Winning Space Race with Data Science

Toni Torrubia

January31, 2023

Outline

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

Executive Summary

- During this project, I used the following methodologies:
  - Data Collection with Webscraping and SpaceX API,
  - Data Wrangling,
  - Exploratory Data Analysis (EDA): with SQL and Data Visualization,
  - Interactive Visual Analytics and Dashboard with Folium and Plotly Dash,
  - Predictive Analysis (Classification) with Machine Learning techniques.
- Summary of all results
  - Raw data was collected and processed to be valuable data.
  - EDA helped to determine which features should be chosen to predict if a launch is successful.
  - Machine Learning techniques showed the best model to predict the success of launches, with all collected data.

Introduction

- The main object of this project feasibility of a new company, called Space Y – of Alan Mask, to compete with Space X.
- Main questions of the project:
  - The best way to estimate the total cost for launches is to predict successful landings of the first stage of rockets.
  - Which place is the best place to take the launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology: data in this project was collected from 2 sources:
  - SpaceX API: api.spacexdata.com/v4/launches/past
  - Webscraping: en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- Perform data wrangling
  - Collected data was used to find some statistics (total launchings on each site, etc).
  - After that, a landing outcome label was added to the data, with value 0 or 1.

Methodology

Executive Summary

- Perform exploratory data analysis (EDA) using visualization and SQL
  - SQL and Data Visualization Libraries in Python (Pandas, Matplotlib, Seaborn) was used to visualize and find some statistics (for example, average payload mass)
- Perform interactive visual analytics using Folium and Plotly Dash
  - Folium was used to plot the launching places on the Earth map.
  - Plotly Dash was used to create dashboards with launching results.

Methodology

Executive Summary

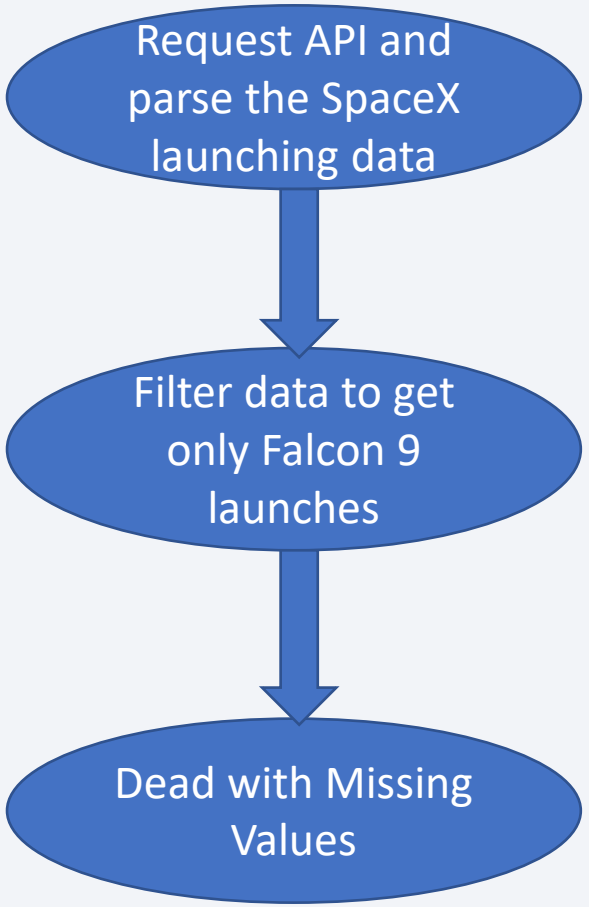
- Perform predictive analysis using classification models
  - Data that was collected until this steps was normalized.
  - Data was split into training and test datasets.
  - Four classification methods (K Nearest Neighbors, Support Vector Machine, Decision Tree, and Logistic Regression) were used to evaluated, find the accuracy to determine the best model.

Data Collection

- Data was collected from:
  - SpaceX API: api.spacexdata.com/v4/launches/past
  - Wikipedia: en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches with webscraping techniques.

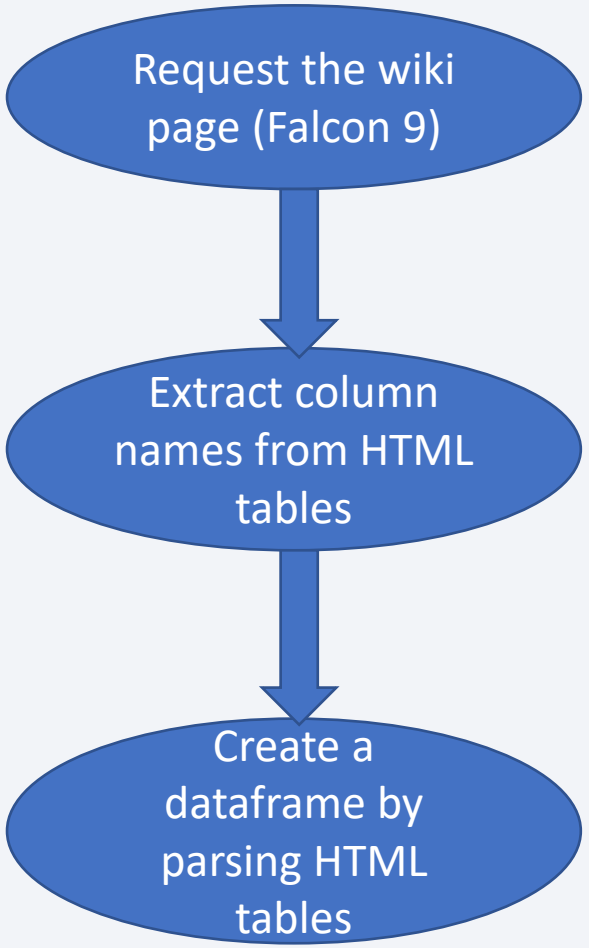
Data Collection – SpaceX API

- SpaceX provides an API from where data can be gotten and used for our purposes.
- This API was used according to the flowchart beside, and then we have data.
- Source code here.



Data Collection - Scraping

- Data of SpaceX launches can also be collected from Wikipedia.
- Data was downloaded from Wikipedia based on the flowchart beside.
- Source code here.



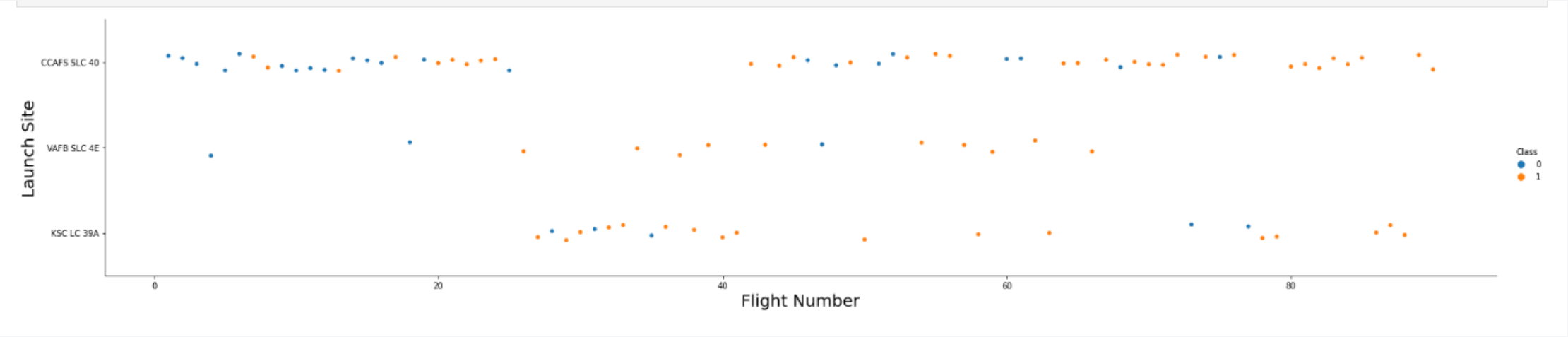
Data Wrangling

- First, some simple EDA was used to determine the characteristics of data.
- After that, the data was summarized by calculating the number of launches on each site, the number and occurrence of each orbit, and the number and occurrence of mission outcome per orbit type.
- Then a new column, which contains the landing outcome labels of launches, was added to the dataframe.

- Source code here.

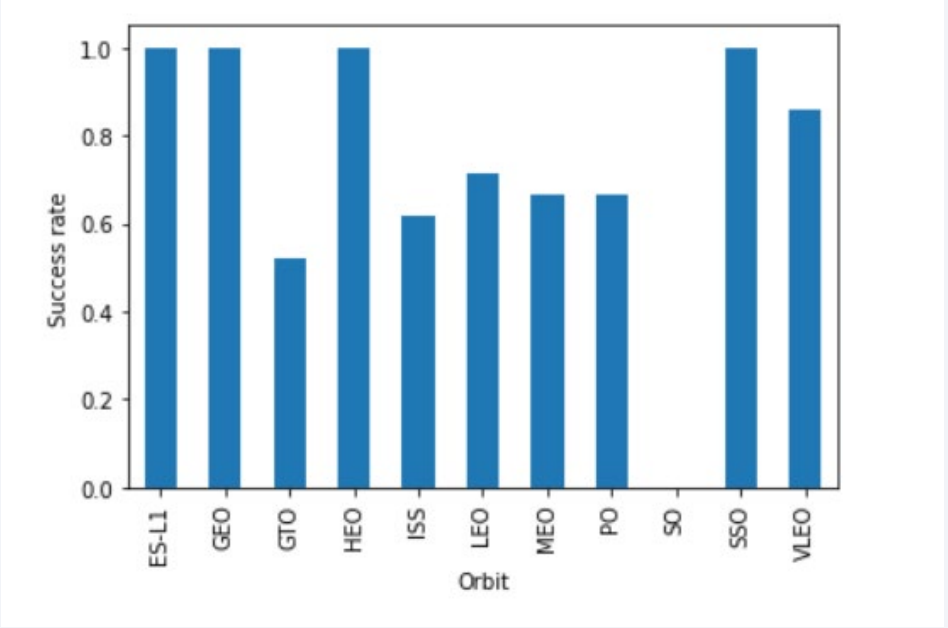
EDA with Data Visualization

- Scatter plots were used to visualize the relationship between Flight Number and Launch Site, Payload and Launch Site, Flight Number and Orbit Type. I chose scatter plot because it was a greate choice for relationship visualization with numerical data.



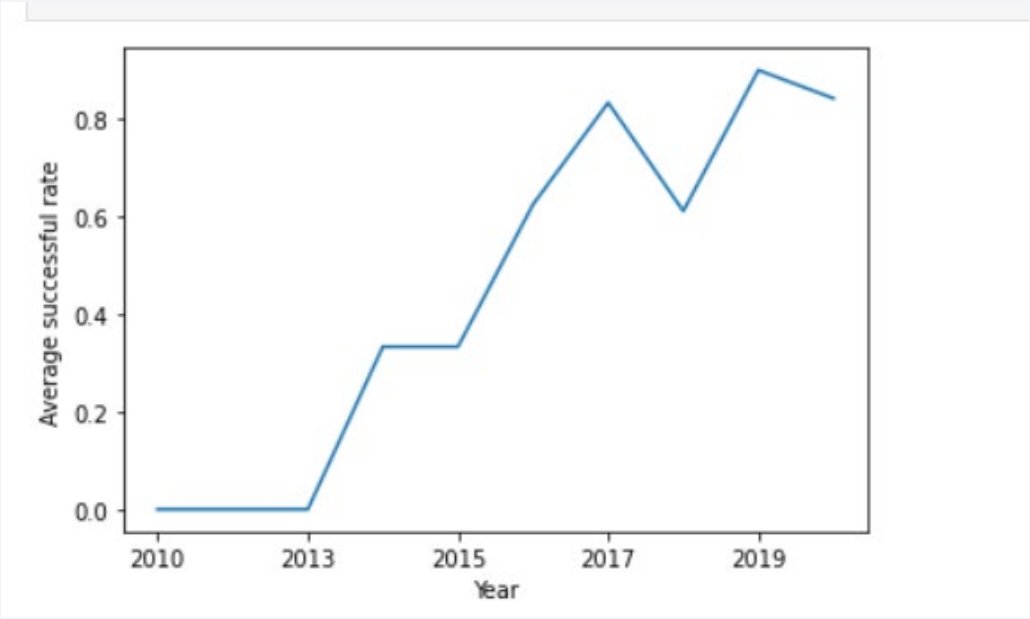
EDA with Data Visualization

- Bar plot best describes categorical data, therefore I used that kind to visualize the relationship between success rate of orbit type.



EDA with Data Visualization

- Moreover, a line plot was drawn to visualize the trend of average success rate over years.



EDA with Data Visualization

- Source code here.

EDA with SQL

- The following SQL queries were executed:
  - Names of launch sites in space missions.
  - Top 5 missions with lauch sites' name begin with 'CCA'.
  - Total payload mass carried by boosters launched by NASA.
  - Average payload mass carried by booster version F9 v1.1.
  - The date when the first successful landing outcome in groud pad was achieved.

EDA with SQL

- The following SQL queries were executed:
  - Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
  - Total number of successful and failure mission outcomes.
  - Names of the booster versions which have carried the maximum payload mass.
  - Failed landing outcome in drone ship mission in year 2015.
  - Rank of count of successful landing outcome missions between the date June 4, 2010 and March 20, 2017.
- Source code here.

Build an Interactive Map with Folium

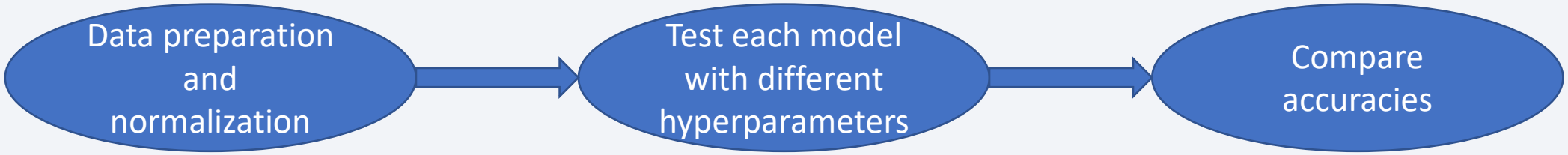
- Folium Maps were used with markers, circles, line and marker clusters.
  - Markers used to mark the points on maps, such as launching sites.
  - Circles used to highlight the regions nearby a specific coordinate, such as NASA Johnson Space Center at Houston, Texas.
  - Lines used to figure out the distance between 2 coordinates, for example from a launch site to the nearest coastline.
  - Marker clusters used to indicate groups of events at a specific coordinate, such as space missions at a launch site.
- Source code here.

Build a Dashboard with Plotly Dash

- For data visualization purposes, the following graphs and plots were added:
  - Percentages of successful and failed missions per launch site.
  - Total successful launches by site.
  - Successful rates in payload ranges.
- These graphs and plots allowed us to have a quick analysis on the relationship between payload mass and successful rate, and to decide which is the best place for launching.
- Source code here.

Predictive Analysis (Classification)

- Until this step, the data was normalized.
- Four classification models, Logistic Regression, Support Vector Machine, K Nearest Neighbors, and Decision Tree, were tested with different hyperparameter sets to find the best one and compared to figure out the best model for predicting the success of a landing.



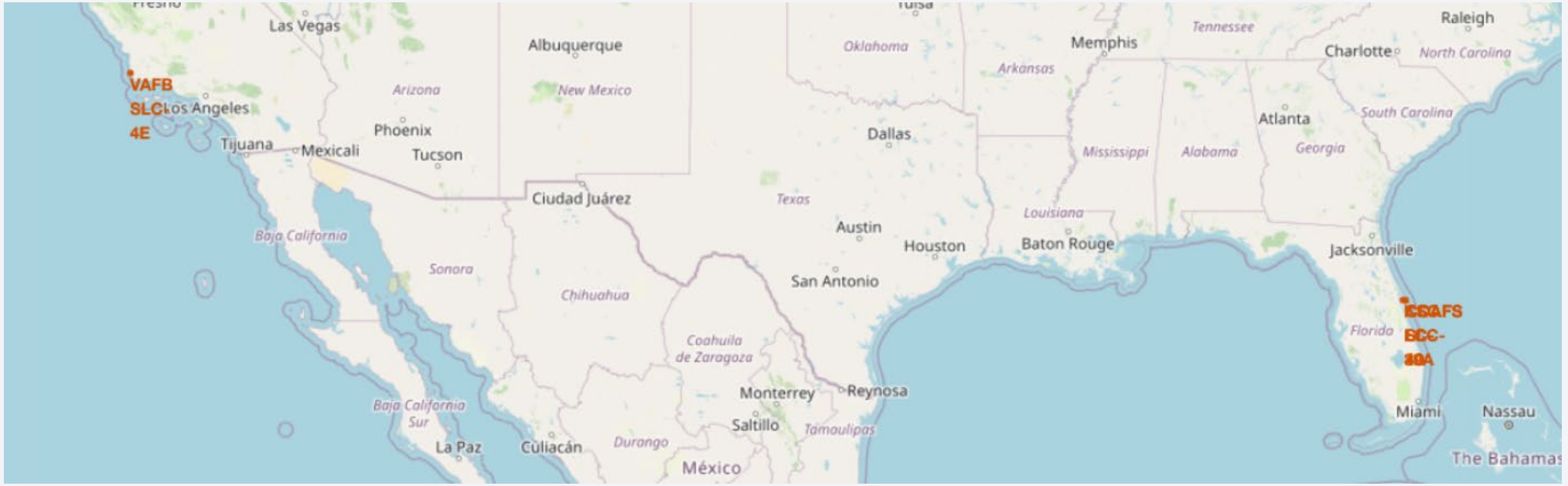
- Source code here.

Results

- Exploratory data analysis results
  - SpaceX uses 4 different launch sites in total.
  - The first launch was done by SpaceX itself.
  - The total payload mass carried by boosters launched by NASA (CRS) is 111,268 kg.
  - The average payload mass carried by booster version F9 v1.1 is 2928 kg.
  - The first landing outcome in ground pad happened in 2015.
  - The KSC LC-39A and VAFB SLC 4E launch site have the higher rate of success than CCAFS LC-40.
  - Almost 100% of missions were successful.
  - The number of landing outcomes increased when time passed.

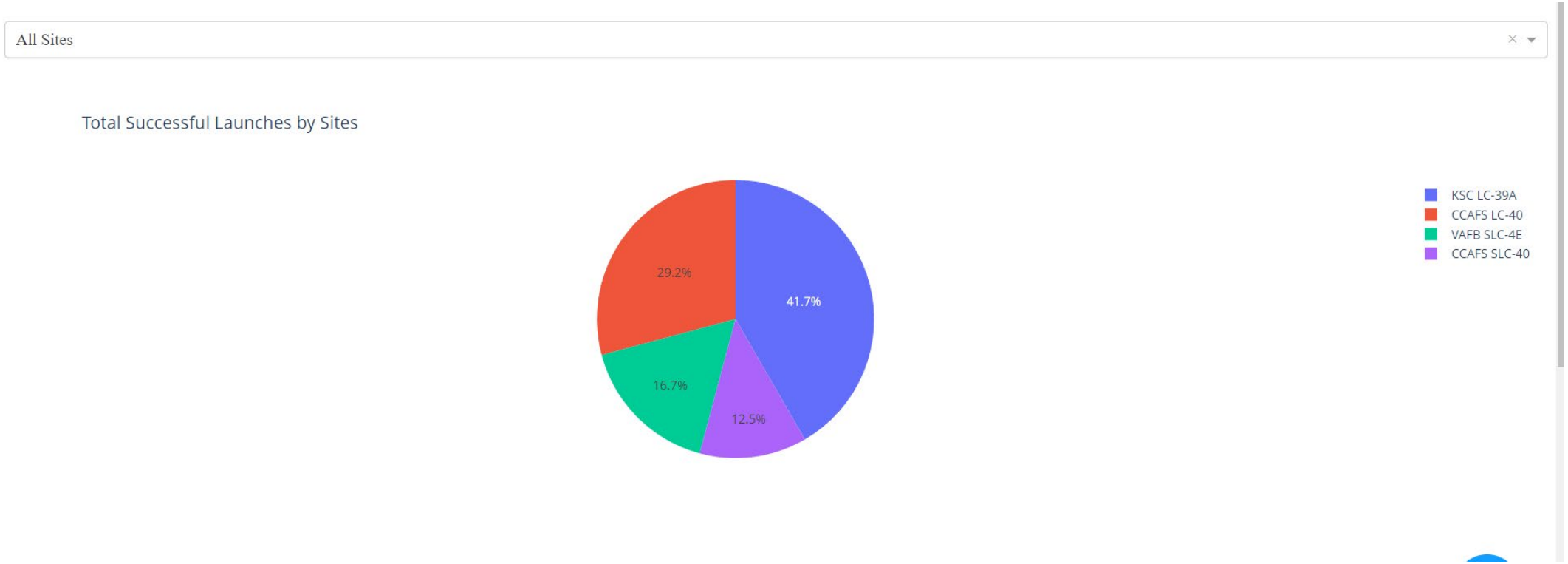
Results

- Interactive analytics demo in screenshots
  - The launch sites are in safe places, have good logistic infrastructure, for example, near the seas.
  - All launch sites are located nearby the coastline, both in the West and the East.



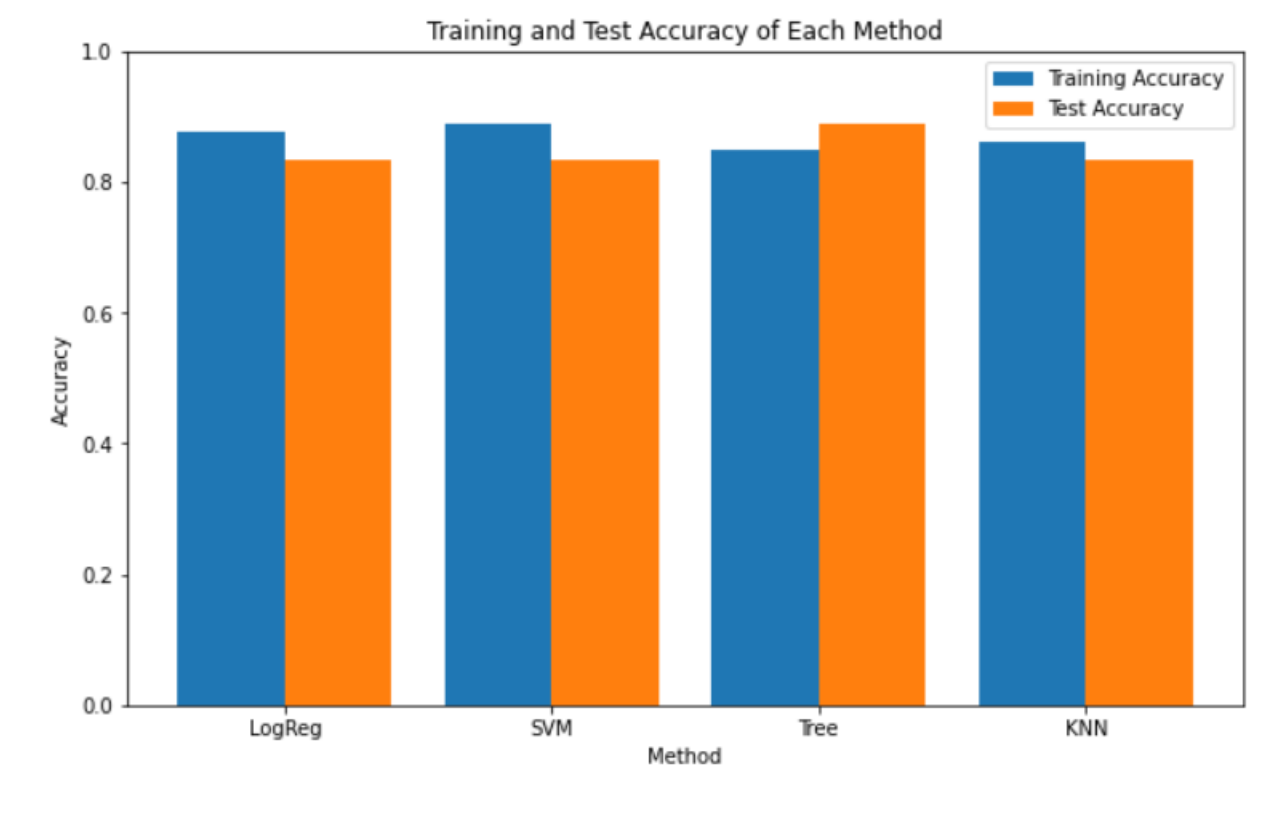
Results

- Interactive analytics demo in screenshots
  - KSC LC-39A has the largest number of successful launches.



Results

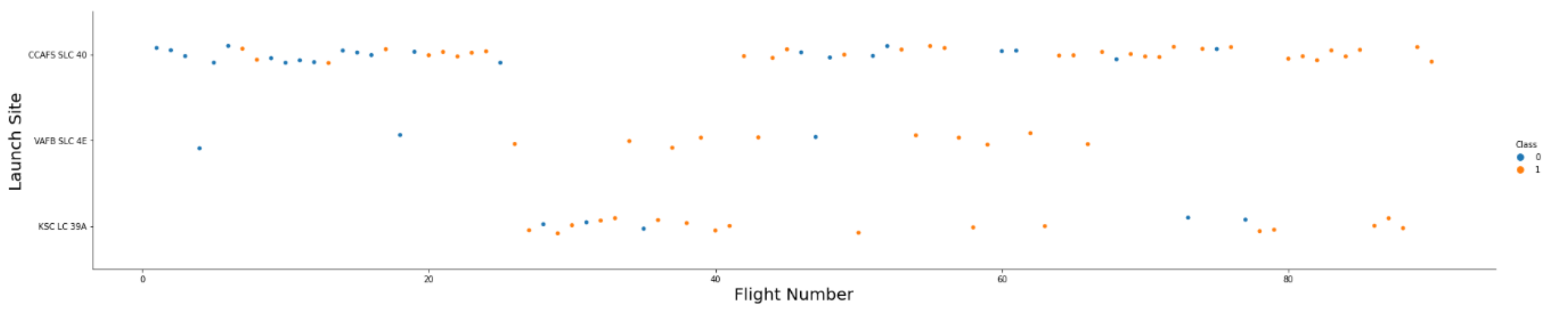
- Predictive analysis results
  - The Decision Tree Classifier is shown to be the best model to predict successful landings, with the accuracy of 84.7% on the training set and 88.89% on the test set.



Section 2

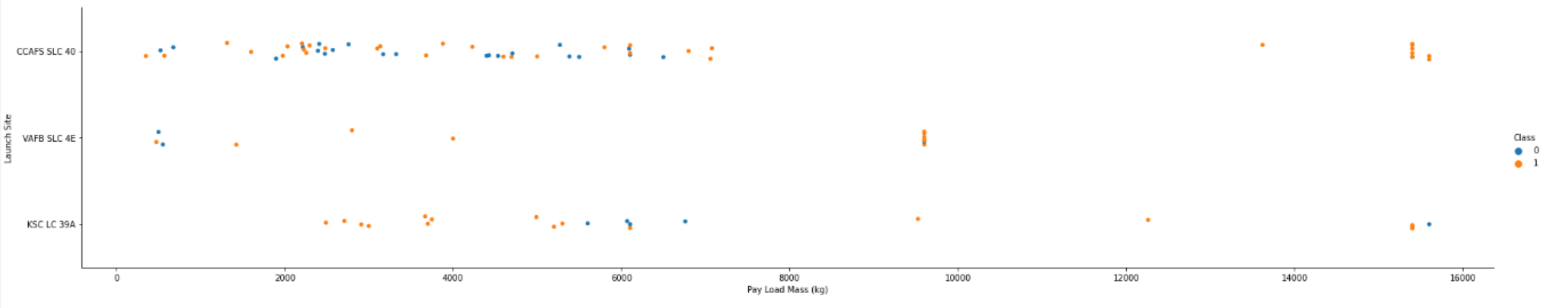
Insights drawn from EDA

Flight Number vs. Launch Site



- It can be seen that the CCAF5 SLC 40 is the best launch site recently, as most of the latest launches were successful.
- VAFB SLC 4E in second place and KSC LC 39A in third place.
- The success rate of missions are growing.

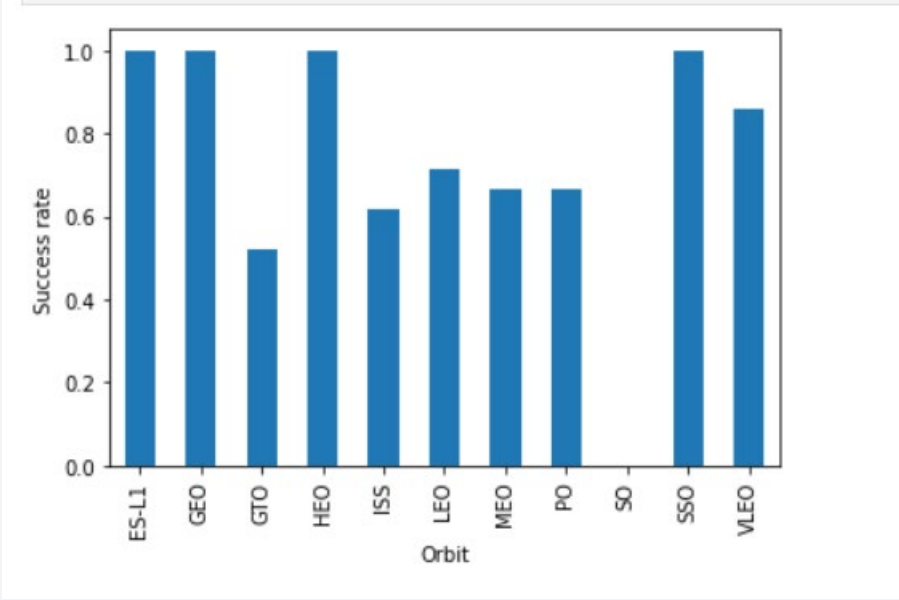
Payload vs. Launch Site



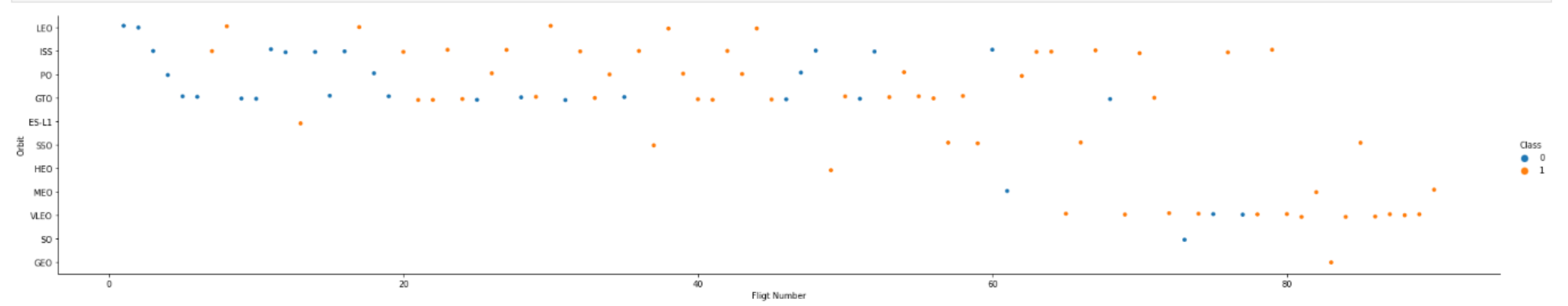
- Missions which have payload mass over 8,000kg (about 4 to 5 an average car) have a high success rate.
- VAFB SLC 4E does not have launches with payload more than 10,000 kg.

Success Rate vs. Orbit Type

- The orbits which have biggest success rate are: ES-L1, GEO, HEO, and SSO.
- SO seems to have no successful landing outcome.

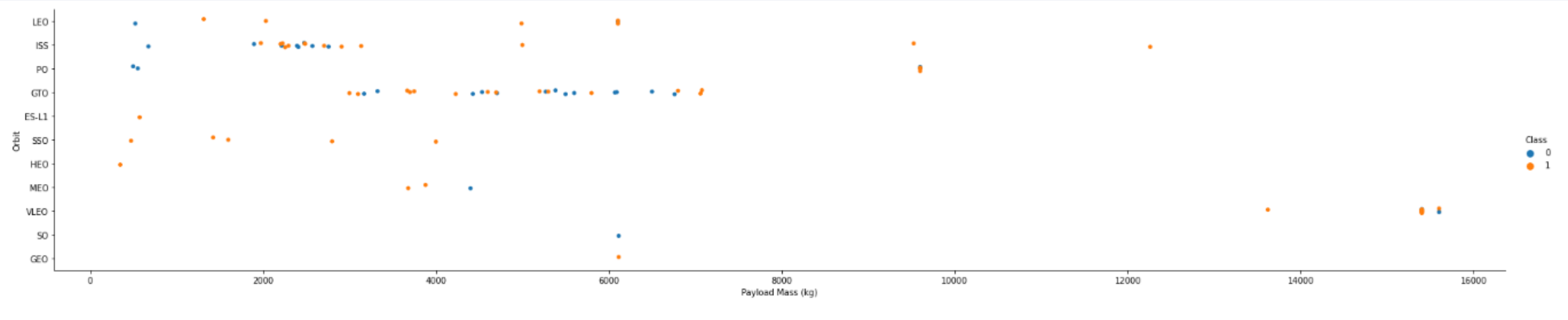


Flight Number vs. Orbit Type



- Generally, success rates of all orbits increase over the period.
- Most of the recent launches happen at VLEO launch site.

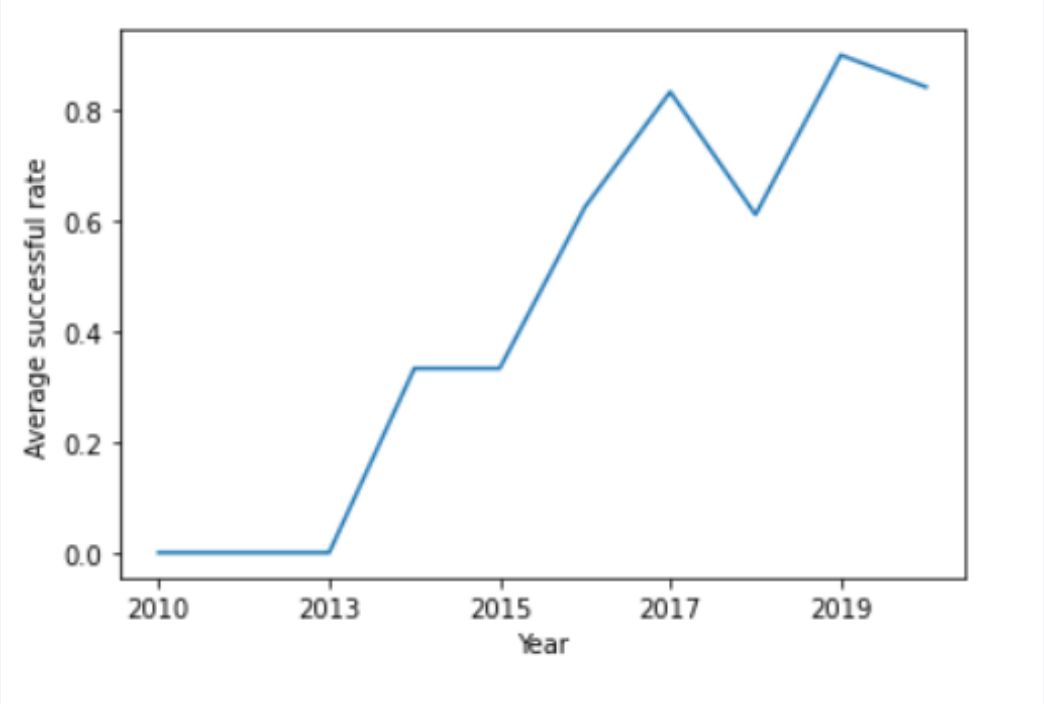
Payload vs. Orbit Type



- It can be seen clearly that there is no relation between payload mass and success rate in orbit GTO.
- SO and GEO has only a few launches.
- ISS has a wide range of payload and a significant rate of success.

Launch Success Yearly Trend

- From 2010 to 2013, the rate remained the same. This period may be used to improve the technology.
- From 2013, the success rate started to grow until 2020, however it had two falling points: in 2018 and 2020. The one in 2020 can be explained by the affect of COVID-19 pandemic, but I cannot think of a reason for the decrease in 2018.



All Launch Site Names

- The names was obtained by selecting the unique occurrences of “launch_site”.
- There are **four** launch sites in total.

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites 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

Total Payload Mass

- By summing all payload mass whose code contain ‘CRS’, corresponding to NASA, we obtain the sum of 111,268 kg.

total_payload
111268

Average Payload Mass by F9 v1.1

- The result was obtained by filtering records that carried by booster version F9 v1.1, and finding the average.
- The average was 2,928 kg.

avg_payload
2928

First Successful Ground Landing Date

- By choosing the earliest between successful ground pad landing missions, we know that it was taken on December 22, 2015, five years after the first launching.

first_successful_date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- By selecting the booster versions which carried the payload mass between 4000 and 6000 kg, we gain 4 results.

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

Total Number of Successful and Failure Mission Outcomes

- Group mission outcomes and we get the result.
- We can see that most of the launches have a successful mission outcome.

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

Boosters Carried Maximum Payload

- There are many booster versions carried 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

- Failed landing outcomes in 2015 with their booster version and launch site.
- There are only 2 results.

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- This is the ranking of counts of successful landing outcome between June 4, 2010 and March 20, 2017.

landing_outcome	total
Success (drone ship)	5
Success (ground pad)	3

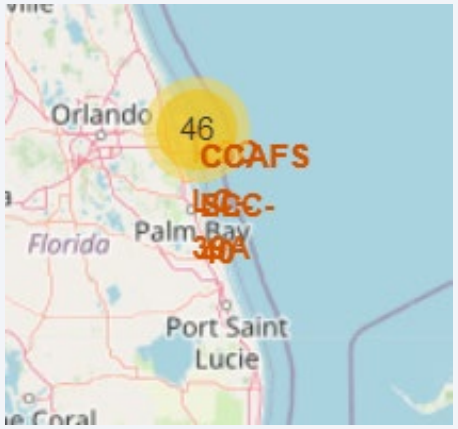

Section 3
Launch Sites
Proximities Analysis

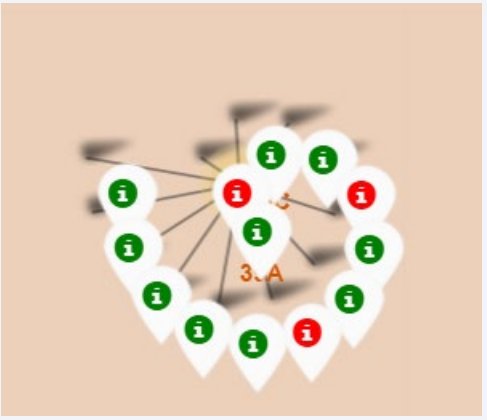
All launch sites

All the launch sites are located in safe places, nearby the oceans, but not far from the roads and railroads.

Landing Outcomes by Launch Site

- Example of KSC LC-39A launch site by landing outcomes

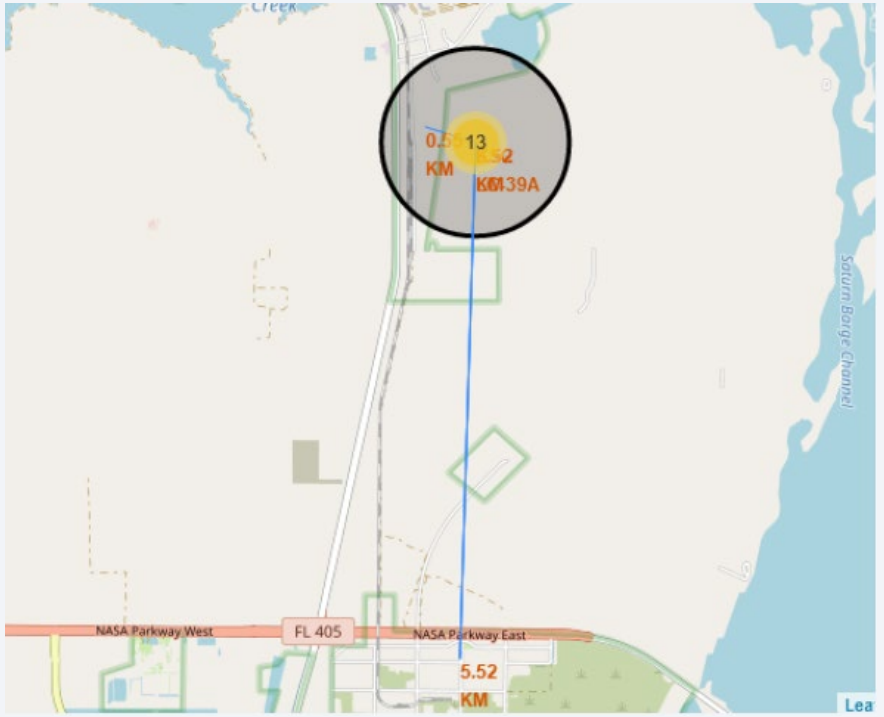


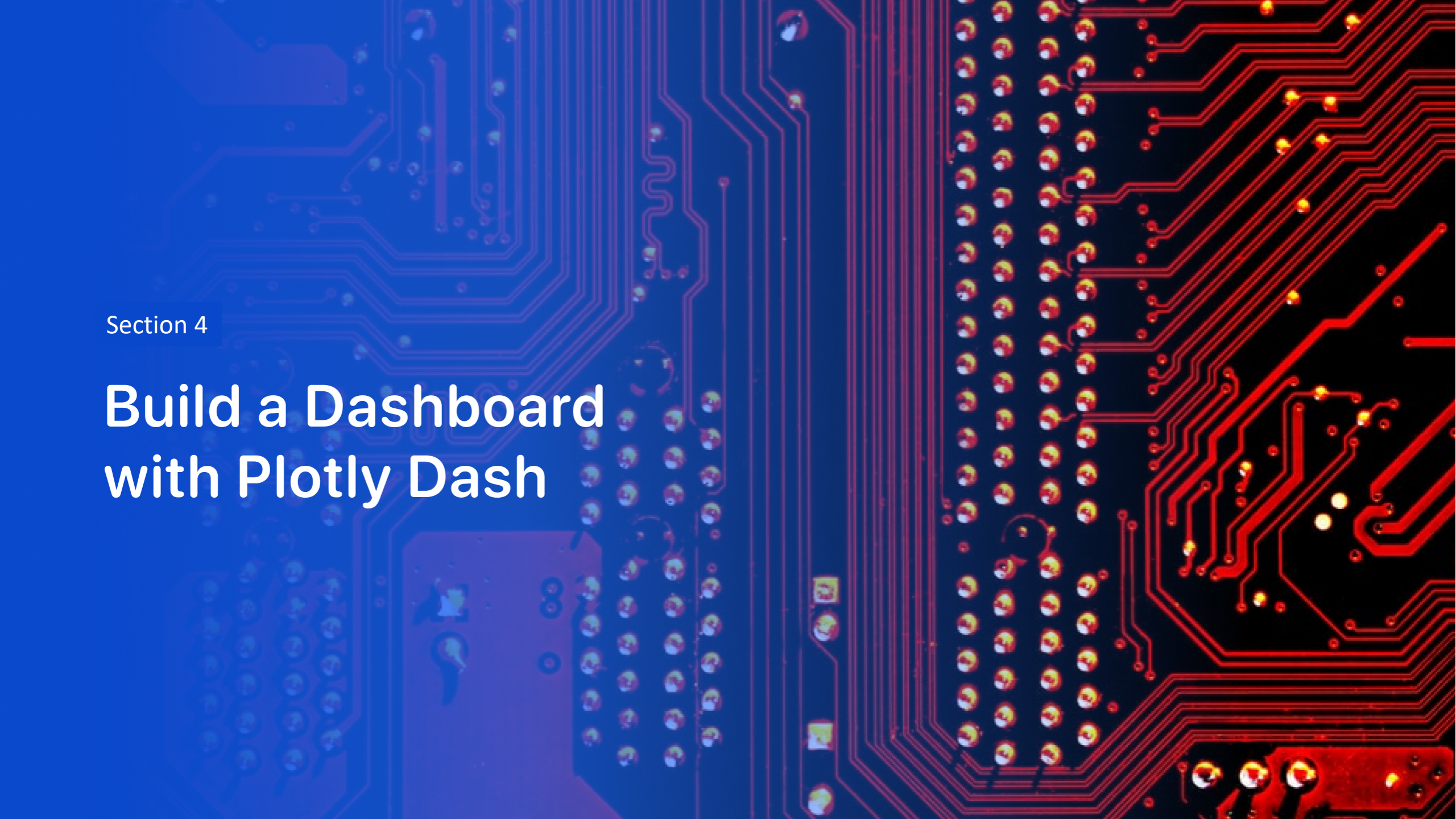


- Green markers and red markers indicate the successful and fail landing outcome, respectively.

Transportation and Safety

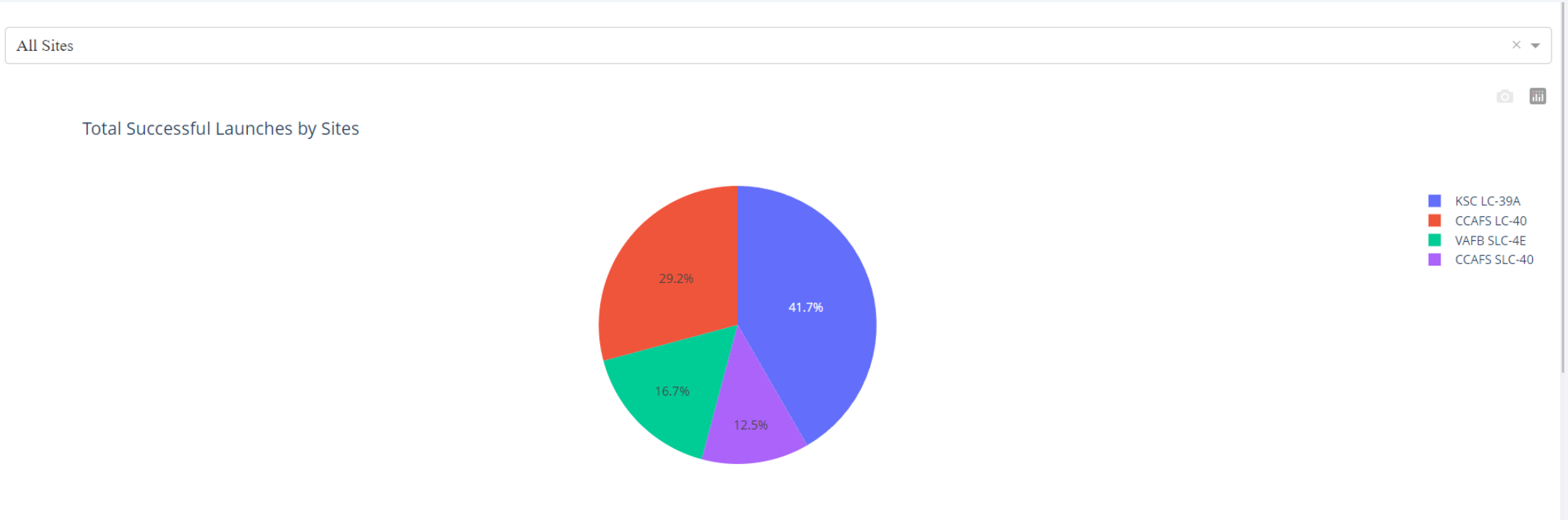
- All the launch sites are located nearby the seas, not far from the roads and railroads, which provides a good logistic condition for them. Moreover, they all have a significant distance to residential areas.



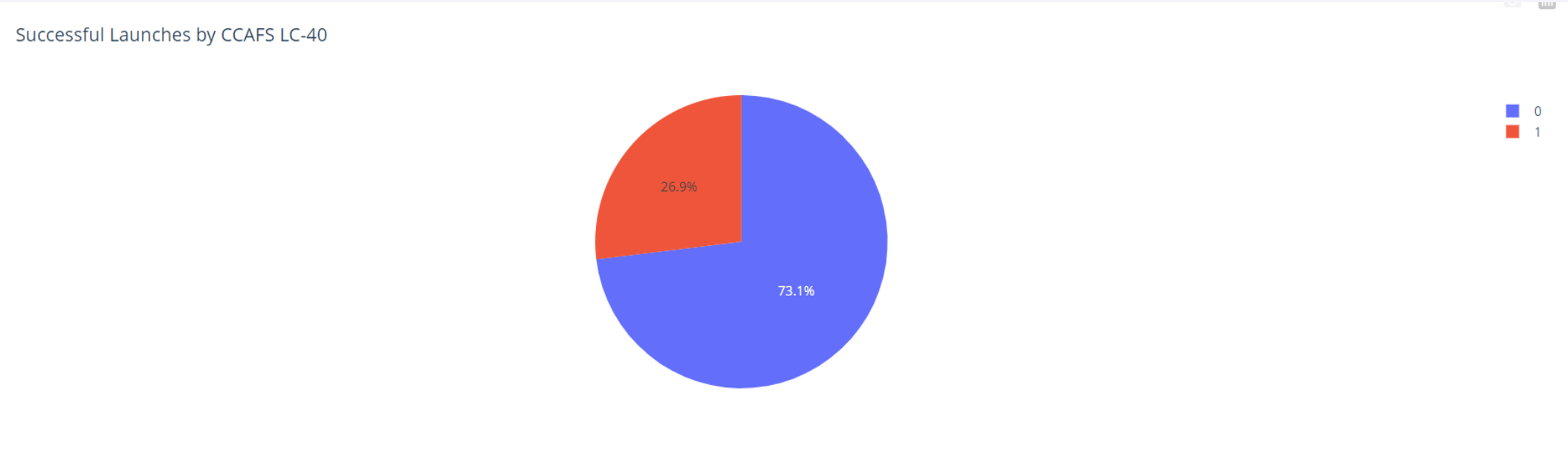

Section 4
Build a Dashboard with Plotly Dash

Successful Launches by Sites

The chart below illustrates the percentage of successful launches by sites.

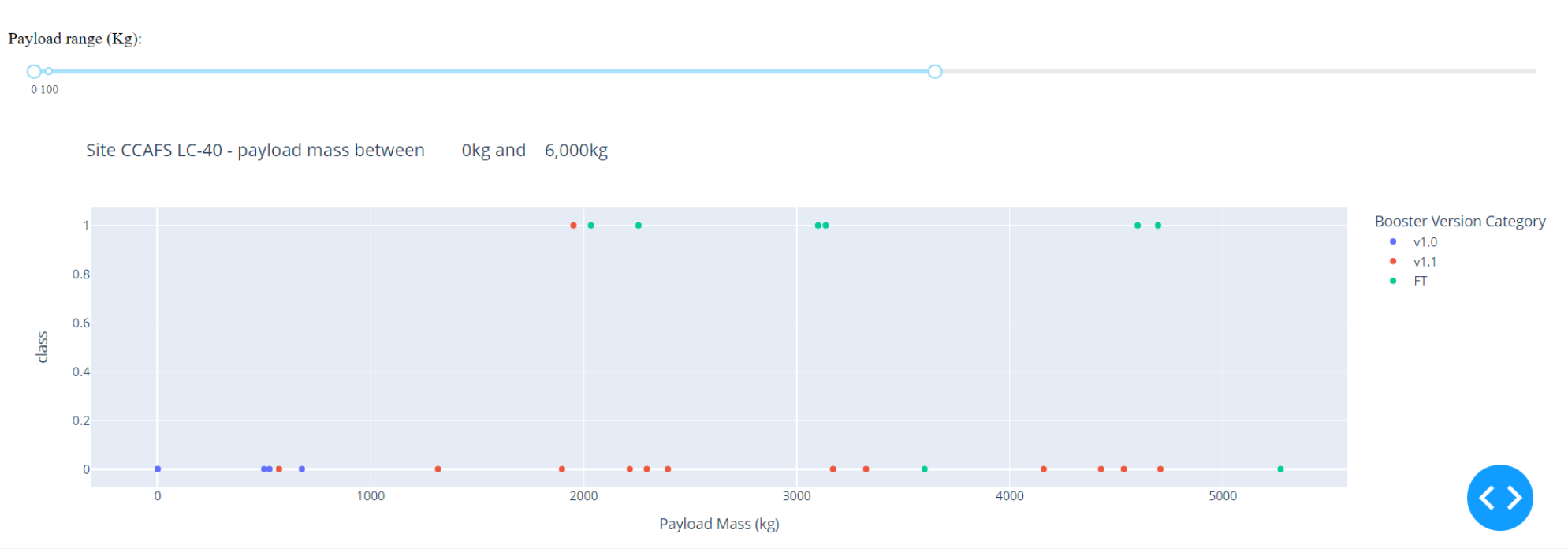


Success Rates of CCAFS LC-40



73.1% launches from CCAFS site are successful.

Payload Mass vs Success Rate



It can be seen from the plot that with the payload mass less than 6000kg at CCAFS LC-40 launch site, the FT booster version tends to get higher rate of success.

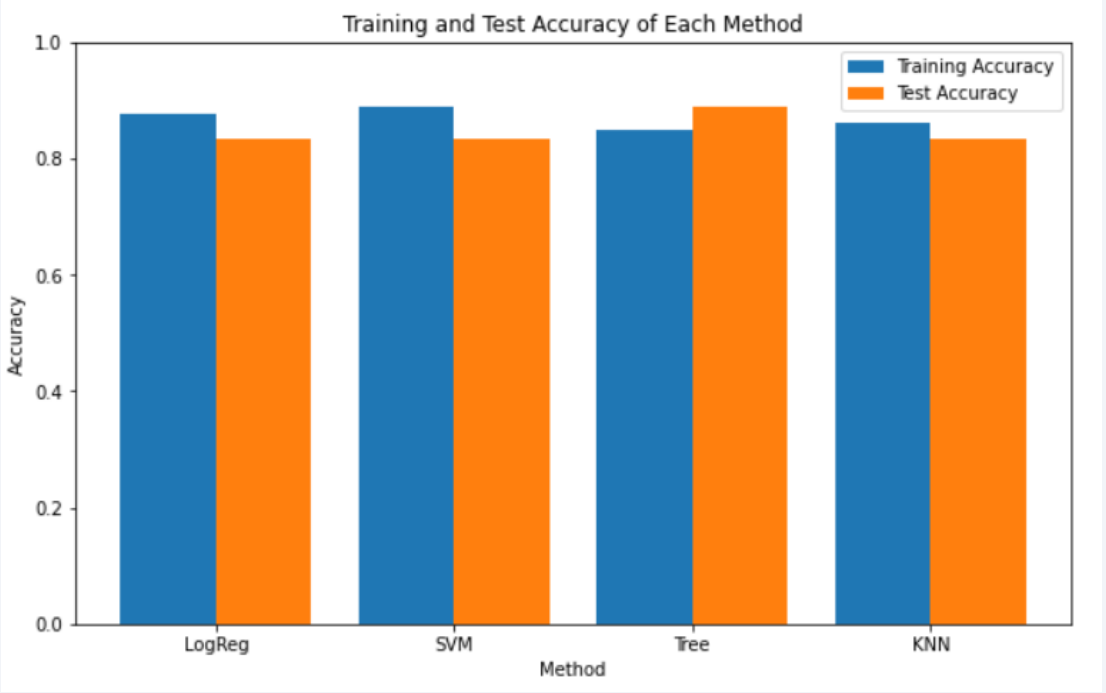
Section 5

Predictive Analysis (Classification)

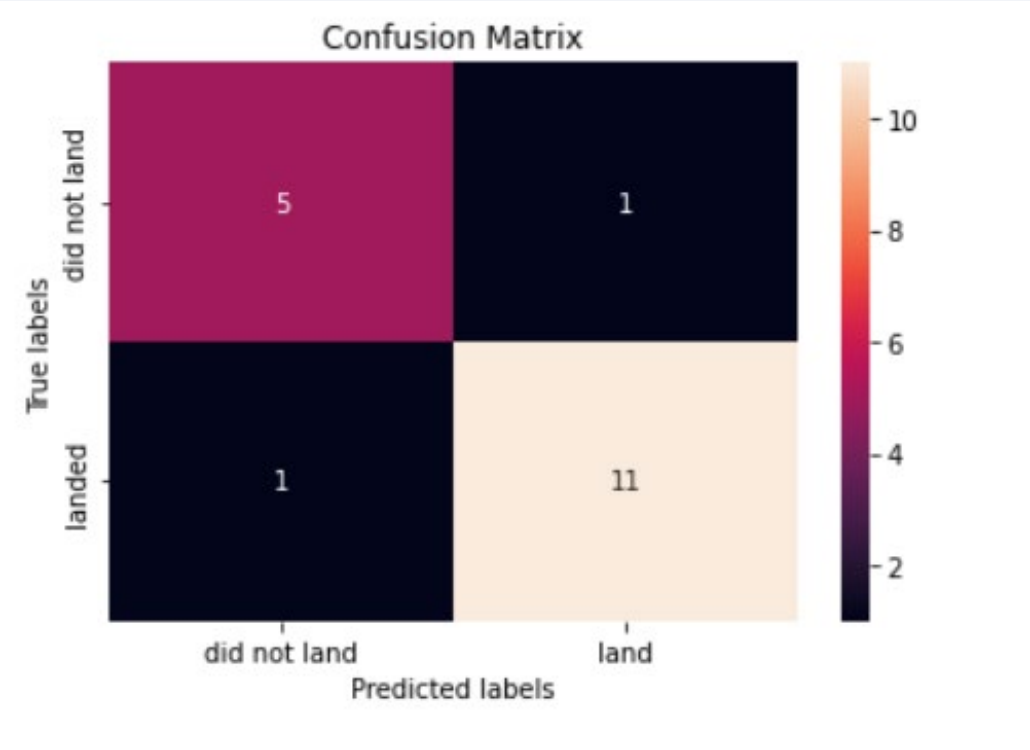
Classification Accuracy

- Training and Testing Accuracy of four models are in the charts beside.

- The Decision Tree has the highest classification accuracy on test set, about 88.89%.



Confusion Matrix



Confusion Matrix of Decision Tree proves its accuracy by showing a big number of true positive and true negative.

Conclusions

- The project collects and analyzes different data to make decision.
- The best launch site is KSC LC-39A.
- Launches with payload mass more than 8000kg tend to be safe.
- Successful landing outcome seems to improve over the time, the number of failed landing are close to zero, due to the evolution of technology.
- Decision Tree Classifier is the best model for predicting the success of a landing, with our data.

Appendix

- All of codes in this project are stored here.
- GitHub does not show maps created by Folium, so if you need to view it, you have to download the notebooks.

Thank you!