IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Simon Goudie
September 2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection - SpaceX API
  - Data Collection - Scraping
  - Data Wrangling
  - Exploratory Data Analysis with Data Visualization
  - Exploratory Data Analysis with SQL
  - Interactive Map with Folium
  - Dashboard with Plotly Dash
  - Predictive Analysis

- Summary of insights drawn from EDA
  - Multivariate analysis
- Launch Sites Proximity Analysis
  - Locations
  - Outcomes
  - Proximities
- Dashboard with Plotly Dash
  - Success and payload analysis
- Predictive Analysis Outcomes
  - Classification outcome and confusion matrix
- Conclusion

Introduction

- Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage.
- Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against Space X for a rocket launch.
- This report analyses many factors involved in successful launches and landings, highlighting the optimal parameters for each.

Section 1

Methodology

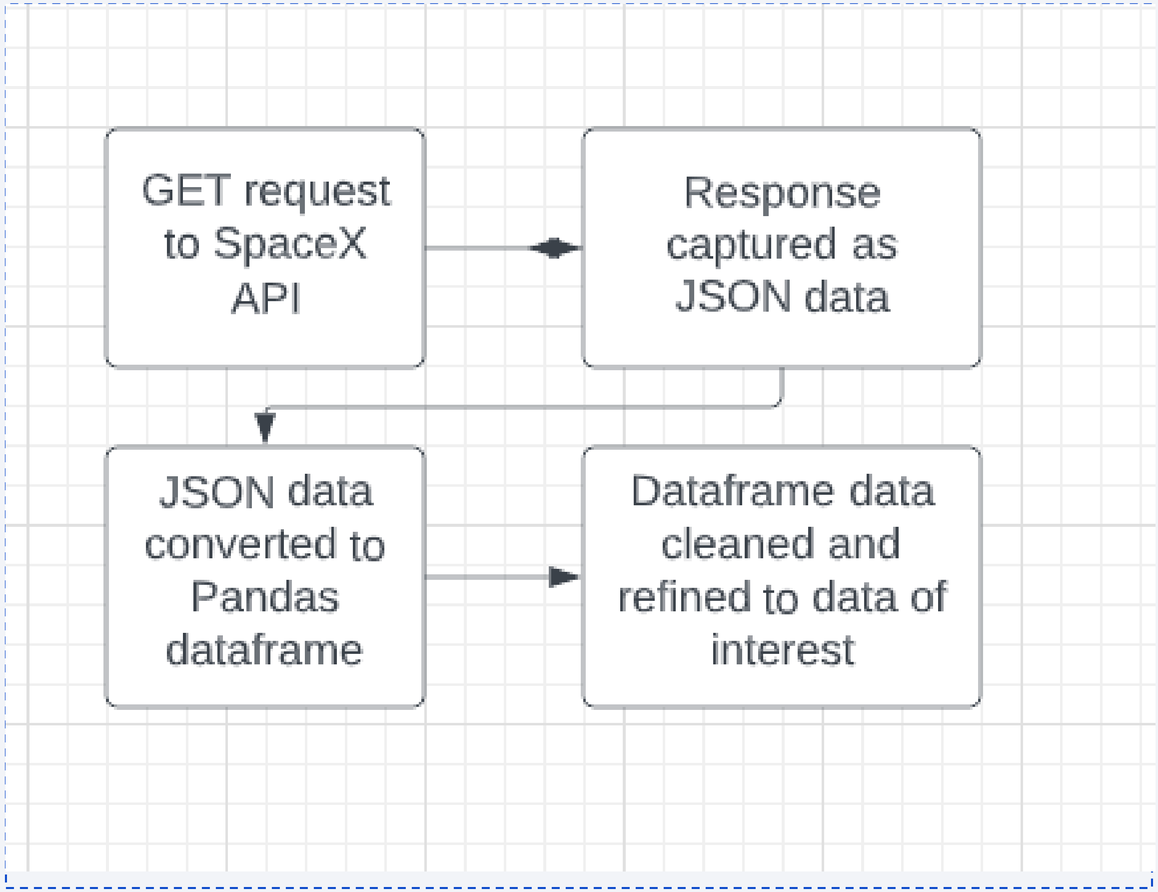
Methodology

Executive Summary

- Data collection methodology:
  - Data was collected via the SpaceX API and via web scraping
- Perform data wrangling
  - Missing data was replaced with means, launch data was aggregated and one hot encoding was used to show outcomes as binaries
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - A number of models were built, then trained on sample data before being tested
  - The best-performing model was then identified based on test scores

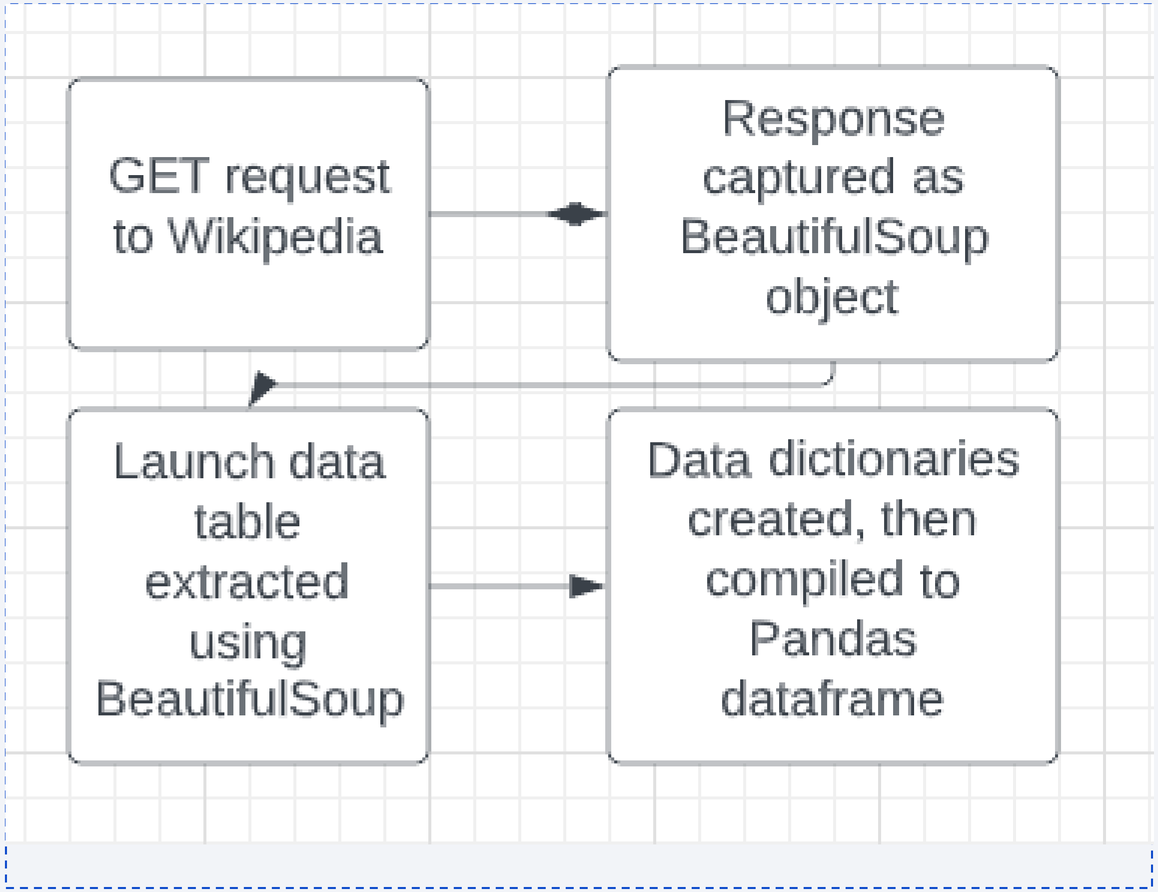
Data Collection – SpaceX API

- SpaceX data was collected via the SpaceX API using REST GET requests. For the actual processing, a static response object was used for consistency.
- The JSON data was then processed into a Pandas dataframe, cleaned and trimmed down to only the items of interest.
- GitHub URL of the completed SpaceX API calls notebook: https://github.com/simongoudie/Applied-Data-Science-Capstone/blob/main/1%20jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- Launch records were scraped via Wikipedia using REST GET requests. For the actual processing, a static response object was used for consistency.
- The scraped data was saved into a BeautifulSoup object, then the table data was extracted to dictionaries and then finally to a Pandas dataframe
- GitHub URL of the completed web scraping notebook: https://github.com/simongoudie/Applied-Data-Science-Capstone/blob/main/2%20jupyter-labs-webscraping.ipynb



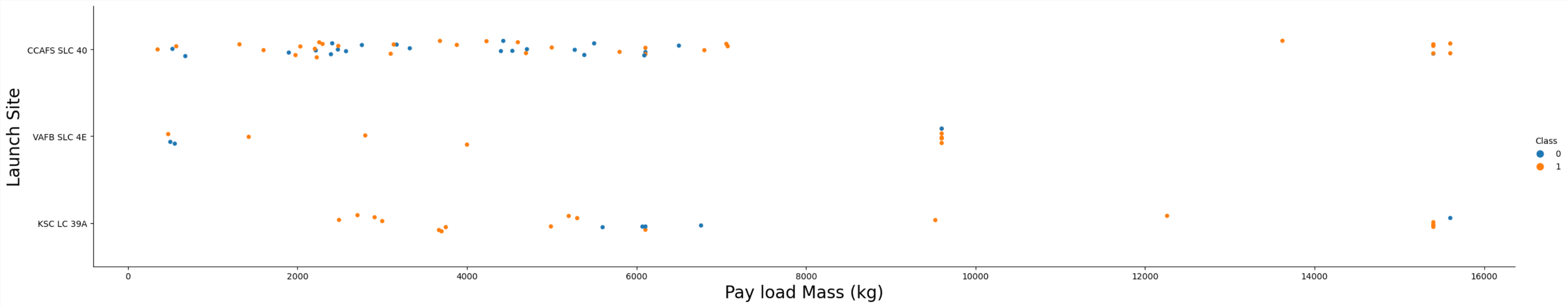
Data Wrangling

- Missing values were replaced in the data by column means.
- Launch data was aggregated to show launches per site, number and frequency of orbit and types of outcome.
- Outcomes were reduced to a binary showing success or failure to enable further analysis of overall success rates, with the new column added to the dataframe.

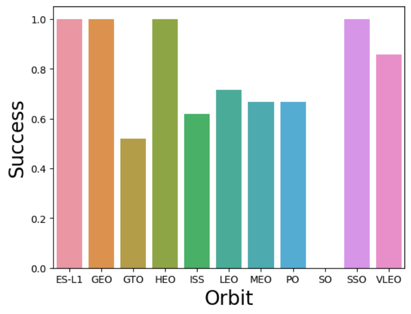
- GitHub URL of the completed data wrangling related notebook: https://github.com/simongoudie/Applied-Data-Science-Capstone/blob/main/3%20labs-jupyter-spacex-Data%20wrangling.ipynb

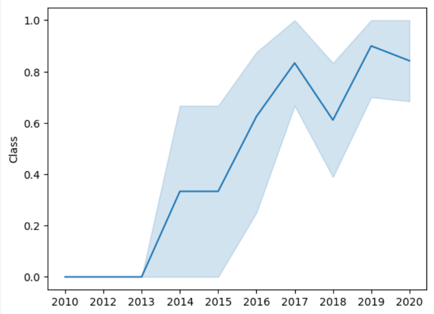
EDA with Data Visualization

- Scatter plots were used to visualize relationships between variables and success rates. e.g. Payload and Launch Site:



- Bar charts and line charts were used to visualize success rates and enable visual comparison variables. e.g.





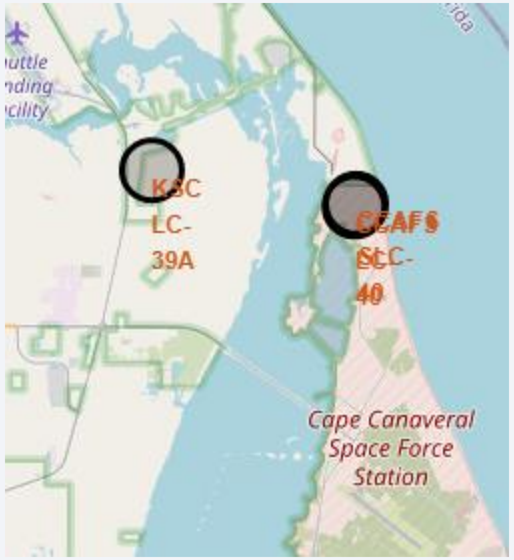
- GitHub URL of the completed EDA with data visualization notebook: https://github.com/simongoudie/Applied-Data-Science-Capstone/blob/main/5%20jupyter-labs-eda-dataviz.ipynb

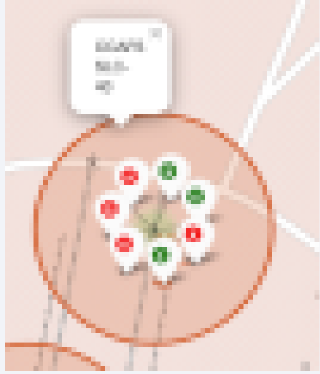
EDA with SQL

- Summary of SQL queries performed:
  - Unique launch sites were listed
  - Five records with launch sites starting with ‘CCA’ were listed
  - Total payload mass was calculated
  - Average payload mass of F9 v1.1 boosters was calculated
  - Date of first successful landing on ground pad was discovered
  - Booster names that had drone ship success and payloads between 4-6,000kg were listed
  - Total number of successful and failing mission outcomes were listed
  - Booster names with the maximum payload mass were listed
  - Month names, drone ship failures, booster versions and launch sites were listed for 2015
  - Successful landing outcomes between July 2010 and March 2017 were ranked
- GitHub URL of the completed EDA with SQL notebook: https://github.com/simongoudie/Applied-Data-Science-Capstone/blob/main/4%20jupyter-labs-eda-sql-coursera_sqllite.ipynb

Interactive Map with Folium

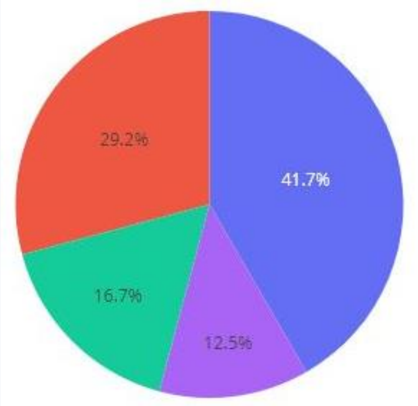
- Maps of the launch sites were created with the following objects:
  - All launch sites were labelled with text labels and circles for clear identification.
  - Launches were added with marker objects, color coded for success or failure to provide a visual indicator of success at each site.
  - Points, lines and labels were added to features such as coasts, cities, railways and highways to indicate proximity to these landmarks. Distance to these landmarks is key for provision of services to the launch sites, and to maintain safe launch distances from population centers.
- GitHub URL of the completed interactive map with Folium map: https://github.com/simongoudie/Applied-Data-Science-Capstone/blob/main/6%20lab_jupyter_launch_site_location.ipynb

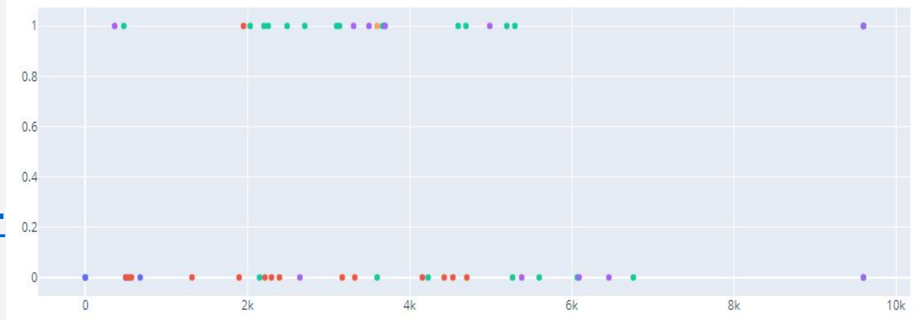




Dashboard with Plotly Dash

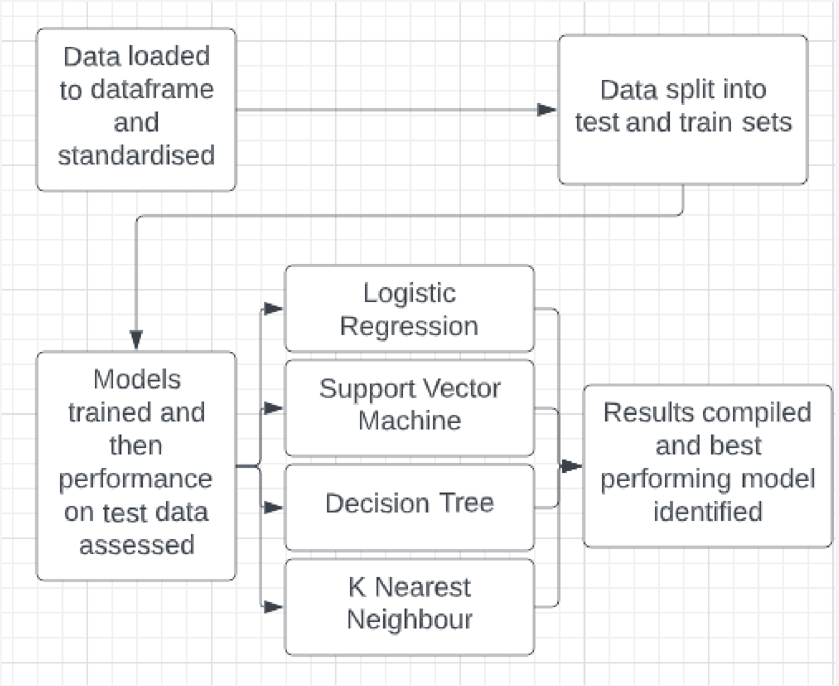
- A dashboard was created using pie charts and scatter plots:
  - The pie chart displayed at the launch success rate by site, filterable by site or 'all sites'.
  - The scatter plots mapped launch success against payload weight, filterable by booster type and by payload weight range.
- These filters allowed for adjustable views of the presented data, enabling the discovery of idea payload weight ranges and best-performing sites/boosters.
- GitHub URL of the completed Plotly Dash lab: https://github.com/simongoudie/Applied-Data-Science-Capstone/blob/main/spacex_dash_app%20(Plotly).py





Predictive Analysis (Classification)

- Multiple models were constructed and evaluated: Logistic Regression, Support Vector Machine, Decision Tree, K Nearest Neighbour
- Data was loaded into a dataframe, standardized and split into train/test sets
- Each model was trained and tested on these sets, with the optimal parameters from each then scored and compared
- On review, SVM gave the best result
- GitHub URL of the completed predictive analysis lab: https://github.com/simongoudie/Applied-Data-Science-Capstone/blob/main/7%20SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



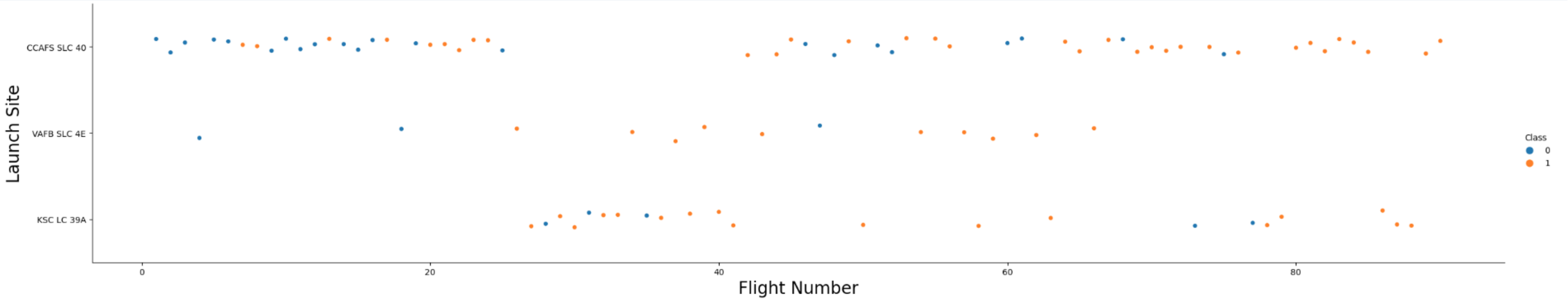
Results

- CCAFS SLC-40 was the most commonly-used launch site, with later launches having a higher success rate
- Payload range of 4,600 – 5,400kg tended to have a better success rate
- ES-L1, GEO, HEO, and SSO orbits all featured a high success rate
- Predictive analysis results indicated that a SVM model would have the best predictive accuracy for future launches

Section 2

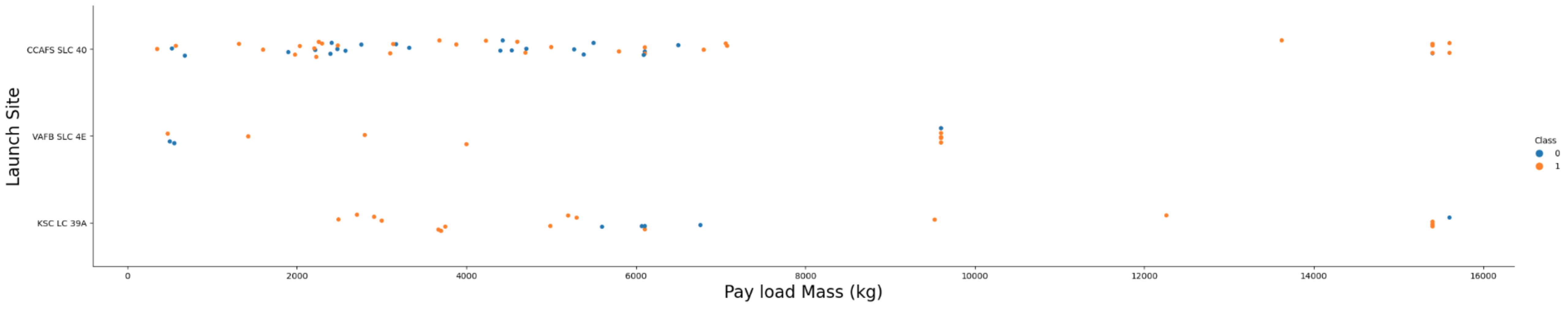
Insights drawn from EDA

Flight Number vs. Launch Site



- Scatterplot indicates that CCAFS SLC-40 was the most commonly-used launch site
- Early launches showed a higher failure rate
- Success rate improved over time and the final cluster of launches were all successful

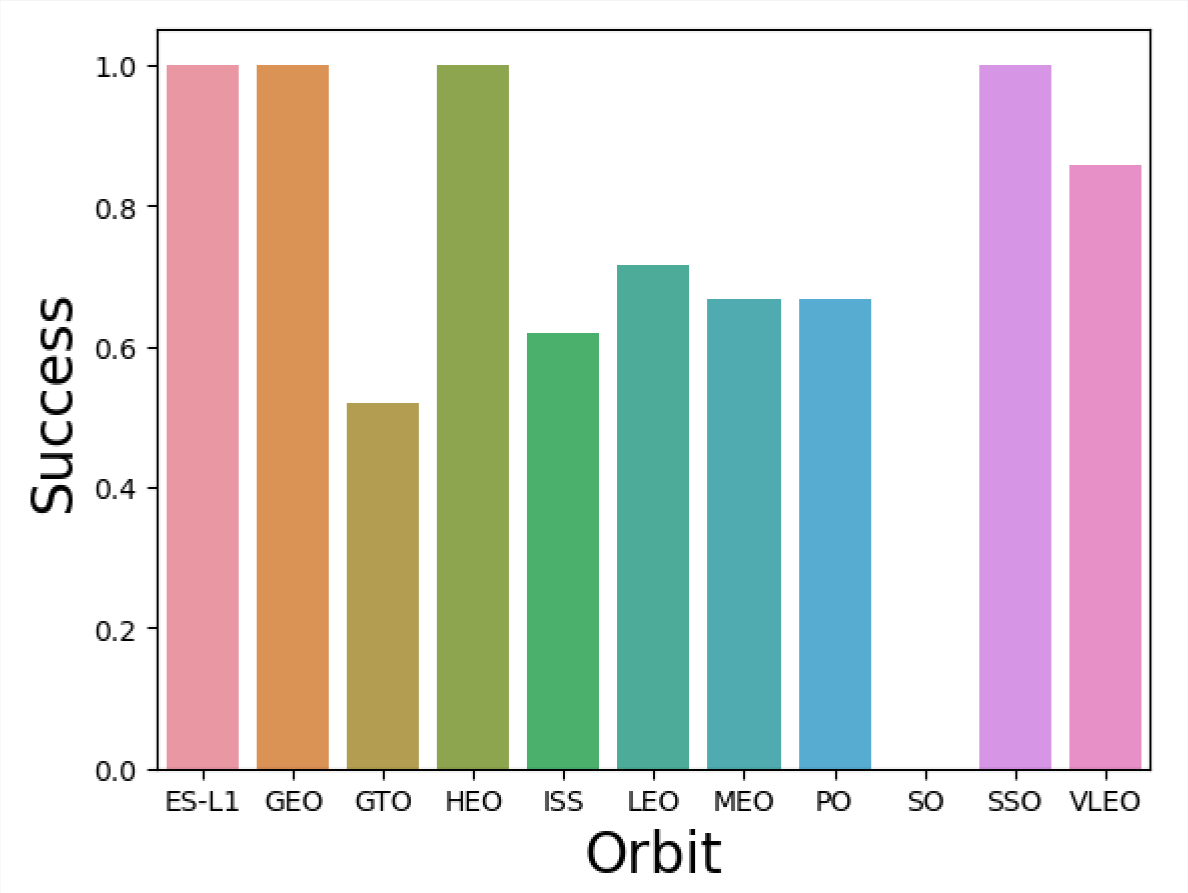
Payload vs. Launch Site



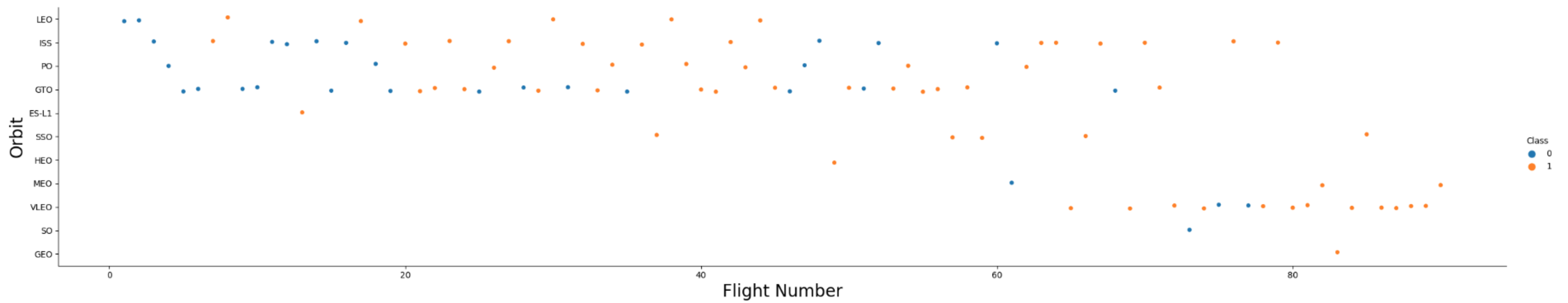
- Scatterplot shows clusters of payload amounts: 0-7,000kg, 9,500kg and ~17,000kg
- Higher payloads tended to have better launch success rates

Success Rate vs. Orbit Type

- Bar chart shows the difference success rates of orbit types
- ES-L1, GEO, HEO, and SSO orbits all featured a high success rate
- GTO, ISS, LEO, MEO, PO and VLEO had success rates of >1.0, but all were higher than 0.5.
- SO recorded a 0.0 success rate

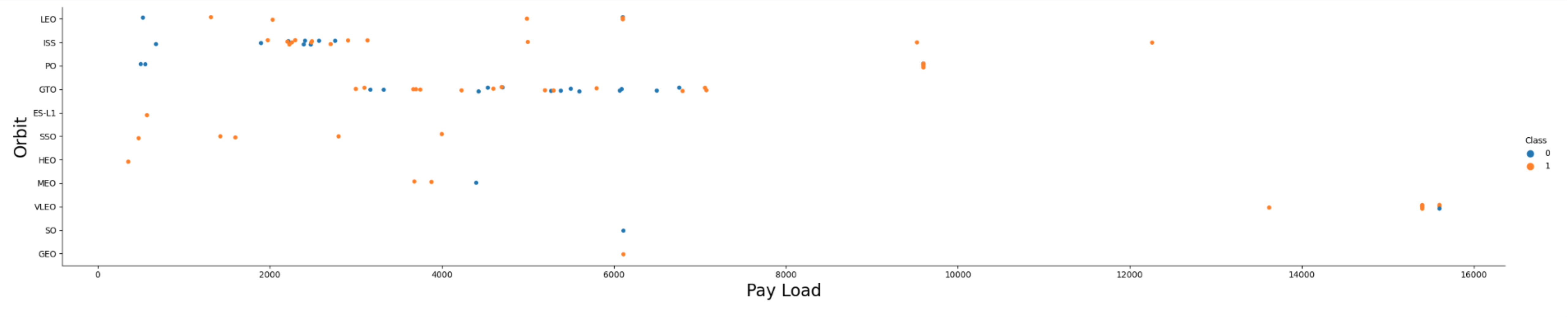


Flight Number vs. Orbit Type



- Scatterplot indicates that the types of flights changed over time
- Early flights tended to be LEO, ISS, PO and GTO
- Later flights tended to be LEO, GTO, MEO, VLEO
- The later flights, particularly VLEO, showed good success rates

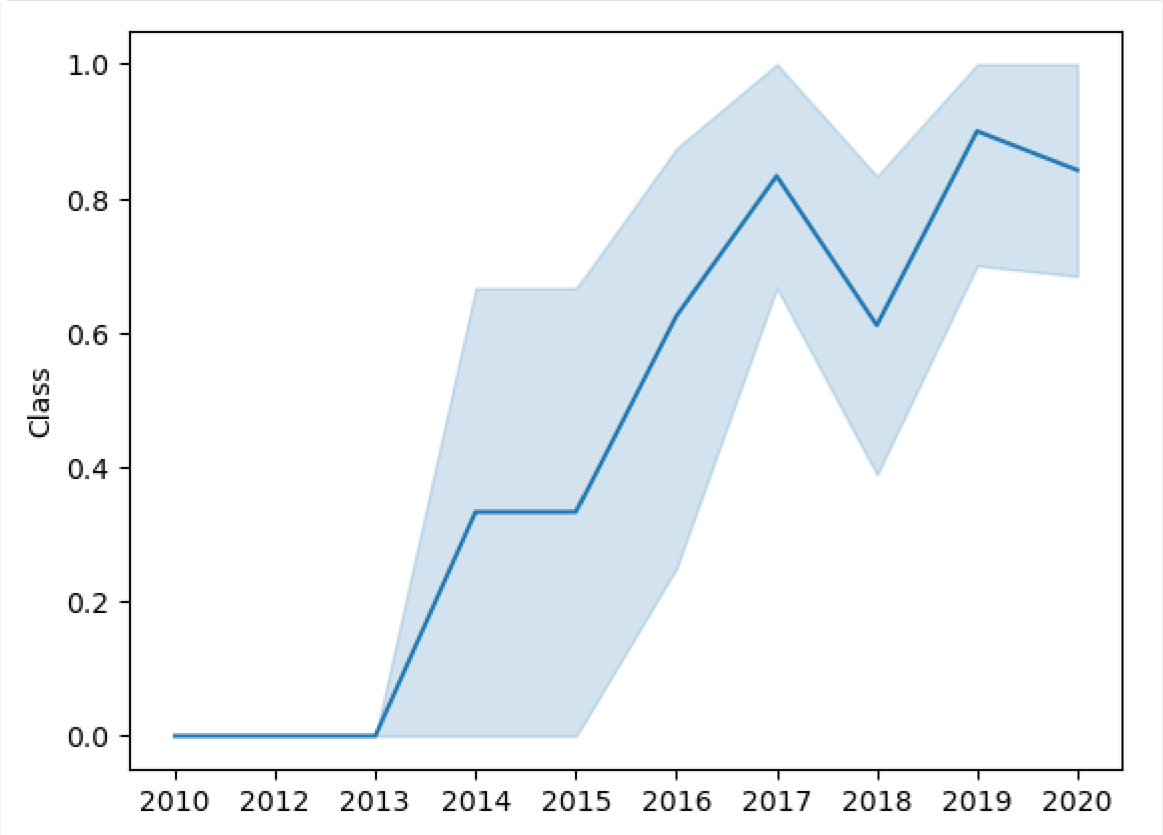
Payload vs. Orbit Type



- Scatterplot shows patterns of behaviour, including:
  - Low payloads on ISS orbits
  - Medium payloads on GTO orbits
  - High payloads on VLEO orbits

Launch Success Yearly Trend

- Line chart indicates significant improvement in success rate over time
- The program started with low success from 2010-2013, but maintained >0.5 success rate from 2016 onward



All Launch Site Names

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

- Each of the distinct launch site names were extracted from the data set
- %sql SELECT DISTINCT Launch_Site FROM SPACEXTBL

Launch Site Names Begin with 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_ MASS__KG_	Orbit	Customer	Mission Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- The five records returned from this query were all from the site CCAFS LC-40
- %sql SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5

Total Payload Mass

SUM(PAYLOAD_MASS__KG_)
45596

- The total payload mass for NASA (CRS) was 45,596kg, calculated by summing the payload mass records for each launch
- %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer = 'NASA (CRS)'

Average Payload Mass by F9 v1.1

AVG(PAYLOAD_MASS__KG_)
2928.4

- The average payload mass for this booster was 2,928.4kg, calculated by performing the AVG function on the payload mass records for each launch
- %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version LIKE 'F9 v1.1'

First Successful Ground Landing Date

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
22-12-2015	01:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)

- The date of the first successful ground landing was 22-12-2015
- This was found by using the MIN function on the date records for the set of launches with successful ground pad landing outcomes
- %sql SELECT MIN(Date) FROM SPACEXTBL WHERE "Landing _Outcome" = 'Success (ground pad)';

Successful Drone Ship Landing with Payload between 4000 and 6000

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

- There are four boosters which have successfully landed on a drone ship with payload mass greater than 4,000 but less than 6,000
- This result was found by filtering the data by those criteria then listing booster names
- %sql SELECT Booster_Version FROM SPACEXTBL WHERE "Landing _Outcome" = "Success (drone ship)" AND "PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MASS__KG_" < 6000

Total Number of Successful and Failure Mission Outcomes

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

- From the provided data, 100 missions were successful and one failed.
- %sql SELECT MISSION_OUTCOME, COUNT(Mission_Outcome) AS COUNT_OUTCOME FROM SPACEXTBL GROUP BY MISSION_OUTCOME

Boosters Carried Maximum Payload

- In total, 12 boosters carried the maximum payload
- This was found by selecting the booster name from records where the payload was equal to the maximum payload across all launches.
- %sql select BOOSTER_VERSION as booster from SPACEXTBL where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTBL)

booster
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

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

- The failed landing outcomes on drone ships, their booster versions, and launch site names in the year 2015 are listed here
- This result was found by selecting record columns after filtering data by landing outcome and the year 2015. Month number was extracted using substr(Date, 4, 2)
- %sql SELECT substr(Date, 4, 2) as Month, "Landing _Outcome", Booster_version, Launch_Site FROM SPACEXTBL WHERE "Landing _Outcome" = 'Failure (drone ship)' AND substr(Date,7,4)='2015'

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

- Landing outcomes between the dates of 2010-06-04 and 2017-03-20 are ranked here in descending order

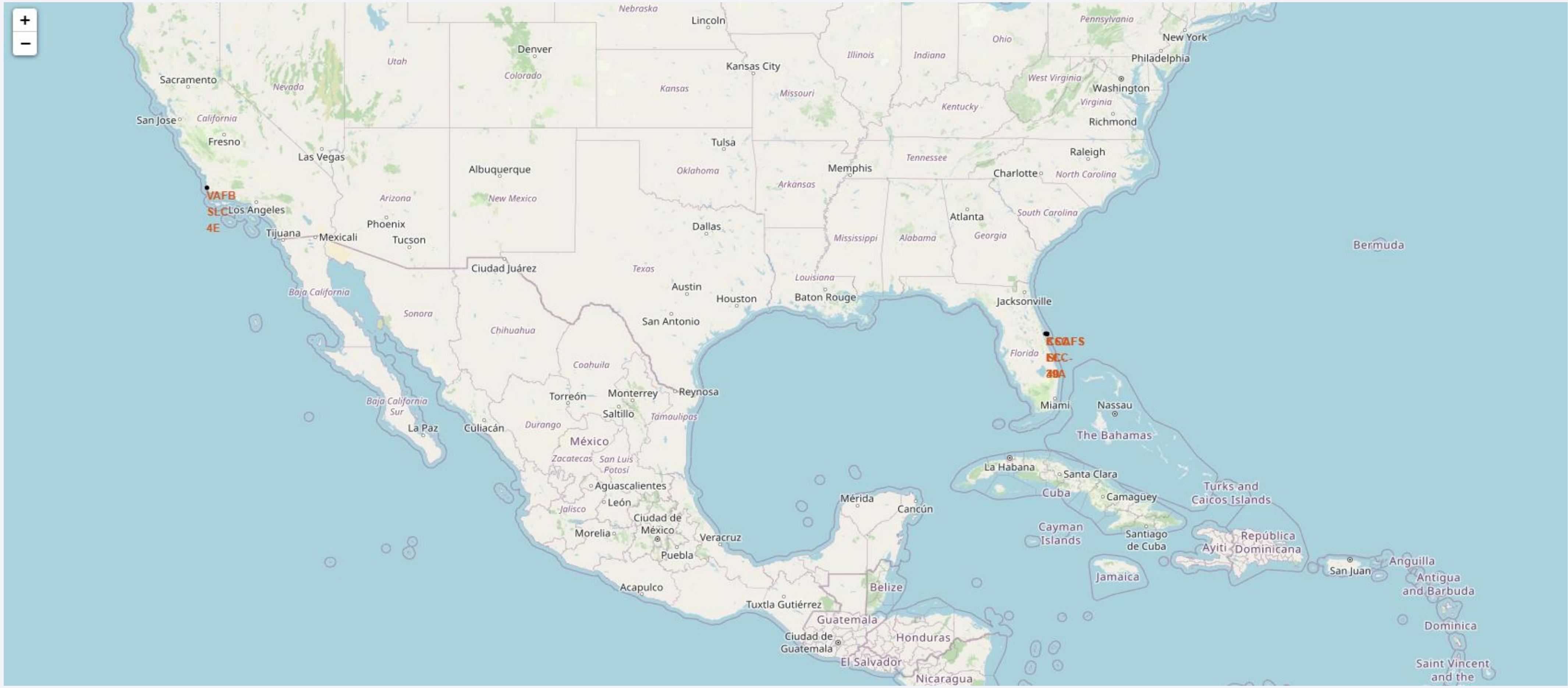
%sql SELECT "LANDING _OUTCOME", COUNT("LANDING _OUTCOME") AS TOTAL FROM SPACEXTBL WHERE DATE BETWEEN '04-06-2010' AND '20-03-2017' GROUP BY "LANDING _OUTCOME" ORDER BY TOTAL DESC

Landing _Outcome	TOTAL
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

Section 3

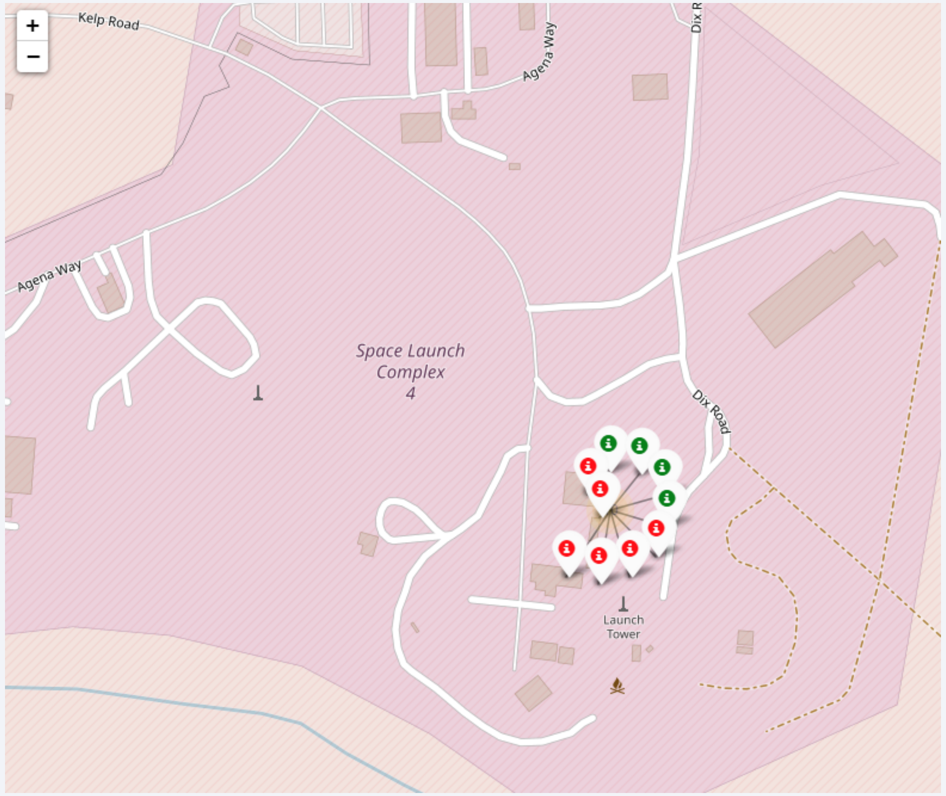
Launch Sites Proximities Analysis

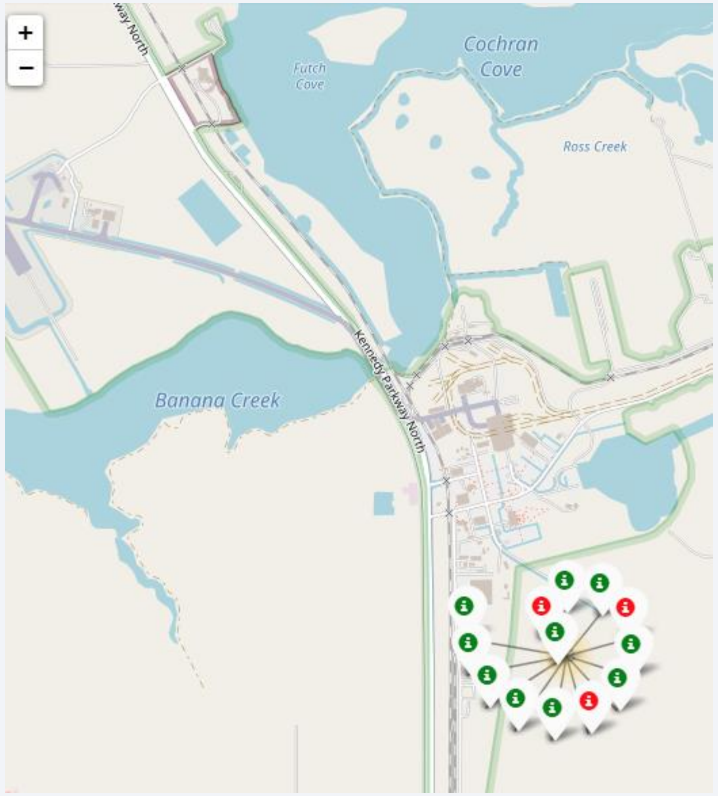
Launch Sites Proximity Analysis – all site locations

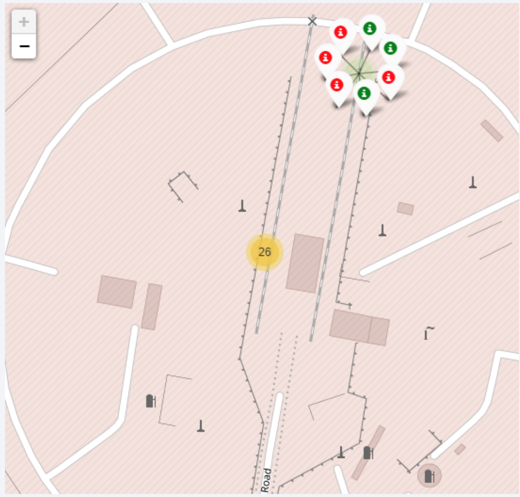


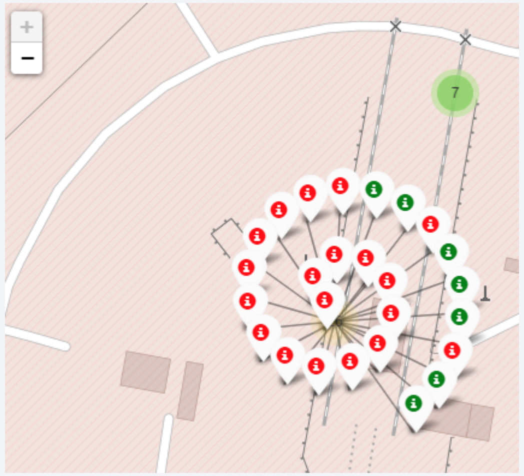
- Each of the launch sites has been labelled and on review, we can see that the launch sites are situated in coastal regions of the USA.

Launch Sites Proximity Analysis – launch outcomes



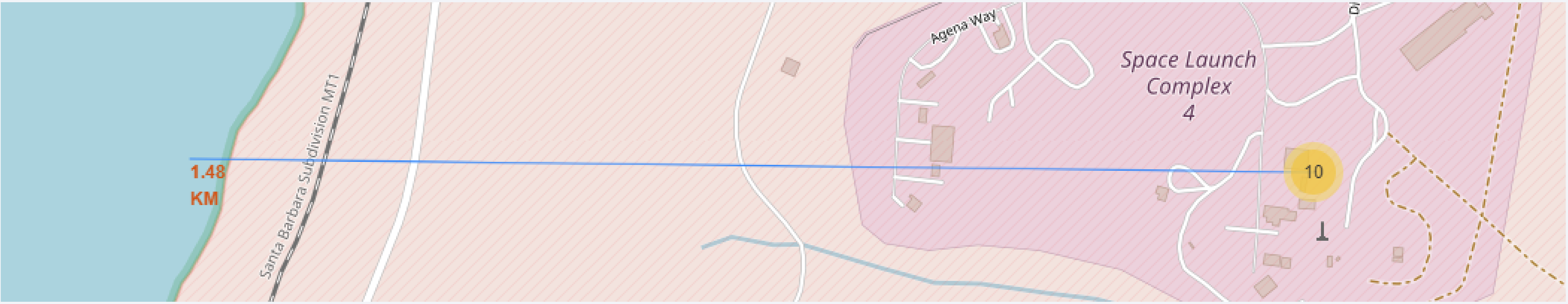




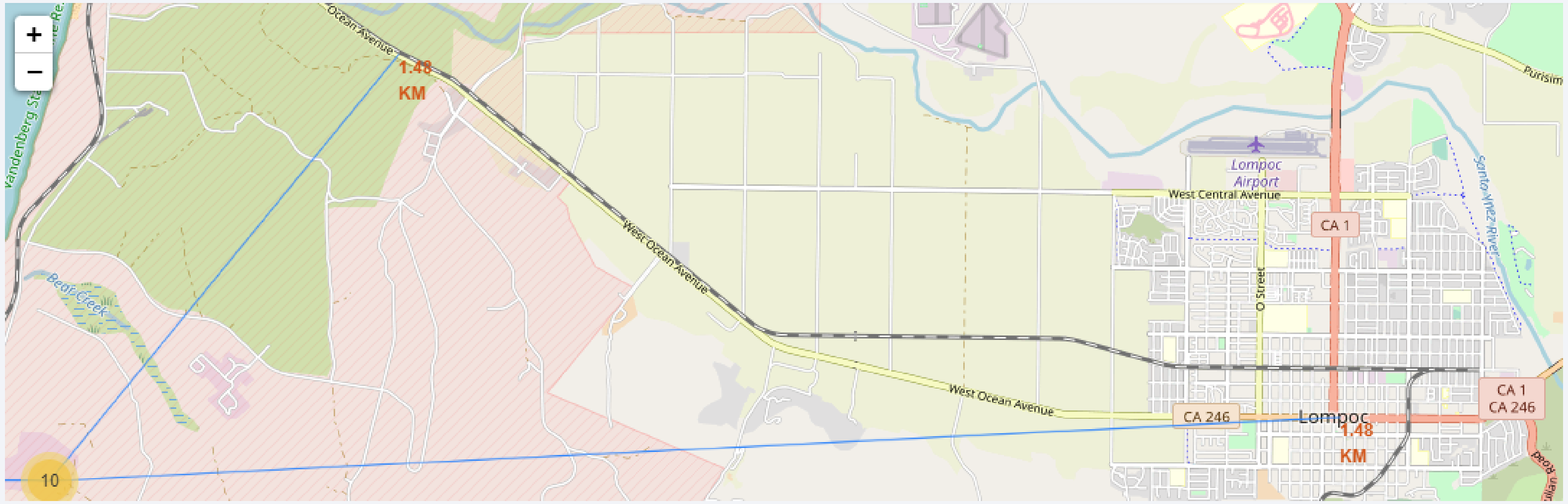


- Coloured indicators in these maps show successful (green) and failed (red) launches. An idea of relative success per site can be obtained quickly by the colour balance.

Launch Sites Proximity Analysis – proximity analysis



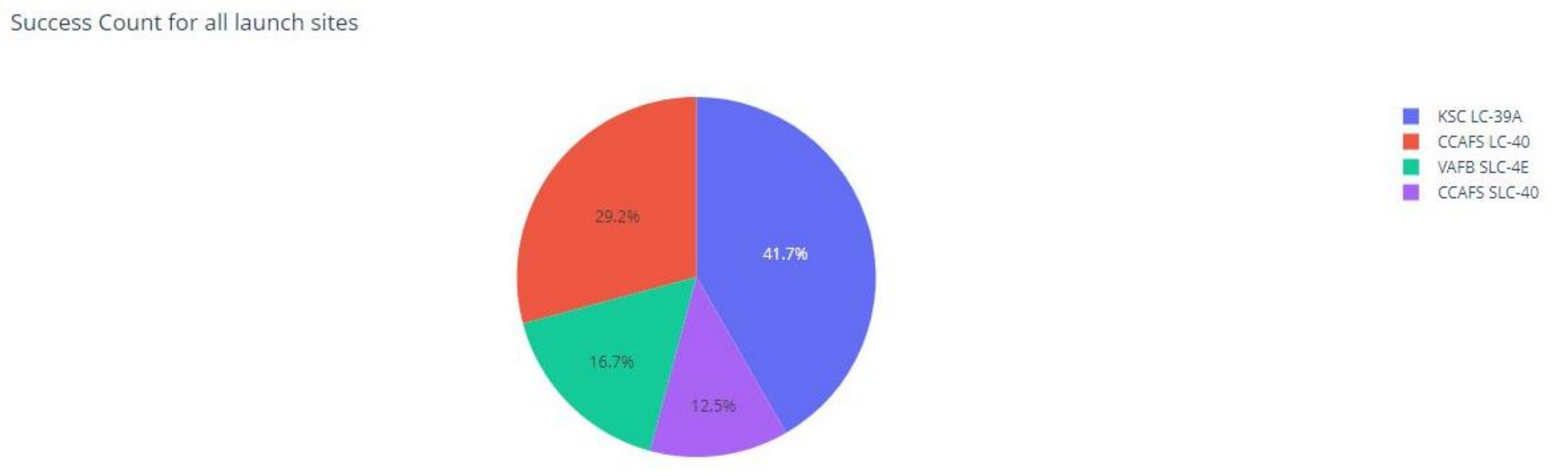
- Mapping distances to major landmarks (coast, rail, highway, city) shows launch sites are situated significant distances away



Section 4

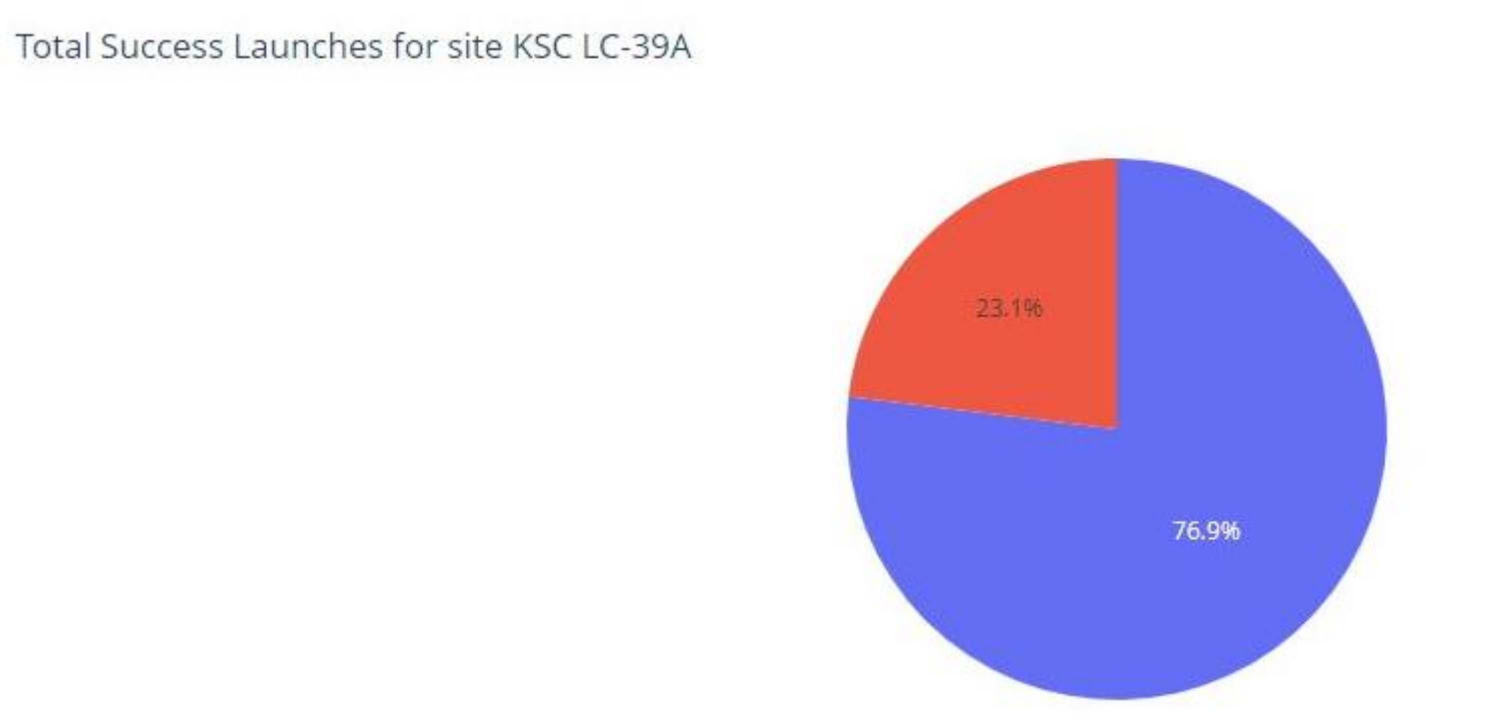
Build a Dashboard with Plotly Dash

Success count for all launch sites



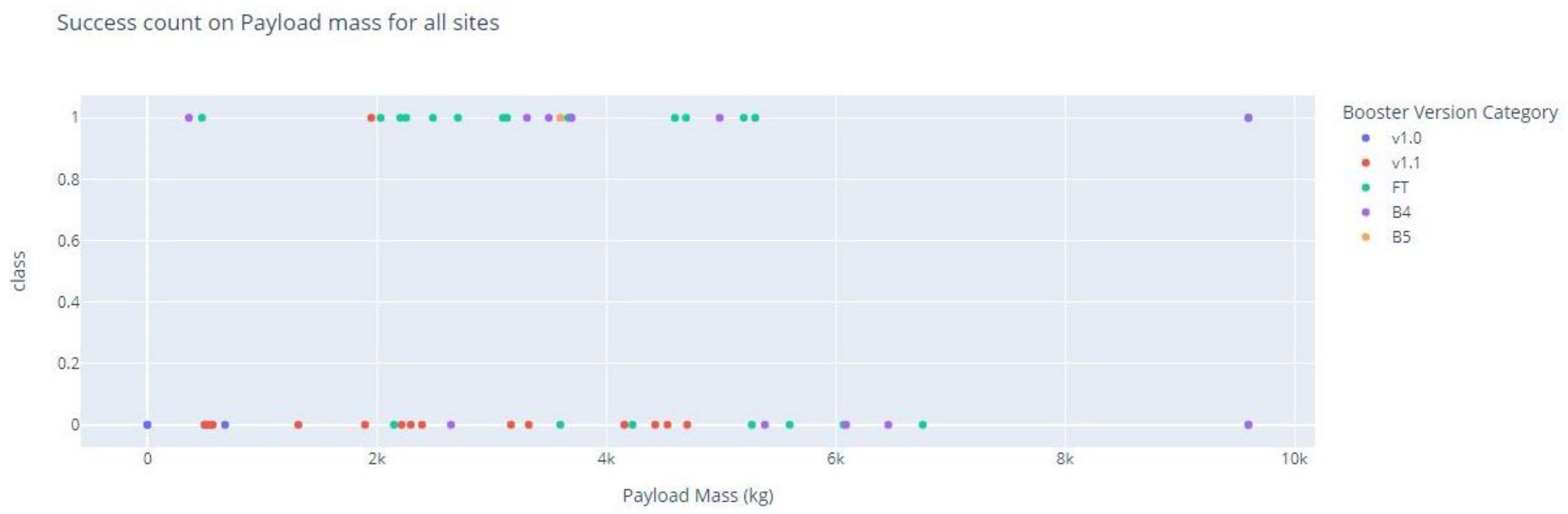
- Notably, KSC LC-39A had the highest success count, with 41.7% of successful launches.
- CCAFS SLC-40 had the lowest success count, with only 12.5%.

Launch success ratio, highest-ratio site



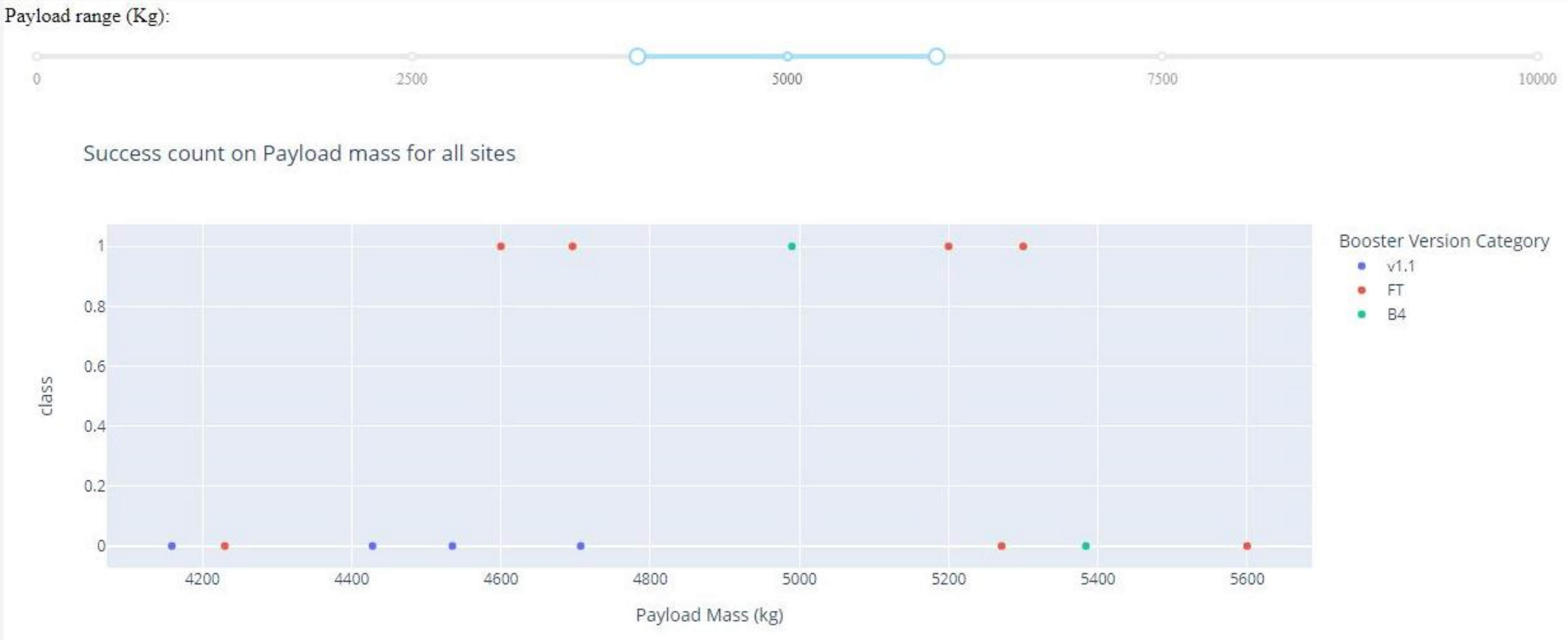
- KSC LC-39A showed the highest ratio of success to failure, with a 76.9% success rate

Payload vs. Launch Outcome – all launches



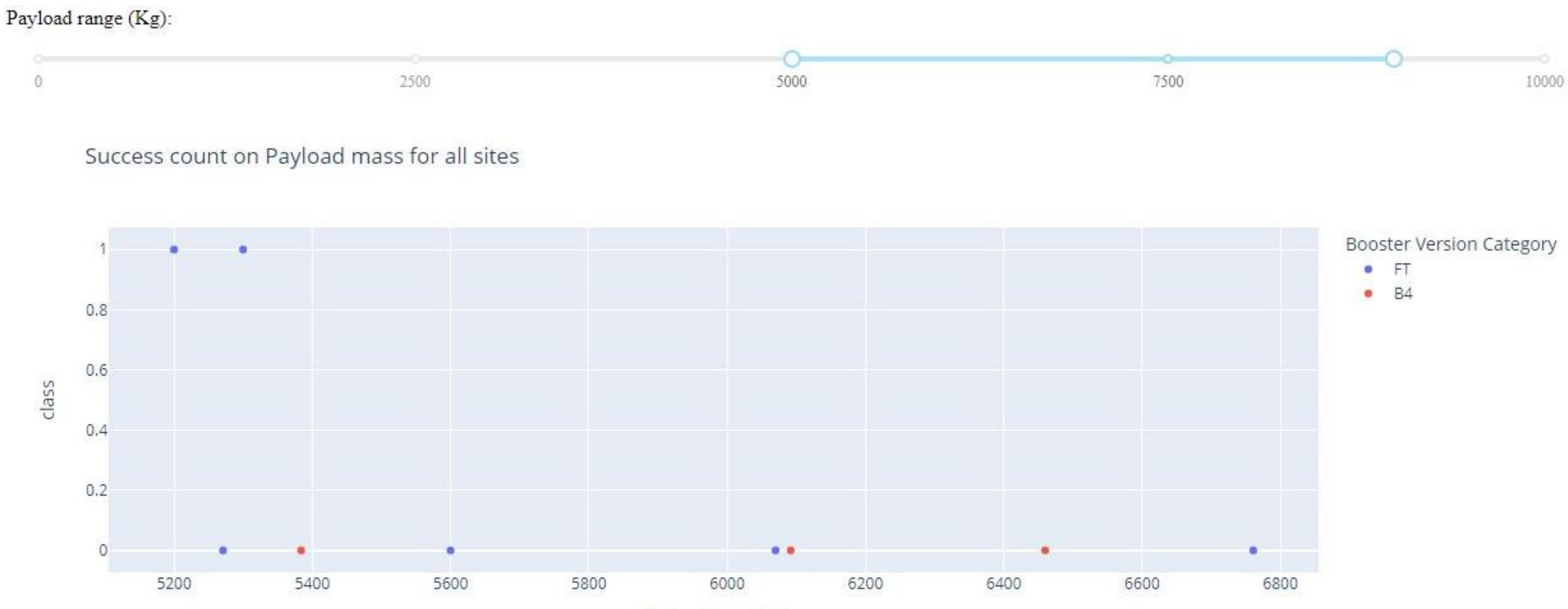
- All launches are shown in this plot, illustrating the trend for more failures at the top-end of the payload mass index.

Payload vs. Launch Outcome – highest success



- Payload range 4,600 – 5,400kg shows the highest proportion of success to failure across all booster types.

Payload vs. Launch Outcome – lowest success



- Conversely, payload range 5,300 – 6,800kg shows as the worst performing range across all booster types.

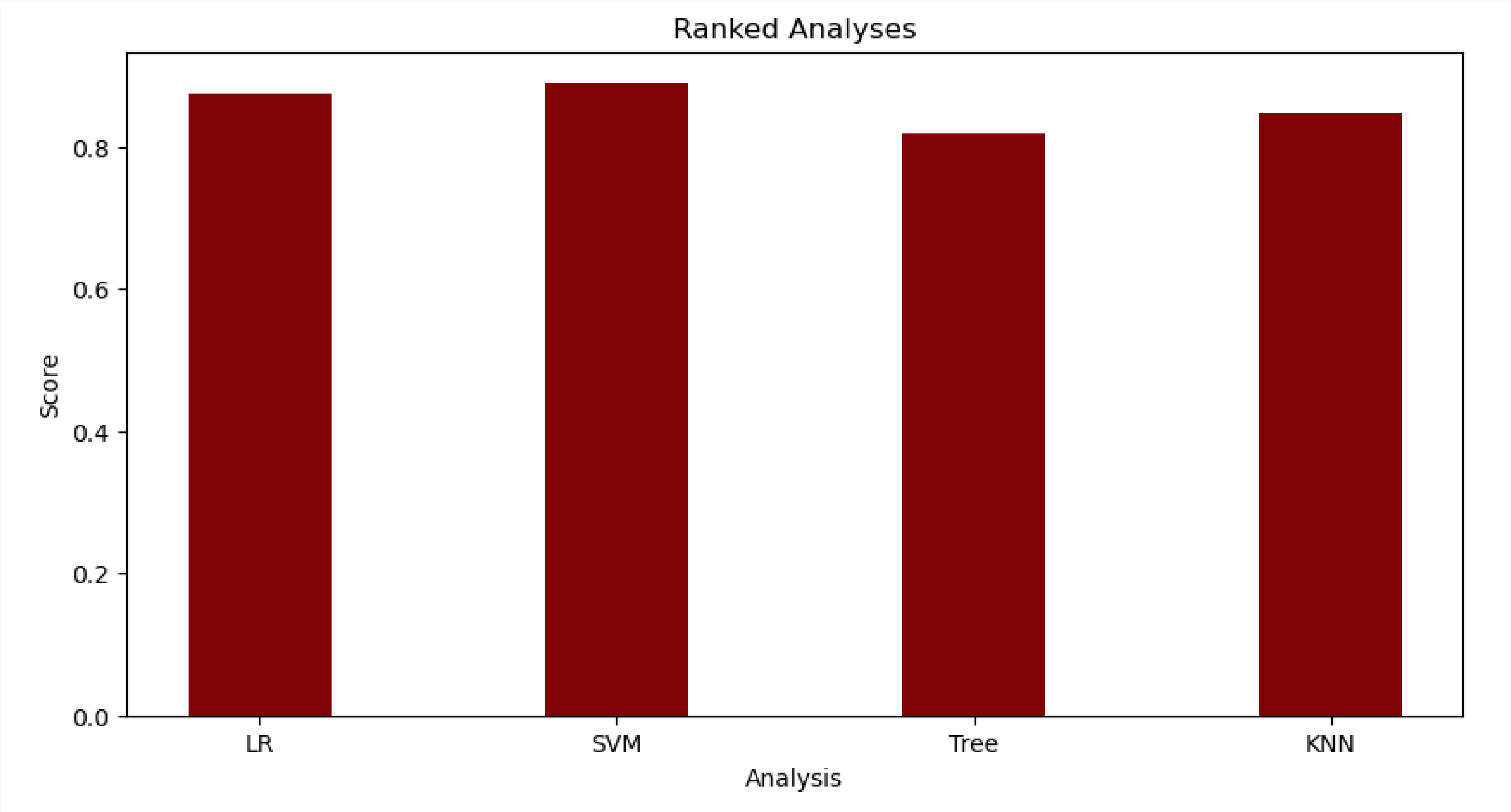
Payload vs. Launch Outcome – booster performance

- Booster B5 is the best performing booster, with 100% success rate from one launch.
- The low sample size should be taken into consideration.

Section 5

Predictive Analysis (Classification)

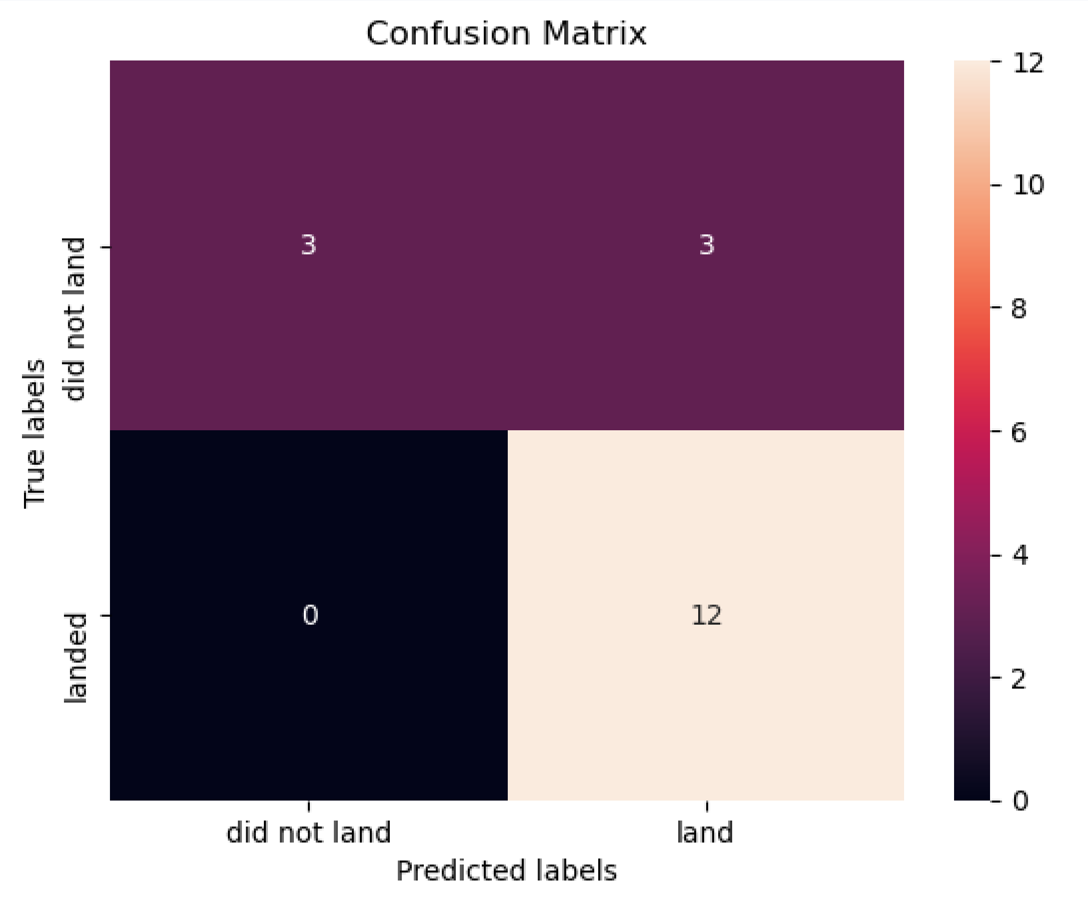
Classification Accuracy




```
lr accuracy is 0.875
svm accuracy is 0.8888888888888888
tree accuracy is 0.8194444444444444
knn accuracy is 0.8472222222222222
```

- Support Vector Machine analysis showed the highest accuracy

Confusion Matrix



- Confusion Matrix shows a high degree of accuracy for failures and a good overall ability to predict success; however, there were a modest number of false positives.

Conclusions

- CCAFS SLC-40 was the most commonly-used launch site while KSC LC-39A was the most successful
- Later launches had a higher success rate, indication improvement and refinement over time
- A payload range of 4,600 – 5,400kg tended to have a better success rate
- ES-L1, GEO, HEO, and SSO orbits all featured a high success rate
- Predictive analysis results indicated that a SVM model would have the best predictive accuracy for future launches
- However, the potential for false positive predictions when using this model should be considered

Appendix

- All notebook and related files can be found on GitHub at:

https://github.com/simongoudie/Applied-Data-Science-Capstone

Thank you!