IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

03.JUNE.2024

Author: Moe Dastranj (Parkway Production)
Position: Data Science
Project: Final Presentation

Contact info: **Linkedin**
GitHub

Outline

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

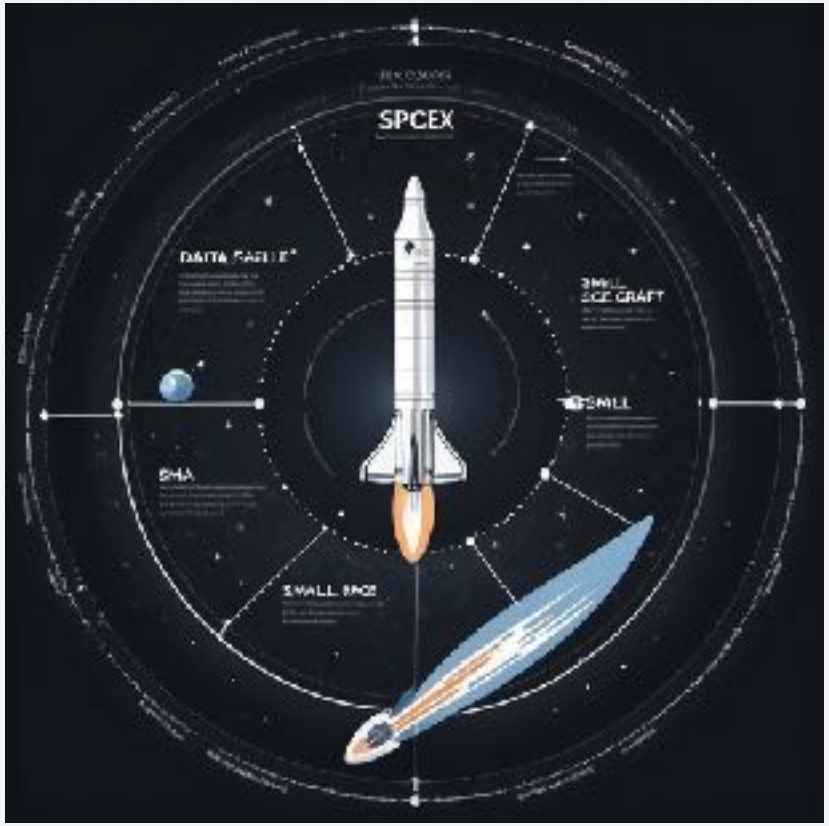
Executive Summary

- **Summary of methodologies**
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Folium
  - Machine Learning Prediction
- **Summary of all results**
  - - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive result

Introduction

- **Project background and context**

Data science is critical in today's economic and intelligence industries, enabling decision-making and innovation. Despite its importance, many individuals are still uninformed of what data is and how it is used. This initiative attempts to demystify Data Science by walking people through its techniques and showing how data-driven approaches can turn challenges into practical solutions. By delving into these principles, we will discover the critical role of data in moulding our modern environment.



In this research, we will investigate the particular aspects that impact whether a rocket lands successfully. We will evaluate the chance of a successful landing by studying how these components interact. In addition, we will determine the best operating parameters required to achieve consistent success in SpaceX's landing program.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - A total of two methods/sources where used for Space X Data Collection
    - Space X REST API (https://api.spacexdata.com/v4/rockets/)
    - Web-Scarping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)

- Perform data wrangling
  - Data wrangling, which entails cleaning, organizing, and transforming raw data into a usable format for analysis, was carried out by summarizing and evaluating characteristics, assigning a landing result label based on outcomes, transforming categorical data using One Hot Encoding for machine learning algorithms, and removing any empty or unnecessary information.

- Perform exploratory data analysis (EDA) using visualization and SQL
  - Identifying all distinct landing sites, Payloads, Boosters, success and failure mission outcomes.

- Perform interactive visual analytics using Folium and Plotly Dash
  - Created a map showing the success/failure rates of each site's launches as well as the distances between launch sites and their proximity. Identified some geographical trends regarding launch places. A Dashboard was designed to graphically present data and conclusions in response to data requests.

- Perform predictive analysis using classification models
  - The data acquired up to this point was normalized, separated into training and test data sets, and assessed by four distinct classification models, with the accuracy of each model tested using different parameter combinations.

Data Collection

Datasets were collected from Space X API and from Wikipedia, using web scraping technics

API

- Gathered Space X past launch data via public API
- Retrieved and processed data with GET request
- Ensured the data included Falcon 9 launches only.
- Filled in missing payload weights from secret missions with average values

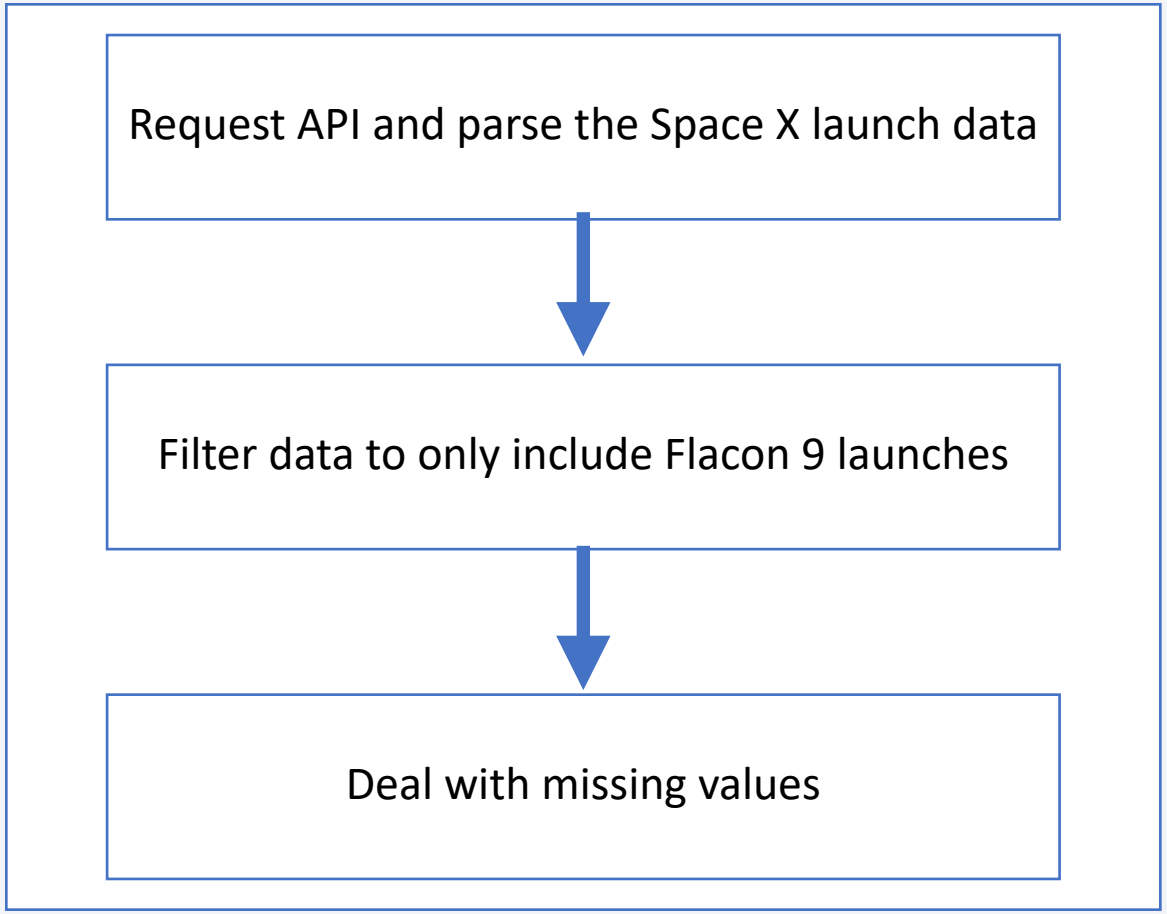
Web Scraping

- Request past Falcon 9 and Falcon Heavy launch data from Wikipedia's relevant page
- Accessed theFalcon 9 Launch page via its direct link
- Extracted all the column names from the HTML table
- Parsed and transformed the table into a Pandas data frame suitable for analysis

Data Collection – SpaceX API

GitHub link: >> GitHub <<

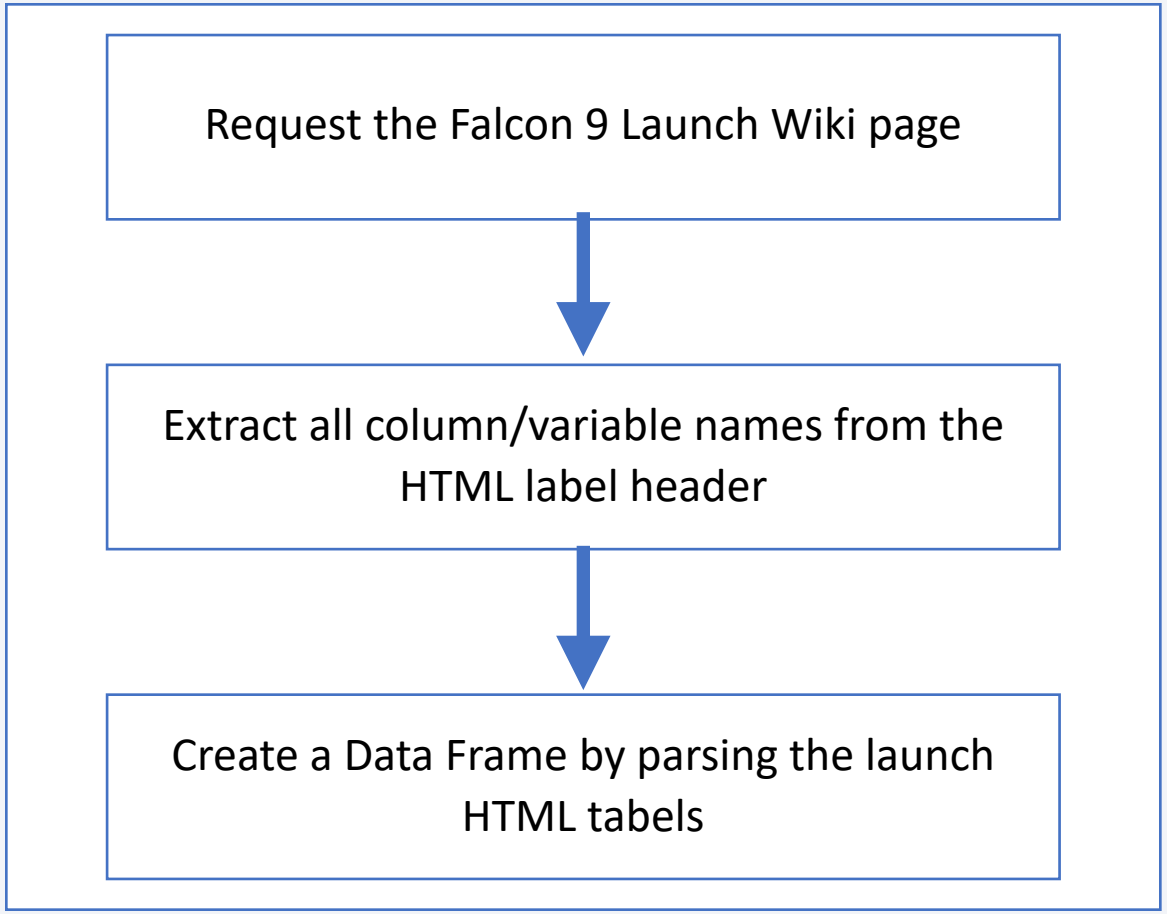
- Space X provides a public API where data is being provided for use
- See FlowChart for API usage for persistent data format
- GitHub URL completed code cell and outcome cell



Data Collection - Scraping

GitHub link: >> GitHub <<

- Data from Space X launches can also be obtained from Wikipedia;
- See FlowChart for WebScraping usage for persistent data format



Data Wrangling

GitHub link: >> GitHub <<

EDA → Summarization → Creation of Landing Outcome Label

- Dataset were processed with EDA (Exploratory Data Analysis) initially.
- Then the summary launches per site, occurrences of each orbit and occurrences of mission outcomes per orbit type were calculated.
- Final stage the label for landing outcome was created from Outcome column.

```
[5] # Apply value_counts() on column LaunchSite
    df['LaunchSite'].value_counts()

LaunchSite
CCAFS SLC 40    55
KSC LC 39A      22
VAFB SLC 4E     13
Name: count, dtype: int64
```

```
[ ] df.head(5)
```

	FlightNumber	Date	BoosterVersion	PayloadM
0	1	2010-06-04	Falcon 9	6104.959
1	2	2012-05-22	Falcon 9	525.000
2	3	2013-03-01	Falcon 9	677.000

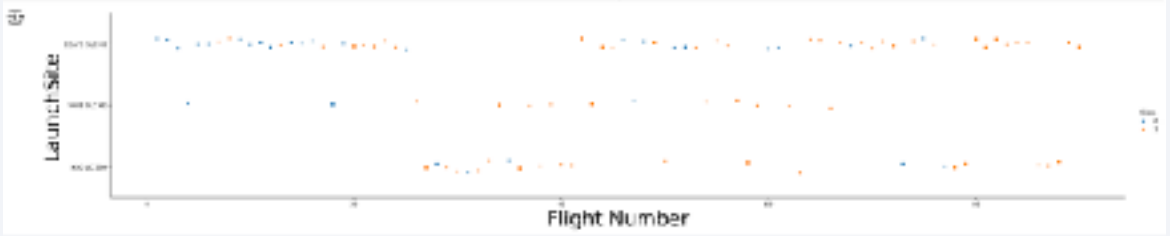
```
# Apply value_counts on Orbit column
df['Orbit'].value_counts()

Orbit
GTO     27
ISS     21
VLEO    14
PO       9
LEO      7
SSO      5
MEO      3
ES-L1    1
HEO      1
SO       1
GEO      1
Name: count, dtype: int64
```

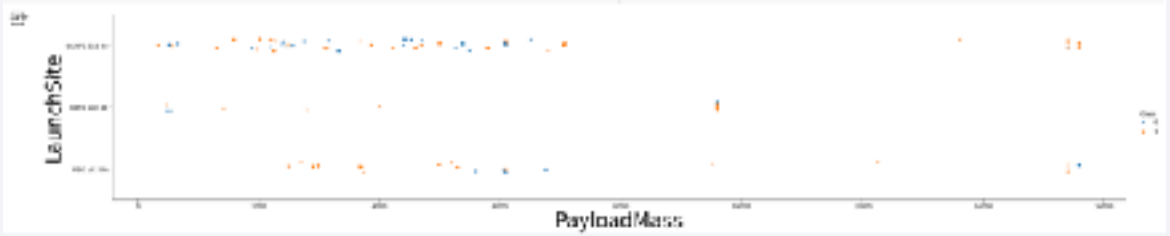
```
# landing_class = 0 if bad_outcome
# landing_class = 1 otherwise
landing_class = []
for outcome in df['Outcome']:
    if outcome in bad_outcomes:
        landing_class.append(0)
    else:
        landing_class.append(1)

landing_class

[0,
 0,
```

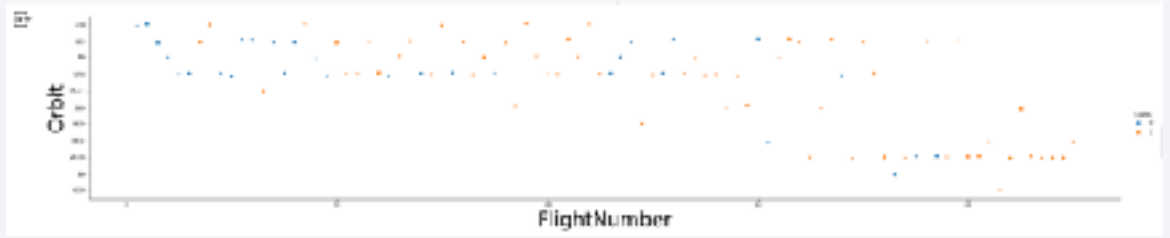
EDA with Data Visualization

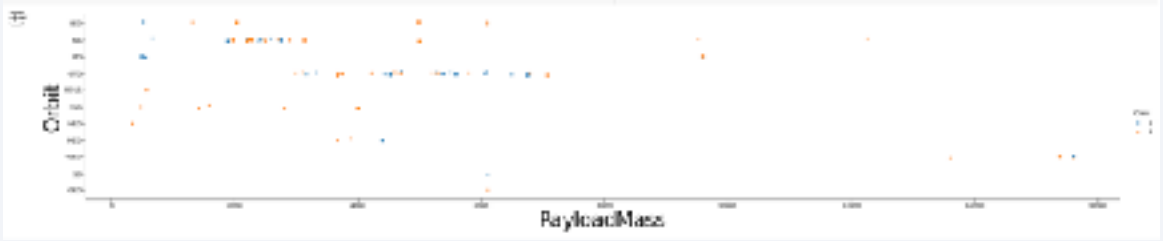
GitHub link: [>> GitHub <<]

Scatter point charts used to help visualize the relations between Flight Number & Launch Site, PayLoad & Launch Site, Flight Number & Orbit type.

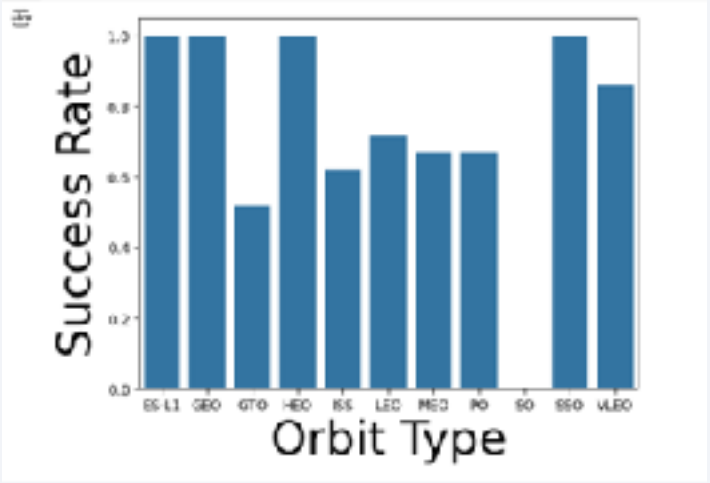


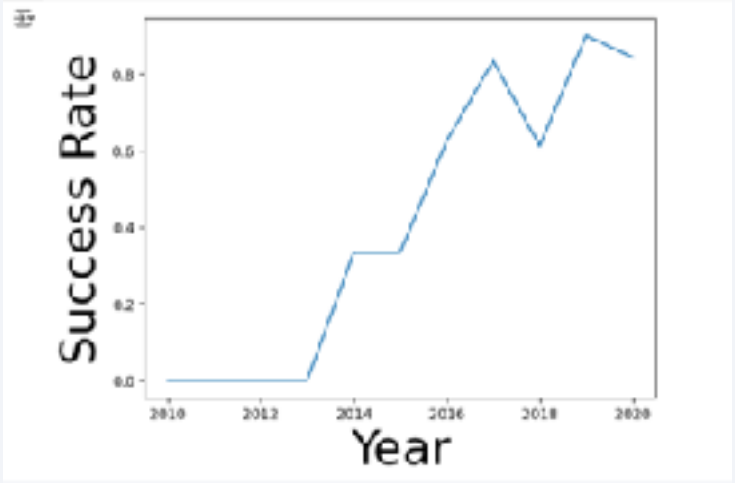






Visualize success rate of each orbit type with the Bar Chart. Line Chart for success year trend





EDA with SQL

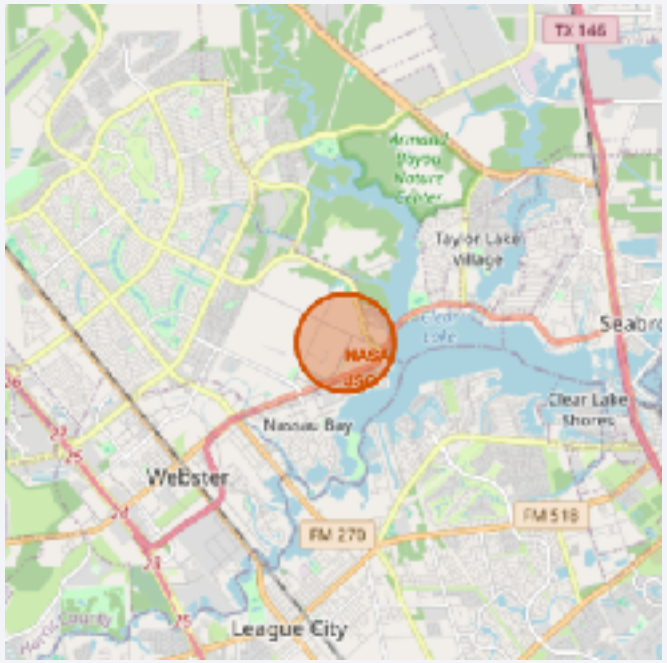
GitHub link: >> GitHub <<

1. Display the names of the unique launch sites in the space mission.
2. Top 5 records where launch site name begins with ‘CCA’
3. Display the total payload mass carried by the booster 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 achieved
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 using a subquery
9. List the records which will display the moony names, failure landing outcomes in drone ship, booster versions, launch site for the months 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.

Build an Interactive Map with Folium

GitHub link: [>> GitHub <<]

A circle marker was created to show NASA Johnson Space Center's coordinate

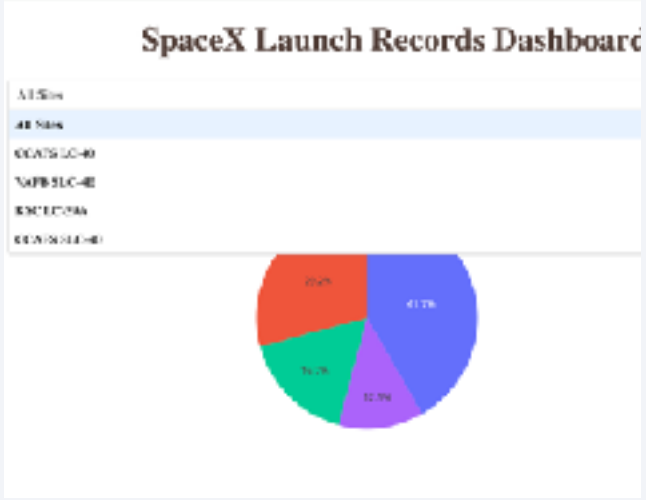


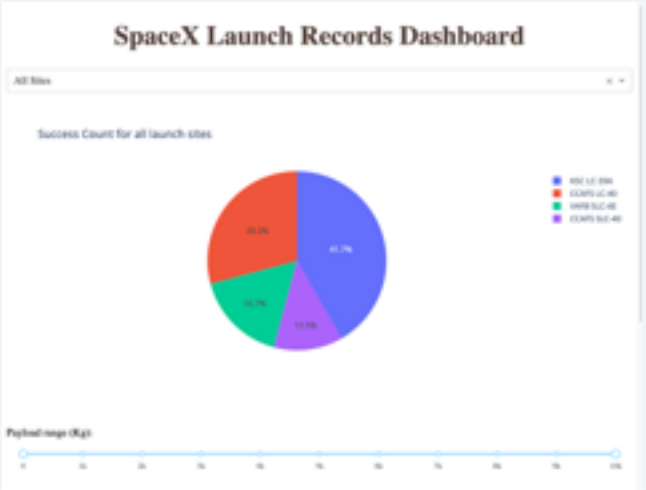
Distance marker was created to show distance between a launch site to its proximities

Build a Dashboard with Plotly Dash

GitHub link: [>> GitHub <<]

Dropdown option of pie chart created to show the success launches of all/each site



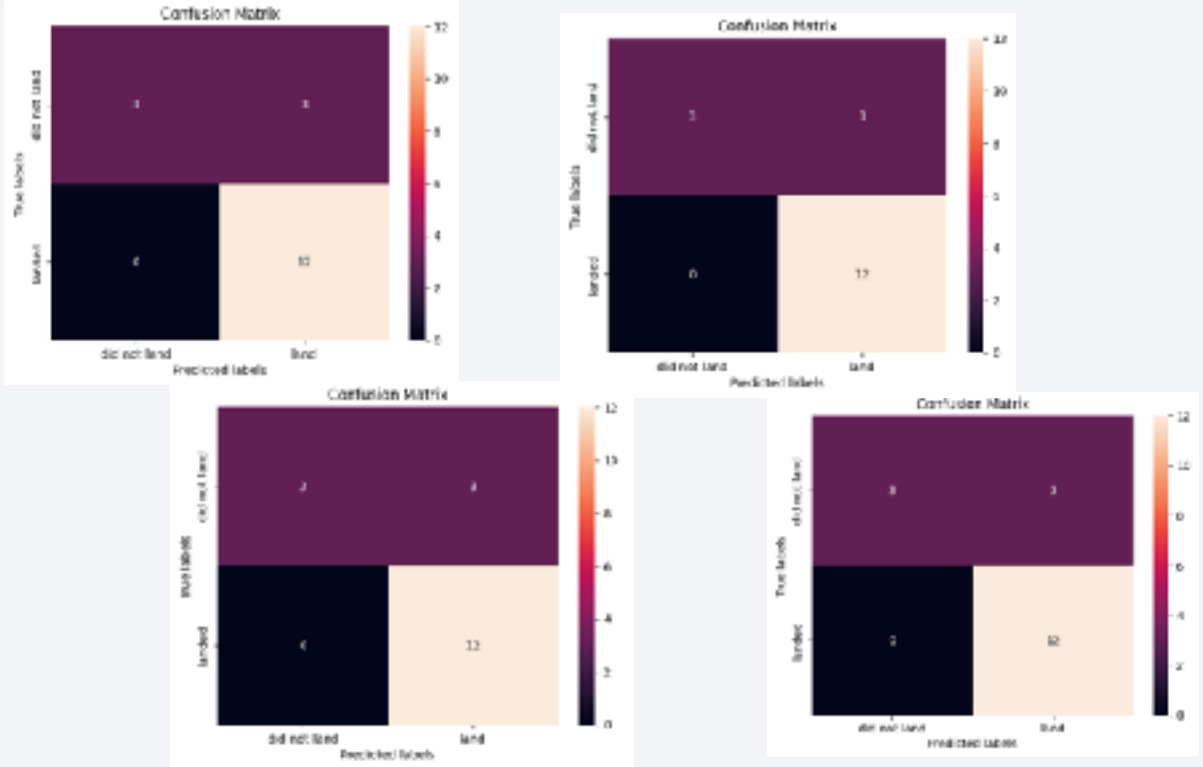


Payload range slider with Scatter plot show the success launches of all / each site by payload

Predictive Analysis (Classification)

GitHub link: [>> GitHub <<]

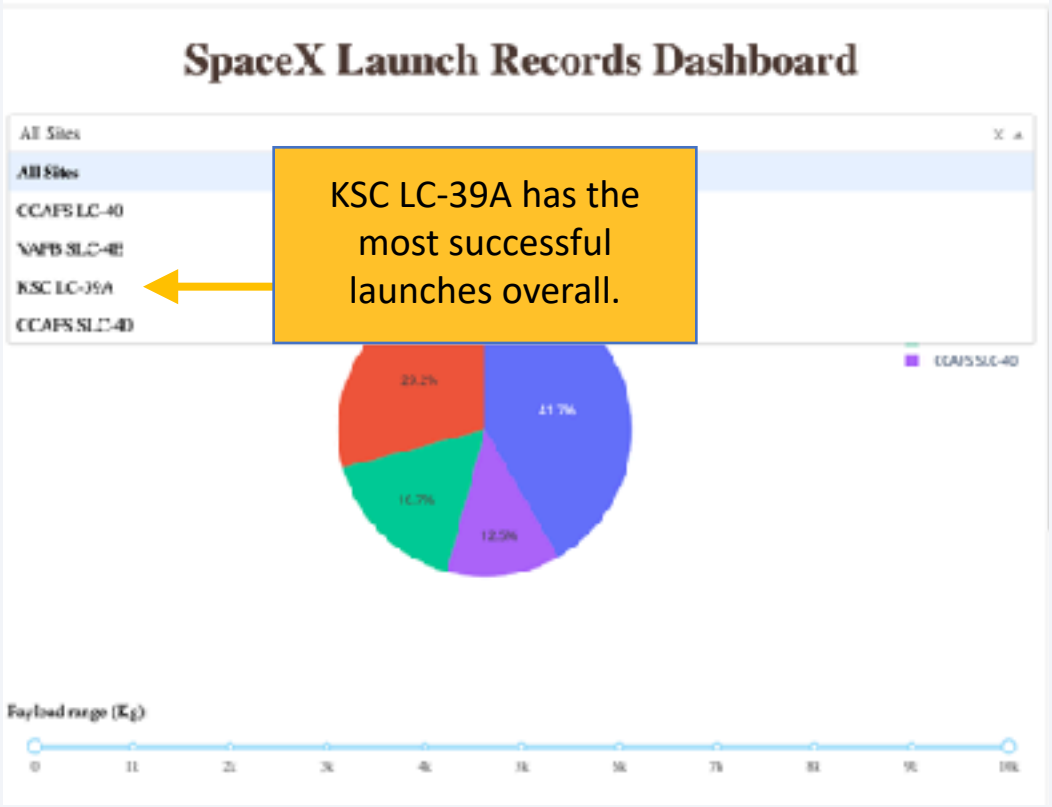
- LR, SVM, Decision Tree and KNN objects are created and fit with GridSearchCV object to find the best parameters, then the models are trained on the training set.
- The accuracy of the test data are calculated for each machine learning model. It is found that the methods performed best where LR, SVM, KNN; all 3 achieved highest accuracy of 83.33%




```
Unique labels in Y_test: [0 1]
Model Evaluation Results:

KNN:
Accuracy: 0.8333333333333334
F1-Score: 0.888888888888889
Jaccard Index: 0.8
LogLoss: 0.36621985248824704

SVM:
Accuracy: 0.8333333333333334
F1-Score: 0.888888888888889
Jaccard Index: 0.8
LogLoss: None

Logistic Regression:
Accuracy: 0.8333333333333334
F1-Score: 0.888888888888889
Jaccard Index: 0.8
LogLoss: 0.4786666068559153

Decision Tree:
Accuracy: 0.7777777777777778
F1-Score: 0.8571428571428571
Jaccard Index: 0.75
LogLoss: 0.5177300891540629

Best Predictor based on Accuracy: KNN with a score of 0.8333333333333334
Best Predictor based on F1-Score: KNN with a score of 0.888888888888889
Best Predictor based on Jaccard Index: KNN with a score of 0.8
Best Predictor based on LogLoss: KNN with a score of 0.36621985248824704
```

Results

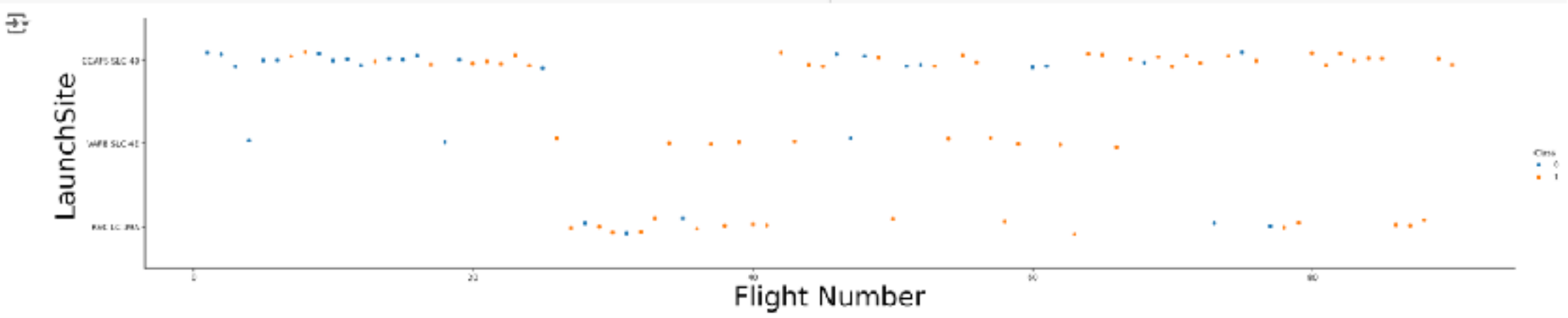
- LR, SVM, KNN are top-performer models in regards to forecasting outcomes in this data.
- Lighter payloads have a higher performance outcome, compared to heavier ones.
- Likelihood of SpaceX launch succeeding increases with number of years and experience, suggesting a trend towards flawless launches over time.
- Launch Complex 39A at Kennedy Space Center has highest number of successful launches compared to other launch sites based on findings.
- GEO, HEO, SSO, ES L1 orbit types exhibit the highest rates of successful launches.



Section 2

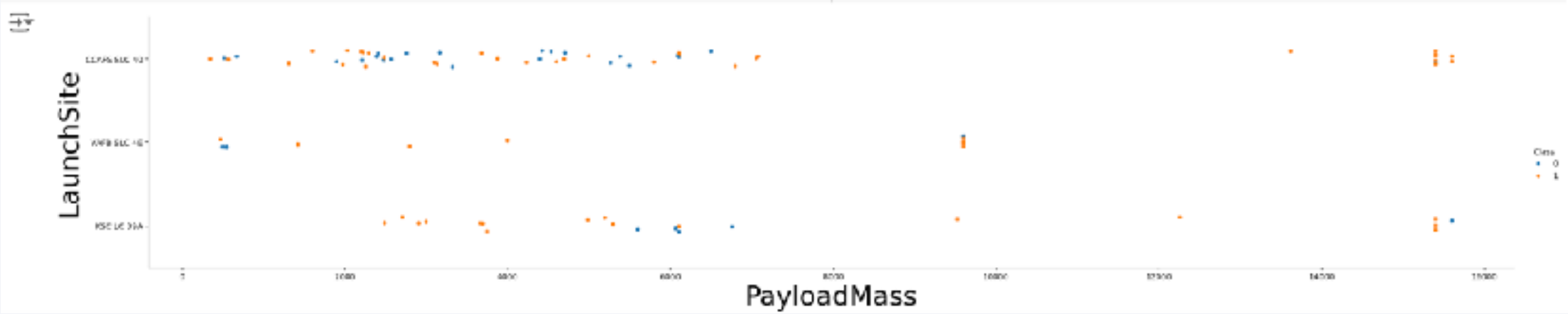
Insights drawn from EDA

Flight Number vs. Launch Site



- Scatterplot shows that CCAFS SLC 40 had significantly more total number of launches compared to the other sites.

Payload vs. Launch Site



- Scatterplot shows that payloads with ***less*** mass had more launches in comparison to payloads with more mass on all 3 launch sites.
- VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).
- We can also observe the heavy payload launches where conducted on CCAFS SLC 40 and KSC LC 39A with both sites conducting 5 flights and resulting in 4 success missions and one unsuccessful.

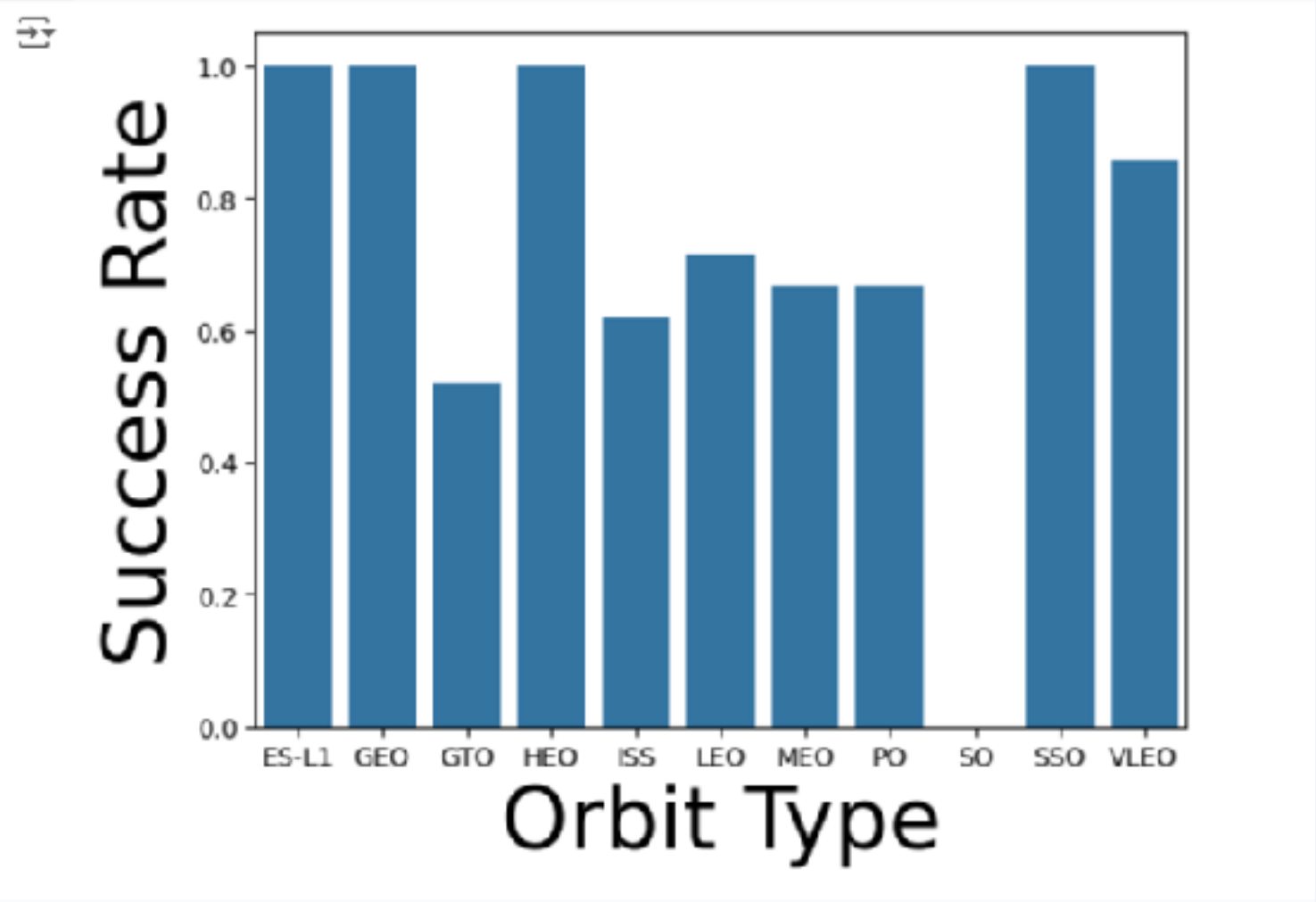
Success Rate vs. Orbit Type

To visually check if there are any relationship between success rate and orbit type a Bar Chart is created.

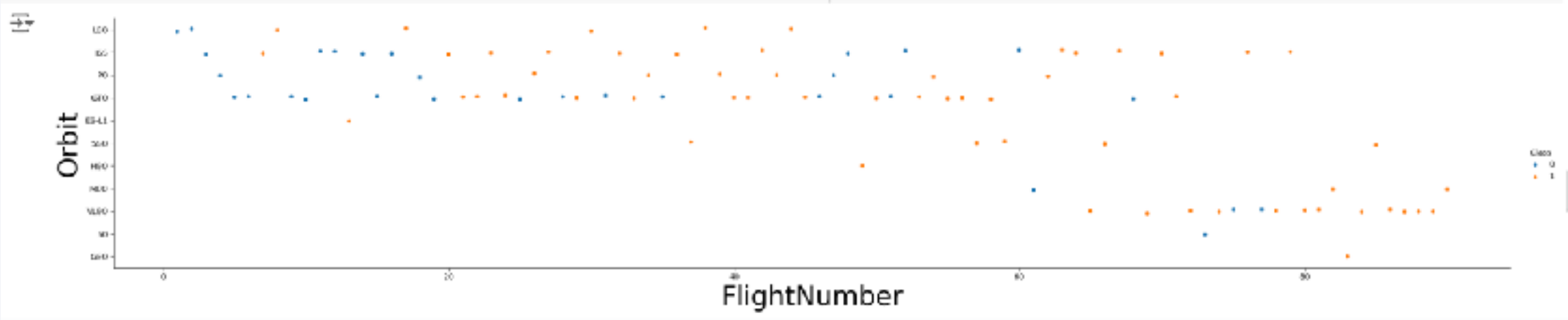
- Orbit types

  ES-L1
  GEO
  HEO
  SSO

  have the highest success rate compared to the other orbits

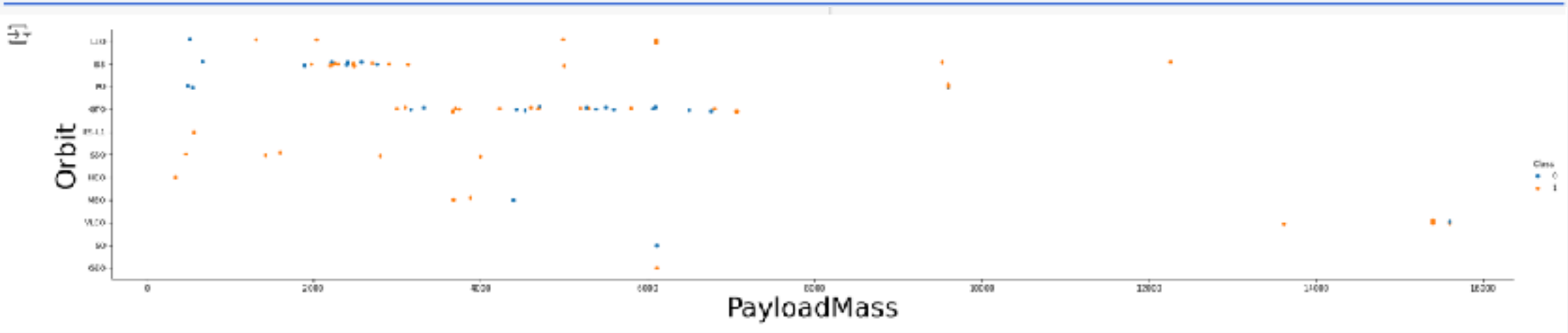


Flight Number vs. Orbit Type



- We can see in the Scatterplot that in the earlier years LEO, ISS, PO, GTO orbits had the most launches, however as we proceed in time we see that the launch site changes to VLEO orbit.

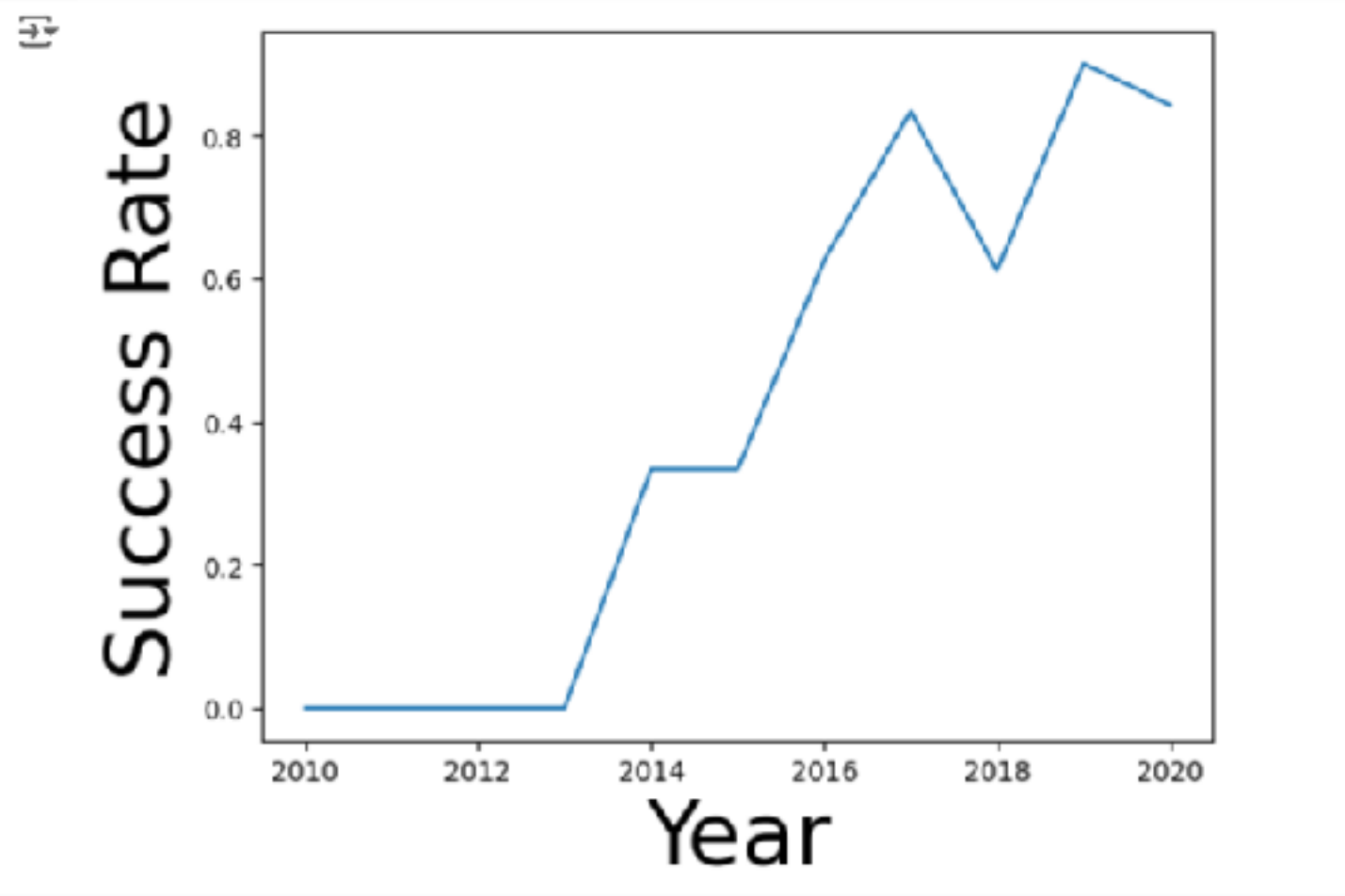
Payload vs. Orbit Type



- With heavy payloads, successful landing or positive landing rate are in favour 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 and here.

Launch Success Yearly Trend

- We can observe that the success rate since 2013 kept increasing till 2017 (stable in 2014) and after 2015 it started increasing.



All Launch Site Names

- Performed a SQL query with the keyword ‘DISTINCT’ to display only unique launch sites from the SpaceX table.

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

```
 * ibm_db_sa://[redacted]?security=SSL
Done.
  launch_site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E
```

Launch Site Names Begin with 'CCA'

- With SQL a query was made to display only 5 launch sites that start with ‘CCA’

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

```
 * ibm_db_sa://[redacted]
Done.
```

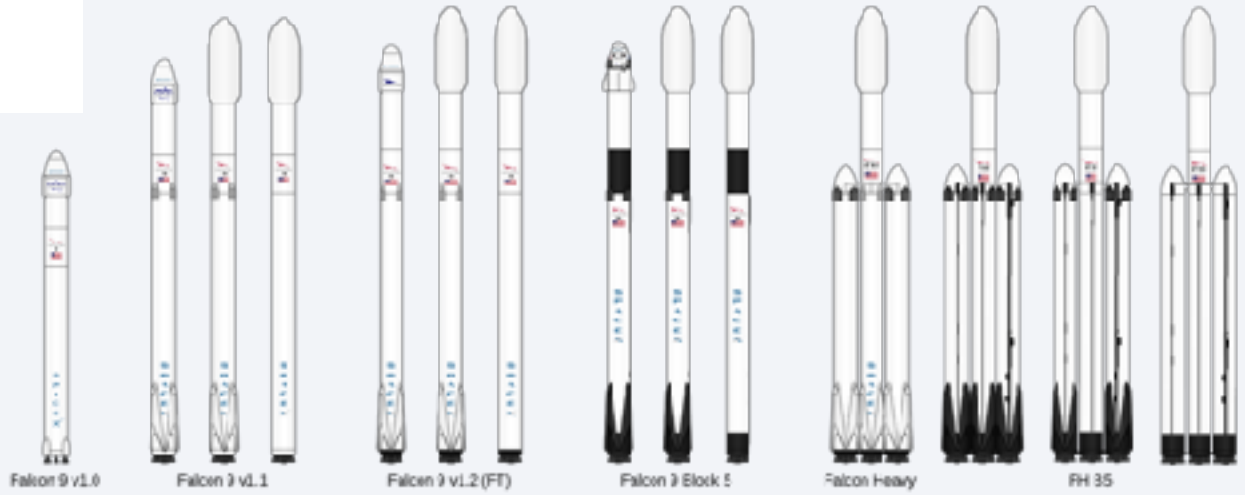
DATE	Time (UTC)	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing_outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS) NRO	NASA (COTS)	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

- Displaying with SQL the total payload mass in KG carried by boosters launched by NASA (CRS)

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

```
 * ibm_db_sa://[redacted]
Done.
  1
45596
```



Average Payload Mass by F9 v1.1

- Calculated the average payload mass in KG carried by booster version F9 v1.1 with SQL query.

```
[43] %%sql
     SELECT AVG(PAYLOAD_MASS__KG_)
     FROM SPACEXTBL
     WHERE Booster_Version LIKE 'F9 v1.1';

 * ibm_db_sa://[redacted]
Done.
  1
2928
```

First Successful Ground Landing Date

- Listing with SQL query the date when the first successful landing outcome in ground pad was achieved.

```
[35] %%sql
     SELECT MIN(DATE)
     FROM SPACEXTBL
     WHERE LANDING_OUTCOME = 'Success (ground pad)'
```

```
 * ibm_db_sa://[redacted]
Done.
    1
2015-12-22
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- Listing the names of boosters with a WHERE clause which have successfully landed on drone ship and had a payload mass greater than 4000 but less than 6000 KG.

```
[46] %%sql
     SELECT BOOSTER_VERSION
     FROM SPACEXTBL
     WHERE LANDING_OUTCOME = 'Success (drone ship)'
         AND 4000 > PAYLOAD_MASS__KG_ < 6000;
```

```
 * ibm_db_sa://xlz28193:***@b1bc1829-6f45-4cd4-bef4
Done.
```

booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1029.1
F9 FT B1021.2
F9 FT B1036.1
F9 B4 B1041.1
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- With the SQL query we group the MISSION_OUTCOME from the Space X table by selecting the counts of Failure and Success mission outcomes.

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

```
 * ibm_db_sa://[redacted]
Done.
```

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

Boosters Carried Maximum Payload

- Listing the names of the booster_versions which have carried the maximum payload mass. This subquery finds the maximum value of PAYLOAD_MASS__KG_ from the SPACEXTBL table. The outer query uses the result of the subquery to filter rows.

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

```
 * ibm_db_sa://xlz28193:***@b1bc1829-6
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

- Listing the failed landing_outcomes in drone ship and booster versions, including launch site names for in the year 2015 with the months. SQLite query using substr() for Date format.

```
%%sql
SELECT substr(DATE, 6, 2) as MONTH, LANDING_OUTCOME, BOOSTER_VERSION, LAUNCH_SITE
FROM SPACEXTBL
WHERE LANDING_OUTCOME = 'Failure (drone ship)'
    AND substr(DATE, 1, 4) = '2015';
```

```
 * ibm_db_sa://[redacted]
Done.
```

MONTH	landing_outcome	booster_version	launch_site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- With a SQL query 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. The GROUP BY clause was called to group the landing_outcome while we request to have it descending to be displayed by TOTAL_NUMBER.

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

```
 * ibm_db_sa://[redacted]
Done.
```

landing_outcome	total_number
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 map

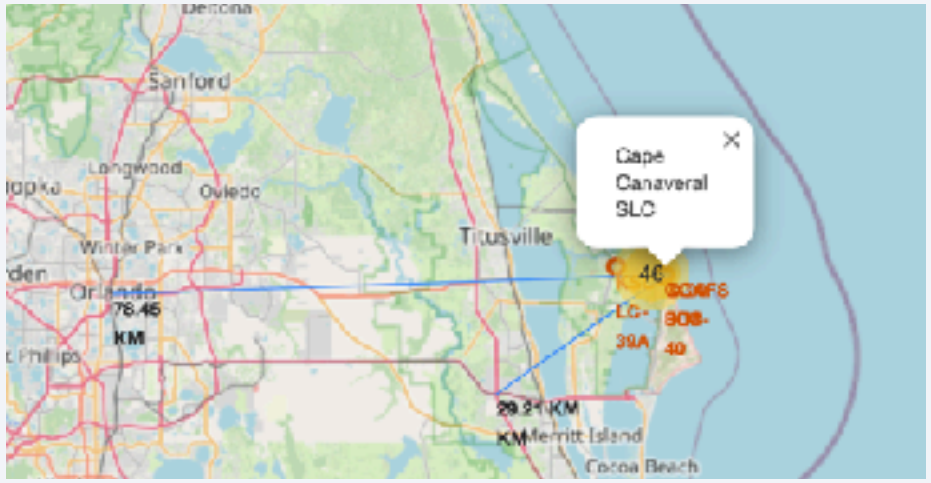
- The launch sites are marked red and are labeled on the interactive map created in Folium

Successful and Failed launches on the map

- The launch records are grouped in clusters on the map. Then labelled by green markers for successful launches and red markers indicate unsuccessful ones.

Distance between launch site to its proximities

- Marking down a point on the closest coastline using MousePosition and calculating the distance between the coastline point and the launch site
- Are launch sites in close proximity to railways? **NO**
- Are launch sites in close proximity to highways? **NO**
- Are launch sites in close proximity to coastline? **YES**
- Do launch sites keep certain distance away from cities? **YES**

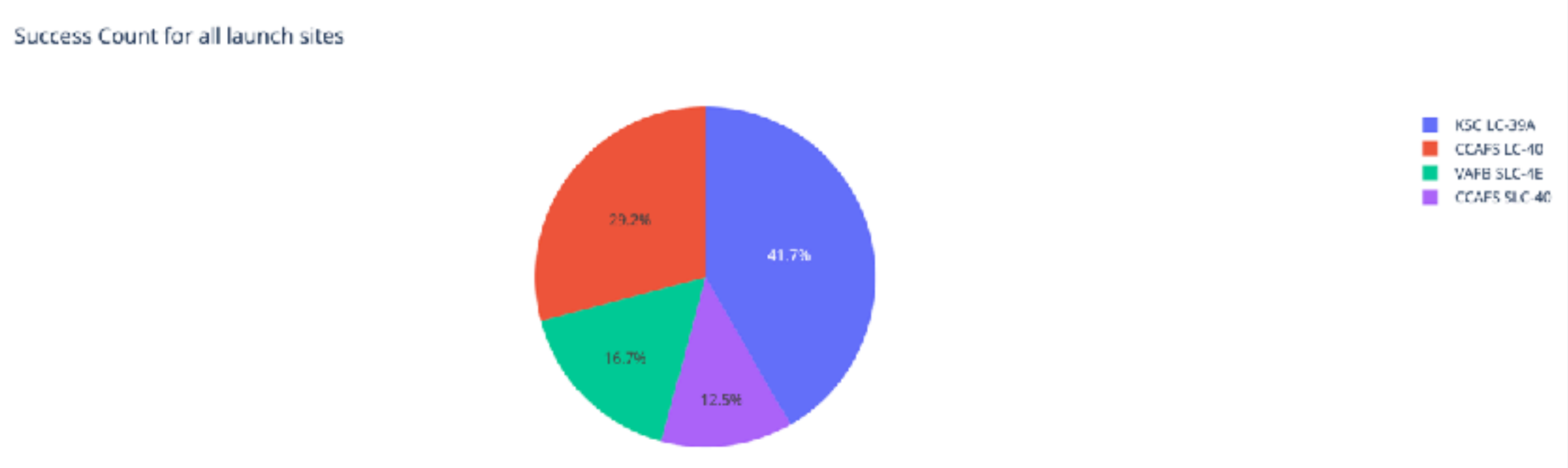


Section 4

Build a Dashboard with Plotly Dash

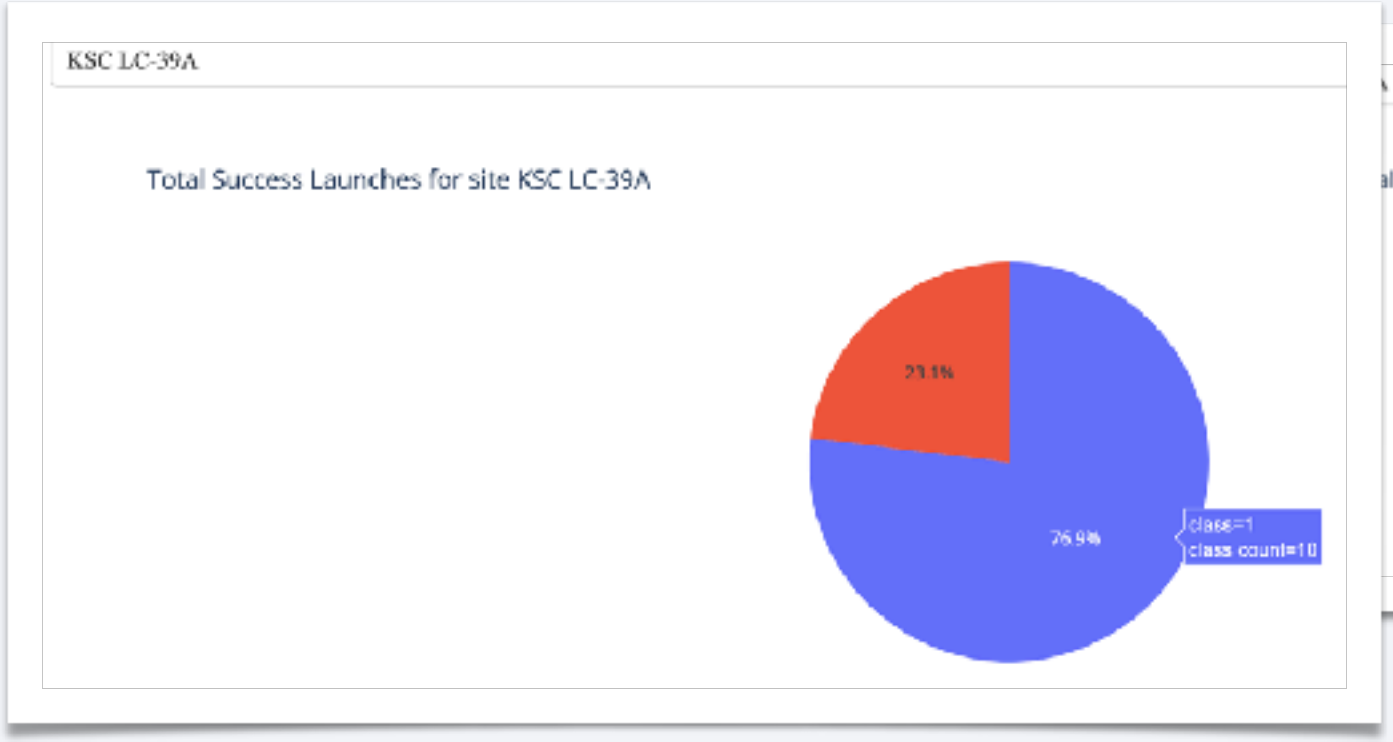
Pie Chart representing Success Count for all launch sites

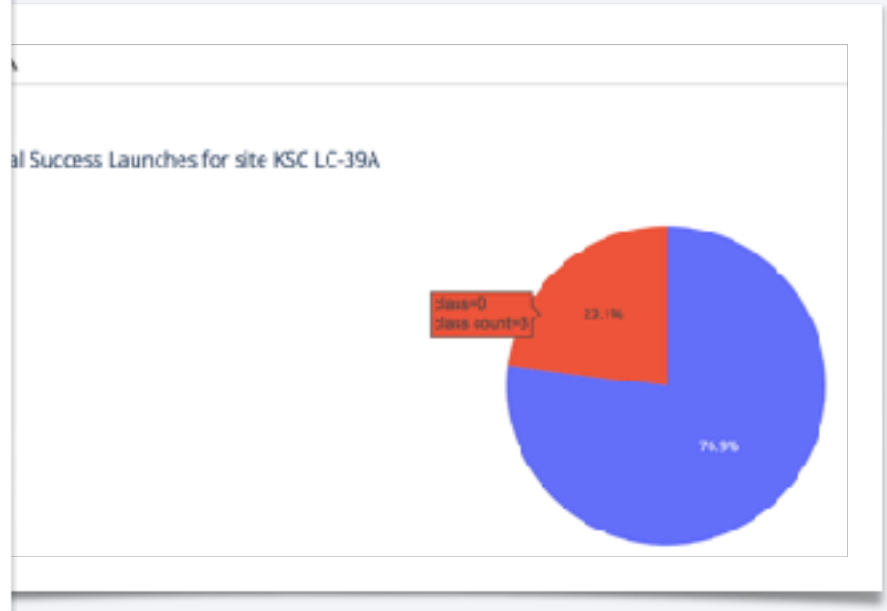
- At first glance we see that **KSC LC-39A** has the most success rate in in our Pie Chart with 41.7% and in second place we see that **CCAFS LC-40** with 29.2% success rate while the launch site with the lowest success rate is **CCAFS SLC-40** with 12.5%



Success ratio of launch site with the highest success launches

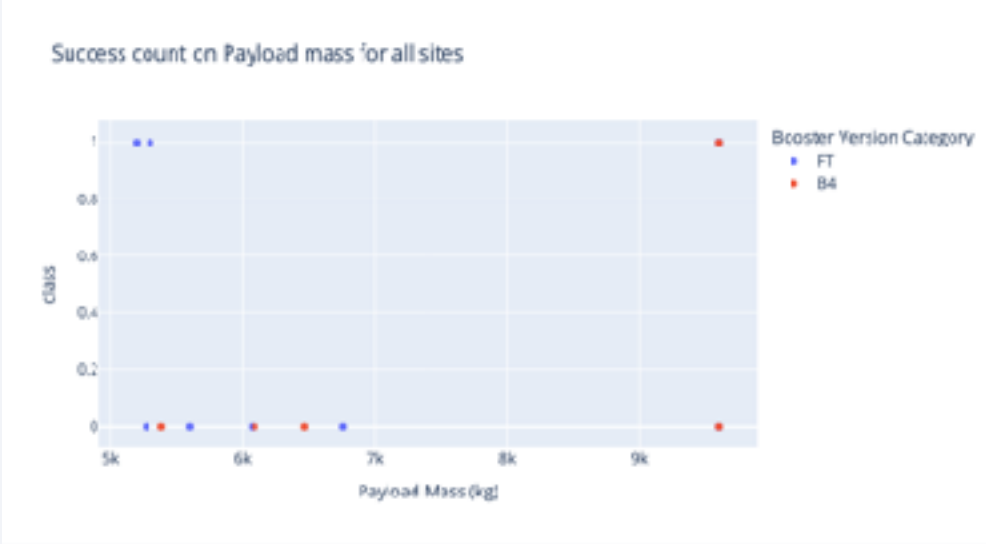
- We can see more detail when hovering over our pie chart to obtain more information, there is a count of 10 success counts which holds 76.9% and 3 unsuccessful ones of a total of 23.1% on the pie chart.





Scatter plot Payload vs. Launch site - All sites with different payloads

- With the range slider on our Dashboard two different payloads were selected to show the Payload vs launch outcome. On the left we have 0 KG to 5000kg payload scatter plot and on the right we have 5000 KG to 10.000 KG payload on our plot.
- Booster FT marked with green dots indicates the highest success rate, in second place we have Booster B4 displayed in purple.

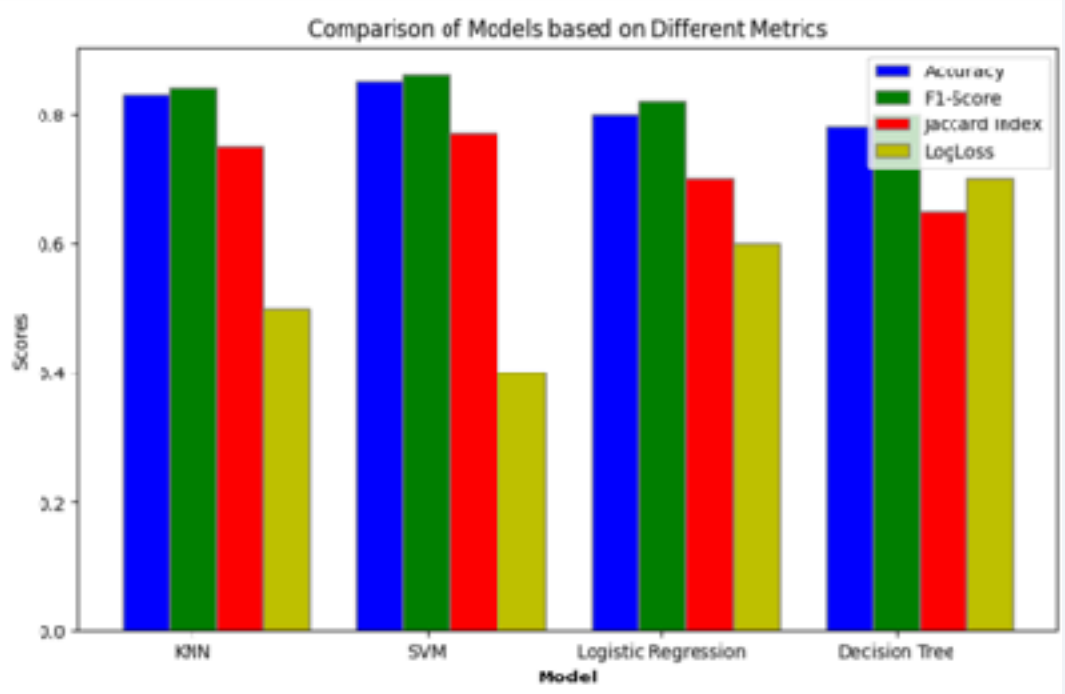


Section 5

Predictive Analysis (Classification)

Classification Accuracy

- All models performed well based on outcome with 83.3%

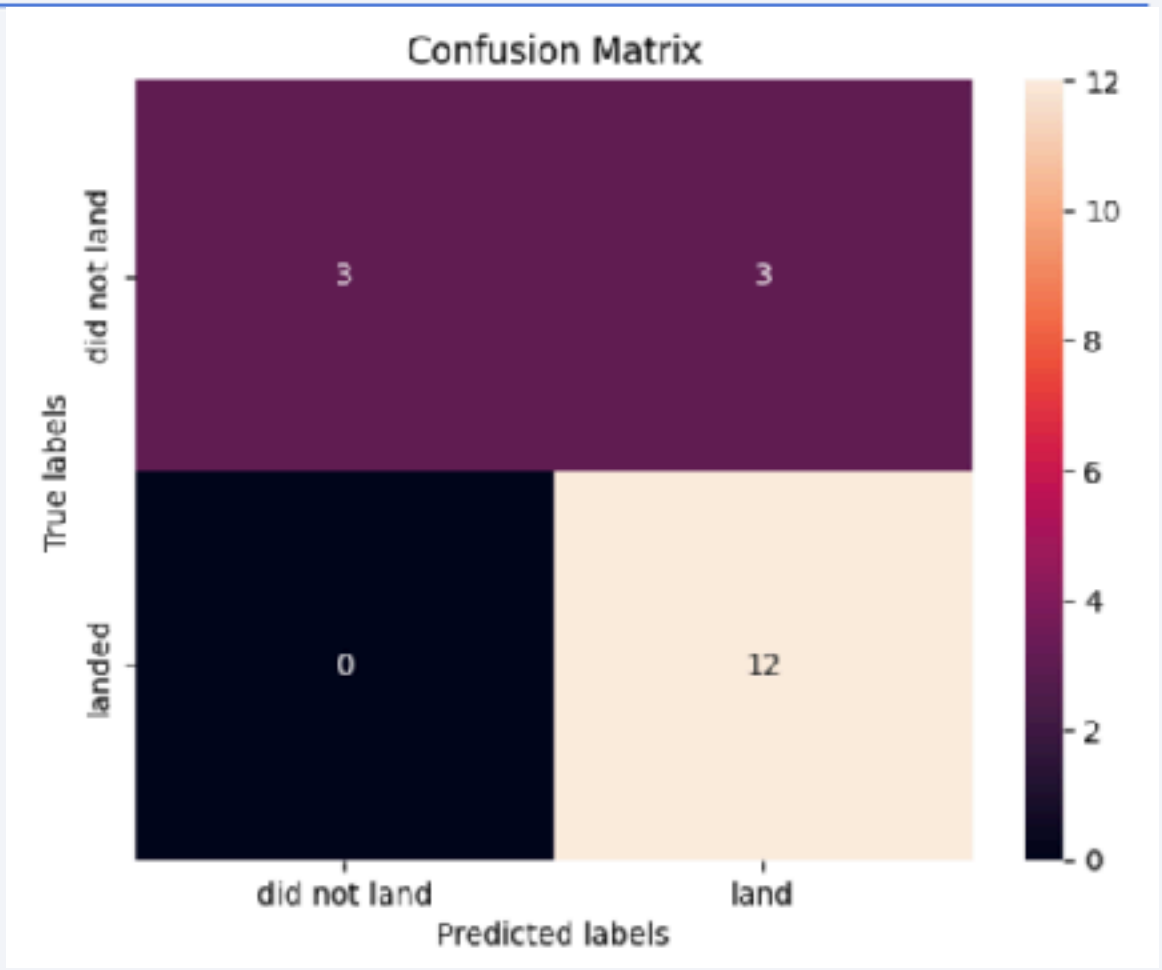


	Model	Accuracy	F1-Score	Jaccard Index	LogLoss
0	KNN	0.833333	0.888889	0.8	0.366219
1	SVM	0.833333	0.888889	0.8	NaN
2	Logistic Regression	0.833333	0.888889	0.8	0.478667
3	Decision Tree	0.833333	0.888889	0.8	0.402821

Best Predictor based on Accuracy: KNN

Confusion Matrix

- Classifier can distinguish between different classes in the confusion matrix for the decision tree classifier.



Conclusions

- LR,SVM and KNN are top performers in regards to other launch sites.
- Orbit types **ES-L1, GEO, HEO, SSO** have the highest success rate compared to the other orbits.
- Kennedy Space Center **KSC LC-39A** had the most highest recorded successful launches compared to other sites.
- Lighter payloads have better performance in comparison to heavier payloads.
- Launch Sites are located in close approximate to water (Florida and California).
- The safety for citizens is taken into account by having a big distance to public infrastructure.
- We can observe that the success rate since 2013 kept increasing till 2017 (stable in 2014) and after 2015 it started increasing.

Thank you!