IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Tiago Flores
2021-10-28

Outline

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

Executive Summary

- Data analysis techniques included the following:
    - Utilizing the SpaceX API and web scraping to collect data.
    - Exploratory Data Analysis (EDA), included data wrangling, data visualization and interactive visual analytics.
    - Ends with Machine Learning Prediction.
- Summary of all results
    - It was possible to gather useful information from public sources.
    - EDA enabled the identification of which features best predict the success of launches.
    - Using all collected data, Machine Learning Prediction demonstrated the best model for predicting which characteristics are important to drive this opportunity.

Introduction

- The objective was to examine the viability of the new company Space Y in competing with SpaceX.
- Favorable responses:
  - The best way to estimate total launch costs is to predict successful first-stage rocket landings.
  - What is the best location for launches?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
    - Data from SpaceX was obtained from 2 sources:
        - SpaceX API (https://api.spacexdata.com/v4/rockets/)
        - Web Scraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)
- Perform data wrangling
    - After summarising and analysing features, collected data was enriched by creating a landing outcome label based on outcome data.
- Use visualisation and SQL to conduct exploratory data analysis (EDA).

Methodology

Executive Summary

- Use Folium and Plotly Dash to perform interactive visual analytics.
- Use classification models to perform predictive analysis.
    - Data collected up to this point were normalised, divided into training and test data sets, and evaluated by four different classification models, with the accuracy of each model evaluated using different parameter combinations.

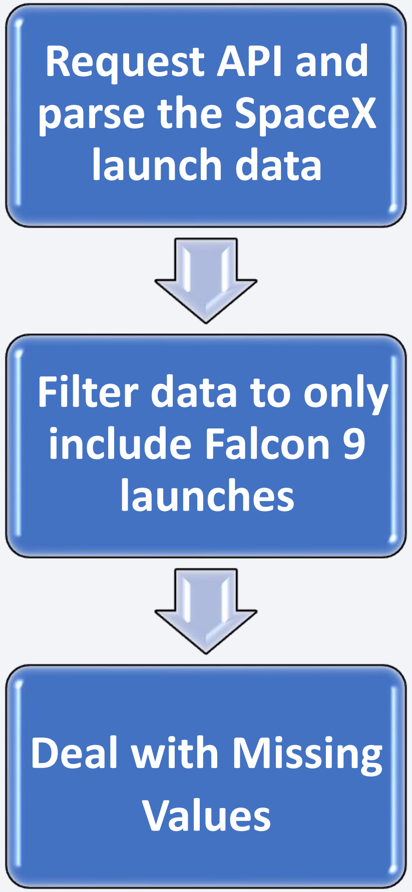
Data Collection

- Using web scraping techniques, data sets were collected from the Space X API (https://api.spacexdata.com/v4/rockets/) and Wikipedia (https://en.wikipedia.org/wiki/List of Falcon/ 9/ and Falcon Heavy launches).

Data Collection – SpaceX API

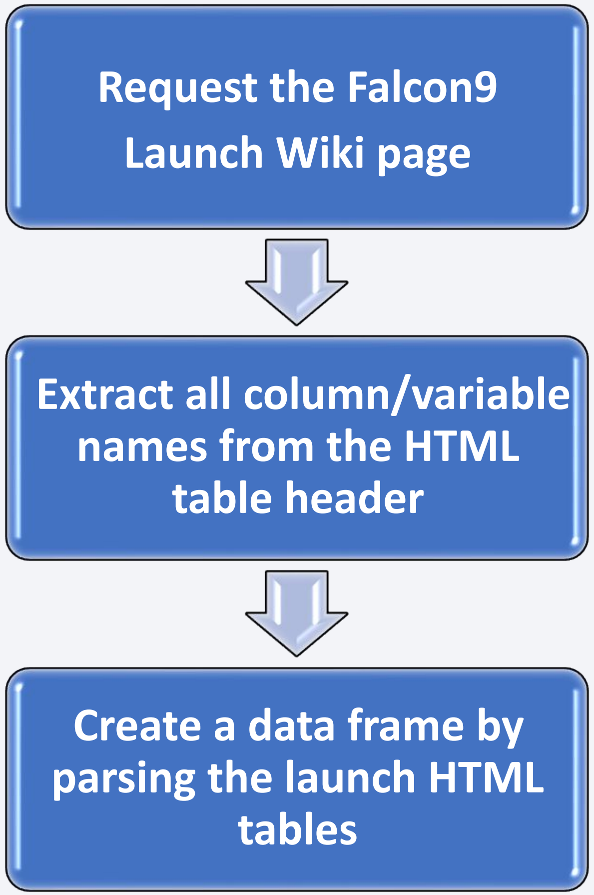
- SpaceX offers a public API from where data can be obtained and then used;
- This API was used according to the flowchart beside and then data is persisted.

- [Source code: https://github.com/Subin-Vidhu/Applied-Data-Science-Specialization/blob/main/Applied%20Data%20Science%20Capstone/Data%20Collection%20API.ipynb](https://github.com/Subin-Vidhu/Applied-Data-Science-Specialization/blob/main/Applied%20Data%20Science%20Capstone/Data%20Collection%20API.ipynb)



Data Collection - Scraping

- Data from SpaceX launches can also be obtained from Wikipedia;
- Data are downloaded from Wikipedia according to the flowchart and then persisted.
- Source code: https://github.com/Subin-Vidhu/Applied-Data-Science-Specialization/blob/main/Applied%20Data%20Science%20Capstone/Data%20Collection%20with%20Web%20Scraping.ipynb



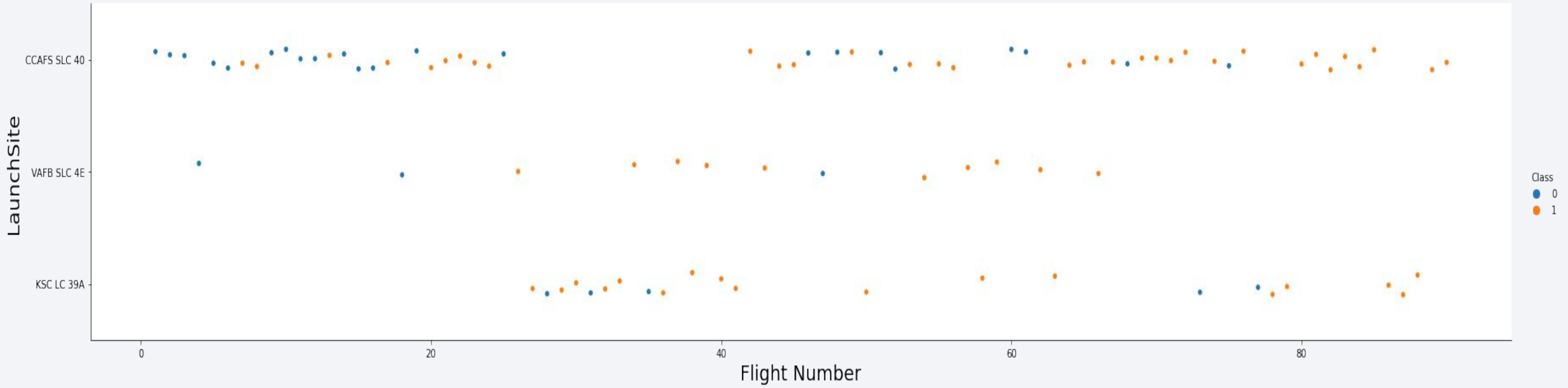
Data Wrangling

- 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: https://github.com/Subin-Vidhu/Applied-Data-Science-Specialization/blob/main/Applied%20Data%20Science%20Capstone/Data%20Wrangling%20EDA.ipynb

EDA with Data Visualization

- Scatter plots and bar plots were used to visualise the relationship between two features when exploring data:
  - Payload Mass X Flight Number, Launch Site X Flight Number, Payload Mass X Flight Number, Payload and Flight Number



- [Source code: https://github.com/Subin-Vidhu/Applied-Data-Science-Specialization/blob/main/Applied%20Data%20Science%20Capstone/EDA%20with%20Data%20Visualization.ipynb](https://github.com/Subin-Vidhu/Applied-Data-Science-Specialization/blob/main/Applied%20Data%20Science%20Capstone/EDA%20with%20Data%20Visualization.ipynb)

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; and
  - 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: https://github.com/Subin-Vidhu/Applied-Data-Science-Specialization/blob/main/Applied%20Data%20Science%20Capstone/EDA%20with%20SQL.ipynb

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; and
  - Lines are used to indicate distances between two coordinates.

- Source code:
- https://github.com/Subin-Vidhu/Applied-Data-Science-Specialization/blob/main/Applied%20Data%20Science%20Capstone/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

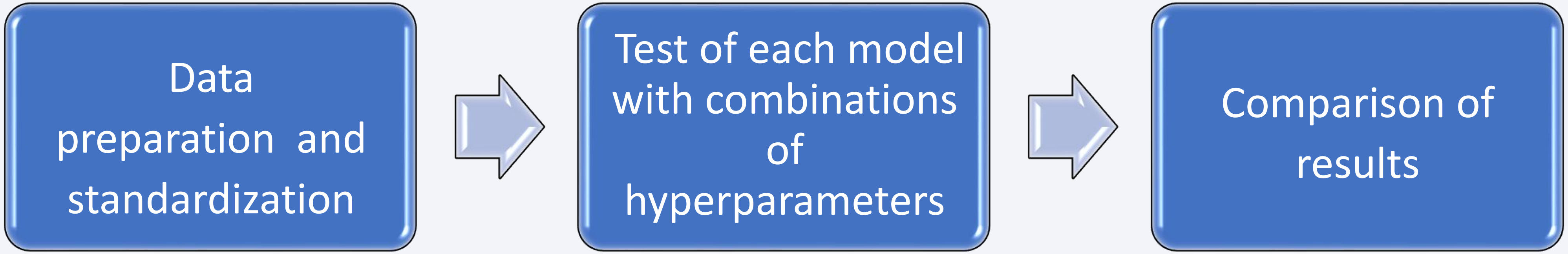
Build a Dashboard with Plotly Dash

- The following graphs and plots were used to visualize data
  - Percentage of launches by site
  - Payload range
- This combination allowed to quickly analyze the relation between payloads and launch sites, helping to identify where is best place to launch according to payloads.

- [Source code: https://github.com/Subin-Vidhu/Applied-Data-Science-Specialization/blob/main/Applied%20Data%20Science%20Capstone/spacex_dash_app.py](https://github.com/Subin-Vidhu/Applied-Data-Science-Specialization/blob/main/Applied%20Data%20Science%20Capstone/spacex_dash_app.py)

Predictive Analysis (Classification)

- The classification models logistic regression, support vector machine, decision tree, and k nearest neighbours were all compared.



- [Source code: https://github.com/Subin-Vidhu/Applied-Data-Science-Specialization/blob/main/Applied%20Data%20Science%20Capstone/Machine%20Learning%20Prediction%20lab.ipynb](https://github.com/Subin-Vidhu/Applied-Data-Science-Specialization/blob/main/Applied%20Data%20Science%20Capstone/Machine%20Learning%20Prediction%20lab.ipynb)

Results

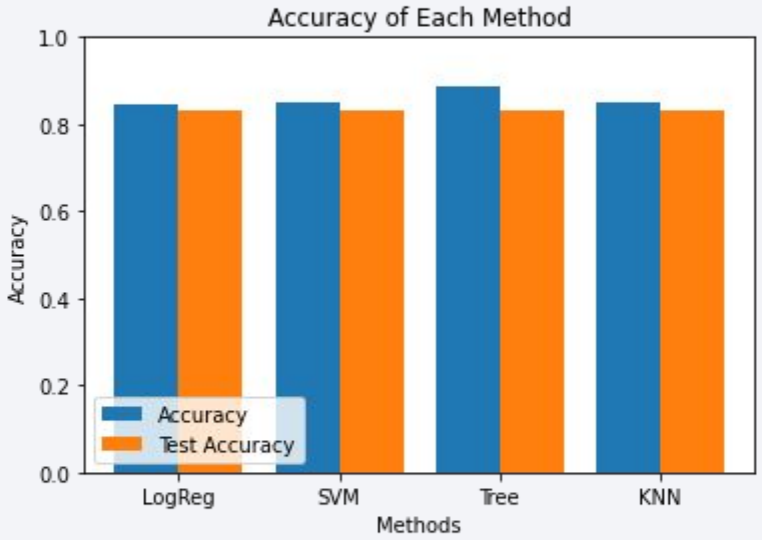
- Exploratory data analysis results:
  - SpaceX uses 4 different launch sites;
  - The first launches were done to Space X itself and NASA;
  - The average payload of F9 v1.1 booster is 2,928 kg;
  - The first success landing outcome happened in 2015 fiver year after the first launch;
  - Many Falcon 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 became as better as years passed.

Results

- Using interactive analytics, it was possible to determine that launch sites used to be in safe locations, such as near the sea, and had a good logistic infrastructure surrounding them.
- The majority of launches take place on the east coast.

Results

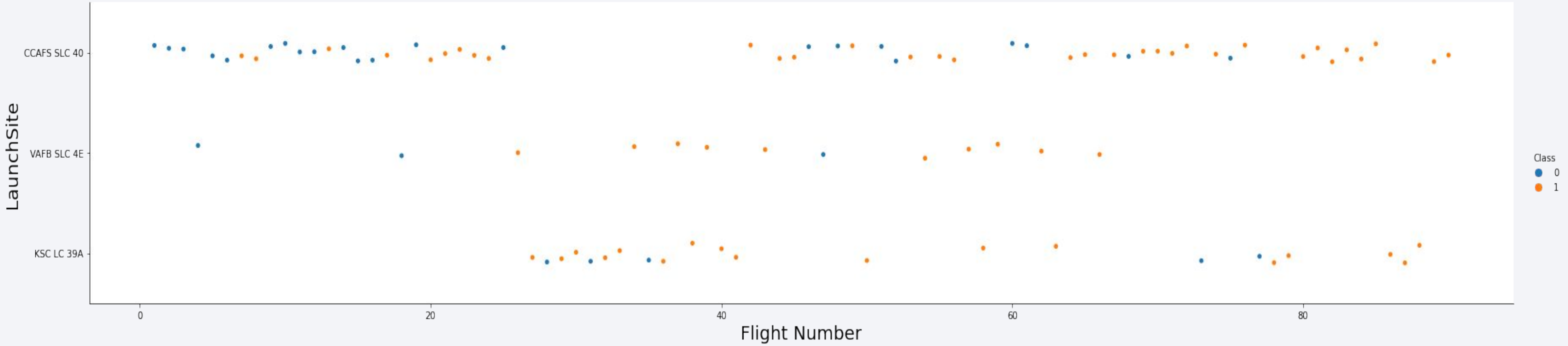
- Predictive Analysis revealed that the Decision Tree Classifier is the best model for predicting successful landings, with an accuracy of over 87 percent and an accuracy of over 83 percent for test data.



Section 2

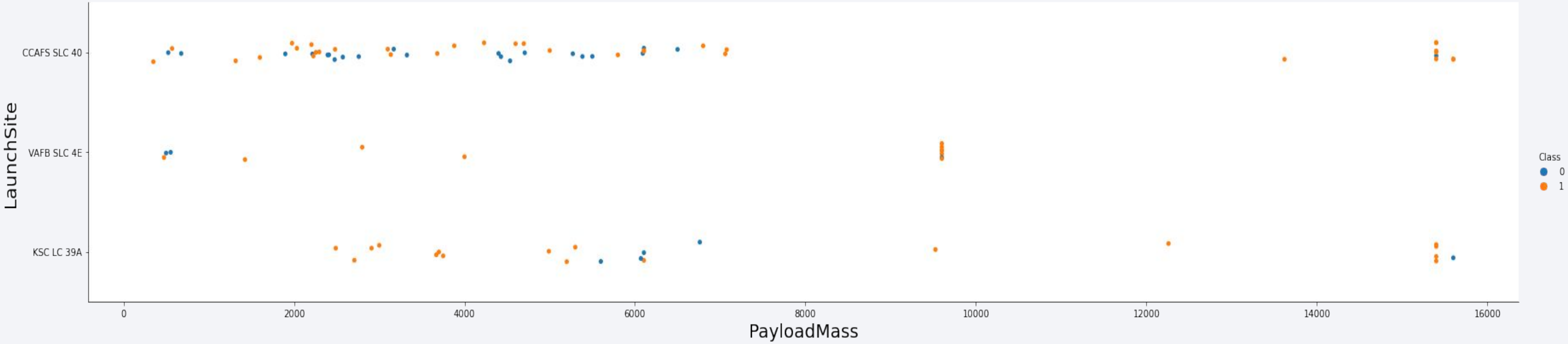
Insights drawn from EDA

Flight Number vs. Launch Site



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

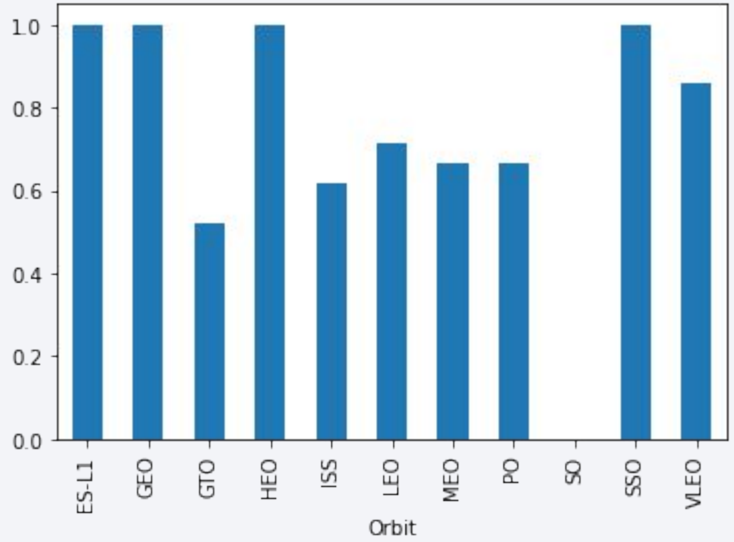
Payload vs. Launch Site



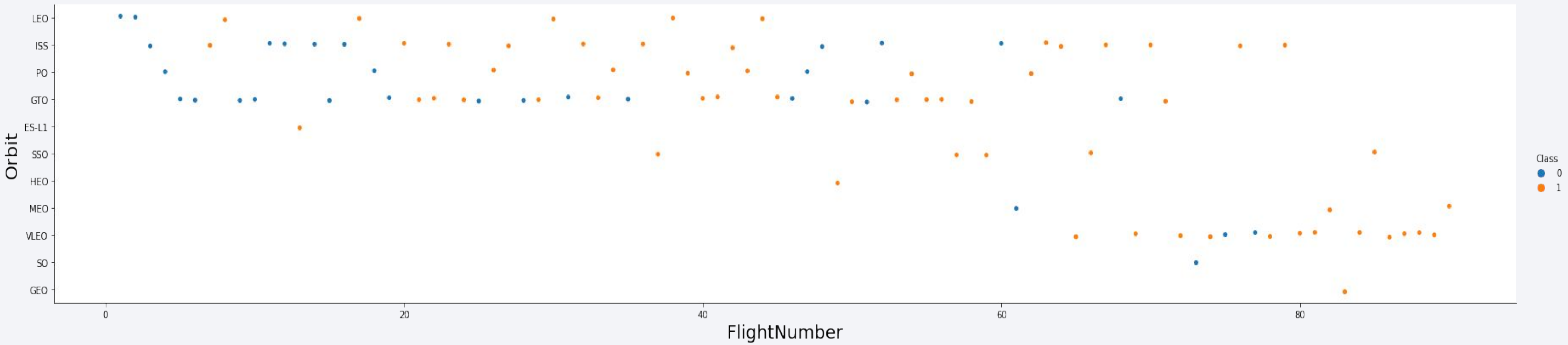
- Payloads over 9,000kg (about the weight of a school bus) have excellent success rate;
- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.

Success Rate vs. Orbit Type

- The biggest success rates happens to orbits:
  - ES-L1;
  - GEO;
  - HEO; and
  - SSO.
- Followed by:
  - VLEO (above 80%); and
  - LFO (above 70%).

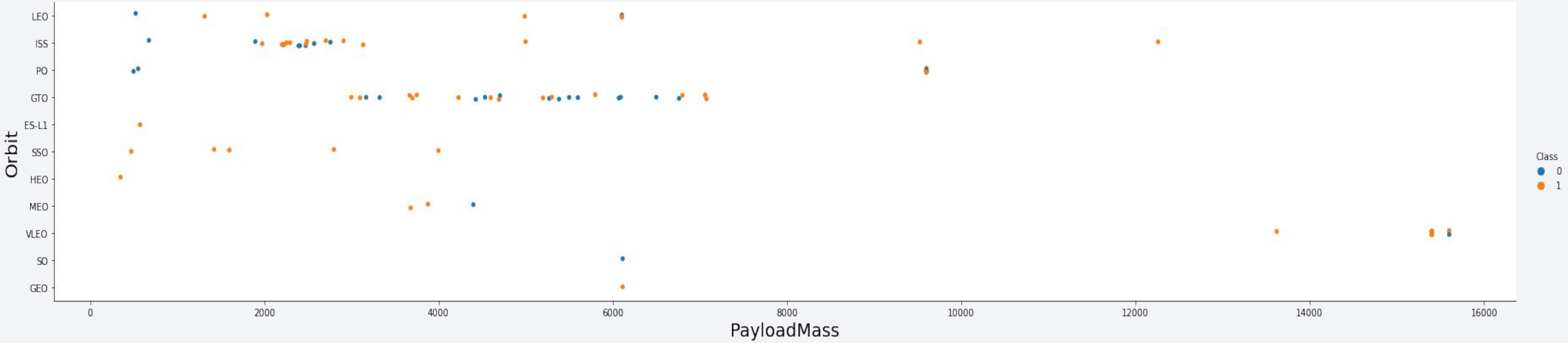


Flight Number vs. Orbit Type



- Apparently, success rate improved over time to all orbits;
- VLEO orbit seems a new business opportunity, due to recent increase of its frequency.

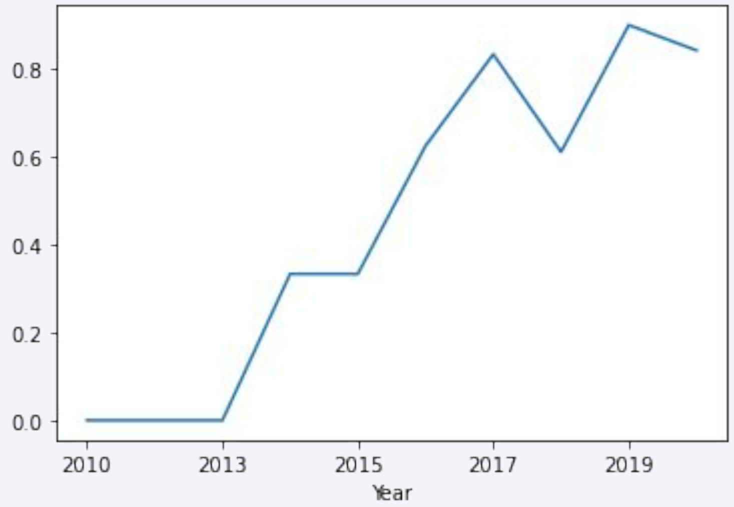
Payload vs. Orbit Type



- Apparently, there is no 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 started increasing in 2013 and kept until 2020;
- It seems that the first three years were a period of adjusts and improvement of technology.



All Launch Site Names

- According to data, there are four launch sites:

Launch Site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

- They are obtained by selecting unique occurrences of “launch_site” values from the dataset.

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
2010-06-04	18:45:00	F9 v1.0 B0003	**CCA**FS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	**CCA**FS 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	**CCA**FS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	**CCA**FS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	**CCA**FS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attemp

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

Total Payload Mass

- Total payload carried by boosters from NASA:

Total Payload (kg)
111.268

Total payload was calculated above by adding all payloads with the code 'CRS,' which corresponds to NASA.

Average Payload Mass by F9 v1.1

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

Avg Payload (kg)
2.928

We obtained the value of 2,928 kg after filtering data by the booster version mentioned above and calculating the average payload mass.

First Successful Ground Landing Date

- First successful landing outcome on ground pad:

Min Date
2015-12-22

- By filtering data by successful landing outcome on ground pad and obtaining the minimum value for date, the first occurrence, which occurred on 12/22/2015, can be identified.

Successful Drone Ship Landing with Payload between 4000 and 6000

- Boosters with a payload mass greater than 4000 but less than 6000 that successfully landed on a drone ship

Booster Version
F9 FT B1021.2
F9 FT B1031.2
F9 FT B1022
F9 FT B1026

- These four are the results of using the filters listed above to select distinct booster versions.

Total Number of Successful and Failure Mission Outcomes

- Number of successful and failure mission outcomes:

Mission Outcome	Occurrences
Success	99
Success (payload status unclear)	1
Failure (in flight)	1

- We arrived at the summary above by grouping mission outcomes and counting records for each group.

Boosters Carried Maximum Payload

- Boosters which have carried the maximum payload mass

Booster Version (...)
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3

Booster Version
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

- These boosters have carried the largest payload mass that has been recorded in the dataset.

2015 Launch Records

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

Booster Version	Launch Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

- The list above has the only two occurrences.

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

- Ranking of all landing outcomes between the date 2010-06-04 and 2017-03-20:

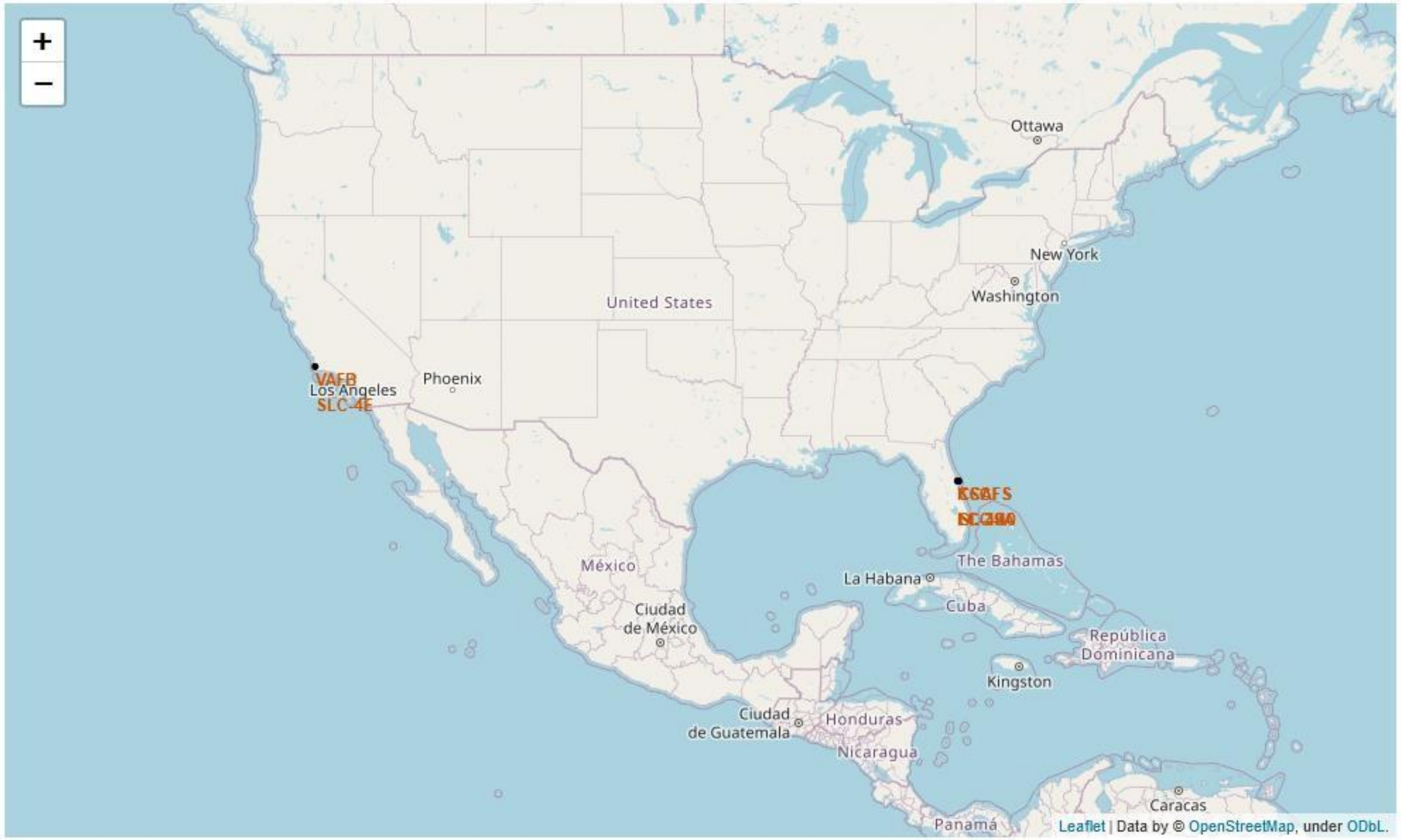
Landing Outcome	Occurrences
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

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

Section 4

Launch Sites Proximities Analysis

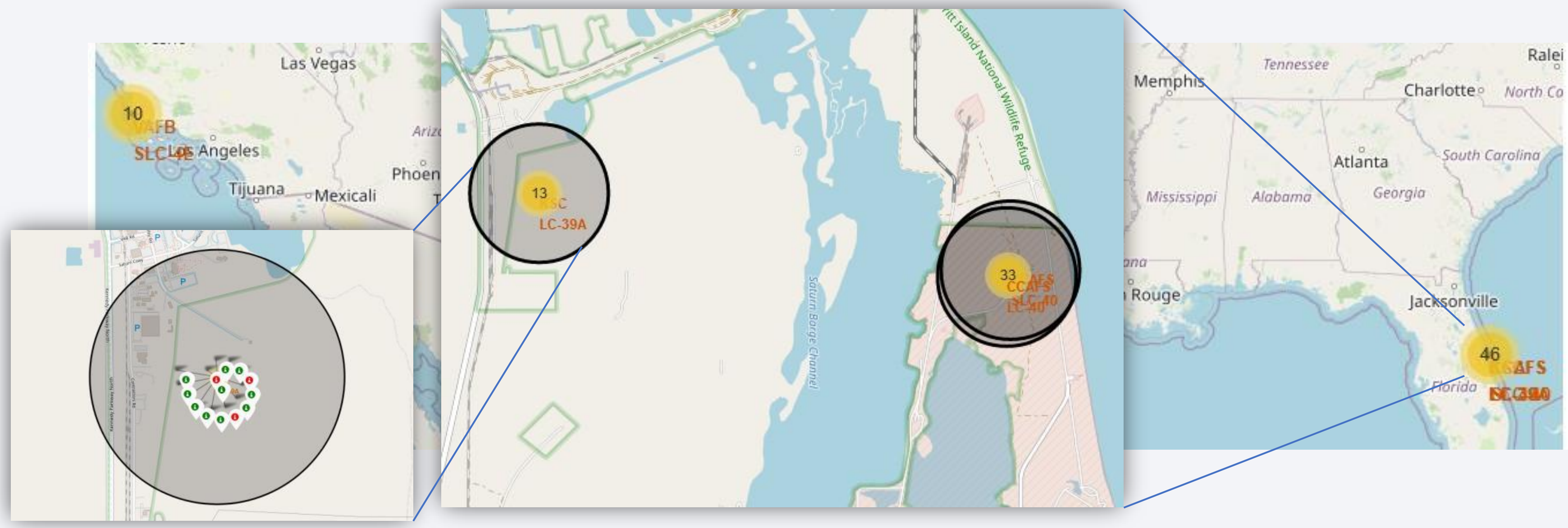
All launch sites



- Launch sites are near sea, probably by safety, but not too far from roads and railroads.

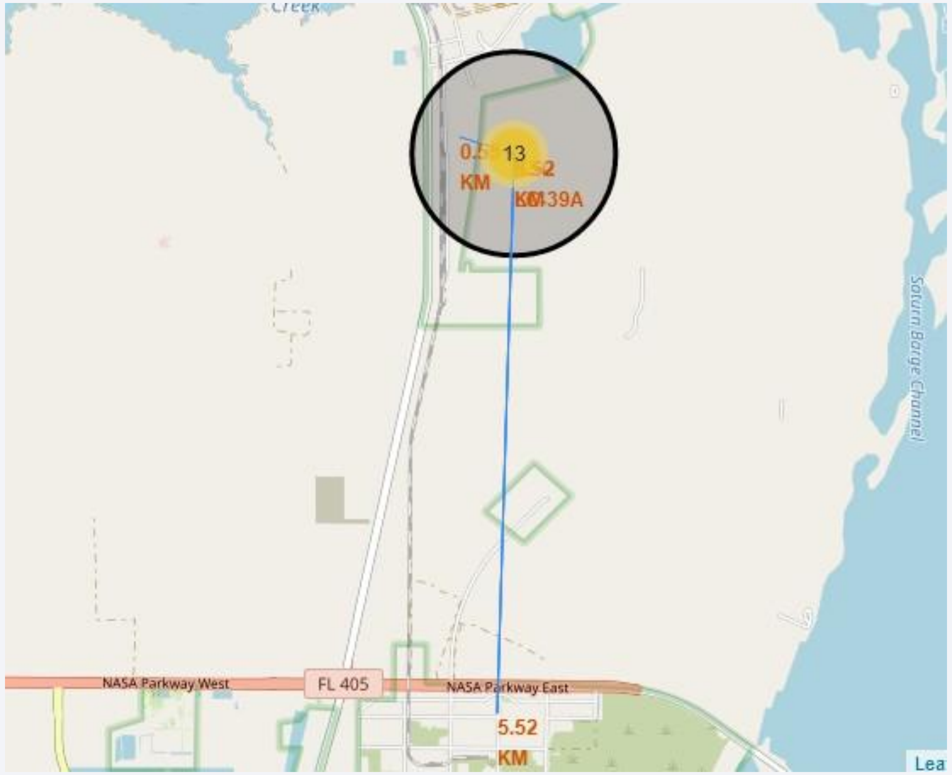
Launch Outcomes by Site

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



- Green markers indicate successful and red ones indicate failure.

Logistics and Safety

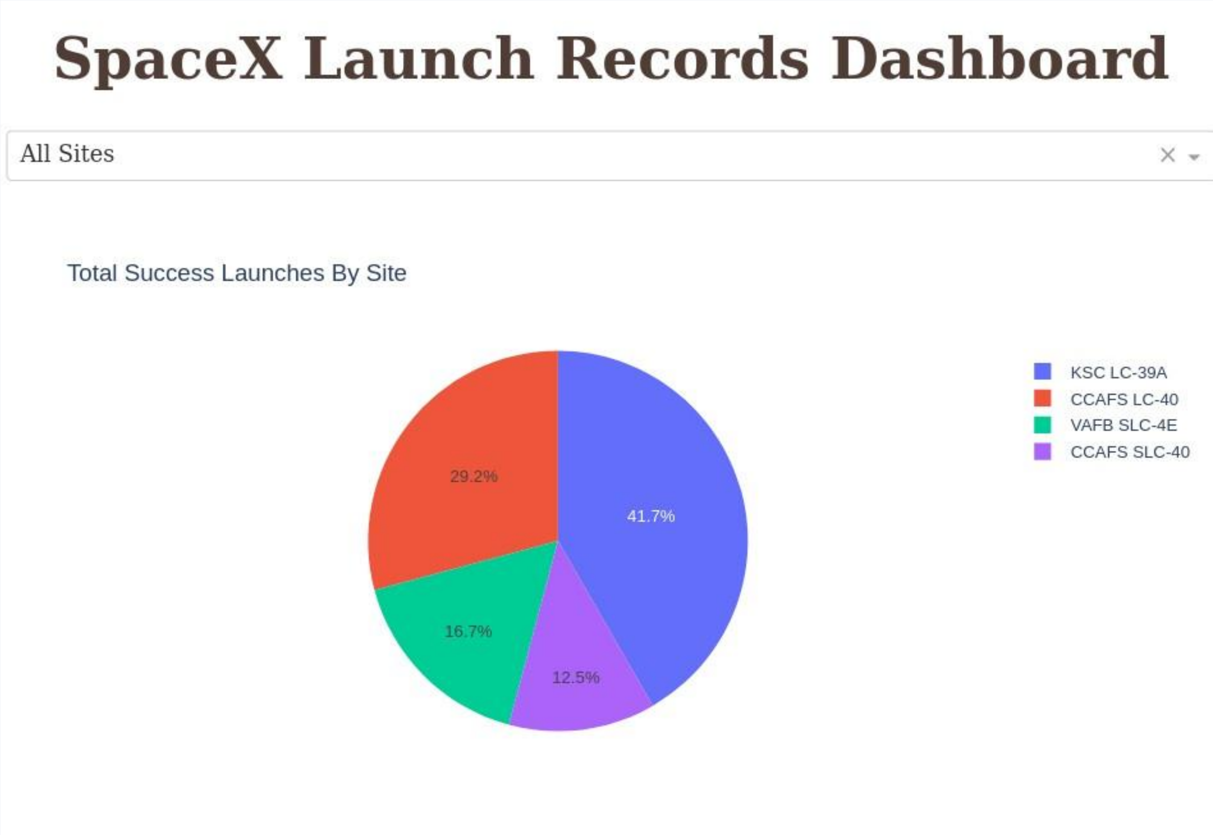


- Launch site KSC LC-39A has good logistics aspects, being near railroad and road and relatively far from inhabited areas.

Section 5

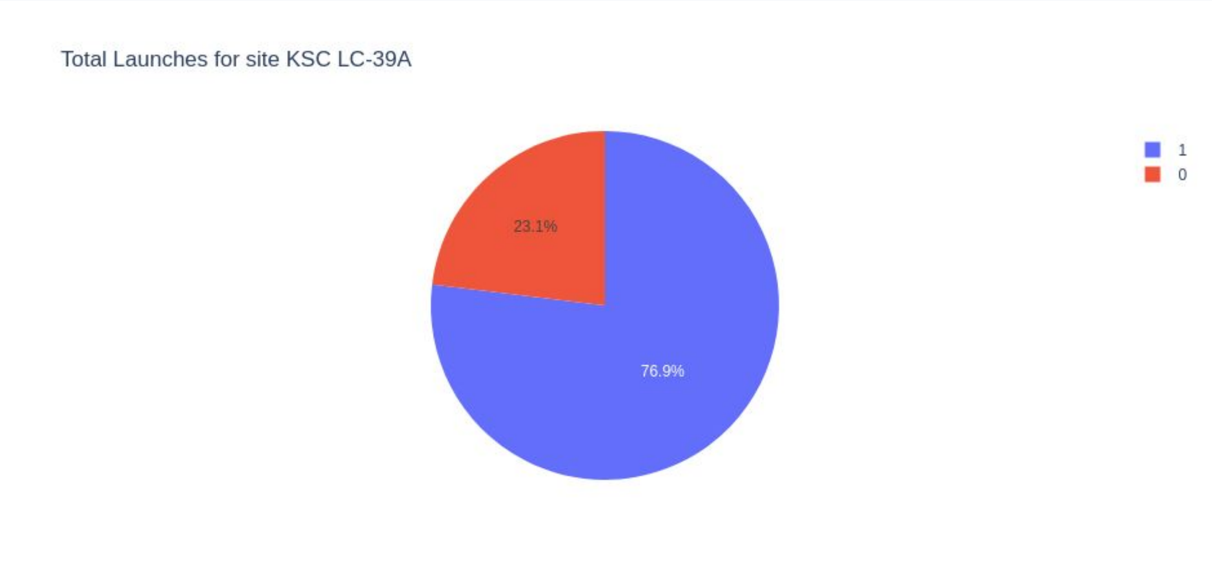
Build a Dashboard with Plotly Dash

Successful Launches by Site



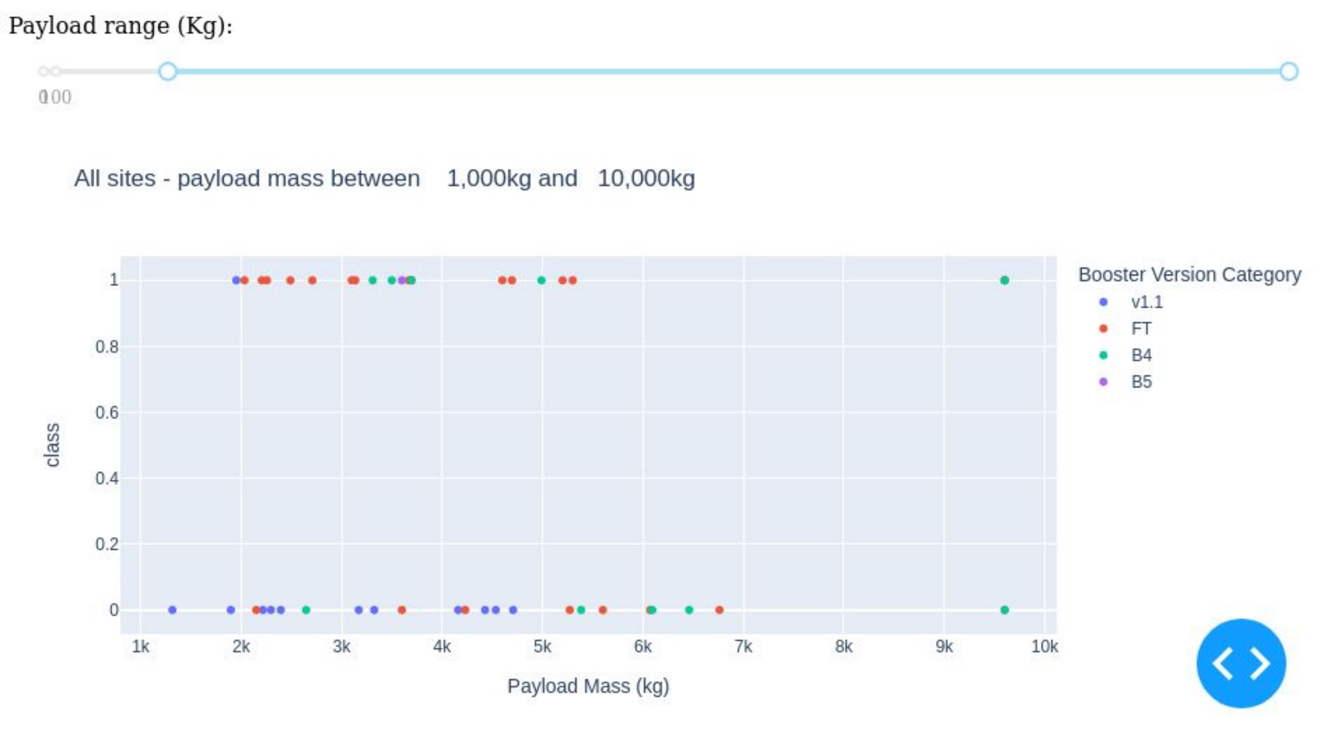
- The place from where launches are done seems to be a very important factor of success of missions.

Launch Success Ratio for KSC LC-39A



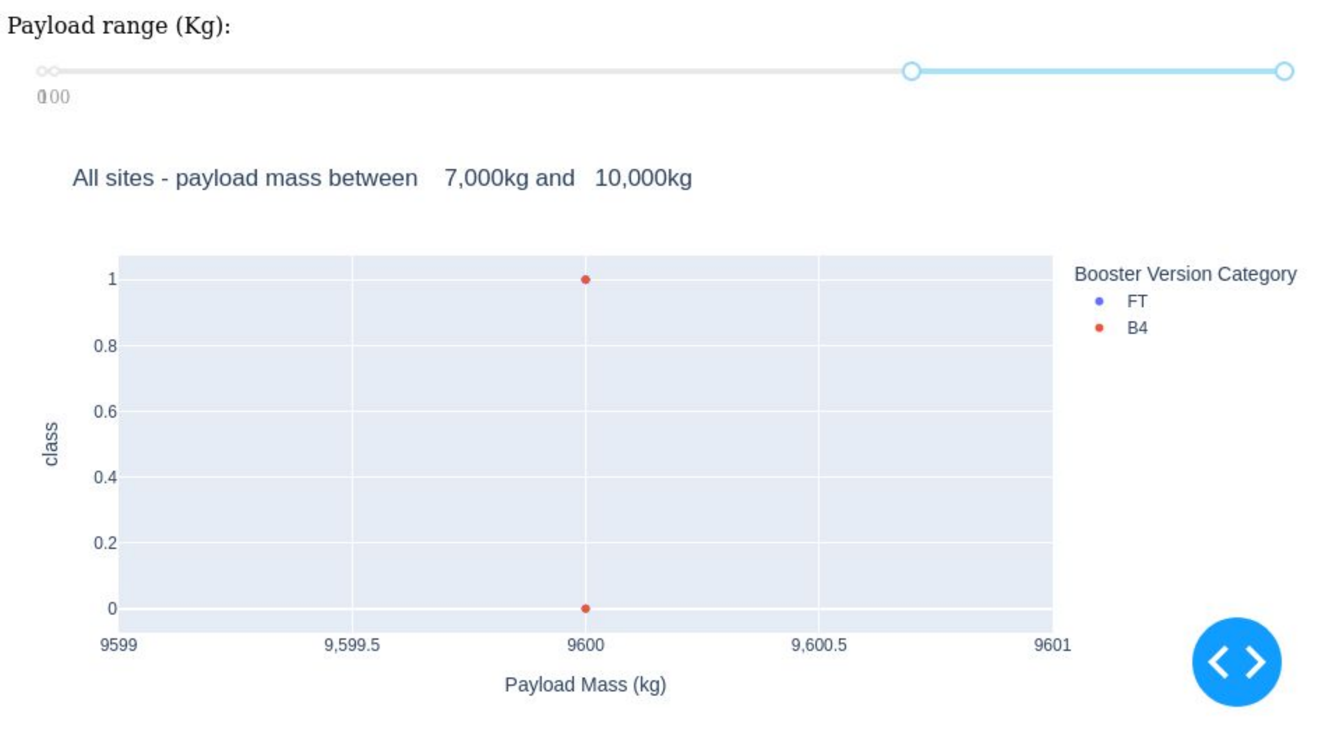
- 76.9% of launches are successful in this site.

Payload vs. Launch Outcome



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

Payload vs. Launch Outcome



- There's not enough data to estimate risk of launches over 7,000kg

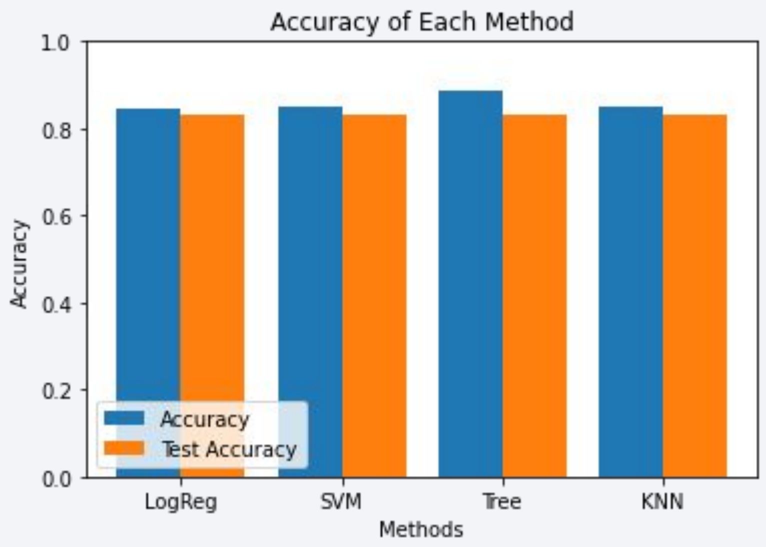
Section 6

Predictive Analysis (Classification)

Classification Accuracy

- Four classification models were tested, and their accuracies are plotted beside;

- The model with the highest classification accuracy is Decision Tree Classifier, which has accuracies over than 87%.



Confusion Matrix of Decision Tree Classifier

- Confusion matrix of Decision Tree Classifier proves its accuracy by showing the big numbers of true positive and true negative compared to the false ones.

Conclusions

- Different data sources were analyzed, refining conclusions along the process;
- The best launch site is KSC LC-39A;
- Launches above 7,000 kg are less risky;
- Although most of mission outcomes are successful, successful landing outcomes seem to improve over time, according the evolution of processes and rockets;
- Decision Tree Classifier can be used to predict successful landings and increase profits.

Appendix

- As an improvement for model tests, it’s important to set a value to `np.random.seed` variable;

Thank you!