IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ricardo Sánchez Sánchez
01/11/2022

Outline

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

Executive Summary

Summary of methodologies

- Data collection
- Data wrangling
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly Dash
- Predictive analysis (Classification)

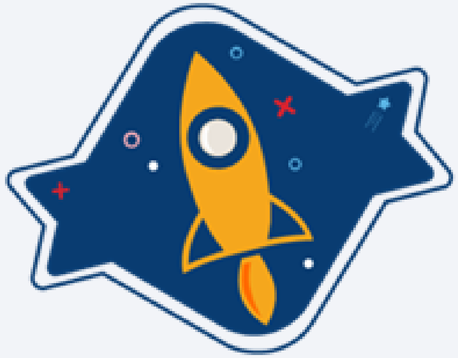
Summary of all results

- It was possible to collected valuable data from public sources;
- EDA allowed to identify which features are the best to predict success of launchings;
- Machine Learning Prediction showed the best model to predict which characteristics are important to drive this opportunity by the best way, using all collected data.

Introduction

Project background and context

- The objective is to evaluate the viability of the new company Space Y to compete with Space X.



Problems you want to find answers

- 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 landings increase over the years?
- What is the best algorithm that can be used for binary classification in this case?

Section 1

Methodology

Methodology

Executive Summary

Data collection methodology:

- Data from Space X was obtained from the following sources:
  - https://api.spacexdata.com/v4/rockets/
  - https://api.spacexdata.com/v4/launches/past
  - https://api.spacexdata.com/v4/launchpads/
  - https://api.spacexdata.com/v4/payloads/
  - https://api.spacexdata.com/v4/cores/
- WebScraping
  - (https://en.wikipedia.org/wiki/List_of_Falcon_/_9/_and_Falcon_Heavy_launches)

Perform data wrangling:

- Filtering the data
- Dealing with missing values
- Using One Hot Encoding to prepare the data to a binary classification

Methodology

Executive Summary

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data that was collected until this step were normalized, divided in training and test data sets and evaluated by four different classification models, being the accuracy of each model evaluated using different combinations of parameters.

Data Collection

Data collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX's Wikipedia entry. We had to use both data collection methods in order to get complete information about the launches for a more detailed analysis.

Data Columns are obtained by using SpaceX REST API:

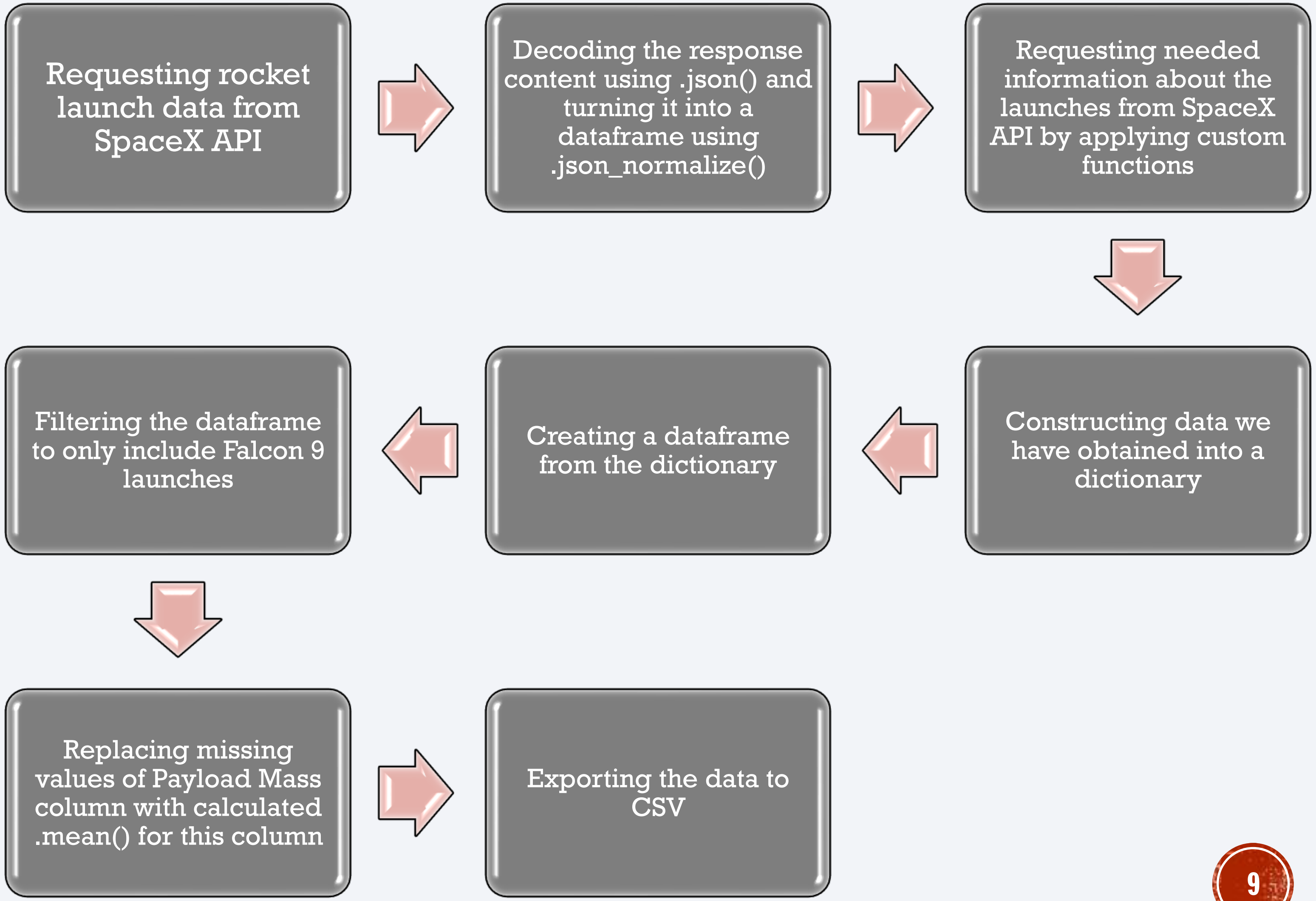
FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude.

Data Columns are obtained by using Wikipedia Web Scraping:

Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time.

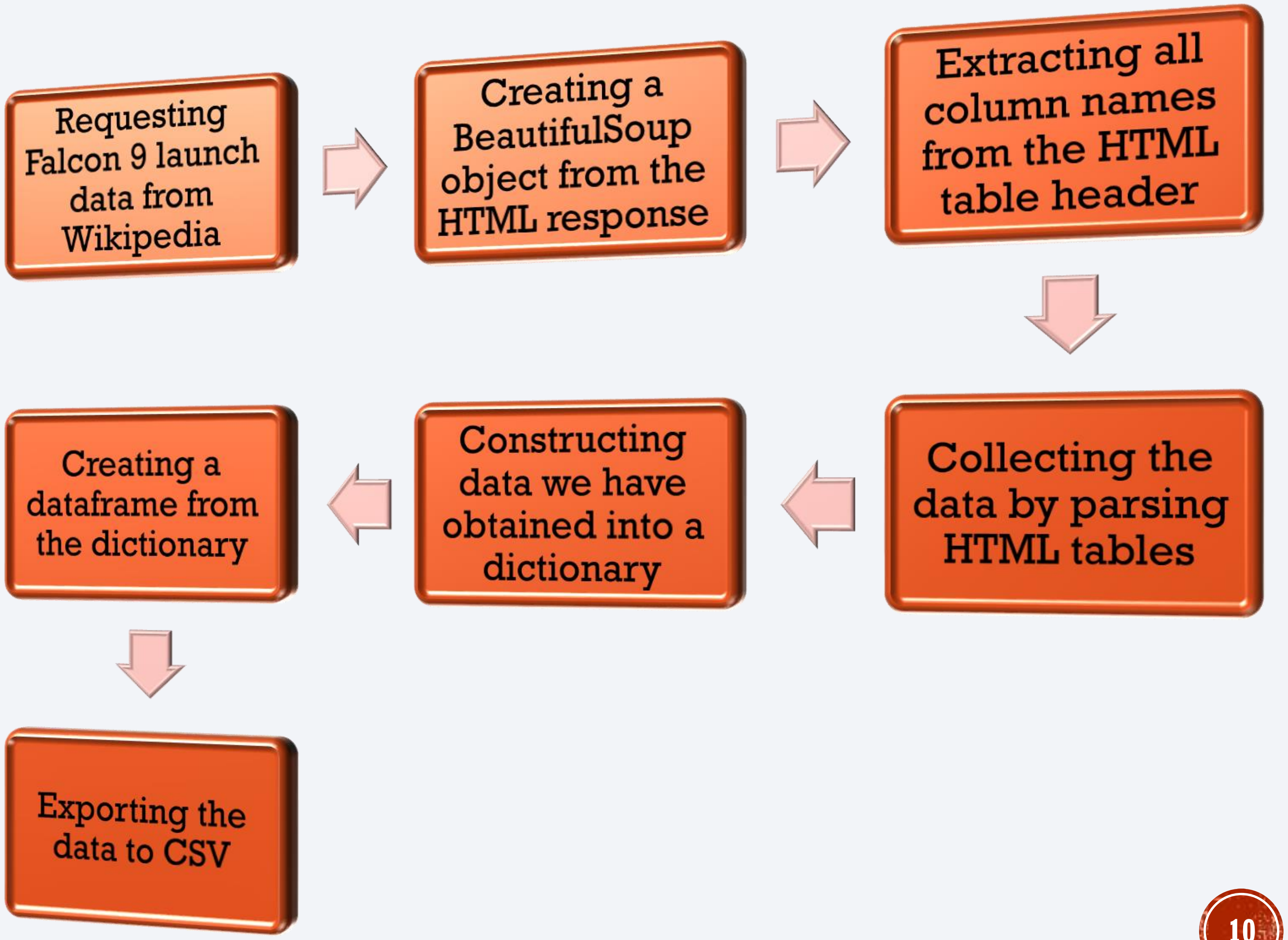
Data Collection – SpaceX API

- SpaceX offers a public API from where data can be obtained and then used.
- This API was used according to the flowchart beside and then data is persisted.
- Source code:
  - https://github.com/ricss125/Applied-Data-Science-Capstone/blob/main/DataCollectionAPI.ipynb



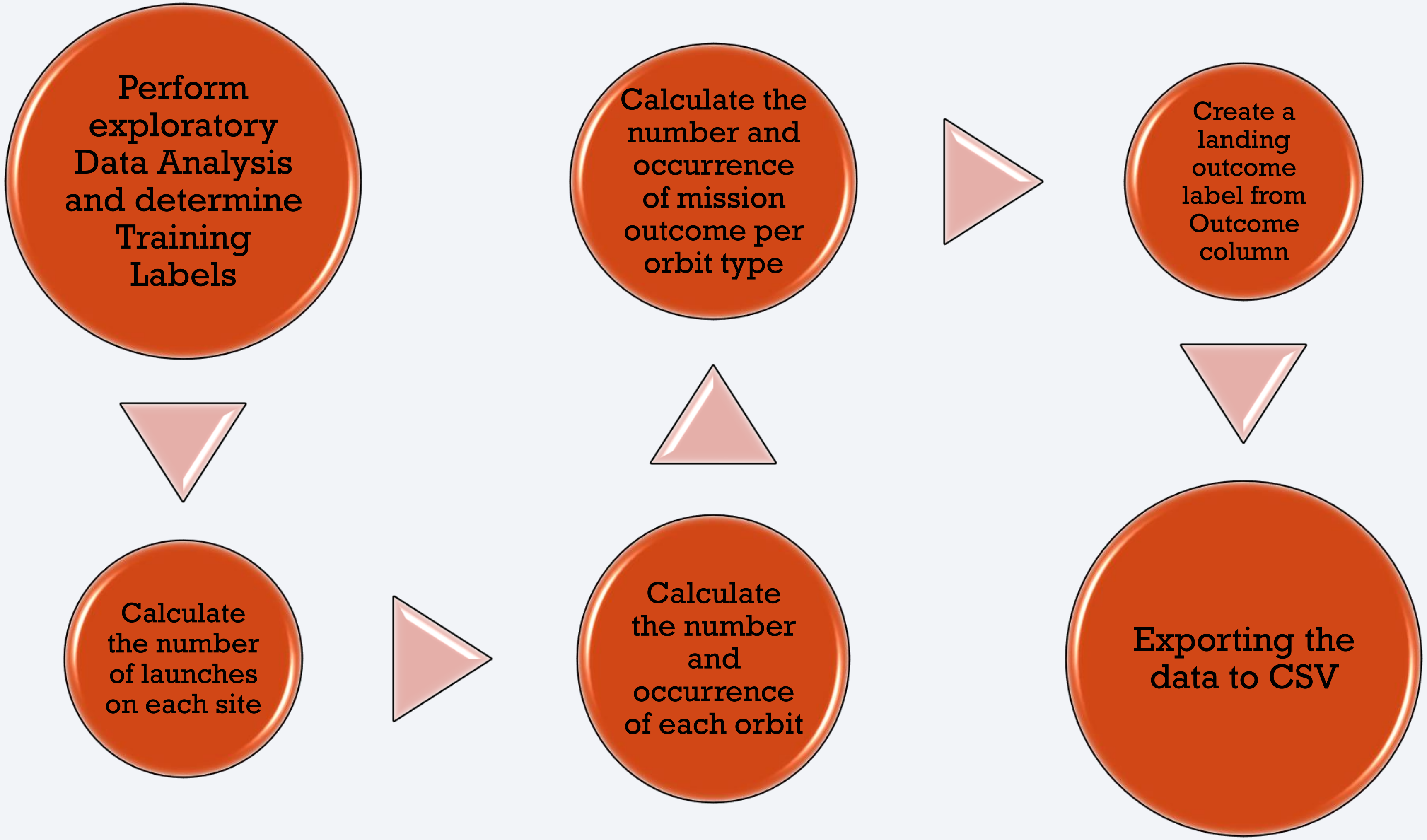
Data Collection - Scraping

- Data from SpaceX launches can also be obtained from Wikipedia
- Data are downloaded from Wikipedia according to the flowchart and then persisted.
- Source Code:
  - https://github.com/ricss125/Applied-Data-Science-Capstone/blob/main/DataCollectionWithWebScraping.ipynb



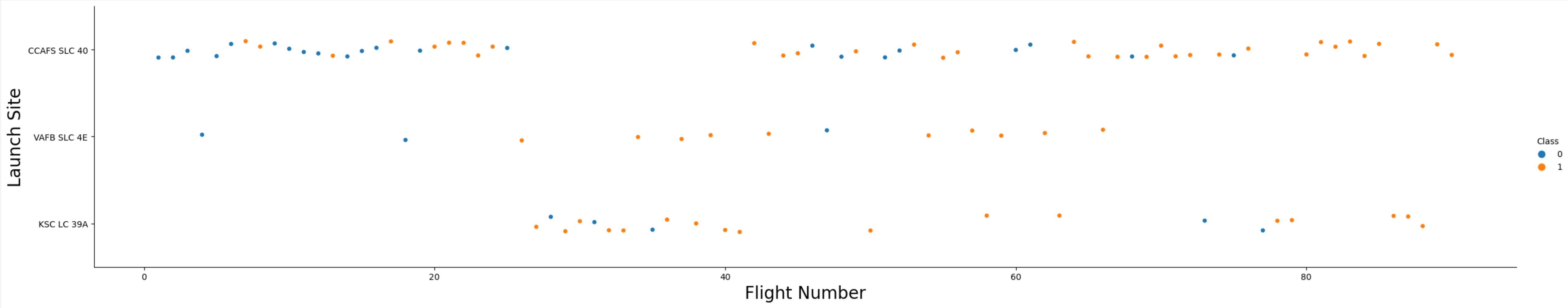
Data Wrangling

- Initially some Exploratory Data Analysis (EDA) was performed on the dataset.
- Then the summaries launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type were calculated.
- Finally, the landing outcome label was created from Outcome column.
- Source Code:
  - https://github.com/ricss125/Applied-Data-Science-Capstone/blob/main/DataWrangling.ipynb



EDA with Data Visualization

- To explore data, scatterplots and barplots were used to visualize the relationship between pair of features:
  - Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs Orbit Type and Success Rate Yearly Trend



- Source Code: https://github.com/ricss125/Applied-Data-Science-Capstone/blob/main/EDA_DataVisualization.ipynb

EDA with SQL

Performed SQL queries:

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string 'CCA'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in ground pad was achieved
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which have carried the maximum payload mass
- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
- Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20 in descending order
- Source Code: https://github.com/ricss125/Applied-Data-Science-Capstone/blob/main/EDA_with_SQL.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/ricss125/Applied-Data-Science-Capstone/blob/main/Visual_Analytics_with_Folium.ipynb

Build a Dashboard with Plotly Dash

- *Launch Sites Dropdown List:*
  - ✓Added a dropdown list to enable Launch Site selection.
- *Pie Chart showing Success Launches (All Sites/Certain Site):*
  - ✓Added a pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected.
- *Slider of Payload Mass Range:*
  - ✓Added a slider to select Payload range.
- *Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions:*
  - ✓Added a scatter chart to show the correlation between Payload and Launch Success.

Source Code: https://github.com/ricss125/Applied-Data-Science-Capstone/blob/main/Interactive_Dashboard_with_Ploty_Dash.py

Predictive Analysis (Classification)

- Four classification models were compared: logistic regression, support vector machine, decision tree and k nearest neighbors.
- Source Code:
  - https://github.com/ricss125/Applied-Data-Science-Capstone/blob/main/Machine_Learning_Prediction.ipynb

Creating a NumPy array from the column "Class" in data

Standardizing the data with StandardScaler, then fitting and transforming it

Splitting the data into training and testing sets with train_test_split function

Creating a GridSearchCV object with cv = 10 to find the best parameters

Applying GridSearchCV on LogReg, SVM, Decision Tree, and KNN models

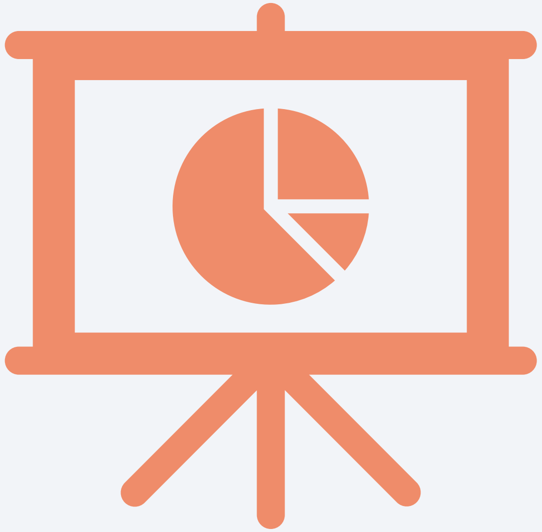
Calculating the accuracy on the test data using the method .score() for all models

Examining the confusion matrix for all models

Finding the method performs best by examining the Jaccard_score and F1_score metrics

Results

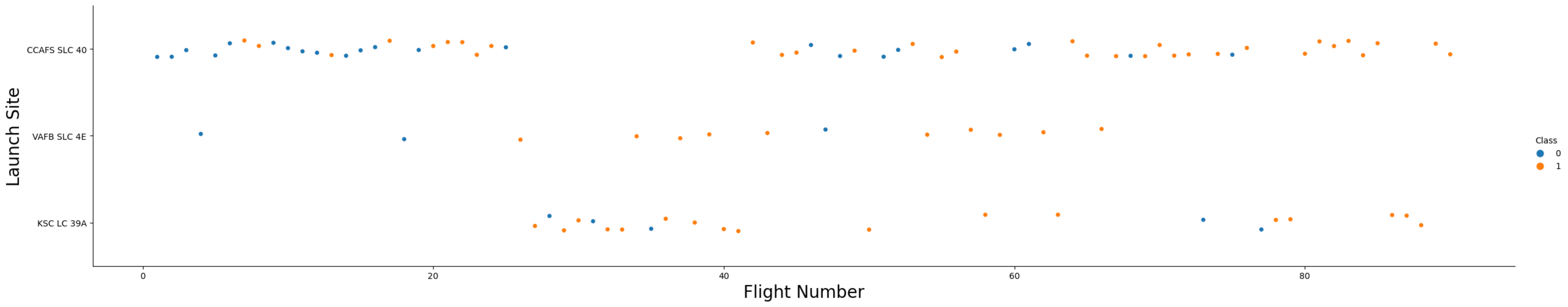
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results



Section 2

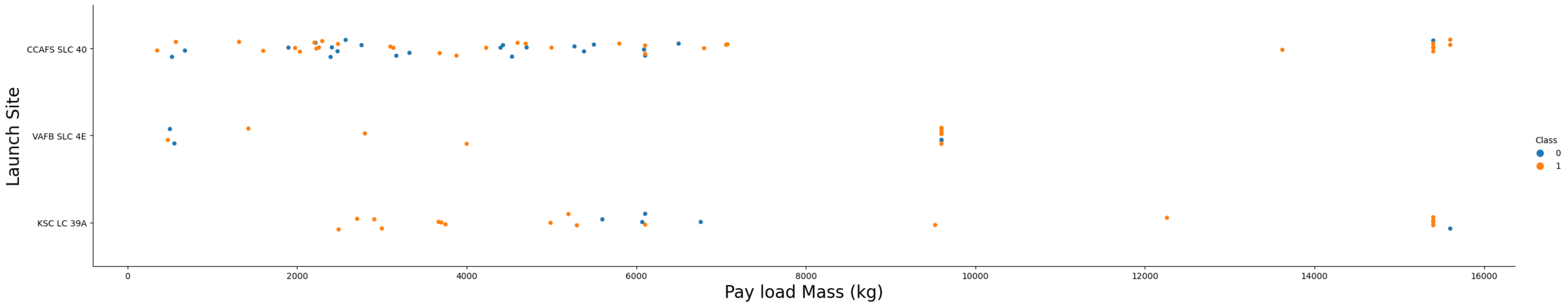
Insights drawn from EDA

Flight Number vs. Launch Site



- Explanation:
  - ❖The earliest flights all failed while the latest flights all succeeded.
  - ❖The CCAFS SLC 40 launch site has about a half of all launches.
  - ❖VAFB SLC 4E and KSC LC 39A have higher success rates.
  - ❖It can be assumed that each new launch has a higher rate of success.

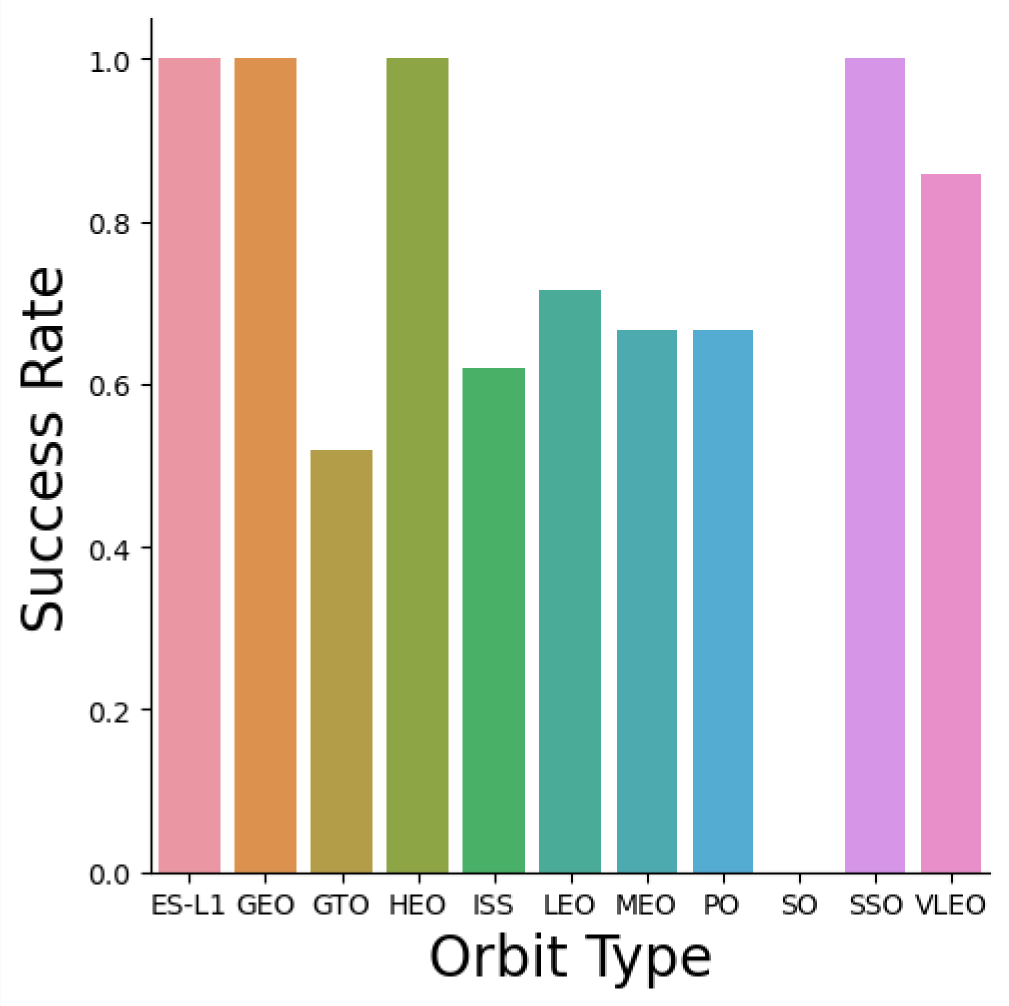
Payload vs. Launch Site



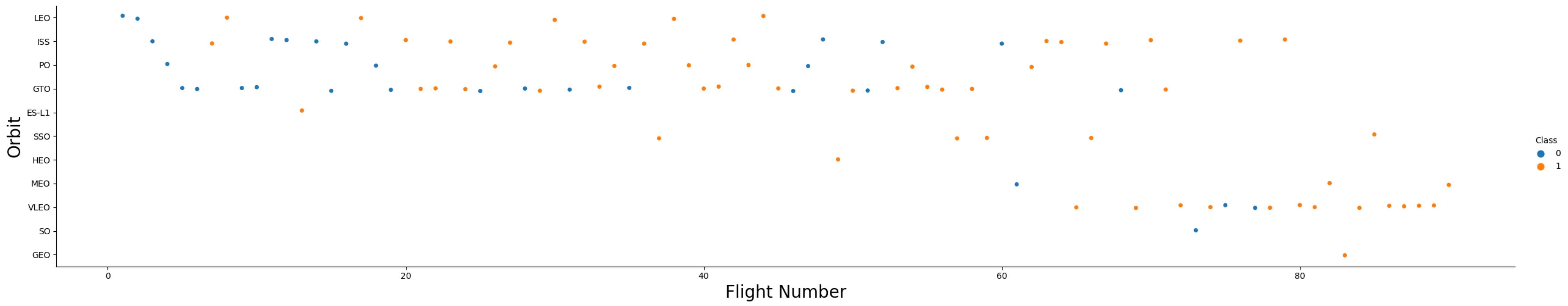
- Explanation:
  - ❖For every launch site the higher the payload mass, the higher the success rate.
  - ❖Most of the launches with payload mass over 7000 kg were successful.
  - ❖KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.

Success Rate vs. Orbit Type

- Explanation:
  - ➢Orbits with 100% success rate:
    - ❖ES-L1, GEO, HEO, SSO
  - ➢Orbits with 0% success rate:
    - ❖SO
  - ➢Orbits with success rate between 50% and 85%:
    - ❖GTO, ISS, LEO, MEO, PO



Flight Number vs. Orbit Type



Explanation

- Apparently, 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

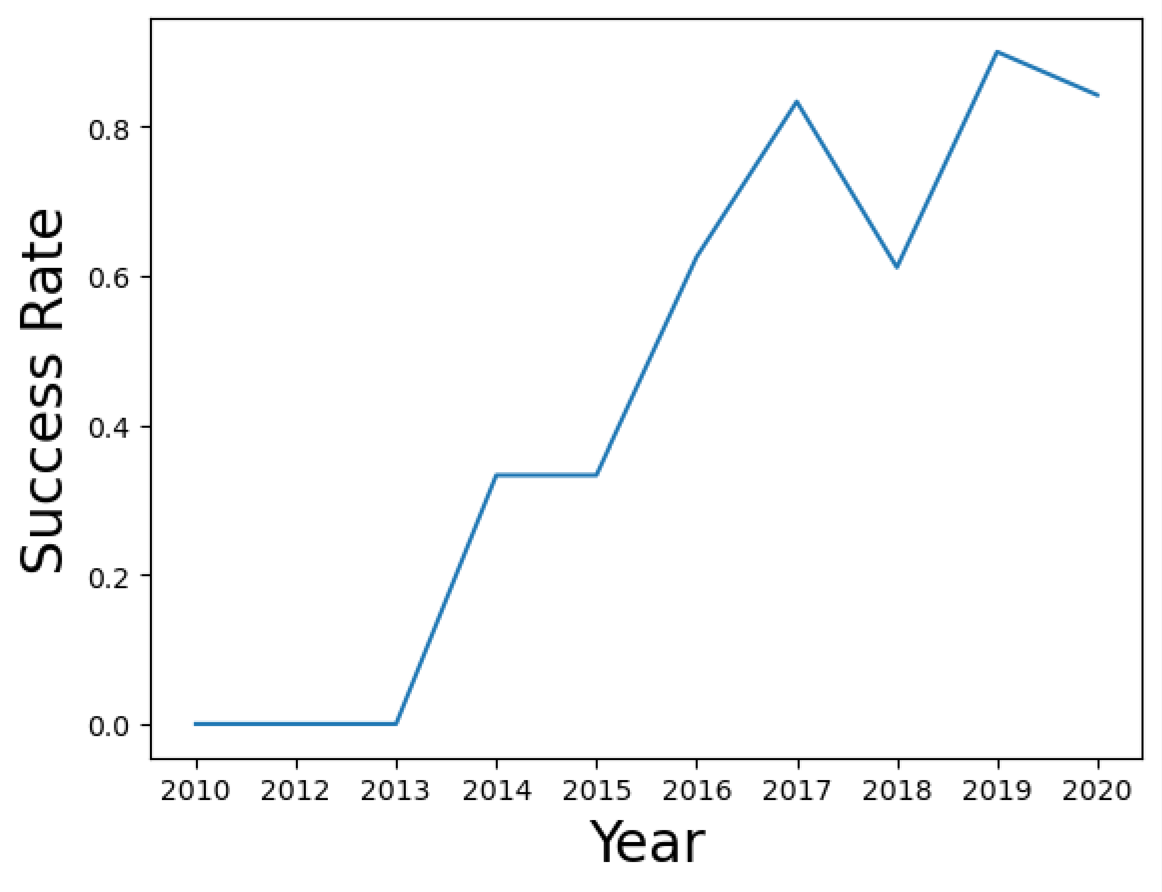
Explanation

- Apparently, 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

Explanation

- Success rate started increasing in 2013 and kept until 2020.
- It seems that the first three years were a period of adjusts and improvement of technology.



All Launch Site Names

- According to data, there are four launch sites:

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

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

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

- Displaying 5 records where launch sites begin with the string 'CCA'.

Total Payload Mass

- Total payload carried by boosters from NASA:

total_payload_mass
45596

- Total payload calculated above, by summing all payloads whose codes contain ‘ CRS’, which corresponds to NASA.

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1:

average_payload_mass
2534.6666666666665

- Filtering data by the booster version above and calculating the average payload mass we obtained the value of 2534.666 kg.

First Successful Ground Landing Date

- First successful landing outcome on ground pad:

first_successful_landing
2015-12-22

- 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.

Successful Drone Ship Landing with Payload between 4000 and 6000

- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.

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

Total Number of Successful and Failure Mission Outcomes

- Number of successful and failure mission outcomes:

Mission_Outcome	total_number
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

- Listing the names of the booster versions which have carried the maximum payload mass.

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

- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015.

MONTH	DATE	booster_version	launch_site	landing__outcome
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

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

landing__outcome	count_outcomes
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

- This view of data alerts us that “No attempt” must be taken in account.

Section 3

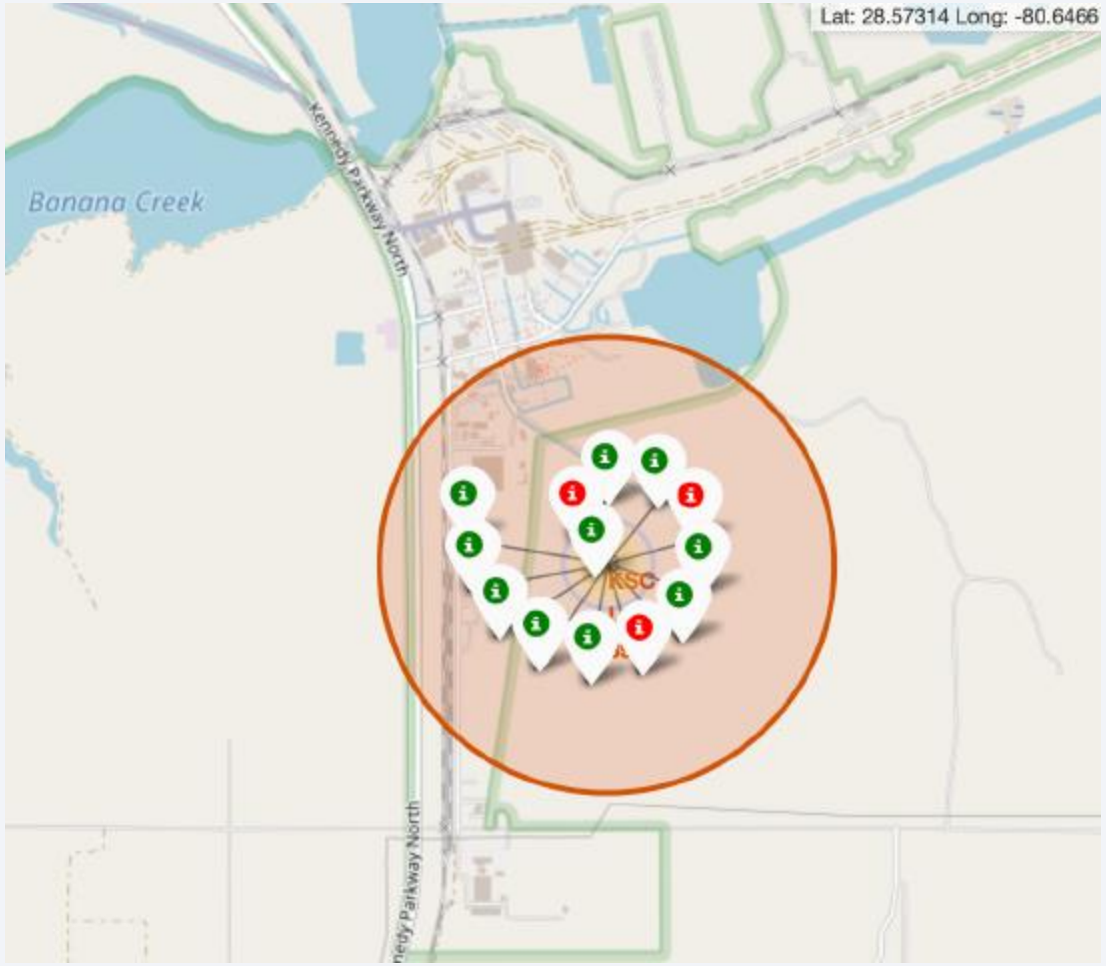
Launch Sites Proximities Analysis

All launch site's location markers on a global map

- Most of Launch sites are in proximity to the Equator line. The land is moving faster at the equator than any other place on the surface of the Earth. Anything on the surface of the Earth at the equator is already moving at 1670 km/hour. If a ship is launched from the equator it goes up into space, and it is also moving around the Earth at the same speed it was moving before launching. This is because of inertia. This speed will help the spacecraft keep up a good enough speed to stay in orbit.
- All launch sites are in very close proximity to the coast, while launching rockets towards the ocean it minimizes the risk of having any debris dropping or exploding near people.

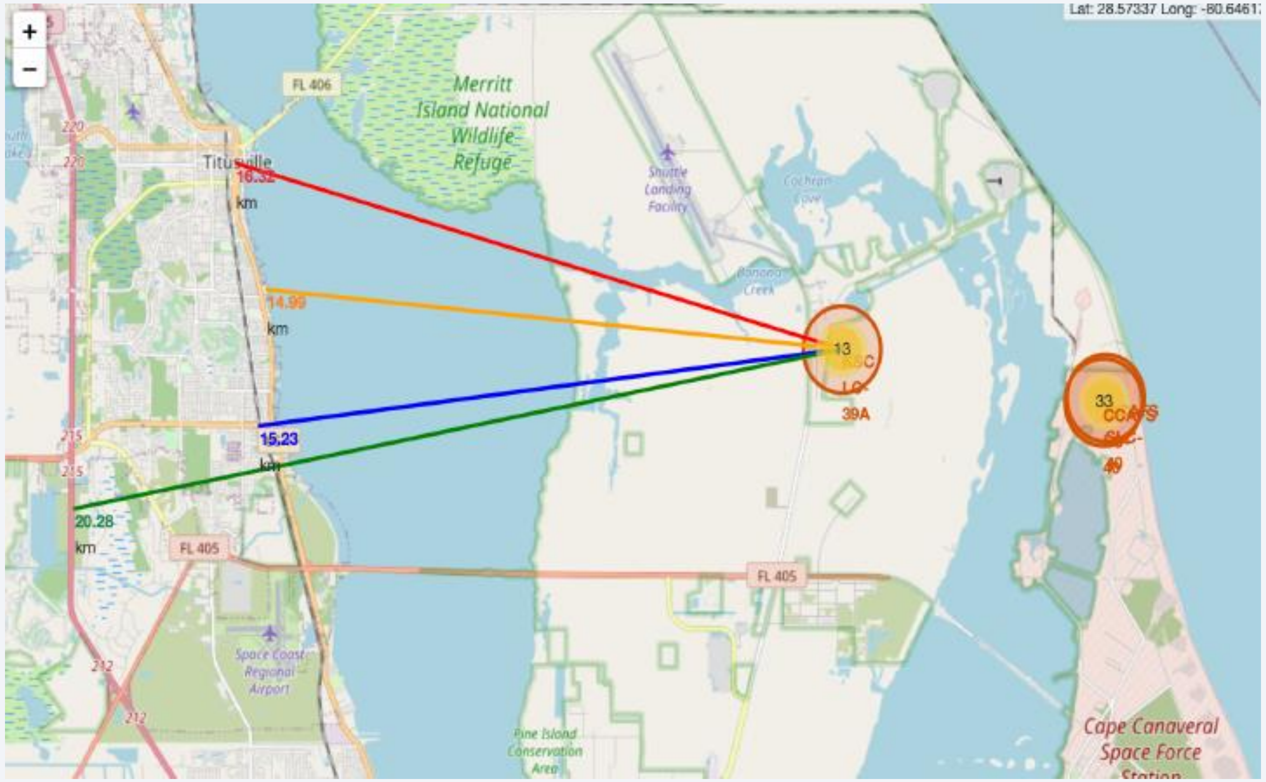
Launch records by Site

- From the color-labeled markers we should be able to easily identify which launch sites have relatively high success rates.
  - ➢ Green Marker = Successful Launch
  - ➢ Red Marker = Failed Launch
- Launch Site KSC LC-39A has a very high Success Rate.



Launch site KSC LC-39A Proximities

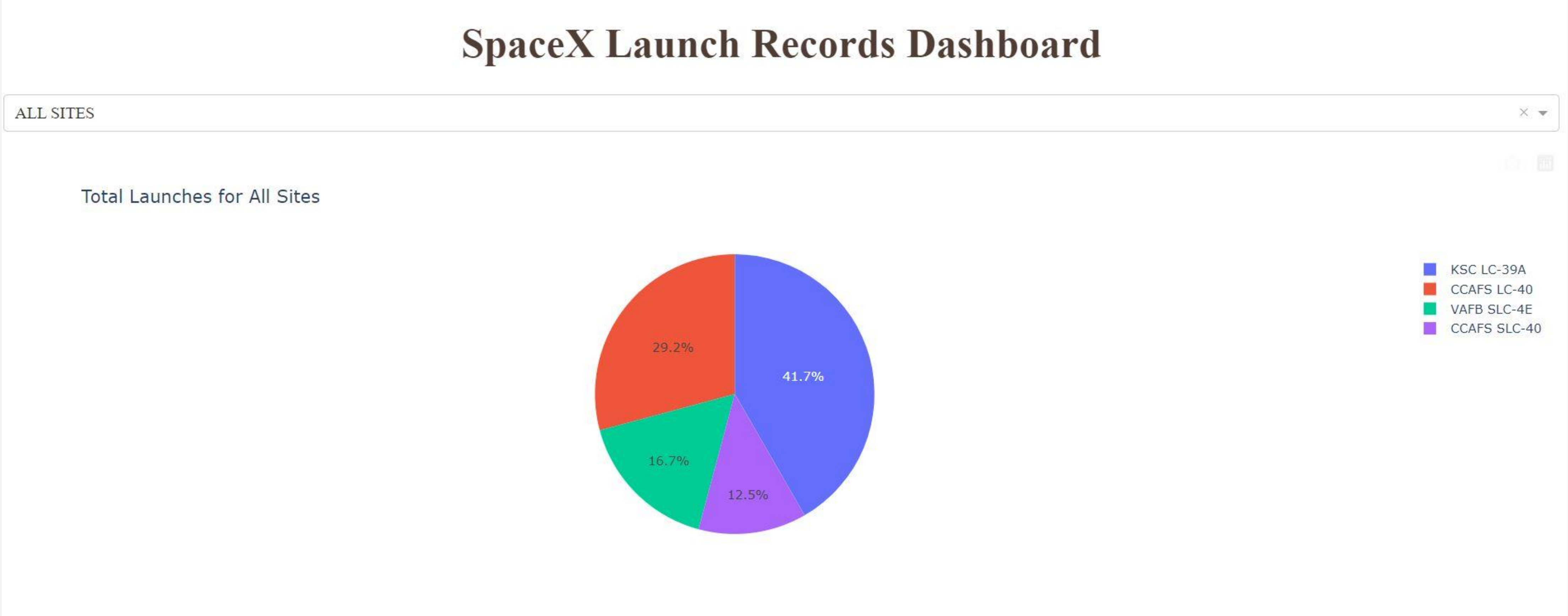
- From the visual analysis of the launch site KSC LC-39A we can clearly see that it is:
  - ❑relatively close to railway (15.23 km)
  - ❑relatively close to highway (20.28 km)
  - ❑relatively close to coastline (14.99 km)
- Also, the launch site KSC LC-39A is relatively close to its closest city Titusville (16.32 km).
- Failed rocket with its high speed can cover distances like 15-20 km in few seconds. It could be potentially dangerous to populated areas.



Section 4

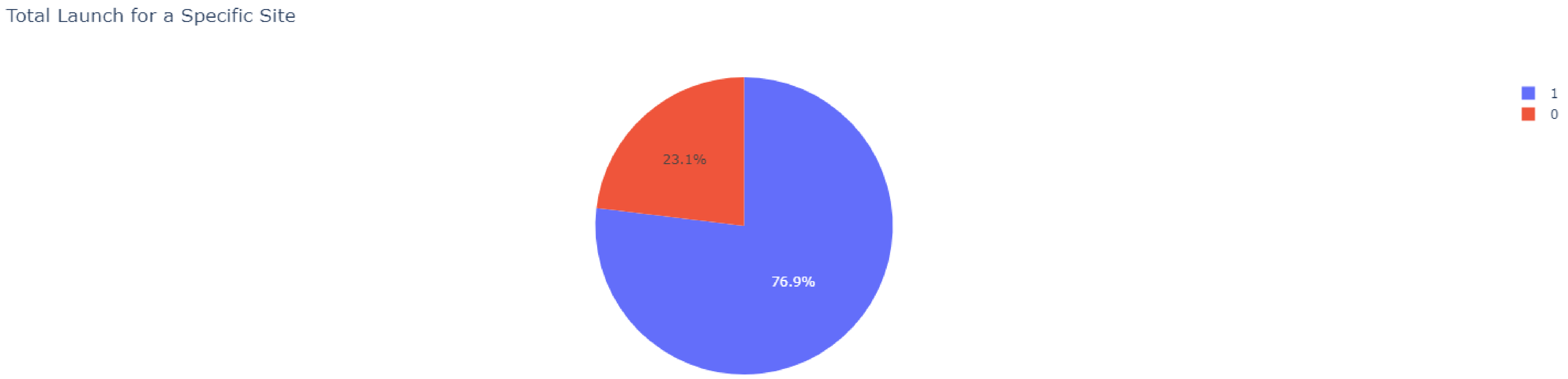
Build a Dashboard with Plotly Dash

Successful Launches by Site



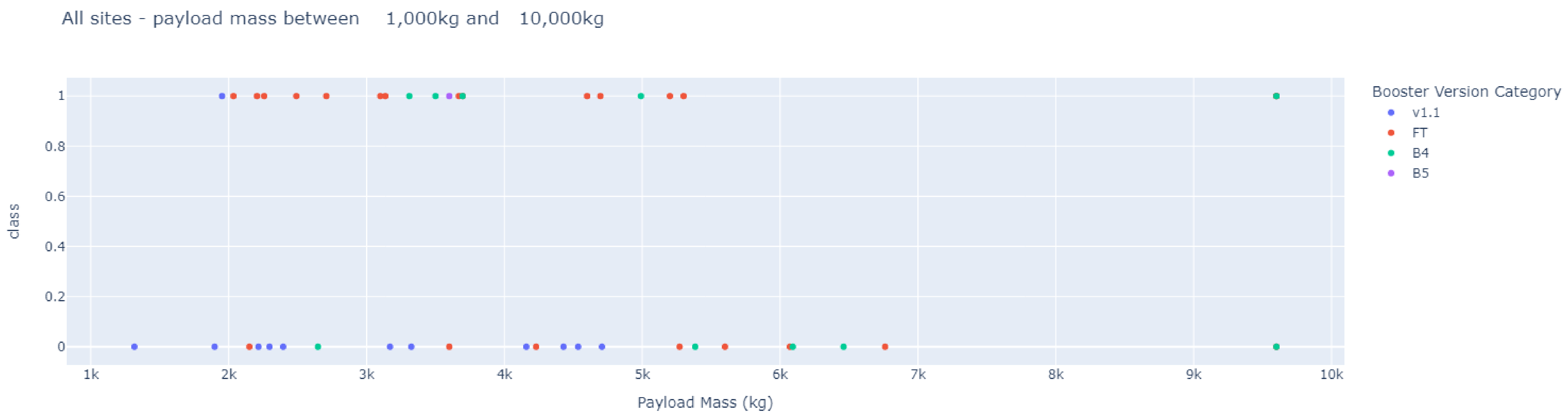
- The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches.

Launch Success Ratio for KSC LC-39A



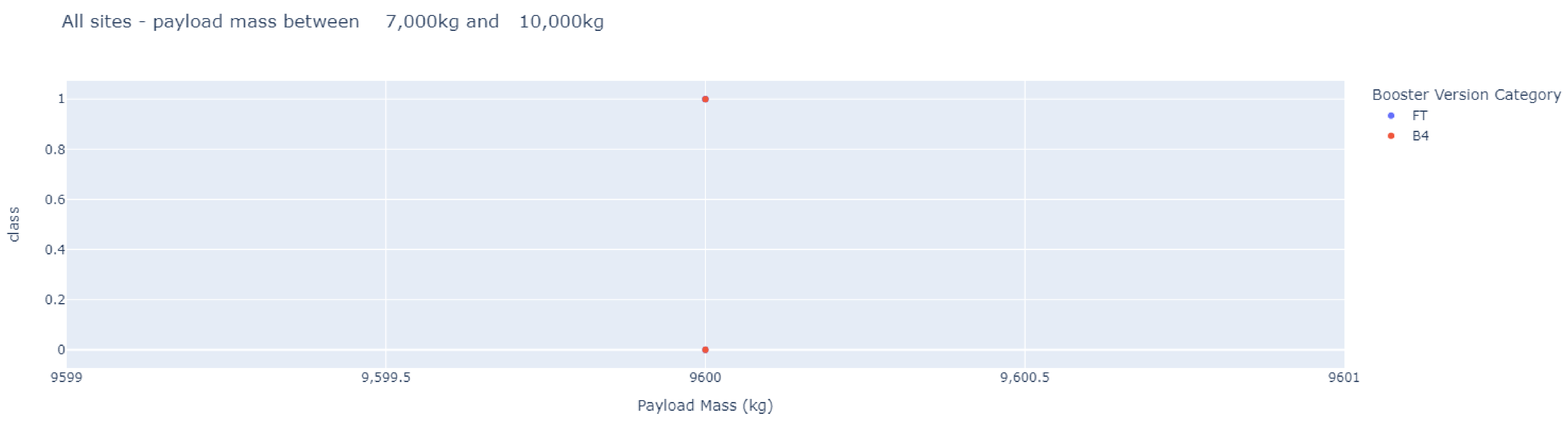
- KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings.

Payload vs. Launch Outcome



- Payloads under 6,000kg and FT boosters are the most successful combination.

Payload vs. Launch Outcome



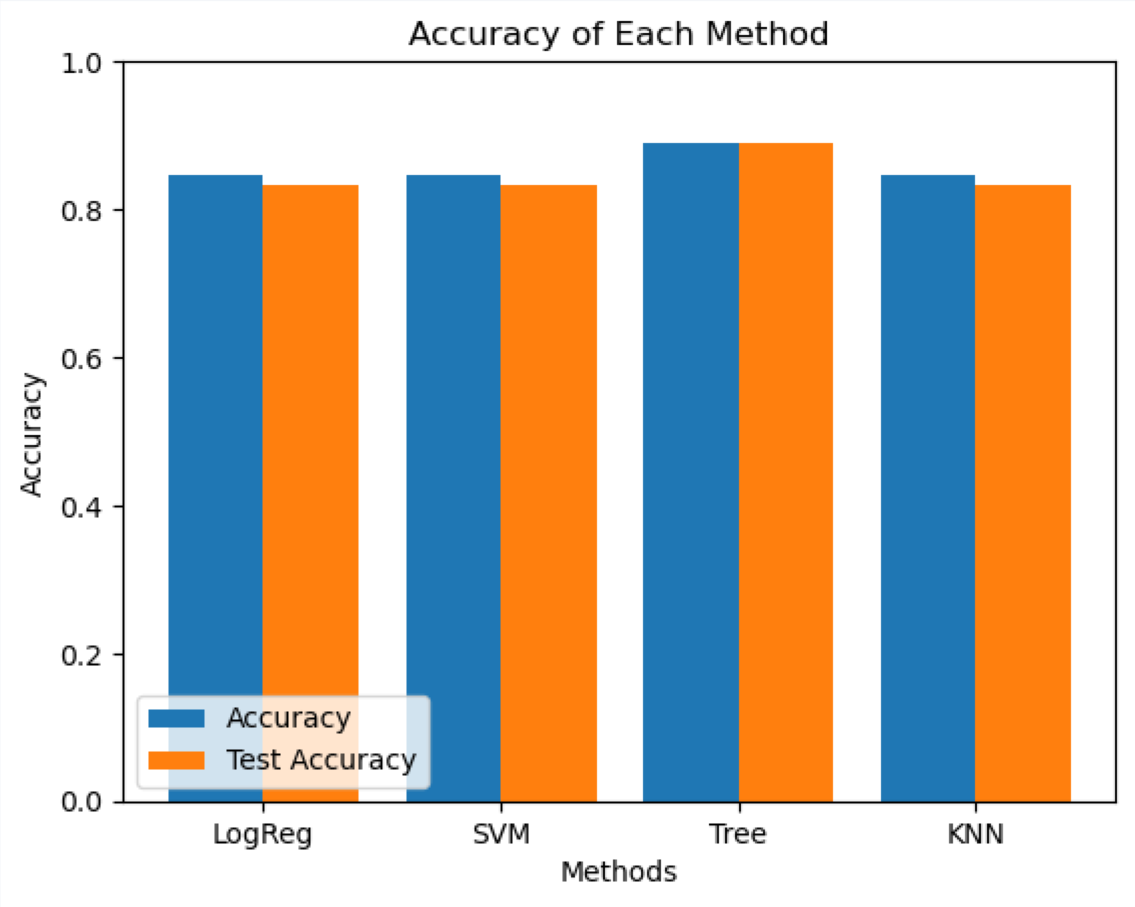
- There's not enough data to estimate risk of launches over 7,000kg

Section 5

Predictive Analysis (Classification)

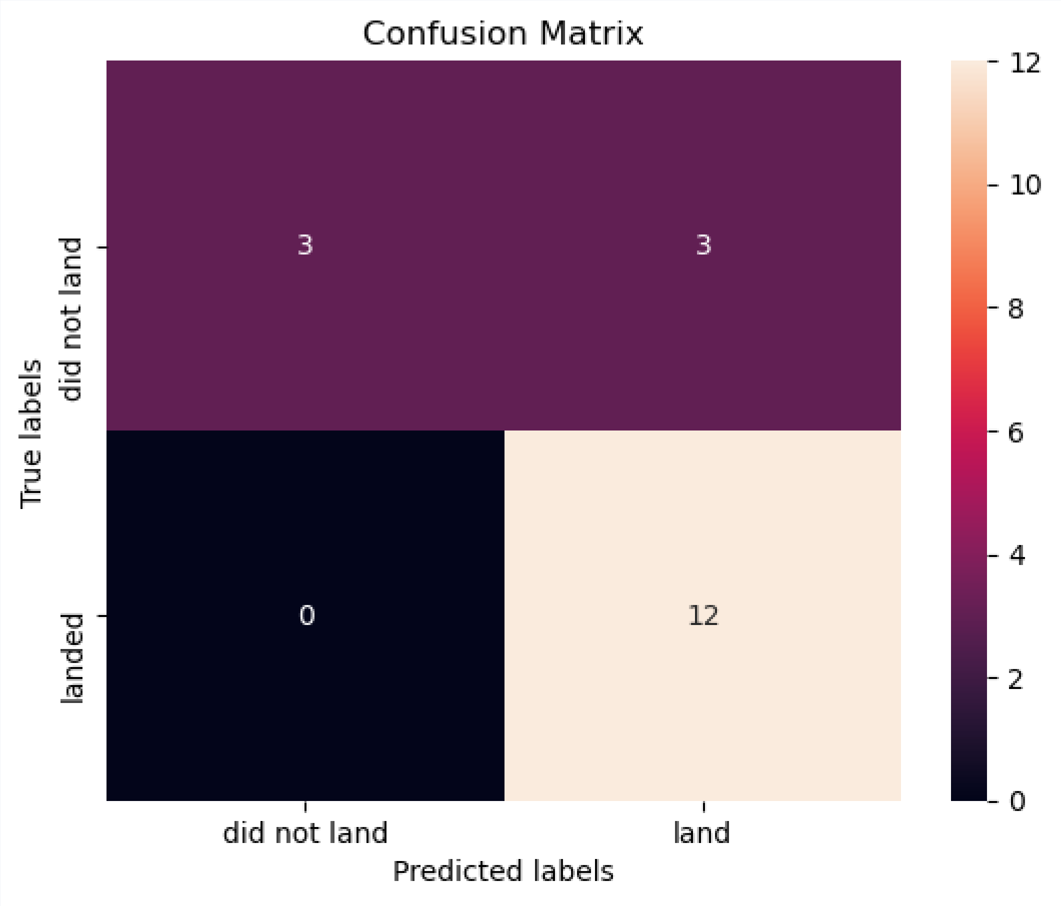
Classification Accuracy

- Four classification models were tested, and their accuracies are plotted beside;
- The model with the highest classification accuracy is Decision Tree Classifier, which has accuracies over than 85%.



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.



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.
- Although most of mission outcomes are successful, successful landing outcomes seem to improve over time, according the evolution of processes and rockets.

Appendix

Notebooks to recreate dataset, analysis, and models:

- https://github.com/ricss125/Applied-Data-Science-Capstone/blob/main/DataCollectionAPI.ipynb
- https://github.com/ricss125/Applied-Data-Science-Capstone/blob/main/DataCollectionWithWebScraping.ipynb
- https://github.com/ricss125/Applied-Data-Science-Capstone/blob/main/DataWrangling.ipynb
- https://github.com/ricss125/Applied-Data-Science-Capstone/blob/main/EDA_DataVisualization.ipynb
- https://github.com/ricss125/Applied-Data-Science-Capstone/blob/main/EDA_with_SQL.ipynb
- https://github.com/ricss125/Applied-Data-Science-Capstone/blob/main/Interactive_Dashboard_with_Ploty_Dash.py
- https://github.com/ricss125/Applied-Data-Science-Capstone/blob/main/Machine_Learning_Prediction.ipynb
- https://github.com/ricss125/Applied-Data-Science-Capstone/blob/main/Visual_Analytics_with_Folium.ipynb

Special thanks to all the instructors

Thank you!