IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Linh Tong
08 September 2023

Outline

EXECUTIVE SUMMARY

INTRODUCTION

METHODOLOGY

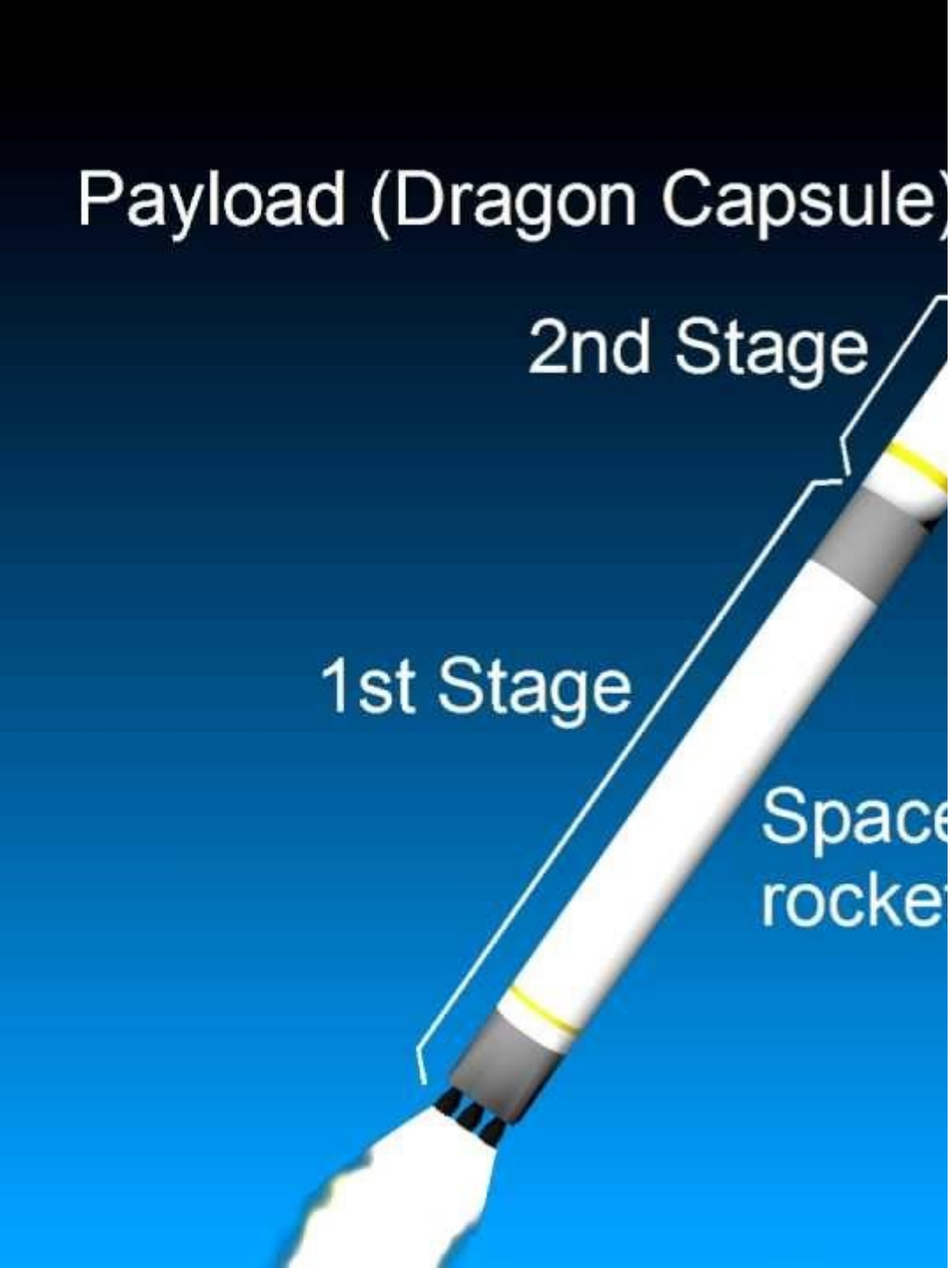
RESULTS

CONCLUSION

APPENDIX

Executive Summary

- Methodologies: data collection, data wrangling, exploratory data analysis (EDA) using visualization and SQL, interactive visual analytics using Folium and Plotly Dash, predictive analysis using classification models
- Results:
  - Launches with higher Payload Mass and Flight Number seemed to have better successful landing rate. However, in case of Payload, it may related to orbit type since only ISS, PO, VLEO have launches of higher load. In addition, VLEO flights all have over 60 attempts which results in good landing rate
  - Launch success yearly increased since 2013 to 2020
  - All 4 F9 Falcon lauch sites are located: near coastline, earth equator, railroad, highway but far away from cities for making use of the earth velocity, safety and effective transportation
  - Out of 4 algorithms tested, Decision Tree performed the best on train set. Nevertheless, the accuracy on the test set is the same for all model



Introduction

- Backgroud and context: SpaceX is a aerospace manufacturer and space transport services company founded by Elon Musk. The company develops the Falcon 9 rocket, which is capable of launching both crew and cargo to the International Space Station (ISS) with lower cost due to the resuse of the first stage.
- Questions:
  - Can we predict if the Falcon 9 first stage will land successfully?
  - If so, what factors affect the likelihood of a successful landing?

Section 1

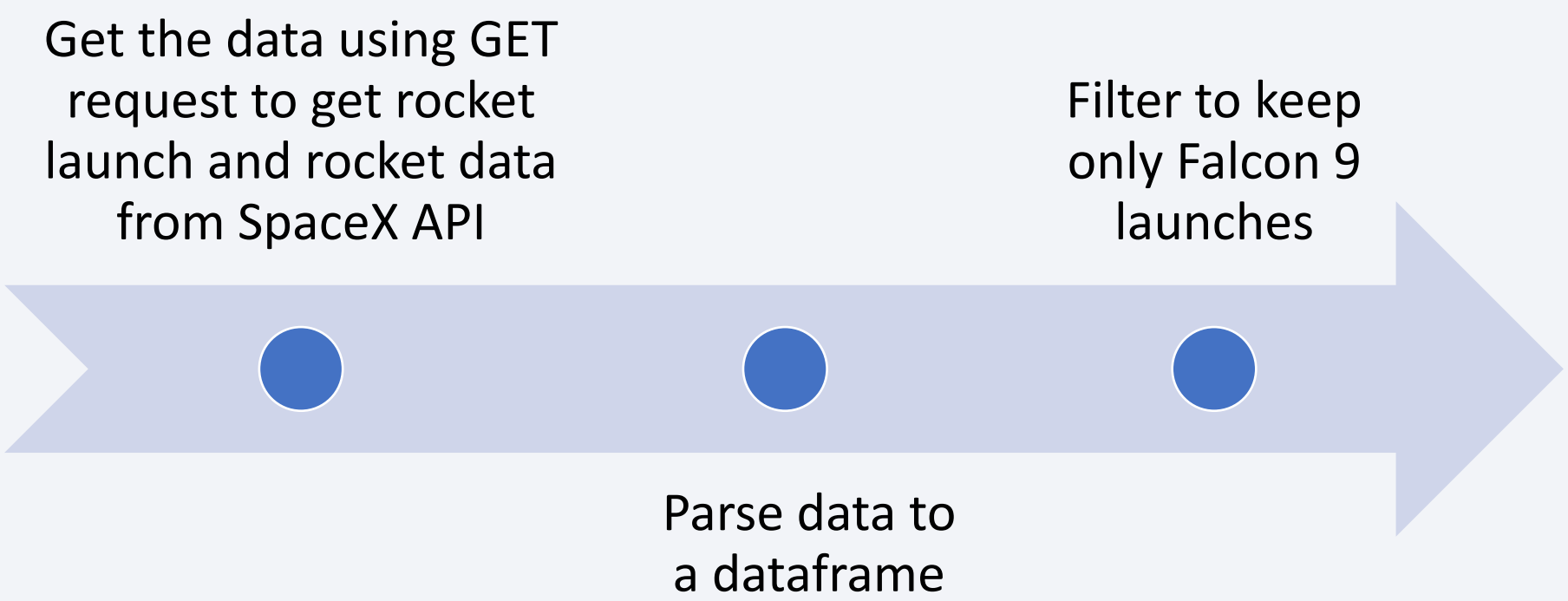
Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected by sending requests to SPACEX API and parse as dataframe
- 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
  - 4 algorithms was used: Decision Tree, k-Nearest Neighbors, Support Vector Machine and Logistic Regression, parameters was tune using GridSearchCV, the models was evaluated by accuracy score

Data Collection – SpaceX API

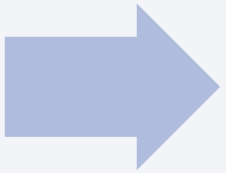


GitHub URL of the completed SpaceX API calls notebook

Data Wrangling

Fill missing values

- PayLoadMass with the mean
- LandingPad with None



Create a landing outcome of the first stage column

- 0: did not land successfully
- 1: land sucessfully

GitHub URL of completed data wrangling related notebooks

EDA with Data Visualization

- Scatter plot shows how Pay Load Mass, Launch Site, Flight Number related to the success of first stage landing
- Scatter plot of how orbit type related to Success Rate, Flight Number and Pay Load Mass
- Line plot shows the trend of yearly success from 2010 to 2020

GitHub URL of the completed EDA with data visualization notebook

EDA with SQL

SQL queries performed to get the following information

- Names of the unique launch sites in the space mission
- 5 records where launch sites 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 succesful landing outcome in ground pad was achieved
- Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Total number of successful and failure mission outcomes
- Names of the booster_versions which have carried the maximum payload mass
- records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
- 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.

GitHub URL of your completed EDA with SQL notebook

Build an Interactive Map with Folium

- The map center at the USA with:
  - 4 main markers of the launch sites
  - The popup markers of success/unsuccess event for each site
  - The lines show distance of one launch site to nearest coastline, railroad, highway and city
- The aim for these maps is to show why the launch sites are located there and the success rate of each site

GitHub URL of completed interactive map with Folium map

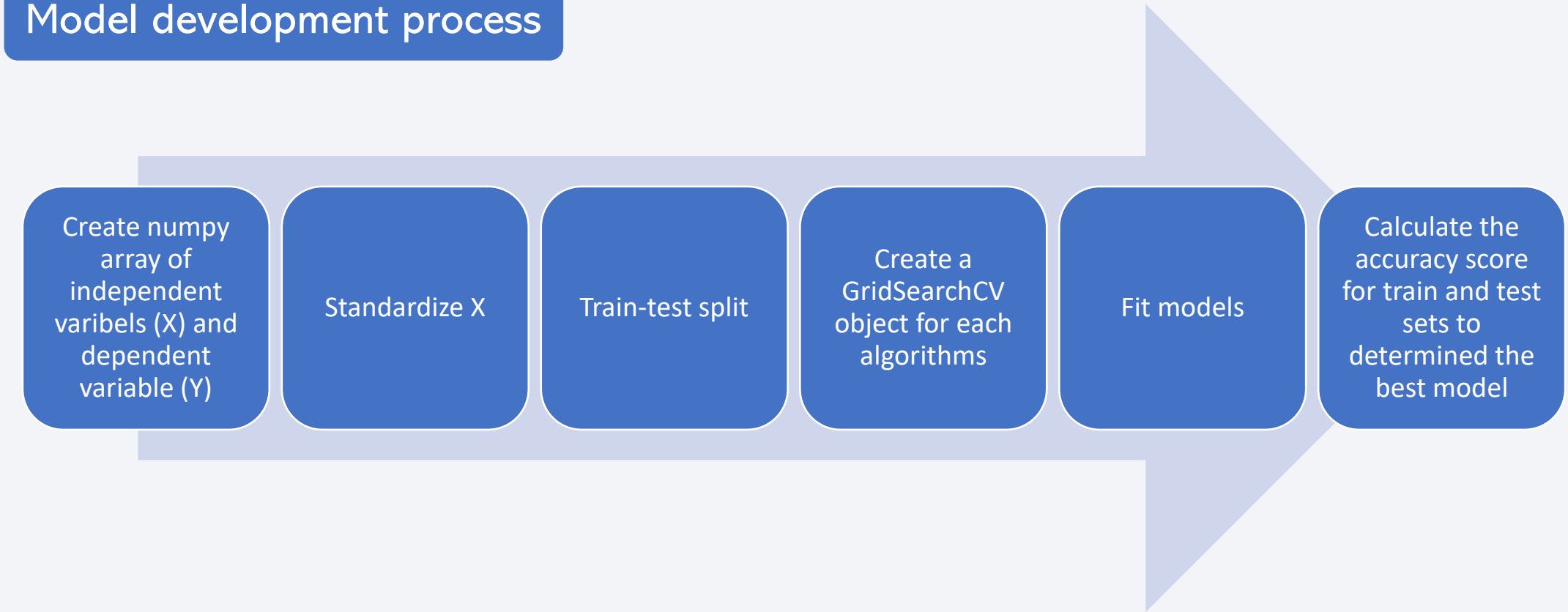
Build a Dashboard with Plotly Dash

- The Potly Dash dashboard include 2 types of charts
  - Pie chart:
    - Success count for all sites
    - Success rate for each site
  - Scatter plot shows correlation between pay load and success launch by booster version
- All the charts have the drop down to choose launch site
- The scatter plot has an extra slider to adjust the pay load range

GitHub URL of completed Plotly Dash lab

Predictive Analysis (Classification)

Model development process



GitHub URL of completed predictive analysis lab

Results

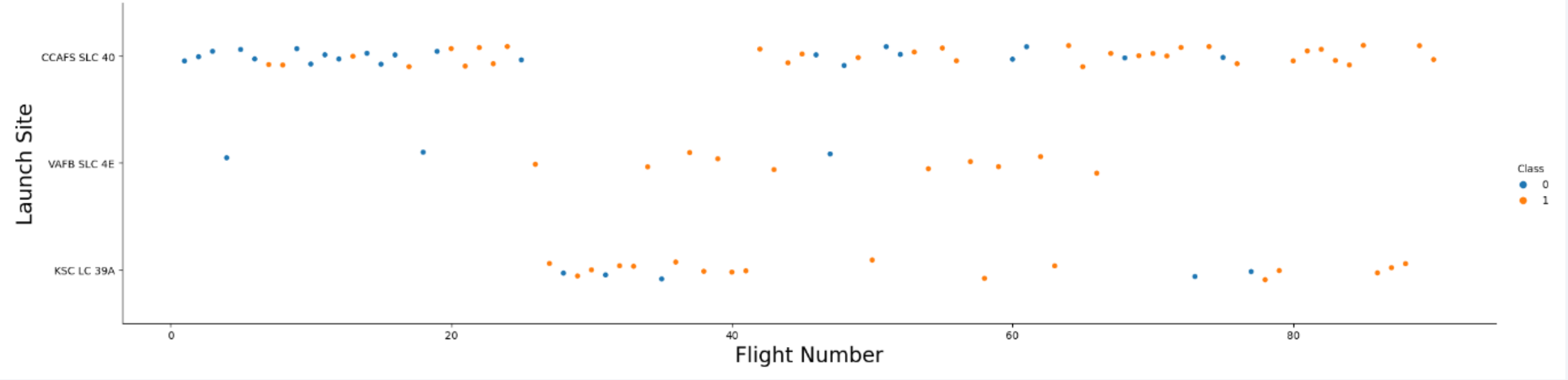
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

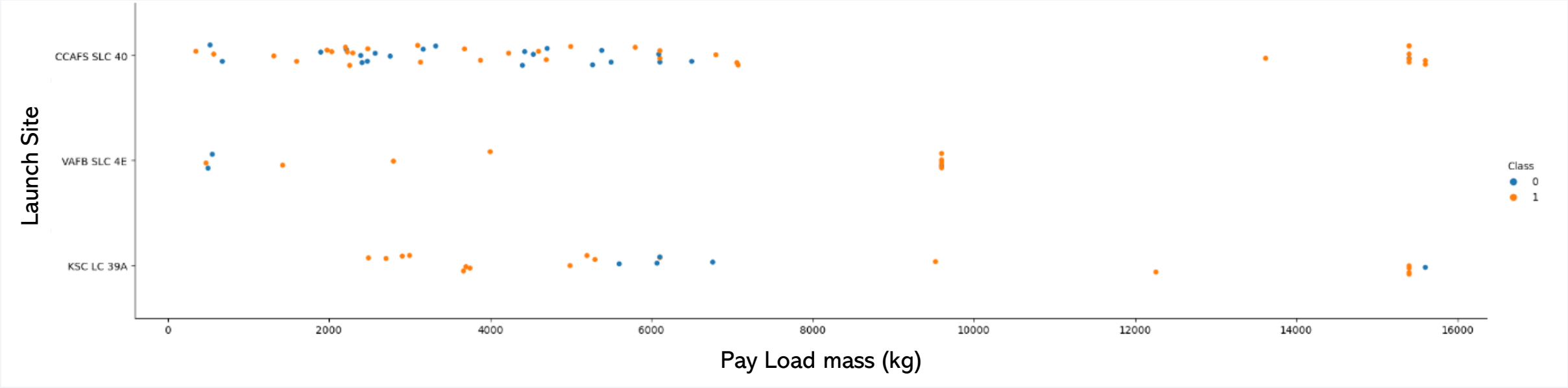
Flight Number vs. Launch Site

- The higher Flight Number has higher successful rate regardless of the Launch Sites
- It seemed liked Launch Site isn't an important factor affecting landing outcomes



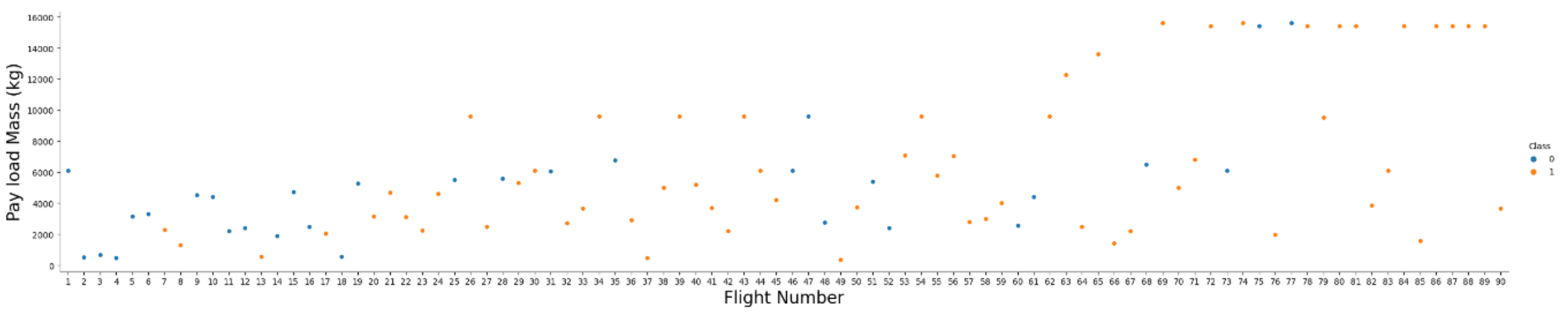
Payload vs. Launch Site

- The higher Pay Load Mass (from 9000 to 16000 kg) has almost 100% successful rate regardless of the Launch Sites



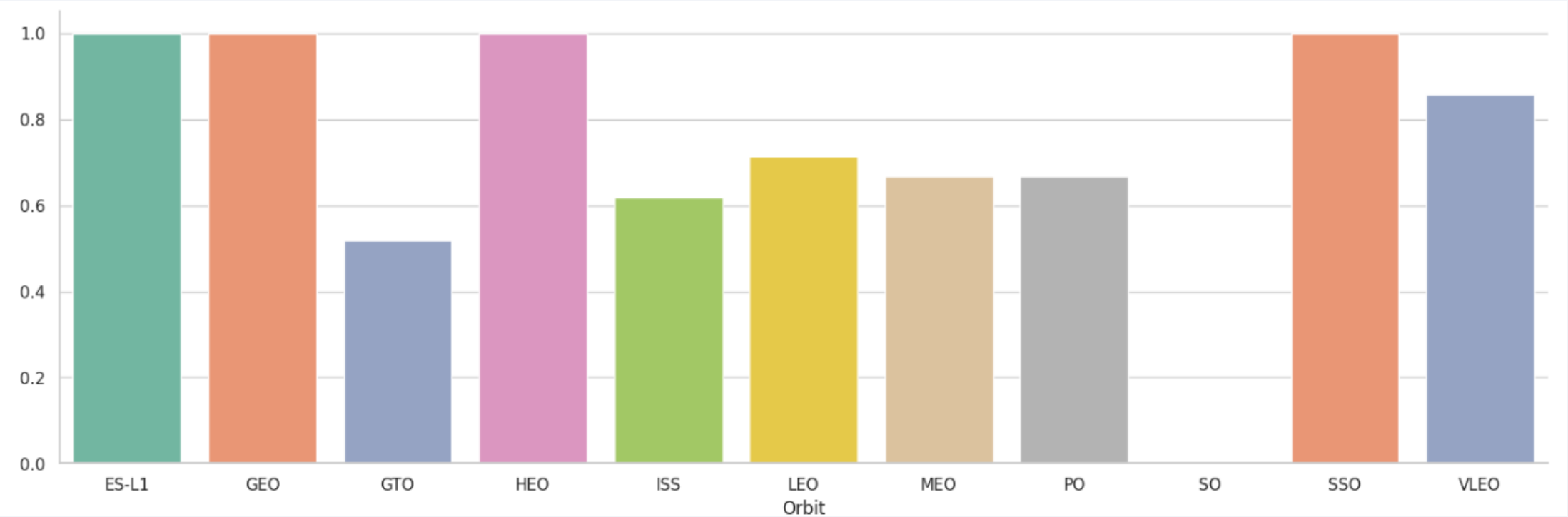
Payload vs. Flight Number

- As the flight number increases, the first stage is more likely to land successfully. The payload mass is also important; it seems the more massive the payload, the less likely the first stage will return.



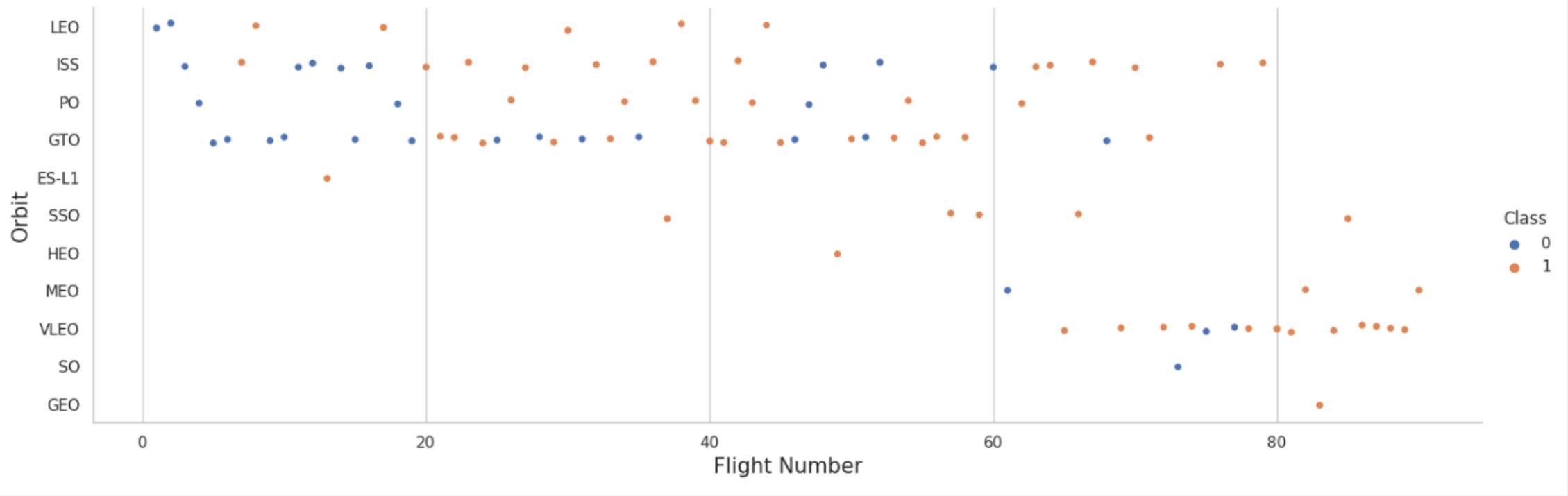
Success Rate vs. Orbit Type

- ES-L1, GEO, SSO and HEO has perfect landing rate
- SO stands last with no successful landing
- Flight Number and Pay Load Mass found to affect landing outcome. How are those 2 features relationship with orbit?



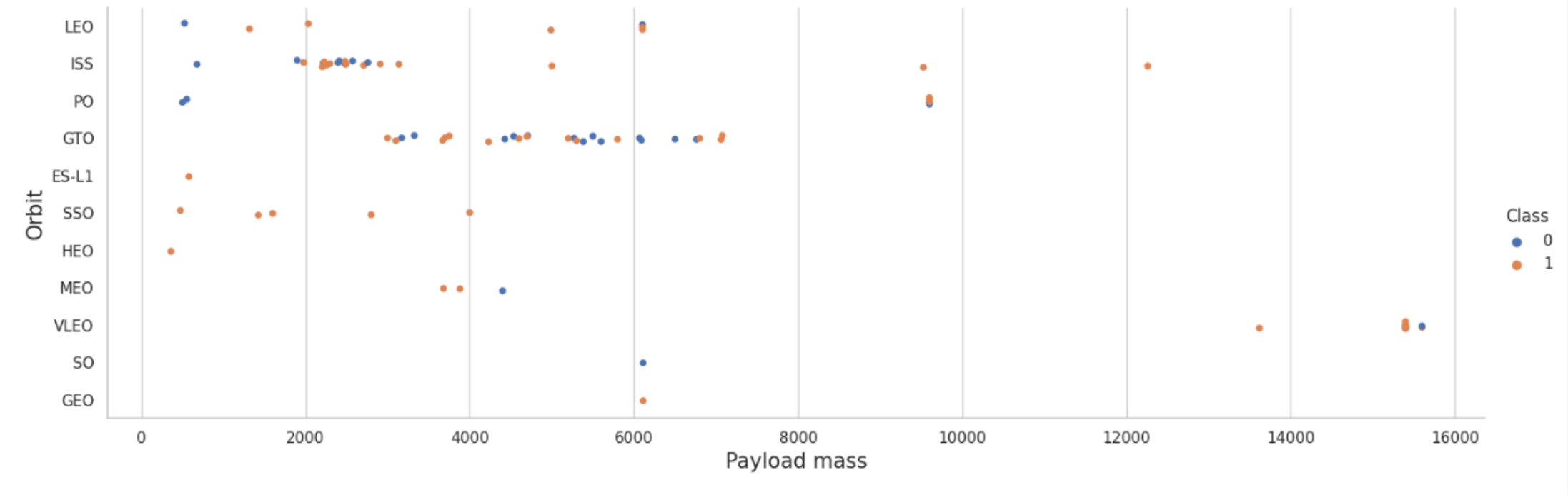
Flight Number vs. Orbit Type

- ES-L1, GEO, HEO, SSO only have few launches compare to others so the perfect successful rates may be irrelevant. And so does SSO with only one launch (0% success)
- With other orbit (except GTO) increasing flight numbers does seemed to boost successful landing

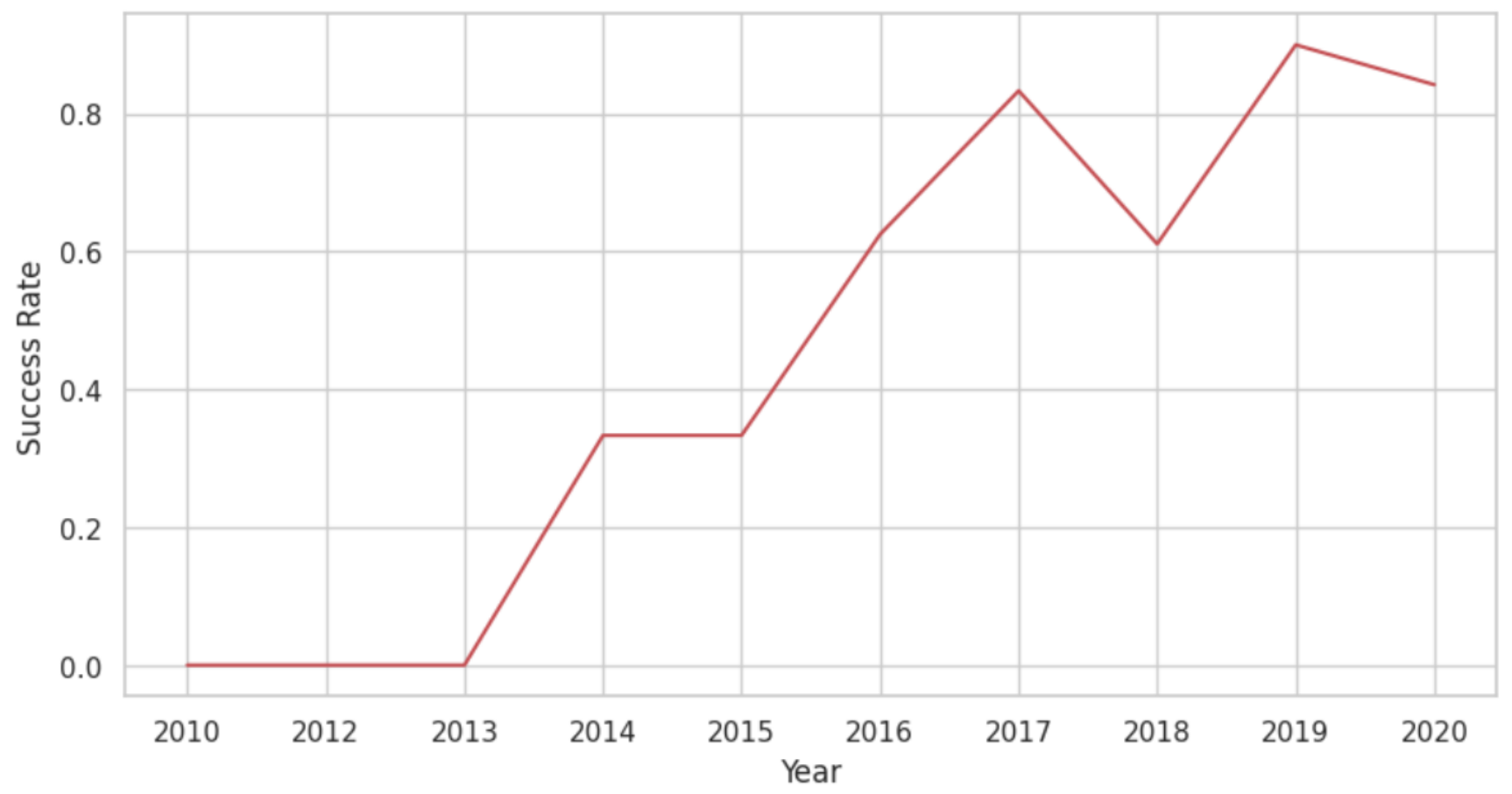


Payload vs. Orbit Type

- With higher Pay Load Mass the positive landing rate are more for LEO and ISS



Launch Success Yearly Trend



The success rate keep increasing since 2013 till 2020

All Launch Site Names

- Names of the unique launch sites:
  - CAFS LC-40
  - VAFB SLC-4E
  - KSC LC-39A
  - CCAFS SLC-40
- Query: %sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`
- Query: %sql 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
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	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 is 45596 kg
- Query: `%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)'`

Average Payload Mass by F9 v1.1

- The total average payload mass carried by booster version F9 v1.1 is 2928.4 kg
- Query: `%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1'`

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad: 2015-12-22
- Query: `%sql SELECT MIN(Date) FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)'`

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
  - F9 FT B1022
  - F9 FT B1026
  - F9 FT B1021.2
  - F9 FT B1031.2
- Query: `%sql SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome='Success (drone ship)' AND (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000)`

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

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes

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

- Query: %sql SELECT Mission_Outcome, COUNT(Mission_Outcome) FROM SPACEXTABLE GROUP BY Mission_Outcome

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Query: `%sql SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_=(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)`

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

- The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

Month	Booster_Version	Launch_Site
10	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

- Query: %sql SELECT SUBSTR(Date, 6, 2) AS Month, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE SUBSTR(Date,1,4)='2015' AND Landing_Outcome='Failure (drone ship)'

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

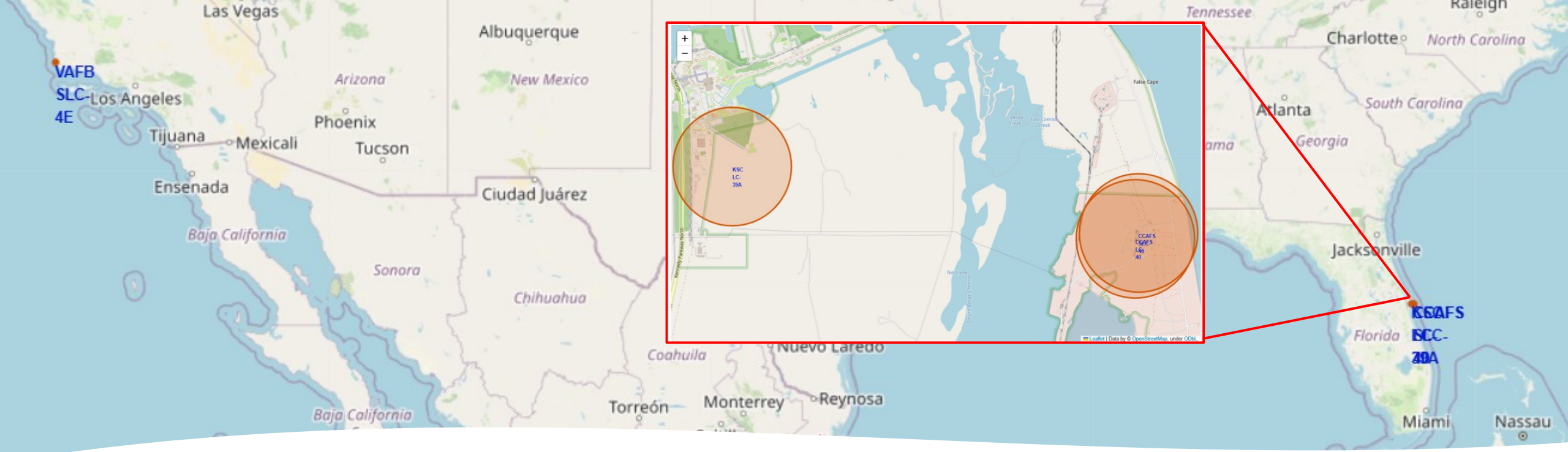
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

Date	Landing_Outcome	Num_outcomes
2010-08-12	Failure (parachute)	32

- Query: %sql SELECT Date, Landing_Outcome, COUNT(Landing_Outcome) AS Num_outcomes FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' ORDER BY Date DESC

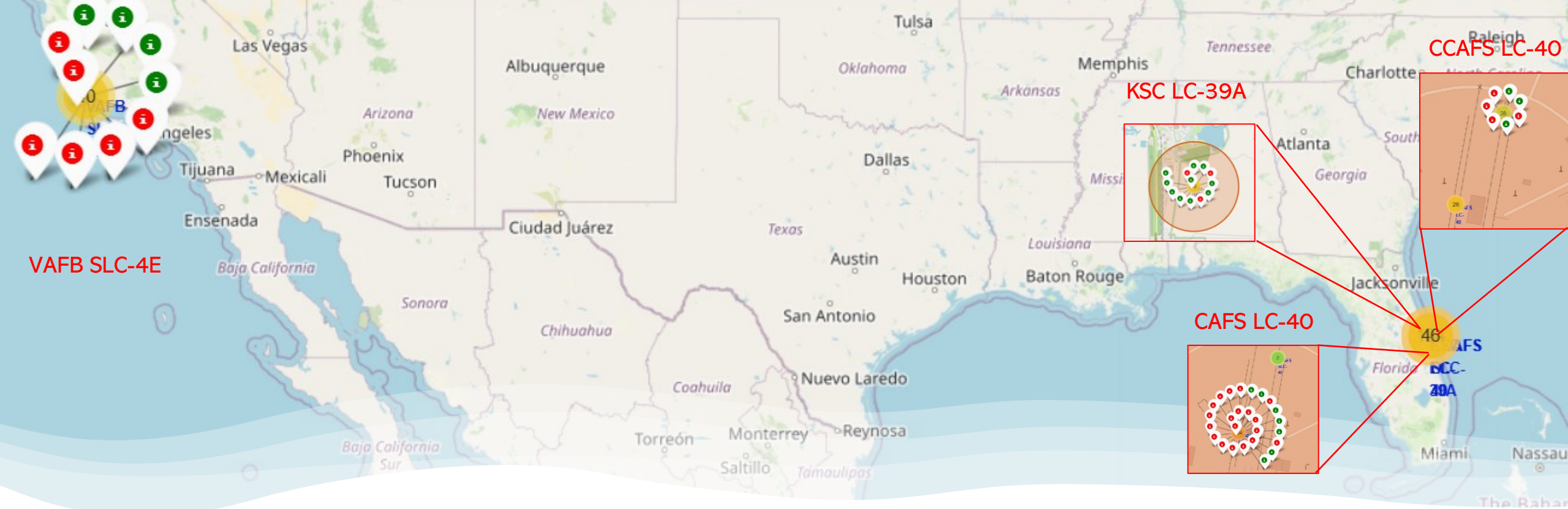
Section 3

Launch Sites Proximities Analysis



Folium Map of all Falcon 9 Launch Sites

- There are 4 launch sites of Falcon 9 in the US:
  - 3 on the the east side: CCAFS LC-40, CCCAFS LC-40, KSC LC-39A
  - 1 on the west side: VAFB SLC-4E
- All lauch sites are in very close proximity to the coast. This maybe be due to safety measures, if the lauch fail, the rocket could landed in the sea and cause no harm for the citizens.
- The lauch sites are in proximity of the Equator line so it can take optimum advantage of the Earth's substantial rotational speed. Sitting on the launch pad near the equator, it is already moving at a speed of over 1650 km per hour relative to Earth's center.

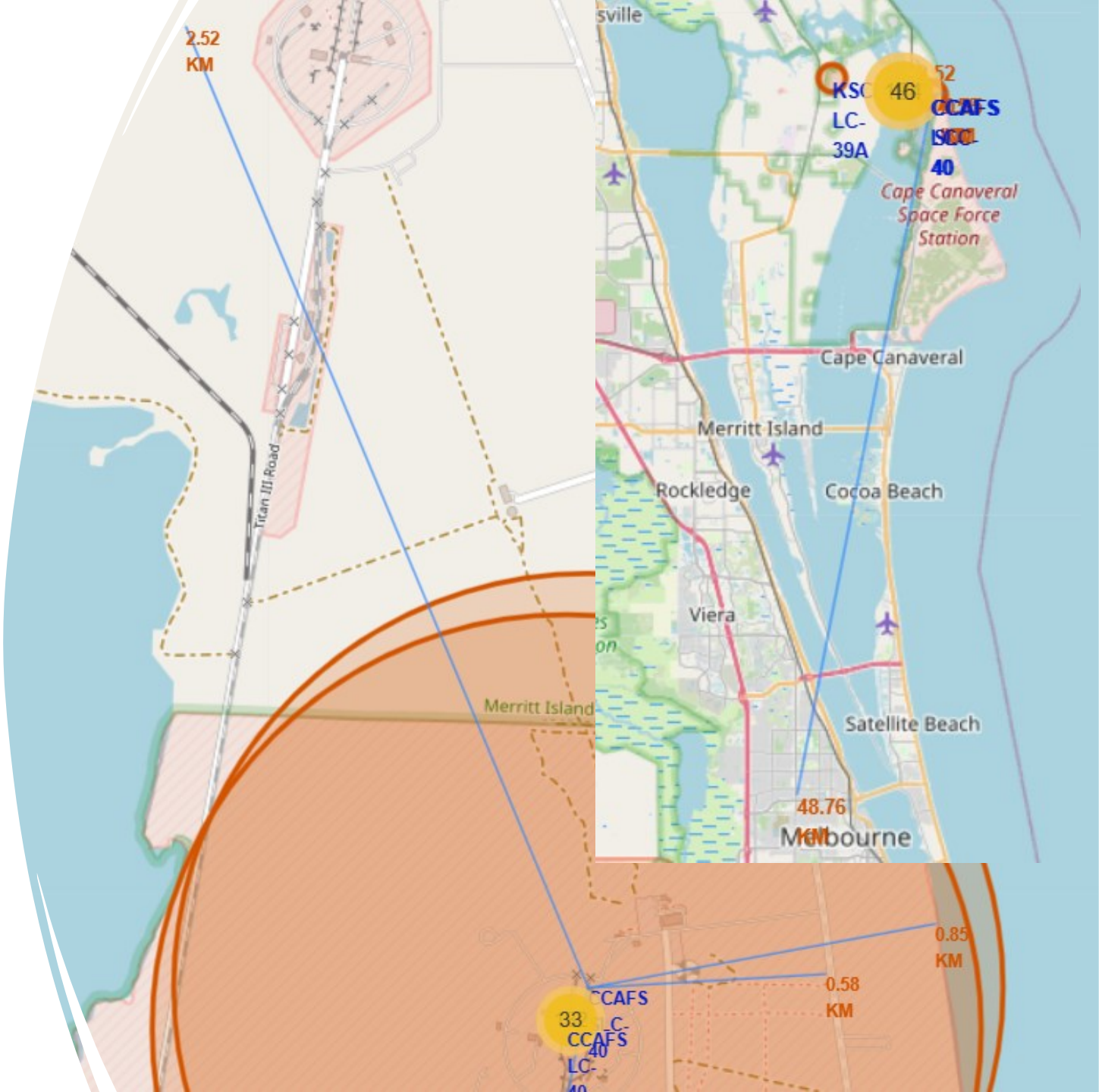


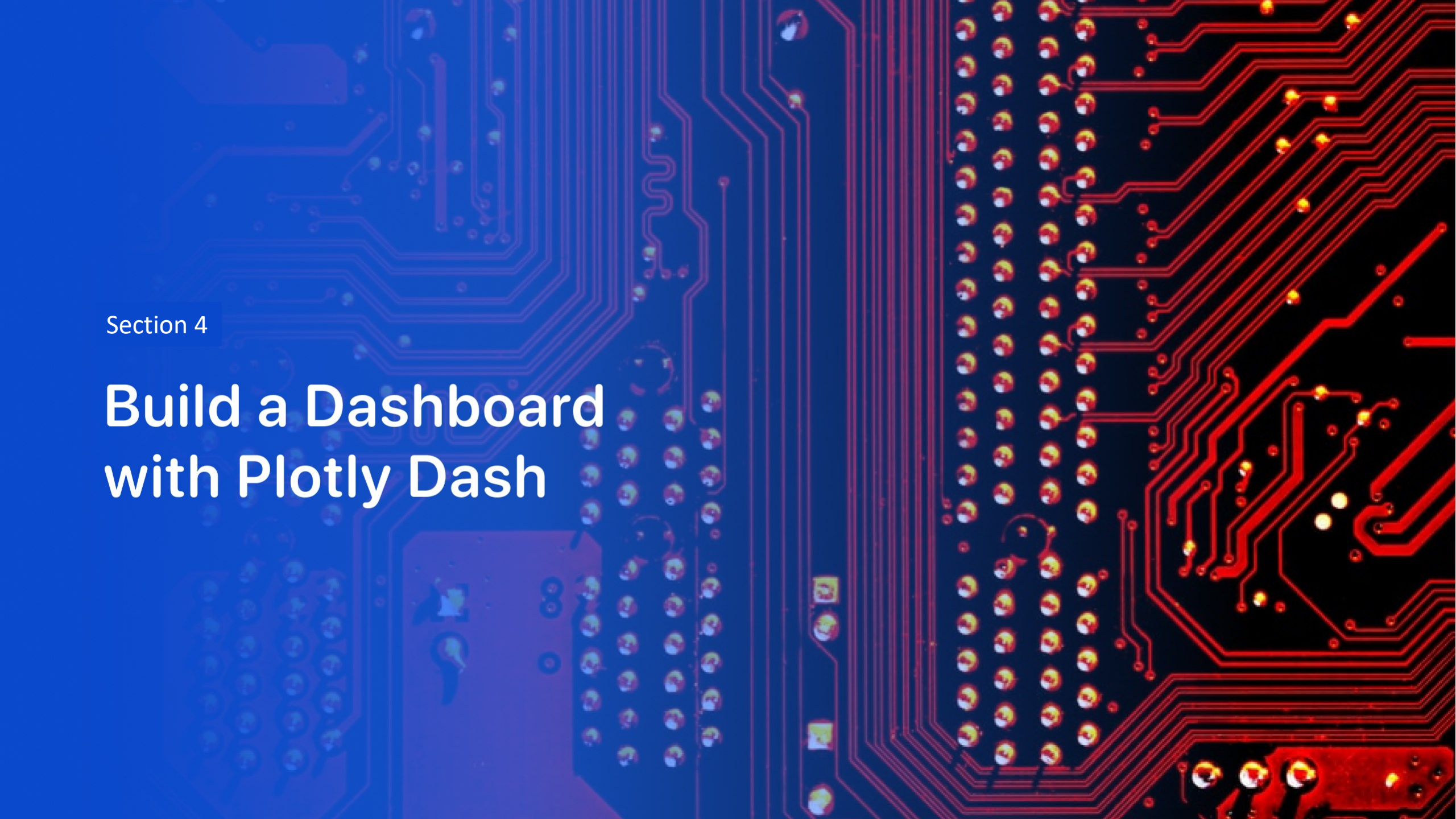
Folium Map of all Falcon 9 Launch Sites outcomes

- The outcomes for each site are colormap:
  - Red: fail landing
  - Green: successful landing
- From the color-labeled markers in marker clusters, we should be able to easily identify the number of launches and which launch sites have relatively high success rates

Distance from CCAFS LC-40 launch site to its proximities

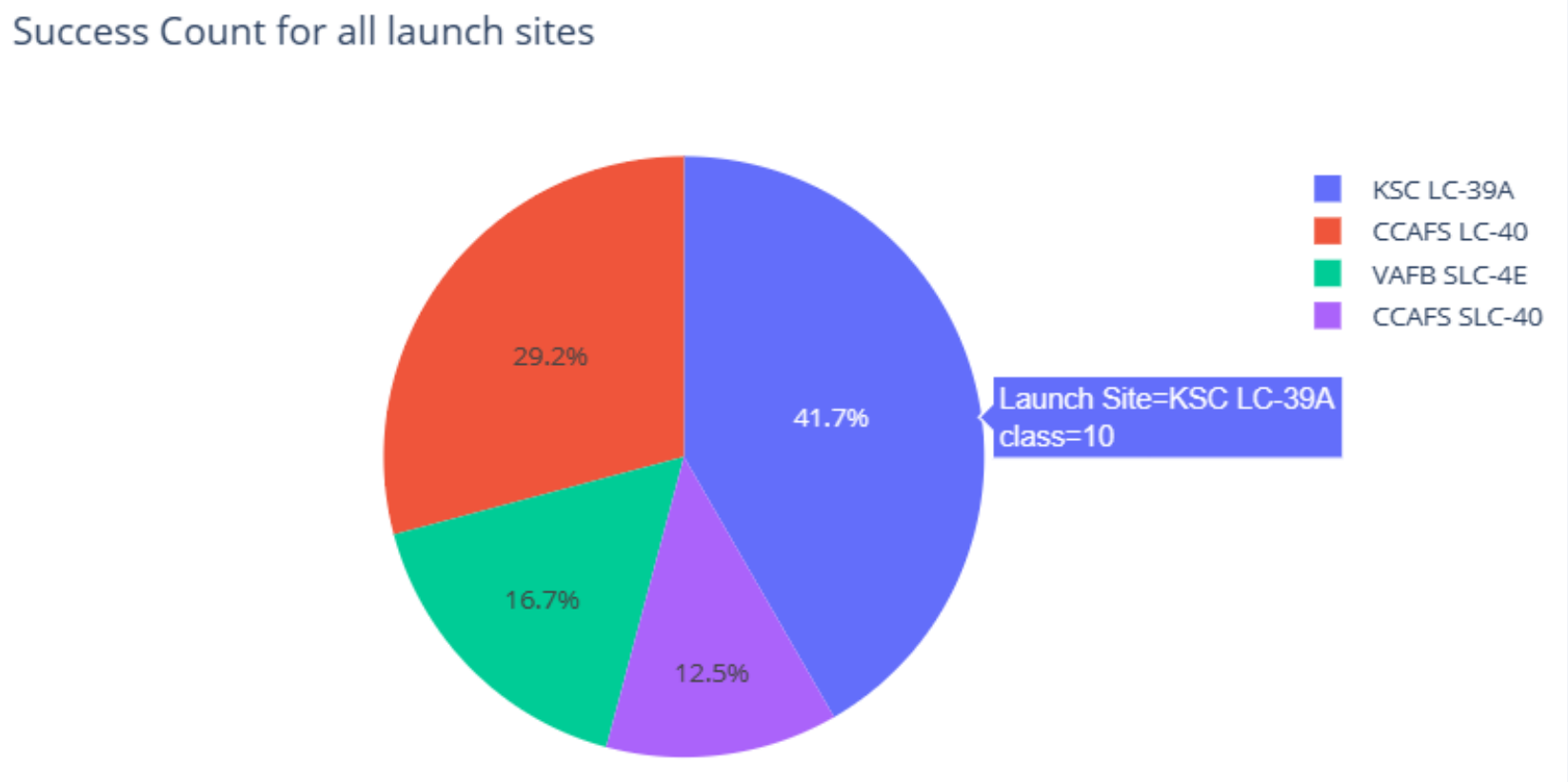
- Distance from CCAFS LC-40 launch site to its proximities including railway, highway, coastline
- Launch sites are close to railways, highways and coastline but far away from cities for 2 reasons:
  - Safety: as being closed to coastline, if the mission fail and the crew need to escape, they can land on water
  - Transportation cost-effiency: as people, space ships and material such as fuel require to be at lauch site, being closed to railways and highways would cut down the cost.



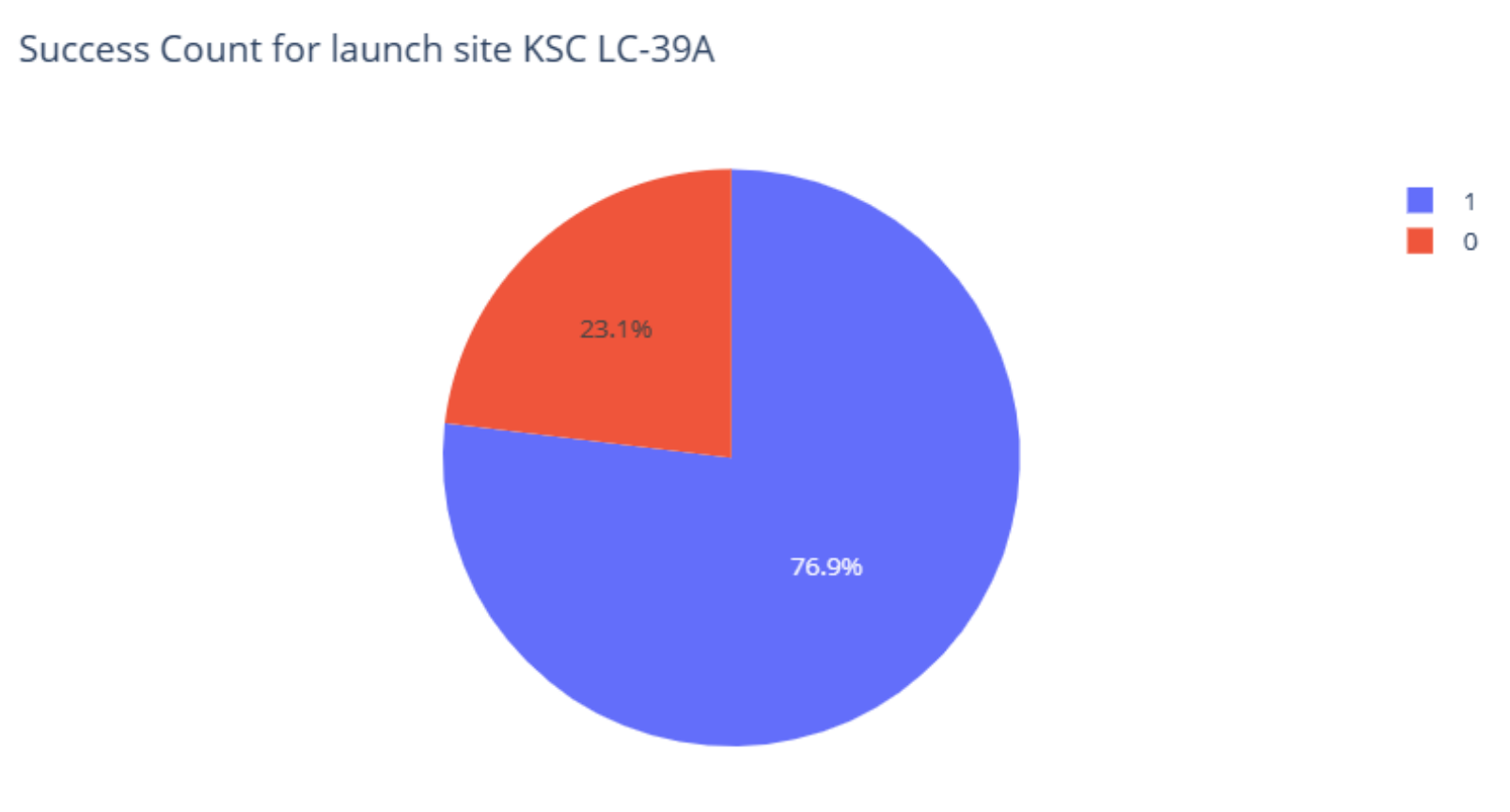

Section 4
Build a Dashboard with Plotly Dash

Launch success count for all sites

KSC LC-39A with 10 success launches and accounted 41.7% of all success launches for all sites



Site with highest launch success ratio



KSC LC-39A is also the site with the highest success ratio of 76.9%

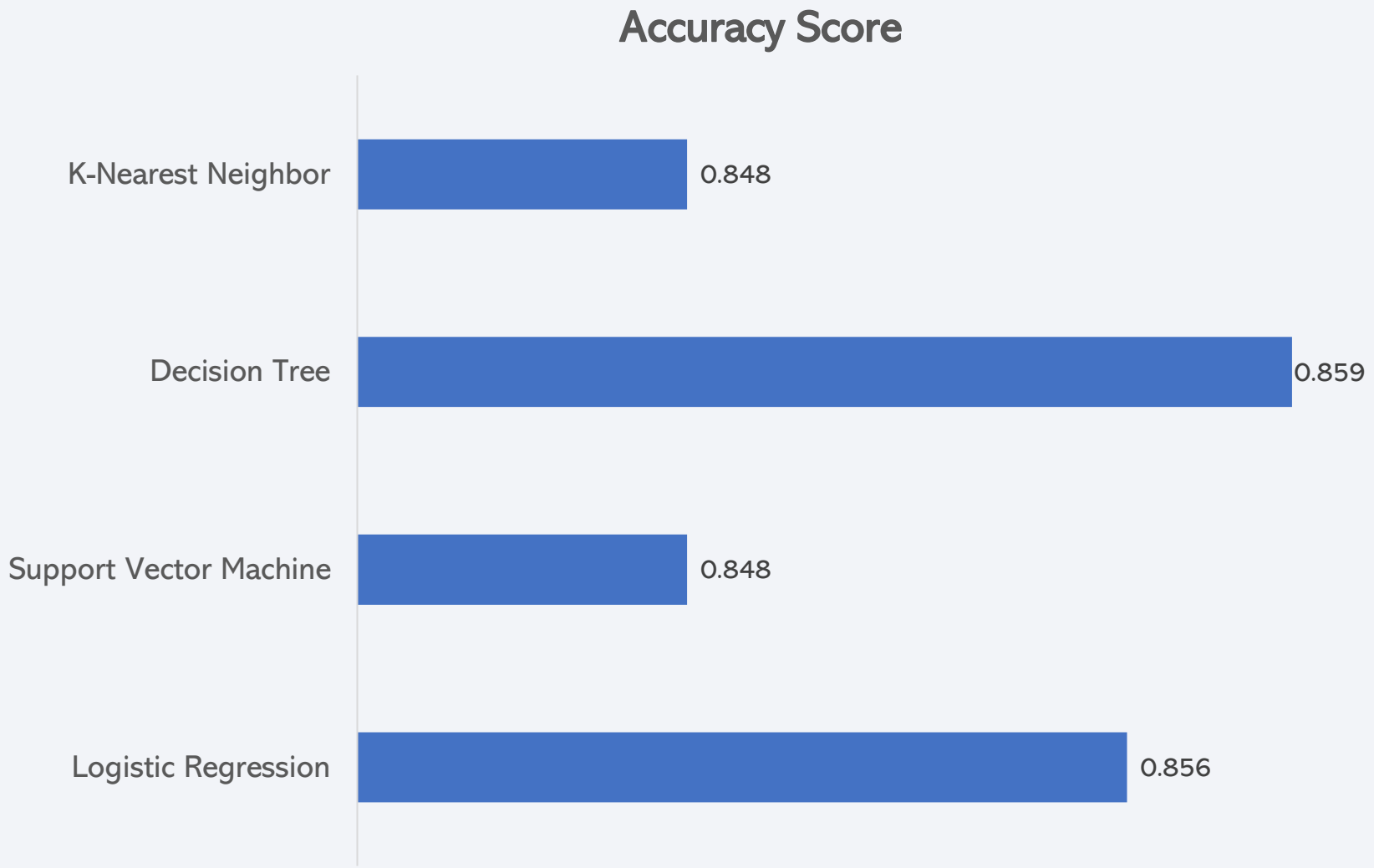
Payload vs. Launch Outcome scatter plot for all sites

- Payload of 4000-4500 kg and 6000-7000 kg has the lowest launch success rate
- Booster version B5 has only one launch and success which makes it 100% success rate. FT Booster have most of its launches successful

Section 5

Predictive Analysis (Classification)

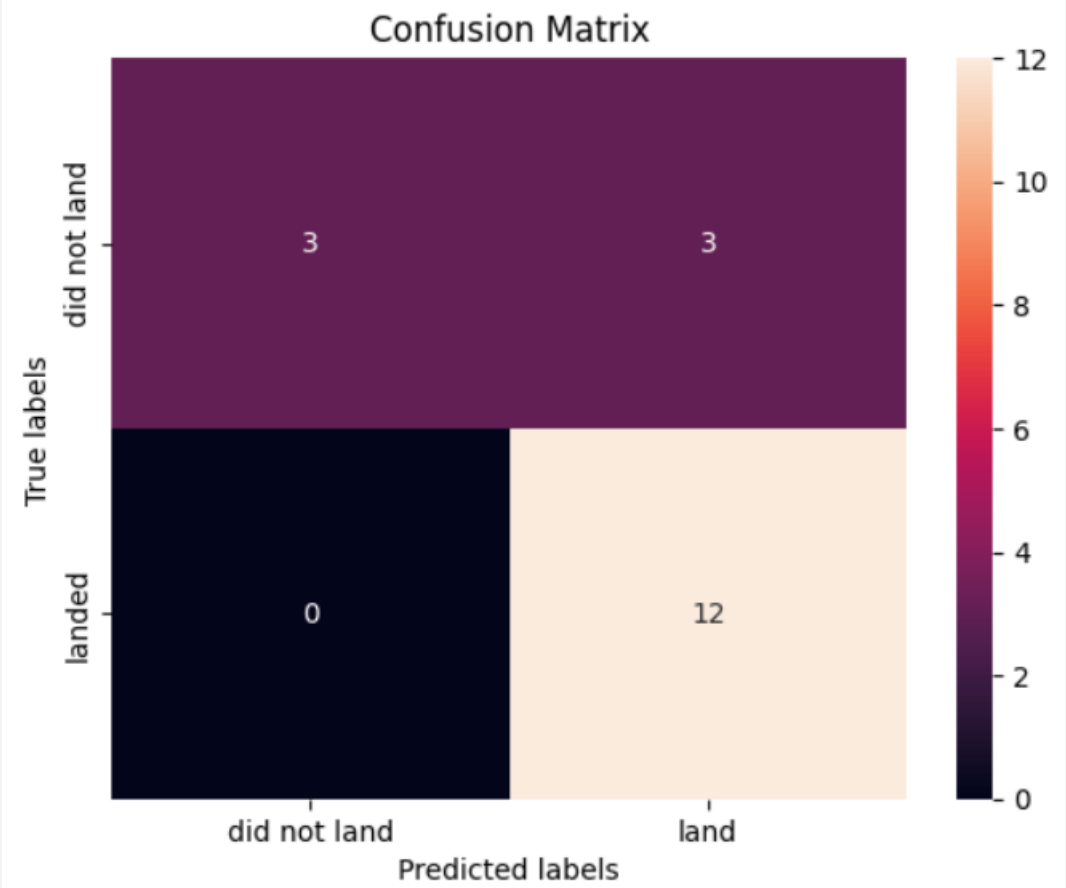
Classification Accuracy



- Decision tree model has the highest accuracy score for train set out of all 4 classification algorithms
- Accuracy for the test set is the same for all models at 0.833.

Confusion Matrix

- Here is the confusion matrix for all 4 models
- Accuracy is calculated as following
  - Accuracy = (TP + TN) / (TP + TN + FP + FN)
  - With TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative
  - For this case:
    - Accuracy = (12+3)/(12+3+0+3) = 0.833



Conclusions

- Launches with higher Payload Mass and Flight Number seemed to have better successful landing rate. However, in case of Payload, it may related to orbit type since only ISS, PO, VLEO have launches of higher load. In addition, VLEO flights all have over 60 attempts which results in good landing rate.
- Launch success yearly increased since 2013 to 2020
- All 4 F9 Falcon lauch sites are located:
  - Near the earth equator so it can take optimum advantage of the Earth's substantial rotational speed
  - Near coastline and far from cities for safety
  - Near railway, highway and coastline to maximize transportation cost-efficiency
- Out of 4 algorithms tested, Decision Tree performed the best on train set. Nevertheless, the accuracy on the test set is the same for all model. This may due to insufficient observations for the test set

Appendix

- All the code and results for this report can be found via this GitHub URL
- This presentation is the final project of IBM Applied Data Science Capstone

Thank you!