IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Alfred Nøland

15 August 2024

Outline

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

Executive Summary

Summary of methodologies

- Data collection
- Data wrangling
- Exploratory Data analysis with Data Viusualization
- Exploratory Data analysis with SQL
- Interactive maps with Folium
- Dashboards with Plotly Dash
- Predictive analysis

Summary of all results

- Exploratory Data Analysis results
- Imteractive analytics demo in screenshots
- Predictive analysis results

Introduction

Project background and Context

SpaceX leads the commercial space age, making space travel more affordable. The company lists Falcon 9 rocket launches on its site, costing 62 million dollars; other providers charge over 165 million each. Much of the savings come from SpaceX reusing the first stage. Thus, by predicting if the first stage will land, we can estimate the launch cost. Using public data and machine learning, we will predict SpaceX's ability to reuse the first stage.

Questions answered

- How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?
- Does the rate of successful landing increase over the years?
- What is the best algorithm tha be used for binary classification in the case?

Section
1

Methodology

Methodology

Executive Summary

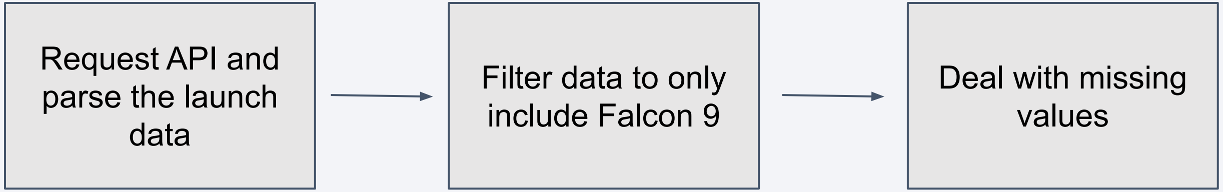
- Data collection methodology:
  - Using SpaceX Rest API
  - Using Web scraping to scrape Wikipedia
- Perform data wrangling
  - Filtering data, Dealing with missing values and One hot encoding
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building, tuning and evaluation of classification models to ensure the best results.

Data Collection

- The data collection process involved using a combination of API requests from the SpaceX REST API and web scraping data from a table in SpaceX's Wikipedia entry. Both methods were necessary to gather complete information on the launches for a more detailed analysis.
- Data columns obtained via the SpaceX REST API include: FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, and Latitude.
- Data columns obtained via Wikipedia web scraping include: Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, and Time.

Data Collection – SpaceX API

- SpaceX offers a public API from where data can be obtained and them used.
- The API was used accordingly to the flowchart below:



Request API and parse the launch data

Filter data to only include Falcon 9

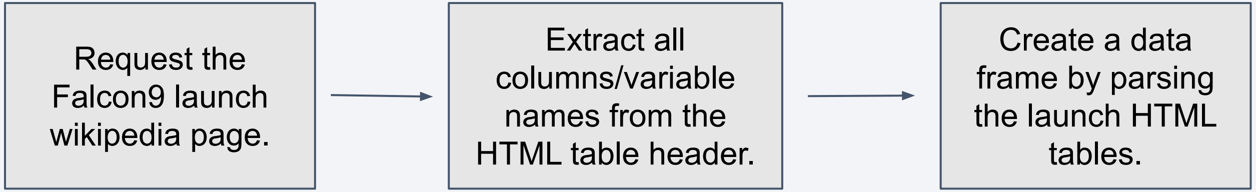
Deal with missing values

Source code:

https://github.com/Alfrednoland/applied-data-science-capstone/blob/main/Data%20Collection%20API.ipynb

Data Collection - Scraping

- Data from SpaceX launches was also scraped from Wikipedia.
- Data was scraped accordingly to the flowchart below:



Request the Falcon9 launch wikipedia page.

Extract all columns/variable names from the HTML table header.

Create a data frame by parsing the launch HTML tables.

Source code:

https://github.com/Alfrednoland/applied-data-science-capstone/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

- Initially some exploratory data analysis (EDA) was performed on the dataset.
- The summaries were then generated for each site, including the number of occurrences for each orbit and the mission outcomes associated with each orbit type.
- Then finally the landing outcome label was created from the Outcome column.

EDA

Summarizations

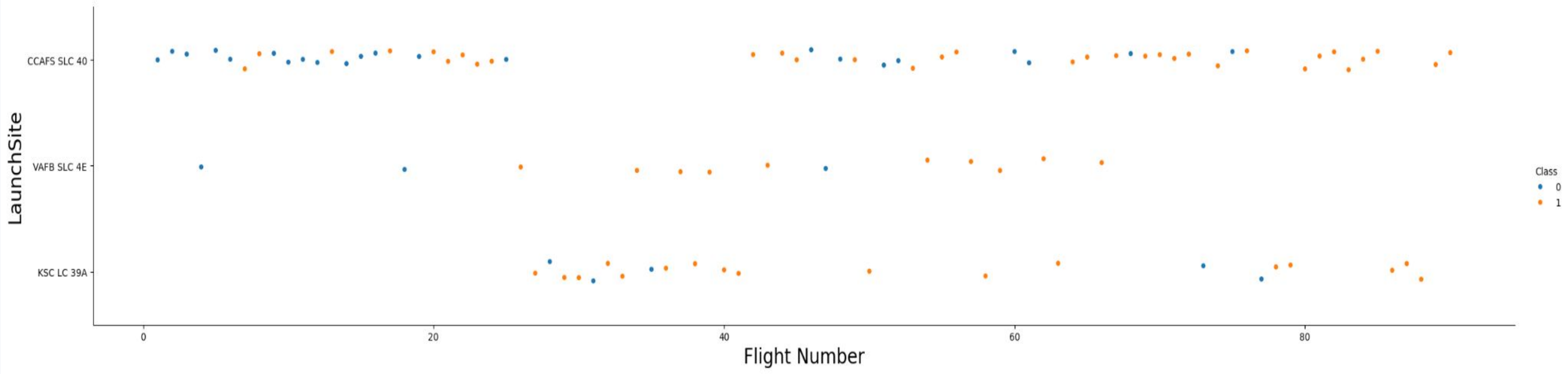
Landing outcome label

Source code:

https://github.com/Alfrednoland/applied-data-science-capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- To explore data, scatterplats and barplots were used to visualize the relationship between pair of features.
  - Payload Mass X Flight Number, Launch Site X Flight Number, Launch Site X Payload Mass, Orbit and Flight Number, Payload and Orbit.



Source code:

https://github.com/Alfrednoland/applied-data-science-capstone/blob/main/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

- Names of the unique launch sites in the space mission;
- Top 5 launch sites whose name begin with the string 'CCA’
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- Date when the first successful landing outcome in ground pad was achieved
- Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg
- Total number of successful and failure mission outcomes
- Names of the booster versions which have carried the maximum payload mass
- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20

Source code:

https://github.com/Alfrednoland/applied-data-science-capstone/blob/main/EDA.ipynb

Build an Interactive Map with Folium

- Markers, circles, lines and marker clusters were used with Folium Maps
- Markers indicate points like launch sites
- Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center
- Marker clusters indicates groups of events in each coordinate, like launches in a launch site
- Lines are used to indicate distances between two coordinates.

Source code:

https://github.com/Alfrednoland/applied-data-science-capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

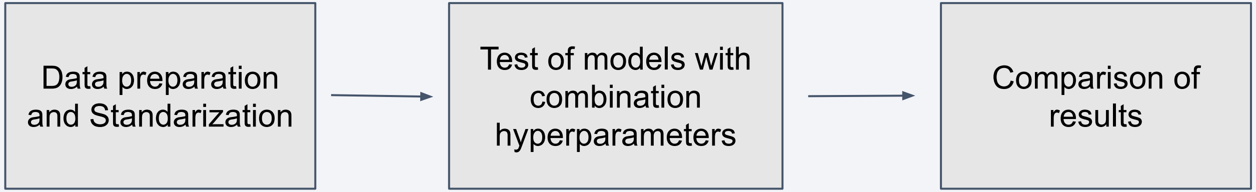
- The following graphs and plots were used to visualize data Percentage of launches by site and Payload range
- This combination enabled a rapid analysis of the relationship between payloads and launch sites, helping to identify the optimal launch locations based on the payloads.

Source code:

https://github.com/Alfrednoland/applied-data-science-capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Four classification models were compared: logistic regression, support vector machine, decision tree and k nearest neighbors.



Data preparation and Standarization

Test of models with combination hyperparameters

Comparison of results

Source code:

https://github.com/Alfrednoland/applied-data-science-capstone/blob/main/SpaceX_Machine%20Learning%20Prediction.ipynb

Results

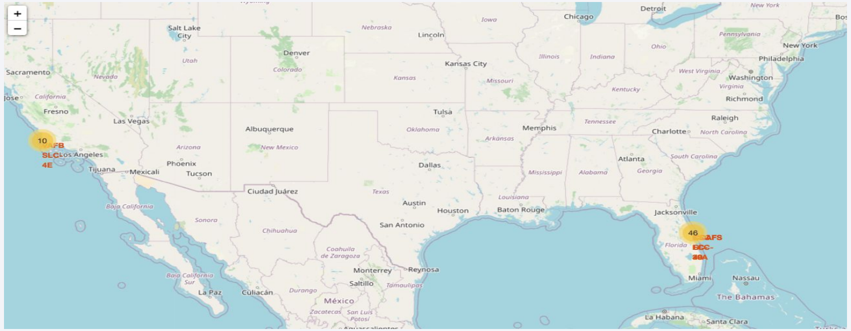
Exploratory data analysis results

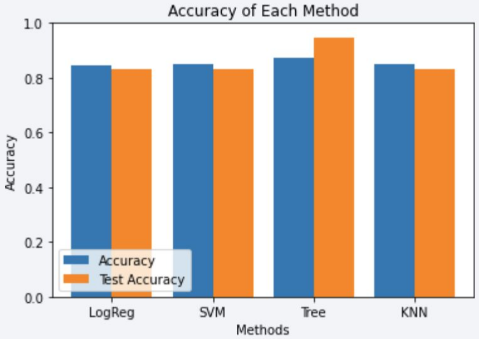
- Space X uses 4 different launch sites
- The first launches were done to Space X itself and NASA
- The average payload of F9 v1.1 booster is 2,928 kg
- The first success landing outcome happened in 2015 fiver year after the first launch
- Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average
- Almost 100% of mission outcomes were successful
- Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015
- The number of landing outcomes became as better as years passed

Results

Using interactive analytics was possible to identify that launch sites use to be in safety places, near sea, for example and have a good logistic infrastructure around.
Most launches happens at east cost launch sites.

Predictive Analysis showed that Decision Tree Classifier is the best model to predict successful landings, having accuracy over 87% and accuracy for test data over 94%.



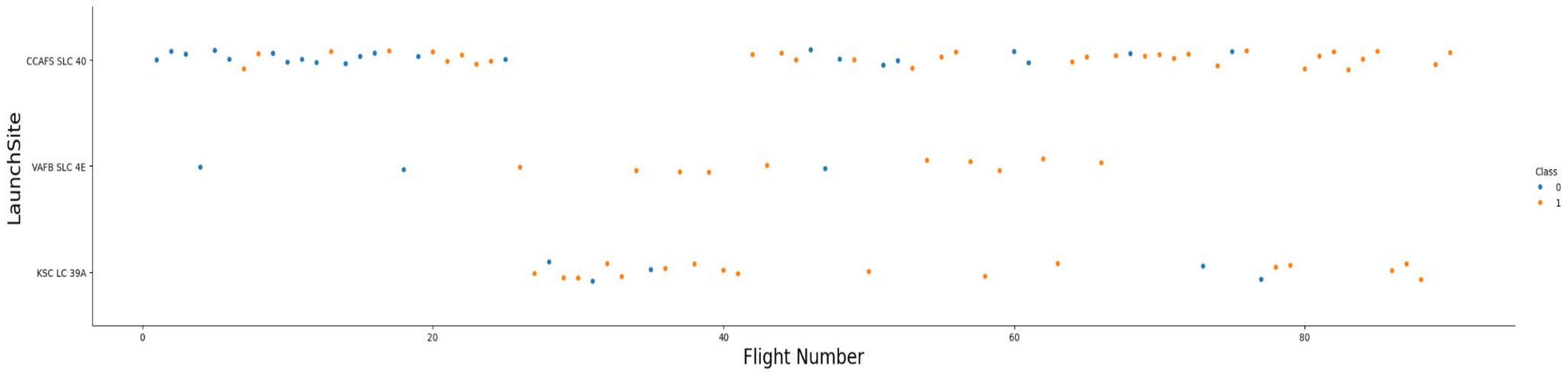


Section
2

Insights drawn from EDA

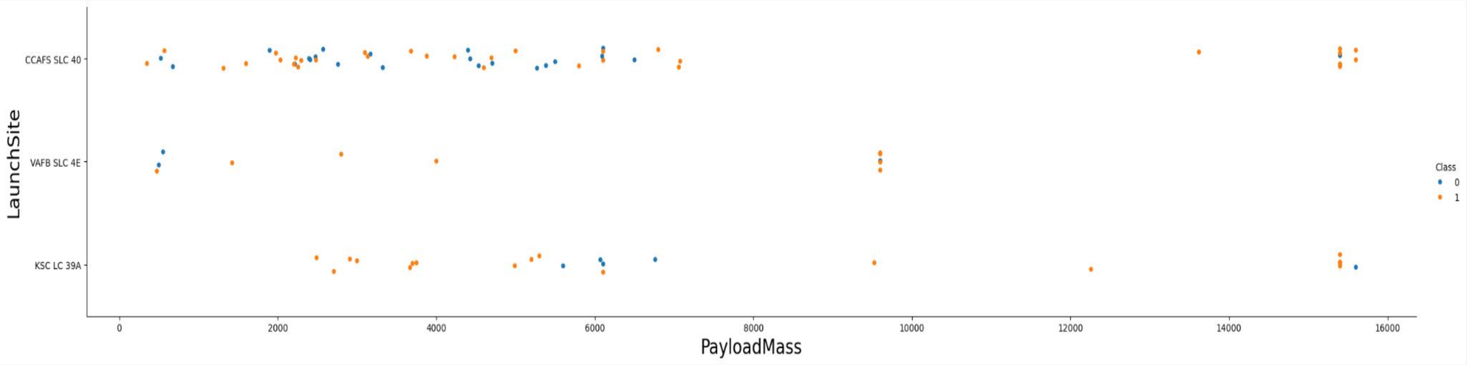
Flight Number vs. Launch Site

- According to the plot above, it’s possible to verify that the best launch site nowadays is CCAF5 SLC 40, where most of recent launches were successful
- In second place VAFB SLC 4E and third place KSC LC 39A
- It’s also possible to see that the general success rate improved over time



Payload vs. Launch Site

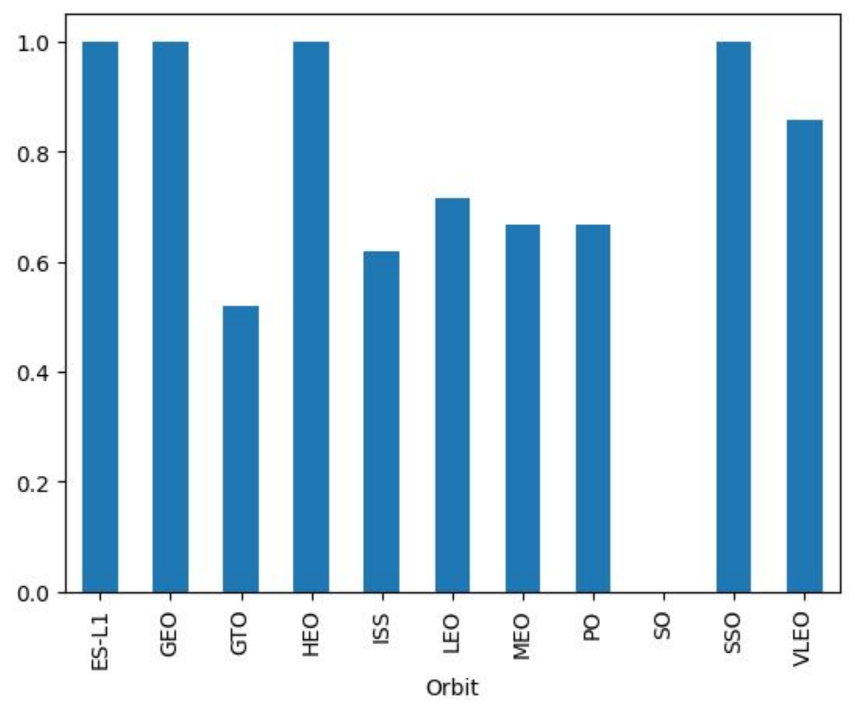
- Payloads over 9,000kg (about the weight of a school bus) have excellent success rate
- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites



Success Rate vs. Orbit Type

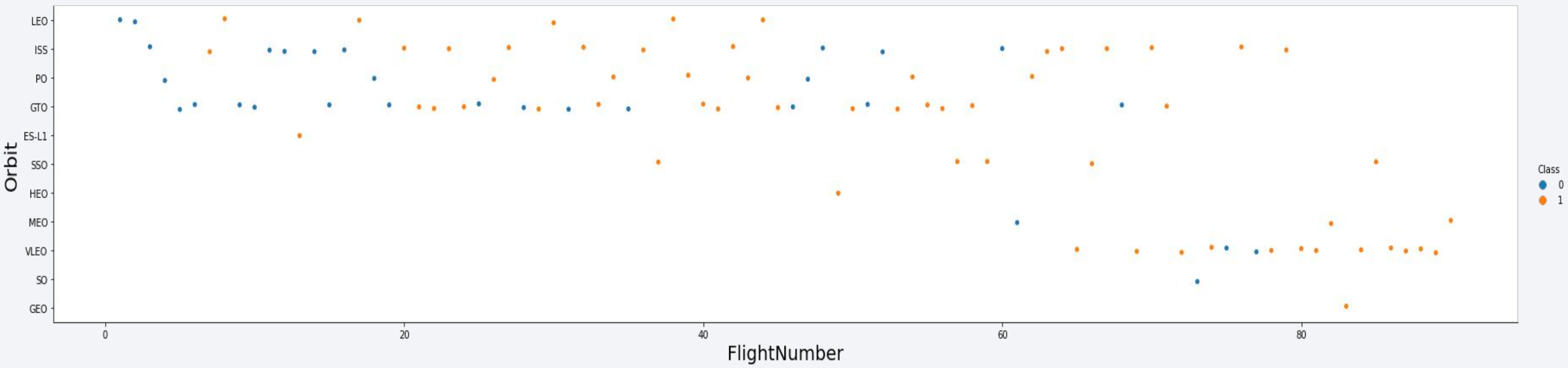
- The biggest success rates happens to orbits:
  - ES:L1
  - GEO
  - HEO
  - SSO

- Followed by
  - VLEO (above 80%)
  - LFO (above 70%)



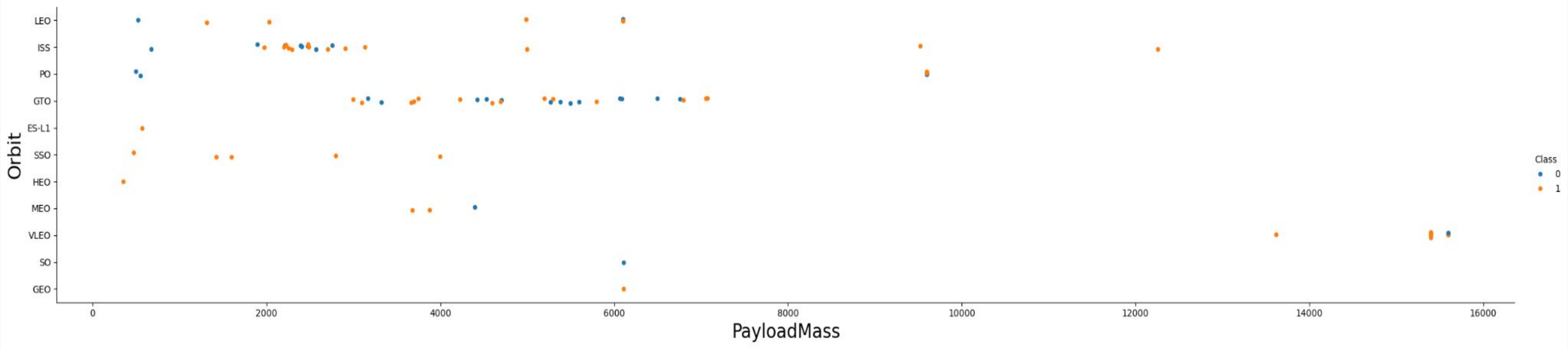
Flight Number vs. Orbit Type

- Success rate improved over time to all orbits
- VLEO orbit seems a new business opportunity, due to recent increase of its frequency



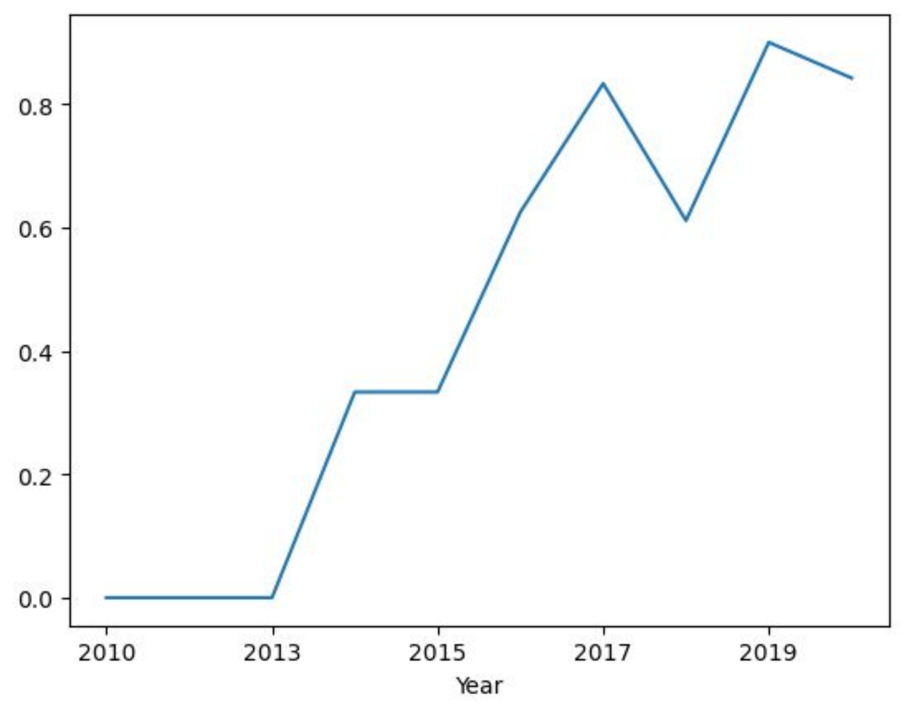
Payload vs. Orbit Type

- There is no relation between payload and success rate to orbit GTO
- ISS orbit has the widest range of payload and a good rate of success
- There are few launches to the orbits SO and GEO



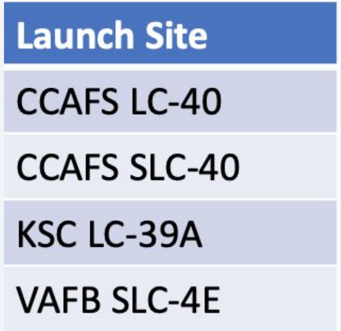
Launch Success Yearly Trend

- Success rate was starting to increase in 2013 until 2020.
- It seems that the first three years were a period of adjusts and improvement of technology.



All Launch Site Names

- There are four launch sites:

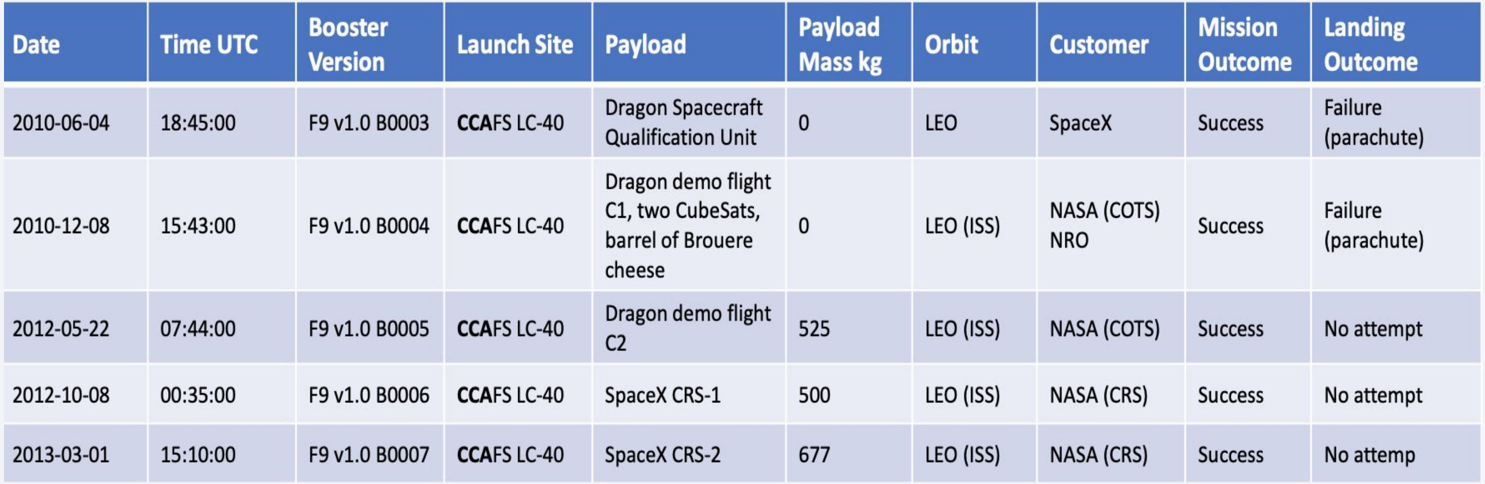


Launch Site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

- They are obtained by selecting unique occurrences of “launch site” values from the dataset.

Launch Site Names Begin with 'CCA'

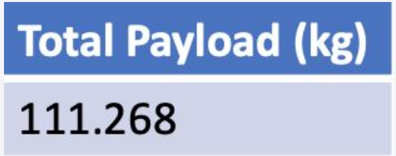
- Here we can see five samples of Cape Canaveral launches



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 attemp

Total Payload Mass

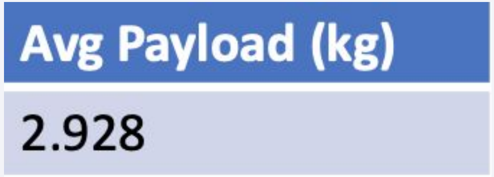
- Total payload calculated by summing all payloads whose codes contain CRS which corresponds to NASA.
- Total payload by NASA boosters:



Total Payload (kg)
111.268

Average Payload Mass by F9 v1.1

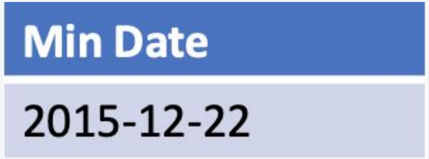
- Average payload mass carried by booster version F9 v1.1
- Filtering data by the booster version above and calculating the average payload mass we obtained the value of 2,928 kg.



Avg Payload (kg)
2.928

First Successful Ground Landing Date

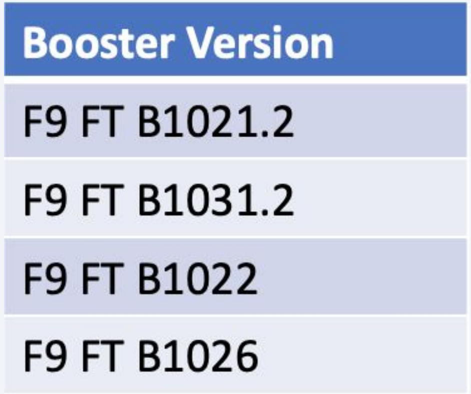
- First successful landing outcome on ground pad:
- By filtering data by successful landing outcome on ground pad and getting the minimum value for date it’s possible to identify the first occurrence, that happened on 12/22/2015.



Min Date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

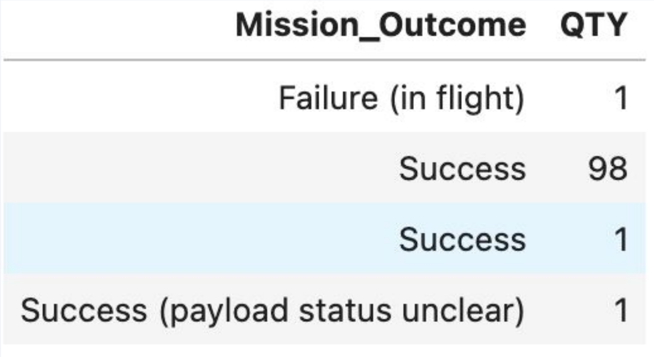
- Boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- Selecting distinct booster versions according to the filters above, these 4 are the result.



Booster Version
F9 FT B1021.2
F9 FT B1031.2
F9 FT B1022
F9 FT B1026

Total Number of Successful and Failure Mission Outcomes

- Results:

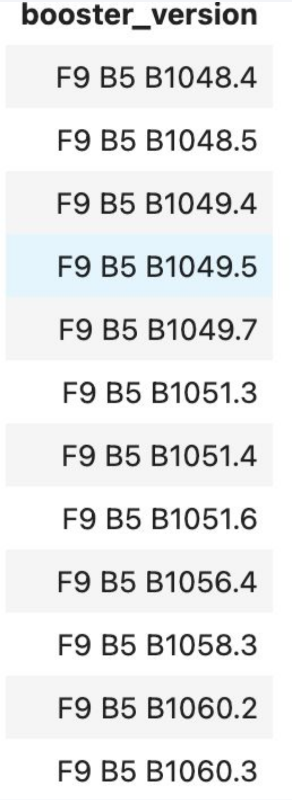


Mission_Outcome	QTY
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Grouping mission outcomes and counting records for each group led us to the summary above.

Boosters Carried Maximum Payload

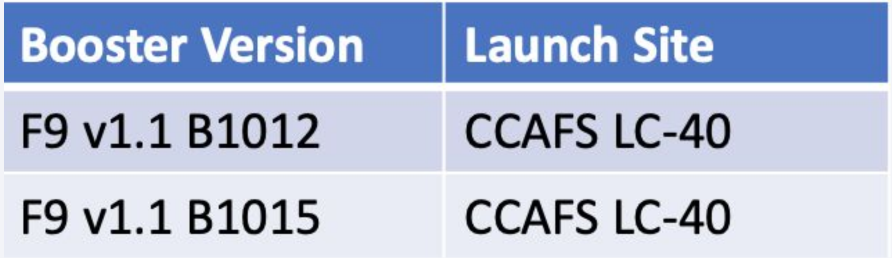
- These are th booster which carried the most payload mass



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

2015 Launch Records

Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015 with only two cccurences:



Booster Version	Launch Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

- Results of Landing outcomes:

Landing Outcome	Occurrences
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

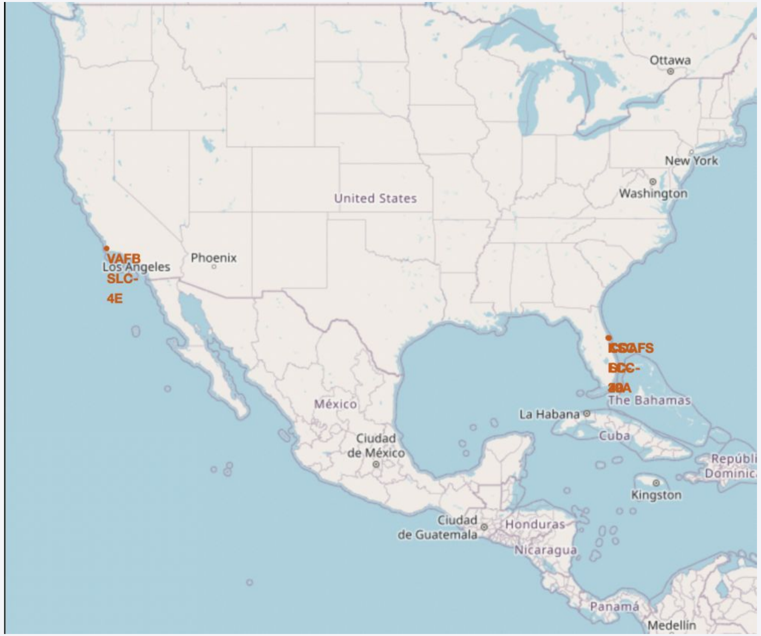
- Ranking the count of landing outcomes such as Failure or Success between the dates 2010-06-04 and 2017-03-20

Section
3

Launch Sites Proximities Analysis

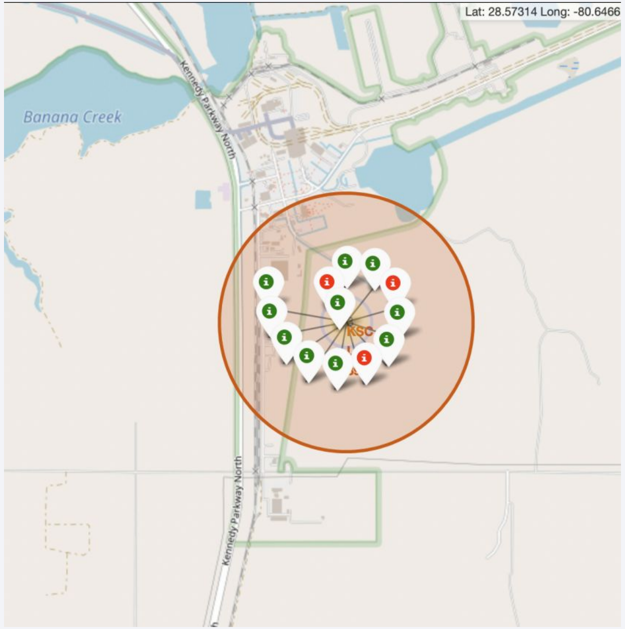
Launch sites location markers

- Close to equator line because the earth is moving at a higher pace.
- They are in close proximity to the coast, while launching rockets towards the ocean it minimises the risk of having debris or exploding exploding near people.



Color-labeled launch outcomes

- Here we can identify the lauch outcomes by the markers on the map:
- Green markers = Successful
- Red markers = Failures
- Launch site KSC LC-39A has high success rate for launches.

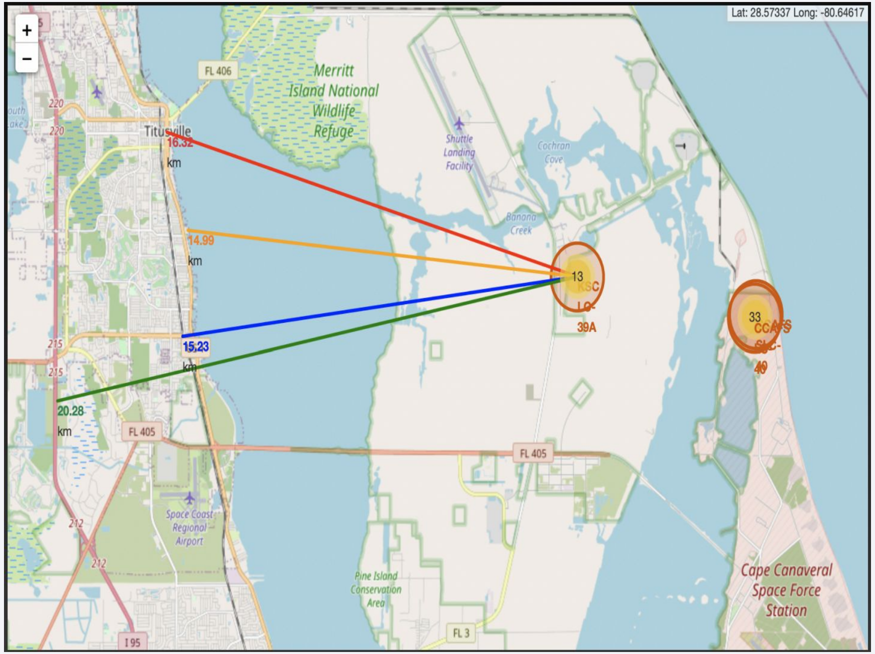


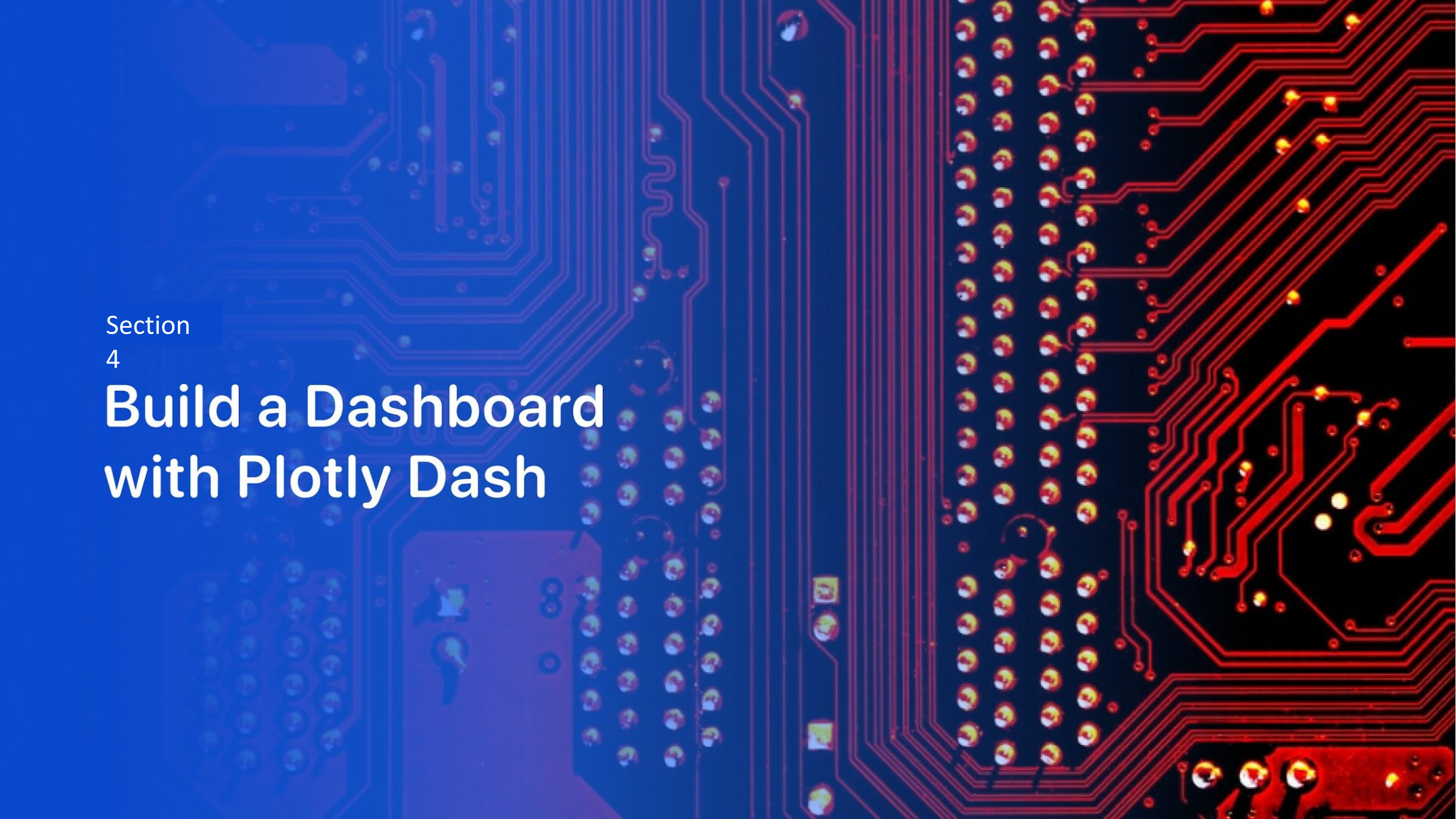
Distances - Launch site KSC LC-39A

From the visual analysis of the launch site KSC LC-39A, we can clearly observe the following:

- It is relatively close to a railway (15.23 km).
- It is relatively close to a highway (20.28 km).
- It is relatively close to the coastline (14.99 km).
- Additionally, the launch site is relatively close to the nearest city, Titusville (16.32 km).

A failed rocket, traveling at high speed, could cover distances of 15-20 km in a matter of seconds. This proximity could pose a potential danger to populated areas.

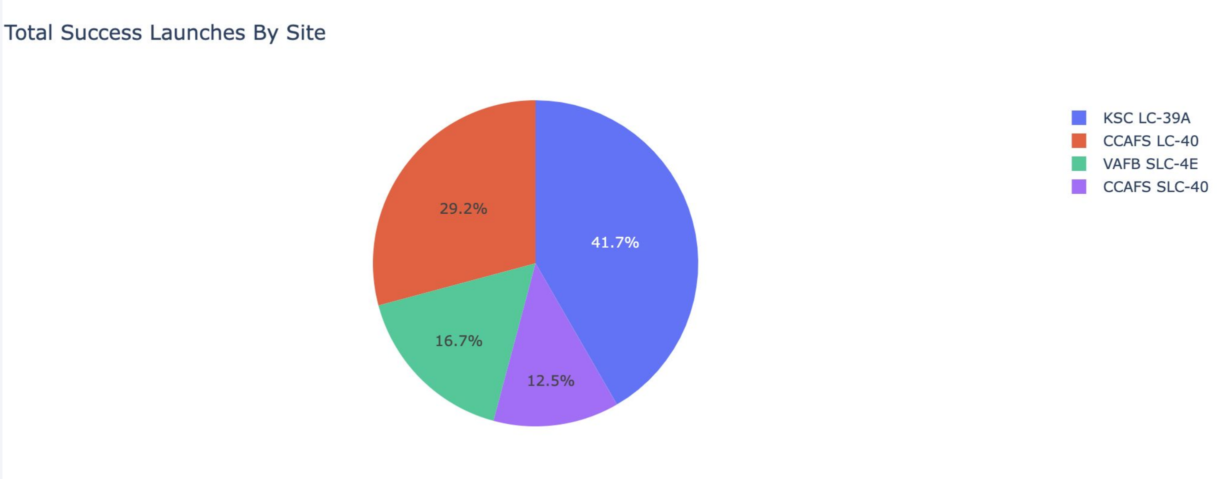




Section
4

Build a Dashboard with Plotly Dash

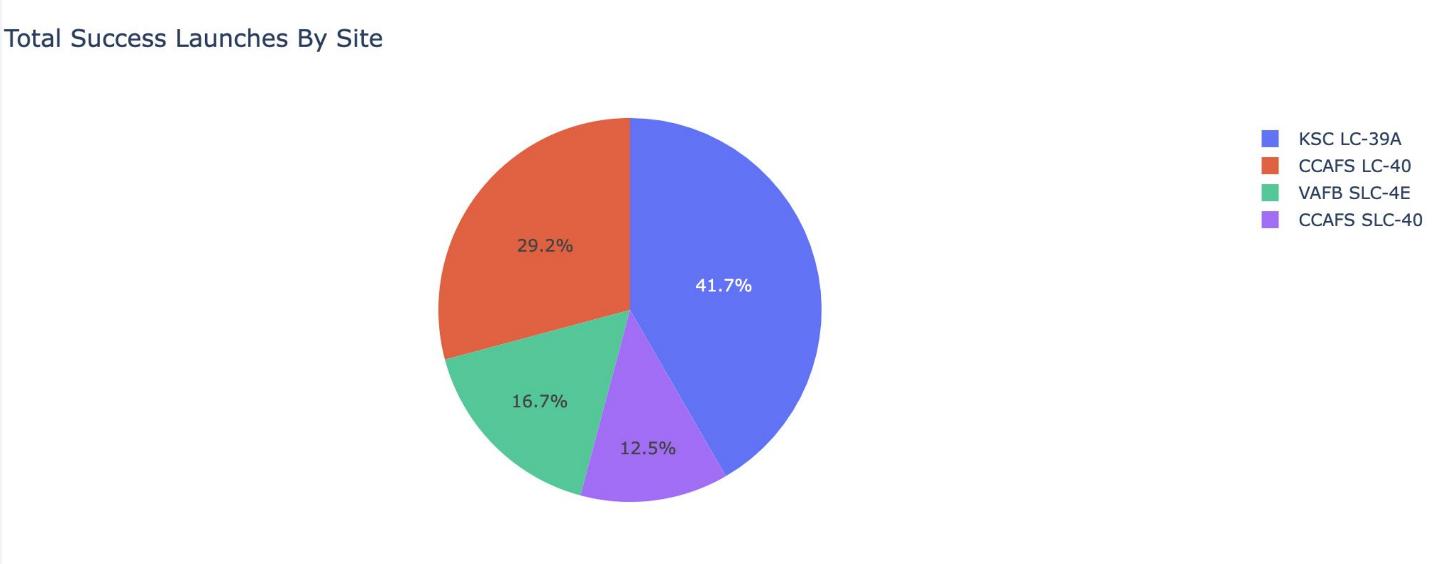
Launch success



Explanation:

- The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches.

Launch site with highest success ratio



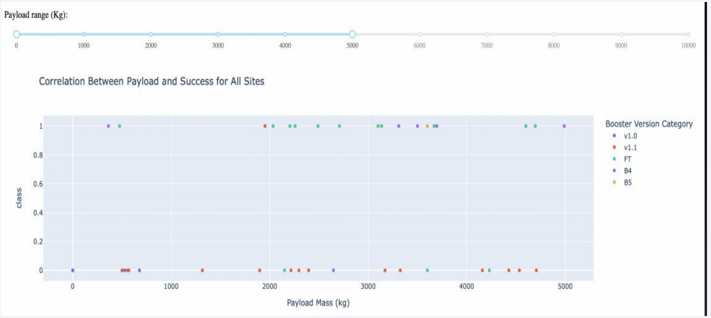
Explanation:

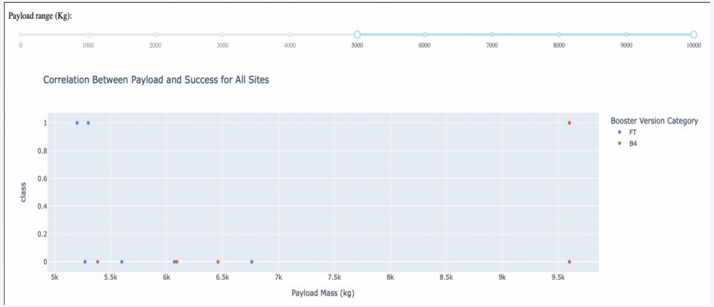
- KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings.

Payload Mass vs Launch Outcome

Explanation:

- The charts show that payloads between 2000 and 5500 kg have the highest success rate.



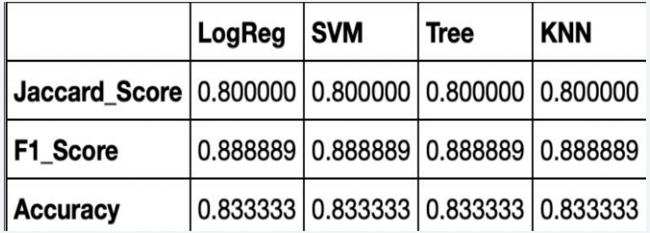


Section
5

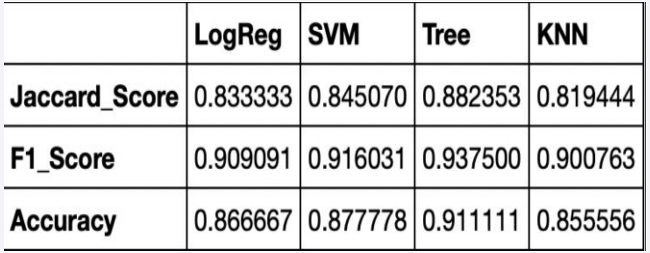
Predictive Analysis (Classification)

Classification Accuracy

- Based on the scores of the Test Set, we can not confirm which method performs best.
- Same Test Set scores may be due to the small test sample size (18 samples). Therefore, we tested all methods based on the whole Dataset.
- The scores of the whole Dataset confirm that the best model is the Decision Tree Model. This model has not only higher scores, but also the highest accuracy.



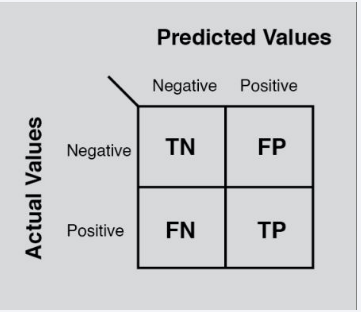
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333



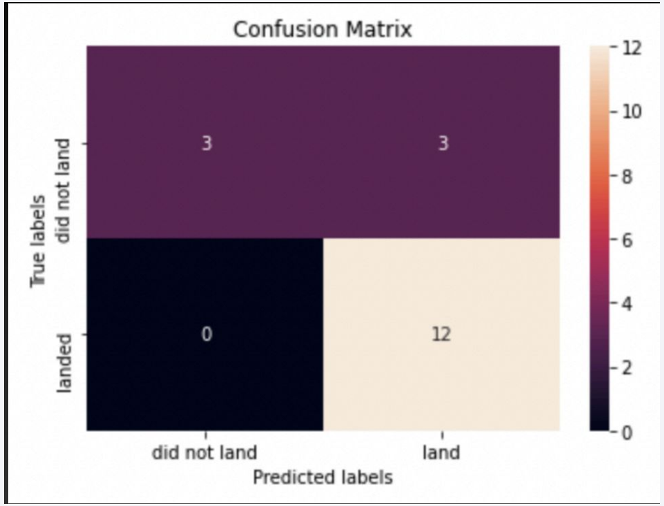
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

Confusion Matrix

- Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the major problem is false positives.



		Predicted Values	
		Negative	Positive
Actual Values	Negative	TN	FP
	Positive	FN	TP



Conclusions

- Decision Tree Model is the best Algorithm for this dataset.
- Launches with a low payload mass show better results than launches with a larger payload mass.
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- The success rate of launches increases over the years.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.

Appendix

- Source code: https://github.com/Alfrednoland/applied-data-science-capstone/tree/main

Thank you!