Winning Space Race with Data Science

Rahul Dange
22 December 2023

Outline

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

Executive Summary

- **Summary of methodologies**
  - Data Collection
  - Data Wrangling
  - Exploratory Data analysis (EDA) with Data Visualization
  - EDA with SQL
  - EDA using Folium (graph)
  - EDA using dashboard with Plotly Dash
  - Predictive analysis
- **Summary of all results**
  - EDA results & Interactive analysis
  - Predictive analysis

Introduction

- **Project background -** Space X 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 Space X can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch.

- **Problems to find answers** – Using this project we will try to find out which all parameters are important for Successful Launch.

Section 1

Methodology

Methodology

Executive Summary

- **Data collection methodology**:

  a) SpaceX Rest API b) Data available on Wikipedia using web scrapping

- **Perform data wrangling**

  a) Data cleaning by removing null values

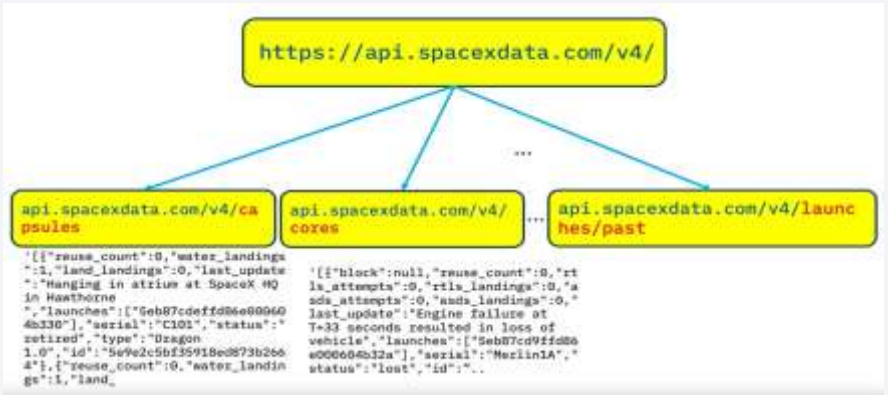
  b) One hot encoding for Machine learning

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

  Different Machine learning models are built like Logistic Regression, Support Vector Machine (SVM), Decision tree (DT), K Nearest Neighbors (KNN) and evaluate these models for correct prediction.

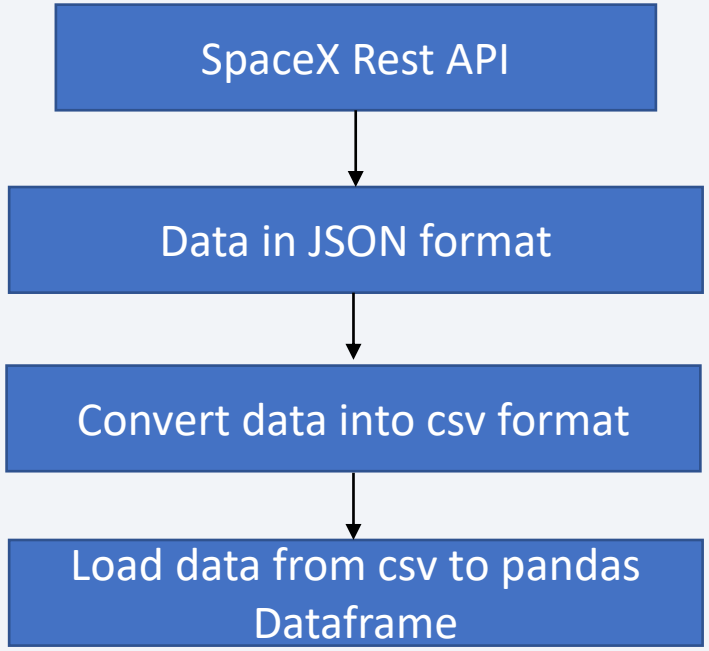
Data Collection

- Data is collected using SpaceX rest API and using data wrangling from Wikipedia (Flowchart on next slide)

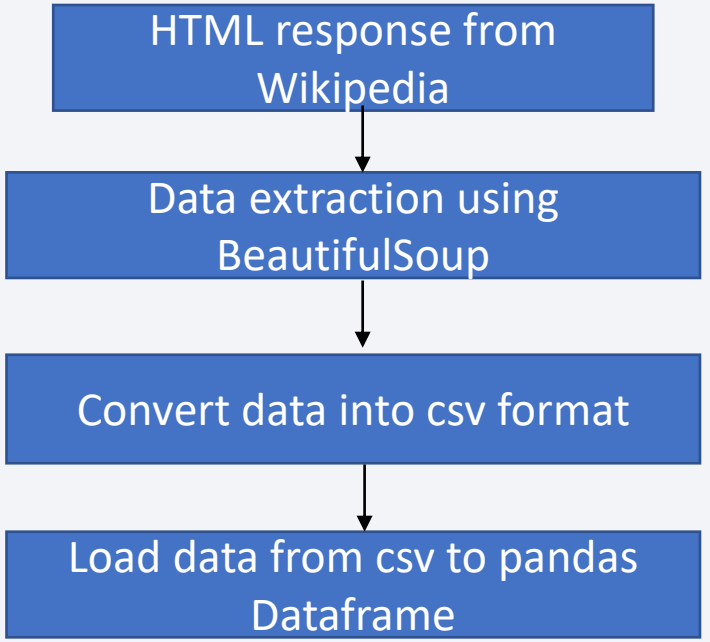


Data Collection – Flowchart of process

- As shown in previous slide data is collected using SpaceX rest API and using data wrangling from Wikipedia. Below are the process used to prepare data for processing.
- Data from SpaceX rest API



Data from Wikipedia



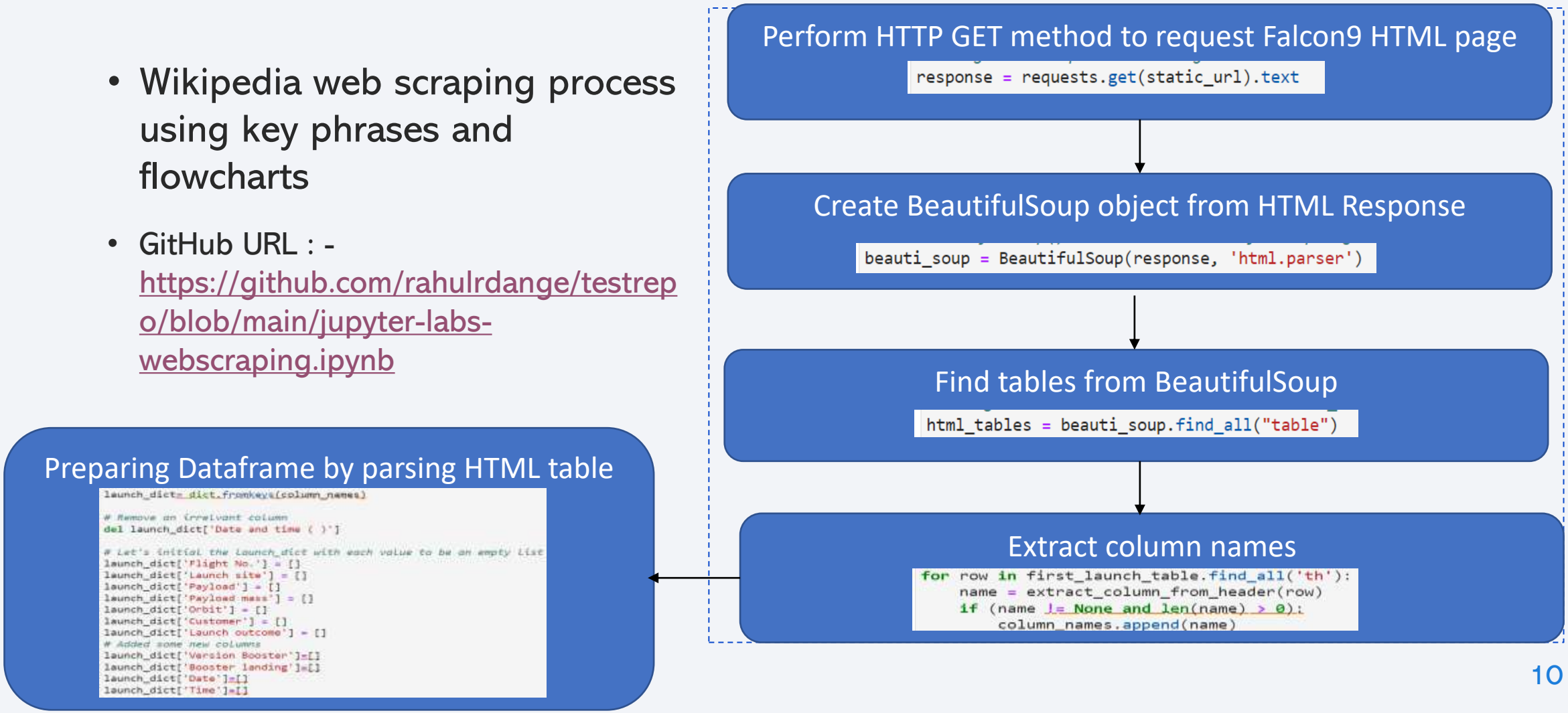
Data Collection – SpaceX API

- Data from SpaceX REST calls using key phrases and flowcharts
- These steps are used to prepare data so that it can be used for prediction

- **GitHub link** - https://github.com/rahulrdange/testrepo/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

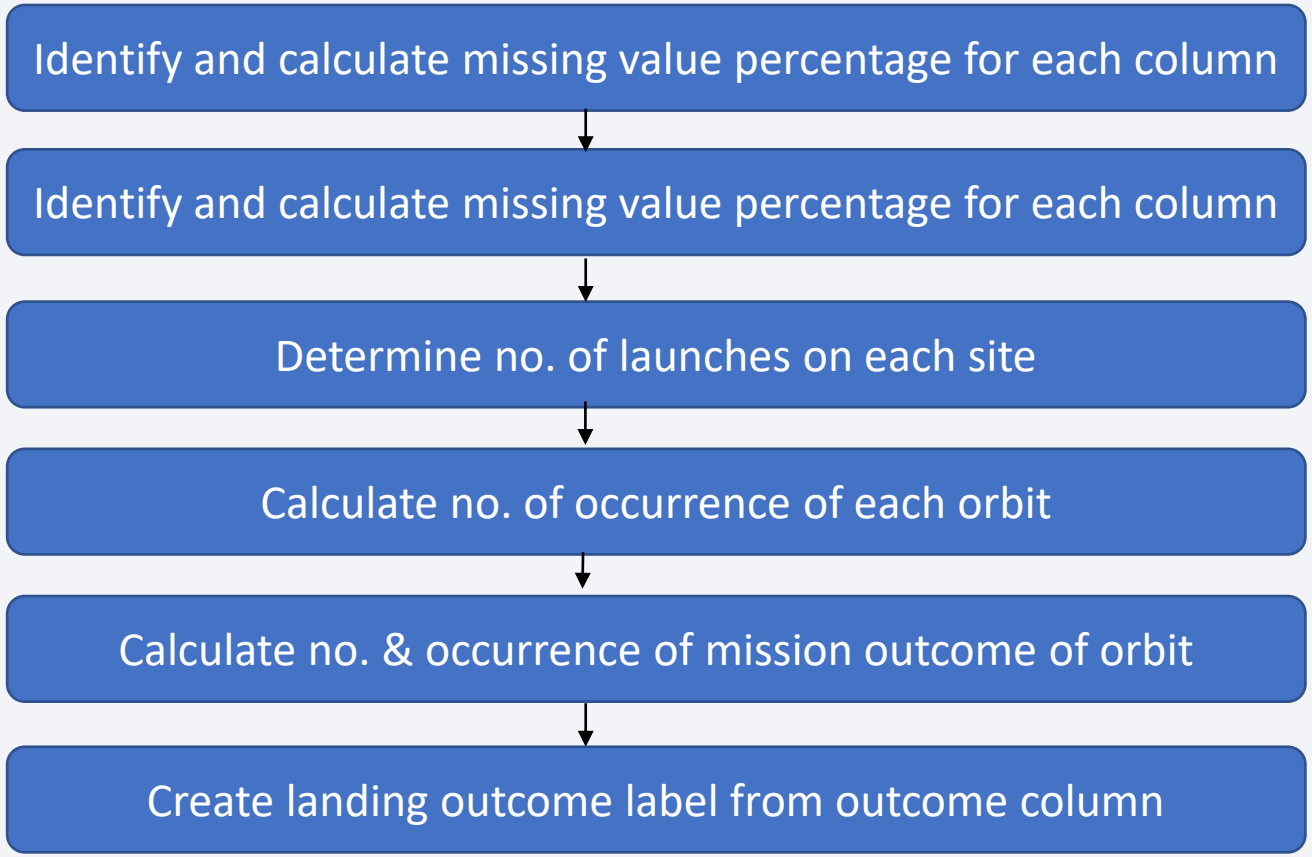
Data Collection - Scraping

- Wikipedia web scraping process using key phrases and flowcharts
- GitHub URL : - https://github.com/rahulrdange/testrepo/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- Steps to perform Data Wrangling

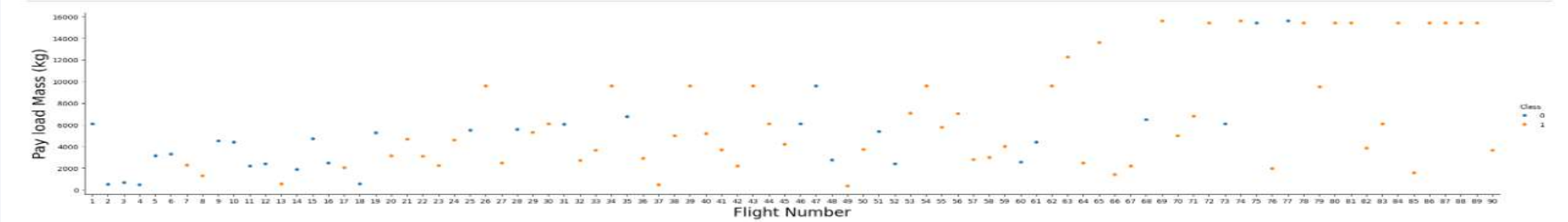


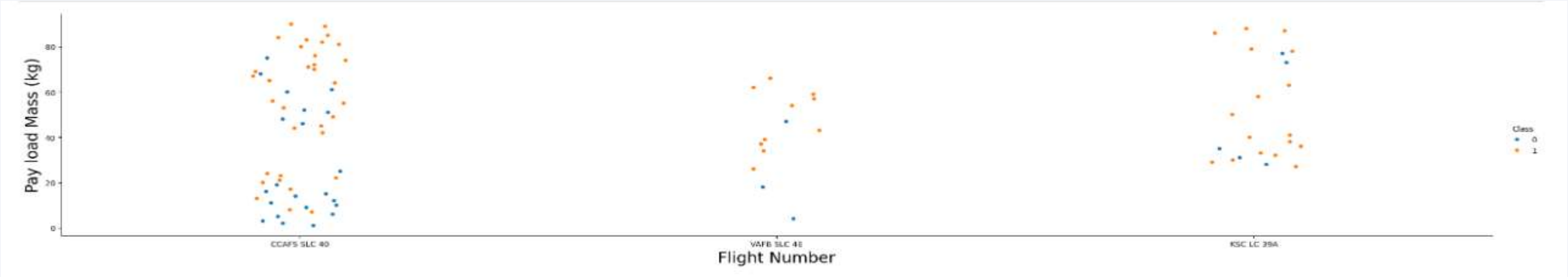
- Github URL : https://github.com/rahulrdange/testrepo/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization (Slide-1)

- Below mentioned charts are plotted

To see how the Flight Number and payload would affect launch outcome





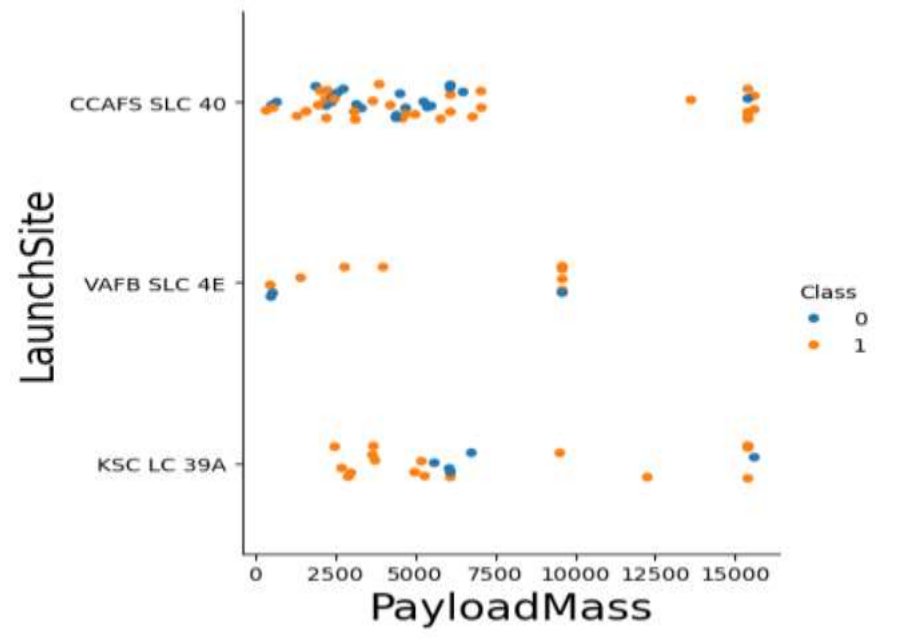
Continue…

GitHub URL: https://github.com/rahulrdange/testrepo/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

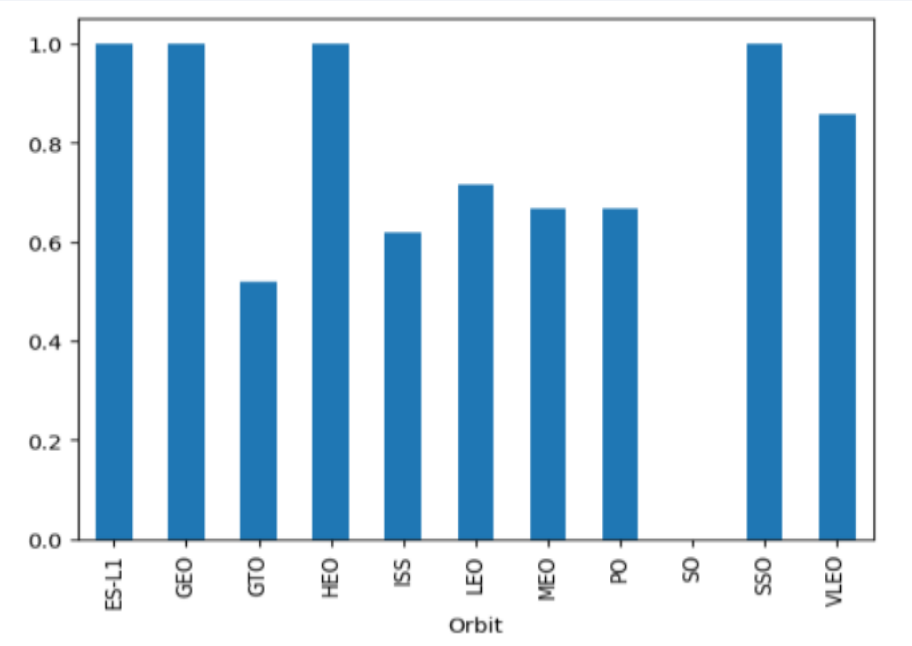
EDA with Data Visualization (Slide-2)

- Below mentioned charts are plotted

Is any relationship between launch site and payload mass



Is any relationship between success rate and orbit type



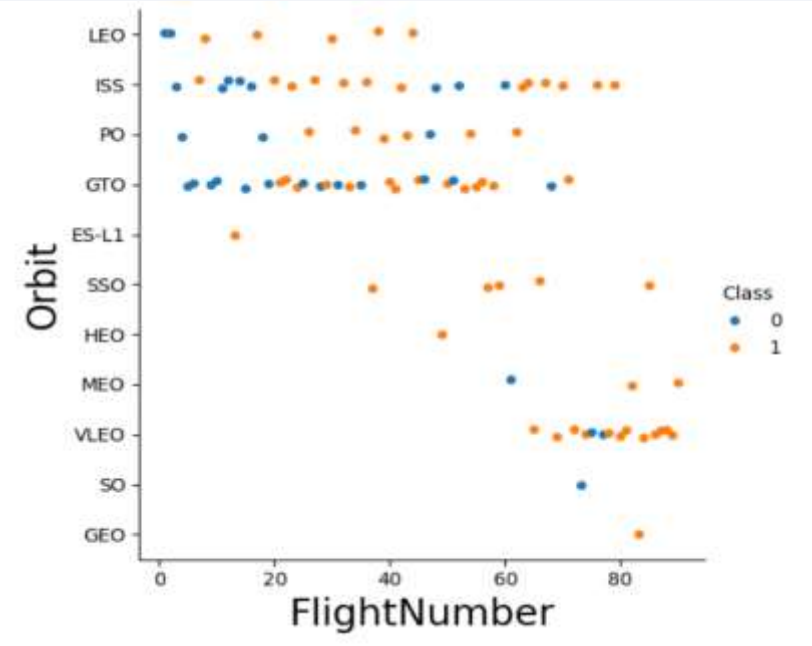
Continue…

GitHub URL: https://github.com/rahulrdange/testrepo/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

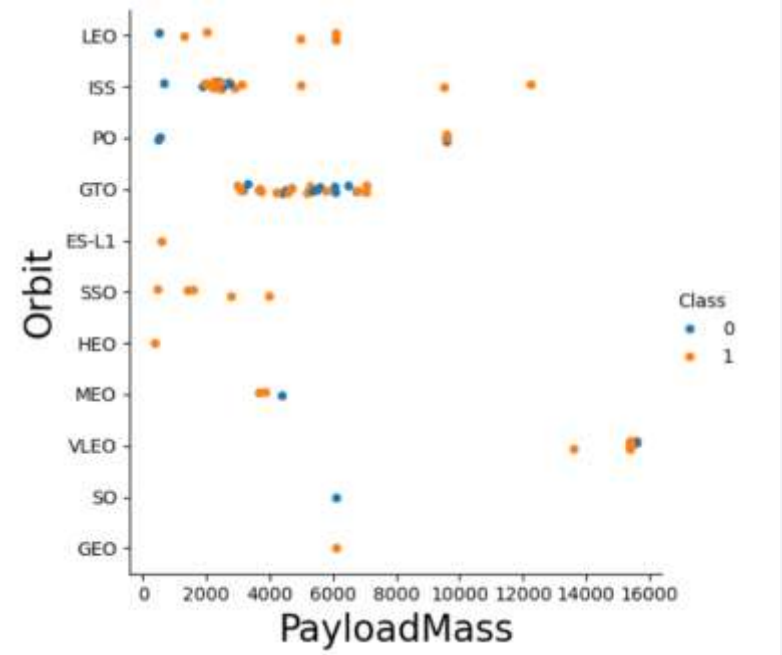
EDA with Data Visualization (Slide-3)

- Below mentioned charts are plotted

Is any relationship between flight no. and orbit type



Is any relationship between payload and orbit type



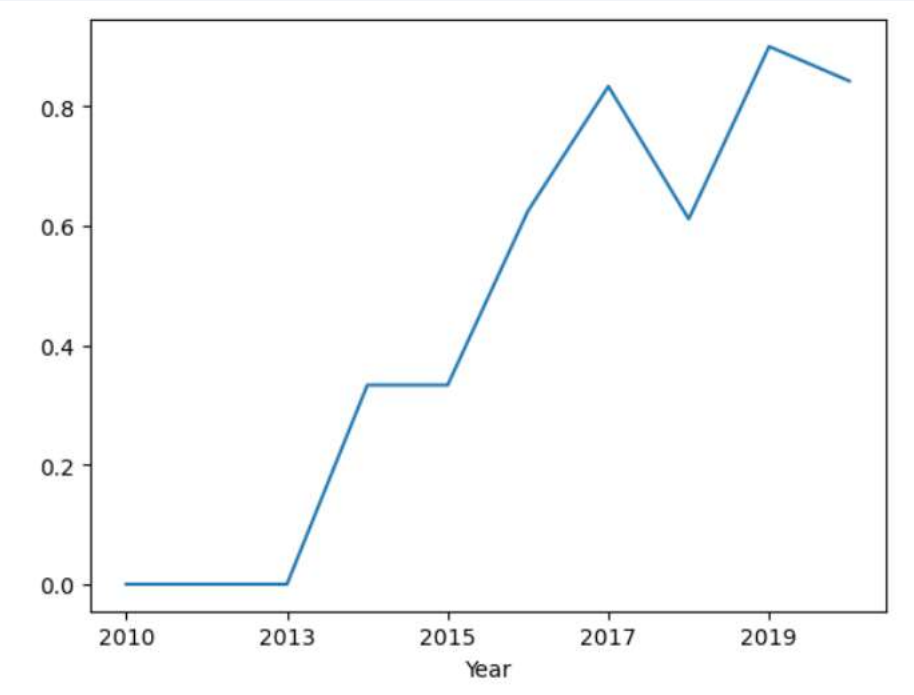
Continue…

GitHub URL: https://github.com/rahulrdange/testrepo/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with Data Visualization (Slide-4)

- Below mentioned charts are plotted

  To get average launch success trend (Success per year)



GitHub URL: https://github.com/rahulrdange/testrepo/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- Below SQL's are executed
  1. Finding the names of unique launch sites in the space mission.
  2. Display 5 records where launch site begin with string 'CCA'.
  3. Display the payload mass carried by boosters 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.
  9. List the records which will display the month 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.

- Github URL: https://github.com/rahulrdange/testrepo/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

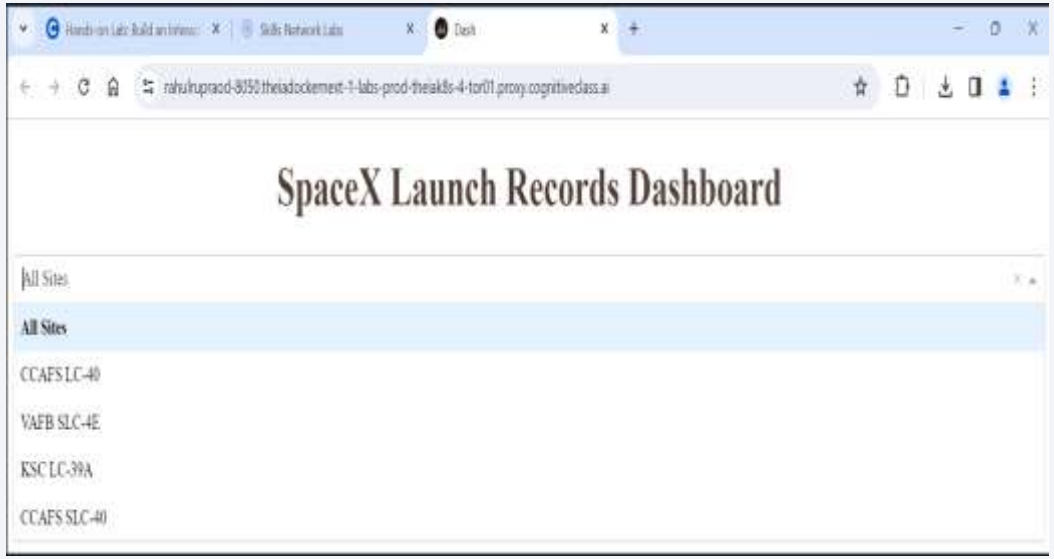
- Map markers have been added to map to find the optimal location for successful launch

- Gitbub URL: https://github.com/rahulrdange/testrepo/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash (Slide-1)

- Below plots/graphs added to a dashboard

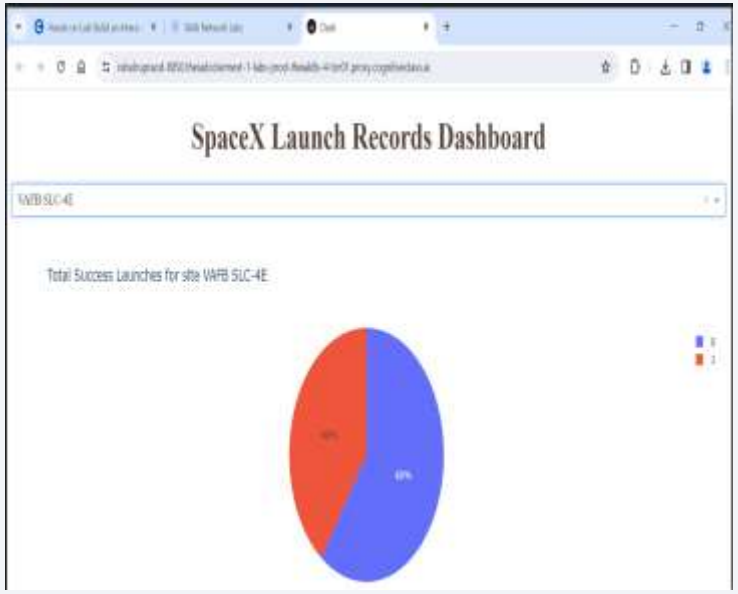
  Dashboard created with dropdown for different launch sites so that we can analyze success rate for all sites as well as each site.



- GitHub URL: https://github.com/rahulrdange/testrepo/blob/main/dash_interactivity.py

Build a Dashboard with Plotly Dash (Slide-2)

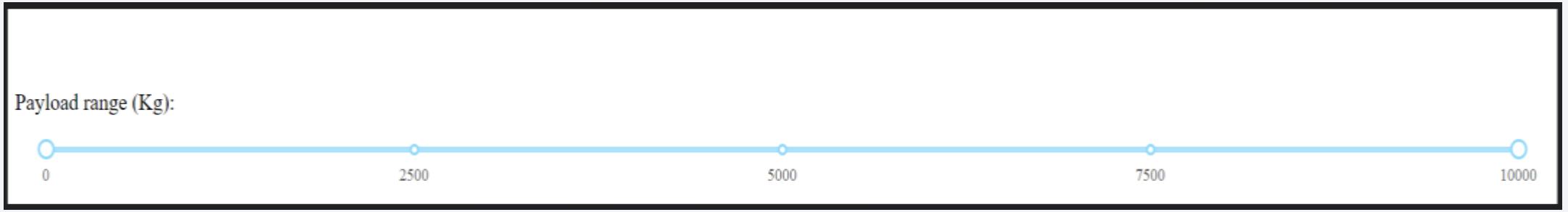
- Dashboard created with dropdown for different launch sites so that we can analyze success rate for each site

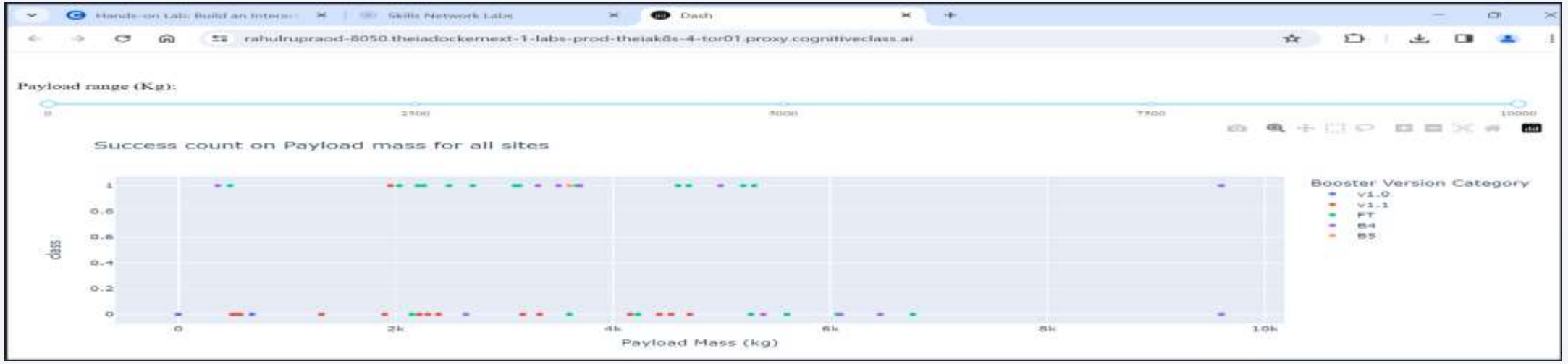


- GitHub URL: https://github.com/rahulrdange/testrepo/blob/main/dash_interactivity.py

Build a Dashboard with Plotly Dash (Slide-3)

- Dashboard created for payload range with slider success count on payload mass for all site



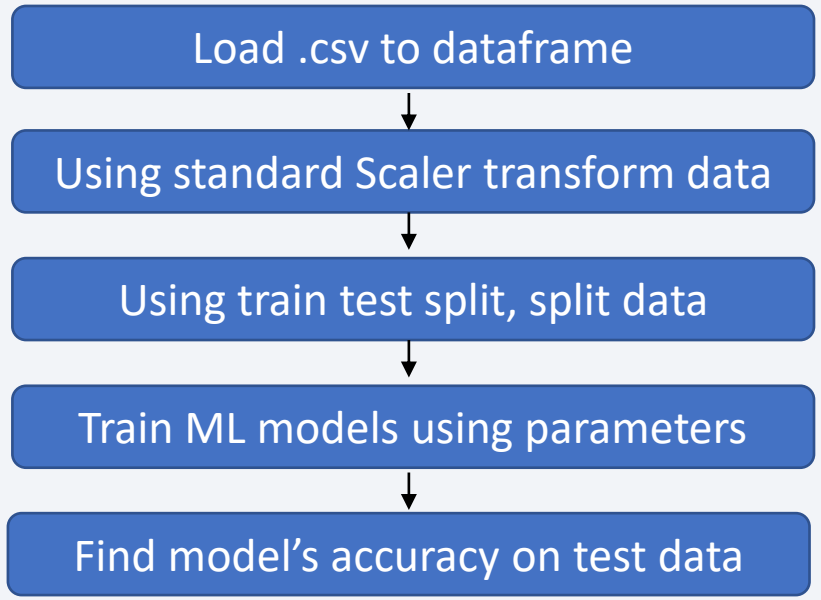


- GitHub URL: https://github.com/rahulrdange/testrepo/blob/main/dash_interactivity.py

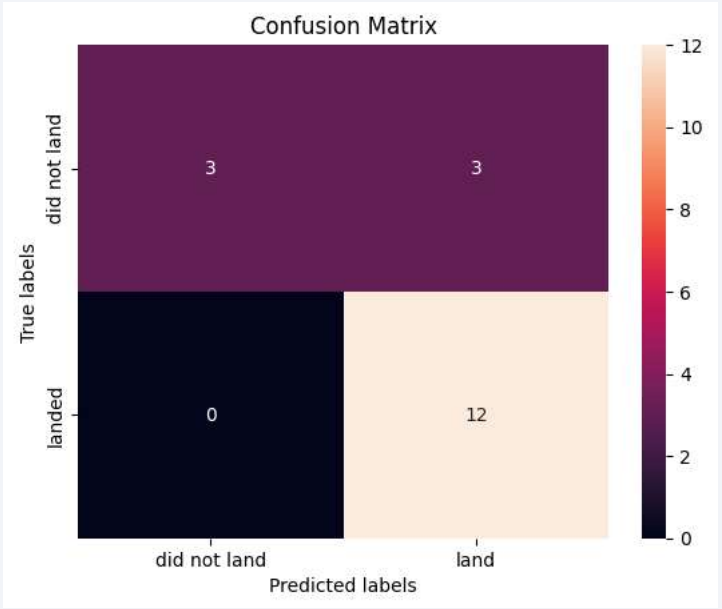
Predictive Analysis (Classification)

- We built different models like Logistic Regression, Support Vector Machine, Decision Tree, K Nearest Neighbor.
- All models almost work similar, like for training 84% accuracy and for testing 83% accuracy.
- Decision tree performs well on training with 86% accuracy while on testing 83% accuracy.

Flowchart



confusion matrix for models



- GitHub URL : https://github.com/rahulrdange/testrepo/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

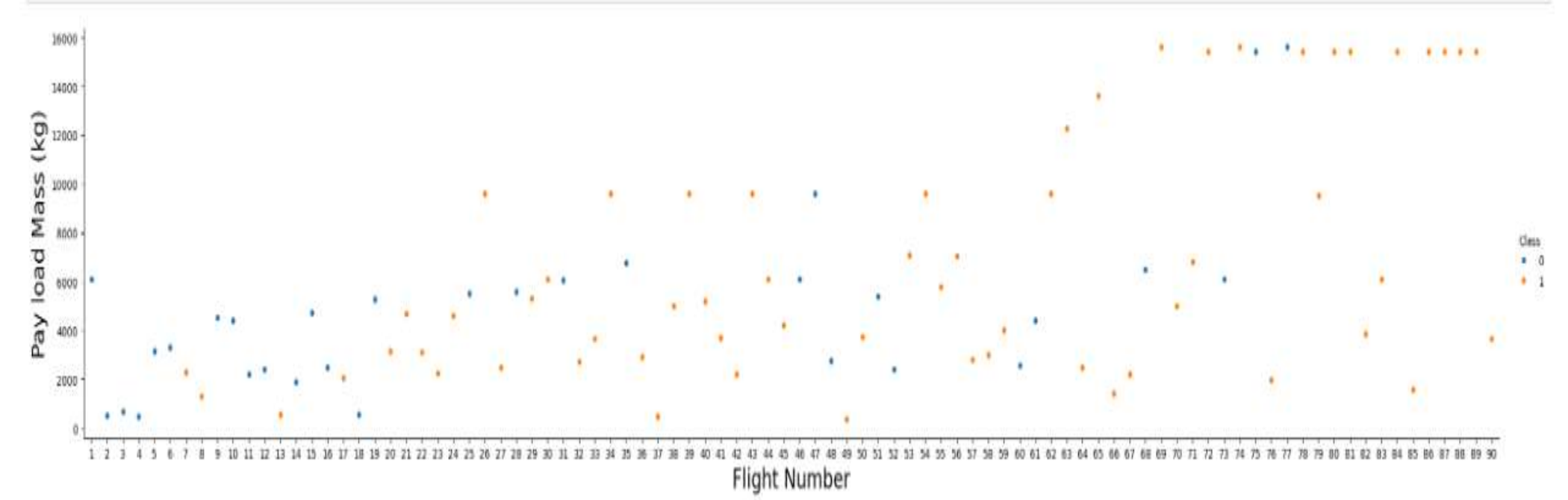
- Success rate increased in recent years.
- Success rate is more for launch site KSC LC-39A which is 41.7%.
- Success rate is more for low weighted payload than high weighted payload
- All the models perform similar for test data with 83% accuracy (For training data DT perform best with 86% accuracy).

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

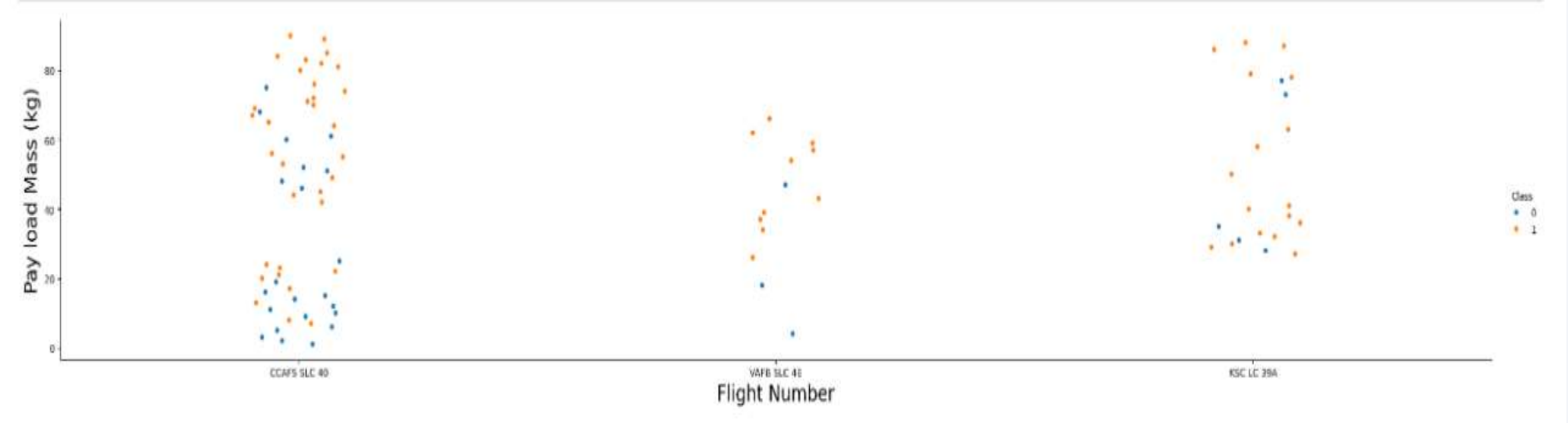
- Scatter plot of Flight Number vs. Launch Site



- Launches from launch site CCAFS SLC 40 is higher than other launch site

Payload vs. Launch Site

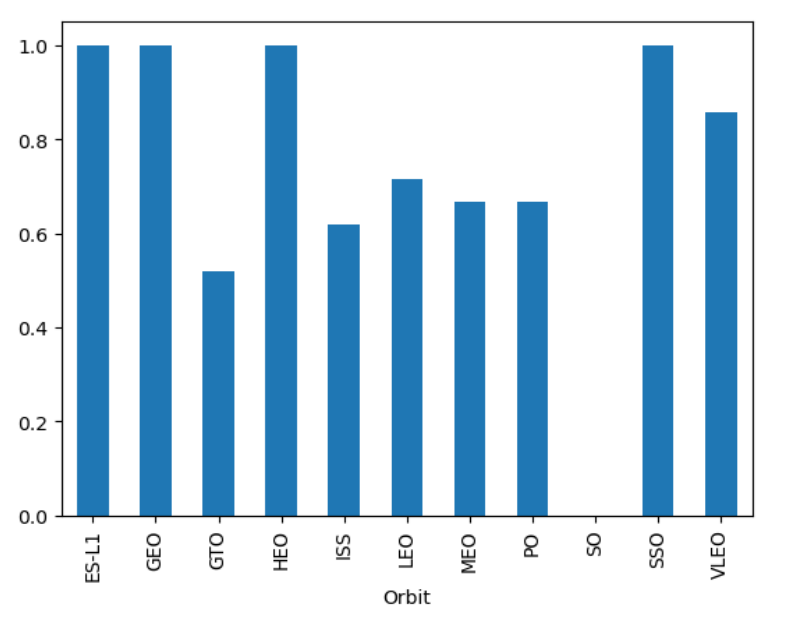
- Scatter plot of Payload vs. Launch Site



- Low mass payloads have been launched more from CCAFS SLC 40 than other launch sites.

Success Rate vs. Orbit Type

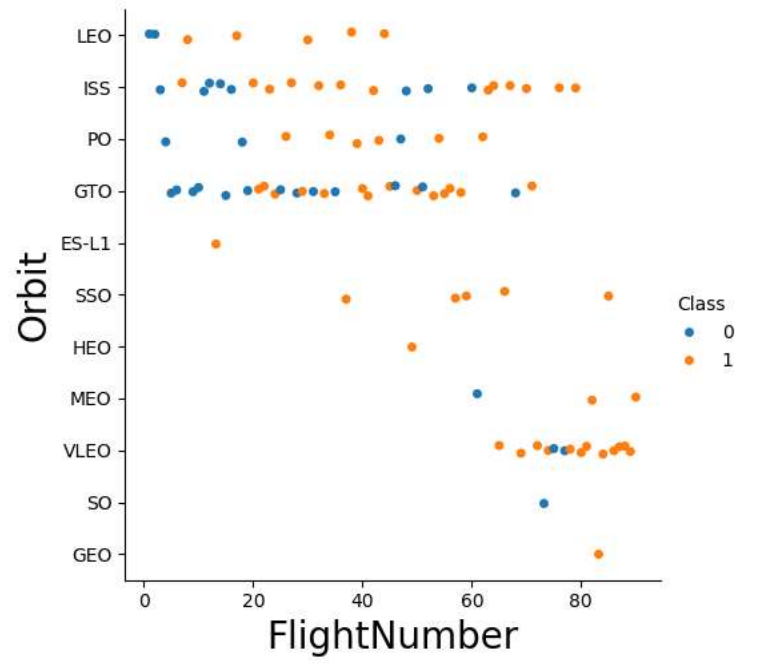
- Bar chart for the success rate of each orbit type



- Success rate is higher for ES-L1, GEO, HEO, SSO orbits.

Flight Number vs. Orbit Type

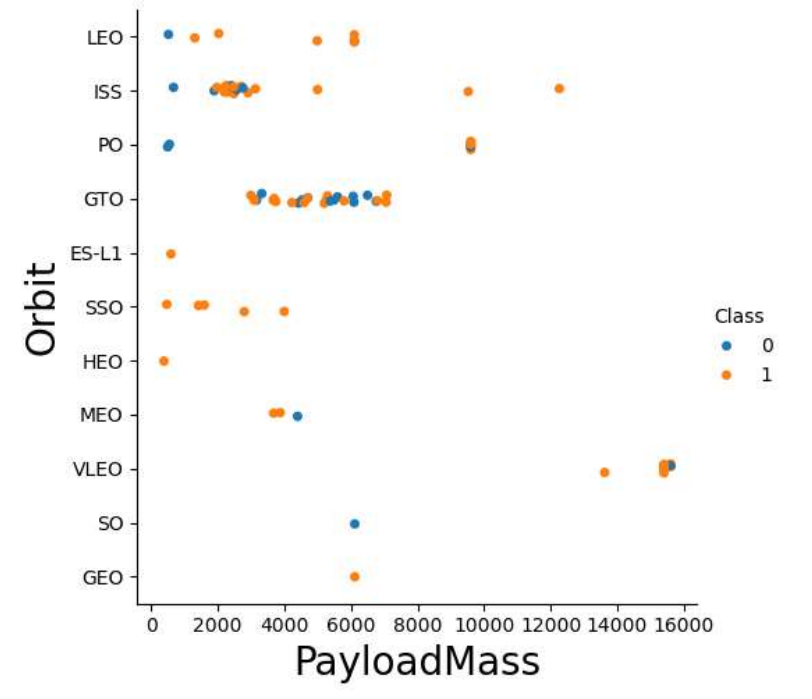
- Scatter point of Flight number vs. Orbit type



- In recent years there are more launches in GTO,PO, LEO orbits

Payload vs. Orbit Type

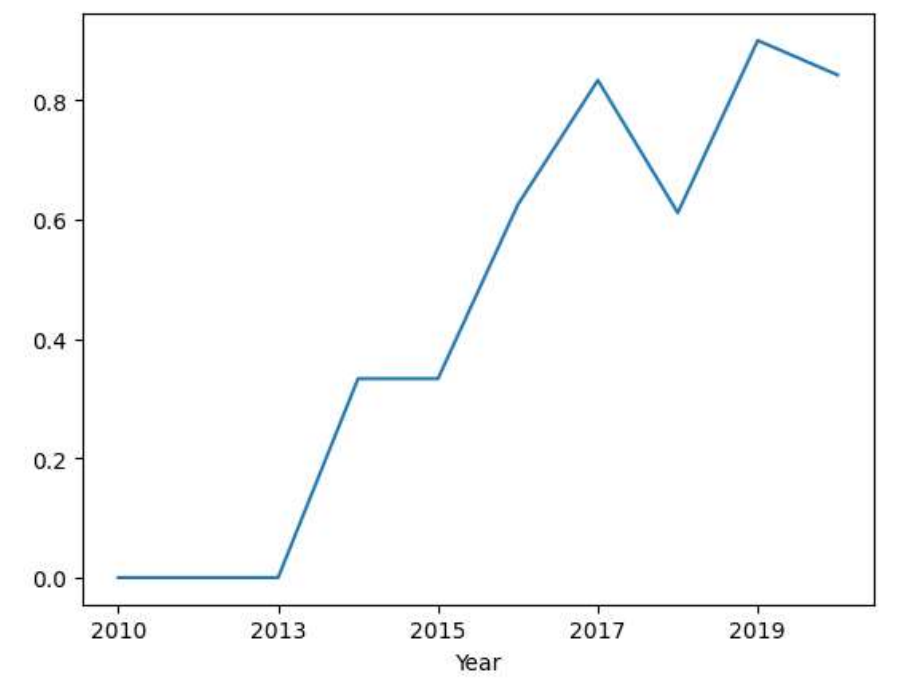
- Scatter point of payload vs. orbit type



- There is strong correlation between ISS orbit and 2000 Payload mass. Similarly for GTO orbit and 3000-7000 payload mass.

Launch Success Yearly Trend

- Line chart of yearly average success rate



- Success rate has significantly increased from 2013. We can see downfall in 2017-18.

All Launch Site Names

- Find the names of the unique launch sites

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

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

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

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

- There are 4 different Launch sites for SpaceX, which are CCAFS LC -40, VAFB SLC -4E, KSC LC-39A and CCAFS SLC-40.

Launch Site Names Begin with 'CCA'

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

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

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

```
 * sqlite:///my_data1.db
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)	NASA (COTS) NRO	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

- Above mentioned are 5 records where Launch site begin with ‘CCA’.

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

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

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)'
```

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

SUM(PAYLOAD_MASS__KG_)
45596

- Total 45596 kg of payload mass carried by busters launched by NASA (CRS).

Average Payload Mass by F9 v1.1

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

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

AVG(PAYLOAD_MASS__KG_)
2928.4

- Average 2928.4 kg of payload mass carried by booster version F9 v1.1.

First Successful Ground Landing Date

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

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

Hint:Use min function

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

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

MIN(DATE)
2015-12-22

- On 22 December 2015 first successful landing was done on ground pad.

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

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 (PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000)
```

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

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

- Four booster versions F9 FT B1022, F9 FT B1026, F9 FT B1021.2, F9 FT b1031.2 which have successfully landed on dron ship and had payload mass greater than 4000 but less than 6000.

Total Number of Successful and Failure Mission Outcomes

- Calculate the 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_NUMBER \
FROM SPACEXTBL GROUP BY MISSION_OUTCOME_
```

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

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

- There are total 100 successful and 1 failure mission outcome.

Boosters Carried Maximum Payload

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

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

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

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

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

- Above mentioned 12 Booster Versions carried the maximum payload mass.

2015 Launch Records

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

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
%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,0,5) = '2015'
```

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

- Two Booster version F9 v1.1 B1012 & F9 v1.1 B1015 along with launch site and month in 2015 which has failed landing outcome in drone ship.

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

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 DISTINCT(LANDING_OUTCOME), COUNT(LANDING_OUTCOME) AS CNT FROM SPACEXTBL\
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING_OUTCOME ORDER BY CNT DESC
```

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

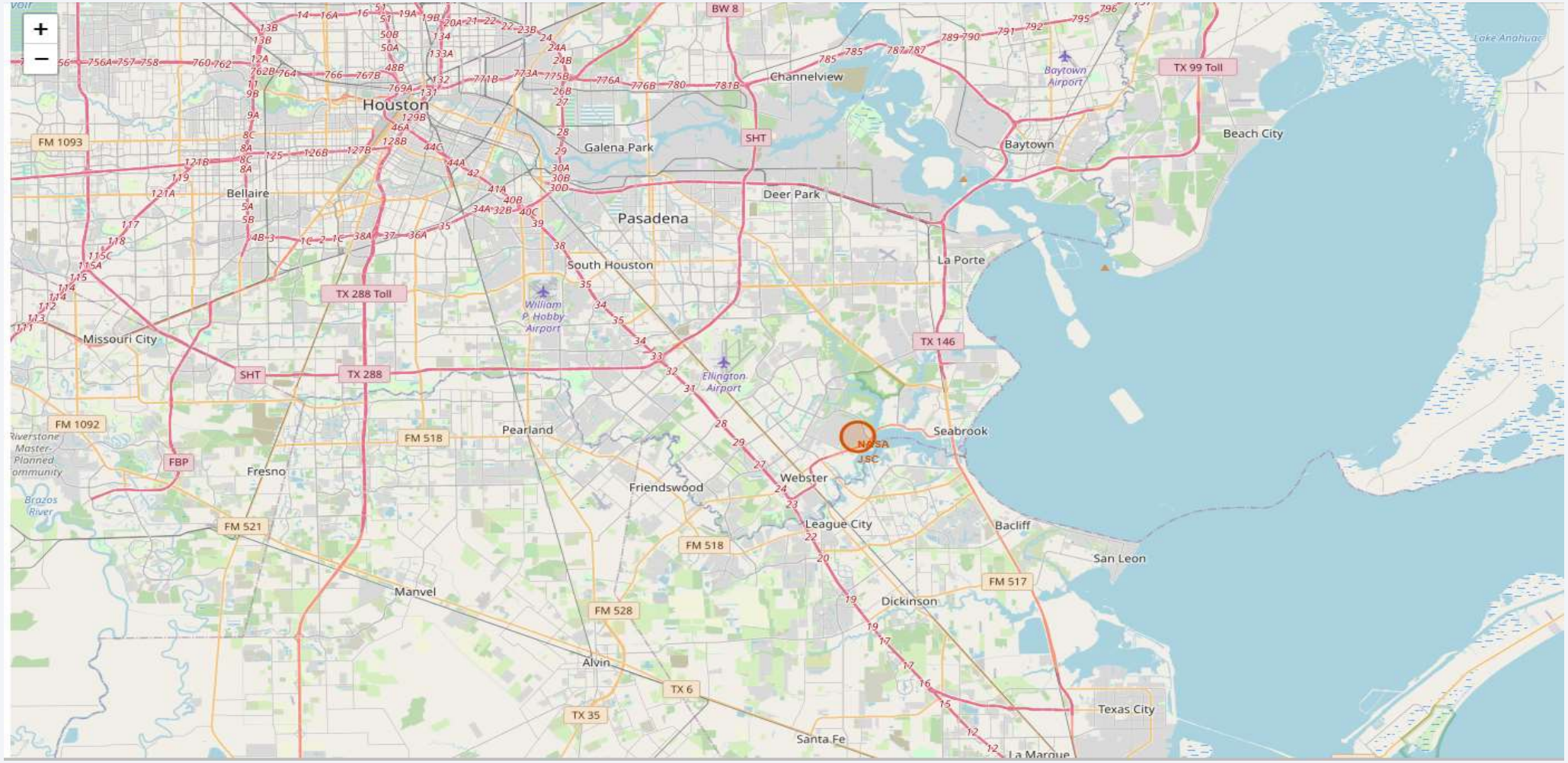
Landing_Outcome	CNT
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

- Above mentioned are count of landing outcome between date 2010-06-04 and 2017-03-20 in descending order.

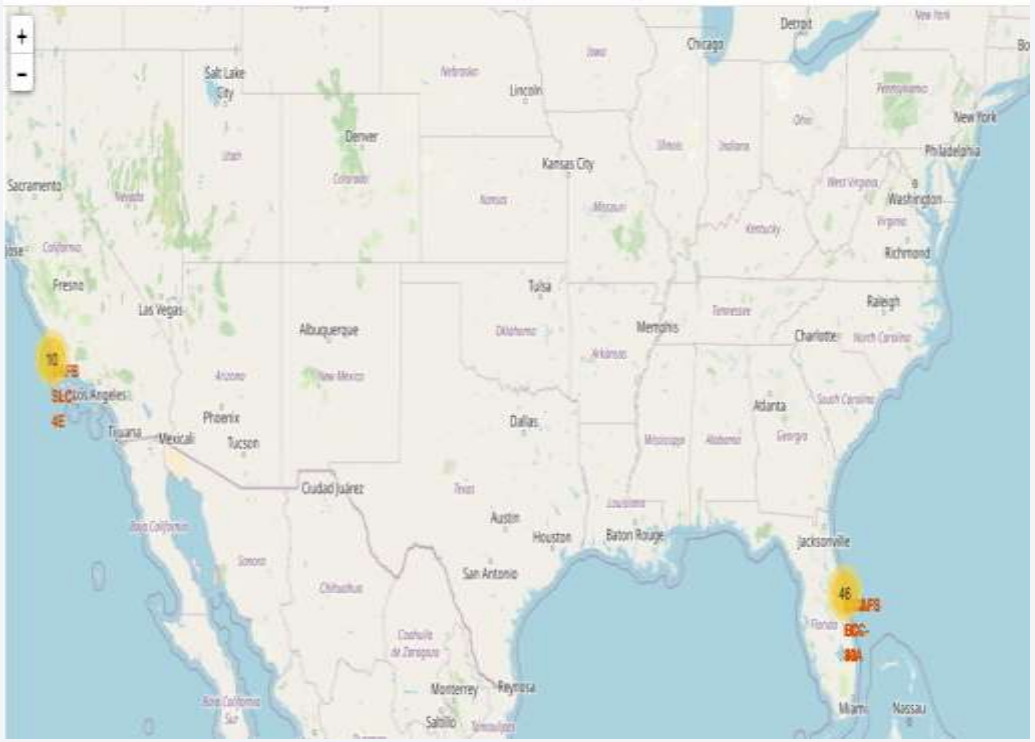
Section 3

Launch Sites Proximities Analysis

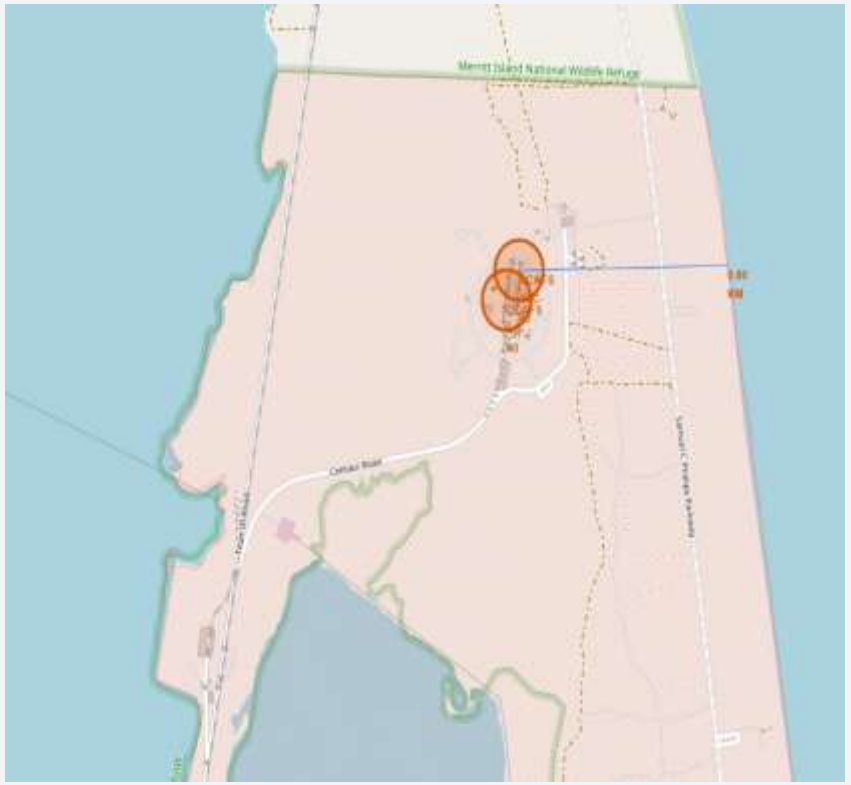
All launch sites



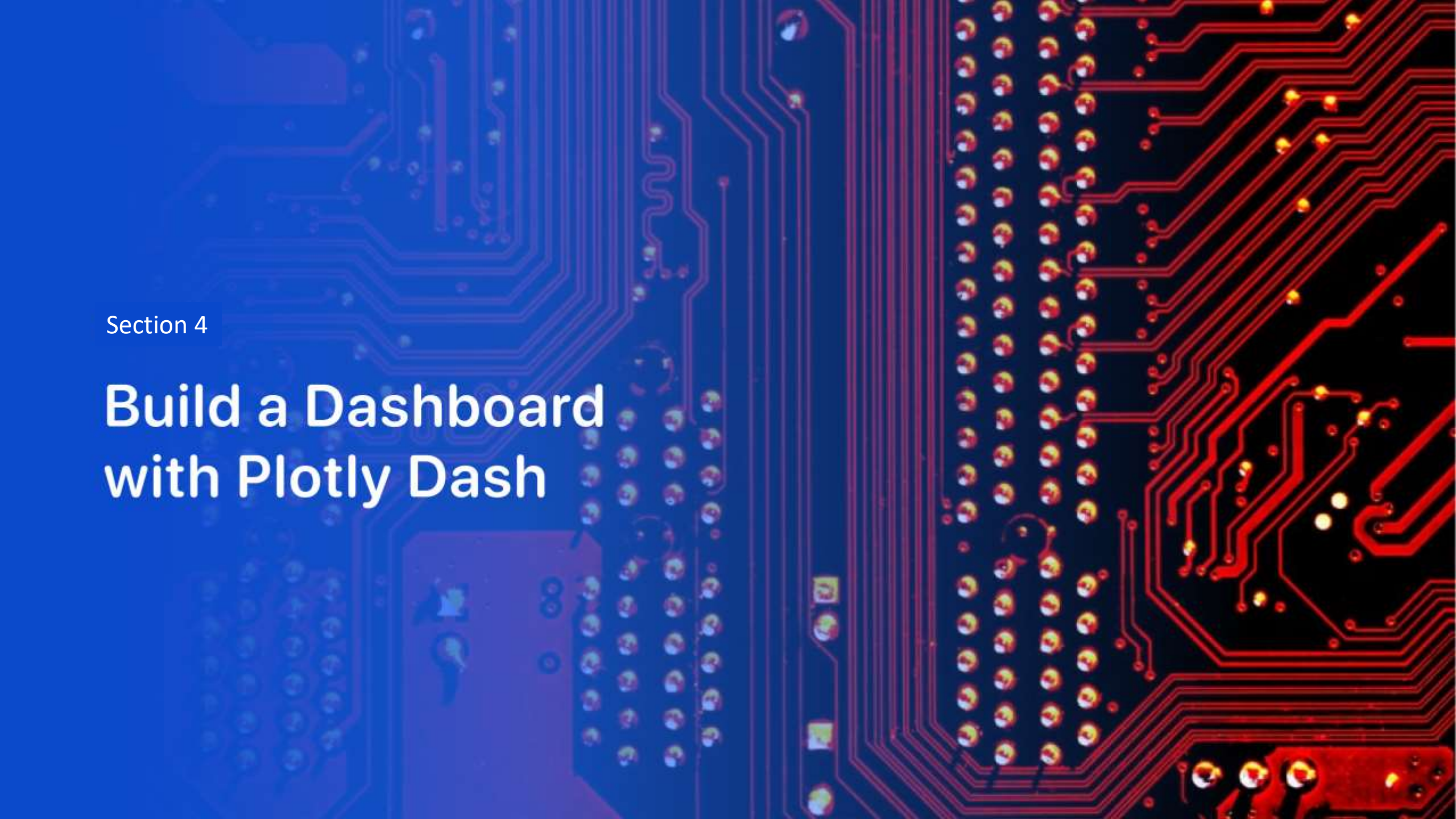
Launch outcomes(Success/Failure)



Distance between launch site and coastline



- As we can see in map coastline is 0.86 km from CCAFS SLC -40


Section 4
Build a Dashboard with Plotly Dash

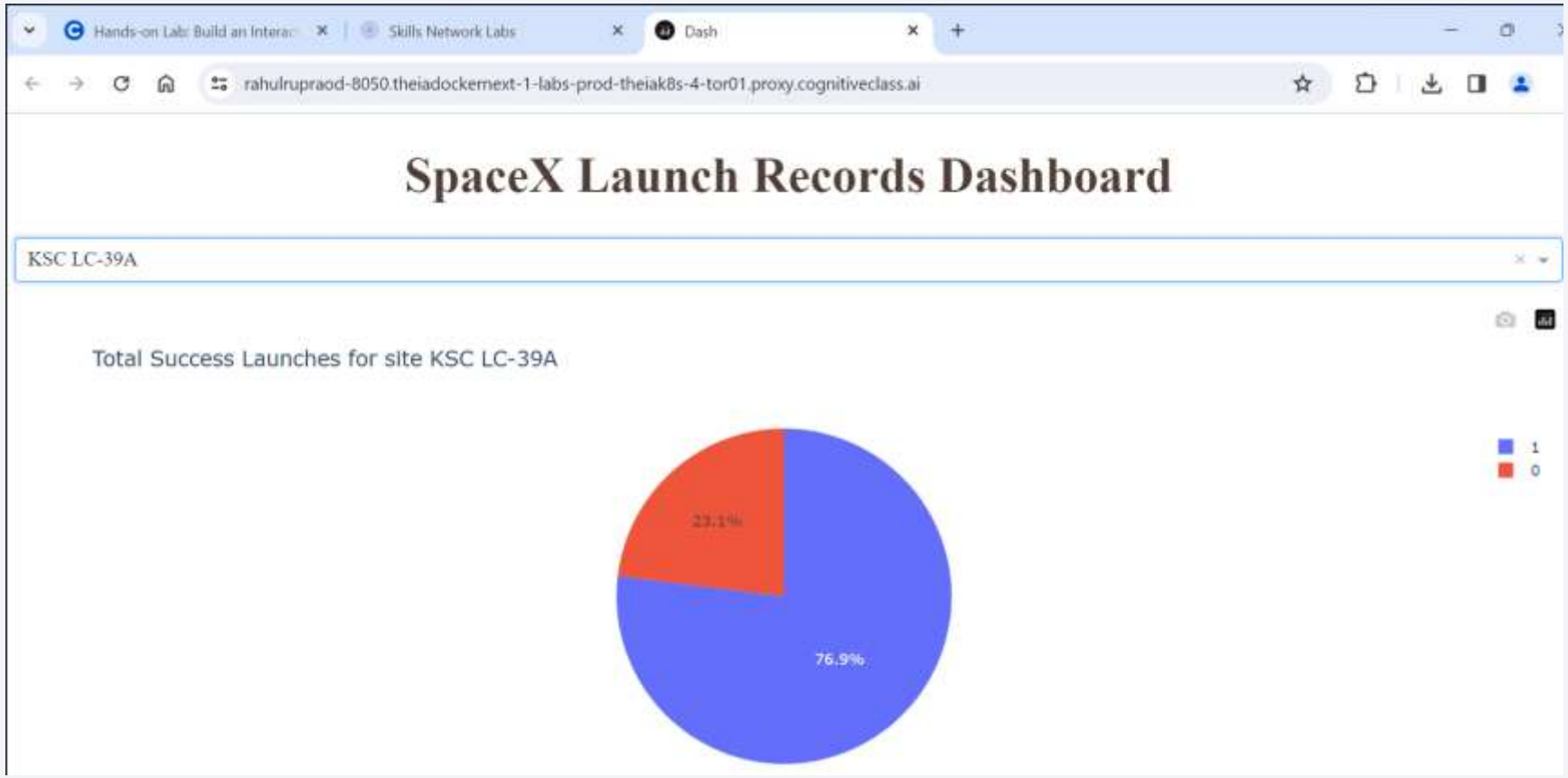
Launch success count for all sites

- Pie chart of launch success count for all sites.
  - We can see success count of KSC LC-39A is highest (41.7%) compared to other sites.

Highest launch success ratio

- Pie chart for the launch site with highest launch success ratio

  We can see success ration is 76.9% & failure rate is 23.1% for KSC LC-39A launch site.



Payload vs. Launch Outcome

- Scatter plot of Payload vs. Launch Outcome for all sites, with different payload selected in the range slider.
  - Success rate is higher for Low Payload Mass than High Payload Mass.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Logistic Regression

```
logreg_cv.score(X_test, Y_test)
```

```
0.8333333333333334
```

Support Vector Machine (SVM)

```
svm_cv.score(X_test, Y_test)
```

```
0.8333333333333334
```

- Decision tree (DT)

```
tree_cv.score(X_test, Y_test)
```

```
0.8333333333333334
```

KNN

```
knn_cv.score(X_test, Y_test)
```

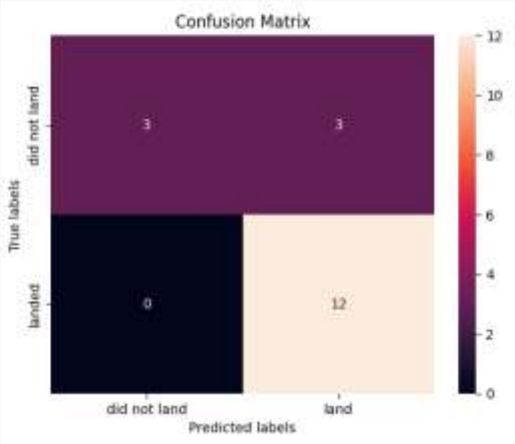
```
0.8333333333333334
```

- All models perform same on testing data with accuracy of 83.33%.

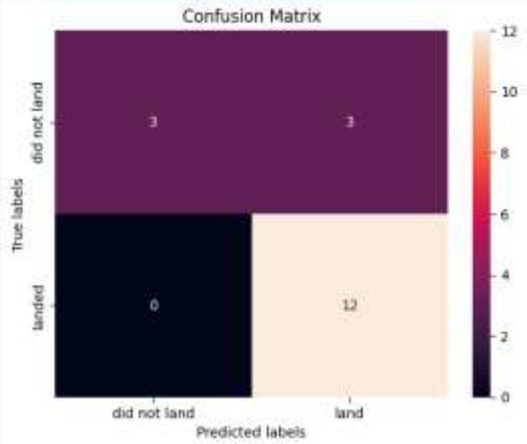
Confusion Matrix

- All models perform same on testing data with the given parameters, gives same accuracy(83.33%), confusion matrix are same.

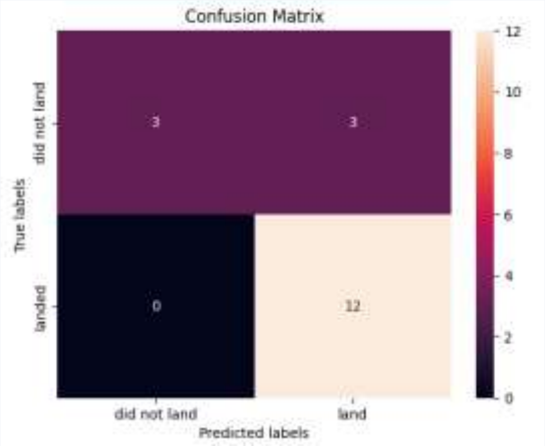
Logistic regression



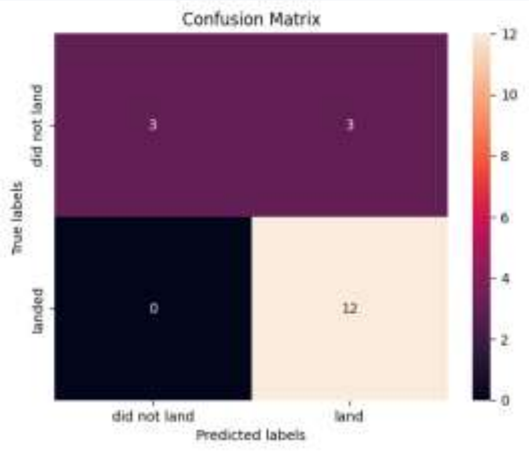
SVM



DT



KNN



Conclusions

- Decision tree is best for training data and gives 86% accuracy. However, all models LR, SVM, DT, KNN perform same on testing data with 83.33% accuracy.
- Low payload mass launches are more successful than high payload mass.
- Launches are more successful in recent years.
- Launch site KSC LC 39A have most successful launches as compared to other launch sites.
- GEO, HEO, SSO and ES L1 orbits has best success rates than other orbits.

Appendix

- Data Collection sites

SpaceX REST API : \

https://api.spacexdata.com/v4/launches/past

Wikipedia :

https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

Thank you!