REACHING NEW SKIES WITH DATA SCIENCE

Bryce A. Haraldsen

March 23, 2024

INDEX

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

Executive Summary

- Summary of Methods Used:
  - *Data Collection using SpaceX API*
  - *Data Collection using Web Scraping techniques*
  - *Data Wrangling*
  - *Exploratory Data Analysis with Data Visualization*
  - *Exploratory Data Analysis with SQL*
  - *Interactive Analysis with Folium and Plotly*
  - *Predictive Analysis with Machine Learning*
- Summary of Results
  - *Exploratory Data Analysis results*
  - *Interactive Visuals*
  - *Predictive analysis results*

Introduction

- Project Scope:
  - *SpaceX advertises the Falcon 9 rocket to cost about $62 million dollars, while competitors are costing almost triple the amount at $165 million. This amount of savings is mainly due to SpaceX's ability to reuse the first stage of the Falcon 9 rocket.*
  - *Our objective of this project is to predict if the first stage of the Falcon 9 rocket will land successfully given Launch site and payload mass, landing area and other operational conditions.*

METHODOLOGY

Section 1

Methodology

Executive Summary

- Data Collection Methodology
  - SpaceX Rest API
  - Web Scrapping through Wikipedia
- Data Wrangling
  - Cleaned Data
  - One-hot encoding on Categorical features
- Exploratory data analysis (EDA) using visualizations and SQL
- Create visual analytics with Folium and Plotly
- Prediction analysis using Classification models
  - Best classifier was tested and evaluated via LR, KNN, DT and SVM

Data Collection – SpaceX API

- Data was collected through GET requests from the SpaceX API, and formed into a data frame using Pandas in Python. The data from requests was then formatted and filtered to only keep launches that were from SpaceX's Falcon 9 rockets.
  - *Any missing values were replaced with mean values for the feature*

- Source:https://github.com/Haraldsenb46/Applied_Data_Science_Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection

Web Scraping

- Another source of data was SpaceX's Wikipedia page about the Falcon 9 rocket.
  - Web scrapping was done via Beautiful Soup to look for columns and values about the Falcon 9.
  - All data was the assimilated into the a Data frame along with the API data.
- Source:https://github.com/Haraldsenb46/Applied_Data_Science_Capstone/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

- Exploratory Data Analysis was performed to determine, the number of launches per site, all outcomes of each launch and the numbers of each occurrence, as well as what type of Orbit and Payload.
  - A binary label called *Landing Outcome* was created from the Outcome column and inserted into the Class column
- Source: https://github.com/Haraldsenb46/Applied_Data_Science_Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

Data Visualization

- Data Visualizations were created to assist with EDA to obtain further insight on how a multitude of factors affected the outcome of the Falcon 9 Rocket launches. The following charts were plotted:
  - *Payload mass vs Flight number*
  - *Flight number vs Launch site*
  - *Payload mass vs Launch site*
  - *Success rate for each Orbit*
  - *Flight number vs Orbit type*
  - *Launch success trend over time*
- Relevant features were filtered and categorical columns were One-hot encoded
- Source; https://github.com/Haraldsenb46/Applied_Data_Science_Capstone/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

Eda using SQL

- SQL queries were used on data to perform further data analysis:
  - List of unique launch sites
  - Display total payload mass by boosters launched by NASA
  - Average payload mass carried by the version 1.1 booster
  - Date of first successful ground pad landing outcome
  - Boosters that landed in drone ship with mass between 4000 and 6000kg
  - Total number of successful and unsuccessful mission outcomes
  - Booster versions that have had the maximum payload mass
  - Month and booster versions of unsuccessful drone ship landing outcomes in 2015
  - Descending order ranking of all landing outcomes between June 6th, 2010 and March 3rd 2017.
- Source: https://github.com/Haraldsenb46/Applied_Data_Science_Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite-Copy1.ipynb

Interactive Maps with Folium

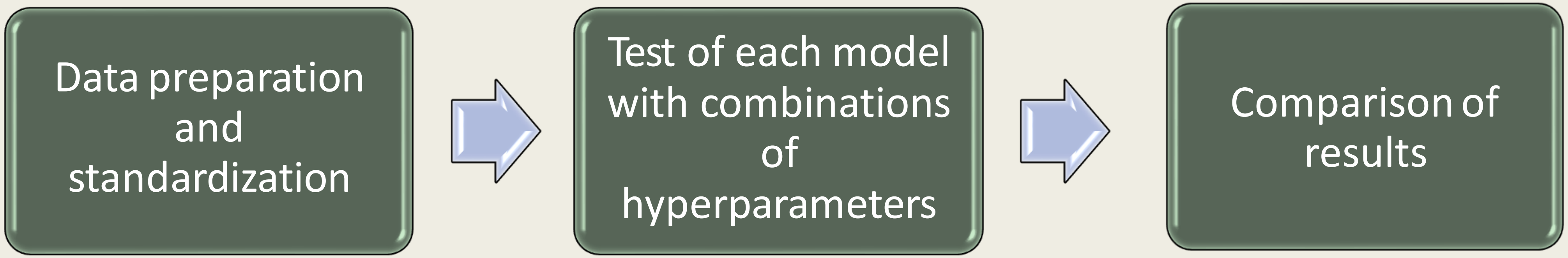
- Folium used to analyze launch site locations, proximities to launch site and other information that could reason why specific launches occurred at each site.
  - Circles and markers used to highlight launch sites with text labels.
  - Marker clusters were added to each site to show launch records and colored to match outcome of each launch. (green or red)
  - Lines represented the distance between the launch site and distinct markers to show distance between site and cities, roads, railways and airports.
- Source: https://github.com/Haraldsenb46/Applied_Data_Science_Capstone/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

Plotly Dashboard

- An interactive visual display was built using a Plotly Dash dashboard
  - Pie chart was created to represent ratios of successful launches, with an ability to view an individual site success rates or the overall success rate.
  - Scatter plot was also created to provide more insight between payload mass and outcome, and sorted by booster version where each color represents a different booster.
- Source: https://github.com/Haraldsenb46/Applied_Data_Science_Capstone/blob/main/SpaceX_Dash_App.py

Predicitve Analysis

- Classification models compared in analysis: K nearest neighbors, Logistic regression, Support vector machine, and Decision tree



Data preparation and standardization

Test of each model with combinations of hyperparameters

Comparison of results

Source:
https://github.com/Haraldsenb46/Applied_Data_Science_Capstone/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

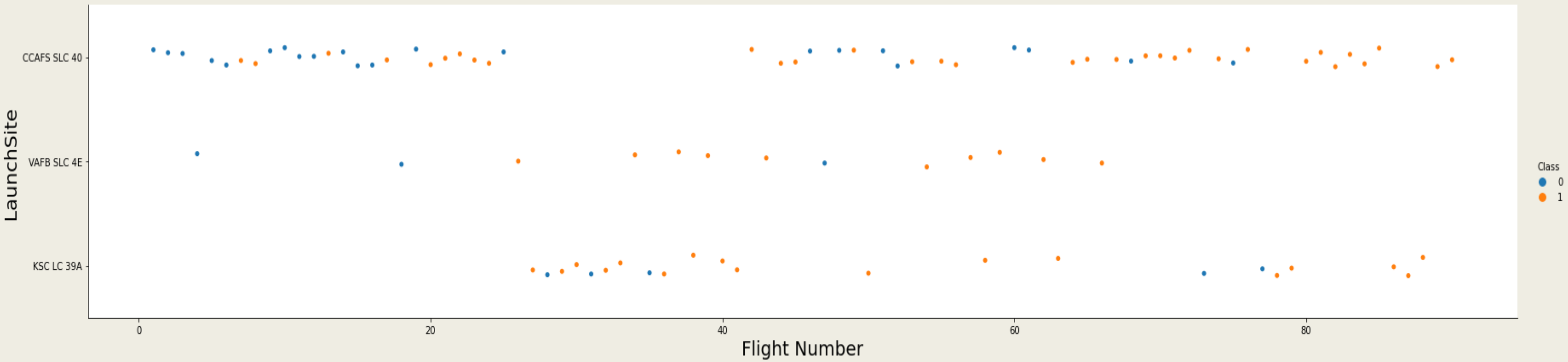
Results of EDA

- 4 different launch sites are used
  - *Launches conducted mainly by SpaceX and NASA*
  - *SpaceX's first successful landing outcome was in 2015.*
- Payload mass tended to increase as time and success ratios increased.
  - Specifically, from 2013 to present.
- Interactive analysis
  - Most sites were on the East Coast and close to the Atlantic
    - Likely for easy access to use drone ships
  - Near areas with sound infrastructure to support launches
- Predictive Analysis
  - Decision tree was most accurate model to use for predictive landing success.
    - > 87% accuracy

EXPLORATORY DATA ANALYSIS INSIGHTS

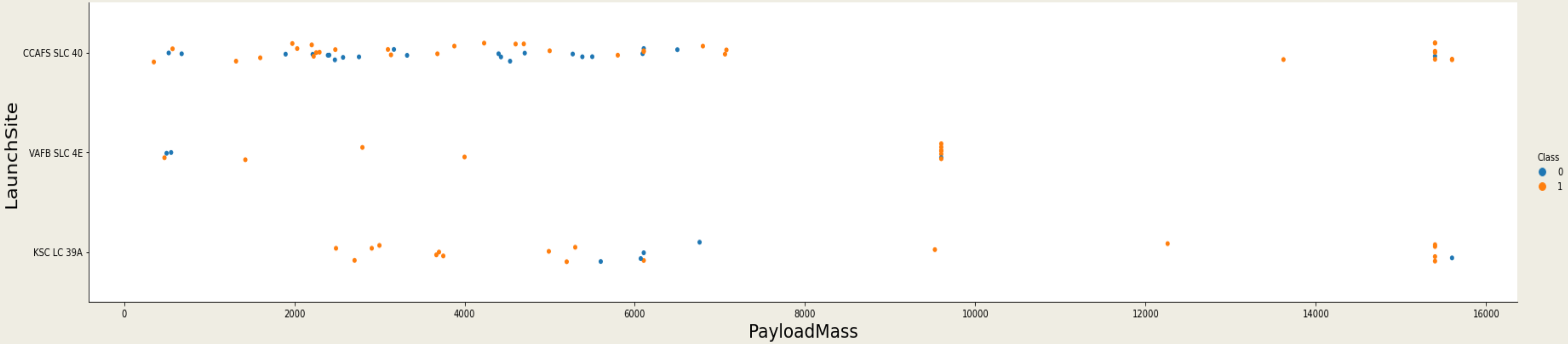
Section 2

Flight number vs Launch Site



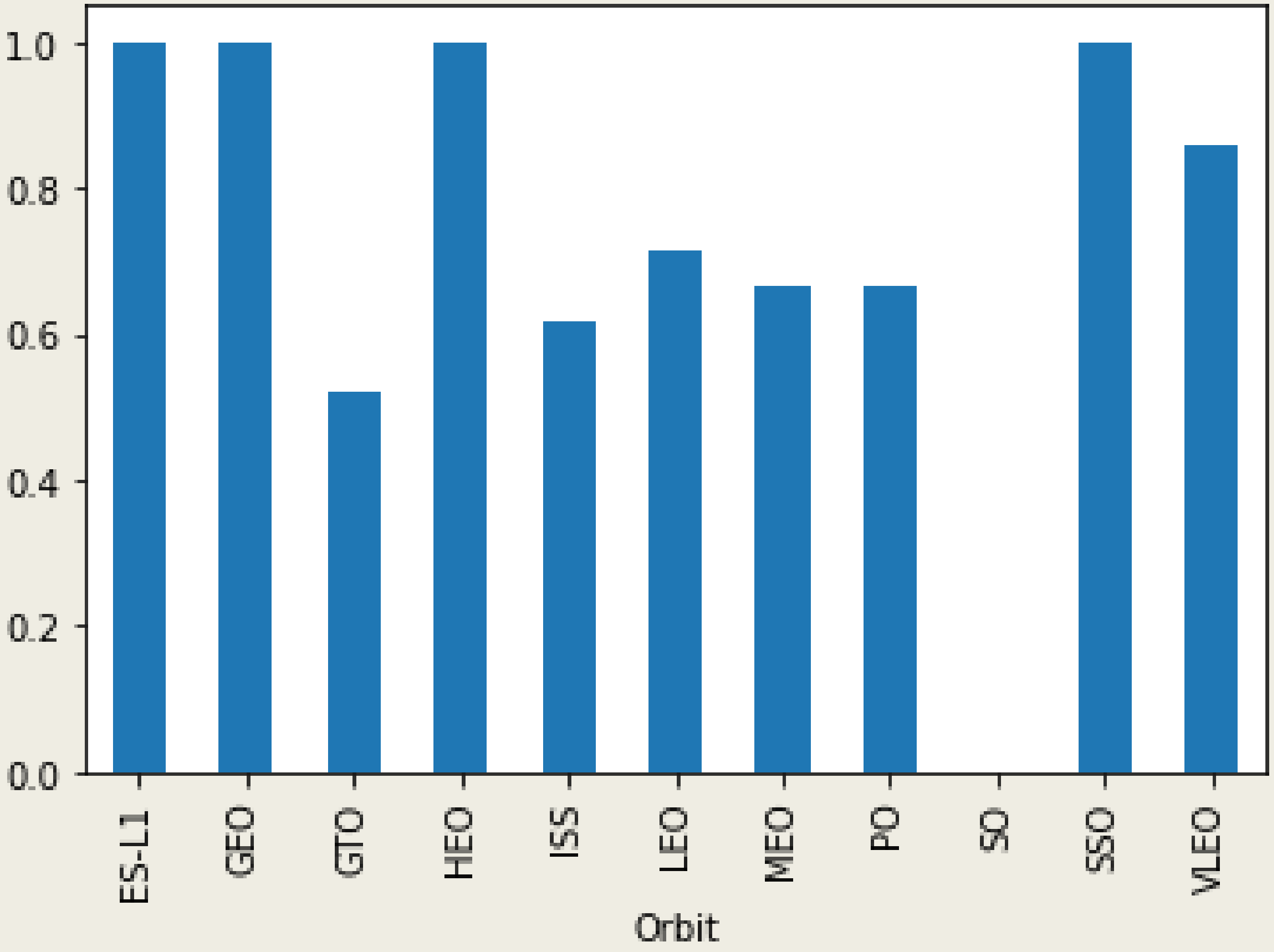
- The graph above illustrates relations between flight numbers and launch sites used. With a trend showing successful flights (class 1) increasing drastically as the flight number increases.

Payload vs Launch site



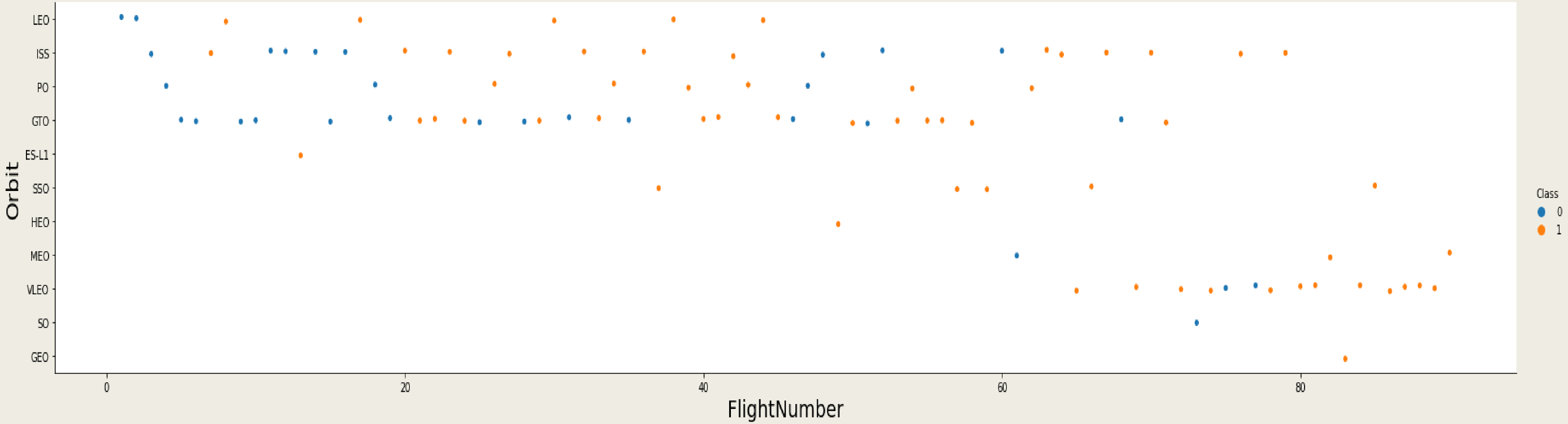
- The above graph shows the relation between payload mass (KG) to the launch site used. A few notable trends from this graph:
  - *VAFB-SLC only launched rockets a payload of 10000 KG and less.*
  - *While the lower payload rockets had a mixed bag of success and failure, large payload rockets had a high success rate.*

Success rate vs Orbit type



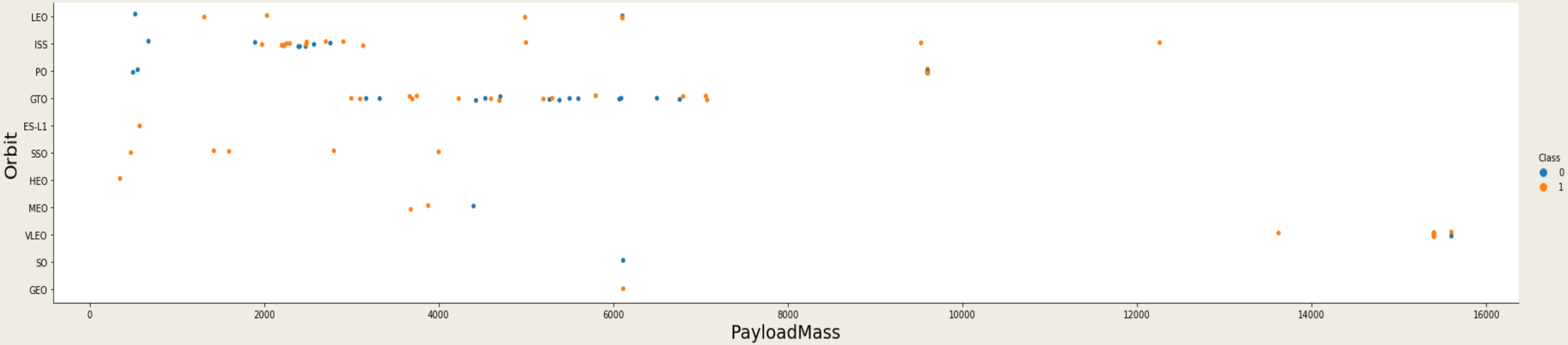
- The bar graph to the left shows relationships between rocket orbits and success outcomes.
  - ES-L1, GEO, HEO, and SSO orbits had near perfect success rates
  - SO orbits experienced a 100% failure rate for landing outcomes

Orbit Type vs Flight Number



- The scatter plot above represents the relation between orbits of each flight mission and the flight number. The colored classes are used to distinguish between successful and unsuccessful mission landing outcomes.
- Notable relations are that VLEO and SSO orbits had significantly high success rates.

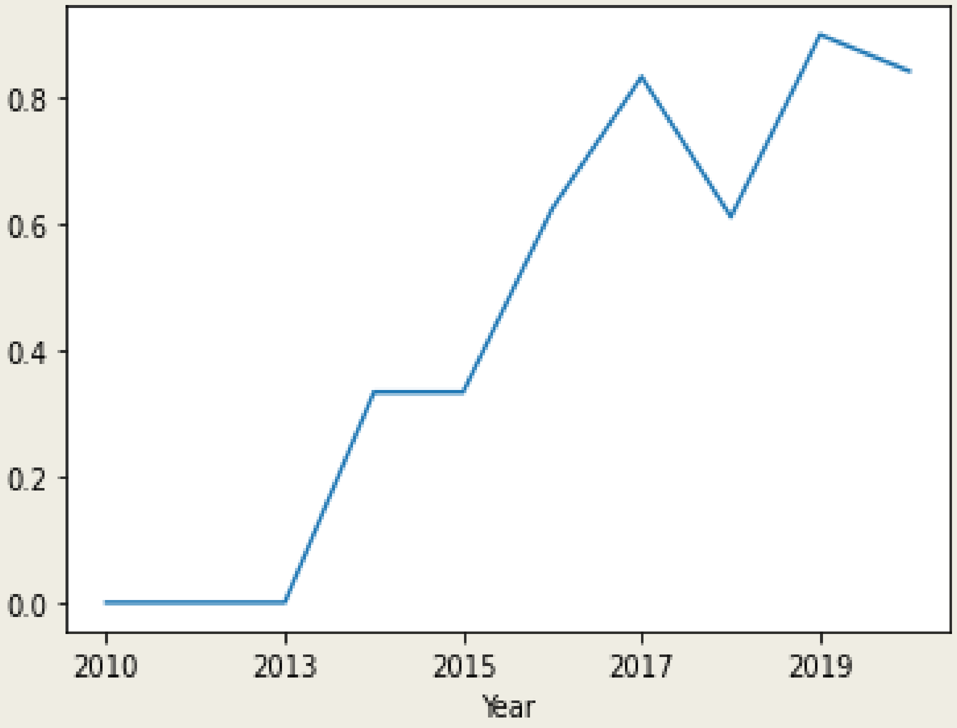
Payload vs Orbit Type



- The scatter plot above displays relations between payload mass (in KG) and mission orbits.
  - *Most failures happened with payloads of < 6000KG.*

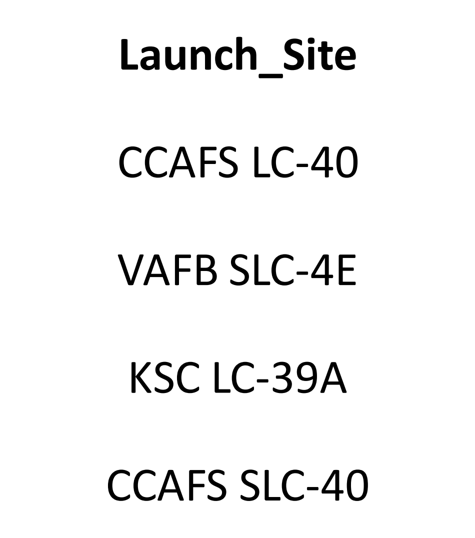
Launch Success Over Time

- The Line graph to the right shows average landing outcome success rate for each year.
  - *Success rate increased significantly after 2013.*
  - *Sharp decrease in success after 2017.*



All Launch Site Names

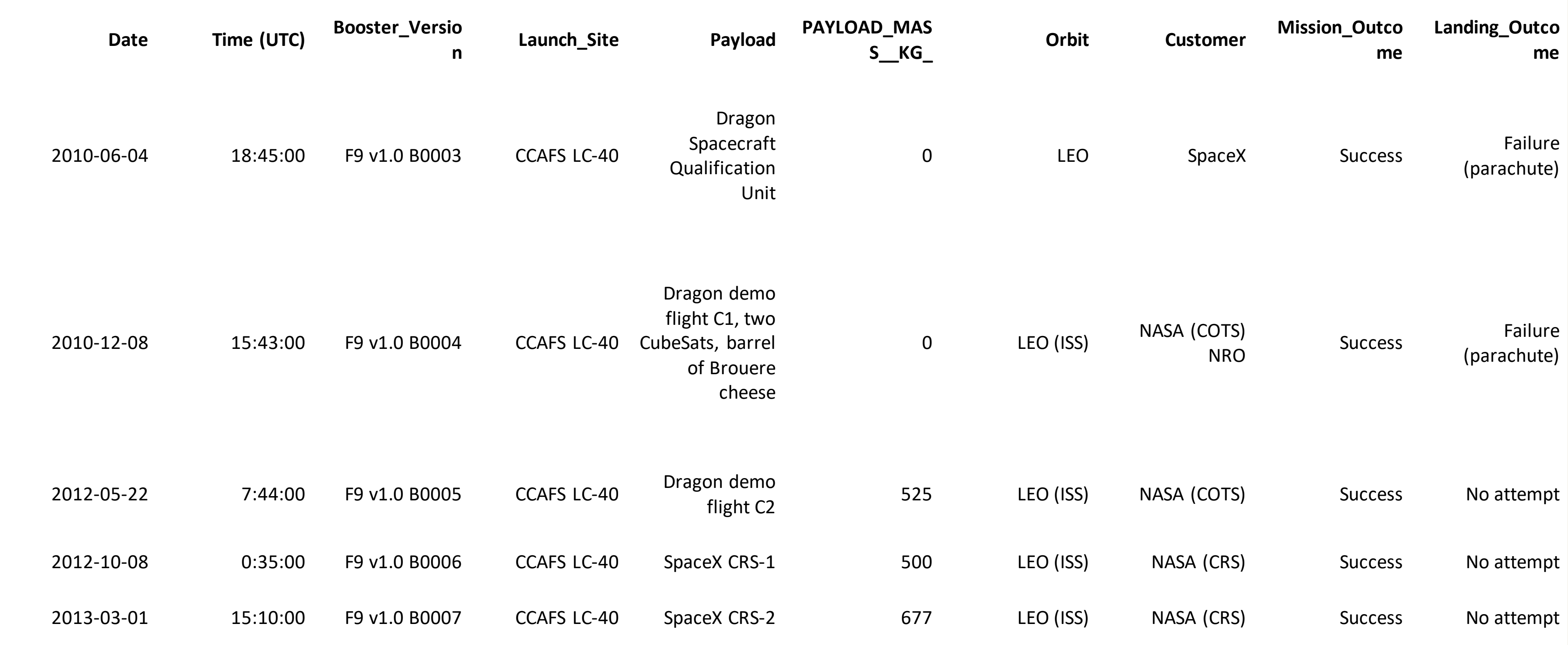
- Query used: sql SELECT DISTINCT(“Launch_Site”) FROM SPACEXTABLE;
  - *This Query was used to identify all unique names of launch sites from the data.*



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

Launch Site Name Begins with ‘CCA’

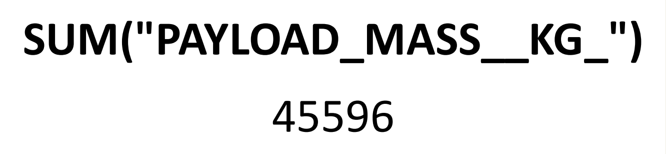
- Query used: sql SELECT * FROM SPACEXTABLE WHERE “Launch_Site” LIKE “CCA$” LIMIT 5:
  - – *This Query was used to find 5 launches from the CCAFS LC-40 Launch site.*



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

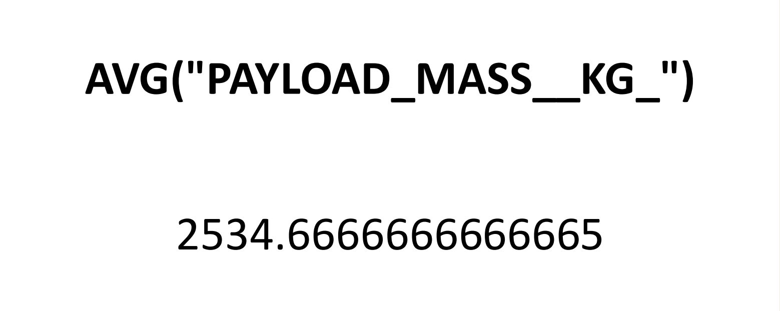
- Query used: sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE PAYLOAD LIKE '%CRS%';
  - *This query was used to find the total p0ayload carried by NASA launched rockets.*



SUM("PAYLOAD_MASS__KG_")
45596

Average Payload Mass From F9 v1.1

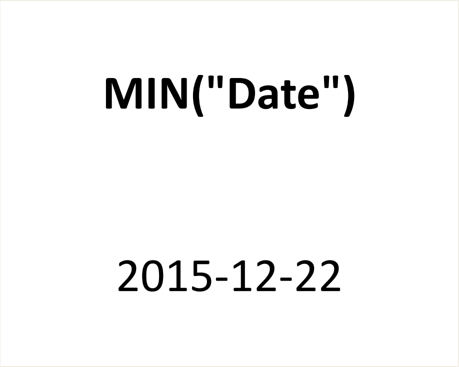
- Query used:
  sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1’;
  - *This Query was used to find the average payload of the Falcon 9 Booster v1.1*



AVG("PAYLOAD_MASS__KG_")
2534.6666666666665

First Ground Landing Successful

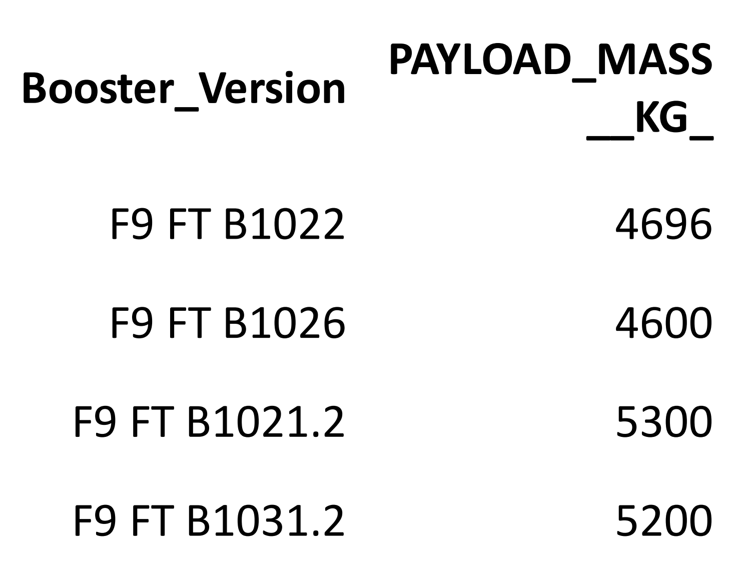
- Query used: sql SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEXTBL WHERE Landing_Outcome = 'Success (ground pad)';
  - *This query was used to find first successful landing outcome.*



MIN("Date")
2015-12-22

Successful Drone Ship Landing with Payloads between 4000 and 6000 KG

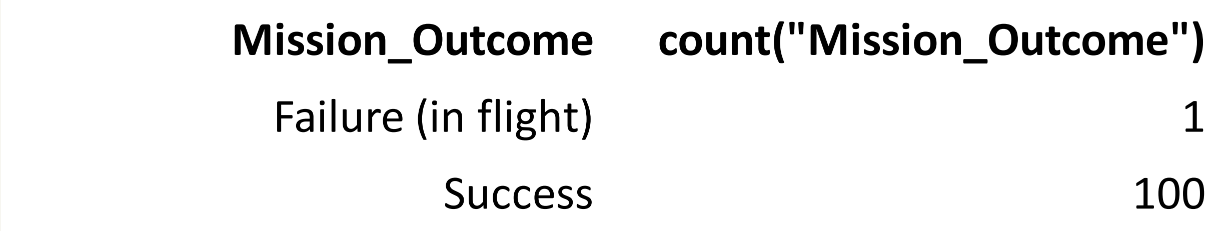
- Query used: sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND Landing_Outcome = 'Success (drone ship)’;
  - *This query helped us find all distinct boosters that successfully landing on a drone ship between the desired payloads.*



Booster_Version	PAYLOAD_MASS__KG_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

Total Successes and failures

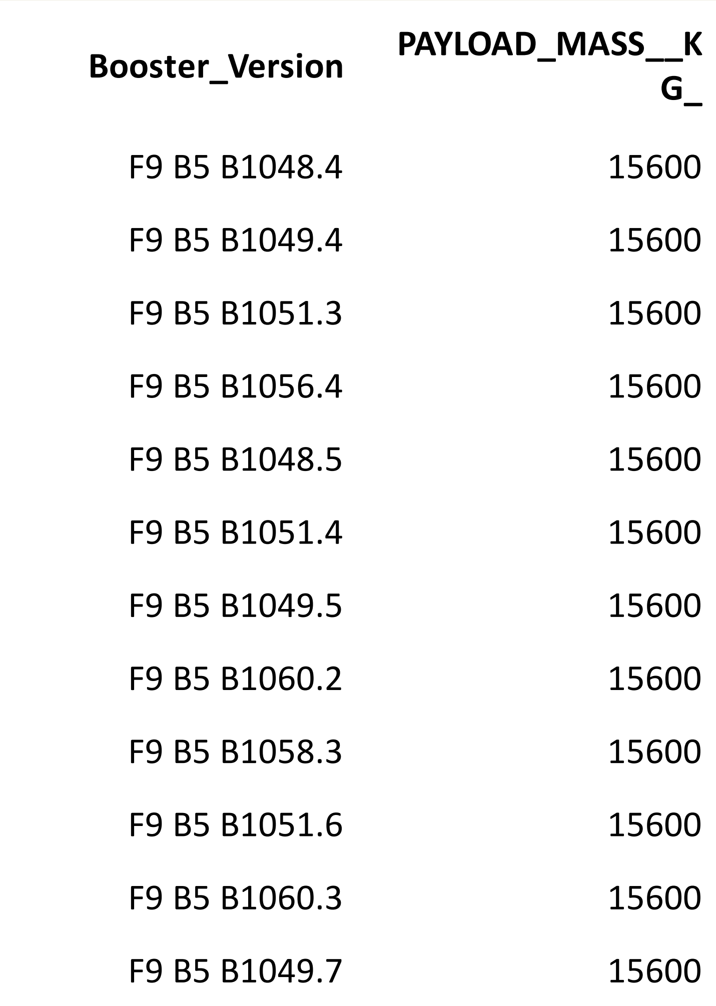
- Query used: sql SELECT MISSION_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL GROUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME;
  - *This query found the number of successful mission outcomes and failures.*



Mission_Outcome	**count("Mission_Outcome")**
Failure (in flight)	1
Success	100

Maximum Payload Boosters

- Query used: %%sql SELECT "Booster_Version", "PAYLOAD_MASS__KG_" FROM SPACEXTABLE WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTABLE)
  - *This query found all Booster versions that carried maximum payloads*



Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 failed outcome records

- Query used: sql SELECT substr(Date,6,2) AS 'Month', "Landing_Outcome", "Booster_Version", "Launch_Site" FROM SPACEXTABLE WHERE substr(Date,0,5)='2015' AND "Landing_Outcome" LIKE 'Failure (drone ship)’;
  - *This query shows all failed landing outcomes in 2015 with the booster version and launch site.*

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

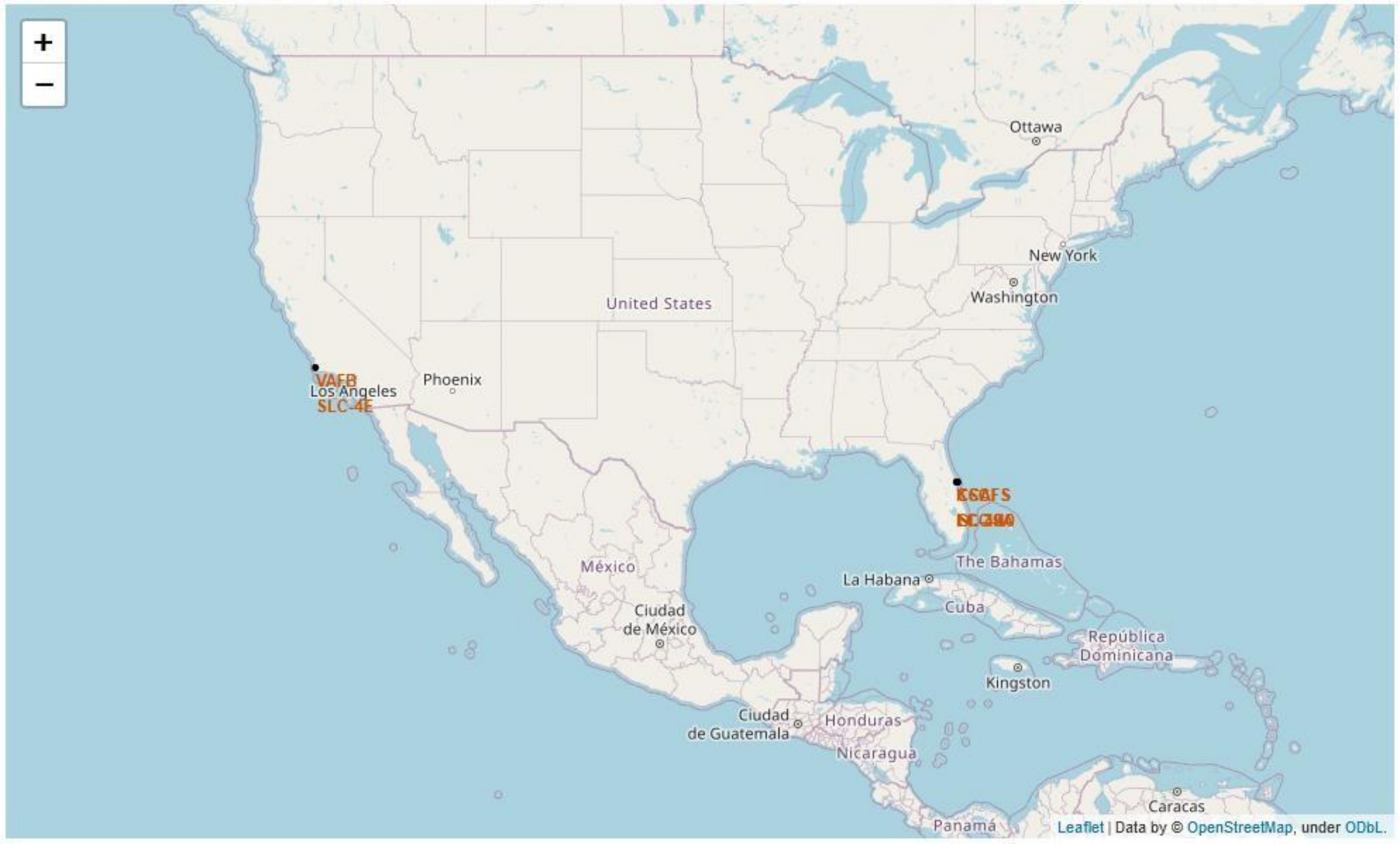
Landing outcomes

- Query used: sql SELECT "Landing_Outcome", COUNT("Landing_Outcome"), "Date" FROM SPACEXTABLE WHERE DATE >= '2010-06-04' AND DATE <= '2017-03-20’ GROUP BY "Landing_Outcome“ ORDER BY COUNT("Landing_Outcome") DESC;
  - *This query ranked lists the number of occurrences of each outcome of rocket launches between 2010-06-04 and 2017-03-20.*

Landing_Outcome	COUNT("Landing_Outcome")	Date
No attempt	10	2012-05-22
Success (drone ship)	5	2016-04-08
Failure (drone ship)	5	2015-01-10
Success (ground pad)	3	2015-12-22
Controlled (ocean)	3	2014-04-18
Uncontrolled (ocean)	2	2013-09-29
Failure (parachute)	2	2010-06-04
Precluded (drone ship)	1	2015-06-28

LAUNCH SITE PROXIMITIES ANALYSIS

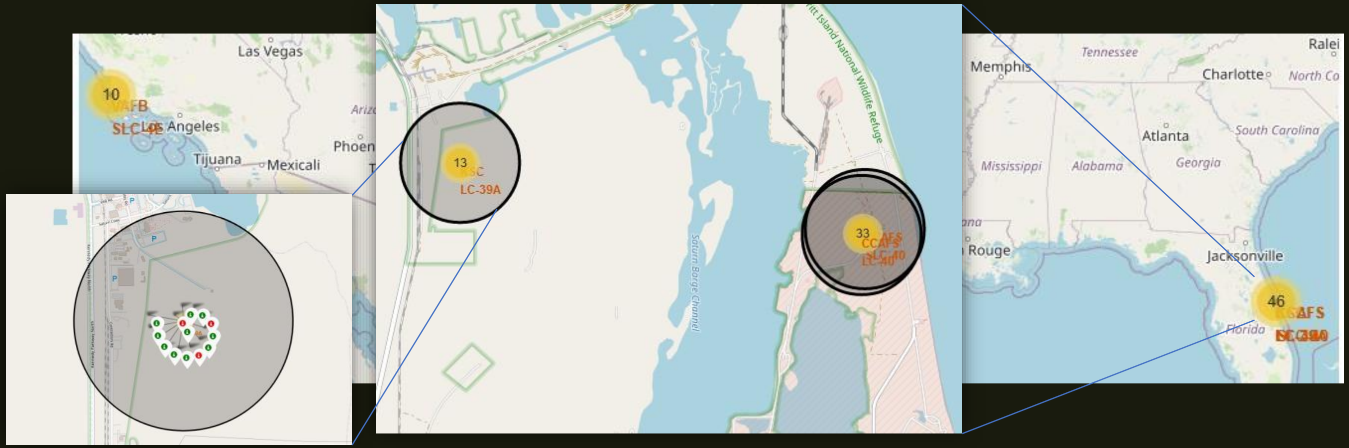
Section 3



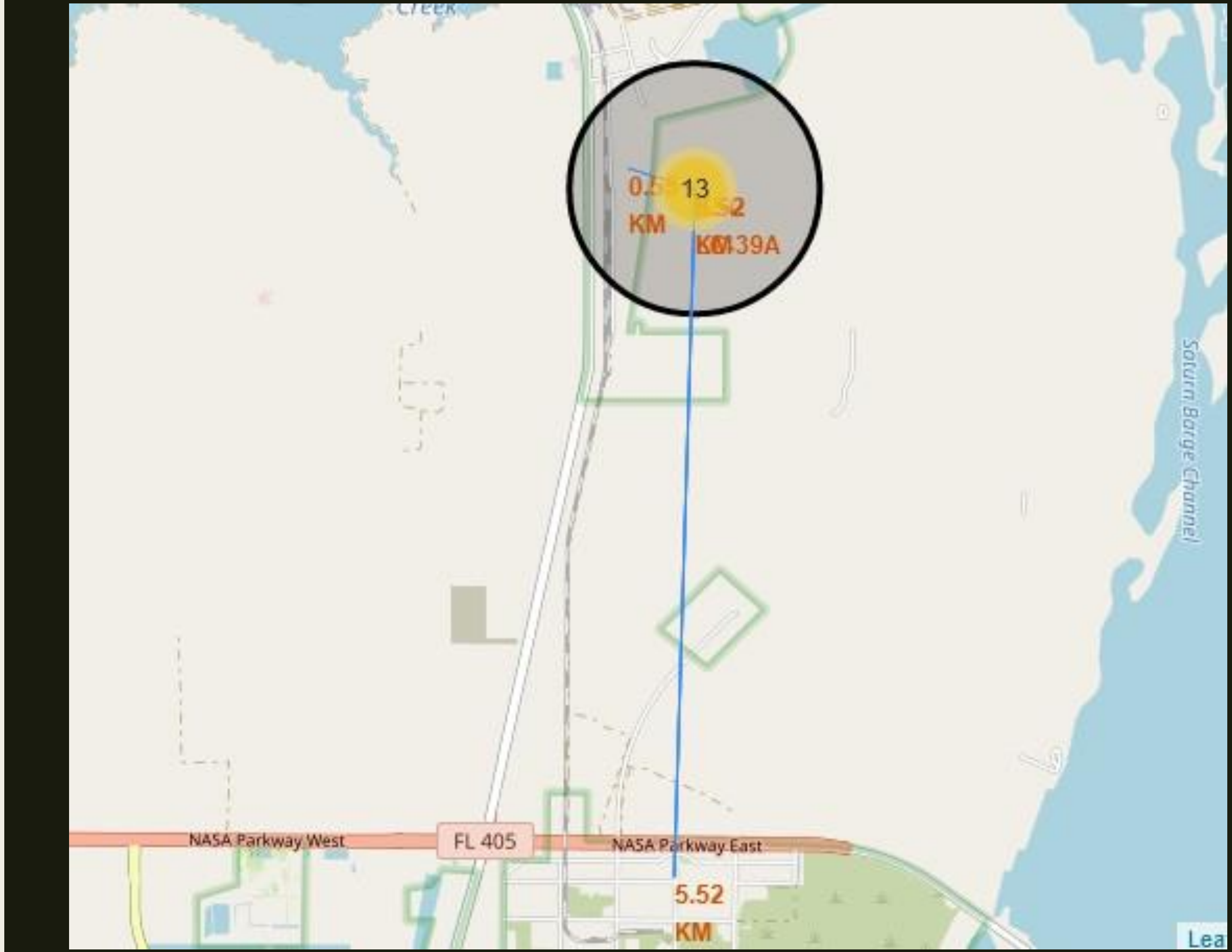
All Launch Sites

- Launch sites located close to oceans , but not far from roads, railroads and cities.

Launch Outcomes per Site



- Shown is the KSC LC-39A Launch site with outcomes marked in green for successful and red for failure.



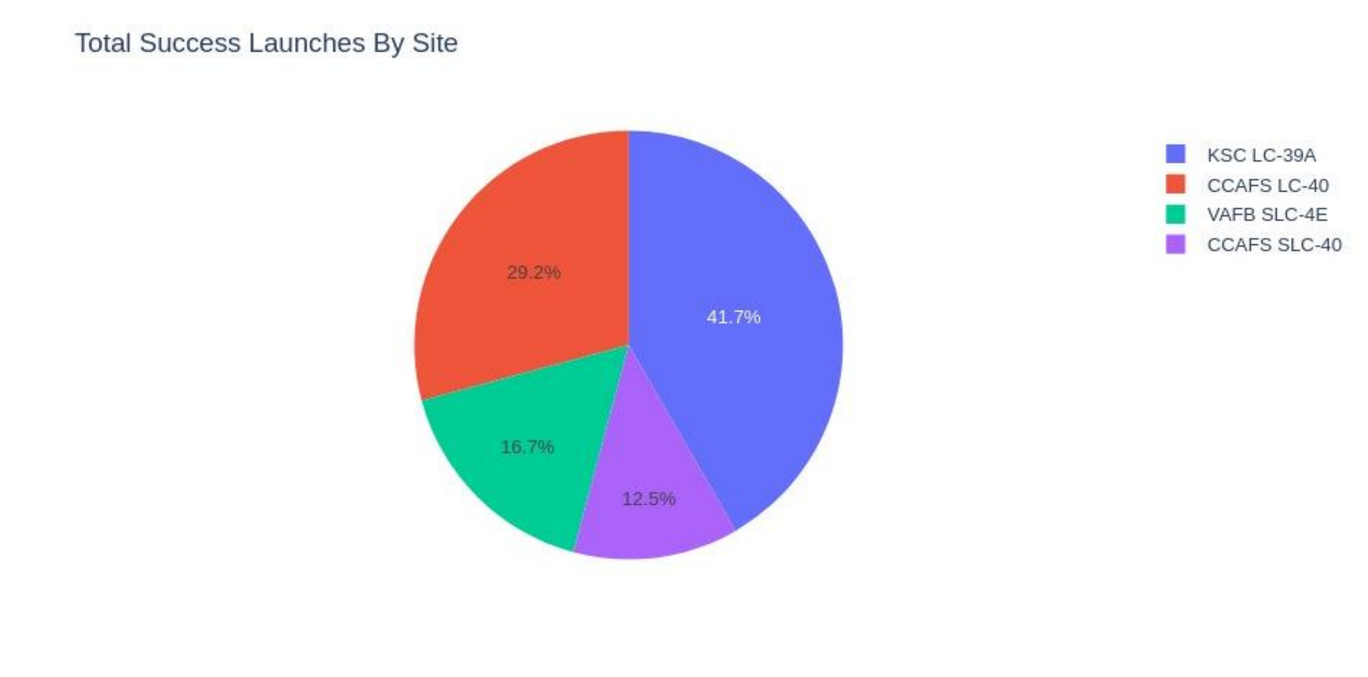
Safety Precautions

- KSC LC-39A is a launch site located right along a railway and highway but not too close to a populated area.

DASHBOARD WITH PLOTLY DASH

Section 4

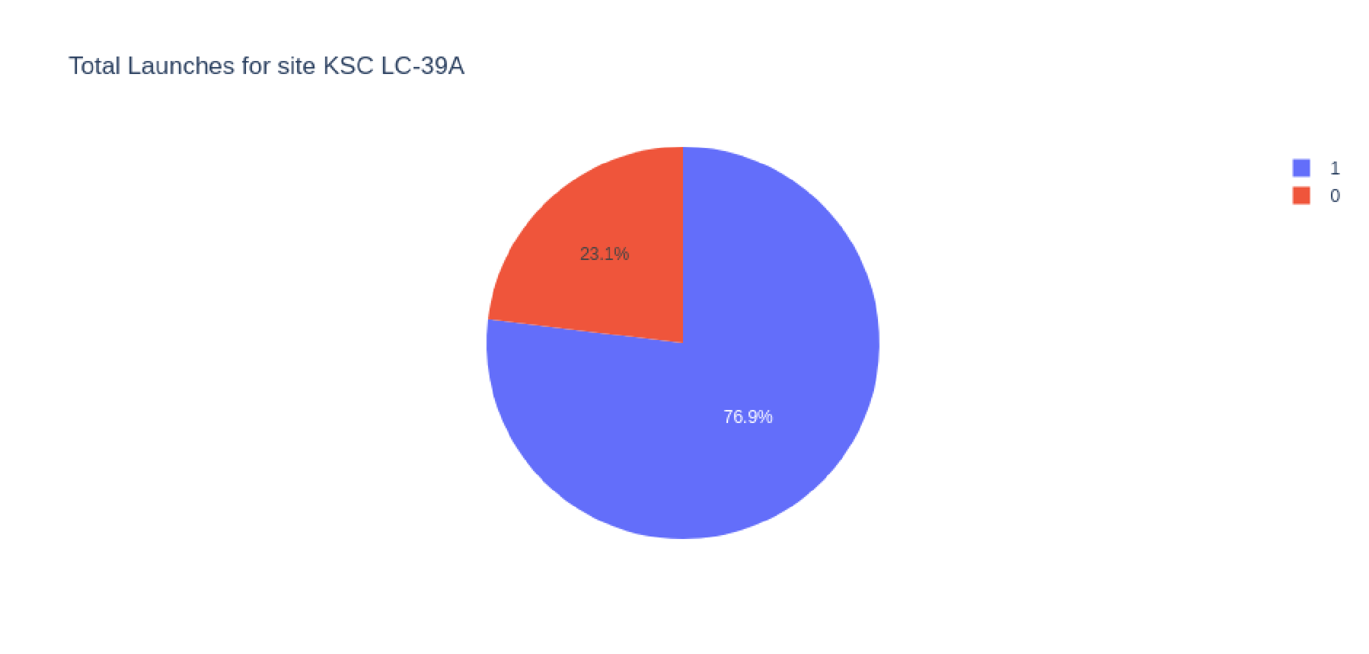
Launch Success ratios



- The Image to the left was a pie chart created in plotly dash to show the successful launch ratio for all launch sites.
  - *KSC LC-39A and CCAFS-LC-40 had the highest ratios.*
  - *CCAFS-SLC-40 had the lowest.*

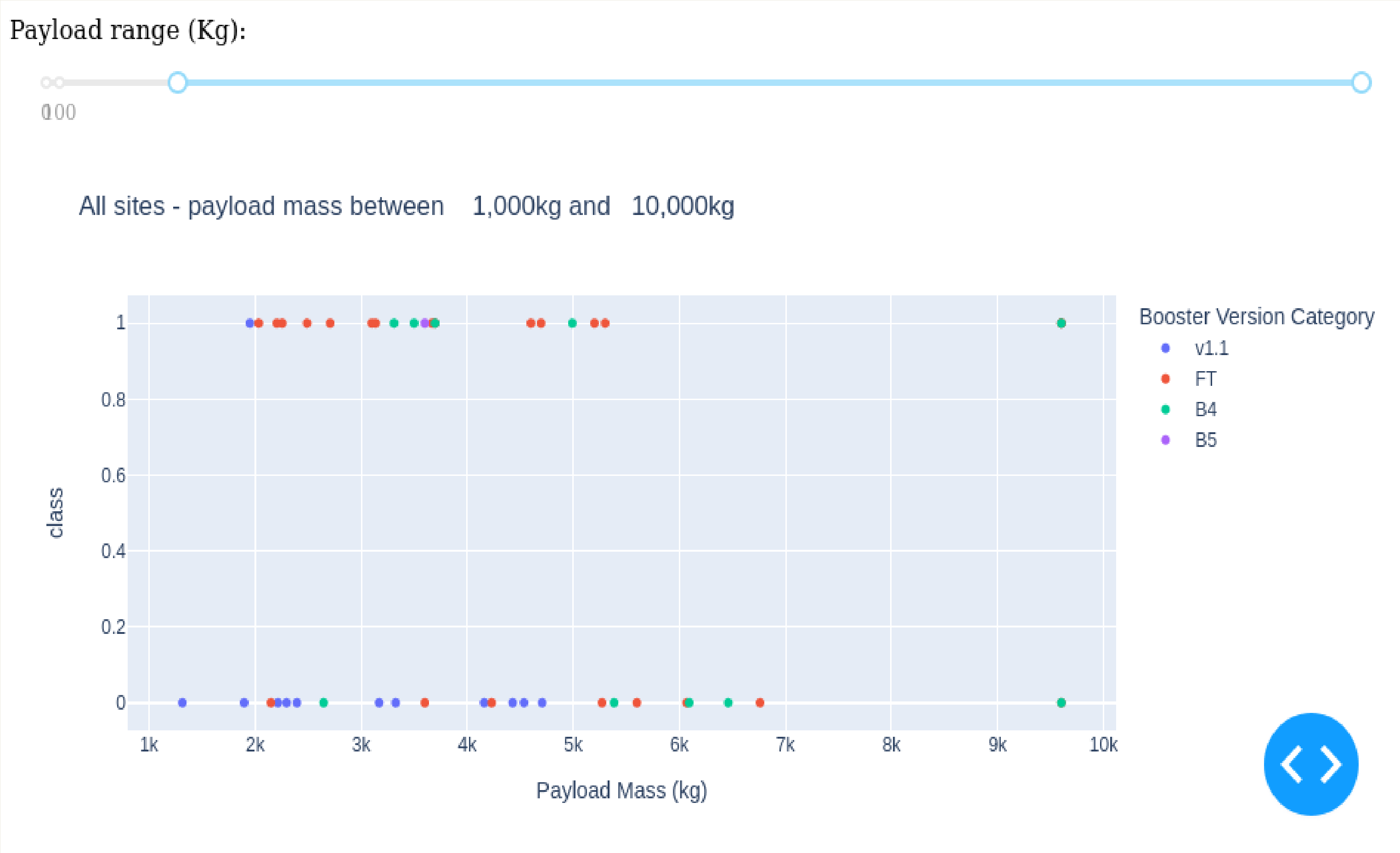
Launch ratio by site

- Shows individual launch site success ratio.
  - *KSC-LC-39A had a launch success rate of almost 77%.*



Payload vs Launch outcome

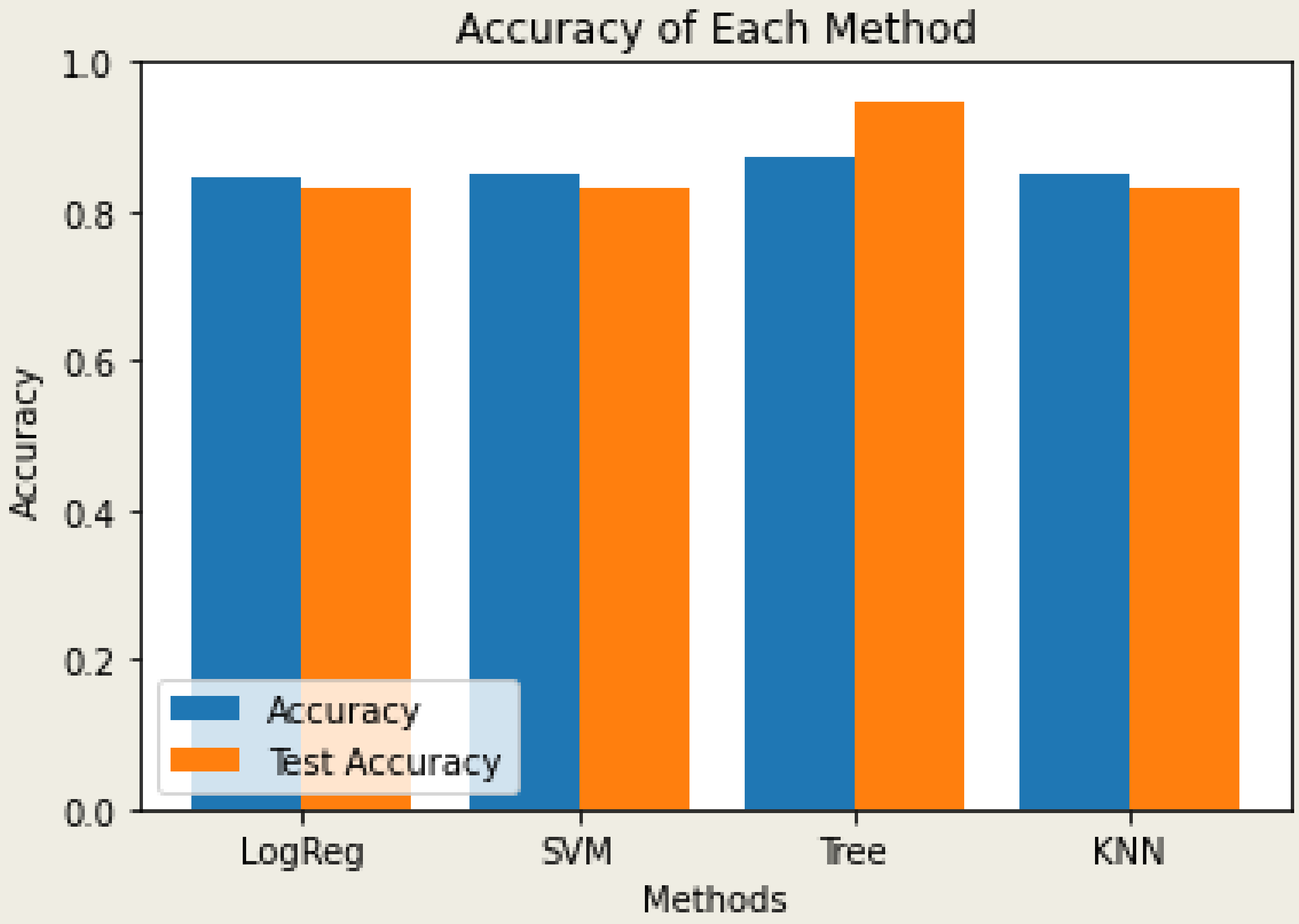
- The Scatter plot to the right represents the correlation between payload mass and success outcome.
- Points are colored to represent different booster versions.
  - *Success rates for payloads under 4000KG is significantly higher than success rates for larger payloads.*



Payload range (Kg):

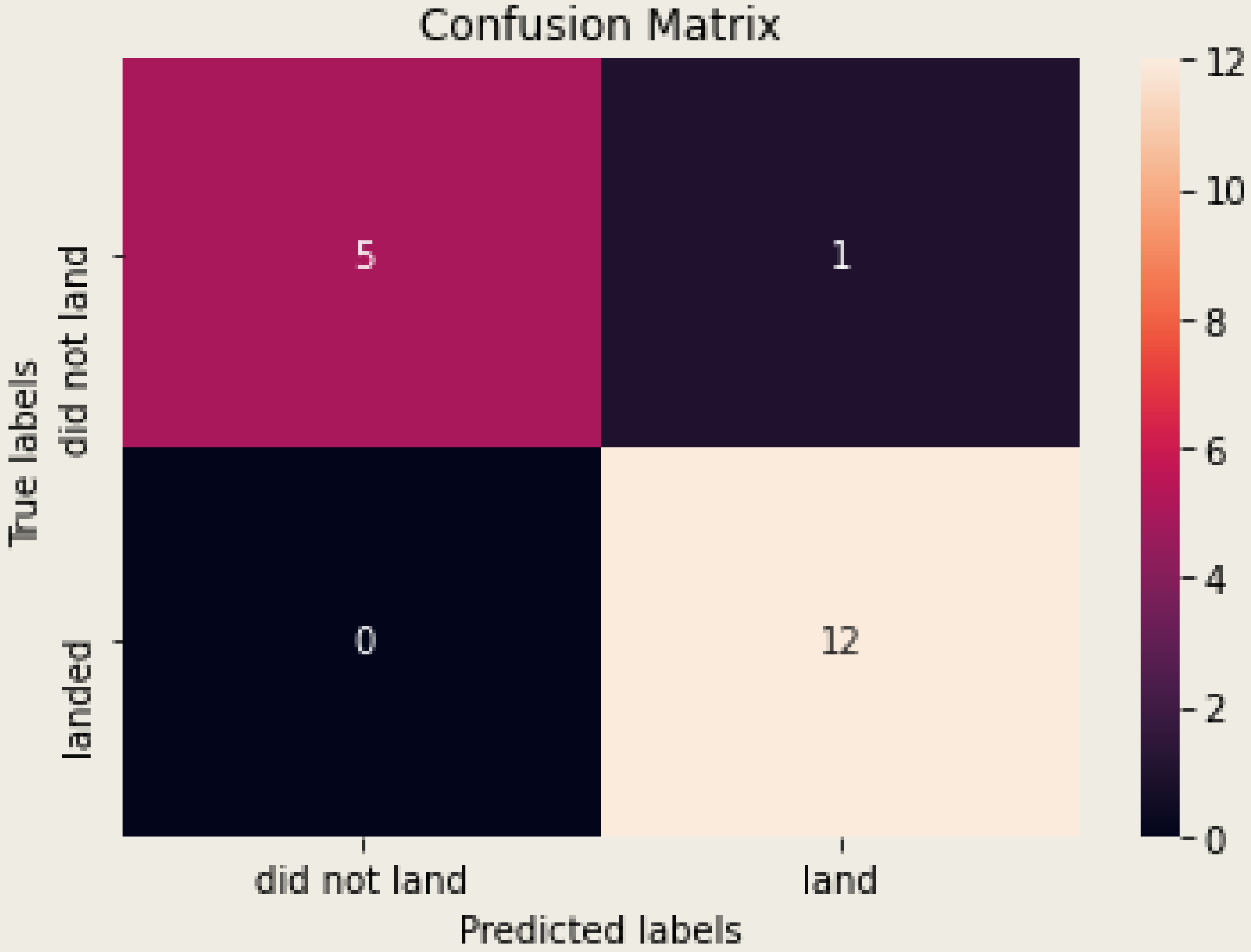
PREDICTIVE ANALYSIS

Section 6



Classification Accuracy

- The bar chart to the left represents the four classification models used with their training model(blue) and Test model(orange).
- The Decision tree model seems to have the highest accuracy in this scenario.



Confusion Matrix

- The Decision tree model when put into a confusion matrix and tests shows that the test only experienced one false positive while all other predictions were true.

Conclusions

- Best launch site is KSC LC-39A
- Most failures happened in lower payloads but many more lower payload rockets have been launched than heavier payload rockets.
  - *Likely due to beginning missions having lower payloads to test rocket booster versions initially and then heavier payloads being used after success with initial tests, which would also explain why heavier tests were often successful.*
- Landing outcomes were more and more successful as time went on as well.
- Decision tree classifier is the best candidate for predicting whether a rocket will have a successful landing outcome or not.