IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Carlo Nesti
3 March 2024

Outline

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

Executive Summary

The main purpose of this report is to evaluate the viability of the new company SpaceY to compete with SpaceX by predicting successful landings of the first stage of rockets.

- Summary of methodologies
  - Data Collection using web scraping and SpaceX Application Programming Interface (API).
  - Exploratory Data Analysis (EDA), including data wrangling, data visualization and interactive visual analytics.
  - Predictive Analysis using Machine Learning.

- Summary of all results
  - Mission success factors: launch site, orbit type, payload mass and the number of previous flights.
  - Key findings: Lighter payloads, specific orbits, and repeated attempts improve success rates. KSC LC-39A is the best launch site. Payloads under 6,000 Kg and FT type boosters are the most successful combination. Decision Tree is the preferred machine learning model for predicting a successful landing (accuracy 87%).

Introduction

- Project background and aim
  - SpaceX advertises Falcon 9 launches at a cost of 62 million dollars each. Other providers charge up to 165 million dollars per launch. A significant portion of these savings is due to SpaceX’s ability to reuse the first stage of the rockets. Accordingly, if it can be predicted whether the first stage will land successfully, the cost of a launch can be estimated. This information could be used by SpaceY to compete with SpaceX, potentially saving millions on each launch.
  - This project aims to leverage data science, specifically machine learning models, to predict the success of Falcon 9's first-stage landings

- Problems to be solved
  - Identifying the factors that influence the successful landing of rockets.
  - Determining the best way to estimate the total cost for launches by predicting the successful landings of rockets.

Section 1

Methodology

Methodology

Executive Summary

- • Data collection
  - The data was gathered from the SpaceX REST API and from web scraping Wikipedia pages.
- • Data wrangling
  - The data was read into a Pandas dataframe, cleaned and one-hot encoded for machine learning analysis.
- • Exploratory data analysis (EDA) using visualization and SQL
  - Scatter and bar graphs were used to show relationships between variables and identify patterns.
- • Interactive visual analytics using Folium and Plotly Dash
  - Launch sites geography and proximities were analyzed to identify an optimal launch site.
- • Predictive analysis using classification models
  - Classification machine learning models were used to predict if the first stage of rockets will land successfully.

Data Collection

SpaceX launch data was collected from the SpaceX REST API. This API provides information about launches, such as the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome. (https://api.spacexdata.com/v4/rockets/)

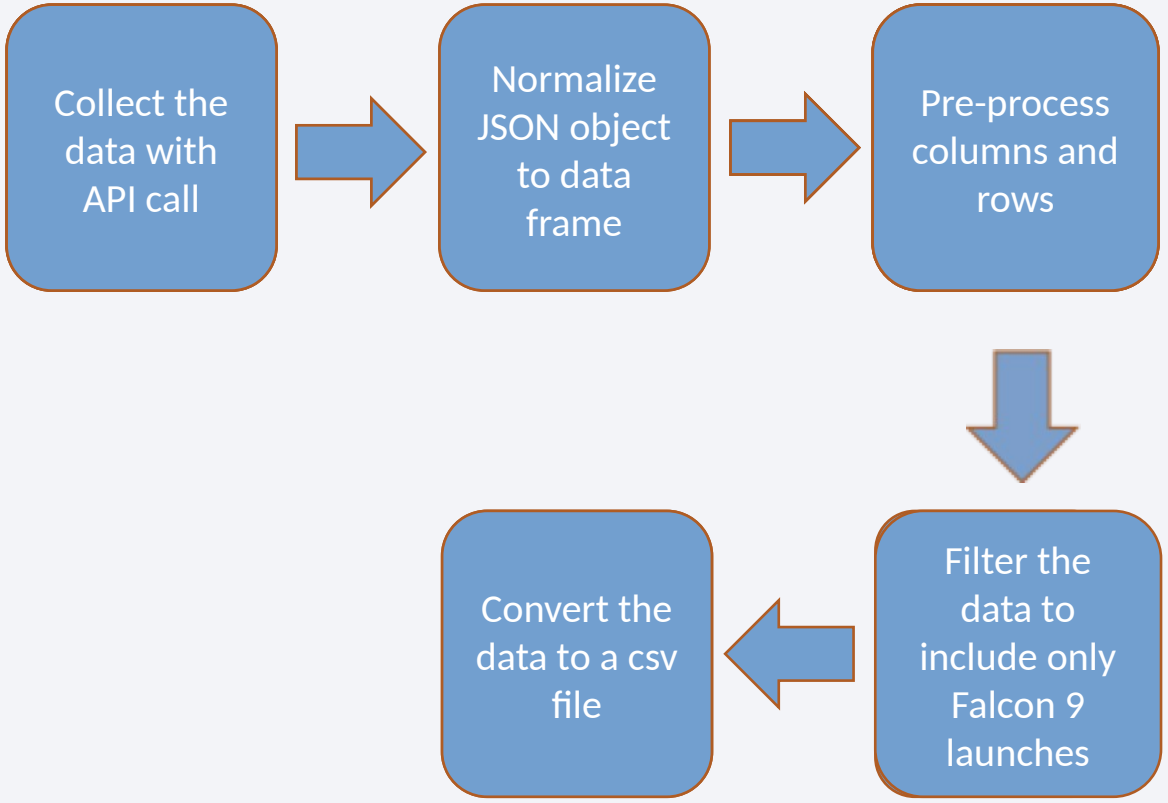
Another data source used for obtaining SpaceX launch data was web scraping Wikipedia pages using Python library BeautifulSoup. (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)

The flowchart shows the main steps involved.

Data Collection – SpaceX API

- Data collection by using SpaceX REST API.

  The flowchart shows the main steps involved.



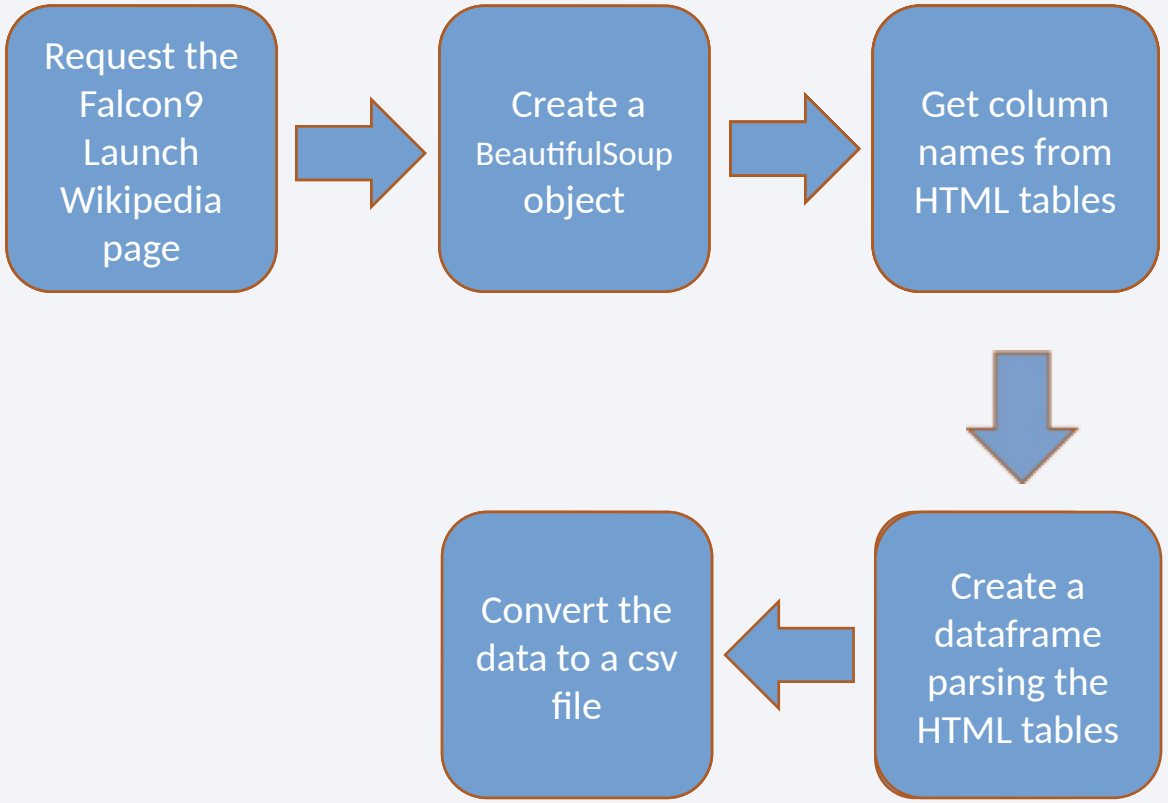
Data collection - SpaceX API notebook

GitHub url: https://github.com/Carlo-bits/IBM-Applied-Data-Science-Capstone-Project/blob/main/Notebooks/Lab1-jupyter-labs-spacex-data-collection-api.ipynb

Data Collection – Web Scraping

- Data collection by scraping Wikipedia web pages with Falcon 9 launches.

  The flowchart shows the main steps involved.



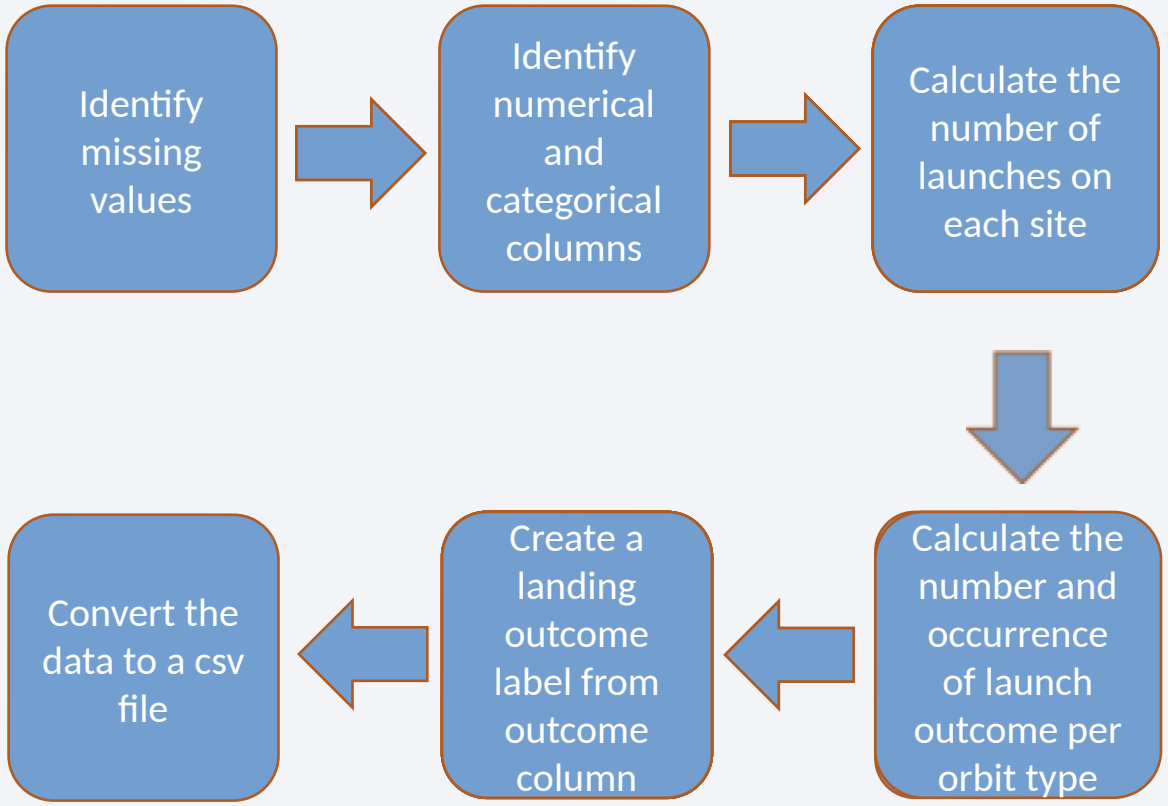
Data collection – Web Scraping notebook

GitHub url: https://github.com/Carlo-bits/IBM-Applied-Data-Science-Capstone-Project/blob/main/Notebooks/Lab2-jupyter-labs-webscraping.ipynb

Data Wrangling

- Data wrangling was carried out to determine launch outcome labels for training machine learning models
  - 1: successful landing
  - 0: unsuccessful landing

The flowchart shows the main steps involved.



[Data collection – Data Wrangling notebook]

GitHub url: https://github.com/Carlo-bits/IBM-Applied-Data-Science-Capstone-Project/blob/main/Notebooks/Lab3-labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Scatter plots were drawn to show how two numerical variables are correlated:
  - Flight Number vs Payload, Flight Number vs Launch Site, Payload vs Launch Site, Flight Number vs Orbit type, and Payload vs Orbit type
- Bar charts were drawn to compare categorical data between different groups:
  - Orbit vs Landing Success Rate
- Line charts were drawn to show trends between variables:
  - Year vs Landing Average Success rate

Data collection – EDA with Data Visualization notebook

GitHub url: https://github.com/Carlo-bits/IBM-Applied-Data-Science-Capstone-Project/blob/main/Notebooks/Lab5-jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- SQL queries performed:
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA’
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first successful landing outcome in ground pad was achieved.
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List the names of the booster versions which have carried the maximum payload mass. Use a subquery
  - List the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between 2010-06-04 and 2017-03-20, in descending order

Data collection – EDA with SQL notebook

GitHub url: https://github.com/Carlo-bits/IBM-Applied-Data-Science-Capstone-Project/blob/main/Notebooks/Lab4-jupyter-labs-eda-sql-coursera_sqllite.ipynb

Interactive Map with Folium

- The interactive map with Folium was created to help in finding an optimal launch site location. The following map objects were created:
  - folium.Marker() was used to create markers on the maps.
  - folium.Circle() was used to create circles above markers on the map.
  - folium.Icon() was used to create an icon on the map.
  - folium.PolyLine() to draw a straight line between a launch site and its closest city, railway or highway.
  - folium.plugins.AntPath() was used to create an animated line between the points.
  - markerCluster() was used to simplify a map containing multiple markers with similar coordinates.

Data collection – Interactive Map with Folium notebook

GitHub url: https://github.com/Carlo-bits/IBM-Applied-Data-Science-Capstone-Project/blob/main/Notebooks/Lab6-lab_jupyter_launch_site_location.ipynb

Dashboard with Plotly Dash

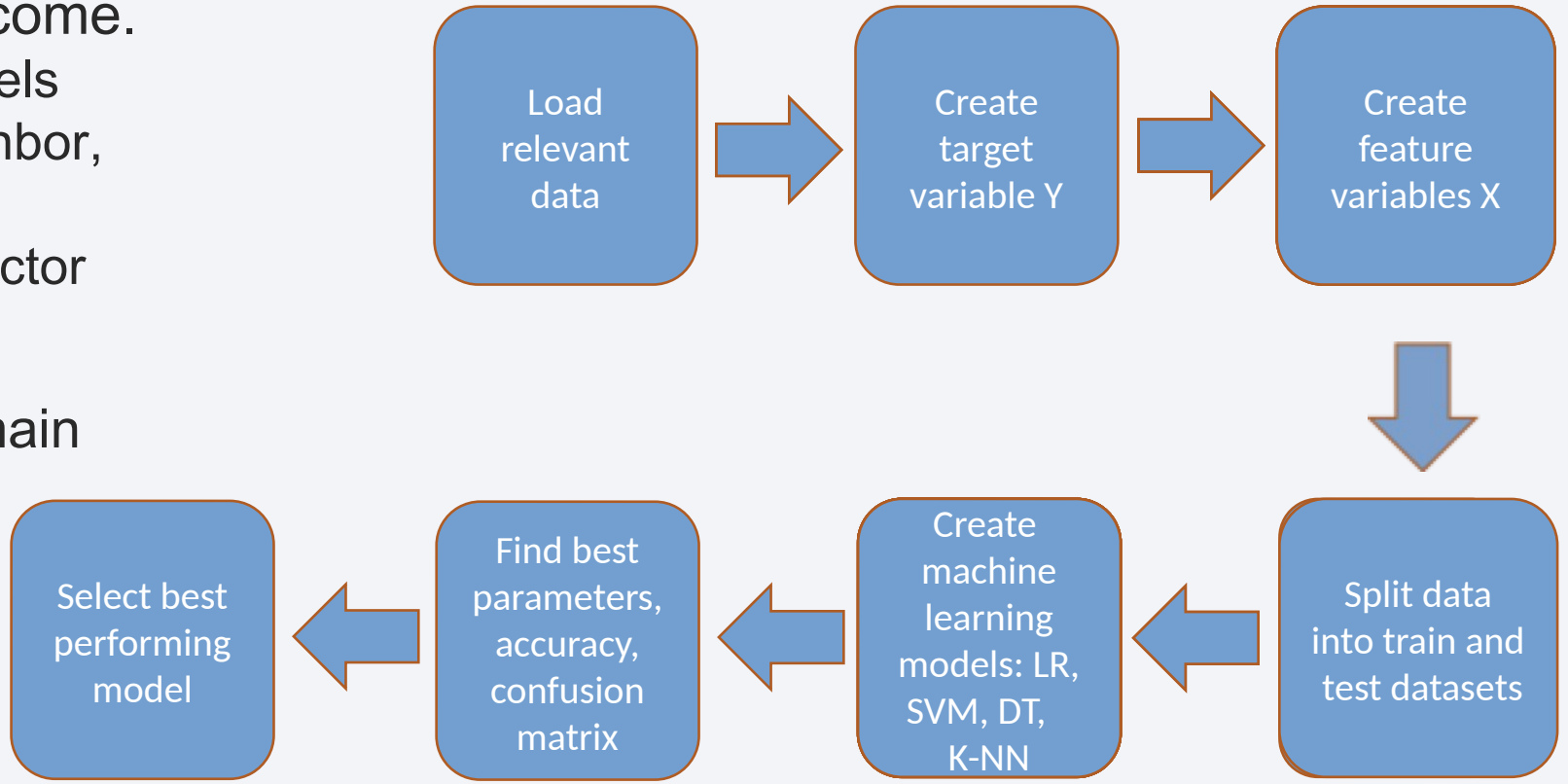
- Graphs and interactions were added to the dashboard to perform interactive visual analytics on SpaceX launch data.
  - The dashboard contains input components such as a dropdown list and a range slider to interact with a pie chart and a scatter plot chart.
  - A launch site drop-down input component allows the selection of different launch sites from a dropdown menu.
  - A callback function generates a pie chart to visualize launch success counts based on the chosen launch site.
  - A range slider component allows the selection of specific payload ranges to check for visual patterns in the data.
  - A callback function generates a scatter plot to visually observe how payloads correlate with mission outcomes for selected sites.

Data collection – Dashboard with Plotly Dash notebook

GitHub url: https://github.com/Carlo-bits/IBM-Applied-Data-Science-Capstone-Project/blob/main/Notebooks/Lab7-Build_a_Dashboard_Application_with_Plotly_Dash.ipynb

Predictive Analysis (Classification)

- Machine learning predictive analysis was carried out to predict rocket launch outcome. Four machine learning models were used: K-Nearest Neighbor, Decision Tree, Logistic Regression and Support Vector Machine.
- The flowchart shows the main steps involved.



Data collection – Predictive Analysis (Classification) notebook

GitHub url: https://github.com/Carlo-bits/IBM-Applied-Data-Science-Capstone-Project/blob/main/Notebooks/Lab8-SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

- Exploratory data analysis results
  - The success rate for the missions is clearly increasing over time, but it is not possible to predict the success or failure based on flight number and launch site alone.
  - The greater the payload mass (greater than 8000) the higher the success rate. But there is no clear pattern to make a decision if the launch site is dependent on payload mass for a success launch.
  - Orbits SSO, HEO, GEO, and ES-L1 show the highest success rate.
  - The mission success rate improved over time for all orbits, especially for LEO, but no clear pattern can be detected.
  - With heavy payloads the successful landing rate are higher for LEO, ISS and PO orbits.
  - The mission success rate increased in 2013 and kept increasing until 2020, with the exception of year 2018.

Results

- Interactive analytics results
  - Launch sites are strategically located near USA coastlines (in Florida and California) for safety reasons, allowing options for aborting launches and minimizing risks to people and property. Proximity to highways facilitates easy transportation of equipment, while railways enable efficient transport of heavy cargos. Importantly, launch sites intentionally avoid densely populated areas to minimize danger to residents.
  - Launch sites are also in proximity to the Equator, as it takes less fuel to get into space from the Equator due to Earth's rotation.
  - Site KSC LC-39A shows the highest success rate of 41.7%. Site CCAFS SLC-40 shows the lowest success rate of 12.5%. The launch site appears to be a relevant factor in deciding the success of a mission.
  - Site KSC LC-39A shows 76.9% of successful launches, and 23.1% of unsuccessful launches.
  - Overall, launch outcomes tends to be less successful with increasing payloads. Payloads under 6,000 Kg and FT boosters appear to be the most successful combination.

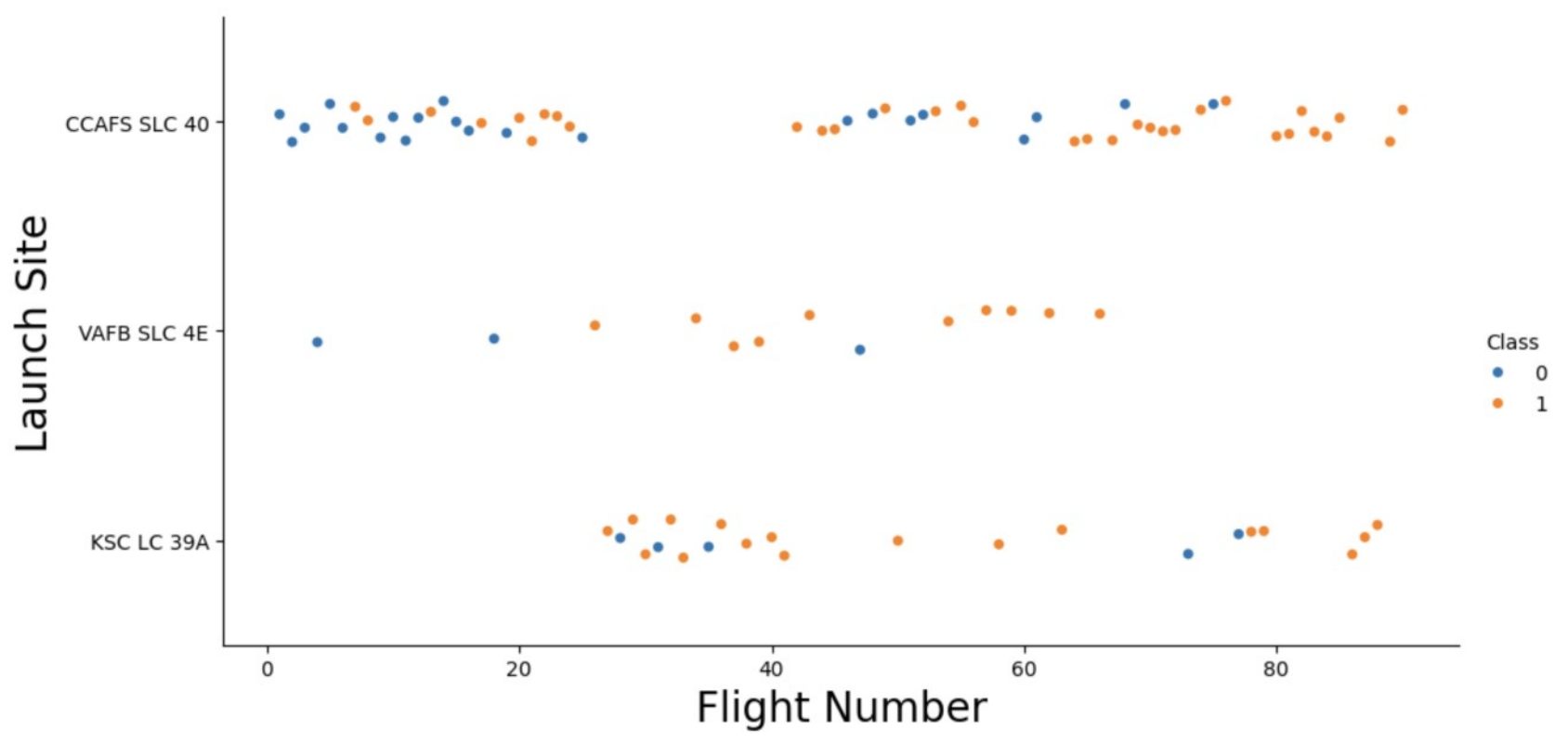
Results

- Predictive analysis results
  - Four classification models were tested: K-Nearest Neighbor, Decision Tree, Logistic Regression and Support Vector Machine. The models resulted in comparable accuracy, with the Decision Tree classifier showing the highest accuracy at approximately 87%.
  - The confusion matrices are identical for all 4 classification models.
    - True Positives: the models seem good at predicting successful landings. This is crucial information for space agencies.
    - False Positives: however, the models sometimes predict successful landings when rockets actually crash. This is a serious issue.
    - True Negatives: the models are good at predicting unsuccessful landings.
    - Overall Performance: while the models seem to perform well in predicting successful and unsuccessful landings, the false positive results need to be carefully considered.

Section 2

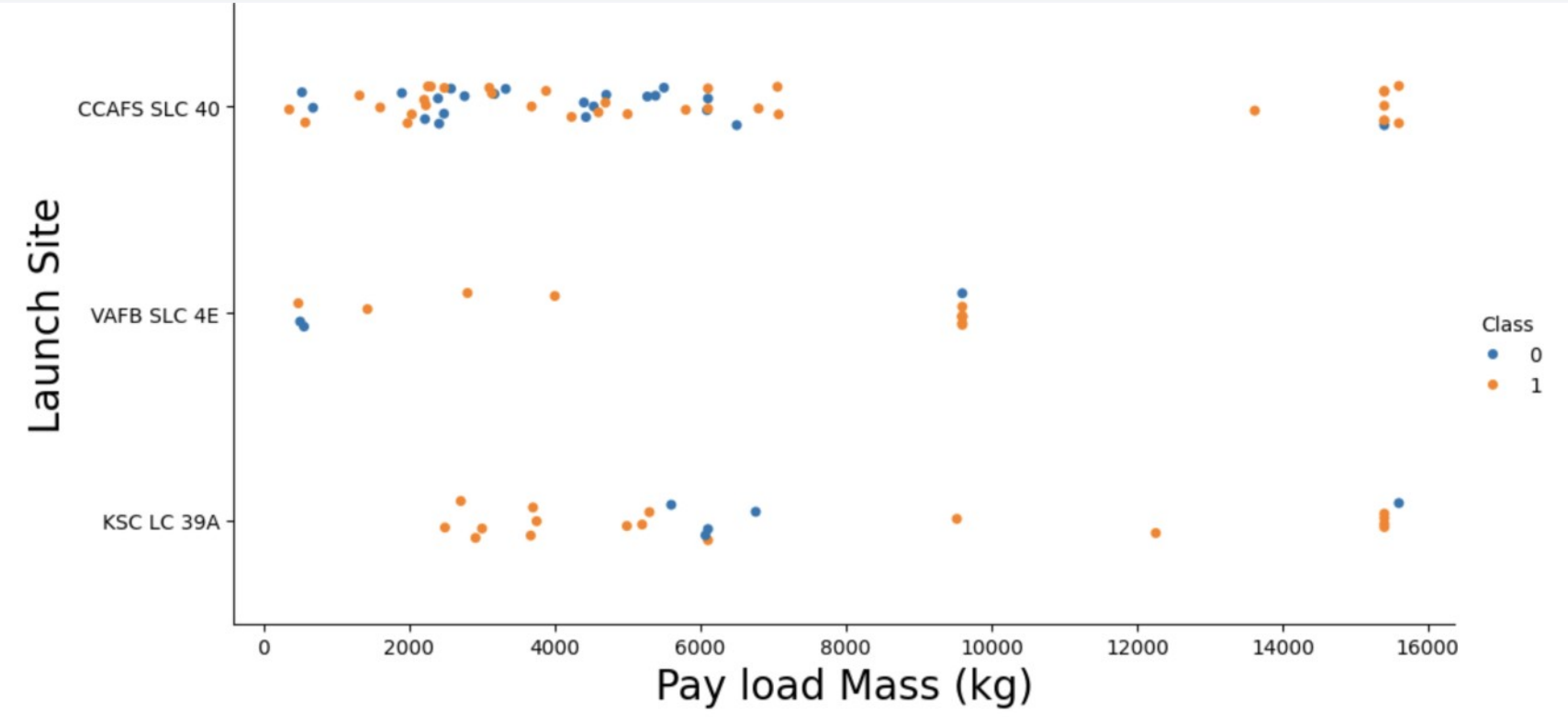
Insights drawn from EDA

Flight Number vs. Launch Site



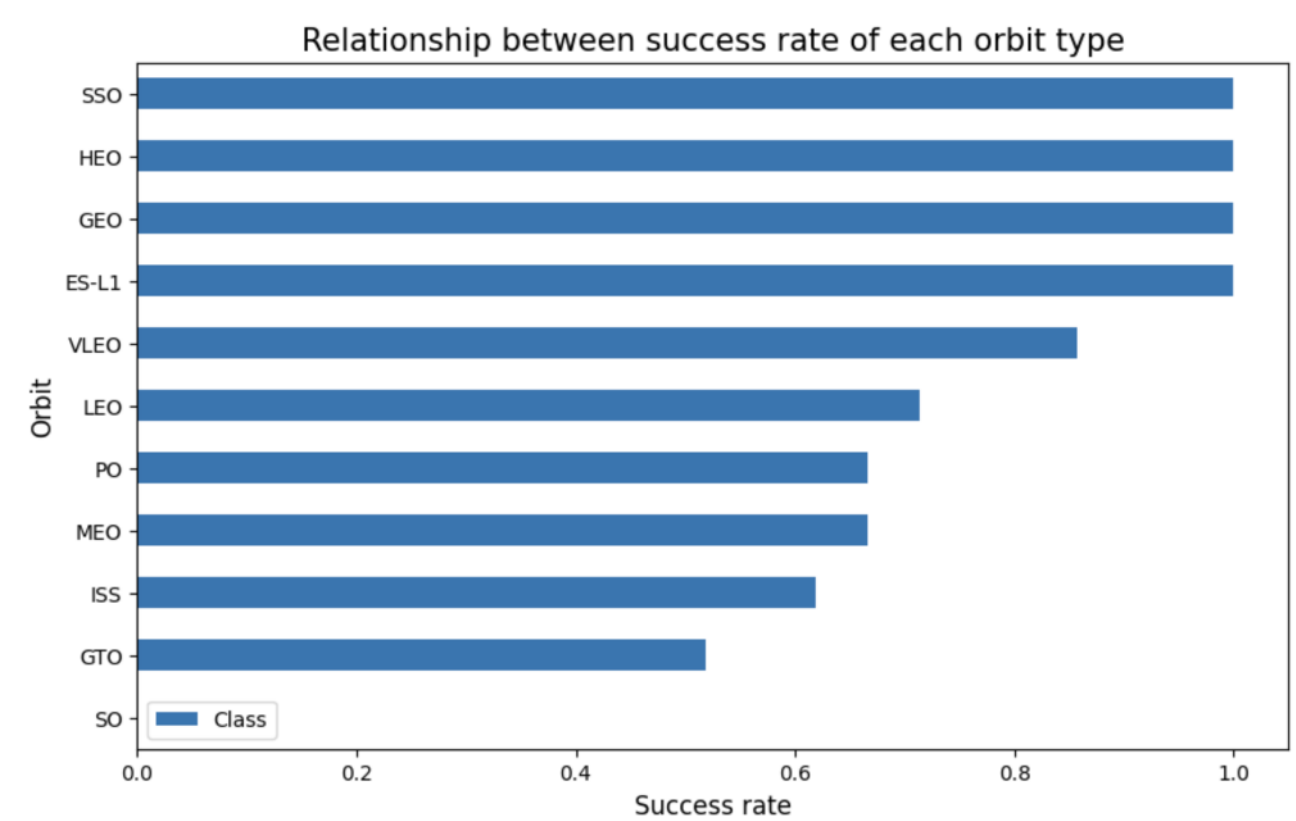
The success rate for the missions is clearly increasing over time, but it is not possible to predict the success (1) or failure (0) based on flight number and launch site alone. The latest launches, after 80, were successful for all 3 sites.

Pay load Mass vs. Launch Site



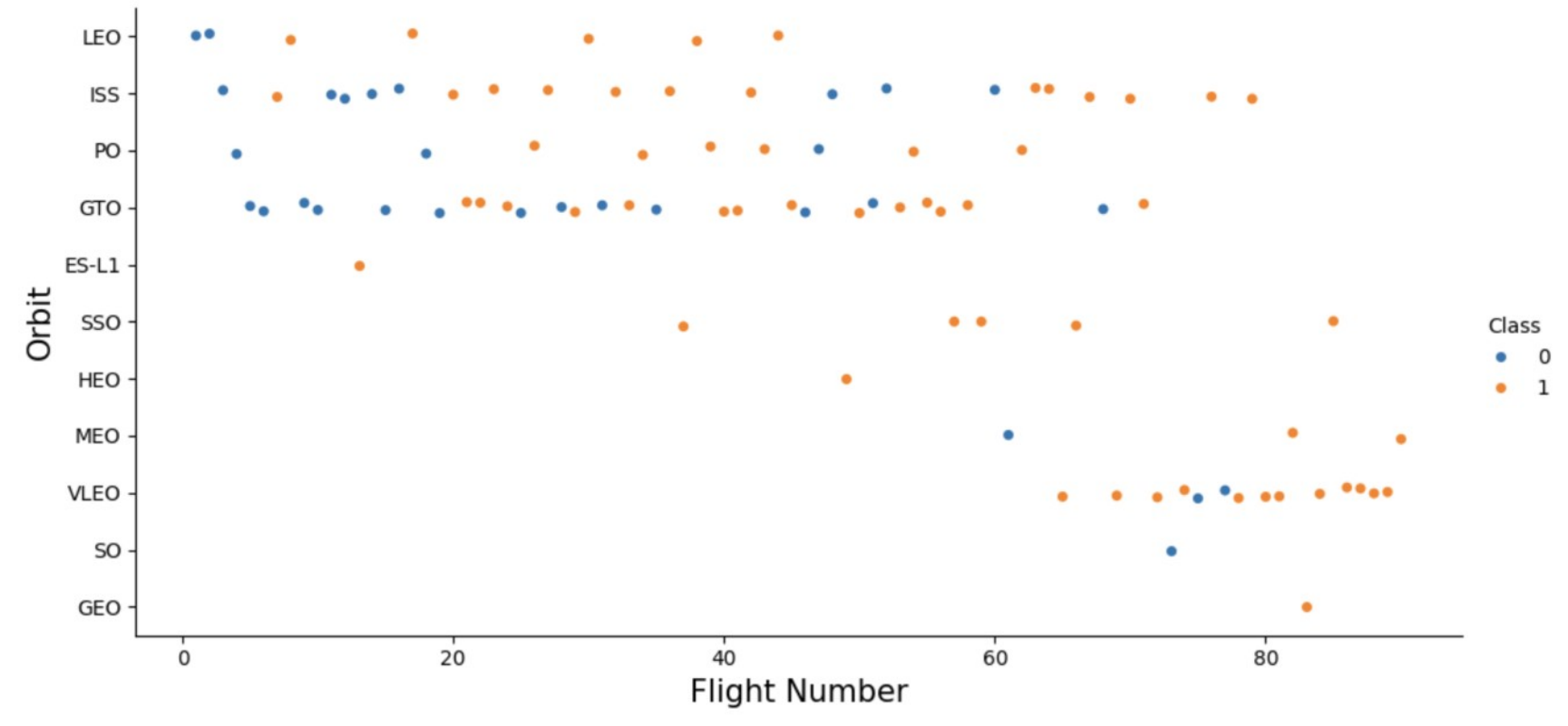
- The greater the payload mass (greater than 8000) the higher the success rate. But there is no clear pattern to make a decision if the launch site is dependent on payload mass for a success launch. For the VAFB-SLC launch site there are no rockets launched for payload mass greater than 10000.

Success Rate vs. Orbit Type



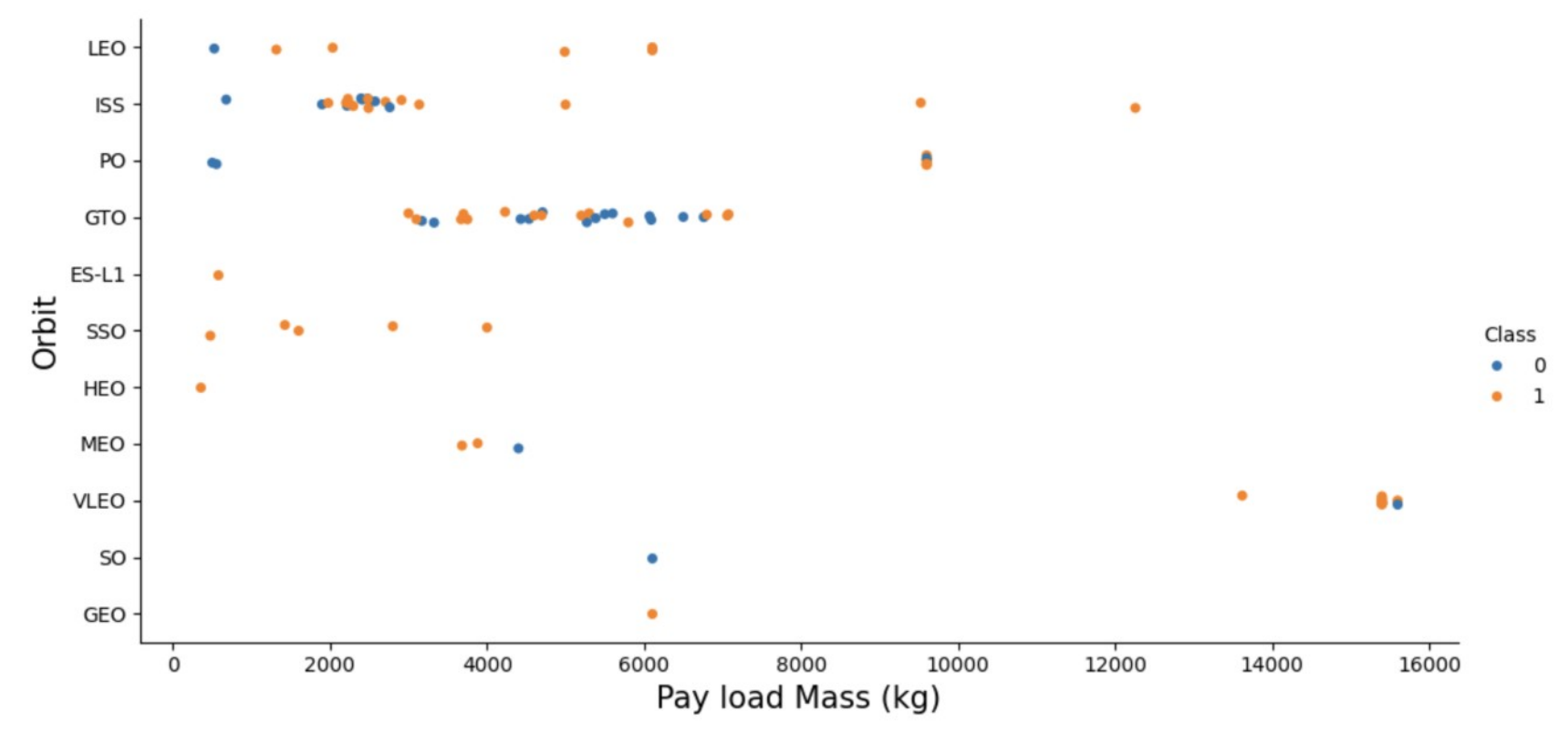
- Orbits SSO, HEO, GEO, and ES-L1 show the highest success rate. No successful mission is reported for site SO.

Flight Number vs. Orbit Type



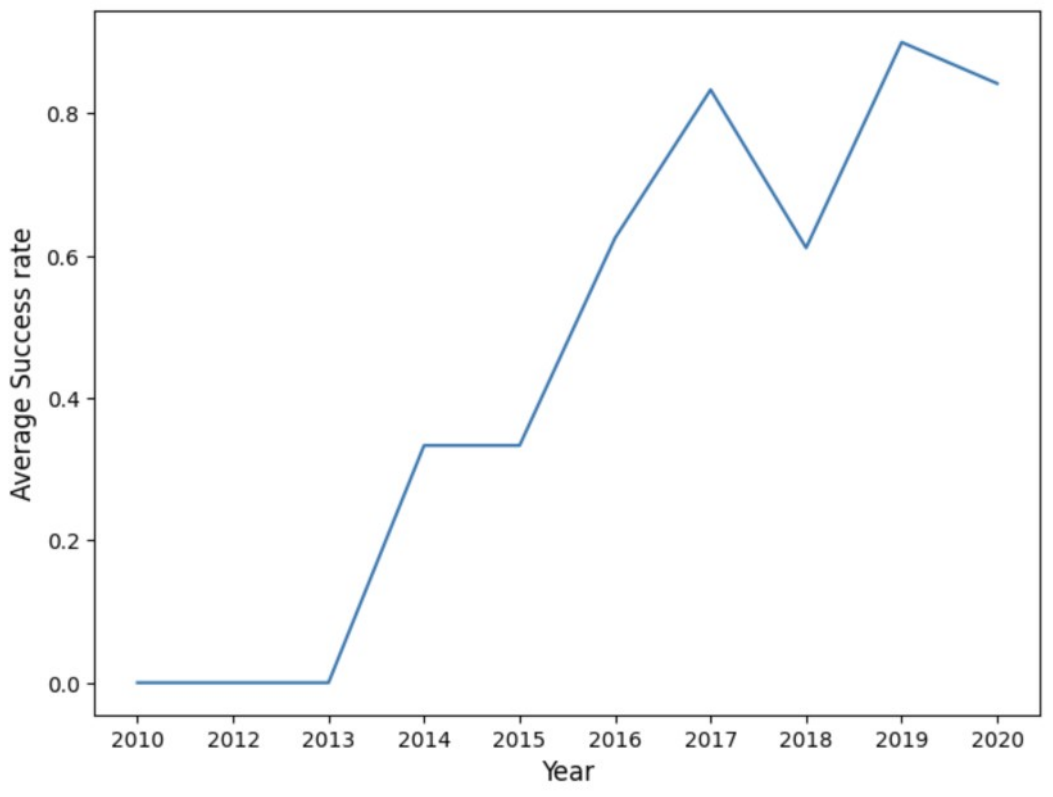
The mission success rate improved over time for all orbits, especially for LEO, but no clear pattern can be detected.

Pay load Mass vs. Orbit Type



With heavy payloads the successful landing rate are higher for LEO, ISS and PO orbits. However, for GTO but no clear pattern can be detected.

Launch Success Yearly Trend



The mission success rate increased in 2013 and kept increasing until 2020, with the exception of year 2018.

All Launch Site Names

```
1  %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

The DISTINCT keyword was used in the SQL query to find the unique values in the Launch_Site column.

Launch Site Names Begin with 'CCA'

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

Python

```
 * 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

The keyword LIKE uses the wildcard % with the string 'CCA%'. This ensures that the Launch_Site names start with 'CCA' by matching any characters following it in the column. The statement LIMIT 5 restricts the SQL query to return only 5 records.

Total Payload Mass

```
1  %sql SELECT SUM(payload_mass__kg_) AS TOTAL_PAYLOAD_MASS FROM SPACEXTBL WHERE customer='NASA (CRS)';
```

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

TOTAL_PAYLOAD_MASS
45596

The function SUM summates the total in the column payload_mass_kg_. The keyword WHERE filters the records to only perform calculations on customer equal to “NASA (CRS)”

Average Payload Mass by F9 v1.1

```
1  %sql SELECT AVG(payload_mass__kg_) AS AVG_PAYLOAD_MASS FROM SPACEXTBL WHERE booster_version='F9 v1.1';
```

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

AVG_PAYLOAD_MASS
2928.4

The function AVG calculates the mean in the column payload_mass_kg_. The keyword WHERE filters the records to only perform calculations on booster_version equal to “F9 v1.1”.

First Successful Ground Landing Date

```
1  %%sql  SELECT MIN(Date) AS First_Successful_Landing_Date FROM SPACEXTBL
2  WHERE Landing_Outcome = 'Success (ground pad)';
```

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

First_Successful_Landing_Date
2015-12-22

The function MIN finds the minimum value in the column Date. The keyword WHERE filters the records to only perform searching on Landing_Outcome equal to “Success (ground pad)”.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
1  %%sql SELECT Booster_Version FROM SPACEXTBL
2  WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;
```

✓ 0.0s

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

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

The keyword WHERE filters the records to only perform searching on Landing_Outcome equal to “Success (drone ship)”. The keyword AND specifies additional filter conditions: payload_mass_kg_ > 4000 and payload_mass_kg_ < 6000.

Total Number of Successful and Failure Mission Outcomes

```
1  %sql SELECT mission_outcome, COUNT(mission_outcome) AS TOTAL FROM SPACEXTBL GROUP BY mission_outcome;
2
```

✓ 0.0s

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

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

The function COUNT() counts the occurrences of each distinct value in the mission_outcome column. The function GROUP BY groups the rows based on the values in the mission_outcome column, with success = 100 and failure = 1.

Boosters Carried Maximum Payload

```
%%sql SELECT DISTINCT BOOSTER_VERSION AS "Booster Versions which carried the Maximum Payload Mass" FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ =(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);

```

✓ 0.0s

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

Booster Versions which carried the Maximum Payload Mass
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

The keyword DISTINCT retrieves unique booster versions. The keywords WHERE and SELECT MAX in the subquery find and retrieve the Booster Versions which carried the Maximum Payload Mass across all records.

2015 Launch Records

```
%%sql SELECT landing_outcome, booster_version, launch_site, DATE FROM SPACEXTBL
WHERE landing_outcome LIKE '%Failure (drone ship)%' AND (DATE LIKE '2015%');
```

✓ 0.0s

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

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

The SQL query returns the failed landing_outcomes in drone ship, booster versions, and launch site names in year 2015. The keyword WHERE filters the records to only perform searching on landing_outcome like “Failure (drone ship)”. The keyword AND specifies additional filter condition for year 2015.

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

```
%%sql SELECT LANDING_OUTCOME as "Landing Outcome", COUNT(LANDING_OUTCOME) AS "Total Count" FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY  LANDING_OUTCOME
ORDER BY COUNT(LANDING_OUTCOME) DESC;
```

✓ 0.0s

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

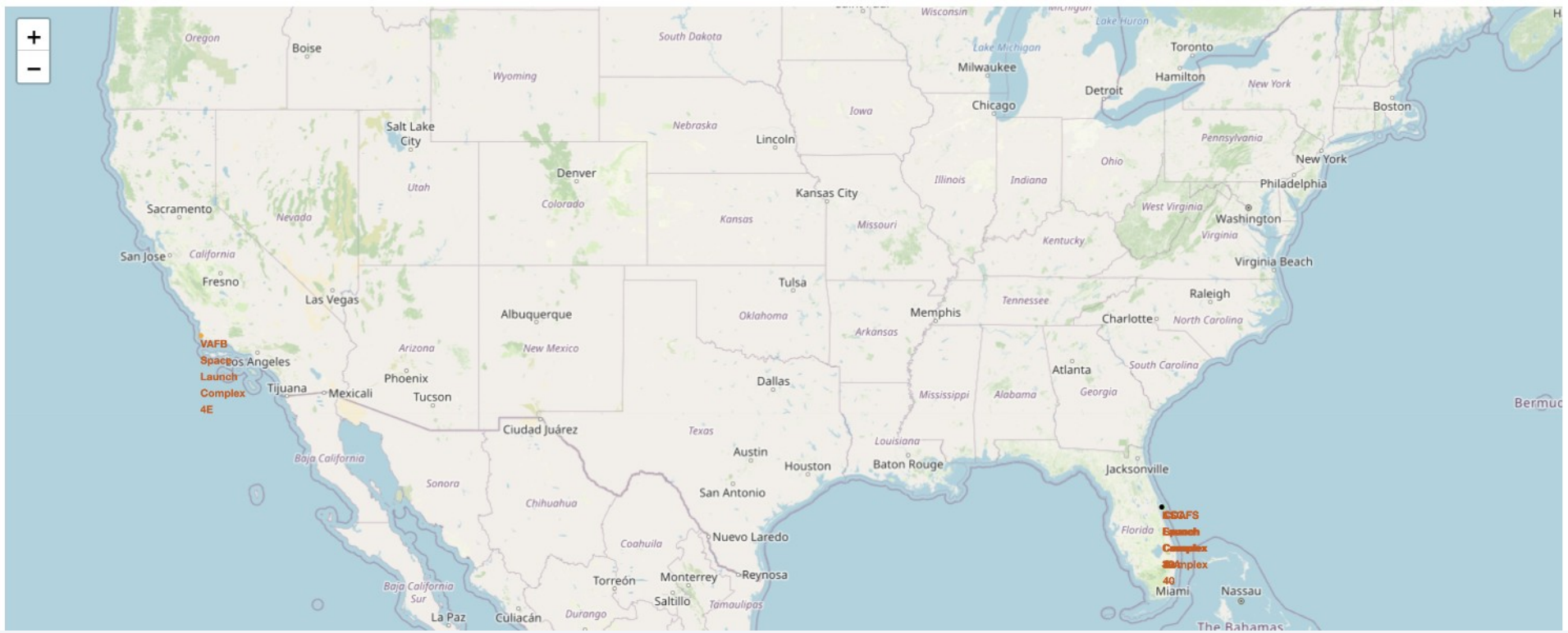
Landing Outcome	Total Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

The SQL query returns 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. The keyword COUNT retrieves the landing outcomes and their total counts. The keyword WHERE and BETWEEN filter the rows to include only those with dates falling within '2010-06-04' and '2017-03-20'. The keyword GROUP BY groups the results by the LANDING_OUTCOME column.

Section 3

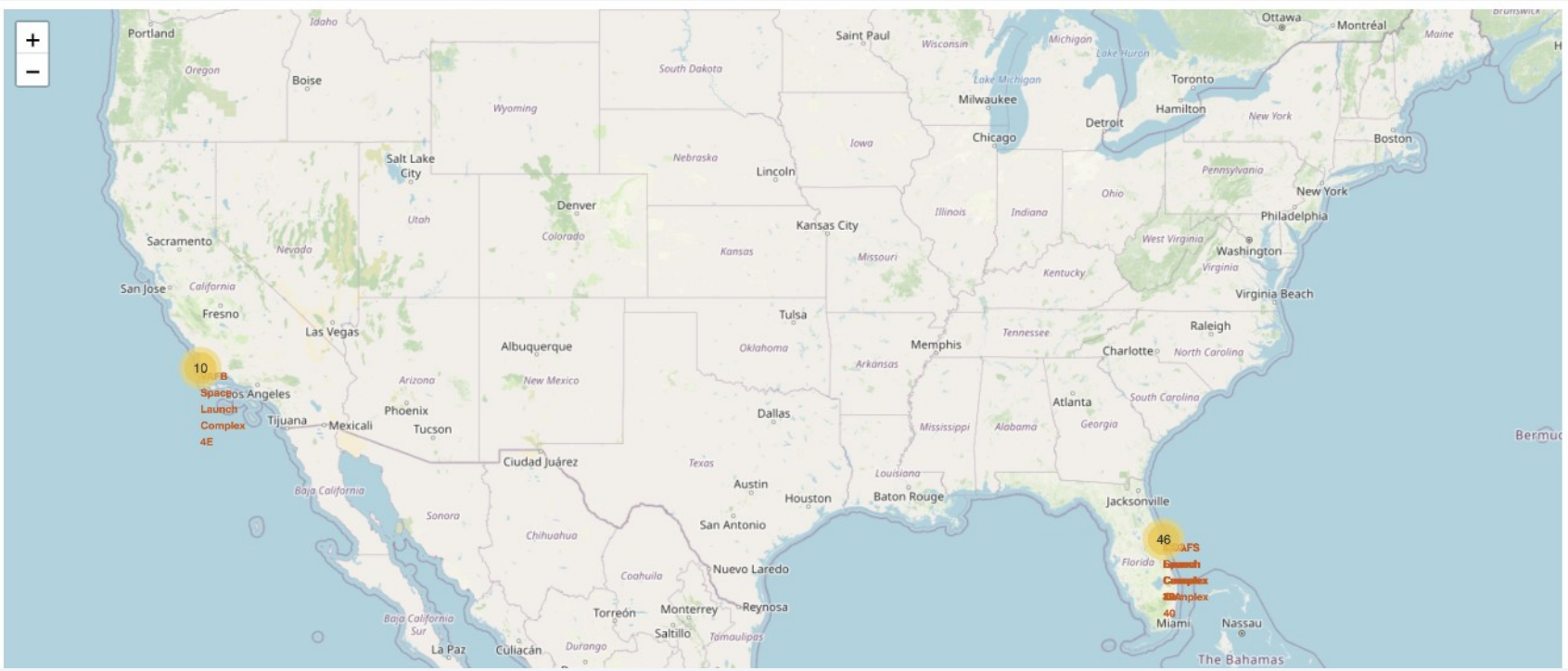
Launch Sites Proximities Analysis

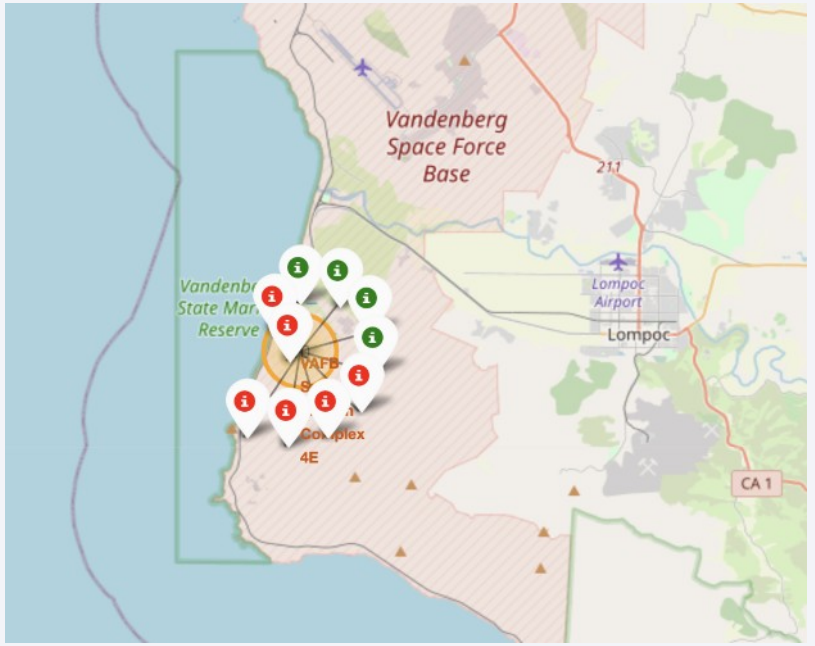
Map of all launch sites



All the launch sites are near to the USA coastal lines (in Florida and California) for safety reasons, e.g.: avoiding densely populated area. Launch sites are also in proximity to the Equator, as it takes less fuel to get into space from the Equator due to Earth's rotation.

Map of launch outcomes by site



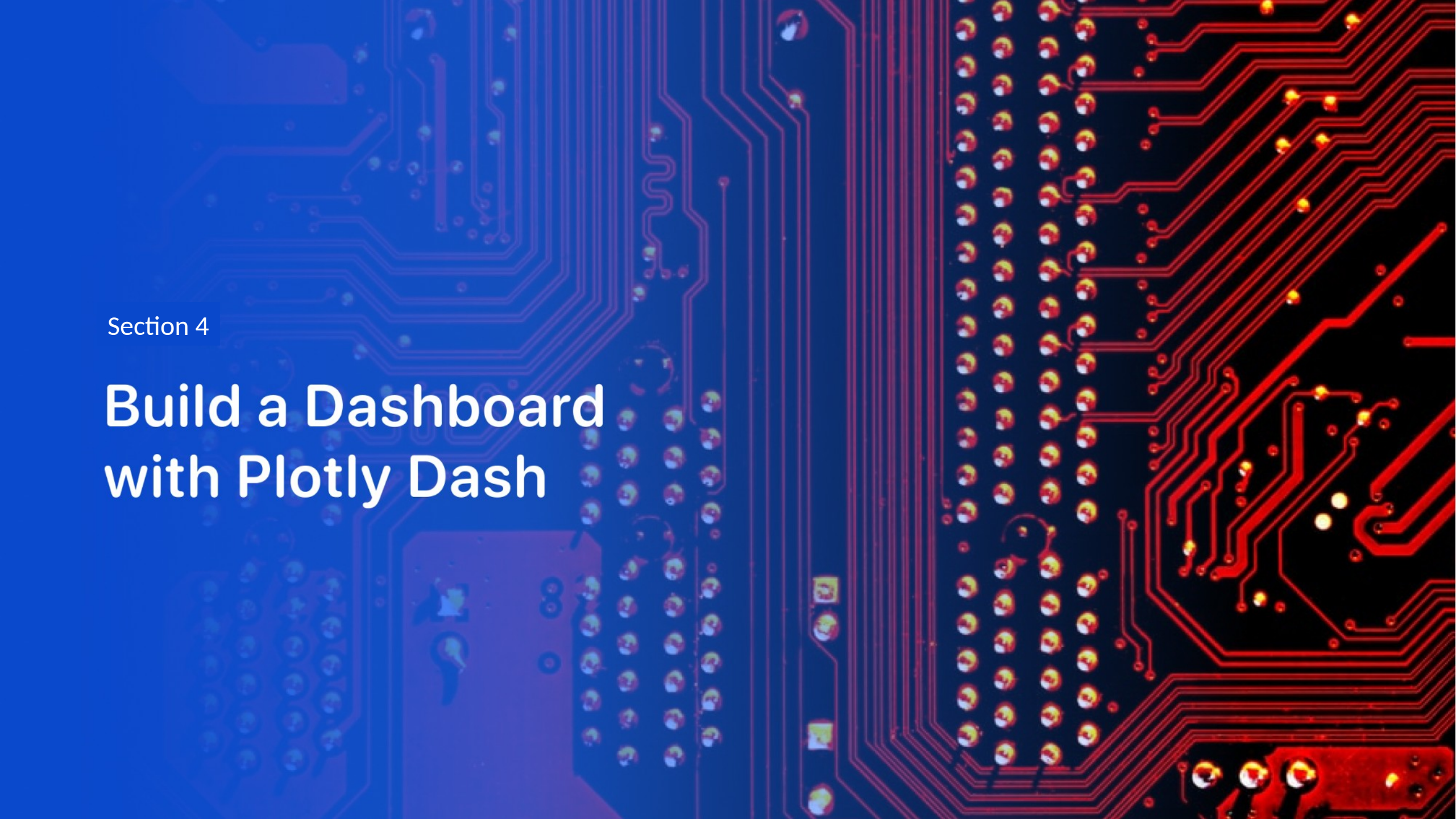


The first map (launch clusters): displays clusters of launches for each launch site. It helps identify areas with high launch activity. The second map (launch outcome markers): the red markers indicate launch failures, the green markers indicate successful launches.

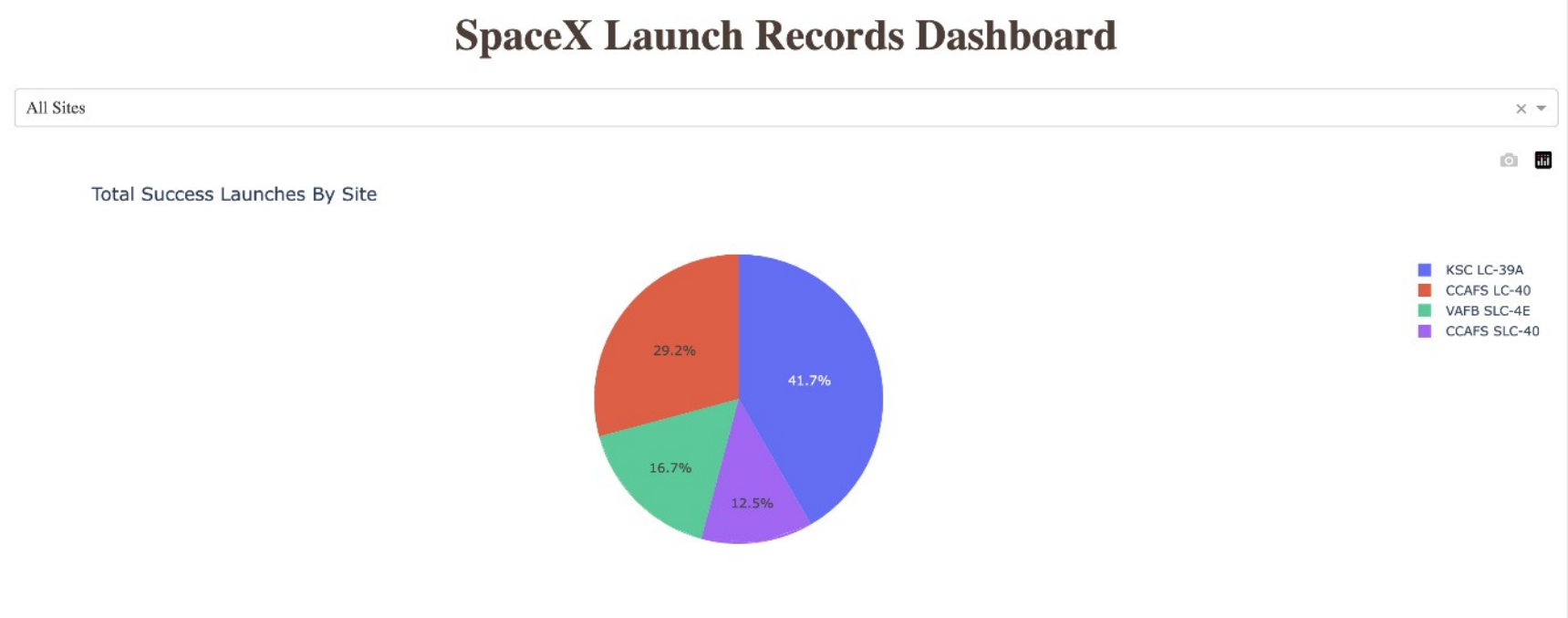
Map of site CCAFS SLC-40 and its proximities

Launch sites are strategically located near coastlines for safety reasons, allowing options for aborting launches over water and minimizing risks to people and property. Proximity to highways facilitates easy transportation of equipment, while railways enable efficient transport of heavy cargos. Importantly, launch sites intentionally avoid densely populated areas to minimize danger to residents.

Distance of site CCAFS SLC-40 from coastal line (0.86 Km) is reported.

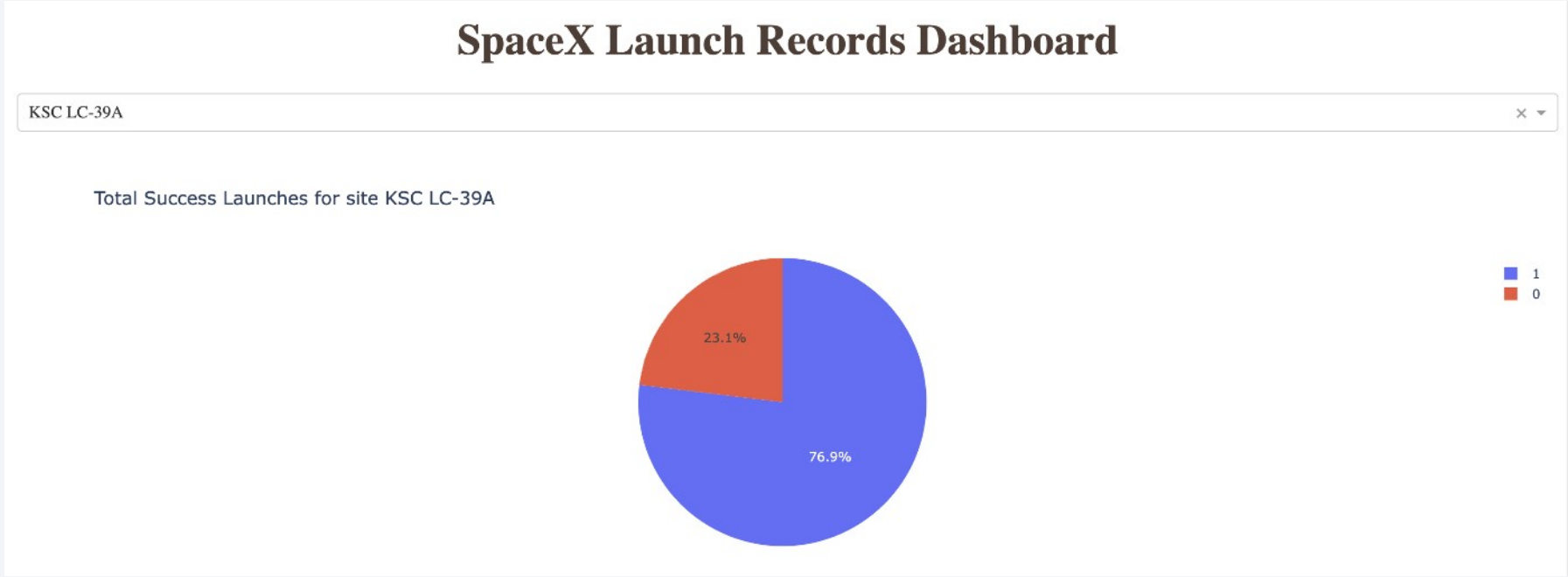

Section 4
Build a Dashboard with Plotly Dash

Total Successful Launches By Site



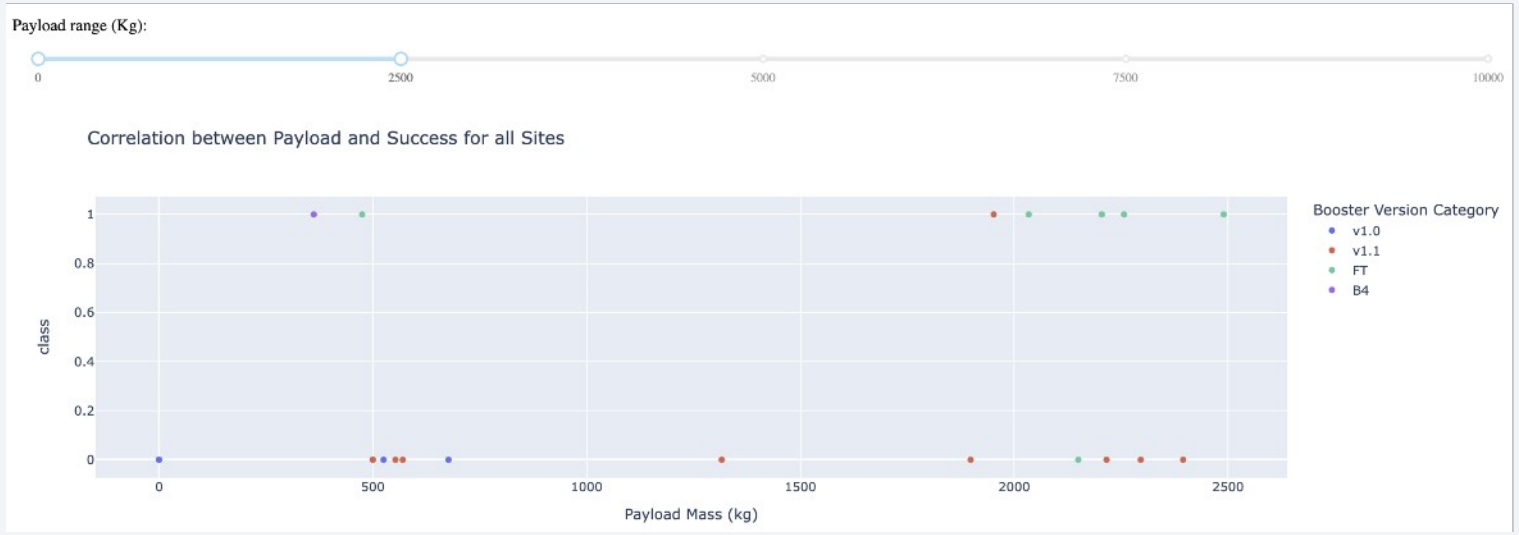
Site KSC LC-39A shows the highest success rate of 41.7%. Site CCAFS SLC-40 shows the lowest success rate of 12.5%. The launch site appears to be a relevant factor in deciding the success of a mission.

Successful Launches for site KSC LC-39A



Site KSC LC-39A shows 76.9% of successful launches, and 23.1% of unsuccessful launches.

Plot of Payload Mass vs Launch Outcome for all sites (1)



Overall, launch outcomes tends to be less successful with increasing payloads.

Payloads under 6,000 Kg and FT boosters appears to be the most successful combination.

1: successful landing

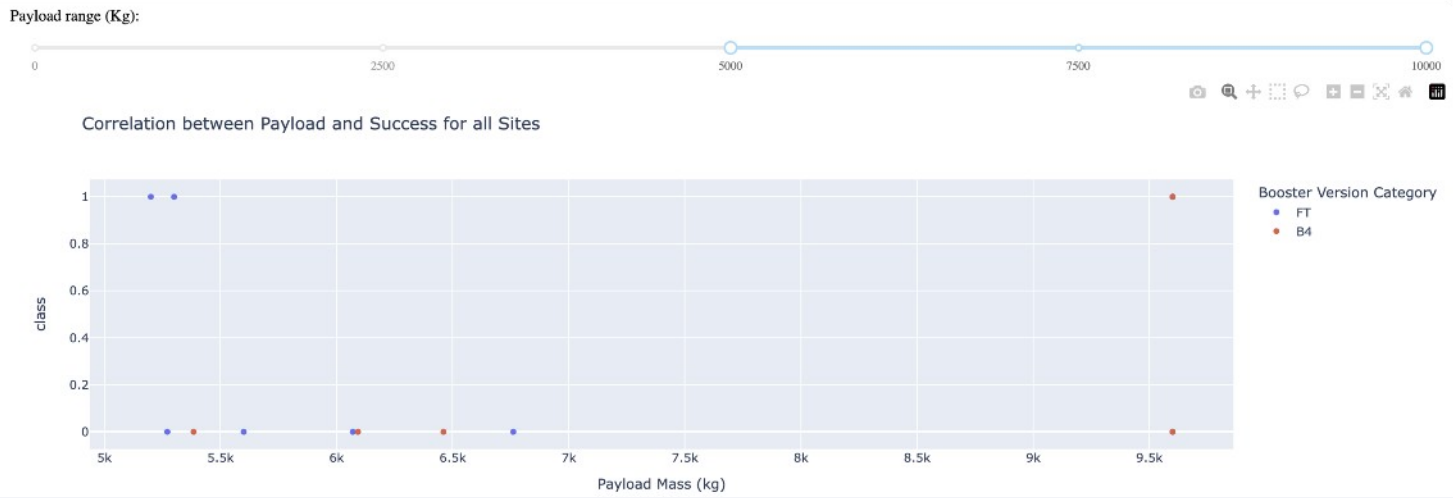
0: unsuccessful landing

Plot of Payload Mass vs Launch Outcome for all sites (2)

No conclusion can be drawn concerning risk of launches over 7,000 Kg.

1: successful landing

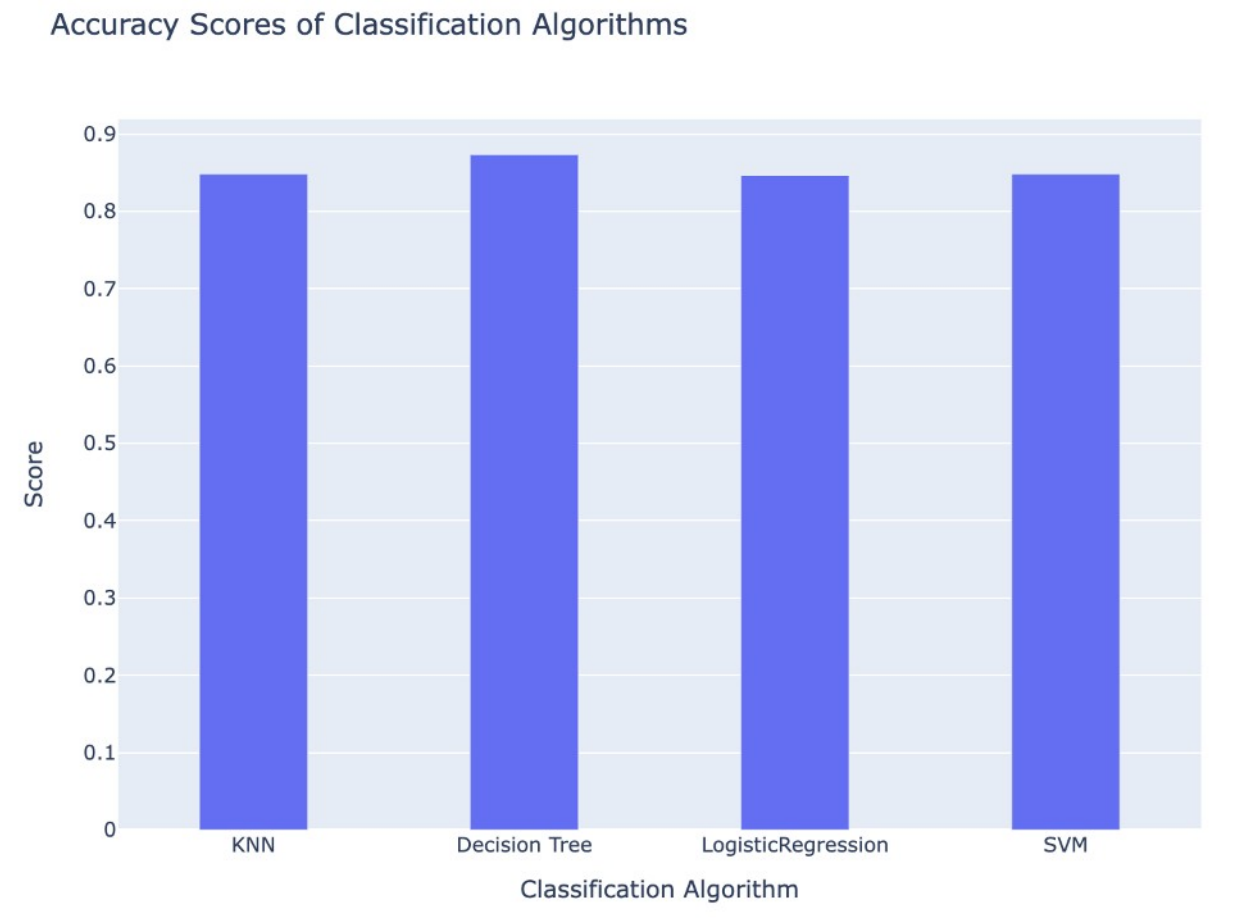
0: unsuccessful landing



Section 5

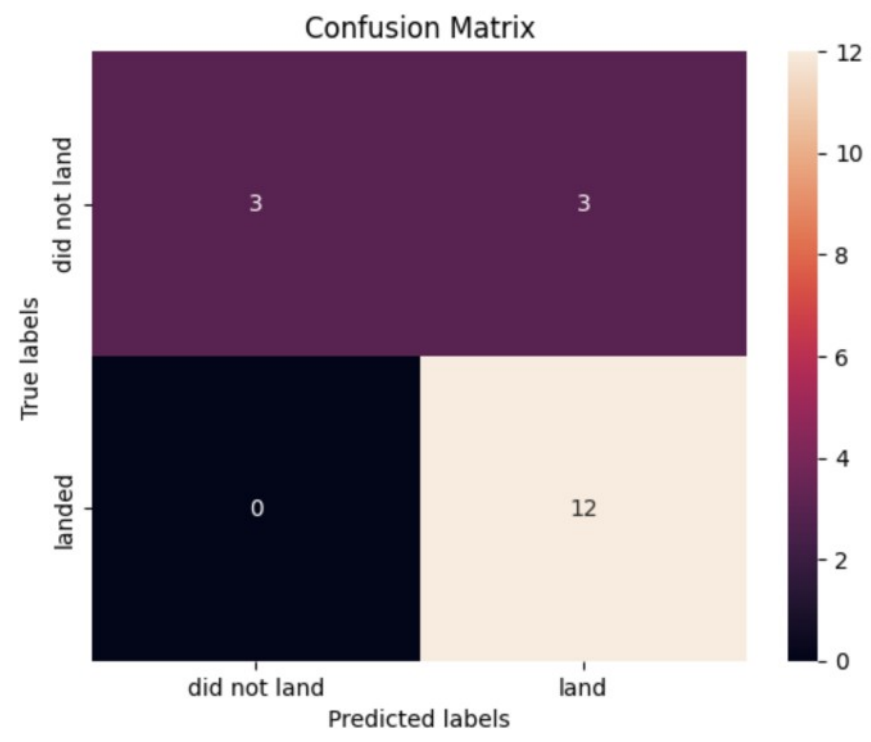
Predictive Analysis (Classification)

Classification Accuracy



Four classification models were used: K-Nearest Neighbor, Decision Tree, Logistic Regression and Support Vector Machine. The models resulted in comparable accuracy, with the Decision Tree classifier showing the highest accuracy at approximately 87%.

Confusion Matrix



The confusion matrices are identical for all 4 classification models.

True Positives: the models seem good at predicting successful landings. This is crucial information for space agencies.

False Positives: however, the models sometimes predict successful landings when rockets actually crash. This is a serious issue.

True Negatives: the models are good at predicting unsuccessful landings.

Overall Performance: while the models seem to perform well in predicting successful and unsuccessful landings, the false positive results need to be carefully considered.

Conclusions

Mission success factors: launch site, orbit type, payload mass and the number of previous flights leading to a higher success rate. This shows a technological improvement over time.

Launch site: KSC LC-39A has the highest success rate, but the reason is unclear.

Orbit type: Orbits SSO, HEO, GEO, and ES-L1 show the highest success rate.

Payload mass: lighter payloads generally perform better. Payloads under 6,000 Kg and FT boosters appears to be the most successful combination.

Number of previous flights: the success rate for the missions increases over time.

Model choice: four classification models were tested: K-Nearest Neighbor, Decision Tree, Logistic Regression and Support Vector Machine. The models resulted in comparable accuracy, with the Decision Tree classifier showing the highest accuracy at approximately 87%.

Key findings: Lighter payloads, specific orbits, and repeated attempts improve success rates. KSC LC-39A is the best launch site. Decision Tree is the preferred model for predicting a successful landing of the first stage of rockets.

Appendix

For notebooks, datasets and presentations follow this GitHub repository link:
https://github.com/Carlo-bits/IBM-Applied-Data-Science-Capstone-Project/tree/main

Thank you!