IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Theodore Psillos

4th March 2024

Outline

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

Executive Summary

- **Data collection methodology:**
  - Data collection using web scraping and SpaceX APIs;
  - Exploratory Data Analysis (EDA), including data wrangling, data visualization, SQL and interactive visual analytics using Folium & Plotly Dash;
  - Implementation of predictive analysis using machine learning classification models
- **Summary of results:**
  - Data was successfully collected from public sources
  - The EDA performed identified features that are best in predicting the success of launches
  - Machine learning Prediction showed the best model to predict which characteristics are important to drive this opportunity by the best way, using all collected data.

Introduction

- The objective is to evaluate the viability of the new company Space Y to compete with with Space X
- Desirable results:
  - The most appropriate way to estimate the cost of launches, which was determined by predicting successful landings of the first stage of rockets;
  - Where the best location is to ensure successful launches

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data from SpaceX was obtained from 2 sources:
    - The SpaceX API - https://api.spacexdata.com/v4/rockets/
    - WebScraping Wikipedia – https://en.Wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches
- Perform Data Wrangling:
  - The collected data was wrangled by creating a creating a landing outcome based on outcome data after analysing features using summary metrics.
- Perform Exploratory Data Analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash

Methodology

Executive Summary

- Perform predictive analysis using classification models
  - The data prepared in the prior steps is split into training and test sets
  - Four classification models are then trained and evaluated using these data sets
  - The accuracy of each model is calculated and used to compare the efficacy of each model using different combinations of hyperparmaters.

Data Collection

- The 2 key data sources were collected from the following sites using Python API calls and web scraping libraries such as *requests* and *beautfiulsoup4*
  - The SpaceX API - https://api.spacexdata.com/v4/rockets/
  - WebScraping Wikipedia – https://en.Wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches

Data Collection – SpaceX API

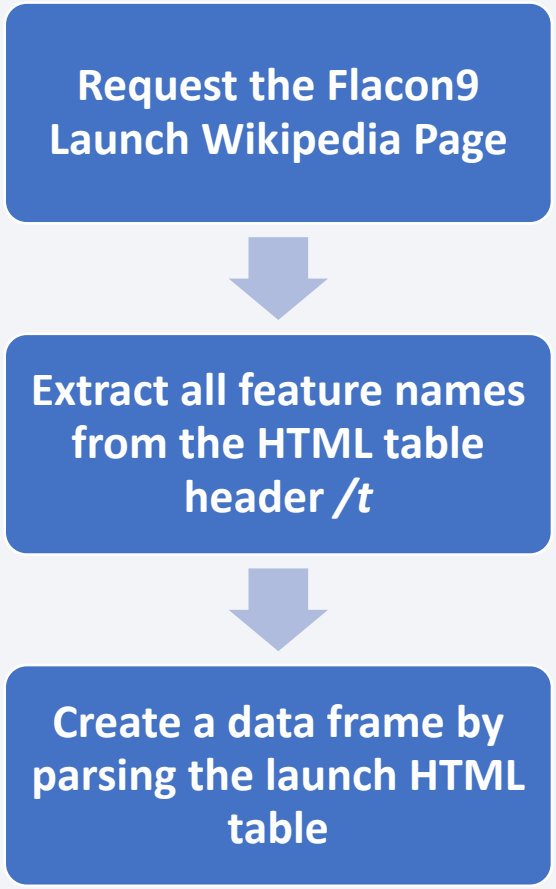
- SpaceX offers a public API from where data can be requested/obtained and then used.
- The API was used according to this flowchart shown on the right, with the requested data being stores in memory.

Source Code - https://github.com/theodorepsi/IBM-Datascience-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection – Scraping

- Additional data on the SpaceX launches can be found on Wikipedia.

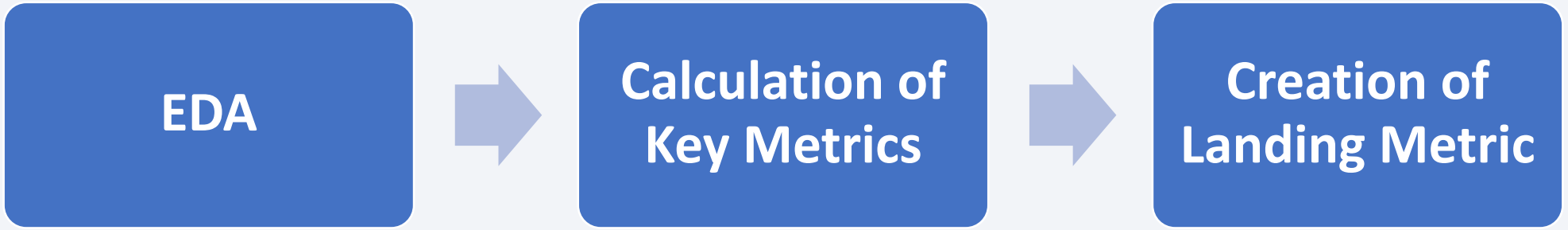
- The data was downloaded and scraped from the webpage according to the flowchart shown on the right, with the extracted data frame stored saved.



Source Code - https://github.com/theodorepsi/IBM-Datascience-Capstone/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

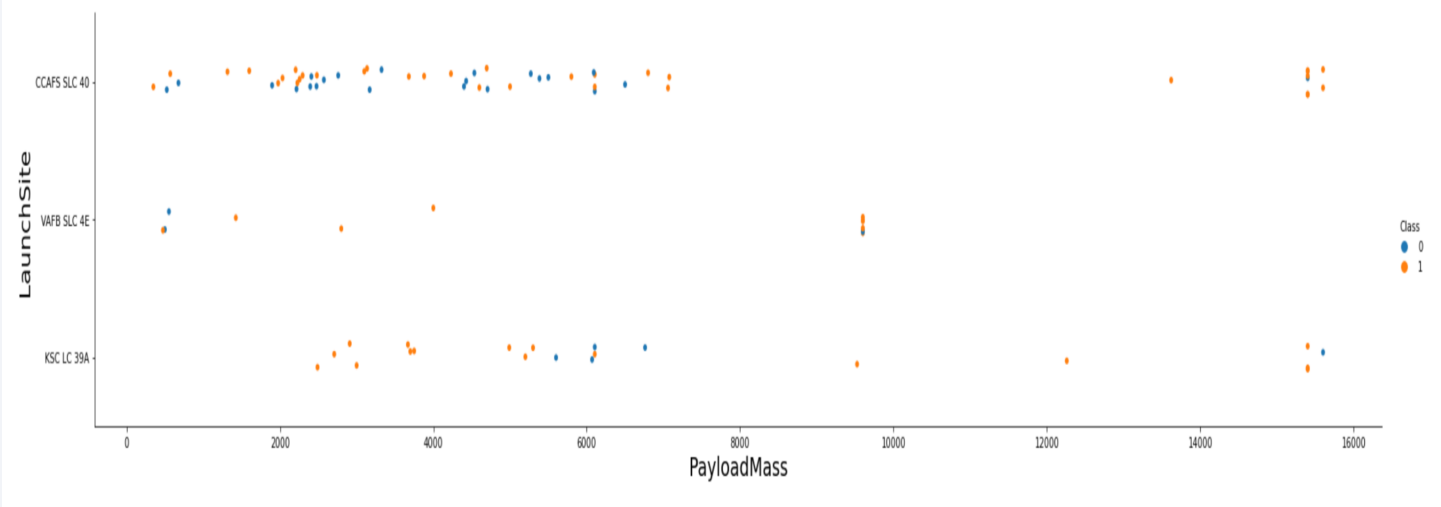
- Exploratory Data Analysis (EDA) was performed on the prior 2 data sets
- A summary of the data sets was obtained by calculating the following key metrics:
  - Launches per site,
  - Occurrences of each orbit and
  - Occurrences of mission outcome per orbit
- A landing outcome label was created using the *Outcome* column



Source Code - https://github.com/theodorepsi/IBM-Datascience-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- To explore data, barplots and scatter plots were used to visualize the relationship between a variety of combinations of feature pairs such as
  - Payload Mass & Flight Number
  - Launch Site & Flight Number
  - Launch Site & Payload Mass
  - Orbit & Flight Number
  - Payload & Orbit



Source Code - https://github.com/theodorepsi/IBM-Datascience-Capstone/blob/main/jupyter-labs-eda-with-data-visualization.ipynb

EDA with SQL

- The following SQL queries were performed:
  - Names of the unique launch sites in space missions
  - Top 5 launch sites whose names began with '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 successful landing outcome in ground pad was achieved
  - Names of boosters which have success in drone ship and have payload mass between 4000 and 6000kg
  - Total number of successful and failed mission outcomes
  - Names of the booster versions which have carried the maximum payload mass
  - Failed landing outcomes in drone ship, followed by their booster versions and launching site names for the year 2015
  - Rank of the count of landing outcomes (like Failure (drone ship) & Success (ground pad)) for the dates between 2010-06-04 and 2017-03-20

Source Code - https://github.com/theodorepsi/IBM-Datascience-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Markers, circles, lines and marker clusters from the Folium Python package were used to indicate information regarding selected launch sites on the Folium Map
  - The markers indicated points of interest such as the Launch Sites
  - Circles were used to highlight areas around specific coordinates such as the NASA Johnson Space centre
  - Marker clusters were then used to indicate groups of events at each coordinate such as the number of launches at a launch site
  - Lines were drawn between launch sites and key infrastructure to indicate the distance between 2 coordinates

Source Code - https://github.com/theodorepsi/IBM-Datascience-Capstone/blob/main/jupyter-lab-interactive-visual-analytics-with-folium.ipynb

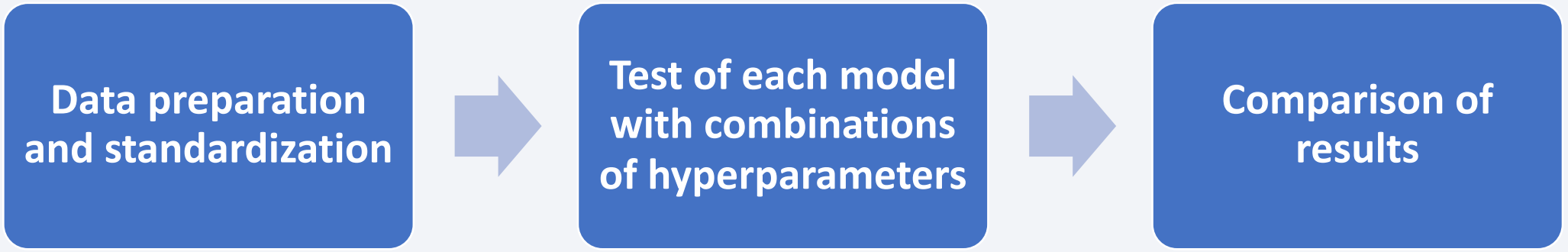
Build a Dashboard with Plotly Dash

- There were 2 key graphs that were shown in the Dashboard namely:
  - The percentage of launches per site
  - The payload range
- A combination of the above graphs and the ability to select a particular launch site allowed for a quick analysis of the relation between the payloads and launch sites.
- This analysis allows one to easily identify the best place to launch.

Source Code - https://github.com/theodorepsi/IBM-Datascience-Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- A total of 4 classification models were trained, tested and then compared:
  - Logistic Regression
  - Support Vector Machine (SVM)
  - Decision Tree
  - K Nearest Neighbours



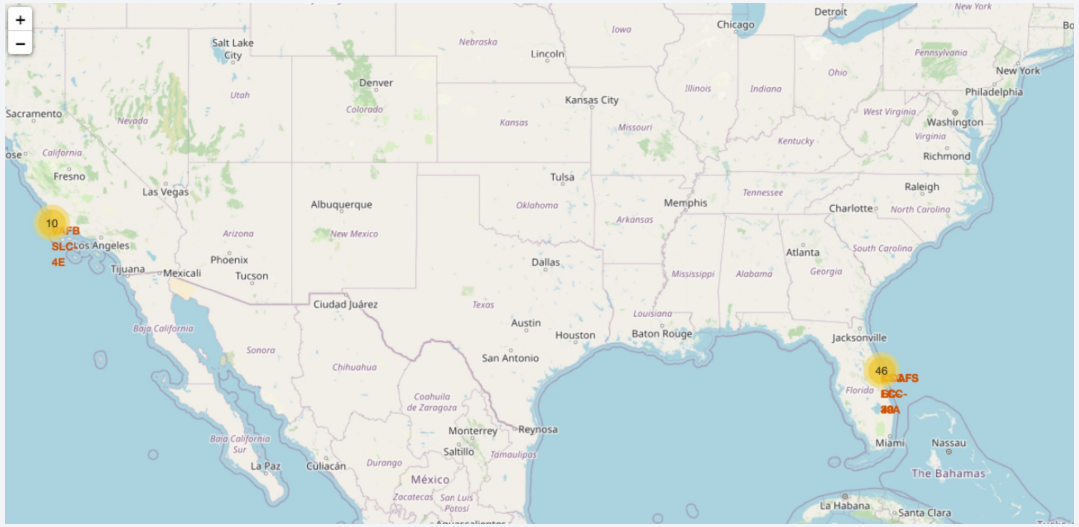
Source Code - https://github.com/theodorepsi/IBM-Datascience-Capstone/blob/main/jupyter-lab-machine-learning-prediction.ipynb

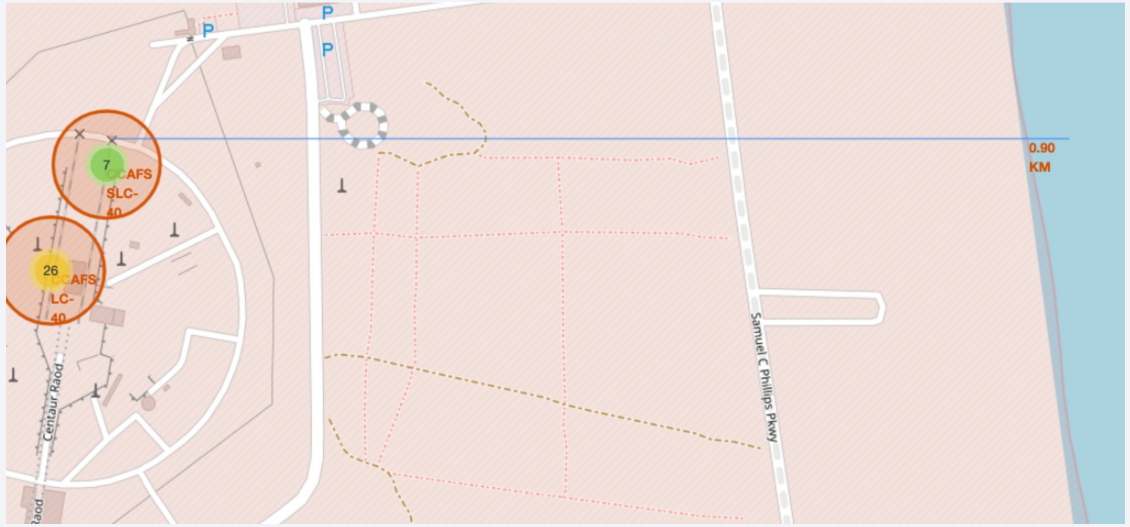
Results

- Exploratory data analysis results
  - SpaceX uses 4 different launch sites
  - The average payload of F9 v1.1 booster was 2,928kg
  - The first successful landing outcome occurred in 2015, which was 5 years after the first launch date
  - Numerous Falcon 9 booster versions were successful in landing on drone ships and had a payload mass greater than the average payload mass mentioned above.
  - Almost 100% of the mission outcomes were successful
  - The two booster versions that failed at landing drone ships in 2015 were:
    - F9 v1.1 B1012
    - F9 v1.1 B1015
  - The number of successful landing outcomes improved with each year

Results

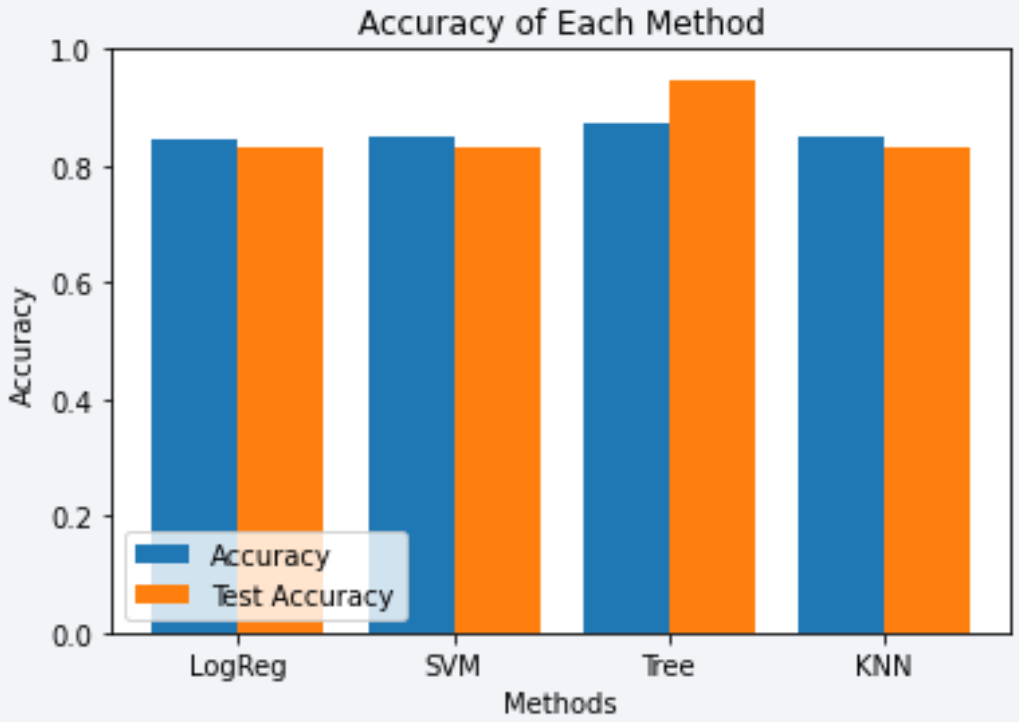
- Using the interactive visual analytics it was possible to identify that launch sites were based in safe locations, near the sea and having access to good infrastructure
- The majority of launches occurred on the east coast of North America

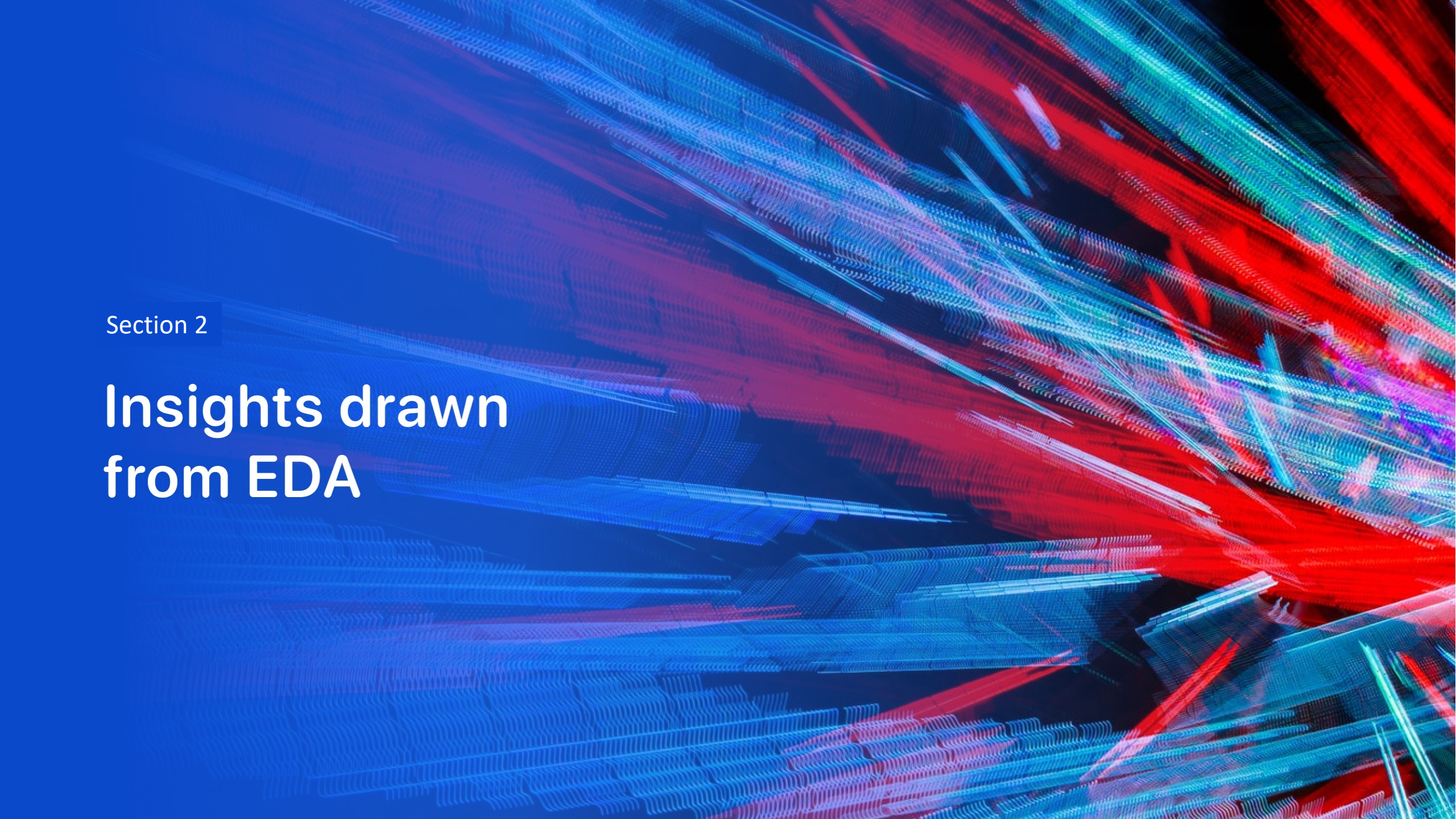




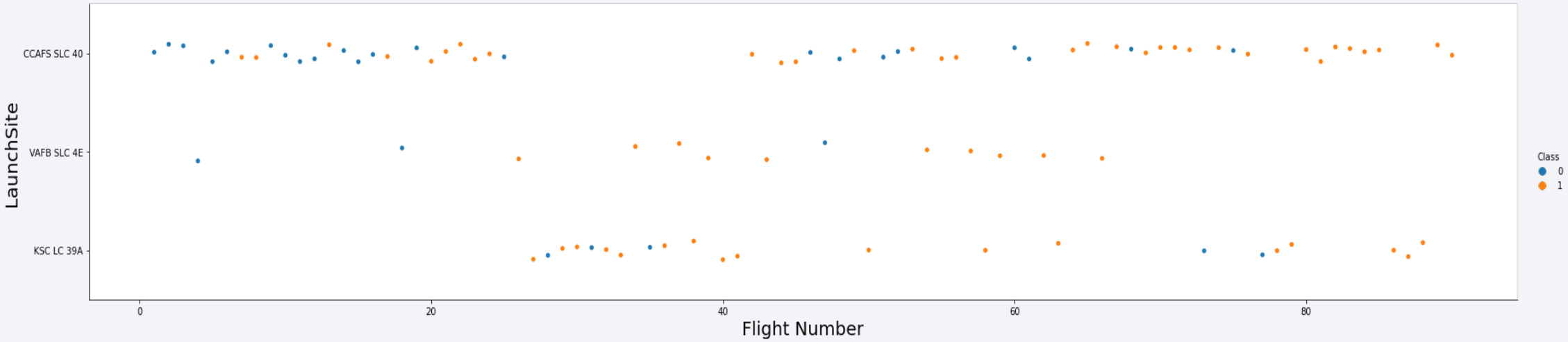
Results

- The predictive analysis performed showed that the Decision Tree Classifier was the best model when it came to predicting successful landings.
- It had a training accuracy of over 87% and a test accuracy of 94%



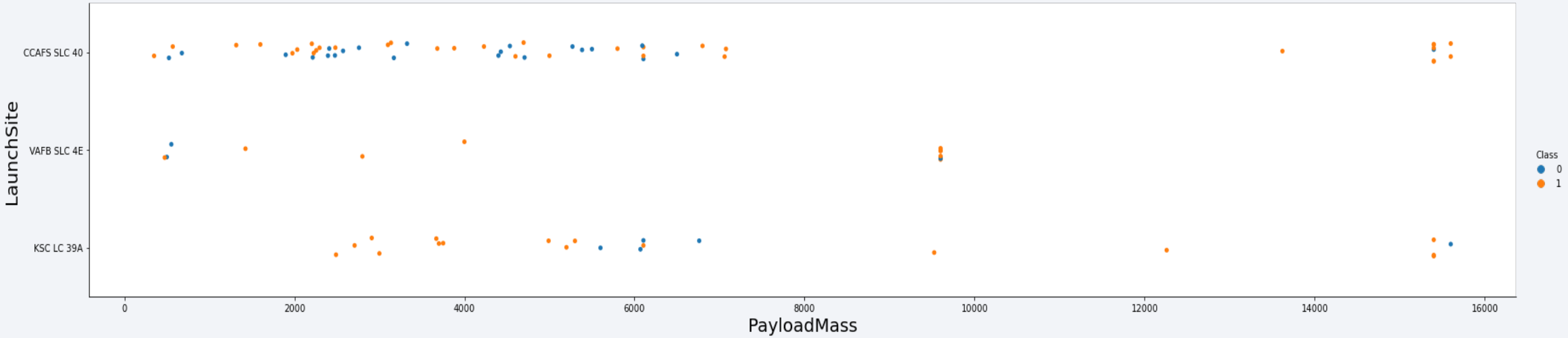

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- It is evident in the graph above that the best launch site was CCAF5 SLC 40 as recent launches were all successful.
- The launch site, VAFB SLC 4E was the next most successful launch site, albeit with fewer launches.
- Launch site, KSC LC 39A was the least successful.

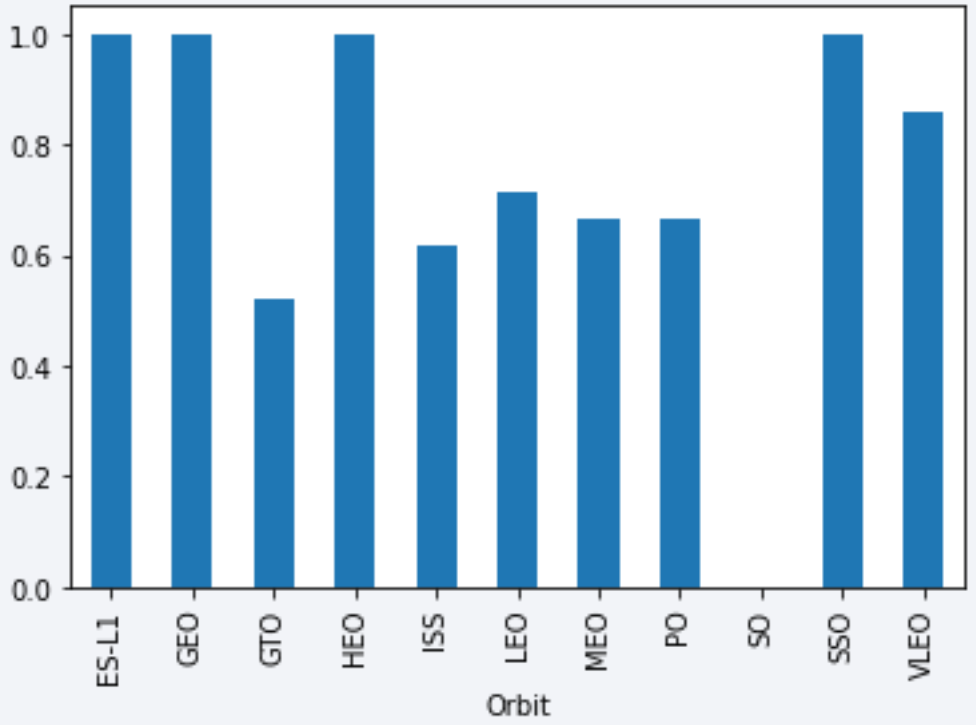
Payload vs. Launch Site



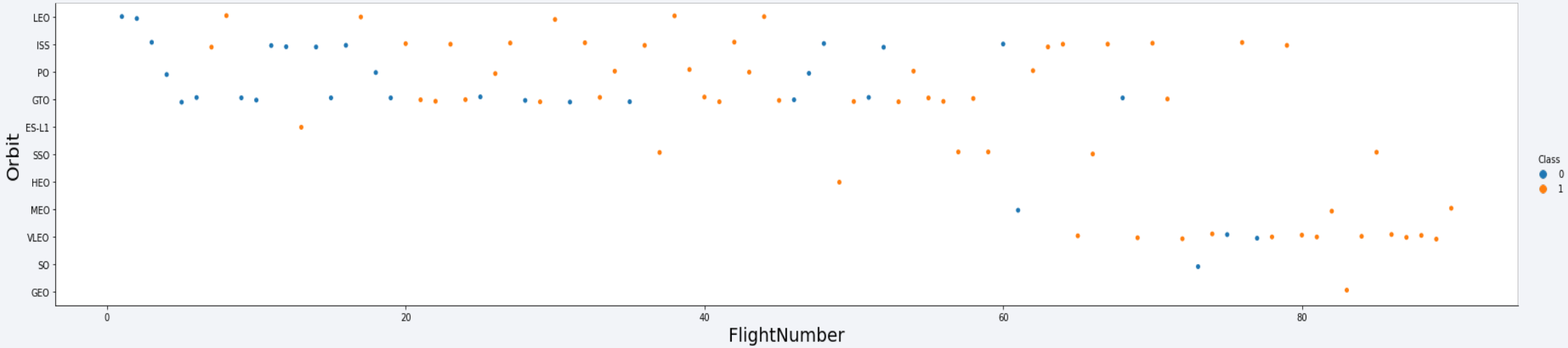
- Payloads with a mass greater than 9,000kg had a greater success rate than those with a lighter payload mass.
- Payloads greater than 12,000kg are only possible at the launch site locations CCAFS SLC 40 and KSC LC 39A.

Success Rate vs. Orbit Type

- The following orbits had a 100% success rate:
  - ES-L1,
  - GEO,
  - HEO and
  - SSO
- Launch site SO had no launches, unsuccessful launches or no data
- The next successful launch sites where:
  - VLEO (80% success) and
  - LFO (70% success)

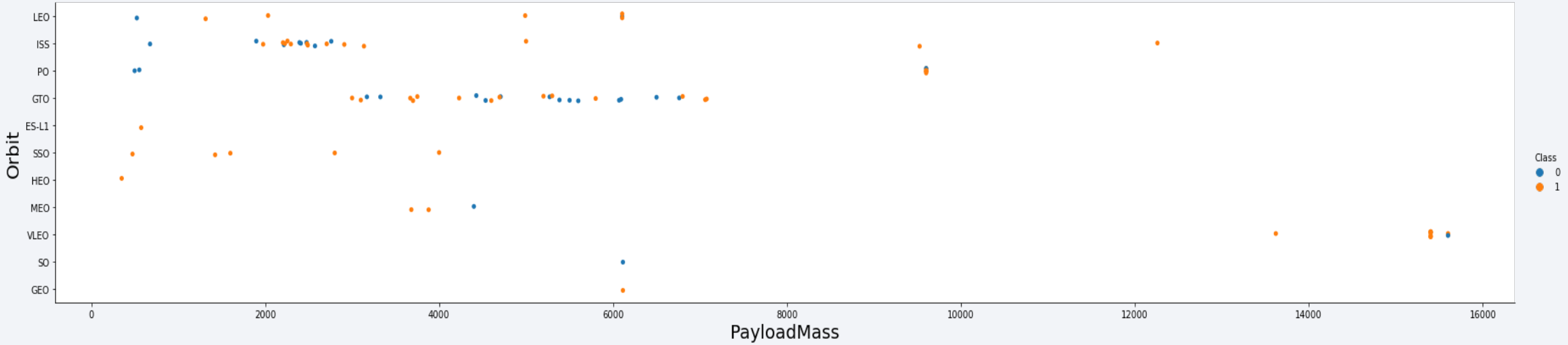


Flight Number vs. Orbit Type



- The graph shows a trend that all orbits improved with time and the increased number in flights
- VLEO is a newer orbit than the other orbits and is showing great return with a high success rate.

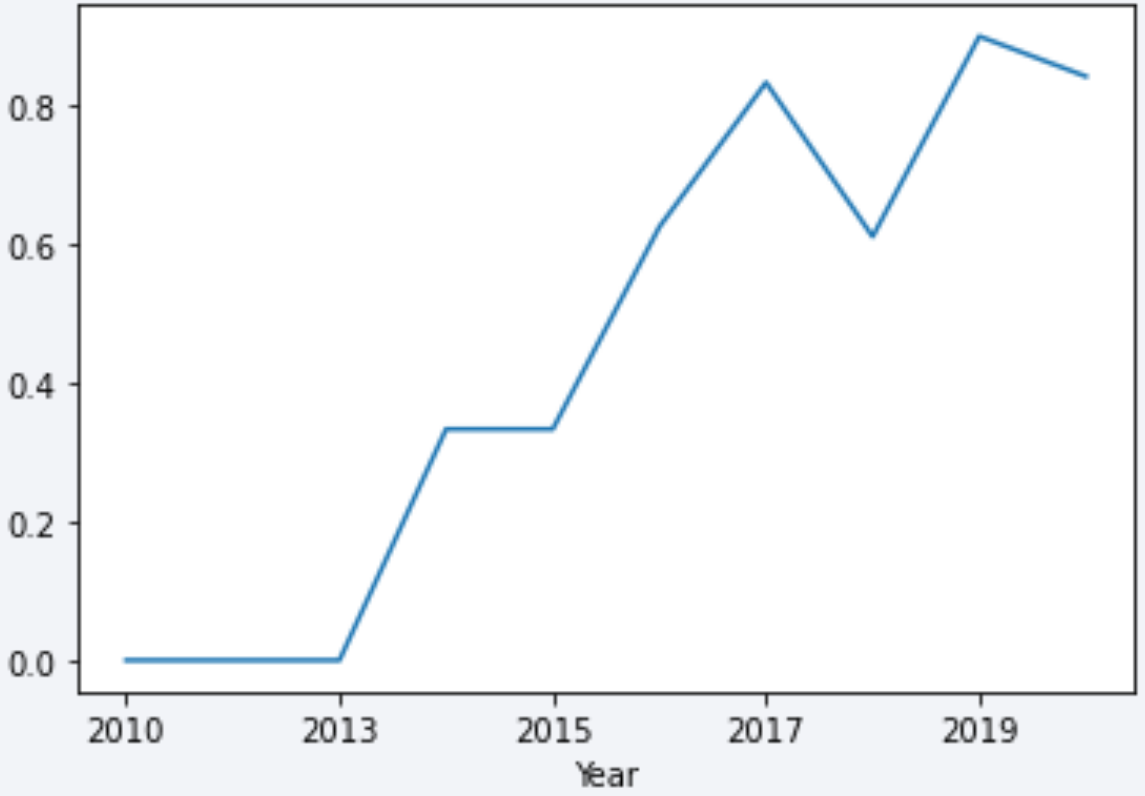
Payload vs. Orbit Type



- There appears to be no relation between the GTO orbit and payload mass
- On the contrary, orbit ISS has a stronger relation for a variety of payload masses as it achieved a high success rate

Launch Success Yearly Trend

- The success rate steadily increases from the year 2013 through till 2020
- There were was the exception of 2018 and 2020
- It appears that the first 3 years no successful landings occurred, which was probably due to initial tests and improvements



All Launch Site Names

- The analysis showed that there are a total of 4 launch sites:

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

- The above results were obtained by performing a distinct sql query on the launch sites feature

Launch Site Names Begin with 'CCA'

- Below is a table showing 5 records where launch sites 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	**CCA**FS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	**CCA**FS 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	**CCA**FS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	**CCA**FS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	**CCA**FS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attemp

- It is clear to see that the launch site CCA stands for Cape Canaveral.

Total Payload Mass

- The total payload carried by boosters from NASA was:

Total Payload (kg)
111.268

- The total payload was calculating by summing all the individual payload masses where the booster code contained ‘CRS’ and the customer was NASA

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 was:

Avg Payload (kg)
2.928

- The data was filtered for the above booster version, with the average payload mass being calculated on the remaining booster payload masses.

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad was:

Min Date
2015-12-22

- The data was first filtered for all successful landings on ground pads where the minimum date value of the resulting data set was retrieved, which was 22/12/2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

- A list of the booster names which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 is shown below:

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

- The distinct sql query function was used to identify the booster versions that fit the above payload mass threshold range.

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes was:

Mission Outcome	Occurrences
Success	99
Success (payload status unclear)	1
Failure (in flight)	1

- The mission outcomes were grouped and the total number of records were counted for each group as shown in the summary above.

Boosters Carried Maximum Payload

- The boosters which have carried the maximum payload mass were:

Booster Version	Booster Version (...)
F9 B5 B1051.4	F9 B5 B1048.4
F9 B5 B1051.6	F9 B5 B1048.5
F9 B5 B1056.4	F9 B5 B1049.4
F9 B5 B1058.3	F9 B5 B1049.5
F9 B5 B1060.2	F9 B5 B1049.7
F9 B5 B1060.3	F9 B5 B1051.3

- The maximum payload mass was identified and then later used to filter for the booster versions that carried this maximum mass.

2015 Launch Records

- The failed landing outcomes in drone ship, their booster versions, and launch site names for the year 2015 are shown below:

Booster Version	Launch Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

- The data was filtered for the year 2015. At the same time, only failed launches were searched for as well as launches that happened on drone ships.

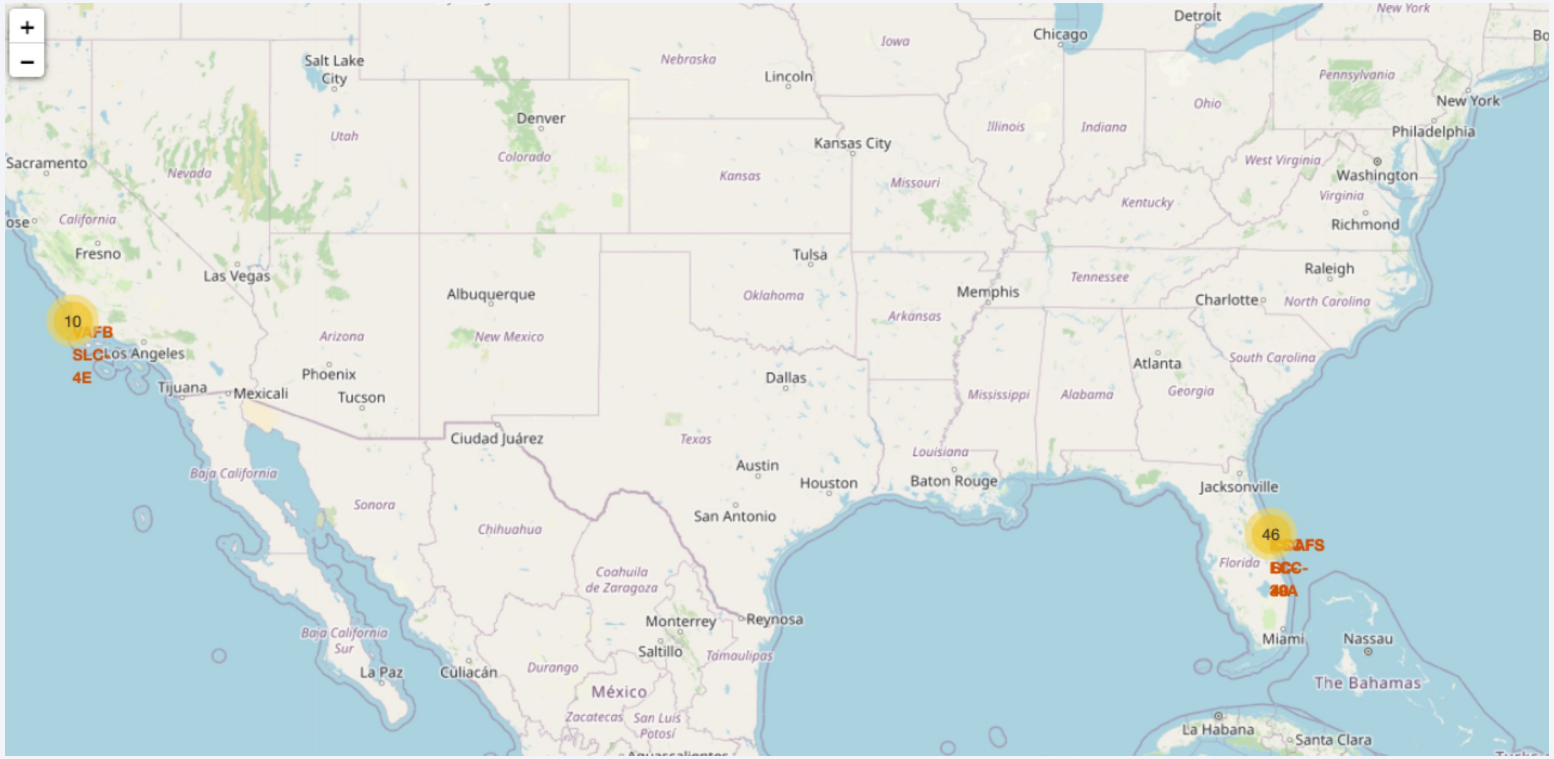
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- A rank of 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 is shown below:

Landing Outcome	Occurrences
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

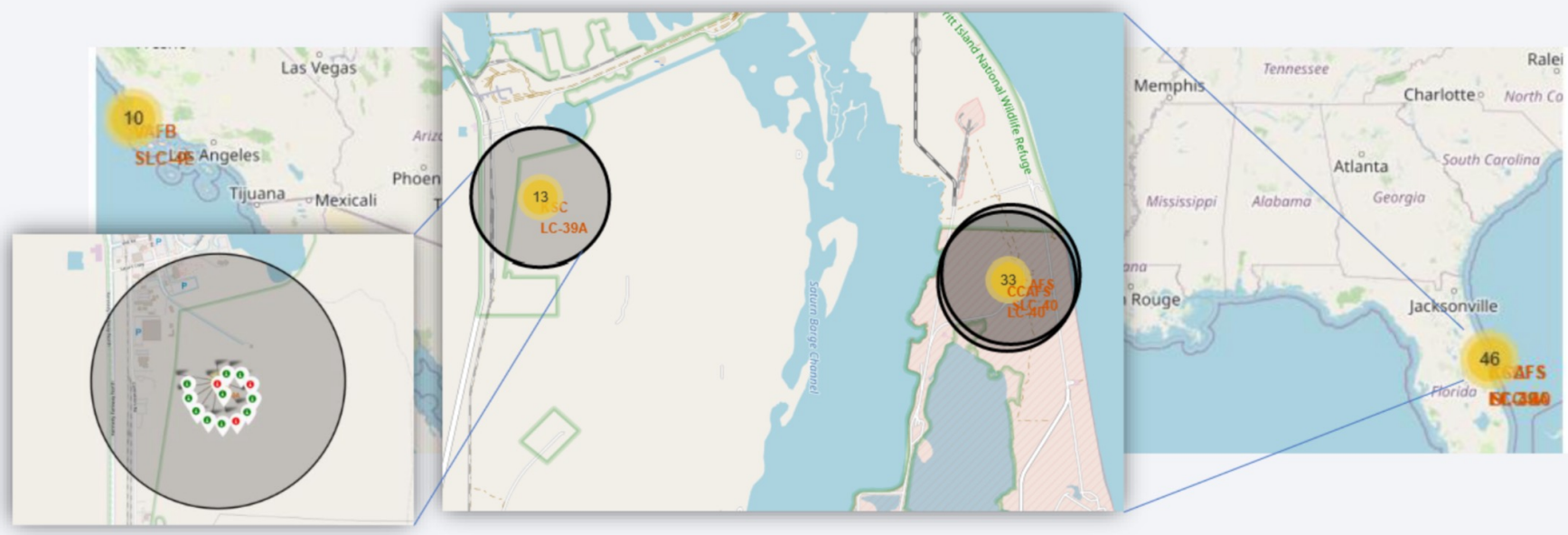

Section 3
Launch Sites
Proximities Analysis

Overview of All Launch Sites



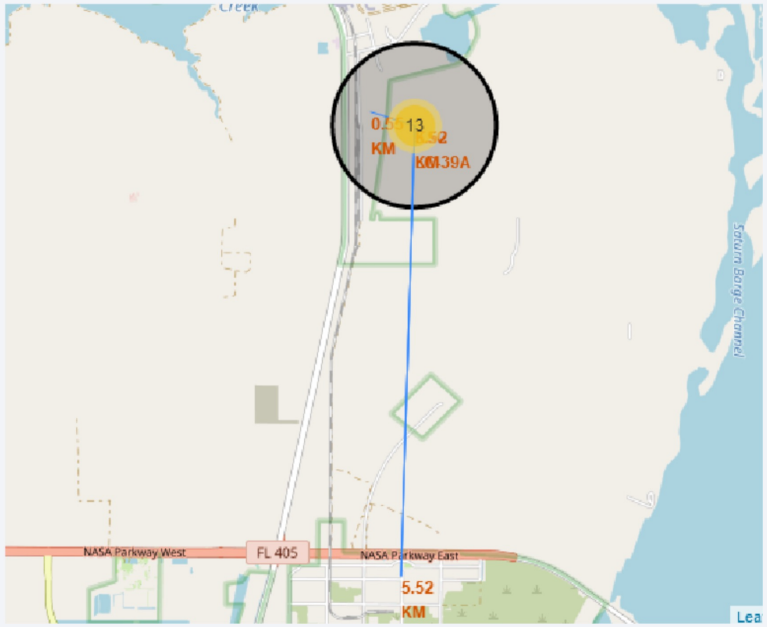
- Launch sites are all located near the sea, with the majority of launch sites being located on the east coast of North America.

Launch Sites by Outcome

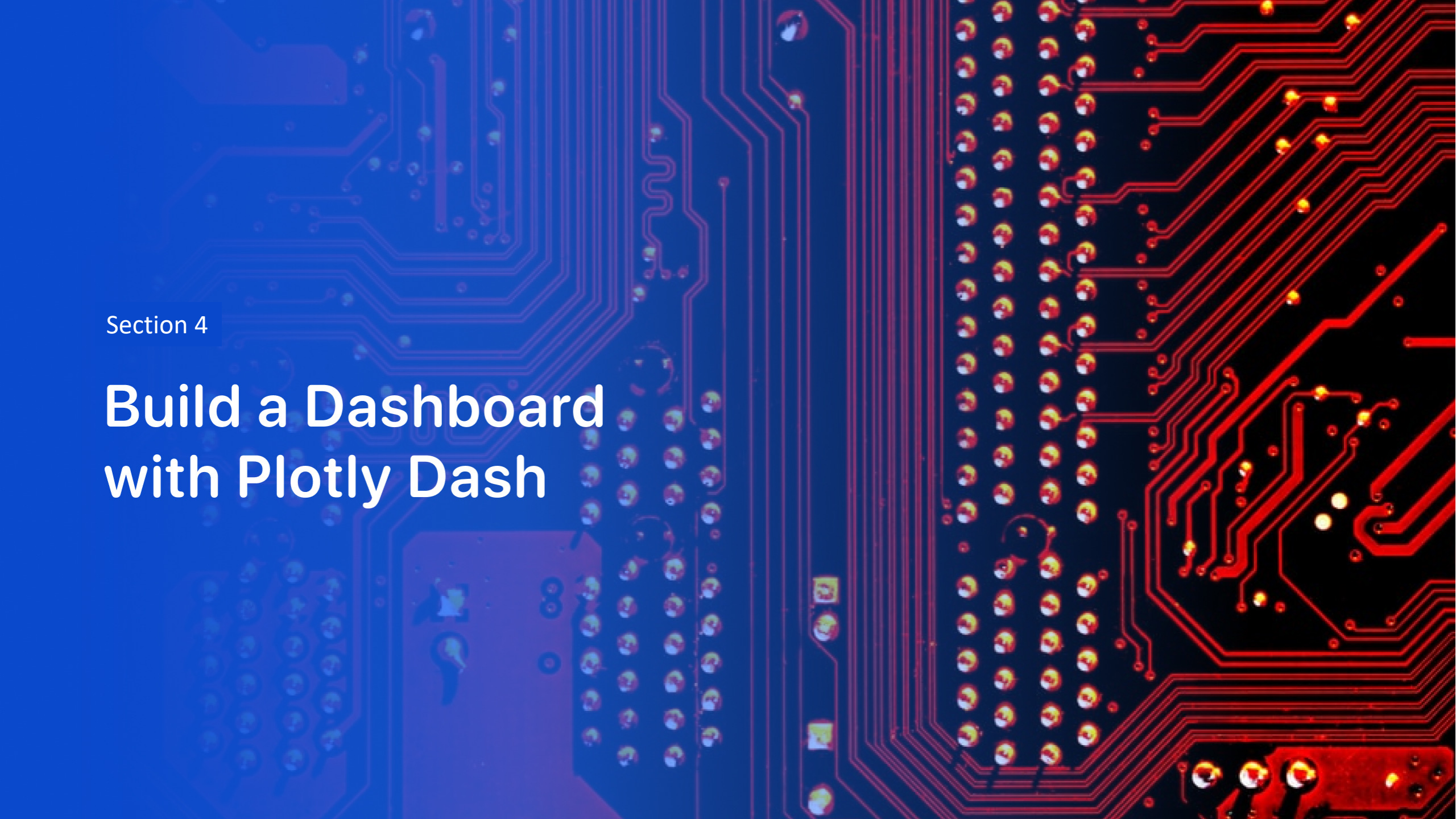


- The above screenshots show the outcomes (both successful and unsuccessful) for launch site KSC LC-39A.
- The green markers, found on the left most screenshot, indicate successful launches, while red markers indicate unsuccessful launches

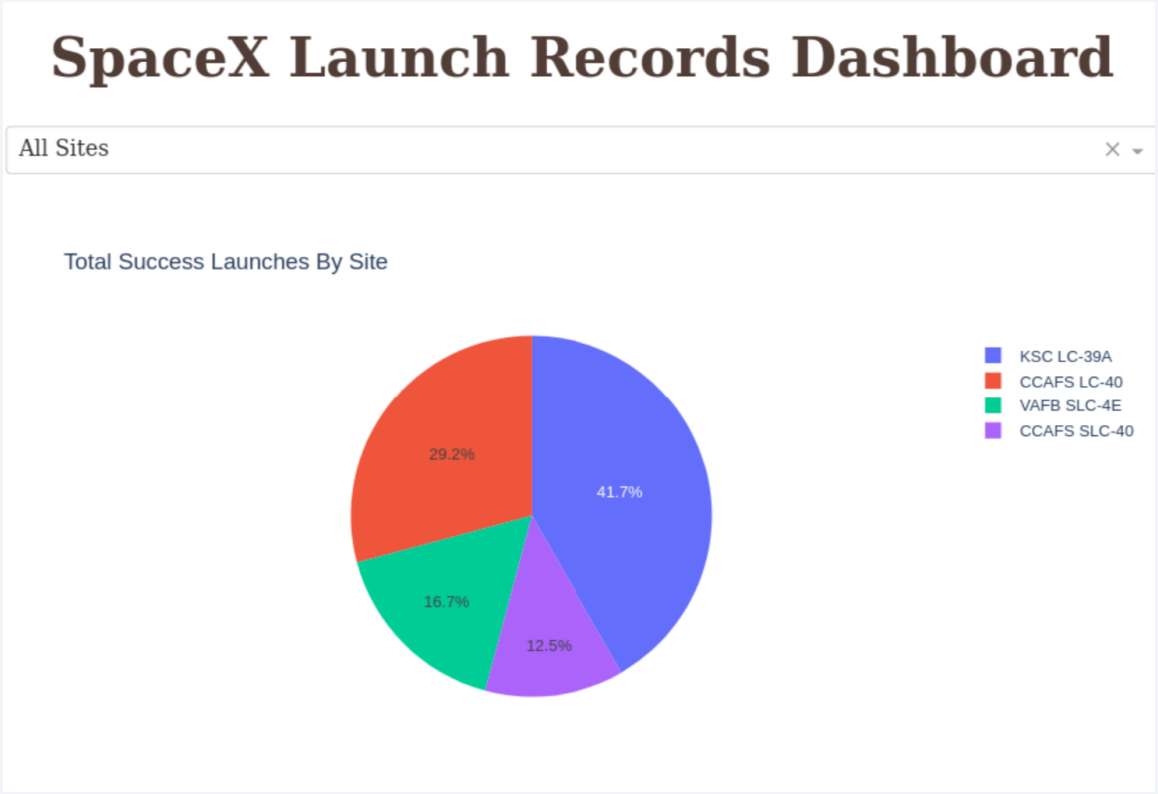
Logistics & Safety Infrastructure



- A screenshot of the logistics and safety infrastructure surrounding the KSC LC-39A launch site is shown above.
- From this image it is clear to see that the launch site is in close proximity to a rail-road (roughly 5km) and other key infrastructures that are necessary in a successfully operating launch site.

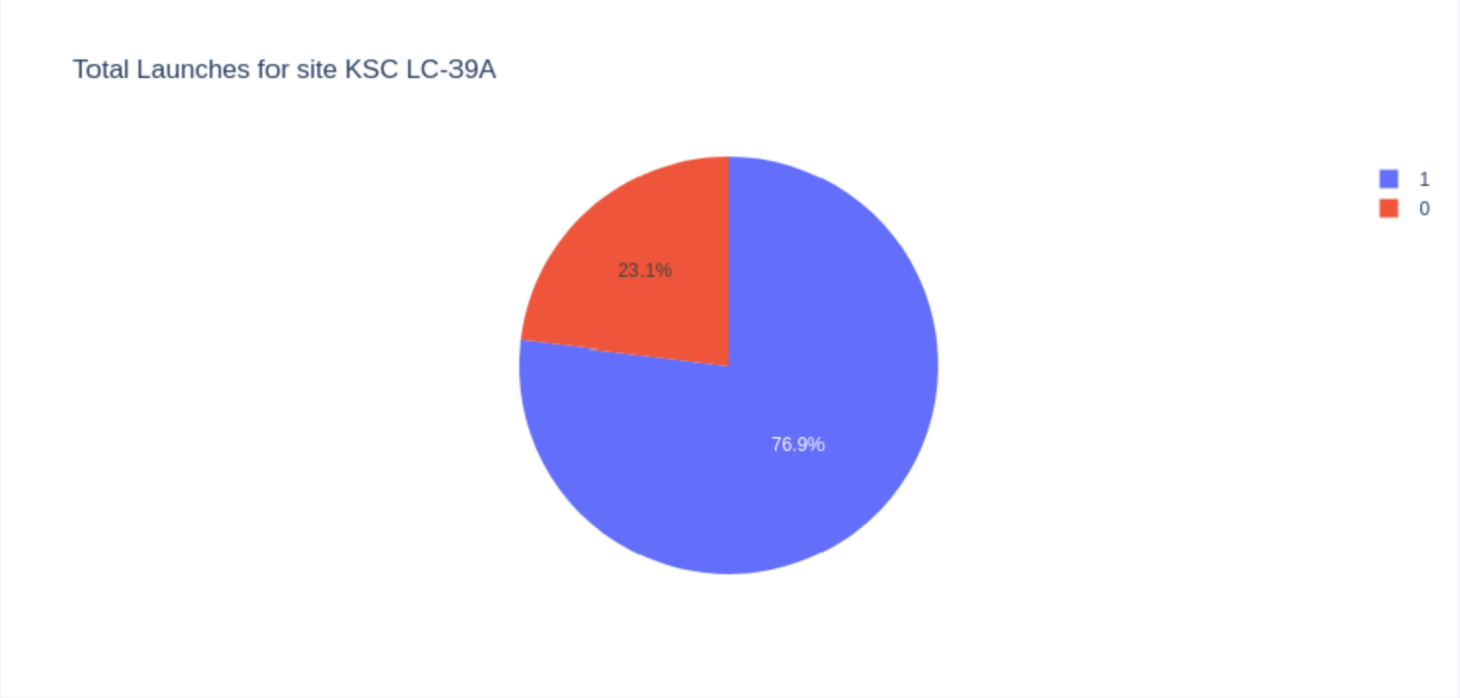

Section 4
Build a Dashboard with Plotly Dash

Total Successful Launches by Site



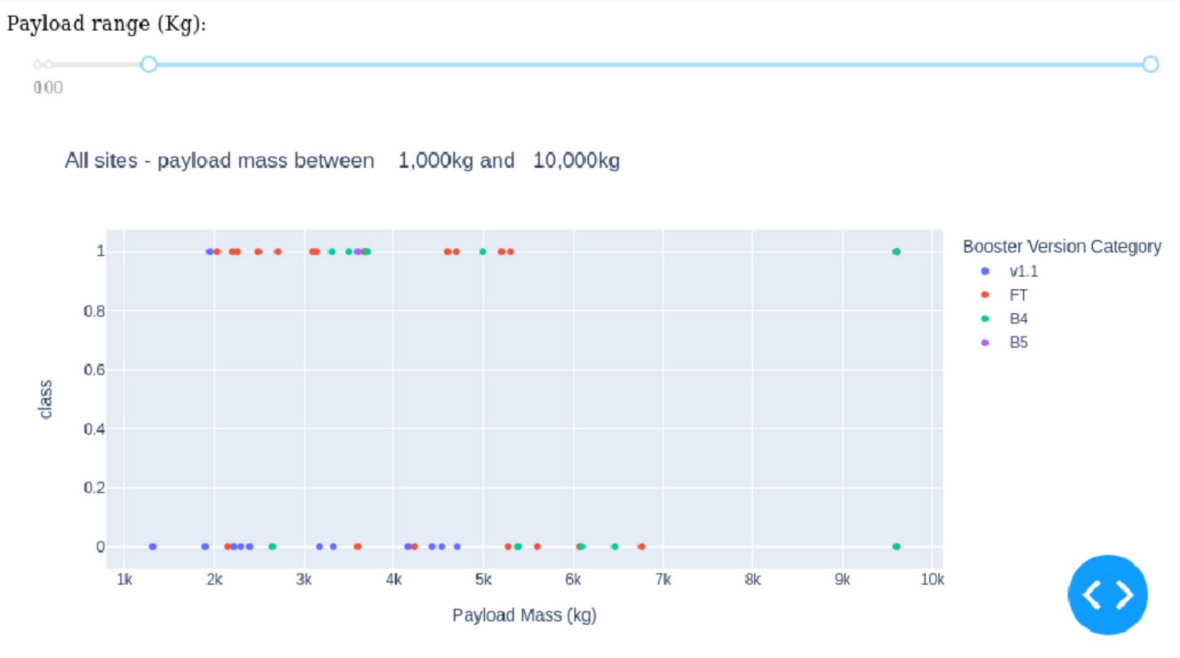
- Launch site KSC LC-39A had the largest percentage of successful launches with the launch site CCAFS LC-40 having the second largest percentage

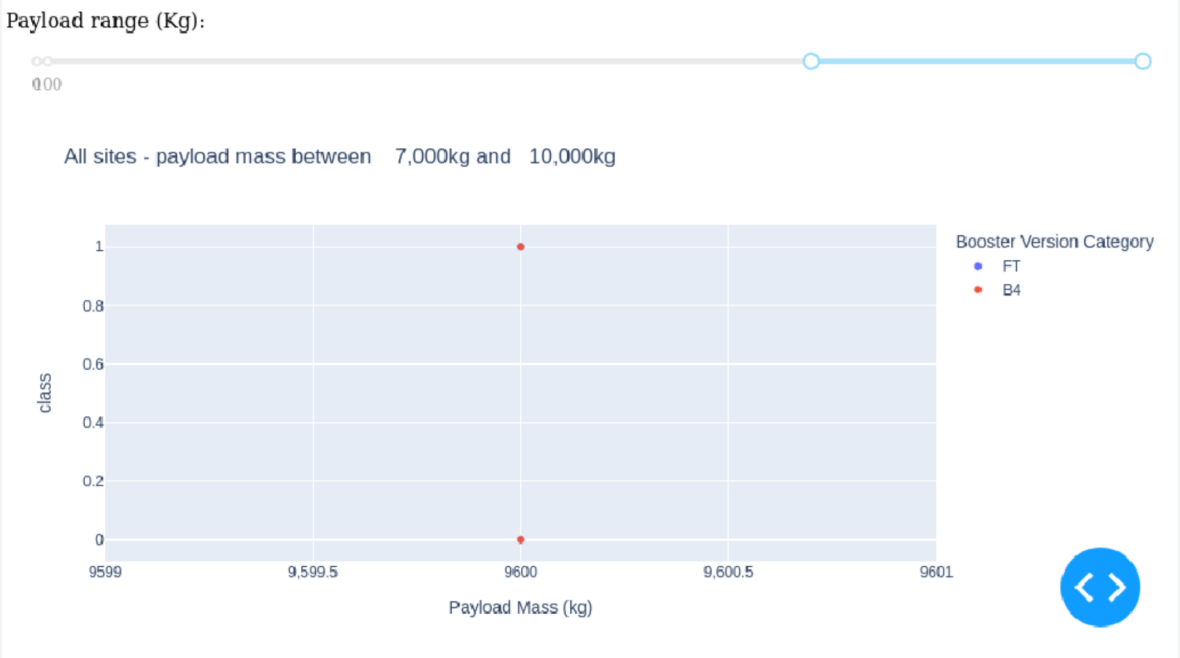
Launch Success Ratio for KSC LC-39A



- The most successful launch site had a total success ratio of 76.9% as seen in the above pie chart.

Payload vs Launch Outcome





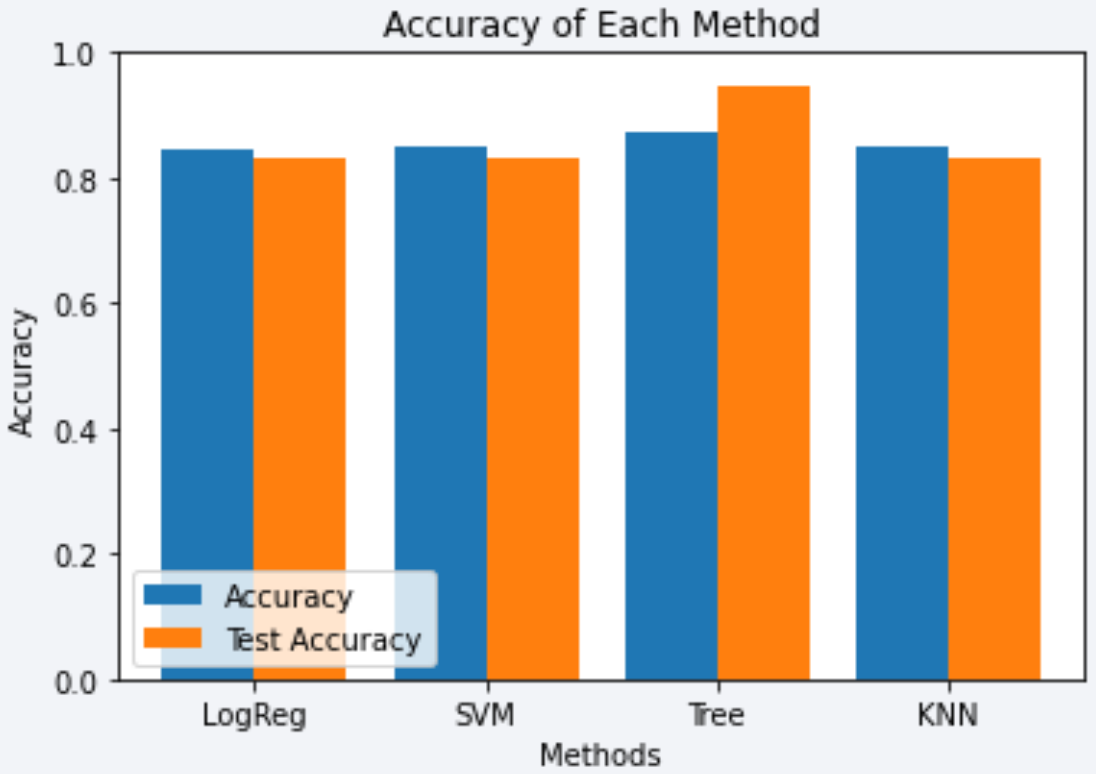
- Payloads that were under 6,000kg and FT booster types were the most successful combination.
- Unfortunately, there was not sufficient data to draw conclusions on the relationship between payload masses and the success class for payloads greater than 7,000kg

Section 5

Predictive Analysis (Classification)

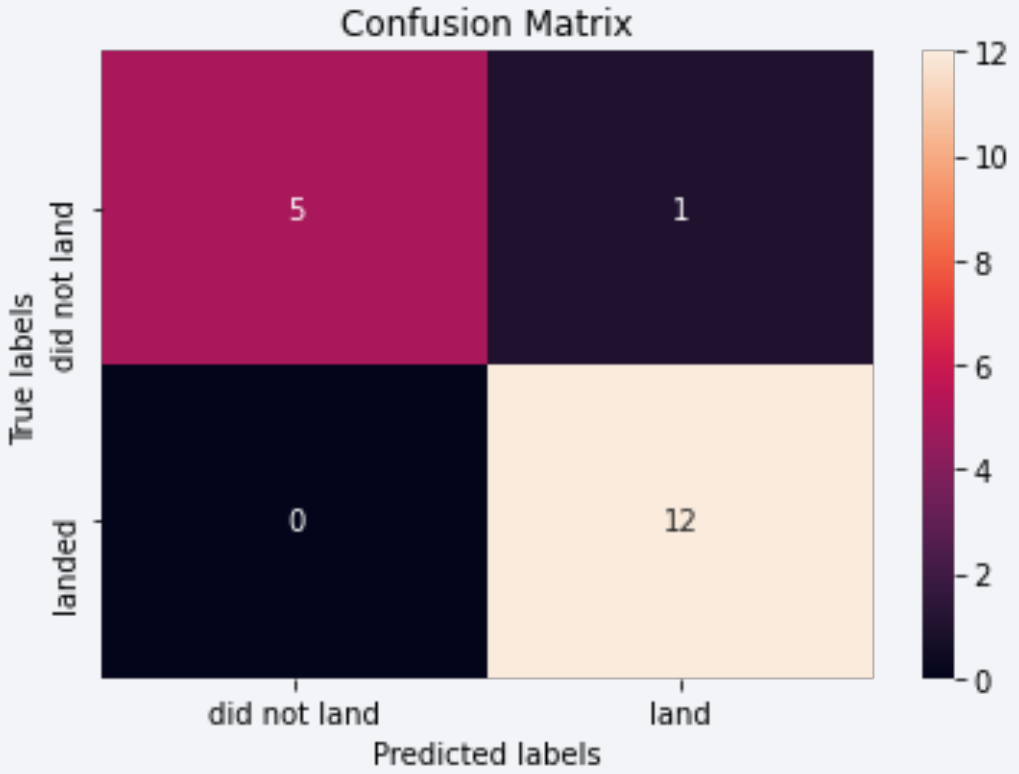
Classification Accuracy

- A total of four classification models were implemented at tested with their accuracies for both training and testing being plotted in the graph on the right
- The model with the highest classification accuracy was the Decision Tree Classifier, with an accuracy of over 85%.



Confusion Matrix – Decision Tree Classifier

- The confusion matrix for the most successful classifier, namely, the Decision tree is shown on the right.
- The confusion matrix supports the accuracy results achieved by this classifier with 17 of the 18 outcomes correctly classified as either a true positive or true negative.



Conclusions

- A variety of data sources were extracted and analysed to refine conclusions throughout the process in identifying a successful launch site.
- The best launch site, given the high number of successful outcomes, among other factors, was KSC LC-39A.
- Launches with a payload mass greater than 7,000kg are deemed to be less risky, with there being a high correlation with successful outcomes.
- Although there are many missions with successful outcomes, there was a general trend that saw the total number of successful outcomes per launch site increasing with time, due to the improvement and evolution of boosters, rockets and equipment.
- The Decision Tree Classifier can be used to predict successful landings and increase profits.

Appendix

- As an improvement for model tests, the *np.random.seed* variable.
- For more information on the code, please refer to the GitHub links that can be found at the bottom of the majority of slides.

Thank you!