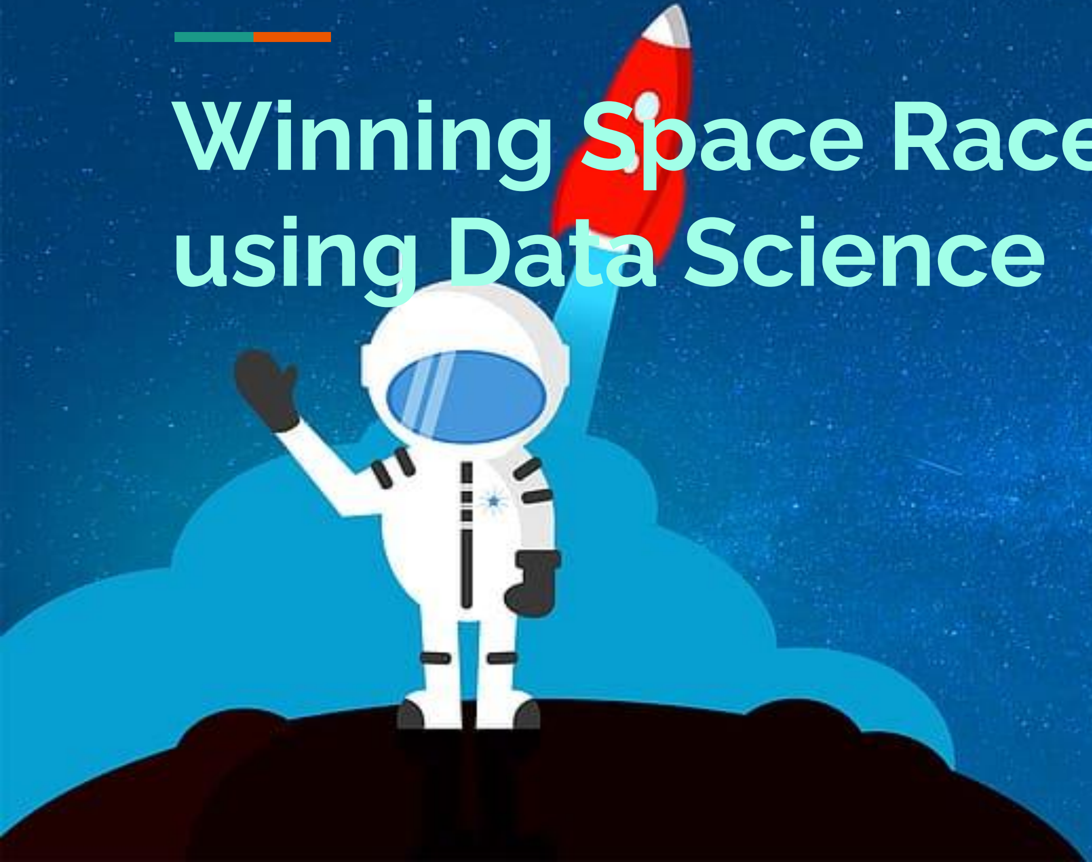
Winning Space Race using Data Science



Lokesh Raja

Jul 04, 2022

Outline

Introduction

Business Problems

Executive Summary

Methodology

Data Collection

Data Wrangling

EDA

Interactive Map

Dashboard

Predictive Modeling

Results EDA 1

Results EDA 2

Results Folium Maps

Results Dashboard

Conclusion

Appendix

Introduction

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.

Problems to solve

1. Can we visualize the factors affecting the success of first stage landing?

2. How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?

3. Does the rate of successful landings increase over the years?

4. What is the best algorithm that can be used for binary classification in this case?

Executive Summary

The study covers various methodologies for data collection & wrangling, data exploration, model building & evaluation. The study also covered designing interactive maps & dashboards to deep dive into success assessment based on various factors. To aid the final model multiple univariate bivariate and multivariate analysis have been carried out.

The best predictive model turned out to be Decision Tree Based Classification & the factors influencing the most turned out to be launch site, Payload, Destination Orbit.

Methodologies followed

1. Data collection methodology -
   - ➔ Request to the SpaceX API & selecting relevant data
   - ➔ Web Scraping using HTML & Beautiful Soup package
2. Perform data wrangling
   - ➔ Relevant data is selected & cleaned from collected data
   - ➔ Outcome variable is labelled
   - ➔ Missing Value handling
3. Perform exploratory data analysis (EDA) using visualization and SQL
4. Perform interactive visual analytics using Folium and Plotly Dash
5. Perform predictive analysis using classification models
   - ➔ Building, tuning and evaluation of classification models to ensure the best results

Data Collection

Data Collection
01 SpaceX Rest API

Using SpaceX's REST API, the relevant data was extracted from Website. The SpaceX API jupyter notebook in Git Repo details the step followed.

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

1 **Request Data using API**

Making use of SpaceX API data can be pulled from different data store e.g. Flight, Payload, etc.

2 **Format Data**

Decode data & make it useful for further analysis using Custom build functions & json library.

3 **Selecting Features**

Creating dataframes with selected features

4 **Filter the Data**

Filter the data for Falcon 9 launches

5 **Missing Data Treatment**

Replace the missing data using mean values of Payload mass

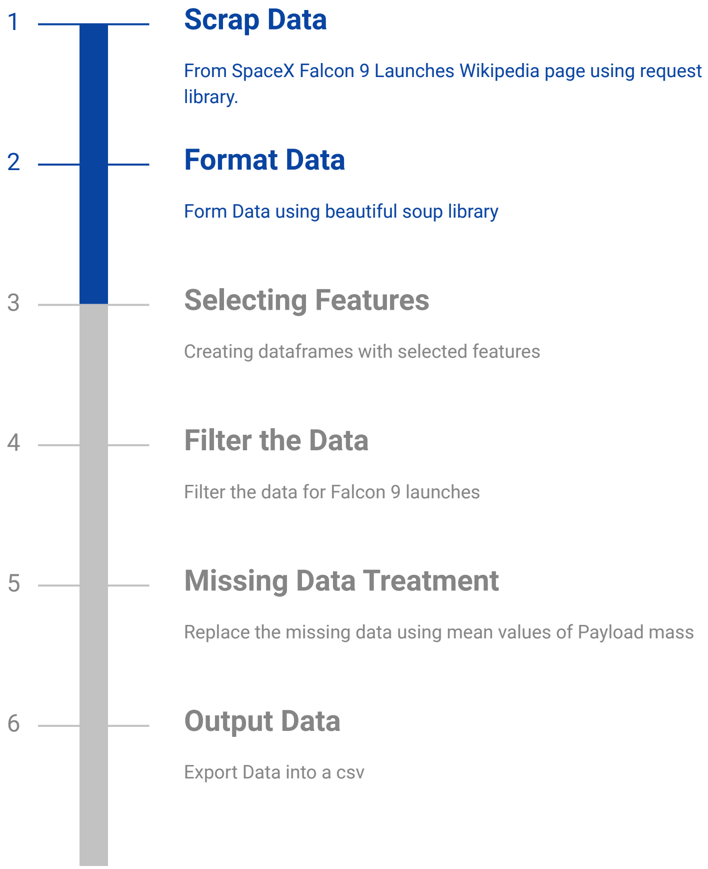
6 **Output Data**

Export Data into a csv

Data Collection
02 Web Scraping

Using SpaceX Wikipedia Page. The Web Scraping jupyter notebook in Git Repo details the step followed.

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



1 **Scrap Data**

From SpaceX Falcon 9 Launches Wikipedia page using request library.

2 **Format Data**

Form Data using beautiful soup library

3 **Selecting Features**

Creating dataframes with selected features

4 **Filter the Data**

Filter the data for Falcon 9 launches

5 **Missing Data Treatment**

Replace the missing data using mean values of Payload mass

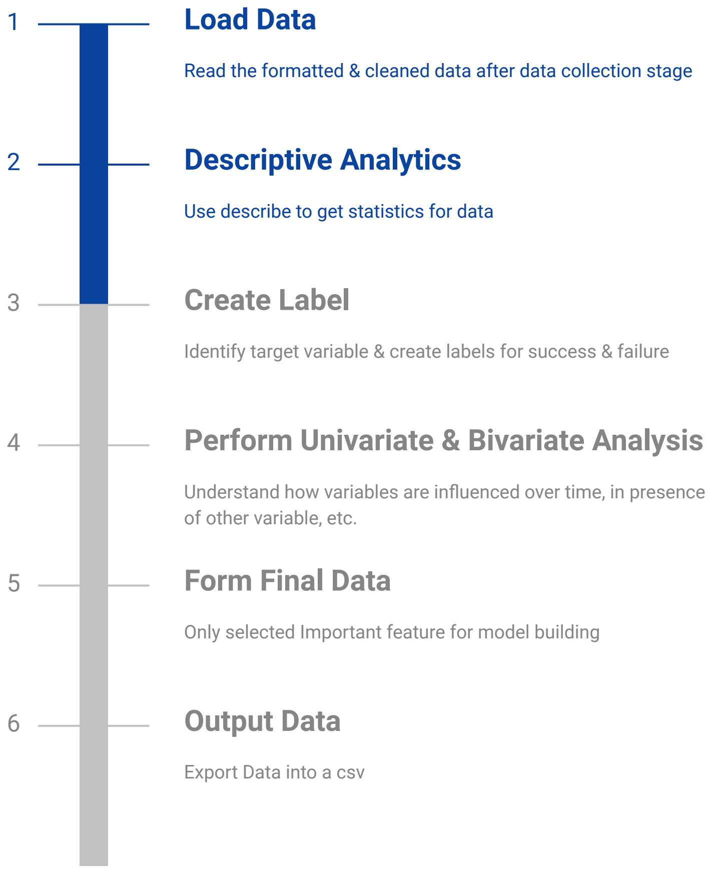
6 **Output Data**

Export Data into a csv

Data Wrangling

The main goal of data wrangling was understanding & cleaning features useful for predicting successful outcome. Labels were created using all possible values of outcome. Data Wrangling notebook details all the steps covered in process.

Takeaways:
Outcomes into Training Labels with “1” means the booster successfully landed, “0” means it was unsuccessful.



1 **Load Data**

Read the formatted & cleaned data after data collection stage

2 **Descriptive Analytics**

Use describe to get statistics for data

3 **Create Label**

Identify target variable & create labels for success & failure

4 **Perform Univariate & Bivariate Analysis**

Understand how variables are influenced over time, in presence of other variable, etc.

5 **Form Final Data**

Only selected Important feature for model building

6 **Output Data**

Export Data into a csv

Exploratory Data Analysis

EDA

01 Data Visualization

1

Charts 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
- ➔ Success Rate Yearly Trend

Refer to Data Viz (EDA) Notebook for detailed steps

2

- ➔ 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)

EDA
02 SQL

1

Written SQL for Displaying -

- ➔ Names of the unique launch sites in the space mission
- ➔ 5 records where launch sites begin with the string ‘CCA'
- ➔ Total payload mass carried by boosters launched by NASA (CRS)
- ➔ Average payload mass carried by booster version F9 v1.1

Refer to SQL (EDA) Notebook for all the queries

2

- ➔ Date when the first successful landing outcome in ground pad was achieved
- ➔ Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- ➔ Total number of successful and failure mission outcomes
- ➔ Names of the booster versions which have carried the maximum payload mass
- ➔ 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

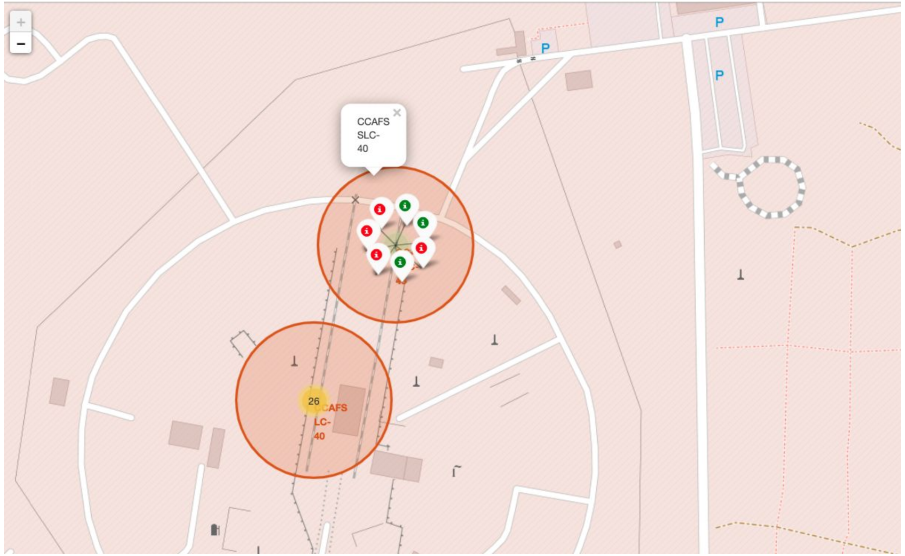
Interactive Map using Folium

Markers for Launch Sites -

- ➔ Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.
- ➔ Marker with Circle, Popup Label and Text Label of other sites using latitude and longitude coordinates.

Refer to [Folium Map](#) Notebook for detailed steps

- ➔ Markers are coloured Red for Failure & Green for Success for each launch site.
- ➔ 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.



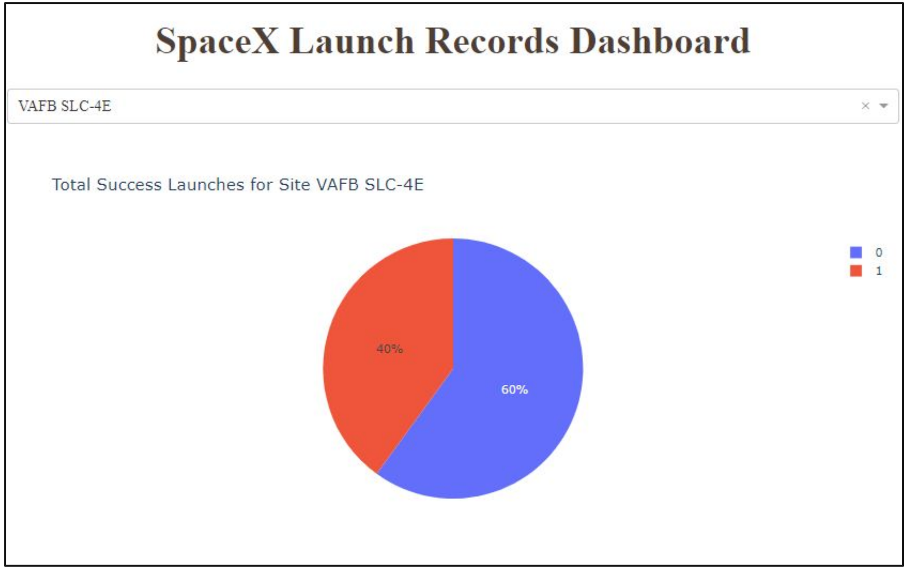
SpaceX Launch Site Dashboard

Multiple Interactions are listed below -

- **Launch Sites Dropdown List -** To enable Launch Site selection.
- **Pie Chart showing Success Launches (All Sites/Certain Site) -** 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.

Refer to App Dash Notebook for detailed steps

- **Slider of Payload Mass Range -** To select Payload range.
- **Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions -** To show the correlation between Payload and Launch Success.



SpaceX Launch Records Dashboard

VAFB SLC-4E

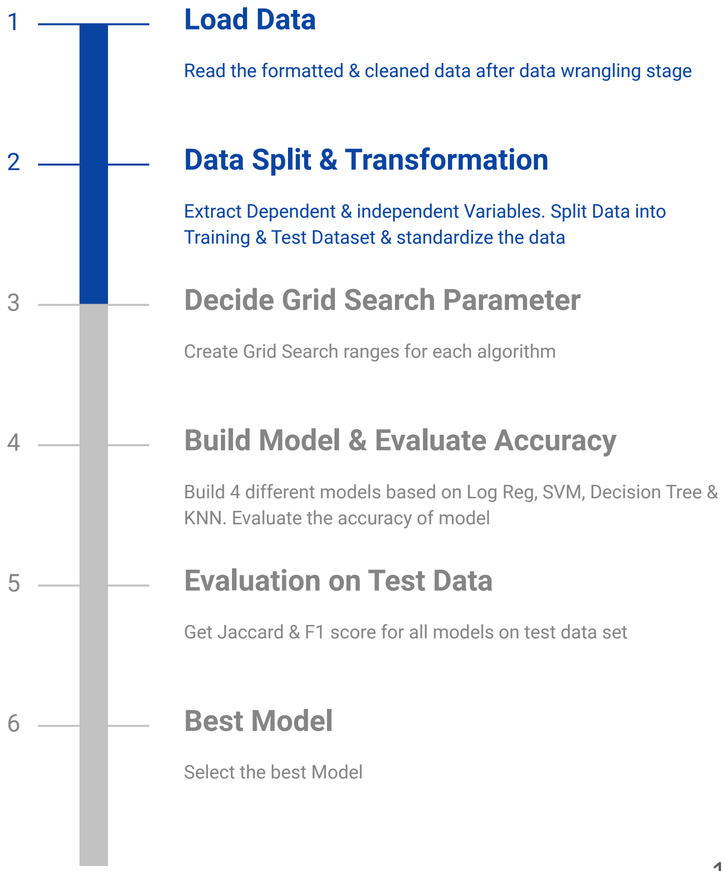
Model Building & Evaluation

Prediction is done using Classification techniques and 4 algorithms were tested -

- ➔ Logistic Regression
- ➔ K Nearest Neighbour
- ➔ Support Vector Machine
- ➔ Decision Tree

All Models were evaluated on Test & Train Dataset using Jaccard score, F1 score, etc.

Refer to Predictive Modeling Notebook for detailed steps



1 **Load Data**

Read the formatted & cleaned data after data wrangling stage

2 **Data Split & Transformation**

Extract Dependent & independent Variables. Split Data into Training & Test Dataset & standardize the data

3 **Decide Grid Search Parameter**

Create Grid Search ranges for each algorithm

4 **Build Model & Evaluate Accuracy**

Build 4 different models based on Log Reg, SVM, Decision Tree & KNN. Evaluate the accuracy of model

5 **Evaluation on Test Data**

Get Jaccard & F1 score for all models on test data set

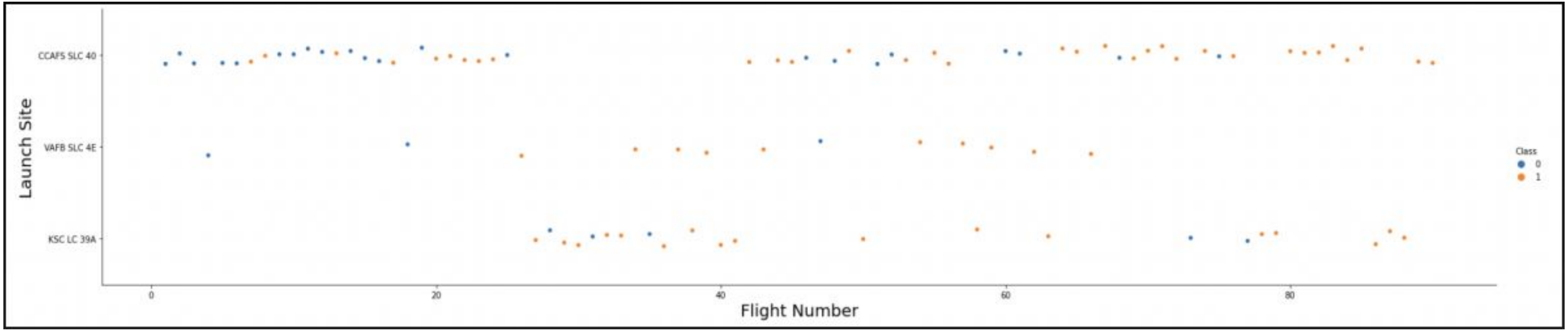
6 **Best Model**

Select the best Model

Results
Data Visualization EDA

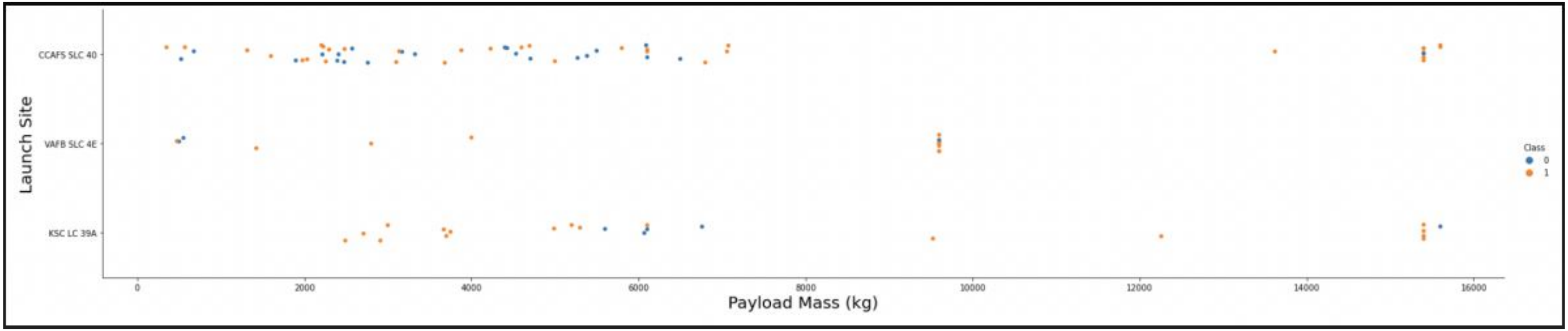
Flight Number vs. Launch Site

- ➔ 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.



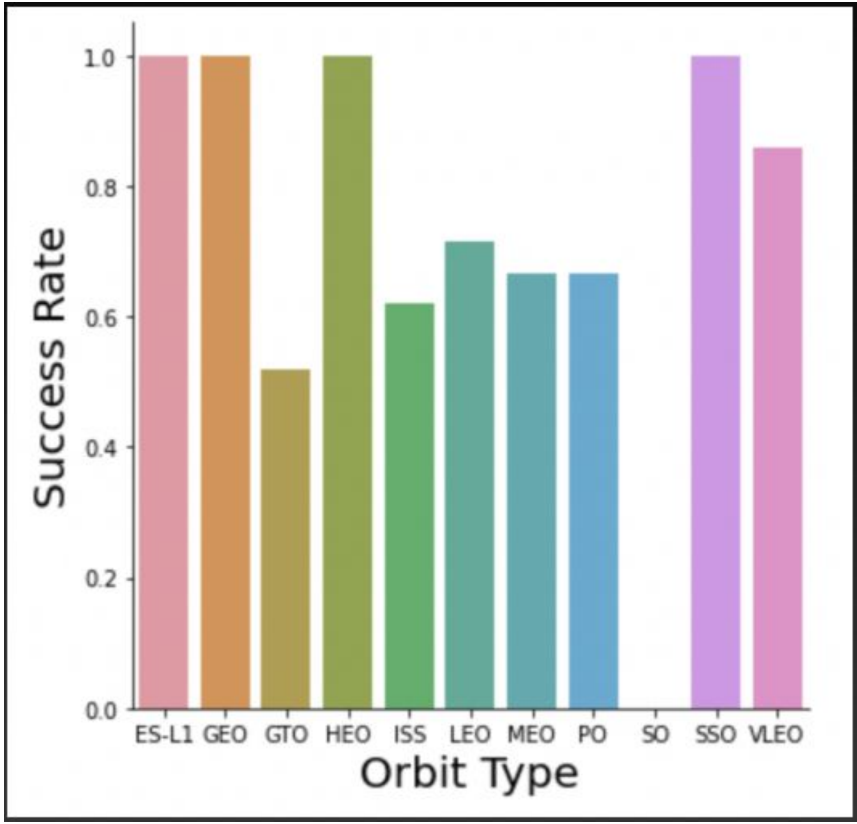
Payload vs. Launch Site

- ➔ 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



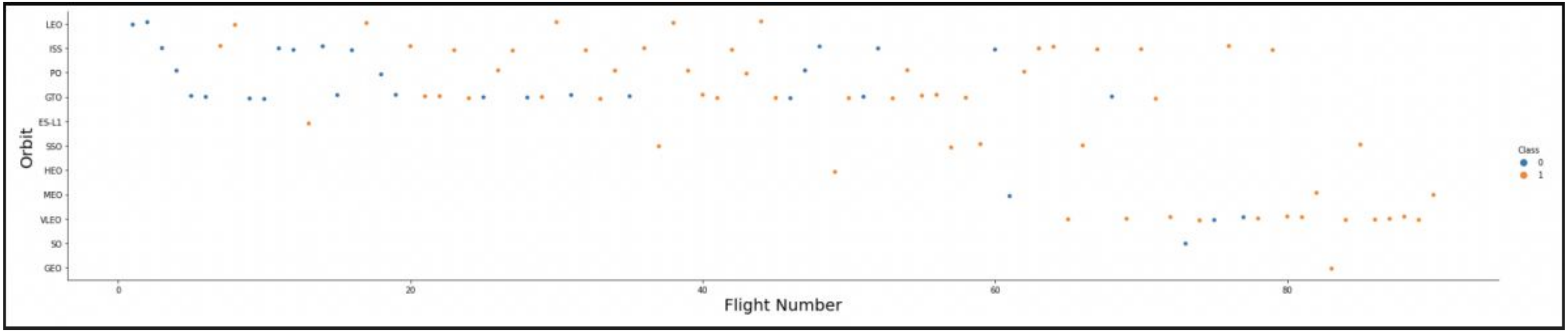
Success rate vs. Orbit type

- ➔ 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



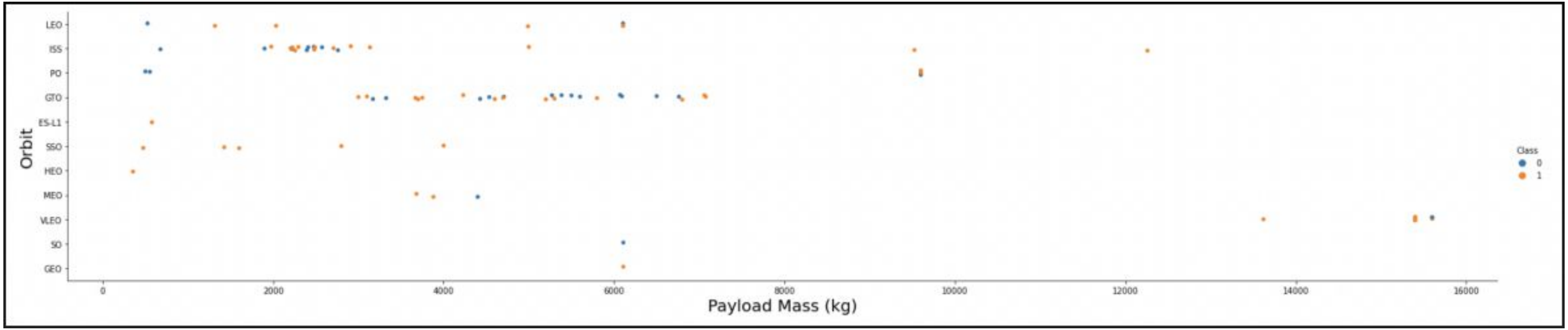
Flight Number vs. Orbit Type

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



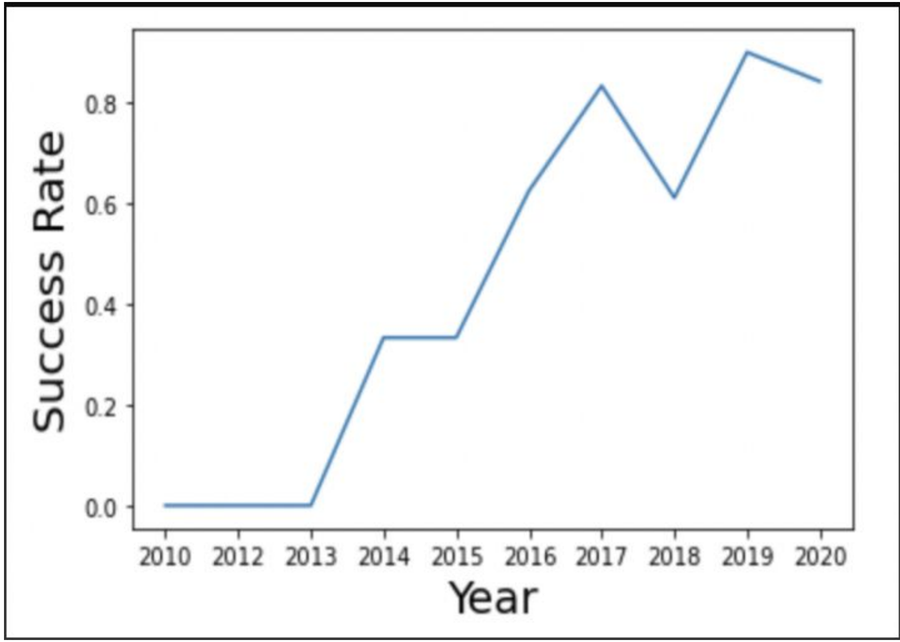
Payload Mass vs. Orbit Type

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



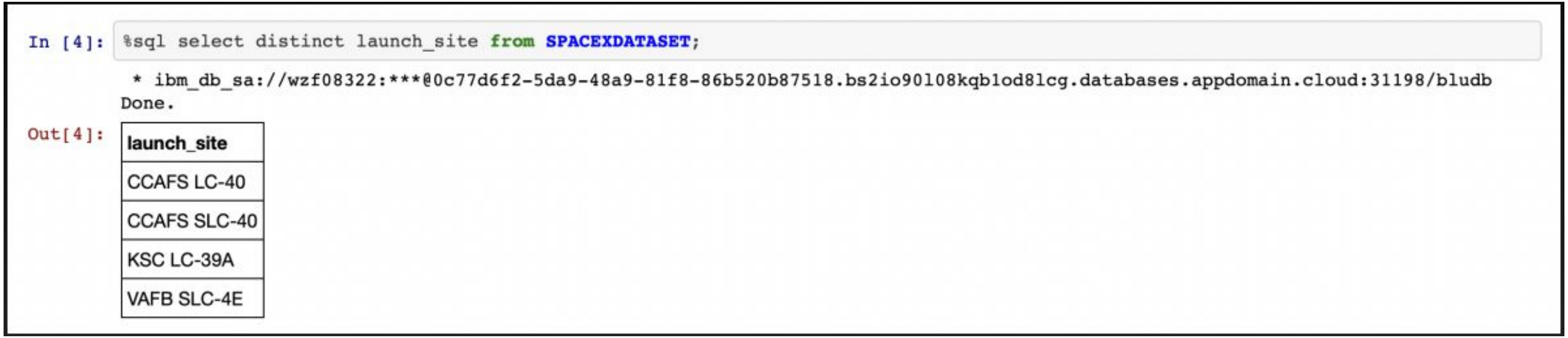
Launch success yearly trend

The success rate since 2013 kept increasing till 2020



Results
SQL Based EDA

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

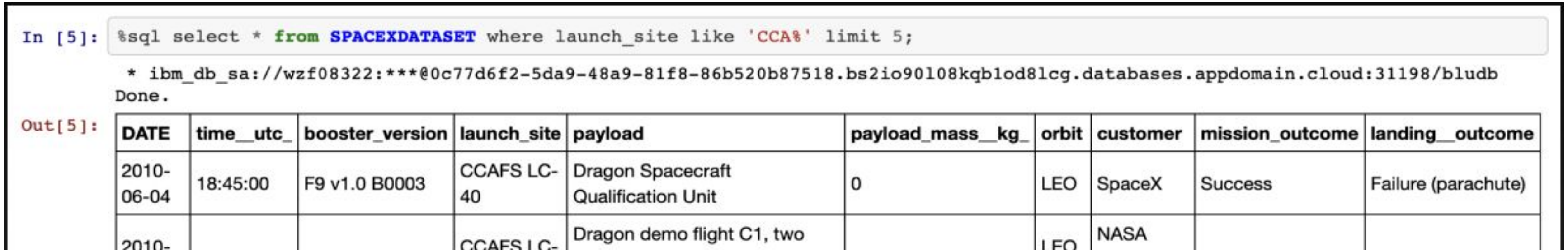


```
In [4]: %sql select distinct launch_site from SPACEXDATASET;

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[4]:
```

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

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

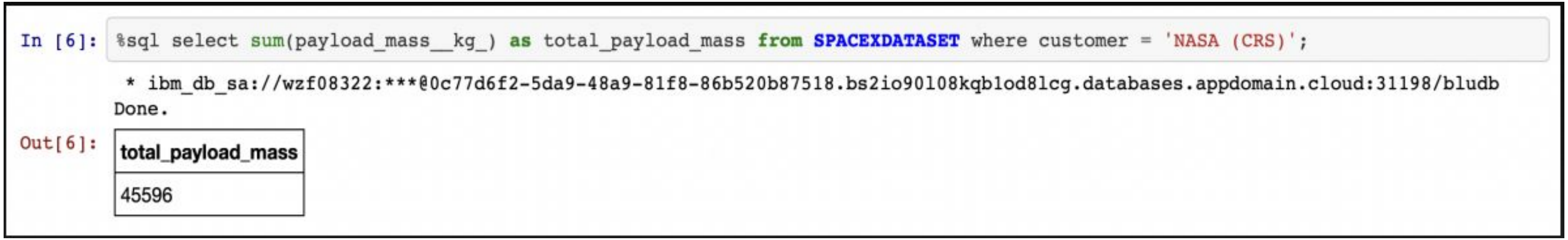


```
In [5]: %sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5;

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[5]:
```

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-			CCAFS LC-	Dragon demo flight C1, two		LEO	NASA		

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



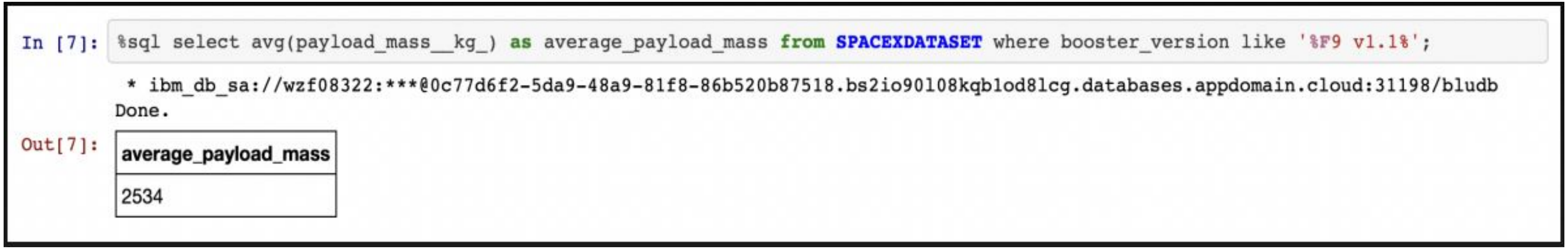
```
In [6]: %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)';

         * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
        Done.
```

Out[6]:

total_payload_mass
45596

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



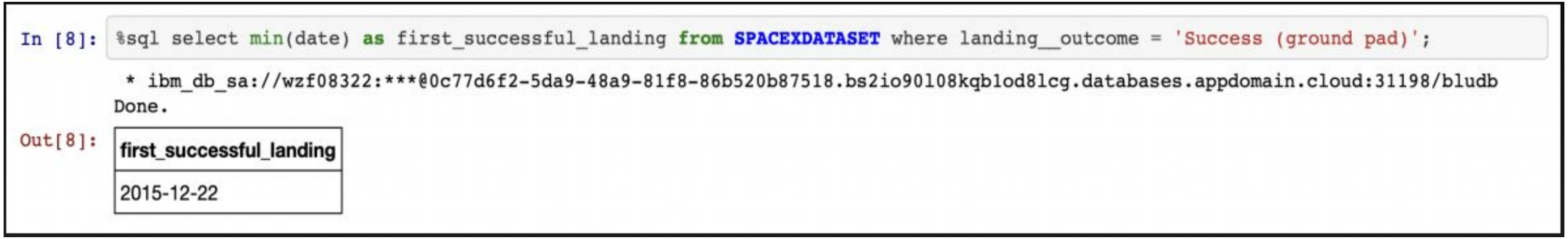
```
In [7]: %sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXDATASET where booster_version like '%F9 v1.1%';

         * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
        Done.
```

Out[7]:

average_payload_mass
2534

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

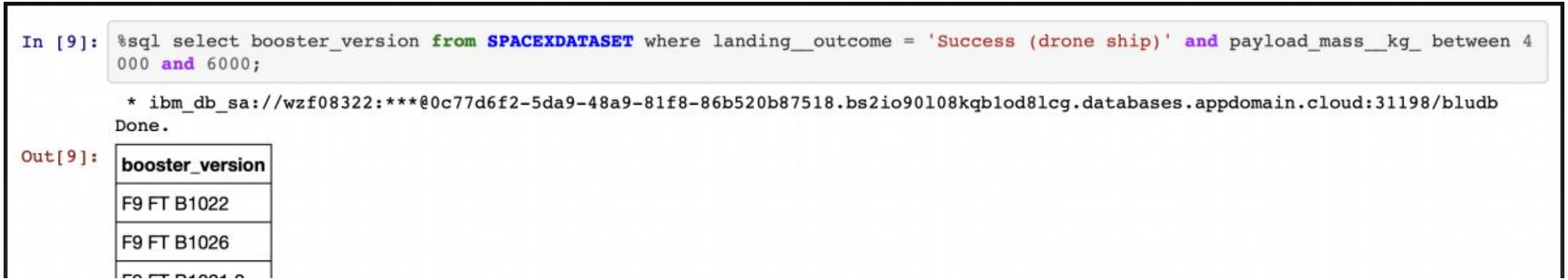


```
In [8]: %sql select min(date) as first_successful_landing from SPACEXDATASET where landing__outcome = 'Success (ground pad)';

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[8]:
```

first_successful_landing
2015-12-22

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

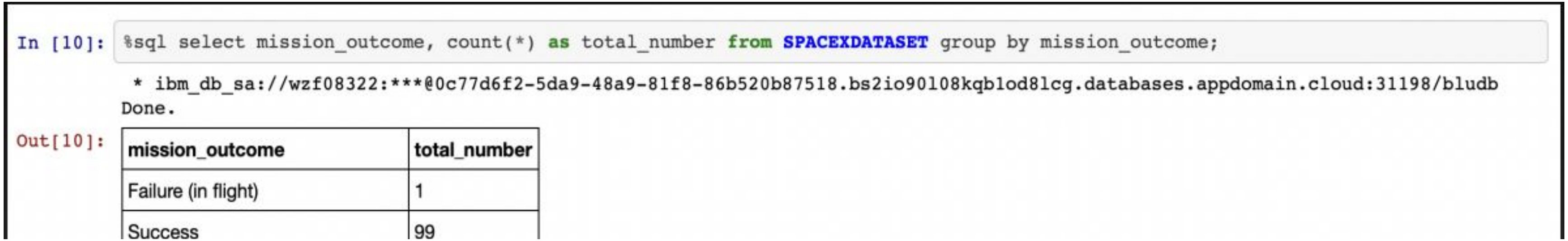


```
In [9]: %sql select booster_version from SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4
        000 and 6000;

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[9]:
```

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

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

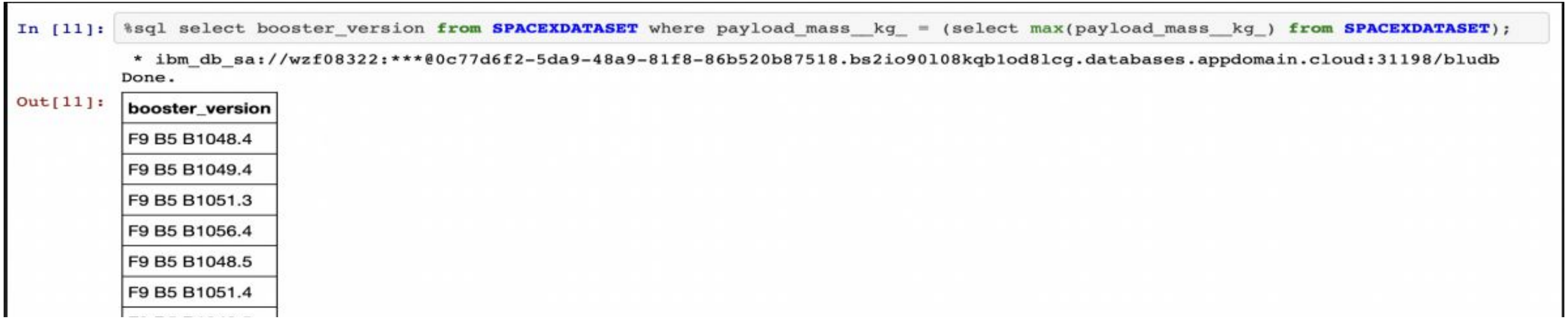


```
In [10]: %sql select mission_outcome, count(*) as total_number from SPACEXDATASET group by mission_outcome;

          * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
         Done.
Out[10]:
```

mission_outcome	total_number
Failure (in flight)	1
Success	99

8.List the names of the booster versions which have carried the maximum payload mass

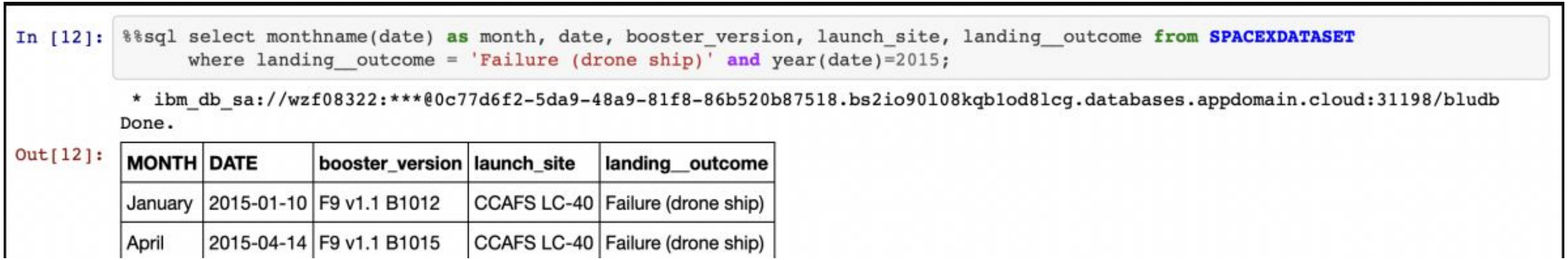


```
In [11]: %sql select booster_version from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXDATASET);

          * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
         Done.
Out[11]:
```

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

9. List the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015

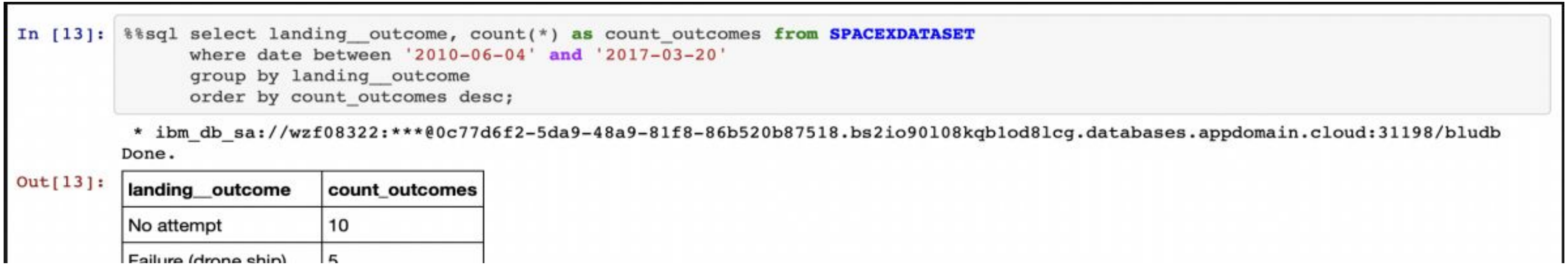


```
In [12]: %%sql select monthname(date) as month, date, booster_version, launch_site, landing__outcome from SPACEXDATASET
                where landing__outcome = 'Failure (drone ship)' and year(date)=2015;

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[12]:
```

MONTH	DATE	booster_version	launch_site	landing__outcome
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

10.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



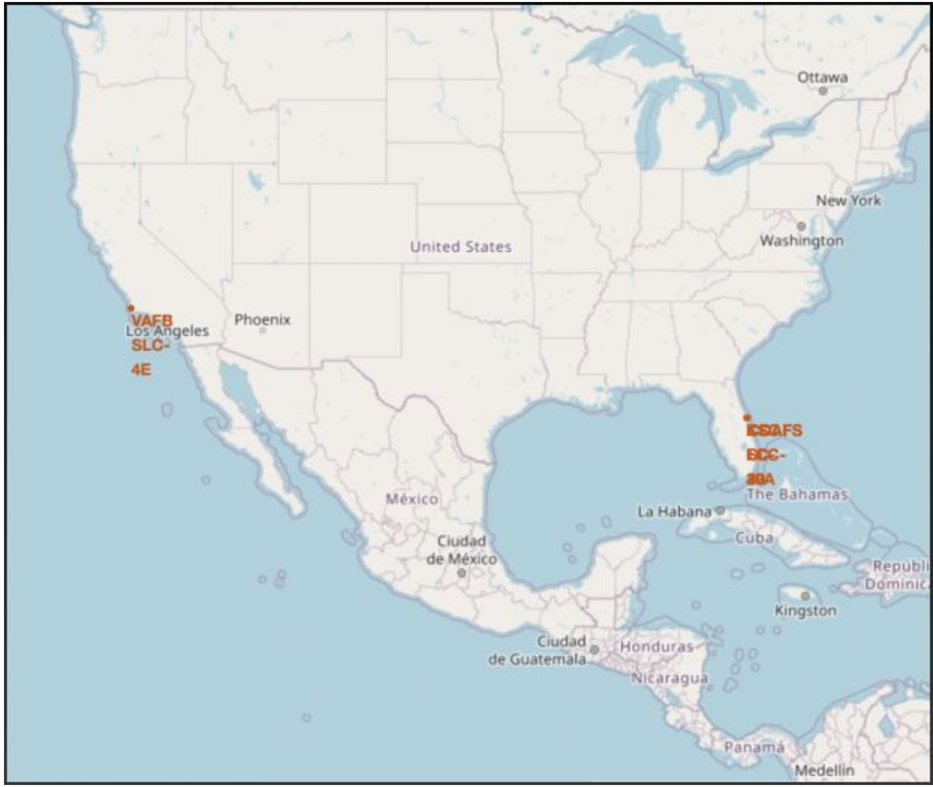
```
In [13]: %%sql select landing__outcome, count(*) as count_outcomes from SPACEXDATASET
                where date between '2010-06-04' and '2017-03-20'
                group by landing__outcome
                order by count_outcomes desc;

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[13]:
```

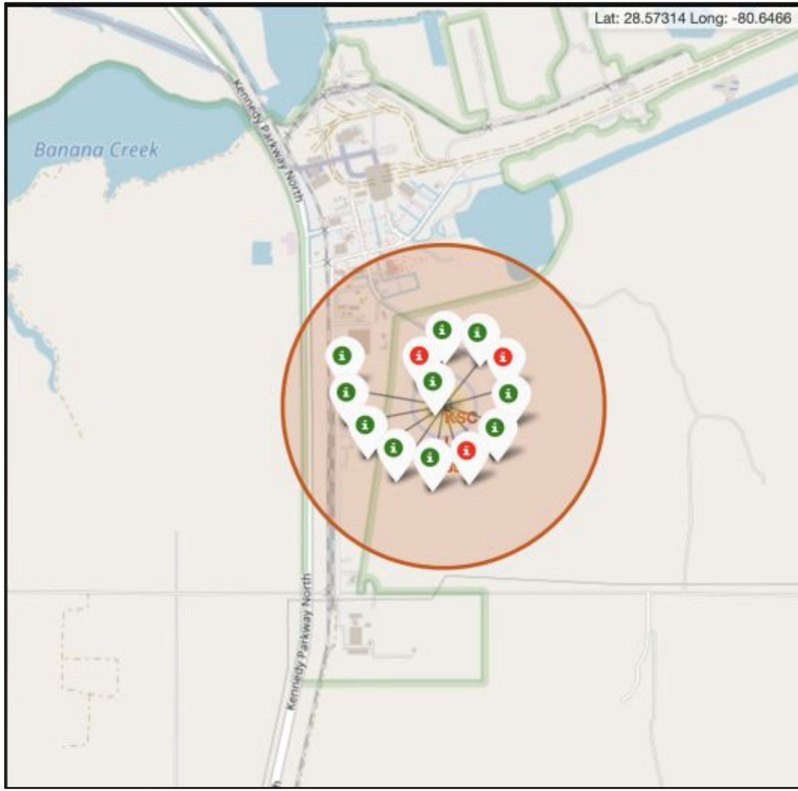
landing__outcome	count_outcomes
No attempt	10
Failure (drone ship)	5

Results
Interactive Folium Map

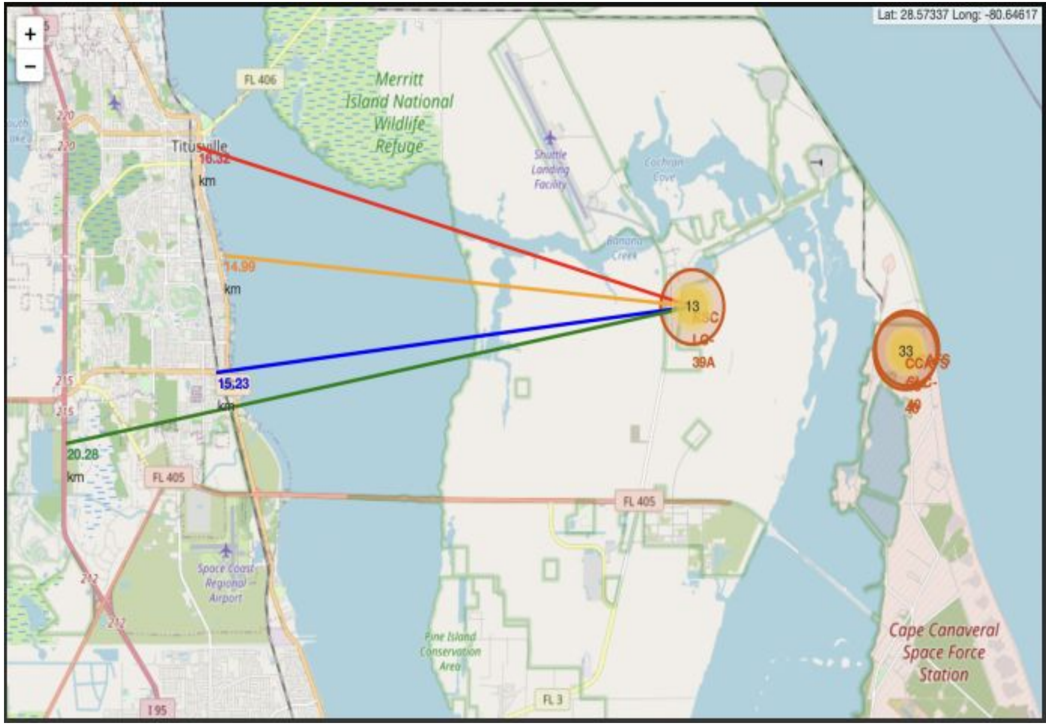
- ➔ Most of Launch sites are in proximity to the Equator line. The land is moving faster at the equator than any other place on the surface of the Earth. Anything on the surface of the Earth at the equator is already moving at 1670 km/hour. If a ship is launched from the equator it goes up into space, 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.



- ➔ 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.

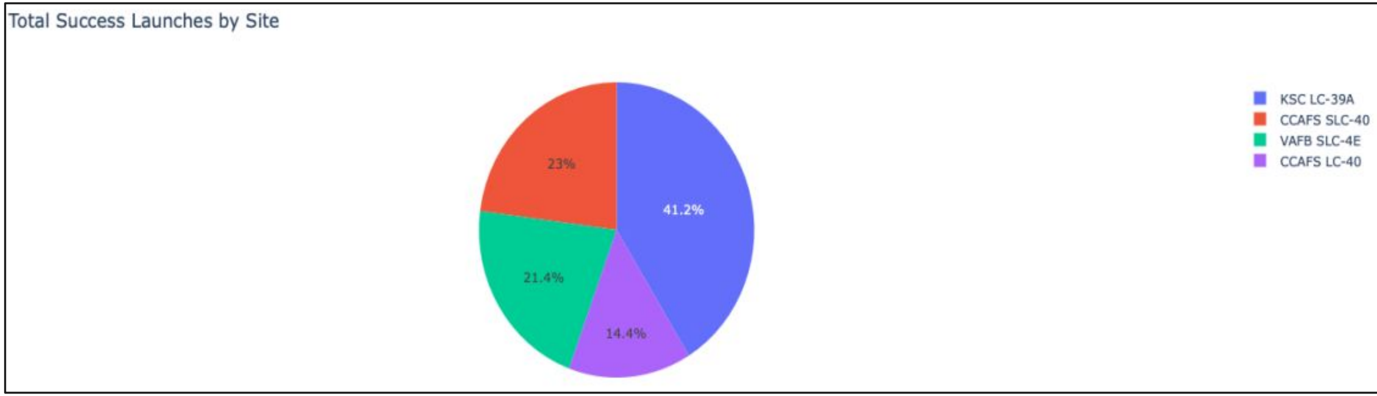


- ➔ From the visual analysis of the launch site KSC LC-39A we can clearly see that it is:
  - ◆ relative close to railway (15.23 km)
  - ◆ relative close to highway (20.28 km)
  - ◆ relative close to coastline (14.99 km)
- ➔ Also the launch site KSC LC-39A is relative close to its closest city Titusville (16.32 km).
- ➔ Failed rocket with its high speed can cover distances like 15-20 km in few seconds. It could be potentially dangerous to populated areas.

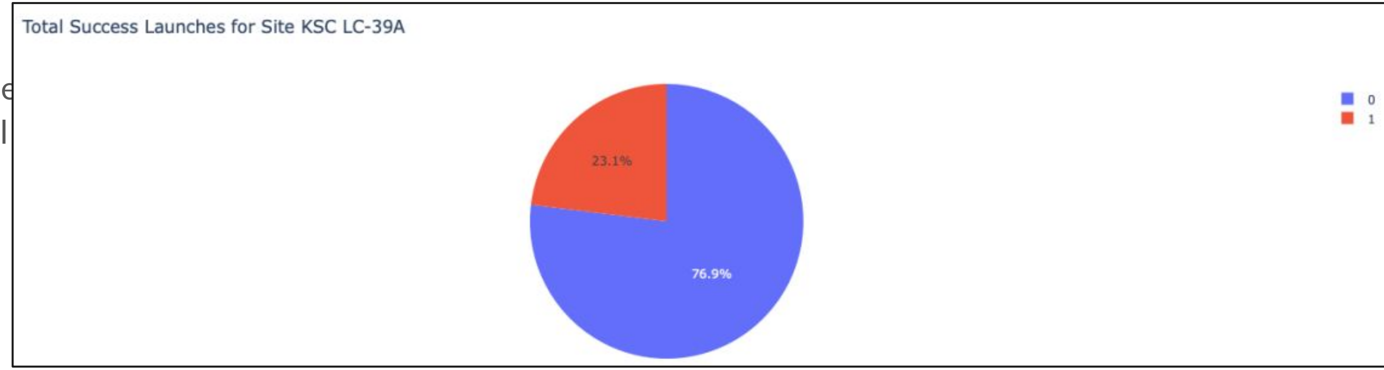

Lat: 28.57337 Long: -80.64617
Merritt Island National Wildlife Refuge
Titusville
16.32 km
14.99 km
15.23 km
20.28 km
13
KSC LC-39A
33
CCAFS SLC-40
Shuttle Landing Facility
Space Coast Regional Airport
Pine Island Conservation Area
Cape Canaveral Space Force Station
FL 406
FL 405
FL 3
I 95

Results Dashboard

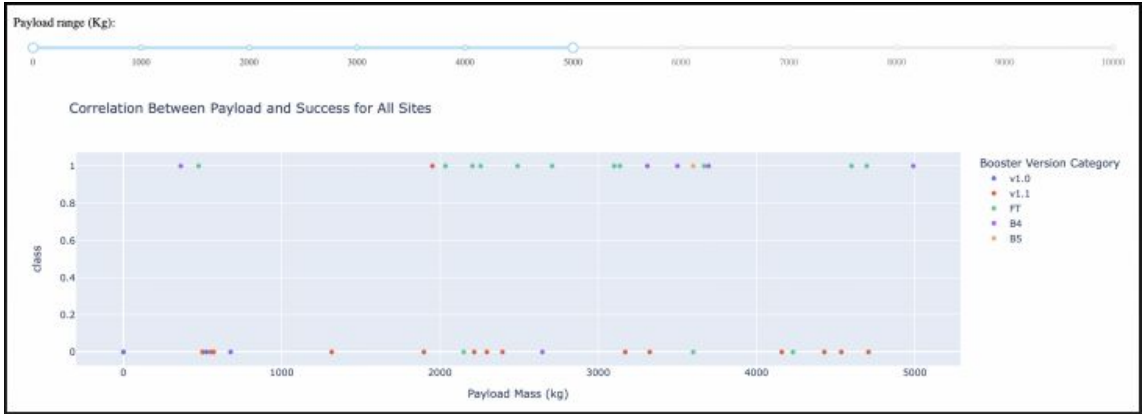
1. The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches

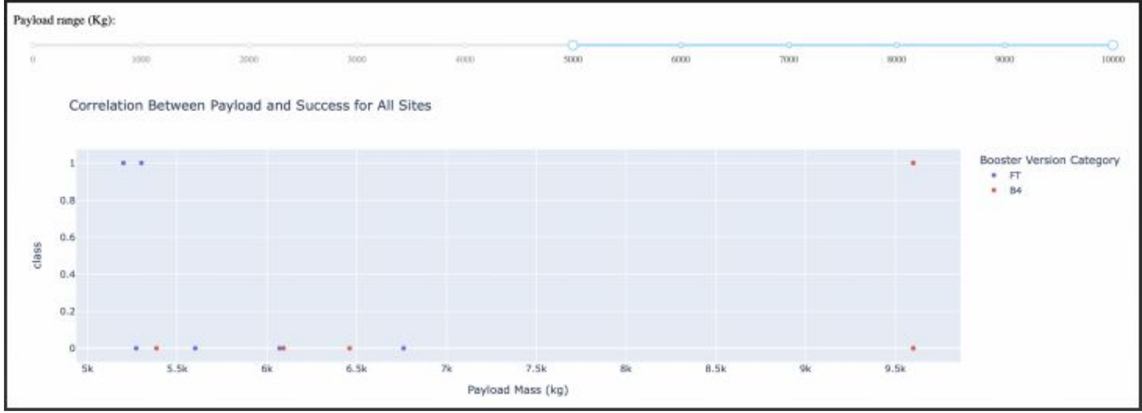


2. KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings



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





Conclusion

Below mentioned points are also sharing some insights in the data & overall spaceX landing effort attempts.

01 | Decision Tree Model is the best algorithm for this dataset

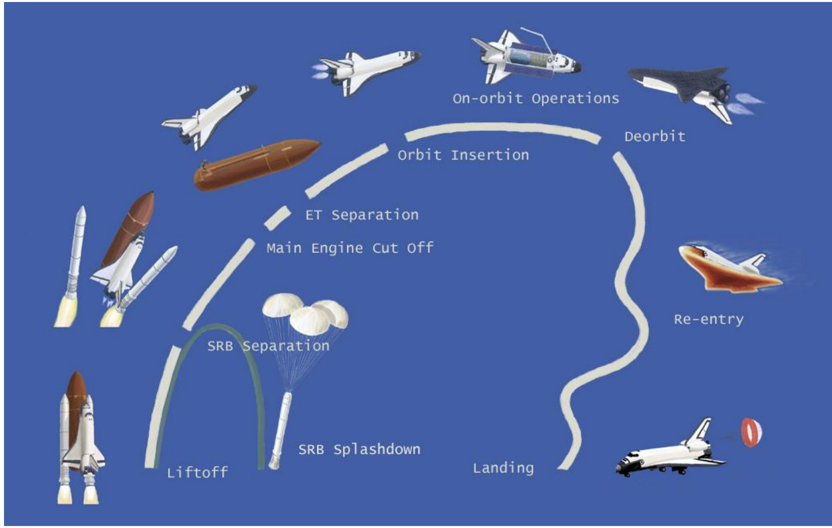
02 | Low payload launches are more successful than high payload launches

03 | Launch sites are near equator & coast line

04 | Success Rate has increased over year

05 | KSC LC-39A has the highest success rate

06 | Orbits ES-L1, GEO, HEO and SSO have 100% success rate



Thank you.

Instructors

IBM

coursera