IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Lucas Hoff Schmidt
13-02-2025

Made by L.H.S.

Outline

- Executive Summary
- Introduction
- Methodology
- Insights drawn from EDA
- Launch Sites Proximities Analysis
- Build a dashboard with Plotly Dash
- Predictive Analysis (Classification)
- When should SpaceY bid for a launch?
- Conclusion
- Appendix

Executive Summary

- This project analyzes the outcome of landing stage one of a rocket in a space launch for SpaceX to determine whether it is possible for a new contender called SpaceY, to compete for a given launch project. Using API data and webscraping we performed exploratory data analysis using visualization libraries and machine learning techniques to find the model that most accurately could predict the outcome of a space launch.
- We learned that the year, payload mass and launch site are key predictors in determining the outcome of landing the stage 1 of a rocket. Testing different classification models yielded an accuracy of 83%.

Introduction

- This project is about determining whether or not SpaceX will reuse the first stage of a rocket in a given scenario, to establish whether a new company called SpaceY can compete with it. If SpaceX can reuse the first stage, they can save about 100 million dollars, making competing against them much harder.
- In this project we seek to answer which features have the greatest impact in determining whether the first stage will land successfully, and use these features to train a machine learning model to predict the outcome of the launch.

Section 1

Methodology

Methodology

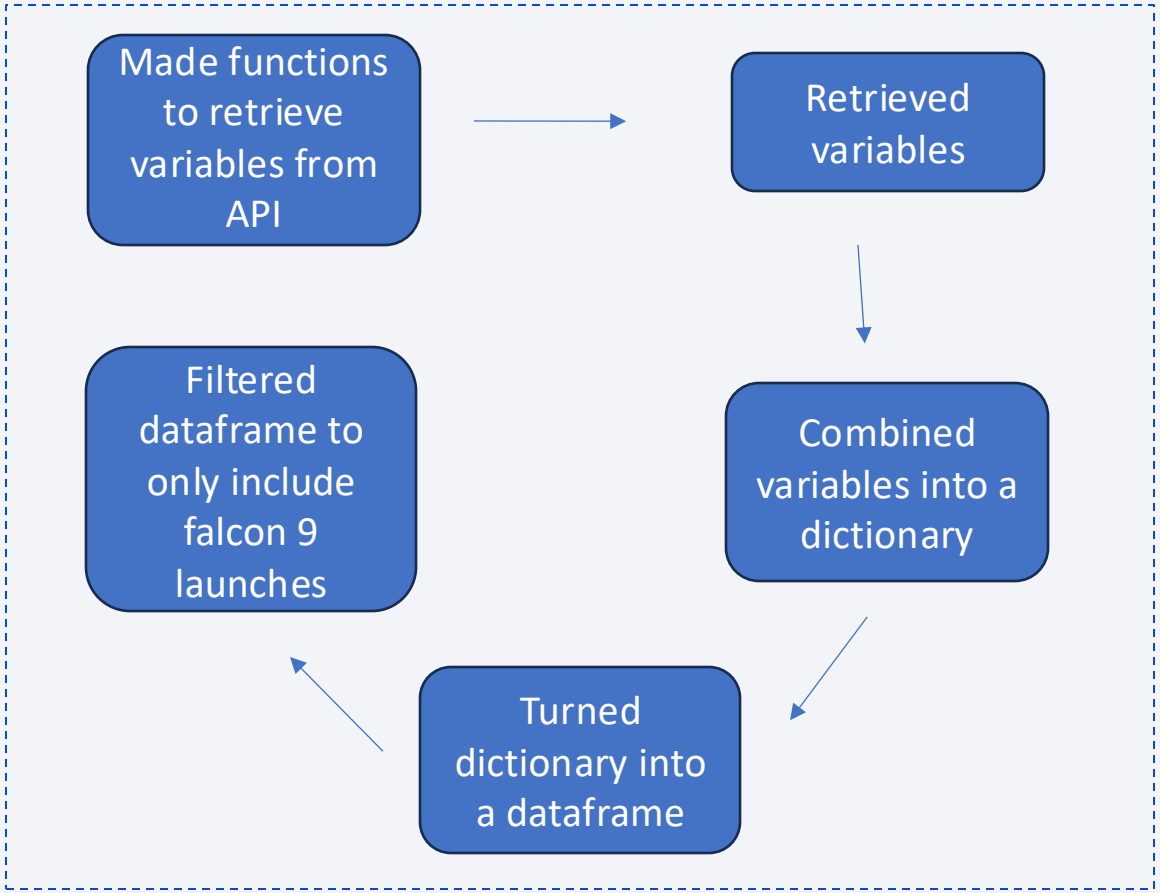
- Data collection
- Data wrangling
- Exploratory data analysis (EDA) using visualizations and SQL
- Interactive visual analytics using Folium and Plotly Dash
- Predictive analysis evaluation of classification models

Data Collection

- Used the API of SpaceX to gather launch data such as launch site, booster version and outcome of the launch and transformed the data into a pandas dataframe featuring falcon 9 launches.
- Performed webscraping from a static wikipedia page from 9th June 2021 featuring a list of Falcon 9 and Falcon Heavy launches and parsed the data to transform an html table into a pandas dataframe.

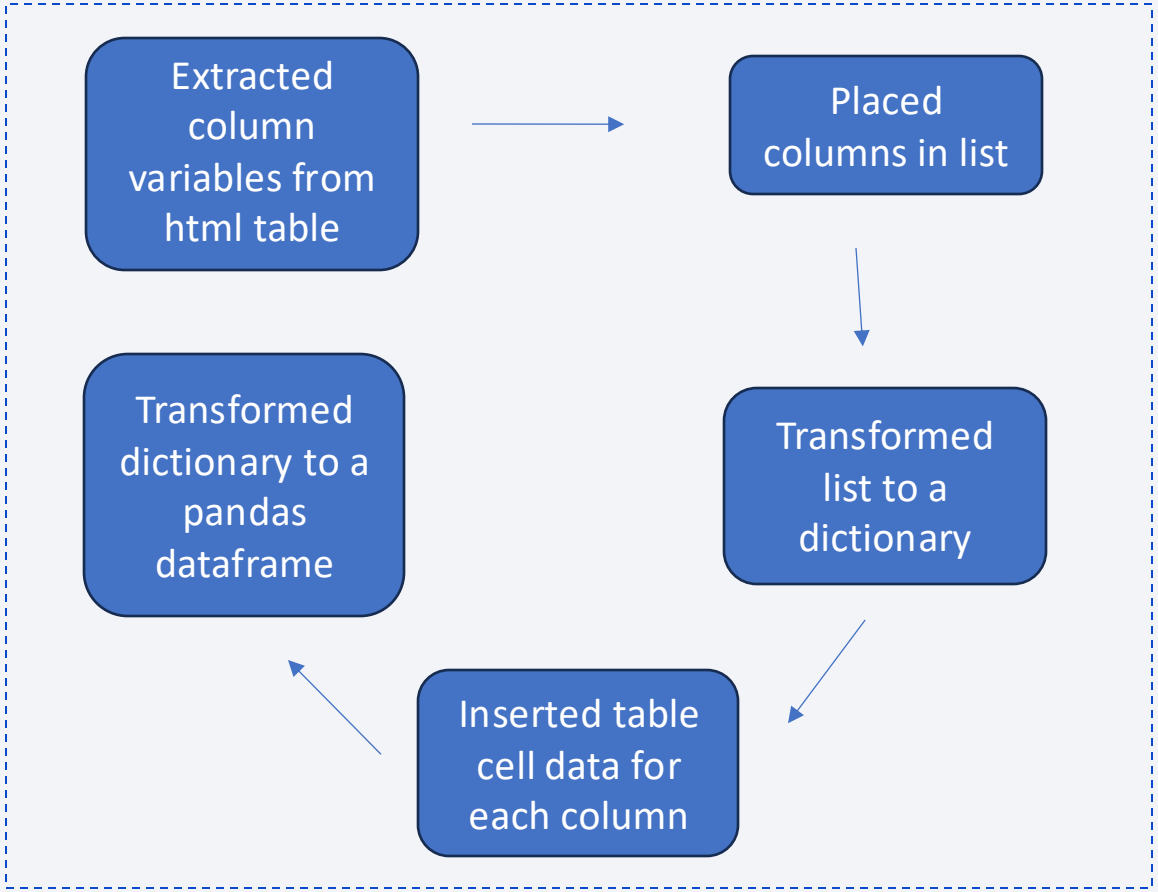
Data Collection – SpaceX API

- REST calls:
  - requests.get("https://api.spacexdata.com/v4/rockets/"+str(x)).json()
  - requests.get("https://api.spacexdata.com/v4/launchpads/"+str(x)).json()
  - requests.get("https://api.spacexdata.com/v4/payloads/"+load).json()
  - requests.get("https://api.spacexdata.com/v4/cores/"+core['core']).json()
  - requests.get(spacex_url)
  - requests.get(static_json_url)
- SpaceX API jupyter notebook:
  - https://github.com/LucasHoffSchmidt/IBM_Data_Science_Capstone_Project_SpaceX/blob/main/Jupyter_Notebooks/Data_Collection_API.ipynb



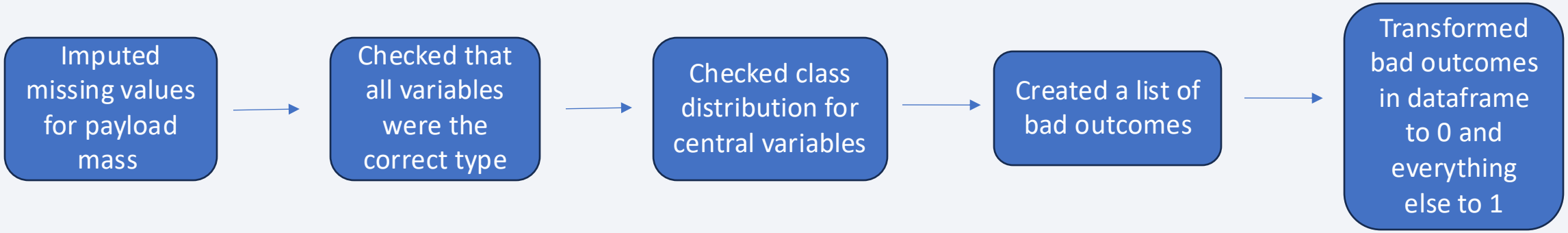
Data Collection - Scraping

- Process
  - Extracted all column variables from html table
  - Placed column variables in a list and transformed the list to a dictionary
  - Inserted table cell data for each column
  - Transformed dictionary to a pandas dataframe
- SpaceX webscraping jupyter notebook:
  - https://github.com/LucasHoffSchmidt/IBM_Data_Science_Capstone_Project_SpaceX/blob/main/Jupyter_Notebooks/Data_Collection_Webscraping.ipynb



Data Wrangling

- Key transformations
  - Imputed missing values for payload mass by using the mean
  - Encoded outcomes to binary values for failed or succeeded outcome



- SpaceX data wrangling jupyter notebooks:
  - Imputing:https://github.com/LucasHoffSchmidt/IBM_Data_Science_Capstone_Project_SpaceX/blob/main/Jupyter_Notebooks/Data_Collection_API.ipynb
  - Encoding:https://github.com/LucasHoffSchmidt/IBM_Data_Science_Capstone_Project_SpaceX/blob/main/Jupyter_Notebooks/Data_Wrangling.ipynb

EDA with Data Visualization

- Charts used
  - Catplots for understanding the correlation between multiple variables and the outcome
  - Barplot for understanding how the success rate changes based on the orbit type
  - Lineplot for understanding how the average success rate changed over the years
- SpaceX pandas and matplotlib exploratory data analysis jupyter notebook:
  - https://github.com/LucasHoffSchmidt/IBM_Data_Science_Capstone_Project_SpaceX/blob/main/Jupyter_Notebooks/Exploratory_Data_Analysis_Pandas_and_Matplotlib.ipynb

EDA with SQL

- Queries
  - Unique launch sites and launch sites beginning with CCA
  - Total payload mass carried by boosters launched by NASA and average payload mass for booster version F9 v1.1
  - Date for the first successful landing on ground pad and Total number of successful and failure mission outcomes
  - Boosters that have landed successfully on a drone ship with a payload mass between 4000 and 6000
  - Names of the booster versions that have carried the maximum payload mass
  - Records for failure landing outcomes on drone ship, months, booster versions and launch site for the months in 2015
  - Ranking of outcomes from 2010-06-04 to 2016-03-20 in descending order
- SpaceX SQL exploratory data analysis jupyter notebook:
  - https://github.com/LucasHoffSchmidt/IBM_Data_Science_Capstone_Project_SpaceX/blob/main/Jupyter_Notebooks/Exploratory_Data_Analysis_SQL.ipynb

Build an Interactive Map with Folium

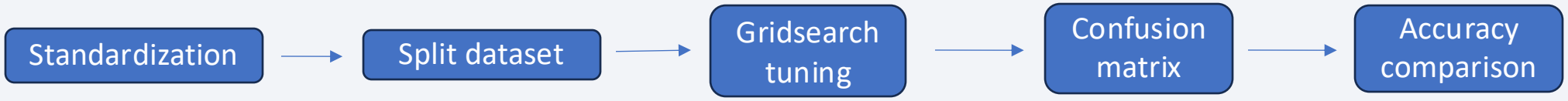
- Process
  - Marked all launch sites on a map with a circle and descriptive text to compare their locations
  - Marked each success or failed launch for each site with a green or red marker, to get an overview of which site had the highest success rate
  - Marked the distances from each launch site to its proximities coast, city, railline and highway with polylines and a text describing the distance to get an overview of average distances to each proximity
- SpaceX folium interactive visual analytics jupyter notebook:
  - https://github.com/LucasHoffSchmidt/IBM_Data_Science_Capstone_Project_SpaceX/blob/main/Jupyter_Notebooks/Interactive_Visual_Analytics_Folium.ipynb

Build a Dashboard with Plotly Dash

- Process
  - Made a dropdown list for displaying data for all launch sites or an individual site
  - Made a pie chart to showcase the success rate for each launch site, allowing for comparisons between sites
  - Made a range slider of the payload mass connected to a scatterplot of the currently selected launch sites to see how different ranges of payload mass influences the outcome for the launch site(s)
- SpaceX dash app interactive visual analytics jupyter notebook:
  - https://github.com/LucasHoffSchmidt/IBM_Data_Science_Capstone_Project_SpaceX/blob/main/Interactive_Visual_Analytics_Dash_App.py

Predictive Analysis (Classification)

- Process
  - Standardized relevant features
  - Split dataset into training and testing
  - Applied gridsearch tuning for the classification models logistic regression, support vector machine, decision tree and k nearest neighbors.
  - Made a confusion matrix for each classification model displaying true and false positives and negatives
  - Compared their accuracy to determine the best classification model in predicting the outcome



- SpaceX machine learning prediction jupyter notebook:
  - https://github.com/LucasHoffSchmidt/IBM_Data_Science_Capstone_Project_SpaceX/blob/main/Jupyter_Notebooks/Machine_Learning_Prediction.ipynb

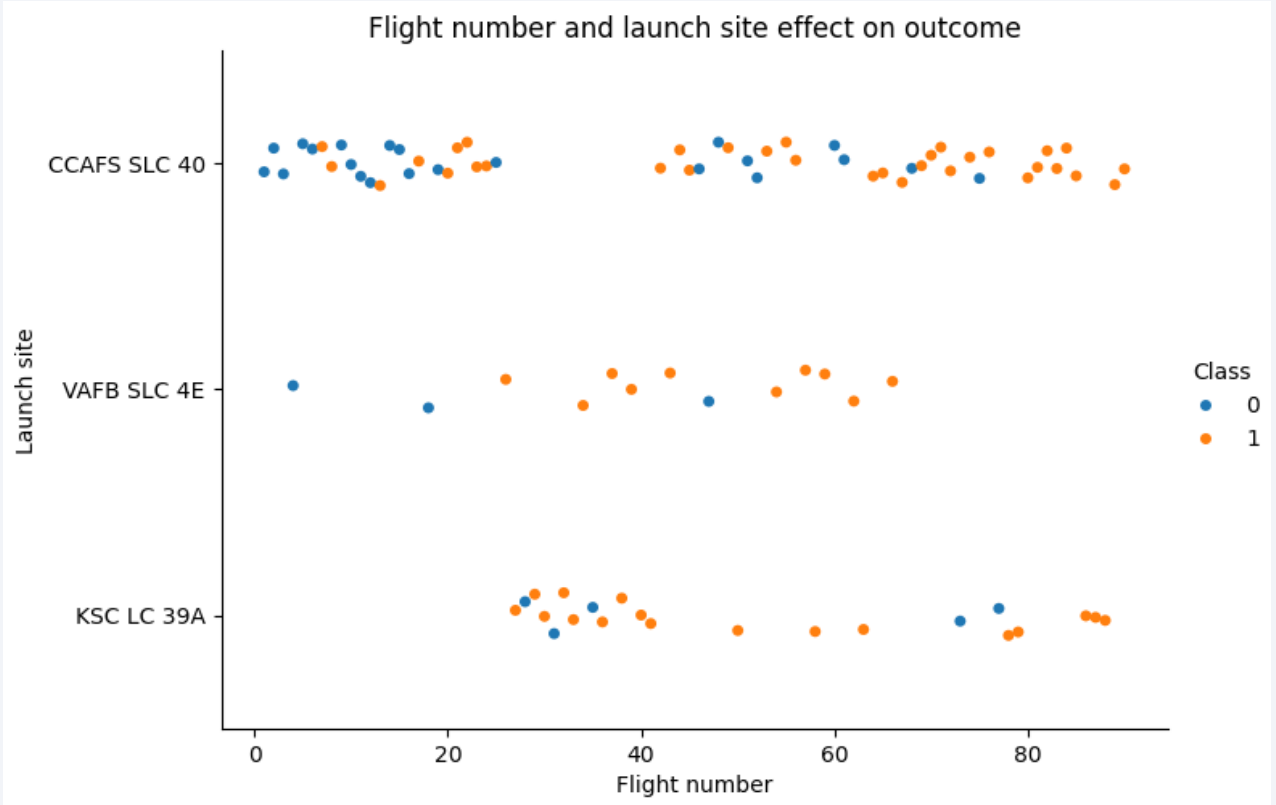
Section 2

Insights drawn from EDA

Made by L.H.S.

Flight Number vs. Launch Site

- Most failed outcomes happen in earlier flights
- The launch site KSC LC-39A has the highest success rate
- CCAFS SLC 40 is by far the launch site that is used most frequently

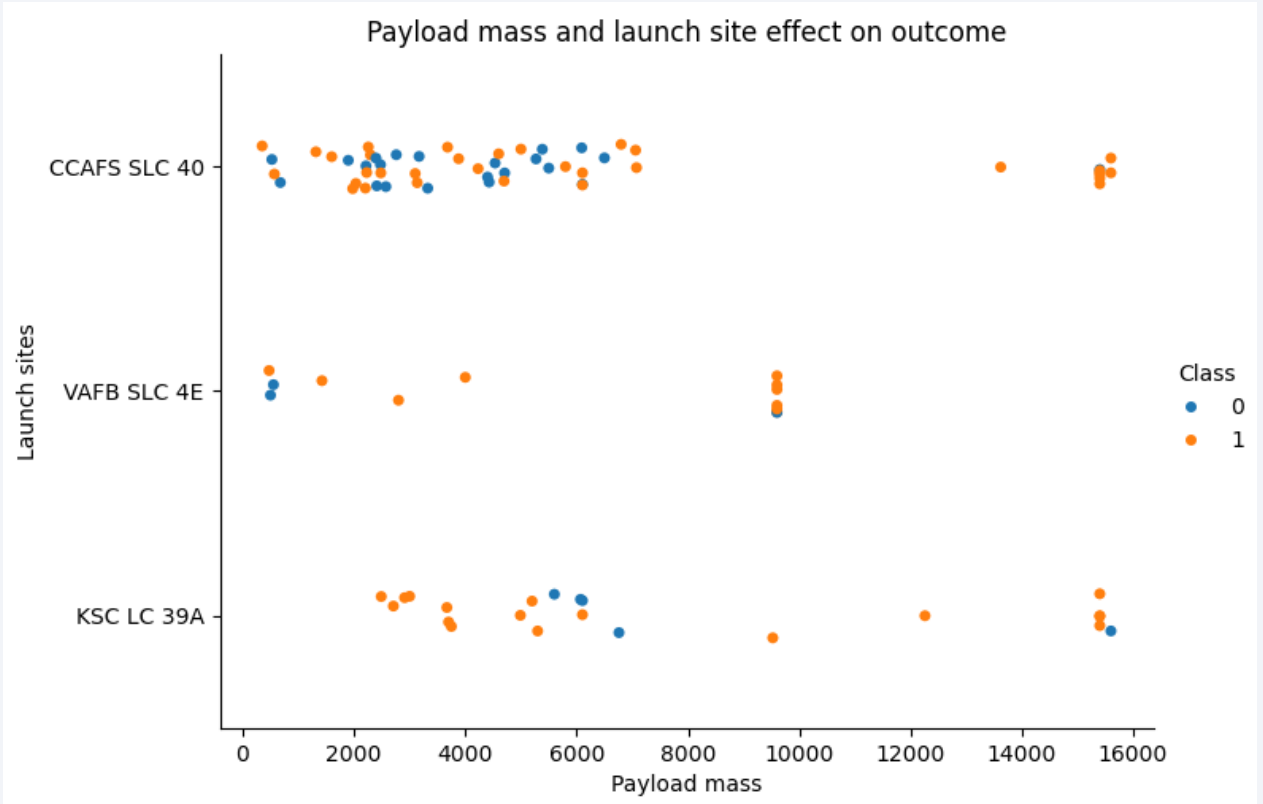


Scatter plot over flight number, launch site and class outcome

Made by L.H.S.

Payload vs. Launch Site

- VAFB has no rockets launched for heavy payload mass greater than 10000 kg
- The likelihood of success is much higher when the payload mass is greater than 8000 kg
- KSC has no rockets with a light payload mass lower than 2000 kg

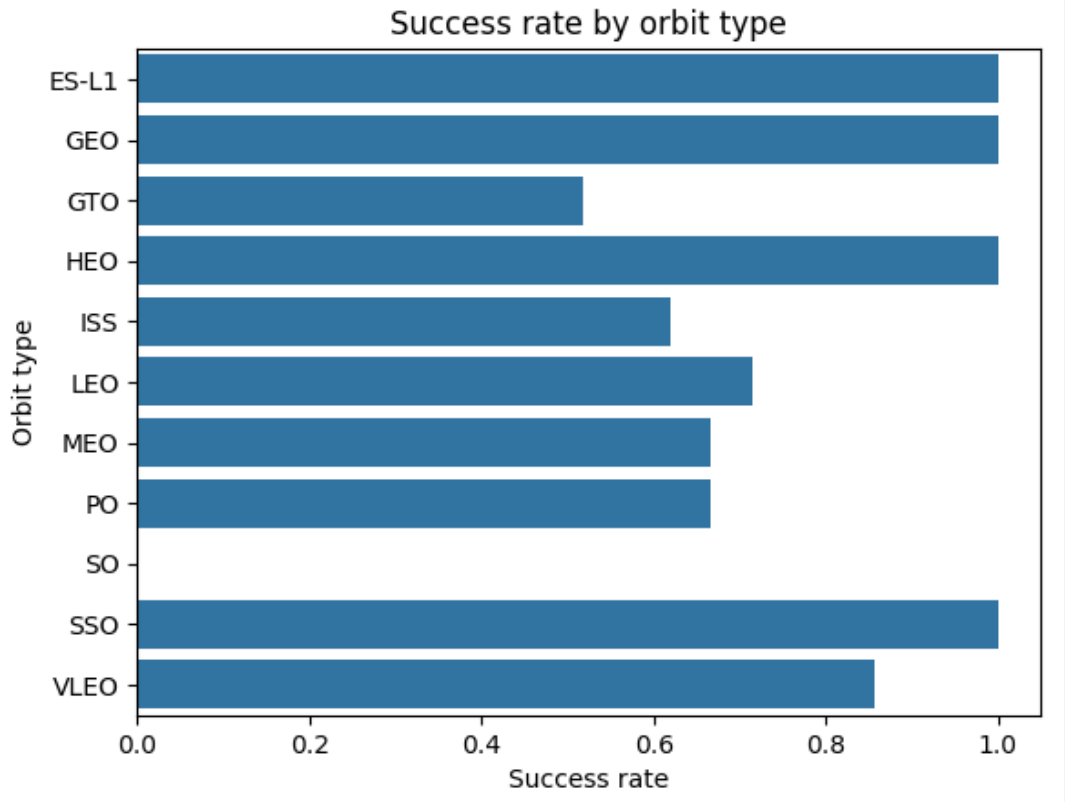


Scatter plot over payload mass, launch site and class outcome

Made by L.H.S.

Success Rate vs. Orbit Type

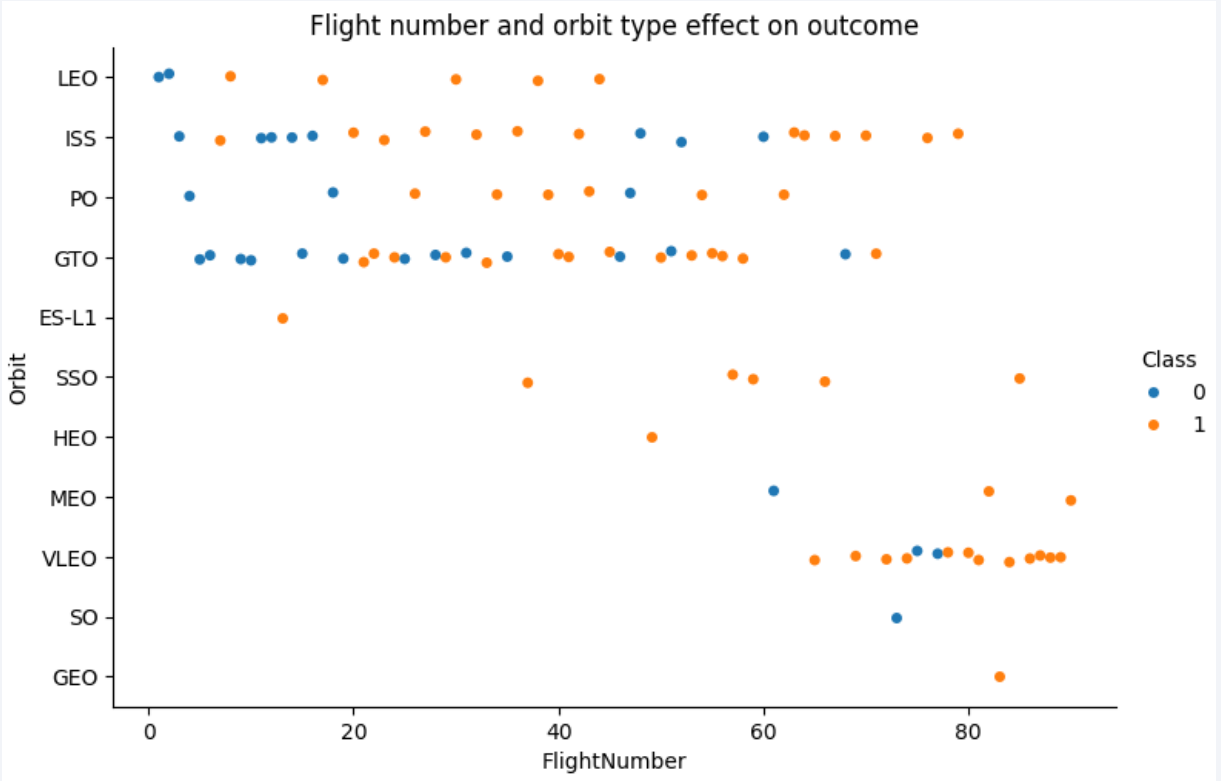
- The orbit types of ES-L1, GEO, HEO and SSO has a 100% success rate
- The orbit type with the lowest success rate is GTO
- The orbit success rate is never less than 50%



bar plot over orbit type and class outcome

Flight Number vs. Orbit Type

- The orbit types with a 100% success rate doesn't have a lot of flights
- The most frequent orbit types are ISS and GTO
- The rarest orbit types are HEO, SO and GEO

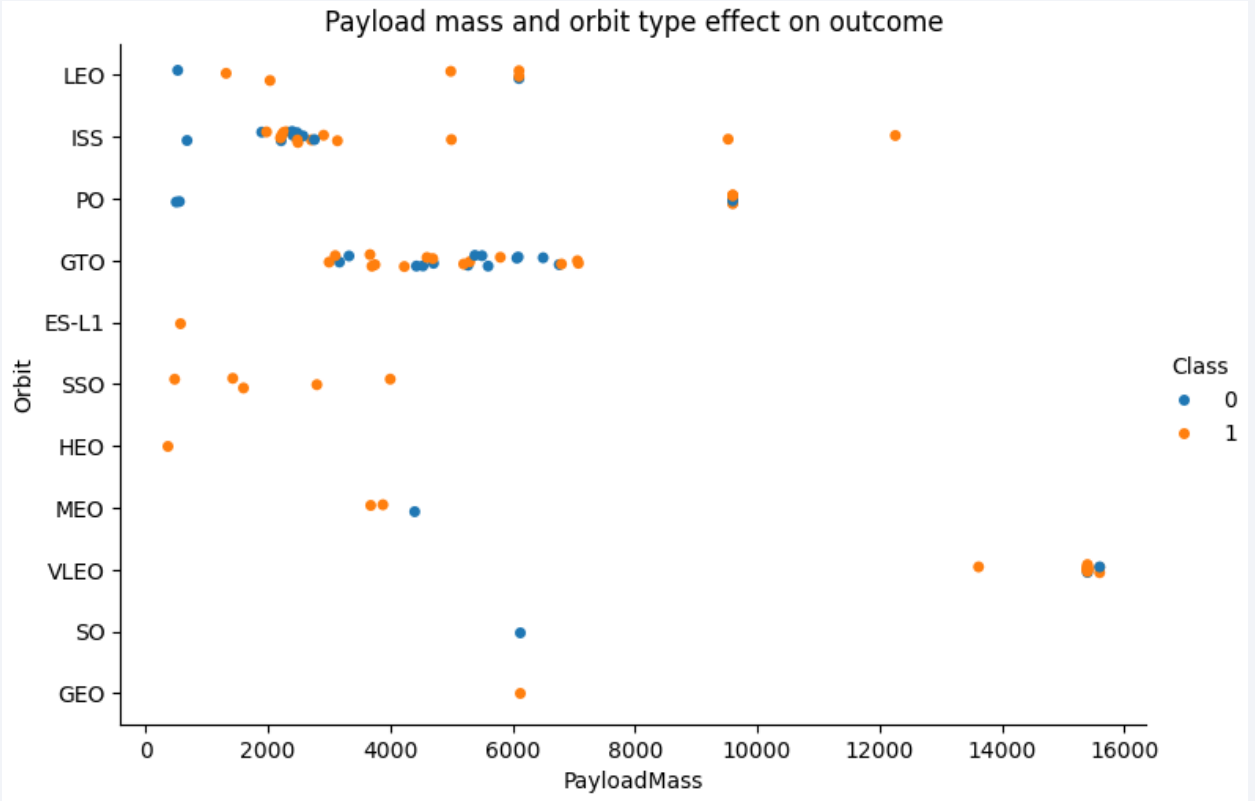


scatter plot over flight number, orbit type and class outcome

Made by L.H.S.

Payload vs. Orbit Type

- Only ISS, PO and VLEO orbits have heavy payloads of more than 10000 kg
- GTO only have payloads in the interval between 2000 and 8000 kg
- ES-L1, SSO and HEO only have light payloads under 4000 kg

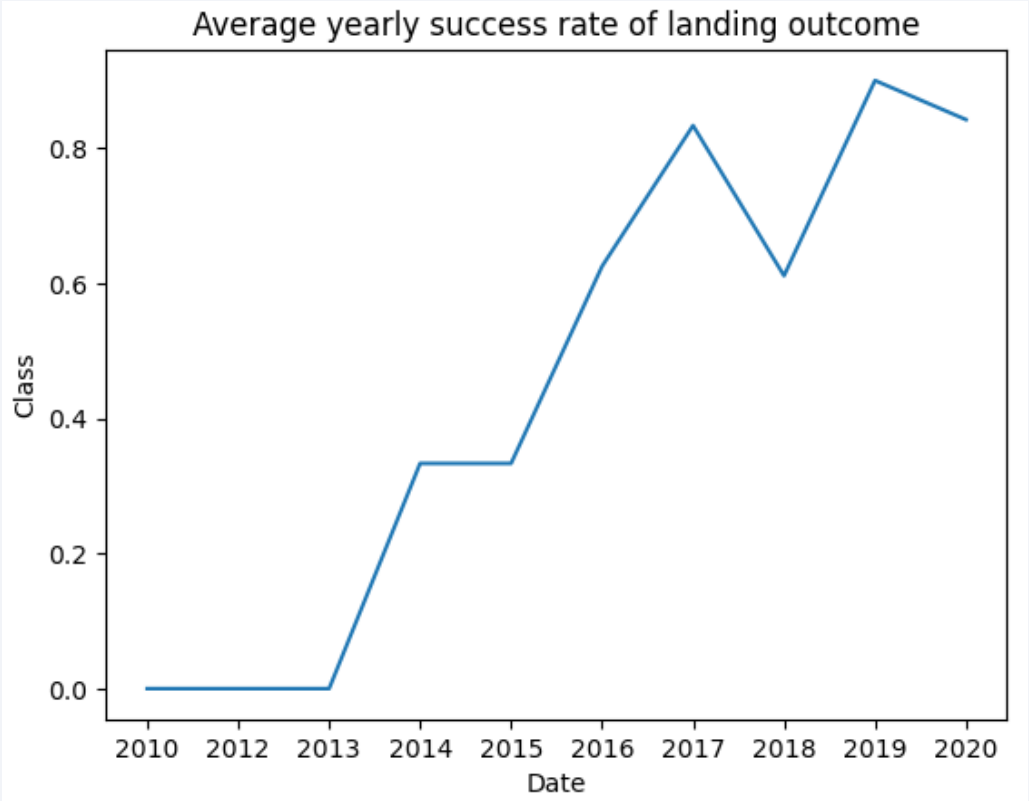


scatter plot over payload mass, orbit type and class outcome

Made by L.H.S.

Launch Success Yearly Trend

- Initially the success rate of landings were 0%
- Since 2013 the success rate of landings have been steadily increasing to past 80%
- 2018 featured a significant drop in success rate of landings



line plot of average yearly success rate of landing outcome

All Launch Site Names

- Query
  - **%%sql SELECT DISTINCT** Launch_Site
  - **FROM** SPACEXTABLE;
- Interpretation
  - This query shows the names of the unique launch sites

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

Query results

Launch Site Names Begin with 'CCA'

- Query
  - **%%sql SELECT** *
  - **FROM** SPACEXTABLE
  - **WHERE** "Launch_Site" **LIKE** "CCA%"
  - **LIMIT** 5;
- Interpretation
  - This query shows 5 records of launch sites beginning with CCA

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

Query results

Total Payload Mass

- Query
  - %%sql SELECT SUM(PAYLOAD_MASS__KG_)
  - FROM SPACEXTABLE
  - WHERE Payload LIKE "%CRS%";
- Interpretation
  - This query shows the total payload mass carried by boosters launched by NASA

SUM(PAYLOAD_MASS__KG_)
111268

Query results

Average Payload Mass by F9 v1.1

- Query
  - %%sql SELECT AVG(PAYLOAD_MASS__KG_)
  - FROM SPACEXTABLE
  - WHERE Booster_Version LIKE "F9 v1.1%";
- Interpretation
  - This query shows the average payload mass carried by booster version F9 v1.1

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

Query results

First Successful Ground Landing Date

- Query
  - **%%sql SELECT MIN**(Date) **AS** "first successful landing in ground pad"
  - **FROM** SPACEXTABLE
  - **WHERE** Landing_Outcome = "Success (ground pad)";
- Interpretation
  - This query shows the date of the first successful landing outcome on ground pad

first successful landing on ground pad
2015-12-22

Query results

Successful Drone Ship Landing with Payload between 4000 and 6000

- Query
  - **%%sql SELECT** Booster_Version
  - **FROM** SPACEXTABLE
  - **WHERE** Landing_Outcome = "Success (drone ship)" **AND** (PAYLOAD_MASS__KG_ > 4000 **AND** PAYLOAD_MASS__KG_ < 6000);
- Interpretation
  - This query shows the names of boosters which have had success landing on a drone ship with a payload mass between 4000 and 6000

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

Query results

Total Number of Successful and Failure Mission Outcomes

- Query
  - %%sql
  -
  - SELECT
  - COUNT(CASE WHEN Landing_Outcome LIKE "%Failure%" THEN 1 END) AS "Total_Failures",
  - COUNT(CASE WHEN Landing_Outcome LIKE "%Success%" THEN 1 END) AS "Total_Successes"
  - FROM SPACEXTABLE;
- Interpretation
  - This query shows the total number of successful and failed mission outcomes

Total_Failures	Total_Successes
10	61

Query results

Boosters Carried Maximum Payload

- Query
  - **%%sql**
  -
  - **SELECT** Booster_Version
  - **FROM** SPACEXTABLE
  - **WHERE** PAYLOAD_MASS__KG_ = (**SELECT MAX**(PAYLOAD_MASS__KG_) **FROM** SPACEXTABLE);
- Interpretation
  - This query shows the names of the booster_versions that have carried the maximum payload mass

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

Query results

2015 Launch Records

- Query
  - **%%sql SELECT**
  - substr(Date, 6, 2) **AS** "Month",
  - substr(Date, 1, 4) **AS** "Year",
  - Landing_Outcome,
  - Booster_version,
  - Launch_Site
  - **FROM** SPACEXTABLE
  - **WHERE** substr(Date, 1, 4) = "2015" **AND** Landing_Outcome **LIKE** "%Failure (drone ship)%"
  - **GROUP BY** "Month";

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

Query results

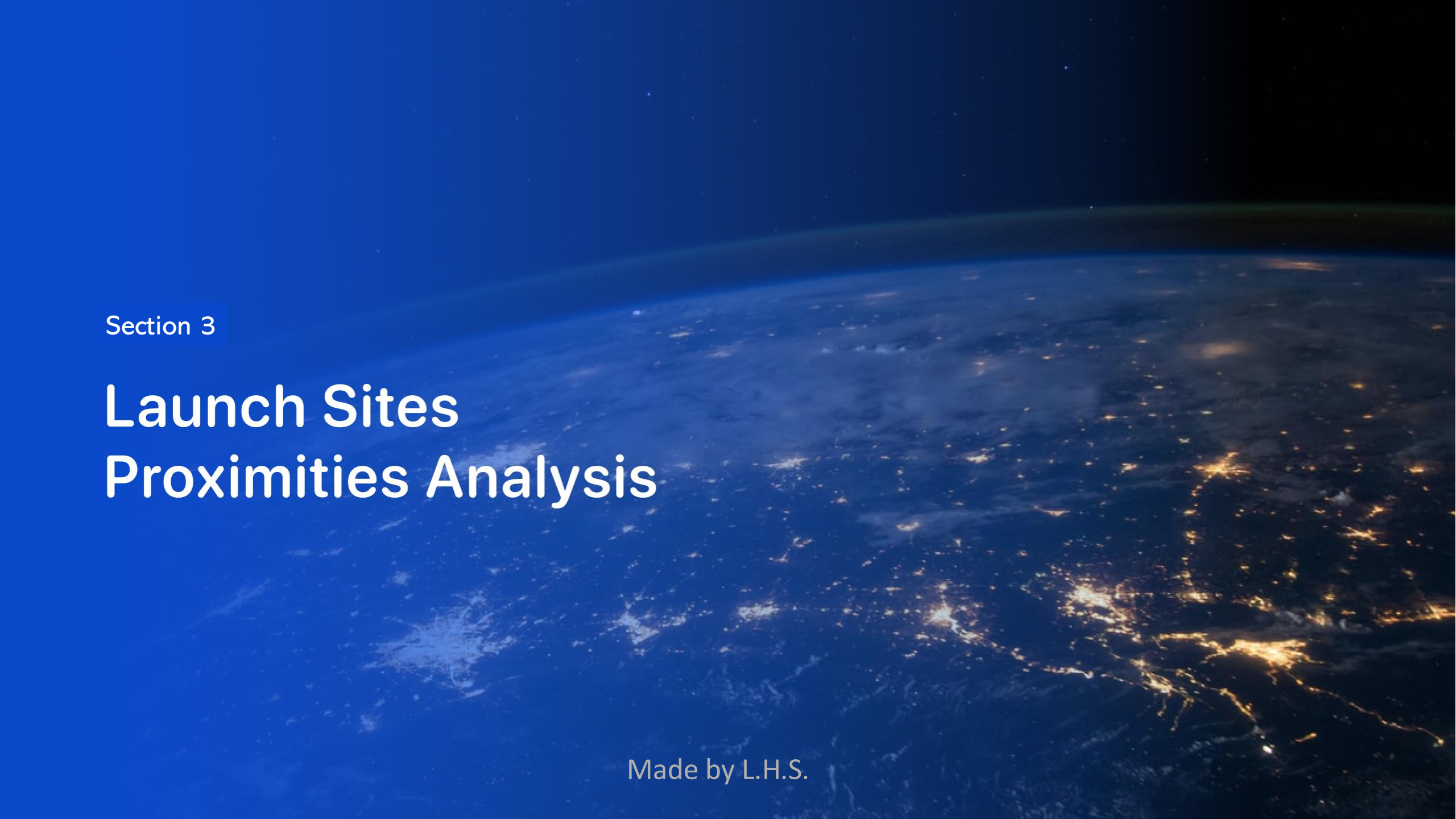
- Interpretation
  - This query shows the list of records for failure landing_outcomes on drone ship, booster versions, launch sites, months and year for the year 2015

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

- Query
  - %%sql
  -
  - SELECT Landing_Outcome, COUNT(*) AS Amount
  - FROM SPACEXTABLE
  - WHERE Date BETWEEN "2010-06-04" AND "2017-03-20"
  - GROUP BY Landing_Outcome
  - ORDER BY COUNT(Landing_Outcome) DESC;
- Interpretation
  - This query shows the ranking of count of landing outcomes between 2010-06-04 and 2017-03-20 in descending order

Landing_Outcome	Amount
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

Query results



Section 3

Launch Sites Proximities Analysis

Made by L.H.S.

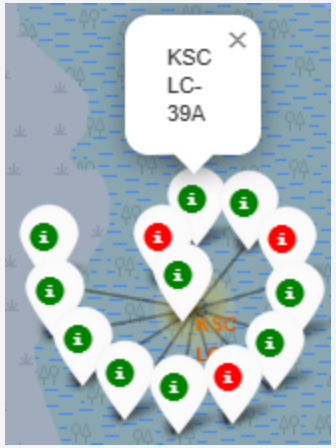
Launch site locations

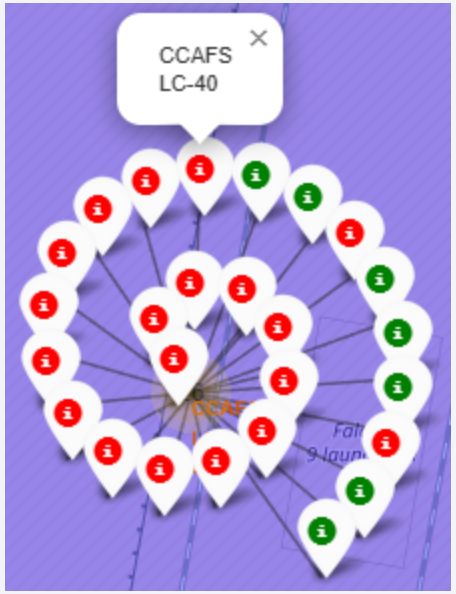
- Here we see the location of launch site for SpaceX in the US. One launch site is on the west coast, while the other three are in close proximity on the east coast.
- We find that all launch sites are in close proximity to the ocean

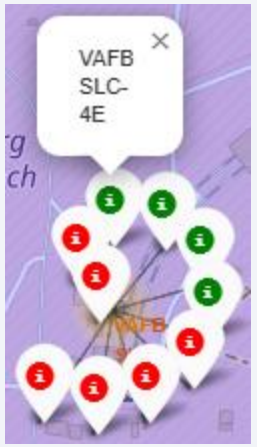
Made by L.H.S.

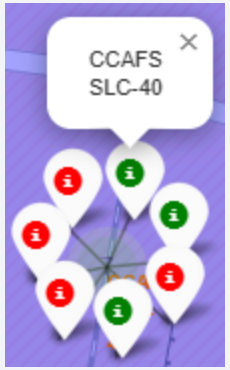
Launch outcomes for launch sites

- Here we see the distribution of successful and failed launch outcomes for the different launch sites
- We see that the launch site with the highest success rate is KSC LC-39A, and the launch site with the lowest success rate is CCAFS LC-40
- We also notice that most launches happen from CCAFS LC-40



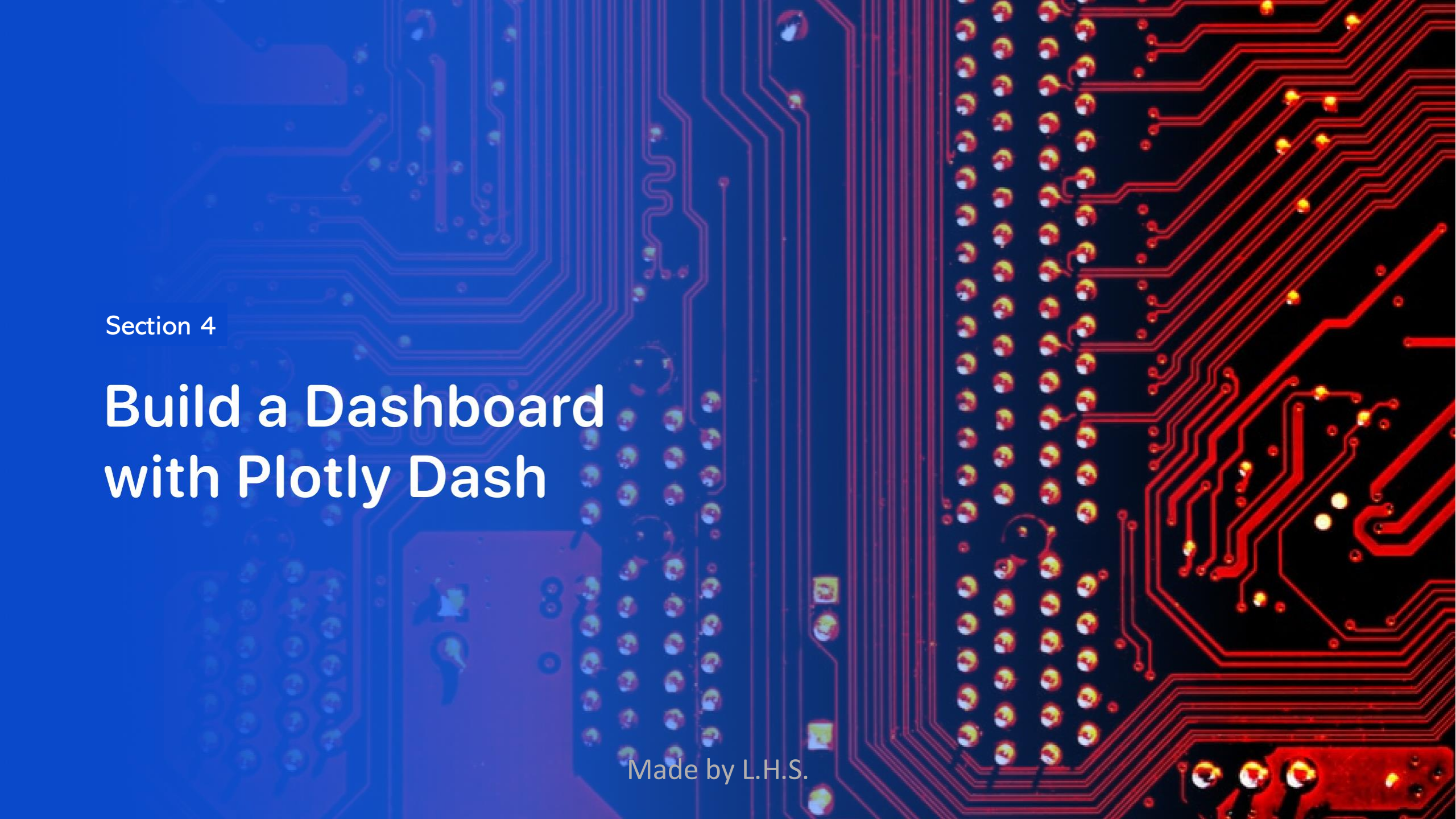






Launch site proximities

- Here we see the proximities of the launch site VAFB SLC-4E
- We see how close the launch site is to the coast and the nearest railline, and how far it is from the nearest city and major highway



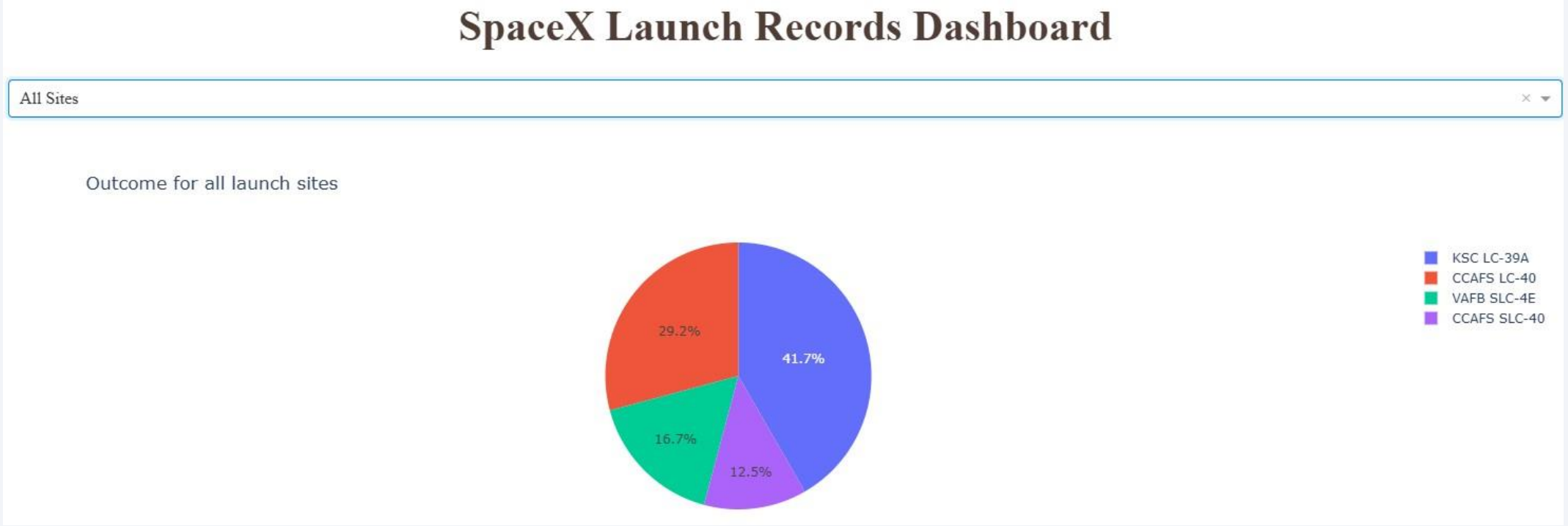
Section 4

Build a Dashboard with Plotly Dash

Made by L.H.S.

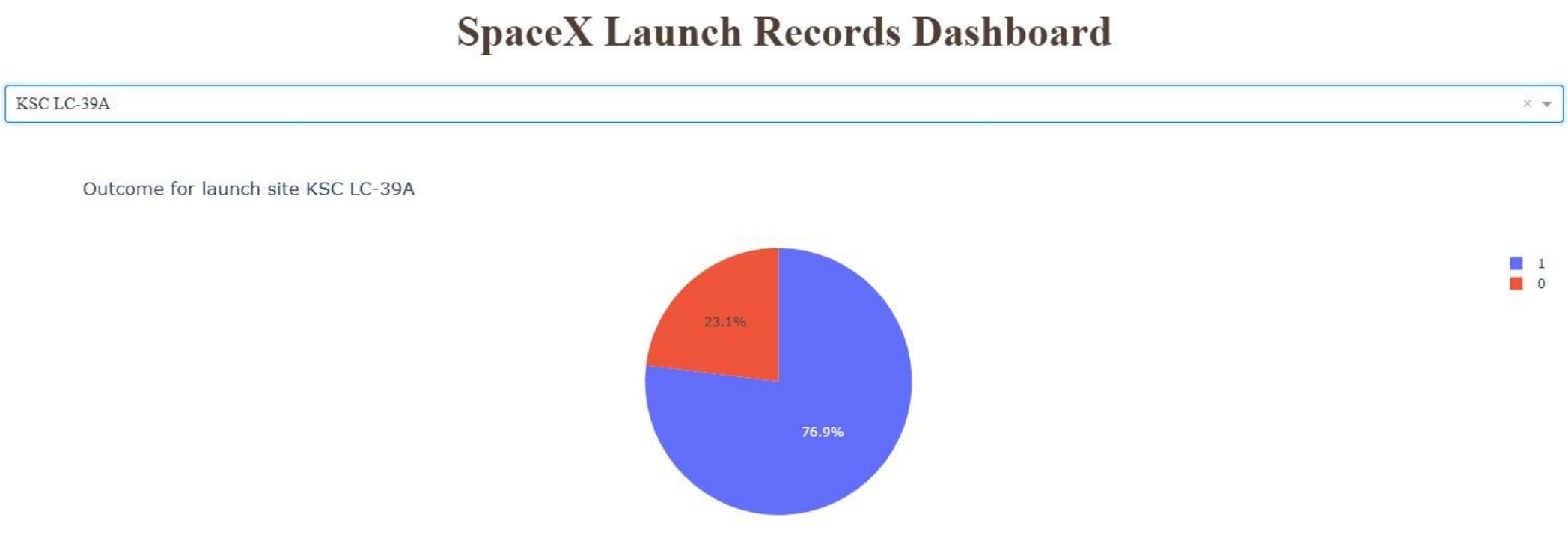
Launch success count for all sites

- This shows the distribution of successful launches between the launch sites.
- We can see how KSC LC-39A has the most successful launches.



Launch site with highest success ratio

- This shows the succeeded and failed launches for the launch site KSC LC-39A, which has the highest success ratio at 76.9%.



Payloads vs launch outcome

- Here we see how the launch outcome changes with payload mass and booster version
- We see that between 6000 and 8000 kg all outcomes fail
- We furthermore see that the FT booster version has the most successful outcomes

Made by L.H.S.

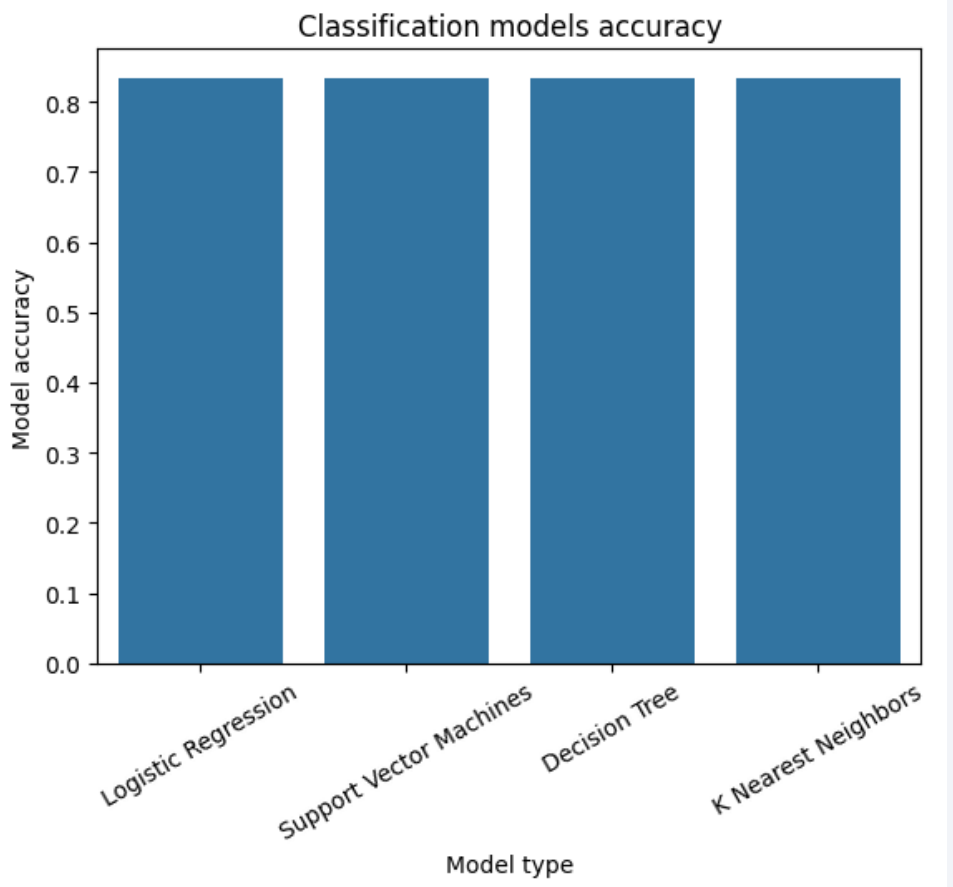
Section 5

Predictive Analysis (Classification)

Made by L.H.S.

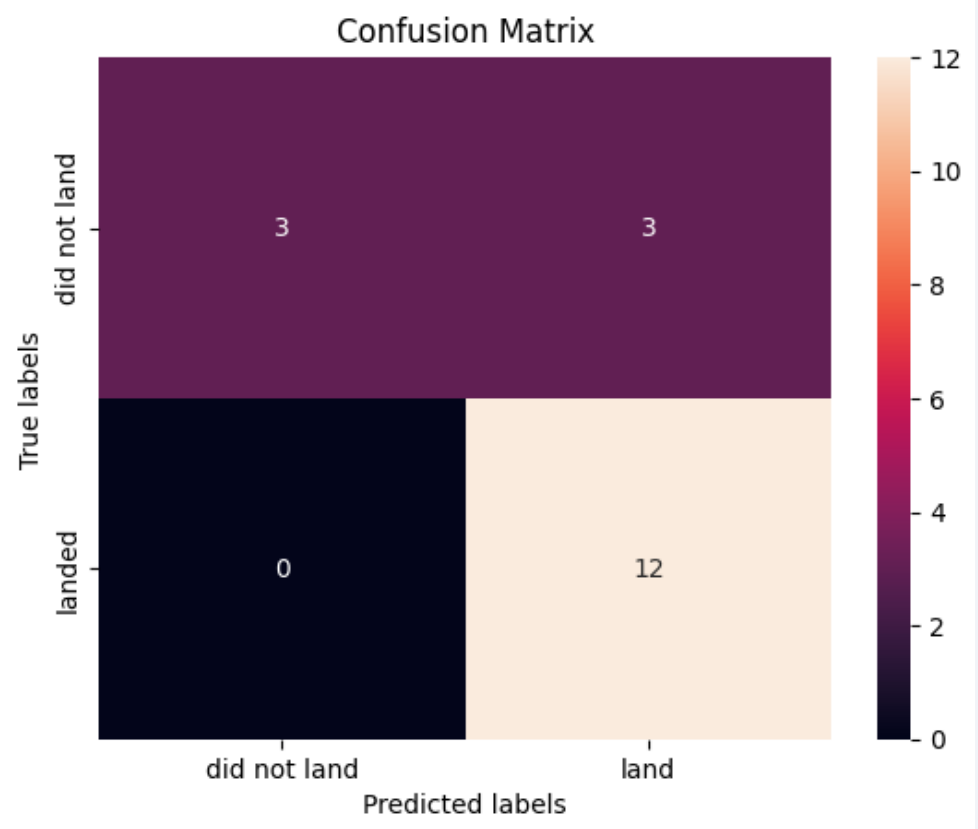
Classification Accuracy

- We can see how all classification models have the same test accuracy,
  which indicates that they all generalize equally well.



Confusion Matrix

- Here we see the confusion matrix of the logistic regression model.
- We see that the model does not have any false negatives, but it has 3 false positives. It therefore has perfect recall at 100%, but lower precision at 80%.



When should SpaceY bid for a launch?

In terms of the mission statement of determining whether the fictional company SpaceY can compete with SpaceX for a given launch, we are interested in under which circumstances the chances of a successful landing is the lowest possible for SpaceX.

Circumstances for maximum chance of landing failure:

- Landing site is CCAFS LC-40
- Payload mass is less than 8000 kg
- The orbit type is GTO
- Booster version is v1.1

Conclusion

Insights:

- Positive landing outcomes are increasing year by year
- The launch site KSC has the highest success rate
- All launch sites are close to the ocean and far away from cities
- The booster FT has the most successful outcomes
- All classification models have the same testing accuracy with high recall, and slightly lower precision

SpaceY should attempt to compete against SpaceX under the following conditions:

- Landing site is CCAFS LC-40
- Payload mass is less than 8000 kg
- The orbit type is GTO
- Booster version is v1.1

Appendix

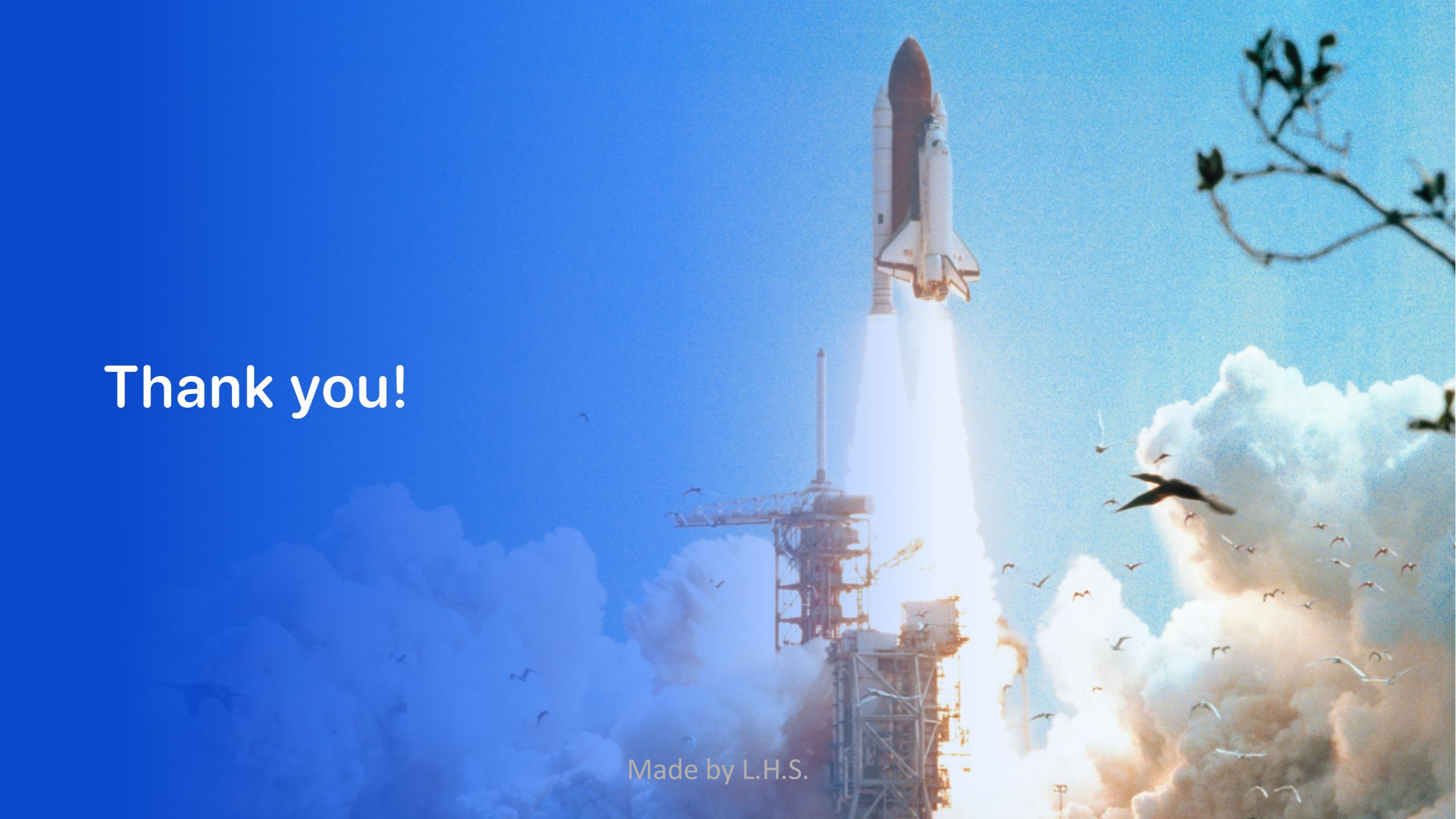
- Classification models accuracy bar plot code

```
X_models = ["Logistic Regression", "Support Vector Machines", "Decision Tree", "K Nearest Neighbors"]
Y_models_accuracy = []


Y_models_accuracy.append(logreg_cv.score(X_test, Y_test))
Y_models_accuracy.append(svm_cv.score(X_test, Y_test))
Y_models_accuracy.append(tree_cv.score(X_test, Y_test))
Y_models_accuracy.append(knn_cv.score(X_test, Y_test))


sns.barplot(x = X_models, y=Y_models_accuracy)
plt.title("Classification models accuracy")
plt.xlabel("Model type")
plt.ylabel("Model accuracy")
plt.xticks(rotation=30)
```

Thank you!



Made by L.H.S.