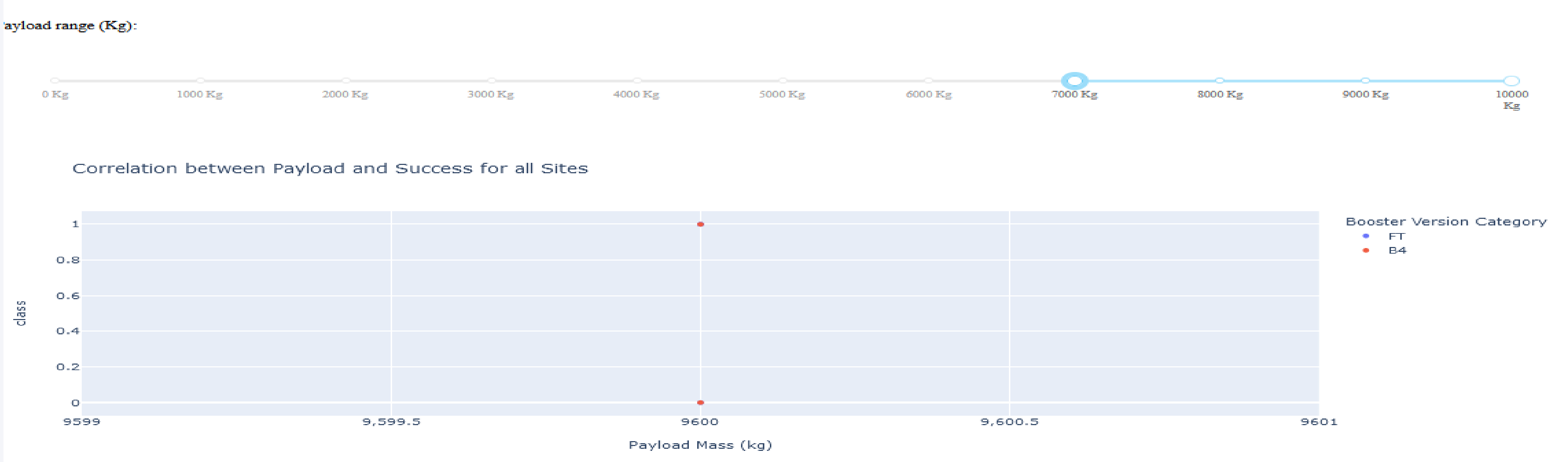


- Low weight = up to 3000Kg
- Low weight payload appear to be strongly correlated with the success rate of the following booster versions:
  - FT (positive correlation)
  - v1.1 (negative correlation)

Payload mass Vs. Outcome by Site (Medium weight)

- Medium weight = between 3000Kg and 7000Kg
- Medium weight payload appear to be strongly correlated with the success rate of the following booster versions:
  - FT & B4 (positive correlation up to 5000-5300Kg ; negative afterwards)
  - v1.1 (negative correlation)

Payload mass Vs. Outcome by Site (Heavy weight)

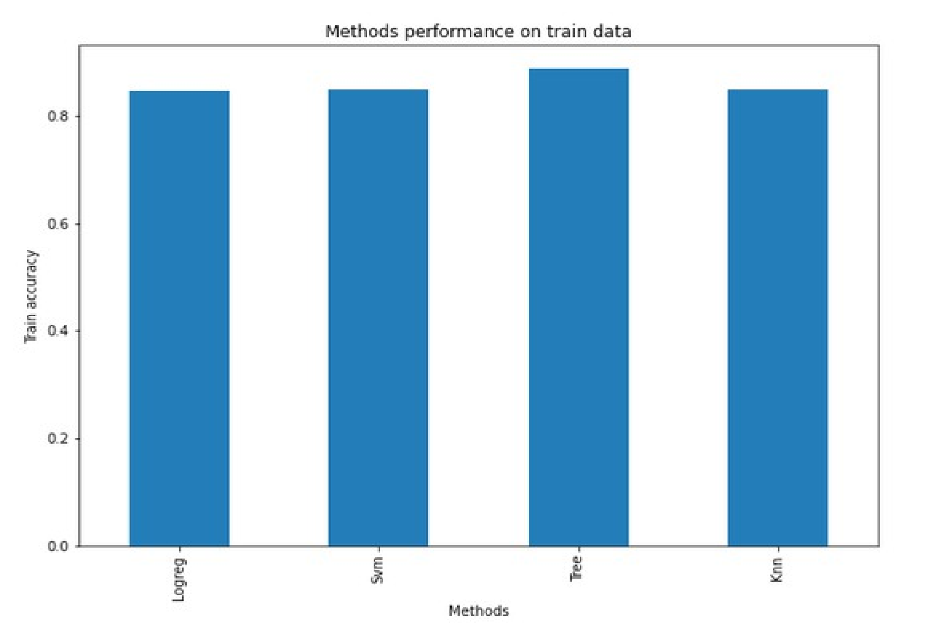


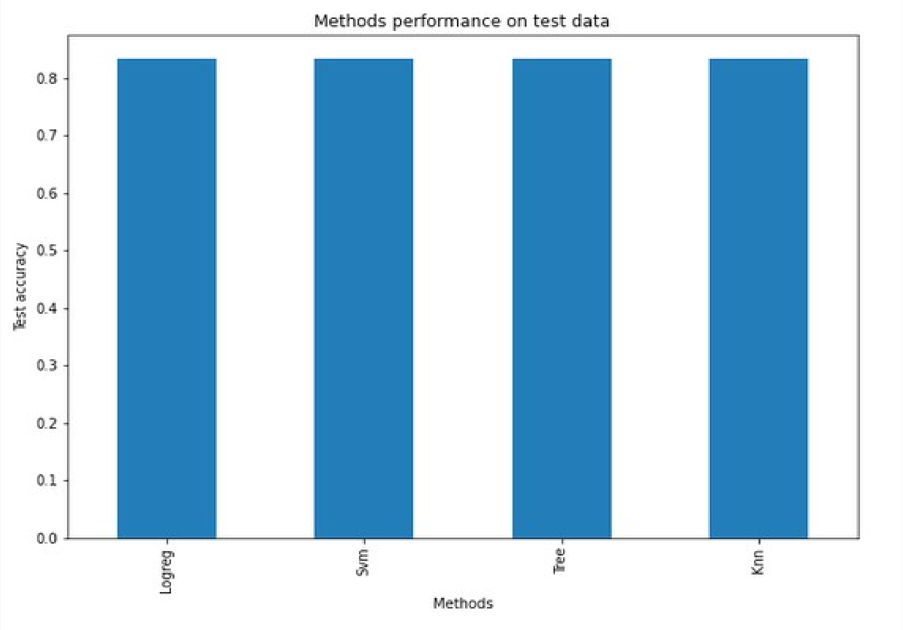
- Heavy weight = greater than 7000Kg
- Not enough data to draw conclusions

Section 5

Predictive Analysis (Classification)

Classification Accuracy



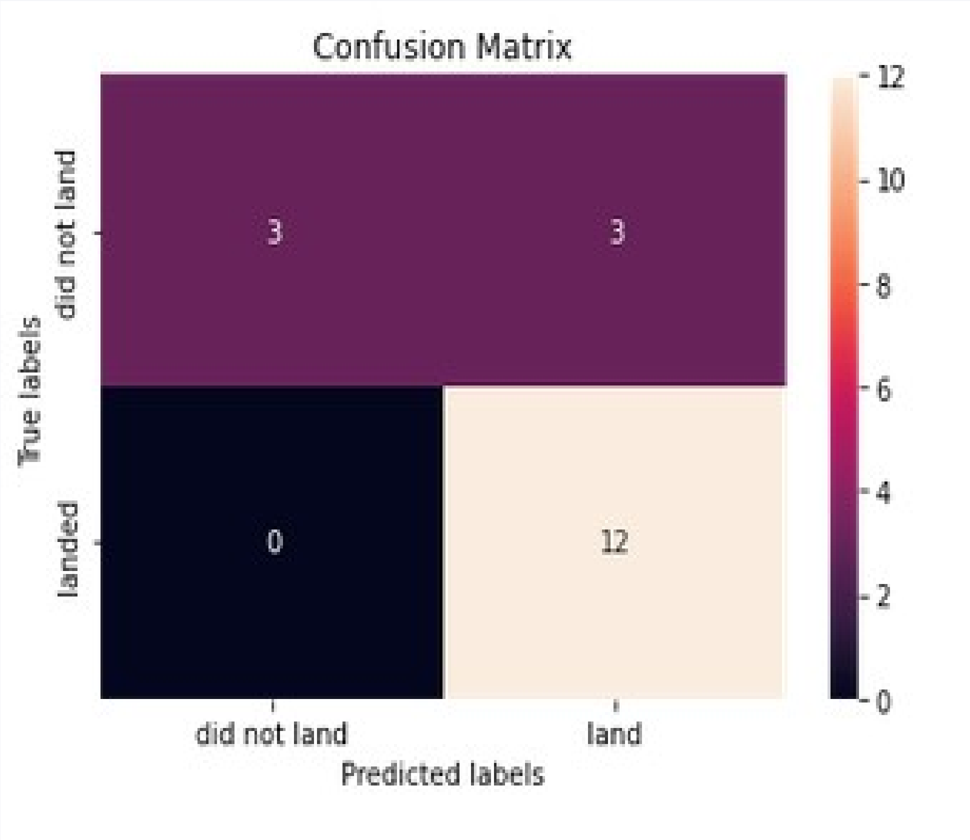


	Accuracy Train	Accuracy Test
Tree	0.887500	0.833333
Knn	0.848214	0.833333
Svm	0.848214	0.833333
Logreg	0.846429	0.833333

While Decision Tree shows slightly better results on the train data, when it comes to the test data all methods shows the same Accuracy results. More data and analysis are required to select the best method

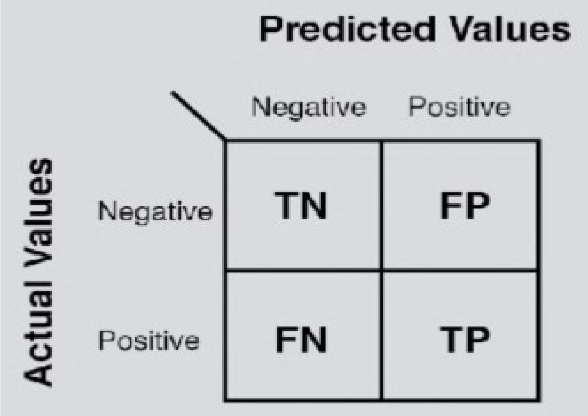
Confusion Matrix

Decision Tree CM



The Decision Tree method scored better results on the train data, however on the test data all methods scored the same result and therefore they all present a similar Confusion Matrix.

- All methods seems to struggle with False Positive (Top-right corner)



Conclusions

- The success rate of a mission appear to be correlated with different factors including Launch Site, Payload, Orbit and particularly the number of previous Launches. It seems reasonable to assume that there was a learning curve between launches that resulted in an increased success rate over time;
- Low-weight Payload shows the best overall success rate;
- KSC LC-39A ranked 1st among Launching Sites (77%); possibly because its the Launch Site with the highest number of low-weigh Payload launches;
- When looking at the different Orbits, the Payload mass appear to be a critical success factor;
- All used model showed similar Accuracy results on test data, however the Decision Tree classifier algorithm showed slightly better results on train data;

Thank you!