Winning Space Race with Data Science

Name: Rindhuja Treesa Johnson
Date: 20th April 2023

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results

- Data Collection: SpaceX API and web Scraping
- Exploratory Data Analysis (EDA): Data Wrangling, Data visualization, SQL
- Prediction: Machine Learning – Found the accuracy of different models

✓ Determined the parameters that could accurately predict successful launchings.
✓ Determined the ML models that will be most efficient and accurate.

Introduction

- Project background and context
- Problems you want to find answers

- Space technology has procured its economic relevance since the first time a spacecraft soared into space.
- The expenses involved in sending a ship out into space have been intense until recently when SpaceX introduced reusing the first stage in launching.
- Data suggests that it cuts the cost by half when a first stage is saved successfully. There were instances when SpaceX successfully retained the first stage of their Falcon 9 launches.
- This work aims to use the data from SpaceX and analyze how the profit can be maximized by adjusting different dependent factors.
- Machine Learning models are tested for accuracy to determine the most effective one.

Section 1

Methodology

Methodology

- Data collection methodology:
  - Web Scraping
  - SpaceX API
- Perform data wrangling
  - The type of orbits, launching sites, and landing sites were summarized to get an idea of how the success in retaining the first stage was related to them.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Dataset was normalized and split into train-test data sets, and classification models were trained.
  - Log regressions, Support Vector Machine, KNN, and Decision tree models were tested for accuracy

Data Collection

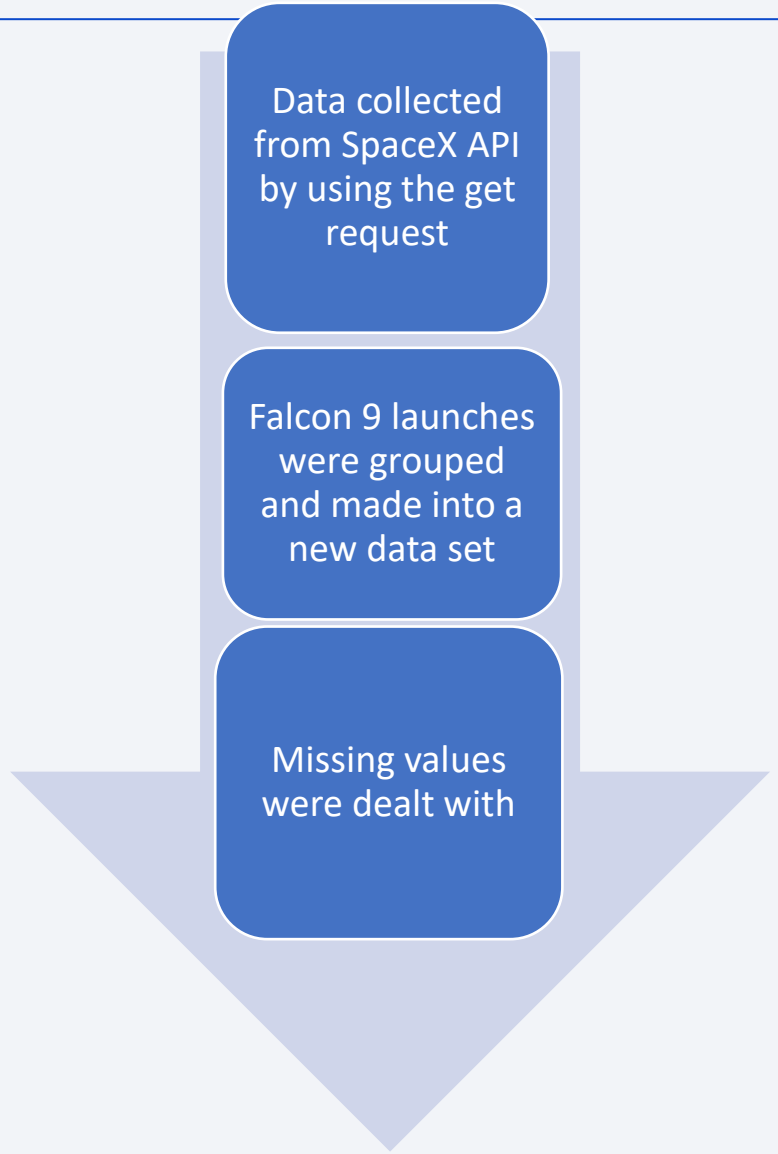
- Describe how data sets were collected.
- You need to present your data collection process using key phrases and flowcharts

- Web Scraping: From Wikipedia
- API: SpaceX API

Data Collection – SpaceX API

- Data Collection and missing values were dealt with

- Data Collection – SpaceX API link - Data Collection using API



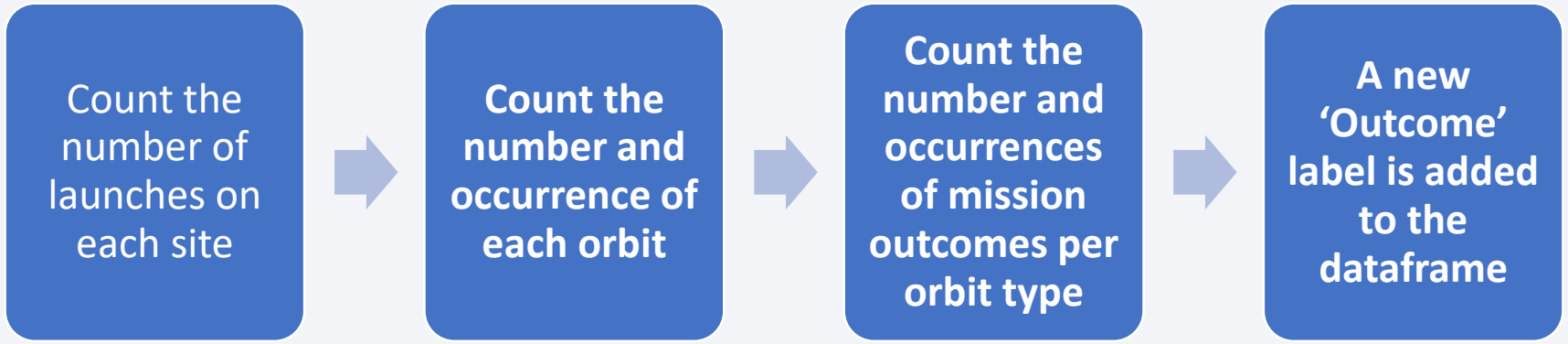
Data Collection - Scraping

- Falcon 9 wiki page was requested for web scraping
- Web Scraping – Wikipedia - Data Collection by Web Scraping

Data Wrangling

- Describe how data were processed
- You need to present your data wrangling process using key phrases and flowcharts
- Add the GitHub URL of your completed data wrangling related notebooks as an external reference and for peer-review purpose

- GitHub Link: Data Wrangling
- The data was analyzed for missing values after loading it into a data frame.



EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts
- Add the GitHub URL of your completed EDA with data visualization notebook, as an external reference and peer-review purpose

- GitHub link: EDA with Data Visualization
- Cat plot is used to visualize the distribution of successful landings with respect to different parameters
- Bar graph: to study the occurrences of a particular event.

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed
- Add the GitHub URL of your completed EDA with SQL notebook, as an external reference and peer-review purpose

GitHub URL: EDA with SQL

- SELECT DISTINCT(LAUNCH_SITE) FROM SPACEXTBL
- SELECT * FROM SPACEXTBL
  WHERE launch_site LIKE('CCA%')
  LIMIT 5
- SELECT customer,SUM(payload_mass__kg_) AS Total_Payload_Mass FROM SPACEXTBL
  WHERE customer = 'NASA (CRS)'
  GROUP BY customer
- SELECT booster_version, AVG(payload_mass__kg_) AS Average_Payload_mass FROM SPACEXTBL
  WHERE booster_version = 'F9 v1.1'
  GROUP BY booster_version
- SELECT * FROM SPACEXTBL
  WHERE Landing_Outcome = 'Success (ground pad)'
  ORDER BY DATE ASC
  LIMIT 1
- SELECT booster_version, payload_mass__kg_ FROM SPACEXTBL
  WHERE landing_outcome = 'Success (drone ship)' AND (payload_mass__kg_ BETWEEN 4000 AND 6000)
- SELECT mission_outcome, COUNT(mission_outcome) FROM SPACEXTBL
  GROUP BY mission_outcome
- SELECT booster_version, payload_mass__kg_ FROM SPACEXTBL
  WHERE payload_mass__kg_ =(SELECT MAX (payload_mass__kg_) FROM SPACEXTBL)
- SELECT substr(Date,4,2) as Month, landing_outcome, booster_version, launch_site FROM SPACEXTBL
  WHERE substr(date,7,4) = '2015' AND landing_outcome = 'Failure (drone ship)'
- SELECT date, landing_outcome FROM SPACEXTBL
  WHERE landing_outcome LIKE('Success%') AND (Date BETWEEN '04-06-2010' AND '20-03-2017')
  ORDER BY substr(date,7,4) DESC

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
- Explain why you added those objects
- Add the GitHub URL of your completed interactive map with Folium map, as an external reference and peer-review purpose

- GitHub Link: Interactive Visual Analytics with Folium
- The map objects used are:
- 1) Circle – it marks the coordinates of a specific location with a pop-up
- 2) Marker - it marks the coordinates of a specific location with its name
- 3) Mark clusters – it reduces confusion on the map marked with many markers closely.
- 4) Polyline – connects to locations on the map

Build a Dashboard with Plotly Dash

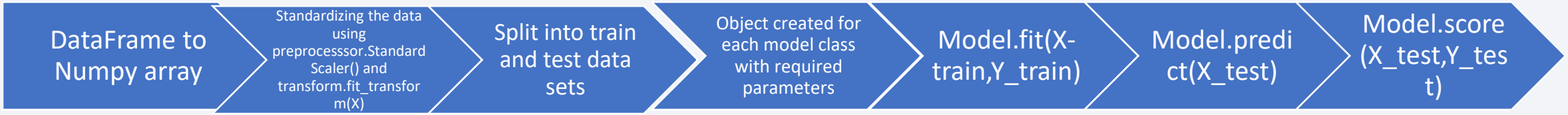
- Summarize what plots/graphs and interactions you have added to a dashboard
- Explain why you added those plots and interactions
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose

- GitHub Link: Interactive Dashboard with Plotly Dash
- Drop Down Menu: this enables the user to choose the desired launch site
- Pie chart: This visualizes the percentage of successful launches from each launch site.
- Range slider: This allows the user to choose the payload mass range of the satellite to od analysis
- Scatter plot: It gives an insight into the number of successful and failed launches with respect to payload mass of different booster versions

Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model
- You need present your model development process using key phrases and flowchart
- Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose

- GitHub link: SpaceX Prediction Models
- The values in the data frame were normalized using the StandardScaler() function after converting the whole thing into a numpy array.
- After this, the dataset was divided into the training and test sets to be used for the modeling.
- The classification methods used are Logistic regression, Support vector machine, decision tree, and K – Nearest neighbor.
- For each of the above models, upon completion of modeling using the training dataset, it was run on the test dataset. The result was compared with the given y_test and accuracy score was calculated.
- The best method could be selected from these models, which has the highest accuracy score.



Results

- Exploratory data analysis results

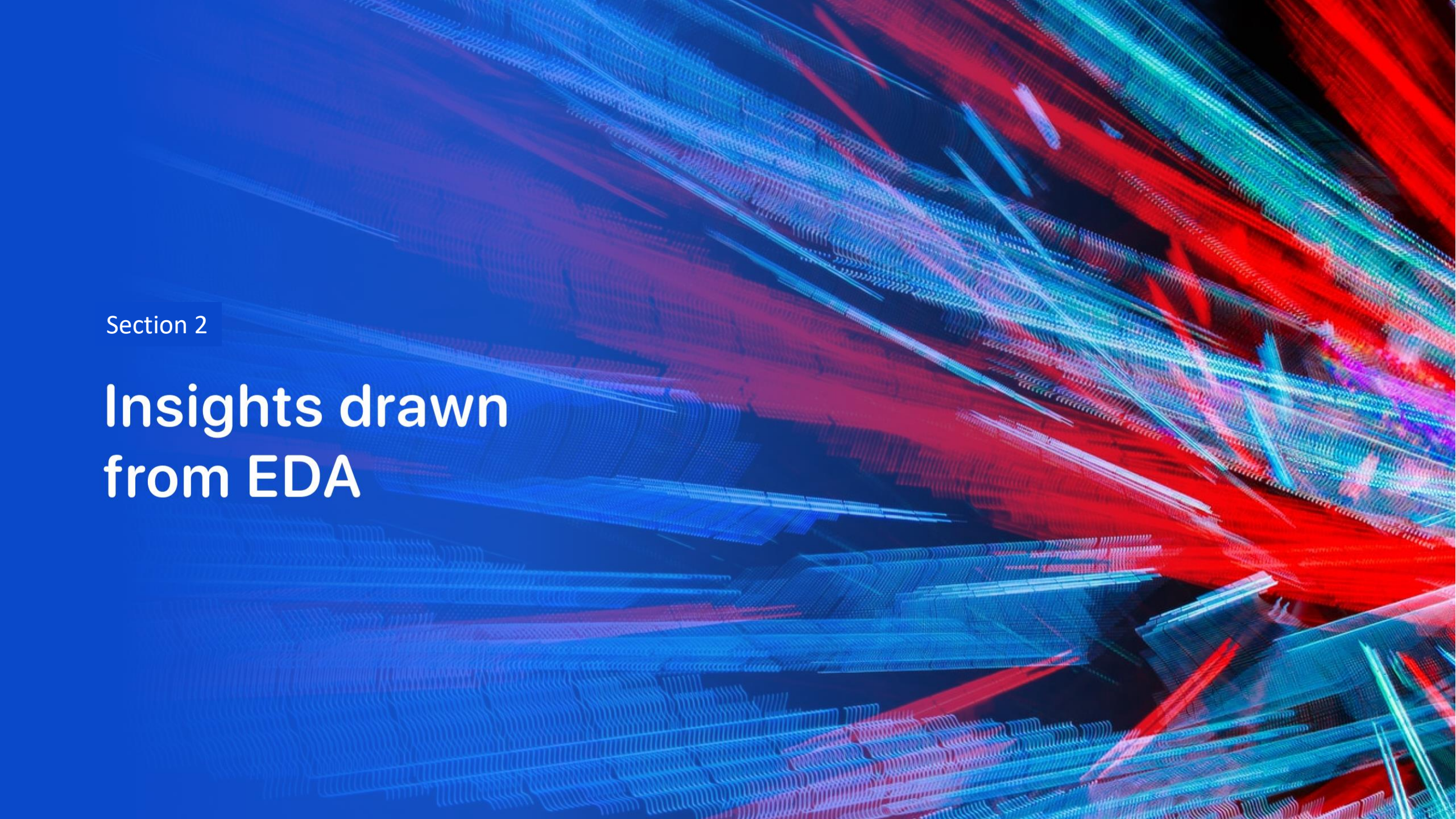
i. There are four different launch sites. Two along each coastline of the US.

ii. The first successful launch was on 01-05-2017 from KSC LC-39A

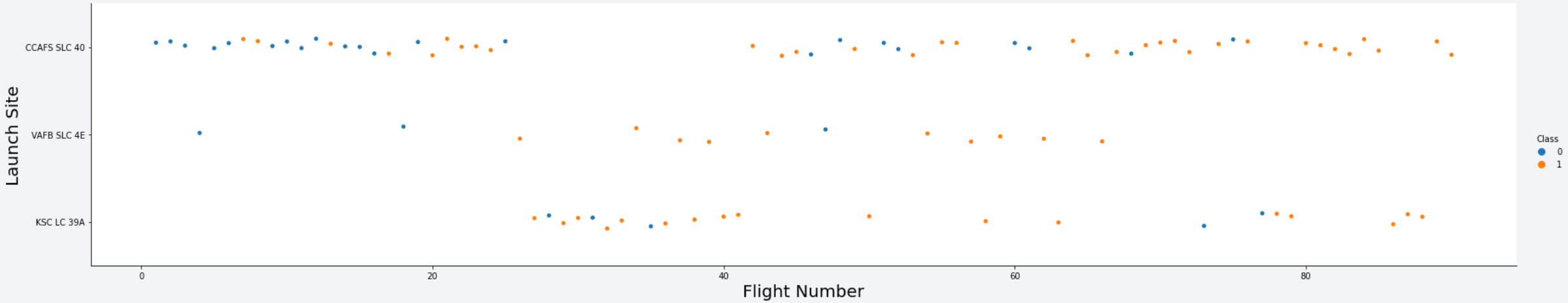
- Interactive analytics demo in screenshots

i. Although there were equal number of launch sites on the two coastlines. The East coast has more launches and are well connected by roads and rails.

- Predictive analysis results
  - The decision tree Classifier tends to have a higher accuracy as a model than the other classification models

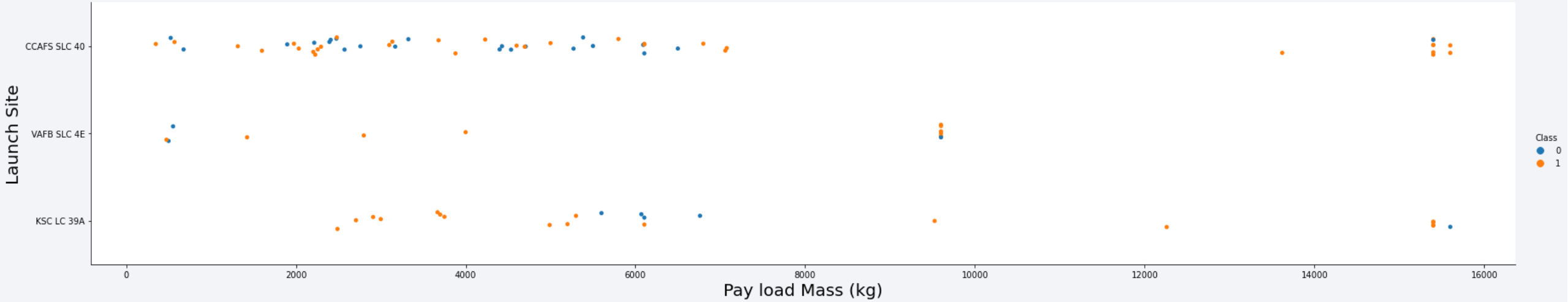

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



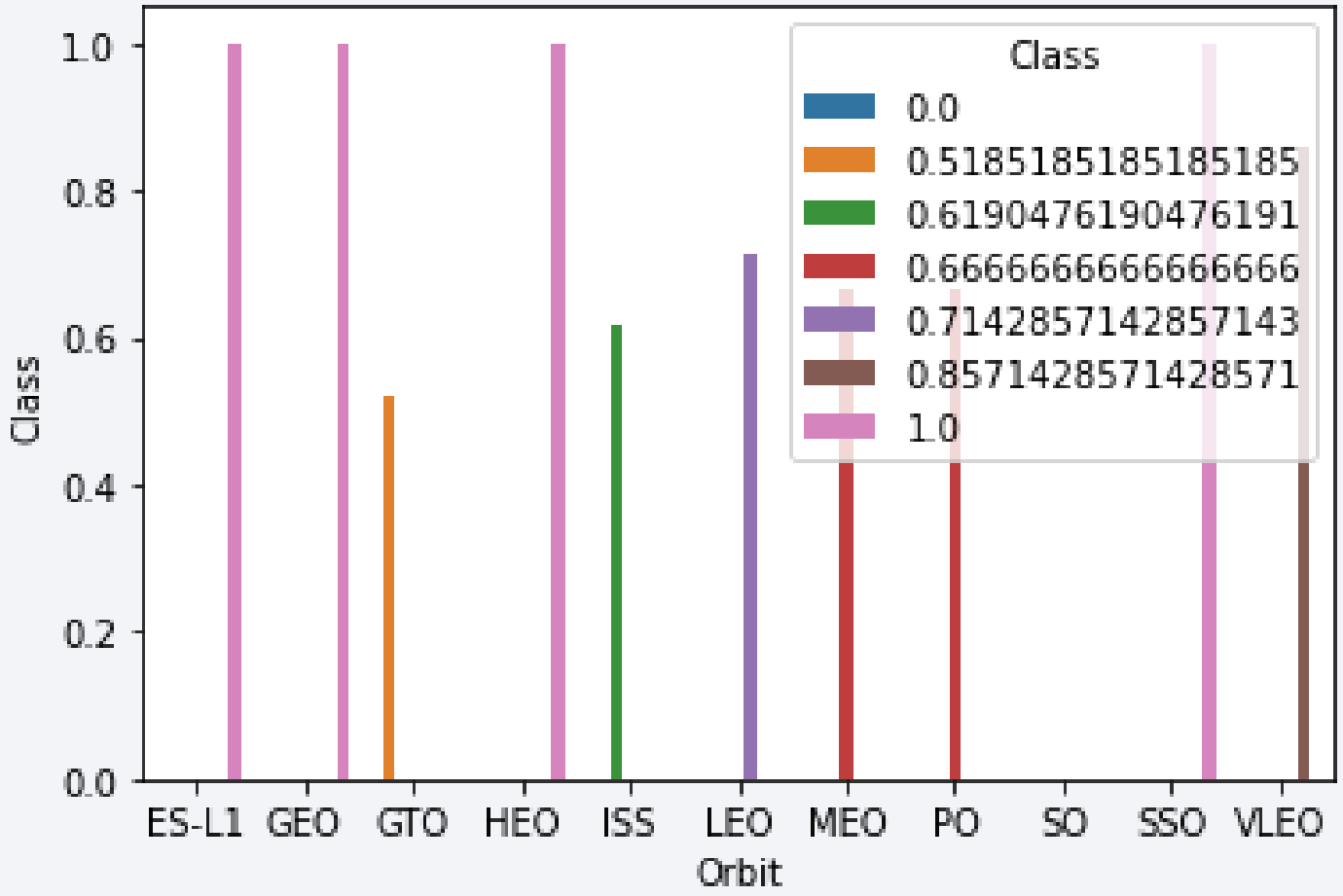
- Most of the launches were from the site CCAF SLC40 and has a high success rate.
- The number of successes has an increase with the Flight number

Payload vs. Launch Site



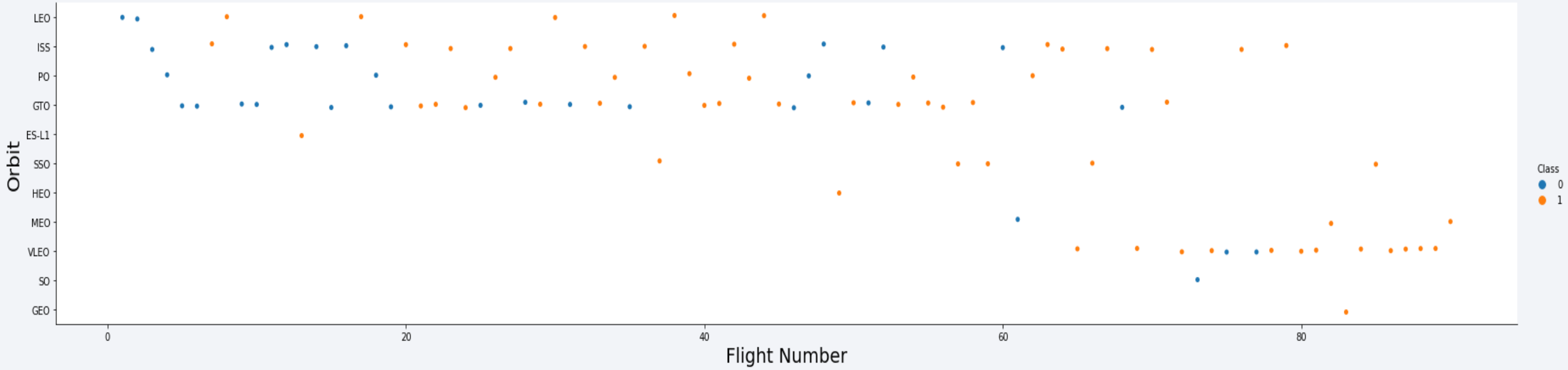
- No rockets were launched from VAFB-SLC with a payload mass greater than 1000.
- The success from the other two launch sites are higher with greater payload mass.

Success Rate vs. Orbit Type



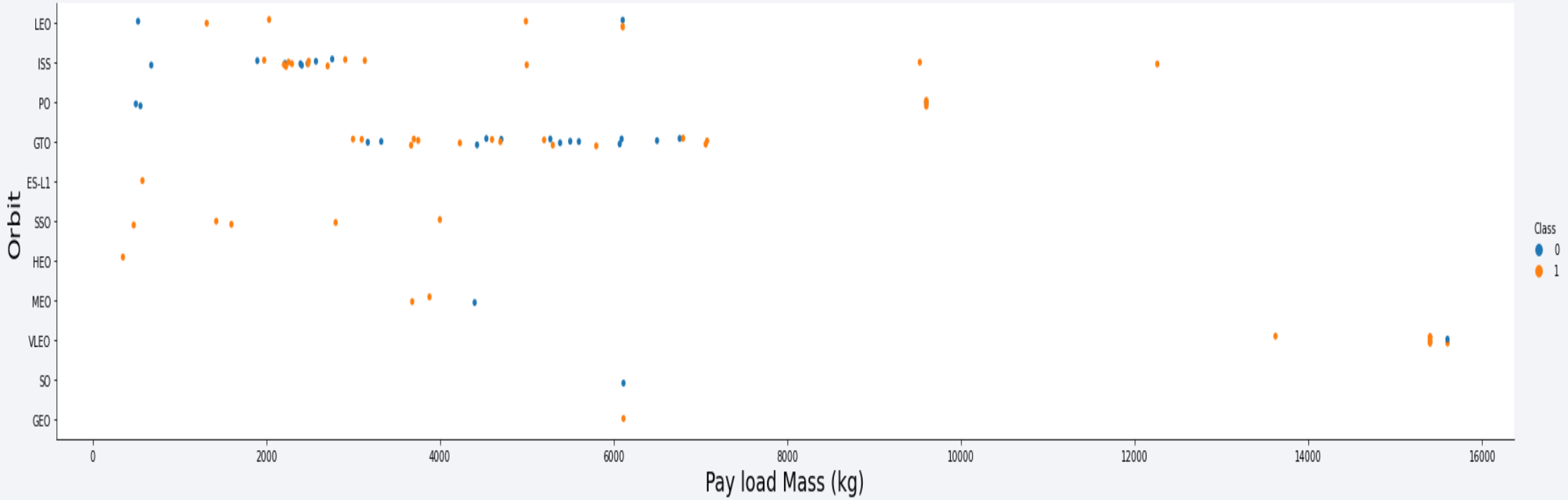
- The orbits with maximum success (class = 1) are ES-L1, GEO, HEO, SSO
- VLEO and LEO also have considerable success.

Flight Number vs. Orbit Type



- Most successful launches happened with Flight numbers from 60 and above with the orbit VLEO

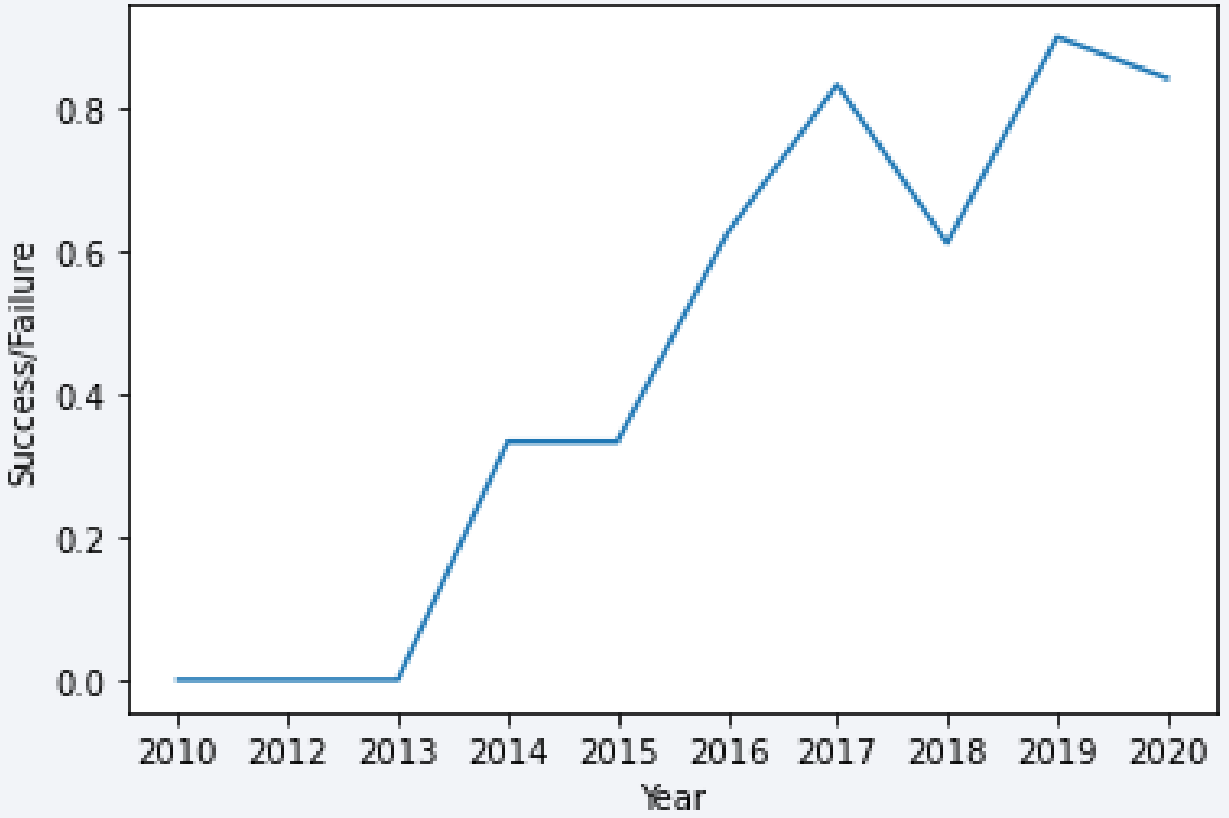
Payload vs. Orbit Type



- Polar, LEO, and ISS has higher success rates with heavy payloads.
- GTO and ISS have more number of launches that others.

Launch Success Yearly Trend

- The success tends to have an increasing trend
- However, it was stagnant in the initial years and had a nose dive in 2018.



All Launch Site Names

- Find the names of the unique launch sites:

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

Present your query result with a short explanation here

These are the unique launch sites

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- Present your query result with a short explanation here

These are the details of the launch sites that begin with ‘CCA’

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

Customer	Total_Payload_Mass
NASA (CRS)	45596

- Present your query result with a short explanation here

The total payload mass carried from the NASA boosters are found using SUM() function.

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1

Booster_Version	Average_Payload_mass
F9 v1.1	2928.4

- Present your query result with a short explanation here

The average of the payload mass carried by version F9 v1.1 is obtained using AVG() function.

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- Present your query result with a short explanation here:

The first successful landing was on 5th January 2017 on a ground pad.

Date	Time (UTC)	Booster_ Version	Launch_ Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_ Outcome	Landing_ Outcome
01-05-2017	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)

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
- Present your query result with a short explanation here

There were only four boosters that have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

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

- Calculate the total number of successful and failure mission outcomes
- Present your query result with a short explanation here

There were 100 successes in total among which one has an unclear payload status.

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Present your query result with a short explanation here

There were 12 versions which carried the maximum payload mass which is 15600kg.

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Present your query result with a short explanation here

There were two booster versions that failed on drone ship.

Month	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

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

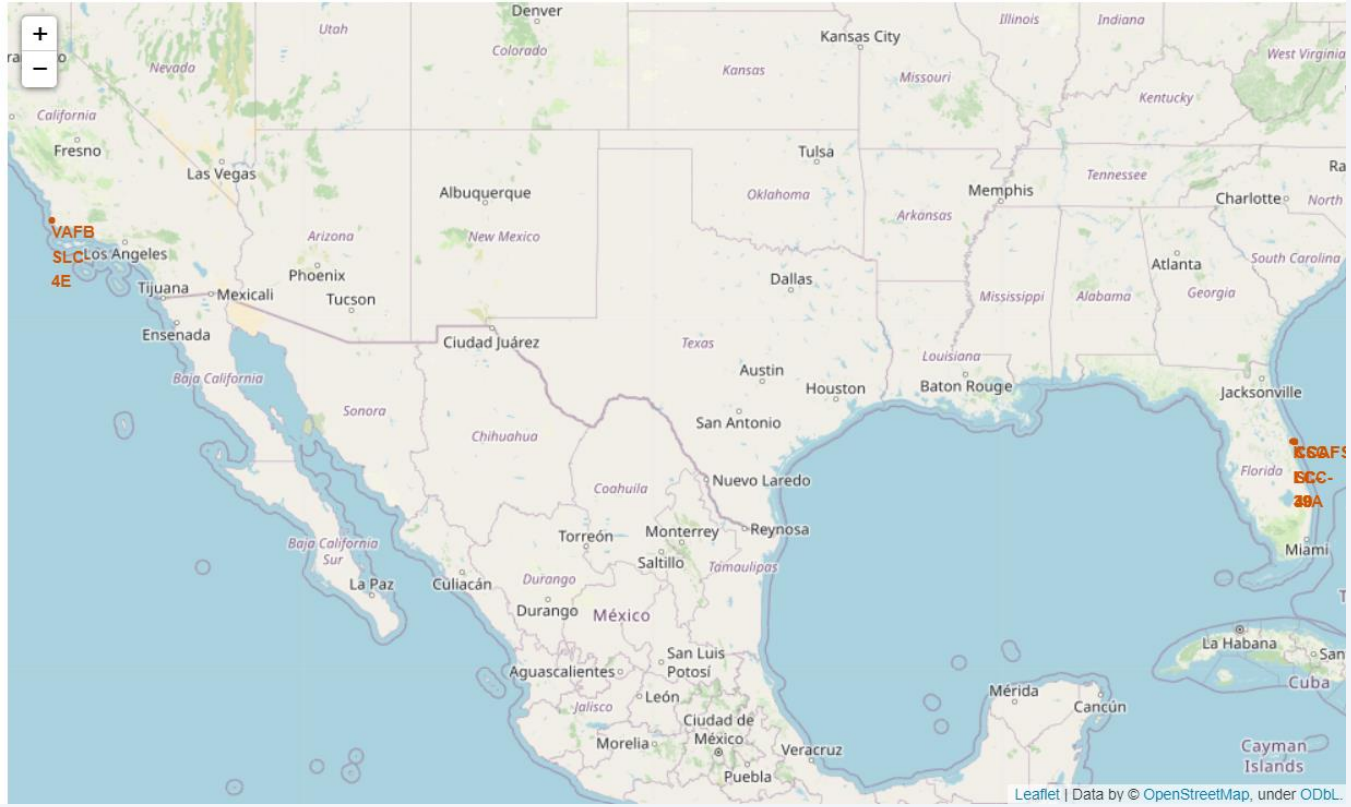
- The data gives a positive insight as the higher numbers of landings were success

Landing_Outcome	Number_of_landings
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

Section 3

Launch Sites Proximities Analysis

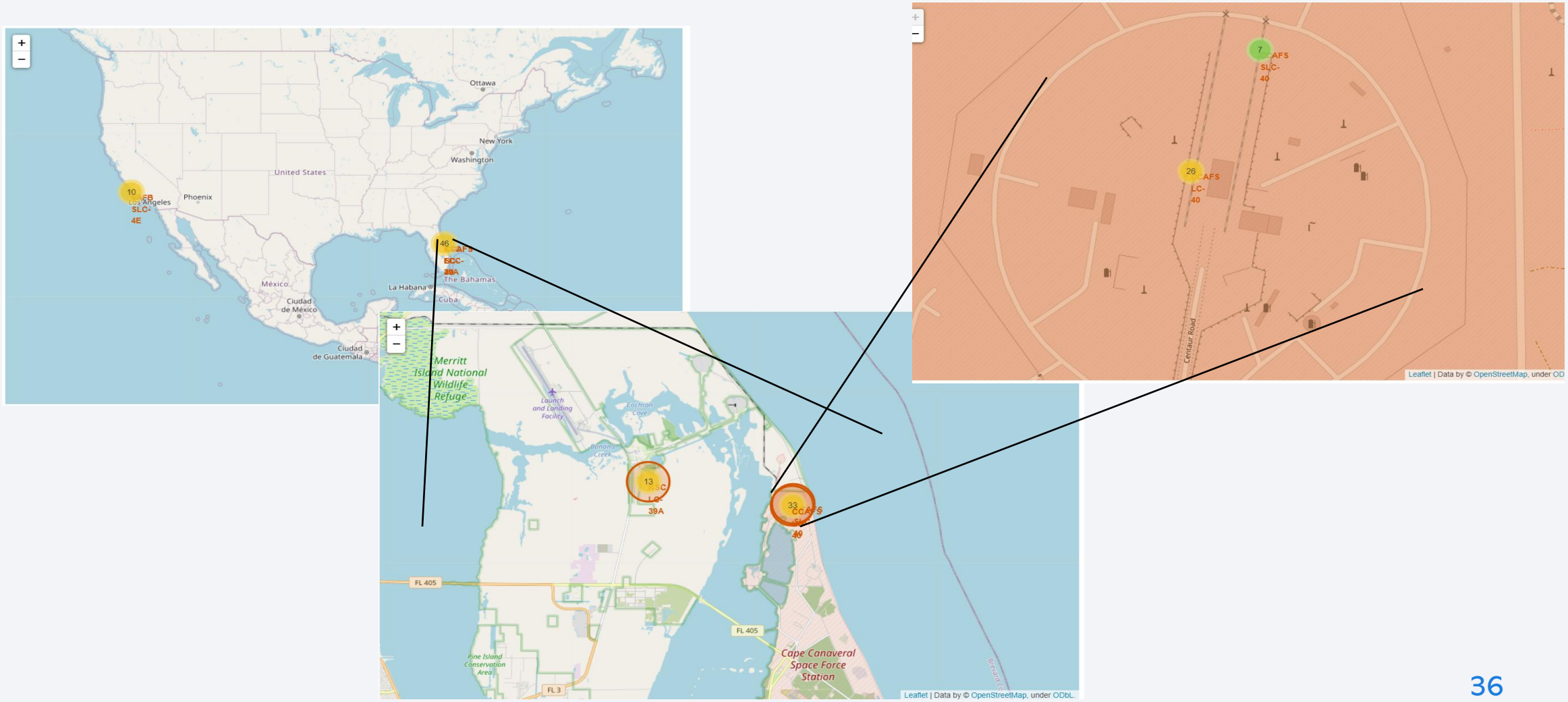
Launch Sites



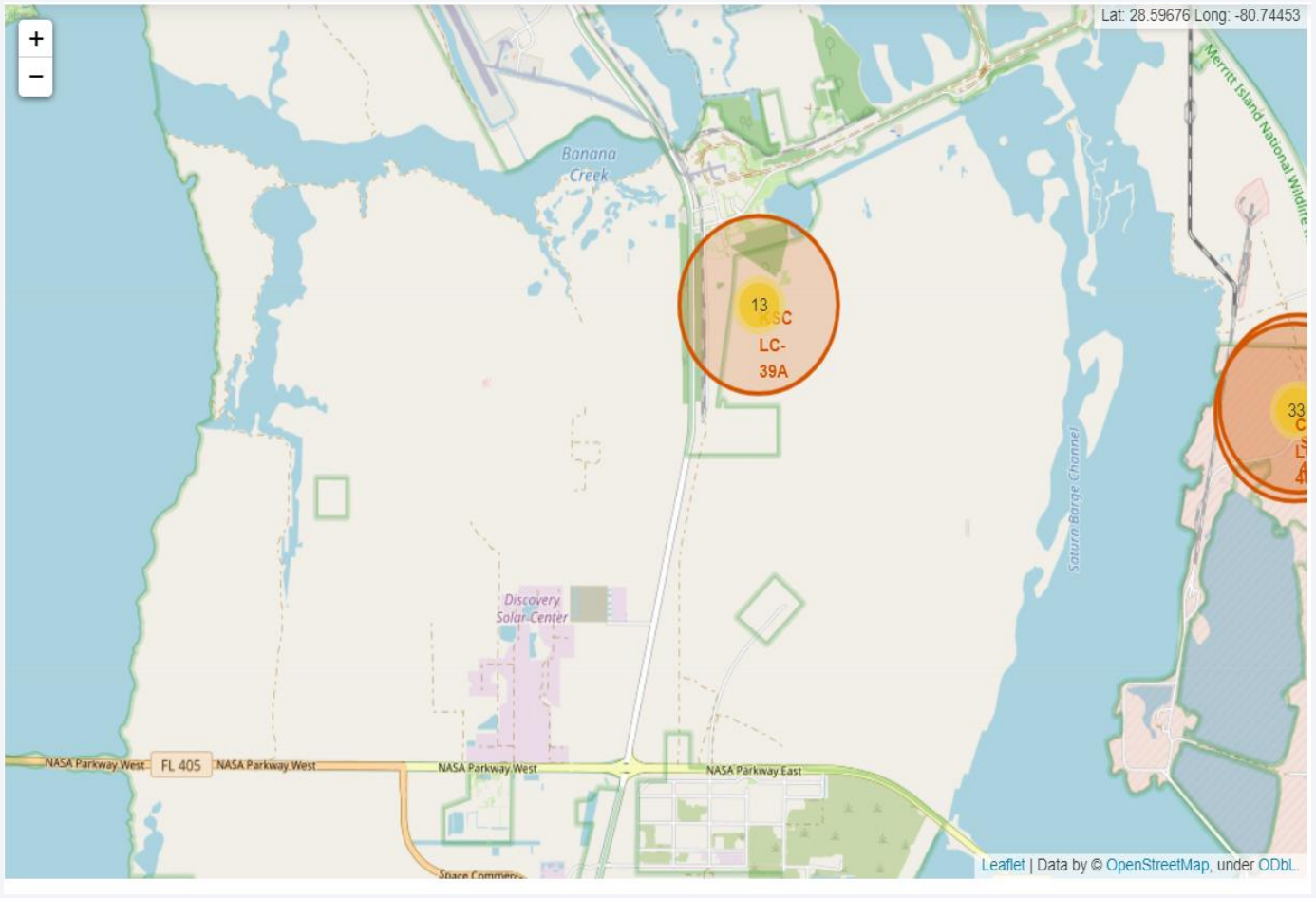
- All the launch sites appear to be along the coastal line.

Was it a successful Launch Site?

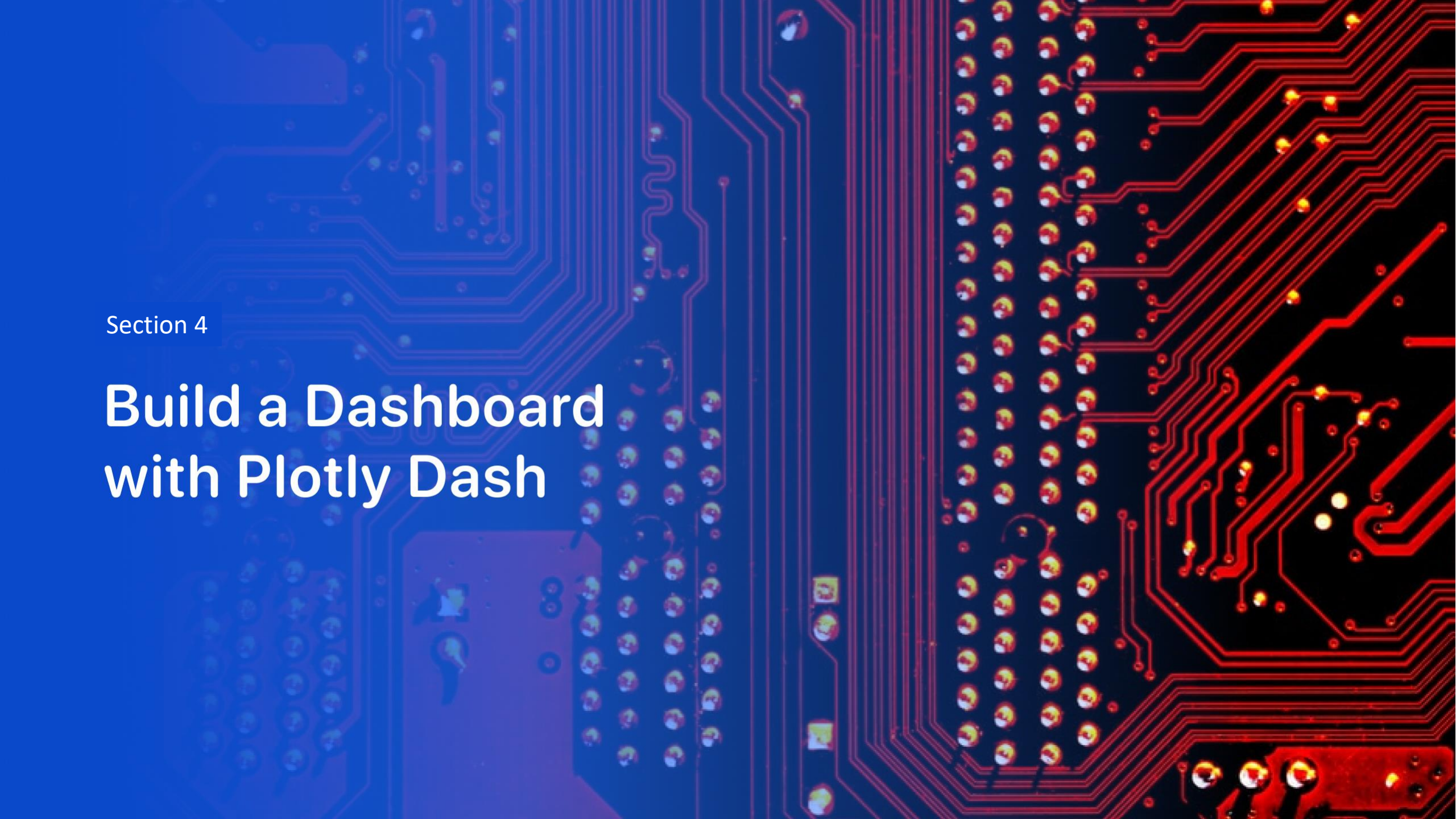
- Green colored labels show success precedes over failed landings.



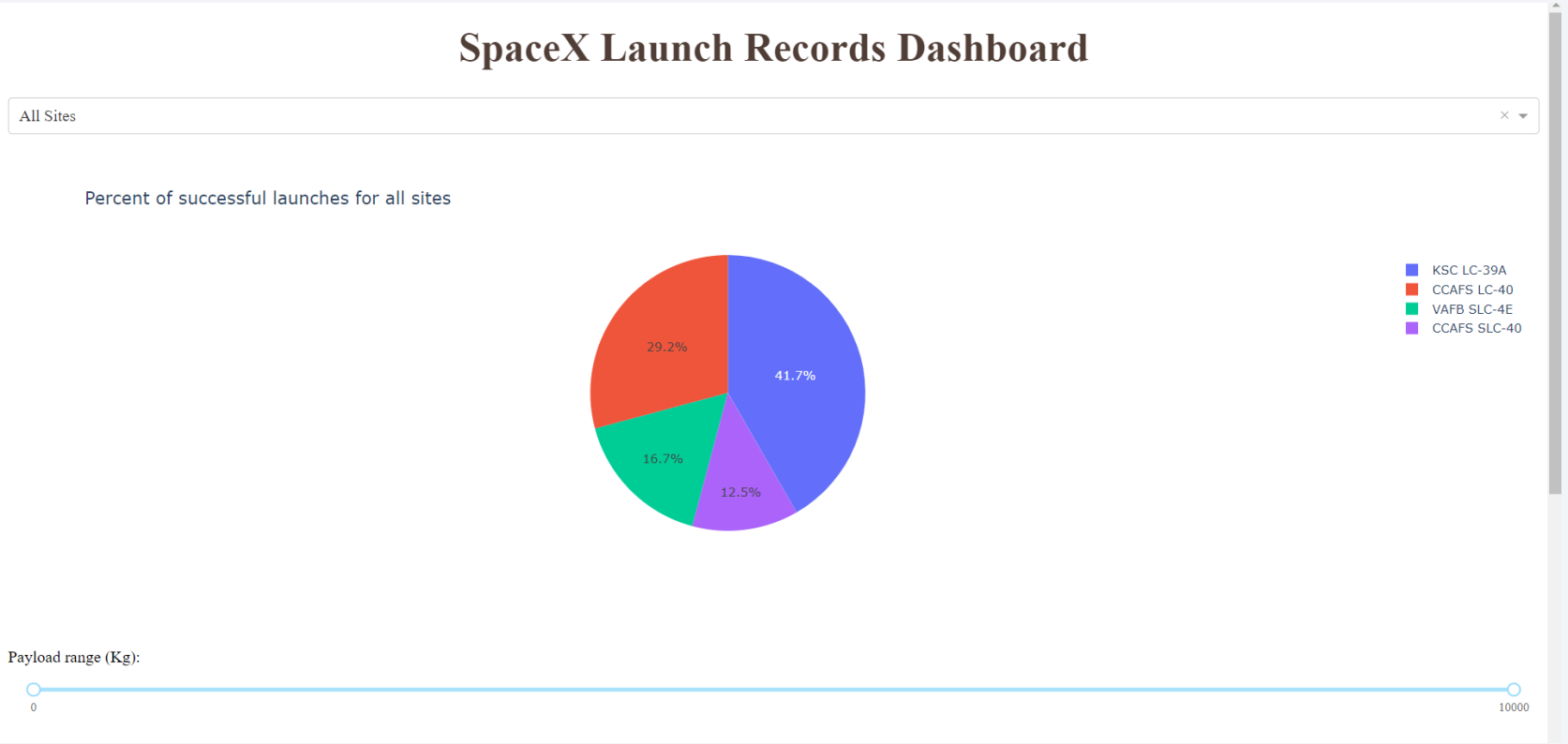
How connected are the launch sites?



- Here, launch site, KSC LC-39A is considered. It is well connected by highways and railways

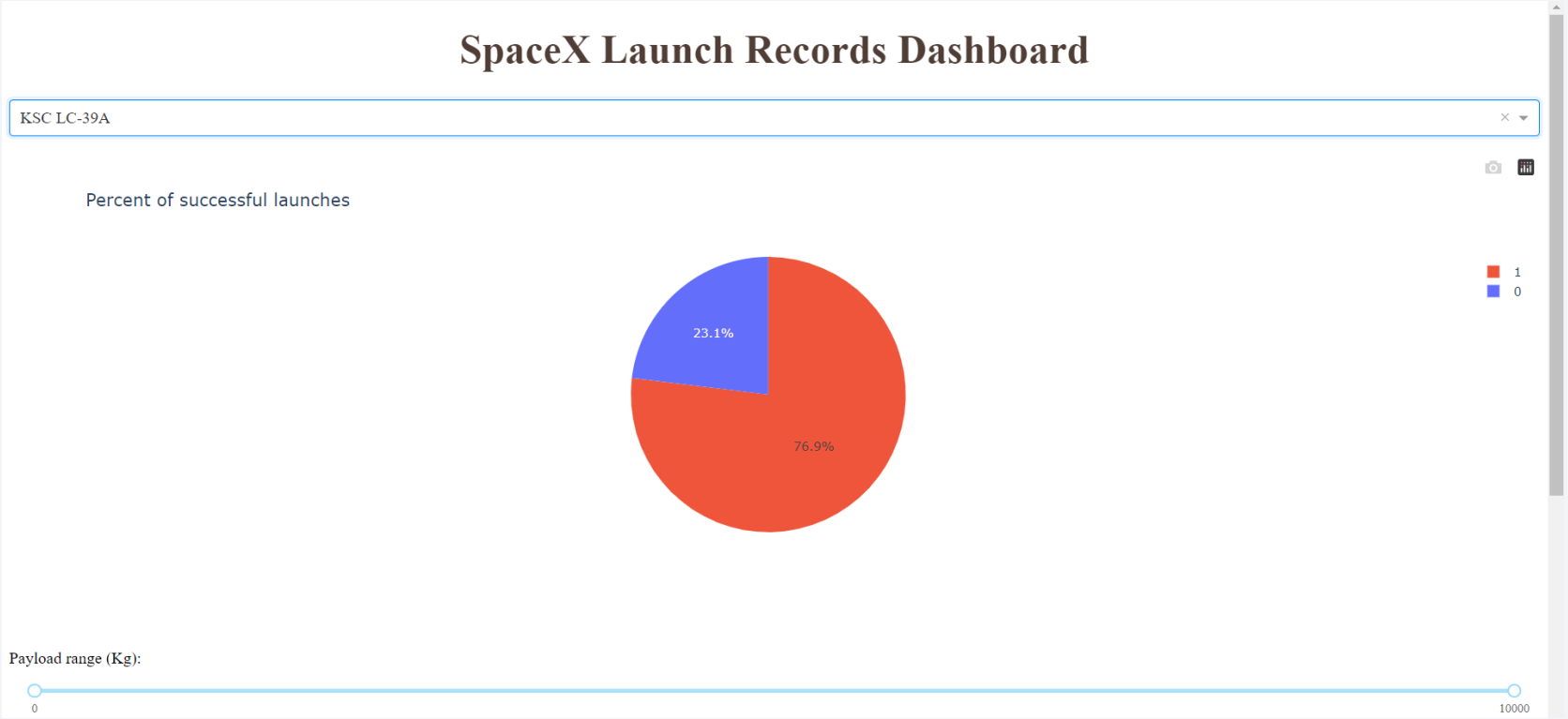

Section 4
Build a Dashboard with Plotly Dash

Count of successful launches by site



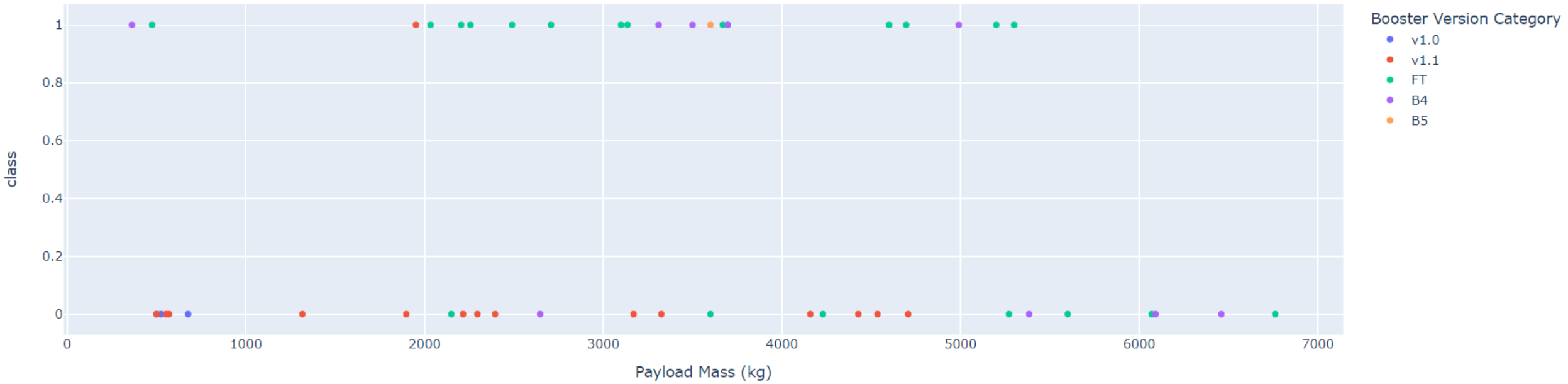
- KSC LC-39A has the maximum number of successful launches followed by CCAPS LC-40

Ratio of Successful Launches – KSC LC-39A



76.9% of the launches were successful from this site.

Payload vs. Launch Outcome scatter plot for all sites



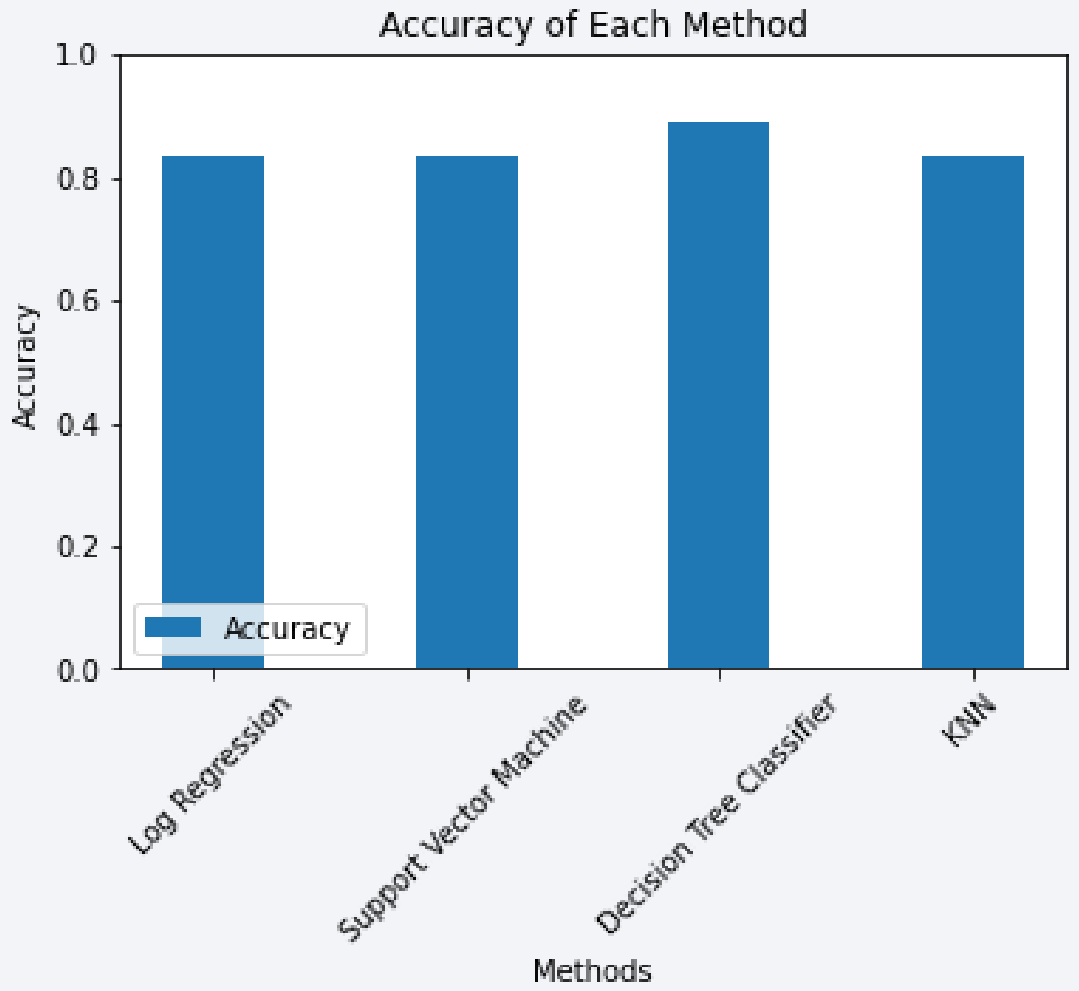
- FT boosters have a high success rate when the payload mass is less.
- v1.1 seems to have many failures at low payload masses

Section 5

Predictive Analysis (Classification)

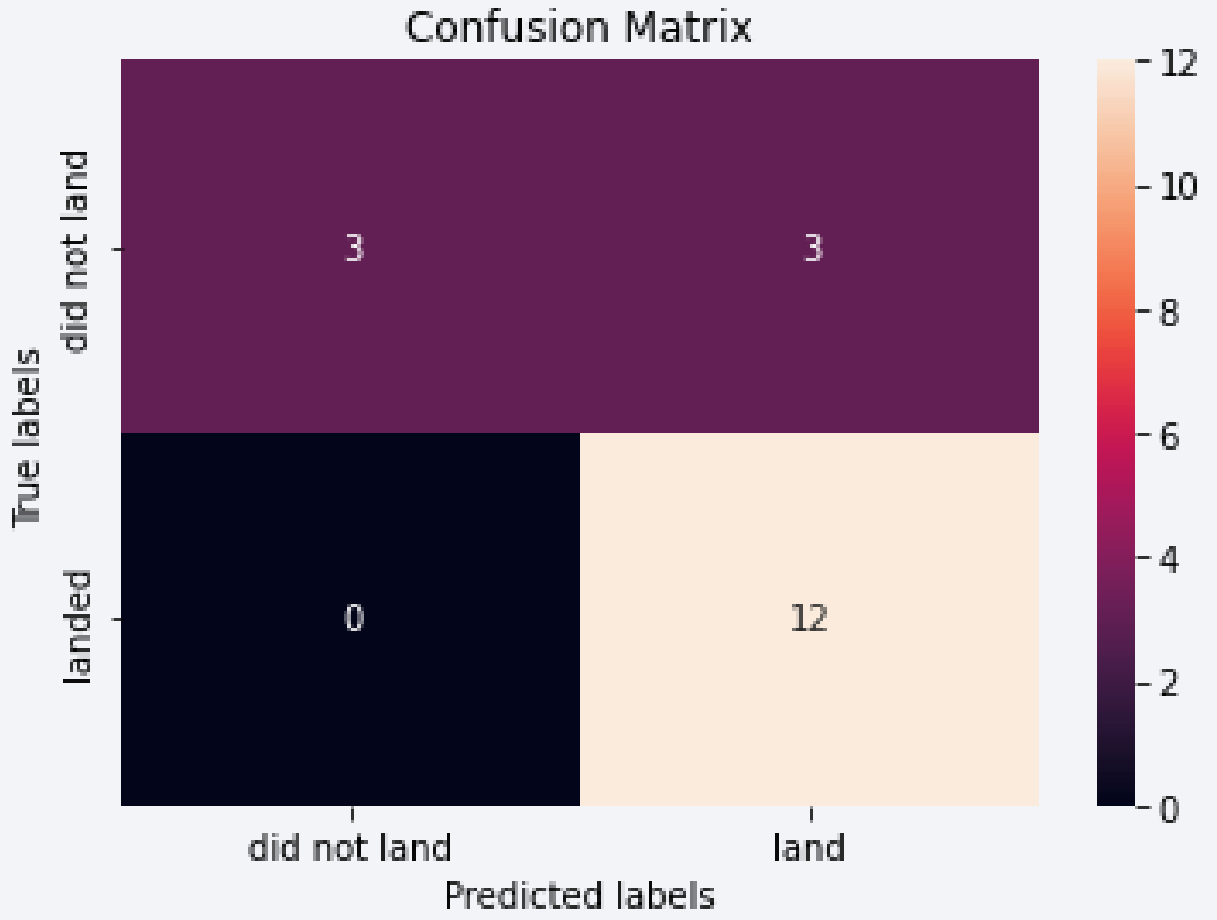
Classification Accuracy

- Decision Tree Classifier has the highest Accuracy.



Confusion Matrix

- As expected, it has a large number of true positives (12), which makes the model good.



Conclusions

- KSC LC-39A seems to be the best launch site when the payload mass is around mid values.
- Polar, LEO, and ISS has higher success rates with heavy payloads.
- GTO and ISS have more number of launches that others.
- Most successful launches happened with Flight numbers from 60 and above with the orbit VLEO
- The trend of successful outcomes has an increasing trend, neglecting the dip in 2018.
- Decision Tree Classifier has the highest accuracy among all the classification models.

Appendix

- I worked on my local Jupyter Notebook instead of the Watson Studio mostly.

Thank you!