IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Anastasia Kobyzeva
March 2022

Outline

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

Executive Summary

- Summary of methodologies:
  - Required data was collected with REST API and Web Scrapping. This data was formatted and transformed to a better form for further usage and understanding
  - Exploratory data analysis (EDA) with visualization and SQL and interactive visual analytics with Folium and Plotly Dash were performed
  - Classification models for predictive analysis were built
- Summary of all results:
  - Orbit types with the best success rate were defined
  - Launch sites proximity to infrastructure objects was evaluated
  - Launch site with the best success rate was identified
  - Best classification models were chosen and main problem in accuracy was identified

Introduction

- This project covers analysis of SpaceX launches to predict possible outcomes of new launches organized by other companies.
- Key problems are:
  - What factors have influence on outcome of the launch?
  - Do some factors influence more than others?
  - What predictive model could be used?

Section 1

Methodology

Methodology

Executive Summary

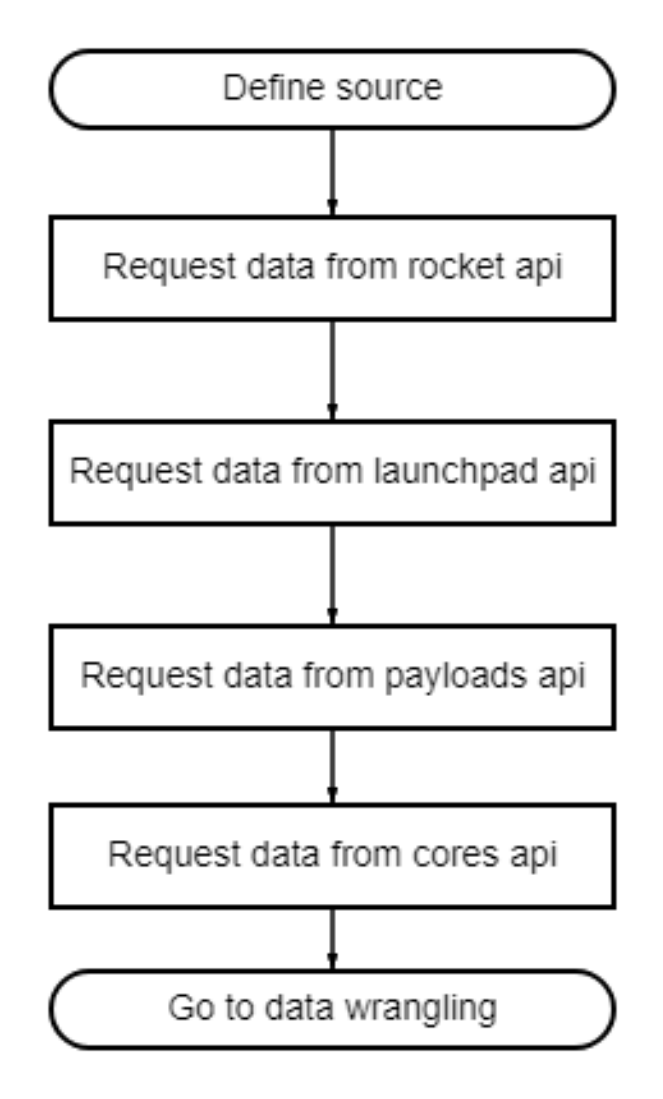
- Data collection methodology:
  - API requests and web-scrapping
- Perform data wrangling
  - Adding new columns based on delivered data for better understanding, dropping null values or replacing it with mean
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building models such as KNN, decision tree, SVM and logistic regression using scikit-learn library, evaluating accuracy of models based on score and confusion matrices.

Data Collection

- To collect data two methods were used:
  - Web scrapping
  - REST API
- With REST API *spacexdata.com* source was used. For web scrapping *wikipedia.org* source was used.
- As a result of Data Collection section we've got two dataframes with useful information for further processing.

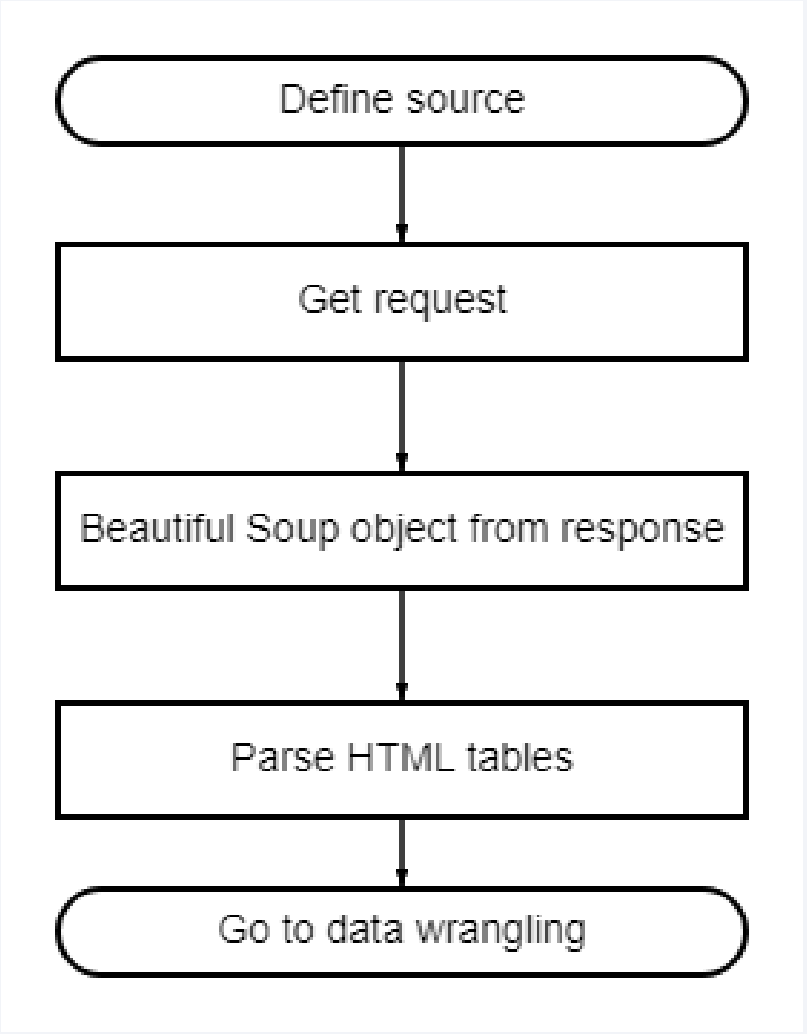
Data Collection – SpaceX API

- For Data Collecting api.spacexdata.com site was used. Requests were made in succession to rocket, launchpad, payloads and cores sections. Further response was processed and formed pandas dataframe.
- GitHub URL: https://github.com/knopin2009/DS_project/blob/master/Data%20Collection%201.ipynb



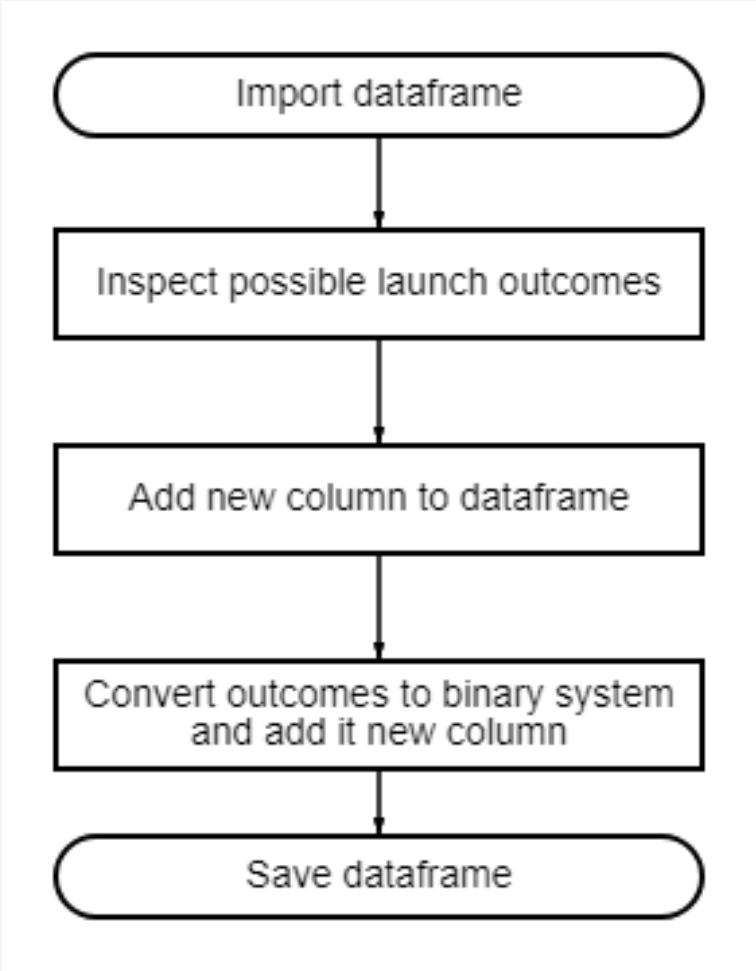
Data Collection - Scraping

- For Data Collection wikipedia.org was used. Get request was made to get data from Wikipedia page. To parse it beautiful soup library was used. Data from parsed tables was recorded to python dictionary and then pandas dataframe was formed.
- GitHub URL: https://github.com/knopin2009/DS_project/blob/master/Data%20Collection%202.ipynb



Data Wrangling

- The main idea of data processing was to simplify detection of successful outcomes. Different types of outcomes were transformed to two possible variables in a new column ‘Class’:
  - 0 for failure
  - 1 for success
- GitHub URL: https://github.com/knopin2009/DS_project/blob/master/Data%20Wrangling.ipynb



EDA with Data Visualization

- Charts that were plotted:
  - Flight number vs. Payload mass (to see if success depends on payload mass)
  - Flight number vs Launch site (to see if success depends on launch site)
  - Success rate of different orbit types (to see the best performing orbits)
  - Flight number vs Orbit type (to see how preferences in choosing orbits were changing in time and if changes were successful)
  - Launch success yearly trend (to see dynamics of successful outcomes)
- GitHub URL: https://github.com/knopin2009/DS_project/blob/master/EDA%20with%20Visualization.ipynb

EDA with SQL

- SQL queries that were performed:
    - Display the names of the unique launch sites
    - Display the total payload mass carried by boosters launched by NASA (CRS)
    - Display average payload mass carried by booster version F9 v1.1
    - Display the date when the first successful landing outcome in ground pad was achieved
    - List the total number of successful and failure mission outcomes
    - List the names of the booster versions which have carried the maximum payload mass
    - List the failed landing outcomes in drone ship
    - Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20
- GitHub URL: https://github.com/knopin2009/DS_project/blob/master/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- On the map circles and markers were created to mark the launch sites locations
- Marker clusters were created to mark the success/failed launches for each site on the map
- Lines and distance markers were added to show proximity of the launch sites to coastline, highway, railway and nearest city.
- GitHub URL: https://github.com/knopin2009/DS_project/blob/master/Visualization%20with%20Folium.ipynb

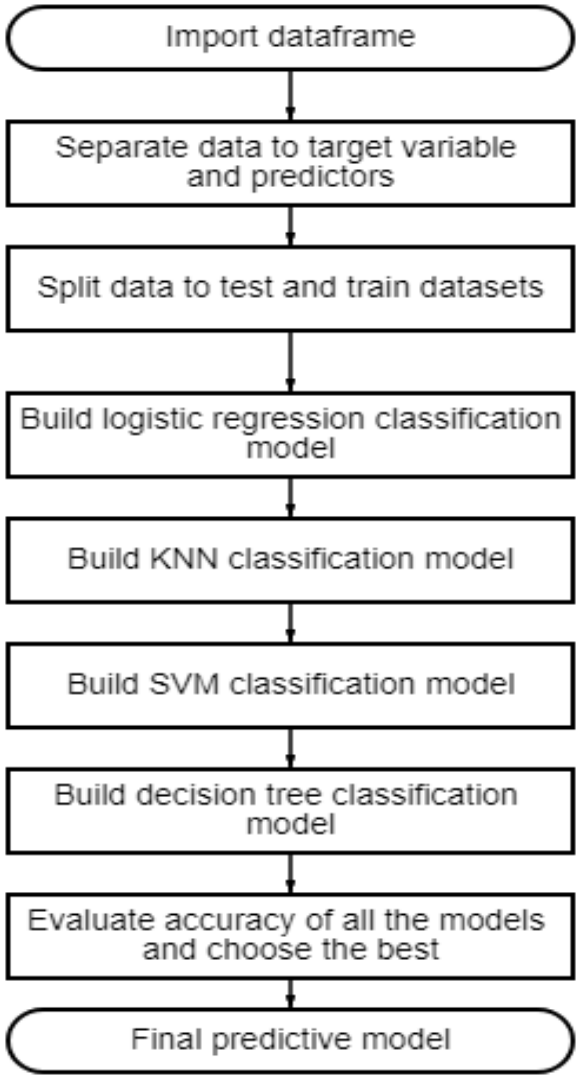
Build a Dashboard with Plotly Dash

- Created plots:
  - A pie chart to summarize the success rate of launch site was created
  - Payload vs. class scatter plot was created to see how payload mass and booster version influence on launch outcome.
- Created interactions:
  - Launch site drop-down input component to change available launch sites
  - Range slider to select payload

- GitHub URL: https://github.com/knopin2009/DS_project/tree/master/Dash

Predictive Analysis (Classification)

- To build the classification model python scikit-learn library was used. Data was separated into train and test datasets, variables were labeled as predictive or target. With this data four types of models were built: logistic regression, KNN, SVM and decision tree.
- To chose the best parameters for each model Grid Search was used.
- The best model was chosen according to best accuracy score and resulting confusion matrix.
- GitHub URL: https://github.com/knopin2009/DS_project/blob/master/ML%20Prediction.ipynb

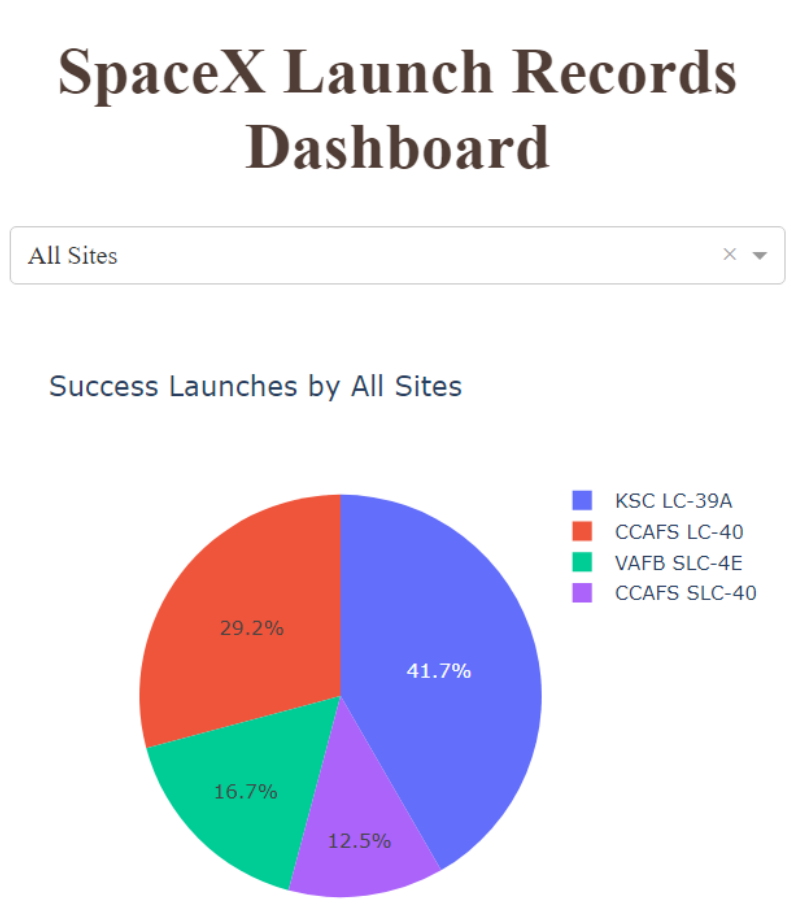


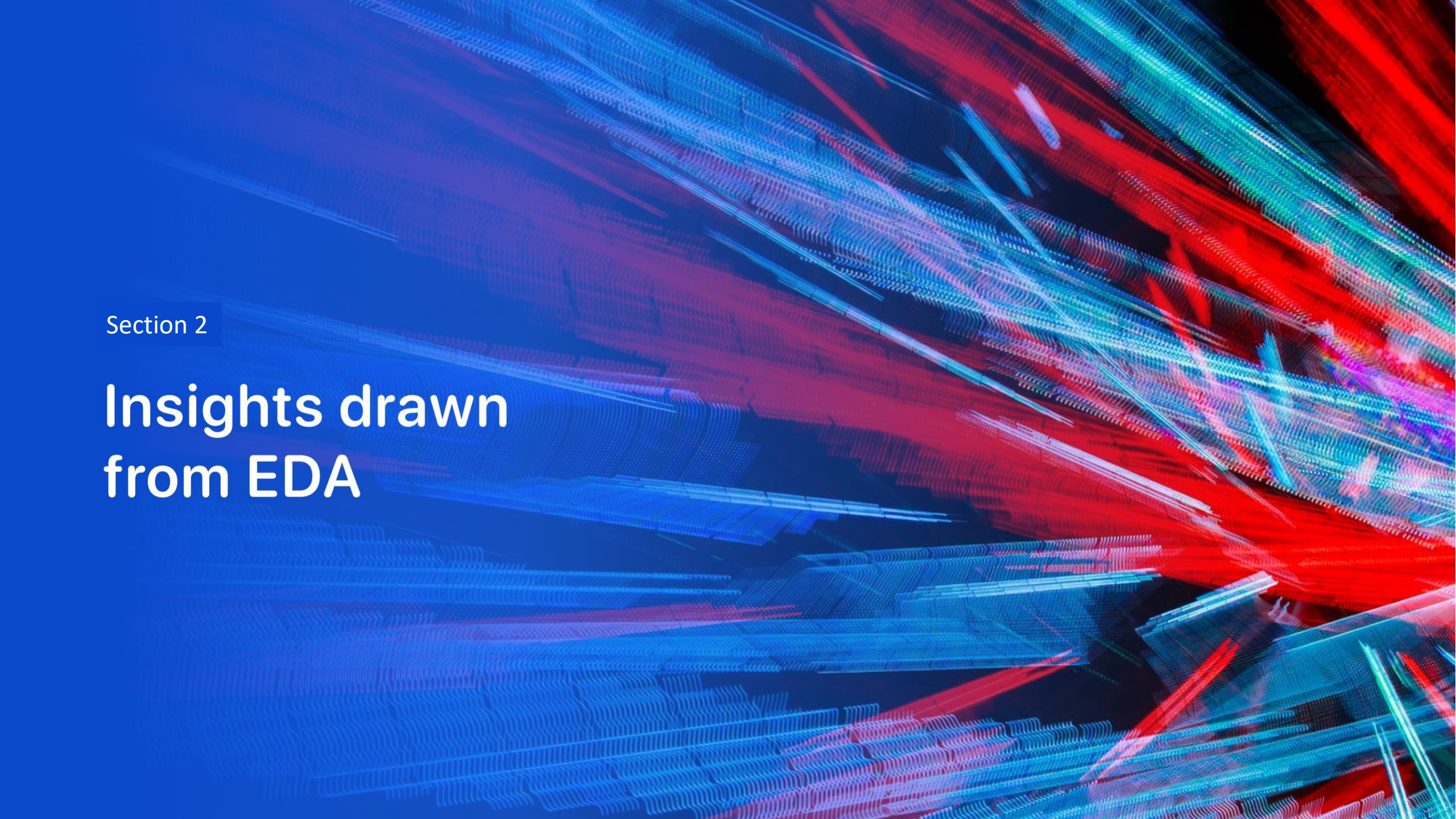
Results

- Exploratory data analysis results:
  - Charts visualizing different relationships between variables were created and gave basic understanding of what parameters influence on launch outcome
  - SQL queries allowed to look at data in a more directed way, look at maximal and average values of payload mass, study the data in connection with date
- Predictive analysis results:
  - Four types of ML models were built and assessed
  - For each model confusion matrix was built and accuracy was calculated

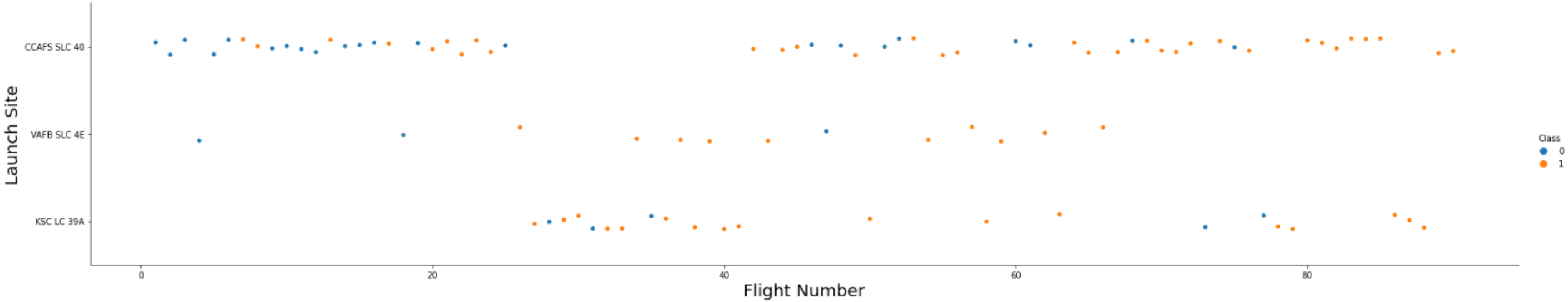
Results

- Interactive analytics demo in screenshots



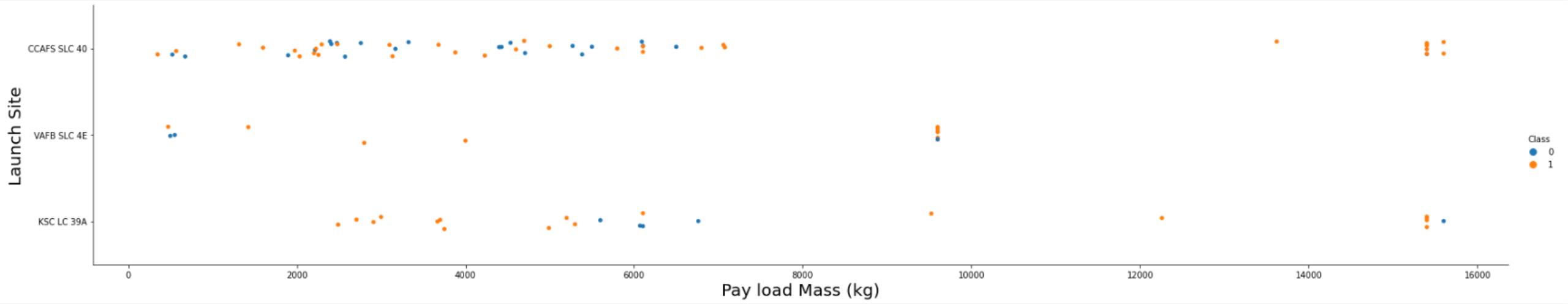

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



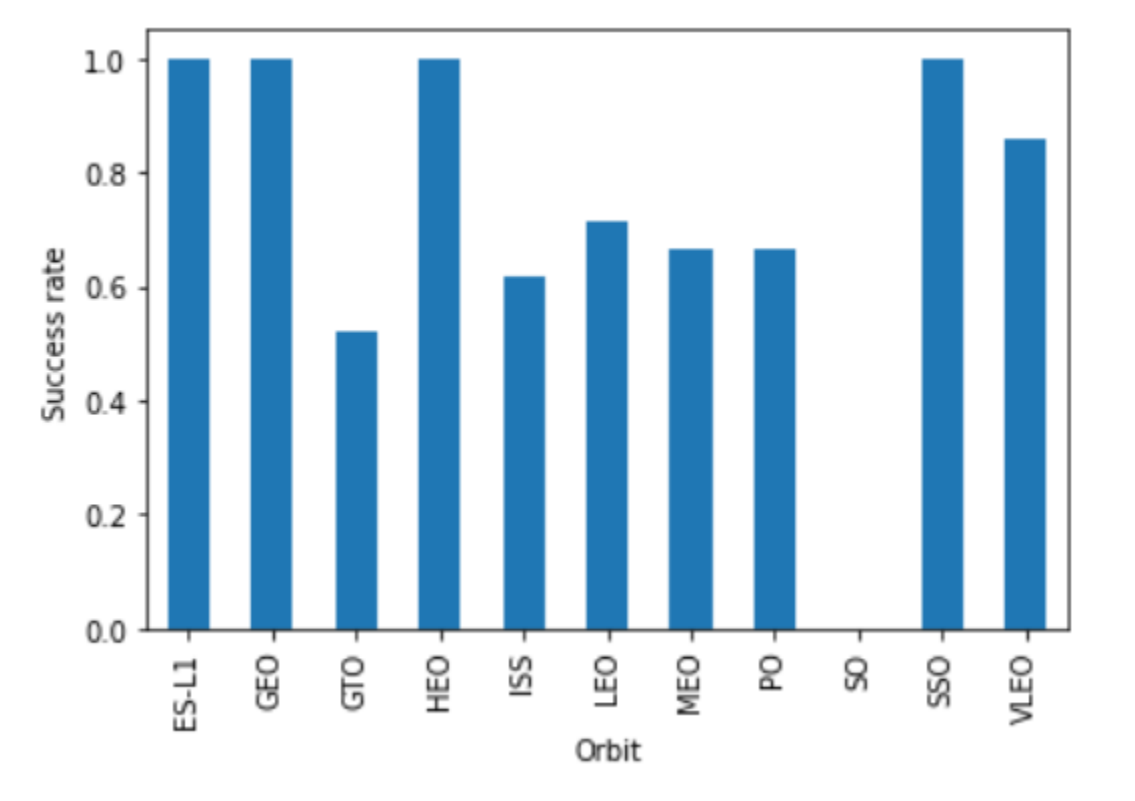
- On this plot we can see that CCAFS SLC 40 launch site has the largest number of launches and first launches were mostly unsuccessful

Payload vs. Launch Site



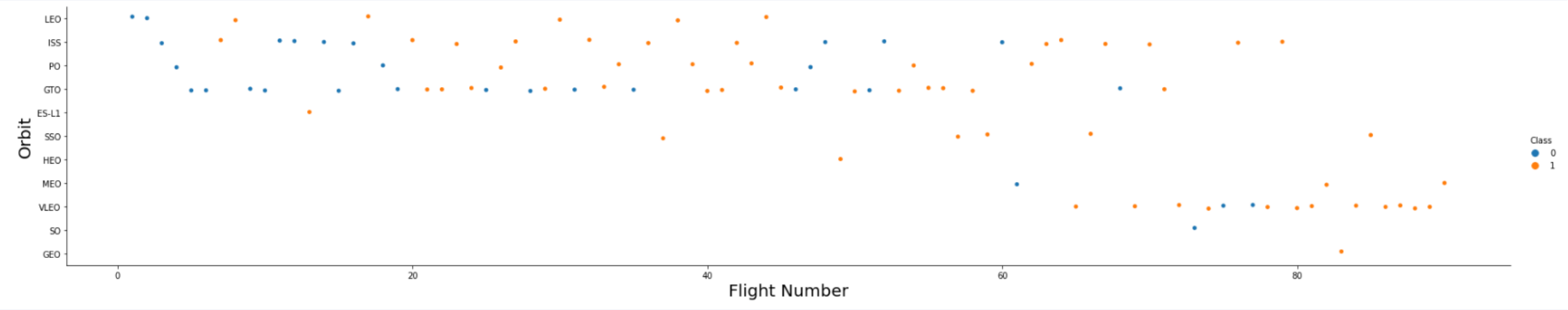
- For the VAFB-SLC launch site there are no rockets launched for heavy payload mass (greater than 10000)

Success Rate vs. Orbit Type



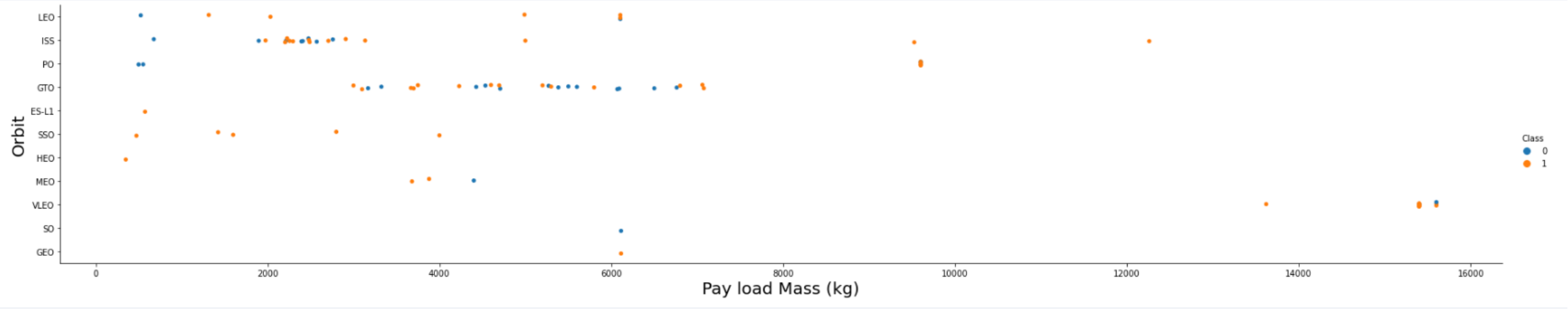
- Here we can see that ES-L1, GEO, HEO and SSO orbits have the highest success rate while SO has the lowest

Flight Number vs. Orbit Type



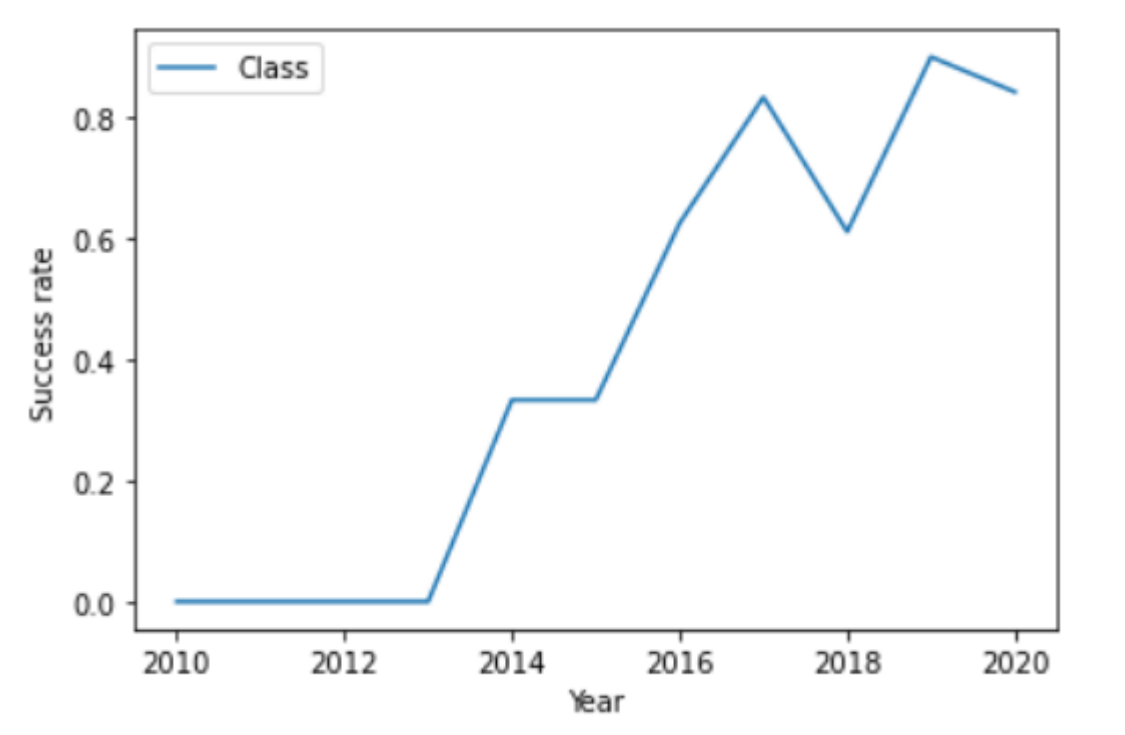
- In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit

Payload vs. Orbit Type



- 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.
- SSO has perfect positive landing rate but used only with payload up to 4000 kg

Launch Success Yearly Trend



- As we can see, success rate keeps increasing from 2013 to 2020

All Launch Site Names

- There are four unique launch site names in dataset:

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`
- We can see that in all this records orbit type is LEO

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

- As sum of all relevant payloads we can find that total payload carried by boosters from NASA is 45596 kg
- Query result:

Total_Payload_NASA
45596

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1 is 2928 kg
- Query result:

AVG_Payload_v1.1
2928

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad is December, 22 of 2015
- Query result:

First_Success_date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- There are four boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- Query result:

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

Total Number of Successful and Failure Mission Outcomes

- There are 61 Successful and10 Failure mission outcomes in total
- Query result:

sucsess
61

failure
10

Boosters Carried Maximum Payload

- There are 12 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

2015 Launch Records

- There are two failed landing outcomes in drone ship in year 2015
- Query result:

DATE	booster_version	launch_site	landing__outcome
2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- The most frequent landing outcomes between the date 2010-06-04 and 2017-03-20 are “No attempt” and “Failure (drone ship)”
- Query result:

landing__outcome	rate
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

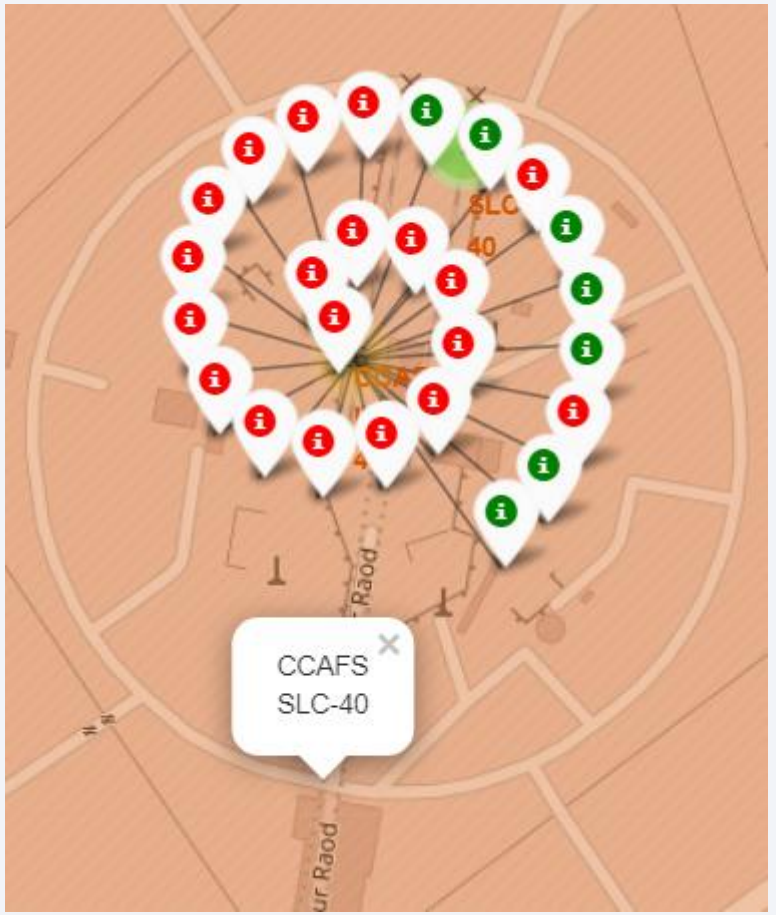
Launch Sites Proximities Analysis

Location of Launch Sites

- Locations of launch sites are marked with orange circles. As we can see all of them are near the coastline

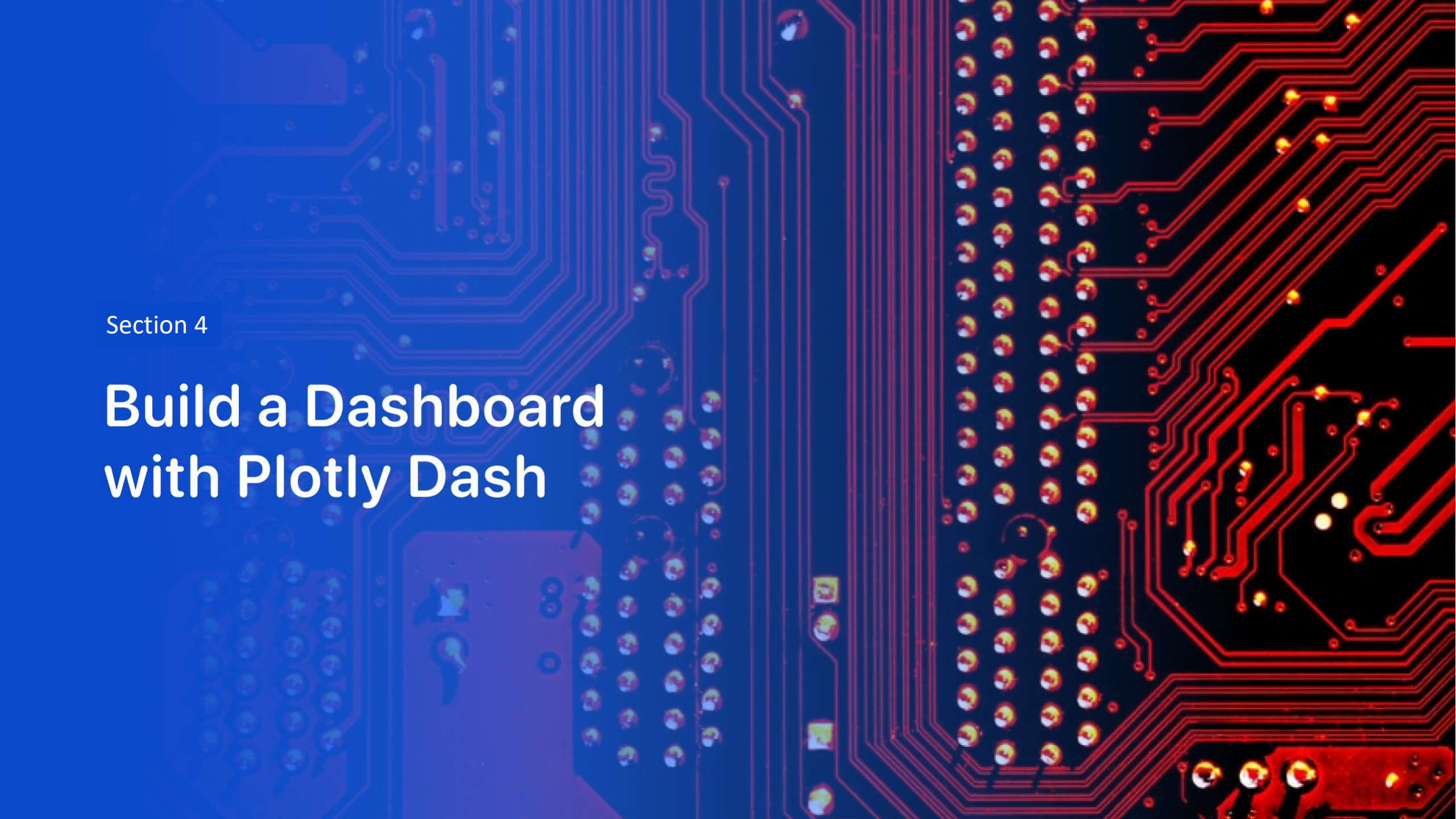
Launch Outcome Labels

- Green and red markers show if launch outcomes were successful or failure respectively for specific launch site
- On the screenshot we can see that CCAFS SLC-40 launch site mostly has negative outcomes

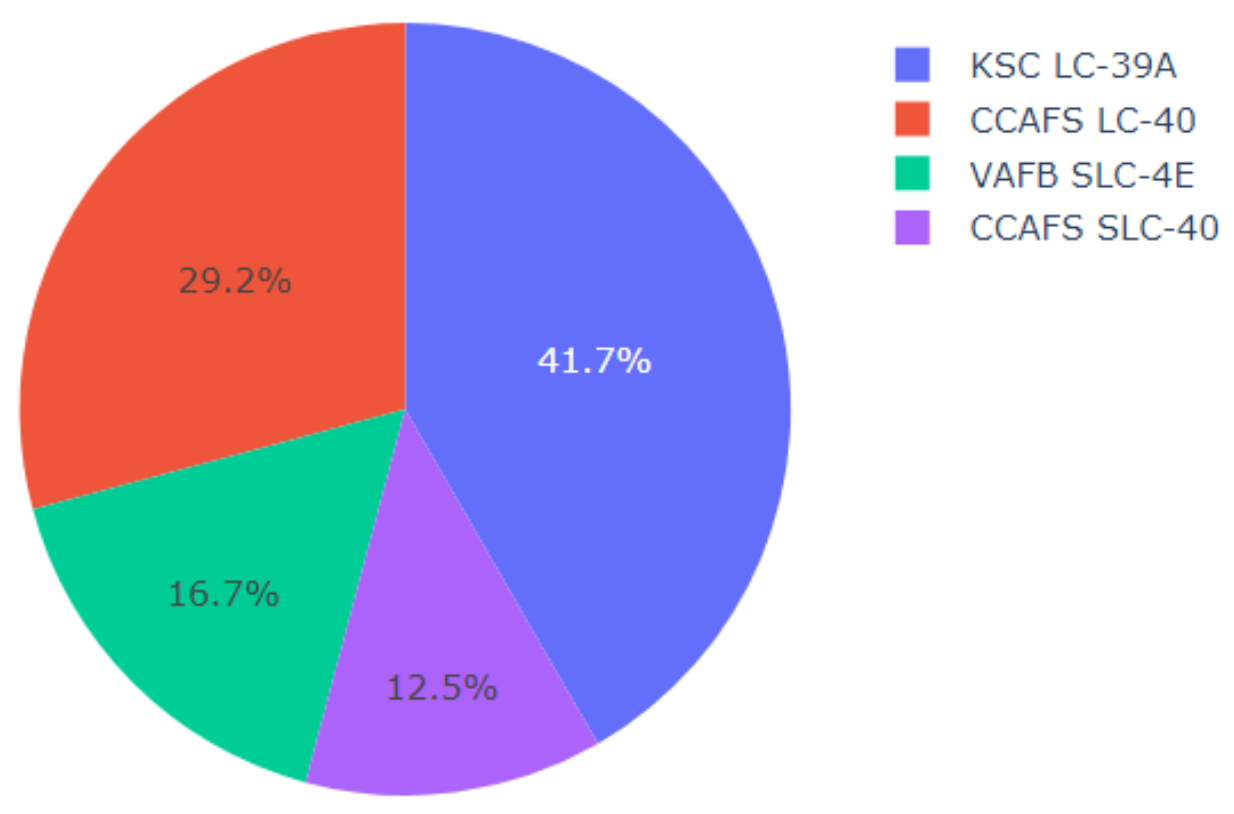


Proximity to Infrastructure

- On the lower screenshot we can see that railway, highway and coastline are closer than 1 km to selected launch site
- On the right screenshot we can see that the nearest city is in 17.59 km i.e. not close to the launch site

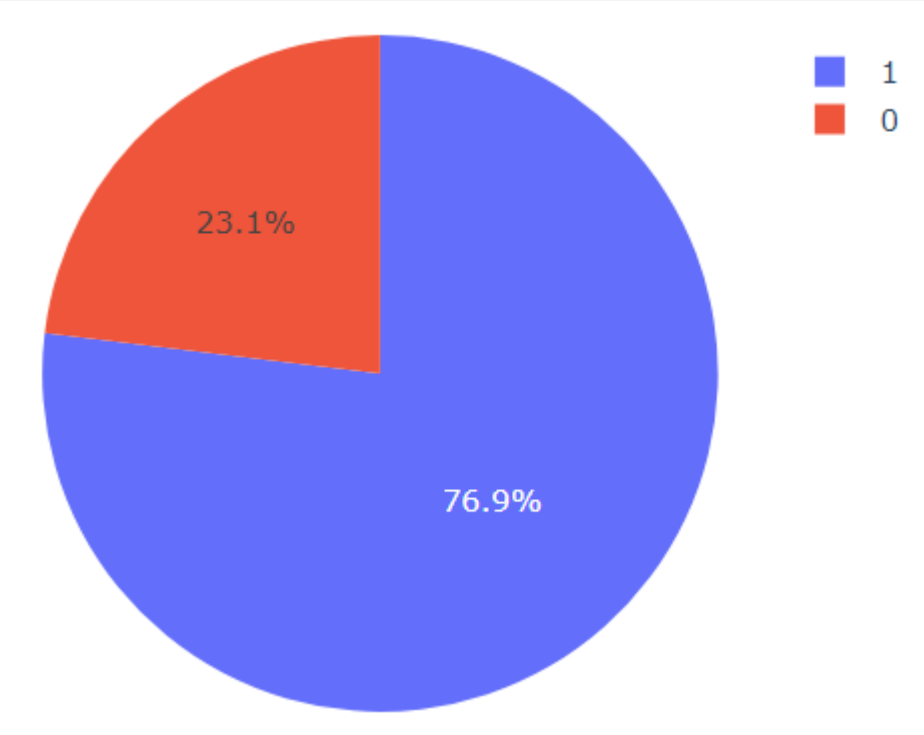

Section 4
Build a Dashboard with Plotly Dash

Success Launches by All Sites



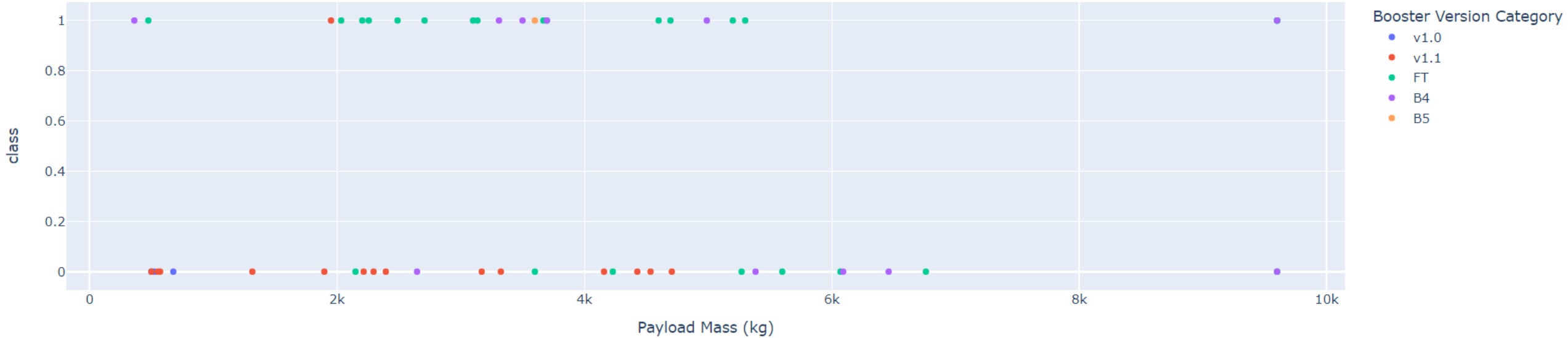
- On this pie chart we can see that the site with the largest amount of successful outcomes is KSC LC-39A it host 41.7% of all successful outcomes
- CCAFS LC-40 is on the second place with 29,2%

Highest Launch Success Ratio



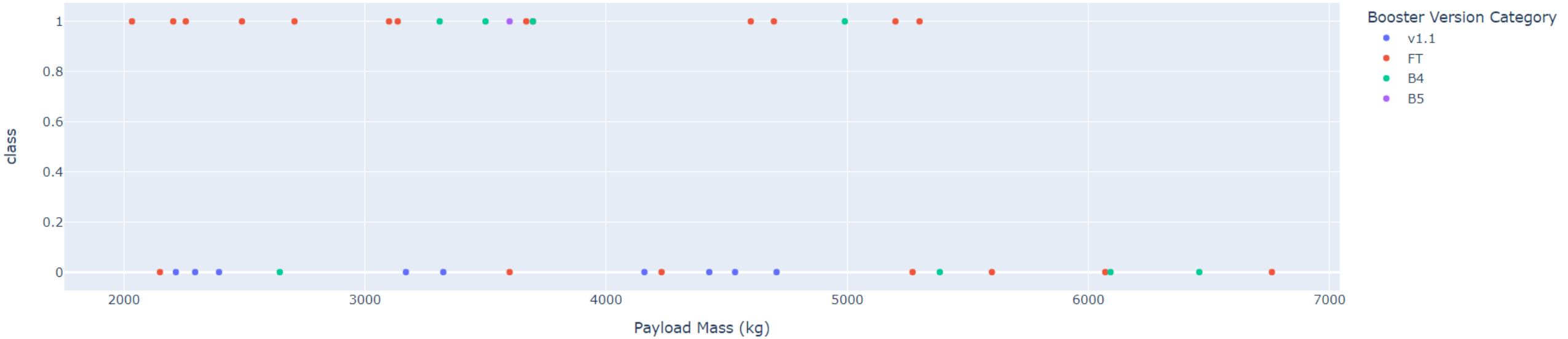
- The highest launch success ratio has KSC LC-39A launch site
- On the chart we can see that it has 76.9% of success launches

Payload vs. Launch Outcome



- From the scatter plot we can see that launches where payload mass is less than 1000 kg and booster version is v1.1 or v1.0 mostly have negative outcome

Payload vs. Launch Outcome

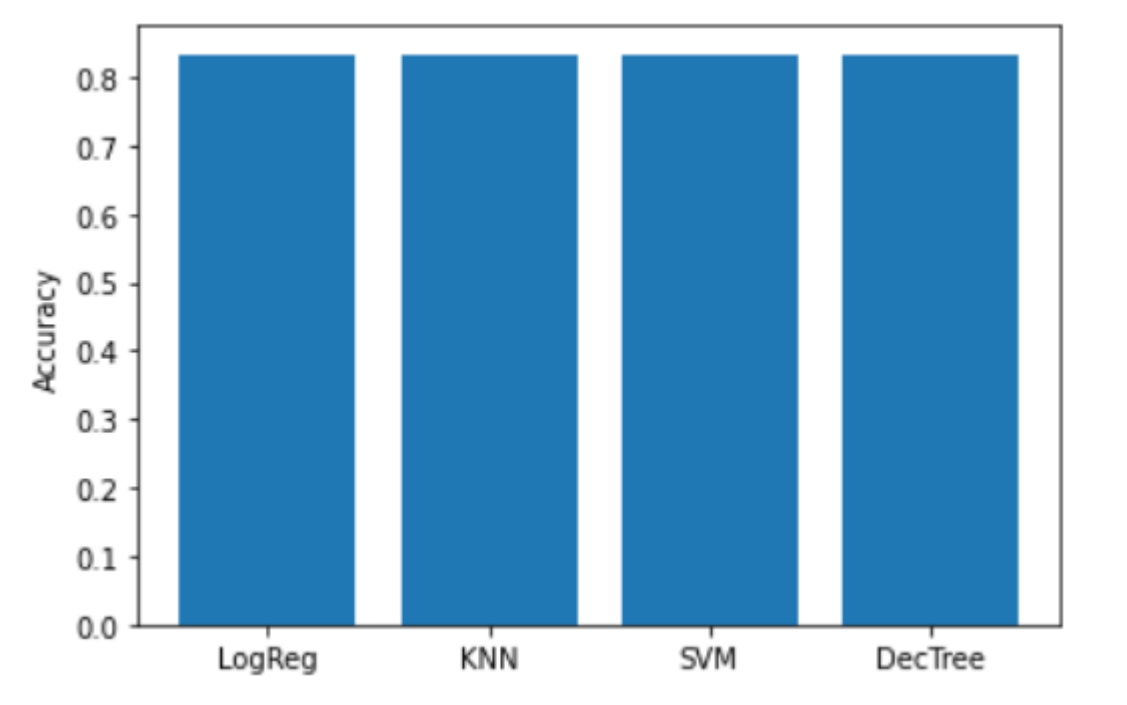


- In range of payload mass between 2000 kg and 5000kg booster version v1.1 is the most frequently failed while FT has the largest number of successful outcomes

Section 5

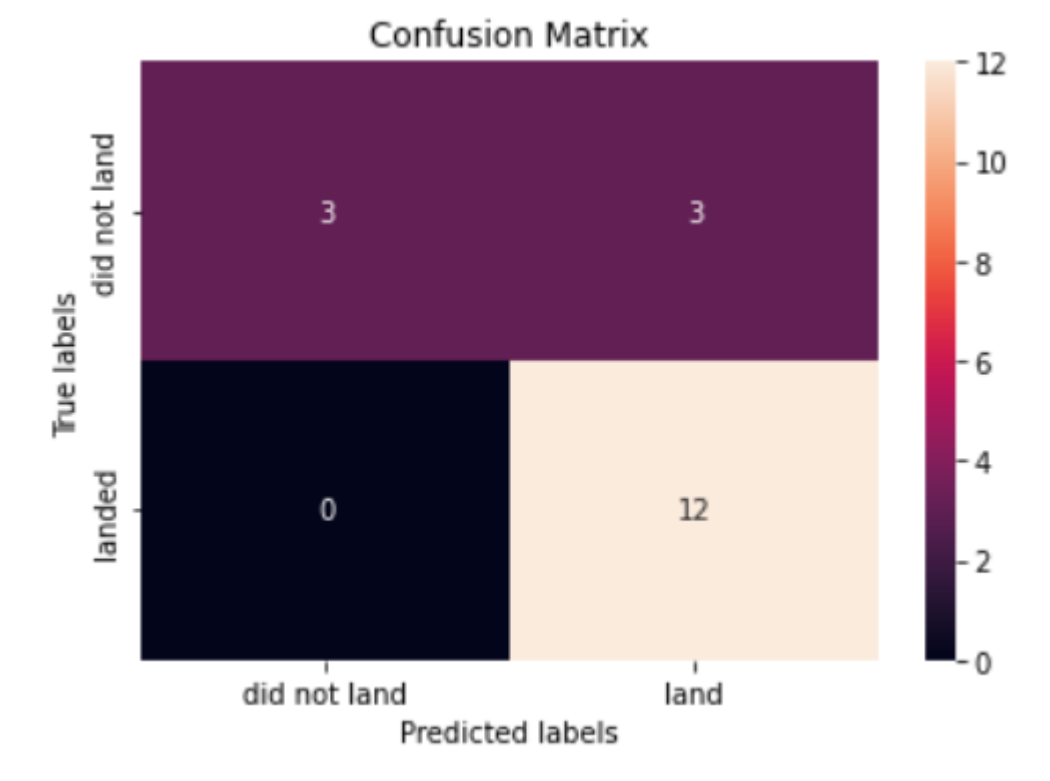
Predictive Analysis (Classification)

Classification Accuracy



- As we can see from the bar chart Logistic Regression, KNN, SVM and Decision Tree models have equal accuracy of 83.3%, so any of them could be used for prediction

Confusion Matrix



- Looking at confusion matrix we can that the biggest problem is false positives
- Successful landings are predicted right but in half of the failed landing cases prediction is wrong

Conclusions

- Launch success rate keeps increasing from 2013
- ES-L1, GEO, HEO and SSO orbits have the highest success rate
- With heavy payloads the successful landing rate are more for Polar, LEO and ISS
- Launch sites are usually located near coastline, railway and highway but far enough from cities
- FT booster is the best with payload up to 5000 kg
- The highest launch success ratio has KSC LC-39A launch site
- In predictive model any type of classifier could be used with accuracy 83.3%
- The problem of created classification models is false positives

Appendix

- First rows of data collected with web scrapping

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10

Appendix

- First rows of data collected with REST API

	Flight Number	Date	Booster Version	Payload Mass	Orbit	Launch Site	Outcome	Flights	GridFins	Reused	Legs	Landing Pad	Block	Reused Count	Serial	Longitude	Latitude
0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003	-80.577366	28.561857
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005	-80.577366	28.561857
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007	-80.577366	28.561857
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.0	0	B1003	-120.610829	34.632093
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1004	-80.577366	28.561857

Thank you!