IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Florian Tönjes
2023/3/7

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

With the aim to estimate the price of a rocket launch by SpaceX and to also predict if the first stage of a rocket can be reused, data from the SpaceX-API, as well as publicly available launch data from Wikipedia, has been first explored, prepared visually, and then been analyzed to finally create a regression model for prediction.

Brief Summary of Results

With every new flight SpaceX makes progress in successful landing outcomes.

Flights to the ES-L1, GEO, HEO, and SSO orbits prove to be the easiest mission objectives.

The launch sites are close to the coast, next to transportation infrastructure, and at a good distance from the nearest city.

Also a dashboard is available to monitor success-to-failure rates at or of different launch sites and a dynamic scatter plot to analyze the impact of different payload weight ranges on successful landings and for different launch sites and booster versions.

With the collected data we could fit a decision tree model which was able to predict a successful landing of the rocket stage with an accuracy of ca. 94%.

Introduction

As a data scientist working for SpaceY my task was to analyze the rocket launch and landing data of SpaceX, to find patterns and finally enable our company to make predictions based on this data.

We wanted to find answers to the following questions:

- What will be the price of a rocket launch by SpaceX based in the future?
- Can we predict if the first stage of a rocket can land successfully and so can be reused reducing the launch price?

Section 1

Methodology

Methodology

Executive Summary

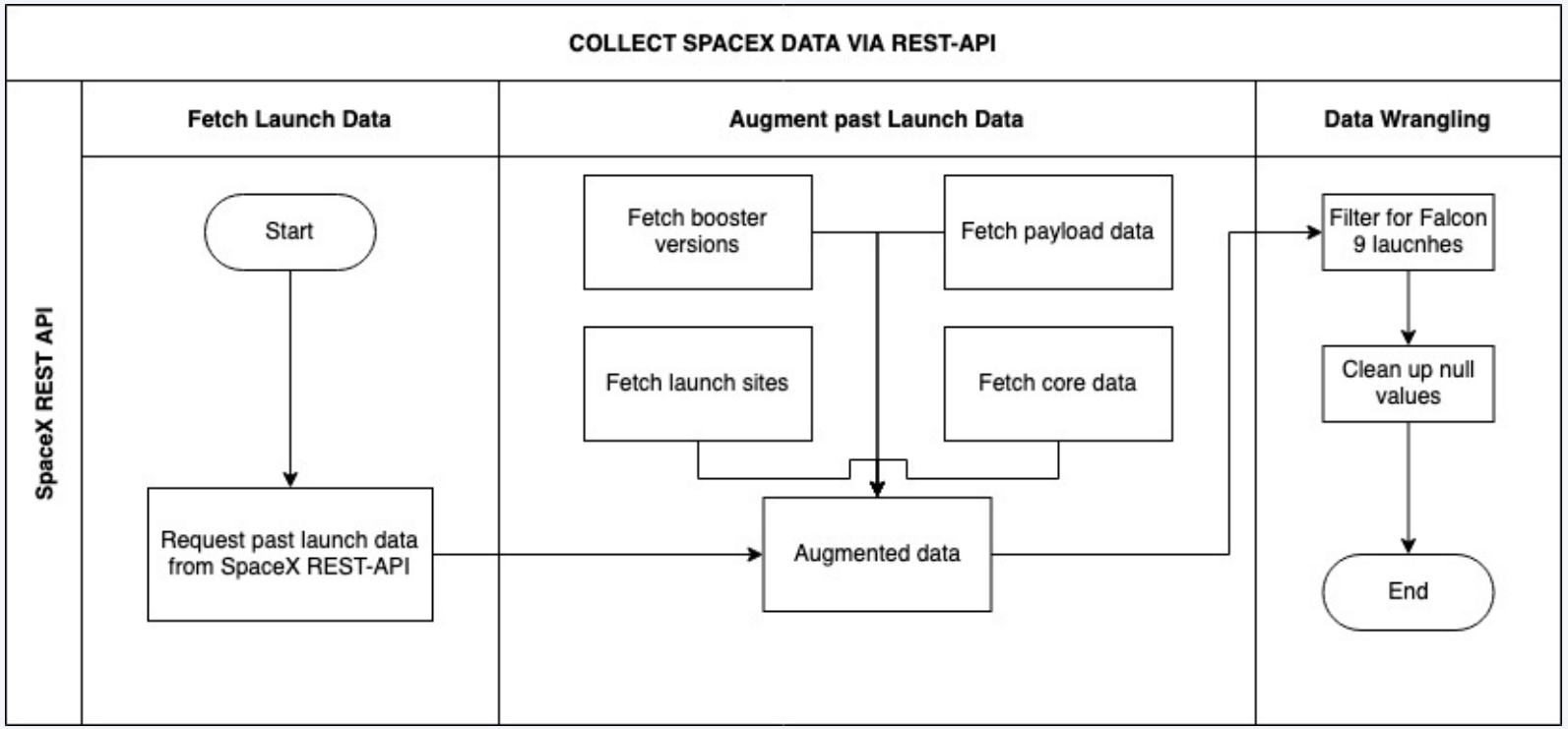
- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- The data used for the analysis was gathered with the following methods:
  - Querying the SpaceX REST API for launch data
  - Scraping data from publicly available Wikipedia websites

Data Collection – SpaceX API

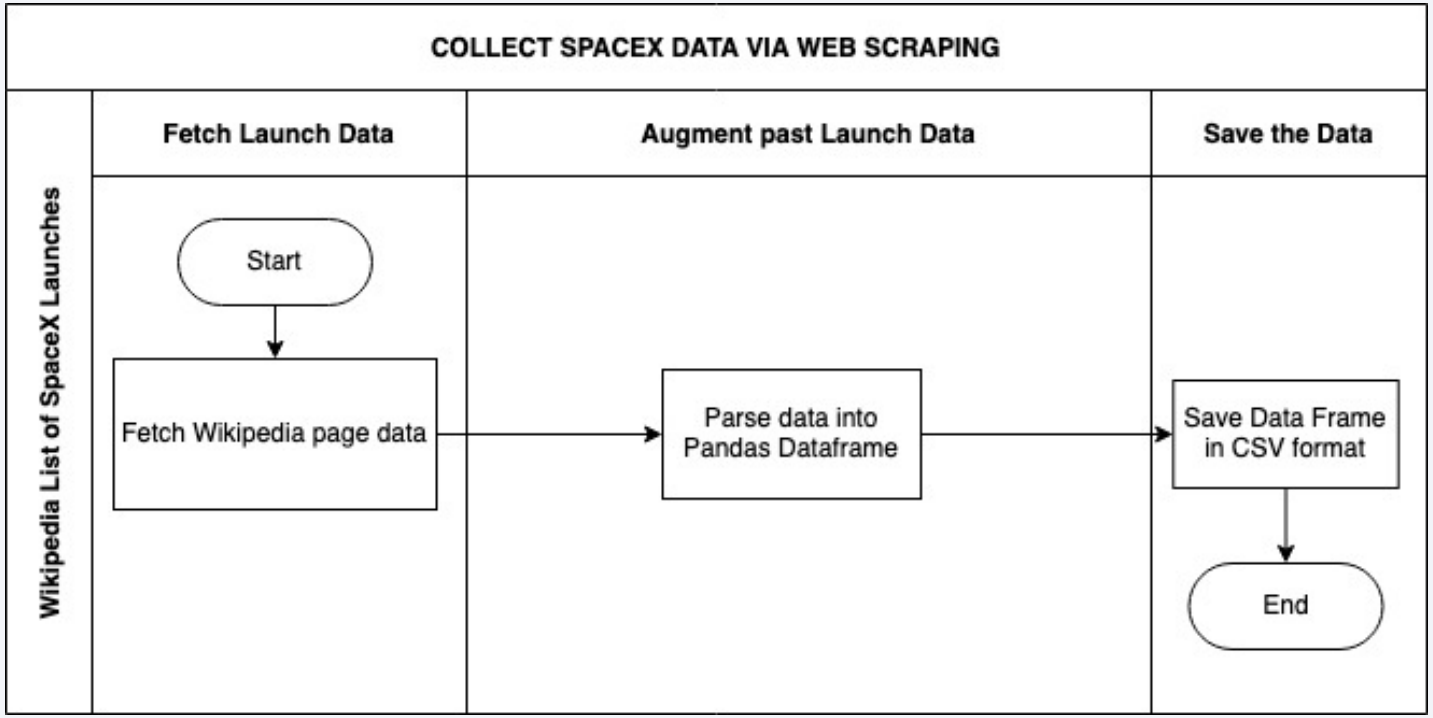
- First SpaceX launch data was gathered via a REST-API call to the SpaceX API
- Next the fetched data was augmented by data from additional calls
- Finally the data was filtered, cleaned, and saved in CSV-Format



Jupyter Notebook Link:
https://github.com/floriantoenjes/IBM_Data_Science_Professinonal_Capstone_Project/blob/main/Data%20Collection%20API.ipynb

Data Collection - Scraping

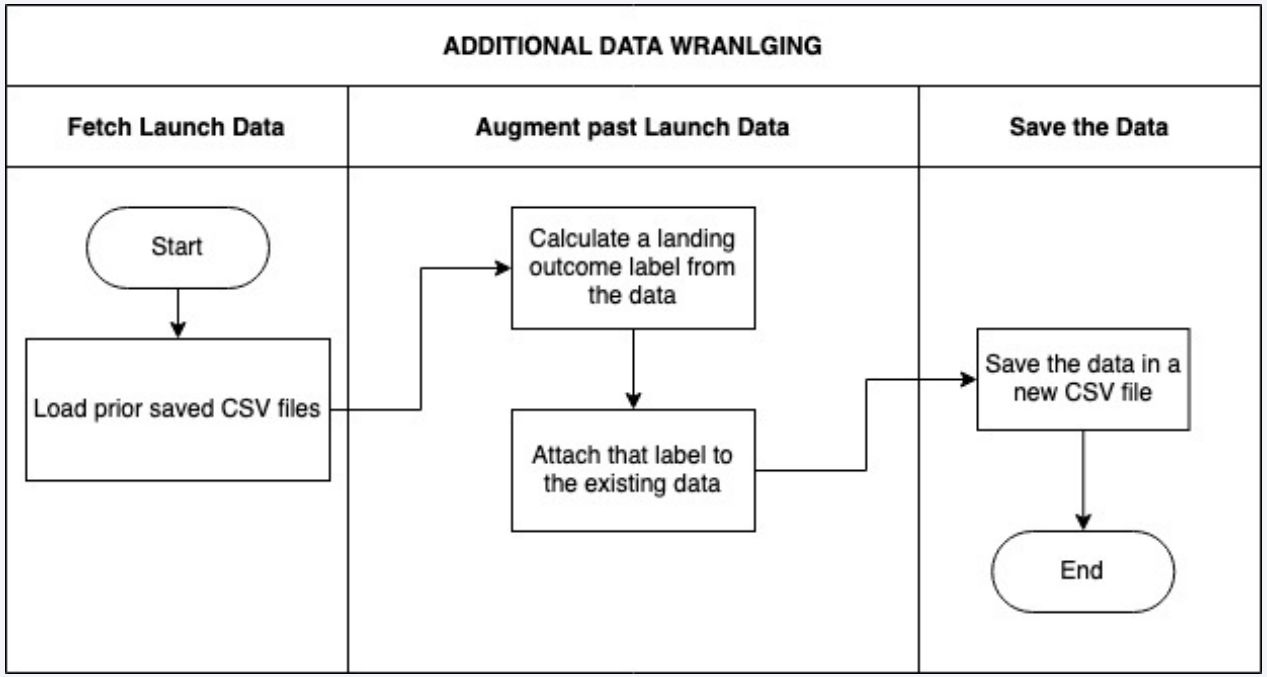
- Additional data was scraped from a Wikipedia page listing past SpaceX launches
- The data was saved in CSV-format, too.



Jupyter Notebook Link:
https://github.com/floriantoenjes/IBM_Data_Science_Professinonal_Capstone_Project/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb

Data Wrangling

- Additional data wrangling was done to calculate a landing outcome label for each row and attach it to the data that has been saved in CSV-file format before



Jupyter Notebook Link:
https://github.com/floriantoenjes/IBM_Data_Science_Professinonal_Capstone_Project/blob/main/Data%20Wrangling.ipynb

EDA with Data Visualization

- Then to explore the data visually following charts were created:
  - Scatter plot to show the relationship between flight number and payload mass
  - Scatter plot to show the relationship between flight number and launch site
  - Scatter plot to show the relationship between payload and launch site
  - Bar chart for the relationship between success rate and orbit
  - Scatter plot to show the relationship between flight number and orbit type
  - Scatter plot to show the relationship between payload and orbit type
  - Line chart to visualize the success yearly trend

Jupyter Notebook Link:
https://github.com/floriantoenjes/IBM_Data_Science_Professinonal_Capstone_Project/blob/main/EDA%20with%20Visualization.ipynb

EDA with SQL

- The following information was gained from the database:
  - The different launch sites were identified
  - A sum for the total payload mass launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1
  - The date when the first successful landing outcome on a ground pad was achieved
  - The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000 were listed
  - The total number of successful and failed mission outcomes was counted
  - The booster versions which carried the maximum payload were identified
  - Launch information for the year 2015 was identified
  - The count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order was ranked

Jupyter Notebook Link:
https://github.com/floriantoenjes/IBM_Data_Science_Professinonal_Capstone_Project/blob/main/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- On an interactive map the following objects were created to visually analyze geo-information and identify common factors on launch site terrain and near infrastructure:
  - Markers for the different launch sites
  - Distance to the nearest coastline
  - Distance to the nearest highway
  - Distance to the nearest railway
  - Distance to the nearest city

Jupyter Notebook Link:
https://github.com/floriantoenjes/IBM_Data_Science_Professinonal_Capstone_Project/blob/main/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- A dashboard was built with Plotly offering the following interactive elements:
  - A dropdown to confine the information for a specific launch sites
  - A pie chart which displays the successful and failed launches for all or only for the the selected launch site
  - A slider to adjust the payload weight range that is being visualized
  - A scatter plot visualizing the relation between the selected payload range for the selected launch site and the payload mass to get an understanding of which payload range is more or less successfully launched

Jupyter Notebook Link:
https://github.com/floriantoenjes/IBM_Data_Science_Professinonal_Capstone_Project/blob/main/PlotyDashboard.py

Predictive Analysis (Classification)

- The following predictive models were built and tested on the prepared data:
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree Classifier
  - K Nearest Neighbors Classifier
- The performance of each model was evaluated using a grid search to find its most effective hyper parameters
- Then the different predictive models were compared by their scores
- All models but the decision tree gave a precision of ca. 83% on the out of sample data set
- Though the decision tree model took the longest for its prediction, apparently the time and processing power was well spent as it reached a precision of over 94%

Jupyter Notebook Link:
https://github.com/floriantoenjes/IBM_Data_Science_Professinonal_Capstone_Project/blob/main/Machine%20Learning%20Prediction.ipynb

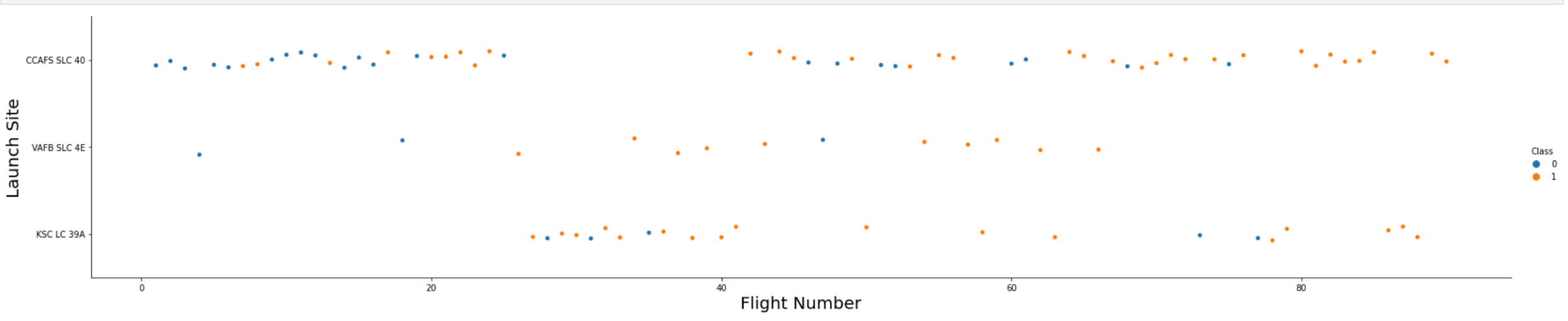
Results

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

Section 2

Insights drawn from EDA

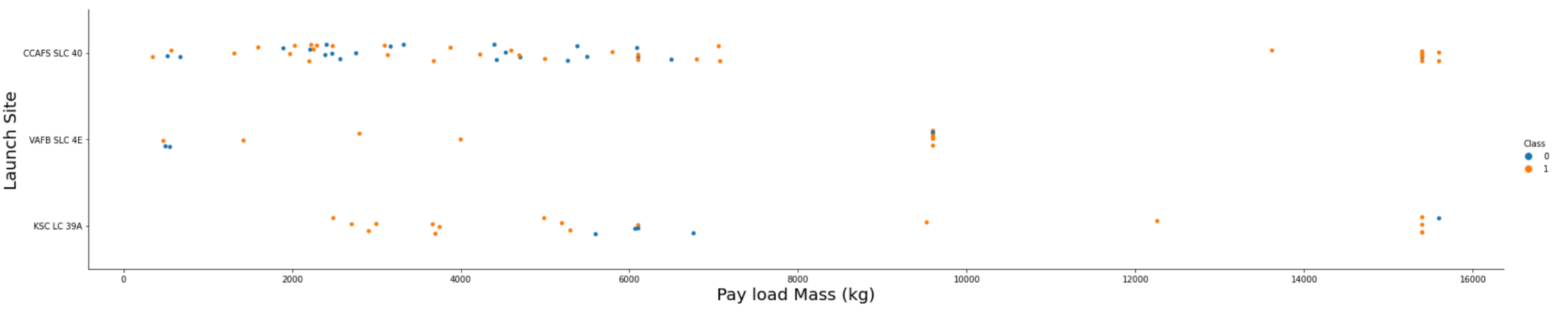
Flight Number vs. Launch Site



The scatter plot relating flight numbers and launch sites to the success rate

Class 0 = Failure
Class 1 = Success

Payload vs. Launch Site



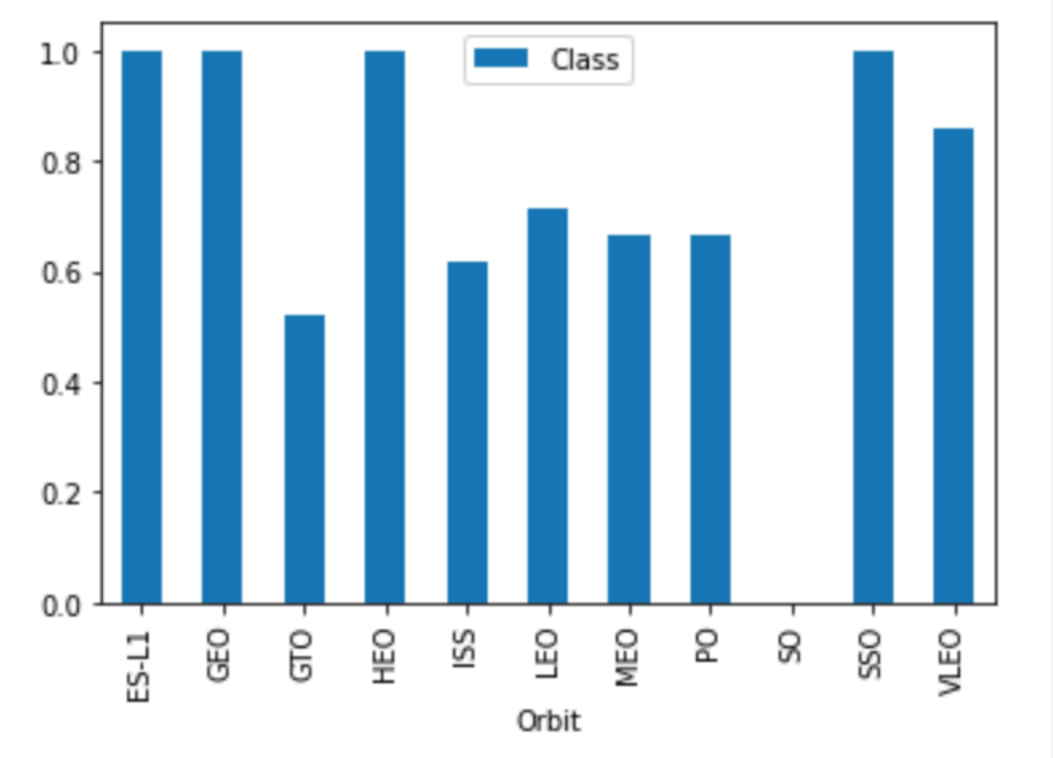
The scatter plot relating launch sites and payload mass to the success rate

Class 0 = Failure

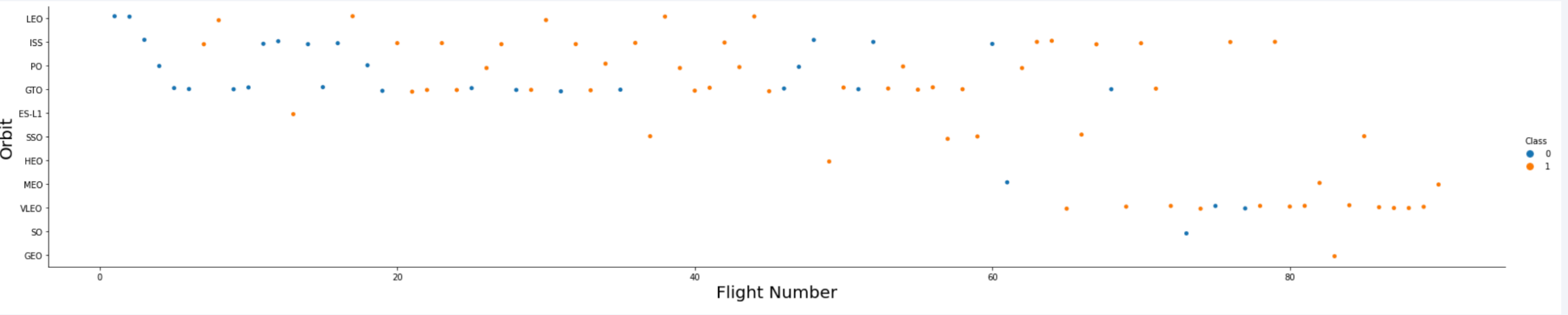
Class 1 = Success

Success Rate vs. Orbit Type

- Bar chart comparing the amount of successful launches for each launch site



Flight Number vs. Orbit Type

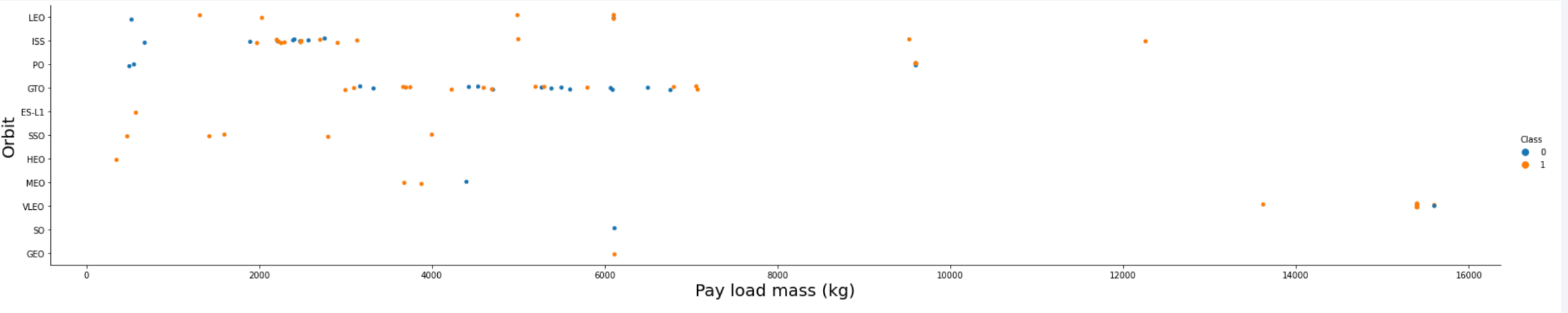


The scatter plot relating flight number and orbit type to the success rate

Class 0 = Failure

Class 1 = Success

Payload vs. Orbit Type



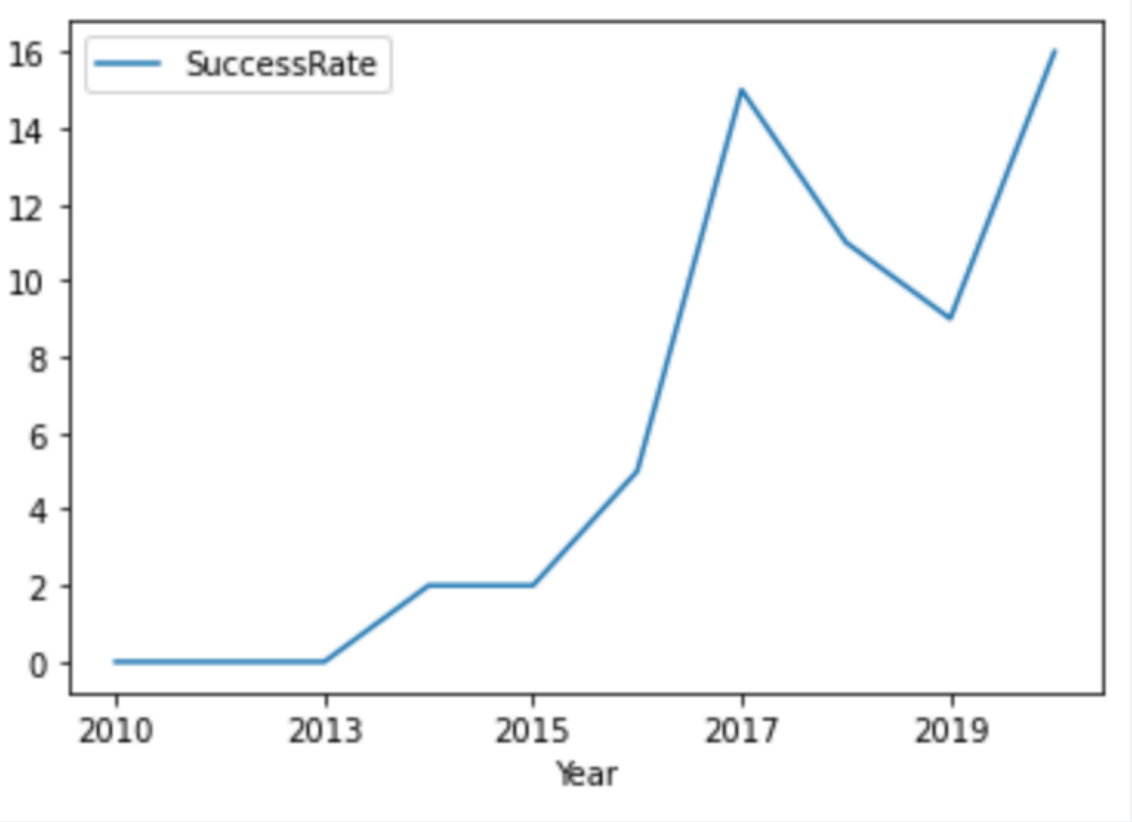
The scatter plot relating orbit type and payload mass in kg to the success rate

Class 0 = Failure

Class 1 = Success

Launch Success Yearly Trend

Line chart representing the successful launches over the last years



All Launch Site Names

The different launch sites that were identified

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

Launch Site Names Begin with 'CCA'

Five rows of the data frame with launch site names beginning with „CCA“

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success
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
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success

Total Payload Mass

The total payload mass in kilograms that was launched by NASA

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

The average payload mass of the Falcon F9 v1.1 boosters

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

The date of the first successful landing outcome on ground pad

MIN(Date)
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

List of names of boosters which have successfully landed on a drone ship and had payload mass greater than 4000 but less than 6000

Booster_Versio
F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1014
F9 v1.1 B1016
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1030
F9 FT B1021.2
F9 FT B1032.1
F9 B4 B1040.1
F9 FT B1031.2
F9 B4 B1043.1
F9 FT B1032.2
F9 B4 B1040.2
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1046.3
F9 B5B1054
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5B1060.1
F9 B5 B1058.2
F9 B5B1062.1

Total Number of Successful and Failure Mission Outcomes

Total number of mission outcomes

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

Boosters Carried Maximum Payload

List of the names of boosters which 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

Landing outcomes on drone ships for year 2015

substr(Date, 4, 2)	Landing _Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
02	Controlled (ocean)	F9 v1.1 B1013	CCAFS LC-40
03	No attempt	F9 v1.1 B1014	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
04	No attempt	F9 v1.1 B1016	CCAFS LC-40
06	Precluded (drone ship)	F9 v1.1 B1018	CCAFS LC-40
12	Success (ground pad)	F9 FT B1019	CCAFS LC-40

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

Data of successful landing outcomes
In a date range between
2010/06/04 and 2017/03/20

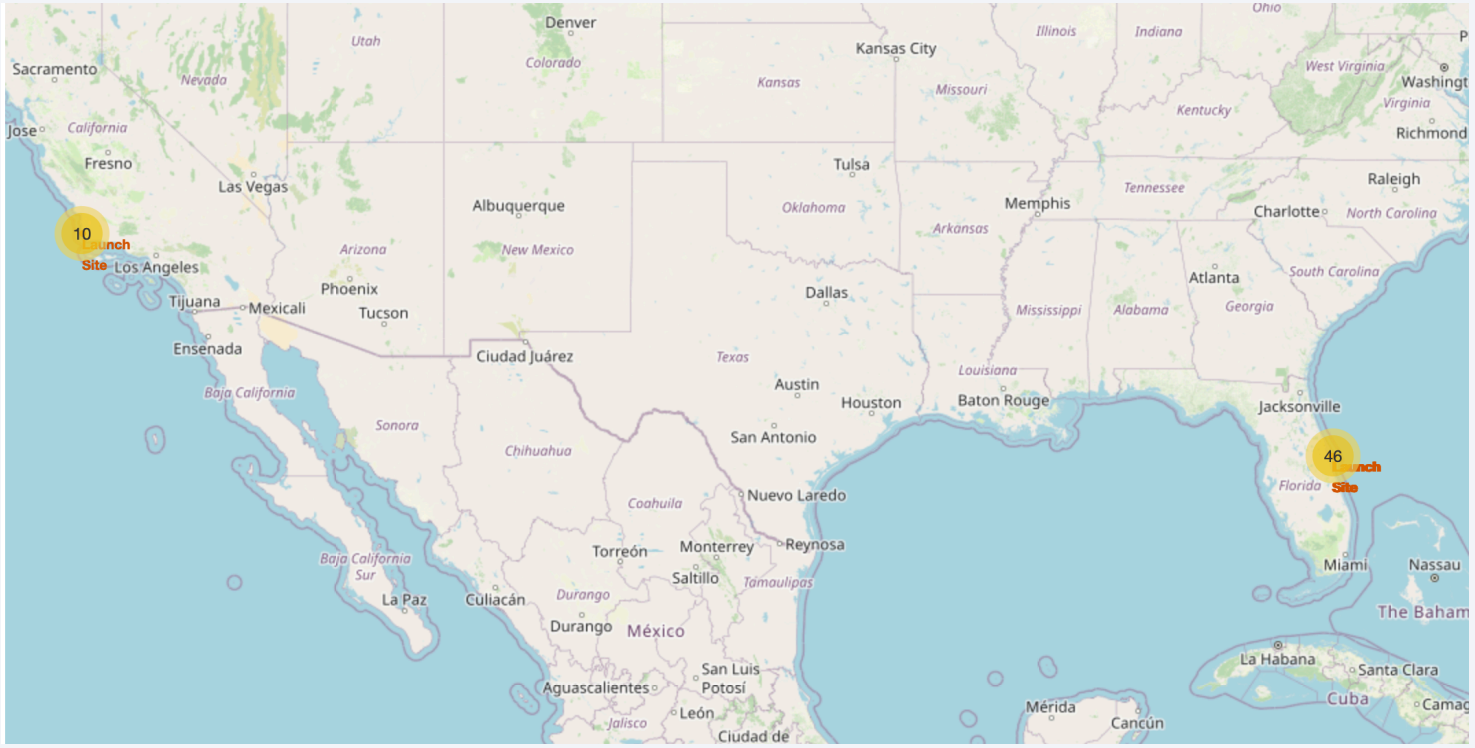
Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outc
19-02-2017	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success
14-01-2017	17:54:00	F9 FT B1029.1	VAFB SLC-4E	Iridium NEXT 1	9600	Polar LEO	Iridium Communications	Success
14-08-2016	05:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success
18-07-2016	04:45:00	F9 FT B1025.1	CCAFS LC-40	SpaceX CRS-9	2257	LEO (ISS)	NASA (CRS)	Success
27-05-2016	21:39:00	F9 FT B1023.1	CCAFS LC-40	Thaicom 8	3100	GTO	Thaicom	Success
06-05-2016	05:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success
08-04-2016	20:43:00	F9 FT B1021.1	CCAFS LC-40	SpaceX CRS-8	3136	LEO (ISS)	NASA (CRS)	Success
22-12-2015	01:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2	2034	LEO	Orbcomm	Success

Section 3

Launch Sites Proximities Analysis

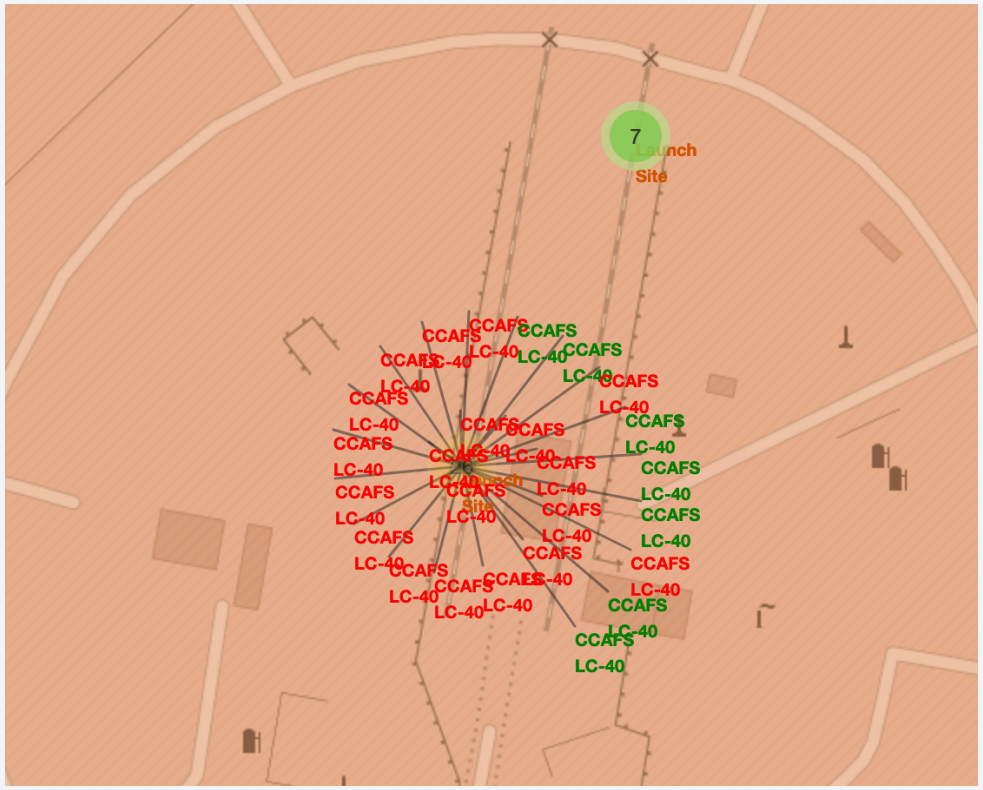
US SpaceX Launch Site Locations

A screenshot of the interactive Folium map showing the different launch sites marked with a yellow radius



CCAF Launch Site Successful vs Failed Attempt Cluster

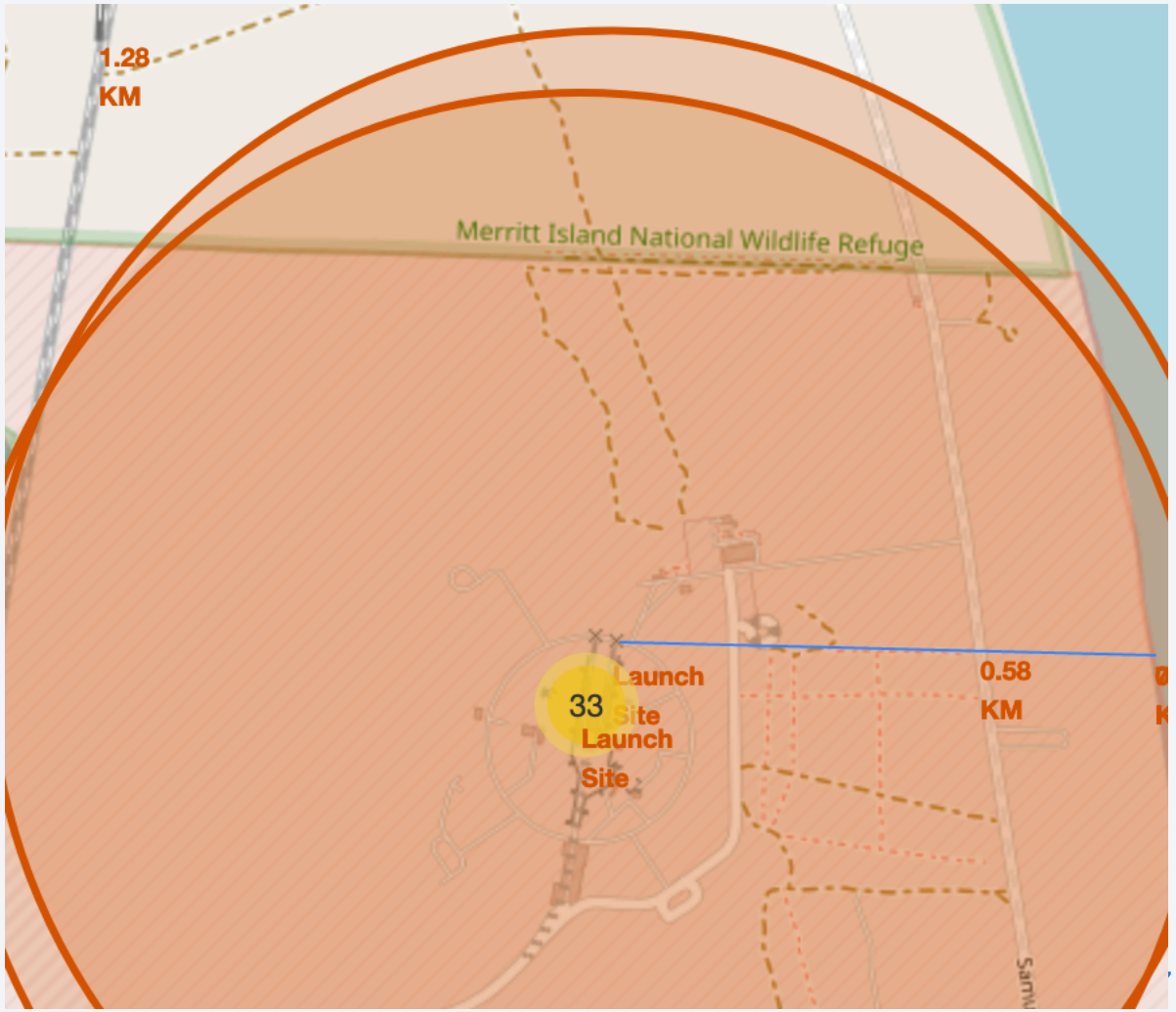
An interactive color coded cluster marker on the map showing the failed launches in red and the successful launches in green

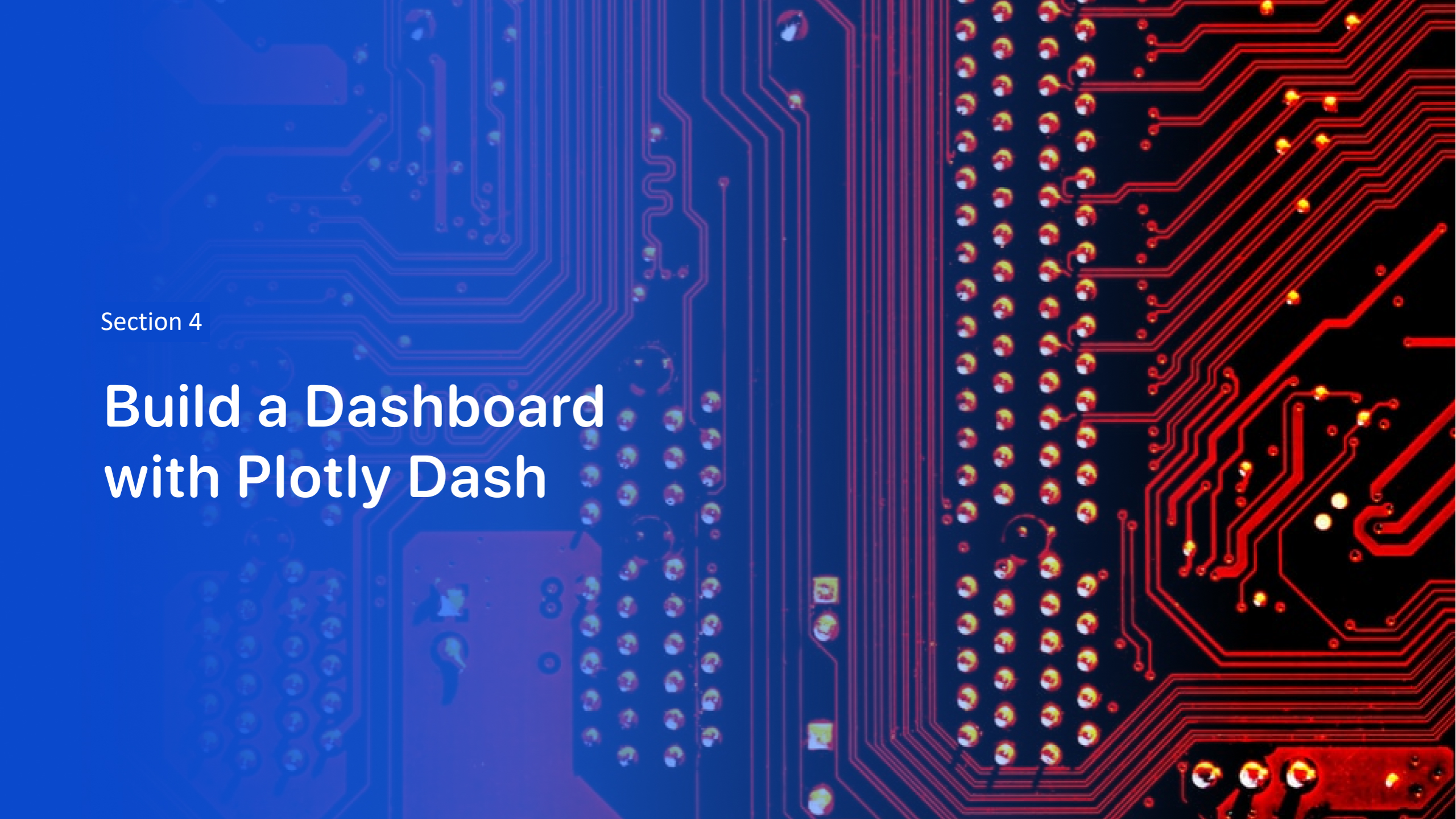


CCAF Launch Site Distances to Nearest Infrastructure and Coastline

The launch site and its distance to the nearest railway, highway, and the cost.

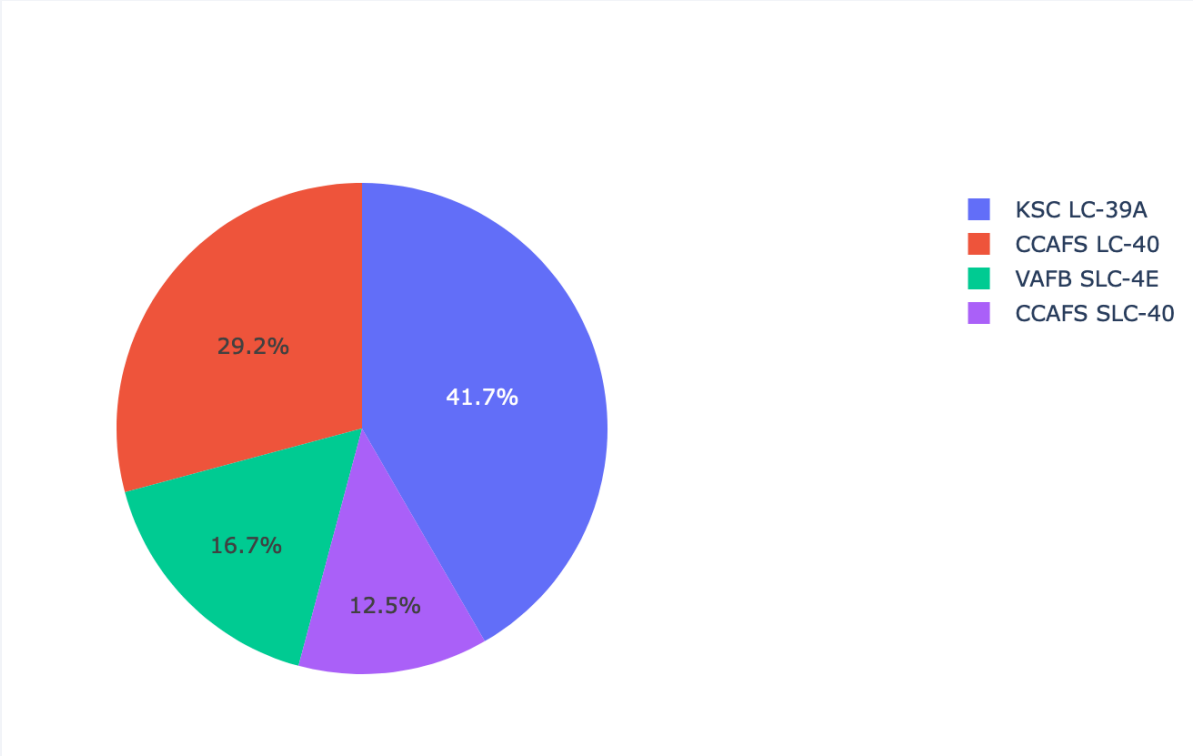
Area	Distance
Railway	1.28km
Highway	0.58km
Coast	0.96km




Section 4
Build a Dashboard with Plotly Dash

Percentages of Successful Launches per Launch Site

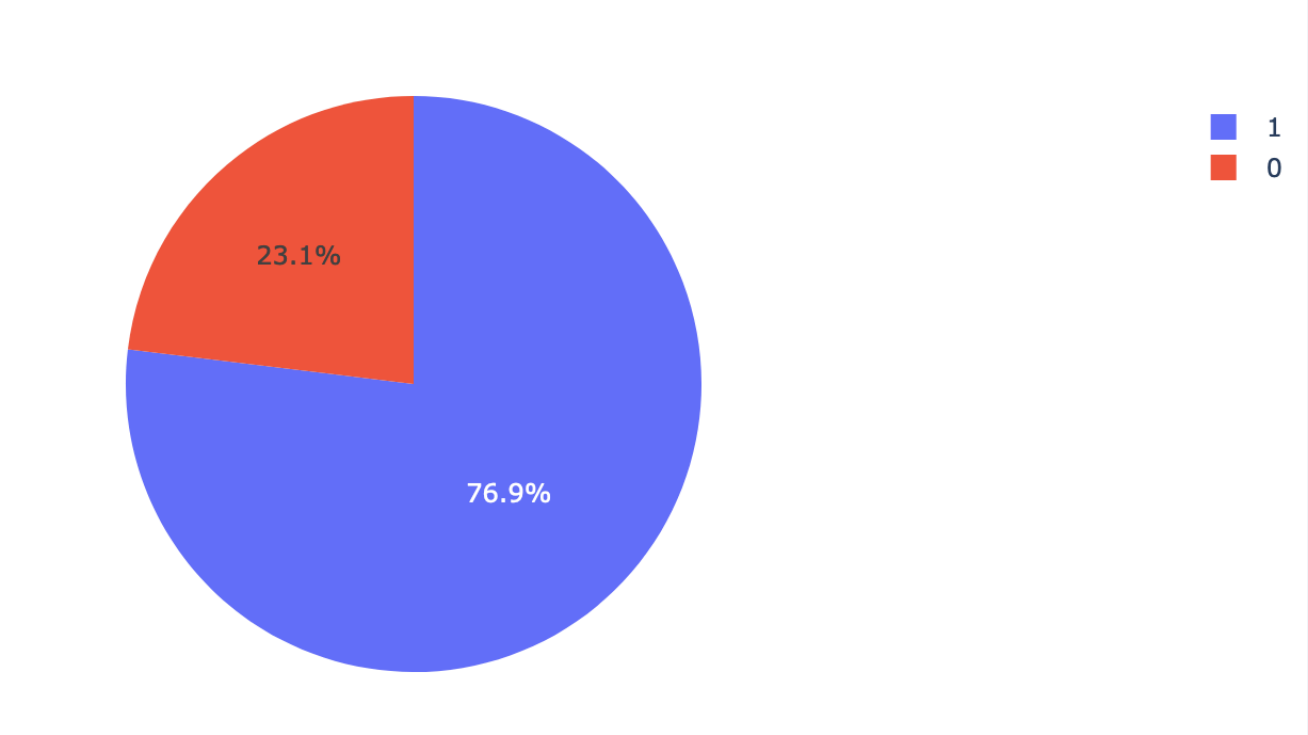
Interactive dashboard pie chart showing the percentages for successful launches



Successful Launches vs Failures for KSC LC-39A

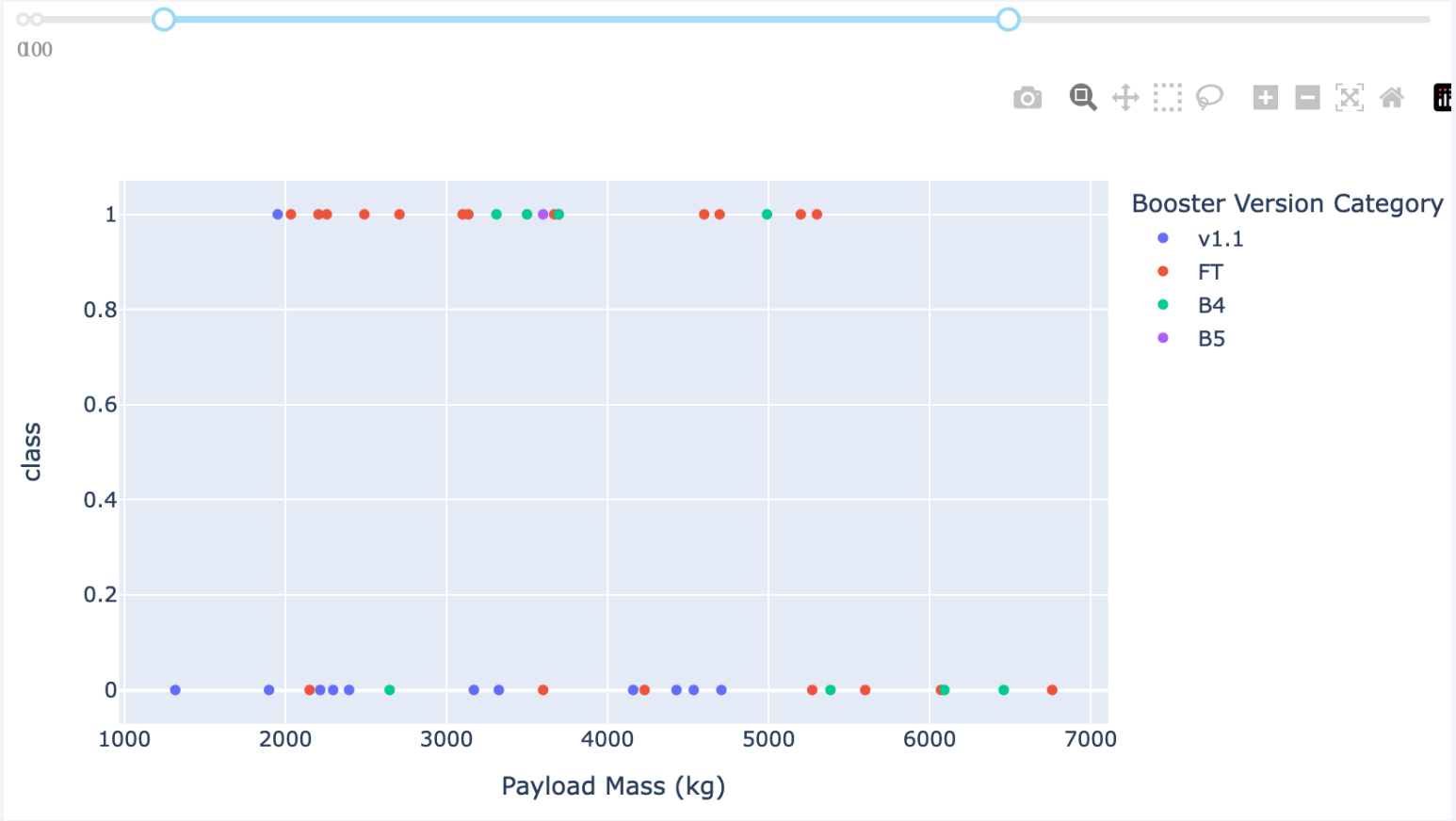
KSC LC-39A has the highest success rate of all launch sites with 76.9%

Class 1 = Success
Class 0 = Failure



Interactive Scatter Plot Relating Success/Failure to Payload Mass

This is a screenshot of the dynamic scatter plot on the dashboard relating success and failure classes to a payload mass range selected with a slider and separating different booster versions with different colors



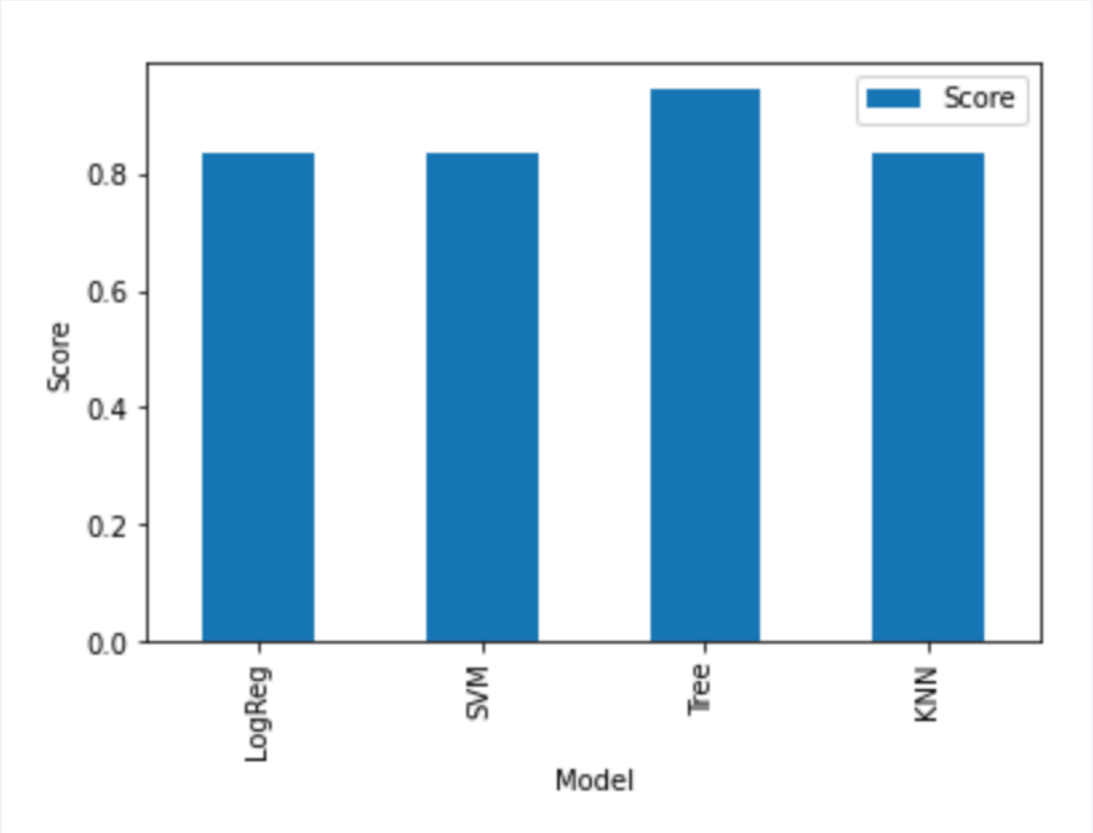
Section 5

Predictive Analysis (Classification)

Classification Accuracy

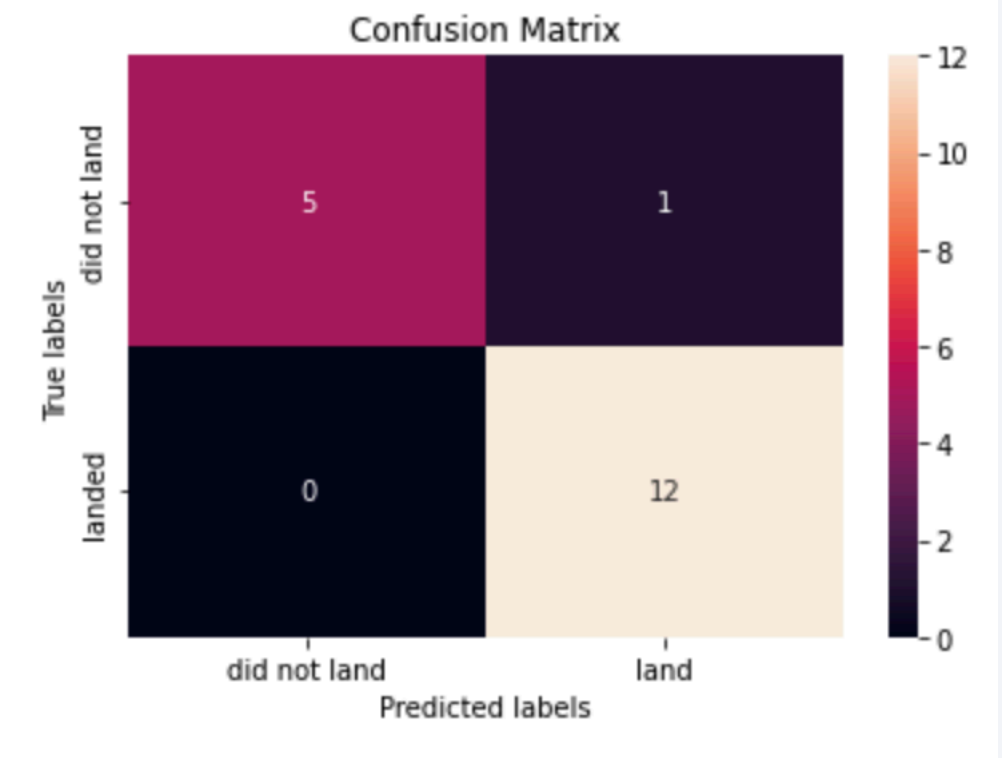
As one can see visualized in the bar chart the Logistic Regression, Support Vector Machine, and K Nearest Neighbors models all reach a similar accuracy of ca. **83%**

Only the **Decision Tree** model is able to reach a score of **94.4%** accuracy



Confusion Matrix

As can be seen in the confusion matrix
the tree model predicted only one false positive
which was a rocket that the model predicted
as landed when it actually did not land successfully
in reality



Conclusions

- With every new flight SpaceX makes progress in successful landing outcomes. Which is to be expected as they are learning from every failure as well as collecting information from every successful launch.
- Flights to the ES-L1, GEO, HEO, and SSO orbits prove to be the easiest mission objectives.
- The launch sites are close to the coast, next to transportation infrastructure, and at a good distance from the nearest city. It makes sense to launch a rocket in the direction of the coast and far away from cities were failures could endanger citizens or buildings. It is also practical to have transportation infrastructure near the launch sites as every launch demands transportation of multiple parts and chemicals which are assembled at the the sites. The information about geo data and the launch sites is now available in an interactive map.
- Landings of rockets with high payload masses are more successful than with lower ones.
- Using our fitted Decision Tree model we are now able to predict the outcome of rocket landings with an accuracy of about 94%, which gives us the ability to give a rough estimation about the price of the rocket launch as if the rocket can be reused the price drastically decreases compared to a launch without a recovered rocket.
- It would makes sense to gather new data and reevaluate the model on a constant, maybe yearly, basis.

Thank you!