IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Szymon
27.08.2024

Outline

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

Executive Summary

- Methodologies used:
  - Data Collection
  - Exploratory Data Analysis
  - Feature Engineering
  - Model Developement
- Summary of all results:
  - Launch site and payload mass were the most influential features in predicting landing success.
  - Higher success rates observed for RTLS and ASDS landings compared to ocean landings.
  - Payload mass and launch site location significantly impact landing success.

Introduction

- Project background and context:
  - SpaceX, a pioneering aerospace manufacturer, has revolutionized the space industry with its Falcon 9 rocket, which is partially reusable. This reusability, particularly of the rocket's first-stage booster, significantly reduces the cost of space launches. While SpaceX advertises launch costs at $62 million, competitors often charge upwards of $165 million. The ability to recover and reuse the first stage is a key factor in SpaceX's cost advantage. As the space launch market becomes more competitive, the ability to predict the success of these landings has become increasingly important. Accurate predictions can provide crucial insights for companies looking to compete with SpaceX, enabling them to estimate launch costs and make informed strategic decisions.
- Problems you want to find answers:
  - Can we accurately predict whether the Falcon 9 first stage will land successfully?
  - What factors most significantly influence landing success?
  - How can these predictions be used to drive strategic business decisions?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- The primary data was collected from the Wikipedia page by web scraping: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches using request library and BeautifulSoup.
- Additional data points, such as Booster Version, Launch Site, and other relevant launch parameters, were collected using SpaceX APIs.
- Collection proces:
  - Identify Data Sources (Wikipedia and APIs)
  - Web Scraping (Extract launch data)
  - API Integration (Enrich data with additional factors)
  - Data Preprocessing (Cleaning, Handling Missing Data, Format Standardization)

Data Collection – SpaceX API

- Data were collected with SpaceX REST API calls using request library.

- https://github.com/SzewczykSzy/IBM_DS_Project/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

```
BoosterVersion = requests.get("https://api.spacexdata.com/v4/rockets/"+str(x)).json() ['name']


LaunchSite = requests.get("https://api.spacexdata.com/v4/launchpads/"+str(x)).json()


PayloadData = requests.get("https://api.spacexdata.com/v4/payloads/"+load).json()


CoreData = requests.get("https://api.spacexdata.com/v4/cores/"+core['core']).json()


spacex_url = "https://api.spacexdata.com/v4/launches/past"
```

Data Collection - Scraping

- At first got page title. Then found important table. Table headers were found. For every row in this table data were colected. All data were appended to lists and transformed to dataframe

- https://github.com/SzewczykSzy/IBM_DS_Project/blob/main/jupyter-labs-webscraping.ipynb

```
static_url = "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922"

response = requests.get(static_url)
bs = BeautifulSoup(response.text)
bs.find_all('table')[2]

first_launch_table.find_all('th')

table.find_all("tr")
```

Data Wrangling

- Data were loaded from previous step.
- Identified and calculated the percentage of the missing values in each attribute
- Calculated the number of launches on each site
- Calculate the number and occurrence of each orbit
- Calculate the number and occurance of mission outcome of the orbits
- Create a landing outcome label from Outcome column

- https://github.com/SzewczykSzy/IBM_DS_Project/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

Charts:

- Scatter plot of FlightNumber, PayloadMass and Class hue to see how Payload affect outcome.
- Scatter plot of FlightNumber, LaunchSite and Class hue to see how LaunchSite affect outcome.
- Scatter plot of LaunchSite and PayloadMass and Class hue to investigate if there is some correlation.
- Bar chart to visualize the relationship between success rate of each orbit type.

EDA with Data Visualization

- Scatter plot of FlightNumber and Orbit type (checking correlation)
- Scatter plot of Payload and Orbit type visualize relationship.
- Line plot to visualize percentage of success lauches each year.
- Bar chart to visualize the relationship between success rate of each orbit type.

- https://github.com/SzewczykSzy/IBM_DS_Project/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- DROP TABLE IF EXISTS SPACEXTABLE
- create table SPACEXTABLE as select * from SPACEXTBL where Date is not null
- SELECT DISTINCT(Launch_Site) FROM SPACEXTABLE
- SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5
- SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)'
- SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version LIKE 'F9 v1.1%'
- SELECT MIN(Date) FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)'
- SELECT DISTINCT(Landing_Outcome) FROM SPACEXTABLE

EDA with SQL

- SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000
- SELECT Mission_Outcome, COUNT(Mission_Outcome) FROM SPACEXTABLE GROUP BY Mission_Outcome
- SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
- SELECT substr(Date, 6,2), Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE substr(Date,0,5)='2015' AND Landing_Outcome ='Failure (drone ship)'
- SELECT Landing_Outcome, COUNT(Landing_Outcome) FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome Order by COUNT(Landing_Outcome) DESC

- https://github.com/SzewczykSzy/IBM_DS_Project/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- I have created and addd `folium.Circle` and `folium.Marker` for each launch site on the site map. Then added the launch outcomes for each site, to see which sites have high success rates. To do this i have used MarkerCluster() and then added Marker()'s. Also added MousePosition() to get Lattitude and Longitude of exact places. By using that i could calculate the distance between places and it to map by Marker() and draw a line by PolyLine() between those places.

- https://github.com/SzewczykSzy/IBM_DS_Project/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- At first i have added a pie chart to show the total successful launches count for all sites. If a specific launch site was selected, the chart shows the Success vs. Failed counts for the site.
- The second graph is a scatter chart to show the correlation between payload and launch success.

- https://github.com/SzewczykSzy/IBM_DS_Project/blob/main/spacex_dash_app.py

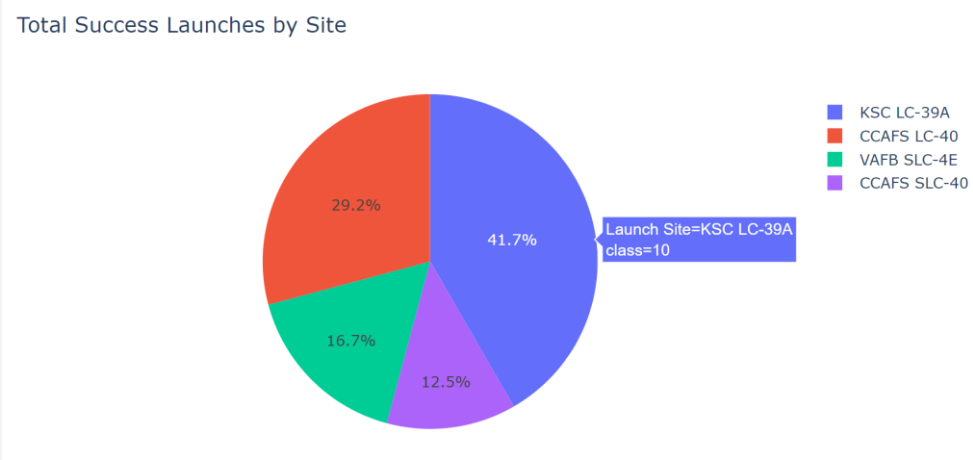
Predictive Analysis (Classification)

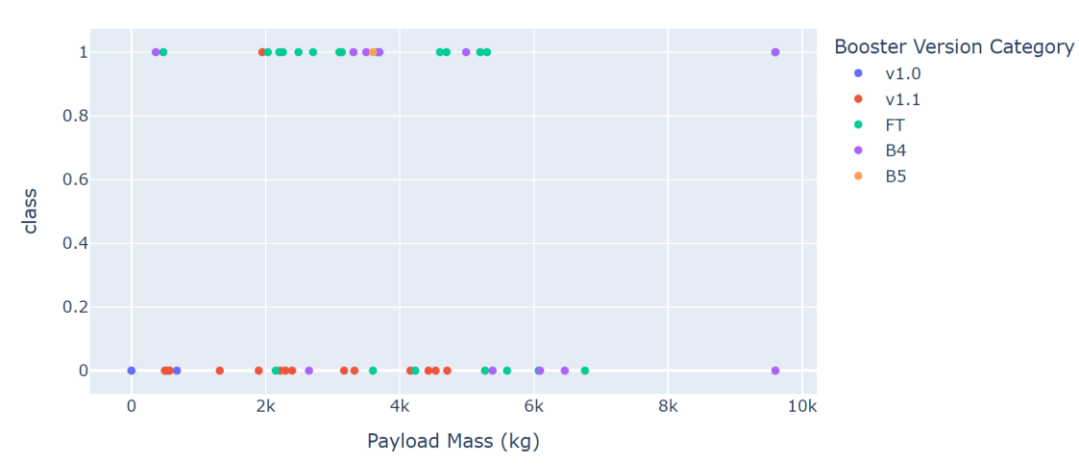
- At first I have created X dataframe and Y array. Then standarized the X by StandardScalers() and divided dataset into train and test sets.
- I have created for differents models: LogisticRegression(), SVC(), DecisionTreeClassifier() and KNeighborsClassifier()
- For each model I have created parameters dictionary and use it to find the best set of parameters by using GridSearchCV() with ``cv=10``.
- I have fitted models on train data checked accuracy on train set and evaluated them on test set. Most of the models get 83.33% accuracy on the test set. DecisionTreeClassifier got worst 72.22% accuracy. SVC and KNN got the best accuracy on train set: 84,82%.
- https://github.com/SzewczykSzy/IBM_DS_Project/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

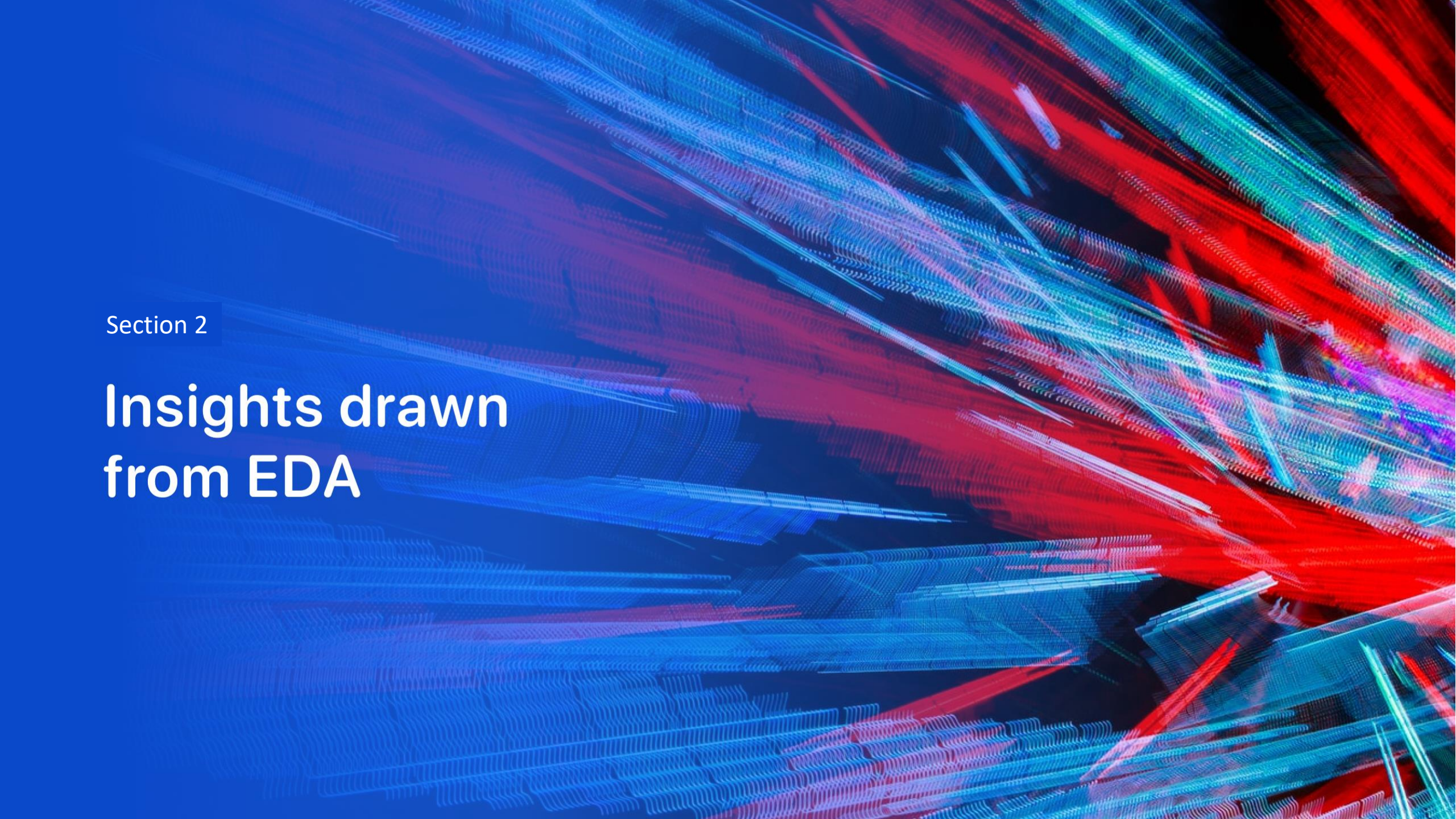
Results

The analysis revealed that the success of Falcon 9 landings is significantly influenced by factors such as the launch site, payload mass, and the type of landing attempt (RTLS, ASDS, Ocean).

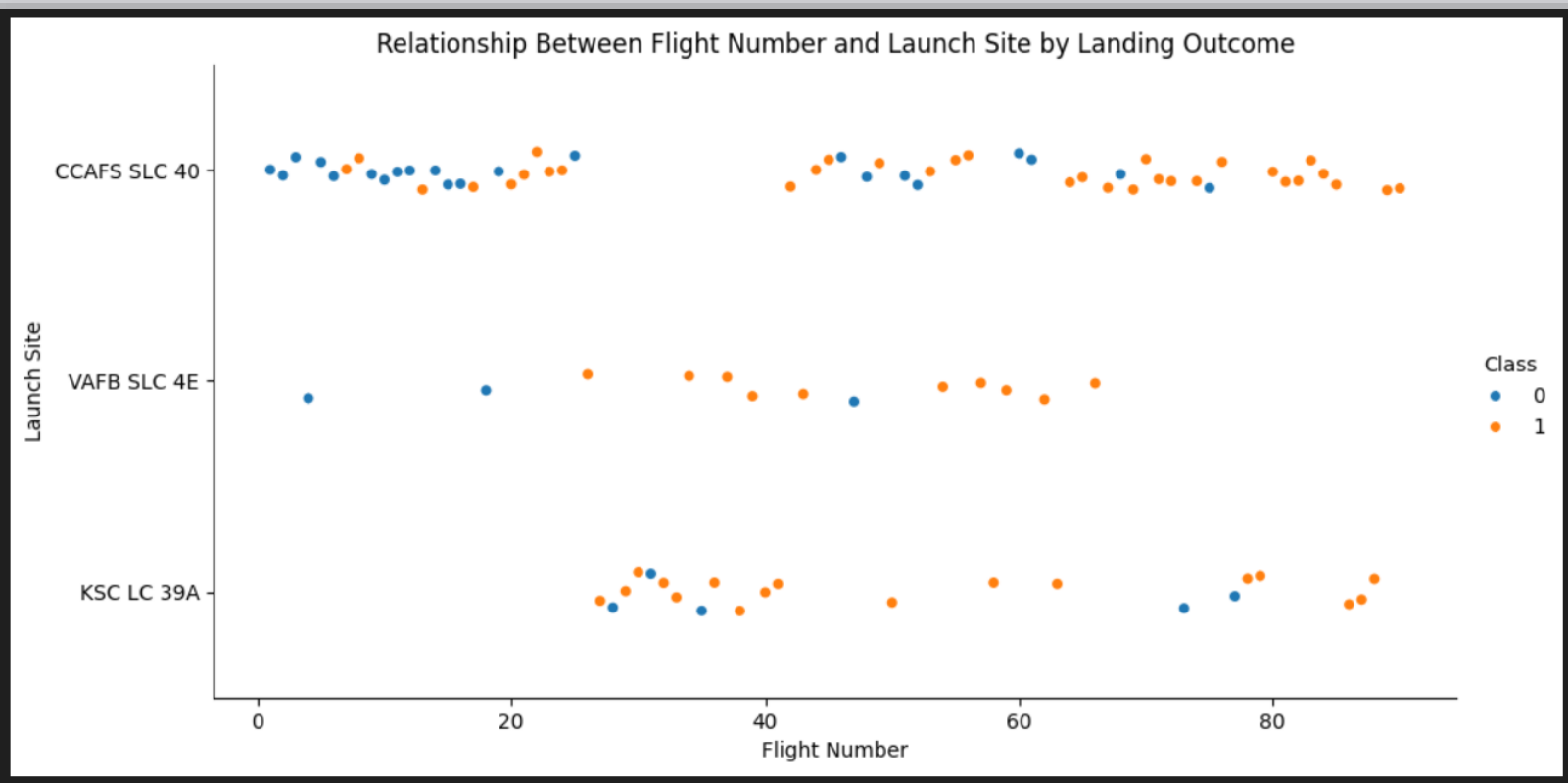
All models had the same accuracy of about 83%. Each of the models had a problem with False Postive predictions - 3 were classified wrongly





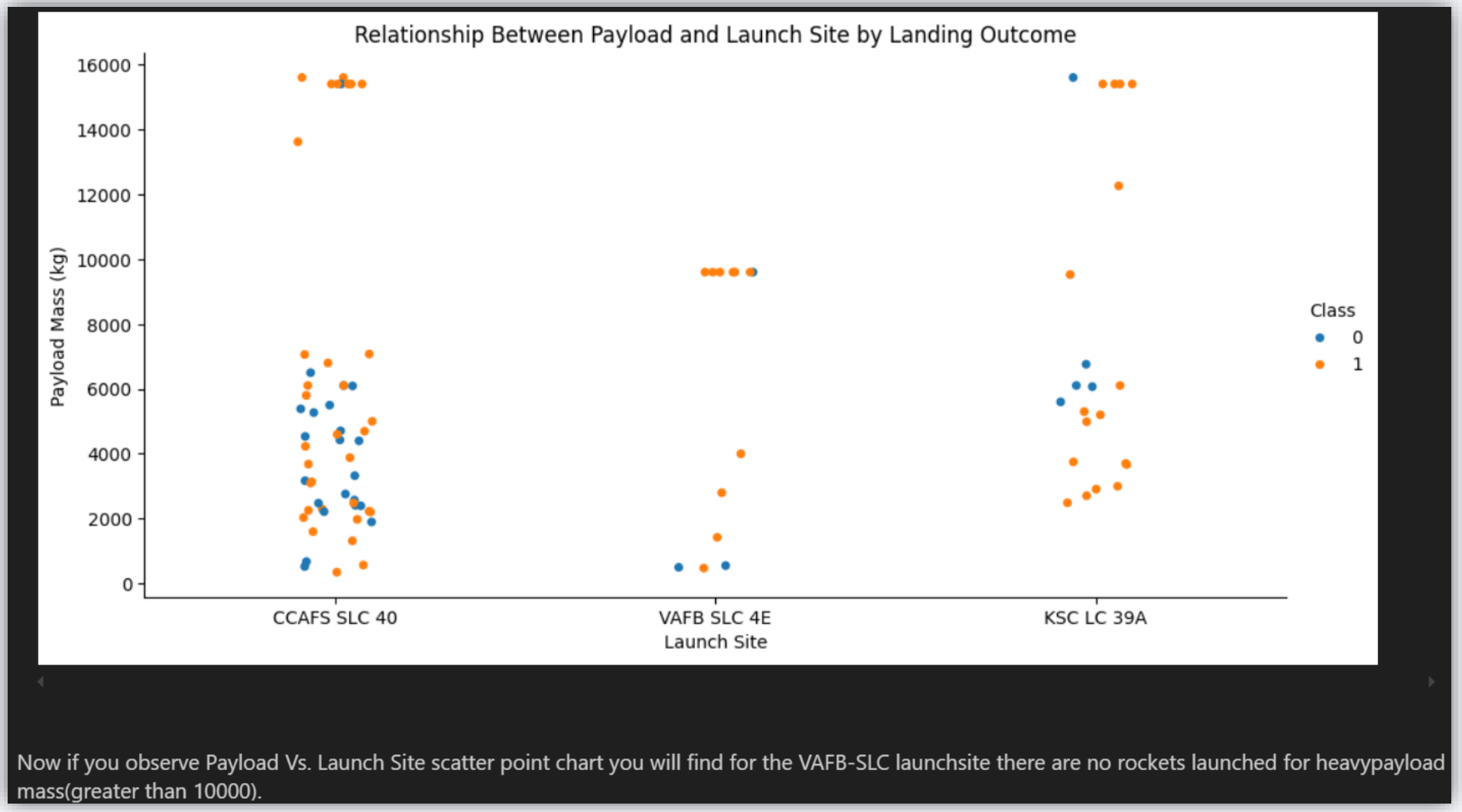

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



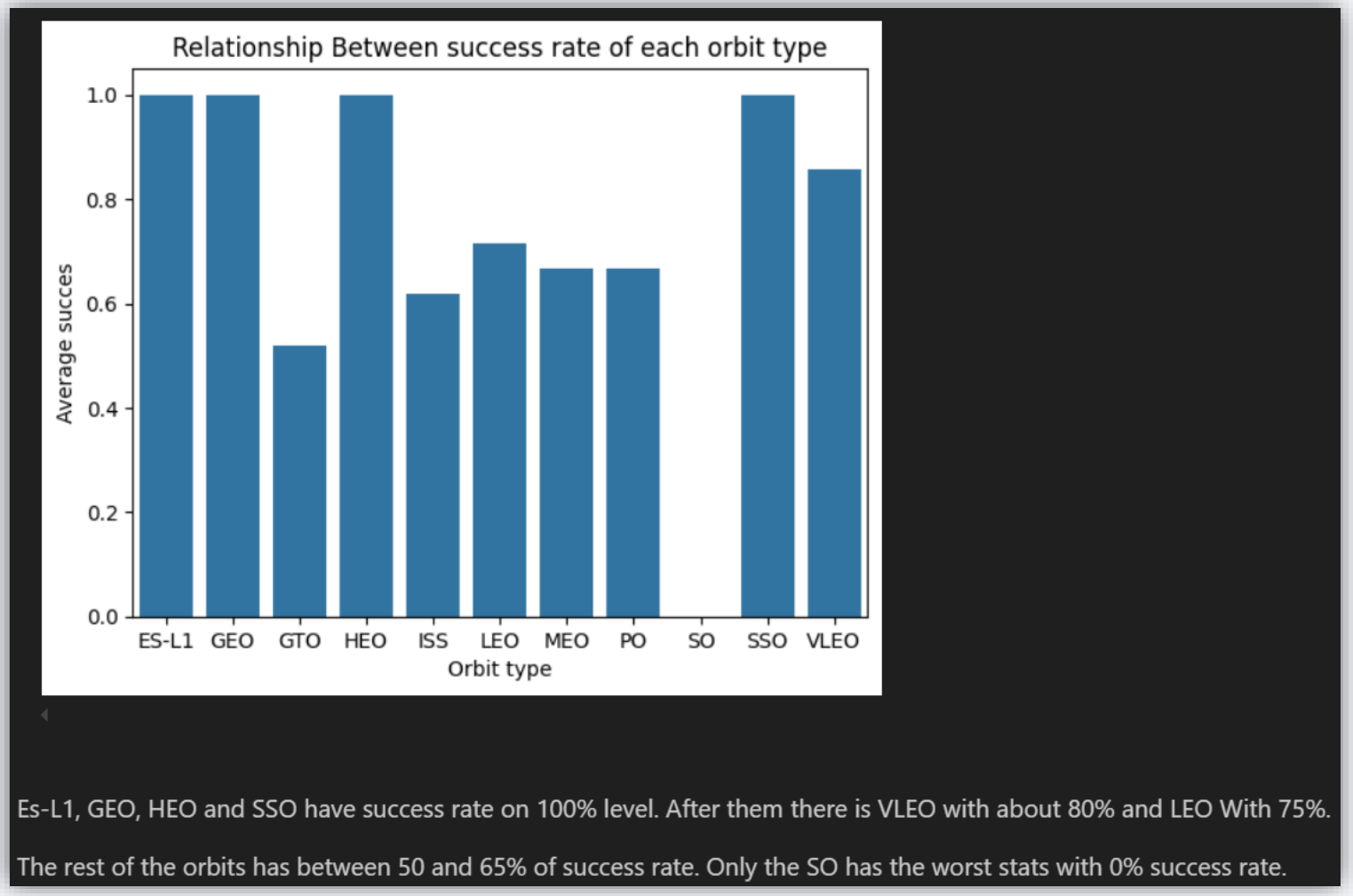
Across all sites, earlier flight numbers (especially below 20) show more variability in landing outcomes, indicating a learning curve as SpaceX refined its technology and procedures. Each launch site has its unique trajectory and payload constraints, which may influence the success rates. For instance, KSC LC 39A shows a strong improvement trend, likely due to refinements in landing technology and site-specific adjustments. There's a clear trend of increasing landing success over time (higher flight numbers), which aligns with SpaceX's reported improvements in technology and operational efficiency.

Payload vs. Launch Site



Now if you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

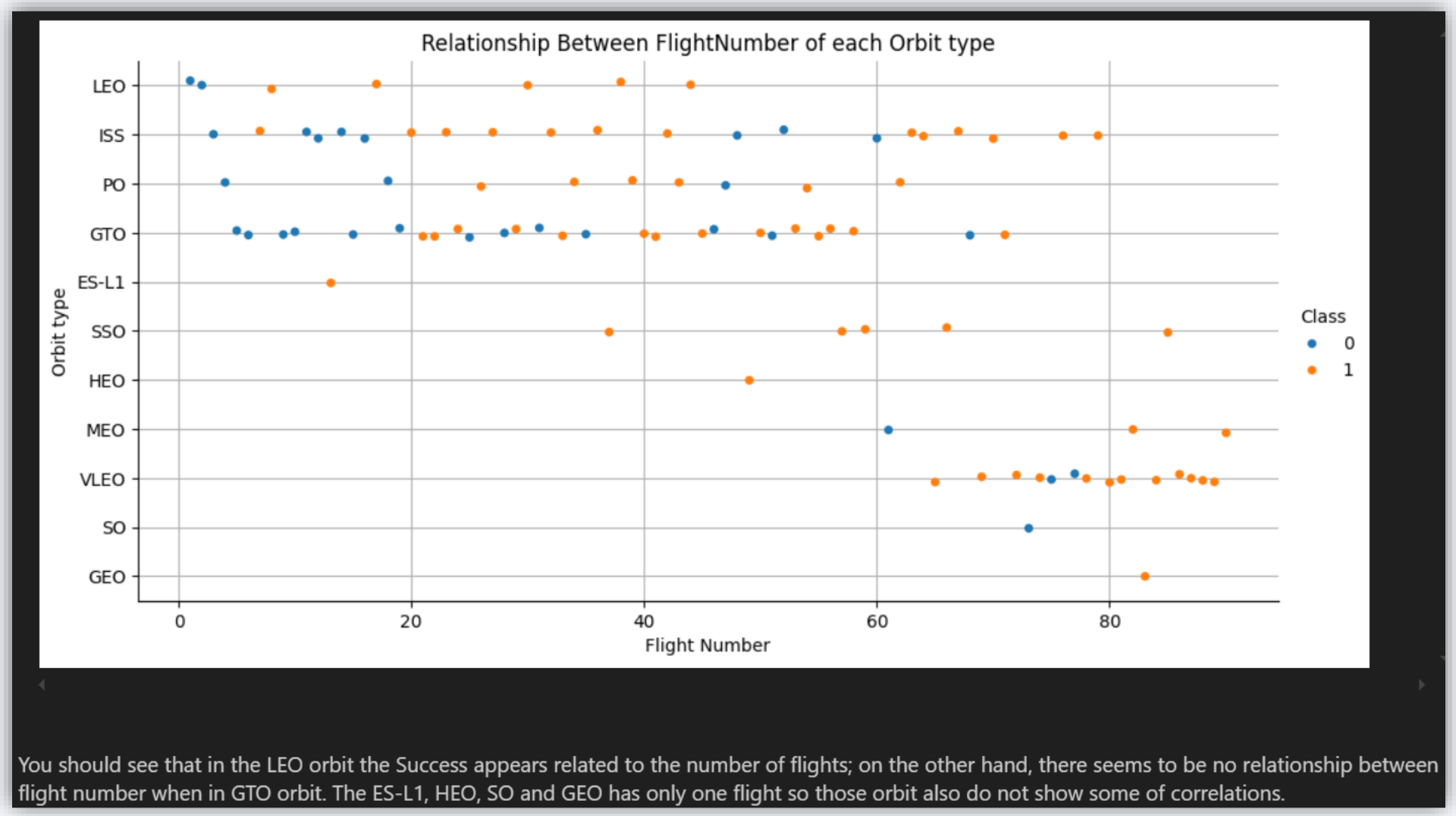
Success Rate vs. Orbit Type



Es-L1, GEO, HEO and SSO have success rate on 100% level. After them there is VLEO with about 80% and LEO With 75%.

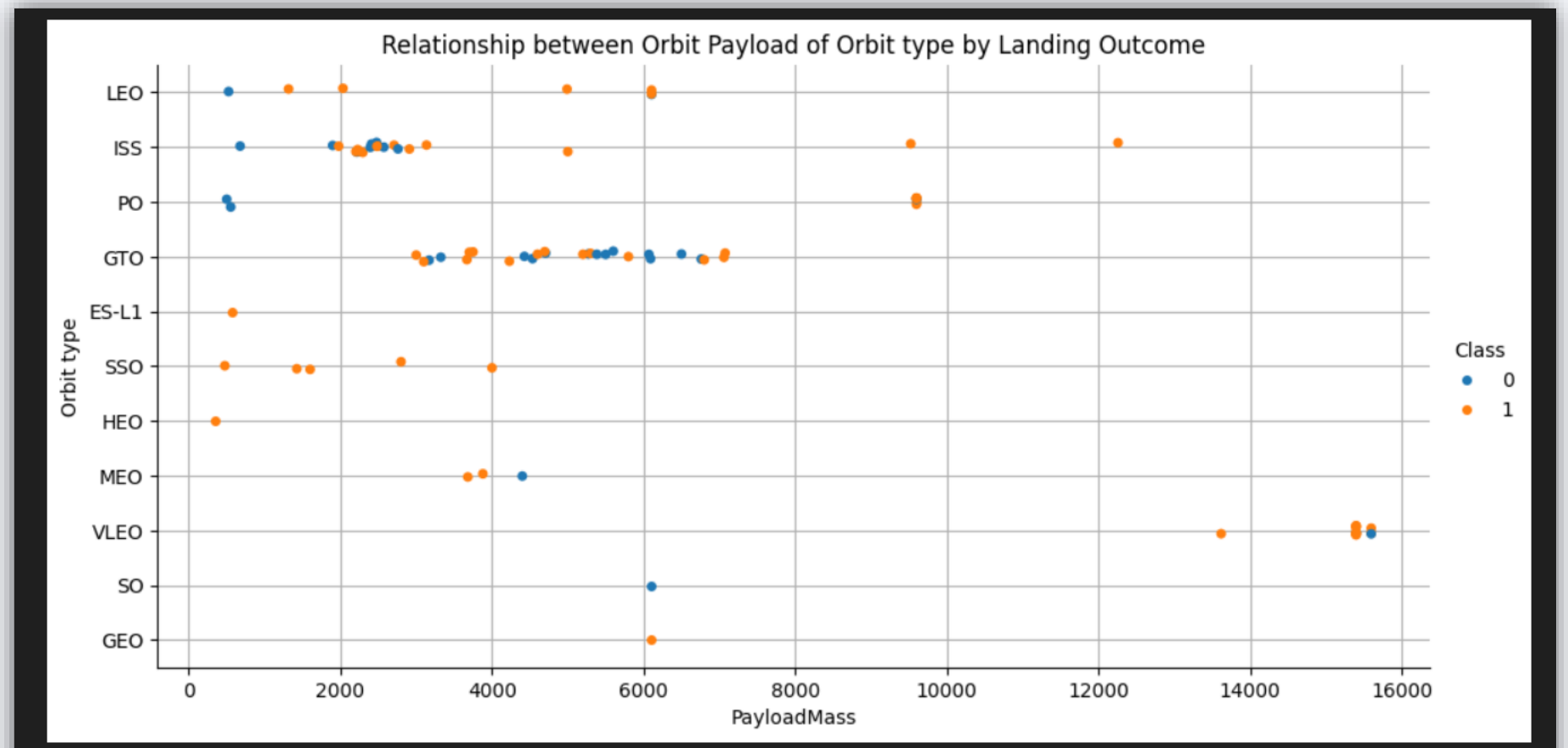
The rest of the orbits has between 50 and 65% of success rate. Only the SO has the worst stats with 0% success rate.

Flight Number vs. Orbit Type



You should see that 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. The ES-L1, HEO, SO and GEO has only one flight so those orbit also do not show some of correlations.

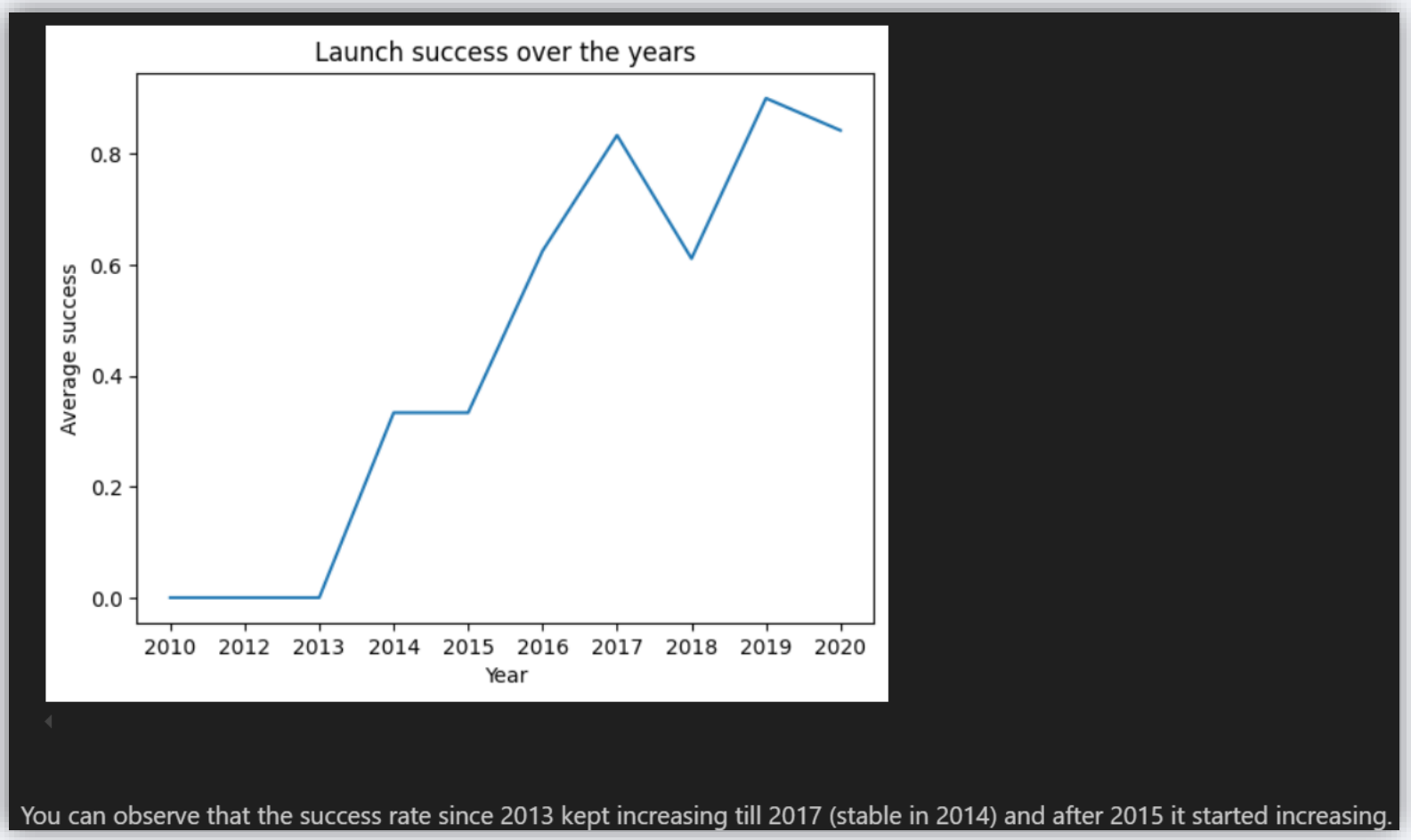
Payload vs. Orbit Type



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



All Launch Site Names

- SELECT DISTINCT(Launch_Site) FROM SPACEXTABLE
- There are only 4 unique (distinct) Launch Sites:
  - CCAFS LC-40
  - VAFB SLC-4E
  - KSC LC-39A
  - CCAFS SLC-40

Launch Site Names Begin with 'CCA'

- SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

- LIKE expression is used to get results with some subsequences.

Total Payload Mass

- SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)'

SUM(PAYLOAD_MASS__KG_)
45596

- Using SUM function with WHERE expression (limitation to ‚NASA (CRS)')

Average Payload Mass by F9 v1.1

- SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version LIKE 'F9 v1.1%’

- Usage of AVG function with LIKE expression (limitation to , 'F9 v1.1’)

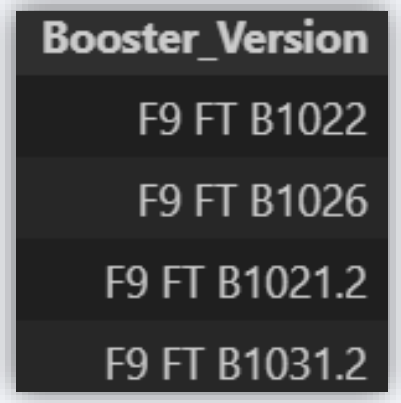
First Successful Ground Landing Date

- SELECT MIN(Date) FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)'

- Finding the first date using MIN Function restricted to Succes ground pad.

Successful Drone Ship Landing with Payload between 4000 and 6000

- SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000



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

- Using WHERE expression with two AND.

Total Number of Successful and Failure Mission Outcomes

- SELECT Mission_Outcome, COUNT(Mission_Outcome) FROM SPACEXTABLE GROUP BY Mission_Outcome

Mission_Outcome	COUNT(Mission_Outcome)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- Using GROUP BY statement to group Mission Outcomes and using COUNT function to count Outcomes.

Boosters Carried Maximum Payload

- SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
- Using WHERE expression with nested SELECT query (max payload)

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

- SELECT substr(Date, 6,2), Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE substr(Date,0,5)='2015' AND Landing_Outcome ='Failure (drone ship)'

substr(Date, 6,2)	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- Substr(Date, 6, 2) – month number. Using WHERE expression with substr() function and AND.

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

- SELECT Landing_Outcome, COUNT(Landing_Outcome) FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome Order by COUNT(Landing_Outcome) DESC

- Using GROUP BY statement to group by Landing Outcome, WHERE expression to filter results by date and ORDER BY COUNT() DESC to get final result.

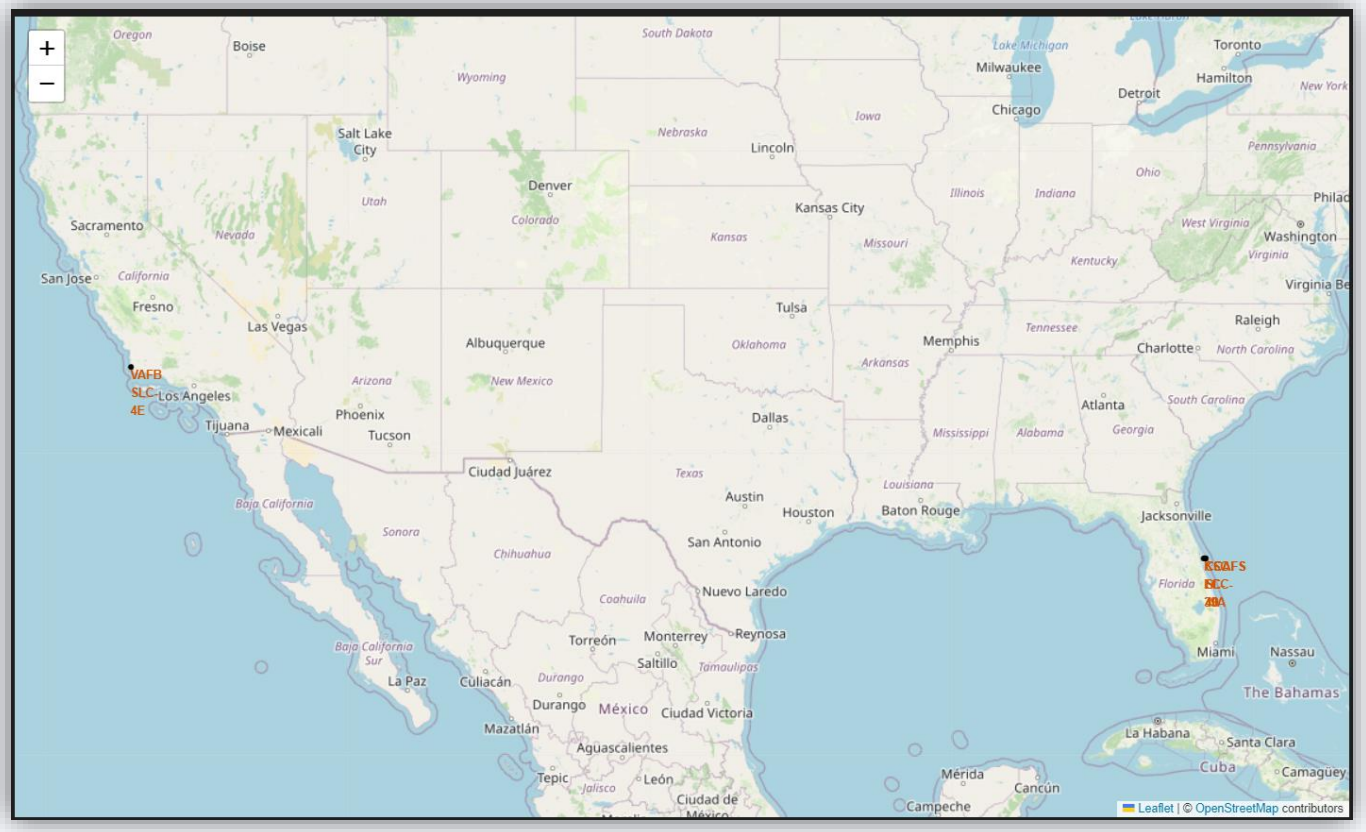
Landing_Outcome	COUNT(Landing_Outcome)
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

Launch sites' locations

The Lauch sites globally take place it two regions. One is on the west coast of the USA and the secont one is on the east cost.



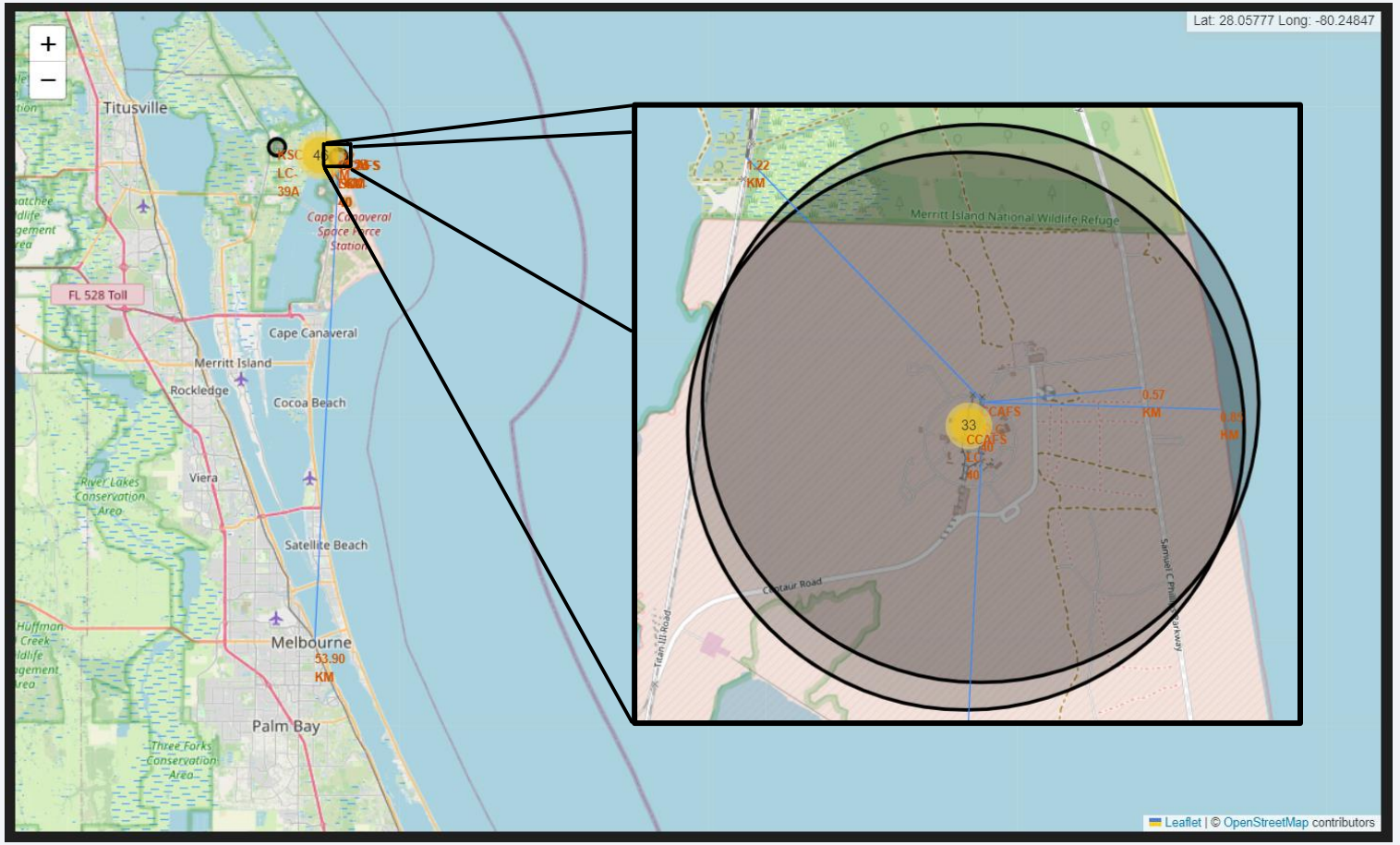
Color-labeled launch outcomes

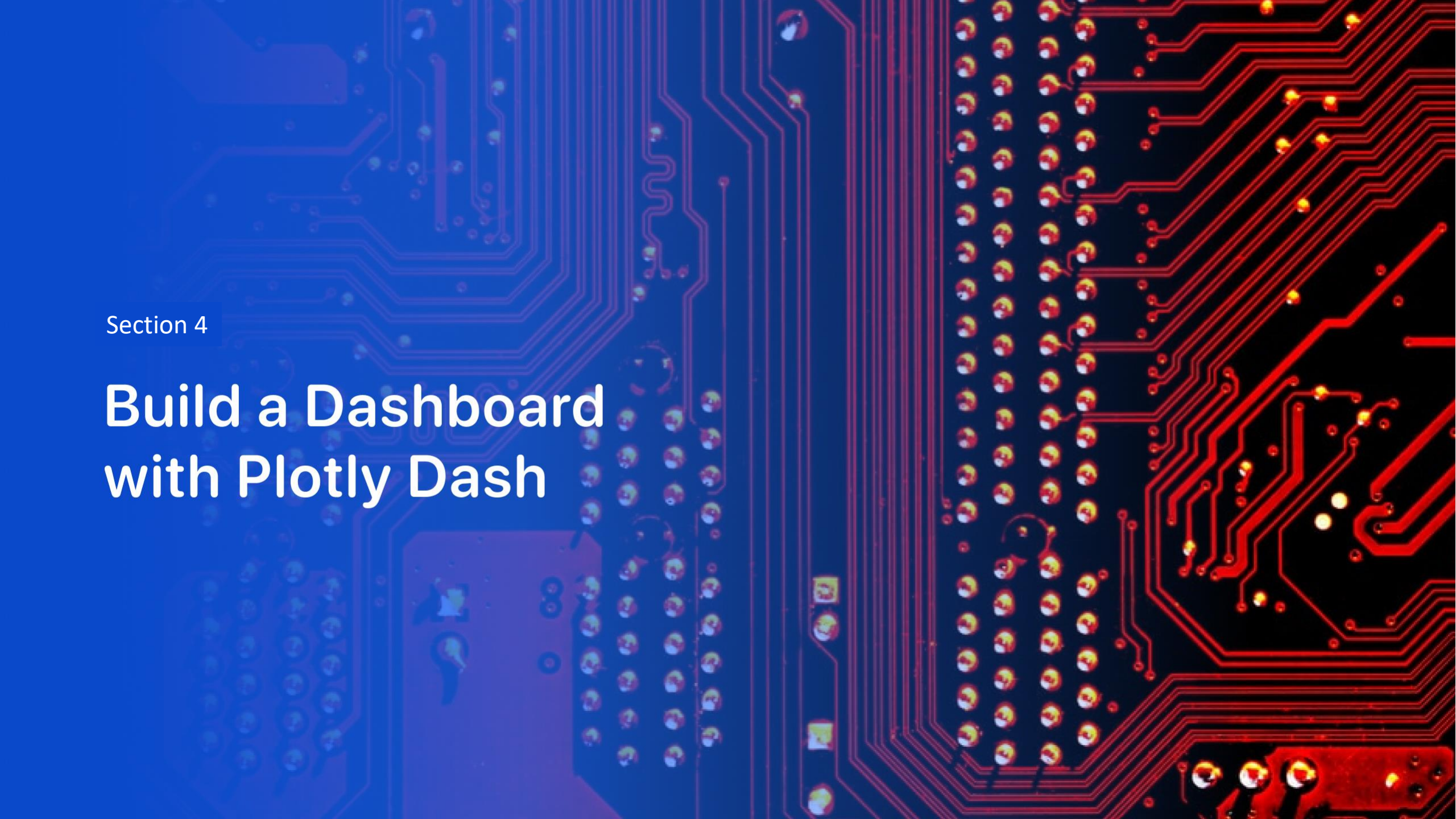
The map shows one area of launch sites. It represand the launchase with color-labeled marks represanting outcome:

- Green – success
- Red – fail

Distance to railway, highway and coastline

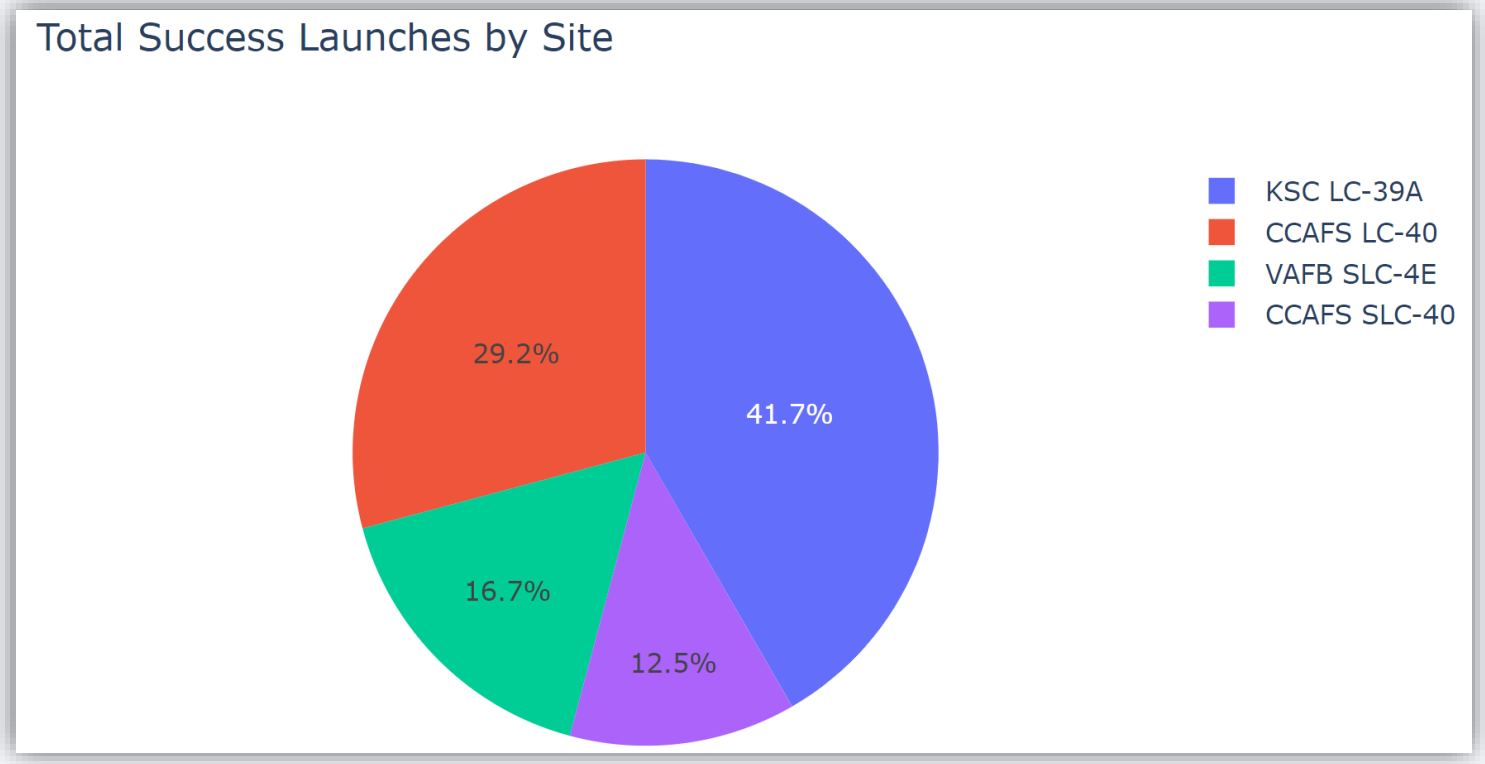
- Map shows distances to the nearest places:
  - City – Melbourne 53.9 km
  - Railway – 1.22 km
  - Highway – 0.57 km
  - Coastline – 0.85 km




Section 4
Build a Dashboard with Plotly Dash

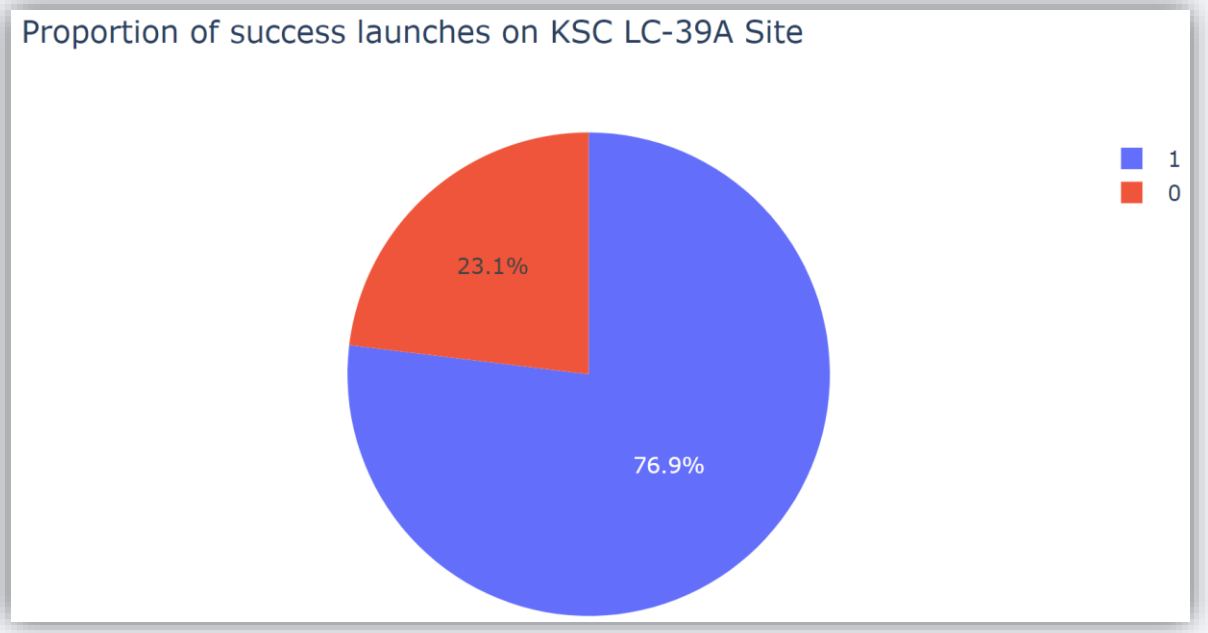
Launch success count for all sites

The most success launches made from KSC LC-39A (41.7%) and the least from CCAFS SLC-40 (12,5%).

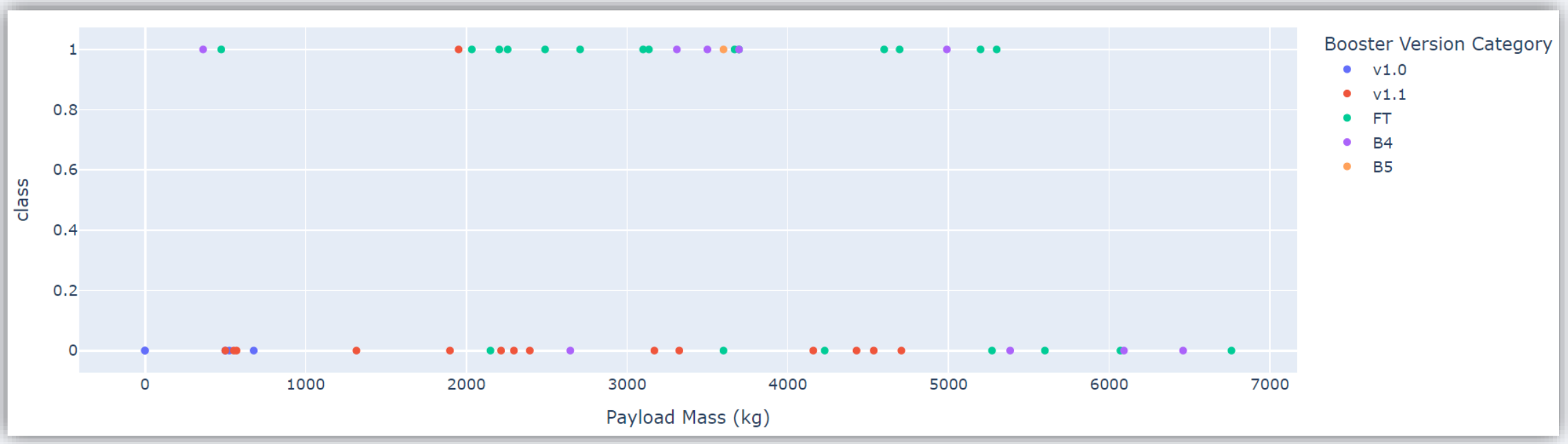


Highest launch success ratio

The most successful launch site was KSC LC-39A with 76.9% of success launches.



Payload vs. Launch Outcome



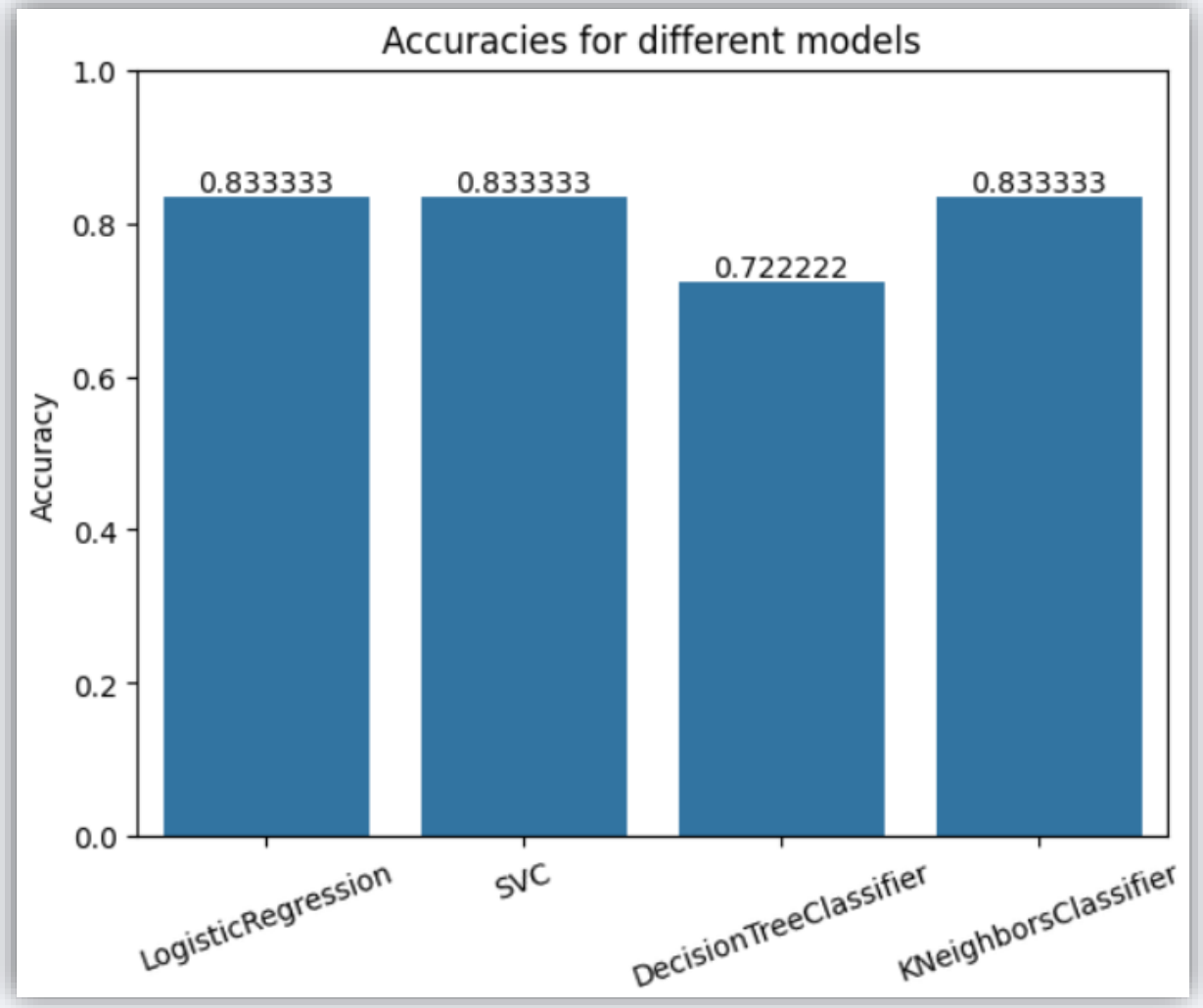
- The highest success ratio has B5 Booster, but it contains least of all launches. At second place there could be FT and B4. At the Payload range (0-4000) the most successfull is FT.

Section 5

Predictive Analysis (Classification)

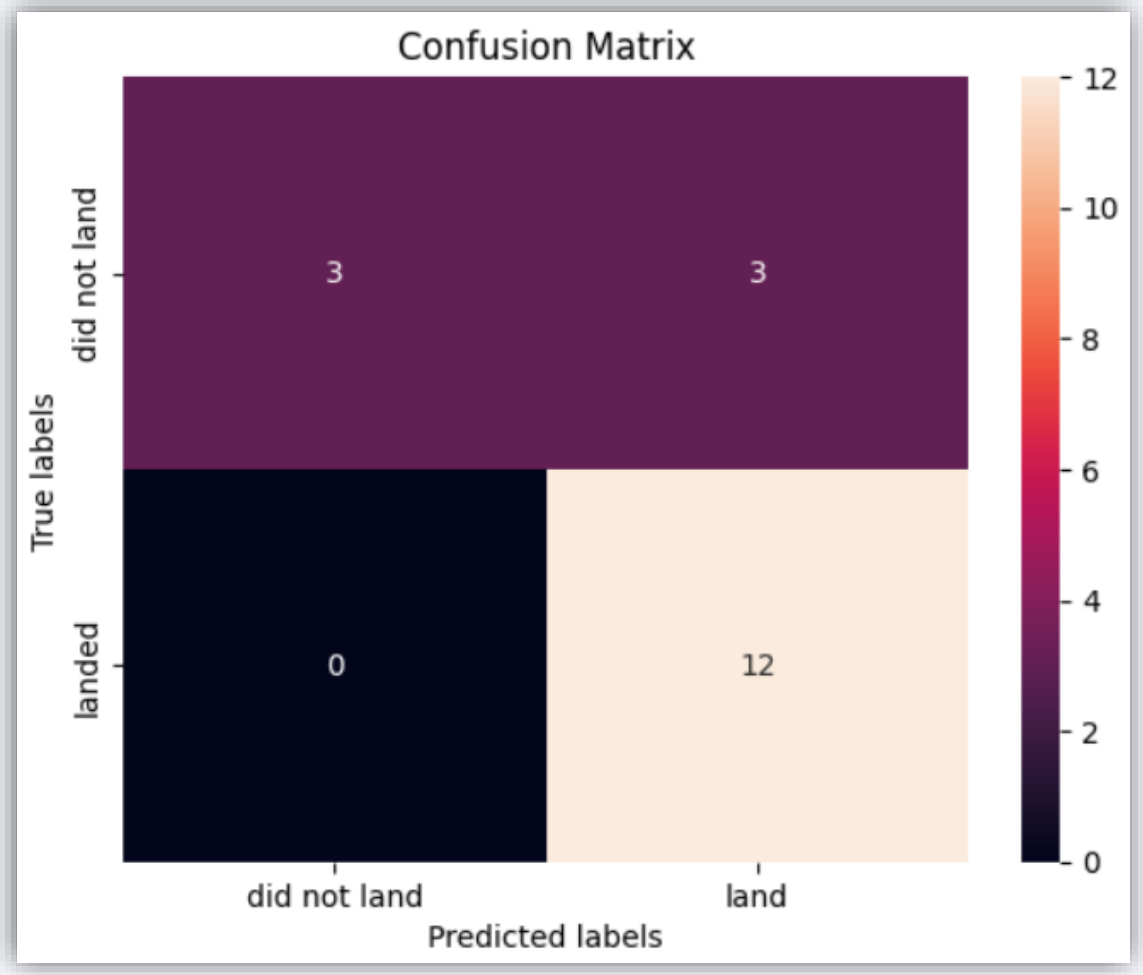
Classification Accuracy

Three out of four models has highest validaion accuracy 83.33%. When it comes to training accuracy the differences are small, but the best is DecissionTree (86%) which could indicate overfitting.



Confusion Matrix

Ale the models with 83.33% validation accuracy made the same mistakes. There is one problem of False Postives – 3 (True label is not landed, Predicted label is landed)



Conclusions

In conclusion, this project demonstrates that the success of SpaceX Falcon 9 landings can be effectively predicted by analyzing key factors such as launch site, payload mass, and the type of landing attempt (RTLS, ASDS, Ocean). The models built showed a consistent accuracy of around 83%, though they encountered challenges with false positive predictions. The insights gained through this analysis are vital for stakeholders in the aerospace industry, as they enable better decision-making processes, particularly in estimating launch costs and strategizing for competitive advantage in the space launch market.

Appendix

The appendix contains various supplementary materials that were crucial for the development and validation of this project. This includes Python code snippets, SQL queries, detailed data visualizations, and outputs from Jupyter Notebooks used throughout the project. Key assets included in this appendix:

- Python scripts for data collection via web scraping and API integration.
- SQL queries used for data preprocessing and exploratory data analysis.
- Data visualizations such as scatter plots, bar charts, and interactive maps created using Plotly and Folium.
- Details of the machine learning models used, including their respective parameters, training, and validation processes.
- Confusion matrices showcasing the performance of each model.

These resources provide a comprehensive understanding of the methodology and tools utilized, ensuring the reproducibility of the study and offering a deeper dive into the technical aspects for interested parties.

Thank you!