IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

EPIFANIO SUFRIR
July 03, 2022

Outline

- Executive Summary.....03
- Introduction.....04
- Methodology.....05
- Results.....17
- Insight Drawn from EDA.....18
- Launch Sites Proximity Analysis.....35
- Build Dashboard with Plotly Dash.....39
- Predictive Analysis(Classification)….....43
- Conclusion.....46
- Appendix.....47

Executive Summary

- Methodologies used to analyze the collected data:
  - Data Collection using web scraping and SpaceX API.
  - Machine Learning Prediction using data test split analysis.
  - Exploratory Data Analysis (EDA), including data wrangling, data visualization and interactive visual analysis.

- Summary of all results
  - The collected data has been possible from readily available public sources.
  - EDA easily identified which features are the best to predict successful launches.
  - Machine Learning Prediction showed the best model to predict successful launches using all collected data.

Introduction

- Project background

  SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62M dollars; other providers cost upward of 165M dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

- Problems needed to find answers
  - What influence the success of first-stage rocket landing?
  - What are the variables that will impact in determining the success rate of first-stage rocket landing?
  - What are the condition does SpaceX have to achieve to get the best result and ensure the best landing success rate.

Section 1

Methodology

Methodology

- Data collection methodology:
  - Data from SpaceX was obtained from the following:
    - SpaceX API (https://api,spacexdata.com/v4/rockets/)
    - Web Scraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)
- Perform data wrangling
  - One Hot Encoding data fields for Machie Learning and dropping irrelevant columns
- Perform exploratory data analysis (EDA) using visualization and SQL

Methodology

- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data that was collected until this step were normalized, divided in training and test datasets and evaluated using 4 different classification models. Accuracy of each model were evaluated using various combinations of parameters and selected the best suited model.

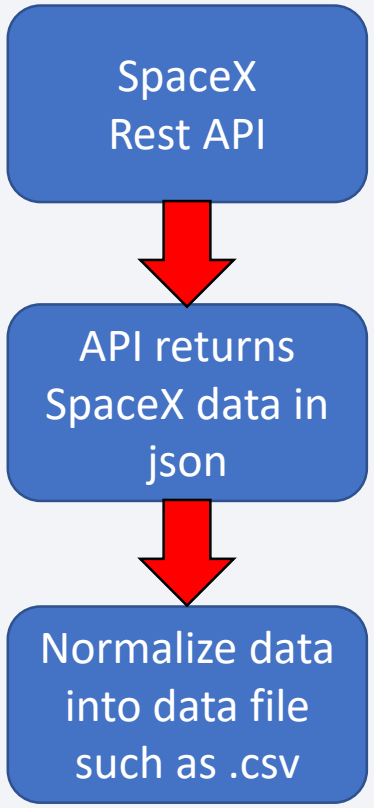
Data Collection

- Datasets were collected from SpaceX API (https://api.spacexdata.com/v4/rockets/) and from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon_Heavy_launches), using web scraping method.

Data Collection – SpaceX API

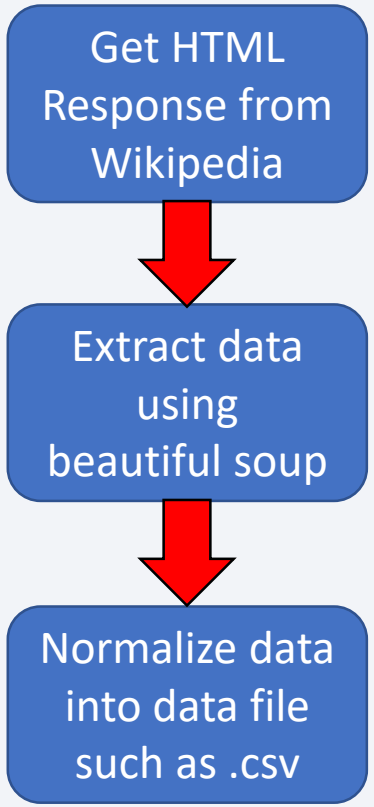
- SpaceX provides a public API from where data can be obtained, analyzed and used.
- This API was used according to the flowchart shown.

- Source code: https://github.com/donatello0214/IBM_Data_Science_SpaceX_Capstone-_Project/blob/main/Final%20Capstone%20Project%20-%20Data%20Collection%20API.ipynb



Data Collection - Scraping

- Data for SpaceX launches can be obtained from Wikipedia.
- Data downloaded from Wikipedia were processed according to the flowchart shown.

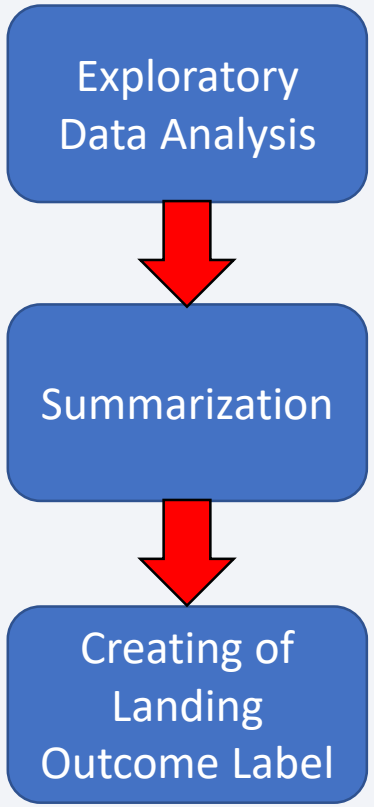


- Source code: https://github.com/donatello0214/IBM_Data_Science_SpaceX_Capstone-_Project/blob/main/Data%20Science%20-%20Data%20Collection%20with%20Web%20scraping.ipynb

Data Wrangling

- Performed Exploratory Data Analysis EDA on Dataset.
- Calculated the number of launches at each site, occurrence of each orbit as well as the occurences of mission outcome per orbit type.
- Create a landing outcome label from outcome column.

- Source code: https://github.com/donatello0214/IBM_Data_Science_SpaceX_Capstone-_Project/blob/main/Aplied%20Data%20Science%20-%20Data%20Wrangling.ipynb



EDA with Data Visualization

- Scatter Plot was used to show how much one variable is affected by the other. The relationship between two variables correlates to each other. Scatter Plots usually consists of large body of data.
- Bar Diagram makes it easy to compare sets of data between different groups. The graph represents categories on one axis and a discrete valus in the other. The goal is to show the relationship between the two axe. Bar charts can also show big changes in the data over time.
- Line Graphs are very useful because they show data variables and trends very clearly and can help to make predictions more accurately.
- Source code: https://github.com/donatello0214/IBM_Data_Science_SpaceX_Capstone-_Project/blob/main/Applied%20Data%20Science%20-%20EDA%20with%20Visualization.ipynb

EDA with SQL

- The following SQL queries were performed
  - Display names of the unique launch sites in the space mission.
  - Display 5 launch sites whose name begin with the string 'CCA'.
  - Display Total payload mass carried by boosters launced by NASA (CRS).
  - Display Average Payload mass carried by booster version F9 v1.1.
  - Display date when the first successful landing outcome in ground pad was achieved.
  - Display names of boosters which have success in drone ship and have payload mass between 4000 and 6000 kg.
  - Display total number of successful and failed mission outcome.
  - Display name of the booster versions which have carried the maximum payload mass.
  - Display failed landing outcome in drone ship, their booster versions, and launch site names for year 2015.
  - Display rank of landing outcomes between dates 2010-06-04 to 2017-03-20.

- Source code: https://github.com/donatello0214/IBM_Data_Science_SpaceX_Capstone-_Project/blob/main/jupyter-labs-eda-sql-coursera.ipynb

Build an Interactive Map with Folium

a. Visualize the launch data into an interactive map.

- Latitude and Longitude coordinate were took at each launch site and added a **circle marker** around with a label name of the launch site.
- Assigned the data from launch outcome to classes 0 and 1 with green and red markers on the map.
- Using Haversine formula, distance from launch site to various landmarks were calculated to find and marked to see the viability of logistics for each launch site.

- Source code: https://github.com/donatello0214/IBM_Data_Science_SpaceX_Capstone-_Project/blob/main/Visualization%20with%20Folium.ipynb

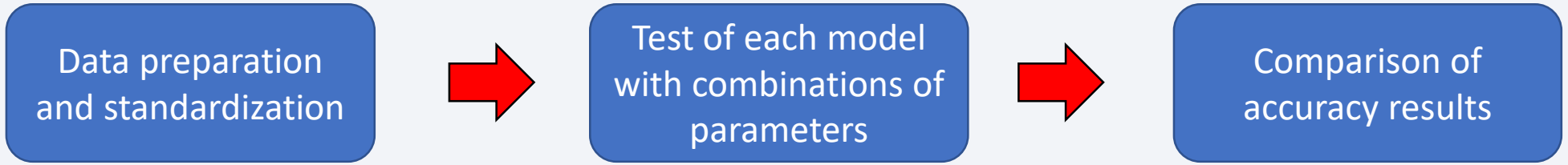
Build a Dashboard with Plotly Dash

- The following graphs and plots were used to visualize data
  - Pie Charts showing the total launches by a certain launch site/all launch site.
  - Scatter Graph showing the relationship between Outcome and Payload Mass for different Booster version.
- This combination allowed to quickly analyze the relation between payloads and launch sites, helping to identify the ideal location to launch according to payloads.

- Source code: https://github.com/donatello0214/IBM_Data_Science_SpaceX_Capstone-_Project/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Logistic Regression, Support Vector Machine, Decision Tree and K-Nearest Neighbor were the Classification models used and compared to find which Classification model will give more accurate result.

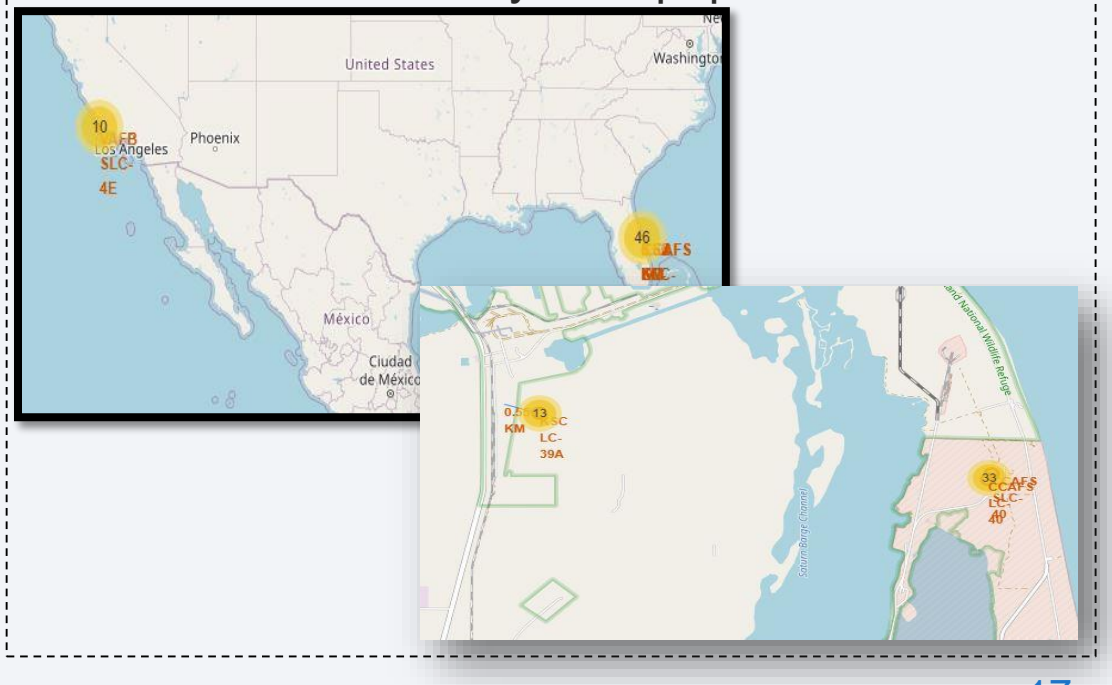


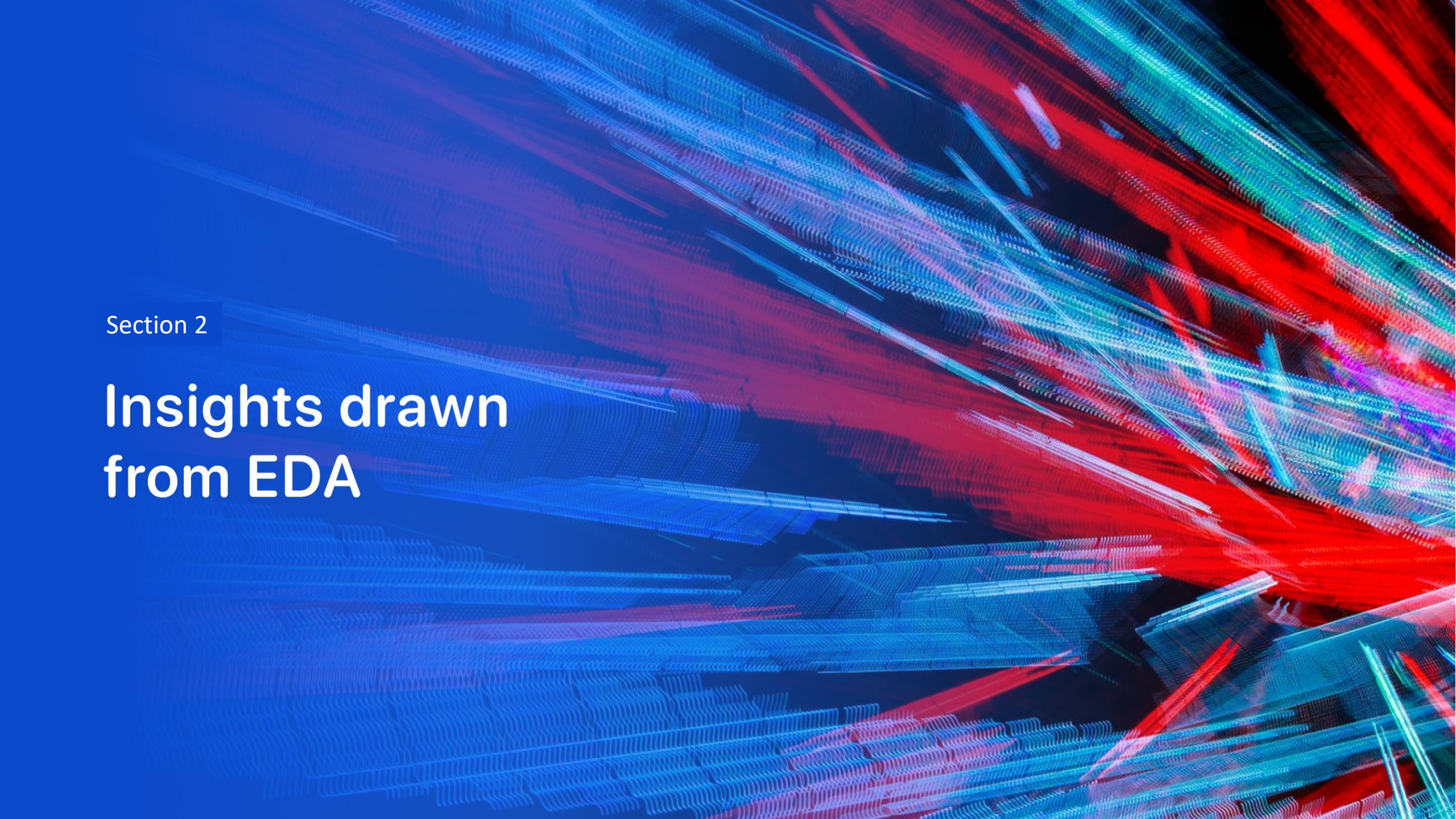
- Source code: https://github.com/donatello0214/IBM_Data_Science_SpaceX_Capstone-_Project/blob/main/Final%20Capstone%20week%204%20-%20ML%20Prediction.ipynb

Results

- Exploratory data analysis results
  - Average payload of F9 v1.1 booster is 2,928 kg.
  - The first launch were done in SpaceX and NASA.
  - The first successful landing happened in 2015, 5 years after the first launch.
  - The number of landing outcome became better as years passed.
  - Almost 100% of mission outcome were successful.
- Predictive analysis results showed that the Decision Tree Classifier is the best model to predict successful landing, having an accuracy over 87%.

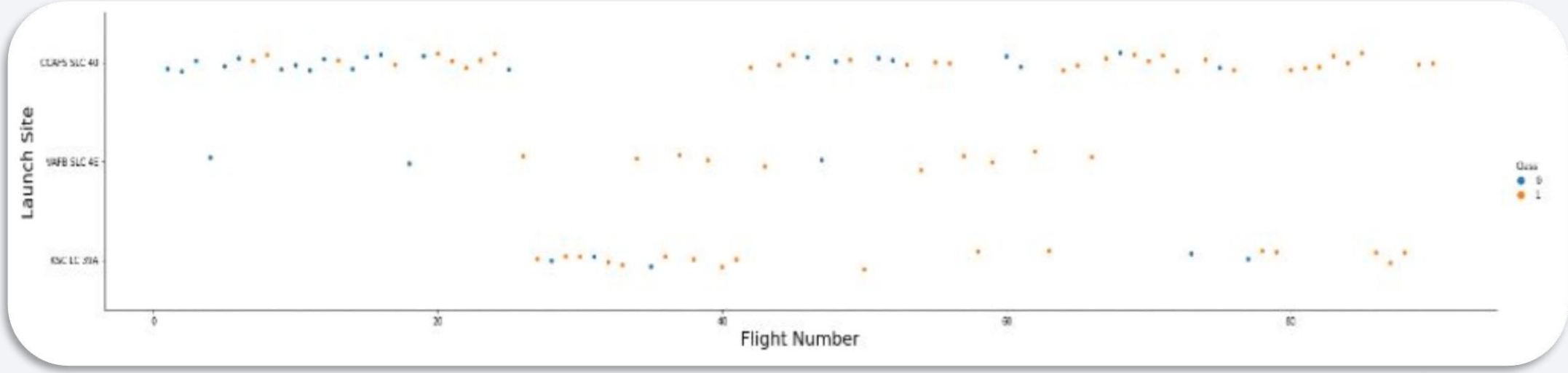
- Interactive analytics
  - Using interactive analytics, it was possible to identify that launch site were having good logistic location, near the coastal line and far away from populated area.




Section 2
Insights drawn from EDA

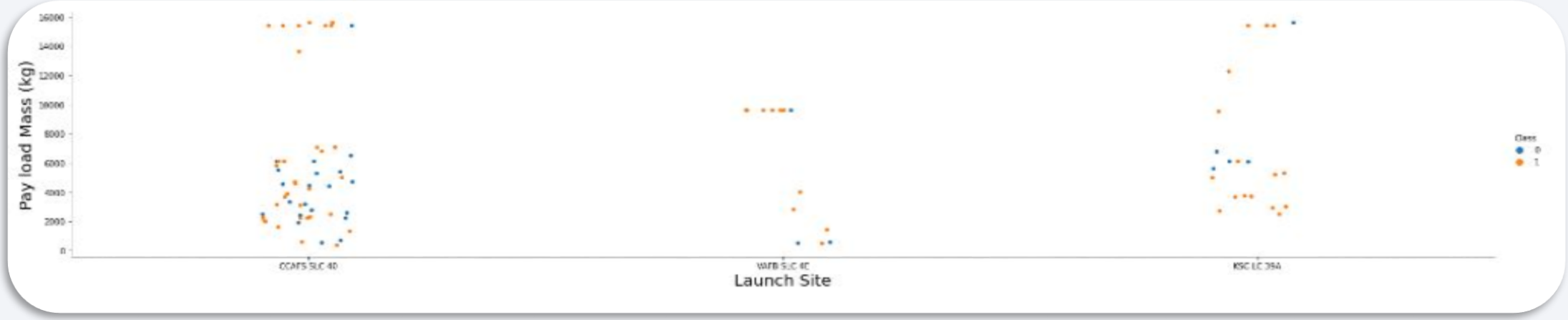
Flight Number vs. Launch Site

- The more flights at a launch site, the greater the success rate it gets.
- The plot shows that the best launch site is CCAF5 SLC 40, where most of recent launches were successful.



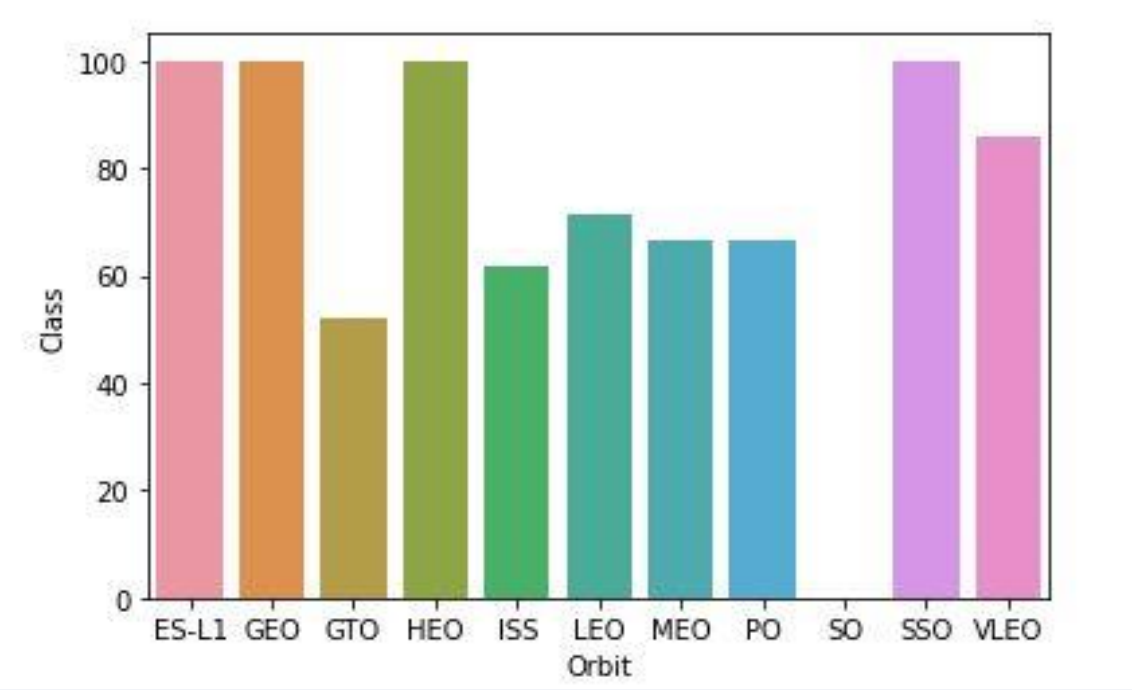
Payload vs. Launch Site

- Payloads over 9000 kg have excellent success rate.
- Payloads over 12000 kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.



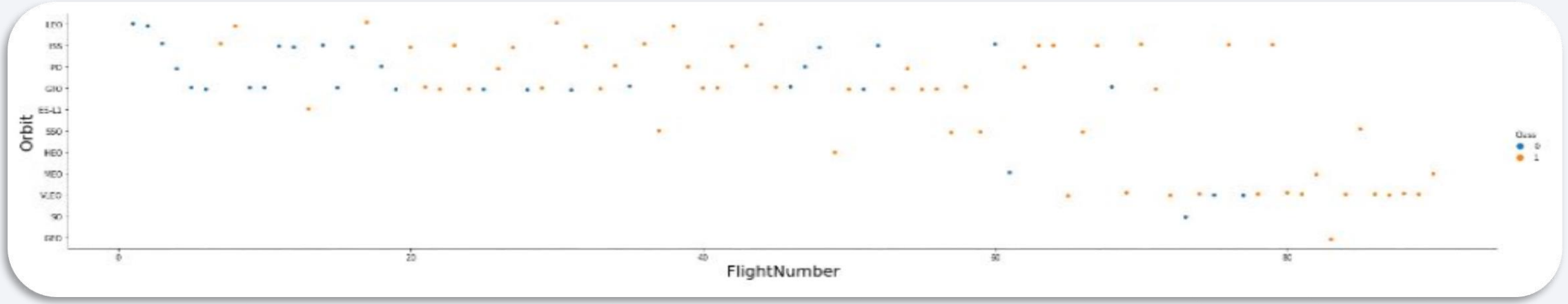
Success Rate vs. Orbit Type

- Orbit GEO, HEO, SSO, ES-L1 has the best success rate and SO having 0 success rate.



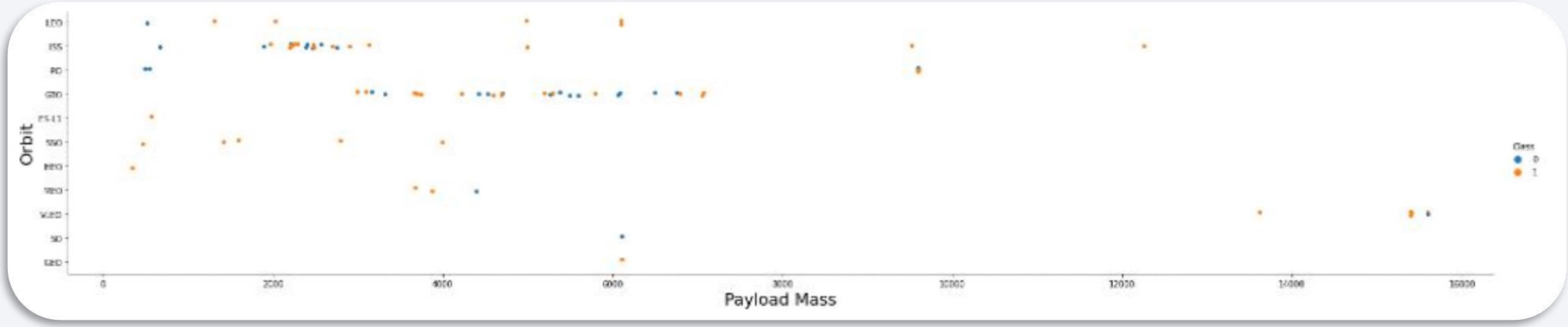
Flight Number vs. Orbit Type

- It is evident that in the LEO orbit the success rate appears to be related to the number of flights. On the other hand, there seems to be no relationship between flight number when in GTO orbit.



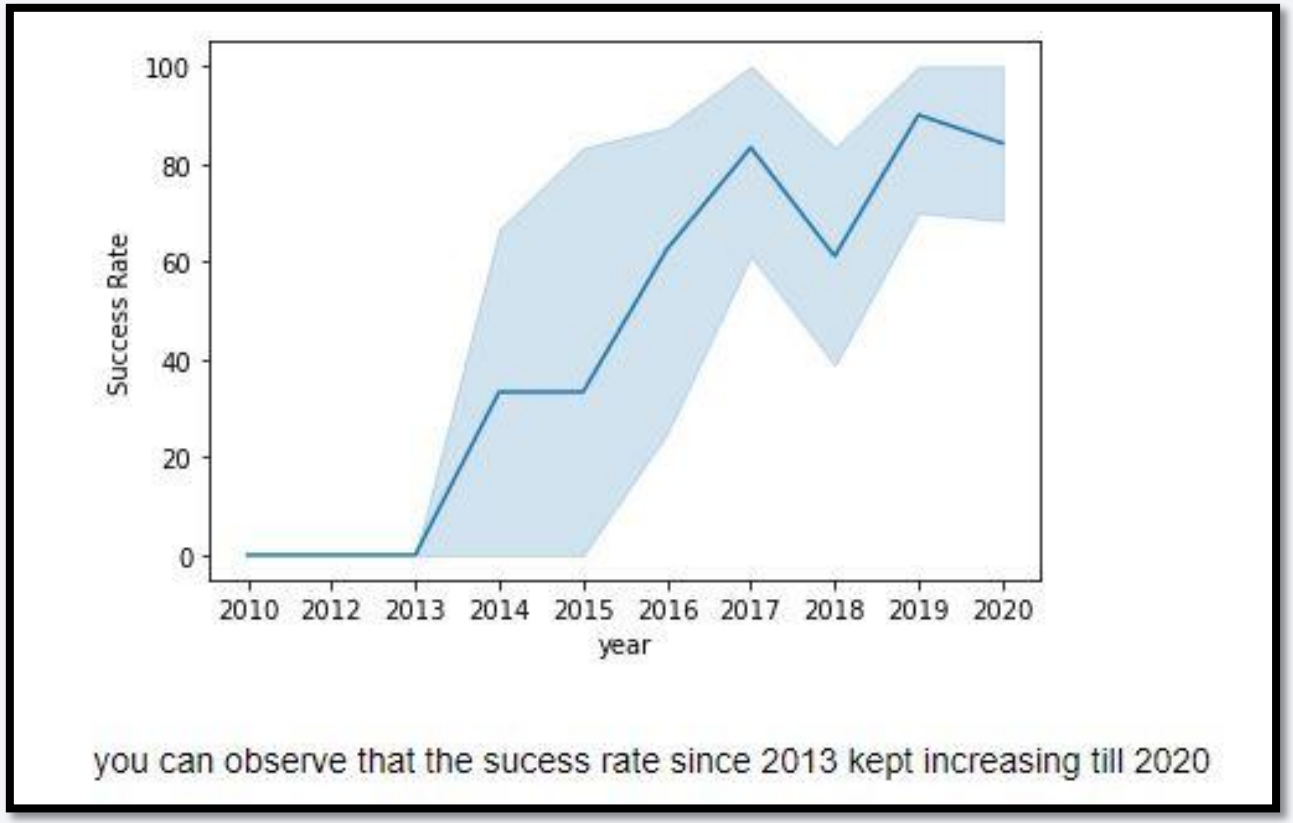
Payload vs. Orbit Type

- The graph shows no relation between payload and success rate to orbit GTO
- ISS orbit has the widest range of payload and good rate of success.
- Few launches are evident to the orbits SO and GEO.



Launch Success Yearly Trend

- Success rate started increasing in 2013 and kept until 2020;
- Success in recent years at around 80%.



All Launch Site Names

- The data showed that there are 4 launch sites as follows below:
- Query was obtained by selecting unique occurrences of "launch_site" values from the database.

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

```
[6]: %sql SELECT DISTINCT launch_site FROM SPACEXDATASET
 * ibm_db_sa://xps79436:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66
Done.
```

[6]:

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

Launch Site Names Begin with 'CCA'

- There were 5 recorded launch sites that begin with `CCA`.
- Here we can see 5 samples of launches that had landing failure or no attempt at all.

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

```
[7]: %sql SELECT * FROM SPACEXDATASET WHERE launch_site LIKE 'CCA%' LIMIT 5
```

```
 * ibm_db_sa://xps79436:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/BLUDB
Done.
```

[7]:

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

- Total payload were calculated below by summing up all payloads whose having 'CRS', which correlate with NASA.

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

```
[8]: %sql SELECT SUM(payload_mass__kg_) AS SUM FROM SPACEXDATASET WHERE customer = 'NASA (CRS)'
```

```
 * ibm_db_sa://xps79436:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/BLUDB
Done.
```

[8]:

SUM
45596

Average Payload Mass by F9 v1.1

- Data were filtered by the booster version and calculated the average payload mass as shown below.

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

```
[9]: %sql SELECT AVG(payload_mass__kg_) AS AVG FROM SPACEXDATASET WHERE booster_version LIKE 'F9 v1.1%'
 * ibm_db_sa://xps79436:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appd
Done.
```

[9]:

AVG
2534

First Successful Ground Landing Date

- Data were filtered by successful landing outcome on ground pad and getting the minimum date value possible to identify the very first occurrence as shown below.

```
[10]: %sql SELECT MIN(date) AS DATE FROM SPACEXDATASET WHERE landing__outcome LIKE 'Success%'

 * ibm_db_sa://xps79436:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.
Done.

[10]:      DATE
      ----------
      2015-12-22
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- This query returns the 4 booster versions that had successful drone ship landings and a payload mass between 4000 and 6000 non-inclusively.

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
[11]: %sql SELECT DISTINCT booster_version FROM SPACEXDATASET WHERE (payload_mass__kg_ BETWEEN 4000 AND 6000) AND (landing__outcome = 'Success (drone ship)')
```

```
 * ibm_db_sa://xps79436:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/BLUDB
Done.
```

[11]:

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

Total Number of Successful and Failure Mission Outcomes

- This query returns a count of each mission outcome.
  SpaceX appears to achieve its mission outcome nearly 99% of the time.

List the total number of successful and failure mission outcomes

```
[12]: %sql SELECT mission_outcome, COUNT(mission_outcome) AS COUNT FROM SPACEXDATASET GROUP BY mission_outcome
```

```
 * ibm_db_sa://xps79436:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/BLUDB
Done.
```

[12]:

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

Boosters Carried Maximum Payload

- This query returns the booster versions that carried the highest payload mass of 15600 kg.
- These booster versions are very similar and all are of the F9 B5 B10xx.x variety.

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

```
[13]: %sql SELECT DISTINCT booster_version, payload_mass__kg_ FROM SPACEXDATASET WHERE payload_mass__kg_ = (SELECT max(payload_mass__kg_) FROM SPACEXDATASET)
 * ibm_db_sa://xps79436:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31505/BLUDB
Done.
```

[13]:

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

2015 Launch Records

- This query returns the Month, Landing Outcome, Booster Version, Payload Mass and Launch site of 2015 launches where stage 1 failed to land on a drone ship which had two such occurrences.

```
[14]: %sql SELECT landing__outcome, booster_version, launch_site, DATE FROM SPACEXDATASET WHERE (landing__outcome LIKE 'Failure (drone ship)') AND (DATE LIKE '2015%')
 * ibm_db_sa://xps79436:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31505/BLUDB
Done.
```

[14]:

landing__outcome	booster_version	launch_site	DATE
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	2015-01-10
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	2015-04-14

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

- This query returns a list of successful landings and between 2010-06-04 and 2017-03-20.
- There are 2 types of successful landing outcome and there were 8 successful landings in total during this time period.

```
[15]: %sql SELECT landing__outcome, COUNT(landing__outcome) AS COUNT FROM SPACEXDATASET WHERE (DATE between '2010-06-04' AND '2017-03-20') GROUP BY landing__outcome ORDER BY COUNT DESC
 * ibm_db_sa://xps79436:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31505/BLUDB
Done.
```

[15]:

landing__outcome	COUNT
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

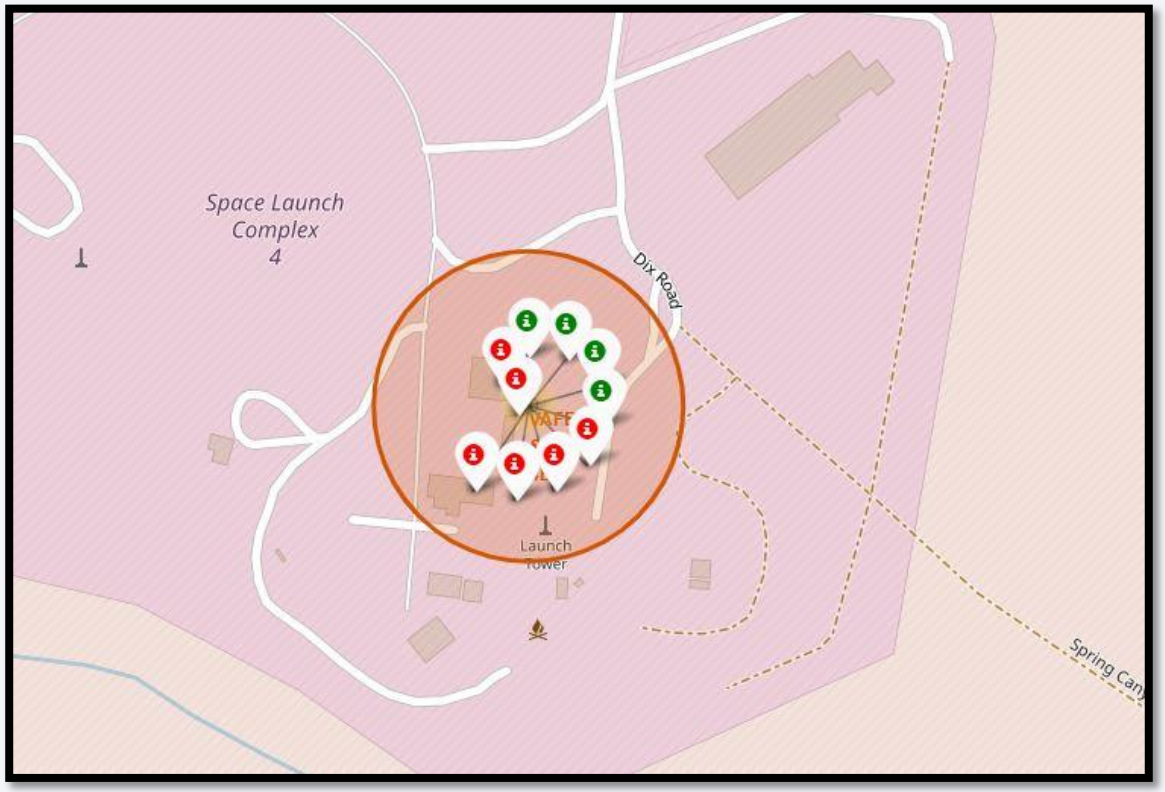
Launch Sites Proximities Analysis

Lauch Site Locations

- Lauch sites are situated logistically sound, not far from roads and railroads and mostly near the ocean but far away from inhabited areas.

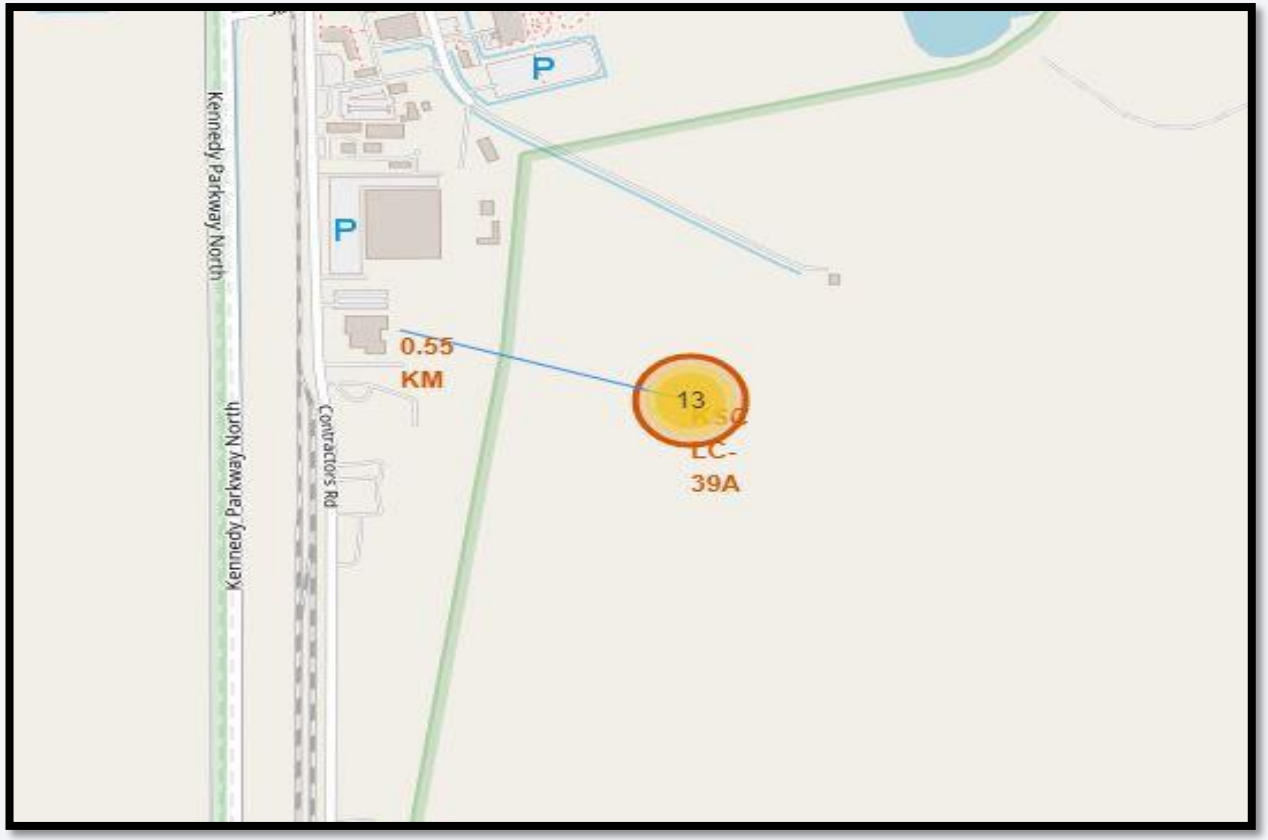
Colored-coded Launch Outcomes

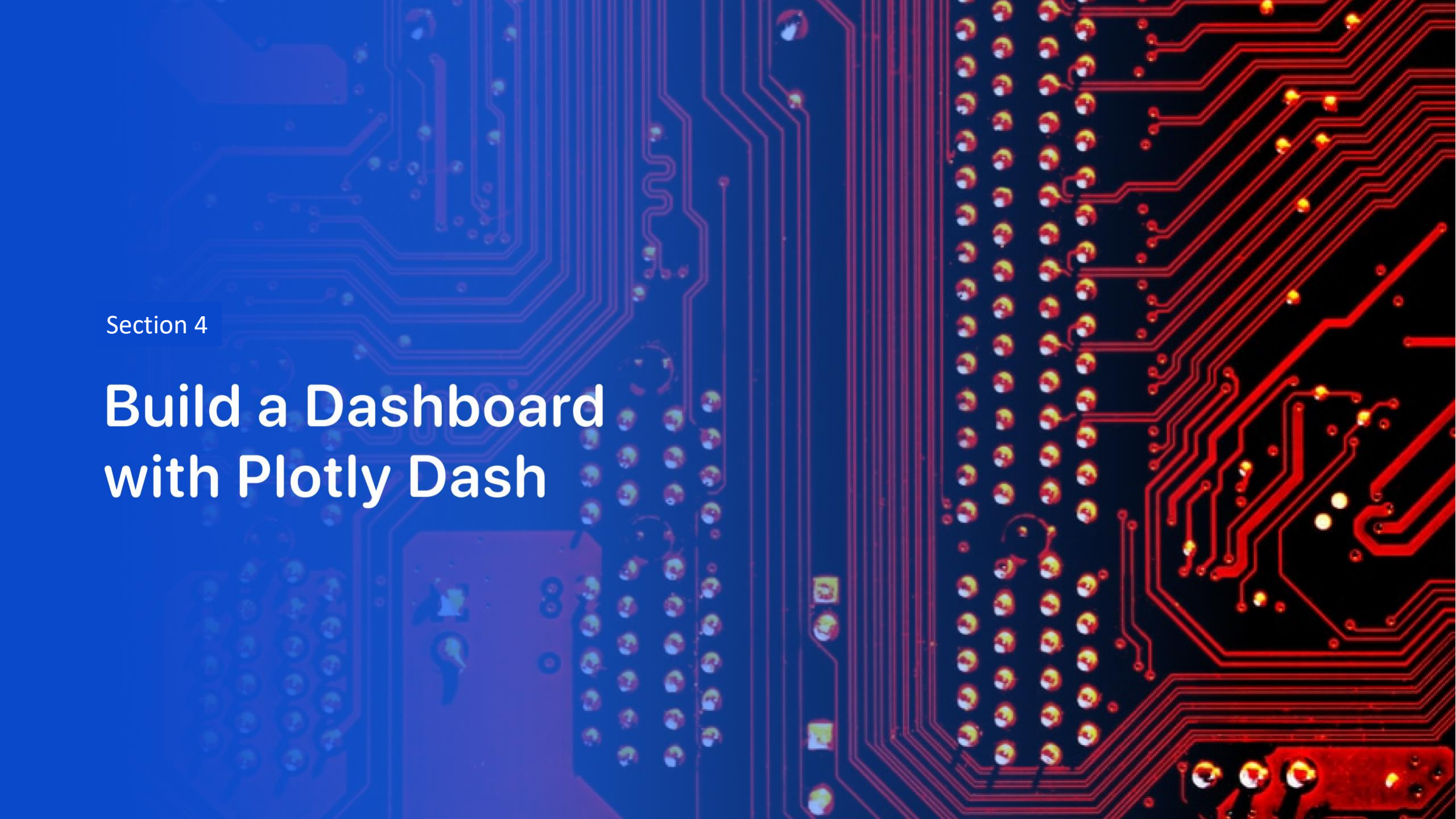
- Clusters on Folium map can be clicked on to display each successful landing (green icon) and failed landing (red icon). In this example VAFB SLC-4E shows 4 successful landings and 6 failed landings.



Logistics and Safety

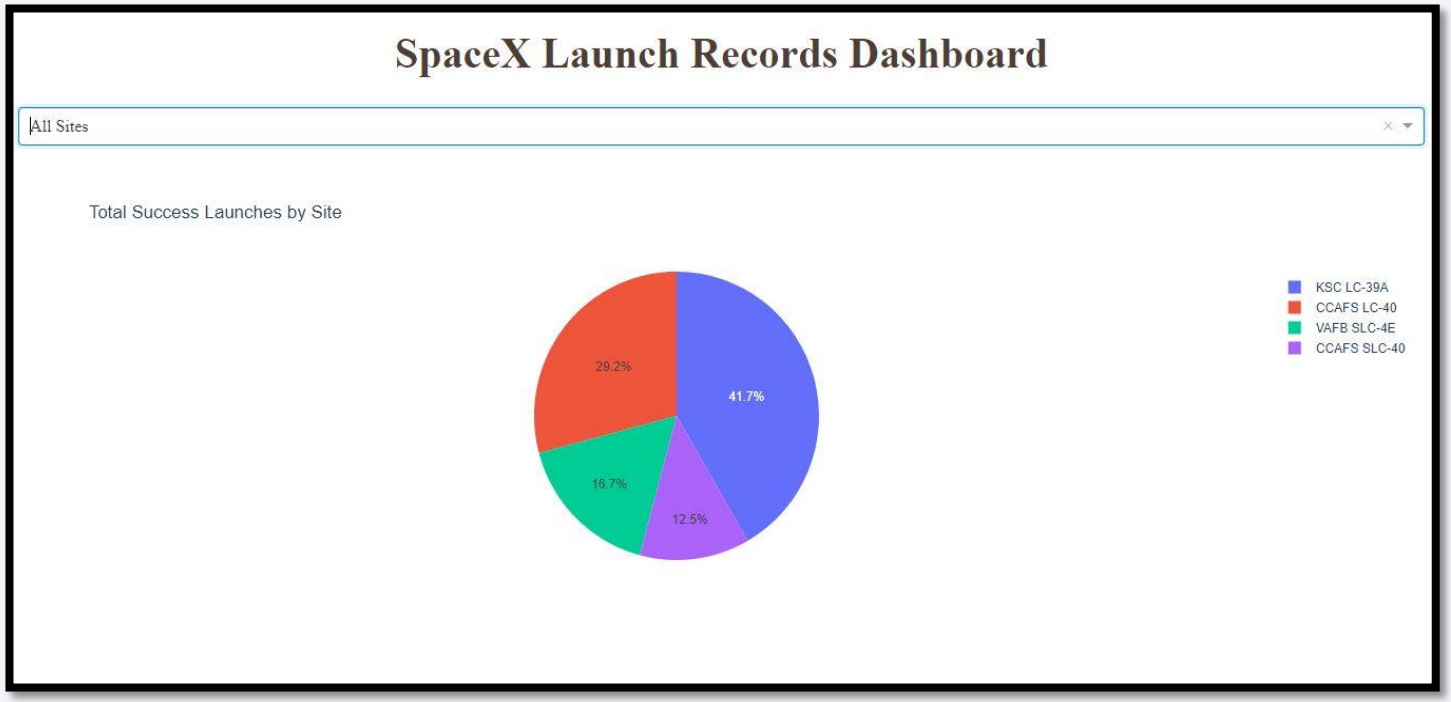
- Launch site KSC LC-39A has very logistics, being near both railroad and roads and relatively far from inhabited areas.




Section 4
Build a Dashboard with Plotly Dash

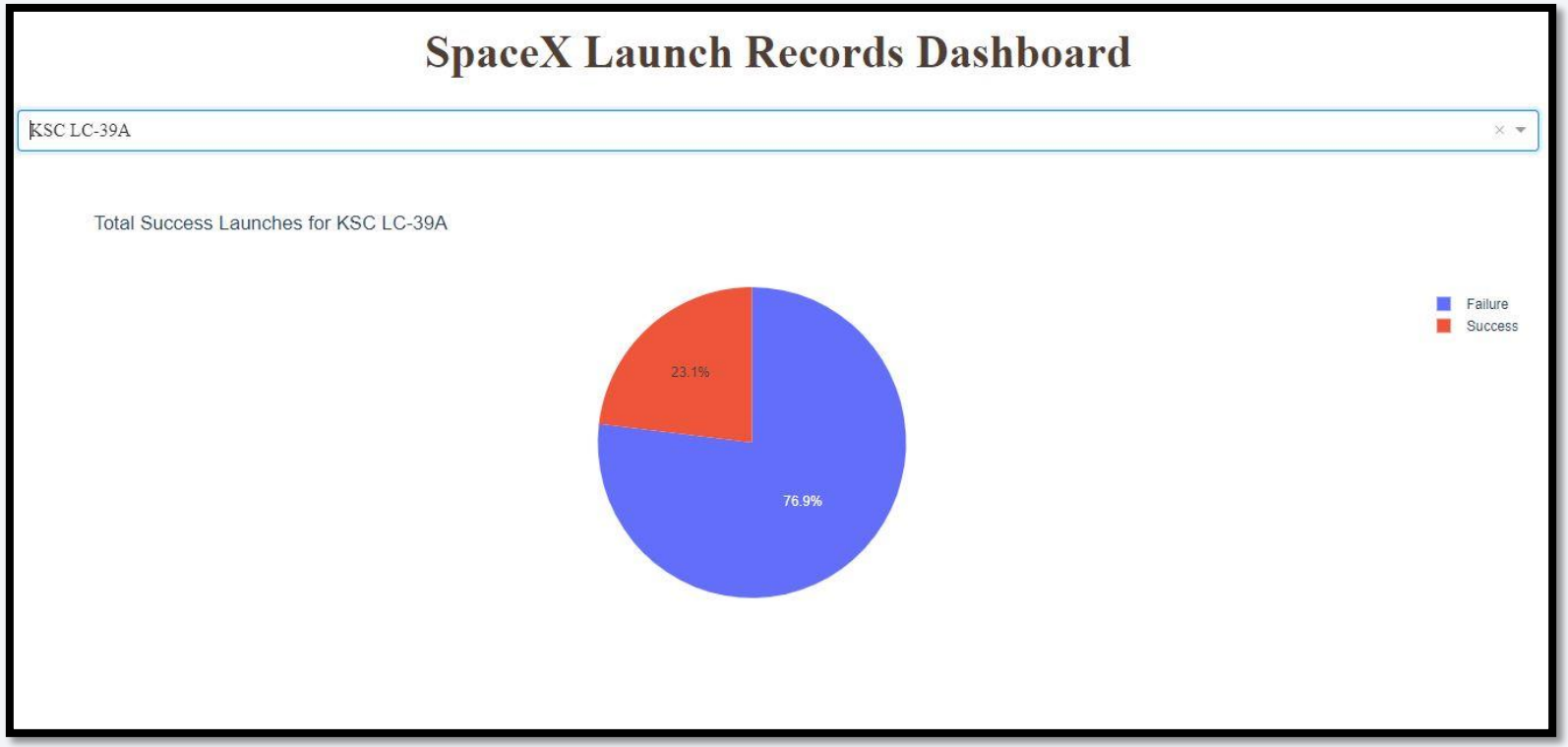
Successful Launches Across Launch Sites

- This is the distribution of successful landings across all launch sites. CCAFS and KSC have the same amount of successful landings. VAFB has the smallest percentage of successful landings. This may be due to smaller sample and increase in difficulty of launching in the west coast.



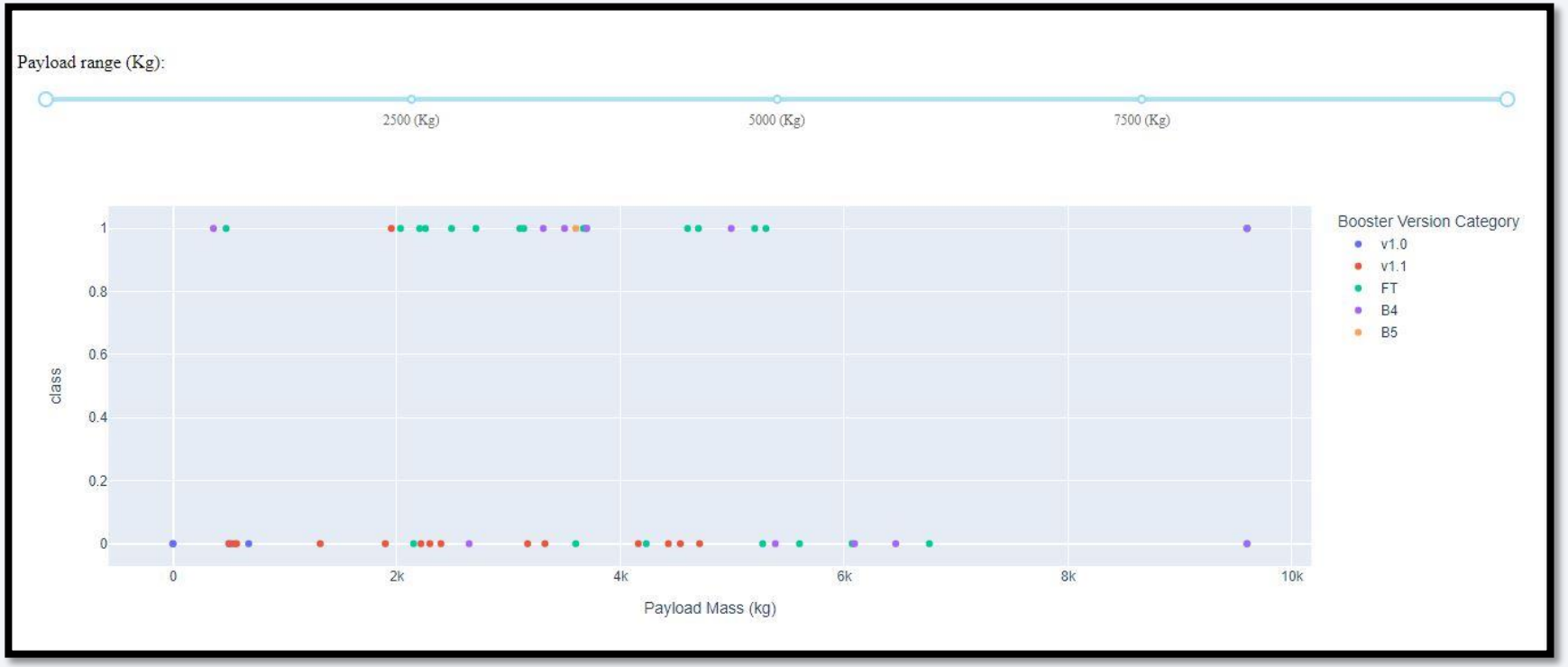
Highest Success Rate Launch Site

- KSC LC-39A has the highest success rate of 76.9% among all the launch sites.



Payload vs. Launch Outcome

- Payloads under 6,000 kg. And FT Boosters are the most successful combination as shown in the scatterplot.



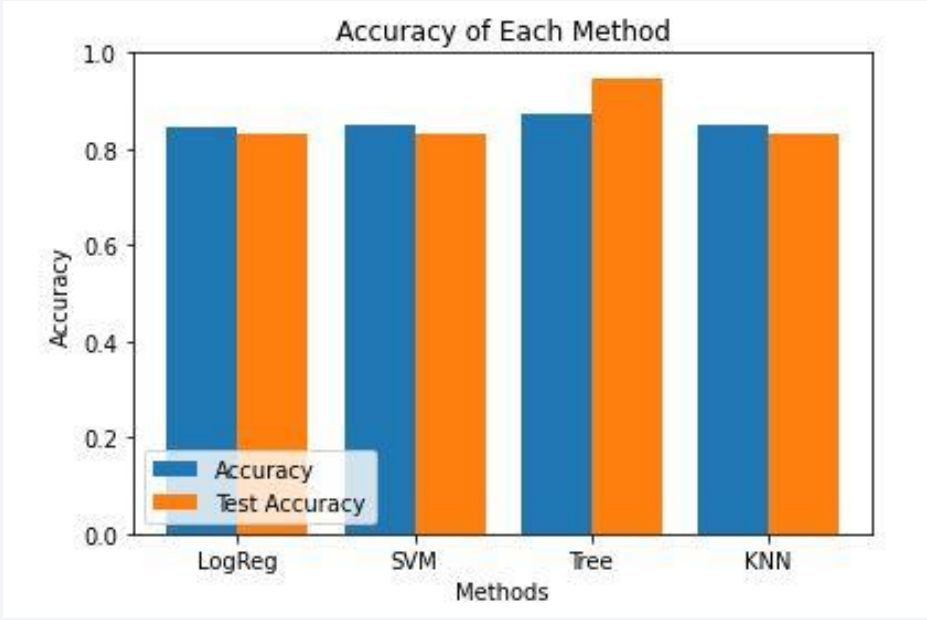
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- There are four classification models tested and the accuracies are shown in the bar chart.

- As shown in the bar graph and chart below, the model with the highest accuracy is the Decision Tree Classifier which having 87.5% accuracy and 94.4% test accuracy.

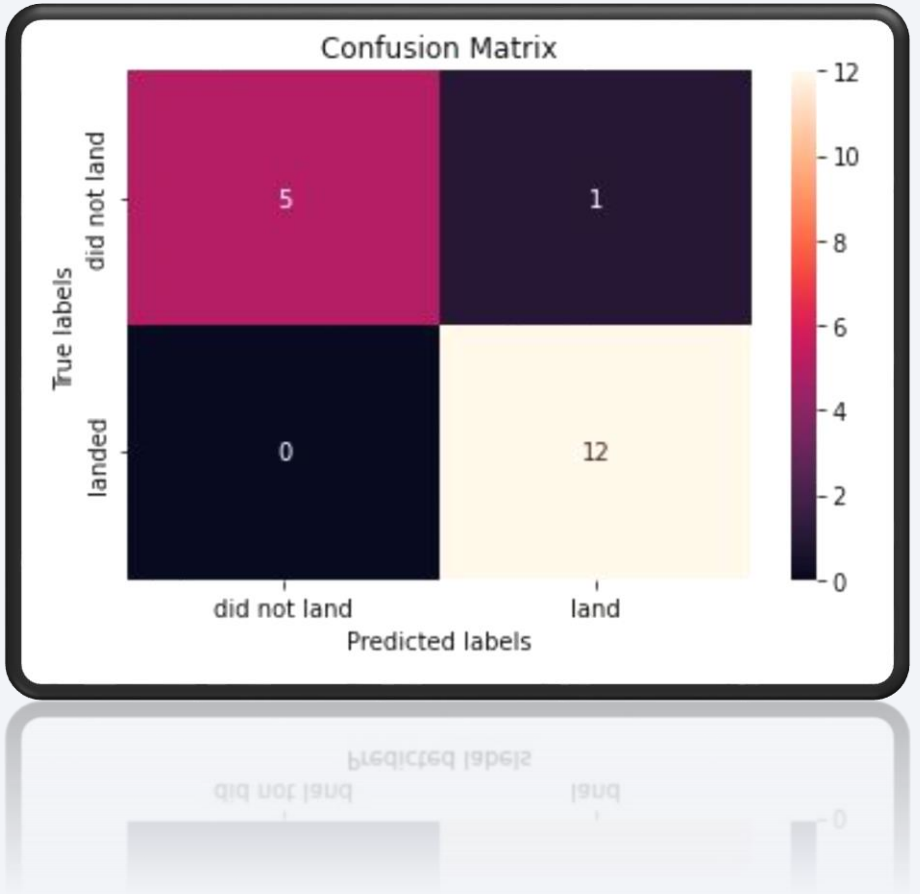


Model	Accuracy	TestAccuracy
LogReg	0.84643	0.83333
SVM	0.84821	0.83333
Tree	0.875	0.94444
KNN	0.84821	0.83333

Confusion Matrix

- Decision Tree Classifier Confusion Matrix clearly show its accuracy by the numbers shown in the true positive and true negative compared to the false one.

- Correct predictions are read diagonally from the top left down to the bottom right.



Conclusions

- Low weight payloads perform better than heavier ones.
- The success rates of SpaceX launches and landings is directly proportional to the numbers of launches and the times in years as they improve future launches.
- As the data have shown, KSC LC-39A had the most success rate of launches done from all the launch sites.
- Orbit GEO, HEO, SSO, ES-L1 has the best success rate.
- Created a machine learning model and come up with the best Classifier with an accuracy of 87.5%.
- If possible more data should be collected to better determine the best machine learning model and improve accuracy.

Appendix

- GitHub repository URL:

https://github.com/donatello0214/IBM_Data_Science_SpaceX_Capstone_Project

- Data Science Instructors:

https://www.coursera.org/professional-certificates/ibm-data-science?#instructors