IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Eric Howard
02/07/2024

Executive Summary

- The following methodologies were used to analyze data:

  - Data Collection using SpaceX API and BeautifulSoup Web Scraping

  - Exploratory Data Analysis (EDA), including data wrangling, visualization and interactive visual analytics

  - Machine Learning

- Summary of all results

  - Data was collected from online public sources

  - EDA allowed to identify which features are the best to predict success of launchings

  - Machine Learning Prediction showed the best model to predict which characteristics are most likely have a successful launch.

Introduction

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

- How are we going to compete?

  - By being able to estimate the total cost for launches, by predicting successful landings of booster rockets

  - Also where is the best places to launch rockets

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

    - Data from Space X was obtained from 2 sources:

        - Space X API (https://api.spacexdata.com/v4/rockets/)

        - WebScraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launc hes)

- Perform data wrangling

    - Collected data was wrangled by creating a landing outcome label and column

    - Which was based on outcome data after summarizing and analyzing features
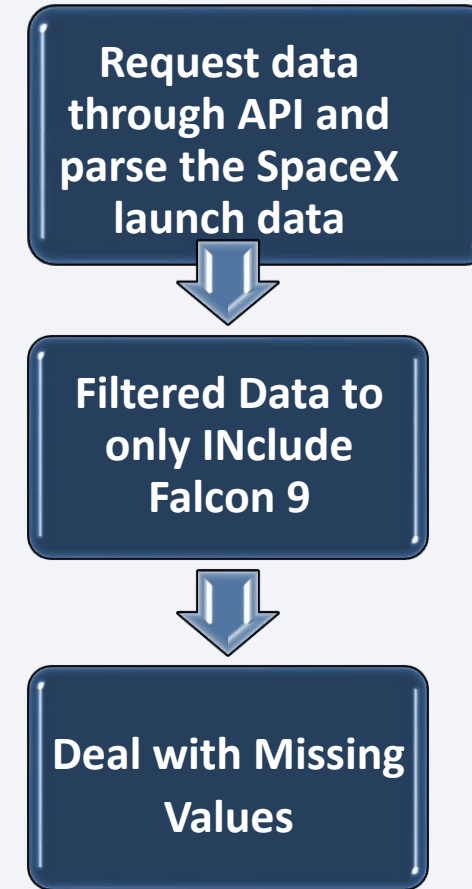
Methodology

Executive Summary

- Performed interactive visual analytics using Folium and Plotly Dash

- Perform predictive analysis using classification models
  - Data that was collected until this step were normalized, divided in training and test data sets and evaluated by four different classification models

Data Collection

- Data sets were collected from Space X API (https://api.spacexdata.com/v4/rockets/) and from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches), using web scraping techniques
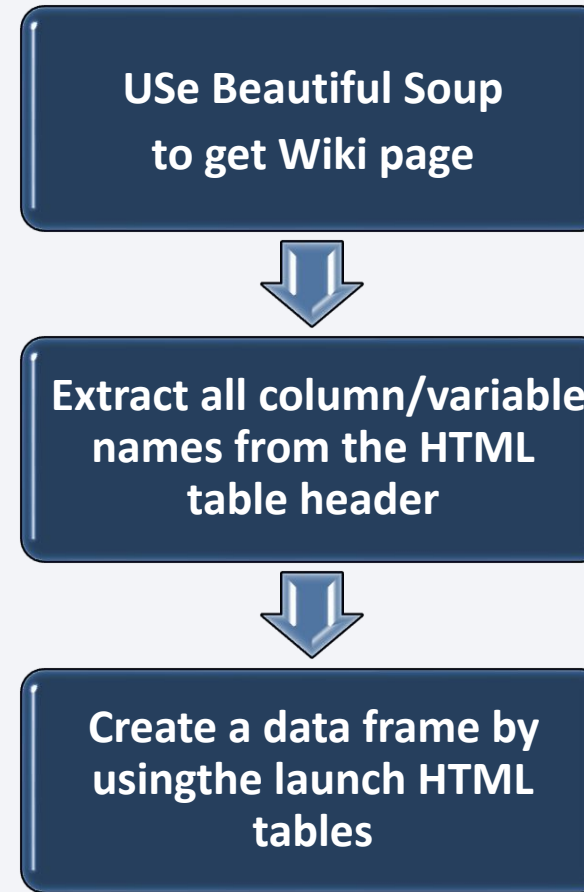
Data Collection – SpaceX API

- SpaceX offers a public API from where data can be obtained and then used;

- GitHub code:
  https://github.com/EricRHoward/Howard-Applied-Data-Science-Capstone-Coursera-IBM/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Request data through API and parse the SpaceX launch data

⬇

Filtered Data to only INclude Falcon 9

⬇

Deal with Missing Values

Data Collection - Scraping

- Data from SpaceX launches can also be obtained from Wikipedia with Beautifulsoup Python Pacakge

- GitHub code: https://github.com/EricRHoward/Howard-Applied-Data-Science-Capstone-Coursera-IBM/blob/main/jupyter-labs-webscraping.ipynb
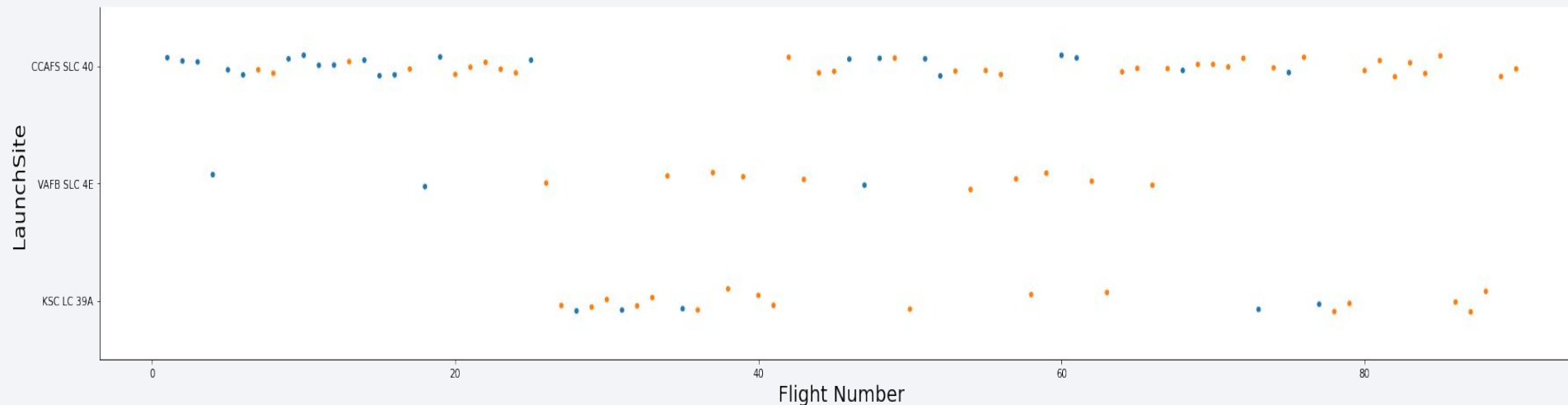
USe Beautiful Soup to get Wiki page

Extract all column/variable names from the HTML table header

Create a data frame by usingthe launch HTML tables

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.

- GitHub code:https://github.com/EricRHoward/Howard-Applied-Data-Science-Capstone-Coursera-IBM/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb
-
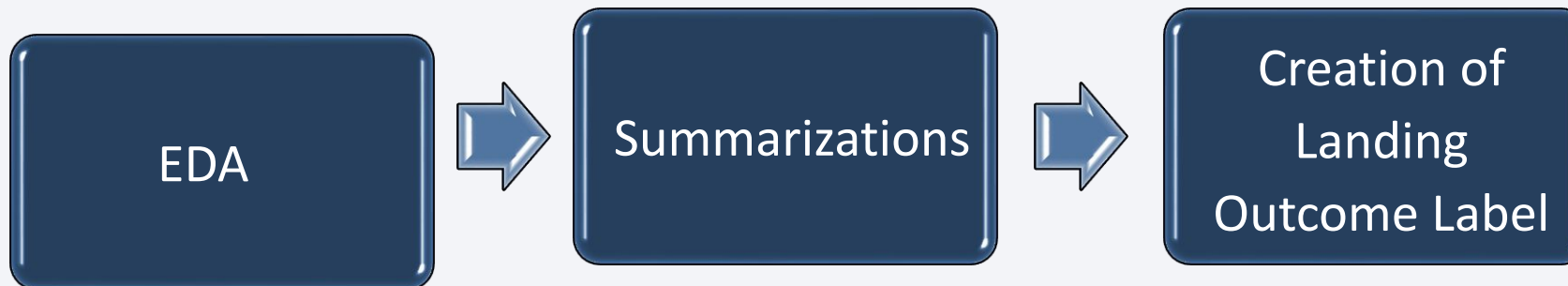
EDA with Data Visualization

- To explore data, scatterplots and barplots were used to visualize the relationship between 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

- Github code:
  https://github.com/EricRHoward/Howard-Applied-Data-Science-Capstone-Coursera-IBM/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

12

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.

GitHub code:
- https://github.com/EricRHoward/Howard-Applied-Data-Science-Capstone-Coursera-IBM/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

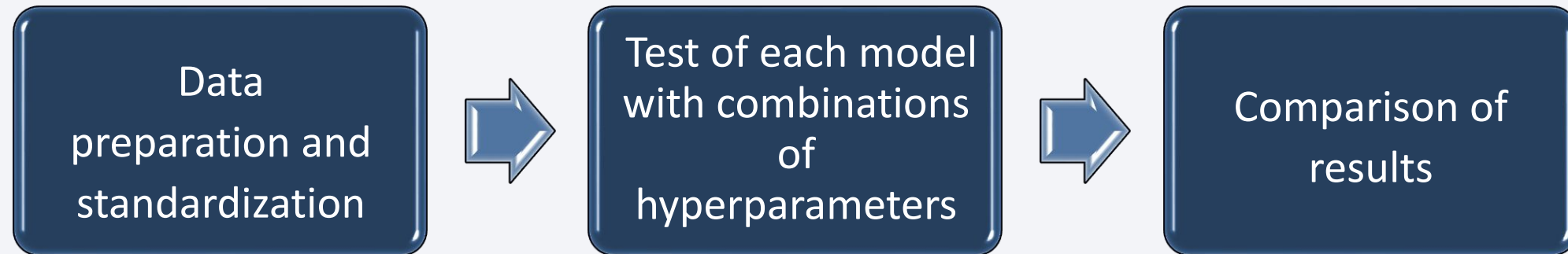# Built an Interactive Map with Folium

- Markers, circles, lines and marker clusters were used with Folium Maps
  - Markers indicate points such as launch sites
  - Circles indicate highlighted areas around specific coordinates, like the Johnson Space Center
  - Marker clusters indicates groups of events in each coordinate, such as launches in a launch site and lines were used to indicate distances between two coordinates

- GitHub code: https://github.com/EricRHoward/Howard-Applied-Data-Science-Capstone-Coursera-IBM/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

Built 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 analyze the relation between payloads and launch sites, helping to identify locate the best place to launch according to payloads

- GitHub code: https://github.com/EricRHoward/Howard-Applied-Data-Science-Capstone-Coursera-IBM/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Four classification models were compared: logistic regression, support vector machine, decision tree and k nearest neighbors

| Data preparation and standardization | → | Test of each model with combinations of hyperparameters | → | Comparison of results |

- GitHub code:
  https://github.com/EricRHoward/Howard-Applied-Data-Science-Capstone-Coursera-IBM/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

- Exploratory data analysis results:

  - Space X 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 was possible to identify that launch sites with key benefits such as safety measures, near ocean, and logistic infrastructure

- Most launches happens on the East Coast
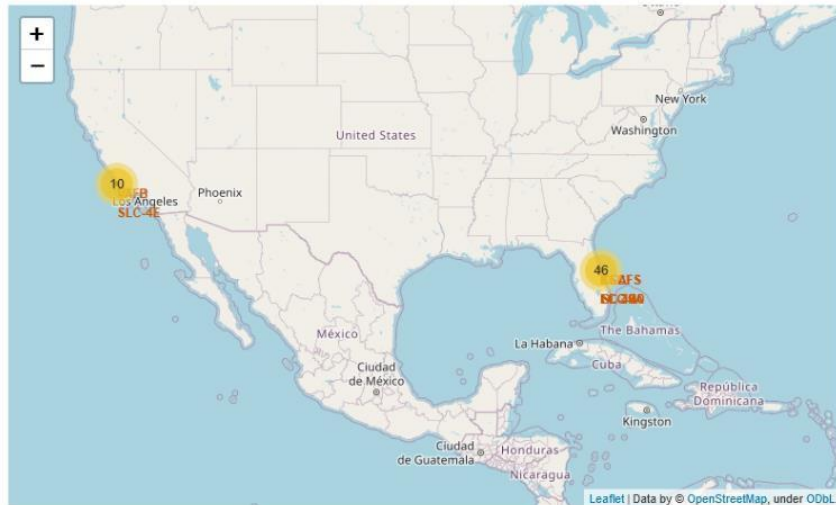
Results

- Predictive Analysis showed that Decision Tree Classifier is the best model to predict successful landings, had an accuracy over 87% and the accuracy for test data was over 94%
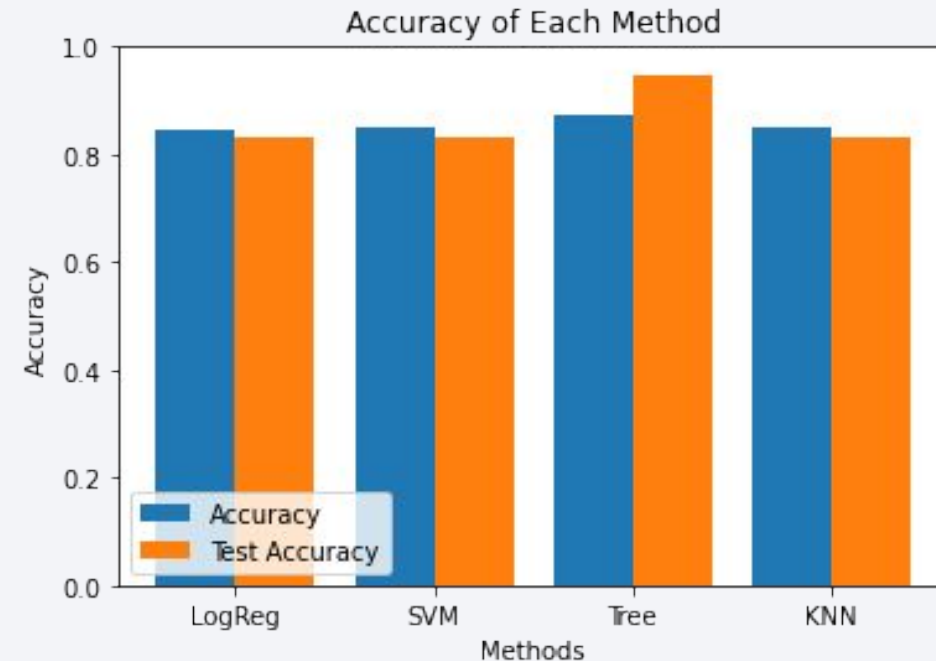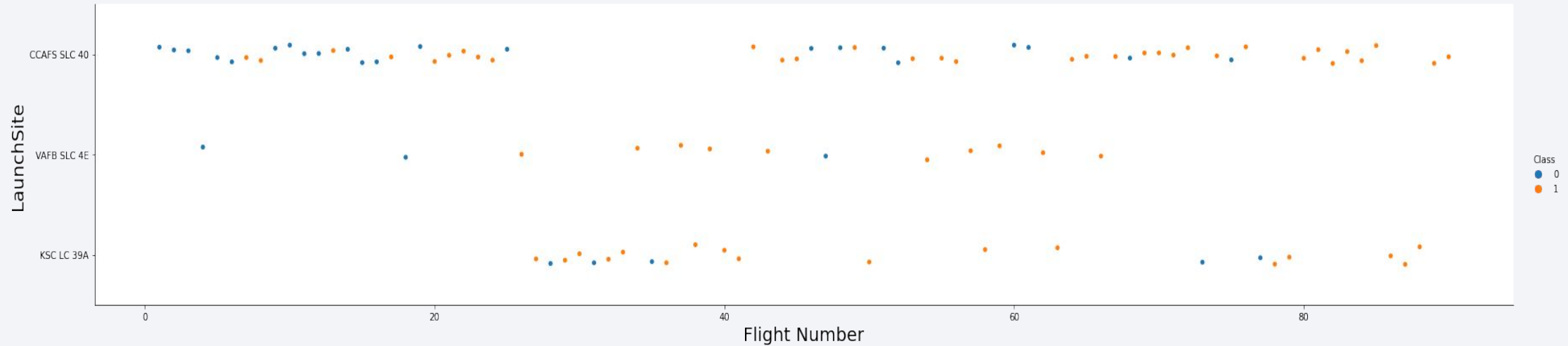

Accuracy of Each Method

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
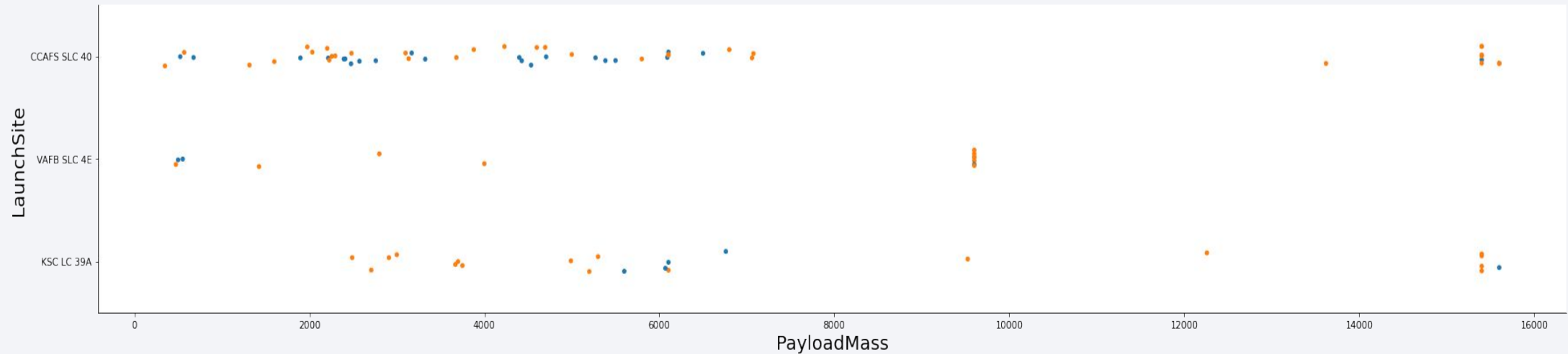
- Payloads over 9,000kg have a good success rate

- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.
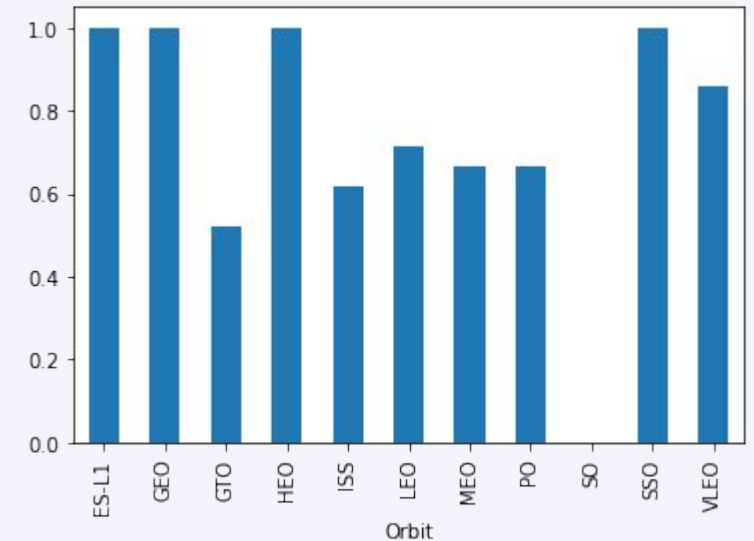
21

Success Rate vs. Orbit Type

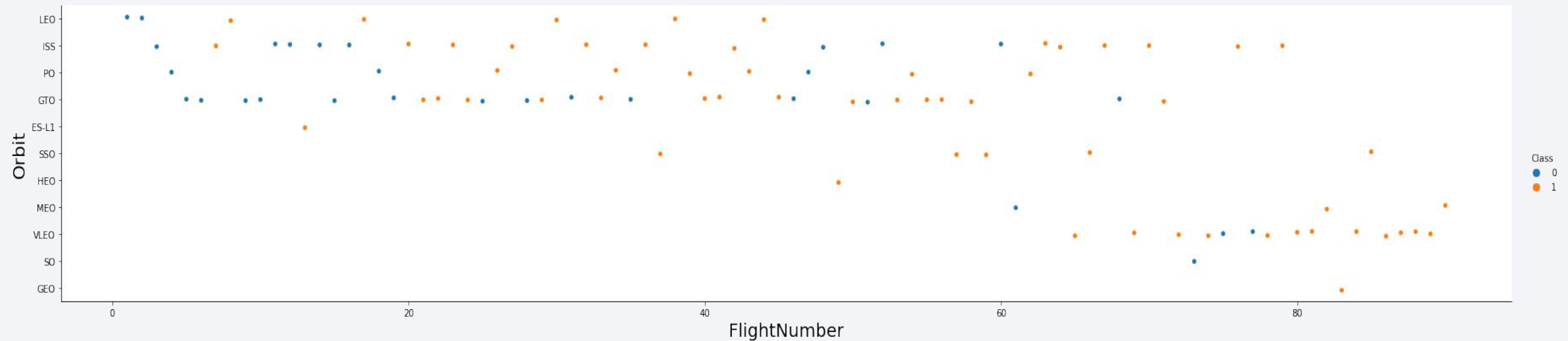- The biggest success rates happens to orbits:

  - ES-L1

  - GEO

  - HEO

  - SSO

- Followed by:

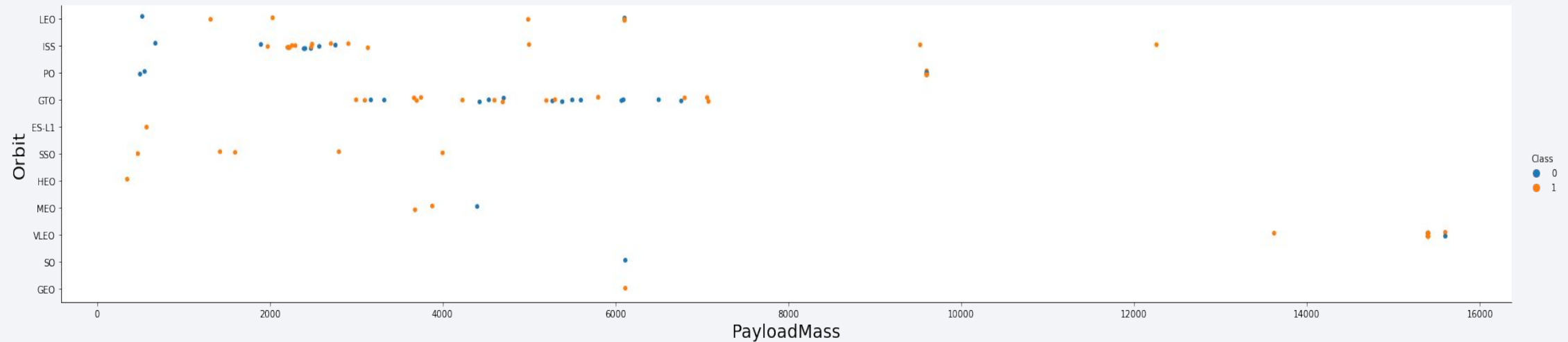 - VLEO (above 80%)

  - LFO (above 70%)

Flight Number vs. Orbit Type

- 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

- There is seemingly 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

24

Launch Success Yearly Trend

- Success rate started increasing in 2013 and seems to be in upward trajectory

- It seems that the first three years were a period of adjusts and improvement of technology.
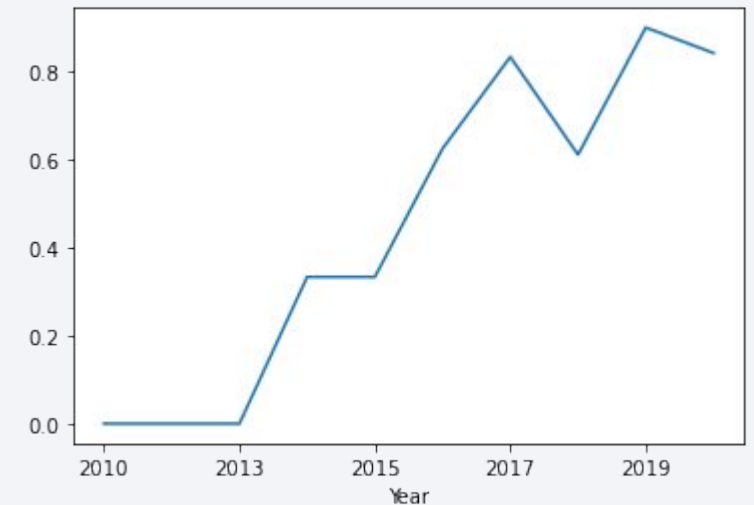
All Launch Site Names

- According to gathered data, there are four launch sites:

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

- They were 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 calculated above, by summing all payloads whose codes contain '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

- 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
2015-12-22

- By filtering data by successful landing outcome on ground pad and getting the minimum value for date it's possible to identify the first occurrence, 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 B1021.2
F9 FT B1031.2
F9 FT B1022
F9 FT B1026

- Selecting distinct booster versions according to the filters above, these 4 are the result

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

- 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 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 are the boosters which have carried the maximum payload

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

- Only two failures

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

Section 4

Launch Sites
Proximities Analysis

All launch sites
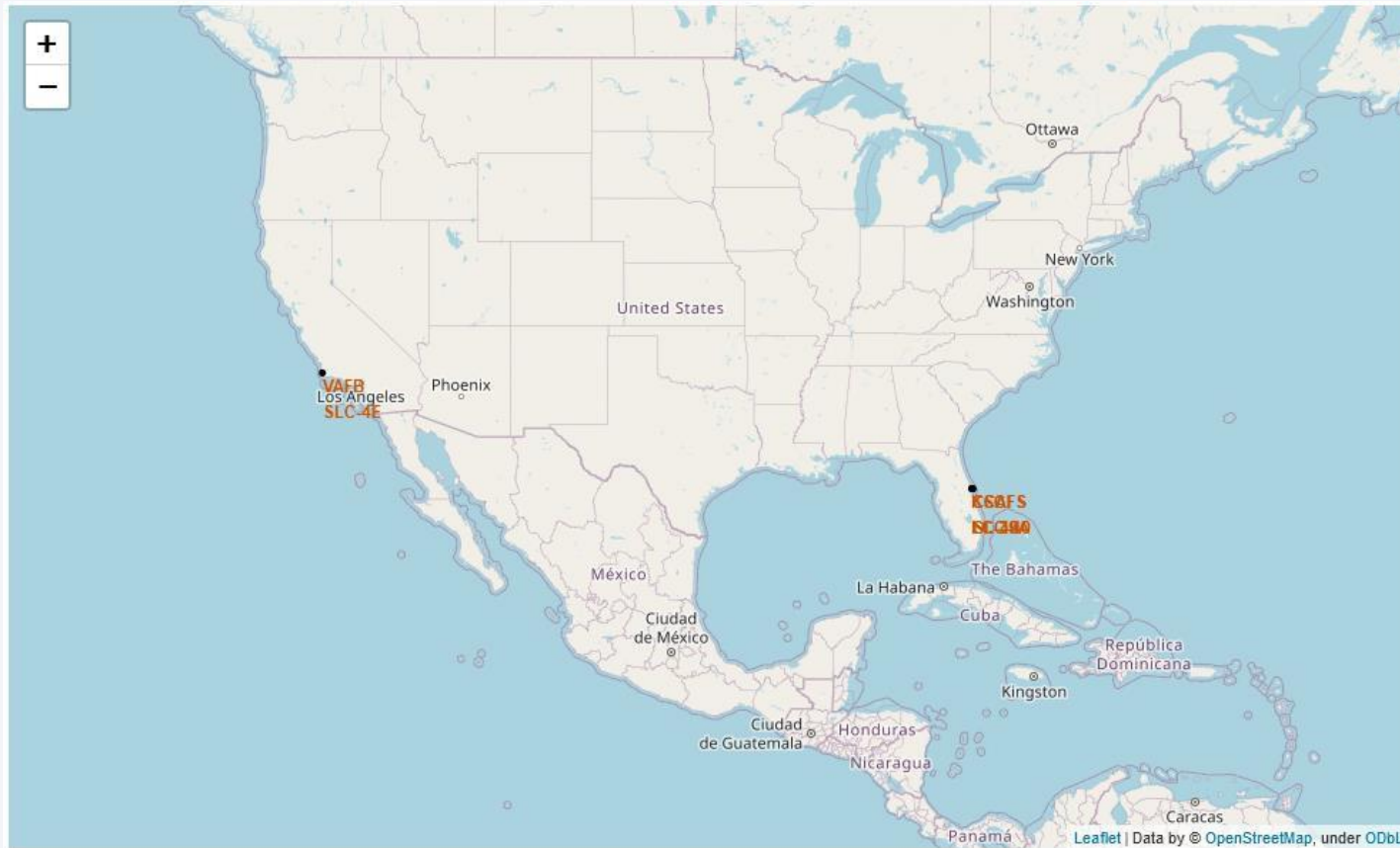
- Launch sites are near sea for safety and near railroad for easy transportation
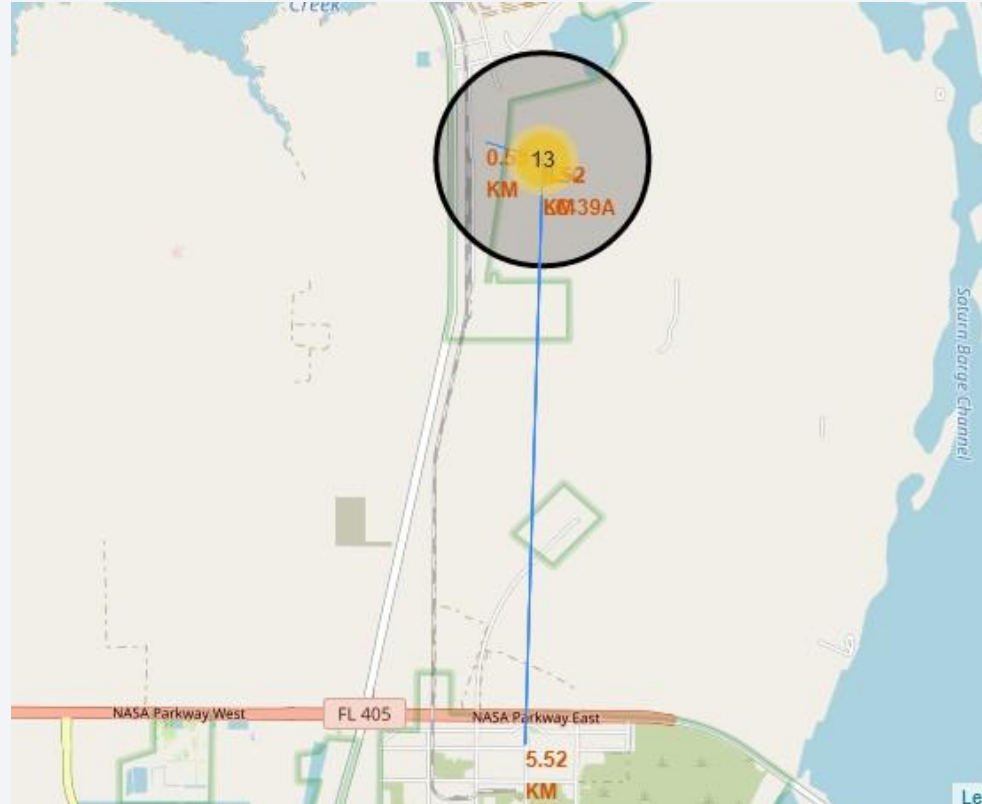
Launch Outcomes by Site

- Example of KSC LC-39A launch site launch outcomes
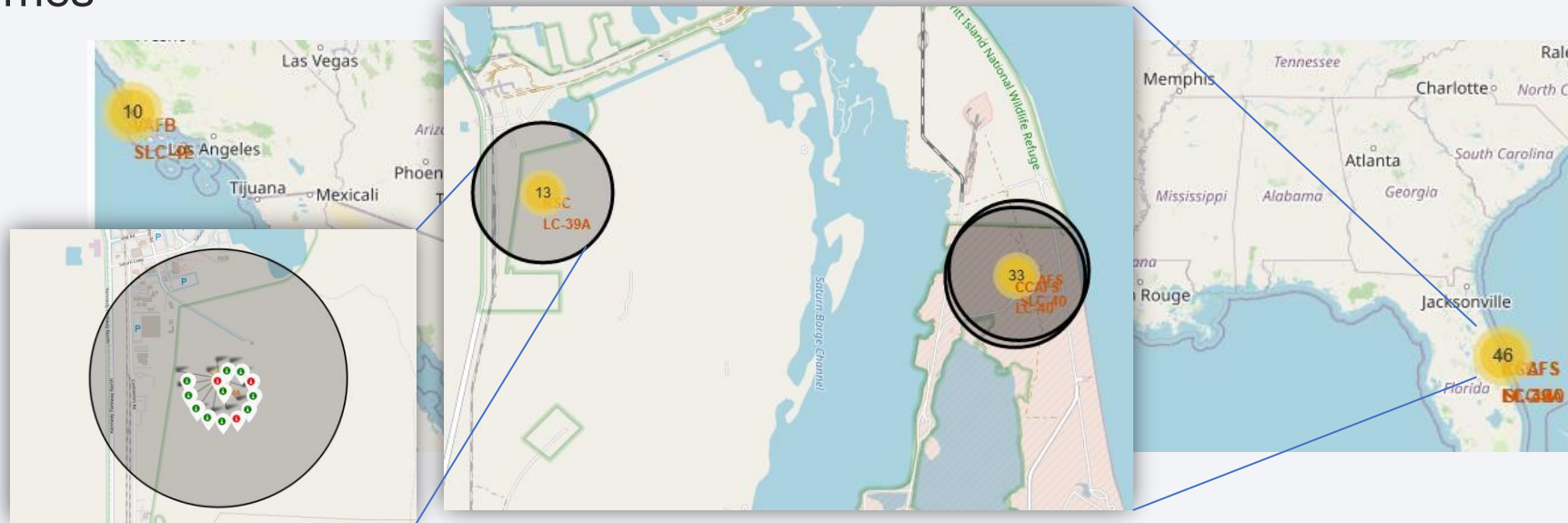
- Green markers indicate successful and red ones indicate failure
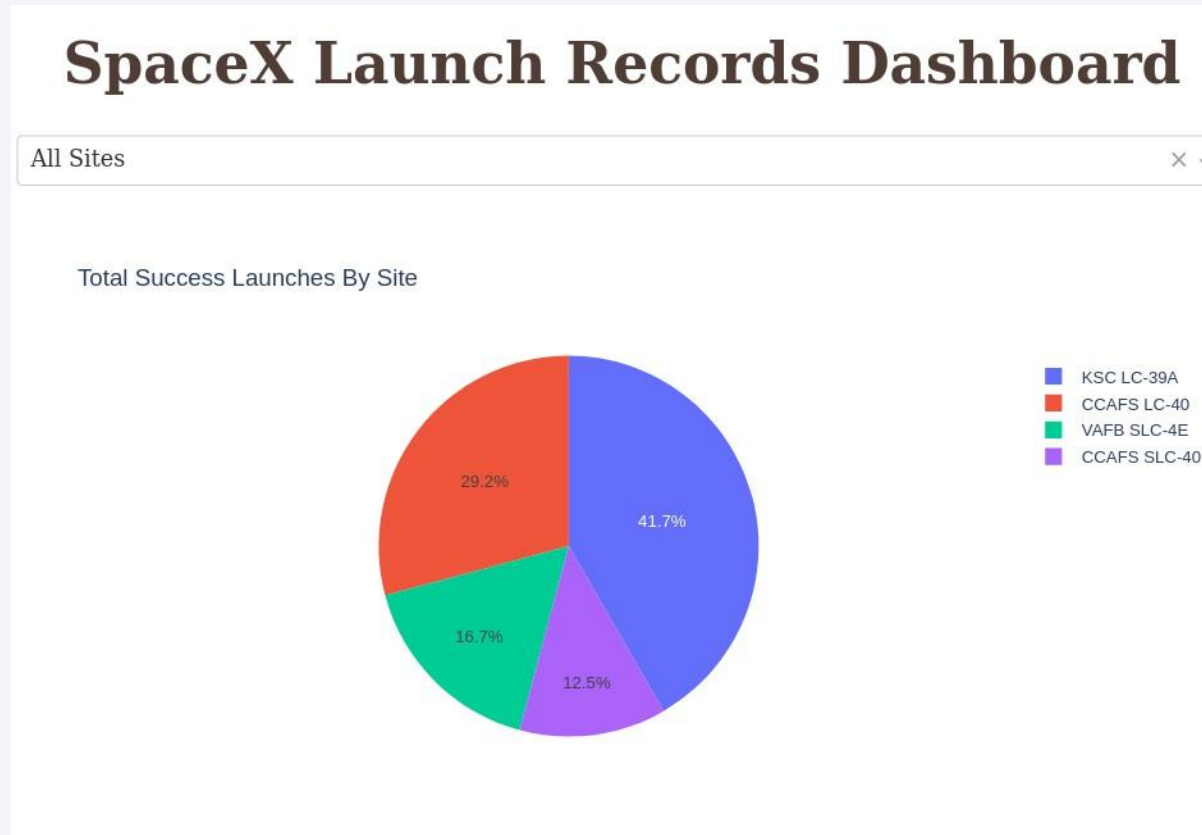
Logistics and Safety

- Launch site KSC LC-39A has good logistics aspects, being near railroad and road and relatively far from inhabited areas
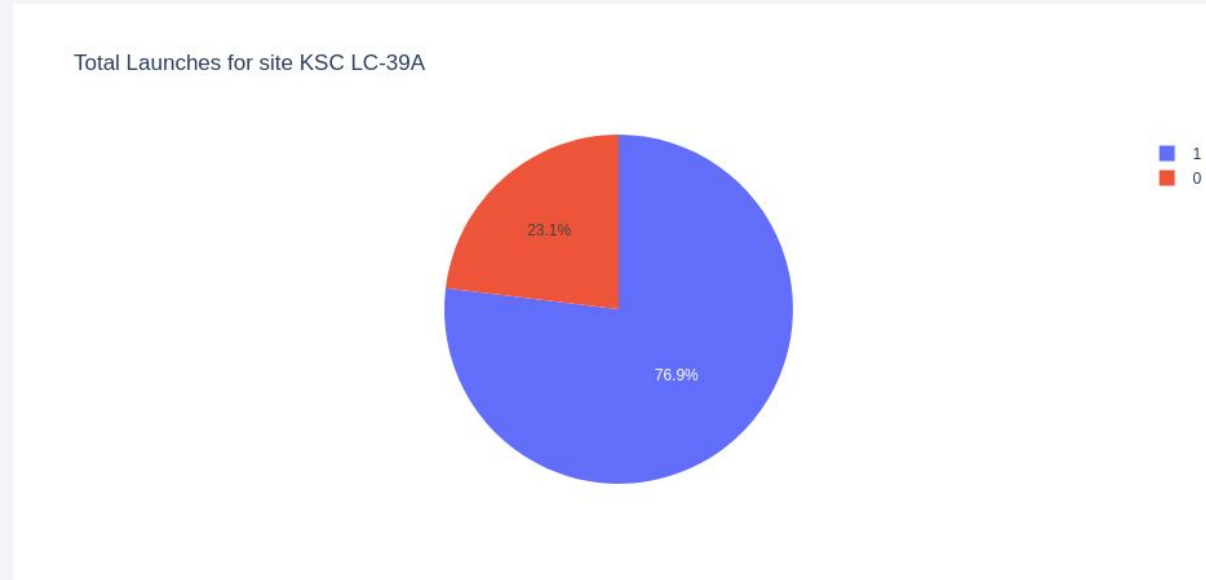
Section 5

Build a Dashboard
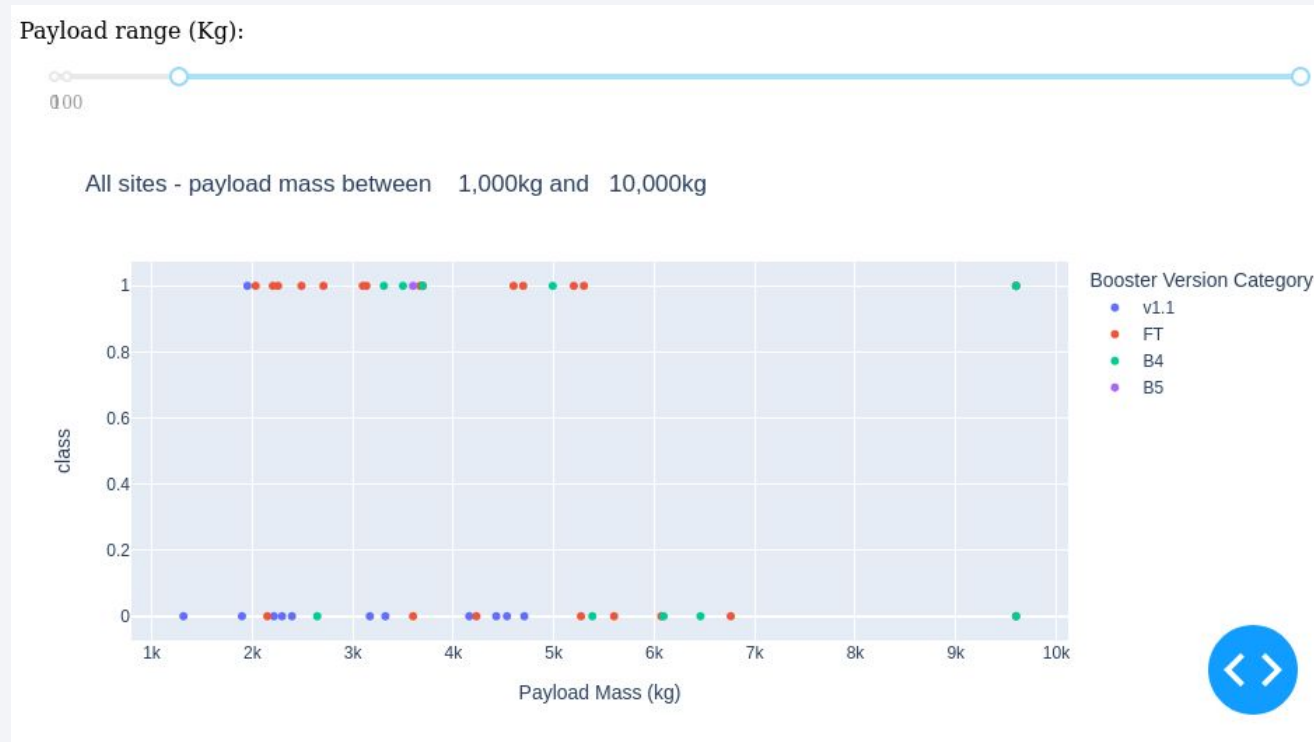# with Plotly Dash

Successful Launches by Site

- The place from where launches are done suggests that launch locations are important

Launch Success Ratio for KSC LC-39A

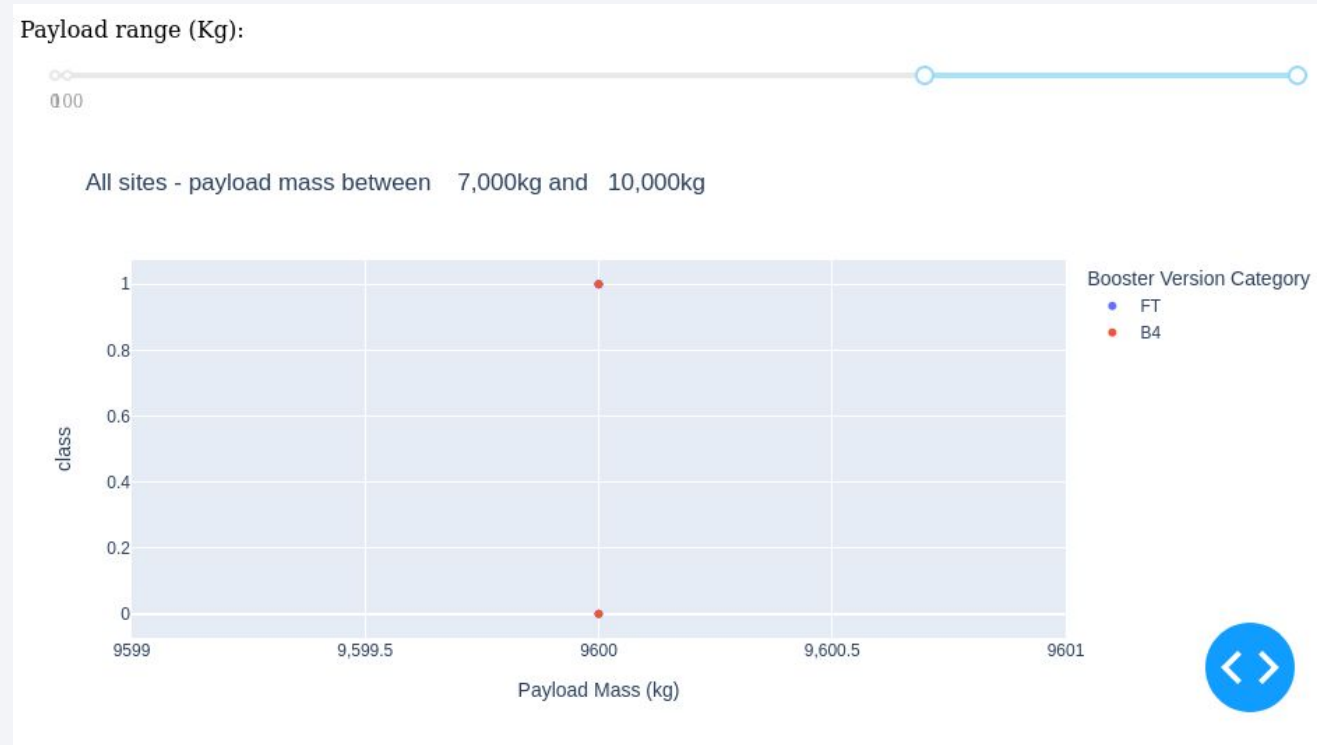Total Launches for site KSC LC-39A

23.1%

76.9%

■ 1
■ 0

- 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
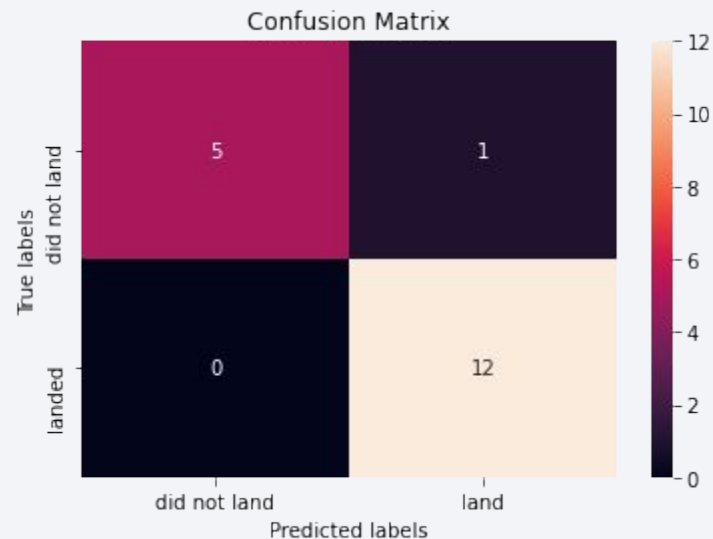
Section 6

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%
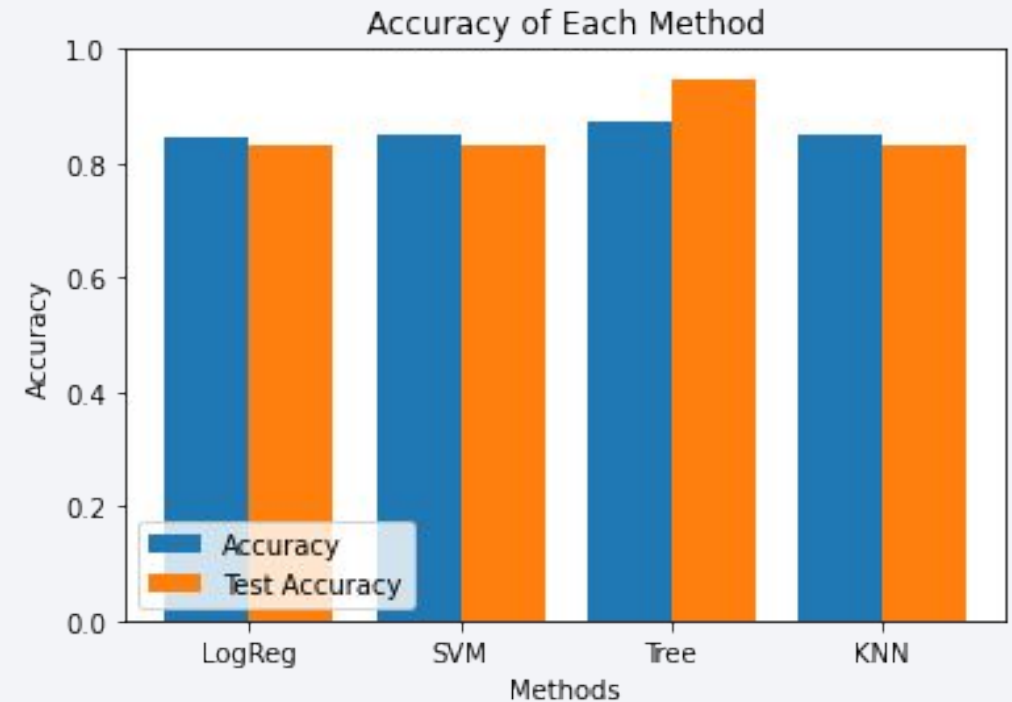
Confusion Matrix of Decision Tree Classifier

- Confusion matrix of Decision Tree Classifier proves its accuracy by showing the 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 7,000kg 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 by having more successful landings

Appendix

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

Thank you!