IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

QUOC-ANH NGUYEN
2021-11-22

Outline

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

Executive Summary

- Summary of methodologies
  - Collect data from SpaceX API and Wikipedia
  - Verify, clean and transform data
  - Exploratory data analysis using SQL and visualization
  - Perform predictive analysis using classification models: KNN, Logistic Regression, SVM and Decision Tree
- Summary of all results
  - Launch sites are located near coastline, highways and railroads, far away from cities
  - KSC LC-39A is the launch site have the most successful launches and highest successful launch rate
  - FT booster version has the highest successful rate
  - Best model can predict the outcome with 83% accuracy

Introduction

Objective:

- Using the Space X launch data to get an insight for the Space Y project
- Information that helps Space Y increase the first-stage reuse rate

Question to be answer:

- Which factor help increasing the reuse rate?
- What is the chance of reusing the first stage of a rocket before it launch?

Section 1

Methodology

Methodology

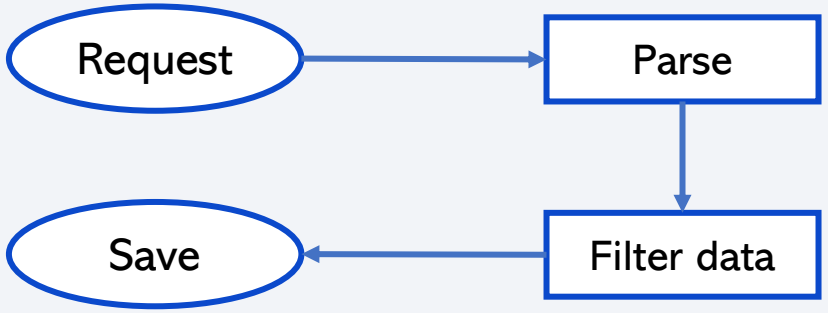
Executive Summary

- Data collection methodology:
  - SpaceX open data API or Web scraping from Wikipedia
- Perform data wrangling
  - Verify and clean data
  - Standardize outcome to value 0 or 1
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models:
  - Using Grid Search and Cross Validation to finetune the models (KNN, Logistic Regression, SVM, Decision Tree)

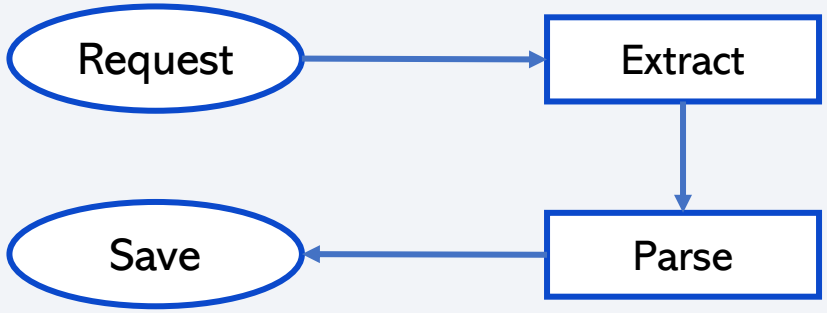
Data Collection

- Dataset is collected using two methods:
  - Using SpaceX API: `https://api.spacexdata.com/v4/launches/past`
  - Web scraping: `https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches`

SpaceX API



Web scraping

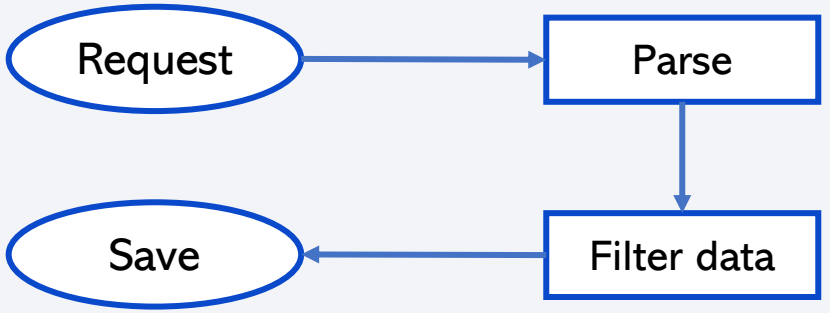


Data Collection – SpaceX API

- Request the SpaceX launch data using the GET request
- Parse the response to create a data table
- Filter the data to only include Falcon 9 launches
- Save the collected data

- Reference file on GitHub

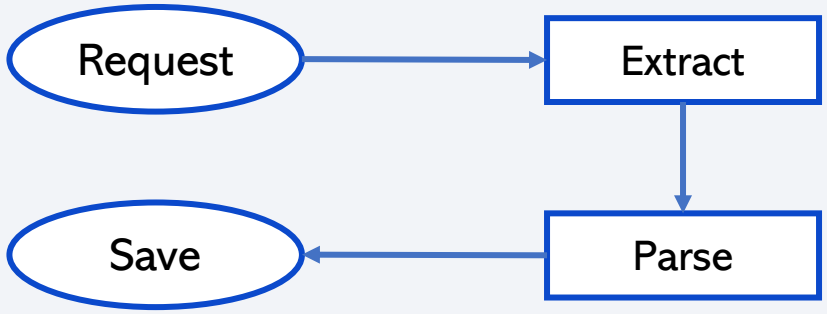
Collect data from SpaceX API



Data Collection - Scraping

- Request the Falcon 9 Launch Wiki page
- Extract the required table from the response
- Parse the HTML table to a data table
- Save the collected data

Collect data using web scraping

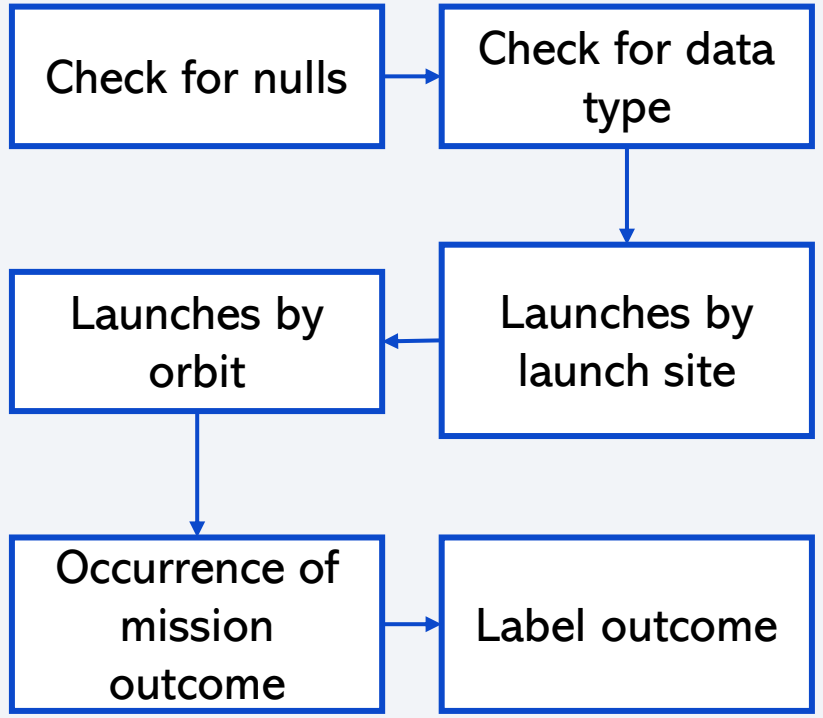


- Reference file on GitHub

Data Wrangling

- Check for null values and data types
- Calculate the number of launches on each launch site and number of each orbit
- Calculate the occurrence of mission outcome
- Create a landing outcome label

- Reference file on GitHub



EDA with Data Visualization

- Scatter chart: visualize the output distribution showing how the variables affect the outcome.
  - Concerned variables: Payload mass, Flight number, Launch site, Orbit type
- Line chart: showing how the success rate change over time.
- Bar chart: visualize the success rate differences between Orbit types

- Reference file on GitHub

EDA with SQL

- List all launch sites
- Display 5 records where Launch site begin with ‘CCA’
- Display total mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- All the boosters which have success in drone ship and payload between 4000 kg and 6000 kg
- The date when the first successful landing in ground pad achieved
- Total number of successful and failure mission outcomes
- The boosters which have carried the maximum payload mass
- Date. launch site, booster of the failed landing in drone ship
- Rank the count of landing outcome between the date 2010-06-04 and 2017-03-20

Reference file on GitHub

Build an Interactive Map with Folium

- Circles and launch site name as marker added to show where all the launch sites are located
- A marker cluster of every launch showing how many launches (successful and failed) at each launch site.
- Lines showing the distances from a launch site to coastline, railroad, highway and city

- Reference file on GitHub

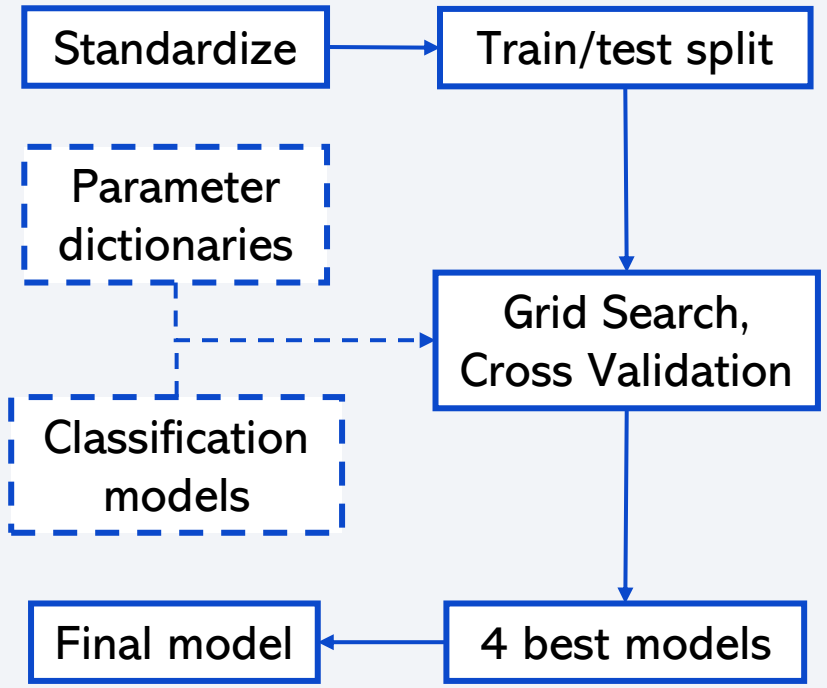
Build a Dashboard with Plotly Dash

- A dropdown menu to select launch site for visualization
- A pie chart showing success launch rate for launch sites to see the highest percentage of outcome compares to the others
- A slider to select the range of payload mass
- A scatter plot showing the distribution of mission outcome by payload mass and booster version.

- Reference file on GitHub

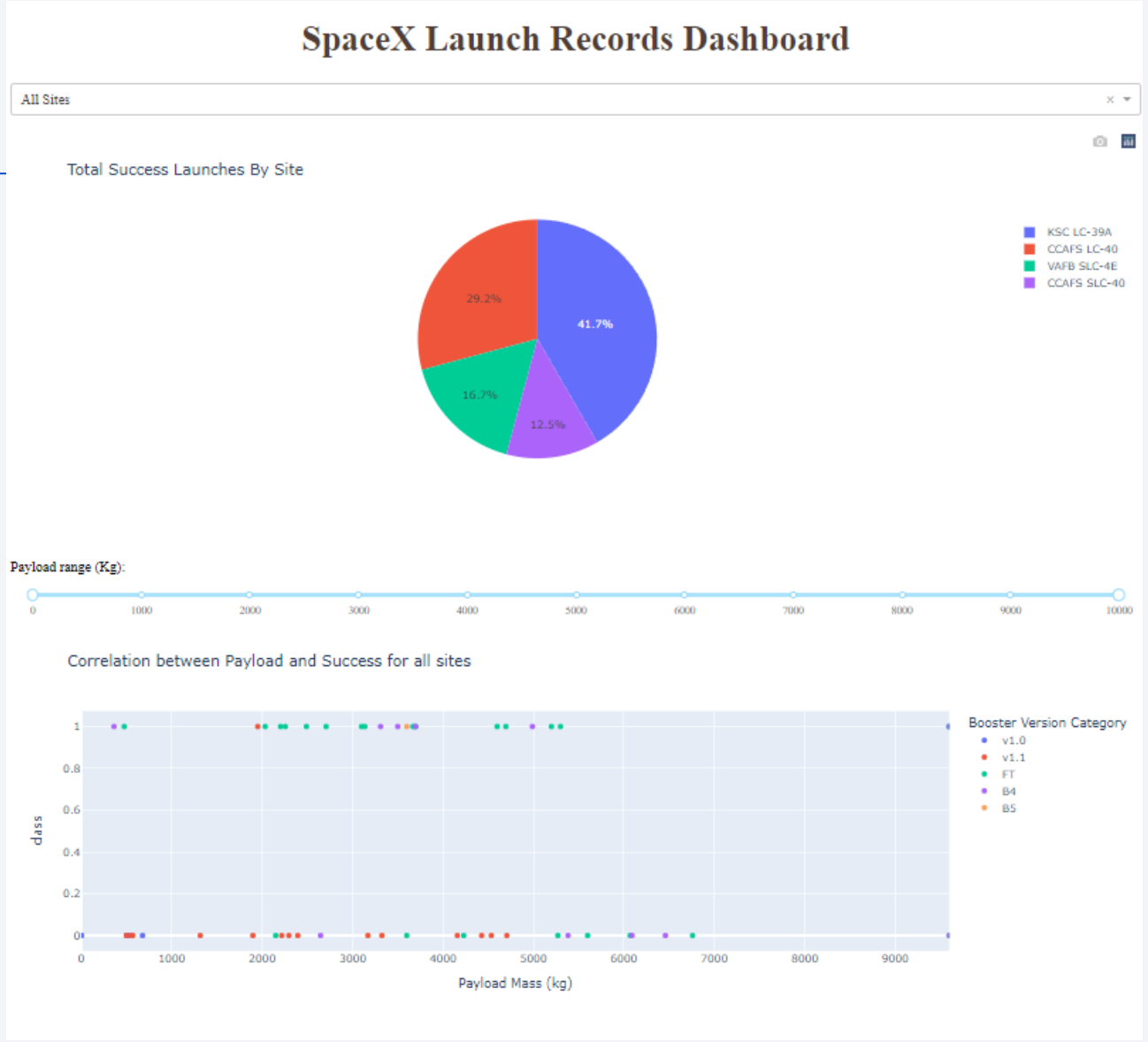
Predictive Analysis (Classification)

- Rescale the input using standardize scaler
- Split train and test set with the ratio 8:2
- Using grid search for each classification models with difference parameter to find the best version of each model.
- Cross validation is used to validate each models
- 4 best models will be compared using test set, the model with the best accuracy will be selected
- Reference file on GitHub



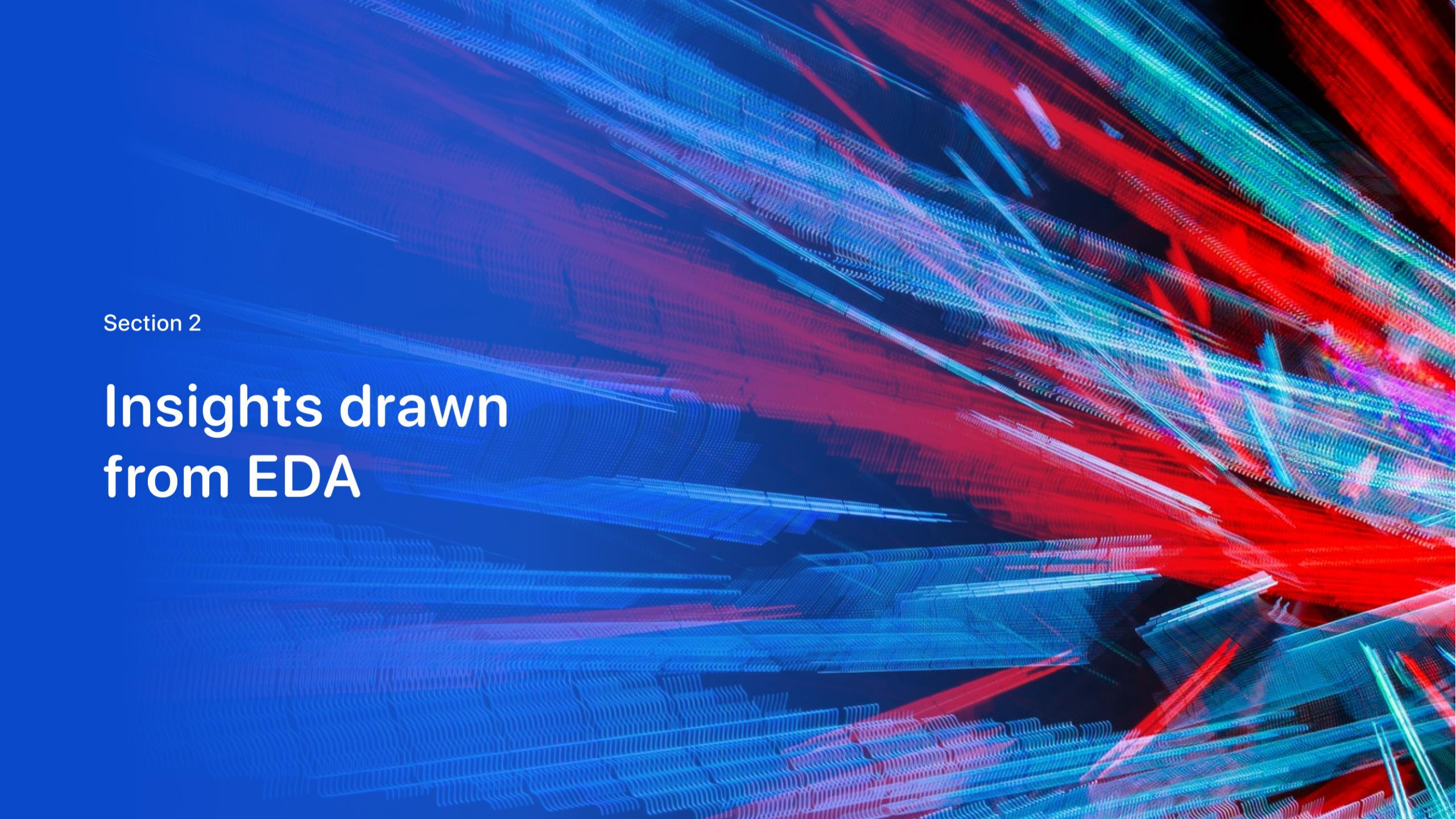
Results

- Launch sites are located near coastline, highways and railroads, far away from cities
- KSC LC-39A is the launch site have the most successful launches and highest successful launch rate
- FT booster version has the highest successful rate
- Best predictive model is LR with test score 83.33%

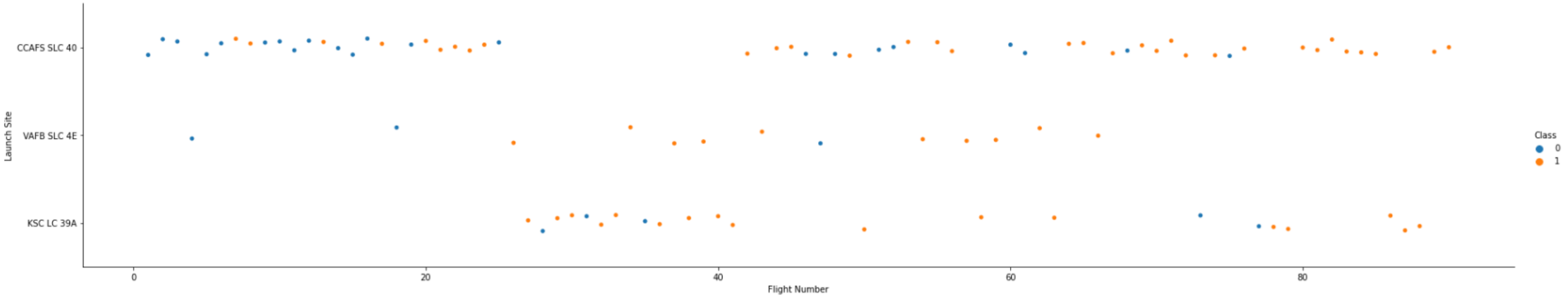


Section 2

Insights drawn from EDA

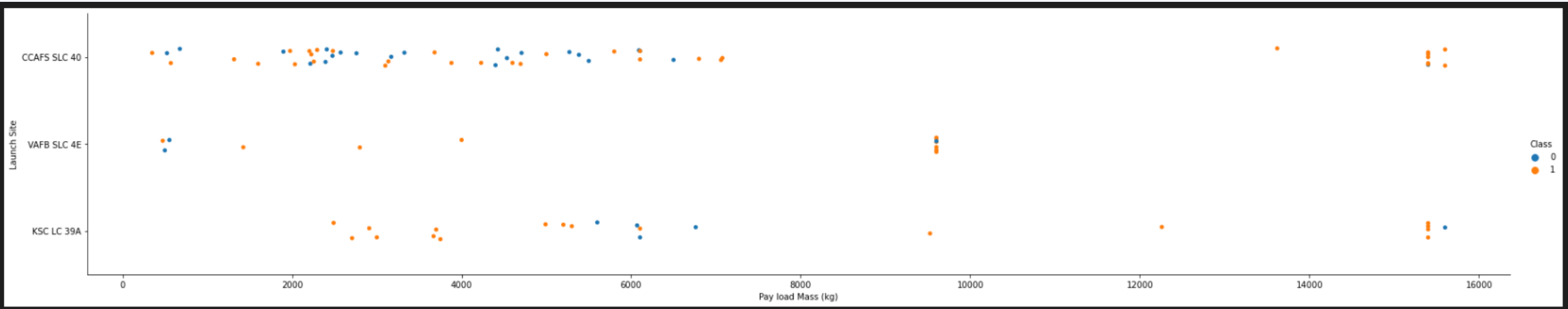


Flight Number vs. Launch Site



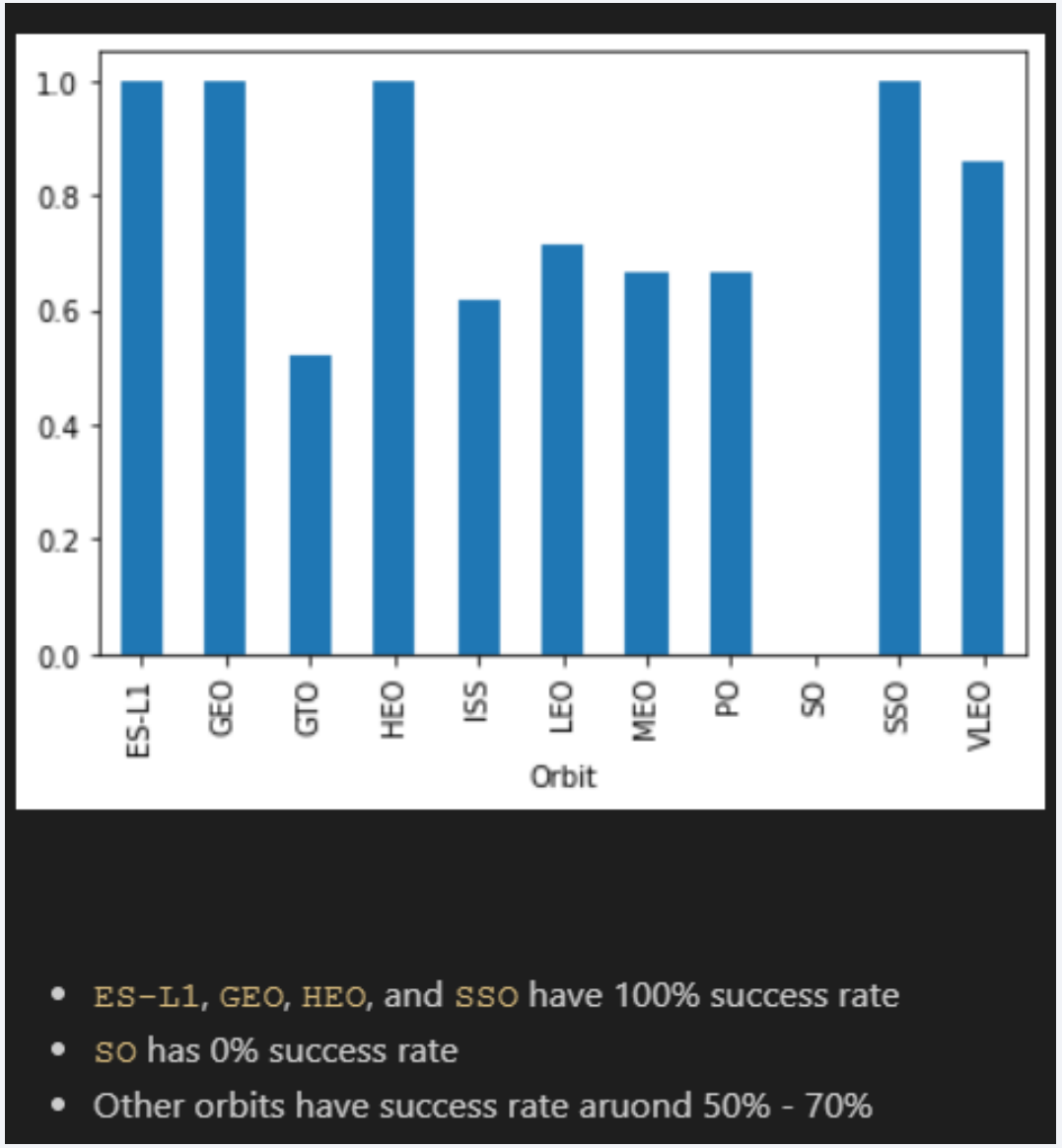
- Initially, `CCAFS SLC 40` has low chance of success. Later on, it is increase greatly. It aslo has the highest number of flight.
- `VAFB SLC 4E` has the lowest number of flight. It also has high chance of success.
- `KSC LC 39A` has a high chance of success with a lot of flight.
- At first, flights mainly occur at `CCAFS SLC 40` and rarely at `VAFB SLC 4E`. Then, flights start to launch at `KSC LC 39A` a lot instead of `CCAFS SLC 40`. Later, there are no more flight at `VAFB SLC 4E`, a few flights at `KSC LC 39A` and most of the flights are launched at `CCAFS SLC 40`
- As `Flight Number` increase, the flight seem to be more successful.

Payload vs. Launch Site



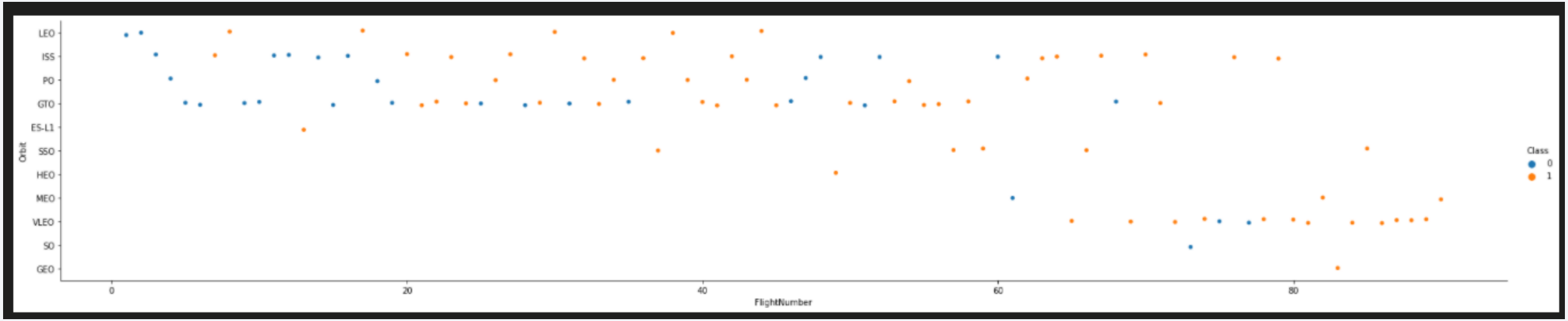
- Most of the rockets have Payload mass below 8000 kg and are mostly launched at `CCAFS SLC 40` and `KSC LC 39A`. There is no rocket has payload below 2000 kg is launched at `KSC LC 39A`.
- For the VAFB-SLC launchsite there are no rockets launched for heavy payload mass (greater than 10000 kg).
- For a payload mass greater than 8000 kg, there are some of the rockets with a certain payload mass values (around 9500 kg and 15500 kg)

Success Rate vs. Orbit Type



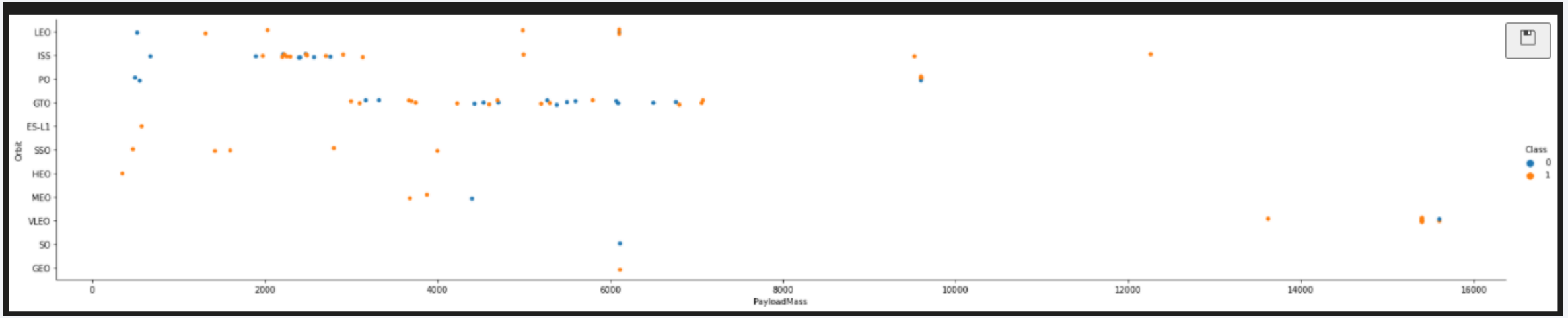
- ES-L1, GEO, HEO, and SSO have 100% success rate
- SO has 0% success rate
- Other orbits have success rate aruond 50% - 70%

Flight Number vs. Orbit Type



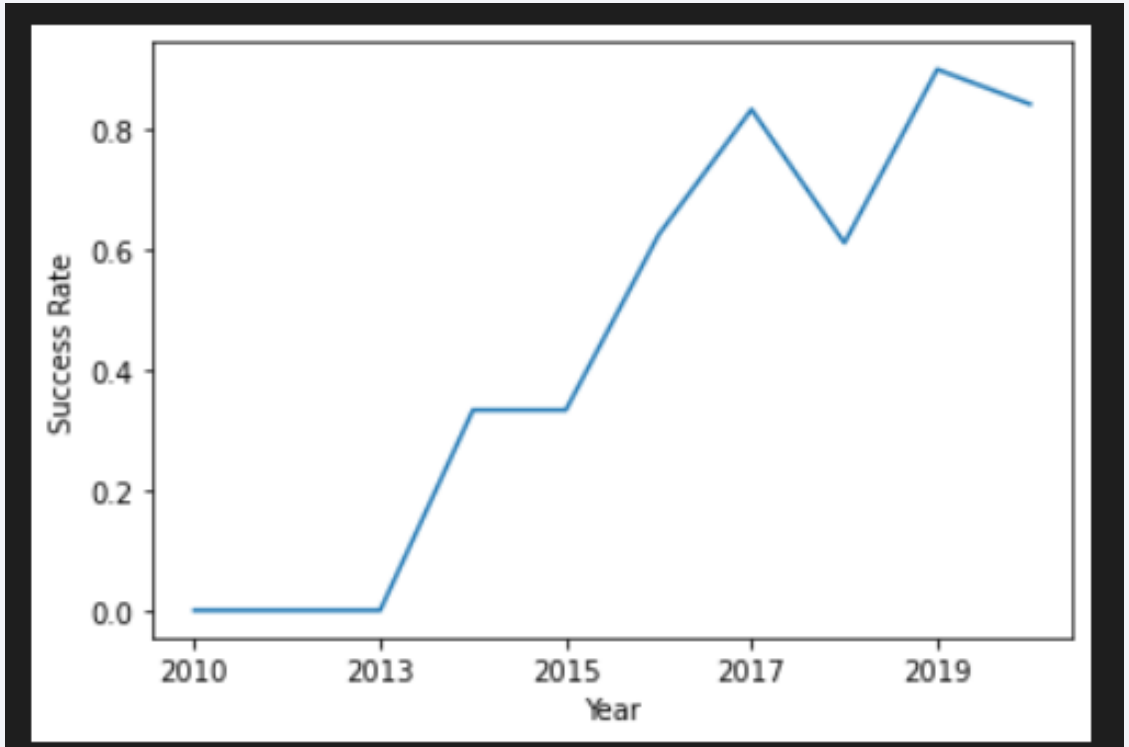
- In the LEO orbit the success appears related to the number of flights.
- There seems to be no relationship between flight number when in GTO and other orbits.
- Most of the flights are in LEO, ISS, GTO, PO, and VLEO orbit.

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.

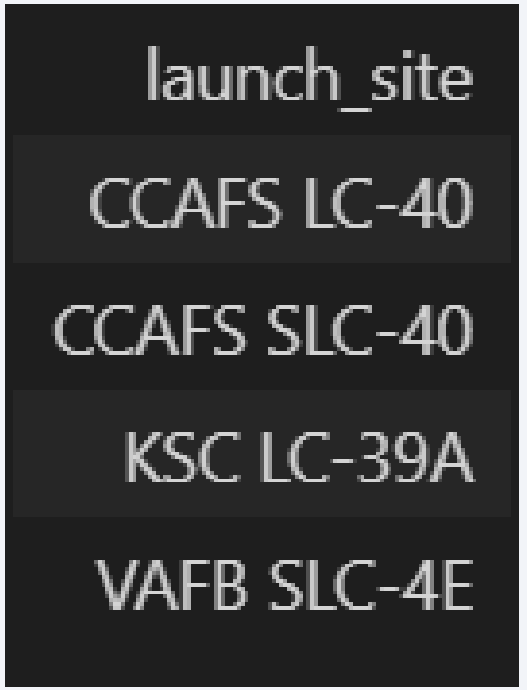
Launch Success Yearly Trend



- The sucess rate since 2013 kept increasing till 2017.
- Then it dropped about 10% in 2018 and increase again in 2019.
- Finally, there's a slight drop in 2020.

All Launch Site Names

- There are 3 launch sites
- CCAFS LC-40 changed its name to CCAFS SLC-40



Launch Site Names 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	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

- Five launches at CCAFS LC-40 are all successful but failed to land
- All the launches have low payload mass and orbit in LEO
- All the launches use different booster verions

Total Payload Mass

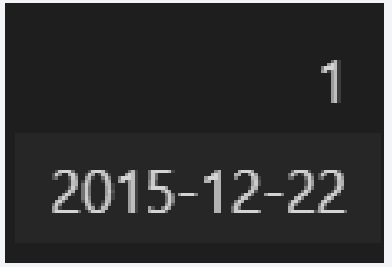
- Total payload mass carried by boosters launched by NASA is 45596 kg

Average Payload Mass by F9 v1.1

- Average payload mass carried by F9 v1.1 is 2534 kg

First Successful Ground Landing Date

- First successful ground landing was 2015, Dec 22nd



Successful Drone Ship Landing with Payload between 4000 and 6000

- There are only 4 boosters which have success in drone ship and have payload mass greater than 4000 kg and less than 6000 kg

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

Total Number of Successful and Failure Mission Outcomes

- There are mission outcome: success and failure
- There's only 1 failure mission which is failure in flight
- 100 successful mission. In that, there's one mission with unclear payload status

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- There are 12 boosters that carried the maximum payload and they're all F9 B5

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

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

- In 2015, there are two failed landing in drone ship. They are all launches at CCAFS LC-40

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

- There are 8 outcomes between 2010-06-04 and 2017-03-20
- The most occurred outcome is “No attempt”
- Two of them are success

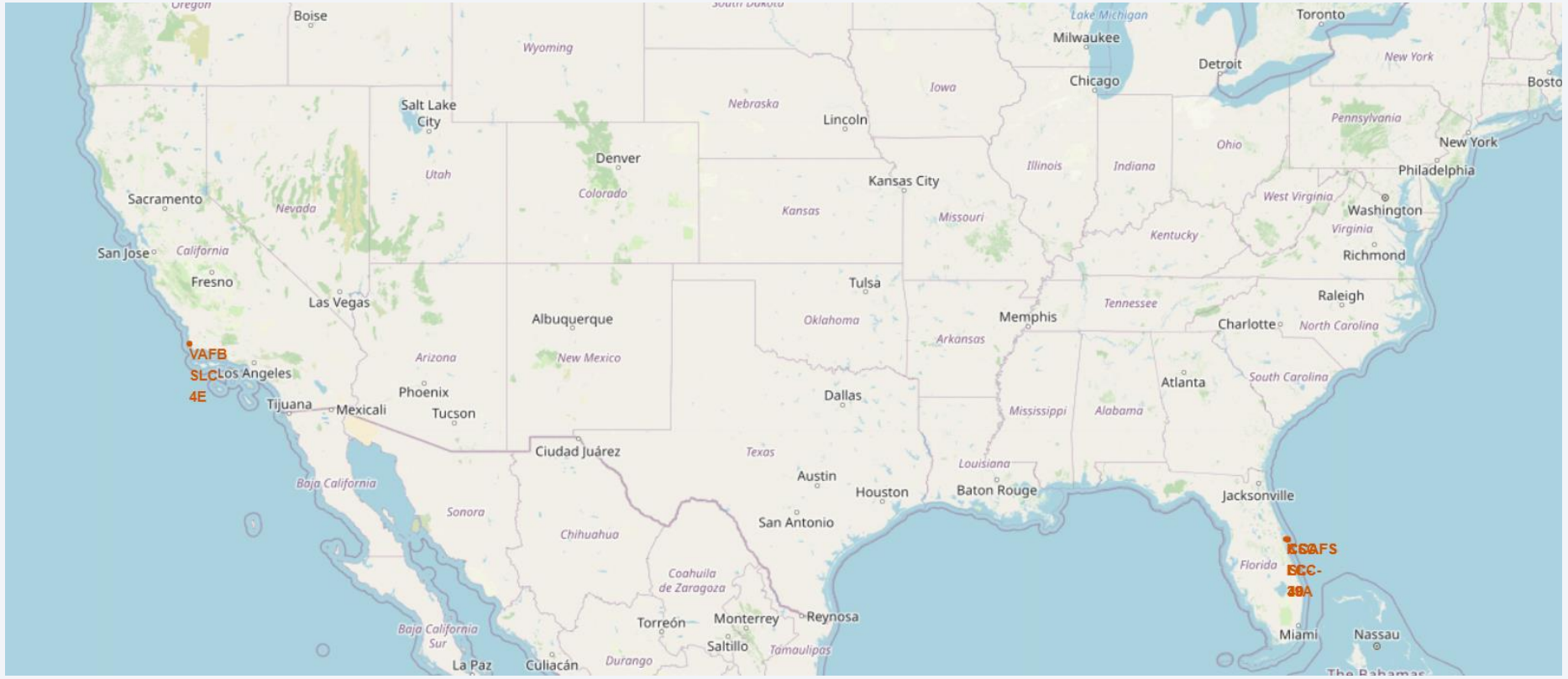
outcome	COUNT	RANK
No attempt	10	1
Failure (drone ship)	5	2
Success (drone ship)	5	2
Controlled (ocean)	3	4
Success (ground pad)	3	4
Failure (parachute)	2	6
Uncontrolled (ocean)	2	6
Precluded (drone ship)	1	8

Section 4

Launch Sites Proximities Analysis

Launch site location

- All the launch site are located near the coast
- KSC LC-39A is close to CCAFS SLC-40/CCAFS LC-40

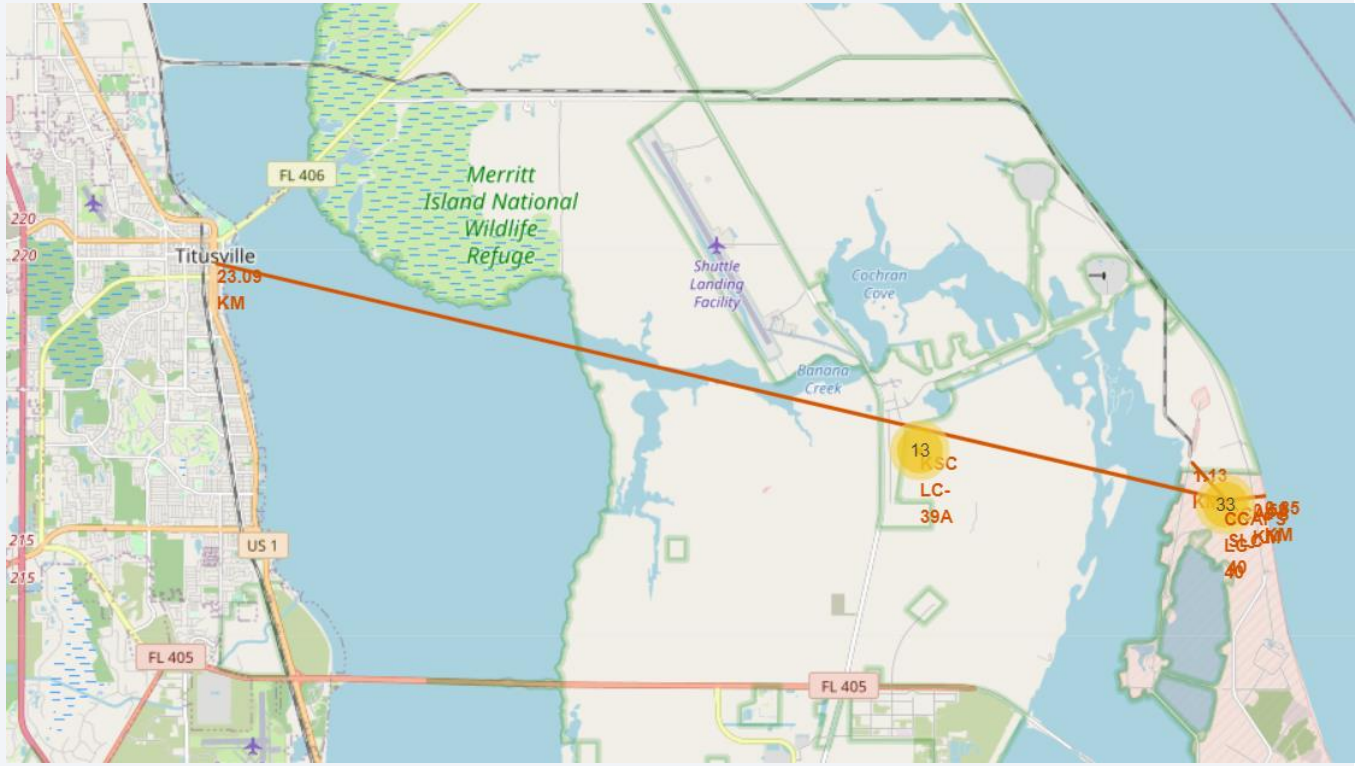


Mark launch result for each site

- Green markers show the successful outcome and red ones shows failed outcome
- CCAFS LC-40 has low chance of success but increase overtime (from inside to outside)

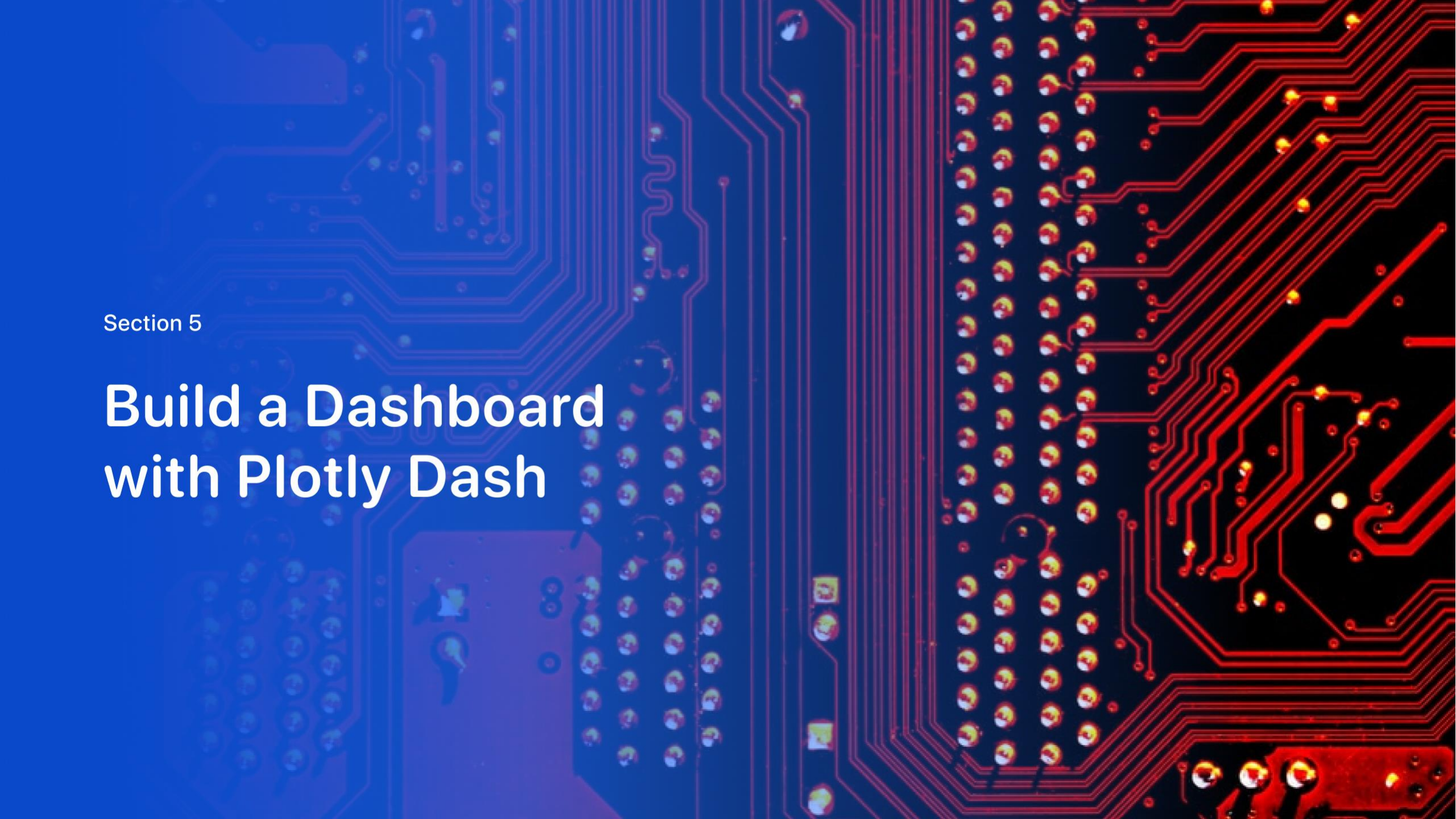
The distances between a launch site to its proximitie

- Distance from CCAFS SLC-40 to
  - Coastline: 0.85 km
  - Railway: 1.13 km
  - Highway: 0.58 km
  - City (residential area): 23.09km



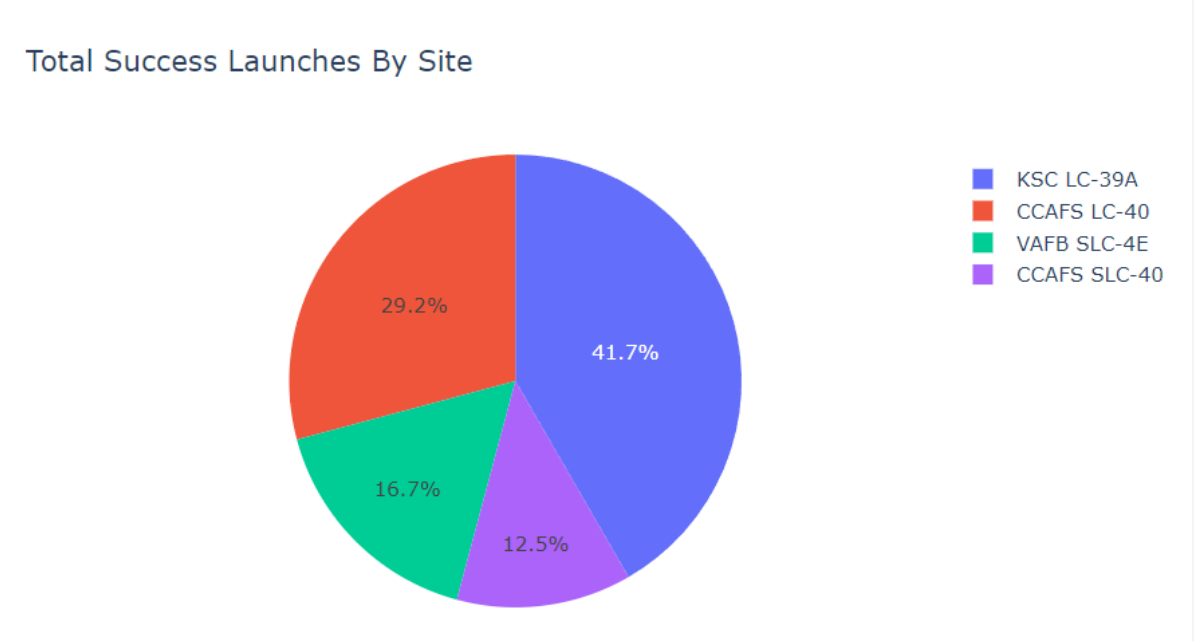
Section 5

Build a Dashboard with Plotly Dash



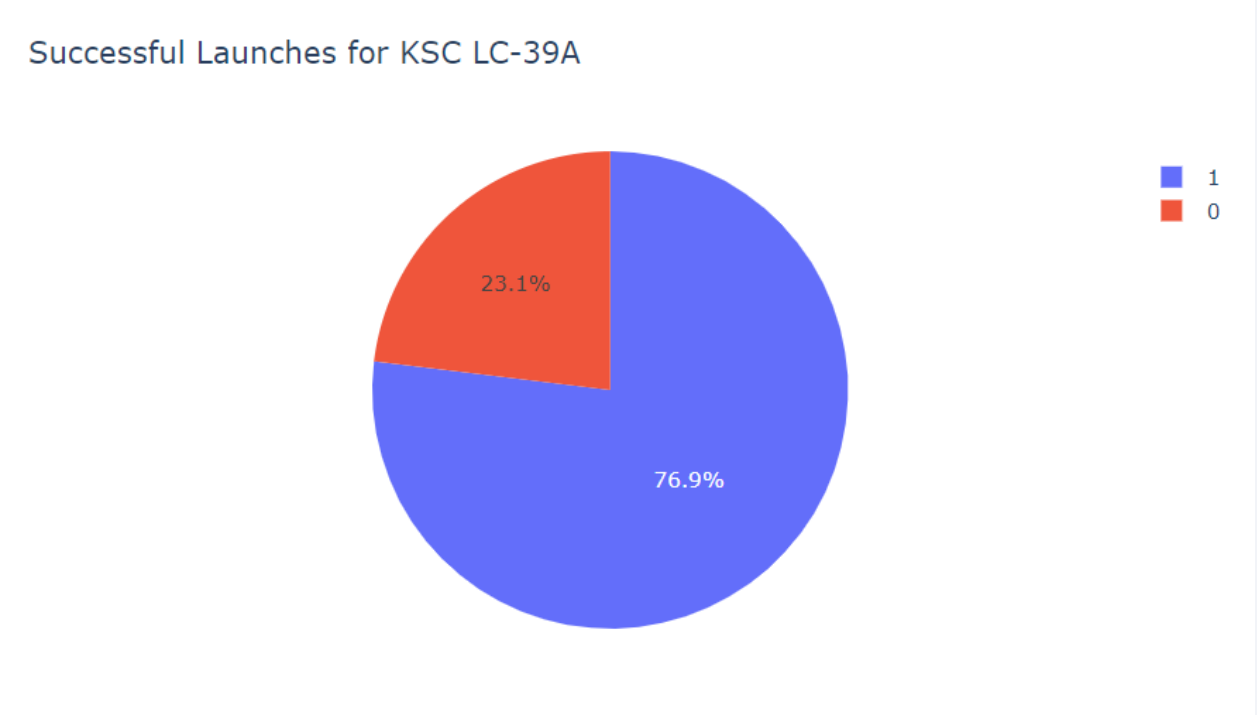
Total success launches by site

- KSC LC-39A has the highest percent of success launches (41.7% of all success launces)
- CCAFS LC-40 and CCAFS SLC-40 also add up to 41.7% of all success launches (CCAFS LC-40 got its name changed to CCAFS SLC-40)



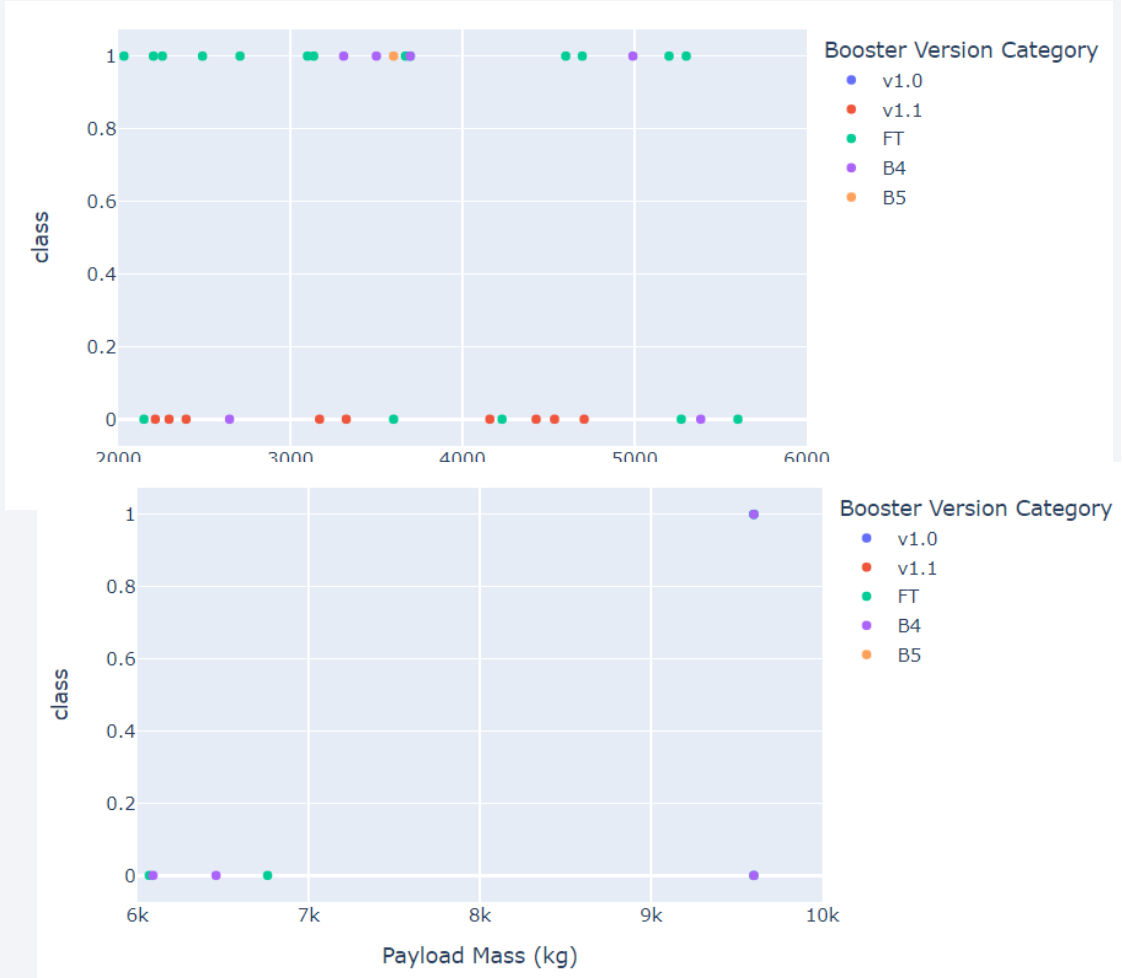
Launch success ratio

- KSC LC-39A also has the highest success launch ratio (76.9% to 23.1%)



Correlation between Payload and Success

- In range between 2000 kg and 6000 kg, boosters with version category of FT seems to have high successful chance while v1.1 seems to have high chance of failing.
- With Payload more than 6000 kg, it is almost certain that the mission outcome will be failed.

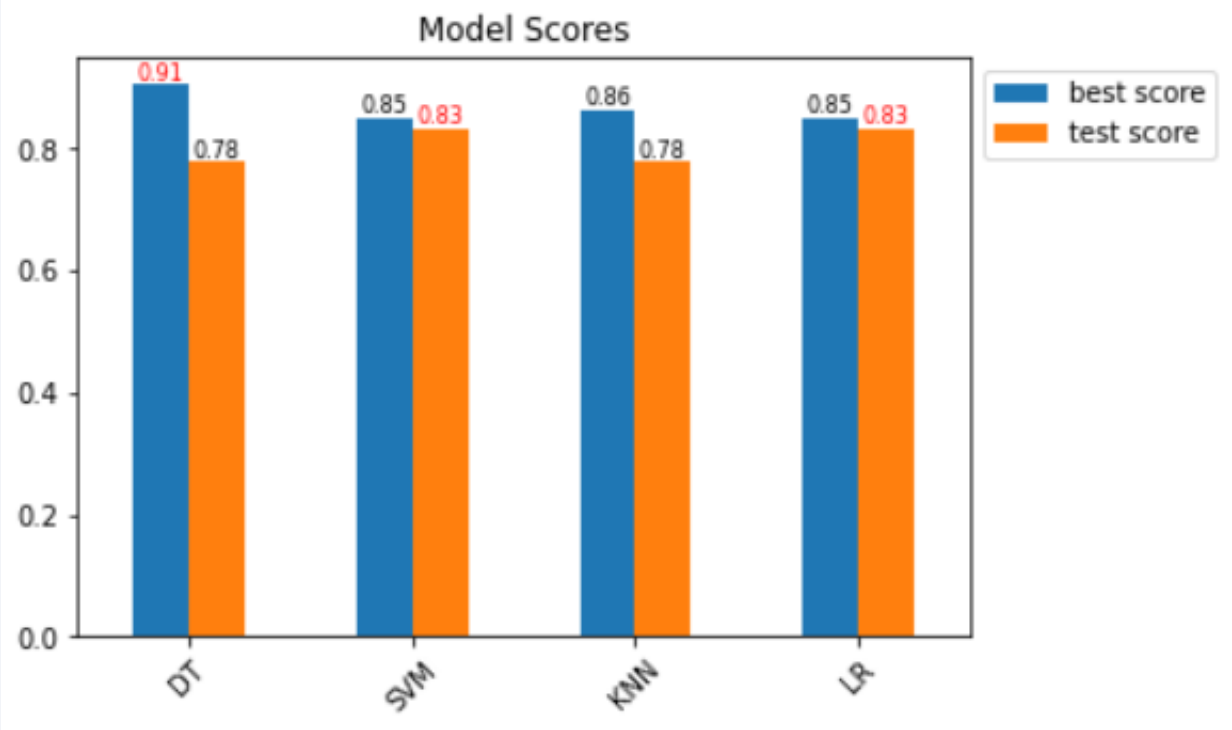


Section 6

Predictive Analysis (Classification)

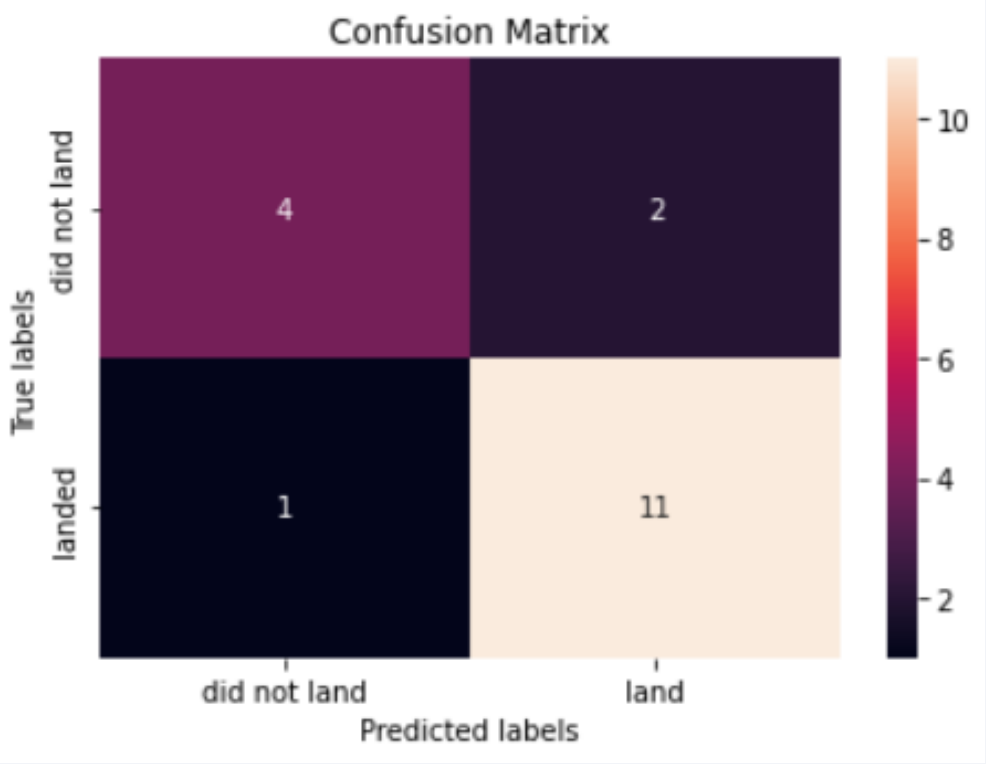
Classification Accuracy

- SVM and LR are the best models with the same performance (83% test accuracy)
- DT has the highest training accuracy but lower test score (overfitted).



Confusion Matrix

- SVM and LR result in the same confusion matrix
- The models performed on the test dataset with 15 correct prediction, 3 incorrect prediction:
  - 11 true positive, 4 true negative
  - 2 false positive, 1 false negative



Conclusions

- Launch sites are located near coastline, highways and railroads, far away from cities
- KSC LC-39A is the launch site have the most successful launches and highest successful launch rate
- FT booster version has the highest successful rate
- Best predictive model is LR with test score 83.33%

Appendix

- https://github.com/ngquocanh1197/Coursera-IBM-DS/tree/main/Course%2010

Thank you!