Data Science Capstone Project

Francesco Paolo De Gregorio

https://github.com/FrancescoDeGregorio

02/09/2021

OUTLINE

EXECUTIVE SUMMARY

- Collected data from public SpaceX API and SpaceX Wikipedia page.
- Created labels column ‘class’ which classifies successful landings.
- Explore data using SQL
- Data visualization, folium maps, and dashboards.
- Selection of features.
- Changed all categorical variables to binary (one hot encoding).
- Standardized data
- GridSearchCV to find best parameters for machine learning models.
- Accuracy score of all models.

- The process produced the following models: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors. All the models had similar outcomes and showed an accuracy of 83.33% circa. All models over predicted successful landings. More data is needed for better model determination and accuracy.

INTRODUCTION

Background:

- Space X has best pricing ($62 million vs. $165 million USD)
- This most to the capacity of recovering part of rocket (Stage 1)
- Space Y is competing with Space X

Problem:

- Space Y tasks us to train a machine learning model to predict successful Stage 1 recovery

METHODOLOGY

- Data collection methodology:
  - Combined data from SpaceX public API and SpaceX Wikipedia page
- Perform data wrangling
  - Classifying true landings as successful and unsuccessful otherwise
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Tuned models using GridSearchCV

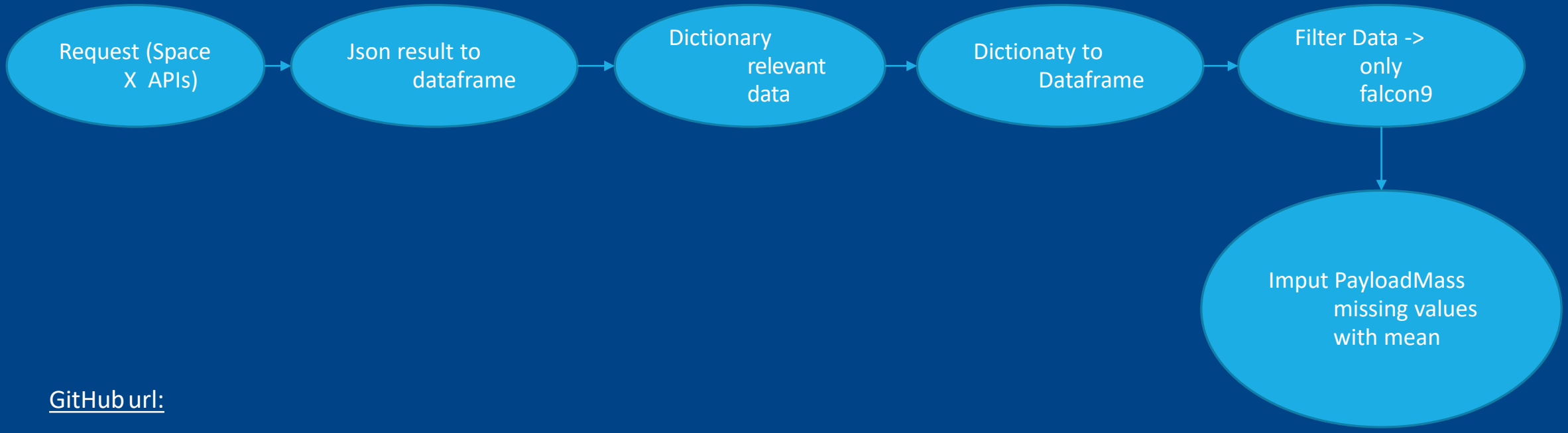
Methodology

DATA COLLECTION OVERVIEW

The data collection process consisted in both API requests to the SpaceX public API and web scraping from a table on the SpaceX's wikipedia page.

The next slide will show the flowchart of data collection from API and the one after will show the flowchart of data collection from webscraping.

Data Collection - SpaceX API



GitHub url:

https://github.com/FrancescoDeGregorio/IBM-Data-Science-Professional-Certificate/blob/main/Applied_Data_Science_Capstone/Week%201%20-%20Introduction/DataCollection_API.ipynb

Data Collection - Web Scraping

GitHub url:

https://github.com/FrancescoDeGregorio/IBM-Data-Science-Professional-Certificate/blob/main/Applied_Data_Science_Capstone/Week%201%20-%20Introduction/DataCollection_WebScraping.ipynb

DATA WRANGLING

Create a training label with landing outcomes where successful = 1 & failure = 0.

Outcome column contains: ‘Mission Outcome’ ‘Landing Location’

New training label column ‘class’ with a value of 1 if ‘Mission Outcome’ is True and 0 otherwise.

Value Mapping:

True ASDS, True RTLS, & True Ocean are set to 1 because they represent a successful landing.

None None, False ASDS, None ASDS, False Ocean, False RTLS are set to 0 because they represent a failed attempt.

GitHub url: https://github.com/FrancescoDeGregorio/IBM-Data-Science-Professional-Certificate/blob/main/Applied_Data_Science_Capstone/Week%201%20-%20Introduction/Data_wrangling%20.ipynb

EDA - DATA VISUALIZATION

Exploratory Data Analysis performed on variables with the help of some interesting plots in order to understand correlations between variables.

The following plots were done:

1) Flight Number vs. Payload Mass

2) Flight Number vs. Launch Site

3) Payload Mass vs. Launch Site

4) Orbit vs. Success Rate

5) Flight Number vs. Orbit

6) Payload vs Orbit

7) Success Yearly Trend

Scatter plots, line charts, and bar plots were used to compare relationships between variables to decide if a relationship exists so that they could be used in training the machine learning model

GitHub url: https://github.com/FrancescoDeGregorio/IBM-Data-Science-Professional-Certificate/blob/main/Applied_Data_Science_Capstone/Week%202%20-%20EDA%20-%20SQL/EDA_Visualization.ipynb

EDA-SQL

1) The dataset was loaded into the IBM DB2 Database through IBM Cloud.

2) Performed 10 queries thanks to the Python integration.

The objective of this task was to obtain a better understanding of the dataset.

Queried information regarding launch site names, mission outcomes, various pay load sizes of customers and booster versions, and landing outcomes

GitHub url: https://github.com/FrancescoDeGregorio/IBM-Data-Science-Professional-Certificate/blob/main/Applied_Data_Science_Capstone/Week%202%20-%20EDA%20-%20SQL/EDA__SQL.ipynb

INTERACTIVE MAP WITH FOLIUM

Folium maps mark Launch Sites, successful and unsuccessful landings, and a proximity example to key locations: Railway, Highway, Coast, and City.

This allows us to understand why launch sites may be located where they are. Also visualizes successful landings relative to location.

GitHub url: https://github.com/FrancescoDeGregorio/IBM-Data-Science-Professional-Certificate/blob/main/Applied_Data_Science_Capstone/Week%203%20-%20Visual%20Analytics%20-%20Dashboards/Interactive_Visuals_with_Folium.ipynb

DASHBOARD - PLOTLY DASH

The dashboard created contains a pie chart and scatterplots.

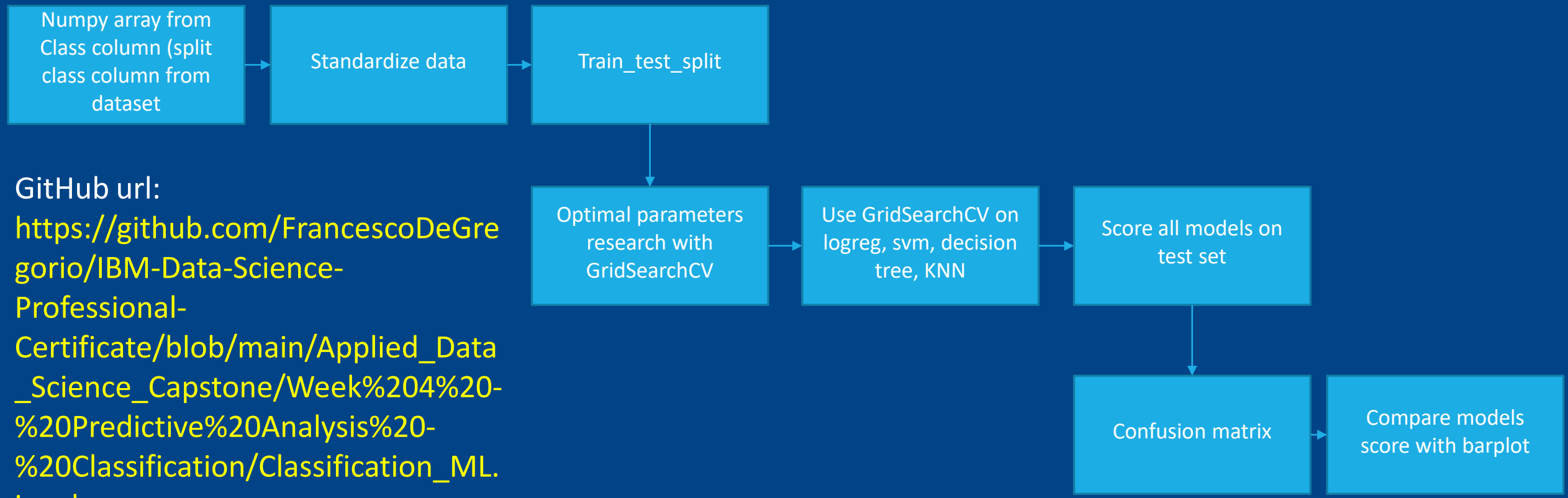
Pie chart can be used to show distribution of successful landings for all launch sites or also for individual launch site.

Scatterplots are made in order to visualize relationship between launch sites and payload mass.

The pie chart is used to visualize launch site success rate.

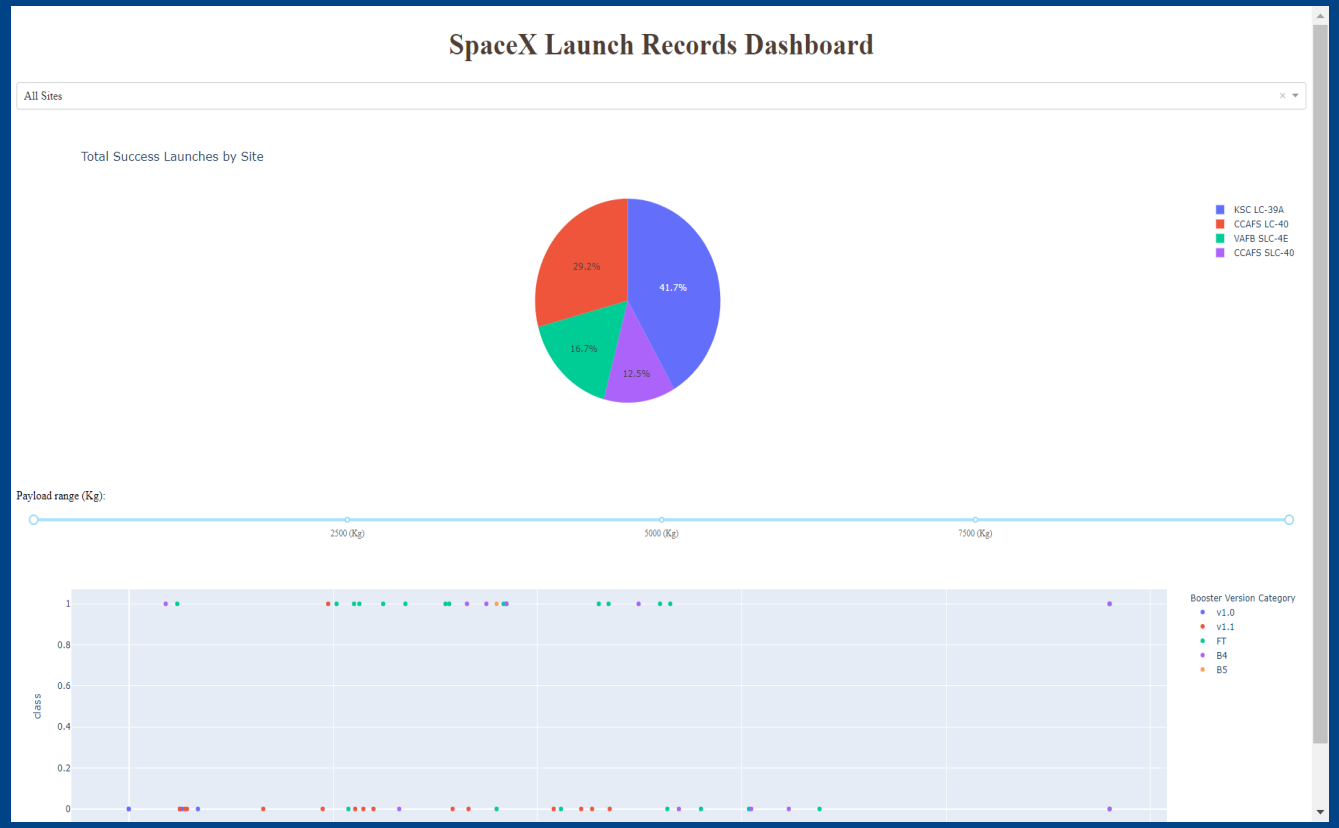
GitHub url: https://github.com/FrancescoDeGregorio/IBM-Data-Science-Professional-Certificate/blob/main/Applied_Data_Science_Capstone/Week%203%20-%20Visual%20Analytics%20-%20Dashboards/spacex_dash_app.py

CLASSIFICATION



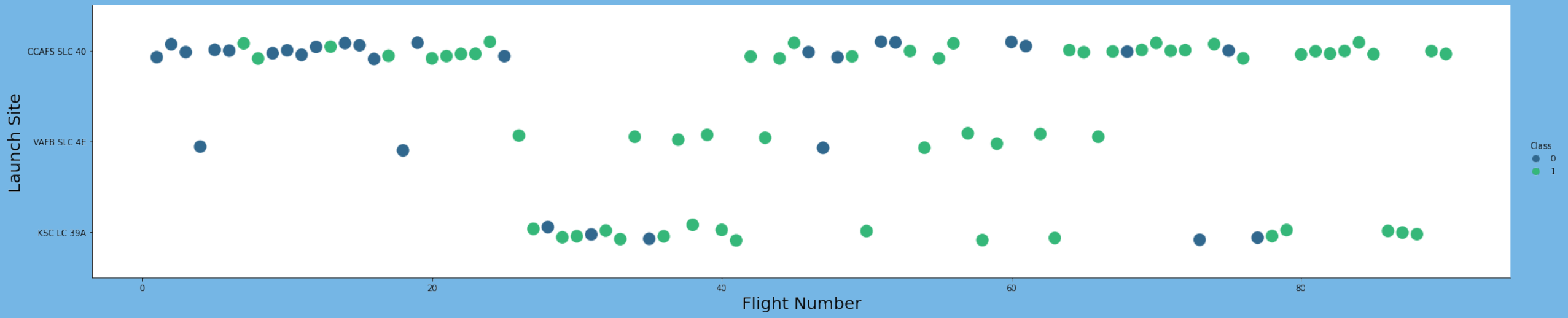
GitHub url: https://github.com/FrancescoDeGregorio/IBM-Data-Science-Professional-Certificate/blob/main/Applied_Data_Science_Capstone/Week%204%20-%20Predictive%20Analysis%20-%20Classification/Classification_ML.ipynb

PLOTLY DASH PREVIEW



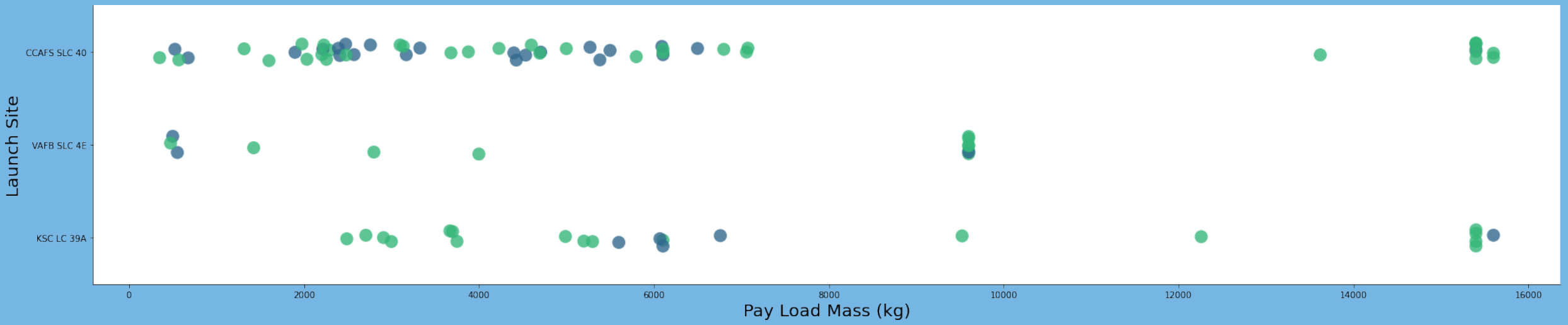
EDA – Visualization

FLIGHT NUMBER VS. LAUNCHSITE



The plot can indicate that there was an increase of success over time(Flight Number). We can observe a breakthrough from flight 20 on which significantly increased success rate. CCAFS can be considered as the most important launch site as it has the most volume.

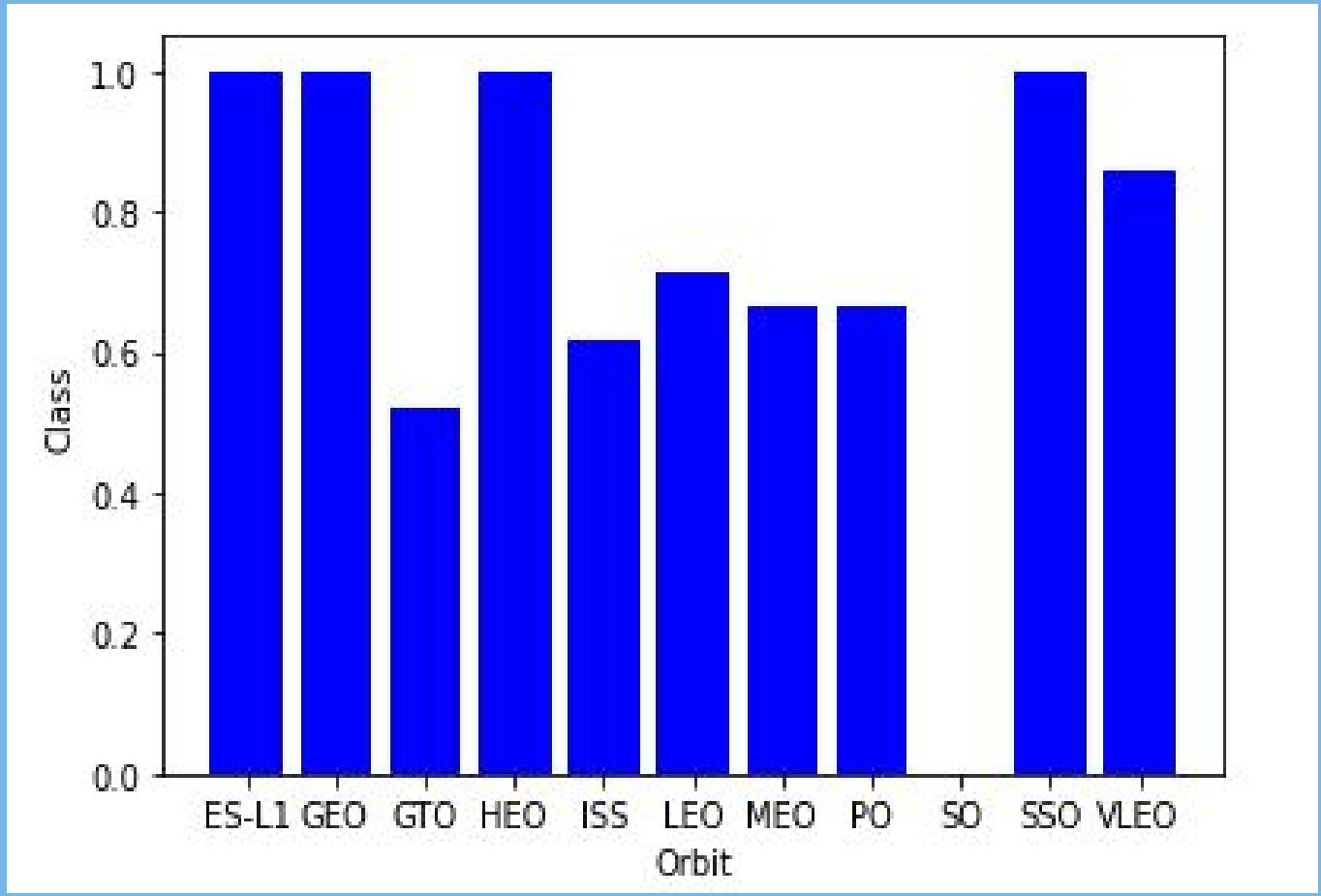
PAYLOAD VS. LAUNCHSITE



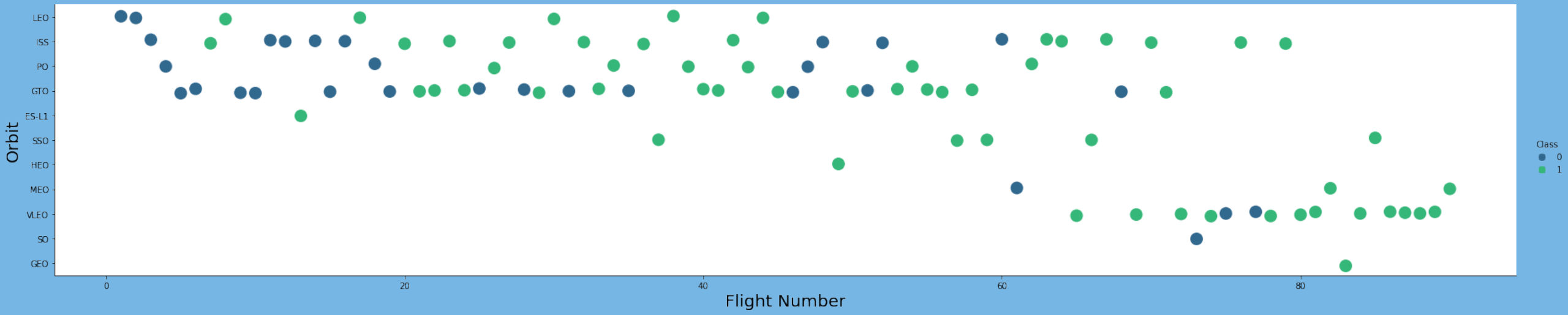
Class 0 = unsuccessful launch
Class 1 = successful launch

SUCCESS RATE VS. ORBIT TYPE

The plot above shows the different success rate for the various orbit types. The orbit types with most higher success rates are: ES-L1, GEO, HEO and SSO.
The other orbit type such as VLEO, MEO, PO, ISS, GTO and SO have less success rate.

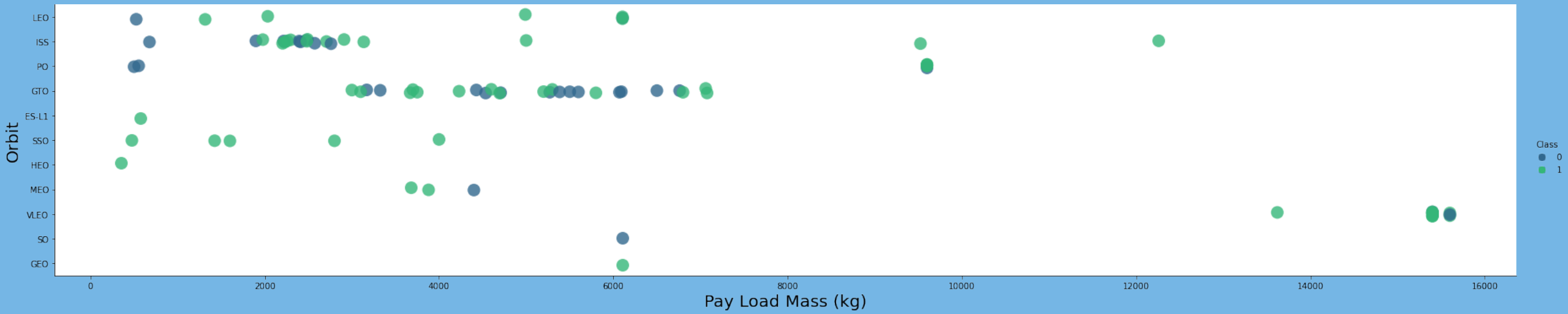


FLIGHT NUMBER VS. ORBITTYPE



Class 0 = unsuccessful launch
Class 1 = successful launch

PAYLOAD VS. ORBIT TYPE



Class 0 = unsuccessful launch
Class 1 = successful launch

LAUNCH SUCCESS YEARLY TREND

We can observe that success increased continuously from 2013 to 2018 when there was a small fall. Then from 2018 to 2019 there was another important increase that then (from 2019 to 2020) slightly slowed.

Success in recent years at around 80%

EDA WITH SQL

EXPLORATORY DATA ANALYSIS WITH SQL DB2

INTEGRATED IN PYTHON WITH SQLALCHEMY

DISPLAY THE NAMES OF THE UNIQUE LAUNCH SITES IN THE SPACE MISSION

This query shows us the unique launch sites names in the space mission.

We can observe that the names are the following:

CCAFS LC-40

CCAFS SLC-40

KSC LC-39A

VAFB SLC-4E

Task 1

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

```
%sql select DISTINCT LAUNCH_SITE from SPACEXDATASET
```

```
 * ibm_db_sa://nxs27972:***@54a2f15b-5c0f-46df-8954-7e38e61
Done.
```

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

DISPLAY 5 RECORDS WHERE LAUNCH SITES BEGIN WITH THE STRING 'CCA'

The following query shows 5 entries of the DB that have launch site name starting with CCA

Task 2

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

```
%sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5
```

```
 * ibm_db_sa://nxs27972:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/BLUDB
Done.
```

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

DISPLAY THE TOTAL PAYLOAD MASS CARRIED BY BOOSTERS LAUNCHED BY NASA (CRS)

This query performs the sum of all the payload mass carried by boosters launched by NASA (CRS)

Task 3

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

```
%sql select sum(payload_mass__kg_) as sum from SPACEXDATASET where customer like 'NASA (CRS)'
```

```
 * ibm_db_sa://nxs27972:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases
Done.
```

SUM
45596

DISPLAY AVERAGE PAYLOAD MASS CARRIED BY BOOSTER VERSION F9 V1.1

This query displays the average payload mass of booster with version F9 v1.1

Task 4

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

```
%sql select avg(payload_mass__kg_) as Average from SPACEXTBL where booster_version = 'F9 v1.1'
```

```
 * ibm_db_sa://vwm02263:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases
Done.
```

average
2928

FIRST SUCCESSFUL GROUND PAD LANDING DATE

This sql query return the date of the first successful landing outcome in ground pad

Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
%sql select min(date) as Date from SPACEXTBL \
where landing__outcome like 'Success (ground pad)'
```

```
 * ibm_db_sa://vwm02263:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l0
Done.
```

DATE
2015-12-22

SUCCESSFUL DRONE SHIP LANDING WITH PAYLOAD BETWEEN 4000 AND 6000

The following query list the names of the boosters which had successful landing outcome (in drone ship) and have payload between 4000 and 6000

Task 6

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

```
%sql select BOOSTER_VERSION \
from SPACEXTBL \
where (LANDING__OUTCOME = 'Success (drone ship)') and (PAYLOAD_MASS__KG_  between 4000 and 6000)
```

```
 * ibm_db_sa://vwm02263:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.clou
Done.
```

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

TOTAL NUMBER OF EACH MISSION OUTCOME

This query returns a count of each mission outcome.

SpaceX appears to achieve its mission outcome nearly 99% of the time.

Only 1 flight had an unsuccessful mission outcome and also 1 other flight had an unclear status.

Task 7

List the total number of successful and failure mission outcomes

```
%sql SELECT mission_outcome, count(*) as Count FROM SPACEXTBL GROUP by mission_outcome ORDER BY mission_outcome
```

```
 * ibm_db_sa://vwm02263:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/b
Done.
```

mission_outcome	COUNT
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

LIST THE NAMES OF THE BOOSTER VERSIONS WHICH HAVE CARRIED THE MAXIMUM PAYLOAD MASS

This query returns the booster versions that carried the highest payload mass.

All the booster version listed are of the type F9 B5 B10xx.x

This may indicates payload mass is correlated with the booster version that is used.

Task 8

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

```
%sql select distinct(BOOSTER_VERSION) \
from SPACEXTBL \
where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL);
```

```
 * ibm_db_sa://vwm02263:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databa
Done.
```

booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

FAILED DRONE SHIP LANDING RECORDS IN 2015

This query returns the Month, Landing Outcome, Booster Version, Payload Mass (kg), and Launch site of 2015 launches where stage 1 failed to land on a drone ship.

There were two such occurrences.

Task 9

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for the in year 2015

```
%sql select MONTHNAME(DATE) as Month, landing__outcome, booster_version, launch_site \
from SPACEXTBL \
where DATE like '2015%' AND landing__outcome like 'Failure (drone ship)'
```

```
 * ibm_db_sa://vwm02263:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.app
Done.
```

MONTH	landing__outcome	booster_version	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

RANKING COUNTS OF SUCCESSFUL LANDINGS BETWEEN 2010-06-04 AND 2017-03-20

Task 10

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

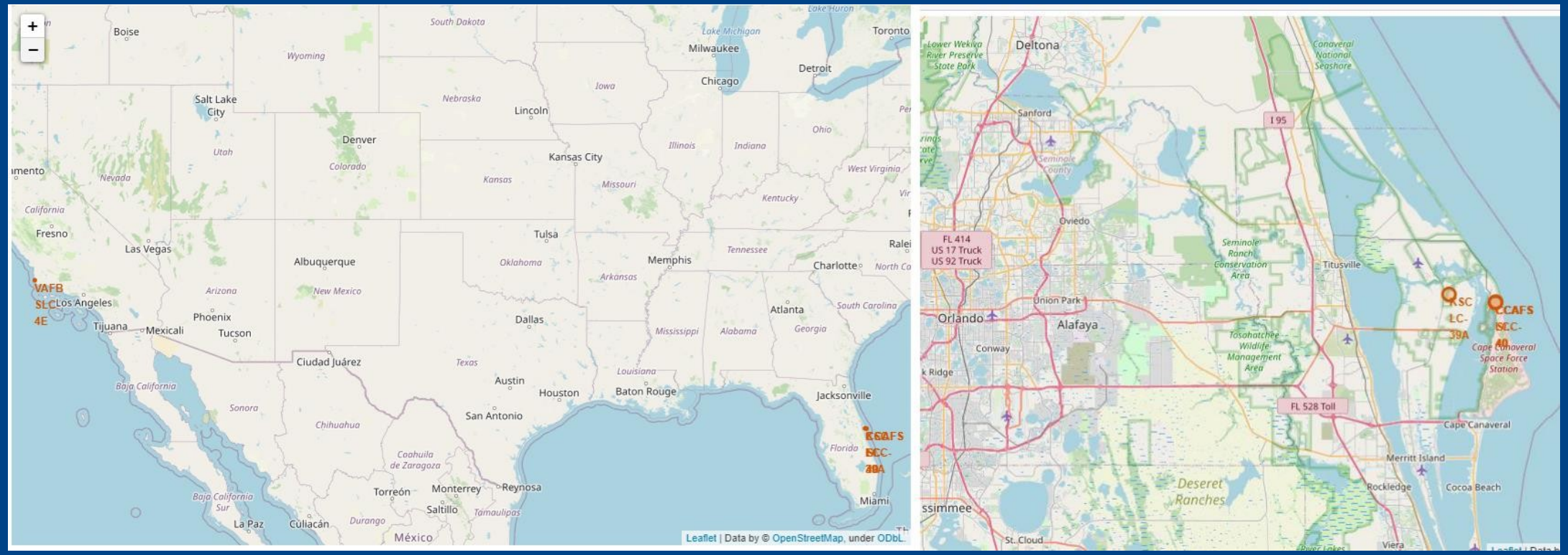
```
%sql select landing__outcome, count(*) as count \
from SPACEXTBL \
where Date >= '2010-06-04' AND DATE <= '2017-03-20' \
GROUP by landing__outcome ORDER BY count Desc
```

```
 * ibm_db_sa://vwm02263:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

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

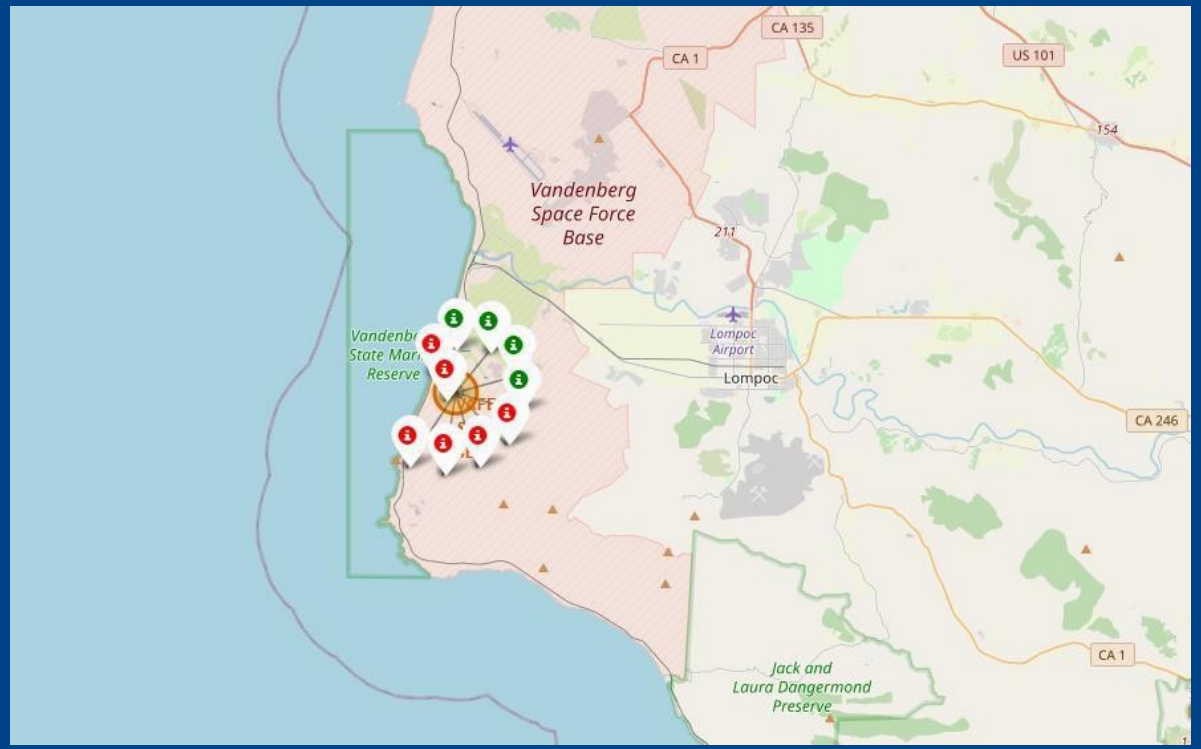
INTERACTIVE MAP - FOLIUM

LAUNCH SITE LOCATIONS



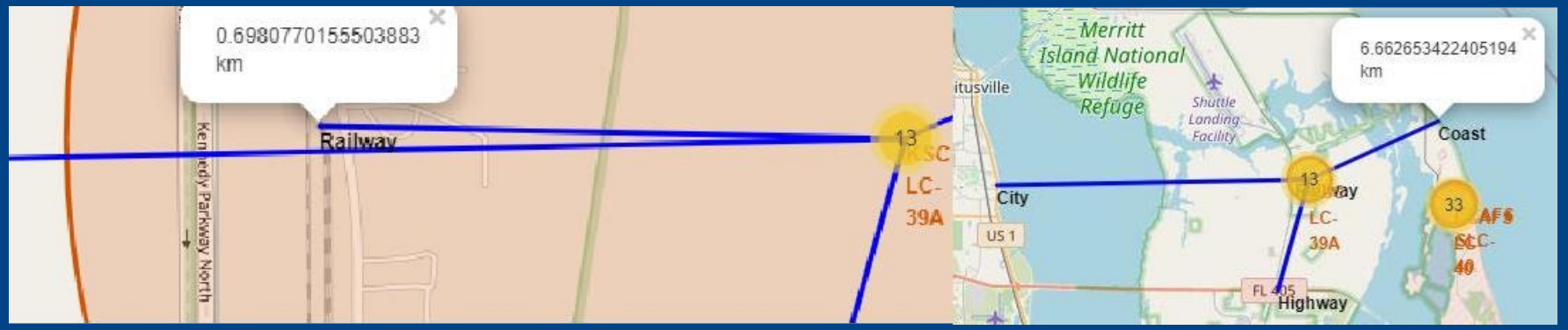
The left map shows all launch sites relative US map. The right map shows the two Florida launch sites since they are very close to each other. All launch sites are near the ocean.

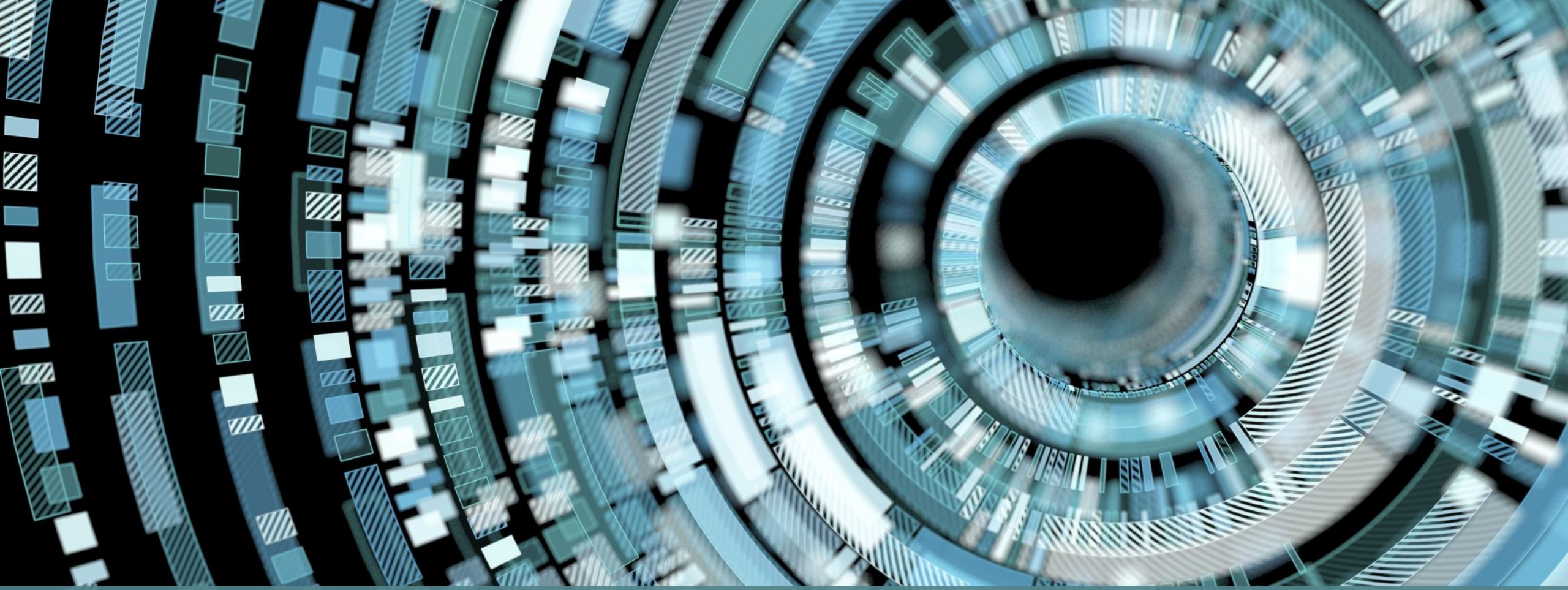
COLOR-CODED LAUNCH MARKERS



Clusters on Folium map can be clicked on to display each successful landing (green icon) and failed landing (red icon). In this example VAFB SLC-4E shows 4 successful landings and 6 failed landings.

KEY LOCATION PROXIMITIES

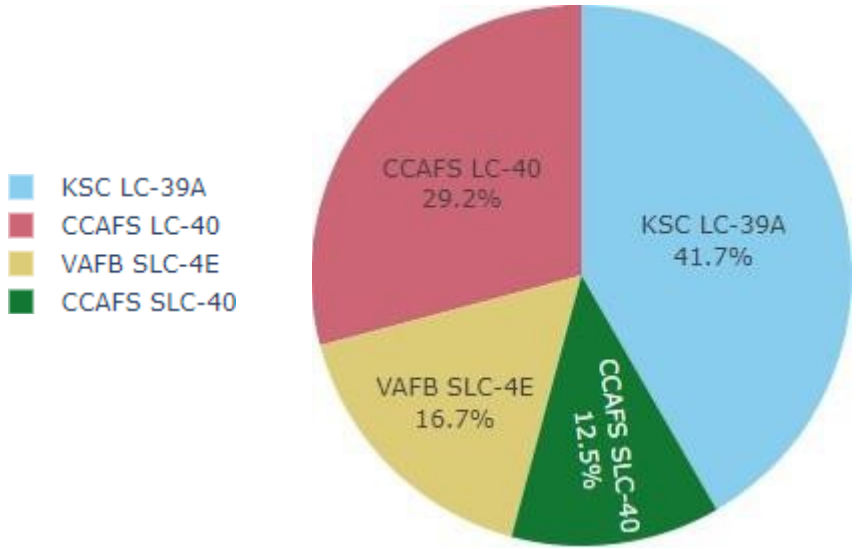




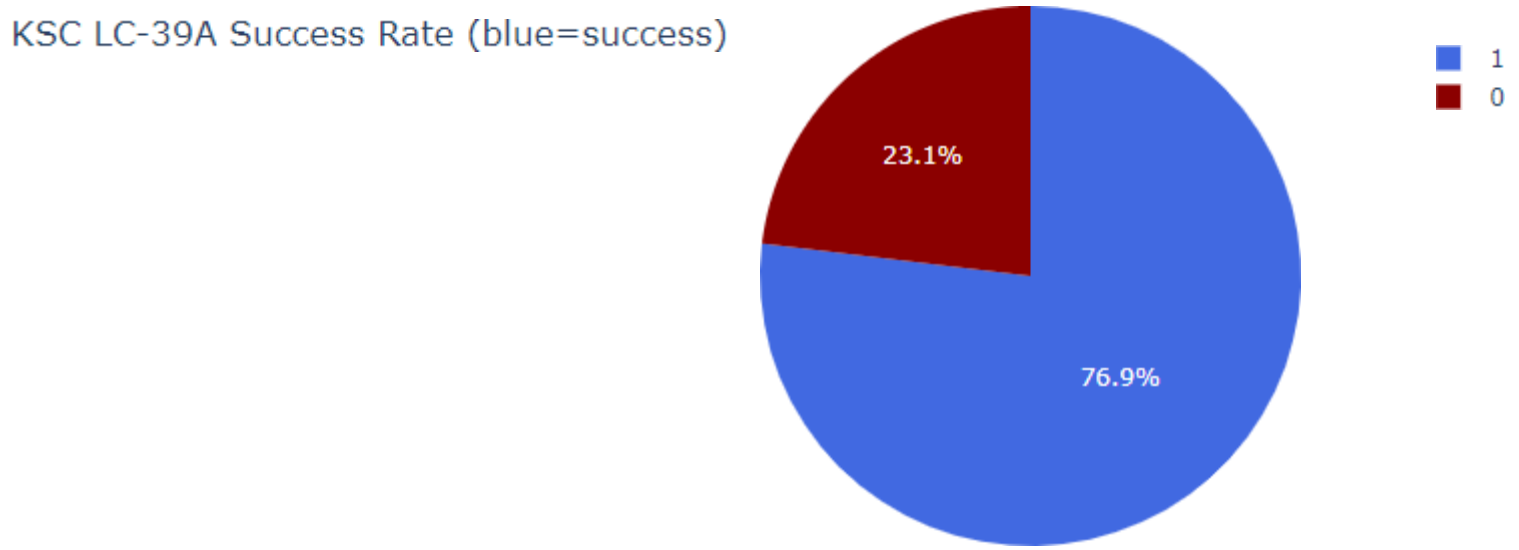
DASHBOARD - PLOTLY DASH

SUCCESSFUL LANDINGS ACROSS DIFFERENT LAUNCH SITES

The launch site with most successful landings id KSC LC-39A, followed by CCAFS LC-40, VAFB SLC-4E



HIGHEST SUCCESS RATE LAUNCH SITE

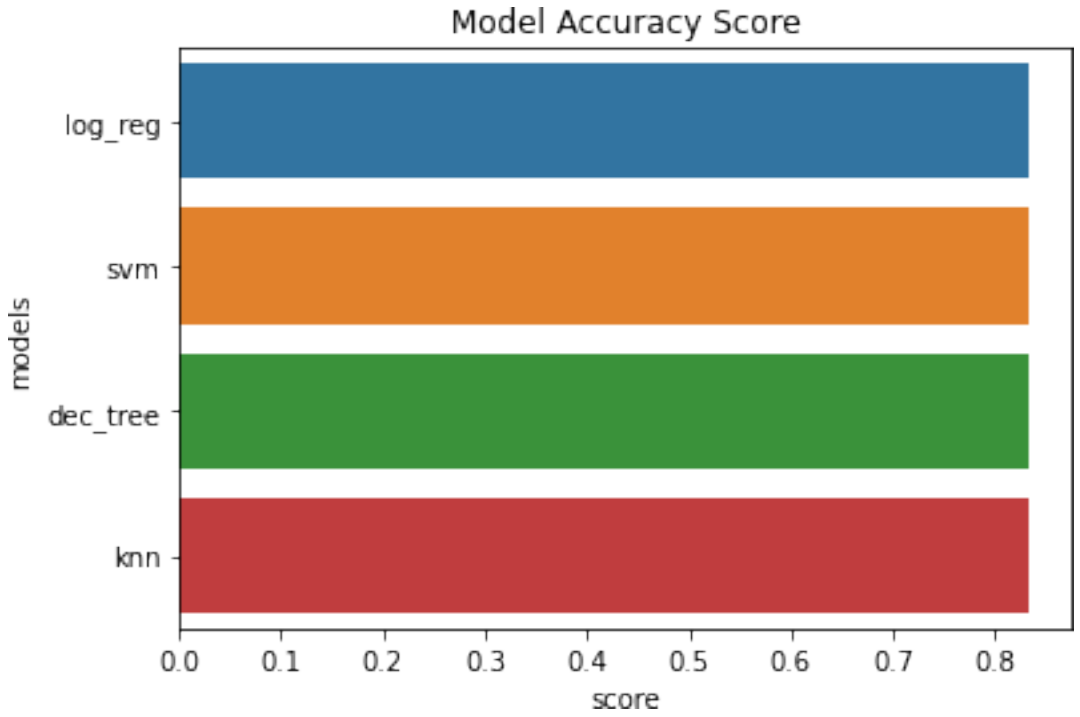


KSC LC-39A has the highest success rate.

PAYLOAD MASS VS. SUCCESS VS. BOOSTER VERSION CATEGORY

CLASSIFICATION

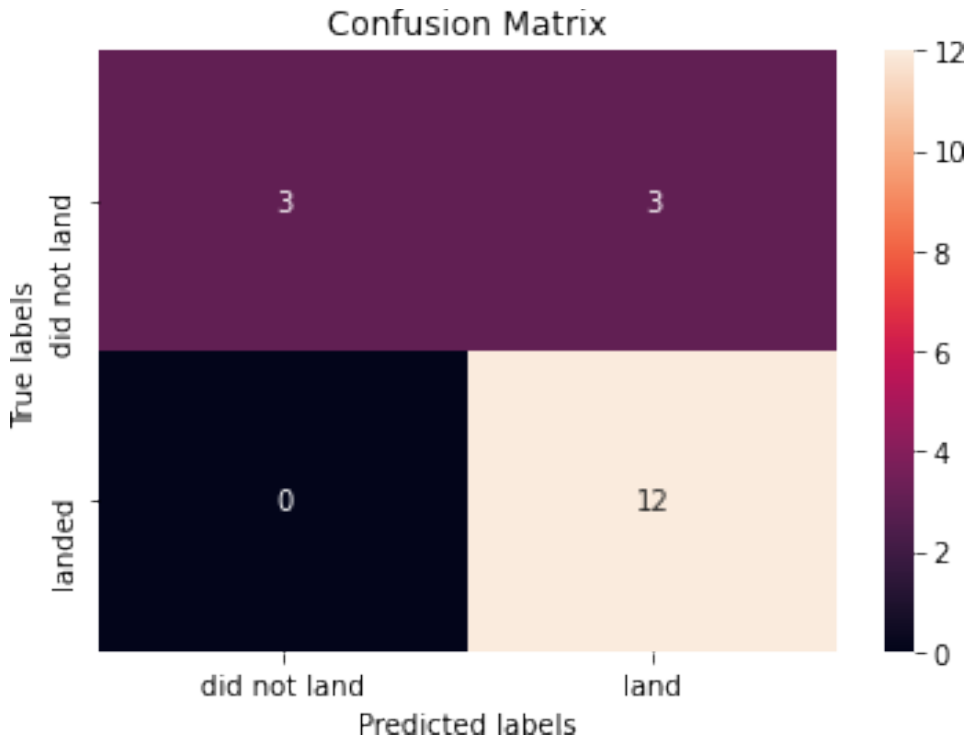
CLASSIFICATION ACCURACY



All models show to have approximately the same accuracy score of 83.33% circa.

We likely need more data to determine the best model.

CONFUSION MATRIX



Correct predictions are on a diagonal from top left to bottom right.

CONCLUSION

1) Our task: to develop a machine learning model for Space Y who wants to bid against SpaceX

2) The goal of model is to predict when Stage 1 will successfully land to save ~$100 million USD

3) Data source -> public SpaceX API and web scraping SpaceX Wikipedia page

4) Stored data into the IBM BD2 database

5) Done visualization and interactive dashboard

6) Trained a machine learning model with an accuracy of 83%

7) Space Y can use the models to predict with high accuracy if stage 1 will land successfully or not, and possibly save money.

APPENDIX

GitHub repository url: https://github.com/FrancescoDeGregorio/IBM-Data-Science-Professional-Certificate/tree/main/Applied_Data_Science_Capstone