IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Kevin Juandi
08.08.2022

Outline

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

Executive Summary

- The following methodologies were used to analyze data:
  - Data collection from SpaceX API and web scraping
  - Exploratory Data Analysis (EDA), including data wrangling, data visualization and interactive visual analytics
  - Machine Learning Prediction
- Summary of all results:
  - It was possible to collect valuable data from public sources
  - EDA helps 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

- The objective is to evaluate the viability of the new company SpaceY to compete with SpaceX.

- Desired answers:
  - optimisation of the total cost for launches, by predicting successful landings of the first stage of rockets;
  - finding the right location for launches

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
    - The data was available from two sources:
        - SpaceX API (https://api.spacexdata.com/v4/rockets/)
        - Webscraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)
- Perform data wrangling
    - Collected data is processed by creating a landing outcome label based on outcome data after summarizing and analyzing features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash

Methodology

Executive Summary

- Perform predictive analysis using classification models
  - Data that was collected until this step were normalized, split into training and test data sets and evaluated by four different classification models, the accuracy of each model is then evaluated using different combinations of parameters.

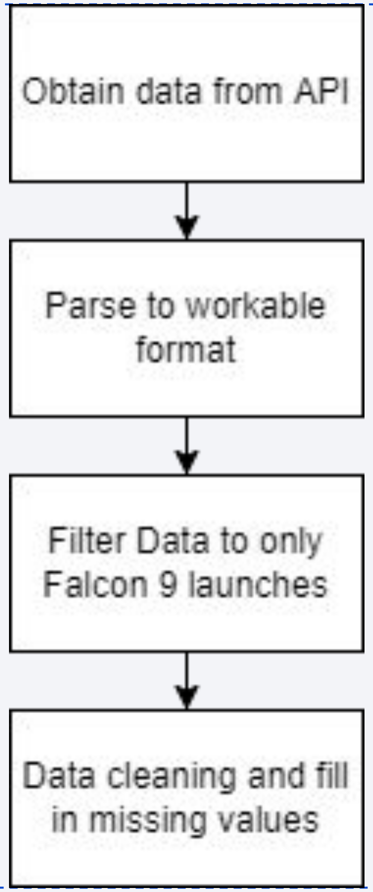
Data Collection

Data sets were collected from SpaceX API (https://api.spacexdata.com/v4/rockets/) and scraped from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)

Data Collection – SpaceX API

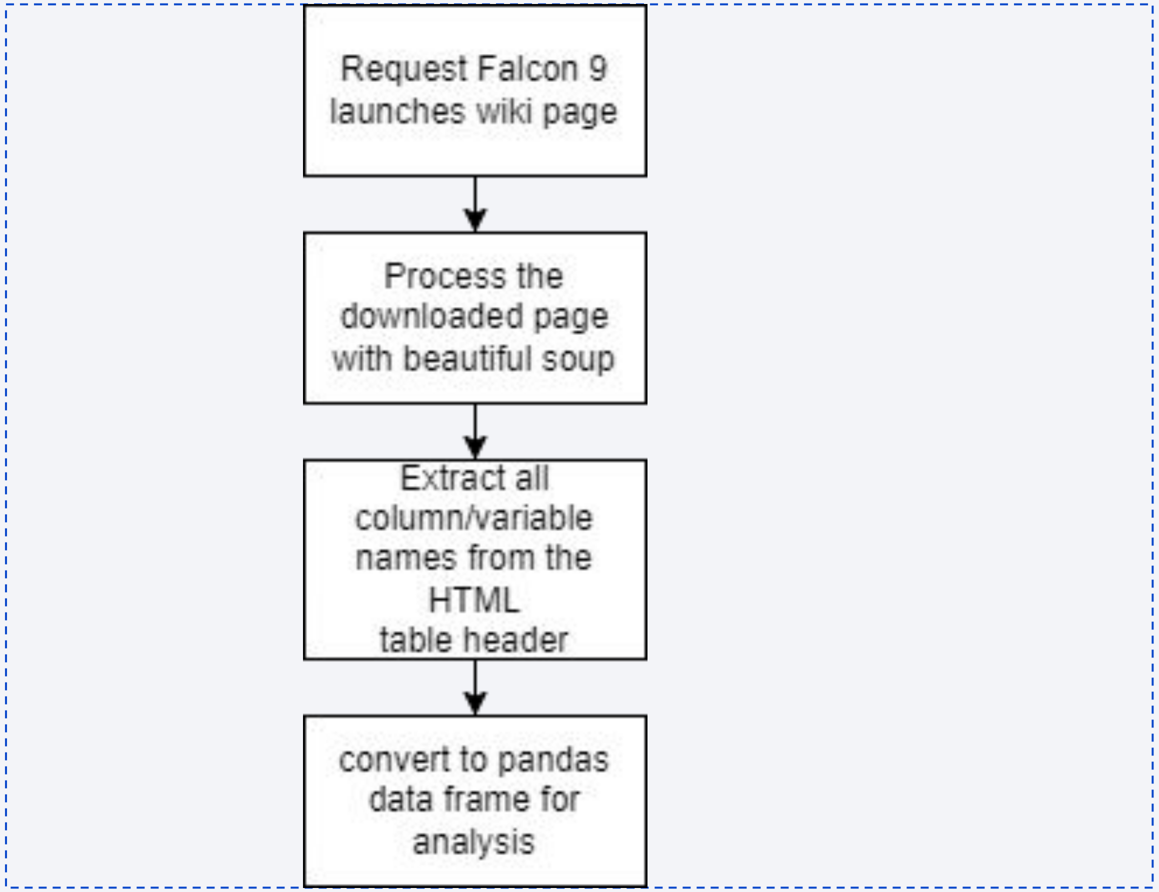
- SpaceX offers a public API from where data can be obtained and then used
- This API was used according to the flowchart beside and then data is persisted.

- [source code]



Data Collection - Scraping

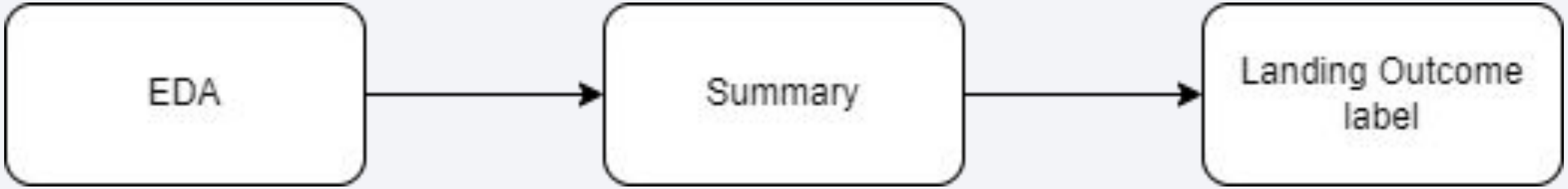
- Data from SpaceX launches can also be obtained from Wikipedia
- [source code]



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]

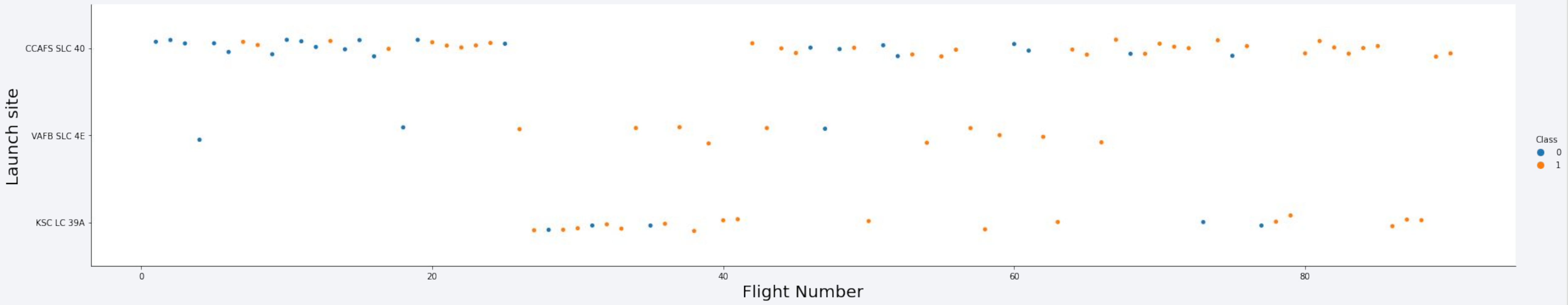


EDA with Data Visualization

- Scatter plots and bar plots were used to visualize the relationship between these pair of features:

  Payload Mass X Flight Number, Launch Site X Flight Number, Launch Site X Payload Mass, Orbit and Flight Number, Payload and Orbit

[source code]



EDA with SQL

The following SQL queries were performed:

- Names of the unique launch sites in the space mission
- Top 5 launch sites whose name begin with the string 'CCA’
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- Date when the first successful landing outcome in ground pad was achieved
- Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg
- Total number of successful and failure mission outcomes
- Names of the booster versions which have carried the maximum payload mass
- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.

source code

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

Build a Dashboard with Plotly Dash

The following graphs and plots were used to visualize data

- Percentage of launches by site
- Payload range

This combination allows to quickly analyze the relation between payloads and launch sites, helping to identify where is best place to launch according to payloads.

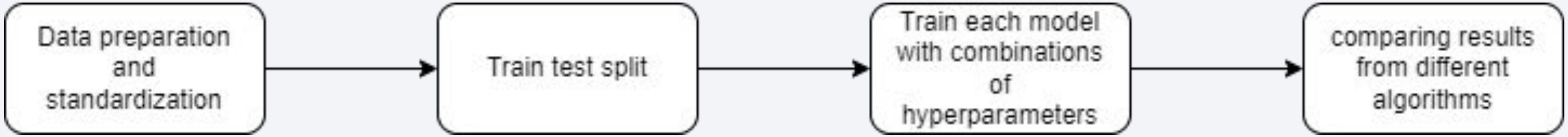
source code

Predictive Analysis (Classification)

Four classification models were used and compared:

- logistic regression
- support vector machine
- decision tree
- k-nearest neighbors

[source code]

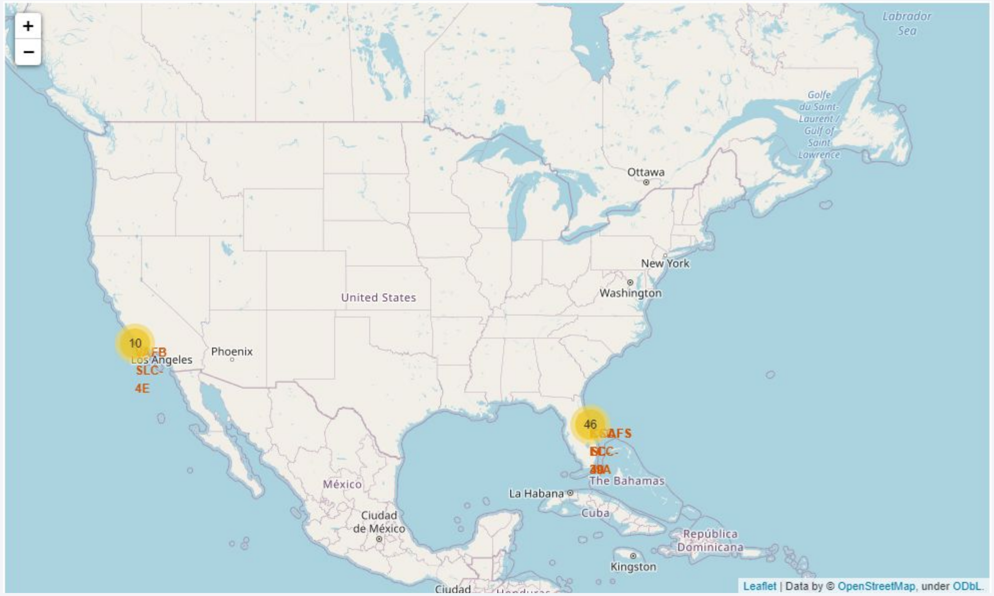


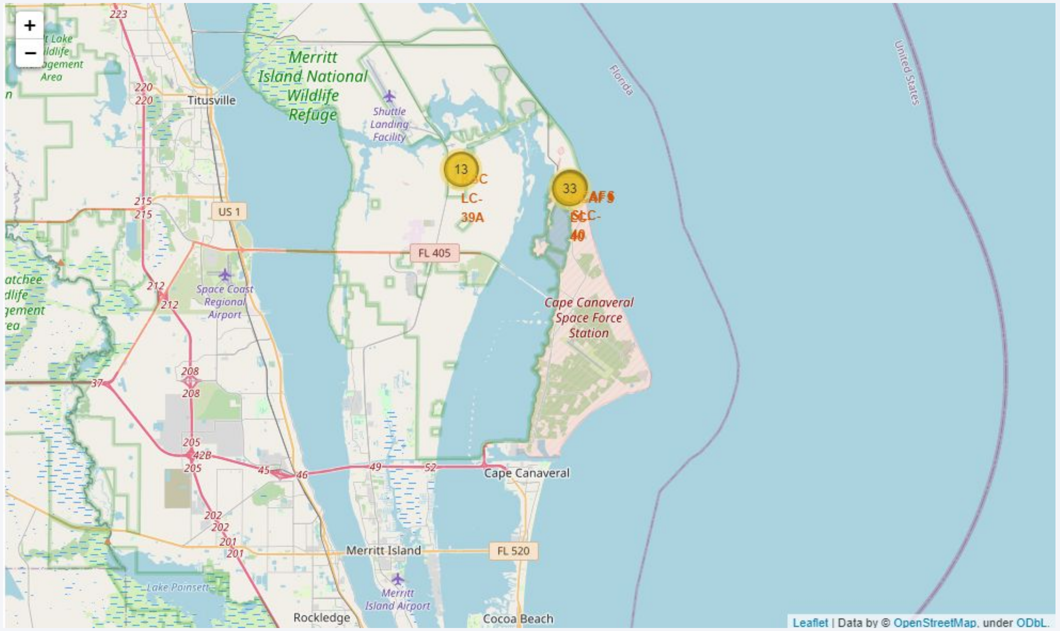
Results

- Exploratory data analysis results:
  - SpaceX uses 4 different launch sites
  - The first launches were done by SpaceX itself and NASA
  - The average payload of F9 v1.1 booster is 2,928 kg
  - The first success landing outcome happened in 2015, five year after the first launch
  - Many Falcon-9 booster versions were successful at landing in drone ships having payload above the average
  - Almost 100% of mission outcomes were successful
  - Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015
  - The number of landing outcomes is progressively better as years passed.

Results

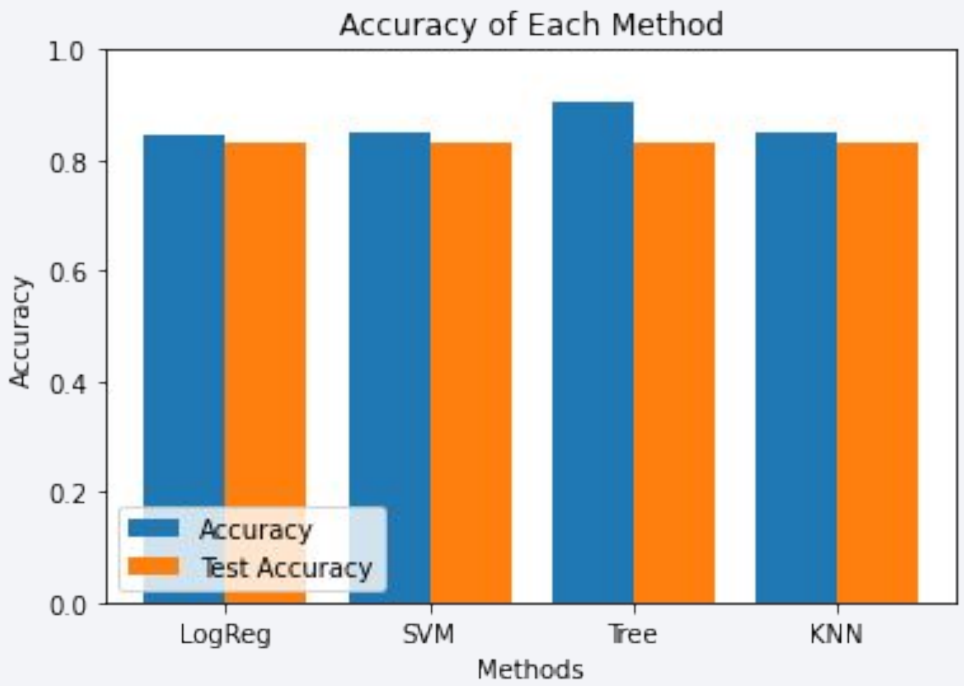
- Using interactive analytics, it is possible to identify that launch sites used are in coastal area, near the sea and have a good logistic infrastructure around.
- Most launches happens at east coast launch sites.





Results

- Decision Tree Classifier is our best model to predict successful landings, having accuracy over 87% and accuracy for test data over 94%.

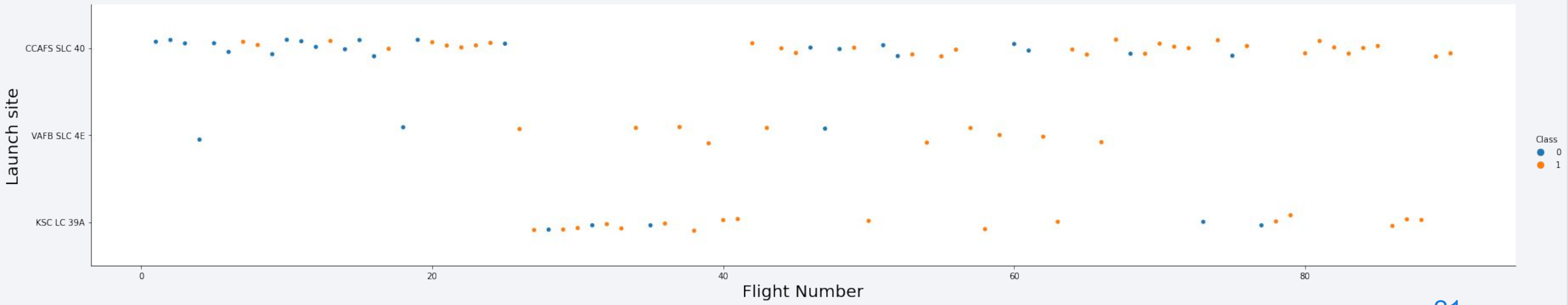


Section 2

Insights drawn from EDA

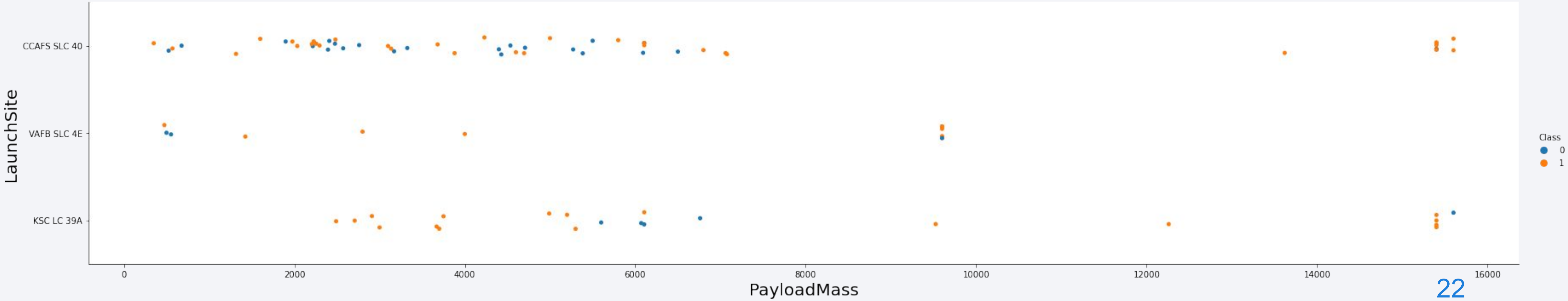
Flight Number vs. Launch Site

- According to the plot, it’s possible to verify that the best launch site is CCAF5 SLC 40, where most of recent launches were successful
- Followed by VAFB SLC 4E and third place KSC LC 39A
- It’s also visible that the general success rate improved over time.



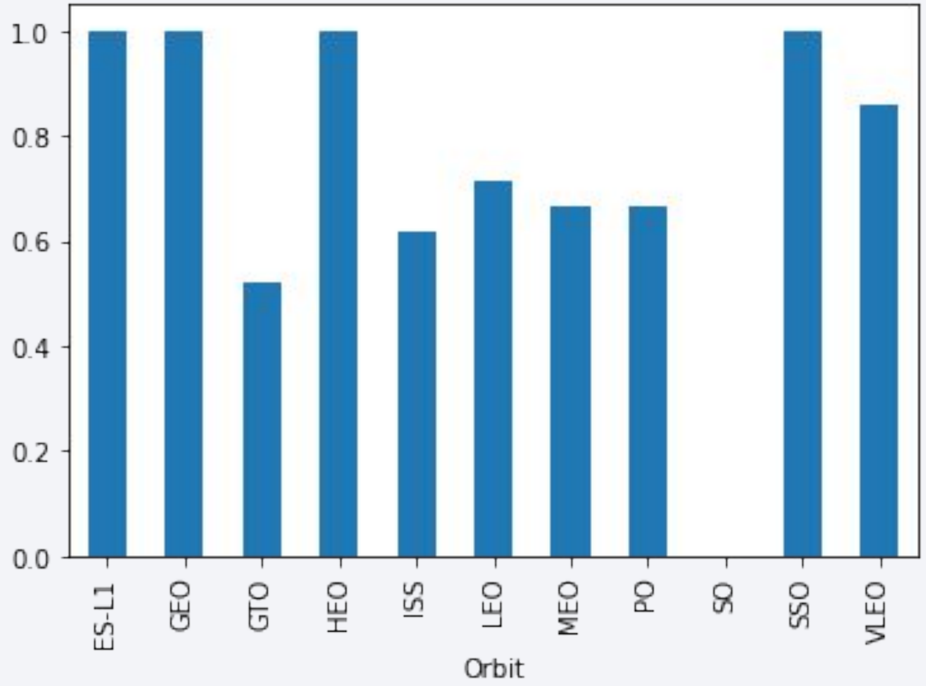
Payload vs. Launch Site

- Payloads over 9,000kg (about the weight of a school bus) have excellent success rate
- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.



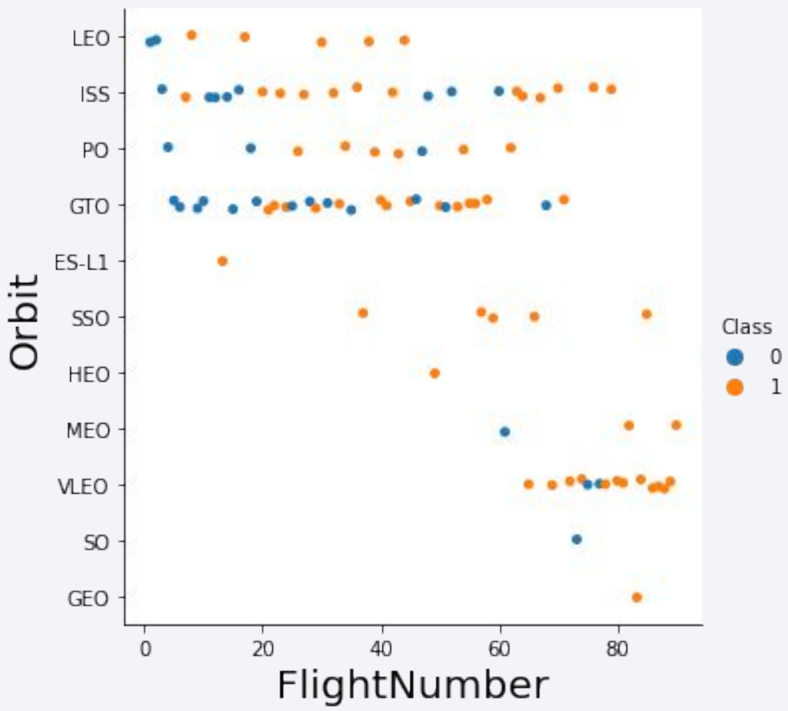
Success Rate vs. Orbit Type

- We have highest success rate for the orbits:
  - ES-L1
  - GEO
  - HEO
  - SSO
- Followed by:
  - VLEO (above 80%)
  - LFO (above 70%)



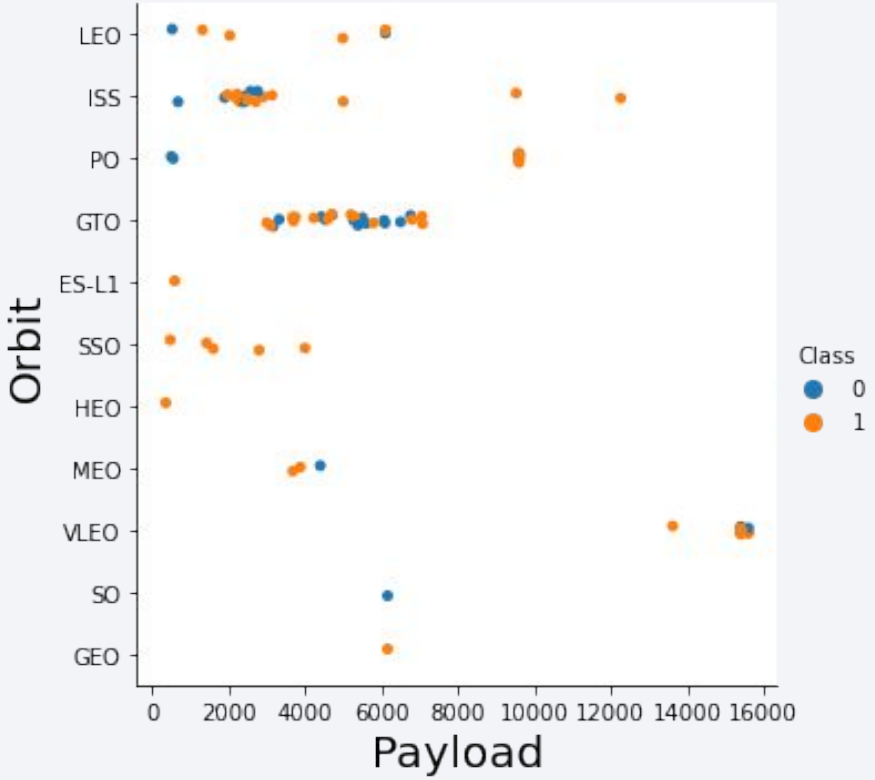
Flight Number vs. Orbit Type

- Success rate improved over time for all orbits
- Interestingly, VLEO orbit recently increase in frequency.



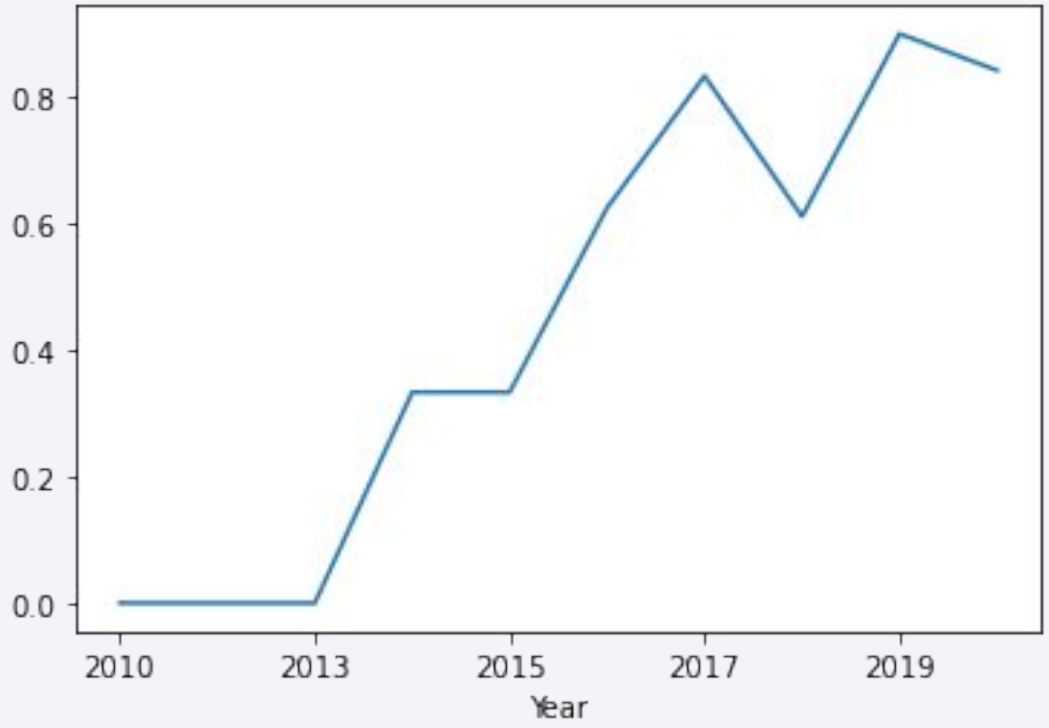
Payload vs. Orbit Type

- There doesn't seem to be any 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

- Success rate is increasing since 2013 with a plunge in 2018
- The first three years is likely the experimental stage of trial and error



All Launch Site Names

- We have 4 launch sites which we obtain by querying unique value of "launch_site" from the data sets

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites 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

- Here we can see five samples of Cape Canaveral launches.

Total Payload Mass

- Total payload carried by boosters from NASA

sum(PAYLOAD_MASS__KG_)
45596

- Total payload calculated above is obtained by summing all payloads whose codes contain ‘NASA CRS’

Average Payload Mass by F9 v1.1

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

avg(PAYLOAD_MASS__KG_)
2928.4

- Filtering data by the booster version above and calculating the average payload mass we obtained the value of 2,928 kg.

First Successful Ground Landing Date

- First successful landing outcome on ground pad:

min(DATE)
01-05-2017

- By filtering data by successful landing outcome on ground pad and getting the minimum value for date it’s possible to identify the first instance, that happened on 12/22/2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

- Boosters which have successfully landed on drone ship and had 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

- Querying distinct booster versions according to the filters above give us this result.

Total Number of Successful and Failure Mission Outcomes

- Number of successful and failure mission outcomes:

Mission_Outcome	QTY
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

- Boosters 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

- These are the boosters which have carried the maximum payload mass registered in the dataset.

2015 Launch Records

- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
10-01-2015	09:47:00	F9 v1.1 B1012	CCAFS LC-40	SpaceX CRS-5	2395	LEO (ISS)	NASA (CRS)	Success	Failure (drone ship)
14-04-2015	20:10:00	F9 v1.1 B1015	CCAFS LC-40	SpaceX CRS-6	1898	LEO (ISS)	NASA (CRS)	Success	Failure (drone ship)

- The list above has the only two occurrences.

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

- The ranking for the time frame is:

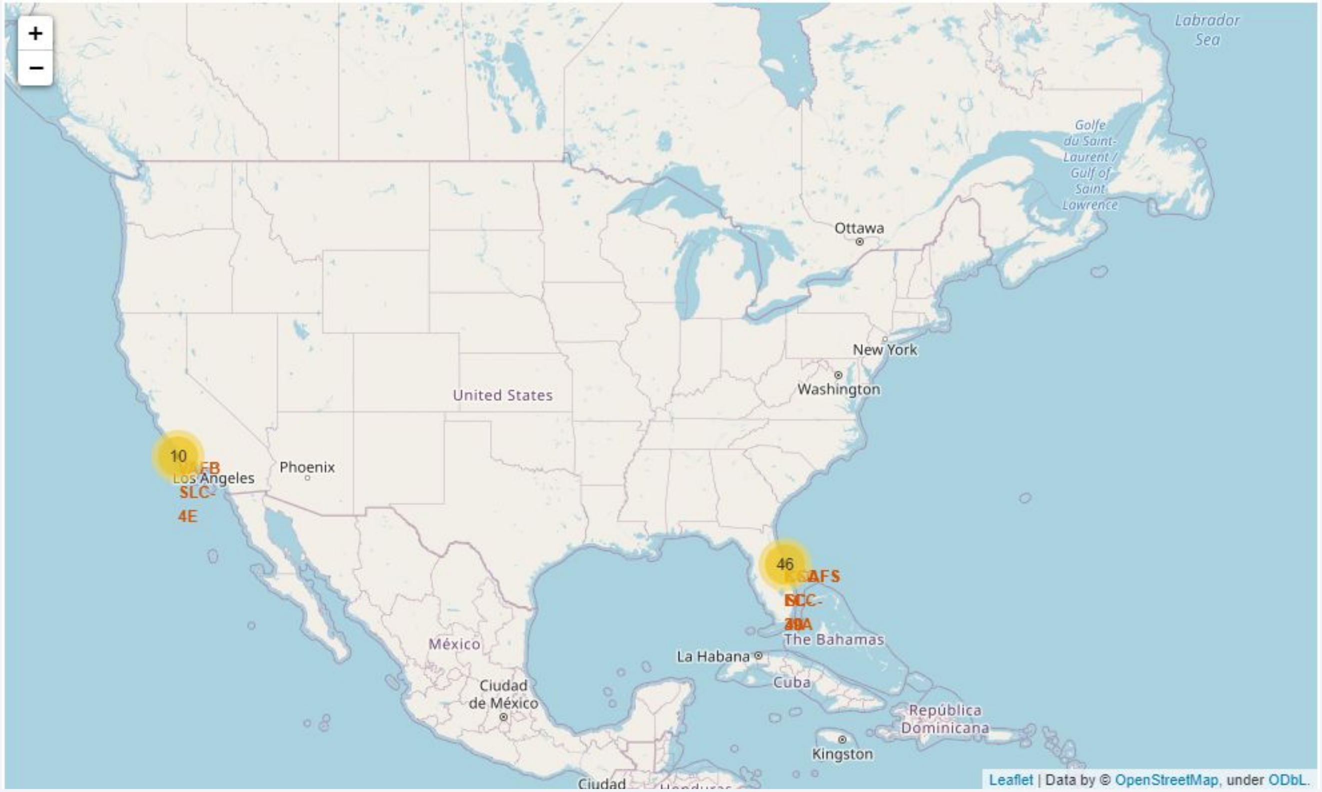
landing_outcome	qty
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

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

Section 3

Launch Sites Proximities Analysis

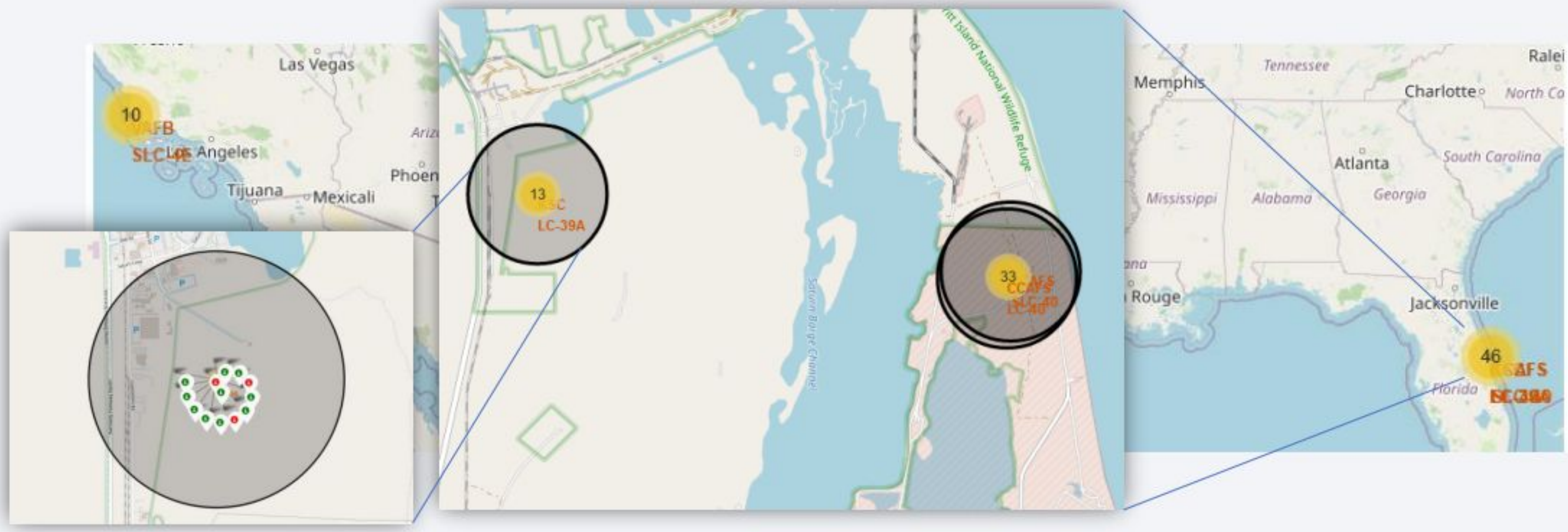
All launch site



Launch sites are near sea, probably for safety and close to roads and railroads

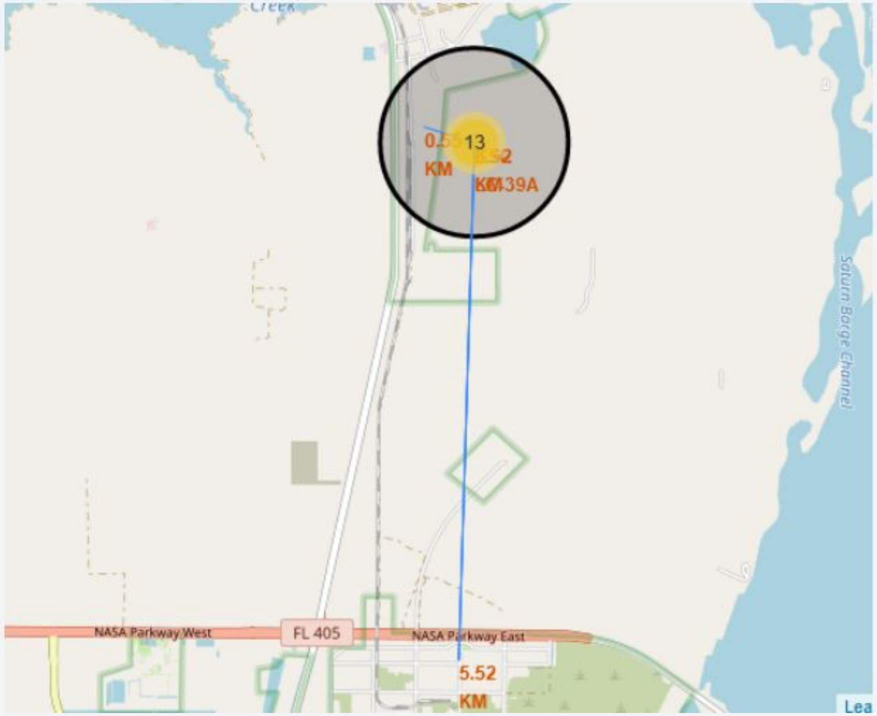
Launch Outcomes by Site

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



- Green markers indicate success and red indicate failure.

Logistics and Safety

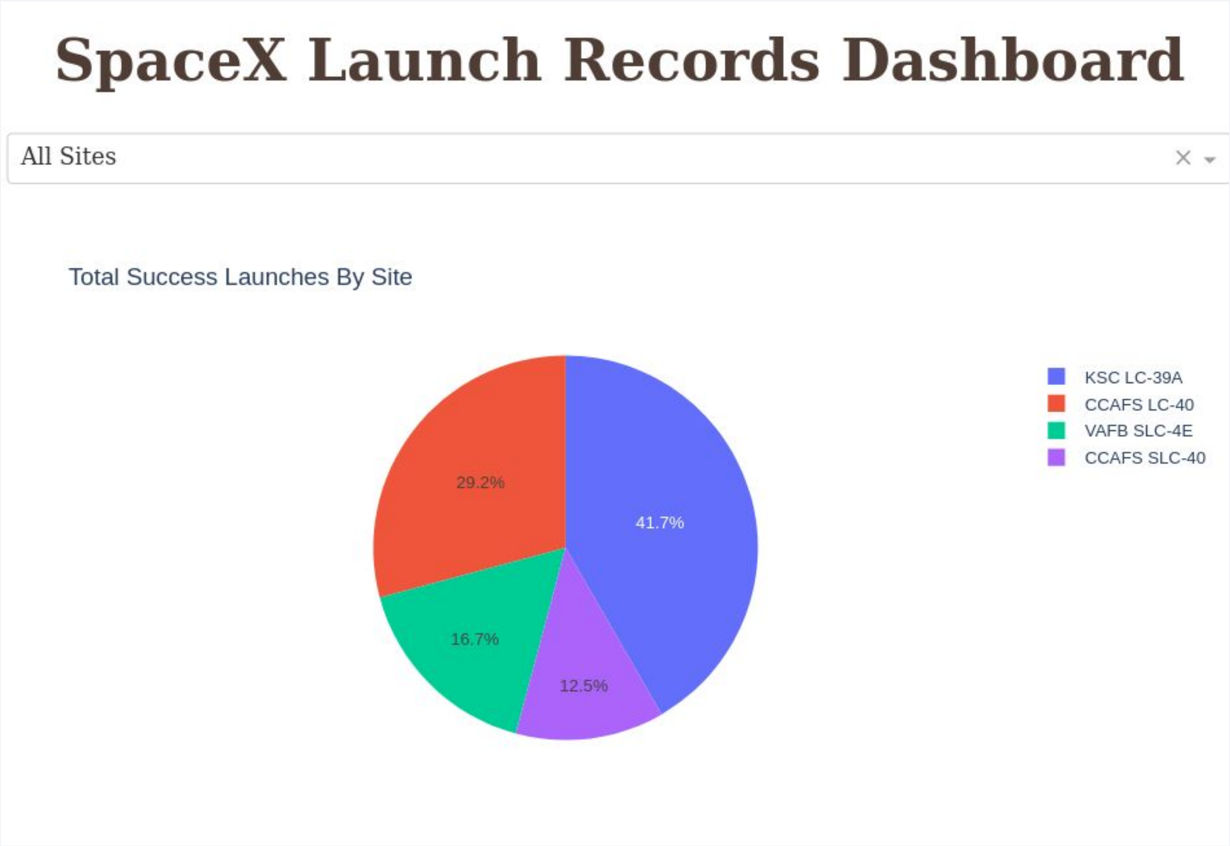


- Launch site KSC LC-39A has good logistics, near railroad and road and relatively far from inhabited areas.

Section 4

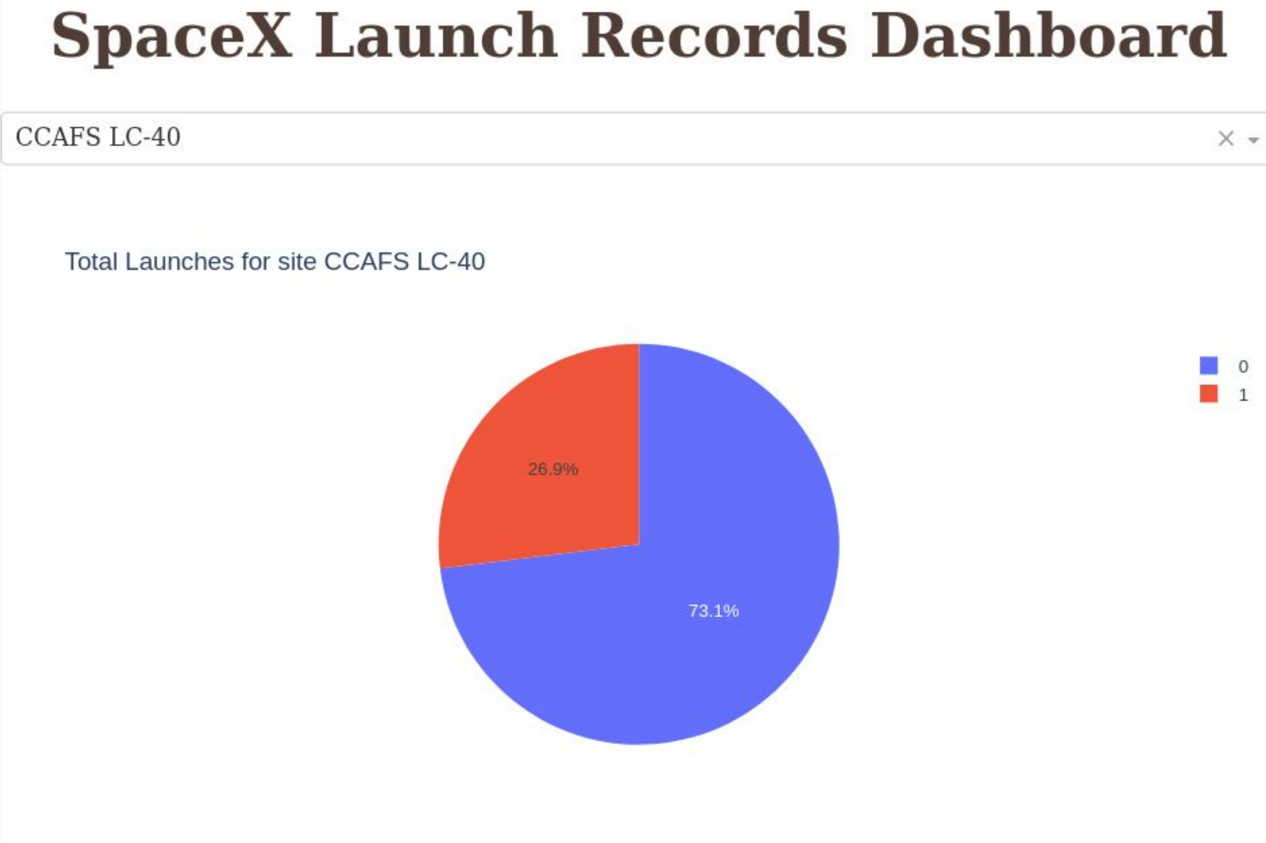
Build a Dashboard with Plotly Dash

Successful Launches by Site



Launch site seems to be important factor for success

Launch Success Ratio for KSC LC-39A



76.9% of launches are successful in this site.

Payload vs. Launch Outcome

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

<Dashboard Screenshot 3>

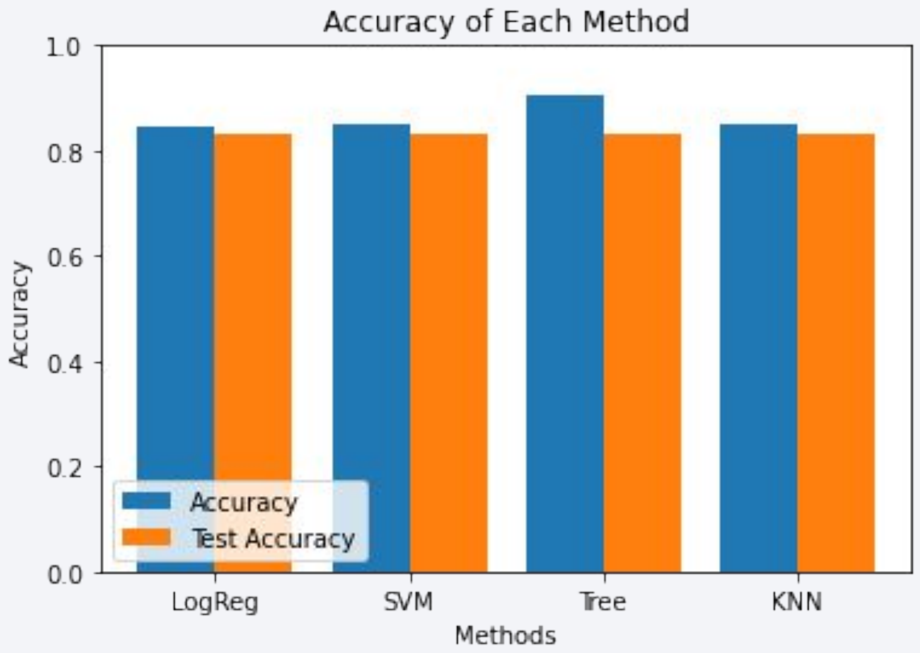
- There is not enough data to make conclusion for payload over 7,000 kg

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Four classification models were tested, and their accuracies are plotted
- The model with the highest classification accuracy is Decision Tree Classifier, which has accuracy over 87%.



Confusion Matrix of Decision Tree Classifier

- The confusion matrix shows the accuracy of the model from the large numbers of true positive and true negative compared to the false ones.

Conclusions

- We analysed different data sources and made conclusions along the process
- The best launch site is KSC LC-39A
- There isn’t enough data to make conclusion about high payload (above 7,000 kg)
- Although most of mission outcomes are successful, success rate seem to improve over time
- Decision Tree Classifier can be used to predict successful landings and increase profits.

Appendix

- Please check github for source code, the code might not be up to date with presentation slides
- Folium output isn’t saved on the notebook, screenshots are taken.

Thank you!