IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Charmi Patel

Outline

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

Executive Summary

- The methodology of this project includes *data collection, data wrangling, exploratory data analysis, data querying, statistical data analysis, data visualisation, and refining predictive models* to uncover meaningful patterns for data-driven business insights.
- The results of this project led to insights to inform strategic price optimisation for rocket development and launch execution for a new space company to compete in the market it occupies.

Introduction

- Billionaire industrialist, Allon Musk is set to compete with Space X founder, Elon Musk to establish his new venture, Space Y.
- Space X has dominance ahead of competitors in the commercial rocket space due to the critical key advantage of rocket reusability. This results in significant cost savings as the first stage of the rocket is traditionally the most expensive, largest, and holds the majority of the critical payload.
- Business Problems we are looking to solve:
    - Predict if Space X will reuse the first stage of the rocket launch?
    - Which rocket launches were the most successful and why?
    - Which rocket launches were failures and why?
    - What are the key takeaways, insights, and findings that can aid Space Y in making informed decisions to become competitive in the rocket industry?

Section 1

Methodology

Methodology

Data collection:

- SpaceX Application Programming Interface (API) calls
- Web scraping Wikipedia page on SpaceX rocket launches

Data wrangling:

- Jupyter Notebooks, Python (Pandas, NumPy)

Exploratory data analysis (EDA) using visualisations and querying:

- Python SQL connection, MySQLite, MatPlotLib, Seaborn

Interactive visual analytics, dashboards, and maps:

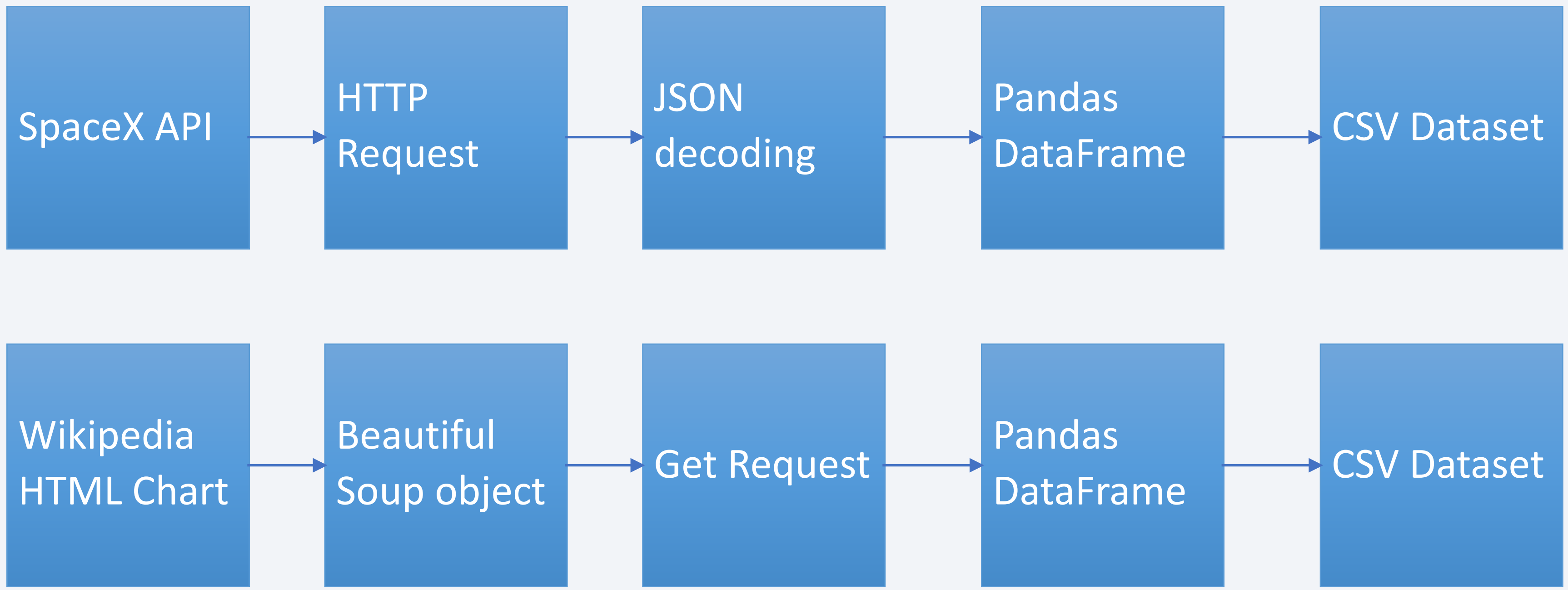
- Plotly Dash and Folium

Predictive analysis using classification models:

- Sci-kit-learn, Seaborn, Matplotlib

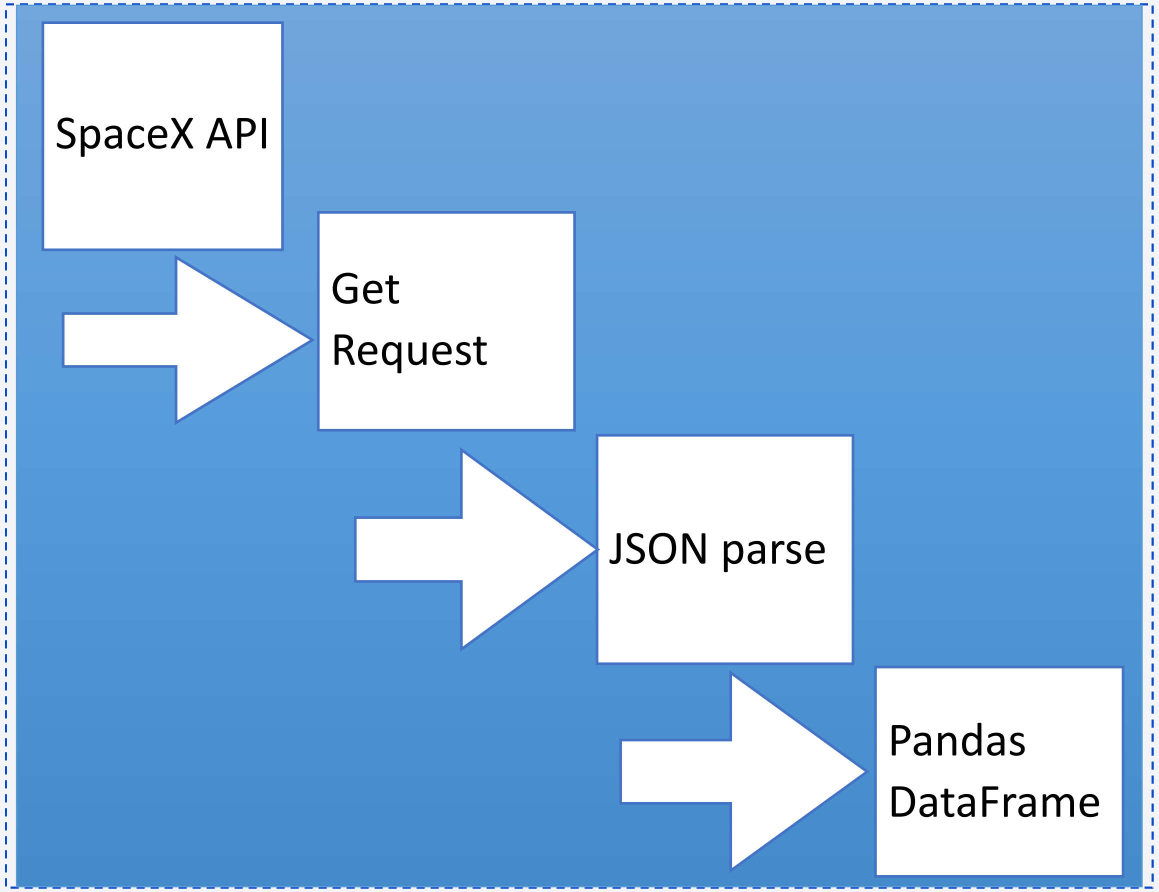
Data Collection

- Datasets were collected by web scraping and API



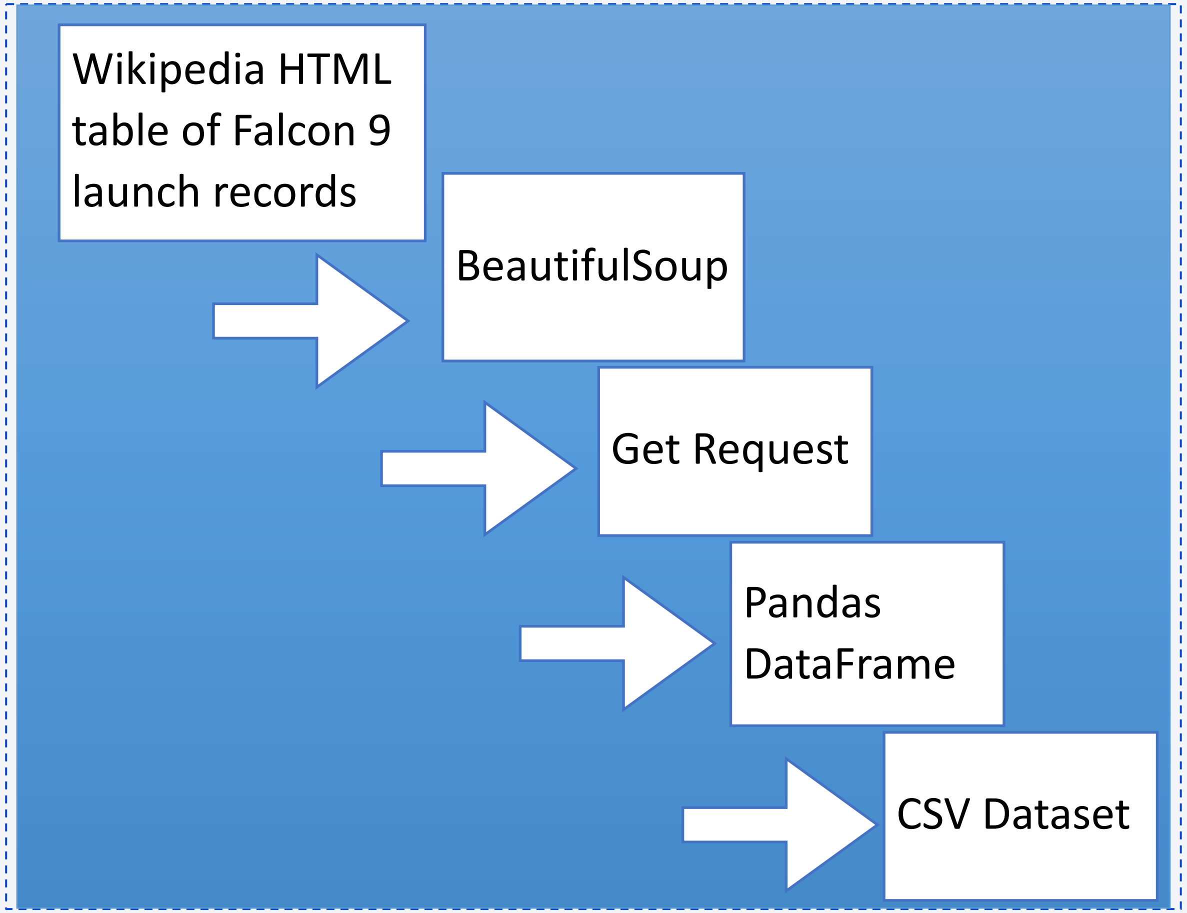
Data Collection – SpaceX API

- Here is the data collection with SpaceX REST calls
- The GitHub of the SpaceX API calls notebook can be found here



Data Collection - Web Scraping

- Here is the web scraping process using the BeautifulSoup object
- GitHub URL of web scraping notebook can be found here



Data Wrangling

- Data was wrangled using Pandas and NumPy
  - Missing values, counts of launches per site, counts of orbit types
  - Landing outcome labels were created to differentiate successful and failure mission outcomes
  - Missing values were replaced with average of column data
- GitHub of data wrangling related notebooks can be found here

Exploratory Data Analysis (EDA) with Data Visualisation

- The following charts were plotted for viewing specific insights
  - Scatter plots to visualise the following relationships (Flight Number vs. Payload Mass, Flight Number vs. Launch Site)
  - Bar chart to visualise the success rate of each orbit type (discrete)
  - Line plot to visualise the launch success yearly trend (time series)
- GitHub of EDA with data visualisations notebook can be found here

Exploratory Data Analysis (EDA) with SQL

- Performed SQL queries:
  - Payload Mass: Total, Average, etc
  - First Successful Landing Outcome
  - Booster Names and specific filters
  - Total number of Successful and Failure Mission Outcomes
  - Rank of the count of landing outcomes between 2010-06-04 and 2017-03-20
- GitHub URL of EDA with SQL notebook can be found here

Interactive Maps with Folium

- Map objects such as markers, circles, lines, pop up labels, text labels were created and added to a folium map
- Coloured markers of successful (Green) and failed (Red) launches were added to visualise clusters of launches at the designated sites
- These objects were added to better identify and visualise the locations of rocket launches, their success rates, and relative proximity to other sites
- GitHub URL of interactive maps with Folium map can be found here

Dashboard with Plotly Dash

- The dashboard displays the following:
  - Interactive pie chart displaying success rates per all sites and each individual site based on user selection using a dropdown
  - Scatter plot displaying outcomes of booster versions compared to payload mass ranges selected by user using a slider
- GitHub URL of completed Plotly Dash notebook can be found here

Predictive Analysis (Classification)

- Summary of how models were built, evaluated, improved, found to be the best performing classification model
    - Created a NumPy array from “Class” column in data
    - Standardised the data using StandardScaler
    - Split the data into training and testing sets with train_test_split function
    - Found best hyper parameters for each model using GridSearchCV object with cv = 10
    - Applied GridSearchCV on all models (Logistic Regression, Support Vector Machine, K-Nearest Neighbours, and Decision Tree Classification)
    - Calculated accuracy of test set data for all models
    - Confusion matrix for each model was investigated
    - Determined best performing model using Jaccard score and F1 score
- GitHub URL of predictive analysis notebook can be found here

Results

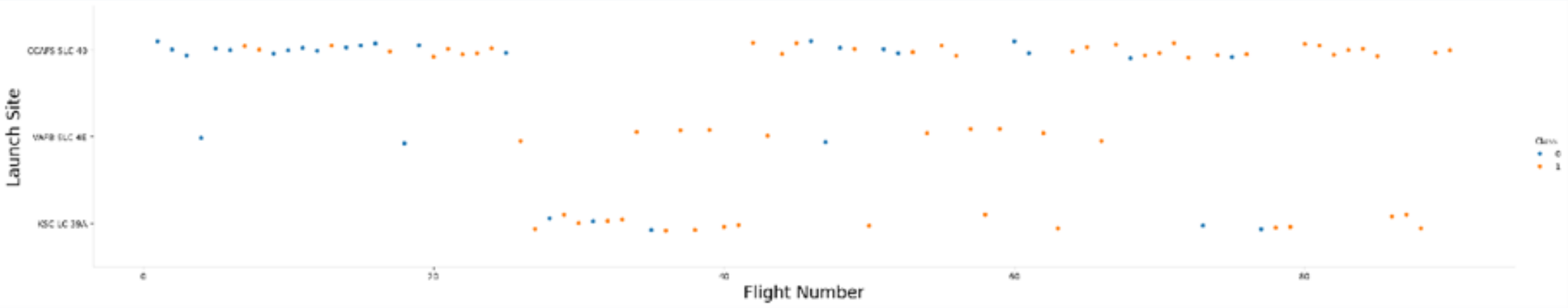
- Exploratory data analysis results
    - Interactive analytics demo in screenshots
- Predictive analysis results are shared in the following slides

Section 2

Insights drawn from EDA

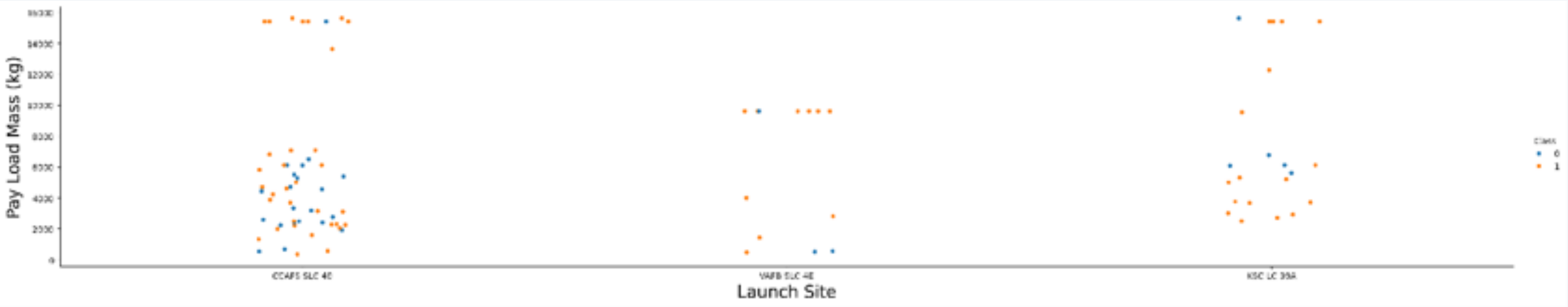
Flight Number vs. Launch Site

- Here is a scatter plot of Flight Number vs. Launch Site mapped by outcome (class)
- You can see the data points distribution of flight number corresponding to their launch sites where the earliest flights all failed while the later flights succeeded



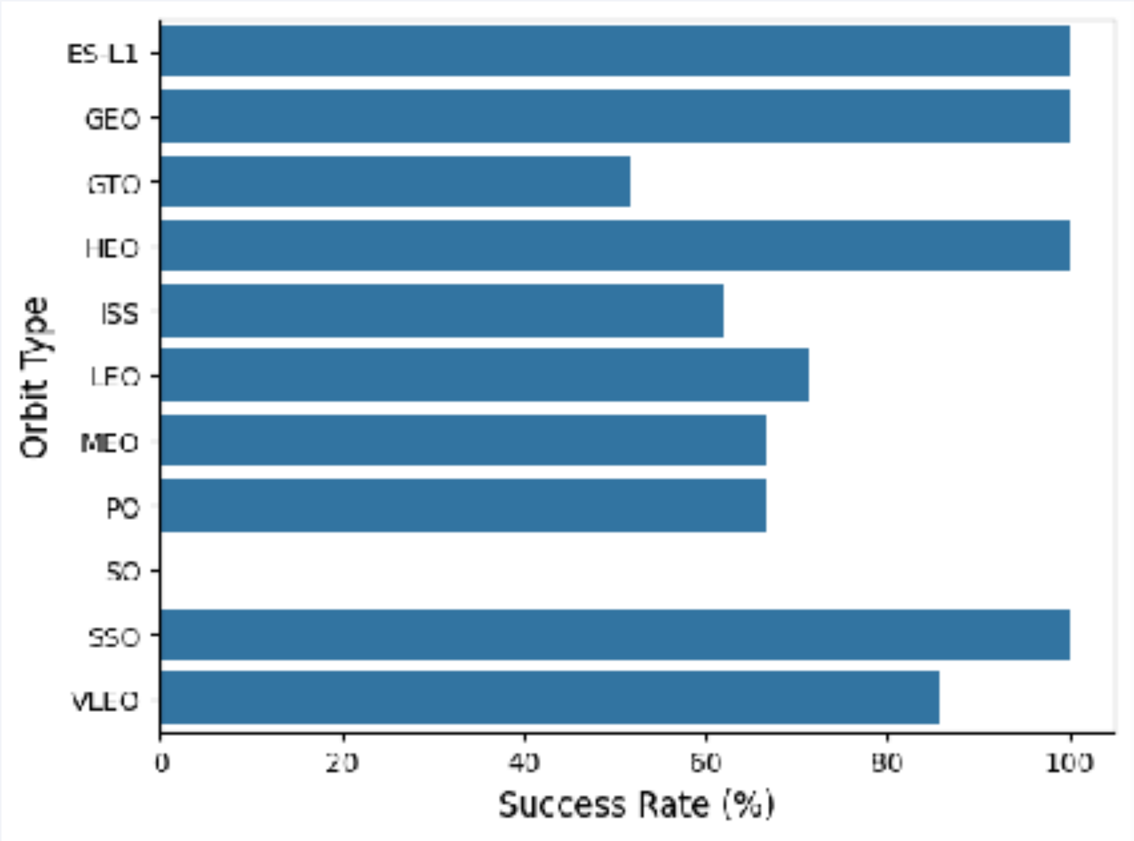
Payload vs. Launch Site

- Here is a scatter plot of Payload vs. Launch Site with outcomes identified
- You can see the majority of missions launching from the site CCAFS SLC-40 that have payload mass under 8000 kg have a mix of successful and unsuccessful outcomes



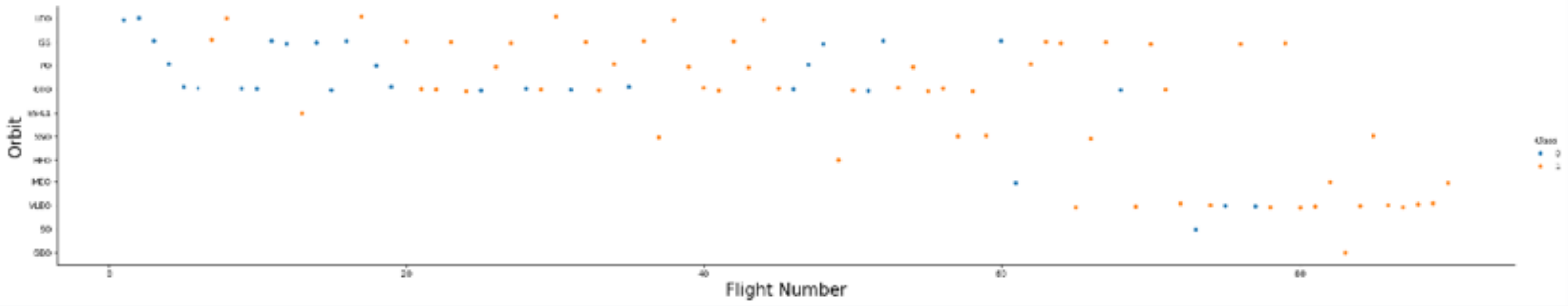
Success Rate vs. Orbit Type

- Here is a bar chart for the success rate of each orbit type
- You can see the Success Rate of the 'SO' orbit type was the lowest and and Success Rates of the 'ES-L1', 'GEO', 'HEO', 'SSO' orbit types were the highest



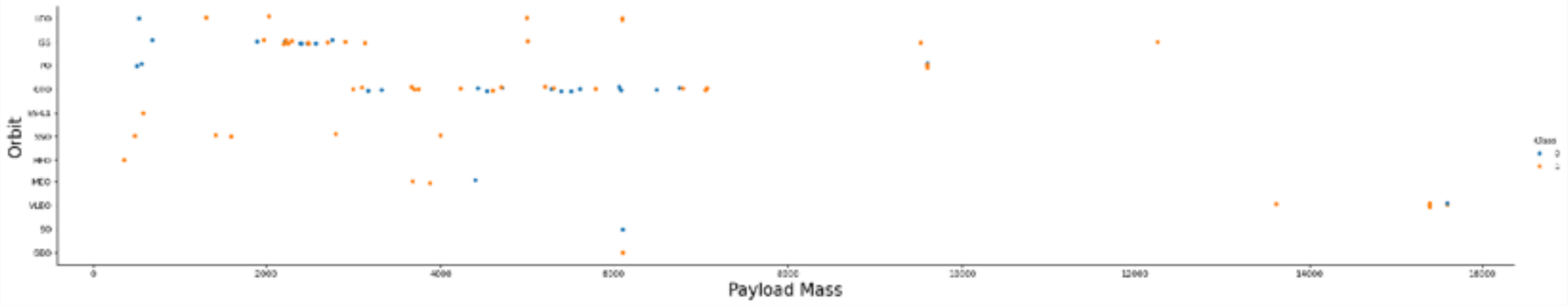
Flight Number vs. Orbit Type

- Here is a scatter plot of Flight Number vs. Orbit type
- You can clearly see the earlier flight numbers are concentrated in one area of the plot for specific orbit types and the later flights are concentrated in a different area of the plot



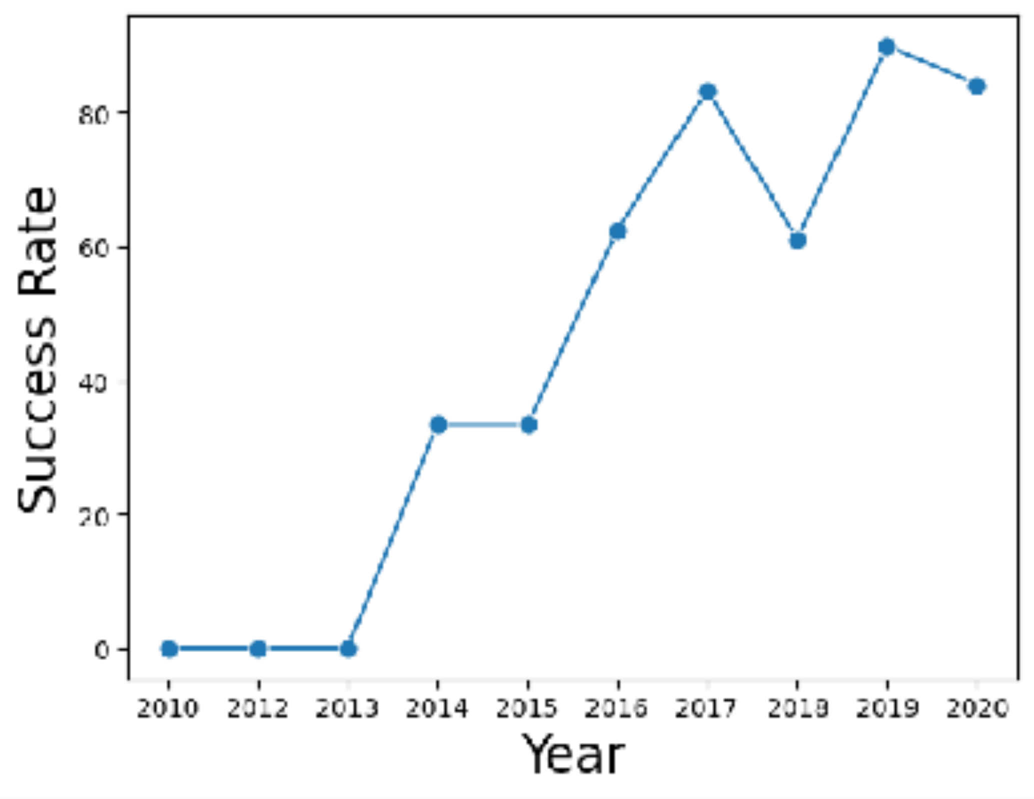
Payload vs. Orbit Type

- Here is a scatter plot of Payload Mass vs. Orbit Type
- The data points are coloured blue for unsuccessful mission outcome and orange for successful mission outcome
- The majority of the launches had a payload mass under 8000 kg



Launch Success Yearly Trend

- Here is a line plot of the yearly average success rate from 2010-2020
- The overall line trends at a bottom of 0% percent in 2010-2013 to under 40% in 2015 and 60% in 2017 before reaching the peaks of above 80% in 2019-2020.



All Launch Site Names

- Here is the SQL query result displaying the names of each unique launch site from the space missions

```
[14]: %sql Select DISTINCT Launch_Site FROM SPACEXTBL LIMIT 25

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

[14]:

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

Launch Site Names Begin with 'CCA'

- Here is the SQL query result of 5 records where launch sites begin with \`CCA\`

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO
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)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)

Total Payload Mass

- Here is the result of the SQL query for total payload carried by boosters from NASA which is 45596 kg

Total_Payload_Mass
45596

Average Payload Mass by F9 v1.1

- Here SQL query result for the average payload mass carried by booster version F9 v1.1

Average_Payload_Mass
2534.66666666667

First Successful Ground Landing Date

- Here is the SQL query result of the date of the first successful landing outcome on ground pads

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Here is the SQL query result of the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

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

Total Number of Successful and Failure Mission Outcomes

- Here is the SQL query result of the total number of successful and failed mission outcomes

Successful_Outcomes	Failure_Outcomes
100	1

Boosters Carried Maximum Payload

- Here is the SQL query result with the names of the Booster Versions from the database that 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

2015 Launch Records

- Here is the SQL query result for the failed landing_outcomes in drone ship, their booster versions, and launch site names in year 2015

Month	Booster_Version	Launch_Site	Landing_Outcome
1	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
4	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- Here is the SQL query result of the rank of the count of landing outcomes between 2010-06-04 and 2017-03-20
- The highest count of landing outcome was no attempt to land

Landing_Outcome	Outcome_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

Section 3

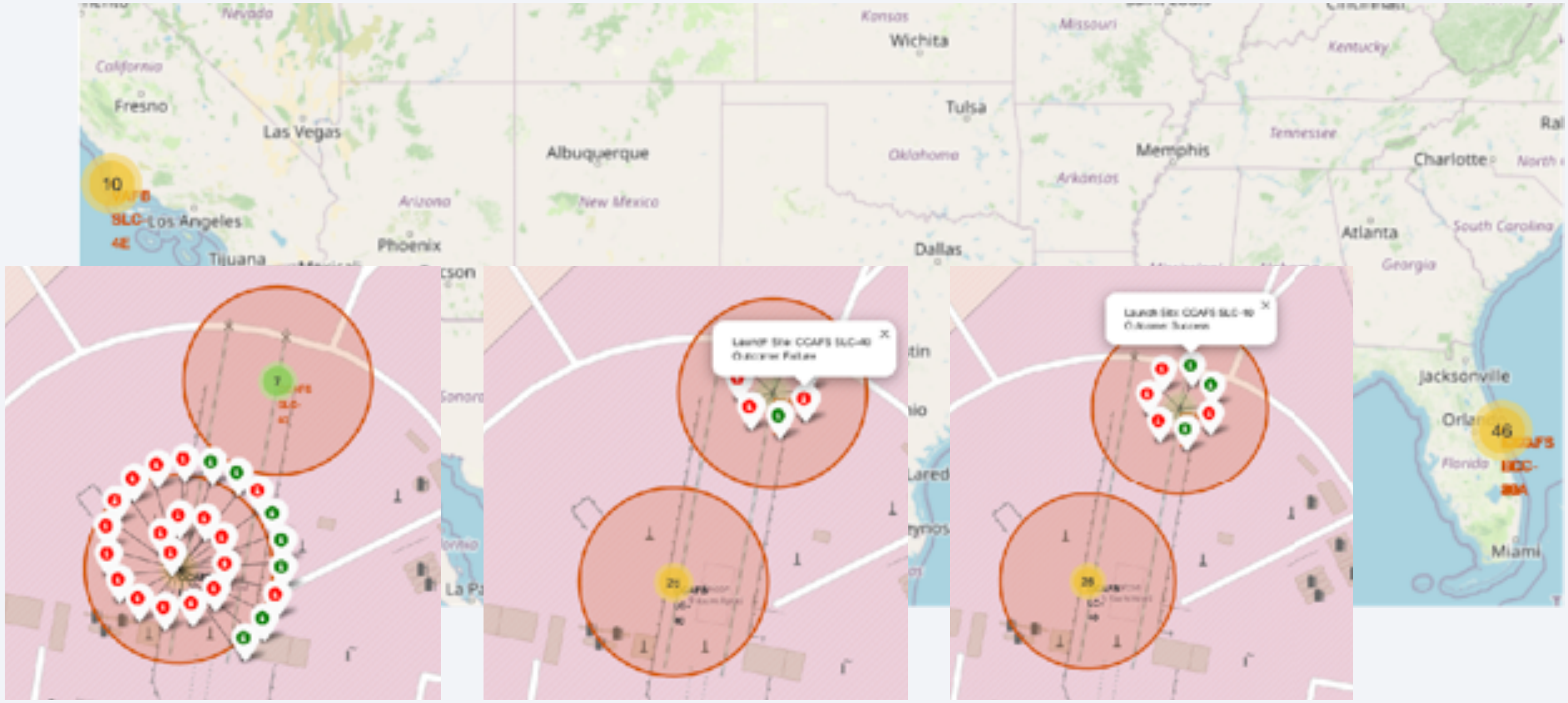
Launch Sites Proximities Analysis

Launch Sites on map

- Generated folium map includes all launch sites locations on a global map showing their proximity to the equator and coastline
- Close proximity to the coastline minimises the risk of debris being dropped in populated areas potentially endangering civilians

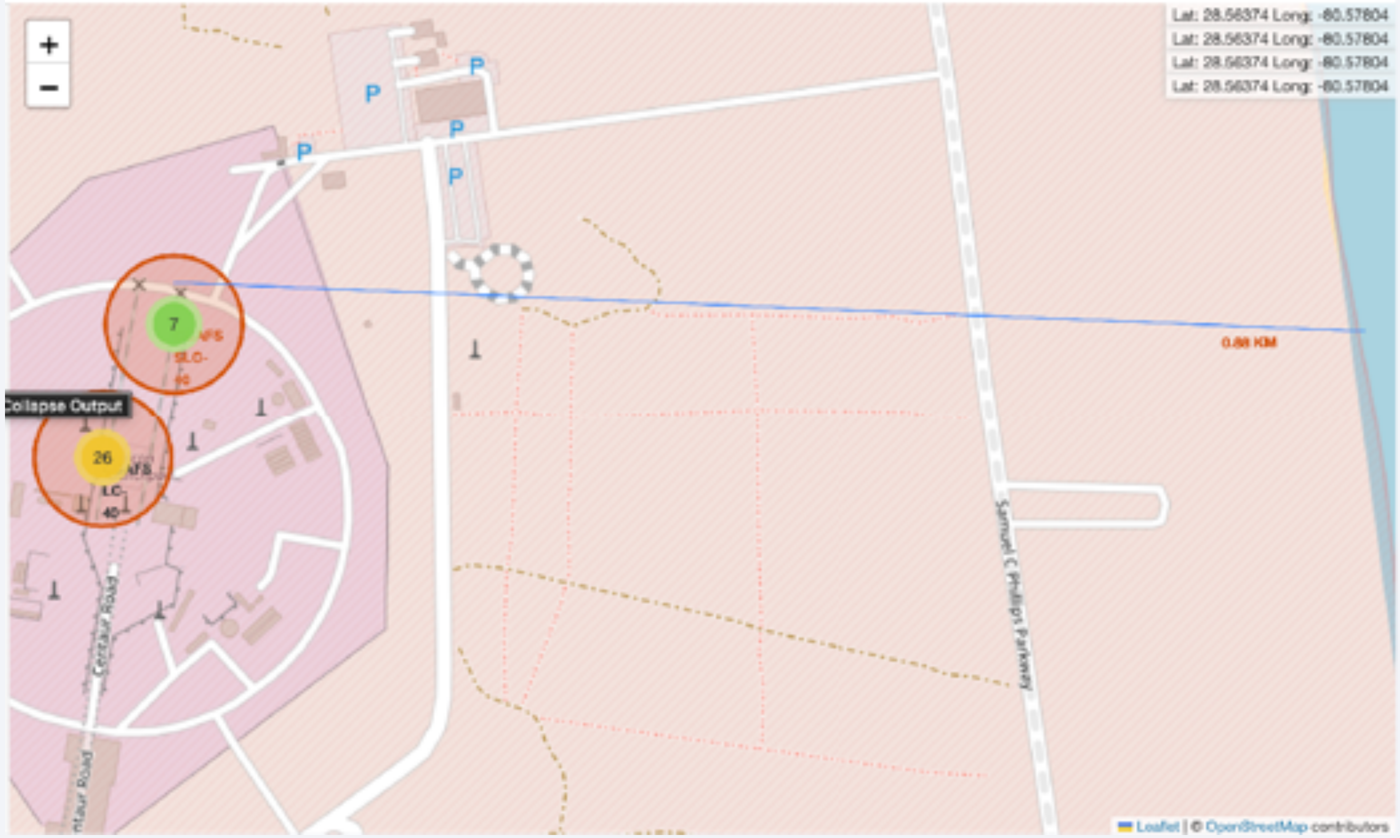
Launch Outcomes of Launch Sites on map

- You can see the launch sites based in Florida had more successful launches based on the colour-labelled launch outcomes on the map
- Green marker signifies a successful launch and a red marker signifies a red marker



Launch Sites proximity to landmarks on map

- Selected launch site and its proximity to the nearest coastline
- Distance calculated is 88KM and shown on the map

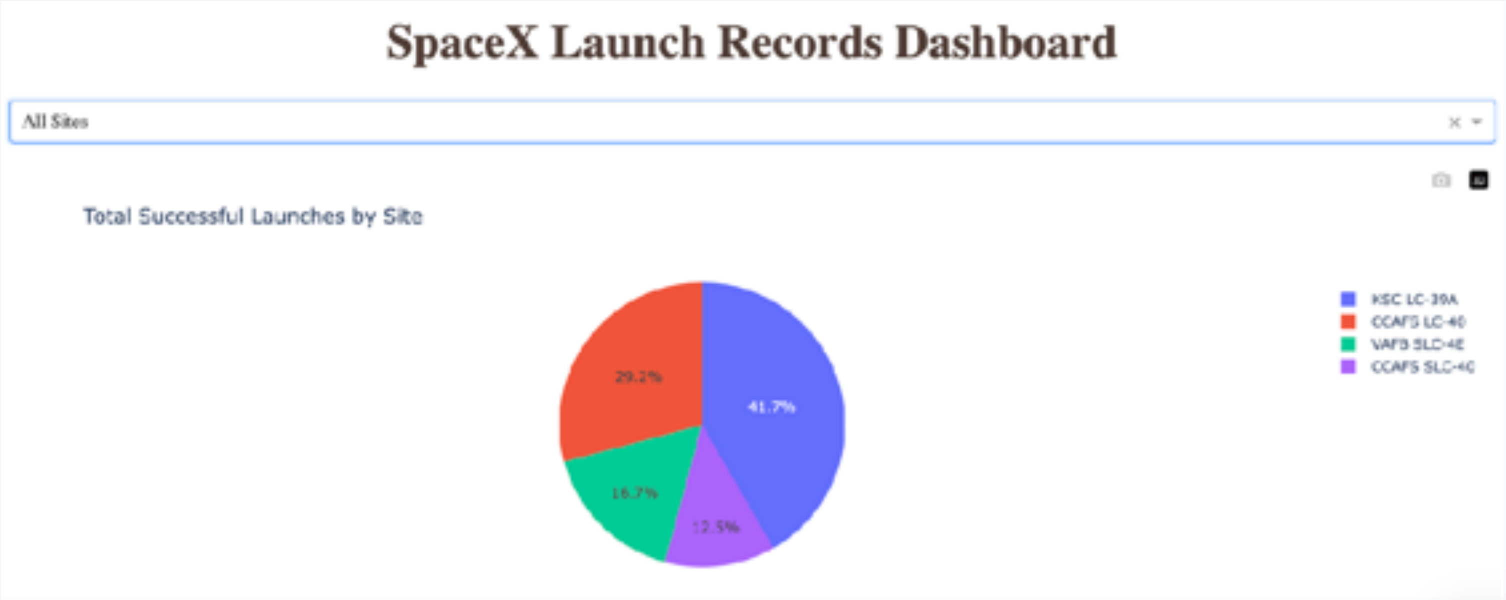


Section 4

Build a Dashboard with Plotly Dash

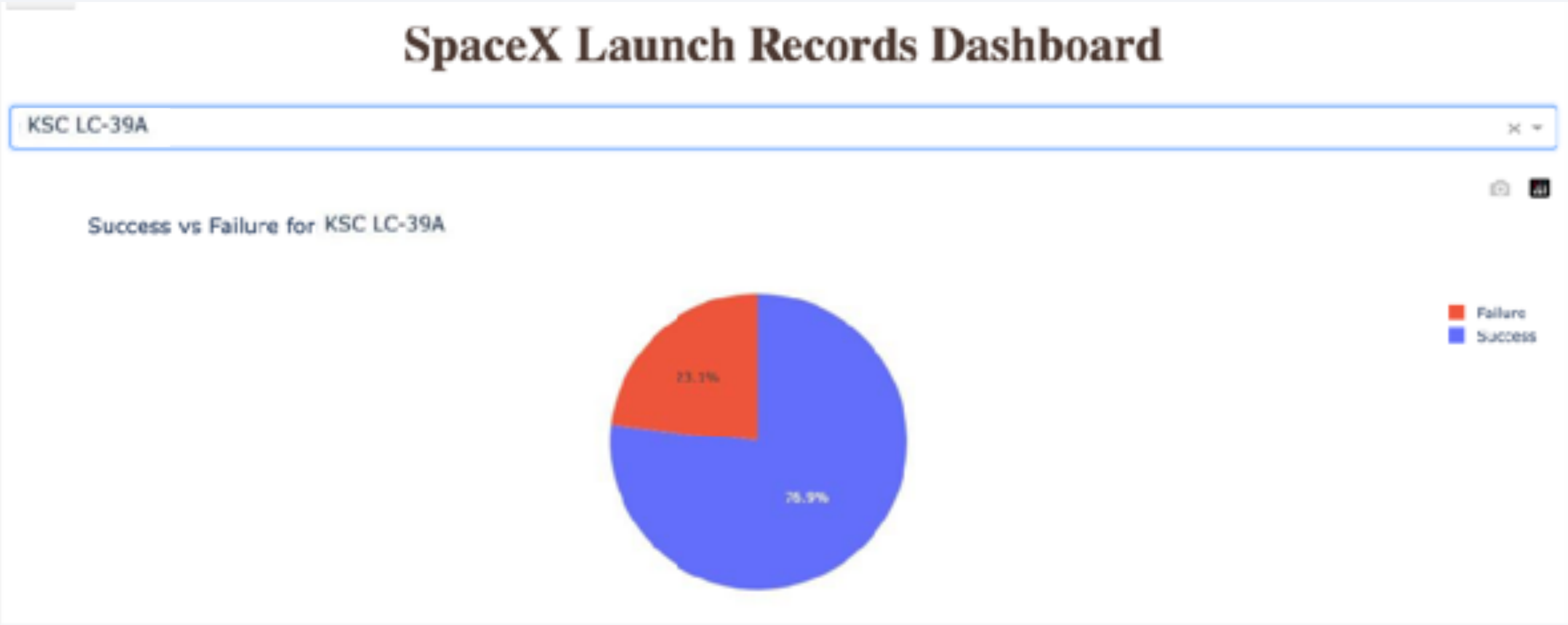
Total Successful Launches by Site

- Here is a screenshot of launch success count for all sites in a piechart
- The CCAFS LC-40 and KSC LC-39A sites had the high percentages of successful launches compared to other sites



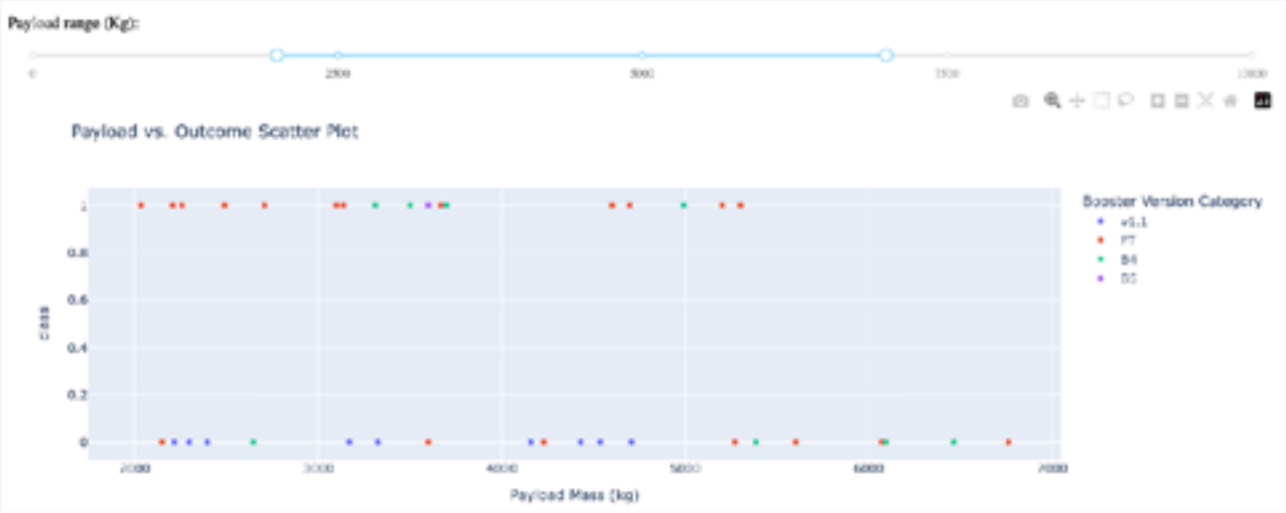
Launch Site with highest launch success ratio

- Here is the pie chart for the launch site that had the highest launch success ratios
- The KSC LC-39A site had a 76.9% success rate with 10 successful landings and 3 failed landings



Payload Mass vs Launch Outcome for all sites

- Scatter plot for Booster Version mapped to Payload Mass vs. Outcome
- A range slider is displayed to let the user select a specified range for the x-axis of the scatter plot
- Payload Mass lower than 4,000 kg had the most instances of successful outcomes
- The B4 and FT booster versions have the most instances of successful outcomes

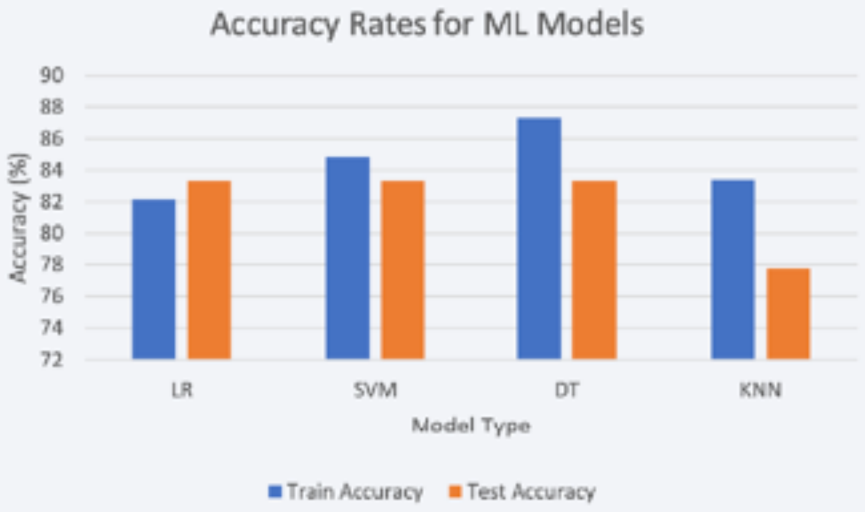


Section 5

Predictive Analysis (Classification)

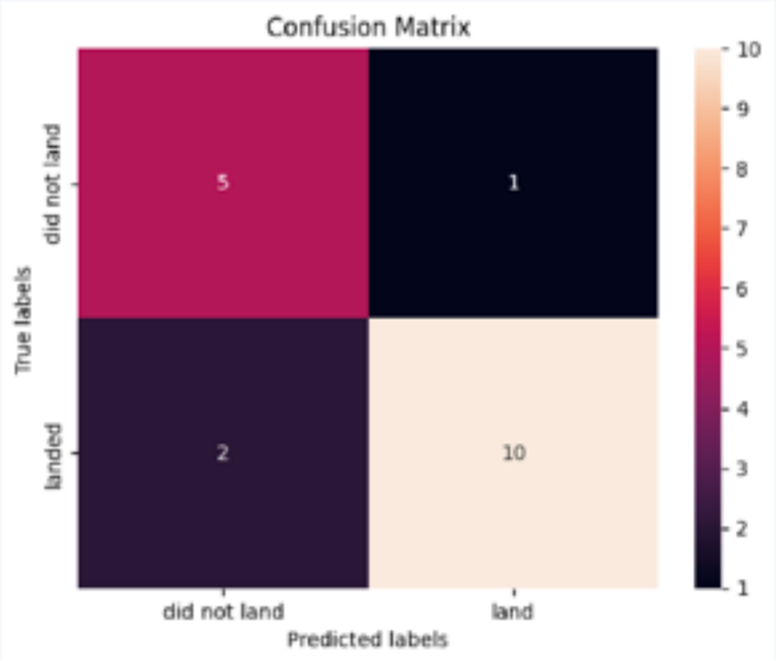
Classification Accuracy

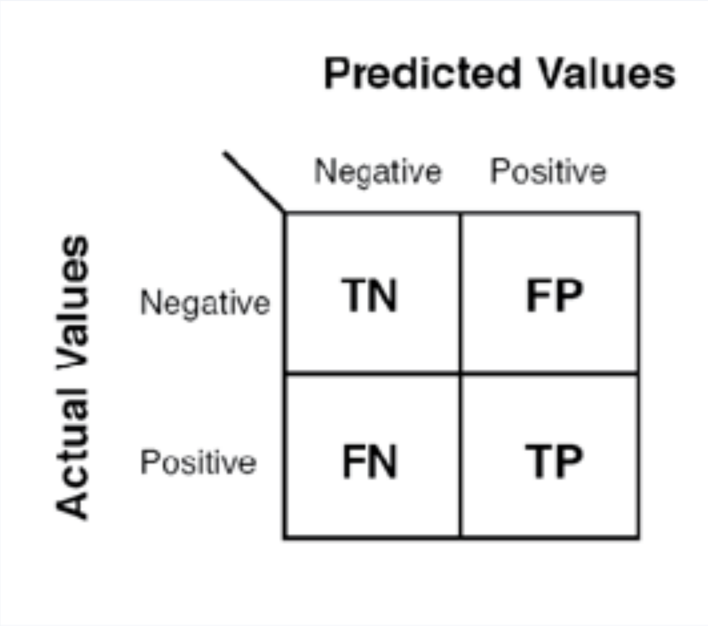
- Logistic Regression, Support Vector Machine, and K-Nearest Neighbours performed very similarly on the test dataset with the highest test accuracy of 0.833
- Decision Tree performed the worst on the test set and showed signs of overfitting as it had the highest accuracy on the training set but the lowest accuracy on the test set



Confusion Matrix

- The following confusion matrix shows that the best performing model was the decision tree model as it classified launch failure at a higher rate Thant he other models. The performance in prediction of launch success rate is similar to the other models.
- This is the primary metric determining the selection of the decision tree model as the best performing model.





Conclusions

- **Point 1** - Based on the available data, the Decision Tree Classification model is the best algorithm for this dataset however the other models were also highly effective
- **Point 2** - Rocket launches with low payload mass showed better results than launches with larger payload mass based on the insights drawn from the analytics and visualisations
- **Point 3** - All of the launch sites are in very close proximity to the coast lines and equator
- **Point 4** - The success rate of launches increases over the years
- **Point 5** - KSC LC-39A had the highest success rate for launches of all of the sites
- **Point 4** - Orbits ES-L1, GEO, HEO, and SSO have 100% success rate

Appendix

- Link to GitHub repository: https://github.com/charm5/Space-X-Rocket-Launch-Insights
- This presentation is also uploaded to the GitHub repository

Thank you!