IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Arman Wirawan
06 June 2023

Outline

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

Executive Summary

- Methodologies used in this analysis:

1. Data Collection using web scraping and SpaceX API
2. Exploratory Data Analysis (EDA), including data wrangling, data visualization and interactive visual analytics
3. Machine learning prediction

- Summary of all results

1. Valuable data was collected through public sources
2. EDA allowed to identify which features are the best to predict success of rocket launch
3. Machine learning prediction showed the best models 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 Space Y to compete with Space X
- Desirable answers:

1. The best way to estimate the total cost for launches, by predicting successful landings of the first stage of rockets
2. Where is the best place to make launches

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data from Space X was obtained from 2 sources:
    1. Space X API(https://api.spacexdata.com/v4/rockets/)
    2. WebScraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)
- Perform data wrangling
  - Collected data was enriched by creating a landing outcome label based on outcome data after summarizing and analyzing features
- Perform exploratory data analysis (EDA) using visualization and SQL

Methodology

- Perform 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, being the accuracy of each model evaluated using different combinations of parameters.

Data Collection

- Data were collected from Space X API and from Wikipedia using web scraping technics

1. Space X API: https://api.spacexdata.com/v4/rockets/
2. Wikipedia: https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches

Data Collection – SpaceX API

- Space X offers a public API from where data can be obtained and then used
- This API was used according to the flowchart as you can see. The data then is persisted.
- Source: https://github.com/armanw96/Arman-Applied-Data-Science-IBM-/blob/main/Data%20Collection%20API%20IBM.ipynb

Request API and parse the Space X launch data

Filter data to only include Falcon 9 launches

Deal with missing values

Data Collection - Scraping

- Data from Space X launches can also be obtained from Wikipedia
- Data are scraped from Wikipedia according to the flow chart and then persisted
- Source: https://github.com/armanw96/Arman-Applied-Data-Science-IBM-/blob/main/Data%20Collection%20with%20Web%20Scraping%20IBM.ipynb

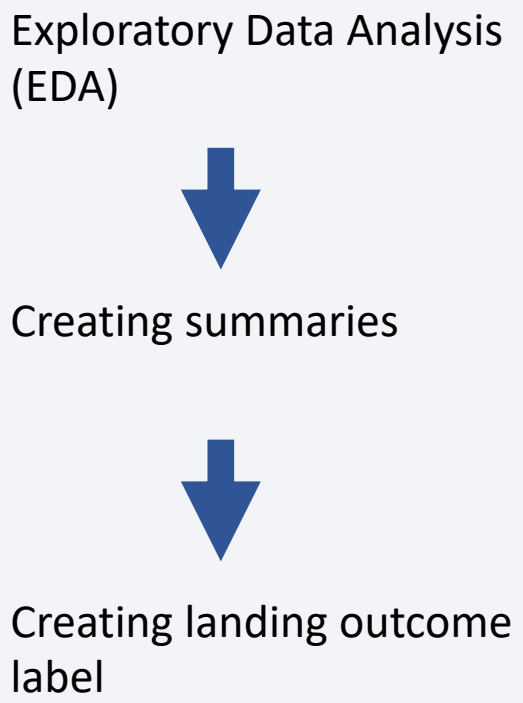
Request the Falcon 9 Launches Wikipedia

Scrape all the column/variable from the website

Create a data frame by parsing the launches HTML table

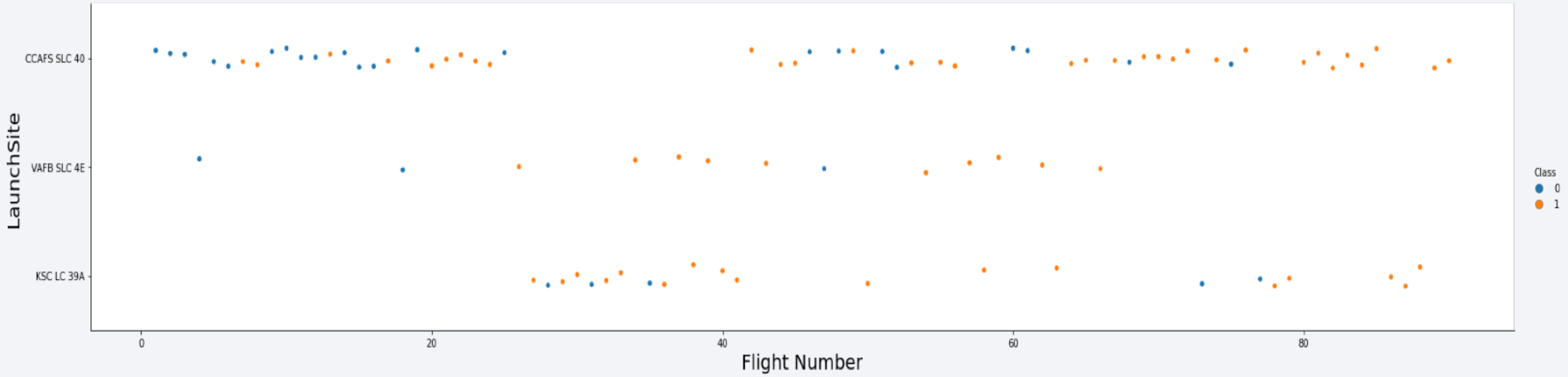
Data Wrangling

- Initially some Exploratory Data Analysis (EDA) is performed on the dataset
- Then the summaries launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type were calculated.
- At last, the landing outcome label was created from the Outcome column
- Source: https://github.com/armanw96/Arman-Applied-Data-Science-IBM-/blob/main/IBM%20Data%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
- Source code: https://github.com/armanw96/Arman-Applied-Data-Science-IBM-/blob/main/IBM%20EDA%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
- 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/armanw96/Arman-Applied-Data-Science-IBM-/blob/main/EDA%20SQL%20IBM.ipynb

Build an Interactive Map with Folium

- Markers, circles, clusters, and lines were used

1. Markers is used to indicate launch sites
2. Circles is used to indicate highlighted areas around specific coordinates, like NASA Johnson Space Center
3. Marker clusters is used to indicate group of events in each coordinate, like launches in a launch site
4. Lines are used to indicate distances between two coordinates

Source code: https://github.com/armanw96/Arman-Applied-Data-Science-IBM-/blob/main/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

- The following graphs and plots were used to visualize data

1. Percentage of launches by site
2. Payload range

The combination allowed us to quickly analyze the relation between payloads and launch site, helping to identify where is best place to launch according to payloads.

Source code: https://github.com/armanw96/Arman-Applied-Data-Science-IBM-/blob/main/IBM%20spacex_dash_app.py

Predictive Analysis (Classification)

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

Source code: https://github.com/armanw96/Arman-Applied-Data-Science-IBM-/blob/main/IBM%20Machine%20Learning%20Prediction.ipynb

Data Preparation and standardization

Test of each model with combinations of hyperparameters

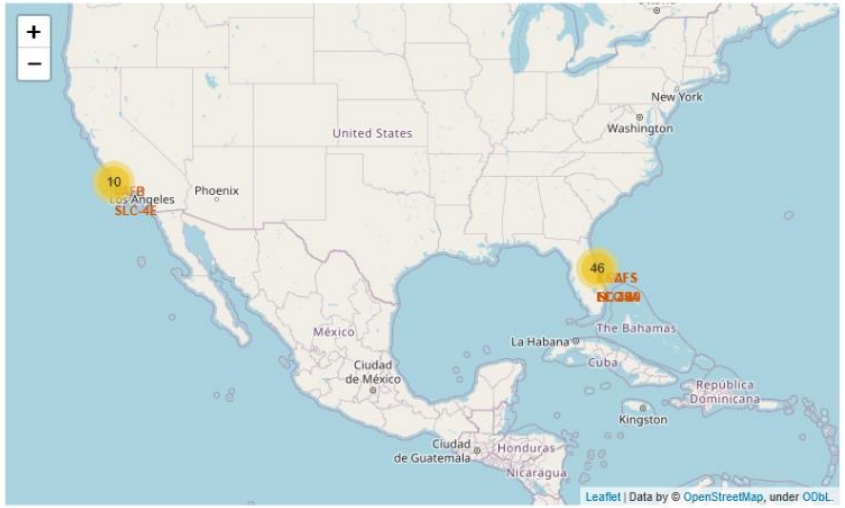
Comparison of results

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 five year after the first launch
- Many falcon 9 booster versions were successful at landing in drone ship 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

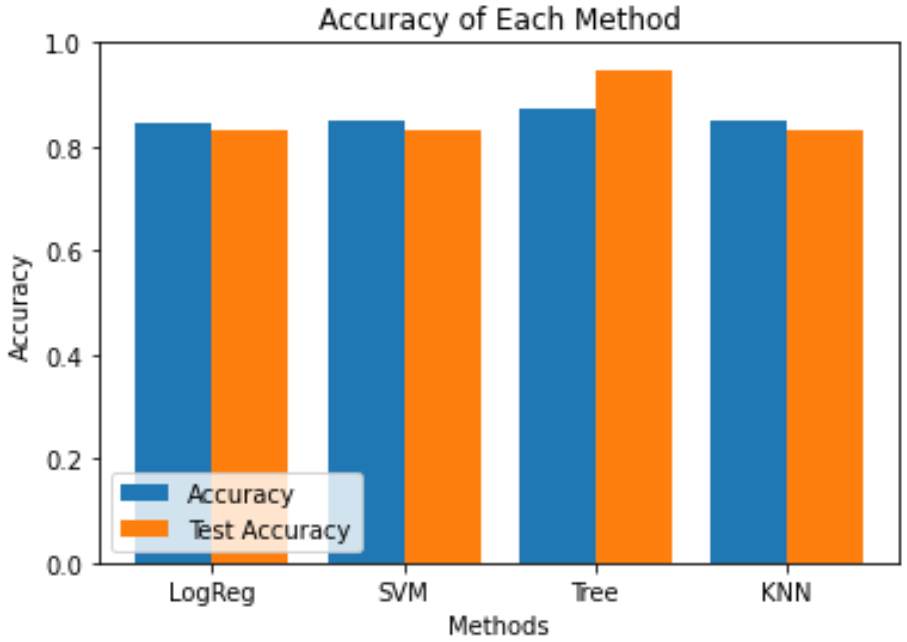
Results

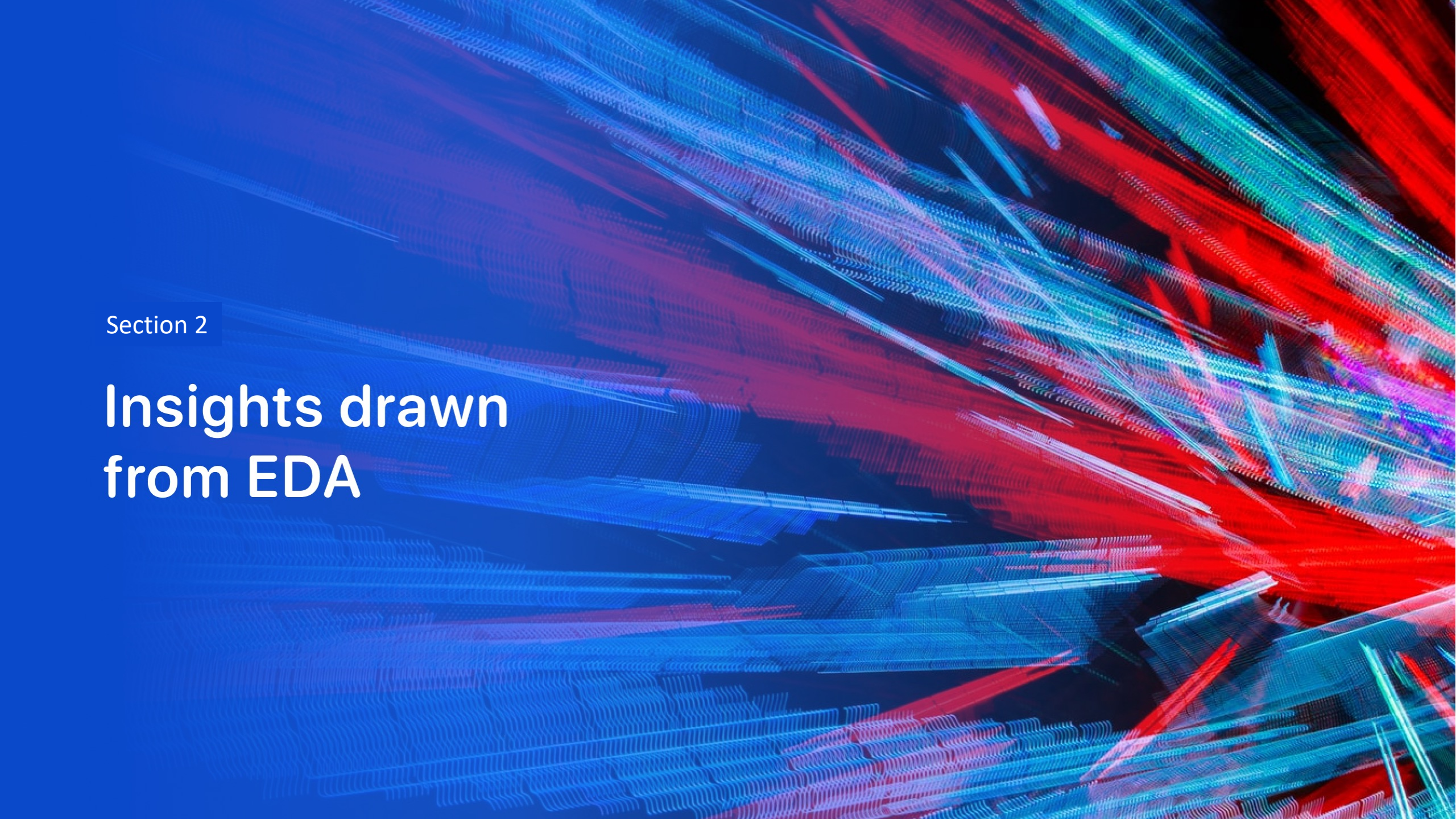
- Using interactive analytics was possible to identify that launch sites use to be in safety places, near sea, for example and have a good logistic infrastructure around.
- Most launches happens at east cost launch sites.



Results

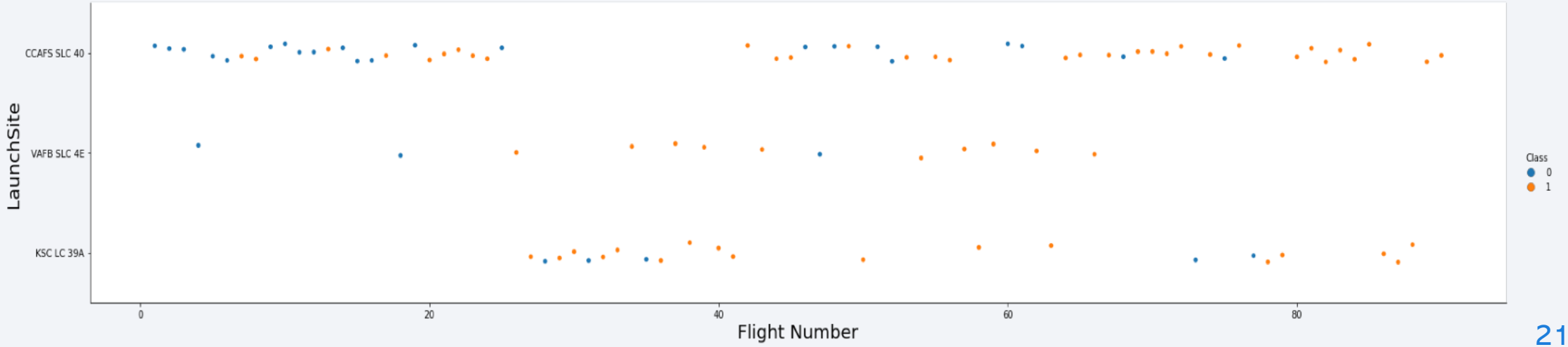
- Predictive analysis showed that Decision Tree Classifier is the best model to predict successful landings, having accuracy over 87% and accuracy for test data over 94%




Section 2
Insights drawn from EDA

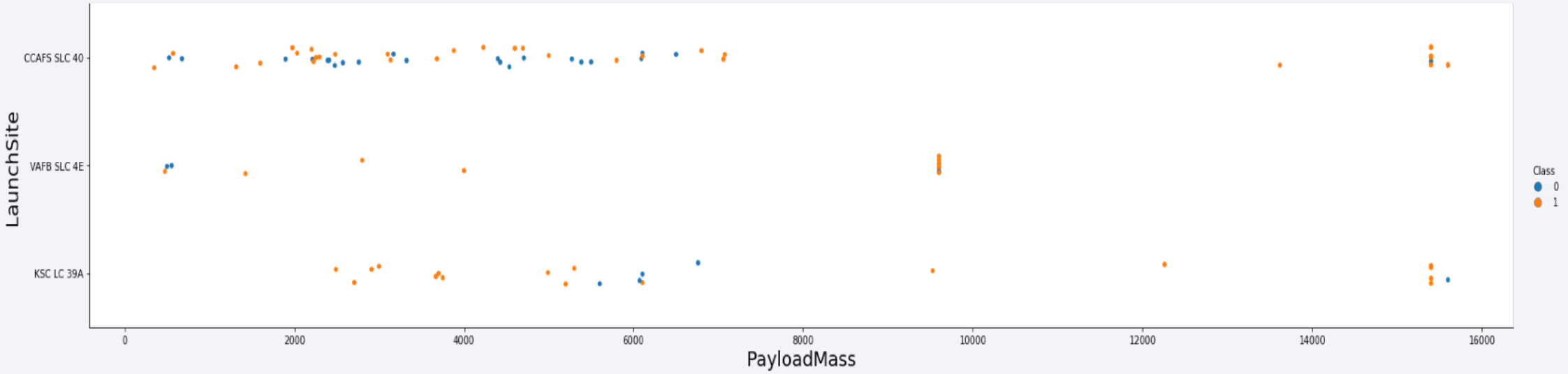
Flight Number vs. Launch Site

- According to the plot below, it's possible to verify that the best launch site nowadays is CCAF5 SLC 40, where most recent launches were successful
- Second place belongs to VABF SLC 4E
- Third place KSC LC 39 A
- Its also possible to see over time that success rate of launches have improved.



Payload vs. Launch Site

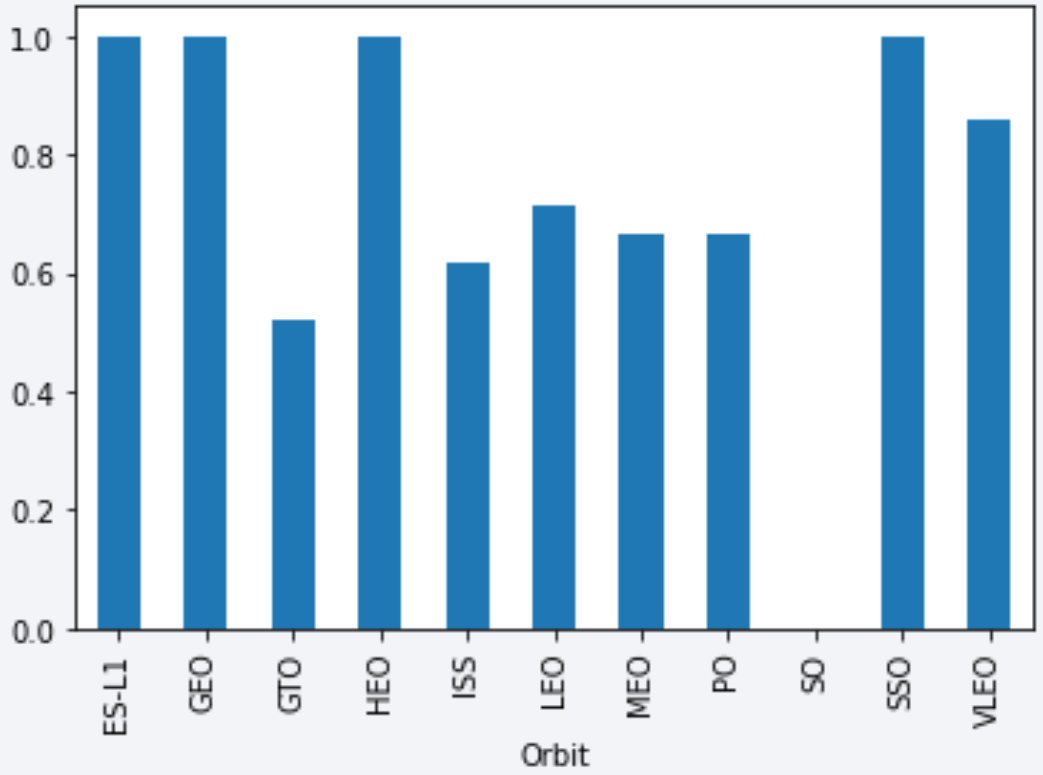
- According to the plot below, payloads over 9,000KG have excellent success rate
- However payload over 12,000 KG seems to be only possible from CCAFS SLC 40 and KSC LC 39A launch sites



Success Rate vs. Orbit Type

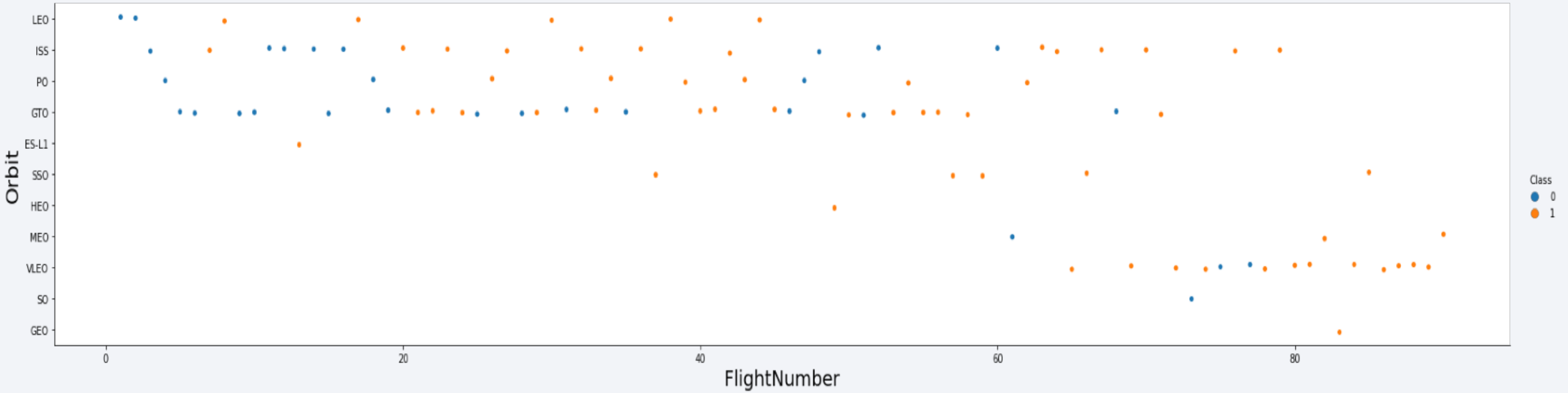
- The most successful launch are to the following orbits:

1. ES-L1
2. GEO
3. HEO
4. SSO
5. VLEO
6. LFO



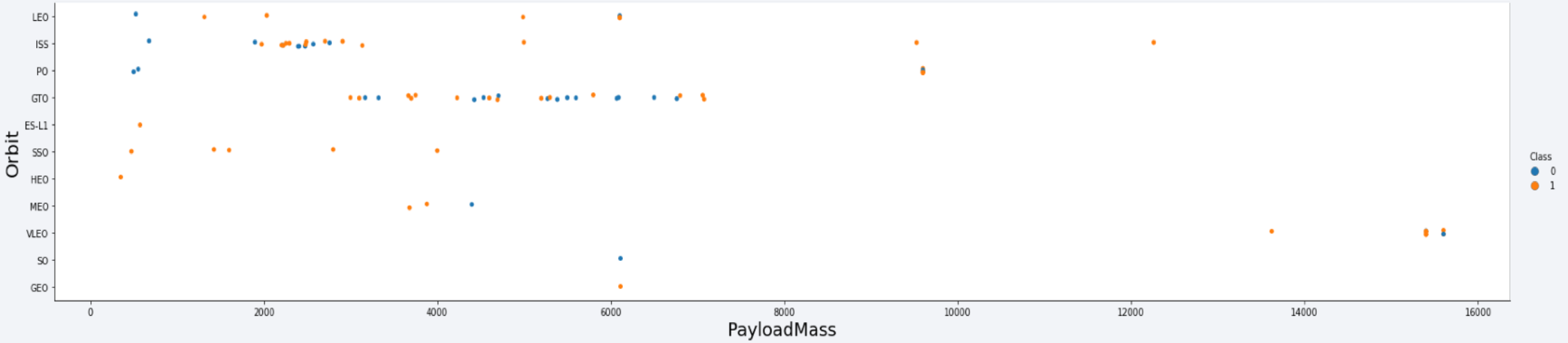
Flight Number vs. Orbit Type

- According to the graph below VLEO orbit seems like a new business opportunity as it is seeing a new rising trend in its orbit. Furthermore the success rate of all orbits have steadily improved over time.



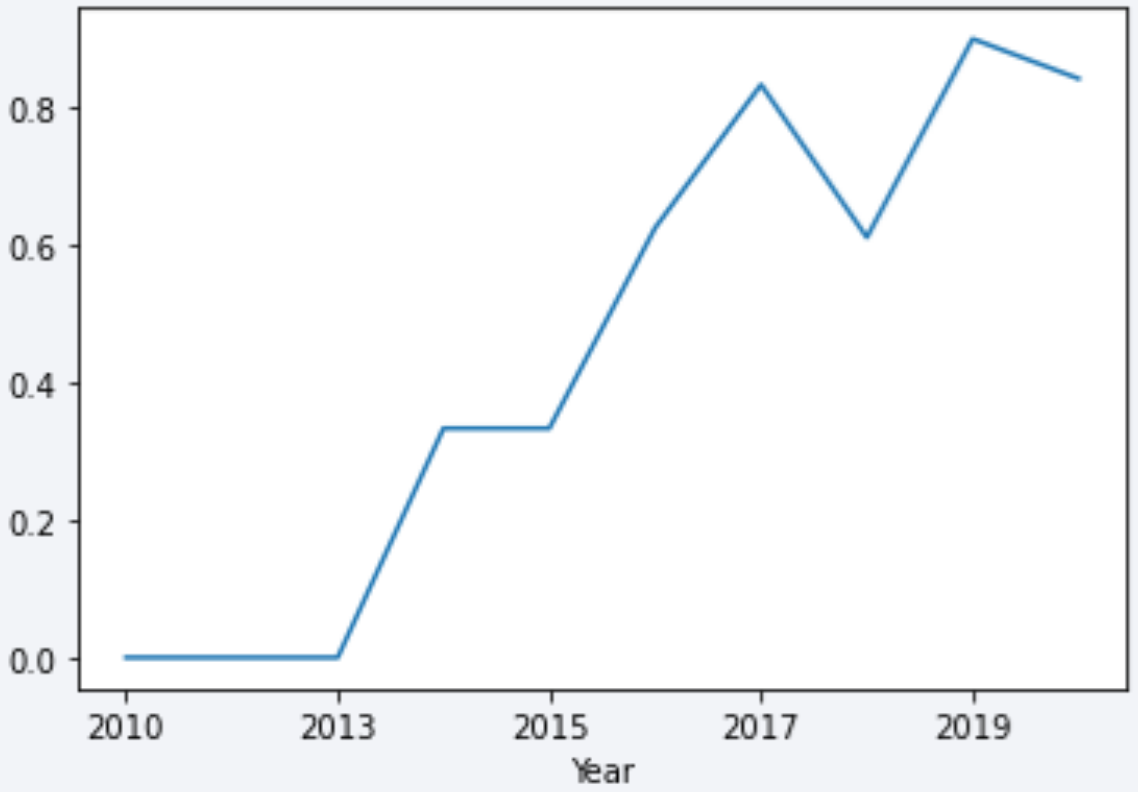
Payload vs. Orbit Type

- According to the chart below ISS orbit has the widest range of payloads
- Moreover there SO and GEO has the fewest launches
- Furthermore, compare to the rest of the orbit there is no relation between payload and success rate to GTO orbit



Launch Success Yearly Trend

- Launch Success continues to increase following 2013 till 2020
- There's a set back during 2017-2019 which could indicate new technology are being tested.



All Launch Site Names

- From the date we could find 4 launch names
- They are obtained by selecting unique of “launch_site” value from the data set

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- These are the 5 examples of launches from Cape Canaveral

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

- The total payload carried by boosters from NASA in Kg
- Total payload calculated below, by summing all payloads whose codes contains ‘CRS’, which corresponds to NASA

total_payload
111268

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1
- We find the average payload by calculating the average payload mass from filtering the data to only include the booster version above in Kg

avg_payload
2928

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad
- First we have to filter the data by successful landing outcome on ground pad. Then we can get the minimum value to figure out the first occurrence which happens in the data below:

first_success_gp
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- Once we filter for the DISTINCT booster versions we are left with the following 4 boosters:

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- To find the right outcome you have to group by the mission_outcome thus the following query results showed:

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- We have to subquery this query because you need to figure out how many boosters have carrier the maximum payload mass before you could DISTINCT the boosters that 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
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- They are the failed landing outcomes in drone ship in year 2015

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

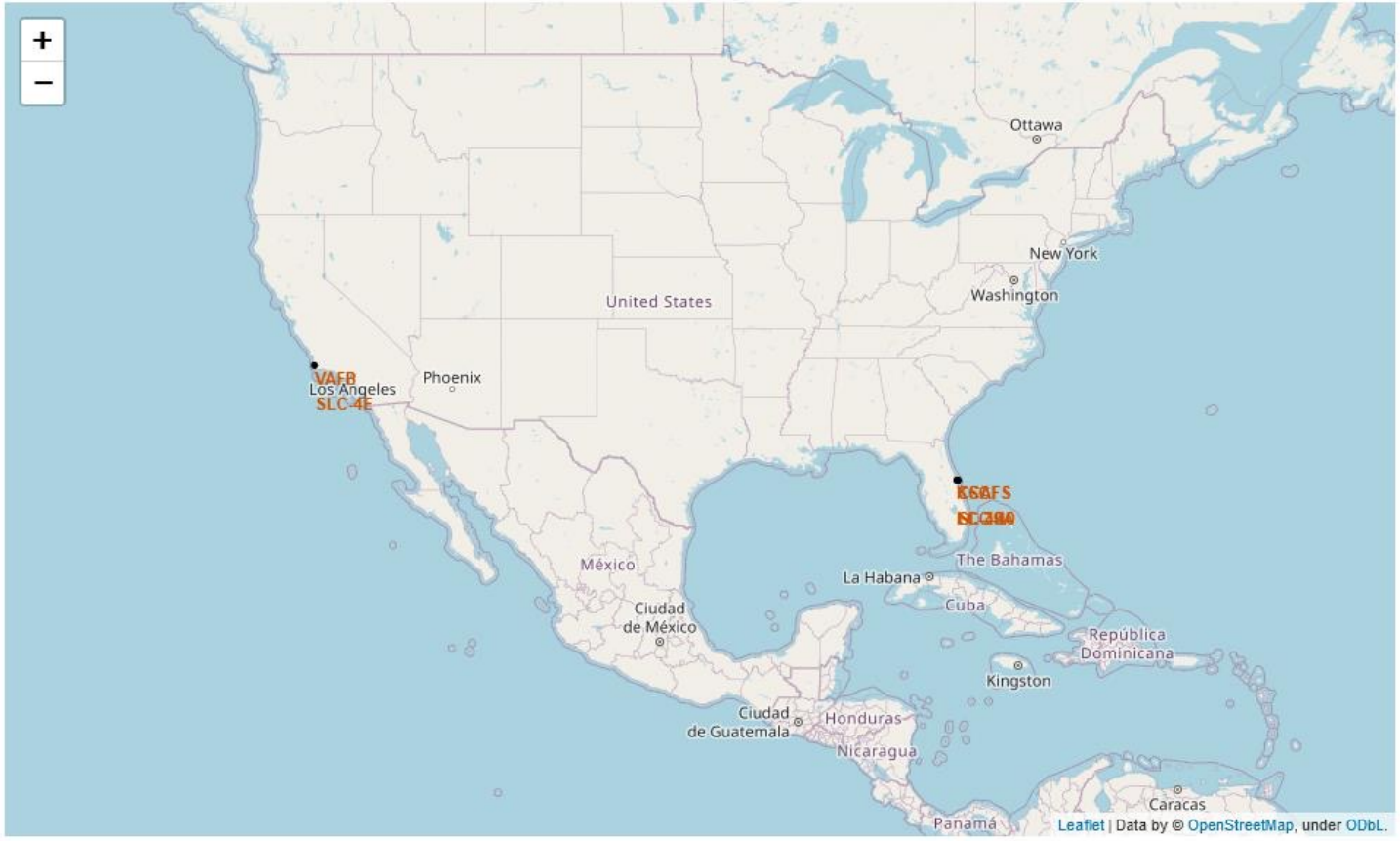
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order
- There is a significant landing_outcome for no attempt.

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


Section 3
Launch Sites
Proximities Analysis

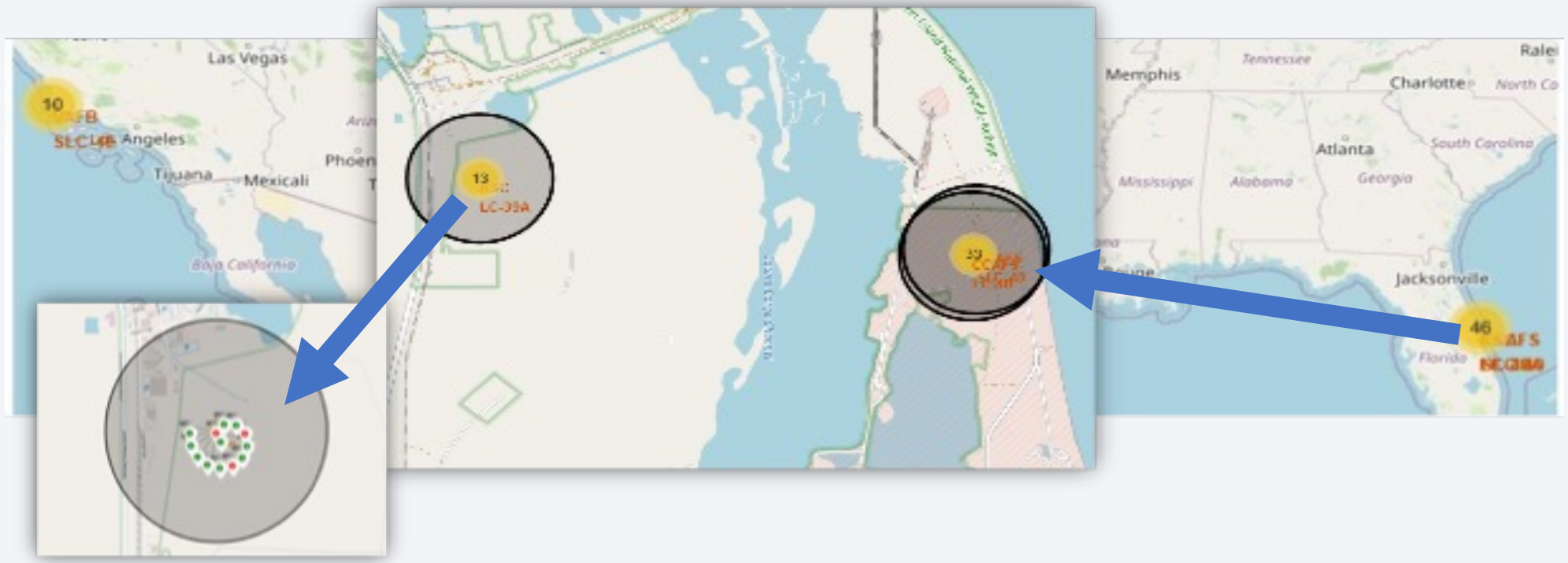
All launch Sites

- Launch Sites are near the sea for safety reasons, but not too far from access roads and railroads probably for access for all the equipment's.



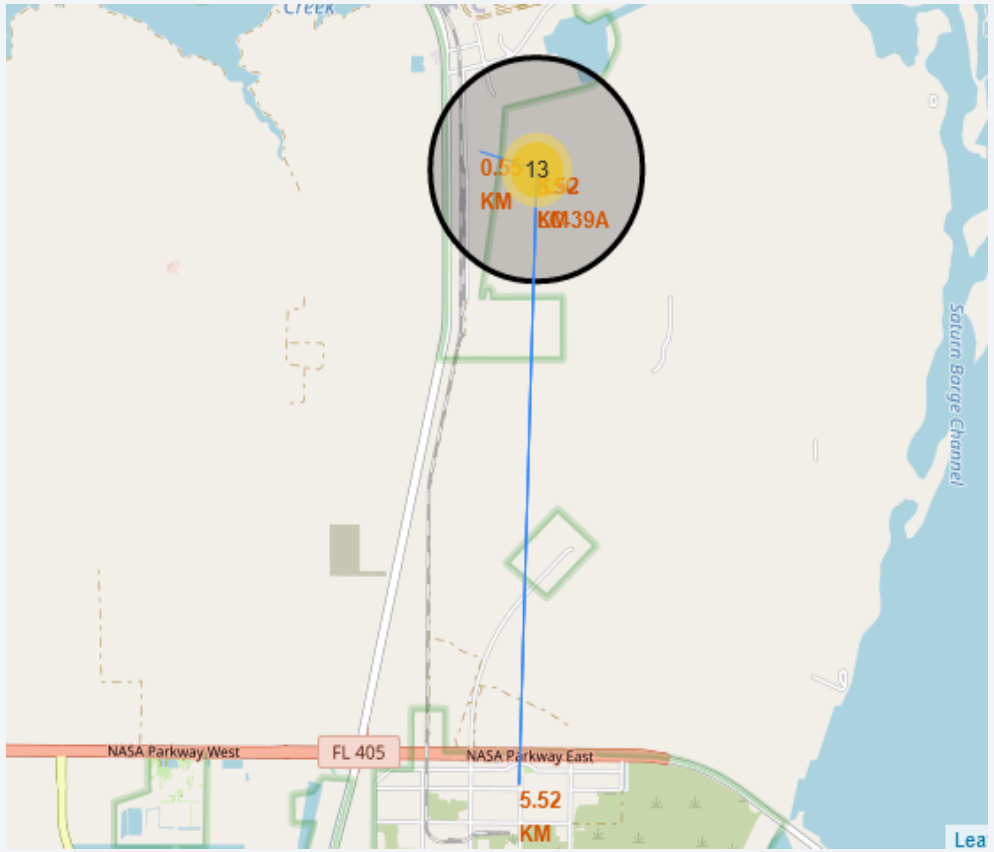
Launch Outcomes by site

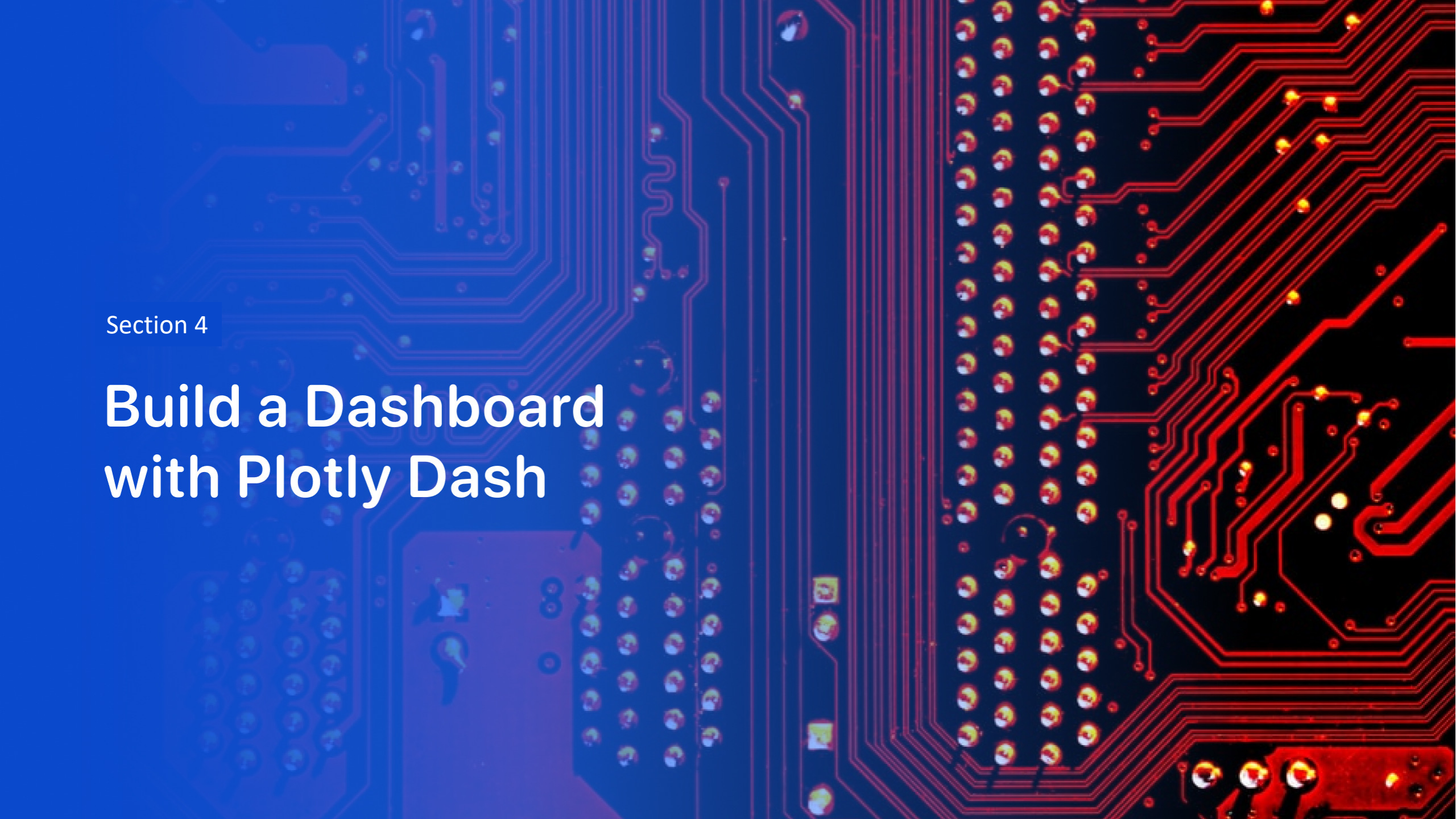
- Example of KSC lC-39A launch site launch outcomes
- Green markers indicate successful launch and red indicate unsuccessful launch



Logistics and Safety

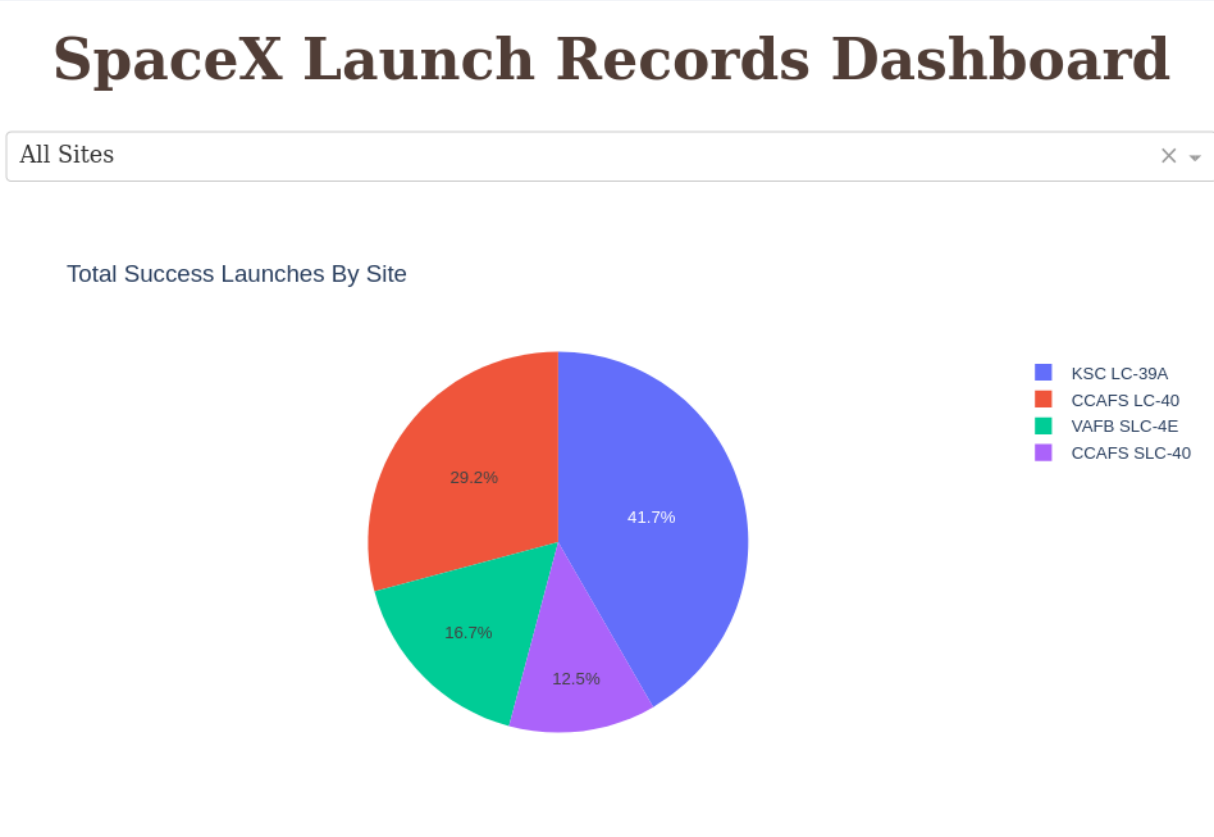
- Launch site KSC LC-39A has good logistics aspects, road and easy raildoad access provide much needed area for incoming shipment as well as near to the sea having easy access to tankers and shipment.
- due to its area being closed off to the public gives it a level of added safety from anyone braving to come in as they might get lost due to the sheer size of the place as well.




Section 4
Build a Dashboard with Plotly Dash

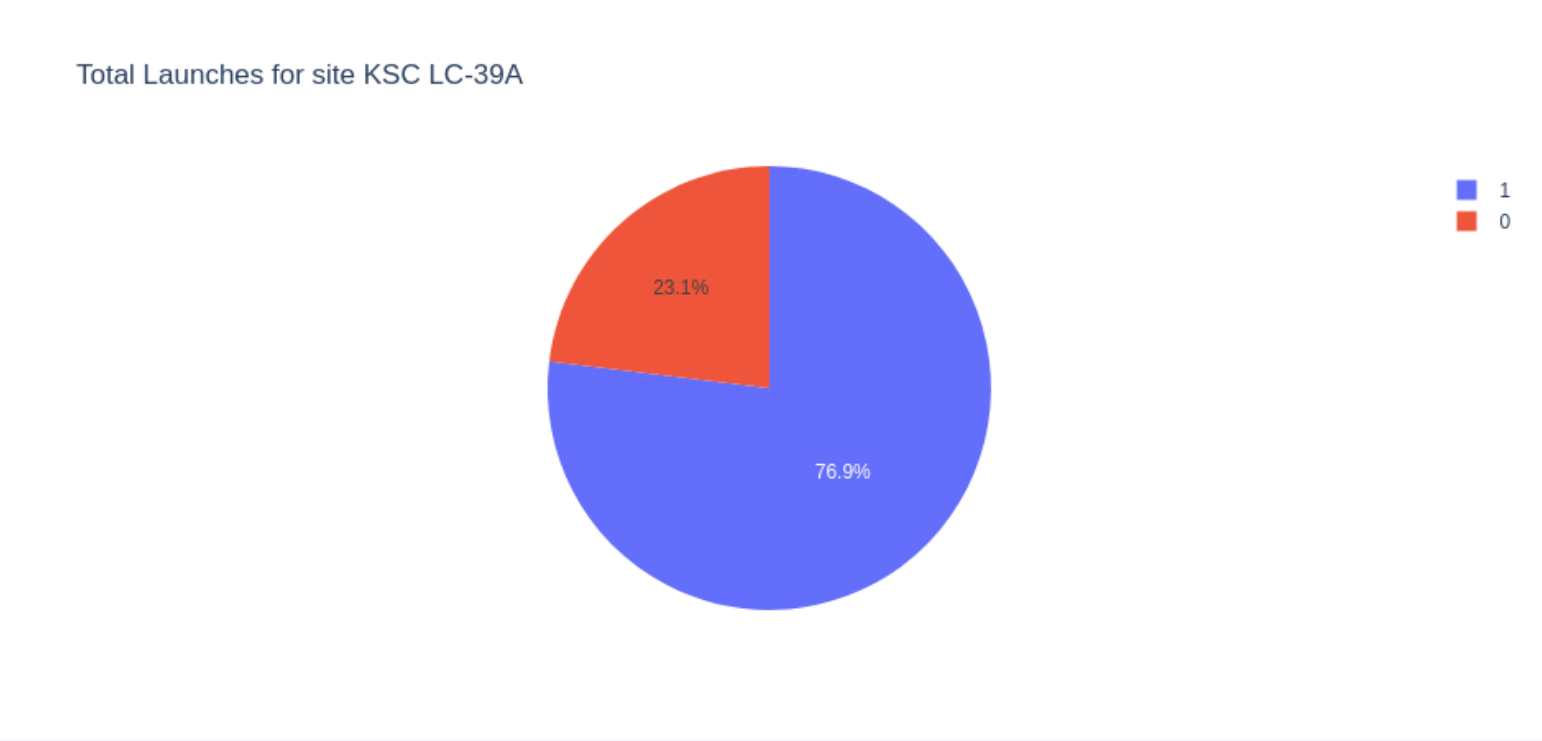
Successful Launches by site

- The location from where the launch takes place plays a very significant role in the success of the launch.



Launch Success Ratio for KSC LC-39A

- Success rate from KSC LC-39A is 76.9%

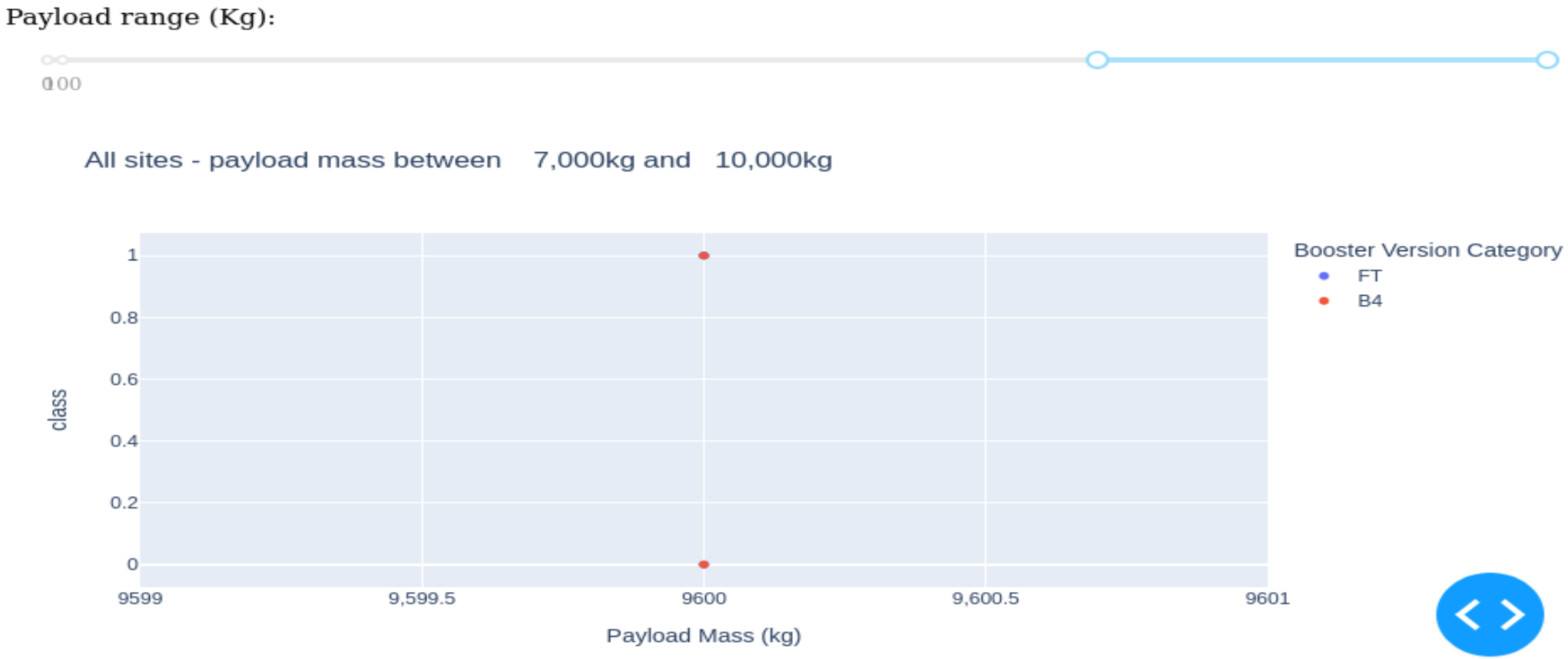


Payload (6000 KG) VS Launch Outcome

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

Payload (7000KG) vs Launch Outcome

- There is not enough data to estimate the risk of 7,000KG and above

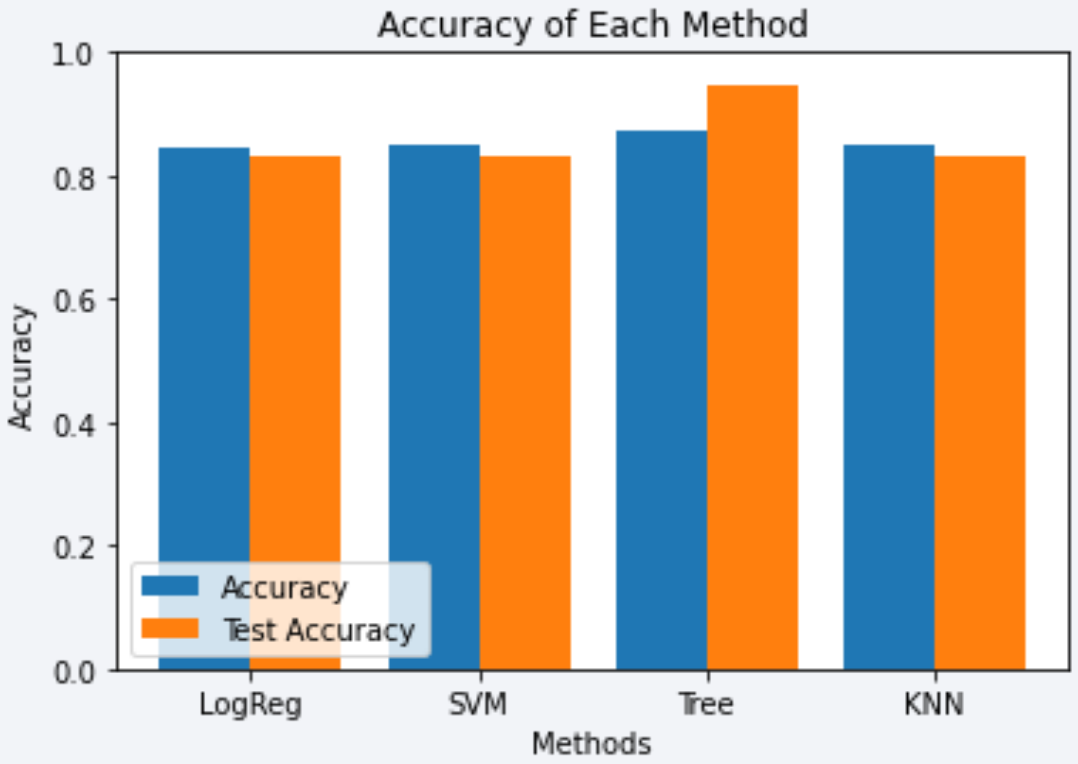


Section 5

Predictive Analysis (Classification)

Classification Accuracy

- There are 4 classification models that were tested you can see the result
- Decision tree has the highest classification accuracy with accuracy around 87%.



Confusion Matrix of Decision Tree

- Confusion matrix of decision tree shows its accuracy by proving the big numbers of true positive and true negative compared to the false positives and false negatives.

Conclusions

- Different data sources were analyzed from the Wiki to the API which helps hone the conclusions
- The best outcome for launches are with boosters for payloads 6,000 KG and below
- Decision tree classifier is the most accurate to predict successful landings
- Rockets and boosters are getting better as the launch outcome has steadily increase over the years
- The dip in successful missions could mean that trial and error is happening due to advancement in technology not so much in loosen safety regulation
- Safety in area are paramount and its better to build launch site near the ocean.

Appendix

- Improvement could be using more recent data and perhaps bigger amounts of data to help understand and predict successful launches
- Folium couldn't show maps
- Model test has to include np.random.seed variable.

Thank you!