IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sergey Aityan
May 28, 2023

Outline

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

Executive Summary (1/2)

Summary of methodologies

- Programming language: Python
- SQL
- IDE: Jupyter Notebook
- File Repository: GitHUB

Summary of methodologies

- Exploratory data analysis
- Reading data from the internet and local files
- Data scrapping, wrangling, and visualization
- Conversion between csv files and database
- Data manipulation using SQL
- Interactive map with Folium
- Dashboards
- Machine Learning and predictive analysis

Executive Summary (2/2)

Summary of all results

- For the educational purpose, the Falcon 9 flights and launching sites were analyzed.
- Predictions were made for successful flights.

Introduction

Project background and context

- SpaceX has introduced a new rocket type Falcon for space exploration.
- Falcon 9 is latest version of Falcon series of rockets.
- Falcon 9 is powerful and cost-effective rocket.
- Information on Falcon launches is publicly available on the SpaceX website.

Problems Addressed

- Falcon 9 flight features like payload, successful and failed flights and landings.
- Information about launching sites and distribution of successful and failed flights over the launching sites.

Section 1

Methodology

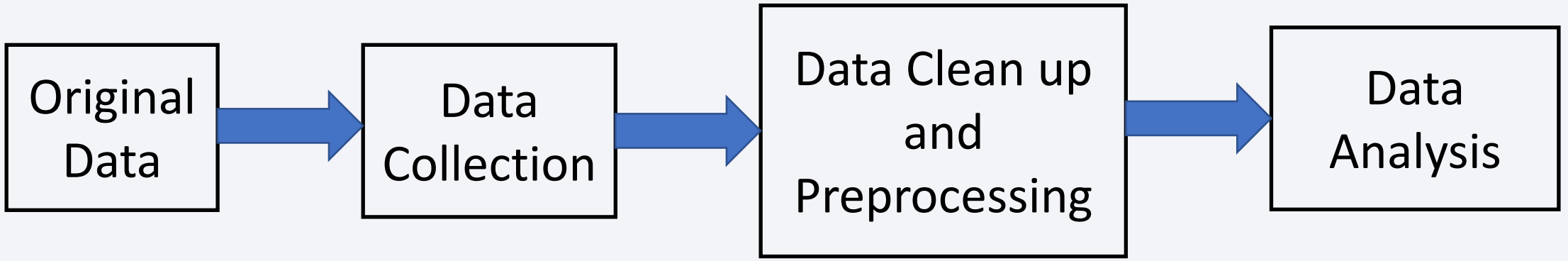
Methodology

Project Scope

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

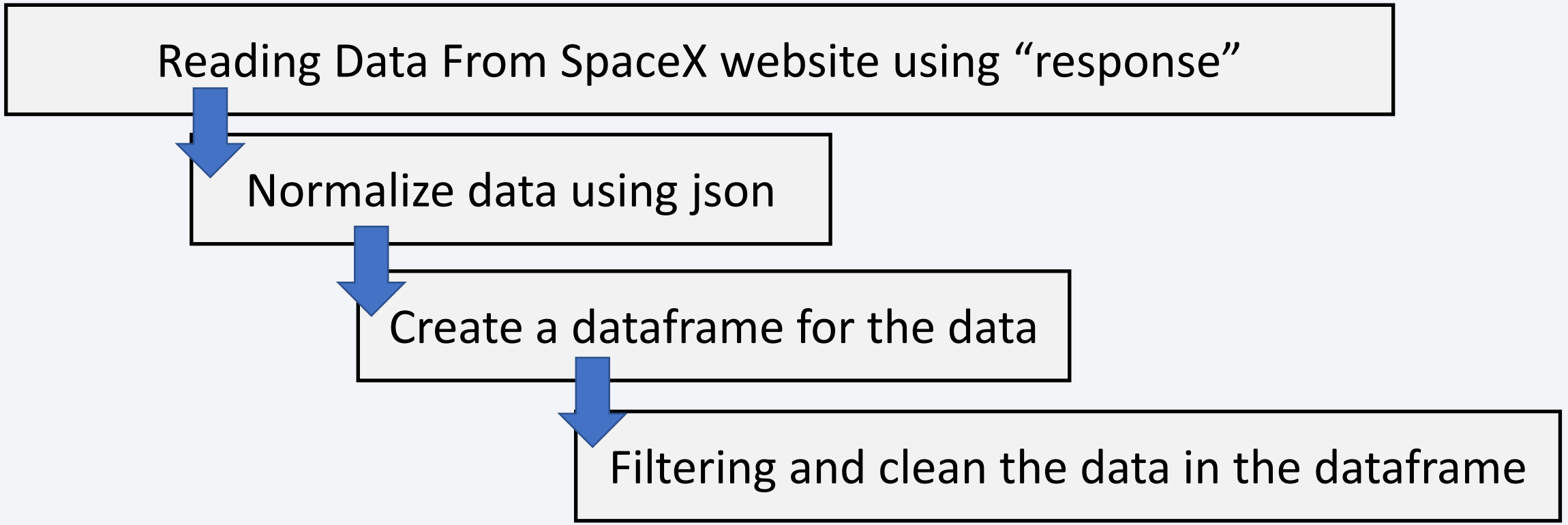
Data Collection

- Data was collected from the respective Wikipedia website using API requests and web scraping.
- The collected data set for SpaceX included flight number, date, booster version, payload mass, orbits, flights, orbits, flight outcomes, landing pads, launch sites including their coordinates as of longitude, latitude.



Data Collection – SpaceX API

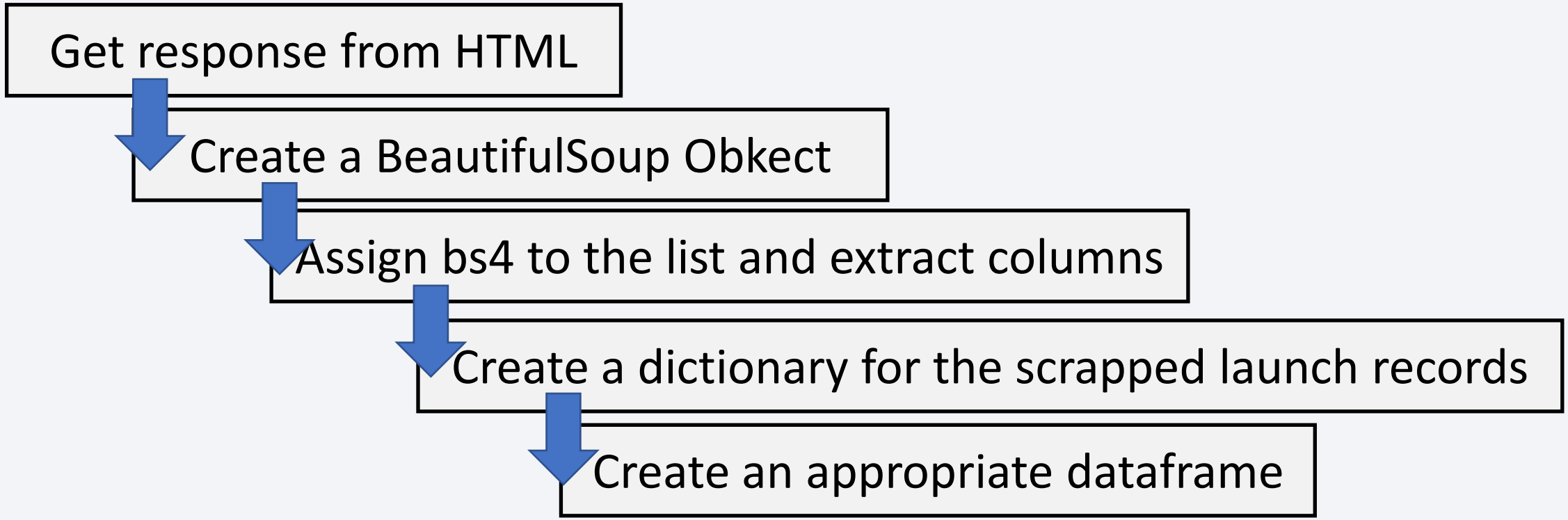
- Data collection with SpaceX REST calls and flowcharts



- GitHub public repository for the completed SpaceX API calls notebook is https://gitHUB.com/paramon22/IBM_10_capstone/
- File: 10_1_1_Falcon_Data_Collection_API.ipynb

Data Collection - Scraping

- Web scraping process flowchart



- GitHub public repository for the completed web scraping notebook is https://gitHUB.com/paramon22/IBM_10_capstone/
- File: 10_1_2_Falcon_Data_Web_Scrapping.ipynb

Data Wrangling

- Data wrangling was conducted woth Pandas and NumPy libraries

Missing values identified and replaced with mean values from other records

Found launch sites and number of flights form each site

Found the number of flights on each orbit

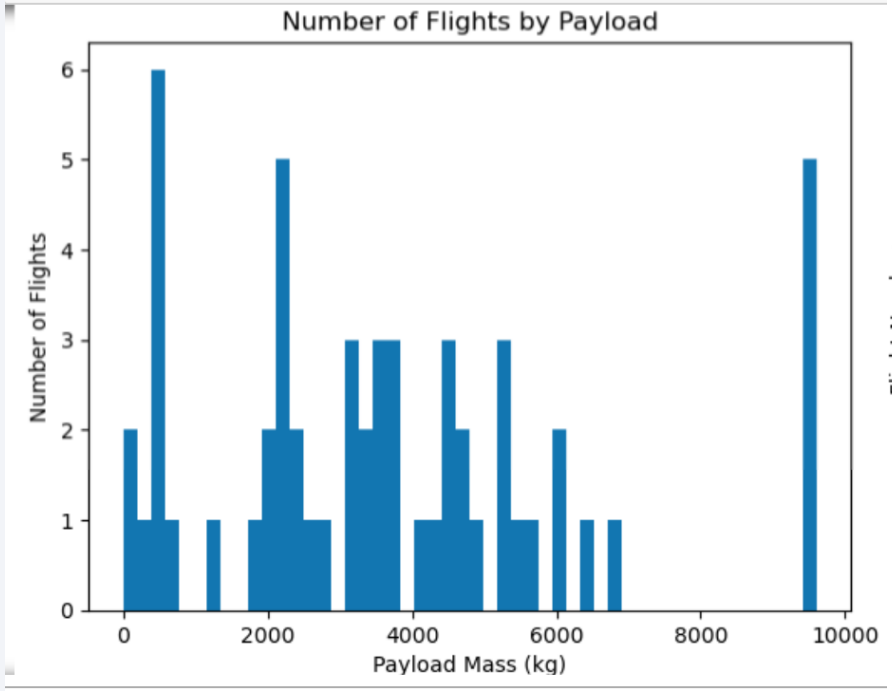
Found different flight outcomes by launch site

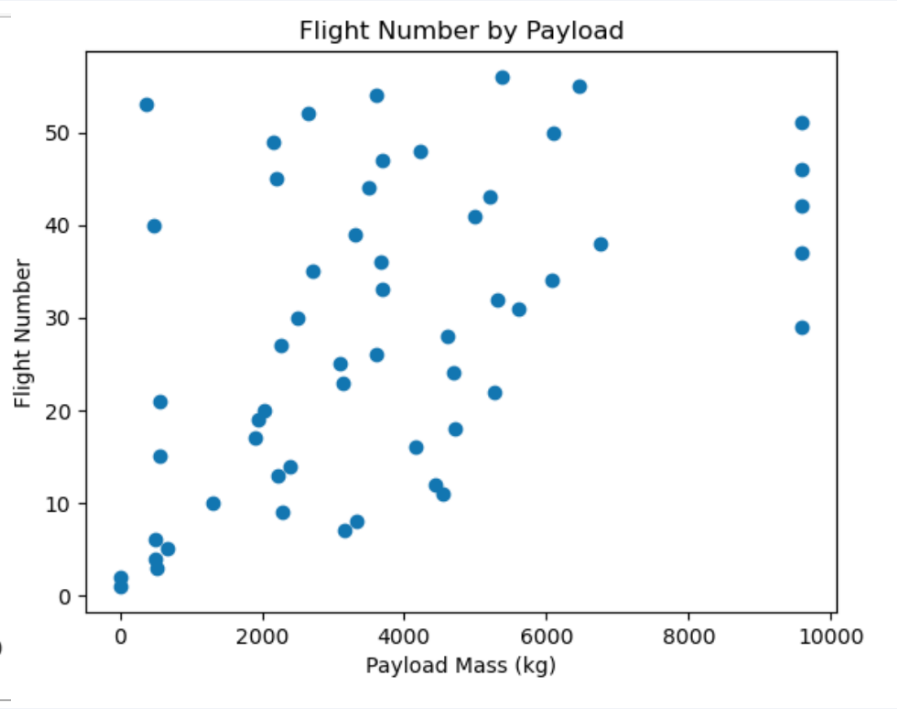
Found different landing outcomes

- The GitHub public repository for the completed data wrangling notebook
  https://gitHUB.com/paramon22/IBM_10_capstone/
- File: 10_1_3_Falcon_Data_Wrangling.ipynb

EDA with Data Visualization

- Various dependencies were identified and visualized.





- The GitHub public repository for the completed data wrangling notebook https://gitHUB.com/paramon22/IBM_10_capstone/
- Files: 10_2_Falcon_EDA_SQL.ipynb and 10_3_2_Falcon_data_visualization.ipynb

EDA with SQL (1/2)

- Loaded the dataset into table SPACEXTBL in my_data1.db database.
- Task 1: Displayed the names of the unique launch sites .
- Task 2: Displayed 5 first records of the launch site with the names begins with 'CCA'. It was CCAFS LC-40.
- Task 3: Displayed the total payload mass carried by boosters launched by NASA (CRS). It was 45,596 kg.
- Task 4: Displayed the average payload mass carried by booster version F9 v1.1 and booster versions begins with 'F9 v1.1'. It was 2,928.4 kg.

Launch Sites:

CCAFS SLC-40	34
CCAFS LC-40	26
KSC LC-39A	25
VAFB SLC-4E	16

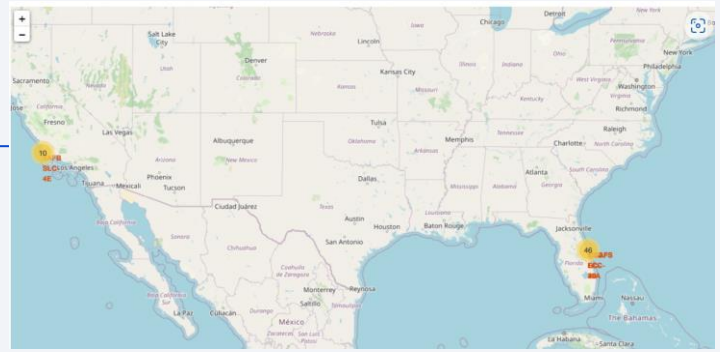
- The GitHub public repository for the completed data wrangling notebook https://gitHUB.com/paramon22/IBM_10_capstone/
- File: 10_2_Falcon_EDA_SQL.ipynb

EDA with SQL (2/2)

- Task 5: Listed the date when the first successful landing outcome in ground pad was achieved. It was 01/06/2014
- Task 6: Listied the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000,
- Task 7: Listed the total number of successful and failure mission outcomes. The launch sites were found and the total number of successful missions was 99.
- Task 8: Listed the names of the booster versions which have carried the maximum payload mass. There were 12 such missions.
- Task 9: Listed the records which will display the month names, failure landing outcomes in drone ship ,booster versions, launch_site for the months in year 2015. There were totally seven such missions, two of which failed.
- Task 10: Ranked 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.

Landing_Outcome	Count
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1

Build an Interactive Map with Folium



Objects created:

- Folium Map
- Markers to show all launch sites on a map
- Markers to show the successful and failed launches for each site on the map
- Lines that show the distances between a launch site to its proximities

Findings

- Launch sites are located close to railroads
- Launch sites are located close to freeways
- Launch sites are located close to coastlines
- Launch sites are reasonable far from populated areas (cities)

- The GitHub public repository for the completed data wrangling notebook
  https://gitHUB.com.paramon22/IBM_10_capstone/
- File: 10_3_1_Launch_Site_Location.s_Folium.ipynb

Build a Dashboard with Plotly Dash

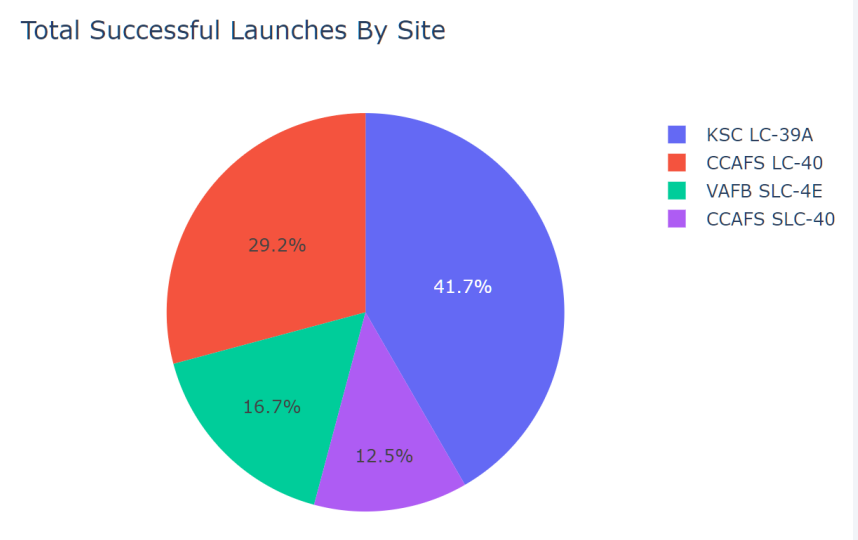
The dashboard application consists of a pie chart and a scatter point chart.

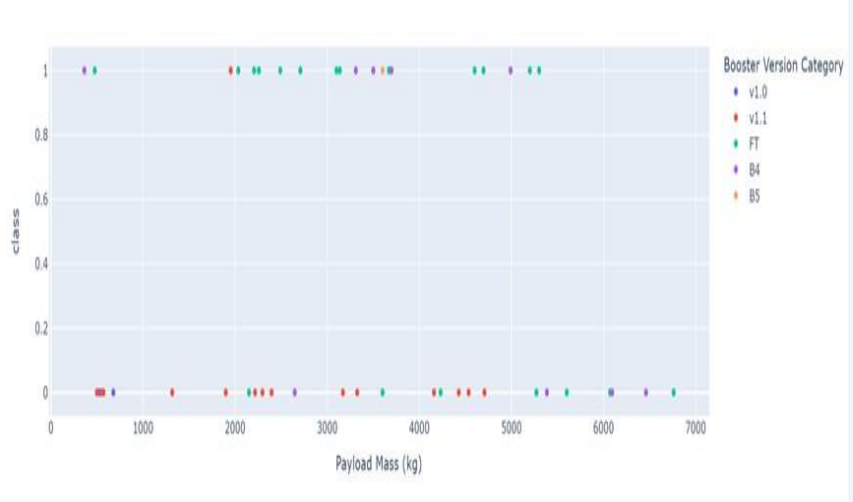
The pie chart:

- shows the total successful launches by sites

The scatter chart:

- shows the relationship between successful outcomes and payload mass(Kg)

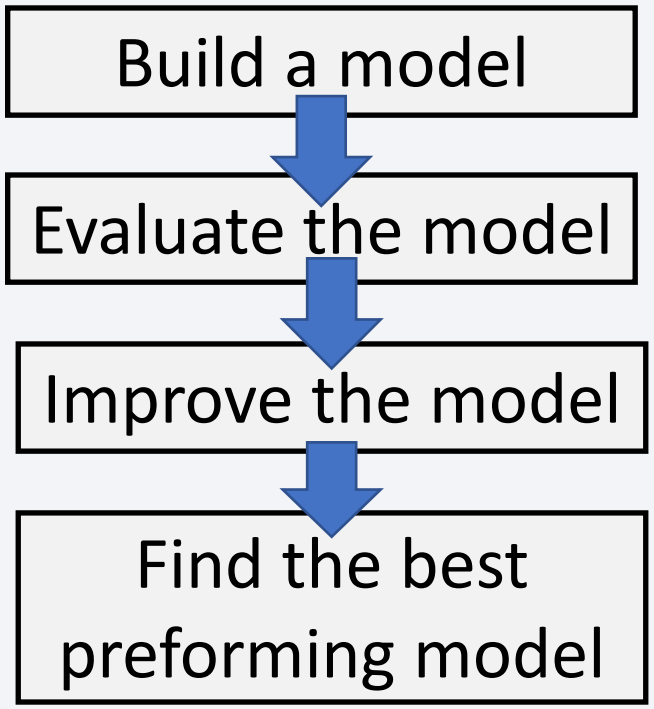




- The GitHub public repository for the dashboard notebook https://gitHUB.com/paramon22/IBM_10_capstone/
- File: 10_3_3_SpaceX_dashboard.ipynb

Predictive Analysis (Classification)

- Classification is a supervised learning approach
- Classification categorizes some unknown items into a discrete set of categories ir classes
- Classification target attribute is a categorical variable.



- First build a model
- Split data set into training and testing subsets
- Ecaluate the model
- Improve the model
- Find the best performing model to make predictions

- The GitHub public repository for the dashboard notebook https://gitHUB.com/paramon22/IBM_10_capstone/
- File: 10_4_SpaceX_ML_prediction_Part_5.ipynb

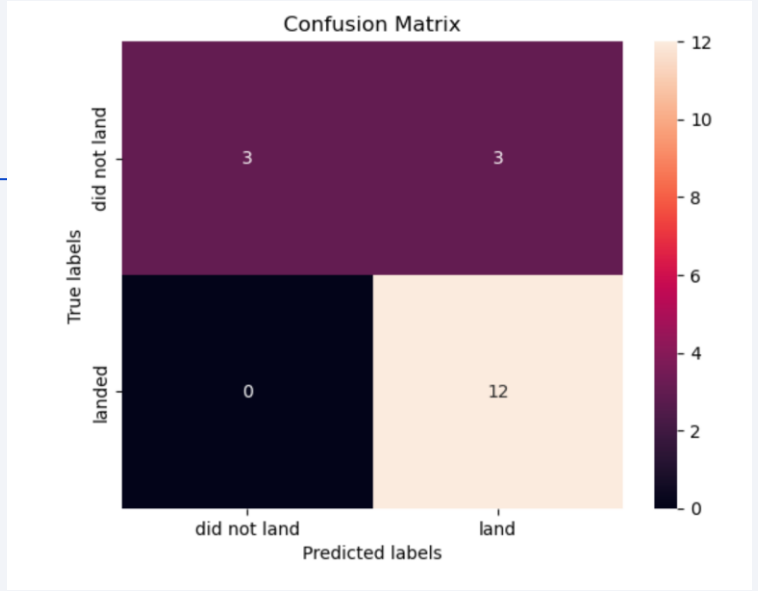
Results

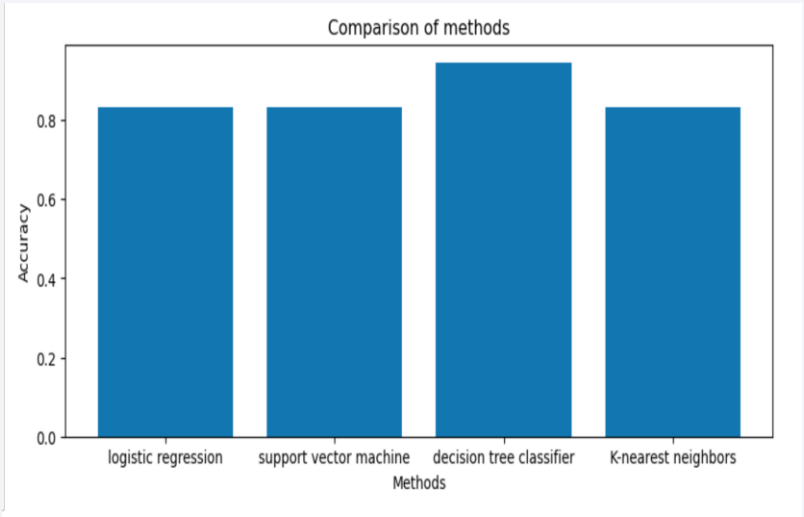
Exploratory data analysis (EDA) was conducyed by using

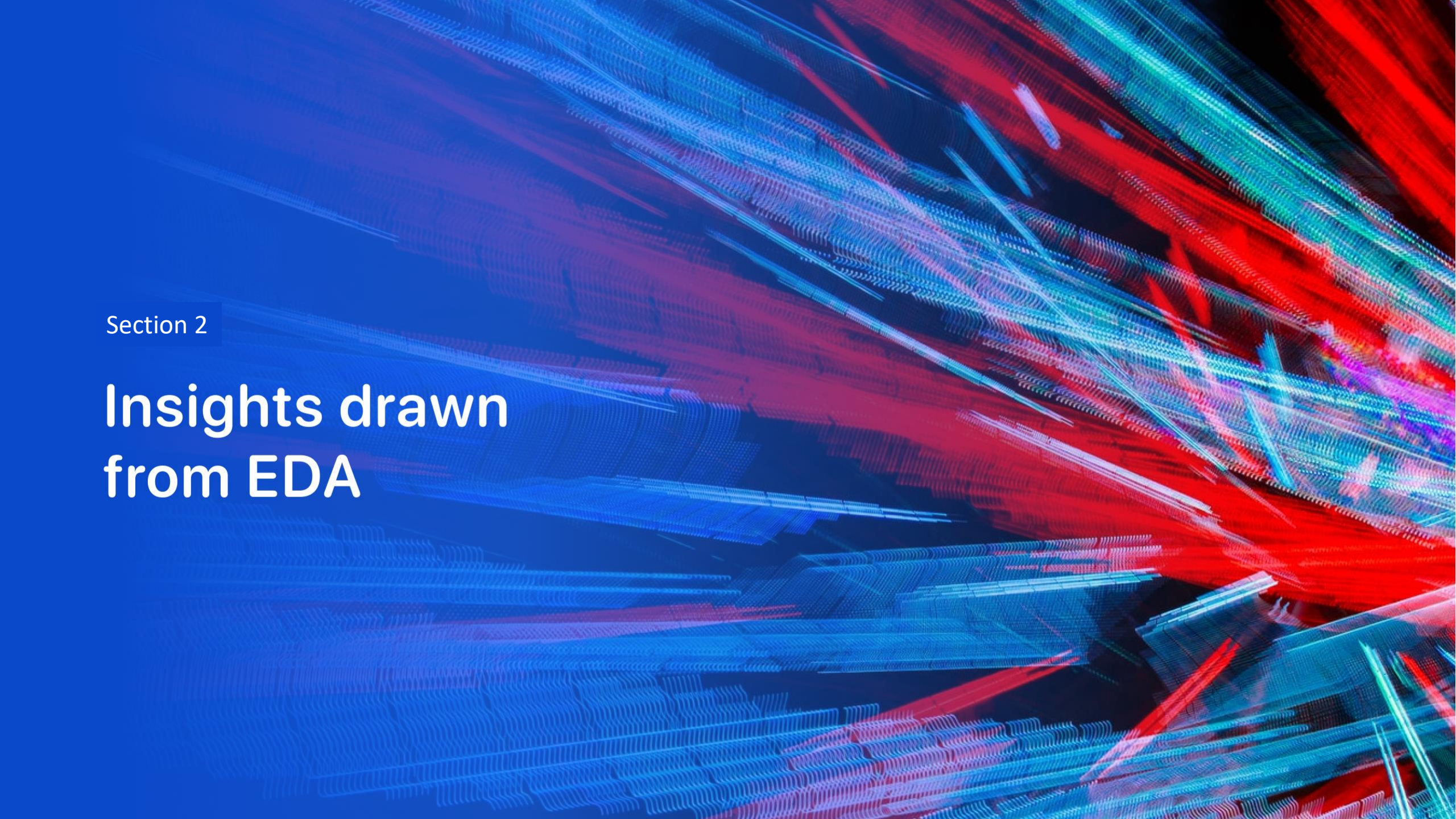
- Charts using MatPlotLib
- SQL queries
- Interactive maps with Folium
- Interactive dashboard

Prediction:

- all prediction methods have shown a similar accuracy of about 83% for test data.

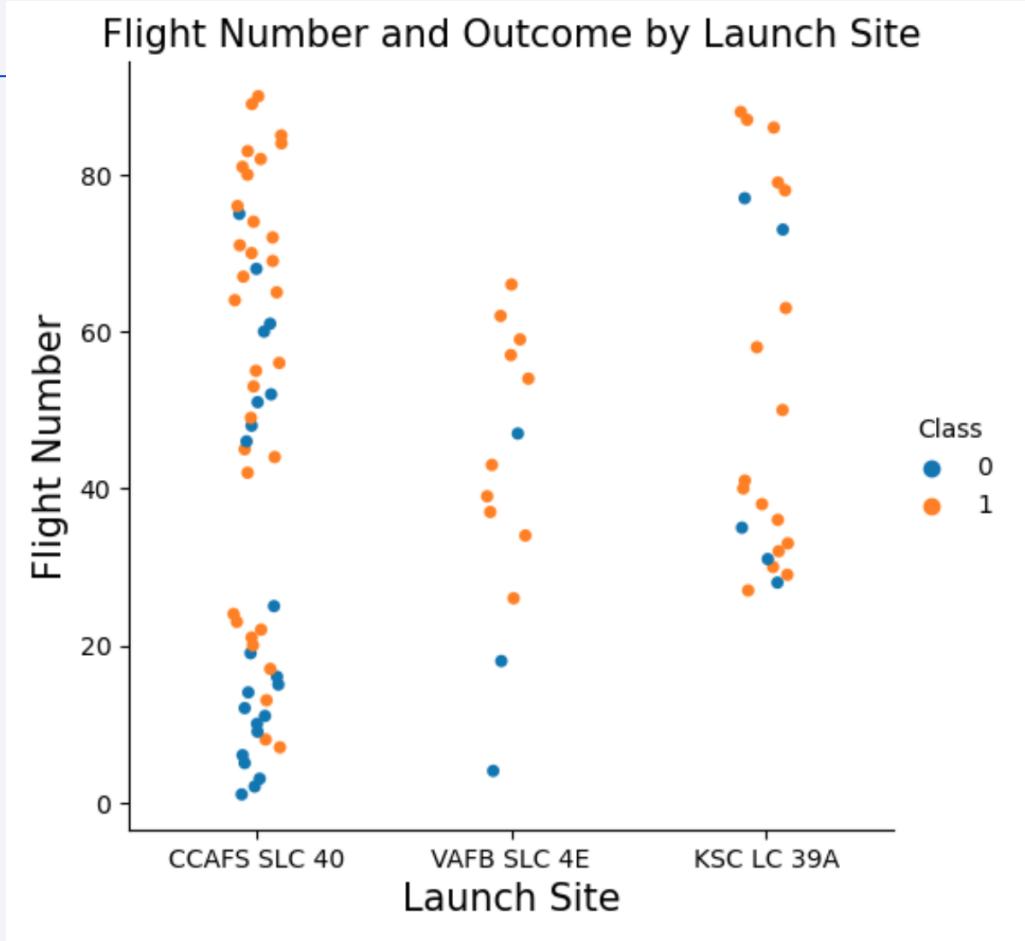





Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

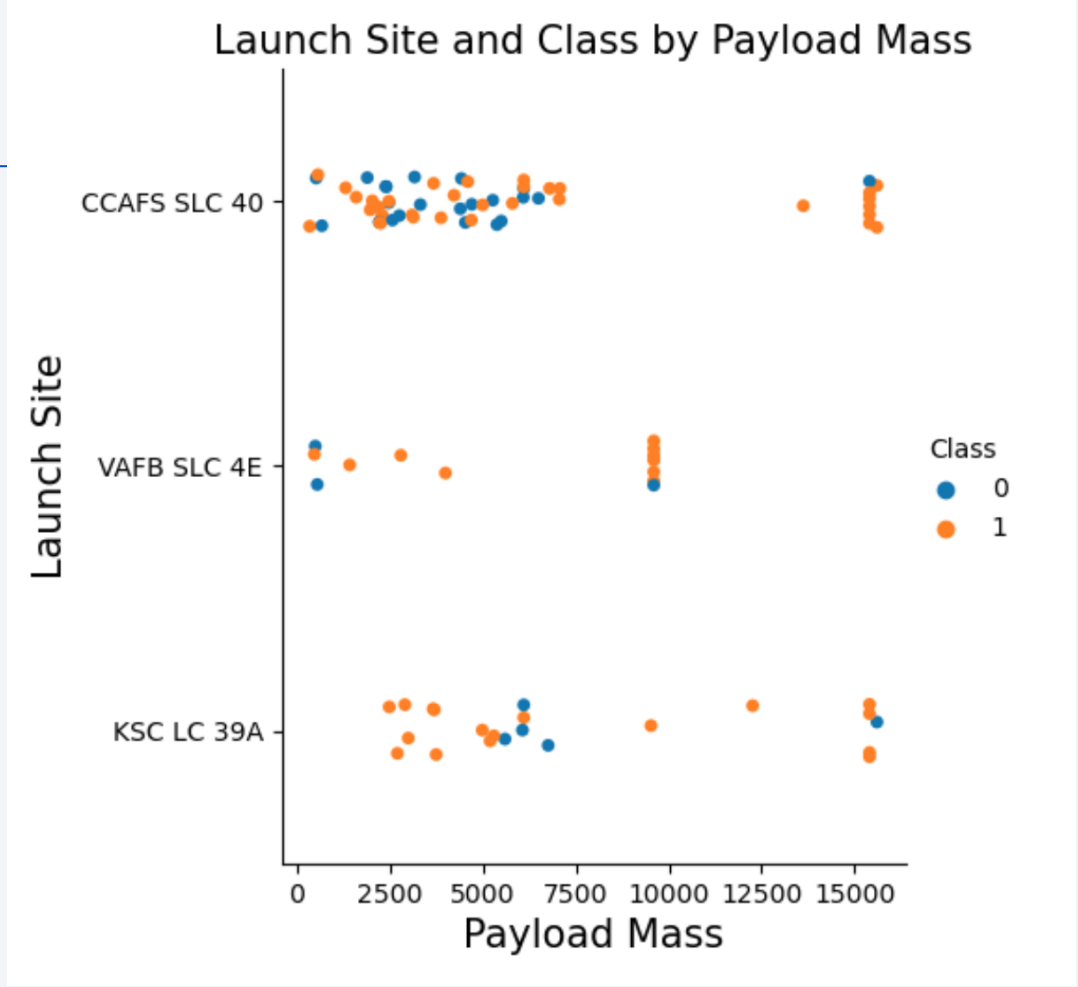
- Class 0 (blue) means an unsuccessful launch and
  Class 1 (orange) means a successful one.
- The figure shows that mission success rate increased with the flight number. i.e. the quality of flight is improving.



- The GitHub public repository for the dashboard notebook
  https://gitHUB.com/paramon22/IBM_10_capstone/
- File: 10_4_1_Falcon_EDA_DF.ipynb

Payload vs. Launch Site

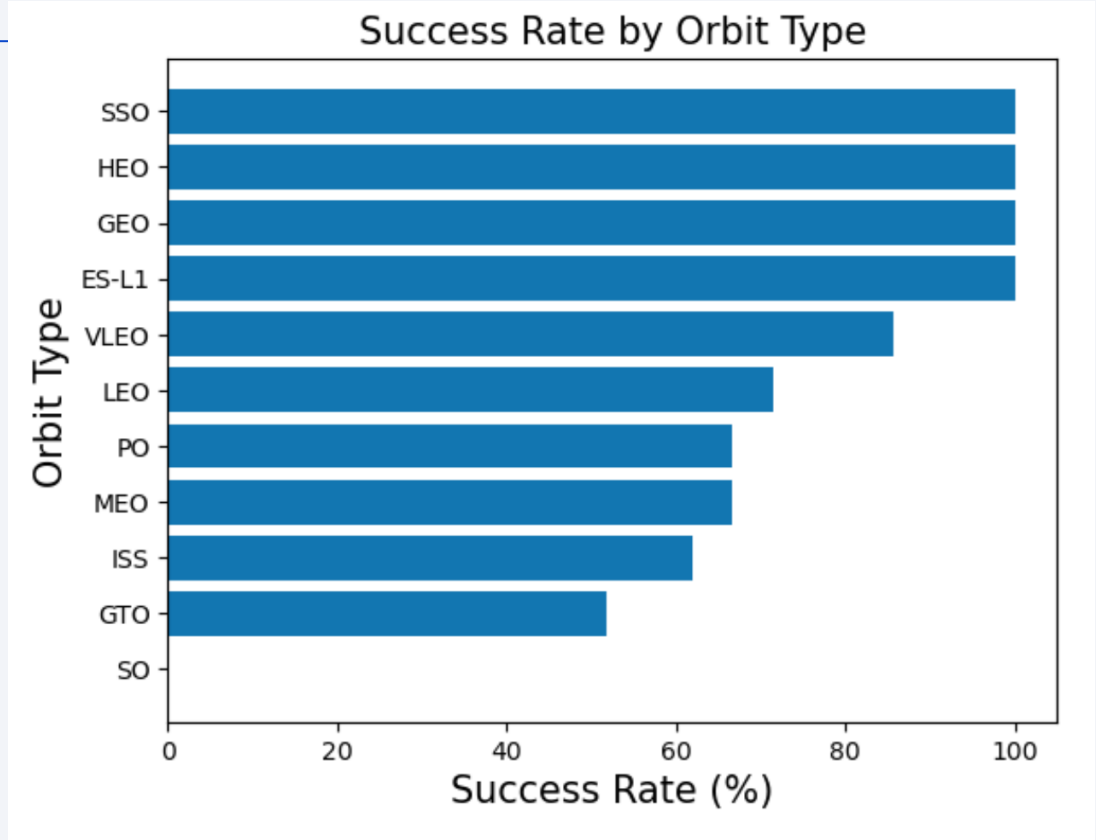
- Class 0 (blue) means an unsuccessful launch and
  Class 1 (orange) means a successful one.
- .According to the chart, the greater payload mass, the higher the mission success rate.
- However, it may be esplaned by the fact that the greater payload mass was used in the later missions which were associated with he better experience and quality.



- The GitHub public repository for the dashboard notebook
  https://gitHUB.com/paramon22/IBM_10_capstone/
- File: 10_4_1_Falcon_EDA_DF.ipynb

Success Rate vs. Orbit Type

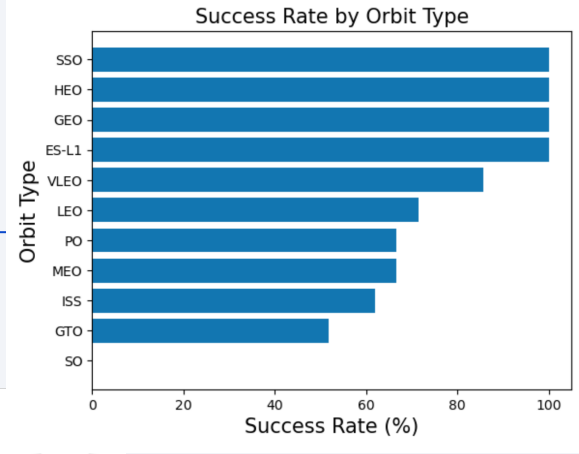
- Missions were 100% successful on orbits SSO, HEO, GEO, and ES-L1.
- The least successful missions were on orbit GTO.

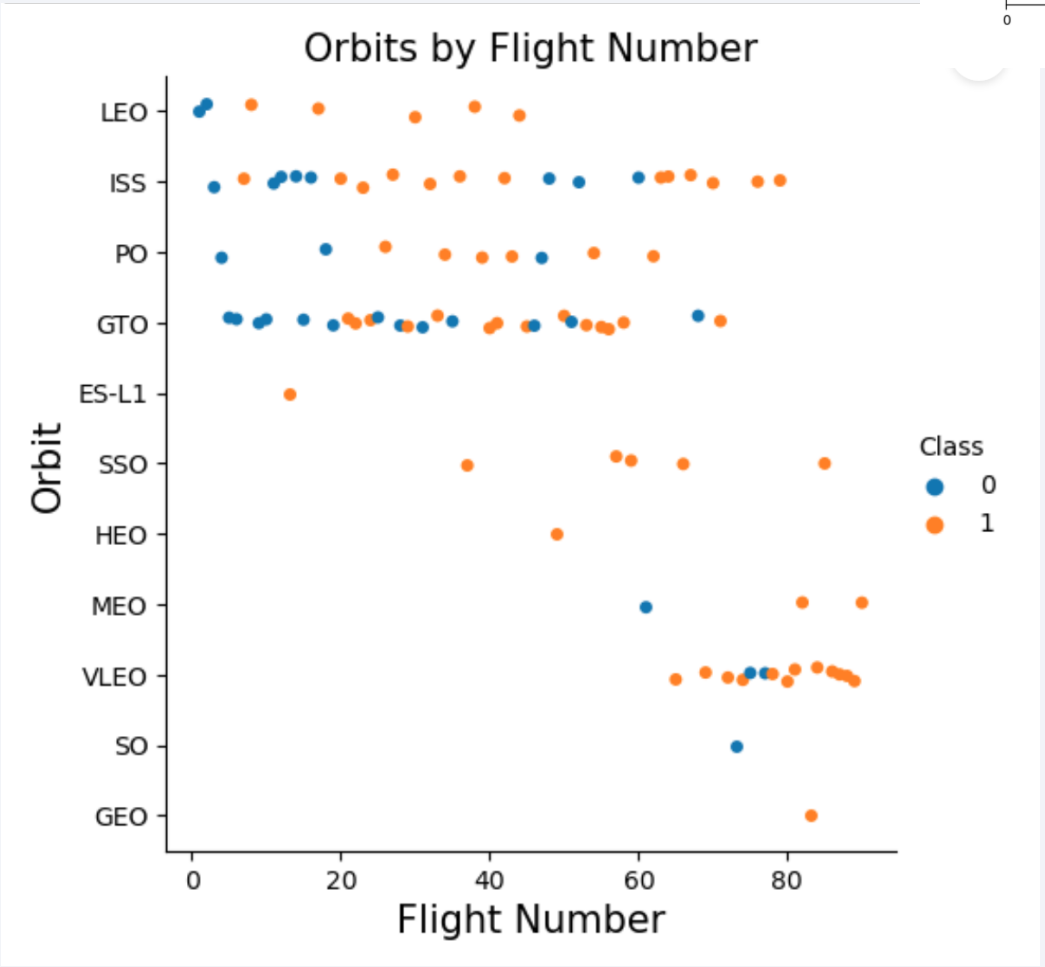


- The GitHub public repository for the dashboard notebook https://gitHUB.com/paramon22/IBM_10_capstone/
- File: 10_4_1_Falcon_EDA_DF.ipynb

Flight Number vs. Orbit Type

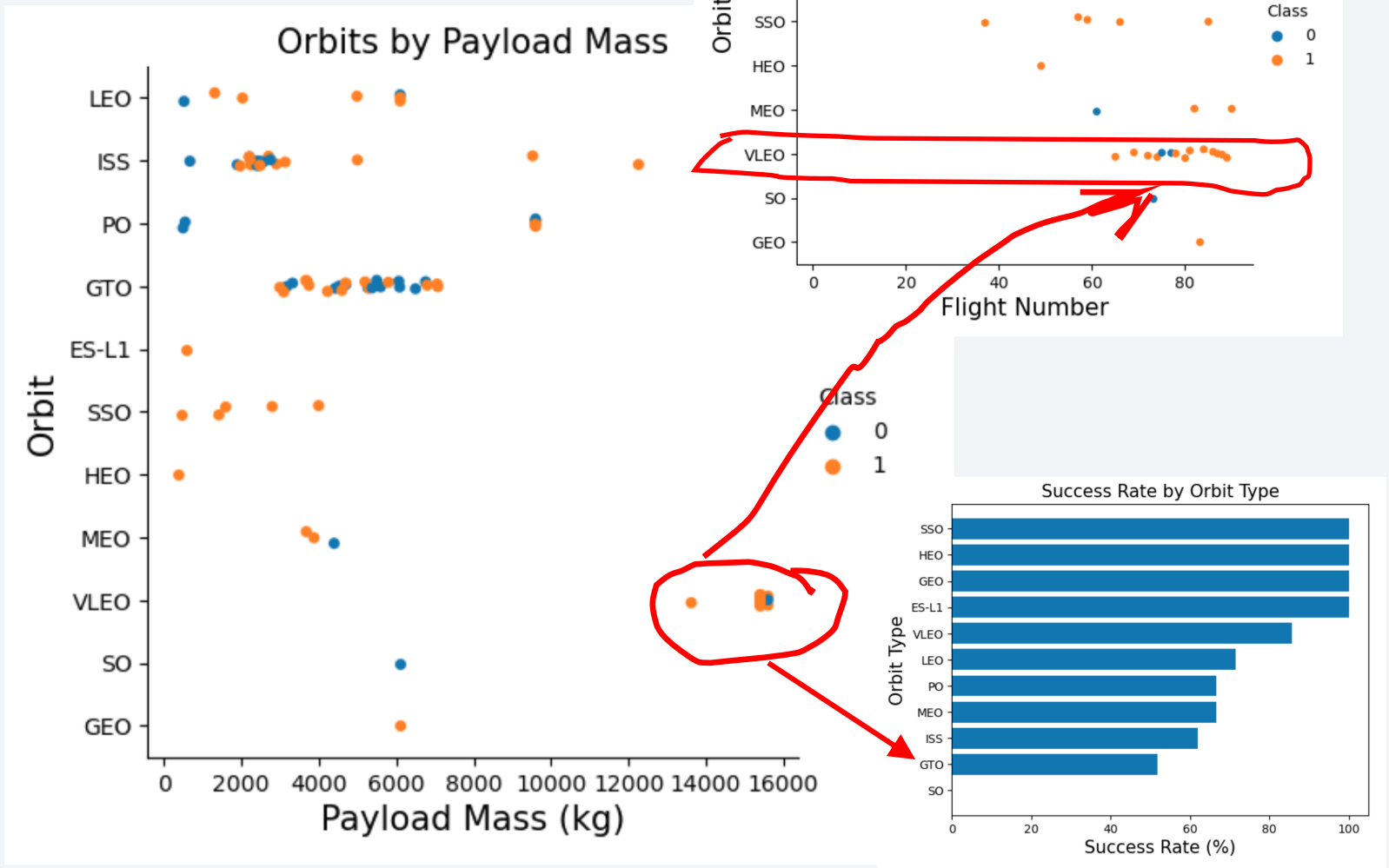
- Most flights were launched on orbits ISS, GTO, and PO.
- Show the screenshot of the scatter plot with explanations.
- According to the previous slide, missions on the orbit GTO, ISS, and PO have the highest failure rate.





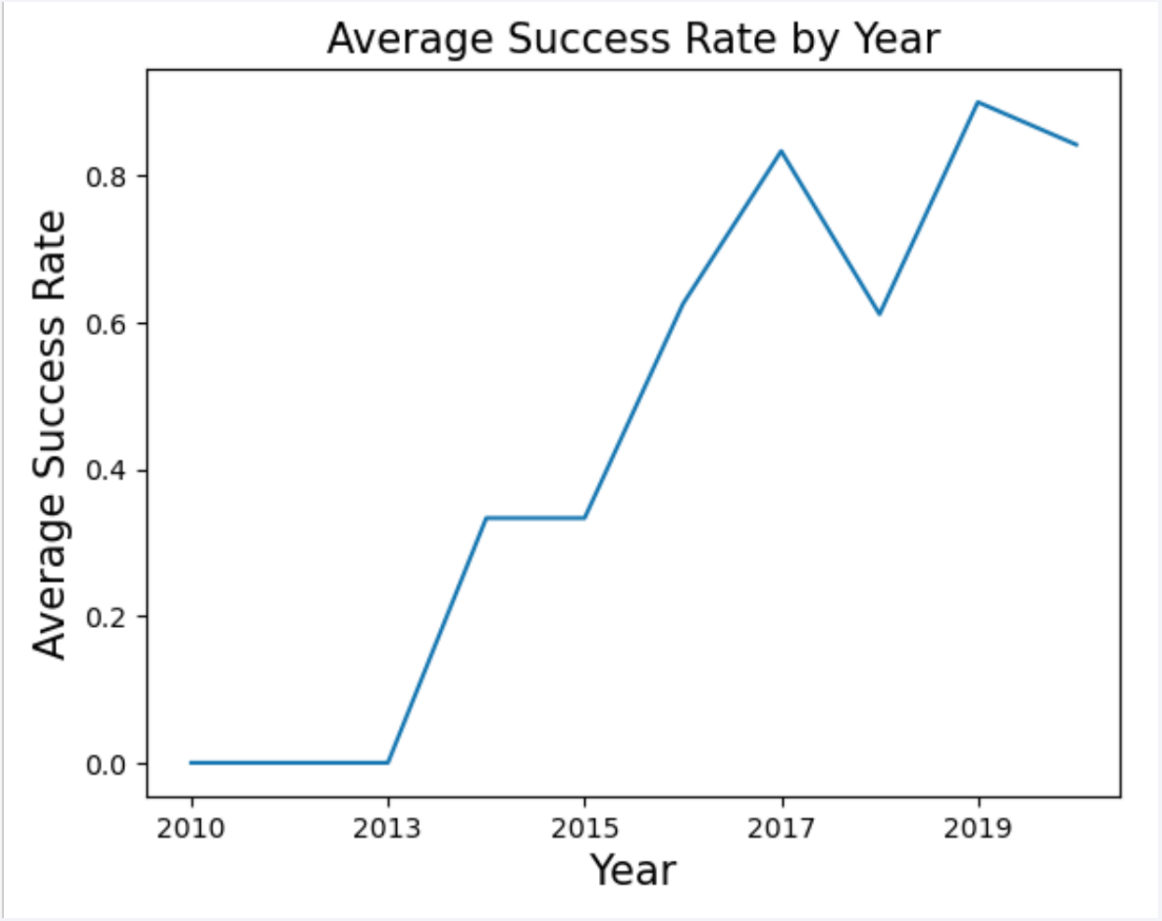
Payload vs. Orbit Type

- Most flights with highest payload were launched on VLEO orbit
- Most flights were launched on orbit GTO with a mddle range of payload, though the success rate of missions on this orbit is lowest.



Launch Success Yearly Trend

- Mission success rate if growing with experience (years).
- An exemption was 2018, the reason for which were beyond the scope of this analysis.
- As a guess, reduced success rate in 2018 might be possibly caused by introducing new designs or equipment or some other changes.



All Launch Site Names

In [13]:
```
# Finding the names of the unique launch sites in the space mission
%sql SELECT distinct Launch_Site from SPACEXTBL
```

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

Out[13]:

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

- The GitHub public repository for the dashboard notebook https://gitHUB.com/paramon22/IBM_10_capstone/
- File: 10_2_Falcon_EDA_SQL.ipynb

Launch Site Names Begin with 'CCA'

Data source: Spacex.csv

In [14]:
```
%sql SELECT Launch_Site from SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5
```

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

Out[14]:

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

```
In [15]: %sql SELECT sum(PAYLOAD_MASS__KG_) from SPACEXTBL WHERE Customer = 'NASA (CRS)'
 * sqlite:///my_data1.db
Done.
```

Out[15]:

sum(PAYLOAD_MASS__KG_)
45596.0

Average Payload Mass by F9 v1.1

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

In [16]:
```
%sql SELECT avg(PAYLOAD_MASS__KG_) from SPACEXTBL WHERE Booster_Version LIKE 'F9 v1.1'
```

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

Out[16]:

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

Task 5

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

Hint:Use min function

In [17]:
```
%sql SELECT min(Date) from SPACEXTBL WHERE Mission_Outcome = 'Success'
```

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

Out[17]:

min(Date)
01/06/2014

Successful Drone Ship Landing with Payload between 4000 and 6000

```
In [18]: %sql SELECT Booster_Version, PAYLOAD_MASS__KG_ from SPACEXTBL WHERE PAYLOAD_MASS__KG_ between 4000 and 6000
 * sqlite:///my_data1.db
Done.
```

Out[18]:

Booster_Version	PAYLOAD_MASS__KG_
F9 v1.1	4535.0
F9 v1.1 B1011	4428.0
F9 v1.1 B1014	4159.0
F9 v1.1 B1016	4707.0
F9 FT B1020	5271.0
F9 FT B1022	4696.0
F9 FT B1026	4600.0
F9 FT B1030	5600.0
F9 FT B1021.2	5300.0
F9 FT B1032.1	5300.0
F9 B4 B1040.1	4990.0
F9 FT B1031.2	5200.0
F9 B4 B1043.1	5000.0
F9 FT B1032.2	4230.0
F9 B4 B1040.2	5384.0
F9 B5 B1046.2	5800.0
F9 B5 B1047.2	5300.0
F9 B5 B1046.3	4000.0
F9 B5B1054	4400.0
F9 B5 B1048.3	4850.0
F9 B5 B1051.2	4200.0
F9 B5B1060.1	4311.0
F9 B5 B1058.2	5500.0
F9 B5B1062.1	4311.0

```
%sql SELECT Booster_Version, PAYLOAD_MASS__KG_ \
     from SPACEXTBL WHERE PAYLOAD_MASS__KG_ \
     between 4000 and 6000
```

Total Number of Successful and Failure Mission Outcomes

One successful mission falls apart from the other ones possibly because of spelling or syntactical difference Such situations may happen due to not normal form of the tables in the database.

Task 7

List the total number of successful and failure mission outcomes

In [18]:
```
# Total number of Missions
%sql SELECT count(*) as 'Total Number of Missions' from SPACEXTBL
```

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

Out[18]:

Total Number of Missions
999

In [19]:
```
%sql SELECT distinct Mission_Outcome, count(*) as 'Number of Missions'
     FROM SPACEXTBL \
     GROUP BY Mission_Outcome
```

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

Out[19]:

Mission_Outcome	Number of Missions
None	898
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

In [20]:
```
%sql SELECT Booster_Version, PAYLOAD_MASS__KG_ from SPACEXTBL \
    WHERE PAYLOAD_MASS__KG_ = (SELECT max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

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

Out[20]:

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600.0
F9 B5 B1049.4	15600.0
F9 B5 B1051.3	15600.0
F9 B5 B1056.4	15600.0
F9 B5 B1048.5	15600.0
F9 B5 B1051.4	15600.0
F9 B5 B1049.5	15600.0
F9 B5 B1060.2	15600.0
F9 B5 B1058.3	15600.0
F9 B5 B1051.6	15600.0
F9 B5 B1060.3	15600.0
F9 B5 B1049.7	15600.0

2015 Launch Records

In [41]:
```
# Sqlite does not support functions to extract month name and year from date.
# strftime('%m', Date) or strftime('%y %m %d', Date) do not work either
# For this reason, I use d month as numerals

%sql SELECT Date, substr(Date,4,2) as Month, substr(Date,7,10) as Year, \
    Booster_Version, Launch_Site, Landing_Outcome from SPACEXTBL \
    WHERE (Landing_Outcome LIKE 'Fail%' AND substr(Date,7,4) = '2015')
```

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

Out[41]:

Date	Month	Year	Booster_Version	Launch_Site	Landing_Outcome
01/10/2015	10	2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
14/04/2015	04	2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

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

In [39]:
```
%sql SELECT landing_outcome, count(*) AS "Count" \
   FROM SPACEXTBL \
   WHERE DATE BETWEEN '04-06-2010' and '20-03-2017' \
   GROUP BY landing_outcome \
   ORDER BY Count DESC
```

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

Out[39]:

Landing_Outcome	Count
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1

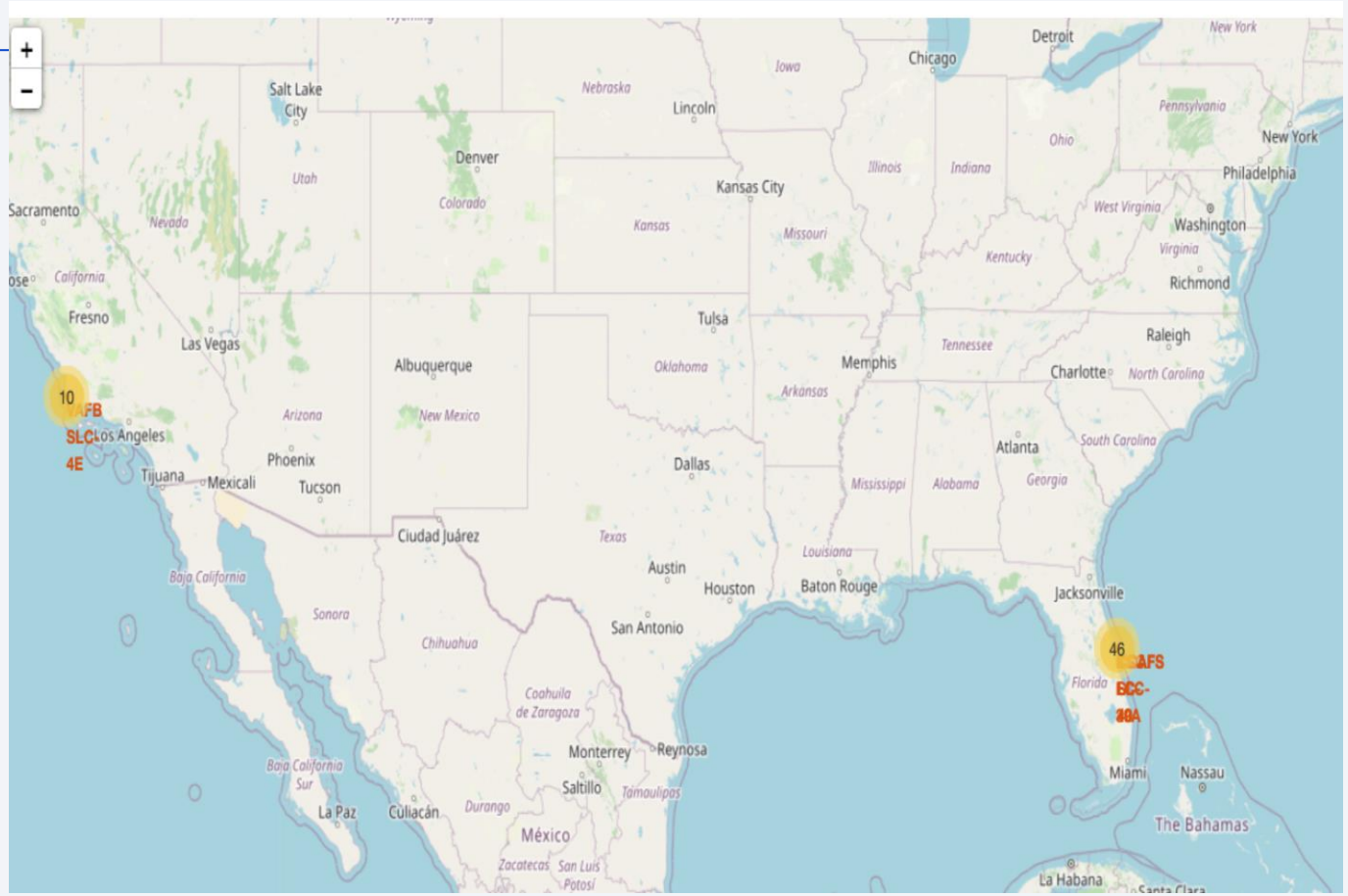
Section 3

Launch Sites Proximities Analysis

SpaceX Falcon Launch Sites

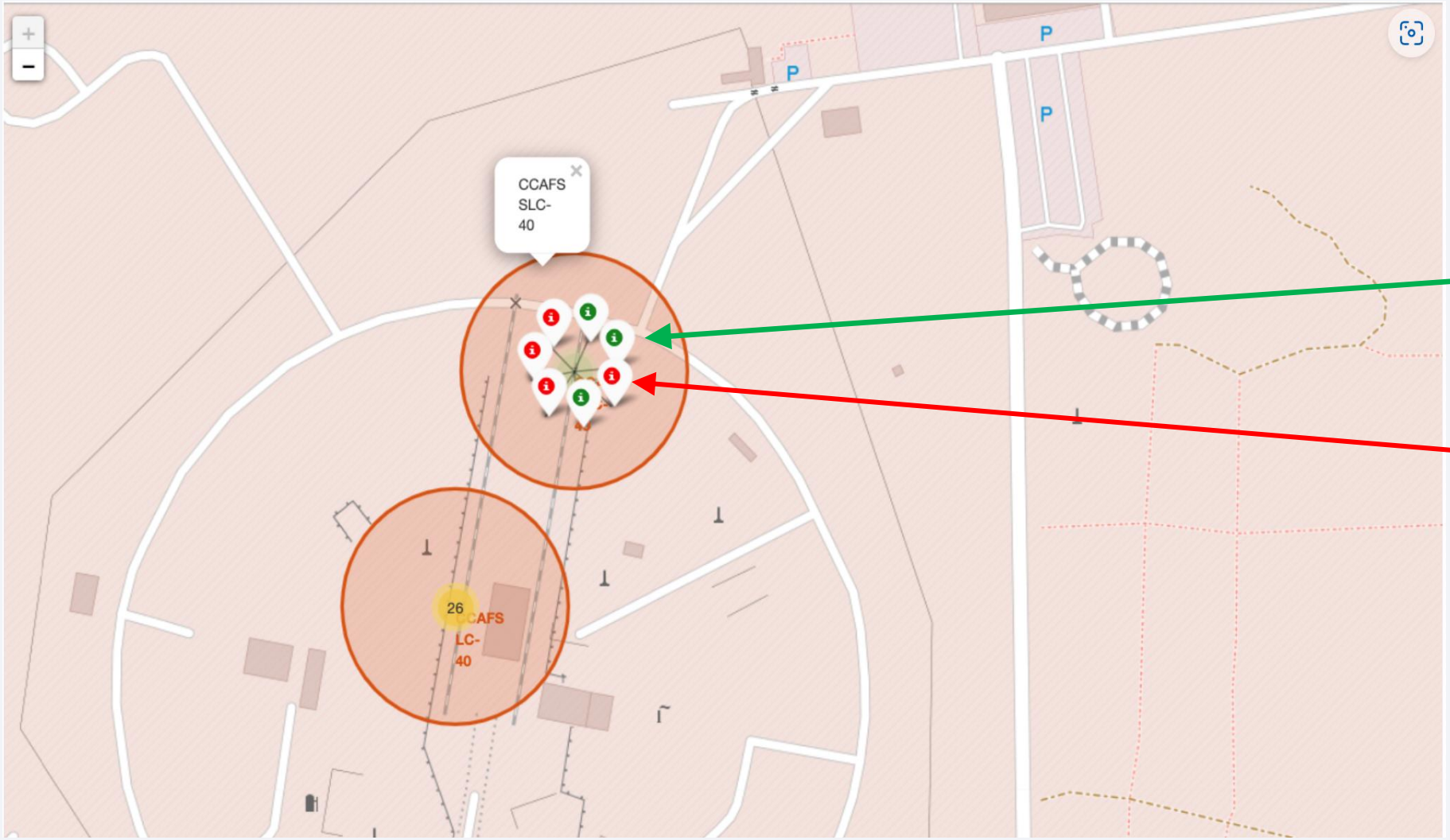
There are four SpaceX Falcon launch sites:

- One in California on the Pacific coast.
- Three in Florida on the Atlantic coast.



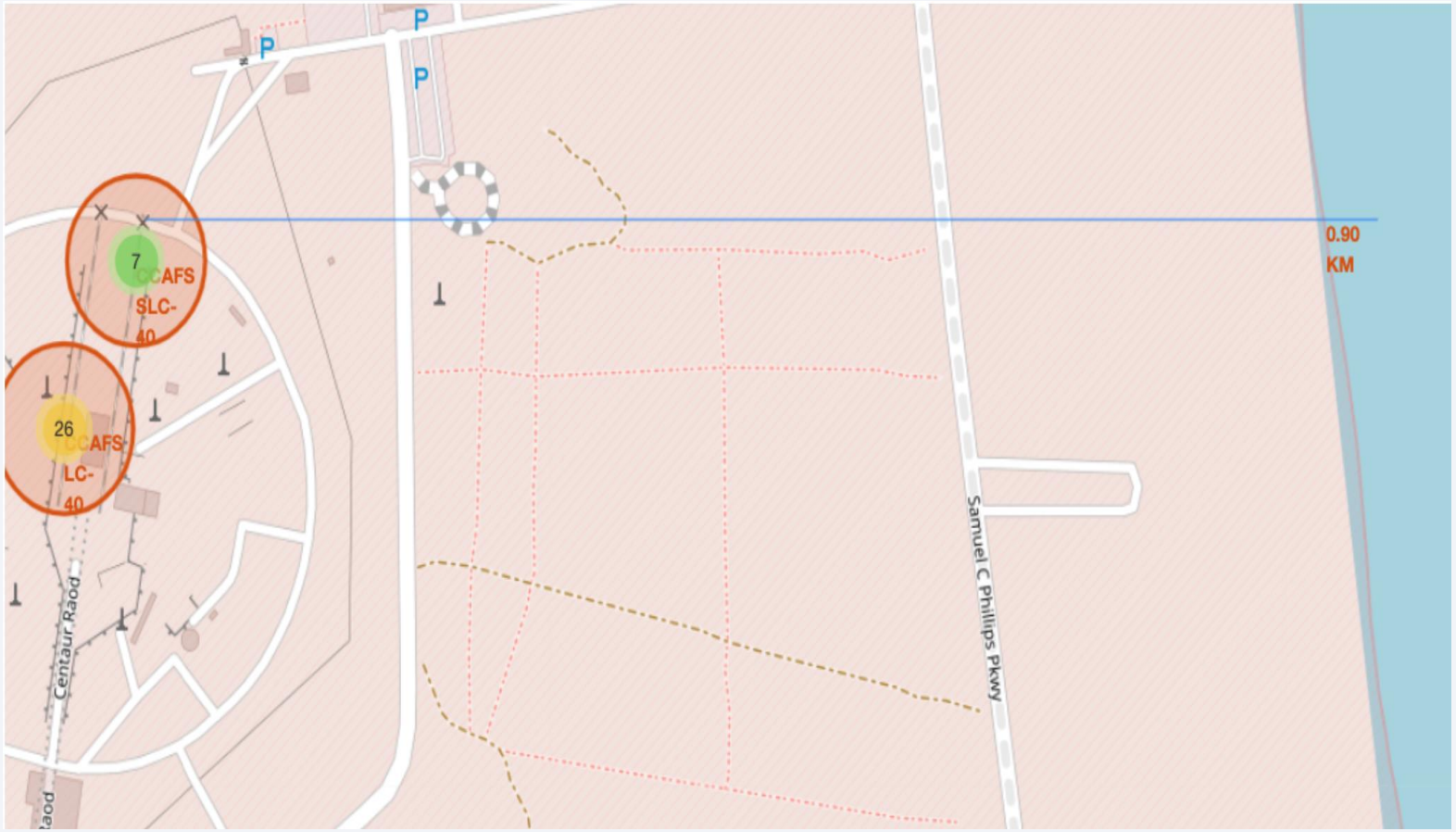
- The GitHub public repository for the dashboard notebook https://gitHUB.com/paramon22/IBM_10_capstone/
- File: 10_3_1_Launch_Site_Location.s_Folium.ipynb

CCAFS SLC-40 Launch Site with Marked Launch Outcomes

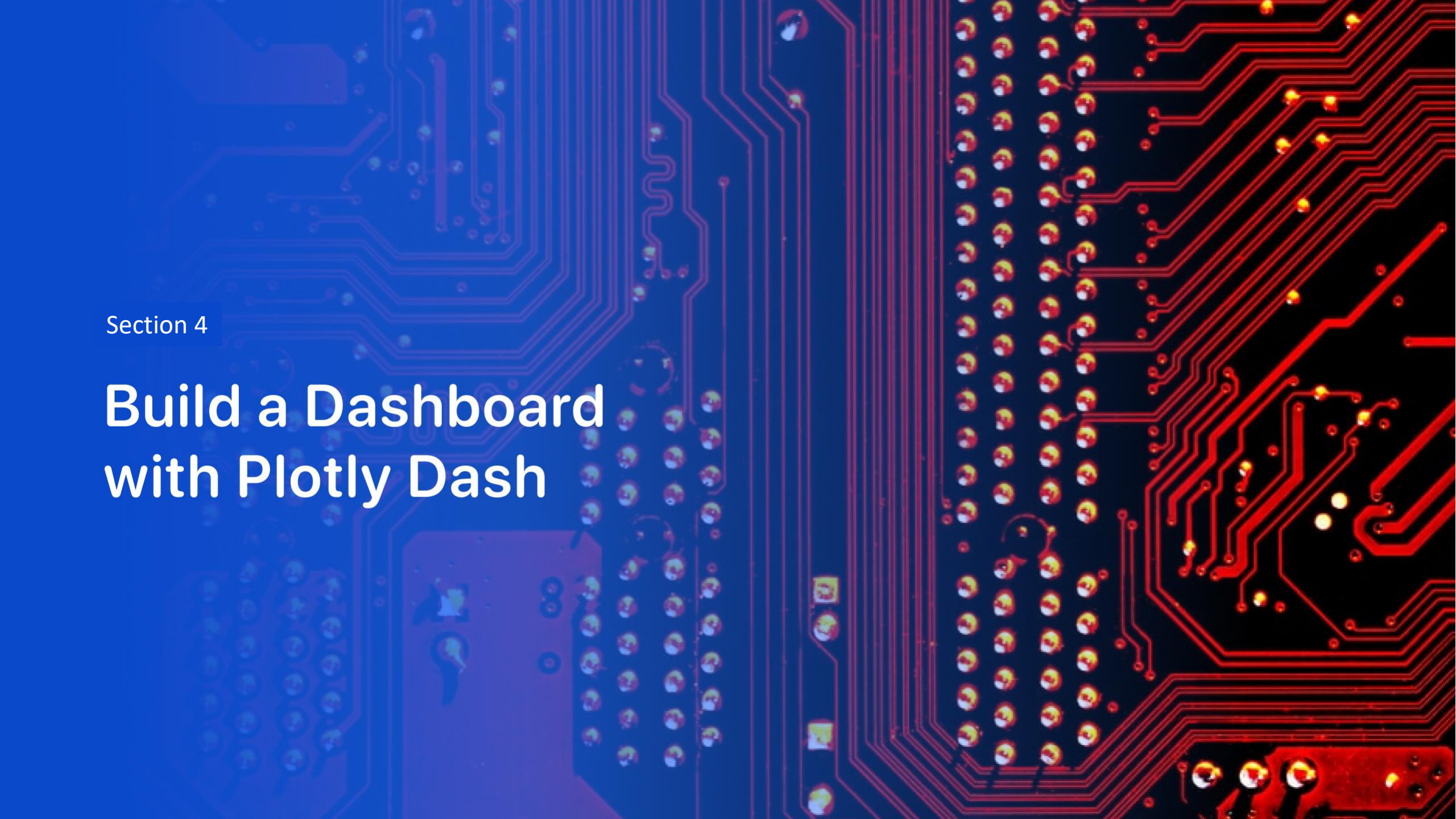


- Successful flight outcomes marked with **green** dots
- Failed flight outcomes marked with **red** dots

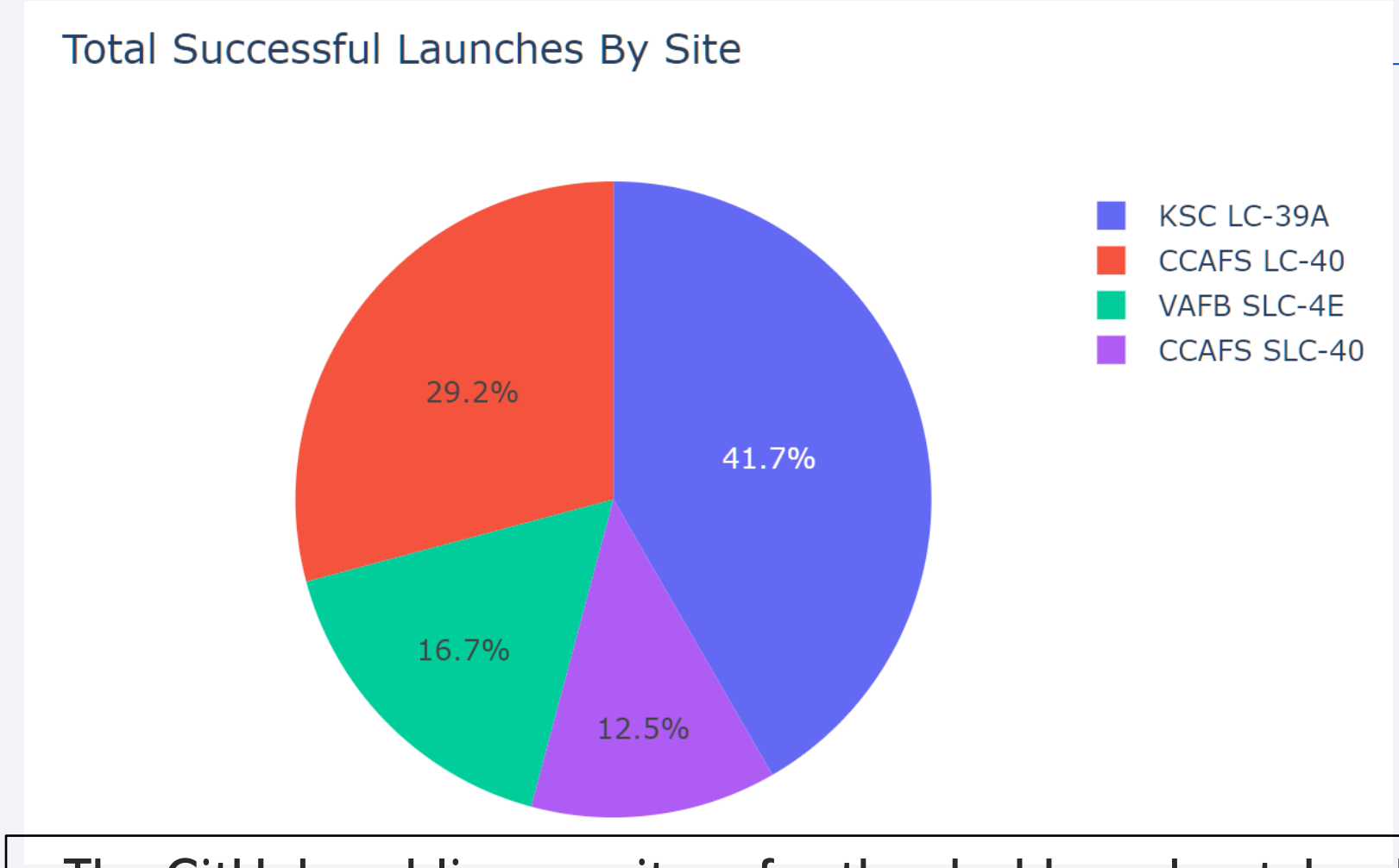
Proximity of a Launch Site ot roads, railroads, and the coastline



- Launch sites are located in close proximity to roads, railroads, and the coastline.


Section 4
Build a Dashboard with Plotly Dash

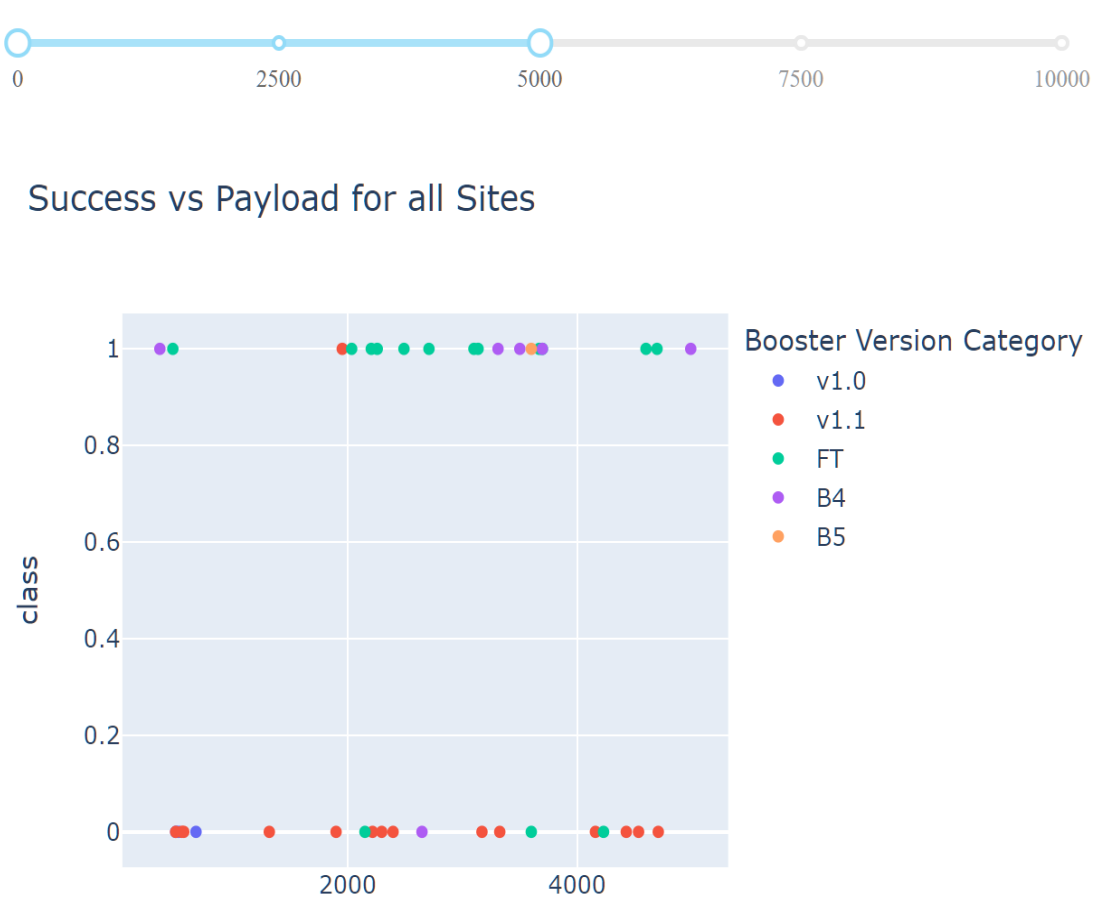
Successful Missions by Launch Site



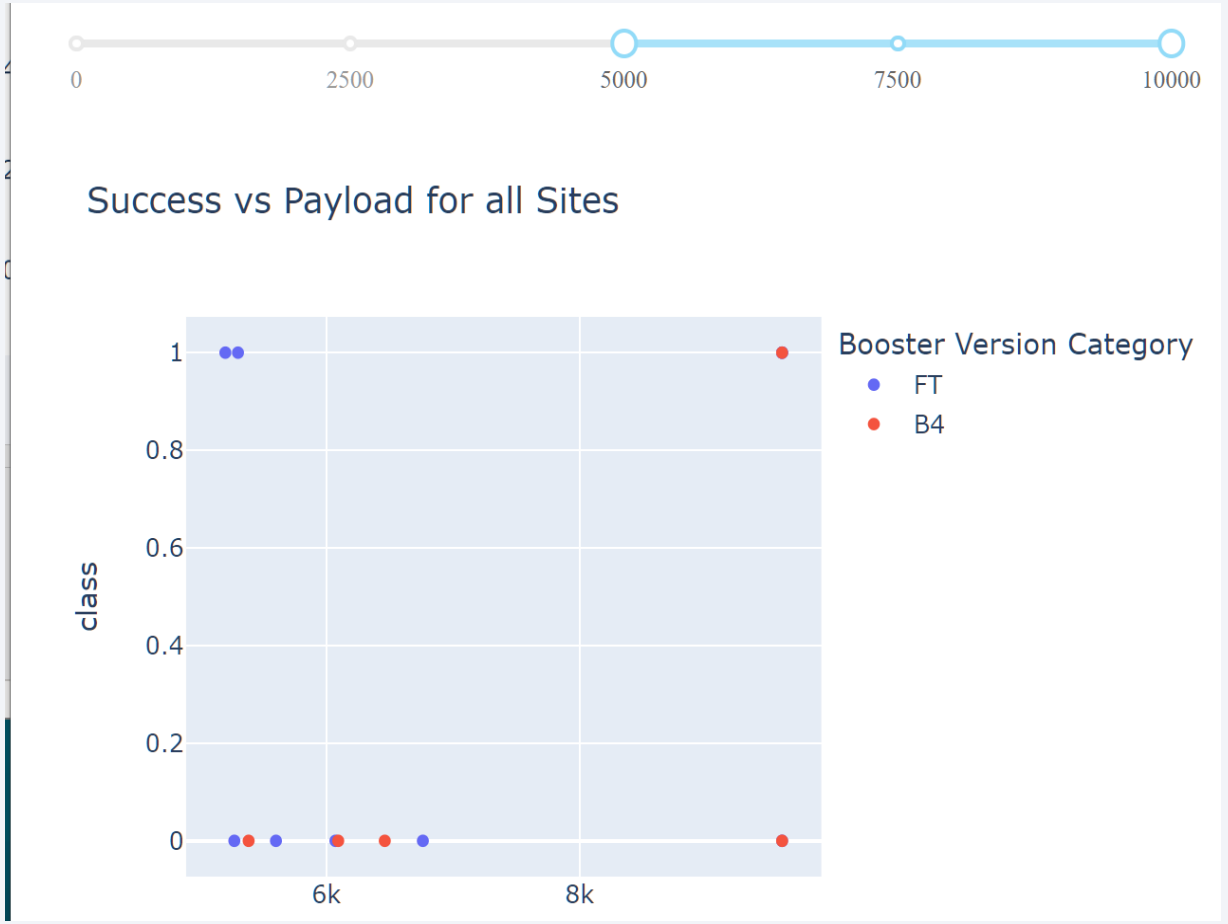
- The majority of the successful flight have been launched from launch site KSC LC-39A.

- The GitHub public repository for the dashboard notebook https://gitHUB.com/paramon22/IBM_10_capstone/
- File: 10_3_3_SpaceX_dashboard.ipynb

Successful Missions vs Payload Mass



Low payload



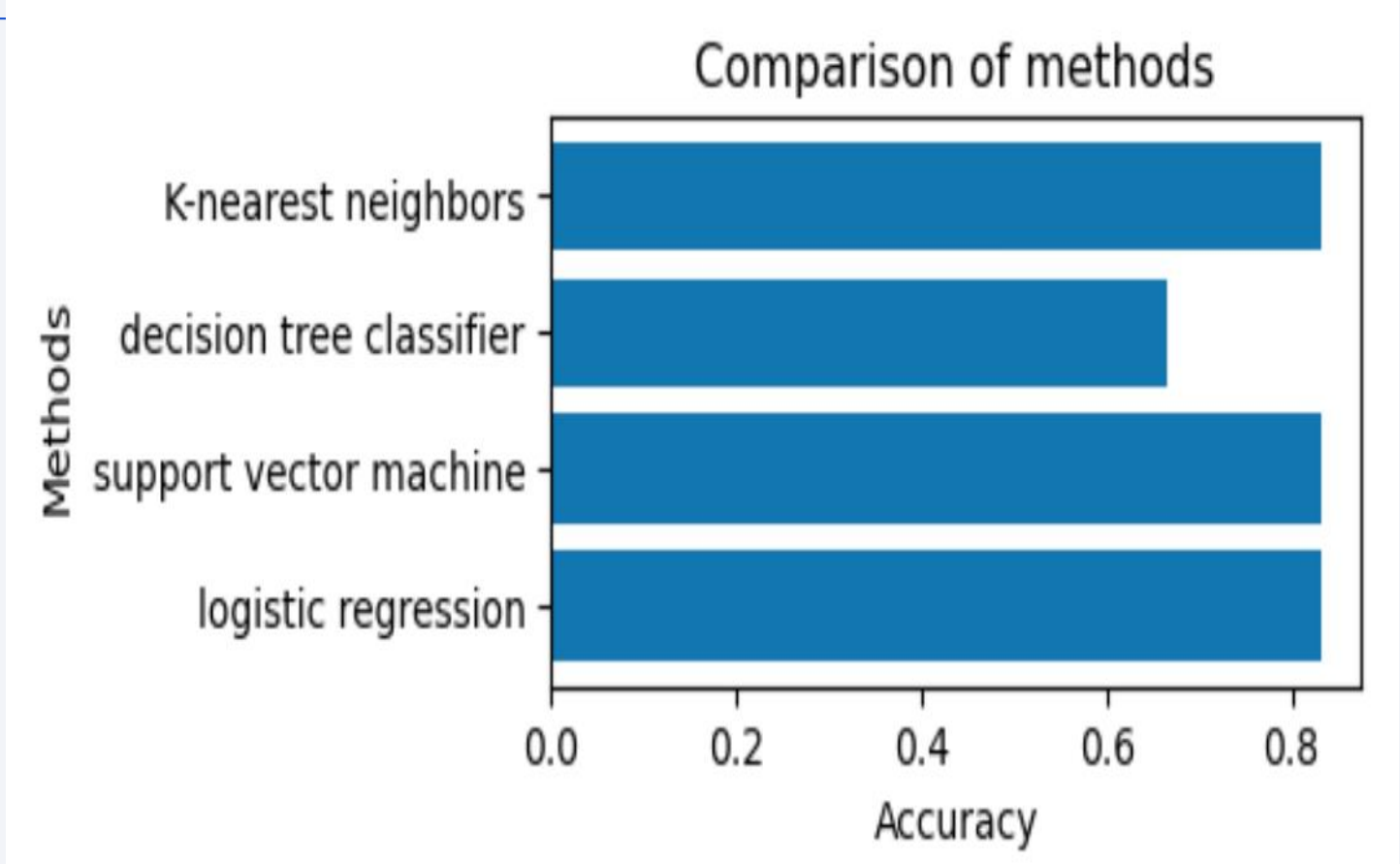
High payload

Section 5

Predictive Analysis (Classification)

Classification Accuracy

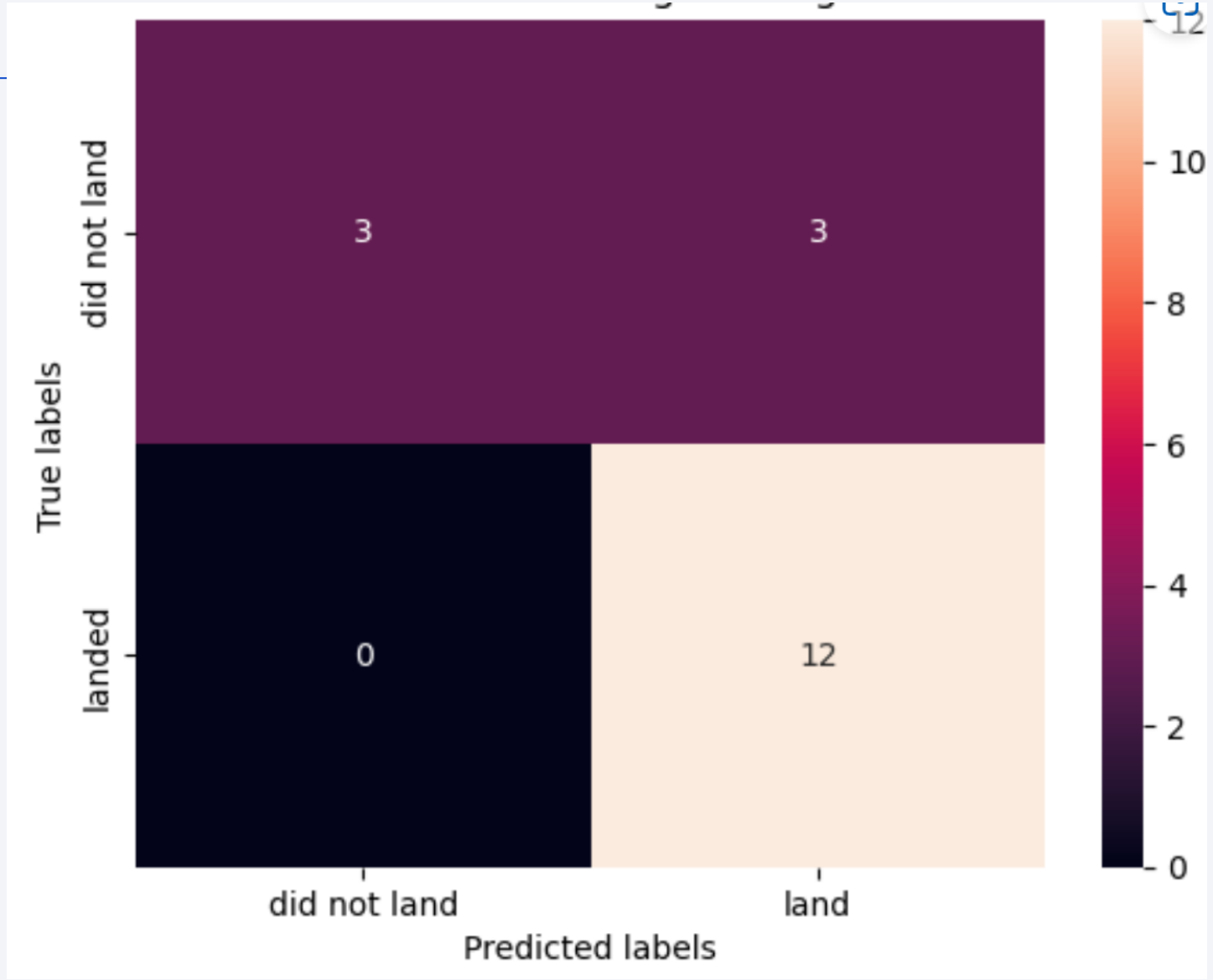
- The accuracy of the three methods such as K-nearest neighbors, support vector machine and logistic regression was almost similarly high around 0.83
- The accuracy of the decision tree classifier was a little lower around 0.67.



- The GitHub public repository for the dashboard notebook https://gitHUB.com/paramon22/IBM_10_capstone/
- File: 10_4_2_SpaceX_ML_prediction.ipynb

Confusion Matrix

- The confusion matrix is the same for all four models.
- All models predicted 12 successful landings while 3 landings were predicted to fail.
- 3 predictions were false positive.



Conclusions

- The success rate increased as the number of flights grew.
- The success rate of the most recent flights exceeds 80%.
- Orbits SSO, HEO, GEO, and ES-L1 have the highest success rate (100%).
- KSLC-39A launch site has the highest number of successful launches.
- All launch sites are close to the coastline, roads, and railroads.
- The success rate of flights with low payloads was higher than that of the heavy payloads.
- weighted payloads.
- More data is needed to get better classification and prediction models.

Appendix

The source cose is available in the GitHub public repository at https://gitHUB.com/paramon22/IBM_10_capstone/

10_1_1_Falcon_Data_Collection_API.ipynb

10_1_2_Falcon_Data_Web_Scrapping.ipynb

10_1_3_Falcon_Data_Wrangling.ipynb

10_2_Falcon_EDA_SQL.ipynb

10_3_1_Launch_Site_Location.s_Folium.ipynb

10_3_2_Falcon_data_visualization.ipynb

10_3_3_SpaceX_da shboard.ipynb

10_4_1_Falcon_EDA_DF.ipynb

10_4_2_SpaceX_ML_prediction.ipynb

Thank you!