IBM Developer
SKILLS NETWORK

Advancing the Space Race through Data Science

Pascal F. Meier
23.Jun.2024

1. Project Outline

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

2. Executive Summary

- Public available data sources from SpaceX API and SpaceX Wikipedia page are used for data wrangling.
- Rocket landings are sorted in an extra column “class” based on their landing outcome.
- Using SQL, visualization, Folium map and dashboards to explore the data.
- Normalizing and standardized the data. Applying GridSearchCV to find the best parameters for machine learning models.
- Accuracy score of all models are used for visualization.
- While four machine learning models were used (Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors), all of them produced similar results, predicting a successful landing with an accuracy rate of ~83.33%.
- For a better prediction and model determination of future flights and landings, more data is necessary.

3. Introduction

Y-Wing: https://overmental.com/content/star-wars-canon-catch-up-what-are-y-wings-40599

- Commercial Space Age is now!
- The governmental monopoly for space flights has been taken over by the private sector. The funding of space flights is no longer solemly in the hand of tax payers but it is business that capitalize on the opportunity of sending rockets to space.
- Reducing the price of a space flight brings an advantage to every company. Due to the ability of recover and reuse parts of the rocket, SpaceX has the best pricing model (US $62 million vs. US $165 million by there nearest competitors).
- SpaceY wants to compete with SpaceX on the bidding of future missions. How can the use of data science advance SpaceY to predict the successful recovery of Stage1 of rockets and therefore win the bid for future missions.

4. Methodology

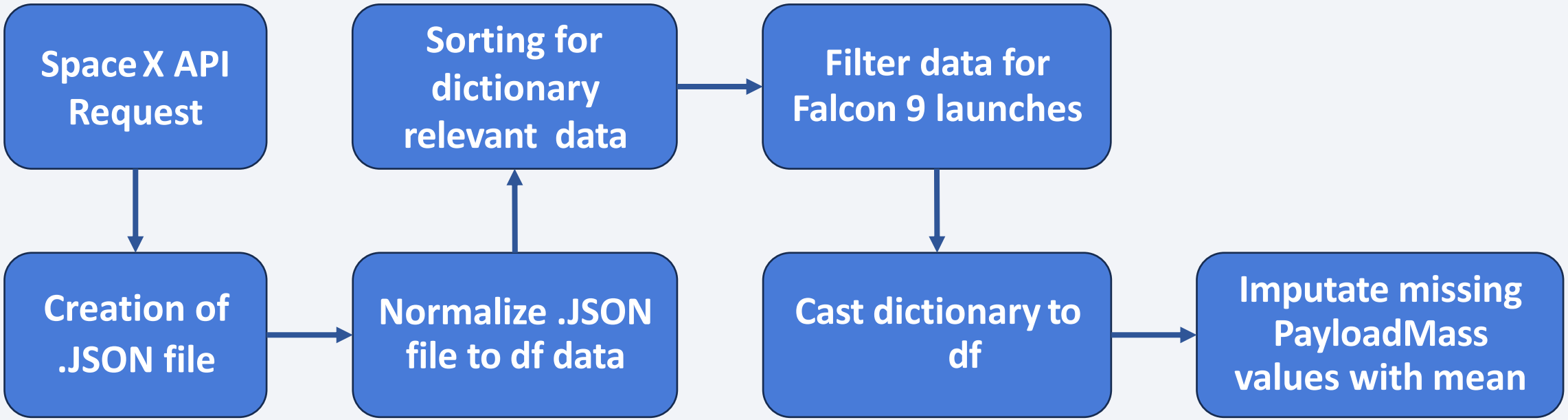
- **Data collection methodology:**
  - Combined data from SpaceX public API and SpaceX Wikipedia page
- **Perform data wrangling**
  - Classifying true landings as successful and unsuccessful otherwise
- **Perform exploratory data analysis (EDA) using visualization and SQL**
- **Perform interactive visual analytics using Folium and Plotly Dash**
- **Perform predictive analysis using classification models**
  - Tuned models using GridSearchCV

5. Data Collection

- Through an API request from SpaceX public database, data was collected. Furthermore web scrapping from SpaceX Wikipedia entry was used. Both database were combined.
- Both processes are visualized in Flow-charts in the following two slides.
- Space X API Database containing the following informations:
  FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
- Webscraping from Wikipedia Entry containing the following informations :
  Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

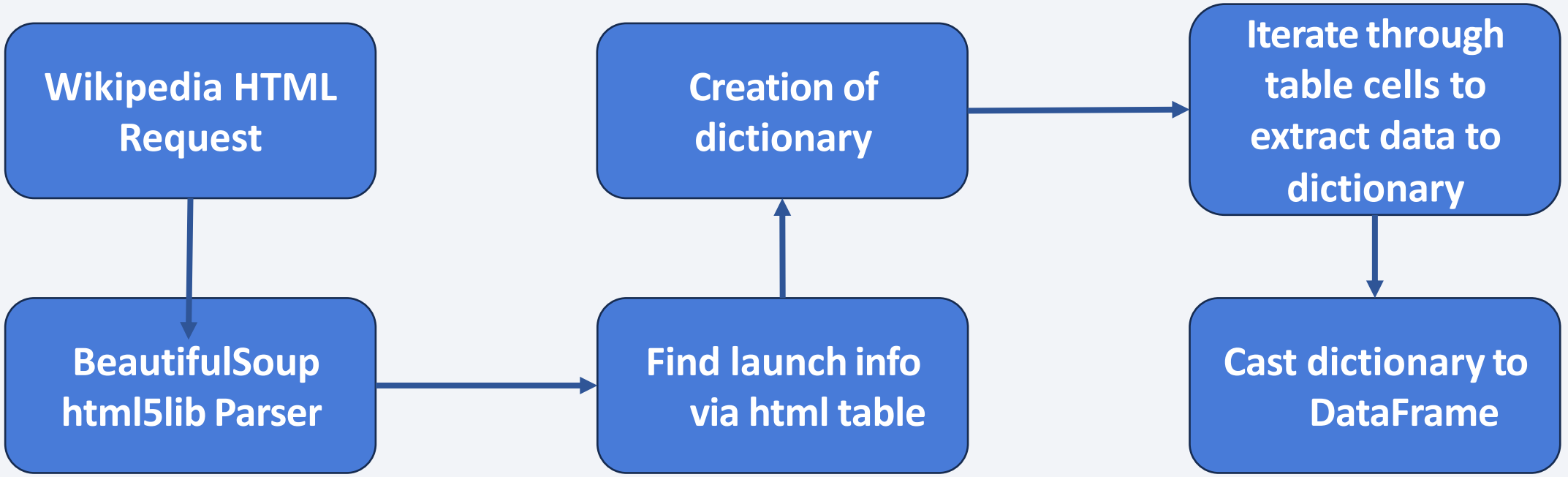
6. Data Collection – SpaceX API

- GitHub URL Reference:
  https://github.com/Mr-Minion/IBM_Data_Science_Professional_Certificate/blob/main/Applied_Data_Science_Capstone/Module1/Data_Collection_API.ipynb



7. Data Collection - WebScraping from Wikipedia

- GitHub URL Reference:
  https://github.com/Mr-Minion/IBM_Data_Science_Professional_Certificate/blob/main/Applied_Data_Science_Capstone/Module1/Data_Collection_Web_Scraping.ipynb



8. Data Wrangling

- GitHub URL Reference:
  https://github.com/Mr-Minion/IBM_Data_Science_Professional_Certificate/blob/main/Applied_Data_Science_Capstone/Module1/Data_wrangling.ipynb

- The Outcome column has two components: “Mission Outcome” and “Landing Location”.
- In order to investigate the landing outcome/class, an additional column “class” was create.
- For each mission, if the “Mission Outcome” was registered as “True”, a value “1” was added to ”class”. For failures a value “0” was added to the ”class”.

Value Mapping of landing outcomes:

- “True ASDS”, “True RTLS”, “True Ocean” were assigned a landing class value of ”1”
- “None None”, “False ASDS”, “None ASDS”, “False Ocean”, “False RTLS” were assigned a landing class value of ”0”.

9. EDA with Data Visualization

- GitHub URL Reference:
  https://github.com/Mr-Minion/IBM_Data_Science_Professional_Certificate/blob/main/Applied_Data_Science_Capstone/Module2/EDA_with_Visualization.ipynb

- Visualizing data:
- To visualize the data, the following charts were plotted to compare the relationship between variables:
- Scatter plots, line charts, bar plots
- By this means it could be decided whether a relationship exist between Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit vs. Success Rate, Flight Number vs. Orbit, Payload vs Orbit, and Success Yearly Trend. Such relationships can be used in training the machine learning model.

10. EDA with SQL

- GitHub URL Reference: https://github.com/Mr-Minion/IBM_Data_Science_Professional_Certificate/blob/main/Applied_Data_Science_Capstone/Module2/EDA_with_SQL.ipynb

- Download dataset and store it as database table in IBM DB2 Database.
- SQL Python integration is used for queries.
- Queries are a helpful too to better understand the dataset.
- The queried information contain launch site names, mission outcomes, various pay load sizes of customers and booster versions, and landing outcomes.

11. Build an Interactive Map with Folium

- GitHub URL Reference:
  https://github.com/Mr-Minion/IBM_Data_Science_Professional_Certificate/blob/main/Applied_Data_Science_Capstone/Module3/Interactive_Visual_Analytics_with_Folium.ipynb

- Folium maps display launch sites, successful and unsuccessful landings, and proximity to key locations such as railways, highways, coasts, and cities. This helps understand the strategic placement of launch sites and visualizes the success rate of landings based on location.

12. Build a Dashboard with Plotly Dash

- GitHub URL Reference:
https://github.com/Mr-Minion/IBM_Data_Science_Professional_Certificate/blob/main/Applied_Data_Science_Capstone/Module3/spacex_dash_app.py

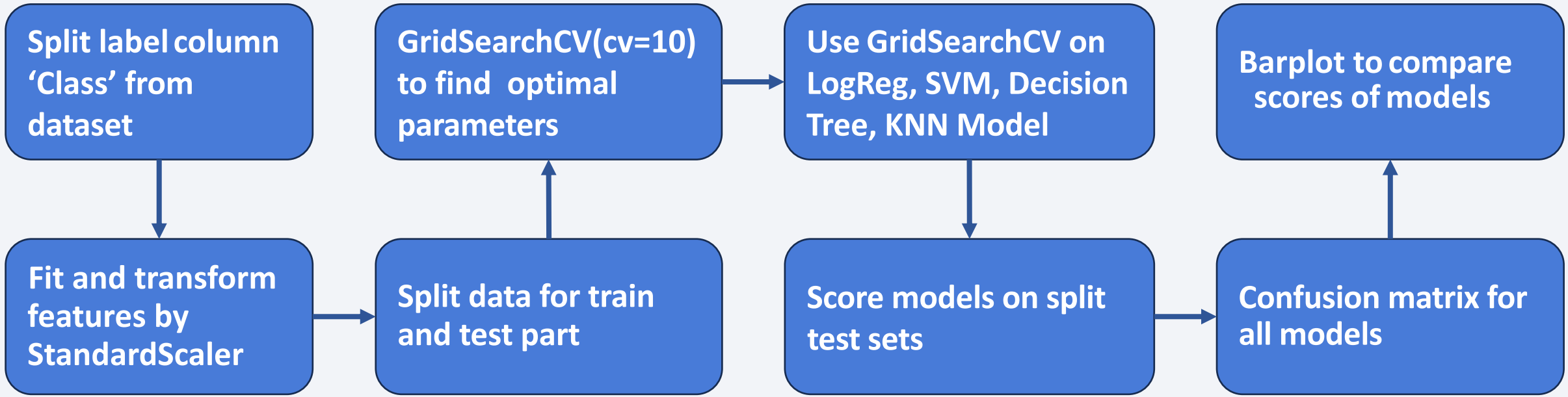
- Creation of a dashboard containing a pie chart and a scatter plot.

- In the pie chart, a selection of the distribution for successful landings across all landing sites can be made. Additionally individual launch sites can be selected to visualize their success rate.

- The scatter plot has two inputs: Sites (all or individual) and PayloadMass on a slider between 0 and 10000 kg. This can help to visualize the success rate across launch sites, payload mass, as well as booster version category.

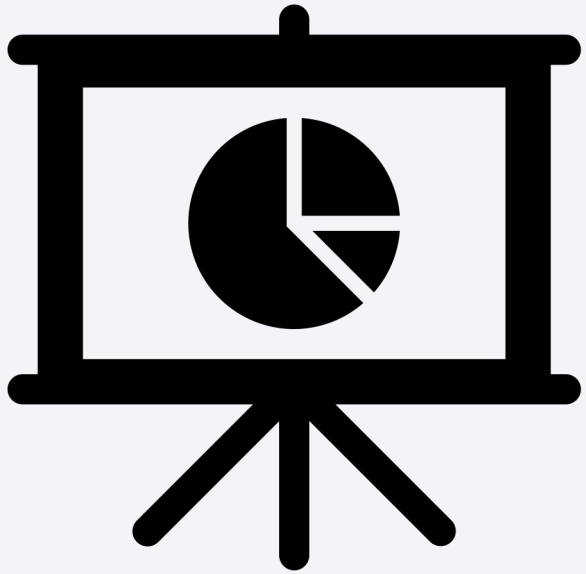
13. Predictive Analysis (Classification)

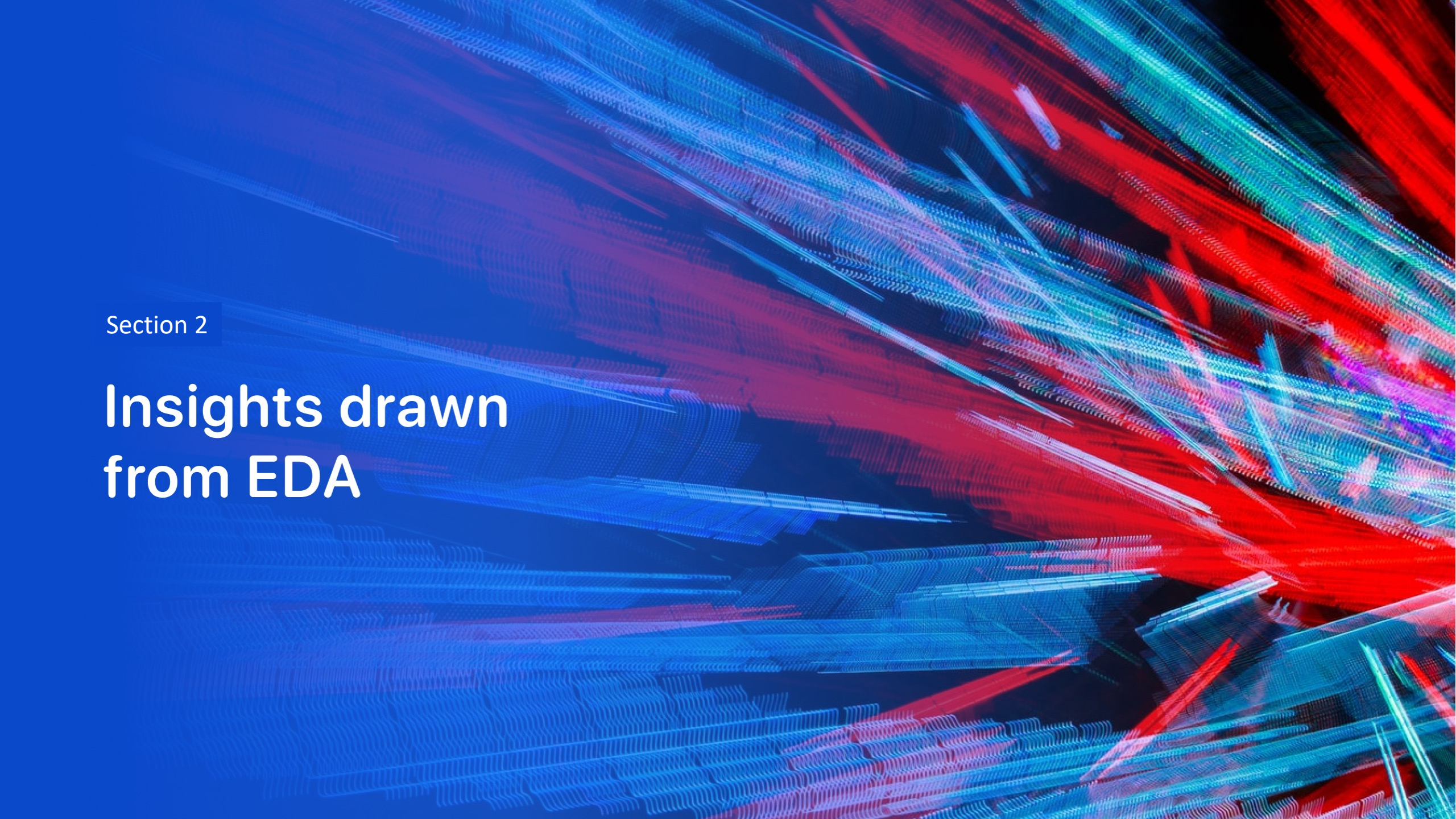
- GitHub URL Reference:
  https://github.com/Mr-Minion/IBM_Data_Science_Professional_Certificate/blob/main/Applied_Data_Science_Capstone/Module4/Machine_Learning_Prediction.ipynb



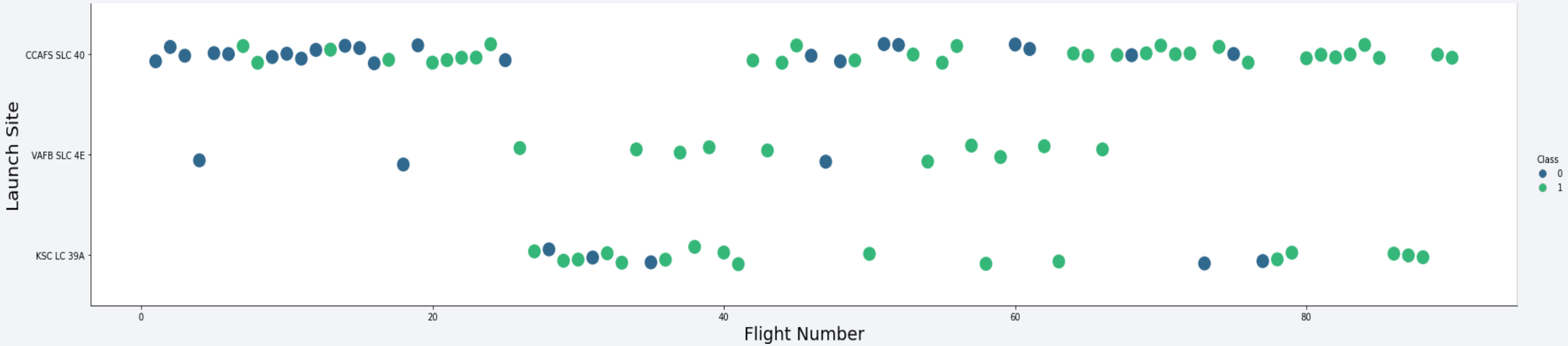
14. Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results




Section 2
Insights drawn from EDA

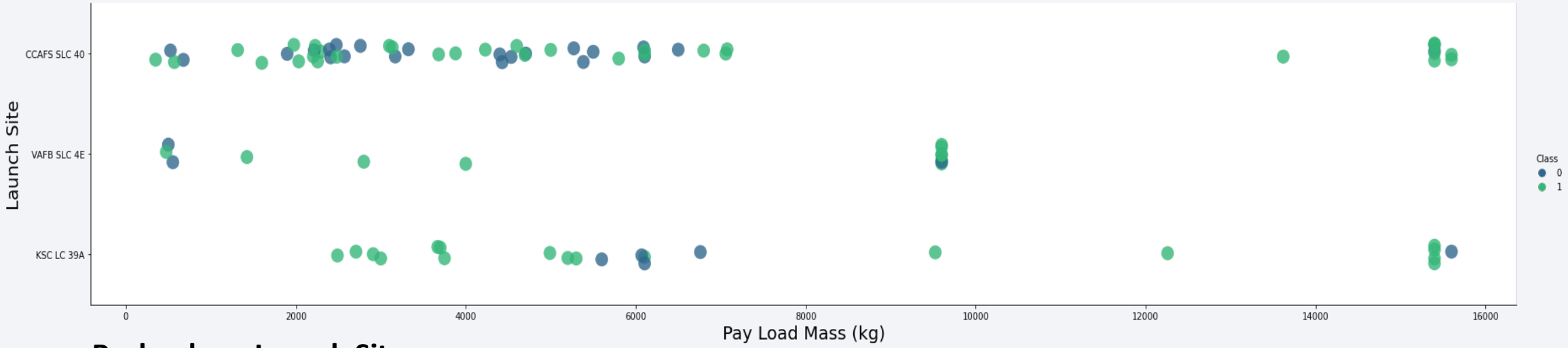
15. Flight Number vs. Launch Site



Flight Number per Launch Site:
Green indicates successful launch; blue indicates unsuccessful launch

- An increased success rate over time (by increased Flight Number) is shown.
- A breakthrough is shown at around flight 20, with a significantly increased success rate.
- CCAFS SLC40 has the most launches with a high success rate

16. Payload vs. Launch Site

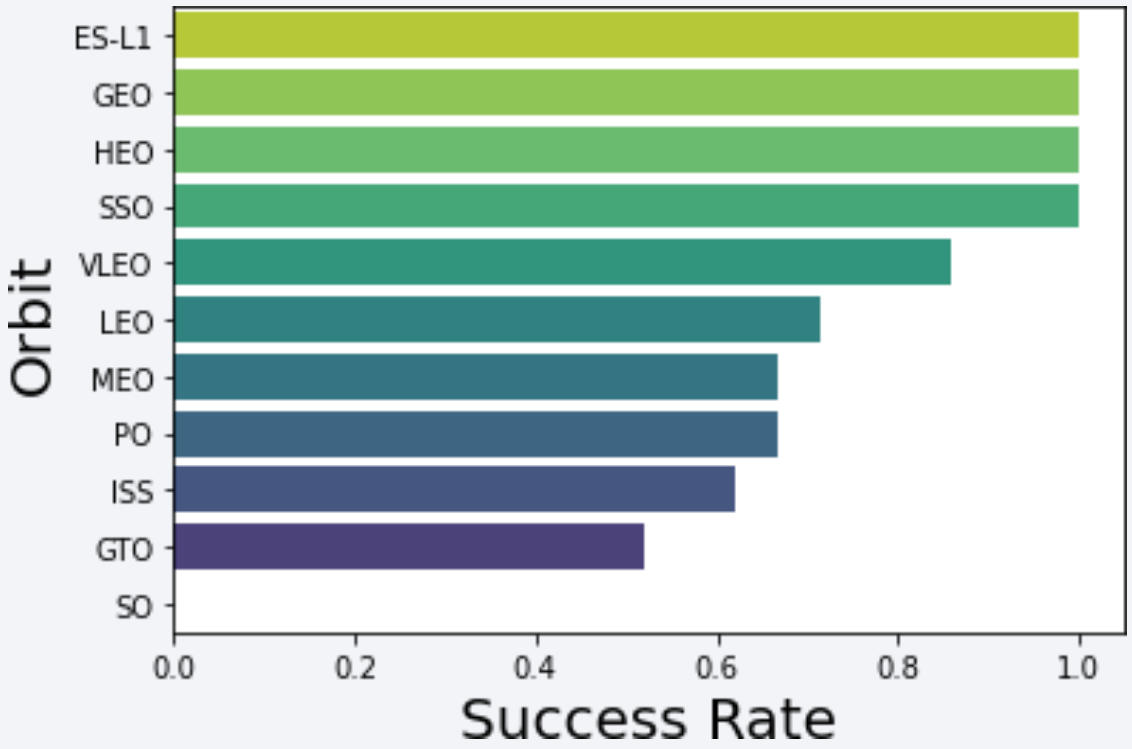


Payload per Launch Site:

Green indicates successful launch; blue indicates unsuccessful launch

- Successful launches of different payload differ for different launch sites.
- Unsuccessful launches occur in the range of ~500kg to 6500 kg payload mass.

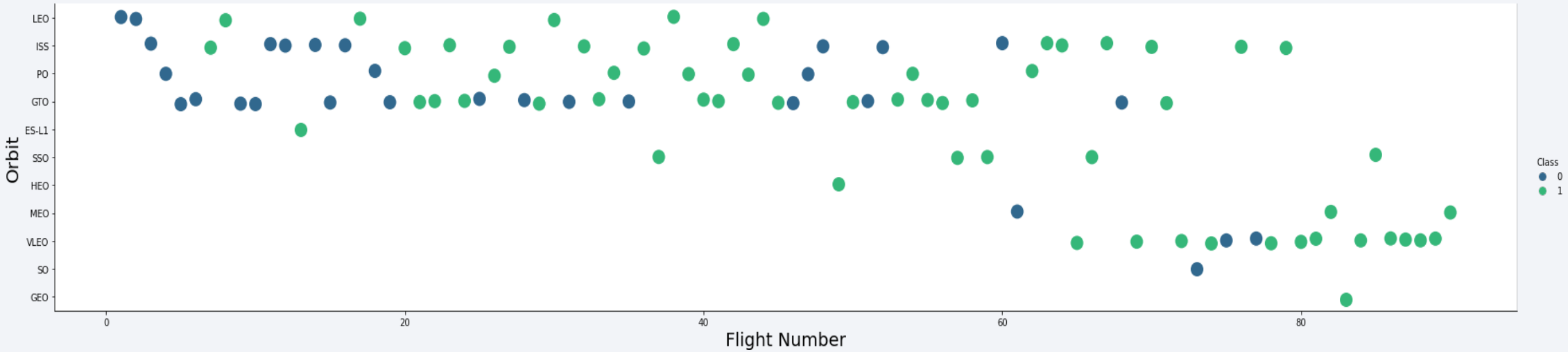
17. Success Rate vs. Orbit Type



- 100% success rate are shown for ES-L1 (1x), GEO (1x), HEO (1x), and SSO (5x) (sample sizes in parenthesis).
- 90% success rate is shown for VLEO (14x).
- GTO (27) has the around 50% success rate but largest sample
- SO (1x) has 0% success rate.

- This analysis shows that the dataset is fairly small for each orbit type.

18. Flight Number vs. Orbit Type

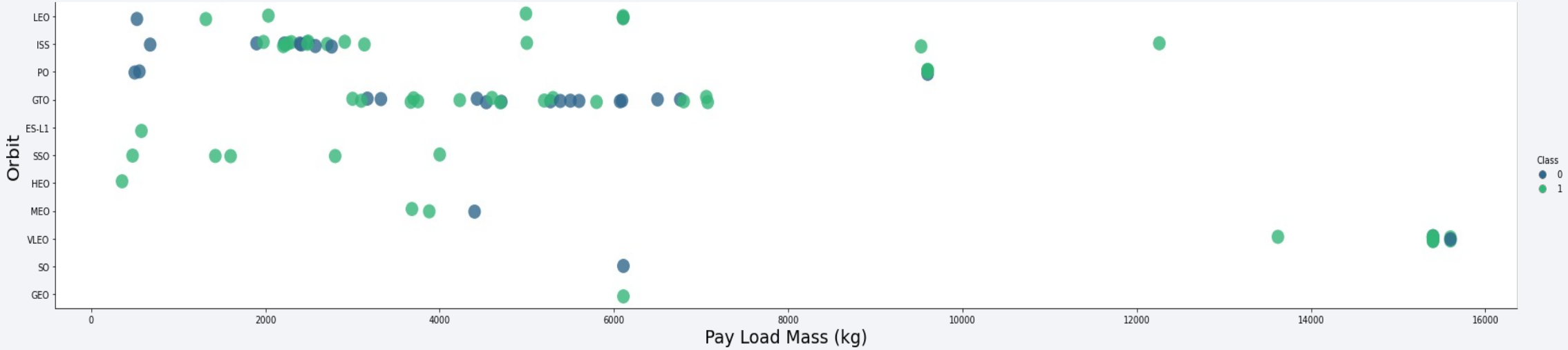


Flight Number per Orbit Type:

Green indicates successful launch; blue indicates unsuccessful launch

- Orbits that the rockets are launched to diversify over flight numbers
- The successful outcome of launches correlate with the flight numbers

19. Payload vs. Orbit Type

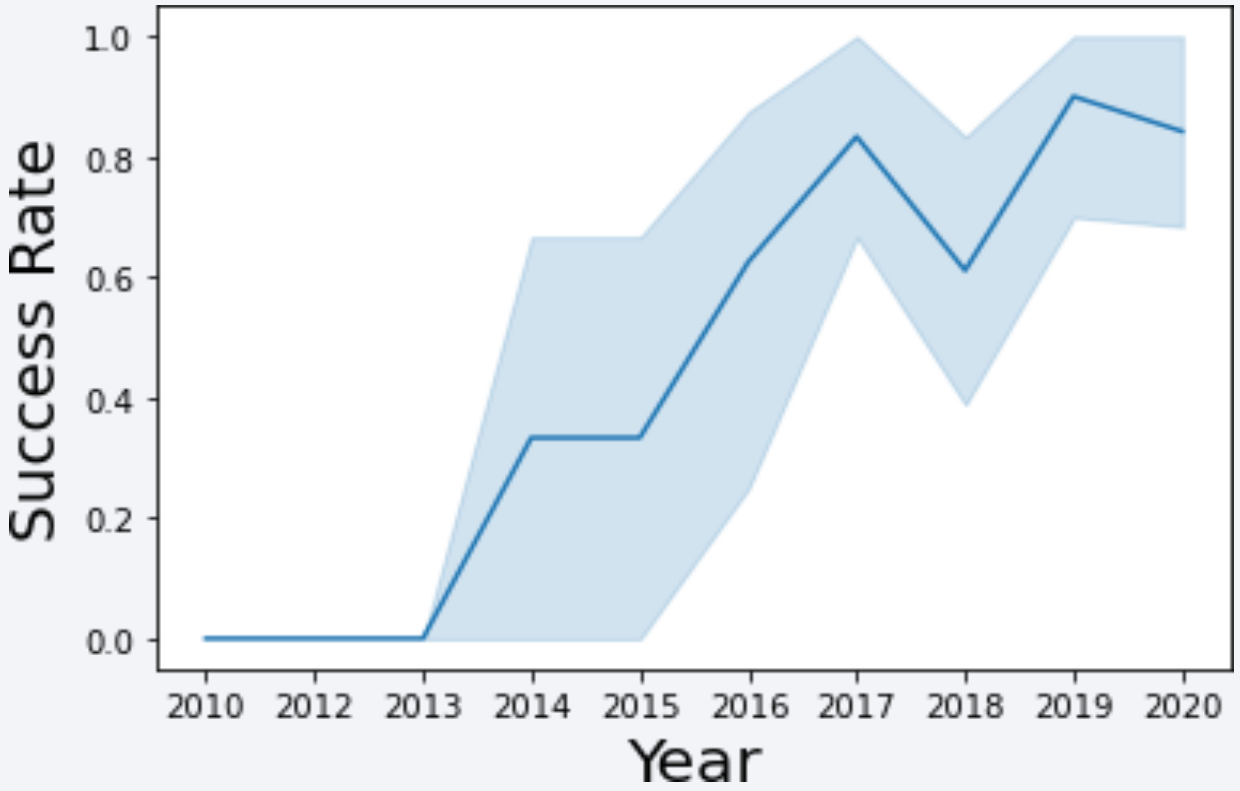


Payload per Orbit Type:

Green indicates successful launch; blue indicates unsuccessful launch

- The payload mass has a correlation with the orbit type based on this chart
- The payload for LEO, SSO and MEO have a successful launch with relatively low payload mass.
- The payload for VLEO is successful with high payload mass.

20. Launch Success Yearly Trend



95% confidence interval
as light blue shading

- Successful launches increase over time since 2013 to 2019 with a slight dip in 2018.
- In recent years the success rate dropped slightly to around 80%

21. All Launch Site Names

```
%%sql
SELECT UNIQUE LAUNCH_SITE
FROM SPACEXDATASET;
```

```
 * ibm_db_sa://ftb12020:***
Done.
```

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

- Name query for unique launch site from database
- The three launch sites CCAFS LC-40, CCAFS SLC-40 and CCAFSSLC-40 look the same and most likely represented the same launch site. This can be due to data entry error and requires data manipulation.
- Therefore we assume that only 3 unique launch_site values exist: CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E

22. Launch Site Names Begin with 'CCA'

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

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
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	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

- The first five entries in the database with the launch_site name beginning with CCA are shown.

23. Total Payload Mass

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS SUM_PAYLOAD_MASS_KG
FROM SPACEXDATASET
WHERE CUSTOMER = 'NASA (CRS)';
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86
Done.
```

sum_payload_mass_kg
45596

- The query sums the total payload mass in kg as sum_payload_mass_kg for the filtered by the customer “NASA”.
- NASA (CRS) are resupply missions (Commercial Resupply Services)to the International Space Station (ISS).

24. Average Payload Mass by F9 v1.1

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD_MASS_KG
FROM SPACEXDATASET
WHERE booster_version = 'F9 v1.1'
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-8
Done.
```

avg_payload_mass_kg
2928

- The query calculates the average payload mass carried by booster version F9 v1.1
- The average payload mass is 2928kg and on the lower end of he payload mass range.

25. First Successful Ground Landing Date

```
%%sql
SELECT MIN(DATE) AS FIRST_SUCCESS
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (ground pad)';
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81
Done.
```

first_success
2015-12-22

- To determine the first successful landing outcome on a ground pad this query was used.
- The resulting date of the first successful landing outcome on ground pad was on 22.Dec.2015.

26. Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
SELECT booster_version
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ BETWEEN 4001 AND 5999;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.database
Done.
```

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

- This query was used to determine the booster types that resulted in a successful landing on a drone ship, with a lift-off payload between 4000 and 6000 kg.
- The following booster versions were used:
  F9 FT B1022, F9 FT B1026, F9 FT B1021.2, F9 FT B1031.2

27. Total Number of Successful and Failure Mission Outcomes

```
%%sql
SELECT mission_outcome, COUNT(*) AS no_outcome
FROM SPACEXDATASET
GROUP BY mission_outcome;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-
Done.
```

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

- This query is used to return each mission outcomes.
- The total number of successful and failure mission outcomes are listed.

28. Boosters Carried Maximum Payload

```
%%sql
SELECT booster_version, PAYLOAD_MASS__KG_
FROM SPACEXDATASET
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXDATASET);
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1
Done.
```

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

- This query is used to return booster versions with the highest payload mass of 15600kg.
- These booster versions look like to be of a Falcon 9 type, indicated by the F9 B5 B10XX.X numbering.
- A correlation between the payload mass and booster version is therefore most likely.

29. 2015 Launch Records

```
%%sql
SELECT MONTHNAME(DATE) AS MONTH, landing__outcome, booster_version, PAYLOAD_MASS__KG_, launch_site
FROM SPACEXDATASET
WHERE landing__outcome = 'Failure (drone ship)' AND YEAR(DATE) = 2015;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.app
Done.
```

MONTH	landing__outcome	booster_version	payload_mass__kg_	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	2395	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	1898	CCAFS LC-40

- This query lists the failed landing_outcomes on drone ships, including their their booster versions, and launch site names for the year 2015.

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

```
%%sql
SELECT landing__outcome, COUNT(*) AS no_outcome
FROM SPACEXDATASET
WHERE landing__outcome LIKE 'Succes%' AND DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY landing__outcome
ORDER BY no_outcome DESC;
```

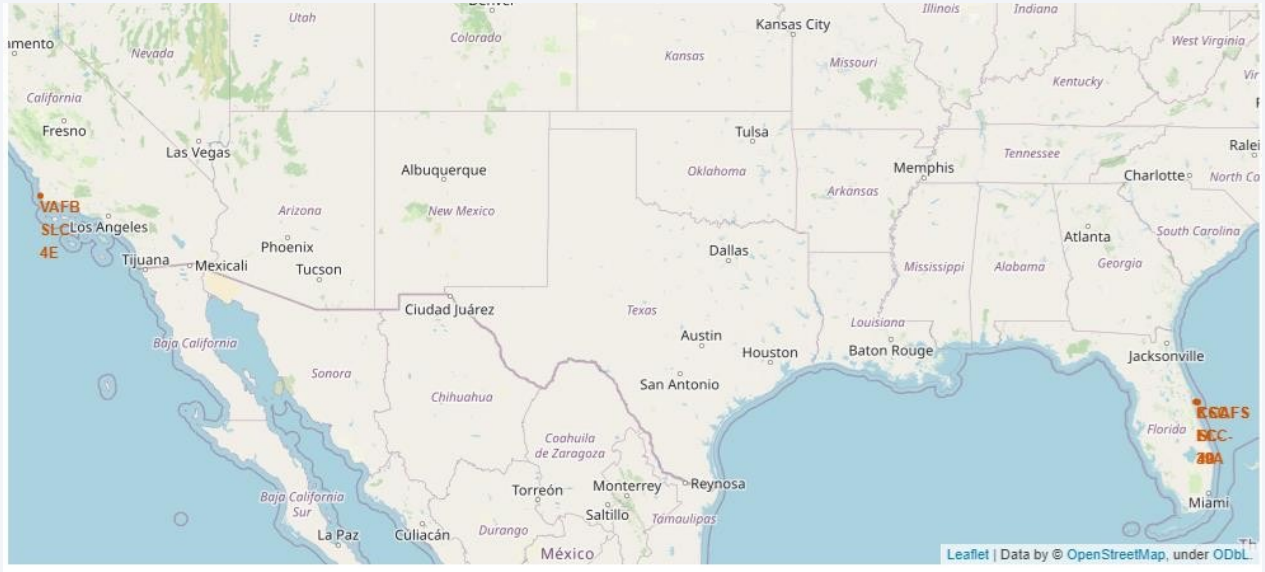
```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg
Done.
```

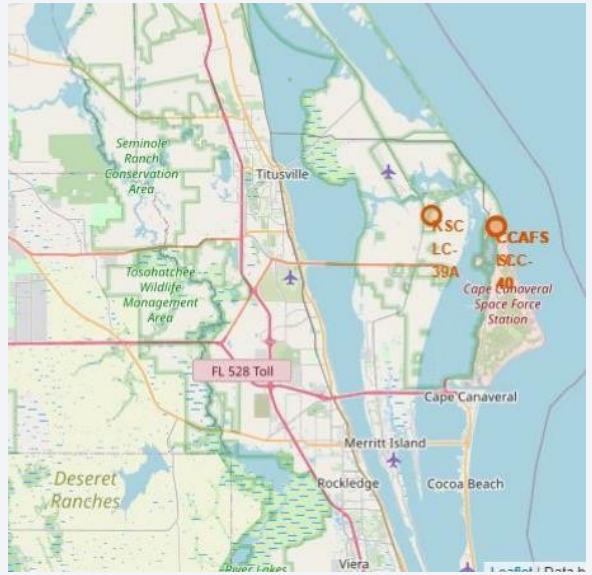
landing__outcome	no_outcome
Success (drone ship)	5
Success (ground pad)	3

- This query ranks the landing outcome by drone ship and ground pad between the 04. Jul.2010 and 20.Mar.2017 in a descending order.


Section 3
Launch Sites
Proximities Analysis

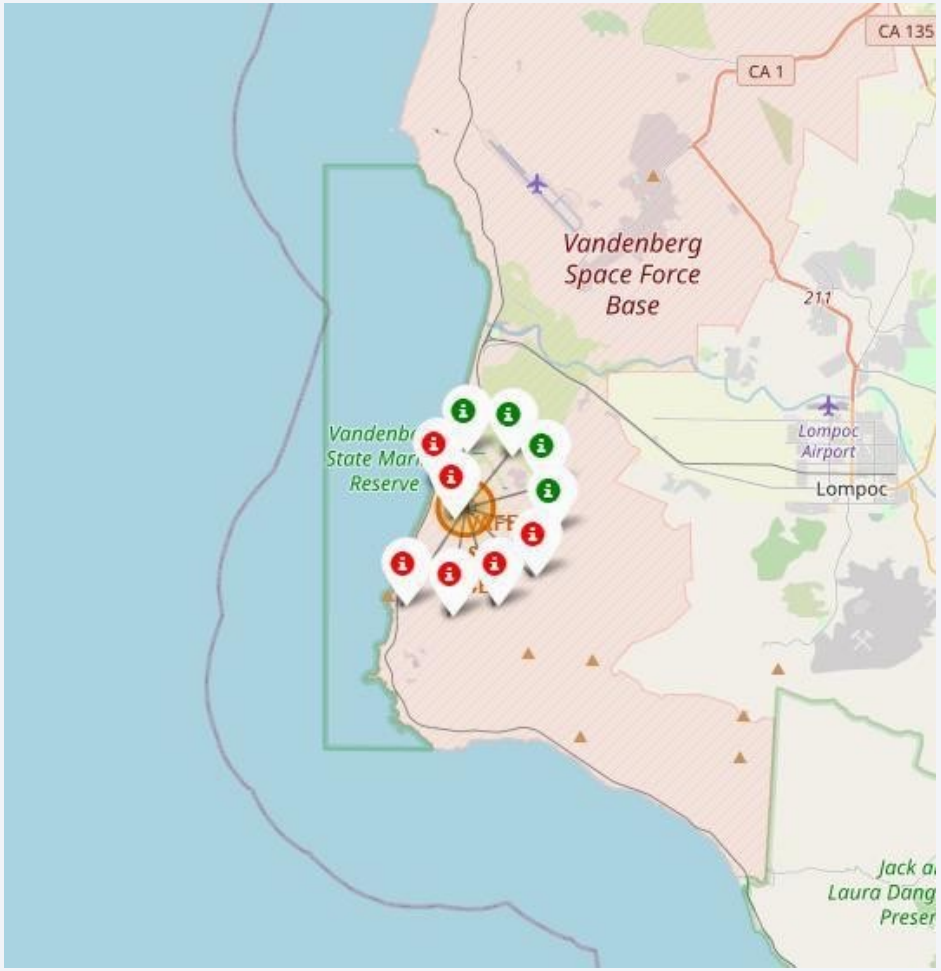
31. Locations of the launch sites in the USA



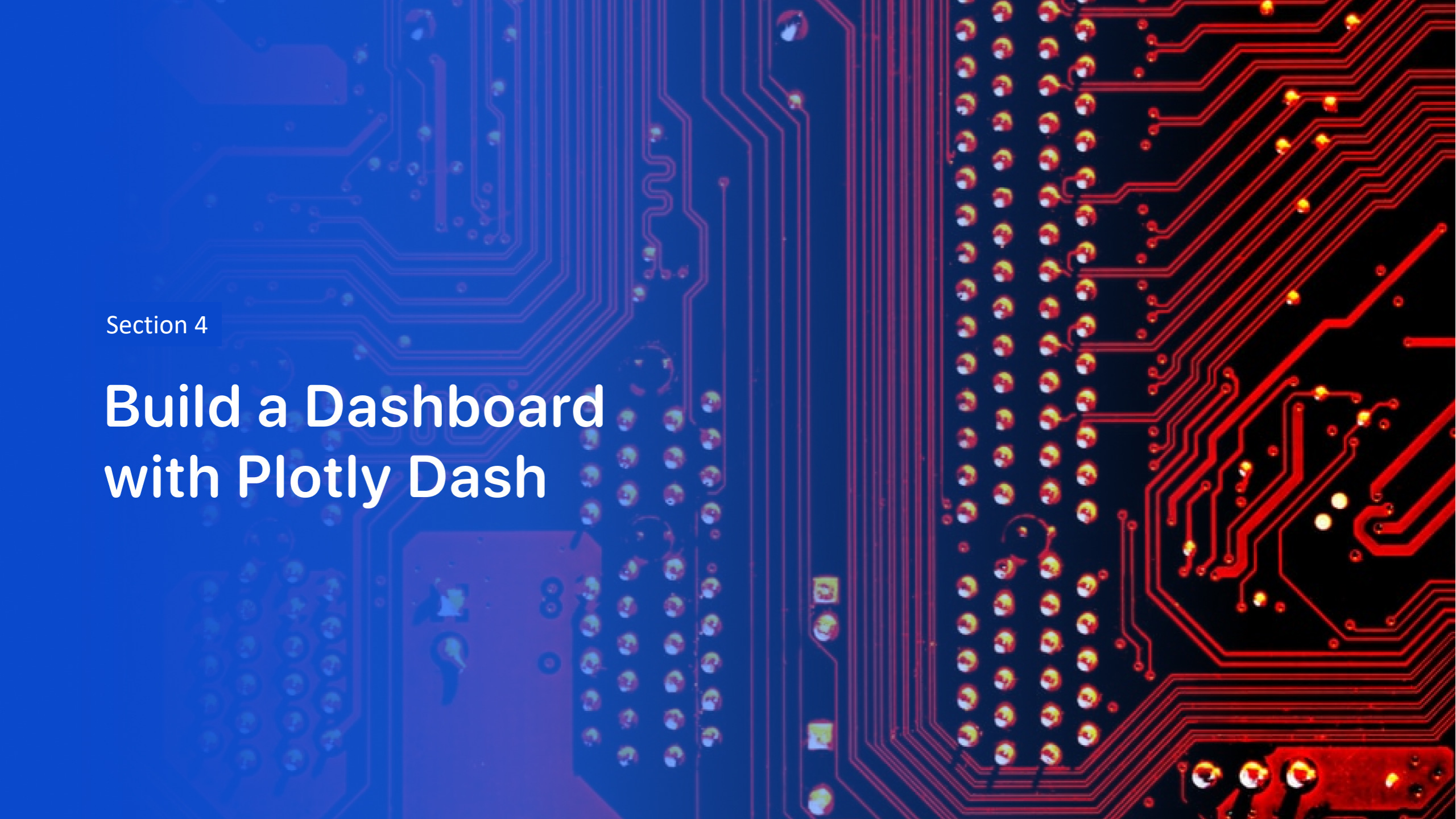


- In the left map, launch sites (in orange) are shown across the USA.
- In the right map, a closer look was done to the launch sites in Florida.
- Multiple launch sites exist in close proximity.

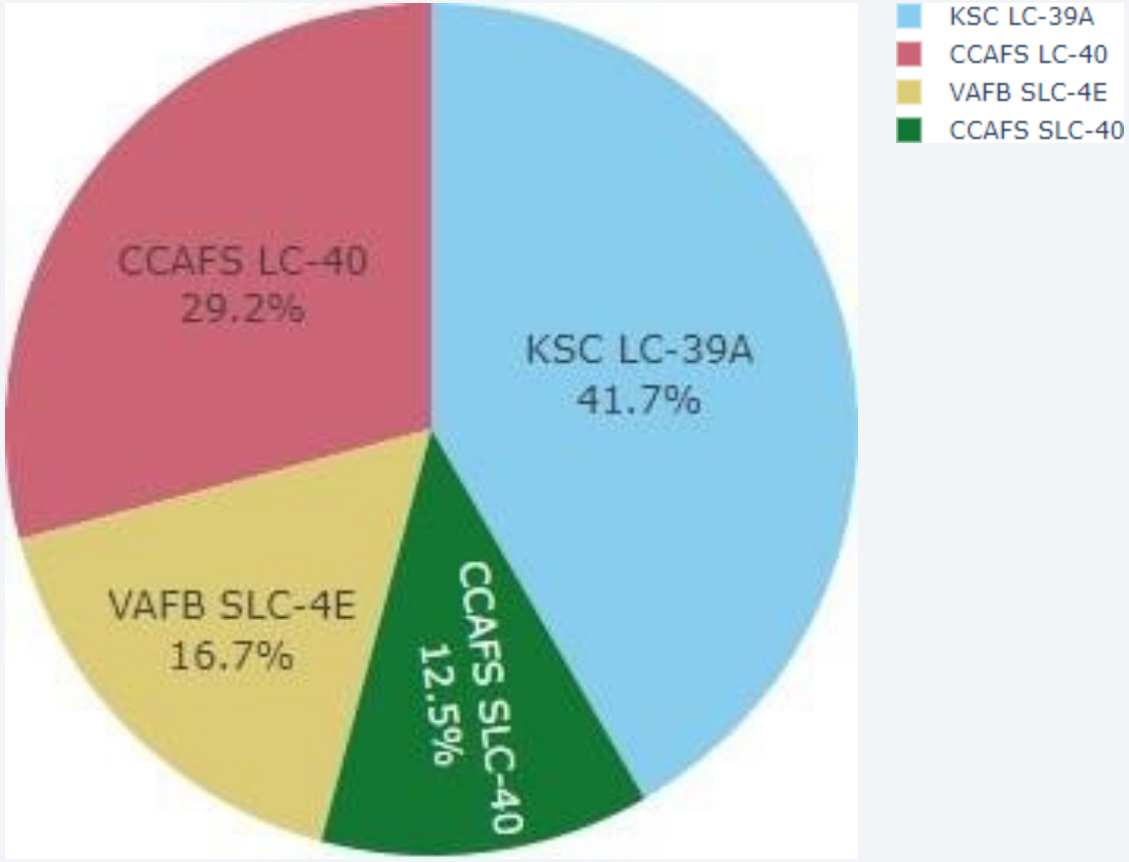
32. Color Coded Markercluster of launch site



- Interactive MarkerClusters in the Folium map was created. Each marker will display a successful landing (green icon) and failed landing (red icon).
- VAFB SLC-4E is shown here with 4 successful landings and 6 failed landings.

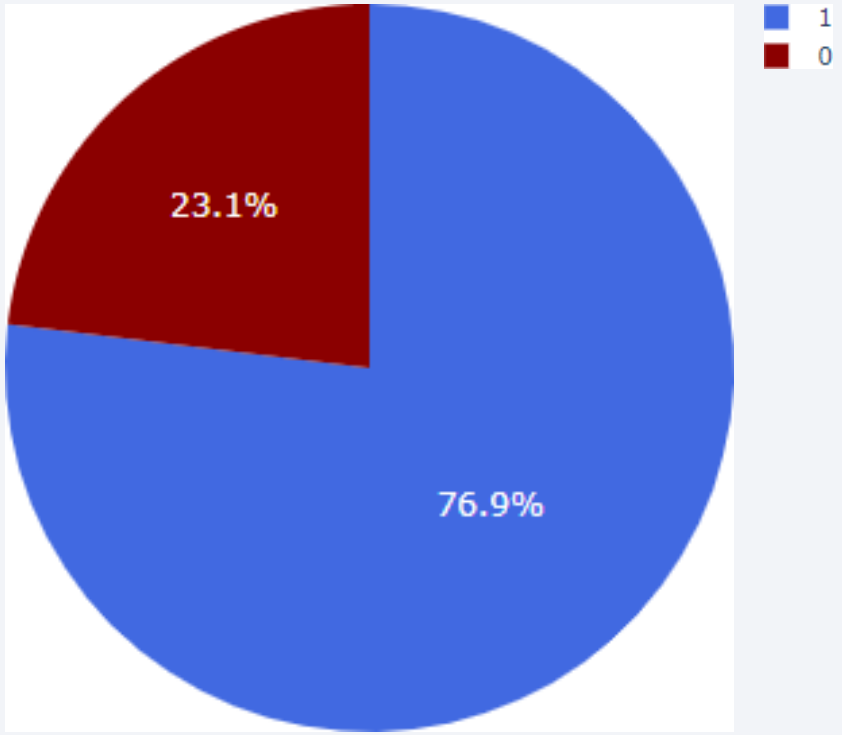

Section 4
Build a Dashboard with Plotly Dash

33. Successful launches at different launch sites



- The distribution of successful landings at different landing sites is shown in a pie chart. Percentage values are indicating the success at the different sites compared to each other.
- CCAFS LC-40 and CCAFS SLC-40 are the same launch sites, resulting in a total success rate of 41.7%.
- The majority of successful landings where done at CCAFS and KSC LC-39A with 41.7%
- However the overall data size is small.

34. Highest success rate at launch site



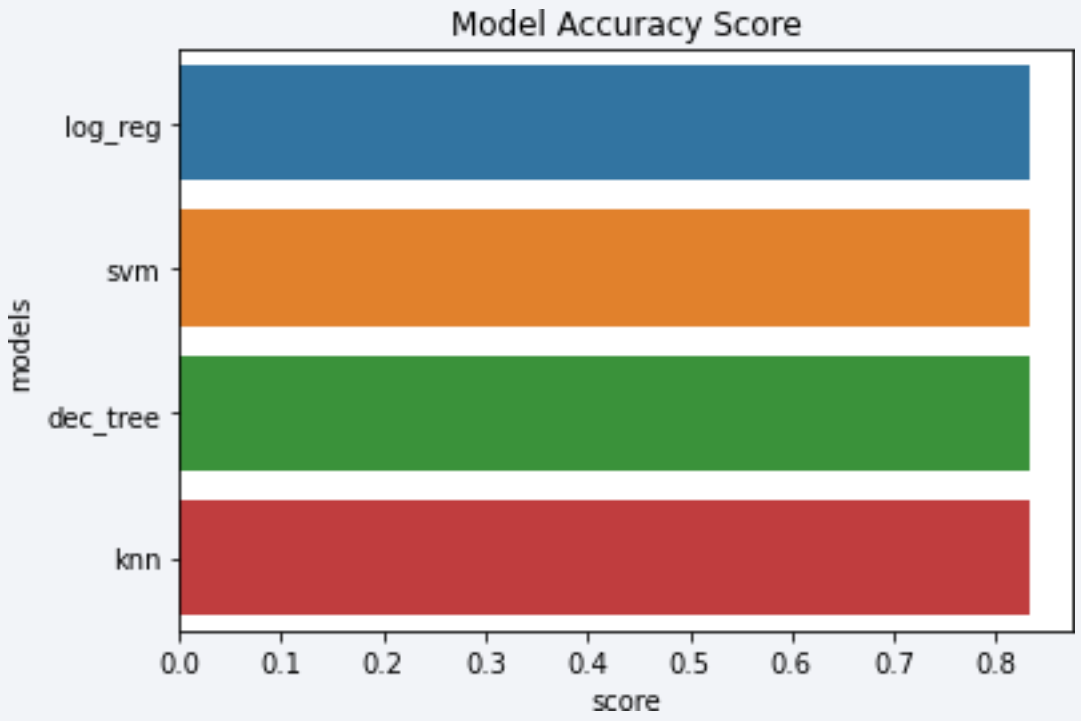
KSC LC-39A Success Rate (blue color)

For KSC LC-39A the success rate is the highest with 10 successful landings and 3 failed landings, resulting in a 76.9% success rate.

Section 5

Predictive Analysis (Classification)

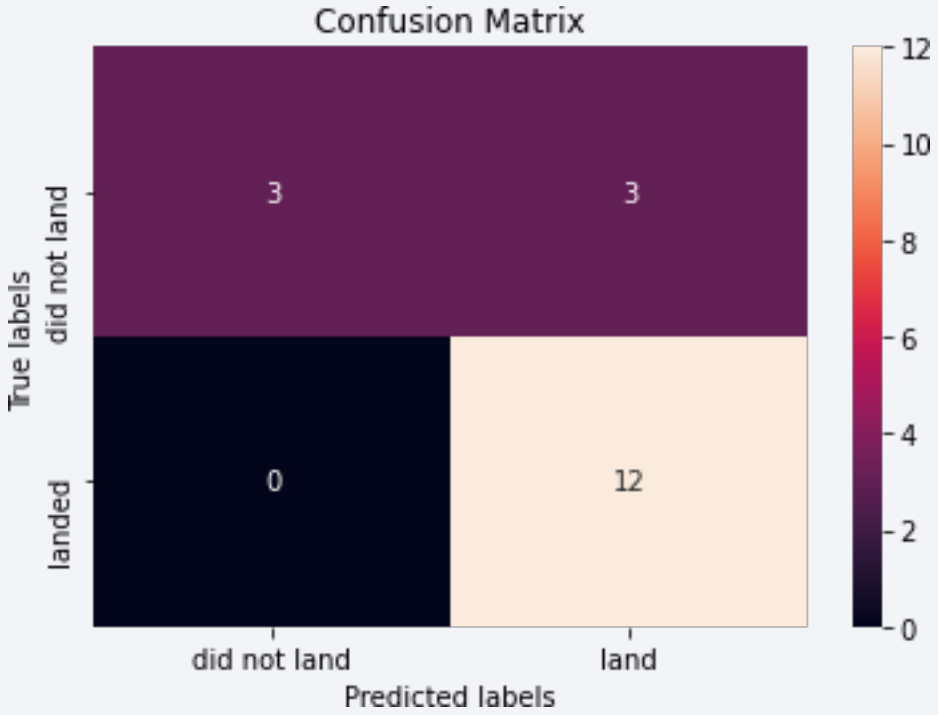
Classification Accuracy



- All built classification models are compared in a bar chart.
- For all models the accuracy on the test set is at the same value of 83.33% accuracy.
- Test samples are however low with a sample size of 18.
- The small test set can therefore cause a large variance in accuracy results in repeated runs.

- To determine the best model with higher accuracy, more data is required.

Confusion Matrix



- Confusion Matrix visualize the landing prediction
- The Confusion Matrix is the same for all calculated models.
- 12 successful landings were predicted when the true label was a successful landing.
- 3 unsuccessful landings were predicted when the true label was unsuccessful landing.
- 3 successful landings were predicted when the true label was unsuccessful landings (false positives).

- All models overpredicted a successful landing.

Conclusions

- The task of this Capstone Project was to develop a machine learning model for SpaceY who wants to bid against SpaceX.
- With our models we predict when and where a Stage 1 of the rocket will successfully land and therefore save ~$100 million USD.
- Public available data from SpaceX API and webscraping from the SpaceX Wikipedia page was used.
- Data labels were created and the data was stored into a DB2 SQL database.
- The creation of a dashboard helped to visualize the data.
- A machine learning model was created with an accuracy of 83%.
- This models is available to the board of SpaceY to predict with relatively high accuracy whether a launch will have a successful Stage 1 landing, even before the launch.
- For more accurate predictions, more data is required to improve the accuracy of the machine learning model.

Appendix

- GitHub repository URL: https://github.com/Mr-Minion/IBM_Data_Science_Professional_Certificate/tree/main/Applied_Data_Science_Capstone

- Special Thanks to all the instructors and the Funding bodies

Thank you!