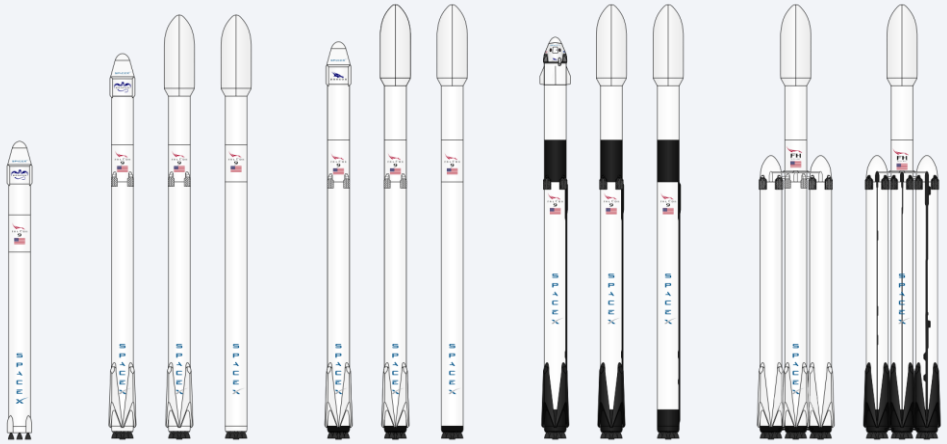
IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Chng Lee Chiat
2021-09-12

Outline

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



Executive Summary

- Summary of methodologies:

  Data collection was done by using get request to SpaceX API and web scraping of Falcon 9 launch information on Wikipedia page. Data wrangling was performed on collected data for further analysis. Exploratory data analysis, visualization using seaborn and matplotlib libraries and SQL were done to understand the data. Interactive maps and dashboard were generated to understand the correlation of each factors to the success rate of the mission. Finally, four machine learning models were used to predict data outcome and the best performing model was identified.

- Summary of results:

  It was found that the decision tree model provides the highest accuracy for predicting the outcome of a launch. The success rate of an outcome seems to be influenced by the payload mass, booster type and orbit of a rocket launch. Most launch sites are located near the coast with some located near a railway for easy transport. The impact of launch site on success rate is inconclusive.

Introduction

- Project background
  - Launching rockets cost up to 165 million dollars each.
  - There can be a cost-saving of up to 100 million dollars if the first stage of a rocket is reused.
  - A rocket company needs a good estimation of launch cost in order to bid against other companies such as SpaceX.
- Objective
  - To determine the cost of launch by analyzing SpaceX data to predict if the first stage of a rocket launch will land successfully.

Successful landing

Failed landing

Section 1

Methodology

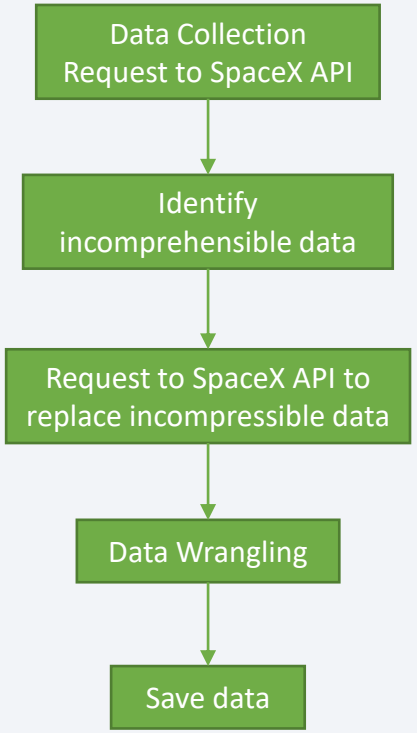
Methodology

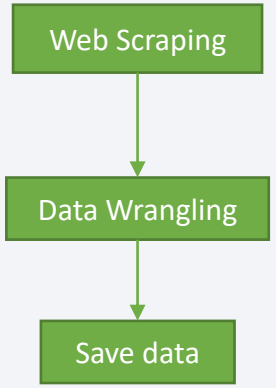
Executive Summary

- Data collection methodology: data were collected by get request of the SpaceX API and web scraping of Wikipedia page.
- Perform data wrangling: data were converted to Json, turned into Pandas dataframe, check for missing values and filling missing values.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using four classification models– logistic regression, support vector machine (SVM), decision tree and k-nearest neighbors (KNN). Collected data were transformed, split into training and test datasets, and each model was used to predict the landing outcome based on selected features. The accuracy of the models were compared to find the best performing model.

Data Collection

- Data sets for predicting launch outcome were collected from two sources – SpaceX API and Wikipedia page that contains Falcon 9 launches.
- The flowcharts on the right show the main process for data collections. The details of the process are shown in slides 8 ~ 11.



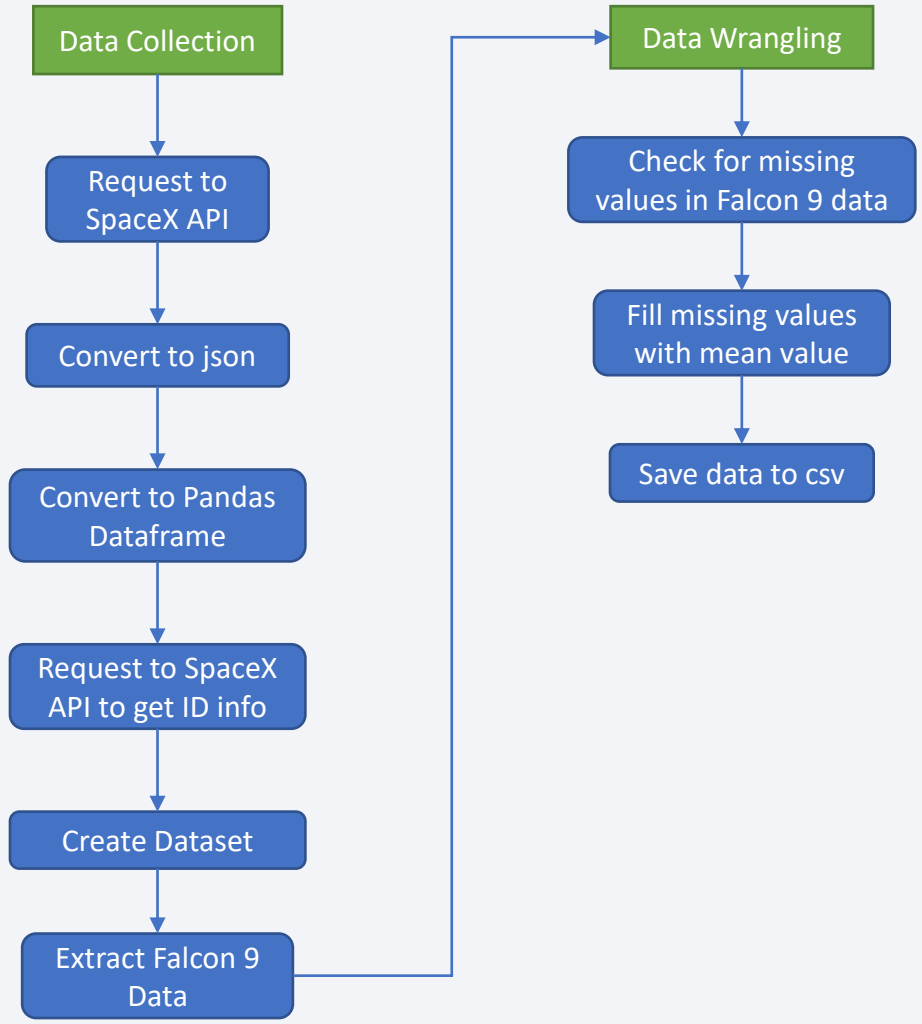


Data Collection – SpaceX API

- Data were collected and processed following the flowchart on the right.

- See code and output on GitHub: Data Collection

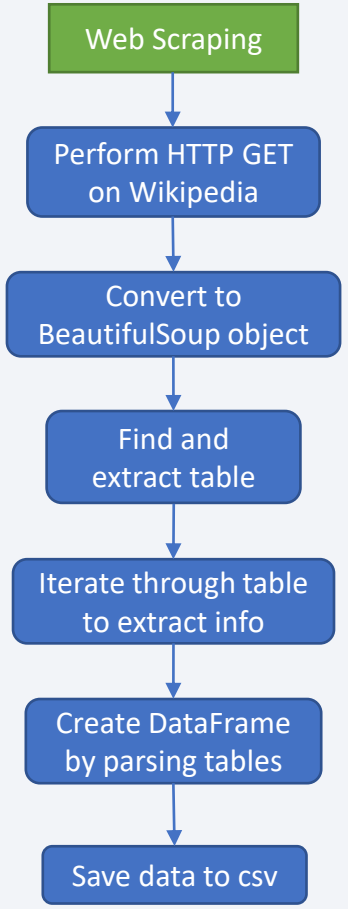
Please download PDF file to access the link.



Data Collection - Scraping

- Web scraping was completed following the flowchart on the right.

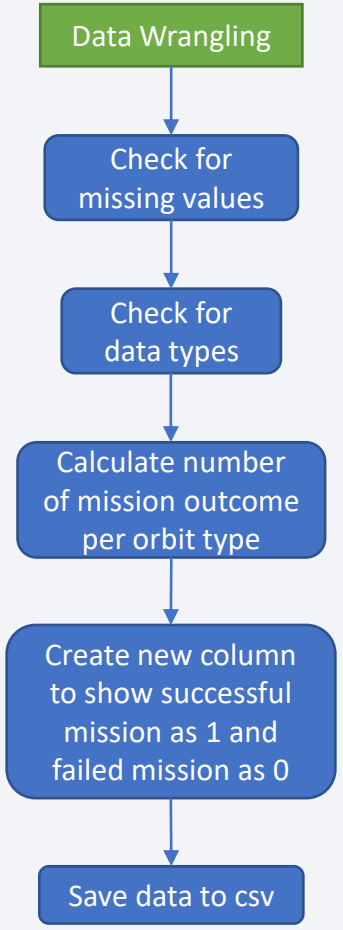
- See code and output on GitHub: [Data Collection - Scraping]



Data Wrangling

- Data wrangling was completed following the flowchart on the right.

- See code and output on GitHub: Data Wrangling



EDA with Data Visualization

- In the EDA section, flight number was plotted against launch site and orbit type in a scatter plot format to see the number of launches done at each launch site and to each type of orbit and their corresponding success rates.
- Payload mass was plotted against launch site and orbit type to observe the influence of payload mass with launch site and orbit type to success rates.
- A bar chart was plotted to show the success rate for rockets launched to each type of orbit was.
- A line chart was plotted to show the yearly trend of success rate.
- The charts were plotted to show how payload mass, launch site, orbit type affect success rate and the trend of success rate.
- See code and output at GitHub: EDA with Data Visualization

EDA with SQL

- Data was analyzed using SQL queries to find the unique launch sites of the space mission and first 5 of the launch sites that contained the string ‘CCA’.
- The sum and average of the payload mass were calculated.
- The first date of success was queried, and the total number of successful and failed missions were calculated.
- The type of booster with payload mass between 4000 and 6000 kg, and those that carried maximum payload mass (15,600 kg) were queried.
- The type of booster and the launch site were found for failed drone ship landing.
- Each type of landing outcome was counted and listed in a table.
- See code and output at GitHub: EDA with SQL

Build an Interactive Map with Folium

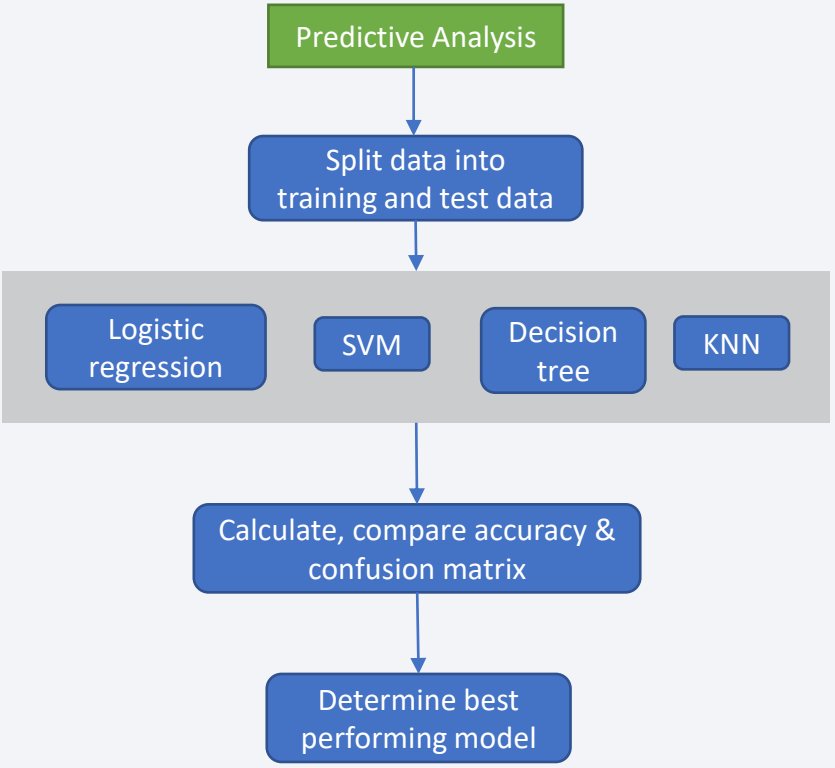
- Names of the launch sites and the respective number of launch missions were added to a global map in orange to show the locations of the sites.
- Launch markers of each record were added to show the outcome of the launch, in which green showed a successful launch and red a failed launch.
- The distance of two launch sites to a nearby railway was calculated and labeled on the map together with a straight line connecting two points to show the proximity of the launch sites to the railway.
- Interaction Map was plotted to show the location of the launch sites to understand why a location was chosen and how it affects the success rate.
- See code and output at GitHub: Interactive Map with Folium

Build a Dashboard with Plotly Dash

- A dashboard built using Plotly Dash shows a pie chart of success rate for all four launch sites and the success : failure ratio when a specific site is selected.
- The dashboards includes a slide bar where users can select payload mass range. The corresponding success and failure cases are shown in a scatter plot with booster versions color-coded.
- This dashboard configuration allows user to identify how launch site, booster version and payload mass affect success rate.
- See code at GitHub: Dashboard with Plotly Dash

Predictive Analysis (Classification)

- Four classification methods– logistic regression, support vector machines (SVM), decision tree and k-nearest neighbors (KNN)– were used to predict the landing outcome.
- The accuracy of the models were compared to find the best performing model.
- See code and output at GitHub: [Predictive Analysis]

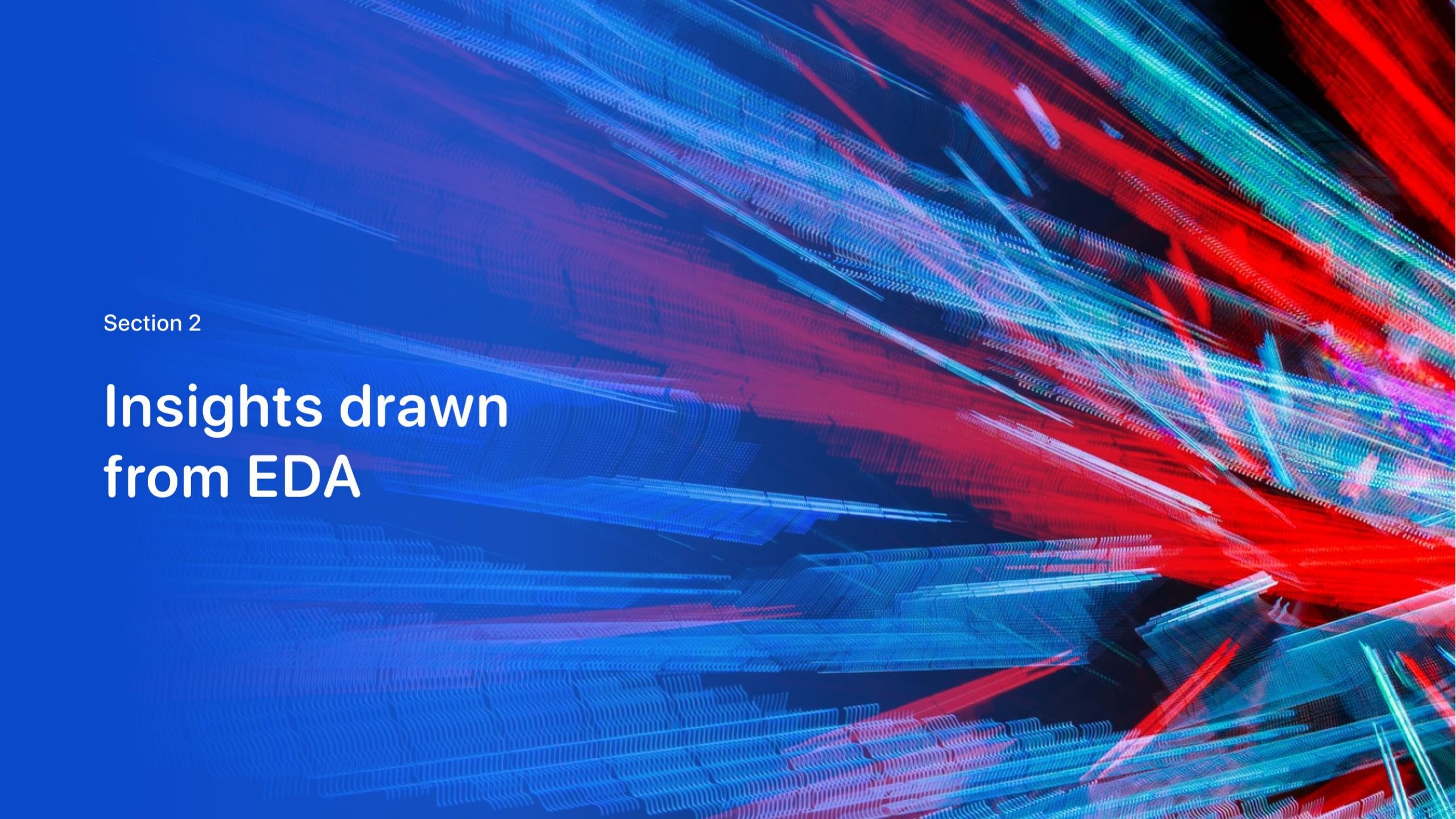


Results

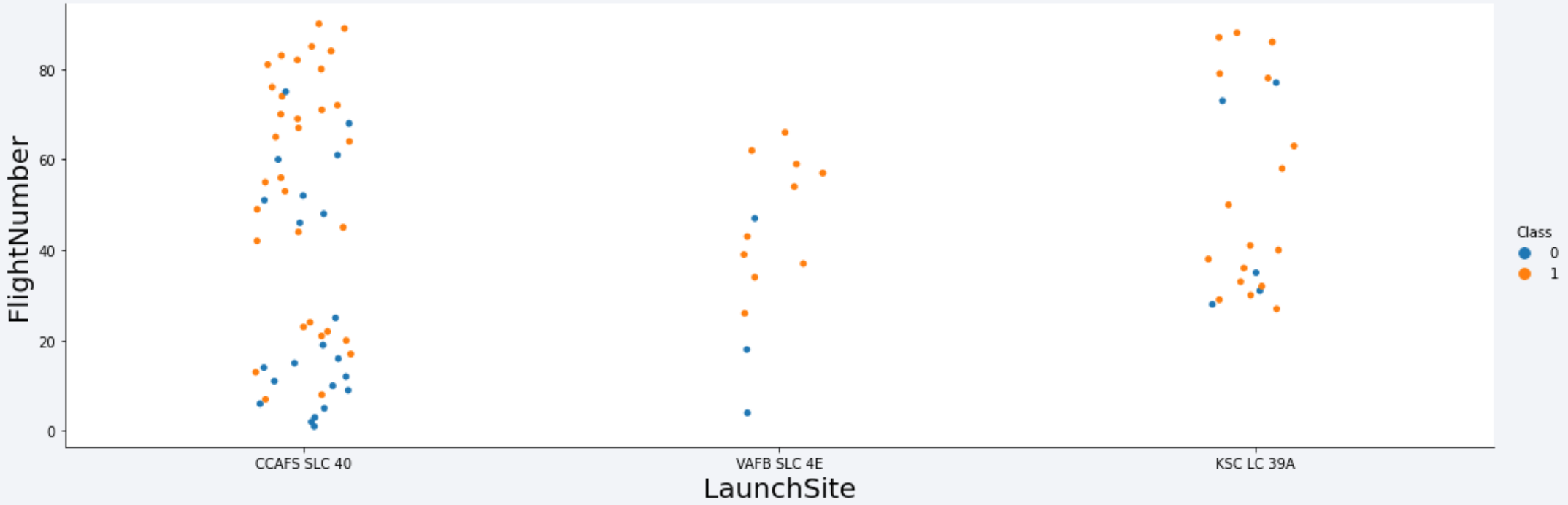
- Exploratory data analysis results:
  - Success rate of landing outcome increases over the years since 2013. There are higher likelihood of success for high flight number (recent launch) compared to low flight number.
  - Payload mass seems to have an influence on success rate depending on the type of orbit the rockets launch to.
- Interactive analytics:
  - Location of launch site, payload mass and booster version seem to have an influence on success rate of landing. Low to medium payload mass and booster version FT have the highest success rate.
- Predictive analysis results:
  - Out of four classification models, decision tree model has the best accuracy.

Section 2

Insights drawn from EDA

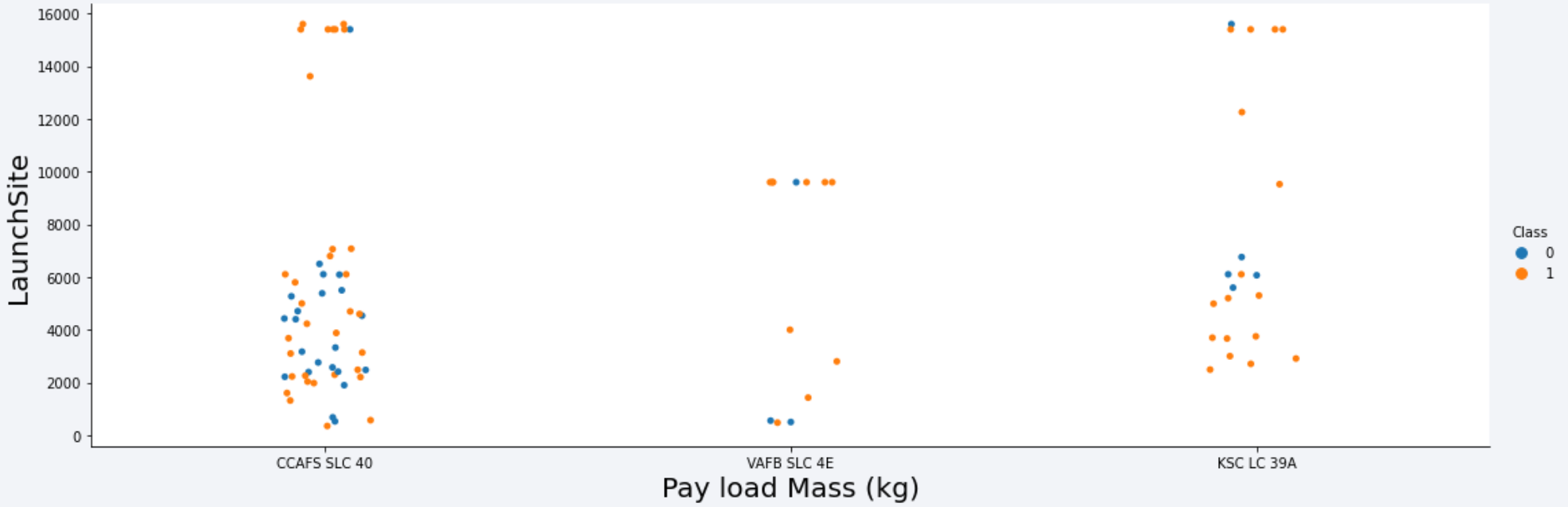


Flight Number vs. Launch Site



- The plot above shows a scatter plot of Flight Number vs. Launch Site.
- Most launches were at CCAFS SLC 40, followed by KSC LC 39A and VAFB SLC 4E.
- Success rate seems to increase with higher FlightNumber.

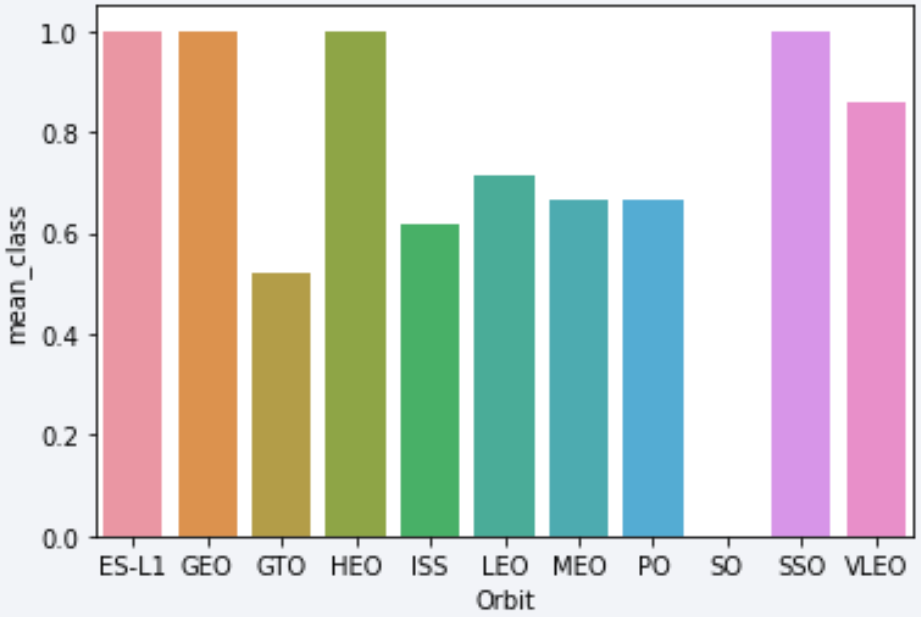
Payload vs. Launch Site



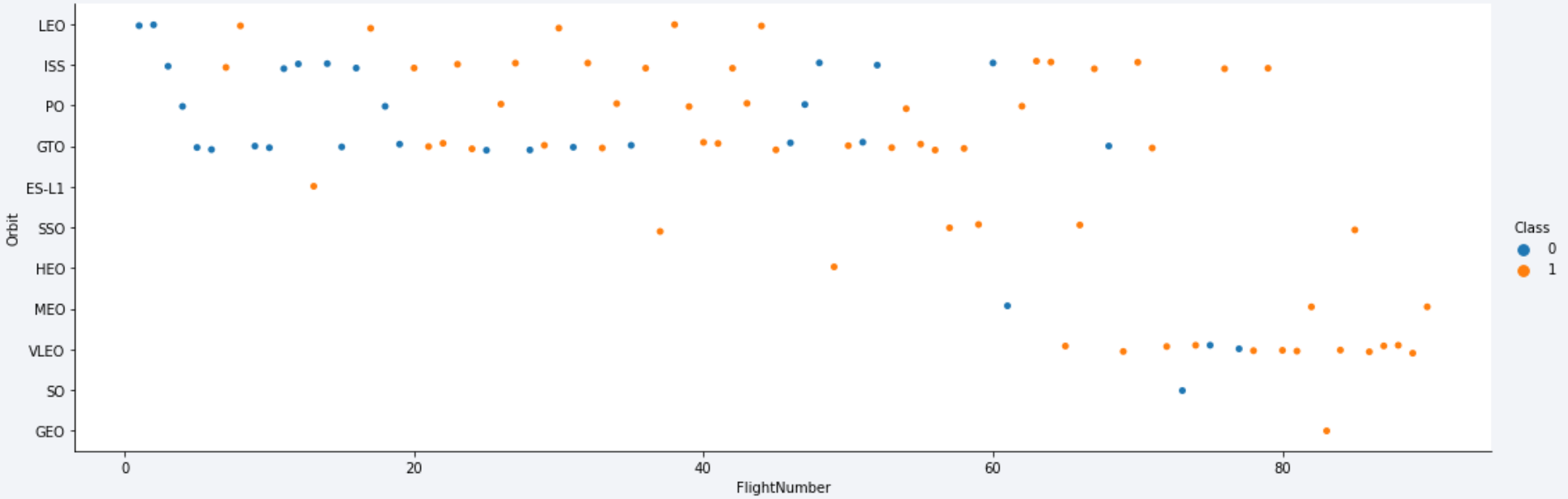
- The plot above shows a scatter plot of Payload Mass vs. Launch Site.
- High payload mass launches at CCAFS SLC 40 seems to have a higher success rate, medium payload mass has a higher success rate at VAFB SLC 4E, and KSC LC 39A launch site has a higher success rate for low payload mass.

Success Rate vs. Orbit Type

- The bar chart on the right shows a comparison of success rate for each orbit type.
- SSO has the highest consistent success rate with 5 launches and 100% success. VLEO has a rather high success rate at above 80% with 14 launches.
- The launches to GTO, ISS, LEO, MEO, and PO were moderately successful (50~70%).
- There was only 1 launch to SO orbit and failed. There was 1 launch each to ES-L1, GEO and HEO, and all were successful.

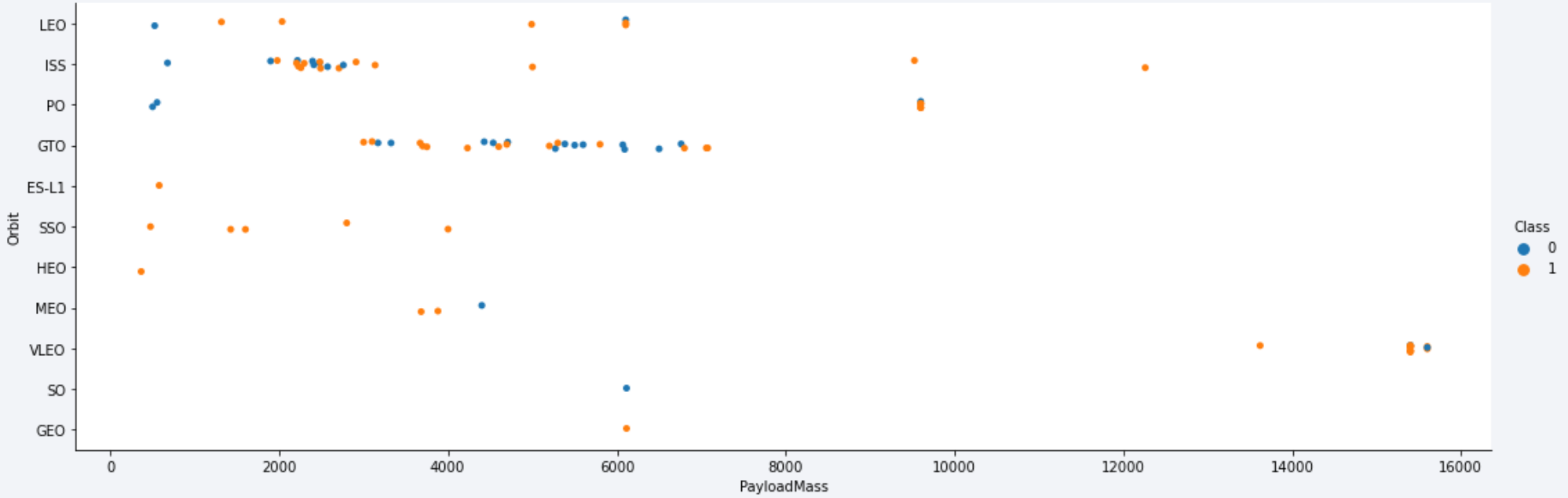


Flight Number vs. Orbit Type



- The plot above shows a scatter plot of FlightNumber vs. Orbit.
- Similar to previous observation, the higher the FlightNumber, the higher the success rate.
- The success rate for LEO, ISS, GTO, etc. may seem lower because most of the early launches were sent to these orbits. The actual success rate for these orbits may be higher now.

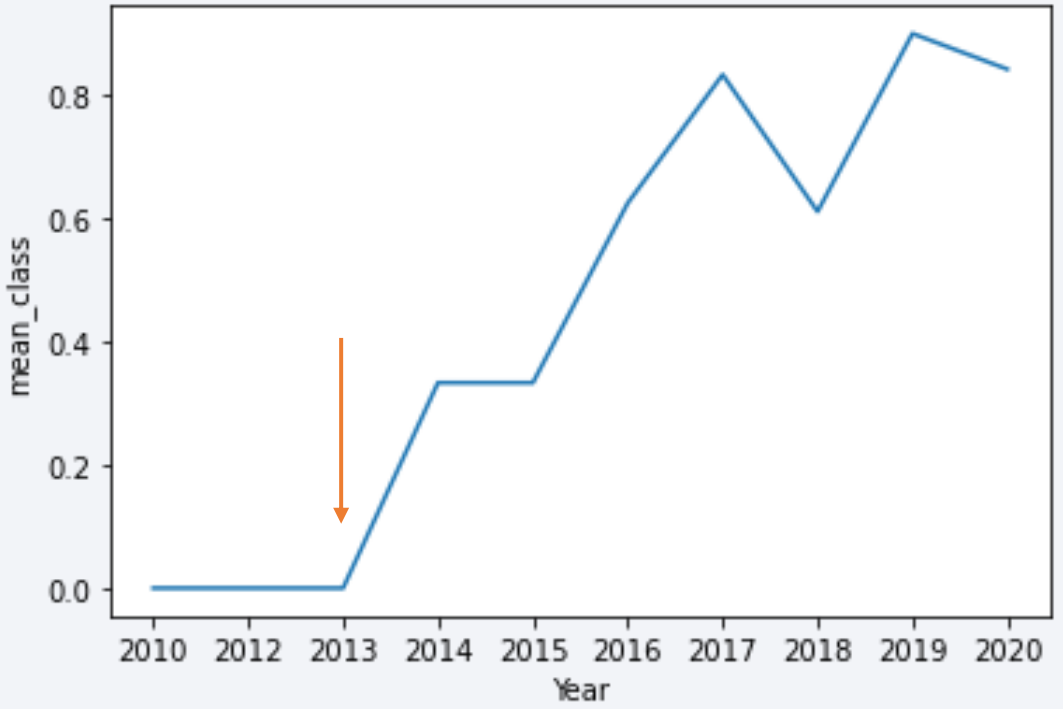
Payload vs. Orbit Type



- The plot above shows a scatter plot of PayloadMass vs. Orbit.
- Higher payload seems to have a negative influence on GTO orbits, and positive on ISS and LEO orbits.

Launch Success Yearly Trend

- Success rate keeps increasing since year 2013.



All Launch Site Names

- There are four unique launch sites:

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

Launch Site Names Begin with 'CCA'

- The first five records where launch sites begin with `CCA` are shown in the following table.
- They all happen to be the same launch site – CCAFS LC40.

launch_site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

- The total payload carried by boosters from NASA is 619,967 kg or around 620 ton.

sum_payload_mass
619967

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2,928.4 kg, which is quite low considering the maximum payload mass is 15,600 kg.

avg_payload_mass
2928.40

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad was on 2010-06-04.

first_success_date
2010-06-04

Successful Drone Ship Landing with Payload between 4000 and 6000

- There were 4 boosters that have successfully landed on drone ship and had payload mass greater than 4000 but less than 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

- There were a total of 100 successful and 1 failed missions.

success_mission	failed_mission
100	1

Boosters Carried Maximum Payload

- The following boosters have carried the maximum payload mass (15,600 kg):

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

- There were 5 failed landing_outcomes in drone ship in year 2015:

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1017	VAFB SLC-4E
Failure (drone ship)	F9 FT B1020	CCAFS LC-40
Failure (drone ship)	F9 FT B1024	CCAFS LC-40

- Most of the failed drone ship landing in 2015 launched at CCAFS LC-40 launch site.

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

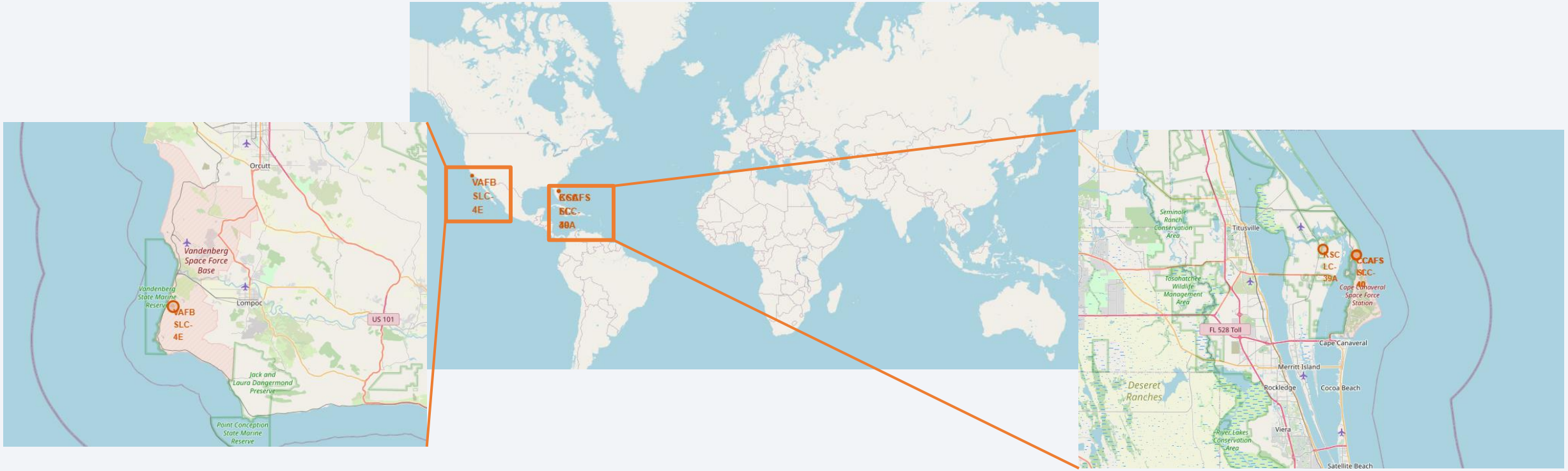
- The table below shows the counts of types of landing between the date 2010-06-04 and 2017-03-20:

landing__outcome	cnt
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

Section 4

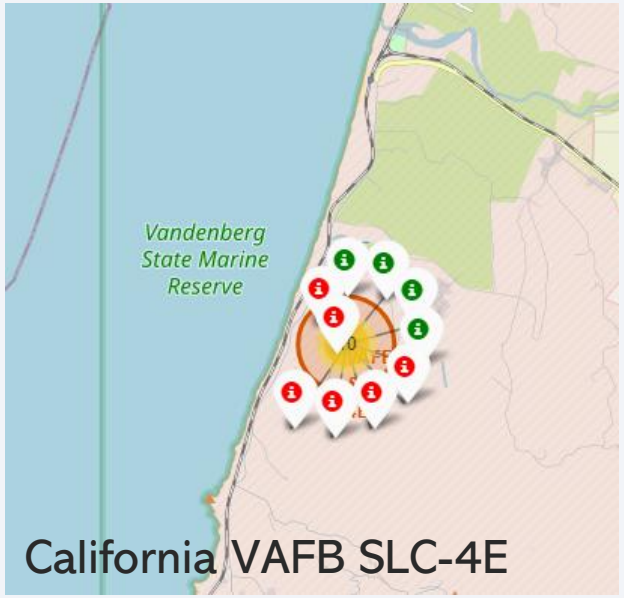
Launch Sites Proximities Analysis

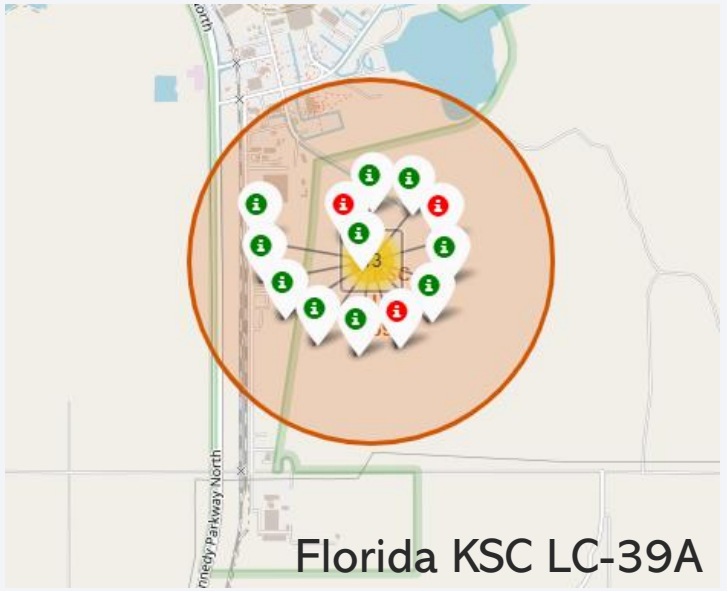
Location of SpaceX Launch Sites

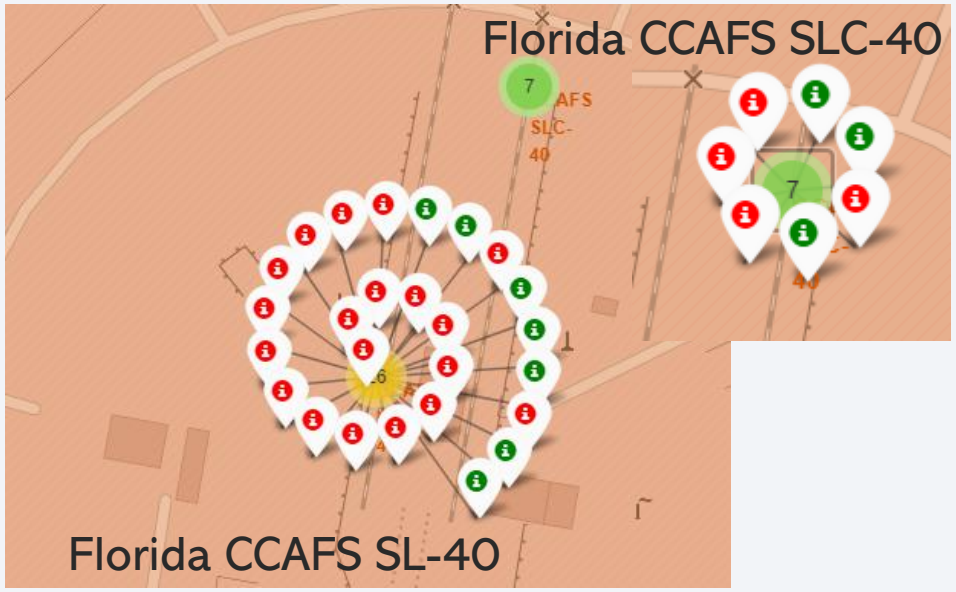


- SpaceX launch sites are all located in the United States and by the coast – one in west coast, California (VAFB SLC-4E) and three in east coasts, Florida (CCAFS LC-40, CCAFS SLC-40 and KSC LC-39A).

Outcome of Mission at Each Launch Site



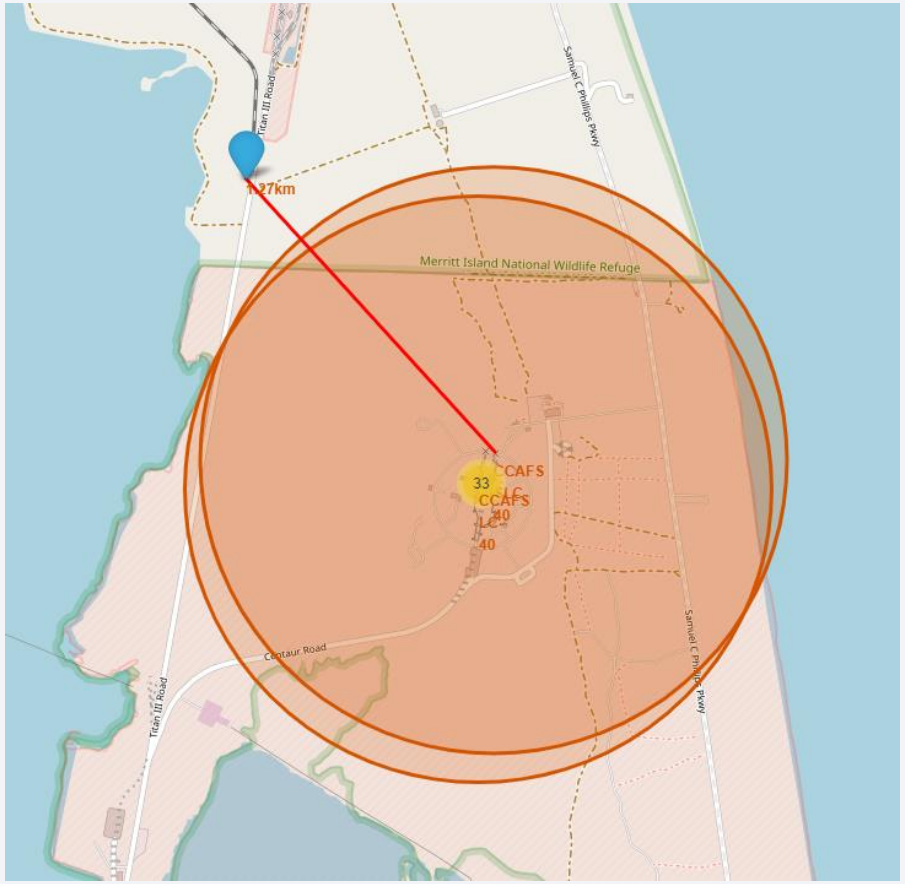




- The launch site at Florida KSC LC-39A produced the best outcome with the highest ratio of success, followed by Florida CCAFS SLC-40, California VAFB SLC-4E.
- The launch site at Florida CCAFS SL-40 had the highest number of launches, but lowest ratio of success.

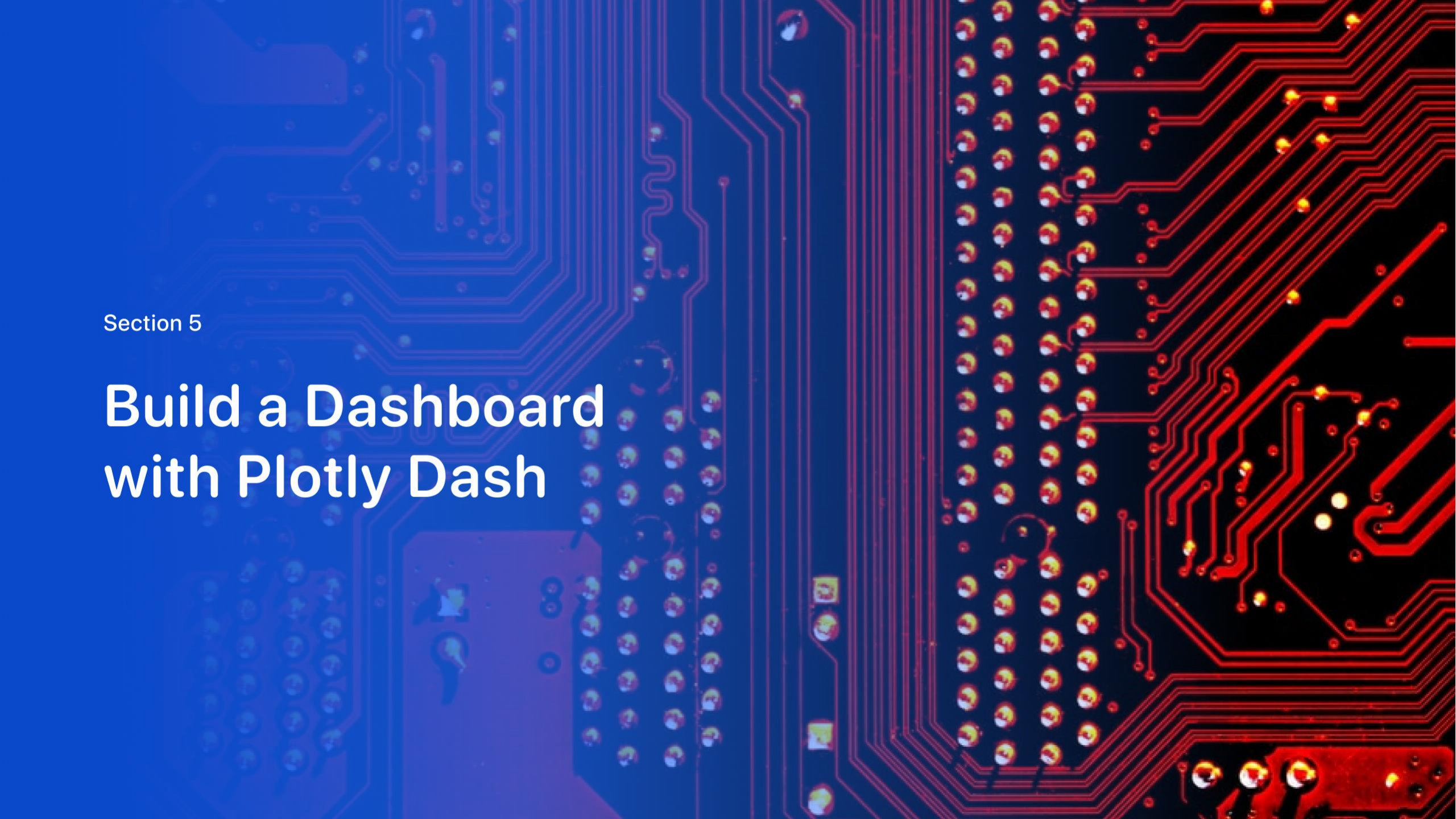
Proximity of Launch Site to Railway

- The two launch sites at Florida CCAFS SLC-40 and CCAFS SL-40 are in very close proximity to a railway at 1.27 km.
- Proximity to railway helps to transport rockets and supplies to the launch sites.

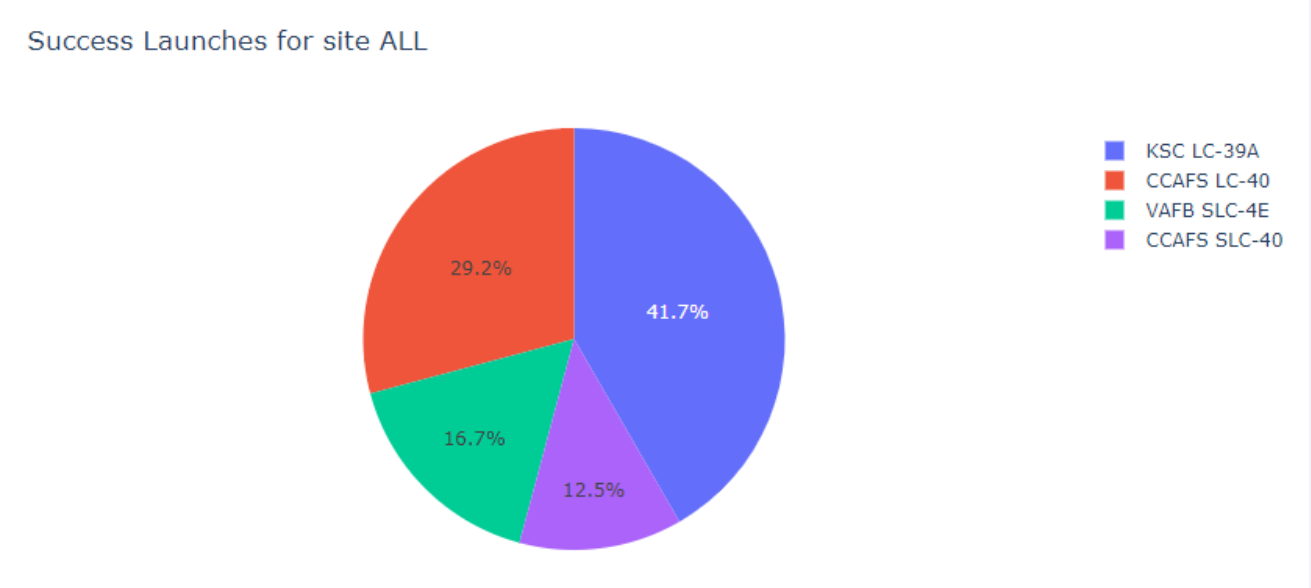


Section 5

Build a Dashboard with Plotly Dash

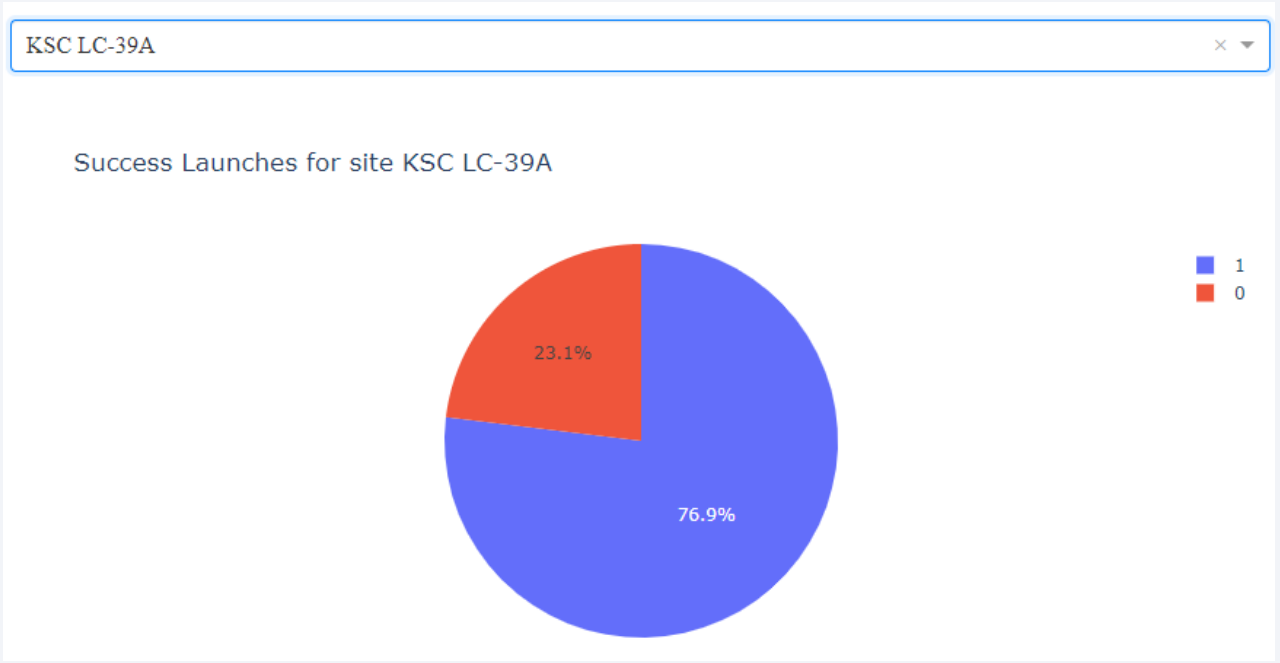


SpaceX Launch Success Count for All Sites



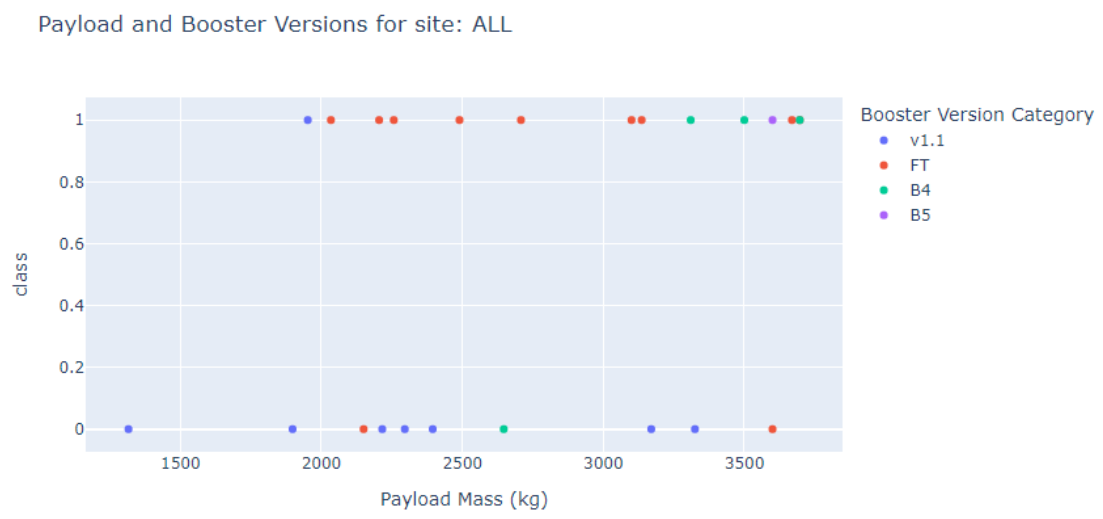
- The launch site success counts are consistent with the observation from the Folium map in the previous section.
  - ➢KSC LC-39A has the highest success count (highest success ratio) followed by CCAFS LC-40 (lowest success ratio), VAFB SLC-4E and lastly CCAFS SLC-40.

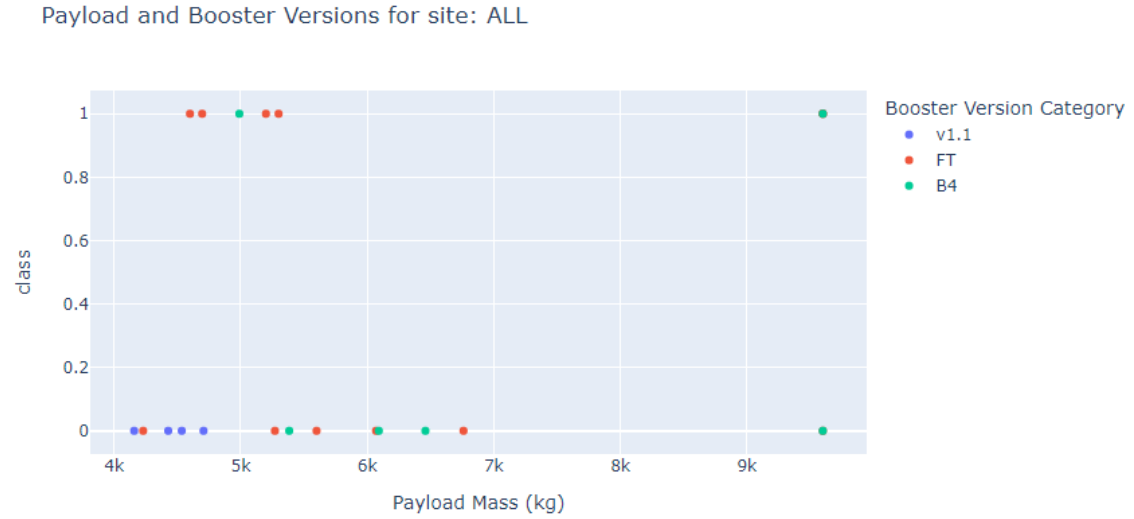
Launch Site with the Highest Success Ratio



- The pie chart above shows the success to failure ratio for the launch site – KSC LC-39A.
- This launch site has the highest success ratio at 77:23 among four launch sites.

Influence of Payload to Success Rate





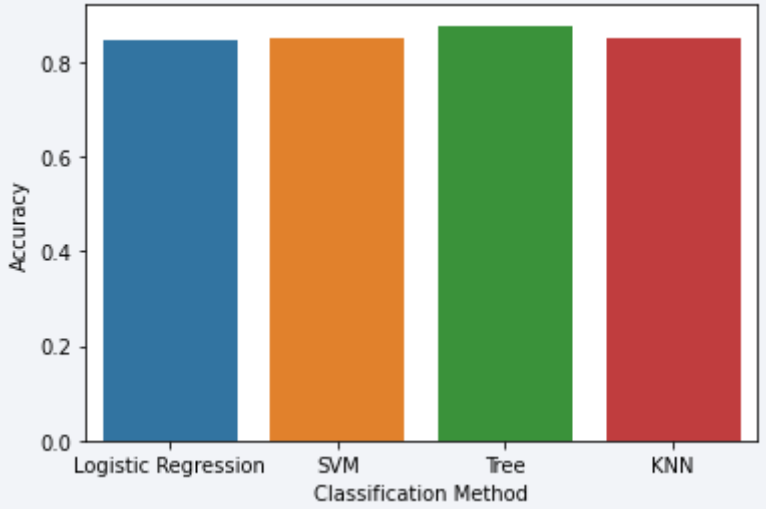
- High payload (4,000 ~ 10,000 kg) seems to have a negative influence on the success rate compared to low payload (0 ~ 4,000 kg).
- The number of failed outcome increases, and successful outcome decreases as payload mass increases.
- The booster version – FT seems to have the highest success rate.

Section 6

Predictive Analysis (Classification)

Classification Accuracy

- The bar chart on the right shows the accuracy of four classification methods:
  - Logistic regression
  - Support vector machines
  - Decision tree
  - k-nearest neighbors
- All four classification methods have comparable accuracy and decision tree has slightly higher accuracy.



Confusion Matrix

- The confusion matrix of decision tree model on the right shows that there are no false negative and three false positive.
- All four classification methods (logistic regression, SVM, decision tree, KNN) resulted in a same confusion matrix.

Conclusions

- Success rate of landing for SpaceX has been increasing since 2013.
- Launch site at Florida KSC LC-39A produced the best outcome with the highest ratio and highest count of successful landing. However, it is inconclusive whether launch site has direct influence on success rate, because some launch site was used more during early launches.
- Launch sites are located within proximity to the coast and railway.
- Payload mass, booster type and orbit type seem to influence the success rate. Low to medium payload mass and FT booster type have the highest success rate for most orbit type. However, payload mass may have a reversed impact on certain orbit type such as high payload has a positive impact on ISS and LEO orbit types.
- Decision tree is the best performing model for predicting the outcome of a launch with 88% accuracy.

Appendix

- Links to files on GitHub:
  - Data Collection API Lab
  - Data Collection with Web Scraping
  - EDA Lab
  - EDA with Data Visualization
  - EDA with SQL
  - Dashboard with Plotly Dash
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
  - Report

Please download PDF file to access the links.

Thank you!