IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Roshini Fernando
17 Jul 2022

Outline

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

Executive Summary

Summary of methodologies

- Data collected using the SpaceX API and Web Scraping.
- Exploratory Data Analysis (EDA) performed with Data Wrangling and Interactive Data Visualization.
- Built four Machine Learning Models and Evaluated.

Summary of all results

- We can collect useful data from publicly available datasets.
- EDA supports to identify which variables have more relation with launches.
- ML models supports to predict outcomes of future launches and get crucial business decisions to improve businesses.

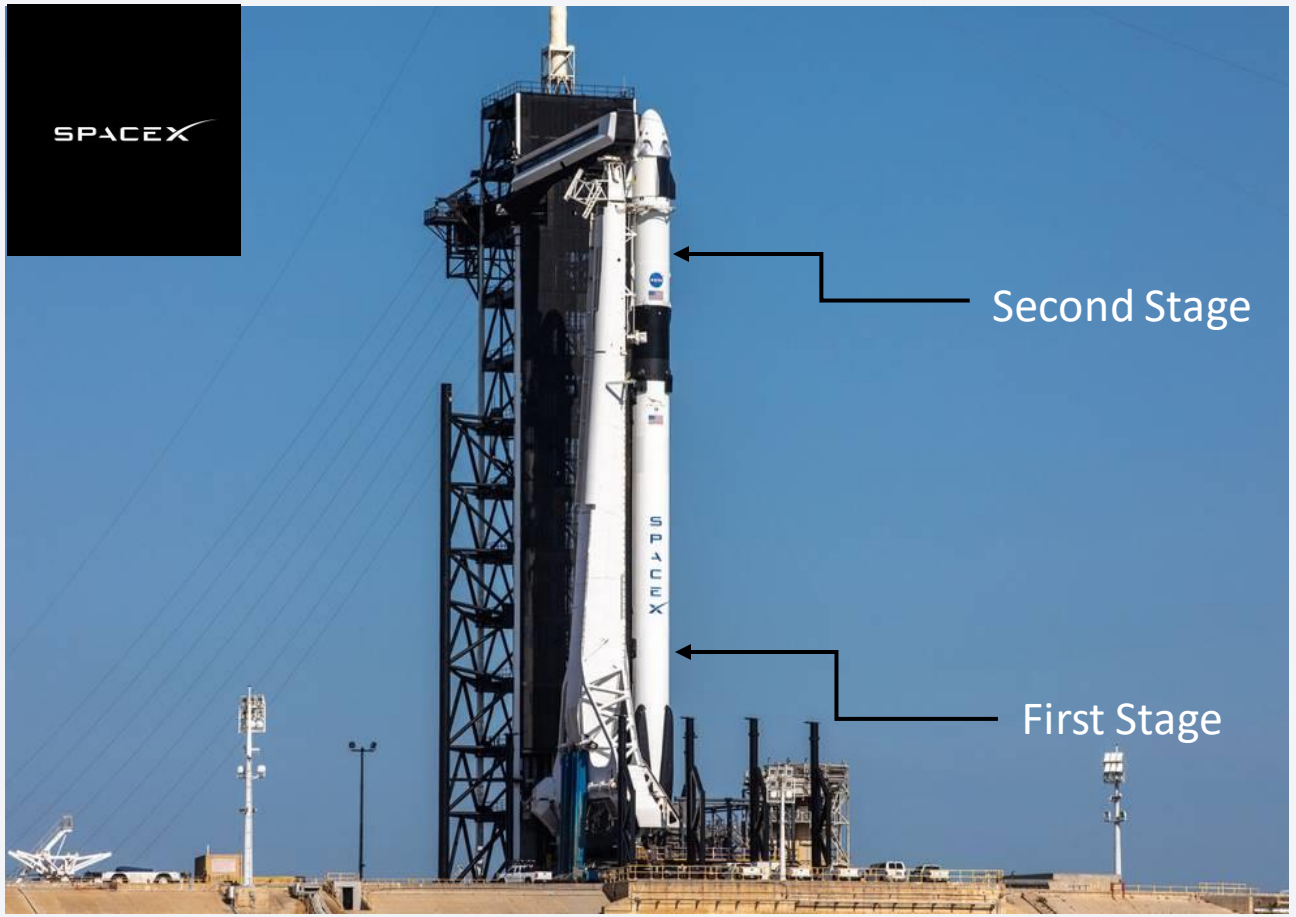
Introduction

Falcon 9

Falcon 9 is a two stage rocket designed and manufactured by SpaceX.

As they said Falcon 9 can safely transport people and payloads to Earth orbit and beyond.

Most crucial fact is they say that first stage component can be reuse.



SpaceX Falcon 9 rocket launches with a 62 million dollars cost while other companies cost upward of 165 million dollars. SpaceX saves more money for launches because Falcon 9 first stage can reuse.

So being informed about the likelihood of getting success of the first stage of Falcon 9 landing and the rocket launch cost is crucial for other competitors to compete.

Main Objective

To support decisions of stake holders of a new startup company who competing with SpaceX in bids against for a rocket launch.

To fulfill the objective need to,

1. Predict successful landings of first stage of Falcon 9 rocket by SpaceX and estimate the Total cost to a launch.
2. Find best locations for Launch Sites.

Section 1

Methodology

Methodology

Executive Summary

- Data collection Methodology:

  Data Acquisition was done using two sources.

  - Via SpaceX API
  - From Web Scraping

- Perform Data Wrangling

  Created a 'Class' feature for landing outcomes of each launch with a dedicated orbit by analyzing and summarizing features whether Falcon 9 first stage landed successfully or not.

- Perform Exploratory Data Analysis (EDA) using Visualization and SQL

Methodology

Executive Summary

- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

  In this step the collected data, normalized, Sliced as Training data and Test data and built four Classification models by fitting them and tuned for getting optimal parameters and evaluated using different techniques.

Data Collection

For predicting the first stage landing status of Falcon 9 launches we needed relevant history data of Falcon 9 launches for Analyzing, Modeling for classification and for evaluation.

PayloadMass, Orbit, LaunchSite, Outcome like feature information were needed to collect for each launch approach to do a classification.

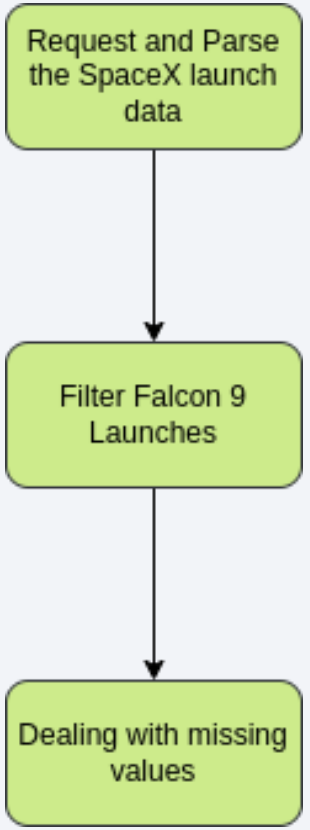
So followed the two easiest ways to do data collection for Falcon 9 launches by using online available data.

1. Using SpaceX API
2. Using Web scraping

Data Collection – SpaceX API

- There's an unofficial open source SpaceX REST API that Provides endpoints to obtain related data.
    1. https://api.spacexdata.com/v4/rockets/
    2. https://api.spacexdata.com/v4/launchpads/
    3. https://api.spacexdata.com/v4/payloads/
    4. https://api.spacexdata.com/v4/cores/
    5. https://api.spacexdata.com/v4/launches/past
- The REST API was used according to the flow chart beside.
- GitHub URL of the completed SpaceX API calls notebook -

https://github.com/RoshiniFernando/IBM-Applied-Data-Science-Capstone-Project/blob/main/Data%20Collection%20API%20Lab.ipynb



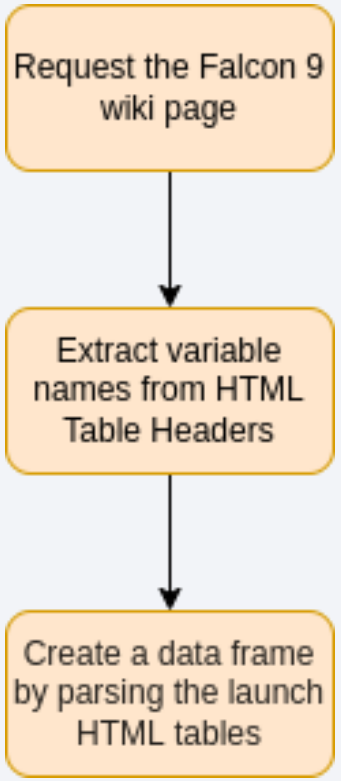
Data Collection - Scraping

- Data related to SpaceX rocket launches can be also found on web pages like Wikipedia.
- Process was conducted as beside flow chart using following Wikipedia page.

https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

- GitHub URL of the completed web scraping notebook -

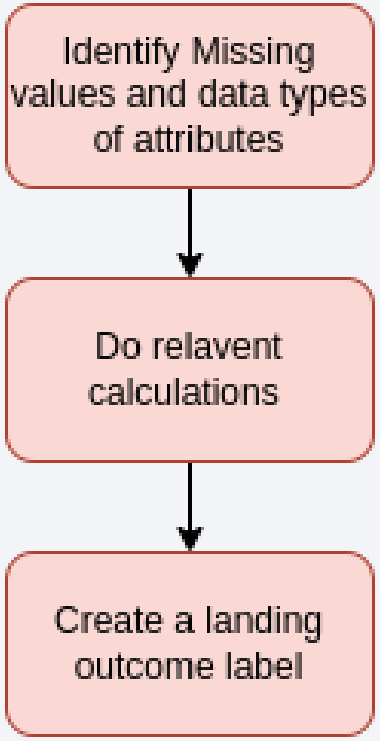
https://github.com/RoshiniFernando/IBM-Applied-Data-Science-Capstone-Project/blob/main/Data%20Collection%20with%20Web%20Scraping%20lab.ipynb



Data Wrangling

- Firstly calculated the Missing value percentages for each attribute and identified Categorical/Numerical columns performing EDA.
- Calculated the number of launches on each site, occurrences of each orbit and occurrences of mission outcome per orbit type.
- Created a landing outcome label from Outcome column.
- GitHub URL of your completed data wrangling related notebook -

https://github.com/RoshiniFernando/IBM-Applied-Data-Science-Capstone-Project/blob/main/EDA%20Lab.ipynb



EDA with Data Visualization

- Here considered about relationships between attributes using three different charts for feature selection.

1. **Scatter Plots**

   Usually show relationship between variables.

2. **Bar Charts**

   Can compare data among different categorical variables.

3. **Line Charts**

   Show how something changes over time.

- GitHub URL of your completed EDA with data visualization notebook -

https://github.com/RoshiniFernando/IBM-Applied-Data-Science-Capstone-Project/blob/main/EDA%20with%20Visualization%20Lab.ipynb

EDA with SQL

The SQL queries performed.

- Names of the unique launch sites in the space mission
- Top 5 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
- The date of first successful landing outcome in ground pad
- 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
- Failed landing_outcomes in drone ship, their booster versions, and launch site names for 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 -

https://github.com/RoshiniFernando/IBM-Applied-Data-Science-Capstone-Project/blob/main/EDA%20with%20SQL%20Lab.ipynb

Build an Interactive Map with Folium

- The Launch site success rate may depends on many factors like PayloadMass, Orbit type and so on but it may depend on the location and proximities too.
- So to analyze the launch site locations built interactive maps using Folium with map objects such as markers, circles, lines, etc.

1. **Circles** - To add a highlighted circle area that shows on a specific coordinate with a label.
2. **Markers** - To indicate launch sites as points.
3. **Marker Clusters** – To show the events (Many Launches) on a same coordinate by using many markers.
4. **Mouse Position** - To get coordinate for a mouse over a point of the map.
5. **Lines** - To indicate distance between two coordinates.

- GitHub URL of the completed interactive map with Folium map -

https://github.com/RoshiniFernando/IBM-Applied-Data-Science-Capstone-Project/blob/main/Interactive%20Visual%20Analytics%20with%20Folium%20Lab.ipynb

Build a Dashboard with Plotly Dash

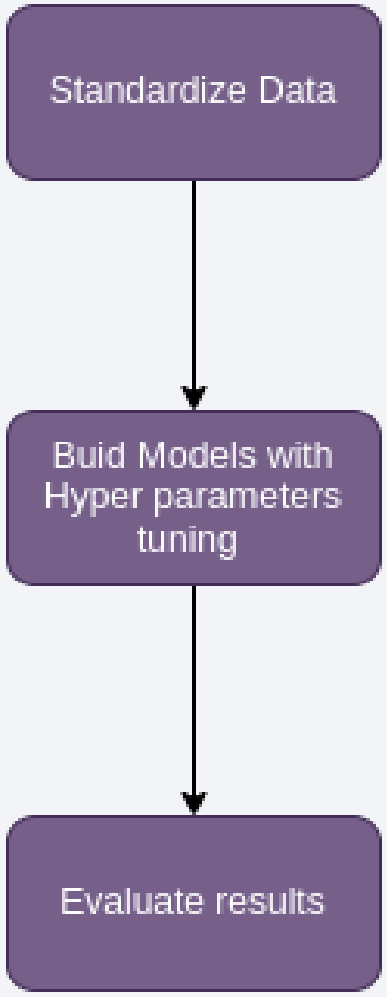
- Drop down, Range Slider, Pie charts and Scatter plots were used to analyze the relation between Launch sites and Payloads by creating a dashboard.

1. **Drop Down** - To select different launch sites.
2. **Range Slider** - To select a Payload range.
3. **Pie Charts** - To visualize success counts of selected launch site.
4. **Scatter Plots** – To observe how payload maybe correlated with mission outcomes for selected site.

- GitHub URL of the completed Plotly Dash lab -

https://github.com/RoshiniFernando/IBM-Applied-Data-Science-Capstone-Project/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)



- Raw data went through some set of preprocessing steps.
- Created Logistic Regression, SVM, Decision Tree Classifier and KNN Classifier.
- Improved models using Hyper Parameter Tuning.
- Evaluated using Confusion matrices, Bar chart and ROC Curves and found the SVM as the best performing model.
- GitHub URL of the completed predictive analysis lab -

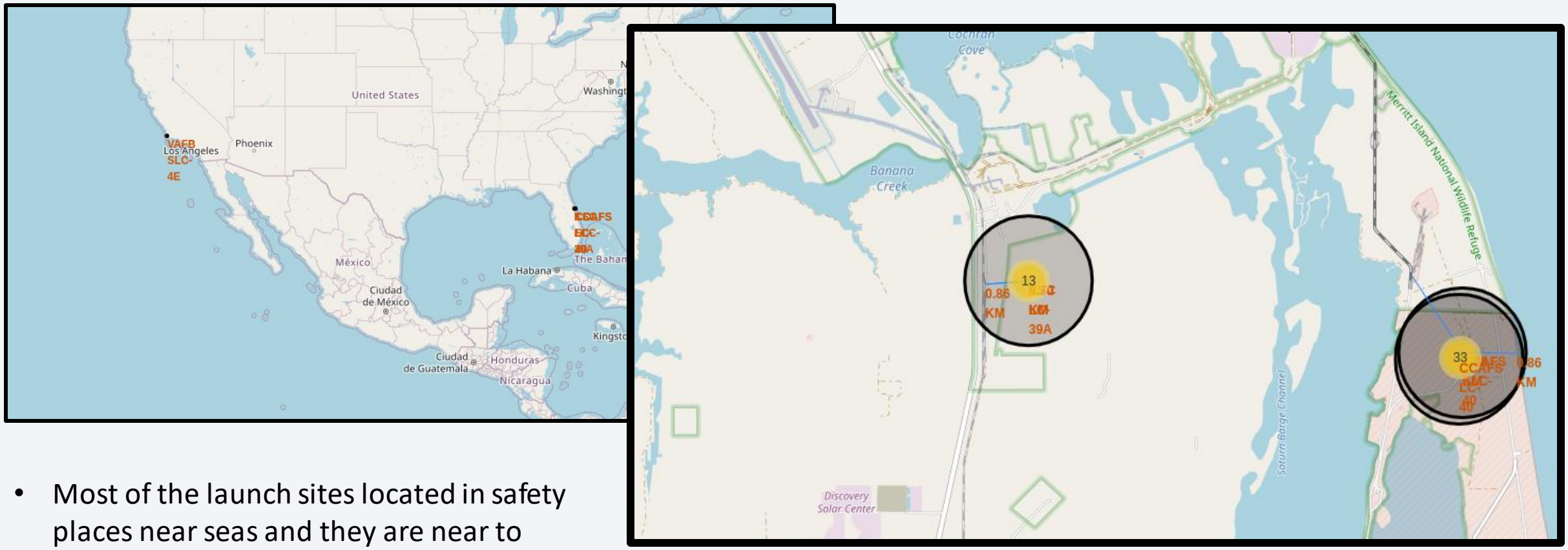
https://github.com/RoshiniFernando/IBM-Applied-Data-Science-Capstone-Project/blob/main/Machine%20Learning%20Prediction%20Lab.ipynb

Results

Exploratory data analysis results

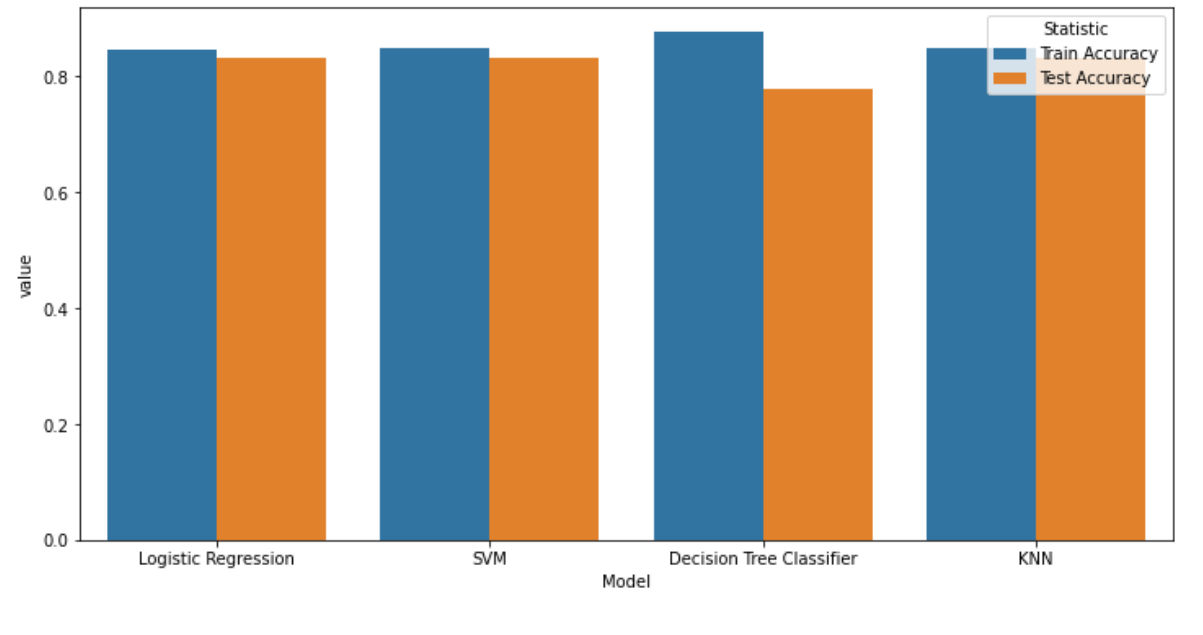
- SpaceX has four Launch sites.
- The first launches are conducted by SpaceX and NASA.
- The Average Payload Mass of F9 v1.1 is 2928kg.
- The very First successful Ground Pad landing happened in 2015.
- In the most of launches, Mission outcome was successful.
- When the years pass, the Landing outcomes were getting successful.

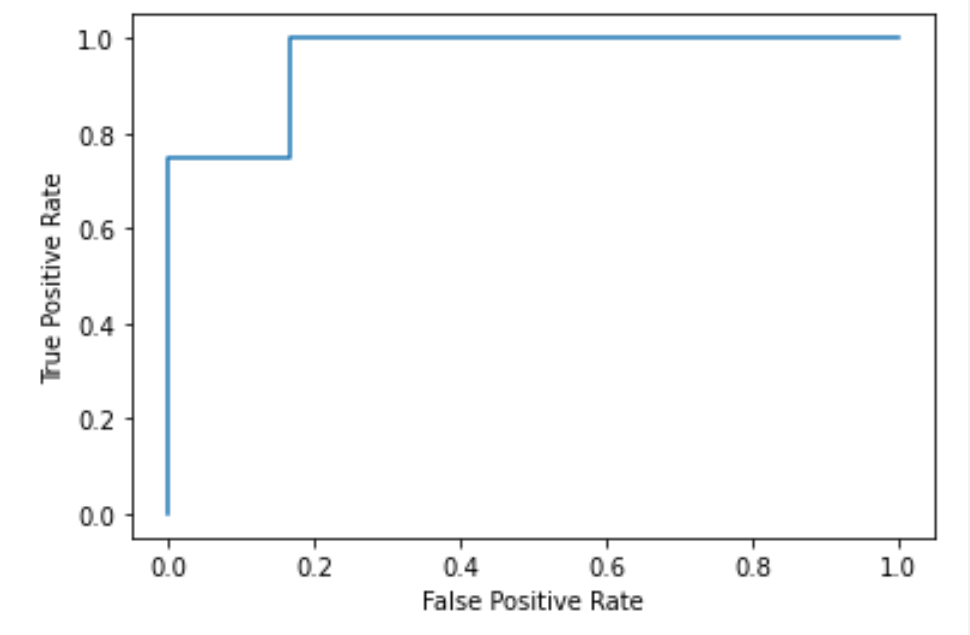
Results



- Most of the launch sites located in safety places near seas and they are near to Railways and roads.
- The areas are inhabited too.

Results

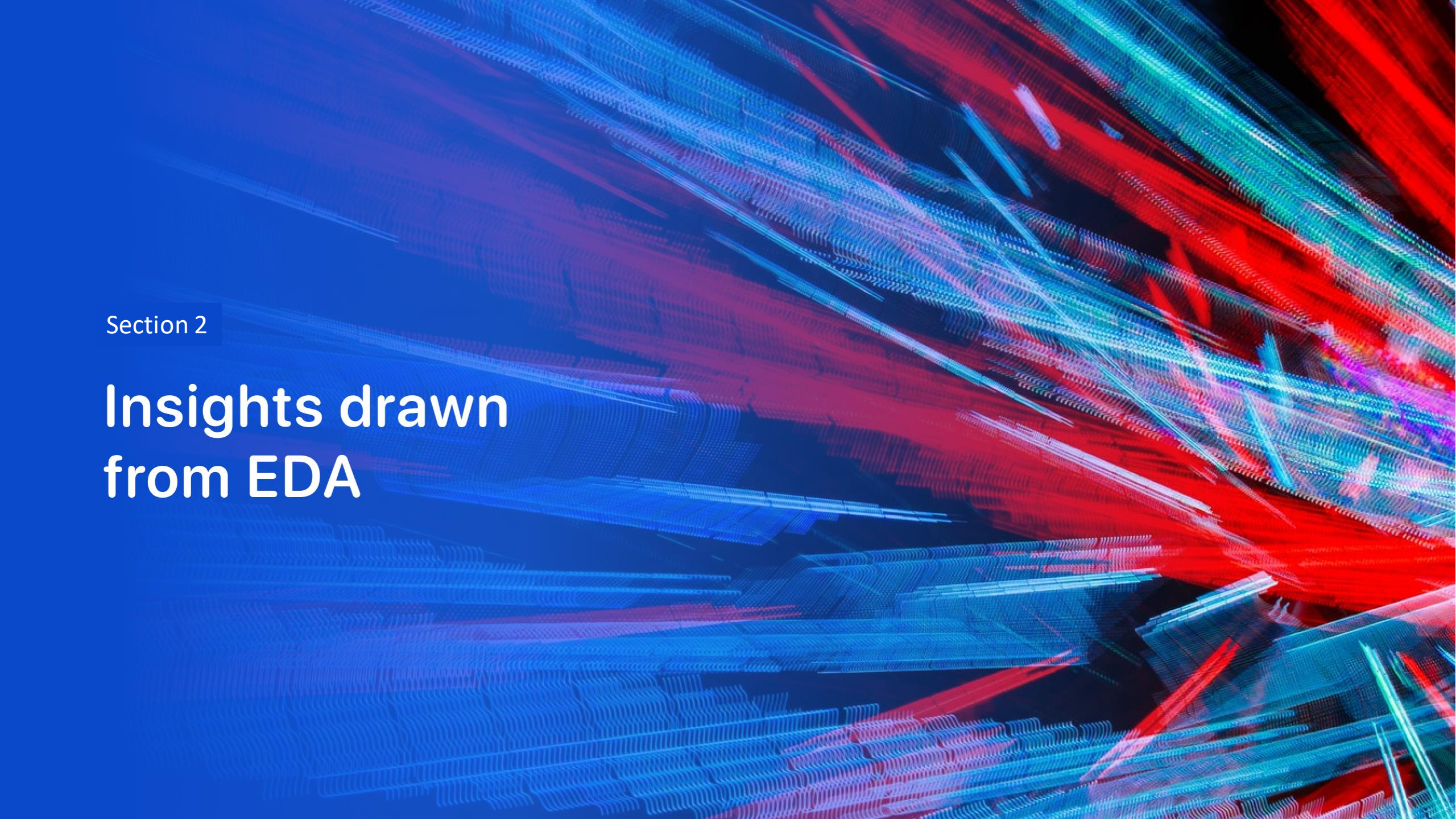




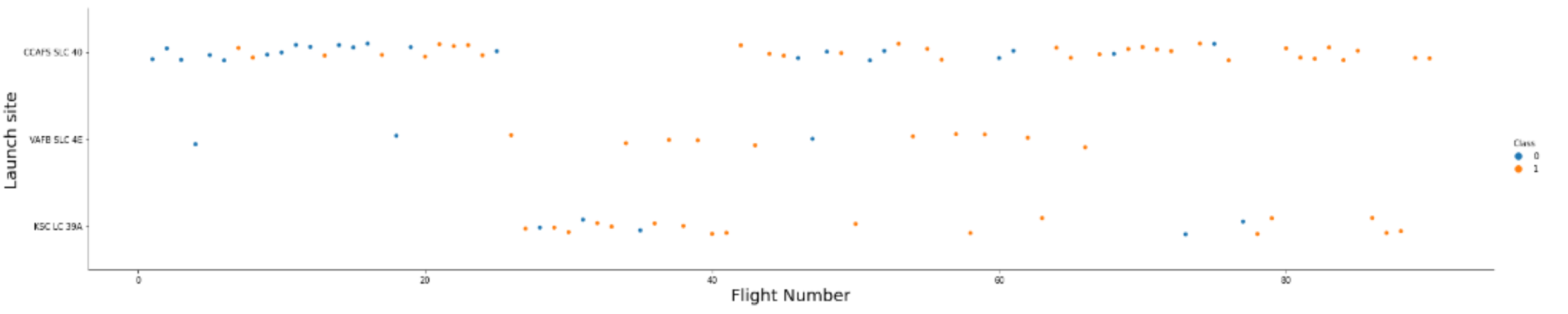
- Predictive Analysis shows that the best classification model is SVM with ~85% of Training Accuracy and ~83% of Test Accuracy.

Section 2

Insights drawn from EDA



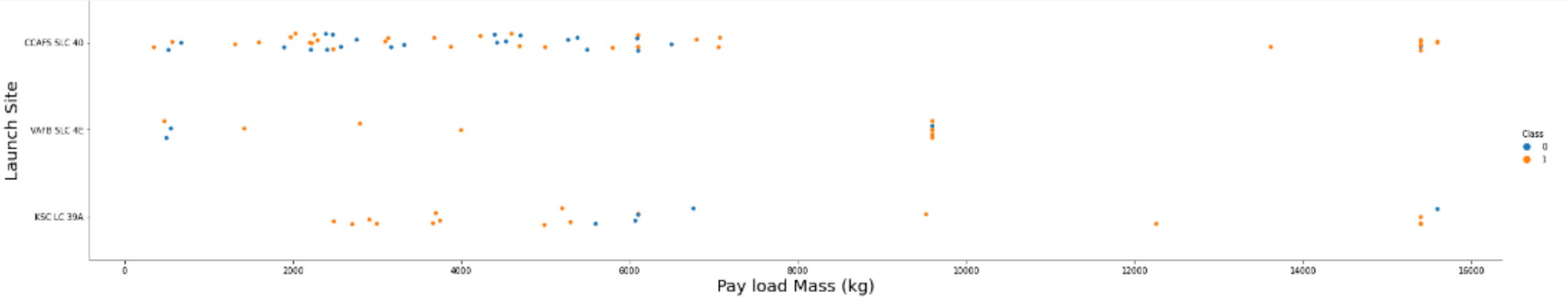
Flight Number vs. Launch Site



According to the scatter plot,

- CCAFS SLC 40 is the best launch site these days because in most of the last launches first stage has landed successfully.
- Also all last launch outcomes of all three launch sites can see with best results means landing first stage is improving.

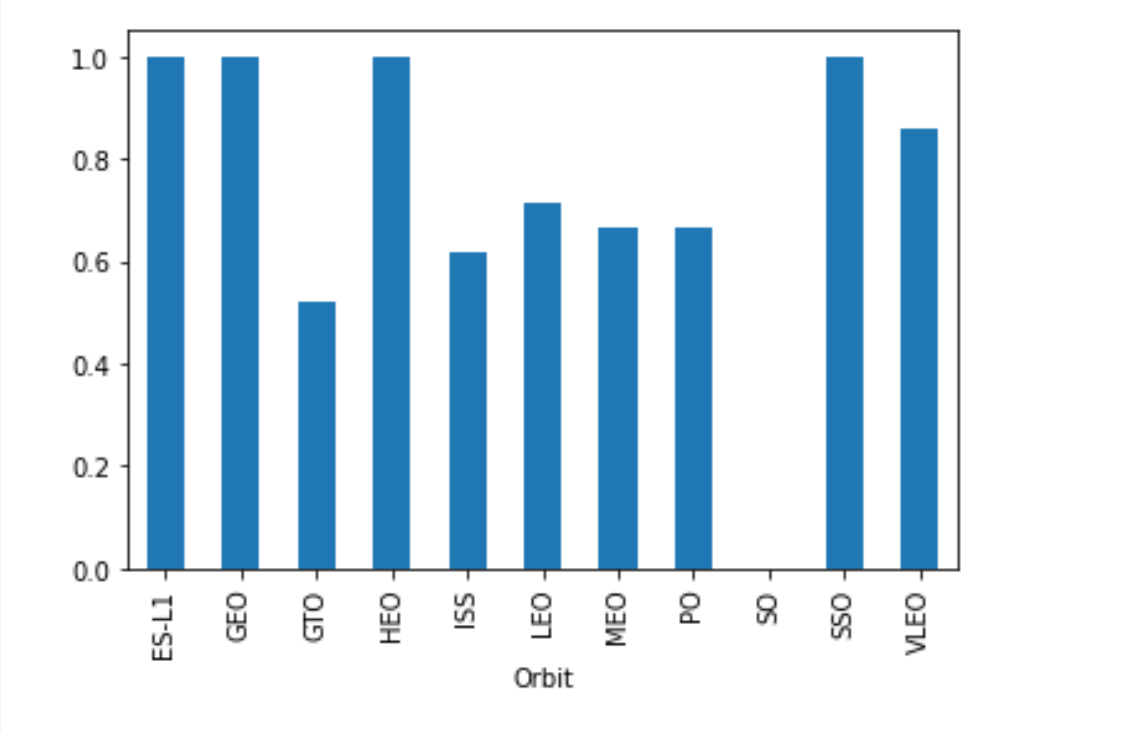
Payload vs. Launch Site



According to the scatter plot,

- Most successful launch outcomes can be seen when Payload is more than 9000kg.
- Also the launches with Payload weigh more than12000kg seems to be tried only in CCAFS SLC 40 and KSC LC-39A launch sites.

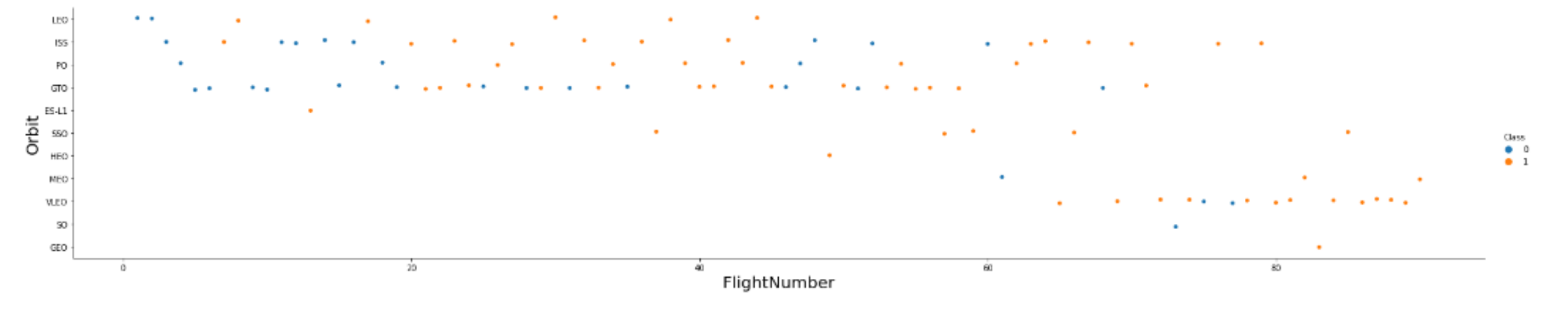
Success Rate vs. Orbit Type



According to the Bar Plot,

- The highest success rate shows in the launches to the Orbit ES-L1, GEO, HEO and SSO.

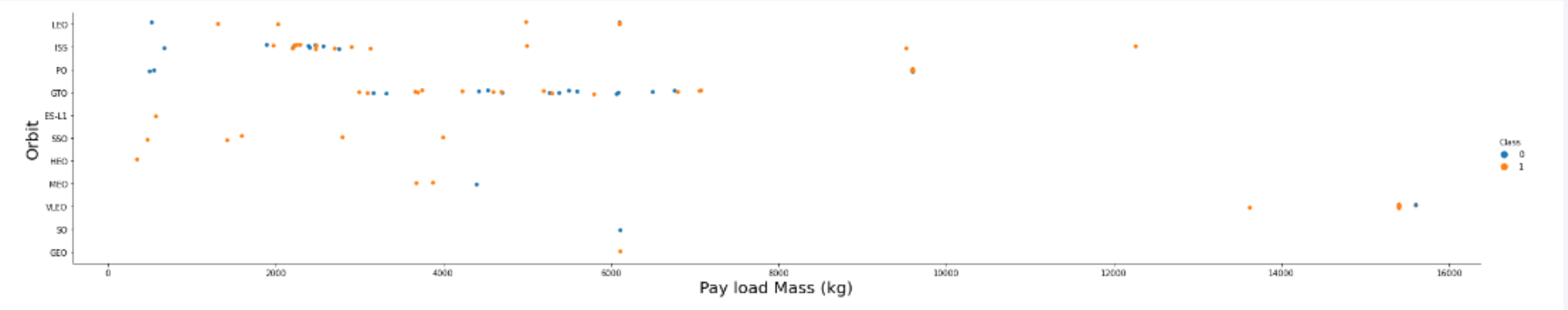
Flight Number vs. Orbit Type



According to the scatter plot,

- There can see success in the LEO orbit related to Flight Number.
- But between GTO orbit and Flight Number can't see a relationship with success.

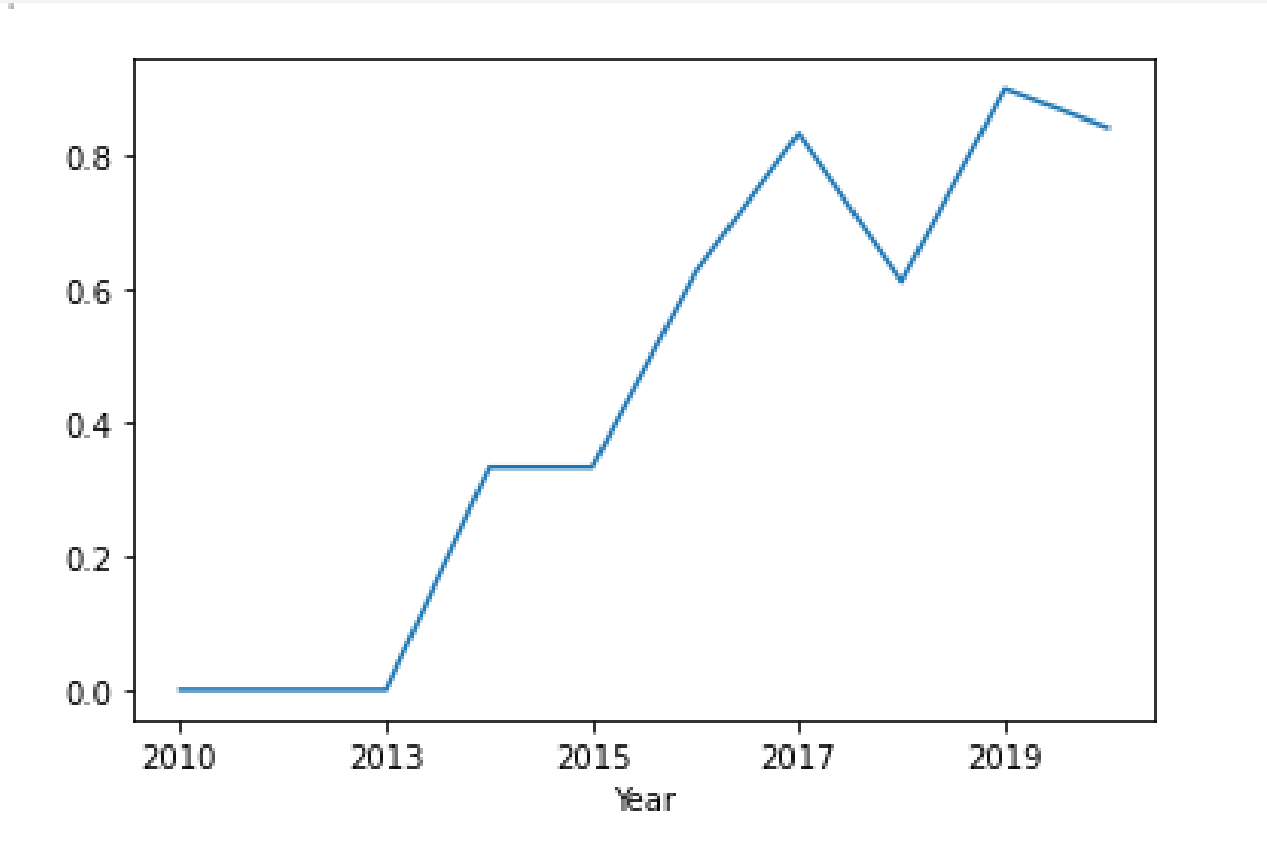
Payload vs. Orbit Type



According to the scatter plot,

- With heavy Payloads PO, LEO and ISS has a positive success rate.
- But GTO has mixed outcomes. So the relationship with Payload is not a clear pattern.

Launch Success Yearly Trend



According to the line chart of yearly average success rate,

- It seems more research and technology improvements has done in first 3 years.
- Success rate has started increasing since 2013.

All Launch Site Names

Launch Sites
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

According to the data,

- There are 4 unique launch sites in the space mission.
- These launch sites collected by taking unique values from column 'launch_site' in the dataset.

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

- These are the first 5 records where launch sites begin with `CCA`
- Also we can see that the mission outcomes are successful in all these five.

Total Payload Mass

Total Payload
111268

- The Total payload carried by boosters from NASA (CRS) has calculated as total_payload in query.
- Average School bus weight is 11,062kg.

Average Payload Mass by F9 v1.1

Average Payload
2928

- The average payload mass carried by booster version F9 v1.1 has calculated as avg_paload in query.
- It's 2928kg.

First Successful Ground Landing Date

First successful date on Ground pad
2015-12-22

- The first successful landing outcome on ground pad is on 2015-12-22.
- This maybe a very happy day because of the success in first stage landing on a ground other than on boats in the sea.

Successful Drone Ship Landing with Payload between 4000 and 6000

Booster Versions
F9 FT B1021.2
F9 FT B1031.2
F9 FT B1022
F9 FT B1026

- These are the boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

Mission Outcomes	Quantity
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- These are the total number of successful and failure mission outcomes.
- We can see the great amount of successful missions than unsuccessful missions clearly.
- These results were taken by grouping mission_outcome and counting records.

Boosters Carried Maximum Payload

Booster Versions
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

- This is the booster versions list which have carried the maximum payload mass.
- There are 12 F9 versions of Boosters.

2015 Launch Records

Booster Versions	Launch Sites
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

- These are the booster versions of failed Landing Outcomes on drone ships and their launch site names that launched in year 2015.

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

landing__outcome	qty
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

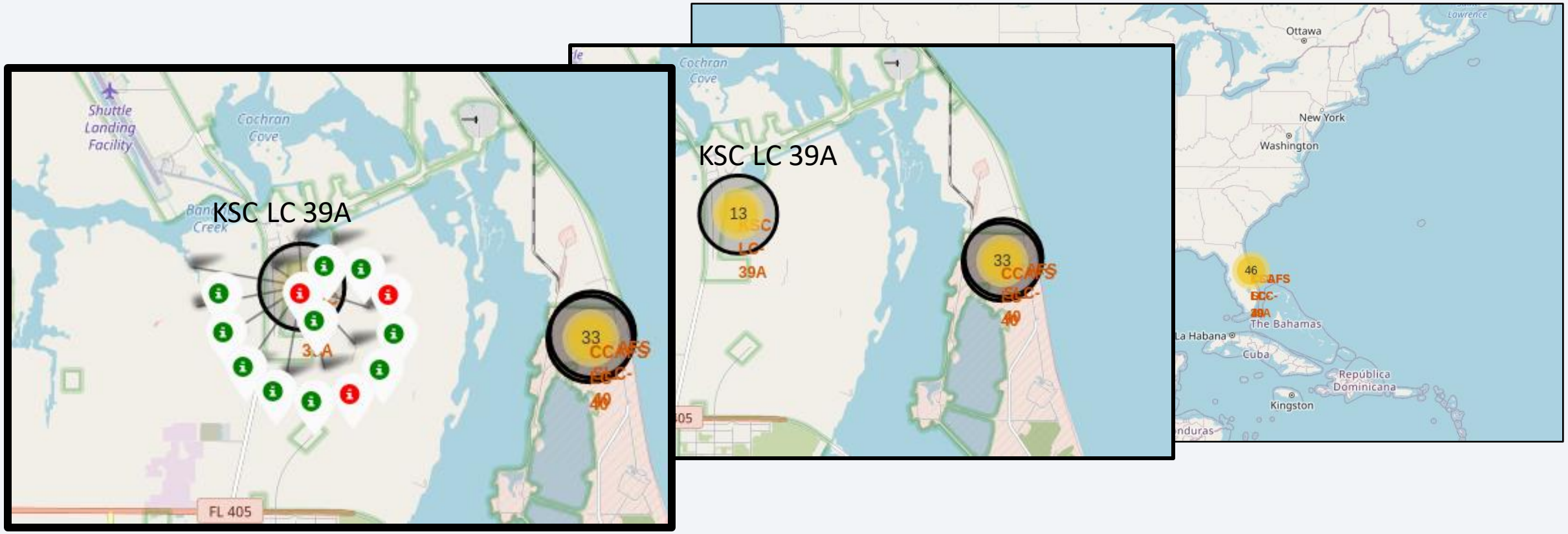
- Ranking of all landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order.
- In here 'No attempt' is also a landing outcome.


Section 3
Launch Sites
Proximities Analysis

All launch sites

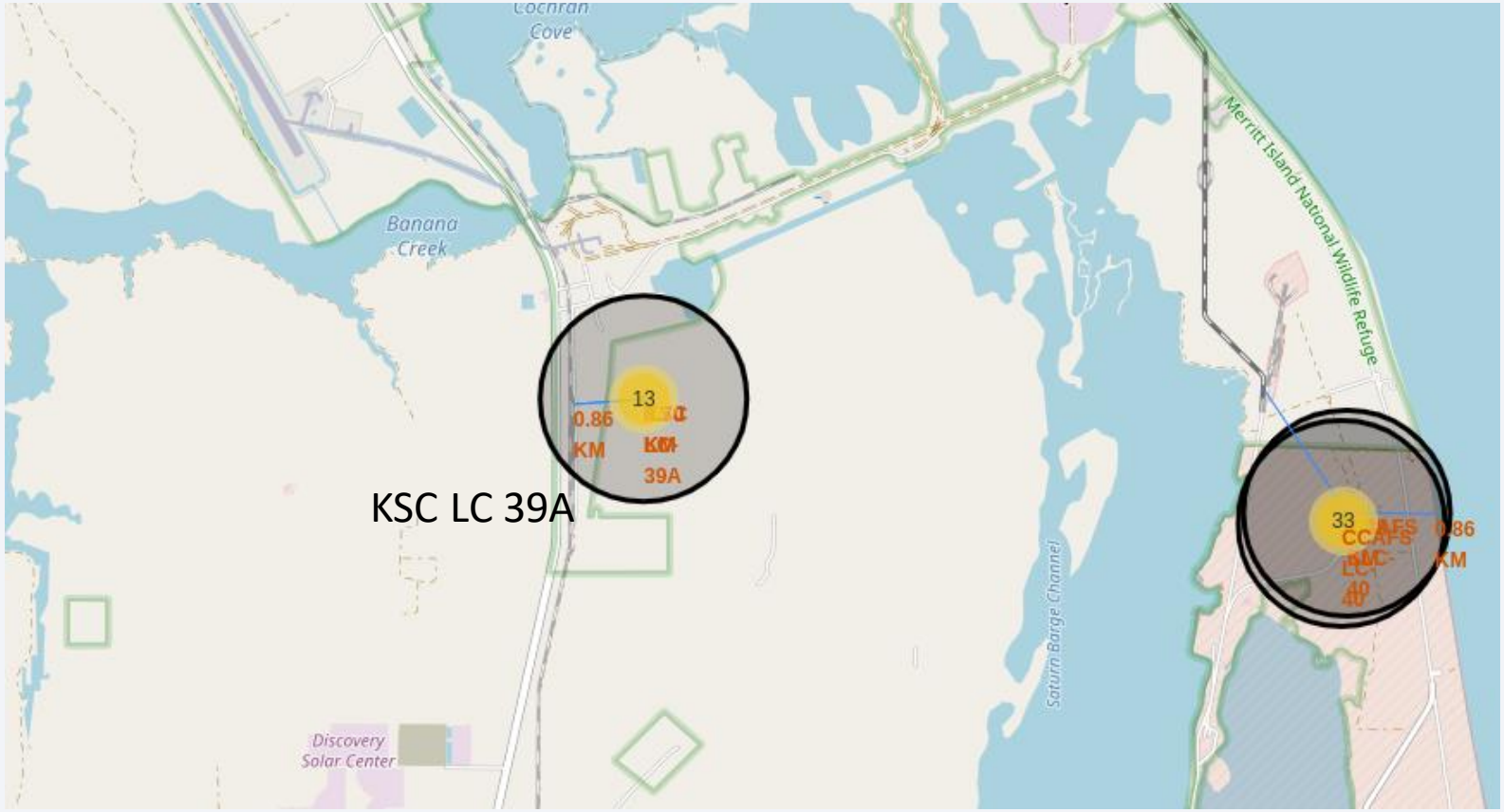
- Most of the launch sites located near sea, rail roads and roads. It can be for safety reasons and easiness to carry needed technical equipment.

Launch Outcomes in a selected Site

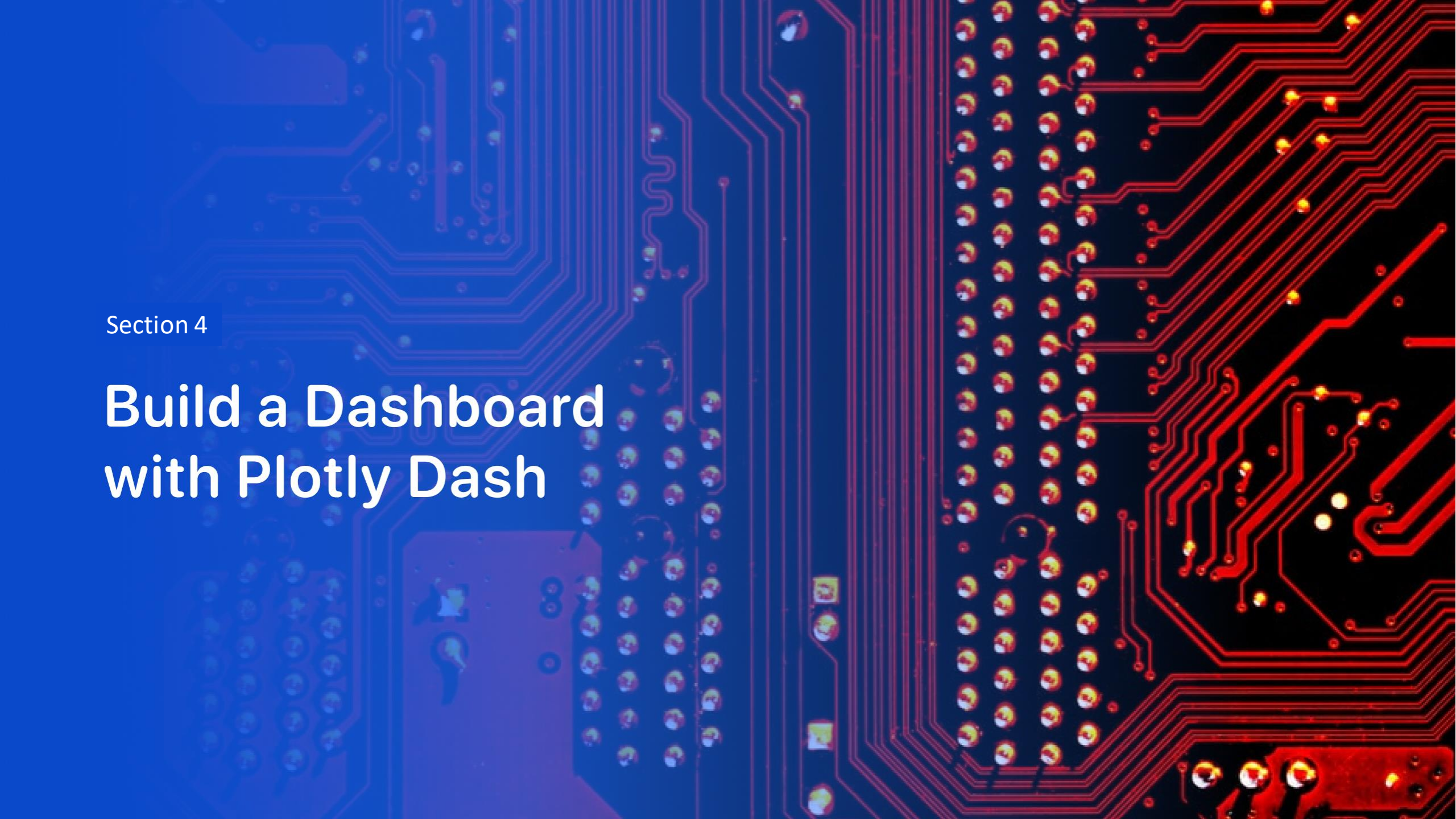


- This shows the KSC LC 39A launch site outcomes. Green markers indicates successful outcomes and red for failures. This site has green markers than other launch sites.

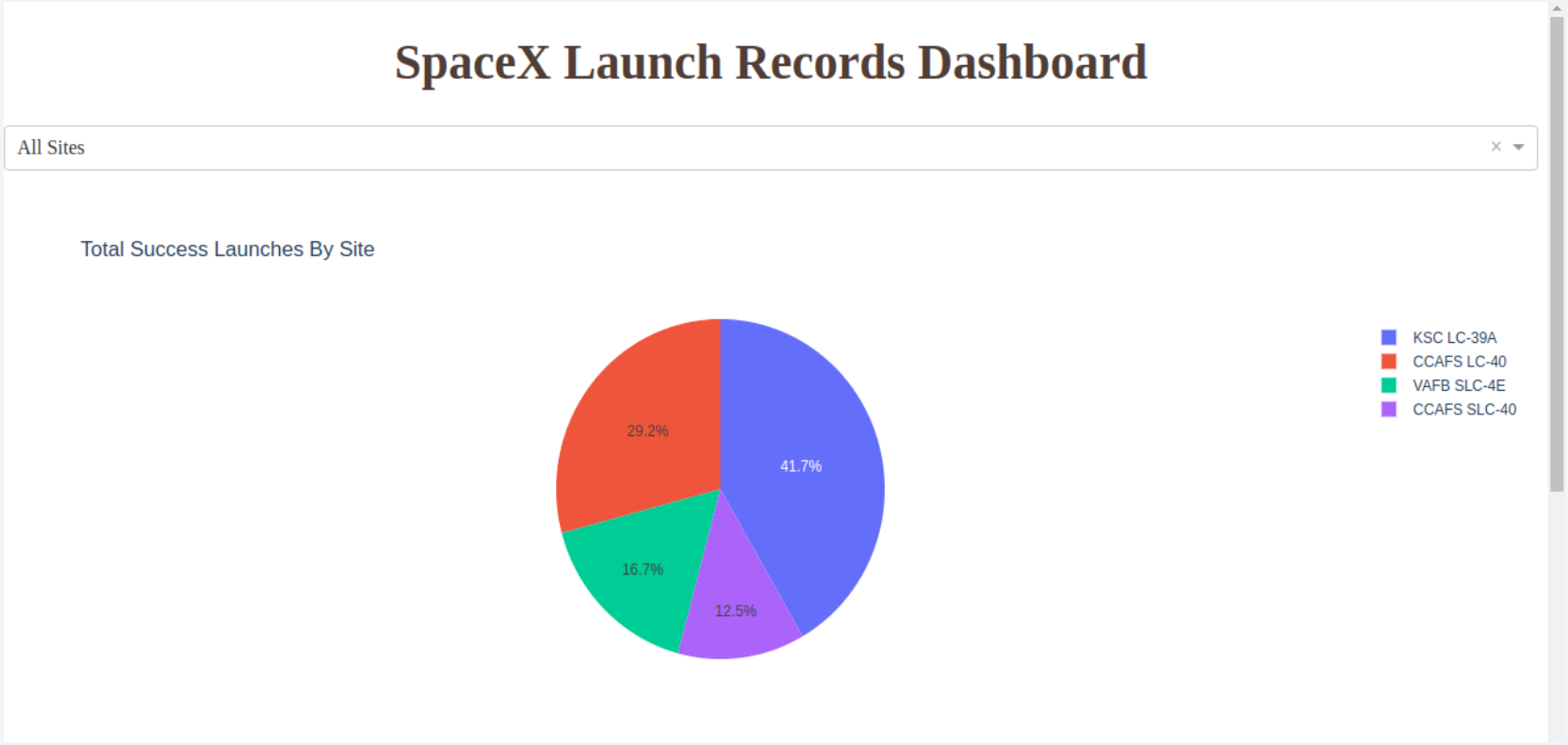
Relationship between Launch Sites and Proximities.



- KSC LC 39A launch site clearly near to the Railways, Road and sea.
- Also around this launch site, the around area is inhabited.
- So it's well suited for rocket launches.

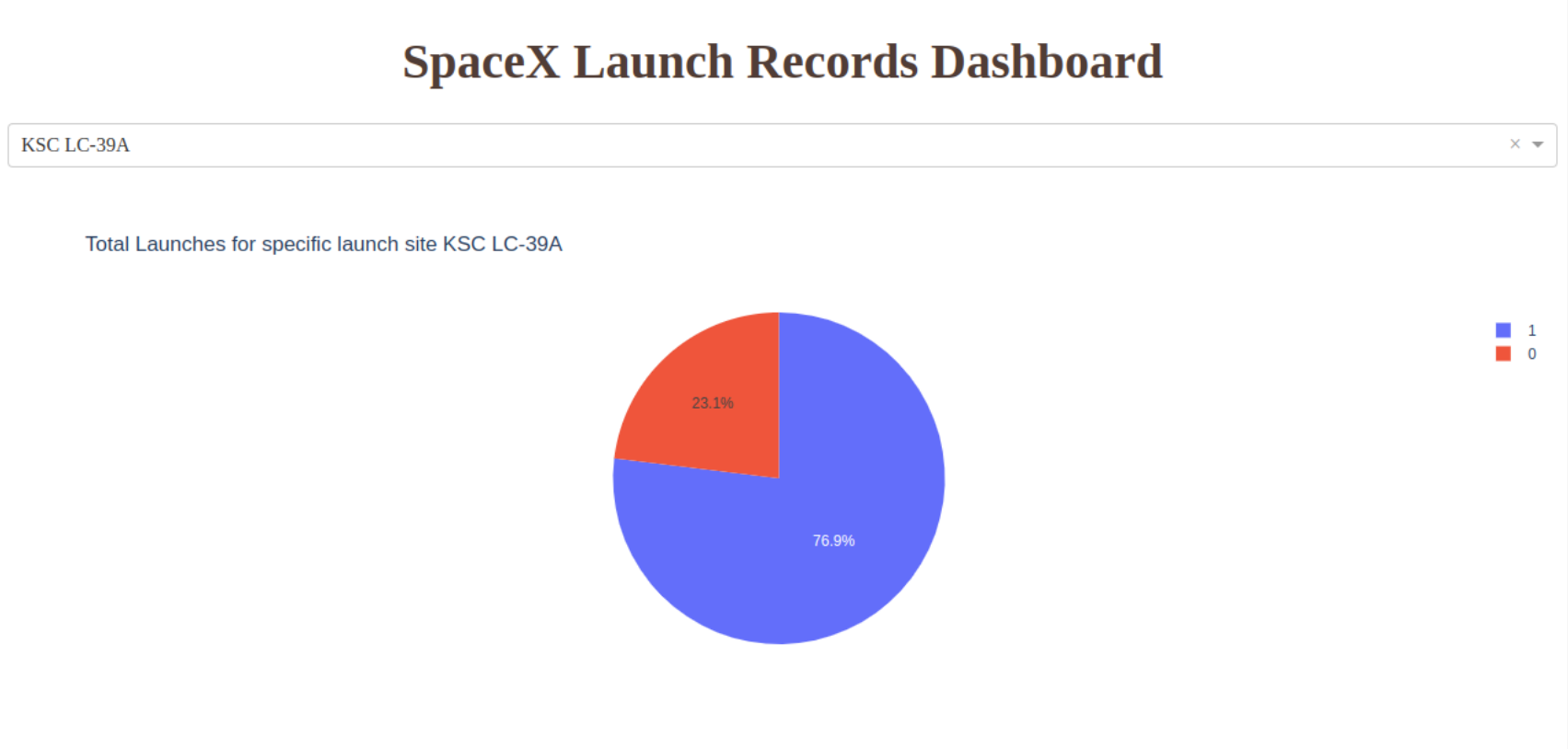

Section 4
Build a Dashboard with Plotly Dash

Total successful Launches count for each site



- KSC LC 39A launch site has the highest percentage. So the location of a launch site located is crucial.

Launch Site with the Highest Success Ratio

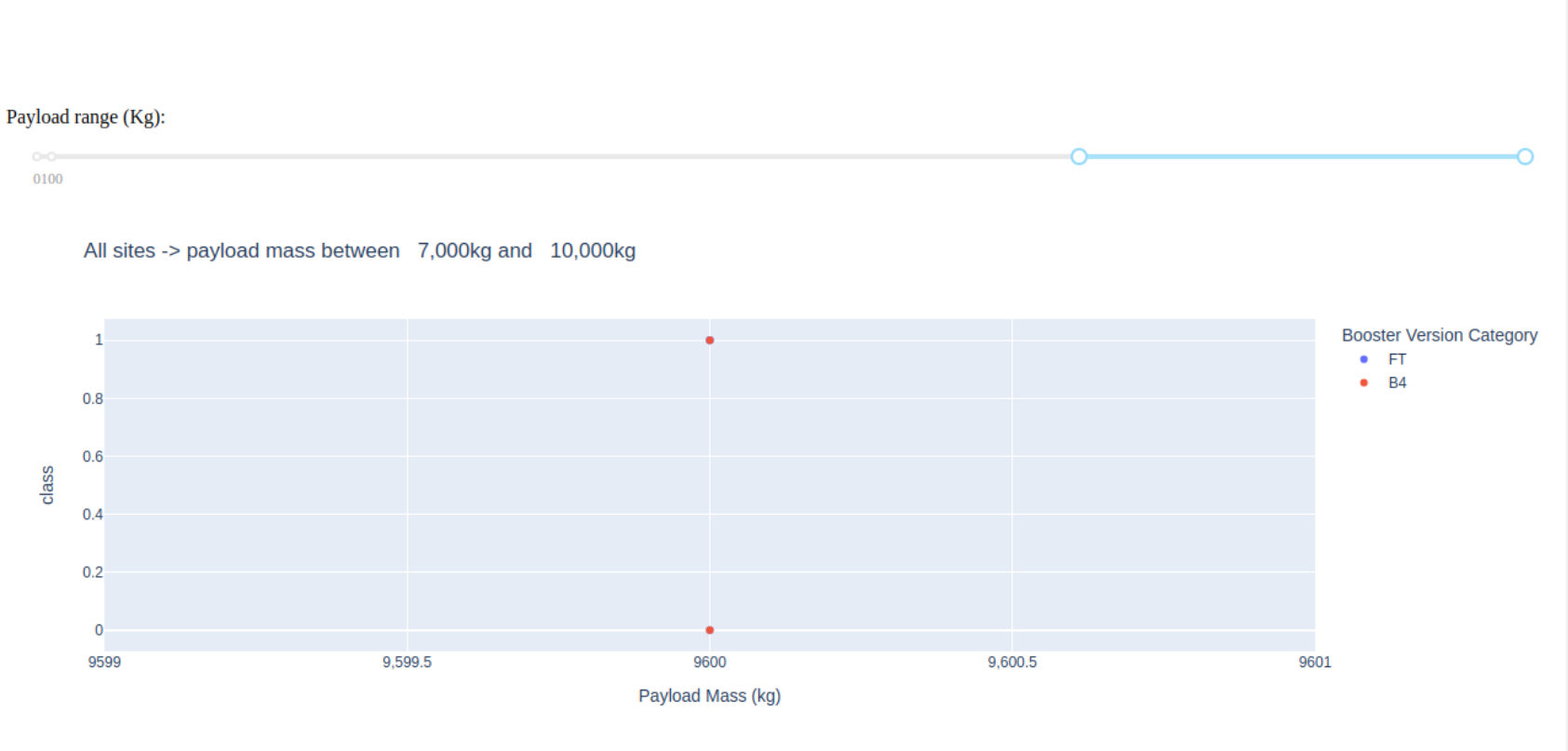


- It's KSC LC 39A launch site with ~76 % of success ratio.

Payload vs Launch Outcome

- F9 Boosters with Payload Mass less than 6000kg have the largest success rate.

Payload vs Launch Outcome

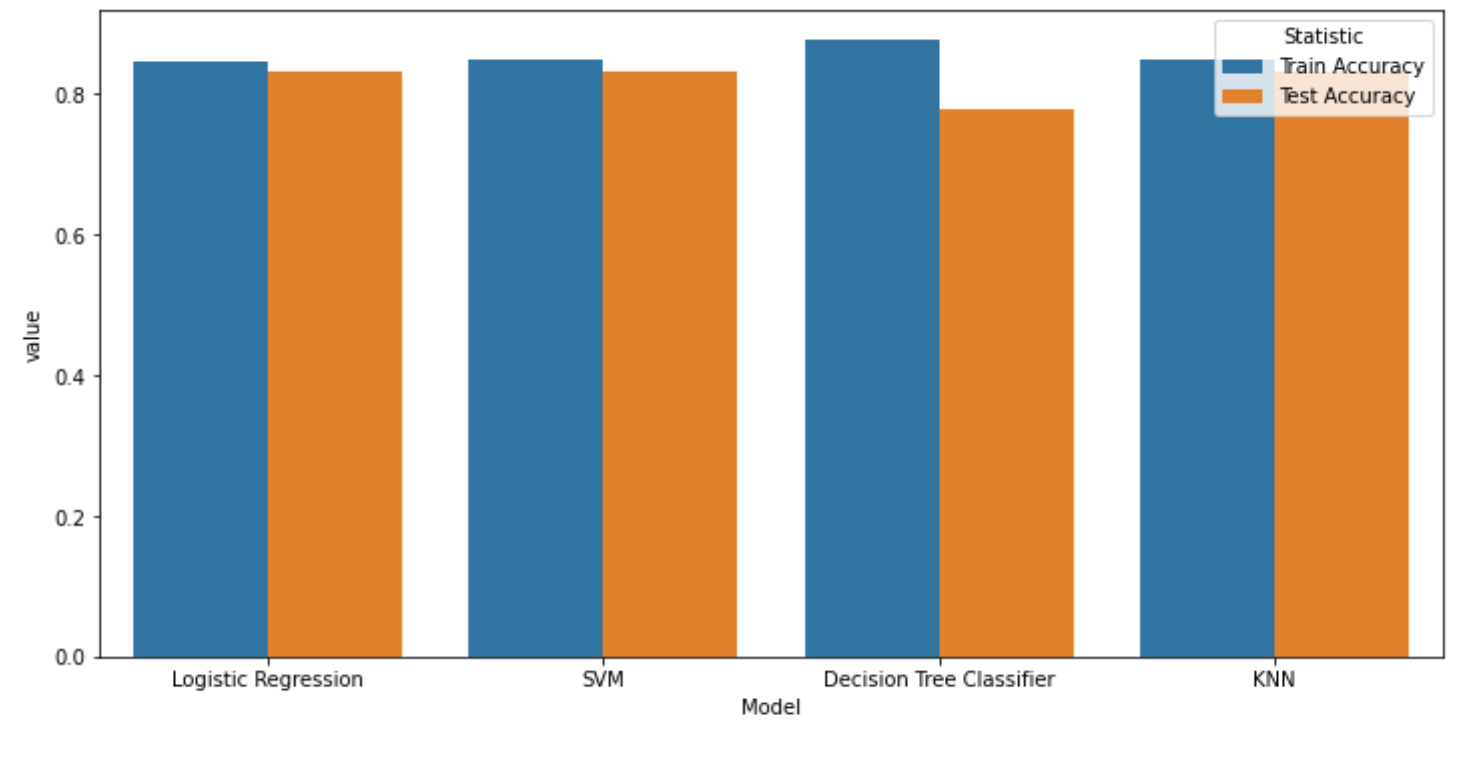


- There are no sufficient records of launches with Payload Mass greater than or equal to 7000kg.

Section 5

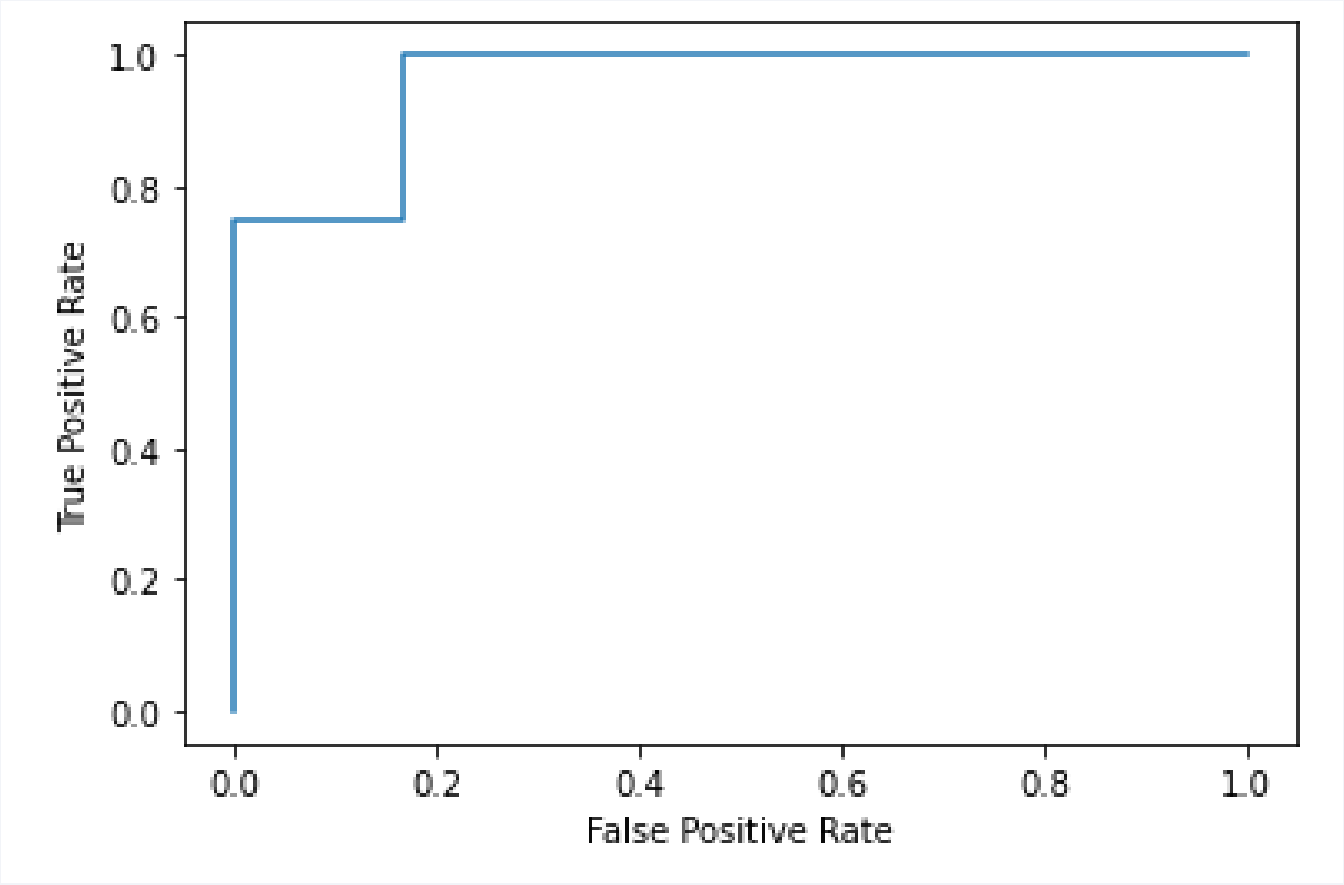
Predictive Analysis (Classification)

Classification Accuracy



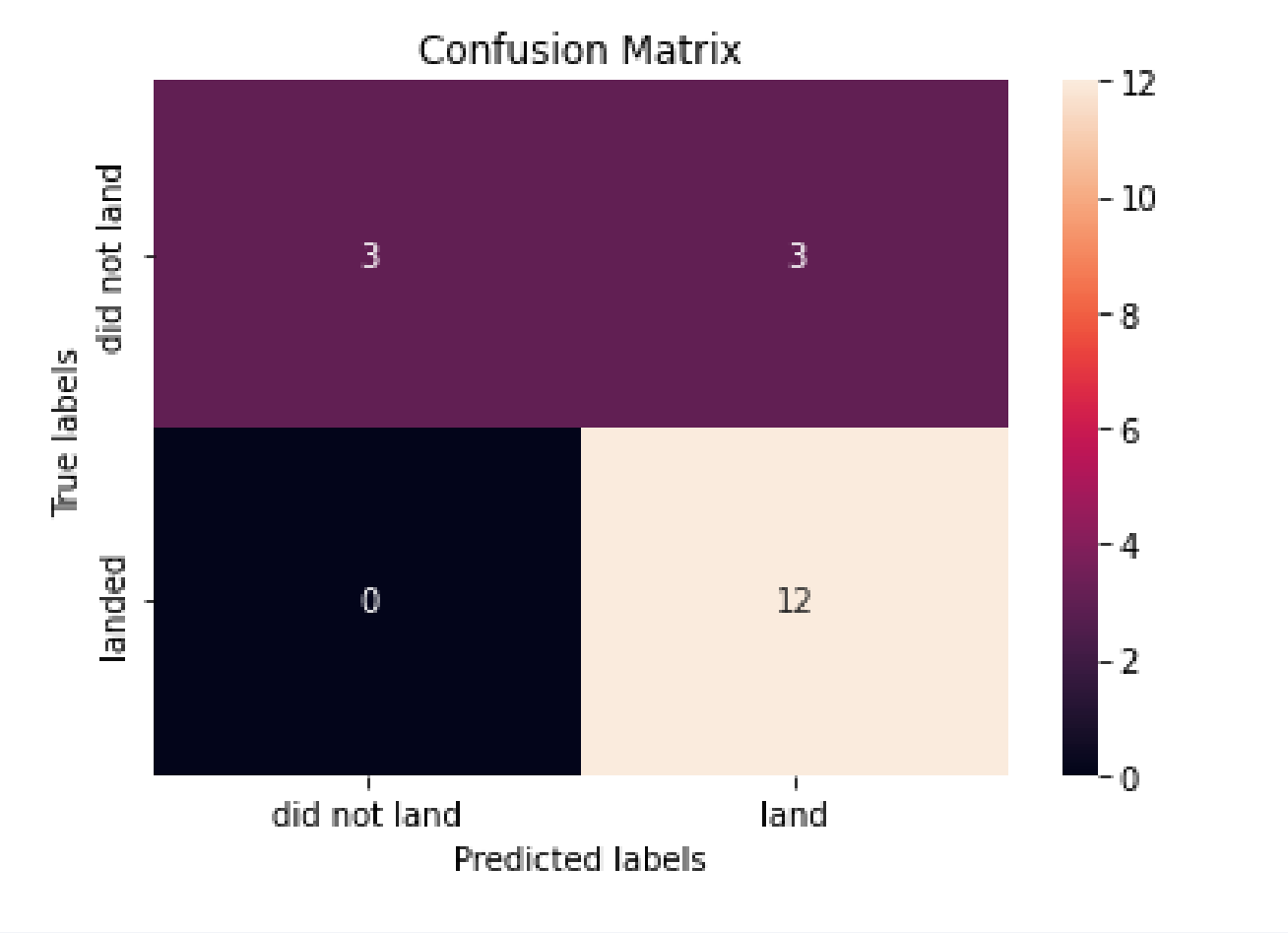
- These are the Train and Test accuracies got from the four models that built and evaluated.
- Here I got same results for Logistic Regression, SVM and KNN models.
- Decision Tree Classifier has a good Training accuracy but has a poor Test accuracy.
- So only with this bar chart can't decide which model has the highest classification accuracy.
- So drew ROC curves for models.

Classification Accuracy



- So this is the ROC curve I got for SVM Classifier.
- It has a good True Positive and False positive rate than other classifiers I built.
- Therefore, I take SVM as the model with highest classification accuracy.
- Other ROC curves for models can be found on the GitHub Notebook link shared on the Appendix 1A.

Confusion Matrix



- So this is the confusion matrix of the SVM Model as the best performing model.
- Here we can see Fully classified True positives and Partially Classified True Negatives by the SVM compared to the False ones.

Conclusions

- Collected data through API and Web Scraping and performed EDA.
- Created Interactive visualizations and Models for Evaluation.
- KSC LC 39A launch site is the ideal launch site for rocket launches.
- When the Payload is above 9000kg it is likely to be get successful outcomes.
- Most of the mission outcomes are successful but first stage landing outcomes seems to be success in future gradually.
- SVM Classifier can be used to predict successful Falcon 9 first stage landing outcomes and from evaluating the results we can make proper business decisions.

Appendix

- 1A – Predictive Analysis Notebook with bar chart and ROC curves for models -

https://github.com/RoshiniFernando/IBM-Applied-Data-Science-Capstone-Project/blob/main/Machine%20Learning%20Prediction%20Lab.ipynb

- GitHub Link for the Main Repository containing all Notebooks with code and visualizations - https://github.com/RoshiniFernando/IBM-Applied-Data-Science-Capstone-Project

Thank you!