IBM Applied Data Science Capstone Project

Capstone Project By

MOHAMMAD ASHRAF ANSARI

OUTLINE

- Executive Summary
- Introduction
- Methodology
  - Data collection
  - EDA (viz & sql)
  - Classification prediction
- Results
  - Visualization – Charts
  - Dashboard
- Discussion
  - Findings & Implications
- Conclusion
- Appendix

EXECUTIVE SUMMARY

- SpaceX is a leading company in making spacecrafts, their major achievement is that they can sometimes land stage one rocket launchers again instead of losing them, bringing the cost way down.
- This report is aimed at looking at the historical launches of SpaceX rockets to investigate the success rate of landing, and if it's correlated to
  - Location
  - Payload Mass
  - Booster type

INTRODUCTION

- SpaceX doesn’t attempt to land each launched rocket
- So, it’s useful to investigate and see if we can predict if SpaceX will have a successful landing for a future launch or not.
- SpaceX historical launch data is a good starting point

METHODOLOGY

- We collect SpaceX launch historical data, through their API
- We perform Exploratory data analysis to see which factors contribute to the success or failure of the landing
- We isolate these features, wrangle the data and standardize it.
- We use classification algorithms to see which one has the highest accuracy.

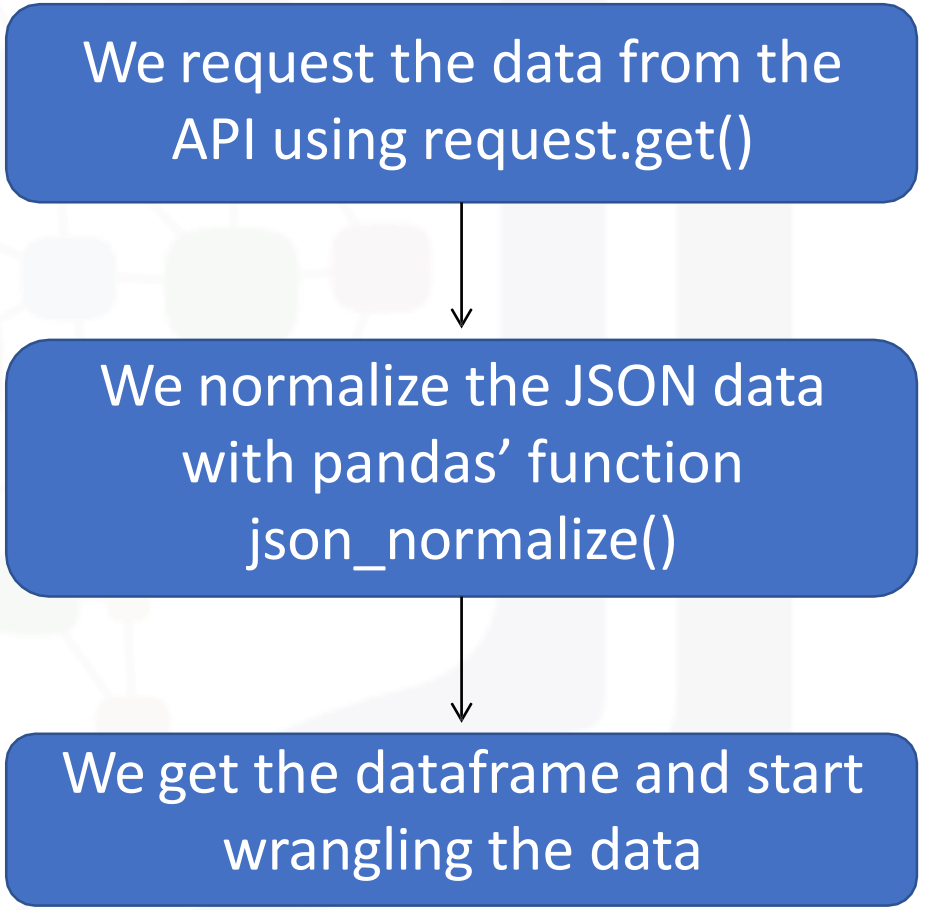
Data collection

SpaceX API

We request the historical data from this url: https://api.spacexdata.com/v4/launches/past

Here's the notebook for data collection:

https://github.com/asrf007/capstone-project



Data Cleaning

SpaceX API

We clean the data and wrangle the data for feature engineering as preparation for the analysis.

Data Cleaning

We select only the columns relevant to our analysis; excluding ids, links, images and other irrelevant columns

We filter the data for Falcon 9 Launches only.

Data Wrangling

SpaceX API

We clean the data and wrangle the data for feature engineering as preparation for the analysis.

Here's the notebook for data wrangling:

https://github.com/asrf007/capstone-project

Data Wrangling

We check missing data, and replace it with appropriate substitute; in our case, we replace missing payload mass with the average.

We define a column for the output, to be 0 if failure and 1 if success.

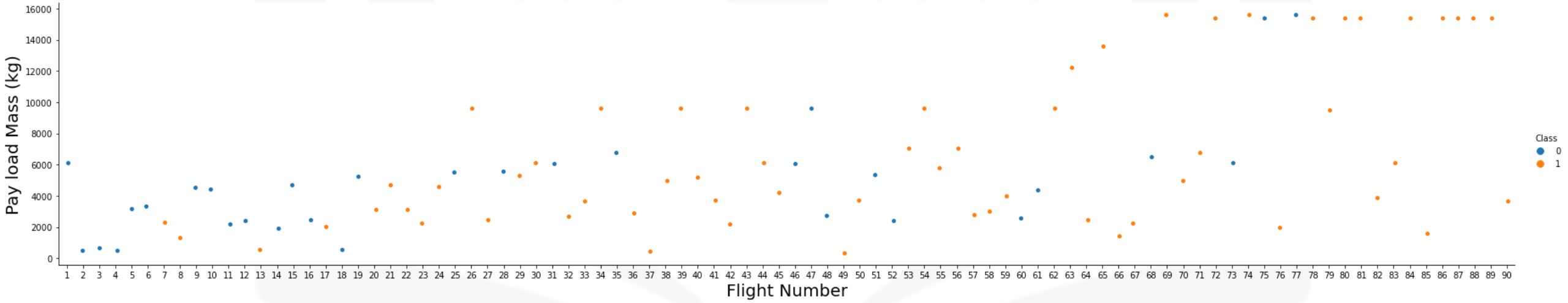
EDA with data visualization

In this notebook, we perform preliminary EDA with some visualization to get an idea about the relationship between different factors and success rate. So, we visualize the following:

- The number of flights and payload mass
- Launch site and flight number
- Payload mass and launch site
- Success rate and orbit
- The number of flights and payload mass
- Flight number and orbit
- Payload and orbit
- Launch success yearly trend

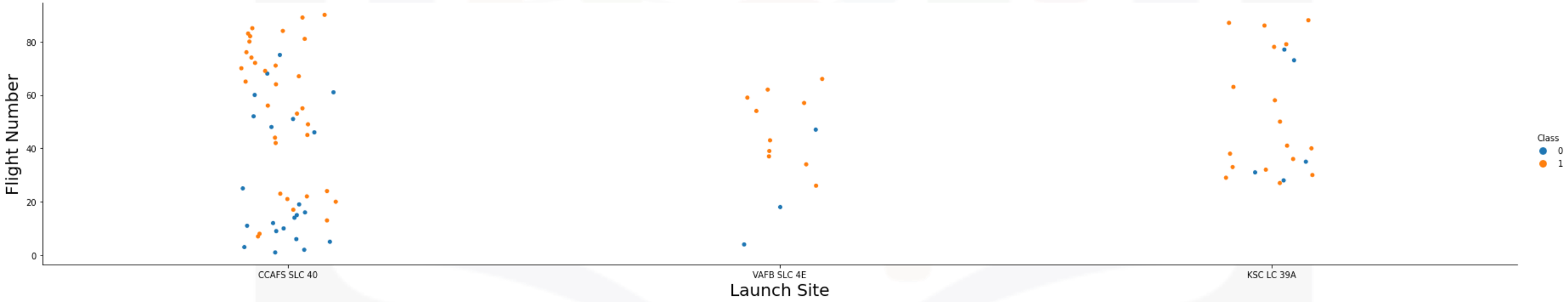
Correlation trends

Relationship between Flight number, and Payload Mass



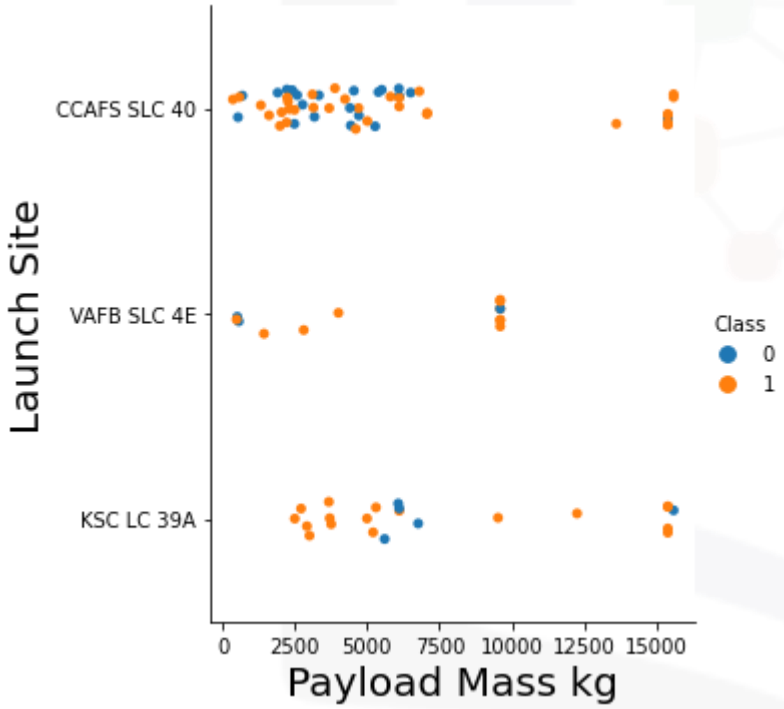
Correlation trends

Relationship between Flight number, and Launch Site



Correlation trends

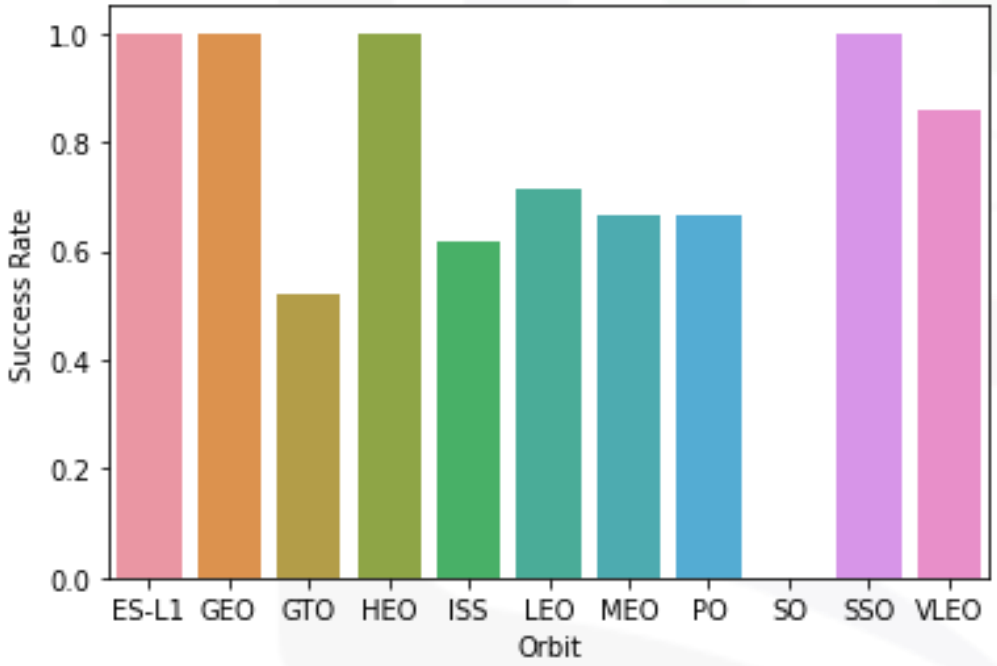
Relationship between Launch Site, and Payload Mass



We can see that the launch site “VAFB SLC 4E” doesn’t have launches with heavy payloads

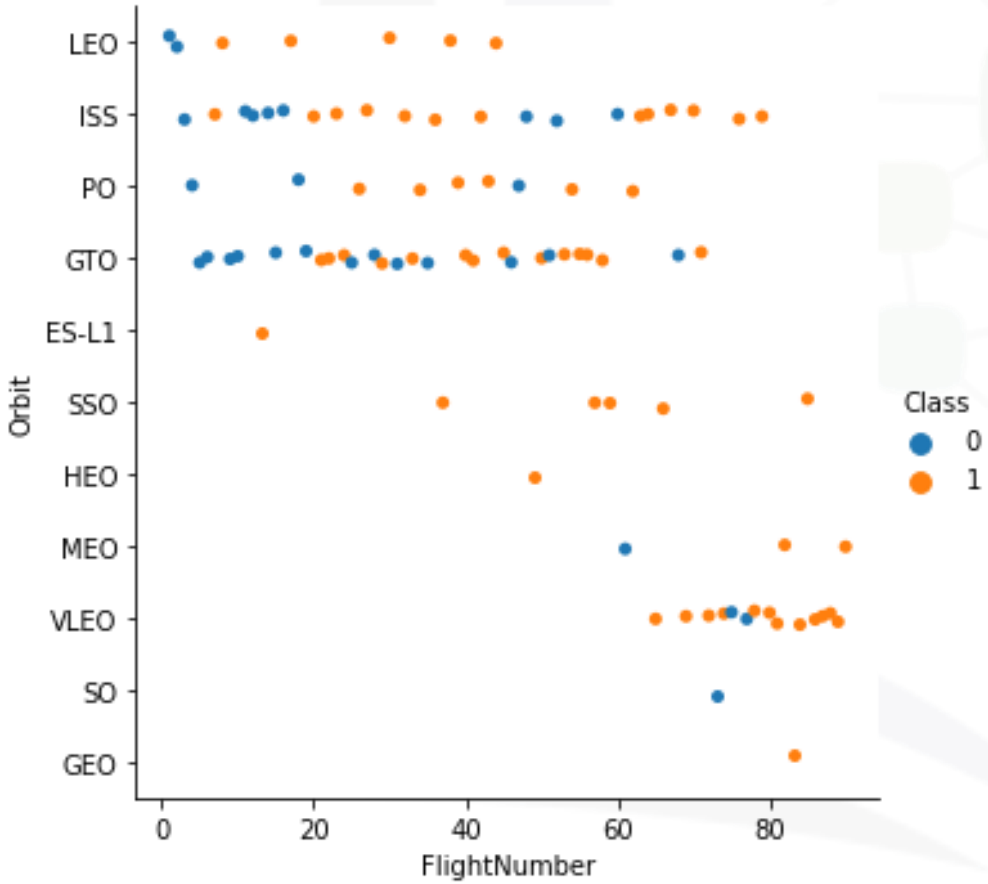
Correlation trends

Success rate for each orbit



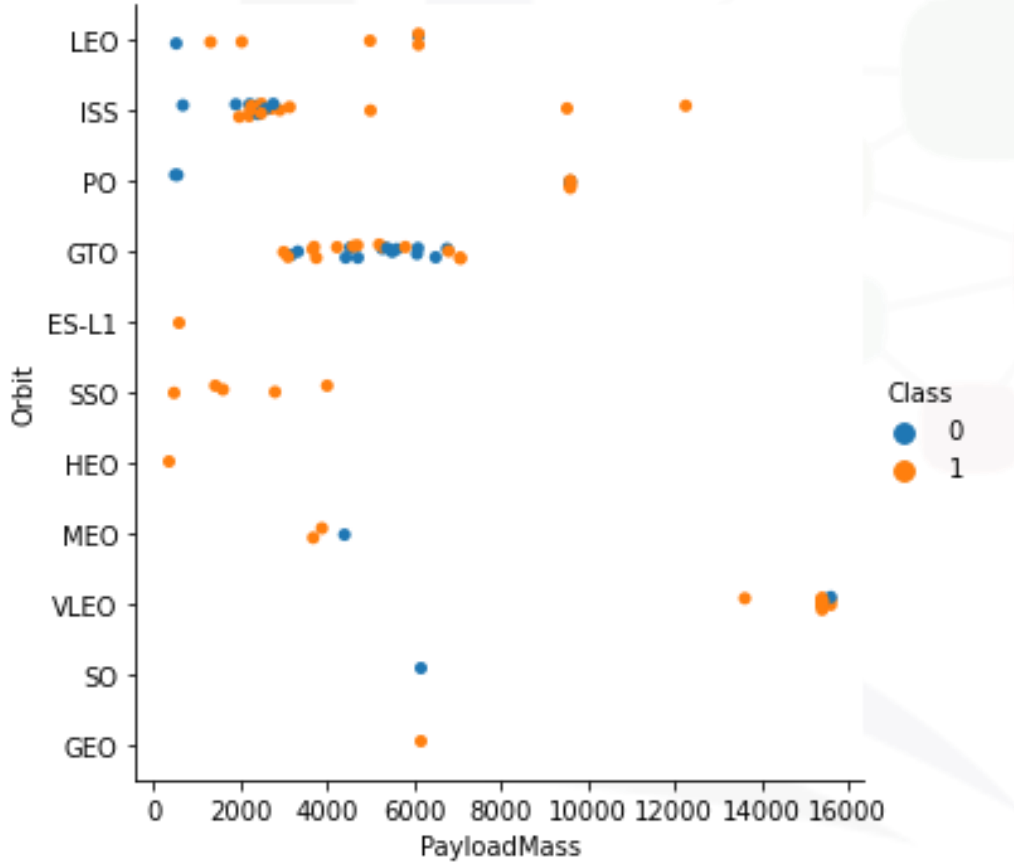
Correlation trends

Relationship between Flight number and orbit type



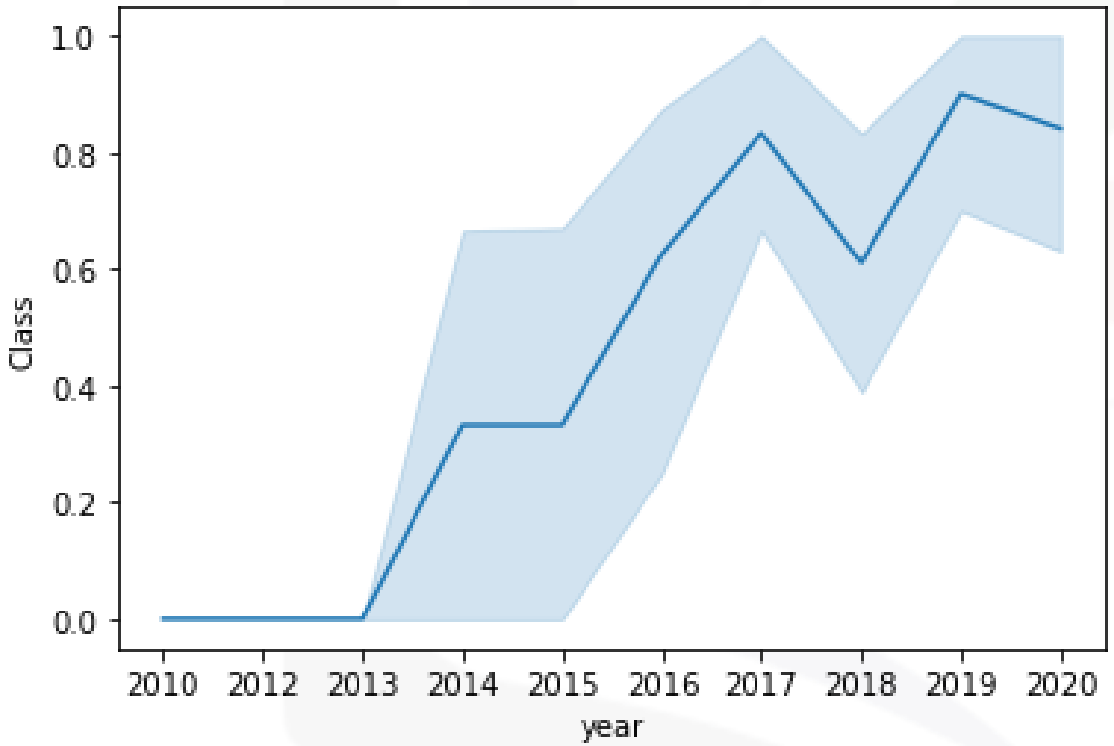
Correlation trends

Relationship between Payload mass and orbit type



Correlation trends

Success rate yearly trend



EDA with SQL

In this notebook, we perform preliminary EDA with some SQL queries. We first upload the data in the database, establish a connection with the database, and then perform our queries, to accomplish things like:

- Listing all unique launch sites
- Listing launch sites that begin with CCA
- Calculating average payload carried by F9
- Getting the date of first successful landing
- Getting total number of success and failure
- Listing booster versions carrying maximum payload

SQL Queries

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

In [3]:
```
%%sql

SELECT DISTINCT LAUNCH_SITE
FROM SPACEX
```

```
 * ibm_db_sa://qkb61028:***@2f3279a5-
Done.
```

Out[3]:

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

SQL Queries

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

In [4]:

```
%%sql

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

```
 * ibm_db_sa://qkb61028:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/bludb
Done.
```

Out[4]:

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

SQL Queries

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

In [11]:
```
%%sql

SELECT SUM(PAYLOAD_MASS__KG_) AS PAYLOAD_MASS_TOTAL
FROM SPACEX
WHERE CUSTOMER = 'NASA (CRS)'
```

```
 * ibm_db_sa://qkb61028:***@2f3279a5-73d1-4859-88f0-a6c3e6b4
Done.
```

Out[11]:

payload_mass_total
45596

SQL Queries

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

In [13]:
```
%%sql

SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD_MASS
FROM SPACEX
WHERE BOOSTER_VERSION = 'F9 v1.1'
```

```
 * ibm_db_sa://qkb61028:***@2f3279a5-73d1-4859-88f0-a
Done.
```

Out[13]:

avg_payload_mass
2928

SQL Queries

5- List the date when the first successful landing outcome in ground pad was achieved

In [19]:

```
%%sql

SELECT MIN(DATE) AS FIRST_GROUND_PAD_SUCCESSFUL_LANDING
FROM SPACEX
WHERE MISSION_OUTCOME = 'Success'
AND LANDING__OUTCOME LIKE '%ground%'
```

```
 * ibm_db_sa://qkb61028:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b90
Done.
```

Out[19]:

first_ground_pad_successful_landing
2015-12-22

SQL Queries

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

In [20]:
```
%%sql

SELECT DISTINCT BOOSTER_VERSION
FROM SPACEX
WHERE MISSION_OUTCOME = 'Success'
AND LANDING__OUTCOME LIKE '%drone ship%'
AND PAYLOAD_MASS__KG_ > 4000
AND PAYLOAD_MASS__KG_ < 6000
```

```
 * ibm_db_sa://qkb61028:***@2f3279a5-73d1-4
Done.
```

Out[20]:

booster_version
F9 FT B1021.2
F9 FT B1031.2
F9 FT B1020
F9 FT B1022
F9 FT B1026

SQL Queries

7- List the total number of successful and failure mission outcomes

In [23]:
```
%%sql

SELECT MISSION_OUTCOME, COUNT(*) AS COUNT
FROM SPACEX
GROUP BY MISSION_OUTCOME
```

```
 * ibm_db_sa://qkb61028:***@2f3279a5-73d1-4859-8
Done.
```

Out[23]:

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

SQL Queries

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

In [28]:
```
%%sql

SELECT DISTINCT BOOSTER_VERSION
FROM SPACEX
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEX)
```

```
 * ibm_db_sa://qkb61028:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cm
Done.
```

Out[28]:

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

SQL Queries

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

In [36]:
```
%%sql

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

```
 * ibm_db_sa://qkb61028:***@2f3279a5-73d1-4859-88f0-a6c3e(
Done.
```

Out[36]:

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

SQL Queries

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

In [53]:
```
%%sql

SELECT * FROM SPACEX where DAYNAME(DATE)='Friday' LIMIT 5
```

```
 * ibm_db_sa://qkb61028:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/bludb
Done.
```

Out[53]:

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)
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt
2014-04-18	19:25:00	F9 v1.1	CCAFS LC-40	SpaceX CRS-3	2296	LEO (ISS)	NASA (CRS)	Success	Controlled (ocean)
2016-03-04	23:35:00	F9 FT B1020	CCAFS LC-40	SES-9	5271	GTO	SES	Success	Failure (drone ship)
2016-04-08	20:43:00	F9 FT B1021.1	CCAFS LC-40	SpaceX CRS-8	3136	LEO (ISS)	NASA (CRS)	Success	Success (drone ship)

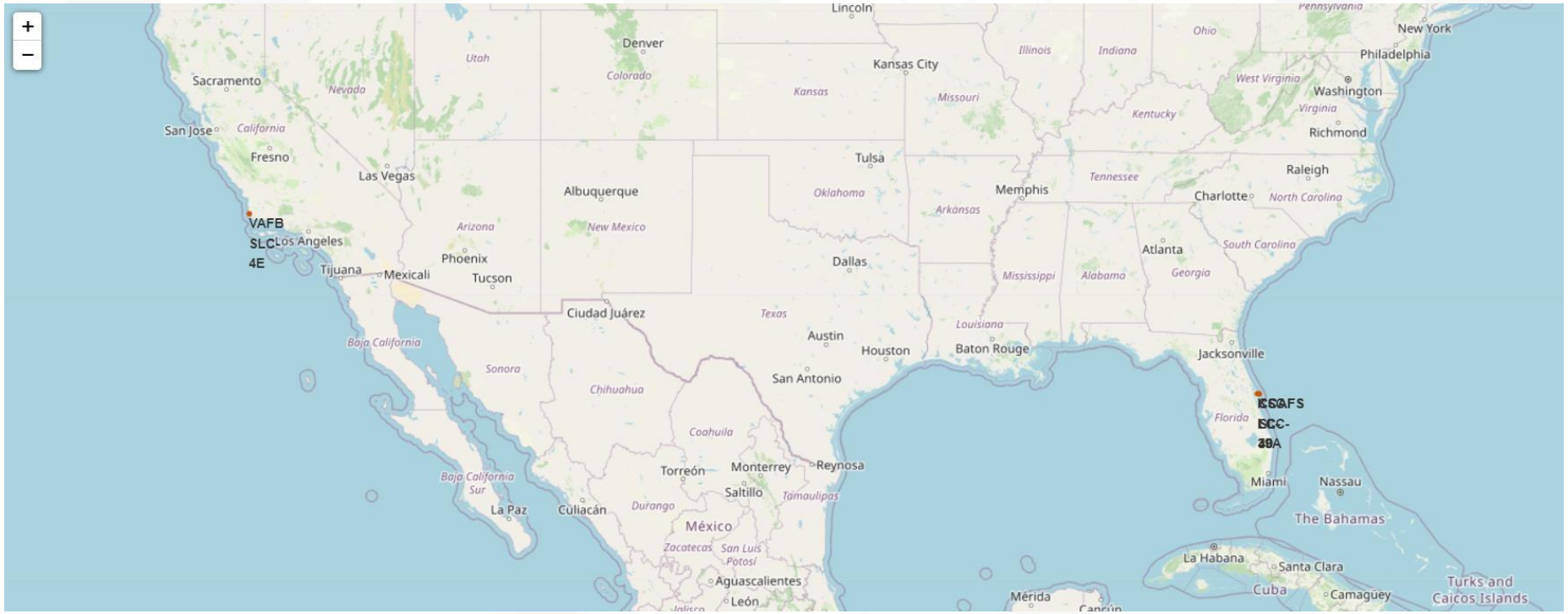
EDA with Dashboards

We use folium to create maps to visualize the launch sites geographically, we also create markers on each site to represent successful and failed landings

We also create an interactive dashboard that has graphs to show the success rate of any site of the user's choosing, and another graph to show the success and failure of different booster types with a payload mass of the user's choosing.

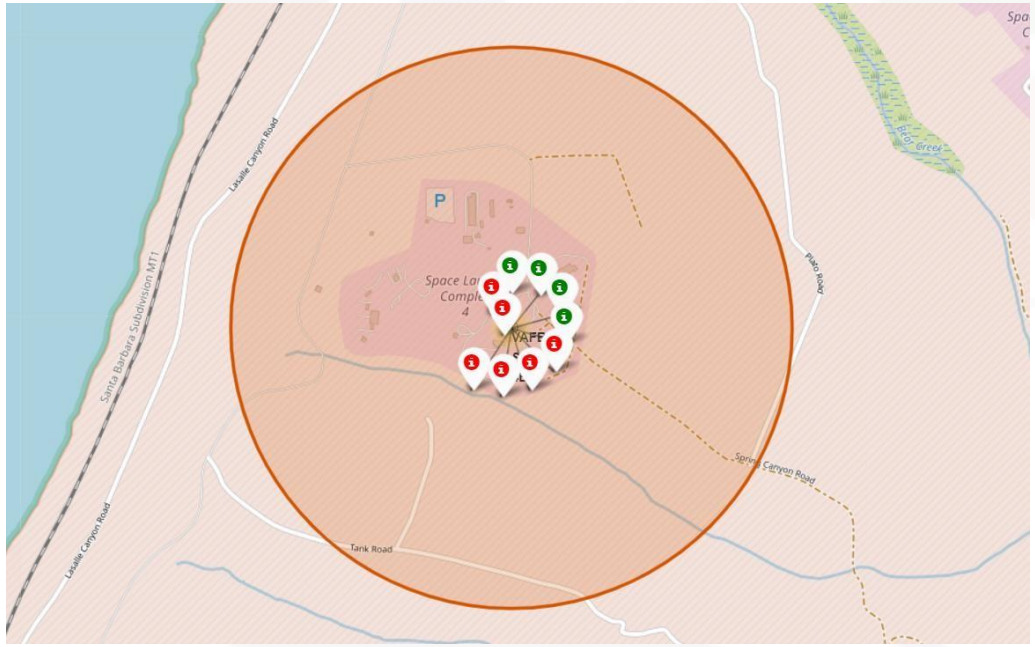
Location characteristics

As we see in the map below, all Launch sites are close to the oceans



Location characteristics

Zooming in on one of the sites, the VAFB SLC 4E to be specific



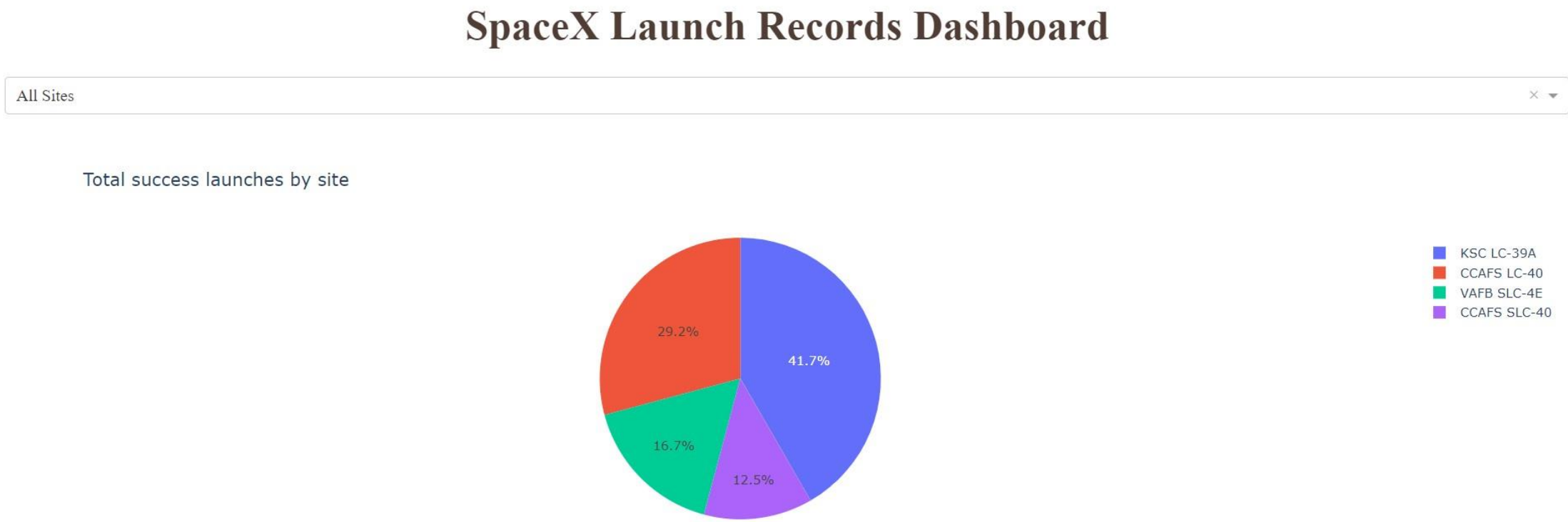
The green marks indicate successful landings, and the red ones indicate failures.

DASHBOARD

The following dashboard presents the relationship between Launch Site, success rate, and payload, with interactive mode so you can focus and customize the graph and relationship you want.

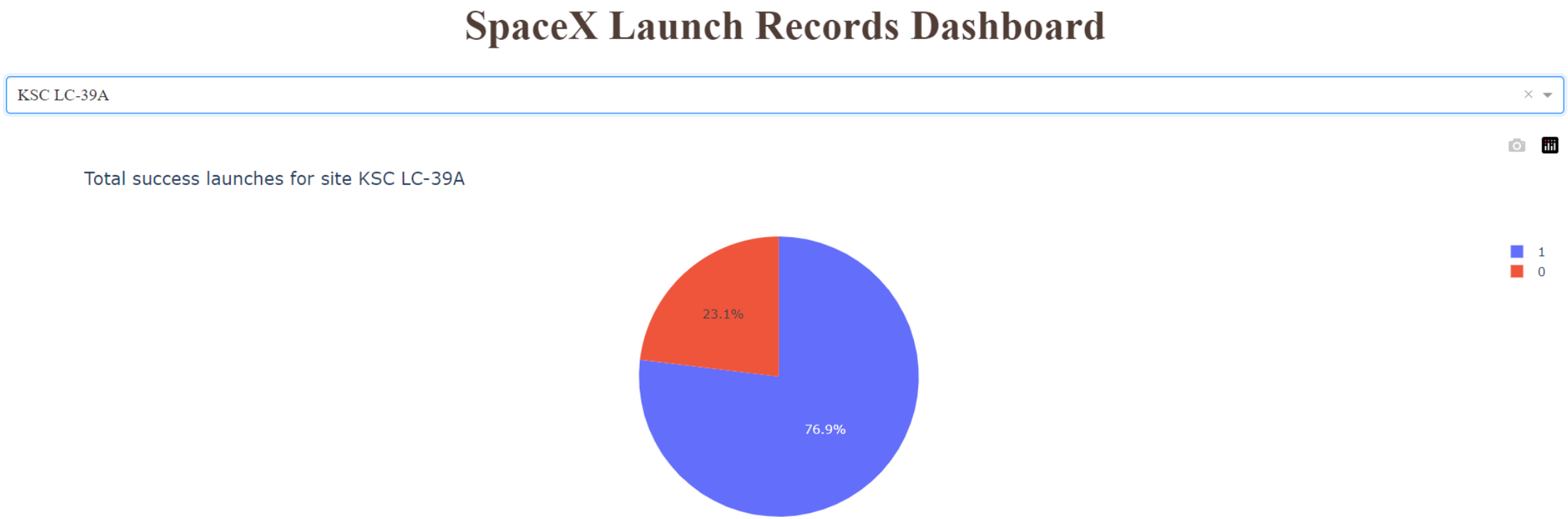
DASHBOARD View 1

Success count for all site



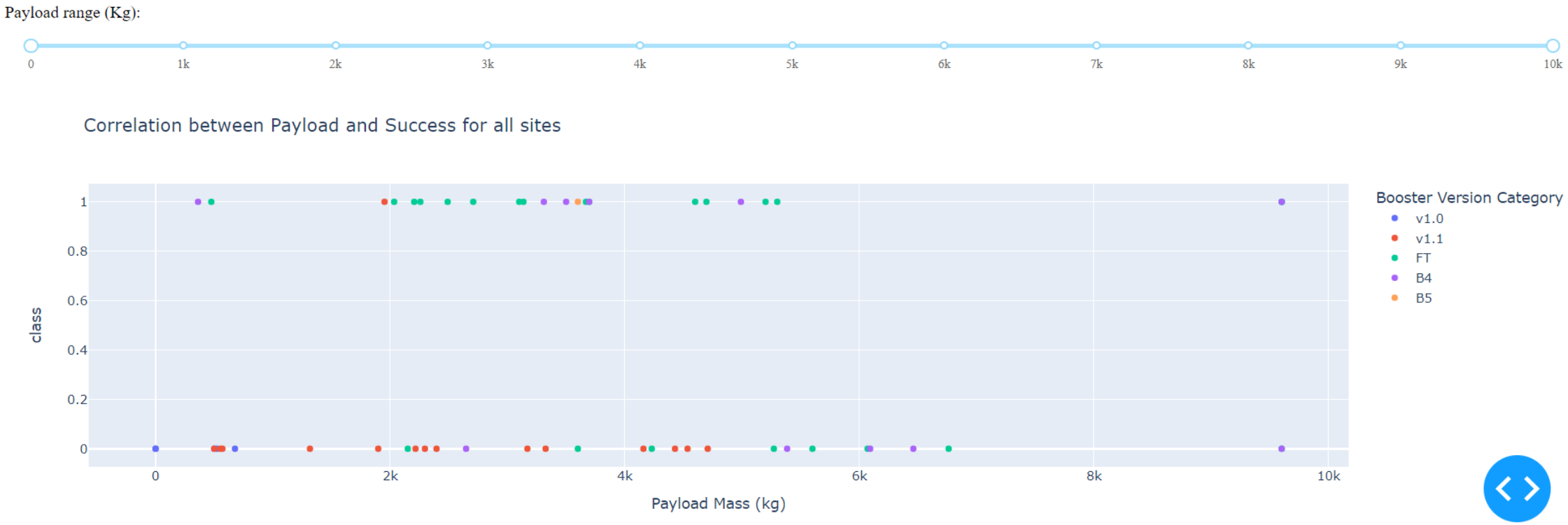
DASHBOARD View 2

Pie chart representing the launch site with the highest success rate



DASHBOARD View 2

Scatter plot of Payload mass vs launch outcome, colored by booster version category



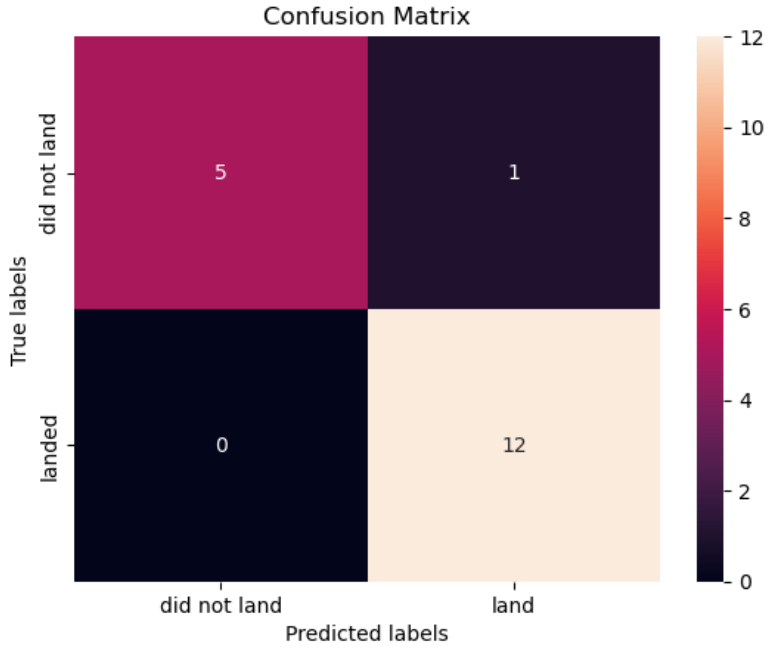
Predictive analysis

We finally build a model using classification algorithms to predict the success or failure of future launches.

We try these algorithms:

- Logistic regression
- Support vector machine
- Decision Tree
- K Nearest Neighbors

Here's the confusion matrix of the predicted values using the best model on test data:



CONCLUSION

- With sufficient data, we can somewhat accurately predict the success or failure of future launches.
- If a competitor wants to enter the market, they should tune to specific payload mass range, stick to certain orbits, and prefer some launch sites to be able to compete with SpaceX.

APPENDIX

Thank you for your attention.

For all the notebooks and files, please visit:

https://github.com/asrf007/capstone-project