IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Nishant
Medpalwar
06.08.2023

Outline

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

Executive Summary

- Used methodologies:
  - Data collection using SpaceX API and web scraping
  - Exploratory Data Analysis, including data wrangling, data visualization and creating interactive dashboard
  - Predictive analysis using machine learning algorithms

- Summary of all results:
  - EDA identified the best features to predict the success of launchings
  - Machine learning prediction found the models with relatively high accuracy

Introduction

- The objective is to evaluate the competitiveness of the new rocket company Space Y with SpaceX

- Desirable outcome:
    - the best way to estimate the cost of launch, by predicting if the first stage of the rocket will be reused

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
    - Data was collected from 2 sources:
        - SpaceX API (https://api.spacexdata.com/v4/rockets/)
        - Wikipedia web scraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)
- Perform data wrangling
    - The data was enhanced by creating a landing outcome label based on the Outcome data and preceding feature analysis
- Perform exploratory data analysis (EDA) using visualization and SQL

Methodology

Executive Summary

- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Collected data was standardized, splitted into train and test sets and evaluated by four different classification models. The best combination of hyperparameters for each model was found by Grid Search and then the accuracies were computed and compared.

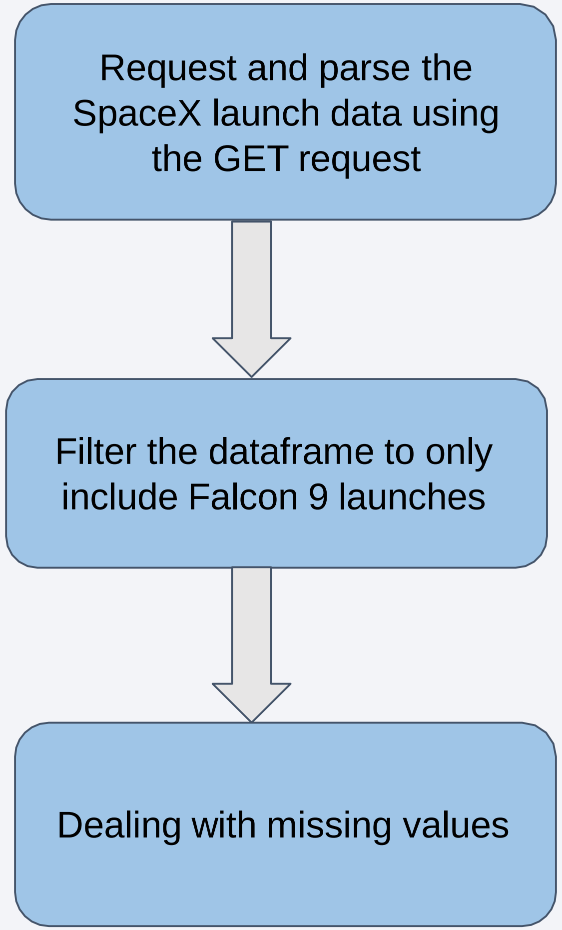
Data Collection

- Data sets were collected from 2 sources:
    - SpaceX API (https://api.spacexdata.com/v4/rockets/)
    - Wikipedia web scraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)

Data Collection – SpaceX API

- SpaceX offers API where the data can be obtained from

•Flowchart shows data collection with SpaceX REST calls

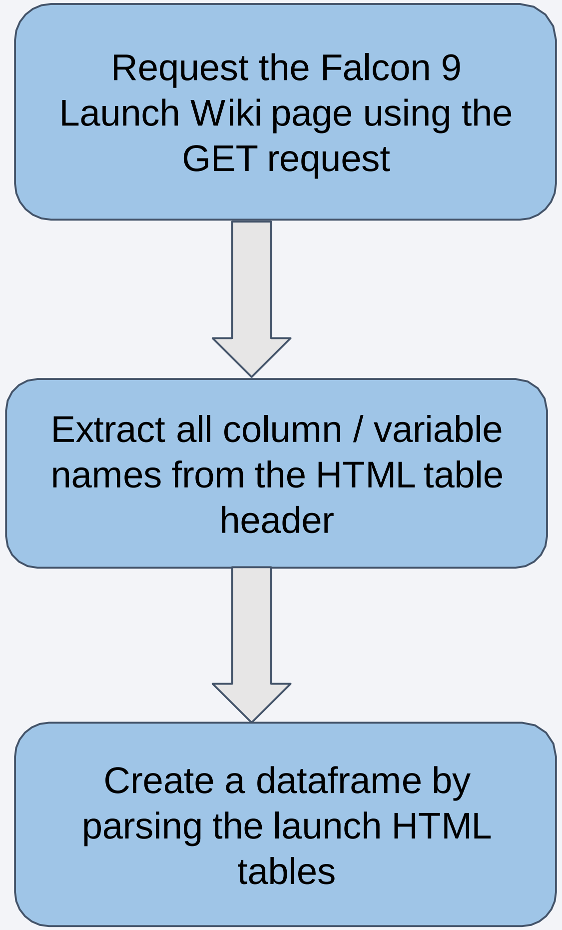


Source code: https://github.com/Nishn3/Data-science-capstone-space-x-falcon-9/blob/main/Data_Collection_API.ipynb

Data Collection - Scraping

- SpaceX data can also be obtained from Wikipedia

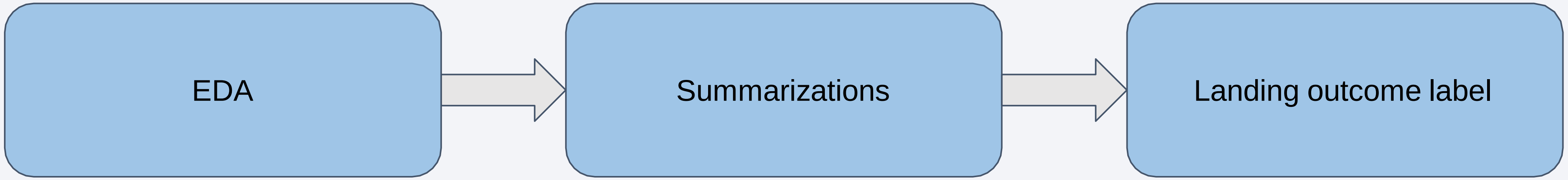
- Flowchart shows data collection and manipulation from Wikipedia



Source code:
https://github.com/Nishn3/Data-science-capstone-space-x-falcon-9/blob/main/Webscraping.ipynb

Data Wrangling

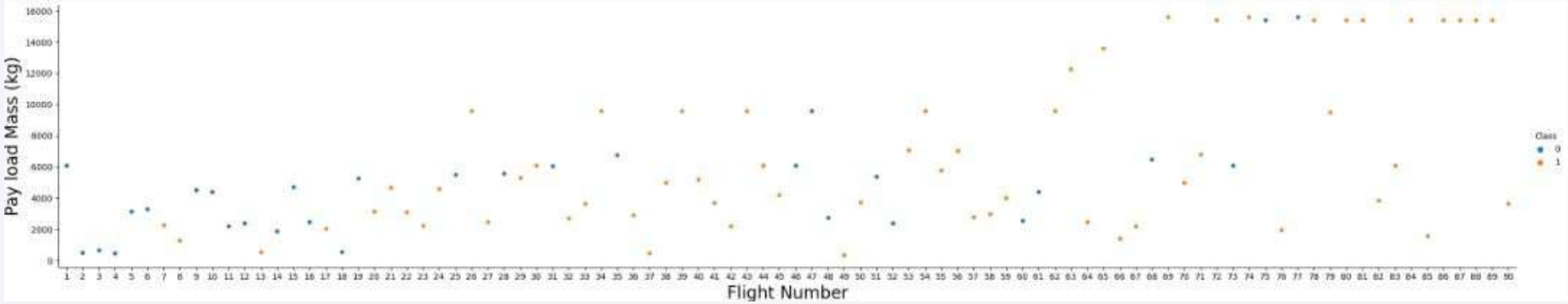
- Initial Exploratory Data Analysis was performed on the dataset
- Performed summarizations - number of launches per site, occurrences of each orbit, occurrences of mission outcome per orbit
- The landing outcome label was created based on Outcome column



Source code:
https://github.com/Nishn3/Data-science-capstone-space-x-falcon-9/blob/main/Data_Wrangling.ipynb

EDA with Data Visualization

- The following charts were used for further analysis and visualization of relationship between the pairs of features:
  - **scatterplot** - Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Flight Number vs. Orbit type, Payload Mass vs. Orbit Type
  - **bar chart** - Orbit Type vs. Success Rate



Source code:
https://github.com/Nishn3/Data-science-capstone-space-x-falcon-9/blob/main/Data_Vizualisation.ipynb

EDA with SQL

- The following SQL queries were performed and displayed:
  - Names of the unique launch sites in the space mission
  - Top 5 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
  - 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
  - Failure landing outcomes in drone ship and their booster versions, launch site names in year 2015
  - Rank of the count of successful landing outcomes between the date 04-06-2014 and 20-03-2017 in descending order

Source code:https://github.com/Nishn3/Data-science-capstone-space-x-falcon-9/blob/main/EDA_SQL.ipynb

Build an Interactive Map with Folium

- Markers, marker clusters, circles and lines were added to a Folium map
    - Markers indicate coordinates / points (like launch sites)
    - Marker clusters indicate a group of markers with the same specific characteristic (like markers having the same coordinate)
    - Circles indicate an highlighted area around a specific coordinate (like NASA Johnson Space Center)
    - Lines indicate the distance between two coordinates

Source code:https://github.com/Nishn3/Data-science-capstone-space-x-falcon-9/blob/main/Launch_Site_Location_Folium.ipynb

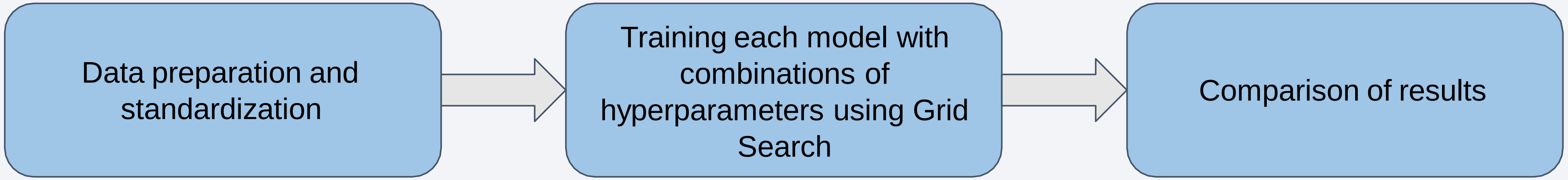
Build a Dashboard with Plotly Dash

- Following graphs were used:
  - **pie chart** to show the total successful launches for all sites
  - a **slider** to select of payload range
  - **scatter plot** to show the correlation between the payload and launch success for each site

Source code:https://github.com/Nishn3/Data-science-capstone-space-x-falcon-9/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Four classification models were used for predictive analysis: Logistic regression, SVM, Decision tree classifier, k Nearest neighbors
- The best combination of hyperparameters for each model was found by Grid Search



Source code:https://github.com/Nishn3/Data-science-capstone-space-x-falcon-9/blob/main/Machine_Learning_Prediction.ipynb

Results

- Exploratory Data Analysis results:
    - SpaceX uses 4 launch sites
    - Average payload mass carried by booster version F9 v1.1 was 2928 kg
    - The first successful landing outcome in ground pad was achieved in 2017
    - Almost 100% of mission outcomes were successful

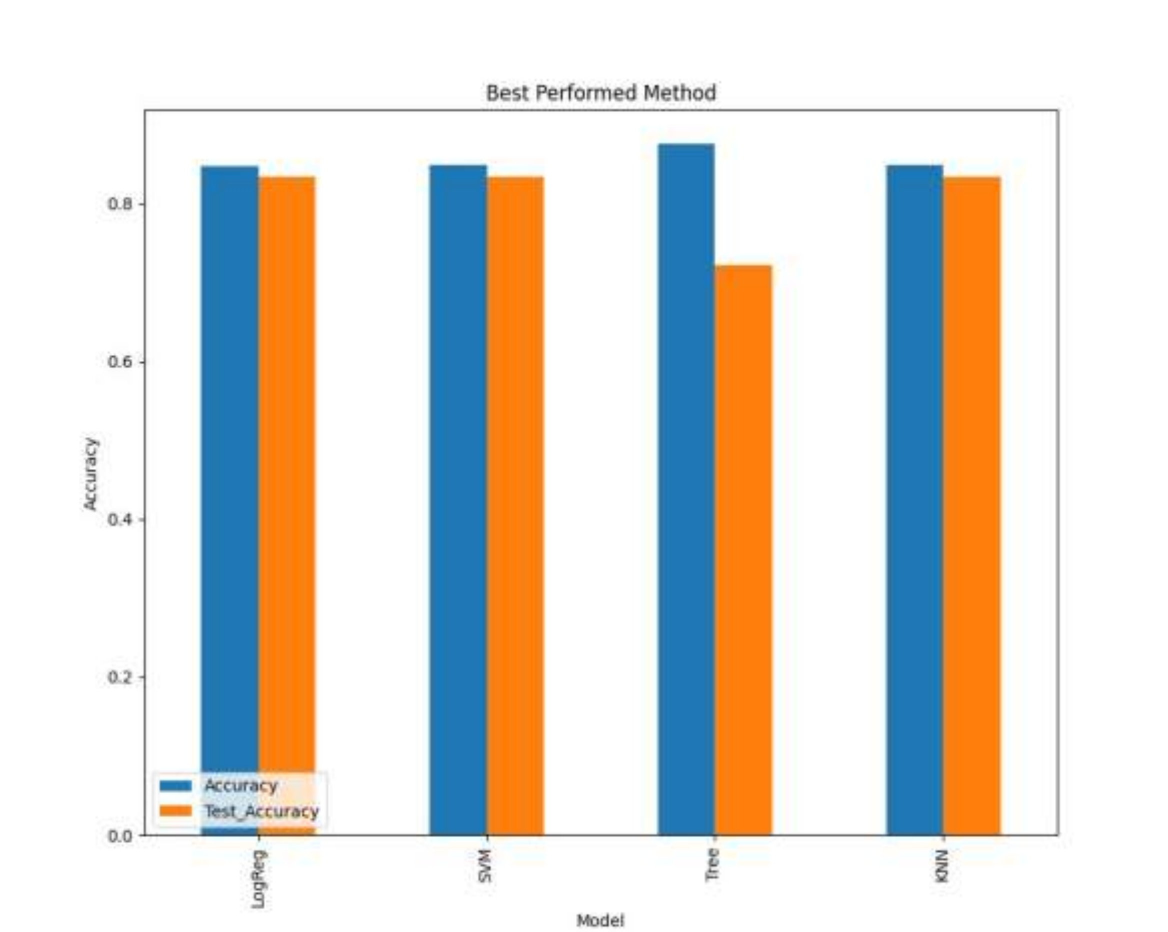
Results

- Using interactive analytics was possible to identify the locations of launch sites
- They are usually located near the sea with a good logistic infrastructure around

Results

- Predictive analysis showed that Logistic regression, SVM and k Nearest neighbors produced similar results with accuracy rate of about 83.33%

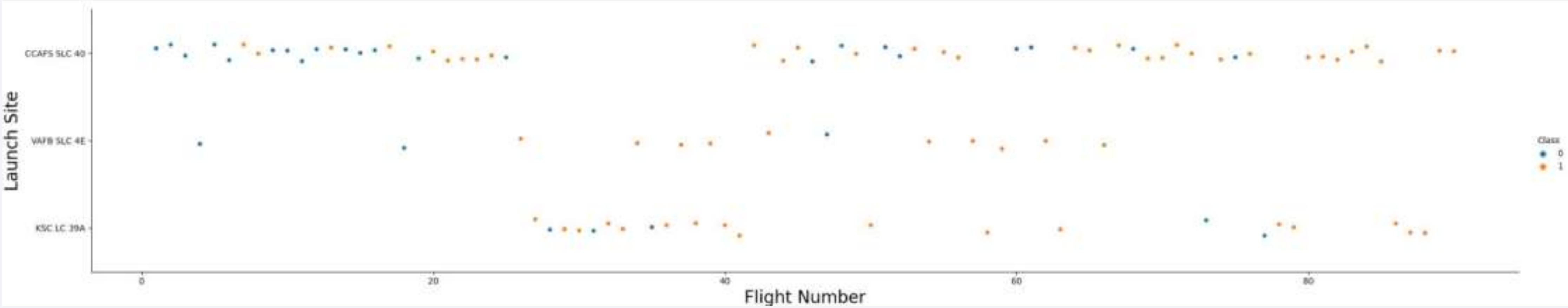
- More data is needed for better model determination and accuracy



Section 2

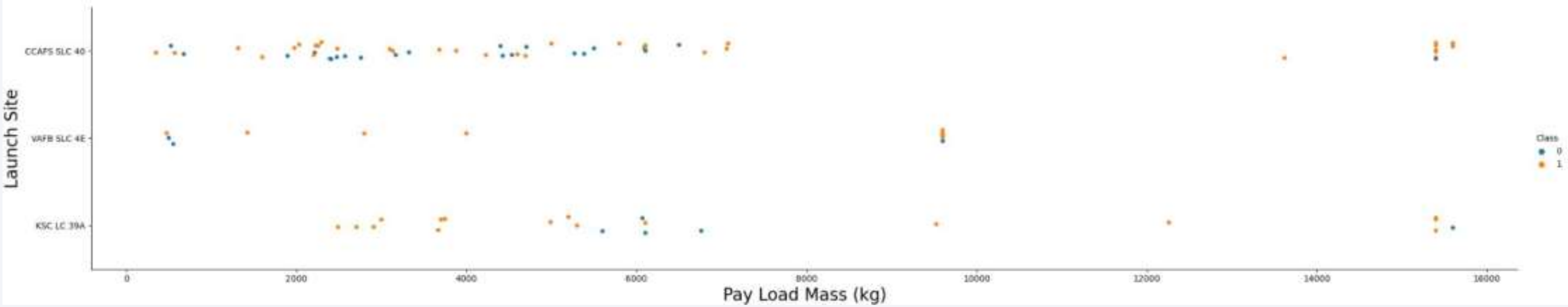
Insights drawn from EDA

Flight Number vs. Launch Site



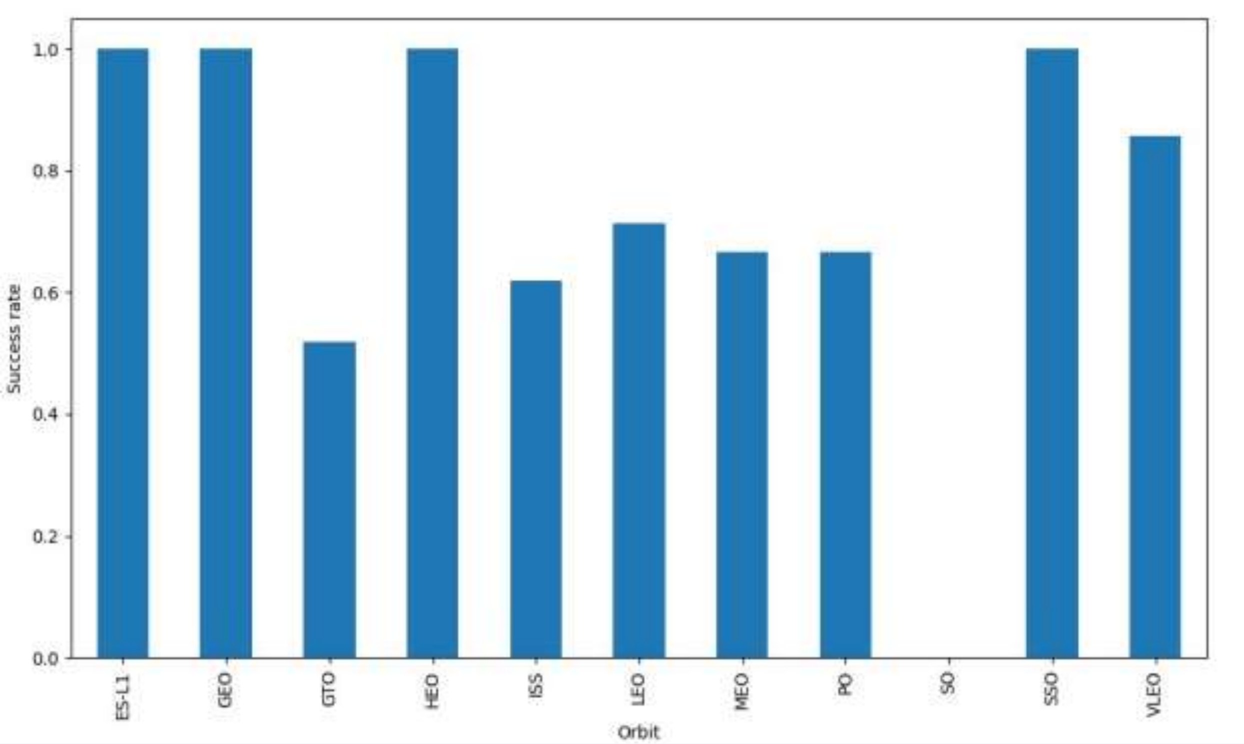
- The higher the flight number, the higher chance of successful landing of the rocket’s first stage
- The most successful launch site is CCAFS SLC 40

Payload vs. Launch Site



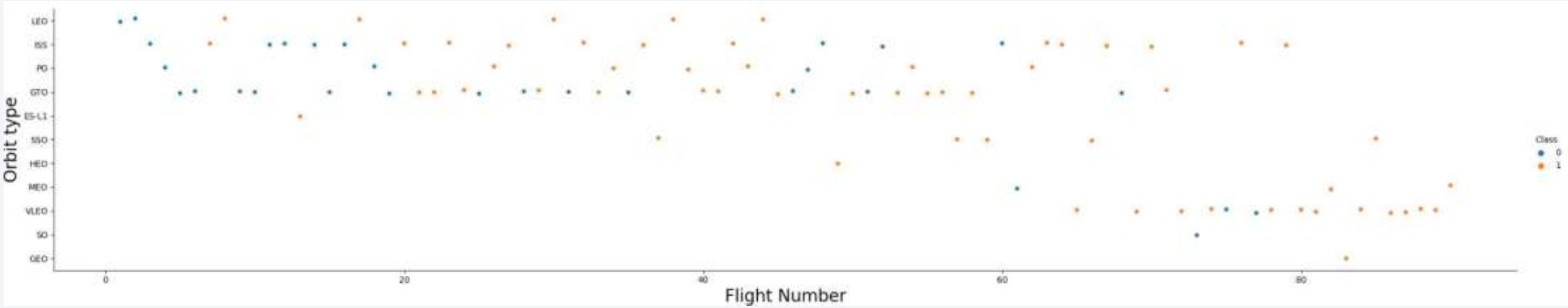
- VAFB-SLC 4E : no rockets launched with payload mass greater than 10000 kg
- Payloads over 9000 kg have high success rate

Success Rate vs. Orbit Type



- the highest success rate : ES-L1, GEO, HEO, SSO orbits
- no success rate - SO orbit

Flight Number vs. Orbit Type



- LEO orbit : strong relation to the number of flights
- GTO orbit : no relation
- in general, success rate improved for almost all orbits

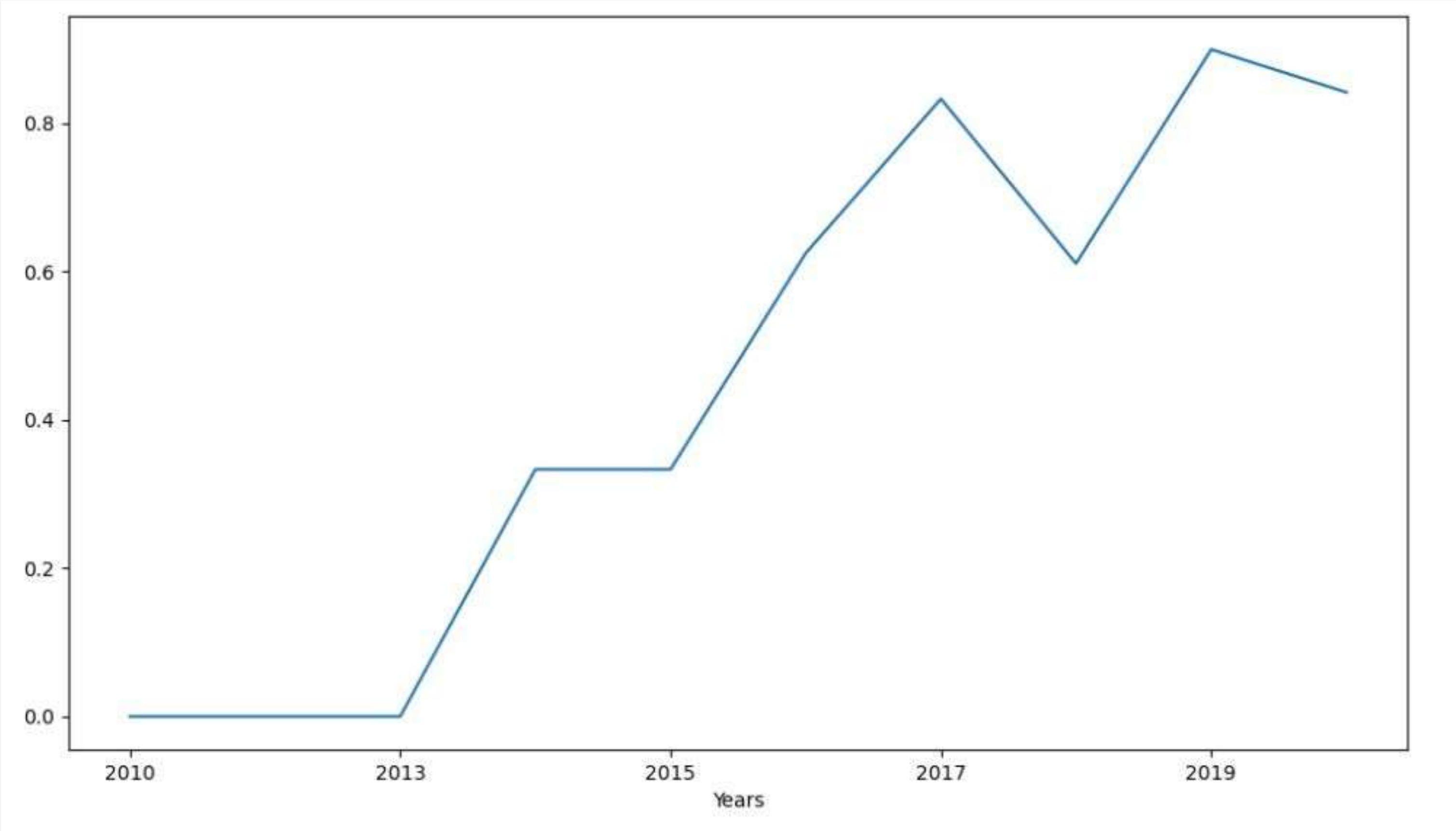
Payload vs. Orbit Type

- GTO : no relation between payload and success rate
- SO, GEO : only few records
- ISS, PO, LEO : high success rate for heavy payloads

Launch Success Yearly Trend

- The success rate kept increasing since 2013 till 2020

- A slight fall in 2018



All Launch Site Names

- Find the names of the unique launch sites

```
In [10]: %sql SELECT DISTINCT "Launch_Site" FROM SPACEXTBL;

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

Out[10]:

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`

In [11]:
```
%sql SELECT * FROM SPACEXTBL WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5;
```

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

Out[11]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Calculate the total payload carried by boosters from NASA (CRS)

```
In [21]: %sql SELECT SUM("PAYLOAD_MASS__KG_") AS "Total Payload Mass" FROM SPACEXTBL WHERE Customer LIKE "NASA (CRS)";

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

Out[21]:

Total Payload Mass
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1

```
In [26]: %sql SELECT AVG("PAYLOAD_MASS__KG_") AS "Average Payload Mass" FROM SPACEXTBL WHERE "Booster_Version" LIKE "F9 v1.1";

 * sqlite:///my_data1.db
Done.

Out[26]:
```

Average Payload Mass
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

In [34]:
```
%%sql SELECT MIN(Date) AS "First succesful landing outcome in GP" FROM SPACEXTBL
WHERE "Landing _Outcome" LIKE "Success (ground pad)";
```

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

Out[34]:

First succesful landing outcome in GP
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
In [42]: %%sql SELECT DISTINCT "Booster_Version" FROM SPACEXTBL
WHERE ("Landing _Outcome" LIKE "Success (drone ship)") AND ("PAYLOAD_MASS__KG_" BETWEEN 4000 AND 6000) ;

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

Out[42]:

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

In [50]:
```
%sql SELECT "Mission_Outcome", COUNT(*) AS "Count" FROM SPACEXTBL GROUP BY "Mission_Outcome" ORDER BY "Mission_Outcome";
```

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

Out[50]:

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

```
In [53]: %%sql SELECT DISTINCT "Booster_Version" FROM SPACEXTBL
         WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTBL);

 * sqlite:///my_data1.db
Done.

Out[53]:
```

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 [55]: %%sql SELECT substr(Date, 4, 2) AS "Month", "Landing _Outcome", "Booster_Version", "Launch_Site" FROM SPACEXTBL
         WHERE ("Landing _Outcome" LIKE "Failure (drone ship)") AND (substr(Date,7,4)='2015');

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

Out[55]:

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

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

```
In [59]: %%sql
SELECT "Landing _Outcome", COUNT(*) AS "Count" FROM SPACEXTBL
WHERE ("Landing _Outcome" LIKE "%Success%") AND (Date BETWEEN '04-06-2010' AND '20-03-2017')
GROUP BY "Landing _Outcome" ORDER BY "Count" DESC;

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

Out[59]:

Landing _Outcome	Count
Success	20
Success (drone ship)	8
Success (ground pad)	6

Section 3

Launch Sites Proximities Analysis

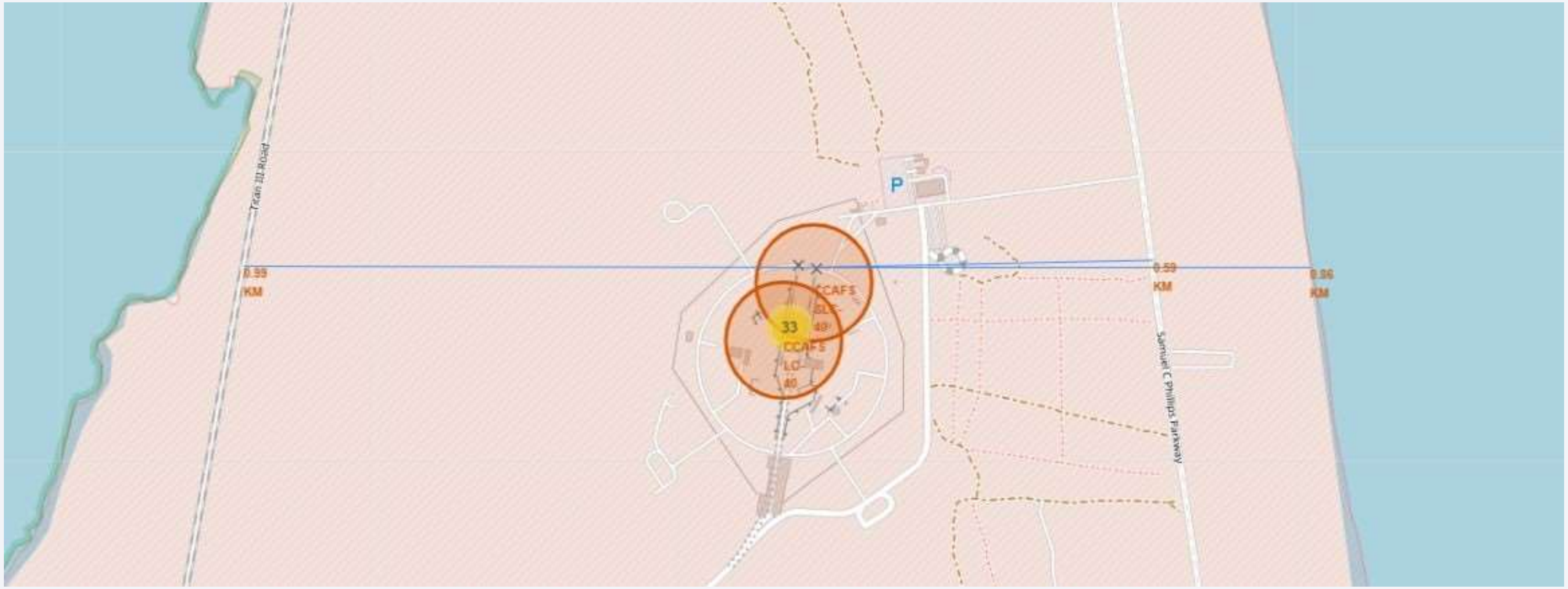
Launch sites

- All launch sites are located close to the sea, probably for safety reasons

Launches for each site

- example of KSC LC-39A launch site
- green markers indicate successful launches, red indicates failed launches

Launch sites proximities

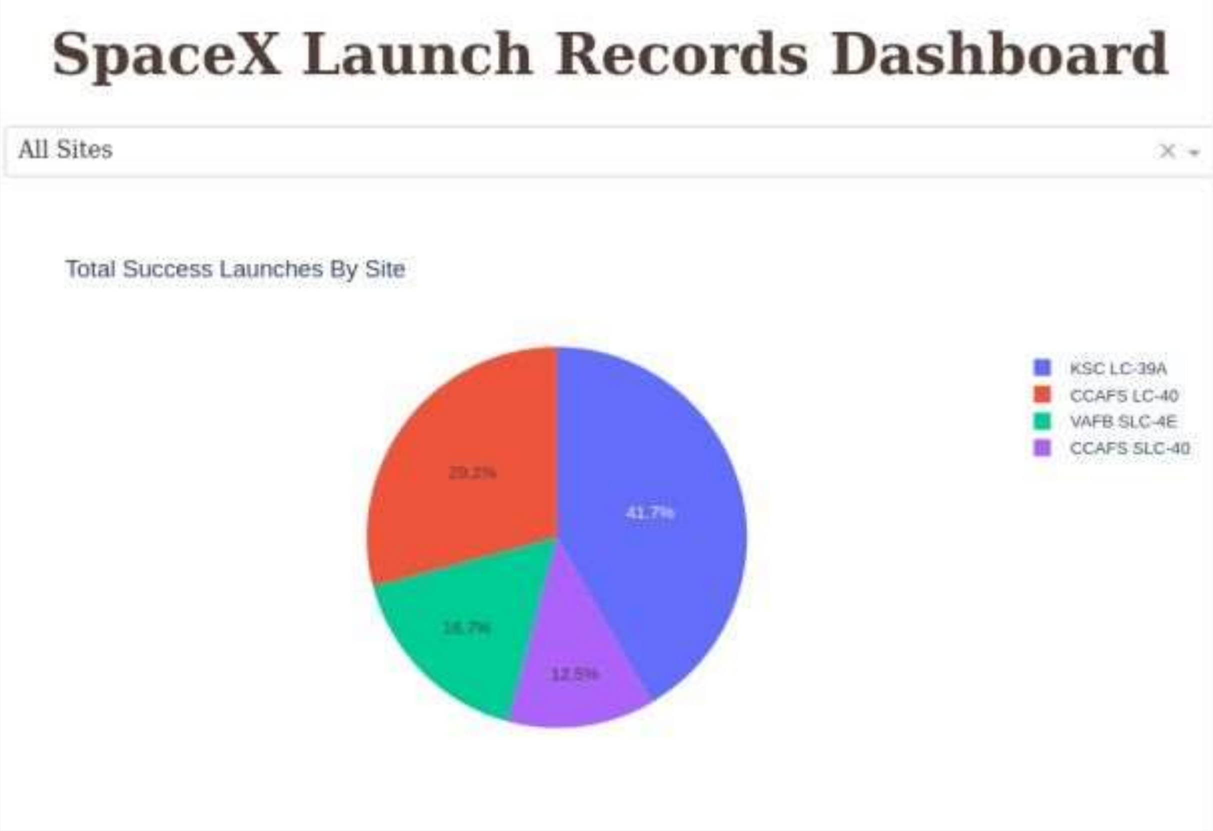


- picture shows that launch sites are located close to the railway, highway and the coastline

Section 4

Build a Dashboard with Plotly Dash

Total successful launches by site

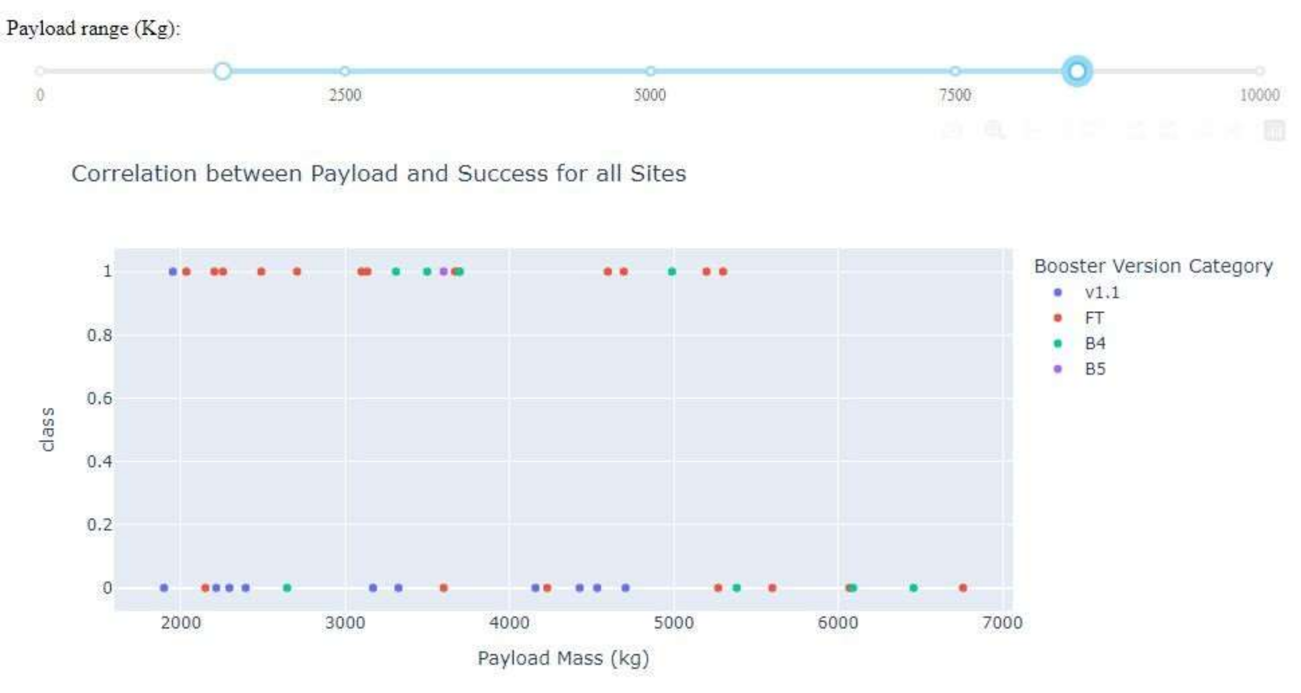


- seems that the location of launch site has an impact on the success of a mission

Launch site with the highest success ratio

- 76.9% of launches are successful for KSC LC-39A

Payload vs. Launch Outcome



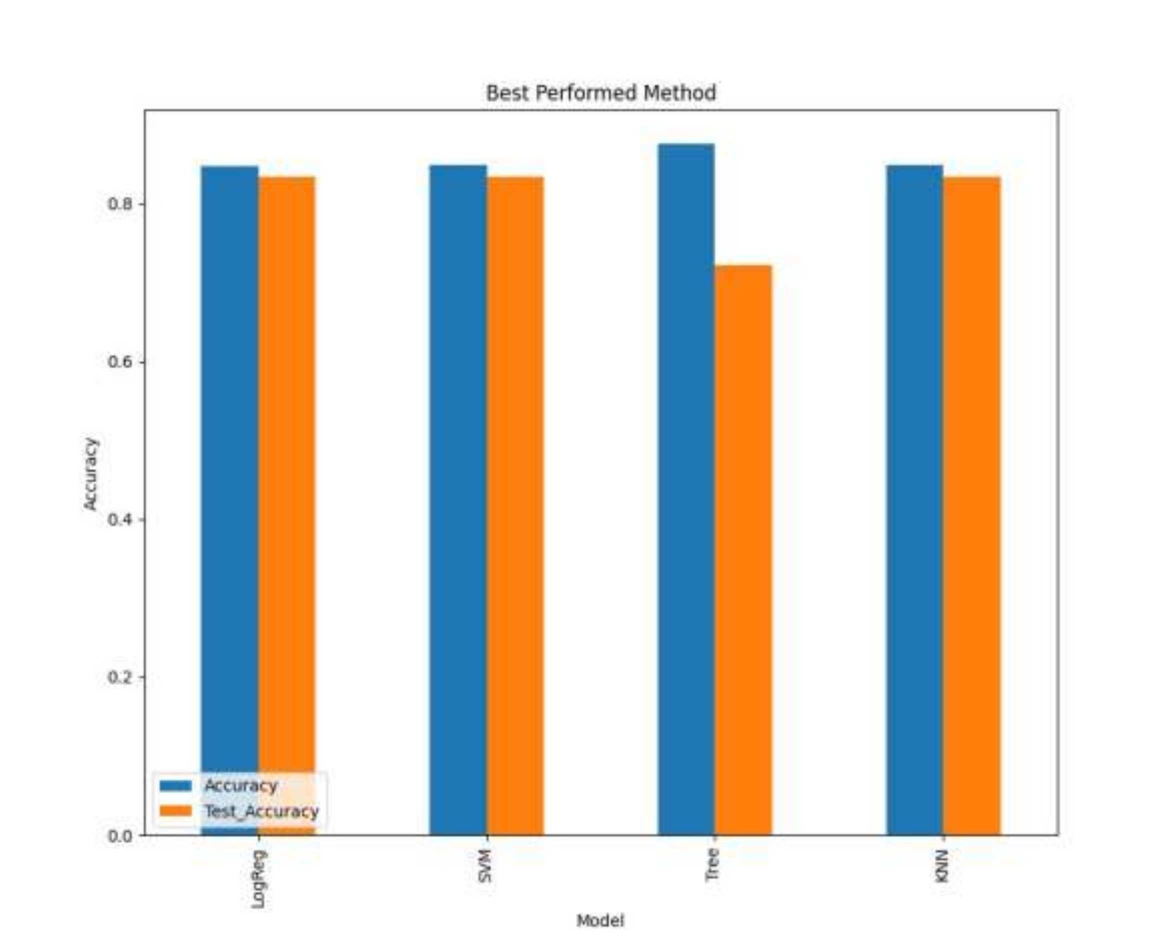
- Payloads under 5500 kg are likely to be successful for FT boosters
- Payloads under 5000 kg are likely to be unsuccessful for v1.1 boosters

Section 5

Predictive Analysis (Classification)

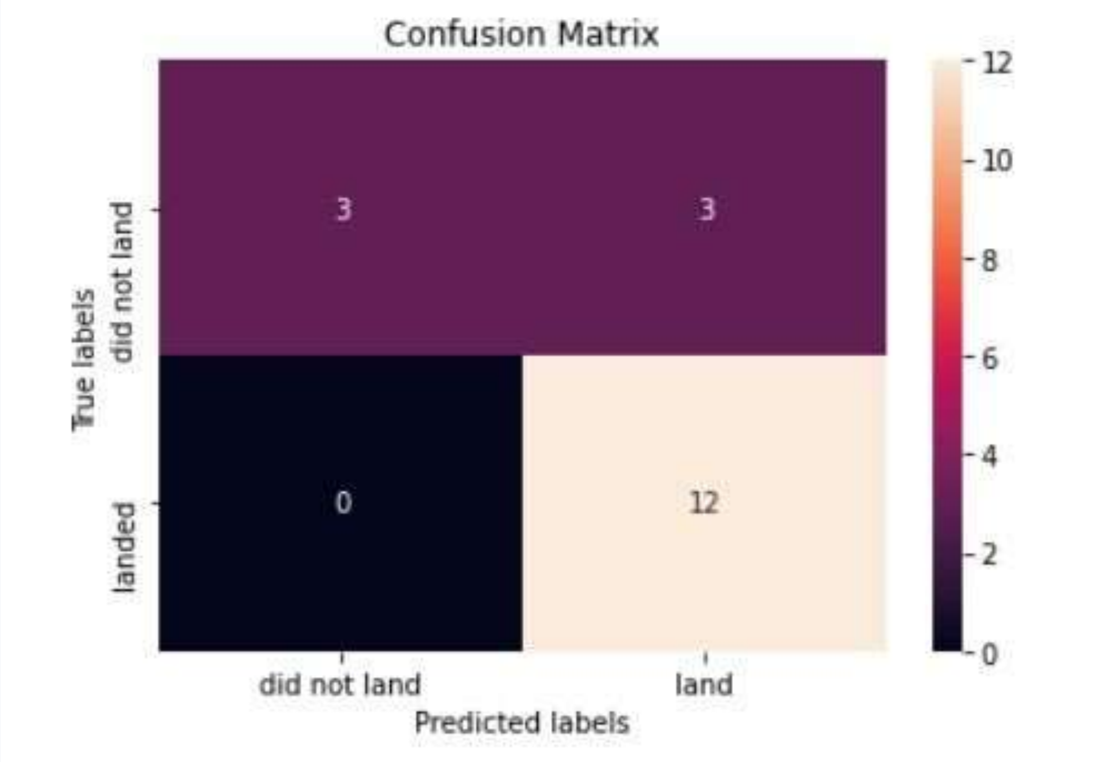
Classification Accuracy

- Four classification models were trained
- Logistic regression, SVM and k Nearest neighbors produced similar results with accuracy rate of about 83.33% on test set
- More data is needed for better determination of the best model



Confusion Matrix

- Same matrix for Logistic Regression, SVM and KNN
- Correct predictions on diagonal: top left - bottom right
- Predictions:
  - 12 true positive
  - 3 true negative
  - 3 false positive
  - 0 false negative



Conclusions

- The most successful launch site is KSC LC-39A
- Payloads over 9000 kg have high success rate
- Successful landing outcomes improved over the time
- We created three machine learning models with an accuracy of 83.33%
- They can be used to predict with a relatively high accuracy whether the first stage of the rocket will land successfully
- More data is needed for better model determination and accuracy improvement

Thank you!