IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ananthu Samanth Kumar

06 April 2023

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

Summary of Methodology

- Data Collection using API
- Data Collection with web scraping
- Data Wrangling
- Exploratory Data Analysis with SQL
- Exploratory Data Analysis with Data Visualization
- Interactive Visual Analytics with Folium
- Machine Learning Prediction

Introduction

- In this capstone I will take the role of a data scientist working for a new rocket company
- to determine the price of each launch of SPACE X.
- Do this by gathering information about Space X and creating dashboards for the team.
- Determine if SpaceX will reuse the first stage.
- Instead of using rocket science to determine if the first stage will land successfully, you will train a machine learning model and use public information to predict if SpaceX will reuse the first stage.

Section 1

Methodology

Methodology

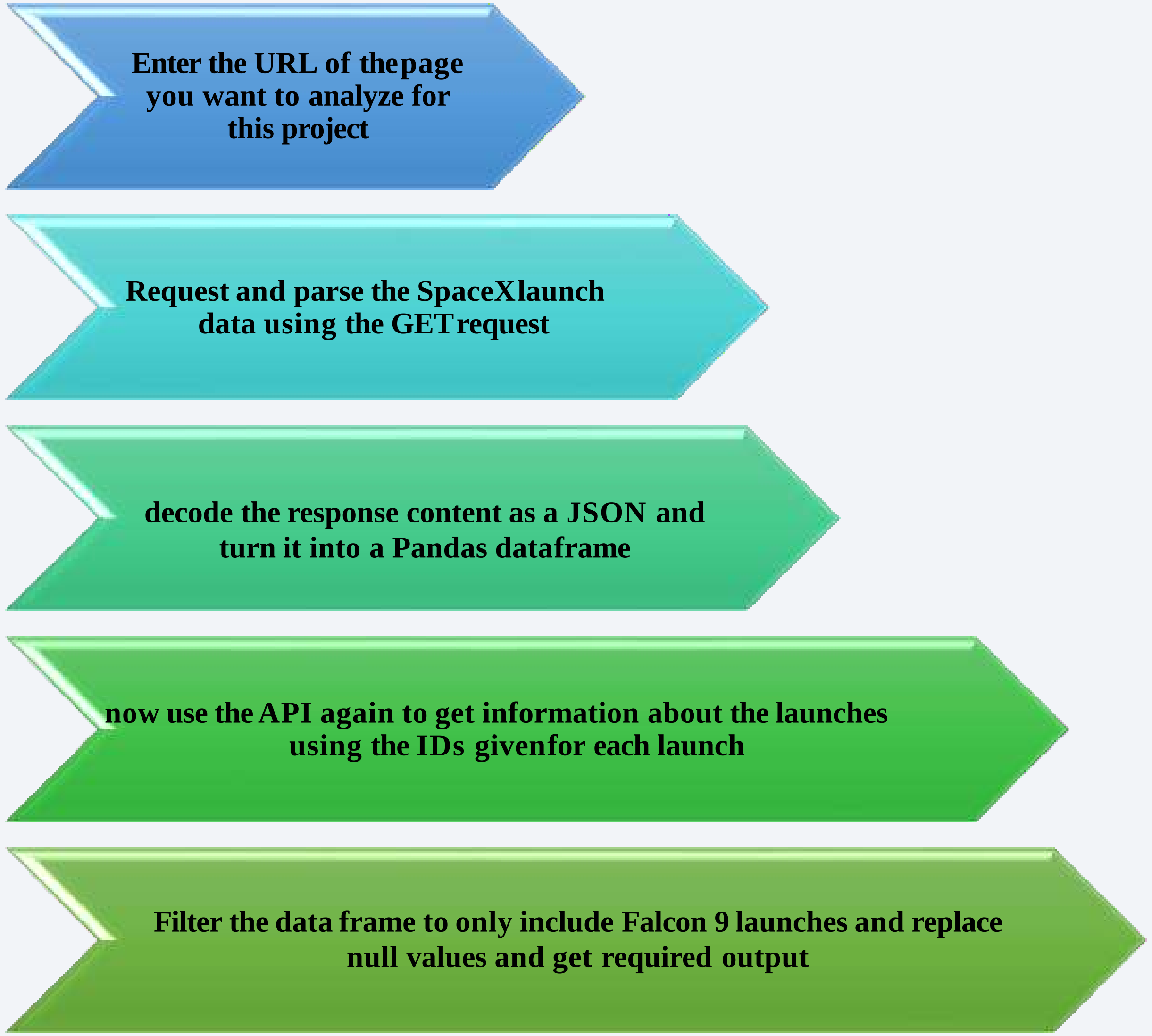
Executive Summary

- Data collection methodology:
  - Get requests to the Space X API and web scraping from wikipedia
- Perform data wrangling
  - Clean the data
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Creating best Machine Learning model.

Data Collection

The Data sets are collected by

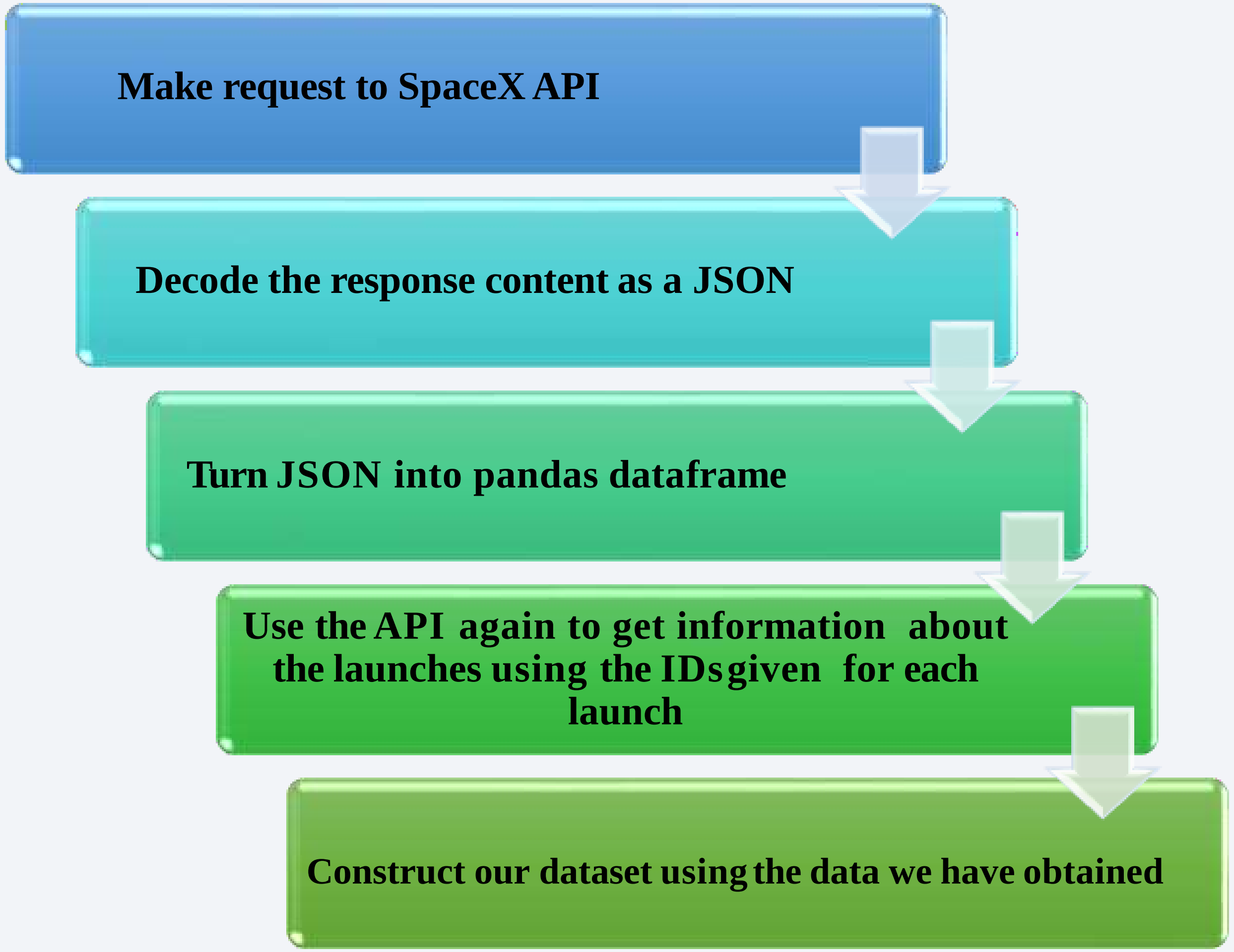
- SpaceX API request.
- Web Scraping



Data Collection – SpaceX API

- How Data Collection has done is given in a form of flow chart for an overview. For completed notebook link is given below

- GitHub : Link



Data Collection - Web Scraping

- Data Collection BY Web Scraping process is given in flow chart for an overview. For Completed Notebook link given below
- GitHub : [Link]

Falcon9 Launch Wikipage from its URL

↓

Create a Beautiful Soup object from the HTML

↓

Extract all column names from the HTML table

↓

Create an empty dictionary with keys from the column names

↓

Fill up the dictionary with launch records extracted from table rows.

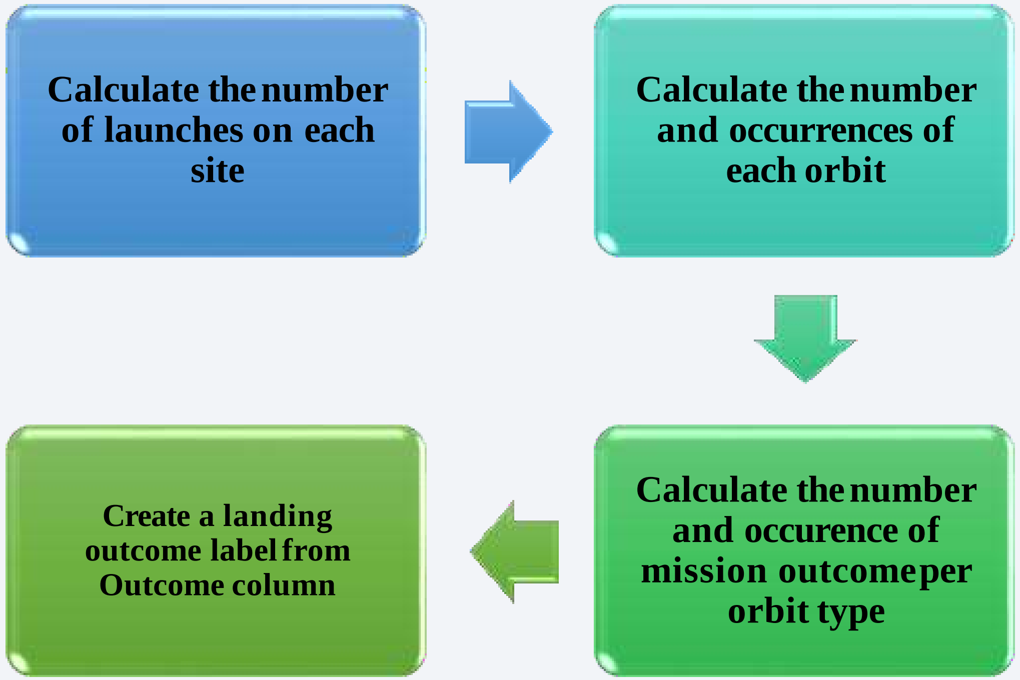
↓

Convert the dictionary into a CSV dataset

Data Wrangling

- Data Wrangling process is given in a flow chart for a over view. For completed notebook link given below

- Git Hub : Link



EDA with Data Visualization

Types of Charts Used :

- *Scatter plot* - Flight Number vs Payload Mass , Flight Number vs Launch Sites , Payload and Launch Sites , Flight Number and Orbit Type , Payload and Orbit Type
- *Bar chart* – Success rate of each orbit
- *Line plot* – success rate and Date

EDA with Data Visualization complete notebook link is given below

GitHub : Link

EDA with SQL

Summary of SQL queries that were used:

1. Display the names of the unique launch sites in the space mission
2. Display 5 records where launch sites begin with the string 'CCA
3. Display the total payload mass carried by boosters launched by NASA(CRS)
4. Display average payload mass carried by booster version F9 v1.1
5. List the date when the first successful landing outcome in ground pad was acheived
6. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
7. List the total number of successful and failure mission outcomes
8. List the names of the booster versions which have carried the maximum payload mass. Use a subquery
9. List the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
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

GitHub : Link

Build an Interactive Map with Folium

Folium Markers were used to show the Space X launch sites and their nearest important landmarks like railways, highways, cities and coastlines.Polylines were used to connect the launch sites to their nearest land marks.

Red represents launch failures

Green represents the successes.

Build a Dashboard with Plotly Dash

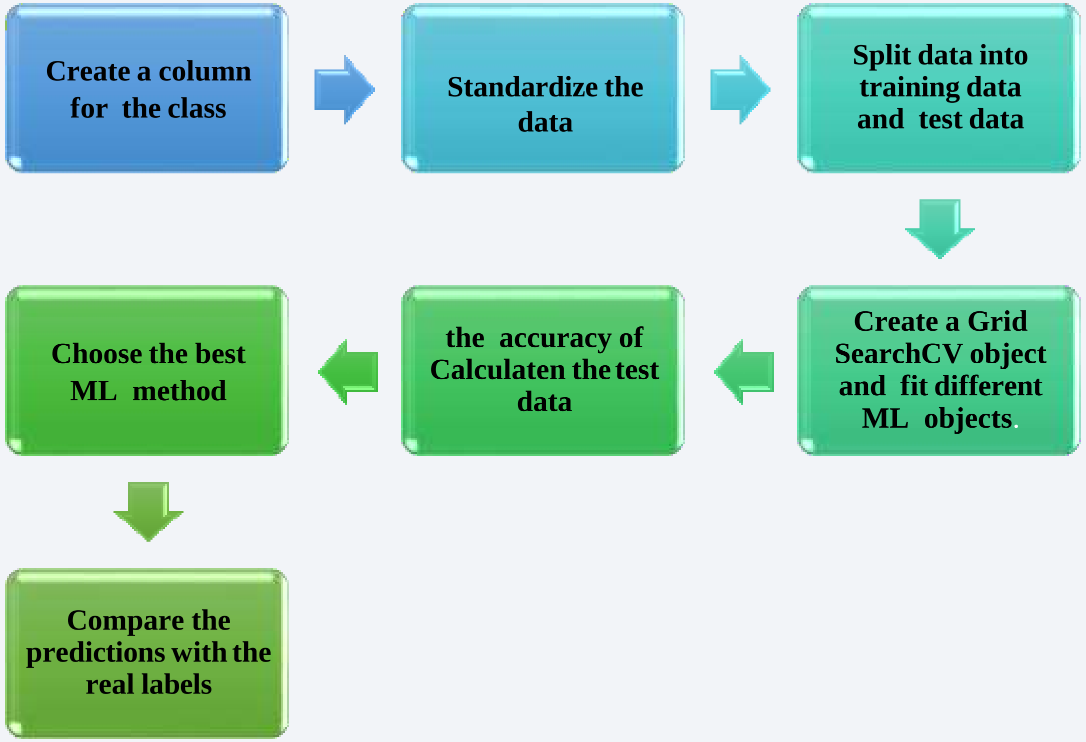
- Pie charts and scatter charts were used to visualize the launch records of Space X.
- These charts displayed the rocket launch success rate per launch site. We are were able to get an understanding of the factors that may have been influencing the success rate at each site. Such as the payload mass and booster versions.
- Successful launches were represented by 1 while failures were represented by 0.

Predictive Analysis (Classification)

Scikit-learn is Machine Learning library that was used for predictive analysis. The following took place:

- Created a machine learning pipeline to predict if the first stage will land given the data.

- GitHub: [Link]



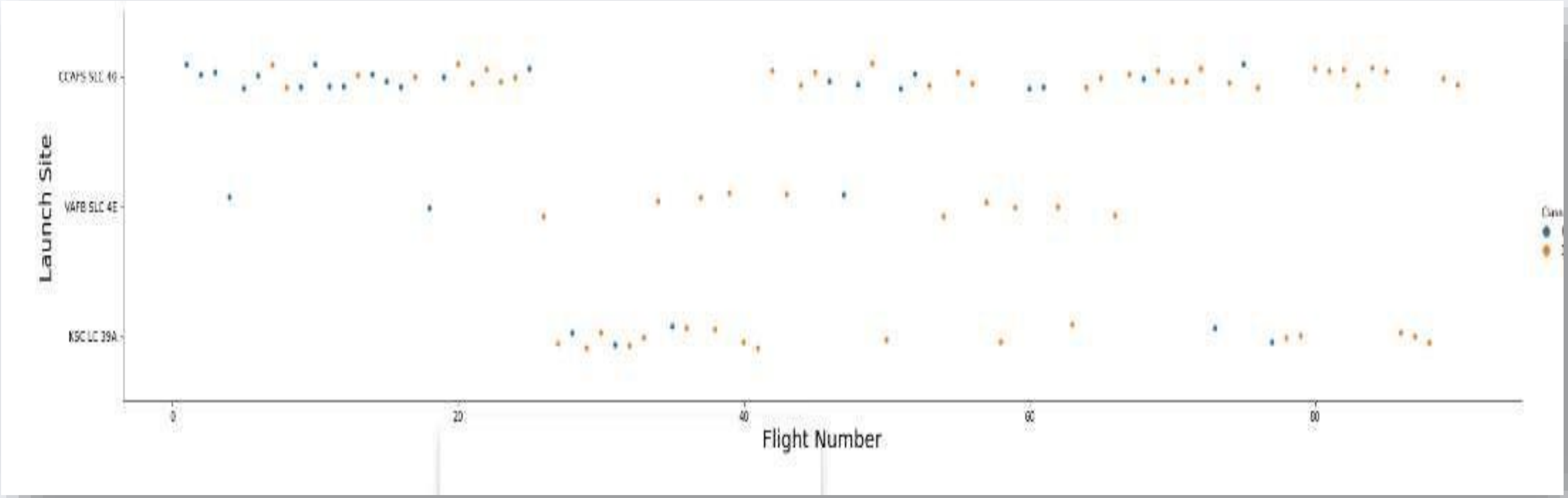
Results

- The exploratory data analysis has shown us that successful landing outcomes are somewhat correlated with flight number. It was also apparent that successful landing outcomes have had a significant increase since the year 2015.
- All launch sites are located near the coast line. Perhaps, this makes it easier to test rocket landings in the water.
- sites are also located near highways and railways. This may facilitate transportation of equipment and research material.
- The machine learning were able to predict the landing success of rockets with an accuracy score of 83.33%.

Section 2

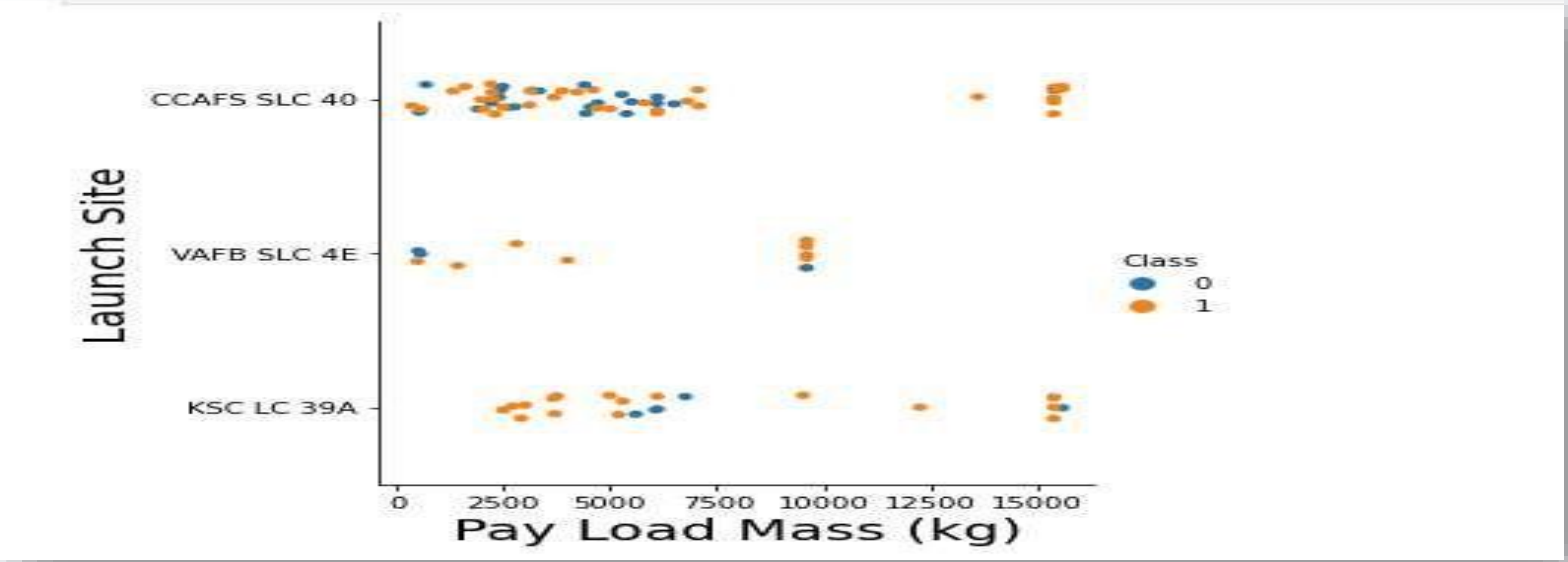
Insights drawn from EDA

Flight Number vs. Launch Site



- It appears that there were more successful landings as the flight numbers increased. launch site **CCAFS SLC 40** had the most number of landing.

Payload vs. Launch Site

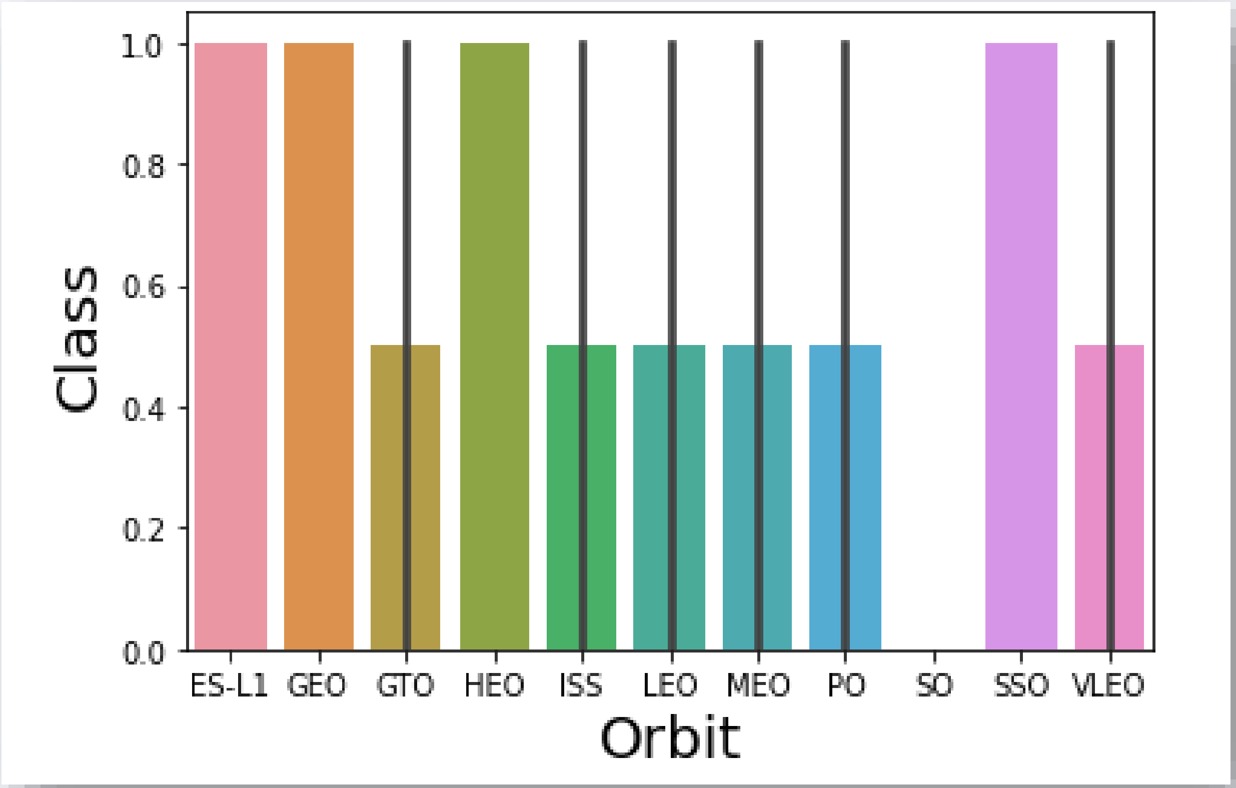


- Now if you observe the scatter point chart, you will find for the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).

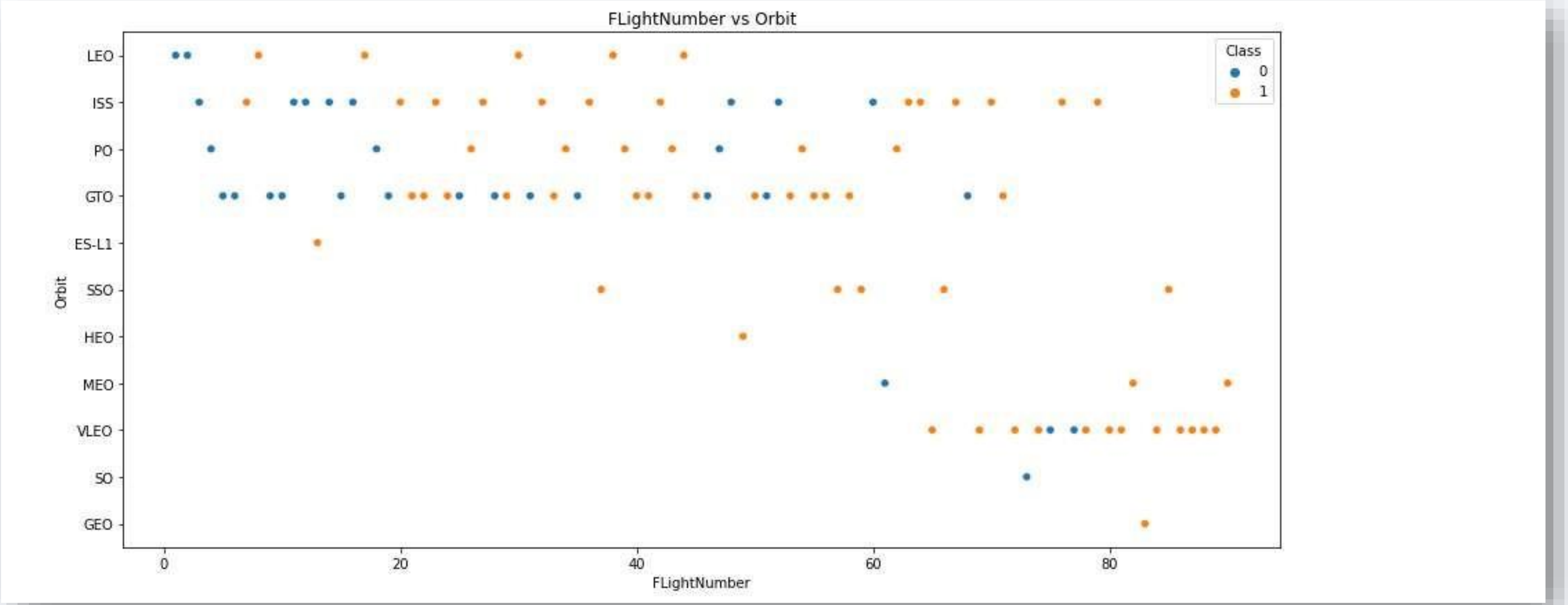
Success Rate vs. Orbit Type

The highest success rate ORBITS are

1. ***ES-L1***
2. ***GEO***
3. ***SSO***
4. ***HEO***

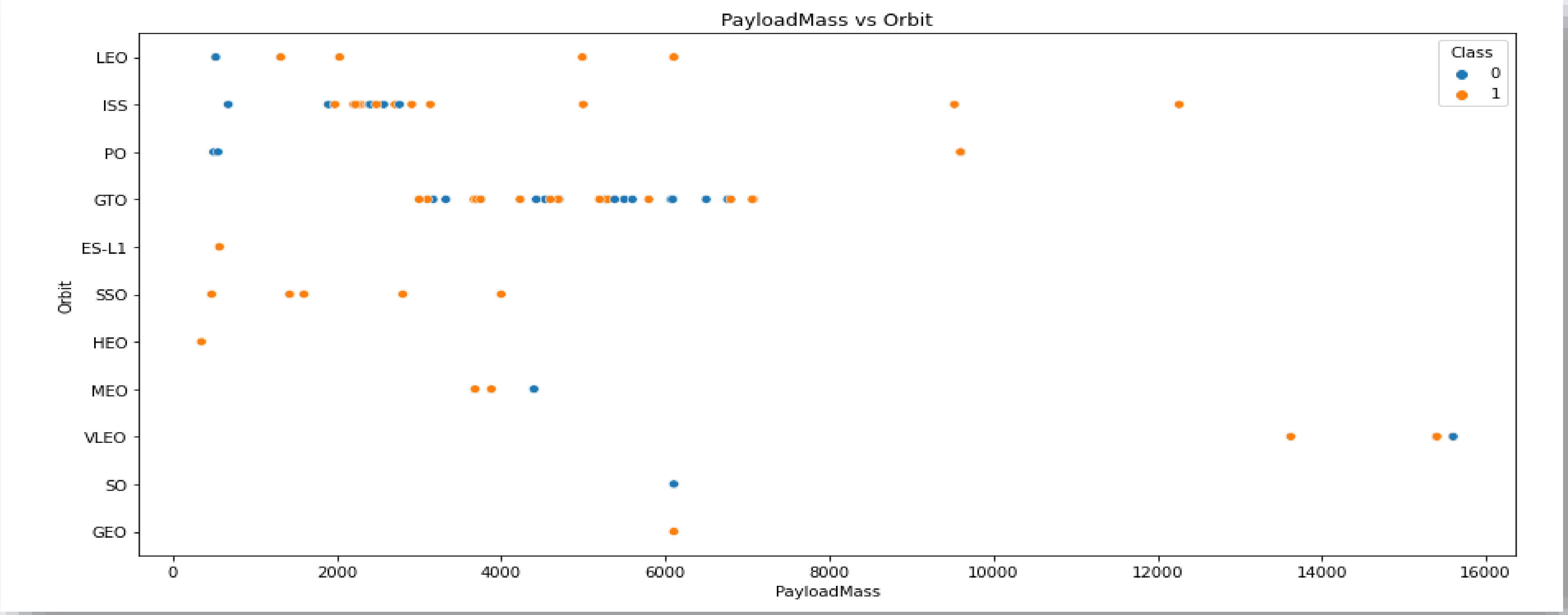


Flight Number vs. Orbit Type



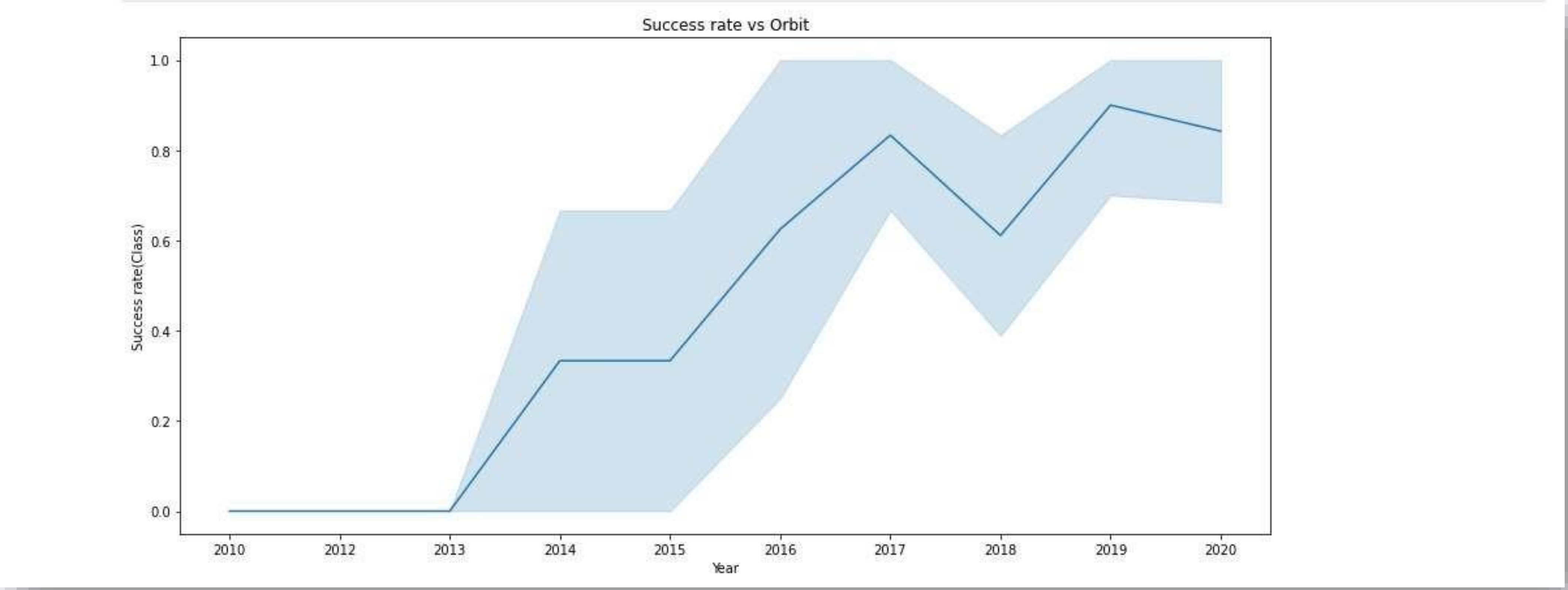
You can see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there.

Launch Success Yearly Trend



- It is apparent that the success rate has significantly increased from 2013 to 2020.

All Launch Site Names

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

```
[15]: %sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL
 * sqlite:///my_data1.db
Done.
```

[15]:

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

Given the data, these are the names of the launch sites where different rocket landings where attempted:

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

Launch Site Names Beginning with 'CCA'

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

```
[16]: %sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5

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

[16]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS_KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- These are 5 records where launch sites begin with the letters 'CCA'. As we can see, there are other organizations besides Space X that were testing their rockets.

Total Payload Mass

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

```
[17]: %sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD, CUSTOMER FROM SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)'
```

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

[17]:

TOTAL_PAYLOAD	Customer
45596	NASA (CRS)

- The information in the picture displays the total payload mass carried by boosters launched by NASA

Average Payload Mass by F9 v1.1

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

```
[41]: %sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD, BOOSTER_VERSION FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1'

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

[41]:

AVG_PAYLOAD	Booster_Version
2928.4	F9 v1.1

- The average payload mass carried by **F9 v1.1** was **2928.4 kg**.

First Successful Ground Landing Date

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

```
[24]: %sql SELECT Max(Date) as Date, "Landing _Outcome" FROM SPACEXTBL WHERE "Landing _Outcome" = "Success (ground pad)"

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

[24]:

Date	Landing _Outcome
22-12-2015	Success (ground pad)

- From the picture given above you can see that the first successful ground pad was in **22 December 2015**.

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

```
[61]: %sql SELECT Booster_Version,"Landing _Outcome", FROM SPACEXTBL WHERE ("Landing _Outcome" = "Success (drone ship)") AND (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000)

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

[61]:

Booster_Version	Landing _Outcome
F9 FT B1022	Success (drone ship)
F9 FT B1026	Success (drone ship)
F9 FT B1021.2	Success (drone ship)
F9 FT B1031.2	Success (drone ship)

It appears that there only 4 Boosters with a payload mass between 4000 and 6000 they are

- **F9 FT B1022**
- **F9 FT B1026**
- **F9 FT B1021.2**
- **F9 FT B1031.2**

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
[81]: %sql SELECT mission_Outcome,COUNT(mission_Outcome) AS TOTAL FROM SPACEXTBL GROUP BY mission_Outcome

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

[81]:

Mission_Outcome	TOTAL
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- The Above picture show the total number of successful and failure mission outcomes

Boosters That Carried Maximum Payload Mass

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

```
[102]: %sql SELECT "Booster_Version",Payload, "PAYLOAD_MASS__KG_" FROM SPACEXTBL WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTBL);

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

[102]:

Booster_Version	Payload	PAYLOAD_MASS__KG_
F9 B5 B1048.4	Starlink 1 v1.0, SpaceX CRS-19	15600
F9 B5 B1049.4	Starlink 2 v1.0, Crew Dragon in-flight abort test	15600
F9 B5 B1051.3	Starlink 3 v1.0, Starlink 4 v1.0	15600
F9 B5 B1056.4	Starlink 4 v1.0, SpaceX CRS-20	15600
F9 B5 B1048.5	Starlink 5 v1.0, Starlink 6 v1.0	15600
F9 B5 B1051.4	Starlink 6 v1.0, Crew Dragon Demo-2	15600
F9 B5 B1049.5	Starlink 7 v1.0, Starlink 8 v1.0	15600
F9 B5 B1060.2	Starlink 11 v1.0, Starlink 12 v1.0	15600

- From the above picture it shows that 12 boosters have carried the maximum payload mass of 15600 kg.

2015 Launch Records - Failed Landing Outcomes

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.

```
[83]: %sql SELECT substr(Date,7,4), substr(Date, 4, 2),"Booster_Version", "Launch_Site", Payload, "PAYLOAD_MASS__KG_", "Mission_Outcome", "Landing _Outcome" FROM SPACEXTBL WHERE substr(Date,7,4)='2015' AND "Landing _Outcome" = 'Failure (drone ship)'
```

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

[83]:

substr(Date,7,4)	substr(Date, 4, 2)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Mission_Outcome	Landing _Outcome
2015	01	F9 v1.1 B1012	CCAFS LC-40	SpaceX CRS-5	2395	Success	Failure (drone ship)
2015	04	F9 v1.1 B1015	CCAFS LC-40	SpaceX CRS-6	1898	Success	Failure (drone ship)

- Two boosters **F9 v1.1 B1012** and **F9 v1.1 B1015** failed to land in 2015.

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

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

```
[26]: %sql SELECT * FROM SPACEXTBL WHERE "Landing _Outcome" LIKE 'Success%' AND (Date BETWEEN '04-06-2010' AND '20-03-2017') ORDER BY Date DESC LIMIT 5
```

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

[26]:

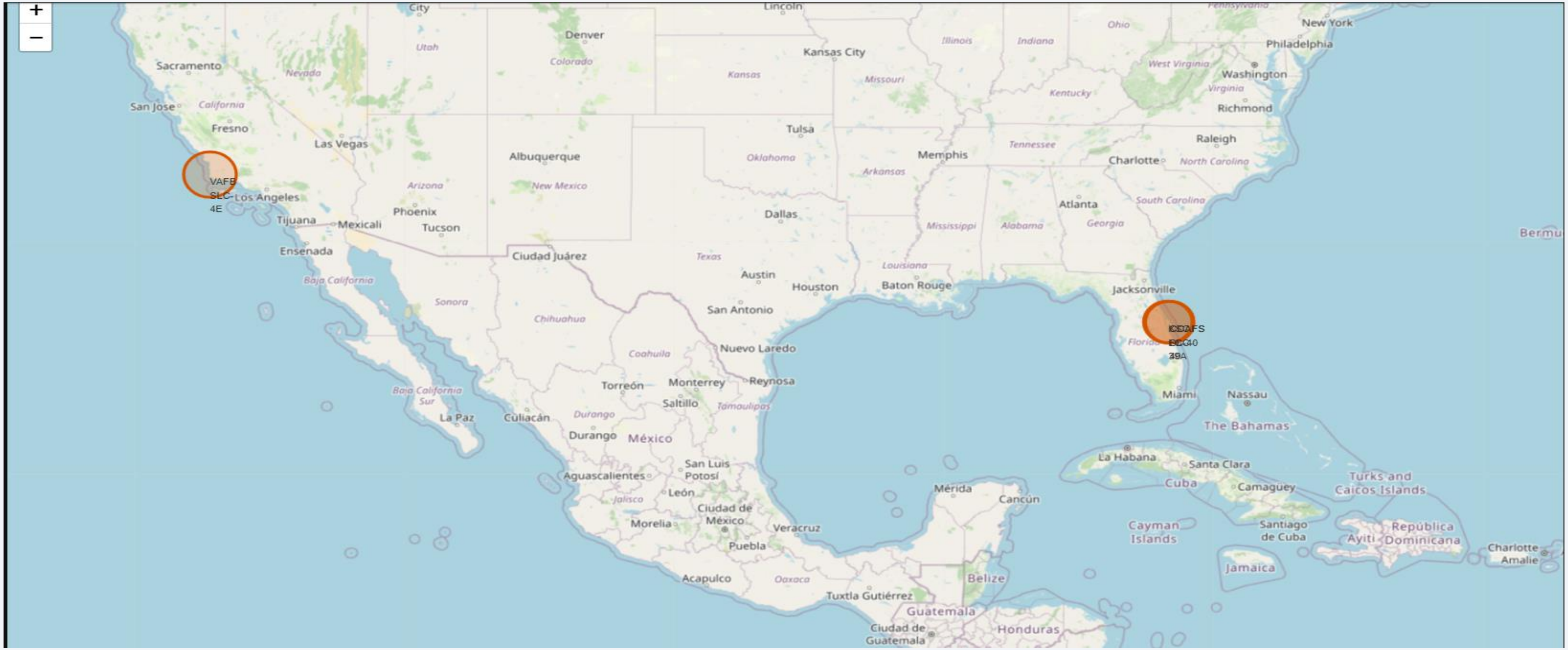
Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
19-02-2017	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
18-10-2020	12:25:57	F9 B5 B1051.6	KSC LC-39A	Starlink 13 v1.0, Starlink 14 v1.0	15600	LEO	SpaceX	Success	Success
18-08-2020	14:31:00	F9 B5 B1049.6	CCAFS SLC-40	Starlink 10 v1.0, SkySat-19, -20, -21, SAOCOM 1B	15440	LEO	SpaceX, Planet Labs, PlanetIQ	Success	Success
18-07-2016	04:45:00	F9 FT B1025.1	CCAFS LC-40	SpaceX CRS-9	2257	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
18-04-2018	22:51:00	F9 B4 B1045.1	CCAFS SLC-40	Transiting Exoplanet Survey Satellite (TESS)	362	HEO	NASA (LSP)	Success	Success (drone ship)

- The number of successful landings have increased since 2015.

Section 3

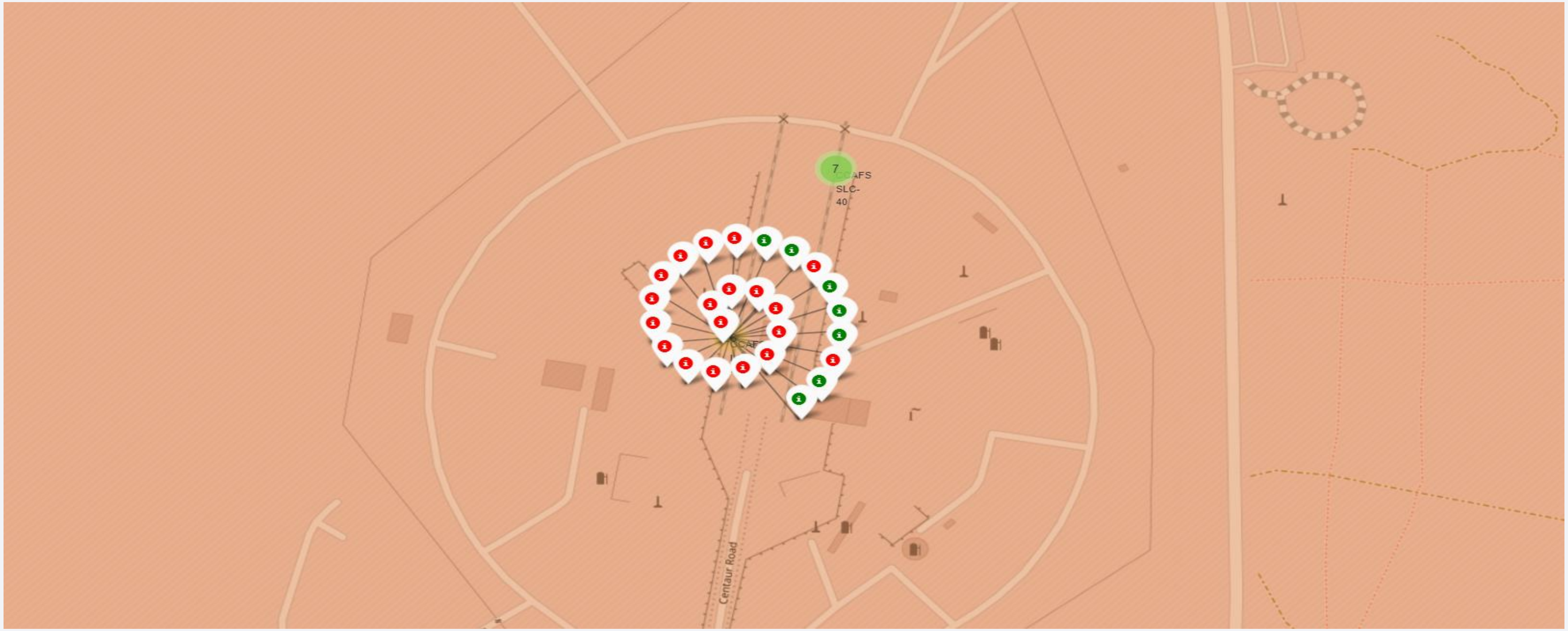
Launch Sites Proximities Analysis

Launch Site Locations



- all launch sites are in very close proximity to the coast and they are also a couple thousand kilometers away from the equator line.

Success Rate of Rocket Launches



- The successful launches are represented by a **green** marker while the **red** marker represents failed rocket launches.

Surrounding Landmarks

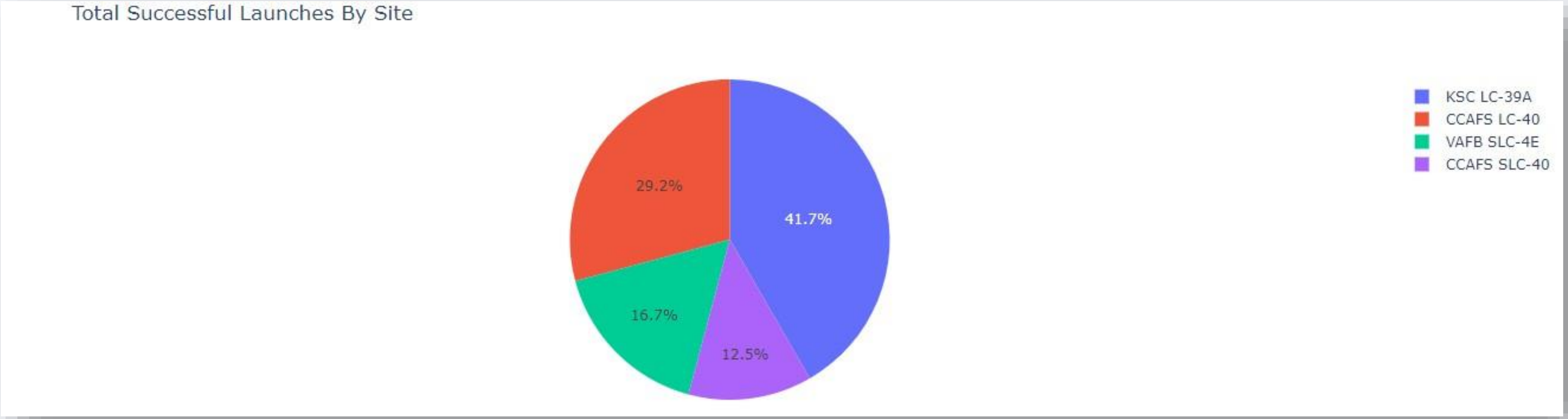
- It appears that launch sites are usually set up at least 18 km away from cities. This may be because of the desire to prevent any crashes near populated areas.
- It is also apparent that launch sites are in very close proximity to railways and highways. Perhaps, due to the necessary transportation requirements for rocket parts.
- The sites are close the coast line. This is evident with the many rocket landing tests on water bodies like the ocean.

Map Object	Colour
Nearest Highway	Green
Nearest Railway	Purple
Nearest City	Crimson
Nearest Coastline	Dark Blue

Section 4

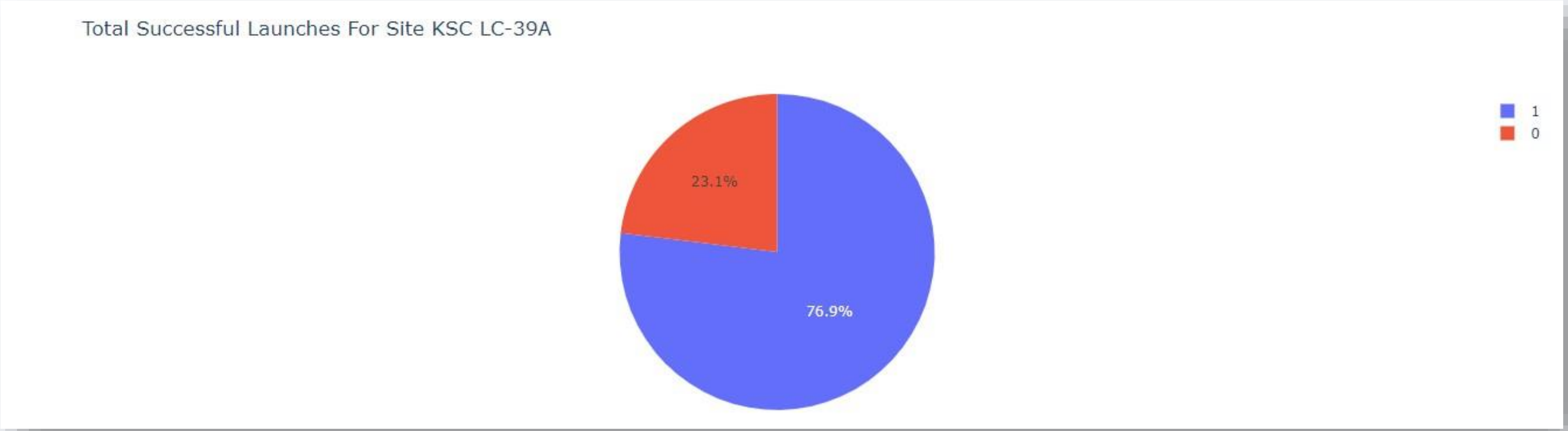
Build a Dashboard with Plotly Dash

Successful Launches by Site



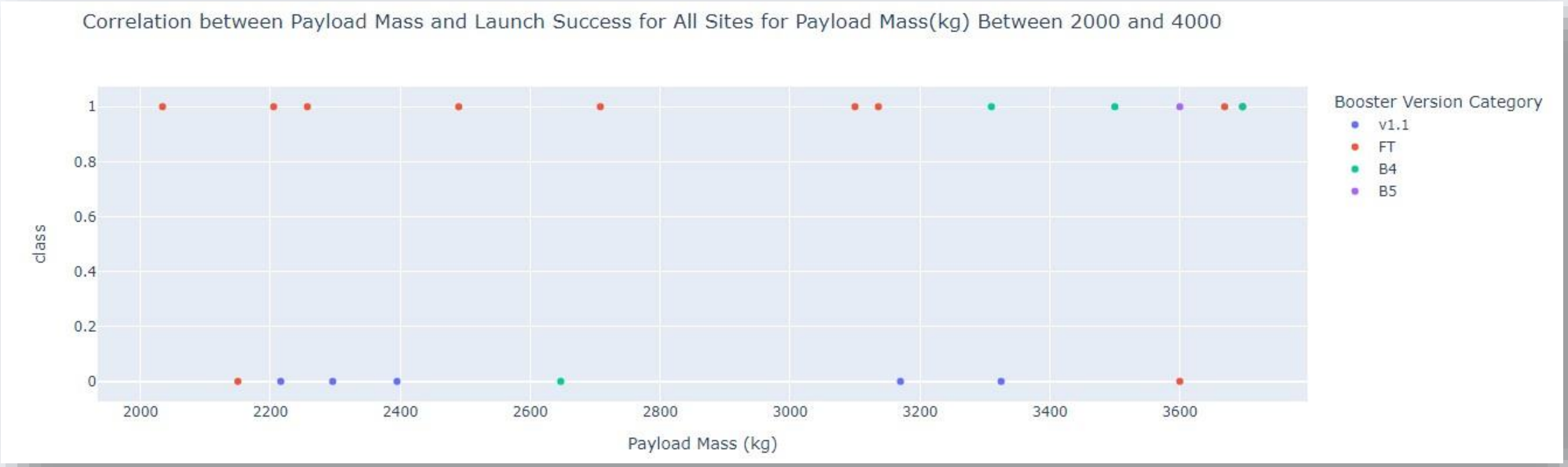
- You can see from the plot that Site KSC LC-39A has the largest successful launches as well the highest launch success rate.

Total Successful Launches for Site KSC LC-39A



- You can see that 76.9% of the total launches at site KSC LC-39A were successful. This is a the highest success rate of all the different launch sites.

Payload Mass vs. Launch Success for All Sites



- It appears that the payload range between 2000 kg and 4000 kg has the highest success rate.

Section 5

Predictive Analysis (Classification)

Classification Accuracy


```
[30]: Report = pd.DataFrame({'Method':['knn_accuracy','Decision_tree_accuracy','SVM_accuracy','Logistic_Regression'],'Test Data Accuracy':[knn,DCT,svm,score]})
Report
```

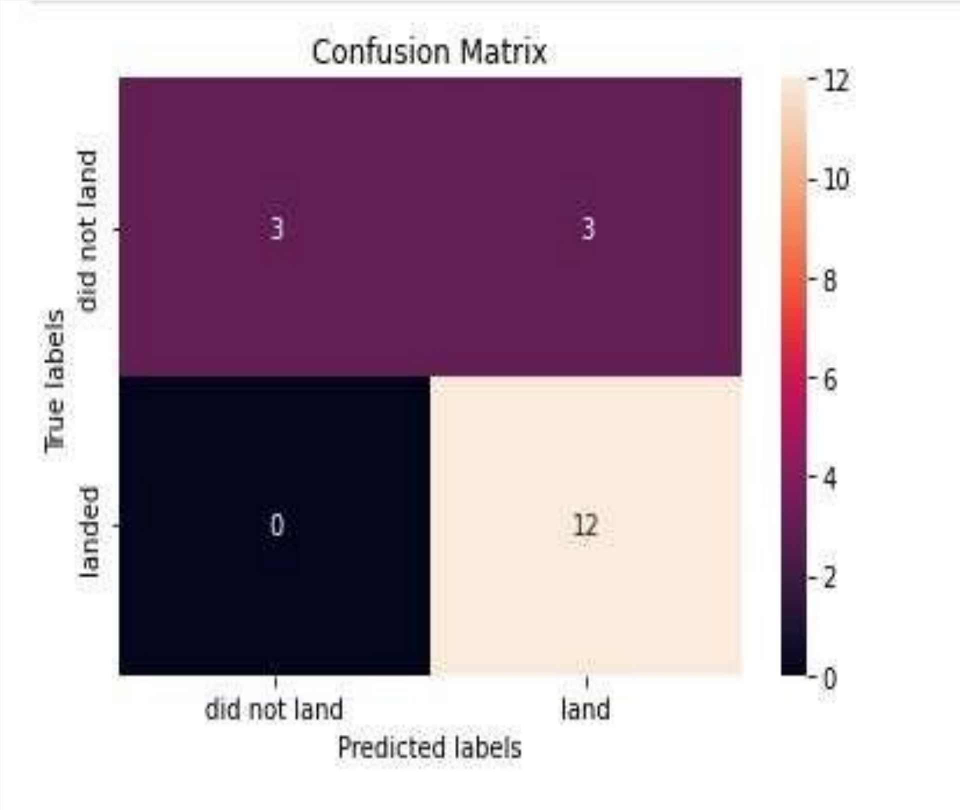
[30]:

	Method	Test Data Accuracy
0	knn_accuracy	0.833333
1	Decision_tree_accuracy	0.833333
2	SVM_accuracy	0.833333
3	Logistic_Regression	0.833333

- You can see that All the methods have an identical accuracy score of 83.33%, so we decided to use Logistic Regression for the classification

Confusion Matrix

- The chart shows the confusion matrix of the Logistic Regression model that was chosen.
- The model only failed to accurately predict 3 labels.



Conclusions

In order to compete with Space X Through this process, a general picture of their success methods are

- All their launch sites are located near the coast, away from nearby cities. This enabled to them to test their rocket landings without much interference.
- Site **KSC LC-39A** had the highest launch success rate out of all the launch sites.
- From 2015 onwards, the success rate of rocket landings significantly increased. It was also apparent that landing success increased with flight number

All this data was used to train a machine learning model that is able to predict the landing outcome of rocket launches with 83.33% accuracy.

Thank you!