DATA SCIENCE
CAPSTONE PROJECT

Abhi Gupta

Outline

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

Executive Summary

Summary of methodologies

- Data collection
- Data wrangling
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly Dash
- Predictive analysis (Classification)

Summary of all results

- Exploratory Data Analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Introduction

Project background and context

SpaceX is the most successful company of the commercial space age, making space travel affordable. The company advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage.

Questions to be answered

- How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?
- Does the rate of successful landings increase over the years?
- What is the best algorithm that can be used for binary classification in this case?

Methodology

Data collection methodology:

- Using SpaceX Rest API
- Using Web Scrapping from Wikipedia

Performed data wrangling

- Filtering the data
- Dealing with missing values
- Using One Hot Encoding to prepare the data to a binary classification

Performed exploratory data analysis (EDA) using visualization and SQL

Performed interactive visual analytics using Folium and Plotly Dash

Performed predictive analysis using classification models

- Building, tuning and evaluation of classification models to ensure the best results

SPACE X

Data collection

Data collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX's Wikipedia entry.

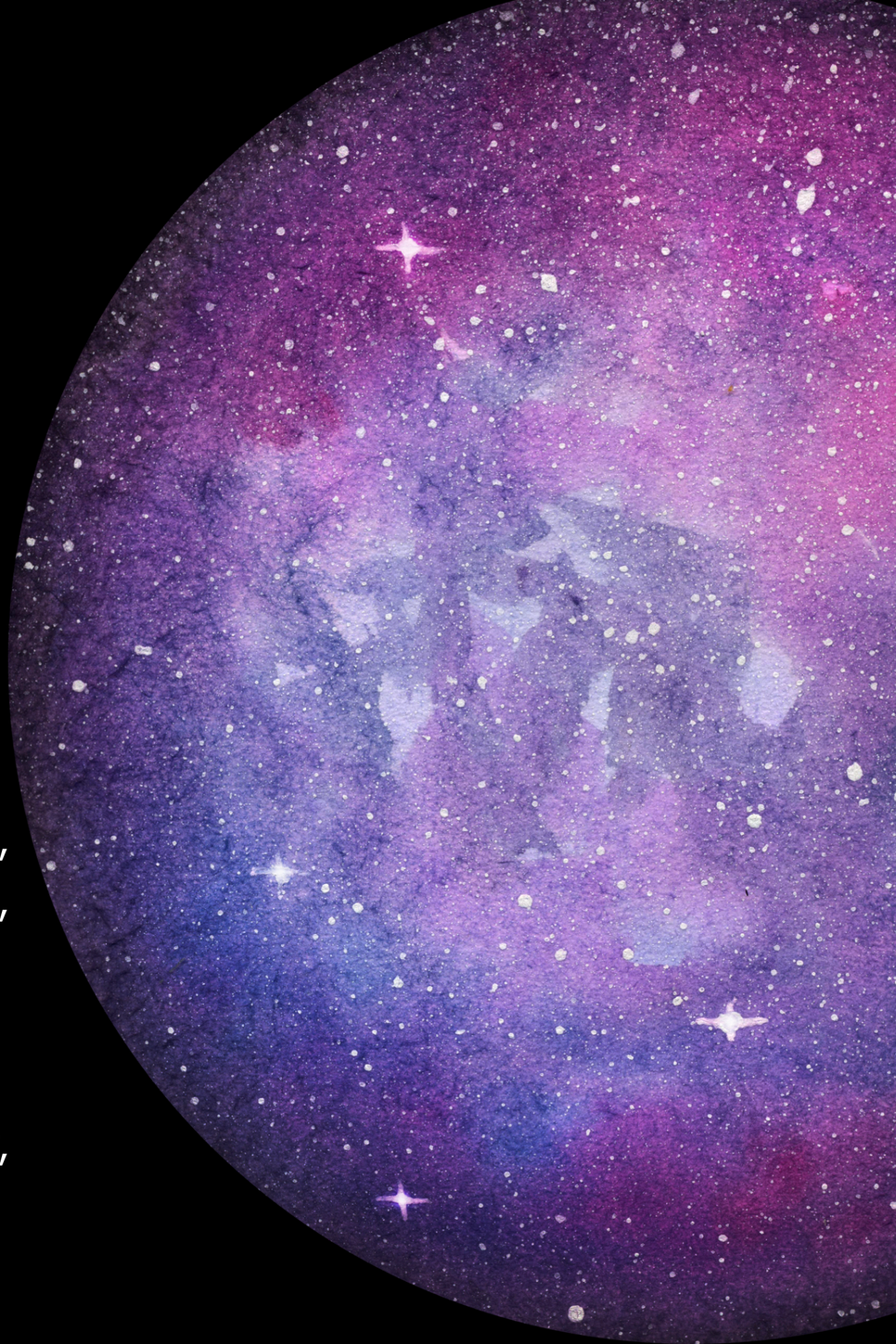
We had to use both of these data collection methods in order to get complete information about the launches for a more detailed analysis.

Data Columns are obtained by using SpaceX REST API:

- FlightNumber,Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount,Serial, Longitude, Latitude

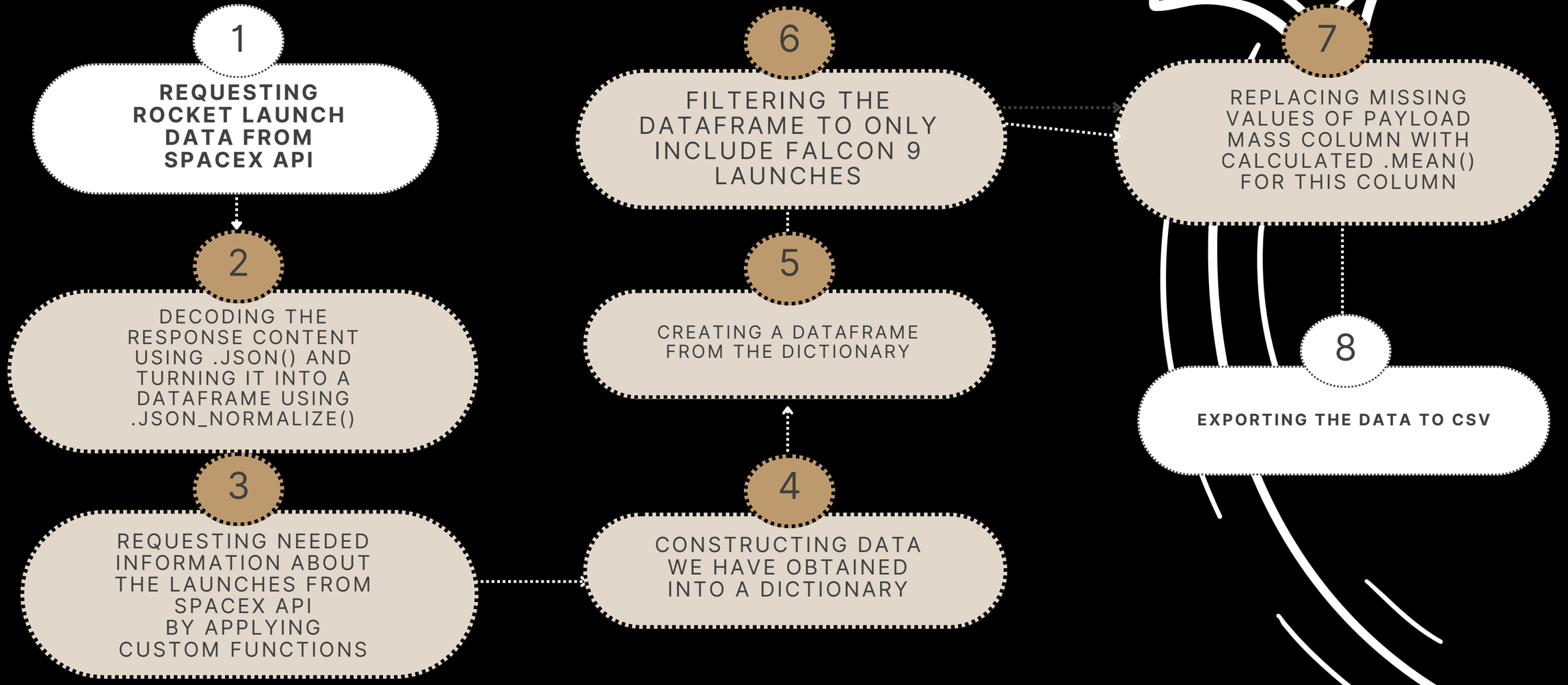
Data Columns are obtained by using Wikipedia Web Scraping:

- Flight No. , Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time



DATA COLLECTION- SPACE X API

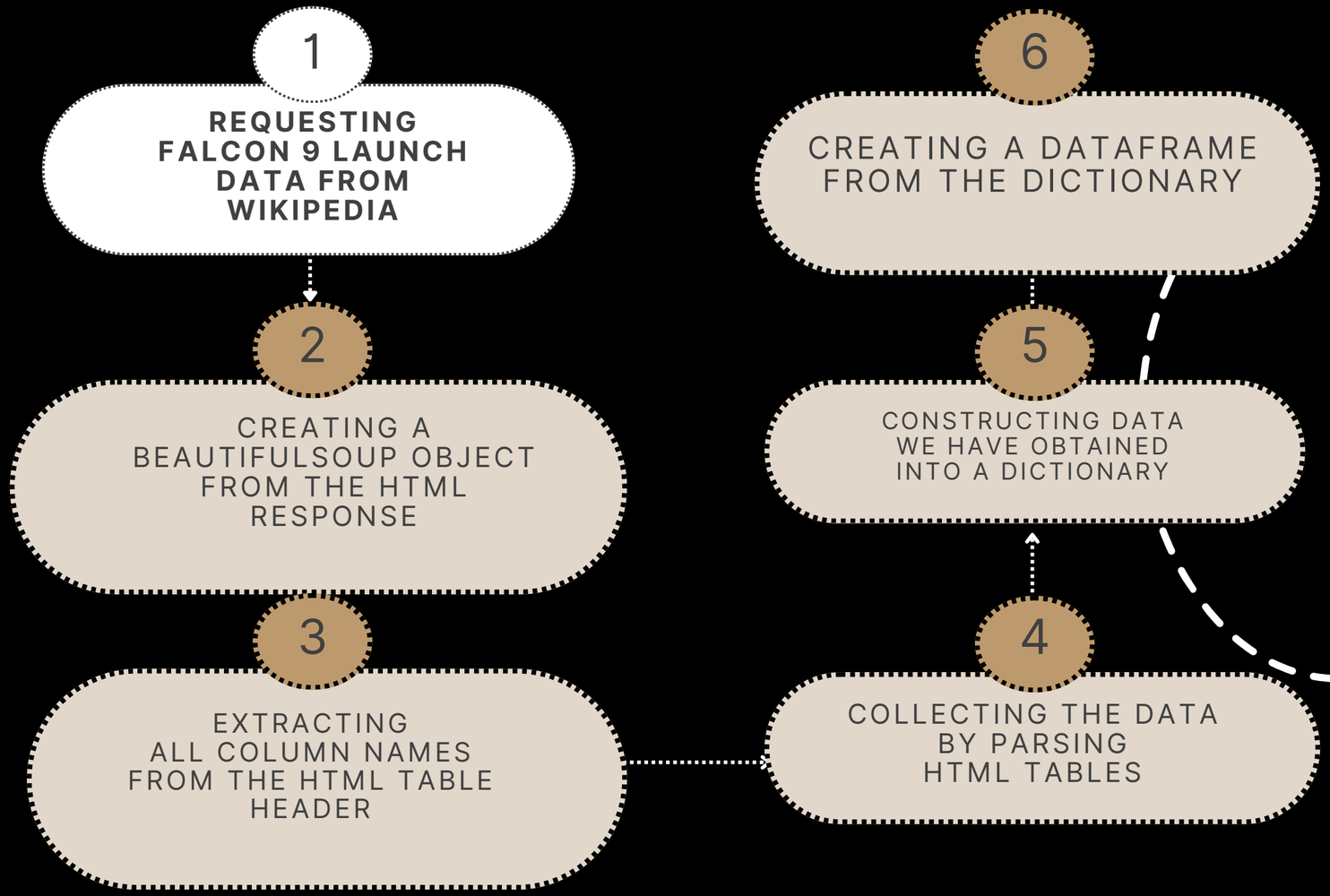
FLOWCHART



GitHub URL: Data Collection API

DATA COLLECTION- Web scraping

FLOWCHART



GitHub URL: Data Collection with Web Scraping

Data wrangling

IN THE DATA SET, THERE ARE SEVERAL DIFFERENT CASES WHERE THE BOOSTER DID NOT LAND SUCCESSFULLY. SOMETIMES A LANDING WAS ATTEMPTED BUT FAILED DUE TO AN ACCIDENT; FOR EXAMPLE, TRUE OCEAN MEANS THE MISSION OUTCOME WAS SUCCESSFULLY LANDED TO A SPECIFIC REGION OF THE OCEAN WHILE FALSE OCEAN MEANS THE MISSION OUTCOME WAS UNSUCCESSFULLY LANDED TO A SPECIFIC REGION OF THE OCEAN. TRUE RTLS MEANS THE MISSION OUTCOME WAS SUCCESSFULLY LANDED TO A GROUND PAD FALSE RTLS MEANS THE MISSION OUTCOME WAS UNSUCCESSFULLY LANDED TO A GROUND PAD.TRUE ASDS MEANS THE MISSION OUTCOME WAS SUCCESSFULLY LANDED ON A DRONE SHIP FALSE ASDS MEANS THE MISSION OUTCOME WAS UNSUCCESSFULLY LANDED ON A DRONE SHIP.

WE MAINLY CONVERT THOSE OUTCOMES INTO TRAINING LABELS WITH“1” MEANS THE BOOSTER SUCCESSFULLY LANDED, “0” MEANS IT WAS UNSUCCESSFUL.

GITHUB URL: DATA WRANGLING

EDA with Data Visualization

CHARTS WERE PLOTTED:
FLIGHT NUMBER VS. PAYLOAD MASS, FLIGHT NUMBER VS. LAUNCH SITE, PAYLOAD MASS VS. LAUNCH SITE, ORBIT TYPE VS. SUCCESS RATE, FLIGHT NUMBER VS. ORBIT TYPE, PAYLOAD MASS VS ORBIT TYPE AND SUCCESS RATE YEARLY TREND.

SCATTER PLOTS SHOW THE RELATIONSHIP BETWEEN VARIABLES. IF A RELATIONSHIP EXISTS, THEY COULD BE USED IN MACHINE LEARNING MODEL.

BAR CHARTS SHOW COMPARISONS AMONG DISCRETE CATEGORIES. THE GOAL IS TO SHOW THE RELATIONSHIP BETWEEN THE SPECIFIC CATEGORIES BEING COMPARED AND A MEASURED VALUE.

LINE CHARTS SHOW TRENDS IN DATA OVER TIME (TIME SERIES).

GITHUB URL: EDA WITH DATA VISUALIZATION

EDA with SQL

PERFORMED SQL QUERIES:

- DISPLAYING THE NAMES OF THE UNIQUE LAUNCH SITES IN THE SPACE MISSION
- DISPLAYING 5 RECORDS WHERE LAUNCH SITES BEGIN WITH THE STRING ‘CCA'
- DISPLAYING THE TOTAL PAYLOAD MASS CARRIED BY BOOSTERS LAUNCHED BY NASA (CRS)
- DISPLAYING AVERAGE PAYLOAD MASS CARRIED BY BOOSTER VERSION F9 V1.1
- LISTING THE DATE WHEN THE FIRST SUCCESSFUL LANDING OUTCOME IN GROUND PAD WAS ACHIEVED
- LISTING THE NAMES OF THE BOOSTERS WHICH HAVE SUCCESS IN DRONE SHIP AND HAVE PAYLOAD MASS GREATER THAN 4000 BUT
- LESS THAN 6000
- LISTING THE TOTAL NUMBER OF SUCCESSFUL AND FAILURE MISSION OUTCOMES
- LISTING THE NAMES OF THE BOOSTER VERSIONS WHICH HAVE CARRIED THE MAXIMUM PAYLOAD MASS
- LISTING THE FAILED LANDING OUTCOMES IN DRONE SHIP, THEIR BOOSTER VERSIONS AND LAUNCH SITE NAMES FOR THE MONTHS IN
- YEAR 2015
- RANKING 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

GITHUB URL: EDA WITH SQL

Build an interactive map with Folium

MARKERS OF ALL LAUNCH SITES:

- ADDED MARKER WITH CIRCLE, POPUP LABEL AND TEXT LABEL OF NASA JOHNSON SPACE CENTER USING ITS LATITUDE AND LONGITUDE COORDINATES AS A START LOCATION.
- ADDED MARKERS WITH CIRCLE, POPUP LABEL AND TEXT LABEL OF ALL LAUNCH SITES USING THEIR LATITUDE AND LONGITUDE COORDINATES TO SHOW THEIR GEOGRAPHICAL LOCATIONS AND PROXIMITY TO EQUATOR AND COASTS.

COLOURED MARKERS OF THE LAUNCH OUTCOMES FOR EACH LAUNCH SITE:

- ADDED COLOURED MARKERS OF SUCCESS (GREEN) AND FAILED (RED) LAUNCHES USING MARKER CLUSTER TO IDENTIFY WHICH LAUNCH SITES HAVE RELATIVELY HIGH SUCCESS RATES.

DISTANCES BETWEEN A LAUNCH SITE TO ITS PROXIMITIES:

- ADDED COLOURED LINES TO SHOW DISTANCES BETWEEN THE LAUNCH SITE KSC LC-39A (AS AN EXAMPLE) AND ITS PROXIMITIES LIKE RAILWAY, HIGHWAY, COASTLINE AND CLOSEST CITY.

GITHUB URL: INTERACTIVE VISUAL ANALYTICS WITH FOLIUM

Build a Dashboard with Plotly Dash

LAUNCH SITES DROPDOWN LIST:

- ADDED A DROPDOWN LIST TO ENABLE LAUNCH SITE SELECTION.

PIE CHART SHOWING SUCCESS LAUNCHES (ALL SITES/CERTAIN SITE):

- ADDED A PIE CHART TO SHOW THE TOTAL SUCCESSFUL LAUNCHES COUNT FOR ALL SITES AND THE SUCCESS VS. FAILED COUNTS FOR THE SITE, IF A SPECIFIC LAUNCH SITE WAS SELECTED.

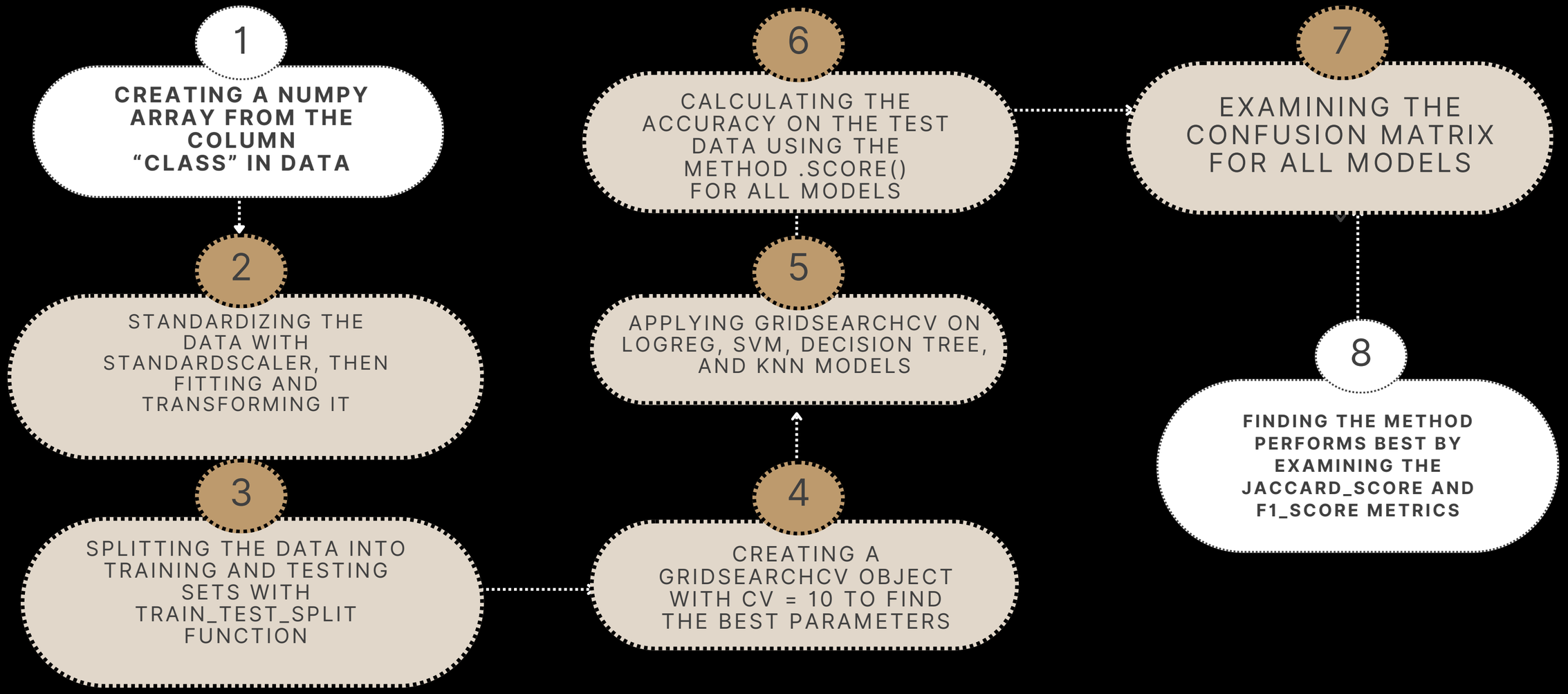
SLIDER OF PAYLOAD MASS RANGE:

- ADDED A SLIDER TO SELECT PAYLOAD RANGE. SCATTER CHART OF PAYLOAD MASS VS. SUCCESS RATE FOR THE DIFFERENT BOOSTER VERSIONS:
- ADDED A SCATTER CHART TO SHOW THE CORRELATION BETWEEN PAYLOAD AND LAUNCH SUCCESS.

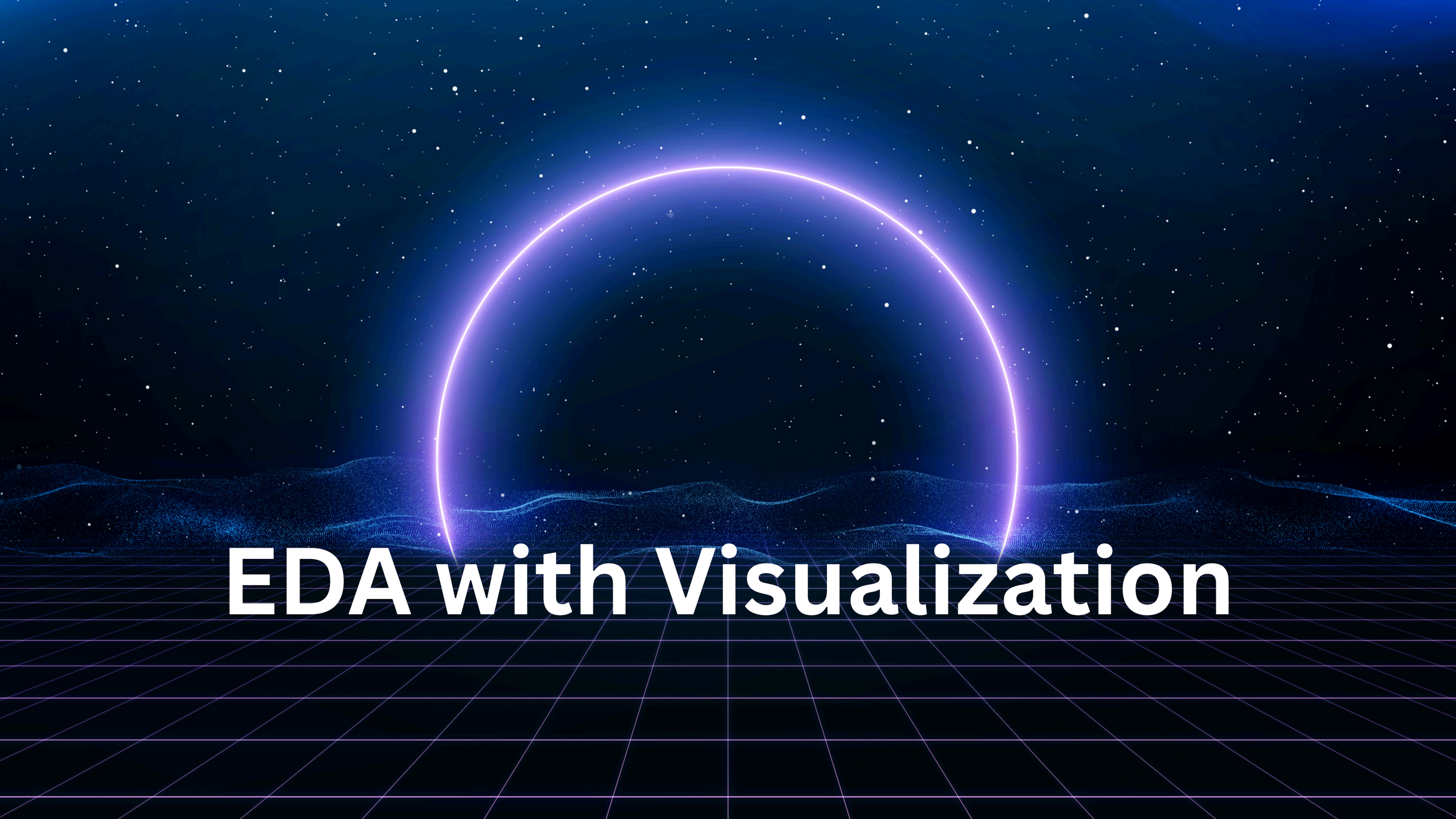
GITHUB URL: SPACEX DASH APP

Predictive analysis (Classification)

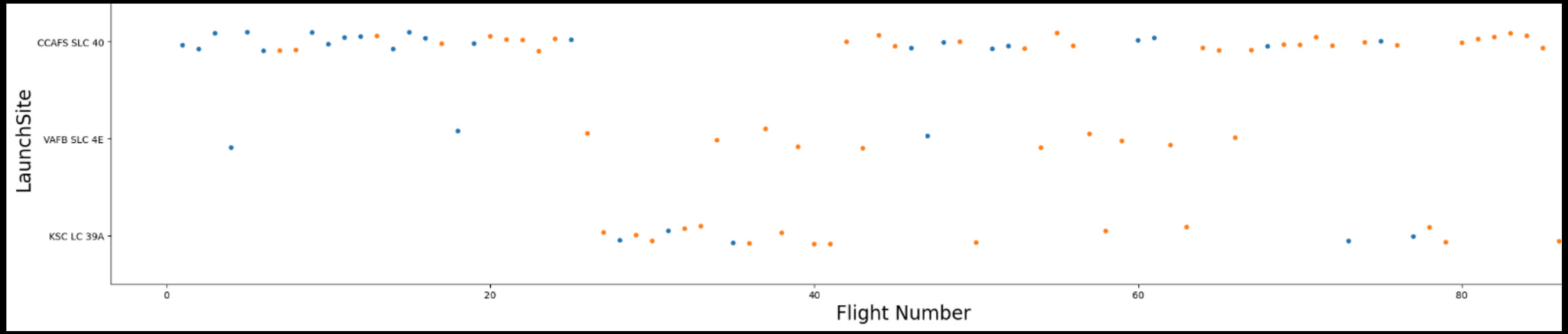
FLOWCHART



GitHub URL: Machine Learning Prediction


EDA with Visualization

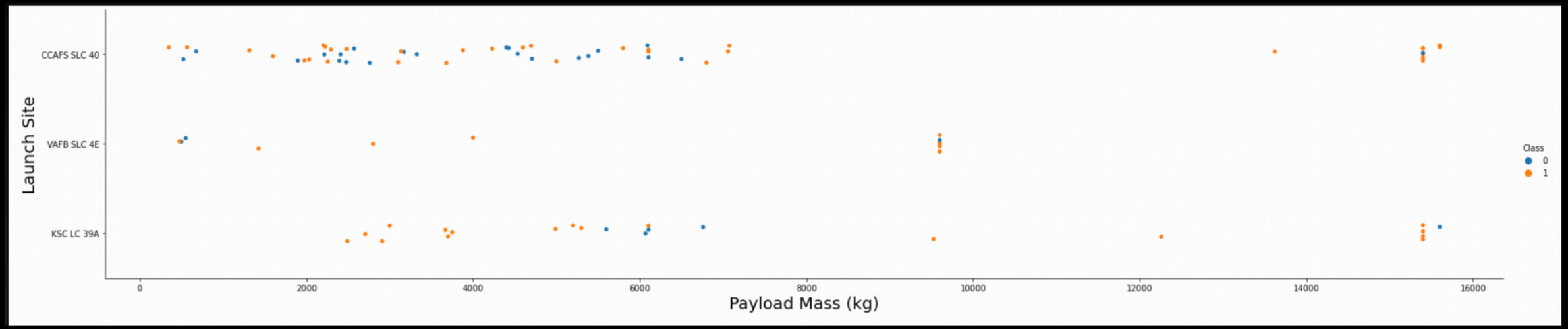
Flight Number vs. Launch Site



Explanation:

- The earliest flights all failed while the latest flights all succeeded.
- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- It can be assumed that each new launch has a higher rate of success.

Payload vs. Launch Site



Explanation:

- For every launch site the higher the payload mass, the higher the success rate.
- Most of the launches with payload mass over 7000 kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.

Success rate vs. Orbit type

Explanation:

• Orbits with 100% success rate:

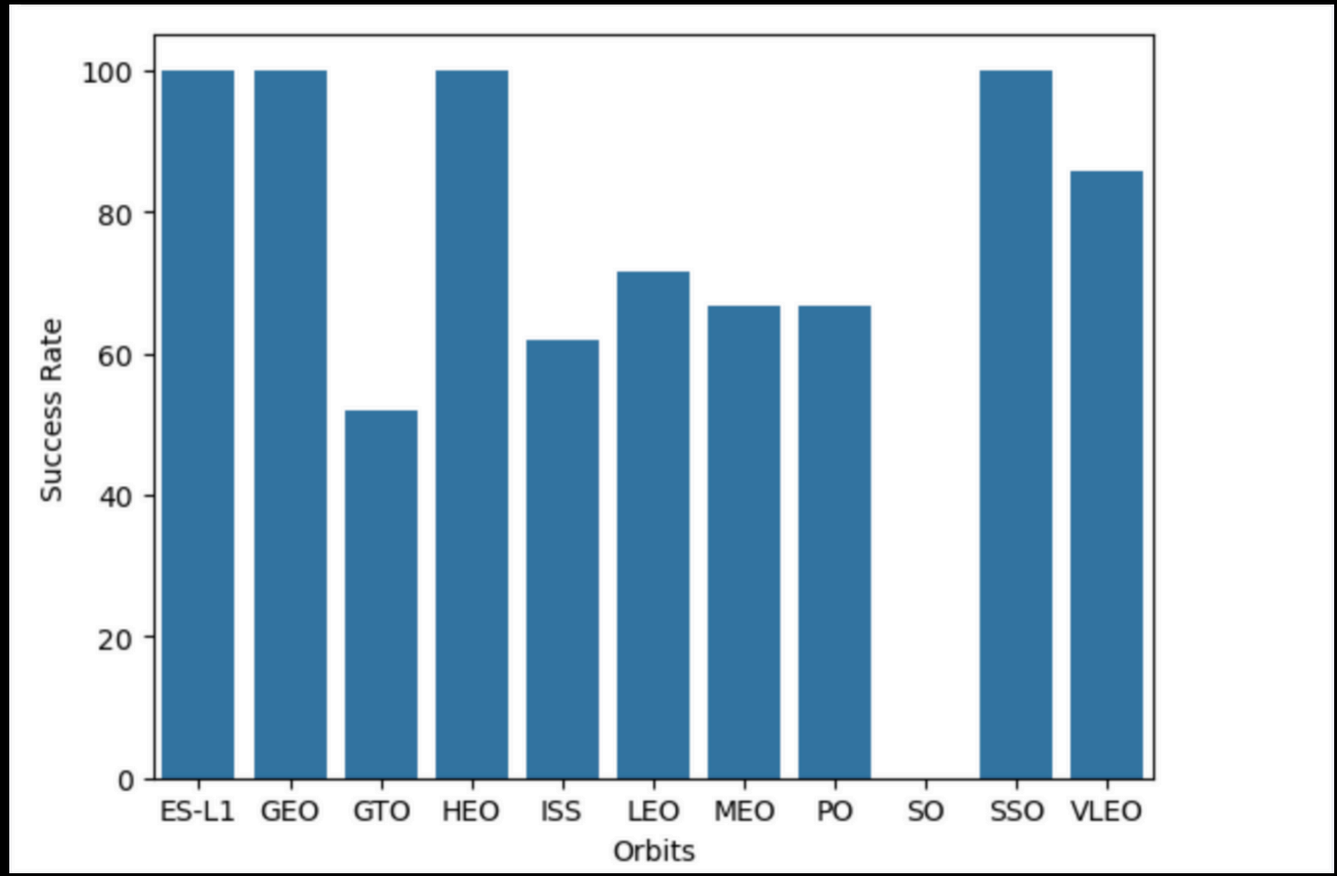
- ES-L1, GEO, HEO, SSO

• Orbits with 0% success rate:

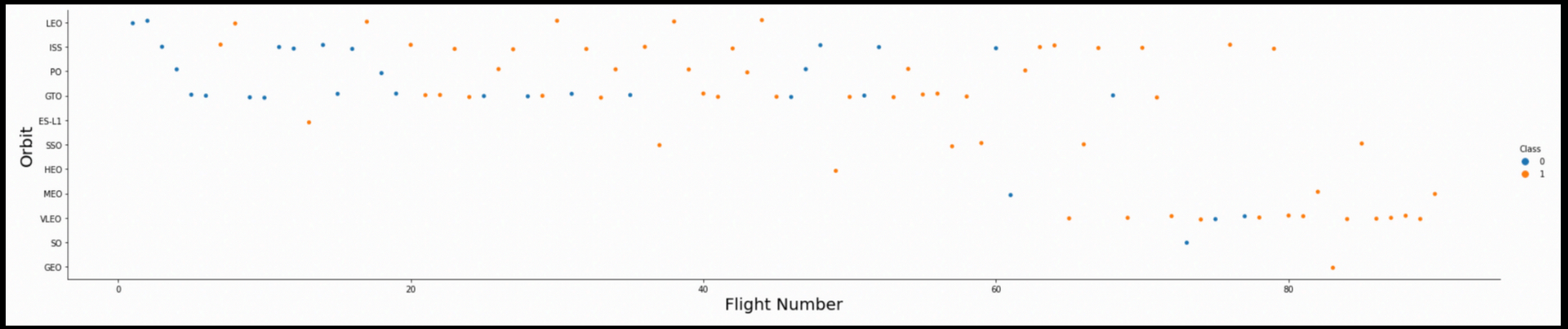
- SO

• Orbits with success rate between 50% and 85%:

- GTO, ISS, LEO, MEO, PO



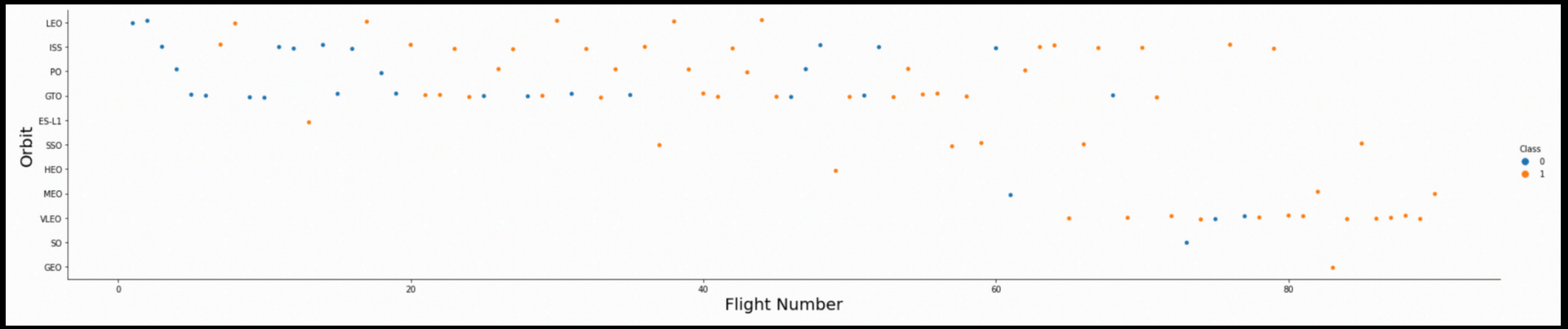
Flight Number vs. Orbit type



Explanation:

• In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

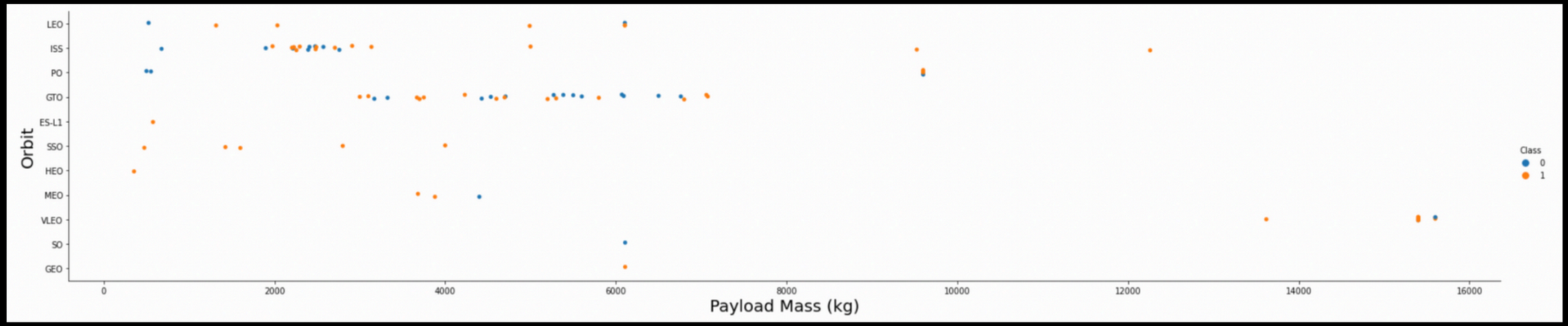
Flight Number vs. Orbit type



Explanation:

• In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

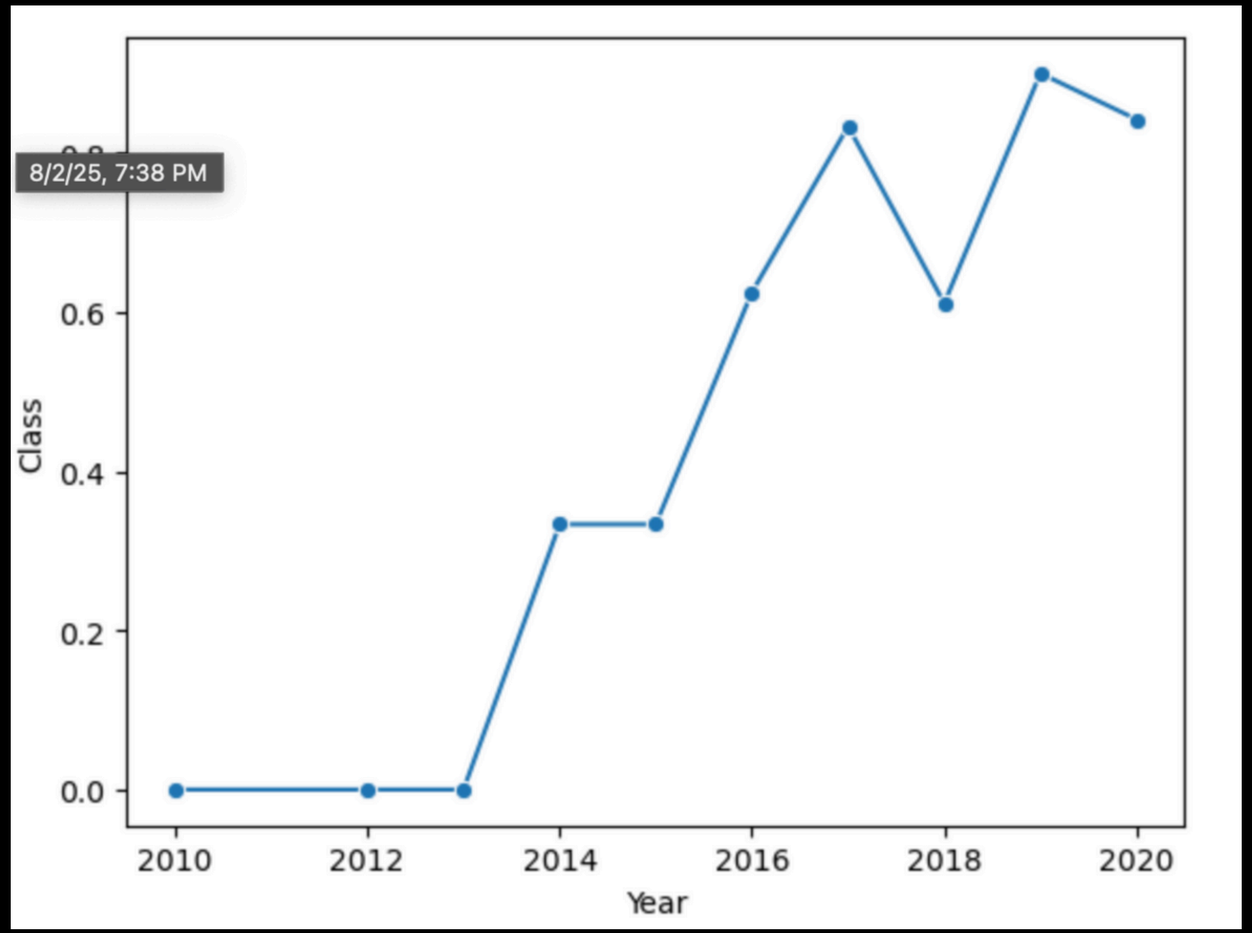
Payload Mass vs. Orbit type



Explanation:

• Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.

Launch success yearly trend



Explanation:

- The success rate since 2013 kept increasing till 2020.

EDA WITH SQL

All launch site names

```
%sql select distinct(Launch_Site) from SPACEXTABLE;
```

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

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

Launch site names begin with `CCA`

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

```
%sql select * from SPACEXTABLE where Launch_Site like "CCA%" limit 5;
```

```
 * sqlite:///my_data1.db
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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

Total payload mass

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

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer = 'NASA (CRS)';
```

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

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

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

```
%sql select round(avg(PAYLOAD_MASS__KG_),2) as Avg_payload_mass from SPACEXTABLE where Booster_Version like 'F9 v1.1%';
```

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

Avg_payload_mass
2534.67

First Successful Ground Landing Date

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

Hint:Use min function

```
%sql select min(Date) from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)';
```

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

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
# df.head()
%sql select * from SPACEXTABLE where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ < 6000 and PAYLOAD_MASS__KG_ > 4000;
# df['Landing_Outcome'].value_counts()
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2016-05-06	5:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2016-08-14	5:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-10-11	22:53:00	F9 FT B1031.2	KSC LC-39A	SES-11 / EchoStar 105	5200	GTO	SES EchoStar	Success	Success (drone ship)

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

```
# df.head()
%sql select Booster_Version from SPACEXTABLE where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ < 6000 and PAYLOAD_MASS__KG_ > 4000;
# df['Landing_Outcome'].value_counts()
```

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

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

Total number of successful and failure mission outcomes

```
%%sql
SELECT
    SUM(CASE WHEN Mission_Outcome LIKE 'Success%' THEN 1 ELSE 0 END) AS success_count
    SUM(CASE WHEN Mission_Outcome LIKE 'Failure%' THEN 1 ELSE 0 END) AS failure_count
FROM SPACEXTABLE;
```

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

success_count	failure_count
100	1

Boosters carried maximum payload

```
%sql select Booster_Version from spacextable where PAYLOAD_MASS__KG_ = 15600;
```

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

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

2015 launch records

```
%sql select substr(Date , 6,2) as Month , date , Booster_Version , Launch_Site ,Landing_Outcome from SPACEXTABLE where Landing_Outcome like 'failure%' and substr(Date , 0,5) = '2015';
```

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

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

Rank success count between 2010-06-04 and 2017-03-20

```
%%sql
SELECT landing_outcome,
       COUNT(*) AS count_outcomes
FROM SPACEXTABLE
WHERE date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY landing_outcome
ORDER BY count_outcomes DESC;
```

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

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

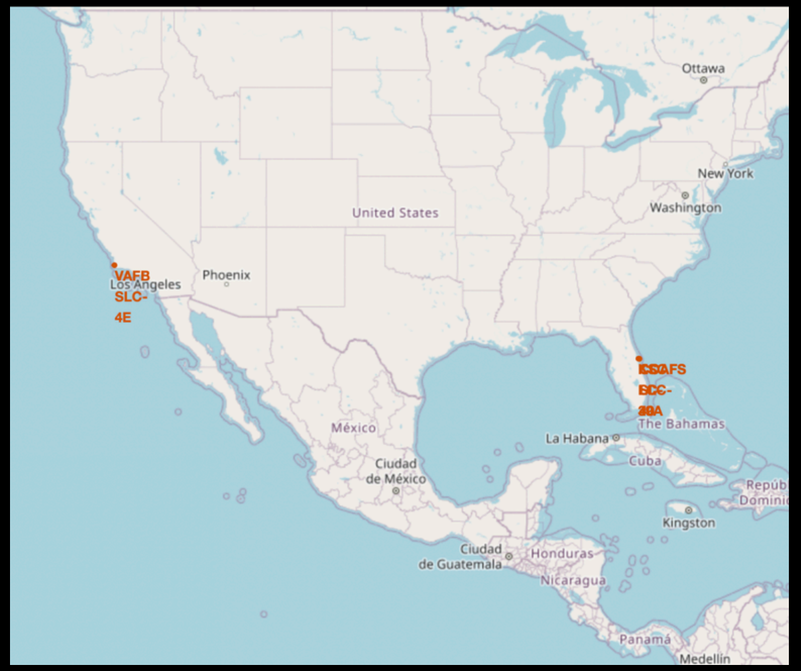

Interactive map with Folium

All launch sites’ location markers on a global map

Explanation:

• Most of Launch sites are in proximity to the Equator line. The land is moving faster at
the equator than any other place on the surface of the Earth Anything on the surface of the Earth at the equator is already moving at 1670 km/hour. If a ship is launched from the equator it goes up into space, and it is also moving around the Earth at the same speed it was moving before launching. This is because of inertia. This speed will help the spacecraft keep up a good enough speed to stay in orbit.

• All launch sites are in very close proximity
to the coast, while launching rockets
towards the ocean it minimises the risk of
having any debris dropping or exploding
near people.



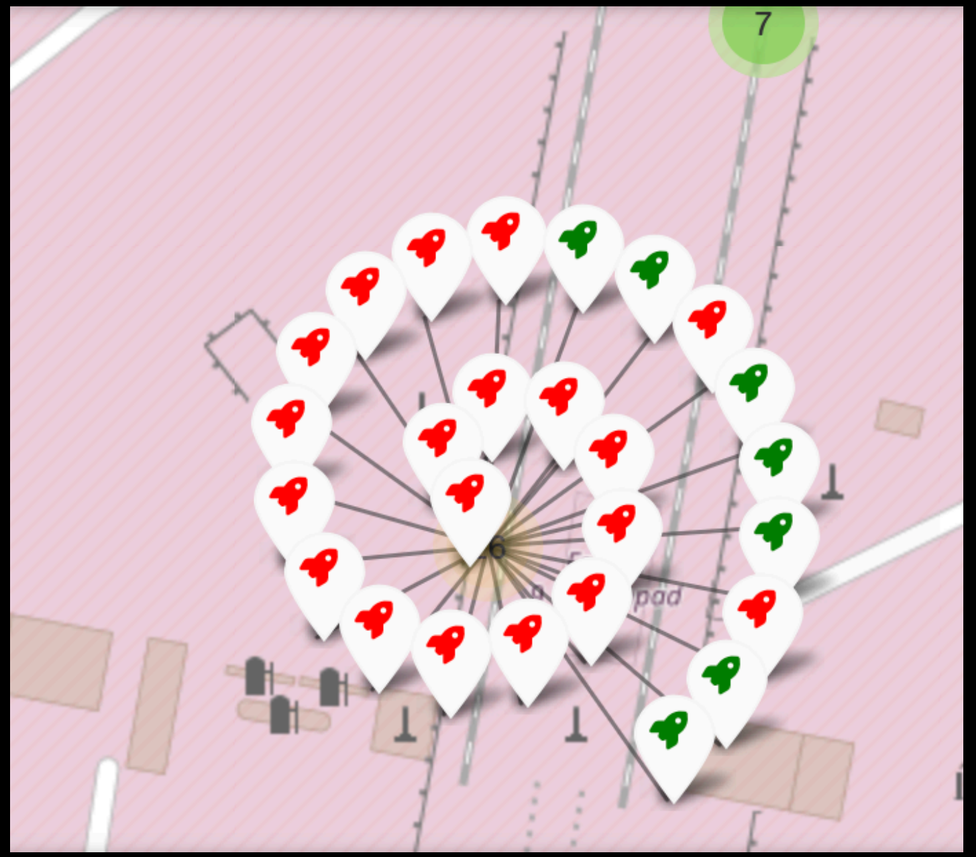
Colour-labeled launch records on the map

Explanation:

• From the colour-labeled markers we should be able to easily identify which launch sites have relatively high success rates.

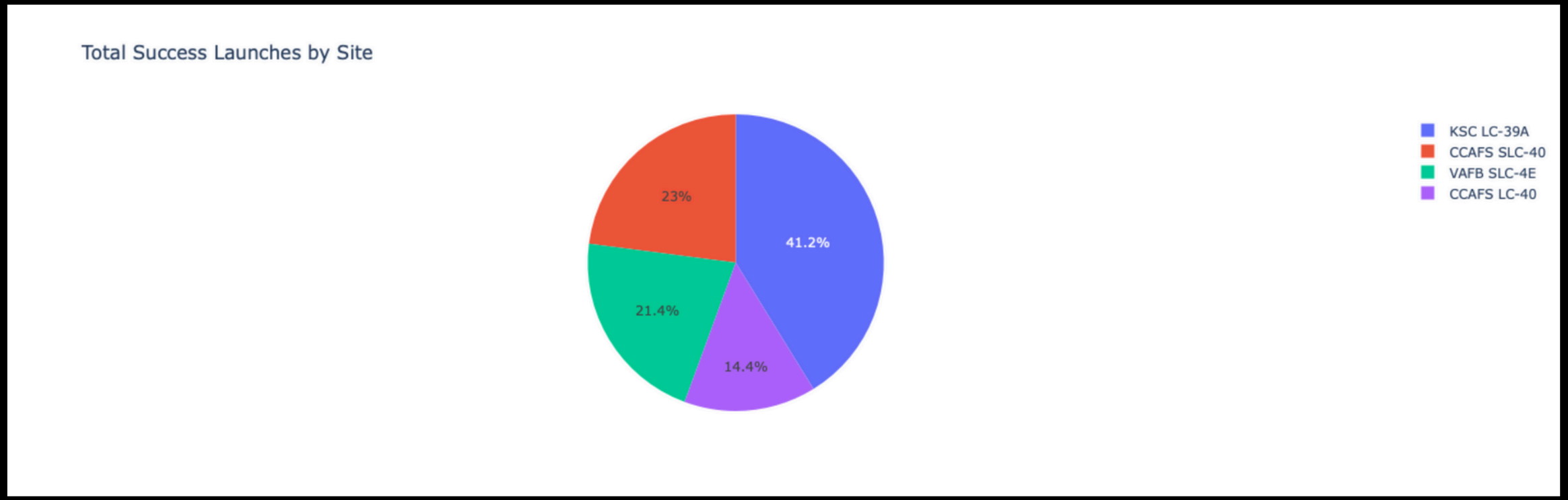
- Green Marker = Successful Launch
- Red Marker = Failed Launch

• Launch Site KSC LC-39A has a very high Success Rate.

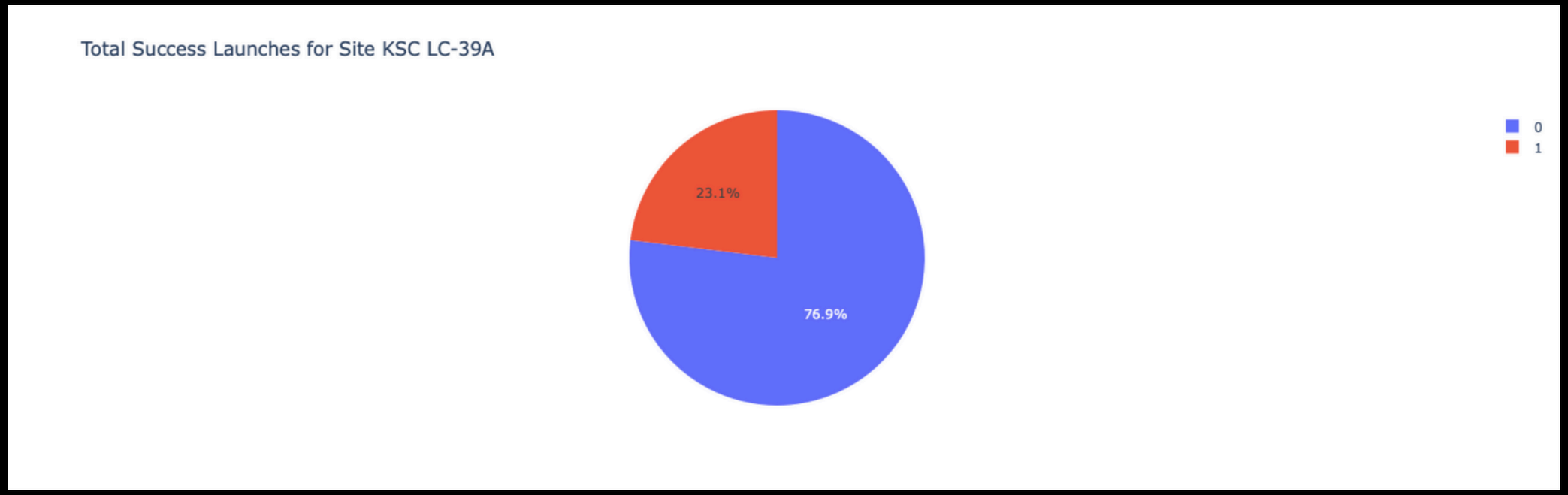


Build a Dashboard with Plotly Dash

Launch success count for all sites



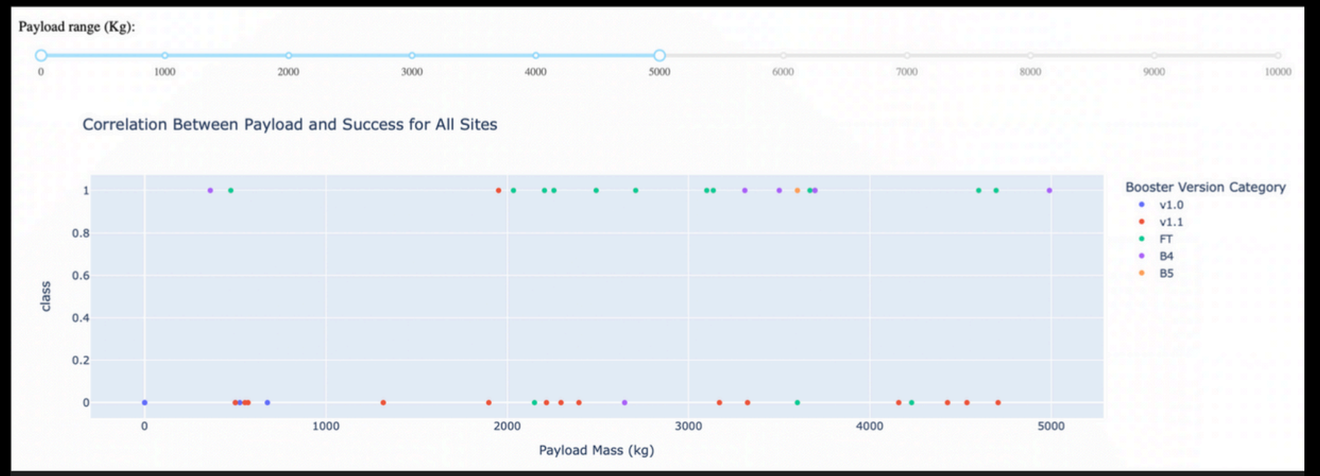
Launch site with highest launch success ratio

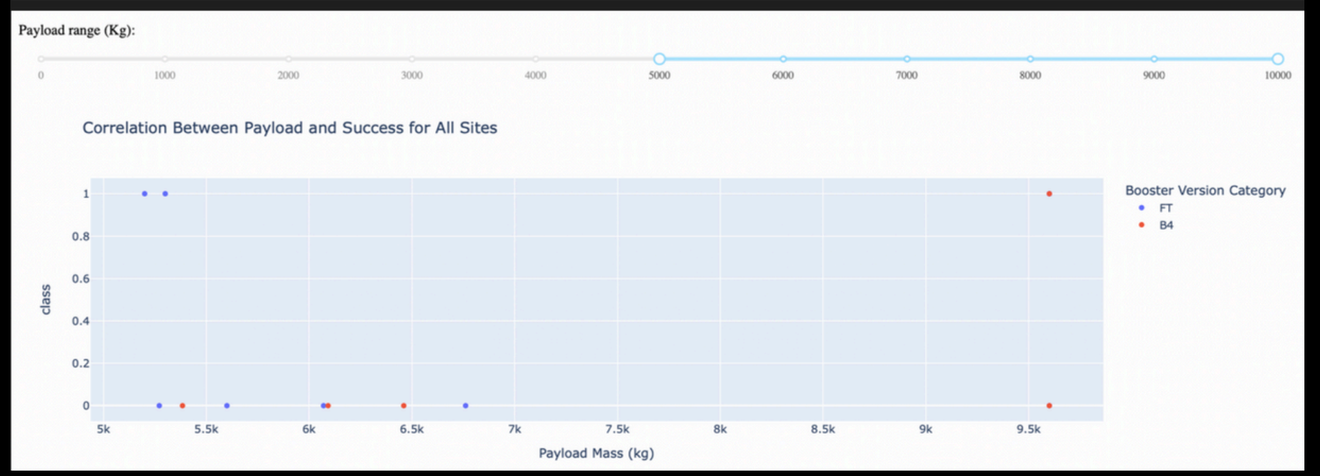


Payload Mass vs. Launch Outcome for all sites

Explanation:

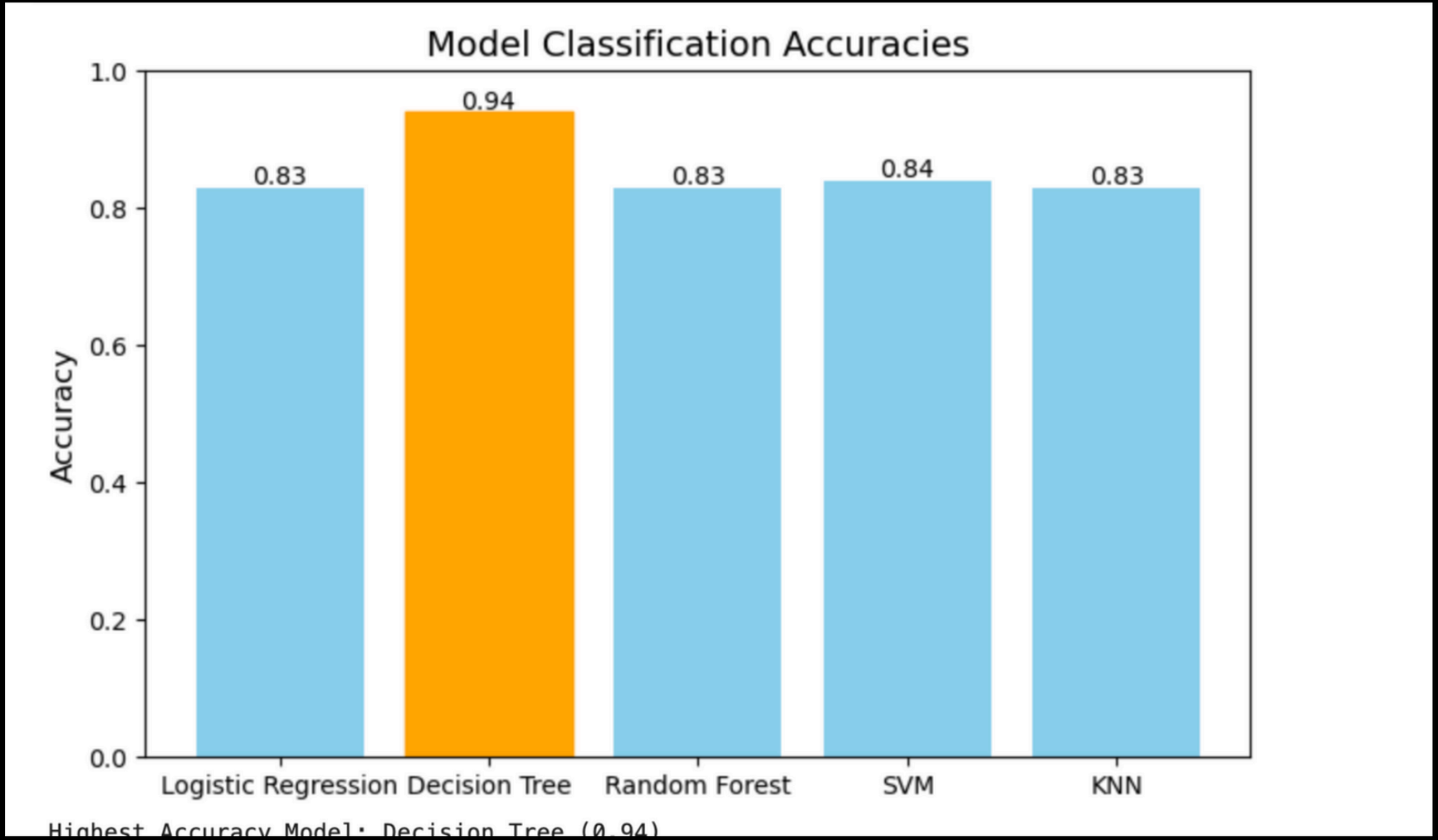
- The charts show that payloads between 2000 and 5500 kg have the highest success rate.



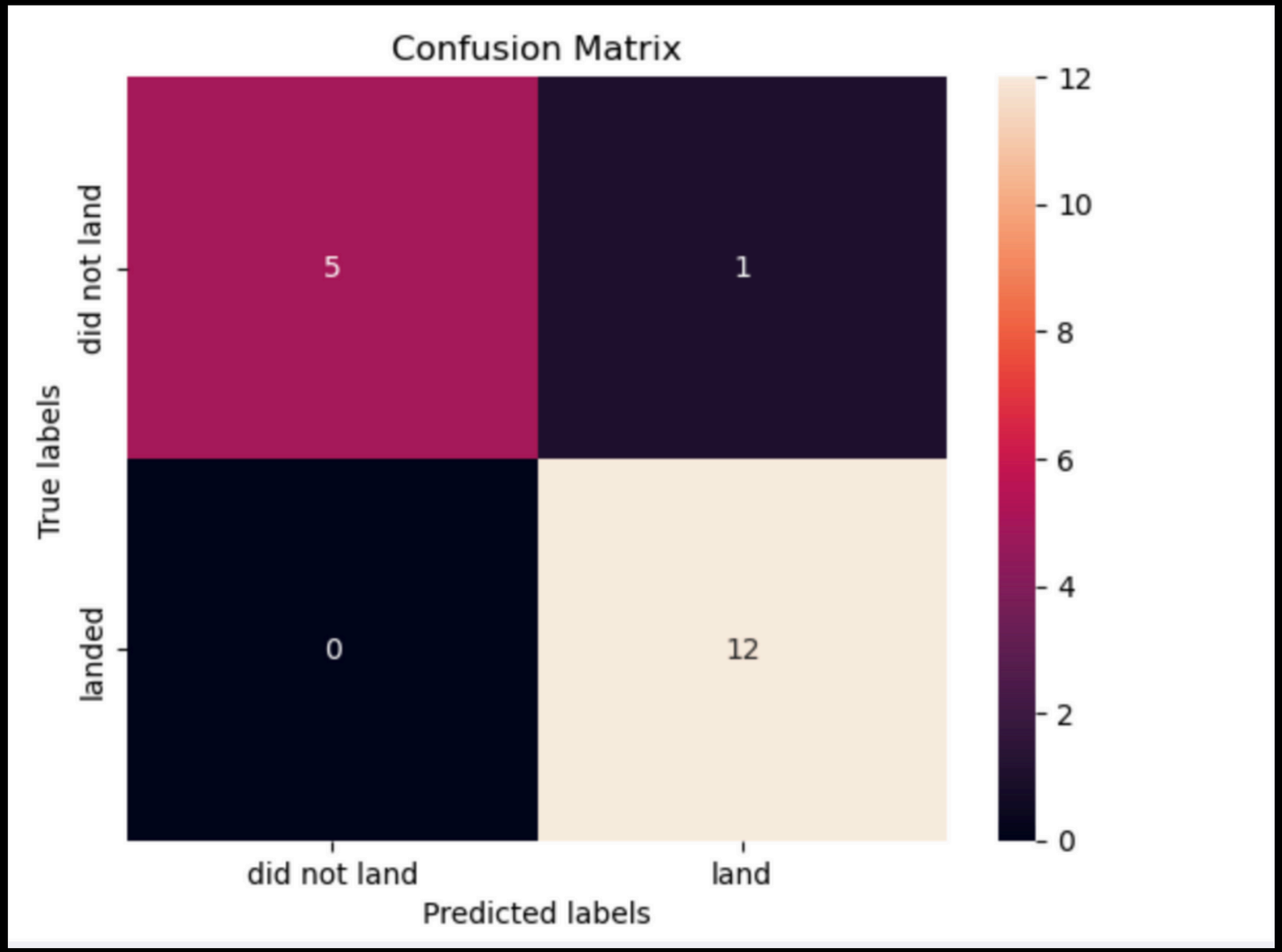


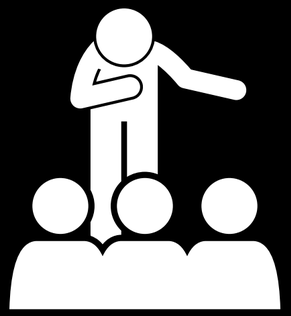
Predictive analysis (Classification)

Classification Accuracy



Confusion Matrix





Conclusion

- Decision Tree Model is the best algorithm for this dataset.
- Launches with a low payload mass show better results than launches with a larger payload mass.
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- The success rate of launches increases over the years.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.