IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jad Saade
09 October, 2023

Outline

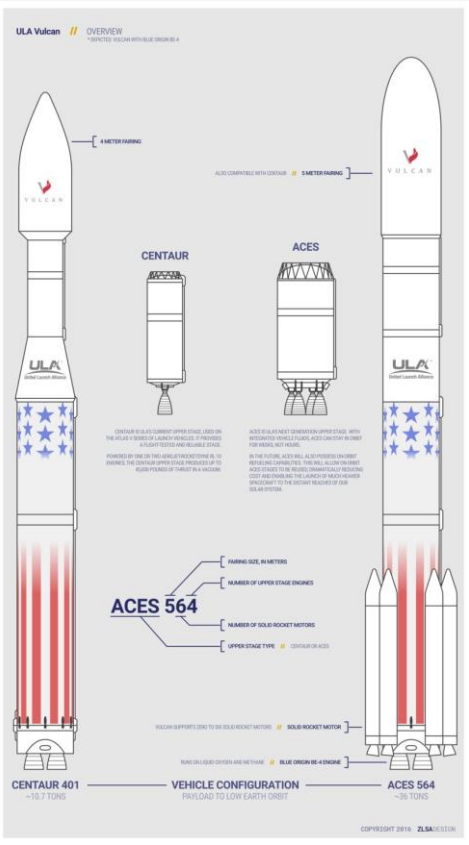
- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- **Objective:** Predicting if a space rocket will land back in earth successfully
- **Methodologies used**:
  - Data collection using (a) SpaceX API and (b) web-scraping;
  - Data wrangling;
  - Exploratory Data Analysis (EDA) using python and SQL Commands;
  - Data visualization and interactive dashboard analytics;
  - Prediction using Machine Learning (ML)
- **Summary of Results:**
  - It was possible to collect valuable data from public sources
  - EDA allowed to identify which features are the best to predict success of launchings
  - Machine Learning prediction showed which is the best supervised ML model to predict

Introduction

- The project background is working for SpaceY fictitious rocket company to compare with SpaceX company.
- SpaceX gained popularity as of December 2010 after being the 1st private company to ever return a spacecraft from low earth orbit
- SpaceX is more advanced and cheaper than other rocket companies since the cost of the 1st stage of its rocket launch using Falcon 9 ($ 62 million) is much cheaper than that of other companies ($ 165 million).
- The project objective is to predict using Machine Learning (instead of rocket science) whether the 1st stage of the launch can be reused.
- This will enable company SpaceY to be able to bid against SpaceX
- [Image Source Link]



Section 1

Methodology

Methodology (1/2)

- Data from SpaceX was collected from 2 sources:
  - SpaceX API (Space X API Link)
  - Web-scraping from a Wikipedia page (Space X Wikipedia Link)
- Data wrangling was conducted:
  - Collected data was enriched by creating a landing outcome label (Class = 0 or 1) based on outcome data after summarizing and analyzing features

Methodology (2/2)

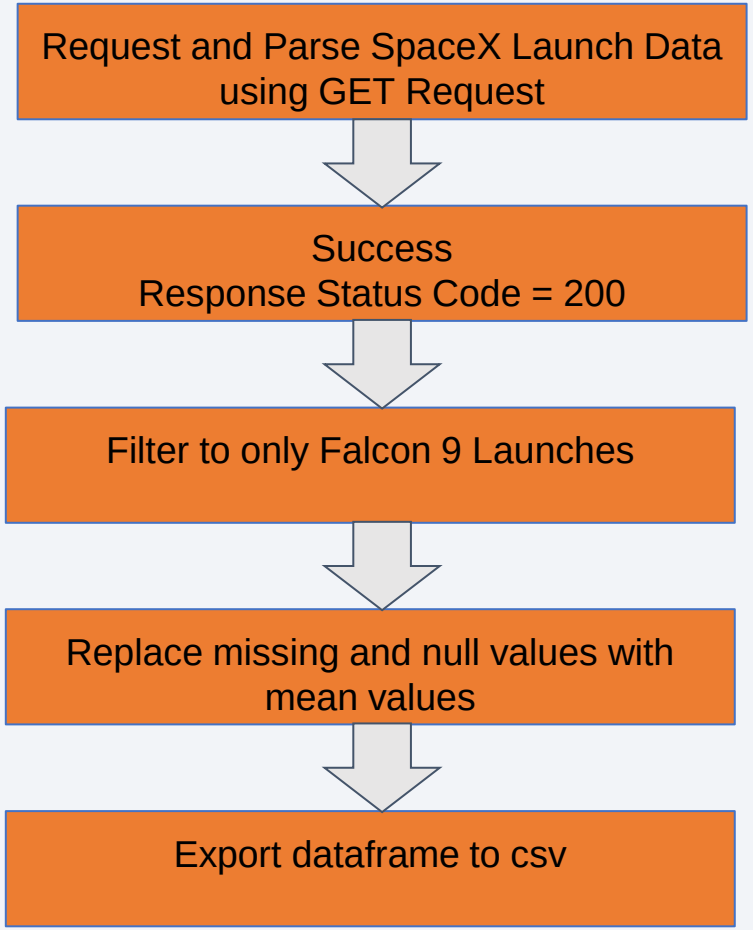
- Perform Exploratory Data Analysis (EDA) using python commands and SQL commands (using **sqlite3** library)
- Perform interactive visual analytics using **Folium** and **Plotly Dash** libraries
- Perform predictive analysis using classification models
  - Data that was collected was encoded and normalized, split into training and testing data sets and evaluated for different ML classification models
  - The performance of the models was evaluated using accuracy on the different input parameter combinations

Data Collection

- Datasets were collected from
  - SpaceX API using **requests** library
    - Space X API Link
  - Wikipedia using web-scraping from **BeautifulSoup** library
    - Space X Wikipedia Link

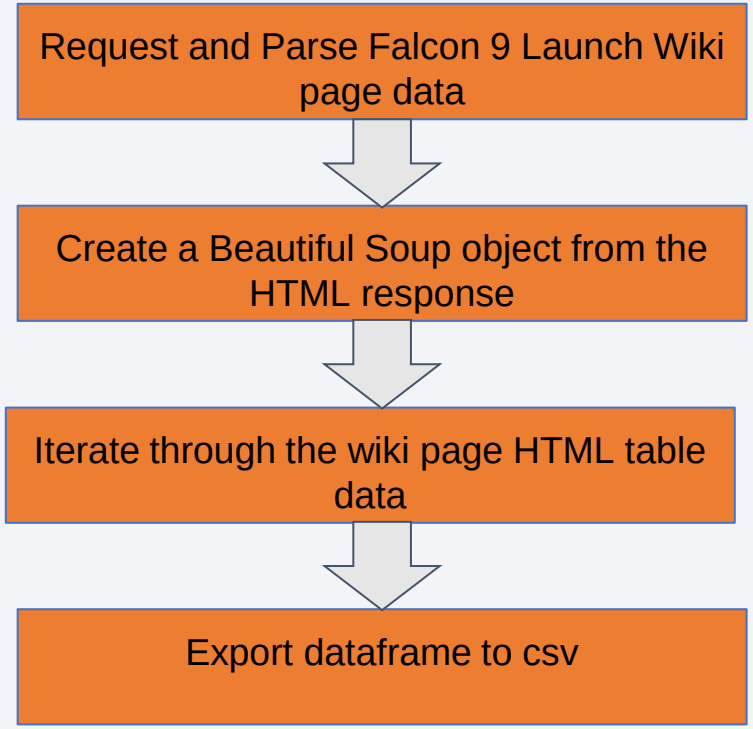
Data Collection – SpaceX API

- SpaceX offers a public API from where data can be obtained using an API request and then used;
- This API was used according to the flowchart beside and then data is stored
- GitHub Link – Data Collection with API



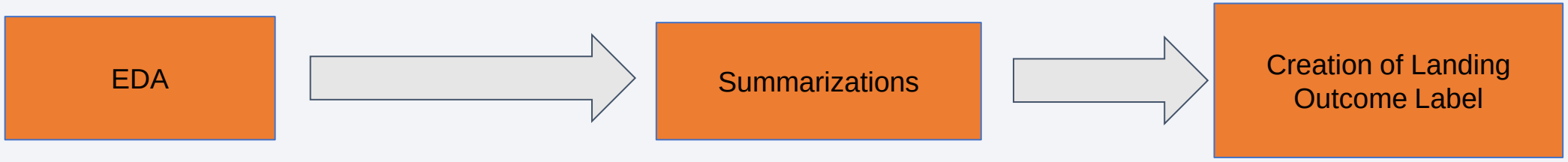
Data Collection - Web-scraping

- Data from SpaceX launches can also be obtained from Wikipedia;
- Data are downloaded from the Wikipedia according to the flowchart and then persisted
- GitHub Link - Data Collection Using Web scraping



Data Wrangling

- Some EDA was performed on the dataset to find patterns in the data for the 90 launches
- Find number of launches on each of the 3 sites
- Find number and occurrence of each of the 12 orbits
- Find number and occurrence of mission outcomes for each of the 12 orbits
- Create a landing outcome label (60 Successful labelled 1 and 30 unsuccessful labelled 2)
    - Success rate 66.66%
- Save the data in csv
- GitHub Link - Data Wrangling



EDA with Data Visualization

- To explore the data, scatter plots were used to visualize the relationship between pairs of features (with success rate as a hue in seaborn)
    - Payload Mass (kg) vs. Flight Number
    - Launch Site vs. Flight Number
    - Launch Site vs. Payload Mass (kg)
    - Orbit vs. Flight Number
    - Orbit vs Payload Mass (kg)
- Bar plot was used to visualize relationship of Success rate vs Orbit
- A line plot was used to visualize relationship of Success rate vs Date
- GitHub Link - EDA with Visualization

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 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 1st successful landing outcome in ground pad was achieved,
  - Names of the boosters which have success in drone ship and have payload mass between 4,000 and 6,000 kg,
  - Total number of success 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 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 Link - EDA with SQL

Build an Interactive Map with Folium

Folium library was used for geospatial visualization (i.e., creating Folium maps with certain features):

- Markers indicated point locations of launch sites
- Circles indicated highlighted areas around specific coordinates like the NASA Johnson Space Center
- Marker clusters indicated groups of events in each coordinate
- Lines indicated distances between points (markers)

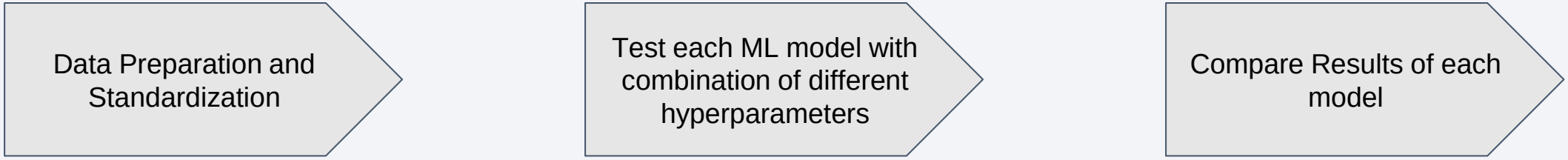
- GitHub Link - Geospatial Visualization with Folium

Build a Dashboard with Plotly Dash

- A dashboard for graphs was used to visualize data and answer the following questions:
  - Which site has the largest successful launches?
  - Which site has the highest launch success rate?
  - Which payload range(s) has the highest launch success rate?
  - Which payload range(s) has the lowest launch success rate?
  - Which F9 Booster version (v 1.0, v 1.1, FT, B4, B5, etc.) has the highest launch success rate?
- GitHub Link - Dashboard

Predictive Analysis (Classification)

- 4 classification supervised ML models were used and compared for different hyperparameters:
  - Logistic regression
  - Support Vector Machine (SVM)
  - Decision Tree
  - k Nearest Neighbors (KNN)
- GitHub Link - Machine Learning



Results – EDA Results

- SpaceX uses 4 different launch sites
- The 1st launches were done to SpaceX itself and NASA
- The average payload of F9 v1.1 booster is 2,929 kg
- The 1st success landing outcome happened in 2015 5 years after the 1st launch
- Many F9 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 better as years passed

Results on Geospatial Visualization

- **Are launch sites in close proximity to railways?**

They all are far from railways

- **Are launch sites in close proximity to highways?**

They ones east in Florida are close to highways. The one easternmost is close to Samuel C Phillips Parkway, Titan III Road. The other right west of that, close to Kenedy-Parkway North. The one west in California isn't close to any major highway

- **Are launch sites in close proximity to coastline?**

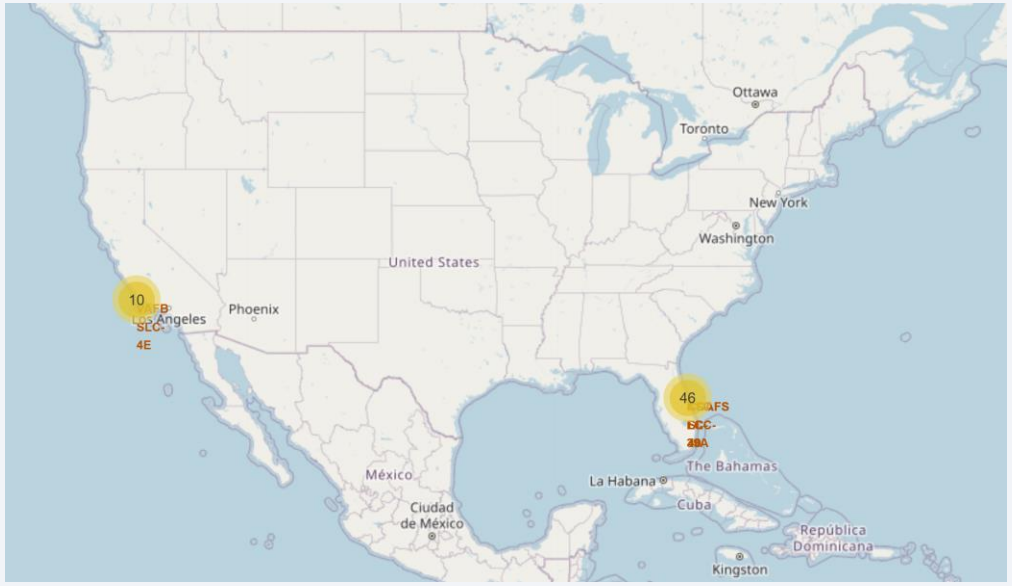
Yes, 2 launch sites in Florida, 9 and 11 km from Atlantic coastline and 1 in California similar distance from Pacific coast

- **Do launch sites keep certain distance away from cities?**

Yes, all 3 of them are located away from any major city to avoid having their launches disrupt to dense urban centers

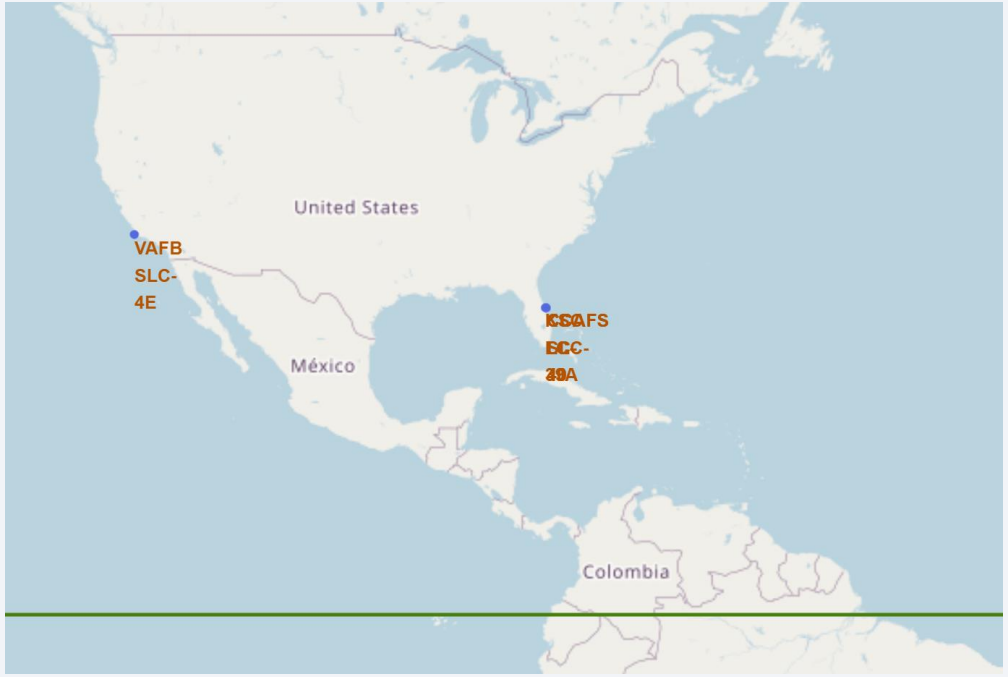
Results on Geospatial Visualization

- Using interactive analytics was possible to identify that launch sites are located in safe places near the sea, for example, and have a good logistic infrastructure around.
- Most launches happen at east coast launch sites



Results on Geospatial Visualization

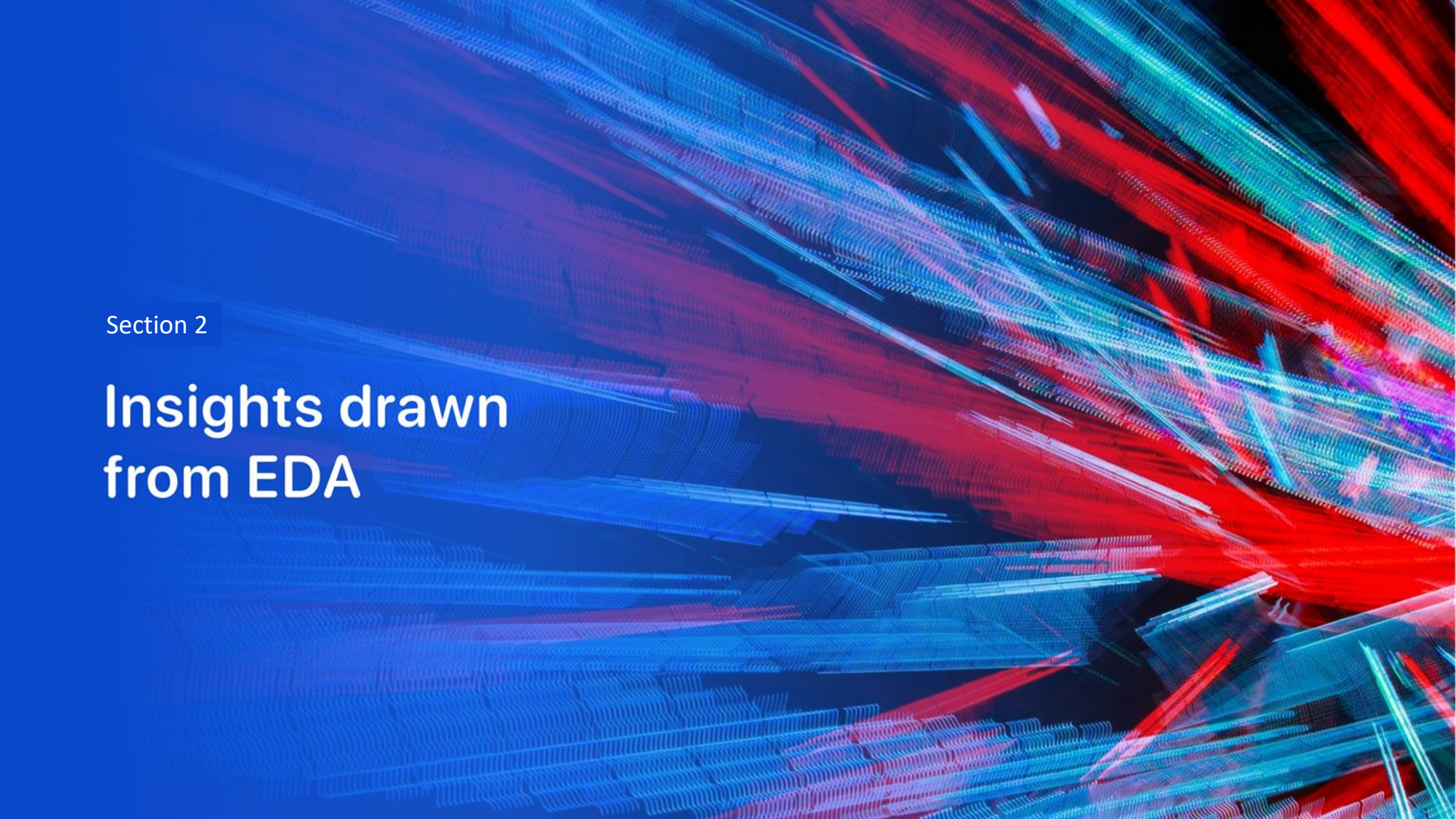
- Sites in Florida are closer to the Equator than those in California



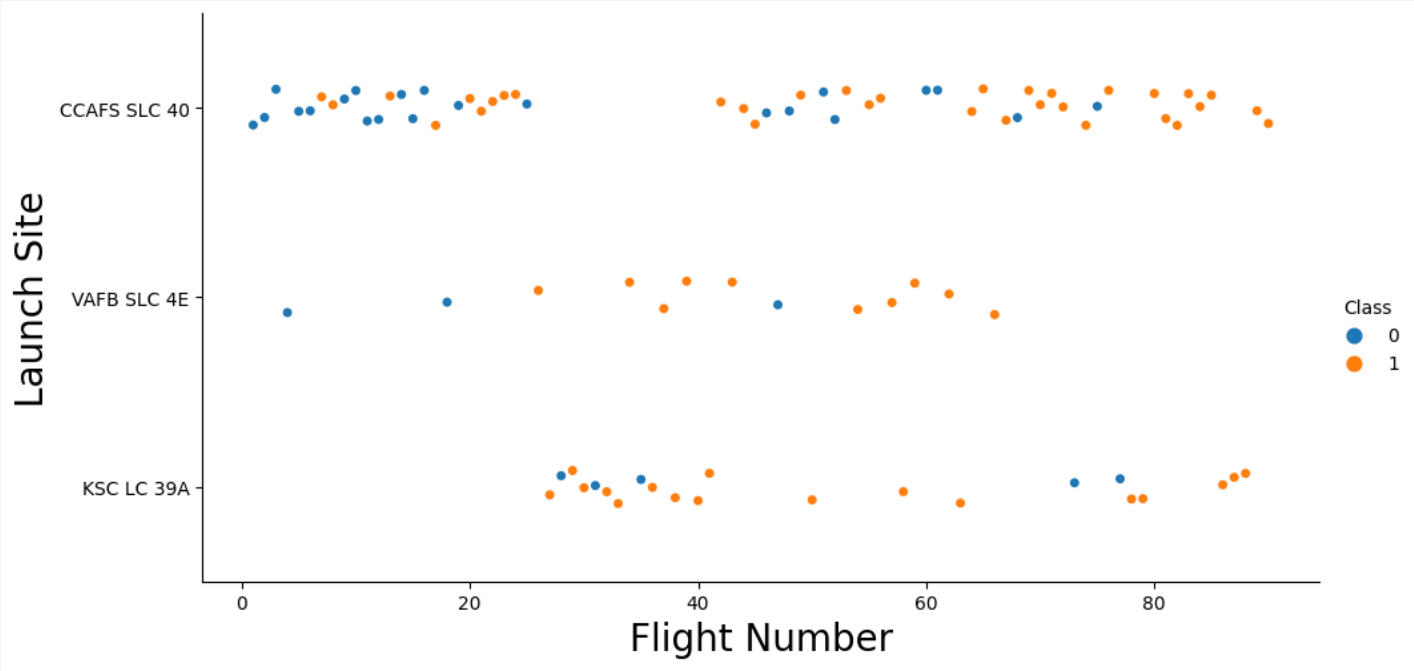
Results – Machine Learning

- Predictive analysis showed that Decision Tree Classifier is the best model to predict successful landings, having accuracy over **83%** and accuracy for test data over **94%.**

Classifier	Train Data Accuracy	Test Data Accuracy
Logistic Regression	70.83%	84.64%
Support Vector Machine	88.8%	83.3%
Decision Tree	83.3%	94.4%
KNN	86.1%	83.3%

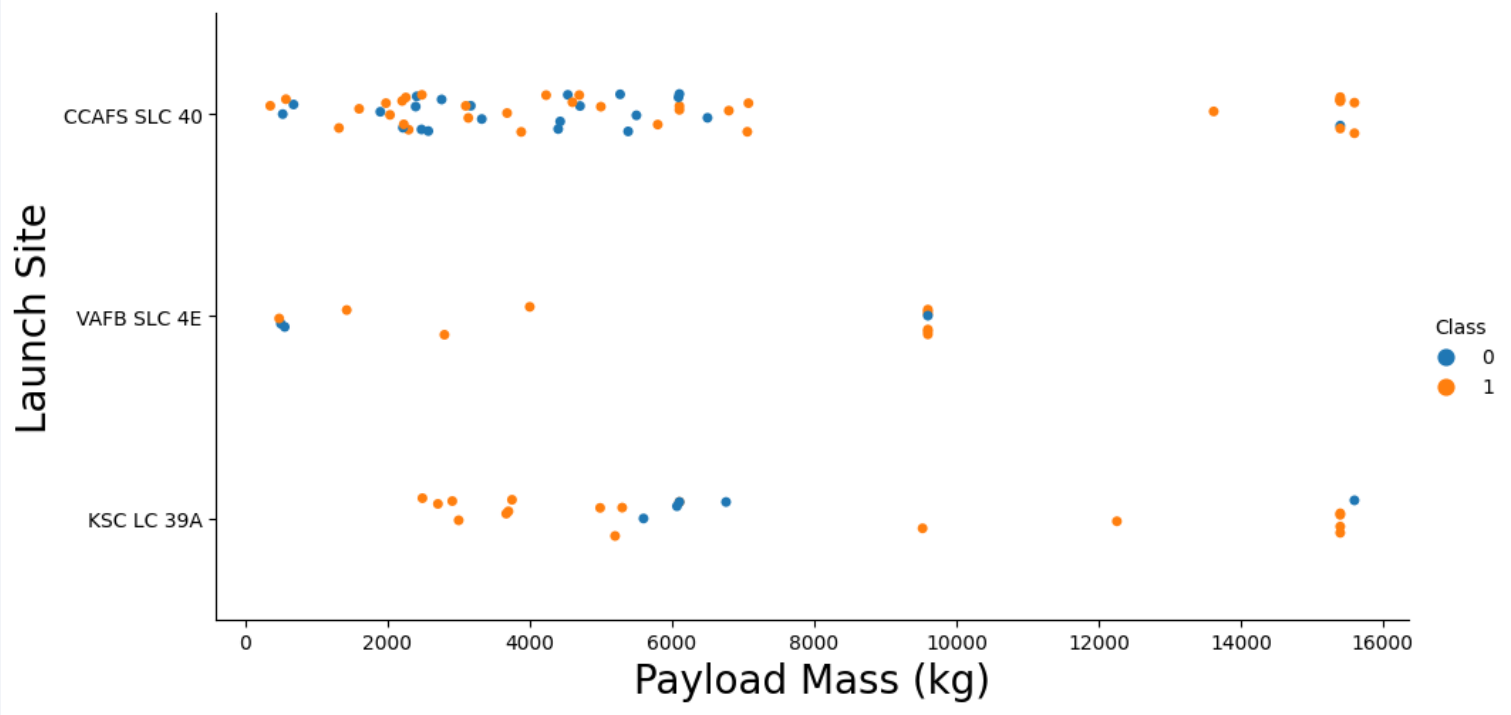

Section 2
Insights drawn
from EDA

Flight Number vs. Launch Site



- According to the plot above, CCAF5 SLC 40 is the best launch site, where most of the recent launches were successful (the ones with higher Flight number in orange)
- VAFB SLC 4E is 2nd and KSC LC 39A is 3rd
- Generally, success rate improved over time for all 3 Launch Sites

Payload vs. Launch Site

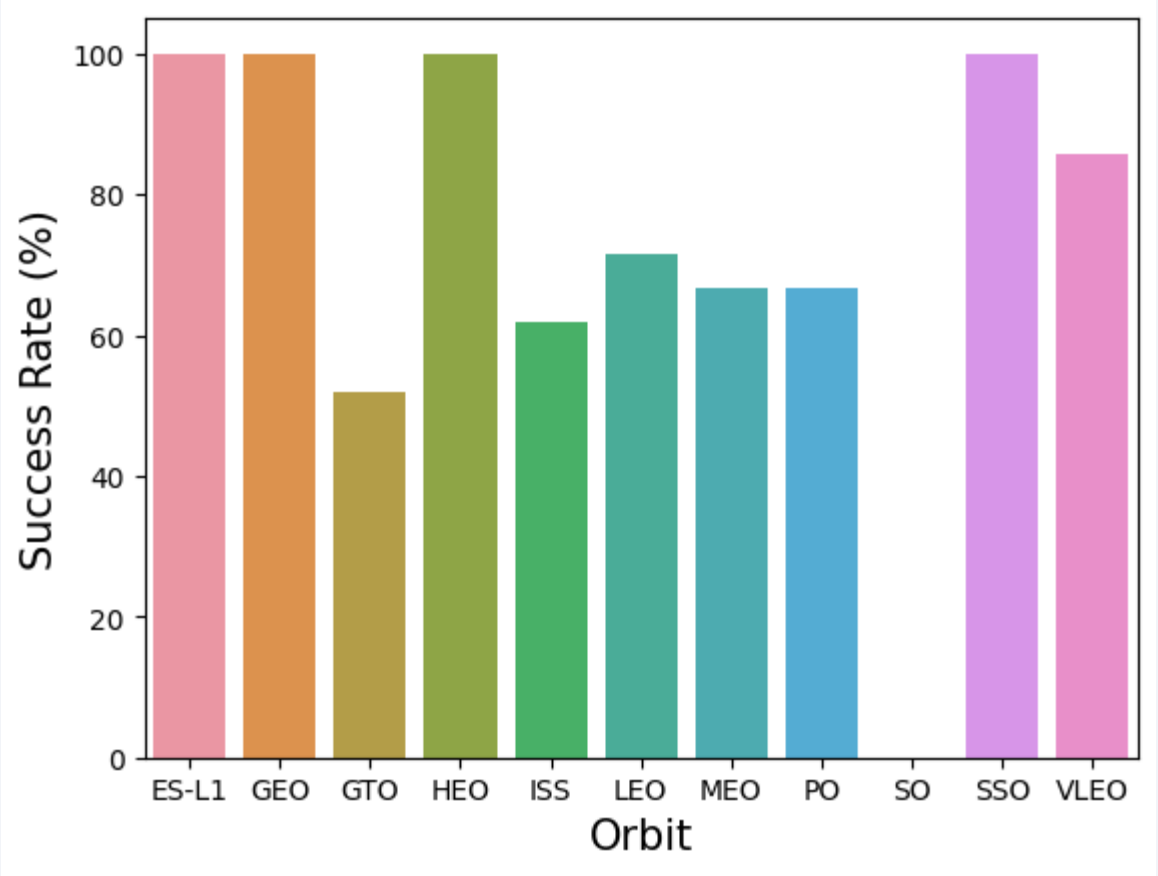


- Payloads over **9,500 kg** (about the weight of a school bus) have excellent success rate (all orange)
- Payloads over **12,000 kg** seem to be possible only for CCAFS SLC 40 and KSC LC 39A Launch Sites

Success Rate vs. Orbit Type

11 orbits

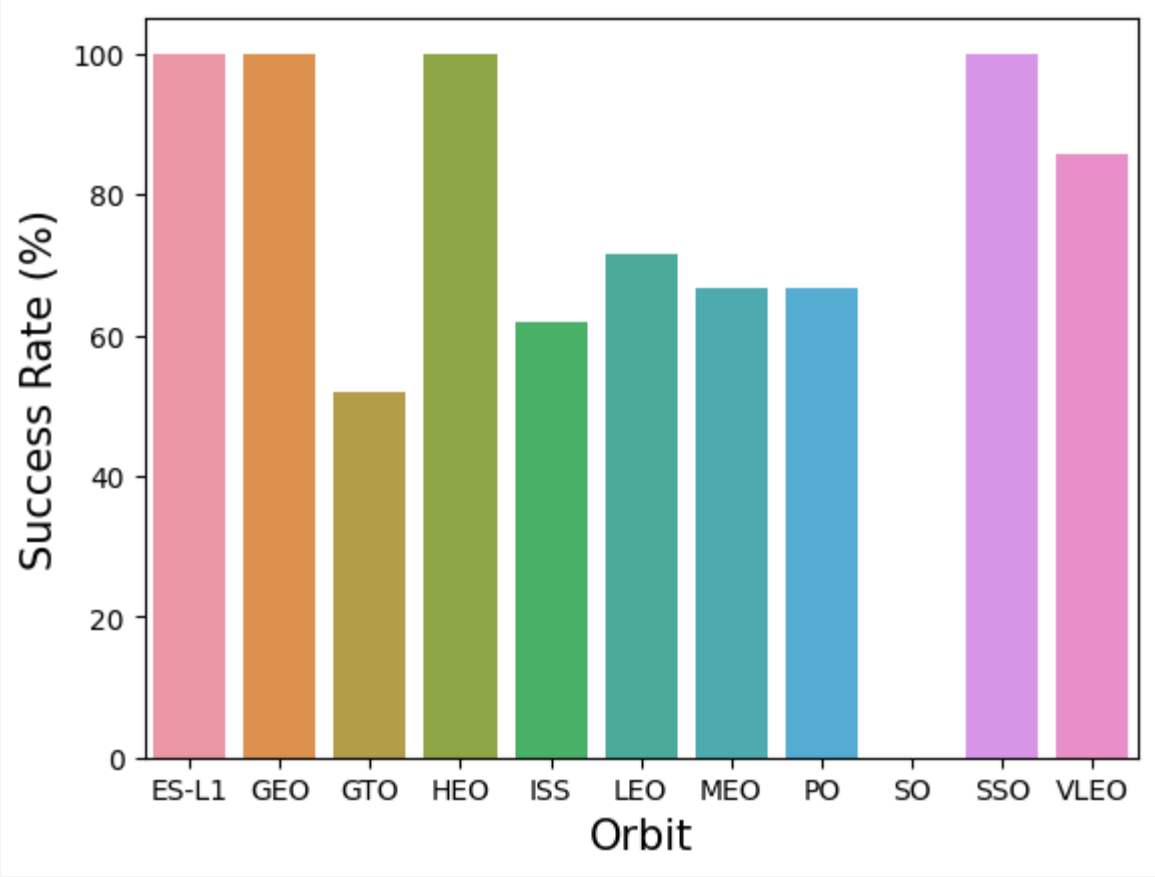
- 4 orbits with 100% success rate:
  - **ES-L1: Earth-Sun Lagrange Point 1,** one of the five Lagrange points in the Earth-Sun system where the gravitational forces of the Earth and Sun balance out
  - **GEO: Geostationary orbits** are orbits located at an altitude where a satellite's orbital period matches the rotation period of the Earth.
  - **HEO: Highly Elliptical Orbit** is an orbit that is not circular but instead has a significant degree of ellipticity. These orbits are often used for missions like certain types of communication, navigation, and reconnaissance satellites.
  - **SSO: Sun-Synchronous Orbit** is a type of near-polar orbit used by many Earth-observing satellites



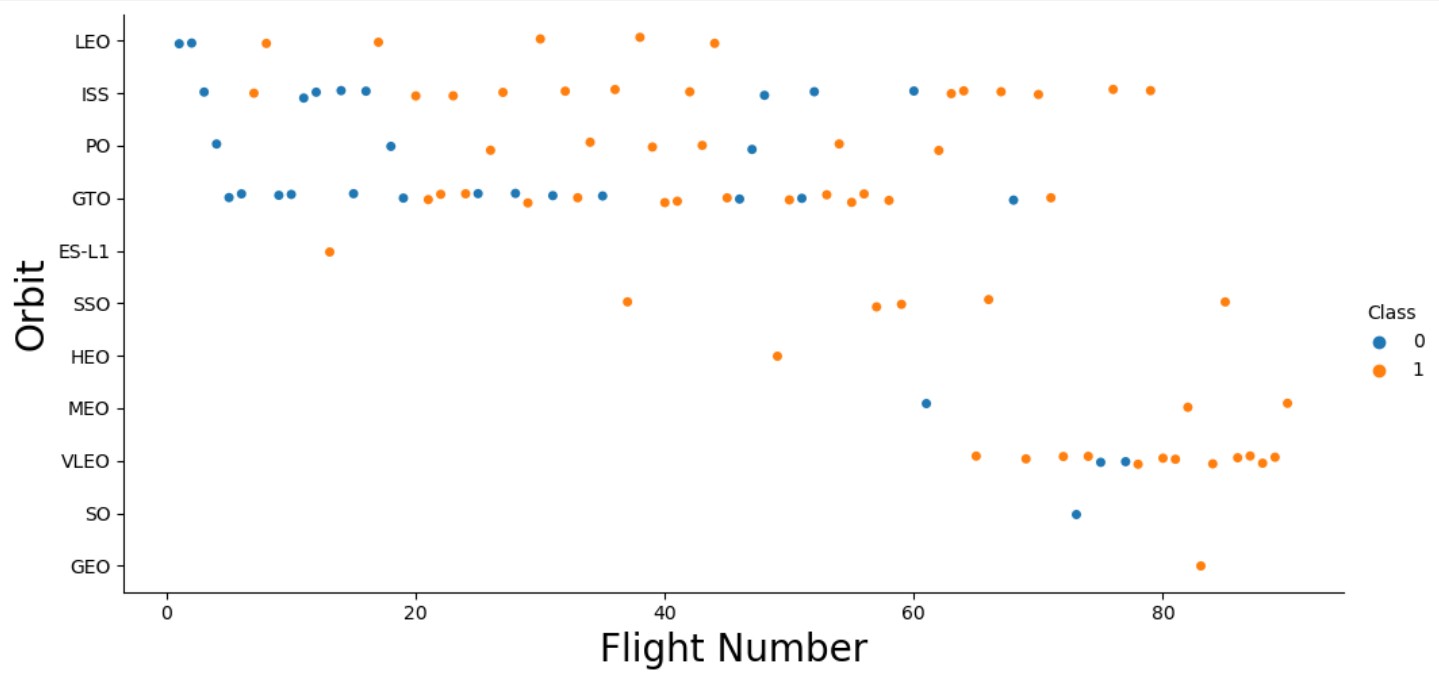
Success Rate vs. Orbit Type

11 orbits

- 7 are below 100% success
  - VLEO (~86%)
  - LEO (~71%)
  - MEO (~66.6%)
  - PO (66.6%)
  - ISS (~62%)
  - GTO (~52%)
  - SO (0%)

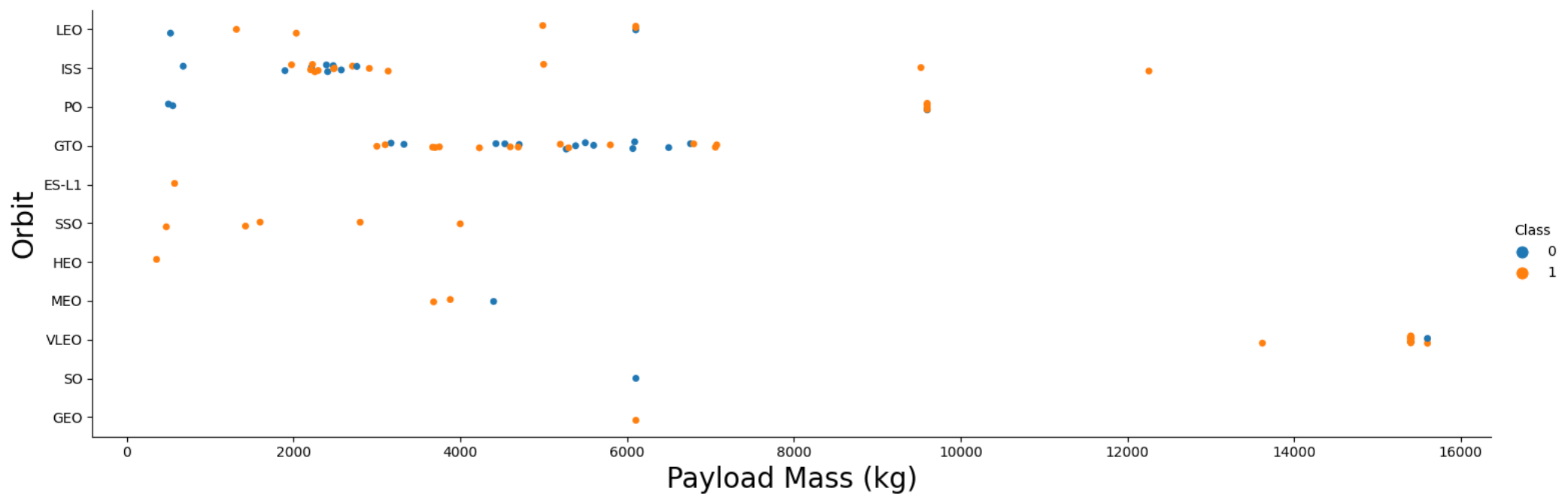


Flight Number vs. Orbit Type



- Success rate gets better over time
- VLEO has frequent success flights

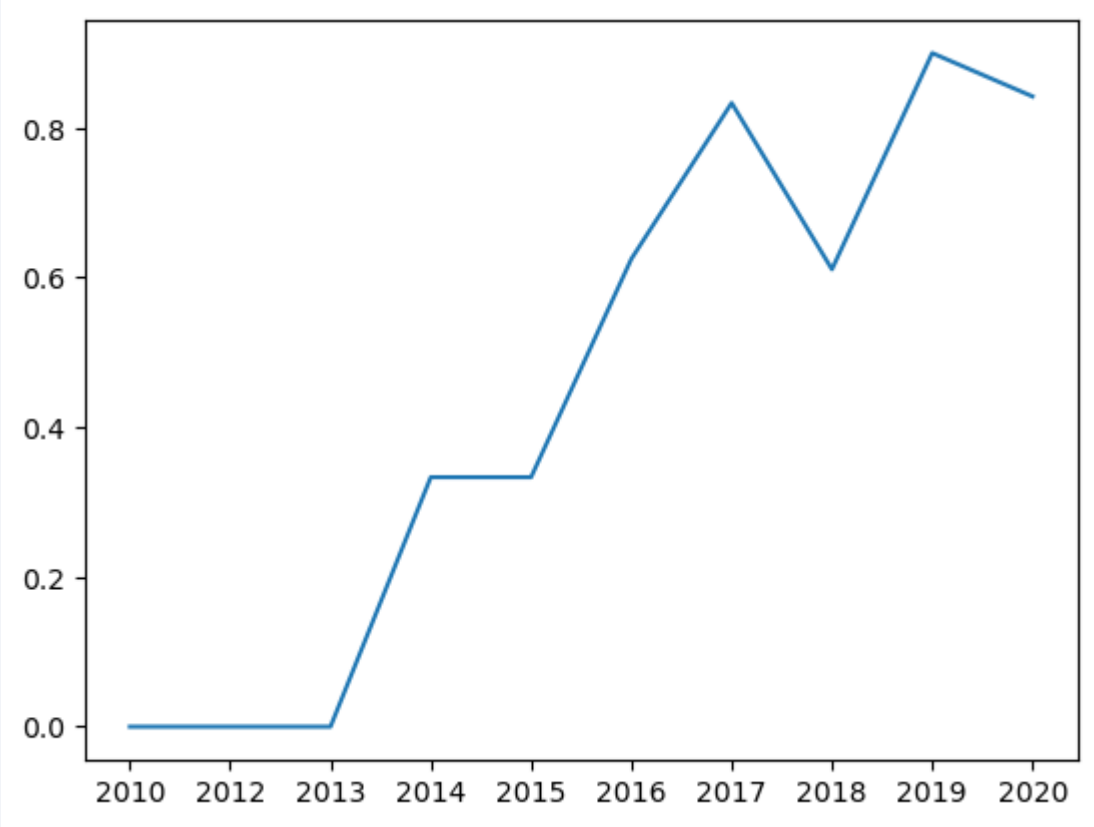
Payload vs. Orbit Type



- For the GTO orbit, success rate is not apparently in any direct relationship with payload mass
- SSO has very good success rate for 5 flights of payload mass <4000 kg
- VLEO is the only orbit with payload mass over 12,000 kg, reaching ~16,000 kg

Launch Success Yearly Trend

Year	Flights
2020	19
2017	18
2018	18
2019	10
2016	8
2014	6
2015	6
2013	3
2010	1
2012	1



- 60 success landings
- 30 failed landings
- Most flights (and landings) happened in 2020 (19 landings)
- Plotting the average yearly success rate for each year, we find that success rate started increasing after 2013 till 2020 with a drop in

All Launch Site Names

- There are 3 unique launch sites. Their unique names and their count are obtained from the following command

df[‘LaunchSite’].value_counts()

Launch Site	Count
CCAFS SLC 40	55
KSC LC 39A	22
VAFB SLC 4E	13

Launch Site Names Begin with 'CCA'

- Launch sites beginning with ‘CCA’ are basically the ones that are for Launch Site “CCAFS SLC 40”
- Below are 5 records for these launch sites

FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude	Class
1	06/04/2010	Falcon 9	6104.959	LEO	CCAFS SLC 40	None None	1	FALSE	FALSE	FALSE		1	0	B0003	-80.5774	28.56186	0
2	05/22/2012	Falcon 9	525	LEO	CCAFS SLC 40	None None	1	FALSE	FALSE	FALSE		1	0	B0005	-80.5774	28.56186	0
3	03/01/2013	Falcon 9	677	ISS	CCAFS SLC 40	None None	1	FALSE	FALSE	FALSE		1	0	B0007	-80.5774	28.56186	0
5	12/03/2013	Falcon 9	3170	GTO	CCAFS SLC 40	None None	1	FALSE	FALSE	FALSE		1	0	B1004	-80.5774	28.56186	0
6	01/06/2014	Falcon 9	3325	GTO	CCAFS SLC 40	None None	1	FALSE	FALSE	FALSE		1	0	B1005	-80.5774	28.56186	0
7	04/18/2014	Falcon 9	2296	ISS	CCAFS SLC 40	True Ocean	1	FALSE	FALSE	TRUE		1	0	B1006	-80.5774	28.56186	1

Total Payload Mass

- With a total payload mass (kg) for all flights of 619,967 kg, the ones with NASA flights are 107,010 kg

Average Payload Mass by F9 v1.1

- There are 97 booster versions, with 5 being F9 v1.1, the mean payload mass is 2,928.4 kg and the total payload mass is 14,642 kg

Flight Number (for F9 v 1.1)	Payload Mass (kg)
6	3,170
7	3,325
8	2,296
9	1,316
10	4,535

First Successful Ground Landing Date

- The 1st date of the first successful landing outcome on ground pad is: ‘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.

Landing Outcome	Count
Success	38
No attempt	21
Success (drone ship)	14
Success (ground pad)	9
Controlled (ocean)	5
Failure (drone ship)	5
Failure	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

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

Booster Version	Payload Mass (kg)
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

Total Number of Successful and Failure Mission Outcomes

- Total Number of Successful mission outcomes are 38+14+9 = 61 Successful
- Grouping mission outcomes and counting records for each group led us to the summary above.

Landing Outcome	Count
Success	38
No attempt	21
Success (drone ship)	14
Success (ground pad)	9
Controlled (ocean)	5
Failure (drone ship)	5
Failure	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Boosters Carried Maximum Payload

- The booster versions with the 10 heaviest payload mass are:

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1060.2	15600
F9 B5 B1051.6	15600
F9 B5 B1051.4	15600
F9 B5 B1048.5	15600
F9 B5 B1056.4	15600
F9 B5 B1049.5	15600
F9 B5 B1051.3	15600
F9 B5 B1058.3	15600
F9 B5 B1048.4	15600
F9 B5 B1060.3	15600

2015 Launch Records

- The booster versions and the launch sites for the 2 failed landing outcomes in drone ship for 2015 are:
- Note that there are only 5 failed landing outcomes in drone ship in 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

- Below is the ranking for the count of landing outcomes between the dates 2010-06-04 and 2017-03-20, in descending order

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

Section 3

Launch Sites Proximities Analysis

Jad – Extra Slide

After you plot distance lines to the proximities, you can answer the following questions easily:

- Are launch sites in close proximity to railways?
- Are launch sites in close proximity to highways?
- Are launch sites in close proximity to coastline?
- Do launch sites keep certain distance away from cities?

All Launch Sites

- Replace <Folium map screenshot 1> title with an appropriate title
- Explore the generated folium map and make a proper screenshot to include all launch sites' location markers on a global map
- Explain the important elements and findings on the screenshot

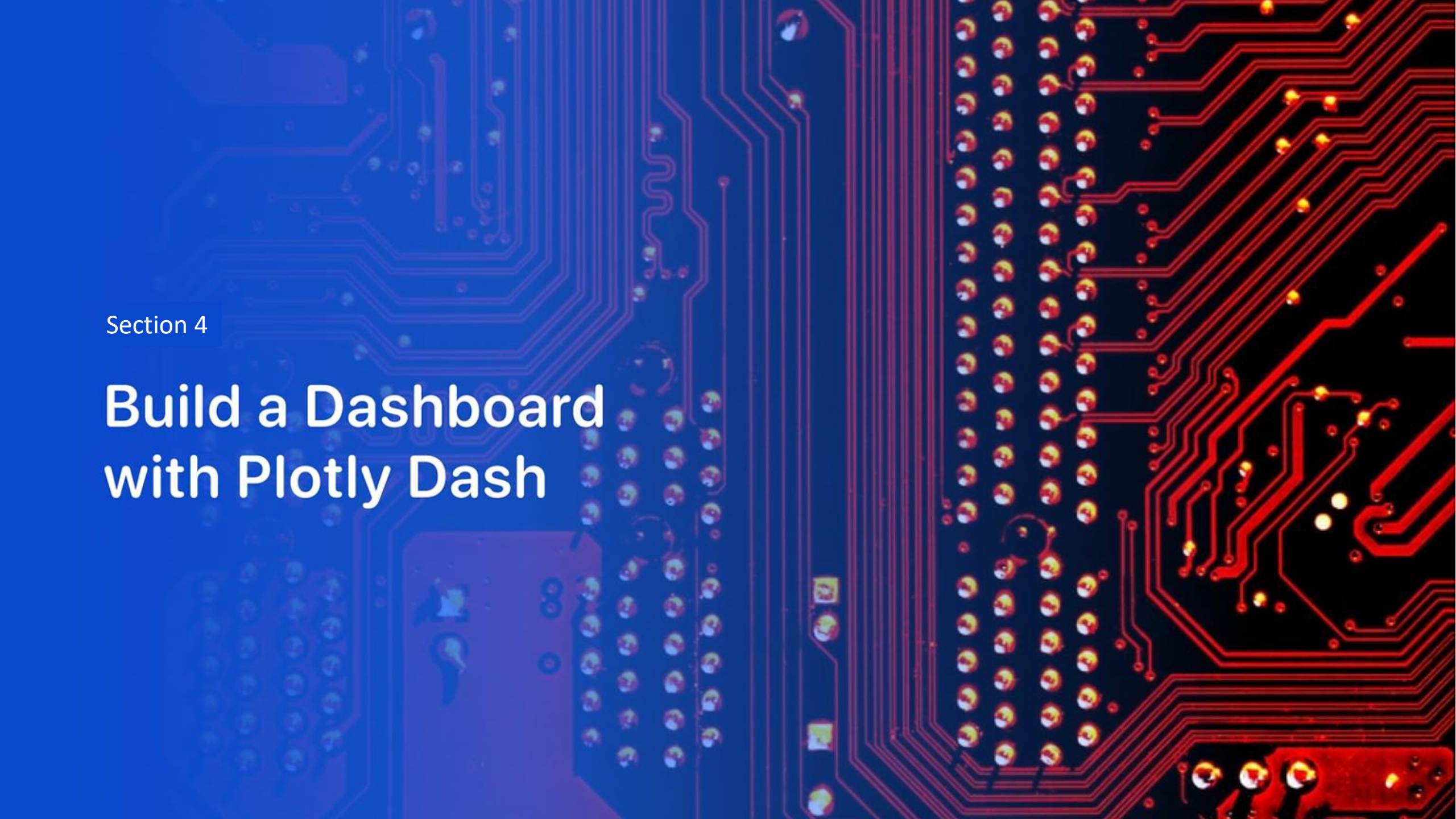
	Launch Site	Lat	Long
0	CCAFS LC-40	28.562302	-80.577356
1	CCAFS SLC-40	28.563197	-80.576820
2	KSC LC-39A	28.573255	-80.646895
3	VAFB SLC-4E	34.632834	-120.610746

Launch Outcome by Site

- Replace <Folium map screenshot 2> title with an appropriate title
- Explore the folium map and make a proper screenshot to show the color-labeled launch outcomes on the map
- Explain the important elements and findings on the screenshot

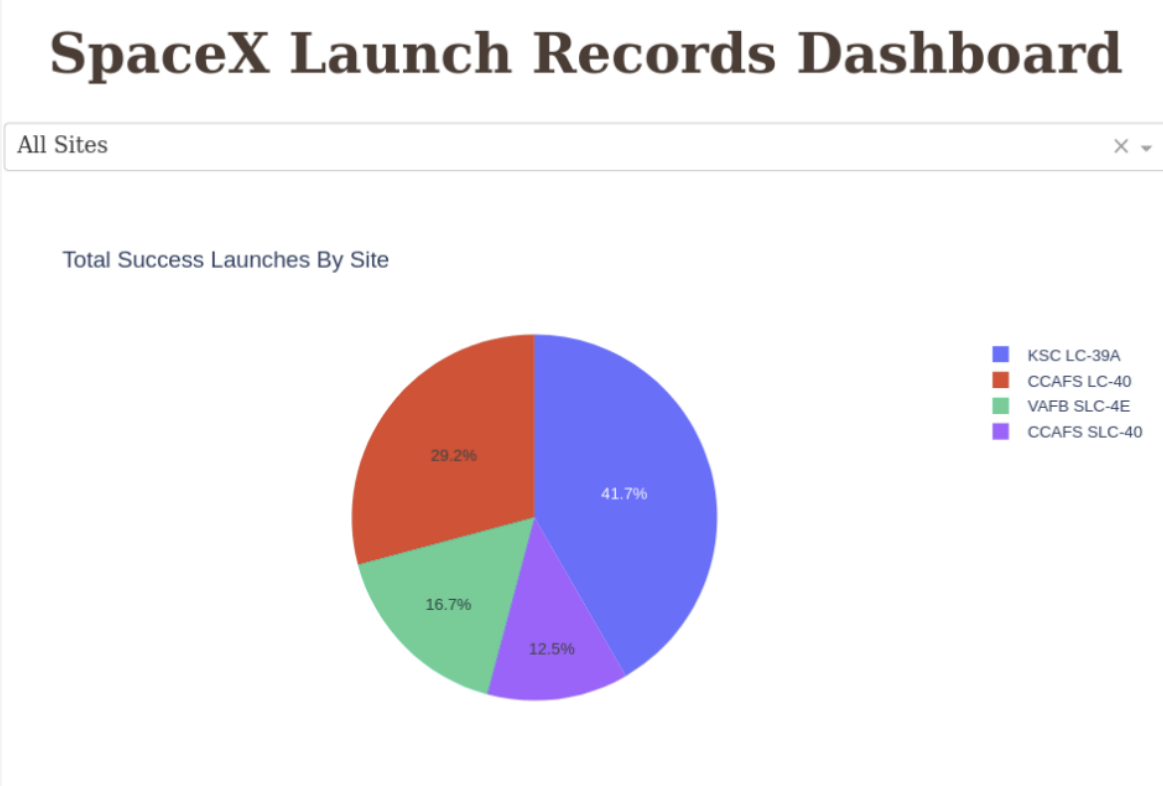
Logistics and Safety

- Replace <Folium map screenshot 3> title with an appropriate title
- Explore the generated folium map and show the screenshot of a selected launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed
- Explain the important elements and findings on the screenshot


Section 4
Build a Dashboard with Plotly Dash

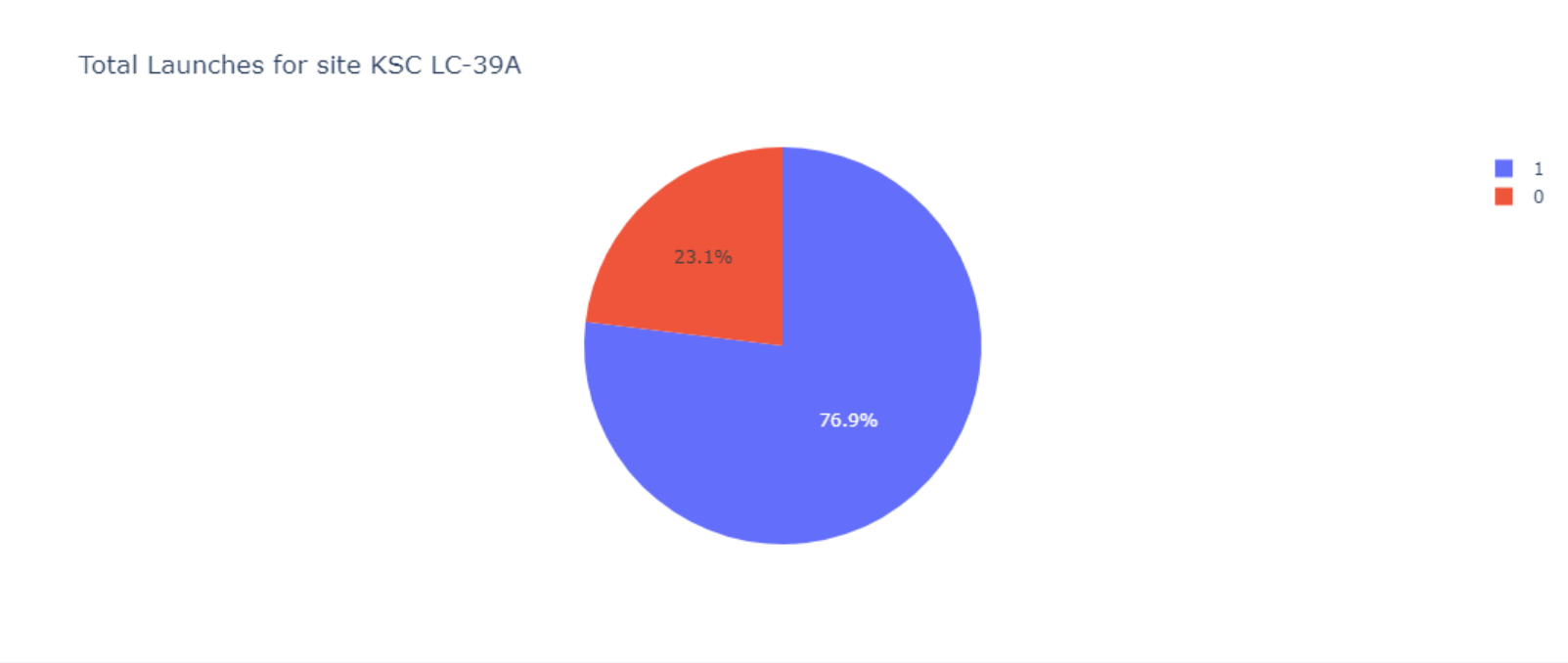
Successful Launches by Site

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



Highest Launch Success Ratio – KSC LC-39A

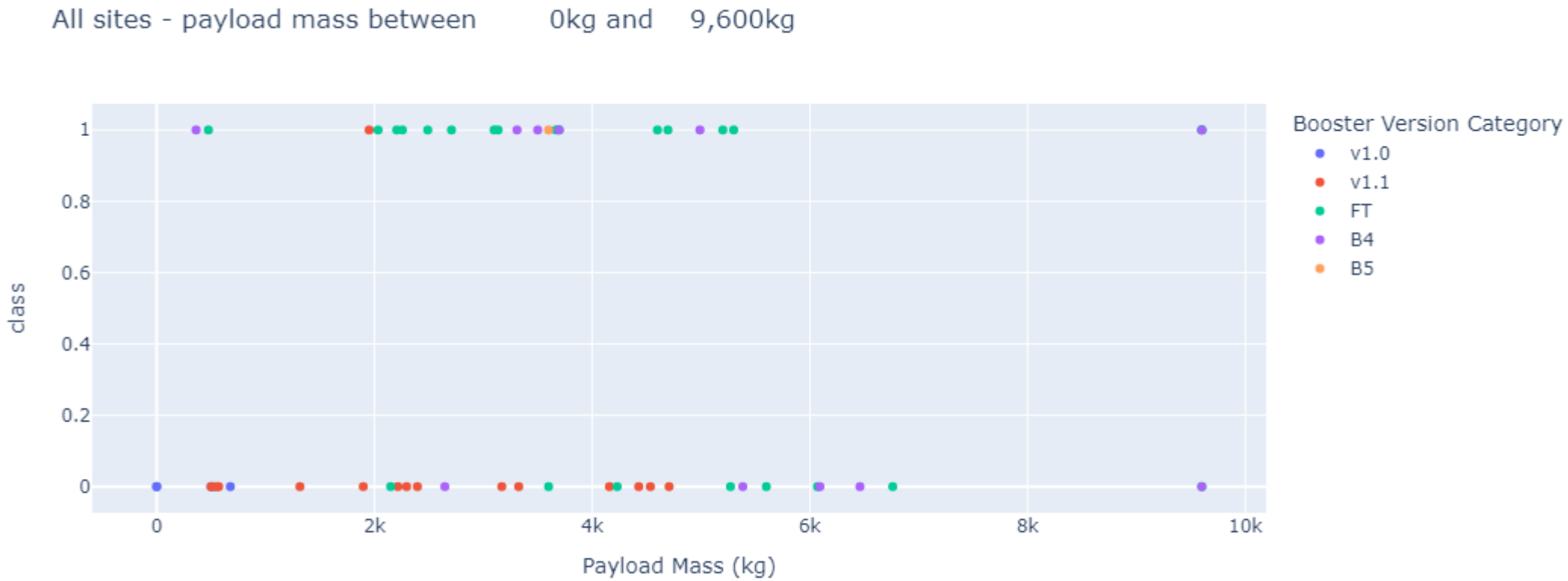
- 76.9% of launches are successful in this site



Launch Site	Success Rate (%)
KSC LC -39A	76.9%
CCAFS LC-40	73.1%
VAFB SLC-4E	60%
CCAFS SLC-40	57.1%

Payload vs Launch Outcome

Payload under 6000 kg for FT Boosters are the most successful

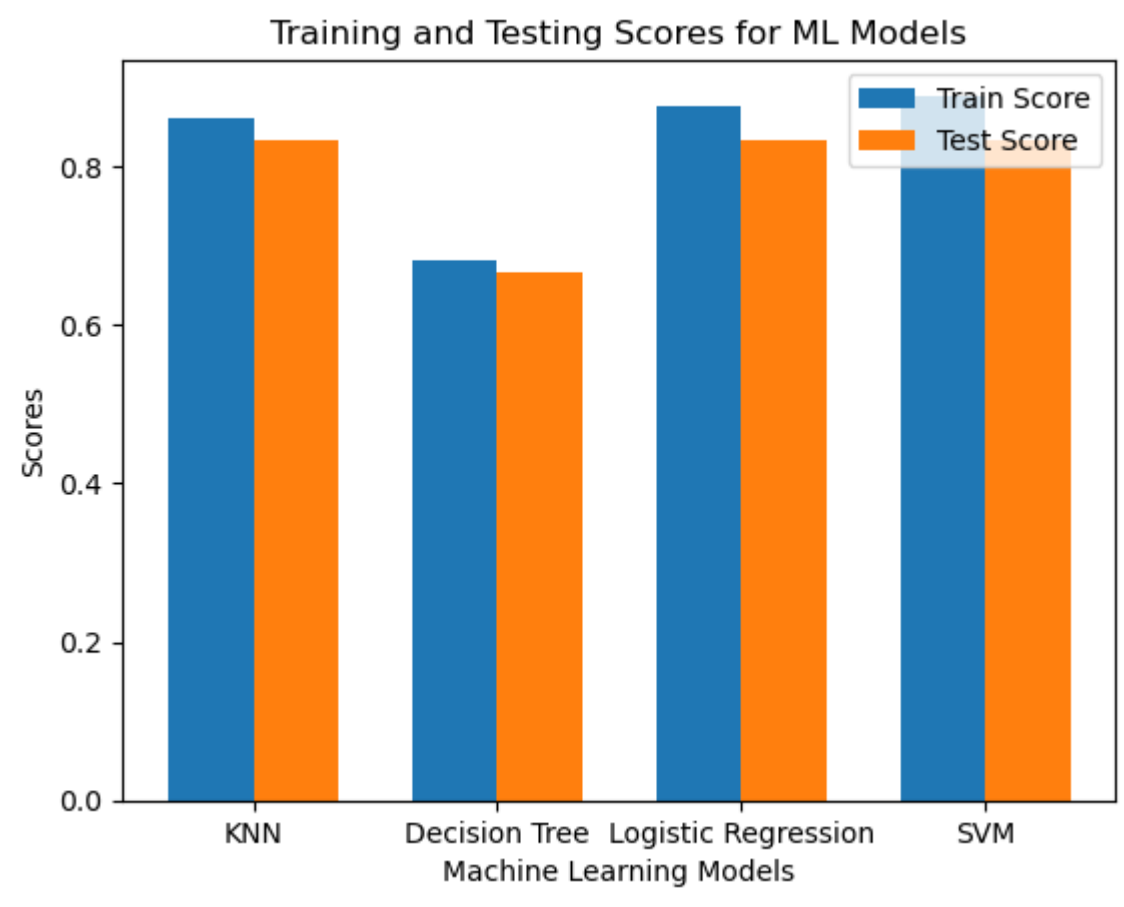


Section 5

Predictive Analysis (Classification)

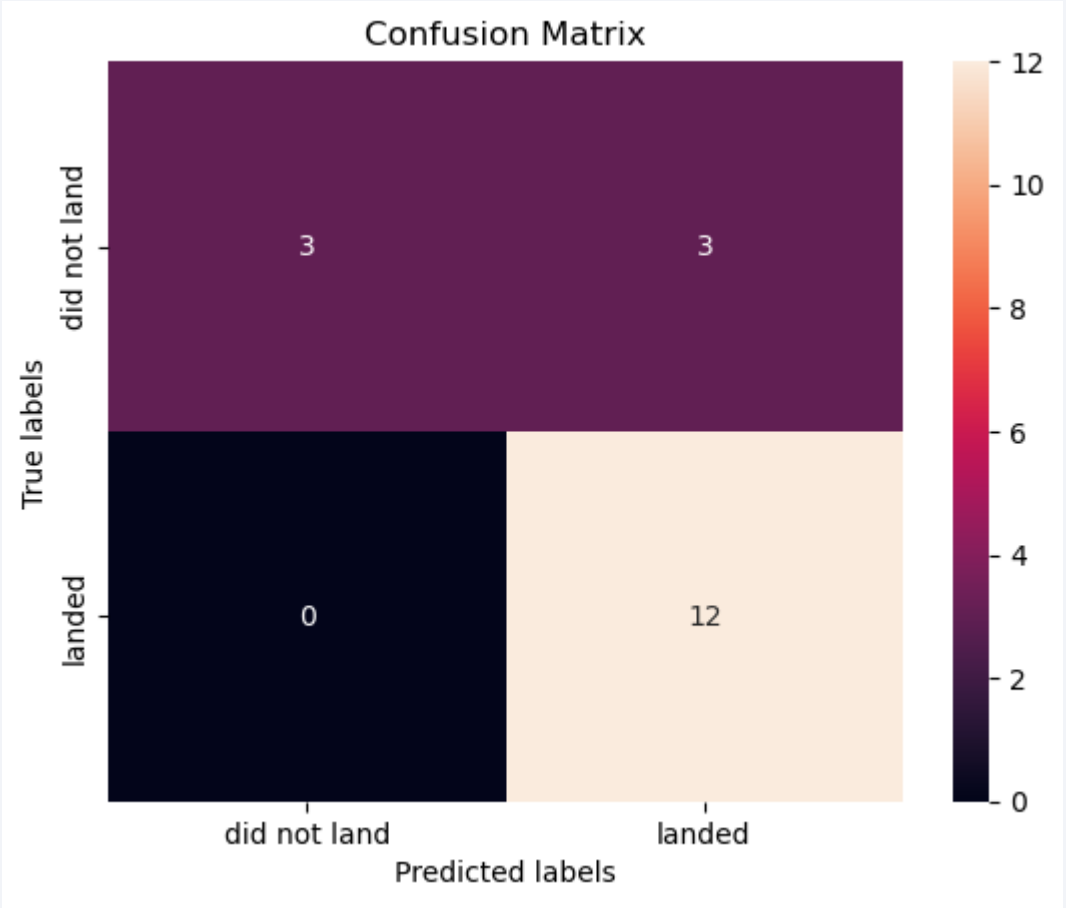
Classification Accuracy

- Four Classification Models were tested and their accuracies are plotted
- The models with the highest classification accuracy is Logistic Regression and SVM



Confusion Matrix of Best Classifiers

- Out of 18 data points for Y test dataset, 12 are True Positives and 3 are True Negatives, with 3 False Positives.



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 to the evolution of processes and rockets
- Decision Tree Classifier can be used to predict successful landings and increase profits

Appendix

- When splitting a dataset into training and testing subsets for machine learning, using a random seed or "random state" can provide reproducibility and consistency in the results. The random state is a parameter in many machine learning libraries, such as scikit-learn in Python, that controls the randomization of the data splitting process.
- When using a fixed random state, the data splitting process will produce the same training and testing sets every time one runs the code, thus having same results.

Thank you!