IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Lawson Pennel
June 6, 2025

Outline

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

Executive Summary

- The Question: Will Stage 1 of a Falcon 9 Rocket Land Successfully.
- Importance: The successful landing of Stage 1 will allow it to be reused, saving millions of dollars.

- The data for this report was gathered from SpaceX itself, and web scrapping Wikipedia pages about Falcon 9 launches.

- Multiple different Machine Learning models were used to attempt to predict a successful Falcon 9 landing, with the best model being Decision Tree with an 89% accuracy rating.

- Using the data in this report, and the models created, it is possible to determine the best conditions to launch a Falcon 9 rocket it, so that its Stage 1 may land successfully.

Introduction

- The SpaceX Falcon 9 Rocket costs only 62 million dollars to launch compared other rockets which cost upwards of 165 million dollars per launch.
- This feature is because the Falcon 9's Stage 1 can land back on the ground, and so it does not need to be rebuilt with each launch, saving approximately 103 million dollars.
- The Question: Determining if a Stage 1 will land successfully given the conditions of the launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The Data was collected from the SpaceX API and with the use of web scrapping from Wikipedia pages about SpaceX launches.
- Perform data wrangling
  - The data was cleaned, so only Falcon 9 launches were factored into analysis
  - Missing data was either removed or replaced with average values.
  - One Hot encoding was preformed to prepared data for predictive analysis.

Methodology

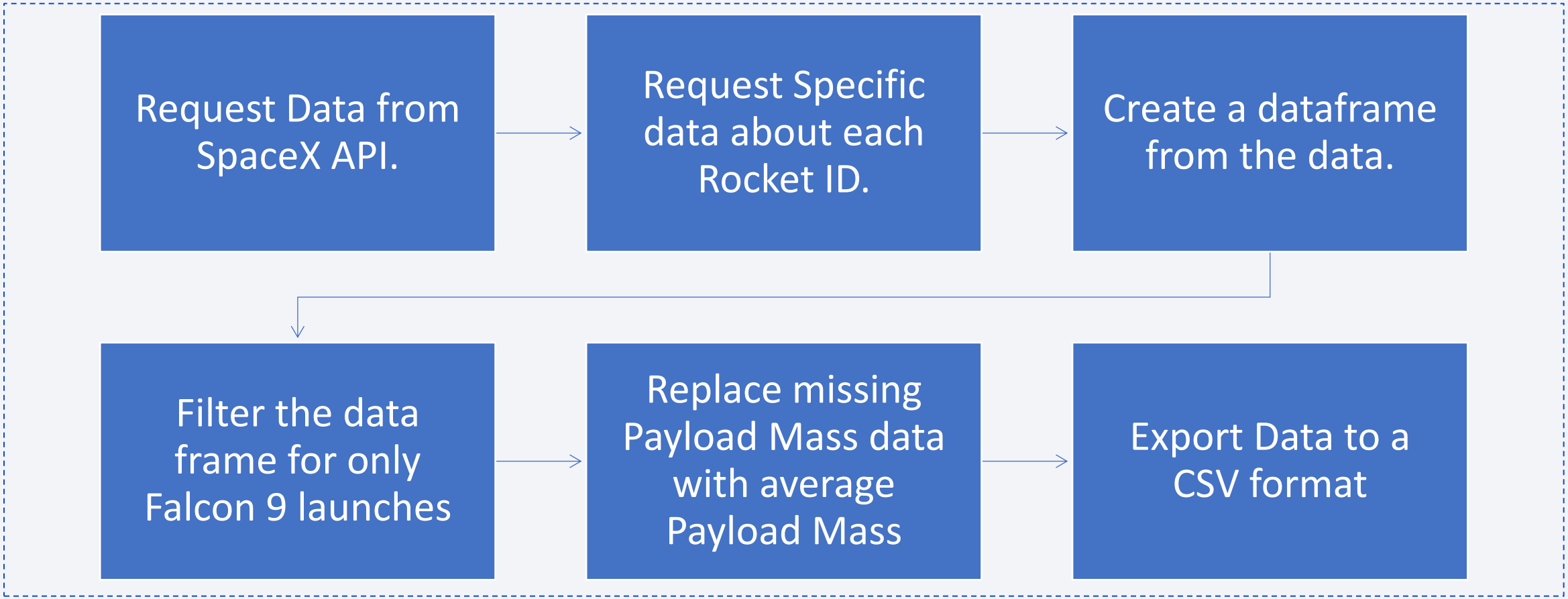
Executive Summary Cont.

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

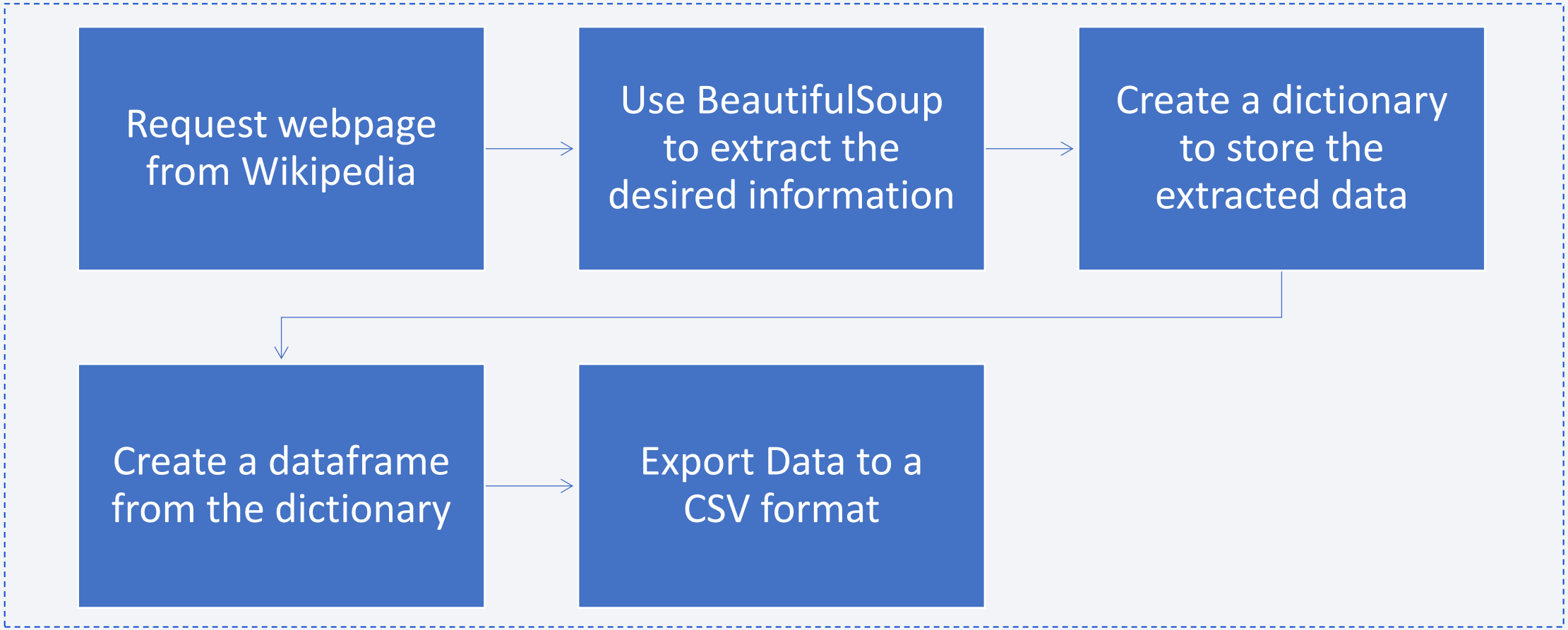
- Data wad collected from the SpaceX API and from Wikipedia pages about SpaceX launches with the use of Web Scrapping
- Data collected from SpaceX API:
  - Booster Version, Payload Mass, Orbit, Launch Site, Outcome, Flights, GridFins, Reused, Legs, Landing Pad, Block, Reused Count, Serial, Longitude, Latitude
- Data collected from Web Scrapping:
  - Flight Number, Launch Site, Payload, Payload Mass, Orbit, Customer, Launch Outcome, Version Booster, Booster Landing, Date, Time

Data Collection – SpaceX API



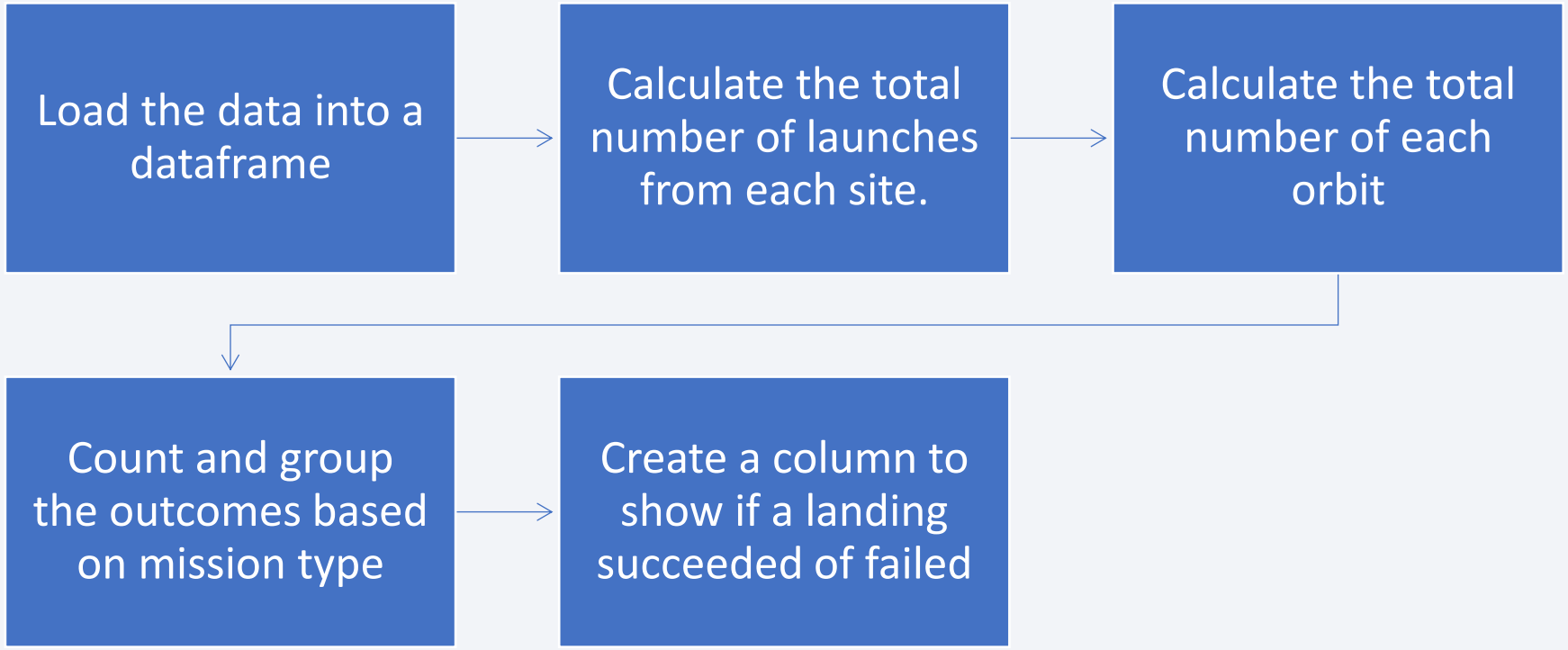
- https://github.com/Lawson-Pennel1127/IBM_CapstoneProject/blob/7b241e7d54fe6aa01aff2be0934ac4ff31ba8556/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping



https://github.com/Lawson-Pennel1127/IBM_CapstoneProject/blob/7b241e7d54fe6aa01aff2be0934ac4ff31ba8556/jupyter-labs-webscraping.ipynb

Data Wrangling



- Results showed that:
  - CCAFS SLC 40 was the most common launch site
  - GTO was the most common orbit
  - Mission type ASDS had the most successful outcomes

- https://github.com/Lawson-Pennel1127/IBM_CapstoneProject/blob/7b241e7d54fe6aa01aff2be0934ac4ff31ba8556/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- A scatter plot was used to compare Launch Sites to the number of Flights. The scatter plot was the optimal choice because it is a clear way to visualize the relationship between two variables. The plot showed that with more flights the success rate tends to improve
- A scatter plot was again used, for the same reasons as above, to compare Launch Sites and Payload Mass. The plot showed that higher Payload Mass tended to have higher success rates, and that the launch site VAFB did not have any payload above 10000.
- A bar chart was used to compare the different orbits average success rates. A bar chart was used because it clearly shows different averages, and the error range of the averages. The plot showed that the orbits ES-L1, GEO, HEO, and SSO have 100% success rates with many other orbits having 50%, with a wide error range.

EDA with Data Visualization Cont.

- A scatter plot was used, for the same reasons as before, the compare Orbit and the number of flights. The plot showed that most orbits success rates improve with more flights, with GTO being an exception, not showing any clear trend.
- A scatter plot was used, for the same reasons as before, to compare Orbit and Payload Mass. The plot showed that success rates improve with higher payload mass, with GTO being an exception, and not showing any clear trend.
- A line graph was used to show the correlation of the date with the success rate. A line graph was used because it clearly illustrates trends over time. It showed that the success rate generally rises with time, with a dip in 2018.
- https://github.com/Lawson-Pennel1127/IBM_CapstoneProject/blob/7b241e7d54fe6aa01aff2be0934ac4ff31ba8556/edadataviz.ipynb
-

EDA with SQL

- Query 1: Determine the different unique launch sites.
- Query 2: Display 5 records where the launch site begins with ‘CCA’
- Query 3: Determine the total payload mass carried by NASA
- Query 4: Display average payload carried by booster version F9 v1.1
- Query 5: Display the date for the first successful outcome in ground pad.

EDA with SQL Cont.

- Query 6: List the boosters that have successfully landed on drone ships with payload masses between 4000 and 6000
- Query 7: Determine the number of successful and failed missions
- Query 8: List the booster versions that have carried the max payload mass
- Query 9: List the failed missions in the year 2015
- Query 10: Rank the landing outcomes from 2010-06-04 to 2017-03-20 in descending order

- https://github.com/Lawson-Pennel1127/IBM_CapstoneProject/blob/7b241e7d54fe6aa01aff2be0934ac4ff31ba8556/jupyter-labs-eda-sql-coursera_sqllite.ipynb

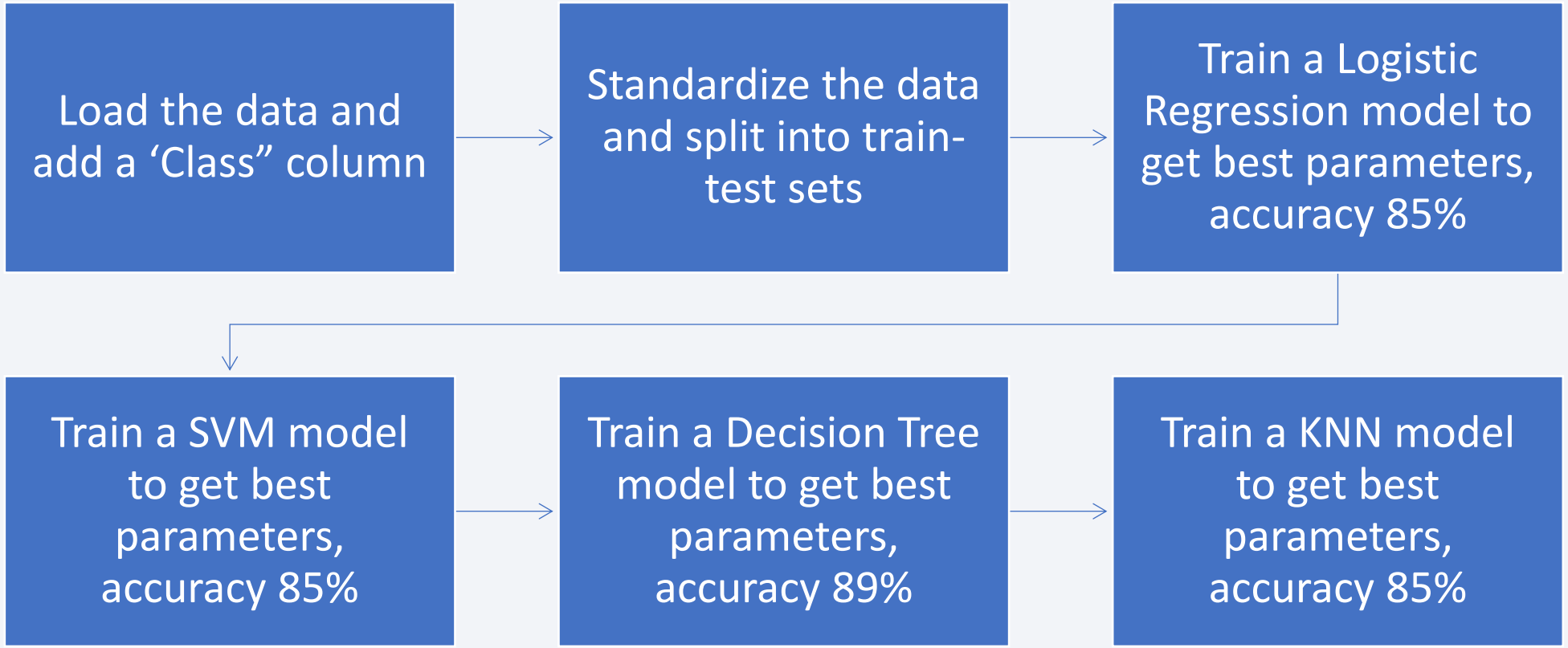
Build an Interactive Map with Folium

- Circles and Markers were used to show the locations of different launch sites.
- Marker Clusters were added to compactly mark the different launches and their success/failure.
- A line was added to show the distance from a certain launch site to the nearby coast.
- These objects where added, so that information about the launch sites could be gathered in a simple visual way on an interactive map.
- https://github.com/Lawson-Pennel1127/IBM_CapstoneProject/blob/7b241e7d54fe6aa01aff2be0934ac4ff31ba8556/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Dashboard Features:
  - Dropdown menu to select different launch sites.
  - Pie chart to quickly display successful launch counts for any and all sites.
  - A slider to select payload mass range.
  - A scatter plot to compare Payload mass and success rate by site.
- This plots and interactions were added for the sake of making a simple, easily navigable, and descriptive dashboard, that would let any user find the specific information they needed.
- https://github.com/Lawson-Pennel1127/IBM_CapstoneProject/blob/7b241e7d54fe6aa01aff2be0934ac4ff31ba8556/spacex-dash-app.py

Predictive Analysis (Classification)



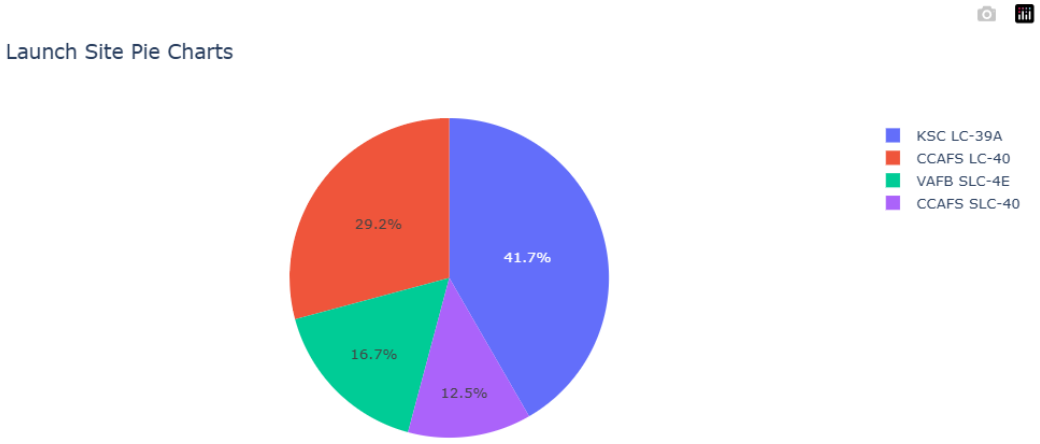
- Based on the best accuracy scores of the models, Decision Tree is the best, and most accurately identifies when a rocket will not land, while remaining highly accurate in predicting when it will land.

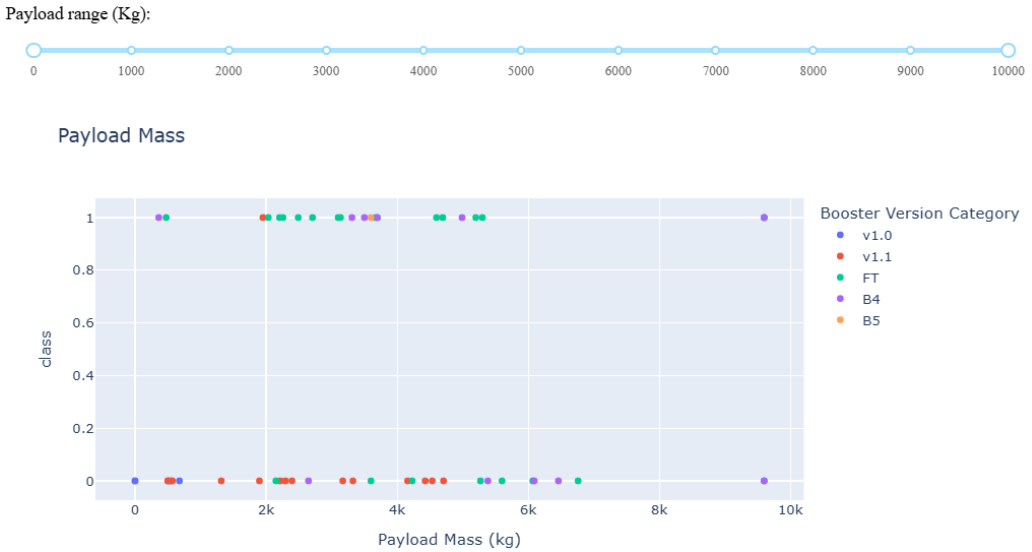
- https://github.com/Lawson-Pennel1127/IBM_CapstoneProject/blob/7b241e7d54fe6aa01aff2be0934ac4ff31ba8556/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

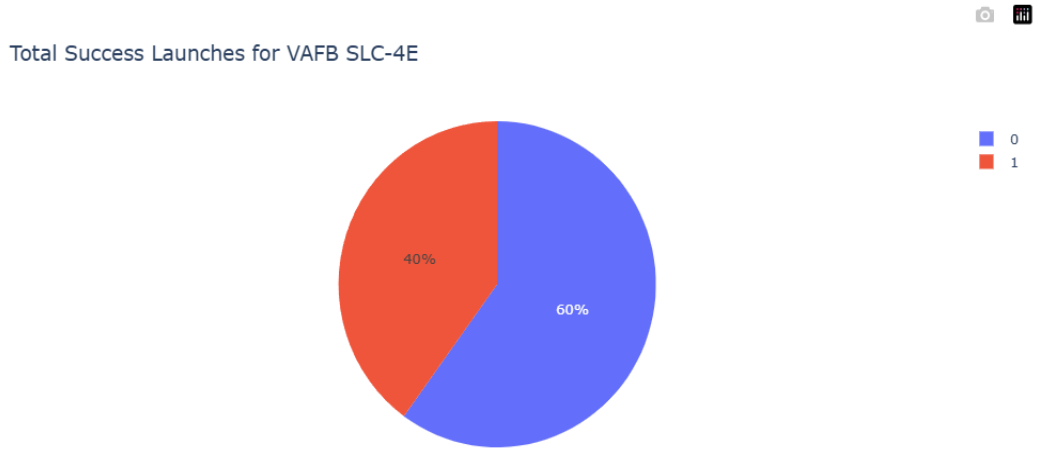
Results (EDA)

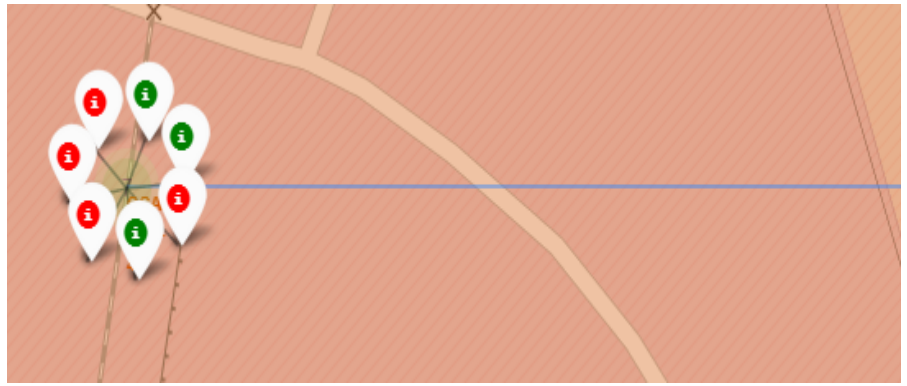
- Analysis showed that:
  - Mission type ASDS has the highest success rate with 41 successful mission and only 6 failed.
  - GTO was the most common orbit type.
  - The current average success rate of Stage 1 landings is 66%.
  - More flights generally results in a higher success rate for all launch sites.
  - A higher payload mass generally leading to a higher success rate.
  - The success rate of Stage 1 landings improves every year.

Results (Map/Dashboard)



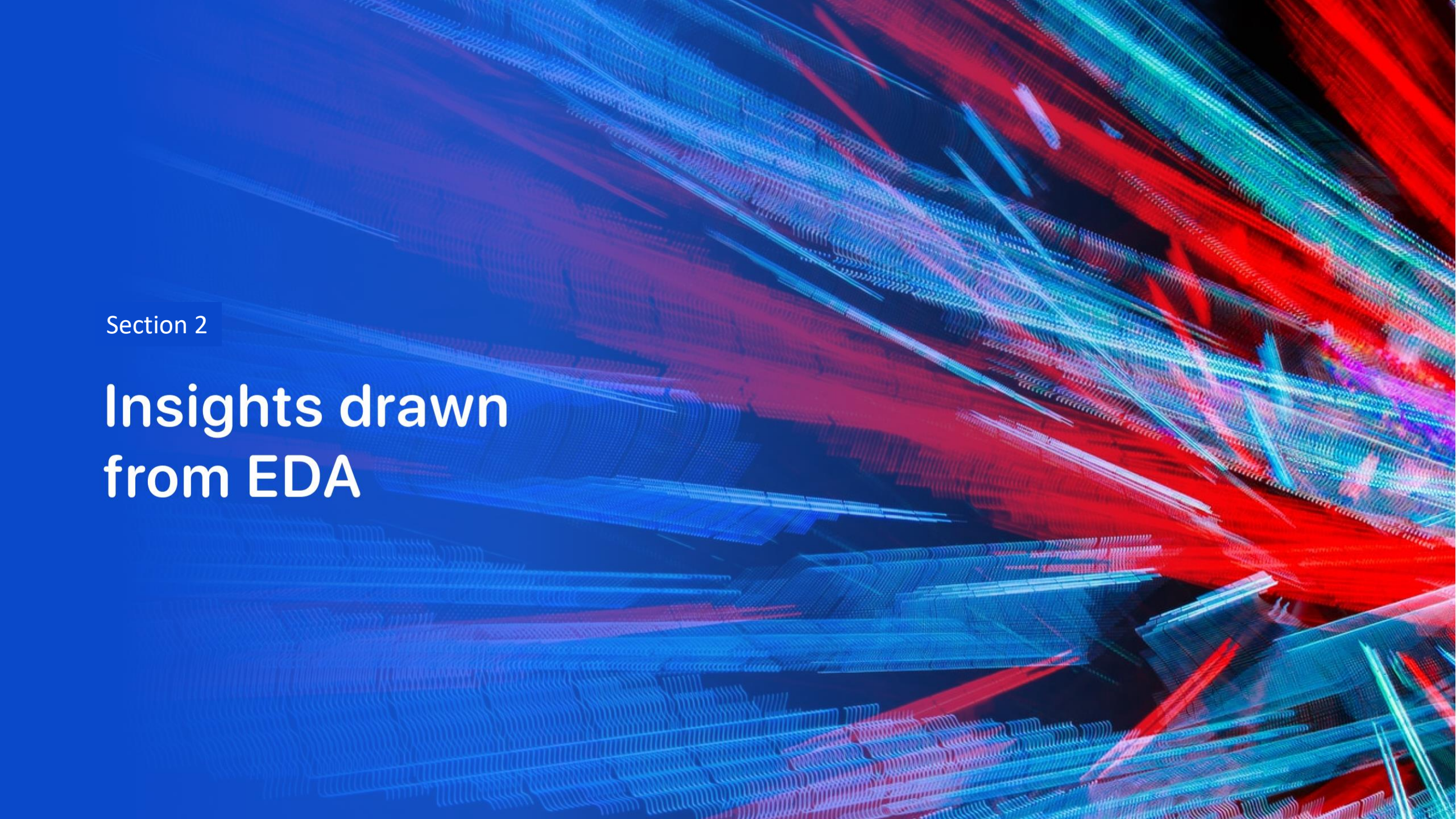






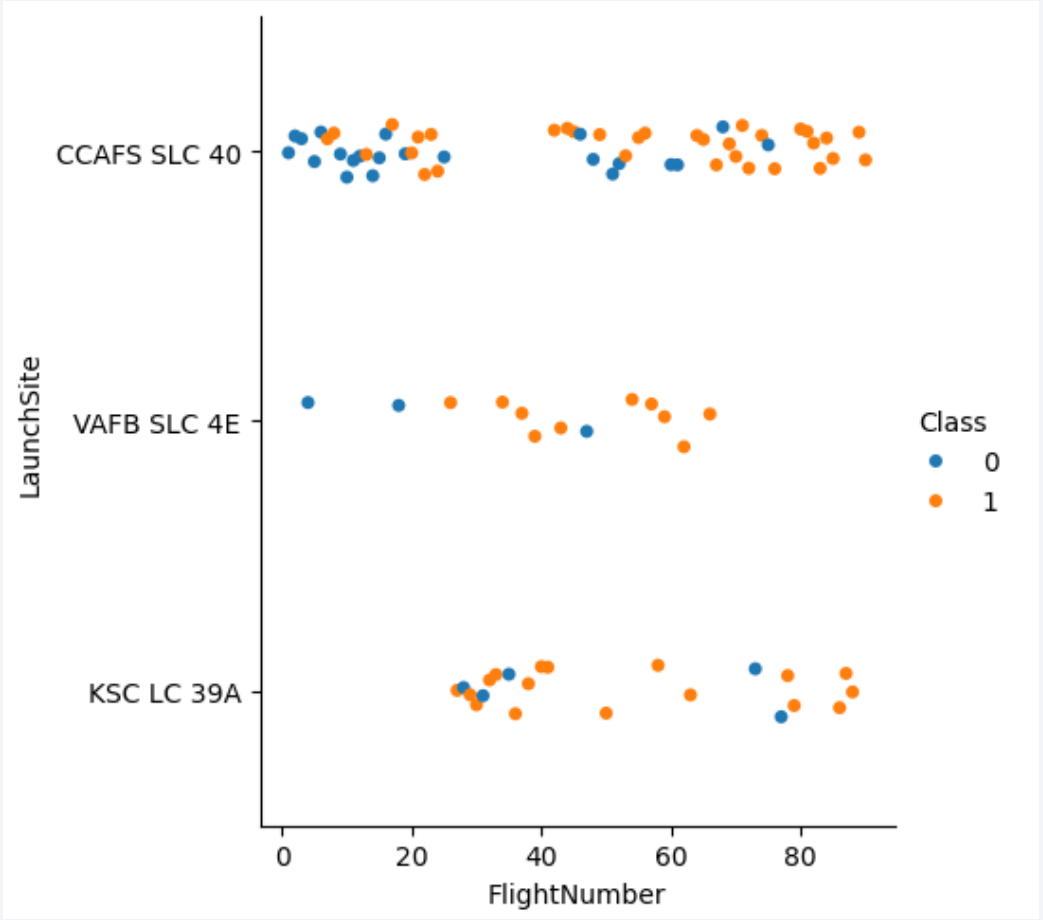
Results (Predictive Analysis)

- Analysis showed that:
  - The Decision Tree had the highest accuracy score with 89%, compared to the other models 85%
  - It was also the most accurate at predicting when a launch will fail to land, with 5 out of 6 predicted correctly, compared to the other models 3-3 split.
  - Compared to other models Decision Tree was worse at predicting when a launch would successfully land, only being correct 10 out of 12 times.


Section 2
Insights drawn from EDA

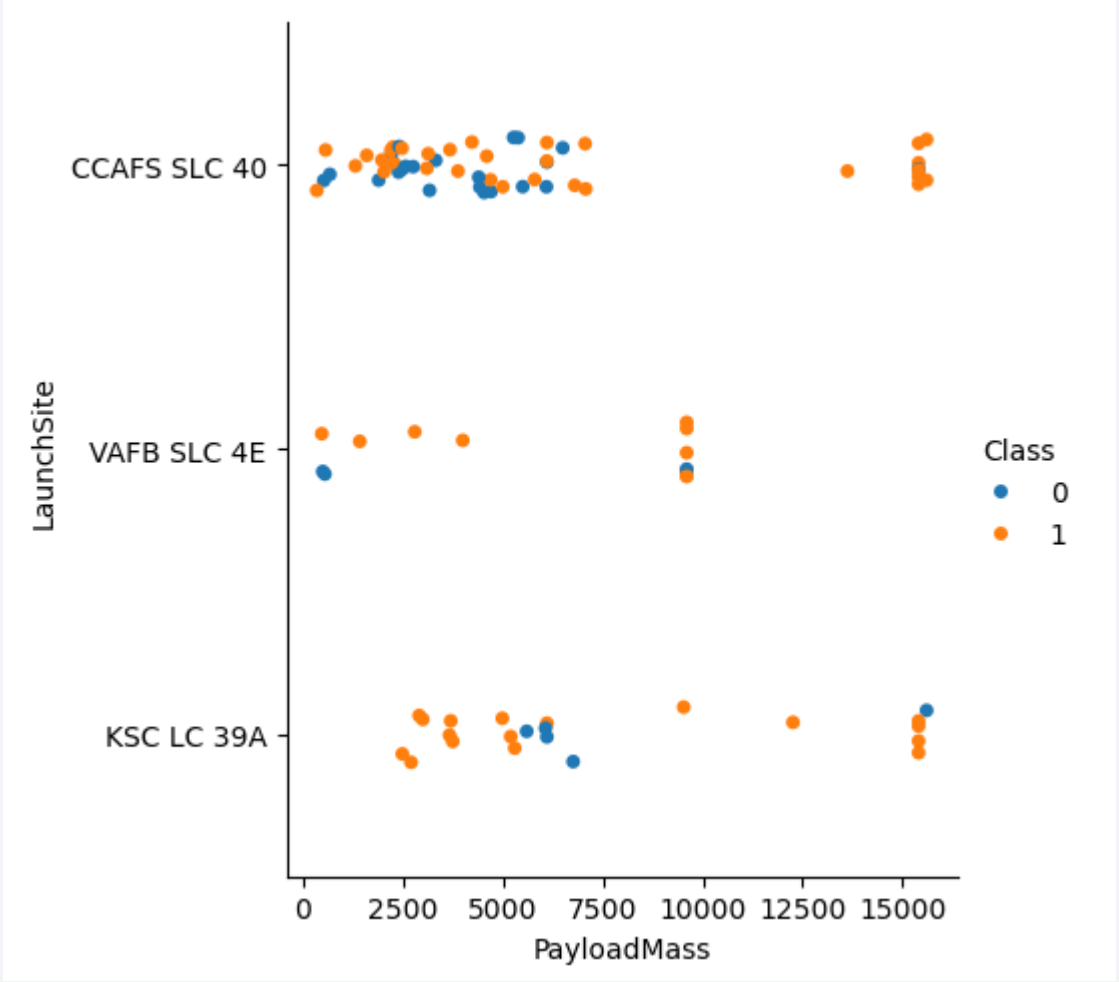
Flight Number vs. Launch Site

- The graph showing Flight Number vs Launch Site shows that, generally, with more flights the success rate increases.



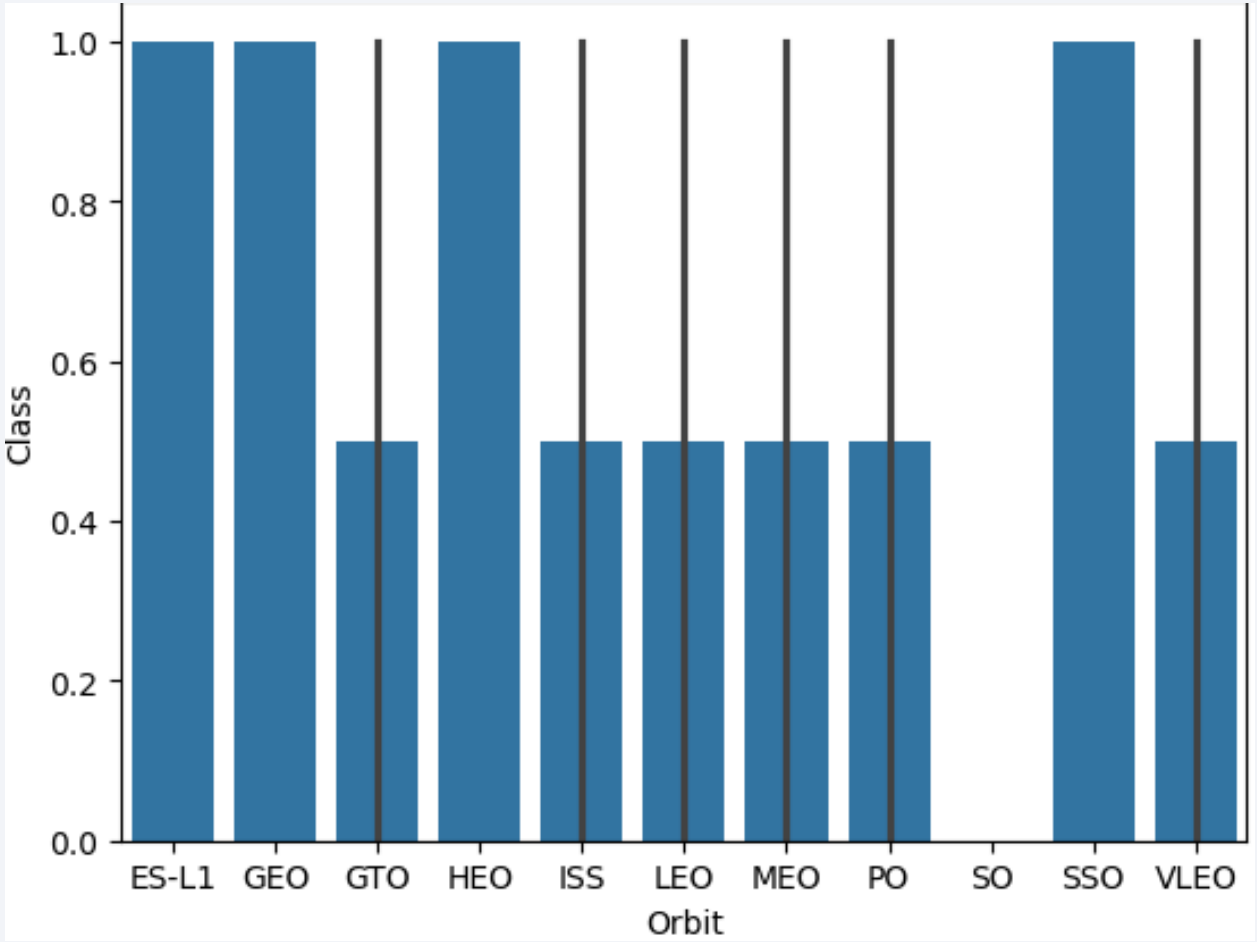
Payload vs. Launch Site

- This scatter plot is of Payload Mass vs Launch Site
- It shows that with a higher payload mass the success rate increases
- It also shows that Launch Site VAFB SLC 4E does not launch with payloads higher than 10000



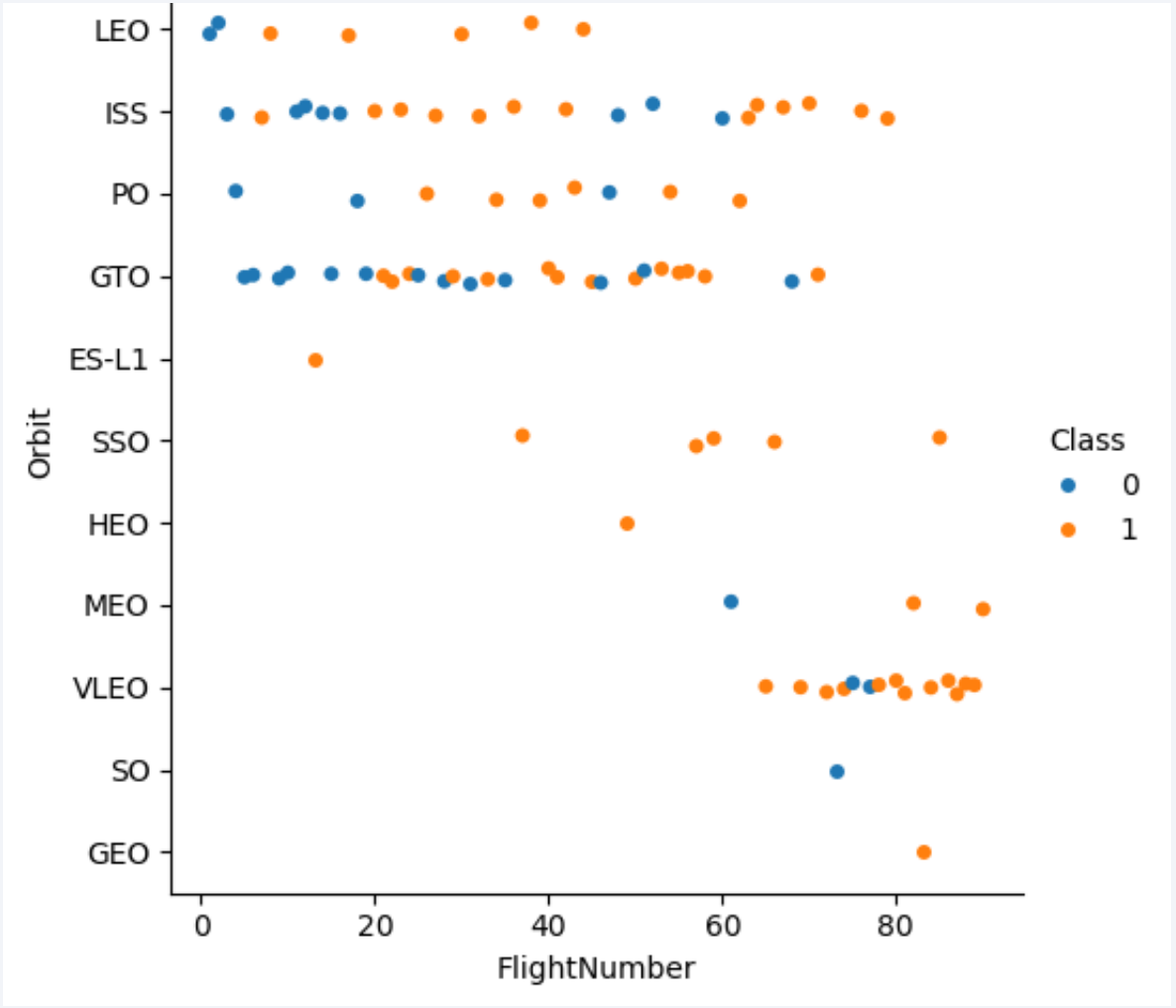
Success Rate vs. Orbit Type

- The bar plot displays Success Rate vs. Orbit Type
- The graph shows that ES-L1, GEO, and SSO have 100% success rated, although some of these orbits have had only one launch.
- All other Orbits, except SO, have a success rate around 50% with a large error range.



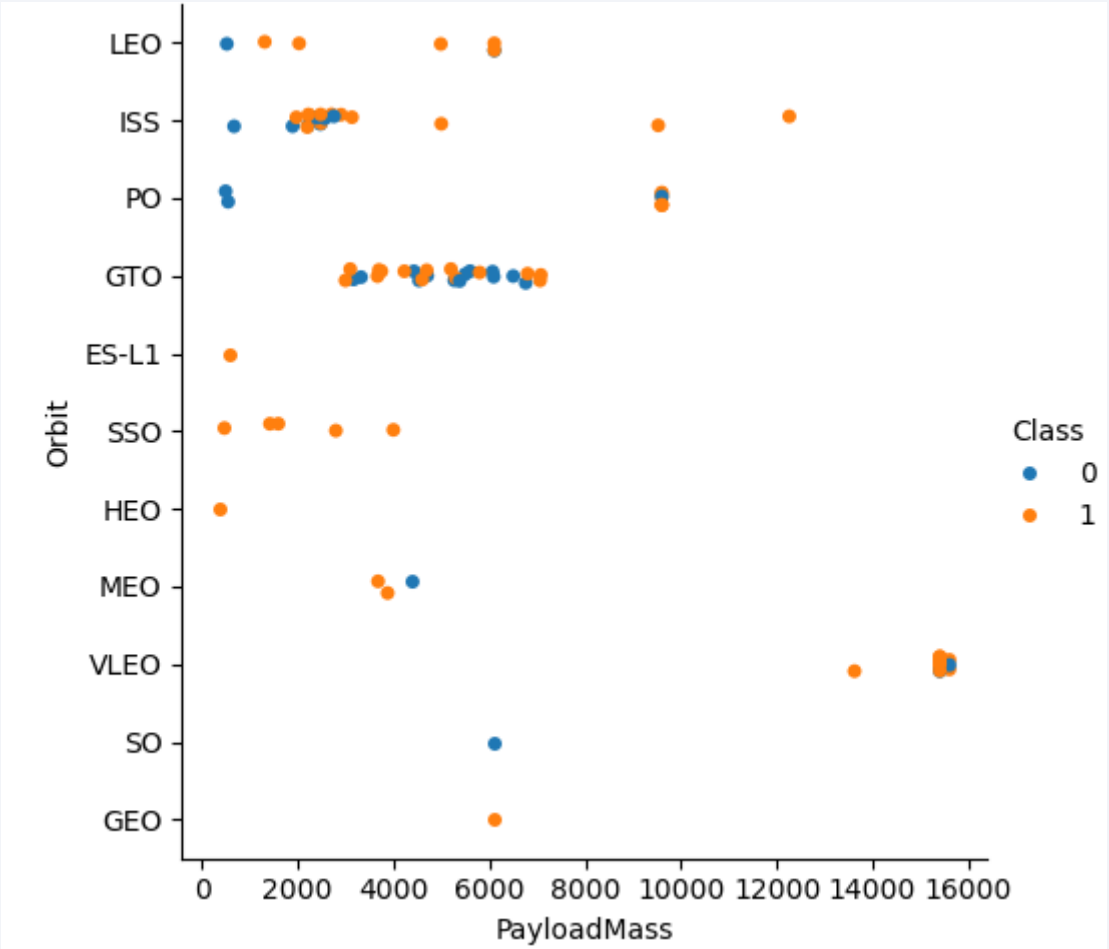
Flight Number vs. Orbit Type

- This graph is of Flight Number vs. Orbit Type
- It shows that, in general, more flights with a certain orbit leads to higher success rates.
- GTO is an exception to the trend.



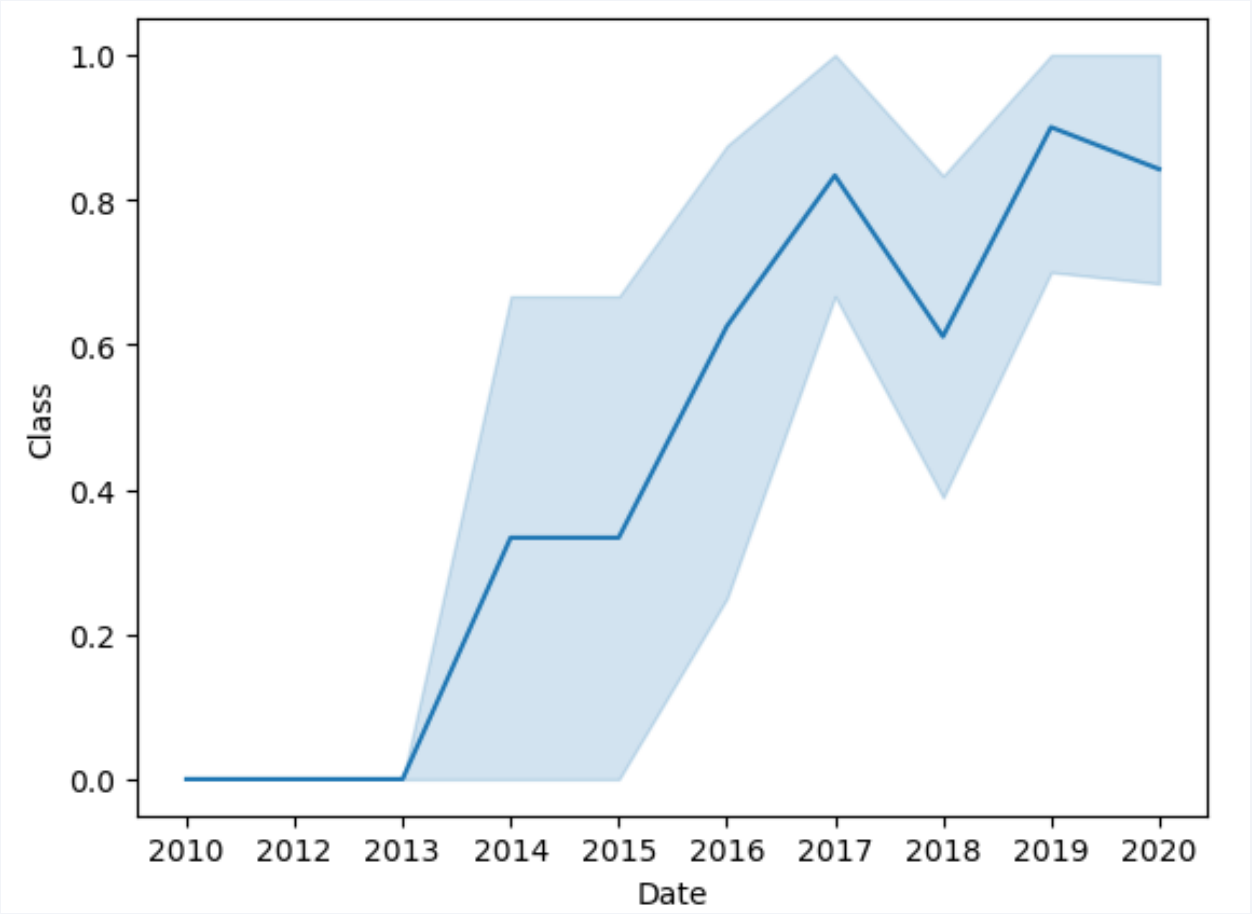
Payload vs. Orbit Type

- This graph shows Payload Mass vs. Orbit
- Displays a trend where higher payload mass increases success rate.
- GTO is once again an exception.



Launch Success Yearly Trend

- This graph shows the average success rate of multiple years.
- Displays a trend of increasing success rates over time.
- There is a clear dip in 2018.



All Launch Site Names

- This Figure shows the four unique Launch Sites
- Gotten using SQL

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

Launch Site Names Begin with 'CCA'

- Query Results show the first 5 launches in the dataset.

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

Total Payload Mass

- Total Payload Mass carried by NASA
- Gotten with SQL

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Total payload mass carried by booster version F9 v1.1
- Gotten with SQL

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

- Date of the first successful landing outcome on ground pad
- Gotten with SQL

MIN(DATE)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- Gotten with SQL

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

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes.
- The 96 denotes successful mission, while the three 1 all denote different types of failed missions.
- 96 successful mission, 3 failed

COUNT("Mission_Outcome")
1
98
1
1

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass
- Gotten with SQL

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

- The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Gotten with SQL

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

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

- Ranking of the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order
- Gotten with SQL

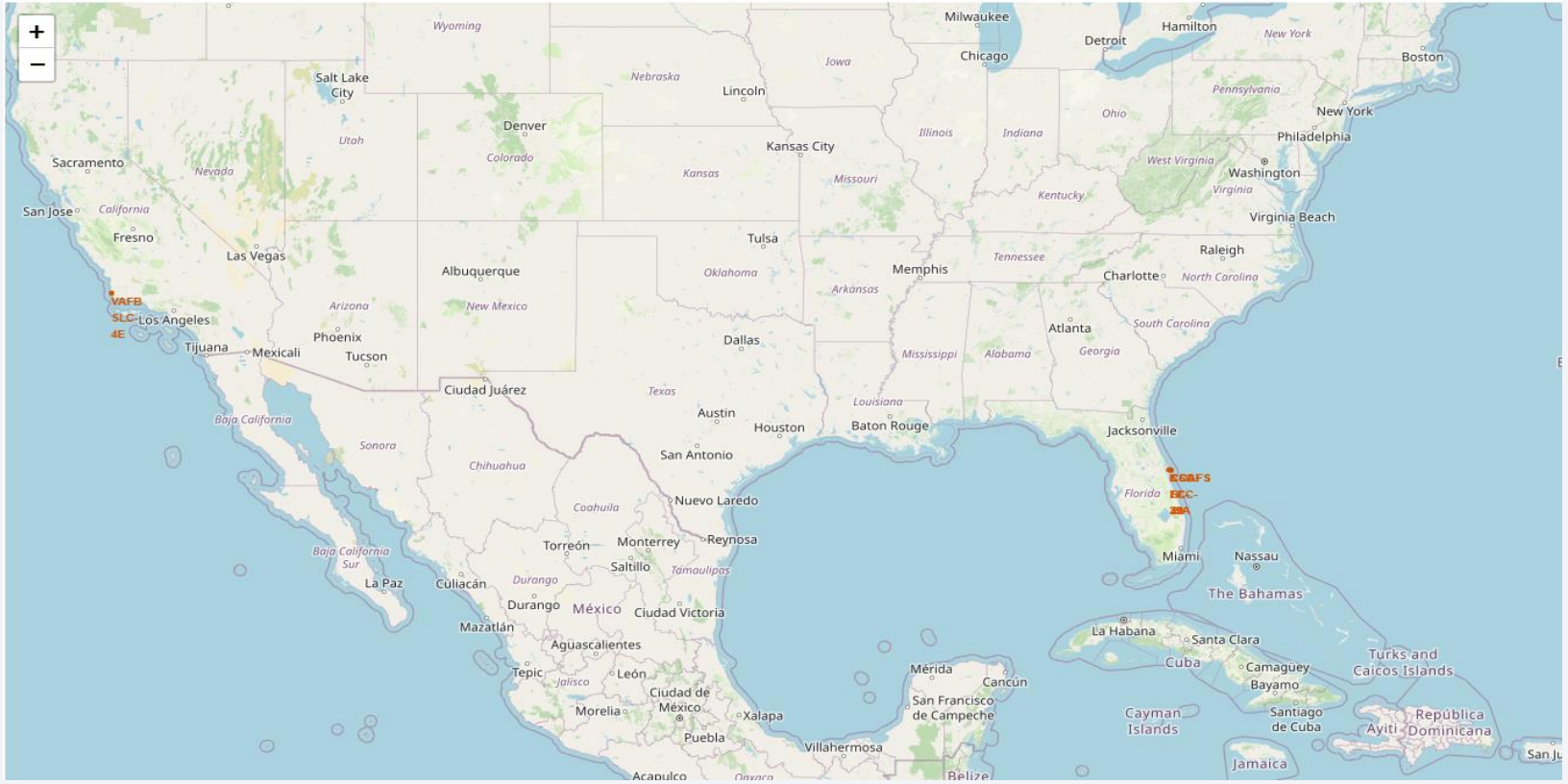
Landing_Outcome	count1
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

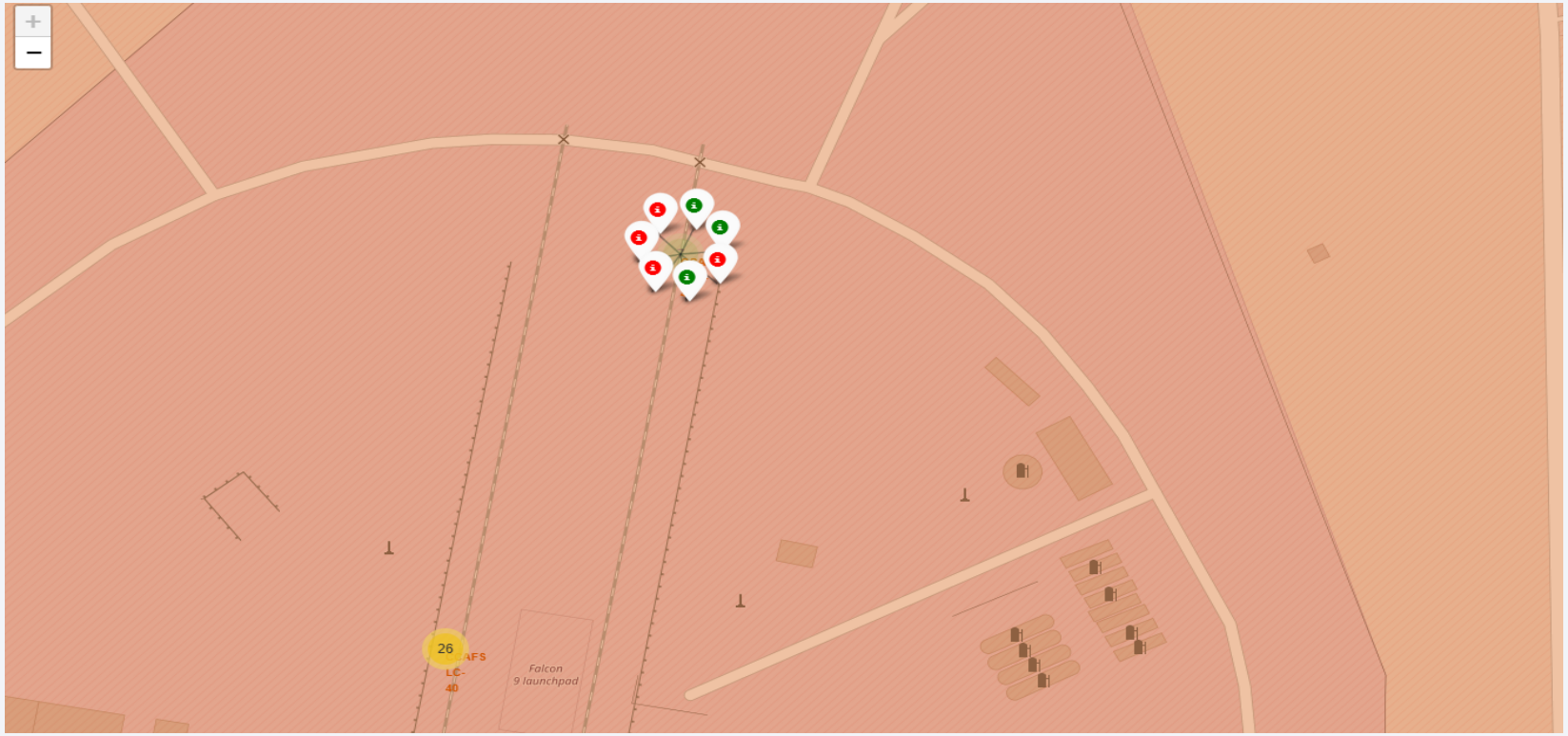
Launch Site Map

- This screenshot shows the different areas for F9 launches, mainly being Florida and California



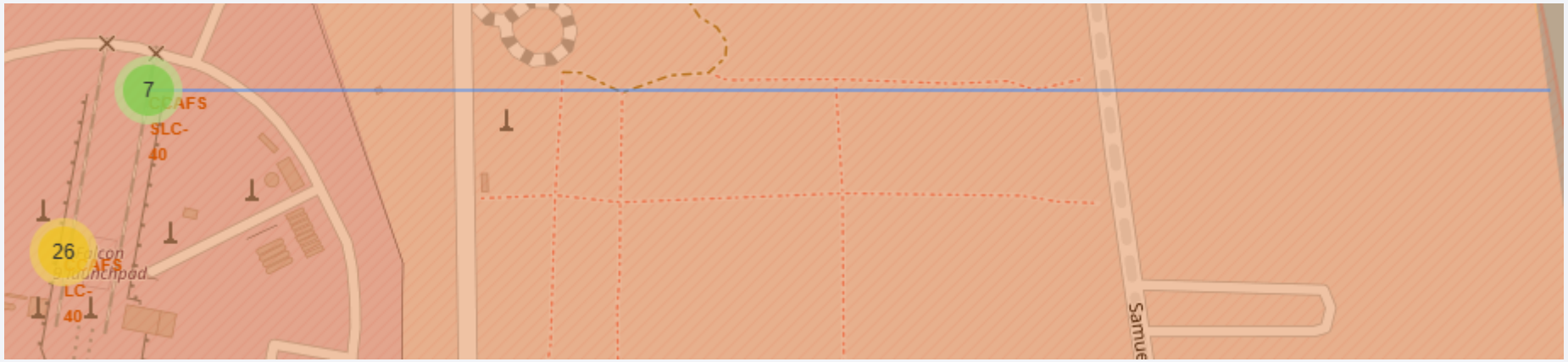
Folium Marker Cluster

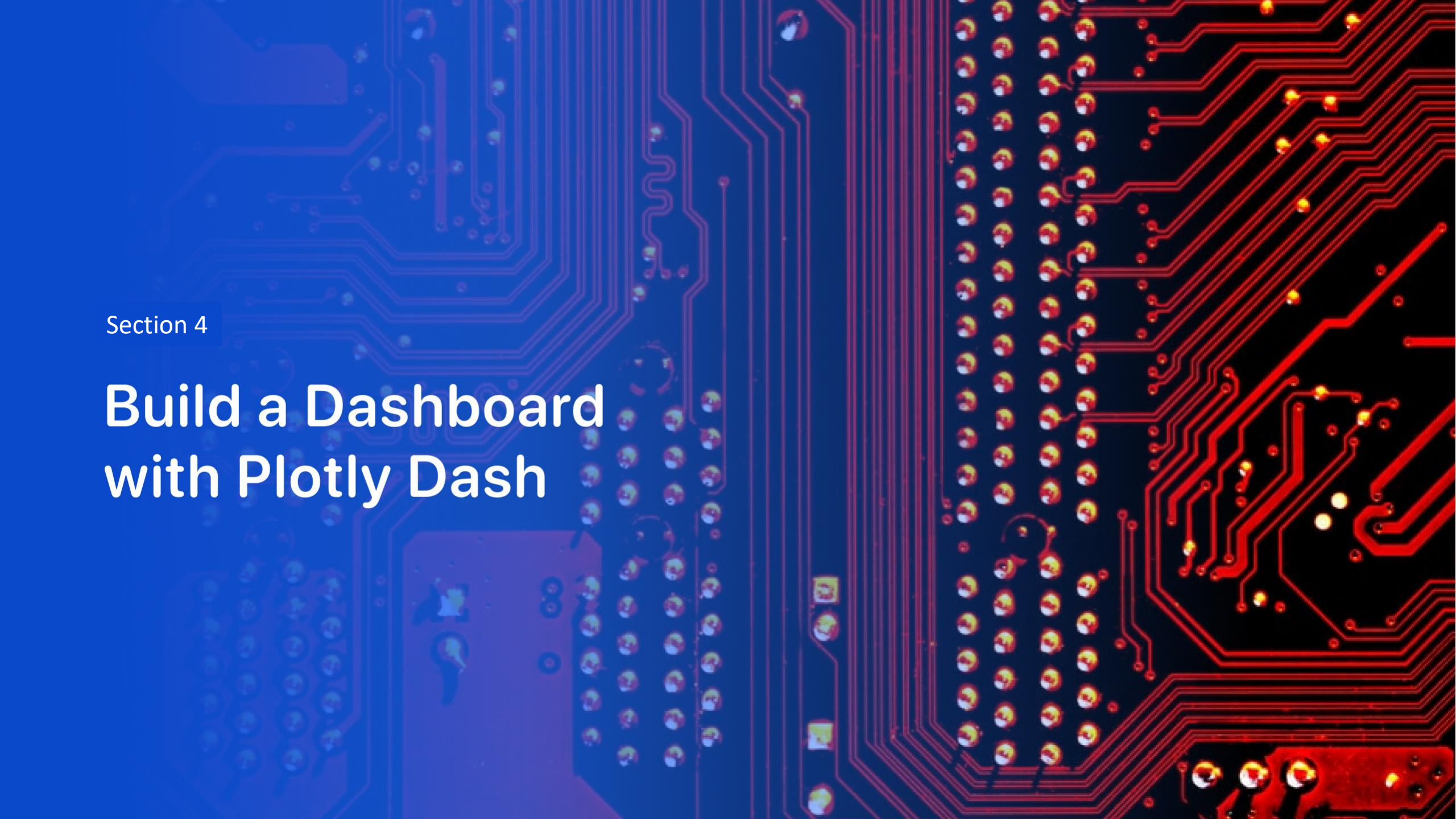
- This screenshot shows a launch site with color coded markers denoting success or failure. Green = Success, Red = Failure



Folium: Distance to Coast

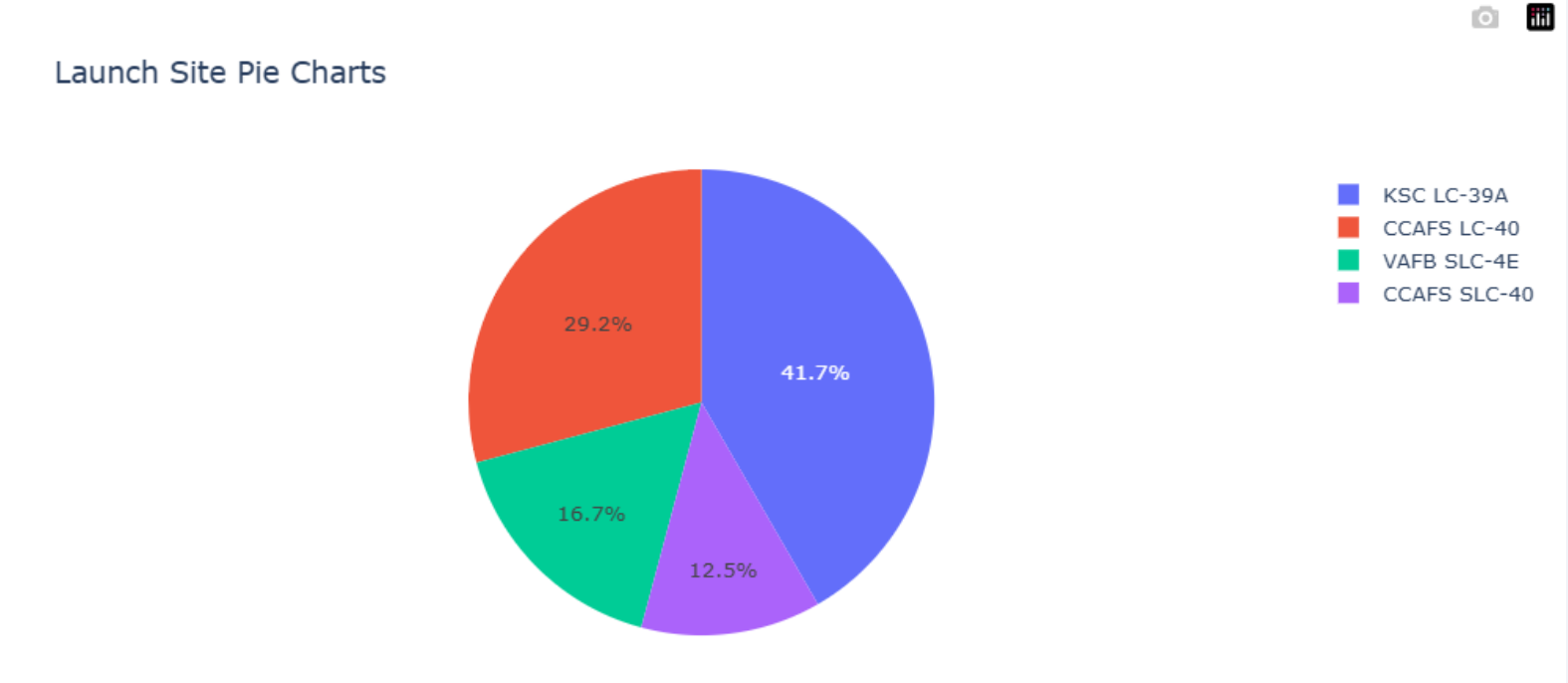
- This screenshot shows a line drawn from a lunch site to the nearest coastline.
- The distance is denoted with a blue line




Section 4
Build a Dashboard with Plotly Dash

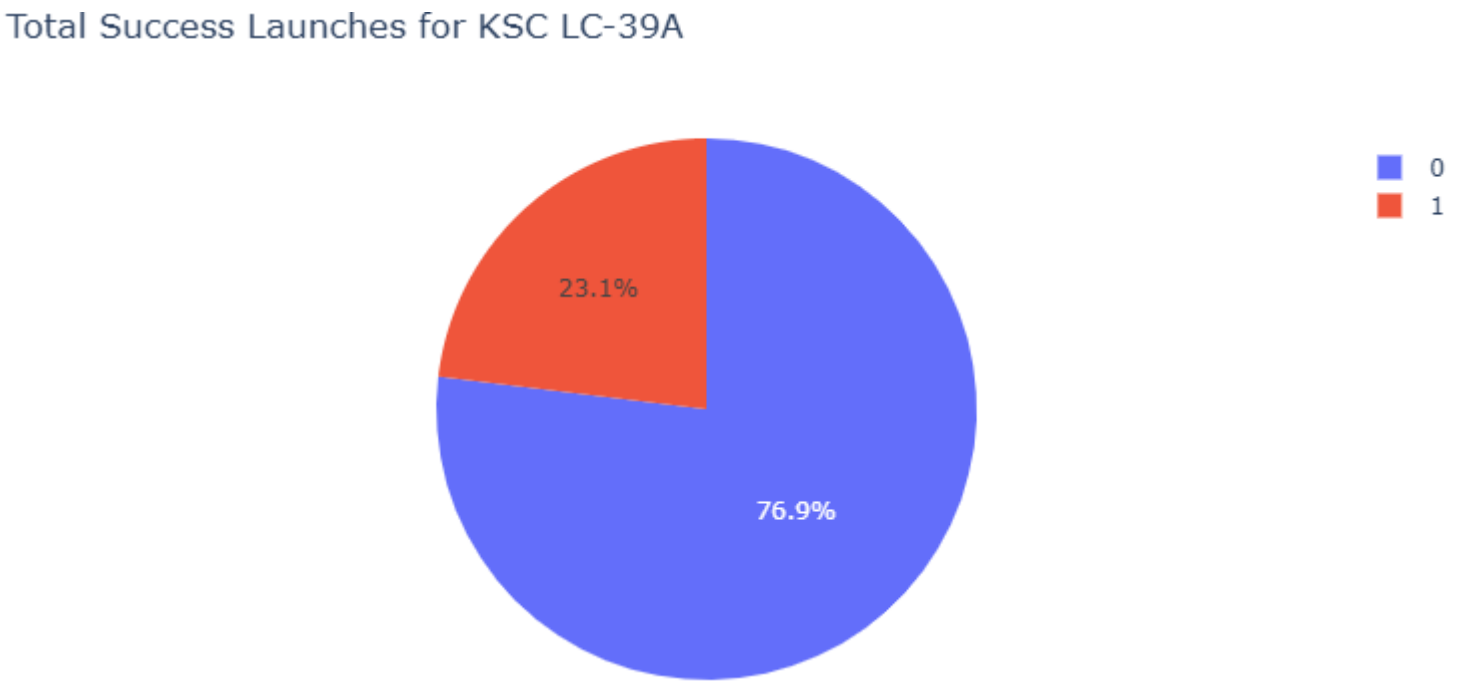
Total Succes Rate Pie Chart

- This Pie Chart shows what percentage of total successful missions each launch site has. With KSC LC-39A having the highest percent with 41.7%



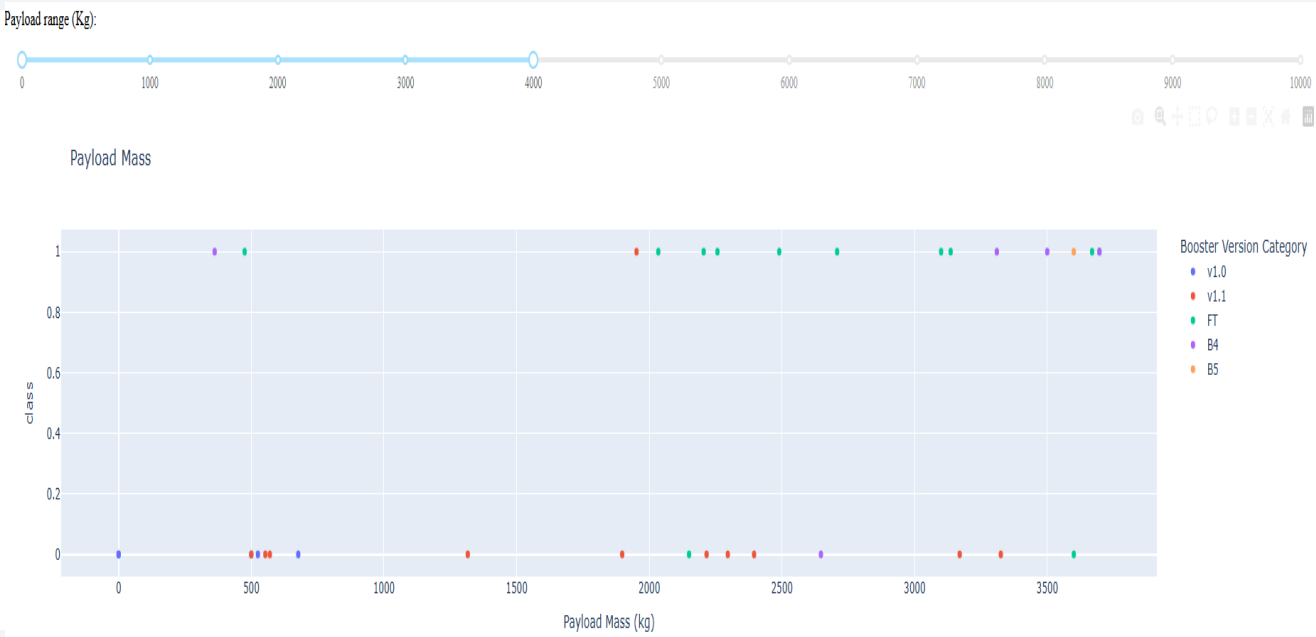
Most Successful Launch Site Pie Chart

- This pie chart shows that KSC LC-39A has the highest success rate of any launch site with a rate of 76.9%



Payload Mass vs Success by Booster Type

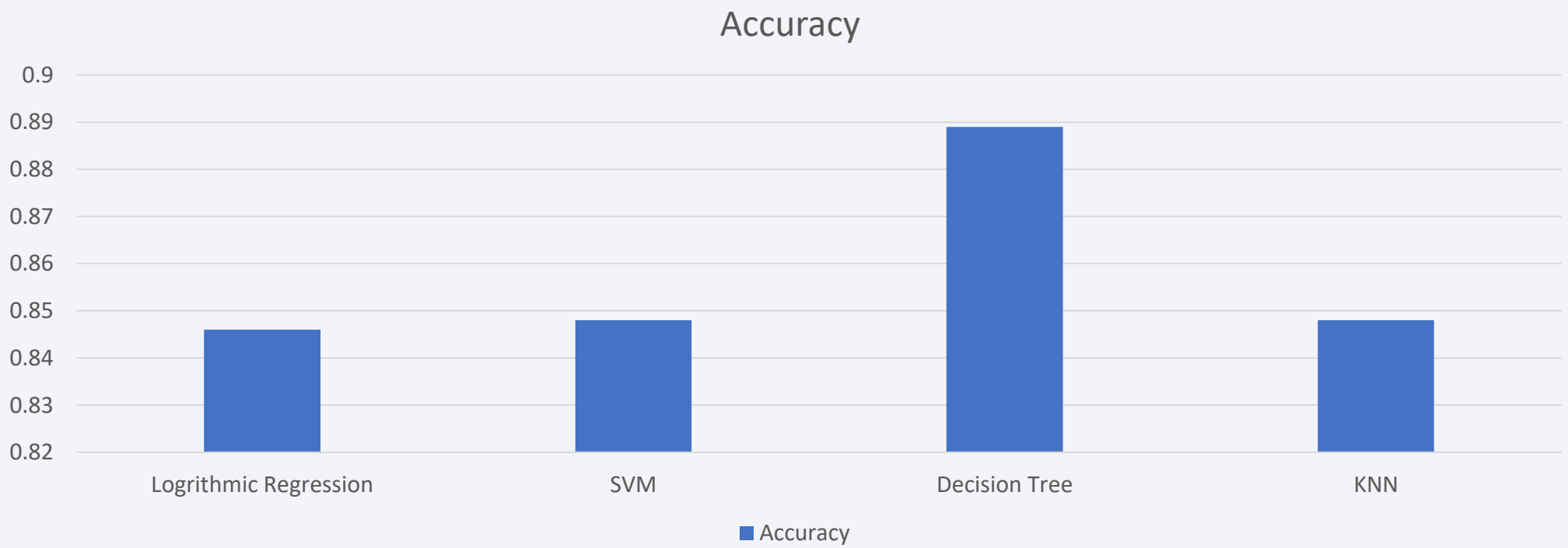
- These screenshots show a scatter plot comparing success and payload mass by booster type. It clearly shows that B4 is able to carry the highest payload mass.



Section 5

Predictive Analysis (Classification)

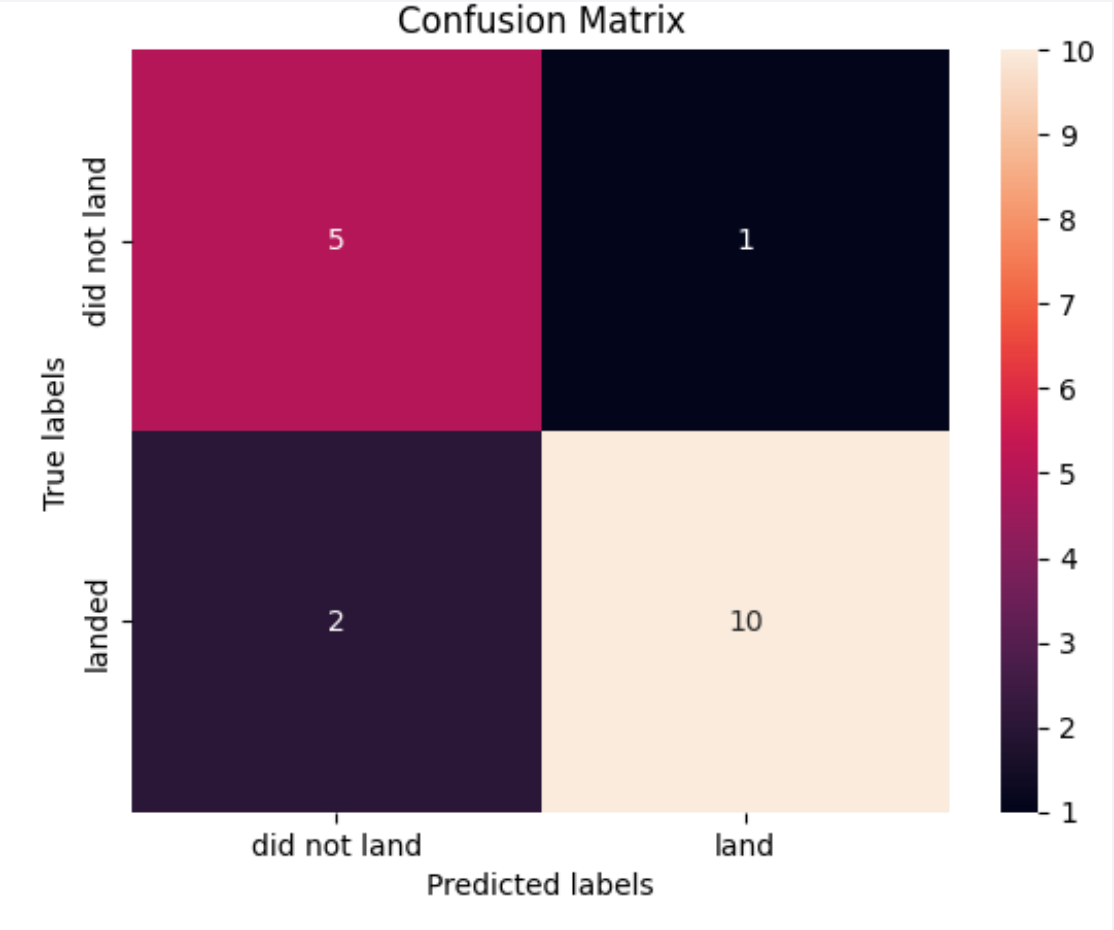
Classification Accuracy



- Decision Tree has the highest accuracy, although only by a slight margin. 87% compared to the other models 84%

Confusion Matrix

- The figure is a confusion matrix. It shows how accurate the Decision Tree model was at predicting if the mission would succeed.
- The matrix shows that it correctly predicted a successful landing 10/12 times and correctly predicted a failed landing 5/6 times.



Conclusions

- Mission type ASDS has the highest success rate.
- A higher payload mass results in a higher chance for success.
- KSC LC-39A is the most successful Launch Site.
- A decision tree model is able to most accurately predict the outcome of a mission.
- Using a decision tree model the success rate of mission could go up significantly, going from the current 66% to 87%

Appendix

- All models had an accuracy of 84% on average but the best accuracy scores set them apart.
- Some launches were missing launch site data.
- The orbits: ES-L1, HEO, GEO, and SO only have one data points so predictions of them would be unreliable.

Thank you!