IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Aïda Piñol
September 1, 2022

Outline

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

1. Executive Summary

IBM Developer SKILLS NETWORK

- SpaceX advertises Falcon 9 rocket with a cost of 62 million dollars; other providers cost upwards of 165 million.

- Much of the savings is because **SpaceX can reuse the first stage**.

- SpaceY wants to determine the price of each launch, training a machine learning model and using public information to **predict if SpaceX will reuse the first stage.**

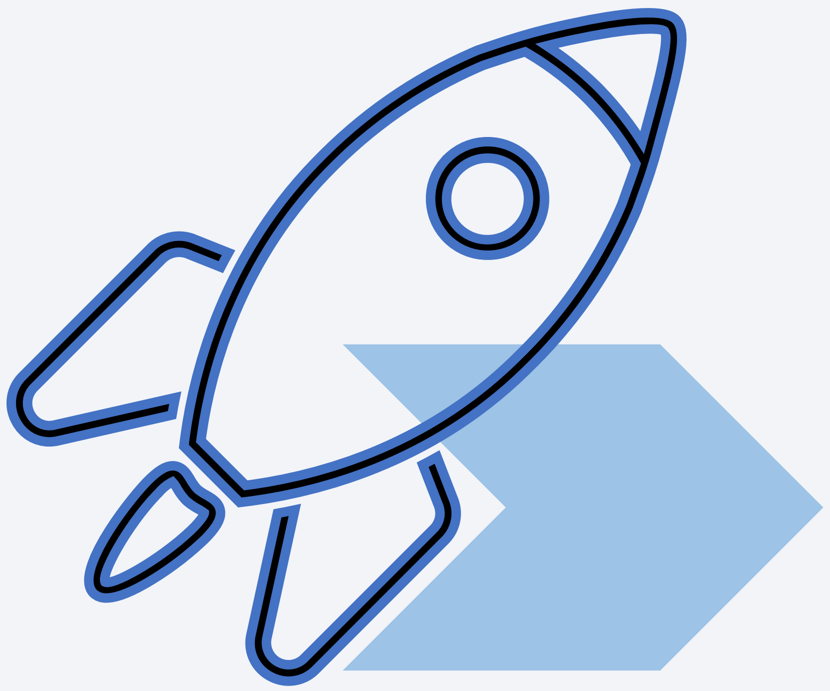
- Public information indicates a first stage Falcon 9 Booster to cost upwards of 15 million.

- The machine learning model can predict, with an **83.3% level of accuracy**, the landing success of the first stage Falcon 9 Booster.

2. Introduction

- PROJECT SCENARIO: Capstone Course

In the Applied Data Science Capstone course, we will **predict if the Falcon 9 first stage will land successfully** as a real-world business problem to solve within our Data Scientist role.



- BUSINESS PROBLEM: Cost of a launch

Determine if the first stage will land in order to **determine the cost of a launch.** This information can be used to bid against SpaceX for a rocket launch.

Section 1

Methodology

3. Methodology: *Executive Summary*

IBM Developer SKILLS NETWORK

a) Data collection methodology: *API & Web Scraping*

b) Perform data wrangling

c) Perform exploratory data analysis (EDA) using visualization and SQL

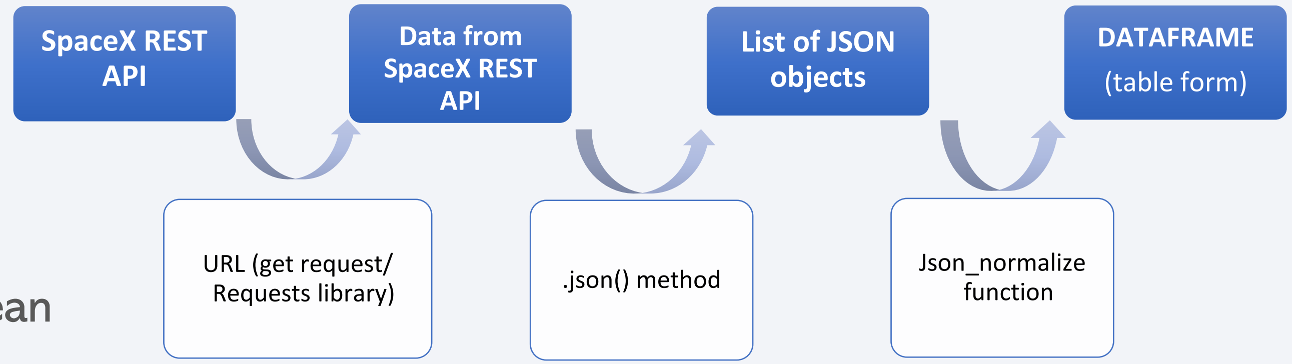
d) Perform interactive visual analytics using Folium and Plotly Dash

e) Perform predictive analysis using classification models

Data Collection – SpaceX API

IBM Developer SKILLS NETWORK

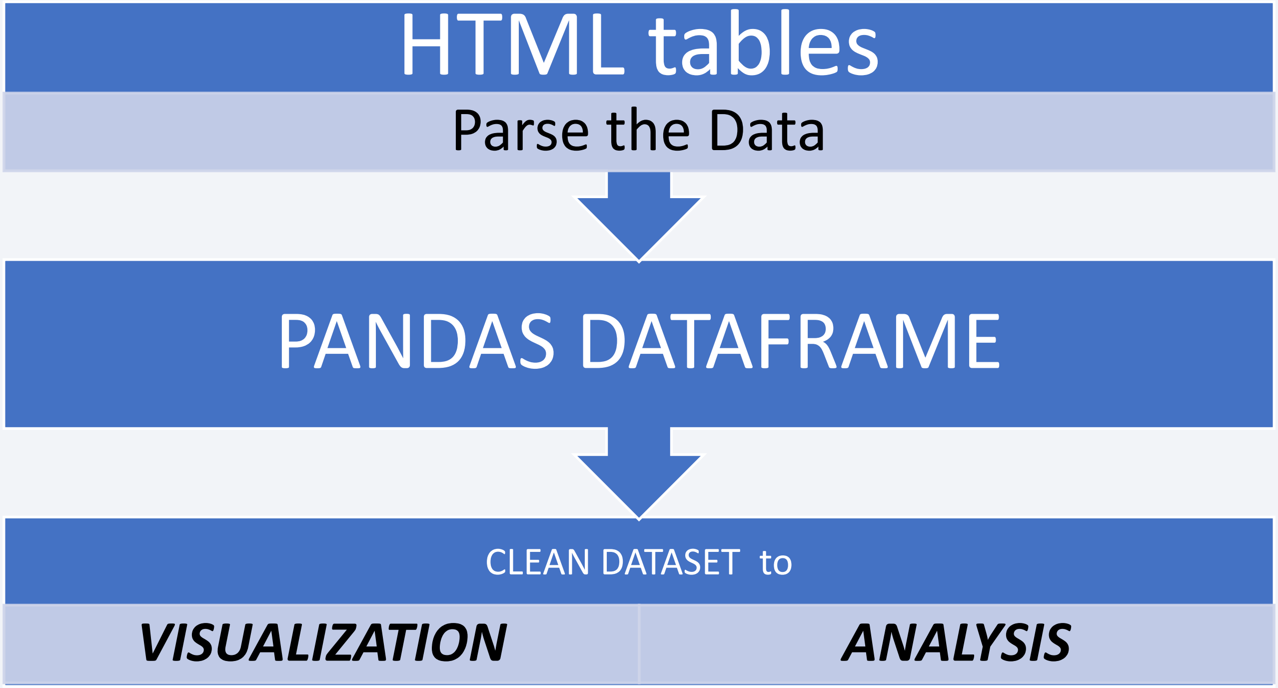
- **Objective:**
  - Obtain a flat table with the structured json data.
- **Additional process:**
  - Filtered the Dataframe to obtain Falcon 9 launches only.
  - Replace missing values with the mean (payload mass).



GitHub URL:
https://github.com/apnoguero/IBM-Data-Science-Capstone_SpaceX/blob/main/Lab%201_%20Collecting%20%20the%20data.ipynb

Data Collection - Scraping

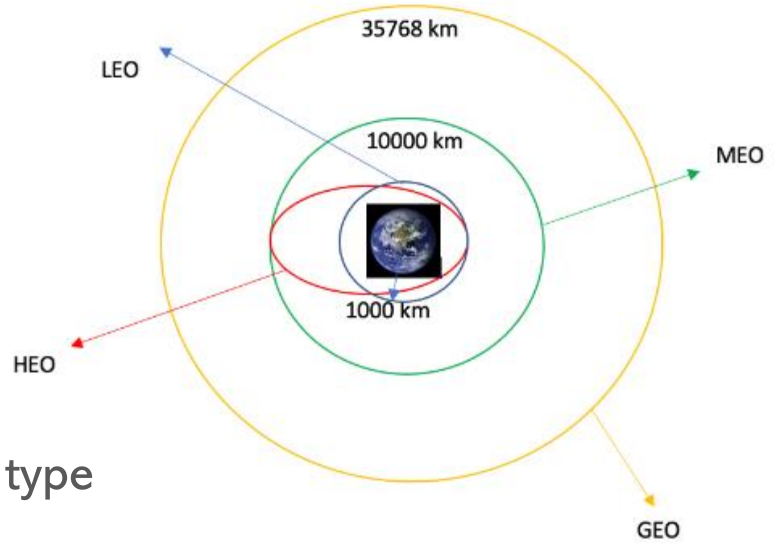
- **Process:**
- Use Python BeautifulSoup package to web scrape some HTML tables.
    - Extract columns and variables names from an ID Number.
    - Filter/sample the data to remove Falcon 1 launches.
    - Deal with NULLs.



Data Wrangling

- **Objectives:**
  - Exploratory Data Analysis
  - Determine Training Supervised Models Labels
- **Process:**
  - Obtain the number of launches on each site
  - Obtain the number and occurrence of each orbit
  - Obtain the number and occurrence of mission outcome per orbit type
  - Create a landing outcome label from “outcome” column

landing_class = 0 → bad_outcome (first stage did not land successfully)
landing_class = 1 → otherwise (first stage landed successfully)



GitHub URL:
https://github.com/apnoguero/IBM-Data-Science-Capstone_SpaceX/blob/main/Lab%202_%20Data%20wrangling.ipynb

EDA with Data Visualization

IBM Developer
SKILLS NETWORK

Objectives:

Exploratory Data Analysis

Preparing Data Feature Engineering

Process:

SpaceX dataset → Pandas Dataframe

Using visualization libraries

- Matplotlib
- Seaborn

Plot out:

FlightNumber ***vs*** PayloadMass

Flight Number ***vs*** Launch Site

Payload ***vs*** Launch Site

Success rate ***vs*** Orbit type

FlightNumber ***vs*** Orbit type

Payload ***vs*** Orbit type

Launch success yearly trend

GitHub URL:
https://github.com/apnoguero/IBM-Data-Science-Capstone_SpaceX/blob/main/Lab%204_%20Exploring%20and%20Preparing%20Data.ipynb

EDA with SQL

- **Objectives:**
  - Execute SQL queries to answer assignment questions:

1. Names of the unique launch sites	`SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL;`
2. 5 records where launch sites begin with 'CCA'	`SELECT LAUNCH_SITE FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;`
3. Total payload mass carried by boosters launched by NASA (CRS)	`SELECT SUM(PAYLOAD_MASS__KG_) AS payload_mass_kg FROM SPACEXTBL WHERE Customer = 'NASA (CRS)';`
4. Average payload mass carried by booster version F9 v1.1	`SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version LIKE 'F9 v1.0%';`
5. When the first successful landing outcome in ground pad was achieved.	`SELECT MIN(DATE) AS Date FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Success (ground pad)';`

EDA with SQL

IBM Developer SKILLS NETWORK

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

```
SELECT BOOSTER_VERSION
FROM SPACEXTBL
WHERE LANDING__OUTCOME = 'Success (drone ship)'
    AND 4000 < PAYLOAD_MASS__KG_ < 6000;
```

7. Total number of successful and failure mission outcomes.

```
SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL
FROM SPACEXTBL
GROUP BY MISSION_OUTCOME;
```

8. Names of the booster_versions which have carried the maximum payload mass.

```
SELECT DISTINCT BOOSTER_VERSION
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (
    SELECT MAX(PAYLOAD_MASS__KG_)
    FROM SPACEXTBL);
```

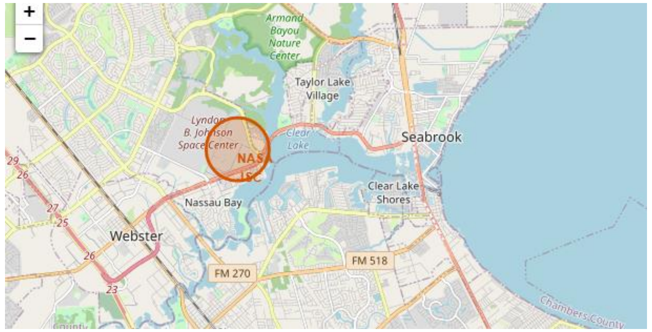
9. Failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015.

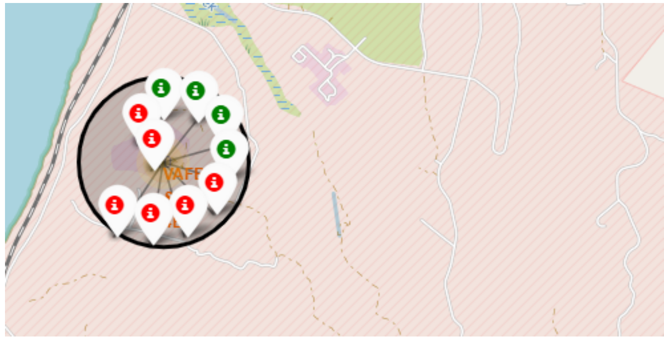
```
SELECT LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE
FROM SPACEXTBL
WHERE Landing__Outcome = 'Failure (drone ship)'
    AND YEAR(DATE) = 2015;
```

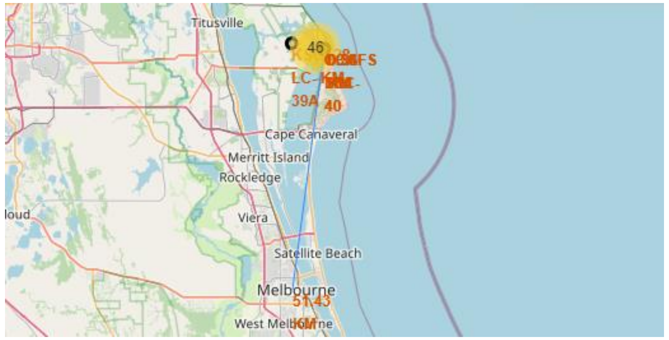
10. Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order.

```
SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS TOTAL
FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LANDING__OUTCOME
ORDER BY TOTAL DESC
```

GitHub URL:
https://github.com/apnoguero/IBM-Data-Science-Capstone_SpaceX/blob/main/Lab%203_%20SQL.ipynb







IBM Developer SKILLS NETWORK

Build an Interactive Map with Folium

Objectives:

TASK 1: Mark all launch sites on a map.

- Using latitude and longitude coordinates (spacex_launch_geo.csv).
- Create a folium Map Object (NASA Johnson Space Center as a start location).
- Add a highlights circle areas and another markers.

TASK 2: Mark the success/failed launches for each site on the map.

TASK 3: Calculate the distances between a launch site to its proximities.

- Railways
- Highways
- Coastlines
- Cities

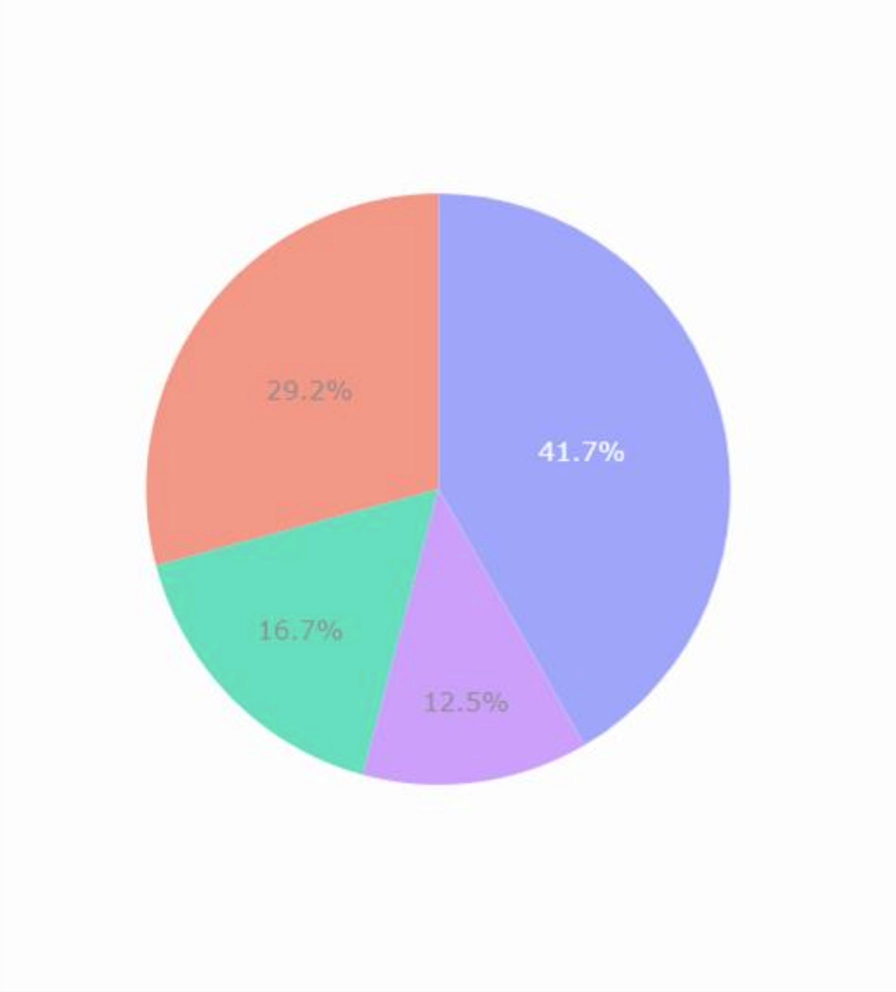
GitHub URL:
https://github.com/apnoguero/IBM-Data-Science-Capstone_SpaceX/blob/main/Lab5_%20Launch%20Sites%20Analysis%20with%20Folim.ipynb

Build a Dashboard with Plotly Dash

<u>Objectives:</u>

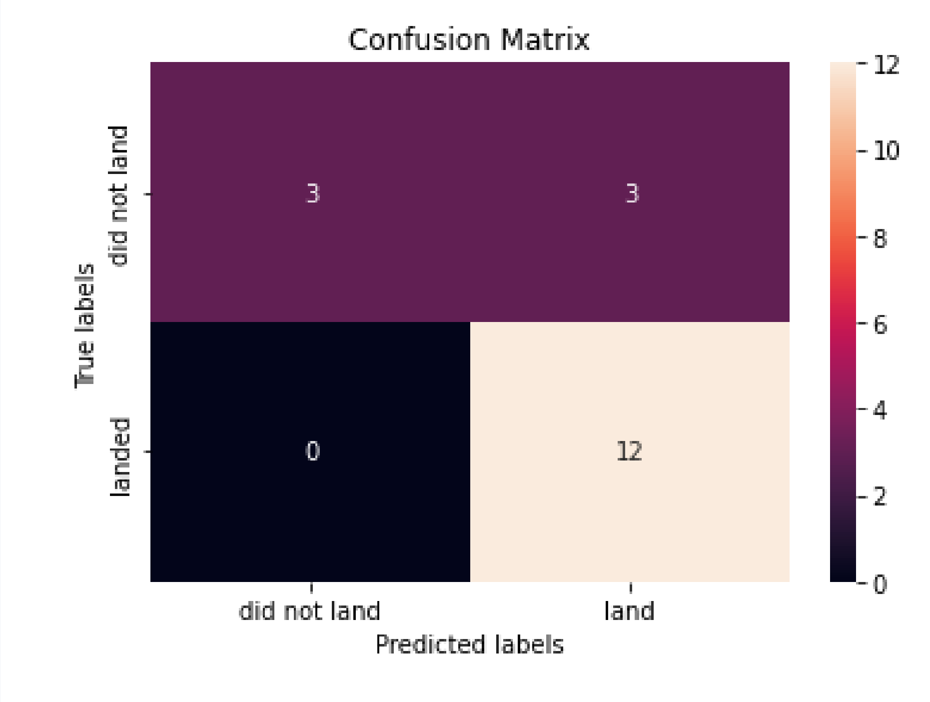
Explore and manipulate Data in an interactive and real-time way.

- ***Pie chart showing success rate***
- ***Scatter chart showing PayloadMass vs landing_outcome***
- ***Drop-down menu to choose individual or all sites launch sites***



Predictive Analysis

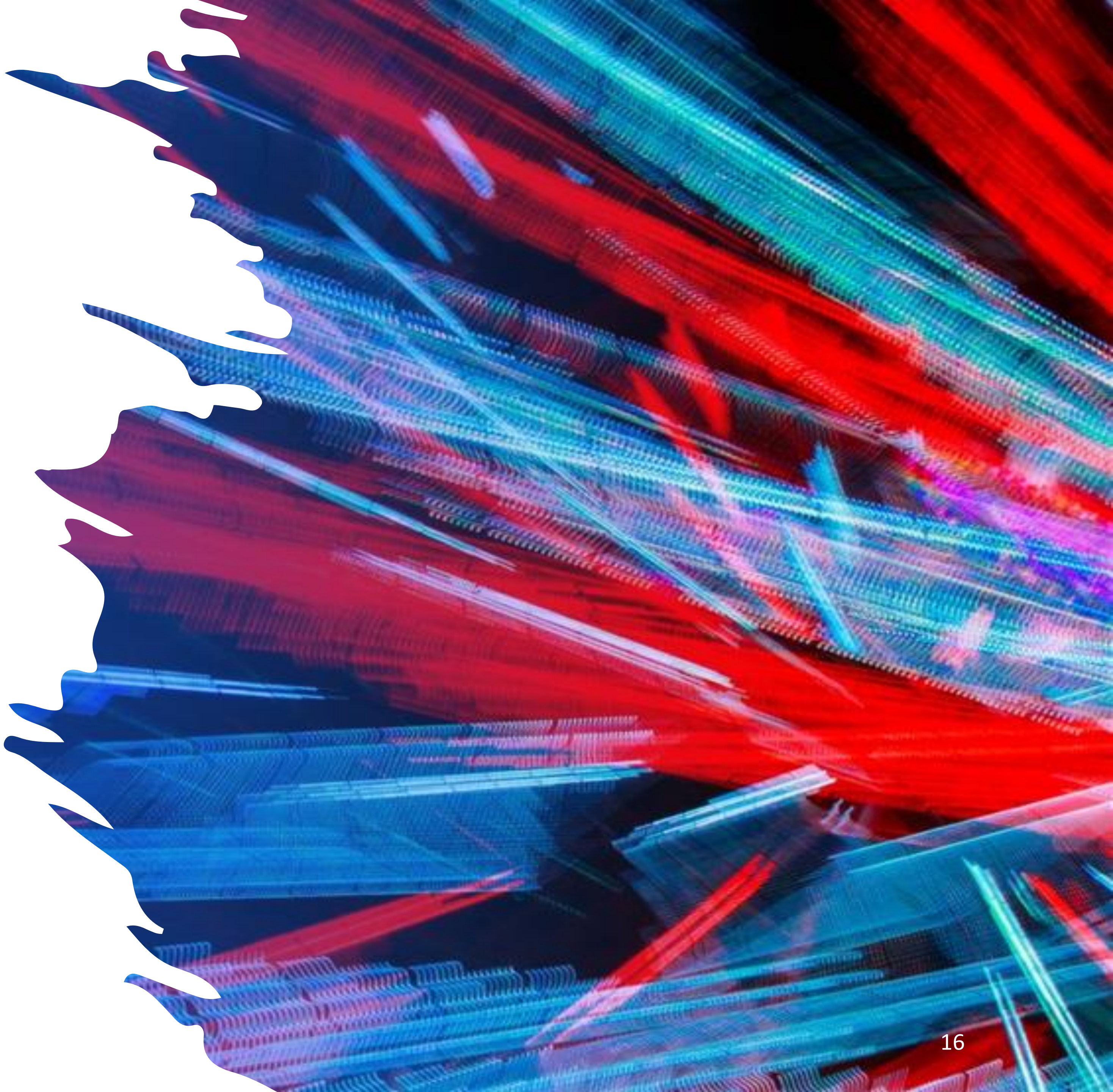
- **Objectives:**
    - Preform exploratory Data Analysis and determine Training Labels.
- **Process:**
    - Import libraries
    - Loaded the Dataframe created during data collection
    - Create a column for the class
    - Standardize the data
    - Split into training data and test data
    - Fit the training data to different model types:
        - LR
        - SVM
        - Decision Tree Classifier
        - K-nearest Neighbors Classifier
    - Select the best one hyperparameters to each model,
    - Evaluate the accuracy of each one.



GitHub URL:
https://github.com/apnoguero/IBM-Data-Science-Capstone_SpaceX/blob/main/Lab7_%20Machine%20Learning%20Prediction.ipynb

Results

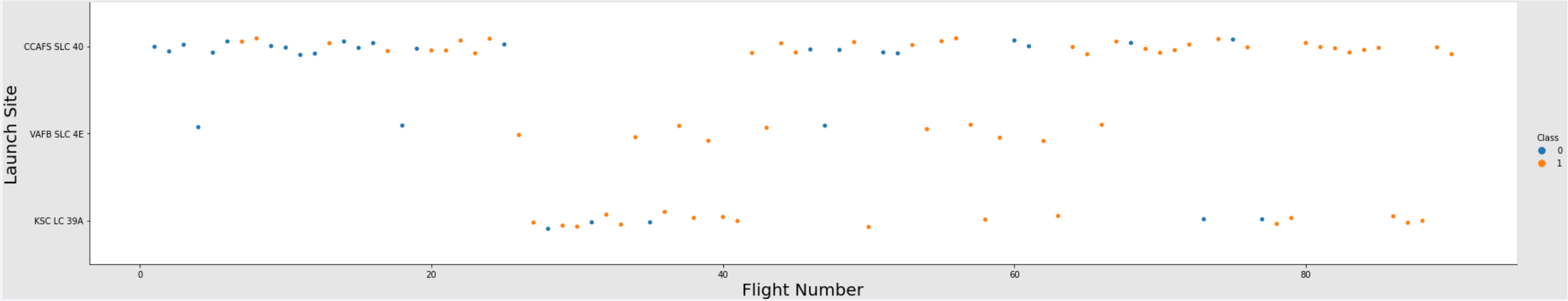
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results



Section 2

Insights drawn from EDA

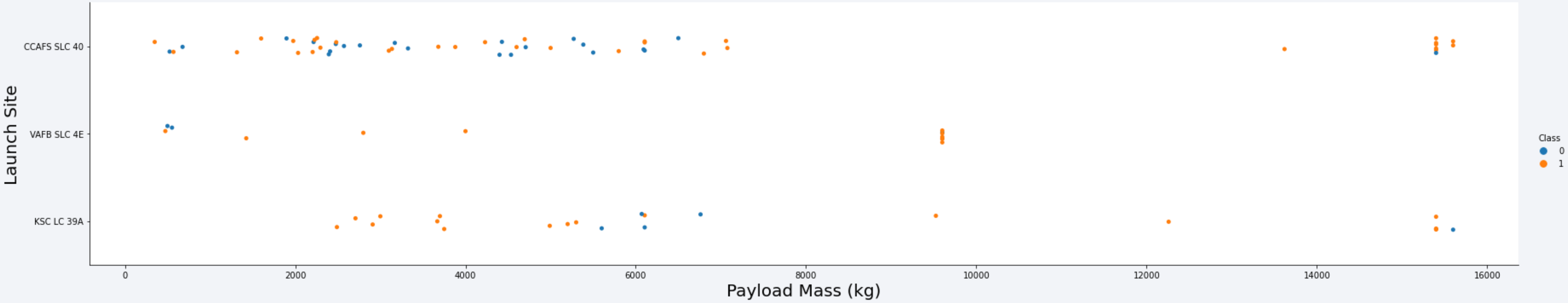
Flight Number vs. Launch Site



Observations:

- Launch Site *CCAFS SLC 40* & Flight Number *< 25*, appears to have most landing failures (colour blue).
- Launch Site *CCAFS SLC 40* report more launches than the others sites.
- As the flight number increases, the first stage is more likely to land successfully (colour orange).

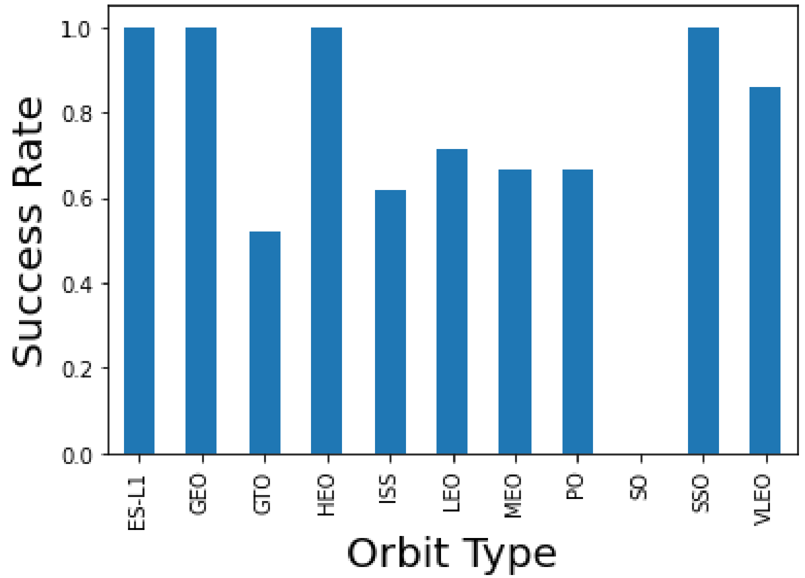
Payload vs. Launch Site



Observations:

- Launch Site *CCAFS SLC 40* appears to concentrate launches when the *Payload Mass is less than 7500 kg.*
- The rockets with heavy payload mass (greater than 10000 kg) are launched from CCAFS SLC 40 and KSC LC 39A sites.

Success Rate vs. Orbit Type

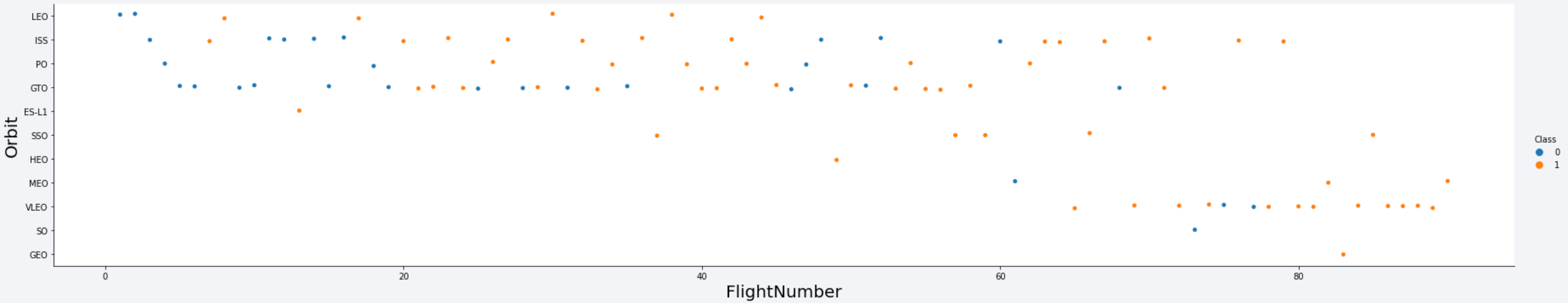


Analyze the ploted bar chart try to find which orbits have high sucess rate. High success rate: ES-L1, GEO, HEO and SSO. Lowest success rate: SO and GTO.

Observations:

- All orbit types except SO have reported successfully 1st stage landings.
- The highest success rates were reported for the ES-L1, GEO, HEO, and SSO orbits.

Flight Number vs. Orbit Type



Observations:

- Visually, no significant relationship is reported between the *Flight Number vs Orbit type*, although in the LEO orbit, the success appears related to the number of flights (higher number reports higher success).

Payload vs. Orbit Type

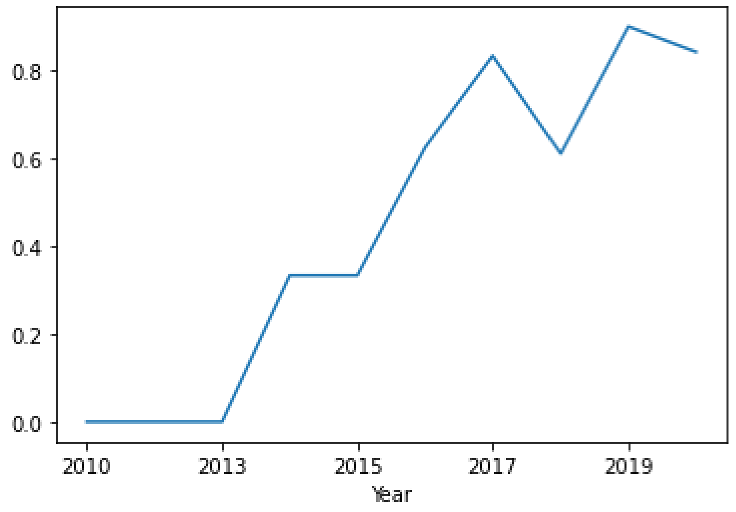
Observations:

- Heavy payloads (higher than 9000) report positive landing rates for *LEO, ISS and PO* orbit.
- For GTO, we cannot distinguish a conclusive result of succession rates.
- We can observe a correlation between *ISS and Payload* when Payload values are around 2000.

Launch Success Yearly Trend

```
In [10]: # Plot a line chart with x axis to be the extracted year and y axis to be the su
         ccess rate
         df2 = df.copy()
         df2['Year'] = year
         df2.groupby('Year')['Class'].mean().plot()

Out[10]: <AxesSubplot:xlabel='Year'>
```



Observations:

- Visually, we can observe that launch success rate is increasing significantly since 2013, although in 2018 it decrease a bit.

All Launch Site Names

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

```
%%sql
SELECT DISTINCT LAUNCH_SITE
FROM SPACEXTBL;
```

```
 * ibm_db_sa://jyl47663:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1
od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

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

Launch Site Names Begin with 'CCA'

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

```
%%sql
SELECT LAUNCH_SITE
FROM SPACEXTBL
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5;
```

```
 * ibm_db_sa://jyl47663:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

launch_site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

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

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS payload_mass_kg
FROM SPACEXTBL
WHERE Customer = 'NASA (CRS)';
```

```
 * ibm_db_sa://jyl47663:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

payload_mass_kg
45596

Average Payload Mass by F9 v1.1

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

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_)
FROM SPACEXTBL
WHERE Booster_Version LIKE 'F9 v1.0%';
```

```
 * ibm_db_sa://jyl47663:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

1
340

First Successful Ground Landing Date

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

Hint:Use min function

```
%%sql
SELECT MIN(DATE) AS Date
FROM SPACEXTBL
WHERE LANDING__OUTCOME = 'Success (ground pad)';
```

```
 * ibm_db_sa://jyl47663:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1
od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

DATE
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

```
%%sql
SELECT BOOSTER_VERSION
FROM SPACEXTBL
WHERE LANDING__OUTCOME = 'Success (drone ship)'
    AND 4000 < PAYLOAD_MASS__KG_ < 6000;
```

```
 * ibm_db_sa://jyl47663:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

booster_version
F9 FT B1021.1
F9 FT B1023.1
F9 FT B1029.2
F9 FT B1038.1
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
%%sql
SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL
FROM SPACEXTBL
GROUP BY MISSION_OUTCOME;
```

```
 * ibm_db_sa://jyl47663:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1
od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

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

Boosters Carried Maximum Payload

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

```
%%sql
SELECT DISTINCT BOOSTER_VERSION
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (
    SELECT MAX(PAYLOAD_MASS__KG_)
    FROM SPACEXTBL);
```

```
 * ibm_db_sa://jyl47663:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

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

2015 Launch Records

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

```
%%sql
SELECT LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE
FROM SPACEXTBL
WHERE Landing__Outcome = 'Failure (drone ship)'
    AND YEAR(DATE) = 2015;
```

```
 * ibm_db_sa://jyl47663:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

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

```
%%sql
SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS TOTAL
FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LANDING__OUTCOME
ORDER BY TOTAL DESC
```

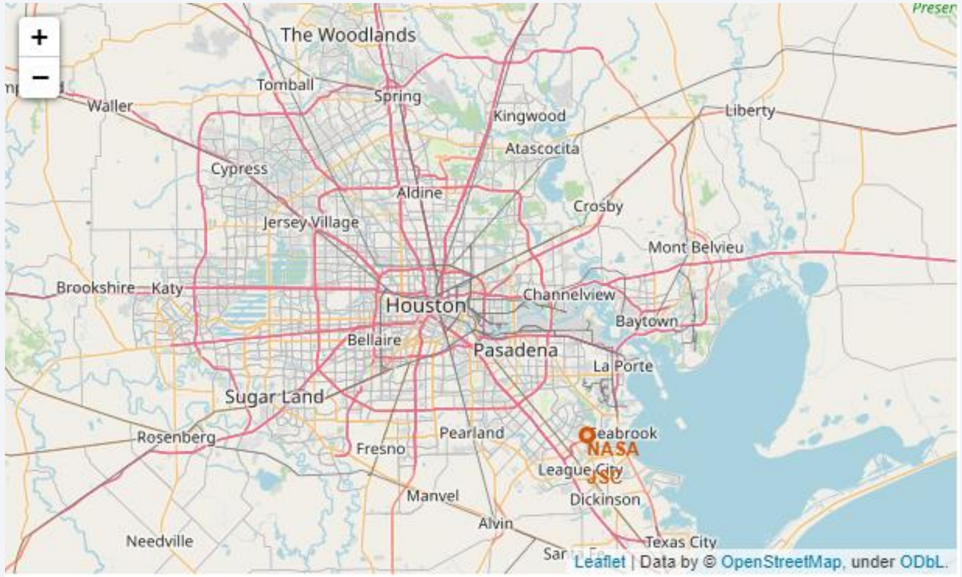
```
 * ibm_db_sa://jyl47663:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

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

Section 3

Launch Sites Proximities Analysis

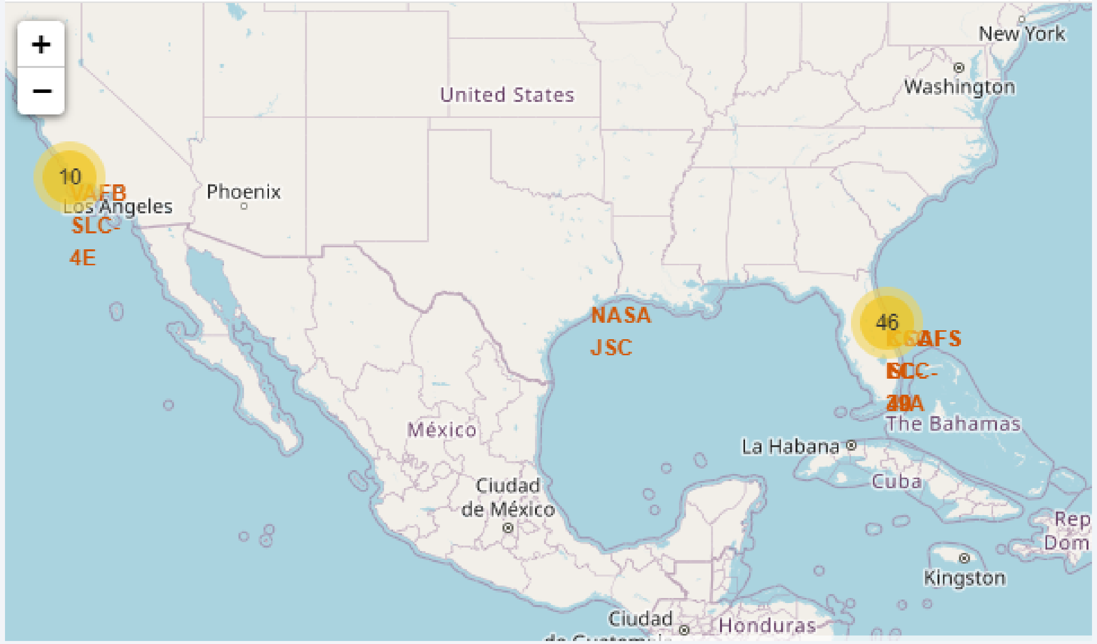
All launch sites on a map

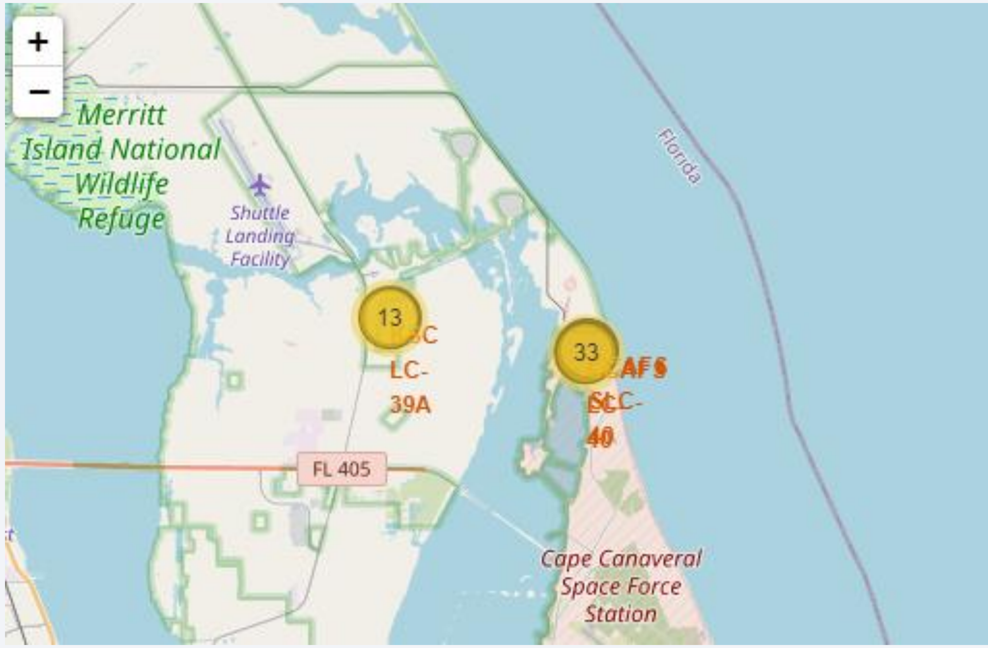


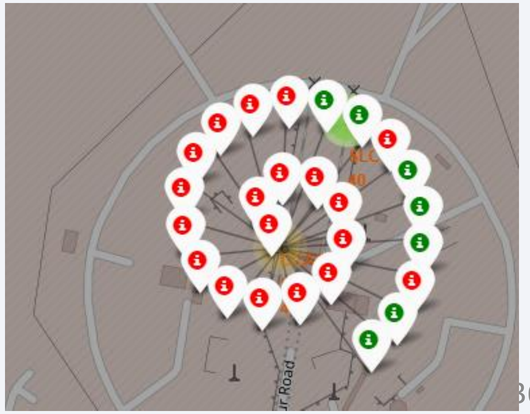
NASA Johnson Space Center's

All launch sites are close to the Equator lane and to the cost.
It’s congruent with the axis of rotation of the earth (less fuel) and with security reasons (areas with less population).

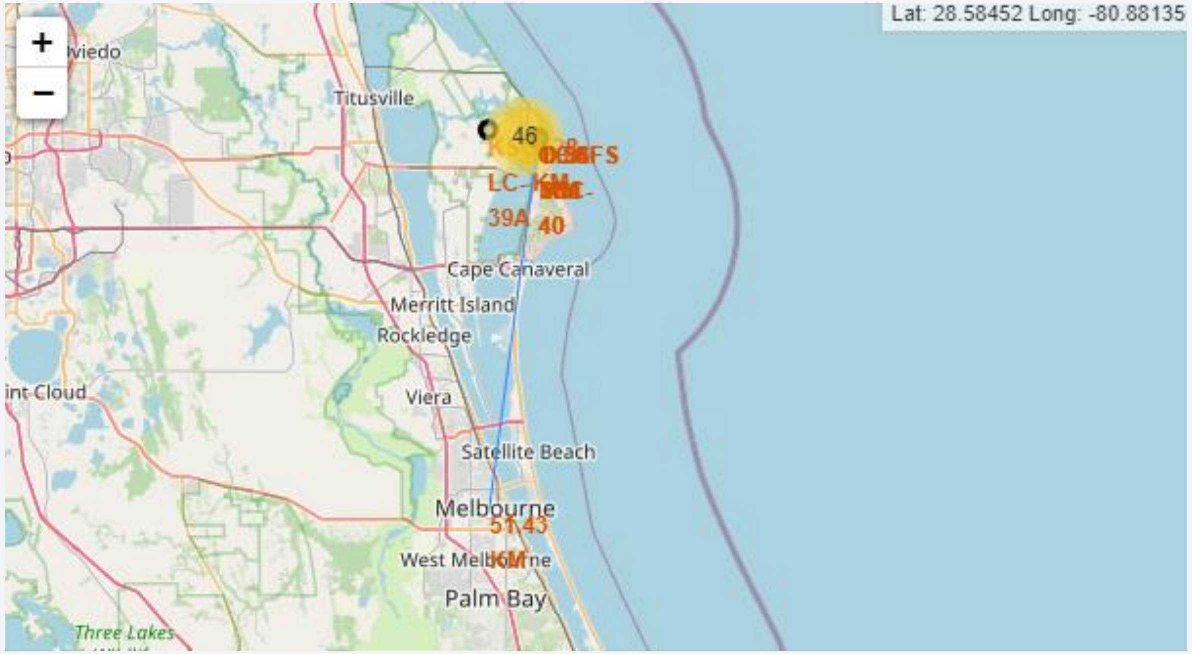
Success/Failed launches for each site







Distances between a launch site to its proximities



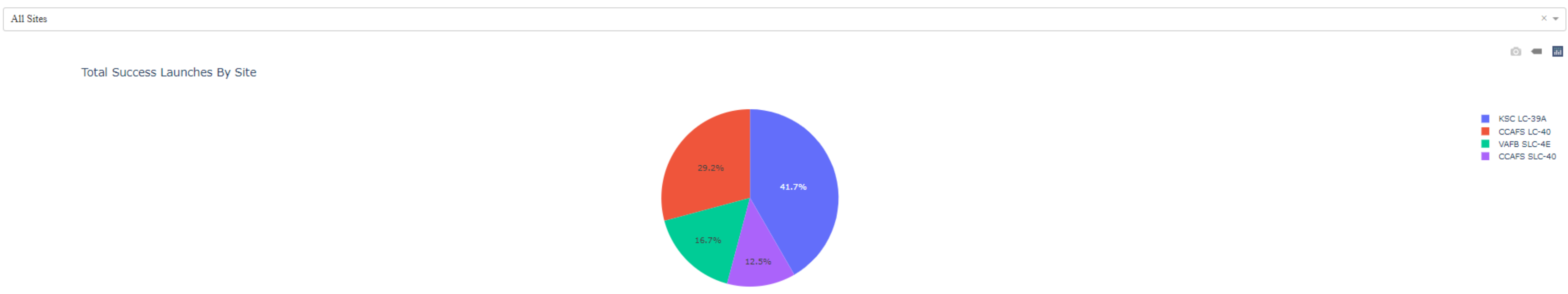
Observations:

- Launch Sites are close to railways and highways, which allows easy transportation.
- They are not in close to proximity to cities, for security reasons (less population less probability of damage).

Section 4

Build a Dashboard with Plotly Dash

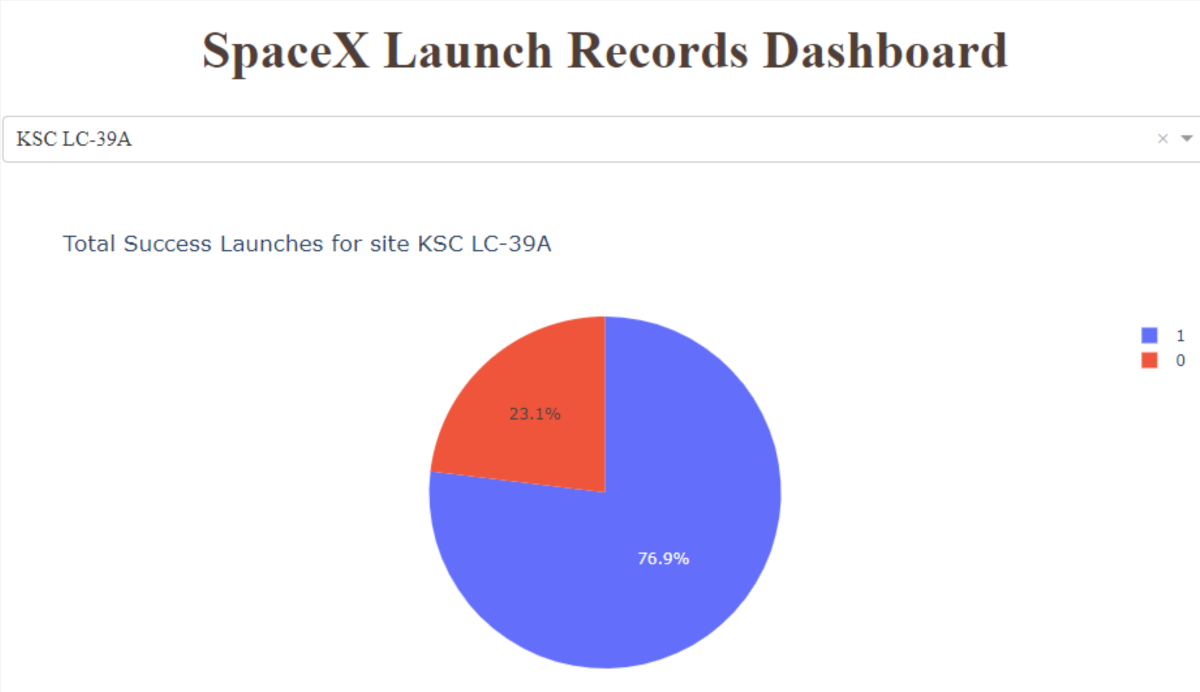
Success launches by all sites



Observations:

- ***KSC LC-39A*** presents the most successful launches from all the sites (41,7%), followed by CCAFS LC-40.
- Lowest ratio of successful launches reported by CCAFS SLC-40.

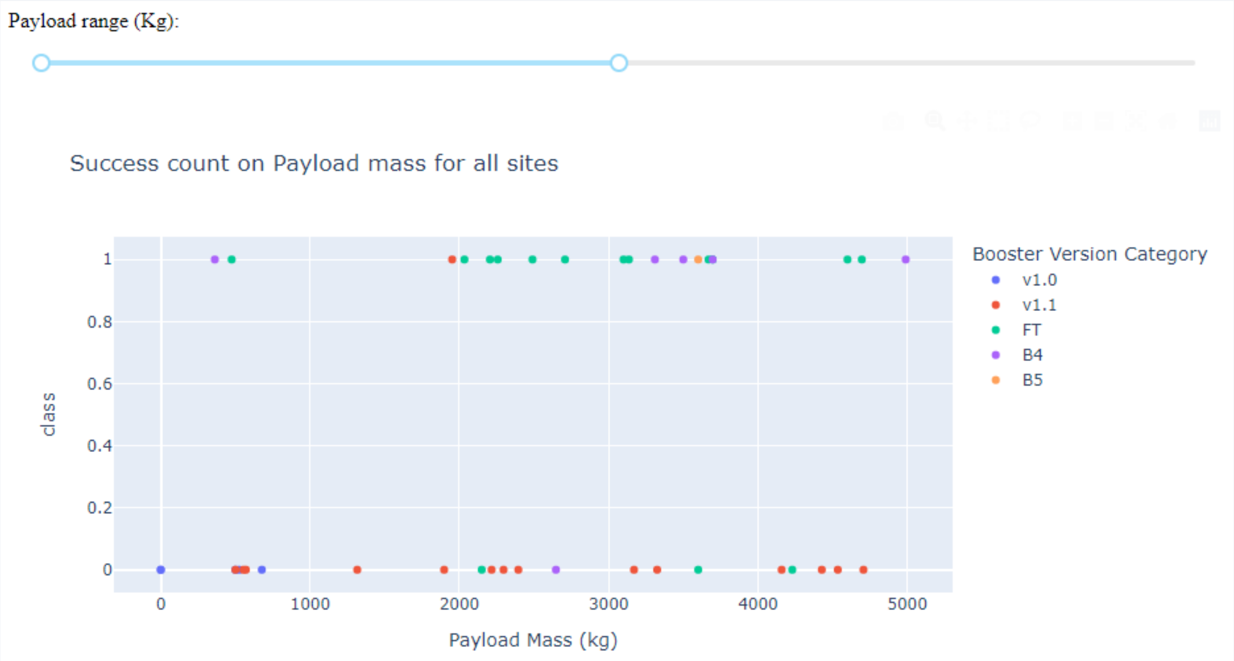
Highest launch success ratio

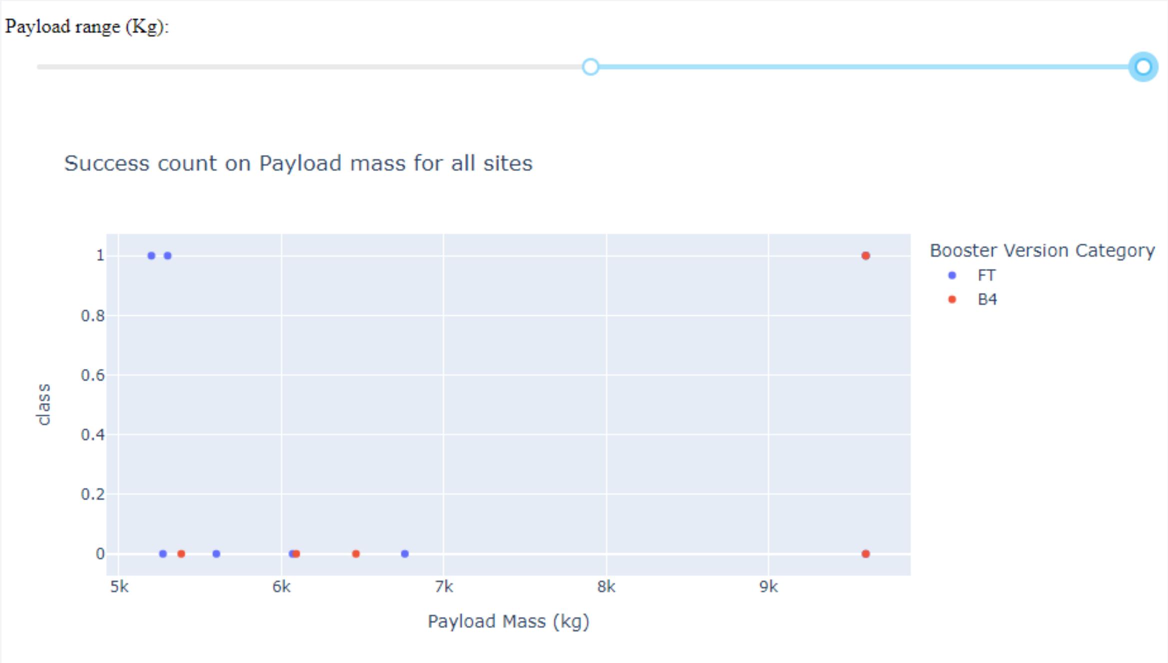


Observations:

- ***KSC LC-39A*** reported a 76,9% success and 23,1% failure rate.

Payload vs. Launch Outcome





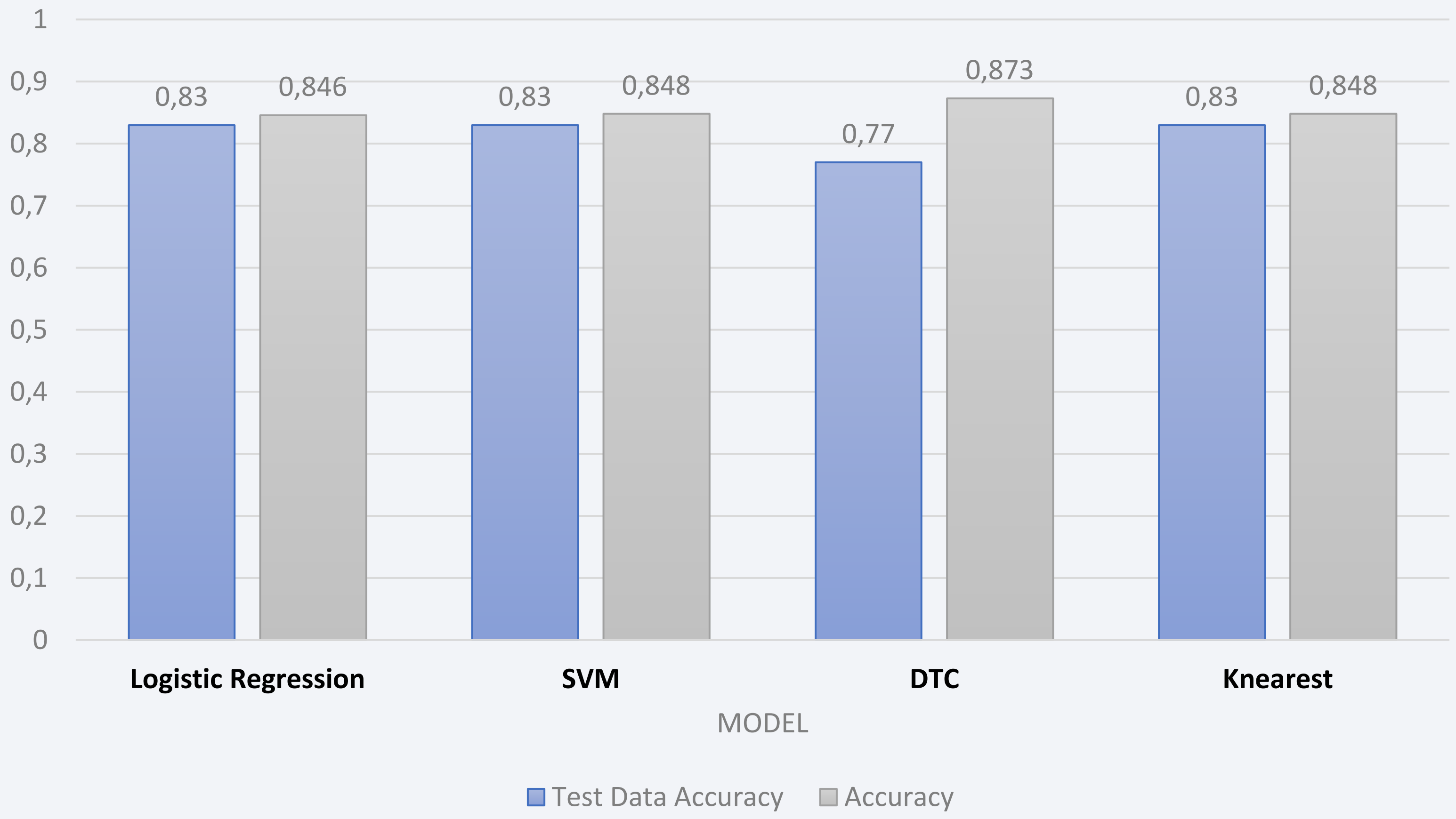
Observations:

- Payload < 5000 kg → Highest booster landing success rate
- Payload > 5000 kg → Lowest booster landing success rate

Section 5

Predictive Analysis (Classification)

Classification Accuracy

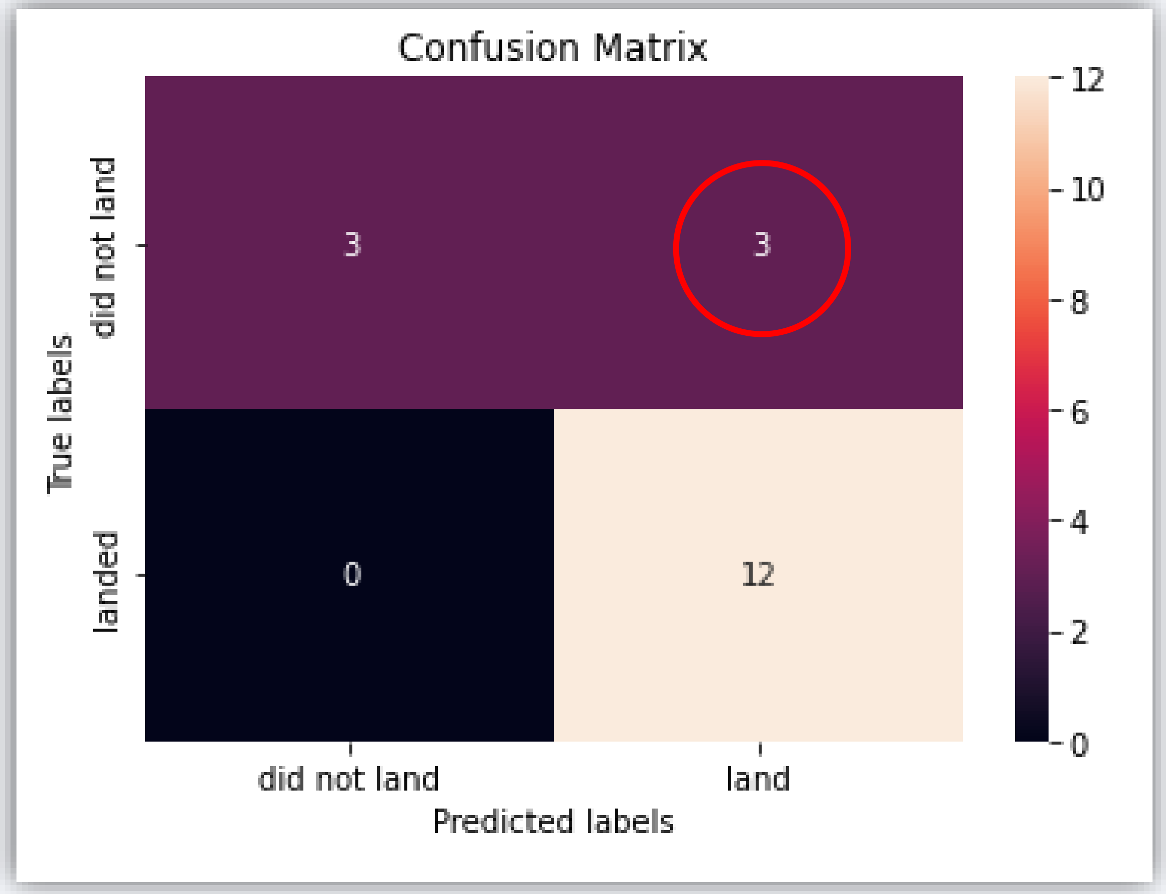


Observations:

- 3 models report the same accuracy score on the test set (0,83).

- The *highest test data* accuracy is reported by the **Decision Tree (0,873)**, although obtain the lowest accuracy on the test set.

Confusion Matrix



Observations:

- All the models report the same confusion matrix.

- We can detect the problem of **False Positives cases** (the models incorrectly predict the positive class in 3 occasions of the 18 samples).

Conclusions

- The model from this report can predict with an 83,3% of accuracy when SpaceX will successfully land the 1st stage booster.
- This **results can help SpaceX to reduce costs**, sacrificing $15+ million predicting and reusing the 1st stage booster.
- Technicities:
    - As the flight number increases, the first stage is more likely to land successfully.
    - The highest success rates were reported for the ES-L1, GEO, HEO, and SSO orbits.
    - Heavy payloads (higher than 9000) report positive landing rates for LEO, ISS and PO orbit.
    - Launch success rate is increasing significantly since 2013.
    - Launch sites are close to the Equator lane and to the cost (logistic and security issues?).
    - KSC LC-39A presents the most successful launches from all the sites (41,7%).
    - Highest booster landing success rate obtained when Payload < 5000 kg.

APPENDIX

- Notebooks to recreate dataset, analysis and models:

https://github.com/apnoguero/IBM-Data-Science-Capstone_SpaceX/blob/main/Lab%201_%20Collecting%20%20the%20data.ipynb

https://github.com/apnoguero/IBM-Data-Science-Capstone_SpaceX/blob/main/Lab%202_%20Data%20wrangling.ipynb

https://github.com/apnoguero/IBM-Data-Science-Capstone_SpaceX/blob/main/Lab%204_%20Exploring%20and%20Preparing%20Data.ipynb

https://github.com/apnoguero/IBM-Data-Science-Capstone_SpaceX/blob/main/Lab%203_%20SQL.ipynb

https://github.com/apnoguero/IBM-Data-Science-Capstone_SpaceX/blob/main/Lab5_%20Launch%20Sites%20Analysis%20with%20Folim.ipynb

https://github.com/apnoguero/IBM-Data-Science-Capstone_SpaceX/blob/main/Lab7_%20Machine%20Learning%20Prediction.ipynb

Thank you!