Data analysis of SpaceX success

Executive summary

- Methodology
  - Data collection – API and web scraping
  - Data Wrangling – NA value and encoding
  - EDA with SQL
  - EDA with visualization
  - Interactive visualization with Folium and Dashboard
  - Predictive models
- Results
  - EDA – descriptive statistics and visualizations
  - Interactive visualizations and implications
  - Predictive analysis and model choice

Introoduction

- In this study, we perform data analysis based on SPACEX launch and landing, trying to find out the factors that could affect success launching and landing
- In the past years, SPACEX has launched several rockets into the orbit around the earth. Some of the rockets are also landed successfully. It is important to figure out the relevant factors which could help reduce waste and cost in future missions
- This analysis aims to identify relevant factors associated with successful launching and landing.

Methodology

- We collected data from SPACEX API and web scraping from Wikipedia.
- We wrangled the data to address null values as well as encoded the categorical variables
- We performed exploratory data analysis EDA with sql and matplotlib
- We built interactive visualizations with folium and Dash.
- Eventually, predictive analysis are done with classification models. Grid search cross fold validation was performed to tune the models.

Data collection

- The data is **collected** from two sources: one is from the API of SPACEX (notebook), the other is scraped from the Wikipedia (notebook)
- We requested the data from API, keeping only the columns we need, filtered the data to only include falcon 9 launches, and imputed the missing PayloadMass values with mean value.
- We used BeautifulSoup to parse Wikipedia page, and parsed launch html tables.
- (notebooks on github are put in the links above)

Data wrangling

- In the **data wrangling** notebook (notebook), we addressed the null values in launching pad, and encoded outcomes (0 for bad and 1 for good), examined the launch site and orbits.
- (notebook on github is put in the link above)

EDA

- SQL query and data visualization are two ways to perform exploratory data analysis (**EDA**).
- With these methods, we are able to get a rough idea about how the range and distribution of the data, as well as some relationships between the data columns.
- In the end of the data visualization, we also performed one-hot encoding of categorical variables
- (notebooks on github are put in the links above)

Interactive Visualization

- With Folium and Dash, the interactive visualization are built to closer examine the factors for success.
- The launch sites and launch outcomes are marked and color-coded on the map.
- On the interactive dashboard, we presented pie chats and scatter graph to explore the launches at sites and factors affecting outcome.
- (codes on github are put in the links above)

Predictive analysis

- Eventually, predictive analysis are done with classification models. Grid search cross fold validation was performed to tune the models.
- (notebook on github is put in the link above)

Results– launch site and orbits

- In the data wrangling part (notebook), we checked the percentage of missing values, where only LandingPad has 28.89% of missing values.
- There are 3 launch site. CCAFS SLC 40 has 55 launches, KSC LC 39A has 22 launches, and VAFB SLC 4E has 13 launches.
- The GTO orbit has the most (27) launches, ISS orbit has 21, VLEO has 14, PO has 9, LEO has 7, SSO has 5, MEO has 3, the rest ones: ES-L1, HEO, SO, and GEO, all have 1 launches.

Results – landing outcome

- There’re 41 successful landings to drone ships (ASDS), 14 successful landings to ground pad (RTLS), and 5 successful landings on a specific region of ocean (Ocean)
- There’re 6 unsuccessful landing for ASDS, 1 for RTLS, and 2 for Ocean.
- Failed landing includes 6 for ASDS and 19 for other occasions.
- With SQL query, it is found that in 2015, there was a failure in January and another in April, both launched at CCAFS LC-40.

Results – landing outcome cont.

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

- In the spaceX table, we found that between 06/04/2010 and 03/20/2017, there are several successful and failed landing on drone ship, ground pad, ocean.

Results – successful missions

- With SQL query, it is found that among the total 101 items in the data table, there was 1 failure in flight, 100 successful missions, with 1 successful mission that doesn’t have clear payload.
- The earliest success mission occurs on 12/22/2015.

Results – loading mass

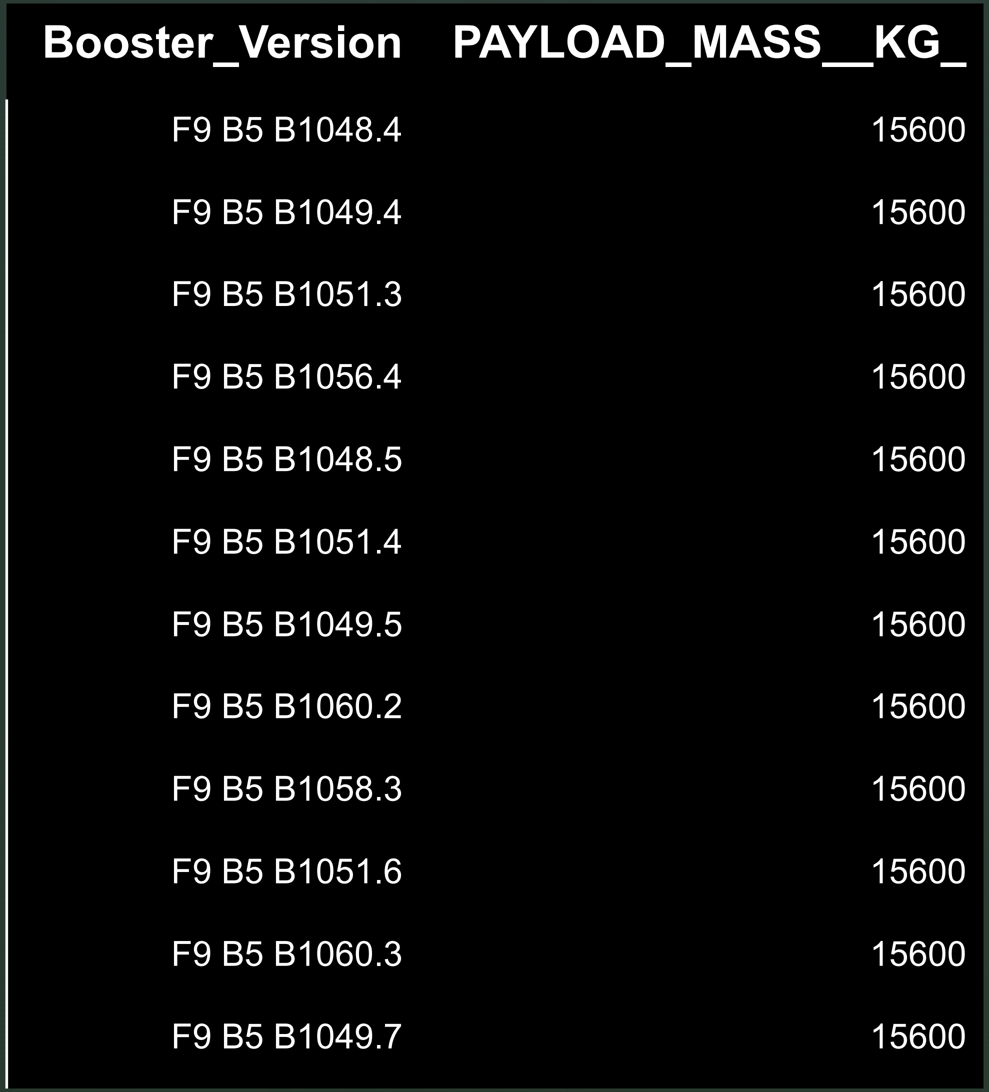
- With SQL, one can list the contents such as the unique names of sites, and items for a particular site.
- One can also perform aggregation calculations. It was found that a total of 44596 kg of loads are launched for NASA (CRS). The average payload mass for the booster F9 v1.1 is 2928.4 kg

Results – loading mass, cont.

- The boosters identified for payloads between 4000 and 6000, with successful drone ship landing outcome are listed in the table

Booster_Version
F9 FT B1021.1
F9 FT B1022
F9 FT B1023.1
F9 FT B1026
F9 FT B1029.1
F9 FT B1021.2
F9 FT B1029.2
F9 FT B1036.1
F9 FT B1038.1
F9 B4 B1041.1
F9 FT B1031.2
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1

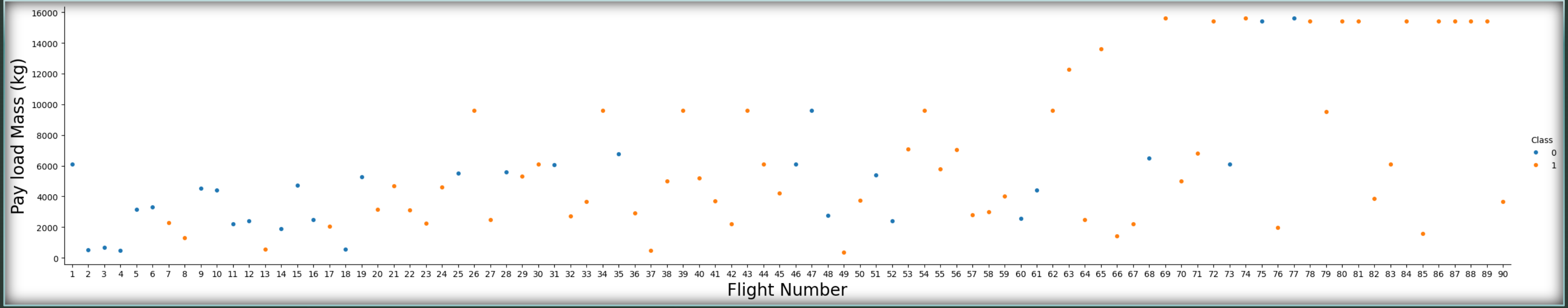
Results – boosters with the max loading mass



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

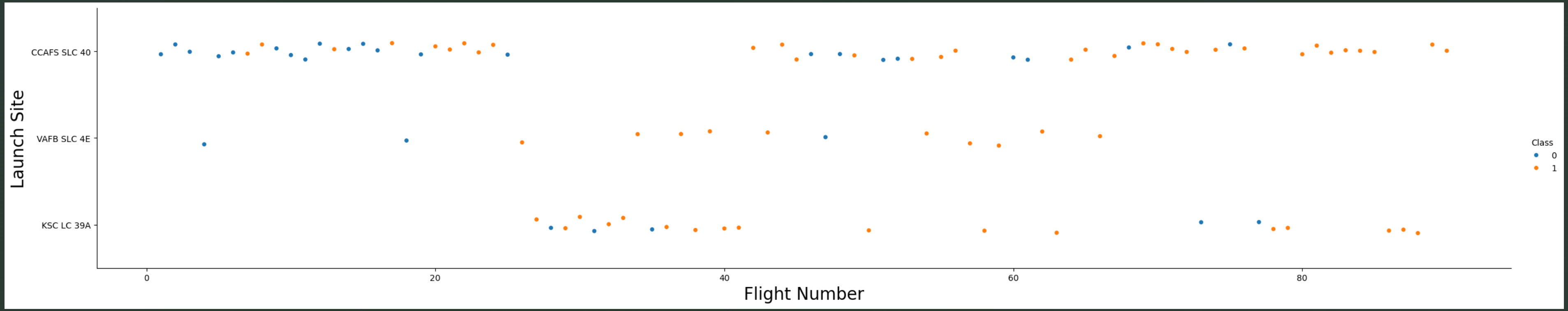
- SQL is also utilized to check which boosters load the maximum payload mass (15600 KG)

Results – payload mass and flight number



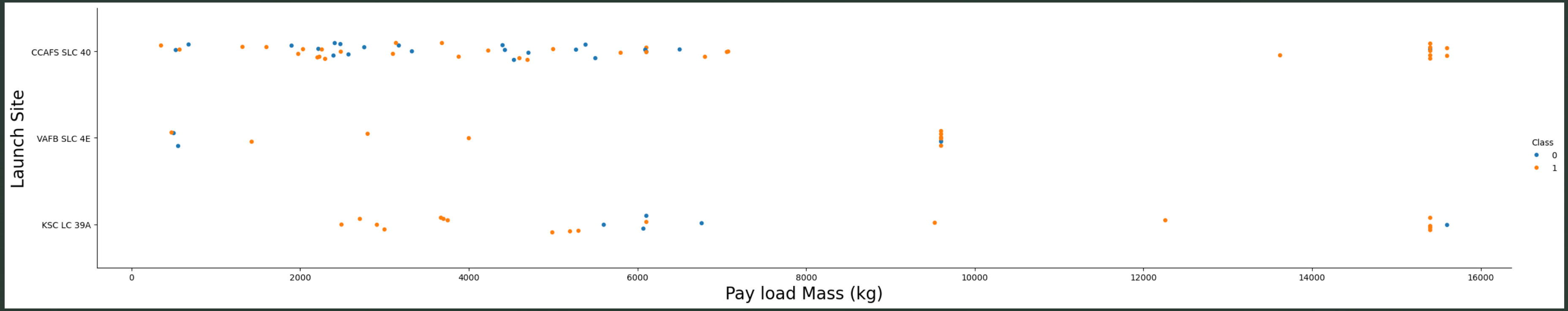
- We used matplotlib to generate scatter plots (notebook)
- We see that different launch sites have different success rates. CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.

Results – Flight Number at launch sites



- We can see that the CCAFS SLC40 launches more flights than any other two. It also has quite good success rate. (notebook)

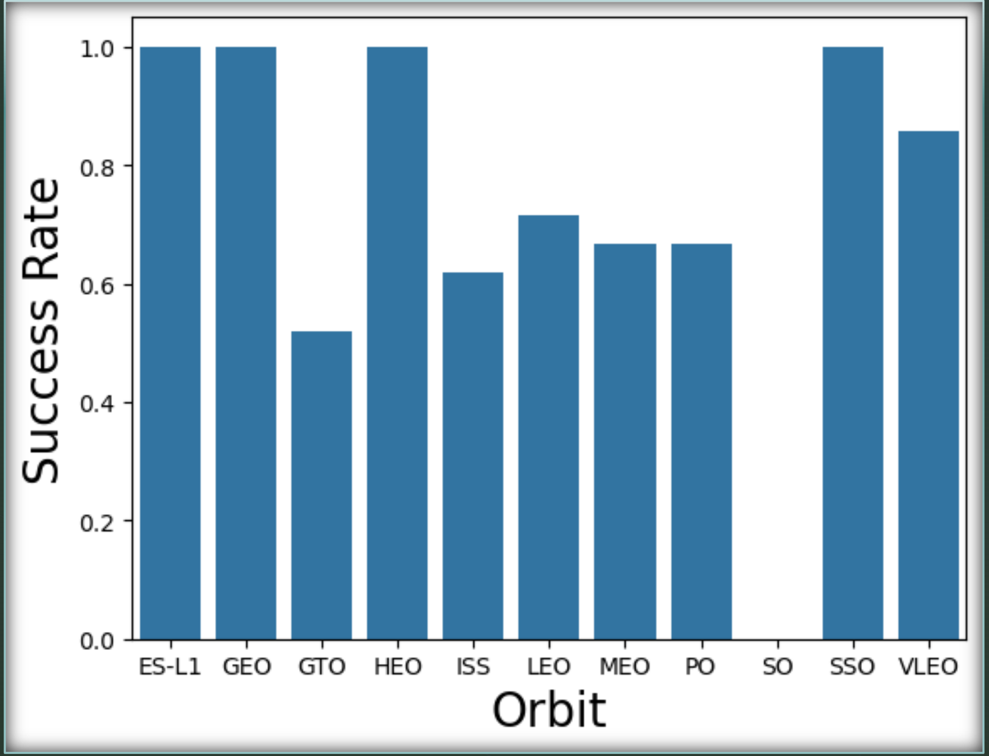
Results – Payload Mass and Launch Site



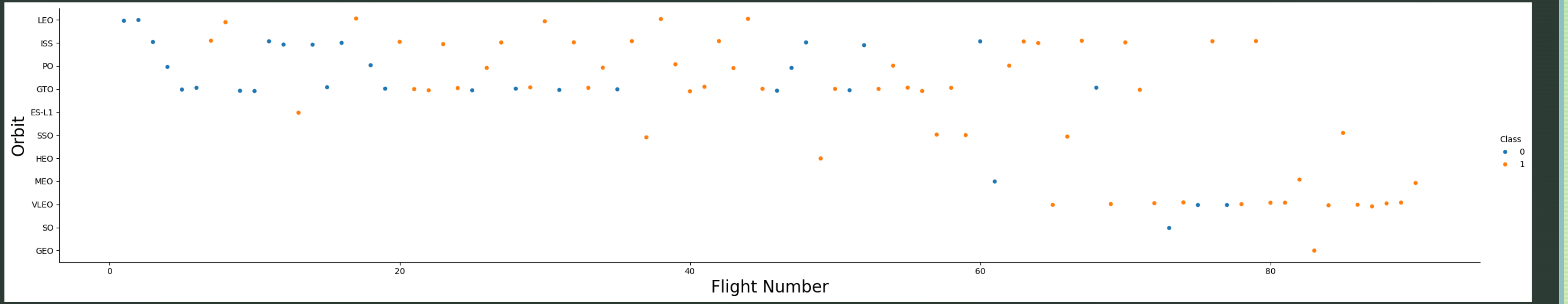
- Most of the launches are at the mass < 8000. At higher mass, only KSC LC39A succeeded once. (notebook)

Results – success vs orbit

- We found that SSO, ES-L1, GEO, and HEO all have the highest successful rate.
- (notebook)

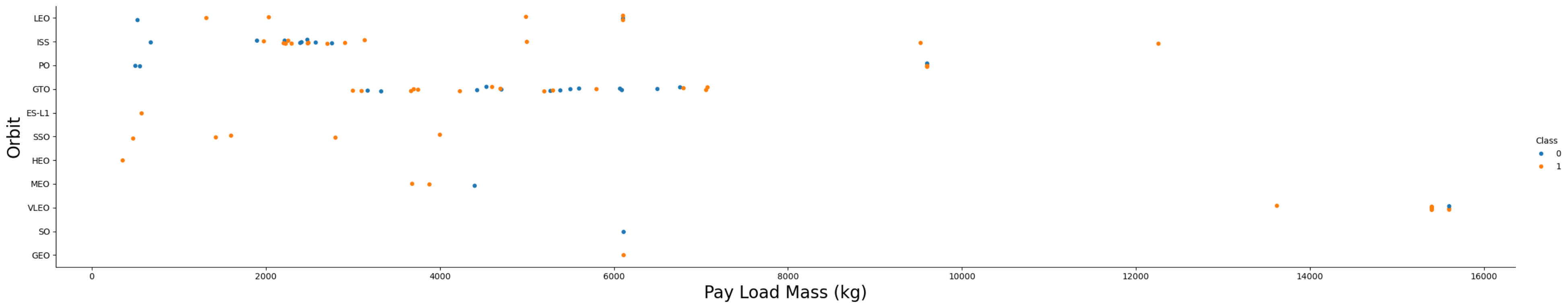


Results – success vs orbit



- 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. (notebook)

Result – Payload mass and orbit



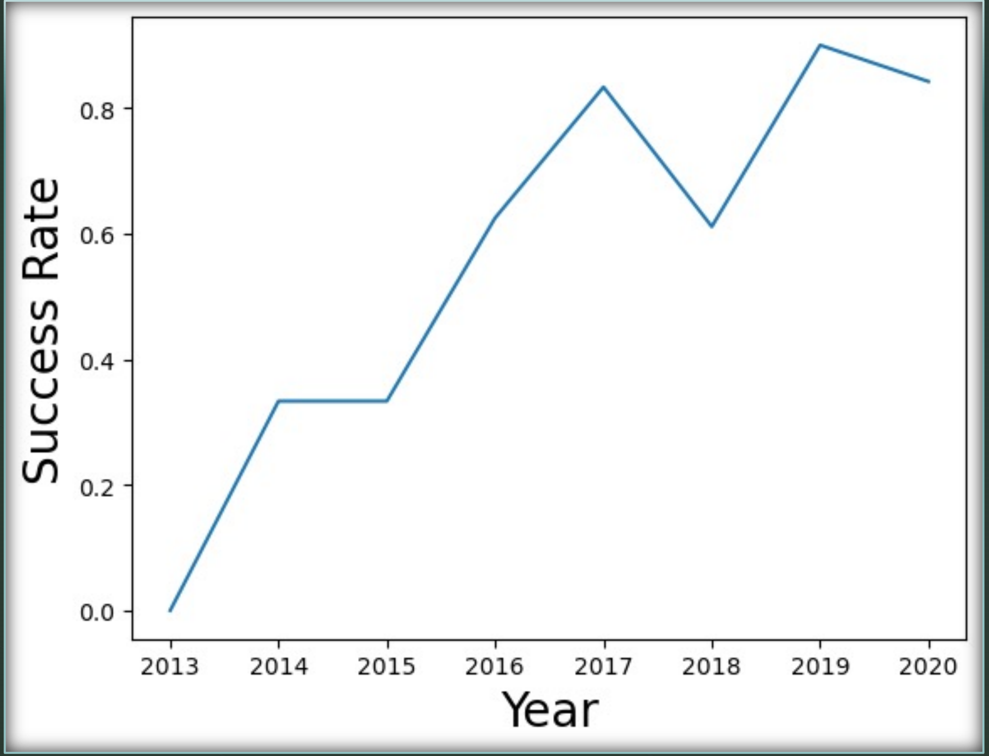
- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.

- However, for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

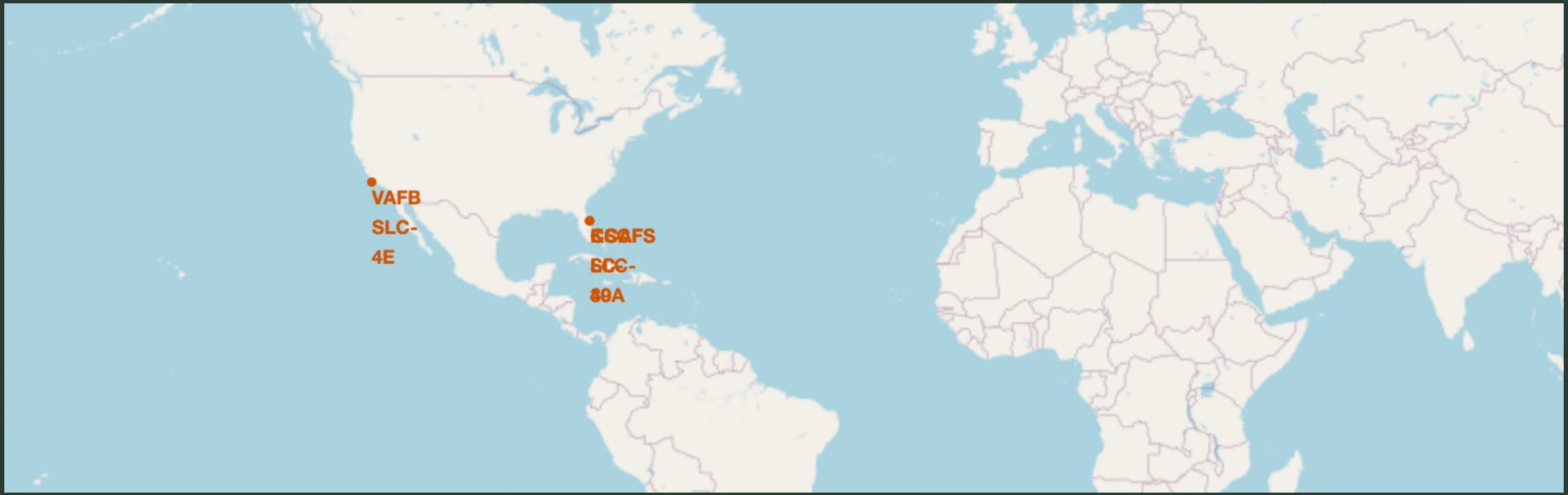
- (notebook)

Result – success year trend

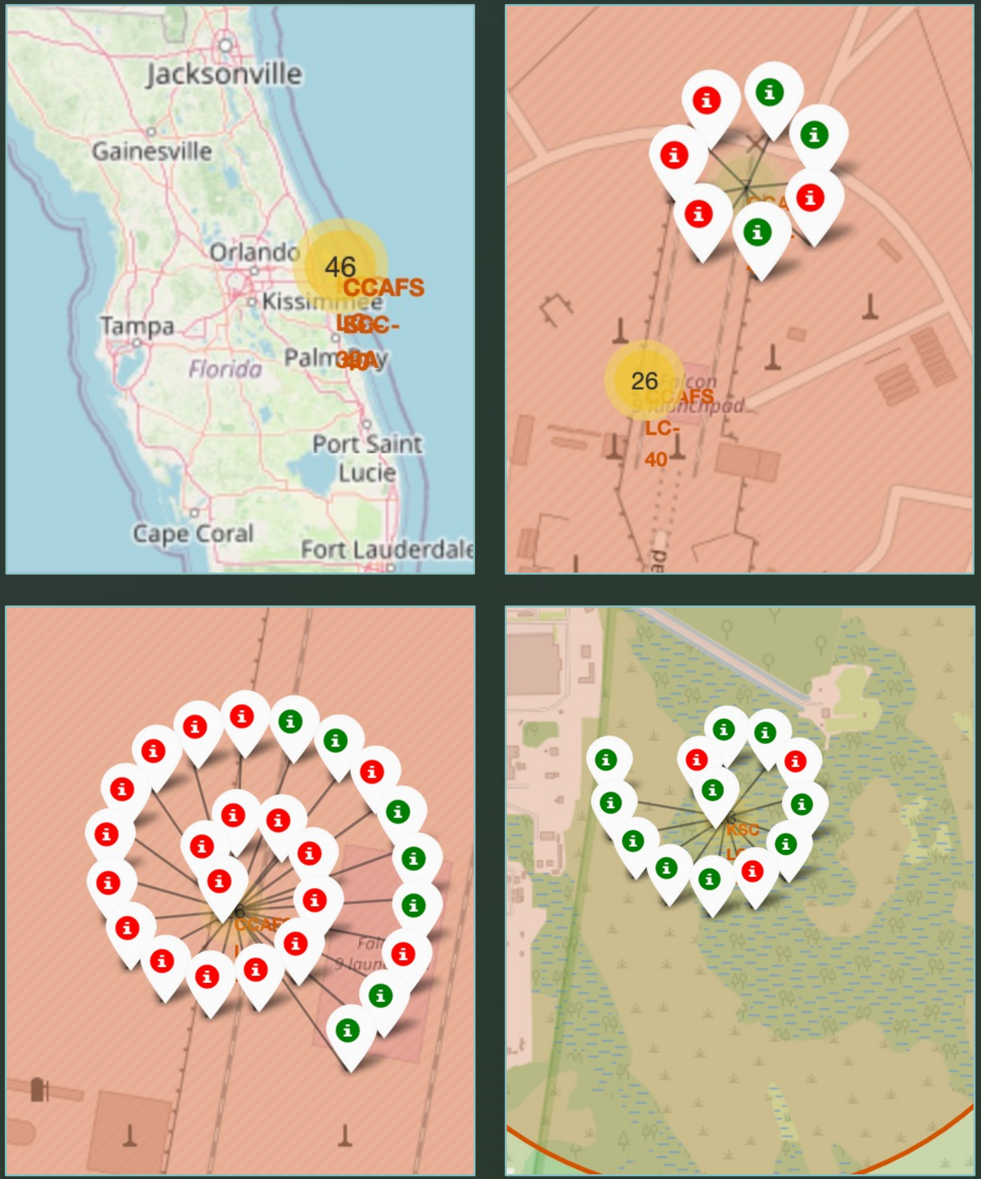
- The success rate is on an overall increasing trend before 2020 (notebook)



Results – Interactive visualization



- In the folium map, we have found that the launch sites are close to the equator



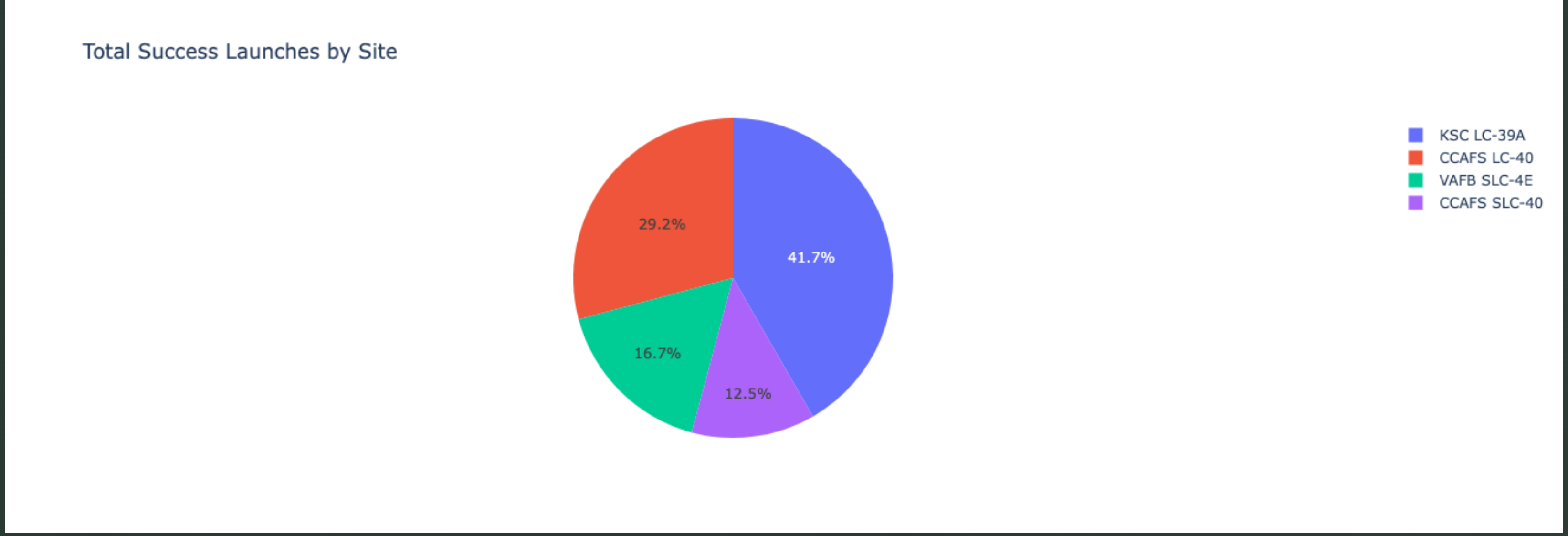
Results - Marker cluster

- By adding marker cluster, we can easily check the success (green) and failure (red) at each site.
- Note: sometimes the code requires restart kernel to remove the error “object is not JSON serializable”

Results - distance

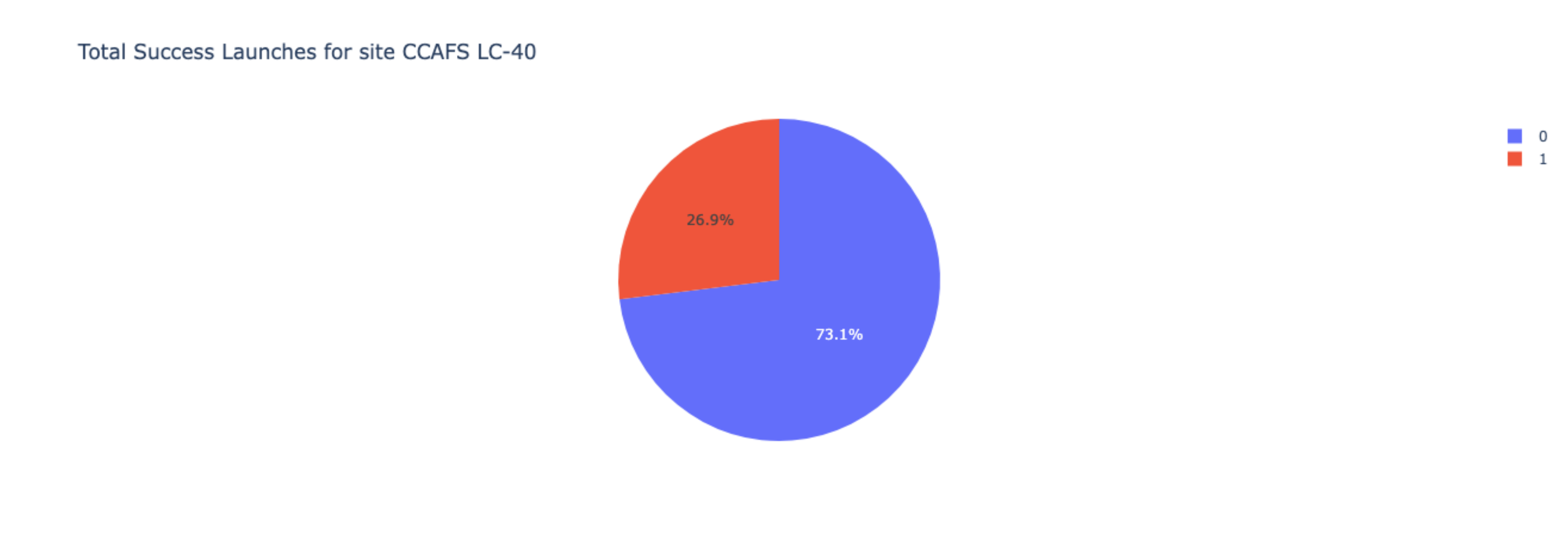
- We also used interactive map to find distances to the nearest coastline, highway, and city,
- We found that there are coastline (0.95km) and highway (0.66 km) close to the launch site, all within 1km, while the closest city, Cape Canaveral, is over 18km far away.

Results -dashboard



- We built an interactive dashboard to explore the percentage of launches at each site and success rate.

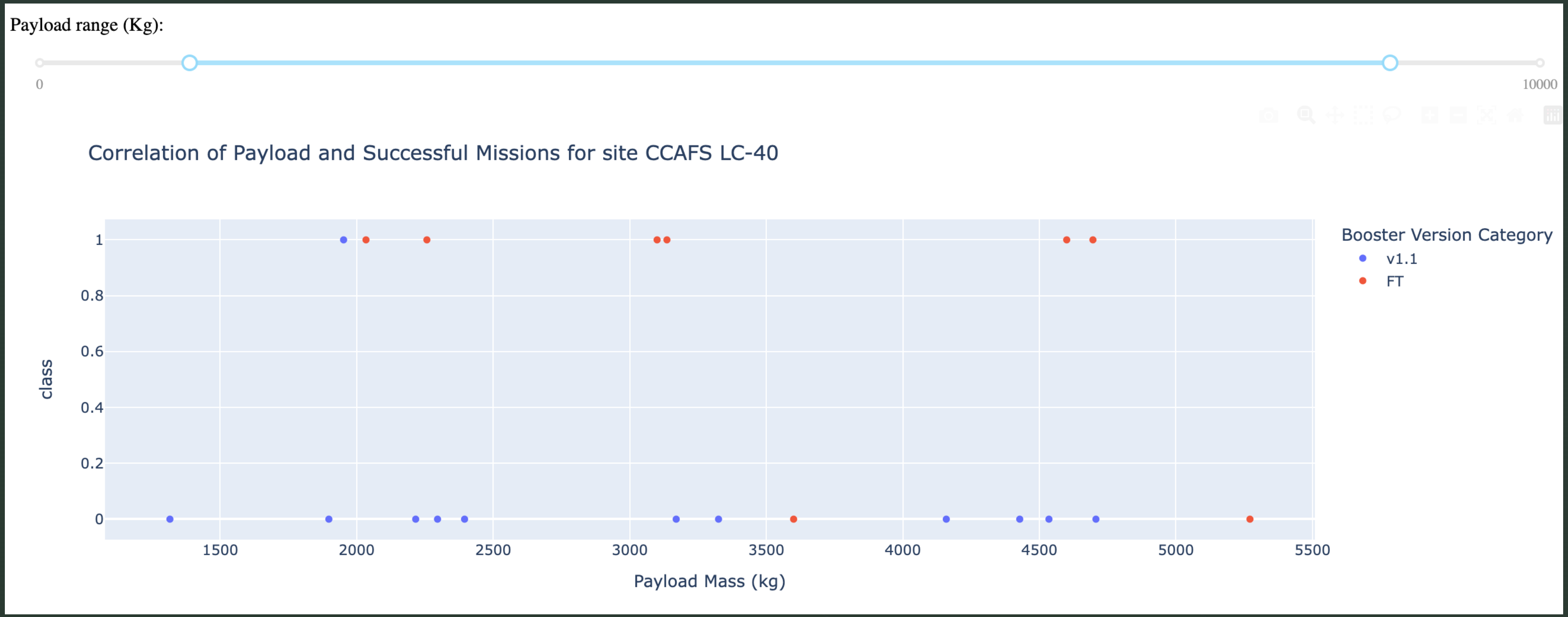
Results -dashboard



Total Success Launches for site CCAFS LC-40

- With drop down, we can check the success rate of every site

Results dashboard



Payload range (Kg):

- For each specific site, we can zoom in payload mass range to examine the success launches

Results – predictive analysis

- We tested several categorization models in the notebook to predict whether a launch is successful.
- From the EDA, we determining training labels by creating columns for classes, standardizing the data, and splitting data into training and test set
- We used the standard scaler to standardize the independent variables. And splitted the data with train_test_split

Logistic regression

- The best hyperparameters are: {'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs’}
- Training accuracy is 0.846
- Test accuracy is 0.833

SVM

- The best hyperparameters are: {'C': 1.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid’}
- Training accuracy : 0.848
- Test accuracy is 0.833

Decision Tree

- The best hyperparameters are: {'criterion': 'gini', 'max_depth': 16, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 10, 'splitter': 'random’}
- Training accuracy : 0.873
- Testing accuracy: 0.833

KNN

- The best hyperparameters are: {'algorithm': 'auto', 'n_neighbors': 10, 'p': 1}
- Training accuracy : 0.848
- Testing Accuracy: 0.833

Discussion

- The models all performs alike in the test set.
- In the training set, the decision tree has the highest accuracy. However, it would be too hasty to say the decision tree is overfit to the training set, because all the accuracy values are close.
- Therefore, in this case, all the predictive models are similar to each other. We cannot say which one is the best.

Conclusions

- We have identified several launch sites and orbits that has the highest success rate.
- The launch success has a positive trend starting from 2013.
- The models performs a like, with decision tree has better accuracy in training set.