Winning Space Race with Data Science

<Name>
<Date>

Outline

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

Executive Summary

- Methodology:
  - Data collection was performed through web scraping and SpaceX's API.
  - The collected data was cleaned, explored and preprocessed before applying ML predictions.
- Summary of results:
  - SpaceX had a successful landing rate of 66.6% between 2010-2020
  - SpaceX's landings improved over the years with 2019 being the most successful with over 80% success.
  - Visualizing launch site clusters, with the most favorable outcomes at KSC LC-39A with a 42% success rate.
  - Out of 4 ML models, the decision tree performs best at predicting outcomes with an accuracy of 89%

Introduction

- Purpose Statement:
  - Evaluating the viability of competing with SpaceX as a startup company.
- Problem Statement:
  - Identifying key metrics from competitors, which provide a good indication of a successful landing.
  - Pin-pointing suitable locations for executing launches.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Utilizing SpaceX's public API
  - Web Scraping from Wikipedia
- Perform data wrangling:
  - The data was cleaned by resolving empty values with the mean, adjusting data types such as dates, and structuring the data in a useable manner.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Built, tuned and evaluate classification models

Data Collection

- The primary source of data was collected through the SpaceX API (https://api.spacexdata.com/v4/rockets/)
- SpaceX launches were also scraped through Wikipedia (https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922)

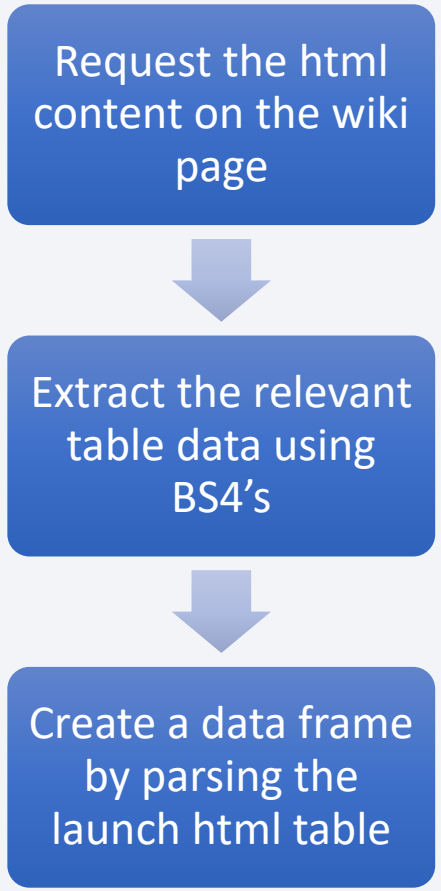
Data Collection – SpaceX API

- SpaceX provides a public API to obtain some data of their launches.
- The flowchart on the right highlights the steps followed after the API call.
- Source code: [IBM-Applied-DS-Capstone/Data Collection API.ipynb at master · EngMarchG/IBM-Applied-DS-Capstone (github.com)]

Data Collection - Scraping

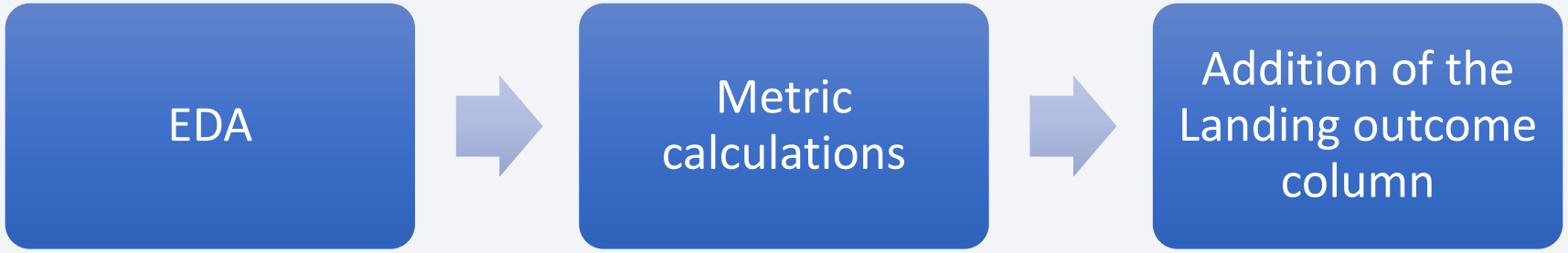
- SpaceX launches were scraped through Wikipedia as well.
- The flowchart on the right illustrates the taken steps.

- Source code: [IBM-Applied-DS-Capstone/Week1 at master · EngMarchG/IBM-Applied-DS-Capstone (github.com)]



Data Wrangling

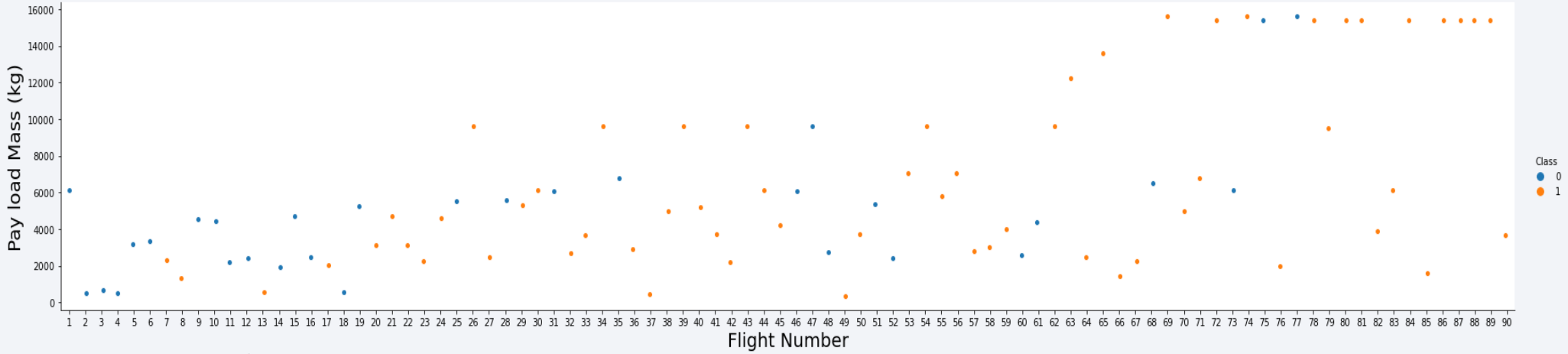
- Exploratory Data Analysis (EDA) was performed on the dataset as a preliminary step.
- Certain metrics were then calculated such as the launches per site, the number of occurrences of each orbit and the occurrences of a mission's outcome per orbit type.
- Lastly, a new column was added to classify the landing outcome.



- Source code: [IBM-Applied-DS-Capstone/Data Wrangling.ipynb at master · EngMarchG/IBM-Applied-DS-Capstone (github.com)]

EDA with Data Visualization

- To further understand the relations and trends between variables, scatterplots and bar plots were utilized to visualize several pairs of features.
- One of these features is shown in the graph below.



- Source code: [IBM-Applied-DS-Capstone/jupyter-labs-eda-dataviz.ipynb at master · EngMarchG/IBM-Applied-DS-Capstone (github.com)]

EDA with SQL

- SQL was used to answer several questions, with the most important being:
  - Top 5 launch sites.
  - Total payload mass carried by boosters launched by NASA.
  - Average payload mass carried by booster version F9 v1.1
  - Total number of both successful and failed mission outcomes.

- Source code: IBM-Applied-DS-Capstone/jupyter-labs-eda-sql-coursera.ipynb at master · EngMarchG/IBM-Applied-DS-Capstone · GitHub

Build an Interactive Map with Folium

- Markers, clusters, circles and lines were used with Folium.
  - Markers highlights launch sites.
  - Circles indicate special areas such as the NASA Johnson Space Center.
  - Clusters indicate groups of events at a location, like launches in a launch site.
  - Lines highlight the distance between two coordinates.

- Source code: IBM-Applied-DS-Capstone/Interactive Visual Analytics with Folium lab.ipynb at master · EngMarchG/IBM-Applied-DS-Capstone · GitHub

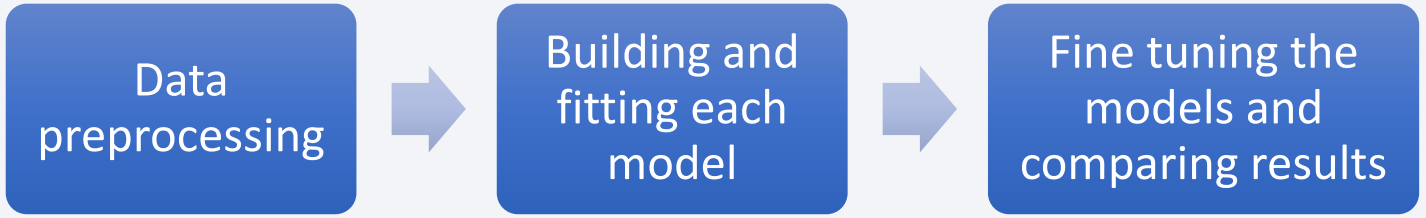
Build a Dashboard with Plotly Dash

- The following graphs and plots were performed using Plotly:
  - Total success launches by site
  - Class vs Pay Load Mass, according to various Booster Version Category
  - Total successful and unsuccessful launches of a site.
- These feature pairs give a quick overview regarding the relation ship between the pay load and launch sites, which in turn identifies the best places to launch.

- Source code: IBM-Applied-DS-Capstone/spacex_plotly.py at master · EngMarchG/IBM-Applied-DS-Capstone · GitHub

Predictive Analysis (Classification)

- Four classification models were built and compared:
  - Logistic regression,
  - Support vector machine (SVM)
  - Decision tree
  - K-nearest neighbors



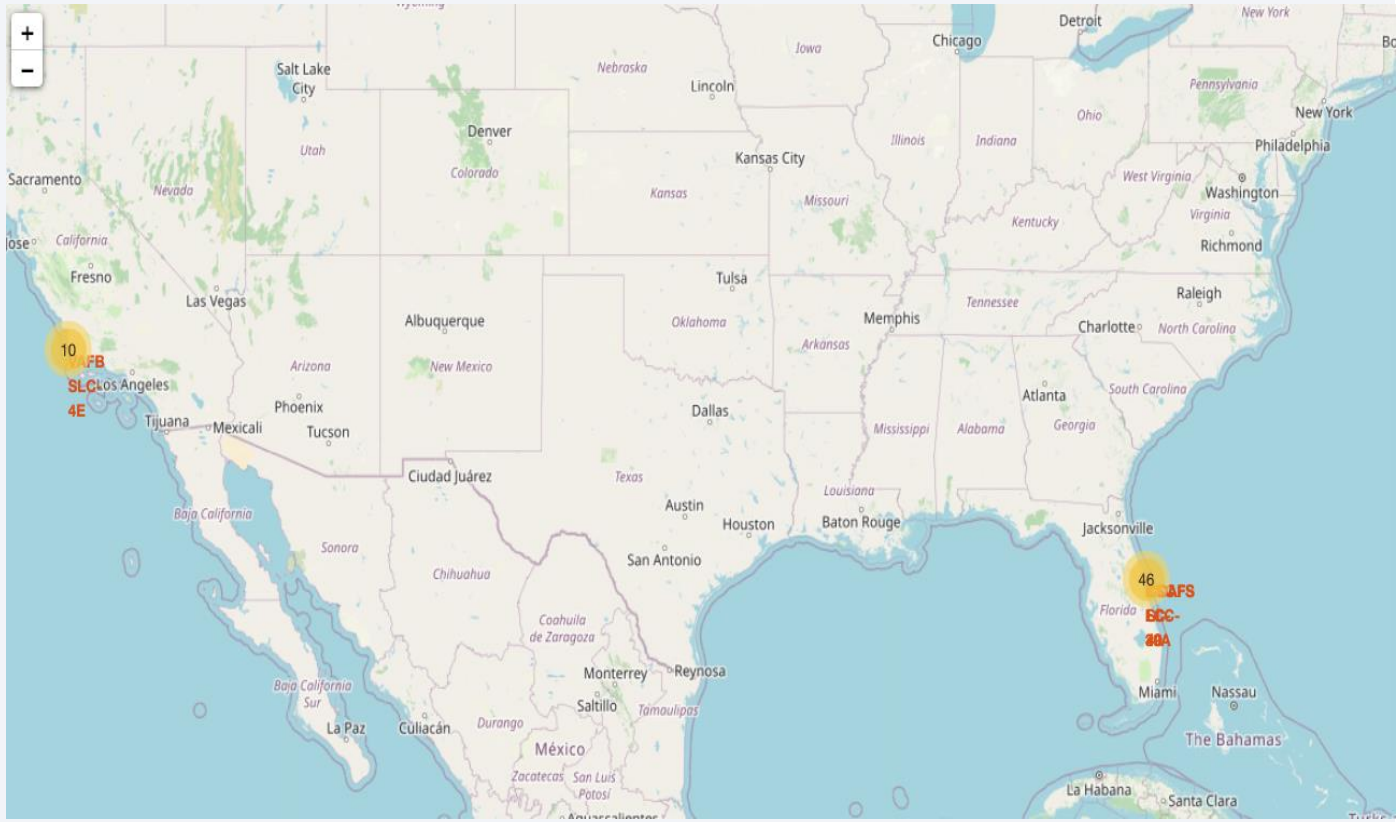
- Source code: IBM-Applied-DS-Capstone/Machine Learning Prediction.ipynb at master · EngMarchG/IBM-Applied-DS-Capstone · GitHub

Results

- Exploratory data analysis results:
  - SpaceX has 4 different launch sites.
  - The average payload of the F9 v1.1 booster is 2,928 kg.
  - Several Falcon 9 booster versions were successful at landing in drone ships, while having an above average payload.
  - Mission outcomes were overwhelmingly successful at nearly 100%
  - The number of landing outcomes kept improving throughout the years.

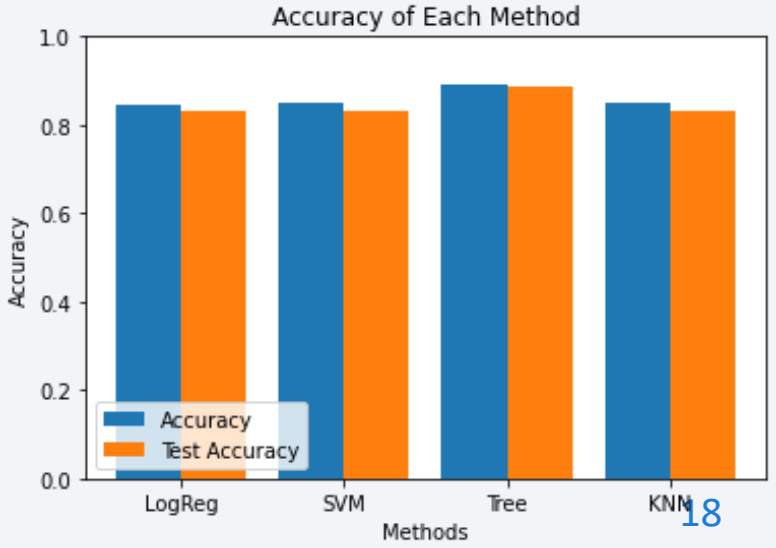
Results

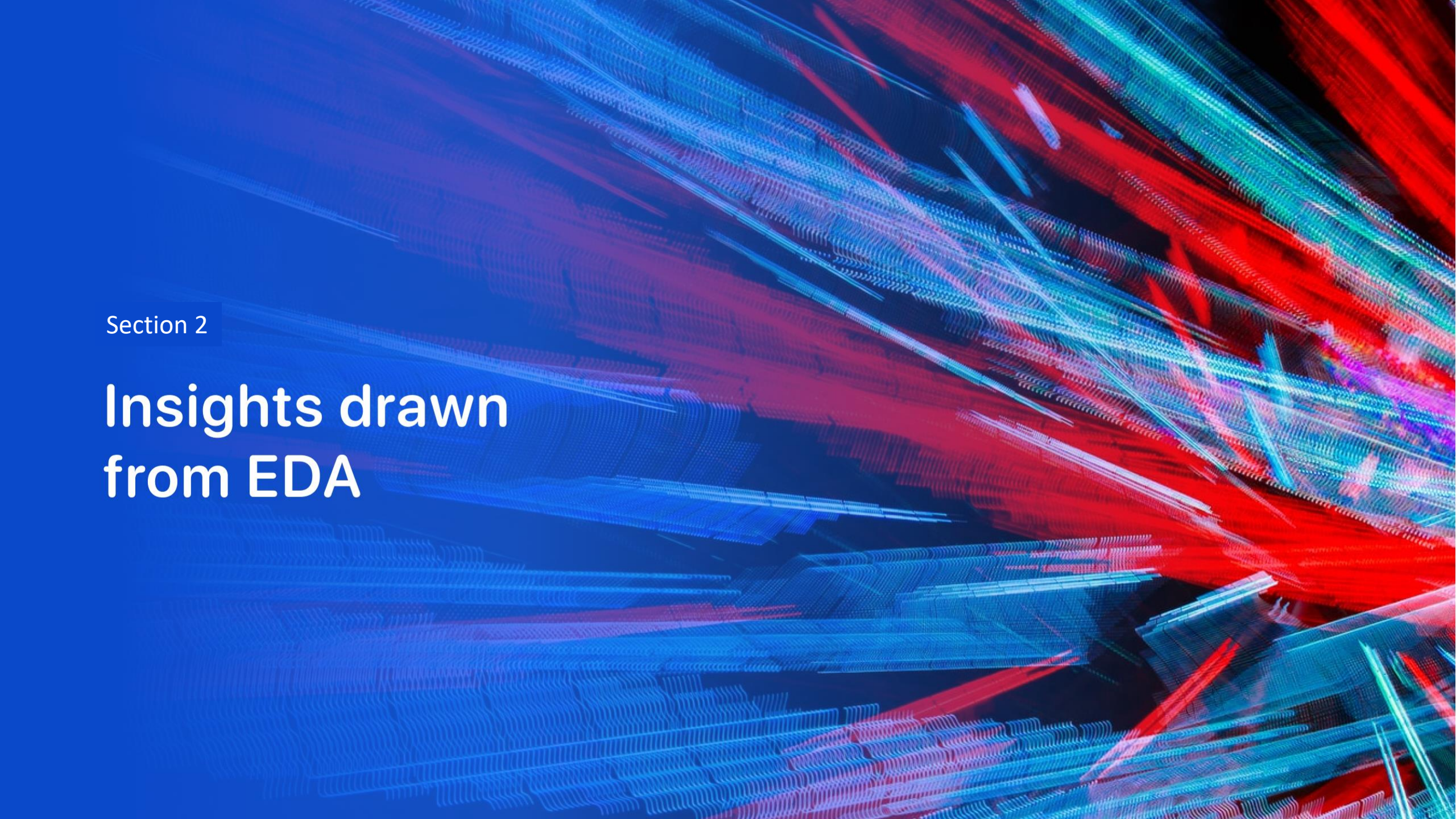
- According to Folium's interactive map, most launches occur near coasts.



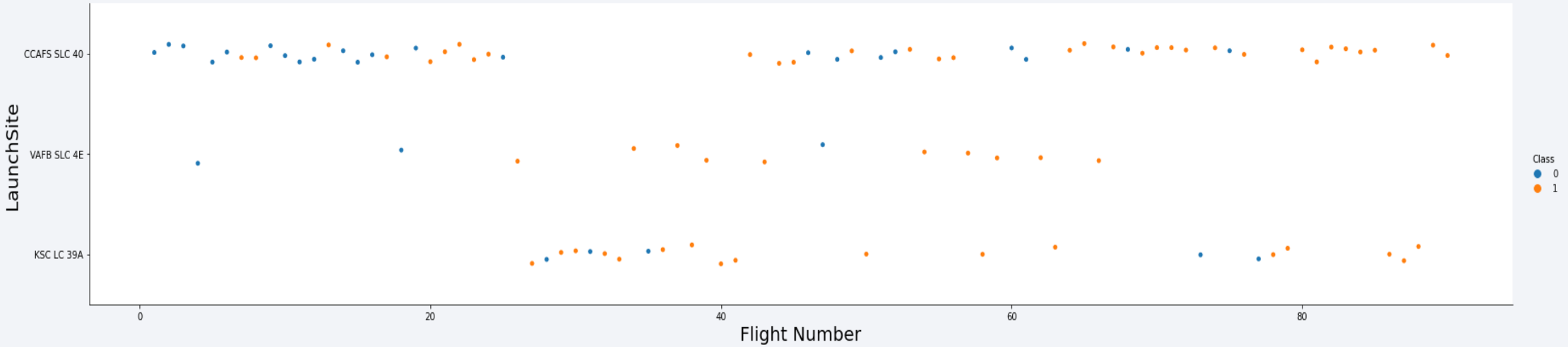
Results

- Out of all the machine learning models, the Decision Tree outperforms every other model at predicting successful landings.
- It boasts an accuracy of 89% for both the test and training datasets.



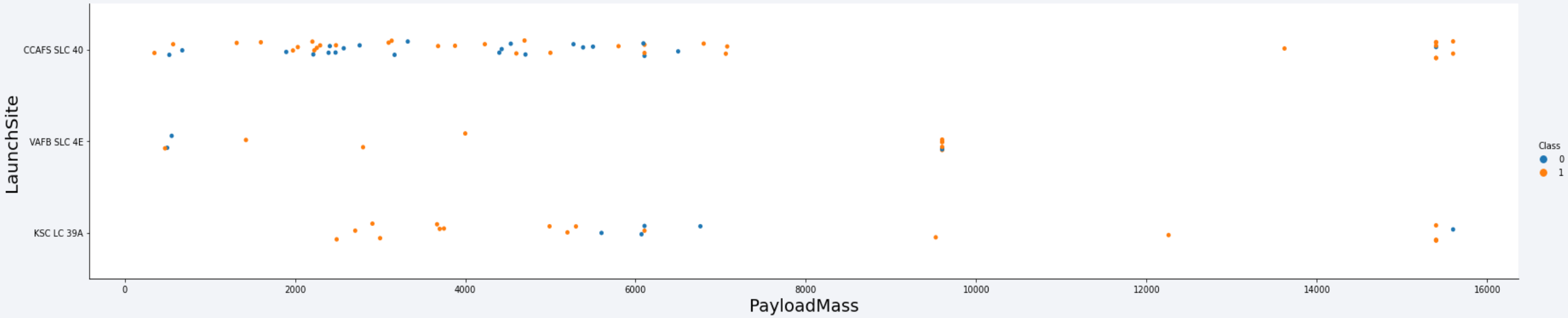

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- The plot above illustrates that the CCAF5 SLC40 launch site appears to be the most successful launch. This is followed by VAFB SLC 4E and KSC LC 39A respectively.
- Additionally, it is clear that the overall success rate continues to improve over time.

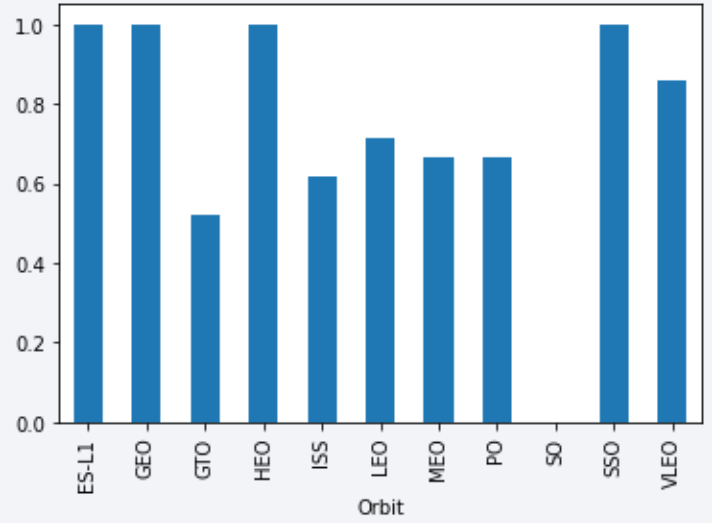
Payload vs. Launch Site



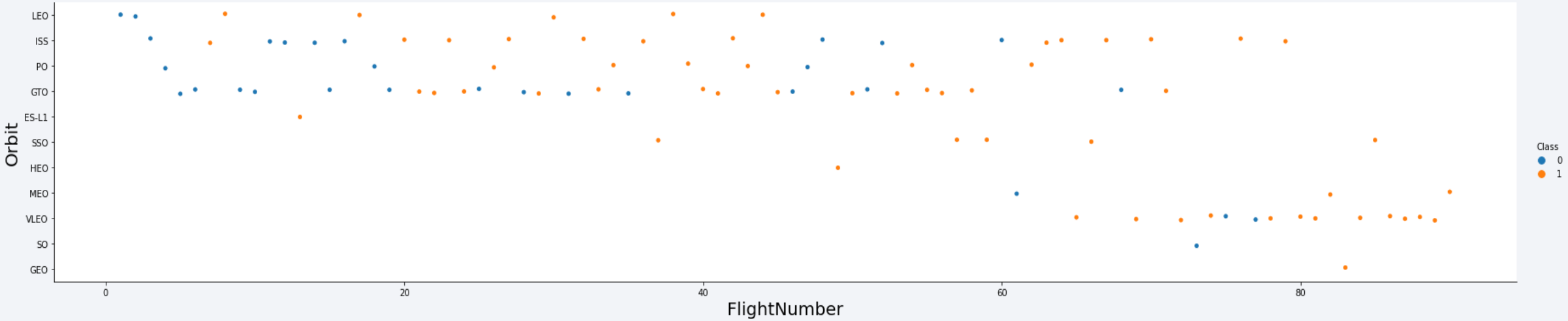
- Payloads over 9,000 kg have an outstanding success rate.
- Payloads over 12,000 kg appear to only be successful at CCAFS SLC 40 and KSC LC 39A launch sites.

Success Rate vs. Orbit Type

- The following orbits appear to have the highest success rates:
  - ES-L1
  - GEO
  - HEO
  - SSO
- Other orbits significantly lower, with the exception of VLEO.

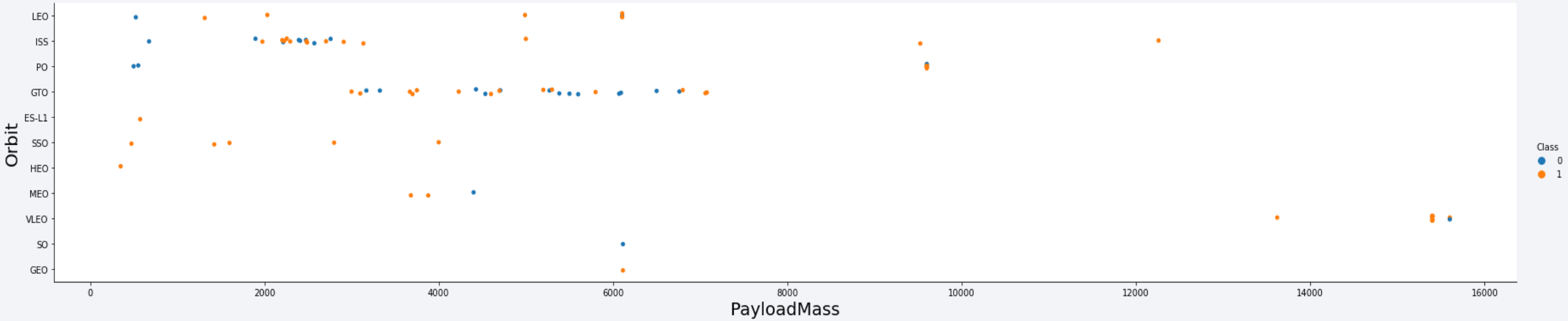


Flight Number vs. Orbit Type



- The success rate of every orbit seem to have improved over the years.

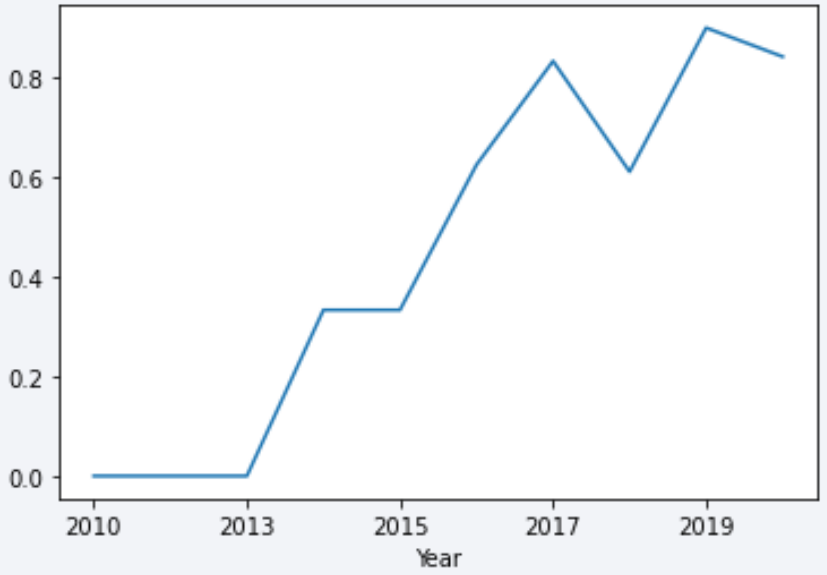
Payload vs. Orbit Type



- No definitive trend can be inferred between these two variables.
- The ISS orbit has largest variance of payload and a good success rate.
- The GTO orbit is the most frequent orbit.

Launch Success Yearly Trend

- The success rate appears to increase throughout 2013 till 2020
- The years 2013-2017 show the most dramatic positive changes.



All Launch Site Names

- According to the gathered data, the four launch sites are as follows:

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

- %sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;
  - Filter the launch sites by distinct names

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin 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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

- %sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
  - The query is used to highlight 5 samples of Cape Canaveral launches

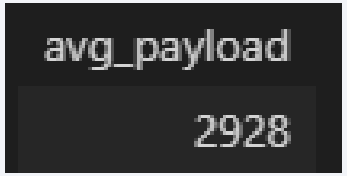
Total Payload Mass

- Calculating the total payload mass carried by boosters from NASA:

- %sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE PAYLOAD LIKE '%CRS%';
  - The total payload found above is achieved by summing up al the payloads, with codes containing “CRS”

Average Payload Mass by F9 v1.1

- Calculating the average payload mass carried by booster version F9 v1.1:



- %sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
  - The data was grouped according to the booster version, the average payload mass is then calculated and found to be 2,928 kg.

First Successful Ground Landing Date

- First successful ground pad landing date:

- %sql SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Success (ground pad)';
  - Data is grouped according to the landing outcome on the ground pad. The minimum value for data is then taken, which in turn corresponds to the first occurrence.

Successful Drone Ship Landing with Payload between 4000 and 6000

- Boosters with a payload greater than 4000 and less than 6000, which have landed successfully on a drone ship are:

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

- %sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND LANDING__OUTCOME = 'Success (drone ship)';
  - Filter the booster versions to distinct ones and add the payload mass restraints.

Total Number of Successful and Failure Mission Outcomes

- Number of both successful and failed mission outcomes:

mission_outcome	qty
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- %sql SELECT MISSION_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL GROUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME;
  - Group the mission outcomes and count the occurrences of each group.

Boosters Carried Maximum Payload

- Boosters that carried the maximum payload mass

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

- %sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL) ORDER BY BOOSTER_VERSION;

2015 Launch Records

- Failed landing outcomes on a drone ship, their booster versions and the launch site name, during the year 2015:

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

- %sql SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Failure (drone ship)' AND DATE_PART('YEAR', DATE) = 2015;

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

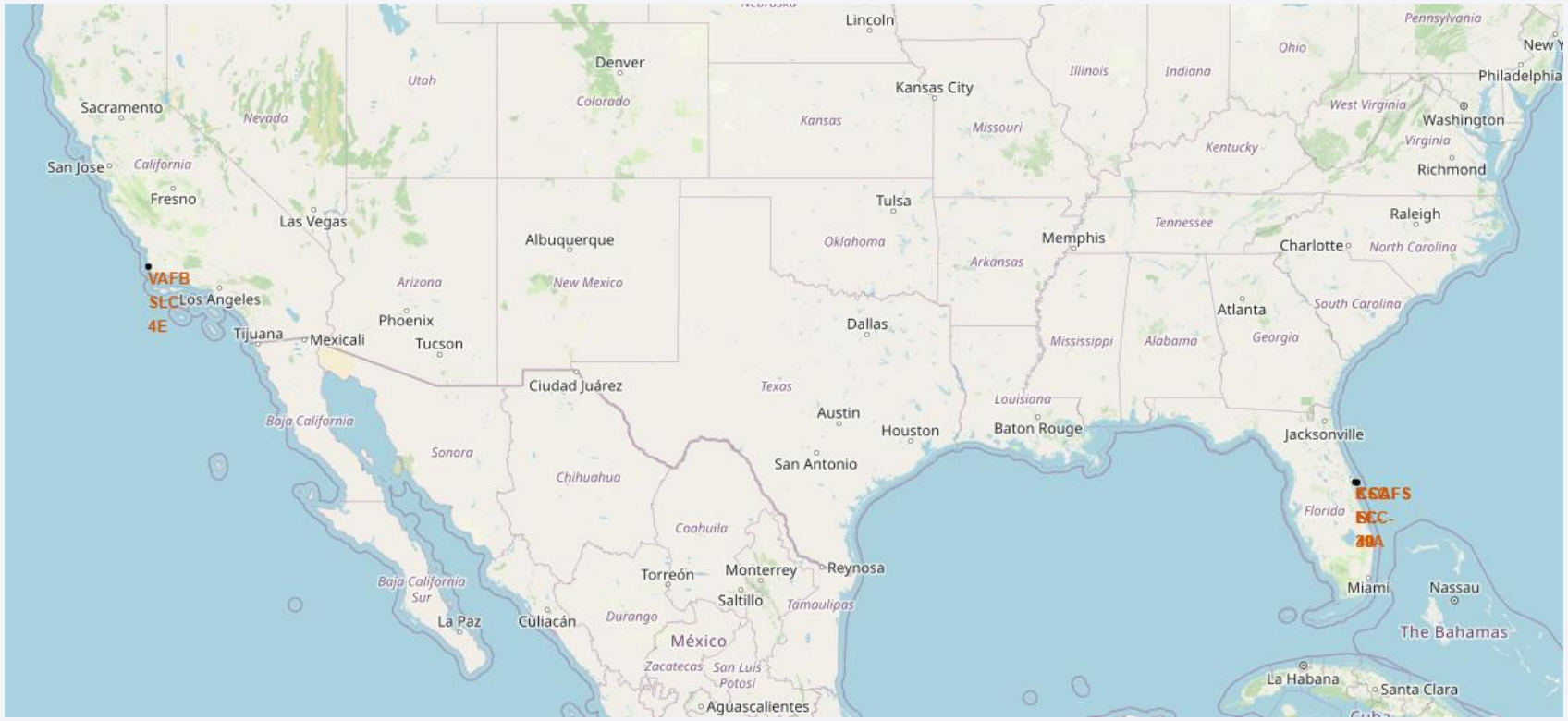
- Ranking of all the landing outcomes between the dates 2010-06-04 and 2017-03-20:

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

- %sql SELECT LANDING__OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING__OUTCOME ORDER BY QTY DESC;
  - The summary table clarifies the importance of considering a 3rd landing outcome, which is “No attempt” as it appears to be the most frequent occurrence.

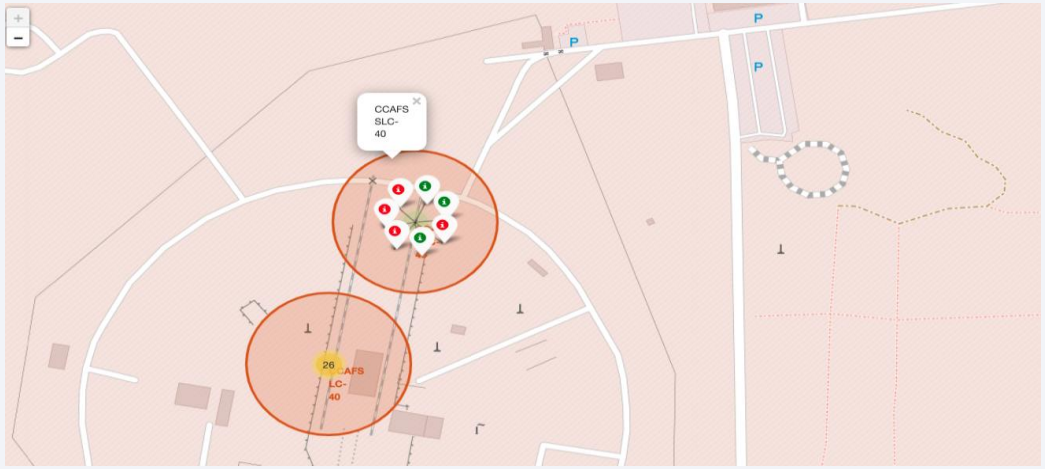

Section 3
Launch Sites
Proximities Analysis

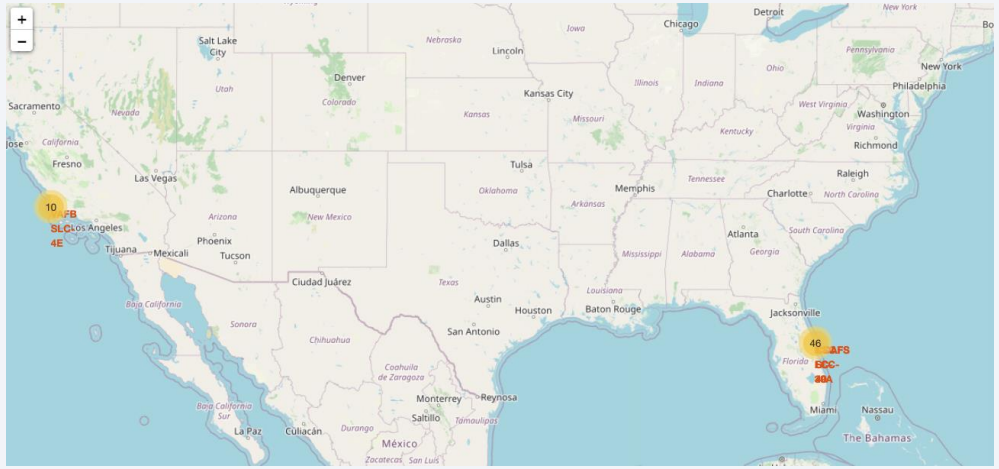
All Launching Sites



- Launch sites are clearly positioned closer to the sea, away from civilization.

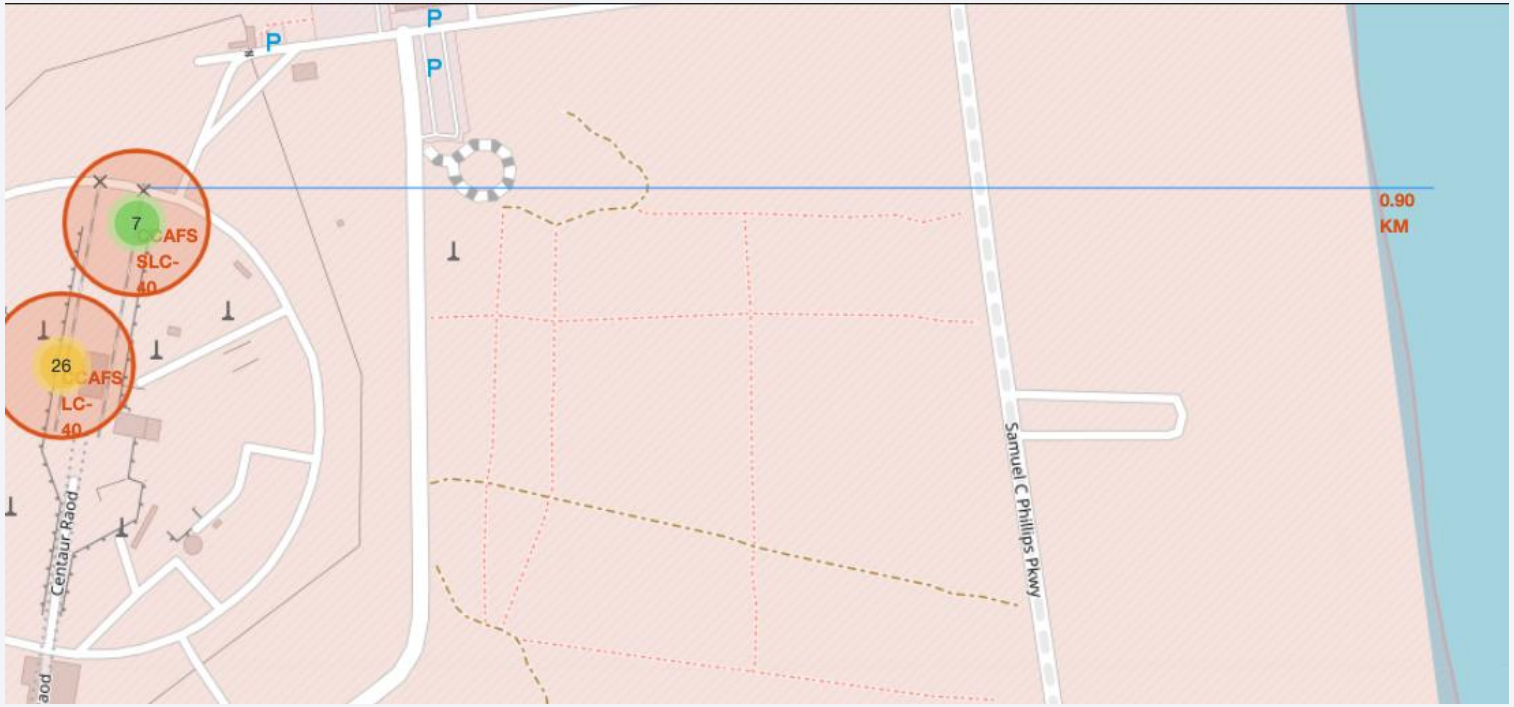
Launch Outcomes by Site



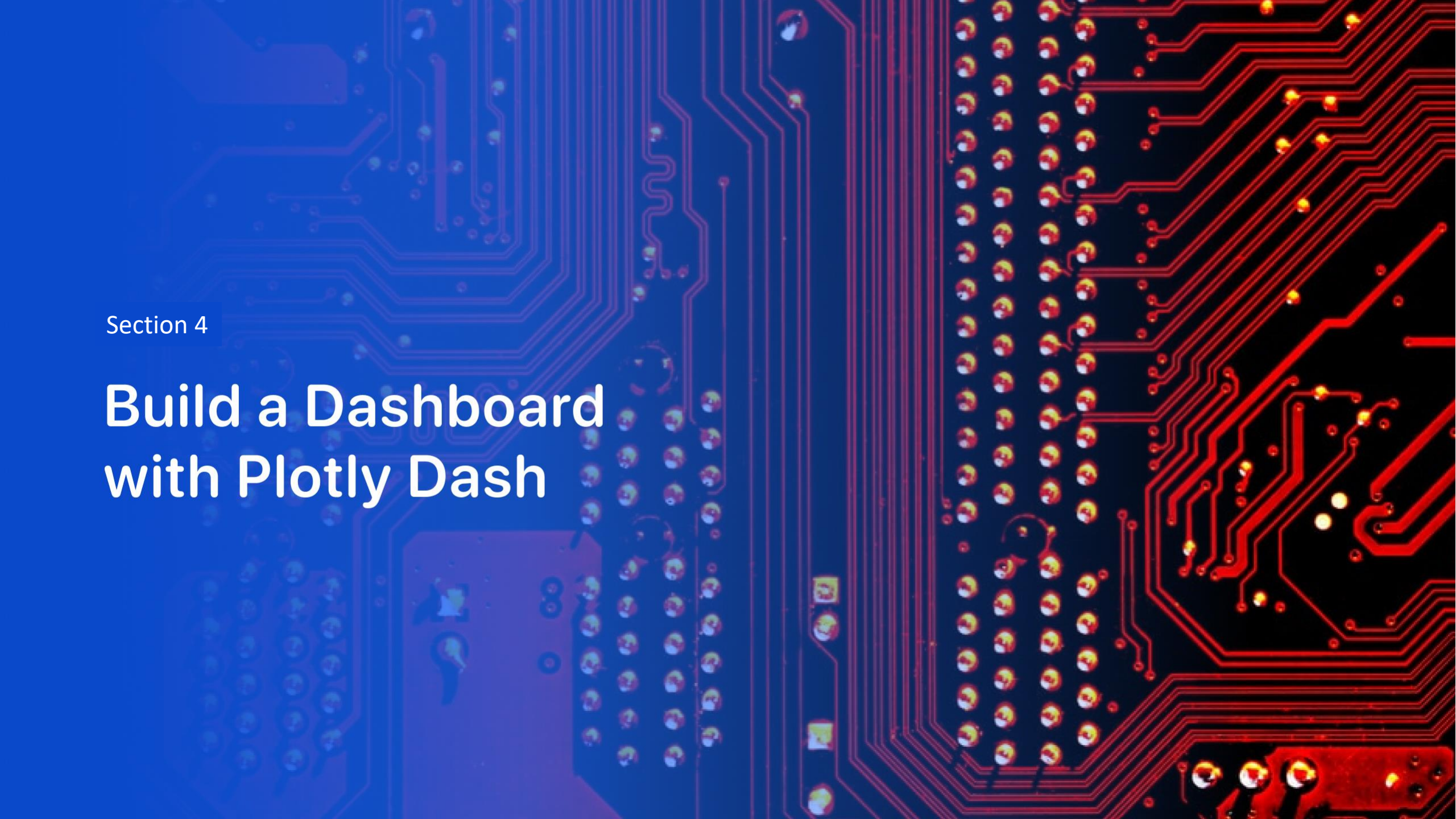


- Example of KSC LC-39A launch site outcomes.
- The green marker indicates a successful outcome, while the red indicates a failure.
- There are 2 main clusters.

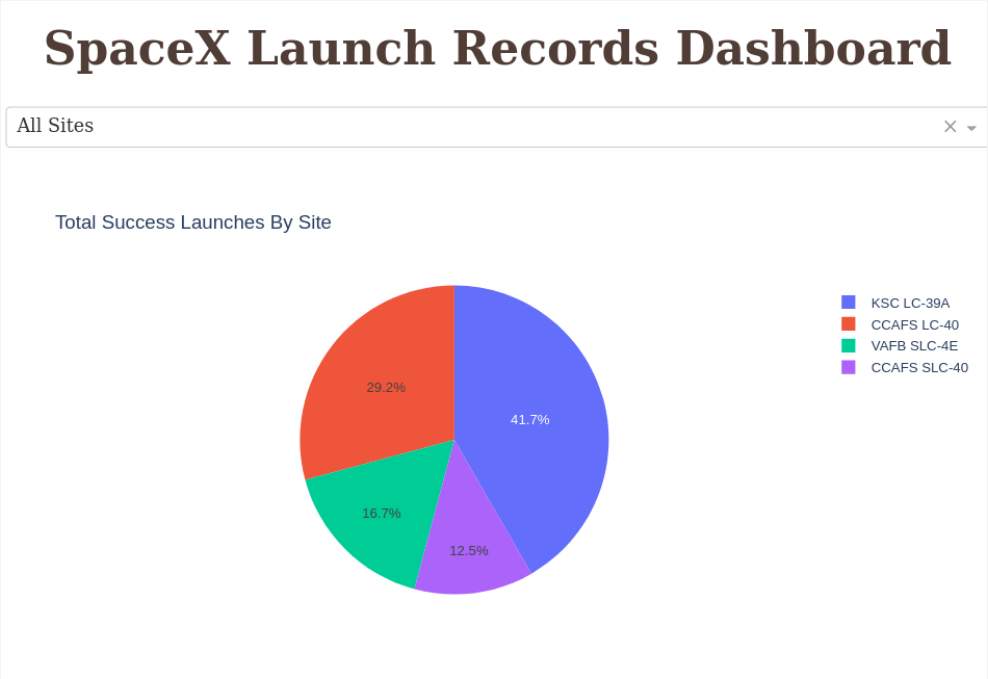
Safety and Logistics



- Launch site KSC LC-39A has good logistics, considering its proximity to a railroad and far from inhabited areas.

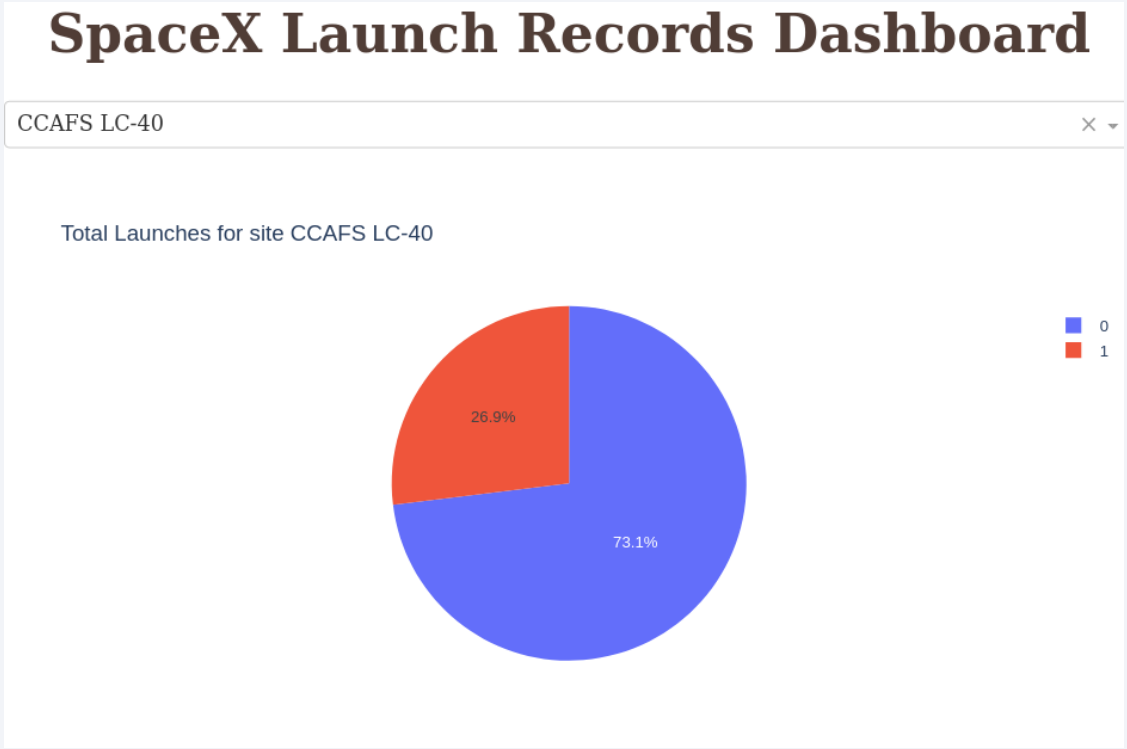

Section 4
Build a Dashboard with Plotly Dash

Successful Launches by Site



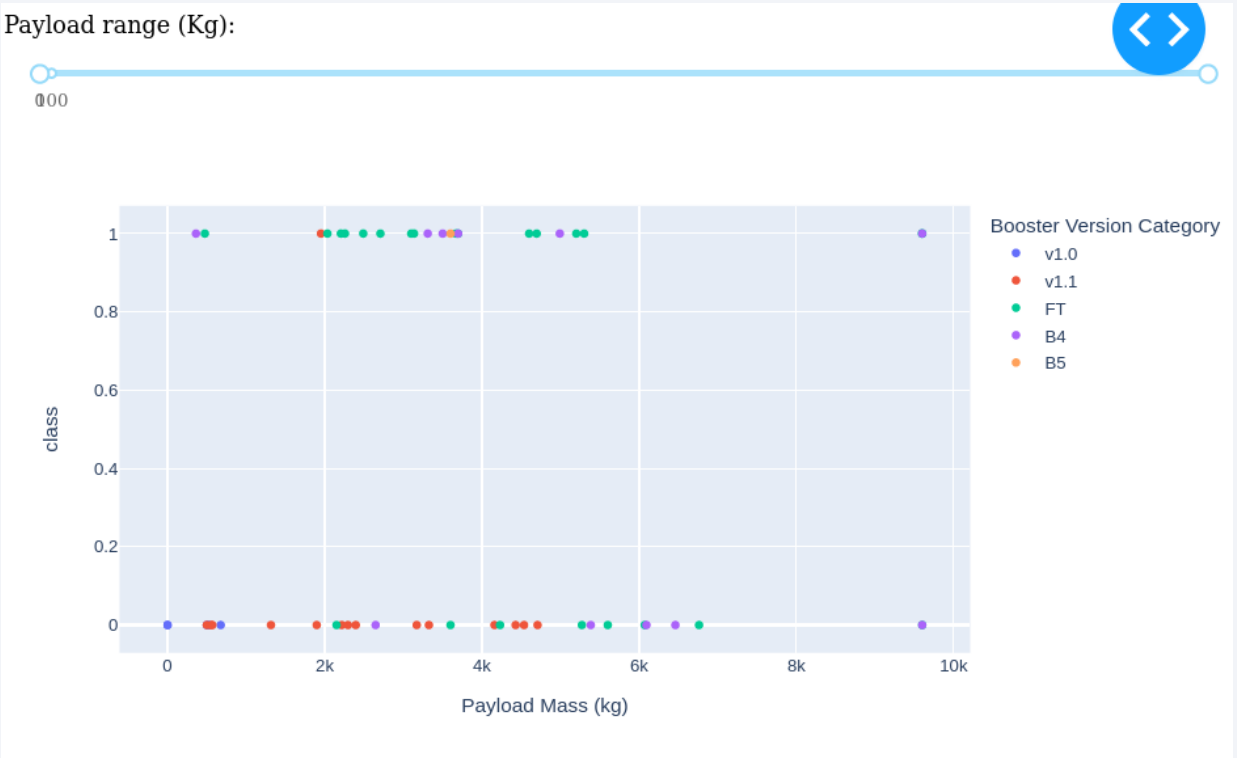
- The location of launching plays an important factor in the success rate.

Launch Success Ratio for KSC LC-39A



- 73.1% of the total launches are successful at the KSC LC-39A launches

Payload vs Launch Outcome



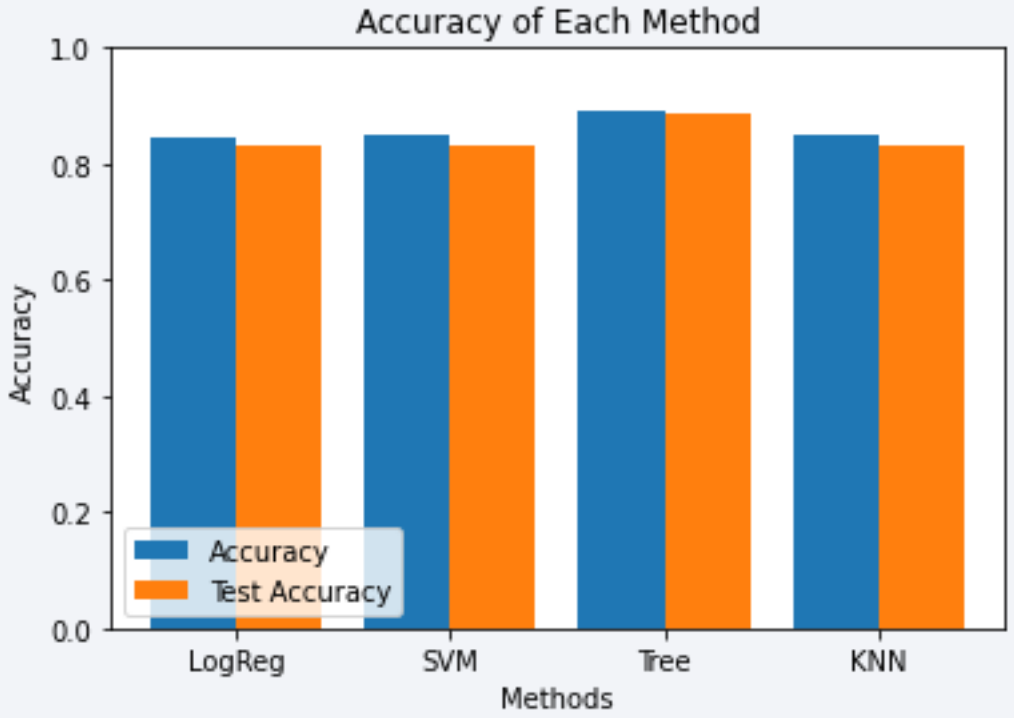
- Payloads under 6,000kg and FT boosters have the highest success.
- Payloads over 7000kg cannot be estimated due to lack of data.

Section 5

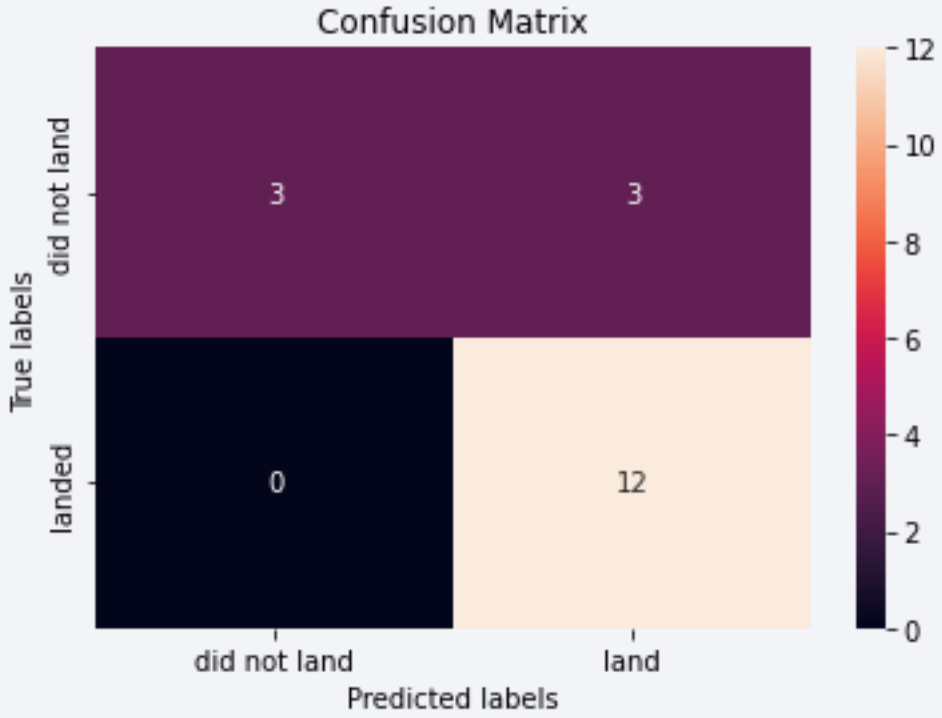
Predictive Analysis (Classification)

Classification Accuracy

- 4 different models were built and the classification accuracy was plotted next to one another for comparison.
- The decision tree achieved the highest accuracy with over 89% accuracy on both train and test data.



Confusion Matrix



- The model appears to have a problem classifying false negatives.

Conclusions

- The most favorable launch site is KSC LC-39A.
- Launches above 7,000 kg are safer.
- Mission outcomes are more successful in general than landing outcomes.
- Landing outcomes improve over the years the company continues operating.
- The decision Tree classifier is the best model at predicting successful landings.

Thank you!