IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Erkin Altuntas
20.01.2023

Outline

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

Executive Summary

- The present data analysis is done by applying following methodologies:
  - Falcon 9 data is collected through web scaping and using the Restful API of SpaceX
  - Exploratory data analysis is done by wrangling the collected data and creating insightful and interactive visualizations
    - Therefore, SQL statements and Python Notebooks and Scripts are developed
  - It is determined if the first stage of Falcon 9 will land successfully, by developing various Machine Learning Models
- Summary of all results
  - Falcon 9 data is collected and prepared successfully for further analysis and prediction
  - Exploratory data analysis showed meaningful insights into launching and landing of Falcon 9
  - Trying different Classification Models showed that the success of launching can be predicted accurately through the collected data

Introduction

- Background of the project is the importance but cost-extensive process of launching rockets.
- Launching rockets encompasses two sages
  - If first stage will land, we can determine the cost of a launch
- In order to reduce the cost of launches, we want to find out:
  - Determine the best place to launch rockets
  - Determine if SpaceX will reuse the first stage
  - Determining the price of each launch

Section 1

Methodology

Methodology

Executive Summary

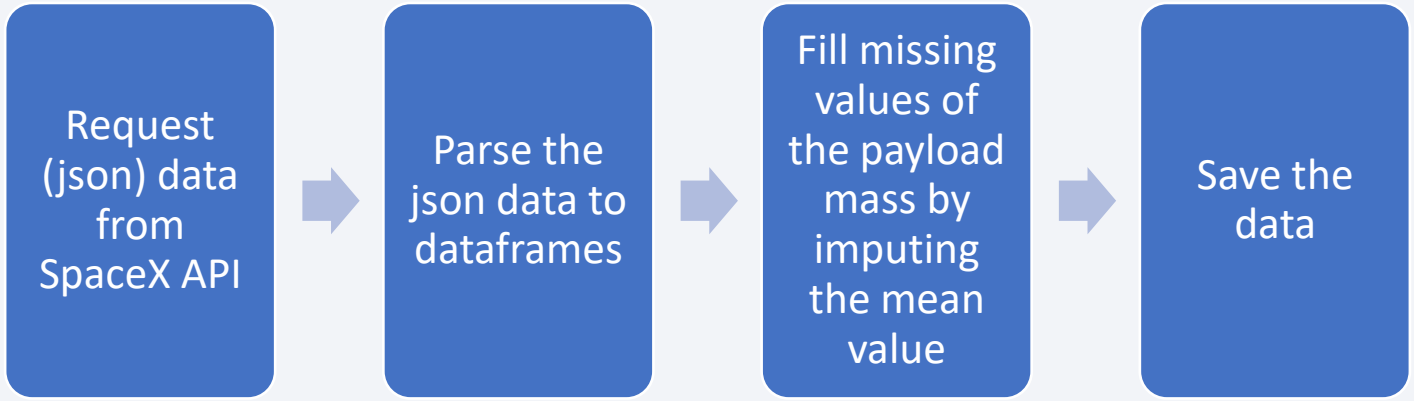
- Data collection methodology:
  - Data from Space X is collected through:
    - Requesting data from the REST API of Space X (https://api.spacexdata.com/v4/rockets/)
    - WebScraping in Wikipedia (https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922)
- Perform data wrangling
  - Creating a landing outcome label (for the predictive analysis) and adding it to the data
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - The wrangled data is normalized and split into a training and testing set (80/20 ratio)
  - Logistic Regression, SVM, KNN, and Tree models are trained on the train set
    - Their parameters are optimized using GridSearchCV
  - The models are evaluated by their accuracy on the test set

Data Collection

- Data from Space X is collected through:
  - Requesting data from the REST API of Space X (https://api.spacexdata.com/v4/rockets/)
  - WebScraping in Wikipedia (https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922)
- Details of the data collection process are shown in the next slides

Data Collection – SpaceX API

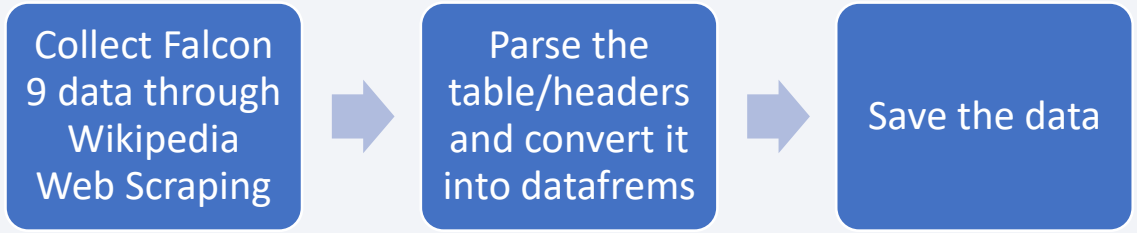
- Rocket launch data of the past is collected from the SpaceX API (https://api.spacexdata.com/v4/rockets/), filtered in order to obtain only Falcon 9 data, and prepared and preprocessed.
- Following flowchart shows the process of data collection via the SpaceX API



- GitHub URL: https://github.com/erkinaltuntas/applied_data_science_capstone/blob/master/Data%20Collection%20API%20Lab.ipynb

Data Collection - Scraping

- Rocket launch data of is additionally collected from Wikipedia (https://api.spacexdata.com/v4/rockets/) through web scraping.
    - Further, it is prepared and preprocessed for the further analysis.
- Following flowchart shows the process of data collection via WebScraping



- GitHub URL: https://github.com/erkinaltuntas/applied_data_science_capstone/blob/master/Data%20Collection%20with%20Web%20Scraping%20lab.ipynb

Data Wrangling

- Exploratory data analysis is done on the collected data
  - Identify the available columns/attributes
  - Calculate the ratio of missing values per attribute
  - Count/Group the data per Launch Site/Orbit/Outcome
- Landing outcome label is created, depending on the value of the Outcome column
  - 0 = No successful outcome/“Bad outcome”
  - 1 = Successful outcome

- GitHub URL: https://github.com/erkinaltuntas/applied_data_science_capstone/blob/master/EDA%20lab.ipynb

EDA with Data Visualization

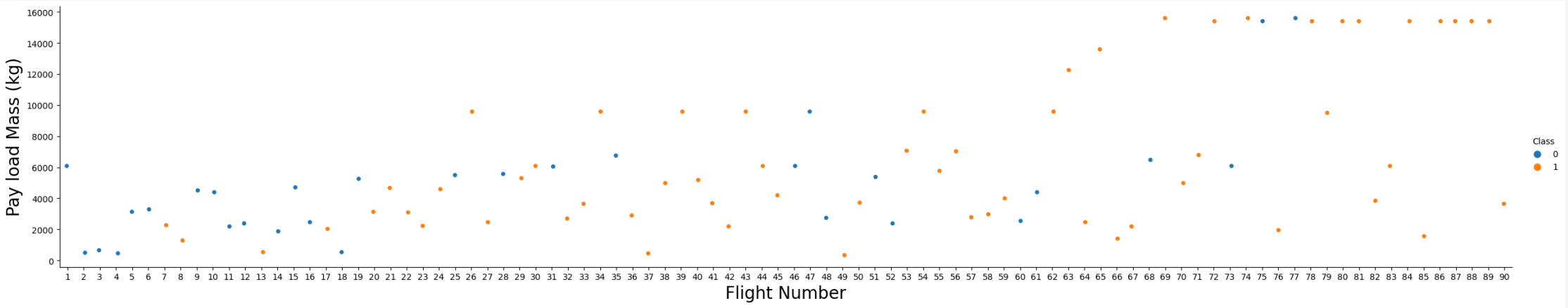
- Exploratory data analysis is done by visualizing the relationship between various features through scatterplots, barplots, and lineplots

- The plots as well as explanations are shown in section 2

- GitHub URL: https://github.com/erkinaltuntas/applied_data_science_capstone/blob/master/DataVis%20EDA.ipynb

EDA with Data Visualization

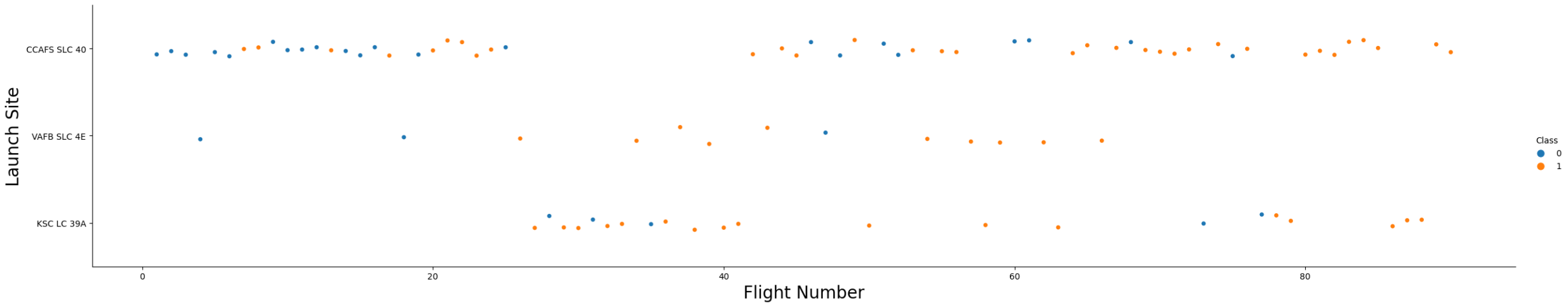
- Following chart shows the the pay load mass per flight number and landing outcome



- GitHub URL: https://github.com/erkinaltuntas/applied_data_science_capstone/blob/master/DataVis%20EDA.ipynb

EDA with Data Visualization

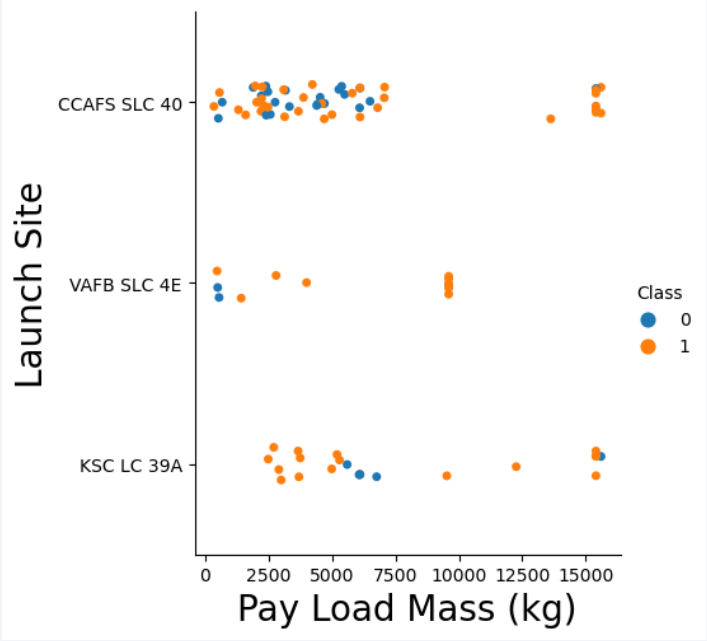
- Following chart shows the relationship between Flight Number and Launch Site



- GitHub URL: https://github.com/erkinaltuntas/applied_data_science_capstone/blob/master/DataVis%20EDA.ipynb

EDA with Data Visualization

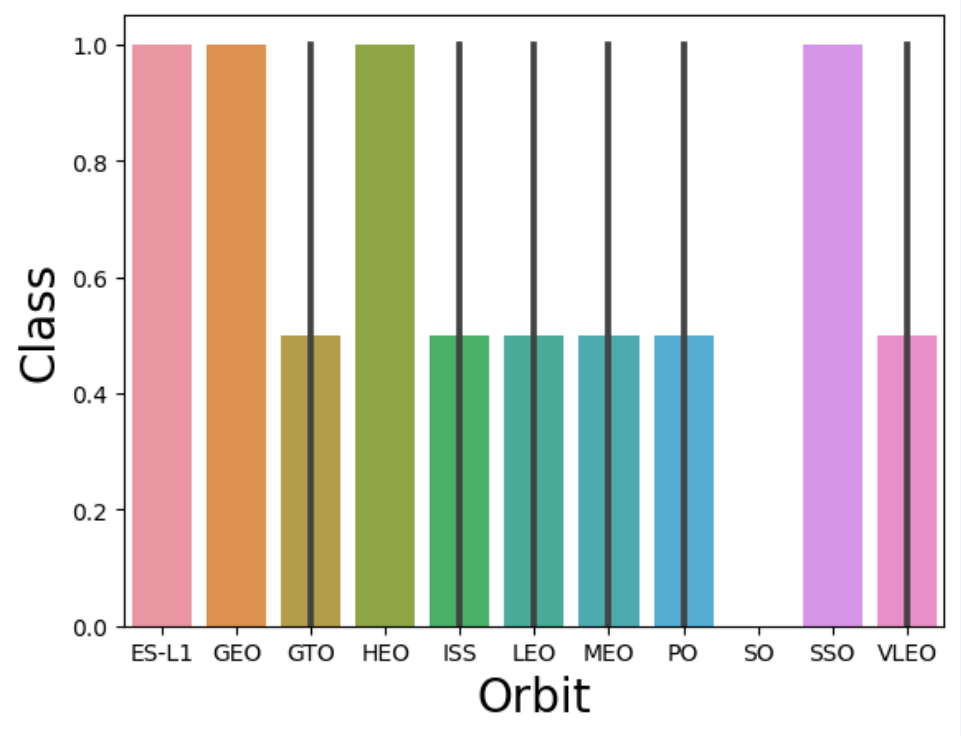
- Following chart shows the relationship between Payload and Launch Site



- GitHub URL: https://github.com/erkinaltuntas/applied_data_science_capstone/blob/master/DataVis%20EDA.ipynb

EDA with Data Visualization

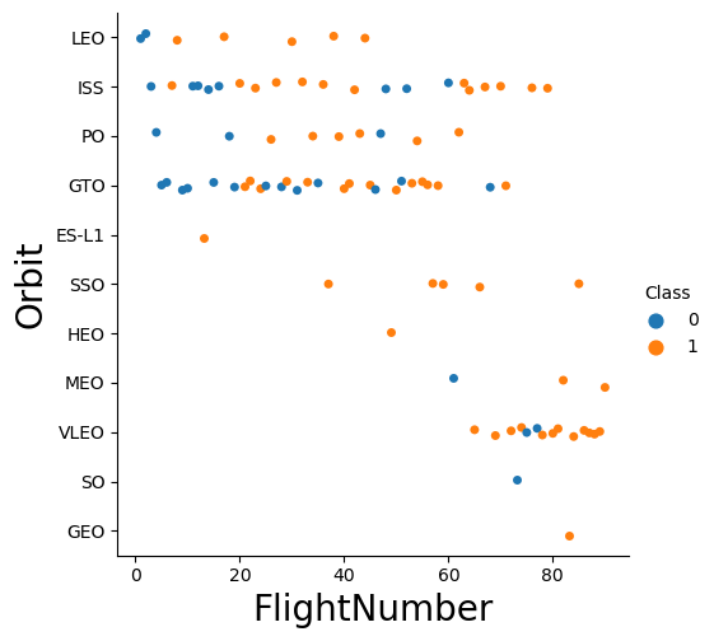
- Following chart shows the relationship between success rate of each orbit type



- GitHub URL: https://github.com/erkinaltuntas/applied_data_science_capstone/blob/master/DataVis%20EDA.ipynb

EDA with Data Visualization

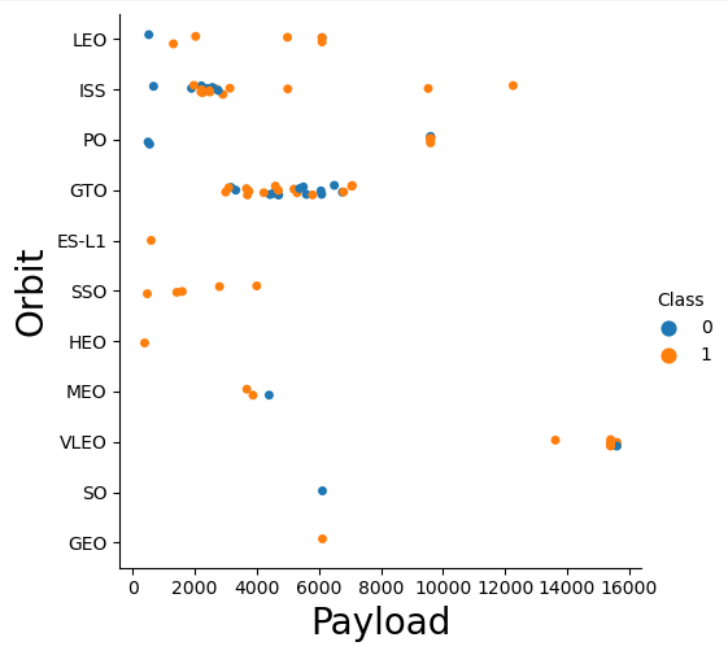
- Following chart shows the relationship between FlightNumber and Orbit type



- GitHub URL: https://github.com/erkinaltuntas/applied_data_science_capstone/blob/master/DataVis%20EDA.ipynb

EDA with Data Visualization

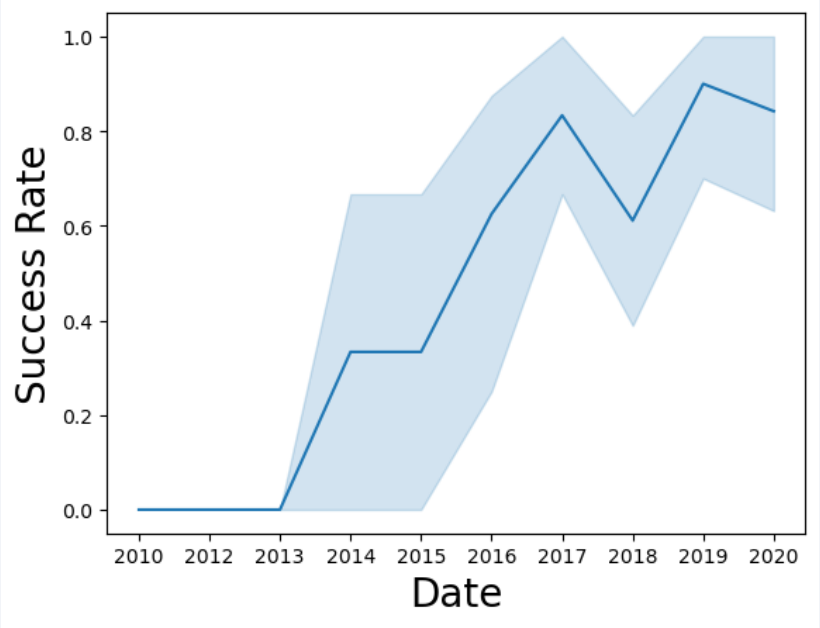
- Following chart shows the relationship between Payload and Orbit type



- GitHub URL: https://github.com/erkinaltuntas/applied_data_science_capstone/blob/master/DataVis%20EDA.ipynb

EDA with Data Visualization

- Following chart shows the launch success yearly trend



- GitHub URL: https://github.com/erkinaltuntas/applied_data_science_capstone/blob/master/DataVis%20EDA.ipynb

EDA with SQL

Using SQL queries, it was identified:

- the names of the unique launch sites in the space mission
- 5 records where launch sites begin with the string 'CCA'
- the total payload mass carried by boosters launched by NASA (CRS)
- average payload mass carried by booster version F9 v1.1
- the date when the first successful landing outcome in ground pad was achieved
- the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- the total number of successful and failure mission outcomes
- the names of the booster_versions which have carried the maximum payload mass
- the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- 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 orde

- GitHub URL: https://github.com/erkinaltuntas/applied_data_science_capstone/blob/master/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

An Interactive Map is built by using the Folium library, with following details:

- Markers are used in order to **point** on specific coordinates
  - All launch sites are marked on the map
- Circles are used in order to highlight **areas** around specific coordinates
  - The NASA Johnson Space Center Area is highlighted
- Lines are used in order to display the **distance** between to locations
  - The distances between launch sites and proximities are identified

- GitHub URL: https://github.com/erkinaltuntas/applied_data_science_capstone/blob/master/Interactive%20Visual%20Analytics%20Folium.ipynb

Build a Dashboard with Plotly Dash

A dashboard is built by using the Ploty library, with following plots/graphs:

- First, a dropdown list to enable Launch Site selection
- Pie charts are added in order to show the total successful launches count for all sites
    - If a specific launch site was selected, show the Success vs. Failed counts for the sit
- Scatter charts are created to show the correlation between payload and launch success
    - As a result, it can be identified the best launch site, depending on the payload

- GitHub URL: https://github.com/erkinaltuntas/applied_data_science_capstone/blob/master/plotly_lab.py

Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model
- Different types of classification models are built, evaluated, and improved in order to find best performing predictive model to predict if the first stage will land given the data from the preceding labs.
  - Logistic Regression Model, Support Vector Machine, K Nearest Neighbor, and a Decision Tree are built
- Therefore, the prepared data is processed through: Standardizing and Splitting it into a train and test set (80%/20% ratio)
- Logistic Regression, SVM, KNN, and Tree models are trained on the train set
- Their parameters are optimized using GridSearchCV
- The models are evaluated by their accuracy on the test set

- GitHub URL: https://github.com/erkinaltuntas/applied_data_science_capstone/blob/master/Prediction%20Lab.ipynb

Results

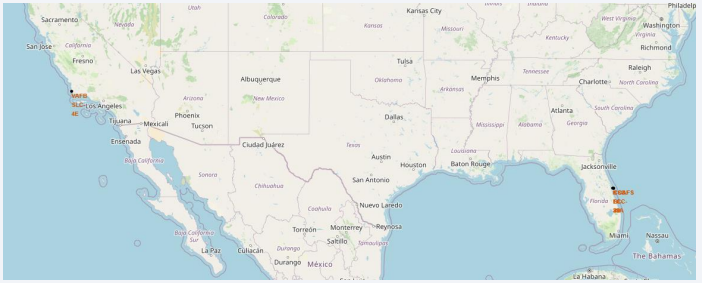
Following insights are derived from the exploratory data analysis:

- There are three different Launch Sites
  - CCAFS SLC 40, KSC LC 39A, VAFB SLC 4E
  - Most launch happen at the launch site “CCAFS SLC 40”
- There are eleven types of Orbits
  - The orbit “GTO” has the highest number of occurrences
- In total, 22007kg payload mass is carried by boosters launched by NASA (CRS)
- The average payload mass carried by booster version F9 v1.1 is 3676kg
- On 2017/01/05 the first successful landing outcome in ground pad was achieved
- The boosters F9 FT B1022 and F9 FT B1031.2 have success in drone ship and have payload mass greater than 4000 but less than 6000
- All 45 mission have a successful outcome
- The bootser versions F9 B5 B1048.4, F9 B5 B1049.4, F9 B5 B1049.5, F9 B5 B1060.2, F9 B5 B1058.3 have carried the maximum payload mass
- The booster version F9 v1.1 B1012 at the launch site CCAFS LC-40 had the only failed landing outcome in 2015

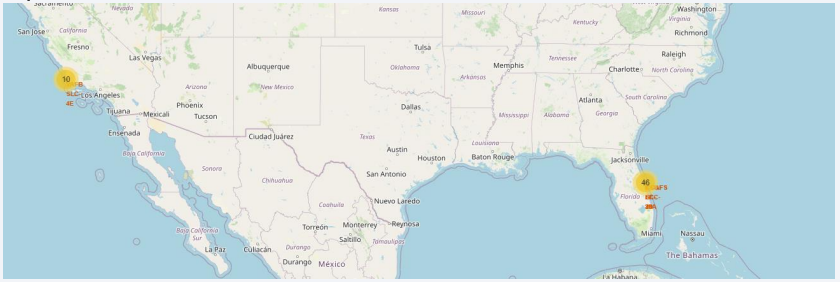
Results

Following insights are derived from the interactive analysis via Folium:

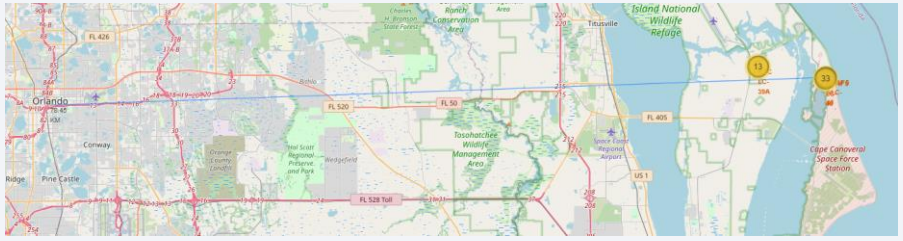
- Launches happen in the east and west near the cost



- Most of the launches (46 in the past) happen in the east, while 10 launches are recorded in the west



- The launch site in the east is near to the city Orlando



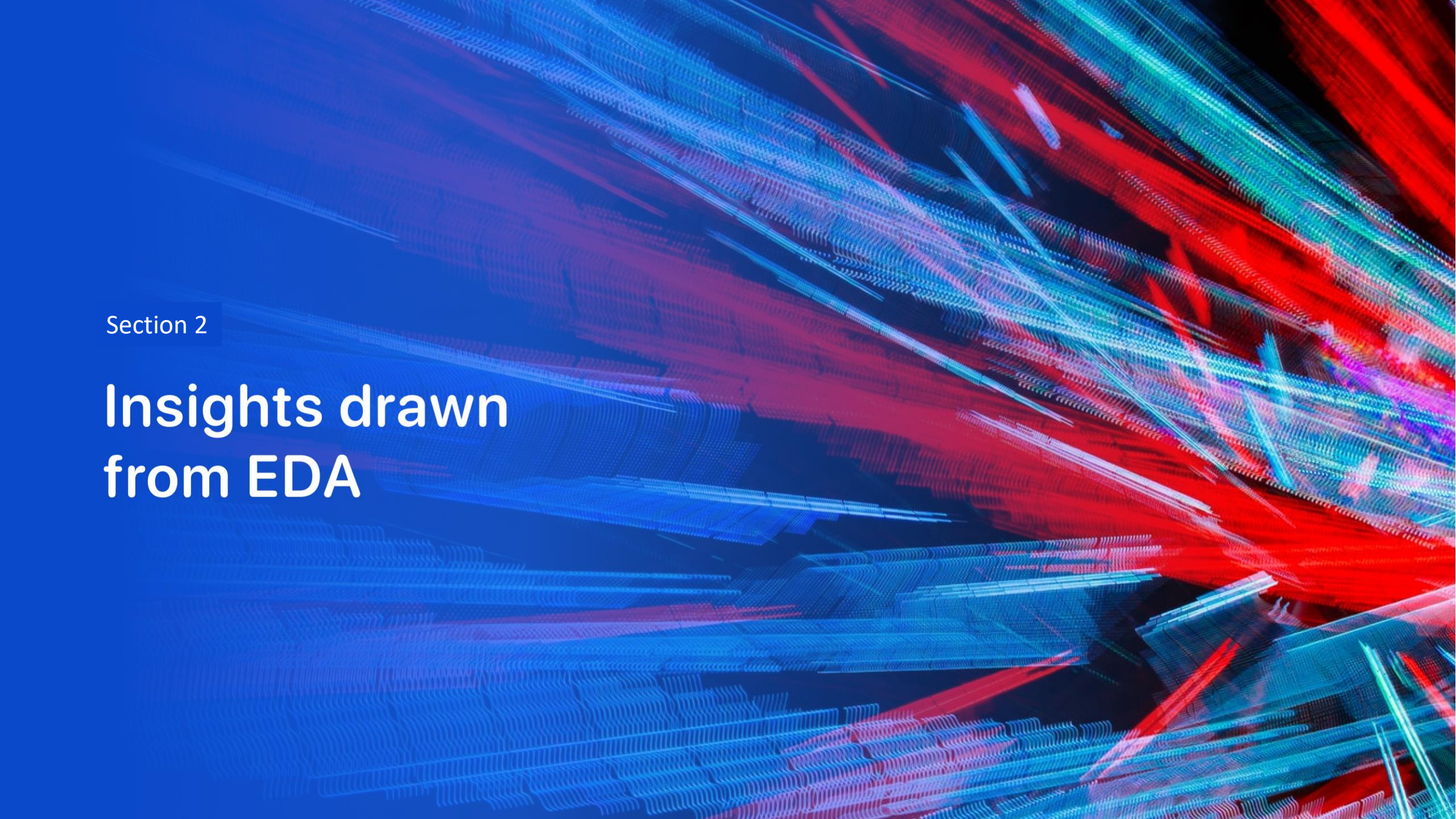
Results

Predictive analysis results

- Accuracy on the train set:
  - Logistic Regression: 0.84
  - SVM: 0.84
  - KNN: 0.84
  - Tree: 0.89

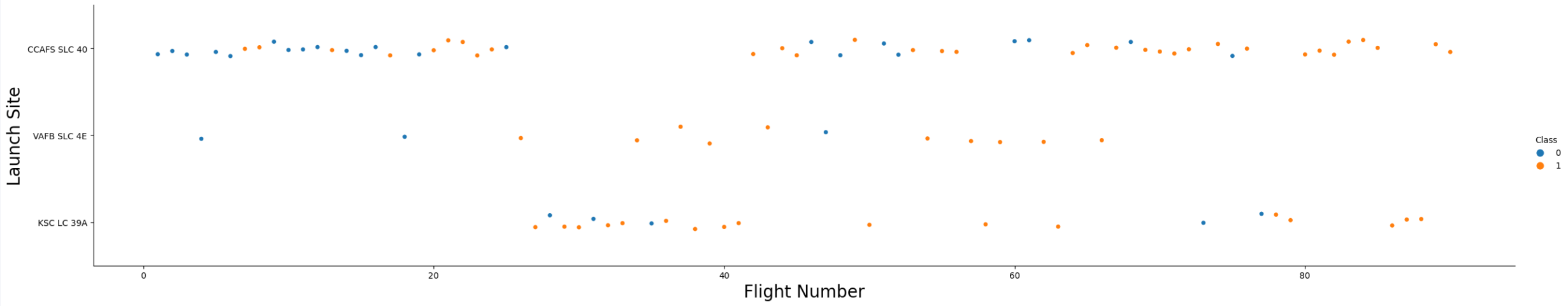
- Accuracy on the test set:
  - Logistic Regression: 0.83
  - SVM: 0.83
  - KNN: 0.83
  - Tree: 0.83

- As the results show, all models indicate the same accuracy on the test set, but the Tree-Classifier has the performance and the train set
- So, the Tree-Classifier can be indicated as the best predictive approach in this project


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

- The chart shows the relationship between Flight Number and Launch Site

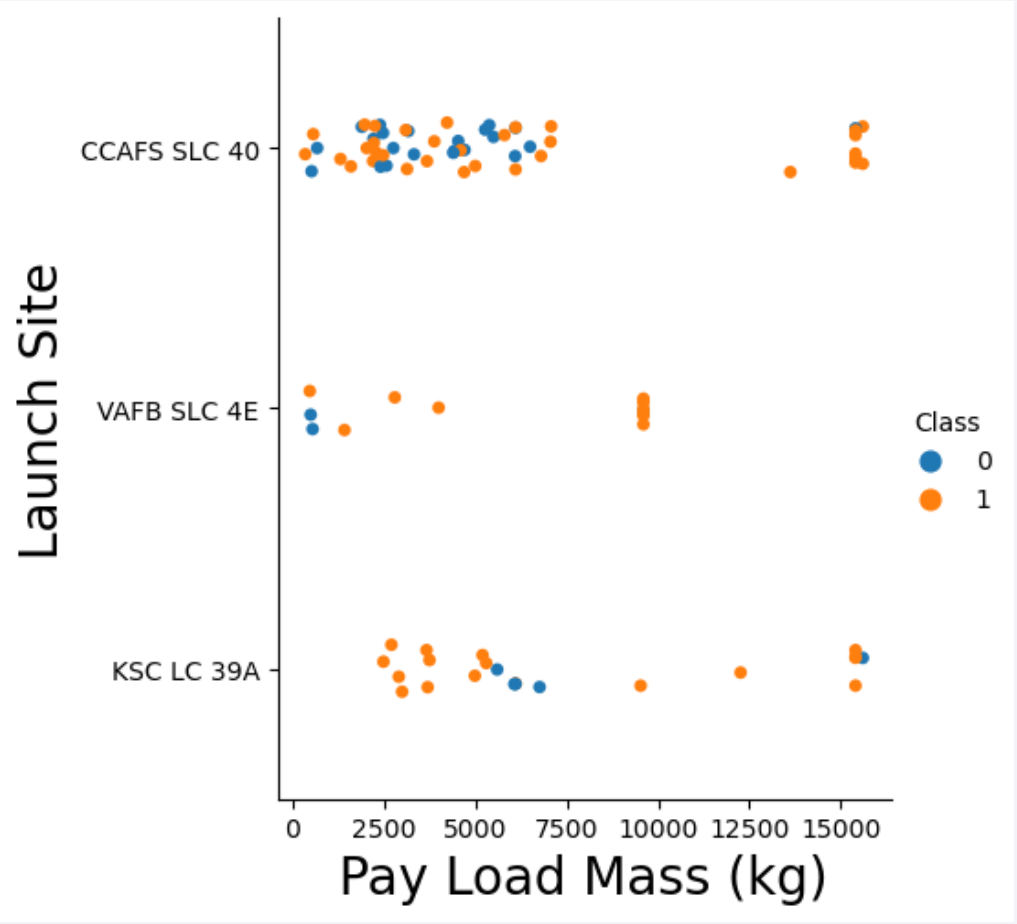


- The most launches are done at the site CCAFS SLC 40
  - Also this launch site indicates the most successful launches
- The launch sites VAFB SLC 4E has the least number of unsuccessful lunches

Payload vs. Launch Site

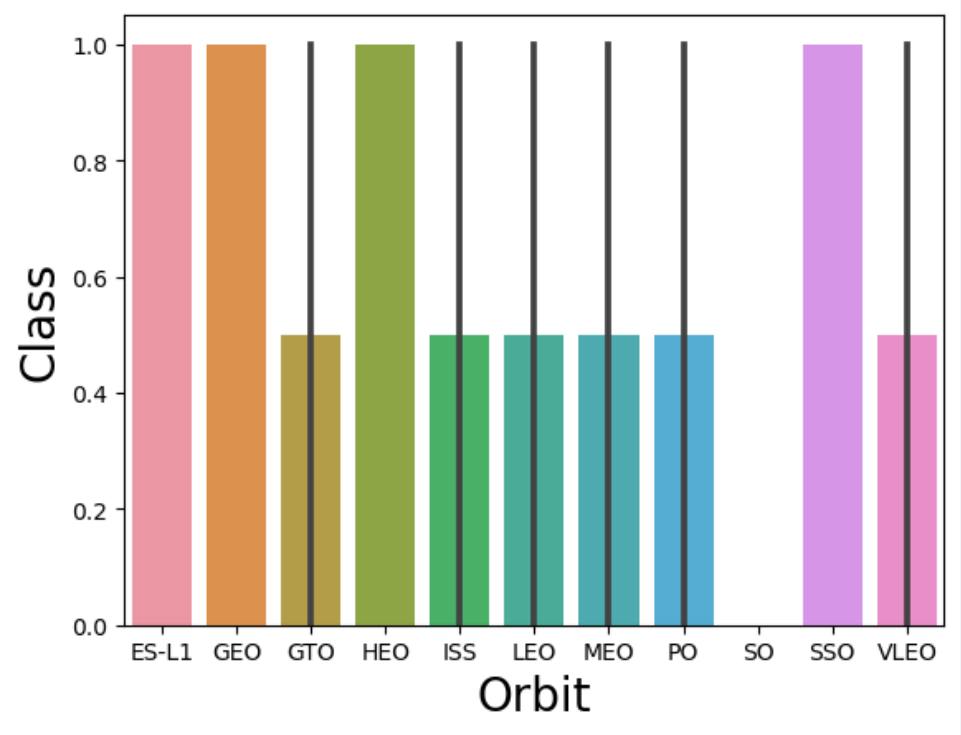
- The chart shows the relationship between Payload and Launch Site

- Except for 2 launches, all launches at all sites with a payload of more than 7500 kg were successful
- CCAFS SLC 40 and KSC LC 39A have launches with very low and high payload masses while VAFB SLC 4E is used for launches with medium pay load masses



Success Rate vs. Orbit Type

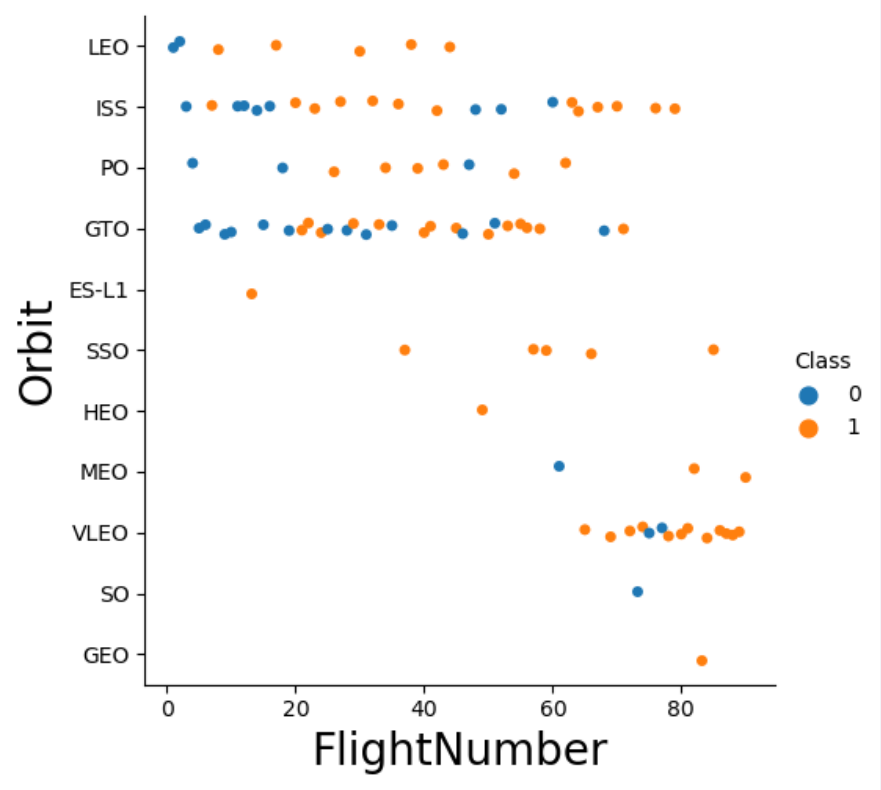
- The chart shows the relationship between success rate of each orbit type
- The orbits ES-L1, GEO, HEO, and SSO have only successful launches
- The success rate of the remaining orbits is at about 50%



Flight Number vs. Orbit Type

- The chart shows the relationship between Flight Number and Orbit type

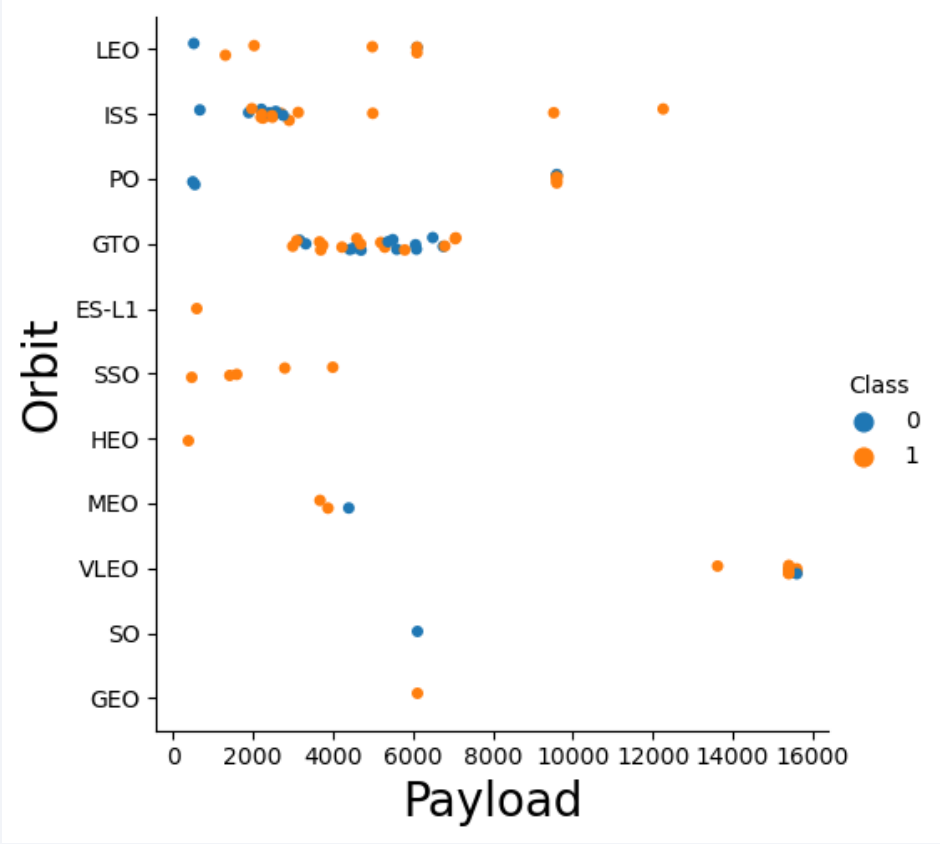
- First attempts of orbits seem to be unsuccessful while the rate increases over time



Payload vs. Orbit Type

- The chart shows the relationship between Payload and Orbit type

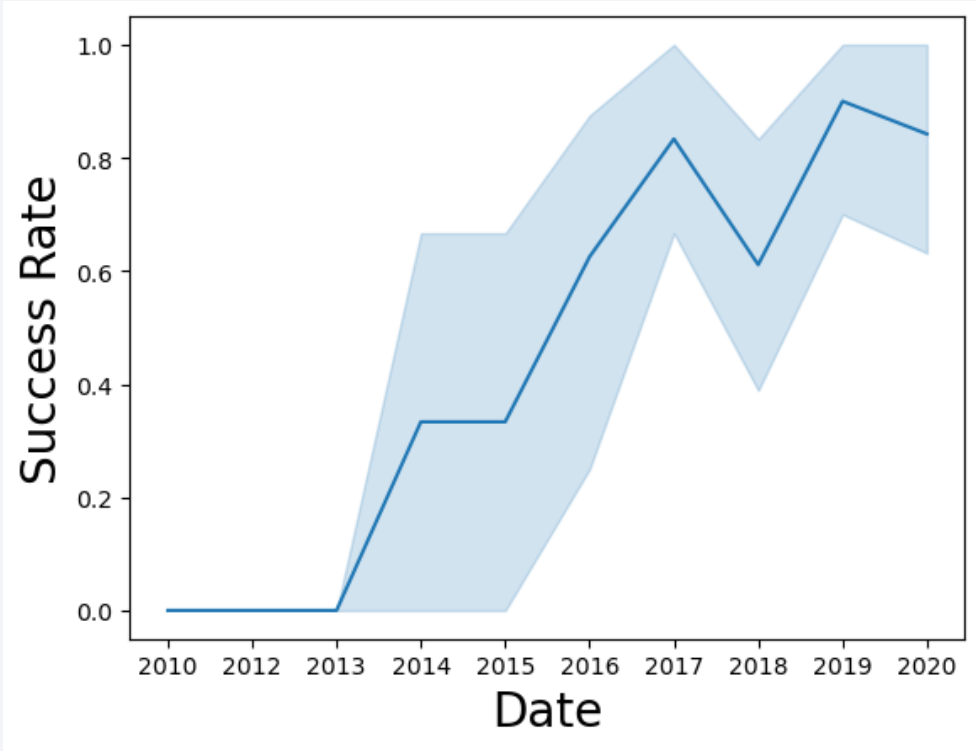
- The orbit VLEO can take a very high payload mass
- The remaining orbits launched with a variety on payload masses
- ES-L1 and HEO had only 1 launch



Launch Success Yearly Trend

- The chart shows the launch success yearly trend

- The success rate increased over time
- The first increase started in 2013
- In 2015 a second increase can be seen until 2017
- While the success rate dropped in 2017, it increased a little bit again
- The last reported success rate is at about 80%



All Launch Site Names

- In sum, the following four launch sites exist

	LAUNCH_SITE
0	CCAFS LC-40
1	CCAFS SLC-40
2	KSC LC-39A
3	VAFB SLC-4E

Launch Site Names Begin with 'CCA'

- Following, 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
0	2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	2010-08-12	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2	2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
3	2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt
4	2013-03-12	22:41:00	F9 v1.1	CCAFS LC-40	SES-8	3170	GTO	SES	Success	No attempt

- The data is filtered for missions at launch sites beginning with “CCA”
  - All refer to CCAFS LC-40

Total Payload Mass

- The total payload mass for boosters launched by NASA (CRS) is 22007 kg

	SUM_PAYLOAD_FROM_NASE
0	22007

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 3676 kg

	AVG_PAYLOAD_FROM_F9
0	3676

First Successful Ground Landing Date

- On 01.05.2017 the first successful landing outcome on ground pad

	MIN_DATE_SUCCESS_GROUND_PAD
0	2017-01-05

Successful Drone Ship Landing with Payload between 4000 and 6000

- Boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are shown in the following

	BOOSTER_VERSION
0	F9 FT B1022
1	F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes is shown in the following:

	MISSION_OUTCOME	AMOUNT
0	Success	44
1	Success (payload status unclear)	1

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass are listed in the following

	BOOSTER_VERSION
0	F9 B5 B1048.4
1	F9 B5 B1049.4
2	F9 B5 B1049.5
3	F9 B5 B1060.2
4	F9 B5 B1058.3

2015 Launch Records

- The only failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 is shown in the following

	MONTH_NAME	LANDING_OUTCOME	BOOSTER_VERSION	LAUNCH_SITE
0	OCTOBER	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40

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

- The ranked the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order, is shown in the following
- There are only two types of landing outcomes between the named dates and both indicate the same occurrence amount

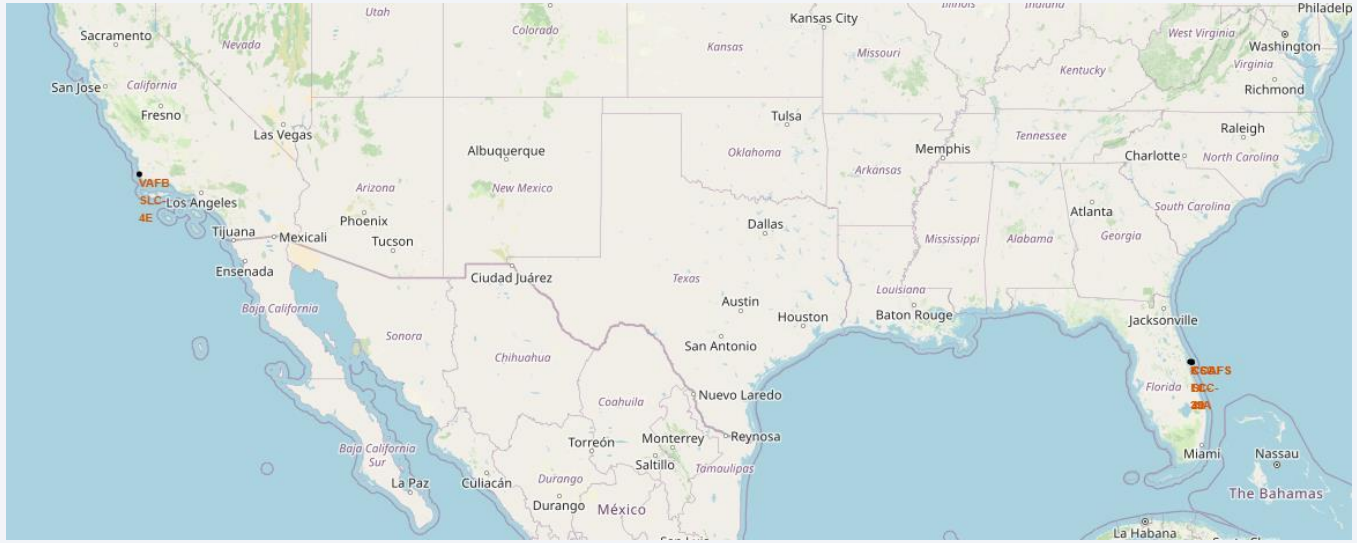
	LANDING_OUTCOME	COUNT
0	Success (drone ship)	2
1	Success (ground pad)	2

Section 3

Launch Sites Proximities Analysis

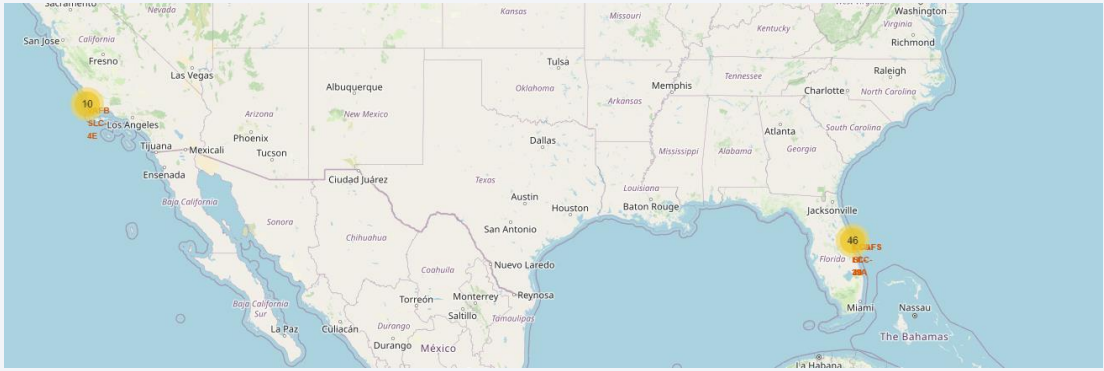
Launch Sites Locations

- The launch sites are located near the water on the east and west coast of the US

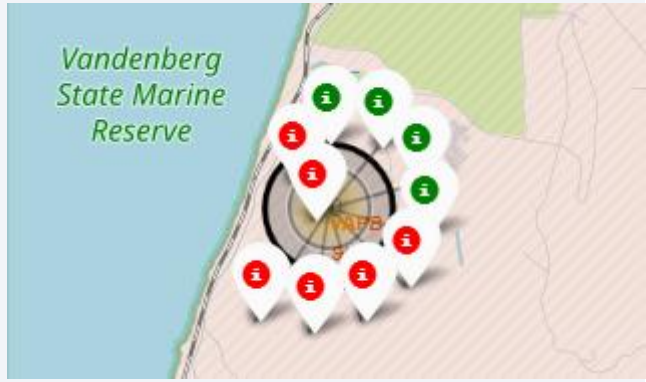


Success of Launch Sites

- Most launches are done in the east cost (46), while only 10 launches are recorded in the west coast



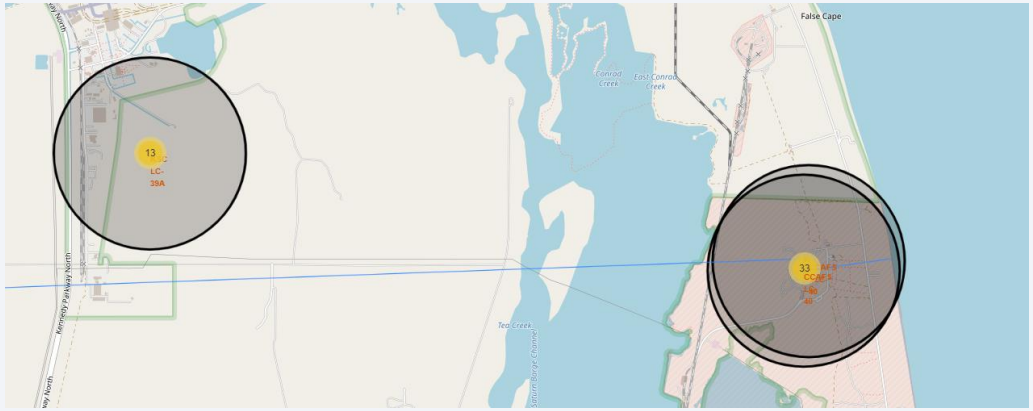
- At the west coast 6 of the 10 had a successful landing outcomes

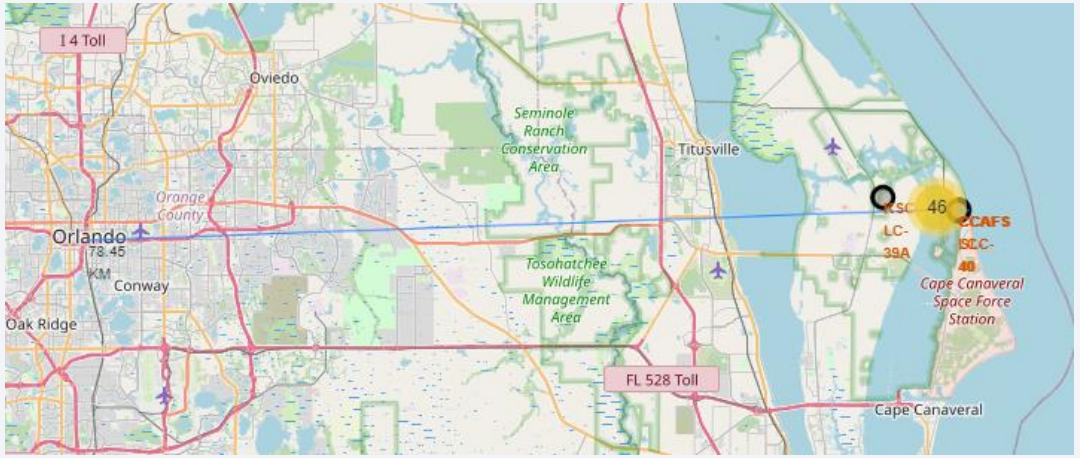


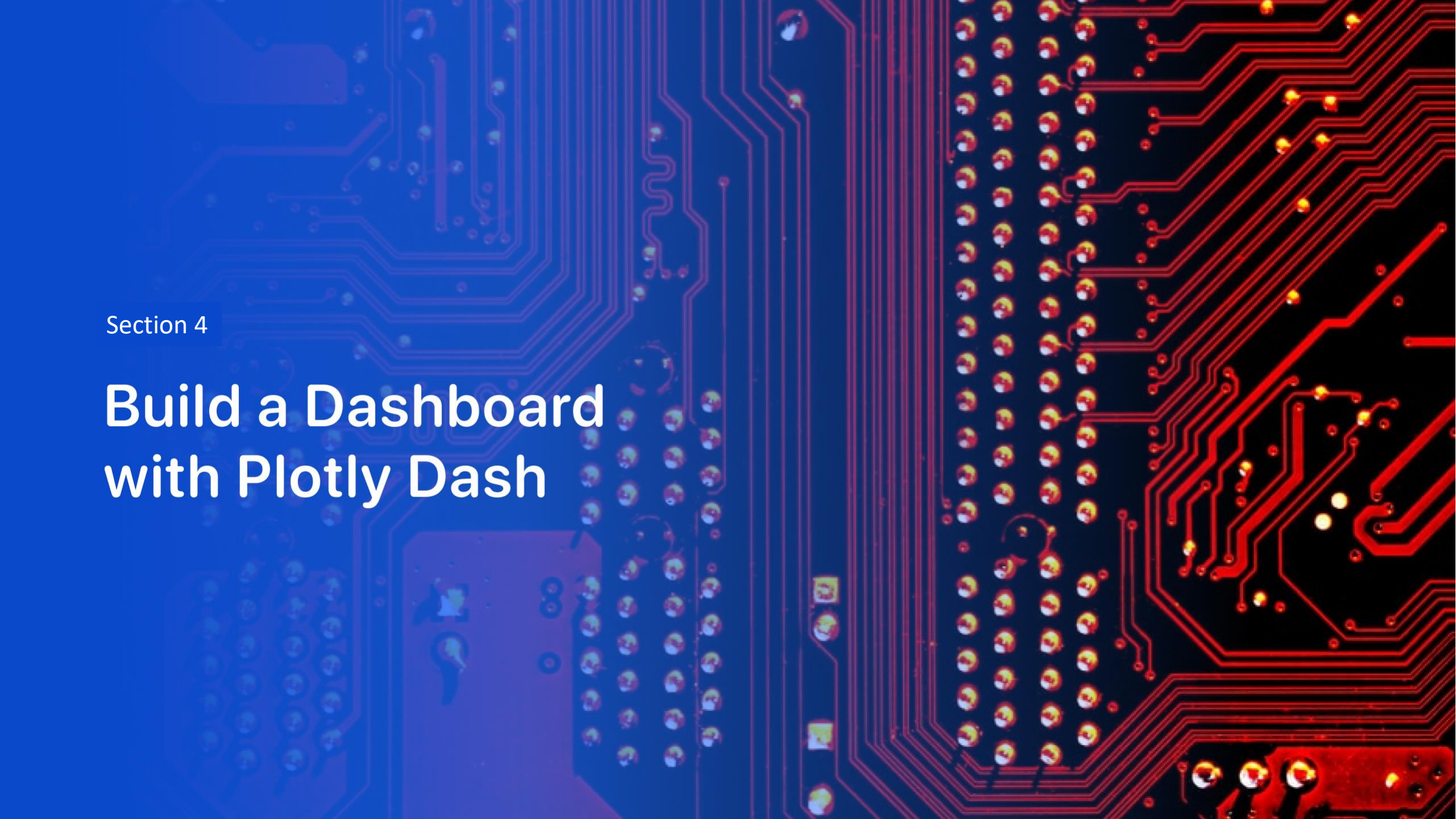
- At the east side the launches have started from different points

Infrastructure

- At the east, the Launch Sites KSC LC 39 A and CCAFS SLC 40 can be seen
- These launch sites are near to the city of Orlando which has an airport
- The launch sites also are near to the railway
- All in all, the launch site in the east has a good infrastructure

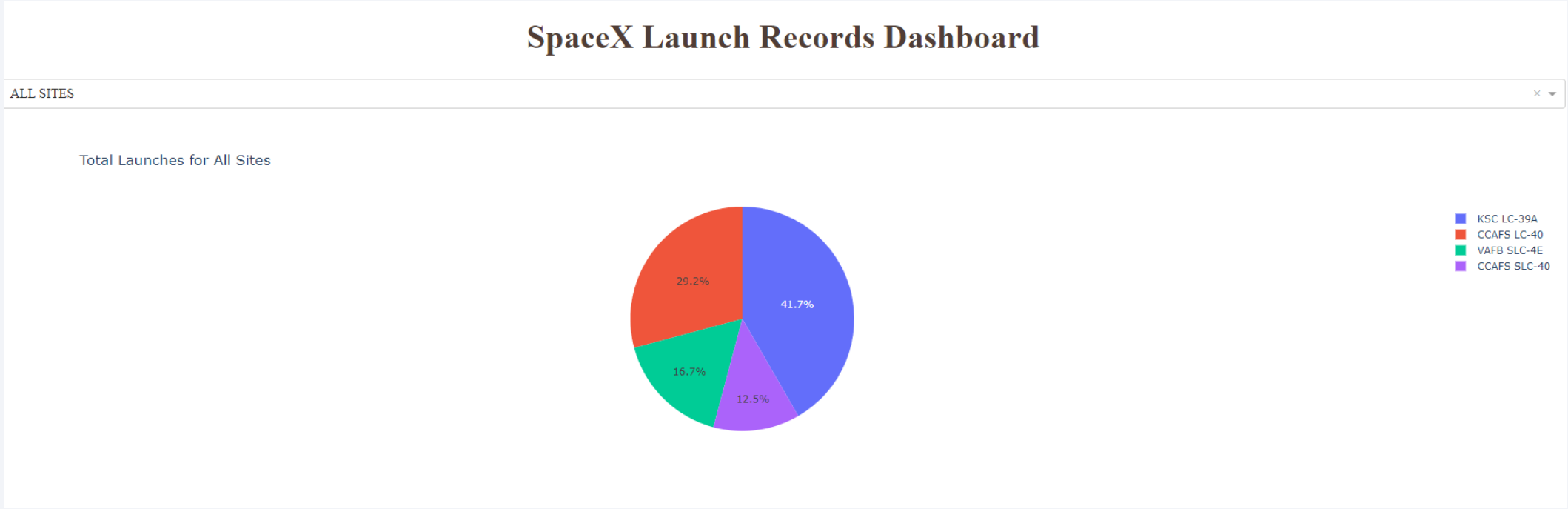





Section 4
Build a Dashboard with Plotly Dash

Total Launches and Success Rate per Site

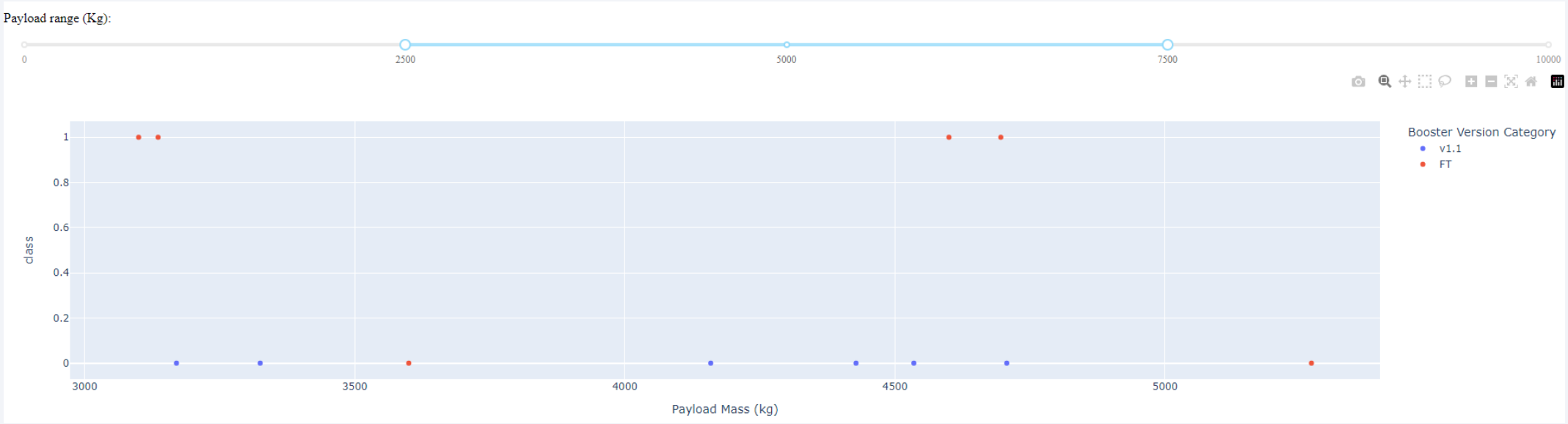
- This pie chart shows the launch success ratio for all sites
- KSC LC 39A has the highest ratio (41,7%)and seems to be the mot important one
    - It is followed by CCAFS LC-40, VAFB SLC-4E and CCAFS SLC-40



Payload vs. Launch Outcome

- The following scatter plot shows thePayload vs. Launch Outcome scatter plot for all sites, with a payload between 2500 and 7500 kg

- The booster version FT indicates a success at four different payload masses
- V1.1 shows no success at all

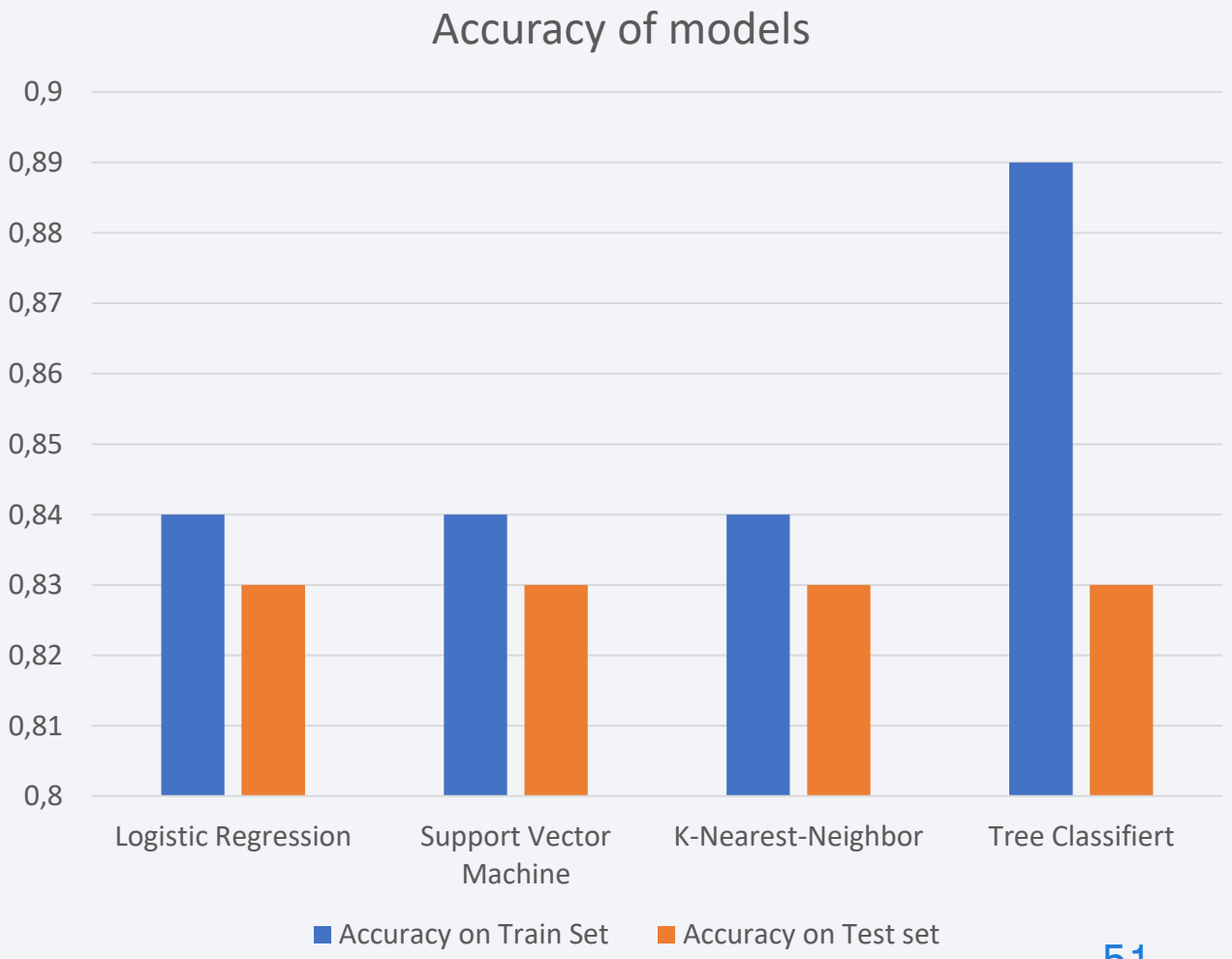


Section 5

Predictive Analysis (Classification)

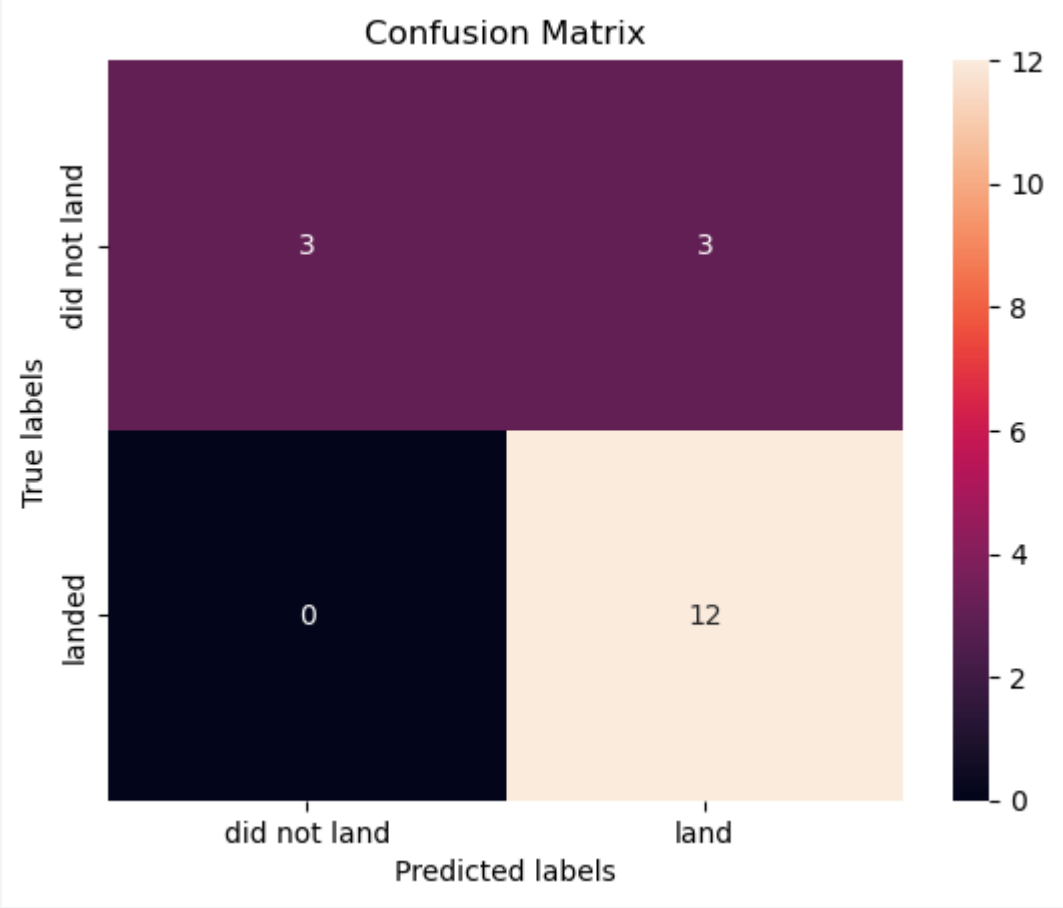
Classification Accuracy

- All models have the same accuracy on the test set
- Tree-Classifier has the highest accuracy on the train set
- As a conclusion, Tree-Classifier is identified as the best model



Confusion Matrix

- On the right side, the confusion matrix of the Tree-Classifier (showing the TP, FP, TN, FN):
- On the set set:
  - 15 of the 18 predictions are predicted correctly
  - 3 of 6 “did not land” are predicted corrected, while the remaining 3 are predicted falsely as “land”
  - All 12 “land” instances are predicted correctly



Conclusions

- SpaceX data is collected successfully from different sources via REST API and Web scraping
- The data is wrangled for further exploratory data analysis as well as prepared for predicting landing outcomes
- A total of 4 launch sites, locates in the east and west are identified
  - The Launch Sites at the east cost have more launches than the one at the west coast
  - Most launch happen at the launch site “CCAFS SLC 40”
- The first successful landing is reported in 2015
  - Overall, over time the success rate increase
- There are eleven types of Orbits
  - The orbit “GTO” has the highest number of occurrences
- Compared with Logistic Regression, SVM, and KNN → Tree-Classifiers show the best performance when predicting the landing outcome

Appendix

- All relevant Jupyter Notebooks, including the SQL statements, as well as the source code for the Plotyly Dashboard can be found at:
  - https://github.com/erkinaltuntas/applied_data_science_capstone

Thank you!