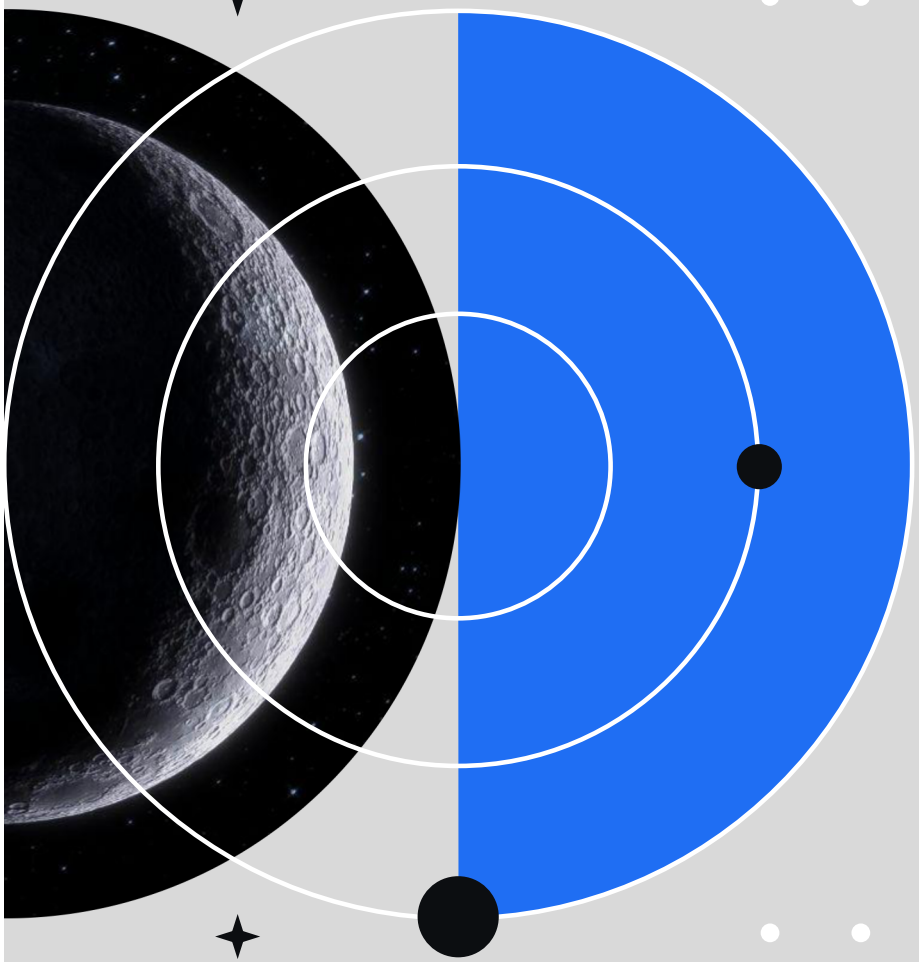
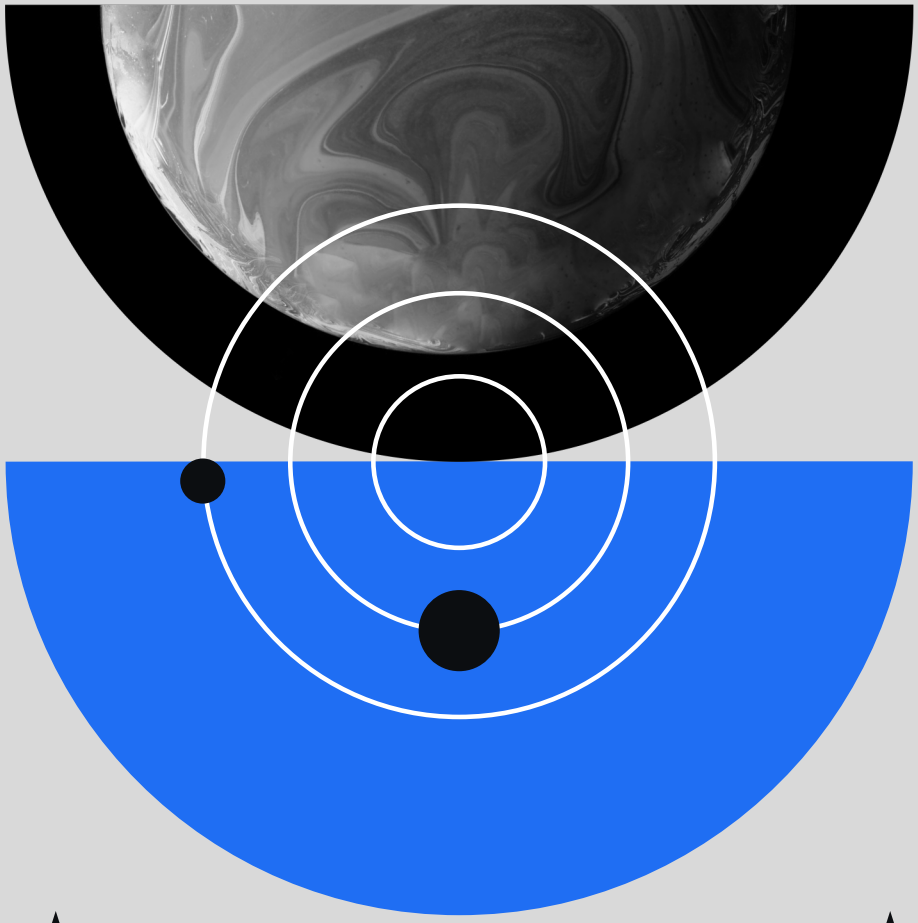


Winning Space Race with Data Science

Teruyoshi Hasegawa
IBM Data Science Capstone Project

Outline

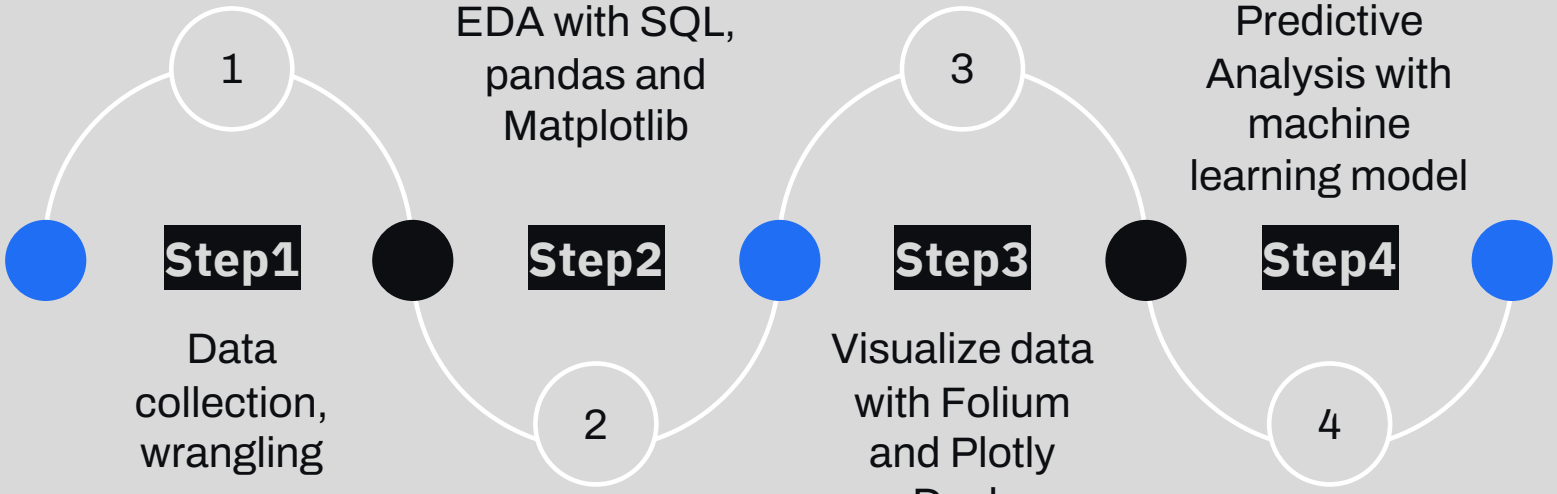


01

Executive Summary

#1 Method

*EDA : Exploratory data analysis



#2 Results

Successfully built a machine learning model with **83.3%** of accuracy to predict if SpaceX will reuse the first stage

02 Introduction

Introduction

Background & Context

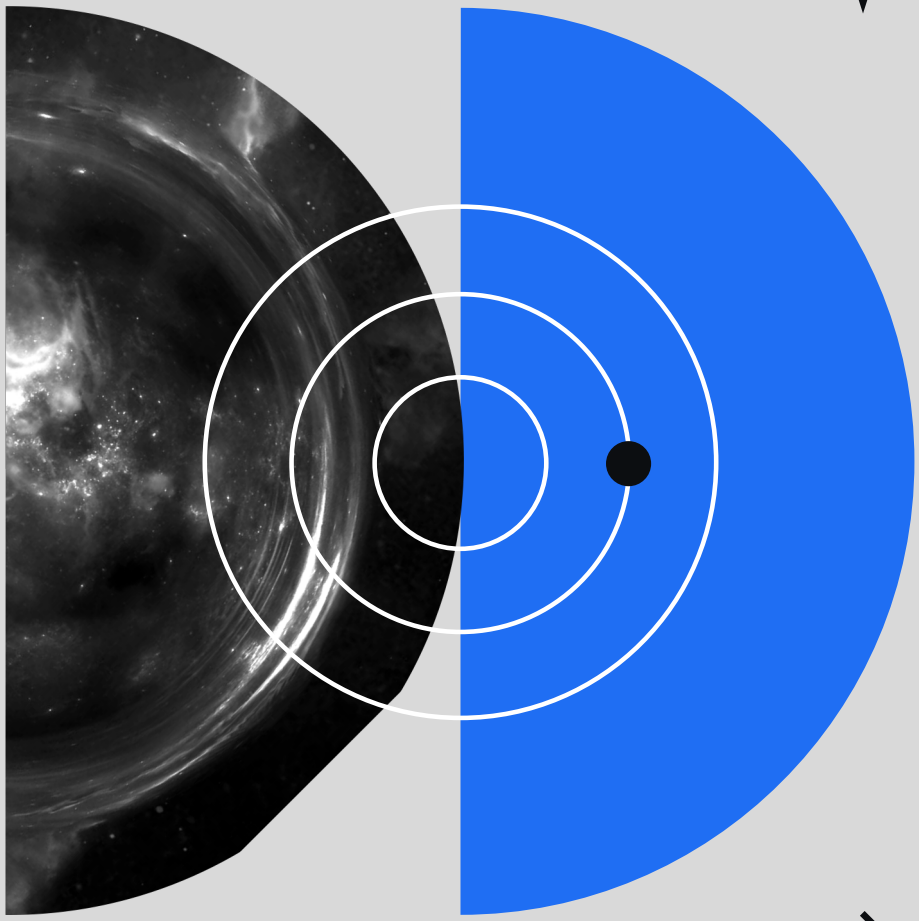
The commercial space age is here. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars, much of the savings is because SpaceX can reuse the first stage.
Our company, SpaceY, would like to compete with SpaceX in order to win the space race.

Problems to solve

In this project, our objective is to build a model in order to determine if SpaceX will reuse the first stage using a machine learning model and public information instead of using rocket science.

03

Methodology



Data Collection

Source#1 : SpaceX open data

- ✓ Request to the SpaceX API
- ✓ Convert JSON format into a Pandas data frame
- ✓ Clean the requested data

Source#2 : Wikipedia data table

- ✓ Extract a Falcon 9 launch records HTML table from Wikipedia
- ✓ Parse the table and convert it into a Pandas data frame

Data Collection Flow – SpaceX API

GitHub URL is here

#1 : Request SpaceX API

#2 : JSON file will be responded

#3 : Convert JSON data into dataframe using “json_normalize()”

#4 : Filter the dataframe to only include Falcon 9 launches

#5 : Dealing with Missing Values (replace np.nan values in the data with the mean)

Data Collection Flow – Web Scraping

GitHub URL is here

#1 : Request the Falcon9 Launch Wiki page from its URL

#2 : Use BeautifulSoup() to create an object from a response text content

#3 : Extract all column/variable names from the HTML table header

#4 : Create an empty dictionary with keys from the extracted column names

#5 : Iterate through table cells to extract data and convert dictionary into dataframe

Data Wrangling

Objective #1 : Exploratory Data Analysis

- ✓ Import data as Pandas dataframe using “.read_csv” method
- ✓ Explore data using pandas module

Objective #2 : Determine Training Labels

- ✓ Use the method “.value_counts()” to confirm contains on each column
- ✓ Create training label with landing outcomes where successful = 1 & failure = 0

Data Wrangling Flow

GitHub URL is here

#1 : Import libraries such as pandas and numpy then load dataset as dataframe

#2 : Explore data using “.isnull()” and ”.dtypes” to identify missing values and data type

#3 : Calculate the number of launches on each site using “value_counts()”

#4 : Calculate the number and occurrence of each orbit

#5 : Create training label with landing outcomes where successful = 1 & failure = 0

EDA with Data Visualization

Objective #1 : Exploratory Data Analysis

- ✓ Import data as Pandas dataframe using ".read_csv" method
- ✓ Explore data using pandas module

Objective #2 : Determine Training Labels

- ✓ Use the method ".value_counts()" to confirm contains on each column
- ✓ Create training label with landing outcomes where successful = 1 & failure = 0

Data Visualization Flow

GitHub URL is here

Plots Used	Objective
Scatter point chart	✓ Relation between Flight Number and Launch Site ✓ Relation between Payload and Launch Site ✓ Relation between Flight Number and Orbit type ✓ Relation between Payload and Orbit type
Bar Graph	✓ Relation between Success rate of each orbit type
Line Graph	✓ Yearly trend of the launch success

EDA with SQL #1

- performed SQL queries to gather dataset

GitHub URL is here

#01 : Display the names of the unique launch sites in the space mission

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

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

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

#05 : List the date when the first successful landing outcome in ground pad was achieved

EDA with SQL #2

- performed SQL queries to gather dataset

GitHub URL is here

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

#07 : List the total number of successful and failure mission outcomes

#08 : List the names of the booster_versions which have carried the maximum payload

#09 : 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 between the date 2010-06-04 and 2017-03-20, in descending order

Map with Folium

- to visualize the launch data into an interactive map

Marker, Circle object

Highlight launch sites with marker and circle

Marker Cluster object

Mark the success/failed launches for each site

PolyLine object

Calculate the distances between a launch site to its proximities and connect with line

GitHub URL is here

Build a Dashboard
- with Plotly Dash flow

GitHub URL is here

#01 : Display the names of the unique launch sites in the space mission

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

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

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

#05 : List the date when the first successful landing outcome in ground pad was achieved

Predictive Analysis
- Classfication flow

GitHub URL is here

#01 : Load datasets then fit and transform features using standard scaler

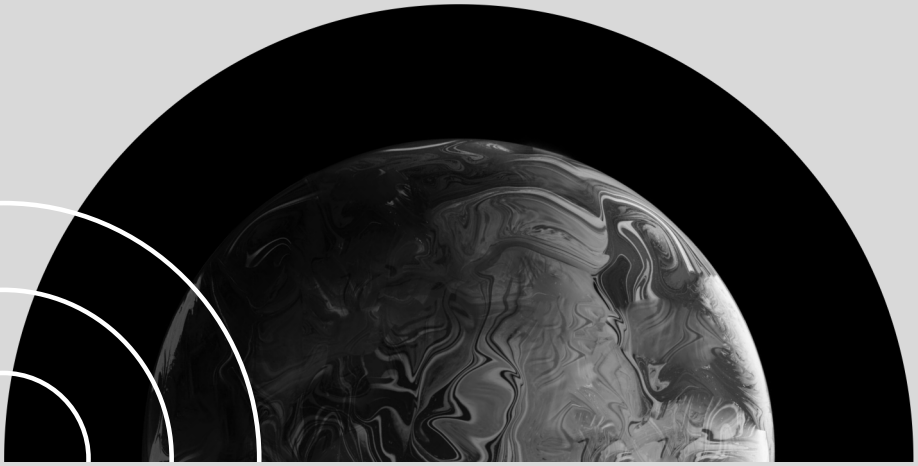
#02 : Split the data X and Y into training and test data using train_test_split function

#03 : Build 4 models which are “Logistic Regression”,“SVM”,”Decision Tree”,”KNN"

#04 : Plot the confusion matrix for all models

#05 : Find the method performs best

Results 04



Data Wrangling

61%

61% of launches were fired off from “CCAFS SLC 40” followed by “KSC LC 39 A” and “VAFB SLC 4E” at 24% and 15%, respectively.

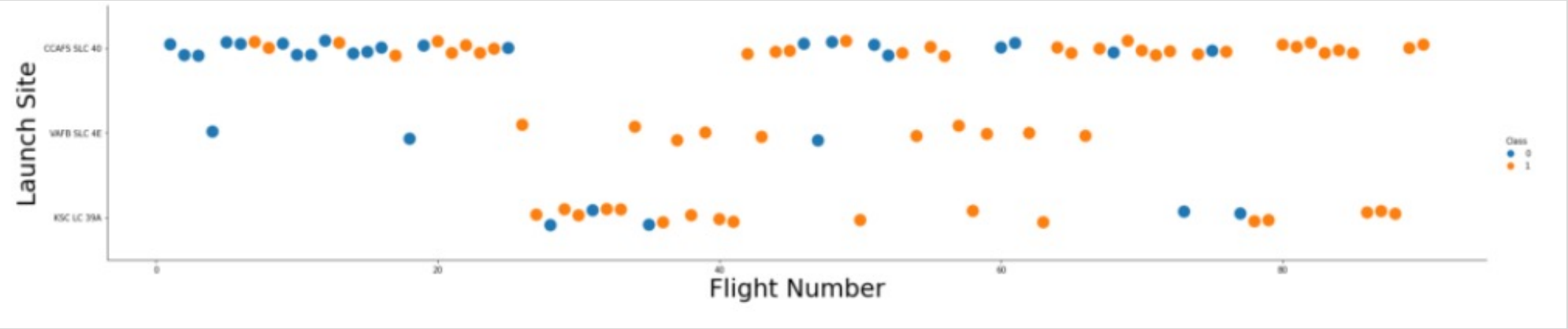
69%

69% of launches towards 3 orbits, “GTO/ISS/VLEO”, out of total 11 orbits.

67%

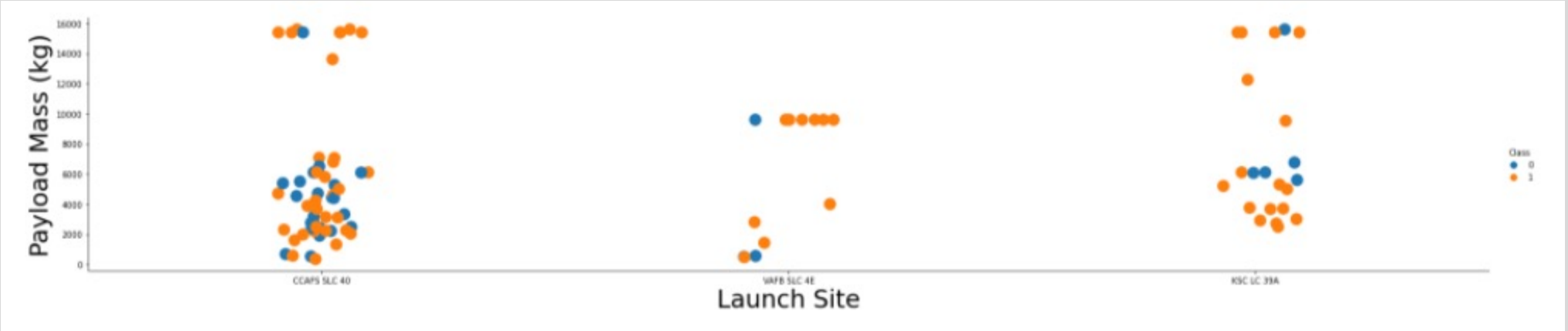
Launch success rate is 67%. In order to calculate a success rate, training label was created with landing outcomes where successful = 1 & failure = 0

Flight Number vs Launch Site



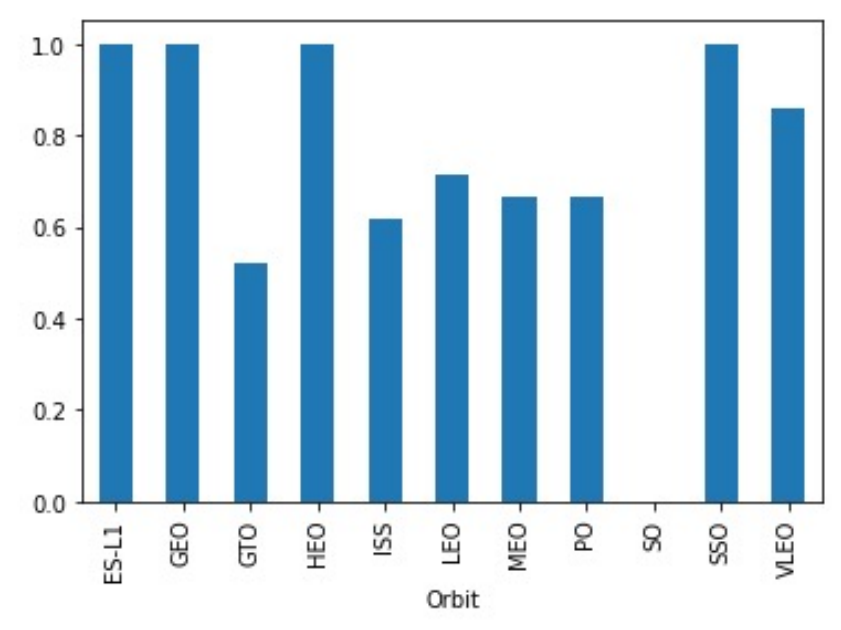
- ✓ “CCAFS SLC 40” is the main launch site and higher success rate can be found in higher flight number
- ✓ Lower success rate was found in “CCAFS SLC 40” especially with flight number is lower than 40

Payload vs Launch Site



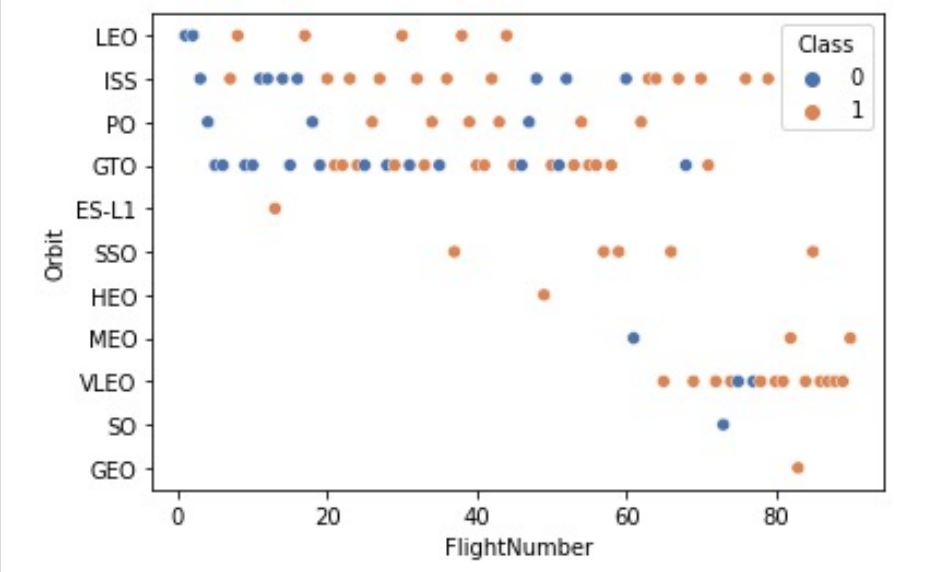
- ✓ Launches with over 10,000 kg payload show higher success rate in all launch site
- ✓ Launches with lower 8,000 kg payload from “CCAFS SLC 40” show lower success rate

Success Rate vs Orbit Type



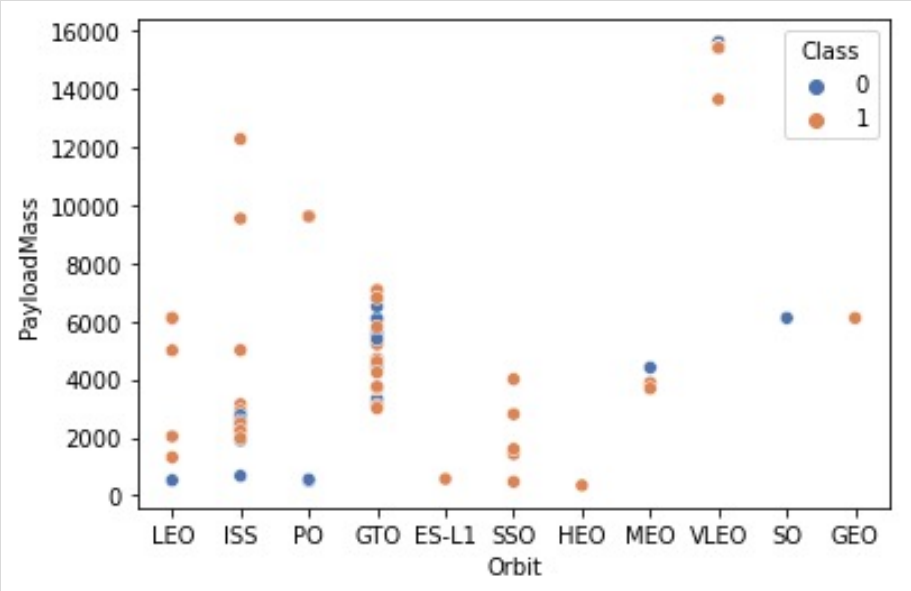
✓ **Launches towards ”ES-L1”, “GEO”, “HEO” and “SSO show 100% success rate**

Flight Number vs Orbit Type



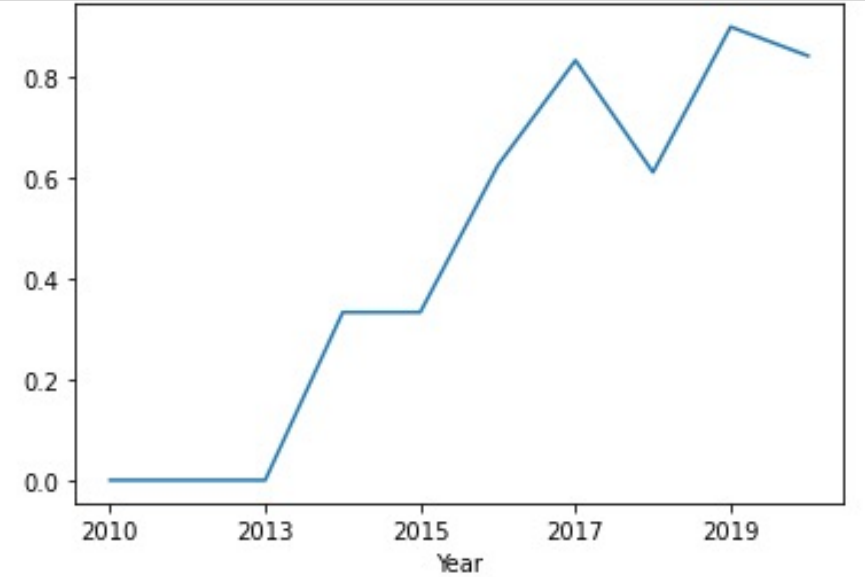
- ✓ LEO orbit the success appears related to the number of flight
- ✓ No relationship between flight number in GTO

Payload vs Orbit Type



- ✓ With heavy payloads the successful landing or positive landing rate are more for LEO and ISS
- ✓ For GTO we cannot distinguish this well as both positive landing rate and negative lading are both there here

Launch Success Yearly Trend



✓ The success rate since 2013 kept increasing till 2020

All Launch Site Name

```
In [12]: %sql SELECT Distinct LAUNCH_SITE FROM SpaceX
```

```
 * ibm_db_sa://ywv99023:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:31249/bludb
Done.
```

Out[12]:

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

Launch Site Name Begin with “CCA”

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

In [13]:
```
%sql SELECT * FROM SpaceX WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5
```

```
 * ibm_db_sa://ywv99023:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:3124
9/bludb
Done.
```

Out[13]:

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

Total Payload Mass

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

In [15]:
```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SpaceX WHERE CUSTOMER='NASA (CRS)'
```

```
 * ibm_db_sa://ywv99023:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:31249/bludb
Done.
```

Out[15]:

1
45596

Average Payload Mass by F9 v1.1

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

In [16]:
```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SpaceX WHERE BOOSTER_VERSION='F9 v1.1'
```

```
 * ibm_db_sa://ywv99023:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:3124
9/bludb
Done.
```

Out[16]:

1
2928

First Successful Ground Landing date

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

Hint:Use min function

In [17]:
```
%sql SELECT min(DATE) FROM SpaceX WHERE LANDING__OUTCOME='Success (ground pad)'
```

```
 * ibm_db_sa://ywv99023:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:3124
9/bludb
Done.
```

Out[17]:

1
2015-12-22

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 [19]:
```
%sql SELECT BOOSTER_VERSION FROM SpaceX WHERE PAYLOAD_MASS__KG_ between 4000 and 6000 AND LANDING__OUTCOME='Success (drone ship)'
```

```
 * ibm_db_sa://ywv99023:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:31249/bludb
Done.
```

Out[19]:

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

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

In [20]:
```
%sql SELECT COUNT(*) FROM SpaceX WHERE MISSION_OUTCOME LIKE '%Success%' OR MISSION_OUTCOME LIKE '%Failure%'
```

```
 * ibm_db_sa://ywv99023:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:3124
9/bludb
Done.
```

Out[20]:

1
101

Booster Carried Maximum Payload

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

In [21]:
```
%sql SELECT BOOSTER_VERSION FROM SpaceX WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SpaceX)
```

```
 * ibm_db_sa://ywv99023:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:31249/bludb
Done.
```

Out[21]:

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

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

In [22]:
```
%sql SELECT TO_CHAR(TO_DATE(MONTH("DATE"), 'MM'), 'MONTH') AS MONTH_NAME, \
    LANDING__OUTCOME AS LANDING__OUTCOME, \
    BOOSTER_VERSION AS BOOSTER_VERSION, \
    LAUNCH_SITE AS LAUNCH_SITE \
    FROM SpaceX WHERE LANDING__OUTCOME = 'Failure (drone ship)' AND "DATE" LIKE '%2015%'
```

```
 * ibm_db_sa://ywv99023:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:3124
9/bludb
Done.
```

Out[22]:

month_name	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

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

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 [23]:
```
%sql SELECT "DATE", COUNT(LANDING__OUTCOME) as COUNT FROM SpaceX \
    WHERE "DATE" BETWEEN '2010-06-04' and '2017-03-20' AND LANDING__OUTCOME LIKE '%Success%' \
    GROUP BY "DATE" \
    ORDER BY COUNT(LANDING__OUTCOME) DESC
```

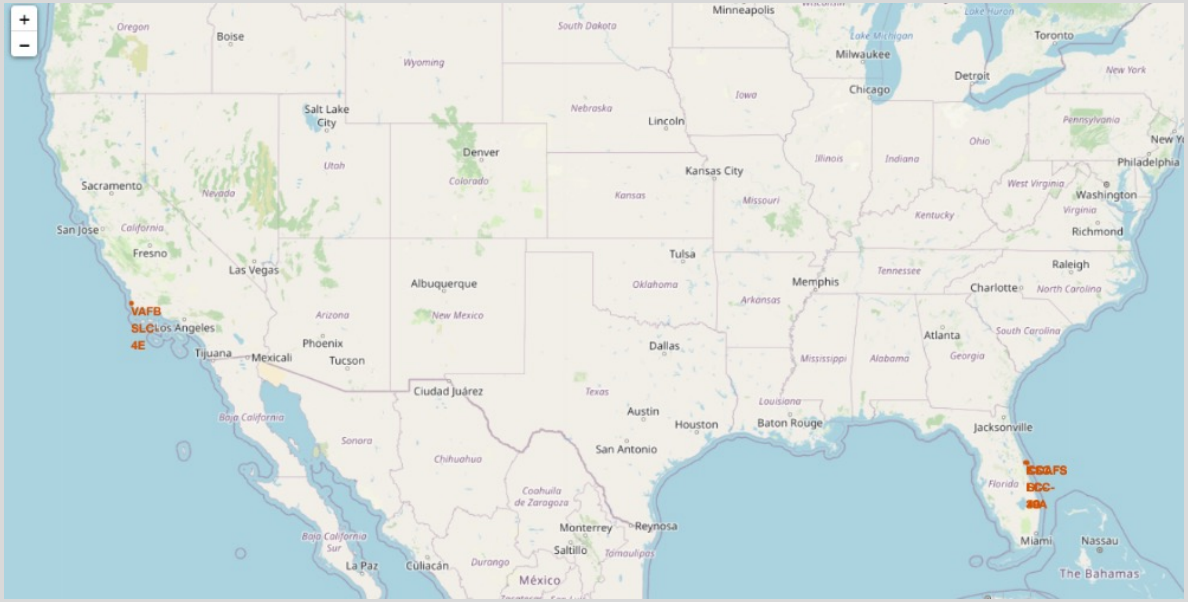
```
 * ibm_db_sa://ywv99023:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:3124
9/bludb
Done.
```

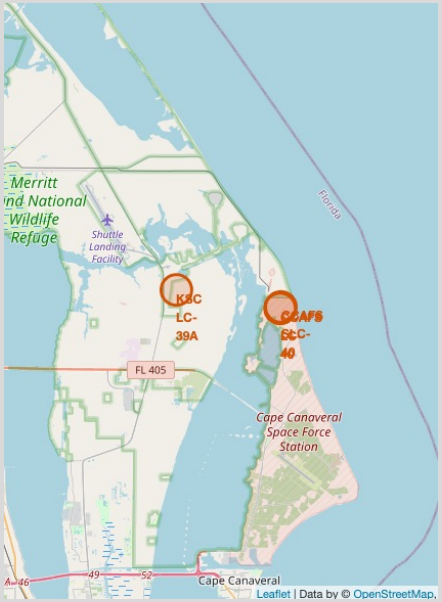
Out[23]:

DATE	COUNT
2015-12-22	1
2016-04-08	1
2016-05-06	1
2016-05-27	1
2016-07-18	1
2016-08-14	1
2017-01-14	1
2017-02-19	1


Launch
Site
Analysis

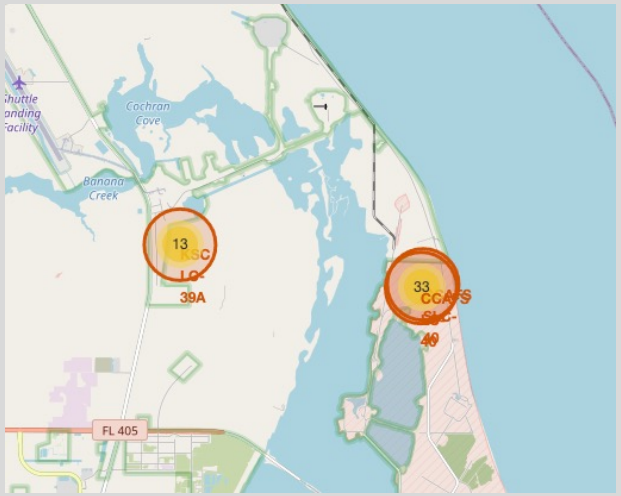
All Launch Site With Folium

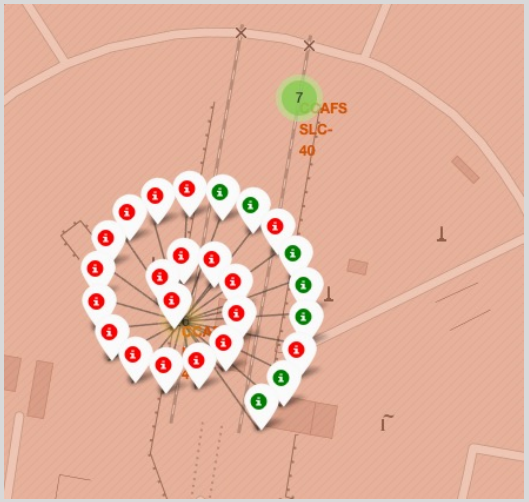




✓ All launch site can be seen in this map
✓ All Launch site is close to the coastline

Launch Site With Success/ Failure Label

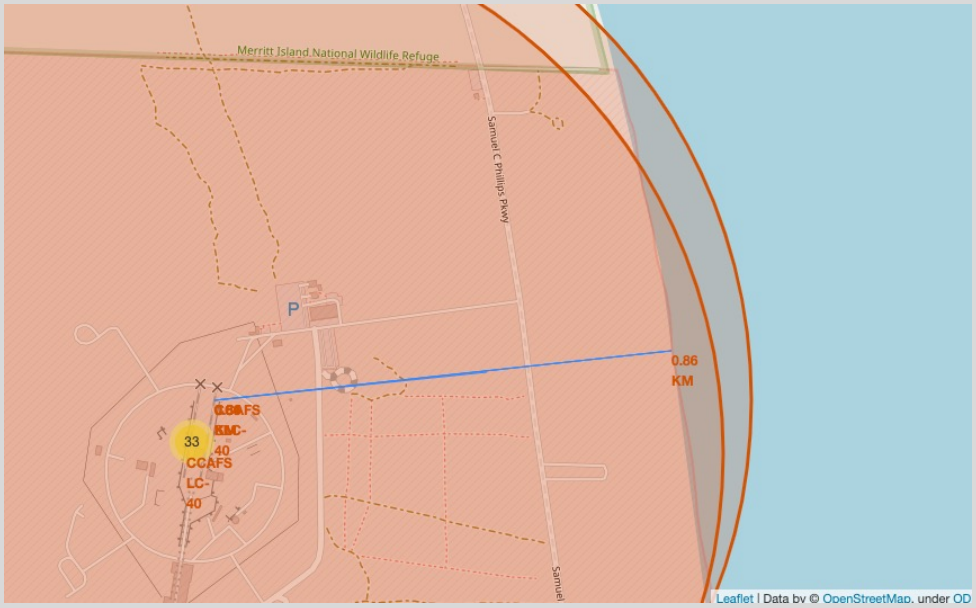




✓ Green = Success Launch , Red = Failure Launch

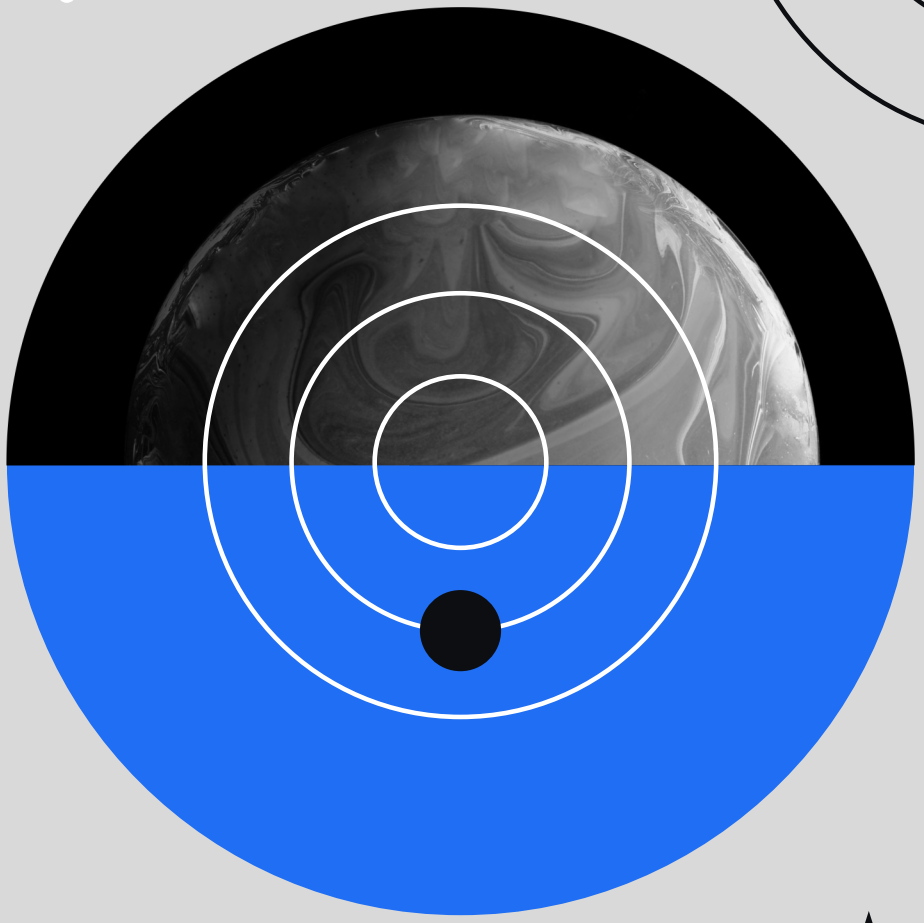
✓ Easy to drill down each launch site to see the success rate

Distances between Launch site to the coastline

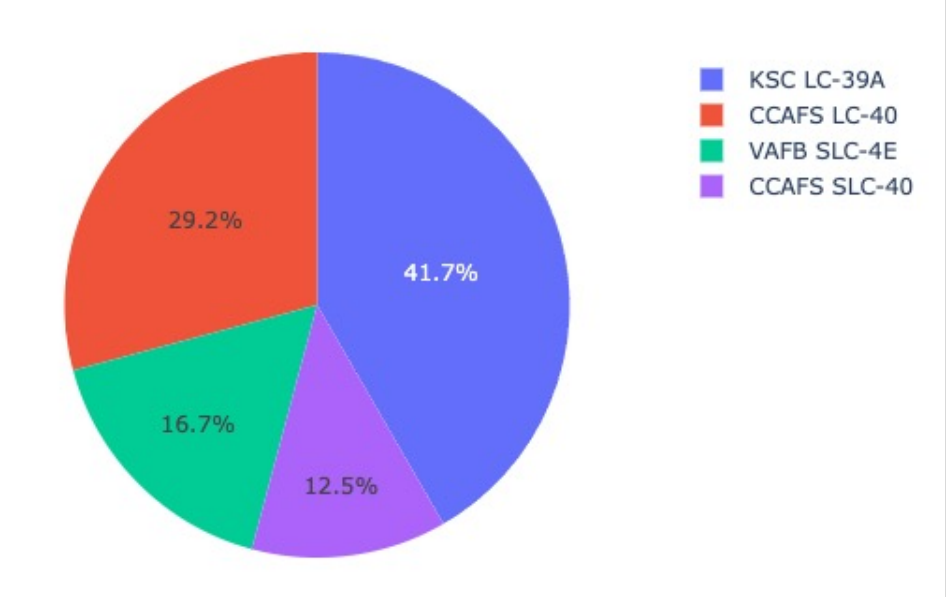


- ✓ Distance can be calculated and displayed In the map
- ✓ Easily draw a polyline between a launch site to the coastline point

Dashboard with Plotly dash

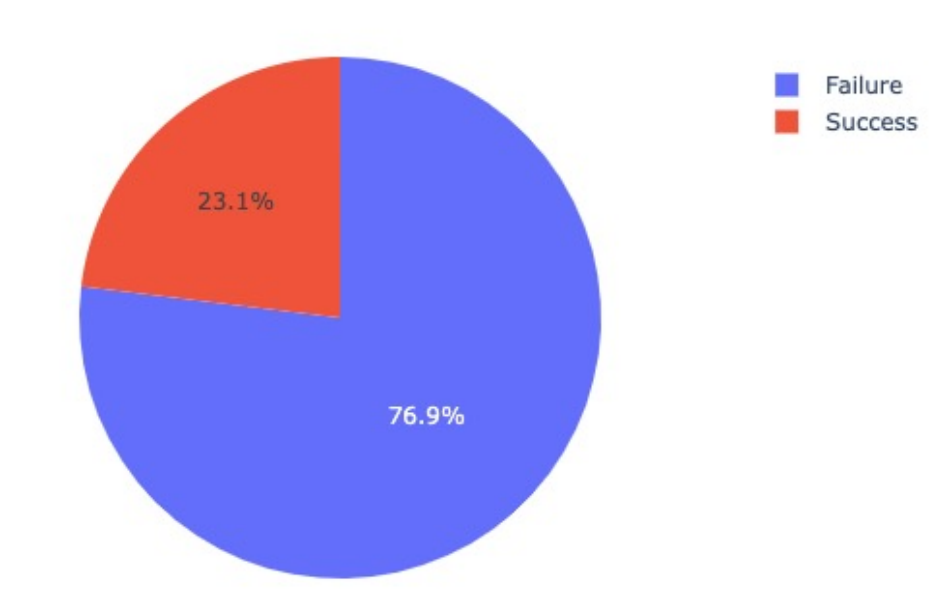


Success Percentage by Site



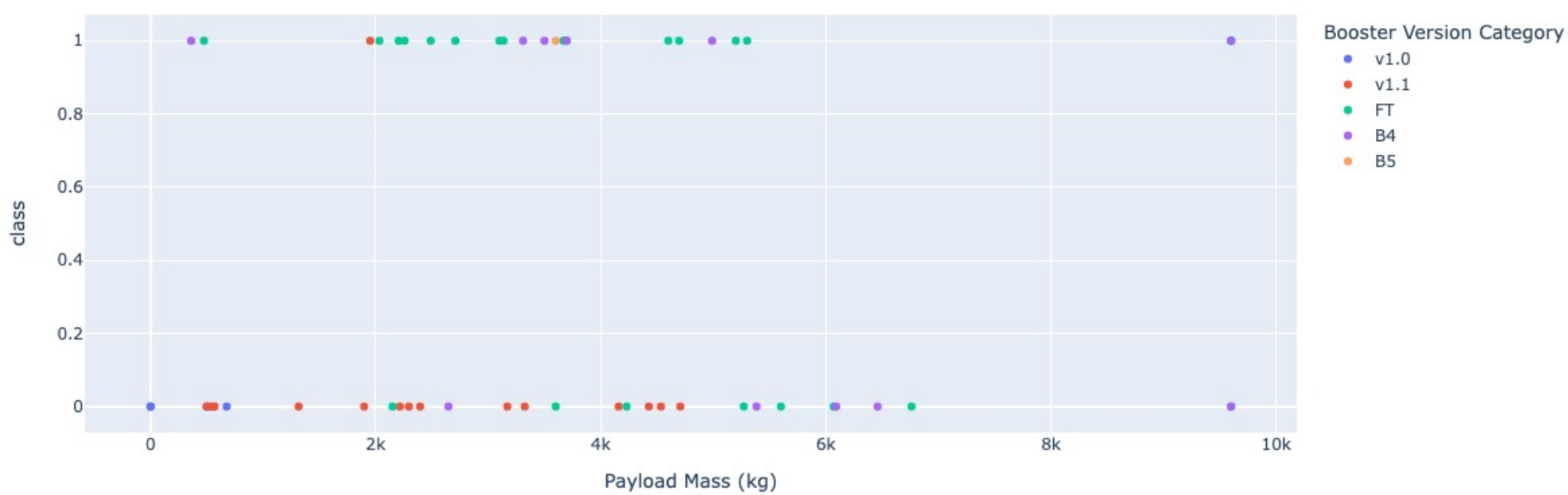
✓ **KSC LC-39 had the most successful launches**

Success Rate in “KSC LC-39”

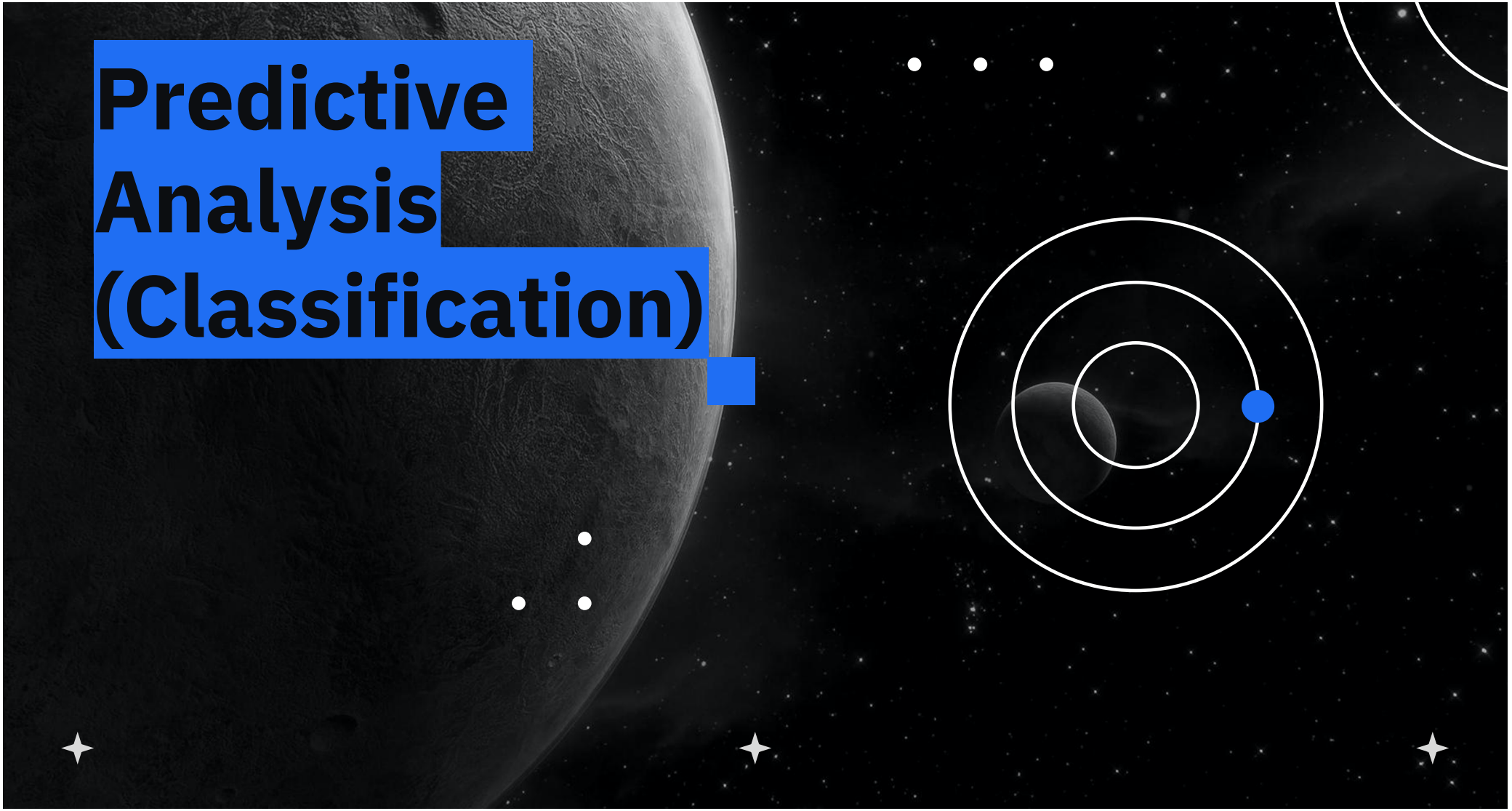


✓ **Success rate in KSC LC-39 is 76.9%**

Payload Mass vs Success by Booster Version Category



✓ **Booster version = FT and payload mass between 2K to 4K shows higher success launch rate**

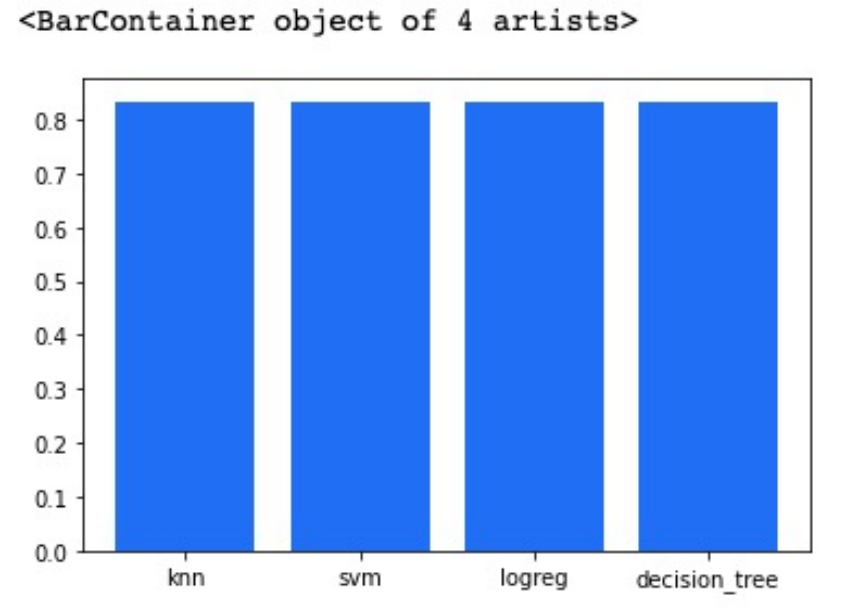

Predictive
Analysis
(Classification)

Classification Accuracy

83.3%

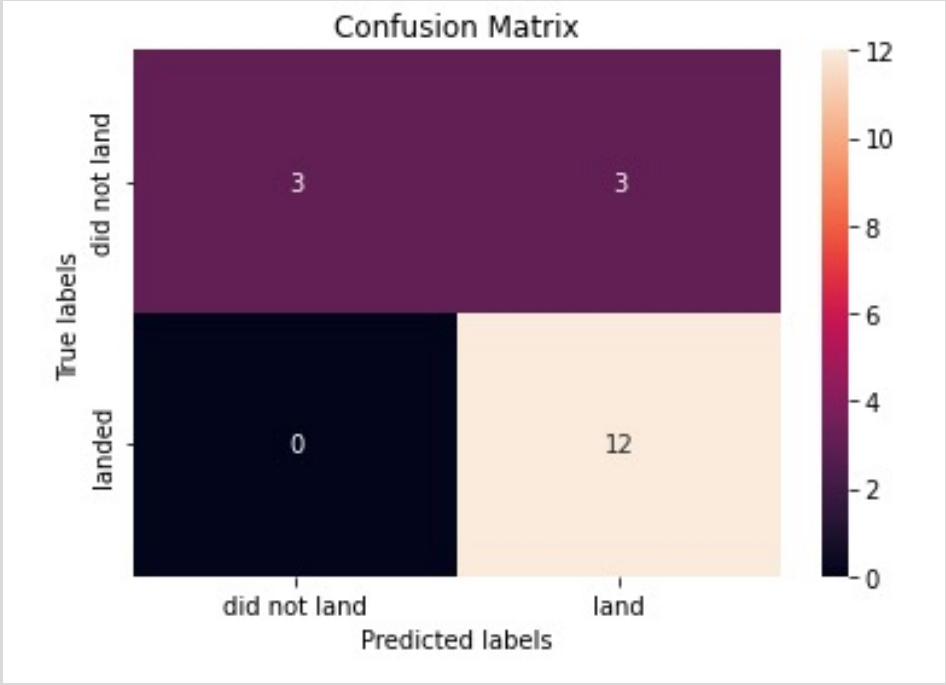
All models show the same accuracy

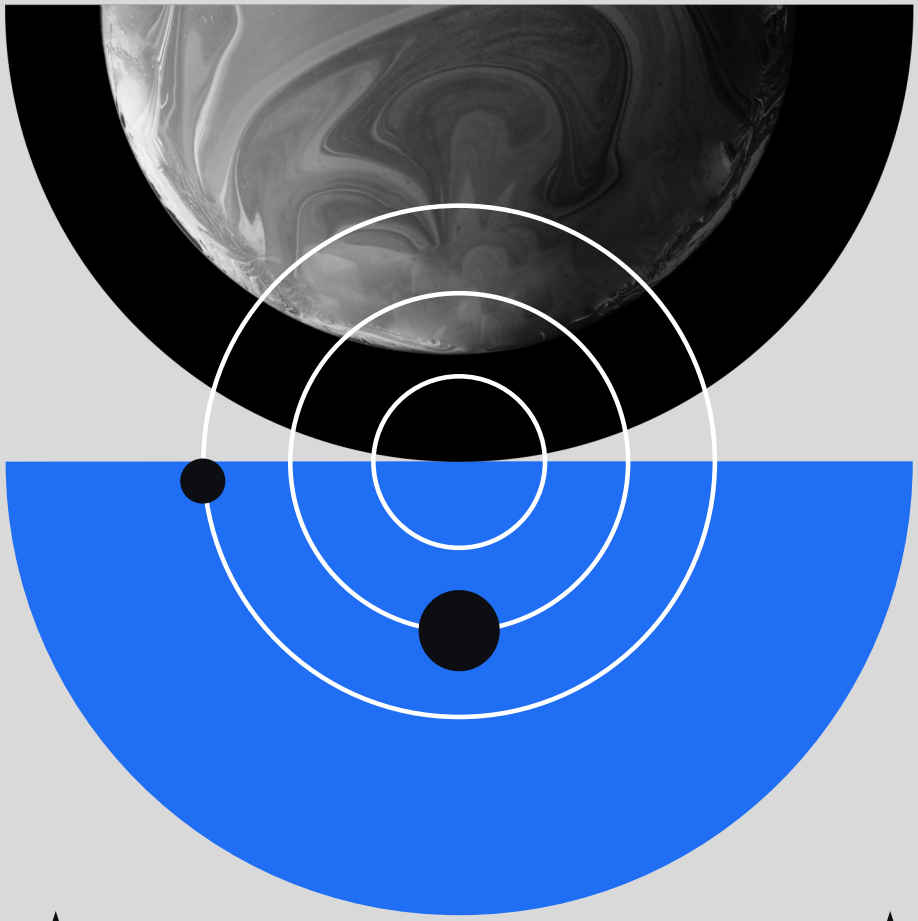
- ✓ KNN
- ✓ SVM
- ✓ Logistic Regression
- ✓ Decision Tree



Confusion Matrix

- ✓ 12 : Predicted correctly as “Land”
- ✓ 3 : Predicted as “Land” but actually ”Did not Land”
- ✓ 3 : Predicted correctly as “Did not Land”





05

Conclusion

Conclusion

- Build a model to predict if SpaceX will reuse the first stage

#01 : Successfully created a machine learning model with 83% of accuracy

#02 : 61% of launches were fired off from “CCAFS SLC 40” with the highest success rate

#03 : Orbit “GEO”,”HEO”,SSO”,”ES-L1” demonstrate 100% success rate

#04 : Success rate is improved rapidly from 2013

#05 : We need more data to build more accurate model

06 Appendix

Links

#01 : Data Collection

#02 : Data collection by scraping

#03 : Data Wrangling

#04 : Dash code

#05 : EDA with SQL

Links

#05 : EDA with Visualization

#06 : Interactive Visualization with Folium

#07 : Machine Learning Model