IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Zin Min Thant
22-Oct-2021

Outline

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

Executive Summary

- Summary of methodologies
    - Data Collection
    - EDA with Data Visualization
    - EDA with SQL
    - Interactive maps with Folium
    - Dashboards with Plotly
    - ML Predictive analysis
- Summary of all results
    - Preliminary analysis based on EDA
    - Interactive maps and dashboards
    - Predictive results

Introduction

- Project background and context
  - The aim is to predict if the Falcon 9 first stage will successfully land on its ground base after being launched. The core of Falcon project is to reuse this first stage propulsion system, which is then traduced into a strong cost saving. By determining the chances of a certain mission to be successfully (to land safely), we can estimate the cost of the missions.
- Problems you want to find answers
  - The main variables that participate into the Falcon 9 first stage success/fail.
  - How does the main variables interact

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX rest API
  - Web scrapping from Wikipedia
- Perform data wrangling
  - Landing outcome was defined successful (1) or failed (0).
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - KNN, Decision Tree, SVM and Logistic Regression models

Data Collection

- The following datasets were collected from Rest SpaceX API:
  - Rocket launch data from SpaceX
  - Rockets - to learn the booster name
  - Launchpads - to know the name of the launch site being used, the longitude, and the latitude
  - Payloads - to learn the mass of the payload and the orbit that it is going to
  - Cores - to learn the outcome of the landing, the type of the landing, number of flights with that core, etc

Data Collection – SpaceX API

- GitHub URL of the completed SpaceX API calls notebook

1. Get data from API
2. Normalize data into data frame
3. Append data for Rockets, Launch Pads, Payloads and Cores
4. Construct new data frame for launches
5. Filter Falcon 9 Launches

Data Collection - Scraping

- [GitHub URL of the completed web scraping notebook]

1. Request the Falcon9 Launch Wiki page

2. Extract all column/variable names

3. Create a data frame by parsing the launch HTML tables

Data Wrangling

- Missing payload data were replaced.
- GitHub URL of the Falcon 9 launches missing payload data wrangling related notebook
- Successful and unsuccessful launches were defined.
- GitHub URL of the Falcon 9 successful and unsuccessful launches data wrangling related notebook

1. **Request the Falcon9 Launch Wiki page**
2. **Extract all column/variable names**
3. **Create a data frame by parsing the launch HTML tables**

EDA with Data Visualization

- Scatter plots to visualize relationship between variables
  - FlightNumber vs. PayloadMass
  - FlightNumber vs LaunchSite
  - Payload and Launch Site
  - FlightNumber and Orbit type
  - Payload and Orbit type
- Bar plot to compare success rate of each orbit type
- Line plot to visualize the launch success yearly trend
- GitHub URL of the completed EDA with data visualization notebook

EDA with SQL

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA’
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was achieved.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster versions which have carried the maximum payload mass. Use a subquery
- List the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Rank 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

- GitHub URL of the completed EDA with SQL notebook

Build an Interactive Map with Folium

- Objects created and added to Folium map:
  - A folium Map object, with an initial center location to be NASA Johnson Space Center at Houston, Texas
  - A blue circle at NASA Johnson Space Center's coordinate with a popup label showing its name
  - A circle for each launch site in data frame launch_sites
  - For each launch site, added a Circle object based on its coordinate (Lat, Long) values
  - Markers for all launch records
  - Marker clusters to group points with the same coordinates but different information
  - Calculate the distances between a launch site to its proximities and plot distance and a line between points
- All the objects were added in order to improve the understanding of the problem, by localizing all ground stations and visualize the number of launches in each one and its outcome.
- GitHub URL of the completed interactive map with Folium map

Build a Dashboard with Plotly Dash

- Plots/graphs and interactions added to a dashboard:
- Plot charts:
  - Showing the total launches by a certain site or all sites
  - Displaying the relative proportions of launches
  - If a single site was selected, the pie chart represents the proportion between success and fails.
- Scatter charts:
  - Relationship between variables to understand how is the relations between them
  - There was added a slider to filter by Payload mass, in order to filter by different ranges of weight.
- GitHub URL of the completed Plotly Dash lab

Predictive Analysis (Classification)

- Build the model by using scikit-learn package
- Find the best parameters using GridSearch
- Evaluate the model using score and confusion matrix
- GitHub URL of the completed predictive analysis lab

1. Transform the data to normalize and therefore improve the performance of machine learning algorithms
2. Split into training and test sets
3. Select several algorithms to be used (Logistic Regression, Support Vector Machine, Decision Tree and K Nearest Neighbors)
4. Set the ranges of variables used for the selected algorithms and load them into GridSearch for parameter optimization
5. Accuracy checks were performed to every model
6. Plot confusion matrices for every model

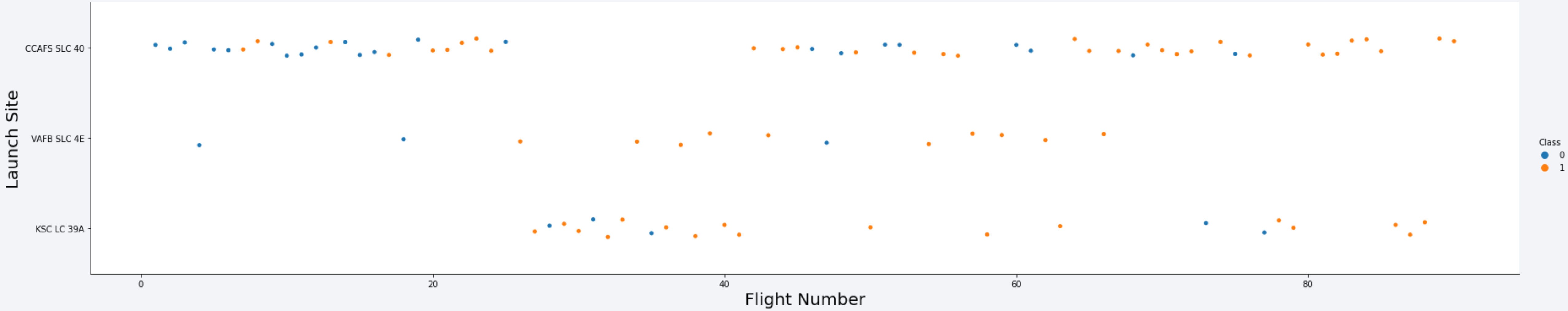
Results

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

Section 2

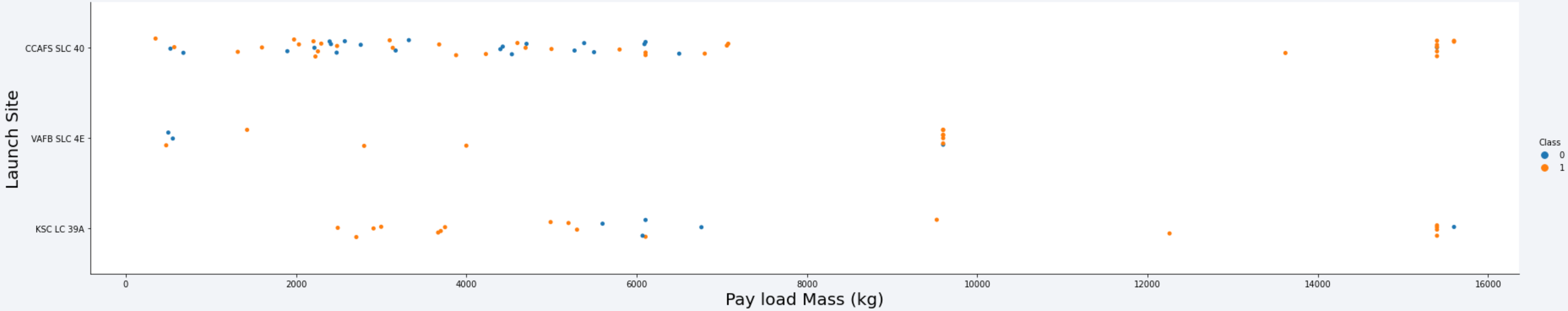
Insights drawn from EDA

Flight Number vs. Launch Site



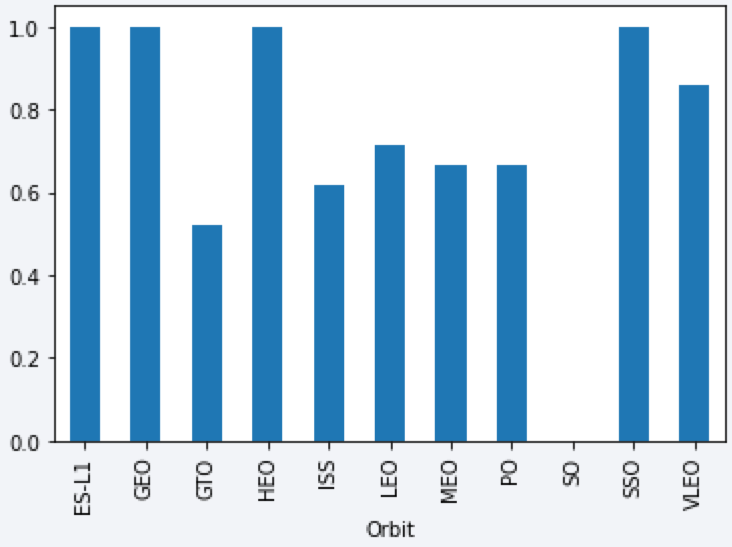
For every site, the outcome improves with the amount of launches

Payload vs. Launch Site



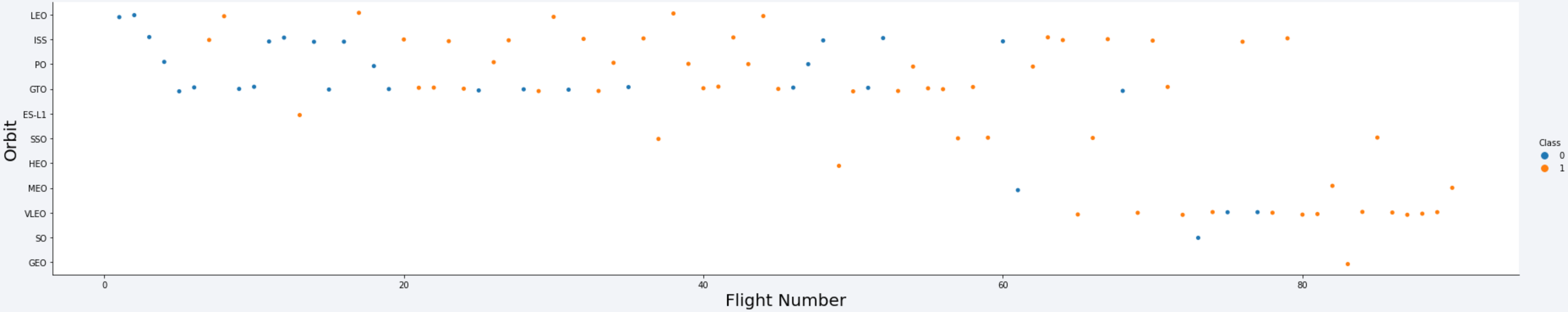
With increasing payload for site SLC 40, the success rate improves. This can be related to the test missions (lower payload) versus the "real" missions, with higher payload.

Success Rate vs. Orbit Type



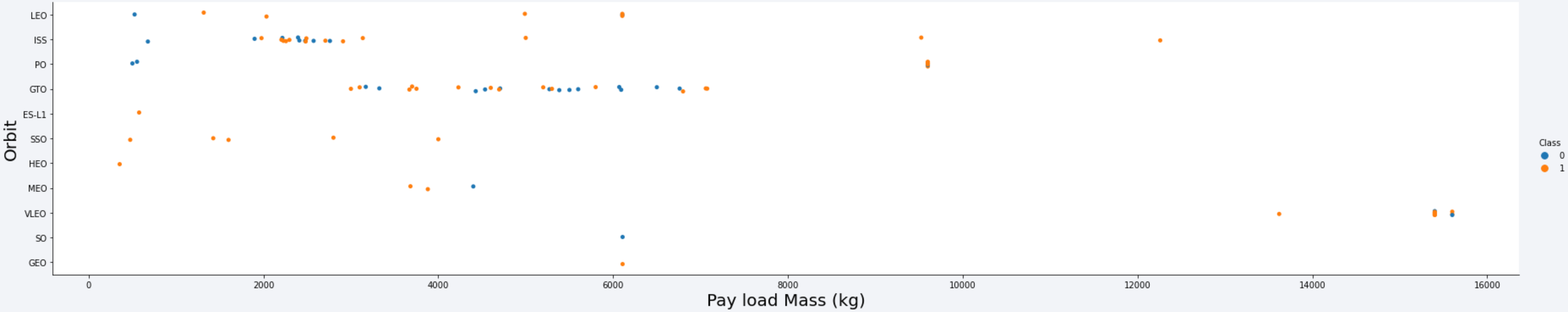
Orbits ES-L1, GEO, HEO and SSO have the better success rates.

Flight Number vs. Orbit Type



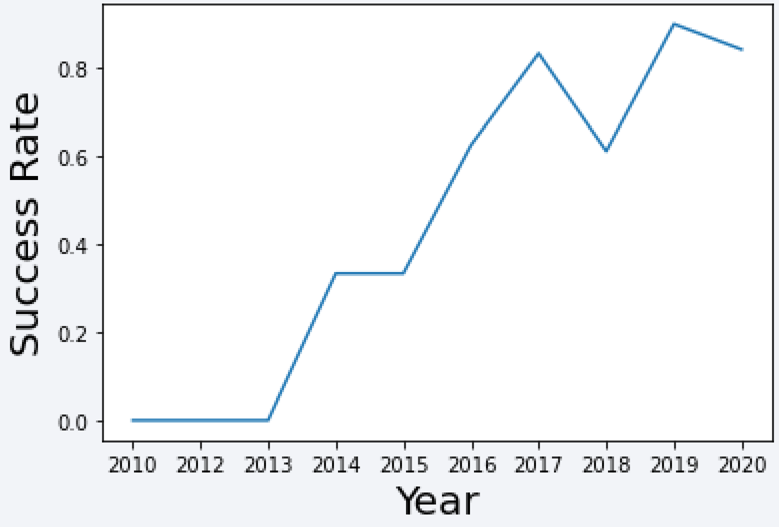
In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS. However, for GTO we cannot distinguish this well as both positive landing rate and negative landing (unsuccessful mission) are both there.

Launch Success Yearly Trend



The success rate since 2013 kept increasing till 2020.

All Launch Site Names

- %sql select Distinct Launch_Site from SPACEXTBL
- By using DISTINCT we just keep the different objects of launch_site table

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

Launch Site Names Begin with 'CCA'

- %sql select * from SPACEXTBL where Launch_Site like 'CCA%' limit 5
- By using the WHERE statement followed by like, we can filter the results to a specific group.

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 NASA boosters

- %sql select Sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer = 'NASA (CRS)'
- By using SUM within the SELECT we can add all records that matches with the WHERE condition.

45596

Average Payload Mass by F9 v1.1

- %sql select Avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version = 'F9 v1.1'
- By using AVG within the SELECT, we can get the average of all records that matches with the WHERE condition.

2928

First Successful Ground Landing Date

- %sql select Min(Date) from SPACEXTBL where Landing__Outcome = 'Success (ground pad)'
- By using min, we can select the oldest date that matches with the WHERE condition. We must select only the success records.

2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- %sql select Booster_Version from SPACEXTBL where Landing__Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ between 4000 AND 6000
- We filter the payload to a desired range by using WHERE statement. We also select DISTINCT in order to obtain just a single record for every booster version.

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

Total Number of Successful and Failure Mission Outcomes

- %sql select count(*) total, Mission_Outcome from SPACEXTBL group by Mission_Outcome
- By using group by, we can count the successful and failure mission outcomes.

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

Boosters Carried Maximum Payload

- %%sql select Booster_Version, PAYLOAD_MASS__KG_ from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
- We used a subquery in order to pre filter the dataset. Then we choose distinct values from the previous selection.

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 Launch Records

- %%sql select Landing__Outcome, Booster_Version, Launch_Site from SPACEXTBL where Landing__Outcome = 'Failure (drone ship)' AND Year(Date) = '2015'
- We used YEAR function to get the year for all dates, and then select all that matches with 2015 and the outcome was a failure.

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

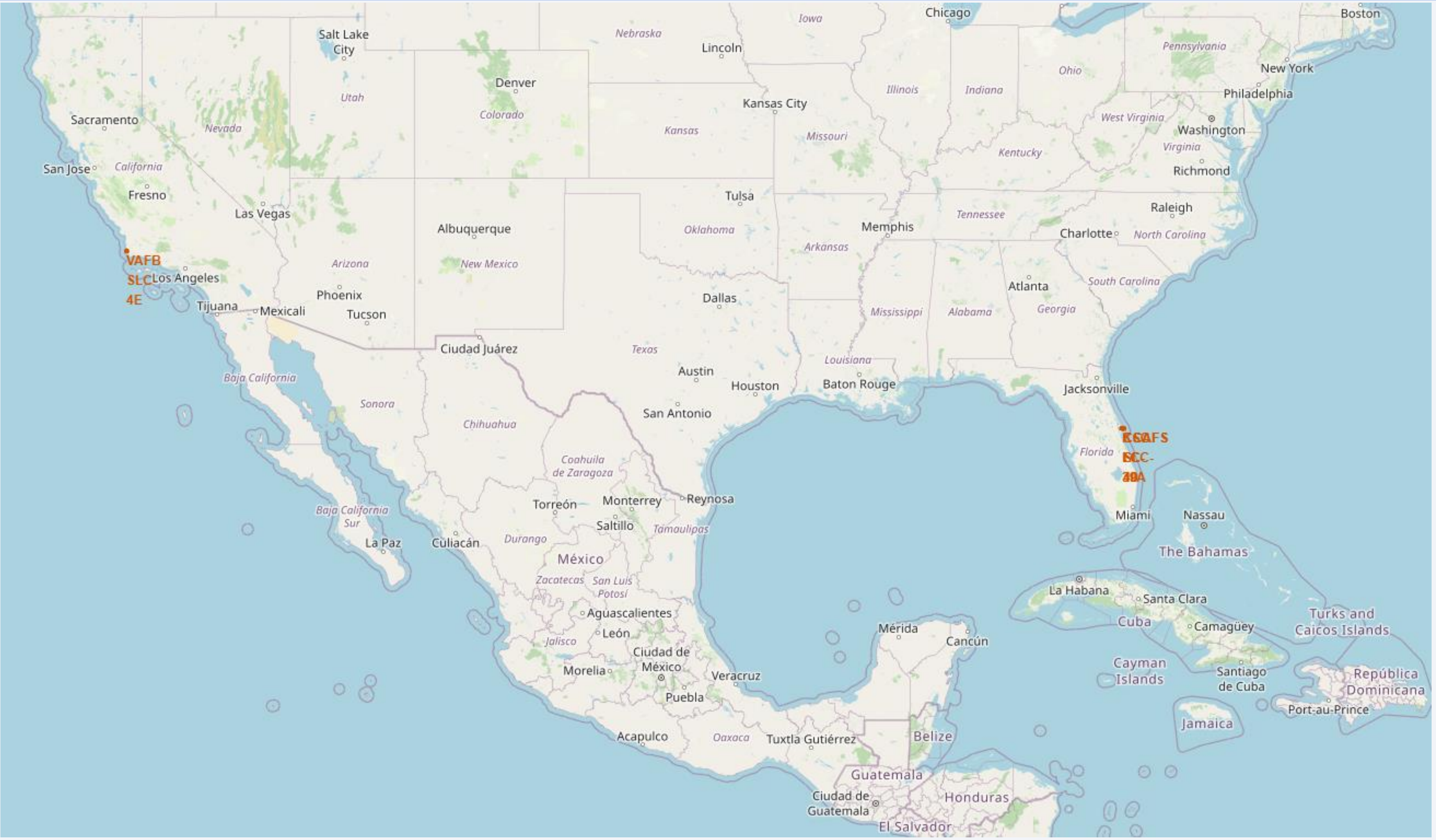
- %%sql select count(*) as total, Landing__Outcome from SPACEXTBL where Date between '2010-06-04' AND '2017-03-20' group by Landing__Outcome order by total DESC
- We used WHERE statement and group the results. Then by using COUNT we can get the matches for the required condition.

total	landing__outcome
10	No attempt
5	Failure (drone ship)
5	Success (drone ship)
3	Controlled (ocean)
3	Success (ground pad)
2	Failure (parachute)
2	Uncontrolled (ocean)
1	Precluded (drone ship)

Section 4

Launch Sites Proximities Analysis

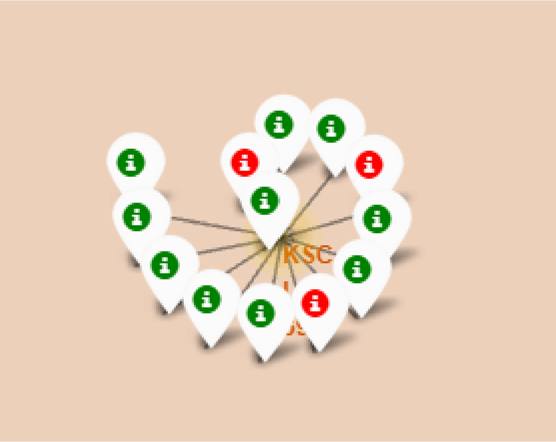
Launch Locations



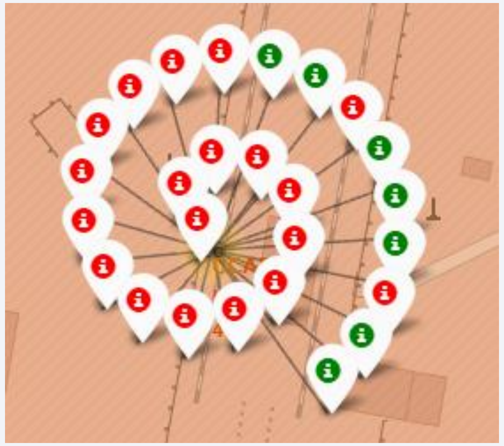
Here are the Launch sites

Mission Outcomes

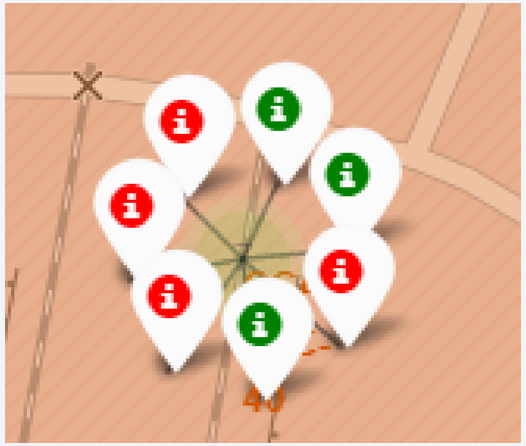
VAFB SLC-4E



KSC LC-39A



CCAFS LC-40

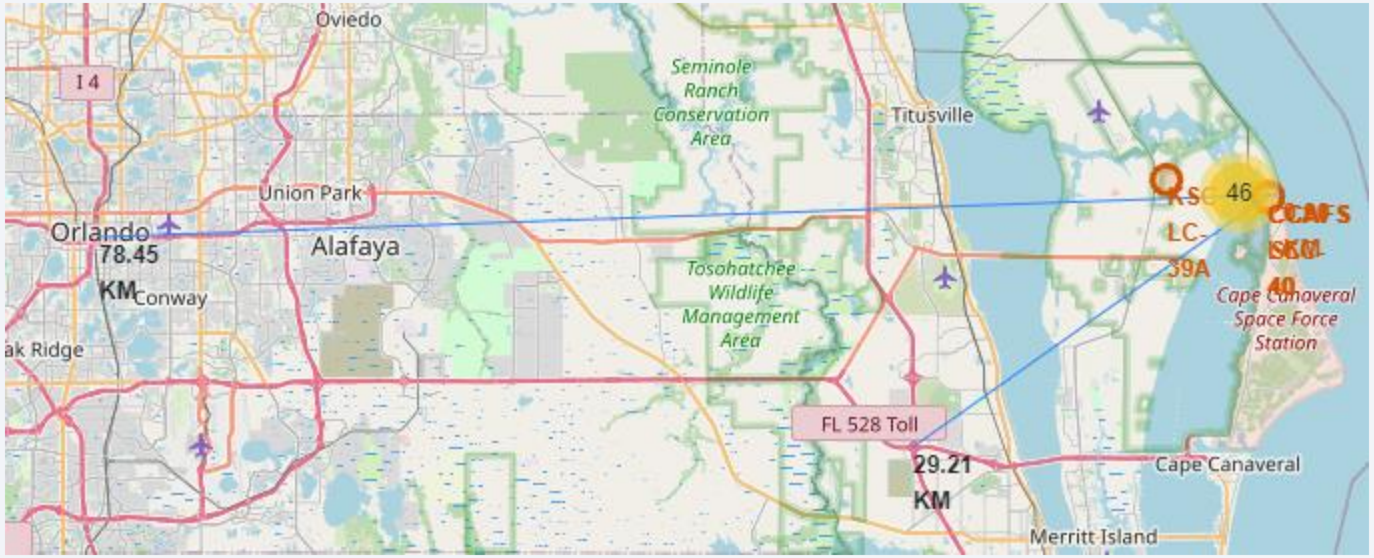


CCAFS SLC-40

Here are all the missions of every ground station and its outcome, being red if it was a failure mission, and green if it was a success one.

Distances from ground stations

Distance to Coast Line



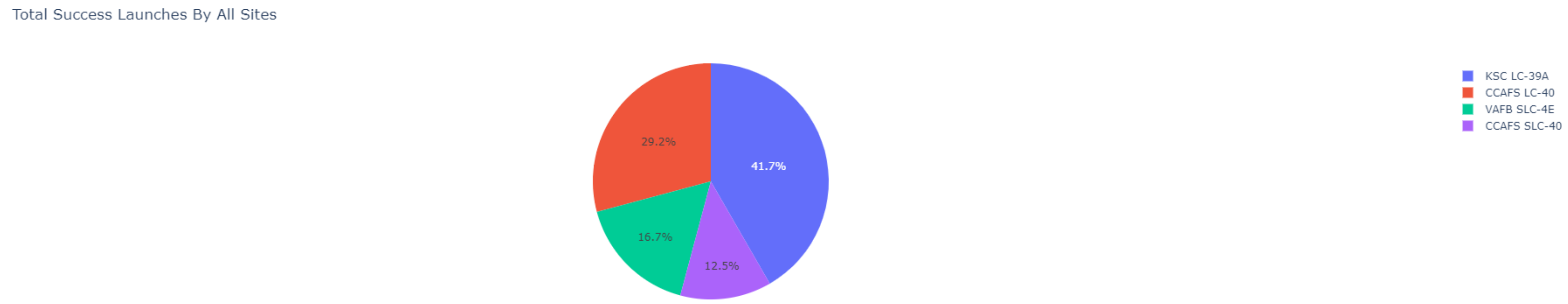
Distance to Highway and City

The distance plot to cost line, highway and Orlando city

Section 5

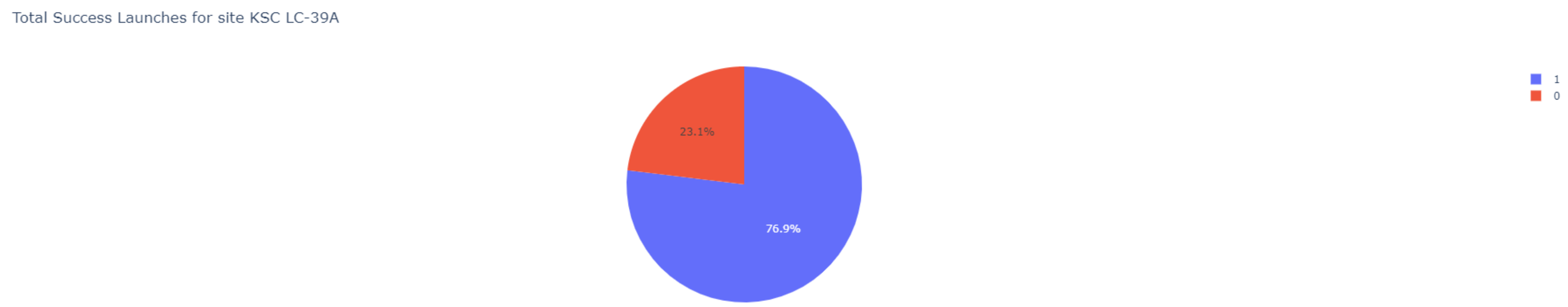
Build a Dashboard with Plotly Dash

Successful Launches for All Sites



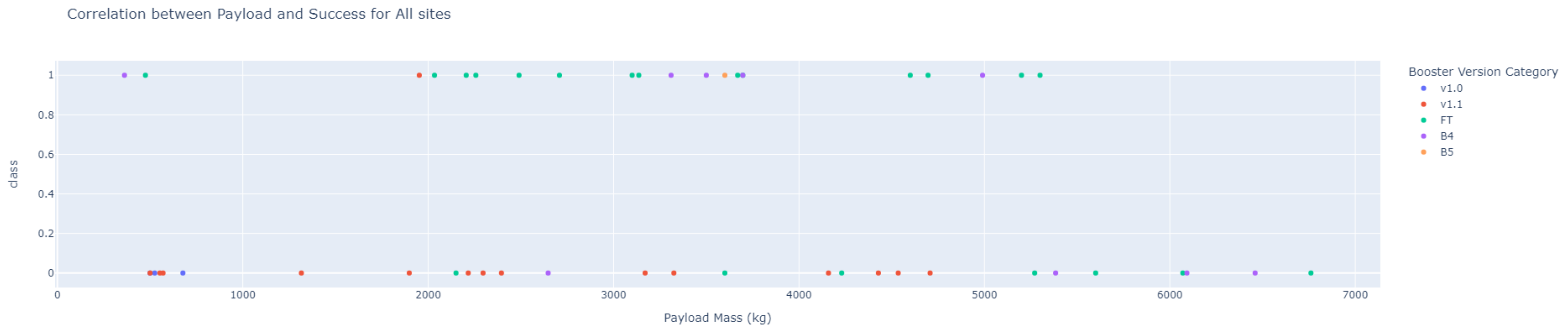
This chart how the success missions are distributed among all stations.

Launch site with highest launch success ratio



Mission outcomes for ground station KSC LC-39A, which is the one that has the greatest share in success missions among all stations.

Payload vs. Launch Outcome

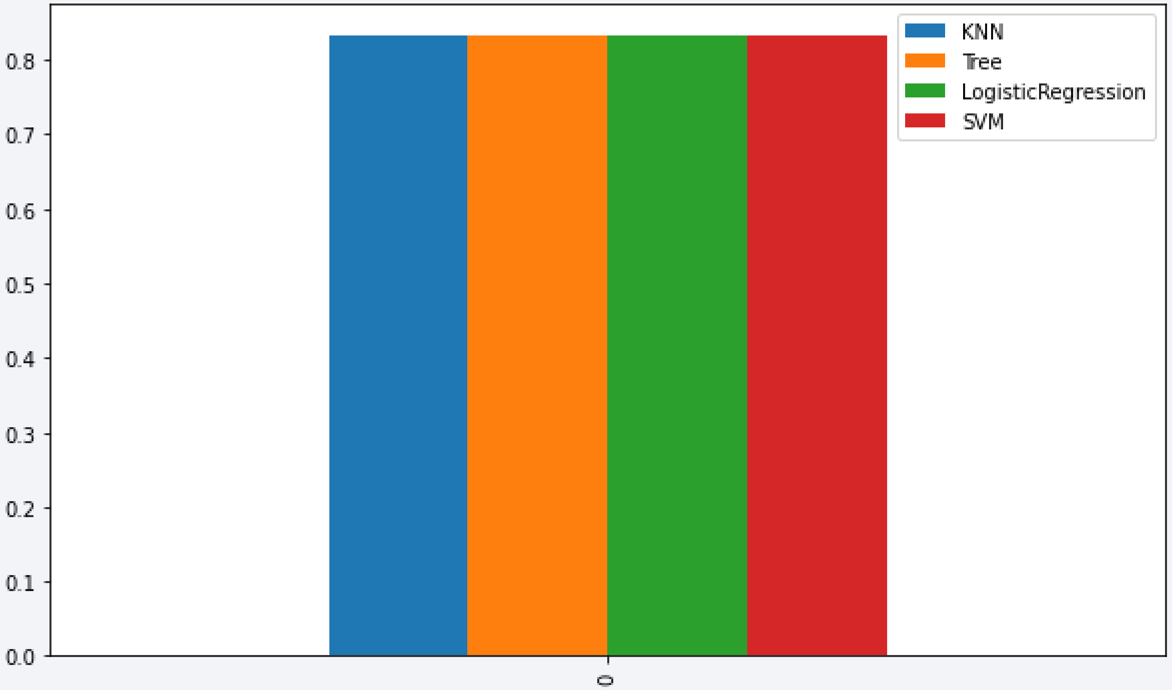


There is a “band” between 2500 and 3500 kg in which we see that Booster Version “FT” shows the best performance.

Section 6

Predictive Analysis (Classification)

Classification Accuracy



All models show the same accuracy.

Confusion Matrix

All models show the same accuracy.

Conclusions

- Mission success shows a clear correlation with number of previous launches, giving a clear indication of the SpaceX engineering team learning curve.
- Some orbits showed best success rate, which at first didn't seem to be some factor that correlates with mission success/failure.

Thank you!